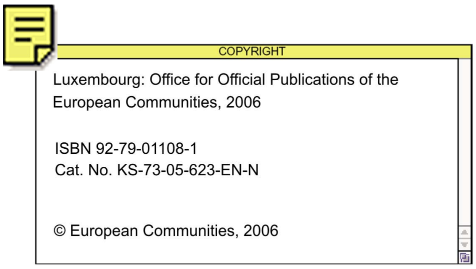
RESEARCH IN OFFICIAL STATISTICS

2006 EDITION

Monographs of official statistics

Work session on statistical data confidentiality

Geneva, 9-11 November 2005



COPYRIGHT

Luxembourg: Office for Official Publications of the European Communities, 2006

ISBN 92-79-01108-1
Cat. No. KS-73-05-623-EN-N

EUROPEAN COMMISSION

UNITED NATIONS
ECONOMIC
COMMISSION
FOR EUROPE

eurostat

THEME
General
and regional
statistics

A great deal of additional information on the European Union is available on the Internet.
It can be accessed through the Europa server (http://europa.eu.int).

Luxembourg: Office for Official Publications of the European Communities, 2006

ISBN 92-79-01108-1

EUROSTAT
L-2920 Luxembourg — Tel. (352) 43 01-1 — www.europa.eu.int/comm/eurostat/

Eurostat is the Statistical Office of the European Communities. Its task is to gather and analyse figures from the different European statistical offices in order to provide comparable and harmonised data for the European Union to use in the definition, implementation and analysis of Community policies. Its statistical products and services are also of great value to Europe's business community, professional organisations, academics, librarians, NGOs, the media and citizens.

To ensure that the vast quantity of accessible data is made widely available and to help each user make proper use of the information, Eurostat has set up a publications and services programme.

This programme makes a clear distinction between general and specialist users and particular collections have been developed for these different groups. The collections *Press releases*, *Statistics in focus*, *Panorama of the European Union*, *Pocketbooks* and *Catalogues* are aimed at general users. They give immediate key information through analyses, tables, graphs and maps.

The collections *Detailed tables* and *Methods and nomenclatures* suit the needs of the specialist who is prepared to spend more time analysing and using very detailed information and tables.

As part of the new dissemination policy, Eurostat has developed its website. All Eurostat publications are downloadable free of charge in PDF format from the website. Furthermore, Eurostat's databases are freely available there, as are tables with the most frequently used and demanded short- and long-term indicators.

Eurostat has set up with the members of the 'European statistical system' a network of support centres which will exist in nearly all Member States as well as in some EFTA countries. Their mission is to provide help and guidance to Internet users of European statistical data. Contact details for this support network can be found on our Internet site.

Eurostat

Table of contents

Acknowledgements

The work session on statistical confidentiality was jointly organized by Ms. Maria João Santos from Eurostat Unit for Methodology and Research and Mr. Juraj Riecan from UN Economic Commission for Europe.

UNECE and Eurostat gratefully acknowledge the valuable contributions of all the participants and particularly the session organizers and discussants, who were the following: Mr. Heinrich Brüngger (UNECE), Mr. Pedro Díaz Muñoz (Eurostat), Mr. Lawrence H. Cox (United States of America), Mr. Josep Domingo-Ferrer (Rovira i Virgili University, Spain), Ms. Luisa Franconi (Italy), Ms. Sarah Giessing (Germany), Mr. Anco Hundepool (Netherlands), Mr. Jean-Marc Museux (Eurostat) and Mr. Eric Schulte Nordholt (Netherlands).
Special thanks to Mr. Anco Hundepool for chairing the session.

The proceedings have been produced by Ms. Maria João Santos and Ms. Aleksandra Bujnowska from Eurostat Unit for Methodology and Research.

Foreword

Statistical confidentiality primarily aims at safeguarding privacy in the field of statistics and is a key to the necessary trust that has to be maintained between statistical bodies and respondents. Mutual confidence ensures accurate and reliable basic information and eventually high quality statistics.

There is a growing appreciation of the benefits of providing access to microdata for research and analysis. At the same time it is vital to protect data confidentiality. It is essential that new approaches are developed at international level to meet these objectives which create conflicting pressures. The risks to confidentiality must be managed effectively. A key challenge is how to minimise the risks to confidentiality, including the perception of threats to confidentiality. Striking the right balance is vital.

The work session covered a wide range of different aspects of statistical confidentiality from remote access to risk management by adequate access procedures to microdata.

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

(i) Web/on-line remote access (techniques, confidentiality protection and organizational issues);
(ii) Disclosure risk, information loss and usability of data;
(iii) Confidentiality aspects of statistical information taking into account register-based data;
(iv) Access to business microdata for analysis;
(v) Confidentiality aspects of tabular data, frequency tables, etc.;
(vi) Software for statistical disclosure control;
(vii) General statistical confidentiality issues (legal framework, political and conceptual aspects, terminology).

Papers presented under topic (i) focused on 2 types of access: remote execution, which is less flexible but provides better disclosure control and where all outputs are checked; and remote access, which is more flexible but disclosure control is more difficult and final output is checked.

The discussion on topic (ii) focused on the release of microdata files that may lead to risk of disclosure. The participants discussed several methods for assessing disclosure risk, as a crucial element of disclosure control and stressed the importance of statistical models.

In topic (iii) aspects of statistical disclosure control were discussed in the presence of accessible registers and archives that may permit re-identification of records.

Several methods were discussed in topic (iv) for secure computation that may allow sharing business data without compromising data confidentiality. These methods included secure summation protocols, secure matrix product protocols, and synthetic data approaches.

Papers presented under topic (v) discussed several methods to protect tabular data from rounding, peturbative methods such as controlled tabular adjustment or the use of fixed intervals as an alternative to cell suppression.

In session (vi) software solutions covering the entire field of statistical disclosure control were presented, amongst others: method producing safe output for complex statistical analysis in a remote access environment, algorithm for controlled tabular adjustment, SUDA program for classifying cell according to their disclosure risk, use of τ-Argus software for cell suppression.

Papers presented in topic (vii) discussed matters such as the balance that needs to be found between the need to provide users with access to microdata and the need to protect the confidentiality of respondents, legal and administrative procedure as part of risk management, harmonization of SDC methods and procedures on international level and production of data confidentiality and microdata access guidelines.

The work session was a great opportunity for official statisticians and researchers to exchange ideas and discuss new methods and tools dealing with confidentiality. The papers presented hereafter constitute a very important contribution to the development of applied procedures in this domain.

Pedro Díaz Muñoz

Heinrich Brüngger

Topic I

Web/on-line remote access

New developments in the Danish system for access to micro data

Lars Borchsenius
Head of Division, Research Services, Statistics Denmark

Summary

To facilitate registerbased research Statistics Denmark has given researchers access to de-identified micro data. The scheme has been changed from on-site to remote access through the Internet. Through the new scheme Danish researchers have a unique possibility to use micro data in their research.

The paper presents the background concerning the relevant legislation, the confidentiality principles of Statistics Denmark and the organisational framework.

The paper presents the new rules for access to micro data. According to these rules access to micro data can be granted to researchers and analysts in authorised environments. The use of micro data has increased markedly under the new rules. As at 25 August, 2005, 132 such environments have been authorised with over 300 active researchers.

Furthermore researchers in an authorised environment have been given (June 2005) the possibility of remote access to micro data from their private address when certain conditions are fulfilled.

The paper finally gives a presentation of the new technical solution.

1. From surveys to registerbased statistics

Denmark introduced the Person Number (the Personal Identification Number) in 1968 and it was used in a census for the first time at the Population and Housing Census in 1970. Accordingly, this became the first Danish register that uses the Person Number as an identification key. During the 1970s the first attempts were made to base the production of statistics on registers. In 1976 a register-based population census was conducted as a pilot project, but the registers were not sufficiently comprehensive and well-established until 1981, when a proper register-based population census was conducted containing most of the conventional population and housing census information.

Like in the other Nordic countries, the person and business registers in Denmark today cover a very substantial part of the production of statistics. The contents of the registers also cover many fields of research such as labour market research, sociology, epidemiology and business economics. The strength of the system is that the identification keys (person number, address, central business register number and property title number) render it possible to correlate the aggregated data both within a specific year and longitudinally across several years.

2. Increased interest in micro data

In the mid-1980s, Statistics Denmark experienced an emerging interest among various research environments and ministerial analysis divisions in applying micro data (individual data) for research and analysis purposes. One reason was that the development in computer technology made it technically possible to process large amounts of data according to advanced statistical models, such as multivariate models.

These environments put pressure on Statistics Denmark to disclose micro data; a request that Statistics Denmark was unable to grant because of the rules of confidentiality lay down by the Management and Board of Statistics Denmark. On the other hand, it was evident already at that time that not only were the registers of enormous importance to he production of statistics by Statistics Denmark, but their research potential was so great that it would be very valuable to actually utilise them for research purposes. Therefore, Statistics Denmark had to find a solution to the problem of access, which complied with the existing legislation on registers while taking into account Statistics Denmark's own confidentiality principles.

During 2001 negotiations between Statistics Denmark, the Ministry of Research and the Research environment resulted in a signing a contract on the establishment of a special unit (the Research Service Unit) in Statistics Denmark with the special duty to improve researchers access to micro data through a better infrastructure and to lower the costs of using the data.

3. Legislation

With the introduction of two acts on registers in 1979, Denmark saw the first statutory regulation concerning, inter alia, disclosure of micro data to researchers. As at 1 July 2000 these acts were replaced by the Act on Processing of Personal Data (lov om behandling af personoplysninger). The Act implements Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and the free movement of such data within the European Union. The former Act primarily governed registration and disclosure of data in registers, while the new Act applies to all forms of processing of personal data. The new term, "processing", covers all types of processing of personal data, including registration, storing, disclosure, merging, changes, deletion, etc.

Previously, the setting up of a register was subject to the so-called register provisions, involving a rather time-consuming and laborious process. These provisions have been abolished, and now the individual authority makes decisions in concrete cases on processing; for example, the authority decides issues of disclosure of data for scientific purposes based directly on the provisions of the Act on the lawfulness of such disclosure.

The new Act introduced a duty of notification to the Danish Data Protection Agency. The purpose is to enable the Agency to supervise the processing of sensitive information carried out.

Accordingly, a scientific project involving processing of sensitive personal data is subject to notification to and approval by the Danish Data Protection Agency before such processing can commence. This applies to all surveys, whether they are conducted by a public administration, individuals or enterprises. The Agency has laid down special provisions on security in connection with the processing of sensitive data.

All in all, the introduction of the Act on Processing of Personal Data has provided potentially more favourable conditions for register-based research in Denmark. In particular, public authorities' basis for disclosing administrative data for research purposes has been enhanced and simplified in terms of administration, as they no longer need to consult the Danish Data Protection Agency; personal data applied for statistical purposes may be disclosed and reused with the permission of the Agency; data from one private research project may be disclosed to another project; there is full access to filing of data in the State archives; both private individuals and public authorities may process data on Person Numbers for scientific or statistical purposes; furthermore, the Act now explicitly stipulates that the data subject's right of access to personal data shall not apply where data are processed solely for scientific purposes.

In addition to the Act on Processing of Personal Data, the Danish Public Administration Act (Forvaltningsloven) is of relevance. Under this Act, a public authority may impose a duty of non-disclosure on persons outside the public administration concerning the data disclosed. Statistics Denmark has applied this provision in connection with researchers' access to micro data, although no disclosure in a formal sense is made. Data - even anonymised data - must be treated as confidential. Breach of the duty of non-disclosure is punishable by simple detention or imprisonment.

4. Confidentiality principles of Statistics Denmark

As it appears from the above, current legislation permits disclosure, to a wide extent, of personal data for scientific purposes. However, the authority in question ultimately decides whether disclosure may take place, meaning that the authority may take other issues into consideration even if the Danish Data Protection Agency has approved the disclosure of data.

That is what Statistics Denmark has decided to do. This decision has been made so that the individual citizen or enterprise can be certain that the data supplied directly or indirectly to Statistics Denmark do not fall into the hands of any unauthorised persons. In the opinion of Statistics Denmark the risk of irreparable damage to the production of statistics outweighs the consideration for more or less convenient access to data by the individual researcher.

Thus, the fundamental principle is that data must not be disclosed where there is an imminent risk that an individual person or individual enterprise can be identified. This does not only apply to identified data, such as Person Numbers, but also to de-identified data, since such data are usually so detailed that identification can be made.

Since Statistics Denmark also considers it important that data can be applied for scientific purposes, special schemes for researchers have been set up.

5. Changing of scheme from on-site arrangement for external researchers at Statistics Denmark to remote access through the Internet

Since its overriding principle is not to disclose individual data, Statistics Denmark set up a scheme in 1986 for the on-site arrangement for external researchers at Statistics Denmark. Under this scheme, researchers got access to anonymised register data from a workstation at the premises of Statistics Denmark. Statistics Denmark creates the relevant datasets on the basis of the researcher's project description, the general principle being that the dataset should not be more comprehensive than necessary for carrying out the project (the "need to know" principle). The researcher signs an agreement which stipulates that data are confidential and that individual data must not be removed from the premises of Statistics Denmark.

From 2001 the users of Statistics Denmark's researcher schemes has been given access to datasets from their own workplaces. The permission is restricted to specially authorised research and analysis environments. Furthermore researchers in an authorised environment has been given (June 2005) the possibility of remote access to micro data from their private address when certain conditions are fulfilled.

A research or analysis environment can apply for an authorisation from Statistics Denmark. As at 25 August, 2005, 132 environments had been granted authorisation. The wording of the authorisation appears from Appendix 1.

Until now the remote access has not been granted for all datasets; particularly data on enterprises are assessed carefully to avoid any problems of confidentiality. It has been emphasised that the data consist of samples. If the researchers request access to total populations, the content of variables must be limited.

With this new development the rules for granting authorisation to micro data are of course of outmost interest.

6. Rules for access to micro data

Access to micro data can only be granted to researchers and analysts in authorised environments.

Authorisations can be granted to public research and analysts environments (e.g. in universities, sector research institutes, ministries etc) and to research organizations as a part of a charitable organization.

Within the private sector following user groups can be granted authorisation if they have a stable research or analyst's environment (with a responsible manager and with a group of researchers/ analysts):

1. Nongovernmental organisations
2. Consultancy firms
3. Enterprises. However single enterprises can not have access to micro data with enterprise data

In order to grant an authorisation, Statistics Denmark will evaluate the proposed organization carefully and especially when it is an organization or firm within the private sector Statistics Denmark will look at credibility of the applicant (as ownership, educational standard among the staff and the research done for others).

Statistics Denmark will not grant authorization to single persons. Furthermore Media organizations are excluded from the scheme.

The "need to know" principle is still in force.

Researchers can have access to relevant business data after the "need to know" principle. Only very few business data are excluded from remote access.

7. Foreign researcher?

Only Danish research environments are granted authorisation as Statistics Denmark is not able effectively to enforce a contract abroad. Foreign researchers from well established research centres can have access to Danish micro data from the on-site arrangement in Copenhagen or Århus. Visiting researchers can have remote access from a workplace in the Danish research institution during their stay in Denmark and under the Danish authorisation.

8. Organisational framework

The scheme is administered centrally by the Division of Research Services. The staff of this unit also create a substantial part of the interdisciplinary datasets and have a general (authorized) access to all relevant data in Statistics Denmark in order to reduce the administrative and bureaucratic work. The scheme requires close cooperation between the Division of Research Services and the individual divisions. The advantage of such central organisation is that the individual researcher is fully aware of whom to negotiate with and who is responsible for the dataset supplied.

In 1996, Statistics Denmark opened a small branch in Århus, Jutland, to grant researchers west of the Great Belt an opportunity to use the scheme on equal terms with researchers in Copenhagen.

9. Research databases

As the researchers almost invariably request datasets linking information from several individual registers in terms of both contents and time, the creation of specific datasets for a project often involves considerable work by Statistics Denmark and often considerable costs for the researcher.

To reduce the cost of datasets for research purposes and solve special data problems, Statistics Denmark has set up a number of research databases. These databases are hardly ever used in the actual production of statistics, but are first and foremost a kind of intermediate products for the benefit of the research process.

The most frequently applied research database is the Integrated Database for Labour Market Research (IDA). One reason for creating the database was to solve a difficult problem of definition: Identity of enterprises over time, a task that individual researchers were unable to handle for reasons of both time and funding. Nine to ten man-years were spent on the task, which was funded by the Danish Social Science Research Council (Statens Samfundsvidenskabelige Forskningsråd) and Statistics Denmark. Since the establishment of IDA, Statistics Denmark has handled the updating of the database against user charges.

Other research databases include the Demographic Database, the Fertility Database, the Prevention Register (health data), the Social Research Register, etc. As the names imply, the databases cover many specialist fields: economy, labour market research, social research, medicine, epidemiology, etc. The last development, where the number of users is growing at a rapid rate, is the Register of Medicinal product Statistics holding information on doctors' prescriptions of medicine sold by the pharmacies in Denmark.

10. Considerable growth

From the modest beginnings in 1986, the use of micro data has increased markedly for researchers at Statistics Denmark. In 1997, 71 researchers used the on-site arrangement, while in 2005 under the scheme for remote access through the Internet the figure had risen to more than 300.

11. Study datasets

Statistics Denmark has prepared some study datasets, so far based on the IDA database, for study programmes in economics/labour market policy and interdisciplinary data material for sociology studies. These datasets follow a few thousand persons over time according to a number of variables. Where possible, the data are scrambled so that the actual register data have been changed in ascending or descending order by a simple mathematical function. However, the fundamental characteristics of the data have been preserved. In this way, students get an opportunity to try out statistical models on realistic data.

Except for the above, Statistics Denmark has not applied scrambling procedures or special grouping techniques to the data that are made available to the researchers. The data appear as in the basic registers.

12. The technical solution

The technical solution is based on the use of the Internet conf. the flow chart at the end of this paper.

The relevant micro data are produced by the staff in Statistics Denmark and the de-identified micro data are transferred to the disk storage connected to the special Unix servers. These Unix servers are only used by researchers and are separated from the production network.

Communications via the Internet is encrypted by means of a so-called RSA SecurID card, a component that secures Internet communications against unauthorised access. In practice the researcher rents a password key (a token) from Statistics Denmark. The token ensures that only the authorised person obtains access to the computer system.

A farm of Citrix Servers ensures that the researchers from their own workplace can “see” the Unix environment in Statistics Denmark. All data processing is actually done in Statistics Denmark and data cannot be transferred from Statistics Denmark to the researcher’s computer. The researcher can work with the data quite freely and can make new datasets from the original data sets. The limit is of course the amount of disk space. Statistics Denmark has just increased the total amount of disk space considerably.

All results from the researchers computer work can be stored in a special file and such printouts are sent to the researchers by e-mail. This is a continuous process (every five minutes) and has shown to be quite effective. The advantage to Statistics Denmark is that all e-mails are logged at Statistics Denmark and checked by the Research Service Unit. If the unit find printouts with too detailed data, contact is taken to the researcher in order to agree on details of the level of output. No severe violation of the rules, establish in the authorisation formula, has taken place.

References

Otto Andersen: From on-site to remote data access, Contributed paper to the Joint ECE/Eurostat work session on statistical data confidentiality (Luxembourg, 7-9 April 2003).

Appendix 1

Statistics Denmark

AUTHORISATION

Statistics Denmark hereby grants

[Institution] represented by [Chief Researcher]

Authorisation for

Remote electronic access to selected datasets at Statistics Denmark

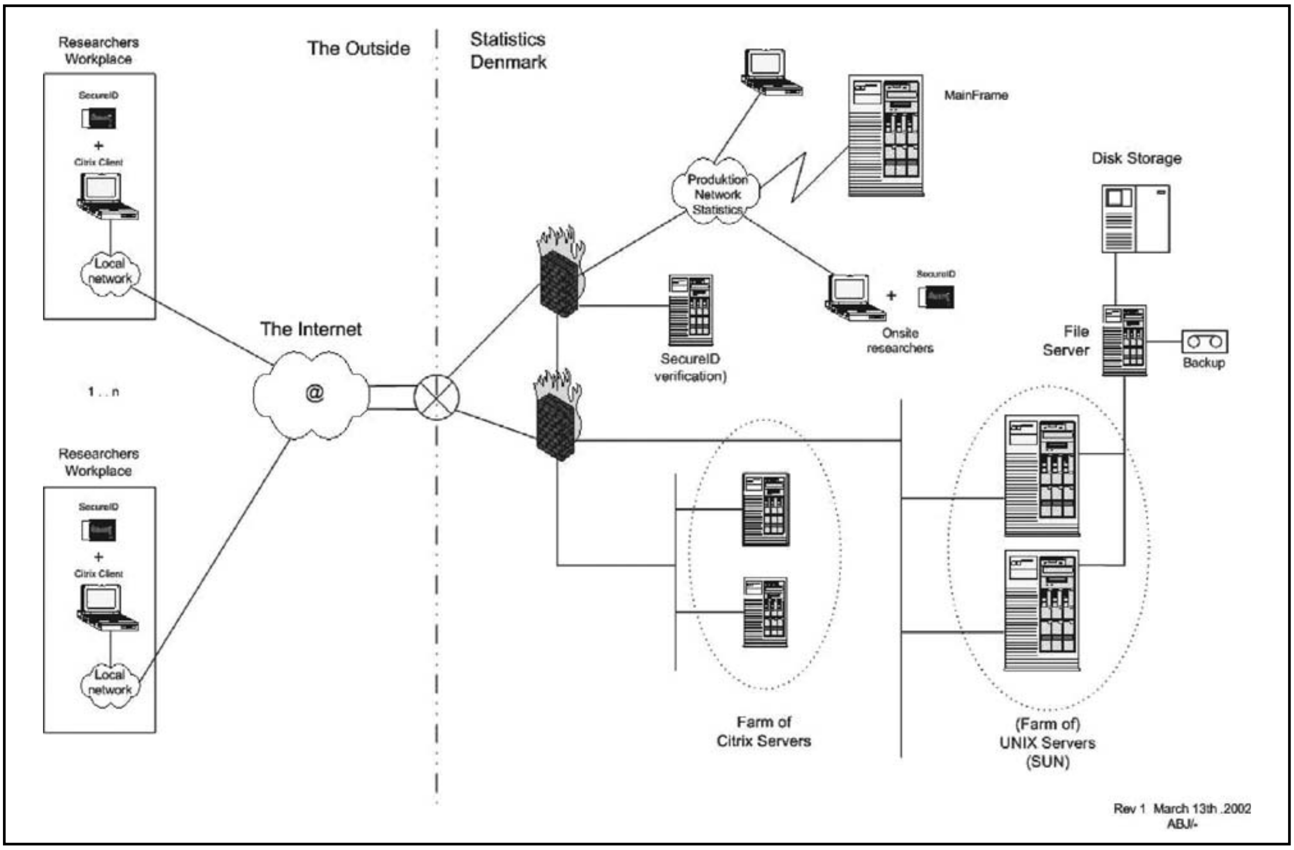
Remote Access via the Internet is subject to the following terms:

1. A project description must be submitted, which states the project objectives and renders it possible to select the data required for successful project execution.
2. Based on the project and data description, Statistics Denmark decides whether external electronic access to data can be granted for the specified project. If the authorisation is not granted, the researcher is referred to use the ordinary scheme for the on-site arrangement for external researchers at Statistics Denmark.
3. The researcher to whom external electronic access is granted shall sign a special agreement with Statistics Denmark, cf. appendix.
4. All datasets are confidential, cf. §27(3) of the Danish Public Administration Act and §152 of the Danish Criminal Code.
5. The researcher obtains access to make batch runs on Statistics Denmark's special researcher machines (UNIX system) from one or more PCs specially assigned for that purpose in the research/analysis environment. Access is denied for batch runs from remote PCs, PCs at home or PCs which cannot be properly supervised.
6. Only the client software assigned by Statistics Denmark may be applied in connection with the RSASecurID card provided. A PC connected to Statistics Denmark may not be made available to unauthorised persons, and when the user leaves the PC, the PC must be either shut down or disconnected, i.e., protected from any unauthorised use.
7. The password of the individual researcher is personal and strictly confidential.
8. The researcher may not, directly or indirectly, download the dataset or any datasets derived there from. All transfers of output for printing or further statistical processing (in spreadsheets or similar) must be executed in accordance with the guidelines and methods laid down by Statistics Denmark. Statistics Denmark will create a log file of such authorised transfers. Furthermore, individual records may not be printed, and all output must be aggregated to an extent that eliminates any risk of direct or indirect identification of persons or enterprises. The researcher may not attempt to make such identification.
9. Statistics Denmark shall be entitled at unannounced visits to check that the rules of this agreement are observed.
10. The person signing this agreement on behalf of the research/analysis environment shall ensure that publications by the environment do not contain any information that may identify individual persons or individual enterprises.

11. The person signing this agreement on behalf of the research/analysis environment undertakes personally to supervise or to appoint a person to supervise that the provisions of this agreement are observed.
12. In case of breach of the provisions of this agreement, the researcher in breach will be excluded from using any researcher schemes of Statistics Denmark permanently or for a period of not less then three years. Furthermore, in the case of breach hereof, this authorisation will be withdrawn for a period.

This agreement, which is signed in two copies, enters into force on [date] and may be terminated by either party at three months' notice.

Remote Access to Statistics Denmark. January 2003. Principles of Operation



MONA - Microdata ON-Line access at Statistics Sweden

Claus-Göran Hjelm

Statistics Sweden, Research and Development, Register Coordination and Microdata Access

Keywords. Microdata, MONA, legislation, Remote Access

1. Introduction

Increased use of microdata requires improved possibilities of providing better data to meet the needs of users. It is vital for National Statistical Institutes (NSIs) to assure that the wealth of stored microdata can be fully utilised by researchers and other authorised users. By and large, the access to microdata means that investments made in official statistics give higher return.

Technological advances in hardware, software, data documentation and the Internet have already largely increased the possibilities to improve the access to microdata, but new possibilities appear every day. However, as the issue of confidentiality protection goes hand in hand with use of microdata, a balance is needed between use of microdata and confidentiality.

2. Legislation

Data confidentiality is guided by two major aspects which both are necessary equirements in order to meet the requests from researchers:

(1) general rules (guidelines, screening procedures, contracts, regulations and laws, etc.), and

(2) technical and practical measures for the same purpose.

The legislation concerning confidentiality and protection of individual's integrity is of importance for the possibility for the NSI to provide access to micro-data. The legislation provides the limits for release of data for e.g. research purposes and underpins and constitutes administrative and technical safeguards for legal founding. Specific legislation of importance is the Statistics Act and the Data Protection Acts. To this specific legislation, the current EU legislation with respect to statistical confidentiality should also be added.

2.1. EU legislation

The Council regulation (EC) No 322/97 of 17 February 1997 on Community Statistics contains rules that are important for the use of information collected for community statistics. According to the regulation data used by the national authorities and the Community authority for the production of Community statistics shall be considered confidential when they allow statistical units to be identified, either directly or indirectly, thereby disclosing individual information.To determine whether a statistical unit is identifiable, account shall be taken of all the means that might reasonably be used by a third party to identify the said statistical unit. Confidential data obtained exclusively for the production of Community statistics shall be used exclusively for statistical purposes unless the respondents have unambiguously given their consent to the use for any other purposes. However it is possible to allow access for scientific purposes to confidential data obtained for Community statistics.

Of importance for the processing, including release of data, is also the 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 (the Data Protection Directive). The object of the Directive is to strengthen data protection, e.g. the legal protection of individuals with regard to automatic processing of personal information relating to them. The Directive has been implemented in all the Nordic countries. The Directve applies to computerised personal data and personal data held in structured manual files. It applies to anything at all done to personal data processing. The new term, “processing”, covers all types of processing of personal data, including registration, storing, disclosure, merging, changes, deletion, etc. According to the Directive data must be: – Processed fairly and lawfully. – Collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes. However, further processing of data for historical, statistical or scientific purposes is not considered as incompatible.

- Adequate, relevant and not excessive in relation to the purposes for which they are collected and/ or further processed.
- Accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that data which are inaccurate or incomplete, with regard to the purposes for which they were collected or for which they are further processed, are erased or rectified.
- Kept in a form, which permits identification of data subjects for no longer than is necessary for the purposes for which the data were collected or for which they are further processed. Personal data can be stored for longer periods for historical, statistical or scientific use.

2.2. The legislation in the Nordic countries

The protection measures applied to confidential data obtained for statistical purposes are based on several legal acts and directives. However, it should be noted that access to statistical micro data for research or other purposes is a part of NSIs duty service and is not an obligation given by the law. In the Nordic countries there are specific and modern Statistics Acts regulating the use of statistical information. Icelandic law with regard to statistical information have not been updated since 1913. However, no specific legal acts deal with Statistics Iceland in cases of access to micro data. In that respect official statistics in Iceland take account of the general acts on data protection and more recently on acts for the protection of individual’s confidentiality. In general, given the lack of rules concerning access to micro data, Statistics Iceland is preparing guidelines in order to meet requests for micro data by external users. Statistics Iceland takes notice of two main aspects in cases of micro data: (1) Common rules internationally on good practices for handling of micro data, and (2) specific rules and practices of Statistics Iceland thus far.

Data collected for statistical purposes, in accordance with any prescribed obligation to provide information, or which is given voluntarily, may in principle only be used for the production of statistics. There are exceptions that enable access to data for research purposes and public planning. However, a condition for the use for research is that there is no incompatibility between the purpose of suchprocessing and the purpose for which the data was collected. The processing of data, which includes release of data, must also be in accordance with the regulation concerning protection of individual’s integrity.

Besides the Statistics Acts there are specific Personal Data Acts2 that apply to the production of statistics and the release of micro data. The Acts are based on the Data Protection Directive and contain rules about the fundamental requirements concerning the processing of personal data. These demands include, inter alia, that personal data may only be processed for specific, explicitly stated and justified purposes.

Very stringent rules apply to the processing of sensitive personal data. Sensitive personal data may be processed for research and statistics purposes, provided the processing is necessary and provided the public interest in the project manifestly exceeds the risks of improper violation of personal integrity. Furthermore in Denmark, Norway and Sweden processing of sensitive data for research purposes needs approval. A scientific project involving processing of sensitive personal data is in these countries subject to notification to and approval by the Data Inspection Agency before such processing can commence. This applies to all surveys, whether conducted by a public administration, individuals or enterprises. (In Sweden the approval of the National Data Inspection Agency is not necessary if a research committee has approved the processing.) If the Data Inspection Agency approves the processing, personal data may be provided to be used in research projects unless otherwise provided by the rules on confidentiality. This means that the NSI may take other issues into consideration even if the Data Inspection Agency (or research committee) has approved the processing of data. Data obtained for statistical purpose are declared as confidential, when they allow statistical units to be identified, directly or indirectly and thereby disclosing individual data. Also anonymous data can be confidential. Statistical data are confidential irrespective of source. Also, data taken from public administrative sources are confidential while in the possession of the NSI. The confidentiality rules are the same irrespective of whether data concerns individuals or enterprises. Under the main rules, access may be granted in forms which do not allow direct or indirect identification of people or other data subjects such as enterprises. However, confidential data may be released to a third party for the purpose of statistical surveys and scientific research. In Finland it is not generally possible to provide access to data when units can be disclosed directly or indirectly. According to the legislation in Denmark, Iceland, Norway and Sweden, statistical data may even be released with identification data for these purposes. In Finland personal data on a person's age, sex, occupation and education may in exceptional cases be released with identification data for research purposes. One condition in all countries is that access to confidential data for statistical or research purposes must not cause any damage or be detrimental to the data subjects. In practice this means that the NSIs only provide access to anonymous data or de-identified data.

The Nordic countries also have special public business registers that contain some common primary information about enterprises. These registers are (except in Denmark) administered by the NSIs and can also be used for other purposes than statistics or research.

When data has been collected in a voluntary survey the respondents in the statistical surveys must give consent to the release of the data.

It is the NSI that decides whether data may be released for research purposes. However in Norway access for other purposes than statistical must be approved by the Data Inspection Agency. The Agency has given general permission to Statistics Norway to provide access to micro data for research purposes and for public planning. The Data Inspection Agency may nevertheless make exceptions to such obligation of confidentiality for certain types of information if they find it in conflict with the Data Protection Act.

The obligation of confidentiality will also – according to the law or by imposition of a duty of non-disclosure – apply to the recipient of the data. The NSI may also impose a restriction limiting the researchers right to re-communicate or use the information. Breach of confidentiality restrictions is punishable by simple detention or imprisonment. In Sweden, however, it is not possible to impose restrictions when data are released to another authority. It is therefore important for Statistics Sweden to take into consideration if the data will be confidential according to the Secrecy Act also at the authority receiving data. If not, any one who so desires can have access to the data because of the authority's obligation under Chapter 2 of the Freedom of the Press Act to provide personal data that are not confidential. However, there are rules providing that confidentiality accompanies data to another authority in special situations e.g. if an authority, for research purpose, receives information from another authority where the data is confidential, the confidentiality will apply also within the receiving authority.

However, there are no such rules concerning release of data for statistical purposes or public planning.

In Finland a new Act and Decree on the Openness of the Government Activities came into force in 1999. This legislation contains comprehensive provisions on good practice on information management. For instance the Decree includes a detailed list of general data protection measures for confidential data. Statistics Finland like all the other government authorities has to implement these measures by the end of the year 2004.

3. Registers and microdata in the Nordic countries

The Nordic countries have a long tradition of collecting administrative data and transforming these data to registers suitable for statistical use. The production system and the statistical information system in the Nordic countries are to a great extent based on a number of large administrative registers. However, much is needed to transform administrative registers into high quality statistical registers. In addition, the register system also includes a number of survey-based registers, known as final observation registers (e.g. results from the Labour Force Surveys).

Microdata suited for researchers must be standardised and of high quality. The Nordic countries have compiled a number of integrated registers based on several registers and suitable for analyses and research purposes. The longitudinal integrated register "Louise" from Sweden containing anonymised microdata on individuals and families regarding their education, income and employment might serve as en example. This register includes annual data on all adults in Sweden from 1990 and is updated each year. Such an integrated database offers rich possibilities to carry out different analyses. In the future we see that via the Statistical Data Warehouse we can offer these types of integrated registers "on demand".

Over a number of years some of the Nordic NSIs (Norway, Sweden and Finland) have distributed anonymous microdata to a large number of research institutions and authorities using magnetic tapes, CD-Rom discs, DVD discs or other formats. The volume has increased at the same time as the number of releases/assignments has increased. Denmark has in the past only allowed access to microdata on-site at Statistics Denmark.

4. Confidentiality

Confidentiality protection of individual and business data is one of the main principles in official statistics and must be addressed when discussing microdata. The individual is entitled to be protected from unacceptable intrusion into personal privacy. At the same time the individual's need for protection must be balanced against legitimate needs for using information connected to society, such as for statistics and research. The legislation concerning confidentiality and protection of privacy of individuals is of importance for the possibility for the Nordic NSIs to provide access to microdata.

The use of statistical information is normally regulated in legislation and/or in a code of practice. In the Nordic countries there are specific legislations regulating the use of statistical information. According to these legislations, as a main principle data collected for statistical purposes, may only be used for the production of statistics. In addition, access can also be provided for research purposes and public planning. The processing of data, which includes release of data, must also be in accordance with the national regulation concerning protection of the individual's privacy and with the current EU legislation with respect to statistical confidentiality.

All data, including anonymous data, obtained for statistical purposes are confidential. Furthermore, statistical data are confidential irrespective of the source. According to the legislation in Sweden and in other Nordic countries, it is prohibited to disclose confidential data to unauthorised users. According to the main principle, confidential data may be released to a third party only for the purpose of statistical surveys and research. Access may only be granted in forms that do not allow direct or indirect identification of individuals or of other data subjects such as enterprises. In practice, the Nordic NSIs only provide access to anonymous data or microdata without name, address and identification number.

Regarding the use of microdata, legislation in the Nordic countries does not contain any specific rules that restrict the way of releasing microdata. As long as the general requirements in the legislation are fulfilled, the most suitable method can be chosen.

5. Recent developments

Some years ago several of the Nordic countries decided to improve access to microdata. A basic goal was to have a functional and secure way of providing microdata from Denmark and Sweden. Furthermore, such a system should be capable of handling large data sources securely for both the NSIs in Denmark and Sweden and the research community.

In 2004 Statistics Sweden formed a new organisational unit called Register coordination and Microdata access at the Department of Research and Development. Furthermore, a development project was started to investigate whether a new technology for remote access to microdata using Server-Based Computing would be feasible at Statistics Sweden, and if it would be in accordance with Swedish law. A close cooperation was established with some representatives of the Swedish research community to find out their needs and objectives regarding access to microdata. These contacts confirmed the need for metadata as well as the importance of testing the security solution at both ends. In addition, the project internally investigated the number of statistical products that could be handled by the new distribution method.

The results of this development project were very positive and since 2005 Statistics Sweden has a new system for remote access to microdata, aka MONA. With this system users are given secure access to databases at Statistics Sweden from practically any place that can provide Internet access. Data are processed and analyzed through a rich set of applications e.g. SAS, SPSS, STATA, GAUSS, Microsoft Office or Super Cross and result sets are then automatically sent to the user’s predefined mailbox.

The main goals for the MONA-system are:

- to increase accessibility to microdata for external users at the same time as security and secrecy is reinforced
- to keep all types of microdata for research on site at Statistics Sweden enforcing control of where, when, who and how data are used
- to have instantly upgraded data when needed without any requirements to produce new sets of disks or tapes for redistribution
- to present an easy to use front end for the end users built on well-known standard techniques and components such as server-based computing
- to present a complete system with powerful servers and a rich set of applications with no requirements on expensive equipment and software costs for end users

The MONA-system is built around communication between a client and a terminal server usually called server-based computing. The main idea for this concept is when a client is connected to the server, the client's computer or terminal performs no application processing. It processes only keyboard input and screen output and functions like an input/output terminal. All application processing is done in the server. A non-expensive PC or specialized terminal can be used as a client computer, running any Windows 9x/NT/XP operating system as well as Linux or MAC.

6. Providing access to micro data to researchers abroad

In the Nordic countries the same regulation concerning data confidentiality, as for release of data outside the NSIs, are in principle also valid when data is delivered to other countries. There are however some restrictions. According to the Data Protection Directive it is in principle forbidden to transfer personal data that is being processed to a third country (a country outside the EU and EEA) unless the third country in question ensures an adequate level of protection. The Data Protection Acts in the Nordic countries contain similar rules about release of data to a third country. In Sweden the Secrecy Act is also of relevance. According to Chapter 1 section 3, the release of confidential data to an authority or an international organisation outside Sweden is not allowed unless it is communicated in accordance with special provisions in legislation. Also, the information in a corresponding case might be given to a Swedish authority and the authority holding the information deems it evidently compatible with Swedish interest that the information is communicated. The EU regulation is such special provisions that make it possible to release micro data to Eurostat. There are no other special provisions concerning statistical micro data.

In Sweden the release of micro data to an authority in other countries for research is therefore possible only if it is compatible with Swedish interest that information is communicated. Micro data may be released to private researchers in other countries if it is evident that the information can be disclosed without the person whom the information concerns suffering loss or being otherwise harmed. In practice Statistics Sweden is restrictive with release of de-identified micro data to researchers in other countries.

Regards to the Statistical Act in Norway, all users of microdata are bound to secrecy. Since the legislation is not valid outside Norway, and Statistics Norway is thus not able to control if researcher in other countries maintains the confidentiality rules, Statistics Norway find it indefensible to release micro data outside Norway. However, the legislation accept transfer abroad if Norway is subject to an obligation to make a transfer pursuant to an international agreement or as a result of membership of an international organization.

In Finland the same regulations concerning data confidentiality as in Sweden, as for release of data outside Statistics Finland, are valid. An applicant must provide a description about how the data confidentiality is secured in the recipient country. Denmark does not release micro data to researchers in other countries but foreign researcher can use the Danish on-site arrangement under the same conditions as Danish researchers. Iceland has no experience in delivering micro data to researchers abroad.

6.1.1. Eurostat

The release of information to Eurostat is regulated in the EU regulations on statistics. According to Regulation 1588/903 the national authorities shall be authorized to transmit confidential statistical data to Eurostat. National rules on 3 Council Regulation (Euratom, EEC) No 1588/90 of 11 June 1990 on the transmission of data subject to statistical confidentiality to the Statistical Office of the European Communities statistical confidentiality may not be invoked to prevent the transmission of confidential statistical data to Eurostat where an act of Community law governing a Community statistic provides for the transmission of such data. This means that NSIs in principle are bound in regulations to release micro data for community statistics. However, transmission of data which are not covered by a specific Community legislative act is voluntary and that national rules can prevent the transmission of confidential data. Transmission of confidential statistical data shall be carried out to Eurostat in such a way that statistical units cannot be directly identified. This does not preclude the admissibility of more far-reaching transmission rules in accordance with the legislation of the Member States.

7. Future challenges

It is clear that the Nordic countries are committed to improving access to high quality microdata. One important strand in future development is to compile several new thematic registers tailored to better meet the needs of the research community. To accomplish this, considerable work is needed, engaging both methodologists and subject matter experts. Another future trend is to develop techniques that allow linkage of data from different sources, both within and outside the Nordic NSIs. In addition, some of the Nordic NSIs are designing Statistical Data Warehouses, which will enable them to build integrated registers and cubes in ways that allow continuous updates of data. It goes without saying that all these development trends are subject to fulfilment of legitimate confidentially requirements.

A close cooperation on use of registers and microdata has been launched with some Nordic universities. A number of seminars and symposiums have been arranged and several postgraduate students and a couple of joint professors are involved in this cooperation, which is expected to grow considerably in volume and importance.

Improved access to microdata involves relatively high costs to be borne by researchers. Because of this, Statistics Sweden has approached The Swedish Research Council arguing that funding from the Council of a system of microdata access would give researchers a lower initial cost when accruing data. This would also facilitate an increasing use of microdata in research. Our view is that the system of microdata access should be regarded as a national facility. Experiences from other areas where basic financing have been arranged and researchers only pay for marginal costs have been very positive. Such a solution would firstly incorporate full IT support for on-line access via the Internet. Secondly, a front office would be installed to serve and advise the researchers involved. Thirdly, this financial support would allow more and better thematic databases, which could be accessed directly by researchers, and could be created at an early stage. A solution along these lines would clearly facilitate an improved access to user friendly, high quality microdata.

8. Concluding remarks

There seem to be good possibilities to improve the access to microdata for researchers and other legitimate users in a radical way without violating confidentiality. In such a statistical system largely based on registers as in the Nordic countries, this really is a major improvement, also bearing in mind the new possibilities for dynamic analysis thanks to longitudinal microdata. Although considerable progress has already been made, systematic work for further improvements pave the way for new opportunities for researchers.

References

MONA Microdata ON-line Access at Statistics Sweden (only in Swedish).

Issues in Designing a Confidentiality Preserving Model Server

Philip Steel and Arnold Reznek***

*** United States Bureau of the Census, Statistical Research Division,**
Rm 3209 FB-4, 20233-9100, USA
philip.m.steel@census.gov
**** United States Bureau of the Census, Center for Economic Studies,**
WP2-206, 20233-6300, USA
arnold.phillip.reznek@census.gov

1. Introduction

Microdata publication has been the avenue of choice for data producers to serve the needs of sophisticated data users. Microdata allows virtually any analysis, but it is the only avenue of publication that allows modeling. To produce microdata in the context of a survey that promises the confidentiality of responses, the identity of respondents must be hidden. This has been accomplished by a variety of means, but the staple method is to suppress low-level geography and coarsen other variables until there is ambiguity as to whom in the population a record corresponds. Evaluating whether microdata is safe is problematic and often reduces to whether or not a high quality, identified, external file with sufficient overlap exists or can be constructed. As public data become more and more accessible, and data of all types on individuals accumulate, the ambiguity we rely on to protect the respondent is reduced. At the Census Bureau, we have undertaken a continual review of external data and have reduced detail on our microdata publications as potential problems are discovered. It is not difficult to project that at some point, data will be unsafe to publish, or what is publishable will be of low utility. On the one hand, the demand for microdata, both for general research and programmatic needs, continues to grow, and on the other, its separation from identified public or commercial data is harder to maintain. At the conference marking the publication of "Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies" this problem was dubbed "the train wreck" [Doyle et al 2001].

Model servers represent one way out of the train wreck problem. The result of the model is the object of interest, not the underlying data. There are indications that most model results are safe, at least when considered in isolation. The Census Bureau operates a number of Research Data Centers (RDCs) where researchers with Special Sworn Status have access to specified microdata. Model-oriented research is encouraged, and the output that is to be removed from the research data center is reviewed by the on-site employee and sometimes by the Census Bureau's Disclosure Review Board (DRB). The model output that we have examined over the past ten years of operation has been virtually without disclosure problems. Can this be reproduced in an automated system?

1.1. Existing model query systems

Several statistical query systems for analysis of sensitive data allow modeling: the Luxembourg Income Study (LIS), the (US) National Center for Health Statistics' ANDRE, and (US) National Center for Education Statistics' DAS are the most proven. All three systems require registration and a statement of purpose. ANDRE has some explicit monitoring capability. The LIS web documentation does not mention monitoring of users for compliance, but does in fact do so. ANDRE operates in conjunction with their Research Data Center (RDC) and provides remote access for the RDC's registered users. LIS is often utilized solely by remote access. Several similar systems are under development in the European Union. DAS takes a different approach by offering correlation matrices instead of running a formal routine. For a description of current systems see [Rowland 2003].

2. Preliminary design issues

The Census Bureau funded the development of a prototype model server suited to its needs. It is a proof of concept, using the Current Population Survey (CPS) public use microdata as its test bed, but designed to accommodate other microdata sets, and run SAS; some adaptation of it should be compatible with American FactFinder, our table server system. It would have exploratory capability, be user friendly, run a variety of models and populations, and provide measures of model fit.[1]

The mail systems employed by LIS and ANDRE have certain vulnerabilities. The submitted code could contain blocks that haven't been anticipated. This could include undocumented options and procedures, code that interacts with the operating systems, new procedures, or convolutions to get around procedures that are explicitly banned. Such a system can work in a monitored environment, where the users and their programming style are under observation, but staffing and user registration become an issue.

A web-based system, where code is built to user specification, avoids many of the problems associated with mail-based servers. This approach is termed an "enabling" system in the programming community: activity is restricted to what has been designed into the system, rather than proscribing certain activities or procedures and allowing all others, i.e. "disabling". It solves the problem of understanding the code being submitted by users and evaluating the output for disclosure. The enabling approach also has its pitfalls: it requires a user-friendly interface and is limited to only those statistical models that it has been programmed to construct. We have opted for an enabling system, since the monitoring problem of the disabling approach seems inescapable and the confidentiality problems in an enabling system can be addressed as capabilities are added to the system.

A computerized system generates its own set of problems, some quite different than what is encountered in the RDC environment. In particular, there is a problem of accumulated results and the inferences one can make from them. This is a variation of what is commonly referred to as "the subtraction problem". In current practice, preliminary results never leave the center and we rely on the RDC administrator or the DRB to recognize when final model results are too similar or when the results are not germane to a researcher's purpose. The process of producing a release from an RDC may involve months or years of work; a computerized system can produce the same volume of results in minutes. With an enabling system the problem of accumulated results is bounded (since the range of queries is known and limited) and can be confronted more directly.

2.1. The Current Population Survey

The system is meant to have general applicability, but we were also interested in seeing the extent of the difficulties that would be encountered on large complex data and what aspects of the system required survey specific programming. The public use dataset for CPS was selected for a test bed, specifically the March 2000 supplement. The CPS is the United States' longest running survey, tracing its roots back to an effort to measure unemployment during the Great Depression. Microdata from 1994 on are freely available at http://www.bls.census.gov/cps/. The March supplement focuses on demographic data and widens the applicability of the test. The data are well documented and have carved out a place both in current research and as an educational tool [Berndt 1990]. Using a public-use file allows us to involve more people in the testing of the software system, without risking confidential data. The data are topcoded and have a prepared geography, and any issues with categorical variables have already been resolved. The system's dependency on this kind of preparation must, at some point, be evaluated.

[1] Synectics is the developer of the system.

2.2. Non-confidentiality problems in design

As with any data tool, the model analysis system has to present the options of a complex task in a simple to use format. Hierarchical data structures, such as geography and the relationship of the individual in groupings such as household and family, affect the ultimate structure of the queries and the confidentiality problems encountered. Yet they are also the items frequently of interest to the user. The CPS is rich in detailed financial data with a data dictionary that is 127 pages long. Variable descriptions must be incorporated into the instrument and often determine the role of the variable. A variable may play several different roles or, for confidentiality reasons, be barred from playing a particular role; for instance a poverty indicator (which combines income thresholds with household size) can be available in the exploratory phase, cannot be combined with income in universe formation, but can be available again in the analysis phase. The handling and display of the survey metadata is a large piece of the overhead for this project.

2.3 Confidentiality strategy in design

The design can be divided into five moderately distinct sections: data preparation, data exploration, universe definition, model statement and results. Because the basic strategy is to "enable", most restrictions will be passive from the user's point of view. They simply have no facility to engage in risky behavior. We try to avoid active restrictions, where the user makes a choice, it is evaluated, and possibly denied. Active restrictions lead to frustration, particularly if the evaluation comes later in the process or otherwise generates delay.

Data exploration will initially be rudimentary–allowing the user to make only a general examination of the data, sufficient to inform decisions for constructing a model. More descriptive or expressive data exploration can be permitted once the confidentiality requirements are known to be effective.

Universe definition, the restriction to the user's desired population, is more directly involved with determining the parameters of the modelling system and may be the most difficult to accommodate. This will be the substantial focus of the initial development effort. The universe definition stage is equivalent to a "coarse" table server.

While restrictions must be imposed on the model statement, those restrictions address some known very specific problems. These constitute a very small fraction of possible models. The user may never encounter those restrictions associated with the model statement, with the exception of an initial ban on large, fully saturated models [Reznek 2003].

The estimates for the model are derived directly from the data in as much detail as the collection and preparation can afford. The values on which they are based may differ from what is encountered in exploration and what is available in the universe formation stage, where the user may encounter a synthetic analogue or a recode. Diagnostic statistics require considerable care and can pose a substantial disclosure risk. For instance, residual values are record level data and the values they are based on are easily recoverable. Diagnostics may also address outliers. Where diagnostics are risky, a synthetic approach will be employed [Reiter 2003]. The emphasis will be on the ability to obtain model results without noise or bias from confidentiality restrictions. Diagnostics will be approached more conservatively and can be the subject of future improvement.

3. Data exploration

It is important that users be able to examine univariate and bivariate distributions before being asked to specify a model. The user may also wish to confirm some aspect of their result with a simple table. The exploratory capability should include such tables for most or all categorical variables, at the user specified geographic level. The exploratory tabulation applies to the entire population, not necessarily the population the user is studying. The confidentiality requirement for this facility is on the data preparation. The preparation must support exploratory two-dimensional tables; for CPS we have preparation primarily through the CPS geographic designation, where no designation shows fewer than 100,000 population. For numeric variables we can offer a categorical analogue or a synthetic representation. The categorical analogue may be a simple indicator or something more detailed. A synthetic representation allows us to extend the exploratory capability to a display of plots. This is particularly suited to displaying transformations of numeric variables. Transformation will be limited initially to log, square and square root, but this list may be expanded at a later date [Reiter 2003].

4. Universe formation

Universe formation, or subsetting down to the model population, is an overt confidentiality problem. It gives the count of the population defined by some set of conditions, i.e. it is gives a cell in a table of counts. With complete freedom to vary the conditions, any table could be constructed, including cells of size one. But the ability to run a model on one additional observation may allow the reconstruction of the record in its entirety. See [Cox 2004] for the construction of dependent variables. Disclosure avoidance techniques used on tables could conceivably be used, but the more sophisticated ones would be difficult to apply. Disguising the number of observations in the user's desired population gets into sequence, retention and additivity problems. For example, controlled rounding is attractive for this sort of problem, but to work in the context of the server it would have to be performed consistently on all possible tables and the appropriate rounded value would have to be presented to any user whose population corresponds to that particular cell. Later values produced by the model would have to be consistent with the number presented when the universe is initially defined. Cell suppression, or in this context, a rejection of the user's universe selection, can also lead to an open-ended problem.

The problem is not insurmountable, however. Models are usually run on fairly substantial populations, and hence equivalent to moderately large table cells. We will assume (or rather, require) that the model universe will have at least 75 observations. The magnitude is such that we should be able to guarantee (in the data preparation) that the balance of the table cells does not fall below a count of 4. The task then becomes to verify that the user is attempting to model a reasonable sized population and the balance is not a confidentiality problem.

4.1. Numeric variables in universe formation

For categorical variables one would use an equality condition variable to extract the subset on which the model is to be run. For example, head of household=1, Hispanic=1, and educational attainment=44. For numeric variables, like income, the condition would be defined in terms of an inequality. Note that the system also allows for "or" conditions, but the simplest case is sufficient for illustration. For example, all households with total income greater than or equal to 17,000. However, for numeric variables the underlying data have full detail. By incrementing the value, or cutpoint, it would be possible to define a universe for all households with total income greater than or equal to 17,001. This could contain just one additional observation. Comparing coefficients of the model run on both universes may indicate the characteristics of that isolated record. By going through a progression of models, it may be possible to reconstruct the entire microdata record. How can one prevent differencing of this type, but still provide some facility for using numeric variables in defining a universe?

Clearly the set of cutpoints must be pre-determined. We would like the points to be evenly spaced, with enough distance between points so that there is a reasonably good chance that the user will not run into a denial based on the "at least 4 observations" rule. When numeric data are being rounded or presented in tables, the schemes used are often ad hoc; but they share a property of graduation, so that the average difference between the true value and the rounded value is proportional to the magnitude. Rounding by 10 up to 200, by 100 up to a 2000 etc. Examination of the CPS data on income reveals a great deal of rounding by respondents, following a pattern that is also responsive to magnitude. Generally the data spikes initially at values divisible by 5, then later by 10 then 50, 100, 500, 1000, 5000, 10000. The peak of observations at 30000 represent some values being rounded up, and some being rounded down. The rounding scale is also varying, with some people rounding 32,500 to 32,000 and some rounding all the way to 30,000. The initial set of cutpoints for the long list will consist of a compromise between an attempt to evenly distribute the data and the incrementing between peaks that occur naturally in the data. The cutpoints will be calculated using the natural increments with the shift in increments sensitive to a threshold on the number of observations. Cut points will be offered at 50, 100, 150 until the number of observations between falls below our parameter, at which point the increment will increase to 100 to continue until it fails to capture enough observations, bump the increment and so on. The set of cutpoints distributes fairly evenly across the data, can be adjusted in the testing phase, and follows a scheme that is similar to the clustering already occurring in the data.

On the high end, the cutpoints should not exceed the value one would use for a topcode in a microdata publication; that is, a half percent of all observations should be above the topcode or 3 percent of the non-zero observations. Note that this restriction is applied to universe formation. No restriction is currently envisioned on the use of large values in the model input. This feature of the system is not applicable to the current test. The CPS is already topcoded and already published as microdata. Since the record structure is already known, we cannot additionally allow access to large values in universe formation and, perhaps, even in the modeling phase.

Note that the burden is shifted from confidentiality to usability. More granularity in the cutpoints leads to more rejected universe formations. The optimum setting on the threshold for the minimum number of points between cutpoints is a function of what “large” is in the context of users running models on “large” populations. Our initial set of cutpoints was generated to accommodate restrictions where the categorical variables define a population approximately one tenth of the CPS universe. The numeric variable adds a further restriction, but one that is unlikely to violate the four observation rule, by design. This long list of cutpoints is appropriate only for users that require a fairly exact threshold on a numeric variable and have few other restrictions for the model’s population. The loss in precision is not as great as it might seem. More hinges on whether a natural rounding point is included or excluded and how much of the contribution to that point comes from respondents rounding up or rounding down, than hinges on the distance from the desired point and the cutpoint actually available. The long list would remain fairly close to a rounded version of the variable, at least for modest increases in the threshold.

For users where categorical restrictions reduce the population to less than one tenth, a less detailed list of cutpoints will be available--a short list. The conditions defining the population can be viewed as a cell in the table of counts obtainable by varying the conditions. The dimensionality of that table is the number of variables involved in the conditions. The size of the table, then is the product of the number of conditions associated with the variables. Our strategy for the short list is that its size will be the square root of the size of the long list and thus two short list variables will be roughly equivalent to the long. The short list is fixed and is a subset of the long list.

4.2. Indicators in universe formation

In addition to short and long lists, an indicator will be available for zero values and perhaps other categorical characterizations. For example, 1 or –1 may have a special meaning. Some variables have a significant range of negative numbers. Whether these are allowable needs to be related to the threshold used in 4.1. An outstanding issue is how to handle derived codes, like poverty. Poverty can be expressed as a function of income, size of household and number of children present. Poverty could be used to subdivide a cutpoint range, since it includes other dimensions implicitly. For such codes an active restriction must be employed, and such restrictions are dataset specific.

5. Confidentiality for the Model Statement: Interactions and Dummies

Disclosure risks may arise from the use of regression models, particularly in the standard linear regression model estimated using Ordinary Least Squares methods as well as in logit and probit models (which use binary (0,1) dependent variables), and other Generalized Linear Models [Reznek 2003, Reznek and Riggs, 2004]. The risks in regression models that contain continuous variables are small if the overall sample is large enough to pass tabular disclosure analysis. Risks are most apparent in models that contain dummy variables as independent variables. Coefficients of models that contain only fully interacted sets of dummy variables on the right-hand sides can be used to obtain entries in cross-tabulations of the dependent variable broken down by the categories defined by the dummy variables. That is, from the disclosure avoidance point of view, these models are equivalent to tables. We will conservatively bar interactions involving 4 or more variables and fully interacted models of 3 variables. It also seems sensible to keep users from specifying an over-determined or nearly over-determined system. We will use either a fixed cap of around 20 variables or a parameter dependent on the number of observations.

In addition, each dummy category should have at least 20 observations. For dummies meeting this threshold, its estimated coefficient will be shown. For dummies failing the threshold they will be absorbed into the constant term along with the last term of the dummy. The choice of the absorbing term will not be available to the user, though this is a product of programming rather than confidentiality constraints. Menuing for such a choice requires an additional population of metadata and some, as yet unresolved, division of tasks between the query build and the query execution. Our initial procedure, “Proc Reg” does not support a full-fledged “absorb” statement, though enabling it for procedures that do, should be relatively straightforward.

6. Confidentiality for Model Results

The restrictions in the formation of the model statement are intended to guarantee that the coefficients can be presented without further restriction. Most summary statistics on residuals can be displayed safely. Our efforts on presenting measures of validity have been focused on devising a way to display residuals. Examination of residuals is a frequently used method for evaluating models, particularly in the early stages of development. Of course, the actual residuals allow the recovery of the underlying values and cannot be presented to the user. Synthetic residuals, provided they convey the same information to the user, are an effective substitute. In the test on the CPS public use data, the synthetic residuals are being presented side by side with the actual residuals to determine if they are adequate.

Density estimation will be done by the SAS KDE routine. The output from this allows us to generate a random set of points with approximately the same density as the original residual data. The random number generator must have a fixed start, since repeated application of the KDE procedure should show convergence to the original set of points. KDE is also sensitive to outliers so the synthetic data are topcoded at four standard deviations from the mean. The number of topcoded values will be provided to the user. A variation on this procedure can be used to generate two-dimensional plots. The procedure for generating synthetic residuals can be adjusted in two ways. The endpoints used by the procedure can be altered. Also the number of grid points can be adjusted. Both may have some impact on the effectiveness of the procedure.

7. Data Preparation

The weaknesses of model servers are strongly related to the problems encountered in table servers. The capability to restrict models to particular populations yields counts of the population. Varying the restriction parameters enables the user to construct table margins or cells. For some models, the estimates of coefficients in a model are equivalent to groups of table cells. Tabular disclosure problems of this type are found in virtually any publication form--a model server does not solve them. What we can hope to accomplish is to use the model server to restrict the table server problem to a manageable dimensionality and to prevent the reconstruction of individual records and most numeric values. That is, we will assume that the data at which the model server points are sufficiently prepared to safely allow publication of most lower dimensional tables on the available geography. This is dependent on the population, sample size, some regularity in variable categories and the creation of an appropriate geography.

The strictest standard for microdata is k-anonymity. For all variables thought to overlap with external files, there are at least k members of the population displaying any combination of characteristics present in the microdata. K-anonymity is usually considered with respect to a limited set of variables and frequently must make the poor substitution of the sample population for the full population. The model server setting introduces a substantial variation. Because the primary function of the server is to hide the record structure and that record structure can be recovered by subtraction for susceptible records in the universe formation stage, the anonymity that is desired is with respect to the sample population. If a record is unique with respect to three variables in the CPS it does not matter what multiplicity it has in the population ... its uniqueness on those variables can be used to recover the rest of the record. The relationship of the three variables to external data is not of consequence. That is, the anonymity we desire is not restricted to key variables, but rather the larger set of variables made available in universe selection. However, in this set only combinations of the dimension allowed in the universe formation need be considered. Uniqueness on six variables is not a problem provided the user cannot specify that combination for a universe. This is critical because it is at the higher dimensions that the computational complexity in standard k-anonymity begins to kick in. K-anonymity with respect to a four-variable key (quasi-identifier) is a computationally feasible problem. K-anonymity with respect to an eight variable key is not currently feasible [LeFevre et al 2005].

It is also worthwhile to note that if the tabular disclosure is not direct but rather derived by inference, it may not allow the isolation of a record by subtraction. You may know that there is a unique record with four particular characteristics but are unable to separate it from other records. Whether tabular disclosure poses a risk for model results is an open and difficult question.

8. Future Application

The degree to which the architecture of the program will support a shift of datasets or an expansion of models is yet to be determined. We would like to add a user selection of the absorbing dummy category in the simple regression procedure. The presentation of long variable lists can be improved. Usability of the tool, independent of confidentiality concerns, is crucial to supporting its further development.

The model server tool is ideal for instruction. Because no programming knowledge is assumed for the user, students with limited programming or even statistical expertise can learn basic regression analysis on live data. For surveys where microdata are a subsample (Decennial Census, ACS), the model server could be pointed at the nonsampled portion and provide model results for data with very little topcoding. For surveys that have developed replicate weights for variance estimation that are unavailable to the public, the model server system may allow results to be evaluated with those weights without compromising the detail (usually geographic) which makes the weights themselves risky.

9. Conclusion

We have specified a query system with parameters that can be adjusted to tighten or relax confidentiality requirements. We welcome any suggestions or justifications for particular settings of those parameters. We hope we have taken a step forward by producing a prototype system that affords some of the advantages of microdata in an environment that is safer than exists with standard microdata. The system described is designed to prohibit the user from reconstructing the underlying microdata. By this we mean primarily that the record-level organization cannot be recovered. The long-term problem with microdata is the number of variables that overlap with external data and, if successful, this tool gives an alternative that denies application of record linkage machinery.

References

Berndt, Ernst R., *The Practice of Econometrics: Classic and Contemporary* (1990) Addison-Wesley.

Cox, L.H., “Confidentiality Problems in Statistical Database Query Systems,” *Research Directions in Data and Applications Security* (C. Farkas and P. Samarati, eds.) (2004) Kluwer.

Doyle, P. et al., editors, “Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies” (2001) North-Holland.

Harris, K.W., Gambhir, V., “National Center for Health Statistics’ Research Data Center”, Proceedings of 2004 FCSM Statistical Policy Seminar (forthcoming). See also http://www.cdc.gov/nchs/r&d/rdcfr.htm

LeFevre, K., DeWitt, D., Ramakrishnan, R., “Incognito: Efficient Full-Domain K-Anonymity”, procedings SIGMOD 2005.

Luxembourg Income Study web site: LISSY V Job Submission Instructions, section A.2 see also http://www.lisproject.org/introduction/userform.htm researcher request form

NCES’ Data Access System web site: http://nces.ed.gov/das/

Reiter, J. (2003), “Model Diagnostics for Remote Access Regression Servers” *Statistics and Computing,* 13, pp. 371-380.

Reznek, A. (2003), “Disclosure Risks in Cross-section Regression Models” Proceedings of the Section on Government Statistics, JSM.

Reznek, A. and Riggs, T. (2004), “Disclosure Risks in Regression Models: Some Further Results”. Proceedings of the Section on Government Statistics, JSM.

Rowland, Sandra (2003), “An Examination of Monitored, Remote Microdata Access Systems” from the National Academy of Science’s Workshop on Access to Research Data: Assessing Risks and Opportunities.

IPUMS-Europe: Confidentiality measures for licensing and disseminating restricted-access census microdata extracts to academic users

Robert McCaa and Albert Esteve***

*** Minnesota Population Center, Minneapolis, MN 55455, USA, rmccaa@umn.edu**

**** Centre d'Estudis Demogràfics, Autonomous University of Barcelona, Barcelona, Spain, aepalos@yahoo.es**

Research for this paper was funded in part by the National Institutes of Health of the United States, grant HD047283'European census microdata harmonization project (IPUMS-Europe)'and CIECM.

> "Inadequate use of microdata has high costs"
> —Len Cook (2003)

Abstract. Confidentiality protections for census microdata depend not only on the sensitivity and heterogeneity of the data, but also on the potential users. It is widely recognized that statistical agencies exert substantial effort to protect microdata from misuse by academics, their most trust-worthy users. The IPUMS-International projects, by disseminating only integrated, anonymized microdata and restricting access to licensed academic users, shifts the risk-utility curve sharply rightward—substantial increasing utility with only marginal increments in risk. The IPUMS-International approach provides access to microdata of high utility at the same time that confidentiality risks are minimized. Many statistical institute partners anonymize the microdata and implement technical measures of confidentiality protection before the data are entrusted to the project. This paper discusses legal, administrative and technical practices of the IPUMS-International project for disseminating harmonized census microdata extracts with specific reference to the IPUMS-Europe regional initiative.

1. Introduction: IPUMS-International

The IPUMS-International is a global initiative led by the University of Minnesota Population Center to confidentialize, harmonize and disseminate high-density census microdata samples on a restricted access basis to academic users (Ruggles et. al. 2003). Begun in 1999 with funding provided by the National Institutes of Health and the National Science Foundation of the United States, to date the initiative enjoys the endorsement of official statistical institutes of more than fifty countries. Marginal costs of constructing and maintaining the database are born by the MPC, its funding agencies, the University of Minnesota and academic partners–not by the statistical institute partners. On the contrary each is paid a modest fee per census to supply microdata and documentation to the project. In May 2002, the first phase of integrated census microdata for Colombia (1964-1993), France (1962-1990), Kenya (1989-1999), Mexico (1960-2000), the United States (1960-2000), and Vietnam (1989-1999) were made available to licensed users, followed by China (1982) in 2003 and Brazil (1960, 1970, 1980, 1991, 2000) in 2004. More than 500 users representing more than 30 countries are currently licensed to obtain custom-tailored extracts free of charge from the project website: https://www.ipums.org/international

Table 1. IPUMS-International Integrated Census Microdata Sample Characteristics, 120 million person records.

Country census		Sample %	No. of Person records	Additional details
Brazil	1960	5.0	3,001,000	Long-form, cluster sample
	1970	5.0	4,954,000	Same
	1980	5.0	5,870,000	Same
	1990	5.0	8,523,000	Same
	2000	6.0	10,136,000	Same
China	1982	0.1	1,003,000	Every thousandth household
Colombia	1964	2.0	350,000	Every fiftieth person
	1972	10.0	1,989,000	Every tenth household
	1985	10.0	2,643,000	Long-form, cluster sample
	1993	10.0	3,247,000	Every tenth household
France	1962	5.0	2,321,000	Every twentieth household
	1968	5.0	2,488,000	Same
	1975	5.0	2,629,000	Same
	1982	5.0	2,714,000	Same
	1990	4.2	2,361,000	Every twenty-fourth household
Kenya	1989	5.0	1,074,000	Every twentieth household
	1999	5.0	1,410,000	Same
Mexico	1960	1.5	503,000	Every 67th individual
	1970	1.0	483,000	Every hundredth household
	1990	10.0	8,028,000	Every tenth household
	2000	10.6	10,099,000	Long-form, cluster sample
USA	1960	1.0	1,800,000	Stratified, random sample
	1970	1.0	2,030,000	Same
	1980	5.0	11,337,000	Same
	1990	5.0	12,500,000	Stratified, cluster sample
	2000	5.0	14,082,000	Same
Vietnam	1989	5.0	2,627,000	Long-form, cluster sample
	1999	3.0	2,368,000	Same

Source: https://www.ipums.org/international/sample_descriptions.html

With the inclusion of the data for Brazil, the IPUMS-International website offers some 120 million person records consisting of more than 100 variables from 28 samples with densities varying from 0.1 to 10 percent (Table 1). Over the next five years, the database will expand to 44 countries with regional initiatives in Europe, Africa, Asia, Oceania and Latin America (McCaa and Esteve 2005). It should be noted that the mode of access to IPUMS-USA samples differs from the International project. The former, a public site, makes data available to anyone and therefore has tens of thousands of users, while the later provides data only to licensees, numbering only in the hundreds.

The IPUMS-International/Europe regional project began in September 2004. Thanks to additional funding by the European Community Sixth Framework Program, the inaugural workshop was held in Barcelona in July 2005. Delegates from the official statistical agencies and academics met to discuss data availability, samples, general harmonization issues, and overall project procedures. A second workshop, hosted by L'Institut National d'Études Démographiques, to be held in June, 2006 will focus on detailed harmonization issues. In 2007, the first European region data release is scheduled for release with a mirror-site at the Centre d'Estudis Demogràfics (Barcelona).

Table 2. IPUMS-Europe: Likely Censuses and Sample Sizes (in 000s), by Country. Bolded census year indicates sample has been drawn and entrusted to project.

	Sample Density (%)	Census	N	Census	N	Census	N	Census	N	Census	N
Austria	**10**	**2001**	**884**	**1991**	**902**	**1981**	**838**	**1971**	**836**		
Belarus	**10**	**1999**	**1,040**			.	.	.	.	.	.
Bulgaria	5	2001	395	1992	425	.	.	.	.	.	.
Czech Rep.	5	2001	515	1991	515	.	.	.	.	.	.
France	**5**	**1999**	**3,005**	**1990**	**2,361**	**1982**	**2,714**	**1975**	**2,629**	**1968**	**2,488**
Germany	?	2001	330	1991	321	1987	tbd	1982	tbd	1973	246
Greece	**10**	**2001**	**1,029**	**1991**	**969**	**1981**	**923**	**1971**	**845**	.	.
Hungary	**5**	**2001**	**511**	**1990**	**518**	**1980**	**536**	**1970**	**515**	.	.
Netherlands	**1**	**2001**	**190**	.	.	.	.	**1971**	**159**	**1960**	**143**
Poland	?	2002	1,930	1995	1,940	1988	1,900	1984	1,850	1978	1,745
Portugal	?	2001	500	1991	495	1981	490	.	.	.	.
Romania	**10**	**2002**	**2,239**	**1992**	**2,138**	.	.	.	.	.	.
Russia	5	2002	7,200	1989	7,400	.	.	.	.	.	.
Slovenia	10	2001	200	1991	tbd			.	.	.	.
Spain	**5**	**2001**	**2,040**	**1991**	**1,940**	**1981**	**1,875**	.	.	.	.
UK	1	2001	600	**1991**	**574**	.	.	.	.	.	.

Notes: Total Person Records ~ 65 million

Micro-censuses: Germany 1982, 1991, 2000; Netherlands 2001; Poland 1974, 1984, 1995.
Samples for 1962 France and 1960 and 1974 Poland are included in the total case count.
Final agreements for Poland, Russia and Turkey are pending, and some of the earliest censuses may not be recoverable.
tbd = to be determined.

2. Dissemination of IPUMS-International "Extracts"

Users of IPUMS-International are not permitted to access microdata containing the original codes provided by the Official Statistical Institutes. Instead, the microdata are integrated, that is, they are transformed into a complex coding scheme which seeks to preserve all significant detail yet assign identical codes to identical concepts. The integrated microdata are provided only in the form of extracts, custom tailored to each researcher's needs. What this means is that there is no distribution of entire datasets in the form of of compact discs, DVDs or otherwise. Since each dataset is custom tailored, "collecting" or "boot-legging"datasets is not only illegal, but effectively curtailed. The database is so enormous and evolving so quickly that users and their institutions have a powerful interest in safe-guarding the data and promoting good use.

To request an extract, the user must first become licensed (see below) and then sign into the project website ("*create an extract*") by entering the registered password. Then a series of selections are made by means of point-and-click menus. The user selects the country or countries, census years, samples, and variables as well as the form of metadata required for the statistical package to be used (SAS, STATA or SPSS are provided). The IPUMS extract engine also makes it possible to select cases (persons, households, or dwellings) with specific characteristics, such as, say, females aged 15-19 in the workforce. Selected cases may also include members of households or families in which the selected case is found.

One of the most valuable enhancements of the database is the "SUBSAMPLE" feature. With SUBSAMPLE, the user may request any of 100 sub-samples each of which is nationally representative and preserves any stratification of the larger sample from which it is drawn. This tool may be used to test procedures, economize resources (where the research does not require large samples), or estimate variances through the replicate method.

Once the selections are complete, there is an opportunity to review or revise before final submission of the request. Then, once submitted, the extract engine registers the request and places it in a data processing queue. When the extract is ready (usually in a matter of minutes), the researcher is notified by email that the data should be retrieved within 72 hours. A link is provided to a password-protected site for downloading the specific extract. The data are encrypted during transmission using 128-bit SSL (Secure Sockets Layer) encryption standard, matching the level used by the banking and other industries where security and confidentiality are essential. The researcher may then securely download the file, decompress it and proceed with the analysis using the supplied integrated metadata consisting of variable names and labels. The metadata are in ASCII format so that a researcher may readily adapt them for use by any statistical software.

3. Confidentiality

IPUMS-International means Integrated Restricted-Access, Anonymized Microdata Extracts. The IPUMS-Europe acronym carries "PUMS" embedded in its name, but in fact the data are available only as "restricted-access extracts" from anonymized, integrated samples. Thus, "IRAAME" would be a more literal acronym, and indeed when the IPUMS was internationalized in 1998, the Principal Investigators discussed replacing "PUMS" with a more accurate moniker. We also discussed inserting "scientific" in place of "public". However, a decade-long, unbroken string of successes in securing monetary resources from the National Science Foundation and the National Institutes of Health dissuaded us then from abandoning the acronym, as it does now with the sister projects, IPUMS-Latin America and IPUMS-Europe.

Nonetheless, it is important to understand that a comprehensive array of additional protections, much greater than those for IPUMS-USA, are in place to guarantee the privacy and statistical confidentiality of census microdata samples incorporated into the IPUMS-International database. 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, etc.) 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.

While much of the literature on statistical confidentiality ignores the legal and administrative environment (and in doing so exaggerates the risk of improper use), we remain firmly persuaded that the strongest system of protections must take into account all three types of guarantees (Thorogood 1999). IPUMS-International confidentiality standards seek to comply with EC Regulation 831/2002, although this regulation encompasses only four datasets at present: European Community Household Panel, Labor Force Survey, Community Innovation Survey, and Continuing Vocational Training Survey (King 2003).

3.1. Legal protections

First, with regard to legal protections, IPUMS-International projects are undertaken only in countries where explicit authorization is forthcoming, usually in the form of a memorandum of understanding endorsed by the official statistical institute and the legal authority of the University of Minnesota (see Appendix A). No work is begun with the microdata of a country without prior signed authorization from the corresponding OSI. The agreement is highly general and uniform across countries. Details specific to each country such as fees and sample densities are negotiated separately with each official agency and do not form part of the agreement. Under a carefully worded legal arrangement, the Regents of the University of Minnesota are responsible for enforcing the terms of the accords. The ten clauses spell out: 1) rights of ownership, 2) rights of use, 3) conditions of access (in which statistical institutes cede their gate-keeping authority to grant individual licenses to the IPUMS-International project), 4) restrictions of use, 5) the protection of confidentiality, 6) security of data, 7) citation of publications, 8) enforcement of violations, 9) sharing of integrated data, 10) and arbitration procedures for resolving disagreements. There are no secret clauses or special considerations. Although minor rewording of clauses is permissible, all members of the consortium are treated equally.

Nonetheless, the protocols are revised, indeed expanded, as OSIs suggest, or request, modifications. Any request for modification is reviewed by the legal cabinet of the University of Minnesota. Compare for example the violations clause in Appendix A (as signed by Statistics Austria in January 2002) with the current text (additions in italics), as follows:

> Violations. Violation of the user license may lead to professional censure, loss of employment, and/or civil prosecution. *The University of Minnesota, national and international scientific organizations, and the [the Statistical Agency of Country X] will assist in the enforcement of provisions of this accord.*

Recently the tenth clause, which establishes jurisdiction for the settlement of a dispute between the University and any signatories to the memorandum, was amended, substituting the International Court of Arbitration for the Chamber of Commerce of Paris. At the same time, an eleventh clause, regarding order of precedence, was added, specifying that the clauses in the letter of understanding supersede any contract, purchase order or other document signed between the parties. Under the agreement, the Minnesota Population Center and its authorized partners are obliged to share the integrated data and documentation with the official statistical institutes and to police compliance by users.

3.2. Administrative measures

Second, researchers must apply for a license to gain access to the microdata extraction system (see: https://www.ipums.org/international/apply_for_access.html). Grounds for approval are based upon three considerations:

1. whether the data are appropriate for the proposed project as stated in the applicant's project description
2. whether the applicant is an academic, non-commercial user
3. whether the applicant agrees to abide by the restrictions on conditions of use (see Appendix B).

The vetting of applications is performed by the Principal Investigators of the IPUMS-International project. It is noteworthy that approximately one-third of applications are denied because of a failure to adequately satisfy one or more of the specified conditions. It is gratifying to report that few users appeal denial of access.

Administrative measures limit access to the extract system to users, who:

1. sign the electronic non-disclosure license;
2. endorse prohibitions against a) attempting to identify individuals or the making of any claim to that effect, b) reporting statistics that might reveal an identity and c) redistributing data to third parties;
3. agree to use the data solely for non-commercial ends and to provide copies of publications to ensure compliance;
4. place themselves under the authority of educational institutions, employers, institutional review boards, professional associations, and other enforcement agencies to deal with any alleged violations of the license.

The license is granted to users, individually, not to research groups, classes, or institutions. The license application instructs the applicant regarding conditions of use (see Appendix B). The license is not transferable. Should the individual change institutions or employment, the license must be updated. Data can be reassigned within an institution, but the person responsible for the microdata must apply for access. Once licensed, the user is permitted to download data extracts of samples and variables according to need. Licensees import the extracts into their statistical software of choice to analyze at the convenience in their own institutional setting.

Since its adoption in 2002, the basic application procedure remains unchanged. Few suggestions for enhancing the application form or approval process have been forthcoming, even though advice is solicited from users, statistical institutes, funding agency review boards, and outside experts. Nevertheless in 2006, we plan to strengthen application and vetting procedures, primarily to guard against fraudulent applications. In addition to requesting additional details about the applicant and institutional affiliation, the form will contain the following statement as a heading:

> **Legal Notice: Submission of this application constitutes a legally binding agreement between the applicant, the applicant's institution, the University of Minnesota, and the relevant official statistical authorities. Submitting false, misleading or fraudulent information constitutes a violation of this agreement. Misusing the data by violating any of the conditions detailed below also constitutes a violation. Violation of this agreement may lead to professional censure, loss of employment, civil prosecution under relevant national and international laws, and to sanctions against your institution, at the discretion of the University of Minnesota and the official statistical authorities.**

In the United State, an Institutional Review Board for the protection of human subjects is required of any academic research institution applying for funding from the National Institutes of Health. IRBs provide a strong mechanism for enforcing of the IPUMS-International license agreement in the United States. Most developed and developing countries have similar mechanisms. Delegates to this conference are invited to provide the names of similar institutions in their country. Oversight boards are nearly universal. It is these boards that provide a strong shield for insuring the highest standards of scientific conduct.

Finally, once these revisions to the application are in place on the website, licenses will be valid for one year and will be renewable. A license may be suspended at any time.

3.3. Technical protections

Third are the technical measures taken to ensure statistical confidentiality. Sampling of datasets alone "provides the additional uncertainty needed to protect many data releases…" (Anderson and Fienberg 2001). Census errors and non-response error also provide their own confidentiality protections.
As Fienberg (2005) has noted the principal threats are geographic detail and extreme values. Many statistical institute partners anonymize the microdata and implement technical measures of confidentiality protection before the data are entrusted to the project. When the OSI provides a sample that is also made available to others–such as public use samples, SARs and the like–no additional protections are implemented by the project. Usually the project is not informed of the precise technical measures imposed on the data. Where the samples are unique, we impose the following technical protections (based on Thorogood 1999):

1. adopt sample density according to official norms or conventions (see tables 1 and 2);
2. limit geographical detail by means of global recoding to administrative units with a minimum number of inhabitants. For some countries, this limit is as high as 100,000 and for others as low as 10,000. For the European project, NUTS3 is likely to be the lowest level of identifiable administrative geography, with the minimum threshold varying from 20,000 to 100,000 inhabitants according to the most recent census;
3. top and bottom code unique categories of sensitive variables (identified by the OSI);
4. round, group, or band age as necessary;
5. suppress date of birth (only age is provided);
6. suppress detailed place of birth (<20/100,000 population);
7. suppress detailed place of residence, work, study, and migration (<20/100,000 population);
8. systematically "swap" (recode) place of enumeration for a fraction of households, inversely proportional to population size at the NUTS3 level; Data swapping protects confidentiality by introducing uncertainty about sensitive data values, yet maintains the strength of statistical inferences by preserving summary statistics (see Fienberg and McIntyre, 2004).
9. randomly order households within administrative units (NUTS3);
10. and, conduct a sensitivity analysis once these measures are imposed to determine what additional measures may be required.

We continue to evaluate emerging methods and technologies for disclosure protection (McCaa and Ruggles 2002). At present we have decided against automatic data protection methods such as *μ-Argus* (Hundepool et al, 1998; Polettini and Seri 2003). It should also be noted that no synthetic data are added to the IPUMS samples.

4. Shifting the R-U Curve Rightward

In practice, disclosure of confidential information from census microdata samples is highly improbable. Moreover, researchers have no interest or incentive to even attempt to identify individuals. There are compelling reasons for jealously guarding confidentiality, both for individual users and the academic community as a whole. Any partially successful effort, such as that by a rogue intruder, will require an enormous investment of resources to obtain rather trivial details invariably with a high degree of uncertainty about whether any one record truly corresponds to a targeted individual or entity

(Dale and Elliot 2001). Indeed, over the past forty years of disseminating census microdata in the United States and elsewhere there are few allegations of misuse or breach of statistical confidentiality by an academic researcher. The IPUMS-International procedures are designed to extend this nearly perfect record.

Len Cook (2003) notes that increased access is not a threat to statistical systems. On the contrary he observes that increasingly there is an expectation that analysis of microdata will inform research and evaluation of policy. Increased access builds trust in statistical systems, while lack of access leads to suspicion. He advocates that different forms of access be granted for different degrees of trust. Moreover academic researchers possess a range and depth of expertise that national statistical institutes cannot replicate.

Julia Lane (2003) highlights five classes of benefits which accrue from broader access to microdata: address more complex questions, calculate marginal effects, replicate findings, assess data quality and build new constituencies or stakeholders. Replication is extremely important because there is an overwhelming temptation for scientists to misrepresent results when the data are unlikely to be available to others. The IPUMS system facilitates replication by providing access to microdata to all approved academic users on an equal basis.

5. Conclusion

Now that the construction of anonymized microdata data samples is becoming an increasingly widespread practice, harmonization of census microdata is an obvious next step to enhancing use. With the emergence of global standards of statistical confidentiality and the massive power of ordinary desktop computers, the major challenge that remains is the actual construction of integrated, anonymized census microdata samples. By restricting access to a class of academic users, high-density microdata extracts can be provided to researchers at vanishingly low risk.

References

Anderson, Margo and Stephen E. Fienberg. (2001). "U.S. Census Confidentiality: Perception and Reality," International Statistical Institute Biennial Meeting (Seoul).

Cook, Len. (2003). "Summary of Discussants' Main Points," in *Statistical Confidentiality and Access to Microdata: Proceedings of the Seminar Session of the 2003 Conference of European Statisticians*. Geneva, pp. 7-10.

Dale, A. and Elliot, M. (2001). 'Proposals for 2001 SARS: An assessment of disclosure risk.' *Journal of the Royal Statistical Society, Series A,* 164, part 3, pp.427-447.

Fienberg, Stephen E. (2005). "Confidentiality and Disclosure Limitation," *Encyclopedia of Social Measurement*, Elsevier, Inc.. Vol. 1, pp. 463-469.

Hundepool, A., L. Willenborg, A. Wessels, L. van Gemerden, S. Tiourine and C. Hurkens. (1998). *μ-Argus User's Manual*. Statistics Netherlands: Voorburg.

King, John (2003). "Recent European Union Legislation for Research Access to Confidential Data: Implementation and Implications," in *Statistical Confidentiality and Access to Microdata: Proceedings of the Seminar Session of the 2003 Conference of European Statisticians*. Geneva, pp. 97-116.

Lane, Julia (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.

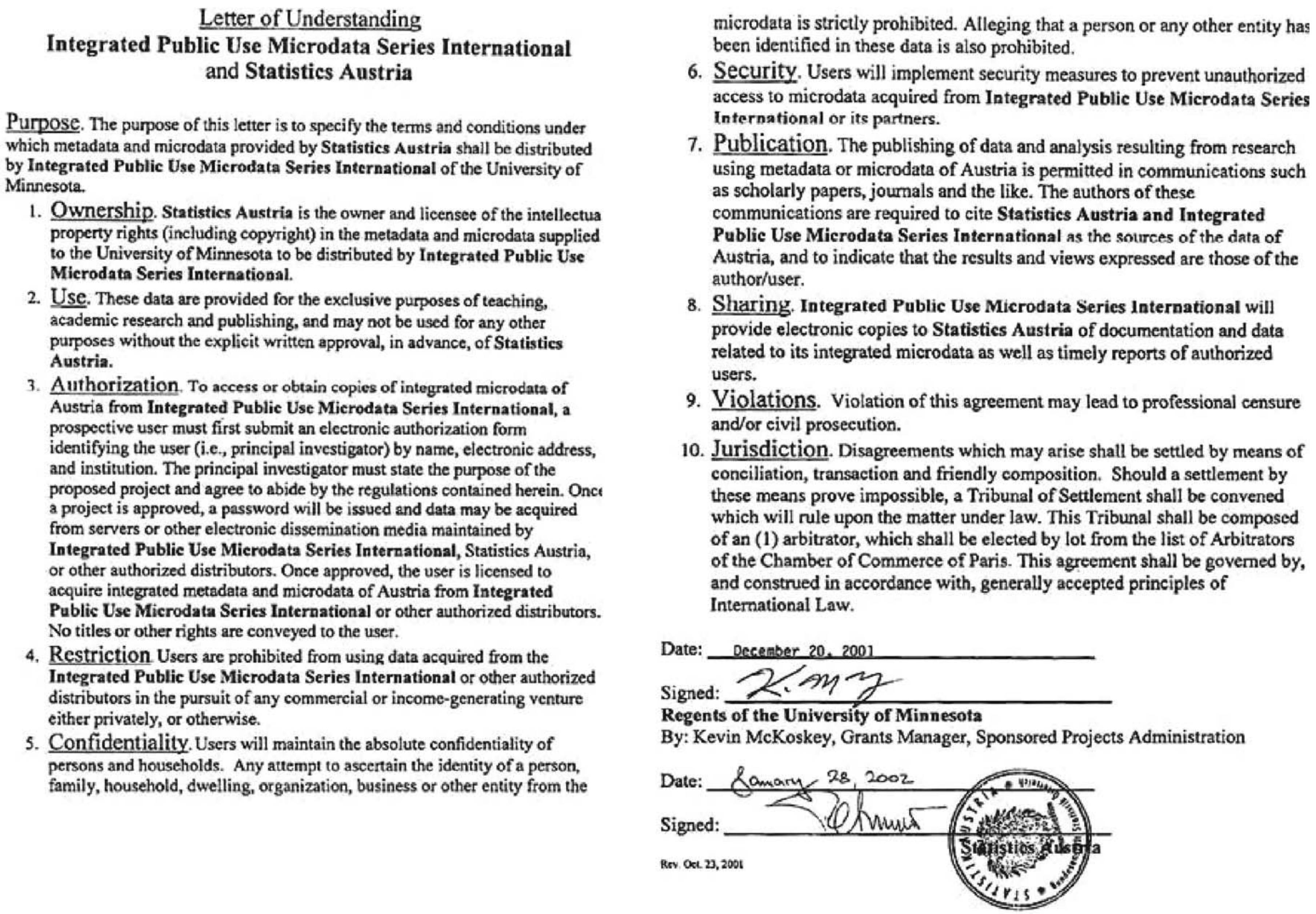
McCaa, Robert and Albert Esteve. (2005). "La integración de los microdatos censales de América Latina: el proyecto IPUMS," *Estudios Demográficos y Urbanos* 20:1(58) 37-70.

McCaa, Robert, and Steven Ruggles. (2002). The Census in Global Perspective and the Coming Microdata Revolution. In Vol. 13, *Nordic Demography: Trends and Differentials, Scandinavian Population Studies*, edited by J. Carling. Oslo: Unipub/Nordic Demographic Society, pp. 7-30.

Polettini, Silvia and Giovanni Seri (2003). "Guidelines for the Protection of Social Microdata Using Individual Risk Methodology: Application within µ-Argus Version 3.2," in *CASC Project: Computational Aspects of Statistical Confidentiality*.

Ruggles, Steven, Miriam King, Deborah Levison, Robert McCaa, and Matthew Sobek. (2003). "IPUMS-International: An Overview". *Historical Methods*, 36: 60-65.

Thorogood, D. (1999). 'Statistical Confidentiality at the European Level.' Paper presented at: Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Thessaloniki, March.



Letter of Understanding
Integrated Public Use Microdata Series International and Statistics Austria

Purpose. The purpose of this letter is to specify the terms and conditions under which metadata and microdata provided by **Statistics Austria** shall be distributed by **Integrated Public Use Microdata Series International** of the University of Minnesota.

1. Ownership. **Statistics Austria** is the owner and licensee of the intellectua property rights (including copyright) in the metadata and microdata supplied to the University of Minnesota to be distributed by **Integrated Public Use Microdata Series International.**
2. Use. These data are provided for the exclusive purposes of teaching, academic research and publishing, and may not be used for any other purposes without the explicit written approval, in advance, of **Statistics Austria.**
3. Authorization. To access or obtain copies of integrated microdata of Austria from **Integrated Public Use Microdata Series International,** a prospective user must first submit an electronic authorization form identifying the user (i.e., principal investigator) by name, electronic address, and institution. The principal investigator must state the purpose of the proposed project and agree to abide by the regulations contained herein. Onc a project is approved, a password will be issued and data may be acquired from servers or other electronic dissemination media maintained by **Integrated Public Use Microdata Series International,** Statistics Austria, or other authorized distributors. Once approved, the user is licensed to acquire integrated metadata and microdata of Austria from **Integrated Public Use Microdata Series International** or other authorized distributors. No titles or other rights are conveyed to the user.
4. Restriction. Users are prohibited from using data acquired from the **Integrated Public Use Microdata Series International** or other authorized distributors in the pursuit of any commercial or income-generating venture either privately, or otherwise.
5. Confidentiality. Users will maintain the absolute confidentiality of persons and households. Any attempt to ascertain the identity of a person, family, household, dwelling, organization, business or other entity from the microdata is strictly prohibited. Alleging that a person or any other entity has been identified in these data is also prohibited.
6. Security. Users will implement security measures to prevent unauthorized access to microdata acquired from **Integrated Public Use Microdata Series International** or its partners.
7. Publication. The publishing of data and analysis resulting from research using metadata or microdata of Austria is permitted in communications such as scholarly papers, journals and the like. The authors of these communications are required to cite **Statistics Austria and Integrated Public Use Microdata Series International** as the sources of the data of Austria, and to indicate that the results and views expressed are those of the author/user.
8. Sharing. **Integrated Public Use Microdata Series International** will provide electronic copies to **Statistics Austria** of documentation and data related to its integrated microdata as well as timely reports of authorized users.
9. Violations. Violation of this agreement may lead to professional censure and/or civil prosecution.
10. Jurisdiction. Disagreements which may arise shall be settled by means of conciliation, transaction and friendly composition. Should a settlement by these means prove impossible, a Tribunal of Settlement shall be convened which will rule upon the matter under law. This Tribunal shall be composed of an (1) arbitrator, which shall be elected by lot from the list of Arbitrators of the Chamber of Commerce of Paris. This agreement shall be governed by, and construed in accordance with, generally accepted principles of International Law.

Date: December 20, 2001

Signed: ______

Regents of the University of Minnesota
By: Kevin McKoskey, Grants Manager, Sponsored Projects Administration

Date: January 28, 2002

Signed: ______
Statistics Austria

Rev. Oct. 23, 2001

OnSite@Home: Remote Access at Statistics Netherlands

Anco Hundepool[*] *and Peter-Paul de Wolf*[**]

[*] **Department of Methods and Informatics, Statistics Netherlands, P.O. Box 4000, 2270 JM Voorburg, The Netherlands, e-mail: ahnl@cbs.nl**

[**] **Department of Methods and Informatics, Statistics Netherlands, P.O. Box 4000, 2270 JM Voorburg, The Netherlands, e-mail: pwof@cbs.nl**

Abstract. In this paper we discuss a pilot project concerning a remote access facility at Statistics Netherlands. We describe some aspects of the technical implementation as well as the functional implementation. Moreover, we will discuss some tentative first experiences of external users.

Keywords. Statistical Disclosure Control, Remote Access, Citrix.

1. Introduction

Statistics Netherlands has a longstanding tradition of releasing safe microdata to researchers. This dates back to the beginning of the nineties of the previous century. The microdata files were made available to researchers at universities under a strict contract. These files were protected against statistical disclosure using a specific set of disclosure control rules that were defined based on the technological circumstances of those years. For many years these files could satisfy the research needs of the universities. The researchers could analyse the microdata files on their own computers.

However, the level of detail in these microdata files made it impossible for certain researchers to perform serious analyses. The restrictions of Statistical Disclosure Control (SDC), enforced by a national Law on Statistics Netherlands, did not allow more detailed microdata files to be made available to researchers outside the premises of Statistics Netherlands. The law demands that the use of and the results from analyses based on detailed microdata files should be under strict control of Statistics Netherlands.

To deal with this situation, the option to work OnSite, i.e., at the premises of Statistics Netherlands, was introduced: the detailed microdata files were made available to selected researchers in a controlled setting. The selected researchers could perform their desired analyses, but the results thereof were checked by Statistics Netherlands' staff for possible disclosure risk, before the researchers were allowed to bring the results outside the controlled setting. This means that only 'final' output (output to be taken home by the researcher) will be checked, whereas intermediate output can be examined by the researcher, but cannot be taken home.

The OnSite facility has proven to be very successful. For quite some time, many researchers have been using the facility. From time to time 5 to 6 researchers were working at the OnSite facility simultaneously.

A major drawback of this facility is that the researchers have to travel to the office of Statistics Netherlands, in order to be able to do their analyses. Even in a small country like the Netherlands this proved to be inefficient in many situations. Moreover, Statistics Netherlands has to organise specially equipped offices for the OnSite researchers.

As more and more facilities became available to use safe internet connections, the question has risen whether an equivalent of the OnSite facility could be build over the internet. This has led to the pilot project OnSite@Home. Only recently we have started a life test of this system as a pilot project with one partner at the University of Tilburg.

In section 2 we will describe the OnSite@Home facility in more detail, both from a functional and a technical point of view. The pilot with the partner at the University of Tilburg started very recently, but we will present some preliminary experiences in section 3. We will conclude in section 4 with some conclusions and remarks on the use of the facility as implemented for the pilot project.

2. OnSite@Home

In this section we will give a description of the Remote Access facility that is used in the pilot OnSite@Home. We will make a distinction between the technical and the functional aspects of the facility.

2.1. Functional aspects

The main idea is that the OnSite@Home facility should resemble the 'traditional' OnSite situation as much as possible, concerning confidentiality aspects. Moreover, it should resemble the look and feel of the OnSite facility without the aspect of having to travel to the premises of Statistics Netherlands. I.e., the following aspects have to be taken into account:

1. Only authorized users should be able to make use of this facility,
2. microdata should remain at Statistics Netherlands,
3. desired output of analyses should be checked on confidentiality,
4. legal measures have to be taken when allowing access.

2.1.1. Only authorised users are allowed

At the OnSite facility access of authorised users only is ensured because researchers cannot leave nor enter the premises of Statistics Netherlands unaccompanied. Moreover, only a selected group of researchers working at universities and similar institutes is allowed to make use of this facility.

The OnSite@Home facility is making use of biometric identification, to ensure that the researcher who is trying to connect to the facility is indeed the intended person. Whenever the researcher wants to access the facility, he will be identified by his fingerprint. For the pilot this is only checked at the start of each session, whereas in the future it is the intention to check the fingerprint at random times during the session as well. For more information about the process of logging on to the facility, see 2.2.1.

Obviously, again only a selected group of researchers will be given permission to make use of this facility.

2.1.2. Microdata remain at Statistics Netherlands

The network that is used by the 'traditional' OnSite facility is not connected to the production network. Moreover, the computers that the researcher can use are such that no removable media can be used (no floppy drive, no USB ports) and no internet connection is possible (no email, no surfing, no ftp, etc.). This means that the microdata used by the researcher can only be accessed using an OnSite computer at the premises of Statistics Netherlands and that the researcher cannot take a copy of the data to his institute. He is able to view the (intermediate) results of his analyses on the screen, but he is not able to send those results to his institute by email or otherwise. Moreover, he is not allowed to take a printout of the results to his institute either, without having it checked by a member of Statistics Netherlands' staff.

The network that is used by the OnSite@Home facility is separate from the production network as well (for the technical implementation, see 2.2). Moreover, the connection that a researcher makes with this facility is a terminal connection: he can only see on his screen what is running on a computer at Statistics Netherlands. He is not able to print or download any of the results of his analyses to his own computer at the institute where he is working. This ensures that the microdata and the intermediate results remain at Statistics Netherlands.

2.1.3. Checking output

Whenever a researcher wants to take output from a session at the 'traditional' OnSite facility to his own institute, the output first needs to be checked by Statistics Netherlands' staff for confidentiality. Only then he will either be allowed to take a printout with him, or the results will be sent to him by email. A researcher that makes use of the OnSite@Home facility is not able to download or print his results either. Whenever he wants to have the results on his own computer at the institute, the results need to be checked by Statistics Netherlands' staff for confidentiality. The results will then be sent to him by email. For the technical implementation, see 2.2.3.

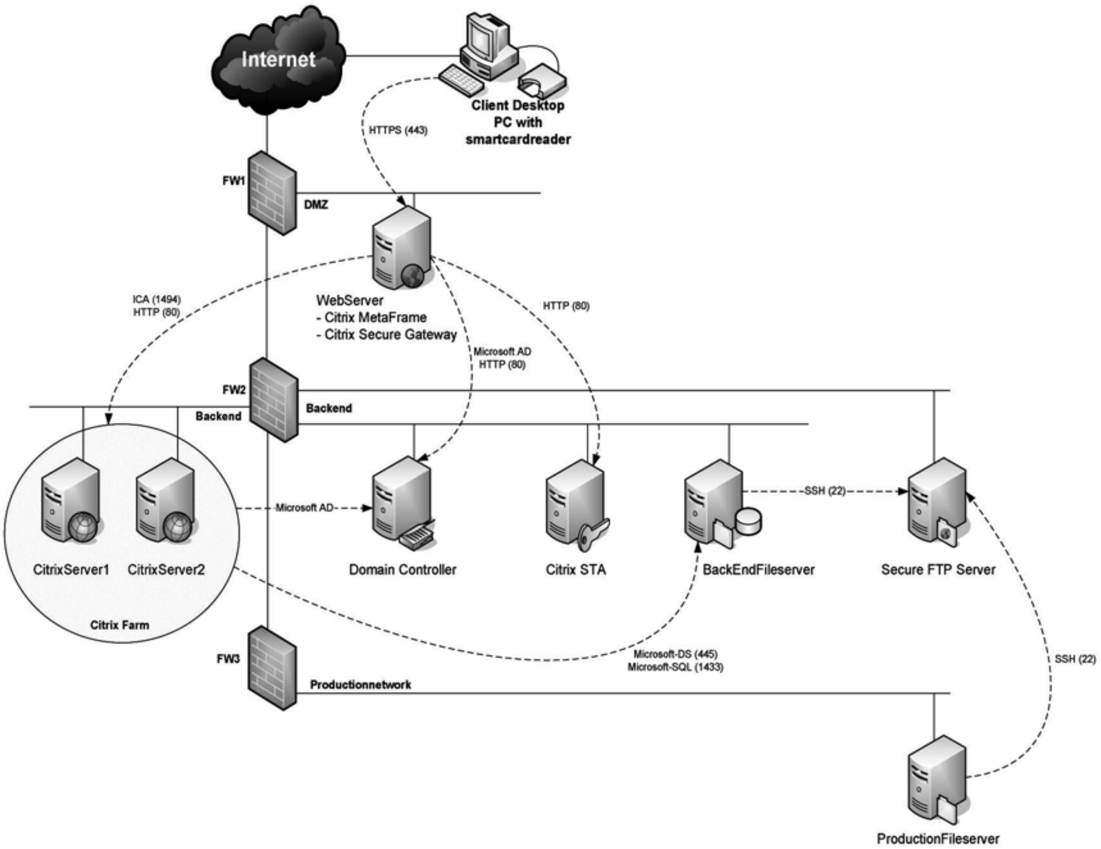
2.1.4. Legal measures

Both in case of the 'traditional' OnSite and the new OnSite@Home facility, legal measures are taken to prevent misuse of the microdata. To that end, a contract will be signed by the institute where the researcher is working. Moreover, a statement of secrecy is signed by the researcher as well as the institute he works for.

2.2. Technical aspects

In Fig. 1 the network representation of the OnSite@Home facility is given. An important fact is that there are three hardware firewalls involved denoted by FW1 through FW3, controlling the connectivity between the 'outside' world, the Onsite@Home facility and the production network of Statistics Netherlands. Each horizontal line extending from a firewall in Fig. 1 is a separate VLAN (virtual local area network). I.e., communication between the different VLAN's is directed through at least one firewall.

Figure 1. Network representation of the OnSite@Home facility



The dashed lines in Fig. 1 denote the only possible connections, where the arrows indicate which computer is allowed to initiate that connection and the text next to those lines denote the allowed type of connection and used port between brackets. E.g., the BackEndFileserver is allowed to set up a secure shell (SSH) connection on port number 22 with the Secure FTP Server. Firewall FW2 ensures that only the BackEndFileserver and the ProductionFileserver are allowed to make a connection with the Secure FTP Server, and that the Secure FTP Server is not allowed to make a connection with any of the other computers in Fig. 1 (not even the BackEndFileserver and the ProductionFileserver).

The three firewalls effectively guard three parts of the complete network of Statistics Netherlands. The first firewall (FW1) controls the access of the 'outside world' to the demilitarized zone (DMZ). The second firewall (FW2) is in between the DMZ and the backend, where several 'intermediate services' are situated, like the Citrix part of the OnSite@Home facility and the e-mail servers for Statistics Netherlands (not displayed in Fig. 1.). The third and final firewall (FW3) separates the backend with the actual production network of Statistics Netherlands. This way it is virtually impossible to directly connect from an external computer to the production network.

2.2.1. The process of setting up a session

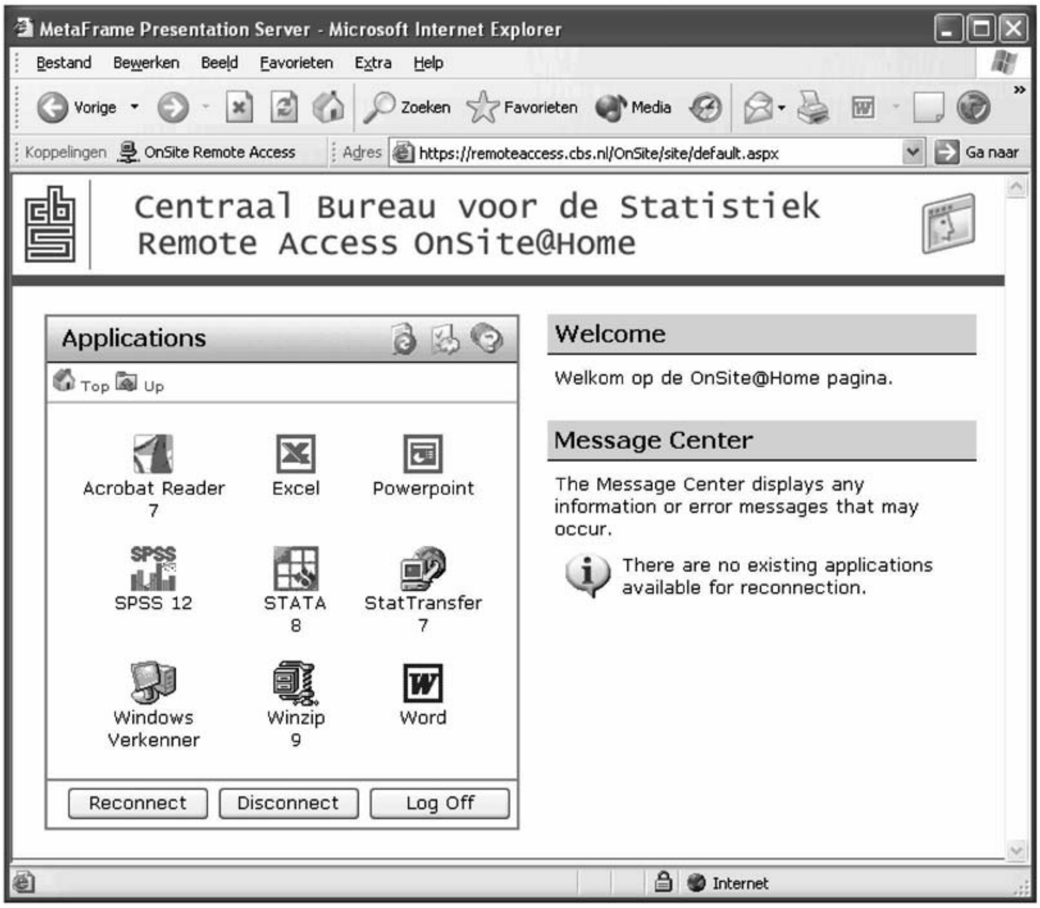
To ensure that only authorised users are allowed to set up a connection with the OnSite@Home facility, biometric identification is used, in combination with PKI[1] certificates. An authorised user is given a smartcard with a personal certificate. He will have to import the public part of that certificate onto his computer; the private part of that certificate is stored on an encrypted section that can only be decrypted by presenting the fingerprint that is also stored on the smartcard. This means that the user will need a smartcard reader that can read the users' fingerprint as well. This reader will be provided by Statistics Netherlands.

Whenever the user wants to start a session, he will have to start the Microsoft Internet Explorer and type the https address of the OnSite@Home facility to try to initiate an SSL[2] connection. Since it is possible that multiple researchers will make use of the same physical computer at the institute to access the OnSite@Home facility, he will then be prompted to choose which certificate he wants to use. Obviously, he is only able to use his own certificate: the private part of his certificate is written on the encrypted section of his smartcard. He will then have to present his finger to the fingerprint/smartcard reader in which the smartcard is inserted. If his fingerprint matches the one on the smartcard, the private part of his personal certificate will be released and sent to the Webserver. This server will check the credentials of the user using the Domain Controller. If everything is correct, the user will be shown the login site of OnSite@Home, using Citrix MetaFrame (a Web Interface). Finally, he has to type in his username and password to enter the main page of the OnSite@Home facility. On that page he will see the applications that he is allowed to use. For an example screenshot of that page, see Fig. 2.

[1] PKI = Public Key Infrastructure

[2] SSL = Secure Sockets Layer

Figure 2. Example screenshot of the main page of the OnSite@Home facility



Welkom op de OnSite@Home pagina.

The Message Center displays any information or error messages that may occur.

There are no existing applications available for reconnection.

2.2.2. Using an application

The researcher now has access to a number of applications and certain microdatafiles needed for his research. Moreover, he will have access to a working environment, in which he can store his intermediate results and/or files.

To start an application, SPSS say, he has to double-click the corresponding icon. The Citrix Secure Gateway then again checks the credentials of the user and asks the Citrix STA (Secure Ticketing Authority) to issue an ICA ticket[3] of limited lifetime. Using that ICA ticket, a secure connection will be established with one of the CitrixServers from the Citrix Farm. Then SPSS will be run on that server within the Citrix Farm. This connection is in effect a terminal connection: only screenshots of the connected server will be transmitted to the computer at the researchers' end.

2.2.3. Checking output for confidentiality

Using the OnSite@Home facility a researcher is able to perform his analyses interactively: he will constantly see what is happening on the terminal server in the Citrix Farm. I.e., he is able to see his intermediate results and to adjust his analyses accordingly. At some point, he would like to have some of his results on his own computer at the institute. Since it is impossible to print or download the results directly, the following procedure is set up.

The researcher places the results in a specific directory within his own working environment. A program, running on the BackEndFileserver, is constantly checking those directories and if anything is found in such a directory, that content will be placed on the Secure FTP Server. At the same time another program, running on the ProductionFileserver, is constantly checking the secure ftp server

[3] ICA = Independent Computing Architecture

for new content. If there is any on the Secure FTP Server, that program will move the content to the ProductionFileserver. In this way, there will never be a direct connection between the BackEndFileserver and the ProductionFileserver.

Once the output arrives on the ProductionFileserver of the production network, a signal is given to specific Statistics Netherlands' staff. The output will then be checked for confidentiality and sent to the corresponding researcher by email within a half working day.

During the pilot, the check on the output for confidentiality is done by hand. Obviously, this is very labour-intensive. In the future, this should ideally be facilitated by some software. However, since the output of the results can be very diverse in format (SPSS, Stata, S-plus, SAS, etc.) the development of such a type of software is very difficult. Moreover, at Statistics Netherlands, no easily automated rules are available at the moment to decide whether or not general analysis' results breach confidentiality.

3. First experiences

The partner at the University of Tilburg (called Netspar) started to participate in this pilot project mid September 2005. At the premises of Netspar, two computers have been prepared for the use of the OnSite@Home facility, i.e., two smartcard/fingerprint readers were installed. Five researchers working at Netspar have been authorised to make use of the OnSite@Home facility and given their own personal smartcards. As far as the workload on the Citrixservers is concerned, this means that a maximum of two users from Netspar can be logged on to the system simultaneously, along with up to six members of Statistics Netherlands' staff (for testing purposes).

So far, no real problems have been encountered with the facility. Both the performance of the system and the look and feel resemble that of working on a state of the art workstation. I.e., it feels like working on your own computer.

4. Conclusions and remarks

The OnSite@Home facility seems to be a promising counterpart of the 'traditional' OnSite facility. Concerning confidentiality issues, both facilities appear to be comparable. The OnSite@Home facility is more flexible in allowing researchers to perform their analyses on microdata from a computer at their own desk, so they can work any time they want. Moreover, no travelling is needed whenever they want to perform additional research. On the other hand, the 'traditional' OnSite facility has the advantage that there will be no intervening colleagues while performing the research.

The technical implementation of the OnSite@Home facility tackles most of the confidentiality issues: the microdata remain at Statistics Netherlands, it is not possible to print or download any results and the final results will be checked for confidentiality before being released to the researcher.

Obviously, during the pilot we will monitor the experiences of the partner at the university of Tilburg. At the end of the pilot, an evaluation report will be written that can be used to further develop the facility and make it more generally available.

This facility can also be used to provide access to Microdata files Under Contract. Currently, those kind of microdata files are protected using statistical disclosure control methods as well as legal measures. These files are provided using CD-ROMs. Using the OnSite@Home facility, these files do not leave Statistics Netherlands, hence the dissemination of the microdata is much more under control. Moreover, a different level of statistical disclosure control might be possible.

ANalytical Data Research by Email and Web (ANDREW)

Vijay Gambhir and Kenneth W. Harris
Research Data Center, National Center for Health Statistics,
Centers for Disease Control and Prevention, Hyattsville,
MD 20782, USA, rdca@cdc.gov

Abstract: The NCHS Research Data Center (RDC), established in 1998, is a facility at the NCHS headquarters in Hyattsville, Maryland, where researchers are granted access to restricted data files, in a secure environment, that are needed to complete approved projects. Restricted data files may contain information, such as lower levels of geography, but do not contain direct identifiers (e.g., name or social security number). Identifiable data include not only direct identifiers such as name, social security number, etc., but also data that can serve to allow inferential identification of either individual or institutional respondents by a number of means.

Although it was envisioned that most of the research work would be performed by the data analysts onsite, with RDC staff closely monitoring to assure that confidential or identifiable data do not leave RDC premises, remote access capabilities were considered an integral part of the fundamental set up of RDC. Thus, the software engineers at RDC designed and developed an e-mail based remote access system, **AN**alytical **D**ata **R**esearch by **E-**mail (**ANDRE)**.

The main objective of ANDRE is to provide a convenient, reliable, economical, and flexible tool for remote data access for statistical analysis. Although ANDRE has served the data users' community very well while strictly adhering to the confidentiality restrictions of NCHS, it does have certain limitations and constraints inherent in its design. Most of these constraints are non-critical but make the system less flexible and less efficient than onsite analysis. Some of these constraints were known at the design and development stage of ANDRE; others have been compiled by RDC staff from interaction with the users and from several years of regular performance analysis of the system. As a result, the RDC now plans to develop and thoroughly test a new remote access system, **AN**alytical **D**ata **R**esearch by **E**-mail and **W**eb (**ANDREW)**. This new system, if successful, will address research needs of data analysts at all levels. It will support multiple statistical languages (SAS, Sudaan, and Stata) and will provide a Graphic User Interface (GUI) for language free statistical analysis. In addition, the system will address the problem of confidentiality risks resulting from cumulative data retrieval through multiple requests from the same user.

1. RDC and data access for statistical analysis

Despite the wide dissemination of its data through publications, CD-ROMs, etc., the inability to release files with, for instance, lower levels of geography, severely limits the utility of some data for research, policy, and programmatic purposes and sets a boundary on one of the Center's goals to increase its capacity to provide state and local area estimates. In pursuit of this goal and in response to the research community's interest in restricted data, NCHS established the Research Data Center (RDC), a mechanism whereby researchers can access detailed data files in a secure environment, without jeopardizing the confidentiality of the respondents.

The NCHS Research Data Center, established in 1998, is a facility at the NCHS headquarters in Hyattsville, Maryland, where researchers are granted access to restricted data files needed to complete approved projects. Restricted data files may contain information, such as lower levels of geography, but do not contain direct identifiers (e.g., name or social security number). Identifiable data include not only direct identifiers such as name, social security number, etc., but also data that can serve to allow inferential identification of either individual or institutional respondents by a number of means.

2. Remote Access/ANalytical Data Research by Email (ANDRE)

Although it was envisioned that most of the research work would be performed by the data analysts onsite, with RDC staff closely monitoring to assure that confidential or identifiable data do not leave RDC premises, remote access capabilities were considered an integral part of the fundamental set up of RDC. The software engineers at RDC designed and developed an email based remote access system that enables researchers to perform **AN**alytical **D**ata **R**esearch by **E**mail. Thus the system was named ANDRE.

The main objective of ANDRE is to provide a convenient, reliable, economical, and flexible tool for remote data access for statistical analysis. Since 1998 ANDRE has served many data analysts flawlessly. A number of researchers have used ANDRE in conjunction with onsite sessions either performing preliminary analysis before onsite sessions or performing post onsite session analysis to wrap up their research. However, most of the analysts have used it for independent statistical research.

3. ANalytical Data Research by Email and Web (ANDREW)

Although ANDRE, the existing remote access system, has served the data users' community very well while strictly adhering to the confidentiality restrictions of NCHS, it does have certain limitations and constraints inherent in its design. This new remote access system will address research needs of data analysts at all levels. It will be built upon the time-tested architecture of its predecessor with a few enhancements to its algorithms. Also, it will incorporate new features to make it robust, very strong on confidentiality issues, more efficient and more flexible.

3.1. Basic Layout

ANDREW will be a fully automated remote access system that will serve registered users around the clock without human intervention. To subscribe to the system, a data user will be required to submit a research proposal. Once the proposal is approved, the user will be provided with login information and guidance as to how to use the system. A registered user may submit data requests from anywhere and at any time. However, results of the data requests will always be released to a specific email address that has been certified as secure and approved by RDC.

The system will use a multilayered authentication procedure to ward off unauthorized access. Upon receipt of a data request, ANDREW will verify the login credentials and subscription status of the requester. Unauthorized communications will be discarded without any response.

A user's program from a validated data request will be scanned online for its suitability for execution by ANDREW. To ensure smooth operations, it will not allow certain commands and words in a user's programs, especially those that can create permanent datasets or files on ANDREW's disk space or interfere in any way with the underlying operating system.

To deal with the issues of disclosure limitation, ANDREW will use prevention as well as suppression techniques. To prevent disclosure violations, it will not allow certain commands (e.g., print command in SAS) that have little, if any, statistical value. Also, it will modify certain commands in the user's program to prevent the output of sensitive information. For example, it will modify the "proc means" command so that it does not produce minimum and maximum values. However, there are certain commands that are important for statistical studies and that do generate output that cannot be released. ANDREW will use a variety of enhanced suppression algorithms to prevent disclosure violations. For example, it will white out extreme values resulting from proc univariate and will use state of the art commercial software packages to suppress certain low values in one-way and two-way tables with a special emphasis on prevention of inferential disclosure violations.

3.2. New Features

3.2.1. Data Security

The technological platform used for the development of ANDRE is almost obsolete. ANDREW will be designed using MS Visual Studio. The GUI front end will be implemented using C# (C Sharp) and the backend will be a product from a leading corporation which has proven technical know-how as well as commitment and resources to keep the platform secure by issuing security patch ups to keep the product safe.

3.2.2. Robust disclosure limitation checks

Although the suppression algorithms employed by ANDRE have performed well, certain limitations and constraints have made the system less effective. For example, ANDRE works directly on the SAS output (.lst files) and has to negotiate labels, formatting characters, and background information in order to evaluate statistical values and apply home grown suppression algorithms. By contrast, ANDREW will develop a set of mapping algorithms that will extract values from the SAS/Sudaan/Stata outputs and save them in external files in an appropriate format without any background information. The system will invoke a well recognized commercial suppression software package. Another set of algorithms will put the validated values back into the original SAS/Sudaan/Stata outputs.

3.2.3. Cumulative data retrieval and confidentiality

No solution has yet been found to the classic problem of a user submitting a series of data requests through a remote access system, each time getting different bits and pieces of data and eventually getting sufficient information to identify entities in the data set. However, ANDREW will start addressing this problem by tracking the amount and type of data released about certain risky/critical variables. A committee of confidentiality experts will examine each data set and identify variables that have disclosure violation potential. For each variable a tolerance level will be defined. ANDREW will examine data being released for each risky variable. As soon as the amount of data released for any variable reaches its tolerance level, ANDREW will issue an alert to the system administrator and generate a report displaying all the data releases for that variable.

3.2.4. Accessibility

Since the user interface of ANDREW will have no confidentiality risks, it will be a Web based component. This feature will not only give wider accessibility to the system but will also allow a user to get his/her program parsed interactively for suitability before it is accepted for execution by ANDREW. However, all of the confidential data along with the main resident component of ANDREW will reside on a set of machines physically located in RDC's secure area. The main resident component of the system will receive the parsed and validated users' programs from the web component, execute them in the secure environment and send the results to the appropriate users.

3.2.5. Multi-language support

Unlike SAS, other statistical languages such as Sudaan produce a very rich variety of output patterns. Remote access systems rely heavily on pattern recognition algorithms to identify disclosure violations. With older technology matching is done at the character level, whereas C# has a built in feature called regular expressions that can detect a variety of patterns at a block level rather quickly. This will allow the system to deal with all kinds of patterns generated by Sudaan and other languages in a timely fashion.

3.2.6. Language free GUI data access

This feature of ANDREW will implement a Graphic User Interface (GUI) that will allow specification of desired variables and the constraints by a few mouse clicks. GUI will be very useful to the users whose SAS skills are rusty/non-existent or who want to get quick results. Use of GUI technology has futuristic implications as it will give a lot more control on what a user can specify compared with the current approach of giving free hand to a user (via his or her SAS code) once the data set is approved. The confidential data will also reside in RDC's secure area and appropriate SAS/Sudaan/Stata codes will be generated and executed in the secure environment.

4. Summary

Despite a number of inherent limitations and constraints, RDC's initial remote access system, ANDRE, has performed very well since its inception in 1998. However, with an ever growing recognition that new and improved technologies are needed, the RDC has undertaken to develop ANDREW, the next generation remote access system. When successfully completed, ANDREW will represent a major advancement in the area of remote data access. Even so, it should not be considered the final product. It is a continuous software engineering process geared towards regular induction of improvements and enhancements to the system as dictated by the feedback from the user community and by the internal system performance analysis.

Topic II

Disclosure risk, information loss and usability data

Assessing Risk in Statistical Disclosure Limitation

Silvia Polettini, Julian Stander***

*** Dipartimento di Scienze Statistiche, Università di Napoli Federico II, V. L. Rodinò 22, I-80138 Napoli, Italy (spolettini@unina.it)**

**** School of Mathematics and Statistics, University of Plymouth, Drake Circus,
Plymouth, PL4 8AA,
UK (jstander@plymouth.ac.uk)**

Abstract. When microdata files for research are released, it is possible that external users may attempt to breach confidentiality. For this reason most National Statistical Institutes apply some form of disclosure risk assessment. Our disclosure risk measure is based on re-identification and is specific to cells of the contingency table built by cross tabulating the variables that allow identification. We discuss five Bayesian hierarchical models for risk estimation. Model I leads to the negative binomial distribution for the population cell counts given the sample cell counts as advocated by Benedetti and Franconi (1998). Model II is discussed in Rinott (2003) and is based on the one proposed by Bethlehem, Keller and Pannekoek (1990). Model III is an extension of Model II due to Polettini and Stander (2004) that allows extra variation by modelling the sample cell probabilities. Model IV is an extension of Model III that takes account of the large number of empty cells and that makes use of the available estimate of sample cell probabilities based on sampling design weights. Each of these models assumes independence across the cells of the contingency table. Model V makes some use of the structure of the contingency table by assuming a Dirichlet-multinomial-multinomial formulation. The parameters of the Dirichlet prior are elicited from available marginal tables by means of log-linear models. We discuss each model in detail and compare their performance by using an artificial sample of the Italian 1991 Census data, drawn by means of a widely used, unequal probability, sampling scheme.

Keywords. Bayesian hierarchical models, confidentiality, disclosure risk, empirical Bayes, MCMC

1. Introduction

When microdata files for research are released, it is possible that external users may attempt to breach confidentiality. For this reason National Statistical Institutes apply some form of disclosure risk assessment. Risk assessment first requires a measure of disclosure risk to be defined; as this is usually cast in terms of population quantities, risk estimation is then achieved by introducing suitable statistical models. If the estimated risk is considered not tolerable, protection measures must be put into practice.

We base our definition of disclosure on the concept of re-identification; see Willenborg and de Waal (2001). Therefore by disclosure we mean a correct record re-identification operation that is achieved by an intruder when comparing a target individual in a sample with an available list of units that contains individual identifiers.

Even when attention is focused on re-identification disclosure, different approaches to risk assessment can be pursued. We focus on individual or combination-level risk measures, as defined in Benedetti and Franconi (1998), Skinner and Holmes (1989), Carlson (2002), and Elamir and Skinner (2004) among others. A routine for computing a measure of individual risk of disclosure is now implemented in the software μ-Argus, developed under the European Union project CASC on Computational Aspects of Statistical Confidentiality.

In social surveys, the observed variables are frequently categorical in nature, and often comprise publicly available variables. If these variables allow identification they are referred to as *key variables*. Risk is usually defined as a function of *combinations* of values of key variables. These correspond to cells of a contingency table built by cross-tabulating the key variables. Records presenting combinations of key variables that are rare in the population clearly have a high disclosure risk, whereas rare or even unique combinations in the sample do not necessarily correspond to high risk individuals.

Benedetti and Franconi (1998) estimate a record-level measure of re-identification risk within a Bayesian framework. Bayesian risk estimation is also presented in Fienberg and Makov (1998), Omori (1999) and Takemura (1999), among others. Benedetti and Franconi noticed that $1/F_k$ is the probability of re-identification of individual i in cell k, $k = 1,\dots, K$, when F_k individuals in the population are known to belong to this cell. In order to infer the population cell frequency F_k from its sample frequency f_k, they then focused on the posterior distribution of F_k given f_k. Finally, they define the risk as the expected value of $1/F_k$ under this distribution. This proposal aroused a large debate; see Di Consiglio, Franconi and Seri (2003), Polettini (2003) and Rinott (2003). In this paper we build on previous work to discuss a variety of Bayesian hierarchical models for risk estimation. For these models we derive the posterior distribution $[F_k \mid f_1,\dots, f_K]$ of the population frequency given the observed sample frequencies for each combination k of values of the key variables. Knowledge of this distribution enables us to obtain suitable summaries that can be used to estimate the risk of disclosure; one such summary is $\mathrm{E}[1/F_k \mid f_1,\dots,f_K]$, but different summaries, such as the mode or the median, can perform better. The methodology adopted in the paper follows a superpopulation approach similar to that used in Bethlehem, Keller and Pannekoek (1990), where a Poisson-gamma model is first proposed (see also Rinott, 2003); Skinner and Holmes (1998) suggest instead using a Poisson-lognormal model, while Carlson (2002) and Elamir and Skinner (2004) adopt a different, yet related approach.

The paper is organised as follows. In Section 2 we discuss the Benedetti and Franconi (1998) approach and show that it is equivalent to a superpopulation model, which we call Model I. In Section 3 we present Model II, which is based on the one discussed by Bethlehem, Keller and Pannekoek (1990), and show that Model I is a limiting case of it. Model III was introduced in Polettini and Stander (2004) and is presented here in Section 4. In order to assess the risk estimates from each model, we use an artificial sample, drawn from the 1991 Italian Census data according to the sampling scheme of the Labour Force Survey, so that we know the population frequencies. This data set is discussed in Section 5. In Section 6 we refine Model III to produce Model IV. In Section 7 we discuss the estimated risks that we obtain when we apply Models III and IV to the data set. In Section 8 we introduce Model V. Unlike Models I to IV, that assume independence across cells, this model makes some use of the structure of the contingency table by assuming a Dirichlet-multinomial-multinomial formulation. The parameters of the Dirichlet prior are elicited from available marginal population tables by means of log-linear models. Estimates from this model are compared with those obtained from Model IV. Finally, in Section 9 we briefly discuss avenues for further work.

2. Model I

Let

$$\pi_k = P(\text{a member of the population falls into cell } k), \text{ and}$$
$$p_k = P(\text{a member of population cell } k \text{ falls into the sample}), k = 1,\dots, K,$$

where K is the number of combinations in the population. Let the microdata file be a random sample of size n drawn from a finite population of N units.

As mentioned above, Benedetti and Franconi (1998) estimate the risk for cell k as a posterior expectation $r_k = E(1/F_k \mid f_k) = \sum_{h \geq f_k} \Pr(F_k = h \mid f_k)/h$. They do not, however, formally define a hierarchical model for the population and cell frequencies. They implicitly assume that, given f_k, F_k is independent of $(f_1,\dots, f_{k-1}, f_{k+1},\dots, f_K)$ and that $F_k \mid f_k \sim$ negative binomial(f_k, p_k). They then estimate p_k using the sampling design weights as

$$\hat{p}_k = \frac{f_k}{\hat{F}_k^D}, \tag{1}$$

where $\hat{F}_k^D = \sum_{i \in C_k} w_i$, in which w_i^{-1} is the probability that unit i is included in the sample and C_k is the set of records in the sample that belong to class k. Sometimes the sampling weights are calibrated to match known population totals on a set of auxiliary variables, that might be different from the key variables; see Deville and Särndal (1992) and Di Consiglio, Franconi and Seri (2003). This can lead to problems with Benedetti and Franconi's estimate of disclosure risk; see Rinott (2003) and Di Consiglio, Franconi and Seri (2003) for detailed discussions.

Benedetti and Franconi's assumption is equivalent to the following superpopulation model, that we shall call **Model I**:

$$\begin{aligned} \pi_k &\sim m(\pi_k) \propto 1/\pi_k, k = 1, \ldots, K, \\ F_k \mid \pi_k &\sim \text{Poisson}(N\pi_k), F_k = 0, 1, \ldots, \\ f_k \mid F_k, \pi_k, p_k &\sim \text{binomial}(F_k, p_k), f_k = 0, 1, \ldots, F_k, \text{ independently across cells.} \end{aligned}$$

The equivalence is due to the fact that $F_k \mid f_k, p_k \sim$ negative binomial(f_k, p_k); see Rinott (2003) and Franconi and Polettini (2004) for a more detailed discussion. In general, any unknown parameters of such a model may be estimated using an empirical Bayesian (EB) approach; see Efron and Morris (1973). For this model, the EB approach would be based on the log-likelihood $\sum_{k=1}^{K} \log[f_k]$, in which $[f_k]$ is the marginal probability mass function. However, since $[f_k]$ is improper, the EB approach is not feasible here.

3. Model II

We now present another superpopulation model, which we shall call **Model II**. This model is an extension of the one proposed by Bethlehem, Keller and Pannekoek (1990) and is also discussed by Rinott (2003). It takes the form

$$\begin{aligned} \pi_k &\sim \text{gamma}(\alpha, K\alpha), k = 1, \ldots, K, \\ F_k \mid \pi_k &\sim \text{Poisson}(N\pi_k), F_k = 0, 1, \ldots, \\ f_k \mid F_k, \pi_k, p_k &\sim \text{binomial}(F_k, p_k), f_k = 0, 1, \ldots, F_k, \text{ independently across cells,} \end{aligned}$$

in which $\alpha > 0$ is an unknown parameter. The assumption that $\pi_k \sim \text{gamma}(\alpha, K\alpha)$ implies $\text{E}[\pi_k] = 1/K$, ensuring that on average the π_ks sum to 1, and $\text{Var}[\pi_k] = 1/(K^2\alpha)$. Since K is usually very large in real applications, the associated small variance means that the gamma hyperprior is very strongly concentrated on its mean, which is itself small. Hence Model II does not allow much variation across cells. It turns out that Model I can be thought of as the limit of Model II as $\alpha \to 0$ (Rinott, 2003). Hence Model I allows for more variation across cells than Model II. Parameter estimation is problematic for Model II as there are $K+1$ parameters and K data points. $[f_k]$ becomes improper as $\alpha \to 0$. So even under the simplifying assumption of equal probability sampling (i.e. $p_k = n/N$ for all k) as adopted by Rinott (2003), the EB approach may not work well.

4. Model III

In an attempt to allow extra variation Polettini and Stander (2004) extended Model II by modelling the p_k. Their model takes the form:

$$\begin{aligned} \pi_k &\sim \text{gamma}(\alpha, K\alpha), \pi_k > 0, k = 1, \ldots, K, \\ F_k \mid \pi_k &\sim \text{Poisson}(N\pi_k), F_k = 0, 1, \ldots, \\ p_k &\sim \text{beta}(a_k, b_k), 0 < p_k < 1, \\ f_k \mid F_k, \pi_k, p_k &\sim \text{binomial}(F_k, p_k), f_k = 0, 1, \ldots, F_k, \text{ independently across cells,} \end{aligned}$$

and will be referred to as **Model III**. Here $a_k > 0$ and $b_k > 0$ are unknown parameters. Polettini and Stander (2003) present the probability mass function of F_k given f_k and the probability mass function of f_k:

Distribution 1 *The probability mass function of* F_k *given* f_k *is*

$$\begin{aligned} &[F_k \mid f_k] = \\ &\frac{(K\alpha)^{\alpha+f_k}\Gamma(a_k+b_k+f_k)}{\Gamma(b_k)\Gamma(\alpha+f_k)\,{}_2F_1\left(\alpha+f_k, a_k+f_k; a_k+b_k+f_k; -\frac{N}{K\alpha}\right)} \times \\ &\frac{N^{F_k-f_k}}{(N+K\alpha)^{\alpha+F_k}} \frac{\Gamma(\alpha+F_k)\Gamma(F_k-f_k+b_k)}{\Gamma(a_k+b_k+F_k)\Gamma(F_k-f_k+1)}, \quad F_k = f_k, f_k+1, \ldots, \end{aligned}$$

where ${}_2F_1\left(a, b; c; z\right)$ denotes the Hypergeometric function (Abramowitz and Stegun, 1965).

Dstribution 2 *The probability mass function of* f_k *is*

$$\begin{aligned} &[f_k] = \\ &\frac{\Gamma(b_k)}{\Gamma(\alpha)B(a_k, b_k)} \left(\frac{N}{K\alpha}\right)^{f_k} \frac{\Gamma(\alpha+f_k)\Gamma(a_k+f_k)}{\Gamma(f_k+1)\Gamma(a_k+b_k+f_k)} \times \\ &{}_2F_1\left(\alpha+f_k, a_k+f_k; a_k+b_k+f_k; -\frac{N}{K\alpha}\right), \quad f_k = 0, 1, \ldots. \end{aligned}$$

Polettini and Stander (2004) set $a_k = a$ and $b_k = b$ so that effectively the beta distribution of p_k – and therefore the risk – is not cell specific. In Section 6 we will discuss a model in which the distribution of p_k is cell specific. The EB approach may be problematic for Model III for the same reasons as for Models I and II. In fact Polettini and Stander (2004) report problems maximizing the associated log-likelihood over (α, a, b). They do, however, arrive at estimates of these parameters that are sensible in terms of goodness of fit criteria.

5. The Data Set

The data that we consider are an artificial sample of $n = 53{,}872$ records drawn from the 1991 Italian Census data according to the complex sampling scheme of the Labour Force Survey, as described in Di Consiglio, Franconi and Seri (2003).

The data come from four administrative Italian regions, namely Campania, Lazio, Val d'Aosta and Veneto. The total number of individuals in the population from these four regions is $N = 15{,}142{,}320$. Among the many variables collected in the Census, we chose the following as key variables: sex (2 categories), age (recoded in 14 classes), region of residence (the 4 regions just mentioned), position in profession (14 categories) and relationship with the head of the household (13 categories), giving $K = 2\times14\times4\times14\times13 = 20,384$. Since this is an instance where the population cell frequencies F_k are known, the data allow the proposed procedures to be assessed by comparing known population quantities with their corresponding estimates.

6. Model IV

We decided to modify Model III for two reasons. First, we wanted to account for the large number of empty cells. Following Skinner and Holmes (1993) and Carlson (2002), we did this by assuming that the p_k s are drawn independently from a mixture of a beta distribution and a point mass at zero, with the weight given to the beta distribution being $\gamma \in [0,1]$. Secondly, in our application we always deal with complex, unequal probability, sampling schemes that we try to approximate by assuming $f_k \mid F_K, p_k \sim \text{binomial}(F_k, p_k)$. These models could be informed by using the known sampling weights. For this reason we made use of the $\hat{p}_k$ defined in equation (1) and used by Benedetti and Franconi (1998). If we set $a_k = v\hat{p}_k$ and $b_k = v(1-\hat{p}_k)$ for some unknown positive parameter v to be estimated, then the $\text{beta}(a_k, b_k)$ distribution has mean $\hat{p}_k$ and variance $\hat{p}_k(1-\hat{p}_k)/(v+1)$. Hence, the $\text{beta}(a_k, b_k)$ is now located around the estimated $\hat{p}_k$ and is thus cell specific. **Model IV** takes the form:

$$
\begin{aligned}
\pi_k &\sim \text{gamma}(\alpha, K\alpha), \pi_k > 0, k = 1,\dots,K,\\
F_k \mid \pi_k &\sim \text{Poisson}(N\pi_k), F_k = 0,1,\dots,,\\
p_k &\sim \gamma\,\text{beta}(v\hat{p}_k, v(1-\hat{p}_k)) + (1-\gamma)\delta_{\{0\}}(p_k), p_k \in [0,1],\\
f_k \mid F_k, \pi_k, p_k &\sim \text{binomial}(F_k, p_k), f_k = 0,1,\dots,F_k, \text{ independently across cells,}
\end{aligned}
$$

in which the delta function $\delta_{\{0\}}$ is such that $\delta_{\{0\}}(0) = 1$ and $\delta_{\{0\}}(p_k) = 0$ for $p_k \in (0,1]$. Clearly when $\gamma = 1$, we recover Model III, if a_k and b_k are as just defined.

