Disclosure detection in research environments in practice

Felix Ritchie*

* Office for National Statistics, Cardiff Road, Newport, South Wales NP10 8XG
Email: felix.ritchie@ons.gov.uk

Abstract: There is an increasing demand for access to raw confidential data, and NSIs have responded by setting up controlled research facilities. However, the most common approaches to statistical disclosure detection and control (SDDC) struggle to accommodate the infinite variety of outputs produced in research environments. The main problems are designing statistical disclosure control (SDC) rules for unknown transformations of the data, and in managing the potential volume of outputs needing review.

Research facilities need a different approach to SDDC. In the UK, ONS has developed an approach based around classes of output, where the time devoted to checking outputs can be concentrated on the more unsafe outputs.

Defining "safe" and "unsafe" outputs based on the functional form of the model improves the efficiency and security of confidentiality checking, but is not straightforward. This paper outlines the broad approach, and then takes specific examples to show how the UK rules on analytical outputs (and the conditions attached to them) have been developed.

1 Disclosure control in research environments

After falling out of popularity, in recent years there has been an increase in the provision of Research Data Centres (RDCs) and other research facilities by National Statistics Institutes (NSIs). These pose problems for disclosure control. RDCs are designed to be places where experts have access to very detailed data; they select, twist, transform and link it in interesting and different new ways; and they produce complex outputs which need to be assessed for disclosiveness.

A disclosure control system should

- be transparent
- be consistent
- guarantee a level of disclosure risk
- not unduly restrict research output

As noted in Ritchie (2007), automatic disclosure control and hard-and-fast rules do not really provide this for RDC outputs. Hence all NSIs operate manual disclosure checking for their RDCs, and have guidelines for NSI staff and researchers (see, for example, Enright et al (2006), or the NORC/NIST website at dataenclave.norc.org). However, the potentially infinite range of outputs is a problem: how can any set of

guidelines cover all types of outputs in enough detail to be secure, with enough flexibility to be useful, and with enough consistency to be fair?

Ritchie (2007) proposed that grouping outputs into certain classes would go a long way towards making a feasible RDC checking system. He noted that the Office for National Statistics (ONS) in the UK was already developing such a classification for outputs from its Virtual Microdata Laboratory.

The aim of this paper is to shed light on how some of the concepts raised in the earlier paper can be used in practice. In particular, it shows

- how definitions of safe and unsafe outputs can be turned into rules
- how those definitions need to be based upon functional form and not data
- some of the steps to define an effective classification for an output

It also shows how the most popular SDC guidelines fit into this model.

2 Classifying the research zoo

As noted in Ritchie (2007), designing a disclosure control mechanism for a research environment is like designing a zoo. There may be uncertainty about the specific animals, but it should be possible to classify the various animals into groups: those that swim, those that fly, those that need water, those that eat unwary keepers. A herpetarium may not be designed with any specific snake, or species of snake, in mind, but should be able to effectively contain and keep healthy most of the snakes the zoo intends to stock.

These types can be allocated to broad classifications of “safe” and “unsafe”. The “safe” animals do not pose a significant danger to themselves or others. Hence, the zoo owner can then concentrate more time on the unsafe ones, in an efficient distribution of resources.

In terms of research outputs, “safe” and “unsafe” have clear interpretations for both researchers and NSI staff

- **Safe outputs**: these **will** be released **unless** the NSI staff can see some reason why they should be held back or adjusted.
- **Unsafe outputs**: these **will not** be released **unless** the researcher can demonstrate to NSI staff that the output meets the detailed criteria for this output

Note that the burden of proof shifts depending upon whether safe or unsafe outputs are being discussed.

For a safe output, the NSI team have decided that a certain class of output holds no disclosure risk in general. They may have concerns about a specific output which is an exception to the general rule. To enable the system to work well, these exceptions should be

- Small in number
- Well defined
- Comprehensible to and communicated to the researchers before research begins

The third bullet is essential. Developing an effective SDDC system for a research environment requires a positive relationship between researchers and NSI staff; there should be no surprises on either side. By clearly specifying exceptions, the researcher can be confident that the results produced will be acceptable for release.

For unsafe outputs, the NSI has decided that scope for disclosiveness in the output is such that it is, in general, unprepared to release the output. However, it leaves the door open for the researcher to argue a case as to why the decision should be changed.

Clearly, a researcher arguing that an unsafe output can be released needs to have a good awareness of the principles of disclosure control as well as the specific data and context. Hence, the researcher training necessary for effective SDDC should be

- focusing on and discouraging these unsafe outputs
- illustrating what can turn an unsafe output into a safe one
- informing researchers as to what needs to be demonstrated to make an unsafe outputs safe

3 Determining “safety”

For historical reasons, the SDC literature focuses on specific aspects of the data being released: dominance, outliers, use of public information etc. This is because the vast bulk of work on SDC has gone into making sure that either datasets have been anonymised effectively, or that aggregate tables are safe.

