Wales Institute of Social & Economic Research, Data & Methods
Sefydliad Ymchwil Gymdeithasol ac Economaidd, Data a Dulliau Cymru

Statistical disclosure detection and control in a research environment

WISERD DATA RESOURCES

WISERD/WDR/006

Felix Ritchie

December 2011

WALES INSTITUTE OF SOCIAL & ECONOMIC RESEARCH, DATA & METHODS
SEFYDLIAD YMCHWIL GYMDEITHASOL AC ECONOMAID, DATA A DULLIAU CYMRU

Authors
Felix Ritchie

Address for Correspondence:

Microdata Analysis and User Support
Office for National Statistics
Cardiff Road
Newport
South Wales
NP10 8XG

Email: felix.ritchie@ons.gsi.gov.uk or felix.ritchie@virgin.net

WISERD Hub Contact:

Cardiff University
46 Park Place
Cardiff
CF10 3BB

Tel: 02920879338

Email: wiserd@cardiff.ac.uk

Abstract

Statistical disclosure control (SDC) in a research environment poses particular problems. Most SDC research is concerned with ensuring that a finite set of tabular outputs are safe from disclosure, or that microdatasets are sufficiently anonymised. By its nature, a research environment is one where confidential data is made available for analysis with very few restrictions. Imposing SDC rules not designed specifically for this environment may lead to excessively complex rules which still fail to achieve the objectives of flexibility and effectiveness.

This paper argues that the research environment requires a different approach to traditional SDC based on a small collection of simple rules with a necessary “fuzziness” in interpretation. This requires (a) clear agreement on the principles and general purpose of SDC (b) the demonstration of classes of safe and unsafe outputs and (c) the active involvement of researchers. However, this does raise a number of practical issues.

Key words

Confidentiality; output protection

Acknowledgements

I am grateful to Rhys Davies for reviewing the text and suggesting the additional section on practicalities. My thanks also to ONS colleagues and participants at the 2006 CAED conference and 2007 Workshop on Data Access. All errors and omissions remain my own. Opinions expressed here are those of the author and do not necessarily represent the views of HM Government.

This paper originally circulated and was referenced as Ritchie F (2007) *Statistical Detetction and Disclosure Control in a Research Environment*, mimeo, Office for National Statistics, June**.**

1. Introduction

Historically the role of national statistical institutes (NSIs) has been to collect large amounts of information on all aspects of individuals and businesses. The publication of tables of aggregate data from these sources is the core function of NSIs. However, in recent years the research potential from using the underlying microdata has grown in importance.

Most NSIs provide some sort of access to microdata, although the extent of this varies considerably across countries and across data types. For example, there is widespread access to social data, as this can be anonymised effectively without damaging the information content significantly. In contrast, the use of business data is typically much more restricted, and little, if any, perturbation or anonymisation is carried out.

In terms of the access methods, circulation of confidential data is often restricted by the use of special licences or by remote job submission models, as in Australia and New Zealand. A number of NSIs also provide dedicated laboratory facilities for research into disclosive microdata. This may be at a physically controlled location (as in the US, Canada, Italy or Germany) or through a "virtual lab" (as in Denmark, Sweden and the Netherlands).

New technology, particularly the development of user-friendly thin client systems, has made the provision of lab facilities increasingly appealing[1]. The result is that demands upon NSIs to improve access to data are increasingly being met by innovative lab solutions. Along with flexible remote job submission systems, the provision of "research environments" (where manipulation of data and the choice of statistical models are both largely unrestricted) is therefore growing strongly.

This growth in use of research environments presents a problem for statistical disclosure control (SDC)[2]. The typical focus of SDC has been on ensuring the non-disclosiveness of aggregates or, in recent years, generating non-disclosive datasets for research use (often

[1] A "thin-client" system is where processing is carried out on a remote computer; the client computer appears to interact fully with the data but only sends instructions and sees the output of the operation. "Fat-client" systems are where processing is carried out by the client machine. "Remote job submission" is when a program is sent to a remote computer for execution and the results of the program returned; there is no direct interaction with the data.
The main benefits of thin client systems are simplified data management, improved security, and the decoupling of location from access. The last two are particularly appealing for NSIs, but while thin-client processing is historically the default operating mode for large computer systems, it is only in the last decade that thin-client solutions for Microsoft Windows™ systems have become viable for suppliers and users. Hence, recent years have seen a strong growth in the provision of lab environments.