It can be shown that for $f_k > 0$ the probability mass function $[F_k \mid f_k]$ remains the same as Distribution 4 with a_k and b_k as just defined. Unless $\gamma = 1$, there is a change to Distribution 2:

Distribution 3 *The probability mass function of* f_k *is now*

$$
[f_k] = \begin{cases}
\gamma \;\; {}_2F_1\left(\alpha, v\hat{p}_k; v; -\dfrac{N}{K\alpha}\right) + (1-\gamma) & \text{if } f_k = 0\\[2ex]
\gamma \dfrac{\Gamma(v(1-\hat{p}_k))}{\Gamma(\alpha)B(v\hat{p}_k, v(1-\hat{p}_k))}\left(\dfrac{N}{K\alpha}\right)^{f_k} \dfrac{\Gamma(\alpha+f_k)\Gamma(v\hat{p}_k+f_k)}{\Gamma(f_k+1)\Gamma(v+f_k)} & \\[2ex]
\quad \times {}_2F_1\left(\alpha+f_k, v\hat{p}_k+f_k; v+f_k; -\dfrac{N}{K\alpha}\right) & \text{if } f_k > 0.
\end{cases}
$$

7. Applying Models III and IV to the Data Set

Because of the already mentioned problems associated with parameter estimation via an EB approach, we perform a fully-Bayesian analysis. For Model III we have to select values for the parameters α, a and b. We can use the expression for $[f_k]$ given in Distribution 2 to assess goodness of fit and so to select model parameters. This led us to choose $\alpha = 0.92$, $a = 0.86$ and $b = 80$. For Model IV the parameters that we have to specify are α and v. Again, we can use the expression for $[f_k]$ given in Distribution 3 for $f_k > 0$ to assess goodness of fit. This led to the choice of $\alpha = 0.1$ and $v = 80$.

Figure 1. Scatter plots of the disclosure risks estimated using Model III (in grey) and Model IV (in black) against the true risk $1/F_k$. The left panel is for the Val d'Aosta region, while the right panel is for three large regions Campania, Lazio and Veneto. Logarithmic scales are used for all axes.

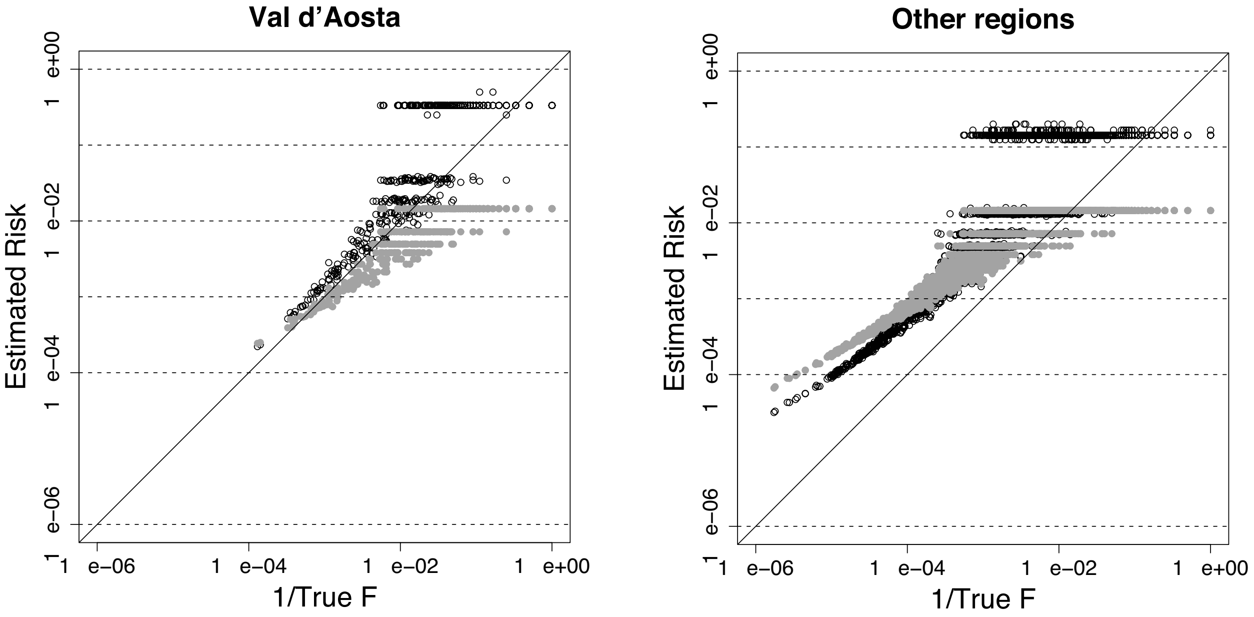


Figure 1 shows the estimated disclosure risk obtained using Model IV and Model III, plotted against the known disclosure risk $1/F_k$. We present the results for the smaller Val d'Aosta region and the three larger regions Campania, Lazio and Veneto separately. This is because the sampling design weights and hence the p_k s have considerably different levels in the smaller and larger regions. Indeed, in order to achieve the same precision of estimation across regions, sampling in Val d'Aosta is more than proportional to region's size. Model IV offers some improvement over Model III. We observe the desirable feature that high risks are generally no longer underestimated, even though we now observe overestimation. There is also a more appropriate spread in the estimated disclosure risk. Small risks tend to be overestimated, although using Model IV can reduce the extent of overestimation especially in the three large regions.

8. Model V

All the above models assume independence across cells. We believe that further improvements can be achieved by making some use of the structure of the contingency table. We could, for example, assume that

$$\underline{\pi} \sim \text{Dirichlet}(\alpha_1,\dots,\alpha_K),$$

where $\underline{\pi} = (\pi_1,\dots,\pi_K)$. We could then take

$$\underline{F} \mid \underline{\pi} \sim \text{multinomial}(N;\pi_1,\dots,\pi_K),$$
$$\underline{f} \mid \underline{F} \sim \text{multinomial}(n;F_1/N,\dots,F_K/N),$$

for example. This approach is exactly the one suggested in Polettini and Stander (2004), except that here the assumption of equality of the parameters in the Dirichlet distribution has been relaxed. We shall refer to this Dirichlet-multinomial-multinomial model as **Model V**. It offers the simplest way of introducing information from external archives within a Bayesian framework.

When no information other than the sample is available, we suggest that the sampling design weights could be exploited by taking $\alpha_k \propto \hat{F}_k^D$. If data collected at a previous census were to be available, we could take $\alpha_k \propto F_k^{\text{previous}}$. If only marginal tables were available, we could specify a conditional independence log-linear model corresponding to these marginal tables to elicit the $(\alpha_1,...,\alpha_K)$ parameters. We have experimented with two such log-linear models:

1. log-linear model 1: `F~sex+(rel+age+posprof)^3`;
2. log-linear model 2: `F~rel+(sex+age+posprof)^3`.

A Dirichlet-multinomial approach is also proposed in Forster and Webb (2005), in which a Bayesian model averaging methodology for disclosure risk assessment is presented. We perform inference about Model V using Markov chain Monte Carlo methods; see Gilks, Richardson and Spiegelhalter (1996), for example. We have implemented these methods in WinBUGS; see `://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml` and Congdon (2005), for example. We have also written our own routines in R (R Development Core Team, 2005) for confirmation.

8.1. Results for Model V

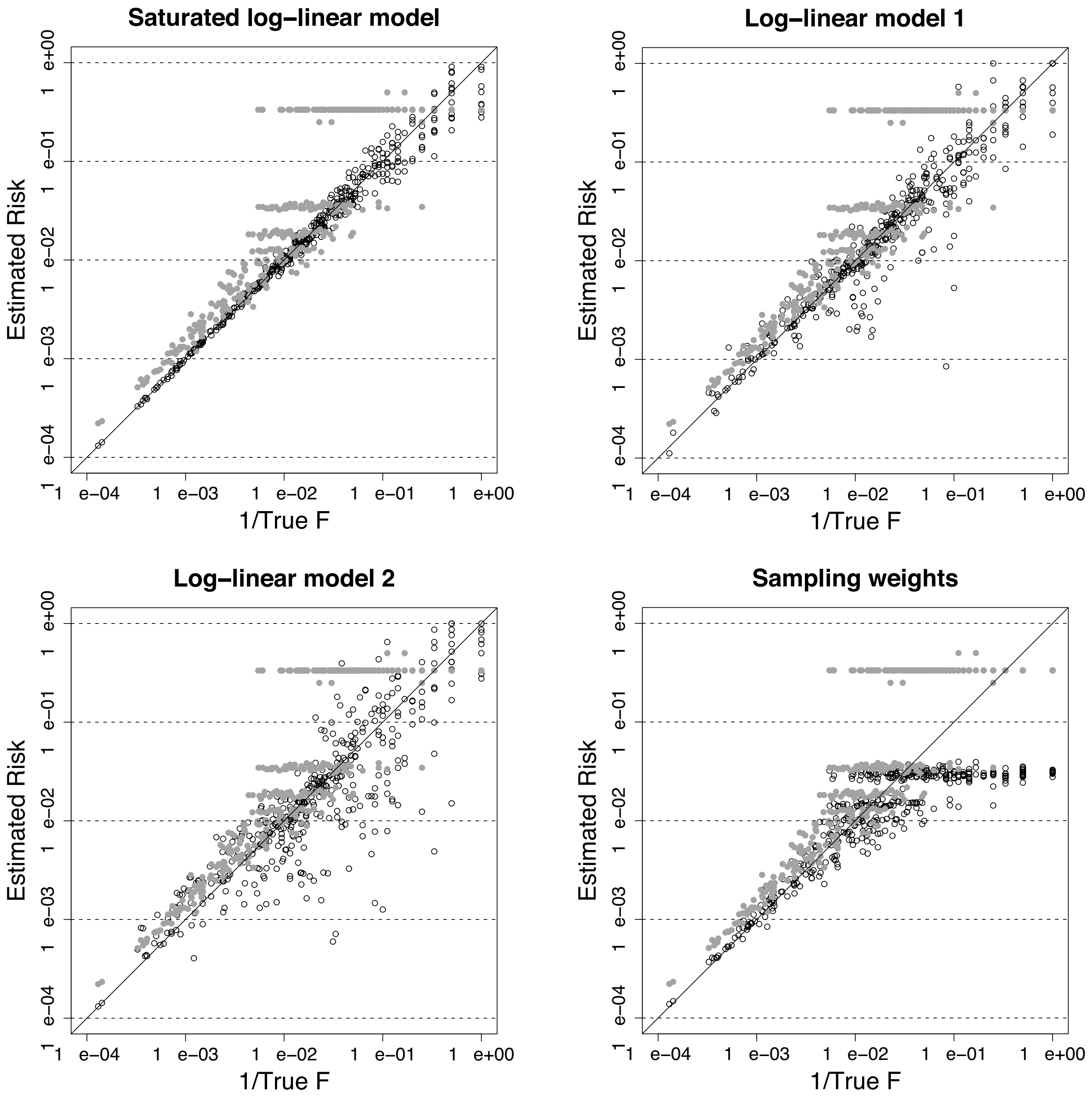
We present estimates of disclosure risk obtained using Model V for the Val d'Aosta region only. Val d'Aosta has a large number of potentially unsafe cells since 44% of the cells in the sample are sample uniques and 77% lie in the range 1 to 5. However, of the sample unique cells, only 4% correspond to a population unique. This large number of small cells is due to high sampling rates which cause individuals from cells with relatively low population frequencies F_k to fall into the sample. Hence, it is important to take proper account of high sampling fractions to achieve accurate risk estimation so as to be able to distinguish between safe and unsafe records in the sample. The top-left panel of Figure 2 was produced using $\alpha_k = F_k$. These are in fact the fitted-values from a saturated log-linear model. Estimation is clearly very good, even for high values of disclosure risk. This result is, however, overly optimistic, as the F_ks would not be known in practice. The top-right panel of Figure 2 was produced using $\alpha_k = \hat{F}_k$, the fitted values from log-linear model 1. Not unexpectedly, the performance is worse. The bottom-left panel of Figure 2 was produced using $\alpha_k = \hat{F}_k$, the fitted values from log-linear model 2. Performance is less good than when using log-linear model 1. Finally, the bottom-right panel of Figure 2 was produced using $\alpha_k = \hat{F}_k^D$, the estimates based on the sampling design weights. Performance is quite poor since high risks are considerably under-estimated. In Figure 2 we have also superimposed the disclosure risk estimates from Model IV in grey.

Following Forster (2005), we assess the performance of the method by considering it as a classifier, defining a cell as unsafe if its risk is greater than 0.05. We present the associated sensitivities and specificities in Table 1. Although it is difficult to recommend a 'best' approach, we prefer Model V because it performs well and allows us to take into account information from external archives.

Table 1. Performance of Models IV and V as risk classifiers. The associated sensitivities and specificities are given.

Classifier	Sensitivity	Specificity
Model V: saturated log-linear	0.94	0.97
Model V: log-linear model 1	0.84	0.96
Model V: log-linear model 2	0.71	0.94
Model V: sampling design weights	0.00	1.00
Model IV	0.96	0.74

Figure 2. Scatter plots of the disclosure risks estimated using Model V against the true risk $1/F_k$. In the top-left panel $\alpha_k = F_k$, the fitted-values from a saturated log-liner model; in the top-right panel $\alpha_k = \hat{F}_k$, the fitted values from log-linear model 1; in the bottom-left panel $\alpha_k = \hat{F}_k$, the fitted values from log-linear model 2; finally, in the bottom-right panel $\alpha_k = \hat{F}_k^D$. The risk estimates from Model IV have been superimposed in grey.



9. Discussion

We have presented and discussed the performance of a variety of models for assessing risk in statistical disclosure limitation. Our assessment of the performance of Model V is optimistic because we have made use of the population frequencies that we are trying to estimate to inform the Dirichlet prior. The idea is, however, to use data from a previous census, if it were accessible. We believe that such data will soon be made available to us. We also plan to use updated projections of population totals from published demographic archives. Finding ways of using all the relevant archive information is a challenging problem.

Often, the assumption of a Dirichlet prior in Model V is criticised as being too restrictive; see O'Hagan and Forster (2004) for discussion. If the prior variance is too low for example, robustness to prior specification may be lacking. In fact, our application shows that the specification of the prior has a large impact on the risk estimates. Accordingly, we plan to examine ways of relaxing the Dirichlet assumption. Finally, we need to improve the performance of our MCMC samplers.

Acknowledgements

The authors would like to thank Luisa Franconi for useful discussions.

References

Abramowitz, M. and Stegun, I.A. (1965). *Handbook of Mathematical Functions*. Dover, New York.

Benedetti, R. and Franconi, L. (1998). Statistical and technological solutions for controlled data dissemination. In *Pre-proceedings of New Techniques and Technologies for Statistics*, volume 1, pages 225–232. Sorrento, June 1998.

Bethlehem, J., Keller, W. and Pannekoek, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association*, **85**, 38–45.

Carlson, M. (2002). Assessing microdata disclosure risk using the Poisson-inverse Gaussian distribution. *Statistics in Transition*, **5**, 901–925.

Deville, J. C. and Särndal, C.E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87**, 367–382.

Di Consiglio, L., Franconi, L. and Seri, G. (2003). Assessing individual risk of disclosure: an experiment. In *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*. Luxembourg, April 2003.

Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors — an empirical Bayes approach. *Journal of the American Statistical Association*, **68**, 117–130.

Elamir, E. A. H. and Skinner, C. J. (2004). Modelling the re-identification risk per record in microdata. Technical report, Southampton Statistical Sciences Research Institute, University of Southampton, UK.

Fienberg, S. E. and Makov, U. E. (1998). Confidentiality, uniqueness, and disclosure limitation for categorical data. *Journal of Official Statistics*, **14**, 385–397.

Forster, J. J. (2005). Bayesian methods for disclosure risk assessment. Joint UNECE/Eurostat work session on statistical data confidentiality. Geneva, Switzerland, 9–11 November 2005.

Forster, J. J. and Webb, E. L. (2005). Bayesian model averaging for disclosure risk assessment. Working paper. URL *http://www.maths.soton.ac.uk/staff/JJForster/paper.html.*

Franconi, L. and Polettini, S. (2004) Individual risk estimation in μ -Argus: a review. In Domingo-Ferrer, J. and Torra, V. (Eds.) *Privacy in Statistical Databases*, pages 262–272. Springer-Verlag, Berlin.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (Eds.) (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

O'Hagan, A. and Forster, J. J. (2004). Bayesian Inference, 2nd edition, volume 2B of *Kendall's Advanced Theory of Statistics*. Arnold, London.

Omori, Y. (1999). Measuring identification disclosure risk for categorical microdata by posterior population uniqueness. In *Proceedings of the International Conference on Statistical Data Protection SDP '98*, pages 59–76. Eurostat, Luxembourg.

Polettini, S. (2003). Some remarks on the individual risk methodology. In *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*. Luxembourg, April 2003.

Polettini, S. and Stander, J. (2004). A Bayesian hierarchical model approach to risk estimation in statistical disclosure limitation. In Domingo-Ferrer, J. and Torra, V. (Eds.) *Privacy in Statistical Databases*, page 247–261. Springer-Verlag, Berlin.

R Development Core Team (2005). **R:** A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rinott, Y. (2003). On models for statistical disclosure risk estimation. In *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*. Luxembourg, April 2003.

Skinner, C. J. and Holmes, D. J. (1998). Estimating the re-identification risk per record in microdata. *Journal of Official Statistics*, **14**, 361–372.

Takemura, A. (1999). Some superpopulation models for estimating the number of population uniques. In *Proceedings of the International Conference on Statistical Data Protection SDP '98*, pages 45–58. Eurostat, Luxembourg.

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

Assessing disclosure risk in microdata using record-level measures

Chris Skinner[*] *and Natalie Shlomo*[**]

[*] **Southampton Statistical Sciences Research Institute, University of Southampton**

[**] **Southampton Statistical Sciences Research Institute, University of Southampton,**
Department of Statistics, Hebrew University,
Office for National Statistics

Abstract: We consider the estimation of measures of disclosure risk using Poisson log-linear models. We focus on the question of how to specify the log-linear model. We develop some procedures related to the assessment of over-dispersion. We evaluate these procedures using simulated samples drawn from the 2001 United Kingdom Census and a real dataset being considered for release by the Office for National Statistics.We find that the procedures do indeed help to select models which provide good estimates of disclosure risk measures.

1. Introduction

Assessing disclosure risk for sample-based microdata is a growing challenge for National Statistical Institutes. Most decisions are based on ad-hoc rules, check lists and experience. There is a need to incorporate consistent and high-quality quantitative measures of disclosure risk in order to obtain more objective criteria for releasing microdata to different users. Since data on businesses are rarely released because of the high risk associated with skewed distributions, the focus in this paper is on microdata from social surveys.

Disclosure risk depends on microdata records that are both unique in the sample and in the population on a set of potentially identifying cross-classified key variables (i.e., a key). The key variables are determined according to disclosure risk scenarios. For example, if we are protecting against the risk scenario of matching the microdata to publicly available external files, we would want to choose key variables that are in common between the sources of data. We assume that the key variables are discrete and that there is no measurement error in the way these variables are recorded. Typical key variables include visible and traceable variables: sex, age, ethnicity, religion, place of residence, and occupation. In general, we will be analysing contingency tables spanned by the key variables. These tables contain the sample counts and are typically very large and very sparse.

We consider individual risk measures for each record in the microdata. By targeting only records with high risk, disclosure control techniques can be applied locally and the information loss to the file minimized. One advantage of the probabilistic method for assessing disclosure risk is that individual risk measures can be aggregated to obtain consistent overall global risk measures for the entire file which are useful to Microdata Release Panels in their decision making processes. The assessment and management of disclosure risk in microdata depends also on the means for disseminating the microdata and the level of protection that is needed. Microdata can be released to on-site data labs, licensed data archives, and public use. Thresholds are set below which the microdata can be released and above which more disclosure control techniques are necessary. The quantitative disclosure risk measures are therefore a necessary tool for ranking files according to their level of disclosure risk.

Let the key define K cells in the contingency table, labeled $k = 1,...,K$. Let the population count in cell k be F_k and the sample count be f_k. We consider the following two global risk measures:

1. Expected number of sample uniques that are population unique, $\tau_1 = E[\sum_k I(f_k = 1, F_k = 1)]$,
2. Expected number of correct matches for sample uniques to the population, $\tau_2 = E[\sum_k I(f_k = 1) / F_k]$.

These measures may be expressed as aggregates of record level measures: $\tau_1 = \sum_{SU} r_{1k}$, $\tau_2 = \sum_{SU} r_{2k}$, where $r_{1k} = P(F_k = 1 \mid f_k = 1)$, $r_{2k} = E[1/F_k \mid f_k = 1]$ with the sums over sample unique cells, denoted SU.

We suppose the f_k are observed but the F_k are unobserved. The measures are estimated using a Poisson model for the f_k, as developed by Bethlehem et al. (1990) and subsequently also used for disclosure risk assessment in the μ-Argus software (Hundepool, 2003; Benedetti et al., 1998; Polettini and Seri, 2003; Rinott, 2003). We shall assume a log-linear model for the underlying means of the Poisson distribution, following Skinner and Holmes (1998) and Elamir and Skinner (2005) and build on their approach by developing methods for the selection of the log-linear model and goodness-of- fit criteria.

In Section 2 we set out the model and its implications for disclosure risk assessment. Section 3 discusses possible criteria for choosing a model. Section 4 covers a model selection algorithm that is implemented taking into account the hierarchical structure of the log-linear models. Section 5 presents examples of how the probabilistic modelling can be implemented on both simulated samples drawn from the 2001 UK Census and a real dataset that is being considered for release by the Office for National Statistics (ONS). Finally, Section 6 contains a discussion and future research.

2. The Poisson Model

Following the earlier notation, a key is defined with cells $k = 1,\dots,K$. Let $N = \sum F_k$ and $n = \sum f_k$ be the population and sample sizes respectively. Based on natural assumptions for estimating rare populations we assume for each cell k: $F_k \mid \gamma_k \sim Pois(N\gamma_k)$ for $\gamma_k > 0$. A sample is drawn by Bernoulli sampling without replacement: $f_k \mid F_k \sim Bin(F_k, \pi_k)$, where the inclusion probability π_k may vary between cells. It follows that $f_k \mid \gamma_k \sim Pois(N\pi_k\gamma_k)$ since a thin Poisson variable is again Poisson. Based on these assumptions, we obtain: $F_k \mid f_k \sim Poisson(N\gamma_k(1-\pi_k)) + f_k$. Denoting $N\gamma_k = \lambda_k$, the record level measures may be expressed as:

$$r_{1k} = e^{-\lambda_k(1-\pi_k)},\ r_{2k} = E(\frac{1}{F_k} \mid f_k = 1) = \frac{1}{\lambda_k(1-\pi_k)}[1 - e^{-\lambda_k(1-\pi_k)}].$$

Elamir and Skinner (2005) propose using log-linear modelling to estimate the parameters λ_k. Assuming a simple random sampling design where $\pi_k = \pi = n/N$ for all cells k, the sample frequencies f_k are independent Poisson distributed with means $u_k = \pi\lambda_k$.

In the Poisson regression log-linear modelling framework, observed counts in a contingency table y_k are independent Poisson distributed given a design vector of regressors $\mathbf{x}_k$ denoting the main effects and interactions of the key variables. The means take the form $u_k = \exp(\mathbf{x}_k'\beta)$, where β is a parameter vector. The maximum likelihood estimator $\hat{\beta}$ is obtained by solving the score equations $\sum(y_k - u_k)\mathbf{x}_k = 0$ using iterative proportional fitting (IPF) or methods such as Newton-Raphson. Fitted values are obtained as $\hat{u}_k = \exp(\mathbf{x}_k'\hat{\beta})$. We calculate $\hat{\lambda}_k = \hat{u}_k/\pi$ for our estimates of the λ_k and plug these estimates into the formulae for the individual and global disclosure risk measures.

3. Criteria for Model Choice

We seek criteria for specifying the vector $\mathbf{x}_k$ in the log-linear model which lead to accurate estimated disclosure risk measures and are robust across different settings. One approach would be to use goodness-of-fit criteria such as Pearson or likelihood-ratio tests. The accuracy of the standard asymptotic approximations involved in the use of these procedures depends on the average cell size n/K being large enough, usually $n/K > 5$ although at least $n/K > 1$. Even this constraint does not hold for the large and sparse contingency tables that are typically used for assessing disclosure risk. For example, the quarterly UK Labour Force Survey individual microdata has 127,200 records in 10,540,000 cells defined by six identifying key variables, and the average cell size is 0.012. Some work on sparse tables (Koehler, 1986) suggests that the Pearson test is preferable to the likelihood ratio test in such circumstances. Nevertheless, our empirical work has suggested that neither of these criteria are very successful in predicting whether the disclosure risk measures will be well estimated and we shall not consider them further in this paper.

Instead, we consider an approach which is motivated more directly by our aim to estimate the disclosure risk measures accurately. Specifically, we seek a criterion for choosing a model which minimises the bias of $\hat{\tau}_1$ as an estimator of τ_1. Treating $\hat{\beta}$ as fixed, the bias may be expressed as:

$$E(\hat{\tau}_1 - \tau_1) = \sum_k \pi\lambda_k e^{-\lambda_k}\{e^{-(\hat{\lambda}_k - \lambda_k)(1-\pi)} - 1\}$$

$$\approx \sum_k \pi\lambda_k e^{-\lambda_k}\{-(\hat{\lambda}_k - \lambda_k)(1-\pi) + (\hat{\lambda}_k - \lambda_k)^2(1-\pi)^2/2\}$$

if the $\hat{\lambda}_k - \lambda_k$ are small. Under the Poisson assumption of equal mean and variance and ignoring estimation error in $\hat{\beta}$, $(y_k - \hat{u}_k)/\pi$ and $[(y_k - \hat{u}_k)^2 - y_k]/\pi^2$ will unbiasedly estimate $(\lambda_k - \hat{\lambda}_k)$ and $(\hat{\lambda}_k - \lambda_k)^2$ respectively. Hence, the bias of $\hat{\tau}_1$ may be expected to be reduced by minimising the absolute value of:

$$T_1 = \sum_k \frac{(1-\pi)}{\pi}\hat{\lambda}_k e^{-\hat{\lambda}_k}\{(y_k - \hat{u}_k)\pi + [(y_k - \hat{u}_k)^2 - y_k](1-\pi)/2\}$$

The statistic T_1 is a weighted mean of the $(y_k - \hat{u}_k)$ and the $[(y_k - \hat{u}_k)^2 - y_k]$. A similar expression applies for $\hat{\tau}_2$ but with different weights. The fitting of the log-linear model by IPF ensures that a weighted mean of the $(y_k - \hat{u}_k)$ is zero so the critical element of T_1 comes from the expression $[(y_k - \hat{u}_k)^2 - y_k]$ (in fact numerical work has shown that the $(y_k - \hat{u}_k)$ term usually only makes a very minor contribution to the value of T_1).

Choosing a model such that a weighted mean of the $[(y_k - \hat{u}_k)^2 - y_k]$ is close to zero may be interpreted as choosing a model which exhibits little under- or over-dispersion. This follows because y_k and $(y_k - \hat{u}_k)^2$ are unbiased estimators of the conditional mean and variance of y_k respectively, again ignoring differences between $\hat{\beta}$ and β and assuming $u_k = \exp(\mathbf{x}'_k\beta)$. Hence, an average of $[(y_k - \hat{u}_k)^2 - y_k]$ is a measure of over or under-dispersion. Cameron and Trivedi (1998, p.78) show that the hypothesis that $(y_k - u_k)^2$ and y_k share the same expectation may be tested by first defining $z_k = [(y_k - \hat{u}_k)^2 - y_k]/\hat{u}_k$ and then using OLS to estimate the regression model: $z_k = \kappa + \varepsilon_k$ where ε_k is an error term. Either the estimate $\hat{\kappa}$ or its associated t-statistic, T_κ, for testing $H_0 : \kappa = 0$ may be taken as a measure of over- or under-dispersion. The statistic, T_κ, is asymptotically normal and can be used to determine a class of models which do not exhibit significant departures from H_0.

An additional reason for avoiding under-dispersed models is that over-fitting may produce too many zeros on the margins leading to expected cell means being too high for the non-zero cells of the table and disclosure risk measures under- estimated. In contrast, under-fitting in over-dispersed models will produce no zeros on the margins and expected cell means may be too low for the non-zero cells of the table and disclosure risk measures over-estimated. Therefore, the model which manages the random and structural zeros of the contingency table will produce the best estimates for the disclosure risk measures.

4. Model Search Algorithm

We consider a model search algorithm that is similar to the TABU method introduced by Drezner, Marcoulides and Salhi (1999) for variable selection in multiple regression analysis. It is a local search method that depends on a criterion for the selection of the variables, a definition of a neighbourhood for each subset of the variables, a starting solution, a consistent method for moving through the neighbourhood and a stopping criterion. The neighbourhood is defined by adding a variable to the subset, removing a variable from the subset, and swapping variables.

As a starting solution, we begin with the all 2-way interactions log-linear model. This is motivated by the experience that it seems to lead to good estimates of the disclosure risk measures in many empirical experiments that we have undertaken. On the other hand, we have found that the independence log-linear model tends to be over-dispersed and leads to over-estimation of the disclosure risk measures. At the other extreme, the all 3-way interactions model tends to be under-dispersed and leads to under-estimation of the risk measures. Thus we expect a reasonable solution to lie between these extremes.

The algorithm is adapted to take into account the hierarchical structure of the log linear models. If we consider up to 2-way interactions for *k* independent variables, we obtain $k+\binom{k}{2}$ possible variables to examine. The variables, however, are highly dependent on each other. In the hierarchical log-linear modelling framework, for example, a model containing the interaction *{a*b}* means that the expected cell counts are fitted to the sample counts in the 2-way table defined by crossing variables *{a}* and *{b}*. Therefore, including into the model the separate variables *{a}* and *{b}* is redundant. Conversely, if we add into a model an independent term *{a},* we need to remove all interactions that involve that specific term *{a*b}*, *{a*c}*, etc. When defining the neighbourhood of a chosen model by dropping terms, swapping terms and adding terms, we need to make sure that we are not checking models that were previously examined or will be examined at a later stage of the search.

Starting with the all 2-way interactions model and considering only independent and all 2-way interaction terms, the first round of the algorithm involves dropping each interaction in turn and then swapping in independent terms and removing the relevant interactions involved with the specific term. Note that there are no terms to add in for the first round since these only produce redundant models. For each model in the neighbourhood the goodness-of-fit criteria will determine the most appropriate model for continuing the search and defining the next neighborhood.

5. Practical Implementation

In this Section we present some results on how the risk assessment can be carried out at a National Statistics Institute such as the ONS. We first demonstrate on samples drawn from the 2001 UK Census where we compare the estimated disclosure risk measures with the true disclosure risk measures and check the performance of the model choice criteria. The second example will present an analysis on a real data set that is being considered for release to the UK data archive.

Example 1: Simulated samples from population census

Table 1 presents true and estimated global risk measures for simple random samples of different sizes drawn from two Estimation Areas of the 2001 UK Census (N=944,793). We demonstrate on two keys defined by cross-classifying six traceable and visible key variables. The first key has 412,080 cells (the number of categories is in parenthesis): Estimation Area (2), Sex (2), Age (101), Marital Status (6), Ethnicity (17), Economic Activity (10). The second key has 73,440 cells and is defined as the first key except that age was banded into 18 groupings. We ran two log-linear models: the independence model and the all 2-way interactions model.

Table 1. Global Risk Measures on Samples Drawn from the 2001 UK Census

Sample Size	Model	True		Estimates		Cameron-Trivedi Test		
		τ_1	τ_2	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\kappa}$	T_κ	T_1
Small Key								
4,724	Indep	23	68.2	54.2	126.9	0.5074	8.55	562.24
4,724	2-way			16.0	52.2	-0.0041	-3.62	-11.695
9,448	Indep	39	127.1	99.3	230.2	1.0316	8.58	1,447.20
9,448	2-way			37.8	117.9	-0.0051	-3.91	-30.952
18,896	Indep	75	215.3	174.3	355.7	2.0622	9.56	3,153.22
18,896	2-way			85.5	222.0	0.0059	2.00	16.891
Large Key								
4,724	Indep	80	183.9	197.4	385.1	0.0881	10.58	1,178.91
4,724	2-way			35.9	112.3	-0.0025	-7.96	-16.822
9,448	Indep	159	355.9	386.6	701.2	0.1846	14.42	3,400.76
9,448	2-way			104.9	280.1	-0.0036	-10.32	-59.257
18,896	Indep	263	628.9	672.0	1170.5	0.3865	16.77	7,269.90
18,896	2-way			252.0	591.3	-0.0030	-5.69	-43.594

In Table 1, the 2-way interactions model always leads to better estimates than the independence model and this is predicted in all cases by values of $\hat{\kappa}$, T_κ and T_1 being closer to 0. For the smaller key, the values of T_κ for the 2-way interactions model are close to the critical values for accepting the null hypothesis of equal dispersion. For the larger key, the values of T_κ suggest that the 2-way interactions model is over-fitting the data. We continue in our model search based on the large key and the 1% sample (n=9,448). Table 2 presents results of the first round of the neighbourhood search.

Table 2. Round 1 of the Neighbourhood Search for n=9,448 and K=412,080

Model	Estimates		Cameron-Trivedi Test		T_1
$\tau_1 = 159 \quad \tau_2 = 355.9$	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\kappa}$	T_κ	
Independent	386.6	701.2	0.1846	14.42	3,400.76
All 2-way	104.9	280.1	-0.0036	-10.32	-59.257
Drop {ea*s}	104.6	279.8	-0.0035	-10.15	-58.969
Drop {ea*a}	105.3	281.3	-0.0032	-9.69	-61.684
Drop {ea*m}	103.8	279.1	-0.0034	-10.92	-63.851
Drop {ea*et}	108.7	290.0	-0.0024	-6.09	-58.230
Drop {ea*ec}	105.2	280.0	-0.0035	-10.60	-60.399
Drop {s*a}	104.5	280.7	-0.0033	-9.87	-60.699
Drop {s*m}	105.5	281.8	-0.0032	-8.53	-57.649
Drop {s*et}	105.2	280.3	-0.0035	-10.26	-58.949
Drop {s*ec}	103.2	281.5	-0.0018	-5.18	-64.670
Model	Estimates		Cameron-Trivedi Test		T_1
$\tau_1 = 159 \quad \tau_2 = 355.9$	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\kappa}$	T_κ	
Drop {a*m}	134.0	328.6	0.0071	9.42	-39.178
Drop {a*et}	147.0	346.2	0.0018	1.52	-38.477
Drop {a*ec}	184.7	419.2	0.0316	13.27	543.90
Drop {m*et}	108.7	287.5	-0.0032	-8.56	-59.692
Drop {m*ec}	108.3	284.0	-0.0028	-6.74	-51.510
Drop {et*ec}	132.3	308.2	-0.0015	-2.24	-20.147
In {ea} Out {ea*s}{ea*a}{ea*m}{ea*et}{ea*ec}	109.5	290.6	-0.0020	-5.72	-64.293
In {s} Out {ea*s}{s*a} {s*m}{s*et}{s*ec}	105.0	284.2	-0.0011	-3.13	-64.734
In {a} Out {ea*a}{s*a} {a*m}{a*et}{a*ec}	285.1	576.3	0.0803	18.43	487.31
In {m} Out {ea*m}{s*m}{a*m}{m*et}{m*ec}	134.3	355.5	0.0181	14.05	-62.752
In {et} Out {ea*et}{s*et}{a*et}{m*et}{et*ec}	190.7	396.5	0.0188	3.25	1,155.74
In {ec} Out {ea*ec}{s*ec}{a*ec}{m*ec}{et*ec}	207.7	464.0	0.0457	17.68	117.29

(Note: Estimation Area–ea, Sex-s, Age–a, Marital Status–m, Ethnicity–et, and Economic Activity-ec)

From Table 2, removing the {et*ec} interaction provides the minimum value of T_1 and is also defining a model that accepts the null hypothesis of equal dispersion with a small estimate for parameter κ. Therefore, we chose this model to continue to the second round of the model search. As mentioned, some models need not be checked in subsequent rounds because of the hierarchical structure of the log linear models. For example, deleting the last interaction {et*ec} means that there is no need to evaluate adding in {et} or {ec} and taking out their relevant interactions since this leads to the same models that were previously checked in round one.

Table 3. Round 2 of a Neighbourhood Search for n=9,448 and K=412,080

Model	Estimates		Cameron-Trivedi Test		
$\tau_1 = 159 \quad \tau_2 = 355.9$	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\kappa}$	T_κ	T_1
Drop {et*ec}	132.3	308.2	-0.0015	-2.24	-20.147
Drop {ea*s} {et*ec}	132.3	308.2	-0.0015	-2.27	-20.594
Drop {ea*a} {et*ec}	133.4	310.4	-0.0011	-1.65	-14.781
Drop {ea*m} {et*ec}	131.9	307.9	-0.0014	-2.28	-28.398
Drop {ea*et} {et*ec}	139.8	320.8	-0.0002	-0.20	-2.909
Drop {ea*ec} {et*ec}	133.7	309.5	-0.0015	-2.33	-22.478
Drop {s*a} {et*ec}	132.1	309.2	-0.0013	-2.17	-32.570
Drop {s*m} {et*ec}	133.4	310.3	-0.0011	-1.58	-14.389
Drop {s*et} {et*ec}	132.4	308.5	-0.0015	-2.24	-21.111
Drop {s*ec} {et*ec}	130.9	310.3	0.0002	0.35	-38.516
Drop {a*m} {et*ec}	159.7	354.2	0.0091	10.27	-29.537
Drop {a*et} {et*ec}	173.4	370.2	0.0066	2.58	161.58
Drop {a*ec} {et*ec}	208.4	442.5	0.0324	13.38	573.72
Drop {m*et} {et*ec}	137.3	315.8	-0.0011	-1.68	-18.588
Drop {m*ec} {et*ec}	134.0	311.1	-0.0008	-1.10	-12.185
In {ea} Out {et*ec} {ea*s} {ea*a} {ea*m} {ea*et} {ea*ec}	141.3	321.7	0.0002	0.28	0.3363
In {s} Out {et*ec} {ea*s} {s*a} {s*m} {s*et} {s*ec}	132.6	313.0	0.0009	1.36	-37.947
In {a} Out {et*ec} {ea*a} {s*a} {a*m} {a*et} {a*ec}	313.4	596.5	0.0830	19.03	656.94
In {m} Out {et*ec} {ea*m} {s*m} {a*m} {m*et} {m*ec}	166.0	386.5	0.0221	12.64	66.937

(Note: Estimation Area–ea, Sex-s, Age–a, Marital Status–m, Ethnicity–et, and Economic Activity-ea)

In Table 3, many models accept the null hypothesis of equal dispersion. The model {In {ea} Out {et*ec} {ea*s} {ea*a} {ea*m} {ea*et} {ea*ec}} has the minimum value of T_1 and also accepts the null hypothesis with a small value T_κ. This model is our preferred model. We note that the models that accept the null hypothesis (with $|T_\kappa| \leq 2.4$) are giving good estimated global risk measures compared to the true measures and therefore the global risk measures seem robust to slight deviations in the model.

Per-record risk measures are very important as a means of isolating high-risk sample uniques or groupings of sample uniques (i.e., particular age groups, etc.) for targeting disclosure control methods. In Figure 1 we examine the marginal distribution of the true and estimated per-record risk measures $\hat{r}_{2k}$ for the sample uniques within bands under the preferred model from Table 3. Table 4 presents their joint distribution.

Figure 1. Per-Record Risk Measures for Preferred Model

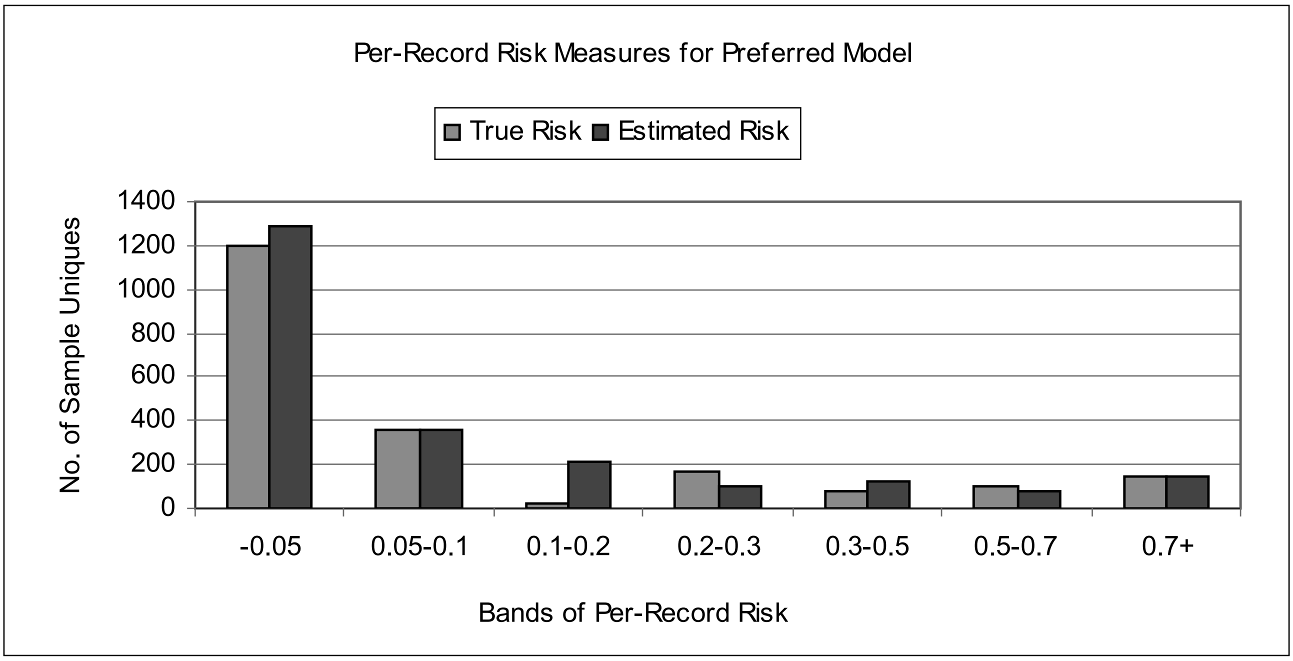


Table 4. Joint Distribution of Per-Record Risk Measures for Preferred Model: Cramer's V=0.4347

True Per-Record Risk Measures	Estimated Per-Record Risk Measures			
	0 – 0.3	**0.3 – 0.7**	**0.7 – 1**	**Total**
0 – 0.3	1,838	97	26	1,961
0.3 – 0.7	75	57	52	184
0.7 – 1	45	49	65	159
Total	1,958	203	143	2,304

We obtain a good fit between the marginal distributions of the true and estimated per-record risk measures, although for the true high risk sample uniques, only 41% obtain a high estimated risk measure.

Example 2: large UK social survey

In this example, we look at an ONS dataset considered for release to the data archives. The sample size is n=530,013 with a sampling fraction of 0.9%. The microdata underwent disclosure control methods based on recoding key variables and eliminating other identifiable variables. We examined several combinations of key variables:

1) The key variables are: Region (20), Sex(2), Age Bands (45), Marital Status (6), Ethnicity (16) and Economic Activity (23). This resulted in a key of K=3,974,400 out of which 13,954 were sample uniques. The results for the all 2-way interactions log-linear model were the following: The estimated number of population uniques that are sample uniques is $\hat{\tau}_1 = 440.6$ which is 3.2% of the sample uniques. The expected number of correct matches is $\hat{\tau}_2 = 1{,}289.5$ which is 9.2% of the sample uniques or 0.2% of the entire sample. The values of the model choice criteria were $T_1 = 310.6$, $\hat{\kappa} = 0.00423$ and $T_\kappa = 7.03$. The slightly high values of T_1 and T_κ leads us to expect some over-estimation of the disclosure risk measures.

2) The key is the same as key 1 except for replacing age bands with single years of age (100). This resulted in a key of K=8,832,000 out of which 39,588 sample uniques. This new key increased the disclosure risk. Based on the all 2-way interactions log-linear model, the esti-

mated number of population uniques that are sample uniques is $\hat{\tau}_1 = 1,985.4$ which is 5.0% of the sample uniques. The expected number of correct matches is $\hat{\tau}_2 = 4,779.9$ which is 12.1% of the sample uniques or 0.9% of the entire sample. The model choice criteria were $T_1 = 568.3$, $\hat{\kappa} = 0.00162$ and $T_\kappa = 5.61$. These are also slightly high values and we expect some over-estimation of the disclosure risk measures.

3) The key is the same as key 1 except for replacing Economic Activity with Occupation (82). This resulted in a key of K=14,169,600 out of which 28,656 sample uniques. Because the key is so large it was necessary to partition the contingency table and carry out the disclosure risk assessment separately on each sub-table. Based on empirical work, it was found that disclosure risk assessment performs best when partitioning the contingency table according to a key variable that is correlated with the other key variables, since the partitioning key variable has an underlying interaction with the other variables. We partitioned the table into two sub-tables according to sex, and within each sub-table carried out an independence log-linear model. After combining the results from the two separate log-linear models, we obtained the following results: the estimated number of population uniques that are sample uniques is $\hat{\tau}_1 = 1,190.1$ which is 4.2% of the sample uniques. The expected number of correct matches is $\hat{\tau}_2 = 3,082.2$ which is 10.8% of the sample uniques or 0.6% of the entire sample. The model choice criteria were $T_1 = 337.0$, $\hat{\kappa} = 0.00021$ and $T_\kappa = 2.43$. These model choice criteria are slightly better than previously obtained based on the other keys.

It is clear that more iterations are needed to determine the recoding of the variables on the final microdata to be released which would manage the disclosure risk while maximising the utility of the data. Also, a model search needs to be carried out in order to obtain a model that indicates acceptance of the null hypothesis of equal dispersion (i.e., the fit of the Poisson Model) and therefore more accurate estimated disclosure risk measures.

6. Discussion

In this paper we have examined the estimation of global and individual disclosure risk measures based on a Poisson log-linear model as developed by Skinner and Holmes (1998) and Elamir and Skinner (2004). We have addressed the implementation of model selection criteria for the large and sparse contingency tables spanned by key variables that are typical in the assessment of disclosure risk in microdata. Empirical results show that the goodness-of-fit criteria do select models that give good estimates for the disclosure risk measures based on the simple random samples that were drawn from the Census. There is a need for further empirical work to assess the impact of the size of the key on the goodness-of-fit criteria and model choice. In addition, since keys can be very large in practice, we need to develop optimal methods for splitting contingency tables since this has implications for the types of models that can be assessed.

Future work on the Poisson model for disclosure risk assessment will focus on applications for hierarchical datasets and more complex survey designs, and in particular stratified samples with varying probabilities. In addition, variance and confidence intervals need to be developed for the estimated risk measures.

References

Benedetti, R., Capobianchi, A, and Franconi, L. (1998) Individual Risk of Disclosure Using Sampling Design Information.

Bethlehem, J., Keller, W., and Pannekoek, J. (1990) Disclosure Control of Microdata, JASA, Vol. 85.

Elamir, E. and Skinner, C. (2004) Record-level Measures of Disclosure Risk for Survey Microdata, Technical Paper, Southampton Statistical Sciences Research Institute, University of Southampton.

Cameron, A. C. and Trivedi. P.K. (1998), *Regression Analysis of Count Data*, Cambridge University Press, Cambridge.

Drezner, T., Marcoulides, G. and Salhi, S. (1999), TABU Search Model Selection in Multiple Regression Models, Communications in Statistics 28(2).

Hundepool, A., et. al. (2003) Mu-Argus Version 3.1 User's Manual, http://neon.vb.cbs.nl/casc/

Koehler, K.J. (1986) Goodness-of-Fit Tests for Log-Linear Models in Sparse Contingency Tables, Journal of the American Statistical Association, Vol. 81, No. 394 pp. 483-493.

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, http://neon.vb.cbs.nl/casc/

Rinott, Y. (2003) On Models for Statistical Disclosure Risk Estimation, Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Luxemburg, April 7-9, www.unece.org/stats/documents/2003/04/confidentiality/wp.16.e.pdf

Skinner, C. and Holmes, D. (1998), Estimating the Re-identification Risk Per Record in Microdata, JOS, Vol.14, 1998.

A neighborhood regression model for sample disclosure risk estimation

Yosef Rinott, Natalie Shlomo***

*** Department of Statistics, Hebrew University, Jerusalem, Israel. (rinott@mscc.huji.ac.il)**

**** Department of Statistics, Hebrew University of Jerusalem, Southampton Statistical Sciences Research Institute, University of Southampton, UK. (N.Shlomo@soton.ac.uk)**

Abstract. The disclosure risk involved in releasing data which consist of a sample from some population depends on both the sample and the population. When the sample is fully known, with only partial or no information on the population, a major problem in *Statistical Disclosure Control* (SDC) is the estimation of *disclosure risk* on the basis of the sample. Considering data in the form of a frequency table, risk arises from non-empty sample cells which represent small population cells (and population uniques in particular). Therefore risk estimation requires assessing which of the relevant population cells are indeed small.

Various methods have been proposed for this task, and we present a new one, in which estimation of population cell frequencies is based on a model connecting the table parameters in neighborhoods defined in natural ways using the table structure and the nature of the variables. At this point this method is under experimentation, and we provide some preliminary comparisons with the *Argus method* in which inference is based on sampling weights, and with a *log-linear models* approach.

1. Introduction

Let $\mathbf{f} = \{f_k\}$ denote an m-way sample frequency table, where $k = (k_1, ..., k_m)$ indicates a cell and f_k is the frequency in the cell, and let $\mathbf{F} = \{F_k\}$ denote the population from which the sample is drawn. We denote the sample and population sizes by n and N respectively, and the number of cells by K. Disclosure risk arises from cells in which both f_k and F_k are positive and small, and in particular when $f_k = F_k = 1$ (a sample and population unique).

Various individual and global risk measures have been proposed in the literature, see e.g., Benedetti, Capobianchi and Franconi (1998), Skinner and Holmes (1998), Elamir and Skinner (2006), Rinott (2003). In this paper we chose to focus only on two global risk measures,

$$\tau_1 = \sum_k \mathbb{I}(f_k = 1, F_k = 1), \qquad \tau_2 = \sum_k \mathbb{I}(f_k = 1)\frac{1}{F_k}$$

where $\mathbb{I}$ denotes the indicator function. Note that τ_1 counts the number of *sample uniques* which are also *population uniques*, and τ_2 is the expected number of correct guesses if each sample unique is matched to a randomly chosen individual from the same population cell. These measures are somewhat arbitrary, and one could consider measures which reflect matching of individuals that are not sample uniques, possibly with some restrictions on cell sizes. Also, it may make sense to normalize these measures by some measure of the total size of the table, by the number of sample uniques, or by some measure of the information value of the data. Such and other measures should also be considered.

When only $\mathbf{f}$ is known, and $\mathbf{F}$ is considered an unknown parameter (on which there is often some partial information) the quantities τ_1 and τ_2 should be estimated. Note that they are not proper parameters, since they involve both the sample $\mathbf{f}$ and the parameter $\mathbf{F}$. Therefore a discussion of the variances of estimates of τ_1 and τ_2 requires special care, see Rinott (2003) for some details, and Zhang (2005) for general theory. We shall discuss this issue in a subsequent paper.

In this paper we describe two known methods of estimation of quantities like τ_1 and τ_2, propose a new one, and compare them by some experiments. The first by Benedetti, Capobianchi and Franconi

(1998) which uses the *Negative Binomial* model, provides the basis to the μ-*Argus* program, and the second, proposed by Skinner and Holmes (1998) and Elamir and Skinner (2006), uses a *Poisson* model and bases estimation on hierarchical *log-linear models*. The new method we propose is based on a different model which we shall explain. We shall present here the main ideas of this method, which is under development, and preliminary experiments .

All the above methods consist of modeling the conditional distribution of $\mathbf{F}\,|\,\mathbf{f}$, estimating parameters in this distribution and then using estimates of the form

$$\hat{\tau}_1 = \sum_k \mathbb{I}(f_k = 1)\hat{P}(F_k = 1\,|\,f_k = 1), \qquad \hat{\tau}_2 = \sum_k \mathbb{I}(f_k = 1)\hat{E}[\frac{1}{F_k}\,|\,f_k = 1] \tag{1}$$

where $\hat{P}$ and $\hat{E}$ denote estimates of the relevant conditional probability and expectation. For a general theory of estimates of this type see Zhang (2005) and reference therein.

2. Models

For completeness we briefly introduce the Poisson and Negative Binomial models. More details can be found, for example, in Bethlehem et al (1990), Cameron and Trivedi (1998), Rinott (2003).

A common assumption in the frequency table literature is $F_k \sim \text{Poisson}(N\gamma_k)$, independently, with $\sum\gamma_k = 1$. Binomial (or Poisson) sampling from F_k means that $f_k\,|\,F_k \sim Bin(F_k, \pi_k)$, π_k being the sampling fraction in cell k. By standard calculations we then have

$$f_k \sim \text{Poisson}(N\gamma_k\pi_k) \text{ and } F_k\,|\,f_k \sim f_k + \text{Poisson}(N\gamma_k(1-\pi_k)), \tag{2}$$

leading to the Poisson model of subsection 1 below.

If one adds the Bayesian assumption $\gamma_k \sim \text{Gamma}(\alpha, \beta)$ indipendently, with $\alpha\beta = 1/K$ to ensure that $E\sum\gamma_k = 1$, then $f_k \sim NB(\alpha, p_k = \frac{1}{1+N\pi_k\beta})$, the Negative Binomial distribution defined for any $\alpha > 0$ by $\text{P}(\text{f}_k = x) = \frac{\Gamma(x+\alpha)}{\Gamma(x)\Gamma(\alpha)}(1-p_k)^x p_k^\alpha$, x = 0,1,2,..., which for a natural α counts the number of *failures* until α successes occur in independent Bernoulli trials with probability of success p_k. Further calculations yield $F_k\,|\,f_k \sim f_k + NB(\alpha + f_k, \frac{N\pi_k + 1/\beta}{N + 1/\beta})$, $(\text{F}_k \geq f_k)$.

As $\alpha \to 0$ (and hence $\beta \to \infty$) we obtain $F_k\,|\,f_k \sim f_k + NB(f_k, \pi_k)$, which is exactly the Negative Binomial assumption in Section 2 below. As $\alpha \to \infty$ the Poisson model of Section 1 is obtained, and in this sense the Negative Binomial with $\alpha \neq 0$ subsumes both models. Applications of this generalization will be given in a subsequent paper.

2.1. The Poisson log-linear method

Skinner and Holmes (1998) and Elamir and Skinner (2006) proposed and studied the following approach. Assuming a fixed sampling fraction, that is, $\pi_k = \pi$, the first part of (2) implies $f_k \sim \text{Poisson}(n\gamma_k)$, where $n = N\pi$. Using the sample $\{f_k\}$ one can fit a log-linear model using standard programs, and obtain estimates $\{\hat{\gamma}_k\}$ of the parameters. Using the second part of (2) it is easy to compute

$$P(F_k = 1\,|\,f_k = 1) = e^{-N\gamma_k(1-\pi_k)}, \; E[\frac{1}{F_k}\,|\,f_k = 1] = \frac{1}{N\gamma_k(1-\pi_k)}[1 - e^{-N\gamma_k(1-\pi_k)}]. \tag{3}$$

Plugging $\hat{\gamma}_k$ for γ_k in (3) leads to the desired estimates $\hat{\tau}_1$ and $\hat{\tau}_2$ of (1). The quantity $E[\frac{1}{F_k}\,|\,f_k = 1]$ is sometimes referred to as the *individual risk measure* at cell k.

2.2. The Negative Binomial Argus method

In this method, proposed by Benedetti, Capobianchi and Franconi (1998), see also Polettini and Seri (2003), it is assumed that $F_k \mid f_k \sim f_k + NB(f_k, \pi_k)$. There is an implicit assumption of independence between cells.

Using the relation $E_{\pi_k}[F_k \mid f_k] = f_k / \pi_k$, the parameters π_k are estimated using sampling weights: if w_i denotes the sampling weight of individual i, then an *initial estimate* of F_k is $\hat{F}_k = \sum_{i \in \text{cell } k} w_i$, and we obtain the moment-type estimate $\hat{\pi}_k = f_k / \hat{F}_k$ Straightforward calculations with the Negative Binomial distributions show

$$P_{\hat{\pi}_k}(F_k = 1 \mid f_k = 1) = \hat{\pi}_k, \quad \text{and} \quad E_{\hat{\pi}_k}[\tfrac{1}{F_k} \mid f_k = 1] = -\tfrac{\hat{\pi}_k}{1-\hat{\pi}_k}\log(\hat{\pi}_k).$$

Plugging these estimates for $\hat{P}$ and $\hat{E}$ in (1) we obtain the estimates $\hat{\tau}_1$ and $\hat{\tau}_2$ of the global risk measures. Note that in this method the cells are treated completely independently, each cell at a time, and the structure of the table plays no role.

2.3. A brief discussion

Estimation of risk measures without a model which restricts the number of parameters, such as a log-linear model, is inherently difficult. To see this just note that if one estimates γ_k in each cell separately without a model by $\hat{\gamma}_k = f_k / n$ then the estimated population cell frequency $N\hat{\gamma}_k$ satisfies Var $N\hat{\gamma}_k \approx N^2 \gamma_k / n$. Typically, risk arises from cells where $\gamma_k = O(1/N)$ since such cells are likely to contain population uniques, and for such k we obtain SD $(N\hat{\gamma}_k) = O((N/n)^{1/2})$ which is usually large.

The situation improves in the presence of a model that reduces the number of parameters, provided of course that the model is valid. In order to see this in a specific example, consider a two-way table and the (log-linear) model of independence. For the Maximum Likelihood estimate $\hat{\gamma}_k$, it can then be shown directly that the variances of the cell frequency estimate $N\hat{\gamma}_k$, or that of $\hat{P}(F_k = 1 \mid f_k = 1) = e^{-N\hat{\gamma}_k(1-\pi_k)}$ which appears in the estimate $\hat{\tau}_1$ is $O(\frac{N^2}{n^2}\gamma_k + \frac{N^2}{n}\gamma_k^{1+\nu})$ for some $\nu \leq 1/2$ which depends on the parameters. Again looking at cells where $\gamma_k = O(1/N)$, and $\nu = 1/2$, the standard deviation SD $(N\hat{\gamma}_k)$ of the cell frequency estimate is like $O(N^{1/2}/n)$, a great improvement, and often small enough. The situation improves further with large higher order tables if simple models, like independence, are valid (see Zhang 2005). The above also shows that dividing the population into smaller parts will increase the variance, and should be done only if it leads to better models.

The estimation question here is essentially the following: given, say, a sample unique, how likely is it to be also a population unique, or arise from a small population cell. The *Argus* method bases its estimation on sampling weights (and the NB model). There is no learning from other cells. However, such learning appears natural. If a sample unique is found in a part of the table where neighboring cells (by some reasonable metric, to be discussed later) are small or empty, then it seems reasonable to believe that it is more likely to have arisen from a small population cell.

As we saw above, when a *log-linear model* is indeed valid, it will reduce the standard deviation of population frequency estimates and hence of risk measures. The *log-linear model* approach indeed uses cells from neighborhoods which depend on the model to determine the risk in a given cell. For example, if the attributes forming the table are assumed independent, then the estimate $\hat{\gamma}_k$ is the product of marginals obtained by fixing one attribute at a time, so that every cell which has a common value with one of the attributes of cell k will contribute to the risk estimate at this cell; thus if one of the attributes is economic status, then inference on the very rich involves also information from the very poor, provided they have some other attribute in common, such as marital status.

This observation led us to trying another type of *neighborhoods*, thinking that log-linear models, which provide explanations to the data when they fit, may not lead to the most natural neighborhoods for the question at hand. Our initial attempts will be described in Section 4 and some experiments are described in Section 3.

Another inherent problem that arises in the *Argus* method is related to the fact that for empty sample cells the *initial estimates* $\hat{F}_k$ vanish. Since the total population estimate should be N, it follows that other population cells tend to be overestimated, and as a result, risk measures are underestimated. A systematic treatment of this hard problem would require identifying *structural zeros* and perhaps replacing other sample zeros by some ε, as sometimes proposed in the literature in the context of model building. It is easy to see that if this is introduced into the *Argus* method, risk measures will increase to the correct values as ε increases, and then will exceed them. However, estimation of the right ε, or similar parameters, appears difficult.

A version of the latter issue appears also in the *log-linear model* approach. If a saturated model is used then $\hat{\gamma}_k = f_k/n$, and for empty sample cells $\hat{\gamma}_k = 0$, leading to the same underestimation problem as above. In fact, for $f_k = 1$ we have $N\hat{\gamma}_k = N/n$ and from (3) we obtain $\hat{P}(F_k = 1 \mid f_k = 1) \approx e^{-N/n}$ and $\hat{E}[\frac{1}{F_k} \mid f_k = 1] \approx n/N$, so that all sample uniques are estimated to have the same very low risk. At the other extreme, if we take a model of independence then $\hat{\gamma}_k$ is obtained as a product of terms where each term is a large sum of frequencies over all attributes except for one, that is, for k=(k$_1,...,k_m$) , $\hat{\gamma}_k = \prod_i (\sum_{k_j, j \neq i} f_k/n)$. The large sum for a given i in the latter product vanishes only if the level k'_i of the attribute i never appears in the sample, and in that case it would probably be omitted from the file. Thus, the model usually has no zero population cell predictions, and in view of the above one should expect higher risk estimates. In fact the independence model often leads to overestimation of risk as expected by this explanation. Intermediate models, such as those of conditional independence involve products of smaller sums, and in general one may expect monotonicity of the risk estimates in the size of the model (number of parameters). So again, as in the choice of ε above, there is usually a model which would give a good risk estimate for a given risk measure. The question of finding goodness of fit measures so that the model chosen provides good risk estimates is studied in Skinner and Shlomo (2005).

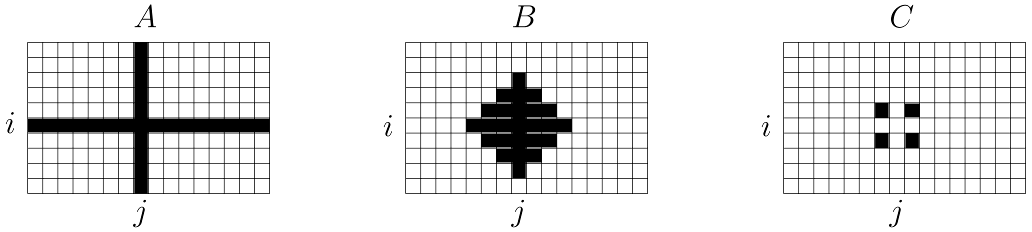
2.4. Neighborhoods

We consider frequency tables in which some of the attributes are ordinal. For such an attribute i we can consider a set of levels S_{k_i} which are close to a given level k_i in the attribute's ordering. Given cell $k = (k_1, ..., k_m)$ we can construct a neighborhood of cells N_k of k, by varying the coordinates k_i of the ordinal attributes in some way in the sets S_{k_i}, and fixing the other, non-ordinal attributes.

More specifically, let O denote the set of ordinal attributes and suppose the attribute $i \in O$ has levels $1, 2, ..., r_i$.

Here we consider neighborhoods of the type $N_c^k = \{h = (h_1, ..., h_m) : \sum_{j \in O} | h_j - k_j | = c, h_j = k_j \text{ for } j \notin O\}$ or the type $M_a^k = \{h = (h_1, ..., h_m) : | h_j - k_j | = a_j \text{ for } j \in O, h_j = k_j \text{ for } j \notin O\}$ for some $a = (a_1, ..., a_m)$.

Figure 1. *Neighborhood of cell* $k=(i,j)$. *A: under independence model. B: the union of neighborhoods* $\bigcup_{c\le 3} N_c^k = \bigcup_{|a|\le 3} M_a^k$. *C: the neighborhood* $M_{(1,1)}^k$



This approach can perhaps be extended to non-ordinal attributes having some metric or a measure of proximity between their levels, such as geographic location.

The neighborhoods are used as follows: we assume as in (2), with $\pi_k = \pi$ for simplicity, $f_k \sim \text{Poisson}(N\gamma_k\pi)$ and $F_k \,|\, f_k \sim f_k + \text{Poisson}(N\gamma_k(1-\pi))$, but now we propose to consider log-linear models of the form $\gamma_k = exp\{\beta_0 + \sum_{c\le C} \beta_c x_c^k\}$ for some C to be determined, where $x_c^k = \sum_{\ell \in N_c^k} f_\ell$. We can estimate the parameter vector β, obtain estimates $\hat{\gamma}_k$ and proceed to estimate risk as before using these estimates in the above Poisson conditional distribution of $F_k \,|\, f_k$. In a similar way, setting $|\,\text{a}| = \sum_{j\in O} |\, a_j \,|$, we tried model with $\gamma_k = exp\{\beta_0 + \sum_{a:|a|\le C} \beta_a z_a^k\}$ for some C to be determined, where $\text{z}_a^k = \sum_{\ell \in M_a^k} f_\ell$.

Other regression models (e.g., Negative Binomial, see Cameron and Trivedi (1998) for Poisson and Negative Binomial regression) and types of neighborhoods, and combinations of the neighborhood approach with weights and other information on the population will be discussed in a subsequent paper.

Regarding the issue of *structural zeros*, we tried declaring a cell to be a structural zero if all its neighborhoods which are used in the regression contain only empty cells.

Some technical issues: The cardinality of N_c^k satisfies $|\text{N}_c^k| = 2^m \sum_{t=\min(m-c,0)}^{m-1} 2^{-t} \binom{m}{t}\binom{c-1}{m-t-1}$ which increases rapidly with m and c (it is smaller for k 's near the boundary of the table, but still many of these neighborhoods are rather large). For $m=4$ (a four-way table) we have for k 's not near the boundary $|\,N_5^k\,| = 360$ and $|\,N_7^k\,| = 856$. On the other hand, the neighborhoods M_a^k are not as large, however, the number of neighborhoods of the type M_a^k with $|\,a\,| = c$ is $\binom{c+m-1}{c}$, so that for $m=4$ and $c=7$, for example, we would have to deal with 120 such neighborhoods and β coefficients in the regression. Therefore our preliminary experiments presented in the Section 3 are quite restricted in size and perhaps not very impressive at this point. There is much room for improving and fine-tuning the method and the programs, and for testing different types of data before conclusions can be drawn.

3. Experiments with neighborhoods

We present a few experiments. They are preliminary as already mentioned and more work is needed on the approach itself and on classifying types of data for which it might work.

In the experiments we used our versions of the Argus and log-linear models approaches, programmed on the SAS system. In all experiments we took a real population data file of size N given in the form of a contingency table with K cells, and from it we took a random sample of size n. Since the population and the sample are known to us, we can compute the *true values* of τ_1 and τ_2 and their estimates by the different methods, and compare.

Example 1 In this example the population consists of an extract of the 1995 Israeli Census Sample File for Individuals with age 15 and over with $N = 746{,}949, n = 14{,}939$, and $K = 337{,}920$. The attributes (with number of levels in parentheses) were Sex (2), Age Groups (16), Groups of Years of Study (10), Number of Years in Israel (11), Income Groups (12), and Number of Persons in Household (8). Since Sex is not ordinal, neighborhoods were constructed with Sex being fixed and the set of ordinal attributes O contains the other five variables. We used neighborhoods of the type N_c^k for $c \le C = 4$, and M_a^k for $|a| \le C = 4$.

In one version of the experiment we ignored the issue of structural zeros, and in another we define structural zeros as all sample cells that have a zero count and the sum of the sample counts in all of the neighborhoods is zero. Out of $K = 337,920$ cells, we obtained 206,655 non-structural zeros. The Poisson regression model with the new types of neighborhoods was run on this file with and without the structural zeros to obtain the expected cell means and risk measure estimates as described above. The weights w_i for the *Argus* method in all our examples were computed by post-stratification on Sex by Age by Geographical location (the latter is not one of the attributes in any of the tables, but it was used for post-stratification). These variables are commonly used for post-stratification, other strata may give different results. Two log-linear models are considered, one of independence, the other including all two-way interactions.

Model	τ_1	τ_2
True Values	430	1125.8
Argus	114.5	456
Log Linear Model: Independence	773.8	1774.1
Log Linear Model: 2-Way Interactions	470	1178.1
Neighborhood method M_a^k	786.8	2146.9
Neighborhood method M_a^k excluding structural zeros	385.4	1674.1
Neighborhood method N_c^k	723.3	2099.6
Neighborhood method N_c^k excluding structural zeros	344.8	1624.2

Example 2. This data consist of an extract of the 2001 UK Census file N = 944,793, *n* = 18,896, *K* = 152,100, with the attributes, Sex (2) Age Groups (25) Number of Persons in Household (9) Education Qualifications (13) Occupation (26). Sex was treated as non-ordinal as above.

Model	τ_1	τ_2
True Values	191	568.0
Argus	79.2	315.6
Log Linear Model: Independence	364.8	862.3
Log Linear Model: 2-Way Interactions	182.3	546.2
Neighborhood method M_a^k	42.8	770.0
Neighborhood method M_a^k excluding structural zeros	6.4	540.2
Neighborhood method N_c^k	38.5	755.2
Neighborhood method N_c^k excluding structural zeros	5.6	529.0
Neighborhood method N_c^k with $c \leq 12$	50.6	748.3

Example 3. This example is from the extract of the 1995 Israeli Census Sample File for Individuals aged 15 and over, $N = 248,983, n = 2,490, K = 8,800$, with attributes Sex (2) Age Groups (16), Years of study (25), and Occupation (11).

Model	τ_1	τ_2
True Values	5	36.9
Argus	7.7	35.5
Log Linear Model: Independence	6.4	44.2
Log Linear Model: 2-Way Interactions	1.1	26.4
Neighborhood method M_a^k	0	30.0
Neighborhood method M_a^k excluding structural zeros	0	25.0
Neighborhood method N_c^k	0	30.1
Neighborhood method N_c^k excluding structural zeros	0	25.5

Discussion of examples In Example 1, the independence log-linear model and the neighborhoods model overestimate the two risk measures. As expected (see Section 3), the log-linear model with two-way interactions, which provides the best estimates here, and the exclusion of structural zeros in the neighborhood method yield lower risk estimates. The neighborhood models which take structural zeros into account yield reasonable estimates, while Argus underestimates risk.

In Example 2, again the two-way interaction model wins, while Argus and the neighborhood model with $C = 12$, which requires heavy calculations and therefore was so far done only once, are doing reasonably well.

In Example 3 Argus comes out best, while here the log-linear independence model does well and it is better than the two-way interaction model, which was the winner in the previous two examples, although it is hard to believe that variables like Age, Years of Study, and Occupation can be independent. A similar phenomenon occurred in another experiment from the same file, with the ordinal attributes Age (71, top coded at 85+), Groups of Years of Study (18), and Income Groups (18). The log-linear model of independence gave the best results, although the variables cannot be independent.

This raises the following question: in a multi-way table, how would one choose the right model? Will the best fitting model by standard measures of goodness of fit provide the best risk estimation results? Skinner and Shlomo (2005) deal with this question. In Example 3, the risk estimates from the two-way interaction model are also quite good, but it seems that in higher dimensional tables, with many possible models, the problem of model selection will be crucial.

Our preliminary **conclusions** are that the new neighborhood approach presented here proposes a natural model which like the other methods needs to be refined and fine-tuned. We expect the new model to work well relative to log-linear models in multi-way tables when simple log-linear models are not valid. We intend to incorporate our approach into a more general regression model, the Negative Binomial Regression, which subsumes the Poisson regression model (Cameron and Trivedi 1998), invoke sampling weights and calibration to partial information on the population, and thus combine the new ideas with known aspects of regression models and ideas of Argus. The burden of proof is still on us.

References

Benedetti, R., Capobianchi, A, and Franconi, L. (1998) Individual Risk of Disclosure Using Sampling Design Information.

Bethlehem, J., Keller, W., and Pannekoek, J. (1990) Disclosure Control of Microdata, *J. Amer. Statist. soc.* , **85**, 38–45.

Cameron, A. C., and Trivedi, P. K. (1998) Regression analysis of count data. *Econometric Society Monographs*, **30**. Cambridge University Press.

Elamir, E. and Skinner, C. (2006) Record-level measures of disclosure risk for survey microdata, *Journal of Official Statistics*, to appear.

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, http://neon.vb.cbs.nl/casc/

Polletini, S. and Stander, J. (2004), A Bayesian Hierarchical Model Approach to Risk Estimation in Statistical Disclosure Limitation, in (J. Domingo-Ferrer and V. Torra, eds.), *Privacy in Statistical Databases*, Springer-Verlag, New York, 247–261

Rinott, Y. (2003) On models for statistical disclosure risk estimation, *Proceedings of the Joint ECE/ Eurostat Work Session on Statistical Data Confidentiality*, Luxemburg , 275-285.

Skinner, C. and Shlomo, N. (2005), Assessing disclosure risk in microdata using record-level measures. *In this volume*.

Skinner, C. and Holmes, D. (1998), Estimating the Re-identification Risk Per Record in Microdata, *J. Official Statist.*, **14**, 361-372.

Willenborg, L. and de Waal T. (2001) Elements of Statistical Disclosure Control , *Lecture Notes in Statistics*, **155**, Springer, New York.

Zhang C.-H. (2005) Estimation of sums of random variables: examples and information bounds, to appear in *Ann. Statist.*, **33**.

ROMM Methodology for Microdata Release

Daniel Ting, Stephen Fienberg*, Mario Trottini***

*** Carnegie Mellon University, Pittsburgh, PA 15213, USA. (dting@stat.cmu.edu; fienberg@stat.cmu.edu)**

**** Departmento de Estadística e I.O , Universidad de Alicante, Spain (mario.trottini@ua.es)**

Abstract. Statistically defensible methods for disclosure limitation allow users to make inferences about parameters in a model similar to those that would be possible using the original unreleased data. We present a new perturbation method for protecting confidential continuous microdata–**R**andom **O**rthogonal **M**atrix **M**asking (ROMM) which preserves the sufficient statistics for multivariate normal distributions, and thus is statistically defensible. ROMM encompasses all methods that preserve these statistics and can be restricted to provide "small" perturbations. We discuss methods for evaluating the disclosure risk and data utility of ROMM.

1. Introduction

Statistical agencies and publicly-funded researchers are under a dual obligation to share data with others, especially in the form of detailed microdata, and at the same time preserving the confidentiality of the respondents who provided these data. To protect the data, they typically must do something beyond removing obvious identifiers from the individual records. When the data come from a sample survey, the sampling rate may be sufficient to provide suitable protection, although not necessarily for local geographic areas. The natural question is then: "How should we modify the data?" To answer this question we need to ask: "What algorithm should one used to modify the data?", "How can we ascertain the extent to which confidentiality is protected?", "How useful are the modified data?".

For continuous microdata, the addition of noise is perhaps the most popular method, e.g., see Kim (1986) and Kim and Winkler (1995), but other proposals include GADP (Muralidhar, et al. (1999)), information preserving statistical obfuscation or IPSO (Burridge (2003)), PRAM (Gouweleeuw, et al. (1998)), CTA (Cox, et al. (2004)), Latin hypercube sampling (Dandekar, et al. (2001)), rank-swapping (Reiss, et al. (1984)), data-shuffling (Muralidhar and Sarathy (2005)), and multiple imputation (Rubin (1993)). All of these techniques succeed at some level in protecting the confidentiality of the data, although there is disagreement as to the extent of the protection. The degree to which they provide useful data for the analyst is, however, a far more contentious is-sue. Fienberg (2005) and others have argued that statistically defensible methods for disclosure limitation need to allow the data analyst to make inferences about param-eters of interest in a model applicable to the original unreleased data. If the data are transformed then one way to achieve this is to provide details of the method to allow the creation of a usable likelihood function for the true unreleased data, as with PRAM. Another strategy is to preserve essential features of the data as part of the transforma-tion process. Burridge (2003) does this for continuous microdata by preserving the minimal sufficient statistics of the data under an assumption of multivariate normality, i.e., the mean and covariance matrix. Such a choice means that the user who sets out to apply a normal-distribution-based multivariate method will get the same estimates for the underlying parameters, but when such estimates are applied to the transformed data, to produce residuals for example, the new results should reflect the added un-certainty associated with the perturbation process. Some methods only approximately preserve features of the data. For example, CTA minimizes changes to sample means and covariances while adjusting cells in tabular data to meet a confidentiality objective while Latin hypercube sampling minimizes differences between sample rank correla-tion matrices of the masked and original data.

In Section 2, we propose a new perturbation method, *Random Orthogonal Matrix Masking* (ROMM), for continuous microdata. ROMM's principal features are:

- It preserves *sample* means and *sample* covariances, the sufficient statistics of a multivariate normal, and hence it exactly preserves linear regression estimates.
- It controls the magnitude of the perturbation, so useful analyses can often be performed on the perturbed data even when the underlying model is not assumed to be a multivariate normal.

We describe the implementation of the method in detail in Section 3, and then, in Section 5, we discuss evaluating the level of confidentiality protection and the usefulness of the perturbed data formally in terms ofrisk and utility.

2. Random Orthogonal Matrix Masking

We begin by introducing **R**andom **O**rthogonal **M**atrix **M**asking (ROMM). We then link ROMM to matrix masking, and we show that it encompasses other previously proposed methods. The procedure for ROMM is as follows:

1. Remove identifying variables such as name, address, and social security or other forms of publicly accessible identification numbers and represent the resulting data as an $n \times k$ matrix x.
2. Generate a random orthogonal matrix, t, from a distribution G defined on the group of $n \times n$ orthogonal matrices which keep 1_n invariant, i.e., $t1_n = 1_n$ where 1_n is the column vector consisting of n 1's.
3. Apply the orthogonal operator, t to the original data x to produce perturbed microdata y: $y = tx$.
4. Release to the users:
   (a) The output of the transformation, y;
   (b) The information that y has been obtained applying to the original data an orthogonal operator randomly generated from a distribution G;
   (c) The exact distribution G.

ROMM is a specific case of matrix masking as described by Duncan and Pearson (1991), which for the $n \times k$ data matrix x involves the transformation

$$x \longrightarrow y = AxB + C.$$

For ROMM, B is the identity, C is the zero matrix, and the class of masks consists of those A's that are random orthogonal matrices drawn from some known distribution G. Mera (1997) earlier proposed the use of orthogonal matrices to perturb data by preserving means and the covariance matrix, but he restricted attention to symmetric orthogonal matrices.

ROMM was designed so that the sample means and sample covariance matrix for x and y are the same. Furthermore, we can show that, for *any* pair of matrices x', y' with the same sample means and sample covariance matrix, there exists an orthogonal matrix t that keeps 1_n invariant and $y' = tx'$. Because the IPSO method of [1] preserves the sample means and sample covariance matrix, it is a special case of ROMM. The distribution G that is equivalent to IPSO is given in the appendix. We give two theorems that formally codify these statements and refer the interested reader to the proofs in Ting, et al. (2005).

Theorem 1. *Let $\overline{x}$ and Σ_x be the sample mean and the sample covariance matrix of the original microdata and let $\overline{y}$ and Σ_y be the corresponding quantities in the masked microdata produced by ROMM. Then*

$$\overline{x} = \overline{y} \quad \text{and} \quad \Sigma_x = \Sigma_y .$$

Theorem 2. *Let M be any data masking procedure that generates a random microdata, y, with the same sample mean and sample covariance matrix as the original microdata. Then M is a special case of ROMM for a suitable choice of the "parameter" G.*

The preservation of the sample means and sample covariance matrix has special practical and theoretical features. On a practical level, simple linear regression estimates are preserved exactly. On a theoretical level, the sufficient statistics are preserved when the underlying distribution is assumed to be multivariate normal.

The second feature of ROMM, namely the ability to control the magnitude of perturbation, is achieved through an appropriate choice of distribution G. A perturbation is considered small if the (Riemannian) distance d of the orthogonal matrix to the identity is close to zero. Section 3 describes some choices of G for which draws from G tend to correspond to small perturbations.

Comparison With Additive Perturbation Methods

Much of the analysis of additive perturbation methods has focused on the effects on regression estimates and estimating the covariance matrix of the underlying distribution. How does ROMM compare with such methods in the regression setting?

Under the naive additive perturbation method, uncorrelated additive noise, the problem of estimating regression coefficients reduces to the well-known problem of estimation with measurement error in the covariates. The usual regression estimates are biased towards 0 and inconsistent. Hence, they must be bias corrected. The resulting bias corrected estimates on perturbed data have greater variance than estimates based on unperturbed data. See Lechner and Pohlmeier (2004) for related discussion.

With perturbed data obtained through bias corrected and correlated additive noise in Kim (1986) or with GADP in Muralidhar, et al. (1999), the sample mean and sample covariance estimates on the perturbed data are unbiased estimates of the true *estimates* obtained from the original data. While this means they are also unbiased estimates of the true parameters in a multivariate normal model, it underscores the fact that additional variability is introduced. The Rao-Blackwell theorem shows that this increase in variance is strict unless the sufficient statistics of the unperturbed data can be recovered exactly from the perturbed data. This is clearly not the case under these additive noise methods. Furthermore, the usual regression estimates under these additive noise methods are not necessarily unbiased. They are, however, consistent.

3. Implementation: Distributions on Orthogonal Operators

In this section we describe some choices for G, the distribution on orthogonal matrices, and show how an appropriate choice of the parameters for G results in small perturbations. To simplify the descriptions, we temporarily ignore the requirement that the vector 1_n must be held invariant. This deficiency is easily corrected by treating an $(n-1)\times(n-1)$ orthogonal matrix as a linear operator on the space orthogonal to 1_n and then extending it to hold 1_n invariant to obtain an $n \times n$ orthogonal matrix. We give details in Ting, et al. (2005).

Coordinate by Coordinate and Uniform Distributions

Using the idea that small perturbations are matrices "close to" the identity, we add a small amount of noise to the identity and then modify this matrix to be orthogonal. The algorithm is as follows.

1. Choose a parameter $\lambda > 0$ corresponding to the magnitude of perturbation.
2. Draw an $n \times n$ random matrix m with entries from a standard normal.
3. Put $p = I + \lambda m$.
4. Apply Gram-Schmidt and normalize the columns of p to obtain an orthonormal matrix t.

When $\lambda = 0$, t is the identity and no perturbation has occured. When $\lambda \to \infty$, t is a draw from the uniform distribution on orthogonal matrices (according to Haar Measure) (according to Haar Measure) (see Eaton (1983, p. 234)).

Block Diagonal Distribution

Another approach to control the magnitude of perturbations is to consider orthogonal matrices with eigenvalues close to 1. For simplicity assume n is even. To sample from the block diagonal distribution, we perform the following steps.

1. Draw s from the uniform distribution on orthogonal matrices.
2. For $j = 1 ... n/2$, independently draw $r_j \sim Beta(\alpha, \beta)$ for some choice of parameters α and β. Put $\theta_j = 2\pi r_j - \pi$.
3. Let b be the block diagonal matrix where each block is the 2×2 matrix

$$\begin{pmatrix} cos\theta_j & -sin\theta_j \\ sin\theta_j & cos\theta_j \end{pmatrix}$$

4. Put $t = sbs^T$. Then t is an orthogonal matrix with eigenvalues $e^{\pm i\theta_j}$. Furthermore, the support of the resulting block diagonal distribution contains all orthogonal matrices except for a set of measure 0.

As $\alpha = \beta \to \infty$, the distribution of each θ_j converges to 0, and t converges to I. Unlike the coordinate by coordinate distribution, however, the block diagonal distribution cannot reproduce the uniform distribution because the eigenvalues of orthogonal matrices drawn from the uniform distribution are highly correlated (see DiaconisEigenvalues and Shahshahani (1994)), and this distribution assumes independence of eigenvalues.

4. Example: Boston Housing Data

We demonstrate ROMM's performance using a subset of 4 variables from the Boston house price data in Harrison and Rubinfeld (1978):

Variable	Meaning
RM	average number of rooms per dwelling
PTRATIO	pupil-teacher ratio by town
LSTAT	% lower status of the population
MEDV	Median value of owner-occupied homes in $1000's

We may treat `LSTAT` as the sensitive variable, i.e., homeowners in a particular tract do not want to disclose what percentage of people in that tract have low socioeconomic status. The regression model of interest has `MEDV` as the dependent variable and the remaining ones as predictors. We drew the following random subset of 13 observations:

Ob	RM	PT-RATIO	LSTAT	MEDV
1	6.630	18.5	6.53	26.6
2	5.986	19.1	14.81	21.4
3	5.709	14.7	15.79	19.4
4	5.877	14.7	12.14	23.8
5	6.402	14.7	11.32	22.3
6	6.782	15.2	6.68	32.0
7	6.433	19.1	9.52	24.5

Ob	RM	PT-RATIO	LSTAT	MEDV
8	6.315	16.6	7.60	22.3
9	6.023	18.4	11.72	19.4
10	6.251	20.2	14.19	19.9
11	5.757	20.2	10.11	15.0
12	5.304	20.2	26.64	10.4
13	6.425	20.2	12.03	16.1

We use two perturbed datasets, one using the coordinate-by-coordinate distribution for ROMM with $\lambda = 1/3$, and the other using the bias corrected, correlated additive noise method described in Kim (1986) with $\sqrt{c} = 1/2$. We give the difference of the perturbed data under each method and the original data below.

ROMM				Additive Noise			
RM	PT-RATIO	LSTAT	MEDV	RM	PT-RATIO	LSTAT	MEDV
-0.253	3.725	4.721	-7.681	-0.337	1.509	1.848	-2.173
-0.529	2.006	3.843	-9.182	0.182	2.249	-1.479	-1.238
0.494	0.738	-12.337	5.930	-0.036	2.335	-0.761	-3.027
0.045	0.880	1.249	-0.095	0.235	-0.132	-3.451	0.908
0.223	1.313	-1.086	3.740	-0.057	1.274	3.057	-3.099
-0.183	1.841	0.338	-3.079	-0.324	0.380	1.999	-3.797
-0.269	2.093	1.883	-4.906	-0.281	-1.957	3.303	0.980
0.139	-0.867	0.348	3.018	0.175	-0.046	-0.831	0.765
0.414	0.367	-1.688	0.357	0.053	-1.423	-2.049	1.165
0.212	-3.778	-5.139	4.602	0.086	-0.116	0.367	0.883
-0.437	-3.218	10.437	-2.359	-0.107	0.195	2.658	0.143
0.601	-2.851	-6.872	12.014	0.390	-0.581	-8.411	4.243
-0.457	-2.249	4.302	-2.361	-0.468	0.254	6.544	-5.830

The magnitude of the differences is in generùal much larger under ROMM and the following sample variances give a rough idea of by how much:

	RM	PTRATIO	LSTAT	MEDV
ROMM	0.145	5.618	33.307	34.327
Additive noise	0.066	1.641	13.950	7.346

The linear regression estimates for the original data set (and ROMM) and for the additive noise version are as follows:

Original Data (and ROMM)			Additive Noise		
Variable	Estimate	S.E.	Variable	Estimate	S.E.
(Intercept)	-5.564	23.652	(Intercept)	-2.470	25.183
RM	7.449	3.366	RM	6.266	3.584
PTRATIO	-0.956	0.369	PTRATIO	-0.393	0.433
LSTAT	-0.177	0.274	LSTAT	-0.678	0.300

The differences between regression estimates are substantial in this case. For `PTRATIO` and `LSTAT` the difference in coefficients is greater than the estimated standard error for each coefficient. Further, we note that the inferences under the additive noise method do not reflect the added uncertainty of the perturbation.

5. Disclosure Risk and Data Utility

Like any data masking procedure aimed at finding a suitable balance between *safety* and *usability* of the perturbed data, ROMM relies on an implicit assumption that targets of potential intruders do not overlap with targets of legitimate data users, c.f. Trottini (2004). The underlying assumption is that researchers, policy makers, and public opinion are interested in statistical analyses aimed at discovering and making inferences about general features of the population represented by the data (e.g., association among variables, or the models for different type of phenomena) while intruders are interested in identifying confidential information about individual respondents. Within this framework, for any given unperturbed data set the knowledge of the distribution that has generated the data is "sufficient" for any statistical analysis that legitimate data users might wish to perform. ROMM exploits this idea by preserving general features of the distribution while increasing the difficulty for an intruder to recover confidential information about individual respondents through perturbation.

A rigorous assessment of disclosure risk and utility requires: (a) A model for *users'* behaviors when the ouput of ROMM is released, (b) An assessment of agency uncertainty about this model's inputs (*users'* targets, prior information, estimation procedure, etc.), (c) A formalization of agency's perception of the consequences of data users' actions and of the agency's preference structure for consequences of *users'* actions (see Trottini (2004)).

Because of space limitations, here we consider a simplified scenario where: (i) the modeling of *users* and agency's behaviors does not take explicitly into account some relevant aspects of the problem, such as agency's perception of usefulness in terms of model checking, diagnostics, and feasibility of the users' inferences under the released data, and (ii) the agency has no uncertainty about the *users*' model inputs. We can extend the results to more realistic scenarios by explicitly incorporating agency's uncertainty on *users'* model inputs as described in Trottini (2004), and by explicitly formalizing agency's perception of usefulness in terms of model checking, diagnostics, and feasibility of users inferences using a suitable structuring of objectives and attributes (see Trottini (2005)).

5.1. Notation and Posterior Distribution

We first fix some notation, and before evaluating risk and utility, we give a formula for the posterior distribution of the unreleased data, x. Then we use these in evaluating utility under non-normality assumptions and in evaluating disclosure risk.

1. Let X and T be random variables representing, respectively, the unperturbed data and the random orthogonal matrix used to transform the unperturbed data.
2. Let x_{orig} be the realized value of X data (i.e., the true values for the unperturbed data) and let t be the realized value for T.
3. Define $Y = TX$ to be the random variable representing the perturbed data. Then $y = tx_{orig}$ is the realization of Y.
4. Define E to be any external knowledge available to the user. In particular E may contain non-confidential values in x_{orig} that may be used for record linkage or to undo the perturbation t.
5. Let $\tilde{m}$ be the masked data set produced by ROMM. Following Trottini (2004) we can represent $\tilde{m}$ as a pair $\tilde{m} = (y, I(T,G))$ where $I(T,G)$ represents the information provided by the agency to the *users* about the transformation (in this case, the information that the released microdata has been obtained by applying to the original data a random orthogonal operator, T, generated from a distribution G.)
6. Let $A(x,y) = \{t : y = tx\}$ be the set of orthogonal matrices such that 1_n is invariant under t and that take x to y.

The posterior distribution is then given by

$$\pi(x \mid \tilde{m}, E) \propto \pi(x \mid E) \cdot L(\tilde{m}; x, E) = \pi(x \mid E) \cdot \int_{A(x,y)\cap Supp(G)} dG(t)$$

where $Supp(G)$ is the support of G. Note that we assume G is given to all users since valid inferences are not possible otherwise. This formula also assumes that T is independent of X which may not be true. We give the full derivation of this result and a discussion of the case when T and X are dependent in Ting, et al. (2005). Another point worth noting is that, for $n \times k$ matrices x and y, the dimension of $A(x,y)$ is $(n-k)(n-k-1)/2$. Thus we need to compute a high-dimensional integral, and this is difficult. For $n = 100$ and $k = 10$ the dimension is 4000! Under some assumptions on the users' prior and on G, however, it is possible to sample from the posterior distribution (see Ting, et al. (2005)).

5.2. Data Utility

Utility Under Normality

To assess data utility associated with ROMM, we first consider the procedure under normality assumptions. For this case, the notation given above is not important, but it will be when we consider the nonnormal case. If the original data are independent and identically distributed (i.i.d.) realizations from a multivariate normal distribution, the output of ROMM is a random sample from the the same multivariate normal distribution (any orthogonal transformation that preserves 1_n preserves the multivariate normal distribution of the original data). Thus, regardless of the distribution G used to generate the orthogonal operator and regardless of the inferences of interest for legitimate data users, under normality the ROMM procedure guarantees maximum data utility. The trade-off dilemma in this case is trivial. The statistical agency should choose the "noise parameter," G, to minimize the risk of disclosure, since data utility is constant (and maximum!) as a function of G.

Utility Under Non-normality

When the data are not normal, ROMM preserves the mean vector and the covariance matrix of the unperturbed data but no longer the distribution. Heuristically, the idea of using small perturbations suggests that using y directly as input may still be useful in exploratory data analyses. For more rigorous analyses, however, legitimate data users interested in inferences other than the mean and covariances, e.g., in quantiles or mixture models, cannot directly use the output y as input to their standard statistical analyses and expect the resulting inferences to be valid. To make valid inferences, the output y is relevant only to the extent that provides information about the transformation, t, that has been applied to the original data and thus, indirectly, about the original data x_{orig}. Formally, a user must make inferences through the posterior distribution of x_{orig} given y and any additional information E.

For a given legitimate data user's target, denote by $Z(x_{orig})$ the inferences that legitimate data users would make if they had access to the original data. His/her inference under the released ROMM data will be:

$$Z(\tilde{m}) = \int Z(x) \cdot \pi(x \mid \tilde{m}, E)dx \tag{1}$$

where $\pi(x \mid \tilde{m}, E)$ is the *user's* posterior distribution for the unperturbed data. Note, however, that the computation of $\pi(x \mid \tilde{m}, E)$ in this case can be very complex.

Depending on the agency's interpretation of *usefulness* of the data, we can define different measures of data utility as a function of (1). If, for example, the agency is concerned with minimizing the difference between data users inferences with the original data and the corresponding inferences with the masked data set, then a generic measure of data utility might be $DU = D(Z(\tilde{m}), Z(x_{orig}))$, where D is a distance metric that depends on the nature of the target Z and on the features of the inferences that should be preserved. See Trottini (2004) for examples.

5.3. Disclosure Risk

There are multiple ways in which a data intruder can pose a risk to a statistical agency in charge of releasing the microdata. The way we consider here is to take disclosure risk as the data intruder's utility. The problem then is similar to that of data utility for legitimate users. A data intruder has target $\zeta(i, x_{orig})$ where i corresponds to an individual with some row of values in x_{orig}. The disclosure risk may then be some distance metric between the target estimate and the true target value. Alternatively, it may be a reflection of the intruder's uncertainty of the target, such as the variance in the estimate of the target.

There is a subtle difference between this disclosure risk case and the case considered under data utility above. In this case, the row number r corresponding to the targetted individual is unknown to the intruder. Thus, exchanging the rows of x_{orig} corresponds to changing the targetted individual when a particular row r is targetted. For a legitimate user's target, in many cases the rows corresponding to individuals may be exchanged without affecting the estimates. For example, the maximum of a list of numbers is invariant under permutation. Because of this difference, we account for the uncertainty of r by writing $x_{perm} = p^T x_{orig}$ where p^T is a permutation matrix that takes the r^{th} row to the 1^{st} row. The intruder's target $\zeta(i, x_{orig})$ then becomes a function $\zeta_{perm}(x_{perm})$ which does not directly depend on i. We may then treat p as the realization of a random variable P with some distribution on permutation matrices (not necessarily the uniform distribution). Since a permutation matrix is an orthogonal matrix that holds 1_n invariant, it follows that TP is an orthogonal matrix that holds 1_n invariant, and it has some distribution G_P. Since $y = (tp)x_{perm}$, the problem of estimating $\zeta(x_{perm})$ reduces to calculating the posterior as in the data utility case, but with G replaced by G_P. Note that when G is the uniform distribution, $G_P = G$ by the left invariance of G.

6. Conclusion

We have introduced the disclosure limitation method of Random Orthogonal Matrix Masking (ROMM), and we demonstrated some of its theoretical properties. In particular, ROMM is designed to preserve the sufficient statistics of a multivariate normal distribution and thus preserves many common statistical quantities favored by users in their analyses, e.g., linear regression estimates. Further, ROMM encompasses the entire class of perturbation methods that preserve these sufficient statistics. The procedures for ROMM we introduced suggest how it utilizes “small perturbations” of the data. We then considered the disclosure risk and data utility associated with ROMM, showing how to assess them within a Bayesian framework using the posterior distribution of the original data given the perturbed data and other information.

There is considerable work to be done to turn ROMM into a complete statistically-defensible disclosure limitation method. For example, while the distributions on orthogonal matrices described here heuristically favor small perturbations, we do not have theoretical results as to what the magnitude of the perturbation is, nor do we have theoretical results about the resulting distribution of perturbed data when the underlying distribution is not normal. From a practical perspective, although we gave a formula for the posterior, we also noted that it is difficult to calculate. In Ting, et al. (2005), we show how to sample from the posterior under certain assumptions about the prior. How these difficulties and assumptions will play out remains to be seen.

Acknowledgements

This research was supported in part by NSF grant EIA-0131884 to the National Institute of Statistical Sciences, and by Army contract DAAD19-02-1-3-0389 to CyLab, at Carnegie Mellon University.

Appendix: IPSO and ROMM

ROMM and information preserving statistical obfuscation (IPSO) are similar in that both preserve the sufficient statistics for a multivariate normal. IPSO can be described as follows:

1. Given an $n \times k$ matrix x, let $\hat{\mu}$ be the vector of sample means and $\hat{\Sigma}$ be the sample covariance matrix.
2. Construct an $n \times k$ matrix r of new “residuals” where the columns of r are orthogonal to each other and also to 1_n, the vector of all 1’s.
3. Let L be the lower triangular matrix in the Choleski factorization $n\hat{\Sigma} = LL^T$.
4. Let $\tilde{r} = rL^T$ and $\tilde{\mu}$ be the $n \times k$ matrix where every row is $\hat{\mu}$. It follows that $y = \tilde{\mu} + \tilde{r}$ has sample means given by $\hat{\mu}$ and sample covariance matrix $\hat{\Sigma}$.

Though Burridge does not explicitly state what distribution r should be drawn from, he describes the case where r is drawn from i.i.d. normals and the columns are orthonormalized. In this case, the perturbed dataset is a draw from a multivariate normal given the sufficient statistics $\hat{\mu}$ and $\hat{\Sigma}$. When r is drawn as described above, IPSO is a special case of ROMM where the orthogonal matrix t is drawn using the coordinate-by-coordinate approach with the parameter $\lambda \to \infty$. For a proof of this refer to Ting, et al. (2005).

Domingo-Ferrer, et al. (2005) report on empirical comparisons of IPSO (their IPSO-C) with two less constrained methods and it appears to perform well for data protection. The ability to choose parameter settings in ROMM to optimize performance would lead us to expect that it would perform even better.

References

Burridge, J. (2003) Information Preserving Statistical Obfuscation, *Statistics and Computing*, 13, 4321–327.

Cox, L.H., Kelly, J., and Patil, R. (2004) Balancing Quality and Confidentiality for Multivariate Tabular Data in (J. Domingo-Ferrer and V. Torra, eds.) Privacy in Statistical Databases 2004, Springer-Verlag, 87–98.

Dandekar, R.A., Cohen, M., and Kirkendall, N. (2001) Applicability of Latin Hypercube Sampling to Create Multivariate Synthetic Microdata. *Proceedings of ETKNTTS 2001*, Eurostat, 839–847.

Diaconis, P. and Shahshahani, M. (1994) On Eigenvalues of Random Matrices, *Journal of Applied Probability*, 10, 49–62.

Domingo-Ferrer, J., Torra, V., Mateo-Sanz. J.M., and Sebé, F. (2005) *Empirical Disclosure Risk Assessment of the IPSO Synthetic Data Generators,* UNECE/Eurostat Work Session on Statistical Data Confidentiality.

Duncan, G.T., Pearson, R.W. (1991) Enhancing Access to Microdata While Protecting Confidentiality: Prospects for the Future, *Statistical Science*, 6, 219–232.

Eaton, M. (1983) *Multivariate Statistics: A Vector Space Approach*, Wiley, New York.

Fienberg, S. E. (2005). Allowing Access to Confidential Data: Some Recent Experiences and Statistical Approaches, Submitted for publication.

Gouweleeuw, J. M., Kooiman, P., Willenborg, L. C. R. J., and Wolf, P. P. de. (1998) Post Randomization for Statistical Disclosure Control: Theory and Implementation, *Journal of Official Statistics*, 14, 463–478.

Harrison, D. and Rubinfeld, D.L. (1978) Hedonic Prices and the Demand for Clean Air, *Journal of Environmental Economics and Management*, 5, 81–102.

Kim, J. J. (1986) A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation, *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 303–308.

Kim, J. J., and Winkler, W. E. (1995) Masking Microdata Files, *Proceedings of the Section on Survey Research Methods*, American Statistical Association,114–119.

Lechner, S. and Pohlmeier, W. (2004) To Blank or Not to Blank? A Comparison of the Effects of Disclosure Limitation Methods on Nonlinear Regression Estimates. *Privacy in Statistical Databases 2004*, Springer-Verlag, 187–200.

Mera, R. (1997) Matrix Masking Methods which Preserve Moments. (1997) *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 445–450.

Muralidhar, K., Parsa, R. and Sarathy, R. (1999) A General Additive Data Perturbation Method for Database Security, *Management Science*, 45, 1399–1415.

Muralidhar, K. and Sarathy, R. (2005) Data Shuffling: A New Masking Approach for Numerical Data. Unpublished manuscript.

Reiss, S. P., Mark, J., and Dalenius, T. (1982) Non-reversible Privacy Transformations, *Proceedings of the ACM Symposium on Principles of Database Systems*, Los Angeles, CA, 139–146.

Rubin, D. B. (1993) Satisfying Confidentiality Constraints Through the Use of Synthetic Multiply Imputed Microdata. *Journal of Official Statistics*, 9, 461–468.

Ting, D., Fienberg, S. E., and Trottini, M. (2005) Unpublished Technical Report.

Trottini, M. (2004) *Decision Models for Data Disclosure Limitation*, Ph.D. thesis, Carnegie Mellon University.

Trottini, M. (2005) *Statistical Disclosure Limitation in Longitudinal Linked Data: Objectives and Attributes*, UNECE/Eurostat Work Session on Statistical Data Confidentiality.

Bayesian methods for disclosure risk assessment

Jonathan J Forster
Southampton Statistical Sciences Research Institute, School of Mathematics,
University of Southampton, UK. (J.J.Forster@soton.ac.uk)

Abstract. Measures of the risk of individual identification in the release of categorical microdata are commonly based on the probability of an intruder correctly matching an individual in the population to a record in the released data. In this paper, we discuss how such probabilities can be interpreted and focus on Bayesian predictive probabilities as risk measures. By utilising a Bayesian approach to estimation under model uncertainty, known as model-averaging, we can provide more realistic estimates of disclosure risk for individual records than are provided by methods which ignore the multivariate structure of thedata set. The method is illustrated with two examples.

1. Introduction

Suppose that an agency releases categorical data on a sample of individuals from a population. Then, the sample data can be expressed as a a multiway contingency table. Identification risk occurs when there are small sample cell counts (particularly uniques) in the marginal table representing the cross-classification of individuals by a subset of *key* variables (those variables whose values in the population are available to a potential intruder from a source external to the released data under consideration). If the intruder can determine, with confidence, that a record in the released contingency table of key variables matches a particular individual in the population, then this record can be identified and the data release allows disclosure of the values of the remaining (non-key) variables for this individual. In this paper we focus on the disclosure risk associated with the release of individual records as part of a larger database.

It is common to quantify individual record disclosure risk as the probability of an individual being identified. Let $f_1,...,f_K$ denote the sample cell counts in the contingency table of key variables and $F_1,...,F_K$ the corresponding population cell counts and let n and N represent the sample and population totals respectively. Then, *given the intruder had knowledge of the population cell counts* $\boldsymbol{F} = (F_1, \dots, F_K)$, they could match any individual in the population whose record belonged in cell j, with a record chosen from the sample, and (in the absence of any other information) would be able to evaluate the probability of a correct match as $1/F_j$. We denote this as the conditional probability

$$P(E_j \mid \boldsymbol{F}) = \frac{1}{F_j} \tag{1}$$

where E_j is used to denote the event that an individual whose record is in cell j is correctly matched. This measure was proposed by Benedetti and Franconi (1998). In practice, the intruder only has knowledge of the sample cell counts $\boldsymbol{f} = (f_1, \dots, f_K)$ and cannot calculate (1). One alternative is to use the sample data to estimate the F_j in (1). Alternatively,it might be considered that the relevant disclosure risk measure is $P(E_j \mid \boldsymbol{f})$, the predictive probability of a correct match given only the sample data. The easiest interpretation of this quantity is within a Bayesian statistical framework, as follows.

Bayesian inference uses probability to quantify uncertainty. Hence, the uncertainty about any unknowns prior to obtaining sample data is encapsulated in a prior probability distribution. On observing data, this is then updated to a posterior distribution using Bayes theorem. In the present context, the unknowns are the population counts $\boldsymbol{F}$, the sample data are $\boldsymbol{f}$ and we obtain the posterior distribution as

$$P(\boldsymbol{F} \mid \boldsymbol{f}) \propto P(\boldsymbol{F})P(\boldsymbol{f} \mid \boldsymbol{F}) \tag{2}$$

where $P(\boldsymbol{F})$ is the prior distribution for $\boldsymbol{F}$ and $P(\boldsymbol{f} \mid \boldsymbol{F})$ represents the sampling distribution of the observed table. Having observed the sampled records it is only the unsampled records about which uncertainty remains, hence (2) can be replaced by

$$P(\boldsymbol{F}-\boldsymbol{f} \mid \boldsymbol{f}) \propto P(\boldsymbol{F}-\boldsymbol{f})P(\boldsymbol{f} \mid \boldsymbol{F}-\boldsymbol{f}) \tag{3}$$

Now, we simply note that our required disclosure risk measure, $P(E_j \mid \boldsymbol{f})$, can be expressed simply by using a standard conditional expectation relationship as

$$\begin{aligned} P(E_j \mid \boldsymbol{f}) &= E[P(E_j \mid \boldsymbol{F}) \mid \boldsymbol{f}] \\ &= E\left[1/F_j \mid \boldsymbol{f}\right] \end{aligned} \tag{4}$$

where the expectation is with respect to the posterior distribution $P(\boldsymbol{F} \mid \boldsymbol{f})$, evaluated as in (2). Hence, the predictive probability of disclosure event E_j, given knowledge only of sample data $\boldsymbol{f}$ is equal to the posterior mean of $1/F_j$, the reciprocal of the relevant population cell count.

In the model of Benedetti and Franconi (1998), subsequently extended by Rinott (2003) and Polettini and Stander (2004), the posterior distribution (2) simplifies to

$$P(\boldsymbol{F} \mid \boldsymbol{f}) = \prod_j P(F_j \mid f_j), \tag{5}$$

where each $P(F_j \mid f_j)$ is a negative binomial probability function in Benedetti and Franconi (1998) and Rinott (2003), and a more complex expression in Polettini and Stander (2004). In (5), not only are the population cell frequencies conditionally independent given the sample cell frequencies, as pointed out by Rinott (2003), but perhaps more notably the posterior distribution for F_j given $\boldsymbol{f}$ can be written as $P(F_j \mid f_j)$, and so F_j is also independent of the *sample* cell frequencies in all other cells. In other words, for estimating the disclosure risk in cell j, the only pertinent information is the sample frequency in that cell.

Where *empirical* Bayes estimation is used, as suggested by Rinott (2003) following Bethlehem et al (1990), the observation above is no longer strictly true, as $P(F_j \mid f_j)$ is replaced by $\hat{P}(F_j \mid f_j)$, where the maximisation is performed over the parameters of the prior distribution $P(\boldsymbol{F})$ in (2). This quantity *does* now typically depend on cell frequencies other than f_j. However, for the models which have been typically proposed, it does so in a way which is completely invariant to any permutation of the cell frequencies in other cells. In other words, all that is relevant are the sizes of the cell frequencies in the other cells, and not their positions in the table. The tabular structure of the data is completely ignorable.

Skinner and Holmes (1998) and Elamir and Skinner (2004) adapt the original model of Bethlehem et al (1990) in a way which respects the table structure and hence allows more of the information in the data to be incorporated into disclosure risk estimation. Their approach is equivalent to proposing a prior distribution $P(\boldsymbol{F})$ in (2) which is based on a *log-linear model* for the underlying contingency table. The parameters of the log-linear model are then estimated in empirical Bayes fashion. The sensitivity surrounding which log-linear model to use is somewhat averted by choosing a relatively simple model, but allowing some divergence from the model.

In this paper, we follow the approach of Forster and Webb (2005). This approach is also based on log-linear models for the contingency table of population cell frequencies, but the requirement to choose a model *a priori* is avoided, and any model uncertainty is coherently incorporated into the resulting inferences. The approach is described in detail in the next section.

2. A Bayesian model

Following Omori (1999), we assume that $\boldsymbol{F}$ has a multinomial(N,π) prior distribution. Then, we assume a log-linear model for π. In the current paper, we shall restrict consideration to those log-linear models which are *decomposable graphical models*. For a broader class of log-linear models, see Forster and Webb (2005). The advantage of considering only decomposable graphical models is that computation is made significantly more tractable and efficient. There is some loss of model flexibility, but the decomposable graphical models still constitute a highly flexible model class. For further details of decomposable graphical models, see Lauritzen (1996). For a decomposable graphical model, we shall write

$$\pi = \pi(\beta_m) \tag{6}$$

where m indexes the particular model under consideration, and β_m is the corresponding vector of model parameters, which is of lower dimension than π. For a decomposable model, β_m may be considered to be a collection of marginal probabilities corresponding to those subtables of the full contingency table which are unconstrained by the model. As a prior distribution for β_m, we use the hyper-Dirichlet family, a class of prior distributions based on the Dirichlet distribution for the saturated model (no log-linear constraints) and developed by Dawid and Lauritzen (1994). A hyper-Drichlet distribution consists of a Dirichlet distribution on each set of marginal cell probabilities (sub-vector of β_m) which are unconstrained by the model. These marginal Dirichlet distributions are dependent, where they have common margins. For example, consider a three-way table with cross-classifiying variables A, B and C. The model AB+BC does not constrain the AB or BC marginal distributions, so the corresponding hyper-Dirichlet distribution is composed of Dirichlet distributions for these two margins, but constrained to give a common set of probabilities for the marginal distribution of B.

Having specified a prior distribution, and observed sample cell frequencies $\boldsymbol{f}$, inference concerning disclosure risk is obtained from the posterior distribution $P(\boldsymbol{F}-\boldsymbol{f}\mid\boldsymbol{f})$. In the presence of a model, (3) is replaced by

$$P(\boldsymbol{F}-\boldsymbol{f}\mid\boldsymbol{f}) = \int P(\boldsymbol{F}-\boldsymbol{f},\beta_m\mid f)d\beta_m = \int P(\boldsymbol{F}-\boldsymbol{f}\mid\beta_m,\boldsymbol{f})P(\beta_m\mid\boldsymbol{f})d\beta_m \tag{7}$$

Assuming a sampling scheme under which records are exchangeable, for example Bernoulli sampling or simple random sampling without replacement (the model can be adapted when this is not appropriate), then $\boldsymbol{f}$ and $\boldsymbol{F}-\boldsymbol{f}$ are conditionally independent given β_m and have multinomial (n,π) and multinomial$(N-n,\pi)$ prior distributions respectively. Hence (7) becomes

$$P(\boldsymbol{F}-\boldsymbol{f}\mid\boldsymbol{f}) = \int P(\boldsymbol{F}-\boldsymbol{f}\mid\beta_m)P(\beta_m\mid\boldsymbol{f})d\beta_m. \tag{8}$$

where the first term in the integrand, $P(\boldsymbol{F}-\boldsymbol{f}\mid\beta_m)$, is a multinomial$(N-n,\pi)$ probability function, with π determined from β_m using (6). Using Bayes theorem, the second term of the integrand is

$$P(\beta_m\mid\boldsymbol{f}) \propto P(\boldsymbol{f}\mid\beta_m)P(\beta_m),$$

the product of a multinomial(n,π) probability function for $\boldsymbol{f}$ and the prior density for β_m.

To this point, we have only described inference under a single log-linear model. In practice, it is unlikely that we will be certain about which model is the most appropriate for building the prior distribution for $\boldsymbol{F}$. A Bayesian approach allows this uncertainty to be coherently incorporated into the prior distribution. Let M denote the set of possible models, and suppose that prior uncertainty about m is encapsulated by a prior distribution over M, involving a set of prior model probabilities $P(m)$. In practice, a discrete uniform distribution over M is commonly used, to represent prior ignorance. The prior distribution over $\boldsymbol{F}, m$ and $\{\beta_m, m \in M\}$ now consists of three components, the multinomial $P(\boldsymbol{F} \mid \beta_m, m)$, the prior for the parameters of each possible decomposable graphical model $P(\beta_m \mid m)$ and the prior model probabilities $P(m)$. Note that the first two distributions are now explicitly conditional on m, as both the form of the log-linear model in (6), and the prior distribution for its parameters, will depend on which model is under consideration.

Under model uncertainty, the posterior distribution for the unobserved cell counts $\boldsymbol{F} - \boldsymbol{f}$ in (8) becomes

$$P(\boldsymbol{F}-\boldsymbol{f} \mid \boldsymbol{f}) = \sum_{m \in M} P(m \mid \boldsymbol{f}) \int P(\boldsymbol{F}-\boldsymbol{f} \mid N-n, \beta_m, m) P(\beta_m \mid \boldsymbol{f}, m)\, d\beta_m. \qquad (9)$$

The posterior model probabilities, which appear in (9) but not (8) are obtained, using Bayes theorem as

$$P(m \mid \boldsymbol{f}) = \frac{P(m)P(\boldsymbol{f} \mid m)}{\sum_{m \in M} P(m)P(\boldsymbol{f} \mid m)} \qquad (10)$$

where $P(\boldsymbol{f} \mid m)$ is the *marginal likelihood* for the sampled cell counts, obtained as

$$P(\boldsymbol{f} \mid m) = \int P(\boldsymbol{f} \mid m, \beta_m) P(\beta_m \mid m)\, d\beta_m. \qquad (11)$$

The posterior distribution (9) under model uncertainty is obtained as a weighted average of the posterior distributions (8) under the various models. This is sometimes referred to as model-averaging. Care is required when performing model-averaging, that the quantity which is being averaged is one which shares a common interpretation across the component models. That is clearly the case here, where we are averaging probabilities for cell frequencies. The posterior model probabilities are not of interest in themselves, as we do not actually believe that the population was generated by a particular multinomial log-linear model. Their function is to indicate the appropriate weight, based on the sample data, to be applied to the various models in any inference required. Consequently, they determine the differential impact of other cells, when making inference about a particular population cell frequency.

Having obtained the posterior distribution of the unobserved cell frequencies $F - f$ as in (9), it simply remains to evaluate the risk measure (4), using

$$E[1/F_i \mid \boldsymbol{f}] = \sum_{i=0}^{N-n} \frac{1}{f_j + i} P(F_j - f_j = i \mid \boldsymbol{f}) \qquad (12)$$

where $P(F_j - f_j = i \mid \boldsymbol{f})$ is the marginal posterior probability obtained from (9) by

$$P(F_j - f_j = i \mid \boldsymbol{f}) = \sum_{\mathbf{F}-\mathbf{f}: F_j - f_j = i} P(\boldsymbol{F} - \boldsymbol{f} \mid \boldsymbol{f}). \qquad (13)$$

3. Computation

There are three computational difficulties associated with calculating the predictive probabilities which are proposed as disclosure risk measures. The first is the evaluation of the integrals in (9) and (11). These integrals are analytically intractable for general log-linear models, but can be straightforwardly evaluated when a hyper-Dirichlet prior distribution is used for a decomposable graphical model. The second problem is evaluation of the sum in (9), in cases where the number of models is so large that evaluation of the summand for every model is infeasible. For example, for a six-way contingency table, as in our second example below, there are many thousands of possible decomposable graphical models. Finally, evaluation of the sum in (13) can also be impracticable, as it involves summing over the sample space for all cells except the one currently of interest.

For decomposable models and hyper-Dirichlet prior distributions, some of the calculations can be performed exactly. For example, provided that the number of models under consideration is not too great, marginal likelihoods (11) are available as ratios of products of gamma functions, and hence (10) may be evaluated directly. For more than a few (3 or 4) cross-classifying variables, it is unlikely to be feasible to calculate posterior probabilities for all models. In such examples, Forster and Webb (2005) propose two possible approaches for decomposable models. The first is a Markov chain Monte Carlo (MCMC) approach to sampling, suggested by Madigan and York (1996). The second approach, which we adopt here, is an efficient search stategy for identifying a subset of posterior models with high probability, based on the 'Occam's window' approach of Madigan and Raftery (1994). Posterior model probabilities are then estimated using (10), assuming $P(m \mid \boldsymbol{f}) = 0$ for all m not in the candidate set.

Having obtained model probabilities, we are required to evaluate (9) and hence (13). The integral in (9) is tractable for decomposable graphical models and hyper-Dirichlet prior distributions. However, the sheer size of the sample space for $\boldsymbol{F} - \boldsymbol{f}$ in practical disclosure risk assessment problems makes complete enumeration infeasible. An alternative is to replace $\boldsymbol{F} - \boldsymbol{f}$ in (9) with $F_j - f_j$, and hence obtain $P(F_j - f_j \mid \boldsymbol{f})$ directly. However, when $P(\boldsymbol{F} - \boldsymbol{f} \mid N - n, \beta_m, m)$ is replaced by the binomial probability $P(F_j - f_j \mid N - n, \beta_m, m)$ in the integrand of (9), the integral is no longer tractable. These calculations may, however, be approximated by Monte Carlo sampling from the predctive distribution $P(F_j - f_j \mid \boldsymbol{f})$. This is easily achieved, and just requires sampling from various Dirichlet and binomial distributions. Then, the probabilities $P(F_j - f_j \mid \boldsymbol{f})$ in (13) are simply estimated by sample proportions, which can then be plugged into (12), avoiding the requirement to evaluate (13) by summation. An alternative 'Rao-Blackwellized' calculation described by Forster and Webb (2005) avoids any binomial sampling and reduces Monte Carlo error.

4. Examples

We present two examples. The first is a small example, to illustrate the methodology. The second example is more realistic in terms of size and complexity and is presented to illustrate that the methodology is practicable in disclosure risk assessment applications.

4.1. Example 1: A three-way table

To illustrate the model-averaging approach, we consider the data used by Fienberg and Makov (1998) to illustrate their approach. It is a three-way table representing cross-classification by gender, race and income for a selected US census tract.

Table 1 Three-way table from Fienberg and Makov (1998).
Income categories are 1. ≤$10000, 2. >$10000 and ≤$25000, 3. >$25000

		Gender					
		Male			Female		
		Income			Income		
		1	2	3	1	2	3
Race	White	96	72	161	186	127	51
	Black	10	7	6	11	7	3
	Chinese	1	1	2	0	1	0

Interest focusses on the three uniques in this sample, which we label (C,M,1), (C,M,2) and (C,F,3). Following Fienberg and Makov (1998) we shall investigate the potential disclosure risk of this release, in the cases that the sample represents 10%, 20% and 50% of the population.

We consider two possible sets of hyper-Dirichlet prior parameters, both of which have been suggested as implying weak information concerning the model parameters. In both cases all marginal priors are derived from a symmetric Dirichlet distribution (all parameters taking common value α) for the saturated model. For the first prior, $\alpha = 1/2$, and for the second prior $\alpha = 1/K$, the reciprocal of the number of cells in the table (here, $K = 18$). Where the number of cells is large, then the second prior is likely to be preferred, as $K\alpha$ is a measure of the information in the prior, which increases with K in the first case, but is fixed at 1 in the second case.

Table 2 presents the measure of disclosure risk $E[1/F_j \mid \boldsymbol{f}]$ for the three sample unique cells, for both priors and all three sampling fractions. The first thing to notice is that the inferences for the three cells are different, although not dramatically so. Hence the approach is having the desired effect of incorporating information about the structure in the table. For both priors, the posterior distribution is concentrated on a small selection of models. For prior 1, models R+IG and RG+IG dominate, while for prior 2, model R+IG dominates.

Table 2. Monte Carlo Estimates of $E[1/F_j \mid f]$

		Prior 1 ($\alpha = 1/2$)			Prior 2 ($\alpha = 1/18$)		
		Sampling fraction			Sampling fraction		
		50%	20%	10%	50%	20%	10%
Cell	(C,M,1)	0.595	0.235	0.109	0.707	0.344	0.173
	(C,M,2)	0.665	0.295	0.142	0.766	0.422	0.226
	(C,F,3)	0.570	0.221	0.105	0.651	0.293	0.143

4.2. Example 2: A six-way table

To test the methodology on a more realistic example, we extracted a six-way table of potential key variables from the 3% Individual Sample of Anonymized Records (SAR) for the 2001 UK Census (Office for National Statistics and University of Manchester, 2005). The table extracted consisted of 154295 individuals living in South West England, cross-classified by sex (2 categories) age (coded into 11 categories), accomodation type (8 categories), number of cars owned or available for use (5 categories), occupation type (11 categories) and family type (10 categories). The full table has 96800 cells of which 3796 are uniques. For the purposes of this exercise, this is considered to be the population. To mimic the selection into the SAR, we took a 3% subsample, containing 4761 individuals.

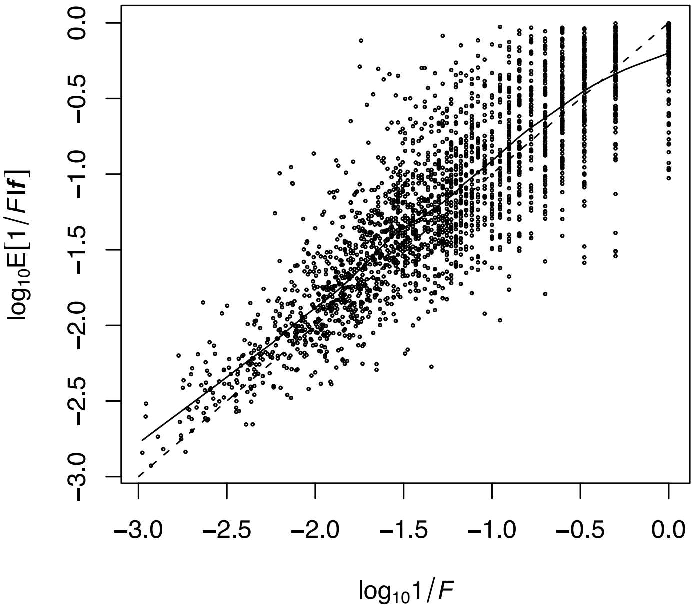
The joint distribution of sample and population cell frequencies for the 96800 cells is summarised in Table 3. Of particular note is the fact that, in the sample data, only 2330 of the 96800 cells are non-empty, and of these 1543 are uniques. Hence, 32% of records, and 66% of cells correspond to sample uniques. Of these cells, only 114 (7%) are population uniques, and the average population total in a sample unique cell is 17, so not all such cells represent disclosure risk.

For each of the 2330 non-empty cells j, we calculated the predictive disclosure probability $P(E_j \mid \boldsymbol{f}) \equiv E[1 / F_j \mid \boldsymbol{f}]$. For this exercise, we are also able to calculate the probability of a disclosure event $P(E_j \mid \boldsymbol{F}) \equiv 1 / F_j$ in the case that full population knowledge was available. We compare these quantities, and hence assess the performance of our disclosure risk assessment procedure by plotting $\log_{10}(1/F_j)$ against the estimated $\log_{10}(E[1 / F_j \mid \boldsymbol{f}])$ for the 2330 non-empty sample cells, in Figure 1.

Table 3. Summary of the joint distribution of sample and population cell frequencies across the 96800 cells of the sampled table.

		Population								
		0	1	2	3	4	5-9	10-19	20+	Total
Sample	0	84867	3682	1694	967	631	1482	757	390	94470
	1	—	114	110	118	104	313	322	462	1543
	2	—	—	0	2	5	28	67	266	368
	3	—	—	—	0	0	1	15	140	156
	4	—	—	—	—	0	0	0	76	76
	5-9	—	—	—	—	—	0	0	125	125
	10-19	—	—	—	—	—	—	0	48	48
	20+	—	—	—	—	—	—	—	14	14
	Total	84867	3796	1804	1087	740	1824	1161	1521	96800

Figure 1. The estimated $\log_{10}(E[1/F_j \mid \boldsymbol{f}])$ against $\log_{10}(1/F_j)$ for the 2330 non-empty sample cells. The dashed line represents equality (no error). The solid line is a loess smooth through the plotted points.



Given that low frequency sample cells correspond to such a wide range of population frequencies, accurate estimation of $1/F_j$, using sample data alone is a difficult task, and without some kind of modelling would be hopeless. Indeed, using any approach which treats the cells as exchangeable, would lead to all 1543 sample uniques having the same estimated risk, in the absence of extra external information. In this context, the model-averaging approach seems to be performing quite well, with perhaps a slight tendency to overestimate risk in this example. This slight overestimation, particularly for low-to-moderate risk cells is apparent when we fit a smooth curve through the points of Figure 1.

We note that without any log-linear modelling, the estimated $E[1/F_j \mid \boldsymbol{f}]$ for any sample unique is evaluated as 0.11, so it is immediately clear that our approach is providing a more accurate measure of risk for the cells with low population counts (genuinely risky records). Indeed, for the 114 genuine population uniques, we computed an average risk of 0.65, while for the 111 sample unique cells with population totals greater than 50, the average risk was estimated as only 0.04. So the method is successfully distinguishing risky and non-risky cells with the same cell counts.

One way of assessing the performance of the method is by considering it as a classifier. Suppose that we define a cell as 'risky' if the probability of a disclosure event is greater than 5%. Then, our method classifies cells as risky if $(E[1/F_j \mid \boldsymbol{f}]) > 0.05$. The 'true' classification is determined using the corresponding (unobserved) value of $1/F_j$. Table 3, shows how our classifier performs. In these terms, the performance seems quite satisfactory, given the small sampling fraction. The sensitivity of the classifier is 88% and its specificity is 76%.

Table 4. Performance of model averaging as a risk classifier, assessed using the 2330 non-empty cells of the sample data table.

		True classification	
		Not risky	Risky
Estimated classification	Not risky	864	140
	Risky	267	1059

5. Discussion

The examples presented in Section 4 illustrate that this approach has potential for identifying cells which may pose a disclosure risk. With the second example, we have started to investigate the performance of the methodology on more realistic examples. In fact, the computational time for this example was not large. It took a few seconds to compute using functions written in R. There therefore remains scope to extend to much more demanding examples. For such examples, it is likely to be neceaary to find further gains in speed of computation. In this context, Forster and Webb (2005) investigate approximations which avoid the necessity for using Monte Carlo computations. The extra error introduced by using such approximations is negligible.

Acknowledgement

This work is based on the SARs provided through the Centre for Census and Survey Research of the University of Manchester with the support of ESRC and JISC

References

Benedetti, R and Franconi, L. (1998). Statistical and technical solutions for controlled data dissemination. In *Pre-Proceedings of New Techniques and Technologies for Statistics, Volume 1* 225–232. Sorrento, Italy.

Bethlehem, J. G., Keller, W. J. and Pannekoek, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association*, **85**, 38–45.

Dawid, A. P. and Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, **21**, 1272–1317.

Elamir, E. A. H. and Skinner, C. J. (2004). Record-level measures of disclosure risk for survey microdata. *S^3 RI Methodology Working Paper*, **M04/02**. Southampton Statistical Sciences Research Institute.

Fienberg, S. E. and Makov, U. E. (1998). Confidentiality, uniqueness and disclosure limitation for categorical data. *Journal of Official Statistics*, **14**, 385–397.

Forster, J. J. and Webb, E. L. (2005). Bayesian Disclosure Risk Assessment: Predicting Small Frequencies in Contingency Tables. *Working Paper*, available from

`http://www.maths.soton.ac.uk/staff/JJForster/paper.html`.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press, Oxford.

Madigan, D. and Raftery, A. E. (1995). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, **89**, 1535–1546.

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review* **63**, 215–232.

Office for National Statistics and University of Manchester. (2005). *Individual Sample of Anonymised Records (Licensed File) [computer file].* Office for National Statistics, Census Division, [original data producer(s)]. University of Manchester, Cathie Marsh Centre for Census and Survey Research [distributor].

Omori, Y. (1999). Measuring identification disclosure risk for categorical microdata by posterior population uniqueness. In *Proceedings of the International Conference on Statistical Data Protection SDP '98*, 59–76. Eurostat, Luxembourg.

Polettini, S. and Stander, J. (2004). A hierarchical Bayesian model approach to risk estimation in statistical disclosure limitation. In *Privacy in Statistical Databases*, J Domingo-Ferrer and V Torra (Eds), 247–261. Springer Lecture Notes in Computer Science, 3050, Berlin.

Rinott, Y. (2003). On models for statistical disclosure risk estimation. In *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*. Luxembourg.

Skinner, C. J. and Holmes, D. J. (1998). Estimating the re-identification risk per record in microdata. *Journal of Official Statistics*, **14**, 361–372.

A graphical framework to evaluate risk assessment and information loss at individual level

Giovanni Seri
Istat – Italian National Statistical Institute, 00184 Rome, Italy, seri@istat.it

Abstract: When dealing with statistical disclosure control (SDC) problems, two aspects have to be considered. Firstly, a rule based on a measure of the risk of disclosure has to be adopted in order to decide if a certain release of data is safe or unsafe. Secondly, protection methods have to be applied to reduce the risk of identification when the release of data is classified as unsafe. The performance of a protection method is usually measured in terms of 'risk of disclosure' and 'information loss'. In this work we present a graphical framework named 'confidentiality plot' for the evaluation of risk of disclosure at individual level. For some extent the tool can be jointly used to evaluate risk of disclosure and information loss.

1. Introduction

National Statistical Institutes (NSIs) facing the problem to release micro data file for research (MFR) usually adopt statistical methods to preserve confidentiality of data. We consider that breach of confidentiality is produced if a unit is re-identified and the value of some sensitive variables is disclosed. A measure of the re-identification risk is then needed in order to classify data as 'at risk' or 'safe'. When data are at risk some disclosure limitation methods have to be applied in order to reduce the level of re-identification risk under a pre-defined threshold assumed as acceptable.

The definition of a disclosure scenario is a first step towards the development of a strategy for producing a "safe" MFR. A scenario synthetically describes (i) which is the information potentially available to the intruder, and (ii) how the intruder would use such information to identify an individual: i.e. the intruder's attack means and strategy. We refer to the information available to the intruder as an *External Archive* containing direct identifier (name, id-number, etc.), and some other variables that are expected to be available also in the MFR. We assume that the intruder tries to match the information in the individual archive with that in the MFR (for instance through record-linkage). We refer to these matching variables as *key* or *identifying variables*.

Statistical limitation methods suggested in literature can be classified on the base of their impact on the data in two categories (Domingo-Ferrer and Torra, 2001): (i) methods based on data reduction; (ii) methods based on data perturbation.

Methods based on data reduction aim at increasing the number of individual in the population sharing the same or similar identifying characteristics presented by the investigated statistical unit in order to avoid presence of unique or rare recognizable individuals. Perturbation methods, on the contrary, achieve data protection from a twofold perspective: (i) if the data are modified, re-identification by means of record linkage or matching algorithms is made harder and uncertain; (ii) even when an intruder is able to re-identify a unit, he/she cannot be sure that the data disclosed are consistent with the original data. A different approach based on synthetic micro data is out of the scope of this work (see for example, Polettini 2003).

Summarising, safety of a record characterised by certain values of the identifying variables is generally evaluated by two aspects: (i) the number of individuals sharing similar identifying characteristics in the population and (ii) the difference between the data released and the original data. In this paper we present a graphical framework, we call 'confidentiality plot', where units are plotted with coordinates representing these two aspects. Rules to classify units as 'safe' or 'at risk' can be represented in the graph, easing classification of units 'at risk'. Confidentiality plot can be used to assess the performance of a disclosure limitation method on the base of the individual risk and, for some extent, of the level of information loss.

The connection between the "confidentiality plot" and the R-U confidentiality map (Duncan *et al.*, 2001) introduced to compare the performances of different disclosure limitation methods, is clear. The main difference is that a point in the R-U confidentiality map represents a disclosure limitation techniques (e.g. a given data release), whereas a point in the confidentiality plot represents individual data.

In Section 2 we outline the framework used for the risk assessment based on confidentiality plot. Empirical results on business perturbed micro data are presented in Section 3. Conclusions and future perspective are discussed in Section 4.

2. Confidentiality plot

We consider the micro data set to be protected as a matrix A with n rows representing units and m+s columns representing the m key variables (x_j, j=1,…,m) and the s confidential variables (c_r, r=1,…,s) respectively:

A=(X,C), where X={$x_{i,j}$, i=1,…,n; j=1,…,m} and C={$c_{i,r}$, i=1,…,n; r=1,…,s}. (1)

The matrix C usually is not involved in the risk assessment and will be ignored in the following. We can assume that the application of protection methods consists in replacing X with a different matrix Y={$y_{i,j}$, i=1,…,n; j=1,…,m}.

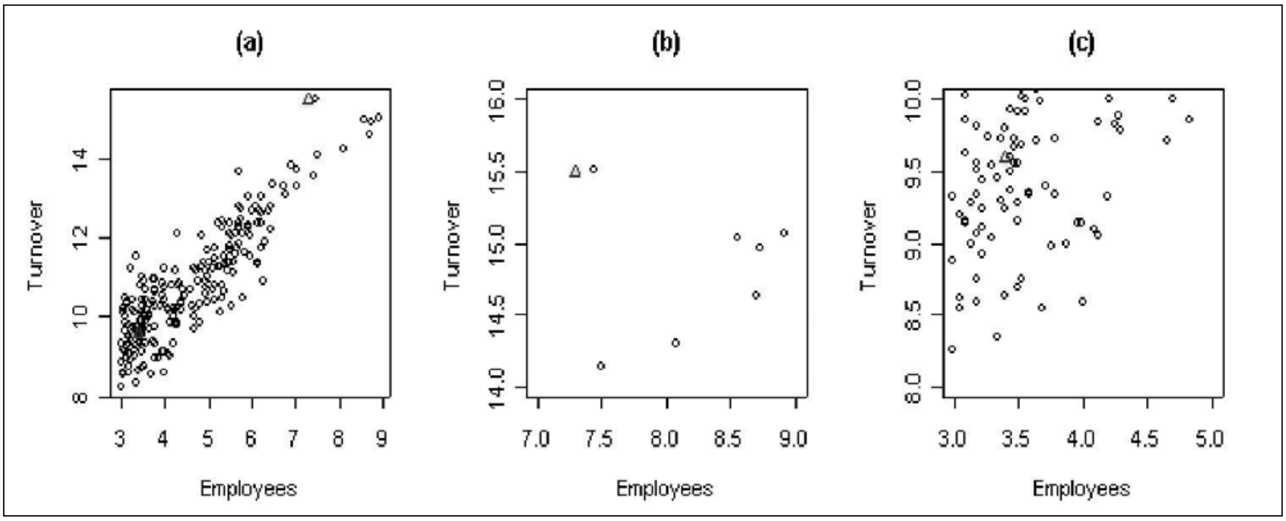
We firstly consider the case of social data in which identifying variables are mainly categorical. As a disclosure scenario we consider the external archive being a complete and reliable population register and the strategy being to link records presenting the same combination of key variables in both the external archive and in the MFR. Let y be the combination of identifying variables presented by a given record 'i' in the micro data file with f_y and F_y respectively the frequency of the same combination in the MFR and in the population. Under such a scenario the risk can be measured as $r_y=1/F_y$: a unit represented by a combination of values of some identifying variables is "at risk" if the same combination is "rare" in the population or, equivalently, the higher is F_y the lower is the risk associated to the combination y. For records at risk, methods such as 'global recoding' or 'local suppression' (Willenborg and de Waal, 2001) reduce data in a way that F_y is higher after that protection is applied. Similar reasoning on the risk measure can be made considering f_y instead of F_y or replacing F_y with a proper estimate if the true value of F_y is unknown (Franconi and Polettini, 2004).

As regards business micro data most of the information collected usually takes the form of quantitative variables with skew distributions. These variables are often representative of enterprise size and are extremely identifying. This means that, even though they are not always publicly available, quantitative variables have to be considered as key variables. The practical consequence of this is that all units are unique (rare) with respect to a small set of quantitative variables. Moreover, in many cases, populations of enterprises are sparse and firms are easily identifiable simply by their economic activity and geographical position. Even the knowledge about the survey design can be used to identify an enterprise, see Cox (1995). As a consequence, many of the protection techniques specifically proposed for business micro data aim at perturbing the original data in such a way that (i) re-identification by means of record linkage or matching algorithms is made harder and uncertain and (ii) even when an intruder is able to re-identify a firm, he/she cannot be sure that the data disclosed are consistent with the original data. Of course, this latter aspect has to be balanced with the need to make the information content of perturbed data as similar as possible to that of the original data in order to preserve the quality of statistical results.

