WP. 46
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Geneva, Switzerland, 9-11 November 2005)

Topic (vii): General statistical confidentiality issues

PROVIDING ACCESS TO DATA AND MAKING MICRODATA SAFE, EXPERIENCES OF THE ONS

Invited Paper

Submitted by the Office for National Statistics, United Kingdom[1]

[1] Prepared by Paul Jackson and Jane Longhurst.

Providing access to data and making microdata safe, experiences of the ONS

Paul Jackson[1] and Jane Longhurst[1]

[1] 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, that is, data sets containing for each respondent the scores on a number of variables. It is in the interest of the users to make the microdata as detailed as possible but this interest conflicts 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. The increasing number of databases containing personal and business data, advances in technology, improvements in matching and linking techniques, the increasing value of identifiable information for non-statistical purposes and an increasing public awareness of privacy issues all contribute to this increasing threat. Thus, NSIs are confronted with the problem of ensuring that the risk of a breach of confidentiality is acceptably low, while at the same time providing as much detail as possible in the microdata that is to be released.

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, and is introduced in the next section of the paper. Section 3 provides an overview of the framework. The rest of the paper is based on the structure of the framework covering the legal, policy and ethical issues that need

to be considered when providing access to microdata as well as risk analysis and management.

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.

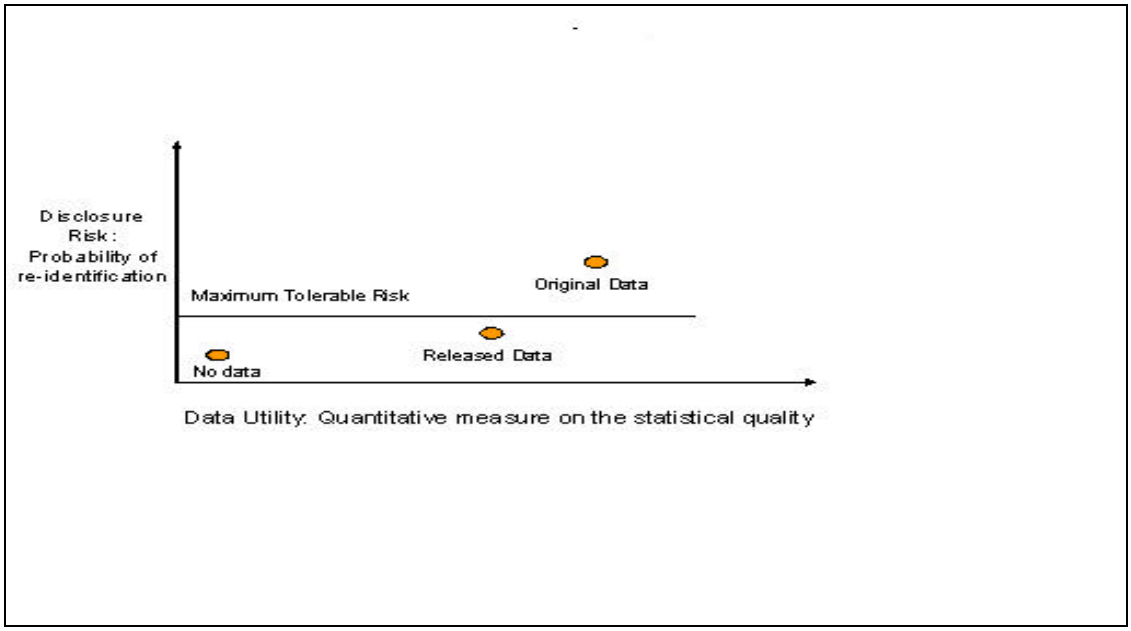

Fig 1. R-U Confidentiality Map (Duncan, et.al. (2001))

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. Note that the maximum tolerable disclosure risk is represented in the R-U Confidentiality Map as a horizontal line. This line can also be sloped representing the fact that the NSI may be willing to release slightly more disclosive data if there is much gain in utility. In addition the maximum tolerable risk for an NSI is not fixed but will change over time

as it is affected by public perception and intruder behaviour. An R-U map is a useful tool that will support informed decision making.

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 (ref). The figure below provides an outline of the framework.