[2] Although disclosure detection and control are two distinct concepts, in this paper for simplicity we use SDC to cover both, distinguishing where necessary.

called “public use” files). There is a large literature on SDC in respect of aggregates and public use files.

However, SDC for disclosive microdata in a research environment requires a different approach. The key problem is the predictability of outputs. This makes the scenario-based modelling used to evaluate the safety of public-use files, for example, difficult to use effectively.

Compared to regular SDC research, there is almost no literature on this. The *Journal of Official Statistics* special edition on disclosure limitation (Feinberg and Willenborg, 1998) did not discuss research environments in any one of its thirteen papers. Recent international conferences have focused on either the physical aspects of safe settings, or on preparing safe files for distribution (see, for example, UN (2004, 2006, 2008); Domingo-Ferrer and Torra (2004), Domingo-Ferrer and Franconi (2006). The European Statistical System programme includes research into output disclosure for the first time in the ESSNet project, commencing in 2008. Apart from Reznek (2004), Corscadden et al (2006) Steel and Reznek (2006), and Ritchie (2006a, 2006b), which all discuss the release of analytical outputs, there appears to be little analysis of some of the general problems that arise when researchers are given free rein over data.

Partly this reflects the set-up of NSI research centres. These are often a small part of the NSI, operating with relative independence and staffed by experts with practical experience of relevant research. SDC knowledge is embodied in research centre staff.

However, there is a need now for a discussion of what constitutes effective SDC in a research environment. This has five drivers. First, with the increasing sharing of international data (particularly in the EU) there is concern over the lack of agreement on SDC standards, which reduces the likelihood of cross-border data sharing. Second, the increasing amount of research work being carried out has raised the profile of research, while the lack of any discussion has led to attempts to take SDC rules designed for aggregate outputs and anonymisation, and apply them to research outputs. This can be ineffective, irrelevant and needlessly bureaucratic; and at worst the blind application of inappropriate rules can be devastating for research. Third, the range of analysis carried out in research environments goes far beyond the traditional models used for designing SDC rules. Fourth, with increasing requests for potentially disclosive data to be made available to off-site facilities, there is a need for transparency in SDC procedures so that data used securely at an NSI retains its confidentiality when management is transferred to, for example, secure research centres at universities. Finally, while SDC for aggregation and anonymisation is regularly tested and developed, the lack of discussion about rules for research outputs means that there is little independent scrutiny of the internal rules the research centre managers have developed; nor is there much sharing of "best practice".

This paper aims to address these issues, particularly the last. It argues that SDC in a research environment requires a fundamentally different approach to proscriptive rules-based methods – the "principles-example" approach. This recognises explicitly the limitations of trying to specify exact rules, and places the focus on an understanding of principles to which rules can be more flexibly tied. This has implications both for the training of researchers and for the use of automated systems.

The next section comments on the research environment. We then look at the problems of applying hard-and-fast rules for disclosure control, and argue that the nature of the research environment means that rules are fundamentally difficult to specify. The following section suggests an approach based around very simple rules but complex application. This requires some education of both researchers and NSIs, and the criteria for approving outputs become necessarily complex. We conclude with an example from the UK, and some comments on sharing information.

2. The characteristics of the research environment

Most SDC is concerned with making aggregate tables safe, or with effectively anonymising microdata. This is a practical objective, because in most cases a finite set of tables, or "intruder" scenarios, can be specified, and the resulting "safe" data can be measured against these targets.

The contradistinction of a research environment is the unpredictability of outputs. Researchers produce tables, but those tables may be a world away from aggregate tables produced from the same data. Data may be stretched, twisted and combined in unexpected ways. Researchers may apply a very personal treatment to missing or out-of-scope variables, or may use unexpected sub-samples of the data. Data can also be combined from a variety of sources.

Moving away from linear aggregates, the range of research outputs expands considerably: linear and non-linear estimation, simulation, probabilistic modelling, Bayesian analysis, factor analysis, dynamic modelling, transition data, et cetera. After all, the reason for providing access to microdata is to allow researchers to explore a range of analysis which is not possible from simple linear aggregation, or which cannot be easily defined by an automatic process.