We define the 'worst' disclosure scenario for a NSI assuming that the external archive available to the intruder coincides with the original file, X. It means that the intruder knows: (i) the target enterprise is included in the released file and (ii) there are no differences between the original data and the external archive due, for example, to classification errors. As data are perturbed, a strategy of attack based on exact record linkage will probably results in failure of matching. Therefore, we assume that the intruder will consider eligible links those records of the external archive which are similar to the target with respect to the set of key variables. This leads us to consider the concept of neighbourhood of a released record. For each record y in the released file, we denote as neighbour of y any unit x in the original sample (the external archive) that is "similar" to y. The level of protection ensured by each perturbed unit will depend on the number of neighbours we can attach to it.

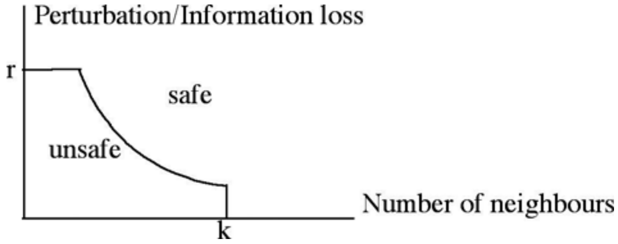
As an instance, we assume that the set of identifying variables consists in Turnover and Number of employees. In Figure 1 (a) two triangles - with coordinates (7.3,15.5) and (3.4,9.6) - representing two units treated with the same amount of perturbation are plotted against the original values (the external archive), logarithmic scale is used. In plot (b) and (c) the positions of those two perturbed units is zoomed in. Figure 1 (b) represents the first point as an outlier in the original data that is weakly perturbed. As the protected record (the triangle) is very close (similar) just to a single isolated point, an intruder trying to compare the released record with the data in his/her archive will have great confidence that the link between two such points is a true link. Figure 1 (c) shows the other protected record confused in a crowded cloud of points, and of course this makes it harder to identify the correct link because a high number of enterprises share similar values of turnover and number of employees.

Figure 1. Quantifying the extent of protection by the number of neighbours



We then argue that the amount of perturbation induced in the data and the number of neighbours, can be jointly exploited to assess the protection of a record. In particular, a graphical tool connecting the above mentioned aspects can be introduced. We denote by "confidentiality plot" a graph in which protected data can be represented with coordinates the "number of neighbours" (horizontal axis) and the "amount of perturbation" (vertical axis). A general scheme for this plot is presented in Figure 2.

Figure 2. A general scheme for confidentiality plot



The threshold "r" means that the released value is safe if it is distorted over the r% of the original value, whatever the risk of re-identification. On the other hand, the threshold "k" means that if the perturbed value is close to more than k units in the population, then no perturbation is required to protect this value. The curve represents the trade-off between these two aspects: the more the released value is confused in the population, the less is the required perturbation, and vice versa. The area under the curve is defined "unsafe" zone, because points in this area represent records that are not protected enough (at risk). The position of each point in the confidentiality plot with respect to the vertical axis can also be interpreted as an index of the quality of representation. In other words "information loss" and "perturbation" induced in a single record are equivalent labels for the vertical axis (see Section 2.1).

In the case of social data described above the confidentiality plot can result as a vertical line in k being the threshold of safety fixed for a given frequency F_y=k. If no perturbation methods are applied the vertical axis is not more representative of safety of records. Nevertheless, a proper measure of information loss at individual level can be defined and represented on the vertical axis of the confidentiality plot.

2.1. A way to measure information loss/perturbation and count neighbours

In this section we outline a proposal to measure the amount of perturbation (information loss) and a way to count the numbers of neighbour for each record.

We assume that 'perturbation' can be represented for each record by the difference between the original data and the corresponding perturbed data independently from the SDC method used. The smaller is the difference, the better the unit is represented in the MFR (the lower is the information loss). In other words, both information loss and perturbation can be suitably represented by a distance, e.g. the Euclidean distance, between the perturbed and the original record. The aim is to measure the error that is to be accepted by a user accessing the released data in place of the original data. Denoting by y the key variables for a generic record in the MFR, and by x the corresponding true values, we compute as the relative error:

Information loss=$||y-x||/||x||$. (2)

Clearly, the measure of information loss in (2) is also a measure of the perturbation induced in the data, as it represents the distance of the released value from the true.

In order to compute the number of neighbours of each record, a measure of similarity between units is needed (depending on the hypothesis about the intruder's strategy of attack) to compare released records y and record x. As the data are numerical, comparisons can be made on the base of a distance Z: $z=d(x,y)$. We define the comparison variable Z as the relative euclidean distance between a released record y and a record x in the external archive: $Z=||y-x||/||y||$; $\forall y\in Y$; $\forall x\in X$. All the possible pairs $(x,y)\in XxY$ can be classified: $(x,y)\in M$, pairs corresponding to correct links; $(x,y)\in U$, pairs corresponding to nonlinks. We then have: $XxY=M\cup U$.

The two distributions:

$m(z) = P(Z = z \mid (x,y) \in M)$ and $u(z) = P(Z = z \mid (x,y) \in U)$ (3)

are the basic ingredients of probabilistic record linkage and their estimation is the main issue in the record linkage literature. In our framework (based on the NSI point of view) the two archives Y and X are completely known and each pairs can be assigned to the set M or U without uncertainty. In other words, we are allowed to compute the two distributions in (3).

For each y to be released, the number of neighbours is computed as the number of original records $x\in X$ such that z is lower than a threshold δ (we remind that X is also the external archive available to the intruder):

Number of neighbours of y = $\#\{ x \in X : z < \delta\}$. (4)

We then denote as “neighbourhood” of y: N(y)={ $x \in X : z < \delta$}. Equivalently, we can consider the probability of the type 1 error, i.e. the probability of designating a pair as a link when it is not. The two densities of the distributions m(z) and u(z) in (3) can be estimated over the set of pairs $(x,y) \in M$ and $(x,y) \in U$ respectively. Assuming that lower distances are likely to be measured in the occurrence of a true link, we have: $\Pr(z<\delta|(x,y)\in U)=\alpha$, where α is the acceptable level for the type 1 error and δ, the critical distance, is fixed accordingly. In practice, the neighbourhood of y consists of all the units $x \in X$ that, for given α, are not rejected as nonlinks according to the probabilistic record linkage procedure. Choosing a smaller α turns into reducing the number of neighbours of the released record. As an alternative, it is possible to use the Fellegi-Sunter (1969) approach to record linkage.

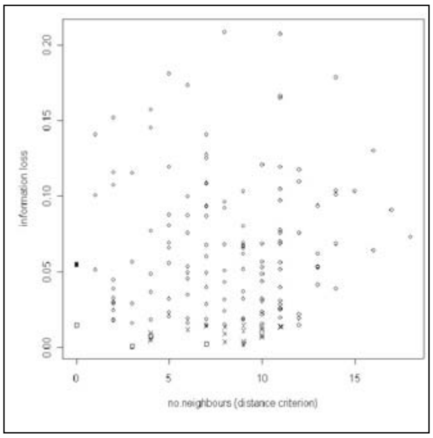
3. Experimental results

Data used in this work come from the Community Innovation Survey (CIS) and are treated with the statistical disclosure control techniques proposed in Polettini *et al.* (2002). We assess the level of protection guaranteed by the method using the above defined confidentiality plot on 157 enterprises belonging to the division 18 of the NACE nomenclature of economic activity We assume the disclosure scenario defined by an external archive equivalent to the original data file containing Turnover and Number of employees as key variables.

Figure 2 shows the confidentiality plot when α=0.05. For each record y in the micro data file under investigation, a point is plotted on the graph with coordinates: the number of neighbours (horizontal axis) and the information loss/perturbation (vertical axis). Squares identify 5 cases for which the nearest-neighbour is the correct link, that is when for a given record y the pair $(x,y)\in M$ and x is the nearest neighbour of y. Crosses identify 33 cases for which the correct link is in the neighbourhood, that is when for a given record y the pair $(x,y)\in M$ and $x\in N(y)$. A filled square highlights the unit presenting the highest Turnover in the original data, which, in many cases, is the most easily re-identifiable unit.

The two lines in the plot represent two different hypothesis of confidentiality policy (no real situation is taken into account). The curves consist of combinations of values of ‘perturbation’ and ‘number of neighbours’ joining the two points representing approximately the following rules for safety of a record: (i) a 15% of difference (on the logarithmic scale) between the original data and the corresponding perturbed record; (ii) a number of neighbours of 15 and null level of perturbation. The straight line defines stronger rules to preserve confidentiality. Anyway, the plot highlights the presence of outliers (enterprises relatively too large and easily identifiable) and the need for higher protection of the records in the unsafe zone of the plot. Nevertheless, results should be interpreted in the light of the severe disclosure scenario we assumed. It is worth noting that most (or all) of the crosses fall down in the unsafe area. For a few units the neighbourhood is empty (points with coordinate Number of neighbours=0). In this case the perturbation applied to records y is higher than the critical distance defining the neighbourhood (δ=0.040838).

Figure 3. Confidentiality plot for perturbed micro data: α=0.05



4. Conclusions

Assessing the performance of a disclosure limitation method is a difficult task particularly for business micro data. We have outlined a way to assess graphically by the so called 'confidentiality plot' the level of protection guaranteed by SDC methods at record level. We also introduce the framework for joint evaluation of the disclosure risk and information loss at record level. We argue that in order to assess the disclosure risk of an individual record in a MFR two aspects have to be taken into account: (i) the number of records in the external archive that share the same or similar identifying characteristics of the investigated record; (ii) the difference between the original data and the protected data. These two aspects are assumed as coordinates of each record represented on the confidentiality plot.

Empirical results are based on a set of enterprises from the Italian sample of the CIS survey protected by the method presented in Polettini *et al*. (2002). The purpose of this work is mainly to present the confidentiality plot as a mean to assess re-identification risk and information loss at record level. We think that the framework can be adopted in many situation even for tabular data. At this purpose, further studies are needed considering as an example: (i) protection methods that results in predictive intervals for numerical variables or (ii) in suppression of cell values in a table (when a feasibility interval for the true value can be evaluated). In both this cases y can be assumed as the interval midpoint. That is, from the intruder's point of view, y is the estimate of x that minimizes the error in (2). Moreover, different intruder's strategies of attack can (or have to) be taken into account. For example: different sets of key variables; different measures of the similarity between units; linkage based on the comparison of ranks for the biggest and therefore more easily identifiable enterprises instead of the nearest neighbour. We will develop these aspects elsewhere.

Acknowledgements

The author would like to thank Luisa Franconi and Silvia Polettini. The views expressed are those of the author and may not reflect the policies of Istat.

References

Cox, L.H. (1995). *Protecting confidentiality in business surveys*. In Business Survey Methods, Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J. and Kott, P.S. (Eds.), New-York: Wiley, 443-476.

Domingo-Ferrer, J., & Torra, V., (2001). *Disclosure Control Methods and Information Loss for Microdata*. In Confidentiality, Disclosure and Data Access, eds. P. Doyle, J. Lane, J. Theeuwes and L. Zayatz. Amsterdam: North-Holland, 91-110.

Duncan, G.T., Fienberg, S.E., Krishnan, R., Padman, R. & Roherig, S.F. (2001). *Disclosure risk vs. data utility: the R-U confidentiality map*. In Confidentiality, Disclosure and Data Access, eds. P. Doyle, J. Lane, J. Theeuwes and L. Zayatz. Amsterdam: North-Holland, 135-166.

Franconi, L. & Polettini, S. (2004). *Individual Risk Estimation in μ-Argus: A Review*. In J. Domingo-Ferrer and V. Torra (Eds), Privacy in Statistical Database. *Springer-Verlag*, Berlin Heidelberg 2004.

Fellegi, I.P. and Sunter, A.B., (1969). *A theory for record linkage*. Journal of the American Statistical Association, 64, (1969), 1183-1210.

Polettini, S. Franconi, L. and Stander, J. (2002). *Model Based Disclosure Protection.* In Domingo-Ferrer, J. (Ed.), Inference Control in Statistical Databases: From Theory to Practice. Lecture Notes in Computer Science, 2316, Berlin: Springer-Verlag, 83-96.

Polettini, S. (2003). *Maximum Entropy Simulation for Microdata Protection*. Statist. Comput., **13**, 307-320.

Willenborg, L. and de Waal, T. (2001). Elements of statistical disclosure control. Lecture Notes in Statistics, 115, New York: Springer-Verlag.

A 'Microdata for Research' sample from a New Zealand census

Mike Camden
Statistical Methods, Statistics New Zealand, PO Box 2922,
Wellington, New Zealand, mike.camden@stats.govt.nz.

Abstract: Statistics New Zealand has provided researcher access to many unit record datasets since 1995 in its three internal data laboratories. It began a programme to produce licensed Confidentialised Unit Record Files (CURFs) from social surveys and censuses in 2004, and it is currently investigating remote access, synthetic datasets and other ways for enabling access to microdata for research.

We will focus on our new 2% census sample CURF, which has just been pilot-tested by a set of New Zealand researchers. We will assume that cell count has a hypergeometric distribution, and use this to quantify sampling error, (a measure of both information loss and usability), justify the sample size and assess disclosure risk from the sample. Our sampling method preserves almost exact census proportions for a few census variables, and decreases sampling error for others. We will outline this method and its effects on cell counts and proportions.

We will summarise Statistics New Zealand's future plans for meeting the growing demand for microdata for research, in a country of four million people.

1. Introduction

In September 2005, Statistics New Zealand released a 2% sample from its 2001 Census to a set of ten researchers. The researchers have been asked to comment on the two core issues in microdata for research: usability and disclosure risk. A few of the researchers were specifically asked to test the disclosure risk by behaving as intruders. This pilot is part of a larger program of licenced Confidentialised Unit Record Files (CURFs).

The 2001 New Zealand census dataset contains 3.8 M records. This population size raises disclosure risk issues of uniqueness and confidentiality. It also raises usability issues for a small sample from it. To help us with decisions involving these issues, we considered cell counts (or proportions) in tables, and used the hypergeometric distribution to predict the behaviour of these counts. This helped us to calculate sampling errors, and hence to decide the sample size. In fact, the sample selection method produces sampling errors, for some tables, that are smaller than the hypergeometric ones. We are in the process of analysing how well the model fits the real behaviour of the sample.

The population is diverse in ethnicity, origin, income and employment. This raises the question of how to lessen uniqueness and minimise damage to usability. Considerable effort has been put into the concepts and practice of confidentiality for census data, including the application of appropriate rules for micro- and aggregate data (Statistics New Zealand, 2005).

An examination of the issues (Dunlop, 2004) recommended that we should proceed with extreme caution to produce a pilot CURF.

We will use the researchers' feedback in deciding the future for licenced CURFs from censuses. It will help us decide which variables to include, what sample sizes to use, what sampling methods to use and which of our five-yearly censuses to use.

2. The provision of microdata for research in New Zealand

In 2004, the New Zealand government completed a review of the official statistics system. This gave Statistics New Zealand a leadership role in collection and storage of datasets, and dissemination of information from them. The Annual Report 2004 (Statistics New Zealand, 2004) states: 'users will be able to use a variety of standard methods' to access information.

The Statistics Act (New Zealand, 1975) obliges Statistics New Zealand to protect confidential information from people and businesses, hold it securely, limit use to statistical purposes and prevent disclosure of identifiable information. The need to preserve respondent trust implies similar obligations.

In response to these two needs, Statistics New Zealand currently is actively extending its provision of microdata for research. Three methods of provision are detailed below. In each, it seeks to provide access to microdata, within its legal and contractual constraints.

2.1. The Data Laboratory

Researchers (academic, government or private) make an application, stating data needs, proposed outcomes and how the research will contribute to the improvement of official statistics. If the application is approved, Statistics New Zealand prepares a customised dataset with identifiers removed, arranges access in the data laboratory room, and checks all output. The three data laboratory rooms are in the Statistics New Zealand offices in Auckland, Wellington and Christchurch. This process has been running since 1995, with census datasets often being used. Sensitive variables, like geographic and household ones, are supplied where the need is clear and the disclosure risk allows.

Government departments can apply to access data in their offices if they can demonstrate a secure environment similar to that of the data laboratories. Statistics New Zealand audits these secure environments regularly.

2.2. The Remote Access system

This was trialled in 2004, on a modified version of the New Zealand Income Survey dataset, and decisions are pending. A modified version of a dataset is prepared, and kept secure. Researchers send in SAS code and receive outputs. Both are checked automatically, and are audited by our staff. Software written by the Australian Bureau of Statistics is used. The trial was with a social survey dataset, but census datasets could be used.

2.3. The CURF programme

The programme began in 2004. CURFs are issued to researchers, who sign a licence agreement. Datasets are modified carefully, and sent out on a CD. So far, we have issued CURFs for the New Zealand Income Survey 2002 and 2003. We are working towards issuing a joint CURF for the Income Survey and Household Labour Force Survey 2004. The samples for these CURFs contain about 28,000 records each. CURFs for further socio-economic surveys are in preparation or planning. The CURFs do not contain sample design variables, like stratum, and hence we attach datasets with 100 replicate weight variables.

3. Size and contents of this census CURF

The pilot CURF has 33 variables, of the approximately 100 output variables available, and 76,415 (2%) of the 3,820,749 records available. All the CURF variables are categorical, and most of them have categories collapsed from their original versions. The variables are about demographics, residence, ethnicity, origin, income and employment. Variables dealing with geographic location and household structure are omitted from the CURF.

The two drivers of CURF design are the needs to maximise usability and minimise disclosure risk.

4. Selection of the 2% sample size

Several issues influenced our choice of sample size. They are outlined below. Given these considerations, for New Zealand 1% is too small and 3% is too large. The sampling method means that a whole-number percentage is more convenient. So the conclusion, for this pilot, is clear!

4.1. Overseas practice

Many other countries produce the equivalent of CURFs, some licensed and some more freely available. We considered several countries, all with populations much larger than ours, and observed small sample proportions: 1% to 5%. We decided to be conservative, and like them, to aim for a small proportion.

4.2. The relationship of disclosure risk to sample size

Most types of disclosure risk depend on the number of records. It is reasonable then to assume that much disclosure risk increases linearly with sample proportion.

4.3. The relationship of usability to sample size

Many types of output have a sampling error which decreases with the square root of sample size. These types include cell counts and proportions, and regression coefficients (this assumes simple random sampling (SRS)). Sampling error is a measure of one form of information loss, and an inverse measure of usability. As sample proportion increases, usability increases, but with a square root law of diminishing returned.

4.4. Existing sample surveys and CURFs

We already have two large ongoing surveys, with nearly 1% of the population in each. The first is the Income Survey/Household Labour Force Survey, and CURFS are being produced from this. The second is the more recent Survey of Family, Income & Employment (SoFIE). Both contain much more socio-economic data per person or household than the census. So we needed a sample bigger than these.

5. Disclosure risk and uniqueness

5.1. Types of disclosure risk

The CURF will go to licensed researchers, some of whom may be research students. It should not, but could, fall into the hands of other persons. We'll assume that some of these recipients may behave as 'intruders'.

These events are possible:

- A researcher spontaneously recognises someone who is unique in the CURF and in the population, on a small number (3-5) of identifying demographic variables.
- An intruder hunts for a person or type, on a larger number of demographic variables.
- A researcher or intruder finds what they think is their own (or a neighbour's) record, because it appears to be unique on all or most of the 33 variables, and these values match their own values.

- A rogue researcher or intruder links this dataset with another one using software.

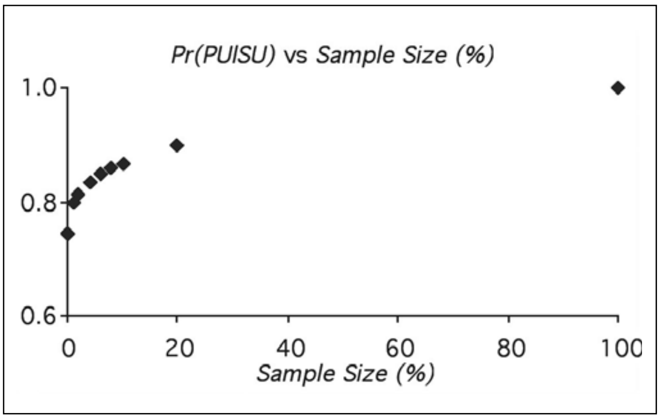
The risk from any of these events increases with sample size, as well as with level of detail in the variables.

5.2. Quantifying uniqueness

We carried out all the processing of variables on the entire census dataset, and drew the sample at the end. Hence we are able to look for population and sample uniqueness, using our 33 variables. Children (22.2%) and visitors from overseas (2.2%) have many structural missing values, so we will examine adult New Zealand residents.

Using these 2.9 M people, and using all our 33 variables, a high percentage of us (74.4%) are population uniques. For the 2% sample, the conditional probability that a person is a population unique, given that the person is a sample unique, is 81.3%. The level of uniqueness, and the size and behaviour of the conditional probability, are further reasons to keep the sample proportion small.

Figure 1. The conditional probability Pr(Population Unique | Sample Unique) rises with sample size.



6. Methods for limiting risk, and their impacts on usability

To lessen all the types of risk listed above, we used the methods below. There is further protection in the licence agreement that limits use and distribution.

6.1. Omission of household and location variables

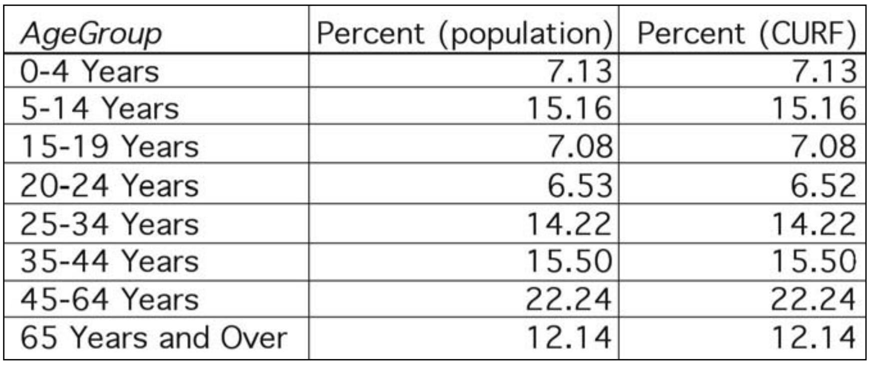
We omitted all variables dealing with family and household structure, and with geographic location. Both these sets of variables are very useful to researchers, and their non-selection is an important form of information loss. Responses to the pilot may suggest that we provide a location variable in future CURFs. We may need to balance this with further collapsing or omissions.

6.2. Collapsing of categories

We aimed to minimise risk and preserve usability. For each variable included, we examined the univariate distribution. Where categories had about 1% or less of the dataset, they were combined with others. When possible, new categorisations followed existing classifications. Most variables have similar proportions in each category. We aimed to use categories that would be useful to researchers. For example, people who cycled to work on Census Day 2001 (1.10%) were put with others to form “bicycle, walked or jogged” (3.60%). Careful design of the classification is a way of both limiting uniqueness and minimising the loss of usable information.

uniqueness and minimising the loss of usable information.

Table 1. In the original dataset, age is in years. The CURF variable *AgeGroup* has eight categories. They match life stages, and are all above 5%. This categorisation aims to preserve useful detail, and remove other detail. The population and CURF percentages are almost the same, due to the sampling method.



AgeGroup	Percent (population)	Percent (CURF)
0-4 Years	7.13	7.13
5-14 Years	15.16	15.16
15-19 Years	7.08	7.08
20-24 Years	6.53	6.52
25-34 Years	14.22	14.22
35-44 Years	15.50	15.50
45-64 Years	22.24	22.24
65 Years and Over	12.14	12.14

We regard *IncomeGroup* as the target variable that an intruder might want to find for a person they have recognised. This has 8 categories, with the top one being $70,001 (€40,000)/year upwards and having 3.43% of the population. The categorisation substantially lessens the variable's usefulness to an intruder, while it minimally lessens the usefulness to a researcher.

6.3. Special Uniques analysis

This process applied the ideas of Elliot et al (2002). We selected a subset of 14 variables that were considered very identifying: ie likely to be known about a neighbour or colleague. One of the variables we used was the five binary ethnicity variables combined into one variable with 32 categories. We took *Sex*, *AgeGroup* and every combination of three of the remaining 12 variables, and marked the records that showed up as uniques.

This process adds a variable (number of occurrences as a unique), and also shows which variables produce them. We decided to treat the 15,000 records (of the 3.8 M) that had two or more occurrences. We set up rules for modifying the value of the "worst" variable for each of these, ran seven iterations of this process and reduced the number of these 'special unique' records from 15,000 to 2,800.

This process reduces risk, but changes a tiny proportion (0.014%) of the values in the dataset. These values are replaced by neutral categories (like NEI), and not by wrong values. The process therefore produces minimal information loss.

7. The Sampling Method and its Consequences

Our sampling method was controlled on three variables. New Zealand is divided into 1,860 'Area Units', which contain on average 2,100 people, but vary widely in size. We added a random-number variable, and sorted the census dataset by *Sex*, *AgeGroup*, *AreaUnit* and the random-number variable. We divided this sorted dataset into groups of 100 records, and sampled two records from each group. *AreaUnit* is not included in the CURF.

The sort on the three named variables divides the census dataset into about 29,000 cells, with an average of about 130 people in each. Some cells are much smaller. The cells are homogeneous on at least the three named variables. Neighbouring cells are usually similar, and hence most of the groups of 100 records are homogeneous. We plan to investigate the effect of small cells on the value of this sampling method.

We can distinguish three types of variable: C: Controlled: *Sex*, *AgeGroup* (and *AreaUnit*); D: Dependent on these Controlled variables: ranging from highly dependent to slightly dependent; I: Independent of these Controlled variables. There are probably no completely independent variables, but they would form a worst case, and hence their properties need to be examined.

This distinction has some use, as it affects sampling error and hence information loss and usability. A sampling method that limits sampling error for some variables is of value, as it increases usability.

For the Type C variables, cell counts are extremely close to 2% of the population cell count; they are about ± 1 person away. The sample is an extremely close image of the population, by sex, age and geographical location. Other types, and combinations of them, are discussed below.

8. The distribution of cell counts under independence

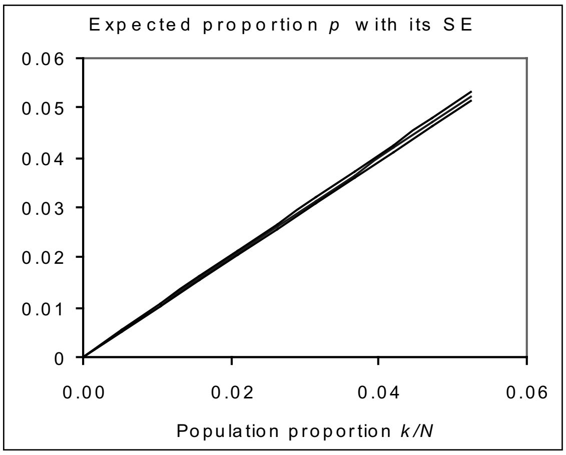
We assume that some researchers will make frequency tables using one, two or more of the variables. Each cell in these tables will have a population count k, which remains unknown to CURF users. It will have a sample count x from the CURF, and a sample proportion p. We assume here that some variables are independent from the three controlled variables. This gives us a 'worst case'; other cases will usually have less variation.

If we assume that the CURF behaves like a simple random sample (without replacement) of *n* people from *N* people, then *x* is the number of people who are both in this cell for the population and in the sample. We will treat *x* as having a hypergeometric distribution, with parameters (*N, k, n*). (In fact, if a table has *c* cells, then only *c*-1 of the *x*-values can be independent, but *c* is large for most tables.)

There are two convenient approximations. If *n/N* is small, then *x* will be approximately binomial, with parameters (*n, k/N*). If *k/N* is small too, then *x* will be approximately Poisson, with parameter *nk/N*. These give simple expressions for the standard deviation of sampling error, for counts ($\sigma x = \surd(kn/N)$ and proportions ($\sigma p = \surd(k/(nN))$). In fact, *n* is small (2%), and *k/N*) is small for most cells of interest to researchers.

All three models give the law of diminishing returns: sampling error decreases with the square root of sample size.

Figure 2. The graph relates size of sampling error for p, for a 2% sample, to the quantity being estimated. If a cell has a population proportion k/N, we locate this value on the x-axis, and move upwards to see its CURF proportion and sampling error (shown as ±1 standard deviation). Even in this this 'worst case', the 2% sample gives relatively small errors. If k/N is say 5%, then the error is ±0.08%.

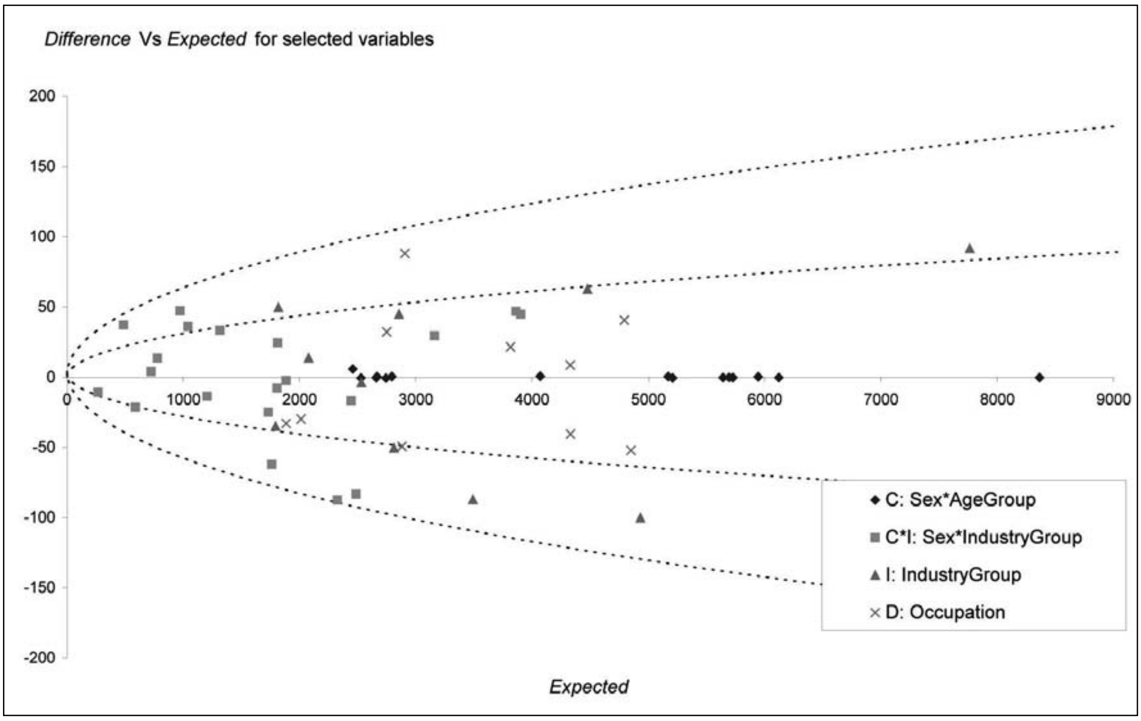


9. Sampling error and the controlled variables

With three types of variables (C, D, I), there are 7 (= 2^3 -1) types of combinations. The sampling error for D and combinations of two or more types will usually be between the best case (C) and the SRS case (I).

For Type D variables, and for combinations of C and I variables, standard error will usually be smaller than for SRS. Unfortunately, for cells of practical usefulness, the improvement is small. The expected behaviour can be studied analytically, but we will instead graph examples from the CURF.

Figure 3. *Expected* is 2% of the population cell count (= *kn/N*), and *Difference* is CURF cell count minus *Expected*. The curves show ±1 and ±2 standard errors, assuming independence and hypergeometric behaviour. The relationship is shown for the cells of four tables that exemplify types C, I, D and C with I. We used *IndustryGroup* as an example of Type I, as it is weakly related to Type C variables. The type C table shows very small values for *Difference*. The other tables have behaviour consistent with, or slightly less variable than, the hypergeometric distribution.



10. Acknowledgements

I wish to acknowledge the work and insights of these Statistic New Zealand Colleagues: Lisa Corscadden, Jamie Pou, Tiri Sullivan, Zoe Wood, Jamas Enright and Jason O'Sullivan (all, at the time, in Statistical Methods), and Adele Quinn, Adam Bedford and Peter Lafferty (Population and Census).

11. Conclusions

Statistics New Zealand already makes information from the 2001 Census freely available via its website, in the form of tables, with counts random rounded to base three. Tables can contain geographic and household variables. This pilot CURF enables a constrained group of people to produce a range of tables that is more limited in some ways and wider in other ways. Counts in these tables, if weighed up, would resemble counts random rounded to base 50. These researchers could perform any other analyses applicable to categorical variables. The CURF is a new form of access, with its own limitations and advantages.

We will use the feedback from the researchers who have trialled the pilot to consider new balances between disclosure control and usability, in possible future census CURFs.

References

Dunlop, A. (2004). "Census CURFs in New Zealand: an examination of the issues". Internal paper. Statistics New Zealand.

Elliot, M.J., Manning, A.M. & Ford, R.W. (2002). "A computational algorithm for handling the special uniques problem", *International Journal of Uncertain.ty, Fuzziness and Knowledge-Based Systems*, 5:10, 493-509.

New Zealand. (1975). Statistics Act 1975

Statistics New Zealand (2004). *Annual Report of the Government Statistician for the year ended 30 June 2004*, Statistics New Zealand, Wellington.

Statistics New Zealand (2005). "2006 census confidentiality rules", http://www.sats.govt.nz/census/2006-census/methodology-papers/confidentiality-rules.htm (3 Oct 2005)

Experience of using a Post Randomisation Method at the Office for National Statistics

Christine Bycroft and Katherine Merrett [1]

Statistical Disclosure Control Centre, Methodology Directorate, Office for National Statistics, UK, email Christine. Bycroft@ons.gsi.gov.uk or Katherine. Merrett@ons.gsi.gov.uk

Abstract: We describe an application of the Post Randomisation Method (PRAM) for disclosure control of categorical microdata. PRAM has been used as one of the protection methods for the 2001 Census Individual SARs file. In contrast to the standard application, only a small proportion of the whole file has been perturbed, with PRAM used where further recoding would lead to a high loss of information. The standard method has been adapted to better preserve multivariate distributions and minimise edit failures.

1. Introduction

Microdata files from sample surveys or extracts from Censuses are of great analytical value to researchers. When surveys and Censuses are undertaken confidentiality guarantees are given to respondents usually saying that information that could lead to their identification will not be released. When a statistical office is considering whether to release and how to release these microdata files they must consider the risks of possible disclosure of confidential information. Once a disclosure risk assessment has been conducted, some protection methods can then be applied to the data. Typically at the Office for National Statistics (ONS) the main disclosure control method used is recoding. However there comes a point where further recoding causes a large decrease in the information released for little decrease in disclosure risk. When this point is reached the only way to protect the remaining high risk records will be to remove them, or to alter one or more of their characteristics. Our preference was for a method that would perturb values of high risk records in a manner that has a small impact on analysis outcomes.

As nearly all of the variables on social surveys are categorical, we have developed a method based on the Post Randomisation Method (PRAM). The Invariant PRAM method especially seemed attractive to us as it conserves the expected values of the frequencies for each category after the perturbation. The PRAM method has been applied to the 2001 Individual Sample of Anonymised Records (SAR) drawn from the Census. In this context we have adapted the Invariant PRAM in the following three ways:

- We were interested in preserving relationships between variables and developed an adaptation of the method which enables some control of the joint distributions between the variables perturbed by the PRAM and other variables in the microdata.
- Our implementation of the method conserves the exact frequencies of the categories after the perturbation and not just the expected values.
- In contrast to the typical PRAM described in Willenborg and De Waal (2001), we are perturbing only the high risk records within the sample, not the whole sample.

In section 2 of the paper we provide a description of the PRAM method as introduced by Kooiman et. al. (1997) and the way in which ONS has adapted the methodology to suit its needs. Section 3 of the paper describes some of the methods that were used to examine the effects that the perturbation has had on data quality and section 4 concludes.

[1] The PRAM methodology developed by the ONS for the use on the 2001 Individual SAR was joint work between Bill Gross, Philippe Guiblin, Joshua Fox and Katherine Merrett. Thanks to Natalie Shlomo who provided helpful comments on an earlier draft.

2. Method: PRAM

PRAM is a disclosure control technique for microdata. It was introduced in 1997 by Kooiman et. al. as a disclosure control method to be applied to categorical data in microdata files. The values of a categorical variable for certain records in the microdata file are changed according to a prescribed probability. Each new value may or may not be different from the original value. For example, a person who is classified as a widow may be re-classified as single under PRAM.

The probability mechanism is described by an invertible transition matrix P per variable.
Let $P = (p_{ij})$ be an L x L matrix for a variable having L categories. The entries of the matrix are the conditional probabilities

p_{ij} = Pr(New_value = j | Old_value = i).

The resulting perturbed file is released along with information about the probability mechanism (transition matrix) used. The researcher can use this information to adjust his or her analysis regarding the perturbation caused by PRAM. The perturbation can be seen as a form of prior misclassification.

As explained in Willenborg and De Waal (2001, Section 5.5.1), PRAM offers protection by inflow and outflow: inflow from safe combinations of values to risky combinations, and outflow from risky combinations to safe combinations. The resulting perturbed file will retain some unusual or high risk combinations, but there will be uncertainty over whether these have been created through the perturbation process or are original values from a respondent.

A problem with PRAM as described above is the possibility of creating invalid or highly unusual combinations, e.g. a 14 year old doctor or 17 year old widow. This is partly a result of allowing inflow as part of the confidentiality protection. Also using the transition matrix in the analysis may be a burden to some researchers as standard statistical applications may be more difficult to implement.

2.1. Invariant PRAM

A specific form of PRAM introduced by Kooiman et. al. (1997) is the invariant PRAM method. In this form, applying PRAM is invariant with respect to the frequencies of the variables. Let $P = (p_{ij})$ be the transition matrix for a variable ξ having L categories, and F be the vector of frequencies containing the sample counts of each category. The matrix P is chosen such that:

$$P^t F = F \tag{1}$$

As a result, frequencies after the perturbation are in expectation equal to the original frequencies of ξ. This relieves the user of the perturbed file from the extra effort of obtaining unbiased estimates of the original data. It is still important to release the transition matrices so that the user can compute the extra variance introduced by using invariant PRAM.

2.2. Adapting Invariant PRAM

The Individual SAR is a 3% sample of some 1.8 million individuals drawn from the 2001 Census. The adaptations to PRAM were motivated by the need to protect the 2001 Individual SAR. In this situation it was possible for us to identify high risk records because we had the full population data from which the sample was derived. Our approach to reducing disclosure risk was to use recoding of variable categories to a point where further recoding would seriously impact on data use without much increase in protection. The remaining subset of high risk records were protected using perturbation through use of our adapted PRAM. PRAM as implemented here has the advantage of being able to target modification of the file directly to high risk records and to the particular variables within each high risk record that contribute most to disclosure risk.

In this situation where we are only perturbing high risk records, protection against disclosure is achieved by largely removing any inflow and relying only on outflow. In contrast to invariant PRAM, the transition matrix P will need to ensure that the probability of changing values is maximised. Thus we need to minimize the probabilities that are on the diagonal of the transition matrix (i.e., the probability that no change occurs). Other constraints on the transition matrix are that it is invariant and that statistical properties of the dataset stay similar after the perturbation. Thus we want to perturb categories to other categories that are both feasible and will not result in highly unusual combinations.

In summary, the method developed for obtaining the transition matrix P ensures three goals:

1. the probabilities of no change are minimised
2. in expectation, the output distributions are the same as the input distributions
3. transition to "similar values" are maximised

To obtain the transition matrix P we used the linear programming feature of SAS. The routine minimises an objective function, subject to constraints. The objective function is defined as follows:

$$\sum_{i} w_{ii} p_{ii} + \sum_{i \neq j} w_{ij} p_{ij} \tag{2}$$

where $W = (w_{ij})$ is a Weight Matrix: a low weight for a preferred transition and a high weight for a non-preferred transition. The Weight Matrix is set up to avoid extreme transitions. Rather than having extreme changes that might create highly unusual individuals or invalid combinations, we prefer to keep the values as they are. Given the Weight Matrix, the optimisation routine finds p_{ij} values that minimise the objective function. This will lead to minimum probabilities on the diagonals, subject to also avoiding extreme transitions.

The constraints of the optimization routine are the following:

- the rows of the transition matrix sum up to 1;
- all the probabilities are positive;
- expected output frequencies are equal to input frequencies;

The constraints are mathematically expressed as:

$$\begin{cases} \sum_{j=1}^{L} p_{ij} = 1 \quad with \;\; i = 1,...,L \\ p_{ij} \geq 0 \quad \forall \quad (i,j) \\ \sum_{j=1}^{L} p_{ij} F_i = F_i \quad with \quad i = 1,...,L \end{cases} \tag{3}$$

2.3. Preserving Univariate Distributions

In applications of the invariant PRAM, the movement of a record from category *i* to category *j* is applied in a way that ensures that the expected values of the frequencies of the categories will be preserved.

At the ONS a method was developed for obtaining the exact frequencies of the categories after the perturbation scheme and not just in the expected values. First, based on the transition matrix, we calculate how many records should be changed from category i to category j. Denote this number by r_{ij} . The records are then sorted randomly within category i and a sub-sample of size r_{ij} is drawn for each j. All records in the sub-sample have their category i changed to category j. The method is repeated for all categories, i = 1,..., L.

2.4. Stratification: A way to preserve multivariate distributions

The modified PRAM method preserves the univariate distributions as shown above. Preservation of multivariate distributions is also important to users.

The method we have developed to preserve the relationship between the prammed variable (e.g. age) and another variable (e.g. marital status) is to PRAM the variable within strata defined by the values of the second variable (e.g. PRAM age within each stratum ‘single’, ‘married’, etc...). This provides some control on the transitions. To make this process more effective, the second variable, if prammed, should be prammed also within strata defined by the first variable. We call the variable used for the definition of the strata a ‘control variable’ Strata can also be defined as a combination of several ‘control variables’. This ensures that the perturbations are constrained for the joint distributions of the prammed variables with the control variables.

A problem occurs when the subset of high risk records defined by a stratum becomes too small. PRAM only changes a category *i* to a category *j* value that appears within the stratum. Strata with too few values or records do not offer enough options for the transitions and may result in undesirable transitions. To avoid this we used broad categories to define some of the control variables and limited the number of control variables used.

2.5. Disclosure risks

As explained above protection is provided by both inflow and outflow. In the ONS implementation of PRAM there is considerable outflow. However little protection is provided in the file by inflow as only those records which are high risk are perturbed and perturbed values are controlled to avoid creating unusual and therefore potentially risky combinations.

To counteract this reduced protection, we are not providing the transition matrices to the user. This in turn limits the information available to the user about the specific perturbation mechanism that has been applied, and means that the user will be unable to calculate the additional variance introduced by PRAM. However only a small proportion of records in the file have been perturbed and we are providing the analyst with some information by flagging records which have been perturbed with the same flag as used for marking imputed records. The implication from an intruder point of view is that an intruder does not know whether a flagged record is a true value, a perturbed value or an imputed value.

2.6. Application to the Individual SAR

Stratified Invariant PRAM as described above was implemented on the 2001 Individual Licensed SAR from the Census. The file is available for researchers to download under a License agreement, and the protection measures reflect ONS assessment of disclosure risk under these conditions, (Gross 2004). We aimed to protect only against attempts at exact matching, so considered that perturbing the value of one variable in a high risk record provided sufficient protection.[2]

We used the results of a special uniques analysis, (Elliot 2004), to efficiently target the perturbation to the highest risk records and highest risk variables.

PRAM was applied to variables sequentially, beginning with the highest risk variable. Using 10 variables allowed us to perturb one variable on all the target high risk records. PRAM was least successful

[2] Because Scottish data has a different risk profile, more than one variable in a record was sometimes prammed.

for variables lowest in the sequence of application. Since only one variable was prammed for each record, relatively few records remained for the variables low down in the sequence. For these variables, no controls were used, and higher proportions remained unchanged.

After applying PRAM to the individual SAR a series of edits were run to ensure that no invalid or extreme combinations had been created.

3. Effect on Data Quality – Measuring the information loss

Investigations of data quality after PRAM found that:

- The univariate distributions for prammed variables remained virtually unchanged.
- The multivariate distributions between variables involved in the PRAM process (prammed and control variables) worked well too. Very little difference was observed in cells of prammed variables by control variables
- Very few perturbed values failed subsequent edits and these were adjusted manually.

The assessment of the damage, or information loss, on the distribution between variables involved and not involved in the PRAM process was measured by comparing the impact of pramming relative to the sampling error[3]. The added error due to PRAM was measured as the relative absolute difference between perturbed and unperturbed cell estimates. The ratio between the PRAM error and the relative sampling error was calculated for each cell. This provided some assessment of the additional variance due to PRAM.

We measured the ratio between the error due to PRAM and the sampling error for 2891 cells from 15 tables. Tables were created to reflect variable combinations that were important to data users. Table 1 shows the percentage of cells with a ratio of greater than 1 and greater than 2 by the size of the (unweighted) cell frequency. The majority (84%) of the cells in our test have a ratio of less than 1. The proportion of cells with a ratio of 2 or more was small at just 5% of the total cells. The effect of perturbation relative to sample error decreases as the cell size increases. Thus the damage done by PRAM is greater for cells with low frequencies.

Just over one third have a ratio of less than 0.1 (not shown), and for these cells the additional error due to PRAM is negligible.

Table 1. Percentage of Cells across all tables with a ratio of the error due to PRAM and the sampling error of greater than 1 and 2

	Cell Frequency Before PRAM								
	0-5	6-10	11-20	21-40	41-90	91-150	150-500	500+	**Total**
Percentage of cells with a ratio >1	35	25	24	13	15	10	17	10	16
Percentage of cells with a ratio >2	9	8	6	4	5	4	7	4	5

These are some basic tests that we conducted to look at the effect PRAM has on data quality. Further work should be carried out to look in more detail at the effect the PRAM has on typical analyses that users are likely to conduct.

[3] If we approximate the sampling process as simple random sampling with replacement, the relative sampling error for a cell is given by: $RSE(\bar{N}_c) = \frac{\sqrt{Var(\bar{N}_c)}}{E(\bar{N}_c)} = \sqrt{\frac{(1-p_c)}{np_c}}$ where $\bar{N}_c$ is the estimated population total of individuals in category *c*, *n* is the sample size (of the SAR) and p_c the probability that a population member falls in this category. In practice this can be estimated by replacing p_c with the observed proportion $\bar{p}_c$ of cases falling in category *c*.

4. Conclusions

The PRAM methodology described above was adapted from the original PRAM in order to meet ONS objectives and constraints when releasing microdata. The purpose of PRAM is to perturb data without damaging the statistical properties. The difficulty resides in finding the right balance between safety and damage.

We feel that the PRAM methodology worked well as one part of the overall confidentiality protection for the 2001 Individual SAR. Recoding allowed us to reduce the number of high risk records to a small percentage (4%). We were then able to target the perturbation only to those records which we considered to be of high risk. We were also able to preserve some of the most important multivariate distributions in the datasets through the use of stratification and preserve exactly univariate distributions of the variables being prammed..

Applying PRAM to a small proportion of the file has allowed us to strike a good balance between recoding and minimising the damage from perturbation.

References

Elliot, M.J and Manning, A (2004) The methodology used for the 2001 SARs Special Uniques Analysis, University of Manchester.

Gross, B, Guiblin, P and Merrett, K (2004) Risk Assessment of the Individual Sample of Anonymised Records (SAR) from the 2001 Census, Office for National Statistics. http://www.ccsr.ac.uk/sars/events/2004-09-30/slides/index.html

Kooiman, P., Willenborg, L.C.R.J., and Gouweleeuw, J.M. (1997). PRAM: a method for disclosure limitation of microdata, Research paper 9705, Voorburg/Heerlen: Statistics Netherlands.

Willenborg, L.C.R.J., and De Waal, T. (2001). Elements of Statistical Disclosure Control, New York: Springer.

Disclosure Risk Assessment through Record Linkage

Sam Hawala[*]*, Martha Stinson*[**]*, John Abowd*[**]

[*] **Statistical Research Division, U.S. Bureau of the Census, 4700 Silver Hill Road, Washington, DC 20233, USA, sam.hawala@census.gov**

[**] **Longitudinal Employer-Household Dynamics Project**

Abstract: We present an example of assessing the disclosure risk of files intended for public release (Public Use Files – (PUF)). These files contain synthetic data, created from a confidential linked longitudinal data file. We used automatic record linkage experiments to assess the risk of disclosure from the PUF. We matched the PUF one by one to the confidential data file from which they were originally constructed. The confidential longitudinal data file contains linked information, matching selected worker and employer records for statistical research.

1. Introduction

The data in question are individual-level data containing demographic information on individuals that were in the 1990-1996 panels of the Survey of Income and Program Participation (SIPP), linked to the Social Security Administration (SSA) administrative earning records data. The Longitudinal Employer-Household Dynamics (LEHD) research group performed the data linkage and the data cleaning for the confidential file, and is preparing the public use files (PUF) planned for release. The confidential data file represents the kind of data that would be compiled for analysis by a researcher working in a protected area at either the Census Bureau or the other federal agencies that supplied the data. The PUF will be released to the general public and will benefit communities of researchers interested in disability and retirement.

2. The Confidential File

The confidential file we worked with contains an attributed number (a person identifier) for each person and their spouse, if the individual is married. There are nine variables on the file that are copied exactly from the SIPP: sex, black/non-black, education (3 categories), marital status, age (3 intervals), marital status and the same variables for the spouse. There are also a host of additional SIPP variables that are subject to confidentiality protection using synthetic data methods. These include: birth date, hispanic/non-hispanic, education (5 categories), whether or not health limits the kind or the amount of work, number of children under 18, marital history, immigrant status, industry and occupation categories, total number of weeks worked in a year, annual total personal and family incomes, annual family total combined benefit dollars from government programs, total net worth, whether or not the individual is a homeowner, home equity, non-housing financial wealth, whether or not individual has a defined contribution or benefit pension plan, and a summary of the individual's annual health insurance status. Neither the confidential data file nor the proposed public use files contain Privacy Act protected identification information such as names, addresses, and social security numbers.

3. The Public Use Files

The PUF consist of several files. Each is a version ("implicate") of synthetic data constructed from the confidential data file, in the spirit of multiple imputation outlined by Rubin (1993). All SIPP variables from the confidential file are synthesized except for a few variables. For more details on the creation of synthetic data, see papers by Abowd and Woodcock (2001, 2004), Rubin (1993), Fienberg (1994), Fienberg, Makov, and Steele (1998), Kennickell (1991, 1997, 1998, 2000), Raghunathan, Reiter, and Rubin (2003), and Reiter (2003).

Synthetic data have the advantage of making re-identification of respondents difficult while still providing analytically valid microdata to researchers in a format that they are accustomed to using. The synthetic public use data files are being prepared to closely mimic the characteristics of the confidential file. They provide analytically useful data sets, while at the same time do not allow for re-identification of individuals in the already published SIPP public use files.

With the PUF, researchers at large, working in their own institutions, will have access to important demographic and economic information but some of the finer details of each person's record are synthesized to help preserve confidentiality. The disclosure avoidance standard is that individuals in the new PUF cannot be re-identified using the already published SIPP public use data products with a greater probability than a false re-identification. Linking of so much detailed administrative data to the SIPP necessitates this high standard of disclosure control.

In creating the synthetic data, LEHD's goal is to refrain from imposing prior beliefs about the relationships amongst variables and instead to allow the data themselves to determine the nature of these relationships. Thus, all variables can potentially be used as explanatory variables for the posterior predictive distributions of all other variables, even when such a relationship might not seem sensible to a social science researcher. In practice, due to feasibility issues, LEHD chooses some subset of variables to go on the right hand side of the predictive regressions but the goal remains to impose as few prior beliefs as possible. In this sense, the modelling done to create synthetic data is different than modelling done in order to predict future outcomes or to analyse cause and affect relationships.

Once the synthetic data are created, however, a different kind of analysis becomes necessary, where prior beliefs become important. Standard economic and demographic models must be tested using the synthetic data and analysts with experience evaluating such results must determine whether the synthetic data are statistically valid. Rubin (1996, p. 474) outlined what is meant by statistical validity:

- First and foremost, for statistical validity for scientific estimands, point estimation must be approximately unbiased for the scientific estimands averaging over the sampling and posited nonresponse mechanisms.
- Second, interval estimation and hypothesis testing must be valid in the sense that nominal levels describe operating characteristics over sampling and posited nonresponse mechanisms.

This definition should be modified to include the phrase “confidentiality protection mechanisms” wherever “nonresponse mechanisms” appears.

Thus in order to assess the quality and usefulness of the synthetic data, LEHD looks at several statistics of interest, calculates these statistics and averages them over the implicates of synthetic data, and then compares them to the best estimate of the same statistics from the confidential data. The estimates must be unbiased and the variances of the estimates must be such that inferences drawn about the estimates are similar to the inferences from the confidential data.

4. The Record Linkage Experiments

We performed the record linkage exercises using preliminary versions of the PUF. The LEHD research group continues to implement improvements for a final version. We matched each implicate PUF to the confidential file. The confidential file played the role of the already published SIPP public use data sets. We obtained similar results for each implicate, so we report on the results obtained for only one of the implicates. We used a matching program based on the standard Census Bureau record linkage software written by Winkler et al. This standard software relies on the frequentist approach taken by Fellegi-Sunter (1969) to the probabilistic model of record linkage. It is used throughout the Bureau to create survey frames, to combine files, or to remove duplicates from files. Background on matching and some of the methods available in the software are described in research reports rr93/08, rr93/12, rr94/05, and rr99/04 at http://www.census.gov/srd/www/byyear.html.

The original purpose of the matching software we used was to extract plausible matching records, from a very large file A, that correspond to records in a smaller file B. The file B is assumed to fit into core memory. In our application, we treat the confidential file as file A, and one of the implicate files as file B. Here, files A and file B have the same size. Every person in the confidential file has one record in each of the implicate files. This fact further raises the bar for the disclosure testing. If file A were a large national file with many millions of records, matching to the smaller implicate files would be less successful. But the existence of the public use SIPP files for all the panels in the 1990s limits the size of file A to just under 250,000 people.

The software compares record pairs from the two files A and B when they agree on a specified blocking criterion. In order for the best matching pairs to be selected, the files must first be sorted according to this blocking criterion. The program outputs a file of records from the PUF that are plausible matches to records in the confidential file. The standard Census Bureau record linkage program features one-to-one matching that result in each record being paired with its most likely match within its blocking group. The matching program we used does not do this; rather, an output file may contain several records from the PUF that were scored as likely matches to the same record in the confidential file.

An input file to the matching software specifies the agreement criterion for each of the matching variables. From the agreement criterion, the software computes a score, or weight. For each record in B, the program determines the matching comparison weights with records in A that share the same values of the blocking variables. If any of the comparison weights exceeds a cut-off value, the A record is written out to an output file. Finally the common person identifier on both the A and B files is compared in order to determine whether the match is true or false. Thus our testing determines how many matches can be obtained by comparing the confidential and implicate file and what percentage of these matches are actually correct.

We report results on a matching run where we chose 5 blocking variables and 10 matching variables. The blocking variables we used were the variables that were not synthesized. They are: sex, black/non-black, education (3 categories), marital status, age (3 categories), plus these same variables for a spouse if one is present. The matching variables were all categorical except birth-date (month, year), which was converted to number of days since earliest birth-date on the confidential file. The other variables were hispanic/non-hispanic, education (5 categories), and immigration status, whether or not health limits the kind or the amount of work, whether or not the individual is a homeowner, number of children under 18, marital history, industry and occupation categories.

For all the categorical matching variables, we used the exact string comparison which assigns either the full agreement or disagreement weight based on whether the variable on the implicate file is the same or different from the variable on the confidential file. If the value of a variable is missing, the record will automatically be considered to agree on that variable. In practice this is unnecessary because the unsynthesized variables are never missing but in principle it ensures that we enable the most matches possible.

The conditional matching agreement probability is defined as the probability that the A and B values for this variable agree given that the records are truly a match. The conditional non-matching agreement probability is defined as the probability that the A and B values for this variable agree given that the records are not a match. These probabilities are used to calculate the weights given to this variable when it agrees or disagrees. The probabilistic record linkage model defined by Fellegi and Sunter (1969) assigns the weight $\log(\frac{m_k}{u_k})$ if the records agree on the k^{th} variable, m_k being the matching agreement probability and u_k being the non-matching agreement probability. It assigns the weight

$\log(\frac{1-m_k}{1-u_k})$ if the records disagree on the k^{th} variable . The software compares the values for each variable designated for matching, decides whether the values agree or not, and then assigns the appropriate weight to the variable based on the user supplied probabilities. Then a cumulative weight is calculated by summing the weights across all the variables designated for matching. This cumulative weight is the ultimate determiner of whether two records match. It is compared to the cut-off values provided by the user and if it passes the stated threshold, a match is declared. In our experiment, all pairs with positive cumulative agreement weights are considered matches. The relative matching and non-matching agreement probabilities chosen by the user control the influence of one variable relative to another on this cumulative weight. The non-matching agreement probability essentially tells how often a variable will agree at random across two files. A high value for this probability will reduce the importance of this variable in the matching by causing the agreement weight to be lower. This is desirable because if the variable is likely to agree at random, any match in values between the A and B files is less likely to signify a true match. At the same time, a high non-matching agreement probability causes the disagreement weight to be less negative or smaller, meaning that the penalty for not matching on this variable is not as high. In contrast, the relative matching agreement probability tells the importance of this variable compared to other variables in determining whether two records are a match. A high matching agreement probability means that a match on this variable is crucial to determining an overall match between 2 records. Thus a high value for m produces a high agreement weight. It also produces a more negative or higher disagreement weight, more severely penalizing non-matching in this variable. Blocking variables are essentially matching variables that have m = 1.

5. Discussion of Results

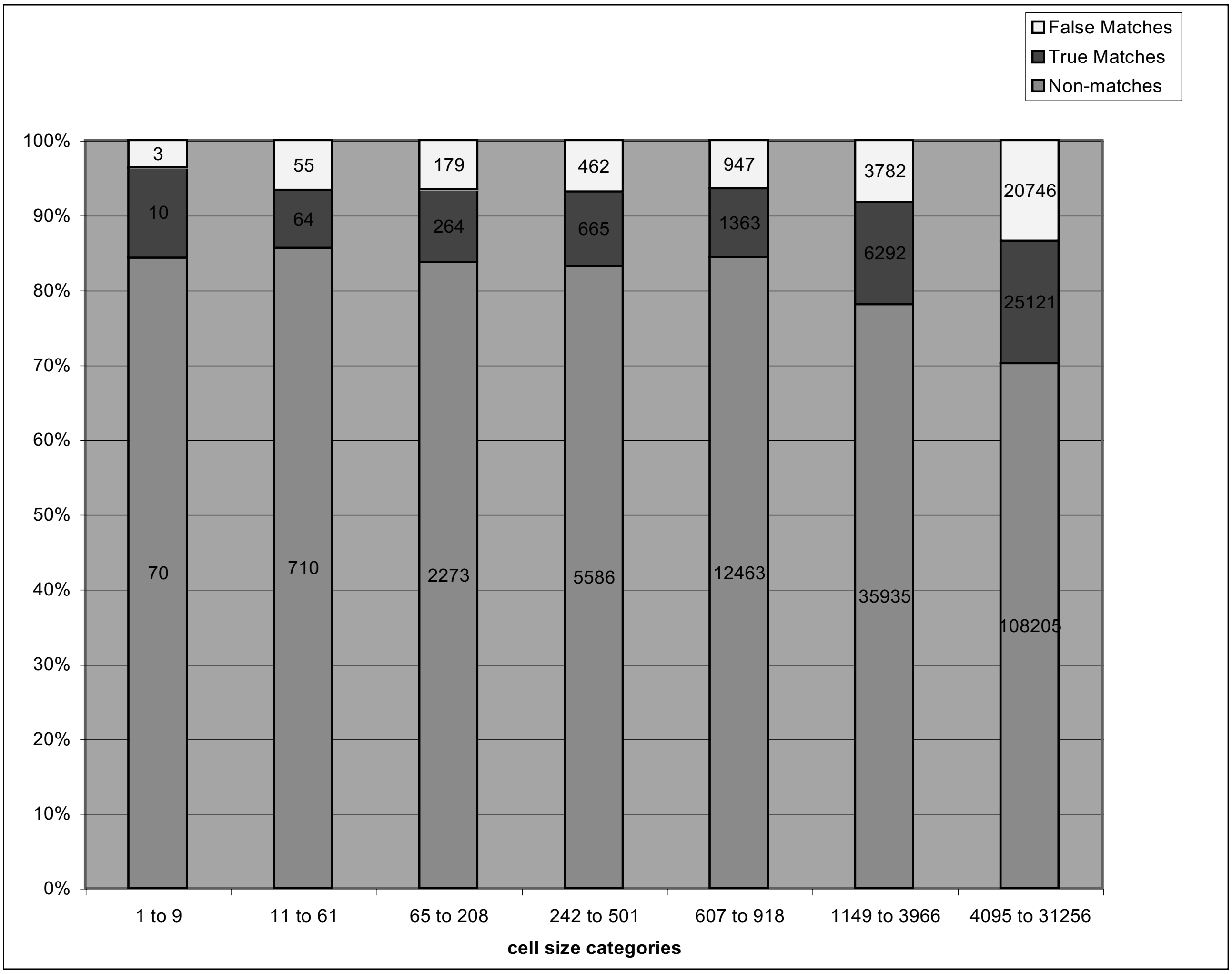
The 5 unsynthesized variables available for all individuals (there are nine for married individuals) create 136 unique combinations (cells). There are some cells that will present disclosure problems simply by virtue of the fact that the cell contains (cell size) only 1 or 2 individuals. Those cells where we could correctly match large numbers of records also represent disclosure problems. There were a total of 33,771 true matches and 26,174 false matches. The numbers of true and false matches vary considerably from one cell to another and do not appear to be tied to the cell size. What we consider informative is the ratio of number of true matches to false matches (tm/fm). When a cell has a tm/fm ratio much greater than 1 then the cell represents a disclosure problem. Indeed, an outside person doing the matching would obtain a total number of matches where a much higher percentage of them would be true matches. The outsider would be right much more often than they were wrong. When the ratio tm/fm is close to 1, the outsider would not be able to distinguish the true from the false matches by just guessing at random. So the ratio tm/fm was the most useful statistic for highlighting cells with problems.

Figure 1 summarizes the findings. We sum the number of non-matches, true matches, and false matches for groups of 20 cells and plot a bar showing the percentage of each type of record. The size cutoffs of the cells are listed on the x-axis. For example, the first bar aggregates across the first twenty cells, which range in size from 1 to 9. Of the 83 total records in this grouping of cells, almost 84 percent do not match across the confidential and the implicate files. Another 12 percent match correctly and 4 percent match incorrectly. Each bar in Chart 1 represents a grouping of 20 cells except for the last bar, which contains only 16. It is evident from Figure 1 that larger cells have both more true and false matches but they do not necessarily have lower tm/fm ratios. Although the smallest cells have a very high number of true matches relative to false matches (12% versus 4%, tm/fm ratio=3.33) and this same comparison is much better for the largest cells (16.3% true to 13.5% false matches, tm/fm ratio=1.21), the group with the closest number of true and false matches is the second bar (7.7% true matches versus 6.6% false matches, tm/fm ratio=1.16) whose cells range in size from 11 to 61. The third, fourth, fifth bars all have tm/fm ratios just under 1.5 and the sixth bar has a ratio just over 1.5. So there is no monotonic change in the tm/fm ratio as the cell sizes increase.

6. Conclusion

The results from the matching procedure performed on the Preliminary Public Use File give cause for some concern and some cautious optimism. There are many cells where the synthetic data are properly perturbing matches between the confidential and implicate files to the point where there are at least as many false matches as true matches. However there are also problematic cells where there are a disproportionate number of true matches. We are working on performing the best matching possible by choosing our conditional matching and non-matching agreement probabilities in an optimal manner. Any strategies employed to reduce disclosure risk will have to be measured against their effect on the analytic validity of the file. Hence at this point it is too early to make decisions about specific steps we will take to handle the problematic cells. We will repeat our matching procedure on the next version of the Preliminary Public Use File and re-evaluate how many cells have a ratio of true to false matches greater than one. At that point, different actions may be necessary to solve any remaining disclosure problems.

Figure 1. Percentage of Non-matches, False Matches, and True Matches



References

Abowd, John M. and Simon Woodcock, *Disclosure Limitation in Longitudinal Linked Data, in Confidentiality, Disclosure and Data Access*: Theory and Practical Applications for Statistical Agencies, P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (eds.), (Amsterdam: North Holland, 2001), 215-277.

Abowd, John M. and Simon Woodcock, *Multiply-Imputing Confidential Characteristics and File Links in Longitudinal Linked Data,* in J. Domingo-Ferrer and V. Torra (eds.) Privacy in Statistical Databases (New York: Springer-Verlag, 2004), pp. 290-297

Fienberg, S. E. (1994), *A radical proposal for the provision of micro-data samples and the preservation of confidentiality*. Carnegie Mellon University Department of Statistics Technical Report No. 611.

Fienberg, S.E., U.E. Makov, and R. J. Steele (1998), *Confidentiality, uniqueness, and disclosure limitation for categorical data.* Journal of Official Statistics 14(4), 485-502.

Fellegi, I. P., and Sunter, A. B., (1969), *A Theory for Record Linkage.* Journal of the American Statistical Association, 64, 1183-1210.

Kennickell, A. B. (1991), *Imputation of the 1989 Survey of Consumer Finances: Stochastic relaxation and multiple imputation.* SCFWorking Paper, prepared for the Annual Meetings of the American Statistical Association, Atlanta, Georgia, August 1991.

Kennickell, A. B. (1997), *Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances.* SCF working paper.

Kennickell, A. B. (1998), *Multiple imputation in the Survey of Consumer Finances.* SCF working paper, prepared for the August 1998 Joint Statistical Meetings, Dallas, TX.

Kennickell, A. B. (2000), Wealth measurement in the Survey of Consumer Finances: methodology and directions for future research. SCF Working Paper, prepared for the May 2000 annual meetings of the American Association for Public Opinion Research, Portland, Oregon.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003), *Multiple imputation for statistical disclosure limitation.* Journal of Official Statistics 19, 1—16.

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

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* Wiley

Rubin, D. B. (1993), *Discussion: Statistical disclosure limitation.* Journal of Official Statistics 9(2), 461-468

A Combined Methodology for Assessing Identity and Value Disclosure Risk for Numerical Microdata

Krish Muralidhar and Rathindra Sarathy***

*** University Of Kentucky, Lexington KY 40506 USA, krishm@uky.edu**

**** Oklahoma State University, Stillwater OK 74073 USA , sarathy@okstate.edu**

Abstract: Numerical microdata are often masked prior to release to prevent disclosure of confidential information. One key aspect of assessing the effectiveness of masking techniques is their ability to prevent *identity* and *value* disclosure. Identity disclosure refers to the ability of an intruder to match a particular released record as belonging to an individual. Value disclosure refers to the ability of an intruder to predict the true value of confidential variable(s) using the released microdata. In this study, we establish a *theoretical basis* to assess whether a masking technique is capable of minimizing (both types of) disclosure risk. For masking techniques that do not provide minimum disclosure risk, we provide a *common basis* for assessing (both types of) disclosure risk.

1. Introduction

Government agencies and other organizations often mask numerical microdata prior to releasing or sharing it with other entities. The *primary* purpose for such masking is to prevent the disclosure of sensitive or confidential information contained in the data. Hence, it is critical that we have the ability measure the actual or the potential risk of disclosure resulting from a particular masking technique.

Prior to actually describing this study, we would first like to define disclosure risk since these definitions play an important role in the evaluation of disclosure risk. Dalenius (1977) provides a general description of disclosure risk as having occurred if an intruder is able to determine the value of a microdata point more accurately with the release of information (than without that information). Similar generic definitions of disclosure risk have been provided by Duncan and Lambert (1986). Our interest in disclosure risk is more specific than the generic (but useful) definitions provided above. We are interested in the specific ability to evaluate a set of masking techniques and disclosure risk resulting from these techniques. Hence, it is necessary to define disclosure risk in more concrete terms.

In practice, there are two types of disclosure, namely, identity disclosure and value disclosure. Identity disclosure refers to the case where, using the released data, an intruder is able to identify a particular released record as belonging to a particular individual. Clearly, this type of disclosure is relevant in situations where the identity of an individual is in itself considered sensitive. Value disclosure occurs if an intruder is able to estimate the value of a confidential variable for a particular record. Whether identity disclosure or value disclosure (or both) are important depends on the particular context. In some situations, being able to identify an individual as belonging to a particular record alone could constitute disclosure (such as when the released data consists of a set of individual with a disease). In others, that an individual belongs to the released data set alone does not constitute disclosure. In these cases, disclosure occurs only when an individual is able to estimate the value of a confidential variable. This situation occurs in the case of organizational databases where that an individual is the employee of the organization does not in itself constitute disclosure. However, if an intruder is able to estimate the value of a confidential variable for this particular individual, then such estimation constitutes disclosure. It is also easy to see that in some situations, it may be necessary for an intruder to first identify the record as belonging to an individual and then estimate the value of a confidential variable in order for disclosure to occur.

In practice, disclosure could also be deterministic (exact) or probabilistic (partial). In exact disclosure, an intruder is able to either identify a particular record as belonging to an individual with certainty or is able to compute the exact value of a confidential variable. As the name implies, in probabilistic disclosure, the intruder is not certain that a identity or value has been disclosed. In terms of identity disclosure, the intruder is able to identify that a record belongs to a particular individual with a high probability and/or estimate the true value of a confidential variable with a greater degree of accuracy. It is very clear that any masking technique that results in deterministic disclosure is unlikely to be used in practice. Hence, in the remainder of the paper, we will use the term “disclosure” to represent “probabilistic” or “partial value” disclosure.

The objective of this paper is to develop a methodology for assessing disclosure risk from two perspectives. First, consistent with the definitions of Dalenius (1977) and Duncan and Lambert (1986), we establish a *theoretical basis* to assess whether a masking technique is capable of minimizing (both types of) disclosure risk. For masking techniques that do not provide minimum disclosure risk, we provide a *common basis* for assessing (both types of) disclosure risk.

Topic III

Confidentiality aspects of statistical information (including panel surveys) taking into account (partly) register-based data

Assessment of Statistical Disclosure Control Methods for the 2001 UK Census

Natalie Shlomo
Southampton Statistical Sciences Research Institute, University of Southampton, Department of Statistics, Hebrew University, Office for National Statistics

Abstract: We define the disclosure risk scenarios that led to the statistical disclosure control (SDC) methods for the 2001 UK Census. We examine the SDC methods that were implemented based on a disclosure risk-data utility framework and assess whether the methods managed the disclosure risk while maintaining the utility and quality of the outputs. We conclude with final remarks and goals for forming strategies for future Censuses.

1. Introduction

Beginning with the 2001 UK Census, the Office for National Statistics (ONS) re-examined its statistical disclosure control (SDC) policies and methods for protecting standard Census tabular outputs. The initial SDC method that was planned for the 2001 Census was random record swapping on the microdata prior to tabulating the data (defined in Section 2) and higher population thresholds for released tables. This method was shown to give about the same level of protection as the method that was used for the 1991 UK Census based on a post-tabular variation of record swapping that was applied to the tables. However, prior to releasing the 2001 Census tables, it was decided that an additional disclosure control method of small cell rounding would also be applied to the tabular outputs. This was due to the following reasons:

- 100% of the questionnaire was coded compared to only 10% in the 1991 Census;
- Increasing IT technologies and the wealth of available public data, including the Neighborhood Statistics Service (NeSS) website which provides detailed small area social and economic statistics from both administrative and census sources, raised the level of disclosure risk compared to 1991;
- Pre-tabular record swapping leaves the perception that no SDC method is applied at all to the tables, thus raising concerns about the impact on future response rates for ONS Censuses and surveys.

Scotland, however, did not include the small cell rounding in their SDC strategies and this led to differential SDC methods across the Statistical Offices of the UK.

In this paper, we examine the SDC methods that were applied to the 2001 UK Census tabular outputs based on a disclosure risk – data utility framework (Duncan, et. al. (2001)). The purpose is to assess whether the methods managed the disclosure risk against the risk of re-identification while maintaining the utility and quality of the Census outputs. Section 2 describes the SDC methods and Section 3 the data used for the analysis. Section 4 presents the disclosure risk and data utility quantitative measures with results as well as an R-U confidentiality map. Section 5 concludes with a discussion of the analysis and goals for forming strategies for future Censuses.

2. SDC Methods Used in the 2001 UK Census

The SDC method implemented on the 2001 Census tables was a combination of a pre-tabular method of random record swapping and a post-tabular method of small cell rounding.

The views expressed are those of the author and do not necessarily reflect the views of the University of Southampton and the ONS.

2.1. Random Record Swapping

The most common pre-tabular method of SDC for Census outputs is record swapping. As defined in Willenborg and de Waal (2001), each record i is partitioned into three sub-vectors: x_i, y_i and z_i. Controlling for x_i, a household is selected for swapping having the same sub-vector x_j. In this case, the distributions of the pairs of values (x_i, y_i) and (x_i, z_i) are preserved after swapping. If X is chosen so that Y and Z are conditionally independent given X then swapping will not affect the joint distribution of X, Y and Z. For example, let Z define geographical variables, X the household characteristics (household size, age-sex composition of the household, ethnic background, etc.), and Y all other variables. The above method will swap households across geographical areas Z while ensuring that swapped households have the same characteristics on X. This protects against disclosure risk by perturbing the relationship between y_i and z_i in the record. Note that this method distorts the joint distribution of Y and Z though marginal distributions are maintained at a higher geographical level. In addition, because of the conditional independence, we obtain less inconsistencies and edit failures as a result of swapping records. This method also gives slight protection for the disclosure risk resulting in differencing two tables which are nested and non-coterminous because of the uncertainty introduced in the data.

For the 2001 UK Census, the random record swapping of households was carried out within a large geographical area defined by the local authority (LA). A random sample within strata defined by control variables was selected using a fixed swapping rate *f*. The control variables that were used were: hard-to-count index, household size, sex and broad age distribution of the household (0-25, 25-44, 45 and over). For each household selected, a paired household is found and all geographical variables are swapped. Note that this has the same effect as swapping all other variables and leaving geography fixed.

For this analysis we carried out random record swapping as implemented for the 2001 UK Census at the following swapping rates: 1%, 10%, and 20%. In addition, we carried out some modifications of the random record swapping in order to compare the disclosure risk – data utility across the different methods:

- As carried out in the 2001 UK Census, records were swapped on imputed records as well as non-imputed records. Imputed records arise from two sources: records that have missing items and whole records that were imputed for correcting the coverage of the Census based on the Census Coverage Survey. Since imputed records are a priori protected records, there is no need to perturb them and therefore we carried out the random record swapping only on the non-imputed records.
- Based on the tables used in the analysis (see Section 3) we identified and flagged all the small cells of the tables. We implemented a targeted record swapping by pairing and swapping households that matched not only on the control variables but also on the flagged variable. If, however, a household that was selected for swapping did not have a match on the control variables from among the flagged households, a match was found outside the flagged households.

Note that on average, about 0.15% of the households selected for record swapping were not swapped because no matching household was found for them. In general, those records would have to be swapped outside the large geographical area (LA) but this was not carried out in this analysis. Table 1 presents advantages and disadvantages of the record swapping.

Table 1. Advantages and Disadvantages of Record Swapping as a Pre-Tabular SDC method for Census Tabular Outputs

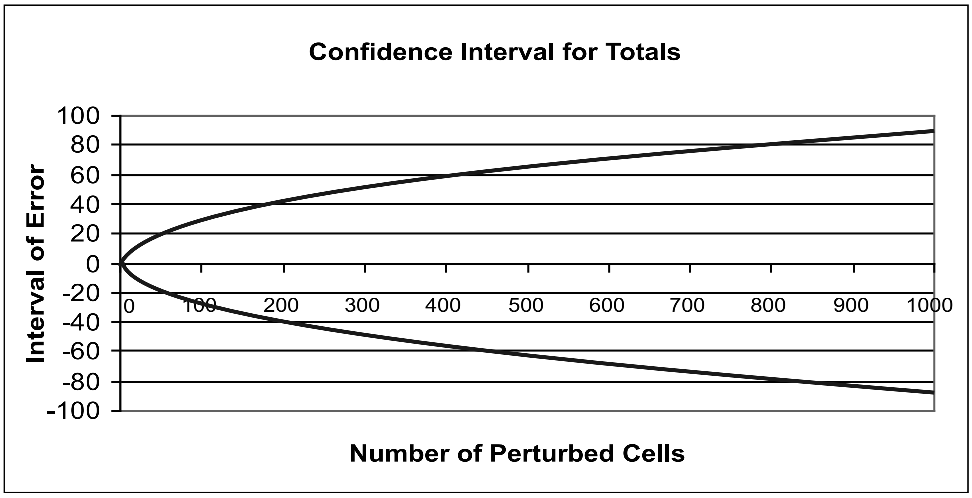
Advantages	Disadvantages
Consistent totals for all tables	Leaves a high proportion of risky (unique) records unperturbed
Preserves marginal distributions at higher aggregated levels	Errors (bias) in data and in particular joint distributions distorted
Some protection against disclosure by differencing two non-coterminous tables	Effects of perturbation hidden and can't be measures or accounted for in statistical analysis, i.e. a number in a table is not the true value
Less inconsistencies and edit failures when swapping geographies	Method is not transparent to users and appears as if no SDC method is used
Targeted swapping lowers disclosure risk	Targeted swapping causes more distortion in the distribution of the table

2.2. Small Cell Rounding

In comparison to pre-tabular record swapping where effects are hidden, the post-tabular rounding procedures are transparent to users and the stochastic forms of rounding can be taken into account when carrying out statistical analysis. For the 2001 Census tables (not including Scotland) small cells were rounded. The method used was an unbiased random rounding. Let x be a small cell and 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. The expected value of the rounded entry is the original entry. Each small 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 realisation 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.

When only small cells are rounded, the margins of the tables are obtained by aggregating the rounded and non-rounded cells, and therefore tables with the same population base will have different totals. The confidence interval for the expected differences from the true totals as a result of the small cell rounding procedures depends on the number of small cells that are adjusted in the table. Figure 1 presents the confidence intervals for the expected differences from true totals when rounding small cells to base 3.

Figure 1. Confidence Intervals for Random Rounding to Base 3



In addition, we also carried out modifications of the random rounding procedure for this analysis in order to compare the disclosure risk – data utility across the different methods:

- Since different totals are obtained for tables with the same population base, we carried out a semi-controlled small cell rounding where the overall total of the table is preserved. This method can also preserve some of the marginal totals in the tables as well (Shlomo and Young (2005)).
- A full (semi-controlled) random rounding was also carried out. This is implemented as described above for the small cells after first converting the entries *x* to residuals of the rounding base *res(x).*

Table 2 presents advantages and disadvantages of the rounding procedures.

Table 2. Advantages and Disadvantages of Rounding as a Post-Tabular SDC method for Census Tabular Outputs

Advantages	Disadvantages
Full protection for the high-risk (unique) cells	Inconsistent totals between tables when margins aggregated from rounded and non-rounded cells
Full rounding protects against disclosure by differencing two non-coterminous tables.	Small cell rounding gives little protection against disclosure by differencing so only one set of geographies and other variables disseminated
Small cell rounding has less information loss	Full rounding has margins rounded separately and tables aren't additive
Methods clear and transparent to users	Stochastic methods of rounding are easier to unpick and tables may need to be audited prior to release
Stochastic methods can be accounted for in statistical analysis	

Note that fully controlled rounding which preserves the marginal totals of the tables as developed within the Tau-Argus framework (Salazar, et al (2004)) is not at the moment a viable option for the size and scope of Census tables and therefore will not be examined further in this analysis.

3. Data Used

To carry out the disclosure risk – data utility analysis on 2001 Census data, we obtained unperturbed microdata from different Estimation Areas of the UK. In this report, we will show results for one Estimation Area: SJ - (Southampton, Eastleigh, Test Valley) 437,744 persons, 182,337 households, 1,487 Output Areas (OA). For this Estimation Area (EA), we defined five standard census tables (the number of categories of the variable are in parenthesis):

(1) Religion(9) * Age-Sex(6) * OA

(2) Travel to Work(12) * Age-Sex(12) * OA

(3) Country of Birth (17) * Sex (2) * OA

(4) Economic Activity (9) * Sex (2) * Long-Term Illness (2) * OA

(5) Health status (5) * Age-Sex (14) * OA

The microdata was perturbed according to the record swapping scenarios (random, random without imputed records and targeted) and then tabulated and rounded according to the rounding procedures: small cell rounding (SCA), semi-controlled small cell rounding (CSCA) and semi-controlled full random rounding (CRND).

4. Disclosure Risk – Data Utility Analysis

In this Section we assess the methods used in the 2001 UK Census based on a disclosure risk-data utility framework. This will examine whether an optimal balance was found between managing the disclosure risk and maximizing the utility of the data for the standard 2001 Census tables.

4.1. Disclosure Risk

The disclosure risk in population-based Census tables arises from small cells or small cells obtained from differencing tables. The record swapping will not inhibit small cells from appearing in the tables and therefore we need a quantitative disclosure risk measure which reflects whether the ones or twos in the table are true values or perturbed values.

The quantitative measure of disclosure risk for assessing record swapping is the proportion of records in the small cells that have not been perturbed. The perturbation comes from two sources: the record swapping procedure and imputation. Imputed records can be viewed as protected records and therefore there is no need to apply SDC methods on those records nor include them in the quantitative risk measures.

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, $|C_1 \cup C_2|$ the number of small cells with a value of 1 or 2. The disclosure risk measure is: $$DR = \frac{\sum_{i \in C_1 \cup C_2} I(R_i \quad perturbed \quad or \quad imputed)}{|C_1 \cup C_2|}$$

Table 3 presents results of the disclosure risk remaining in the Census tables as defined in Section 3 after implementing the different scenarios of record swapping.

Table 3. Percent Records in Small Cells of the Tables that were Swapped or Imputed

Method	EA SJ		
	1%	10%	20%
Original	84.2%	84.2%	84.2%
Random	82.3%	66.3%	49.7%
Rand/Imp	82.0%	63.4%	43.6%
Targeted	80.6%	45.9%	18.0%

In Table 3, we see that without any disclosure control method, there is a priori protection against disclosure risk because of the imputation. For EA SJ, there were about 16% imputed records. We see almost no impact on the disclosure risk from the 1% record swapping and it is about the same as if no SDC method was applied at all. Even for the targeted record swapping at the 1% swapping rate, we obtain about an 80% chance that a small cell in the table (one or a two) is the true value. This leaves a high probability of identifying uniques in the Census tables. For the higher swapping rates (10% and 20%), we are able to bring the disclosure risk down to lower levels of disclosure risk, especially if the records to be swapped are targeted from among the records in small cells of the tables. Note that if the random record swapping as carried out for the 2001 UK Census had not included the imputed records, the disclosure risk could have been lowered as will be shown in the R-U confidentiality map in Figure 4 at no cost to the utility of the data.

Forms of rounding eliminates all small cells in the table and therefore disclosure risk is zero with respect to the re-identification of small cells. For attribute and group disclosure, zeros in the table may not be true zeros since small cells can be rounded down to zero in the rounding procedure. The disclosure risk remains however when applying the method of small cell rounding and tables can be differenced. This is because only small cells are affected by the rounding procedure and large cells are left intact. Therefore, large counts in tables that are differenced can lead to disclosive small cells. For the 2001 UK Census, disclosure by differencing was managed and minimized by allowing only one set of geographies and other variables to be disseminated. Therefore, we won't assess the disclosure by differencing problem in this analysis and assume that for the rounding methods, there is no disclosure risk and we only need to examine one dimension of the decision problem and that is data utility.

4.2. Data Utility - Measuring distortions to distributions

Utility measures that measure distortions to distributions are based on distance metrics between the original and perturbed cells. Some useful metrics were presented in Gomatam and Karr (2003). Since the basic unit of most of the Census tables is a geography, i.e. Output Areas (OA), we are interested in a measure of distortion at this level of geography. Therefore, we will calculate the distance metric for each OA separately in the tables and then take the overall average across the OA's as the utility measure for the whole table. Note that we can also look at the table as a whole and measure distortions to distributions across all the cells.

Let D^k represent a table for OA k and let $D^k(c)$ be the cell frequency c in the table. Let $|OA|$ be the number of OA's in the estimation area. The distance metrics are:

- Hellinger's Distance:

$$HD(D_{orig}, D_{pert}) = \frac{1}{|OA|}\sum_{k=1}^{|OA|}\sqrt{\sum_{c\in k}\frac{1}{2}(\sqrt{D^k_{pert}(c)} - \sqrt{D^k_{orig}(c)})^2}$$

- Relative Absolute Distance:

$$RAD(D_{orig}, D_{pert}) = \frac{1}{|OA|}\sum_{k=1}^{|OA|}\sum_{c\in k}\frac{|D_{pert}^{k}(c) - D_{orig}^{k}(c)|}{D_{orig}^{k}(c)}$$

- Average Absolute Distance per Cell:

$$AAD(D_{orig}, D_{pert}) = \frac{1}{|OA|}\sum_{k=1}^{|OA|}\frac{\sum_{c\in k}|D_{pert}^{k}(c) - D_{orig}^{k}(c)|}{|k|} \text{ where}$$

$|k| = \sum_{c} I(c \in k)$ the number of non-zero cells in the k^{th} OA

Table 4 presents results of the three distance metrics for the record swapping scenarios for EA SJ and Tables 5 and 5a the results for the rounding procedures.

Table 4. Average Distance Metrics Between Original and Perturbed Cells per OA for Record Swapping

Method	EA SJ		
	1%	10%	20%
Random			
HD	1.035	3.721	5.279
RAD	4.302	32.437	53.001
AAD	0.138	0.726	1.053
Rand/Imp			
HD	1.044	3.714	5.238
RAD	4.337	32.345	52.433
AAD	0.136	0.722	1.036
Targeted			
HD	1.376	4.787	6.372
RAD	6.215	43.375	63.135
AAD	0.160	0.845	1.173

Table 5. Average Distance Metrics Between Original and Perturbed Cells per OA for Rounding

Method	SCA	CSCA	CRND
HD	5.272	5.279	5.416
RAD	76.804	76.824	84.641
AAD	0.629	0.630	1.021

Table 5a. Average Distance Metrics Between Original and Perturbed Cells per OA for Rounding Combined with Record Swapping

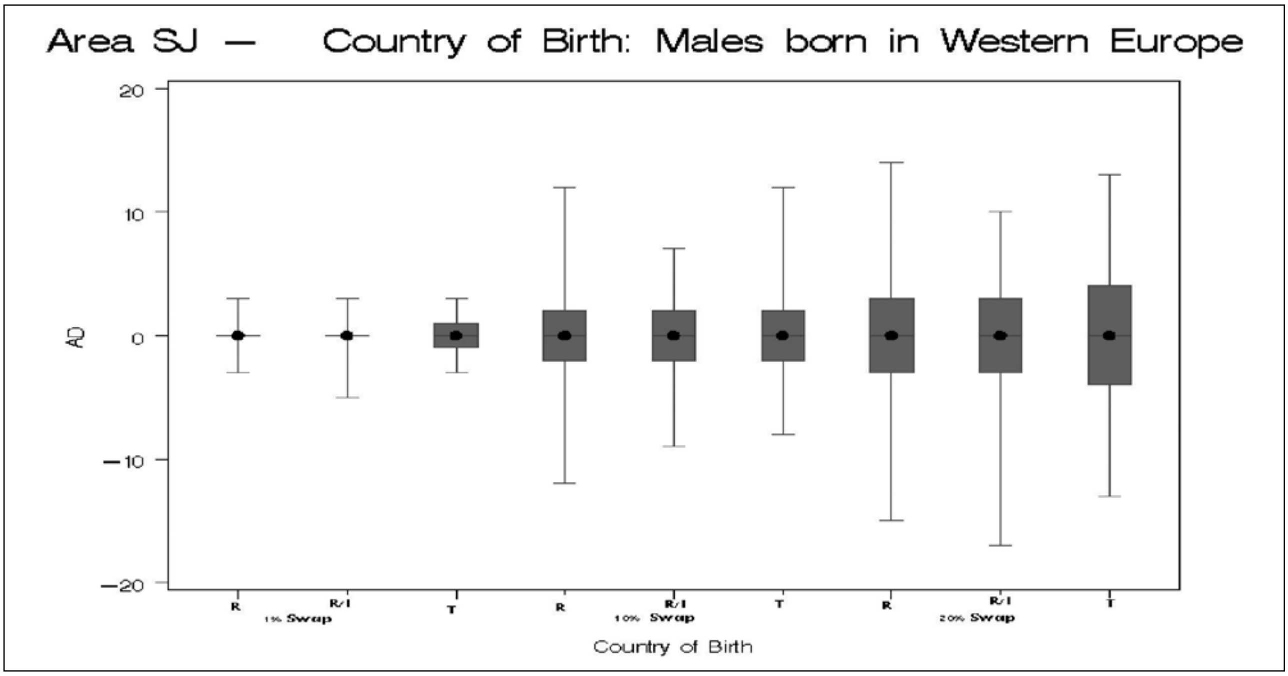
Method	1%			10%			10%		
	SCA	CSCA	CRND	SCA	CSCA	CRND	SCA	CSCA	CRND
Random									
HD	5.383	5.390	5.524	6.313	6.299	6.421	7.228	7.226	7.311
RAD	78.630	78.687	85.546	89.446	89.478	92.848	97.599	97.570	100.003
AAD	0.745	0.746	1.074	1.119	1.119	1.251	1.337	1.335	1.418
Random/Imputed									
HD	5.390	5.393	5.524	6.305	6.302	6.406	7.173	7.183	7.297
RAD	78.636	78.636	85.474	89.162	89.152	92.740	96.836	96.986	99.381
AAD	0.745	0.745	1.073	1.114	1.114	1.245	1.315	1.318	1.403
Targeted									
HD	5.444	5.442	5.575	6.791	6.764	6.872	7.800	7.818	7.899
RAD	78.709	78.721	85.530	89.157	89.048	92.339	96.271	96.326	98.651
AAD	0.753	0.752	1.080	1.165	1.161	1.292	1.383	1.386	1.469

In Table 4, we see a consistent pattern of small distance metrics for the 1% swapping rate and larger distance metrics for the 20% swapping rate. The measure of *AAD* tells us by how much the cells are perturbed on average for each OA. For example, for the random record swapping, each cell is perturbed by about 0.7 for the 10% swap and about 1.1 for the 20% swap. Similarly, for the targeted record swapping, each cell is perturbed by about 0.8 for the 10% swap and about 1.2 for the 20% swap. Between the random record swapping and the random record swapping without imputed records, we see almost no difference in the distance metrics. The targeted record swapping has the highest distance metrics showing that more distortion occurs as the swapping rate increases.

Tables 5 and 5a show the same distance metrics for the rounding procedures. In the table, we see much higher levels of information loss based on the distance metrics for the rounding procedures and even higher distance metrics when combining rounding procedures with the record swapping. One can argue that if we assume zero disclosure risk because no small cells are left in the table and there is no risk of disclosure by differencing, then we lower the utility of the tabular outputs by combining the record swapping with the rounding procedures. This loss of utility is minimal for the 1% swap, but increases for the higher swapping rates. Note in Tables 5 and 5a that the HD metric does not pick up differences between the full random rounding and small cell rounding (controlled or not), whereas the other distance metrics are more sensitive to the perturbation of the internal cells. This is because HD is heavily influenced by small cells. As seen in the table, the difference between the independent small cell rounding (SCA) and the controlled small cell rounding (CSCA) has about the same distance metrics for all the measures and therefore this utility measure is not sensitive to the totals which will be examined next.

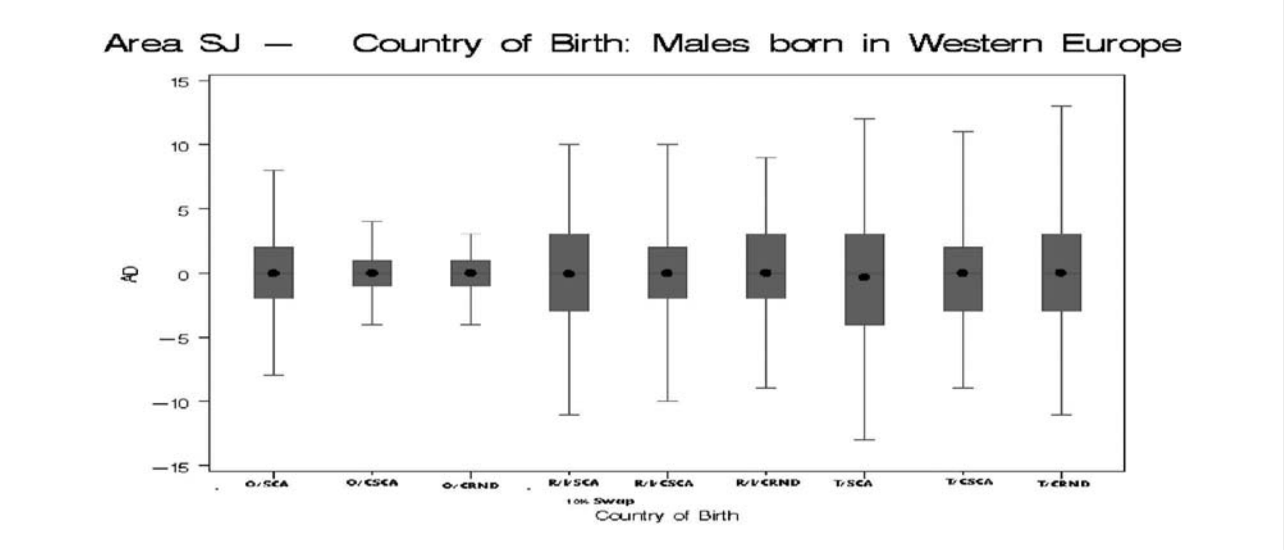
One problem for the rounding procedures was that different totals were obtained in tables with the same population base. This was particular problematic for users of Census tables who are mainly concerned about obtaining high quality aggregated level data for specified and non-standard geographical areas, for example school districts. The OA level tables are typically used as building blocks to construct higher level geographies. Because the tables are highly perturbed at the OA level of geography, aggregating OA data results in much information loss to the totals. In order to evaluate the range of the differences for sub-totals on specific Census target variables, we use the statistical graphing tool of a box plot on the differences between the perturbed sub-total and the original sub-total: $AD(N^k_{orig}, N^k_{pert}) = N^k_{pert}(C') - N^k_{orig}(C')$ where $N^k(C') = \sum_{c \in C'} D^k(c)$ is a subtotal in the k^{th} OA. Each box in the plot contains the inter-quartile range (between the 25th and 75th percentile) of the *AD*'s. The lower 25th percentile and the upper 25th percentile are represented by the whiskers of the box. The line in the middle of the box is the median of the *AD*'s and the dot represents the mean. The length of the box and the length of the whiskers gives an indication of how wide spread the perturbed totals are from the original totals. Figure 2 presents the box plot of the *AD's* in EA SJ based on the number of Males born in Western Europe within ten consecutive groupings of OA's for the different scenarios of record swapping.

Figure 2. Box Plot of AD's for the Number of Males Born in Western Europe in Ten Consecutive OA's of EA SJ for the Record Swapping Average Original Total in 10 OA's=24.6



From Figure 2 we see almost no loss of utility for the 1% swapping rate. The 10% and 20% swapping rates had more loss of utility with wide spread whiskers. For this particular example, there is a slight loss of utility between the random and targeted record swapping for each swapping rate. As shown, for this particular sub-total, the error in the total for ten consecutive OA's could be as much as ±15, which is about 61% of the average original value. This lowers the utility of the Census data, especially since users are not able to take the perturbation error into account in their analysis. Figure 3 presents the box plot for the post-tabular methods of rounding and combined with the 10% random without imputed (R/I) and targeted (T) record swapping.

Figure 3. Box Plot of AD’s for the Number of Males Born in Western Europe in Ten Consecutive OA’s of EA SJ for the Rounding and 10% Record Swapping Average Original Total in 10 OA’s = 24.6

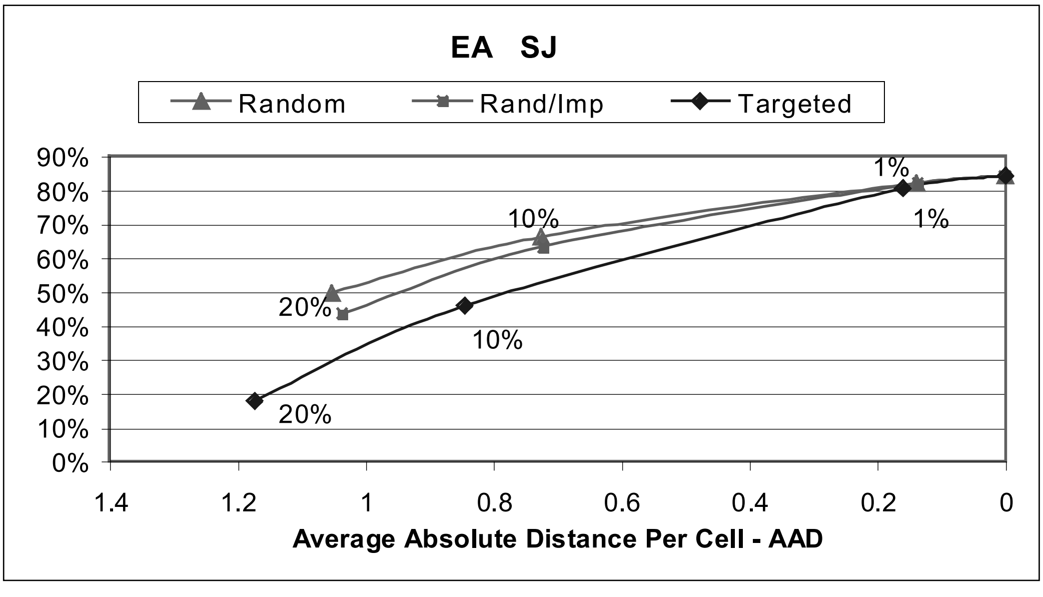


In Figure 3 we see that the boxes are narrower for the rounding methods on the original data compared to the rounding method when combined with record swapping. The effect of the random and the targeted record swapping on the *AD’*s is about the same. What is interesting to note is that the length of the boxes and whiskers are narrower for the semi-controlled small cell rounding compared to the independent small cell rounding. This means that there is more utility in the tables since the perturbed sub-totals do not differ from the original totals as much as the independent small cell rounding. Even the semi-controlled full rounding of all entries in the table shows slightly higher utility than the independent small cell rounding.

4.3. R-U Confidentiality Map

In Figures 4 we present an empirical R-U confidentiality map for the record swapping methods based on the risk measure *DR* and the distance metric AAD.

Figure 4. R-U Confidentiality Map for Estimation Area SJ



As seen from Table 3, the 1% swapping rates for all methods of record swapping have high utility but also very high disclosure risk (about 80% of the small cells in the table (ones and twos) are true values). The 10% targeted record swapping has about the same disclosure risk as the 20% random record swapping (about 45% of the small cells are true values). However, we gain much more utility in the data with the 10% targeted record swapping compared to the 20% random record swapping. This graph clearly shows that the 10% targeted record swapping is preferable for a given disclosure risk. Another conclusion from the R-U Confidentiality Map is that had the random record swapping as implemented in the 2001 Census not included the imputed records, we would have obtained higher utility in the data for the same disclosure risk.

5. Discussion

Based on the risk-utility analysis, we see that the SDC methods of record swapping and rounding used for the 2001 UK Census managed the disclosure risk. As a stand alone method, the random record swapping gives little protection against disclosure risk but could have been improved had a targeted record swapping taken place. When combined with the small cell rounding, we obtain full protection of the Census tabular outputs taking into account that there is no risk from differencing tables because of the standard geographies and other variables that were disseminated. The loss of utility mostly resulted from the bias that occurred because of the record swapping and the fact that totals were different across tables with the same population base. In particular, the very large and sparse 2001 origin-destination tables were badly affected by the SDC methods. Utility could have been improved by placing more controls into the rounding algorithm and preserving overall totals of the tables and benchmarking.

Based on these results, it is clear that when planning for future censuses there needs to be consistent SDC methods across all of the UK Statistical Offices that disseminate Census data. The methods need to ensure that sufficient statistics (totals, averages and variances) are not compromised. Flexible table generating software should be developed so that users can design and customize their own Census tables. The SDC method would then be applied only once on the final outputted table as opposed to the standard system today where hard-copy tables are disseminated on paper or on CD's and non-standard geographies are aggregated from perturbed lower level geographies. Improved GIS systems may allow more flexible dissemination of geographies in the future and further development of the SDC tool Tau-Argus may automate better the SDC processes. Finally, SDC methods should be tailored and coordinated between the types of Census outputs, such as standard tables, microdata, origin-destination tables, in order to maximize utility while managing the disclosure risk.

6. Acknowledgements

A special thank you to Alexa Courtney for her tremendous help in preparing the data for analysis and to Joan Holland for her help in obtaining and interpreting the Census data.

References

Duncan, G., Keller-McNulty, S., and Stokes, S. (2001) “Disclosure Risk vs. Data Utility: the R-U Confidentiality Map”, Technical Report LA-UR-01-6428, Statistical Sciences Group,Los Alamos, N.M.:Los Alamos National Laboratory.

Gomatam, S. and Karr, A. (2003) “Distortion Measures for Categorical Data Swapping”, Technical Report Number 131, National Institute of Statistical Sciences.

Salazar, J.J., Lowthian, P., Young, C., Merola, G., Bond, S. and Brown, D. (2004) “Getting the Best Results in Controlled rounding with the Least Effort”, (J. Domingo-Ferrer and V. Torra, eds.), *Privacy in Statistical Database*, Springer: New York.

Willenborg, L. and de Waal, T. (2001), *Elements of Statistical Disclosure Control,* Lecture Notes in Statistics, 155 (Springer Verlag, New York).

EU SILC anonymisation: results of the Eurostat Task Force

Museux Jean-Marc
Living Conditions and Social Protection Statistics, Unit F3, Eurostat, European Commission,
L-2920 Luxembourg, jean-marc.museux@cec.eu.int

Abstract: The European Instrument on Income and Living Conditions (EU-SILC) is gathering ex post output harmonised household and individual, cross sectional and longitudinal micro data income and living conditions on collected in 25 MS and 2 EEA countries. Data are a mix of register and survey data. Anonymised micro data will be released to researchers. In order to reflect on best practices, Eurostat convened a task force of experts in data protection and/or in the EU-SILC instrument. The paper presents the results of the Task Force. The methodological problems (panel/cross sectional, register/survey, individual/household) related to the anonymisation of a European database are reviewed, problems enhanced by the multiplicity of perceptions and realities of disclosure risk encountered in the different countries. The paper details both the methodological solutions proposed by the Task Force and the strategic and operational options retained to achieve a good trade-off between, on one side, information content and usability, and, on the other side, monitoring disclosure risk.

1. Background

The European Instrument on Income and Living Conditions (EU-SILC) is gathering ex post output harmonised micro data collected in 25 MS and 2 EEA countries. It aims to provide comparable annual cross sectional data on income and living conditions and longitudinal data on income across Europe. The main operation started in 2004 for 10 MS and will reach the almost full regime (25 countries) in 2005. The data collection is based on the European Parliament and Council Regulation n°1177/2003 concerning Community statistics on income and living conditions. The instrument allows for flexibility and MS can collect data directly from a new survey or compile data from existing surveys and registers.

The EU-SILC micro data is a unique information source for studying poverty in its relation to socio-economic variables. It will be the primary source of data used by Eurostat for the calculation of many indicators in the field of Income, Poverty & Social Exclusion such as the Structural Indicators of Social Cohesion; indicators adopted under the Open Method of Coordination such as the 'Laeken' indicators of Social Inclusion and indicators of Pensions Adequacy; Sustainable Development Indicators of poverty and of ageing; and many other indicators published on the Eurostat New Cronos database. It is therefore a key tool for policy makers in particular, for monitoring Lisbon strategy. It will be indubitably of great interest for the research community in order to carry out detailed studies on poverty and living conditions.

The EU-SILC data are cleaned and imputed by the MS and then individual records are be transmitted to Eurostat without any direct identifiers (e.g. name, address, fiscal numbers). MS deliver a cross sectional dataset annually and a longitudinal dataset in which up to 4 years individual trajectories are compiled.

EU-SILC individual records are likely to be considered as confidential data in the sense of Article of Council Regulation 322/97 (Statistical Law) because they would allow indirect identification of statistical units (individuals or households). With this respect they should only be used for statistical purposes or for scientific research.

Commission Regulation 831/2002 granted the Commission to provide access to confidential data in the Eurostat premises and to release anonymised micro data for instance via CD-ROM to researchers.

Anonymised micro data are defined as individual statistical records which have been modified in order to minimise, in accordance to best practices, the risk of identification of the statistical units to which they relate.

Provision for the release of anonymised micro data to researchers is present in the EU-SILC framework Regulation n°1177/2003. The first data set to be released from EU-SILC will contain 2004 cross sectional data and will be available in March 2006.

EU-SILC is the successor of the European Community Household Panel (ECHP) which was the main source of data for Income, Poverty & Social Exclusion for the reference years between 1994 and 2001. The ECHP anonymised user data base has been widely released to researchers. In this respect, this initiative was pioneering. Many contracts have been signed between Eurostat and research bodies ruling the access to ECHP micro data.

In order to come up with best practices and recommendations for anonymisation of EU-SILC user data base (UDB), Eurostat has convened a Task Force (TF) bringing together experts in the domain of anonymisation and experts in the SILC instrument[1] This paper covers only the work of the TF on the research release. Public release was addressed by the TF on a pilot basis..

2. Main orientations

The work of the TF initially drew upon three different approaches for assessing the disclosure risk when releasing micro data :

(1) the ONS approach based on population uniques and sample uniques (Elliot and Skinner) implemented in the SUDA software package,

(2) the CBS approach based on sample counts and implemented in the software package Mu-Argus.

(3) the ISTAT approach based on individual and household measure of disclosure risk (Franconi and Polettini, 2004), also implemented in Mu-Argus

The objective of the TF was not to reconcile the different methods but to benefit from each of them in order to issue recommendation for the anonymisation of the EU-SILC micro-data sets.

EU-SILC is a typical household survey. By nature key income variables are measured at household level. The household dimension carefully developed in EU-SILC makes it a primary source for household studies. It is characterised by the presence of hierarchies in the micro data file where the link between individuals and household is always present. From a disclosure point of view, this structure complicates the disclosure risk assessment since individuals can be disclosed through the identification of household characteristics and vice versa, households can be disclosed by characteristics of individuals. Given this hierarchical structure, it was necessary to consider the disclosure risks at both the individual and household level when assessing the disclosure risk of EU-SILC database.

One characteristic of EU-SILC instrument is the coexistence of a longitudinal dimension together with a cross sectional component. Although, the framework does not impose these to be linkable, many MS have used an integrated design (a rotational panel as recommended by Eurostat) and in this case, the components are not independent. Therefore it was necessary to ensure consistency between the anonymisation methods for the longitudinal and the cross-sectional data files.

The TF has pointed the specificities of so called register countries. For these countries, some of the income variables available in the EU-SILC may come directly from registers (DK, NO, SE, FI, SI, IS). If this register information is available together with direct identifiers to external users, the risk of disclosure is greatly increased. A specific section of this paper is dedicated to it.

The recommendations proposed in this document draw upon the analysis of the disclosure risk in the EU-SILC 2003 data base for two countries. Analysis has been conducted for one small country, Luxembourg which gets the highest sampling fraction and in one medium size country, Greece, which gets a small sample fraction.

[1] (B. Bruno (Eurostat), L. Coppola (Istat), P. Feuvrier (INSEE), Ph. Gublin/J. Longhurst (ONS), N. Jukic (Stat Of. Slovenia), H. Minkel (Stat Bun), JM Museux (Eurostat), E. Schulte Nordholt (CBS), H. Sauli (Stat Fin))..

3. Identifying variables in EU-SILC UDB and intruder scenarios

For research release, the list of variables and the corresponding structure of the User Data Base (UDB) is likely to be very close to the structure of the data bases transmitted to Eurostat and described in Commission Regulation N°1983/2003. The data base is likely to be complemented with several derived variables (e.g. household size measured with the modified OECD equivalence scale, household activity status).

When releasing micro data files, statistical offices want to protect against standard disclosure risk scenarios by which an intruder, possessing a few variables (called identifying variables) about individuals in the population, is able to re-identify individual records from the micro data file thereby disclosing the content of other variables. These other variables can be classified as "sensitive" if they are perceived as (strictly) confidential.

The initial EU-SILC variables have been reviewed with respect to the identifying potential (the availability for intruders). They have been classified as Identifying, Sensitive or Others. Some variables have been classified as problematic regarding their specific nature: design weights and strata can lead to disclosure of detailed geographic information; precise timing variables such as month of birth or month on moved in or out are likely to create too fine classification which can lead to rare combinations; detailed fieldwork information, although not strictly disclosive, may contain personal information. In addition, according to the CBS methodology, the identifying variables are grouped into **E**xtremely identifying, **V**ery identifying, (simply) **I**dentifying. This grouping refers to the specific methodology used by CBS and aims to introduce a hierarchy (in terms of availability) between the different variables. The subgroups are nested: one extremely identifying variable is considered automatically as very identifying and so on. This grouping exercise was based on the experience of the TF members.

The result is as follow:

Extremely identifying: REGION, DEGREE of URBANISATION.

Very identifying: SEX; COUNTRY OF BIRTH; CITIZENSHIP 1; CITIZENSHIP 2.

Identifying YEAR OF BIRTH; BASIC ACTIVITY STATUS; EDUCATION AT PRE-SCHOOL /COMPULSORY SCHOOL; CHILD CARE VARIABLES; DWELLING TYPE; TENURE STATUS; NUMBER OF ROOMS AVAILABLE TO THE HOUSEHOLD; DO YOU HAVE A CAR?; BATH OR SHOWER IN DWELLING; MARITAL STATUS; CONSENSUAL UNION; CURRENT EDUCATION ACTIVITY; HIGHEST ISCED LEVEL ATTAINED; SELF-DEFINED CURRENT ECONOMIC STATUS; STATUS IN EMPLOYMENT; OCCUPATION (ISCO-88 (COM)); NACE; MOST FREQUENT ACTIVITY STATUS (EMPLOYED, UNEMPLOYED, RETIRED); HOUSEHOLD TYPE, HOUSEHOLD SIZE.

Because, in some sense, it is difficult to unify national disclosure perception and the availability of variables, the subset of identifying variables can be seen as the union of national identifying variable sets.

The potentially identifying variables are then grouped into different "scenarios" which represent the information set the intruder has to hand to attack the database in different situations.

The TF has collected thirteen specific scenarios based on national experience. They can be classified according to

- whether they relate to research release only or to both public and research release (by definition public scenarios apply to the release to researchers)
- whether they consider attack at household level or only at individual level

In order to assess the disclosure risk of research release of the EU-SILC database, 3 generic scenarios have been generated representing the common core of the different scenarios collected by the TF.

EU1 (Simple attack with HH information (individual and household level))

REGION x SEX x YEAR OF BIRTH x MARITAL STATUS x HH SIZE x HH TYPE

EU2 (Nosy neighbour individual attack– minimum scenario for nosy neighbour)

REGION x URBANISATION x SEX x YEAR OF BIRTH x BASIC ACTIVITY STATUS x BATH OR SHOWER x DO YOU HAVE A CAR? x EDUCATION x OCCUPATION x SECTOR OF ACTIVITY x HH SIZE x HH TYPE

EU3 (Occupational group address book individual attack)

REGION x URBANISATION x SEX x YEAR OF BIRTH x MONTH OF BIRTH x STATUS IN EMPLOYMENT x OCCUPATION x SECTOR OF ACTIVITY

These scenarios include quite a large number of variables. Their conjoint availability in attackers' data base depends on the national situation. The disclosure risk analysis relies on detection of rare combination in the population based on sample estimates.

The CBS (NL) methodology for assessing the disclosure risk for research release does not use scenarios but considers all 3 way-combinations of one Extremely identifying variable, one Very identifying variable and one Identifying variable focusing on rare (unique, two's, three's) combinations in the sample.

4. Measure of risk and threshold

The three different approaches for assessing disclosure risk introduced in Section 2 can be distinguished by the measure of risk they rely on.

The ONS method is based on a measure of risk developed by Elliot and Skinner: the probability that a unique match of identifying variables with a sample unit is correct. This risk is estimated by computational intensive resampling methods implemented in the SUDA software. Special uniques in the sample (sample uniques which correspond to population uniques up to a high level of aggregation) are detected by the software. In addition to the individual and global measures of risk the results of the method also provides variables and value contribution to the risk Up to now, the ONS has only implemented this method for assessing the risk of microdata samples from the population censuses.

In the CBS approach, considering research release only rare combinations of identifying keys are considered as problematic. The following table gives the threshold for the number of replications of the given combination of the identifying keys above which the records are considered as potentially disclosive, e.g in the majority of MS a record is considered potentially disclosive if it is unique in the sample.

Sampling fraction	Countries	Threshold
1/50 – 1/2	LU (f=2.5%)	5 (1+114 f)
1/100 – 1/50	MT, IS, CY	3
1/200 – 1/100	EE, SI	2
< 1/200	All other 21 MS	1

ISTAT approach provides a measure of risk for individual record based on the scenario where an attacker has a database with identifiers and key variables and tries to link it with the records in the sample. A match is given when the combination of the key variables is the same in both the sample record and the attacker's data base. The disclosure risk is defined as the probability that a match is correct (i.e. the sample record actually corresponds to the individual in the attacker's data base). This risk is estimated taking into account the sampling design, and is implemented in Mu Argus. Once the risk is estimated, a threshold has to be fixed to decide whether a record has to be considered safe or unsafe. Such a threshold is chosen depending of several aspects: (i) availability of data base providing identifiers and key variables; (ii) data base level of completeness and quality; (iii) comparability between the data base and the micro-data to be released (in terms of classification of the key variables); (iv) sampling fraction…etc.

The three methods are based on different risk measures, it is hoped however that they produce datasets protected against the main disclosure risk.

The ONS method has not been used for this analysis because the software was not available at Eurostat at the moment of the study, the approach implemented did not take into account the sample design weights and the algorithm was known to be computer intensive[2] (one run could last for one week on a PC).

5. Household and individual records

In EU-SILC data base, two levels of information (household and individual) will always coexist. Even if household identifiers are removed from individual files, the presence of household information (e.g. equivalised income) at individual level allows for household clustering.

The ISTAT approach allows for estimating individual risk of disclosure (named BIR in Mu-Argus) as well as household risk of disclosure (named BHR in Mu-Argus). To some extent, BHR implicitly takes into account the individual level of risk because it is based on the assumption that an household member is identifiable through the identification of another member of the same household. The ISTAT approach takes into account the household size through including it in the set of key variables. Moreover, an individual is considered safe only if BHR is lower than a chosen threshold, and at the same time BIR is lower than the same threshold divided by the household size. This implies an increasing level of protection, according to the household size. Such a strategy is sensible when the scenario of attack assumes that the intruder tries to match individual records, knowing which household each individual in the file to be released belongs to.

In other words, in order to integrate both individual and household levels, disclosure risk analysis is carried out on a file of individual data where identifying household variables are collected at individual level. Alternatively, a household file where data on individuals (such as age and sex of household members) are brought together at household level might be considered.

On the basis of the latter approach, UK studies on disclosure risk of household data have shown that most households of size 5+ are unique in the UK population and that they are a non-ignorable part of the population, The UK pattern has been reproduced for Slovenia and is likely to occur for all MS. These studies underline the fact that the disclosure risk of large households when using the ISTAT approach is likely to be underestimated. The household character of EU-SILC data are taken into account in scenario EU1 which take into account household and individual level attack and partly in scenario EU2 where household characteristics have been included at an individual level. The ISTAT model for household risk estimation, as implemented in Mu-Argus 4.0, has been used for assessing the risk.

It is proposed to control the actual disclosure risk linked to large households by carrying out uniform recoding as described in the section 7 and to only release household files under strict licence as was the case for ECHP.

[2] At the work session, a new version of the software was presented which proves to remove most of the concerns expressed here.

Others solutions could have been:

- removal of households above a threshold and some uniform recoding
- perturbation with some uniform recoding

Given large households are crucial for the analysis of poverty their removal is not conceivable for research release.

Perturbation was not adopted here since the method requires a high level of competence both for designing a perturbed dataset and for analysing perturbed datasets. In addtion the method risks altering key indicators derived from the perturbed dataset.

6. Longitudinal data

One component of the EU-SILC instrument is longitudinal: individuals (and corresponding households) are traced for a minimum period of four years.

The release of cross sectional and longitudinal components to researchers is crucial.

Many countries are using an integrated design where the longitudinal and the cross sectional component may share the same individuals, as well as the same variables. Therefore, the matching between the files will be possible even with the absence of common identifiers. So there is a need to have a common strategy of anonymisation for both datasets so that one cannot be used to disclose the other (and vice versa).

The tracing of individuals over time increases the disclosure risk because transitions in identifying variables are likely to determine rare patterns. This type of risk has been highlighted for the different identifying variables and is detailed in the table of annex 4.

At the same time, this presupposes that the attacker is able to detect this change equally. For registers it is actually difficult to get access to data that exactly correspond to the reference period of EU-SILC data.

In conclusion, it is very difficult to fully protect longitudinal data and keep relevant information for researchers at the same time. Following practice in MS, the TF recommends to consider the release of longitudinal files only under strict license.

7. Fieldwork and sampling information

The release of sampling design information is potentially problematic because it may reveal geographical information or delineate subpopulations. In a first approach, it is recommended to remove the design information from the file. This issue if further addressed when discussion researchers needs in section 11.

8. Global recoding

The aim of global recoding and top coding of identifying variables is to reduce the number of unsafe records by reducing the level of information that can be used to identify them. The TF considers that an appropriate choice of global recoding could achieve a significant decrease of the disclosure risk of the EU-SILC data base. In addition global recoding methods can be harmonised for all MS and also are more easily implemented at a centralised level. The harmonisation of the anonymisation methods is crucial for usability and usefulness of the released database. The details of the recodes are based on a systematic examination of the distributions of the identifying variables and the identification of rare sample combinations in 3 ways combinations of variables. The choices made are benchmarked against each other using the number of remaining unsafe records. On the basis of the analysis carried out with the software Mu-Argus (4.0) and presented in section 12, the TF proposes the following recoding as a first stage of risk reduction:

Label	Code	Global/top coding 1st step
REGION	DB040	not considered at the first step
DEGREE OF URBANISATION	DB100	not considered at the first step
SEX	RB090	None
COUNTRY OF BIRTH	PB210	Local/EU/non EU/world
CITIZENSHIP 1	PB220A	Local/EU/non EU/world
CITIZENSHIP 2	PB220B	Removed
YEAR OF BIRTH or AGE[3]	RB080	Bottom recode (1923 and before)
MONTH OF BIRTH		Removed
DWELLING TYPE	HH010	Modality 5 put to missing
TENURE STATUS	HH020	None
NUMBER OF ROOMS AVAILABLE TO THE HOUSEHOLD	HH030	Top coding (6 and more)
BATH OR SHOWER IN DWELLING	HH080	None
DO YOU HAVE A CAR?	HS110	None
MARITAL STATUS	PB190	None
CONSENSUAL UNION	PB200	None
EDUCATION (ISCED)	PE040	Isced 5 and 6 regrouped
ECONOMIC STATUS	PL030	None
STATUS IN EMPLOYMENT	PL040	None
OCCUPATION (ISCO-88 (COM))	PL050	None
NACE	PL110	Regrouped at 1 one letter (19 level)
HOUSEHOLD TYPE		Derived
HOUSEHOLD SIZE		Derived from hierarchical structure

For EU-SILC, it is of primary importance not to hamper the scientific interest of the data base. For this reason, special attention has been put into keeping the year of birth/age at the current level of aggregation.

Geographical information

At this first stage no geographical information (NUTS code and degree of urbanisation) is considered for inclusion in the data base. EU-SILC was not primarily designed for providing regional information. Moreover, the NUTS 2 information as available in the original data sets might not be useful because sample might not have been designed to be representative at this geographical level. The same level of NUTS code encompasses different geographic realities depending of the country. In small countries, the first NUTS breakdown is confounded with the country itself. In some other countries NUTS classes are not homogeneous and are not relevant for statistical analysis. There is thus a danger to rely only on NUTS code to define geographic desegregation. The degree of urbanisation, on his side, is a complex concept that might not be readily available to an attacker.

For some MS (most likely the large MS), the impact of reintroducing some geographical information might be limited. Under the hypothesis that regional information is statistically relevant and taking into account that it could be of primary importance for researchers and policy makers to carry out regional studies, these MS should have the possibility to allow for the release of this information in the research files.

[3] Providing only AGE instead of YEAR of BIRTH would provide an additional safeguard against disclosure because it depends on the date on which it is calculated which is not always available to the attackers

9. Local suppressions

It is expected that the global recoding and top/bottom coding that have been proposed so far will significantly decrease the re-identification risk associated with EU-SILC. If the number of records for which the risk measure is considered too high (the so called "unsafe" records) remains limited (less than a few percents), the datasets can be released to researchers under licence as mentioned above. Alternatively, the unsafe records can be protected by carrying out local suppression or random perturbations of key variables.

Different patterns for local suppression exist. The suppression pattern can be controlled by the use of suppression weights which can help to penalise local suppressions for some variables. Ideally, suppression should concentrate on the least crucial variables for researchers and variables that will not affect the politically relevant estimates. Age, gender, activity status, household type and tenure status are particularly important in this respect.

In addition, local suppressions may alter the comparability between output of Official Statistics providers and results of research and policy evaluation. Local suppression and basic perturbation may thus hamper the interest of researchers in the data. Local suppression will also break out the calibration of the files released. Calibration is crucial to ensure consistency with other sources (demographic…). Eventually, the coherence of the local suppression pattern between the different releases of the datasets will be very cumbersome to implement.

On the other hand, local suppressions might be embedded in the bulk of the "natural" missing values in the data files resulting from item non response. In some situations local suppressions may allow the release of more detailed information for several critical variables (e.g. geographical information). The right balance between the two aspects has to be obtained on a case by case basis.

Because of selectivity of the suppression, imputation of suppressed values seems not to bring an appropriate solution to this problem.

10. Register information

In register countries, some EU-SILC variables (mainly some income components) could come directly from register, which under certain conditions can be public or accessible to researchers.

The TF surveyed the situation of the register countries. The most difficult situation is encountered in Norway, where a public file on individuals exists in Internet available for anyone. For all citizens included in the tax register, the file contains the following variables: name, address, postcode, net assets, income and tax. The variables are not identical with the variables used in Norwegian SILC, but they can be of use for the possible attacker.

For other countries, the situation is better because the access to register information is usually restricted and controlled. Although it is possible - at least for researchers - to match different registers with identifying variables to EU-SILC files, it takes knowledge of the data sources, resources and skills to attain these registers.

When EU-SILC variables can be obtained by an attacker from register sources, the TF recommends applying rounding techniques to EU-SILC variables. For instance, the base for rounding could be tuned to the data and vary along the measurement scale. If rounding did not offer sufficient protection micro – aggregation could also be considered.

These aspects require additional studies which were beyond the scope of this TF. They have to be addressed and carefully monitored at national level.

11. Research community needs

To validate the a priori choices discussed above made by the TF, Eurostat has carried out a large consultation of researcher on the basis of the TF proposal. Comments have been collected in a free format and the following needs have immerged.

Geographic information is required not only because regional analysis is important but mainly for the coupling of macro information at the level of the region (employment rates …) in order to develop explanation model of individual behaviour. NUTS1 seems a minimum requirement. The degree of urbanisation also appears as an important explanatory variable.

Age and date of birth are critical variables especially for analysis of life transitions (child care, education, work, retirement). The need to locate precisely in time the event recorded by the survey is crucial. For instance, it is required to define household equivalized scales that take change of composition during the reference period. Year coarsening is far from sufficient with this respect: the possibility to identify period with a precision of a quarter seems to be a minimum requirement. With this respect the withdrawal of move in/out information would have also important consequences. The impact of top coding of age has also to be carefully assessed on the basis of the interest of developing studies on ageing people. 80+ to coding might prevent some interesting analysis on elderly.

Researchers pay a lot of importance on the quality of the inference they can draw from observed data. With this respect, not providing design weight will not allow to develop alternative weighting schemes. Worst, the masking of clustering effect (PSU, SSU) will not allow to develop correct (embedded multi level) modelling of observed behaviour.

Masking fieldwork information will not allow detailed analysis of respondent behaviour and quality checking of the analysis.

The prevalence of ISCO on NACE with respect to level of details is validated by researchers. It is underlined that the rough coarsening of country of birth and citizenship would prevent migration analysis and the inclusion of migration trajectories in human behaviour.

12. Results of experiments

The disclosure risk of EU-SILC databases has been studied using the data available at Eurostat for Luxembourg and Greece. Luxembourg is characterised by the highest sampling fraction (f=2.5 %) among all countries. Greece is characterised by a small sampling fraction (f= 0.1 %) and should be representative of medium size countries and regions in large countries.

The impact on disclosure risk of the global, top/bottom coding described in section 8 is studied using the different approaches and scenarios described in section 3 and 4. For the ISTAT approach different levels of risk and different levels of attack (household/individual) are considered. The focus is put on the number of "unsafe" records and the structure of the local suppressions proposed by the software package.

The CBS and the ISTAT approach have been tentatively compared by first, selecting the variables that concentrates the risk in the CBS approach and then measuring the ISTAT risk associated to the combination of variables ("CBS scenario – individual attack"). The results of the two approaches are roughly comparable for individual attacks.

For the EU2 complex scenario, the software is limited and does not allow the consideration all the variables simultaneously [4]. The scenario has thus been subdivided in overlapping sub scenarios so to capture the dependencies between the set initial variables.

[4] In addition, the whole set of variables might not be available at the same time in all MS

For Greece, for attacks developed at the individual level, the number of suppressions is always less than 1 % of the number of records in the datasets.

The household attack always implied a higher number of suppressions. For scenario 2 with a level of risk of 0.04 the number of suppression is less than 2% of the number of records. The analysis of the structure of suppression among the different types of households have shown (not visible in the excel sheet) that the suppression preferentially affects the large households (1 over 2). Some large households are considered are not protected according to the method. This could be a characteristic of the structure of household population in Greece.

For Luxembourg, the results clearly demonstrate that the level of coding considered is not enough to declare the file as safe. The level of "unsafe" records can reach 10% depending on the approach and level of risk. The impact of recoding age by 5 years classes is briefly studied and seems to improve the situation. Further studies might be needed. The low performance of the coding for Luxembourg is due to the relatively high sampling fraction in that country. This corresponds to a more important disclosure risk typical of small regions/countries.

In conclusion, for large countries (sampling fraction lower than 0.01%), the global recodings described in section 8 are likely to significantly decrease the disclosure risk (measured in terms of local suppression) regardless of the approach (sampling fraction lower then 0.1 %). Further coding should be envisaged for small countries (LU (f=2.5 %) EE (f=:0.6%) CY (f=1.1%) MT (f= 1.8%) SI (f=:0.5%) and IS (f=1.3%)

13. EUROSTAT strategy for the design of Anonymised EU-SILC UDB for research release

In order to implement the here above recommendations, trying to find the right balanced between the need for harmonisation on one side and the need for some flexibility to adapt to MS sensitivity., Eurostat proposes the following strategy emphasising the prevalence of objective risk measure and good practices.

It can be split in two steps:

1st step:

- The global recoding envisaged so far should be carried out uniformly for all national data sets (longitudinal and cross sectional)
- For large countries this should maintain the number of records for which the risk measure is too high to a few percents of the number of records;
- For small countries, further recoding are not unlikely, most likely for the variable Year of Birth;
- At least NUTS1 and degree of urbanisation geographic coarsening should be introduced whenever the level of risk remain limited. Homogeneity of geographical grouping should prevail and more detailed breakdown could also be envisaged. It is likely that region of size equivalent to small countries can be made explicit in the files.
- MS should have the possibility to propose limited number of additional coding/grouping (regrouping of rare modalities) adapted to their national specificities. However for the usability of the UDB, their number and their extent should remain limited and nested;
- Depending of the shape of the distribution of the income variables, grouping/top coding of these variables should be envisaged in order to protect "outliers".

In view of the researchers needs, Eurostat advocates to maintain minimum design information (anonymised PSU, anonymised SSU, order of selection and rotational group). However, neither strata nor design weights should be released. Fieldwork information such as contact information, proxy would be released in order to allow respondent behaviour studies which could help to improve the instrument.

In the longitudinal files, month of move in/out would be released because they are crucial longitudinal information and the increase of risk is inherent part of the risk of releasing longitudinal information.

The increase of risk induced by these decisions can be monitored by the contractual link and the close follow up mechanism Eurostat has put in place. The procedure and the practical conditions of the data release are intimate part of the risk management and reduction. They must be considered at the same time as the methods used for protecting data. They should be part of the agreement to be reached between MS and Eurostat.

Statistical Confidentiality in Longitudinal Linked Data: Objectives and Attributes

Mario Trottini
University of Alicante, Spain (mario.trottini@ua.es)

Abstract. Researchers and practitioners interested in applications of multiple objectives decision theory generally agree that the task of structuring the objectives and identify suitable attributes is the most important step in any formal analysis of a complex decision problem. In this paper we briefly review some of the relevant literature on the topic and we argue on its relevance for the current research on Data Disclosure Limitation, with particular emphasis on the dissemination of longitudinal linked data.

1. Introduction

A number of European and U.S. official statistical agencies have undertaken initiatives to develop longitudinal linked data sets. These are defined as “microdata that contain observations from two or more related sampling frame, with measurements for multiple time periods for all units of observation” (Abowd and Woodcock 2001, page 1). They are obtained integrating (through record linkage techniques) existing (administrative, or survey) microdata possibly collected by different agencies. Longitudinal linked data are essential to a wide range of research efforts and provide exceptionally rich source of information to address complex policy issues in key areas such as health care, education, and economic, to mention just a few. From the perspective of an official statistical agency, in addition, longitudinal linked data have great potential to enhance existing official statistics, improve data accuracy and reduce data collection redundancies (see Mackie and Bradburn 2000). The information content of longitudinal linked data, however, makes them vulnerable to disclosure. Requirements designed to protect confidentiality in the native data sets, i.e. those data sets from which the links were made, can drastically reduce the potential research (and non-research) benefits of linking.

This makes dissemination of longitudinal linked data a complex decision problem. An ideal data dissemination procedure, in fact, should: (i) allow legitimate data users to perform the statistical analyses of interest *as if* they were using the data set originally collected; (ii) reduce the risk of misuses of the data by potential intruders aimed to disclose confidential information about individual respondents, harm the data providers, or embarrass the statistical agency; (iii) be operational (it should be possible for the agency to implement the data dissemination procedure given the agency’s resources - budget, time, people skill, technology etc.). This identify three conflicting objectives (that we call “maximize usefulness”, “maximize safety” and “minimize cost”) that no data dissemination procedure can fully achieve simultaneously. Improvement in an arbitrary subset of these objectives usually requires to reduce achievement in some of the objectives in the complementary set and there is no data dissemination procedure which is obviously the best. In addition the above objectives are too ambiguous to be of operational use and there is no obvious measure that can be used to quantify the extent to which they are achieved by different candidates for data dissemination.

The research literature and current practice in Data Disclosure Limitation, have addressed these issues only in part and to a different extent. *Decision theory*, we believe, might provide a suitable framework to think about these problems. Within this framework a sensible choice of the data dissemination procedure requires the agency/ies responsible for it to:

a) Identify a set of suitable alternatives (candidate data dissemination procedures);

b) Defining the fundamental *objectives* in more operational terms;

c) Define suitable *attributes* that can measure the extent to which objectives are achieved when an arbitrary alternative is considered;

d) Assess the trade-off between the fundamental objectives of the problem. This means that for any arbitrary subset of the objectives the agency has to make a decision about how much of those objectives is willing to sacrifice in order to improve achievement in the others.

The research literature on decision theory have proposed guidelines for the implementation of the four steps decision analysis described above. In this paper we briefly review the most relevant results for structuring objectives and defining suitable attributes (points (b)-(c) above) and we argue on their relevance for increasing the values of existing research efforts in statistical confidentiality[1].

Section 2 reviews current alternatives for data dissemination and proposes a broader definition of "alternative". Strategies for a suitable structuring of the objectives and their relevance for the dissemination of (longitudinal linked) microdata are described in section 3. Section 4 deals with the problem of attributes definition and attributes selection. Section 5 summarizes the main results of the paper.

2. Identifying the Alternatives

The research literature on statistical confidentiality has identified three ways of disseminating (longitudinal linked) microdata: (1) releasing a *masked* version of the data obtained through a suitable transformation of the original data; (2) *restricting access* by reducing the set of users or the modality of access to the data; and (3) generate *synthetic data* through multiple imputation (or other) methods.

Strengths and limitations of these approaches have been extensively studied (see, for example Abowd and Woodcock 2004). The problem of selecting an alternative for data dissemination is often reduced to compare different instances of (1) to (3) the discussion focusing on whether broader access (data masking) is more important than greater data details (restricted access) and the ability of imputation method (syntetic data) to accomplished both for targeted complex analyses.

Despite the fact that several official statistical agencies disseminate longitudinal linked data using a combination of these three basic approaches (see for example Abowd and Lane 2003a), identification of alternatives is still understood as a comparison of the three basic approaches. We argue here that the value of each method would increase if we start thinking about data dissemination procedures as combinations of them. This idea has several rationales.

First of all, it is well recognized that data users and data users needs are very diverse. Thus a data dissemination procedure that relies only on one approach is unsatisfactory since it would likely produce a data set of insufficient detail for the more sophisticated data users while perhaps unnecessarily disclosing information not needed for more basic research (Mackie and Bradburn 2000). This is specially true for longitudinal linked data that are essential both for complex modeling (often of nonlinear relationships) and for the production of official statistics targeted to a much broader and less specialized audience.

In addition the definition of data dissemination as a combination of different dissemination modes, forces the agency to think of the problem as an optimal portfolio problem. The amount of resources devoted to each dissemination mode being the "parameter" that differentiates candidate alternatives. As noted by Abowd and Lane (2003b) "Any two protection methods are correlated in their risk of disclosure of confidential information, but not perfectly. Combining the two methods can, then, produce greater data utility for any level of disclosure risk in exactly the same way that an investor can achieve greater expected return for any given level of investment risk by combining the risky assets into a portfolio".

[1] The paper complements a companion paper presented by the author at the UNECE-EUROSTAT workshop held in Luxembourg on April 2003. In that paper (see Trottini 2003) the author discussed the trade-off problem (point (d) above) assuming as given attributes and objectives.