The first three stages involve establishing why confidentiality protection is need, this could be associated with legal or policy issues. The next three stages relate to the risk-utility approach that is required. A detailed understanding of the microdata files must be established including the users and uses of the data. A risk analysis must then be undertaken. These assessments of risk and utility must be taken into consideration at the risk management stage where statistical disclosure control methods or restricting access (or a combination of both) can be used to reduce the risk in the data 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.

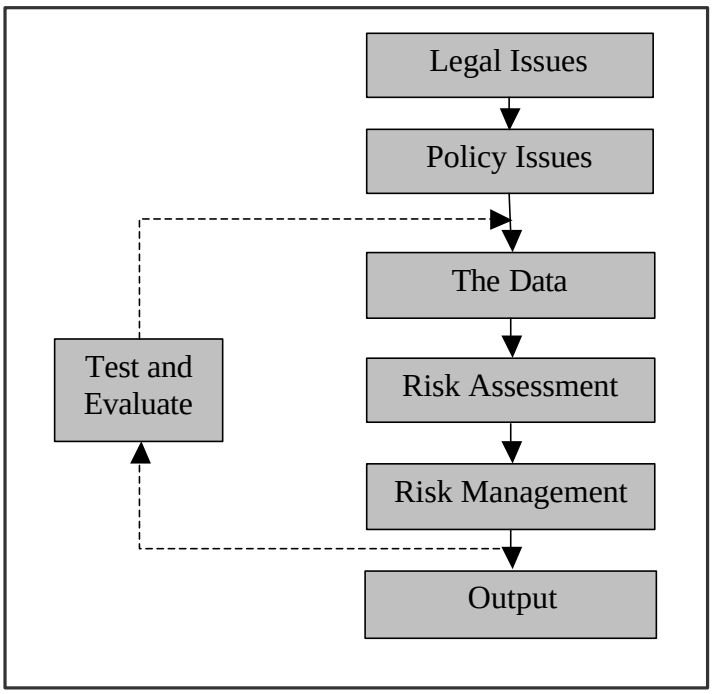

Fig 2. A Framework for Protecting and Providing Access to Microdata

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. Powers to share micro-data may be express (found in statutory law) or implied (where it is reasonable to imply suitable powers in more general legislation, or within the general authority of a Minister of a department.)

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. A duty of confidence is presumed to exist where the information in question has the necessary quality of confidence (i.e. not public domain and of some value) and where the information was provided in circumstances giving rise to an obligation of confidence (i.e. under a pledge, or under an unwritten but reasonable assumption of confidentiality). 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. 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. This makes the wording of pledges used on ONS social surveys of critical importance for future use of identifying data.

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 Code also recognises that these enhanced datasets may be more disclosive and that maintaining confidentiality is paramount. Data sharing must therefore take place within a framework designed to honour this commitment. 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 standard for the 'Confidentiality Guarantee' to be met in publications, and the standards for governance when sharing statistical micro-data:

> ***The National Statistics Confidentiality Guarantee***
> *Statistical disclosure control methods may modify the data or the design of the statistic, or a combination of both. They will be judged sufficient when the guarantee of confidentiality can be maintained, taking account of information likely to be available to third parties, either from other sources or as previously released National Statistics outputs, against the following standard:*
>
> *It would take a disproportionate amount of time, effort and expertise for an intruder to identify a statistical unit to others, or to reveal information about that unit not already in the public domain.*

The minimum requirements for governance set out in the Protocol are to be captured in an operational record and shall include :

- The organisation/Data Owner granting access to the data and the Responsible Statistician/Data Owner's representative within that organisation who has care of the data.
- 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
- A description of the arrangements by which the data access agreement is to be reviewed.
- Details of any legislation enabling access and any legal constraints on the processing of the data once access has been granted
- Details of any ethics or steering committee that will oversee the access
- Details of how any commitment made to respondents to surveys is being upheld and the confidentiality of their information safeguarded
- Procedures to be followed if the accessed data is to be matched to another data source.

- Any intended duplication of the data
- The process that will be followed to resolve any disputes
- 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 is the means by which 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.

[1] Full DCA report http://www.dca.gov.uk/majrep/rights/mori-survey.pdf

6 The Data

Sections 4 and 5 outline why an NSI needs to maintain the confidentiality of respondents when providing access to microdata. Therefore, when releasing microdata, the NSI must undertake a risk assessment of the file. Before this can take place a detailed understanding of the microdata file must be established.