A basic statistical competency on the part of the researchers can be expected. All NSIs apply some level of checking into the background and qualifications of researchers. This is done partly to ensure that the work carried out on the data is scientifically valid, but also to lower the demands upon the NSI. While NSIs assist researchers in data-related questions, they would not normally expect to offer statistical mentoring.

In summary, we define a research environment as one where expert researchers have largely unrestricted access to disclosive data to produce an unpredictable set of outputs; and where it is neither desirable nor practical to fully specify ex ante modelling methods or data transformations to be used

We assume, for the purposes of this paper, that the researchers in the lab can be trusted not to deliberately misuse the data; and that the technical security of the lab is acceptable. These are important, but separate concerns; for a discussion, see for example Desai (2004) or Ritchie (2006b).

3. Difficulties with rules-based methods in a research environment

All SDC is based upon rules which are intended to guarantee the level of disclosure protection. These are designed to provide a clear, independent and verifiable set of standards, and are essential for production of non-disclosive aggregates or anonymised datasets.

Our purpose is not to argue that rules *per se* are inappropriate; instead, we argue that the nature of a research environment is such that trying to define an SDC strategy based largely upon rules which do not take full account of the range of transformations available is almost doomed to failure. This is because the unpredictability of outputs inevitably turns any general rules into a complex set of special cases.

We illustrate this by considering simple primary disclosure (that is, the risk of disclosure in a cell without reference to any other cells). A typical *threshold rule* could be:

> *a table for release must have a frequency of at least five observations underlying any displayed cell*

This is the sort of rule applied to aggregate data: for example, total turnover by industry. The cell limit might be based upon what the NSI thinks are the possibilities for collusion – in this case, a limit of five implies that the NSI believes that at most three respondents will collude to determine the implied values for a fourth party. On this assumption (and ignoring any possibility of secondary disclosure and dominating values for the moment), this rule guarantees the confidentiality of the microdata.

While this may be appropriate when the data is itself disclosive and can be easily identified with the data donor, this is overly restrictive when these conditions do not hold.

First, consider the disclosiveness of the data. A transformation may render this rule irrelevant. For example, if productivity per employee is being displayed, small numbers may not be a cause for concern: the ratio does not allow individual survey responses to be unpicked.

The threshold rule can then be amended:

> *…unless the data has been transformed*

However, this information might still be potentially useful. Suppose *growth* in turnover per employee was being displayed. While the original survey returns could not be determined from such a complex variable, the information on how a company's productivity changes may be commercially sensitive. The NSI may well consider this a breach of confidentiality, and so once more the rule needs to be amended:

> *…and the resulting information does not breach confidentiality*

However, this information may already be in the public domain. Growth in productivity per employee could be approximated by growth in gross profits per employee; if the company is incorporated, then this information is likely to be available from published company accounts. As the information being gleaned from the survey returns is qualitatively identical to that available from public documents, the confidentiality criterion is not being breached:

> *…by providing information which is not available from public sources*

However, if the information is not readily available then the NSI may be under an obligation to not provide commercially sensitive information:

> *…easily…*

Moreover, even if similar information is available publicly and easily, the NSI may still feel that allowing any inferences to be drawn from survey responses (for example, which could corroborate uncertain public information) would breach its confidentiality protocols. There may also be legal restrictions – that information supplied in confidence, even if ratified by public knowledge, may not be published by the NSI.

Turning to the issue of identification, this at least seems amenable to a simple rule. To an extent this is the case, but again there are hidden issues. First, the range of direct identifiers (name, address, industry, location) varies across data sets. The context of that identifier is also important. For example:

- in the UK a postcode is sufficient to identify any medium-sized business, but an individual or household only in very exceptional circumstances
- a five-digit Standard Industrial Classification (SIC) code may have hundreds of companies in one industry, and yet only contain one company in another industry, such as a government monopoly
- in health statistics, certain events (such as rare cancers) are strong identifiers because of their rarity; others (such as birth) are strong identifiers because of their ubiquity in other datasets
- geography *per se* is rarely disclosive; but in combination with other variables it almost always becomes one of the key identifiers (see Elliot (2004) for an example)

More intractably, the underlying data may not be collected at the relative identification level. Consider the case of UK New Earnings Survey data. This is a 1% sample of employees, but collected from companies. Although tables may have over five observations in each cell, this only counts the number of employees. It is quite possible that the employees in a cell might all come from one company (for example, if the table shows specialised occupations in a nationalised industry). If the NSI's disclosure rules are based upon identification of company returns, a cell with high-frequency data may still violate the NSI rules.