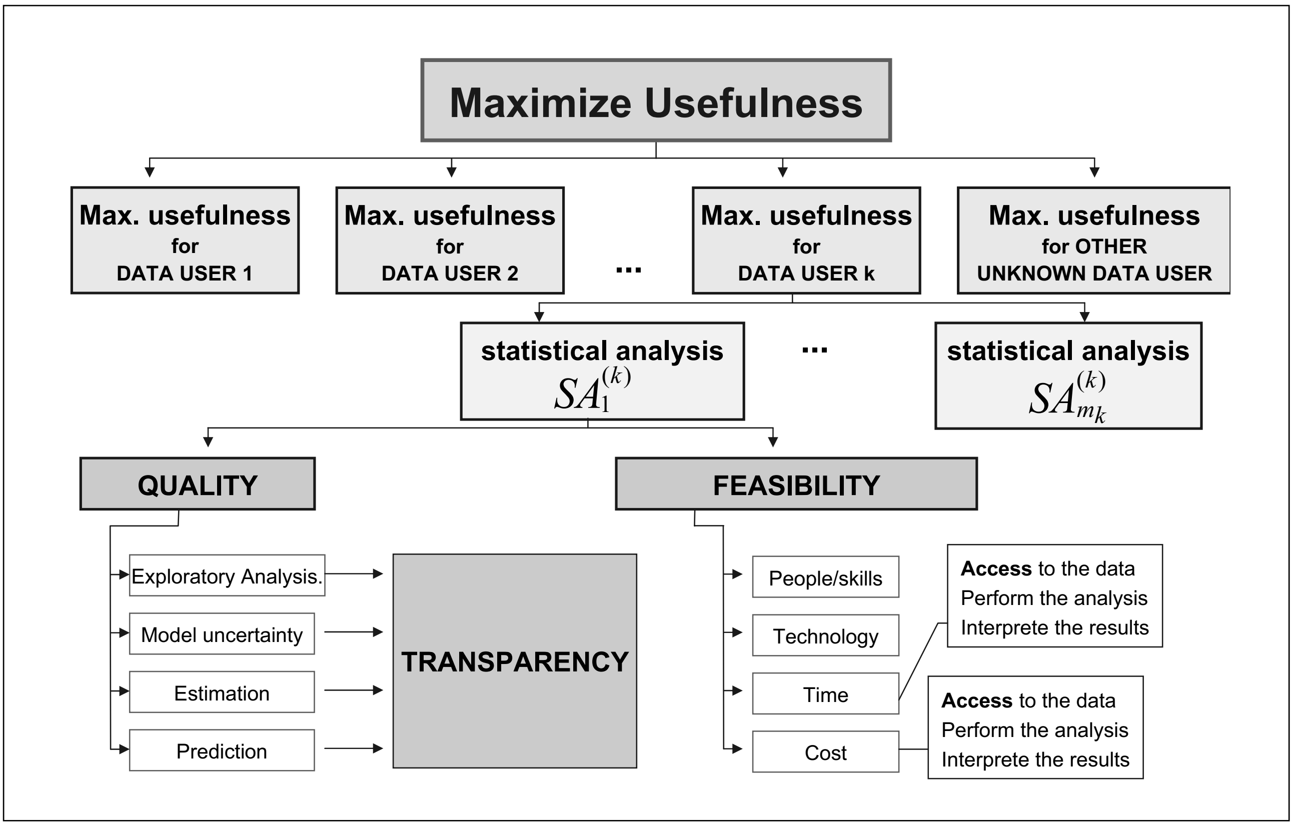
3. Structuring the Objectives

In complex decision problems the fundamental objectives are usually too broad and ambiguous to be of operational use. A useful strategy is then to divide an objective in *lower level objectives* that clarify the interpretation of the broader objective. As an illustration, consider the objective "maximize usefulness". According to the definition in section 1,

> "[A data dissemination procedure is useful] to the extent that it allows *legitimate data users* to perform the *statistical analyses* of interest *as if* they were using the data set originally collected".

In the above definition there are at least three sources of ambiguity. In order to make the definition of operational use we need to specify: a) The set of legitimate data users; b) For a given user in a) the statistical analyses relevant for the user; c) For a given user in a) and statistical analysis in b) the interpretation of "as if", that is a formal definition of what we mean by "preserving statistical inferences". It can be noted that the specification of objectives is a hierarchical process. In the above example, we need to specify a) first, and then b) given a), and then c) given a) and b). Not surprising, in decision analysis the output of the specification process is called a *hierarchy*. On the top of the hierarchy there are the fundamental objectives and at the bottom all the lowest level objectives that specify all aspects that matter to assess achievement of the fundamental objectives. Figure 1 shows a possible hierarchy for the objective "maximize usefulness".

Figure 1. Example of hierarchy for the objective "maximize usefulness".



In the hierarchy in figure 1, "maximize usefulness" is specified as maximize usefulness for $k+1$ different data users, $U_1,\ldots,U_{k+1}$. For a given data user, U_h, maximize usefulness is interpreted as preserving inferences of m_h statistical analyses. For a given user and statistical analysis "preserving inferences" is further specified as preserving "Quality" and "Feasibility" of the analysis. The hierarchy is then completed by identifying lower level objectives for "Quality" and "Feasibility". The objective

“Quality” is specified by taking into account different features of the inferences that data users might want to preserve (exploratory analysis, estimation, model uncertainty etc.) and for each feature the “Transparency” of the data dissemination procedure, that is the extent to which the data dissemination procedure provides direct or even implicit information on the added bias and variability induced by the procedure. “Feasibility”, on the other hand, is specified by considering time, people, skills, technology required to access and analyze the disseminated data (as compared to the original data).

The “splitting” process that generates a hierarchy can be repeated several times, if necessary. The stopping point depends on several considerations. First of all at each split we should make sure that the set of the lower level objectives that is produced represents all the relevant aspects of the broader objective that has been split. This constraint being satisfied, the number of lower level objectives should be kept as small as possible. Test of importance can be used to minimize the number of lower level objectives (see Keeney and Raiffa 1976, chapter 2). A second aspect that must be taken into account is the balance between “objective” representation of the fundamental objectives and feasibility of preferences elicitation. The more we split an objective the easier is to define “objective” attributes for the lowest level objectives and thus an “objective” representation of the fundamental objective. However, continuing splitting increases the dimension of the attribute that represents the fundamental objective and makes harder to formalize sensible trade-offs for different alternatives (since the number of elements that are involved in the trade-off increases as well). A good compromise is to build a hierarchy as detailed as possible in order to have a representation of all the relevant aspects of the problem but use the extended hierarchy as a qualitative tool to define quantitative attributes only at higher levels. The idea of the hierarchy, in its simplicity, provides a great tool to clarify what really matters in the decision analysis. Building a hierarchy can be very helpful in decision problems where several decision makers have to reach a joint decision (this is often the case in the dissemination of longitudinal linked data). Different decision makers can merge their hierarchies and the merged hierarchy will provide a suitable framework for comparison and constructive criticisms.

The use of hierarchy in statistical confidentiality is null, and, we believe, that’s unfortunate. Despite the fact that the research literature and current practice on disclosure limitation, as a whole, have identified many relevant aspects of the three fundamental objectives “maximize usefulness”, “maximize safety” and “minimize cost”, just few of those aspects are taken into account at the decision stage in the applications. For instance, the relevance of the lower level objectives in the hierarchy in figure 1, has been extensively discussed in the research literature of Data Disclosure Limitation (see, for example, Mackie and Bradburn 2000, and Fienberg 2003 and 2004). However, how many applications do really refer to such hierarchy when try to assess “usefulness” of the disseminated data? But few exceptions, as Abowd and Woodcock (2001), the common approach is to focus on very few items of those represented in figure 1. “Accessibility”, “Feasibility”, and “Transparency”, for example, are often not considered and “Quality” is replaced by “Quality of parameter estimates” ignoring other relevant aspects of “Quality” already identified in the research literature like “Model Uncertainty” or “Quality of residual analysis”, to mention just a few. Such incomplete hierarchies can compromise any posterior effort aimed to define suitable attributes and assessing sensible trade-offs.

We do not believe on a universal hierarchy appropriate for any arbitrary disclosure limitation problem. In decision theory it is well understood, in fact, that even for a specific decision problem with a single decision maker, hierarchies are not unique and different hierarchies can lead to different courses of action.

However we do believe that statistical agencies that disseminate data collected under a pledge of confidentiality would obtain great benefits by building their own hierarchy for their specific data dissemination problems. The hierarchy would help the decision maker to clarify the interpretation of the relevant objectives, to check that no relevant aspects of the decision have been ignored, facilitating the communication of all parts involved in the problem.

4. Defining and Selecting Attributes

Assuming that a set of objectives has been specified and that it is appropriate for the data dissemination problem of interest, the next step, in the decision analysis, is to define a suitable set of attributes. In this section we review the main issues related with attribute definition and attribute selection[2] and we discuss their relevance for data dissemination problems.

4.1. Types of Attributes

According to the research literature on decision theory, we can distinguish three types of attributes (see Keeney and Gregory 2005): *natural attributes*, *constructed attributes*, and *proxy attributes*.

When there exist an obvious scale that can be used to measure the extent to which an objective is achieved such a scale represents a *natural attribute*. For example, the objective "minimize cost" has the natural attribute "cost measured in euros". Natural attributes directly measure the extent to which an objective is achieved in a natural scale commonly understood. But for the example described above and few others, the use of natural attribute in disclosure limitation is quite unusual due to the complexity of the objectives involved in the problem. For example, for the objective "maximize usefulness", for which we have described a possible "splitting" in section 3, no natural attribute can be defined even after trying to decompose the objective in lower level objectives and searching for natural attributes for each of the lower level objective.

When no obvious scale for an objective exists, we could still try to directly measure the extent to which an objective is achieved by constructing a "subjective scale" or "subjective index". The scale, which is called *constructed attribute*, should take into account the relevant aspects of the objective as described by the hierarchy of the decision problem. A panel of experts usually takes the responsibility for it. Two illustrative examples of constructed attributes - both discussed by Keeney and Gregory (2005)- are the Dow Jones Industrial Average that measures movement in the stock market, and the Michelin rating system for restaurants. Note that although defining a constructed attribute that directly measures the extent to which an objective is achieved is not always possible (or successful) "interpretability" is a priority for constructed attributes. "Interpretability" here means that the decision maker should be able to associate to each "level" of the constructed attribute a clear description of the consequences for the objective of interest and viceversa, the decision maker should be able to describe consequences for the objectives in terms of levels of the attribute (in the terminology used in subsection 2, "interpretability" requires the attribute to be *comprehensive* and *understandable*). To the extent of our knowledge there are not examples of constructed attributes in disclosure limitation and that's unfortunate as we explain in the next section. For the moment we turn our attention to proxy attributes.

A *proxy attribute* is an attribute that reflects the degree to which an associated objective is achieved but does not directly measure the objective (Keeney and Raiffa 1976).

The value of a proxy matters only to the extent that it serves as predictor of the objective of interest. It's usefulness depends on the "prediction error" or, which is the same, on the relationship that exists between the objective of interest and the associated objective measured by the proxy and on the decision maker's understanding of such relationship. Note, for example, that *monotonicity* of the relationship (the greater/smaller the value of the attribute the better the achievement of the objective) is not sufficient. Monotonicity allows the decision maker to rank different alternatives in terms of the attribute but not to make sensible trade-offs. Trade-off assessment, in fact, requires to understand how differences in the proxy attribute values translate into different degrees of achievement of the objective of interest (for a formal argument in terms of utility functions se Keeney and Raiffa 1976, chapter 2).

[2] Our review of attribute definition and attribute selection is a very short summary of a more detailed discussion on the topic by Keeney and Gregory (see Keeney and Gregory 2005).

The research literature on disclosure limitation presents several (we believe too many!) examples of proxy attributes. For example, for the objective “maximize usefulness” different authors have discussed proxy attributes based on (Hellinger, Kullback-Leibler and other) distances between a density estimation under the perturbed and unperturbed data (see Gomatam et al. 2003, for example). Others, for the same objective, have proposed proxy attributes based on measures of discrepancy between summary statistics for the perturbed and unperturbed data (see G. Crises 2004 for a review). The intuition underlying all these proxies is that low distortion of the original data implies approximately correct inferences for most of the statistical analyses. However it is hard to see how perturbations of the original data expressed by any of the proxies listed above can be translated by the decision maker into meaningful statements about degradation of relevant statistical inferences. Not even monotonicity is guaranteed to be preserved. Being this the case how can one responsibly think whether, for example, a decrease of the Hellinger distance, say from 0.4 to 0.3, is worth an increase, say from 2% to 3%, in the percentage of records correctly re-identified? Wouldn’t be better in this case trying to define a constructed attribute? We believe so and we explain why in section 3. To make the argument we need a preliminary description of the desirable properties of an attribute.

4.2. Desirable Properties of an Attribute

Keeney and Gregory (2005) identify five sufficient properties of good attributes. Because of space limitations, here we discuss just three of them: *comprehensiveness*, *understandability* and *operationality*.

An attribute is *comprehensive* if satisfies two properties: (a) it takes into account all the relevant aspects of the objective that is meant to measure; (b) the values judgments embedded in the attribute are appropriate for the decision problem. A constructed attribute that takes into account only differences in parameter estimates, for instance, is not comprehensive for the objective “Quality” according with the hierarchy outlined in section 3 since the users cost of the inference (which includes time to access the data, software and people skills necessary to analyze the disseminated data) are relevant aspects of the objective not considered in the attribute. On the other hand, if for the same objective we use an attribute based on discrepancies between summary statistics evaluated using the perturbed and unperturbed data (as proposed in G. Crises (2004)) we are making the value judgement that data users will ignore the information provided about the masking and will use the released masked data sets *as if* they were the original data (otherwise the attribute should compare summary statistics under the original data with the corresponding *estimates* under the masked data). Assuming that the selected statistics considered in the attribute reflect the relevant aspects of “Quality” (we really doubt that such statistics do exists in real applications), the chosen attribute will be comprehensive to the extent to which this value judgment is appropriate for the decision problem. In general all the attributes that involve counting, such as “number of records re-identified”, implicitly assume that all items are equally important and we should ask the question whether this is an appropriate assumption in the decision problem under study.

A comprehensive attribute takes into account all the relevant aspects of the corresponding objective but is not of much help in the decision analysis if all the parts involved in the decision problem do not have a clear understanding of the levels of the attribute. An attribute is *understandable* if the decision maker and anyone else interested in the decision process understands what each level of the attribute means in terms of the objective of interest. In Data Disclosure Limitation understandability is a key property for two reasons. First of all, if any of the attributes for the fundamental objective “maximize safety” and “maximize usefulness” is not understood by the decision maker, then no sensible trade-off can be made. In addition understandability is a necessary condition to maintain data users’ confidence on the agency’s data dissemination procedures. This relates to the discussion on the “transparency” objective in section 3. Data users understanding of the perturbation that the data dissemination has introduced into the statistical analysis of interest is crucial for the acceptance of the statistical agency’s data dissemination procedure.

Unfortunately, even an attribute that meets all the previous properties is not sufficiently good if it's practical use in the decision problem generates a cumbersome work for the analyst who is in charge to implement the decision analysis. A fundamental property of an attribute is thus operationality. An attribute is *operational* to the extent that it is possible to obtain the values of the attribute for the set of different alternatives.[3] For the fundamental objective "maximize safety", for example, a natural attribute for a Bayesian would be a function of the intruders' posterior distribution for the sensitive variables given the disseminated data. However it could be the case that evaluation of such posterior distribution is too cumbersome or even infeasible and constructed or proxy attributes should be considered instead.

4.3. Selecting an Attribute: a Decision Problem

The evaluation of a candidate attribute in terms of the properties described in the previous section is not a dichotomic outcome (presence, absence). Different attributes are comprehensive (understandable, and operational) to different degree. In complex decision problems it is quite unusual to find attributes that fully satisfy all the properties. Rather the choice requires a decision about how much of a subset of properties we are ready to sacrifice to improve achievement of the others. A usual trade-off is the one that involves "understandability and comprehensiveness" on one hand, and "operationality" on the other. If a comprehensive and understandable attribute is not operational we might choose an alternative attribute which is not as much as understandable and comprehensive as the original but can be evaluated for the different alternatives. Note, however, that comprehensiveness and understandability are the priority. Meaning that we should reduce these properties as little as possible and stop as soon as we get an attribute that within the constraints of the problem (time, money, people skills, technology) is operational. The preference structure in the trade-off that we have just described, has a natural explanation in the discussion of the desirable properties in section 2. Comprehensiveness and understandability are a necessary condition for the decision maker to be able to make sensible trade-offs which is the core of the decision problem.

These ideas have a direct application on the prescriptive order in attribute selection. As described by Keeney and Gregory (2005), the choice of an attribute for a given objective should start with natural attributes. If, even after trying to decompose the objective in lower level objectives, no natural attributes can be found (or can be found but are not operational) then we should try to define a constructed attribute. Only when this turns out to be an infeasible task we should look for proxy attributes.

The discussion on section 4 shows that, having checked that no natural attributes can be defined, too often in disclosure limitation problems we choose the easiest "solution". We identify a proxy attribute. The non "interpretability" of proxies, and thus the practical impossibility to make sensible trade-offs do not seem a sufficient deterrent for their use nor a motivation to invest on constructed attributes. Part of the reason, we believe, is that quantitative proxies (such as those described in G. Crises (2004)) are perceived as "more objective" than subjective indices as constructed attributes are. The argument, however, seems weak. As commented before, knowing the value of some measures of discrepancy between distributions (or between summary statistics) evaluated for the perturbed and unperturbed data is, in general, of little or no value to understand the degradation of relevant statistical inferences. This is especially true for the complex statistical modeling of interest in the analysis of longitudinal linked data sets. We believe that in these cases, constructed attributes based on a panel of experts (that could certainly contain representatives of legitimate data users) would allow much more sensible trade-offs.

This is not meant to say that proxy attribute are useless. Rather than constructed attributes should receive more attention that they did so far.

[3] Note that in Keeney and Gregory (2005) the definition of "operational" attribute addresses the additional concern of whether the attribute allows the decision maker to make informed value trade-offs. The definition that we use here does not address this additional concern and rather refers to the definition of "'measurability" described in Keeney and Raiffa (1976). The reason for this choice will be apparent in the next section.

5. Conclusions

On page 9 of the book, *Value-Focused Thinking. A Path to Creative Decision Making*, Keeney says:

> “There is a tendency in all problem solving to move quickly away from the ill-defined to the well-defined, from constraint-free thinking to constrained thinking. There is a need to feel, and perhaps even to measure, progress toward reaching a “solution” to a decision problem.”

To get that feeling of progress, in Data Disclosure Limitation, we often quickly identify objectives, attributes and some viable alternative and proceed to evaluate them, without making the effort that a comprehensive definition of the problem, in terms of alternatives, objectives and attribute would require. This paper has addressed these concerns with particular emphasis on the importance of a proper structuring of objectives and the prescriptive order in the selection of attributes. The discussion hasn’t focused on longitudinal linked data, as much as desired, but, we believe, it is particularly relevant for this type of data, given: (i) the complexity of the modeling usually associated to the analysis of longitudinal linked data; (ii) the multiple decision makers involved in the problem; and (iii) the different perspectives and perceptions of risk and utility that must be accommodated in the final decision.

Acknowledgements

Preparation of this paper was supported by the U.S. National Science Foundation under Grant EIA-0131884 to the National Institute of Statistical Sciences. The contents of the paper reflects the authors’ personal opinion. The National Science Foundation is not responsible for any views or results presented.

References

Abowd, J. M. and Lane J. (2003a), “Synthetic Data and Confidentiality Protection", Workshop on Microdata, August 2003, Stockholm, Sweden.

Abowd, J. M. and Lane J. (2003b), “The Economics of Data Confidentiality". Unpublished paper presented at the National Research Council’s Committee on National Statistics Workshop on Confidentiality and Access to Research Data Files, October 2003, Washington DC.

Abowd, J. M. and Woodcock S. D. (2001), “Disclosure Limitation in Longitudinal Linked Data". In *Confidentiality, Disclosure, and Data Access. Theory and Practical Applications for Statistical Agencies*. P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (Eds.), North-Holland, Amsterdam, 135–166.

Abowd, J. M. and Woodcock S. D. (2004), “Multiply-Imputing Confidential Characteristics and File Links in Longitudinal Linked Data". In *Privacy in Statistical Databases*. Domingo-Ferrer J. and Torra, V. (Eds.), Springer-Verlag, 290–297.

Fienberg, S. E. (2003), “Allowing Access to Confidential Data: Some Recent Experiences and Statistical Approaches". Workshop on Microdata, August 2003, Stockholm, Sweden.

Fienberg, S. E. (2004), “Datamining and Disclosure Limitation for Categorical Statistical Databases", IEEE International Conference on Data Mining, Workshop on Privacy and Confidentiality, November 2004, Brighton, England.

Gomatam, S., Karr, A. F., and Sanil, A. (2004), “Data Swapping as a Decision Problem", *Journal of Official Statistics*, to appear.

G. Crises (2004), “Information Loss Measures for Microdata in Database Privacy Protection", Research Report CRIREP-04-004, September 2004, Dept. of Computer Engineering and Mathematics, Rovira i Virgili University of Tarragona, Spain.

Keeney, R. L. (1992), *Value-Focused Thinking. A Path to Creative Decision Making*, Harward University Press 1992.

Keeney, R.L. and Raiffa H. (1976), *Decisions with Multiple Objectives*, New York:Wiley 1976.

Mackie, C. and Bradburn, N. (2000), *Improving Access to and Confidentiality of Research Data: Report of a Workshop*. Committee on National Statistics (CNSTAT). Commission on Behavioral and Social Sciences and Education. Washington DC: National Academies Press, 2000.

Trottini, M. (2003), “Assessing Disclosure Risk and Data Utility: A Multiple Objective Decision Problem", UNECE/Eurostat Work Session on Statistical Data Confidentiality, Luxembourg April 2003.

Confidentiality aspects of household panel surveys: the case study of Italian sample from EU-SILC

Lucia Coppola and Giovanni Seri

Istat – Italian National Statistical Institute, 00184 Rome, Italy, {lcoppola, seri}@istat.it

Abstract: In this paper we discuss some of the disclosive features to deal with when releasing data collected through a household panel survey. The discussion and the empirical analyses are based on provisional data from the EU-SILC Italian survey. In particular, two structural characteristics are considered: (i) the hierarchical data structure, providing information simultaneously about household and individual characteristics; (ii) the longitudinal data structure, providing information about household and individual specific patterns of change during the period of observation. The disclosive power of these information depends on the nature of the information available to the intruder. We firstly point out a few intruder's attack scenarios. Secondly, we propose an anonymisation strategy to protect micro data against intruders' attack under all these scenarios at the same time. Such a strategy is based on the estimates of re-identification risk at individual and household level, and on a reduction of household and individual information through global recoding and/or local suppression.

1. Introduction

The Italian National Statistical Institute (Istat) is releasing so called Microdata Files for Research (MFR) since more then ten years. MFRs consist of individual records representing a sample of the population (MFRs are released only for social surveys). Statistical confidentiality is preserved reducing the information contents of the files, and minimising the risk of identification of statistical units. Users requiring MFR are asked to sign an agreement with Istat.

In order to release a MFR, a reasonable evaluation of the risk of disclosure is needed. Istat recently adopted an approach consisting in estimating for each record the 'risk of re-identification' at individual level (Franconi and Polettini, 2004). A threshold for the re-identification risk is then fixed at a reasonable low level. On the base of this threshold, records are classified as "at risk" or "safe". Finally, protection methods are applied in order to reduce the risk of each record under the threshold. Only protection methods based on data reduction are considered, particularly "global recoding" and "local suppression" (Willenborg and de Waal, 2001). Data reduction implies loss of information with respect to the original contents of the file. Therefore, main purpose is combining protection methods in order to minimise loss of information, given that the fixed level of risk is respected.

As a case study, we consider provisional data from the Italian EU-SILC survey. This is a panel survey, carried out in different EU member states, and providing every year cross-sectional and longitudinal data on income, poverty, social exclusion and living conditions. Information is collected about both households and household members at the same time.

In order to measure the re-identification risk for EU-SILC micro data we need to keep into account (i) the hierarchical data structure (individuals belonging to the same household can be associated), and (ii) the longitudinal data structure (Abowd and Woodcock, 2000). The re-identification risk has to be measured under an appropriate disclosure scenario, namely the quality and quantity of the information assumed to be available to the intruder and the strategy to re-identify a statistical unit.

In this paper we discuss the disclosure figures belonging to the EU-SILC survey and propose a strategy to produce a MFR from the Italian sample. In particular, in Section 2, we briefly describe the EU-SILC survey, and address the main disclosive figures it implies: household data structure and longitudinal data structure. In Section 3, we describe some intruder's attack scenarios worth of consideration in Italy. In Section 4, we introduce the SDC method we apply, in order to estimate household and individual re-identification risk. In Section 5, we show the empirical results, and propose a strategy for protecting the EU-SILC micro data. Finally, in Section 6, conclusions are discussed.

2. The EU Survey on Income and Living Conditions (EU-SILC)

The EU-SILC data are organized into four datasets: (i) the *Household Register* file, containing information about every sampled household; (ii) the *Household Data* file, containing information about each interviewed household; (iii) *Personal Register* file, containing information about every household member; (iv) *Personal Data* file, containing information about each interviewed household member. These files can be linked together, through country, household and individual identification codes. It is worth noting that household and individual files can be linked also longitudinally, so that the amount of information increases yearly.

When data are organized in a hierarchical structure, that is households and household members are explicitly linked through identification codes, the following issues have to be considered: (i) household characteristics might be used for identifying an individual (for instance, a household composed by 15 individuals might be very rare); (ii) household members' characteristics might be used for identifying a household or other household members (as an example a widow 18 years old might be easily re-identified, for being rare or unique in the population, and probably the household would be easily identified as well).

In the Italian case, cross-sectional and longitudinal data belong to the same sample of households. As these files might be easily linked, the same protection criteria have to be applied simultaneously to both data set. However, some anonymisation criteria often applied to cross-sectional data, might not be used when dealing with longitudinal data. An example is provided by the aggregation of age in classes (as soon as an individual moves form one class to the next one, the exact age of the individual might be easily deduced). If a variable is protected in a specific record and at a given year of survey, the same treatment has to be coherently applied in the following years. Thus, if local suppression is chosen, the suppression of a variable for some records during the whole period of observation should be applied. Similarly, when perturbing the value of a variable, we have to be aware of the consequences on the analyses of these variables over the period of observation. Thus, when dealing with longitudinal data, using local suppression or perturbation methods might be not convenient.

3. The intruder's scenarios: available information and attack strategy

A scenario synthetically describes (i) which is the information potentially available to the intruder, and (ii) how the intruder would use such information to identify an individual: i.e. the intruder's attack means and strategy. We refer to the information available to the intruder as an *External Archive*, where information is provided jointly with directly identifying data (name, surname, etc.). The anonymised data set is called the *EU-SILC user data base*.

In the *Nosy Neighbour Scenario*, we assume that the intruder has many information about a single (or few) individual, and the information is based on personal knowledge. We are not able to know how many and which individual or household characteristics the intruder knows, but we assume that he/she does not know that the individual is in the data set we want to protect. The intruder's attack would be the spontaneous recognition of the individual in the EU-SILC user data base. For protecting the micro data against this kind of attack we propose to reduce the information in a way that the intruder can not be confident that a given combination of information is unique or rare in the population.

The *Individual Archive Scenario* is based on the assumption that the external archive available to the intruder provides individual directly identifying variables, and some other variables. Some of these latter are assumed to be available also in the data set we want to protect. The intruder's strategy would be matching the information in the individual archive with that in the EU-SILC user data base. A "match" would be considered only if all the matching variables assume the same value in both data sets. We refer to these matching variables as *key* or *identifying variables*. The individual archive we consider to be worth of attention is represented by the *Electoral Registers*. These are based on the population register, and provide information about individuals having electoral right. Electoral registers are public and provide: place of birth, place of residence, date of birth, sex, marital status, occupation and educational level. We suggest to drop information about the place of birth and recode the current place of residence at least at regional level. Date of birth is reliable, and has a strongly identifying power. Thus we recommend to reduce the information, through recoding it in age. Marital status is usually considered as reliable but it is not public any longer. We will discuss alternative scenarios that respectively include it or exclude it as a key variable. We do not consider occupation and education reliable in this scenario, but recoding of these variables is suggested under the nosy neighbour scenario.

In the *Household Archive Scenario*, we assume the intruder would use as matching variables not only individual characteristics, but also household characteristics. We also assume the external archive to have the same structure of the EU-SILC user data base: each record in the file represents a single individual and a household identifier is associated to each record, allowing for household recognition. The external archive we consider is the *Population Register*. This is not public, but an individual might ask for information about one or a few households. The intruder's chance of access to the external archive is lower than in the previous case. Information provided by the population register are the same individual variables provided by the electoral register, as well as (i) the household size and (ii) the parental relationship (we use six categories coherently with the information provided by the survey).

Longitudinal data structures provide a same set of identification variables several times, in a given period of observation (say four years in EU-SILC survey). An intruder might use the specific key variables pattern of change in order to identify individuals. Rare patterns of change might ease individual spontaneous recognition. Nevertheless, as in the nosy neighbour scenario, we assume that once extremely identifying variables are properly recoded, the intruder should not be able to know whether the pattern of change is unique (or rare) in the population. Concerning the individual or household archive scenario under a longitudinal prospective, we assume that the intruder is not likely to have access to external archives several times, and at the same reference periods of the survey. Anyway, further studies might be worth, considering particularly electoral registers for been more easily accessible. However, these provide key variables that are not expected to change, apart from the case of the place of residence and the marital status.

4. The individual and household risk of re-identification

We consider a measure of the disclosure risk based on a probabilistic estimation of the individual re-identification risk (Franconi and Polettini, 2004). The individual approach allows to apply protection methods only to those records that present a risk higher than a pre-fixed threshold 'α'. Protection methods taken into account are mainly global recoding and local suppression. Usually a preliminary step of global recoding is used in order to reduce the number of suppressions to an acceptable level. The idea behind the method is that a statistical unit, represented by a combination of identification variables, is "at risk" if the same combination is rare in the population. Thus, the relation between the frequency of a combination of identification variables in the sample data and the frequency of the same combination in the population is considered. The true value of the latter is often unknown, and consequently we estimate it using sampling information (i.e. sample weights).

The individual approach has been implemented in Argus (available at http://neon.vb.cbs.nl/casc/), allowing for two alternative risk computations (Polettini and Seri, 2003): *Base Individual Risk* (BIR) and *Base Household Risk* (BHR). The former is based on the approach just described. The latter is intended to be computed when data structure is characterised by a household identifier associated to each record allowing for household recognition. BHR estimate, as implemented in Argus, is based on the individual risk assuming that, if an individual is correctly linked and identified, all household components might be identified as well. The value of BHR is the same for each household member. In order to apply local suppression to the records at risk (BHR>α) we consider that: let hhs be the household size, if BIR is lower than α/hhs for all the members of the household then BHR is lower than α. Thus, we only apply local suppression to records showing a BIR higher then α/hhs. In other words, the higher the household size, the lower is the threshold considered. This certainly represents and advantage because higher levels of safety are asked for larger households.

5. Empirical results

So far, only the first wave (2004) of the EU-SILC survey has been carried out in Italy. Consequently, we cannot empirically address the disclosive features implied by the longitudinal data structure. The following analyses are based on provisional data, organized in 61750 individual records. A first step to reduce the disclosure risk consists in dropping or recoding identifying variables (see Table 1).

Table 1. EU-SILC variables to be recoded or dropped

		Dropped	*Recoded*
Sample Variables	Primary Strata	X	
	Psu-1 (First Stage)	X	
	Psu-2 (Second Stage)	X	
	Order Of Selection Of Psu	X	
Individual Variables	Month Of Birth	X	
	Year Of Birth		X
	Month Moved Out The Household Or Died	X	
	Month Moved In The Household	X	
	Day Of The Personal Interview	X	
	Month Of The Personal Interview	X	
	Citizenship 1		X
	Citizenship 2	X	
	Highest Isced Level Attained		X
	Parental Relationship		X
Household Variables	Day Of Household Interview	X	
	Month Of Household Interview	X	
	Number Of Rooms Available To The Household		X
	Place Of Residence		X

In a second step, we estimate the individual and household risk of re-identification to evaluate the number of suppressions needed respectively under the individual and household archive scenario. Provided that key variables might be recoded according to different levels of aggregation, we propose and discuss some alternative solutions.

The individual archive scenario has been defined through the following key variables: sex, age, and place of residence. *Age* is recoded according to two standards: (i) top coded at 85 years (Age85); (ii) top coded at 85 years and simultaneously recoded (0-2, 3-5, 6-10) on the base of the first levels of the educational system (Age85_edu). The *place of residence* is recoded according to two standards: (i) Region: 19 modalities; (ii) Macro region: 11 modalities. The four alternative solutions are tested including *marital status* as a key variable (Mar Stat).

The threshold is fixed at 0.01, that is an individual is considered as "safe" when in the population there are at least other 100 individuals showing the same combination of key variables. Records at risk are treated through local suppression. Table 2 shows the distribution of suppressions and the maximum of individual risk, by "solution" and key variable.

Results show that when considering Sex, Age85 and Macro Region as key variables (solution (1) and (3)), all individuals have a risk of identification lower than 0.01 (i.e. no suppressions have to be applied). In contrast, when marital status is added (solution (2) and (4)), 192 individuals have a risk of identification higher than the threshold. Most of the suppressions are applied to widow and divorced individuals, or to never married individuals but older age. When the region is considered as a key variable, instead of macro region, we notice that the number of suppressions is still low if marital status is disregarded (solution (5) and (7)) and the maximum risk is not extremely high (0.036). In both cases, if marital status is added as a key variable (solution (6) and (8)), the number of suppressions increases as well as the maximum risk.

Table 2. Individual archive scenario: distribution of suppressions by solution and key variable (threshold fixed at 0.01).

Solutions	*Sex*	*Age85*	*Age85_edu*	*Region*	*Macro Region*	*Mar Stat*	*Ind. at risk*	*Suppressions*	*Max Ind. Risk*
(1)	0	0	----	----	0	----	0	0	0.008
(2)	0	0	----	----	0	192	192	192	0.078
(3)	0	----	0	----	0	----	0	0	0.008
(4)	0	----	0	----	0	192	192	192	0.078
(5)	7	1	----	0	----	----	8	8	0.036
(6)	7	1	----	0	----	598	606	606	0.093
(7)	7	----	0	0	----	----	7	7	0.036
(8)	7	----	0	0	----	598	605	605	0.093

As far as the household archive scenario is concerned, the same variables as in the previous scenario are considered, and *parental relationship* and the *household size* are included (named respectively Rel Par and HHsize). Under this scenario, the threshold is fixed at 0.04. It is higher than in the previous case because the population register (i.e. the intruder external archive) is not public. Thus, the intruder is not likely to have access to it for a region (or macro region).

Results of three different combinations of key variables are shown in Table 3. We firstly consider Sex, Age85, Macro Region, Mar Stat, HHsize and Rel Par as key variables. Households estimated as at risk are 531, consequently 3041 suppressions are applied. Clearly, in some unsafe households there are more than one individual showing a risk higher than the threshold divided by household size. Thus, more than one suppression per household is applied. The second solution (2) shows that substituting Age85 with Age85_edu the information loss due to the aggregation of some ages in classes is not compensated by a significant reduction of suppressions. In solution (3) we use Age85 and Region instead of Macro Region, increasing the geographical information. As a consequence, the number of households at risk and of suppressions increases. Comparing these solutions, we notice that the information loss due to the use of macro region instead of region actually strongly reduces the risk of identification, and the number of suppressions. Households of bigger size are more protected (see Table 4).

Table 3. Household archive scenario: distribution of suppressions by solution and key variable (threshold fixed at 0.04).

Solutions	*Sex*	*Age85*	*Age85_edu*	*Region*	*Macro Region*	*Mar Stat*	*Hous. Size*	*Rel Par*	*Hous. at risk*	*Suppr.*	*Max Hous. Risk*
(1)	259	1131	----	----	0	720	0	931	531	3041	0.51
(2)	217	----	1038	----	0	720	0	864	490	2839	0.51
(3)	485	2039	----	2	----	745	0	1162	828	4433	0.51

Table 4. Suppressions by household size, Solution (1).

Household size	1	2	3	4	5	6	7	8	9	10	11	Total
Number of records with suppression	32	192	414	621	613	559	302	60	39	16	22	2870
Number of records in the sample	6342	13644	15612	18143	5840	1501	483	88	45	30	22	61750
% of records with suppression	0.5	1.4	2.7	3.4	10.5	37.2	62.5	68.2	86.7	53.3	100	

According to the analyses carried out, a proposal to protect the EU-SILC micro data can be exploited as follows: (i) variables in Table 1 are dropped or recoded; (ii) age is top-coded at 85 years, and place of residence in 11 Macro Regions according to NUTS nomenclature; (iii) BIR is estimated, and no suppressions have to be applied (solution (1) in Table 2); (iv) BHR is estimated and key variables are suppressed for individuals belonging to households at risk (solution (1), Table 3).

6. Conclusions

In this work we propose an approach to define a MFR from provisional EU-SILC micro data. Particularly we highlight statistical disclosure control problems when dealing with data presenting both hierarchical and longitudinal structure, as is the case of the EU-SILC micro data. We described the individual approach to the risk of disclosure based on a probabilistic estimate of the re-identification risk. The risk of disclosure has been analysed, taking into account some disclosure scenarios in the Italian context. In particular, we consider (i) a nosy neighbour scenario where disclosure is possible by spontaneous recognition, and (ii) two scenarios where re-identification may arise by record linkage techniques (individual and household archive scenarios). In this last cases re-identification risks can be estimated and a threshold can be fixed in order to classify record “at risk” or “safe”. Consequently, protection method can be applied in order to minimise information loss, guaranteeing the respect of the fixed acceptable level of risk.

We argue that high levels of risk are estimated when both hierarchical and longitudinal structure of data are taken into account. However, we observe that the disclosure scenario for such a situation may occur rarely, because of the low chance to access reliable external archive with household information and in different points in time, coherently with the observation period of the survey. Nevertheless, analyses on re-identification risk may be conducted when other waves of the survey will be available. Suitable disclosure scenarios taking into account longitudinal structure of the data may be defined at least on the basis of the individual scenario. At this stage, we suggest to consider these aspects mainly under the nosy neighbour scenario. A MFR can be proposed on the basis of the experimental results presented in Section 5, provided that recoding and dropping of variables reported in Table 1 are applied. Furthermore, we recommend, when releasing the EU-SILC user data base, to ask the researcher for signing an agreement, in order to guarantee the data protection on legal basis too.

Acknowledgements

We thank Luisa Franconi e Maurizio Lucarelli for helpful comments and suggestions. The views expressed are those of the authors and may not reflect the policies of Istat.

References

Abowd, J. M. & Woodcock, S. D. (2000). *Disclosure Limitation in Longitudinal Linked Data.* In P. Doyle, J. I. Lane, J. J. Theeuwes and L. V. Zayatz (Eds), Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies 215-278. North-Holland.

Franconi, L. & Polettini, S. (2004). Individual Risk Estimation in μ-Argus: A Review. In J. Domingo-Ferrer and V. Torra (Eds), Privacy in Statistical Database. *Springer-Verlag*, Berlin Heidelberg 2004.

Polettini, S.& Seri, G. (2003). *Guidelines for the Protection of Social Micro-data Using Individual Risk Methodology: Application with* μ*–Argus Version 3.2.* CASC-Computational Aspects of Statistical Confidentiality Deliverable No: 1.2-D3.

Willenborg, L. & de Waal, T. (2001). Elements of Statistical Disclosure Control. Springer, New York.

A ranking approach to confidentiality in survey data

Johan Heldal
Statistics Norway, P.O. Box 8131 Dep, N-0033 Oslo, Norway

Abstract. This paper suggests a method for confidentiality protection of datasets when (continuous) numeric identifying variables have been linked to the dataset by exact matching with registers. Such variables may be highly identifying, in particular if the register is publicly available. The method, in this paper called *rank matching*, takes advantage of the registers themselves to mask the original data and eliminate the confidentiality hazard.

Keywords. *Superpopulation distribution, information loss, density estimation, rank matching, disclosure, re-identification probability.*

1. Introduction

In some countries, the national statistical office includes some variables in surveys not by asking the respondents, but by exact matching of information from registers files comprising the entire population. Such data, for instance income from tax registers, are often of high quality and may be of high value both for researchers and for an intruder trying to disclose the identity of a statistical unit, in particular if the register itself is available to him or her. Data from registers has raised concern in the context of providing anonymous datasets to researchers under EU Regulation 831/2002. The EU-SILC Anonymisation Task Force Report (Museux 2005) writes:

> ”*The TF has pointed the specificities of so called register countries. For these countries, some of the income variables available in the EU-SILC may come directly from registers (DK, NO, SE, FI, LT, LI, CZ, SI, IS). If this register information together with direct identifiers is available to external users, the risk of disclosure is greatly increased. This specific issue should be carefully studied. A specific section of this report is dedicated to it.* ”

This is however also a situation that opens an opportunity to apply disclosure control methods which are not otherwise available.

This paper is basically a representation of an idea called *rank matching* (rm) earlier presented in Fosen and Heldal (2001) and Heldal (2001), an idea that Statistics Norway now wishes to follow up in the context of EU-SILC. Carlson and Salabasis (2002) have (independently) worked on the same idea and in greater detail using theory of order statistics. Because of space constraints I refer to these papers for study of the statistical properties of the method from a user viewpoint. This paper will concentrate on some intruder scenarios associated with the method.

Section 2 outlines the ideas behind rank matching. In section 3 simulations and small examples are used to discuss some intruder scenarios. More work is needed to study the scenarios in more realistic settings.

2. Basic ideas

Consider a finite population $\boldsymbol{\mathcal{U}}$ consisting of units $u_1,\dots,u_N$ indexed by a variable *j*. To each unit a vector $\boldsymbol{X}_j^T = (X_{j1},\dots,X_{jK})$ of absolutely continuous numeric variables is attached that for the moment can be considered as generated by a (cumulative) superpopulation distribution $F(\boldsymbol{x})$. The $N\times K$ matrix $\mathbb{X} = (\boldsymbol{X}_j^T, j \in \boldsymbol{\mathcal{U}})$ is termed a *register*. Examples of numeric register variables are income from the tax assessment and age of individuals. Turnover or other economic variables stored in business registers are other examples.

A sample $\boldsymbol{s}$ of size n is drawn from the finite population $\boldsymbol{\mathcal{U}}$ with some sampling design $p(\boldsymbol{s})$. $\boldsymbol{s}$ gives rise to a dataset matrix $\mathbf{X} = (\boldsymbol{X}_j^T, j \in \boldsymbol{s})$. The joint non-singular density of $\boldsymbol{X}_j$ is called $f(\boldsymbol{x})$. To keep concepts as simple as possible, assume that $p(\boldsymbol{s})$ is simple random so that the $\boldsymbol{X}_j$'s are identically distributed also in the sample.

Let R_{jk} be the rank of the observed value X_{jk} in the j-th column of $\mathbb{X}$ where $R_{jk} \in \{1, \dots ,N\}$. Further, let $\boldsymbol{R}_j^T = (R_{j1},\dots,R_{jK})$ and $\mathbb{R} = (\boldsymbol{R}_j^T, j \in \mathcal{U})$ the $N \times K$ rank matrix corresponding to the register $\mathbb{X}$. Let i index the sample units and j_i (stochastic) be the population index (label) of sample unit i and let $\boldsymbol{x}_i = (x_{i1},\dots,x_{ik}) = \boldsymbol{X}_{j_i}$. Let $r_{jk} \in \{1,\dots,n\}$ be the rank of x_{ik} in $\mathbf{X}$, $\boldsymbol{r}_i = (r_{i1},\dots,r_{iK})^T$ and let $\mathbf{R} = (\boldsymbol{r}_1^T,\dots,\boldsymbol{r}_n^T)$ be the sample rank matrix. The latter should be distinguished from $\mathbb{R}_s = (\boldsymbol{R}_j^T, j \in \boldsymbol{s})$ which contains the population ranks for the sample units. The continuity assumption guarantees uniqueness of the ranks. For the mapping from the ranks to the labels the (somewhat simplified) notation $i= (r_{ik})$ and $j = (R_{jk})$ is being used, k=1, ... , K.

Rank matching (rm) now goes as follows: Draw a new sample $\boldsymbol{s}_2$ independently of $\boldsymbol{s}$, according to the same design and sample size as $\boldsymbol{s}$. $\boldsymbol{s}_2$ gives rise to a new sample dataset $\mathbf{X}^{(2)} = (\boldsymbol{X}_j^T, j \in \boldsymbol{s}_2) = (x_1^{(2)T},\dots,x_n^{(2)T})$ with the same variables as before and generated by the same superpopulation distribution F. Replace $x_{ik} (= x_{(R_{ik})k})$ in the original sample with the value $x^{(2)}_{(R_{ik})k}$ having the same rank on the same variable in $\mathbf{X}^{(2)}$. This produces a synthetic dataset $\mathbf{X}^*$ with rows $x_i^{*T} = \boldsymbol{x}^{(2)T}_{(\boldsymbol{R}_i)k} = (x^{(2)}_{(R_{i1})1},\dots,x^{(2)}_{(R_{iK})K})$. This version will be called *joint* rm. The distribution of $\boldsymbol{x}_i^*$ will depend on the original $\boldsymbol{x}_i$ only through its rank vector $\boldsymbol{r}_i$. The method can be applied for an entire sample or within strata or domain. It is also possible to draw one sample for each variable. This is computationally more intensive, but analytically somewhat simpler to deal with. This will be called *independent* rank matching. It will however not matter much from the point of view of the intruder (see section 3).

Generally there is information loss associated with rm. In the original dataset, the marginal distribution of x_{ik} given $\boldsymbol{r}_i$ can depend on components of $\boldsymbol{r}_i$ other than r_{ik}. In other words, for independent rm,

$$F_k^*(x_k \mid \boldsymbol{r}) = F_k^*(x_k \mid r_k) = F_k(x_k \mid r_k) \neq F_k(x_k \mid \boldsymbol{r})$$

For joint rm the first equality will not be exact.

An option to rank matching with register is rank *swapping* (rs) and related methods (Moore 1996). Contrary to rm, rs preserves exactly the observed marginal distributions of all variables as in the original dataset, but not the exact multivariate rank order structure. Simulations and theoretical considerations in Carlson and Salabasis (2002) based on correlations between normally distributed variables indicate that attenuation due to independent rm is slightly larger than due to joint rm. The attenuation is larger for smaller sample sizes than for big ones. A proper comparison to various versions of rs remains, but Heldal (2001) indicates that simple half sample rs is inferior.

A similar approach can be attempted on discrete ordinal variables. Then an artificial ordering must be introduced between units having the same value on the discrete variable. Care must be taken to avoid illegal edits. Some variables, and income components in particular, take both discrete and continuous values. For many income components, there are typically many zero values and otherwise positive values. Improper ordering of the original zeroes in $\mathbf{X}$ can easily introduce positive values on units that according to the values of other variables cannot be positive, like giving a 20 year old man retirement pension. Detailed discussions about how to handle discrete values will not be given here.

In most cases, the variables available from registers only make up some of the variables in a survey dataset. Non-register variables, usually collected in the survey are not affected by the register rank matching, but may be target variables for a disclosure attempt.

3. Inference about population units

While information loss should be considered at superpopulation level, probability of disclosure is definitely a finite population matter. The samples $\boldsymbol{s}$ and $\boldsymbol{s}_2$ are (simple random) samples from a labelled set of units $\boldsymbol{\mathcal{U}}$ to which realisations from the population distribution F have been associated. Identity disclosure is inference about the label in this finite population. Such inference is possible only when someone with access to the dataset **X** has information about some of these variable values associated to given labels.

It is clear that an intruder having accurate information about the value of at least one absolutely continuous numeric variable for some unit drawn to the sample will be able to identify that unit. If intruder's information or the measured values of the variables in the sample is not quite accurate, inference about a label can never the less very often be done with high degree of confidence.

Question 1: Which information on labels associated with the units in $\boldsymbol{s}$ is still present in the rank matched dataset $\mathbf{X}^*$?

Question 2: How can an intruder make use of this information to make a disclosure?

The answer to these questions will depend on the intruder scenario. Two worst-case scenarios will be discussed:

a. The intruder knows that some members in her Identification File (IF) are in $\boldsymbol{s}$ and their true values on some X_{ij}.

b. The intruder has access to the entire population register, but does not know which units were drawn to $\boldsymbol{s}$.

Case a will be studied in a simulation experiment presented in section 3.1. This shows that with an increasing number of variables available for disclosure the probability of doing correct identification using distance techniques increases rapidly. Case b will be illustrated with an example in section 3.2. This is an extreme case, but is interesting. Someone having access to the entire register can extract its rank structure (population ranks) and from that identify all possible samples whose sample rank matrix $\mathbf{R}$ equal the rank matrix of $\mathbf{X}$. On such a basis the probability that an individual with a given sample rank vector corresponds to a given population unit can be computed exactly for every records in the sample.

3.1. Situation a, a simulated intrusion

With what confidence can an intruder identify the original record number associated with the synthetic record $\boldsymbol{x}^*$? Assume that the intruder in her identification file has access to an original record $\boldsymbol{x}$ from **X** and knows that the owner of $\boldsymbol{x}$ is in **X**. To disclose the corresponding record in $\mathbf{X}^*$ (and $\mathbf{X}^+$), she uses discriminant analysis and decides for the following decision rule: Choose the record $\boldsymbol{x}_i^*$ in $\mathbf{X}^*$ that minimizes a distance

$$\left\|\boldsymbol{x}-\boldsymbol{x}_i^*\right\|_{\mathbf{W}}^2=(\boldsymbol{x}-\boldsymbol{x}_i^*)'\mathbf{W}(\boldsymbol{x}-\boldsymbol{x}_i^*)\,. \qquad (3.1)$$

A thorough discussion of the use of discriminant analysis in the context of disclosure control is given in Paaß and Wauschkuhn (1985). In order to test the capacity of this decision rule, **W** was taken as the inverse of the diagonal of $\hat{\Sigma}^*$ and $\hat{\Sigma}^+$, the obvious estimates of the covariance matrices based on $\mathbf{X}^*$ and $\mathbf{X}^+$. All 63 possible combinations of one to six variables were tested and the number of correct hits recorded. The results are summarized in table 1.

Table 1. Minimum and maximum numbers of correct identifications of records in $\mathbf{X}^*$ (rm rows) and $\mathbf{X}^+$ (rs rows) with various numbers of identification variables.

The number of variables used	Number of correct hits		The number of variables used	The number of variables used	
One (of 6 variables)	rm	6-41	Four (15 combs.)	rm	845-989
	rs	0		rs	722-945
Two (of 15 pairs)	rm	137-545	Five (6 combs.)	rm	983-996
	rs	93-321		rs	924-981
Three (20 triples)	rm	472-933	Six (1 comb.)	rm	996
	rs	244-720		rs	987

Table 2 shows that the identifying capacity of combinations of variables increases rapidly with the number of variables available for disclosure for both methods. This is no surprise. The number of correct identifications with the same number of variables shows large variations. The tendency is, as expected, that among the combinations with the same number of variables, those showing higher correlations produce the smallest number of correct hits and vice versa. The results in table 2 may seem discouraging. But this was for an intruder knowing that the target is there. An intruder not knowing that the target unit is in the dataset will need to verify that. For some discussion of that case, see Heldal (2001).

3.2. Case b.

Consider an intruder with access to a population register $\mathbb{X}$ described in the beginning of section 2. This intruder can extract the population rank matrix $\mathbb{R} = (\boldsymbol{R}_1, \dots , \boldsymbol{R}_N)^T = \rho(\mathbb{X})$ from $\mathbb{X}$. Without loss of generality we can take the ranks in the first column of $\mathbb{R}$ and $\mathbf{R}$ as population and sample labels, setting $R_{j1} = j$ and $r_{i1} = i$. Let j_i be the stochastic variable that maps sample label i to a population label. The intruder observes $\mathbf{X}^*$ and $\mathbb{X}$ and wishes to calculate $P(j_i = j \mid \mathbf{X}^*, \mathbb{X})$ for all i and all $j \in \boldsymbol{\mathcal{U}}$. With a little algebra we prove that $\mathbf{R}$ and $\mathbb{R}$ are sufficient for the intruders inference.

$$P(j_i = j \mid \mathbf{X}^*, \mathbb{X}) = P(j_i = j \mid \mathbf{R}, \mathbb{R})$$

The sample version of $\mathbb{R}$, $\mathbb{R}_s = (\boldsymbol{R}_j^T, j\in \boldsymbol{s})$, will not be directly observable in the sample. Never the less, there is a 1-1 correspondence between the sample space $\mathcal{S}$ and $\{\mathbb{R}_s : s\in \mathcal{S}\}$. $\mathbb{R}_s$ uniquely determines $\mathbf{R} = \rho(\mathbb{R}_s) = \rho(\mathbf{X}) = \rho(\mathbf{X}^*)$, and the structure of $\mathbb{R}$ determines the probability structure of $\mathbf{R}$. There are $(n!)^{K-1}$ possible (unordered) sample rank matrices $\mathbf{R}$. They define a partition of $\mathcal{S}$ into disjoint subsets $\mathcal{S}_{\mathbf{R}}$, some of which may be empty by the configuration of $\mathbb{R}$. If for an observed matrix $\mathbf{R}$, $\mathcal{S}_{\mathbf{R}}$ is identified, then the probability $P(j_i = j \mid \mathcal{S}_{\mathbf{R}})$ that a given sample unit i corresponds to a given population unit j can be calculated exactly. However, it does not seem to be feasible to do this by formula except when $K = 1$. For large N and n efficient algorithms will be necessary to identify $\mathcal{S}_{\mathbf{R}}$.

Example: Assume $N = 7$, $K = 1$ and $n = 3$. Then $\mathbb{R} = [1, 2, 3, 4, 5, 6, 7]^T$ and $\mathbf{R} = [1, 2, 3]^T$. Then

$$P(j_i = j \mid \mathcal{S}_{\mathbf{R}}) = \binom{j-1}{i-1}\binom{N-j}{n-i}\Big/\binom{N}{n} = \binom{j-1}{i-1}\binom{7-j}{3-i}\Big/35$$

Table 2. Tabulation of the distribution of $P(j_i{=}j \mid \mathbf{R}, \mathbb{R})$

i\j	1	2	3	4	5	6	7
1	15/35	10/35	6/35	3/35	1/35	0	0
2	0	5/35	8/35	9/35	8/35	5/35	0
3	0	0	1/35	3/35	6/35	10/35	15/35

Assume $K = 2$ and that the population rank matrix is

$$\mathbb{R} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 4 & 5 & 2 & 3 & 1 & 7 & 6 \end{bmatrix}^T,$$

The sample space still consists of 35 samples. Now there are 6 possible sample rank matrices **R**. The 6 sample rank matrices, their associated partition sets and the probabilities $p(j \mid i) = P(j_i = j \mid \mathcal{S}_\mathbf{R})$ are given in table 3. The table shows large variation of the number of samples in each partition. The cases where $p(j \mid i) = 1$ define identity disclosure with probability one. This occurs for at least one unit in eleven samples in three partition subsets, meaning that before sampling the probability of a disclosure producing dataset is 11/35.

Table 3. The partition of $\mathcal{S}$ generated by the sample rank matrices **R** and the induced identification probabilities and disclosure probabilities given **R**. * marked **R**s generate some certain disclosures, indicated by italic sample and population labels.

$\mathbf{R}^T$	$\{\mathbb{R}_s^T : s \in \mathcal{S}_\mathbf{R}\}$	$p(j \mid i) = P(j_i = j \mid \mathbf{R})$
$\begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}$	$\begin{bmatrix} 1 & 2 & 6 \\ 4 & 5 & 7 \end{bmatrix}, \begin{bmatrix} 1 & 2 & 7 \\ 4 & 5 & 6 \end{bmatrix}, \begin{bmatrix} 3 & 4 & 6 \\ 2 & 3 & 7 \end{bmatrix}, \begin{bmatrix} 3 & 4 & 7 \\ 2 & 3 & 6 \end{bmatrix}$	$p(1 \mid 1) = p(3 \mid 1) = 1/2$ $p(2 \mid 2) = p(4 \mid 2) = 1/2$ $p(6 \mid 3) = p(7 \mid 3) = 1/2$
$\begin{pmatrix} 1 & \mathit{2} & \mathit{3} \\ 3 & 1 & 2 \end{pmatrix}*$	$\begin{bmatrix} 1 & \mathit{3} & \mathit{4} \\ 4 & 2 & 3 \end{bmatrix}, \begin{bmatrix} 2 & \mathit{3} & \mathit{4} \\ 5 & 2 & 3 \end{bmatrix}$	$p(1 \mid 1) = p(2 \mid 1) = 1/2$ $p(\mathit{3} \mid \mathit{2}) = p(\mathit{4} \mid \mathit{3}) = \mathbf{1}$
$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}$	$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 2 \end{bmatrix}, \begin{bmatrix} 1 & 2 & 4 \\ 4 & 5 & 3 \end{bmatrix}, \begin{bmatrix} 1 & 2 & 5 \\ 4 & 5 & 1 \end{bmatrix}, \begin{bmatrix} 3 & 4 & 5 \\ 2 & 3 & 1 \end{bmatrix}$	$p(1 \mid 1) = 3/4, p(3 \mid 1) = 1/4$ $p(2 \mid 2) = 3/4, p(4 \mid 2) = 1/4$ $p(3 \mid 3) = p(4 \mid 3) = 1/4, p(5 \mid 3) = 1/2$
$\begin{pmatrix} 1 & \mathit{2} & \mathit{3} \\ 1 & 3 & 2 \end{pmatrix}*$	$\begin{bmatrix} 1 & \mathit{6} & \mathit{7} \\ 4 & 7 & 6 \end{bmatrix}, \begin{bmatrix} 2 & \mathit{6} & \mathit{7} \\ 5 & 7 & 6 \end{bmatrix}, \begin{bmatrix} 3 & \mathit{6} & \mathit{7} \\ 2 & 7 & 6 \end{bmatrix},$ $\begin{bmatrix} 4 & \mathit{6} & \mathit{7} \\ 3 & 7 & 6 \end{bmatrix}, \begin{bmatrix} 5 & \mathit{6} & \mathit{7} \\ 1 & 7 & 6 \end{bmatrix}$	$p(1 \mid 1) = p(2 \mid 1) = p(3 \mid 1)$ $= p(4 \mid 1) = p(5 \mid 1) = 1/5$ $p(\mathit{6} \mid \mathit{2}) = p(\mathit{7} \mid \mathit{3}) = \mathbf{1}$
$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}$	$\begin{bmatrix} 1 & 3 & 6 \\ 4 & 2 & 7 \end{bmatrix}, \begin{bmatrix} 1 & 3 & 7 \\ 4 & 2 & 6 \end{bmatrix}, \begin{bmatrix} 1 & 4 & 6 \\ 4 & 3 & 7 \end{bmatrix}, \begin{bmatrix} 1 & 4 & 7 \\ 4 & 3 & 6 \end{bmatrix}$ $\begin{bmatrix} 1 & 5 & 6 \\ 4 & 1 & 7 \end{bmatrix}, \begin{bmatrix} 1 & 5 & 7 \\ 4 & 1 & 6 \end{bmatrix}, \begin{bmatrix} 2 & 3 & 6 \\ 5 & 2 & 7 \end{bmatrix}, \begin{bmatrix} 2 & 3 & 7 \\ 5 & 2 & 6 \end{bmatrix}$ $\begin{bmatrix} 2 & 4 & 6 \\ 5 & 3 & 7 \end{bmatrix}, \begin{bmatrix} 2 & 4 & 7 \\ 5 & 3 & 6 \end{bmatrix}, \begin{bmatrix} 2 & 5 & 6 \\ 5 & 1 & 7 \end{bmatrix}, \begin{bmatrix} 2 & 5 & 7 \\ 5 & 1 & 6 \end{bmatrix}$ $\begin{bmatrix} 3 & 5 & 6 \\ 2 & 1 & 7 \end{bmatrix}, \begin{bmatrix} 3 & 5 & 7 \\ 2 & 1 & 6 \end{bmatrix}, \begin{bmatrix} 4 & 5 & 6 \\ 3 & 1 & 7 \end{bmatrix}, \begin{bmatrix} 4 & 5 & 7 \\ 3 & 1 & 6 \end{bmatrix}$	$p(1 \mid 1) = p(2 \mid 1) = 3/8$ $p(3 \mid 1) = p(4 \mid 1) = 1/8$ $p(3 \mid 2) = p(4 \mid 2) = 1/4$ $p(5 \mid 2) = 1/2$ $p(6 \mid 3) = p(7 \mid 3) = 1/2$
$\begin{pmatrix} 1 & 2 & \mathit{3} \\ 3 & 2 & 1 \end{pmatrix}*$	$\begin{bmatrix} 1 & 3 & \mathit{5} \\ 4 & 2 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 4 & \mathit{5} \\ 4 & 3 & 1 \end{bmatrix}, \begin{bmatrix} 2 & 3 & \mathit{5} \\ 5 & 2 & 1 \end{bmatrix}, \begin{bmatrix} 2 & 4 & \mathit{5} \\ 5 & 3 & 1 \end{bmatrix}$	$p(1 \mid 1) = p(2 \mid 1) = 1/2$ $p(3 \mid 2) = p(4 \mid 2) = 1/2$ $p(\mathit{5} \mid \mathit{3}) = \mathbf{1}$

4 . Future work

In 2006 the ideas presented in this paper will be attempted on the Norwegian SILC survey will start up in Statistics Norway. An application of this kind will require further work on the method and may require use of other methods as well. Questions related to sample design and rm-domains, balance of information loss versus disclosure risk and data integrity must be addressed. What about discrete or mixed mode variables?

We know that the methods suggested in this paper can be relevant for other countries as well and the so-called ‘register countries' in particular. We wish to do our work in an international context and we hereby invite workers who may be interested in this kind of problems for collaboration.

References

Carlsson, M. and Salabasis, M. (2002): *A data-swapping technique using ranks – A method for disclosure control.* Research in official Statistics, Vol. 4 no. 2 pp 35- 67 (with comment by S.E. Fienberg).

Duncan, G. T. and Lambert, D. (1989): *Risk of Disclosure for Microdata.* J. of Business & Economic Statistics, Vol 7., no. 2 pp 207-217

Fuller, W. A. (1993): *Masking Procedures for Microdata Disclosure Limitation.* Journal of Official Statistics, vol 9 no. 2 pp 383-406.

Hurkens, C.A.J. and Tiourine, S.R. (1998*): Models and Methods for the Microdata Protection Problem.* Journal of Official Statistics, 14, pp 437-447.

Little, R.J.A. (1993): *Statistical Analysis of Masked Data*. Journal of Official Statistics, vol 9 no. 2 pp 407-426.

Moore, R. (1995): *Controlled Data Swapping Techniques For Masking Public Use Data Sets,* U.S. Bureau of the Census, Statistical Research Division Report rr96/04, (available at http://www.census.gov/srd/www/byyear.html).

Museux, J-M. (2005): *EU-SILC anonymisation: Results of the Eurostat Task Force.* UNECE/Eurostat work session in Confidentiality 2005, WP no. 20.

Paaß, G. and Waushkuhn, U. (1985): Datenzugang, Datenschutz und Anonymisierung; Analysepotential und Identifizierbarkeit von Anonymisierten Individualdaten. München: Oldenburg Verlag

Paaß, G. (1988): *Disclosure Risk and Disclosure Avoidance for Microdata*. J. of Business & Economic Statistics, Vol. 6., no. 4 pp 487-500.

Reiss, R.-D. (1989): *Approximate Distributions of Order Statistics.* With applications to Nonparametric Statistics. Springer Verlag.

Skinner, C.J., Marsh, C., Openshaw, S. and Wymer, C. (1994). *Disclosure Control for Census Microdata.* Journal of Official Statistics, 10, pp 31-51.

Strudler, M., Oh, H. L. and Scheuren, F. (1986): *Protection of Taxpayers Confidentiality With Respect to the Tax Model*. Proceedings of the Section on Survey Research Methods, American Statistical Assoc. pp 375-381

Estimated record level risk for the CVTS

Liv Belsby[*] *and Alexander Stuart McAllister*[**]
[*] **Statistics Norway, lbe@ssb.no**
[**] **Eurostat, Alexander.MC-ALLISTER@cec.eu.int**

Abstract: We estimate the record level disclosure risk for the anonymised EU *Continuing Vocational Training Survey*, (CVTS). CVTS covers companies in all the MS and in the EFTA countries. NACE and size group of the company are regarded as identifying variables and these two variables are also in business registers. Consequently, the total number of companies for the combinations of these two variables will be known. Additionally we include the variable *has been involved in a take over or not during the reference year*, i.e., up to three identifying variables. The estimates are the conditional expectations of the inverse of the totals, given the totals in the strata and the sample. Our estimates indicate the data is not sufficiently anonymised.

1. Introduction

The CVTS is one of the four surveys covered by the Commission Regulation no. 831/2002 on "access to confidential data for scientific purposes". Moreover, Article 6 in the Regulation requires" … that the methods of anonymisation applied to these microdata sets minimise in accordance with current best practice the risk of identification of the statistical units concerned, in accordance with Regulation (EC) No 322/97".

This analysis aims to assess the degree of anonymisation that was agreed on between Eurostat and the National Statistical Institutes, The goal of the anonymisation was to produce a Microdata File for Researchers (MFR) and not a public use file.

The approaches to assess the disclosure risk are generally based on estimating the number of "rare" observations with respect to characteristics given in the both the data file and are known for the population.

The disclosure risk is the probability of identifying a company correctly in the dataset. This is often denoted the *record level risk*. The person who attempts to do disclose data is called an *intruder*, see e.g. Benedetti *et al* (2004). By *intruder scenario* we mean the conditions and the type of information under which the identification occurs. We assume that the intruder has available an external database or public registers, e.g. via Internet, with identifiers such as name of the company and other identifying variables which are also in the CVTS dataset. The NACE codes, size of the company measured by number of employees or turnover are examples of identifying variables for companies. The European Business Registers (Internet site www.ebr.org.) is an example of such a register. This register include business registers from Belgium, Austria, Denmark, Estonia, Finland, France, Germany, Greece, Ireland, Italy, Latvia, Norway, Spain and Sweden. The type of information and detail level vary to some extent from country to country.

Furthermore, we assume that the identifying variables in the register or the database are identical to the identifying variables in the CVTS dataset, i.e., that they are reported without measurement error and refer to the same period. In this analysis we apply the common practice of combining up to three identifying variables at one time. This is motivated by the assumption that if the intruder knows more identifying variables, then he or she is one who knows the company and the information we seek to protect.

In the following we will denote the combination of the identifying variables as a *key*. The individual risk is then the probability of linking the company in the register or database correctly with the company in the CVTS file, given the key.

As pointed out by Polettini (2003), the record level risk has the advantage of allowing for selective protection. The estimated disclosure risk for all the companies in the CVTS file gives a detailed picture of the how safe the data is. Additionally the estimated risks indicate which variables should be further recoded or whether some of the observations should be suppressed to avoid disclosure. The record level risk approach is suggested by Benedetti and Franconi (1999) and is implemented in μ-Argus (2002).

On the other hand the *global risk* approach focuses on population uniques, see Bethlehem *et. al.* (1990). Moreover the global risk approach seeks to classify the whole data file as safe or not. For the CVTS data we would like to obtain a more detailed picture. For example we shall try to determine if there is a difference between countries. Both the record level risk approach by Benedetti and Franconi (1999) and the global risk approach by Bethlehem et. al. (1990) include estimation of the population totals, F_k, as well as similar model assumptions, see Rinott (2003) and Skinner *et. al.* (1994) for more details. Moreover Skinner and Elliot (2002) propose a new measure: the probability that a unique match between a microdata record and a population unit is correct.

The CVTS dataset also has a legal protection in that contracts are signed both by Eurostat and the Institution where the researcher is employed. Before the signing of a contract all data providing countries are consulted. Thus the anonymisation intends exclusively to protect against spontaneous recognition. Note that the legislation in Netherlands does not allow business data to be given out as MFR.

2. The data

All MS and EFTA countries provide data to Eurostat. The sampling unit is the company. The strata are defined by NACE and *size*, i.e., how many employees the company has. The size variable is coded into the groups of 10-49, 50-249 and 250 and more employees. The NACE variable was collected with four digits NACE-code, but anonymised by recoding it into 20 groups. See Appendix A for more details. The data, which we base the analysis on, was sampled in 2000/2001 with reference year 1999. The CVTS data from this year is called CVTS2.

The sampling fraction is the same within one stratum, but varies for the different strata. The number of companies in the strata is known from registers. This will be utilized in the estimation.

Table 1. The structure of the CVTS data

NACE	*Size group*	*Stratum*	*Take over or not*	*Number of observations*	*Number of observations sampled from the stratum*	*Number in the strata*
Mining	1	1	no	14	15	279
"	1	1	yes	1	15	279
"	2	2	no	7	11	134
"	2	2	yes	4	11	134
"	3	3	no	6	10	79
"	3	3	yes	4	10	79
.	.		.	.	.	.
.	.		.	.	.	.

3. Record level risk and the estimation

We illustrate the record level risk with a simple example: Assume that the intruder finds a company in the file, which he suspects is a company he knows. Furthermore, assume that there are two companies in his register (the population) with the same key as the company, which has caught his interest. We assume that both these two companies are in his register, and have the same probability of being the identical company as in his CVTS file. Consequently, we assume that the probability of linking the company in the CVTS file with the right one in his register is simply ½. More generally, we assume that the record risk is the inverse of the total number in the population with the same key as this record. Often the total number in the population with a certain key k, say F_k, is unknown. Denoting the number in the sample with key k f_k, a common estimate for the record level risk is the conditional expectation $E\left\{\hat{F}_k^{-1} \middle| f_k\right\}$. Benedetti and Franconi (1999) base the estimation on the assumption that the population total with key k, F_k, given f_k is negative-binomial distributed. Moreover different models have been discussed, see Stander (2003).

Bethlehem *et.al.* also assume that the population totals F_k 's are stochastic variables, and that the parameters, say $\Pi_{k,}$ in the distribution for F_k are stochastic. Furthermore, they suggest a gamma distribution for the Π_k. The conditional distribution $F_k | \Pi_k$ is assumed to be a Poisson distribution. Consequently the marginal distribution for F_i is a negative-binomial distribution. Rinott (2003) shows that when the selection probabilities are equal, the model by Benedetti and Franconi (1999) can be regarded as embedded in the model by Bethlehem *et.al.*.

In our estimation approach we will utilize the fact that, as shown in the table above, the total number of companies within each stratum defined by NACE and size denoted by F_s, is known. Furthermore, the two stratification variables are assumed to be identifying variables. The third identifying variable is if variable *has been involved in a take over or not during the reference year 1999*. This means that the key k, will consist of these two stratification variables and a third identifying variable, indicated by $k=\{i,s\}$, where $i=0$ or $i=1$ and where s is the index for the stratum. Consequently, the estimator we use is the conditional expectation, given both the number in the sample with key k, i.e., f_k and the number in the stratum F_s, expressed $E\left\{\hat{F}_k^{-1} | f_k, F_s\right\}$

Given the totals in the strata it is not advantageous to model them as stochastic variables as they are then fixed numbers. This is different from the two approaches described above, and it simplifies the estimation of the record level risk.

We illustrate with an example for $s=1$ in table 1. For $k=\{0,1\}$ and $k=\{1,1\}$, respectively we have that f_{01} equals 14, f_{11} equals = 1, F_1 equals 279 and $f_{01}+f_{11}$ is the number in the sample and equals 15. The possible values for F_{01} are {14, … , 278}. To simplify we rather consider the difference F_{01}-f_{01} and denote it by x. This variable can be considered as a sum of 264 independent, Bernoulli experiments. The binomial variable is zero if the company has not been involved in "a take over" and 1 otherwise. The probability of being involved in a "take over" or "not" is estimated from the sample by MLE from the sample.

Thus we assume that the companies in the strata, which are not in the sample, are binomially distributed. Furthermore we estimate the record level risk by the conditional expectation given f_k and F_s as follows,

$$
\begin{aligned}
&E_{p_k}\left\{ 1/_{F_k} \mid f_k, F_s \right\} \\
&= \sum_{x=0}^{F_s - f_s} 1/_{(f_k + x)} \cdot \Pr(F_k - f_k = x) \\
&= \sum_{x=0}^{F_s - f_s} 1/_{(f_k + x)} \cdot \Pr(X = x) \\
&= \sum_{x=0}^{F_s - f_s} 1/_{(f_k + x)} \cdot Bin(F_s - f_s, x, p_k).
\end{aligned}
$$

As mentioned above, the p_k is estimated by the ratio $f_k / \sum_k f_{k,}$ which is the MLE, i.e., by the relative frequency with key k in the sample selected from stratum s. Of course the variance of the estimator will be strongly influenced by the size of the sample. We have not performed any estimation of the variance in this study. The estimate will generally not be an unbiased estimate of $1/F_s$.

The estimation has been conducted using SAS, utilizing among other things the cumulative binomial formula implemented in SAS BASE. Our estimator is simple to implement, as is illustrated by the program in the appendix.

4. Estimated record level risk

Table 2 below shows the estimates for the key NACE**size,* and also for the key, which in addition includes *has been involved in a take over or not during the reference year 1999*.

Table 2. The disclosure risks for the companies in the CVTS2, using NACE and the size variable *number of employees* grouped into 10-49, 50 - 249 and 250 and more as identifying variables. Additionally, an overview of the *estimated* disclosure risks also including *has been involved in a take over or not during the reference year 1999* as an identifying variable.

Country	Identifying variables NACE, nr of employees				Identifying variables NACE, nr of employees and has been involved in a take over or not during the reference year 1999.		
	<10 %	[10, 50] %	<50,100> %	Singeltons, i.e., 100 % risk	< 10%	[10, 50] %	50 % <
Austria	2527	84	0	1	2370	222	19
Belgium	1107	21	0	1	1066	59	2
Bulgaria	2659	12	2	2	2563	101	11
Czech Republic	4284	94	0	0	4119	239	20
Denmark	1165	83	0	2	1139	99	12
Estonia	1427	65	0	2	1328	150	16
Finland	1698	10	0	0	1589	110	9
France	4548	33	0	0	4449	126	6
Germany	3177	7	0	0	3143	41	0
Greece	2372	177	0	10	2277	248	23
Hungary	2798	52	0	1	2745	98	8
Ireland	377	23	0	0	353	40	7
Italy	6720	118	0	2	6600	219	21
Latvia	3287	63	0	2	3164	183	5
Lithuania	2901	43	0	1	2757	164	24
Luxembourg	739	69	0	3	684	118	9
Netherlands	3993	66	0	3	3887	158	17
Norway	1803	19	0	1	1731	89	3
Poland	1279	37	0	1	1244	60	13
Portugal	5708	26	0	0	5638	93	3
Romania	5821	140	0	7	5584	356	28
Slovenia	1089	28	0	1	1005	107	6
Spain	9733	3	0	0	9706	28	2
Sweden	2759	12	0	0	2732	38	1
UK	941	0	0	0	941	0	0

First, we see from table 2 that many companies have a disclosure risk of more than 10% when the key consists of NACE and *number of* employees grouped into 5-49, 50 - 249 and 250 and more. These probabilities are based on known figures from the registers. Thus they do not have the uncertainty of the estimates. There are also some singletons, which of course correspond to record level risk equal to one.

Second, when we add *has been involved in a take over or not during the reference year 1999* as an identifying variable, there are also quite a few, which have a disclosure risk above 50%.

5. Some concluding remarks

Our estimates clearly show that with a standard approach such as record level risk, the data cannot be considered safe for all countries. The estimated record level risk for NACE**size*has been involved* *in a take over or not during the reference year 1999* indicate that spontaneous recognition may occur. Also when the key consists of only NACE and *size* there are already many records with high risk. As mentioned before, for this key the totals are known for the population.

It is recommended that the data could be more extensively anonymised by for example not releasing either *size* or NACE. But of course both these variables are an important basis for many analyses of *CVTS* data. Another possibility is to suppress some variable observations.

The disclosure scenario selection of key variables is of course a very important factor in the assessment of how safe the data is. These depend heavily on the availability of registers, which varies to a large extent from country to country, due in part to differences in national legislation. In Norway the NACE code (a modified version is used) is by law public and available in "The central coordinating register of legal entities" together with an identity number and the name of the company. The number of employees is also often available too, and there is a proposal from the authorities to classify this as public information. Additionally many companies have their financial report available on their Internet site so that potential investors have easy access to this data for their financial analysis.

Another discussion centres around the question "if data is older than2000/2001 is it of any interest to an intruder". For an intruder, who is interested in gaining information useful for improving his financial analysis for investing in the stock market, the data is certainly too old. On the other hand the age of the data is not so important for a journalist acting as an intruder to spread negative publicity for some NSI.

Our results actualise the discussion to what degree the data protection should rely on the anonymisation and on the legal protection of the data, respectively. Business data tend to have higher risk of disclosure than personal and household data. This may be taken, as an argument for the case that access to business data should be treated differently. Currently Regulation (EC) No 831/2002 covers access to both these types of data. One possibility is to adjust the legal framework by increasing the screening of the researches and put more weight on the legal protection and so be less strict with the anonymisation of the business data.

Appendix A. NACE codes used

NACE-categories in CVTS2 based on NACE Rev. 1

NACE 20	Section/ Sub-section	Division	Description
01	C/CA, CB	10-14	Mining and quarrying
02	D/DA	15-16	Manufacture of food products, beverages and tobacco
03	D/DB, DC	17-19	Manufacture of textiles and textile products; Manufacture of leather and leather products
04	D/DE	21-22	Manufacture of pulp, paper and paper products; Publishing, printing and reproduction of recorded media
05	D/DF to DI	23-26	Manufacture of coke, refined petroleum products and nuclear fuel; Manufacture of chemicals, chemical products and man-made fibres; Manufacture of rubber and plastic products; Manufacture of other non-metallic mineral products
06	D/DJ	27-28	Manufacture of basic metals and fabricated metal products
07	D/DK, DL	29-33	Manufacture of machinery and equipment n.e.c.; Manufacture of electrical and optical equipment
08	D/DM	34-35	Manufacture of transport equipment
09	D/DD, DN	20, 36-37	Manufacture of wood and wood products; Manufacturing n.e.c.
10	E	40-41	Electricity, gas and water supply
11	F	45	Construction
12	G	50	Sale, maintenance and repair of motor vehicles and motorcycles; retail sale of automotive fuel
13	G	51	Wholesale trade and commission trade, except of motor vehicles and motorcycles
14	G	52	Retail trade, except of motor vehicles and motorcycles; repair of personal and household goods
15	H	55	Hotels and restaurants
16	I	60-63	Land transport; transport via pipelines; Water transport; Air transport; Supporting and auxiliary transport activities; activities of travel agencies
17	I	64	Post and telecommunications
18	J	65-66	Financial intermediation, except insurance and pension funding; Insurance and pension funding, except compulsory social security
19	J	67	Activities auxiliary to financial intermediation
20	K; O	70-74; 90-93	Real estate, renting and other business activities; Other community, social, personal service activities

Appendix B. SAS program to estimate record level risk[1]

```
data checksafe;
merge strata help;
by country NACE_SP SIZE_SP ;
drop x1;
status='dontknow';
if fk gt T then status='safe';
else if fk le T then status='unsafe';

pk=fk/nsample;
N_est=pk*NSTRA_SP;
risk=1/N_est;

* Estimating the E(1/N ¦ fk,NSTRA_SP), p 277 MOS ws on conf;

data checksafe;
set checksafe;

nn=NSTRA_SP-nsample;

prob=probbnml(pk,nn,0);
sumprob=prob;
Erisk=prob/fk;

do x=1 to nn by 1;
   xmin=x-1;
   probx=probbnml(pk,nn,x);
   probxmin=probbnml(pk,nn,xmin);
   prob= probx-probxmin;
   Erisk=Erisk+prob/(fk+x);

   sumprob=sumprob+prob;
   end;

* The record level risk is the inverse of the number in the strata;
data checksafe;
set checksafe;
risk_str=1/NSTRA_SP;

proc sort;
by country;

proc print;
var NACE_SP A5C Nsample  fk NSTRA_SP risk_str Erisk risk;
by country;

proc sort data=checksafe;
by country;

proc plot;
plot risk_str*erisk;
by country;

file 'riskest.out';
Put country NACE_SP A5C Nsample  fk NSTRA_SP risk_str Erisk risk;

proc univariate data=checksafe;
var Erisk risk;
by country;
run;
```

[1] This is the version to check the NACE*SIZE*TAKEOVER

References

Benedetti, R. and Franconi, L. (1998), 'An estimation method for individual risk of disclosure based on sampling design'

Bethlehem, J. G., Keller, W. J., and Pannekoek, J. (1990), Disclosure control of microdata*, Journal of the American Statistical Association* Vol 85, 38-45.

Di Consiglio, L., Franconi, L. & Seri, G. (2003). Assessing individual risk of disclosure: an experiment, *Proceedings from the Joint ECE/Eurostat work session on Statistical Data Confidentiality (Luxembourg, 7-9 April 2003*). MOS Eurostat.

Hundepool, A., Wetering, A. van de, Franconi, L., Capobianchi, A. & Wolf, P.P. de (2002b). μ-*ARGUS, user's manual*, version 3.1.

Pannekoek, J., (1996). 'Statistical methods for some simple disclosure limitation rules', Statistica Neerlandica, Volume 53, Nr. 1 - March 1999, sider 55-67.

SAS Institute (1999): SAS Language References: Dictionary, Version 8.

Skinner, C. J. , and Elliot, M. J. (2002), *A measure of disclosure risk for microdata*, Journal of the Royal Statistical Society, Series B, Methodological, 64 (4) , 855-867

Skinner, C. J., Marsh C., Openshaw S., and Wymer, C. (1994), ``Disclosure control for Census microdata'', *Journal of Official Statistics, Vol.10* , 31-51

Stander, Julian (2003). Discussion of Topic (v) : Risk Assessment, *Proceedings from the Joint ECE/ Eurostat work session on Statistical Data Confidentiality (Luxembourg, 7-9 April 2003).* MOS Eurostat.

Willenborg, L.C.R.J. & Waal, T. de (1996), Statistical disclosure control in practice,*Lecture Notes in Statistics.* New York: Springer-Verlag.

Willenborg, L.C.R.J. & Waal, T. de (2001), Elements of S*tatistical disclosurecontrol, Lecture Notes in Statistics.* New York: Springer-Verlag.

Topic IV

Access to business microdata for analysis

A standard for the release of microdata

Rainer Lenz[*], *Daniel Vorgrimler*[**] *and Michael Scheffler*[***]

[*] **Research Data Centre, Institute for Research and Development in Federal Statistics, Statistisches Bundesamt, 65180 Wiesbaden, Germany, rainer.lenz@destatis.de**

[**] **Tax Statistics, Statistisches Bundesamt, 65180 Wiesbaden, Germany, daniel.vorgrimler@destatis.de**

[***] **IT User Service/Statistical and Geo Information Systems, Statistisches Bundesamt, 65180 Wiesbaden, Germany, michael.scheffler@destatis.de**

Abstract: Statistical Offices in Germany may provide microdata to the scientific community, if these data are sufficiently anonymized. We present a standard for evaluating the degree of protection of a confidential data file. In a first step distance based record linkage is used to re-identify statistical units of the confidential target data. Besides re-identification of the unit it is also important to look at the benefit to a potential data intruder in case he reveals information. The more the information in the data disseminated is disturbed the lower is the benefit a data intruder derives from re-identification. For this reason, in a second step the re-identified units are analyzed if they contribute benefit to potential data intruders.

The paper shows how the standard mentioned can be applied to real world examples, taking the German Turnover Tax Statistics (almost full survey, about 3 million units), the German Structure of Costs Survey (a sample containing about 18 000 units) and the German Retail Trade Statistics (a sample containing about 23500 units) as a basis. Recently, so called Scientific-Use-Files of these surveys have been made available for the scientific community.

1. Introduction

For German statistics legislation, a data set is anonymous (as far as scientific uses are concerned) if the costs of identification exceed the benefits of identification. Those data bases are called Scientific-Use-Files as such data can be provided exclusively to scientists. Costs and benefits depend on how “sure” a data intruder can be to reveal useful information. In practice, a data intruder faces several problems:

- divergence between additional knowledge and anonymized data set,
- lack of knowledge as to whether the target individual is covered by the data,
- uncertainty as to whether an assignment is correct,
- uncertainty about the quality of the data revealed.

While in the area of households and individuals the anonymization of microdata has been practised for several years, an anonymization of business microdata is notably more difficult: Business surveys are based on essentially smaller sample universes than individual-related surveys so that the cell frequencies of individual groups are often also smaller. The distributions of quantitative variables are by far more heterogeneous, and dominating cases do occur. Compared to individual-related surveys, the sampling fractions of business surveys are generally much larger while with respect to some strata, they are even equal to complete counts. Besides, the number of units differs largely between the individual business size classes. Due to the businesses’ obligation to publish data, on the one hand, and to the opportunity to retrieve information from data bases against payment, on the other, an external who intends to assign microdata to the respective carrier has at his disposal a substantially larger and much better processed additional knowledge about businesses than he has about individuals or households. And finally, the advantage gained from knowing data on enterprises and local units is rated by far more highly than that achieved from obtaining information about individual- or household-related surveys. Surveys of local units also include items which may be of interest to competing enterprises, such as information on investments. A rational data intruder will therefore accept higher expenses for deanonymization provided they are offset by the advantage gained from the information obtained.

2. Simulation of a data attack

In this chapter we discuss the concepts of additional knowledge and the most important scenarios of data attack.

2.1. Additional knowledge

In order to re-identify a statistical unit (e.g. a specific enterprise), several assumptions concerning the data intruder are necessary for successful attempts (see also Brand et al. 1999):

- Additional knowledge about the object (in our case in the form of an external database and knowledge obtained by internet research)
- Knowledge about the participation of the organization in the target survey (response knowledge)
- Key variables contained in both target and external data (making a unique assignment possible)

Moreover, the data intruder must be personally convinced of the correctness of the assignment, for which he seems to be asking the impossible in the case of simulating a database cross match described in subsection 2.2.1.

2.2. Scenarios of data attack

In Elliott and Dale (1999) several scenarios of data attack are mentioned, two of them are the so called database cross match and the match for a single individual (see also Vorgrimler and Lenz 2003).

2.2.1. Database cross match

Within a database cross match a data intruder matches an external database with the confidential target data. In order to enhance his external data, he tries to assign as many true pairs of records as possible.

In a first phase, we generate a distance measure covering all common key variables of the records in the two databases. As in a real attack scenario data intruders tend to prefer a few selected variables, supposed to include less deviations from the original data, to other, less reliable variables, it is left to the user to assign concrete weights w_i to variables *i*, although, for the sake of simplicity, standardised weight intervals of *[0, 1]* were laid down.

The objective of the second phase is to make assignments of records on the basis of the previously calculated distances. For that purpose, we minimize the sum of distances for all assignments to be made (total deviation). For the purpose of comparison, we use an algorithm firstly presented in Lenz (2003a) and developed further in Lenz (2004).

2.2.2. Match for a single individual

The intention behind a single individual match is to gain information about a specific target individual. The data intruder collects information about the individual searched for, using several sources of information. For instance, he can generate additional information by commercial databases and generally accessible information (e.g. annual reports of enterprises). The collected information is then used to re-identify the target individual in order to get further information about it.

2.2.3. Combination of scenarios

In order to adequately evaluate the protection effect of an anonymization method, both scenarios of data attack have to be taken into account. Let $R_{SIM}(u)$ denote the estimated re-identification risk associated with a single individual match applied to some unit *u* and $R_{DCM}(u)$ denote the corresponding estimator for the re-identification risk associated with a database cross match. Then, the re-identification risk *R(u)* can be estimated by the maximum of both estimators, $R(u) := max\{R_{SIM}(u) , R_{DCM}(u)\}$.

The re-identification risk for some unit strongly depends on the data blocks to which it belongs. For instance, if an enterprise is assigned to a small branch of economic activity and/or to an upper employee size class, re-identification appears much easier than in the general case. Here, the re-identification risk $R_{SIM}(u)$ associated with a single individual match is expected to be higher than the corresponding one $R_{DCM}(u)$ associated with a database cross match. On the other hand, the database cross match stands above the single individual match in areas of data with high density, since in general there are many units with similar parameter values.

If by a data attack a set of additional knowledge was successfully assigned to an anonymized data set, all target variables which are contained in this data set were revealed. The benefit of a successful assignment hence arises from the „useful“ information which a data intruder can reveal by a successful identification. An information revealed is only useful if the values revealed correspond to the “true values” or at least if the values revealed are similar to the true values to a certain extent. Some anonymization methods modify the values of the data so that the values of the data disseminated differ from the corresponding original („true“) values. Above a certain deviation (between the value revealed and the “true“ value) a data intruder will not obtain a benefit from the information revealed. In our case, deviation is defined as the relative difference between the disseminated value and the “true“ value of a variable.

This means that individual data will fulfill the criterion of being “anonymous“ if the correctly assigned data set provides mainly useless information (the value revealed is outside a “deviation threshold“ of the „true value”). It is a task of the statistical office to specify this deviation threshold. In the following examples, the deviation threshold has been set to 0.1 (that is, a value is considered to contribute useful information to a data intruder if its relative difference from the true value is less than 10 percent) and the risk of revealing useful information is called *disclosure risk.*

3. Application to real world examples

In this chapter we describe how the above-described concepts can be applied to the German Turnover Tax Statistics 2000 (TTS), the German Structure of Costs Survey 1999 (SCS) and the German Retail Trade Statistics 1999 (RTS).

3.1. German Turnover Tax Statistics

Turnover tax statistics are based on an evaluation of monthly and quarterly advance turnover tax returns to be provided by entrepreneurs whose turnover exceeds in the year 2000 € 16,617 and whose tax amounts to over € 511 per annum. Also excluded are enterprises with activities which are generally non-taxable or where no tax burden accrues (e.g. established medical doctors and dentists without laboratory, public authorities). Nearly all economic branches are presented in the survey. The evaluation of the year 2000 contains almost 3 million records. The survey has been conducted annually since 1996 (until then, every two years). The Federal Statistical Office of Germany published the following selected survey characteristics in tables:

- Deliveries and other performances (= taxable and non-taxable turnover)
- Branch of economic activity
- Legal form
- Bases of turnover tax (deliveries and other performances, intra-community acquisitions, input tax by tax rates, etc.)

In this section, we consider four ways to anonymize the TTS. General descriptions of these and other anonymization methods - independent from some specific survey - can be found in Höhne (2003).

1. The first constitutes the weakest possible form of anonymization, formal anonymization, consisting in the deletion of the direct identifiers like name, address and so on. (FORMAL)
2. The second is the use of traditional methods (like truncation and coarsening) of anonymization. Since the German turnover tax statistics determine a rather large data set, an application of traditional methods could produce reasonable results concerning confidentiality. (Traditionally anonymized)
3. The third is the weakest variant of the so-called micro aggregation, where each numerical variable defines its proper group. (MA 21G)
4. The fourth is the strongest variant of multidimensional micro aggregation, where all numerical variables are grouped together. (MA 1G)

3.1.1. Database cross match

For our purposes, the most important variables are:

- Branch of economic activity (NACE)
- Total turnover
- Legal status
- Regional key

The above variables are the key variables of the TTS and external data (additional knowledge). The external data contains nearly 9300 enterprises with 20 or more employees, classified within NACE codes 10 - 37 (manufacturing industry). The corresponding subset of the target data contains nearly 37000 enterprises. We carried out database cross matches with different anonymizations of the categorical variables. In the original data, the NACE code has four digits. Through truncation the NACE code is reduced to zero (in this case the data intruder possesses no information on the branch of economic activity), one, two and three digits, so that we obtain four non-trivial forms of the code. Furthermore, the legal status is re-coded. In the original data, the legal status has a range of eight values, after re-coding it is coarsed to four values.

The following table contains the results obtained by blocking data using the four levels of the NACE code, with the 0-digit cases indicating that the variable was left out of consideration. That is, the data intruder does not have additional knowledge of the branch of economic activity.

Table 1. Matching TTS: re-identification risk (disclosure risk) distributed to employee size classes

TTS	NACE	Total	Employee size class*					
			1	2	3	4	5	6
FORMAL	4 digits	40.1 (40.1)	35.3 (35.3)	35.7 (35.7)	45.7 (45.7)	54.9 (54.9)	57.7 (57.7)	70.0 (70.0)
	3 digits	40.1 (40.1)	35.5 (35.5)	36.1 (36.1)	45.1 (45.1)	52.6 (52.6)	65.4 (65.4)	60.0 (60.0)
	2 digits	35.4 (35.4)	31.6 (31.6)	31.5 (31.5)	39.5 (39.5)	54.2 (54.2)	61.5 (61.5)	80.0 (80.0)
	1 digit	21.0 (21.0)	17.9 (17.9)	18.5 (18.5)	23.1 (23.1)	42.5 (42.5)	34.6 (34.6)	40.0 (40.0)
	0 digits	13.6 (13.6)	11.5 (11.5)	11.9 (11.9)	14.7 (14.7)	28.9 (28.9)	42.3 (42.3)	40.0 (40.0)
MA 21G	4 digits	40.1 (39.6)	35.3 (34.9)	35.7 (35.3)	45.6 (45.1)	55.2 (53.0)	57.7 (39.9)	70.0 (56.1)
	3 digits	39.9 (39.5)	36.1 (35.8)	36.1 (35.7)	44.6 (44.2)	53.3 (52.8)	57.7 (39.7)	80.0 (65.6)
	2 digits	35.4 (35.0)	31.6 (31.3)	31.5 (31.2)	39.3 (39.1)	55.2 (51.2)	61.5 (42.4)	80.0 (65.2)
	1 digit	20.8 (20.6)	17.7 (17.5)	18.3 (18.1)	22.8 (22.6)	42.2 (40.5)	34.6 (23.7)	50.0 (41.6)
	0 digits	13.7 (13.6)	11.3 (11.2)	12.0 (11.9)	14.5 (14.3)	32.5 (31.2)	30.8 (21.3)	40.0 (33.7)
MA 1G	4 digits	27.9 (5.6)	21.4 (3.6)	21.8 (4.8)	35.0 (7.4)	53.9 (8.1)	65.4 (7.3)	60.0 (5.1)
	3 digits	23.4 (4.5)	15.8 (2.6)	17.3 (3.5)	30.1 (6.5)	52.6 (7.6)	73.1 (9.0)	80.0 (7.2)
	2 digits	14.4 (2.9)	6.9 (1.2)	9.5 (1.8)	18.9 (4.2)	46.8 (7.2)	69.2 (7.6)	60.0 (4.7)
	1 digit	5.4 (1.0)	2.1 (0.4)	3.3 (0.7)	7.0 (1.6)	21.2 (3.1)	34.6 (4.5)	30.0 (2.8)
	0 digits	2.6 (0.5)	0.9 (0.2)	1.5 (0.3)	3.4 (0.7)	11.7 (1.9)	26.9 (2.9)	30.0 (2.5)
Traditionally anonymized		30.0 (29.9)	26.7 (26.7)	27.0 (27.0)	34.0 (34.0)	41.2 (39.7)	19.2 (11.2)	20.0 (8.1)

* 1 = less than 25; 2 = 25-100 ; 3 = 100-1 000; 4 = 1 000-5 000 ; 5 = 5 000-15 000 ;
6 = more than 15 000.

Obviously, the weakest variant MA 21G provides lesser protection than the other variants of anonymization. The great deviations between the two data sources are more decisive for this phenomenon than the slight (almost negligible) modifications to the TTS. While only about 1% of the enterprises have been classified differently with regard to the regional information, nearly 25% of the enterprises covered by the German turnover tax statistics have been assigned to another branch of economic activity than their respective records of the external data. With regard to the variable *Number of employees* there also are significant differences in both surveys. *Total turnover* figures match relatively well. Only some 18.8% of the enterprises show deviations of more than 10% between both data sources. As had to be expected in the authors' opinion, the variant MA 1G produces safe microdata. On the other hand, this variant is connected with an unbearable abatement of statistical properties. The matching results obtained by coarsening the NACE code to 3 or 4 digits are comparable. In the case of NACE 4 the increase in the number of enterprises protected due to deviations in both sources is compensated by the decrease in the re-identification risk in the case of NACE 3 due to larger blocks.

An improved effect of protection is achieved by reducing the NACE code to 2 digits. Regarding the traditional method, it is observed in contrast to the other methods that this method – roughly spoken - protects the larger insecure enterprises much better. All in all, the disclosure risks (obtained by involving the concept of useful information) are slowed down in line with an increasing growth of enterprises.

3.1.1. Match for a single individual

We repeated the single individual match for 15 enterprises with the target data set being only formally anonymized. The key variables were the regional key, the business classification, the legal status and the turnovers of the years 1999 and 2000 (note that the key variables were not available for the observations as a whole). Using these key variables, only 6 out of 15 enterprises could be re-identified.

Hence, the results are in accordance with the database cross match, where the influence of deviations in both surveys (irrespective of the method of anonymization decided for) were the main reason for unsuccessful attempts. But we can also observe that in contrast to other statistics (like the German structure of costs survey SCS) the structure of the German turnover tax statistics does not offer a data intruder more key variables within a single match scenario than in the scenario of a database cross match. Therefore, the risk of re-identification of a specific enterprise with respect to a single match scenario is not higher than the risk regarding a database cross match.

3.2. German Structure of Costs Survey

The German structure of costs survey of the year 1999, limited to the manufacturing industry, is a projectable sample and includes a maximum of 18000 enterprises with 20 or more employees. All enterprises with 500 or more employees or those in economic sectors with a low frequency are included. That is, a potential data intruder has knowledge about the participation of large enterprises in the survey. We consider the survey of the year 1999, covering 33 numerical variables (among which are *Total turnover, Research and Development* and the *Number of employees*) and two categorical variables, namely the *Branch of economic activity* (abbreviated: NACE), broken down to the 2-digit level, and the *Type of administrative district* (abbreviated: BBR9), which has 9 values depending on the degree of urbanisaton of the region considered.

In this section, we consider five ways to anonymize the SCS. A detailed description of these methods can be found in Lenz (2003b).

1. The first constitutes the weakest possible form of anonymization, formal anonymization, consisting in the deletion of the direct identifiers like name, address and so on. (FORMAL)
2. The second is the weakest variant of the so-called micro aggregation, where each numerical variable defines its proper group. (MA 30G)
3. In the third variant, the set of variables is textually divided into three-element groups. (MA 10G)
4. Grouping highly correlated variables leads to groups of size between two (smallest group) and twelve elements (largest group). (MA 8G)
5. The fifth is the strongest variant of multidimensional micro aggregation, where all numerical variables are grouped together. (MA 1G)

3.2.1. Database cross match

For our purposes, the most important variables are:

- Branch of economic activity (NACE 2), reduced to two digits
- Type of administrative district (BBR9), containing 9 categories
- Total turnover
- Number of employees

The above variables are the key variables of the SCS and external data (additional knowledge). The external data contains nearly 9400 enterprises with 20 or more employees, classified within NACE codes 10 - 37 (manufacturing industry).

We carried out database cross matches with five different degrees of perturbation of the categorical variables *Total turnover* and *Number of employees*. The results obtained are shown in table 2.

Table 2. Matching SCS: re-identification risk (disclosure risk) distributed to employee size classes

SCS	Total	Employee size class*					
		1	2	3	4	5	6
FORMAL	24.4 (24.4)	15.6 (15.6)	19.0 (19.0)	26.5 (26.5)	36.1 (36.1)	41.8 (41.8)	44.9 (44.9)
MA 30G	24.4 (24.2)	15.6 (15.5)	19.0 (18.9)	26.5 (26.4)	35.9 (35.8)	41.6 (41.4)	44.7 (43.8)
MA 10G	24.2 (19.8)	15.6 (12.8)	19.2 (16.9)	26.5 (21.5)	34.4 (26.1)	41.3 (29.7)	44.0 (24.1)
MA 8G	19.6 (10.8)	12.7 (7.7)	14.9 (8.9)	21.9 (12.5)	27.0 (14.6)	35.5 (18.4)	36.2 (16.3)
MA 1G	3.8 (1.1)	2.2 (0.7)	1.5 (0.4)	3.1 (0.8)	5.8 (1.5)	9.0 (2.1)	16.7 (2.5)

* 1 = 20-49 ; 2 = 50-99 ; 3 = 100-249; 4 = 250-499 ; 5 = 500-999 ; 6 = more than 999

As to be expected, the frequency of correct assignments grows with the number of employees. Although it is normal that for larger enterprises the micro aggregation procedures cause more pronounced changes in the variables, the column on the right of table 2 shows a notably high risk of re-identification and disclosure for enterprises with at least 1000 employees.

While the deviation amounting to about 24% for all enterprises in the *Branch of economic activity* is in line with the preceding section as are the slight deviations in the regional data of less than 2%, there are much more marked differences regarding *Total turnover*. About 50% of the enterprises deviate from each other by more than 10% in the two data sources.

3.2.2. Match for a single individual

We repeated the single individual match for 41 enterprises, without consideration of commercial databases. In general, the key variables were the same as in the previous subsection. In some instances, the variables *Total revenue*, *Research and development investments* (yes or no), *trade activity* (yes or no) appeared as further key variables. With these keys, 19 of the 41 enterprises searched for could be re-identified. Only one enterprise could be re-identified among the 15 enterprises with less than 250 employees. On the other hand, among the larger enterprises a total of 18 out of 26 could be re-identified.

3.3. German Retail Trade Statistics

The German Retail Trade Statistics of the year 1999 is a projectable sample containing about 23500 enterprises. In each branch of economic activity, the dominant enterprises have been included into the survey. The RTS consists of 33 numerical and 3 categorical variables. The results of this annual survey yield important information to economic-political problems concerning the structure, profitability and productivity of enterprises of this sector. In this section, we consider four ways to anonymize the RTS. A detailed description of these methods can be found in Scheffler (2005).

1. The first constitutes the weakest possible form of anonymization, formal anonymization, consisting in the deletion of the direct identifiers like name, address and so on. (FORMAL)
2. The second is the weakest variant of the so-called micro aggregation, where each numerical variable defines its proper group. (MA 31G)
3. The third was obtained by groupwise application of micro aggregation to 9 groups of numerical variables. (MA 9G)
4. The fourth is the strongest variant of multidimensional micro aggregation, where all numerical variables are grouped together. (MA 1G)

3.3.1. Database cross match

In order to simulate database cross matches with the RTS, we generated additional knowledge containing about 12100 enterprises, classified within NACE codes 521 - 527 (retail trade) on a three-digit level. Hence, the key variables are

- Branch of economic activity (NACE 2), reduced to three digits
- Type of administrative district (BBR9), containing 9 categories
- Total turnover

Table 3 below contains the re-identification and disclosure risks associated with the four variants of anonymization distributed to employee size classes.

Table 3. Matching RTS: re-identification risk (disclosure risk) distributed to employee size classes

SCS	Total	Employee size class*						
		1	2	3	4	5	6	7
FORMAL	22.2 (22)	20.9 (20.9)	23.7 (23.7)	29.3 (29.3)	27.4 (27.4)	39.6 (39.6)	25.0 (25.0)	48.1 (48.1)
MA 31G	22.0 (21.8)	20.9 (20.8)	23.6 (23.6)	28.1 (28.1)	29.3 (29.3)	38.4 (38.4)	30.0 (29.9)	45.1 (42.8)
MA 9G	3.1 (2.4)	3.2 (2.7)	3.1 (2.6)	6.0 (4.9)	10.4 (8.0)	17.2 (12.9)	13.1 (9.9)	30.2 (19.9)
MA 1G	2.1 (1.3)	2.0 (1.6)	2.2 (1.4)	3.5 (2.1)	4.8 (2.1)	7.6 (4.2)	9.1 (5.2)	24.8 (11.4)

* 1 = 1-19 ; 2 = 20-49 ; 3 = 50-99; 4 = 100-249 ; 5 = 250-499 ; 6 = 500-999 , 7 = more than 999.

As was to be expected, the protection effect of the weakest variant of micro aggregation, MA 31G, is similar to the effect of formally anonymized data. For enterprises with 500-999 employees, this method even has a disclosive impact. In accordance with the previous sections, the relative frequencies of correct assignments grow with the number of employees.

3.3.2. Match for a single individual

We repeated single individual matches for a sample of 20 enterprises drawn by the size class of enterprises with more than 999 employees. In several passes, the variable *Number of branch offices* turned out to be a key variable between additional knowledge (mainly generated by internet research) and the target enterprise.

At first, the matches were carried out using only the internet as additional knowledge. In doing so, 8 of the 20 enterprises searched for could be uniquely and correctly assigned to their corresponding target individuals (re-identified). In a second step, the matches were carried out using only the external database described in 3.3.1. Here, 11 of the 20 enterprises participated in the external survey, where 6 of them could be re-identified using the external data and 4 of them using the Internet.

Finally, the matches were carried out using both, internet and external database, as additional knowledge. In this simulation, 8 of the 11 enterprises searched for could be re-identified. This means an increase from 4 (Internet) over 6 (external database) to 8 re-identifications.

3.4. Scientific-Use-Files

For each of the above-described surveys a so called Scientific-Use-File has been generated, i.e. data available for scientific purposes. Since the TTS consists of many records (about 2.9 million) and less numerical variables (most of them strongly correlated with *Total turnover*, a strong emphasis was put on anonymization of categorical variables (essentially information reducing methods). Anonymizing the SCS, consisting of less records (about 18.000) and about 30 numerical variables, a stronger emphasis was put on numerical variables (data perturbing methods) as well as in the case of the RTS. Detailed descriptions of the Scientific-Use-Files can be found in Lenz et al. (2005), Vorgrimler et al. (2005) and Scheffler (2005).

4. Conclusion

In this paper we examined the risk a data intruder must take into account when he conducts an identification attempt. Economic rationale suggests that if the risk to fail is too high, an intruder will refrain from an identification attempt and the data sets can be regarded as protected. The concepts have been applied to three different business surveys of German official statistics.

Currently, similar approaches are made in order to anonymize further business statistics like the Continuing Vocational Training Survey 1999 and the Structure of Earnings Survey 2001.

This work was partially supported by the EU project IST-2000-25069, Computational Aspects of Statistical Confidentiality, and by the German national project De Facto Anonymization of Business Microdata.

References

Brand, R., Bender, S., Kohaut, S. (1999). *Possibilities for the creation of a scientific-use-file for the IAB-Establishment-Panel*. Proceedings of the Joint Eurostat/UN-ECE Work Session on Statistical Data Protection, Thessaloniki, 57-74.

Domingo-Ferrer, J., Torra, V. (2003): *Record linkage methods for multidatabase data mining. Information Fusion in Data Mining.* Springer-Verlag, Berlin, 99-130.

Elliot, M., Dale, A. (1999). *Scenarios of attack: the data intruder's perspective on statistical disclosure risk.* Netherlands Official Statistics, 6-10.

Höhne, J. (2003): *Methoden zur Anonymisierung wirtschaftsstatistischer Einzeldaten (German).* Forum der Bundesstatistik, 42, Wiesbaden, 69-94.

Lenz, R. (2003a). *A graph theoretical approach to record linkage.* Monographs of Official Statistics – Research in Official Statistics, Luxembourg, 324-334.

Lenz, R. (2003b). *Disclosure of confidential information by means of multi-objective optimisation.* Proceedings of the Comparative Analysis of Enterprise Micro Data Conference (CAED), London, 2003.

Lenz, R. (2004). *Measuring the disclosure protection of micro aggregated business microdata – An analysis taking the example of German Structure of Costs Survey*. Appears in: Journal of Official Statistics, 2004.

Lenz, R., Vorgrimler, D., Rosemann, M. (2005). *Ein Scientific-Use-File der Kostenstrukturerhebung im Verarbeitenden Gewerbe (German).* Wirtschaft und Statistik 2, 91-96.

Rosemann, M., Vorgrimler, D., Lenz, R. (2004). *Erste Ergebnisse faktischer Anonymisierung wirtschaftsstatistischer Einzeldaten (German).* Journal of the German Statistical Society, Vol. 88, 73-99.

Scheffler, M. (2005): *Ein Scientific-Use-File der Einzelhandelsstatistik 1999 (German).* Wirtschaft und Statistik 3, 197-200.

Vorgrimler, D. (2003): *Re-Identifikationsmöglichkeiten am Beispiel eines konkreten Angriffsszenarios (German).* Forum der Bundesstatistik, 42, Wiesbaden, 40-58.

Vorgrimler, D., Dittrich, S., Lenz, R., Rosemann, M. (2005): *Ein Scientific-Use-File der Umsatzsteuerstatistik 2000 (German).* Wirtschaft und Statistik 3, 197-200.

Vorgrimler, D., Lenz, R. (2003): *Disclosure risk of anonymized business microdata files – Illustrated with empirical key variables.* Bulletin of the 54th International Statistical Institute (ISI), book 2, Berlin, 594-595.

Estimation of the Probit Model From Anonymized Micro Data

Gerd Ronning, Martin Rosemann***

*** Department of Economics, University of Tuebingen, Mohlstrasse 36, D-72074 Tuebingen. (gerd.ronning@uni-tuebingen.de)**
**** Institute for Applied Economic Research (IAW), Ob dem Himmelreich 1, D-72074 Tuebingen, (martin.rosemann@iaw.edu)**

Abstract. The demand of scientists for confidential micro data from official sources has created discussion of how to anonymize these data in such a way that they can be given to the scientific community. We report results from a German project which exploits various options of anonymization for producing such ”scientific-use” files. The main concern in the project however is whether estimation of stochastic models from these perturbed data is possible and - more importantly - leads to reliable results. In this paper we concentrate on estimation of the probit model under the assumption that only anonymized data are available. In particular we assume that the binary dependent variable has undergone post-randomization (PRAM) and that the set of explanatory variables has been perturbed by addition of noise. We employ a maximum likelihood estimator which is consistent if only the dependent variable has been anonymized by PRAM. The errors-in-variables structure of the regressors then is handled by the simulation extrapolation (SIMEX) estimation procedure. Alternatively, we consider estimation of our model starting from the generalized linear model (GLM).

1. Introduction

Empirical research in economics has for a long time suffered from the unavailability of individual ”micro” data and has forced econometricians to use (aggregate) time series data in order to estimate, for example, a consumption function. On the contrary other disciplines like psychology, sociology and, last not least, biometry have analyzed micro data already for decades. The software for micro-econometric models has created growing demand for micro data in economic research, in particular data describing firm behaviour. However, such data are not easily available when collected by the Statistical Office because of confidentiality. On the other hand these data would be very useful for testing microeconomic models. This has been pointed out recently by KVI commission.[1] Therefore, the German Statistical Office initiated research on the question whether it is possible to produce scientific use files from these data which have to be anonymized in a way that re-identification is almost impossible and, at the same time, distributional properties of the data do not change too much. Results from this project have been published quite recently. See Ronning et al. (2005) where most known anonymization procedures have been rated both with regard to data protection and to informational content left after perturbation. In particular we found (rank) swapping procedures not acceptable from user’s point of view.

Published work on anonymization of micro data and its effects on the estimation of microeconometric models has concentrated on *continuous* variables where a variety of procedures is available. See, for example, Ronning and Gnoss (2003) for such procedures and the contribution by Lechner and Pohlmeier (2003) also for the effects on estimation when anonymizing data either by microaggregation or addition of noise. Discrete variables, however, mostly have been left aside in this discussion. The only stochastic-based procedure to anonymize discrete variables is post-randomization (PRAM) which switches categories with prescribed probability.

In this paper we concentrate on estimation of the probit model for which only anonymized data are available. In particular we assume that the binary dependent variable has undergone post-randomization (PRAM) and that the set of explanatory variables has been perturbed by addition of noise. We employ a maximum likelihood estimator which is consistent if only the dependent variable has been anonymized by PRAM. The errors-in-variables structure of the regressors then is handled by the simulation extrapolation (SIMEX) estimation procedure.

In Section 2 we consider the probit model. We assume that the binary dependent variable has been anonymized by PRAM whereas right-hand regressor variables have been left in original form. Consistent estimates are available from an adapted estimation procedure. We then turn to the situation

[1] See KVI (2001).

that the continuous regressors have been anonymized by noise addition (section 3). An attractive procedure for handling such situations is the simulation extrapolation (SIMEX) estimator which will be briefly described. Section 4 then presents some estimation results for the probit model when both the dependent and the independent variables have been anonymized. We present results from a simulation study where the PRAM adapted probit estimator is combined with the SIMEX approach.

2. The probit model under post randomization

2.1. The probit model

Consider the following linear model:[2]

$$Y^* = \alpha + \beta x + \varepsilon \qquad (1)$$

with $E[\varepsilon] = 0$ and $V[\varepsilon] = \sigma_\varepsilon^2$. Here the $*$ indicates that the continuous variable Y is latent or unobservable. This model asserts that the conditional expectation of Y^* but not the corresponding conditional variance depends on x. However we observe only a binary variable Y which is related to the latent variable by the ”threshold model”:

$$Y = \begin{cases} 0 & \text{if } Y^* \leq \tau \\ 1 & \text{else} \end{cases} . \qquad (2)$$

It can be shown that two of the four parameters α, β σ_ε^2 and τ have to be fixed in order to attain identification of the two remaining ones. Usually we set $\tau = 0$ and $\sigma_\varepsilon^2 = 1$ assuming additionally that the error term ε is normally distributed. This is the famous probit model. Note that only the probability of observing $Y = 1$ for a given x can be determined. If we alternatively assume hat the error term follows a logistic distribution, we obtain the closely related binary logit model.

2.2. Randomized response and post randomization

Randomization of the binary variable Y can be described as follows: Let Y^m denote the ’masked’ variable obtained from post randomization. Then the transition probabilities can be defined by $p_{jk} \equiv P(Y^m = j | Y = k)$ with $j, k \,\varepsilon\, \{0,1\}$ and $p_{j0} + p_{j1} = 1$ for $j = 0,1$. If we define the two probabilities of no change by $p_{00} \equiv \pi_0$ and $p_{11} \equiv \pi_1$, respectively, the probability matrix can be written as follows:

$$\mathbf{P}_y = \begin{pmatrix} \pi_0 & 1-\pi_0 \\ & \\ 1-\pi_1 & \pi_1 \end{pmatrix}$$

Since the two probabilities of the post randomization procedure usually are known and there is no argument not to treat the two states symmetrically, in the following we will consider the special case

$$\pi_0 = \pi_1 \qquad (3)$$

When the variable Y has undergone randomization, we will have a sample with n observations y_i^m where y_i^m is the dichotomous variable obtained from y_i by the randomization procedure.

In the handbook on anonymization (Ronning et al 2005) we also discuss the extension of PRAM to more than two categories. If the categories are ordered as, for example, in the case of ordinal variables or count data, switching probabilities for adjoining categories should be higher since otherwise the ordering would be totally destroyed. Of course, PRAM could also be extended to joint anonymization of two or more discrete variables.

[2] See, for example, Ronning (1991).

2.3. Estimation of the model under PRAM

Under randomization of the dependent observed variable we have the following data generating process:

$$Y_i^m = \begin{cases} 1 & \text{with probability} \quad \Phi_i \pi + (1-\Phi_i)(1-\pi) \\ 0 & \text{with probability} \quad \Phi_i (1-\pi) + (1-\Phi_i)\pi \end{cases} \tag{4}$$

Here Φ_i denotes the conditional probability under the normal distribution that the unmasked dependent variable Y_i takes on the value 1 for given x_i, i.e. $\Phi_i \equiv \Phi(\alpha + \beta x_i) = P(Y_i^* > 0 \mid x_i)$.
From (4) we obtain the following likelihood function:

$$L(\alpha, \beta \mid (y_i^m, x_i), i = 1, \dots, n)$$

$$= \prod_{i=1}^{n} \left(\Phi_i \pi + (1-\Phi_i)(1-\pi)\right)^{y_i^m} \left(\Phi_i (1-\pi) + ((1-\Phi_i)\pi\right)^{(1-y_i^m)} \quad . \tag{5}$$

Global concavity of this function with respect to α and β may be checked by deriving first and second (partial) derivatives of the log-likelihood function. Ronning (2005) derives the Hessian matrix of partial derivatives. A simple formula for the information matrix can be derived from which it is immediately apparent that maximum likelihood estimation under randomization is consistent but implies an efficiency loss which is greatest for values of π near 0.5. See Ronning (2005) for detailed results.

3. Addition of noise and the simulation extrapolation approach

3.1. Data protection by addition of noise

Consider the linear model which we write in usual way as follows: $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$. Let $\mathbf{e}_y$ be a vector of errors with expectation zero and positive variance corresponding to $\mathbf{y}$ and let $\mathbf{E}_X$ be a matrix of errors corresponding to $\mathbf{X}$. Addition of noise means that we have to estimate the unknown parameter vector from the model

$$\mathbf{y} + \mathbf{e}_y = (\mathbf{X} + \mathbf{E}_X)\beta + \mathbf{u} \, . \tag{6}$$

This is the well-known errors-in-variables model for which anonymization of right-hand variables creates estimation problems whereas anonymization of the dependent variable only increases the error variance[3] which should be compared with the case of microaggregation where (separate) anonymization of the dependent variable creates problems. Lechner and Pohlmeier (2005) consider nonparametric regression models where the regressors are anonymized by addition of noise. They show that from the simulation-extrapolation method (SIMEX) reliable estimates can be obtained. However for microeconometric models such as logit and probit models general results regarding the effect of noise addition and the suitability of the SIMEX method are not yet available.

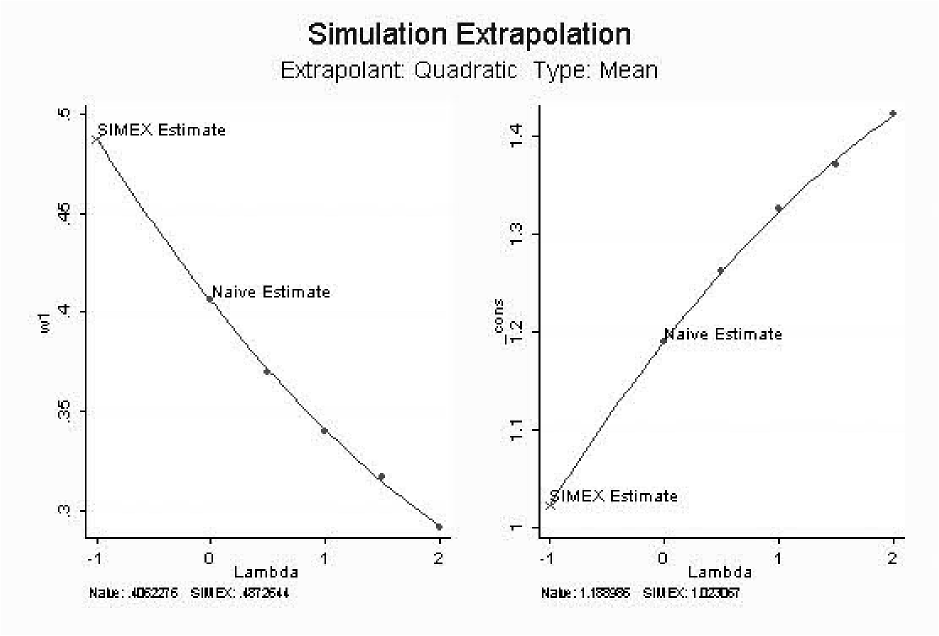
Additive errors have the disadvantage that greater values of a variable are less protected. Take as an example sales of firms. If one firm has sales of 1 million and another sales of 100 million then addition of an error of 1 doubles sales of the first but leaves nearly unchanged sales of the second firm. Therefore research has been done also for the case of multiplicative errors which in this case should have expectation one. Formally this leads to

$$\mathbf{y} \odot \mathbf{e}_y = (\mathbf{X} \odot \mathbf{E}_X)\beta + \mathbf{u}$$

where $\odot$ denotes element-wise multiplication (Hadamard product). For results regarding estimation of this linear model see Ronning et al (2005). In the following we consider only the additive case.

[3] See Lechner and Pohlmeier (2003) for details.

Figure 1. SIMEX estimator – quadratic extrapolation function



3.2. The SIMEX approach

We will only sketch the idea of this approach[4] for the simple linear regression model which is a special case of the linear model considered above with only one regressor and a constant term. It is well known from econometrics that estimation of the regression coefficient β by least squares leads to

$$plim\,\hat{\beta} = \beta \frac{\sigma_x^2}{\sigma_x^2 + \sigma_e^2}. \tag{7}$$

if the regressor variable x can only be observed with error e_x where σ_x^2 is the variance of x and σ_e^2 is the variance of this error. This corresponds to equation (6) with $P\left[\mathbf{e}_y = 0\right] = 1$. Now assume that this variance is known and that another error λe_x with $\lambda > 0$ is added to the error affected regressor variable by purpose. Then we obtain

$$plim\,\hat{\beta}(\lambda) = \beta \frac{\sigma_x^2}{\sigma_x^2 + (1+\lambda)\sigma_e^2} \tag{8}$$

so that a consistent estimator would be obtained for $\lambda = -1$. Of course $\hat{\beta}(\lambda)$ can be evaluated for any positive λ using simulation whereas results for $\lambda < 0$ have to be guessed. Usually M simulation runs are averaged for each λ so that

$$\overline{\hat{\beta}(\lambda)} = \frac{1}{M} \sum_{j=1}^{M} \hat{\beta}_j(\lambda)$$

is the estimate actually used. Cook and Stefanski (1994) suggested an extrapolation procedure which fits a curve to the various points and extrapolates it for $\lambda = -1$. This is illustrated in figure 1 for the case of a quadratic extrapolation function showing results for both coefficients from the simple regression model. Moreover, it can also be shown that for *nonlinear* models this extrapolation approach is appropriate at least approximately!

[4] For details see, for example, Carroll et al (1995).

4. Simulation results

In this subsection we will estimate the two parameters α and β of the probit model defined in (1) and (2) assuming that the dependent variable y has been anonymized by PRAM and that the regressor variable x has been protected by addition of noise. We also assume that the PRAM parameter π and the error variance σ_x^2 are known.[5] Simulated data will be used for estimation.[6] The two unknown parameters are given by $\alpha = -2.5$ and $\beta = 0.6$. The regressor variable is generated from a normal distribution $N(4.35;1.75^2)$ and the error variable satisfies $\varepsilon \sim N(0;1)$ the latter recognizing the identification constraint of the probit model.

4.1. Simex approach combined with PRAM-corrected ML

We first pursue the idea that the PRAM-corrected maximum likelihood (ML) estimator should also work if it is applied to the SIMEX approach, that is $\hat{\beta}(\lambda)$ as defined in subsection 2 now is the ML estimator as described in subsection 3. We assume that $n = 500$ observations are available. (In some cases we also use $n = 3,000$.) The maximum likelihood (ML) estimator of the probit model based on the likelihood function (4) is evaluated by a GAUSS programme written by the first author using the subroutine MAXLIK from the GAUSS library.[7]

We use $R = 50$ iterations in this simulation study which may be considered as too small but was chosen to keep computing time within acceptable limits. In each iteration the ML estimator of the probit model is employed in the SIMEX procedure: First for each $\lambda \varepsilon \{0, 0.5, 1.0, 1.5, 2.0\}$ we computed $M = 250$ values of this estimator from which $\hat{\beta}(\lambda)$ was determined. Using the five different estimates we then fitted a quadratic function to these five points and obtained the final estimate of both α and β from evaluating this function at $\lambda = -1$. From the $M = 50$ estimates we computed mean, standard deviation, median and both the minimal and the maximal value which are presented in the following table.

Since we know from earlier simulation experiments that values of the PRAM parameter π create computational problems if π is far away from 1.0 we confined simulation to the interval $\pi \varepsilon [0.8;1.0]$. Noise addition is done by a normally distributed variable with $\sigma_e^2 = 0.01$. Additionally we considered noise variances of $\sigma_e^2 = 0.04$ and $\sigma_e^2 = 0.09$ but results are not shown here since they alter the - unsatisfactory - results only marginally.

The simulation results are given in table 1. First thing to note is that for $\pi = 1.00$ (no post randomization) both parameters show a remarkable bias ”away from zero” which becomes smaller and switches its sign for decreasing values of π. In particular for $\pi = 0.90$ we get estimates which are almost perfect.

Since there is a monotonicity of the bias with respect to π it might be possible to correct for bias using this relation. We analyzed the scatter plots of the type as given in figure 1 from these estimation results and found an almost linear behaviour so that the quadratic approaximation should work well. However, evaluation of the extrapolating function at $\lambda = -1$ leads to an bias which can be both positive and negative depending on π. See table 1.

[5] It is possible to extend the estimation procedure to the case that π is unknown. See Hausman et al (1998) and Ronning (2005).

[6] The same design has been used in Ronning et al (2005) where only the dependent variable was anonymized.

[7] Many thanks to Sandra Lechner for providing us with a SIMEX routine!

Table 1. Probit model - PRAM adapted ML and SIMEX procedure $\sigma_e^2 = 0,01$

π		estimate	stand.dev.	variance	minimum	median	maximum
1,000	α	-3.581985	0.379676	0.144154	-4.417265	-3.555691	-2.880216
	β	0.860513	0.090857	0.008255	0.720343	0.846306	1.061476
0.975	α	-3.323051	0.368318	0.135658	-4.386024	-3.340966	-2.524121
	β	0.799967	0.088630	0.007855	0.641313	0.786402	1.046858
0,950	α	-3.005847	0.342184	0.117090	-3.870007	-2.980716	-2.274590
	β	0.726057	0.076550	0.005860	0.560696	0.728072	0.889345
0.925	α	-2.750990	0.309059	0.095518	-3.794785	-2.772832	-2.188305
	β	0.660091	0.070754	0.005006	0.532053	0.660100	0.919395
0,900	α	-2.498118	0.257171	0.066137	-3.073287	-2.505895	-2.013187
	β	0.597422	0.051819	0.002685	0.485514	0.600677	0.704354
0.875	α	-2.304071	0.200737	0.040295	-2.994306	-2.280229	-1.982457
	β	0.553141	0.049268	0.002427	0.473322	0.545844	0.700764
0,850	α	-2.051473	0.207291	0.042970	-2.791021	-2.043950	-1.729020
	β	0.488270	0.041388	0.001713	0.422237	0.490443	0.632154
0.825	α	-1.789513	0.190780	0.036397	-2.349686	-1.819519	-1.398894
	β	0.431177	0.040328	0.001626	0.349578	0.434013	0.549036
0,800	α	-1.543171	0.136212	0.018554	-1.882205	-1.530583	-1.272405
	β	0.372731	0.029233	0.000855	0.310063	0.369021	0.451543
Remarks: Simulation design: $\alpha = -2.5, \beta = 0.6, \sigma_e^2 = 0,01, n = 500, R = 50, M = 250$							

4.2. SIMEX in generalized linear models

It should be noted that evaluation of the SIMEX estimator at $\lambda = -1$ is motivated by formula (8) which assumes a *linear* model whereas the probit model is of course nonlinear. We therefore now use the formulation of the probit model provided by the generalized linear model.

It is well known that the probit model can be regarded as a special case of the generalized linear model (GLM) introduced by McCullagh and Nelder (1989). Since the (conditional) expected value of the observed binary variable Y is given by (see subsection 2.3)

$$\mu(x) \equiv E[Y \mid x] = P(Y_i^* > 0 \mid x_i) = \Phi(\alpha + \beta x_i) \tag{9}$$

we will get a linear relation when we consider

$$\Phi^{-1}(\mu(x)) = \alpha + \beta x_i$$

where the inverse distribution function Φ^{-1} is called the ”link function” for this special model.[8]

[8] See McCullagh and Nelder (1989) for details.

The STATA software package offers a simex routine for the probit model for the case that regressors are anonymized by addition of noise.[9] Some simulation runs already showed that the SIMEX procedure applied to GLM estimation of the probit model worked perfectly well if the regressor variable x is protected by noise addition and the PRAM parameter is set to $\pi = 1.00$ (no post randomization). See also table 3 discussed further below! This is in sharp contrast to results obtained from the ML approach reported above! Apparently for this (linear) formulation the evaluation at $\lambda = -1$ is adequate.

4.3. "Naive" GLM estimation of the probit model under PRAM

We then considered the case where the dependent variable y is anonymized by PRAM ($\pi < 1$) whereas the regressor x is observed without error.

Table 2. GLM estimation of probit model under PRAM (No noise addition , $\sigma_e^2 = 0$)

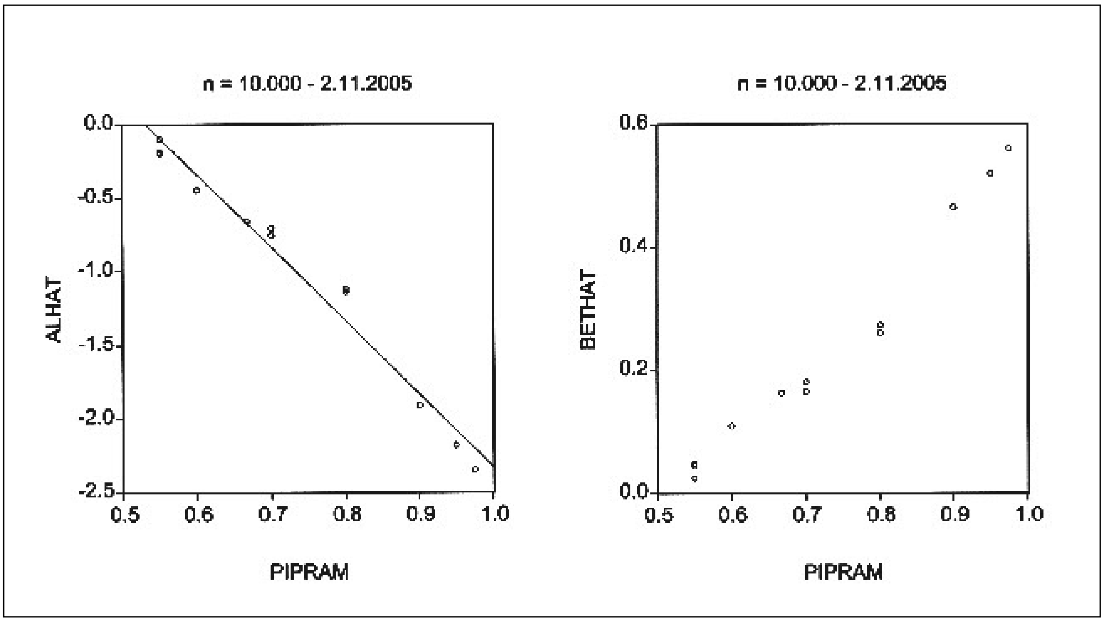
π	$\hat{\alpha}$	$\frac{\hat{\alpha}}{\alpha}$	$\hat{\beta}$	$\frac{\hat{\beta}}{\beta}$	$2\pi - 1$
0.975000	-2.348000	0.939200	0.560000	0.933333	0.950
0.950000	-2.180000	0.872000	0.520000	0.866667	0.900
0.900000	-1.910000	0.764000	0.465000	0.775000	0.800
0.800000	-1.120000	0.448000	0.260000	0.433333	0.600
0.800000	-1.140000	0.456000	0.272600	0.454333	0.600
0.700000	-0.759000	0.303600	0.180000	0.300000	0.400
0.700000	-0.710000	0.284000	0.164600	0.274333	0.400
0.667000	-0.660000	0.264000	0.162300	0.270500	0.333
0.600000	-0.447000	0.178800	0.109000	0.181667	0.200
0.550000	-0.098300	0.039320	0.023980	0.039967	0.100
0.550000	-0.185760	0.074304	0.043750	0.072917	0.100
0.550000	-0.197700	0.079080	0.047300	0.078833	0.100

Results from some simulation runs (with $n = 3,000$) are shown in table 2. The results are also displayed in figure 2. Since in these estimations no provision is made for taking account of post randomization[10], it is not at all surprising that we obtain biased estimates. However these estimates show a *monotonically decreasing* bias with respect to π (greatest bias for π near 0.50) and , more astonishingly, have approximately the same relative bias for both parameters (see third and fifth column of the table).

[9] This routine can be downloaded from STATA home page by the following commands: *net from http://www.stata.com/meror* and *net install merror*. However, there is no full documentation available.

[10] Ronning(2005) termed this approach "naive estimation of the probit model".

Figure 2. SIMEX/GLM estimates



We conjecture that this relative bias (defined by the ratio "estimate divided by true parameter value") is related to the factor[11]

$$\frac{1}{2\pi - 1}$$

which is shown in the last column of table 2. Our conjecture stems partly from the fact that the (conditional) expected value of the post randomized variables Y^m is given by[12]

$$E(Y^m \mid x) = 1 - \pi + (2\pi - 1)\Phi(\alpha + \beta x) \tag{10}$$

from which we obtain

$$\frac{E(Y^m \mid x)}{(2\pi - 1)} - \frac{(1-\pi)}{(2\pi - 1)} = \Phi(\alpha + \beta x).$$

A reasonable estimate would then be obtained from

$$\tilde{\alpha} = \frac{\hat{\alpha}}{2\pi - 1} \quad \text{and} \quad \tilde{\beta} = \frac{\hat{\beta}}{2\pi - 1} \quad .$$

[11] Neuhaus (1999) presents a detailed discussion of bias from 'misclassification' in binary regression models. Table 2 in this paper has an formula of the (approximated) bias also for the probit model although he considers the case of a *binary* regressor. His formula reads (in our terminology) as

$$\text{bias factor} = \frac{(2\pi - 1)\phi(\alpha)}{\phi\left[\Phi^{-1}\{(2\pi - 1)\Phi(\alpha) + 1 - \pi\}\right]} \quad .$$

Note that this expression contains the factor $2\pi - 1$. In particular the bias factor reduces to this expression if we set $\alpha = 0$. We plan to compare this formula with our results.

[12] See Ronning (2005).

4.4. SIMEX/GLM estimation of the anonymized probit model

Now let us turn to the STATA probit SIMEX routine which we applied to the simulation design as described above, that is, with both PRAM and noise addition. Results are summarized in table 3.

Table 3. Probit model - STATA SIMEX procedure (GLM) ($\sigma_\varepsilon^2 = 0.09$)

π		estimate	stand.dev.	minimum	maximum
1.000	α	-2.497237	.1005857	-2.734488	-2.182807
	β	.5991187	.0227196	.5286119	.6517459
	s_α	.0892389	.0054539	.0755891	.1074023
	s_β	.0203171	.0012342	.0174236	.0241485
0.90	α	-1.654937	.0753766	-1.832121	-1.490205
	β	.3979878	.0175081	.3523222	.4398246
	s_α	.075049	.0045581	.0652093	.0868179
	s_β	.0169182	.0010268	.0143047	.0196001
Remarks: Simulation design: $\alpha = -2.5, \beta = 0.6, \sigma_e^2 = 0,09, n = 3000, R = 100, M = 50$ 'stand. dev.' is obtained from bootstrap estimates.					

For $\pi = 1.00$ the results are satisfactory as mentioned already above. However for $\pi = 0.90$ we get results which do not fit to the scheme expected (see table 3): If we multiply the estimates by $1/(2 \cdot 0.9 - 1) = 1.25$ we obtain values which are considerably *lower* than the true values. For α we obtain $1.25 \cdot (-1.6549) = -2.0686$ and for β we obtain $1.25 \cdot 0.3980 = 0.4975$.

We have to analyze this problem in more detail in the future. Among other things we want to use the modified SIMEX approach in case of misclassification or post-randomization as proposed by Küchenhoff et al (2005).

References

Carroll, R.J. Ruppert, D., and Stefanski, L.A., (1995) *Measurement Error in Nonlinear Models*, Chapman and Hall, London

Cook, J.R. and Stefanski, L.A.(1994) "Simulation-Extrapolation Estimation in Parametric Measurement Error Models".*Journal of the American Statistical Association* **89**, 1314–1328.

Hausman, J.A., J. Abrevaya and F.M. Scott-Morton (1998): 'Misclassification of the Dependent Variable in a Discrete-Response Setting.' *Journal of Econometrics 87*, 239-269.

Kommission zur Verbesserung der informationellen Infrastruktur (editor) (2001). *Wege zu einer besseren informationellen Infrastruktur*. Nomos, Wiesbaden , cited as KVI(2001).

Küchenhoff, H., Mwalili, S.M., and Lesaffre, E. (2005). "A general method for dealing with misclassification in regression: The misclassification SIMEX". *Biometrics* (to appear)

Lechner, and S., Pohlmeier, W. (2003) "Schätzung ökonometrischer Modelle auf der Grundlage anonymisierter Daten" In: Gnoss, R. und G. Ronnning (editors) *Anonymisierung wirtschaftsstati-*

stischer Einzeldaten. Forum der Bundesstatistik, volume 42, 115-137.

Lechner, and S., Pohlmeier, W. (2005) ”Data Masking by Noise Addition and the Estimation of Nonparametric Regression Models”' *Jahrbücher für Nationalökonomie und Statistik* **255**, 517-528.

McCullagh, P., and Nelder, J.A. (1989). *Gemeralized Linear Models*. Chapman & Hall, London, second edition.

Neuhaus, J. (1999). ”Bias and efficiency loss due to misclassified responses in binary regression.” *Biometrika 86*, 843-855.

Ronning, G. (1991). *Mikroökonometrie*. Springer, Berlin.

Ronning, G. (2005) ”Randomized response and the binary probit model”.

Economics Letters **86**, 221-228.

