Disclosure control issues associated with access to business microdata in safe settings

by

David Brown

Because of the difficulty of applying disclosure control treatment to business microdata that gives adequate confidentiality protection whilst not damaging the data for essential economic analysis purposes, access is sometimes allowed within a 'safe setting'. The granting of such access usually requires confidentiality agreements, as well as the checking of all outputs that are removed from the safe setting to ensure that confidentiality of individual businesses is adequately protected. Standard disclosure control measures are not usually adequate for this task. In this presentation, we discuss the problems, and some initial approaches to potential solutions.

Introduction

Due to increasing requirements for access to business microdata while still protecting the confidentiality of individual businesses, national statistical institutes have resorted to various forms of safe setting (Marsh et al, 1994) to provide researchers with access to sensitive business microdata for research purposes. There are three main types of safe setting: research data centres, remote access, and special licensing agreements. At first sight, these seem to provide good solutions to the confidentiality problems associated with sensitive microdata, especially business microdata, but also other highly disclosive microdata, for example, those from longitudinal studies. Much work has been done on practical and procedural aspects (e.g. Centre for Economic Studies, 2003, Statistics Canada, 2003), and there has been good progress towards solutions of many of the operational problems.

The remainder of this paper is mainly concerned with research data centres (RDCs), although there are similar issues and solutions for other forms of safe setting. The disclosure risks associated with direct handling of the microdata are managed by allowing access only within the RDC, and prohibiting removal of the microdata from the RDC. Some confidentiality undertaking is usually required. For example, Statistics Canada insists on researchers taking an Oath of Office to become what they term 'a deemed employee of Statistics Canada' before obtaining entry to their RDCs, in which they agree to

'not disclose or knowingly cause to be disclosed, by any means, any information obtained under the Statistics Act in such a manner that it is possible from the disclosure to relate the particulars obtained from any individual return to any identifiable person, business or organization.'

This oath is legally binding with repercussions if violated. Some anonymisation of the microdata accessible within the RDC is usually performed (e.g. removal of direct identifiers,
and broad banding of highly disclosive variables that are not essential for the research being undertaken).

There is, however, one area where progress has been rather slow. A requirement for the effective protection of confidentiality of the businesses contributing to the microdata is the checking of outputs removed from the safe setting, and papers to be published, for acceptable disclosure risk. Frequently this is the responsibility of the data custodian, or the manager of the RDC. There are still numerous unsolved problems in this area, and this means that Managers of RDCs are often rather restrictive in what is allowed out of the RDCs, especially for business data.

**Statistical agency guidelines**

Some statistical agencies have published guidelines for assessments of outputs and publications in RDCs (Centre for Economic Studies, 2003, and in draft form in Statistics Canada, 2003), although these vary substantially in detail, and in strictness of disclosure protection required, probably because they provide access to different types of microdata.

The US Centre for Economic Studies (CES, 2003) paper provides some guidance about disclosure avoidance. The following statements give an idea of the care with which disclosure control is applied by the CES.

'All research output must be reviewed for disclosure before it can be released for use outside Census Bureau facilities.'

'The range of information that can be released without violation of confidentiality is much greater for analytic results (e.g. regression coefficients) than for tabular data. Indeed, we emphasise that researchers should minimize tabular output because secondary disclosure is very difficult: the operating divisions release as much tabular data as possible, which limits the amount of additional tabulations that can be released.'

'Typically, the descriptive tabulations you should expect to remove will consist of small one or two-dimensional tables of variables that describe the samples that appear in the models. Projects that emphasise tabular output (as opposed to statistical models) will not be approved.'

'...tabulations present significant risk of what is called "complementary disclosure." By combining information from the released table with other sources of information, it may be possible to infer information on an individual survey respondent. The risk is almost always much greater for tabular descriptive output than for output from statistical models. Moreover, preventing complementary disclosure on clearing large tabulations for release virtually always imposes significant costs - both on you and on the RDC lab administrator.'

The Statistics Canada (2003) draft guidelines appear to be for access to sensitive sample survey data primarily from surveys with a longitudinal element, mainly relating to household and individual data, not to businesses. In this document, Statistics Canada provide detailed guidance about disclosure protection procedures for tables, often involving suppression of certain types of cells. They provide more detailed guidance for other outputs, e.g. related to analytical outputs from parametric and non-parametric models, and analyses using hierarchical data. They also discuss confidentiality problems associated with geography and indirect identifiers. In particular they emphasise that

'you limit the printing of your tabular output to the minimum necessary to describe the sample used in your models, and how it might compare to an underlying population.'
Even with this restriction, other precautions as mentioned above must be followed. In particular, there is an emphasis on the risk of residual disclosure.

'While a table on its own might not disclose confidential information, disclosure can occur by combining information from several sources, including external ones. When released information can be combined to obtain confidential data this is called residual disclosure (e.g. suppressed data in one table can be derived from other tables).'

Both agencies emphasise a readiness to engage with researchers to facilitate the research they intend to carry out while still adequately protecting confidentiality. Off-the-peg solutions are frequently inadequate. Discussion is then required to tailor or adapt disclosure control methods to facilitate the primary purposes of the research project, while still maintaining confidentiality.