A similar example could be drawn for plant-level (as opposed to company-level) data, or for personal data where the characteristics of individuals may lead to identification through the family unit. The cell count may be irrelevant; what matters is the frequency of the unit of identification.

Without identification, data releases cannot be disclosive. But a combination of factors contribute to identifiability, which is very dependent upon context.

In summary, the simple rule has now grown to:

> *a table for release must have a frequency of at least five observations of the relevant disclosure control unit underlying any displayed cell unless the data has been transformed and the resulting information does not breach confidentiality by providing information which is not easily available from public sources*

This is a good deal more complex and addresses some of the above issues. Unfortunately, as a model for disclosure detection this is difficult to make operational. A phrase such as “not easily available” is an essential part of the rule, but impossible to specify in the general case. The phrasing is deliberately vague to cover all cases, but as a result does not cover explicitly any one case.

The definition also embodies a tautology: the data is non-disclosive when it has been transformed, and the data has been transformed when it is non-disclosive. There is no independent line which says “*this* is transformed data”.

The rule only mentions identification implicitly in the minimum cell count, as this is difficult to specify in a general rule which is meaningful.

Finally, the rule explicitly recognises that the relevant disclosure control unit may not even be part of the table.

In short, this “rule” has become a guideline which needs to be interpreted.

Disclosure control of linear aggregates is of course extremely difficult because of the potential for disclosure by differencing. The aim of the paper is not to set up straw men; threshold rules are the core of identification. However, this paper argues that the threshold rule should be not be seen as an end in itself, but as encapsulating the principles of the SDC – and hence needing to be evaluated in context.

4. Deriving of rules: the research zoo

Part of the difficulty with developing ever more complex rules is the manner in which they are determined. While fundamental rules such as the simple threshold rule above can be derived from first principles, the more complex derivations required a sequence of "what-if" scenarios.

This approach is typically used when testing the disclosiveness of public-use datasets. A dataset believed to be safe may be subject to testing by applying a number of "attack" scenarios. If potentially dangerous cells or observations are identified, then the control mechanism may be adjusted and re-applied. Alternatively, the result of the analysis may lead to rules determining safe tabulations.

The key to the use of scenarios is that the data under consideration form something approaching a "closed" system. In the case of aggregate results, the form of the output is known. For public-use datasets, the final form of outputs derived from data is not known, but the level of uncertainty around each observation can be assessed. Estimates of the probability of re-identification can be derived (see, for example, Elliott and Manning 2004), and the appropriate recoding or rules defined. While scenario testing cannot obviously cover every possibility, a finite set of plausible attacks can be defined.

A research environment with disclosive data is an "open" system. The person responsible for deriving rules must not just test the safety of results, but must also predict what form those results take. *Ex ante*, this is a much more difficult proposition.

An analogy might be to imagine disclosure control as providing an enclosure for animals which keeps the animals safe and alive. In respect of aggregate data and public-use datasets, the aim of disclosure detection is to probe the strength of the fences, walls etc, and to make sure the contained animals are prospering. The problem with research environments is that the SDC personnel must try to do this without knowing in advance whether the residents will be birds, fish, insects… Hence, all rules become contingency rules. SDC in a research environment is designing a zoo, not assessing a cage.

5. The principle-example approach to SDC in a research environment

The above discussion is necessarily an oversimplification of SDC development. Nevertheless, the implication is clear: trying to derive hard-and-fast rules to cover all the eventualities of a research environment is almost certainly doomed to failure.

But, as discussed in the Introduction, there is a need for some sort of “standard” which can be applied in a research environment. How can this circle be squared?

The solution lies in a different approach to SDC. This is based around four key issues: understanding on principles; few and simple, but flexible, rules; the explicit modelling of functional forms rather than data, wherever possible; and the education of researchers. The first two are to some extent already embodied in SDC, but it is in the latter two that the difference in approach needed by a research environment becomes important.