Ronning, G., Gnoss, R. (Editors) (2003) *Anonymisierung wirtschaftsstatistischer Einzeldaten*. Statistisches Bundesamt. Forum der Bundesstatistik, volume 42, Wiesbaden.

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

Methods of Secure Computation and Data Integration

Alan F. Karr, Xiaodong S. Lin**, Jerome P. Reiter***, Ashish P. Sanil*****

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

**** University of Cincinnati, Cincinnati, OH, USA (xiaodong.lin@uc.edu).**

***** Duke University, Durham, NC, USA (jerry@stat.duke.edu).**

****** Bristol-Myers Squibb, Princeton, NJ, USA.**

Abstract. Reluctance of statistical agencies and other data owners to share their possibly confidential or proprietary data with others who own related databases is a serious impediment to conducting mutually beneficial analyses. This paper reviews methods for secure computation that potentially allow agencies to share data without compromising data confidentiality. The methods discussed include secure summation protocols, secure matrix product protocols, and synthetic data approaches.

1. Introduction

In many contexts, statistical agencies, survey organizations, businesses, and other data owners (henceforth all called agencies, to save writing) with related databases can benefit by integrating their data. For example, statistical models can be fit using more records or more attributes when databases are combined than when databases are analyzed separately. However, agencies may not be able or willing to combine their databases because of concerns about data confidentiality. These concerns can be present even when the agencies cooperate: all may wish to perform integrated analyses, but no one wants to break the confidentiality of others' data. In this paper, we review some approaches to data integration that aim to limit the risks of disclosures while maintaining the utility of the integrated data. In particular, we review secure computation techniques and approaches based on synthetic, i.e. simulated, data.

Data integration can be categorized into two general settings. Horizontally partitioned databases comprise the same attributes for disjoint sets of data subjects. For example, several local educational agencies might want to combine their students' data to improve the precision of analyses of the general student population. Vertically partitioned databases comprise the same data subjects, but each database contains different sets of attributes. For example, one agency might have employment information, another health data, and a third information about education, all for the same individuals. A statistical analysis predicting health status from all three sources of attributes is more informative than, or at least complementary to, separate analyses from each data source.

Various assumptions are possible about the participating agencies, for example, whether they use "correct" values in the computations, follow computational protocols, or collude against one another. We assume the agencies wish both to cooperate and to preserve the privacy of their individual databases. We assume that the agencies are "semi-honest:" each follows the agreed-on computational protocols properly, but may retain the results of intermediate computations. The results of analyses of horizontally or vertically partitioned data are to be shared among all participating agencies and possibly disseminated to the broader public.

2. Horizontally partitioned data

Several algorithms have been developed for performing secure analyses of horizontally partitioned data. Among them, Evfimievski *et al.* (2004) and Kantarcioglu and Clifton (2002) present methods for data mining with association rules; Lin *et al.* (2005) present methods for model based clustering; and, Karr *et al.* (2005b,c) present methods for secure regression analyses, including model diagnostics. The literature on privacy-preserving data mining (Lindell and Pinkas, 2000; Agrawal and

Srikant, 2000) contains related results. Here we summarize the approach of Karr *et al.* (2005b,c), who use the secure summation protocol (Benaloh 1987) to perform regression and other analyses on horizontally partitioned databases.

2.1. Secure summation protocol

Consider $K > 2$ cooperating, semi-honest agencies, such that Agency j has a value v_j. The agencies wish to compute $v = \sum_{j=1}^{K} v_j$ so that each Agency j learns only the minimum possible about the other agencies' values, namely the value of $v_{(-j)} = \sum_{\ell \neq j} v_\ell$. The secure summation protocol (Benaloh 1987) can be used to effect this computation.

Following the presentation in Karr *et al.* (2005b), let m be a very large number—which is known to all the agencies–such that $0 \leq v < m$. One agency is designated the master agency and numbered 1. The remaining agencies are numbered $2, ..., K$. Agency 1 generates a random number R from $[0, m)$. Agency 1 adds R to its local value v_1 and sends the sum $s_1 = (R + v_1) \bmod m$ to Agency 2. Since the value R is chosen randomly from $[0, m)$, Agency 2 learns nothing about the actual value of v_1.
For the remaining agencies $j = 2, ..., K-1$, the algorithm is as follows. Agency j receives

$$s_{j-1} = (R + \sum_{s=1}^{j-1} v_s) \bmod m,$$

from which it can learn nothing about the actual values of $v_1, ..., v_{j-1}$. Agency j then computes and passes on to Agency $j+1$

$$s_j = (s_{j-1} + v_j) \bmod m = (R + \sum_{s=1}^{j} v_s) \bmod m.$$

Finally, agency K adds v_K to $s_{K-1} (\bmod m)$, and sends the result s_K to agency 1. Agency 1, which knows R, then calculates v by subtraction:

$$v = (s_K - R) \bmod m$$

and shares this value with the other agencies.
For cooperating, semi-honest agencies, the use of arithmetic $\bmod m$ may be superfluous. It does, however, provide one layer of additional protection: without it, a large value of s_2 would be informative to Agency 2 about the value of R.

This method for secure summation faces an obvious problem if some agencies collude. For example, agencies $j-1$ and $j+1$ can together compare the values they send and receive to determine the exact value for v_j. Secure summation can be extended to work for an honest majority. Each agency divides v_j into shares. The sum for each share is computed individually. However, the path used is altered for each share so that no agency has the same neighbor twice. To compute v_j, the neighbors of agency j from every iteration would have to collude.

2.2 Secure regression via secure summation

Suppose the agencies wish to combine their data to fit the usual linear regression model:

$$Y = X\beta + \epsilon, \tag{1}$$

where

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np-1} \end{bmatrix}, \qquad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \tag{2}$$

and

$$\beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{bmatrix}, \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}. \tag{3}$$

Under the condition that $Cov(\varepsilon) = \sigma^2 I$, the least squares estimate for β is of course $\hat{\beta} = (X^T X)^{-1} X^T Y$. When the data are horizontally partitioned across K agencies, each agency j has its own share of data

$$X^j = \begin{bmatrix} x_{11}^j & \cdots & x_{1p}^j \\ \vdots & \ddots & \vdots \\ x_{n_j 1}^j & \cdots & x_{n_j p}^j \end{bmatrix}, \qquad y^j = \begin{bmatrix} y_1^j \\ \vdots \\ y_{n_j}^j \end{bmatrix}. \tag{4}$$

Here n_j denotes the number of data records for agency j.
Using (4) and altering indices as appropriate, we can rewrite (2) in partitioned form as

$$X = \begin{bmatrix} X^1 \\ \vdots \\ X^K \end{bmatrix} \qquad Y = \begin{bmatrix} Y^1 \\ \vdots \\ Y^K \end{bmatrix} \tag{5}$$

and (3) as

$$\beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \qquad \epsilon = \begin{bmatrix} \varepsilon^1 \\ \vdots \\ \varepsilon^K \end{bmatrix}. \tag{6}$$

Note that β does not change.

To compute $\hat{\beta}$, it is necessary to compute $X^T X$ and $X^T Y$. Because of the partitioning in (5), this can be done locally and the results combined entry-wise using secure summation. Specifically,

$$X^T X = \sum_{j=1}^{K} (X^j)^T X^j. \tag{7}$$

Each agency j can compute locally its own $(X^j)^T X^j$, and the results can be added entry-wise using secure summation to yield $X^T X$, which then can be shared among all the agencies. Similarly, since

$$X^T Y = \sum_{j=1}^{K} (X^j)^T Y^j,$$

$X^T Y$ can be computed by local computation of the $(X^j)^T Y^j$ and secure summation. This provides all the pieces necessary for each agency to compute $\hat{\beta}$.

The least squares estimate of σ^2 also can be computed securely. Since

$$S^2 = \frac{(Y - X\hat{\beta})^T (Y - X\hat{\beta})}{n - p}, \tag{8}$$

and $X^T X$ and $\hat{\beta}$ have been computed securely, the only thing left is to compute n and $Y^T Y$, again using secure summation. The agencies then can compute the estimated covariance matrix of the $\hat{\beta}$, which equals $S^2 (X^T X)^{-1}$.

It is also possible to share via secure summation statistics useful for model diagnostics, including correlations between predictors and the residuals, the coefficient of determination R^2, and the hat matrix $X(X^T X)^{-1} X^T$. Values of residuals are risky to share, since they reveal information about the dependent variable. Karr *et al.* (2005b) describe an approach for simulating plots of residuals versus predictors that mimic the real-data plots, based on the techniques of Reiter (2003a), which can be used for model diagnostics without releasing genuine residuals.

3. Vertically partitioned data

For vertically partitioned data, secure analysis methods exist for association rule mining (Vaidya and Clifton, 2002), K-means clustering (Vaidya and Clifton, 2003), and linear discriminant analysis (Du *et al.*, 2004). Du *et al.* (2004) and Sanil *et al.* (2004) present approaches to computing regression coefficients in vertically partitioned data, using methods that do not share the sample mean and covariance matrix. Here we review the approach of Karr *et al.* (2005a), which assumes the agencies are willing to share sample means and covariances of the integrated database but not the raw data. For simplicity, we describe the secure computation protocol for matrix products as a two-agency protocol. It is readily applicable to multi-agency cases.

3.1. Secure matrix product protocol

Following Karr *et al.* (2005a), let Agency A possess p data vectors $\{X_1, X_2, ..., X_p : X_i \in \Re^n\}$ and Agency B have q vectors $\{Y_1, Y_2, ..., Y_q : Y_i \in \Re^n\}$. Let $X = [X_1, X_2, ..., X_p]$ and $Y = [Y_1, Y_2, ..., Y_q]$ denote the respective data matrices, and assume $p < q$. We assume the matrices are of full rank; if not, the agencies remove any linearly dependent columns. We also assume the attributes in X and Y are disjoint; if not, the agencies coordinate so that any common attributes are included in only one matrix. Lastly, we assume that X and Y have disjoint attributes (columns) but the same data subjects (rows).

Agency A and Agency B wish to compute securely the $(p \times q)$ matrix $X^T Y$ and share it. It is necessary for the participating agencies to align their common data subjects in the same order. We assume each agency possesses a primary key, for example social security numbers, that is shared to facilitate this ordering.

In the interest of fairness to each participating agency, and to encourage trust among the agencies, we desire a protocol for secure matrix products that is symmetric in the amount of information exchanged. That is, the agencies should learn roughly the same amount about each other's data from the information shared in the protocol. A protocol that accomplishes this approximately is described by following procedure:

1. Agency A generates a set of $g = \left\lfloor \frac{n-p}{2} \right\rfloor$ orthonormal vectors $\{Z_1, Z_2, ..., Z_g : Z_i \in \Re^n\}$ such that $Z_i^T X_j = 0$ for any i, j. Agency A then sends the matrix $Z = [Z_1, Z_2, ..., Z_g]$ to Agency B.
2. Agency B computes $W = (I - ZZ^T)Y$, where I is an identity matrix. Agency B sends W to Agency A.
3. Agency A calculates $X^T W = X^T(I - ZZ^T)Y = X^T Y$ since $X_j^T Z_i = 0$ for any i, j.

The vector dot-product protocol is a special case of the matrix product. A method for generating Z is presented in Karr *et al.* (2005a).

It might appear that Agency B's data can be learned exactly since Agency A knows both W and Z. However, W has rank $(n-g) = (n-2p)/2$, so that Agency A cannot invert it to obtain Y.

Exact data values are not revealed in this protocol, but each agency can learn about the others' data from the constraints on the data values imposed by the values of the shared statistics. For any matrix product protocol where $X^T Y$ is learned by all agencies, each agency knows at minimum pq constraints, i.e those implied by the values of $X^T Y$. In addition, Agency A knows the g dimensional subspace that the Y_i lie in (as given by $W = (I - ZZ^T)Y$). Thus, Agency A has a total of $g + pq$ constraints on Y. Agency B knows the $(n-g)$ dimensional subspace that the X_i lie in (the subspace orthogonal to Z). Thus, Agency B has a total of $n - g + pq$ constraints on X.

In most settings involving vertically partitioned data, the $n \gg pq$, so that $g \approx \frac{n}{2}$. Hence, we can say that both agencies can place the other agencies' data in an approximately $\frac{n}{2}$ subspace, so that the protocol is approximately symmetric in the information shared.

The protocol is not immune to breaches of confidentiality if the agencies do not cooperate in a semi-honest fashion. For example, suppose Agency A sends to Agency B a Z such that $(I - ZZ^T)$ contains one column with all zeros except for a non-zero constant in one row. Agency A then learns the value of Agency B's data for the data subject in that row through $X^T W$. Other bogus Z could yield similar disclosures.

Even when the agencies are semi-honest, disclosures might be generated because of the values of the attributes themselves. As a simple example, suppose X includes a variable that equals zero for all but one of the data subjects. Even with a legitimate Z, the $X^T Y$ will reveal that subject's value of Y. Similar problems could arise when some X_i contains non-zeros for only a small number of records, particularly when reliable prior information on those records' values of some Y_j is known. For example, suppose two firms are the only ones in a certain industry in a certain city, with one being large and the other being small. Let X_i be an indicator with ones for those two firms and zeros for other firms. Let Y_j be some sensitive attribute positively correlated to the size of a firm. The $X_i^T Y_j$ equals the sum of the two firms' values, but most of that sum is contributed by the large firm. Thus, $X_i^T Y_j$ may be sufficiently close to the one firm's value of Y_j as to be a disclosure.

Disclosures resulting from subject matter considerations can be difficult to prevent. If Agency B does not know that Agency A has a variable like the X_i above, there is almost no way for Agency B to prevent disclosing some values in the matrix multiplications. A related problem occurs if one agency has attributes that are nearly linear combinations of the other agency's attributes. When this happens, accurate predictions of the data subjects' values can be obtained from linear regressions built from the securely computed matrix products.

3.2. Linear Regression with arbitrary subsets of attributes

In this section, we apply the secure matrix product protocol to conduct secure linear regression analyses. Let the matrix of all variables in the possession of the agencies be $D=[D_1,\cdots,D_p]$, with

$$D_i = \begin{pmatrix} d_{i1} \\ \vdots \\ d_{in} \end{pmatrix}, \quad 1 \le i \le p \ . \tag{9}$$

The data matrix D is distributed through K agencies: $A_1, A_2, \cdots, A_K$. Each agency, A_j, possesses p_j disjoint columns of D, where $\sum_K p_j = p$.

A regression model of some dependent variable, say $D_i \subset D$, on a collection of the other variables, say $D_0 \subseteq D - D_i$, is of the form

$$D_i = D_0\beta_0 + \in_0 \tag{10}$$

where $\in_0 \sim N(0,\sigma_0^2)$. Typically, the model includes an intercept term. This is achieved by including a column of ones in D_0. Without loss of generality, we assume that $D_1^T = (1,1,...,1)$ and that it is owned by Agency A_1.

Our goal is to regress any D_i on some arbitrary subset D_0 using secure computations. It is well known that the maximum likelihood estimates of σ_0^2 and β_0, as well as the standard errors of the estimated coefficients, can be easily obtained from the sample covariance matrix of D, for example using the sweep algorithm (Beaton, 1964). Hence, the agencies need only the elements of the sample covariance matrix of D to perform the regression. Each agency computes and shares the block-diagonal elements of the matrix corresponding to its variables, and the agencies use secure matrix computations to compute the off-diagonal elements, thus completing the sample covariance matrix.

The types of model diagnostic measures available in vertically partitioned data settings depend on the amount of information the agencies are willing to share. Diagnostics based on residuals require the predicted values, $D_0\hat{\beta}_0$. These can be obtained using the secure matrix product protocol, since

$$D_0\hat{\beta}_0 = D_0(D_0{}^T D_0)^{-1} D_0{}^T D_i. \tag{11}$$

Alternatively, once the $\hat{\beta}_0$ is shared, each agency could compute the portion of $D_0\hat{\beta}_0$ based on the variables in its possession, and the vectors can be summed across agencies using the secure summation protocol outlined in Section 2.1.

Once the predicted values are known, the agency with the dependent variable D_i can calculate the residuals $E_0 = D_i - D_0\hat{\beta}_0$. If that agency is willing to share the residuals with the other agencies, each agency can perform plots of residuals versus its independent variables and report the nature of any lack of fit to the other agencies. Sharing E_0 also enables all agencies to obtain Cook's distance meas-

ures, since these are solely a function of E_0 and the diagonal elements of $H = D_0({D_0}^T D_0)^{-1} {D_0}^T$, which can be securely computed.
The agency with D_i may be unwilling to share E_0 with the other agencies, since sharing essentially reveals the values of D_i. In this case, one option is to compute the correlations of the residuals with the independent variables using the secure matrix product protocol. Additionally, the agency with D_i can make a plot of E_0 versus $D_0 \hat{\beta}_0$, and a normal quantile plot of E_0, and report any evidence of model violations to the other agencies. The number of residuals exceeding certain thresholds, i.e. outliers, also can be reported.

3.3. Synthetic data approach for vertically partitioned data

The secure matrix protocol requires that agencies pre-specify the regression analyses of interest. In some settings this could be problematic. For example, it may be necessary to transform some variables to obtain a regression that fits the data appropriately. Agencies can apply the secure matrix protocol more than once, e.g. on the original data to enable model checking and then on transformed data to improve the model, but repeating the protocol generates additional constraints on X and Y that reduce confidentiality protection.

To introduce flexibility of modeling, Kohnen and Reiter (2004) and Kohnen (2005) propose that agencies share synthetic, i.e. simulated, data that mimic the relationships in the real data. To motivate their idea, consider the case where Agency A is willing to share its X with Agency B, but Agency B is not willing to share its Y with Agency A. The approach proceeds as follows:

1. Agency A sends X to Agency B.
2. Agency B fits a model $f(Y | X)$ that relates Y to X, based on the passed X and its genuine Y.
3. Agency B simulates a new value of Y from the model $f(Y | X)$ and passes these simulated data to Agency A. Agency B repeats this M times, so that M versions of the synthetic Y are passed to Agency A.
4. Agency A analyzes the M datasets formed by combining the X with each version of Y using the methods for analyzing multiply-imputed, partially synthetic datasets (Reiter, 2003b).

At stage 2, Agency B can either (i) send $f(Y | X)$ and its parameters to Agency A, or (ii) simulate new values of Y from the model $f(Y | X)$ and pass these simulated values to Agency A. The latter strategy is preferred when the model and its parameters represent a disclosure risk (e.g., parameters in log-linear models for categorical data correspond to cell counts in tables, which may be sensitive) or when the model is too complicated to send (e.g., a semi-parametric model). We assume that Agency B will generate and pass new values of Y.
The multiple versions of Y are needed to enable Agency A to estimate uncertainties in parameter estimates correctly. One version of Y is insufficient, because the process of drawing values of Y from a distribution introduces additional variability into parameter estimates that is not easily estimated from one dataset. The prescription for releasing multiple copies follows the rationale for generating multiply-imputed, partially synthetic datasets (Reiter, 2003b, 2004).
As an extension to this case, the agencies may be willing to share X with each other but not with the broader public. To release data to the public, Agency A can simulate completely synthetic data (Raghunathan, Reiter, Rubin, 2003). That is, it can simulate values of X and values of Y using its original

values of X and the simulated values of Y it received from Agency B. Methods for doing this, as well as methods for obtaining inferences from such datasets, are described by Kohnen (2005).

We next move to the more general case where Agency A is not be willing to share X with Agency B. The key difference in the algorithm is in step 2, since X cannot be passed without some way of protecting it. Kohnen (2005) proposes that Agency A generate disguiser copies of X–that is, new values of X that mimic the distribution of the genuine X–and send them to Agency B along with the genuine X. Agency B then fits models for $Y \mid X$ for each of the copies of X and sends simulated values of Y back for each back to Agency A. Agency A discards all the simulated Y except for the ones that correspond to the genuine X. With L perfect disguisers, Agency B has a $1/L$ chance of guessing which of the L datasets contains the true X. For sufficiently large L, this may provide adequate protection.

Obviously, the protection of X is compromised if Agency B can distinguish the genuine X from the disguisers. This could be accomplished if Agency B knows certain values of X and therefore can hunt for them in the passed copies. To prevent this in a semi-honest setting, Agency B can tell Agency A which values it has, so that these values can be included in all passed copies of X. Agency B also might be able to determine the genuine X by looking for unusual results in the various versions of $f(Y \mid X)$. For example, it may be the case that the genuine X has the strongest correlations with Y. Kohnen (2005) describes several such risks, as well as some methods for reducing them.

Ideally, the disguiser X values are generated from $f(X \mid Y)$. This is not easy to do, since Agency A does not know Y. It may be possible to approximate this distribution, perhaps using methods from standard disclosure limitation strategies. Research on generating good disguisers is a high priority item for this approach.

4. Conclusion

In this paper, we summarized several approaches to secure data integration. These approaches generate many practical challenges, which as of this writing have not been fully met. Some of these include:

- How do we specify models without viewing the data, which is implicit in the secure computation methods?
- How do we perform secure computation for models that don't have sums and products as sufficient statistics?
- How do we incorporate errors when matching records in vertically partitioned data?
- How do we account for differences in data quality and definitions?
- How do we account for disclosure risks from models that fit too well?

Statisticians have only recently started investigating the data integration setting (computer scientists have been active in this area for longer). It is an area that is likely to grow in relevance, as data owners of all types seek to gain the benefits from data integration. And, as the questions listed above indicate, it is an area rich in topics for statistical research.

Acknowledgements

This research was supported by NSF grant IIS–0131884 to the National Institute of Statistical Sciences.

References

Agrawal, R. and Srikant, R. (2000) "Privacy-Preserving Data Mining", in *Proceedings of the 2000 ACM SIGMOD on Management of Data*, 439–450.

Beaton, A. (1964) "The Use of Special Matrix Operations in Statistical Calculus", *Research Bulletin RB-64-51*, Educational Testing Service, Princeton, NJ.

Benaloh, J. (1987) "Secure Sharing Homomorphisms: Keeping Shares of a Secret Secret", in *Advances in Cryptography: CRYPTO86*, ed. A. M. Odlyzko, New York: Springer-Verlag, **263**, 251–260.

Du, W., Han, Y., and Chen, S. (2004) "Privacy-Preserving Multivariate Statistical Analysis: Linear Regression and Classification", in *Proceedings of the 4th SIAM Conference on Data Mining*, 222–233.

Evfimievski, A., Srikant, R., Agrawal, R., and Gehrke, J. (2004) "Privacy-Preserving Mining of Association Rules", *Information Systems*, **29**, June 2004.

Kantarcioglu, M. and Clifton, C. (2002) "Privacy-Preserving Distributed Mining of Association Rules on Horizontally-Partitioned Data", in *Proceedings of Eigth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Karr, A., Lin, X., Sanil, A., and Reiter, J. (2005a) "Privacy-Preserving Analyses of Vertically Partitioned Data Using Secure Matrix Product Protocols", Technical Report, National Institute of Statistical Sciences.

Karr, A., Lin, X., Sanil, A., and Reiter, J. (2005b) "Secure Regressions on Distributed Databases", *Journal of Computational and Graphical Statistics*, **14**, 263–279.

Karr, A., Lin, X., Sanil, A., and Reiter, J. (2005c) "Secure Statistical Analyses of Distributed Databases." in *Statistical Methods in Counterterrorism*, ed. D. Olwell and A. Wilson, ASA-SIAM Series on Statistics and Applied Probability, to appear.

Karr, A., Feng, J., Lin, X., Sanil, A., Young, S., and Reiter, J. (2005) "Secure Analyses of Distributed Chemical Databases without Data Integration", *Journal of Computer-Aided Molecular Design*, to appear.

Kohnen, C. (2005) "Using Multiply-Imputed, Synthetic Data to Facilitate Data Sharing", PhD Dissertation, Institute of Statistics and Decision Sciences, Duke University.

Kohnen, C. and Reiter, J. (2004) "Sharing Confidential Data Among Multiple Agencies Using Multiply-Imputed, Synthetic Data", *Proceedings of the Joint Statistical Meetings*, American Statistical Association.

Lin, X., Clifton, C., and Zhu, Y. (2004) "Privacy-Preserving Clustering with Distributed EM Mixture Models. *Knowledge and Information Systems*, **8**, 68–81.

Lindell, Y. and Pinkas, B. (2000) "Privacy-Preserving Data Mining" in *Advances in Cryptology: Crypto2000*, New York: Springer-Verlag, **1880**, 36–54.

Raghunathan, T., Reiter, J., and Rubin, D. (2003) "Multiple Imputation for Statistical Disclosure Limitation" *Journal of Official Statistics*, **19**, 1–16.

Reiter, J. (2003a) "Model Diagnostics for Remote Access Regression Servers", *Statistics and Computing*, **13**, 371–380.

Reiter, J. (2003b) "Inference for Partially Synthetic, Public Use Microdata Sets", *Survey Methodology*, **29**, 181–188.

Reiter, J. (2004) "Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation", *Survey Methodology*, **30**, 235–242.

Sanil, A., Karr, A., Lin, X., and Reiter, J. (2004) "Privacy-Preserving Regression Modeling Via Distributed Computation." *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 677–682.

Vaidya, J. and Clifton, C. (2002) "Privacy-Preserving Association Rule Mining Over Vertically Partitioned Data", *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 639–644.

Vaidya, J. and Clifton, C. (2003) "Privacy-Preserving k-Means Clustering Over Vertically Partitioned Data", *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Empirical Disclosure Risk Assessment of the IPSO Synthetic Data Generators

Josep Domingo-Ferrer, Vicenç Torra**, Josep M. Mateo-Sanz*, Francesc Sebé**

*** Rovira i Virgili University of Tarragona, Dept. of Computer Engineering and Maths, Av. Països Catalans 26, E-43007 Tarragona, Catalonia, ({josep.domingo,josepmaria.mateo,francesc.sebe}@urv.net)**

**** IIIA-CSIC, Campus UAB, E-08193 Bellaterra, Catalonia, (vtorra@iiia.csic.es)**

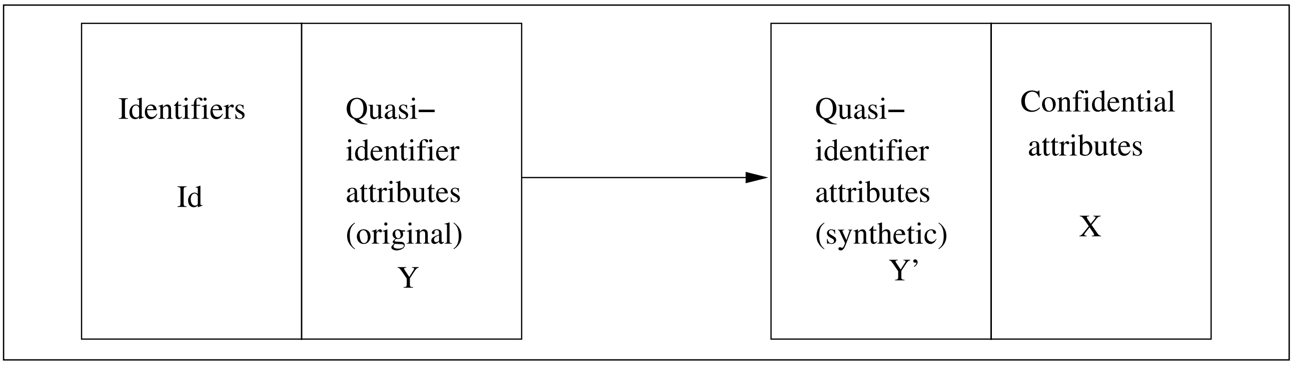
Abstract. Information Preserving Statistical Obfuscation (IPSO) is a family of three methods IPSO-A, IPSO-B, IPSO-C for numerical synthetic data generation designed by Burridge in 2003. This paper reports on empirical work carried out to assess the re-identification risk of each method in different worst-case disclosure scenarios, with different datasets and using different record linkage methods. The conclusions of this study give some insight on how IPSO and other synthetic data generators can be tuned to minimize re-identification risk. Further, we discuss how a similar analysis could be conducted for synthetic generators of categorical data.

1. Introduction

Synthetic microdata generators usually care about preserving a model or some statistics, but they seldom pay attention to disclosure risk. The usual alibi is to argue that, since released microdata are synthetic, no real re-identification is possible. While this may be reasonable if synthetic generation is performed on the confidential outcome attributes, it is an unrealistic assumption if synthetic data generation is performed on the quasi-identifier attributes. In the latter case, re-identification can indeed happen if a snooper is able to link an external identified data source with some record in the released dataset using the quasi-identifier attributes: coming up with a correct pair (identifier, confidential attributes) is indeed a re-identification.

The disclosure model we work with is depicted in Figure 1. We assume that the released dataset on the right-hand side of the figure consists of confidential attributes X and non-confidential quasi-identifier attributes Y'; quasi-identifier attributes Y' have been masked using a partially synthetic data generation method. A snooper has obtained the external identified dataset on the left-hand side of the figure, which consists of one or several identifier attributes Id and several quasi-identifier attributes Y. Attributes Y are original and are not necessarily the same as attributes Y' in the released dataset. The snooper attempts to link records in the released dataset with records in the external identified dataset. Linkage is done by matching quasi-identifier attributes Y and Y'. The snooper's goal is to pair identifier values with confidential attribute values (*e.g.* to pair citizens' names with health conditions).

Figure 1. Re-identification scenario. Quasi-identifiers Y and Y' can have shared attributes or not


Identifiers
Id
Quasi–identifier attributes (original)
Y
Quasi–identifier attributes (synthetic)
Y'
Confidential attributes
X

1.1. Contribution and plan of this paper

For the sake of concreteness, this paper focuses on a particular family of synthetic data generators, namely Information Preserving Statistical Obfuscation (IPSO, Burridge (2003)). IPSO is a family of three methods IPSO-A, IPSO-B, IPSO-C for numerical synthetic data generation which preserve, to a varying extent, a multivariate multiple regression model taking confidential attributes as independent variables and quasi-identifier attributes as dependent variables.

We have run IPSO-A, IPSO-B and IPSO-C on two different datasets and we report on the results of record linkage experiments on those datasets using different quasi-identifiers and different record linkage methods. In particular we consider the case where no quasi-identifier attributes are shared between the released dataset and the external identified source. The purpose of this study is to give some insight about re-identification which helps data protectors tune their synthetic data generators to make life more difficult for snoopers. We also discuss extensions of our study for synthetic generators of categorical data.

Section 2 briefly recalls IPSO-A, IPSO-B and IPSO-C. Section 3 describes the two datasets used. Record linkage methods employed in our analysis are explained in Section 4. Experimental results are given in Section 5. Conclusions and extensions are listed in Section 6.

2. The IPSO methods

Three variants of a procedure called Information Preserving Statistical Obfuscation (IPSO) are proposed in Burridge (2003). The basic form of IPSO will be called here IPSO-A. Informally, suppose two sets of attributes X and Y, where the former are the confidential outcome attributes and the latter are quasi-identifier attributes. Then X are taken as independent and Y as dependent attributes. A multiple regression of Y on X is computed and fitted Y'_A attributes are computed. Finally, attributes X and Y'_A are released by IPSO-A in place of X and Y.

In the above setting, conditional on the specific confidential attributes x_i, the quasi-identifier attributes Y_i are assumed to follow a multivariate normal distribution with covariance matrix $\Sigma = \{\sigma_{jk}\}$ and a mean vector $x_i B$, where B is the matrix of regression coefficients.

Let $\hat{B}$ and $\hat{\Sigma}$ be the maximum likelihood estimates of B and Σ derived from the complete dataset (y, x). If a user fits a multiple regression model to (y'_A, x), she will get estimates $\hat{B}_A$ and $\hat{\Sigma}_A$ which, in general, are different from the estimates $\hat{B}$ and $\hat{\Sigma}$ obtained when fitting the model to the original data (y, x). The second IPSO method, IPSO-B, modifies y'_A into y'_B in such a way that the estimate $\hat{B}_B$ obtained by multiple linear regression from (y'_B, x) satisfies $\hat{B}_B = \hat{B}$.

A more ambitious goal is to come up with a data matrix y'_C such that, when a multivariate multiple regression model is fitted to (y'_C, x), *both* sufficient statistics $\hat{B}$ and $\hat{\Sigma}$ obtained on the original data (y, x) are preserved. This is done by the third IPSO method, IPSO-C.

3. The test datasets

We have used two reference datasets (Brand, et al. 2002) used in the European project CASC:

1. The "Census" dataset contains 1080 records with 13 numerical attributes labeled $v1$ to $v13$. This dataset was used in CASC and in several other works (Domingo-Ferrer, et al. 2001, Dandekar, et al. 2002, Yancey, et al. 2002, Laszlo & Mukherjee 2005, Domingo-Ferrer & Torra 2005, Domingo-Ferrer, et al. 2005).
2. The "EIA" dataset contains 4092 records with 15 attributes. The first five attributes are cat-

egorical and will not be used. We restrict to the last 10 numerical attributes, which will be labeled $v1$ to $v10$. This dataset was used in CASC, in Dandekar et al. (2002), Domingo-Ferrer et al. (2005) and partially in Laszlo & Mukherjee (an undocumented subset of 1080 records from "EIA", called "Creta" dataset, was used in the latter paper).

4. Record linkage methods tried

The record linkage methods used fall into two paradigms:

- *Record linkage with shared attributes*. We assume that the external identified dataset $\mathbf{A}$ and the released dataset $\mathbf{B}$ share some attributes which are used for re-identification. Two methods corresponding to this approach have been tried:
    - Distance-based record linkage
    - Probabilistic record linkage
- *Record linkage without shared attributes*. No common attributes between the external identified dataset and the released dataset are assumed. A new correlation-based record linkage method has been designed and tried here.

We describe distance-based record linkage, probabilistic record linkage and correlation-based record linkage in the sections below. More details on distance-based and probabilistic record linkage can be found in Torra & Domingo Ferrer 2003.

4.1. Distance-based record linkage

This approach, originally described in Tend92 and Full93, consists of computing distances between records in $\mathbf{A}$ and $\mathbf{B}$. Then, pairs of records at minimum distance are considered linked pairs. Of course, the distance between a pair of records must be computed based on shared attributes between those records, so that this approach does not work without shared attributes between the external data source and the released dataset.

Naturally, the application of this method depends on the existence of the distance function. Thus, a distance is assumed in each attribute V_i. We denote this distance by d_{V_i}. Assuming equal weight for all attributes, a record-level distance between records a and b can be constructed as:

$$d(a,b)=\sum_{i=1}^{n} d_{V_i}(V_i^A(a),V_i^B(b))$$

Depending on the data type of attributes, different within-attribute distances must be used. For numerical attributes, the Euclidean distance is a reasonable choice. See Domingo-Ferrer & Torra (2001) and Dom ingo-Ferrer & Torra (2002) on distances for categorical attributes. Whatever the distance and attribute type, one should use some kind of standardization to avoid scaling problems and give equal weight to attributes when combining them. For numerical data, one can

- Standardize each attribute before computing distances (this is done by subtracting the attribute mean and dividing by the attribute standard deviation). This type of distance-based record linkage will be called DRL1 in what follows.
- Compute distances on the unstandardized attributes and standardize distances by subtracting their average and dividing by their standard deviation. This approach will be called DRL2 in what follows.

4.2. Probabilistic record linkage

Probabilistic record linkage, called PRL in what follows, is described in Fellegi & Sunter (1969), Jaro (1989) and Winkler (1995). See the above mentioned references for details. Like distance-based record linkage, PRL assumes that the datasets to be linked share at least one quasi-identifier attribute.

The distinguishing features of PRL with respect to DRL1 and DRL2 are that: i) PRL can work on any data type (numerical or categorical) without any adaptation; ii) PRL does not require any assumptions on the relative weight of attributes (in particular, it requires no standardization). Its main drawback is its computational burden.

4.3. Correlation-based record linkage

This is a new proposal, called CRL in what follows, that we make for record linkage between numerical datasets without shared attributes. We assume that both datasets $\mathbf{A}$ and $\mathbf{B}$ have their own numerical quasi-identifier attributes. We also assume that both datasets consist of n records corresponding to the same set of individual respondents.

The method finds the pair (i, j) of quasi-identifier attributes in $\mathbf{A}$ and $\mathbf{B}$ with highest correlation. Then $\mathbf{A}$ is sorted by its i-th quasi-identifier attribute and $\mathbf{B}$ is sorted by its j-th quasi-identifier attribute. If there remain subsets of records with equal rank in either dataset, find the pair of attributes with the second highest correlation and use them to decide the ordering within those subsets of records. This process can be iterated until no two records in either dataset have the same rank or we have used all quasi-identifier attributes; in the latter case, use a random ordering for any remaining records with equal rank. At the end of this process, all n records in $\mathbf{A}$ and $\mathbf{B}$ are ranked. The final step is to link the k-th record in $\mathbf{A}$ with the k-th record in $\mathbf{B}$, for $k=1$ to n.

In practice, the method can be applied even without knowledge of the exact correlations between $\mathbf{A}$ and $\mathbf{B}$. The semantics of the attributes in both files may give indications that a certain pair of attributes has a higher correlation than another pair.

5. Experimental results

We implemented IPSO-A, IPSO-B and IPSO-C above for generation of partially synthetic data. We then applied them to the “Census” and “EIA” datasets to obtain several versions of partially synthetic data. Next, we considered re-identication scenarios with shared and non-shared attributes and tried distance-based, probabilistic and correlation-based record linkage on them. This section describes in detail this experimental work and the results that were obtained.

5.1. Results on “Census”

We took the “Census” dataset and used the correlations between its 13 attributes to compute a dendrogram. We followed the dendrogram rather than the semantics of attributes in “Census” to select quasi-identifier attributes and confidential attributes. The rationale of this is that we were looking for worst-case scenarios to test the safety of the synthetic generators IPSO-A, IPSO-B and IPSO-C: the worst case (most likely to yield correct re-identifications) happens when the snooper uses quasi-identifier attributes which are highly correlated to the remaining attributes in the dataset. Thus, we chose quasi-identifier attributes with central positions in the dendrogram; this strategy led us to two different choices of confidential outcome attributes X and quasi-identifier attributes Y which gave two different scenarios $S1$ and $S2$. Table 1 summarizes the attributes in each dataset for each scenario.

Table 1. Splittings of "Census" into datasets $\mathbf{A}$ and $\mathbf{B}$ and attributes per dataset. In individual experiments, several subsets of quasi-identifier attributes Y were considered

Scenario	Data set	Shared attributes		Non-shared attributes	
		Quasi-id. Y	Conf. attr. X	Quasi-id. Y	Conf. attr. X
S1	**A**	$v1, v3, v4, v6, v7$ $v9, v11, v12, v13$		$v3, v4, v9, v12$	
	B	$v1, v3, v4, v6, v7$ $v9, v11, v12, v13$	$v2, v5, v8, v10$	$v1, v6, v7$ $v11, v13$	$v2, v5, v8, v10$
S2	**A**	$v4, v7, v12, v13$		$v4, v12$	
	B	$v4, v7, v12, v13$	$v1, v2, v3, v5, v6$ $v8, v9, v10, v11$	$v7, v13$	$v1, v2, v3, v5, v6$ $v8, v9, v10, v11$

We then took the quasi-identifier attributes in datasets $\mathbf{B}$ in Table 1 and used methods IPSO-A, IPSO-B and IPSO-C on them. In other words, we fitted a multivariate multiple regression model to them by taking as independent attributes the confidential attributes X and as dependent attributes the quasi-identifier attributes Y.

We first explain the notation used in the tables of results in this section:

- A, B, C as a subscript denote that the attribute was generated using IPSO-A, IPSO-B or IPSO-C, respectively; no subscript means that the attribute is original.
- $S1$ as a superscript means that this attribute was obtained by fitting a multivariate multiple regression model taking as independent attributes four confidential attributes X (specifically, $v2, v5, v8, v10$, see scenario $S1$ in Table 1).
- $S2$ as a superscript means that this attribute was obtained by fitting a multivariate multiple regression model taking as independent attributes nine confidential attributes X (specifically, $v1, v2, v3, v5, v6, v8, v9, v10, v11$, see scenario $S2$ in Table 1).

Table 2 shows the results of record linkage experiments between the "Census" dataset and a partially synthetic version of it generated using IPSO-A. The table shows only the quasi-identifiers used in each experiment, which are subsets of those specified in Table 1.

Quasi-identifiers in Table 2 were selected using the cross-correlation matrix between the original quasi-identifier attributes and the quasi-identifier attributes generated using method IPSO-A. The rationale of our quasi-identifier choices is that at least some of the quasi-identifiers in datasets $\mathbf{A}$ and $\mathbf{B}$ should be highly correlated. Note that this strategy in quasi-identifier selection can be followed by a real snooper, since he can compute the cross-correlation matrix between the external identified dataset and the released, partially synthetic datasets.

Table 2. Re-identification experiments using dataset "Census" and method IPSO-A. Results in number of correct re-identifications over an overall number of 1080 records. Percentage of correct re-identifications between parentheses. DRL1: attribute-standardizing implementation of distance-based record linkage (DRL); DRL2: distance-standardizing implementation of DRL; PRL: probabilistic record linkage; CRL: correlation-based record linkage

Quasi-identifier in external **A**	Quasi-identifier in released **B**	DRL1	DRL2	PRL	CRL
$v7, v12$	$v7_A^{S1}, v12_A^{S1}$	144 (13.3%)	144 (13.3%)	144 (13.3%)	7 (0.6%)
$v4, v7, v11, v12$	$v4_A^{S1}, v7_A^{S1}, v11_A^{S1}, v12_A^{S1}$	85 (7.8%)	82 (7.5%)	68 (6.2%)	7 (0.6%)
$v4, v7, v12, v13$	$v4_A^{S1}, v7_A^{S1}, v12_A^{S1}, v13_A^{S1}$	104 (9.6%)	106 (9.8%)	116 (10.7%)	7 (0.6%)
$v4, v7, v11, v12, v13$	$v4_A^{S1}, v7_A^{S1}, v11_A^{S1}, v12_A^{S1}, v13_A^{S1}$	79 (7.3%)	80 (7.4%)	85 (7.8%)	7 (0.6%)
$v1, v3, v4, v6, v7$ $v9, v11, v12, v13$	$v1_A^{S1}, v3_A^{S1}, v4_A^{S1}, v6_A^{S1}, v7_A^{S1}$ $v9_A^{S1}, v11_A^{S1}, v12_A^{S1}, v13_A^{S1}$	36 (3.3%)	31 (2.8%)	82 (7.2%)	7 (0.6%)
$v7, v12$	$v7_A^{S2}, v12_A^{S2}$	79 (7.3%)	79 (7.3%)	79 (7.3%)	40 (3.7%)
$v4, v13$	$v4_A^{S2}, v13_A^{S2}$	50 (4.6%)	50 (4.6%)	50 (4.6%)	5 (0.4%)
$v7, v12, v13$	$v7_A^{S2}, v12_A^{S2}, v13_A^{S2}$	82 (7.5%)	81 (7.5%)	85 (7.8%)	40 (3.7%)
$v4, v7, v12, v13$	$v4_A^{S2}, v7_A^{S2}, v12_A^{S2}, v13_A^{S2}$	85 (7.8%)	86 (7.9%)	93 (8.6%)	40 (3.7%)
$v4$	$v7_A^{S1}$	N/A	N/A	N/A	7 (0.6%)
$v7$	$v4_A^{S1}$	N/A	N/A	N/A	4 (0.3%)
$v4, v12$	$v7_A^{S1}, v13_A^{S1}$	N/A	N/A	N/A	37 (3.4%)
$v3, v4, v9, v12$	$v1_A^{S1}, v6_A^{S1}, v7_A^{S1}, v11_A^{S1}, v13_A^{S1}$	N/A	N/A	N/A	37 (3.4%)
$v1, v6, v7, v11, v13$	$v3_A^{S1}, v4_A^{S1}, v9_A^{S1}, v12_A^{S1}$	N/A	N/A	N/A	4 (0.3%)
$v4, v12$	$v7_A^{S2}, v13_A^{S2}$	N/A	N/A	N/A	43 (3.9%)
$v7, v13$	$v4_A^{S2}, v12_A^{S2}$	N/A	N/A	N/A	8 (0.7%)

The results for IPSO-B were very similar to those for IPSO-A, and will not be reported here for the sake of brevity. The results for IPSO-C are different and are shown in Table 3.

Table 3. Re-identification experiments using dataset "Census" and method IPSO-C. Results in number of correct re-identifications over an overall number of 1080 records.

Quasi-identifier in external **A**	Quasi-identifier in released **B**	DRL1	DRL2	PRL	CRL
$v7, v12$	$v7_C^{S1}, v12_C^{S1}$	32 (2.9%)	32 (2.9%)	32 (2.9%)	13 (1.2%)
$v4, v7, v11, v12$	$v4_C^{S1}, v7_C^{S1}, v11_C^{S1}, v12_C^{S1}$	39 (3.6%)	39 (3.6%)	36 (3.3%)	13 (1.2%)
$v4, v7, v12, v13$	$v4_C^{S1}, v7_C^{S1}, v12_C^{S1}, v13_C^{S1}$	35 (3.2%)	35 (3.2%)	33 (3.0%)	13 (1.2%)
$v4, v7, v11, v12, v13$	$v4_C^{S1}, v7_C^{S1}, v11_C^{S1}, v12_C^{S1}, v13_C^{S1}$	40 (3.7%)	40 (3.7%)	43 (3.9%)	13 (1.2%)
$v1, v3, v4, v6, v7$ $v9, v11, v12, v13$	$v1_C^{S1}, v3_C^{S1}, v4_C^{S1}, v6_C^{S1}, v7_C^{S1}$ $v9_C^{S1}, v11_C^{S1}, v12_C^{S1}, v13_C^{S1}$	19 (1.7%)	19 (1.7%)	50 (4.6%)	13 (1.2%)
$v7, v12$	$v7_C^{S2}, v12_C^{S2}$	42 (3.9%)	42 (3.9%)	42 (3.9%)	12 (1.1%)
$v4, v13$	$v4_C^{S2}, v13_C^{S2}$	17 (1.6%)	17 (1.5%)	17 (1.5%)	6 (0.5%)
$v7, v12, v13$	$v7_C^{S2}, v12_C^{S2}, v13_C^{S2}$	31 (2.8%)	31 (2.8%)	36 (3.3%)	12 (1.1%)
$v4, v7, v12, v13$	$v4_C^{S2}, v7_C^{S2}, v12_C^{S2}, v13_C^{S2}$	26 (2.4%)	26 (2.4%)	33 (3.0%)	12 (1.1%)
$v4$	$v7_C^{S1}$	N/A	N/A	N/A	10 (0.9%)
$v7$	$v4_C^{S1}$	N/A	N/A	N/A	3 (0.3%)
$v4, v12$	$v7_C^{S1}, v13_C^{S1}$	N/A	N/A	N/A	3 (0.3%)
$v3, v4, v9, v12$	$v1_C^{S1}, v6_C^{S1}, v7_C^{S1}, v11_C^{S1}, v13_C^{S1}$	N/A	N/A	N/A	3 (0.3%)
$v1, v6, v7, v11, v13$	$v3_C^{S1}, v4_C^{S1}, v9_C^{S1}, v12_C^{S1}$	N/A	N/A	N/A	18 (1.7%)
$v4, v12$	$v7_C^{S2}, v13_C^{S2}$	N/A	N/A	N/A	6 (0.5%)
$v7, v13$	$v4_C^{S2}, v12_C^{S2}$	N/A	N/A	N/A	10 (0.9%)

It can be observed that, for the same quasi-identifier attributes, method IPSO-C results in less re-identifications than methods IPSO-A and IPSO-B. Since, IPSO-C preserves more statistics than the other two methods, it is clearly the best choice.

5.2. Results on “EIA”

We took the “EIA” dataset and computed a correlation-based dendrogram of its 10 numerical attributes $v1,\cdots,v10$. Like for “Census”, we used the “EIA” dendrogram rather than the semantics of “EIA” attributes to select quasi-identifier attributes and confidential attributes. A single scenario (choice of confidential attributes X) was defined. Table 4 summarizes the quasi-identifiers considered in each dataset for the paradigms with shared and non-shared attributes.

We then took the quasi-identifier attributes in dataset $\mathbf{B}$ in Table 4 and used methods IPSO-A, IPSO-B, IPSO-C on them. In other words, we fitted a multivariate multiple regression model to $\mathbf{B}$ by taking as independent attributes the confidential attributes and as dependent attributes the quasi-identifier attributes. The notation in Table 5 below is the same used in the analogous tables for the “Census” dataset, except that no scenario superscript is used. The table shows the results of record linkage experiments between the “EIA” dataset and partially synthetic versions of it generated using IPSO-A, IPSO-B and IPSO-C. Only the quasi-identifiers used in each experiment are listed, which are subsets of those specified in Table 4.

Table 4. Splittings of “EIA” into datasets $\mathbf{A}$ and $\mathbf{B}$ and attributes per dataset

Data set	Shared attributes		Non-shared attributes	
	Quasi-id. Y	Conf. attr. X	Quasi-id. Y	Conf. attr. X
A	$v1, v2, v7, v8, v9$		$v1, v7$	
B	$v1, v2, v7, v8, v9$	$v3, v4, v5, v6, v10$	$v2, v8, v9$	$v3, v4, v5, v6, v10$

Quasi-identifiers in Table 5 were selected using the cross-correlation matrix between the original quasi-identifier attributes and the quasi-identifier attributes generated using methods IPSO-A, IPSO-B, IPSO-C. The rationale of our quasi-identifier choices is that at least some of the quasi-identifiers in datasets $\mathbf{A}$ and $\mathbf{B}$ and should be highly correlated. Note that this strategy in quasi-identifier selection can be followed by a real snooper, since he can compute the cross-correlation matrix between the external identified dataset and the released, partially synthetic datasets.

Table 5. Re-identification experiments using dataset "EIA" and methods IPSO-A, IPSO-B and IPSO-C. Results in number of correct re-identifications over an overall number of 4092 records.

Quasi-identifier in external **A**	Quasi-identifier in released **B**	DRL1	DRL2	PRL	CRL
$v1$ $v1, v7, v8$ $v1, v2, v7, v8, v9$	$v1_A$ $v1_A, v7_A, v8_A$ $v1_A, v2_A, v7_A, v8_A, v9_A$	10 (0.2%) 23 (0.5%) 186 (4.5%)	10 (0.2%) 24 (0.5%) 171 (4.1%)	10 (0.2%) 11 (0.2%) 189 (4.6%)	32 (0.8%) 30 (0.7%) 46 (1.1%)
$v1$	$v9_A$	N/A	N/A	N/A	9 (0.2%)
$v1, v7$	$v2_A, v8_A, v9_A$	N/A	N/A	N/A	7 (0.2%)
$v2, v8, v9$	$v1_A, v7_A$	N/A	N/A	N/A	6 (0.1%)
$v1$ $v1, v7, v8$ $v1, v2, v7, v8, v9$	$v1_B$ $v1_B, v7_B, v8_B$ $v1_B, v2_B, v7_B, v8_B, v9_B$	10 (0.2%) 23 (0.6%) 187 (4.6%)	10 (0.2%) 24 (0.5%) 171 (4.1%)	10 (0.2%) 11 (0.2%) 189 (4.6%)	26 (0.6%) 25 (0.6%) 47 (1.1%)
$v1$	$v9_B$	N/A	N/A	N/A	9 (0.2%)
$v1, v7$	$v2_B, v8_B, v9_B$	N/A	N/A	N/A	10 (0.2%)
$v2, v8, v9$	$v1_B, v7_B$	N/A	N/A	N/A	8 (0.2%)
$v1$ $v1, v7, v8$ $v1, v2, v7, v8, v9$	$v1_C$ $v1_C, v7_C, v8_C$ $v1_C, v2_C, v7_C, v8_C, v9_C$	7 (0.2%) 10 (0.2%) 42 (1.0%)	7 (0.2%) 10 (0.2%) 42 (1.0%)	7 (0.2%) 6 (0.1%) 71 (1.7%)	8 (0.2%) 9 (0.2%) 28 (0.7%)
$v1$	$v9_C$	N/A	N/A	N/A	7 (0.2%)
$v1, v7$	$v2_C, v8_C, v9_B$	N/A	N/A	N/A	6 (0.1%)
$v2, v8, v9$	$v1_C, v7_C$	N/A	N/A	N/A	5 (0.1%)

6. Conclusions and extensions

It can be seen that, among the methods tried, IPSO-C is the safest one, in that it is the one allowing less re-identifications. Apparently, this is perfect, because IPSO-C also preserves more regression statistics that IPSO-A and IPSO-B. However, at a closer look, it can be seen that the individual values generated by IPSO-C for the quasi-identifier attributes are more different from the original values than in the case of IPSO-A and IPSO-B. This can easily be seen by computing the average Euclidean distance between original records and records generated by the three IPSO methods; the largest average distance is between original and IPSO-C records. The explanation of the above is that, in order to preserve more statistics, IPSO-C resorts to "injecting" more perturbation at the record level than IPSO-A and IPSO-B.

We now examine the influence of the number of independent confidential attributes X. In Scenario S1 ("Census" dataset, Table 1), the multivariate multiple regression model uses only four confidential attributes X as independent variables. In Scenario S2, nine confidential attributes X are used. In fact, the X in Scenario S1 are a subset of the X in Scenario S2. Thus, the synthetic quasi-identifier attributes Y in Scenario S1 are generated based on less X attributes than in Scenario S2. Surprising enough, the differences between both scenarios as to the number of re-identifications are less straightforward than one would expect (see Tables 2 and 3). By focusing on identical quasi-identifiers across both scenarios S1 and S2 (that is, $(v7, v12)$ and $(v4, v7, v12, v13)$) we can see that, for IPSO-A and IPSO-B, distance-based and probabilistic record linkage re-identify more when the regression model has been fitted on few independent attributes. For those two methods, correlation-based record linkage works better when the regression model has been fitted on a greater number of independent attributes. IPSO-C displays exactly the opposite behavior: more DRL1, DRL2 and PRL re-identifications and less CRL re-identifications are obtained when there are more independent attributes.

Another important point to be analyzed is the influence of the quasi-identifier length. A longer quasi-identifier does not necessarily result in more re-identifications. Indeed, it can be seen in Table 2 than more re-identifications are obtained with $(v7, v12)$ than with longer quasi-identifiers also including $v7$ and $v12$. The reason is that, as it can be checked in the cross-correlation matrix between the original quasi-identifier and the quasi-identifier generated by IPSO-A, it turns out that $v7$ and $v12$ are good representatives of the other quasi-identifier attributes: $v7$ is highly correlated with $v4_A$ (0.9778), $v6_A$ (0.9807) and $v7_A$ (0.9812); $v12$ is highly correlated with $v3_A$ (0.9509), $v11_A$ (0.9788), $v12_A$ (0.9793) and $v13_A$ (0.9792). Thus $v7$ and $v12$ complement each other in sort of "covering" nearly all quasi-identifier attributes generated by IPSO-A (only $v1_A$ and $v9_A$ stay "uncovered"). This is no surprise, given the central position that $v7$ and $v12$ hold in the dendrogram of the "Census" dataset. Thus, the lessons learned are:

1. If a snooper can find via cross-correlation matrix a few quasi-identifier attributes that are highly correlated to the all partially synthetic quasi-identifier attributes, she should use only those few attributes for re-identification; using longer quasi-identifiers will only add noise and reduce the number of successful re-identifications.
2. *The data protector should generate partially synthetic microdata in such a way that no such small set of original quasi-identifier attributes are highly correlated to all synthetic quasi-identifier attributes.* In doing so, the data protector will force potential snoopers to use longer quasi-identifiers, which makes life more difficult for them (more external identified information required).

We can also compare the performance of the record linkage methods used. It seems that the overall performance of DRL1, DRL2 and PRL in terms of the number of re-identifications is similar. Nonetheless, while both distance-based methods DRL1 and DRL2 stay similar for any quasi-identifier length, probabilistic record linkage PRL seems to clearly outperform DRL1 and DRL2 for longer quasi-identifiers. Correlation-based record linkage (CRL) behaves clearly worse than PRL, DRL1 and DRL2 and should not be used in the shared-attributes paradigm. However, it is the only method among those considered that is still applicable without shared attributes.

Finally, a few words on the influence of the dataset size. We used two datasets with differents sizes ("Census", 1080 records; "EIA", 4092 records) to attempt an assessment of the influence of the dataset size on the number of re-identifications. By comparing Table 5 with Tables 2 and 3, we see that the percentage of re-identifications is lower for the larger "EIA" dataset, as one would expect. However, the *absolute number of re-identifications* is not lower in "EIA" when a sufficiently long quasi-identifier is used. In fact for quasi-identifier $(v1, v2, v7, v8, v9)$ and shared attributes, we obtain between 170 and 190 re-identifications for IPSO-A and IPSO-B, and between 40 and 70 for IPSO-C, which is more than the number of re-identifications we obtained when using the "Census" dataset.

Only numerical attributes have been considered in this work. To deal with categorical quasi-identifier attributes one would need:

- To use methods which, unlike IPSO-A, IPSO-B and IPSO-C, are appropriate for generation of categorical synthetic microdata.
- To use distance-based record linkage with ordinal or nominal distances rather than the Euclidean distance.
- To use Spearman's rank correlations instead of Pearson's correlations to adapt correlation-based record linkage to ordinal attributes (for nominal attributes there is no obvious adaptation).

Probabilistic record linkage is the only record linkage method among those used that can directly work on categorical data without any adaptation.

Acknowledgments

This work was partly funded by Cornell University under contracts no. 47632-10042 and 47632-10043, by the Catalan government under project 2002 SGR 00170 and by the Spanish government under project SEG2004-04352-C04-01/03 "PROPRIETAS".

References

R. Brand, et al. (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.

J. Burridge (2003). 'Information preserving statistical obfuscation'. *Statistics and Computing* **13** :321–327.

R. Dandekar, et al. (2002). 'LHS-based hybrid microdata vs rank swapping and microaggregation for numeric microdata protection'. In J. Domingo-Ferrer (ed.), *Inference Control in Statistical Databases*, vol. 2316 of *LNCS*, pp. 153–162, Berlin Heidelberg. Springer.

J. Domingo-Ferrer, et al. (2001). 'Comparing SDC methods for microdata on the basis of information loss and disclosure risk'. In *Pre-proceedings of ETK-NTTS'2001 (vol. 2)*, pp. 807–826, Luxemburg. Eurostat.

J. Domingo-Ferrer, et al. (2005). 'A polynomial-time approximation to optimal multivariate microaggregation'. *submitted manuscript* .

J. Domingo-Ferrer & V. Torra (2001). 'A quantitative comparison of disclosure control methods for microdata'. In P. Doyle, J. I. Lane, J. J. M. Theeuwes, & L. Zayatz (eds.), *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, pp. 111–134, Amsterdam. North-Holland. http://vneumann.etse.urv.es/publications/bcpi.

J. Domingo-Ferrer & V. Torra (2002). 'Validating distance-based record linkage with probabilistic record linkage'. In F. T. M. T. Escrig & E. Golobardes (eds.), *Topics in Artificial Intelligence*, vol. 2504 of *LNCS*, pp. 207–215, Berlin Heidelberg. Springer.

J. Domingo-Ferrer & V. Torra (2005). 'Ordinal, continuous and heterogenerous k -anonymity through microaggregation'. *Data Mining and Knowledge Discovery* **11**(2). (to appear).

I. P. Fellegi & A. B. Sunter (1969). 'A theory for record linkage'. *Journal of the American Statistical Association* **64**(328):1183–1210.

W. A. Fuller (1993). 'Masking procedures for microdata disclosure limitation'. *Journal of Official Statistics* **9**:383–406.

M. A. Jaro (1989). 'Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida'. *Journal of the American Statistical Association* **84**(406):414–420.

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

P. Tendick (1992). 'Assessing the effectiveness of the noise addition method of preserving confidentiality in the multivariate normal case'. *Journal of Statistical Planning and Inference* **13**:273–282.

V. Torra & J. Domingo-Ferrer (2003). 'Record linkage methods for multidatabase data mining'. In V. Torra (ed.), *Information Fusion in Data Mining*, pp. 101–132, Germany. Springer.

W. E. Winkler (1995). 'Advanced methods for record linkage'. In *Proc. of the American Statistical Association Section on Survey Research Methods*, pp. 467–472. ASA.

W. E. Yancey, et al. (2002). 'Disclosure risk assessment in perturbative microdata protection'. In J. Domingo-Ferrer (ed.), *Inference Control in Statistical Databases*, vol. 2316 of *LNCS*, pp. 135–152, Berlin Heidelberg. Springer.

Access to business microdata in the UK: dealing with the irreducible risks

Felix Ritchie

Social and Economic Micro Analysis and Reporting Division, Office for National Statistics, Government Buildings, Cardiff Road, Newport, South Wales, NP10 8XG.

A longer version of this paper is available from the author.

Abstract. The UK Office for National Statistics provides a thin-client remote laboratory service for secure research on confidential microdata. This technological solution is allied to tight procedural environment and compulsory training of researchers to achieve an effective mix of practicality and security. The result has been a ten-fold increase in the use of the ONS business data by external researchers, and a significant increase in ONS' in-house capabilities and projects.

This solution raises several potential problems: the irreducible person risk from providing access to identifiable data; a need for "intelligent" disclosure control policies; the management of off-site access to data; and the possibility of "unwanted" recreation of official data. This paper discusses how these have been addressed.

Keywords. Remote access, thin clients, disclosure control, confidentiality, identifiable data, research use of microdata

1. Introduction

Since January 2004 the UK Office for National Statistics (ONS) has been providing remote and onsite access to business microdata for research purposes in a controlled environment. This is achieved through thin-client technology and a tight procedural framework, within which researchers have complete access to identifiable (if not identified) data. This has allowed data to be used by researchers in academia, government and the private sector to produce analyses ranging from the impact of innovation on productivity to the calculation of labour cost adjustments for the health service. Researchers are also allowed to link their own data with ONS data in the lab, and this has been taken up with enthusiasm by other government departments, often in collaboration with academics. This has helped to keep the UK at the forefront of policy impact analysis and programme evaluation in Europe.

This solution raises several issues.

First, although the technical solution provides the highest protection from a technical attack, giving access to identifiable microdata creates an irreducible "person risk".

Second, the lack of restrictions on analysis means that disclosure control mechanisms need to be sufficiently flexible to cover an unknown variety of outcomes. As a result, automatic disclosure control methods are not feasible, and training in statistical disclosure becomes necessary for both researchers and lab managers.

Third, a remote access system potentially allows access to data from any location. Thus access is governed by perceptions of acceptable risk rather than technical feasibility; and these perceptions are more subject to criticism by outside bodies.

Fourth, the provision of complete datasets to researchers means that there is a risk of the ONS's own figures coming under attack from researchers using the same data.

In all these cases, there remains an element of risk which cannot be managed away. This paper describes how ONS has addressed these issues within a common corporate framework which is now being applied to non-business data.

2. Providing secure access to confidential data for research

2.1. Background

In 2003 ONS created the Business Data Linking branch (BDL) to address the issue of providing access to its business microdata for research purposes. This posed several significant problems, including the legality of access and the fitness of survey data for the purpose. For a detailed description of this, see Ritchie (2004).

BDL developed a solution based upon a four-part model of security: safe projects, safe people, safe settings, and safe outputs. This composite strategy is designed so that the elements reinforce one another. This model is now being adopted for other parts of ONS wishing to provide access to microdata for research[1].

2.2 The Virtual Microdata Laboratory (VML)

The concepts mentioned above are familiar to many national statistical institutes (NSIs). One unusual feature is the technological solution implemented to secure the system electronically. This system has been in place in Denmark for some years, and is now being considered in test implementations in a few other countries.

The VML is a thin-client system; that is, researchers log on to a computer in a remote location which processes all requests centrally and returns information about the results. Hence, no data travels over the network, save in the form of statistical results. In contrast, a “fat client” such as a PC downloads data over the network and processes it locally.

There are advantages and disadvantages of both systems in terms of hardware, which are not discussed here. However, for security a thin-client system has inherent advantages. As all processing is carried out on the central server; the security of the system is determined largely by the security of that server, rather than clients. The only software required at the user end is the thin client interface, which may be as simple as a web browser; all other software can be managed on the central server.

2.3. Results: use of business microdata at ONS

Since January 2004, all BDL researchers have been using the lab. Because of the security and efficiency of the lab, BDL has been able to expand output enormously. In eighteen months, research output has expanded from roughly 10 projects and 15 accredited researchers to over 90 projects and over 150 researchers.

Whilst a large part of this is due to an increase in general academic research, there has been a significant increase in the use of the data by other government departments, either directly or indirectly through academics. There is a significant movement towards more evidence-based policymaking in the UK, and the BDL has been able to support a large amount of this. A large part of this government work is on programme evaluation (or policy impact analysis, as it is sometimes called). Recent projects have included the effectiveness of small business support programmes, export subsidies, tax changes, tourism promotion, and the National Minimum Wage, all supported by the relevant government departments.

Of more direct relevance to the ONS is the increasing use of the lab by internal staff. Recent projects have included estimates of investment in intangibles, the impact of R&D capitalisation, micro-macro integration, and reconstruction of the main business register on a historical basis. These are major long-term projects with significant implications for ONS, only now made feasible by the combination of secure easy access to a wide variety of microdata sets and advanced econometric skills.

[1] To clarify roles, ONS set up Business Data Linking (BDL); and BDL set up the Virtual Microdata Lab (VML) as the technical solution. The VML is now used by other areas in ONS as well as BDL, but the resource continues to be managed and developed by BDL. Hence references to ONS are to the overall policy; BDL, to how the business data setion operates; and VML to the underlying technology.

Finally, the VML has also provided a secure facility for other types of microdata: it currently houses Census and Labour Force Survey data at levels of detail not available on externally distributed datasets; and is being evaluated as a potential home for personal, medical and mortality data[2].

3. Issues

The ONS system of microdata access provides a flexible and secure system for the analysis of microdata. However, the consequences of this system are a series of risks which are relatively new to ONS. Not all the risks involved with providing such open access to microdata to can be managed away, and so ONS has had to re-evaluate and explicitly define a new series of risks.

3.1. Access to microdata and the irreducible person risk

The VML provides an extremely secure solution. The electronic removal of data from the VML by researchers is not practically possible. Independent verification of the technology and procedures described the VML as meeting or exceeding best practice in almost every area; in some areas, such as disclosure control, BDL procedures far outstripped alternative methods across the UK government.

Nevertheless, providing access to identifiable microdata does give rise to an irreducible risk; that is, that a researcher could identify a company, and then remove information about the company through non-electronic means. This risk cannot be reduced for company data. First, this data cannot be anonymised effectively. Second, even if all writing materials were banned from the VML environment, it is not possible to stop a researcher remembering items of information.

As technological risk reduction is not possible beyond this point, ONS therefore concentrates on the “safe people” part of the security framework. In particular, researchers must come from “trusted” organisations (ie those where ONS is reasonably confident there is no conflict of interest), must themselves have a credible research background, and are made aware, through the BDL training programme, of the consequences of abusing the trust.

For this last point, the credibility of any sanction is important. Prior to 2002, researchers were brought in on “£1 contracts” of the type common in many countries. These were stopped, partly because they did not meet the spirit of the law, but also because they gave no grounds for credible and meaningful sanctions. However, the new contracts tie researchers to their institutions, and enforce a form of collective responsibility. This means that, in the case of a breach of confidentiality, ONS will approach the researcher’s institution who will be responsible for disciplinary action in all but the most serious of cases. This gives a much wider range of sanctions, and is also a credible response by ONS. In addition, BDL is in discussions with key academic funding bodies about tying funding to the “trustworthiness” of the institution.

This has not resolved all problems. Some researchers reject the implication that they might not be trustworthy, whilst there is still some suspicion within government circles that academics might not show an appropriate awareness of the confidentiality of data. In addition, BDL has not yet resolved the position of private sector consultants; there, the possibility of conflict of interest is felt to be too high in general to allow direct access to commercial data. Finally, there have been no agreements with international bodies, because of the difficulty of finding an effective legal framework. Nevertheless, the system as it stands seems to be generally accepted by most parties.

[2] Note that the VML is a last-resort solution for situations where the data cannot be released. ONS uses a variety of distribution channels (such as the UK Data Archive), and the default is to anonymise and release data if possible; for example, social survey data is typically anonymised and distributed under licence, which has meant a much longer and wider tradition of analysis of social data

3.2. Disclosure control in a research environment

Statistical disclosure control (SDC) in a research environment differs fundamentally from methods used to produce aggregate tables or anonymised datasets. In both those cases, there is a finite set of outputs and a well-defined structure for the data.

In contrast, the purpose of a research environment is to combine data in innovative ways and produce a range of outputs which would not normally be generated by the statistical body. This means that

- Automatic disclosure control of most outputs is not possible
- SDC rules, whilst often extremely detailed, are rarely comprehensive and are almost always open to challenge for particular examples

For a more detailed discussion of the issue, see Ritchie (2005).

As manual checking of all outputs is required, and as it is difficult to define in advance acceptable outputs, BDL makes training in SDC compulsory for all researchers and staff involved in the lab. The training method is based on an understanding of principles, backed up by examples to illustrate particular issues; there is very little discussion of rules per se. For example, although the BDL threshold of a minimum 10 units for each cell is given, the main discussion of this rule illustrates cases where the rule can be adjusted or ignored, where it will be tightened, and what information researchers need to provide to BDL to allow them to make a meaningful judgement.

This requires some commitment on the part of both researchers and lab managers, as researchers are required to physically attend the training sessions. However, the response among researchers has generally been positive, with almost all taking the view that the time spent on the course has been productive. Certainly the experience of BDL before and after the introduction of the training course, and of other areas in ONS, is that this makes a significant difference to the time taken to clear outputs.

Again, there are still unresolved questions. One is that, under the BDL model, there are few absolute rules as to what is allowed, and researchers have requested more clarity over acceptable outputs. In practice, once researchers become used to the data, the major problem for BDL is the quantity of output produced.

It has been argued that, by giving researchers a detailed insight into how disclosure control is carried out, the risk of them subverting SDC procedures is raised. At BDL, the view is taken that, should a researcher really want to remove output surreptitiously, no practical method of checking output is going to prevent this. On the other hand, involving researchers in the process of checking for outputs encourages them to take a pro-active approach in avoiding problematic outputs. The results can be seen in that several of the examples BDL uses in its training have arisen from questions posed by researchers.

3.3. The possibilities of remote access: what is a “safe place”?

The VML is potentially accessible from any location connected to the internet or a phone line. However, at the moment, technically access is limited to ONS sites. An investigation is under way to put in place equipment to allow secure access across other government sites. In the longer term, it is possible to envisage secure sites being set up at universities, along the lines of the US Census Research Data Centres.

Whilst there is a technical element to ensuring the security of access, the major concern here of providing access on non-ONS sites is procedural. Off-site access to a research lab involves devolving responsibility for the physical security of the lab site to a third party; and because of the irreducible person risk noted in section 3.1, this means that some of the risk is also being devolved.

There are a variety of models for this. In the US, Census Bureau employees are stationed at each research centre, which is costly but ensures the Census Bureau keeps the risk in-house. At the other end of the scale, Statistics Denmark relies entirely on its "safe people" policy. Access is available from the desktop of any researcher, subject to both the researcher being approved and technological methods being in place to ensure that the researcher can only log on from an approved institution. This is a cheap, flexible and very secure option, save for the problem of ensuring that only approved users are physically in front of the terminal.

In preparation for extended access, the ONS site has begun providing lab access at its Southport office. This site was chosen as a test case because there is no local expertise in any of the datasets available in the lab. The facility is managed locally purely for access to researchers in a measure halfway between the US/Danish models. The local managers are responsible for escorting researchers on to the premises, and observing researchers to make sure there are no attempts to write down confidential data. A set of protocols for "safe places" and "safe kit" (that is, a standard working environment and technology for a remote lab) has been agreed.

This solution is not universally supported. It has been argued that devolving the responsibility of supervising researchers increases ONS' risk unacceptably. Researchers meanwhile tend to view restrictions to certain physical sites as an unwarranted limitation on research. This solution is more costly than the Danish solution; it is also potentially less secure than the US system, as non-specialist local managers may not understand whether confidential data is being removed or not. To address this point, BDL authorises the local manager to "remove first, question later"; that is, in the event of any suspicious activity, the local managers will err on the side of confidentiality. This semi-attended lab facility has been available since the summer of 2005, and so is still under review but initial results are encouraging.

3.4. Are ONS official statistics liable to attack?

One concern expressed with research access to microdata is the risk to the reputation of official statistics. If researchers are using the same data as that used to generate aggregate statistics, what happens if

- Researchers produce different aggregate statistics?
- Researchers discover significant errors in aggregate statistics or the underlying data?

This latter point has been accepted as a risk with potentially beneficial consequences. Although there is a risk of embarrassment of ONS, this is felt to be outweighed by the quality-control aspect of letting a large number of researchers stretch and twist data. BDL is currently formalising feedback arrangements so that queries and comments from researchers can be relayed to the data providers effectively.

The former problem is more subtle. "National Statistics" (ONS official non-experimental outputs) are generated by weighting survey and administrative data to reflect population characteristics. This is a complex process requiring large teams of people, and goes well beyond the simple weighting which researchers would typically use to produce descriptions of the data. Researchers' interest is in marginal analysis and sample description, rather than population totals. Hence, when researchers do produce figures which are comparable to "official" figures, they are unlikely to agree, and the reason for the difference may be quite technical. It is possible that ONS' reputation may suffer from a misinformed reading of outputs produced by different methods, particularly as the outputs of external researchers are not subject to the same stringent quality checks as ONS official statistics.

BDL took the decision not to restrict outputs, partly because it was difficult to police, but mainly because it was thought extremely unlikely that research papers would be compared to official statistics.

The readership for these publications is quite different, and most analysts reading technical papers would be familiar with the differences between aggregate and marginal analysis. The solution BDL has adopted is to explicitly exclude (by contract) research outputs from being described as National Statistics; a standard rubric is given to researchers in data releases. Although not without criticisms, the current position of BDL is to continue with this policy as there is no evidence yet that this is inappropriate or that there is a practical alternative.

4. Conclusion

ONS has spent some time and thought building a research facility for confidential microdata which is believed to offer an optimal combination of very high security, simplicity in operation and management, and flexibility in use. The lab and the procedures developed have proved extremely popular, and use of the data has rocketed with significant benefits to ONS, other parts of government, and academia.

However, even when the problems of providing access to data have been solved as far as possible, a number of significant risks remain. These can be seen to be irreducible risks; that is, an organisation has to decide whether the remaining risk is acceptable or not – there is no practical possibility of reducing the risk further without significantly affecting the operation of the solution. For example, it is necessary to trust researchers with identifiable data; otherwise, it is not possible to analyse business microdata except through costly and inefficient proxies.

In each of the four security aspects on which BDL and ONS base their access models, there is at least one irreducible risk: for safe people, the trust risk; for safe outputs, the disclosure risk; for safe settings, the location risk; and for safe projects, the reproduction problem. These are all contentious, and all are under continuous review. However, at least ONS does have a clear perspective on exactly how much risk is inherent in its microdata operation.

References

Ritchie, FJ (2004). Business Data Linking: Recent UK Experience, *Austrian Journal of Statistics* v33:1/2

Ritchie, FJ (2005) *Statistical Disclosure Control in a Research Environment*, mimeo, Office for National Statistics, London

Topic V

Confidentiality aspects of tabular data, frequency tables, etc.

Safety Rules in Statistical Disclosure Control for Tabular Data

Giovanni M. Merola[1]

Winton Capital Management, 1-5 St. Mary Abott's Place, London W8 6LS, UK g.merola@wintoncapital.com

Abstract. We extend the safety rules used for the Statistical Disclosure Control of magnitude tables to include an intruder who models the ignorance about an unknown confidential quantity with a Uniform distribution. By applying this extension to the generalised p-rule we obtain the safety rules useful also in the presence of groups of respondents. The corresponding disclosure rules for different prior knowledge of the intruder. The different safety rules are then compared to each other by considering some real Structural Business Statistics.

1. Introduction

Statistical Disclosure Control (SDC) consists of a variety of methods used to protect the privacy of respondents when confidential data are published. SDC is mainly applied by National Statistical Institutes (NSIs), but it is also applied by other entities that disseminate confidential data. In order to enforce the confidentiality agreements safely, values that can be estimated "closely" or that can be attributed with "high probability" are considered disclosed, so the values to be published are assessed in terms of *risk of disclosure*. Data-sets are cleared by *safety rule* that sets a level of acceptable risk for for each datum. Once disclosive values have been identified, the whole set of data is then protected with different techniques. Details on SDC theory and methods can be found, for example, in Willenborg and de Waal (2000). Protection of disclosive data unavoidably leads to the suppression or the distortion of some values, so the adoption of an appropriate measure of risk can avoid unnecessary damage to the data while protecting against disclosure (see, for example, Fienberg, 2000; Trottini, 2001; Ducan et al., 2001).

In this paper we consider the assessment of the risk of disclosure for non-negative values released as sums, which are often released in magnitude tables. Magnitude tables are published in large number and they can disclose contributions more easily than other tables. Therefore, SDC for these tables has received great attention in the literature and a computer package mainly devoted to the protection of this type of tables, τ-Argus (Hundepool, 2004), has been developed[2].

In SDC for magnitude tables it is assumed that an intruder with some *prior knowledge* is interested in learning some of the individual responses, in this context called *contributions*, that form a published sum. Cox (1981) defined four different measures of risk for magnitude tables and their properties and mutual relationships are considered, for example, in Willenborg and de Waal (2000); Federal Committee on Statistical Methodology (1994); Cox (2001); Loeve (2001); Merola (2003b). For the SDC of some data-sets it is necessary to consider the existence of groups, that is respondents that are connected and can communicate, must be taken into account when measuring the risk of disclosure. Natural examples of groups of respondents are households and industrial holdings. In this case an intruder may know the contributions of a group and be interested in the total of another group. Ways of including groups of respondents in Cox's rules are proposed in the papers cited above. Merola (2003a,b) extends one of these rules, the p -rule, to groups and shows that all the existing rules can be derived from this generalisation.

In the p-rule it is assumed that an intruder with the knowledge of one of the contributions estimates the largest contribution with its maximum possible value. In this paper, instead, we assume that the same intruder uses the prior knowledge to determine an interval of possible values for the largest

[1] This paper was mostly written while at ISTAT and was partially supported by the European Union project IST-2000-25069 "CASC". We would like to thank Dr. Luisa Franconi of ISTAT for making the data available and for her useful comments. The views expressed are those of the author only.

[2] τ-Argus was created within the Computational Aspects of Statistical Confidentiality (CASC) project. It can be freely downloaded from the CASC Web Page at *http://neon.vb.cbs.nl/casc/*

contribution and that estimates it by minimizing the expected error. We hypothesise that the ignorance about the unknown quantities is modelled with a Uniform distribution and derive the safety rules for different specification of the prior knowledge. Since the rules so obtained consider safe also contributions with large vales, which are more identifiable than others, we extend the requirements to include large dominating contributions. The rules so obtained are stricter versions of the generalised p-rule.

In the following section we relate the existing safety rules to the identification of the respondents. In Section 3 we recall the generalised p -rule. In Section 3 we derive the new rules and in the following one we give a numerical comparison of the different rules. Finally, in Section 6 we give some final remarks.

2. Disclosure rules and identification of respondents

The identification of a respondent constitutes disclosure by itself when one or more of the categories defining a cell are confidential. For example, if one of the categories is "being infected with HIV". Respondents in a cell can be identified because they are known to have the characteristics defining the cell. The probability of this type of identification depends on the number of respondents in a cell. The safety rule that tackles this risk is the *threshold rule*, by which all the respondents belonging to a cells with less respondents than a given threshold are considered identifiable and the cell is considered disclosive.

Respondents can also be identified because they are known to carry a particularly large contribution. For example, it may be known which respondents have the two largest contributions in a cell. In some cases the shear presence of large contributions may lead to identification; for example, a very high total income for a group of people may give away the presence of a person with a much larger income than the others. Such large contributions are said to *dominate* the others and the rule that tackles this type of identification is called *Dominance rule*. By this rule a cell is considered disclosive if the sum of few of the largest contributions exceeds a certain percentage of the total, regardless of how closely the identifiable contributions can be estimated.

The rule that considers the precision of the estimation of a contribution is the p-rule. In this rule it is assumed that the largest respondent of a cell is identifiable and that the intruder knows the second largest contribution and estimates the largest one by its maximum possible values, that is by subtracting the known contribution from the total. The cells in which this estimate gives a relative error smaller than a given level, typically denoted by p -where from the name,- are considered disclosive.

We would like to stress that the p-rule can be applied only for the protection of the largest contribution as this estimating procedure cannot be extended to other contributions, as sometimes suggested. In fact, if T is the total and $z_1 \geq z_2 \geq z_3$ are the three largest contributions, then z_2 will not be estimated by $T - z_3$, simply because $z_2 \leq T/2$ and $T - z_3 > T/2$. Hence, the p rule properly protects the estimation of the largest contribution. The last of the rules currently used, the so called pq-rule, can be considered a stricter version of the p-rule (*e.g.* Merola, 2003a).

3. The *M*-rule

In Merola (2003a) we generalise the p -rule to the existence of groups, considering the subtotals of each group as a confidential datum. It could be the case of medical expenses grouped for household, for example. Let T be the published cell total and $z_1 \geq z_2 \geq ... \geq z_n$ be the n ordered contribution, so that $T = \sum_{i=1}^{n} z_i$. We assume that the intruder wants to estimate the sub-total of the m largest contributions, denoted with $t_m = \sum_{i=1}^{m} z_i$, knowing the total of the subsequent l largest ones, denoted by

$R_{m,l} = \sum_{i=m+1}^{m+l} z_i$. The largest estimate of t_m is given by

$$\hat{t}_m = T - R_{m,l} = t_m + r_{m+l}, \tag{1}$$

where the remainder $r_m = \sum_{i=m+1}^{n} z_i$ (with $r_0 = 0$) is the estimation error. Like in the p-rule we require that the relative estimation error, denoted by $RE(t_m;l)$, is larger than the level p, with $0 \le p < 1$, that is:

$$RE(t_m;l) = \frac{|t_m - \hat{t}_m|}{t_m} \tag{2}$$

The generalised p-rule is obtained substituting the estimate (1 into requirement (2:

$$M_p(m;l): \frac{t_m}{T - R_{l,m}} \le \frac{1}{1+p}, \tag{3}$$

where the subscript p is used to denote the protection level (but will be omitted when not needed) and l denotes the number of known contributions and will be omitted when equal to zero. The symbol M denotes that the estimate is the maximum possible value. Henceforth we will refer to this rule as the M-rule.

As shown in Merola (2003a), all the existing rules are special cases of the M-rule. The *threshold* rule for m respondents protects against exact disclosure, that is $p = 0$, when the intruder knows l contributions and wants to estimate $m - l$ contributions. The requirement for the $M_0(m-l;l)$ rule is $RE(l;m-l) = (T - R_{l,m-l} - t_m)/z_1 = r_m/t_{m-l} > 0$. It is satisfied if the respondents are more than m or if the reminder r_m is greater than zero. So, this formulation, sensibly, extends the Threshold rule to cells with all zero contributions after the m-th.

The Dominance rule can be obtained by assuming that the intruder does not know any of the contributions, hence by setting $l = 0$. For this case the $M_p(m)$ rule is $t_m/T \le 1/(1+p)$. As already noted (*e.g.* Cox, 2001), in this way it is possible to express the requirement of the Dominance rule in terms of the minimum relative error of estimation.

The p-rule can be obtained straightforwardly by setting $m = l = 1$. The $M_p(1;1)$ rule requires that $z_1/(T - z_2) \le 1/(1+p)$. The pq-rule corresponds to the p-rule with protection level equal to p/q, that is $M_{p/q}(1;1)$. One desirable property of safety rules is sub-additivity, introduced by Cox (1981). By transforming rules in *linear sensitivity measure* he shows that a rule is sub-additive if and only if the corresponding sensitivity measure has nonincreasing coefficients. It can be easily shown that the generalised p-rule is sub-additive for all values of m and l (Merola, 2003a) .

The maximizing estimation procedure assumed in the M-rule, in some cases, may not be realistic. In the next section we derive safety rules under the assumption of a different estimating procedure allowing different prior knowledge to the intruder.

4. The MU-rules

Let us assume that the intruder is interested in estimating t_m and uses the prior knowledge to restrict its possible value within bounds, say $t_m^- \le t_m \le t_m^+$. If $F(t_m)$ is the distribution of t_m over this interval, the estimate can be obtained by minimizing the mean squared error (MSE), that is $\int_{t_m^-}^{t_m^+} (t_m - \hat{t}_m)^2 \, dF(t_m)$. In this paper we assume that the intruder does not know the distribution and that models this ignorance by taking it to be a Uniform over $[t_m^-, t_m^+]$ (for a discussion on modeling ignorance over a finite interval see, for example, Bernardo and Smith, 1994). Then, the estimate that minimises the MSE is

$$\hat{t}_m = \frac{t_m^- + t_m^+}{2}, \tag{4}$$

for a well known property of the mean. Of course, other distributions may lead to different estimates but this estimate would be equally optimal for other symmetric distributions, such as the truncated Normal, for example.

Given estimate (4), the safety rules are derived by requiring that RE is not less than the safety level, p, that is $RE = |\hat{t}_m - t_m|/t_m \geq p$. In Merola (2003b) we show that these conditions are not satisfied by values of t_m within an interval, say $lb \leq t_m \leq ub$. This means that large values of t_m would be considered safe. Since it is plausible to assume that an intruder may know that some values dominate, - that is may have assumptions on the distribution of t_m - s/he could take a maximising estimate. Therefore, we extend the rule to include also this case by dropping the requirement that $t_m < ub$. The resulting rule is in the form $t_m < ub$. We name the resulting rules *MU-rules* as they protect against absolute relative error larger than p for any estimate of t_m that lays between its maximum and the estimate (4).

We now give the MU-rules for different possible prior knowledge of the intruder, without details of the derivation, which can be found in Merola (2003b). We always assume that the intruder has the *basic knowledge* of the cell total, T, and that the number of contributions, n, is larger than m. On top of this we consider four other cases given by the combinations of whether the number of contributions is known and whether one or more contributions are known. Since there is not a generalised solution for all the cases, we will consider the different scenarios separately. The safety rules will be generically denoted by $MU(m;\cdots)$, where "$\cdots$" are parameters specifying the extra prior knowledge, if present.

MU (m): the MU-dominance

This is the same prior knowledge as for the Dominance rule. With this knowledge t_m can only be bounded by $0 \leq t_m \leq T$. Substituting these values in (4) gives $\hat{t}_m = T/2$. The resulting $MU(m)$ rule is not satisfied when

$$\frac{t_m}{T} \geq \frac{1}{2(1+p)}. \qquad (5)$$

Hence, this is a stricter version of the Dominance with half minimum level for the ratio t_m/T.

MU (m;n): the MU-dominance when *n* is known

This is the prior knowledge of the Dominance with the addition of the number of contributions. As it can be easily verified $nT/m \leq t_m \leq T$, so an intruder knowing T, n and m can bound t_m by $m\overline{T} \leq t_m \leq T$, where $\overline{T} = \sum_{j=1}^{n} z_j/n$ is the average contribution. Substituting these values in (4) gives $\hat{t}_m = T + m\overline{T}/2 = ((n+m)/2n)T.$ The resulting $MU(m;n)$ rule is not satisfied when

$$\frac{t_m}{T} \geq \left(\frac{n+m}{2n}\right)\frac{1}{(1+p)}. \qquad (6)$$

This rule is slightly less strict than the MU (m) but it is stricter than the Dominance. In this rule, appropriately, the requirement on the concentration of the cell changes with the number of contributions.

MU (m;l): the MU-*p*-rule

When $l > 0$ contributions are known, z_{m+1} is the largest contribution known to the intruder. Hence, the value of t_m can be bounded by $mz_{m+1} \le t_m \le T - R_{l,m}$. Substituting these values in Equation (4) gives $\hat{t}_m = (T - R_{l,m} + mz_{m+1})/2$. So the the $MU(m;l)$ rule is not satisfied when

$$\frac{t_m}{T - R_{l,m} + mz_{m+1}} \ge \frac{1}{2(1+p)}.$$

This condition is stricter than the corresponding ones for the M-rule. The equivalent of the simple p-rule, $l = m = 1$, the rule reduces to the $MU(1)$ rule, the MU analogous of the Dominance $MU(1)$, because $R_{l,m} = mz_{m+1} = z_2$.

MU (m;l;n): the MU-*p*-rule when *n* is known

In this scenario the intruder has maximum knowledge. However when $mz_{m+1} \ge T - R_{l,m} - (n-m-l)z_{m+l}$, the rule reduces to the $MU(m;l)$ because the additional knowledge of n does not improve the bounds on t_m. In the other case, as t_m is estimated by $\hat{t}_m = T - R_{l,m} - ((n-m-l)z_{m+l})/2$ and $MU(m;l,n)$ rule is not satisfied when

$$\frac{t_m}{T - R_{l,m} - (\frac{n-m-l}{2})z_{m+l}} \ge \frac{1}{1+p}.$$

This bound is active for $m = l = 1$.
The resulting MU-rules are shown in Table (1) together with the corresponding *linear sensitivity measures* (LSM), which are a way of representing the safety rules as a linear combinations of the single contributions.

Table 1. Safety bounds and corresponding sensitivity measures for different rules. The asterisk denotes that the bound is active only if conditions are satisfied.

rule	bound	Sensitivity measure
[Thresh] $M_0(l;m-l)$	$\frac{t_m}{T} < 1$	$-r_m$
[Dom] $M(m)$	$\frac{t_m}{T} < \frac{1}{1+p}$	$pt_m - r_m$
[Gen. p] $M(m;l)$	$\frac{t_m}{T-R_{l,m}} < \frac{1}{1+p}$	$pt_m - r_{m+l}$
$MU(m)$	$\frac{t_m}{T} < \frac{1}{2(1+p)}$	$(1+2p)t_m - r_m$
$MU(m;n)$	$\left(\frac{n}{n+m}\right)\frac{t_m}{T} < \frac{1}{2(1+p)}$	$(n(1+2p)-m)t_m - (n+m)r_m$
$MU(m;l)$	$\frac{t_m}{T-R_{l,m}+mz_{m+1}} < \frac{1}{2(1+p)}$	$(1+2p)t_m - mz_{m+1} - r_{m+l}$
$MU(m;l,n)^*$	$\frac{t_m}{T-R_{l,m}-(\frac{n-m-l}{2})z_{m+l}} < \frac{1}{1+p}$	$pt_m + \frac{(n-m-l)z_{m+l}}{2} - r_{m+l}$

From the sensitivity measures it can be seen that the *MU (m)* and *MU (m;n)* are subadditive, while the *MU (m;l)* and *MU (m;l;n)* rules, that assume some contributions are known, are not. Thus, these last two rules might not be appropriate for the protection of whole tables. However, the cases in which merging two cells safe with respect to these rules results in a disclosive one, can be considered rare. One of the reasons is that it seems unlikely that the contributions known to the intruder will maintain the same rank in the merged cell. This is to say that, in general, the knowledge on specific cells may not be the same as the one on two merged cells and, therefore, in some cases the use of non subadditive rules can be justified.

5. Numerical Comparison of Different Safety Rules

As an example, consider a cell with n = 12 contributions (970, 376, 274, 253, 203, 169, 161,121, 86, 62, 21, 10), so that T = 2706 and z_1/T = 0.36. The estimates considered in the Dominance and the p-rule give relative error of estimation for z_1 equal to 1.8 and 1.4 respectively. However, RE for the *MU* (1) rules and MU (1;1) rules is equal to 0.4 while it is 0.5 for the *MU* (1;0,*n*) rule.

We compared the different rules on the turnover classified by geographical region and NACE with two and three digits from the Italian Structural Business Statistics surveys of enterprises with 20 or more employees for the years 1994 and 1997 (ISTAT, 1997, 2001). First we considered the protection of a single contribution from an intruder with the knowledge of at most one contribution, hence ignoring enterprise groups. The average REs obtained for the different rules are shown in Figure (1); clearly the MU-estimates yield a much lower average RE in all cases.

Figure 1. Average relative absolute error committed by one intruder estimating z_1 for different rules.

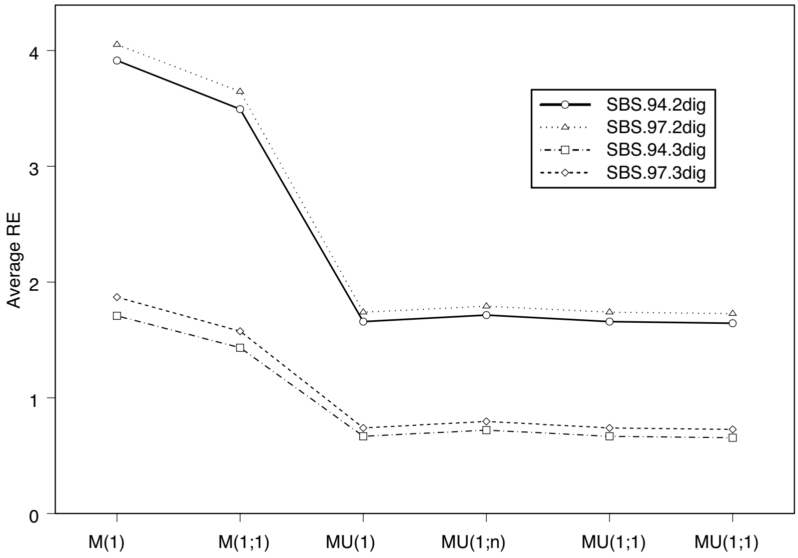
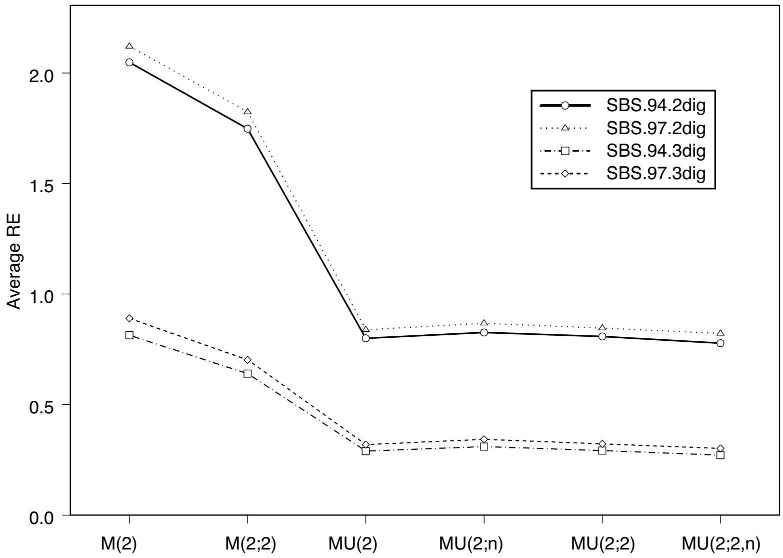


Table (2) shows the percentage of nonempty cells disclosive at a safety level p = 0.5 for the different safety rules. We included also the Dominance *M* (2) because this is sometimes considered alternative to the p-rule - but it is stricter. The table clearly shows that the number of disclosive cells increases drastically when using the MU-rules. The difference among these is small, though. As expected the tables with finer partition (NACE with 3 digits) present a higher number of disclosive cells.

Table 2. Percentage of unsafe cells for different safety rules requiring that *RE* (z_1) > 0.5.

Rule	NACE 2 dig.		NACE 3 dig.	
	SBS 94	SBS 97	SBS 94	SBS 97
M_0 (1; 4) (Threshold)	14.07	13.52	29.61	29.16
M (1)	7.73	7.26	6.68	6.81
M (1; *l* = 1)	13.49	14.02	15.48	15.22
M (2)	19.26	18.65	21.63	20.68
MU (1)	30.33	28.54	32.97	32.05
MU (1; *n*)	25.61	24.28	26.33	25.45
MU (1; *l* = 1)	30.33	28.54	32.97	32.05
MU (1; *l* = 1, *n*)	30.33	28.54	32.72	31.94

Figure 2. Average relative absolute error committed by one intruder estimating t_2 for different rules.



We also compared the rules assuming the existence of two groups of two respondents in every cell. Hence we applied the rules with $m = 2$ and $l = 2$, together with the threshold rule $M_0(5)$ on the same data. Figure (2) compares the *RE* obtained with the different estimating procedures and table (3) shows the percentages of nonempty disclosive cells. Again, it is evident how the estimates obtained with the Uniform distribution give much lower average *RE* than the maximal estimates assumed in the M-rules, and the number of disclosive cells found with the MU-rules is much larger than for the corresponding M-rules.

Table 3. Percentage of unsafe cells for different safety rules requiring that $RE(t_2) > 0.5$ when it is estimated by a coalition of two intruders.

Rule	NACE 2 dig.		NACE 3 dig.	
	SBS 94	SBS 97	SBS 94	SBS 97
M_0 (2; 3) (Threshold)	22.38	21.78	43.29	42.92
M (2)	12.11	11.14	10.14	8.84
M (2; l = 2)	22.49	20.53	20.92	21.11
MU (2)	39.56	37.67	36.64	35.97
MU (2; n)	36.10	33.54	32.47	31.59
MU (2; l = 2)	39.22	36.55	36.29	35.40
MU (2; l = 2, n)	39.22	36.55	36.29	35.40

6. Conclusions

The protection provided by a safety rule depends on the assumptions taken to measure the risk. We show that contributions can be disclosed more precisely than what assumed in the M-rules by adopting a simple ignorance distribution. By clearly specifying the intruder's prior knowledge, safety rules can be adapted to different scenarios. In some situations it is appropriate to expand the safety rules to include groups of respondents. However, taking more stringent hypothesis lead to stricter rules, thus it is important to choose the rules for plausible hypothesis, rather than mechanically apply them, as sometimes may happen. The Uniform distribution used to derive the MU-rules is likely to be inappropriate for many data-sets, other, possibly skewed, distributions could be used and other rules may be derived.

References

Bernardo, J., Smith, A., 1994. Bayesian Theory. Wiley, NY.

Cox, L. H., 1981. Linear sensitivity measures in statistical disclosure control. Journal of Statistical Planning and Inference 5, 153–164.

Cox, L. H., 2001. Disclosure risk for tabular economic data. In: Doyle, P., Lane, J., Theeuwes, J., Zayatz, L. (Eds.), Confidentiality, disclosure and data access: theory and practical application for statistical agencies. Elsevier Science.

Duncan, G., Keller-McNulty, S., Stokes, S., 2001. Disclosure risk vs. data utility: the R-U confidentiality map. technical Report LA-UR-01-6428, Los Alamos National Laboratory.

Federal Committee on Statistical Methodology, 1994. Report on statistical disclosure limitation methodology, working paper 22.Subcommittee on Statistical Limitation Methodology, Washington, DC.

Fienberg, S. E., 2000. Confidentiality and data protection through disclosure limitation: Evolving principles and technical advances. The Philippine Statistician 49, 1–12.

Hundepool, A., 2004. The argus software in the casc project. privacy in statistical databases. Proceedings of the CASC Project International Workshop, June 9-11, 2004, Barcelona, Spain, 323–335.

ISTAT, 1997. Conti economici delle imprese con 20 addetti ed oltre. Anno 1994. Vol. 41 of Collana Informazioni. Istituto Nazionale di Statistica, Roma.

ISTAT, 2001. Conti economici delle imprese. Anno 1997. Vol. 19 of Collana Informazioni. Istituto Nazionale di Statistica, Roma.

Loeve, J. A., 2001. Notes on sensitivity measures and protection levels. research paper no. 0129. Methods and Informatics Department, Statistics Netherlands, Voorburg.Available at the CASC project home page *http://neon.vb.cbs.nl/casc/.*

Merola, G. M., 2003a. Generalized risk measures for tabular data. Proceedings of the 54th Session of the International Statistical Institute.

Merola, G. M., 2003b. Safety rules in statistical disclosure control for tabular data. Contributi Istat 1, istituto Nazionale di Statistica, Roma.

Trottini, M., 2001. A decision-theoretic approach to data disclosure problems. Researchin Official Statistics 4, 7–22.

Willenborg, L., de Waal, T., 2000. Elements of Statistical Disclosure Control. Springer-Verlag, New York.

Effects of Rounding on Data Quality

Jay J. Kim, Lawrence H. Cox, Myron Katzoff and Joe Fred Gonzalez, Jr
U.S. National Center for Health Statistics, Center for Disease Control and Prevention, Hyattsville, MD 20782, USA. Contact: pzk3@cdc.gov

Abstract: Integer data such as frequency counts may be rounded to integer values for purposes including disclosure limitation. It may be necessary to round noninteger data to integer data (*base 1 rounding*) for various statistical purposes, e.g., rounding expected sample counts (noninteger) to actual sample counts (integer). We evaluate the effects of four methods of rounding data on data quality and utility in two ways: (1) bias and variance (increase in total mean squared error) and (2) effects on the underlying distribution of the data (as measured, e.g., by the distance measure which can be considered a proxy chi-square statistic). The four rounding methods are conventional rounding, modified conventional rounding, zero-restricted 50/50 rounding, and unbiased rounding.

1. Introduction

Data are often rounded. Sometimes it is necessary to round noninteger values to integer values for statistical purposes. For example, at the end of sample weighting, the fractions are rounded to integers, since the number of persons or establishments cannot be fractions. Also data are rounded to enhance readability of the data, to protect confidentiality of records in the file, or to keep only the important digits.

Integers can be expressed as $x=q_xB+r_x$, where q_x is the quotient, integer B is the *rounding base*, and r_x is the remainder. B is a constant, but q_x and r_x are random variables. When the subscript x is not needed, it will be ignored. Four rounding rules are considered for rounding the remainder *r*. Note we will use R(x) to denote the rounded number of x and subscript i can be added whenever needed. This implies that R(x) = qB + R(r). For concreteness, we illustrate the rules for B = 10, so that r = 0, 1, 2, . . . ,9. Two important properties for evaluating and comparing rounding methods are as follows. *Unbiased rounding* satisfies: E[R(r) | r] = r. A weaker but still useful property is *sum-unbiasedness*: E[R(r)] = E[r].

The first rounding rule is *conventional rounding*: any r greater than or equal to B/2 = 5 is rounded up to B = 10; otherwise it is rounded down to zero. Conventional rounding is not unbiased but is sum-unbiased if and only if B is odd. The second rule is *modified conventional rounding*. This rule is the same as conventional rounding, except when r = B/2 (e.g., 5), r is rounded up to B = 10 or down to zero each with probability ½. Modified conventional rounding is sum-unbiased. The third is *zero-restricted 50/50 rounding*: r = 0 is rounded down and all nonzero r are rounded up or down with probabilities ½. It, too, is sum-unbiased. The last rule is *unbiased rounding* proposed by Nargundkar and Saveland. According to this rule, r is rounded up with probability r/10 and down with probability 1-r/10. Consequently unbiased rounding is unbiased and therefore also sum-unbiased. These rules are easily restated for any positive integer rounding base B.

We evaluate the effects of these rounding methods on data quality and utility in two ways: (1) bias and variance (increase in total mean squared error) and (2) effects on the underlying distribution of the data as evaluated by a distance measure.

2. Bias and Variance of the Rounded and Unrounded Numbers

We consider various distributions for the data, but assume that the remainders r follow a discrete uniform distribution.

2.1. Mean and Variance of Unrounded Data x and Remainders r

Since r takes values 0, 1, 2, ... , B-1 with uniform probability:

$$E(r) = \frac{B\text{-}1}{2} \tag{1}$$

$$V(r) = \frac{B^2 - 1}{12} \tag{2}$$

Therefore,

$$E(x) = E\left[E(x \mid q)\right] = BE(q) + \frac{B-1}{2}\,. \tag{3}$$

In general,

$$V(x) = V\left[E(x \mid q)\right] + E\left[V(x \mid q)\right] \tag{4}$$

Formula (4) will be used for deriving variance formulas for all cases. We obtain:

$$V(x) = B^2 V(q) + \frac{B^2 - 1}{12}\,^{.} \tag{5}$$

2.2 . Mean and Variance of R(x) for Conventional Rounding

If B is even, under conventional rounding, r is rounded up to B if r is greater than or equal to B/2; otherwise, it is rounded down to 0. If B is odd, we require: r rounds up to B if $r \geq \frac{B+1}{2}$, and rounds down to 0 otherwise.

Case 1. B is an Even Integer

$$E[R(r)] = B/2 \tag{6}$$

Comparing the expression in equation (6) with that in equation (1), we can see that they differ by 1/2. The rounded data overestimate the mean by ½, viz., the *absolute bias* is ½. The variance of the rounded data is:

$$V[R(r)] = B^2/4\,. \tag{7}$$

Note that the expression in equation (7) is approximately three times the variance of r in equation (2) where r is not rounded.

Using equation (4), we have

$$V\left[R(x)\right] = B^2 V(q) + \frac{B^2}{4}\,. \tag{8}$$

Hence,

$$MSE\left[R(x)\right]=B^2V(q)+\frac{B^2+1}{4} \tag{9}$$

Case 2. B is an Odd Integer

$$E\left[R(r)\right]=\frac{B\text{-}1}{2} \tag{10}$$

This expected value is exactly the same as in equation (1). Thus the rounded data provide a sum-unbiased estimator of the original data.

$$V[R(r)]=\frac{B^2-1}{4} \tag{11}$$

The variance of the rounded remainders in equation (11) is exactly three times that of the unrounded r in equation (2).

$$V\left[R(x)\right]=B^2V(q)+\frac{B^2-1}{4} \tag{12}$$

Since R(x) provides an unbiased estimator, the MSE of R(x) is the same as the variance of R(x) above.

2.3. Mean and Variance of R(x) for Modified Conventional Rounding

This rule is the same as conventional rounding rule, except that it allows for rounding r = B/2 up to B and down to 0, each with probability ½. It can be shown that:

$$P\left[R(r)=B\right]=\frac{B\text{-}1}{2B}$$

$$P\left[R(r)=0\right]=\frac{B+1}{2B}$$

Thus,

$$E\left[R(r)\right]=\frac{B\text{-}1}{2} \tag{13}$$

which is the same as for the unrounded r in equation (1), and

$$V[R(r)]=\frac{B^2-1}{4}. \tag{14}$$

This variance is exactly three times that for the unrounded remainders in (2). This rounding rule has the same mean, variance, and mean square error as those of conventional rounding when B is odd.

2.4. Mean and Variance of R(x) for Zero-Restricted 50/50 Rounding

Except for zero, all remainders r are rounded up or down with probability ½. Of course, zero remains zero after rounding. The probability the rounded remainder is B or 0 is the same as that observed with the modified conventional rounding. Hence this rounding rule has the same mean, variance, and mean square error as those of conventional rounding when B is odd.

2.5. Mean and Variance of R(x) for Unbiased Rounding

According to Nargundkar and Saveland's unbiased rounding rule, r is rounded up with probability r/B and rounded down with probability (B-r)/B. Thus,

$$P(r)=\frac{1}{B},\ \ P\big[R(r)=B\,|\,r\big]=\frac{r}{B}\ \text{and}\ P\big[R(r)=0\,|\,r\big]=\frac{B-r}{B},\ \text{for r}\ \geq 1,$$

$$P\big[R(r)=B\big]=\sum_{r=1}^{B\text{-}1}P(r)\,P[R(r)=B\,|\,r]=\frac{B\text{-}1}{2B},$$

$$P\big[R(r)=0\big]=\sum_{r=0}^{B\text{-}1}P(r)\,P[R(r)=0\,|\,r]=\frac{B+1}{2B}.$$

Since the above probabilities are the same as those observed with the modified conventional rounding, this rounding rule again has the same mean, variance, and mean square error as those of the conventional rounding rule when B is odd.

3. Distance Measure

The quality of the rounded data can be measured by the values of the distance measure of the rounding rules mentioned above. In comparing the rounded number with the original number, we can use the following measure for every number or cell subject to rounding:

$$U=\frac{[R(x)-x]^2}{x} \tag{15}$$

The numerator can be re-expressed as $[R(r_x)-r_x]^2$. Since $R(r_x)$ can be either B or 0, the above can be further re-expressed as $(\delta_x B-r_x)^2$, where δ_x is an indicator variable: $\delta_x = 0$ means round down and $\delta_x = 1$ means round up. We assume $U = 0$, when $x = 0$. The conditional expected value of U over δ_x is:

$$E_\delta(U\,|\,x)=\sum_{\delta_x=0}^{1}\left[\frac{(\delta_x B-r_x)^2}{x}\,|\,x\right]P(\delta_x) \tag{16}$$

Given x, δ_x is the only random variable in the above expression. When B = 10, with conventional rounding, $P(\delta_x=1)=1$ with r_x = 5, 6, . . ., 9. Otherwise, $P(\delta_x=0)=1$ with r_x = 0, 1, 2, 3, 4.

The expected value of U can be expressed as

$$E_q\left[E_r\left\{E_\delta(U\,|\,x)\right\}\right]=\sum_{q_x}\sum_{r_x}\sum_{\delta_x}[\frac{(\delta_x B-r_x)^2}{x}\,|\,x]P(\delta_x)P(r_x)P(q_x) \qquad (17)$$

In the above, the expectation over r is conditional on qB.

3.1. Conventional Rounding

For conventional rounding, assuming B even, let

$$U_1=E_r\left\{E_\delta(U\,|\,x)\,|\,q\right\}=\sum_{r_x}\sum_{\delta_x=0}^{1}[\frac{(\delta_x B-r_x)^2}{q_x B+r_x}\,|\,q_x]P(\delta_x)P(r_x)\,, \qquad (18)$$

which is

$$[\sum_{r_x=1}^{B/2-1}\frac{r_x^2}{q_x B+r_x}+\sum_{r_x=B/2}^{B-1}\frac{(B-r_x)^2}{q_x B+r_x}]\frac{1}{B} \qquad (19)$$

By separating the term with $q_x=0$ from those with $q_x\ge 1$, we have

$$U_1=[\sum_{r_x=1}^{B/2-1}r_x+\sum_{r_x=B/2}^{B-1}\frac{(B-r_x)^2}{r_x}]\frac{1}{B}$$

$$+\ [\sum_{r_x=1}^{B/2-1}\frac{r_x^2}{q_x B+r_x}+\sum_{r_x=B/2}^{B-1}\frac{(B-r_x)^2}{q_x B+r_x}]\frac{1}{B} \qquad (20)$$

The second component above is intractable, thus replacing r_x in the denominator with zero, we obtain the upper bound for the component. Thus U_1 is bounded as follows.

$$U_1\le[\sum_{r_x=1}^{B/2-1}r_x+\sum_{r_x=B/2}^{B-1}\frac{(B-r_x)^2}{r_x}]\frac{1}{B}$$

$$+[\sum_{r_x=1}^{B/2-1}\frac{r_x^2}{q_x B}+\sum_{r_x=B/2}^{B-1}\frac{(B-r_x)^2}{q_x B}]\frac{1}{B}\,. \qquad (21)$$

The expected value of the second component of the above equation over q_x reduces to

$$E_{q_x}[\frac{1}{q_x}][\sum_{r_x=1}^{B/2-1}r_x^2+\sum_{r_x=B/2}^{B-1}(B-r_x)^2]\frac{1}{B^2} \qquad (22)$$

The product of the second and third factors above, i.e., the expression in equation (22) excluding the expected q-reciprocal, is denoted by V as seen below.

$$V = [\sum_{r_x=1}^{B/2-1} r_x^2 + \sum_{r_x=B/2}^{B-1} (B-r_x)^2]\frac{1}{B^2} \qquad (23)$$

After some algebra, the above equation (23) reduces to

$$V = \frac{B^2+2}{12B} \qquad (24)$$

The first component of the upper bound of E(U) is the expected value of the upper bound of U_1 over q. However, the first component for U_1 in equation (21) does not involve q_x, hence the first component for U_1 is identical with that of E(U). The first component of U_1 turns out to be a harmonic series. An upper and lower bounds for a harmonic series whose last integer is n are: $\ln(n+1) < H_n \le 1+\ln(n)$. Using the above upper bound, we obtain the upper bound for the first component for U_1 as:

$$B\ \ln[\frac{2(B-1)}{B-2}] - \frac{B+1}{2}$$

Of course, $E_q(\frac{1}{q_x})$ varies depending on the distribution of q_x. We examine the expected q-reciprocal for three distributions in a separate section.

3.2 . Modified Conventional Rounding Rule

For modified conventional rounding, V in equation (23) is:

$$V = [\sum_{r_x=1}^{B/2-1} r_x^2 + \frac{B^2}{4} + \sum_{r_x=B/2+1}^{B-1} (B-r_x)^2]\frac{1}{B^2} \qquad (25)$$

This reduces to the same expression as the one in equation (24).

The first component for U_1 in equation (21) for this rounding rule is the same as that for the conventional rounding rule.

3.3. Zero-restricted 50/50 Rounding Rule

For the zero-restricted 50/50 rounding rule, we have:

$$V = \sum_{r_x=1}^{B-1} [r_x^2 + (B-r_x)^2]\frac{1}{2B^2} \qquad (26)$$

$$V = \frac{2B^2-3B+1}{6B} \qquad (27)$$

The first component for U_1 in equation (21) is

$$\frac{B}{2}\ln(B-1) + \frac{1}{2}.$$

3.4. Unbiased Rounding

For the unbiased rounding, the expectation of the numerator of U over δ_x is:

$$\sum_{\delta_x}(\delta_x B - r_x)^2 P(\delta_x)$$
$$=(B-r_x)^2\frac{r_x}{B}+r_x^2(\frac{B-r_x}{B}).$$

Thus

$$V=\sum_{r_x=1}^{B-1}[(B-r_x)^2\frac{r_x}{B}+r_x^2(\frac{B-r_x}{B})]\frac{1}{B^2} \quad (28)$$

This reduces to

$$V=\frac{B^2-1}{6B} \quad (29)$$

The first component of U_1 in equation (21) is $\frac{B-1}{2}$.

The first and second components for the expected U can be summarized for three rounding rules as follows. Note the modified conventional rounding rule has the same expected value of U as the conventional rounding rule.

Table 1. Two Terms of Upper Bound of E(U) for Three Rounding Rules

	Conventional rule	50/50 rounding	Unbiased rounding
1st term	$B\ \ln[\frac{2(B-1)}{B-2}]-\frac{B+1}{2}$	$\frac{B}{2}\ln(B-1)+\frac{1}{2}$	$\frac{B-1}{2}$
2nd term	$\frac{B^2+2}{12B}E(\frac{1}{q_x})$	$\frac{2B^2-3B+1}{6B}E(\frac{1}{q_x})$	$\frac{B^2-1}{6B}E(\frac{1}{q_x})$

The first and second terms for the three rounding rules are evaluated for two B values: 10 and 1,000.

Table 2. Comparison of Three Rounding Rules with B=10

	Conventional rule	50/50 rounding	Unbiased rounding
1st term	2.61	11.49	4.5
2nd term	$.85\,E_q(\frac{1}{q_x})$	$2.85\,E_q(\frac{1}{q_x})$	$1.65\,E_q(\frac{1}{q_x})$

In Table 2, it can be observed that conventional rounding has the lowest first and second terms. The first term of the unbiased rounding is 1.72 times that for conventional rounding. The first term of the zero-restricted 50/50 rounding rule is 4.4 times that for the conventional rounding. The second term of the unbiased rounding is just below twice that for conventional rounding. The second term of the 50/50 rounding is more than three times that for the conventional rounding.

Table 3. Comparison of Three Rounding Rules with B=1,000

	Conventional rule	50/50 rounding	Unbiased rounding
1st term	194	3,454	500
2nd term	$83\,E_q(\frac{1}{q_x})$	$323\;E_q(\frac{1}{q_x})$	$166\,E_q(\frac{1}{q_x})$

In Table 3, it can again be observed that conventional rounding has the lowest first and second terms. The first term of the unbiased rounding is 2.6 times that for conventional rounding. The first term of the zero-restricted 50/50 rounding rule is 18 times that for the conventional rounding. The second term of the unbiased rounding is twice that for conventional rounding. The second term of the 50/50 rounding is a little less than four times that for the conventional rounding.

Tables 2 and 3 clearly show that the conventional rounding or the modified conventional rounding has the lowest distance between the rounded and unrounded numbers.

4. Expected Value of 1/q or An Upper Bound

For three distributions, we derived the expected value formula for1/q, or the upper bound for the expected value of 1/q when the expected value formula is intractable. They are shown below.

4.1. E(1/q) for the Lognormal Distribution

Let the normal density function g(y) be

$$g(y\,|\,\mu,\sigma^2)=\left(\sigma\sqrt{2\pi}\right)^{-1}e^{-½(\frac{y-\mu}{\sigma})^2}, \qquad -\infty<y<\infty.$$

The lognormal distribution f(x) has the following form:

$$f(x\,|\,\mu',\sigma'^2)=(\sigma'\sqrt{2\pi}\,x)^{-1}e^{-½(\frac{\ln x-\mu'}{\sigma'})^2}, \qquad 0<x<\infty,$$

where μ' and σ' are the mean and variance of ln x. Since q cannot be zero, we use truncated lognormal distribution, truncated from below at one (1).

Let

$$c=\int_1^\infty f(x\,|\,\mu',\sigma'^2)dx.$$

Then

$$E(\frac{1}{q}\,|\,q\geq 1,\mu',\sigma'^2)=\frac{1}{c}\int_1^\infty\frac{1}{q}f(q)\,dq$$

$$=\frac{1}{c}\frac{1}{\sigma'\sqrt{2\pi}\,x}\int_1^\infty\frac{1}{x}\,e^{-½(\frac{\ln x-\mu'}{\sigma'})^2}dx \qquad (30)$$

To integrate by parts, set $\nu = \frac{\ln x - \mu'}{\sigma'}$, so that $d\nu = \frac{1}{x\sigma'} dx$ and $x = e^{\sigma'\nu + \mu'}$. Equation (30) becomes

$$E(\frac{1}{q} \mid q \geq 1, \mu', \sigma'^2) = \frac{1}{c} e^{\frac{1}{2}\sigma'^2 - \mu'} [1 - \Phi(\sigma' - \frac{\mu'}{\sigma'})] \qquad (31)$$

For example, when μ'=.25 and σ'=.25, the probability of the truncated lognormal distribution is .52283.

4.2. E(1/q) for the Pareto Distribution

The Pareto distribution is sometimes used to fit the size of firms, personal incomes and stock price fluctuations, etc. The Pareto distribution of the second kind has the following form:

$$f(q) = \frac{ak^a}{q^{a+1}}, \quad a > 0, \ q \geq k > 0 .$$

Again, for the distribution which is truncated from below at 1, we derive E(1/q).

$$E(\frac{1}{q}) = \int_k^\infty \frac{1}{q} \frac{ak^a}{c\, q^{a+1}} dq$$

$$= \left. \frac{a k^a \, q^{-a-1}}{-c(a+1)} \right|_k^\infty = \frac{a k^{-1}}{c(a+1)} \qquad (32)$$

where k above is the minimum value of q, and c is the cumulative probability from 1 to infinity of the Pareto distribution.

The method of moment’s estimator of a, a^* is

$$a^* = \frac{n\bar{x} - x_1'}{n(\bar{x} - x_1')},$$

where x_1' is the smallest sample value.

The maximum likelihood estimator of a is

$$\hat{a} = n[\sum_{j=1}^{n} \log(x_j / \hat{k})]^{-1},$$

where $\hat{k} = \min_i x_i$, or the smallest sample value of x.

4.3. Upper Limit for E(1/q) for the Multinomial Distribution

Let $q_1, q_2, q_3, \;.\;.$ follow multinomial distribution. That is,

$$f(q_1, q_2, \,.\,.\, q_k \mid p_1, p_2, \,.\,.\, p_k) = \frac{n!}{\prod_{i=1}^{k} q_i\,!} \prod_{i=1}^{k} p_i^{q_i}\;,\;\; q_i = 0, 1, 2, \;\;.\;\;.\;\;.$$

Once again, q’s cannot be zero, and we use a multinomial distribution truncated from below at 1. Since the expected value of the inverse q for the multinomial distribution is intractable, we obtain the upper bound for the expected value.

Note

$$E(\frac{1}{q_i}) \le E(\frac{1}{q_i+1}) + E[\frac{3}{(q_i+1)(q_i+2)}] \;\text{ for all i.} \qquad (33)$$

The expected value of interest is

$$E(\frac{1}{q_1\,q_2\;.\;.\;q_k}) = \sum\sum\;.\;.\;.\;\sum \frac{1}{q_1\,q_2\,.\,.\,q_k} \frac{n!}{\prod_{i=1}^{k} q_i\,!} \prod_{i=1}^{k} p_i^{q_i}\;.$$

The upper limit of the above expected value can be derived using equation (33). Let the size of the category i be n_i, $\sum_{i=1}^{k} n_i = n$ and n’ = $n - \sum_{i=1}^{j-1} n_i$, j = 2, 3, . Then we have

$$E(\frac{1}{q_1\,q_2\;.\;.\;q_k}) \le \prod_{i=1}^{k} [\frac{5(n+1)(n+2)p_i^2 r_i^{\sum_{j=1}^{i-1} n_j} + 2(n+2)p_i + 6}{2(n+1)(n+2)p_i^2 (1 - r_i^{\sum_{j=1}^{i-1} n_j})}] \qquad (34)$$

5. Concluding Comments

It is often necessary to transform statistical data, both count and continuous data, to integer values. Various methods of rounding and in some applications various choices for the rounding base B typically are available. The question becomes: which method and/or base is expected to perform best in terms of data quality and preserving distributional properties of original data and, quantitatively, what is the expected distortion due to the rounding? This paper provides a preliminary analysis towards answering these questions. In terms of bias, unbiased rounding is of course optimal. In terms of the distance measure, conventional or modified conventional rounding performs best.

References

Grab, E.L & Savage, I.R. (1954), Tables of the Expected Value of 1/X for Positive Bernoulli and Poisson Variables, *Journal of the American Statistical Association* **49**, 169-177.

N.L. Johnson & S. Kotz (1969). **Distributions in Statistics, Discrete Distributions**, Boston: Houghton Mifflin Company.

N.L. Johnson & S. Kotz (1970). **Distributions in Statistics, Continuous Univariate Distributions-1**, New York: John Wiley and Sons, Inc.

Kim, Jay J., Cox, L.H., Gonzalez, J.F. & Katzoff, M.J. (2004), Effects of Rounding Continuous Data Using Rounding Rules, *Proceedings of the American Statistical Association, Survey Research Methods Section*, Alexandria, VA, 3803-3807 (available on CD).

Vasek Chvatal. Harmonic Numbers, Natural Logarithm and the Euler-Mascheroni Constant. See www.cs.rutgers.edu/~chvatal/notes/harmonic.html.

Confidentiality Protection by Controlled Tabular Adjustment Using Metaheuristic Methods

Lawrence H. Cox
National Center for Health Statistics, Hyattsville, MD 20782, USA (lcox@cdc.gov)

Abstract. Controlled tabular adjustment is an SDL methodology based on a mixed integer linear programming model. We develop new hybrid heuristics and new meta-heuristic learning approaches for solving this model, and examine their performance. Our new approaches are based on partitioning the problem into its discrete and continuous components, and first creating a hybrid that reduces the number of binary variables through a grouping procedure that combines an exact mathematical programming model with constructive heuristics. We then replace the MILP with an evolutionary scatter search approach that extends the method to large problems with over 9000 entries. Finally, we introduce a new metaheuristic learning method that significantly improves the quality of solutions obtained.

Keywords. statistical disclosure limitation, mixed integer linear program, scatter search, adaptive learning

1. Introduction

The need to safeguard the confidentiality of survey data presents a monumental task, and government agencies must wrestle with this problem on a continuing basis. The continuing challenge is to maximize data quality and usability while preserving confidentiality. The focus of this paper is on *controlled tabular adjustment*, a method for confidentiality protection for tabular data, recently extended by Cox et al. (2004) to preserve data quality in addition to protecting confidentiality (*Q-P CTA*).

The importance of the confidentiality protection problem is confounded by its computational complexity. The primary mechanisms, cell suppression, controlled data rounding and perturbation, and controlled tabular adjustment, are expressed as decision problems subject to linear constraints involving potentially many binary variables. Moreover, NSOs must solve such problems on an ongoing basis and in many (survey) settings. Thus, the confidentiality problem for tabular data in most cases is not solvable optimally or even feasibly by standard algorithmic approaches.

In this paper, we analyze and augment the mixed integer linear programming formulation for controlled tabular adjustment. We conduct an empirical investigation of alternative methods for handling the underlying mixed integer/continuous optimization formulation that derives from this model. Our study creates new methods: a hybrid approach, a combined hybrid scatter search approach, and a metaheuristic learning approach. Our computational investigations disclose the difficulty of solving the problem due to the inherent combinatorial complexity of effective confidentiality protection, and illustrate how the new procedures can provide advances. Most significantly, we show that metaheuristic learning succeeds in improving the solutions to a degree that establishes these models as both a theoretical contribution and a truly practical advance in safeguarding sensitive data.

Controlled tabular adjustment (CTA) affords an opportunity to overcome many of the problems associated with traditional cell suppression and perturbation methods. CTA introduces controlled perturbations (*adjustments*) into tabular data that satisfy the protection ranges and tabular constraints (*additivity*) while minimizing data loss as measured by one of several linear measures of overall data distortion, such as the sum of the absolute values of the individual cell value adjustments. CTA typically replaces each sensitive cell by either of the two endpoints of its protection range, referred to as the *minimally safe values*. Selected nonsensitive cell values are then adjusted from their true values to restore additivity. Subject to assuring feasibility, nonsensitive cell perturbations are constrained to be small, such as within sampling variability, and cell values deemed undesirable for adjustment can be held fixed. Cox (2000) provides an early MILP formulation for CTA. The end result of CTA is a tabular system without suppressions meeting the disclosure rule, and close to original data with respect to a distortion measure.

A more extensive discussion of results reported here, and additionally comparison of simple heuristic procedures to exact solutions computing using the *ILOG CPLEX* ™ solver and limitations on computing exact solutions, are reported in Cox et al. (2006). Here we summarize development and analysis of new methods based on strategies of grouping and evolutionary scatter search. Scatter search offers particular advantages by running far more efficiently than CPLEX, and significantly extending the size of problems that can be addressed, yet still encounters limitations shared with its predecessors in generating solutions of high quality. For that reason, we develop a new metaheuristic learning algorithm that performs far more effectively than all other methods and provides a reliable and efficient approach for producing high quality solutions for problems of practical size.

2. Mixed Integer Linear Programming Model for CTA

The underlying concept of CTA is simple: the value of each sensitive cell is replaced by an *adjusted value* selected to be at a safe distance from the original value. Often, adjustment is to either of the sensitive cell's minimal safe values. Some or all nonsensitive cell values are then adjusted from their true values by small amounts to restore additivity to totals within the tabular system. Tabular data systems with marginal entries can be represented by their system of linear equations in matrix form: **MX = 0**. Column vector **X** represents the tabulation cells of the system; **x*** represents the original data. Matrix **M** is the *aggregation matrix* representing the tabular structure among the cells. The entries of **M** are –1, 0 or +1; each row of the **M** corresponds to one *aggregation* (tabular equation) in which "+1" denotes a contributing internal cell and "–1" a total (*marginal*) cell. With this notation, the mathematical structure of optimal synthetic tabular data is specified below by a mixed integer linear programming (MILP) formulation, containing binary and continuous variables, analogous to that introduced in Cox (2000). Our notation:

$i = 1,\dots, p$: denotes the p sensitive cells

$i = p+1,\dots, n$: denotes the (n-p) nonsensitive cells

B_i = binary (zero/one) variable denoting selection of the lower/upper limit for sensitive cell $i = 1,\dots,p$

L_i = lower adjustment required to protect sensitive cell $i = 1,\dots,p$

U_i = upper adjustment required to protect sensitive cell $i = 1,\dots,p$

y_i^+ = nonnegative continuous variable identifying a positive adjustment to cell value i

y_i^- = nonnegative continuous variable identifying a negative adjustment to cell value i

UB_i, LB_i = upper/lower cell capacities on change to cell i

c_i = cost per unit change in cell i

MILP for Optimal Controlled Tabular Adjustment (to Minimal Safe Values)

$$\text{Min} \sum_{i=1}^{n} c_i \, (y_i^+ + y_i^-) \qquad (1)$$

Subject to:

For $i = 1,\dots, n$:

$\mathbf{M\,(\,y^{+} - \,y^{-}\,) = 0}$ (2)

$0 \leq y_i^{+} \leq UB_i$ (3)

$0 \leq y_i^{-} \leq LB_i$ (4)

For i = 1,…, p:

$y_i^{+} = U_i\, B_i$ (5)

$y_i^{-} = L_i\,(\,1 - B_i\,)$ (6)

After solving the MILP, the adjusted tabular data **t** = (t_i) are: $t_i = x^*_i + y_i^{+} - y_i^{-}$. The objective function (1) minimizes the cost due to cell deviations. Linear costs are typically defined over the net adjustment $y_i^{+} + y_i^{-}$. Two cost functions commonly used are: all $c_i = 1$, to minimize total absolute adjustment, and $c_i = 1/\,x^*_i$ for nonzero cells, to minimize total percent absolute adjustment.

It is possible that (2) – (6) gives rise to an infeasible problem. Relaxing the sensitive cell constraints eliminates a large number of these types of problems:

$y_i^{+} \geq U_i\, B_i$ (7)

$y_i^{-} \geq L_i\,(1 - B_i\,)$ (8)

3. Hybrid Heuristic

Because computation for the MILP roughly doubles with the addition of each binary variable, a sensible approach towards a computationally efficient, near-optimal algorithm is to group the sensitive cells, assign a unique binary variable to the group, and adjust all cells in a group in the same direction. We first tried random grouping, which performed poorly. We suggest ordering sensitive cells from largest to smallest, and assigning variables to different groups successively. This encourages between-group homogeneity, so large cells are less likely to be adjusted predominantly in one direction, expected to improve the solution. An exception: if a sensitive cell value equals one of its totals, both are assigned to the same group.

Let M ≥ 2 be the number of groups. Add these constraints to the mathematical program: For i=1 to M, $B_i = B_{i+M} = B_{i+2M} = \ldots B_{i+kM}$, for (i+kM) ≤ p. This reduces the number of binary variable to M. If M = p then the solution is optimal and if M < p then the solution may or may not be optimal. The mathematical program can be enhanced with additional constraints to improve the statistical characteristics of the solution (Cox et al. 2004). The Hybrid may be run multiple times and the best solution selected: we used groups of size = M, M-1, M-2, …., to produce a range of results and chose a superior solution. The Hybrid is more sophisticated than simply ordering cells by size and assigning directions alternately, as it does not predefine directions and evaluates M^2, not just one, assignments.

To evaluate the effectiveness of the Hybrid, sets of 2- and 3-dimensional test tables were randomly generated using the following specifications:

- 2-dimensional tables ranging in size from 4x4 to 25x25.
- 3-dimensional tables having sizes: nxnxn for n = 5,6,…11,12…20
- 3-dimensional tables having sizes: 10x10xn for n= 3,4,…,19,20
- Data values for internal tabular entries range from 0 to 1000 and are selected from a uniform distribution.
- 10% of the internal entries are selected randomly (uniformly distributed) and are assigned a value of 0.
- For all tables, 30% of the internal entries are defined as sensitive. The sensitive cells are distributed randomly (uniform) throughout the table. Marginal cells are not defined as sensitive.

- Sensitive entries must be assigned a value 20% greater than the original value or 20% smaller than the original value. All nonsensitive cells can be modified to values within 20% of their original values.
- In all tables, the sum of absolute changes is minimized.

Figure 1 shows the performance of heuristics compared to the optimal solution for moderately sized 2-dimensional tables. The heuristics are: Hybrid with M = 16, Ordering-With-Alternate-Assignment, and Best-Among-Random-Assignment over 100 and 1000 repetitions. The optimal solution curve is not displayed because its information is embodied in the report of the percent error of heuristic solutions with respect to optimal. M=16 was chosen to provide solutions in approximately the same time as required by Random-1000. The results indicate that the Hybrid is superior.

Figure 2 shows results for 3-dimensional tables. Optimal solutions could not be obtained for the larger tables, so Percentages are those relative to the Best-Heuristic solution, which, in almost every case, is achieved by the Hybrid heuristic. These results indicate that creating groupings of sensitive cells can significantly extend the applicability of the integer-programming model.

Finally, we explore an advanced approach for building groups. The principle is to minimize the number of potential within-group conflicts so that assignments do not produce large perturbations to totals. First, M groups are formed using the previous approach. For each group, we calculate the number of totals that are in common with each pair of cells, called the *group score*. We then *swap* cells between groups to decrease the grand total of all group scores. Swaps are continued until no further score reduction is possible. The resulting groups are then used to populate the mixed integer program. This procedure is referred to as Hybrid-With-Swaps. This strategy improved solutions approximately 10% on average.