The use of these methods in research environments is inappropriate. The standard techniques are designed for a fixed input dataset and a finite set of outputs, against which a range of intruder scenarios can be tested. In research environments the input dataset and output data are not known when the SDC rules are being drawn up; and it is not practical to provide the same sort of detailed analysis of each output as is done for aggregate finite results.

The key to determining the “safety” of a dataset is to study the underlying functional form of an output. If there is no disclosure risk in an arbitrary dataset, then there cannot be any additional risk from having, for example, an “identifying” set of variables.

Note that this technique also helps to narrow down precisely where the risk arises. For example, in the linear regression model, the apparent risk arises from the coterminous publication of means and frequencies. Hence the linear regression model itself is safe, but supporting statistics may be problematic.

Each output type needs to be assessed for primary disclosure – that is, whether something can be inferred directly from the single output – as well as disclosure by differencing. Assessment should consider both cardinal and categorical variables.

It may not be easy to classify results. If something is fundamentally safe but has a large number of exceptions, it may be better to classify it as unsafe. For example, Table 1 shows examples of current classifications used in the ONS Virtual Microdata Laboratory (VML):

Safe	**Unsafe**	**Uncertain**
General linear regression[1]	Tables	General non-linear aggregations of data
Panel regression	Graphs	
Herfindahl indices[1]	Quantiles	Large high-frequency aggregate tables
Covariance matrices[1]	Cross-product matrices	

1 Restrictions apply; see below

Table 1 Examples of safe and unsafe outputs at ONS

Most of the "safe" outputs have further restrictions. These are where the exceptions come from, which the NSI uses to decide whether the output can be released. These are, as noted, limited in number and made known to the researcher. If those two conditions cannot be met, then the output would have been classified as "unsafe" or, at best, "uncertain".

The "uncertain" elements here arise from several factors. It may be that the model hasn't been studied yet; or that there is no simple statement of the exceptions for a safe output; or that there is no agreement yet on how to demonstrate safety in a way which is not labour intensive.

Before studying practical examples, two further considerations are needed. First, it is clear that, given a specific functional form, in theory a specific combination of data always exists that would allow a data point to be identified. A "safe" output is one where this theoretical possibility has no practical counterpart in analysis.

Following on, it needs to be assumed that the outputs are genuine statistical outputs. A malignant researcher could construct a statistic which appears a valid statistical result, but which has in fact been constructed simply to avoid detection. Dealing with deliberate cheating is outside the scope of this paper.

4 Examples

In this section we investigate some specific assessments of outputs . Only a selection of outputs is covered, to illustrate different aspects of the method. Further examples can be found or referenced in VML(2007).

4.1 Linear transformations of the data

For any linear combination of data,

$$\partial f(x)/\partial x = c$$

$$f(x) - f(y) = f(x - y)$$

where c is some constant. The first equation tells us that an individual data point can be assessed without reference to any other variable. Therefore all data points are a potential disclosure risk, and need to be assessed individually. The second equation notes that, if f(x) is a function which generates useful data when applied to a single observation, then there is a disclosure risk in the differencing of f(x).

All linear aggregates must therefore be classified as "unsafe": there is a high requirement on data checking, and a realistic risk of disclosure by differencing; and both of these are inherent in the mathematical form of linear combinations. This classification refers to all linear aggregates: tables, graphs, means, frequencies. It also covers quantiles, maxima and minima, which can be recast as tables.

This is why most SDDC literature in respect of the release of aggregate tables focuses on data problems, population uniques etc. The tables are linear combinations of data, and so cannot be made safe in their structure: safety must come through an appropriate choice of variables and sample. The alternative is to break the linear relationship between source data and output tables by, for example, recoding or rounding.

4.2 Linear regression coefficients

For a simple linear regression, consider the functional form of the estimated coefficients:

$$f(X, y) = \hat{\beta} = (X'X)^{-1} X'y$$

As Ritchie (2006) demonstrates, there is, in general, no danger from differencing; and the non-linear interactions mean that individual data points cannot be analysed. Hence this counts as a safe output.

This holds true for categorical variables as well as cardinal values. Although there appears to be a potential danger from differencing of models with categorical variables orthogonal to all others, the ability to identify observations relies on having the means available; and with the means available there are more direct ways to identify values.

There are some limited exceptions to be considered:

- If the explanatory variables are all categorical, then this is clearly a table and needs to be evaluated as such; or, if there are insufficient degrees of freedom for this to be a valid statistical model, exact values can be determined.
- If all the explanatory variables could be known to an intruder, then a value for an individual could be predicted; if the fit was particularly good, then potentially this could breach confidentiality restrictions by being close enough to a true value
- If the data comes from repeated observations on single unit, this could be informative, particularly in comparison with another unit

The first is simply a misclassification of a table as an analytical output. The second provides a theoretical problem, but in practice it seems that the fit needs to be infeasibly good (work by Statistics New Zealand suggests R^2 approaching 99%). Moreover, both a simple test for the accuracy of prediction and a counter-measure are easily available; see Ritchie (2006) for details.