5.1 Understanding of principles

SDC in a research environment is not something that can be carried out automatically. It requires understanding of the outputs being checked, the potential disclosure risks, and the level of acceptable risk. Therefore a key issue is that there is agreement on the aims and objectives of SDC. This is not the same as agreeing rules; the principles may be common across an NSI, but different areas may implement the rules in different ways.

For example, the UK Office for National Statistics (ONS) Code of Practice (ONS 2002a) defines a Statement of Principles for “Protecting Confidentiality”; the associated *Protocol on Data Access and Confidentiality* suggests how the principles might be interpreted in practice:

> …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.
>
> ONS (2004) pp7-4

This is intended to give a generally comprehensible view of why results may not be released. Note that it does not specify any absolute standard of SDC, but uses “likely” and “disproportionate” to indicate where judgement is needed. There are no specific rules or parameters, and nothing about preferred control measures. Nevertheless, this is the ultimate standard against which all SDC activities must be measured.

5.2 Soft rules

Flexibility in the application of rules is essential for effective SDC in a research environment. One option is to have strict rules which may be “waived” at the discretion of the SDC reviewers; an alternative is to have rules which are inherently flexible. What is important is that the uncertainty in outputs is incorporated into the rules; for example, the threshold rule (section 3) could be replaced by:

> *Table cells will normally be considered non-confidential if the frequency of units is at least five; lower frequencies can be released if it can be demonstrated that the confidentiality principles (see…) would not be broken; higher frequencies may be required if there is insufficient variation in the data or the data can be identified with a small number of statistical units.*

This uses “may”, “insufficient”, “normally”, “small” to develop a rule of thumb, and outlines explicitly where “grey areas” arise. It puts the onus on the researcher to argue for a lower limit; in contrast, the responsibility for arguing for a higher limit is implicitly the SDC guardian.

Most importantly, the direct reference to some principle of confidentiality contrasts with the mechanistic threshold rule developed in Section 3. It acknowledges that this rule cannot exist independently, and that any output needs to be seen in an appropriate context.

So, although it may be desirable to have an independent standard, rules no longer have to stand by themselves. The problem then is how to avoid every release of data needing to be scrutinised to make sure it complies with these “fuzzy” rules. This is where the education of researchers and SDC practitioners becomes important.

5.3 Model-based reasoning

In research environments a large part of output comes in some form of "analytical outputs" (defined as non-linear aggregates of data). The difficulty is the infinite potential of researchers to manipulate data. It was noted above that this approach is similar to trying to guess how to build a home for an unknown animal. Do an infinite set of rules need to be developed?

Our approach is to realise that, in practice, there are a relatively small number of "animal types". If we can group animals into classes (things that fly; things that dig; things that climb; things that eat people), then the development of appropriate procedures is greatly simplified. While there are an unlimited set of actual animals, the characteristics of an animal are, by and large, all known.

This is the major change in SDC required for research environments: to look at the process of producing outputs, rather than the outputs themselves. We do not look at data for disclosiveness, but for the way that data is used. We call this "model-based" reasoning, and contrast it with "data-based" reasoning.

Consider a simple linear regression. A traditional approach would be to draw up rules concerning outliers, influential points, categorical variables, goodness-of-fit, use of public data, and so on. This rapidly becomes very complicated: for example, the inclusion of influential points in a regression may make for bad statistics but it does not necessarily make the regression disclosive.

In contrast, an analysis of the functional form of a linear regression reveals that regressions are almost always non-disclosive; that the problem cases are a small identified set; that most of the problems are due to the co-publication of means and totals; that a simple check on the disclosiveness of data exists; and so does a simple correction to make any regression non-disclosive irrespective of the data used (see Ritchie (2006a) for details). The analysis is straightforward, and the conclusions are clear. Specific examples of outputs of the type "Regression" can then checked relatively easily.

The problem, then, is not as bad as it seems. It is possible to shuffle whole swathes of potential output into relatively few classes whose properties can be studied. This does not mean that disclosure control becomes easier; for example, although percentiles can be treated as tables, the ordering of the categories presents a different issue of identification.

However, it does mean that the rules used for SDC can be kept small, manageable, and comprehensible.