Figure 1. Performance of Hybrid on 2-dim tables based on percent error; 30% sensitive cells

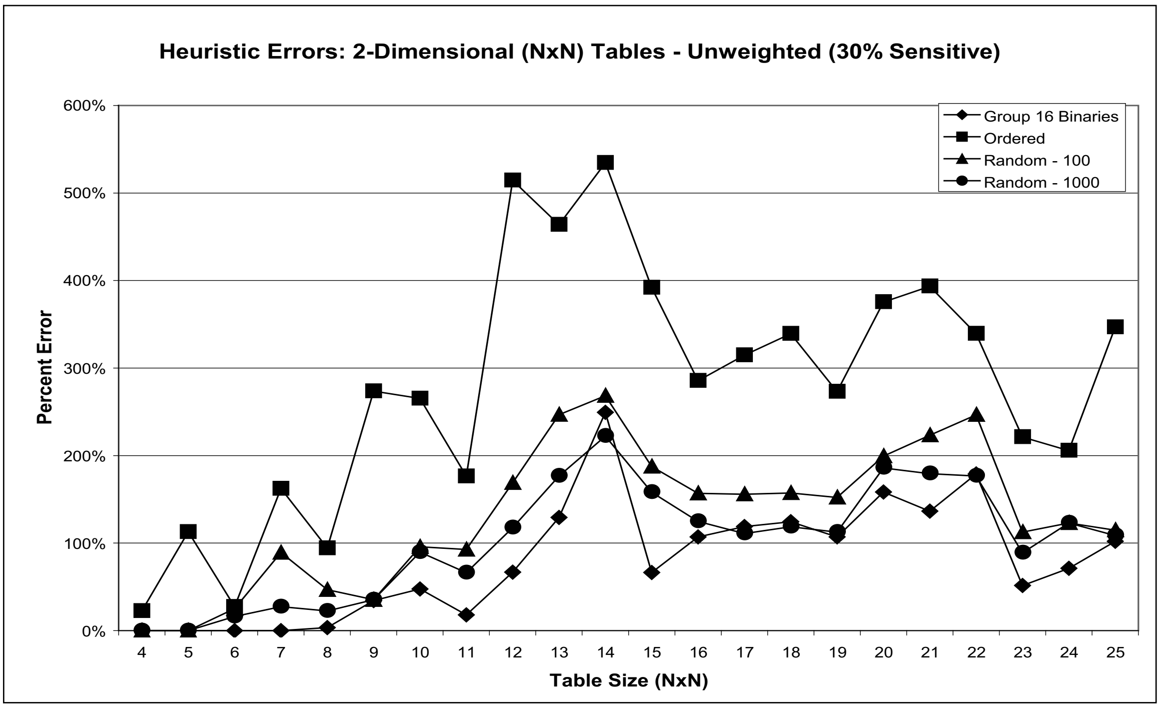
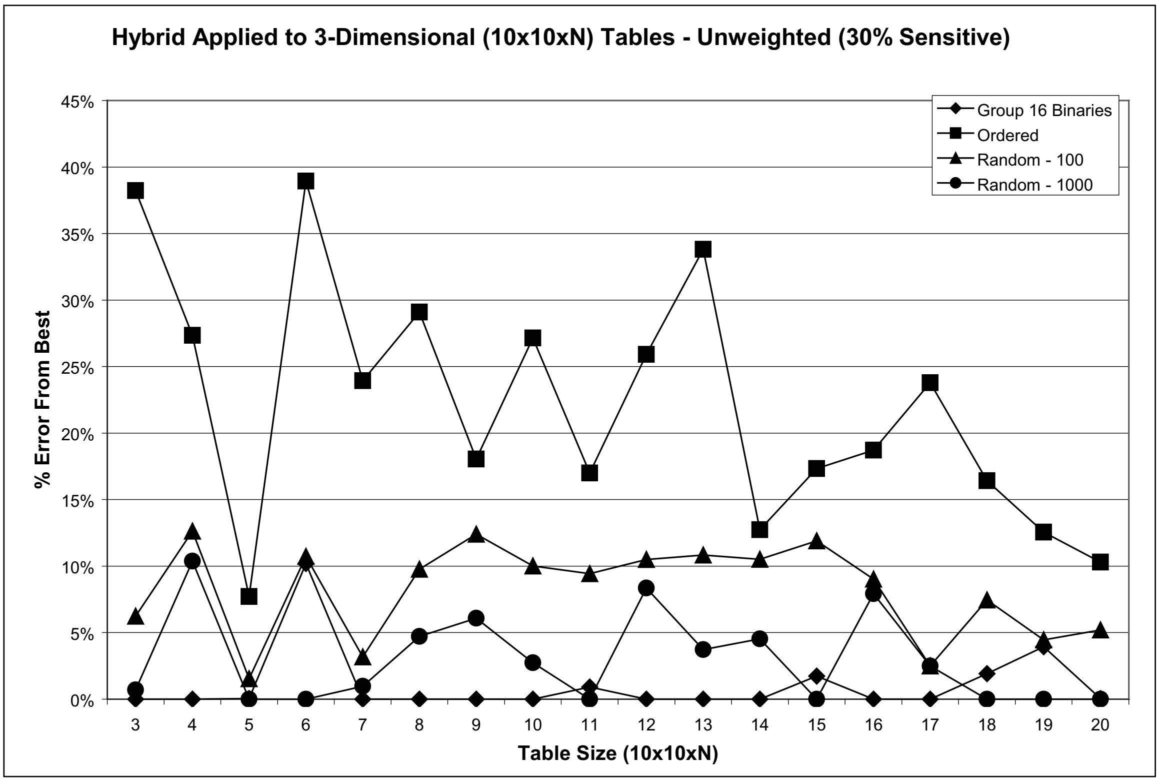


Figure 2. Performance of Hybrid on 3-dim tables based on percent error; 30% sensitive cells



4. Scatter Search to Enhance Hybrid Heuristic

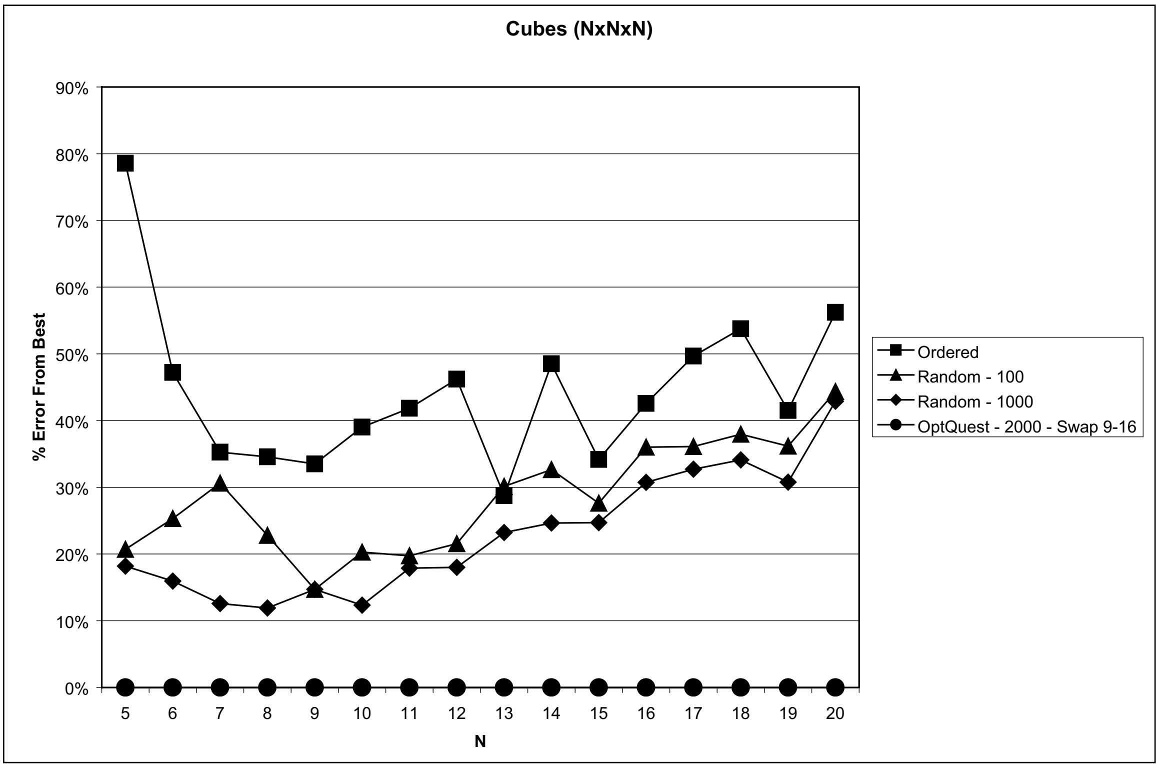
Using the mixed integer programming based approach becomes impractical when the number of tabular entries exceeds 1000, e.g., the 10x10x20 table in Figure 2 required 76 minutes of computational time on 2.8GHz, Pentium 4 , 512 MB PC. To overcome this limitation, we used an evolutionary *scatter search* procedure (Laguna and Marti 2003). Scatter search is designed to operate on a set of points, called *reference points*, which constitute good solutions obtained from previous efforts. The basis for "good" includes criteria, e.g., diversity, that go beyond the objective function value. Scatter search then generates new points as combinations of the reference points. Combinations are generalized forms of linear combinations, accompanied by processes to adaptively enforce feasibility.

Points are considered *diverse* if their elements are "significantly" different from one another. The optimizer uses Euclidean distances to determine how close a potential new point is from those in the reference set, in order to decide whether the point is included or discarded. The number of solutions created depends on the quality of the solutions being combined, viz., combining the best two reference solutions generates up to five new solutions, while combining the worst two generates only one.

Combination may not generate solutions of enough quality to join the reference set, in which case a diversification step is triggered. The reference set is rebuilt to balance solution quality and diversity. Quality is preserved by seeding the reference set with a small subset of *elite solution*s; diversification is used to repopulate the reference set with solutions diverse relative to the elite set.

We used the *OptQuest*™ solver to implement the scatter search method for the CTA problem. Figure 3 shows the results of the scatter search method used in combination with hybrid and swap. Figure 3 provides results from taking the best solution obtained from using M= 9, 10, …, 16 (encompassing cases M = 1, …, 8). This experiment provided the best solutions in all cases and only doubled the computation time required for the M=16 run. It should also be noted that for all tables N≤ 10 the scatter search heuristic solutions were optimal. For larger tables, CPLEX was unable to run to optimality of the scatter solutions.

Figure 3. Performance of scatter search in combination with Hybrid-With-Swaps on cubic 3-dim tables based on percent error; 30% sensitive cells



5. Metaheuristic Learning Algorithm for Binary Variables

5.1. Learning algorithm

The grouping heuristics proposed in the previous section significantly reduced the problem size and thereby quickly solved the resulting integer program. However, these methods failed to produce satisfactory solutions for problems beyond a relatively limited size. The Best heuristic solution was at least 50% inferior to the optimal solution for all moderately large 2 dimensional tables. Moreover, heuristics exhibited considerable variation in the solution quality. These experiments demonstrate the importance of reducing the size of the integer programs for gaining computational efficiency. Inferior performance of these methods is attributed to their inability to predict and set appropriate values for a subset of variables. In this section we show that a metaheuristic learning strategy for fixing a subset of variables can be exceedingly useful for generating high quality solutions without consuming vast amounts of computer time to discover such solutions. This is based on the *proximate optimality principle*, which implies that a good solution at one level is likely to lead to good solutions at adjacent levels (Glover and Laguna 1997).

5.2. Parametric image

Our approach creates a strategic image of part of the problem to generate information about problem characteristics. Such processes have been used successfully in the fixed charge context (Glover et al. 2003), and are the basis for a class of metaheuristics procedures for mixed integer programming proposed in Glover (2003). Adapted to the present setting, the basic idea is to introduce parameters that penalize a variable's violation of integer feasibility, and to drive selected subsets of variables in preferred directions, viz., towards 0 or 1.

We are interested in identifying appropriate directions for selected subsets of binary variables, which are then tentatively fixed at their preferred values. The resulting reduced problem is then solved much more readily than the original problem providing an iterative process that results in high quality (optimal or near-optimal) solutions while expending only a small fraction of the computational effort required by a more traditional integer programming solution approach. We utilize this strategy to develop a parametric objective function approach to generate information on behavior of binary variables in the following manner.

We represent the objective in the compact form: minimize $x_o = cx$, where **x** is set of binary variables used to protect sensitive cells. We refer to "1" direction as (UP) and "0" direction as (DN) direction in our framework. These are called goal conditions (denoted as $x_j^{'}$) because we do not seek to enforce (UP) and (DN) directions by imposing them as constraints in the manner of customary branch and bound method but rather indirectly by incorporating them into the objective function of the linear programming relaxation. N^+ and N^- denote selected subsets of N containing UP and DN goal conditions; their union is $N^{'}$. $x^{'}$ denotes the associated goal imposed solution vector and M a very large positive number used as a penalty:

$$(LP') \text{ Minimize } x_o^{'} = \sum_{j\in N^-}(c_j + M)x_j + \sum_{j\in N^+}(c_j - M)x_j + \sum_{j\in N/(N^+ + N^-)} c_j\, x_j \qquad (9)$$

(*LP'*) targets goal conditions by incentive driven by penalty M. Binary variables of N^- are induced to go DN and those in N^- to go UP. Remaining variables are free to select direction. We are solving a linear program with penalty coefficients in the objective to gain insight about good values for binary variables.

5.3. Goal infeasibility and resistance

If a variable favors a particular direction, then it will achieve its targeted goal; otherwise, it will show some resistance to its imposed goal. We say that an optimal LP solution **x** = $x^{''}$ is *goal infeasible* if: for some $j \in N^+, x^{''}_j < x^{'}_j$ **(V-UP)**, or, for some $j \in N^-, x^{''}_j > x^{'}_j$ **(V-DN)**

We call a variable x_j associated with violation (V-UP) or (V-DN) a *goal infeasible variable*. We create a measure of *overt resistance* (β UP, β DN), based on goal conditions, to learn about variable predilection for either direction:

$$\text{For (V-UP), } \beta UP_j = x_j^{'} - x_j^{''} \qquad (10)$$

$$\text{For (V-DN) } \beta DN_j = x_j^{''} - x_j^{'} \qquad (11)$$

No goal violation means zero overt resistance. If a variable does not violate its goal condition, it may *potentially resist* it: potential resistance = (δ *UP*, δ *DN*):

$$\delta UP_j = M + c_j + RC_j \qquad (12)$$

$$\delta DN_j = -(-M + c_j + RC_j) \qquad (13)$$

where *RC,* is a suitably reduced cost for variable x_j.

The trial solution vector may contain variables without penalties. We use their solution values for the problem (LP) to create free resistances (αUP,αDN):

$$\alpha UP_j = 1 - x_j \qquad (14)$$

$$\alpha DN_j = x_j \qquad (15)$$

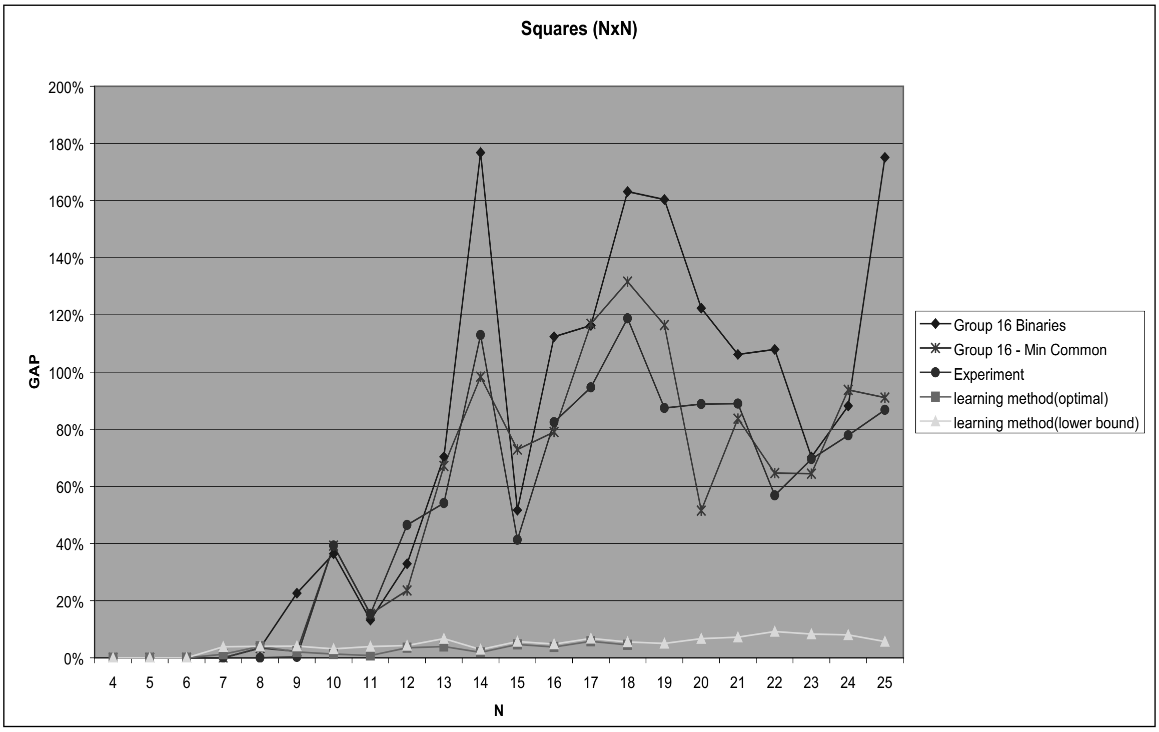
The parametric image of the objective is generated using a goal vector. A diversified sample of goal vectors is generated and resistance measures recorded to estimate directional effects. See Cox et al. (2006) for specification of the *parametric image learning algorithm*. The parametric image of the objective is:

$$x_o' = \sum_{j\in N^-}\left(c_j + M\right)x_j + \sum_{j\in N^+}\left(c_j - M\right)x_j + \sum_{j\in N/(N^+ + N^-)} c_j\, x_j \qquad (16)$$

5.4. Performance of the learning algorithm for 2-dimensional tables

We implemented the learning algorithm using C++, ILOG- Concert Technology 1.2, and ILOG-CPLEX 8.1. Figure 4 shows the performance of our proposed learning method compared to other variable fixing heuristics.

Figure 4. Performance of metaheuristic learning algorithm on optimality gap



The 25x25 problem exhibited an optimality gap of 9.6 %, but direct verification of the optimum was prohibitive. We needed a computationally efficient good lower bound on the optimum to measure the gap. (Cox et al. 2005) proposed a set partitioning based method, which we used as a proxy optimum for computing the gap in larger problems. These lower bounds were consistently very close to the optimum, e.g., for 2-dimensional tables restricted in size to no more than 18 rows and columns, the optimality gap was approximately 1%. In Figure 4, the "learning method (optimal)" curve identifies the optimality gap with respect to the known optimal value, and the "learning method (lower bound)" curve identifies the optimality gap with respect to the lower bound.

6. Concluding Comments

This study has undertaken an extensive set of comparative computations tests and analyses to evaluate the relative performance of alternative methods for the controlled tabular adjustment (CTA) model. Our preliminary tests of previously proposed heuristics compared to the exact CPLEX method confirm that the exact procedure yields superior solutions, but is unable to solve problems of modest size within a reasonable amount of time. To overcome limitations of existing approaches, we introduced a Hybrid heuristic that combines the exact mathematical programming approach with constructive heuristics. Numeric simulations indicate that the Hybrid has the ability to produce better solutions than previous heuristics in reasonable time, and has the added advantage of being able to find reasonable solutions to highly constrained problems, but is limited to problems that remain of modest size. We then show that an evolutionary scatter search approach in place of the exact CPLEX solver yields improved results and makes it possible to handle problems of much greater size, though still is unable to overcome the combinatorial complexity of these problems to achieve solutions that appear attractive in relation to optimality bounds. Finally, we demonstrate that a metaheuristic learning method based on parametric image processes leads to significant additional improvements by generating solutions of greatly improved quality. We anticipate that opportunities exist to improve our results further. Interactions between binary variables are likely to be present, especially variables corresponding to cells sharing the same tabular equation, and plan to pursue this in the next phases of our research. We also plan to extend our investigations to the quality-preserving controlled tabular adjustment model of Cox et al. (2004).

References

Cox, L.H. (2000), “Discussion (on Session 49: Statistical Disclosure Control for Establishment Data),” in: **ICES II: The Second International Conference on Establishment Surveys-Survey methods for businesses, farms and institutions**, Invited Papers, Alexandria, VA: American Statistical Association, 904-907.

Cox, L.H., Glover, F., Kelly, J.P. & Patil, R.J. (2006), “Confidentiality Protection By Controlled Tabular Adjustment: An Analytical and Empirical Investigation of Exact, Heuristic and Metaheuristic Methods,” *Decision Sciences Institute*, to appear.

Cox, L.H. & Kelly, J. P. (2004), “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., 2004, 11-23.

Cox, L.H., Kelly, J.P. & Patil, R.J. (2004), “Preserving Quality and Confidentiality for Multivariate Tabular Data,” in: **Privacy in Statistical Databases 2004 (PSD 2004), Lecture Notes in Computer Science 3050** (J. Domingo-Ferrer and V. Torra, eds.) New York: Springer-Verlag, 87-98.

Cox, L.H., J.P. Kelly & Patil, R.J. (2005), “Computational Aspects of Controlled Tabular Adjustment: Algorithm and Analysis,” in: **The Next Wave in Computer, Optimization and Decision Technologies** (B. Golden, S. Raghavan and E. Wasil, eds.), Boston: Kluwer, 45-59.

Glover, F. (1977), “Heuristics for Integer Programming Using Surrogate Constraints,” *Decision Sciences* **8**, 156-166.

Glover, F. (2004), “Parametric Tabu Search Methods for Mixed Integer Programming,” Leeds School of Business, University of Colorado, Boulder.

Glover F. & Laguna M.(1997), **Tabu Search**, Kluwer, Boston.

Karger, D.R. (1999), “ Random Sampling in Cut, Flow, and Network Design Problems,” *Mathematics of Operations Research* **24(2),** 383-413.

Laguna, M. & Marti, R. (2003), **Scatter Search: Methodology and Implementations in C**, Kluwer, Boston.

Lewis, M.W. (2004),”Solving Fixed Charge Multi-Commodity Network Design Problems using Guided Design Search,” University of Mississippi, Hearin Center Technical Report, HCES-01-04.

Montgomery, D.C.(1984), **Design and Analysis of Experiment***s*, John Wiley and Sons, New York.

Information Loss Measures for Frequency Tables

Natalie Shlomo and Caroline Young***

*** Southampton Statistical Sciences Research Institute, University of Southampton, Department of Statistics, Hebrew University, Office for National Statistics**

**** University of Southampton, Office for National Statistics**

Abstract: In order to manage the disclosure risk in frequency tables containing population counts, the tables undergo statistical disclosure control (SDC) methods. This results in information loss. We examine quantitative information loss measures for frequency tables and compare them across different SDC methods. We show examples of the information loss measures on real UK 2001 Census tables after they have been perturbed. We study the relationship between the results of the information loss measures, the perturbation method and the characteristics of the table (sparsity, skewness, uniformity, etc.).

1. Introduction

The Office for National Statistics (ONS) is leading the development of a new Internet service, Neighborhood Statistics (NeSS), which provides access to tables containing administrative and census data for small areas. The object is to supply the information needs for the National Strategy for Neighborhood Renewal and deliver small area statistics for policy evaluation, informing new developments in areas of deprivation and for addressing issues arising in local areas. The statistical disclosure control (SDC) methods at the ONS for protecting NeSS tables containing population counts include post-tabular methods: controlled rounding and cell suppression implemented using the Tau-Argus Statistical Disclosure Control Software (Hundepool (2003)), and stochastic unbiased forms of random rounding and small cell adjustments. Each of these methods modify the original data in the table in order to reduce the disclosure risk resulting in small cells of the tables. Reducing disclosure risk however results in information loss. In this paper we develop and evaluate quantitative information loss measures for determining the impact of the SDC methods on the original table.

Information loss measures can be split into two classes: measures for data suppliers in order to make informed decisions about optimal SDC methods which depend on the characteristics of the tables, and measures for users in order to allow adjustments to be made when carrying out statistical analysis on protected tables. In this paper, we focus on measures for data suppliers who have access to the raw tables and the aim is to choose the best SDC method which minimizes the information loss.

The SDC methods reviewed in this paper all give adequate protection against disclosure by identification since the small cells are eliminated from the tables. However, small cells also result from differencing nested non-coterminous tables. The cell suppression and small cell adjustments do not protect against disclosure by differencing whereas the full rounding methods do. Therefore, in order to obtain the same level of protection for all the SDC methods in this analysis, we assume that only one set of coding of the variables and geographies are disseminated in the tables and that there is no risk of disclosure by differencing.

Section 2 introduces the SDC methods that will be compared in the paper and Section 3 the data used for analysis. Section 4 presents information loss measures with numerical and graphical results on the data. We conclude in Section 5 with a discussion.

2. Data Masking Techniques for Frequency Tables

Some methods for protecting frequency tables against the disclosure risk of small cells in tables are:

2.1. Small Cell Adjustments (SCA)

Small cell adjustments is an unbiased random rounding procedure carried out on the small cells of the tables (ones and twos). Let *x* be a small cell and 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. The expected value of the rounded entry is the original entry. Each small 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. For this analysis, we randomly rounded to base 3. For each cell, the mean of the perturbation is 0 and the variance is 2. When only small cells are rounded, the margins of the tables are obtained by aggregating the rounded and non-rounded cells, and therefore tables with the same population base will have different totals due to the stochastic process.

2.2. Full Random Rounding (RaRo)

Random rounding is carried out on all entries in the table. This is implemented as described above for the small cells after first converting the entries *x* to residuals of the rounding base *res(x).* Because of the large number of perturbations in the table, the margins are rounded separately from the internal cells and therefore tables are not additive. We implemented random rounding to base 3 for this analysis.

Although we implemented the small cell adjustments and the full random rounding independently in each cell for this analysis, we note that the random rounding procedure can be improved by controlling for some of the marginal (and overall) totals of the table. A very simple algorithm for semi-controlling the random rounding procedure which preserves row totals and the overall total (or column totals after first transposing the table) is as follows:

1. Convert the entire table so that the entries are residuals of the rounding base.
2. Select first row of the table and randomly sort the entries.
3. For those entries having $res(x)$, select first $\frac{res(x)}{b}$ of the entries and round upwards, the rest of the entries round downwards. Repeat for all $res(x)$.
4. Sort entries back into their proper order.
5. Repeat on next row.

2.3. Controlled Rounding (CR(3))

Controlled rounding is a complex procedure carried out in Tau-Argus which is intended to be used as an ONS standard tool for disclosure control on frequency tables. In particular, it is largely supported by NeSS for protecting administrative register based tables disseminated over the internet. The procedure uses sophisticated linear optimization programming techniques to round entries, where the constraint is the equality of the rounded margins to the sum of the interior rounded cells. The algorithm for controlled rounding can also be carried out on the small cells only of the table, thus preserving totals and marginal distributions while only perturbing the small cells. All tables were controlled-rounded to base three for this analysis.

2.4. Suppression (S-A and S-WA)

Tau-Argus can also be used to suppress sensitive cells in frequency tables. Sensitive cells are defined as having counts of a one or a two. To apply secondary suppressions, the Hypercube method was chosen since a solution could be obtained for all tables in this analysis. This ensured a fair assessment of performance across tables. Note that due to the nature of the Hypercube method, occasionally some relatively large counts in the tables were secondary suppressed (Geissing (2003)).

In order to assess information loss from the perspective of what a user might do with suppressed cells in a table prior to analyzing the data, we implemented two very simple methods of imputation for the suppressed cells. Note that more sophisticated techniques for filling in missing data were not carried out in this analysis, but will be developed in future research. Let m_{kj} be a cell count in a two way table $k = 1,...,K$ rows and $j = 1,...J$ columns. For NeSS and Census tables at the ONS, rows are typically geographies: outputs areas or wards. The columns are defined by cross-classified variables, for example sex*long term illness*economic activity. Let the marginal totals be defined as: $m_{k.}$ and $m_{.j}$. The margins appear in the table without perturbation unless they have a small value and are primary suppressed. In that case, we define the margin to take a value of 1 for the following imputation schemes. Let z_{kj} be an indicator taking on the value of 1 if the cell was suppressed (primary or secondary) and a 0 otherwise.

In the first method (S-A), the user replaces all suppressed cells of row k with an average total of the suppressed values, i.e. $\dfrac{m_{k.} - \sum_{j=1}^{J} m_{kj}(1 - z_{kj})}{\sum_{j=1}^{J} z_{kj}}$. In the second method (S-WA), we use a weighted average to replace suppressed cells in a row *k*. The weights are based on the average cell size of the columns *j*: $w_j = \dfrac{m_{.j}}{J}$. A low frequency column will result in a smaller imputed cell frequency and a high frequency column will result in a larger imputed cell frequency. Each suppressed cell in row k is replaced by:

$$\frac{w_j \times (m_{k.} - \sum_{j=1}^{J} m_{kj}(1 - z_{kj}))}{\sum_{j=1}^{J} w_j z_{kj}} .$$

3. Data Used

For the purpose of this analysis we used three 2001 Census tables from one Estimation Area of the UK in the Southwest part of the country. The area included 437,744 persons in 182,337 households in 70 wards (on average 6,250 persons to a ward for this Estimation Area). The tables were the following:

(1) Tenure(3) * Age (7) * Health(4) * Ward

(2) Ethnicity (17) * Ward

(3) Economic Activity (9) * Sex (2) * Long-Term Illness (2) * Ward

The Economic Activity table only includes employed persons. The different SDC methods as described in Section 2 were implemented on the tables.

Table 1 provides summary statistics for each of the tables. Tenure is the largest table in terms of number of cells but also the sparsest with many small cells. The Ethnicity table contains large cell counts for the ethnic 'white' group defined in one column of the table and this is reflected in high skewness and high standard error of the cell counts. In comparison, the Employment table consists of both large and small cell counts.

Table 1. Summary Statistics of Tables

		Table		
		Tenure	**Ethnicity**	**Employment**
Number of Person in Table		433,817	433,817	317,064
Number of Cells		5,880	1,120	2,520
Average cell size and Standard Error		73.8 (3.3)	387.3 (51.3)	125.8 (6.6)
Average cell size in row	**Minimum**	0.0	0.1	0.0
	Maximum	171.4	899.9	309.3
Average cell size in column	**Minimum**	0.2	2.8	3.0
	Maximum	943.7	5,729.4	1,411.7
Percentage of Zero Cells		26%	23%	17%
Percentage of Small Cells		12%	9%	9%

4. Information Loss Measures

Information loss measures can be divided into several subsets according to the statistical aspect that is to be measured: Measures for distortion to distributions; Impact on the variance of estimates; Impact on measures of association; Statistical hypothesis tests for bias; Impact on statistical analysis, i.e. Goodness of fit criteria, Rank correlations.

4.1. Measuring distortion to distributions

Information loss measures that measure distortion to distributions are based on distance metrics between the original and perturbed cells. Some useful metrics were presented in Gomatam and Karr (2003). Since the basic unit of most of the Census and NeSS tables is a geography, i.e. ward, we calculate a distance metric for each ward separately in the table and then take the overall average across all of the wards for the information loss measure. When comparing the average distance metric across wards, we need to take into account the level of dispersion as expressed by the standard error (confidence interval).

Changing the notation from the previous section, let D^k represent a table for a ward k and let $D^k(c)$ be the cell frequency c in the table. Let $|W|$ be the number of wards in the estimation area. The distance metrics are:

– Hellinger's Distance:

$$HD(D_{pert}, D_{orig}) = \frac{1}{|W|}\sum_{k=1}^{|W|}\sqrt{\sum_{c\in k}\frac{1}{2}(\sqrt{D^k_{pert}(c)} - \sqrt{D^k_{orig}(c)})^2}$$

– Relative Absolute Distance:

$$RAD(D_{pert},D_{orig})=\frac{1}{|W|}\sum_{k=1}^{|W|}\sum_{c\in k}\frac{|D_{pert}^{k}(c)-D_{orig}^{k}(c)|}{D_{orig}^{k}(c)}$$

– Average Absolute Distance per Cell:

$$AAD(D_{pert},D_{orig})=\frac{1}{|W|}\sum_{k=1}^{|W|}\frac{\sum_{c\in k}|D_{pert}^{k}(c)-D_{orig}^{k}(c)|}{|k|} \qquad \text{where} \qquad |k|=\sum_{c}I(c\in k)$$

the number of cells in the k^{th} ward.

These distance metrics can also be calculated for sub-totals and totals of the tables. In this report, we use a distance metric defined by the individual perturbations for sub-totals: $PA(N_{pert}^{k},N_{orig}^{k})=N_{pert}^{k}(C')-N_{orig}^{k}(C')$ where $N^{k}(C')=\sum_{c\in C'}D^{k}(c)$ is a sub-total for group C'. Table 2 presents results of the information loss measures based on average distance metrics across wards for the tables and their confidence intervals.

Table 2. Average Distance Metrics Between Original and Perturbed Tables per Ward (95% confidence intervals in parentheses)

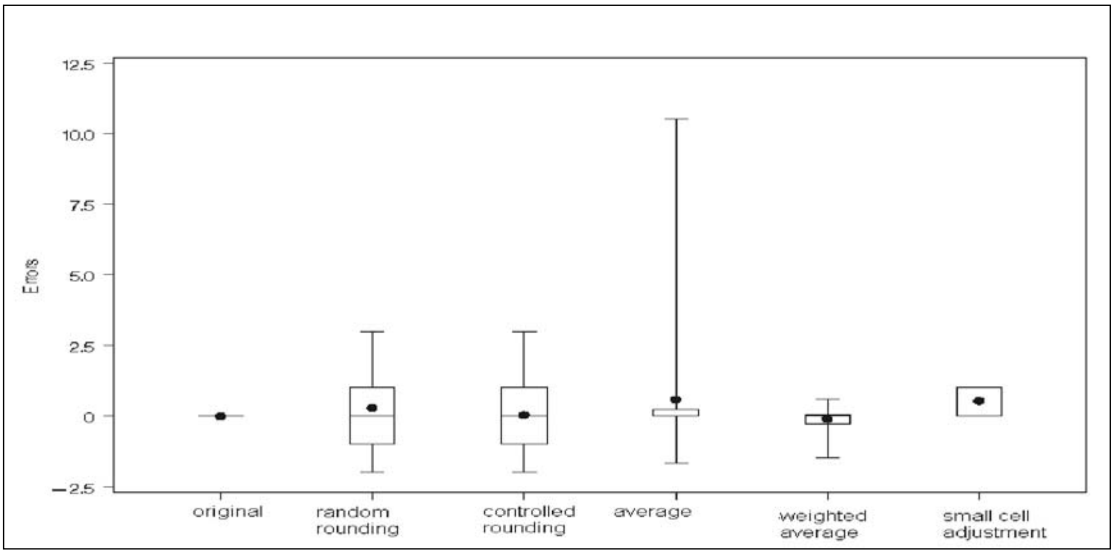
		SCA	RaRo	CR(3)	S - A	S - WA
Tenure	**HD**	1.97 (±0.16)	2.07 (±0.15)	1.78 (±0.15)	0.79 (±0.30)	1.20 (±0.14)
	RAD	10.79 (±1.55)	14.27 (±1.77)	10.64 (±1.29)	7.04 (±3.83)	8.36 (±1.57)
	AAD	0.16 (±0.03)	0.70 (±0.08)	0.52 (±0.06)	0.16 (±0.14)	0.15 (±0.05)
Ethnicity	**HD**	0.55 (±0.13)	0.72 (±0.12)	0.59 (±0.10)	0.79 (±0.89)	0.34 (±0.20)
	RAD	1.59 (±0.0.47)	2.38 (±0.59)	1.98 (±0.47)	12.06 (±20.76)	1.42 (±0.57)
	AAD	0.12 (±0.04)	0.69 (±0.08)	0.56 (±0.06)	3.13 (±5.80)	0.20 (±0.12)
		SCA	**RaRo**	**CR(3)**	**S - A**	**S - WA**
Employment	**HD**	0.93 (±0.15)	1.09 (±0.12)	0.93 (±0.10)	0.61 (±0.43)	0.45 (±0.11)
	RAD	3.24 (±0.65)	4.28 (±0.64)	3.56 (±0.52)	5.58 (±0.56)	1.87 (±0.49)
	AAD	0.12 (±0.02)	0.75 (±0.07)	0.59 (±0.06)	0.47 (±0.66)	0.14 (±0.09)

HD is based on information theory and less intuitive than the other distance metrics. From Table 2, we see that HD doesn't pick up differences between small cell adjustments (SCA) and full rounding (RaRo and CR(3)) as the other measures do. This is because HD is more influenced by small cells than the other metrics. For the Tenure Table and Ethnicity Table, the distance metrics show consistency with respect to the order of the information loss according to the SDC methods. The Employment Table has a slightly mixed order of information loss for the SDC methods depending on the distance metric. This shows that several metrics should be used when assessing bias due to SDC methods and that the impact on the distance metrics are driven by the characteristics of the table.

The minimum distance metric for the Employment Table and Ethnicity Table is obtained by cell suppression with imputed weighted averages (S-WA). This method is nearest to obtaining the original table since there is less error when imputing for suppressed small and large cells. For the Tenure Table, the minimum distance metric is suppression with simple averages (S-A). This is because of the uniformity of the table. The maximum distance metric for the Employment Table and the Tenure Table is random rounding (RaRo). For the Ethnicity Table, the maximum distance metric is cell suppression with imputed simple averages (S-A) because of the fact that the table is highly skewed with one very large column. When a cell in this column is suppressed, the simple average imputation does not take into account the differential cell sizes. Note that controlled rounding (CR(3)) is always better than the random rounding (RaRo), and small cell adjustments (SCA) is on the whole doing better than both.

Users typically want to aggregate tables of lower level geographies in order to obtain statistics for non-standard higher level geographies. The lower level tables however have many small cells and are therefore greatly perturbed because of the SDC methods applied. This leads to more information loss when aggregating sub-totals. In order to evaluate the range of the perturbations for sub-totals of specific target variables obtained by aggregating lower level geographies, we use the statistical graphing tool of a box plot on the differences between the perturbed sub-total and the original sub-total (*PA)*. For unbiased SDC methods, we expect the average and median to be centered around zero. The length of the box and the length of the whiskers gives an indication of how widespread the perturbed totals are from the original totals. Figure 1 presents the box plot of the differences between the original and perturbed sub-totals (*PA's)* for the number of unemployed females with long term illness after aggregating the variable for three consecutive wards across the Estimation Area.

Figure 1. Box Plot of PA's for the Number of Unemployed Females with Long Term Illness in Three Consecutive Wards Average Original Total in combined 3 wards = 14.4



Focusing on one target variable, we see the impact of the different SDC methods on the sub-totals. Full rounding (RaRo and CR(3)) have the same effects with respect to the differences in the perturbed and original sub-totals. Suppression with weighted averages (S-WA) has less information loss than simple averages (S-A). Note that the *PA*'s for S-A can differ by 76% of the average original total in three consecutive wards. The small cell adjustments (SCA) result in less information loss than the other rounding procedures since only small cells are affected.

4.2. Impact on Variance of Estimates

SDC methods will have an impact on the variances that are calculated for estimates based on the frequency tables. We first examine the variance of the cell counts across the geographies (wards) before and after the SDC methods as follows: For each ward *k*, we calculate: $V(D_{orig}^k) = \frac{1}{|k|-1}\sum_{c\in k}(D_{orig}^k(c) - \overline{D}_{orig}^k)^2$ where $\overline{D}_{orig}^k = \frac{\sum_{c\in k} D_{orig}^k(c)}{|k|}$ and $|k| = \sum_c I(c \in k)$ the number of cells in the k^{th} ward. Next we take the average of the variance across all the wards: $V(D_{orig}) = \frac{1}{|W|}\sum_{k=1}^{|W|} V(D_{orig}^k)$. This is repeated for the perturbed table. The information loss measure is: $VR(D_{pert}, D_{orig}) = 100 \times \frac{V(D_{pert}) - V(D_{orig})}{V(D_{orig})}$. Table 3 presents results of the measure VR for the different SDC methods on the three Census tables after removing outlying wards that had very small cells.

Table 3. Percent Relative Difference of Average Variance of Cell Counts (VR) between Original and Perturbed

VR	SCA	RaRo	CR(3)	S - A	S- WA
Tenure	0.003%	0.009%	0.006%	-1.278%	-0.179%
Ethnicity	0.003%	- 0.160%	-0.168%	-2.298%	-0.069%
Employment	0.006%	0.003%	0.138%	-0.266%	-0.111%

For all tables, cell suppression with imputed simple averages (S-A) and weighted averages (S-WA) result in smaller overall variance compared to the original tables. This indicates that these SDC methods, especially the S-A method, are producing more uniform cell counts. The stochastic methods of rounding (SCA and RaRo) have little impact on the cell counts for the Tenure and Employment Tables. The Ethnicity Table, which has one large column and very many sparse columns, have more uniform small cells based on full rounding procedures (RaRo and CR(3)), and therefore a smaller overall variance is obtained.

Another variance that we will focus on is the "between" variance used in regression (ANOVA) analysis for a specific target variable. A typical statistical analysis would be to carry out a regression analysis and model a target variable based on a set of explanatory variables (geography, sex, age, etc.). For a regression analysis the goodness of fit criterion is expressed by the measure R^2. This measure is based on a decomposition of the variance of the target variable. For categorical explanatory variables, the total sum of squares SST can be broken down into two components: the "within" sum of squares SSW which measures the variance of the target variable within the groupings defined by the combination of the explanatory variables and the "between" sum of squares SSB which measures the variance of the target variable 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 proportions within 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: Let $P^k_{orig}(c)$ be a target proportion for a cell c in ward k, i.e. $P^k_{orig}(c) = \frac{D^k_{orig}(c)}{\sum_{c \in k} D^k_{orig}(c)}$ and let $P_{orig} = \frac{\sum_{k=1}^{|W|} D^k_{orig}(c)}{\sum_{k=1}^{|W|} \sum_{c \in k} D^k_{orig}(c)}$ be the overall proportion. The "between" variance is defined as: $BV(P_{orig}) = \frac{1}{|W|-1} \sum_{k=1}^{|W|} (P^k_{orig}(c) - P_{orig})^2$ and the information loss measure is: $BVR(P_{pert}, P_{orig}) = \frac{BV(P_{pert})}{BV(P_{orig})}$.

The target variables in this example are the proportion of full time and part time employed males and females with no long term illness out of the total number of employed persons and the explanatory variable defining the groupings are the wards as obtained in the Employment Table.Table 4 presents the results of the measure *BVR* for the different SDC methods. The overall proportion out of the total in the table is in parentheses.

Table 4. Percent Difference of "Between" Variance (BVR) for the Proportion of Full and Part-Time Employed Males and Females With No Long-Term Illness Within Groupings Defined by Wards

BVR	**SCA**	**RaRo**	**CR(3)**	**S – A**	**S- WA**
Part Time Males NLTI (1.9%)	0.47	6.12	3.52	1.96	0.37
Full Time Males NLTI (31.2%)	0.99	1.61	3.07	0.90	1.42
Part Time Females NLTI (11.1%)	1.01	3.62	1.01	1.11	0.51
Full Time Females NLTI (15.7%)	0.88	1.98	0.58	0.46	0.46

This information loss measure is showing mixed results, sometimes showing more homogenizing of the target proportions between the wards (BVR less than one) and sometimes showing less. It appears that the full rounding procedures (RaRo and CR(3)) have larger "between" variances and more differences in the proportions across the wards. It's interesting to note that cell suppression with imputed averages (simple or weighted) has conflicting effects on the "between" variance of the target proportions. The small cells that are modified due to the SDC methods in particular have an impact on the proportion of the target variable in the ward, since a small number adjusted down or adjusted up can produce either a proportion of 0 or even a proportion of 1 for some wards. Therefore, the effects on the "between" variance are heavily influenced by the way the small cells are perturbed. The small cell adjustments seem to have the least impact on the "between" variances. This information loss measure therefore is difficult to interpret since no consistent pattern emerges and it seems to be driven by the realization of the SDC methods on the target proportions. Future work will investigate this information loss measure further and ways of improving it.

4.3. Impact on Measures of Association

Another statistical analysis that is frequently carried out on contingency tables are tests for independence between categorical variables that span the table. The test for independence for a two-way table is based on a Pearson Chi-Squared Statistic $\chi^2 = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$ where o_{ij} is the observed count and $e_{ij} = \frac{n_{i.} \times n_{.j}}{n}$ is the expected count for row *i* and column *j*. If the row and column are independent then χ^2 has an asymptotic chi-square distribution with *(R-1)(C-1)*and for large values the test rejects the null hypothesis in favor of the alternative hypothesis of association. We use the measure of association, Cramer's V: $CV = \sqrt{\frac{\chi^2 / n}{\min(R-1),(C-1)}}$ and define information loss by the percent relative difference between the original and perturbed table: $RCV(D_{pert}, D_{orig}) = 100 \times \frac{CV(D_{pert}) - CV(D_{orig})}{CV(D_{orig})}$.

Table 5 presents the information loss measures RCV for the three Census tables. We produced two way tables from each of the Census tables where the rows are the wards cross classified with the demographic variables and the columns the cross classified target variables. For example, for the Employment Table, the rows are ward*sex and the columns are economic activity* long term illness.

Table 5. Percent Relative Difference in Cramer's V (RCV) Between Perturbed and Original Two-way Tables

RCV	SCA	RaRo	CR(3)	S - A	S- WA
Tenure	0.26%	0.29%	0.27%	0.20%	-0.13%
Ethnicity	0.11%	0.11%	0.00%	48.27%	-0.33%
Employment	0.10%	0.13%	0.06%	2.36%	-0.09%

All methods except for suppressed cells with imputed weighted averages (S-WA) indicate that the perturbed tables have artificially more association than the original table. The skewed Ethnicity Table is particularly affected when imputing simple averages for the suppressed cells (S-A) as seen in Table 5 since if a large cell is secondary suppressed along with a very small cell, they are both replaced with the simple average resulting in a distribution that is more "flat". This apparently raises the level of association with the geography variable in the table. The weighted averages (S-WA) however seem more consistent with the true values and there is a slight loss of association.

4.4. Statistical Hypothesis Tests for Bias

We first carry out an exact Binomial Hypothesis Test to check if the realization of the random rounding procedures on the tables followed the Binomial rounding scheme. The null hypothesis is: $H_0 : p = 2/3$. The realized proportions and p-values are presented in Table 6 where small p-values means that we reject the null hypothesis and the random rounding procedure was biased. Based on the results, we see a slight bias in the Tenure Table with respect to the small cell adjustments.

Table 6. Exact Binomial Test for Random Rounding Procedures

	Test for Ones		Tests for Twos	
	Proportion	**p-value**	**Proportion**	**p-value**
Tenure				
SCA	0.707	0.0403	0.628	0.0758
RaRo	0.663	0.3756	0.655	0.1756
Employment				
SCA	0.705	0.1851	0.673	0.4449
RaRo	0.685	0.1535	0.655	0.2596
Ethnicity				
SCA	0.677	0.4306	0.692	0.3670
RaRo	0.701	0.0997	0.656	0.3501

For the other SDC methods, we can use a Wilcoxen Signed Rank Test to check whether the location of the empirical distribution has changed. The null hypothesis for the test is no change. The standardized statistic is based on ranking the cells in the table and testing whether the sum of the ranking scores for the original cells deviates from the expected average under the null hypothesis of equal location. If there is a large deviation (small p-value), then one can say that the location of the distribution has been shifted. Table 7 presents p-values for the Wilcoxon Signed Rank Test.

Table 7. p-Values for Wilcoxon Signed Rank Test for Same Location

	Wilcoxen Sign Rank Test p- values		
	S-A	**S-WA**	**CR(3)**
Tenure	<0.001	0.0221	0.0017
Ethnicity	0.2166	0.3888	0.9383
Employment	0.0184	0.9559	0.9883

The Tenure Table is showing significant p-values and we reject the null hypothesis of same location. Since the table is more uniform, it appears that the SDC methods have a larger impact on the distribution of the cell counts. The other tables are not significant except for the cell suppression with imputed simple averages on the Employment Table.

4.5. Impact on statistical analysis

We previously examined the impact of the perturbation schemes on regression analysis through the "between" variance. Another statistical tool for inferences is the Spearman's Rank Correlation. This is a technique that tests the direction and strength of the relationship between two variables. The statistic is based on ranking both sets of data from the highest to the lowest. Therefore, one important assessment of the impact of the perturbation of statistical data is whether we are distorting the rankings of the variables. In the following example, we take target variables that are particularly sparse and therefore are subject to much perturbation: Male and Female students with long term illness (N=544 and N=380, respectively). We sort the original cell counts across wards according to their size and define deciles (10 equal groupings) v^{orig}. This is repeated for the perturbed cell counts which are sorted across wards according to their size and the original order to maintain consistency for the tied variables. Deciles v^{pert} are then defined for the perturbed variable after the sort. The information loss measure is the percent of wards that have changed deciles: $RC = \frac{100 \times \sum_{k=1}^{|W|} I(v_k^{orig} \neq v_k^{pert})}{|W|}$ where I is the indicator function and is 1 if the statement is true and 0 otherwise, and $|W|$ is the number of wards.

Table 8 presents results of the percentage of deciles that have changed due to the perturbation method. Because of the sparseness of the target variables (70% of the cells in the tables take values less than 4), many cells were suppressed or small cells rounded which distorted the rankings of the cell counts. The imputation methods for the cell suppressions (S-A and S-WA) in particular distorted the rankings. An interesting result shown for these target variables is that the controlled rounding causes less distortion to the rankings than the other methods, including random rounding.

Table 8. Percent Changes in Deciles for Male and Female Students with Long Term Illness (N=544)

	SCA	RaRo	CR(3)	S-A	S-WA
Male Students with LTI	10.0%	25.7%	0%	35.7%	20.0%
Female Students with LTI	10.0%	5.7%	2.9%	20.0%	18.6%

Another statistical analysis frequently carried out on contingency tables is log linear modeling. For a 2-way table this narrows down to a test for independence and the Cramer's V statistic. For more variables in a contingency table, one can examine conditional dependencies and calculate expected cell frequencies based on the theory of log-linear modeling. The goodness of fit test for assessing the best fitting parsimonious model is the deviance or likelihood ratio L^2. This is the statistic that is minimized when calculating the maximum likelihood estimates of the parameters of the model. The information loss measure will be based on the ratio of the deviance between the perturbed table and the original table for a given model: $\boldsymbol{LR} = \frac{L^2_{pert}}{L^2_{orig}}$.

Table 9 presents the *LR* measure for the table: Economic Activity (9) * Sex (2) * Long-Term Illness (2) * Ward. The model that is compared is:

$$\log(N_{ijkl}) = \mu + \lambda_i^{Econ} + \lambda_j^{Sex} + \lambda_k^{LTI} + \lambda_l^{Ward} + \lambda_{ij}^{Econ*Sex} + \lambda_{ik}^{Econ*LTI} + \lambda_{il}^{Econ*Ward}$$

Table 9. Ratio of L^2 Statistic Between Perturbed and Original Table of Economic Activity* Sex*Long Term Illness*Ward

LR	Original	SCA	RaRo	CR(3)	S-A	S-WA
Deviance	4,486	5,283	5,316	5,214	6,404	4,744
Ratio	1.00	1.09	1.10	1.08	1.32	0.98

From Table 9, we see that the S-A method increased the deviance by 32%. It is likely that a different model would have been chosen based on the perturbed table as compared to the original table. Future work for this information loss measure will take a more in depth analysis on the impact of choosing different minimal sufficient statistics for original and perturbed tables.

5. Discussion

From this analysis, it is clear that the impact of the SDC methods with respect to information loss depends on the type of table and some general guidelines have emerged:

- A table that has only one or two columns of small values and the remaining columns with large values should not be suppressed since inevitably the secondary suppressions will involve some of the larger cells. A rounding procedure would be preferred.
- A table that is uniform has less information loss regardless of the SDC method so choose a method that causes the least changes to the table.
- A sparse table must have controlled totals so control round if possible, or apply semi-controlled random rounding.

Besides the characteristics of the table, the information loss measures perform differently depending on the outcome of the stochastic processes of the SDC methods. A more robust approach needs to be developed for assessing information loss.

The SDC methods should be tailored to the specific type of table. However, in a large Census context, one (or a combination) of SDC methods are usually applied across all tabular outputs regardless of the tables and the needs of the users. For example, small cell adjustments were implemented for all 2001 Census tabular outputs for England and Wales. This method had little impact on standard tables but had a large negative impact on the very large and sparse origin – destination tables. For the NeSS website which has localized (and less linked) tables, each type of tabular output should have an appropriate SDC method which minimizes the information loss of the data. In the future, we can envision on-line SDC methods tailored according to the users input as to the type of analysis that will be carried out and the variables of interest.

Future work will examine more closely the relationship between the information loss measures and the characteristics of the table and developing a set of guidelines on best practices for designing and protecting frequency tables. We aim to deliver a software tool for suppliers of NeSS tables in order to assess information loss prior to disseminating the tables over the internet. In addition, we need to develop information loss measures with respect to the users of the data and provide more guidance on analyzing perturbed tables, such as the effects of the SDC methods on statistical inferences, how to cope with tables having the same population base with different totals and sub-totals, and how to take into account suppressed cells.

6. Acknowledgements

We wish to thank Stephen Bond of the ONS who initiated this project and developed ideas for the imputation methods for suppressed cells and other related work.

References

Geissing, S. (2003), Coordination of Cell Suppressions: Strategies for Use of GHMITER, Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Luxembourg, April 2003 www.unece.org/stats/documents/ 2003/04/confidentiality/wp.36.e.pdf

Gomatam, S. and A. Karr (2003), Distortion Measures for Categorical Data Swapping, Technical Report Number 131, *National Institute of Statistical Sciences.*

Hundepool, A., et. al. (2003) Argus Version 3.1 User's Manual, http://neon.vb.cbs.nl/casc/

Protecting tables with Cell Perturbation

Juan-José Salazar-González
DEIOC, University of La Laguna, Tenerife, Spain. (jjsalaza@ull.es)

Abstract. This paper presents a new methodology to protect sensitive information in tabular data. It is named Cell Perturbation and can be modelled as a linear programming problem suitable to be solved through a cutting-plane approach. The solutions will satisfy all the protection level requirements, as in other classical approaches like Cell Suppression or Controlled Rounding. Additionally optimal solutions of Cell Perturbation will have smaller loss of information. The paper concludes with computational results on benchmark instances publicly availably for comparisons with other approaches.

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. By law, the agencies are obliged to preserve the confidentiality of data pertaining to individual entities such as persons or businesses. There are various methodologies to preserve confidentiality. We refer the reader to Willenborg and de Waal [13] for a wider introduction to this area, called *Statistical Disclosure Limitation*.

In this area, experts typically distinguish two different problems. The *primary problem* concerns the problem of identifying the sensitive data, i.e., the cell values corresponding to private information that cannot be released within a prescribed exactitude. In this problem also the set of potential attackers and their a-priori knowledge must be identify. The *secondary problem* (also named the *complementary problem*) consists in applying methods to guarantee protection requirements on each sensitive cell against each attacker, while minimizing the overall loss of information. This paper concerns only the secondary problem. The most popular methodologies for solving the secondary problem are variants of the well-known Cell Suppression and Controlled Rounding methods. These two fundamental methodologies will be described next. In practice, some implementations cannot inherently guarantee the protection requirements and great computational effort must be applied to check the proposed output before publication. This checking is called the *Disclosure Auditing* phase and basically consists in computing lower and upper bounds on the original value for each sensitive cell; in the literature there are several techniques to perform this third phase, including linear and integer programming, the Frechet and Bonferroni bounds, and the Buzzigoli and Giusti's shuttle algorithm (see, e.g., Duncan, Fienberg, Krishnan, Padman and Roehrig [5] for references).

Cell Suppression is a methodology that allows the practitioner to do not publish the values in some cells while publishing the original values of the others. In particular, once the primary problem was solved, the cells containing sensitive information must be also not published and they are the *primary suppressions*. Due to the existence of the total marginals in a table, other cells must be also unpublished to guarantee protection of the values under the primary cells, leading to the *secondary suppressions*. They must be identified by solving the so-called *Cell Suppression Problem*, which is a very interesting combinatorial problem widely addressed in the literature. Apart from satisfying the protection requirements, the output of the problem must have a minimum loss of information, which for this methodology could be considered as the sum of the unpublished cell values. See, e.g., [13] for more details on this methodology.

Controlled Rounding is an alternative classical methodology that has not been extensively analyzed in the literature. When applying a rounding procedure the experts are given a base number and they are allowed to modify the original value of each cell by rounding it up or down to a near multiple of the base number. An output pattern must be associated with the minimum loss of information, which for this methodology can be considered as the distance between the original and the modified tables.

In the *Random Rounding* version the experts decide to round up or down each cell by considering a probability that depends on its original cell value, without taking care of the marginal cell values. Therefore, the Random Rounding produces output tables where the marginal values are not the sum of their internal cells, which is a disadvantage of this rounding version. Another version is the so-called *Controlled Rounding*, where probabilities are not considered and the expert should round up or down all cell values such that all the equations in the table hold in the published table. In the so-called *zero-restricted Controlled Rounding* the original values which are already multiple of the base number cannot be modified. Even not considering protection level requirements, a Controlled Rounding solution may not exist for a given table (e.g., Causey, Cox and Ernst [1] showed a simple infeasible 3-dimensional instance). Kelly, Golden and Assad [9] proposed a branch-and-bound procedure for the case of 3-dimensional tables, and Fischetti and Salazar [7] extended this procedure to 4-dimensional tables. Heuristic methods for finding solutions of this problem on multi-dimensional tables have been proposed by several authors, including Kelly, Golden and Assad [9,10]. The problem was first introduced in a statistical context by Cox and Ernst [2].

Salazar [11] presents a common framework to apply Controlled Rounding and Cell Suppression. In addition, two other closely related techniques are described. One tecnique is named *Interval publication*, and it is (in a sense) the linear programming variant of the Cell Suppression method. More details and computational experiments are presented in Fischetti and Salazar [8]. The other technique is named *Cell Perturbation*, which similarly can be seen as the linear programming variant of the Controlled Rounding method. We present in this paper more details and computational results.

Section 2 introduces the main concepts of the Statistical Disclosure Limitation problem. These concepts are fundamental for comparing Cell Perturbation with other similar approaches, like for example the *Controlled Tabular Adjustment* introduced by Cox and Dandekar (see, e.g., Cox, Kelly and Patil [3]). Section 3 describes the *Cell Perturbation* method, with a cutting-plane procedure to find the optimal solution which runs in polynomial time. Results from computational experiments using the proposed methods are analyzed in Section 4.

2. Basic concepts and notation

A statistical agency is provided with a set of n values a_i for $i \in I := \{1,...,n\}$.

Vector $a = [a_i : i \in I]$ is known as "nominal table" and satisfies a set of m equations $\sum_{i \in I} m_{ji} y_i = b_j$ for $j \in J := \{1,...,m\}$. For convenience of notation the linear system will be denoted by $My = b$, thus $Ma = b$ holds. Each solution y of $My = b$ is called *congruent table*.

Statistical tables typically contain sensitive data. We denote the subset of sensitive cells by P. In a general situation, all the sensitive cells in a table must be protected against a set K of *attackers*. The attackers are the intruders or data snoopers that will analyze the final product data and will try to disclose confidential information. They can also be coalitions of respondents who collude and behave as single intruders. The aim of the Disclosure Limitation Methods is to reduce the risk of them succeeding. Each attacker knows the set of linear system $My = b$ plus extra information that bound each cell value. For example, the simplest attacker is the so-called *external intruder* knowing only that unknown cell values are, say, nonnegative. Other more accurate attackers know tighter bounds on the cell values, and they are called *internal attackers*. In general, attacker k is associated with two bounds lb_i^k and ub_i^k such that $a_i \in [lb_i^k ... ub_i^k]$ for each cell $i \in I$. The literature on statistical disclosure control (see, for example, Willenborg and de Waal [13]) typically addresses the situation where $|K| = 1$, thus protecting the table against the external intruder with only the knowledge of the linear system and some external bounds; nevertheless this is a simplification of the real problem in Disclosure Limitation and statistical offices are interested in protecting tables against several intruders.

To protect the sensitive cell p containing value a_p in the input table, the statistical office is interested in publishing an output containing several congruent tables, including not only the original nominal table but also others so that no attacker can disclose the private information a_p (neither a narrow approximation). The output of a Disclosure Limitation Method is generally called a *pattern*, and it can assume a particular structure depending on the methodology considered.

The congruent tables associated to a pattern must differ so that each attacker analyzing the pattern will not compute the original value of a sensitive cell within a narrow approximation. For each potential intruder, the idea is to define a protection range for p and to demand that the a posteriori protection be such that any value in the range is potentially the correct cell value. To be more precise, by observing the published pattern, attacker k will compute an interval $[\underline{y}_p^k ... \overline{y}_p^k]$ of possible values for each sensitive cell p. The pattern will be considered *valid* to protect cell p against attacker k if the computed interval is "wide enough". To set up the definition of "wide enough" in a precise way, the statistical office gives three input parameters for each attacker k and each sensitive cell p with nominal value a_p:

- Upper Protection Level: it is a number UPL_p^k representing a desired lower bound for $\overline{y}_p^k - a_p$;
- Lower Protection Level: it is a number LPL_p^k representing a desired lower bound for $a_p - \underline{y}_p^k$;
- Sliding Protection Level: it is a number SPL_p^k representing a desired lower bound for $\overline{y}_p^k - \underline{y}_p^k$.

The values of these parameters can be defined by using common-sense rules. In all cases, the protection levels are assumed to be unknown to the attackers. An elementary assumption is that

$$lb_p^k \leq a_p - LPL_p^k \leq a_p \leq a_p + UPL_p^k \leq ub_p^k$$

and

$$ub_p^k - lb_p^k \geq SPL_p^k,$$

for each attacker k and each sensitive cell p. For notational convenience, let us also define

$$lpl_p^k := a_p - LPL_p^k, \quad upl_p^k := a_p + UPL_p^k, \quad LB_i^k := a_i - lb_i^k, \quad UB_i^k := ub_i^k - a_i.$$

Given a pattern, the mathematical problems of computing values $\underline{y}_p^k$ and $\overline{y}_p^k$ are known as *attacker problems* for cell p and attacker k. The overall problem of solving the attacker problems for all cells is called *Disclosure Auditing Problem*. This should not be confused with the Disclosure Auditing Phase mentioned in Section 1 and which is an unnecessary phase for the methodologies proposed in this paper since they will implicitly guarantee the protection requirements on the output pattern.

Finally, among all possible valid patterns, the statistical office is interested in finding one with minimum information loss. The *information loss* of a pattern is intended to be a measure of the number of congruent tables in the pattern. A valid pattern must always allow the nominal table to be a feasible congruent table, but it must also contain other different congruent tables so as to keep the risk of disclosure controlled. In practice, since it is not always easy to count the number of congruent tables in a pattern from the point of view of an intruder k, the loss of information of a pattern is replaced by the sum of the loss of information of its cells. In this case, the individual cost for cell p is generally proportional to the difference between the worse-case situations (i.e., to $\overline{y}_p^k - \underline{y}_p^k$), it is proportional to the number of respondents contributing to the cell value a_p, or it is simply a positive fixed cost when a_p is not published (i.e., when $\overline{y}_p^k - \underline{y}_p^k > 0$).

It is very important to observe that these concepts do not always coincide with the one used in other articles in the literature. This observation is fundamental to compare the methodology introduced in this paper with the methodology introduced by other authors. For example, in the Controlled Tabular Adjustment described in Cox, Kelly and Patil [3] the concept of "protected output" is different. In our framework an output is protected if, for each sensitive cell and each value in its protected range, an attacker must deduce the existence of a congruent table assuming this value in this cell. When considering different attackers, this congruent table may not be the same for all the attackers. Also when considering one attacker, different sensitive cells and different values may show different tables. In the framework used in [3] an output is protected if there is a congruent table valid which satisfy *one* of the two protection levels for *all* the sensitive cells. Of course, the reader should not understand from this words that the basic concepts used in [3] is wrong, but only different than the concepts for which Cell Perturbation has been proposed.

3. Cell Perturbation Methodology

The main disadvantage of the Controlled Rounding methodology is that a protected pattern does not always exist due to the tight requirement of rounding each cell value either down or up. A wat of ensuring the existence of protected patterns is to relax this requirement in the Controlled Rounding model and to look for a congruent table $v = [v_i : i \in I]$ such that

$$v_i \in [\lfloor a_i \rfloor \ldots \lceil a_i \rceil]. \tag{1}$$

where $\lfloor a_i \rfloor$ and $\lceil a_i \rceil$ are given in advance from the statistical office such that $\lfloor a_i \rfloor \leq a_i \leq \lceil a_i \rceil$. These extreme values can be defined as the nearest numbers to a_i which are multiples of a given number (i.e., defined as in the standard Controlled Rounding methodology from a given base number), but they can also be the two values within a given difference with respect to a_i (i.e., $\lfloor a_i \rfloor := a_i - t_i$ and $\lceil a_i \rceil := a_i + t_i$ for a given base number $t_i > 0$). Table v is then a pattern in the *Cell Perturbation methodology* and the novelty with respect to the Controlled Rounding is that now v_i can be any value between the two extremes of the interval $[\lfloor a_i \rfloor \ldots \lceil a_i \rceil]$. As in the Controlled Rounding methodology, the loss of information of a cell i could be defined to be proportional to $|v_i - a_i|$, and the "loss of information" of a pattern is the sum of the loss of information of all the cells.

Obviously, if the requirement of rounding up or down is removed for all the cells, and no new one is added to the continuous relaxation of a model minimizing the non-linear function $\sum_{i \in I} |v_i - a_i|$, then the valid pattern with minimum loss of information is the nominal table a. A way to avoid this disappointing solution is to keep some requirements (for example, concerning the sensitive cells) or simply require that the published values in each sensitive cell must be equal to some given values (for example, $v_p = \lceil a_i \rceil$ for all $p \in P$). Still these additional constraints may lead to infeasible problems. Practitioners in statistical offices prefer another way of avoiding the nominal table as published table: it consists in defining a different objective function. Indeed, by considering the objective as the distance between each published value v_i and the value in $\{\lfloor a_i \rfloor, \lceil a_i \rceil\}$ closest to a_i we get the same criteria used in the classical Controlled Rounding methodology, and allow the objective function to be linear on the variables x_i.

Let $r_i := \lceil a_i \rceil - \lfloor a_i \rfloor$ a (possibly) known information for attackers. Then the attacker problems associated with attacker k are now exactly the same as in the Controlled Rounding methodology, i.e.

$$\begin{aligned} My &= b & & \\ v_i - r_i \le y_i &\le v_i + r_i & &\text{for all } i \in I \\ lb_i^k \le y_i &\le ub_i^k & &\text{for all } i \in I. \end{aligned}$$

As in the Controlled Rounding methodology, a necessary (but not sufficient) condition for feasibility is that $\max_{k \in K}\{SPL_i^k, UPL_i^k + LPL_i^k\} \le 2r_i$ for all $i \in I$.

In the literature there are several methodologies to protect tables by data perturbation (see, for example, Evans, Zayatz and Slanta [6]) but, as far as we know, they all concern the direct modification of the microdata and, therefore, there is less control on the final protection interval of each cell in the published pattern.

To write a first model for the Cell Perturbation model, it is convenient to introduce two continuous variables z_i^- and z_i^+ for each cell i, with the following meaning:

$$z_i^- := \max\{0, a_i - v_i\} \quad \text{and} \quad z_i^+ := \max\{0, v_i - a_i\}.$$

Note that $v_i = a_i + z_i^+ - z_i^-$. Let w_i^- be the given cost for each unit of z_i^-, and w_i^+ be the given cost for each unit of z_i^+. Hence the objective function is

$$\sum_{i \in I} w_i^+ z_i^+ + w_i^- z_i^-$$

as in the Controlled Rounding methodology. One way to write the protection level requirements is to introduce additional variables f^{kp} and g^{kp} for each attacker k and each sensitive cell p.

It is again possible to avoid the explicit introduction of the auxiliary variables f^{kp} and g^{kp} ($k \in K$ and $p \in I$) along with the associated linking constraints, by using the standard LP Duality Theory. See Salazar [11] for a full technical description of two LP models for Cell Perturbation. Briefly, a first model is a compact formulation using a large (but still polynomial) number of variables. More precisely, the first model is a linear program using the auxiliary variables f^{kp} and g^{kp}. The second model replaces these variables by an exponential number (but polynomially separable) of linear inequalities. Although the two model are in equivalent in theory, in practice the second one is preferred. The reason is because it works with an small number of variables (z_i^+ and z_i^-, two for each cell), while the inequalities are generated on-the-fly only when required through an iterative procedure. In practice the number of iterations is small. Section 4 empirically supports this claim. What is more, under some hypothesis on the magnitude of the protection levels, the number of linear inequalities in this second model can be strongly reduced. An example is when the table is a frequency table, and the sensitive cells, the external bounds and the protection levels are set in accordance with the criteria proposed in [12]. This is a situation produced when protecting a frequency table with τ -ARGUS. In other words, a cell i with $\lfloor a_i \rfloor < \lceil a_i \rceil$ in a table generated by τ -ARGUS always satisfies

$$lb_i \le \lfloor a_i \rfloor \le lpl_i \quad \text{and} \quad upl_i \le \lceil a_i \rceil \le ub_i$$

because $lb_i = \lfloor a_i \rfloor = lpl_i = 0$ and $upl_i = \lceil a_i \rceil < ub_i$. Under these hypothesis the external bounds and the protection levels are useless, and therefore the rounder called by τ -ARGUS will have the task of finding a fractional solution satisfying all the $|J|$ equations, optimal according to the objective function.

Cell Perturbation has some similarities with the Partial Cell Suppression introduced in Fischetti and Salazar [8]. Both methodologies can be formulated as a Linear Programming (LP) model with an exponential number of constraints that can be efficiently separated in a cutting-plane approach. They are closely related to the LP relaxations of two standard methodologies |Cell Suppression and Controlled Rounding| and they differ in several aspects. For example, an important requirement in the Cell Perturbation is the additivity of the output data, which should be a congruent table. This requirement is not present in Partial Cell Suppression, where the output is a table of intervals. Another requirement in the Cell Perturbation methodology is that each cell value cannot be modified by more than a given base number, which is not an input parameter of the Partial Cell Suppression methodology. The used base numbers, released to the public together with the output data when using Cell Perturbation, have a large impact in the utility of this data. From the practical point of view, it is preferred to use small base numbers subject to the existence of a solution. This consideration does not apply to Partial Cell Suppression.

Table 1. Short description of the benchmark instances

Name	Type	$\|I\|$	$\|J\|$	$\|P\|$	nzeros
bts4	hierarchical from $54 \times 54 \times 4 \times 4$	36570	36310	2260	136912
hier13	hierarchical from $13 \times 13 \times 13$	2020	3313	112	11929
hier16	hierarchical from $16 \times 16 \times 16$	3564	5484	224	19996
nine12	linked from $10 \times 6 \times 6 \times 6 \times 6 \times 6 \times 6 \times 6 \times 6$	10399	11362	1178	52624
nine5d	linked from $4 \times 29 \times 3 \times 4 \times 5 \times 6 \times 5 \times 4 \times 6$	10733	17295	1661	58135
ninenew	linked from $10 \times 6 \times 6 \times 6 \times 6 \times 6 \times 6 \times 6 \times 6$	6546	7340	858	32920
two5in6	linked from $6 \times 4 \times 16 \times 4 \times 4 \times 4$	5681	9629	720	34310

4. Computational Results

We have implemented the cutting-plane algorithm for solving the Cell Perturbation Problem. The implementation has been done in ANSI C using the Microsoft Visual C 6.0 compiler and the branch-and-cut framework of CPLEX 9.0. The experiments have been executed on a PC Pentium IV 2.5 Ghz. under Microsoft Windows XP.

For benchmarking purposes, we have run our codes on a collection of artificial instances close to being realistic. It consists of seven test cases of magnitude data created by Ramesh Dandekar (U.S. Department of Energy), described in Dandekar [4] and available through the webpage http://webpages.ull.es/users/casc. Three of the seven instances (“bts4”, “hier13” and “hier16”) are hierarchical tabulations, while the remaining four are linked tabulations. For each instance of the collection, Table 1 gives the name, the number of cells, the number of equations, the number of sensitive cells, and the number of non-zero elements in M. The protection against one attacker is assumed, i.e. $|K|=1$.

Table 2. Results applying Cell Perturbation with $r_i = 250$

Name	mult	down	frac	up	distance	cost	max	time
bts4	226	11138	14146	11060	1825973.4	2725081.6	247.0	85.8
hier13	10	142	1749	119	33599.2	161975.1	106.6	0.8
hier16	10	255	3050	249	69988.1	291378.0	105.1	7.2
nine12	57	1852	6547	1943	383235.9	804652.1	177.0	164.3
nine5d	73	1736	7221	1703	355980.7	844324.4	204.1	31.3
ninenew	36	1119	4229	1162	224053.2	506292.6	171.0	34.5
two5in6	67	806	3971	837	167860.3	432765.1	141.9	5.8

Table 2 shows the results of applying the Cell Perturbation approach on this collection of data. We considered $r_i = 250$ and $s = 0$. The results are very similar to the ones obtained when solving the LP relaxation of the Controlled Rounding model.

We have also run the algorithms on a benchmark instance provided by Anco Hundepool (CBS), available at http://webpages.ull.es/users/casc. It is a frequency table described by 30886 cells and 39800 equations in a 6-level hierarchical structure. There are 10680 sensitive cells and 120819 non-zero elements in M. An optimal Controlled Rounded solution was found in 279.7 seconds using our computer after exploring 842 nodes, and the optimal objective value is 42600. The solution of the LP relaxation at the end of the root node had 2772 variables with fractional values and the objective value is 42545.5. The root node was solved in 3.6 seconds and had a heuristic solution with objective value 42744. An optimal Cell Perturbation solution was found in 4.2 seconds, with 4121 fractional values and with objective value 42253.2.

Each equation in these instances determines a marginal cell by adding a subset of other cells, thus all the cells of each instance can be considered linked in a hierarchical structure. Since the cell values are assumed to be non-negative ($lb_i^k = 0$), starting from the grand total (the cell with the largest value) one can automatically assign levels of the hierarchical structure to the cells. The grand total cell is assigned to level 0. We have prioritized the cell variables so the branching phase selects a variable associated with a cell with higher level first. This consideration improved the performance of the algorithm in our experiments.

Acknowledgement

Work supported by the European project IST-2000-25069 (CASC) and by the Spanish project TIC2002-00895 (TEIDE). We also thank the Office for National Statistics (contract IT-O3/0763).

References

Causey, B.D., Cox, L.H. and Ernst, L.R. (1985) “Applications of Transportation Theory to Statistical Problems”, *Journal of the American Statistical Association*, **80**, 903–909.

Cox, L. H. and Ernst, L. R. (1982) “Controlled Rounding”, *INFOR*, **20**, 423–432.

Cox, L.H., Kelly, J.P. and Patil, R. (2004) “Balancing Quality and Confidentiality for Multivariate Tabular Data”, 87–98, in *Privacy in Statistical Databases*, Lecture Notes in Computer Science 3050, Springer.

Dandekar, R. A. (2004) “Maximum Utility-Minimum Information Loss Table Server Design for Statistical Disclosure Control of Tabular Data", 121–135 in *Privacy in Statistical Databases*, Lecture Notes in Computer Science 3050, Springer.

Duncan, G. T., Fienberg, S. E., Krishnan, R., Padman, R. and Roehrig, S. F. (2001) "Disclosure Limitation Methods and Information Loss for Tabular Data", 135–166 in Doyle, P., Lane, J., Theeuwes, J. and Zayatz, L. (Eds.) *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, Elsevier Science.

Evans, T., Zayatz, L. and Slanta, J. (1998) "Using Noise for Disclosure Limitation of Establishment Tabular Data", *Journal of Official Statistics*, **14/4**, 537–551.

Fischetti, M. and Salazar-González, J.J. (1998) "Computational Experience with the Controlled Rounding Problem in Statistical Disclosure Control", *Journal of Official Statistics*, **14/4**, 553–565.

Fischetti, M. and Salazar-González, J.J. (2003) "Partial Cell Suppression: a New Methodology for Statistical Disclosure Control", *Statistics and Computing*, **13**, 13–21.

Kelly, J.P., Golden, B.L. and Assad, A.A. (1990) "Using Simulated Annealing to Solve Controlled Rounding Problems", *ORSA Journal on Computing*, **2**, 174–185.

Kelly, J.P., Golden, B.L. and Assad, A.A. (1993) "Large-Scale Controlled Rounding Using TABU Search with Strategic Oscillation", *Annals of Operations Research*, **41**, 69–84.

Salazar-González, J.J. (2005) Controlled Rounding and Cell Perturbation: Statistical Disclosure Limitation Methods for Tabular Data. To appear in *Mathematical Programming*.

Salazar-González, J.J., Lowthian, P., Young, C., Merola, G., Bond, S. and Brown, D. (2004) "Getting the Best Results in Controlled Rounding with the Least Effort", 58–72 in *Privacy in Statistical Databases*, Lecture Notes in Computer Science 3050, Springer.

Willenborg, L.C.R.J. and de Waal, T. (2001) *Elements of Statistical Disclosure Control*. Lecture Notes in Statistics 155, Springer.

A proposed method for confidentialising tabular output to protect against differencing

Bruce Fraser and Janice Wooton

Data Access and Confidentiality Methodology Unit, Australian Bureau of Statistics, Locked Bag 10, Belconnen ACT 2616 Australia, bruce.fraser@abs.gov.au, janice.wooton@abs.gov.au

Abstract: The differencing problem puts increased demands on a system of tabular confidentiality. Methods currently in use at the Australian Bureau of Statistics and many other national statistical offices only target small cell values for treatment, and allow large cells values to be released without any perturbation to protect confidentiality. Such methods are vulnerable to differencing attacks which can derive unprotected small cell values as the difference of two unprotected large cell values. This paper proposes a cell perturbation method for confidentialising Australian population Census tables to protect against differencing attacks and any other attempts at identification. The method is a two stage process. At the first stage a perturbation is added to all cells of all tables, including the independent perturbation of table marginals. The perturbation is set to zero for a pre-determined set of key output (e.g. age by sex population counts). This perturbation process produces a non-additive protected table. At the second stage additivity is restored. Record keys are assigned to the microdata and are used to produce consistent perturbations at the first stage of the process, although consistency is lost when additivity is restored.

1. Introduction

The Australian *Census and Statistics Act, 1905* provides the authority for the Australian Bureau of Statistics (ABS) to collect statistical information, and requires that statistical output shall not be published or disseminated in a manner that is likely to enable the identification of a particular person or organisation. This requirement means that the ABS must take care with any statistical information that relates to very small subpopulations or subsamples.

Output from Australia's 5-yearly Population Census is extensively used for studying small subpopulations in Australia. As a complete enumeration of people in Australia on census night, it is one of the few statistical datasets in Australia that can be used to compile meaningful statistics for small subpopulations. It is therefore important that rigorous procedures and techniques are in place to ensure that Population Census output is released in a manner that is not likely to enable identification of any individual, household or family.

The current technique used to guard against identification or disclosure of confidential information in census tables of counts is random rounding to base 3 for cells with small values. However with the 2006 Population Census this method alone is no longer adequate to protect census tables against disclosure. This is due to a number of factors, in particular the introduction of a new small area geographical unit - the mesh block, and a proposed web-based table-builder product which would allow users to produce tailored tables according to their own specifications. The small cell perturbations produced by the random rounding base 3 method are not effective as the sole protection for ad hoc output where a user can specify tailored tables using fine level geography or fine disaggregations of other variables. A further problem is that it is becoming increasingly difficult to keep track of all the Population Census tables released across different mechanisms, and so difficult to protect against differencing problems by tracking the release of output.

An example of a differencing problem is where a user specifies a table for a user-defined geography, compiled from a number of small area building blocks. For example, the Statistical Local Area (SLA) "Remainder of ACT" had a population of approximately 430 at the time of the 2001 Population Census. The SLA consists of 7 Collection Districts (CDs) with populations of approximately 210, 115, 65, 25, 10 and two with populations of less than 5 each. If a user can specify tables for a tailored geography made up from CD building blocks, then they can specify a table for the full SLA, as well as a table for the amalgam of the six CDs with the greatest populations. Differencing the two tables provides information for a single CD with a population of less than 5 persons. The confidentiality protections applied to the SLA table, and the 6 CDs table, must therefore be sufficient to ensure that no information is disclosed through differencing the two tables. For further details regarding the definition of CD's and SLA's see ABS Cat. no. 1216.0 *Australian Standard Geographical Classification (ASGC)- Electronic publication*, July 2004.

The differencing problem is not limited to geography. User-specified tables could be requested for the whole population of "Remainder of ACT", and for the population aged less than 70 years, or for the population born in English-speaking countries, or for the population who only speak English at home. Differencing would again produce results in respect of small subpopulations.

To solve the differencing problem, we propose an alternative to the current census tabular confidentialisation methodology. The new methodology will be discussed in the following sections and is a cell perturbation technique. This technique has advantages in that it provides protection against disclosure through differencing and disclosure through single contributor cells. Like the current method, the new method will ensure small cells are perturbed to protect against disclosure through single contributor cells. But unlike the existing method a small amount of perturbation is introduced to all cells instead of just small cells. This ensures that when two large cells are differenced to produce a small difference enough perturbation has been introduced so that users cannot have much confidence in the accuracy of the differenced value.

The method has been developed in the context of Population Census tables, and has only been designed to protect tables of non-negative integer counts.

2. Deriving a New Cell Perturbation Method

By perturbing all cells instead of just small cells, we can protect census tables against disclosure through differencing.

Denote the i^{th} cell count of a multi-way table as n_i. For each non-zero n_i an independent perturbation d_i is generated from an integer-value distribution that satisfies the following criteria:

(a) mean of zero;

(b) $d_i \geq - n_i$;

(c) fixed variance V for all i and all n_i; and

(d) $| d_i | \leq c$ for some small positive integer c.

d_i is added to n_i to give n_i*, the cell value for the protected table.

Criterion (a) ensures that the perturbations do not add a bias to the table, criterion (b) ensures that no negative numbers are created as a result of perturbation, criterion (c) ensures that any cell derived by differencing two perturbed cells has a fixed variance, and also that relatively more noise is added to the smallest cells (smallest n_i), and criterion (d) is applied to ensure that no perturbation is ever greater than c in magnitude.

Note that no perturbation is added to zero cells ($n_i = 0$) in order to maintain any structural zeroes. The method we propose also independently perturbs every non-zero cell in a table, in particular this means table margins are perturbed independently from interior cells and so table additivity is lost. An alternative approach would be to perturb interior cells only, and then generate marginals by adding interior cells. However this approach would result in a final perturbation of marginal cells with a high level of variance, and indeed a variance that increases as the number of interior cells increases. In order to prevent large perturbations on the margins we instead perturb them independently of interior cells to create a non-additive table, then restore table additivity as a separate step.

3. Improving consistency by assigning random numbers or record keys to microdata records

A technique for improving consistency between tables is to assign permanent random numbers to each record on the microdata file, and to use these permanent random numbers to generate the random perturbations. In the following we discuss two possible ways of doing this.

The first way is to assign each record a key in the form of a 32-bit binary number (the keys are assigned randomly to each observation on the census microdata file). This record key can be used as a seed for a pseudo-random number generating function, which in turn could be used for generating d_i at record level. Record keys can also be combined across records to guarantee consistent results are applied to aggregates of records. This can be done using the XOR (exclusive or) function. The XOR function will return another 32-bit binary number, and will always return the same result from the input, regardless of the order in which the individual keys are XORed together. This means any aggregate of n records will correspond to a unique 32-bit aggregate key, obtained by XORing the keys of the individual records. The key of the aggregate can be used to seed a pseudo-random number, which in turn can be used to determine the aggregate's perturbed value. This gives the property that whenever the same set of units are together in an interior cell, the perturbed cell value will always be the same.

The second method is to assign an independent random discrete uniform number to each unit in the microdata file. The interval for the discrete uniform random variables will be from 0, 1, 2, ...m-1, where m is a sufficiently large integer value. Let y_i denote the discrete uniform random variable assigned to the j^{th} unit on the microdata file It can be proved that the sum mod m of any combination of the y_i values is also a discrete uniform random variable on the interval 0, 1, 2, ..., m-1. Whenever a set of units are present together in an internal cell, we can combine their discrete uniform random numbers using the mod function described above and use the result to decide on the perturbation to be applied to the cell. This guarantees that whenever the same units contribute to an interior cell, the perturbation will always be the same.

4. Allowing for zero perturbation counts

The assignment of record keys or random numbers to microdata records discussed in section 3 can be modified slightly to allow for a set of predefined counts for which no perturbation error will be introduced. For example, age by sex population counts at a particular level of geography can be defined as zero perturbation counts. (In this example the level of geography chosen must be broad enough to ensure that it is not vulnerable to differencing).

This can be done by assigning record keys or random numbers independently to all but one of the records contributing to each zero perturbation count. The final record in each group is assigned the particular record key or random number that is required to ensure that the aggregation of the record keys or random numbers of all records in the group (through mod m addition, XOR, or some other method of aggregation), gives a final result of zero. The perturbation process is then constrained to ensure that a result of zero will always produce a perturbation of zero.

5. Restoring additivity to perturbed tables

There are a number of ways in which additivity might be restored to the non-additive table that results from the independent perturbation of interior and marginal cells. A key criterion for an additivity algorithm is that it can process large tables quickly. This is necessary in order to be able to offer a responsive web-based table-builder service, whereby a user can specify a tailored table, and have it constructed, confidentialised and delivered to the user via the web interface in a short period of time. A secondary criterion is to only make small adjustments to the table marginals, with larger adjustments in the interior cells, if necessary.

Work is currently in progress to determine a suitable algorithm for restoring additivity to perturbed tables.

6. Balancing information loss and disclosure risk

The Australian Bureau of Statistics is required by law not to release information in a manner that is likely to enable the identification of any respondent. Furthermore, it is essential to retaining the trust of providers that they are confident their personal information will be safeguarded. A tabular confidentiality system therefore must reduce the disclosure risk to a low level. But this should be done in a way that minimises the information loss in the data, and preserves as far as possible the analytical value and integrity of the results.

The method we have outlined has flexibility, primarily in the distribution chosen for the cell perturbation, and to a lesser extent in the algorithm chosen to restore additivity to a table. These characteristics can be varied to change both the amount of protection provided (reduction in disclosure risk) and the amount of damage done to the output (amount and characteristics of information loss). Work is still in progress assessing disclosure risk and information loss resulting from different choices of perturbation distributions. While the intention is to measure information loss and disclosure risk using a number of metrics, there are two key measures being used in the evaluation.

A measure of disclosure risk is the probability that an observed count of 1 corresponds to a true count of 1. It is a simple matter to specify a perturbation distribution that will ensure any cell value of 1 in a table will be perturbed to a value other than 1, however counts of 1 can also be observed through the differencing of two table. Therefore one measure that is used in the evaluation is the probability that an observed difference of 1 equates to a true difference of 1.

The primary measure of information loss is provided by a comparison of the results of a chi-squared test of association on each of the protected and unprotected tables.

References

Australian Bureau of Statistics, (2004). *Australian Standard Geographic Classification (ASGC): Electronic Publication*. Cat. no. 1216.0.

The Controlled Rounding Implementation

*Juan-José Salazar-González**, Christine Bycroft**, Andréa Toniolo Staggemeier***

*** DEIOC, University of La Laguna, Tenerife, Spain. (jjsalaza@ull.es)**

**** Office of National Statistics, U.K.**

Abstract. Rounding methods are common techniques in many statistical offices to protect disclosive information when publishing data in tabular form. Classical versions of these methods do not consider protection levels while searching patterns with minimum information loss, and therefore typically the so-called auditing phase is required to check the protection of the proposed patterns. This paper presents a mathematical model for the whole problem of finding a protected pattern with minimum loss of information, and describe an algorithm to solve it. The base scheme is a branch-and-bound search. If time enough is allowed, the algorithm stops with an optimal solution. Otherwise, an heuristic approach aims at finding a feasible zero-restricted solution. On complicated or infeasible tables where finding a zero-restricted feasible solution cannot be found in a reasonable time, the algorithm generates a non zero-restricted solution, here referred to as *rapid table*. This paper presents a summary of findings from some computational experiments.

1. Basic Concepts and Notation

A statistical agency is provided with a set of n values a_i for $i \in I := \{1,...,n\}$. Vector $a = [a_i : i \in I]$ is known as a "nominal table" and satisfies a set of m equations $\sum_{i \in I} m_{ji} y_i = b_j$ for $j \in J := \{1,...,m\}$. For convenience of notation the linear system will be denoted by $My = b$, thus $Ma = b$ holds. Each solution y of $My = b$ is called a *congruent table* since it is fully additive and coherent with the structure of the original table.

Statistical tables typically contain disclosive data. We denote the subset of disclosive cells by P. In a general situation, all the disclosive cells in a table must be protected against a set K of *attackers*. The attackers are the intruders or data snoopers that will analyze the final product data and will try to disclose confidential information. They can also be coalitions of respondents who collude and behave as single intruders. The aim of the Statistical Disclosure Control is to reduce the risk of them succeeding. Each attacker knows the published linear system $My = b$ plus extra information that bounds each cell value. For example, the simplest attacker is the so-called *external intruder* knowing only that unknown cell values are, say, nonnegative. Other more accurate attackers know tighter bounds on the cell values, and they are called *internal attackers*. In general, attacker k is associated with two bounds lb_p and ub_p such that $a_p \in [lb_p ... ub_p]$ for each cell $p \in I$. The literature on Statistical Disclosure Control typically addresses the situation where $|K| = 1$, thus protecting the table against the external intruder with only the knowledge of the linear system and some external bounds. Our implementation also considers a single-attacker protection.

To protect the disclosive cell p containing value a_p in the input table, the statistical office is interested in publishing a table that is congruent with a collection of several different possible ones. The output of a Statistical Disclosure Control is generally called a *pattern*, and it can assume a particular structure depending on the methodology considered. The congruent tables associated with a pattern must differ so that each attacker analyzing the pattern will not compute the original value of a disclosive cell to within a narrow approximation. For each potential intruder, the idea is to define a protection range for a_p and to demand that the a posteriori protection be such that any value in the range is potentially the correct cell value. To be more precise, by observing the published pattern, attacker k will compute an interval $[\underline{y}_p ... \overline{y}_p]$ of possible values for each disclosive cell p. The pattern will be considered *valid* to protect cell value a_p against attacker k if the computed interval is "wide enough". To set up the definition of "wide enough" in a precise way, the statistical office gives two input parameters for each disclosive cell with nominal value a_p: an Upper Protection Level, which is a number UPL_p representing a desired lower bound for $\overline{y}_p - a_p$; and a Lower Protection Level, which is a number LPL_p representing a desired lower bound for $a_p - \underline{y}_p$. The values of these parameters can be defined by using common-sense rules. In all cases, the protection levels are assumed to be unknown by the attacker. An elementary assumption is that

$$lb_p \le a_p - LPL_p \le a_p \le a_p + UPL_p \le ub_p$$

for each attacker k and each disclosive cell p. For notational convenience, let us also define

$$lpl_p := a_p - LPL_p, \quad upl_p := a_p + UPL_p, \quad LB_p := a_p - lb_p, \quad UB_p := ub_p - a_p.$$

Figure 1. Diagram of parameters

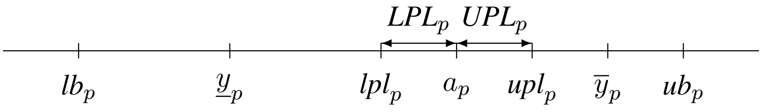


Figure 1 illustrates the position of the parameters in a line. Given a pattern, the mathematical problems of computing values $\underline{y}_p$ and $\overline{y}_p$ are known as *attacker problems* for cell p and attacker k.

An important observation is that we are assuming that our original table contains real numbers. In other words, the value a_i are not necessarily integer. This is a common situation when working with business data. All this documentation is for the general case where a_i are real.

Finally, among all possible valid patterns, the statistical office is interested in finding one with minimum information loss. The *information loss* of a pattern is intended to be a measure of the number of congruent tables in the pattern. A valid pattern must always allow the nominal table to be a feasible congruent table, but it must also contain other different congruent tables so as to keep the risk of disclosure controlled. In practice, since it is not always easy to count the number of congruent tables in a pattern from the point of view of an intruder k, the loss of information of a pattern is replaced by the sum of the loss of information of its cells. In this case, the individual cost for cell p is generally proportional to the difference between the worse-case situations (i.e., to $\overline{y}_p - \underline{y}_p$), it is proportional to the number of respondents contributing to the cell value a_p, or it is simply a positive fixed cost when a_p is not published (i.e., when $\overline{y}_p - \underline{y}_p > 0$).

2. Controlled Rounding Methodology

In *Controlled Rounding Methodology* we are provided with an input base number r_i for each cell i. In practice, the statistical office uses a common base number r_i for all cells, but the method can also be applied when there are different base numbers, as required by some practitioners (e.g., when protecting some hierarchical tables, bigger base numbers are preferred on the top levels than on the low levels). However, in our implementation all base numbers r_i are identical, so from now on all $r_i = r$.

Let us denote by $\lfloor a_i \rfloor$ the multiple of r_i obtained by rounding down a_i, and by $\lceil a_i \rceil$ the multiple of r_i obtained by rounding up a_i. To follow the well-accepted *zero-restricted* version of the Controlled Rounding methodology, if r_i is such that $\lfloor a_i \rfloor = \lceil a_i \rceil$ then we redefine $r_i := 0$, thus $r_i = \lceil a_i \rceil - \lfloor a_i \rfloor$ for all $i \in I$. In other words, cell values which are multiple of the base number are unchanged.

A pattern in the Controlled Rounding methodology is a congruent table $v = [v_i : i \in I]$ such that

$$v_i \in \{\lfloor a_i \rfloor, \lceil a_i \rceil\}. \tag{1}$$

The values r_i are published with the output pattern by the statistical office, thus they are assumed to be known by the attackers. The feasible region for the attacker problems associated with attacker k is defined by

$$\begin{array}{ll} My = b & \\ v_i - r_i \le y_i \le v_i + r_i & \text{for all } i \in I \\ lb_i \le y_i \le ub_i & \text{for all } i \in I. \end{array}$$

We are not using the general concept of information loss as defined at the end of Section 1. Instead, in controlled rounding the natural concept of "loss of information" of a cell is defined as the difference between the nominal value and the published value. Then, the loss of information of a pattern is the weighted sum of all the individual loss of information:

$$\delta(v,a) = \sum_{i \in I} w_i \,|\, v_i - a_i \,| \tag{2}$$

The optimization problem is referred as *Controlled Rounding Problem* (CRP). For more technical details, we refer the reader to [1]. Here we show results of a simplified mathematical model illustrated in Figure 2. This model keeps the additivity requirement, and the minimization of the loss of information through the objective function. Additionally, the external bounds and the protection levels are considered on each rounded cell value. Note that the objective function (2) is a linear function in the variables x_i though parameters c_i, each one measuring the relative cost of rounding up value a_i (i.e. $x_i = 1$) instead of down (i.e. $x_i = 0$).

Figure 2. Basic ILP model for Controlled Rounding.

$$\min \sum_{i \in I} c_i x_i$$

subject to:

$$\begin{array}{ll} \sum_{i\in I} m_{ji}(\lfloor a_i \rfloor + r_i x_i) = b_j & \text{for all } j \in J \\ x_i = 1 & \text{if } \lfloor a_i \rfloor < lb_i \ \text{ or } \ upl_i > \lceil a_i \rceil \\ x_i = 0 & \text{if } \lceil a_i \rceil > ub_i \ \text{ or } \ lpl_i < \lfloor a_i \rfloor \\ x_i \in \{0,1\} & \text{otherwise.} \end{array}$$

Recall that when a cell value a_i is multiple of the base number $x_i = 0$ and so remains fixed. Note also that the conditions fixing a variable either to 0 or to 1 are required in, for example, a magnitude table with $a_i = 4$, $LPL_i = 2$, $UPL_i = 2$, $lb_i = 0$, $ub_i = +\infty$ and $r_i = 5$. Indeed, these given parameters will try to guarantee a protection interval $[2,6]$, and therefore all protected zero-restricted patterns must have $x_i = 1$. Exceptionally, these conditions are unnecessary in some special situations. This is a situation produced when protecting a frequency table with τ -ARGUS. In other words, a cell i with $\lfloor a_i \rfloor < \lceil a_i \rceil$ in a table generated by τ -ARGUS always satisfies $lb_i \le \lfloor a_i \rfloor \le lpl_i$ and $upl_i \le \lceil a_i \rceil \le ub_i$ when $lb_i = \lfloor a_i \rfloor = lpl_i = 0$ and $upl_i = \lceil a_i \rceil < ub_i$. Under these hypothesis the external bounds and the protection levels are useless, and therefore the rounder called by τ -ARGUS will have "only" the task of finding a 0-1 solution satisfying all the $|J|$ equations. We remark the word "only" as still the resolution of the ILP model remains very difficult.

3. Heuristic Approach

Section 3 has outlined the mathematical formulation of the controlled rounding problem, this section will describe how the methods have been implemented in the HCRP, in practical terms. The aim of this section is to produce documentation to explain the general structure of the HCRP. The implemented algorithm can be summarized as follows. It consists of two methods:

Sophisticated Method : This is a near-optimal approach to find a proper solution for the Controlled Rounding Problem, including the protection requirements. It also tries to find a solution by rounding each cell value to a closer multiple of the base number, up or down. The method basically solves the described mathematical model on figure 2 through a branch-and-bound procedure where the bound is computed by solving a linear-programming relaxation. The current implementation should be observed as a multi-start greedy procedure where, at each node of the branch-and-bound tree, the fractional information is used to build a (potential feasible) integer solution. This is explained in more detail below. The whole method will be referred to here as HCRP. Due to the difficulty of the combinatorial problem, a solution may not exist or, even if a solution exists, the approach could required a very long computational time. Therefore, a potential output of this method is "no solution found".

Simple Method : This is a fast approach to build a rounded table, called RAPID. RAPID is applied by rounding each internal cell to the nearest multiple of the base number. Then, marginal cell values are obtained by summing the rounded internal cells and the final rounded table is saved on a solution file. The disadvantage of this approach is the quality of the solution, since any of the marginal cells can have rounded values far from the original values. On the other hand, it ensures a solution in case that the HCRP does not find a better one.

The two methods have been combined in a single algorithm, which is the new rounder. The RAPID method is executed while HCRP does not have a feasible integer solution. In this way, even if the user stops the execution of the whole algorithm, a rounded table will be available. This combination of the RAPID method inside the overall HCRP method has been implemented as follows:

1. First, a data structure is built so the sophisticated method can easily go from linear relations to cells and vice-versa. This step is reading the table from disk.

2. Cells with pre-fixed values are identified. Fixed cells are cells with values that are multiples of the base number (including zeros) and any additional cells fixed by the model constraints in Figure 2. This preprocessing is done by the linear-programming solver. This phase can require at most 2 minutes on the large tables used in our experiments. When time is an issue, the first two steps cannot be aborted, and the user is forced to wait until the end.

3. LP relaxation of the model in Figure 2 is built and a branch-and-bound algorithm is started. As a result of this procedure a fractional solution to the problem is built and the search for the integer solution is started by using the "best-bound first":

   (a) Using the fractional solution of the linear-programming relaxation the solver tries to build an integer solution at each node. Parameters for the solver are defined so that variables with fractional values are set to integer one by one. This is called in our paper LOCAL, however for the solver users it can be found as HEURFREQ in XPRESS, for example. When the solution of the linear-programming relaxation is integer then this is a LOCAL solution. The integer LOCAL solution provides an upper bound to the minimization of the information loss. The upper bound value is updated as new LOCAL solution are found. If there is no integer solution branching is required.

   (b) Branching means that we will proceed with a recursive approach where an open subproblem from a list L is solved and replaced by at most two new open subproblems fixed at either $x_i = 0$ or 1. The first open problem is named *father*, while the subproblems are named *children*. A subproblem is a linear program solved at each node. Each new subproblem is solved before being saved in L, so there is a *lower bound* associated with each subproblem. To solve each subproblem we use a linear-programming solver, like XPRESS or CPLEX.

(c) We select a cell variable with value closest to 0.5 in the fractional solution since we give priority to variables that are farthest from the desired values of 0 or 1. Internal cells are preferred first than marginal cells, and ties are broken by selecting the cell variable in the largest number of linear equations. Once a variable x_i has been selected, the branching consists in creating one subproblem by adding $x_i = 0$ and another subproblem by adding $x_i = 1$. The variables that have been fixed previously in the father problem (i.e. node) continue fixed in the children.

(d) If the list L is non-empty (i.e., there is an open subproblem created by a previous branching step), we select one problem from this list with the smallest lower bound, thus ensuring a global lower bound on the optimal solution value. This is a selection criteria called *best-bound first*.

4. A fractional solution derived from the LP relaxation problem (Figure 1) is then passed to the RAPID whenever we do not have an integer solution from HCRP. The fractional values associated with the internal-cell variables in the model are rounded to their closest integer, thus defining a rounded value for each internal cell. These rounded values for the internal cells are used by the RAPID to generate rounded values for the marginal cells. This is a new RAPID solution. However the quality of the obtained rounded table can be poor, although it is a better rounded table than that computed by the RAPID algorithm in the first step. Therefore, the RAPID solution file is upgraded. This is to guarantee that even if the program stops, either by intervention of the user or because of time-limit specifications, an additive and integer solution will be available.

5. The stopping criteria of this algorithm are:

(a) the list L is empty. In this case there are two possibilities. Either this is the best integer solution found, and in this case we say it is optimal. Or no integer solution has been found by LOCAL, and in this case it is infeasible.

(b) the time limit inserted by the user is achieved. In this case we do not have optimality proof. In the best case, the LOCAL approach was successfully run and found an integer solution (this is a *feasible solution*). Otherwise, we only have a RAPID solution.

(c) The user decides to stop with the first LOCAL solution. Again, this will be a feasible solution, but will not be optimal.

The algorithm described above is for the zero-restricted model. For the general non-zero-restricted mode, the implementation is the same, with the observation that branching can now be done also on x_i^+ and x_i^- variables. Indeed, one can get a fractional solution from a linear-programming relaxation which has all the x_i integers, but still the solution is not feasible if x_i^+ and/or x_i^- contains a non-integer value. In this case the branching phase creates new subproblems by selecting and fixing one of these fractional variables. Since there are more possibilities for each cell i, the search space is larger and therefore the overall resolution may take longer.

4. Computational Experiments

Table 1. Computational results on 23 real-World instances.

	dim	cells	eqns.	Hierar	Sol.Type	Time (sec)	GAP/MaxDist/Jumps
1	2	102051	2052	No	Optimal	17.34	-/-/-
2	2	204051	4052	No	Optimal	96.00	-/-/-
3	2	408051	8052	No	Optimal	387.32	-/-/-
4	2	459051	9052	No	Optimal	493.60	-/-/-
5	2	510051	10052	No	Optimal	1449.61	-/-/-
6	2	1080051	20052	No	Optimal	2260.24	-/-/-
7	2	655793	125067	Yes	Optimal	1014.96	-/-/-
8	2	437532	83314	Yes	Optimal	504.15	-/-/-
9	2	276675	54691	Yes	Optimal	185.91	-/-/-
10	2	118932	24568	Yes	Optimal	26.97	-/-/-
11	3	148960	58548	No	Rapid	1200.00	-/8/8
12	3	133448	52454	No	Optimal	116.28	-/-/-
13	3	124880	49088	No	Optimal	83.20	-/-/-
14	3	46396	18255	No	Optimal	10.44	-/-/-
15	3	38922	15318	No	Optimal	202.63	-/-/-
16	3	38922	15318	No	Optimal	18.50	-/-/-
17	3	181804	104295	Yes	Rapid	1200.00	-/3777/4349
18	3	121296	69548	Yes	Rapid	1200.00	-/2271/2449
19	3	65296	37860	Yes	Feasible	1200.00	0.0010/-/-
20	3	56616	32490	Yes	Feasible	1200.00	0.0013/-/-
21	3	56616	32490	Yes	Feasible	1200.00	0.0010/-/-
22	3	297388	173447	Yes	Rapid	1200.00	-/31544/1120
23	3	787780	461385	Yes	Rapid	1200.00	-/83283/591

Table 1 shows results from testing a small number of tables of the kind typically produced for NeSS. All tables have one geography variable where there may be several thousand categories for small areas, and one or two other variables. The hierarchy is through the geography variable. All tables have $<1\%$ zero cells. Testing was done on an Intel Pentium IV machine with 2.8 GHz processor and 2 Gb RAM. XPRESS-MP 2005 was used as a Mathematical Programming library in our implementation, performed using the C programming language with Microsoft Visual .NET. The larger instance has about 1,000,000 cells, and it was solved to optimality in about 40 minutes.

References

J.J. Salazar, “A Unified Mathematical Programming Framework for Different Statistical Disclosure Limitation Methods”, *Operations Research* 53/5 (2005) 819–829.

A process for writing Standards and Guidance for tabular outputs from ONS

Christine Bycroft and Philip Lowthian
Statistical Disclosure Control Centre, Methodology Directorate, Office for National statistics, UK
e-mail: christine.bycroft@ons.gov.uk, philip.lowthian@ons.gov.uk

Abstract: As a National Statistics Institute, ONS publishes a vast array of tabular outputs, with a requirement to protect the confidentiality of respondents whose information is combined to produce those outputs. Disclosure control in some form has always been applied to tabular outputs, but there has been no overall agreed standards or methods. We wish to provide a systematic approach to disclosure control that will result in consistent methods being applied to similar types of outputs. To achieve this goal, ONS is now developing Standards and Guidance for confidentiality protection of tabular outputs. This paper describes the approach taken to producing Standards and Guidance for ONS Business Surveys and Household Surveys.

1. Introduction

ONS publishes a vast array of tabular outputs, and has a requirement to protect the confidentiality of respondents. Disclosure control in some form has always been applied to these outputs. Current practice is the result of a variety of approaches historically, but there has been no overall agreed standards or methods. We wish to provide a systematic approach to disclosure control that will result in an agreed assessment of disclosure risks and consistent disclosure control methods being applied to similar types of outputs.

The UK situation is complicated because there is no single piece of legislation governing statistical data collection, but different Acts under which specific data collections must operate, as well as Common Law obligations. ONS has a set of operational principles set out in the Code of Practice, and a Protocol on Data Access and Confidentiality that expands considerably on the brief statements about confidentiality in the Code of Practice. However the Protocol still requires interpretation and judgement and is not aimed at providing specific details of how confidentiality protection should be applied in particular situations.

ONS has a programme for developing Standards and Guidance for confidentiality protection. These are formal corporate documents that describe the legal basis for confidentiality protection, contain explicit statements of what disclosure risks are to be protected against and describe standard methods and tools that should be used. In preparation for writing the Standards and Guidance, existing methods were reviewed, followed by a resolution of issues arising from these reviews.

The paper is organised as follows. Section 2 presents a Confidentiality Framework we have developed for carrying out reviews of existing practice, and for structuring the Standards documentation. Section 3 shows how we carried out the review of existing methods whilst section 4 briefly describes the results of the reviews. Finally Section 5 describes the contents of Standards and Guidance, with some examples from those in preparation.

2. A Framework for Confidentiality Protection

A Framework for Confidentiality protection has been developed to guide the reviews of existing practice in individual surveys, and to structure the Standards and Guidance documents. This generic framework has also been used as the basis for protection of microdata (Jackson 2005). The idea of balancing disclosure risk with data utility forms the basis for the framework (see for example Duncan et al), whilst also recognising that one is working within particular legal, ethical and practical constraints. A disclosure control method should first of all reduce the risk of disclosure to a level acceptable to the agency. Theoretically, the best method is one that provides sufficient protection in a way that best preserves the utility of the data. However in practice we must choose a method that can be implemented with available resources and software and within time constraints.

We identified five key aspects critical to producing soundly based confidentiality protection. These are:

1. Why is confidentiality protection needed?
2. Data: what are the key characteristics and uses of the data?
3. Disclosure risk: what disclosure risks need to be protected against?
4. Disclosure control method
5. Implementation

The reviews of existing practices to be discussed in the following section obtained information on all these aspects, and the Standards and Guidance documents are set out under the five Framework headings.

Confidentiality protection measures impact upon respondents (or their representatives), data producers and data users, and there are inevitably tensions between the different viewpoints. We have found that working systematically through the Framework in a collaborative manner has helped to understand where differences in opinion arise and in finding a resolution to those differences. We have engaged in a partnership process with those involved in producing outputs, experts in statistical disclosure control techniques, legal and policy advisors and considering key users and uses of the data..

3. Review of Existing Methods

3.1 Obtaining details on current methods

ONS is responsible for many surveys in a large number of areas, including over 80 business surveys and up to 20 household surveys. While confidentiality protection has always been important, different areas have developed their own disclosure control methods independently. Our aim is to ensure that confidentiality protection is carried out in a consistent and coherent manner across the office.

The full project has been broken into stages covering common families of outputs:

- Business Surveys
- Household Surveys
- Administrative data, and other "whole population" data sources[1]
- Analytic Outputs
- Indices and other complex derived outputs

[1] The 2011 population Census is dealt with separately

To date, draft Standards and Guidance for tabular outputs from Business Surveys and Household Surveys have been prepared. The development of internal Standards and Guidance was led by the Statistical Disclosure Control (SDC) Centre within the central Methodology Directorate. As a small central unit within a large organisation, we had to develop our own knowledge of ONS surveys and their disclosure control practices. As a starting point for writing Standards we carried out a review of existing disclosure control practices and an assessment of their appropriateness.

Given the large number of sample surveys to consider it would have been impractical to review them all. Ten business and household surveys were selected to be reviewed in depth covering a range of different outputs. These were surveys regarded either as key ONS outputs or idiosyncratic ones with unique disclosure control problems. It was expected that these would provide a good understanding of current methods used and highlight any issues that would need to be dealt with. Examples of surveys reviewed are the Annual Business Inquiry (the largest business survey, obtaining employment and financial information) and the General Household Survey (multi-purpose continuous survey of people living in private households in the UK).

3.2. The questionnaire

The review was performed by means of an open ended questionnaire which was initially filled in by the SDC Centre using existing documentation. There then followed a meeting with the relevant business area with further follow up discussions until the completed questionnaire was agreed. On completion of the individual survey investigations, a report summarised findings, and highlighted where changes to existing practice was recommended and areas where further work was needed. The report of the review then formed the basis for writing the Standards and Guidance.

The questionnaire contained detailed questions under the following two groupings reflecting the Framework described in section 2.

General background of group and characteristics of data used: This included specific legislation applicable to the data and assurances of confidentiality given to the respondents, the main outputs and data user needs.

Rules/Methods of disclosure control: This included disclosure risks they wished to protect against, rules used to check whether the data are disclosive or not along with details on the methods used to protect the tables and the implementation (*e.g.* software used) of these rules and methods.

Evaluation: A third section contained the SDC Centre’s assessment of these rules and methods and whether we thought their rules were understood, applied correctly and were addressing the disclosure control problems.

4. Results from the Reviews

This section provides some of the main results from the reviews of business surveys and household surveys.

4.1 Business Surveys

We found a consistent approach across business surveys with the aim being to protect against revealing information about individual business respondents. In this they are following the 1947 Statistics of Trade Act. Respondents are told that no information relating to an individual company will be released.

The rules for determining unsafe cells for magnitude tables typically use a threshold rule (*i.e.* to be safe, a cell must have a minimum number of Enterprise Groups[2]) and a p% rule applied at Reporting Unit level. For some outputs there are simpler rules, such as a very high threshold rule, which have a tendency to overprotect. These were in place where the IT system was not able to implement the more sophisticated p% rule. The p% rule was applied at Reporting Unit level because of software limitations.

Table protection following these safety rules is a combination of table redesign and secondary suppression. Some surveys released only tables designed to ensure there were no unsafe cells, while some provided much more detail that required sometimes complex secondary suppressions. For all but the simplest one-dimensional tables, secondary suppression was carried out by hand. For frequency tables conventional rounding was used.

4.2 Household Surveys

The review found that until recently, for many surveys cells were suppressed on quality grounds, generally based on a certain number respondents, and that these quality suppressions also provided protection of confidentiality. A change of policy concerning quality has resulted in much lower quality thresholds now often used. Another group of surveys had never used suppressions for quality or confidentiality reasons, but indicated unreliable estimates using symbols. Some surveys gave detailed unweighted sample base numbers, publishing the exact number of respondents contributing to a cell estimate as a quality indicator. In the past some surveys had published only unweighted estimates. Thus, in contrast with the business surveys there have been no explicit rules or consistent approach to confidentiality protection, and there is an interdependence between the approach to publishing poor quality estimates and confidentiality.

5. Standards and Guidance

One of the decisions to be made when writing the Standards is how much detail to provide. We envisage two main audiences. The first simply wants to know the rules or methods of disclosure control for a particular type of output. The second audience wants to know the standard rules and methods to be applied, but also wishes to understand the disclosure risks and reasons why a particular method has been chosen. To cater for both groups, the document is split into two main parts. The first provides a basic summary of the disclosure control rules. The longer part of the document then follows the headings of the Confidentiality Framework described above. An Appendix provides examples from real published tables to further illustrate and explain the disclosure risks and application of disclosure control methods. For the internal ONS audience, the body of the standard can also be reasonably concise, and more discussion and explanation can be provided in other papers.

Each of the Framework headings is now discussed in more detail. Space restrictions prevent us providing full examples from the standards.

[2] The basic legal unit of business structure is the enterprise. An Enterprise Group is a grouping of enterprises that have some association. Enterprises may be split into Reporting Units, which in turn consist of one or more Local Units.

5.1. Why is confidentiality protection needed?

The question of why confidentiality protection is needed is fundamental to the whole process and is a key element in obtaining agreement on defining disclosure risks. To answer this we must consider any relevant legislation or policy requirements that must be met and any statements made to respondents. Also it should be established whether there are any particular ethical issues, any variables that might be highly sensitive, or on the other hand any situations where confidentiality protection is not required.

5.2. Data: Key Characteristics and Uses

Once the reasons for needing confidentiality protection are clearly stated, the data and its main uses are described. This includes the type of data, *e.g.* full population or sample survey; the sample design; an assessment of the quality *e.g.* the level of non-response and coverage of the data; the variables and whether categorical or continuous; types of outputs produced, *e.g.* count or magnitude tables. All of these factors influence both the disclosure risks and appropriate disclosure control methods. Then it is important to understand the main uses and users of the data. For example, are there important government planning or policy uses, what are researchers main interests, are exact values needed.

5.3. Disclosure Risk: What disclosure risks need to be protected against?

Disclosure risk assessment then combines the understanding gained above with an intruder scenario analysis to provide an explicit statement of what the disclosure risks are, and what elements of the outputs pose an unacceptable risk of disclosure. We have found in some cases that extensive discussion has been needed to reach agreement on what constitutes a disclosure risk. Writing down intruder scenarios similar to the process described in Elliot and Dale for microdata has proved very useful, as one must then consider the possible situations where confidentiality might be breached. Disclosure risks are heavily dependent on whether data is from whole population sources or samples, (and if so, the sample design) and the sensitivity and value of the data.

For business surveys, responses may be commercially valuable to competitors, extending to imperfect estimates. The main intruder scenario for magnitude tables is the example often found in the literature, where one business contributing to a cell attempts to discover the response of a competitor in the same cell, either exactly or to a close approximation. Our recommendation for the business survey standards was to retain the current threshold and p% rules for definition of unsafe cells for magnitude tables. Application of rules for negative values needed to be investigated further.

Following the Protocol on Data Access and Confidentiality, we are looking for an acceptable minimum disclosure risk, not an absolute guarantee of safety. Judgements have to be made to interpret Protocol phrases such as “likely to identify” and “disproportionate time, effort and expertise”, in the context of the particular circumstances for which each Standard applies. Decisions are made based on the disclosure risk assessment. For example, parameters for the minimum threshold and p% rules encapsulate the judgements made by ONS of what constitutes an acceptable level of risk. Discussions were held with senior management on the appropriate level at which the p% rule should be applied, so that these subjective decisions are endorsed at a senior corporate level.

5.4 Disclosure Control Method

Once a clear understanding of disclosure risk is reached one is in a position to consider possible disclosure control methods. Several factors must be balanced when deciding on the “best” disclosure control method in a given situation (e.g. Massell). Some measure of information loss and the impact on main uses of the data can be used to compare alternatives. Any method must be implemented within a given production system so available software and efficiency within demanding production timetables must be considered.

Previous practice and the availability of software, rather than any quantitative comparisons of information loss have largely fixed the choice of disclosure control method for regular outputs. We focus the discussion on the Business Surveys.

Business surveys have always used table redesign and cell suppression for magnitude tables, and will continue to do so, despite high information loss and the difficulty of maintaining secondary suppression patterns in ad hoc releases. Implementation of secondary suppression should change from the current manual methods, to use of the Tau-Argus software. There is some interest in using perturbative methods for magnitude tables but any new methods will need to be proven to provide adequate confidentiality protection, and data users will need to be convinced that more information is in fact provided. Some change is occurring with the availability of Tau-Argus. Controlled rounding via Tau-Argus will be used rather than conventional rounding for Business Demography count tables.

5.5. Implementation

The final aspect is implementation of those methods. This will include where responsibility for ensuring confidentiality protection lies, the software to be used along with any options and parameters, any exclusions or exemptions, and a process for approval to use other methods. Standard wording to be used as footnotes to tables or textual information will be included. These should then form part of the survey metadata.

6. Conclusion

There has been a clear need within ONS for a consistent approach to confidentiality protection. In the absence of a single Act covering Official Statistics, the Code of Practice and Protocol on Data Access and Confidentiality state general principles and guidelines that must be adhered to for all National Statistics. The Standards and Guidance described here are the next layer of documentation for tables, where these principles are interpreted for common families of outputs. When complete, they will provide a clear explanation of the disclosure risks and the standard rules and methods to be applied.

The Framework for Confidentiality described above has been most useful, giving a sequence of steps to work through in the reviews of existing practice, providing a focus for discussions and as a structure for the Standards themselves. The ONS Standards will also be a valuable resource for other UK Government agencies producing National Statistics that must also comply with the Protocol.

References

Cox L(2001) *Disclosure Risk for Tabular Economic Data* Chpt 8 Doyle et al (Eds) *Confidentiality, Disclosure and Data Access Elsevier Science* BV.

Duncan, Fienberg, Krishnan, Padman, and Roehrig (2001) *Disclosure Limitation Methods and Information Loss for Tabular Data* Chpt 6 Doyle et al (Eds) *Confidentiality, Disclosure and Data Access. Elsevier Science* BV.

Elliot, M J and A Dale, (1999) Scenarios of attack: The data intruders perspective on statistical disclosure risk. *Netherlands Official Statistics* Vol 14 Spring

Massell, P.(2004) Comparing statistical disclosure control methods for tables: Identifying the key factors. *Research report US Bureau of Census*

Jackson P and J Longhurst J (2005), Providing access to data and making microdata safe, experiences of the ONS. *UN-ECE Work session on statistical data confidentiality.*

Using Fixed Intervals to Protect Sensitive Cells instead of Cell Suppression

Steve Cohen and Bogong T. Li
[*] **Bureau of Labor Statistics, 2 Massachusetts Ave. NE., Washington, D.C. 20212,**
e-mail: cohen_steve@bls.gov and li_t@bls.gov,

1. BLS QCEW Proposed Publication Change

BLS Quarterly Census of Employment and Wages (QCEW) is a census that collects data under a cooperative program between BLS and the State Employment Security Agencies. The data contain broad employment and wage information for all U.S. workers covered by state unemployment insurance laws and federal workers covered by the Unemployment Compensation for Federal Employee program. Tabulations of QCEW outcomes are available by 6-digit NAICS industry, by county, by ownership sectors and by size groups, in the form of print, automatic e-mail, fax or plain text file directly from BLS Internet ftp servers. The detailed coverage and readily availability of the QCEW tabular data make it especially vulnerable to confidentiality disclosure risks. Cell suppression is used as for the tabular data confidentiality protection schema.

Since cell suppression methods currently implemented suppress a large number of cells in order to protect QCEW publication tables, an alternative method is sought. Using QCEW data analyzed in this paper, following the BLS confidentiality sensitivity measures, we found for this data set containing employment of five major industry sectors (2-digit NAICS sectors) within a medium-sized U.S. State, 9979 or 59% of 16,878 publication cells have to be completely suppressed using network method, 10631 or 62% of all cells using the hypercube method (for a description of the hypercube method see Repsilber (1994)). The level of employment represented by the suppressed cells is relatively small in comparison to the number of cells suppressed, ranging from 10% to 15% of the total value. Similar results of this magnitude for cell suppression have been also reported by other researchers. Much detail on industry employment distribution at various geographic levels and other cross-classifications is lost due to confidentiality protection

One alternative to complete suppression considered by QCEW would be to publish primary cells in pre-defined, fixed intervals (FIs). Instead of suppressing the value of the sensitive cells, this method would publish all primary suppression cells in FIs which contain the exact value of the sensitive cell value. The consistency of the definition of these pre-defined intervals is kept across tables so that the users can compare values between various industries, geographic locations and other classifications by establishment characteristics, by just looking at the intervals. This method of publication can be used for employment and earnings data, though our discussion in this paper will only focus on employment level data.

Similar to the issues surrounding the cell suppression problem (CSP), if QCEW data is published with FIs replacing primary suppression cells, to prevent outside intruders gaining identifiable information of individual contributors to a cell, additional protecting cells (PCs) may have to be published in FIs. Otherwise an intruder may be able to utilize this additional information and the additive relationships existing in the table to estimate the value of primary cells now in FIs and therefore the value of some contributors to the cell. Intruders can produce better estimates now than before with the added information of published FI bounds. The problem of minimizing the amount of cell values now expressed in FIs by selecting the right set of PCs while still preserving the protection of primary cells is what we call the fixed interval publication problem (FIPP). We will use the following fixed interval ranges for employment levels: 0-19, 20-99, 100-299, 250-499, 500-999, 1000-2499, 2500-4999, 5000-9999, 10000-24999, 25000-49999, 50000-99999, 100000 or more.

Since this risk arises from the additive relationships in the table and is similar to CSP solutions that have been implemented in some BLS survey programs, we start searching solutions made to solve CSP. Our current knowledge indicates CSP problem has been established by researchers as a MILP problem, see Kelly (1990). Exact solution to MILP model belongs to the class of the strong *NP*-hard problem. Heuristic solution procedures such as the network flow method, see Cox (1980 and 1995), for 2-dimensional tables, multi-commodity network flow method for n-dimensional tables, see Castro and Nabona (1996) and hypercube method by Repsilber (1994) and Giessing (2001) have been proposed. These heuristic methods only provide sub-optimal solutions as pointed by Castro (2001). Fischetti and Salazar (1999) proposed a solution using branch-and-cut algorithm as one of the mathematical programming techniques to reach a solution with proven optimality on 2-dimensional tables with up to 500 rows and 500 columns. The problem is solved in a few minutes on a standard PC. Fischetti and Salazar-Gonzales (2000) extended their work to other tabular data including k-dimensional table with k>2, hierarchical tables, linked tables etc., using branch-and-cut based procedures. Alternatively, instead of completely suppressing table cells, Salazar (2001); Fischetti and Salazar (2003) proposed a "partial cell suppression" method that will publish a subset of table cells with variable estimation intervals. Though FIPP and CSP shares the same MILP model, *unfortunately*, so far we think all of the above mentioned secondary cell selection methods do not apply directly to selecting protecting cells (PCs) that are to be published in FIs, neither optimally nor heuristically. The reason is that these models can not accommodate the knowledge of the FI bounds.

In this research we will propose an iterative "selection-improvement" algorithm, which improves cell selection upon each previous suppression pattern until all primary cells are sufficiently protected. All of the selection-improvement steps begin with procedures already implemented in BLS QCEW program. Though no claim of optimality is made in this paper, this method does make publication of tables with FIs realistic, and, as the evaluation at the end shows, there aren't significantly more cells published as FIs than the number of cells completed suppressed. After describing our procedure, we will provide an evaluation study using actual employment data from a U.S. state. We will compare the results with current suppression methods, look into convergence rates, level of information loss and computer programming difficulties associated with various cell selection methods.

2. The Selection-Improvement Algorithm

The iterative selection-improvement algorithm has two stages at each iteration, (1) selecting PCs and (2) conducting an audit on the publication table with the newly selected PCs in FIs. If the audit finds any primary cell is still at risk, the algorithm re-iterates by selecting more PCs and conducting another audit until all primary cells are protected. The initial set of PCs is the set of cells selected through one of the CSP methods. In case the iterations fail at the end, i.e. no candidate PCs available for selection while there are still unprotected cells, the method defaults back to the usual CSP solutions targeting only the remaining exposed cells. The steps of the algorithm are summarized as follows:

Step 1. Identify primary and secondary cells in a table via a CSP method and publish them in pre-defined FIs.

Step 2. Apply linear constrained optimization to identify those primary cells with disclosure risks.

Step 3. For those primary cells at risk, select additional cells that have not been selected previously from the publication table and publish them in FIs. Three specific methods are proposed for this research and will be briefly described in following paragraph and sections. This is the 'selection step".

Step 4. Apply linear constrained optimization again to check if any primary cell in the original table is still at risk. If yes, return to step 3; otherwise EXIT the algorithm, the table is successfully protected. This is the “audit step”.

Step 5. If the step 2 – 4 iteration fails to protect every primary cells, i.e. no further unsuppressed cells available for selection while there are still disclosed primary cells, use any solution method to CSP, i.e. completely suppress these exposure primary and corresponding secondary cells.

There are several alternative methods can be used to select additional PCs in Step 3. We can randomly select cells that are within the same row or column of the exposed primary cells, or we can select through more complex MILP models and mathematical programming techniques. We would like to minimize either the number of cells to be selected or the total value of the selected cells. In this paper we studied the following three methods in the selection step: the Systematic, Single-Source Shortest Path (SSSP) and the Random Selection methods.

1. Systematic Method. To minimize values published in intervals, this method selects the smallest cell among all cells that form additive relationship with two selected exposure cells that need further protection that has not been suppressed during the previous iteration(s). This cell is published as a pre-defined FI. Default to Random Selection Method (see 3 next) at the end if this method fails.
2. Single-Source Shortest Path (SSPS) Method. This method models the table as a network similar to Traveling Salesman’s Problem (TSP), treat all primary exposure cells on a table as destinations of a traveling map. The method aims to find the shortest path through these destinations, to minimize the total cell values expressed in FIs. To make this TSP solvable for all tables, the method fixes the order of the destinations or vertices on the table network. The method only needs to find the shortest path connecting the order-fixed set of vertices to form a closed “loop” with minimized path. Publish all cells that are not already selected in previous iterations on the chosen loop in FIs. Default to Random Selection Method if this method fails at the end.
3. Random Selection Method. This method randomly selects a cell among all cells that form additive relationship with the primary exposure cells. The candidate cells are cells that are either in the same row or column as the primary cell. If all cells forming additive relationships are already selected during previous iteration(s), or it by itself is the only decent from the higher hierarchy, go one hierarchy step higher until additional protecting cells can be found through additive relationships. Randomly select protecting cells among the candidates, publish these and all cells along the hierarchical searching path as FIs.

In addition to providing a valid solution, the FIPP algorithm introduced here is easy to implement in production, since it simply combines separate existing confidentiality protection procedures, such as the complementary cell suppression techniques and auditing of tabular data through linear programming. It requires less software changes in the survey production environment because the only change to current complementary cell selection procedures is the addition of auditing cycles. The difficulty of selecting additional PCs could be simple, for the Random Selection Method, or modestly complex, for the SSSP. The auditing of a table during any stage of the process can be done through available table auditing software tools. Programming work for selecting PCs is only need to be done once and be reused later. More importantly, this method does not alter the actual micro data behind the tabular publication, as that of methods like adding noise to the micro data, which may add unwelcome noise to even safe cells.

3. Evaluation of the Method Using a Subset of Actual QCEW Data

We used actual QCEW employment publication tables for evaluating our stated FIPP procedure. This subset of QCEW data contains eight major 2-digit NAICS industry sectors in a medium-sized U.S. State. In actual BLS publications, these data are published in tabular form separately in multi-dimensional table format classified by county and 6-digit NAICS industry, as well as by establishment size group, metropolitan statistical area (MSA) and ownership types. We used only the 2-dimensional employment table classified by county and hierarchical NAICS code, from 2 to 6-digit, to demonstrate our algorithm. Uses of 2-diminensional table may limit our evaluation conclusion, since multi-dimensional publication tables are "connected", or in other words there are more additive relationships existing than what we considered. Nevertheless these additional additive relationships are identifiable. Once they are correctly identified, we can always add them in the model. Therefore we believe with some modification our method applies to tables with any dimensions and we should expect the number of cells in FIs somewhat more than we report here. Table 1a displays a portion of this publication table currently a user sees in BLS publications. In this table the cells marked with "x" are suppressed cells due to primary and secondary suppressions. In this evaluation, we will apply our FIPP procedure to the data and compare their performance with that of the complete suppression. Table 1b shows the results treated with our FIPP procedures.

For a quick note about how we process the data through some computing tools: we first put the raw micro data through primary and secondary suppression selection using software tool Tau-Argus, see Hundepool, Willenborg et al. (2004). The suppressed table is then formatted to lp format in S-plus® to be used in Matlab®. In Matlab® we called solver lp_solve to conduct the audit of the table, lp_solve is a MILP solver available from the Internet community. If the iteration is not finished, the audited table is passed back to S-plus® and again we select additional PCs with the three methods we stated earlier in this paper. S-plus® and Matlab® were used to convert the publication table between publication tables and LP model input formats. Additional PCs are selected within S-plus® where additive relationship of the entire publication table is kept. Unless the cycles successfully protected all primaries, the cycle should reiterate itself continuously. The convergence is guaranteed through the Random Selection method.

Table 1a. A sample evaluation data set as published perturbed for confidentiality

NAICS code	Counties of a U.S. State								
	Total	County 1	County 2	County 3	County 4	County 5	County 6	County 7	etc.
…	…	…	…	…	…	…	…	…	
…	…	…	…	…	…	…	…	…	
451	13940	113	1758	2691	111	X	241	64	
4511	9070	82	1121	1699	x	X	166	x	
45111	4187	26	703	773	89	-	51	51	
451110	4187	26	703	773	89	-	51	51	
45112	2648	x	274	451	x	X	x	-	
451120	2648	x	274	451	x	X	x	-	
45113	1237	x	110	302	-	X	x	x	
451130	1237	x	110	302	-	X	x	x	
45114	998	x	35	173	-	-	38	x	
451140	998	x	35	173	-	-	38	x	
4512	4870	31	637	992	x	-	75	x	
45121	3415	x	504	444	x	-	x	x	
451211	3193	x	x	438	x	-	x	x	
451212	222	x	x	6	-	-	-	x	
45122	1455	x	133	548	x	-	x	-	
451220	1455	x	133	548	x	-	x	-	
…	…	…	…	…	…	…	…	…	
…	…	…	…	…	…	…	…	…	
Total	1166388	15589	98129	190226	7524	5018	22485	12171	etc.

"x" are nondisclosable data due to primary and secondary suppressions

Table 1b. The same section of the evaluation data set as it is published under FIPP method

NAICS code	Counties of a U.S. State								
	Total	County 1	County 2	County 3	County 4	County 5	County 6	County 7	etc.
…	…	…	…	…	…	…	…	…	
…	…	…	…	…	…	…	…	…	
451	13940	113	1758	2691	111	0-19	241	64	
4511	9070	82	1121	1699	20-99	0-19	166	20-99	
45111	4187	26	703	773	89	-	51	20-99	
451110	4187	26	703	773	89	-	51	20-99	
45112	2648	0-19	274	250-499	0-19	0-19	0-19	-	
451120	2648	0-19	274	250-499	0-19	0-19	0-19	-	
45113	1237	0-19	110	302	-	0-19	20-99	0-19	
451130	1237	0-19	110	302	-	0-19	20-99	0-19	
45114	998	20-99	20-99	173	-	-	38	0-19	
451140	998	20-99	20-99	173	-	-	38	0-19	
4512	4870	31	637	992	0-19	-	75	0-19	
45121	3415	20-99	504	444	0-19	-	20-99	0-19	
451211	3193	20-99	250-499	438	0-19	-	20-99	0-19	
451212	222	0-19	20-99	6	-	-	-	0-19	
45122	1455	0-19	133	548	0-19	-	20-99	-	
451220	1455	0-19	133	548	0-19	-	20-99	-	
…	…	…	…	…	…	…	…	…	
…	…	…	…	…	…	…	…	…	
Total	1166388	15589	98129	190226	7524	5018	22485	12171	etc.

4. Summary of Evaluation Results

Each method was carried out to the end without having to apply the Random Selection method. For cell suppression, there are 9979 (59% of total) cells, or 4,535 (7.5% of total) establishments and 162,368 (14% of total) of employment value that are completely suppressed. With the Systematic Selection method, the entire publication table is successfully protected at the end with only two additional iterations beyond the traditional secondary suppression stage. However the number of cells published in FIs is quite large, not surprisingly since the procedure incorporates the existing secondary suppression methods. For the Systematic Selection method, 10,199 or 60% of all publication cells are selected for FI publication, they account for 6,337 establishments or 10% of all establishments in the table and 180,742 or 15% of total employment in the table. However, if taking into consideration of the number of publication cells suppressed, the number of establishments and total values in FIs, the difference between FIPP solution and complete suppression solution is not very large.

Separately for SSSP method and Random Selection method there are about 64% and 69% of all cells in FIs respectively. In terms of the number of iterations required to reach complete protection of the publication table, Systematic Selection takes 2, SSSP takes 3 and Random Selection takes 5. The reason for the difference in the number of iterations could be attributed to the relatively inefficient methods of picking addition PC cells used by the latter two methods. In particular, the Random Selection method does not taken into consideration of the magnitude of all qualified cells.

See Table 2 for summaries and comparisons of the total number of cells and values contained in FIs under each of the three different selection methods as compared to cell suppression.

5. Conclusions

We developed this FIPP solution with the goal to minimize either the total number of cells selected or the total value contained in the cells selected to be FIs. However since the initial step is built upon secondary cells suppression through CSP solutions and subsequent ad hoc PC selection steps, we probably do not achieve this goal truly. The good news is that all confidentiality rules imposed on the table are well preserved and the number of cells released as FIs is reasonable at the conclusion of the algorithm. The last audit step on the table clearly demonstrates all primary cells on the table are well protected. With reasonable effort a feasible solution can be found to a seemingly unsolvable optimization problem. The success of these methods relies on the assumption that the number of iteration cycles is not large, since current CSP solutions tend to over suppress in the first place. Even with the least inefficient selection method, the Random Selection method, only a maximum of five cycles are needed. The complexity of programming, computer usage time and manual intervention varies depending on selection method used. We found the Random Selection takes the least amount of programming time and manual intervention, SSSP takes longer to run on computer and needs more overhead programming effort, and Systematic method requires more manual interaction during the process than any of the other two methods, therefore is the most cumbersome to use. It is possible with more effort put into the computer programming in the future, we can integrated various parts of software tasks into a single program. This is necessary if our proposal is to be adopted in regular publication production environment

Table 2. Cells published as FIs by three difference selection methods compared to CSP method

	Systematic	*SSSP*	*Random*	*Cell Suppression*
Number of iterations to reach convergence	2	3	5	NA
Total number of cells in Fis or completely suppressed	10,199 (60%)	10,772 (64%)	11,615 (69%)	9979 (59%)
Total employment level in FIs or completely suppressed	180,724 (15%)	184289 (17%)	188,955 (16%)	162,368 (14%)
Total number of establishments in Fis or completely suppressed	6,337 (10%)	7,362 (12%)	7,971 (13%)	4535 (7.5%)

One other advantage of our method is that a user can specify cells that he or she does not want to be published in FIs. Once specified, these cells will be treated as if they are constants in the model. The method also allows a user do global coding, i.e. combining categorical variables such that the result will be a table with fewer unsafe cells, though this may need to be done before the selection-audit cycles begin.

We also noticed the following problems with our methods during evaluation of the test data:

1. For the Systematic and SSSP selection methods, the order of the exposure primary cells during each iteration affect the additional PCs selected. In other words, the final set of FI cells could possibly be different if the process is run more than once, since the order of exposure primary cells entered the local protection cycle may be in a different order. Unless the order of cell entry is fixed, which is possible, the process is not repeatable.
2. The Random Selection method produces a different set of selection cells every time it runs, due to the random nature of its selection of PCs in local cycles. Setting the random seeds during iterations will be intractable. Therefore the PC selection process is not repeatable.
3. Though in theory the methods apply to table with any dimensions and hierarchical structures, as long as the additive relationships in the table is expressible, the time allowed us to conduct the study so far limit ourselves to only 2-dimentional tables with hierarchical structure in one dimension. Higher dimensional tables require us decompose the table into lower dimensional

tables and process lower dimensional tables separately then “back-track” separate results at the end. We chose not to experiment that in this study.

Since the test data we used in this study are in reality published as multi-dimensional tables, i.e. there are other additive relationships in the table we actually did not take into consideration, indubitably more cells will be published in FIs and the programming working will be more demanding if the multi-dimensionality is taken into consideration. This is stated in the limitation 3 above. It is convincible that with other practical issues surrounding publishing sensitive cells in FIs, more works have to be done before we can adopt this method for QCEW regular publication.

References

Castro, J. (2001). “Using Modeling Languages for the Complementary Suppression Problem Through Network Flow Models.” Joint ECE/Eurostat Work Session on Statistical Data Confidentiality.

Castro, J. and N. Nabona (1996). “An Implementation of Linear and Nonlinear Multi-commodity Network Flows.” European Journal of Operational Research **92**: 37-53.

Chartrand, G. (1977). Introductory Graph Theory, Dover Publications, Inc.

Cox, L. H. (1980). “Suppression Methodology and Statistical Disclosure Control.” Journal of the American Statistical Association **75**: 377-385.

Cox, L. H. (1995). “Network Models for Complementary Cell Suppressions.” Journal of the American Statistical Association **90**: 1453-1462.

Fischetti, M. and J. J. Salazar-Gonzales (2000). “Models and Algorithms for Optimizing Cell Suppression Problems in Tabular Data with Linear Constraints.” Journal of the American Statistical Association **95**: 916-928.

Fischetti, M. and J. J. Salazar (1999). “Models and Algorithms for the 2-Dimensional Cell Suppression Problem in Statistical Disclosure Control.” Mathematical Programming **84**: 283-312.

Fischetti, M. and J. J. Salazar (2003). “Partial Cell Suppression: A New Methodology for Statistical Disclosure Control.” Statistics and Computing **13**: 13-21.

Giessing, S. (2001). Nonperturbative Disclosure Control Methods for Tabular Data. Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies. Doyle, Lane, Theeuwes and Zayatz, North-Holland.

Hundepool, A. J., L. C. R. J. Willenborg, et al. (2004). Tau-Argus User’s Manual, Version 3.2.

Huo, H. W. (2004). Exercises & Solutions on Algorithms. Beijing, China, China Higher Education Press.

Kelly, J. P. (1990). Confidentiality Protection in Two and Three-Dimensional Tables. College Park, Maryland, University of Maryland, College Park, Maryland. Ph.D. Thesis.

Repsilber, R. D. (1994). Preservation of Confidentiality in Aggregated Data. Second International Seminar on Statistical Confidentiality. Luxembourg.

Salazar, J. J. (2001). “Improving Cell Suppression in Statistical Disclosure Control.” Joint ECE/Eurostat Work Session on Statistical Data Confidentiality.

Note that the views expressed in this paper are those of the authors and do not necessary represent the policy of the Bureau of Labor Statistics.

