

WP. 37
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 (v): Confidentiality aspects of tabular data, frequency tables, etc.

A PROCESS FOR WRITING STANDARDS AND GUIDANCE FOR TABULAR OUTPUTS FROM ONS

Supporting Paper

Submitted by the Office for National Statistics, United Kingdom¹

¹ Prepared by Christine Bycroft and Philip Lowthian.

A process for writing Standards and Guidance for tabular outputs from ONS

Christine Bycroft* and Philip Lowthian^{*1}

* 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. In recent years considerable effort has been given to providing a consistent set of principles for the operation of the ONS, including confidentiality protection, that are now set out in a Code of Practice,. As the next layer of documentation, the Protocol on Data Access and Confidentiality expands considerably on the brief confidentiality statements 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 is now developing Standards and Guidance for confidentiality protection of tabular outputs from Business Surveys, Household Surveys and some Health Statistics. They include 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.

1 Introduction

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. 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 situation is complicated in the UK because there is no single piece of legislation governing statistical data collection carried out by ONS. There are different Acts under which specific data collections must operate, as well as Common Law obligations. In recent years considerable effort has been given to providing a set of

¹Thanks to Jane Longhurst and Marta Haworth for excerpts from the Abortions Standard and for helpful comments on earlier drafts of this paper, and to Neil Higgins for work on the reviews of business and household surveys.

principles for the operation of the ONS, including confidentiality protection, that are now set out in a Code of Practice. The Protocol on Data Access and Confidentiality 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 that include 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.

To date, work is in progress writing drafts of Standards and Guidance for tabular outputs from Business Surveys and Household Surveys, and a Standard for Abortion Statistics has been published. This will be extended to cover statistics from whole population administrative data, analytic outputs and other more complex derived outputs such as indices.

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 draft 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 (Longhurst 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**

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.

The reviews of existing practices to be discussed in the following section obtained information on all these aspects. Following on from this the Standards and Guidance will give a clear statement of the disclosure risks, define unsafe cells in tabular outputs, and state the disclosure control method to be used.

Section 5 describes in more detail what is meant by each of the headings and gives some illustrative examples from the work so far.

3 Review of Existing Methods

3.1 Obtaining details on current methods

The ONS came into existence as a means of centralising many statistical aspects of UK government. This meant that it was responsible for many surveys in a large number of areas, including over 80 business surveys and over 40 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 sources²
- Analytic Outputs
- Indices and other complex derived outputs

The first stage covers Business Surveys and Household Surveys, and others are planned to follow.

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 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 on 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 of the Review summarised these findings and highlighted where changes to existing practice was recommended, and areas where further work was needed. The review document 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.

² The 2011 population Census is dealt with separately

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.

A third section contained the Disclosure Control 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 for the associated outputs.

3.3 Abortion Statistics

In addition to these internally produced outputs, ONS was asked to carry out a review of disclosure issues around health statistics and to produce guidelines for handling health statistics across the health community. The health statistics review has been carried out as a separate process from the internal ONS reviews described above. The same Confidentiality Framework was used, and focussed first on a specific example, the abortion statistics.

However there has been a slightly different approach because we are working with external agencies. Abortion statistics are released by the Department of Health (DH). DH, as well as ONS and other agencies release many other health related statistics. A Project Board led by ONS, with representatives from DH and users of health data was established. Rather than use a questionnaire, ONS worked closely with DH who have detailed knowledge of the data source, disclosure risks and outputs. Through discussion and examination of proposed tables, intruder scenarios and the parts of the outputs that posed a disclosure risk were identified, possible disclosure control methods were discussed, and agreement reached on confidentiality protection. A published report provides standards and guidance for protecting the confidentiality of abortion statistics.

4 Results from the Reviews

This section provides some of the main results from the reviews of business surveys, household surveys and the abortion statistics.

4.1 Business Surveys

There is 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 Groups³) 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, cells based on a small number respondents were suppressed on quality grounds, and that these quality suppressions also provided protection of confidentiality. Some surveys published unweighted sample base numbers, or with a few surveys, unweighted estimates. A change of policy has seen much lower quality thresholds often used, and therefore there was a need to provide explicit rules for disclosure control.

4.3 Abortion Statistics

In 2003 the Department of Health rules stated that no cell could be published if the true count in the cell was less than 10, (zeros could be published), percentages and rates were to be rounded and rates could be published if the denominator was greater than 1000.

The review examined this approach considering both who was to be protected (the woman herself, the General Practitioner and the Hospital) and the different disclosure scenarios such as self-identification and the potential disclosure of information to third parties.

5 Examples from 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

³ 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.

or methods of disclosure control for a particular type of output. The second audience also 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 rules. The main part of the document then follows the headings of the Confidentiality Framework described above. 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. However for the Abortion Statistics, we are writing for an external audience, with no ready access to further advice, and the standard is long and detailed.

Brief examples⁴ from the Standards are now given under each heading of the Framework

5.1 Why is confidentiality protection needed?

The question of why confidentiality protection is needed is fundamental to the whole process and is key to 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.

The example below is an extract from the draft Household Surveys Standard that describes the legal, policy and ethical requirements for protecting confidentiality for household sample surveys.

Why is confidentiality protection needed?

In summary, the requirement for confidentiality protection of ONS household surveys is considered within the context of the Census Act (where applicable), the National Statistics Code of Practice, the Protocol on Data Access and Confidentiality, the survey specific assurances of confidentiality given to respondents and the common law duty of confidence relating to these public statements. These are all designed to ensure that respondents trust ONS to keep their data safe, so that we can maintain good response rates to our surveys, and continue supplying high quality information in our official statistics.