Model-based reasoning also can help to direct attention onto the most “dangerous” outputs. Model analysis will demonstrate that some outputs are inherently liable to disclosure problems. In these cases, the particular instance needs to be reviewed in detail. For “safer” outputs, a limited checklist might be sufficient to demonstrate confidentiality is met. Returning to the zoo analogy, each lion and sheep is different. Resource-strapped zookeepers concentrate on understanding the individual lions, because the potential for damage from failure to observe properly is so much greater.

Finally, rules are based upon functional form may be able to stand independently:

> *A linear regression is non-disclosive if at least one coefficient is effectively suppressed; that is, it could not reasonably be determined from published information (Ritchie 2006b).*

Within this context we can also put tabular analysis in its proper place. Linear aggregations are inherently unsafe because of the potential for disclosure by differencing. As the method of generating the tables cannot be approved, the outputs must meet the appropriate standard:

> *Tables and other linear aggregations may not be released unless they can be shown to meet confidentiality guidelines.*

This makes clear that tables *per se* are unacceptable; a reason needs to be given for their release, involving direct application of the confidentiality principle. Note that it <u>does not</u> preclude outputs being released; it just shifts the burden of proof. No longer is there a rule stating that tables will be released if it meets certain criteria. Now a table will not be released unless it can be demonstrated to meet the criteria. This is a subtle but important change in emphasis.

5.4 Education

If the SDC rules to be applied embody an element of judgement, it is essential that the researchers are well-informed about disclosure detection and control. This education needs to include the principles, any rules and how they are derived from the principles, and how the rules are applied and interpreted. Without guidelines on interpretation, it may be difficult to achieve consistency, and researchers may be irritated or confused by apparently arbitrary decisions. In contrast, educated researchers will be more able to predict acceptable outputs, should understand the reason for non-approval of outputs, and should avoid burdening the output checkers with large amounts of unacceptable outputs.

It is clear that this makes SDC much more of a co-operative effort between researchers and the SDC team. This is deliberate: the aim is to make both parties share the same goal, the efficient release of non-disclosive data. Researchers want results to be cleared quickly. The SDC team wants results to be cleared accurately. These objectives are not incompatible if both understand and agree the principles and standards to which outputs must adhere.

There are other advantages. First, when new situations arise (for example a novel functional form which the SDC team has no rules or examples for), it means the SDC team and the researchers can work together to develop appropriate guidelines. Second, by drawing in researchers to develop the framework, it provides instant feedback on the appropriateness of SDC methods. Finally, the research environment provides direct access to experienced and proficient analysts. It seems a shame to ignore this source of ongoing peer reviews.

There are dangers in integrating researchers into the SDC framework. Most importantly, the SDC team needs to have the confidence and ability to defend its position. One could envisage a drift towards increasingly relaxed control as an ill-prepared SDC team is browbeaten into accepting lower and lower standards. One part of the solution is to make clear that the responsibility for final decisions rests with the SDC team, so that in matters of risk and interpretation of principle the NSI has the final say. A second part is to ensure that the NSI’s rules are reviewed periodically and independently.

This does not mean that statistical differences cannot be debated; but researchers wishing to challenge rules need to be aware that it is their responsibility to prove that a better method exists. Neither does it give the NSI licence to ignore suggestions for change. If the SDC team has insufficient technical knowledge, it needs to make a reasonable attempt to bridge the gap in understanding; otherwise the trust between the parties breaks down.

Hence, an important function of education is to ensure that there is a positive relationship between the NSI and researchers. As Desai (2004) notes

> "The best form of security is a good relationship with your users, if they feel they have a stake rather than being in the supplicant position they are more likely to act responsibly."
> Desai (2004, p5)

It has been argued that the involvement of researchers in SDC is dangerous: it gives them useful information about how to break the system. We do not consider this a valid argument. First, an ill-intentioned researcher can find much easier ways to remove data from a lab than by trying to get results past disclosure control[3].

Second, if all output goes through SDC, than a malevolent researcher could edit outputs enough to make result looks acceptable under any rules. Would a hard-pressed SDC team notice a deliberately fraudulent output?

Third, and most importantly, the discussion here is about involving researchers in the detection of disclosive results, and educating them in some of the things that can be done. Most NSIs provide some information about detection and control methods, and only decline to discuss details of particular controls applied to specific outputs. The same applies here. Moreover, the aim of schooling researchers in detection is to avoid control becoming necessary through better outputs.