Topic VI

Sotware for statistical disclosure control

The “Jackknife” Method: Confidentiality Protection For Complex Statistical Analyses

Jobst Heitzig
**Federal Statistical Office Germany, IT User Service / Statistical and Geo-Information Systems,
Gustav-Stresemann-Ring 11, 65189 Wiesbaden, Germany. (jobst.heitzig@destatis.de)**

Abstract. The “jackknife” method of confidentiality protection is a kind of compromise between protection methods based on the data (anonymisation) and protection methods based on results (like those used for tabular data). Like the former, it allows to perform all kinds of statistical analyses with the confidential micro-data, but like the latter, it allows us to release results of higher accuracy than can be computed from traditionally anonymised micro-data.

The idea is to publish not the precise analysis result but a small interval containing it, where the interval’s width is chosen as small as possible but still large enough to ensure confidentiality. More precisely, this protection width is chosen so that a potential data snooper cannot distinguish between his sought target value and some random replacement value.

Depending on the actual analysis, the protection width can be determined in different ways: For robust statistics with bounded influence functions (e.g., the median), that function can be used. For non-robust or fairly robust statistics (e.g., the mean or trimmed mean), the effect of replacing each micro-data cell one at a time by a random replacement is determined or estimated, respectively. At the moment, efficient algorithms exist to protect most classical univariate and bivariate statistics and some non-linear model fitting algorithms. Prototypical implementations in SAS® are available for this.

1. Introduction

The sciences’ growing demand for all kinds of statistical analyses with confidential micro-data can be answered in several ways. Recently, there seems to be some shift from releasing anonymised micro-data files to providing remote access facilities. This paper describes a new method for confidentiality protection which can in principle be applied to arbitrarily complex statistical analyses of confidential micro-data, and presents some prototypical software tools which implement the method for certain-important kinds of analysis.

The proposed method achieves confidentiality protection by publishing analysis results only with some imprecision. An essential feature is that the imprecision is kept as small as possible but still large enough to ensure confidentiality even when the potential data snooper has large amounts of additional information. The basic idea of this “jackknife” method is that the necessary amount of imprecision can in principle be determined in a way similar to the jackknife estimation of standard errors: compute a set of approximate analysis results, each based on a slightlymodified micro-data file which coincides with the original data in all but one position. The published result is then an interval containing all these approximate results. Section 2 of this paper motivates and justifies the basic principle.

In practice, these approximate results can often be efficiently determined or estimated by adjusting the true result after replacing a single value in the micro-data. In case of a robust statistic (such as the median) whose influence function is bounded, it is even more efficient to use this bound directly to compute an interval that can be published safely. Section 3 gives various examples of how the jackknife method can be used to publish all kinds of descriptive statistics and test statistics, showing also that the relative imprecision introduced by the jackknife method is usually of order $O(1/N)$ or $O(ln(N)/N)$, whereas relative standard errors are usually of the larger order $O(1/\sqrt{N})$. Section 4 then describes two corresponding prototypical SAS® macros, `%jk_means` and `%jk_freq`, which have been developed by the Federal Statistical Office Germany.

As an example of how the method works also with advanced statistical analyses, Section 5 presents a macro `%jk_nlin` providing jacknife protection for the leastsquares parameter estimators of non-linear regression models.

The current status of our research is described in Section 6.

2. Motivation, basic principle and rationale

2.1. Disadvantages of anonymised micro-data files

Releasing anonymised micro-data has certain well-known problems. First of all, anonymisation as it is currently performed must often be tailored to each single data file and is almost always based on certain assumptions as to which variables are sensitive, which are possible key variables, what amount of additional knowledge the snooper might have, which observations the snooper might be interested in, which degree of uncertainty would render the disclosed data useless for the snooper, and so on. Furthermore, their level of protection is often measured by some aggregate measure of risk (e.g., the estimated percentage of re-identifiable units in certain subgroups), implying that there can easily remain a small percentage of observations which couldstill be at a high *individual* risk of disclosure.

At the same time, raising this level of protection costs a lot. For one thing, global anonymisation methods (like sub-sampling, global recoding, additive or multiplicative perturbation, etc.) increase the error of most analysis results by a constant factor, independently of the actual number of observations entering the specific analysis, and independently of how problematic these results are with respect to confidentiality. For example, relative standard errors of statistics computed from a 70% subsample are usually about 20% higher than in the whole sample. And recoding both the row and column variables for a χ^2-test from $R = C = 9$ to $R = C = 3$ categories increases the relative standard error ($\sqrt{2/(R-1)(C-1)}$) of the test statistic even by 300%. When variables are removed or recoded to too coarse a level,their quantitative analysis becomes impossible. In addition, anonymisation often introduces a bias into many multivariate and/or non-linear analyses. Even a simple estimation of the sum gets biased when top-coding is used.

All these disadvantages can be justified when the goal is to hand out micro-data to researchers so that they can look at them or run simulations, for example. But when the goal is to provide researchers with a remote access facility for complex but somewhat standardised statistical analyses, they seem to be a great price.

2.2. Difficulties in judging by the number of observations

One approach to protect confidential data in a remote access facility could be to publish precise results when the number of used observations seems large enough, and to suppress the output altogether when it is not – just as it is often done in case of frequency tables. However, already simple examples show that it is not at all obvious what number of observations should be considered safe, even when we only want to provide the researcher with some standard descriptive statistics.

Assume there were seven persons with two variables X and Y, and we were to publish the (sample) mean, variance, skewness and kurtosis of both, together with their covariance. Then any two persons in that group could easily compute the Y-value of anythird person in the group of whom they know the X-value. They only need to subtract their own values and then solve the nine equations for the nine unknown values. Or assume there were even 20 persons and we only wanted to publish the (sample) means m_X and m_Y, variances s_X^2 and s_Y^2, and the covariance of X and Y. Then it might turn out that in this particular group the Pearson correlation between X and Y is 0.99, which would allow anyone who knows the X-value of any of the persons to infer with certainty that the corresponding Y-value is in an interval with centre $m_Y + 0.99 s_Y \frac{x-m_X}{s_X}$ and a width $\leqslant 1.2 s_Y$ (note that this is not a confidentiality region but certain). This interval becomes even narrower the more x deviates from m_X, that is, the more unusual the target person is (see [Heitzig 2004]).

The existing literature on "statistical databases" shows that this approach of judging by the number of observations is complicated enough already when one wants to publish only sums.

2.3. Definition of the jackknife method

From a mathematical point of view, almost every statistical method of analysis can be formalised as a function f defined on the set D of possible micro-data files, whose value $f(M)$ (often a real number or vector) is in some set W (e.g., the set of real numbers).

Now, in analogy to the jackknife estimation of standard errors, the jackknife method for confidentiality protection is based on the principal idea to calculate a number of *approximate results* $f(M_i)$ instead of the true result $f(M)$, where for each approximate result we use a slightly *modified micro-data file* M_i instead of the true file M. The file M_i is produced from M by replacing an individual value e_i at exactly one position i of the original file by a *replacement value* z_i (drawn with a pseudo-random number generator, for instance) from which e_i cannot be determined. The index i runs through all positions of the individual *values* of the original file M (that is, i is not a row but a cell index!), and the distribution of z_i is independent of e_i and sufficiently widespread. For example, this *replacement distribution* could be a uniform distribution on the domain of the corresponding variable, or a normal distribution centred at the mean and with twice the standard deviation of the variable, or a conditional distributionfrom some regression model, or derived by some imputation method, etc. When the variable can contain missing values in principle, a missing value should also be used for z_i with some probability. In addition, it is necessary in some cases to use not only one but several replacements. For instance, one could use three replacements in case of "dummy" 0-1-variables and two replacements in case of variables with three to seven possible values, so that in each case the probability that all these replacements equal the true value is at most $\frac{1}{8}$.

The set $F(M)$ of all thus computed approximate results $f(M_i)$ is the basis for what we publish. In some cases, the true result can be estimated from $F(M)$ with large certainty (e.g., if $f(M)$ is a frequency, then $F(M) = [f(M)-1, f(M)+1]$) with large probability, hence $F(M)$ cannot be published directly without revealing $f(M)$. Therefore, $F(M)$ gets moderately enlarged in some random way, giving a *publishing set* $V(M)$ which is finally published instead of the accurate result $f(M)$.

In case of metric or ordinal result values, $V(M)$ would be an *interval* $V(M) \supset F(M)$. For a metric result, a good choice for $V(M)$ seems to be the interval

$$f(M) + 4(b-a)\delta \quad \pm \quad \max\{2, 1+4a+4b\}\}\delta$$

where $\delta := \max_i |f(M_i) - f(M)|$ is the maximal error of the approximate results, and a and b are drawn independently from a $\mathrm{Beta}(2,3)$ distribution. In this way, the published interval's centre $f(M) + 4(b-a)\delta$ has a nearly normal distribution with mean $f(M)$ (so that unbiasedness is preserved), standard deviation 1.13δ and slightly non-normal kurtosis 2.68, but its maximal deviation from $f(M)$ is 4δ. In the special case of frequency tables, this is roughly comparable to adding a $\mathrm{Normal}(0, 1.13^2)$-distributed noise. Note that, even if $a = b = 0$, the published interval's width is at least 4δ, so that the snooper cannot infer $f(M)$ even if he guesses δ.

Although most analysis results are real numbers or vectors, or are at least on an ordinal scale, the method also works for non-ordinal results (e.g., the mode of a categorical variable). In this case, one could publish a set $V(M) \supset F(M)$ of less than thrice the cardinality of $F(M)$, by adding a number of $2|F(M)|-2$ values which are drawn independently and with replacement from the uniform or marginal distribution on all possible result values.

2.4. Mechanism of confidentiality protection

The rationale behind using $F(M)$ to estimate how much imprecision is sufficient for confidentiality protection is that, in this way, the published result is not only compatible with the true data, but also with data in which just the single value the snooper might be interested in was replaced by a random value. Thus the snooper should not be able to distinguish between the true and the replaced value.

Let us assume that the data snooper is interested in some individual *target value* e_t in the micro-data. We can formalise his additional knowledge by assuming he knows that $M \in A$, where A is the (usually infinite) set of all possible micro-data files which are not in conflict with his additional knowledge. If $f(M)$ was published, the snooper could try to calculate e_t from $f(M)$ using his additional knowledge. In principle, this corresponds to determining the pre-image $U := f^{-1}(f(M))$, then computing the intersection $S := U \cap A$, and finally determining the set E_t of all values e'_t occurring at the *target position* t in some of the micro-data files $M' \in S$. The snooper would then know at best that $e_t \in E_t$. The risk of (attribute) disclosure consists in the fact that, given sufficiently detailed additional knowledge, that is, given that A is sufficiently small, the set S may contain only files M' in which $e'_t = e_t$, so that the set E_t would only contain the true value e_t.

Now assume that, following the jackknife method, we publish the set $V(M)$ instead of $f(M)$, and that the snooper tries an attribute disclosure as above. Then he gets a considerably larger pre-image $U := f^{-1}(V(M))$ which contains all the files M_i (but might not contain M). Even in the extreme case where the snooper would already know *all* individual values of M other than e_t, the only file in U compatible to this knowledge is M_t, from which he can at best determine the replacement value z_t, but this tells him nothing about e_t.

Using an alternative strategy, the snooper might also try to determine a set of micro-data files not containing M but some specific replacement file M_j with $j \neq t$, since this file *would* contain the true value e_t instead of z_t. In order to study this strategy, we can again formulate the additional knowledge of the snooper as $M_j \in \tilde{A}$. But because M_t and M_j only differ in the two positions j and t, and since the snooper has no knowledge about the randomly chosen values z_t and z_j, we can conclude that $M_t \in \tilde{A}$, i.e., the snooper cannot distinguish M_t from M_j. Again he cannot determine whether e_t or z_t is the true value, even when he already knows all values except e_t.

For a metric result $f(M)$ and $V(M) = f(M) + 4(b-a)\delta \pm \max\{2, 1+4a+4b)\}\delta$ with $a, b \sim \mathrm{Beta}(2,3)$, as it was suggested above, one can show that, with at least 91% probability, the same interval $V(M)$ had been published if z_t instead of e_t had been the true value and if δ would not be affected by this change. For the usually vast majority of target positions t for which $|f(M_t) - f(M)| < \frac{\delta}{3}$, this probability increases to at least 99%.

Despite these considerations, a formal proof of the sketched protection mechanism remains to be found.

3. Performance and simple examples

The effort needed to apply the described jackknife method to some specific kind of analysis depends on how much the computation of all the $f(M_i)$ costs. Many statistical analyses can be implemented in a way which makes the effort of determining the difference $\delta_i := f(M_i) - f(M)$ independent of N, the number of observations, so that the total effort is some small constant multiple of that needed to compute $f(M)$. For example, this is the case for statistics based on moments or power-sums, like frequencies, sums, square sums, means, standard deviations, (co-)variances, (partial) product-moment correlations, skewness, kurtosis, Cronbach's alpha, test statistics of t- and F-tests, all kinds of table statistics for (fixed size) contingency tables, etc.

Alternatively, if the δ_i are known to fulfil some upper bound, this bound can also be used directly to compute an interval $V(M)$. This is closely related to basic concepts from the theory of robustness (see [Hampel 1986]). If the gross error sensitivity γ^* of f is finite, then δ_i is asymptotically bounded by $2\gamma^*/N$. Otherwise, it usually has asymptotic upper confidence limits of order $\ln(N)/N$. This indicates that the width of $V(M)$ will be of smaller order than the standard error. Consequently, the jackknife method can be expected to clearly out-perform anonymisation methods for large N.

For example, let $f(M)$ be the sample mean of a sample of N values, drawn independently from the standard normal, and let the replacement values also come from that distribution. Then $\delta < \ln(N)\frac{\sqrt{2}}{N}$ with probability at least 95% for all $N \geqslant 15$, and some further calculation shows that $V(M)$'s centre $f(M) + 4(b-a)\delta$ differs from $f(M)$ by at most $16\ln(N)/5N$ with probability at least 90% whenever $N \geqslant 15$. When we compare this to the 90% confidence limit of the sampling error of $f(M)$, which is about $2/\sqrt{N}$, we find that for $N \geqslant 30$, the confidence limit of the additional imprecision is smaller than that of the sampling error, while for $N < 30$, the former is still at most 1.18 times the latter. This shows that even for the extremely non-robust sample mean, the imprecision that is introduced additionally by the jackknife method is acceptable when compared to the sampling error.

Here are some examples in which an upper bound for δ can be used to construct $V(M)$ without actually computing the $f(M_i)$:

- Order statistics: $f(M) = x_{(k)}$, $\delta \leqslant \max\{x_{(k)} - x_{(k-1)}, x_{(k+1)} - x_{(k)}\}$.
- k-times trimmed mean: $\delta \leqslant \max\{x_{(N-k+1)} - x_{(k+1)}, x_{(N-k)} - x_{(k)}\}/N$.
- Kendall's and Spearman's rank correlation: $\delta \leqslant \frac{6}{N-3}$ resp. $\delta \leqslant \frac{6}{N-1}$.
- Sign test: $f(M) = (N^+ - N^-)/2$ (which is of order N), $\delta \leqslant 1$.
- Wilcoxon's signed rank test: $f(M) = \sum\{\text{rank}(|x_i - \mu_0|) : x_i > \mu_0\}$ (of order N^2), $\delta \leqslant N$.
- Wilcoxon's test for two samples: $f(M) = \sum\{\text{rank}(x_i) : x_i \in \text{sample 1}\}$ (of order N^2), $\delta \leqslant \max\{N_1, N_2\}$.
- Kolmogorov–Smirnov test of fit: $f(M) = \sup_x |F_N(x) - F(x)|$ (F_N being the empirical distribution function), $\delta \leqslant 1/N$.
- Bowker's test for $R \times R$-tables: $f(M) = \sum\sum_{i<j}(n_{ij} - n_{ji})^2/(n_{ij} + n_{ji})$ (of order N), $\delta \leqslant 4(2R-3)$.
- Entropy: $f(M) = \sum_i \frac{n_i}{N}\log_2\frac{n_i}{N}$, $\delta \leqslant 2\frac{\log_2 N}{N}$.
- Greenwood's G (Sum of Squares of Spacings): $f(M) = \sum_i(x_{(i)} - x_{(i-1)})^2$, $\delta \leqslant \max\{\max_i(x_{(i)} - x_{(i-1)})^2/2, 2\max_i(x_{(i+1)} - x_{(i)})(x_{(i)} - x_{(i-1)})\}$.

4. Implementation for univariate statistics and contingency table statistics

For most of the above-mentioned statistics, the Federal Statistical Office Germany has implemented the jackknife method prototypically as SAS® macros. These macros `%jk_means` and `%jk_freq` provide essentially the same functionality (and similar syntax) as the original SAS procedures means and freq, with some additional robust statistics.

```
%jk_means ( data = dataset, where = optional condition,
            by = optional classifying variables,
            var = analysis variables,
            weight = optional weight variable,
            stats = requested statistics,
            jk_cntl = control dataset )

%jk_freq  ( data = dataset, where = optional condition,
            by = optional classifying variables,
            row = row variable, col = column variable,
            jk_cntl = control dataset )
```

Their basic syntax is with some additional advance d options.

The *control dataset* specifies the re placement distributions for the variables in a certain way. `%jk_means` currently reports intervals for these statistics:

`N, SumWgt`	No. of observations and sum of weights
`Mean, StdErr, LCLM, UCLM`	Mean with standard error and confidence limits
`Sum, USS, CSS`	Sum and [un]corrected square sum
`StdDev, LCLStd, UCLStd`	Standard deviation with confidence limits
`Var, CV`	Variance, coefficient of variation
`T, ProbT`	t-test for $mean = \mu_o$, with p-value
`Skew, Kurt`	Skewness and kurtosis
`Min, Max, Range`	Extremes and their difference
`Q1, Q3, QRange`	Quartiles and their difference
`P1, P5, P10, P90, P95, P99`	Further percentiles
`Median, Biweight, Trimean`	Robust location estimators
`MAD`	Median absolute deviation from the median
`QSkew, MSkew`	Bowley's and Pearson's measures of skewness
`H10Skew, H5Skew, H1Skew`	Hinkley's robust measures of skewness
`KurtB, M5Kurt, CS5Kurt`	Some robust measures of kurtosis (see [Blest 2003])

`%jk_freq` currently computes:

χ^2-tests (classical, likelihood-ratio, continuity-adjusted, Mantel-Haenszel)
Derived statistics (Phi, contingency coefficient, Cramer's V)
Measures of association with asymptotic tests (Gamma, τ_b, τ_c, Somers' D)
Measures of association with asymptotic confidence limits
(Spearman, [a]symmetric Lambda and uncertainty coefficients)
Tests for agreement or trend (Bowker, Kappa coefficient, Cochran-Armitage)
p-values for all tests (one- and two-sided)

For all reported values, the published interval's centre is an unbiased and consistent estimator of the precise sample value of the statistic. Note that anonymisation methods, in contrast, do not usually guarantee unbiasedness.

5. Example: non-linear OLS regression

A number of statistical analyses, such as methods of model fitting, are based on parameter estimation by numerical optimisation. Although the effect of replacing a single e_t by z_t on the estimators found by such algorithms cannot in general be determined exactly without repeating the optimisation, it (or some upper bound for it) can still often be estimated quite accurately, for example by using the first and second derivatives of the objective function at the found optimum. Assume that ϑ is a k-dimensional vector of real parameters, $L(M;\vartheta)$ is the objective function (e.g., least-squares loss), and $\frac{\partial}{\partial\vartheta}L(M;\vartheta_{\text{opt}})=0$. Now, under certain smoothness assumptions, the Theorem on implicit functions implies that, when $\tilde{M}$ is sufficiently near M, then $\frac{\partial}{\partial\vartheta}L(\tilde{M};\tilde{\vartheta}_{\text{opt}})=0$ for some $\tilde{\vartheta}_{\text{opt}}$ which fulfils

$$\tilde{\vartheta}_{\text{opt}}-\vartheta_{\text{opt}}\approx\left(\frac{\partial^2}{\partial\vartheta^2}L(M;\vartheta_{\text{opt}})\right)^{-1}\frac{\partial}{\partial\vartheta}L(\tilde{M};\vartheta_{\text{opt}}).$$

That is, the change of the estimated parameters due to a small change in the data is approximately the inverse Hessian matrix of the objective function times the gradient of the objective function at the original estimators but with the new data. Using this approximation, δ can be estimated quickly in $O(Nk^2)$ time. Although this estimation of δ is somewhat less thorough than the exact computation, additional confidentiality protection arises from the fact that from such numerical optimisation results $f(M)$, it is even more difficult to determine the pre-image $U=f^{-1}(f(M))$ than in case of other kinds of analysis.

The above technique is used in the macro `%jk_nlin` which reports parameter estimates for non-linear least-squares regression:

```
%jk_nlin( data = dataset, where = optional condition,
          model = model equation without error term,
          parms = parameters with start values,
          jk_cntl = control dataset )
```

As in the SAS procedure `nlin`, also probit models (and analogously logit and complementary log-log models) can be estimated by specifying as model equation

$$0=\sqrt{-2\ln\begin{cases}\Phi(h) & \text{if } y=0\\ 1-\Phi(h) & \text{if } y=1\end{cases}},$$

where Φ is the standard normal distribution function, y is the dependent variable and h is some function of the predictors. This is because minimising the least-squares loss of this model corresponds to maximising the (log-)likelihood of the actual probit model.

6. Status and further steps

At the moment, the Research Data Centres of the Federal Statistical Office Germany and the Statistical Offices of the Länder offer researchers who want to analyse confidential data three ways of access: they can use Scientific Use Files, or come to one of the offices and work at so-called “safe scientific workstations”, or submit program code for manual execution (“controlled remote data processing”). For the latter two ways of access, confidentiality protection is performed manually on a per-case basis, using traditional protection methods as far as possible.

The jackknife method of confidentiality protection is *not* yet used for any requests from researchers. Currently, the Federal Statistical Office is trying to evaluate the practical quality of results produced with the jackknife method and compare them with results from anonymised business-data files from the project “De-facto anonymisation of business micro-data” (see [Ronning et al. 2005]), and later on also with anonymised household-data files, both for small and large N .

We plan to proceed to prototypically implement the method for further types of analyses like (partial) correlations, principal components analysis, forecasting, plots, ANOVA, etc. For evaluation purposes, all prototypes are/will be available to experts on confidentiality protection, upon request.

The main task, however, will be to find a thorough proof of the conjectured level of protection. We would be grateful for any helpful comments in this direction.

Our hope is that, eventually, the method could be integrated into a remote access facility for micro-data from German official statistics, so that researchers would be able to comfortably perform many kinds of statistical analyses with confidential data at their home office and still get high-quality results.

References

Blest, D. C. (2003), “A new measure of kurtosis adjusted for skewness”, Australian and New Zealand Journal of Statistics, 45 (2) 175–179.

Hampel et al. (1986), Robust Statistics, Wiley.

Heitzig, J. (2004) “Protection of Confidential Data when Publishing Correlation Matrices”, in: Proceedings in Computational Statistics (16th COMPSTAT Symposium), 1163–1170.

Ronning et al. (2005), Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten (Statistik und Wissenschaft, Vol. 4), Statistisches Bundesamt, Wiesbaden.

Testing variants of minimum distance controlled tabular adjustment

*Jordi Castro**[1], Sarah Giessing***

*** Department of Statistics and Operations Research, Universitat Politècnica de Catalunya Jordi Girona 1–3, 08034 Barcelona, Catalonia, Spain (jordi.castro@upc.edu)**

**** Federal Statistical Office of Germany, 65180 Wiesbaden, Germany (sarah.giessing@destatis.de)**

Abstract. Controlled tabular adjustment (CTA), and its minimum distance variants, is a recent methodology for the protection of tabular data. Given a table to be protected, the purpose of the method is to find the closest one that guarantees the confidentiality of the sensitive cells. This is achieved by adding slight adjustments to the remaining cells, preferably excluding total ones, whose values are preserved. Unlike other approaches, this methodology can efficiently protect large tables of any number of dimensions and structure. In this work, we test some minimum distance variants of CTA on a close-to-real data set, and analyze the quality of the solutions provided. As another alternative, we suggest a restricted CTA (RCTA) approach, where adjustments are only allowed in a subset of cells. This subset is a priori computed, for instance by a fast heuristic for the cell suppression problem. We discuss benefits of RCTA, and suggest several approaches for its solution.

1. Introduction

Data collected within government statistical systems must be provided as to fulfill requirements of many users differing widely in the particular interest they take in the data. For data in tabular form, this implies that most tables made publicly available belong to a system of multiple, hierarchically structured, overlapping tables which are all publicly available. Usually, some cells of these tables contain information on single, or very few respondents. Especially in the case of establishment data, given the meta information provided along with the cell values (typically: industry, geography, size classes), those respondents could be easily identifiable. Therefore, measures for protection of those data have to be put in place. Traditionally, agencies suppress part of the information (cell suppression). Efficient algorithms for cell suppression are offered f.i. by the software package τ -ARGUS (Hundepool et al., 2004). Cell suppressions, however, must be coordinated between tables. This implies certain restrictions on the release of tabular data which is in some contrast to the flexibility and capacity of modern (OnLine) Data Base systems. Cell perturbation, as alternative to, or in combination with cell suppression may offer a way out of the dilemma.

Minimum distance controlled tabular adjustment (or CTA for short) (Dandekar and Cox, 2002; Castro, 2006) is a recent technique to generate synthetic, i.e. perturbed values that may be used to replace original entries of tables provided for a publication. Although CTA is very efficient from a computational point of view, NSAs are still reluctant to use it, because offering synthetic data might be in conflict to their responsibility to produce data that are ’as accurate as possible’. In order to introduce CTA into practice, it is therefore essential to prove that data sets protected by CTA can provide a *sufficient* amount of *accurate* information, compared to the standards set by cell suppression. Instead of considering how to preserve second order statistics, like variance and covariance, proposed in Cox et al. (2004), in this paper we focus on the following simple criteria for a robust CTA that allow comparison to, or combination with cell suppression to some extent:

[1] Work supported by the Spanish MCyT project TIC2003-00997.

- The number of cells with a large relative deviation (i.e., over 5%, 10%, or any other predefined threshold value) should be as low as possible (hopefully, zero). Such large deviations are in some sense equivalent to the suppression of the cell, which is exactly the technique we plan to replace by using CTA.
- Cells that provide aggregated information on a high level (for geography, for instance, state, or whole country level), should remain unchanged, or only slightly modified.
- CTA should be able to provide a feasible solution if deviations are only allowed in a reduced subset of cells. For instance, this enables to filter through CTA data previously protected by other techniques like cell suppression: in this case the suppressed cells would be the subset of cells allowed for deviations, as suggested in Giessing (2004).

The structure of the paper is as follows. Section 2 sketches the minimum distance CTA family of methods. Section 3 reports and analyzes the results obtained with some close-to-real instances. In Section 4 we discuss a restricted CTA procedure, which improves the quality of the protected tables, although it significantly increases the solution time. Some strategies are discussed for the efficient solution of the restricted CTA procedure.

2. Outline of minimum distance controlled tabular adjustment

Any problem instance, either with one table or a number of tables, can be represented by the following elements:

- A set of cells $a_i, i = 1, \ldots, n$, that satisfy some linear relations $Aa = b$ (a being the vector of a_i's).
- A lower and upper bound for each cell $i = 1, \ldots, n$, respectively $\underline{a}_i$ and $\overline{a}_i$, which are considered to be known by any attacker. If no previous knowledge is assumed for cell i $\underline{a}_i = 0$ ($\underline{a}_i = -\infty$ if $a \geq 0$ is not required) and $\overline{a}_i = +\infty$ can be used.
- A set $\mathcal{P} = \{i_1, i_2, \ldots, i_p\} \subseteq \{1, \ldots, n\}$ of indices of conf1dential cells.
- A lower and upper protection level for each confidential cell $i \in \mathcal{P}$, respectively lpl_i and upl_i, such that the released values satisfy either $x_i \geq a_i + upl_i$ or $x_i \leq a_i - lpl_i$.

CTA attempts to find the closest safe values $x_i, i = 1, \ldots, n$, according to some distance L, that makes the released table safe. This involves the solution of the following optimization problem:

$$\begin{array}{rl} \min\limits_{x} & ||x - a||_L \\ \text{subject to} & Ax = b \\ & \underline{a}_i \leq x_i \leq \overline{a}_i \quad i = 1, \ldots, n \\ & x_i \leq a_i - lpl_i \text{ or } x_i \geq a_i + upl_i \quad i \in \mathcal{P}. \end{array} \tag{1}$$

Problem (1) can also be formulated in terms of deviations from the current cell values. Defining $z_i = x_i - a_i,\quad i = 1,\dots,n$ —and similarly $\underline{z}_i = \underline{x}_i - a_i$ and $\overline{z}_i = \overline{x}_i - a_i$—, (1) can be recast as:

$$\begin{array}{ll} \min\limits_{z} & ||z||_L \\ \text{subject to} & Az = 0 \\ & \underline{z}_i \le z_i \le \overline{z}_i \quad i = 1,\dots,n \\ & z_i \le -lpl_i \text{ or } z_i \ge upl_i \quad i \in \mathcal{P}, \end{array} \tag{2}$$

$z \in \mathrm{IR}^n$ being the vector of deviations.

It has been observed that the best quality solutions are obtained with the L_1 and L_2 distances (Castro, 2006). Using the L_1 distance, and after some manipulation, (2) can be written as

$$\begin{array}{ll} \min\limits_{z^+,z^-} & \sum\limits_{i=1}^{n} w_i(z_i^+ + z_i^-) \\ \text{subject to} & A(z^+ - z^-) = 0 \\ & 0 \le z_i^+ \le \quad \overline{z}_i \quad i = 1,\dots,n \\ & 0 \le z_i^- \le -\underline{z}_i \quad i = 1,\dots,n \\ & \left\{ \begin{array}{lcl} z_i^+ & \ge & upl_i \\ z_i^- & = & 0 \end{array} \right\} \text{ or } \left\{ \begin{array}{lcl} z_i^- & \ge & lpl_i \\ z_i^+ & = & 0 \end{array} \right\} \quad i \in \mathcal{P}, \end{array} \tag{3}$$

z^+ and z^- being the vector of positive and negative deviations in absolute value. For L_2, we have

$$\begin{array}{ll} \min\limits_{z} & \sum\limits_{i=1}^{n} w_i z_i^2 \\ \text{subject to} & Az = 0 \\ & \underline{z}_i \le z_i \le \overline{z}_i \quad i = 1,\dots,n \\ & z_i \le -lpl_i \text{ or } z_i \ge upl_i \quad i \in \mathcal{P}. \end{array} \tag{4}$$

Combinations of L_1 and L_2 were tested in Castro (2004).

In practice the sense for the “or” constraint is heuristically fixed a priori (Dandekar and Cox, 2002). In the computational results of Section 3 we set the “upper level protection” for all the sensitive cells. This can lead to infeasible problems, as it will be discussed in Section 4. An alternative that overcomes the infeasibility at the expense of increasing the computational complexity, is to include the “or” decision within the mathematical model (1), adding a binary variable y_i and two extra constraints for each confidential cell:

$$\begin{array}{lcll} x_i & \ge & -M(1-y_i) + (a_i + upl_i)y_i & i \in \mathcal{P}, \\ x_i & \le & My_i + (a_i - lpl_i)(1-y_i) & i \in \mathcal{P}, \\ y_i & \in & \{0,1\} & i \in \mathcal{P}, \end{array} \tag{5}$$

M in (5) being a large value. In terms of deviations, the equivalent constraints for the L_1 model (3) are

$$\begin{array}{rcccll} upl_i y_i & \le & z_i^+ & \le & My_i & i \in \mathcal{P}, \\ lpl_i(1-y_i) & \le & z_i^- & \le & M(1-y_i) & i \in \mathcal{P}, \\ y_i & \in & \{0,1\} & & & i \in \mathcal{P}; \end{array} \tag{6}$$

and for the L_2 model (4) we should add

$$\begin{array}{rcll} z_i & \geq & -M(1-y_i) + upl_i y_i & i \in \mathcal{P}, \\ z_i & \leq & M y_i - lpl_i(1-y_i) & i \in \mathcal{P}, \\ y_i & \in & \{0,1\} & i \in \mathcal{P}. \end{array} \tag{7}$$

The above constraints result in a combinatorial optimization problem, which is discussed in Section 4.

3. Computational testing

From the perspective of a data provider, it is essential to avoid that in the released table there are large deviations in cells that provide aggregated information on a high level, and at the same time we want to keep the number of cells with large relative deviations (e.g., over 5% or 10%) low. These are contradictory objectives. Large absolute deviations in total cells are avoided if we choose cell weights $w_i = 1$ in (3) or (4). On the other hand, relative deviations are kept small for $w_i = 1/a_i$ (if $a_i = 0$ the cell can not be perturbed, and we set w_i to any value, e.g., 1). Both weights belong to the family $w_i = 1/a_i^\gamma$, for$\gamma = 0$ and $\gamma = 1$. Weights with $\gamma = 0.5$ are also a reasonable choice, since in theory they should balance relative and absolute deviations.

Table 1. Dimensions of the complex instances

Name	n	$\lvert\mathcal{P}\rvert$	m	N.coef
bts4	36570	2260	36310	136912
destatis	5940	621	1464	18180
five20b	34552	3662	52983	208335
five20c	34501	4022	58825	231345
hier13	2020	112	3313	11929
hier16	3564	224	5484	19996
nine12	10399	1178	11362	52624
nine5d	10733	1661	17295	58135
ninenew	6546	858	7340	32920
two5in6	5681	720	9629	34310

We tested the three weights for $\gamma = 0,1/2,1$ and the L_1 and L_2 distances with a set of complex instances: the seven most complex instances used in (Dandekar, 2003; Castro, 2006) (named “bts”, “hier13”, “hier16”, “nine12”, “nine5d”, “ninenew”, and “two5in6”) which seem to present frequency counts, and a close-to-real instance provided by Destatis (named the “destatis” instance in the following). The latter instance represents a tabulation of a strongly skewed variable (like “turnover”, f.i.), typical for business statistics. We also attempted the recently released “five20b” and “five20c” twenty-dimensional tables (Dandekar, 2005). However, unlike the former, which are solved in seconds, these two instances are computationally challenging. For instance, the protection procedure was stopped after 10 hours of CPU time without a solution for “five20b”, using either the dual or primal simplex algorithm of Cplex 9.1 on a Pentium-4 at 1.8GHz; “five20c” was not attempted with simplex algorithms. On the other hand, interior-point algorithms seem to be a more efficient choice for large multidimensional tables. For instance, the interior-point option of Cplex 9.1 protected the “five20b” and “five20c” instances in, respectively, 10 and 20 minutes of CPU using the L_1 distance, and 5 and 10 minutes of CPU using the L_2 distance. In principle there is room for improvement using specialized interior-point methods, as done for three-dimensional tables in Castro (2005). Table 1 provides the dimensions of each instance: number of cells (column “n”), number of sensitive cells (column

“ $|\mathcal{P}|$ ”), number of constraints (column “ m ”), and number of nonzeros in constraints matrix (column “N.coef”). Table 2 shows the number of cells with relative deviations between 2% and 5% and over 5% for each value γ . It is observed that, in general, the number of cells with large relative deviations increases when γ tends to zero. Another observation is that for the business data instance the choice of the cost function seems to have a stronger effect as with the other instances.

Table 2. Number of cells with a relative deviation between 2% and 5% (a)), and greater than 5% (b)), for $\gamma = 0,1/2,1$ and the complex instances

	$\gamma = 0$		$\gamma = 1/2$		$\gamma = 1$	
Instance	L_1	L_2	L_1	L_2	L_1	L_2
bts4	1402	1515	1016	1184	962	933
destatis	164	396	125	416	119	309
five20b	2841	3013	2478	2815	2426	2605
five20c	3218	3477	2769	3096	2777	2822
hier13	101	103	75	82	79	68
hier16	127	145	108	124	112	95
nine12	787	889	685	787	695	709
nine5d	875	999	947	993	978	918
ninenew	613	646	521	598	531	510
two5in6	451	529	388	499	424	384

a) relative deviation between 2% and 5%

	$\gamma = 0$		$\gamma = 1/2$		$\gamma = 1$	
Instance	L_1	L_2	L_1	L_2	L_1	L_2
bts4	741	799	353	521	279	292
destatis	352	1012	11	524	7	70
five20b	1284	1434	650	1161	445	579
five20c	1352	1542	699	1202	559	706
hier13	32	32	26	27	26	24
hier16	60	69	29	46	17	112
nine12	378	427	162	310	120	149
nine5d	606	724	223	523	163	128
ninenew	298	360	154	258	107	131
two5in6	244	80	128	163	90	86

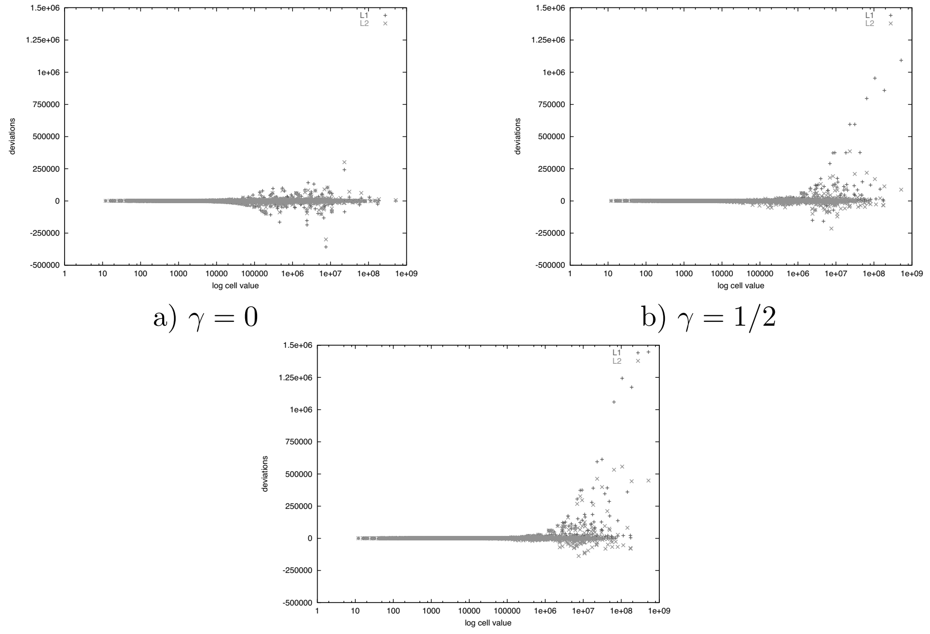
b) relative deviation greater than 5%

In the following, we analyze in more detail this instance “destatis”. It is a 3 dimensional table where one of the 3 variables is hierarchical with 3 levels. Plots a), b) and c) of Figure 1 show the deviations obtained for the cell values (in log scale). As expected the pattern for $\gamma = 0$ provides the lowest variability, and most deviations concentrate around 0. The number of cells by ranges of relative deviations is shown in Table 3. From that table it is clear that $\gamma = 0$ gives the greatest number of cells with large relative deviations. The opposite behaviour is observed for $\gamma = 1$. For $\gamma = 1/2$ we get a small number of cells with large relative deviations, although, from Figure 1, deviations are still fairly large for the highest-valued cells, mainly for L_1.

Table 3. N. of cells by ranges of relative deviation for $\gamma = 0,1,1/2$ for “destatis” instance

	$\gamma = 0$		$\gamma = 1/2$		$\gamma = 1$	
Range	L_1	L_2	L_1	L_2	L_1	L_2
0%	2164	0	2407	0	2439	0
(0%,2%]	540	1812	677	2280	655	2841
(2%,5%]	164	396	125	416	119	309
(5%,10%]	78	233	7	195	4	61
(10%,100%]	274	779	4	329	3	9

Figure 1. Deviations for a) $\gamma = 0$, b) $\gamma = 1/2$ and c) $\gamma = 1$ in the “destatis” instance



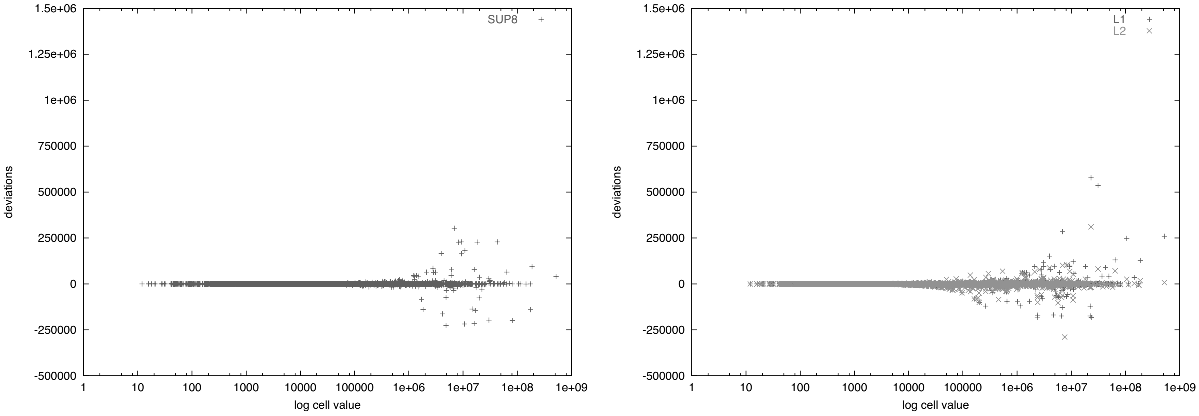
a) $\gamma = 0$

b) $\gamma = 1/2$

c) $\gamma = 1$

As a compromise closer to $\gamma = 0$ we considered weights $w_i = 1/\log a_i$, both for L_1 and L_2; corresponding results are shown in Figure 2.b. Another alternative approach has been suggested in Giessing (2004), a heuristic implementation of a ’restricted CTA’ (RCTA) procedure which is presented in the following section 4. Table 4 proves that this particular RCTA heuristic, referred to as SUP8 in the figures, outperforms the CTA variant with weights $w_i = 1/\log a_i$ in the sense that it reduces the number of cells with a relative deviation beyond 10% from 98 (for L_1; 709 for L_2) to 1. Comparison of Figures 2.a (referring to SUP8) and 2.b shows that large changes in large values are also prevented more efficiently as by the L_1 variant.

Figure 2. a) Deviations for SUP8. b) Deviations for weights $w_i = 1/\log a_i$, for L_1 and L_2



a) SUP8

b) $w_i = 1/\log a_i$

Table 4. N. of cells by ranges of relative deviation for $w_i = 1/\log a_i$ and SUP8

	$w_i = 1/\log a_i$		SUP8
Range	L_1	L_2	
0%	2300	0	2341
(0%,2%]	644	1857	641
(2%,5%]	136	402	169
(5%,10%]	42	252	88
(10%,100%]	98	709	1

However, the patterns of Figures 1 and 2 only give a first impression of the performance with respect to the quality issue we are actually interested in, e.g. that cells on a high level of aggregation should remain unchanged, or be only slightly modified. Most of these cells are among the cells with the largest values, but some are not. A more direct approach to achieve the goal of small deviations for high-level cells is to choose the parameter γ adaptively according to the cell hierarchy, such that cells with large hierarchies (i.e., national cells) have γ close to 0 (i.e., absolute deviations minimized), and low hierarchy cells have γ close to 1 (i.e., relative deviations minimized). Assuming that $h_i, i = 1, \ldots, n$ gives the hierarchy of cell i, and that $\overline{h} = \max\{h_i, i = 1, \ldots, n\}$ the rule considered was

$$\gamma_i = \frac{(\overline{h} - h_i)}{\overline{h}}.$$

Figure 3 shows the deviations by cell value for these adaptive γ values. We observe that the adaptive γ outperforms $\gamma = 1$ and $\gamma = 1/2$, and provides deviations closer to those obtained with $\gamma = 0$. As for the relative deviations, Table 5 reports the number of cells by ranges of relative deviations. The adaptive γ provides better results than $\gamma = 0$, but the number of cells with large relative deviations is still greater than for $\gamma = 1$.

Figure 3. Deviations for adaptive γ according to cell hierarchies for L_1 and L_2

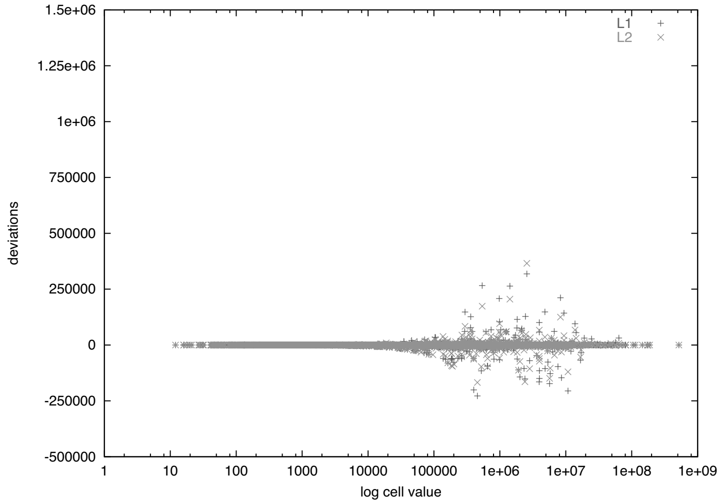


Table 5. Number of cells by ranges of relative deviation for adaptive γ

Range	L_1	L_2
0%	2320	0
(0%,2%]	577	2233
(2%,5%]	124	423
(5%,10%]	63	223
(10%,100%]	136	341

We imagine now that data providers request that on the top levels of a hierarchical table, CTA should present as many *reliable* results as cell suppression. For such highly aggregated data, even a change of 1% is usually considered far too much. For our instance "destatis" we consider as top levels the 2 top levels of the hierarchical variable which are *inner* cells with respect to at most one of the non-hierarchical variables. Within this set of 111 cells, the modular method of τ-ARGUS selects 20 secondary suppressions. For the following analysis, we consider a high-level cell value a as *changed too much for publication*, when the amount of change exceeds $\sqrt{a}$. With this concept, only adaptive γ for L_1 leads to an *acceptable* result: 11 cells change *too much*, while all other CTA variants lead to more than 20 cells *lost* because they lack precision (see Table 6).

Table 6. Number of high level cell values changed too much for publication

$\gamma = 0$		$\gamma = 1/2$		$\gamma = 1$		adaptive γ		$w_i = 1/\log a_i$		SUP8
L_1	L_2	L_1	L_2	L_1	L_2	L_1	L_2	L_1	L_2	
33	65	76	86	83	82	11	28	48	66	40

In the next section we present ideas to combine cell suppression and CTA methodology which may turn out to be of special interest in the context of protecting linked tables.

4. The restricted CTA method

Large relative deviations, independently of the value γ used for weights, can be avoided by imposing constraints

$$(1-\alpha_i)a_i \leq x_i \leq (1+\beta_i)a_i \quad i = 1,\dots,n, \tag{8}$$

for some $\alpha_i, \beta_i \geq 0$, to the general model (1), or, equivalently,

$$\begin{array}{ll} 0 & \leq z_i^+ \leq \beta_i a_i \qquad i = 1,\dots,n \\ 0 & \leq z_i^- \leq \alpha_i a_i \qquad i = 1,\dots,n \end{array} \tag{9}$$

for the L_1 model (3), and

$$-\alpha_i a_i \leq z_i \leq \beta_i a_i \quad i = 1,\dots,n \geq 0, \tag{10}$$

for the L_2 model (4). The parameters $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$ bound the relative deviations on cell values. Imposing, e.g., $\alpha_i = \beta_i = 0.05$ for all $i = 1,\dots,n$ we avoid relative deviations larger than 5%. Imposing $\alpha_i = \beta_i = 0.0, i \in F$ for some subset of cells F, we guarantee that cells of F will remain unchanged in the protected table. Such a set could f.i. be the set of cells a table has in common with another table that has already been protected in the case of linked tables. In the procedure SUP8 presented in Giessing (2004) this set has been determined by a fast heuristic for the cell suppression problem, i.e. the GHMITER hypercube algorithm (Repsilber, 2002; Giessing, 2003) considering +/-8% a priori bounds on the cell values. For the CTA step deviations in this subset of cells were allowed with at most $\alpha_i = \beta_i = 0.09$. For the cost function we used weights with $\gamma = 1$ and L_1 distances. The resulting procedure is more restrictive than the original CTA method, since deviations are only allowed in some cells, and such deviations are confined within some bounds. We call the new procedure the Restricted Controlled Tabular Adjustment (RCTA for short).

The main benefit of RCTA is that we can precisely control through constraints, instead of through the weights, the relative deviations of the cells. The drawback is that small values for α_i and β_i result in infeasible problems, at least if the sense of protection ("upper" or "lower") is a priori fixed. For instance, imposing $\alpha_i = \beta_i = 0$ in the subset of cells previously computed by the GHMITER hypercube heuristic for cell suppression, instance "destatis" becomes infeasible, naturally, when we use the "upper protection sense" for all primary cells. Even when we allow deviations in all cells with $\alpha_i = \beta_i = 0.1$, instance "destatis" remains infeasible using the "upper protection sense" for all primary cells. For $\alpha_i = \beta_i = 0.5$ the instance becomes feasible, again with the "upper protection sense" for all primary cells. However a 50% of relative variation is impractical.

To avoid infeasibility problems with RCTA we are forced to include in the optimization problem the binary decision for the "upper" or "lower" protection sense, either adding constraints (6) to the L_1 model (3) or adding (7) to the L_2 model (4). Unfortunately this transforms the linear and quadratic models for L_1 and L_2 to combinatorial ones, significantly increasing the solution time. For instance, for L_1 we attempted the optimization problem (3,6), using the mixed-integer-programming solver of Cplex 9.1 on a Pentium-4 at 1.8GHz. We stopped the procedure after 10 hours of CPU without a solution. The same model without the binary constraints (6) is solved in about 1 second.

Possible solution strategies to overcome the excessive time of RCTA with the binary variables are:

- Optimal solution through Bender's decomposition, moving binary decisions to a master problem, and solving a sequence of the easy continuous subproblems (3) or (4).
- Use of a heuristic for a good initial choice of the protection senses (either "lower" or "upper"). Once fixed, only one solution of either (3) or (4) is needed.

- Metaheuristic, as genetic algorithms, for adjusting the binary decisions, which involves the solution of a sequence of subproblems (3) or (4).
- The last option consists of removing the binary decisions, and to allow deviations go beyond their bounds, penalizing such bound violations in the objective function by a large penalty term. This guarantees an always feasible problem, at the expense of providing a table with some unprotected sensitive cells. Only one easy linear or quadratic problem has to be solved in that case, but some kind of post processing is eventually required to fix underprotection problems.

The SUP8 procedure of Giessing (2004) makes a heuristic choice of the protection senses (Rabenhorst, 2003), solving infeasibility problems by penalizing bound violations. In the "destatis" instance, this resulted in 3 significantly underprotected sensitive cells.

All the previous approaches are currently being investigated by the authors.

5. Summary and final conclusions

In this paper, we have compared several variants of CTA with a special focus on an instance from business statistics. Our experiments show that at least in the context of strongly skewed business data, the parameters of a CTA approach, such as the choice of a particular cost function, have considerable effect on the output data quality. Spending some effort here on fine tuning of a method seems to be worthwhile.

As CTA is discussed as an alternative to well established cell suppression, we also included a quality criterion that allows direct comparison of the performance of CTA to cell suppression, to some extent. First results are promising, indicating that it may be possible to make CTA procedures provide at least as much data meeting the high data quality standards of official statistics for data of a certain relevance as cell suppression. We also suggested restricted RCTA as an option to combine cell suppression and CTA, or to facilitate use of CTA in the context of linked tables. RCTA allows to control relative and absolute deviations more precisely than CTA. Unfortunately, RCTA is more sensible to the protection sense ("upper" or "lower") of sensitive cells than CTA, leading to infeasibility problems. Several strategies have been discussed for a proper choice of protection sense, leading to both optimal and heuristic solutions. Heuristic solutions are likely to be the best practical option, since they will provide a reasonable quality protected table within reasonable time. All these approaches for RCTA are currently under development by the authors.

References

Castro, J. (2004), Computational experiments with minimum-distance controlled perturbation methods, *Lecture Notes in Computer Science. Privacy in statistical databases* 3050, 73–86. Volume Privacy in statistical databases, J. Domingo-Ferrer and V. Torra, Springer, Berlin.

Castro, J. (2005). Quadratic interior-point methods in statistical disclosure control, *Computational Management Science* 2, 107–121.

Castro, J. (2006). Minimum-distance controlled perturbation methods for large-scale tabular data protection, *European Journal of Operational Research*, 171, 39–52 .

Cox, L. H., Kelly, J. P., and Patil, R. (2004). Balancing quality and confidentiality for multivariate tabular data, *Lecture Notes in Computer Science. Privacy in statistical databases* 3050, 87–98. Volume Privacy in statistical databases, J. Domingo-Ferrer and V. Torra, Springer, Berlin.

Dandekar, R.A. (2003), Cost effective implementation of synthetic tabulation (a.k.a. controlled tabular adjustments) in legacy and state of the art statistical data publication systems, Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Luxembourg. Available from http://www.unece.org/stats/documents/2003.04.confidentiality.htm.

Dandekar, R.A. (2005), personal communication.

Dandekar, R.A., and Cox, L.H. (2002), Synthetic tabular data: an alternative to complementary cell suppression, manuscript, Energy Information Administration, U.S. Department of Energy. Available from the first author on request (Ramesh.Dandekar@eia.doe.gov).

Giessing, S. (2003), Co-ordination of cell suppressions: strategies for use of GHMITER, Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Luxembourg. Available from http://www.unece.org/stats/documents/2003.04.confidentiality.htm.

Giessing, S. (2004), Survey on methods for tabular data protection in ARGUS, *Lecture Notes in Computer Science. Privacy in statistical databases* 3050, 1–13. Volume Privacy in statistical databases, J. Domingo-Ferrer and V. Torra, Springer, Berlin.

Hundepool, A., van de Wetering, A., Ramaswamy, R., de Wolf, P.P., Giessing, S., Fischetti, M., Salazar, J.J., Castro, J., Lowthian, P. (2004), τ-ARGUS users's manual, version 3.0.

Rabenhorst, A. (2003), Bestimmung von Intervallen und Ersatzwerten für gesperrte Zellen in statistischen Tabellen, Diploma Thesis, Manuscript, University Ilmenau (in German).

Repsilber, D.(2002), Sicherung persönlicher Angaben in Tabellendaten' - in Statistische Analysen und Studien Nordrhein-Westfalen, Landesamt für Datenverarbeitung und Statistik NRW, Ausgabe 1/2002 (in German).

Complementary Cell Suppression Software Tools for Statistical Disclosure Control - Reality Check

Ramesh A. Dandekar
Statistics and Methods Group, U. S. Department of Energy, Washington DC
(Ramesh.Dandekar@EIA.DOE.gov -http://mysite.verizon.net/vze7w8vk/)

1. Introduction

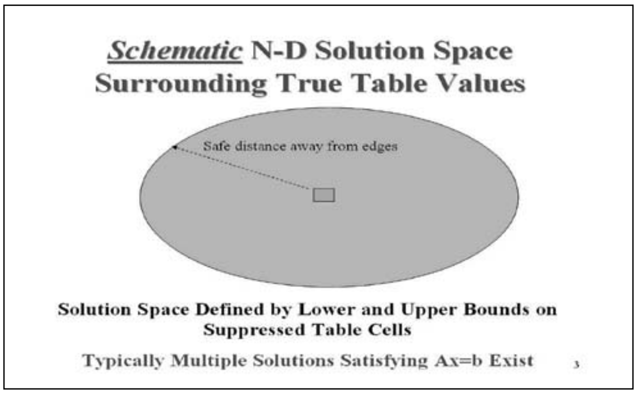
Currently, complementary cell suppression software tools are mostly used by statistical agencies to protect sensitive tabular data from disclosure. It is generally believed that the linear programming (LP) based complementary cell suppression procedures offer the best protection from wrongful disclosure of statistical information. In recent years LP-based cell suppression auditing software tools have been advocated and are being used to ensure the adequacy of protection offered by cell suppression patterns. LP-based lower and upper bounds for suppressed tabular cells are typically used to determine the adequacy of disclosure control measures. This paper identifies limitations of conclusions drawn using LP-based auditing software tools. We use widely employed analytical procedures to demonstrate the relative ease with which statistical disclosure of sensitive tabular data could occur. We conclude by providing additional safeguard measures required to avoid such disclosures.

2. Current Practice

The complementary cell suppression methods, as currently practiced by national statistical offices (NSO), enable data users to determine a multi-dimensional solution space surrounding the “incomplete” tabulation available in the public domain. Linear programming (LP) based lower and upper bounds on the withheld tabular cells are used to establish the boundaries for the solution space.

NSOs are required to ensure that the real complete table containing sensitive cells is well hidden inside the solution space a safe distance away from the edges of the solution space. The solution space typically contains multiple feasible solutions that satisfy the equality constraints associated with the complete real table structure.

Feasible solutions residing close to the edges of the solution space tend to yield poor estimates of the values of withheld cells. On the other hand, feasible solutions located away from the edges of the solution space and toward the “centroid” of the solution space tend to be of better quality and more closely resemble the hidden real complete table. ***This phenomenon has the potential to cause the disclosure of sensitive tabular data protected by complementary cell suppression methods***



Schematic **N-D Solution Space Surrounding True Table Values**

Solution Space Defined by Lower and Upper Bounds on Suppressed Table Cells

Typically Multiple Solutions Satisfying Ax=b Exist

Typically in an attempt to minimize the information loss, NSOs are under pressure to avoid over protection of sensitive tabular cells. The over protection of sensitive tabular cells results in an increase in the size of the solution space.

As per current practice, the solution space is expected to be "just right" in size. Smaller than a minimum required solution space, determined by LP-based lower and upper bounds, is known to be unacceptable. Larger than a minimum required solution space, determined by LP-based lower and upper bounds, is thought to cause unnecessary information loss. As a result, in recent years much of the efforts in tabular data protection area have been concentrated in keeping the cell suppression related solution space to a bare minimum.

3. Current Tools

Optimization Technology Center of Northwestern University and Argonne National Laboratory at http://www-unix.mcs.anl.gov/otc/Guide/faq/ describes linear programming tools as follows:
"Two families of solution techniques are in wide use today. Both visit a progressively improving series of trial solutions, until a solution is reached that satisfies the conditions for an optimum. ***Simplex methods***, introduced by Dantzig about 50 years ago, visit "basic" solutions computed by fixing enough of the variables at their bounds to reduce the constraints Ax = b to a square system, which can be solved for unique values of the remaining variables. Basic solutions represent extreme boundary points of the feasible region defined by Ax = b, x >= 0, and the simplex method can be viewed as moving from one such point to another along the edges of the boundary. ***Barrier or interior-point methods, by contrast, visit points within the interior of the feasible region. ……."***

The increased potential for statistical disclosure of the withheld sensitive tabular data is directly related to the basic property of interior-point methods to ***visit points within the interior of the feasible region,*** where the real complete table containing sensitive tabular cells resides.

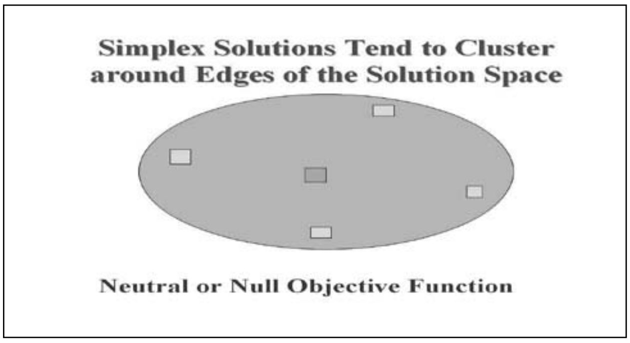
We use the following simple illustrative example supplied by Prof. Jordi Castro http://www-eio.upc.es/~jcastro/ to further clarify the difference in the working of two families of LP solvers.

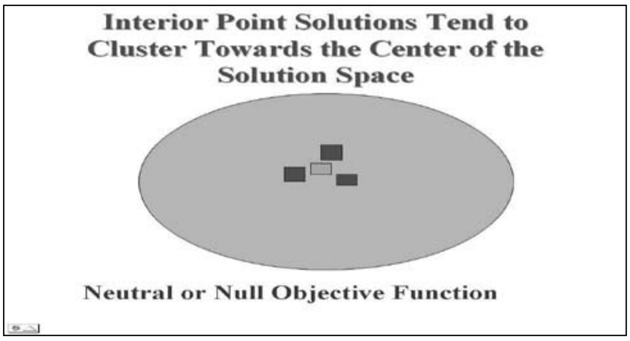
min 0

st. x1 + x2 + x3 = 3

x1, x2, x3 > = 0

Interior point methods will provide the solution x1 = x2 = x3 = 1
The simplex methods will provide some xi = 3, the other two xj = 0.


Simplex Solutions Tend to Cluster around Edges of the Solution Space
Neutral or Null Objective Function


Interior Point Solutions Tend to Cluster Towards the Center of the Solution Space
Neutral or Null Objective Function

A knowledgeable individual can easily exploit the working knowledge of interior-point methods to obtain "high quality" additive point estimates for missing tabular cells by (1) not specifying the objective function (or by using a dummy objective function) and (2) capturing the first feasible solution that satisfies the tabular data equality constraints. A moderately sized solution space, in combination with the tendency of interior point methods to the visit interior of the feasible region, will always ensure high precision estimates. These estimates are most likely to cause the statistical disclosure of withheld sensitive cells.

4. Illustrative Example

In Table 1 we have used the 3-D tabular data example from Dandekar/Cox (2002) paper available from http://mysite.verizon.net/vze7w8vk/ to illustrate the severity of the disclosure problem associated with current SDL practice. The table contains 24 sensitive cells. The table is protected by using 44 complementary cell suppressions. Table 2 shows the LP-based lower and upper bounds for the 24 sensitive cells. The p percent rule (p=10%) was used to identify the sensitive cells. Except for two minor violations for sensitive cell #6 and #18, the suppression pattern associated with the 44 complementary cells fully satisfies the current requirement for "safe table".

5. Statistical Estimation

Typically, statistical estimates for missing table cell values can be derived by using 1) additive point estimates 2) method of averages and 3) peak densities associated with frequency distributions. The last two methods, by themselves, do not provide additive tabular estimates. However, when combined with the controlled tabular adjustment (CTA) method of Dandekar/Cox (2002), the last two methods are capable of providing additive tabular estimates.

We have used the interior-point based, PCx linear programming solver available from http://www-fp.mcs.anl.gov/otc/Tools/PCx/ to illustrate the severity of the disclosure problem resulting from statistical estimates for sensitive table cells.

Table 3 provides additive point estimates for missing sensitive cells[1] by using the conventional simplex method and the PCx solver. The null-objective function was used to derive the additive point estimates. Three of the simplex estimates and 14 of the PCx estimates violate protection level for the sensitive cell causing statistical disclosure. These findings are consistent with the properties associated with the two families of solution techniques as described on the Argonne National Laboratory web site above.

Table 4 provides statistics based on averages from 138 LP solutions obtained by using the PCx software. Half of the LP solutions (sixty-nine) were for a minimization of the objective function. The remaining LP solutions were for a maximization of the objective function. Sixty-eight solutions in each group were obtained by using only one variable in the objective function. One solution in each group included all the sixty-eight variables in the objective function. Based on Table 4 statistics, sixteen of the twenty-four averages are within the prohibited protection range causing the statistical disclosure of 16 sensitive cells.

Table 5 uses the outcome from the same 138 LP solutions to generate the frequency distribution of estimates for missing sensitive cells. The table contains three lines of output for every sensitive cell. The first line in the table displays the true cell value of the sensitive cell (714 for the first sensitive cell) and the LP-based audit range (409 for the first sensitive cell).

In the next two lines we divide the audit range into ten equal intervals and summarize the frequency count resulting from the 138 LP runs. The first line shows the actual count, while the second line shows the interval values associated with the count. For the first sensitive cell, the peak density of 97 is within the sixth interval ranging from 697 to 738. The comparison of the location of the peak of the density function relative to the true cell value reveals statistical disclosure for almost all of the twenty-four sensitive cells.

[1] Space limitations prohibit us from providing values for non-sensitive tabular cells.

6. Targeting the Centroid of the Solution Space

Knowing that the real complete table is typically hidden some where in the vicinity of the centroid of the solution space, a knowledgeable individual can also use any general purpose LP solver (not necessarily interior point solver) to derive "high precision" additive point estimates for the suppressed tabular cells. Related mathematical formulation requires that each suppressed tabular cell ($X_{estimate}$) be represented by three variables in the tabular data equality constraints, namely $X_{centroid}$, Y_{plus} and Y_{minus}.

Where $X_{centroid} = 0.5 * X_{lower\_LP\_bound} + 0.5 * X_{upper\_LP\_bound}$,

$X_{estimate} = X_{centroid} + Y_{plus} - Y_{minus}$ and

Y_{plus} and Y_{minus} are minimal plus or minus corrective adjustments to ensure additivity of the tabular cells.

An individual with advanced computation skills could even go further and use either random Monte Carlo simulations or some sophisticated stratification scheme to obtain density functions (and peak density values) for the missing table cell values by using the following simple equation:

$X_{centroid} = R * X_{lower\_LP\_bound} + (1.0 – R) * X_{upper\_LP\_bound}$

Where R = Random Number between zero and one

If the individual further decides to restrict the search for the feasible solution, say to within a 10 percentile range around the centroid of the solution space, then the values for the random number could be restricted to within 0.4 and 0.6 to achieve that objective.

7. Conclusion and Recommendations

As a result of the easy access to the interior-point methods, such as PCx software tool, the LP-based lower and upper bounds of tabular data cell suppression patterns can no longer be used *alone* to judge the adequacy of the cell suppression pattern.

Conventional statistical analytical measures such as additive point estimates, method of averages and peak density values associated with frequency distributions, in combination with interior point methods, could be used with trivial efforts to cause a statistical disclosure of sensitive tabular data.

Contrary to current belief, over protection of the sensitive tabular data ***reduces*** the possibility of statistical disclosure resulting from use of interior point LP solvers. As a result, the over protection of sensitive tabular data is no longer an undesirable property of cell suppression pattern.

The current practice of using relatively small size cells as complementary suppression cells has a ***tendency*** to produce tighter LP bounds with sharp peak density functions. Therefore, this practice should be used with caution.

Use of cost functions such as reciprocal of cell value or log(cell value)/cell value to develop complementary cell suppression pattern targets large size cells. Complementary cell suppression pattern based on these functions has a ***tendency*** to produce wider protection intervals with flatter density functions. For this reason, these cost functions should be given a serious consideration.

With new technical challenges arising from the easy access to interior point methods, NSOs might want to explore the possibility of switching form the complementary cell suppression methods to other tabular data protection methods.

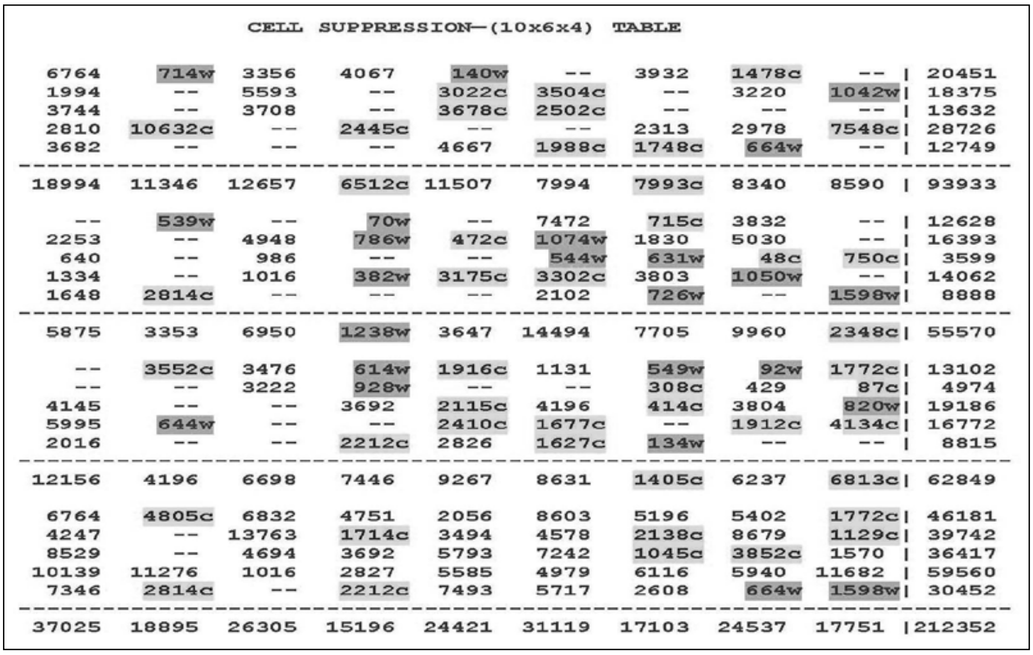
Emerging methods such as synthetic tabular data, which also is referred to as controlled tabular adjustment (CTA), offers sensitive tabular data required protection from disclosure without disclosing the solution space associated with the CTA pattern. The lack of complete information pertaining to the solution space associated with CTA pattern eliminates the possibility of the outside user deploying standardized external procedures to estimate true value for sensitive cells on a massive scale.

References

Dandekar R. A. and Cox L. H. (2002), Synthetic Tabular Data: An Alternative to Complementary Cell Suppression, manuscript available from ramesh.dandekar@eia.doe.gov or from URL http://mysite.verizon.net/vze7w8vk/

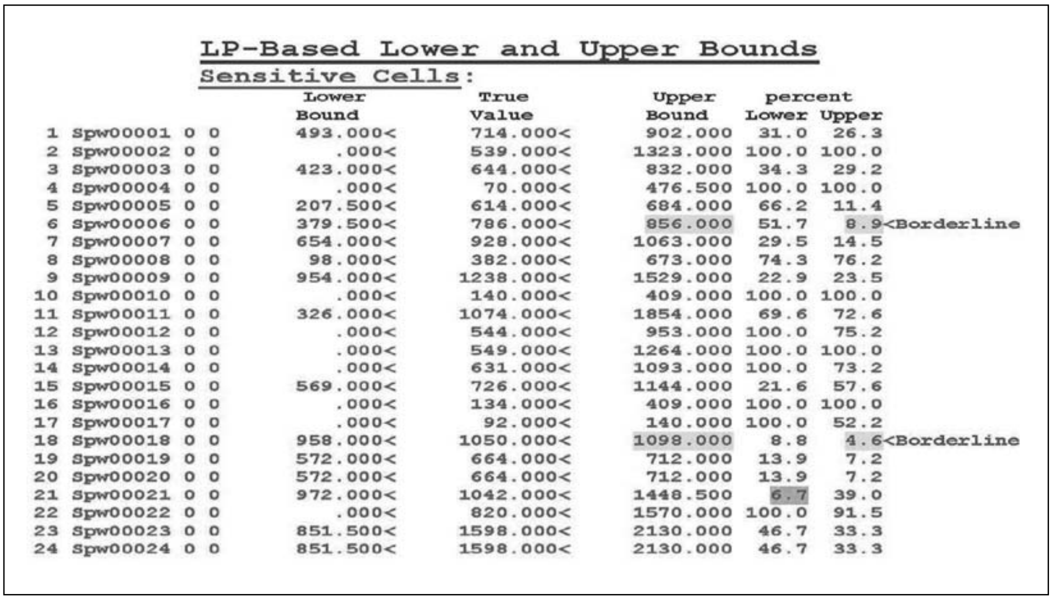
Dandekar, R.A (2003), Cost Effective Implementation of Synthetic Tabulation (a.k.a. Controlled Tabular Adjustments) in Legacy and New Statistical Data Publication Systems, working paper 40, UNECE Work session on statistical data confidentiality (Luxembourg, 7-9 April 2003) http://epp.eurostat.cec.eu.int/portal/page?_pageid=1073,1135281,1073_1135295&_dad=portal&_schema=PORTAL&p_product_code=KS-CR-03-004-3

Table 1.



```
                 CELL  SUPPRESSION—(10x6x4)  TABLE

   6764    714w   3356    4067    140w      --    3932   1478c      --  | 20451
   1994      --   5593      --   3022c   3504c      --    3220   1042w| 18375
   3744      --   3708      --   3678c   2502c      --      --      --  | 13632
   2810  10632c     --   2445c      --      --    2313    2978   7548c| 28726
   3682      --     --      --    4667   1988c   1748c    664w      --  | 12749
-------------------------------------------------------------------------------
  18994   11346  12657   6512c   11507    7994   7993c    8340    8590  | 93933

     --    539w     --     70w      --    7472    715c    3832      --  | 12628
   2253      --   4948    786w    472c   1074w    1830    5030      --  | 16393
    640      --    986      --      --    544w    631w     48c    750c|  3599
   1334      --   1016    382w   3175c   3302c    3803   1050w      --  | 14062
   1648   2814c     --      --      --    2102    726w      --   1598w|  8888
-------------------------------------------------------------------------------
   5875    3353   6950   1238w    3647   14494    7705    9960   2348c| 55570

     --   3552c   3476    614w   1916c    1131    549w     92w   1772c| 13102
     --      --   3222    928w      --      --    308c     429     87c|  4974
   4145      --     --    3692   2115c    4196    414c    3804    820w| 19186
   5995    644w     --      --   2410c   1677c      --   1912c   4134c| 16772
   2016      --     --   2212c    2826   1627c    134w      --      --  |  8815
-------------------------------------------------------------------------------
  12156    4196   6698    7446    9267    8631   1405c    6237   6813c| 62849

   6764   4805c   6832    4751    2056    8603    5196    5402   1772c| 46181
   4247      --  13763   1714c    3494    4578   2138c    8679   1129c| 39742
   8529      --   4694    3692    5793    7242   1045c   3852c    1570  | 36417
  10139   11276   1016    2827    5585    4979    6116    5940   11682  | 59560
   7346   2814c     --   2212c    7493    5717    2608    664w   1598w| 30452
-------------------------------------------------------------------------------
  37025   18895  26305   15196   24421   31119   17103   24537   17751  |212352
```

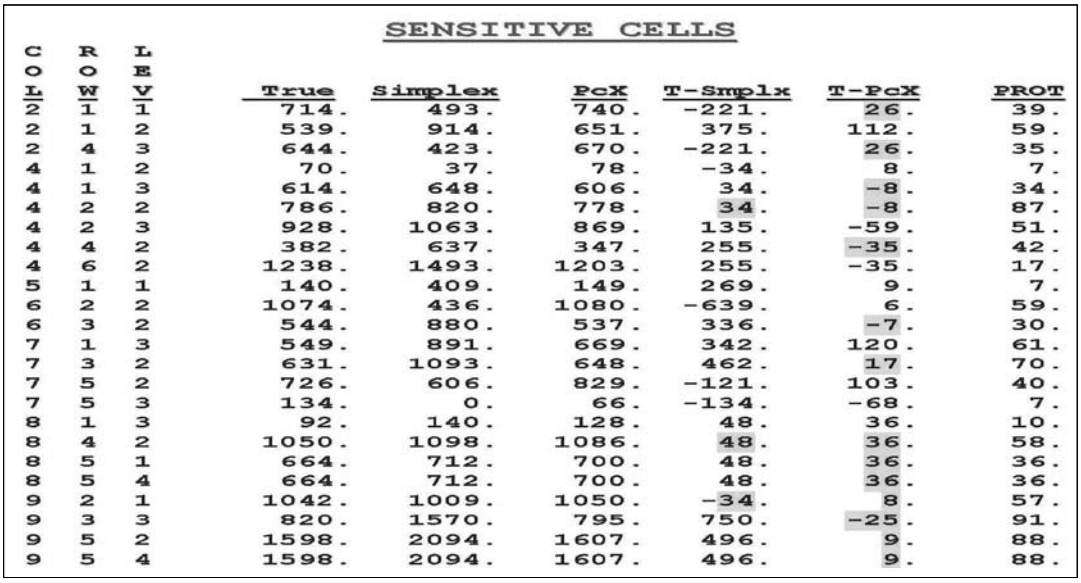
Table 2.



```
              LP-Based Lower and Upper Bounds
              Sensitive Cells:
                        Lower          True            Upper     percent
                        Bound          Value           Bound   Lower Upper
 1 Spw00001 0 0       493.000<       714.000<        902.000   31.0  26.3
 2 Spw00002 0 0          .000<       539.000<       1323.000  100.0 100.0
 3 Spw00003 0 0       423.000<       644.000<        832.000   34.3  29.2
 4 Spw00004 0 0          .000<        70.000<        476.500  100.0 100.0
 5 Spw00005 0 0       207.500<       614.000<        684.000   66.2  11.4
 6 Spw00006 0 0       379.500<       786.000<        856.000   51.7   8.9<Borderline
 7 Spw00007 0 0       654.000<       928.000<       1063.000   29.5  14.5
 8 Spw00008 0 0        98.000<       382.000<        673.000   74.3  76.2
 9 Spw00009 0 0       954.000<      1238.000<       1529.000   22.9  23.5
10 Spw00010 0 0          .000<       140.000<        409.000  100.0 100.0
11 Spw00011 0 0       326.000<      1074.000<       1854.000   69.6  72.6
12 Spw00012 0 0          .000<       544.000<        953.000  100.0  75.2
13 Spw00013 0 0          .000<       549.000<       1264.000  100.0 100.0
14 Spw00014 0 0          .000<       631.000<       1093.000  100.0  73.2
15 Spw00015 0 0       569.000<       726.000<       1144.000   21.6  57.6
16 Spw00016 0 0          .000<       134.000<        409.000  100.0 100.0
17 Spw00017 0 0          .000<        92.000<        140.000  100.0  52.2
18 Spw00018 0 0       958.000<      1050.000<       1098.000    8.8   4.6<Borderline
19 Spw00019 0 0       572.000<       664.000<        712.000   13.9   7.2
20 Spw00020 0 0       572.000<       664.000<        712.000   13.9   7.2
21 Spw00021 0 0       972.000<      1042.000<       1448.500    6.7  39.0
22 Spw00022 0 0          .000<       820.000<       1570.000  100.0  91.5
23 Spw00023 0 0       851.500<      1598.000<       2130.000   46.7  33.3
24 Spw00024 0 0       851.500<      1598.000<       2130.000   46.7  33.3
```

Table 3

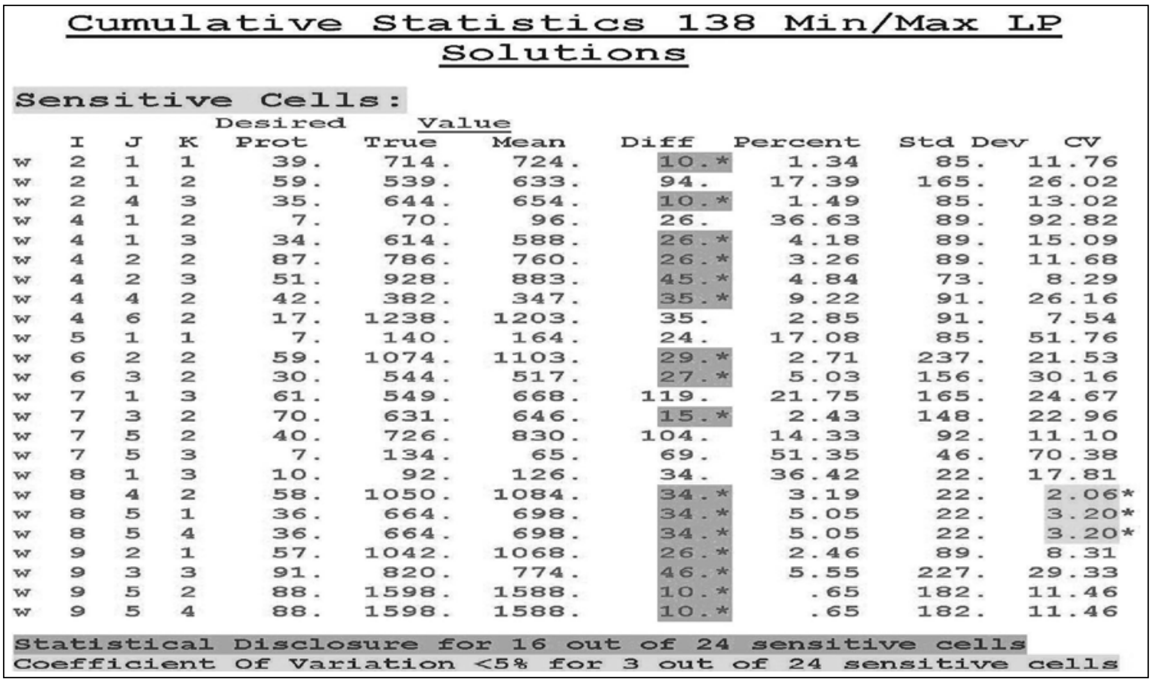
Additive Estimates Simplex versus Interior Point Method



SENSITIVE CELLS

COL	ROW	LEV	True	Simplex	PcX	T-Smplx	T-PcX	PROT
2	1	1	714.	493.	740.	-221.	26.	39.
2	1	2	539.	914.	651.	375.	112.	59.
2	4	3	644.	423.	670.	-221.	26.	35.
4	1	2	70.	37.	78.	-34.	8.	7.
4	1	3	614.	648.	606.	34.	-8.	34.
4	2	2	786.	820.	778.	34.	-8.	87.
4	2	3	928.	1063.	869.	135.	-59.	51.
4	4	2	382.	637.	347.	255.	-35.	42.
4	6	2	1238.	1493.	1203.	255.	-35.	17.
5	1	1	140.	409.	149.	269.	9.	7.
6	2	2	1074.	436.	1080.	-639.	6.	59.
6	3	2	544.	880.	537.	336.	-7.	30.
7	1	3	549.	891.	669.	342.	120.	61.
7	3	2	631.	1093.	648.	462.	17.	70.
7	5	2	726.	606.	829.	-121.	103.	40.
7	5	3	134.	0.	66.	-134.	-68.	7.
8	1	3	92.	140.	128.	48.	36.	10.
8	4	2	1050.	1098.	1086.	48.	36.	58.
8	5	1	664.	712.	700.	48.	36.	36.
8	5	4	664.	712.	700.	48.	36.	36.
9	2	1	1042.	1009.	1050.	-34.	8.	57.
9	3	3	820.	1570.	795.	750.	-25.	91.
9	5	2	1598.	2094.	1607.	496.	9.	88.
9	5	4	1598.	2094.	1607.	496.	9.	88.

Table 4



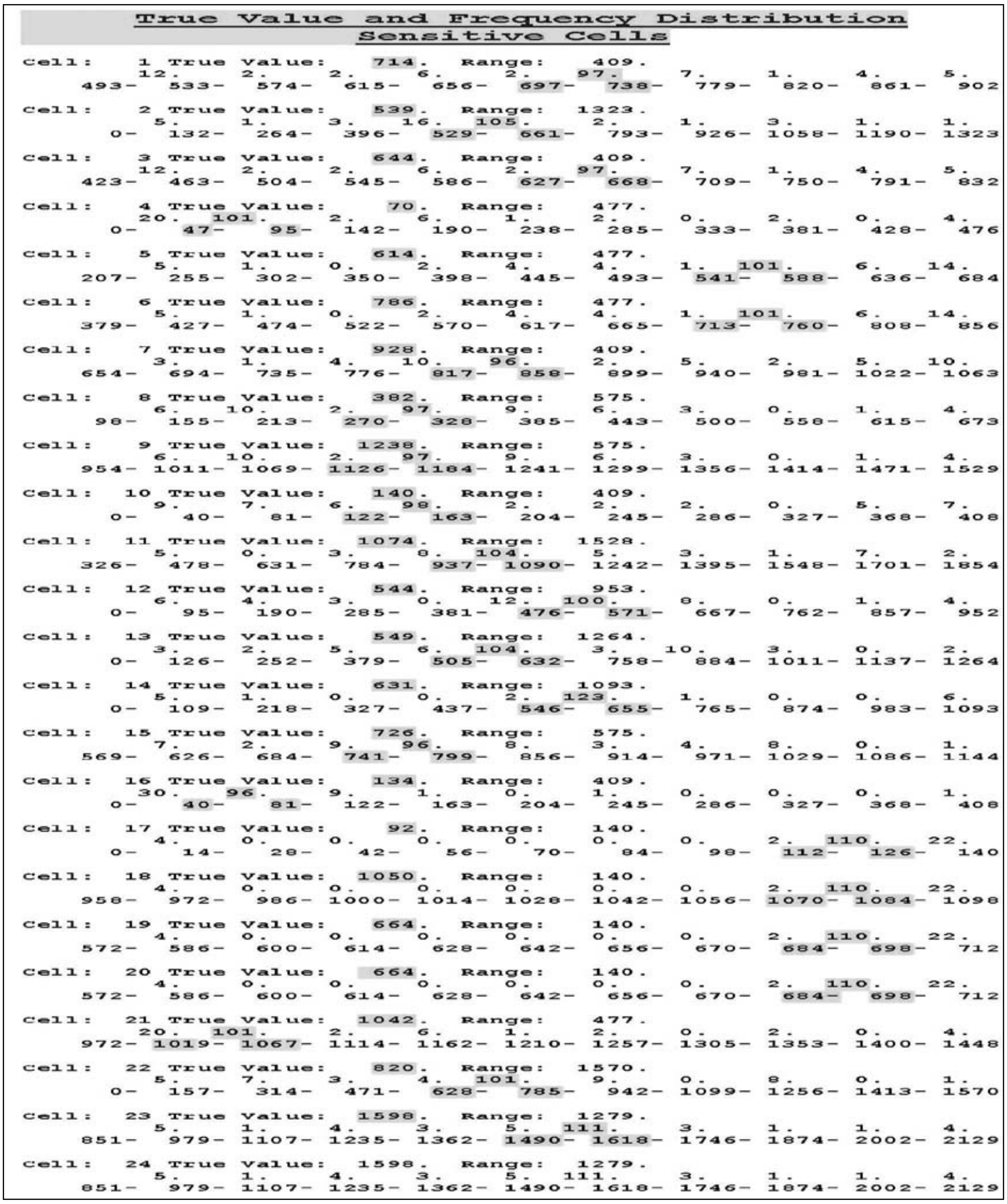
Cumulative Statistics 138 Min/Max LP Solutions

Sensitive Cells:

				Desired	Value					
	I	J	K	Prot	True	Mean	Diff	Percent	Std Dev	CV
w	2	1	1	39.	714.	724.	10.*	1.34	85.	11.76
w	2	1	2	59.	539.	633.	94.	17.39	165.	26.02
w	2	4	3	35.	644.	654.	10.*	1.49	85.	13.02
w	4	1	2	7.	70.	96.	26.	36.63	89.	92.82
w	4	1	3	34.	614.	588.	26.*	4.18	89.	15.09
w	4	2	2	87.	786.	760.	26.*	3.26	89.	11.68
w	4	2	3	51.	928.	883.	45.*	4.84	73.	8.29
w	4	4	2	42.	382.	347.	35.*	9.22	91.	26.16
w	4	6	2	17.	1238.	1203.	35.	2.85	91.	7.54
w	5	1	1	7.	140.	164.	24.	17.08	85.	51.76
w	6	2	2	59.	1074.	1103.	29.*	2.71	237.	21.53
w	6	3	2	30.	544.	517.	27.*	5.03	156.	30.16
w	7	1	3	61.	549.	668.	119.	21.75	165.	24.67
w	7	3	2	70.	631.	646.	15.*	2.43	148.	22.96
w	7	5	2	40.	726.	830.	104.	14.33	92.	11.10
w	7	5	3	7.	134.	65.	69.	51.35	46.	70.38
w	8	1	3	10.	92.	126.	34.	36.42	22.	17.81
w	8	4	2	58.	1050.	1084.	34.*	3.19	22.	2.06*
w	8	5	1	36.	664.	698.	34.*	5.05	22.	3.20*
w	8	5	4	36.	664.	698.	34.*	5.05	22.	3.20*
w	9	2	1	57.	1042.	1068.	26.*	2.46	89.	8.31
w	9	3	3	91.	820.	774.	46.*	5.55	227.	29.33
w	9	5	2	88.	1598.	1588.	10.*	.65	182.	11.46
w	9	5	4	88.	1598.	1588.	10.*	.65	182.	11.46

Statistical Disclosure for 16 out of 24 sensitive cells
Coefficient Of Variation <5% for 3 out of 24 sensitive cells

Table 5



```
            True Value and Frequency Distribution
                       Sensitive Cells
Cell:    1 True Value:    714.  Range:    409.
          12.    2.     2.     6.     2.    97.     7.     1.     4.     5.
    493-   533-   574-   615-   656-   697-   738-   779-   820-   861-   902
Cell:    2 True Value:    539.  Range:   1323.
           5.    1.     3.    16.   105.     2.     1.     3.     1.     1.
      0-   132-   264-   396-   529-   661-   793-   926-  1058-  1190-  1323
Cell:    3 True Value:    644.  Range:    409.
          12.    2.     2.     6.     2.    97.     7.     1.     4.     5.
    423-   463-   504-   545-   586-   627-   668-   709-   750-   791-   832
Cell:    4 True Value:     70.  Range:    477.
          20.  101.     2.     6.     1.     2.     0.     2.     0.     4.
      0-    47-    95-   142-   190-   238-   285-   333-   381-   428-   476
Cell:    5 True Value:    614.  Range:    477.
           5.    1.     0.     2.     4.     4.     1.   101.     6.    14.
    207-   255-   302-   350-   398-   445-   493-   541-   588-   636-   684
Cell:    6 True Value:    786.  Range:    477.
           5.    1.     0.     2.     4.     4.     1.   101.     6.    14.
    379-   427-   474-   522-   570-   617-   665-   713-   760-   808-   856
Cell:    7 True Value:    928.  Range:    409.
           3.    1.     4.    10.    96.     2.     5.     2.     5.    10.
    654-   694-   735-   776-   817-   858-   899-   940-   981-  1022-  1063
Cell:    8 True Value:    382.  Range:    575.
           6.   10.     2.    97.     9.     6.     3.     0.     1.     4.
     98-   155-   213-   270-   328-   385-   443-   500-   558-   615-   673
Cell:    9 True Value:   1238.  Range:    575.
           6.   10.     2.    97.     9.     6.     3.     0.     1.     4.
    954-  1011-  1069-  1126-  1184-  1241-  1299-  1356-  1414-  1471-  1529
Cell:   10 True Value:    140.  Range:    409.
           9.    7.     6.    98.     2.     2.     2.     0.     5.     7.
      0-    40-    81-   122-   163-   204-   245-   286-   327-   368-   408
Cell:   11 True Value:   1074.  Range:   1528.
           5.    0.     3.     8.   104.     5.     3.     1.     7.     2.
    326-   478-   631-   784-   937-  1090-  1242-  1395-  1548-  1701-  1854
Cell:   12 True Value:    544.  Range:    953.
           6.    4.     3.     0.    12.   100.     8.     0.     1.     4.
      0-    95-   190-   285-   381-   476-   571-   667-   762-   857-   952
Cell:   13 True Value:    549.  Range:   1264.
           3.    2.     5.     6.   104.     3.    10.     3.     0.     2.
      0-   126-   252-   379-   505-   632-   758-   884-  1011-  1137-  1264
Cell:   14 True Value:    631.  Range:   1093.
           5.    1.     0.     0.     2.   123.     1.     0.     0.     6.
      0-   109-   218-   327-   437-   546-   655-   765-   874-   983-  1093
Cell:   15 True Value:    726.  Range:    575.
           7.    2.     9.    96.     8.     3.     4.     8.     0.     1.
    569-   626-   684-   741-   799-   856-   914-   971-  1029-  1086-  1144
Cell:   16 True Value:    134.  Range:    409.
          30.   96.     9.     1.     0.     1.     0.     0.     0.     1.
      0-    40-    81-   122-   163-   204-   245-   286-   327-   368-   408
Cell:   17 True Value:     92.  Range:    140.
           4.    0.     0.     0.     0.     0.     0.     2.   110.    22.
      0-    14-    28-    42-    56-    70-    84-    98-   112-   126-   140
Cell:   18 True Value:   1050.  Range:    140.
           4.    0.     0.     0.     0.     0.     0.     2.   110.    22.
    958-   972-   986-  1000-  1014-  1028-  1042-  1056-  1070-  1084-  1098
Cell:   19 True Value:    664.  Range:    140.
           4.    0.     0.     0.     0.     0.     0.     2.   110.    22.
    572-   586-   600-   614-   628-   642-   656-   670-   684-   698-   712
Cell:   20 True Value:    664.  Range:    140.
           4.    0.     0.     0.     0.     0.     0.     2.   110.    22.
    572-   586-   600-   614-   628-   642-   656-   670-   684-   698-   712
Cell:   21 True Value:   1042.  Range:    477.
          20.  101.     2.     6.     1.     2.     0.     2.     0.     4.
    972-  1019-  1067-  1114-  1162-  1210-  1257-  1305-  1353-  1400-  1448
Cell:   22 True Value:    820.  Range:   1570.
           5.    7.     3.     4.   101.     9.     0.     8.     0.     1.
      0-   157-   314-   471-   628-   785-   942-  1099-  1256-  1413-  1570
Cell:   23 True Value:   1598.  Range:   1279.
           5.    1.     4.     3.     5.   111.     3.     1.     1.     4.
    851-   979-  1107-  1235-  1362-  1490-  1618-  1746-  1874-  2002-  2129
Cell:   24 True Value:   1598.  Range:   1279.
           5.    1.     4.     3.     5.   111.     3.     1.     1.     4.
    851-   979-  1107-  1235-  1362-  1490-  1618-  1746-  1874-  2002-  2129
```

SUDA: A program for Detecting Special Uniques[1]

Mark J Elliot[**] *Anna Manning*[*] *Ken Mayes*[*] *John Gurd*[*] *and Michael Bane*[***]

[*] **School of Computer Science, University of Manchester, Manchester M13 9PL UK.**

[**] **Centre for Census and Survey Research, University of Manchester, Manchester M13 9PL UK.**

[***] **School of Earth, Atmospheric and Environmental Sciences, University of Manchester, Manchester M13 9PL UK.**

Abstract: The importance of being able to classify records according to disclosure risk is well understood; Skinner and Holmes (1998), Fienberg and Makov (1998). One concept for so classifying records is called special uniqueness; see Elliot (2000), Elliot et al (2002), Manning and Haglin (2005). This paper describes SUDA (Special Uniques Detection Algorithm) which is both a set of computer science algorithms and indeed a fully functioning software system for detecting and grading special uniques. Section 1 describes the basic design principles behind the sequential SUDA algorithm. Section 2 describes the software (now in use at the UK Office for National Statistics and Australian Bureau of Statistics). Section 3 describes recent advances (i) in parallelising SUDA and improving the algorithm so that cross-classifications of up to 60 variables can be comprehensively analysed (ii) in developing a version of SUDA for Grid computing.

1. Introduction

The principle of being able to classify microdata records according to their disclosure risk is now axiomatic within the SDC field Skinner and Holmes (1998), Fienberg and Makov (1998). Within this paper we describe a software system entitled "SUDA" that provides such record detailed assessment broken down by record, variable, variable value and by interactions of those.

The basic principles behind the SUDA system are described in section 2 and the current version of the window implementation of this software (available as freeware under restricted license) is described in section 3.

Through collaboration of SDC researchers and computer scientists in Manchester efficient search algorithms have been produced which enable special uniques analyses of very large keys at unlimited variable interaction levels, in real time. The use of grid computing to further improve the efficiency is also being investigated. These new methods are described in section 4.

2. The Special Uniques Methodology

The concept of the "special unique" was coined by Elliot et al (1998). The principle is that a microdata record which is sample unique on coarser, less detailed information is more risky than one which is unique on finer, more detailed information. A particular case of that is where a record which is sample unique on a set of variables K and is also unique on a subset of K. Such a record is called a *special unique*, with respect of variable set K.

Extensive empirical work (Elliot 2000, Elliot and Manning 2001, Merrett et al 2005) has shown that special uniques are more likely to be population unique than random uniques. Further work (e.g. Elliot et al 2002) has shown that it is possible to classify special uniques according to the size and number of subsets which are unique minimal sample uniques (MSU) and that such classifications are correlated with the reciprocal population equivalence class, which is a generally accepted measure of underlying risk.

[1] The research described and software development in this paper has been supported by the ESRC, EPSRC and the UK Office for National Statistics.

2.1. The basic SUDA method.

Table 2.1 Notation

Notation	Description
ATT	Total number of attributes in the dataset
REC	Total number of records in the dataset
M	User-specified maximum size of attribute set
***(n-1)*-subset**	Subset of size *n-1*
R	Position of record in the dataset, where 1 ≤ R ≤ REC

SUDA is designed around the observation that 'Every superset of a unique attribute set (minimal or otherwise) is itself unique' (referred to as the *Superset Relationship;* Elliot et al. 2002). SUDA will be described in the following sections using the Superset Relationship as a basis for classifying the risk associated with each measure.

SUDA incorporates the Superset Relationship into the attribute set generation process in order to reduce the amount of record comparisons that are necessary. All attribute sets with the same prefix of size P[2], where 1≤P≤M, are generated in succession so that any superset of a unique prefix at a given record can be ignored immediately without the need to revert to stored information. Given ATT=6 and M=4, with attributes labelled A, B, C, D, E, F, the beginning of the attribute set generation process for SUDA would be as follows:

A, AB, ABC, ABCD, ABCE, ABCF, ABD, ABDE, ABDF, ABE, ABEF, ABF, AC, ACD etc ...

An example of the incorporation of the Superset Relation can be given as follows. Consider attribute sets with prefix ABC: that is ABC, ABCD, ABCE, ABCF from the above listing. If attribute set ABC is found to be unique at record R all supersets with ABC as their prefix can be ignored for R as they are not minimally unique. As all such supersets appear directly after ABC in the above sequence this process can be carried out immediately without the need to check stored information and has the effect of reducing the number of records that need to be considered for each attribute set while at the same time minimising memory usage.

2.2. Record grouping procedures

A grouping method is used in SUDA to collect together records with identical values for a given attribute set. This localises the records required for any given search and minimises the amount of memory usage that is necessary to identify minimal uniques. Figure 2.1 shows a dataset with 20 records and three attributes (A, B, C) and illustrates the check for potential uniques of attribute sets A, AB and ABC.

The records are initially grouped in terms of attribute A, by placing values of R in a 2 dimensional matrix according to their value of A, as shown in the bottom left of Figure 2.1. The records are then rearranged according to the results of this process into partitions, as shown in the second dataset configuration in Figure 2.1. Any partition with one member represents a minimal unique for A and this record can be removed from the grouping process; as A is a single attribute no check is required for the uniqueness of its subsets.

Attribute set AB is then checked by considering attribute values for B in each partition of the second dataset configuration in Figure 2.1.

[2] In general, for an attribute set A containing attributes $a_1,...a_r$ (with 1 ≤ r ≤ M), a prefix of size P of A where 1 ≤ P ≤ r contains the first P attributes $(a_1,...,a_P)$ of A.

If the dataset had more than three attributes (i.e. ATT≥M>3) the grouping procedure described above would be applied recursively to each of the partitions in the second dataset configuration in terms of attribute B and the resulting partitions placed in the third dataset configuration. Any partition containing just one record number would represent a potential unique and this information would be saved (in order to check for minimal uniques later) and the record number would not be placed in the third dataset configuration (as all supersets of this attribute set for this record would be unique).

However, ATT=M=3 and Figure 2.1 demonstrates how the uniqueness of attribute sets of size M-1 and M can be found more quickly.

For attribute sets of size M-1 a one-dimensional matrix *ONE* is used to identify potential uniques. For the (M-1)th attribute value of each record (in this case, attribute B) the corresponding record number (R) is placed in ONE according the value of B: e.g. a record with B=1 has its number (R) placed in the first cell of ONE and in the second cell if B=2. If more than one record is placed in any one cell its value cannot be unique and a value such as '-99' (or 'x' in Figure 2.1) represents this information. When all records in a partition have been checked all cells of ONE are scanned and any that don't contain 'x' or '0' contain values of R for records that are potentially unique for AB. The contents of the partition are then copied to the third dataset configuration, leaving out any records that were found to be potentially unique for AB. This procedure is repeated for all partitions in the second dataset configuration.

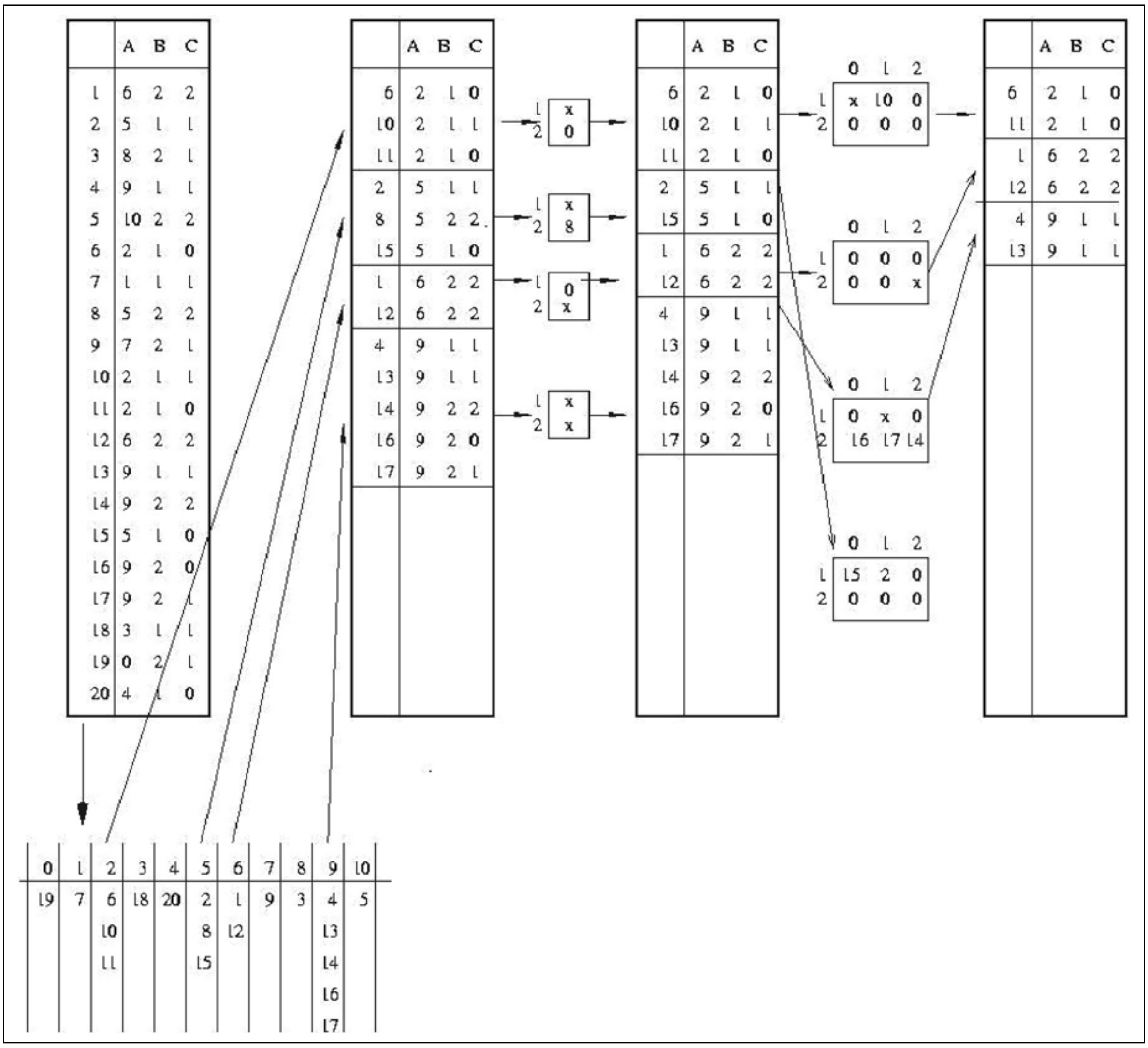
Attribute set ABC is then addressed. Records in dataset configuration 3 are considered on a partition level basis as before. For attribute sets of size M, record numbers are placed in a two-dimensional matrix *TWO* in which each cell corresponds to the values for the (M-1)st and Mth attributes (B and C) of each record. For example, if B=1 and C=2 for a record its number (R) is placed in the cell at the intersection of the first row and the second column of matrix TWO. As with ONE, when all records of a partition have been considered all cells of TWO are checked and if any do not contain 'x' or '0' then these record numbers represent records that are potentially unique for attribute set ABC. All records from this partition are then copied to the fourth dataset configuration in Figure 2.1 omitting the potential uniques (which are stored as before). This procedure is repeated for each partition of the third dataset configuration.

Figure 2.1 has been designed to explain the grouping procedures used by SUDA but suggests that the entire dataset is re-grouped for each attribute set. However, SUDA does not physically re-group the records of the dataset at any stage but uses a matrix (referred to as *Group Matrix*) to store values of R within each partition [Elliot et al. 2002].

2.3. The check for minimal uniqueness

All potential uniques of size $n≥2$ must be checked for the non-uniqueness of all their *(n-1)*-subsets to ensure that a minimal unique has been found. This has the potential to lead to very high memory requirements. However, due to the generation of attribute sets according to their prefixes the information can be retrieved from the Group Matrix and involves the use of a hash table Elliot et al. 2002.

Figure 2.1 Record grouping process for SUDA



	A	B	C
1	6	2	2
2	5	1	1
3	8	2	1
4	9	1	1
5	10	2	2
6	2	1	0
7	1	1	1
8	5	2	2
9	7	2	1
10	2	1	1
11	2	1	0
12	6	2	2
13	9	1	1
14	9	2	2
15	5	1	0
16	9	2	0
17	9	2	1
18	3	1	1
19	0	2	1
20	4	1	0

0	1	2	3	4	5	6	7	8	9	10
19	7	6 10 11	18	20	2 8 15	1 12	9	3	4 13 14 16 17	5

	A	B	C
6	2	1	0
10	2	1	1
11	2	1	0
2	5	1	1
8	5	2	2
15	5	1	0
1	6	2	2
12	6	2	2
4	9	1	1
13	9	1	1
14	9	2	2
16	9	2	0
17	9	2	1

	A	B	C
6	2	1	0
10	2	1	1
11	2	1	0
2	5	1	1
15	5	1	0
1	6	2	2
12	6	2	2
4	9	1	1
13	9	1	1
14	9	2	2
16	9	2	0
17	9	2	1

	A	B	C
6	2	1	0
11	2	1	0
1	6	2	2
12	6	2	2
4	9	1	1
13	9	1	1

2.4. Combining information from the lattice

2.4.1 Generating the intermediate SUDA metric

Once all minimal uniques have been found the following characteristics are important in the detection and grading of special uniques:

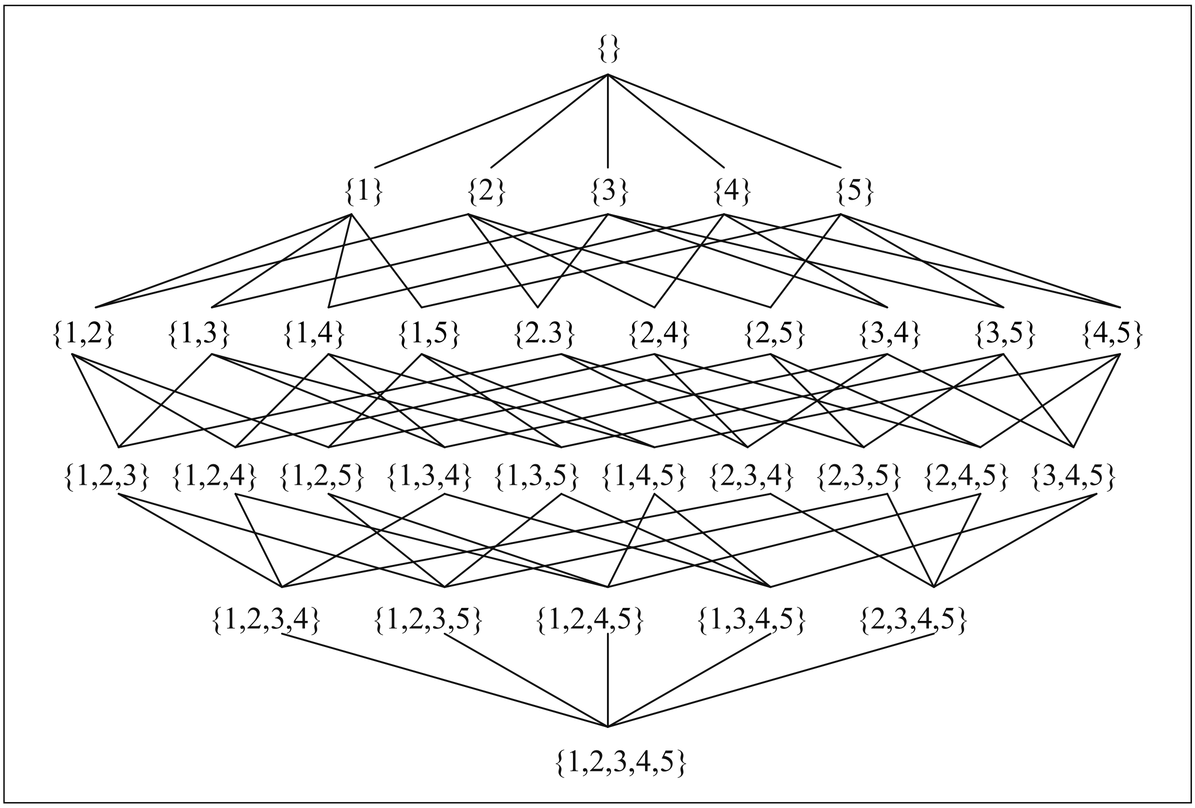
The size of minimal uniques: The smaller the size of the MSU within a record the more 'risky' the record is likely to be.

The number of minimal uniques per record: The larger the number of MSUs contained within a record the more likely the record is to be 'risky'.

These observations are used to code records according to their potential 'risk' as follows:

Let Ξ = {x1,…, xn} be a set of distinct literals, or attributes. The space of possible sets X ▯▯ can be visualised as a lattice - Figure 2.2 shows the case when Ξ = {1, 2, 3, 4, 5} and ATT=5. (This approach is used to describe the search space for association rule discovery [Zaki et al. 1997]).

Figure 2.2 Lattice for Ξ = {1, 2, 3, 4, 5}



Such a lattice can be used to describe the space of all possible subsets of a record.

A judgment needs to be made about the relative risk of MSUs of different sizes. For example, a large number of MSUs of size 3 may be regarded as more risky than one MSU of size 2. The above lattice structure is used to allocate a weight for each record so that the MSUs can be compared.

For each MSU X of size |X|=k contained in a given record R, where 1 ≤ k ≤ ATT-1, the Intermediate SUDA metric (IS metric) can be computed by counting the number of distinct 'paths' from X to the bottom of the lattice.[3] This can be represented as:

$$\# paths = \prod_{i=k}^{ATT-1}(ATT- \ i) = (ATT-k)!$$

If k=ATT the number of distinct paths is zero (i.e. no supersets). To avoid giving zero scores to records containing MSUs of size ATT a value of '1' is applied.

In SUDA, the MSUs often have a user-specified maximum size (M). Figure 2.3 shows an adjustment to the lattice in Figure 2.2 when M=2. Here, all the distinct paths from MSUs of size 2 are considered - this has the effect of cutting Figure 1 below the sets of size 2 and only including paths through the lattice from this point. In this case, the number of distinct 'paths' below a given set X of size k where 1≤k≤M can be represented as:

$$\# paths = \prod_{i=k}^{M}(ATT-i)$$

The above treats each record-level MSU independently – the scores for each record-level MSU are added together to give the final score for the record.

[3] Clearly this is just one way in which the MSU information could be combined. It is computationally principled.

2.4.2 Using the combined information

There are several ways that the IS metric can be used. One is to generate a proportion of lattice measure from the number of possible paths through the lattice structure given by ATT! The proportion of lattice statistic represents the IS metric as a proportion of this total:

Proportion of lattice at record R = (IS metric at R / ATT!)

An alternative is to use the data intrusion simulation output metric (see Skinner and Elliot 2002) to generate the total number of population units corresponding to the sample uniques and then to distribute them in some manner dependent upon the IS metric. This method known as DIS-SUDA produces estimates of intruder confidence in a match against a given record being correct. This is closely related to the probability that the match is correct given assumption of zero data divergence. See Elliot (2002), for a further discussion of the interpretation of this metric. The advantage of this method is that it relates to a practical model of data intrusion, and it is possible to compare different values directly. The disadvantage is that it is sensitive to the level of the max MSU parameter and is calculated in a heuristic manner. However, the method has been extensively tested and produces very good results when the max MSU size is large (Merrett et al 2004) and the number of key variables moderate. The proportion of lattice measure is more robust when the max MSU size is lower (for example when conducting a comprehensive rather than scenario based analysis.)

2.4.3 Description of risk at record-level and database-level

Record Level

In many records there are a small number of attributes that occur in a large proportion of the MSUs as illustrated by the following example.

Example: Table 2.1 shows the MSUs for and imaginary record with twelve attributes.

Table 2.1 MSUs for imaginary record in table

Size 2	Size 3	Size 4	Size 5
1 2	1 6 9		2 5 6 8 11
1 5	5 8 12		
1 8			

Table 2.2 Relative impact of attributes for record 1080324

Variable	Occurrence of MSUs of size:			% of MSUs affected of size:		
	2	3	5	2	3	5
1	3	1	0	100.00	50.00	0
8	1	1	1	33.33	50.00	100.00
5	1	1	0	33.33	50.00	100.00
2	1	0	1	33.33	0	100.00
6	0	1	1	0	50.00	100.00
9	0	1	0	0	50.00	0
11	0	0	1	0	0	100.00
3	0	1	0	0	50.00	0
4	0	0	0	0	0	0
7	0	0	0	0	0	0
10	0	0	0	0	0	0

Attribute 1 is the most prevalent, occurring in all 3 MSUs of size 2 and one of the two MSUs of size 3. The percentage contribution to each attribute to MSUs of each size is shown in Table 2.2.

Let S_i be the IS metric for MSUs of size i.

The total IS metric, S_R, for record 1080324 is given by: $S_R = 3\times S_2 + 2\times S_3 + S_5$

The contribution to the IS metric for each of the attribute values in record 1080324 are calculated by using the contribution percentages in Table 2.3. For example, the contribution to the IS metric of attribute 1 at record 1080324 (S_{R1}) is given by:

$S_{R1} = 1.0\times 3\times S_2 + 0.5\times 2\times S_3$

Database Level

The percentage contribution of each attribute value to the IS metric at database level is found by:

1. summing the contributions of this attribute value at record level over the whole file (call this value T_V) – for example, the contribution to the IS metric at record-level of every occurrence of AGE=24
2. summing the IS metric (S_R) for each record over the whole file (call this value T_S)
3. finding T_V as a percentage of T_S.

The percentage contribution of each attribute to the IS metric at database level is found by:

1. summing the contributions of this attribute at record level over the whole file (call this value T_A) – for example, the contribution to the IS metric at record-level of every occurrence of AGE.
2. summing the IS metric (S_R) for each record over the whole file (call this value T_S).
3. finding T_A as a percentage of T_S.

3. The Software

The SUDA algorithms described above have been implemented as a windows application. The application has a simple two window interface; input and output.

The input window, allows the user to specify the dataset, key variables, and the parameters for the run, such as the Maximum MSU size (smaller is faster but less accurate), sampling fraction of the dataset and so on. Output is sent to the output window and also to user specified file.

3.1. Description of Output

The output is divided into three parts. The first part contains summary information on the run, the most useful part of the output is the DIS score which provides a file level measure of the disclosure risk.

The second part of the output is the record by record output. The important columns of the record level output are: (i) IS metric: This is total IS metric calculated as described in section 2. (ii) Scoring metric: The 3rd column contains either the Proportion of lattice metric or the DIS-IS metric depending on which the user asked for.(3) MSUs: the sequence of columns after the output metrics give the number of MSUs for the record of each size up to the number the user specified. (iv) Variable Contribution percentage: The final set of columns are headed with the variable name with each of the variables the user has chosen. These columns record the percentage contribution of each variable to the total IS metric. This is simply the IS metric for the MSUs involving that variable over the IS metric for the record as described in section 2.4.3.

The third part of the SUDA output is the cross-file breakdown of the IS metric by variable and value. This allows the user to assess where the risk is concentrated within the file. For both types of output the contribution is the percentage of the total IS metric across the whole file which arises from MSUs involving the attribute or attribute value.

4. Current and Future Work

4.1. The SUDA 2 Algorithms

The SUDA system has greatly increased the depth of risk assessment possible; this was demonstrated by its application to data releases from the 2001 British Census. However, due to the demanding levels of execution time required to find all MSUs in stage one of SUDA, this algorithm is restricted to small datasets, particularly in terms of the number of columns that they possess. This problem formed the motivation for the development of a new algorithm, SUDA2.

SUDA2 improves SUDA using several methods. Firstly a new approach is used to provide a more dynamic representation of the search space for MSUs. Secondly, further properties of MSUs are identified and are used to design improved pruning strategies. Thirdly, a more efficient traversal of the search space is employed.

SUDA2 has the ability to identify the boundaries of the search space for MSUs with an execution time which is several orders of magnitude faster than that of SUDA. Not only will these developments provide statistical agencies with a much faster tool to work with, but the ability to assess microdata with many more variables than before will now be possible.

4.2 Grid Hiperstad[4]

The efficiency of the SUDA2 algorithm means that it becomes feasible for large amounts of data to be searched for large patterns. However, as searches increase in size, it is likely that execution time on a single machine will ultimately become prohibitive. Thus it is sensible to provide an infrastructure that allows such applications to execute in a distributed fashion over a heterogeneous network of computers. The aim of the *GridHiPerStaD project* is to produce a prototype software framework for running statistical disclosure applications on a Grid of computers. It is based on the approach of the PerCo performance control system (Mayes et al., 2005).

The nature of the suda2 algorithm allows the entire search to be split into subsearches, each of which can execute on a separate machine (though in the present incarnation of the algorithm the data must be replicated). In general, work on what might be termed ``divisible work" applications fall into two paradigms: master-worker (e.g. Condor MW; Goux et al 1995) and divide-and-conquer (e.g. Blumofe et al (1995). On the whole, the existing systems seek to be paradigm-specific rather than application-specific. That is, they represent efforts to allow application developers to fit a suitable application into the provided paradigm framework.

There is a set of potential problems when considering an application such as Suda2 for distributed execution on a Grid. The available machines are of diverse architectures and capabilities, and may have varying load. On the other hand, the Suda2 subsearches are of unpredictable duration, being related to the nature of the search subspace data rather than its size. Thus in order to optimise application performance, the GridHiPerStaD framework must be adaptive. For example, it must cope with the situation where an unexpectedly large subsearch is executing slowly on a computer that has a heavy multi-user load.

[4] GRID based HIgh PERformance computing STAtistical Discolure risk analysis

There is some evidence in the literature of a trend to recognise complexities introduced by heterogeneous and unpredictable platforms and applications. For example, in the master-worker system of Kee and Ha (1998) the master is able to redistribute allocated work at runtime. In the work-stealing paradigm there is, for example, the topology-aware random stealing of the SALSA actor-based system, which migrates actors according to communication overhead; Desell et al (2004).

The GridHiPerStaD framework is attempting to make available both master-worker (i.e. centralised scheduling) and work-stealing (i.e. distributed scheduling) *mechanisms*. The *policies*, which determine how these mechanisms are used, will be application-specific. Additionally, there are facilities for recovering from sub-optimal deployment of work. In the case where the search of a subspace is taking too long, the GridHiPerStaD system can cause the sub-search to be "checkpointed", and the remaining search migrated for resumption on a faster machine. It should also be possible to divide, at runtime, computationally demanding subspace searches.

Such a flexible approach may be necessary where both the computational demands of the application and the computational capabilities of distributed resources may be unpredictable. That is, in such a dynamic scenario, a single scheduling algorithm or paradigm may not be consistently optimal. The performance-orientated scheduling policy of the system may have to adapt, and this must be underpinned by a number of mechanisms.

Summary

The SUDA system provides increasingly sophisticated methods for disclosure risk assessment of microdata. The method has now been implemented as a windows software package which is in use in three national statistical agencies.

The method is in a continual sate of refinement and enhancement, both in terms of its Computer Science and the SDC algorithms on which it is based. New sophisticated versions are close to completion. With the possibility of GRID enabled versions in the offing it is plausible to envisage more sophisticated risk evaluations explicitly taking into account the co-presence of other datasets in the data environment. In harness with web crawling software it is even possible to envisage comprehensive data environment analyses.

References

Blumofe, R, Joerg, C., Kuszmaul, B., Leiserson, C., Randall, K. & Zhou, Y (1995), 'Cilk: An Efficient Multithreaded Runtime System' In *Proceedings of the 5th Symposium on Principles and Practice of Parallel Programming*, 207-21, Santa Barbara, Calif.

Desell, T. and El Maghraoui, K. and Varela, C. (2004) 'Load Balancing of Autonomous Actors over Dynamic Networks' In *Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences*, HICSS'04, 90268.1, Washington, DC, USA.

Elliot, M, J, (1999). 'DIS: Data Intrusion Simulation - a method of estimating the worst case disclosure risk for a microdata file'. In *Proceedings of an international symposium on linked employee-employer records.* Washington: Bureau of the Census.

Elliot, M. J., Manning, A. M.& Ford, R. W. (2002**)**. 'A Computational Algorithm for Handling the Special Uniques Problem'. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems* 5(10), pp 493-509.

Elliot, M. J., Skinner, C. J., and Dale, A. (1998). 'Special Uniques, Random Uniques, and Sticky Populations: Some Counterintuitive Effects of Geographical Detail on Disclosure Risk'. *Research in Official Statistics* 1(2), pp 53-67.

Fienberg, S. E. and Makov M. M. (1998) Confidentiality, Uniqueness and Disclosure Limitation for Categorical Data, *Journal of Official Statistics* 14 (4), 395-98.

Goux, J-P and Kulkarni, S., Yoder, M. & Linderoth, J.,(1995) 'Master\\u2013Worker: An Enabling Framework for Applications on the Computational Grid' *Cluster Computing*, 4(1), 63—70.

Kee, Y. & Ha, S(1998) 'A Robust Dynamic Load-balancing Scheme for Data Parallel Application on Multicomputer Systems' In *Proceedings of International Conference on Parallel and Distributed Processing Techniques and Applications*, 974-980, Las Vagas, USA.

Manning A. M. and Haglin, D. J. (2005) "A new algorithm for finding Minimal Sample Uniques for use in Statistical Disclosure Assessment", *Proceedings of The Fifth IEEE International Conference on Data Mining*, New Orleans, Louisiana, U.S.A., November 27-30, 2005.

Mayes, K.R., Lujan, M Riley, G.D. Chin, J., Coveney P.V., & Gurd, J.R. (2005) 'Towards Performance Control on the Grid' *Philosophical Transactions of the Royal Society of London Series A*, 363 (1833), 1793-1805.

Merrett, K, (2005) 'Report on the validation of the Initial special uniques analysis of 2001'. *Office for National Statistics internal working document.*

Skinner, C. J. and Elliot, M. J. (2002). 'A measure of disclosure risk for microdata', *Journal of the Royal Statistical Society Series B*, 64(4) pp 855-867.

Skinner, C. J. and Holmes, D. J.(1998) Estimating the Re-identification Risk per Record in Microdata. *Journal of Official Statistics* 14 (4), 361-372.

Improving confidentiality with τ-Argus by focussing on clever usage of microdata

Roland van der Meijden MSc
Statistics Netherlands

Abstract: Users of tabular data want more andù more detailed information, which has huge consequences for confidentiality. In order to meet the users' needs without losing a lot of information (caused by primary and secondary suppressions) one should focus on clever usage of microdata and available tools. Clever usage of microdata can be divided into several different areas. First, reconstructing hierarchies of classifications from narrow to wide hierarchies will diminish primary suppressions as well as secondary suppressions on higher hierarchical levels when making publications on lower hierarchical levels. Second, bringing the different publication obligations into accordance with each other will diminish the existence of unnecessary suppressions. Clever usage of available tools can be divided into two areas. First, (mis)using the history file in τ-Argus in order to direct the confidentiality pattern. Second, using different information loss weights for improving secondary suppressions. When using all these clever adjustments together, considerably less information will be lost due to confidentiality reasons.

1. Introduction

At Statistics Netherlands the software tool τ-Argus is used for statistical disclosure control (SDC) of tabular data for (most) business statistics. τ-Argus uses cell suppression as SDC technique in order to make it impossible to exactly or approximately recalculate sensitive cells in published tables. The cell suppression technique comprises two steps. First, suppressing the primarily sensitive cells and second suppressing a number of cells in order to prevent disclosure due to the additive relationship between the cells of the table (the so-called secondarily sensitive cells).

This paper will give an overview of the possibilities in directing cell suppression and improving confidentiality by clever usage of base material and using some specific features of τ-Argus. This paper will also shortly mention what can be done in order to improve confidentiality without the need for tooling or clever usage of base material. Hopefully this paper can give enough practical guidance to diminish the occurrence of overprotected tabular data.

This paper starts in Section II with an explanation of why software tools are important when working on confidentiality. This section also mentions the different confidentiality rules that τ-Argus offers. In Section III the different tuning possibilities of τ-Argus are extensively described. This section also descibes how tuning is done and shows several results of the different tuning techniques. In Section IV the coordination of publication obligations is mentioned.

2. Why one should work on confidentiality

2.1. General reasons

More and more users of tabular data want more detailed information. This is a problem, since the availability of more detail in tables increases the chance of extracting information about individuals. The more detail is given, the more important confidentiality protection becomes. For more information on confidentiality protection in general we refer to Willenborg and De Waal (2001).

If time and energy is put in elaborate preparation before statistical disclosure control is applied, more detailed information can be published without compromising confidentiality (as will be shown in this paper).

2.2. Why use a tool like τ-Argus?

Protecting tabular data against disclosure is an inevitable part of statistics. Since the development of automated cell suppression software, the design of a useful suppression pattern is no longer a time consuming (and error prone) activity. It is also possible to calculate alternative suppression patterns.

To show the complexity of protecting tabular data and the necessity of automated cell suppression software a few examples will be given successively. See also Van der Meijden and Schalen (2004).

Table 1. Turnover of business sector X in a region is completely produced by one enterprise.

Region	State "Groningen" (NUTS 2 level)	State "Friesland" (NUTS 2 level)	State "Drente" (NUTS 2 level)	North-Netherlands (NUTS 1 level)
Turnover	40 (enterprise A)	60 (enterprise B, C en D)	-	100

According to τ-Argus the cells in "Groningen" and "Friesland" (Table 1) are primarily confidential because both cells have less than four contributors to the turnover (see also Section IIc). τ-Argus can also "see" that the enterprise in "Groningen" is able to recalculate the turnover in "Friesland" based on the total turnover in "North Netherlands" and the enterprise's own turnover. Since the turnover in "Friesland" must stay confidential, the enterprise in "Groningen" should not be able to disclose this information. Therefore τ-Argus will also make "North Netherlands" secondarily confidential. See also Salazar-González (2004), where the mathematical model underlying τ-Argus is explained.

τ-Argus must also protect against recalculation of "North Netherlands" based on other dimensions (Table 2). See also Giessing (2001) and Hundepool (2001). When taking into account the other dimensions of a table, more secondary suppressions will occur than expected at first sight.

τ-Argus takes two steps in order to prevent disclosure of "North Netherlands". First another region in the same business sector will be made confidential. Second another business sector in the same region will be made confidential.

When all the different dimensions (for example NACE, size class and region) in the table are made confidential sequentially, the chance exists that the resulting table is still not completely safe. Therefore, τ-Argus calculates the confidentiality effects on all dimensions simultaneously. See also Van der Meijden and Schalen (2004) and De Wolf (2002).

Table 2. The secondarily confidential region "North Netherlands" can be disclosed if τ-Argus does not make another region as well as another business sector (or totals) secondarily confidential.

	North	East	South	West	Total
Business sector X	Secondarily confidential	300	100	400	1000
Business sector Y	200	…	…	…	…
Business sector Z	500	…	…	…	…
Total	800	…	…	…	…

It can be concluded that a cell with one contributor (singleton) has effects on confidentiality in other dimensions that cannot always (easily) be overseen. Especially when working on large tables with three or more dimensions, statistical disclosure control becomes difficult. See also Feuvrier and Faes-

Cannito (2003). A tool like τ-Argus does oversee all the effects and eases the task of protecting the confidentiality of respondents.

2.3. Confidentiality rules in τ-Argus

τ-Argus offers four confidentiality rules, of which three are "dominance" rules and one is a minimum frequency rule. The "dominance" and frequency rules work independent of each other and determine together how the primary suppressions are selected. The frequency rule states how many contributors a cell must have in order to be safe. The "dominance" rules (Loeve (2001)) are:

- (*n*,*k*)-rule – a cell is primary unsafe if a number of *n* contributors is responsible for more than *k* percent of the total value of that cell.
- *p*%-rule – a cell is primary unsafe if an individual contribution can be recalculated within *p* percent of the actual value.
- *p*-*q*-rule – the assumption is made that individual contributions are known with a margin of *q* percent. A cell is primary unsafe if an individual contribution can be recalculated within *p* percent of the actual value.

3. Tuning possibilities for τ-Argus

3.1. General

The statistical disclosure control provided by τ-Argus is based on input from four different domains. Tuning of cell suppression patterns is therefore only possible by making changes to these four different domains. See Van der Meijden et al (2004).

The four different domains that τ-Argus uses as input are:

1. Hierarchies: The way hierarchies are built is of influence on how secondary suppressions are applied.
2. History file: A preference can be given for which cells may or must be secondarily confidential.
3. Information loss weights: Information will be lost when applying secondary suppressions. The way τ-Argus calculates this information loss can be adjusted.
4. Base material: The way the microdata and preferred output are composed is of influence on the way secondary suppressions are applied.

The order of the above mentioned domains indicates at which domains generally the best results are expected when making changes to the four domains. The four different domains will be discussed successively in the next sections. For more information about τ-Argus one should read the user guide of Hundepool et al (2004).

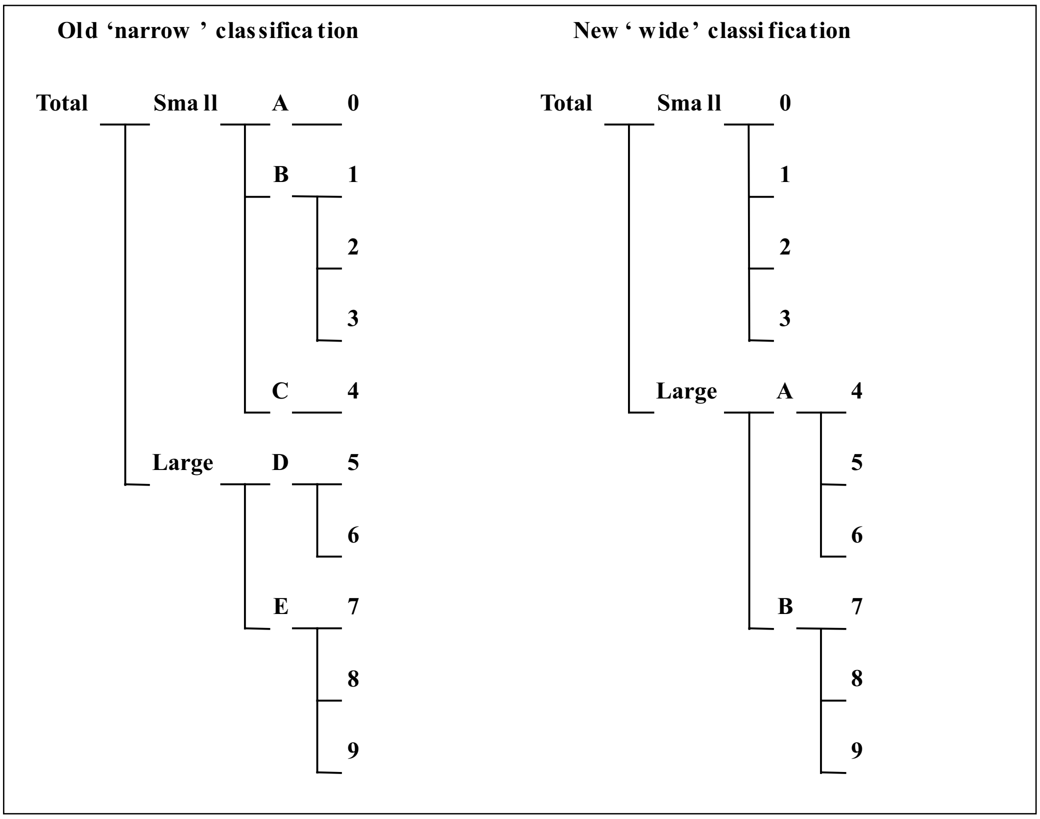
3.2. Hierarchies

τ-Argus needs to know how the hierarchies of classifications (for example NACE, size class and region) in the base material are constructed in order to choose counterparts for primary suppressions, that is secondary suppressions.

It is easier to determine secondary suppressions within a level of a hierarchy (and not influence the higher levels of the classification) if there are more subcategories within the level. Furthermore "singletons" (cells with one contributor) have less influence if τ-Argus can find enough counterparts within a level. In Figure 1 is shown what is meant by more subcategories. Especially the hierarchical parts “A-0”, “C-4” and “D-5/6” in the “narrow” classification may give more primary and secondary suppressions at higher hierarchical levels because there are no or not enough subcategories at the same level that can act as counterpart for primary suppressions.

In order to keep the secondary suppressions at the same level where the primary suppressions occur, the hierarchy has to be “widened”. This is only possible in ordinal classifications like NACE and size class. A change in an ordinal classification can be carried out without influencing the existing obligations for publication. A “widening” of the classification can supply for enough counterbalance for primary suppressions within a level of a hierarchy. If and how a classification can be “widened” depends (of course) on the obligations for publication and the type of classification. See also Westlake (2003).

Figure 1. A rearrangement of subcategories within a size class classification



The results (and differences) of using a “narrow” or “wide” classification (Figure 1) can be seen in Table 3. In this table we used the narrow and wide hierarchies from Figure 1 in order to calculate the number of primary and secondary sensitive cells for NACE “Industry”. Notice that the percentage of frequency unsafe and secondary unsafe cells has been reduced in the higher levels of the hierarchy when using the “wide” hierarchy. Also notice that the percentage of safe cells in the higher levels of the hierarchy has increased. In general a “wide” hierarchy also requires less secondary suppressions than a “narrow” hierarchy.

Because the NACE classification is an international standard, it is not desirable and possible to make changes to this classification. However, it is possible to harmonise the lowest level of the NACE classification that is used with the publication obligations in order to minimize the number of suppressed cells.

Table 3. The number and percentage of primary and secondary confidential cells as well as safe cells when using different (narrow or wide) hierarchies of the size class classification. The numbers in bold italics are the most interesting changes when using a wide hierarchy.

	Status narrow size class					**Status wide size class**				
	size class total	**size class S - L**	**size class A - E**	**size class 1 - 9**	**total**	**size class total**	**size class S - L**	**size class A - B**	**size class 1 - 9**	**total**
	%	%	%	%	%	%	%	%	%	%
A frequency unsafe	32,8	42	52,9	61,4	52,9	32,8	***41,1***	***49,3***	61,4	***51,9***
B dominance unsafe	4,9	1,5	0,5	0,3	1	4,9	1,6	0,7	0,3	1,1
C history file	0,1	0	0	0	0	0,1	0	0	0	0
D secondary unsafe	27,2	31,5	25,1	17,9	23,1	***26,5***	31,5	28,4	18,1	23,8
V safe	35,1	25,1	21,5	20,4	23	***35,7***	***25,8***	***21,6***	20,2	***23,1***
Total	100	100	100	100	100	100	100	100	100	100

When more detail is needed, a lower hierarchical level within a classification is used. However, because the lower hierarchical level will contain smaller cells the chance becomes higher that the number of enterprises in a cell will be below the threshold of the minimum number of enterprises in a cell (for frequency sensitivity). Therefore more detail in the lower hierarchical level can result in more primary (and indirectly in more secondary) suppressions.

When more detail in a lower hierarchical level is combined with a "narrow" classification it may result in secondary suppressions on a higher hierarchical level than the level used. The result is exactly contrary to the desire to show more detail.

Sometimes it is better to use a higher hierarchical level within a classification because it can result in showing more information/detail than when using a lower hierarchical level. The reason for this is less (or even no!) occurrence of primary and secondary suppressions at both the same and higher hierarchical levels as used. See also De Wolf and Mulder (2002).