When making decisions on confidentiality protection it is important to have a good understanding of the data that requires protection. This includes knowing the source of the data, the main uses and users of the data and the 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, variables on the file and the quality of the data 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. For example if the microdata file is to be used as an example data set for students then the level of detail required (and therefore risk involved) is likely to be less than the detail required by an academic involved in a particular research project.

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 uniqueness and hence for microdata disclosure risk comes from individuals that are unique for a certain combination of identifying variables. If an individual is unique for a certain combination of characteristics in the sample and in the population then the risk of identification will be high.

The variables in each record of a microdata file are of two types: identifying or sensitive variables. Identifying variables are those variables that allow one to identify a record, e.g. age, sex, occupation, place of residence, country of birth, family structure, etc. Sensitive variables are variables that an intruder will discover following identification. A combination of identifying variables is defined as a key and provides the basis for identification of a respondent. A key is used to measure uniqueness in the sample and in the population.

The key has K cells and each cell $k = 1,...,K$ is the cross-product of the categories of the identifying key variables. The score combinations of the key are denoted by 1, 2

….., K. For example, if the key is composed of age (in six categories) and sex (in two categories), there are 12 different score combinations, so K=12. The number of elements in the population with key value k is denoted by F_k (k = 1,2,….,K), and the corresponding number of elements in the sample is f_k (k = 1,2,……,K). All F_k are strictly positive, but some f_k may be equal to 0. The value of K need not necessarily be equal to the product of the numbers of categories of the key variables. If some combinations are impossible, K is less that the product of the categories.

Keys can be used to quantify the disclosure risk of a microdata file. Individual risk measures for each record are based on calculating the probability that it can be re-identified. These individual risk measures can be aggregated to obtain global risk measures for the entire file.

Here two global risk measures are defined:

Number of sample uniques that are population uniques:

$$t_1 = \sum_k I(f_k = 1, F_k = 1)$$

Expected number of correct matches for sample uniques to the population:

$$t_2 = \sum_k I(f_k = 1)\frac{1}{F_k}.$$

This section outlines the different approaches to assessing the risk of microdata that are currently implemented or in development at the ONS. Disclosure risk scenarios are used to determine key variables, a checklist can be implemented to identify risky records or variables in a microdata file and quantitative risk assessment methods based on the measures defined above can also be implemented.

7.2 Disclosure Risk Scenarios

As outlined above the 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.

A disclosure risk scenario is in fact a model for the actions and knowledge that a hypothetical intruder is assumed to apply when attacking a data set and provides a rational basis to attempt to prevent the disclosure attempts of an intruder.

The ONS currently has identified six likely disclosure risk scenarios that should be considered when releasing microdata. Each scenario has between 8 and 13 identifying key variables some of which are in common between the scenarios. The scenarios cover topics such as possible political attacks, private database cross

match, journalist, local search and nosy neighbours. For example, the private database cross match includes the following key variables: region, age, sex, marital status, number of cars, number of dependent children, workplace, distance of journey to work, number of residents, number of earners, tenure, and primary economic status.

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. In order to keep disclosure risk scenarios up to date NSIs must monitor the increasing number of databases containing personal and business data that could be available to a potential intruder. 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 such exercises 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 aim of the checklist is to provide a structured procedure that is flexible, objective, and, as far as is possible reasonably safe.

For each bullet point below the data supplier must provide the necessary information, justify why the level of detail is needed and any analysis of associated disclosure risk that has been carried out:

- level of geography used
- detail of ethnicity coding
- details of occupational coding
- information on any other visible or traceable variable released in the dataset
- sampling fraction of the survey (and of the data released)
- sampling design of the survey (e.g. details of cluster sampling etc)
- inclusion of hierarchical (e.g. individual/household) or longitudinal variables
- any measures used to assess and treat outliers in the data

- any assurances given to respondents before their consent was obtained
- information on previous releases of the same or similar data

The information provided on the checklist is used in the MRP process to make judgements about the risk posed by different microdata sets.

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 other techniques and developing the necessary software tools.

The basic quantitative measures of risk were introduced in 8.1. 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.

Quantitative measures for global disclosure risk 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. Acceptable levels of risk will be established through benchmarking against released files, use of the R-U confidentiality maps and will depend upon different access options (see Section 9.2).