5.5 Practicalities

The implementation of this approach does raise three particular concerns, relating to the volume of outputs and the skills of the NSI.

First, this strategy implies little scope for automatic SDC methods, even for linear aggregates. This implies that the volume of SDC work increases linearly with the amount of research done. As one aim of having effective SDC procedures is to encourage research, this potentially could be counter-productive for the NSI.

Second, manual SDC checking requires a level of statistical expertise on the part of the checker. Even for those with a statistical background, this requires some time to develop. For dealing with advanced queries on releasable outputs (is a Herfindahl index safe? A Gini coefficient?), statistical knowledge needs to be similarly developed. However, it is likely that

[3] This argument may not hold for remote job submission

those with a sufficiently developed statistical knowledge would not find SDC of other people's work particularly interesting or motivating. It may be difficult to fill and fund posts.

Third, the success of the model of SDC presented here depends to a large extent upon the relationship between the NSI and the researchers. The development of new methods, the avoidance of conflict over unresolved issues, the acceptability of outputs being submitted, are all facilitated by a good working relationship. This can founder on the NSI unawareness of how researchers work, or on researchers' lack of knowledge of the restrictions under which the NSI operates. Both parties need to put some effort into this relationship.

6. An example: business survey data research in the UK

We conclude with an example from the Virtual Microdata Laboratory (VML) at the UK Office for National Statistics (ONS). This thin-client laboratory facility provides access to sensitive data, mainly business microdata, to ONS, government and academic researchers. The work is largely analytical economics and econometrics. All outputs pass through the VML team for disclosure checking.

One important feature of the VML is that it is designed and managed by active researchers, and so the development of SDC guidance by the team is informed by practical experience of typical outputs. This has ensured that the VML has a thriving research base despite operating, on the face of it, an improbably strict regime. It also helps to develop the relationship between the VML team and researchers, and to give the former more authority in their arguments.

All researchers undergo a short training session. The bulk of this is taken up by SDC, and includes the principles, the dominance and threshold rules, and interpretation in the context of business data. Participatory examples are the primary teaching method both for researchers and VML staff, many derived from observed problems. The VML's principles of disclosure control are the same as ONS, but restated in a manner more appropriate for researchers.

Researchers are shown that outputs are grouped into “safe” and “unsafe” categories, and are also given guidance about what factors influence whether a “safe” output would be refused, or an “unsafe” output be approved.

As a result of the training, BDL researchers are relatively competent in assessing the disclosiveness of their own outputs. It was noted in Section 5.2 that the use of soft rules appears to lead to every output being subject to lengthy scrutiny. In practice, this is not the case, as researchers made aware of the SDC framework quickly learn the messages of safe outputs and produce results which can be cleared quickly. However, this has in itself caused some problems.

First, an essential part of this approach is giving researchers information about the parameter values used in SDC (eg dominance/uncertainty limits). Concern was expressed that this information could be used in other contexts to attack ONS outputs. The VML already used higher threshold limits than regular ONS outputs to guard against disclosure by differencing. This was extended to all the parameter values. The VML training now only discusses the VML-specific values with researchers. Researchers are informed that VML rules for SDC are stricter than ONS rules, and that requests for output are judged against VML rules only.

Second, outputs are sometimes presented without the necessary data to check results (such as tables without underlying frequencies). These are returned to researchers with a request for more information, and over time, researchers learn to provide the necessary information.

Third, the volume of output has increased. Some output files presented to BDL have been so large that the time to check the files has been significant. However confident BDL may feel in the capacity of the researcher to produce safe results, it retains the legal responsibility for ensuring that no disclosive outputs leave ONS, and as a result has refused outputs on the grounds of volume rather than disclosiveness. While number of outputs is a valid reason for refusing to release results (due to the potential for disclosure by differencing), this is not a very satisfactory outcome, and so BDL has had to adjust its training programme to increase the emphasis on the minimal set of outputs.

Overall, this co-operative approach to SDC required some investment in staff and some effort into getting the VML message across. In the longer-term, however, it has delivered a low-cost, scalable, transparent SDC system, with high degree of data security and acceptable response times.