Legislation

There are no legal statutes directly relating to the confidentiality of household surveys. Some household surveys are carried out under the general powers contained in the Census Act (1920). Then they must comply with Census obligations. The 1991 amendment to the Census Act says that it is an offence to

⁴ Please note that those relating to Business Surveys and Household surveys are draft, and have yet to go through an approval process

"disclose any personal census information to another person, without lawful authority"(Census (Confidentiality) Act 1991 (c.6))

For all household surveys, the public statements made about protection of confidentiality define a common law duty of confidence that ONS must legally comply with. The main public statements are set out below. Common law means that we must do what we say we will do. Breach of the common law creates a right for the individual respondent to sue for damages.

ONS Code of Practice and Protocols

The National Statistics Code of Practice and the Protocol on Data Access and Confidentiality (DA & C) provide the ONS policy framework for confidentiality of Official Statistics. They in turn conform to principles set out in the UN Economic Commission report 'Fundamental Principles for Official Statistics', April 1992.

Of particular relevance to disclosure control is the following provision:
'The National Statistician will sets standards for protecting confidentiality, including a guarantee that no statistics will be produced that are likely to identify an individual unless specifically agreed with them.' (section 5(i), National Statistics Code of Practice)

The protocol expands on the Code of Practice and gives guidance on how the principles set out in the Code can be made operational. In particular section 1. b sets the standard that must be met:

'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:
i) 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.' Section 1.b.

The disclosure control methods presented in this Standard are judged to have met the requirements of the confidentiality guarantee.

Trust of respondents

In a very real and practical sense the ONS relies on the cooperation and goodwill of respondents to provide the data that is the basis of our official statistics. Fundamental to maintaining that cooperation and goodwill is ensuring that identifiable information is held securely, is only used for statistical purposes and is not revealed in published outputs. If respondents do not trust us to keep their data safe, they may be reluctant to respond or may supply poor quality information.

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.

The draft standard for ONS business surveys describes the data as follows.

Data: What are the key characteristics and uses of the data?

Table properties

Most tabular outputs consist of magnitude tables of financial variables (*e.g.* turnover, capital expenditure, sales) or employment. Some financial variables are net of two components and may have negative values (*e.g.* capital expenditure = acquisitions- disposals). Count tables (number of businesses) are produced from the whole population data on the Business Register. Table categories include geography, industry, sector, product code and employment size. Tables are produced at a range of geographical levels and industrial classifications. The Industry SIC variable may be at a 2, 3 or 4 digit level. The level of geography varies with the size of the survey, some producing only UK level data, through to outputs from the Business Register at small area level. Some surveys are used to create Indices such as the Average Earnings Index

Publication may be Monthly, Quarterly or Annual. Some surveys support ad hoc customer requests for tables, while others do not.

Data uses

The data are used by the ONS in the production of National Accounts and Balance of Payments and may be provided to Eurostat for combined EU tables. Other published tables have a wide range of uses such as for government policy formulation, allocation of funding, local body planning, by industry groups and academic researchers. Key data users include the Bank of England, Treasury and other government departments.

Access to business survey microdata is provided only within a safe setting environment at ONS.

Sample Design

The Interdepartmental Business Register (IDBR) covers almost the entire population of businesses in the UK. Most business surveys have a similar sample design using the IDBR as the sampling frame. They are generally based on a stratified sample design that includes a full-coverage strata for the larger

businesses. Other strata have differing sample fractions. The businesses in the full-coverage strata are often well known and easily identifiable.

There are a number of very small surveys targeted to specific industry groups such as financial services.

Quality

Because business surveys are compulsory under the Statistics of Trade Act, response rates are generally high, especially for large businesses.

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, though we wish to further investigate the appropriate level at which the p% rule should be applied. Application of rules for negative values also needs to be investigated further. The same rules should be applied for all survey outputs. Parameters are set according to judgements made by ONS of what constitutes an acceptable level of risk, and are not published.

The Abortion Statistics are an extreme example of highly sensitive personal information, which also represents high political value and strong motivation for intruders. The Abortions Standard describes intruder scenarios for identification leading to attribute disclosure, for identification (without revealing further information) and for self-identification. They find that protection should be provided for all these scenarios, and not just for the women involved, but for the practitioner carrying out the abortion, and the hospital or institution in which it takes place. The abortions standard defines unsafe cells as follows:

From consideration of the risk scenarios the unsafe cells for abortion statistics are those for which:

- the count is zero unless no other value is logically possible, or
- the count is below 5 for a Government Office Region in England, the country of Wales or a larger area, or
- it is less than 10 and the area concerned is smaller than Government Office Region in England or the country of Wales, or
- it is less than 10 and the variables are considered highly sensitive, or
- the count is associated with at most 2 practitioners, or
- it is associated with at most 2 hospitals.

The variables that are considered highly sensitive are:

- Young ages (<15)
- Late gestation (over 24 weeks)
- Procedure by gestation
- Medical conditions

Simple calculations such as rates or percentages do not necessarily make an unsafe cell safe and should not be used to protect data unless it can be demonstrated that one cannot work back to the original count. For abortion statistics rates or percentages should only be calculated from safe cells.

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

Much discussion has centred on the values of parameters, for example, minimum threshold values, and the unit and value of p for application of the $p\%$ rule. 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 the Protocol phrases such as “likely to identify”, “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. We have found that explicit definition of intruder scenarios has provided clarity for decision making on where disclosure risks for tables are likely to occur, and this leads to the setting of appropriate parameters.

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 software. 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: Theory and Practical Application for Statistical Agencies*. 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: Theory and Practical Application for Statistical Agencies*. 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*