The third exception is more interesting. While it is not clear what useful information could be derived, on a precautionary basis the VML currently bans regressions based on a single unit.

4.3 Cross-product and covariance matrices

Consider a cross-product matrix

$$M = X'X$$

This is an unsafe output. Frequencies and totals are identified by interactions with any constant or categorical variables. Hence this should be viewed as a linear aggregation.

Now consider the variance-covariance matrix generated by a simple regression

$$V = (X'X)^{-1}\hat{\sigma}^2$$

Should this be released? On the assumption that the estimated σ is available to the researcher, then it is a simple matter to turn V into a cross-product matrix, which is not safe. So this simple covariance matrix is unsafe.

However, this is not the case for the more general form

$$V = \left[(X'WX)(Z'Z)^{-1}(X'WX)\right]^{-1}\hat{\sigma}^2$$

Unless Z=X and W is the identity matrix, this cannot be unpicked. This holds even if W is known. This is a useful result because, for example, W will not be the identity matrix in any robust regression, let alone more complex models.

Can anything be inferred by combining V with the estimated coefficient vector? As:

$$V\hat{\beta}/\sigma^2 = (X'X)^{-1}(X'X)(X'y) = (X'y)$$

this is potentially a problem as a linear combination has been generated. Again, however, this is in general only true in the case of $V=(X'X)^{-1}\sigma^2$. For more complex forms of V, then the convolution of variables cannot be unpicked.

In summary then, variance-covariance matrices appear to be safe unless the model is simple unweighted OLS.

4.4 Herfindahl indices

The Herfindahl index reflects the dominance of one firm in an industry, as measured by turnover, employment etc:

$$H = \sum_i s_i^2 \qquad s_i = x_i / \sum_i x_i$$

On the face of it, this seems a safe output. As long as there are more than two firms in the market, individual values cannot be ascertained.

However, the use of the quadratic term causes a problem: unless the second largest firm is of a significant size, √H is a good approximate of the largest firm's share. The difficulty for SDC assessors is that the goodness of this approximation depends upon the relative sizes of the firms and the size of the tail. Table 2 illustrates this, with six sets of simulated values for the share of the two largest firms (S1 and S2) and the 'tail':

S1	***S2***	***S3-S50***	***H***	***S1-S2***	***Closeness of √H***
27%	1.5%	1.5%	.08	26%	8%
32%	20.0%	1.0%	.15	12%	20%
37%	15.0%	1.0%	.16	22%	10%
S1	***S2***	***S3-S10***	***H***	***S1-S2***	***Closeness of √H***
30%	10.0%	7.5%	.15	20%	27%
37%	15.0%	6.0%	.19	22%	17%
56%	40.0%	0.5%	.47	16%	23%

Table 2 H as an approximation to S1, the largest firm's share

Table 2 shows a range of values for the two largest firms and other firms in an industry, with 10 or 50 firms in the industry. There is not a simple relationship. Moreover, the last entry, which is safe in terms of the value of approximation of the largest value through √H, would usually fail a dominance test.

Therefore, although H is likely to be a safe statistic, it is difficult to state this categorically just on the value of H. The VML therefore allows Herfindahl indexes as long as the researcher demonstrates that

- there are more than two observations
- √H exceeds the largest value by a given percentage
- the dominance criterion is met

This is a pragmatic state of affairs. But it is not ideal: although these additional conditions do guarantee the safety of H, it requires three more pieces of information for researchers to provide and SDDC staff to check.

5 Conclusion

This paper has fleshed out some of the ideas in Ritchie(2007) about how to combine security, consistency and efficiency in a practical SDDC system. The examples here have demonstrated how a relative transparent assessment method can be applied to classes of output.

Many of the ideas here are already implicit in the SDDC manuals produced by NSIs; to some extent, the key purpose of this paper is to stimulate the development of a common framework for evaluating SDDC approaches. In the light of ongoing developments in creating RDC standards, it is intended that this approach be a step forward towards giving research outputs a 'risk rating', with the advantages that would give for establishing greater co-operation and transparency in RDC design.

Acknowledgements

I am grateful to Rhys Davies, Paul Allin and Philip Lowthian for comments.

References

Enright, J., McDonald, S., Corscadden, L., Jewell, E., O'Sullivan, J., Zeng, I., Nair, B., and Bentley, A. (2006) *Confidentiality Best Practices Manual*. Mimeo: Statistics New Zealand

ONS (2007) *Disclosure Control Standard for Business Surveys*. Mimeo: Office for National Statistics

Ritchie, F (2006) *Disclosure Control of Analytical Outputs*. Mimeo: Office for National Statistics

Ritchie, F (2007) *Statistical Disclosure Control in a Research Environment*. Mimeo: Office for National Statistics

VML (2007) *VML Default SDDC Methods*. Mimeo: Office for National Statistics