**Disclosure control of outputs other than tables**

It might be thought that all that is required is to ensure that some standard disclosure control treatments have been applied to analyses being removed from the safe setting or published. This is, however, not always so easy. One variant of the problem occurs when a researcher requests permission to remove from the safe setting analyses that are less usual or are in an unfamiliar form, so that they don't obviously present high disclosure risks, when they are in fact potentially or actually disclosive. This can happen, for example, when a tabulation is expressed as a series of model fits. The remedy then is to only allow model fits that include a continuous covariate, and to ensure that there is adequate variation in the covariates, and that the estimated regressions coefficients are not close to zero.

In other cases, disclosive information might be released in association with some traditional statistical diagnostics or other pieces of supplementary information, for example lists of influential points or outliers when fitting a regression model. Such diagnostics can then be accessed within the safe centre, but not removed from it. Tables or graphs of residuals cannot be removed, since these enable the original values of the microdata to be reconstructed.

Where the distributions of the $x$-variable is highly skew, sufficiently accurate information about a large business can sometimes be determined merely from a fitted regression line and information already in the public domain. It would then seem prudent to examine very carefully any outputs where the $x$-variates in a fitted model exhibit substantial skewness.

Disclosure risk management for fitted models is a relatively new area of research, and few guidelines exist. Disclosure can also occur for other outputs whose construction involves fitting models, such as small area estimates. Until this area has been adequately researched, it will not be possible to fully protect such outputs from disclosure risk, other than by prohibiting their removal from the RDC.

Potentially more serious problems occur with series of releases of information taken out at different times. How does the checker manage the resulting disclosure risks? There are many potential examples involving the problems just discussed, but we first discuss a conceptually simple problem that appears to be difficult to manage without substantial restrictions on outputs that can be removed from the safe setting.
Why the strictness about user-defined tabular outputs?

Suppose a researcher takes away a magnitude table disclosure protected by the usual suppression methods for a business sector, A, involving a number of standard industry codes. The table gives some aggregate statistic from a census, as well as frequencies on which the aggregate statistic in cells are based. The table has been protected by the standard disclosure control method adopted by the agency for magnitude tables of business data (primary cell suppression based on threshold and dominance rules, followed by secondary suppressions where needed), and so there is no risk of disclosure from this table when considered on its own.

Suppose later the researcher then decides that 2 of the industry codes need to be removed from the tabulation because the businesses in this sector were possibly aberrant. He then takes out of the safe setting the same tabulation protected by the same suppression method but for the reduced sector, B. All the outputs were assessed by the checker as being individually safe, since the standard disclosure control routines had been applied to each output individually. It would therefore appear at first sight that there is no particular disclosure problem with these tables.

However, the researcher subsequently finds the results puzzling, for example, there might be much bigger differences than he expected between the two tabulations. He then decides legitimately to look more closely at the results for the industry codes excluded from the second tabulation. He does not need to revisit the safe centre, since he has the tables for the two sectors; all he needs to do is to difference the two tables already produced. For those cells not suppressed in either tabulation, he can obtain the aggregates and associated frequencies for the cells of the table corresponding to the difference between the two sets of industry codes. In some cells, the resulting aggregate statistics for the differenced table might have been suppressed had the usual suppression algorithm been applied to them: for example, the frequencies on which the aggregate statistics are based in the two tabulations A and B differ by only one or two. In other words, by this means, the researcher has obtained access to what the statistical office would consider disclosive data.

Two questions immediately arise. How is it that standard disclosure control techniques allow such a disclosure-by-differencing to occur? Secondly how can this occurrence be prevented?

We briefly consider each question in turn. Standard disclosure control treatments are designed to provide confidentiality protection for a fixed set of standard tabulations, such as for example are listed as the standard (ward and higher level) tables for the 2001 UK Census of Population. These are provided only for a fixed set of administrative boundaries and definitions of categorical variables defining the table margins. Any other customised tables from the Census would require the application of quite different disclosure control methods to control the risk of disclosure-by-differencing. Most 'standard' disclosure control methods are not designed to protect against the risks associated with release of series of closely related tables.

A safe approach to managing this risk is to adopt a strict policy that prevents series of related tabulations being removed from the safe setting (as for example is applied in CES, 2003). An alternative is to develop methods of disclosure control that provide some protection against disclosure-by-differencing. This is an inherently difficult problem. Such methods are not very well developed, and appear to have been very little used so far. One potential method discussed next would probably be strongly resisted by users as it would involve sufficiently large perturbations being applied to every cell of the table.
The method, that requires further testing and development before it can be used in practice, is controlled cell perturbation, in which the aggregate statistics in each cell are perturbed to ensure a margin of uncertainty around each cell. The method maintains additivity of the table, while providing a stated level of protection to any set of cells (Salazar-Gonzalez, 2003). This is similar in principle to the controlled tabular adjustment method advocated by Dandekar (2003), and Cox and Kelly (2003). The Salazar-Gonzalez method can ensure any stated level of protection for any specified set of cells of the table under consideration, in the sense that the true values in the cells cannot be determined to a greater accuracy than intervals specified in advance. However the protection does not extend to disclosure-by-differencing, using other tables, unless these tables are protected at the same time. Clearly this is impossible for tables that have not yet been defined. However, perturbations of sufficient amplitude applied to every cell will induce a greater relative error on the differences between cells in two closely related tabulations, and thereby provide some protection against disclosure-by-differencing. Further work is needed in this area.

References