This research is still in the early stages of development. Work will need to be undertaken to incorporate these risk measures with checklists and scenarios into a decision process for microdata releases. This is vital for the ONS as more and more demands are placed on access to microdata. The project will also consider the practical details of implementing different methods which often involve complex modelling or processing of large data sets and the interpretation of different measures and how they can be explained to internal and external users.

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 (especially samples drawn from frames based on addresses) the population base is unknown or only partially known through marginal distributions. 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. The research project at

the ONS has been established to evaluate different methods for achieving this, in particular heuristic and probabilistic models for modelling the population.

7.4.1 Probabilistic Modelling

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).

The probabilistic disclosure risk assessment depends on log-linear models to obtain expected cell means for a contingency table spanned by the key variables in the microdata. The smaller the expected cell mean the higher the risk that the record is a population unique.

In order to assess the robustness of different probabilistic models and to evaluate 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).

This research focused on model selection techniques for the log-linear models and the robustness of the estimates of the disclosure risk measures to deviations from the best fitted models. In addition goodness of fit criteria are developed for the very large and sparse contingency tables spanned by the identifying key variables containing the sample counts. Robust criteria that will identify the most appropriate model for obtaining accurate estimated disclosure risk measures need to be developed.

This work is ongoing at the ONS and more research is required to consider the analysis of hierarchical microdata, the impact of more complex survey designs on disclosure risk assessment, new and more robust goodness of fit criteria and model selection techniques and confidence intervals for the global disclosure risk measures.

7.4.2 Heuristics

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. The method assigns a per record matching probability to a sample unique based on the number and size of minimal uniques. A minimal unique is a set of key variable values which is unique in the sample and of which no subset is unique. For example, a record may be unique due to its combination of age, sex and

ethnicity but not on age and sex, or age and ethnicity, or sex and ethnicity. The minimal unique size in this case is three. A record may have many minimal uniques, so this same record may also be unique due its combination of age and occupation (minimal unique of size 2). The smaller the size of the minimal uniques and the more minimal uniques within a record the more risky the record.

By using the fact that the DIS method produces an implicit estimate of the total number of population units corresponding to the sample uniques in combination with the SUDA's grading system, it is possible to obtain a per-record estimate of each record's risk in terms of the certainty that an intruder would have in inferring a match was correct (given certain assumptions). This combined method is called the DIS-SUDA. The score obtained from the algorithm is heuristically linked to the estimate of $t_2 = \sum_k I(f_k = 1)\frac{1}{F_k}$ which is the expected number of correct matches, so each individual risk measure estimates $\frac{1}{F_k}$ as is the case for the probabilistic models.

This method was implemented by the ONS in the risk assessment of the Sample of Anonymised Records (SAR), microdata files Gross et al (2004). These files were sampled from the 2001 Census and so the population was known and hence the disclosure risk known. 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.

Work needs to be undertaken to establish whether the DIS-SUDA global risk measure can be used to make comparisons between microdata files as is required for the MRP decision making process.

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 though a combination of both design and controlled use. The choice of approach should take account of users needs. A high level description of different disclosure control methods that can be used to protect microdata files are outlined in this section as well as discussion of different access options.

8.1 Statistical Disclosure Control

Statistical disclosure control techniques for microdata are classified into perturbative and non-perturbative methods. Perturbative methods include: adding random noise to continuous variables or changing values of categorical variables according to a prescribed probability matrix and data swapping. These methods alter the data and usually have hidden effects. Non-perturbative methods preserve the integrity of the data. These methods include global recoding, sub-sampling and suppression.

For the majority of microdata files released by the ONS the non-perturbative method of recoding is used as a preliminary protection method. Two main types of recoding are implemented, either broad banding (where the number of categories for a variable are reduced or coarsened) or top/bottom coding (where top/bottom categories are combined together, e.g. over 80s).

Perturbative disclosure control methods have also been implemented by the ONS for the SARs microdata files. The DIS-SUDA was used to identify the variable or variable values in the key that significantly contributed to the disclosure risk. This was used to implement global recoding. This iterative process was undertaken in consultation with key users. 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. The individuals identified to be risky are turned into non-risky individuals or individuals that are not in the file. An intruder can therefore not be certain of making a correct identification. 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.