There is generally a period of trial-and-error for all new researchers, which is often a frustrating time and needs to be managed carefully. Nevertheless, the overall impact of having an educated research group has been to significantly reduce the target release time for research results from two weeks in 2003 to two days in 2004. In practice, in 2007 results are turned around in one business day in 90% of cases, the rejection rates are roughly one paper per month, and the main reason for rejection is simply volume of output[4].

8. Conclusion

The need to develop SDC standards for a wider range of situations is clear. The case of a research environment is particularly difficult because of the unpredictability of outputs. This makes a dependence upon an absolute standard untenable in many situations, as it does the use of automated tools except in a very limited number of cases. However, by concentrating on the structure of outputs, results can be grouped into classes of varying sensitivity: tables are inherently unsafe and need to be assessed individually, panel data estimates inherently safe, and so on. Crucially, researchers need to be aware of these criteria and able to apply them.

While it may be hard to specify absolute rules, principles are much easier to determine and agree upon. These form an overarching framework against which particular cases can be assessed. They can also provide a cross- and inter-organisational consistency. Although the in-built flexibility makes a principles-based system more opaque than a rules-based one, there remains a common standard of judgement against which procedures can be tested.

This approach, of defining principles and modelling the mathematical structure of potential outputs, implies a knowledgeable SDC team – one that is aware not just of data, but of functional forms, and how to assess novel situations. Teams need to explicitly recognise that the scope of SDC will expand over time, and to have systems in place to incorporate new developments. Learning by example therefore becomes a key part of the training program for SDC staff. SDC staff need to have a general familiarity with statistics and a specific competency in the commoner functional forms used by the relevant researchers.

Finally, it is crucial that researchers are also involved in the SDC process. First, researchers and SDC teams have an interest in getting outputs cleared quickly, safely, and easily, and this is best achieved when all parties are familiar with the framework and rules; researchers can see a return on time invested in SDC. Second, in the more flexible world of principles-based systems, the involvement of researchers increases the transparency of the systems and hence reduces the scope for confusion and disagreement. Third, researchers have an

[4] One researcher did ask for a 300,000 line log file to be cleared. After some thought, this was rejected.

incentive to co-operate in the development of new rules and procedures, and are less likely to request novel outputs without also presenting an appropriate solution. Fourth, SDC training can be used to build a community of trust between researchers and SDC staff.

In summary, while SDC in a research environment may not be as cleanly controlled as in other situations, there is ample scope to develop transparent, accessible procedures which can be compared against a common standard.

References

Corscadden, L, Enright, J., Khoo, J., Krsinich, F., McDonald, S., and Zeng, I. (2006). Disclosure assessment of analytical outputs, mimeo, Statistics New Zealand, Wellington

Desai, T. (2004) “Providing remote access to data: the academic perspective” in UN(2004)

Domingo-Ferrer, J. and Torra, V. (2004) Privacy in Statistical Databases: CASC Project International Workshop Proceedings, Springer-Verlag, Berlin

Domingo-Ferrer, J. and Franconi, L. (2006). Privacy in Statistical Databases: CENEX-SDC Project International Conference Proceedings, Springer-Verlag, Berlin

Elliot, M.E. and Manning, A. (2004). The methodology used for the 2001 SARS Special Uniques Analysis. Mimeo. University of Manchester.

Feinberg S.E., and Willenborg, L.C.R.J. (1998). Introduction to the Special Issue: Disclosure Limitation Methods for Protecting the Confidentiality of Statistical Data. Journal of Official Statistics, 14:4, 337-345

Reznek, A (2004). Disclosure risks in cross-section regression models. Mimeo, Center for Economic Studies, US Bureau of the Census, Washington

Ritchie, F.J. (2006a). Disclosure control for regression outputs. Mimeo, Office for National Statistics, Newport

Ritchie, F.J. (2006b). Access to business data: dealing with the irreducible risks in UN(2006)

Steel, P and Reznek, A. (2006) Issues in designing a confidentiality-preserving model server, in UN (2006)

UN (2004) Monographs in Official Statistics: Work session on Statistical Data Confidentiality Luxembourg 2003, Eurostat, Luxembourg

UN (2006) Monographs in Official Statistics: Work session on Statistical Data Confidentiality Geneva 2005, Eurostat, Luxembourg

UN (2008) Monographs in Official Statistics: Work session on Statistical Data Confidentiality Manchester 2007, Eurostat, Luxembourg