Consider, for example, the Transport statistics (NACE 634); this is a very "narrow" classification. NACE 634 only consists of NACE 6340 which in its turn consists of NACE 63401 and 63402. For the annual Transport statistics at Statistics Netherlands one GK-NACE combination at the fifth digit NACE-level was primarily confidential, while there was not enough counterbalance at the fifth digit NACE-level, which in combination with the "narrow" classification resulted in secondary suppressions at a higher level; in this case even up to the third digit NACE. If NACE 63402 is primarily confidential, while NACE 63401 cannot provide for enough counterbalance, NACE 6340 and NACE 634 will become secondarily confidential. Publishing on the fourth digit NACE, provided that it is not generally known that most of the turnover is in NACE 63402, will result in safe cells.

Publishing at the fourth digit NACE instead of the fifth digit does not lead to disclosure of any sensitive information. Thus it can be beneficial to publish on a higher hierarchical level because of less primary and secondary suppressions in the higher hierarchical levels when using τ-Argus.

It can be concluded that it is beneficial to tune the hierarchies of classifications. When classifications are "widened", the effect of "singletons" will be diminished. Therefore secondary suppressions at higher hierarchical levels will not exist anymore. If classifications cannot be "widened", publishing on a higher hierarchical level can result in less suppressions in the higher hierarchical levels and thus possibly more information is available.

3.3. History file

The statistician can adjust the status of confidentiality of cells by hand in the history file. Originally the history file was meant for τ-Argus to take into account the suppression pattern of former years (hence the name history file). See also De Wolf (2003). However, the history file can also be used for directing the confidentiality pattern towards the wishes of the statistician.

The history file can be used for both giving a preference for suppressing and not suppressing a cell. Through the history file both the primary and secondary suppressions can be directed. τ-Argus tries to grant the settings in the history file as much as possible.

However, there are some rules when assigning status codes:

1) Cells without any contributors cannot get a status by hand.
2) Codes 'Confidential', 'Preferably suppress secondarily' and 'Preferably don't suppress secondarily' can only be applied on cells that are primarily safe.
3) Code 'Publishable' can only be applied on cells that are primarily unsafe.

The retail trade for example has used the history file in an optimal way. The "department stores" were primarily confidential and because of their minor total turnover the "retail sale of antiques" could not be a proper counterpart for secondary suppression. Therefore instead of the not so important "retail sale of antiques" some important cells became secondarily confidential. With the help of a subject matter expert in retail trade the history file was complemented with a list of cells (NACE/size class combinations) that could be used for secondary suppressions in order to get enough turnover as counterpart for the primarily confidential cells. Some important cells were also given a somewhat "protected" status by giving them the status "preferably don't suppress secondarily". Because of the use of the history file the important cells stayed publishable.

3.4. Information loss

Secondary suppressions are inevitable in order to make it (almost) impossible to retrieve the information from cells that are primary confidential. The results of the method used for determining the secondary suppressions depend on how information loss is measured.

τ-Argus will try to keep the information loss at a minimum while finding the cells for the secondary suppressions. In order to keep the information loss at a minimum τ-Argus must know how the information loss should be measured. For that purpose every cell in the table must be assigned a value which reflects the amount of information in that cell.

τ-Argus offers the following standards for measuring information loss:

Cell value: the amount of information in a cell equals the cell value. The consequence of minimizing this information loss is that cells with a higher value will be used less frequently as secondarily confidential cells than cells with a lower value.

Frequency: the amount of information in a cell equals the number of contributors to that cell. Minimizing this kind of information loss results in secondary suppressions applied on cells with fewer contributors.

Equal: the amount of information in a cell is exactly the same for each cell. Minimizing this kind of information loss results in producing the smallest number of cells that are suppressed secondarily. The consequence of this method is that important cells are suppressed as easily as unimportant cells.

Distance: the amount of information in a cell is related to the distance of that cell to a primarily confidential cell. Minimising this kind of information loss results in a clustering of secondarily confidential cells around the primarily confidential cells.

Table 4. The number and percentage of primary and secondary confidential cells as well as safe cells when using different methods for determining information loss. The numbers in bold italics are the most interesting differences when using another method.

	Methods for determining information loss							
	Cell value				Frequency			
Status	2nd digit NACE	3rd digit NACE	4th digit NACE	5th digit NACE	2nd digit NACE	3rd digit NACE	4th digit NACE	5th digit NACE
A frequency unsafe	0	195	2686	7995	0	195	2686	7995
B dominance unsafe	0	26	216	641	0	26	216	641
D secondary unsafe	4	321	2840	6816	4	***298***	***2638***	***6423***
V safe	285	1382	4805	8331	285	***1405***	***5007***	***8724***

It should be clear that the different information loss measures for obtaining secondary suppressions result in different suppression patterns. In Table 4 the results of the “cell value” and “frequency” methods for determining information loss are shown for different NACE levels. The results are for the wholesale business. The experience is that it is very useful to try the different information loss measures in order to get the best possible way of secondary suppression.

4. Coordination of publication obligations

Although this section is not about tuning base material or using specific features in τ-Argus, it is about improving confidentiality given the tools and microdata. Statistical offices are working so hard on developing new methodologies and software tools for disclosure control that one almost forgets that thinking about confidentiality starts way before a publication is made.

Thinking about confidentiality starts when making agreements with customers about what information to deliver. What are the publication obligations? What obligations are enforced (for example by Eurostat and government)? And what obligations are agreed on with specific customers? The more agreements are made about different groupings of variables, classifications, etcetera, the bigger the confidentiality problem becomes. See also De Wolf et al. (2002).

The Industry statistics at Statistics Netherlands, for example, are made for different customers (Eurostat and EIM – Economical Institute for small to Medium-sized enterprises). For both customers another size class classification is used. Because EIM is especially interested in SME (small to medium-sized enterprises), the size class categories are a little different from that of Eurostat. The result is that the different size class classifications are partly overlapping, which is especially limiting the number of cells that can be published without disclosing information. Although the customer is happy to get the exact classification that they wanted; they also get a lot of unwanted disclosed information. The question is whether the customer is not better off with the other classification but with less disclosed information!

Using different classifications or variables for different customers leads to inevitably much unnecessary confidentiality for all customers, thus not only for the customer who wants something else than other customers. Because the publications for EIM are agreed upon, they should be brought into accordance with the enforced obligation of Eurostat in order to minimize the confidentiality and improve the amount of information released for all customers. See also Samuelson (2001).

References

De Wolf, P.-P. (2002), *HiTaS: A heuristic approach to cell suppression in hierarchical tables*, Inference Control in Statistical Databases, Springer-Verlag, Berlin Heidelberg, pp. 74 – 82.

De Wolf, P.-P. (2003), *Guideline τ-Argus interface for business statistics* (internal note), Statistics Netherlands, Voorburg.

De Wolf, P.-P., Hundepool, A., Loeve, A. and De Waal, T. (2002), *Securing tabular data, now and in the future* (internal note), Statistics Netherlands, Voorburg.

De Wolf, P.-P. and Mulder, A. (2002), *Comparison of different table protection tools* (internal note), Statistics Netherlands, Voorburg.

Feuvrier, P. and Faes-Cannito, F. (2003), *Cell suppression in Eurostat on structural business statistics – an example of statistical disclosure control on tabular data*, paper presented at the joint ECE / Eurostat work session on statistical data confidentiality, Luxembourg, 2003.

Giessing, S. (2001), *New tools for cell suppression in tau-Argus: one piece of the casc project work draft*, Federal Statistical Office of Germany, paper presented at the joint ECE / Eurostat work session on Statistical Data Confidentiality, Skopje, 2001.

Hundepool, A. (2001), *The CASC project*, Statistics Netherlands, paper presented at the joint ECE / Eurostat work session on Statistical Data Confidentiality, Skopje, 2001.

Hundepool, A., Van de Wetering, A., Ramaswamy, R., De Wolf, P.-P., Giessing, S., Fischetti, S., Salazar, J.J., Castro, J. and Lowthian, P. (2004), *τ-Argus user's manual version 3.0.*

Loeve, J.A. (2001), *Dominance rule versus P-Q rule* (internal note), Statistics Netherlands, Voorburg.

Salazar-González, J.-J. (2004), *Mathematical models for applying cell suppression methodology in statistical data protection,* European Journal of Operations Research 154, pp. 740-754.

Samuelson P. (2001), *Statistics – a balance between official and confidential data*, Ministery of Justice, Sweden.

Van der Meijden, R., Schalen, J., De Wolf, P.-P. and Hundepool, A. (2004), *Study results and guidelines for tuning τ-Argus* (internal note), Statistics Netherlands, Voorburg.

Van der Meijden, R. and Schalen, J. (2004), *Confidentiality issues for τ-Argus* (internal note), Statistics Netherlands, Voorburg.

Westlake, A. (2003), *Security and disclosure for statistical information*, Survey & statistical computing, London.

Willenborg, L. and De Waal, T. (2001), *Elements of Statistical Disclosure Control*, Lecture Notes in Statistics 155, Springer-Verlag, New York.

Witvliet, H. and Heerschop, M.J. (2005), *small area estimations in ESB* (internal note), Statistics Netherlands, Voorburg.

Disclosure Analysis for the Census of Agriculture

Robert T. Smith

Census and Survey Division, National Agricultural Statistics Service, U.S. Department of Agriculture

Abstract: The National Agricultural Statistics Service used the network-flow disclosure methodology for the 2002 Census of Agriculture. This paper discusses how the statisticians and computer programmers worked closely together to achieve a very successful application of this disclosure methodology. The paper describes how the magnitude and complexity of the agricultural data structure were the impetus for the creation of a system to assist the analyst in the development and analysis of the input parameter files. Enhanced diagnostic tools allowed the analyst to continually review disclosure patterns and provide feedback that helped the computer programmers tailor the system specifically to the agriculture data set. This paper presents the details on these tools and discusses some of the modifications to the program logic that was a result of their use.

1. Background

The National Agricultural Statistics Service (NASS) used the network flow disclosure methodology for the Census of Agriculture. The network flow programs were originally developed at the U.S. Census Bureau for the economic censuses. The programs were modified by NASS for application to the 2002 Census of Agriculture and its follow-on programs including the Census of Puerto Rico, the Farm and Ranch Irrigation Survey and the Census of Aquaculture.

In the agency's first application of this methodology it was important that the agriculture analyst provide feedback to the disclosure system so that it could be adapted more effectively to agricultural data. The publication tables were very complex, and it was very difficult to create the data file of linear relations which specified how an entry in one publication table could be derived from entries in that table or in other tables. The analysts were involved in the project from the beginning specification of the linear relations to the review of the final suppression patterns. It was this involvement and our ability to modify the system that made this a very successful project.

The application of this methodology to the census consisted of three broad areas; the creation of the input parameters and files, the modification and running of the disclosure programs, and the review of the diagnostics that identified any necessary modifications of the parameters or the program itself. Other than a brief overview of the system, this paper will not discuss the internal workings of the network flow system since that has been documented elsewhere. It will discuss how we achieved a very successful application of the methodology by a joint effort between agricultural and disclosure analysts. We were able to design an adaptable system that could respond to the inquiries and needs of the agricultural analysts. This resulted in a continual cycle of review and modification that ultimately yielded a high quality product. This paper will give the highlights of this cooperative effort in the context of the development of the input parameters, the diagnostics and the resulting modifications to the system.

2. Project Scope

The census of agriculture is a very large and complex project which is taken to obtain agricultural statistics for each county or county equivalent, state, and the Nation. The census of agriculture is the leading source of facts and statistics about the Nation's agricultural production and provides a detailed picture of U.S. farms and ranches every five years.

The final census product comprises publications for each of the 50 states and the U.S. The published data products are both hard-copy and web-based and include 61 U.S. tables, 51 all states tables, 3050 state level tables and 2550 all counties by state tables. These tables contain more than 18 million cells. These are the data cells that required disclosure analysis.

3. System Overview

You could think of the agriculture census data structure as a large two-dimensional table where the rows refer to the geographic areas (U. S., states, and counties) and the columns refer to different agricultural statistics. In this structure, there are 3129 rows and 6002 columns. There are 1885 linear relations which define how a column is the sum of other columns.

3.1. Program Logic

Fortunately, we can divide this data set into subsets and process them separately. For example, we can process the U. S. and states in one computer run and then process the counties in following runs. In addition, each linear relation can be used to create an individual sub-table that can be checked for disclosures. When we do this, the largest single sub-table we process has 255 rows and 50 columns.

These sub-tables are dependent because a statistic found in one sub-table often appears in other sub-tables. If it is suppressed in one table, it must be suppressed in the other tables in which it appears, and additional complementary suppressions may be chosen in the other tables to make sure the data cell is protected.

The disclosure methodology is applied to these sub-tables. Four input files define their structure. Two files provide the geographic row information which comprises the U.S. total, a grouping of states, an individual state, or the counties in a state. The third file defines in terms of matrix numbers the columns as the linear relations within the publication tables. A fourth file contains a record for each cell in the table and includes its value, initial suppression flag, and other information necessary for the disclosure analysis.

From these files a disclosure table (sub-table) is created and converted into a network, initial suppressions are identified, capacities and cost are calculated and the minimum cost flow subroutine called to determine the complementary suppressions. The data file records are updated to correspond to the new complementary suppressions. This entire process is repeated until all linear relations have been processed once. If we either suppress a cell or increase the protection on a cell which appears in an earlier processed table, backtracking must be done to recheck the earlier table for disclosures. The disclosure run is completed when all backtracking has been done.

3.2. Order of Runs

The census data are published at the U.S., New England Region, States and County levels. All of these levels could not be processed in a single disclosure run. The first run of the program assigned suppressions to the U.S., New England Region and the States. The rows of the disclosure tables referred to the U.S. total, the New England regional subtotal, and the fifty state totals; the columns referred to the different agricultural statistics. The six New England states summed to the New England subtotal.

After the suppression patterns for the states were finalized, we did a disclosure run for each of the fifty states to assign suppressions to the counties. The first row contained the state data and the other rows had the data for the counties; the columns again referred to the agriculture statistics. Since the state data in the first row had already been processed, all of the cells in the first row were frozen; that is, no new suppressions were added to the first row in these disclosure runs. Had these cells not been frozen, new state suppressions may have been added which would have required redoing the U.S. and state disclosure run.

4. Development and Analysis of the Linear Relations

The development of the linear relations was the most time consuming and labor intensive activity of the entire disclosure process. The linear relations describe the summation relationships that exist among the cells of the published census tables and are defined in terms of numeric cell identifiers.

The development of the linear relations could not begin until there was a stable draft of the census publication table shells. An inter-divisional team of agriculture analysts developed the publication table shells that described all aspects of the census tables including row and column descriptors, detailed tabulation instructions for each cell, and other information needed to program the summary and tabulation system. This effort took more than 18 months and produced 120 table shells.

Concurrent with this process was the creation of the numeric cell identifiers, referred to as matrix numbers, which were used to specify the linear relations. Each unique data cell was assigned a six-digit matrix number; if the cell appeared in multiple tables, it was assigned the same number in those tables. Similar agricultural commodities and characteristics were assigned numbers within a predetermined range to assist the analyst in their review of the disclosure data.

The linear relations were composed of a single matrix number that represented the summation cell followed by matrix numbers representing the component or interior cells that summed to the summation cell. The agricultural data relations were more complex than in earlier applications of the disclosure methodology where the structure was defined by North American Industry Classification System (NAICS) codes which had an inherent logical additive structure. Many of the agriculture relations are unstructured. Groups of cells which summed were not necessarily contiguous within a table and sometimes occurred over multiple tables which made the identification process difficult and required detailed subject matter knowledge. Because of these complexities and the large size of the project, a system of linear relation programs was developed to assist the agricultural analysts in their development of the relations. This aspect of the system was crucial to the successful application of network flow methodology to the census.

The linear relation system took two forms: output to assist the analyst during the development of the relations and programs to analyze the relations for efficient structure and organization. Since the relations were developed by various analysts over an extended period of time, it was important to consolidate the updating of the relation file through a single source; an easier way to add new relations and to update existing relations accomplished this. Data products were created to assist the analyst in developing the relations. They included a complete set of table shells with each cell populated with its matrix number, a matrix number dictionary that gave the description and detailed tabulation instructions in terms of the census questionnaire item numbers, a complete listing by matrix number of all publication tables containing the cell, and a listing of all linear relations giving their matrix numbers, verbal descriptions and tabulation instructions. Another listing provided by matrix number the published table number in which it appears. This listing sometimes revealed a difference in table numbers for related relations that upon investigation revealed a missing relation. Another listing gave for each matrix number the relations that use each number which was an additional help in identifying inconsistent patterns in matrix numbers and relations that could indicate missing or incomplete relations. All of these files were used by the analysts to develop and validate the linear relations logic.

Two files were extremely useful while developing the relations. Analysts reviewed a file of matrix numbers and their verbal descriptions which were not used in any relation; these were called independents. In most situations, it was clear from the verbal description whether the matrix number had been omitted from a relation. The other file was a listing of matrix numbers that are in relations but are not in the file of published matrix numbers. This file should be empty; if it were not, it indicated a problem with the relations.

The relations were grouped into independent blocks of matrix numbers; matrix numbers composing relations within a block were not used in relations outside the block. Initially, this was done to speed computer processing by isolating the run to self-contained groups of matrix numbers; however, this was not needed for speed. The blocking did allow for more efficient testing of the programs by isolating runs to single blocks. The disclosure table output was also blocked to make it easier for analysts to review since similar types of data appeared together.

While these activities assisted the analyst in their development of a complete relations list, the following were some of the activities used to analyze the listings and create a file of relations that would work best with the disclosure program to reduce the potential for extraneous complements.

4.1. Restructuring the Relations

Even though technically correct when considered individually, the relations may not be in the best structure for disclosure processing when considered as a group. Identifying these relations and modifying their structure helped reduce the number of complementary suppressions.

Assume we have these two linear relations, where the numbers in the relations refer to matrix number mentioned earlier.

Relation 1: $10=20+30+40+50+60$

Relation 2: $70=40+50+60$

To reduce the number of complementary suppressions, it would be better to change the first relation in this way.

Relation 1a: $10=20+30+70$

Suppose that matrix number 40 is a primary suppression; we may suppress matrix number 30 as a complement in the first relation. However, if the first relation is restructured as shown, there is no suppression to protect.

4.2. Reordering Relations

In certain situations, the relation processing order may contribute to the number of complements selected by the program. When a matrix number is a summation in one relation and an interior number in another relation, the relation in which it is interior should be processed first. As much as possible, these types of relations were identified and resorted in a logical top-down order to make sure the relations were ordered in this manner.

If the relations are not ordered in this way, it is easy to construct examples showing how the improper order of the relations may cause over-suppressions. Suppose that in relation 1a and 2 above the matrix numbers 30 and 50 are primary suppressions and the value of matrix number 20 is much larger than the value of matrix number 70. If relation 1a is processed first the matrix number 70 is chosen as a complement and no new complements are required in relation 2. However, if relation 2 had been processed first then either matrix number 40 or 60 would have been chosen and matrix number 70 would still have been chosen in relation 1a.

4.3. Combining Linear Relations

Some relations in the census have subcategories that are published together with an embedded subtotal. This occurs frequently with size breakouts such as for acreage when the higher acreage categories are subtotaled and all are published together in the same table. Under certain conditions this situation could create a disclosure. Suppose the subcategories are expressed as two relations and there is a one-respondent primary suppression in each relation. There is a risk that if the subtotal is chosen as a complement then the two one-respondent suppressed cells are left unprotected since either respondent can calculate the others value. If this had been checked as a combined relation, another cell would have been selected as a complement since the program will not let one-respondent primary suppressions protect each other. To avoid this potential problem, relations of this type are combined by the program prior to disclosure processing.

4.4. Complete File of Linear Relations

At the conclusion of this process, 1885 relations had been identified; the largest containing 50 matrix numbers. When combined with the geographic dimension this relation created the largest single disclosure table of more than 12,000 cells. These relations generated more than 96,000 disclosure tables that contained more than 18 million cells awaiting disclosure analysis.

The relations file showed all of the relations grouped into homogeneous blocks for easier access and review by analysts. The matrix numbers in each relation were printed along with a complete verbal description and detailed tabulation instructions.

5. Enhanced Features of the System

5.1 Capacity and Cost Parameters

When suppression patterns were unacceptable, rather than modifying the cost function in the minimum cost flow algorithm, we chose to modify its input parameters of capacity and cost. The capacity of a cell is the amount of protection a cell can give the initial suppression and the cost is the value that is assigned to each cell. The parameters were set so that the algorithm gave suppressions patterns that both adequately protected the initial suppressions while still preserving the most important statistics we wanted to publish. Some of these modifications are discussed below.

5.2. Freeze/Maximum Capacity Program

The agriculture census data set was too large to process in a single computer run, so we had to break it into subsets and process them separately. The first computer run checked the U. S. and state data for disclosures and assigned suppressions to many of the state-level statistics. Then we did disclosure runs to assign suppressions to the data for the counties within each state. In these computer runs, we still had to check each linear relation separately and make sure a value suppressed in one relation had complementary suppressions chosen to protect it in every other relation in which it appeared.

Given the way the program works, this would sometime lead to the program wanting to suppress additional values at the state level. Of course, we could not allow new state suppressions to be chosen because we would then have to redo the U.S. and state disclosure run. To address this problem, we decided to freeze all of the state-level suppressions after the U.S. and states disclosure run. Then we used a computer program to decide if any data for the counties must be frozen before the next set of disclosure runs were done.

For example, if a data cell for a state had not been suppressed and if it is equal in value to one of the counties in the state, then the county must also be frozen. If the newly frozen county cell is in a linear relation and if it is equal in value to the total of the relation, the total must also be frozen. If the data for a county is close in value to an unsuppressed state cell, then the county data may be suppressed during the computer run but its capacity to protect another suppression must be limited. As a result, the computer program for doing all of this is quite complicated, but it must be done when processing large data sets that are divided into subsets.

5.3. Cost Function Options

There are two cost function options that could be selected during the parameter settings for the disclosure run. The standard option does not modify the cost to take into consideration other linear relations when processing a specific relation. This is the setting that has been used in previous applications of the methodology. The modified option adjusts the cost based on whether suppressing that cell would help or hurt other tables that contain that cell; in other words, it attempts to 'look ahead' to see if a suppression would cause problems in other column relations. For example, if the other table had no suppressions in the row that contains the cell, then suppressing that cell would hurt the other table because we would then have to suppress another cell in that row; a much higher cost is given in this situation. If the other table had only one cell suppressed in that row, then suppressing the new cell would help the other table and a lower cost is assigned.

5.4. Preference Codes & Cost Adjustment Factors

We used the preference codes and cost adjustment factors to make the suppression patterns more to our liking. The preference codes were used to identify unpublished cells that can be suppressed with no harm to the publications or to identify cells that must not be suppressed, usually because they are frozen. The cost adjustment factors were used to assign relative levels of importance to cells based on the analysts' recommendations. The costs for cells of lower importance were decreased by incremental factors so that the cells would be more likely chosen as complementary suppressions. The costs for important cells were increased so that they would be less likely to be chosen. Sometimes analysts requested that the summation be suppressed before the interior cells. This was done by decreasing their cost relative to the interior cells' costs.

During the disclosure review analysts identified commodities that were of such importance at a specific geographic level that they should not be suppressed as complements if other patterns of complements could adequately protect the initial suppression. These commodities included such items as Valencia oranges in Florida or grapes and olives in California. The cost for these commodities was increased by a factor to discourage their selection as complements; however, their selection was not strictly forbidden in a situation where there were not other potential complements available. Numerous commodities were processed with this technique.

For some disclosure tables, analysts identified cells which were selected as complements that, because of their size relative to their row total, were important for that row. To decrease the chance that these types of cells were selected as complements their costs were increased by a factor if their disclosure table had more than three columns and the value of the cell was more than 80 percent of the row total.

6. Diagnostic Tools

The diagnostic files gave the analysts the ability to review the suppression patterns and to identify the situations that resulted in the modifications that have been described above. These files were developed to provide detailed information on the suppressions and why they were selected. The two major files were the disclosure table file and the suppression pattern file. The table file was used by both agricultural and disclosure analysts while the suppression pattern file was used mainly by the disclosure analysts. Both were used extensively throughout the disclosure process and were invaluable when explaining the reason why a particular suppression occurred.

The table file gave the disclosure tables for each of the 1885 relations and identified all the suppressed cells within each table. The table files were created for the US/State disclosure run and for each of the 50 State/County runs. The rows of the disclosure tables were the U.S./State or State/County geographies depending on the disclosure run and the columns were the matrix numbers for the specific relation. Each table contained the relations number, block number and verbal descriptions for the matrix numbers composing the relation. The disclosure tables were sorted by block so that similar agricultural commodities were grouped together for the analysts to review. The core of each table contained the published census values followed by the suppression flags which identified the primary and complementary suppressions.

Many relations shared the same matrix numbers due to the cell overlap between publication tables and the complexity within those tables. The table file showed all suppressions for the matrix numbers composing the specific relation regardless of when the suppression was selected. To facilitate the analytical review, the suppressions were coded to indicate whether they were selected when the current relation was processed, during processing of an earlier relation, a later relation or during backtracking. This was of particular help to the analyst as they reviewed the patterns. Prior to implementing this coding scheme we would receive questions about why such a large cell was chosen to protect a very small cell; when in fact, the large cell was required for protection of an initial suppression in another relation that shared a matrix number with the currently processed relation. The coding scheme increased the confidence among analysts that the system was operating properly.

Additional codes were included adjacent to the suppression flags to indicate the preference code or cost adjustment factor. Since these affect the cost of suppressing a cell, knowing these codes helped analysts understand why a particular cell was chosen as a complementary suppression. Each table file provided counts for the number and value of primary and complementary suppressions. Separate counts were given for primary suppressions with one or two respondents to help analysts determine whether the data were being sliced too thinly.

The suppression pattern file was used jointly with the disclosure table file and mainly by a disclosure analyst to determine the suppression patterns for complex situations. For each initial suppression the file lists all cells in the suppression pattern. The data included on the pattern file allows the analyst, even in the most complex situations, to reconstruct the actual suppression pattern. The selection of each complementary suppression can be justified. One of the most important pieces of information gave the suppression pattern in which the cell received its maximum protection or carried its greatest flow. This information was used to address the analysts' questions on why the protection was so high on an initial suppression. By examining the suppression pattern, one can determine which initial suppression pattern gave that protection to the cell. The file also gave the number of units flowing though the complementary suppression. This information was used to determine the importance of the cell relative to other cells in the pattern in protecting the initial suppression.

The joint use of these two files was essential to the success of the disclosure project. The ability to quickly respond to the inquiries from the agricultural analysts gave them confidence in the disclosure system. The analyst may not have always liked our answers but they could see why it happened. On numerous occasions their use of these files identified issues that resulted in a parameter modification or a change in the program. The benefit of this interaction to the success of this project cannot be overstated.

7. Performance Issues

The disclosure system is written in FORTRAN. During census production the disclosure system ran on an IBM R50 UNIX server with 1997 architecture. To run the disclosure system for the entire census which consisted of more than 18 million cells took 3.5 hours. We were able to run the entire census twice each day and review diagnostics which were extremely beneficial during the development and analytical review process. If we were testing specific aspects of the program we could isolate the run to specific relations or blocks and reduce the run time to minutes.

After the census we moved the disclosure system to an IBM P690 UNIX server with 32 processors and 132 gigabytes of memory. Because of the greater power of this machine, the entire census disclosure system ran in 20 minutes.

8. Future Work

All of the FORTRAN programs are being converted to SAS to make it easier for the agency to support the disclosure system in the future. The two-dimensional programs have been completed and run on the entire census with identical results as the FORTRAN programs. As expected, the SAS version runs slower taking approximately 10 times longer on the IBM P690 machine than the original FORTRAN version. However, this is approximately the run time experienced on the slower IBM R50 machine for the FORTRAN version during the census and is deemed to be acceptable. Currently efforts are continuing to convert the remaining programs to SAS.

References

Jewett, R.S. (2003) “Developing the Linear Relations”, *unpublished internal manuscript*, Washington, D.C.: USDA, National Agricultural Statistics Service, Census and Survey Division.

Jewett, R.S. (2004) “Disclosure Analysis for the 2002 Agriculture Census”, *unpublished internal manuscript*, Washington, D. C.: USDA, National Agricultural Statistics Service, Census and Survey Division.

Jewett, R.S. (2004) “Description of the 2002 Agriculture Census Disclosure System”, *unpublished internal manuscript*, Washington, D. C.: USDA, National Agricultural Statistics Service, Census and Survey Division.

National Agricultural Statistics Service, USDA (2004) “2002 Census of Agriculture, Summary and State Data, Volume 1, Geographic Area Series”.

Topic VII

General statistical confidentiality issues

Glossary on statistical disclosure control

Mark Elliot[*], *Anco Hundepool*[**], *Eric Schulte Nordholt*[**], *Jean-Louis Tambay*[***] *and Thomas Wende*[****]

[*] **University of Manchester**

[**] **Statistics Netherlands**

[***] **Statistics Canada**

[****] **Destatis, Germany**

Version September 2005

1. Introduction

At the Joint ECE / Eurostat Work Session on Statistical Data Confidentiality (7-9 April 2003) in Luxembourg the idea for a glossary on Statistical Disclosure Control was launched. The five people who produced this glossary were present at that Work Session and also met on 18 August 2003 at the ISI Session in Berlin. This new glossary on Statistical Disclosure Control will be presented at the next Joint ECE / Eurostat Work Session on Statistical Data Confidentiality (9-11 November, 2005) in Geneva. In the meantime preliminary versions have been presented so that experts in the field from all over the world could comment on these versions. The aim of this glossary is twofold: firstly, it should help people who are new in the field to get acquainted with the terminology used in Statistical Disclosure Control and secondly it can be used in courses on Statistical Disclosure Control as a back-up facility. We hope that this glossary will be useful and the two aims will be reached. If you have any comments or questions, please forward them to Eric Schulte Nordholt (e-mail: ESLE@CBS.NL) so that they can be taken into account for future versions.

Acknowledgements

The authors acknowledge the input of many people during the drafting of this document, especially: Paul Feuvrier (INSEE), Kingsley Purdam (University of Manchester), Barry Schouten (Statistics Netherlands), Duncan Smith (University of Manchester) and Peter-Paul de Wolf (Statistics Netherlands).

A

Ambiguity rule: Synonym of (p,q) rule.

Analysis server: A form of **remote data laboratory** designed to run analysis on data stored on a safe server. The user sees the results of their analysis but not the data.

Anonymised data: Data containing only anonymised records.

Anonymised record: A record from which direct identifiers have been removed.

Approximate disclosure: Approximate disclosure happens if a user is able to determine an estimate of a respondent value that is close to the real value. If the estimator is exactly the real value the disclosure is exact.

Argus: Two software packages for Statistical Disclosure Control are called Argus. μ-Argus is a specialized software tool for the protection of **microdata**. The two main techniques used for this are **global recoding** and **local suppression**. In the case of **global recoding** several categories of a variable are collapsed into a single one. The effect of local suppression is that one or more values in an unsafe combination are suppressed, i.e. replaced by a missing value. Both **global recoding** and **local suppression** lead to a loss of information, because either less detailed information is provided or some information is not given at all. τ-Argus is a specialized software tool for the protection of tabular data. τ-Argus is used to produce safe tables. τ-Argus uses the same two main techniques as μ-Argus: global recoding and local suppression. For τ-Argus the latter consists of suppression of cells in a table.

Attribute disclosure: Attribute disclosure is **attribution** independent of identification. This form of disclosure is of primary concern to **NSIs** involved in **tabular data** release and arises from the presence of empty cells either in a released table or linkable set of tables after any subtraction has taken place. Minimally, the presence of an empty cell within a table means that an intruder may infer from mere knowledge that a population unit is represented in the table and that the intruder does not possess the combination of attributes within the empty cell.

Attribution: Attribution is the association or disassociation of a particular attribute with a particular population unit.

B

Barnardisation: A method of disclosure control for tables of counts that involves randomly adding or subtracting 1 from some cells in the table.

Blurring: Blurring replaces a reported value by an average. There are many possible ways to implement blurring. Groups of records for averaging may be formed by matching on other variables or by sorting on the variable of interest. The number of records in a group (whose data will be averaged) may be fixed or random. The average associated with a particular group may be assigned to all members of a group, or to the "middle" member (as in a moving average). It may be performed on more than one variable with different groupings for each variable.

Bottom coding: See **top and bottom coding**.

Bounds: The range of possible values of a cell in a table of frequency counts where the cell value has been perturbed or suppressed. Where only margins of tables are released it is possible to infer bounds for the unreleased joint distribution. One method for inferring the bounds across a table is known as the **Shuttle algorithm**.

C

Calculated interval: The interval containing possible values for a suppressed cell in a table, given the table structure and the values published.

Cell suppression: In tabular data the cell suppression SDC method consists of **primary and complementary (secondary) suppression**. Primary suppression can be characterised as withholding the values of all risky cells from publication, which means that their value is not shown in the table but replaced by a symbol such as '×' to indicate the suppression. According to the definition of risky cells, in frequency count tables all cells containing small counts and in tables of magnitudes all cells containing small counts or presenting a case of **dominance** have to be primary suppressed. To reach the desired protection for risky cells, it is necessary to suppress additional non-risky cells, which is called **complementary (secondary) suppression**. The pattern of complementary suppressed cells has to be carefully chosen to provide the desired level of ambiguity for the risky cells with the least amount of suppressed information.

Complementary suppression: Synonym of **secondary suppression**.

Complete disclosure: Synonym of **exact disclosure**.

Concentration rule: Synonym of **(n,k) rule**.

Confidentiality edit: The confidentiality edit is a procedure developed by the U.S. Census Bureau to provide protection in data tables prepared from the 1990 Census. There are two different approaches: one was used for the regular Census data; the other was used for the long-form data, which were filled by a sample of the population. Both techniques apply statistical disclosure limitation techniques to the **microdata** files before they are used to prepare tables. The adjusted files themselves are not released; they are used only to prepare tables. For the regular Census microdata file, the confidentiality edit involves "data swapping" or "switching" of attributes between matched records from different geographical units. For small blocks, the Census Bureau increases the sampling fraction. After the microdata file has been treated in this way, it can be used directly to prepare tables and no further disclosure analysis is needed. For long form data, sampling provides sufficient confidentiality protection, except in small geographic regions. To provide additional protection in small geographic regions, one household is randomly selected and a sample of its data fields are blanked and replaced by imputed values.

Controlled rounding: To solve the additivity problem, a procedure called controlled rounding was developed. It is a form of **random rounding**, but it is constrained to have the sum of the published entries in each row and column equal to the appropriate published marginal totals. Linear programming methods are used to identify a controlled rounding pattern for a table.

Controlled Tabular Adjustment (CTA): A method to protect tabular data based on the selective adjustment of cell values. Sensitive cell values are replaced by either of their closest safe values and small adjustments are made to other cells to restore the table additivity. Controlled tabular adjustment has been developed as an alternative to **cell suppression**.

Conventional rounding: A disclosure control method for tables of counts. When using conventional rounding, each count is rounded to the nearest multiple of a fixed base. For example, using a base of 5, counts ending in 1 or 2 are rounded down and replaced by counts ending in 0 and counts ending in 3 or 4 are rounded up and replaced by counts ending in 5. Counts ending between 6 and 9 are treated similarly. Counts with a last digit of 0 or 5 are kept unchanged. When rounding to base 10, a count ending in 5 may always be rounded up, or it may be rounded up or down based on a rounding convention.

D

Data divergence: The sum of all differences between two datasets (data-data divergence) or between a single dataset and reality (data-world divergence). Sources of data divergence include: data ageing, response errors, coding or data entry errors, differences in coding and the effect of disclosure control.

Data intruder: A data user who attempts to disclose information about a population unit through **identification** or **attribution**.

Data intrusion detection. The detection of a **data intruder** through their behaviour. This is most likely to occur through analysis of a pattern of requests submitted to a **remote data laboratory**. At present this is only a theoretical possibility, but it is likely to become more relevant as **virtual safe settings** become more prevalent.

Data Intrusion Simulation (DIS). A method of estimating the probability that a **data intruder** who has matched an arbitrary population unit against a sample unique in a target microdata file has done so correctly.

Data protection: Data protection refers to the set of privacy-motivated laws, policies and procedures that aim to minimise intrusion into respondents' privacy caused by the collection, storage and dissemination of personal data.

Data swapping: A disclosure control method for microdata that involves swapping the values of variables for records that match on a representative key. In the literature this technique is also sometimes referred to as "multidimensional transformation". It is a transformation technique that guarantees (under certain conditions) the maintenance of a set of statistics, such as means, variances and univariate distributions.

Data utility: A summary term describing the value of a given data release as an analytical resource. This comprises the data's **analytical completeness** and its **analytical validity**. Disclosure control methods usually have an adverse effect on data utility. Ideally, the goal of any disclosure control regime should be to maximise data utility whilst minimising disclosure risk. In practice disclosure control decisions are a trade-off between utility and **disclosure risk**.

Deterministic rounding: Synonym of **conventional rounding**.

Direct identification: Identification of a statistical unit from its **formal identifiers**.

Disclosive cells: Synonym of **risky cells**.

Disclosure: Disclosure relates to the inappropriate attribution of information to a data subject, whether an individual or an organisation. Disclosure has two components: **identification** and **attribution**.

Disclosure by fishing: This is an attack method where an intruder identifies risky records within a target data set and then attempts to find population units corresponding to those records. It is the type of disclosure that can be assessed through a **special uniques analysis**.

Disclosure by matching: Disclosure by the linking of records within an identification dataset with those in an anonymised dataset.

Disclosure by response knowledge: This is disclosure resulting from the knowledge that a person was participating in a particular survey. If an intruder knows that a specific individual has participated in the survey, and that consequently his or her data are in the data set, identification and disclosure can be accomplished more easily.

Disclosure by spontaneous recognition: This means the recognition of an individual within the dataset. This may occur by accident or because a data intruder is searching for a particular individual. This is more likely to be successful if the individual has a rare combination of characteristics which is known to the intruder.

Disclosure control methods: There are two main approaches to control the disclosure of confidential data. The first is to reduce the information content of the data provided to the external user. For the release of tabular data this type of technique is called **restriction-based disclosure control method** and for the release of microdata the expression disclosure control by data reduction is used. The second is to change the data before the dissemination in such a way that the disclosure risk for the confidential data is decreased, but the information content is retained as much as possible. These are called **perturbation based disclosure control methods**.

Disclosure from analytical outputs: The use of output to make attributions about individual population units. This situation might arise to users that can interrogate data but do not have direct access to them such as in a **remote data laboratory**. One particular concern is the publication of residuals.

Disclosure limitation methods: Synonym of **disclosure control methods**.

Disclosure risk: A disclosure risk occurs if an unacceptably narrow estimation of a respondent's confidential information is possible or if exact disclosure is possible with a high level of confidence.

Disclosure scenarios: Depending on the intention of the intruder, his or her type of a priori knowledge and the microdata available, three different types of disclosure or disclosure scenarios are possible for microdata: **disclosure by matching, disclosure by response knowledge** and **disclosure by spontaneous recognition**.

Dissemination: Supply of data in any form whatever: publications, access to databases, microfiches, telephone communications, etc.

Disturbing the data: This process involves changing the data in some systematic fashion, with the result that the figures are insufficiently precise to disclose information about individual cases.

Dominance rule: Synonym of **(n,k) rule**.

E

Exact disclosure: Exact disclosure occurs if a user is able to determine the exact attribute for an individual entity from released information.

F

Formal identifier: Any variable or set of variables which is structurally unique for every population unit, for example a population registration number. If the formal identifier is known to the intruder, identification of a target individual is directly possible for him or her, without the necessity to have additional knowledge before studying the microdata. Some combinations of variables such as name and address are pragmatic formal identifiers, where non-unique instances are empirically possible, but with negligible probability.

G

Global recoding: Problems of confidentiality can be tackled by changing the structure of data. Thus, rows or columns in tables can be combined into larger class intervals or new groupings of characteristics. This may be a simpler solution than the suppression of individual items, but it tends to reduce the descriptive and analytical value of the table. This protection technique may also be used to protect microdata.

H

HITAS: A heuristic approach to **cell suppression** in hierarchical tables.

I

Identification: Identification is the association of a particular record within a set of data with a particular population unit.

Identification dataset: A dataset that contains formal identifiers.

Identification data: Those personal data that allow direct identification of the data subject, and which are needed for the collection, checking and matching of the data, but are not subsequently used for drawing up statistical results.

Identification key: Synonym of **key**.

Identification risk: This risk is defined as the probability that an intruder identifies at least one respondent in the disseminated microdata. This identification may lead to the disclosure of (sensitive) information about the respondent. The risk of identification depends on the number and nature of **quasi-identifiers** in the microdata and in the a priori knowledge of the intruder.

Identifying variable: A variable that either is a formal identifier or forms part of a formal identifier.

Indirect identification: Inferring the identity of a population unit within a microdata release other than from **direct identification**.

Inferential disclosure: Inferential disclosure occurs when information can be inferred with high confidence from statistical properties of the released data. For example, the data may show a high correlation between income and purchase price of home. As the purchase price of a home is typically public information, a third party might use this information to infer the income of a data subject. In general, NSIs are not concerned with inferential disclosure for two reasons. First, a major purpose of statistical data is to enable users to infer and understand relationships between variables. If NSIs equated disclosure with inference, no data could be released. Second, inferences are designed to predict aggregate behaviour, not individual attributes, and thus often poor predictors of individual data values.

Informed consent: Basic ethical tenet of scientific research on human populations. Sociologists do not involve a human being as a subject in research without the informed consent of the subject or the subject's legally authorized representative, except as otherwise specified. Informed consent refers to a person's agreement to allow personal data to be provided for research and statistical purposes. Agreement is based on full exposure of the facts the person needs to make the decision intelligently, including awareness of any risks involved, of uses and users of the data, and of alternatives to providing the data.

Intruder: A data user who attempts to link a respondent to a microdata record or make attributions about particular population units from aggregate data. Intruders may be motivated by a wish to discredit or otherwise harm the NSI, the survey or the government in general, to gain notoriety or publicity, or to gain profitable knowledge about particular respondents.

J

K

Key: A set of **key variables**.

Key variable: A variable in common between two datasets, which may therefore be used for linking records between them. A key variable can either be a **formal identifier** or a **quasi-identifier**.

L

Licensing agreement: A permit, issued under certain conditions, for researchers to use confidential data for specific purposes and for specific periods of time. This agreement consists of contractual and ethical obligations, as well as penalties for improper disclosure or use of identifiable information. These penalties can vary from withdrawal of the license and denial of access to additional data sets to the forfeiting of a deposit paid prior to the release of a **microdata** file. A licensing agreement is almost always combined with the signing of a contract. This contract includes a number of requirements: specification of the intended use of the data; instruction not to release the **microdata** file to another recipient; prior review and approval by the releasing agency for all user outputs to be published or disseminated; terms and location of access and enforceable penalties.

Local recoding: A disclosure control technique for microdata where two (or more) different versions of a variable are used dependent on some other variable. The different versions will have different levels of coding. This will depend on the distribution of the first variable conditional on the second. A typical example occurs where the distribution of a variable is heavily skewed in some geographical areas. In the areas where the distribution is skewed minor categories may be combined to produce a courser variable.

Local suppression: Protection technique that diminishes the risk of recognition of information about individuals or enterprises by suppressing individual scores on **identifying variables**.

Lower bound: The lowest possible value of a cell in a table of frequency counts where the cell value has been perturbed or suppressed.

M

Macrodata: Synonym of **tabular data**.

Microaggregation: Records are grouped based on a proximity measure of variables of interest, and the same small groups of records are used in calculating aggregates for those variables. The aggregates are released instead of the individual record values.

Microdata: A microdata set consists of a set of records containing information on individual respondents or on economic entities.

Minimal unique: A combination of variable values that are unique in the **microdata** set at hand and contain no proper subset with this property (so it is a minimal set with the uniqueness property).

N

NSI(s): Abbreviation for National Statistical Institute(s).

(n,k) rule: A cell is regarded as confidential, if the n largest units contribute more than k % to the cell total, e.g. n=2 and k=85 means that a cell is defined as risky if the two largest units contribute more than 85 % to the cell total. The n and k are given by the statistical authority. In some **NSIs** the values of n and k are confidential.

O

On-site facility: A facility that has been established on the premises of several NSIs. It is a place where external researchers can be permitted access to potentially disclosive data under contractual agreements which cover the maintenance of confidentiality, and which place strict controls on the uses to which the data can be put. The on-site facility can be seen as a 'safe setting' in which confidential data can be analysed. The on-site facility itself would consist of a secure hermetic working and data storage environment in which the confidentiality of the data for research can be ensured. Both the physical and the IT aspects of security would be considered here. The on-site facility also includes administrative and support facilities to external users, and ensures that the agreed conditions for access to the data were complied with.

Ordinary rounding: Synonym of **conventional rounding**.

Oversuppression: A situation that may occur during the application of the technique of cell suppression. This denotes the fact that more information has been suppressed than strictly necessary to maintain confidentiality.

P

Partial disclosure: Synonym of **approximate disclosure**.

Passive confidentiality: For foreign trade statistics, EU countries generally apply the principle of "passive confidentiality", that is they take appropriate measures only at the request of importers or exporters who feel that their interests would be harmed by the dissemination of data.

Personal data: Any information relating to an identified or identifiable natural person ('data subject'). An identifiable person is one who can be identified, directly or indirectly. Where an individual is not identifiable, data are said to be anonymous.

Perturbation based disclosure control methods:

Techniques for the release of data that change the data before the dissemination in such a way that the disclosure risk for the confidential data is decreased but the information content is retained as far as possible. Perturbation based methods falsify the data before publication by introducing an element of error purposely for confidentiality reasons. For example, an error can be inserted in the cell values after a table is created, which means that the error is introduced to the output of the data and will therefore be referred to as output perturbation. The error can also be inserted in the original data on the **microdata** level, which is the input of the tables one wants to create; the method will then be referred to as data perturbation - input perturbation being the better but uncommonly used expression. Possible perturbation methods are:

- rounding;
- perturbation, for example, by the addition of random noise or by the **Post Randomisation Method**;
- disclosure control methods for microdata applied to tabular data.

Population unique: A record within a dataset which is unique within the population on a given **key**.

P-percent rule: A **(p,q) rule** where q is 100 %, meaning that from general knowledge any respondent can estimate the contribution of another respondent to within 100 % (i.e., knows the value to be nonnegative and less than a certain value which can be up to twice the actual value).

(p,q) rule: It is assumed that out of publicly available information the contribution of one individual to the cell total can be estimated to within q per cent (q=error before publication); after the publication of the statistic the value can be estimated to within p percent (p=error after publication). In the (p,q) rule the ratio p/q represents the information gain through publication. If the information gain is unacceptable the cell is declared as confidential. The parameter values p and q are determined by the statistical authority and thus define the acceptable level of information gain. In some **NSIs** the values of p and q are confidential.

Post Randomisation Method (PRAM): Protection method for microdata in which the scores of a categorial variable are changed with certain probabilities into other scores. It is thus intentional misclassification with known misclassification probabilities.

Primary confidentiality: It concerns tabular cell data, whose dissemination would permit attribute disclosure. The two main reasons for declaring data to be primary confidential are:

- too few units in a cell;
- dominance of one or two units in a cell.

The limits of what constitutes "too few" or "dominance" vary between statistical domains.

Primary protection: Protection using disclosure control methods for all cells containing small counts or cases of dominance.

Primary suppression: This technique can be characterized as withholding all disclosive cells from publication, which means that their value is not shown in the table, but replaced by a symbol such as '×' to indicate the suppression. According to the definition of disclosive cells, in frequency count tables all cells containing small counts and in tables of magnitudes all cells containing small counts or representing cases of dominance have to be primary suppressed.

Prior-posterior rule: Synonym of the **(p,q) rule**.

Privacy: Privacy is a concept that applies to data subjects while confidentiality applies to data. The concept is defined as follows: "It is the status accorded to data which has been agreed upon between the person or organisation furnishing the data and the organisation receiving it and which describes the degree of protection which will be provided." There is a definite relationship between confidentiality and privacy. Breach of confidentiality can result in disclosure of data which harms the individual. This is an attack on privacy because it is an intrusion into a person's self-determination on the way his or her personal data are used. Informational privacy encompasses an individual's freedom from excessive intrusion in the quest for information and an individual's ability to choose the extent and circumstances under which his or her beliefs, behaviours, opinions and attitudes will be shared with or withheld from others.

Probability based disclosures (approximate or exact): Sometimes although a fact is not disclosed with certainty, the published data can be used to make a statement that has a high probability of being correct.

Q

Quasi-identifier: Variable values or combinations of variable values within a dataset that are not structural uniques but might be empirically unique and therefore in principle uniquely identify a population unit.

R

Randomized response: Randomized response is a technique used to collect sensitive information from individuals in such a way that survey interviewers and those who process the data do not know which of two alternative questions the respondent has answered.

Random perturbation: This is a disclosure control method according to which a noise, in the form of a random value is added to the true value or, in the case of categorical variables, where another value is randomly substituted for the true value.

Random rounding: In order to reduce the amount of data loss that occurs with suppression, alternative methods have been investigated to protect sensitive cells in tables of frequencies. Perturbation methods such as random rounding and controlled rounding are examples of such alternatives. In random rounding cell values are rounded, but instead of using standard rounding conventions a random decision is made as to whether they will be rounded up or down. The rounding mechanism can be set up to produce unbiased rounded results.

Rank swapping: Rank swapping provides a way of using continuous variables to define pairs of records for swapping. Instead of insisting that variables match (agree exactly), they are defined to be close based on their proximity to each other on a list sorted on the continuous variable. Records which are close in rank on the sorted variable are designated as pairs for swapping. Frequently in rank swapping the variable used in the sort is the one that will be swapped.

Record linkage process: Process attempting to classify pairs of matches in a product space A×B from two files A and B into M, the set of true links, and U, the set of non-true links.

Record swapping: A special case of **data swapping,** where the geographical codes of records are swapped.

Remote access: On-line access to protected microdata.

Remote data laboratory: A virtual environment providing remote execution facilities.

Remote execution: Submitting scripts on-line for execution on disclosive microdata stored within an institute's protected network. If the results are regarded as **safe data**, they are sent to the submitter of the script. Otherwise, the submitter is informed that the request cannot be acquiesced. Remote execution may either work through submitting scripts for a particular statistical package such as SAS, SPSS or STATA which runs on the remote server or via a tailor made client system which sits on the user's desk top.

Residual disclosure: Disclosure that occurs by combining released information with previously released or publicly available information. For example, tables for nonoverlapping areas can be subtracted from a larger region, leaving confidential residual information for small areas.

Restricted access: Imposing conditions on access to the **microdata**. Users can either have access to the whole range of raw protected data and process individually the information they are interested in - which is the ideal situation for them - or their access to the protected data is restricted and they can only have a certain number of outputs (e.g. tables) or maybe only outputs of a certain structure. Restricted access is sometimes necessary to ensure that linkage between tables cannot happen.

Restricted data: Synonym of **safe data**.

Restriction based disclosure control method: Method for the release of **tabular data**, which consists in reducing access to the data provided to the external user. This method reduces the content of information provided to the user of the **tabular data**. This is implemented by not publishing all the figures derived from the collected data or by not publishing the information in as detailed a form as would be possible.

Risky cells: The cells of a table which are non-publishable due to the risk of statistical disclosure are referred to as risky cells. By definition there are three types of risky cells: small counts, dominance and complementary suppression cells.

Risky data: Data are considered to be disclosive when they allow statistical units to be identified, either directly or indirectly, thereby disclosing individual information. To determine whether a statistical unit is identifiable, account shall be taken of all the means that might reasonably be used by a third party to identify the said statistical unit.

Rounding: Rounding belongs to the group of disclosure control methods based on output-perturbation. It is used to protect small counts in **tabular data** against disclosure. The basic idea behind this disclosure control method is to round each count up or down either deterministically or probabilistically to the nearest integer multiple of a rounding base. The additive nature of the table is generally destroyed by this process. Rounding can also serve as a **recoding** method for microdata.

R-U map: A graphical representation of the trade off between disclosure risk and data utility.

S

Safe data: Microdata or **macrodata** that have been protected by suitable **Statistical Disclosure Control** methods.

Safe setting: An environment such as a **microdata** lab whereby access to a disclosive dataset can be controlled.

Safety interval: The minimal **calculated interval** that is required for the value of a cell that does not satisfy the primary suppression rule.

Sample unique: A record within a dataset which is unique within that dataset on a given **key**.

Sampling: In the context of disclosure control, this refers to releasing only a proportion of the original data records on a **microdata** file.

Sampling fraction: The proportion of the population contained within a data release. With simple random sampling, the sample fraction represents the proportion of population units that are selected in the sample. With more complex sampling methods, this is usually the ratio of the number of units in the sample to the number of units in the population from which the sample is selected.

Scenario analysis: A set of pseudo-criminological methods for analysing and classifying the plausible risk channels for a data intrusion. The methods are based around first delineating the means, motives and opportunity that an intruder may have for conducting the attack. The output of such an analysis is a specification of a set of **keys** likely to be held by **data intruders**.

Secondary data intrusion: After an attempt to match between identification and target datasets an intruder may discriminate between non-unique matches by further direct investigations using additional variables.

Secondary disclosure risk: It concerns data which is not primary disclosive, but whose dissemination, when combined with other data permits the identification of a microdata unit or the disclosure of a unit's attribute.

Secondary suppression: To reach the desired protection for risky cells, it is necessary to suppress additional non-risky cells, which is called secondary suppression or complementary suppression. The pattern of complementary suppressed cells has to be carefully chosen to provide the desired level of ambiguity for the disclosive cells at the highest level of information contained in the released statistics.

Security: An efficient disclosure control method provides protection against exact disclosure or unwanted narrow estimation of the attributes of an individual entity, in other words, a useful technique prevents exact or partial disclosure. The security level is accordingly high. In the case of disclosure control methods for the release of **microdata** this protection is ensured if the identification of a respondent is not possible, because the identification is the prerequisite for disclosure.

Sensitive cell: Cell for which knowledge of the value would permit an unduly accurate estimate of the contribution of an individual respondent. Sensitive cells are identified by the application of a dominance rule such as the (n,k) rule or the (p,q) rule to their microdata.

Sensitive variables: Variables contained in a data record apart from the key variables, that belong to the private domain of respondents who would not like them to be disclosed. There is no exact definition given for what a 'sensitive variable' is and therefore, the division into key and sensitive variables is somehow arbitrary. Some data are clearly sensitive such as the possession of a criminal record, one's medical condition or credit record, but there are other cases where the distinction depends on the circumstances, e.g. the income of a person might be regarded as a sensitive variable in some countries and as quasi-identifier in others, or in some societies the religion of an individual might count as a key and a sensitive variable at the same time. All variables that contain one or more sensitive categories, i.e. categories that contain sensitive information about an individual or enterprise, are called sensitive variables.

Shuttle algorithm: A method for finding lower and upper cell bounds by iterating through dependencies between cell counts. There exist many dependencies between individual counts and aggregations of counts in contingency tables. Where not all individual counts are known, but some aggregated counts are known, the dependencies can be used to make inferences about the missing counts. The Shuttle algorithm constructs a specific subset of the many possible dependencies and recursively iterates through them in order to find bounds on missing counts. As many dependencies will involve unknown counts, the dependencies need to be expressed in terms of inequalities involving lower and upper bounds, rather than simple equalities. The algorithm ends when a complete iteration fails to tighten the bounds on any cell counts.

Special uniques analysis: A method of analysing the per-record risk of **microdata**.

Statistical confidentiality: The protection of data that relate to single statistical units and are obtained directly for statistical purposes or indirectly from administrative or other sources against any breach of the right to confidentiality. It implies the prevention of unlawful disclosure.

Statistical Data Protection (**SDP):** Statistical Data Protection is a more general concept which takes into account all steps of production. SDP is multidisciplinary and draws on computer science (data security), statistics and operations research.

Statistical disclosure: Statistical disclosure is said to take place if the dissemination of a statistic enables the external user of the data to obtain a better estimate for a confidential piece of information than would be possible without it.

Statistical Disclosure Control (SDC): Statistical Disclosure Control techniques can be defined as the set of methods to reduce the risk of disclosing information on individuals, businesses or other organisations. Such methods are only related to the dissemination step and are usually based on restricting the amount of or modifying the data released.

Statistical Disclosure Limitation (**SDL):** Synonym of **Statistical Disclosure Control**.

Subadditivity: One of the properties of the (n,k) rule or (p,q) rule that assists in the search for complementary cells. The property means that the sensitivity of a union of disjoint cells cannot be greater than the sum of the cells' individual sensitivities (triangle inequality). Subadditivity is an important property because it means that aggregates of cells that are not sensitive are not sensitive either and do not need to be tested.

Subtraction: The principle whereby an intruder may attack a table of population counts by removing known individuals from the table. If this leads to the presence of certain zeroes in the table then that table is vulnerable to **attribute disclosure**.

Suppression: One of the most commonly used ways of protecting sensitive cells in a table is via suppression. It is obvious that in a row or column with a suppressed sensitive cell, at least one additional cell must be suppressed, or the value in the sensitive cell could be calculated exactly by **subtraction** from the marginal total. For this reason, certain other cells must also be suppressed. These are referred to as **secondary suppressions**. While it is possible to select cells for secondary suppression manually, it is difficult to guarantee that the result provides adequate protection.

SUDA: A software system for conducting analyses on population uniques and special sample uniques. The **special uniques analysis** method implemented in SUDA for measuring and assessing disclosure risk is based on resampling methods and used by the ONS.

Swapping (or switching): Swapping (or switching) involves selecting a sample of the records, finding a match in the data base on a set of predetermined variables and swapping all or some of the other variables between the matched records. Swapping (or switching) was illustrated as part of the confidentiality edit for tables of frequency data.

Synthetic data: An approach to confidentiality where instead of disseminating real data, synthetic data that have been generated from one or more population models are released.

Synthetic substitution: See **Controlled Tabular Adjustment**.

T

Table server: A form of **remote data laboratory** designed to release safe tables.

Tables of frequency (count) data: These tables present the number of units of analysis in a cell. When data are from a sample, the cells may contain weighted counts, where weights are used to bring sample results to the population levels. Frequencies may also be represented as percentages.

Tables of magnitude data: Tables of magnitude data present the aggregate of a "quantity of interest" over all units of analysis in the cell. When data are from a sample, the cells may contain weighted aggregates, where quantities are multiplied by units' weights to bring sample results up to population levels. The data may be presented as averages by dividing the aggregates by the number of units in their cells.

Tabular data: Aggregate information on entities presented in tables.

Target dataset: An anonymised dataset in which an intruder attempts to identify particular population units.

Threshold rule: Usually, with the threshold rule, a cell in a table of frequencies is defined to be sensitive if the number of respondents is less than some specified number. Some agencies require at least five respondents in a cell, others require three. When thresholds are not respected, an agency may restructure tables and combine categories or use cell suppression, rounding or the confidentiality edit, or provide other additional protection in order to satisfy the rule.

Top and bottom coding: It consists in setting top-codes or bottom-codes on quantitative variables. A top-code for a variable is an upper limit on all published values of that variable. Any value greater than this upper limit is replaced by the upper limit or is not published on the **microdata** file at all. Similarly, a bottom-code is a lower limit on all published values for a variable. Different limits may be used for different quantitative variables, or for different subpopulations.

U

Union unique A sample unique that is also population unique. The proportion of sample uniques that are union uniques is one measure of file level disclosure risk.

Uniqueness: The term is used to characterise the situation where an individual can be distinguished from all other members in a population or sample in terms of information available on **microdata** records (or within a given **key**). The existence of uniqueness is determined by the size of the population or sample and the degree to which it is segmented by geographic information and the number and detail of characteristics provided for each unit in the dataset (or within the key).

Upper bound: The highest possible value of a cell in a table of frequency counts where the cell value has been perturbed or suppressed.

V

Virtual safe setting: Synonym of **remote data laboratory**.

W

Waiver approach: Instead of suppressing tabular data, some agencies ask respondents for permission to publish cells even though doing so may cause these respondents' sensitive information to be estimated accurately. This is referred to as the waiver approach. Waivers are signed records of the respondents' granting permission to publish such cells. This method is most useful with small surveys or sets of tables involving only a few cases of dominance, where only a few waivers are needed. Of course, respondents must believe that their data are not particularly sensitive before they will sign waivers.

X

Y

Z

Providing access to data and making microdata safe, experiences of the ONS

Paul Jackson and Jane Longhurst
Office for National Statistics (ONS), UK

Abstract. This paper provides an overview of how the ONS is tackling the problem of balancing the need to provide users with access to microdata and the need to protect the confidentiality of respondents. Issues involved with the process of providing access to microdata are addressed and the interaction between these processes. The legal and policy framework in the UK, risk analysis and management, SDC methods and the development of different access options are covered.

Keywords. Microdata, risk-utility plot, risk assessment, licence arrangements

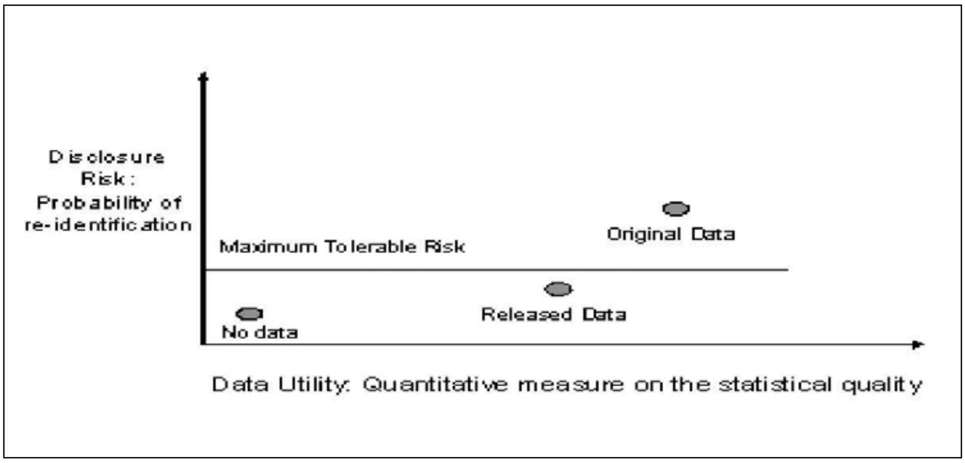
1. Introduction

There is a strong, widespread and increasing demand for National Statistics Institutes (NSIs) to release microdata files. It is important to meet this need with microdata that is as detailed as possible in order to support a wide range of valuable research. However, this interest may appear to conflict with the obligation that NSIs have to protect the confidentiality of the information provided by the respondents. As well as demand increasing it is internationally accepted that the threats to the confidentiality of microdata are also increasing. This paper provides an overview of how the Office for National Statistics (ONS) in the UK is tackling this problem of balancing the need to provide users with access to microdata and the need to protect the confidentiality of respondents. A framework has been developed for protecting and providing access to microdata at ONS covering key issues that must be addressed when making decisions on confidentiality protection for microdata. The idea of balancing disclosure risk with data utility forms the basis for the framework.

2. Risk-Utility Approach

The framework for protecting and providing access to microdata is based on a disclosure risk-data utility decision problem approach. This approach determines optimal methods that minimize the disclosure risk while maximizing the utility of the data. Figure 1 contains an R-U confidentiality map developed by Duncan, et. al. (2001) where R is a quantitative measure of disclosure risk and U is a quantitative measure of data utility.

Figure 1. R-U Confidentiality Map (Duncan, et.al. (2001))


Disclosure Risk: Probability of re-identification
Maximum Tolerable Risk
Original Data
Released Data
No data
Data Utility: Quantitative measure on the statistical quality

In the lower left hand quadrant of the graph low disclosure risk is achieved but also low utility, where no data is released at all. In the upper right hand quadrant of the graph high disclosure risk is achieved but also high utility, represented by the point where the original data is released. The NSI must set the maximum tolerable disclosure risk based on standards, policies and guidelines.

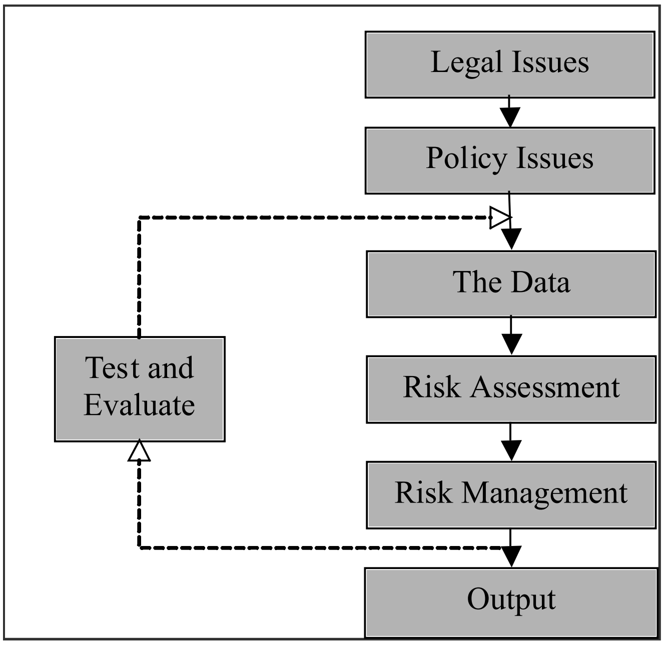
Approaches to risk analysis for microdata are described in Section 7 of the paper and Section 6 covers the importance of assessing data utility. The goal in the disclosure risk–data utility decision problem is to find the balance in maintaining the utility of the data but reducing the risk below the maximum tolerable threshold. The tools used to reduce this risk can involve applying statistical disclosure control methods or restricting access to the data or a combination of both. These are all covered in Section 8 of the paper.

3. A Framework for Protecting and Providing Access to Microdata

A framework that has been developed for protecting and providing access to microdata at ONS covers the key issues that must be addressed when making decisions on confidentiality protection for microdata. This is based on a generic framework that has been developed by the ONS for decisions on confidentiality protection, Bycroft and Lowthian (2005). The figure below provides an outline of the framework.

The first two stages involve establishing why confidentiality protection is needed - usually legal or policy factors. The next three stages relate to the risk-utility approach. The assessments of risk and utility must be taken into consideration at the risk management stage where the risk in the data is reduced to an acceptable level. Note, this process is iterative following the application of a method to reduce the disclosure risk further assessments of risk and utility should be undertaken until a solution or balance is found.

Figure 2. A Framework for Protecting and Providing Access to Microdata



4. Legal Issues

Whether micro-data can be shared lawfully in the UK depends in the first instance on the status of the data to be shared. It is useful to place data into one of three categories:

- *Identified* - allowing the direct identification of individual people, households, businesses, or other unit records.
- *Identifiable* - anonymised but detailed micro-data or aggregates that may allow for the indirect identification of individual unit records.
- *Non-disclosive* - data that is not likely to allow for the identification of an individual unit record, without using disproportionate time, effort and expertise.

For example, in UK law data must be considered 'personal data', and therefore subject to the Data Protection Act, if it is identified or identifiable data that relates to living individuals. The processing of such data to make it non-disclosive is caught by the Act, but once non-disclosive the further use of the data is outside the remit of the Act.

4.1. 'Vires' – statutory and implied powers

The data owner must have the administrative power (*vires*) to share micro-data with others. Sharing data beyond the owner's administrative power is *ultra vires* and therefore unlawful.

ONS has powers to share micro-data relating to the number and condition of the population from the Census Act (1920) which states :

> "S5 - It shall be the duty of the Registrar-General from time to time to collect and publish any available statistical information with respect to the number and condition of the population in the interval between one census and another, and otherwise to further the supply and provide for the better co-ordination of such information, and the Registrar-General may make arrangements with any Government Department or local authority for the purpose of acquiring any materials or information necessary for the purpose aforesaid."

Other powers to share ONS data can be found in the Statistics of Trade Act (1947), the Population and Statistics Act (1960), and elsewhere. There is no single, consolidating statistics act for the UK.

4.2. Statutory prohibitions and limitations on disclosure

The statutory framework under which data were originally collected may limit or prohibit its further use for statistical purposes. The prohibition or limitation may be on the users authorised to have access, and/or the permissible uses of the data. Vast data resources in the UK are themselves unavailable to ONS for this (and other) reasons.

Identifiable business survey micro-data collected by ONS are not subject to any statutory limitation of use (statistical or otherwise) where the user is another government department. Local authorities are authorised to have access to identified ONS business survey data for local planning purposes only. Academia is not authorised to have any access to identifiable business micro-data, unless they have a contract of employment with ONS. ONS social surveys are conducted outside of any statutory regime. The further use of identifiable social survey data is therefore free of any statutory limitations or prohibitions on disclosure.

4.3. Duty of Confidence

Within the UK there is no express privacy legislation but there is a right to confidentiality found in our common law. A breach of the common law duty of confidentiality may provide grounds for a civil action for damages. ONS considers all the statistical information provided to it to be subject to a common law duty of confidence.

It is this duty of confidence in common law that determines the availability of social survey micro-data for statistics and research by others. The express obligations in the survey pledge, and any obligations reasonable to imply from ONS' status as a government department and a statistics institution, are assessed by ONS in every case before access to micro-data is authorised.

For ONS social surveys, consent is sought for sharing identifiable data with others for research purposes. Sharing micro-data in a manner consistent with the consent obtained is not a breach of the common law duty of confidence owed to the respondents.

4.4. Data Protection

The fundamental features of the UK Data Protection Act will be familiar to all who are subject to the EU Data Protection Directive. Processing for statistics and research enjoy only limited exemptions from the data protection principles. Processing for statistics in the UK is exempt from the obligation to provide data subjects with access to their personal information. Processing for statistics is not incompatible with the purposes for which the data were obtained. And personal data used for statistics can be retained for as long as the purpose for which the data were obtained requires. Other than these few exemptions, the whole of the rest of the Act applies. This has a limiting effect on the ability of UK departments to share data for statistical purposes. Compliance with the first principle is perhaps the hardest to achieve – it requires that processing is to be for specified purposes only, and that information about these purposes must be provided to data subjects fairly. When the further disclosure to others of personal data for statistics and research is a secondary purpose, it is often the case that these purposes are not specified at the time of collection in a fair manner. ONS' ability to obtain personal data for its statistics is inhibited by this, and it also affects our ability to further disclose information for research by others.

4.5. Human Rights

Under the Human Rights Act, there is a right to a private and family life which may be interfered with only where necessary, and then only in a proportionate manner. If a household has supplied information in a survey, a census, or for the administrative functions of a public authority, the sharing of this information with others for statistical purposes is preferential to additional data collection, because this minimises interference with private and family life. The collect once / use many times philosophy that lies behind good statistical practices for sharing micro-data should be seen as an inherently Human Rights compliant approach.

5. Policy Issues

5.1. National Statistics Code of Practice

The Framework for National Statistics sets out the roles and responsibilities for Ministers. The Framework says:

> '4.1.7. Departmental Ministers, including the Minister responsible for ONS...authorise Heads of Profession for statistics and their staff to make a full professional contribution to National Statistics activities and authorise access to all data within their control for statistical purposes across government subject to confidentiality considerations and statutory requirements'

Access to statistical data across government is thus clear requirement of the UK government.

The UK National Statistics Code of Practice recognises that sharing and combining data is one way of reducing the burden on data suppliers and extending the range of statistics available. The Protocols to the Code make it clear that sharing and combining extracts of data under suitable governance arrangements might be less intrusive to privacy than the alternative of additional large statistical surveys. In the right circumstances, these principles lead to both public services and the privacy of individuals being improved by sharing data.

5.2. Protocol for Data Access and Confidentiality

The Protocol underpins the Code of Practice. It sets out the standards for governance when sharing statistical micro-data, which include:

- The organisation/Data Beneficiary who is being granted access and the Responsible Statistician/Data Beneficiary's representative within that organisation responsible for the data whilst being accessed by that organisation.
- The name of the main contact overseeing the statistical research - the Data Manager
- Details of the data being provided and whether this is a one off or on-going arrangement
- The statistical purpose for which the data are being accessed
- The outputs that will arise from access to the data and details of how the confidentiality of the data will be protected, that is, details of disclosure control techniques to be applied, such as suppression or rounding
- The period of access and arrangements in place for the return or destruction of the data
- Physical and technical measures in place to protect the confidentiality of the data whilst being transmitted to and being used by the beneficiary

5.3. Departmental Policy

In the past it has been perceived that the public have reservations about the data they provide being passed from one organisation to another. This is not always a correct perception. For example, research carried out by the Department for Constitutional Affairs (DCA) has shown that the public expect data to be shared provided those granted access to the data use it for a purpose consistent with its original collection. [1]

It is important to share statistical data with trusted users in the commercial and academic sectors. Much policy development in UK government is founded upon research done outside central government departments, and such research feeds back vital quality information to the producers of National Statistics.

ONS has developed mechanism for authorising access to its micro-data that aims to enable access data to those who need it within a risk management regime. This Micro-data Release Procedure was established in January 2003 and is our response to the obligations of the Code of Practice, the complexity of UK law for data and statistics, the high and increasing demand for research data, and the need for trust in National Statistics. The Micro-data Release Panel (MRP) is the means by which

[1] Full DCA report http://www.dca.gov.uk/majrep/rights/mori-survey.pdf

ONS policy – that confidentiality can be maintained by control over use and user, by control over the design of data, or a combination of both – can be carried out.

Providing access to identifying micro-data is not a way of avoiding the guarantees of confidentiality found in UK law and the Code of Practice. The issue is simply deferred. The organisation of official statistics in the UK requires that ONS often has to trust others to apply adequate statistical disclosure control methods to protect the shared data. This makes the effective dissemination of the standards and guidance for those methods critical to the continued success of ONS' data sharing arrangements.

6. The Data

Sections 4 and 5 outline why an NSI needs to maintain the confidentiality of respondents. In order to ensure this whilst providing users access to microdata the NSI must undertake a risk assessment of the file. Before this can take place a detailed understanding of the data must be established. This includes knowing the source, quality and coverage of the data. Details of how the microdata file was created are key to understanding the risks involved. Information on the sample design, estimation procedure and variables on the file should be summarised. As introduced in Section 2 the framework for protecting and providing access to microdata is based on a disclosure risk-utility decision problem approach. This stage of the framework involves assessing the utility of the data through the identification of the main users and uses of the data.

7. Risk Assessment

7.1. Introduction

Disclosure risk occurs when there is a possibility that an individual can be re-identified by an intruder, and on the basis of that, confidential information is obtained. Identification is made possible by access to uniqueness and hence for microdata part of the disclosure risk comes from individuals that are unique for a certain combination of identifying variables, and part from any access to data an intruder can achieve. ONS protects confidentiality in micro-data by a combination of limits and controls to access, and the perturbation of the data that is accessed.

This section outlines the different approaches to assessing the risk of microdata that are currently implemented or in development at the ONS.

7.2. Disclosure Risk Scenarios

Disclosure risk for microdata comes from individuals that are unique in the sample and population for a certain combination of identifying or key variables. Disclosure risk scenarios are used to define the variables that should be included in the key. Scenarios are assumptions about what an intruder might know about respondents and what information will be available to him to match against the microdata and potentially make an identification and disclosure.

The ONS currently has identified six likely disclosure risk scenarios that should be considered when releasing microdata. The scenarios cover topics such as possible political attacks, private database cross match, journalist, local search and nosy neighbours.

Awareness of the type of identifying data that is available to potential data intruders is important in developing and maintaining disclosure risk scenarios and greatly facilitates the task of maintaining the confidentiality of released data. Work by the Confidentiality and Privacy group (CAPRI www.capri.

man.ac.uk) at the University of Manchester, investigating, classifying and documenting individual data in the public domain and in restricted access databases, has highlighted the value of monitoring personal and business data that could be available to a potential intruder as part of the disclosure risk assessment process (see Elliot (1998), Elliot and Purdam (2002), Purdam, Mackey and Elliot (2003)).

7.3. SDC Checklist for Microdata Release

Disclosure risk scenarios are used to define the identifying variables within a microdata file. In order to provide an objective basis for the risk assessment of microdata the ONS has developed a checklist of criteria that can be used when considering applications coming before the MRP (as introduced in Section 5). The checklist includes identifying variables, visible and traceable variables and information concerned with the survey design. The information provided on the checklist is used in the MRP process to make judgements about the risk posed by different microdata sets the aim is to provide a structured procedure that is flexible, objective, and, as far as is possible reasonably safe.

7.4. Quantitative Risk Assessment

As described above the current risk assessment procedure for microdata files being released by the MRP at the ONS is based on a checklist criteria, subjective judgement and past experience. There is a need to incorporate quantitative measures for the risk of re-identification in the microdata in order to gain more objective criteria for their release. A project has been initiated by the ONS for implementing new research on the assessment of disclosure risk in microdata based on probabilistic modelling and heuristics.

The individual risk measures enable the evaluation of risky records and identify those that need protection. The individual measures can be aggregated to provide global risk measures or file level measures of risk which provide an overall evaluation of the risk of the microdata set and will be important for ranking microdata files by the risk of re-identification. Depending on the type of global disclosure risk measure used and the level of protection needed for the microdata, thresholds are set below which the microdata can be released and above which more disclosure control masking techniques are necessary.

The quantitative measures of risk are based on the probability of re-identification. For microdata files based on censuses or registers this disclosure risk is known. However, for microdata based on surveys the population base is unknown or only partially known. The majority of microdata files being released by the ONS are based on survey samples. In order to quantify the risk of such microdata files one needs to estimate or model the population given the sample.

ONS are carrying out research into the use of probabilistic models for estimating disclosure risk measures for microdata files based on survey samples. Research has focused on two methodologies: the ARGUS Model for risk assessment developed by Benedetti, Capobianchi, and Franconi (1998), Polettini and Seri (2003) and Polittini and Stander (2004), and the Poisson Model developed by Skinner and Holmes (1998) and Elamir and Skinner (2004).

A heuristic method for evaluating the risk of a microdata file has been developed by Elliot et al (2004). The method consists of two elements. The first, called the Data Intrusion Simulation (DIS) is a method for file level risk assessment for microdata which produces estimates of correct matching probabilities averaged over the whole of a microdata file. The second element called the Special Uniques Detection Algoritm (SUDA) grades and orders records within a microdata file according to the level of risk. This method was implemented by the ONS in the risk assessment of the 2001 Census Sample of Anonymised Records (SAR) microdata files, Gross et al (2004). The DIS-SUDA individual level risk measure was used to identify the high-risk records and enabled data masking techniques to be targeted since it provides an estimate of variable and variable value contribution to risk.

In order to assess the robustness of these different methods and to investigate the practical implementation of the methods research has made use of simulated samples drawn from the UK 2001 Census data where the population base is known as well as real microdata files previously released by the MRP, Skinner and Shlomo (2005).

8. Risk Management

Section 7 of the paper outlined the general principle of understanding a microdata file and assessing its utility. The previous Section provided details on different approaches to assessing the risks associated with a file. This Section describes the approaches that can be used to manage risk. The risk within the data is not entirely eliminated but is reduced to an acceptable level, this can be achieved either through the design of the microdata or through the controlled use of microdata, or through a combination of both. The choice of approach should take account of users needs.

8.1 Statistical Disclosure Control

A wide range of statistical disclosure control techniques are available for microdata. For the majority of microdata files released by the ONS recoding is used as a protection method. Perturbative disclosure control methods have also been implemented by the ONS for the SARs microdata files. At the stage where any further recodes would have severely compromised the utility of the data a perturbative method developed by the ONS and based on the Post Randomisation Method (PRAM) was applied, Bycroft and Merrett (2005). This method modifies some characteristics (e.g. age, class, marital status) of individuals in the microdata file and these changes are made according to a controlled random process. PRAM preserves the univariate distributions within the microdata files and some multivariate distributions since the method is applied within defined strata.

8.2. Access Options

ONS policy has allowed for a spectrum of data access arrangements to be provided. The factors of an approval for access are complex – including the purpose of the access, the status of the user, the legal framework, the status of the data, the availability of facilities, and the history of access.

All ONS social surveys generate a micro-data product suitable for widespread use with only limited controls over use and user. Statistical disclosure control techniques are applied to the extent that these controls do not need to be relied upon to protect the data. These datasets are placed with the UK Data Archive (UKDA) and can be downloaded by the user from there under a basic user license administered by the UKDA. The license requires published outputs to meet the Confidentiality Guarantee.

Some ONS social surveys generate a more detailed micro-data product which is suitable for academic research use under significant controls over use and user. Some statistical disclosure control measures are applied, but this remains potentially identifiable data. These datasets are placed with the UK Data Archive (UKDA) and can be downloaded by the user from there under a 'Special License' obtained from ONS. The license requires published outputs to meet the Confidentiality Guarantee.

ONS business and social surveys generate detailed micro-data products for other central and local government departments and authorities. These datasets are transferred to those users in identifiable or identified form, and the use is controlled through a Data Access Agreement. The agreement specifies that compliance with the Confidentiality Guarantee is required for any publications.

ONS business and social surveys are also available in identifiable or identified form through a data laboratory. Access is determined only on the basis of the user being able to demonstrate a need for access to data of this detail – the only checks and prohibitions ONS imposes are on the disclosure control standards for outputs, to meet the Confidentiality Guarantee.

ONS will be preparing a user interface for our website, whereby users can explore data source and access options, and register their requests for data for statistical and research work. We hope this will widen the use of our valuable data sources, and will provide adequate information at the earliest possible stage about the commitments to maintaining confidentiality required when a beneficiary of access to ONS micro-data.

9. Conclusion

There is a strong, widespread and increasing demand for NSIs to release microdata files. This paper has provided an overview of how the ONS is tackling this demand while balancing the need to provide users with access to the data and the need to protect the confidentiality of the respondents. A framework has been developed for protecting and providing access to microdata at the ONS. The framework involves establishing the need to protect confidentiality, understanding the microdata file, assessing and managing the disclosure risk through the use of statistical disclosure control methods and/or restricting access. Underlying the framework is the need to adopt a disclosure risk-data utility decision problem approach.

References

Benedetti, R., Capobianchi, A., and Franconi, L. (1998) *Individual Risk of Disclosure Using Sampling Design Information*.

Bycroft, C. and Lowthian, P. (2005) *Producing Standards and Guidance for Tabular Outputs from ONS*, United Nations Economic Commission for Europe Work Session on Statistical Data Confidentiality.

Bycroft, C. and Merrett, K. (2005) *Experience of using a Post Randomisation Method at the Office for National Statistics*, United Nations Economic Commission for Europe Work Session on Statistical Data Confidentiality.

Duncan, G., Keller-McNulty, S., and Stokes, S. (2001) *Disclosure Risk vs. Data Utility: the R-U Confidentiality Map, Technical Report LA-UR-01-6428*, Statistical Sciences Group,Los Alamos, N.M.:Los Alamos National Laboratory

Elamir, E. A. H. and Skinner, C. (2004), *Analysis of Re-identification Risk based on Log-Linear Mode*l, in (J. Domingo-Ferrer and V. Torra, eds.), Privacy in Statistical Databases, Springer-Verlag, New York, pp. 273-281.

Elliot, M. J. (1998), *DIS: Data intrusion simulation - a method of estimating the worst case disclosure risk for a microdata file*. Proceedings of international symposium on linked employee-employer records, Washington; May 1998.

Elliot, M.J and Manning, A (2004) The methodology used for the 2001 SARs Special Uniques Analysis, University of Manchester.

Elliot, M. and Purdam, K. (2002) *An evaluation of the availability of public data sources which could be used for identification purposes – A Europe wide perspectiv*e, CASC Report. [See http://neon.vb.cbs.nl/casc/].

Gross, B, Guiblin, P and Merrett, K (2004) *Risk Assessment of the Individual Sample of Anonymised Records (SAR) from the 2001 Census*, Office for National Statistics.

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, http://neon.vb.cbs.nl/casc/

Polletini, S. and Stander, J. (2004), *A Bayesian Hierarchical Model Approach to Risk Estimation in Statistical Disclosure Limitation*, in (J. Domingo-Ferrer and V. Torra, eds.), Privacy in Statistical Databases, Springer-Verlag, New York, pp. 247-261

Purdam, K., Mackey, E. and Elliot, M. (2004) *The Regulation of the Personal: Individual Data Use and Identity in the UK*, Policy Studies, Oxfordshire

Skinner, C. and Shlomo, N. (2005) *Assessing disclosure risk in microdata using record-level measures*, United Nations Economic Commission for Europe Work Session on Statistical Data Confidentiality.

Legal, political and methodological issues in confidentiality in the ESS

Maria João Santos[*] *and Jean Marc Museux*[**]

[*] **Methodology and Research Unit, Eurostat, L-2920 Luxembourg**

[**] **Living conditions and social protection statistics, Eurostat, L-2920 Luxembourg**

Abstract. All member countries in Europe face similar problems with respect to Statistical disclosure control (SDC). They all need to find a balance between preservation of privacy for the respondents and the very legitimate requests of society, researchers and policy makers to provide more and more detailed information. This growing demand, due to developments of the information age and knowledge society is a common problem of the European Statistical System (ESS). SDC is also a critical issue for Eurostat because it is at the core of the delicate trust data providers have towards statistics compilers. It influences greatly the quality of EU statistics and consequently the relationship between Eurostat and ESS. In addition, the regulatory framework on statistics includes strict rules to ensure that the information provided by respondents is adequately protected from disclosure. In order to meet the European challenge the SDC problems connected to it have to be approached by all countries in the coming years. In the paper current Eurostat confidentiality issues and strategy are discussed.

1. Introduction

The objective of this paper is to provide an overview of the various issues related to confidentiality in a European wide perspective. It aims to give technical experts an idea of the difficulties raised by the multinational and administrative perspective which might not be perceived at first sight. The variety of perception, the lack of well defined standard is a source of diversity that renders standard confidentiality problem much more problematic at European level. This paper calls for a closer partnership between administrative and research community and for a strong scientific research input and responsibility in order to design best practices to feed legal reflection at European level.

2. Confidentiality legal framework

2.1. General framework

The right to privacy is a fundamental right. It includes the protection of the person in the context of personal data processing. That means for instance the right to receive certain information, the right to access the data, the right to have the data corrected, etc. Statistical confidentiality primarily aims at safeguarding privacy in the field of statistics and is a key to the necessary trust that has to be maintained between statistical bodies and respondents. Mutual confidence ensures accurate and reliable basic information and eventually high quality statistics.

At EU level, statistical confidentiality is addressed in the following legal acts:

- Council Regulation (EEC, Euratom) No 1588/90 of 11 June 1990 on the transmission of data subject to statistical confidentiality to the Statistical Office of the European Communities;
- Council Regulation (EC) No 322/97 of 17 February 1997 on Community statistics;
- Commission Decision 97/281/EC of 21 April 1997 on the role of Eurostat as regards the production of Community statistics;
- Commission Regulation (EC) No 831/2002 of 17 May 2002 implementing Council Regulation (EC) No 322/97 on Community Statistics, concerning access to confidential data for scientific purposes;
- Commission Decision 2004/452/EC of 29 April 2004 laying down a list of bodies whose researchers may access confidential data for scientific purposes.

Statistical confidentiality is regulated at EU level only to the extent to which statistical activities carried out by Eurostat and the national statistical authorities for the production of Community statistics are concerned. Specific confidentiality regimes still coexist at national level and differences may appear with the EU statistical confidentiality regime. These differences are less on the substance (the general concepts are common to a very large extent) than on the perception of the issue (the national framework remains the frame of reference), which is equally important.

Thus, the existent statistical confidentiality regime is not unified in one regulation, which leads to difficulties of interpretation between MS and the Commission and renders difficult current work in different sectors. Improving the existing framework should contribute to avoiding repeated discussions and even in some cases obstacles when dealing with confidentiality issues in the context of the negotiation of sectoral regulations.

At the moment there is an ongoing reflection at Eurostat and at MS level on the need to revise the legal framework based in the principles of maximising the quality of European Statistics both produced by Member States and European Institutions and increase the possibility of secondary use of the data by the research community and the general public; while at the same time respecting the confidentiality mandate to preserve the direct or indirect disclosure of individual information.

The proposed revision could pass by proposing amendments of the legal framework in several domains. In what concerns the transmission of confidential data, could be envisaged the modification of the provision given by art. 14 of Reg 322/97 on the transmission of confidential data without direct identifiers, towards a regime where the transmission and exchange should cover confidential data as defined objectively by Article 13 of Regulation 322/97, covering thus the full range of confidential data. This transmission and exchange should be allowed: between MS and between MS and Eurostat and whenever it concerns and to the extent it is necessary for the production and the quality of Community statistics.

The concept that publicly available information should not be considered confidential already covered by Art. 13 of Regulation 322/97 should be more systematically implemented, possibly through a specific legal act defining variables and fields that are publicly available according to accounting EU directives. In parallel Article 13 §, could be amended in order to ease its implementation by withdrawing the specification: “and remain available to the public at the national authorities”, which is seen as additional constraint for its implementation.

The wide acceptation of an objective basis for declaring data confidential and measuring disclosure risk would definitively ease legal progress in the field of statistical confidentiality. Scientific researcher’s authority in certainly required to put a cut off to the endless subjective discussion. Lawyers are waiting for a strong technical input in order to design harmonised legislation.

2.2. Access to researchers

There is a growing appreciation of the benefits of providing access to microdata for research and analysis. At the same time it is vital to protect data confidentiality. It is essential that new approaches are developed to meet these objectives which create conflicting pressures. The risks to confidentiality must be managed effectively. A key challenge is how to minimise the risks to confidentiality, including the perception of threats to confidentiality. Striking the right balance is vital.

Complex policy making requires multivariate causal thinking about policy alternatives, which in turn, require complex, multivariate, often longitudinal data. As the economy grows more complex and the population becomes more diverse, increasingly detailed data and data analysis are required for policies to match well with economic and demographic alternatives.

An effective public-private partnership between data collection institutes and the research community is a critical element in bringing analyses of complex data, particularly microdata, to bear on policy design and assessment. This partnership between NSI and research is of mutual benefit and is strengthened by continuous improvements in data access, both through public use data sets and through restricted data access modalities. The relationship between data use and data quality is the essential foundation for the common interest of the statistical system and the broader research community in broad and responsible access to data.

There is a need to explore new avenues of access of data to researchers and in parallel improving the current instruments.

Streamlining the implementation of Commission Regulation 831/2002
A detailed description and analysis of this legal act in the paper presented by John King and Jean Louis Mercy in the Work Session on Statistical Data Confidentiality held in Luxembourg on 7-9 April 2003. While this regulation sets important hopes for the availability of microdata to the research community, its implementation has faced several difficulties which have made its development progress at a slow pace.

The committee statistical confidentiality (CSC) of December 2004 has analysed the progress in the implementation of this Regulation and has agreed on the development of quick procedures to process the requests of researchers and to grant the eligibility of research institutions. These fast track procedures will the presented to the CSC on the next meeting in December 2005; their adoption will improve the timeliness and efficacy of the regulation.

There are two levels of access to microdata:
Level one: **Confidential data as obtained from the national authorities.** They allow only indirect identification of the statistical units concerned. This access is done through the use of a safe centre at Eurostat.

Level two: **Sets of anonymised microdata extracted from the above data**. They are individual statistical records which have been modified in order to minimise, in accordance with current best practice, the risk of identification of the statistical units to which they relate.
This access is done via distribution of encrypted CD-ROM according to contracts established between Eurostat and the corresponding institutions.

At present microdata for researchers for level two can only be provided for three statistical domains. These are the European Community Household Panel (ECHP, CVTS (Continuing Vocational Training Survey) and the Labour Force Survey (LFS). In addition, the Community Innovation Survey Working Group is now discussing criteria to distribute microdata files of this investigation. Furthermore, a task force has been set up to do the same exercise for the coming Survey on Income and Living Conditions (EU-SILC).

The necessary measures are going to be taken to propose adding other microdata sets to the ones mentioned in Commission Regulation 831/2002 such as SES (Structure of Earnings Survey).

The advantage of possibilities offered by this regulation is that researchers now have the possibility to have access to harmonized datasets spanning all Member States (MS), before gaining access to data for each of the MS has involved a lengthy process of making requests to each MS. This gives researchers opportunities for pan-European Union research and analyses. The table below presents a synthesis of the projects reported by those research institutions which, during 2004, submitted to Eurostat requests of micro-data of the European Household panel (ECHP).

Research contracts using ECHP data. Year 2004. Main Topics	
Studies of specific sub-populations	*Studies of specific phenomena*
Elderly	Mobility
Poor	Income inequality
Regions	Transition employment <-> unemployment
Long-term unemployed	Taxation, subsidies
Married women	Intra-family transfers
Female participation in labour	Inequality in income and education
Divorced	Wage changes
Temporary Workers	Education and Health
Persons at end of working life	Labour market participation and fertility
Youth	Childcare
	Discrimination

Regulation 831/2002 foresees (article 3) a fairly straightforward and simple request process for researchers from two categories of organisations:

1(a), i.e. universities and other higher education organisations established by Community law or by the law of a Member State; or

1(b), i.e. organisations or institutions for scientific research established under Community law or under the law of a Member State.

For "other bodies", article 3 of regulation 831/2002 lays down the condition that they must first be approved by the CSC if they wish to make requests to access confidential data for scientific purposes. Commission Decision 2004/452/CE list other bodies that have been considered admissible. The prerequisite to achieve admissibility is that the institution has demonstrated that it fulfils a set of criteria. The CSC has approved these criteria at its meeting of 10 December 2004. Specific services of EU Institutions, which carry out statistical activities, may be considered eligible as researchers for access to specific confidential micro-files provided that the equivalent guarantees are provided. This follows the precedent established with the ECB and the Central Banks of Spain and Italy. Universities based outside Europe can also be considered as admissible; the University of Cornell (USA) was the first to be included in this list. The efforts will continue to extend the list of other bodies than can be regarded as admissible.

Establishment of bilateral agreements on licensing and delocalisation of safe centres

An important component of developing a new confidentiality protection system is the development of a safe centre network. At the moment the safe centre for the data sets mentioned under Commission Regulation 831/2002 is localized at Eurostat. Eurostat will discuss with the MS the possibilities to delocalise via the establishment of bilateral agreements the safe centres to MS or to create the conditions to establish licensing agreement with established institutions.

3. Methodological issues

In general the legislation at national and European levels is fairly harmonised with respect to what is considered as confidential data. However, when implementing this legislation, the criteria used differ considerably from country to country. These criteria have sometimes an important historical weight; sometimes do not have a solid scientific basis; and in many cases lead to conservative solutions because real risks are not well mastered.

This diversity of interpretations is a consequence of the fact that there is no harmonised approach of disclosure risk. To agree on disclosure risk, one should agree first on the sensitivity of the data (how "private" are the variables in the file) and on the possibility to match these data with external sources, that is, to the presence of key variables or identifying variables. Second, there is a need to find a harmonised way to measure the risk. Methodological work is needed to reconcile the different approaches or to express preference for one of them.

It is obvious here the need to have common core criteria which, while providing a satisfactory harmonisation level, allow for a degree of flexibility to adapt to the specific perception of the society in each country. This will also have the advantage of having a more solid internationally agreed basis that better justifies national choices made in the release of microdata.

Disclosure protection of EU aggregates
Most of the time, Eurostat compiles EU aggregates on the basis of national aggregates. These are accompanied with a confidentiality flag informing Eurostat that the information should be treated as confidential. In the best situation, Eurostat is also informed on the presence of dominance in these aggregates. However, meta information is not standardised and even sometimes there exists confusion between not publishable because of lack of reliability and confidential as meant in the legal framework.

To declare information as (primary) confidential, MS use measures of risk of disclosure of individual information (dominance rules, threshold rules) which are not harmonised. The level of protection can varies between MS depending on different perceptions of the level of disclosure risk and also simply of the perception of the damage of disclosure itself. Distinction is rarely made between variables themselves: some variables might be considered as non sensitive whereas other from the same record could be.

The lack of harmonisation of primary confidential rules causes major methodological problems at Eurostat level. Software packages for handling secondary confidentiality are not designed to deal with such a situation. For instance, the input required, mainly micro data, does not fit Eurostat situation which deals with aggregated data. Consequently, Eurostat, following the most stringent rules used by national authorities to protect EU data, is led to over protect data and not to release useful information for the user. The lack of harmonisation of disclosure protection measures between MS hampered thus the release of European data. This situation could be improved by rising the awareness of disclosure control issues and if statistical disclosure experts would issue a unified set of best practices accompanied with practical hints to implement them. This would be developed in a European perspective.

Disclosure protection of micro data
To some extend the same holds when Eurostat has to design, in collaboration with MS, anonymisation of micro data to be released to researchers. Despite they share common objectives:

- the need to follow Regulation principles on the right for privacy,
- the need to maintain the trust the respondent have in the statistical system,
- the need to monitor the release so to avoid confidentiality breach),

the differences in the perception of the risk and the lack of a universal measure of risk render the possibility of a consensus very thin. Part of the problems lies in the absence of knowledge of real risk.

This situation would be improved if once again, European experts would agree on a set of measure and threshold to be used by practitioners. This probably needs more comparative research to be developed on the existing measures and on the tuning of methods. In parallel, more research could be carried out on the measure of the actual measure of risk. Computer scientist could design protocol to crack released European database, which in turn could be used to develop appropriate protection measures.

4. Conclusions and future perspectives

With respect to medium term perspectives, three main components are already identified:

SDC in the ESS
The Setting up of a Centres and Networks of Excellence (CENEX) in Statistical Disclosure Control. CENEX originates from the idea of sharing the work between different institutions within the ESS (European Statistical System) more efficiently, by providing adequate organisational solutions and institutional framework for modern types of cooperation and specialization of work.

In the latter case, the sharing of work between different Member States in a CENEX will creates synergies since each participating NSI will concentrate on specific areas and the product of this work will be beneficial to all NSI ultimately leading to the increase of the quality of ESS statistics. It is moreover essential for the generation of comparable statistical information across countries, and on the European level that similar methods and tools are used to protect confidentiality in the published information. As long as member states compile their statistics using different statistical disclosure control (SDC) methods, the compilation of European statistics is very much hampered.

The pilot CENEX on SDC was defined to address in a first phase the following objectives:

- Set standards for the protection of micro-data sets, based on disclosure risk assessment methods and criteria.
- Improve tabular data protection techniques and develop harmonized criteria.
- Extend and develop SDC software tools, both for micro and tabular data, so as to fit the specific production and dissemination environments of ESS.

Eurostat plans to evaluate and further develop the CENEX approach to harmonise SDC practices in the ESS, promote the definition and use of best practices, bring up to level SDC software tools in the ESS and the remote access to microdata.

Public use files
Public use files (PUF) are the most accessible; widely and freely used microdata products made available by statistical institutes, but their value for policy for much policy relevant research is limited. Nevertheless these files are useful for some research purposes, as teaching aids and are a good advertisement of a statistical institute. Continued distribution of public use files is threatened by the increased re-identification risk associated with both technological advances in linking software and widespread availability of administrative records. During the last decade researchers have developed increasingly sophisticated methodologies for restricted data products. The development of a methodology for generating synthetic or virtual data is a relatively recent activity. A key objective of the method is to preserve faithful representations of the original data so that inferences from the synthetic data are as consistent as possible with the inferences that would be drawn from the original data. One attractive feature of the synthetic data approach is that it can be used to create multiple public use files from the same underlying data – targeted at different audiences. The methodology of synthetic files as a measure to replace public use files need to be further researched.

The work at the sectoral level to establish the criteria for establishing public use files (such as the on going work of the EU-SILC TF on anonymisation and establishment of public use files) will continue to be promoted in the future via the establishment of sectoral TFs that will define PUF for each survey.

Monitored remote access to microdata
A sensible approach for facilitating high quality research is to maintain the data in a secure, restricted remote access environment.

Monitored remote access has the advantage that a researcher does not have to go to a safe centre to make use of confidential data and output is returned relatively quickly. This approach to develop remote access procedures, which has the advantage of reducing researcher burden, involves substantial investment in hardware and software. This approach has been gathering momentum and is now operational in Europe in the Netherlands, Denmark and Sweden. It will be studied with MS the possibility offered by the 7th Research Framework Program in the field of research infrastructures to further develop such an approach at European level.

Some of the requirements and targets specified in laws are not fixed but are moving over time. There is thus a requirement on NSIs and on Eurostat to review practices and methods from time to time. It has been presented some of the more long term threads to be followed in the future regarding the modification of the current legal framework. In parallel were described concrete axes of implementation reflecting the orientations of Eurostat in short to medium term with respect to confidentiality. Eurostat hopes to develop fruitful synergies with experts and NSI along these axes.

References

Jean-Louis Mercy and John King "Developments at Eurostat for research access to confidential data" Joint ECE/Eurostat work session on statistical data confidentiality (Luxembourg, 7-9 April 2003) Working Paper 12.

Developing the core principles and guidelines on managing confidentiality and access to microdata

UNECE secretariat

The paper will give a short overview of the work of the CES Task Force on Confidentiality and Microdata and of the survey on international access to microdata that the UNECE carried out at the request of the Task Force in autumn 2005.

CES Task Force on Confidentiality and Microdata

The Plenary Session of the Conference of European Statisticians[1] (CES) in June 2003 agreed that it is needed to agree on core principles for releasing microdata and to harmonise the confidentiality approaches internationally. The Conference decided to set up a Task Force to develop international guidelines and core principles for managing confidentiality and access to microdata. The group is chaired by Mr. Dennis Trewin, the Chief Statistician of the Australian Bureau of Statistics. Its members are: Canada, Denmark, Georgia, Italy, Poland and UNECE The documentation of the Task Force is available at the following website:

http://www.unece.org/stats/documents/tfcm.htm

The core principles of confidentiality are based on the sixth fundamental principle of official statistics: "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"[2]. The core principles develop this rule further by defining under which conditions statistical offices can provide access to microdata, so that the confidentiality of respondents' data is guaranteed.

The Prinicples are accompanied by Guidelines that provide the good practices in the implementation of the principles. The Guidelines present the perspective of the statistical office and the research community and look at the possibilities to solve the tension between these perspectives. The document gives an overview of the methods of supporting the research community (anonymised microdata files, remote access facilities, data laboratories), management issues associated with the release of microdata (decision making, metadata, breaches by researchers) and some special issues (international access, data linking). An important part of the paper are numerous case studies on the implementation of confidentiality in countries.

The Guidelines recognise that the precise arrangements for access to microdata vary from country to country, depending on legislation, public attitudes and the capacity to support the research community. Therefore, the document gives general guidance and it is up to the countries to make their own specific confidentiality arrangements. The Guidelines are also expected to help in discussions with user community and other government agencies, and to help countries that are in the process of setting up their confidentiality laws and procedures.

The Task Force has worked through several versions of the Guidelines which have been extensively commented and discussed by the countries and the CES Bureau[3]. The Task Force has made an attempt to take into account all comments, as much as possible. The Guidelines are planned to be finalised in

[1] The Conference of European Statisticians consists of the Heads of the statistical offices of the UNECE member countries and the Heads of statistical departments or divisions of all major international organizations. In addition, it includes all the remaining OECD countries and some other countries that have been interested in its work.

[2] The Fundamental Principles of official statistics were adopted by the UN Statistical Commission in 1994, see http://unstats.un.org/unsd/goodprac/bpabout.asp

[3] Any questions or comments on the Guidelines can be addressed to Tiina Luige (tiina.luige@unece.org), the focal point of the Task Force in UNECE.

the beginning of 2006, so that they can be submitted for adoption to the Plenary Session of the Conference of European Statisticians in June 2006. After adoption, the Core Principles and Guidelines for Managing Confidentiality and Microdata Access will be published by the UNECE.

International access to microdata

While working on the Guidelines, the Task Force considered that more information is needed on the practices concerning international access to microdata. Therefore, the UNECE secretariat carried out a small survey among the CES member countries in autumn 2005. A short questionnaire was sent out to 61 countries. Forty three countries[4] responded to the survey (70% response rate).

The survey dealt with two aspects of the international access: access to microdata by international organisations, and access by researchers from other countries. The key findings from the survey are summarised below.

The results showed that the vast majority of countries release microdata to international organizations. About 60% do not place limitations on the type of organizations to which they can release data although some countries indicated that they can provide access to Eurostat only.

In the survey, three main methods of releasing microdata were suggested and countries were asked to indicate the forms in which they *could release* microdata. Of those countries that release microdata, 55% can release anonymised microdata files as public use files and 82% can release as licensed microdata files. Only a relatively small number of countries (5 countries among those who replied) have remote access facilities, although others are developing or have plans to set up these facilities.

When asked about the forms in which countries *actually released* microdata to international organizations, most countries (67%) provide data to international agencies as licensed files (i.e. there is a signed undertaking as part of the provision of the microdata) although many (about 40%) provide public use files.

About 50% of countries that provide data to international organizations, specify conditions. These are usually formulated to ensure that use must be for statistical, research or scientific purposes. Also, another common condition is that microdata must not be passed on by the international agency.

Countries were requested to specify whether they were able to release data to the whole organisation, a specific division or a designated individual. Of the countries that release data to international organisations, 52% can release data to a whole organisation, 50% to a division and 52% to a designated individual. Seven countries can release data to all three audiences.

About two-thirds of countries do not have different arrangements for household data and business data, while the remaining countries make a difference. Although not reported in the survey, it may be the case that many countries do not provide any business microdata because of confidentiality concerns.

Concerning allowing access to individual researchers, most countries (85%) reported that they can release microdata to a designated researcher in their own country who can work collaboratively with an international organization. Some reported they cannot release data outside their own country. Some reported they can only release to research institutions.

About three-quarters of countries can release microdata to researchers in other countries. In some cases, these are public use files only and some countries do not release data directly but through agencies. Most countries do not have special arrangements with statistical offices of other countries, even though the other office may have legislation that protects microdata.

[4] Andorra, Armenia, Australia, Austria, Belarus, Bulgaria, Canada, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Israel, Italy, Japan, Kazakhstan, Latvia, Lithuania, Mongolia, New Zealand, Norway, Poland, Portugal, Republic of Korea, Republic of Moldova, Romania, Russian Federation, San Marino, Serbia and Montenegro, Slovakia, Slovenia, Sweden, Switzerland, The fYR of Macedonia, Turkey, Ukraine, United Kingdom, United States.

In conclusion, it may be said that countries are more relaxed about the release of microdata internationally than previously thought. The Task Force was not aware of any significant cases where release of microdata to international organizations would have been abused. On this assumption, the Guidelines recognise the importance of microdata to international agencies and outline good practices on the provision of microdata to such agencies. However, it would still be the decision of individual countries as to whether to provide these data or not.

The Institut de la statistique du Québec's Approach to the Confidentiality of Microdata Files and Tabular Data

Jimmy Baulne, Éric Gagnon and Lyne Des Groseilliers
Institut de la statistique du Québec, Direction de la méthodologie, de la démographie et des enquêtes spéciales, 200 chemin Sainte-Foy, 3rd floor, Québec, Quebec, Canada G1R 5T4

Abstract: Under its act of incorporation, the Institut de la statistique du Québec (ISQ) is required to protect the confidentiality of the information it collects. Accordingly, the Institute has adopted policies enabling it to fulfil its confidentiality obligations. Two of these policies involve the confidentiality of statistical products disseminated by the Institute, i.e. microdata files and tabular data. The first part of this article deals with microdata files more specifically, as well as the statistical disclosure control (SDC) rules applied to them. The stringency of SDC rules can vary according to the environment in which the microdata files will be used. The second part of the article deals with tabular data, and outlines different aspects of the Institute's dissemination of tables. The ISQ's integrated approach is then discussed, highlighting the different components examined: SDC rules applying to social surveys and business surveys.

1. Introduction

The Institut de la statistique du Québec (ISQ) is the official statistical agency of the Quebec government. Its mission is to provide reliable and objective statistical information on all aspects of Quebec society. To this end, it conducts several social surveys and business surveys every year. In keeping with its mission, the Institute must utilize the entire statistical potential of the information gathered through its different activities. The Institute lacks the internal resources to do this, however, and so it made a strategic decision to maximize the use of its statistical products by third parties. It must ensure that these data are used in accordance with its act of incorporation, which requires that the Institute preserve the confidentiality of the information it collects. It therefore adopted an approach for sharing its microdata files and tabular data, aimed at ensuring flexibility in making these products accessible while protecting their confidentiality.

Section 2 of this article describes the approach adopted for disseminating microdata files resulting from social surveys. Then section 3 describes the approach used for disseminating tabular data from social surveys and business surveys.

2. Microdata files

2.1. Dissemination of microdata files

To ensure maximum productivity of the information gathered through its surveys, the Institute offers outside researchers access to different types of microdata files with variable analytical potential, while protecting the confidentiality of the data provided by respondents. In this section we will look at different types of microdata files produced from social surveys, as well as disclosure control measures applied to these files. Access to files produced from business surveys will not be addressed in this article. The nature of the information in such files calls for more stringent disclosure control measures than those set out here.

There are two methods of rendering microdata files available to researchers. The first is to request that respondents give their consent ahead of time for these data to be shared with researchers. The second method gives researchers access to microdata files when there is no prior consent. In this article we will consider only files shared using this latter method.

2.2. Access to microdata files without prior consent

The Institute can give researchers access to microdata files without prior consent for various reasons. Consent may not have been sought, for instance, because a researcher wants access to all respondents in the file, so that his or her results will be consistent with those of the Institute. Consent may have already been sought for researchers of one agency, but then researchers of another organization also request access to the file during a project. Prior consent clearly does not apply in this case, and another type of access must be suggested.

The Institute suggests different approaches when there is no prior consent, so as to make it possible to share microdata while protecting respondents' privacy. Different disclosure risk control measures are suggested for this purpose:

- statistical measures: statistical disclosure control (SDC);
- legal and administrative requirements;
- physical and computer security measures.

The combined application of these measures makes it possible to control the risk adequately. It is possible to vary the stringency of each measure, while still ensuring adequate risk control. For instance, if a decision is made to apply less stringent statistical measures, then legal and administrative requirements and physical and computer security measures should be more strict.

2.2.1. Public-use microdata files (PUMFs)

The first type of access suggested when no prior consent exists is to give researchers access to a PUMF at their workplace. Before producing this kind of file, the variables in the file must be classified into three categories: direct identifiers, indirect identifiers and non-identifying variables. Direct identifiers are variables that can be used to identify a person directly, e.g. a name, address or telephone number. Indirect identifiers can be used to identify a person when they are cross-referenced, e.g. sex, age and profession. Finally, all other variables in the file are non-identifiers and consequently are excluded from SDC analysis. To create a PUMF, SDC rules are applied to direct and indirect identifiers in the file. As a first step, direct identifiers are removed, and then very strict SDC rules are applied to indirect identifiers. SDC rules are applied in two stages: risk identification followed by masking to minimize the risk. The following criteria are used to identify the risk:

- A region that can be distinguished in a file must have at least 80,000 inhabitants.
- Each cell obtained by combining the categories of three indirect identifiers must comprise at least 800 individuals in the population.
- One of the indirect identifiers in the combination must be the distinguishable region mentioned above.

The minimal criterion of 80,000 inhabitants in a region has been used in the past by Statistics Canada for creating PUMFs (Béland, 1999). Under certain circumstances, this criterion and the minimal number of individuals in a cell may be relaxed or tightened. In using varied thresholds it is important to take into account the sensitivity of information in a file and the type of population covered by the survey. Moreover, the use of lower thresholds must not substantially increase the risk.

For the second stage in applying SDC rules, the following masking techniques may be applied:

- global recoding of regional variables and indirect identifiers at risk;
- removal of an indirect identifier at risk from the file;
- removal of the indirect identifier at risk for certain respondents;

- top-coding and bottom-coding;
- rounding off or adding random noise.

Applying these SDC rules permits minimization of the disclosure risk and, consequently, relaxation of the other risk control measures. For instance, it is not necessary for researchers to apply SDC rules to tabular data produced from this type of file. When using such files, however, researchers must agree:

- to use the file for analysis and research purposes;
- not to combine the file with another file or attempt re-identification;
- not to make back-up copies of the file.

Researchers who do not comply with these requirements may be denied access to the PUMF.

2.2.2. Scientific-use microdata files (SUMFs)

A second type of proposed access is to provide researchers with a SUMF at their workplace. This allows them to work on a file with greater analysis potential than that offered by a PUMF. To obtain access, however, researchers must sign an agreement with the Institute, agreeing to protect the confidentiality of the data provided. A SUMF is created by classifying the file variables into the same categories as in a PUMF; direct identifiers are removed and SDC rules applied to indirect identifiers are less stringent than those used in creating a PUMF. As with PUMFs, SDC rules are applied in two stages: risk identification and masking. Risk identification for SUMFs uses the following criteria:

- A region that can be distinguished in the file must have at least 10,000 inhabitants.
- Each cell produced by combining the categories of three indirect identifiers must comprise at least 100 individuals in the population.
- One of the indirect identifiers in the combination must be the distinguishable region mentioned above.

These risk identification criteria, inspired by the methods developed by Statistics Netherlands (Schulte Nordholt, 2001), are less stringent than those described for PUMFs. Moreover, under certain circumstances these criteria can be relaxed or tightened using the same conditions that apply to PUMFs.

In the second stage of applying SDC rules, masking is used for SUMFs just as it is used for PUMFs. However, the masking is less stringent than that applied to PUMFs, given the lower risk identified at the first stage.

Applying SDC rules to create SUMFs makes reduction of the disclosure risk possible, but does not eliminate it. Consequently, adequate control of this risk requires stricter legal, physical and computer measures than those used for PUMFs:

- Users may not transport the microdata.
- Paper copies must be kept in a secure location.
- Access to the copy of the original microdata file or its subproducts must be controlled and restricted to authorized individuals.
- The file must be kept in a secure location and encrypted.
- Once the project is finished, the copy of the original microdata file must be destroyed and a note confirming its destruction sent to the Institute.
- Researchers must apply SDC rules to the tables produced from the file. Details on these rules are given in section 3.3.1.
- Etc.

Researchers who fail to comply with these requirements may be denied access to the SUMF. The Institute may even take legal action against them.

Finally, the Institute is currently testing remote access as a new way to provide access to SUMFs. In future, the Institute would like to make files with less masking than SUMFs available by remote access.

2.2.3. Microdata files without direct identifiers but non-masked, available at CADRISQ

A third type of suggested access is to provide researchers with a microdata file, without direct identifiers but non-masked, on the premises of the Institute's research data consultation centre (Centre d'accès aux données de recherche de l'Institut de la statistique du Québec – CADRISQ). This approach may be preferable for researchers who are not satisfied with the analysis potential of SUMFs.

In these files, only direct identifiers are removed. No SDC rules are applied to indirect identifiers. Consequently, the disclosure risk for such files is considerable. To make up for the lack of SDC rules, however, legal, physical and computer measures relating to the use of such files are more stringent than those used for SUMFs:

- The microdata file remains on CADRISQ premises.
- Analyses are conducted under the supervision of the CADRISQ supervisor.
- Researchers are sworn to secrecy and subject to the same confidentiality obligations as ISQ employees.
- SDC rules must be applied by researchers to tabular data they wish to remove from the CADRISQ. Section 3.3.1 gives further details concerning these rules. The CADRISQ supervisor ensures that SDC rules have been applied adequately by the researcher, so as to safeguard the confidentiality of the tables.

3. Tabular data

3.1. Dissemination of tabular data

The approach described in the paragraphs below concerns the dissemination of tabular data produced from a microdata file belonging to the Institute. This data may be disseminated by non-ISQ researchers using a file from the Institute, but also by ISQ employees when publishing survey findings.

Unlike the approach used when disseminating microdata, which concerns only social survey data, the approach for disseminating tabular data concerns both social and business survey data.

In the section outlining the approach for the dissemination of microdata files, we made a distinction between different types of files that can be made accessible to users. This distinction has a direct impact on the SDC rules applied to tabular data. That is why the approach concerning the dissemination of tables takes into account the type of file used to produce the table (i.e. whether it is a non-masked file or SUMF). Once again, SDC rules applied to tables also depend on the type of user who disseminates the table (i.e. an ISQ employee or a non-ISQ user).

Keeping all these distinctions in mind, we will now take a closer look at the ISQ approach concerning SDC rules applicable to the dissemination of tables.

3.2. Developing a policy

Whether research is being done for its own publications or for a publication by a researcher using one of its microdata files, the Institute must provide users of its files with a procedure that lays out rules to be followed. The purpose of these rules is to ensure the confidentiality of the information disseminated. Furthermore, if a researcher uses an Institute file, failure to comply with this procedure could result in legal action against the individual and his or her employer.

Since the Institute is obliged to protect the confidentiality of information published, it has developed a policy setting out guidelines governing the confidentiality of tabular data of survey results for dissemination.

This policy covers different types of tables: frequency count tables, tables of magnitude – mean, total or ratio –, percentile and model analysis results (regression). In addition, tables may be produced from social or business survey files.

Thus there are a variety of circumstances that the Institute must take into account when developing its policy on the dissemination of tabular data. For example, who wants to disseminate the table: an ISQ employee or an external researcher? What kind of file (non-masked or SUMF) was used to produce the table? What kind of data (social or business) does the table contain? All these considerations help determine the choice of SDC rules to be applied to tables. The Institute has had to come up with policies, each complemented by a separate procedure, to ensure that it can respect its commitment to confidentiality in all situations involving the dissemination of tabular data.

3.3. Organization of guidelines

Guidelines on the confidentiality of tabular data for dissemination have been split into guidelines for social surveys and business surveys. Each section has a procedure for every different situation involving the dissemination of tabular data.

3.3.1. Social surveys component

The social surveys component has three procedures. The first deals with tables produced by non-ISQ users, using non-masked files. Such files can be made available to researchers either at the CADRISQ, or on the premises of the researcher's public organization, if respondents have given their prior consent. Since no SDC rules have been applied to the indirect identifiers in microdata files of this type (see section 2.2.3), there is a very high disclosure risk and strict SDC rules are applied to the tabular data.

For this procedure, a table represents a disclosure risk if there is not a minimum number of respondents in each of the cells of the table, or if there are zero cells or full cells. A cell is full when it contains all the respondents; a zero cell contains no respondents. The masking techniques applied to tables considered at risk depend on the variables they contain. Indeed, this procedure uses two important concepts: the presence of a variable related to ethnicity and the size of the geographic classification. Distinguishing tables on the basis of ethnicity is justified by the fact that this is a very sensitive concept in Quebec, and sub-populations formed by different cultural communities are relatively small, with consequently higher risk of identification. The same observation applies to the sub-populations defined by certain geographic territories, which makes such tables highly specific.

Accordingly, SDC rules are stricter when there is a variable linked to ethnicity and when the geographic classification is small. Among the masking techniques used in this procedure are:

- table redesign;
- local suppression of data (including secondary cell suppression);
- limiting the number of cross-referenced variables used in a table;
- prohibiting the regional dissemination of tables (in certain cases).

The second procedure concerns tabular data produced from SUMFs used by non-ISQ researchers. The disclosure risk associated with these tables is less than for tables produced from non-masked files, for SDC rules have been applied to the microdata. Thus less stringent SDC rules can be applied to tables. This procedure uses the same concepts as the first (i.e. the presence of a variable linked to ethnicity and the size of the geographic classification). Disclosure risks are also identified in the same way, and only some of the masking techniques are used to reduce this risk. This is more or less what distinguishes the first two procedures, and this distinction comes from the fact that SDC rules are applied to the microdata of a SUMF, but not to the microdata in a non-masked file. For example, a table may be disseminated regionally if it is produced from a SUMF file and complies with the rules of the second procedure, while producing such a table from a non-masked file (first procedure) may not be allowed.

The third procedure in the social surveys component concerns tables produced by ISQ employees. Non-masked microdata files are used to produce such tables, of course. The presence of a variable linked to ethnicity is once again an important concept for determining the SDC rules to be applied to the tables. However, the second concept used in this procedure is the presence or absence of a delicate variable in the table. A variable is considered delicate if it contains information relating to the respondent's private life, which is not generally known and the respondent does not wish to disclose, such as sexual behaviour or the cause of a disability. Tabular data must therefore be classified into one of the following four categories:

- table containing a delicate variable cross-referenced with a variable linked to ethnicity;
- table containing a delicate variable not cross-referenced with a variable linked to ethnicity;
- table containing a non-delicate variable cross-referenced with a variable linked to ethnicity;
- table containing a non-delicate variable not cross-referenced with a variable linked to ethnicity.

The status assigned to variables (i.e. whether they are delicate or not) is up to the survey project leader, to be approved by his or her manager. This strategy makes it possible to relax the SDC rules applied to tabular data in certain cases. The tables in the fourth category are an example. The disclosure risk identification methods are the same as for the other two procedures. Once again, the strictness of the rules depends on the classification of the table. Tables containing delicate variables or variables linked to ethnicity will be subject to stricter SDC rules, whereas the other tables will be subject to less strict measures, in particular allowing low-frequency cells in the tables. The masking techniques used in this procedure are combining categories and local suppression of data deemed confidential, including secondary cell suppression.

3.3.2. Business surveys component

The business surveys component includes just one procedure, involving tabular data produced by ISQ employees. Like their counterparts in the social surveys component, these tables are produced using non-masked files, meaning that no SDC rules have been applied to the indirect identifiers in microdata files of this type.

The procedure for this component uses a concept equivalent to the third procedure in the social surveys component (i.e. delicate variables, but adapted to the business context). Tables in this component are categorized according to whether they contain a strategic variable or a non-strategic variable. Any information likely to give a business a competitive advantage may be considered a strategic variable.

The SDC rules applied to tables containing a strategic variable are stricter than those applied to other tables. In the only procedure in this component, disclosure risk is identified by the absence of a minimum number of respondents in each cell or the presence of zero or full cells and, for table of magnitude data, a sensitivity measure such as the dominance rule (n,k) or the p-percent rule (Willenborg, 2001).

The following masking techniques are used to limit this risk:

- local suppression of data (including secondary cell suppression);
- able redesign;
- adding random noise;
- controlled or random rounding.

Just as for the social surveys component, the choice of strategic and non-strategic variables is up to the survey project leader, subject to approval by his or her manager. For the business surveys component, however, a committee consisting of ISQ employees was struck specially to draw up a list of variables, grouped into themes, that must be considered strategic. Employees wishing to disseminate tabular data are required to use this list to determine the status of variables.

4. Conclusion

As the Quebec government's official statistical agency, the Institute has an obligation to protect the confidentiality of the information it releases. The approaches described in this article allow it to fulfil its obligations while offering researchers access to data with satisfactory analytical potential.

References

Béland, Y. (1999). "Release of Public Use Microdata Files for NPHS? Mission… Partially Accomplished!" *Proceedings of the Survey Research Methods Section.* American Statistical Association, p. 404-409.

Schulte Nordholt, E. (2001). "Statistical Disclosure Control (SDC) in Practice. Some Examples in Official Statistics of Statistics Netherlands." Paper presented at the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality. Skopje, The former Yugoslav Republic of Macedonia.

Willenborg, L. and T. De Waal (2001). *Elements of Statistical Disclosure Control. Lecture Notes in Statistics 155.* New York: Springer-Verlag.

Statistics and confidentiality in the Portuguese case

Ana Dulce Pinto

National Statistical Institut, Statistical Council Secretariat, Lisbon, Portugal, ana.dpinto@ine.pt

Abstract:

Title: “Statistics and Confidentiality in the Portuguese Case”.

Abstract: The paper will detail the cross legal references that should be taken into account when dealing with statistical confidentiality applying to the Portuguese case: The link will be made with the organization of the Portuguese Statistical System. The growing social pressure to meet users’ needs keeping untouched the trust of the respondents is an old challenge but still main question for statisticians which will be pointed out in the presentation. In particular, three questions and three answers that will define a new point of balance between information freedom and confidentiality in the Portuguese case: 1) What exists? 2) What may be improved? 3) What can be done?

1. Cross legal references concerning confidentiality

Statistical confidentiality has always asked for the attention of everyone working in statistical production mainly for two reasons: on one hand, its importance to guarantee the respondents trust (families, enterprises and others), and on the other, the practical difficulty to disseminate solid and relevant statistics at a detailed level, without disappointing the confidence of the respondents.

Restating this was recently approved the Code of Practice of European Statistics. This code refers Statistical Confidentiality as one of the main principles thus conferring it identical importance as others such as statistical impartiality or the need of a sound statistical methodology.

Nowadays, in Portugal, two legal main devices may be applied to confidentiality issues:

Firstly and with higher importance, the National Statistical Law that establishes the main principles as well as the working methods and composition of the Statistical System. It is a 1989 Law which is no longer enough to assure a prompt reply to all situations and doubts concerning the limits and scope of statistical confidentiality. This difficulty, amongst others, has been unanimously appointed by statisticians and statistical users as the reason for the revision of this law, which has recently started.

Being certain that the processes of revision of legal procedures are quite slow, the truth is that concerning confidentiality issues, the new text that will be adopted could always benefit from the discussions and conclusions around the concept and boundaries of statistical confidentiality, undertaken at international level, namely in EUROSTAT and more widened circles as the OCDE.

Secondly, but also very important is a specific diploma that establishes the access rules, collection and processing by INE of personal data from administrative sources used for statistical purposes. This diploma results from a specific need, felt both by INE and the national entity responsible for the control and supervision of the application of the rules concerning data protection, to legalize all the mechanisms adopted whenever personal data are collected for statistical purposes, taking advantage of administrative sources.

Still at the level of the national law, even if not directly applicable to statistics, it is inevitable the intersection of the Personal Data Protection Law (resulting from the transposition of Directive 95 / 46/ CE) with the statistical legislation. In fact, there are many personal data collected for statistical purposes, either through specific surveys or through the exploitation of administrative data. The Portuguese Law that transposes the Directive did not consider statistics as a special purpose of the personal data collection that could justify differentiated treatment of other entities. In fact the Statistical Law in Portugal does not constitute “*lex speciallis*” in relation to the text that transposes the community Directive. This means that INE needs a special authorization from the competent entity to collect or treat personal data, such as any private enterprise who intends to conduct a survey and process data about physical persons.

Finally, at the internal level, INE has specific regulations concerning statistical confidentiality that intend to simplify the management of the principle. Ideally these regulations must be dynamic instruments in permanent update. This is due to the constant technical evolutions not only on the right approach of the confidentiality principle – and to situations that social and economic developments impel to consider as new exceptions to the general rule – but also on the evolution of information technologies permitting to pull off the "brakes" of the principle as concerns statistical dissemination.

It should be considered that the Portuguese Law does not foresee the access to statistical confidential data for scientific research under any circumstances. As a way to surpass the emptiness of the Portuguese Law and due to the principle of direct applicability of community regulations in Portugal, the Regulations 322/97 and 831/2002 are being considered and all the criteria established in this last one are being applied.

Considering that the current structure of the Portuguese Statistical System states that INE can delegate the production of specific statistical areas in other entities of the public administration, these are equally obliged to apply the same principles of INE namely statistical confidentiality.

2. Organization of the Portuguese Statistical System – INE, Statistical Council (SC) and other public bodies

The Portuguese Statistical System, such as stated in the Law of 1989, is composed of two entities and specific public bodies for which INE transfers through a legal act the possibility of producing statistics in specific areas.

INE is the "executive" or operational branch of the system with responsibilities in three main axes: statistical production, from collection to dissemination; promotion of statistics teaching and cooperation with other countries aiming to statisticians training.

SC is the other entity of the system with advisory functions, formally responsible for the orientation and coordination of the Statistical System. It has currently a multiple, heterogeneous and sufficiently widened composition that intends to improve the critical analysis of the functioning of the system. In the future however the SC resources and composition should be reviewed.

SC has a main role in the guarantee of the application of statistical confidentiality, namely through the creation and monitoring of the mechanisms allowing its control. The Council has also the specific ability to analyze and decide when to disclose confidential information produced by INE or other public bodies of the system. Considering the heterogeneous composition of the SC, this analysis as well as the decisions taken on confidential statistical data, is thus the result of the consensus of the sectors represented in the SC, from governmental structures to private users, and has the legal boundaries referred in point 4 below.

The other public bodies that produce statistics due to a legal act from INE must have specific knowledge in the area on which they will produce statistics. A successful example is the area of justice and criminal statistics which are completely transferred to a service of the Ministry of Justice especially qualified in the collection and processing of this information.

3. The definition, the importance and the boundaries

The definition: the principle of statistical confidentiality is not fully defined in a positive way. Nevertheless, in face to the established in the Law, it is possible to give a definition: *obligation of the statistical producer to guarantee that the information specifically collected for statistical purposes does not suffer any deviation from its purpose during the statistical process, since the data collection until the dissemination.*

The importance: this principle is fundamental. It guarantees to data suppliers that the information they provide will not be used differently from its initial purpose, unless a specific authorization from them has been given. In this manner the trust of respondents in the system is assured.

The boundaries: it is in a sufficiently restrictive context of the possible exceptions to the principle that the Portuguese Law is drawn. Looking "from outside to inside" the law explicitly denies to any external service or authority the possibility to command or authorize the access to confidential statistical data collected by INE or other public bodies.

On the other hand, looking "from inside to outside" the current Portuguese Law admits very little cases as exceptions to the general rule of confidentiality of statistical data. As a main rule it says that *all the individual statistical data are confidential* and not able to be disclosed either for public or private purposes. It is in force thus a general rule that prohibits the transmission of individual data.

An exception is done to all public data concerning Public Administration on which the opposite rule is in force, that is, the public information when used for statistical purposes is not subject to statistical confidentiality. This rule is in accordance with the general principle established in the Portuguese legal system, of free access to the administrative documentation.

4. The future of statistical confidentiality in the Portuguese system

4.1. What exists?

The impossibility of access to individual statistical data as a general principle applies to all individual persons, physical or legal, even in the second case in a mitigated form. The truth is that in case of physical persons the law absolutely prohibits the access or dissemination of data, while in the case of legal persons the law allows that SC analyzes specific requests having as limits to its decision the following variables:

- preserving the competition of the economic agents;
- guaranteeing the trust of respondents in the statistical system;
- assuring that statistical data requests refer exclusively to juridical created persons;
- assuring that all requested statistical information is intended to planning and economical purposes or external economical relations;

It can be concluded, therefore, that the current law includes very few exceptions. This situation places each time more difficulties in dealing with some requests whose purposes do not fit in the settings of the law. This creates to SC situations of some "uneasiness" for the discrepancy between the relevance of attending to specific requests and the impossibility to find a legal justification for disclosure.

Concerning specific requests made by researchers for scientific purposes the national law does not state anything. In fact, for these cases and considering the existence of community Regulations on this subject, though specifically conceived for the analysis of requests made to Eurostat, SC has analyzed Portuguese requests using these Regulations criteria. For instance it is very important that researchers belong to recognized scientific institutions, namely Universities, which are co-responsible for the use of the requested data by the researcher. Also very important is a detailed description of the project and all methodological aspects as well as its duration and the kind of results dissemination that will be done.

The analysis of researcher's requests is made by a group of SC members recruited according to the specific matter, always in the presence of the researchers and the institution they belong to, so that all possible doubts are immediately solved.

In the last years the number of requests made by researchers and analyzed by the SC intended mainly to economical studies and has increased due to the need of statistical data at its maximum detail.

In the case of scientific research, the possibility to create solutions for releasing information is verified, even so applying community legislation. However there are many other situations, analyzed by the SC and deserving a special attention (as for instance public health or special interests of criminal inquiry), where the Law does not allow the release of statistical data.

4.2. What may be improved?

After presenting the Portuguese Statistical System and the legal framework of statistical confidentiality and the main lines defined in the Portuguese Law, it matters now trying an answer to an important question concerning all the statistical systems:

- how to make compatible the increasing requests of statistical detailed data necessary to political decision and other purposes with the need of keeping and assuring the respondents trust?

The problem, at least in the Portuguese case, does not present an immediate solution, even though some changes are now beginning.

It will be fundamental to take advantage of what exists and can be improved developing three main lines:

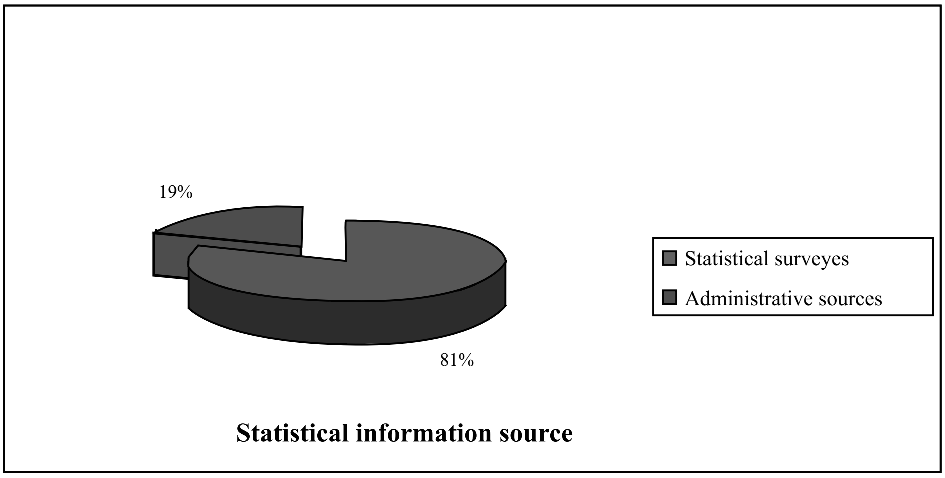
1 – The knowledge of international good practices, mainly through increasing the participation in events as the present conference.

2 – The reinforcement of the relation between INE and the national authority responsible for personal data protection (National Commission for Data Protection).

3 – The progressive use of more data from administrative sources.

Concerning personal data, the statistical Portuguese system is very conditioned by the Personal Data Protection Law because this text – though in principle is a faithful transposition of a community Directive – does not consider statistics as an exception to any other personal data collection, processing or dissemination. This situation is strongly conditioning INE activities specially when personal data are collected trough a survey and specifically used for statistical purposes.

On the other hand, and concerning personal or other data, the Portuguese system strongly depends on the collection through specific inquiries – as it can be seen in the graphic below – not taking complete advantage of the existing public administrative data because of legal difficulties. If these administrative data were used, statistics would appear as a second use of the collection and some of the confidentiality problems would be solved.



4.3. What can be done?

The Portuguese system and its 1989 Law need a deep reformulation. The solution to be found should conceive new ways that could articulate the growing political and social pressure for getting easily reliable and accurate statistics with the maintaining of the respondents trust. This is not an easy commitment if we consider that simultaneously, different and sometimes opposite values are involved as for example the right to the information and the right to privacy.

A first step might consist on the exploitation of the current attempts to change the legal framework of the Portuguese Statistical System by making use of the diagnosis of constraints and solutions already internationally identified concerning statistical confidentiality. The question is to take advantage of the international best practices that fit to the Portuguese case with some adaptation work.

The solution eventually found to move ahead with the commitment between respondents and INE should specifically review the following situations:

Data concerning physical persons – overtake the established in the Personal Data Protection Law and reconsider the specific need of authorization of some statistical procedures.

A possible solution could consist of creating, in the new Law, an element to assure continuously the linkage between INE and the National Commission for Data Protection. This would allow an immediate and quick reaction that would differentiate INE demands from other requests made to the Personal Protection Data Commission. This would also avoid the current long waiting periods not compatible with the regular and on time statistical production.

Scientific research – include this purpose among the exceptions to confidentiality, so that researchers could reach an easier access to statistical data.

Other specific purposes – enlarge the current group of exceptions considering other purposes as public health or criminal inquiry, in order to better adequate the exceptions of the Law to the user's needs.

The second step consists of finding the balance between the guarantee of maintaining the confidentiality of the data providers and the obligation of satisfying users needs even when they ask for high levels of detailed data.

This implies an enlarged concept about the boundaries of statistical confidentiality resulting of a new "agreement" between statisticians and respondents about their respective expectations.

The third step concerns an efficient control and management of all the confidentiality questions. This could be solved by the creation of a unit that could provide integrated proposals and answers. This unit could function as a liaison between INE and SC. INE could propose new rules and follow their application and also clarify all the doubts concerning internal regulations or the legal framework on statistical confidentiality. It could also coordinate and harmonize procedures in application of automatic solutions based on mathematical methods.

Only if these solutions can not be applicable at all and simultaneously the users have specific needs of detailed data, should the requests be sent to SC for analysis.

References

Decree nº 294/2001 of 24 November - Access to personal data for INE)

Eurostat, Statistical Confidentiality Committee, (2003) *Protection of Confidential Data*

Law nº 6/89 of 15 of April - Basic Law of the National Statistical System

Law 67 / 98 of 26 October - Personal Data Protection Law

Pinto, Ana Dulce, (2004) *Structuring principles of the new internal regulations about statistical confidentiality,* Seminário sobre o Princípio do Segredo Estatístico, Lisboa

Regulation (EC) nº 322 / 97 of the Council, 17 February

Regulation (EC) nº 831 / 2002 of the Commission, 17 May

UN, (2003) *Statistical Confidentiality and micro-data*, Conference of European Statisticians