UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
Expert Meeting on Statistical Data Confidentiality
26-28 September 2023, Wiesbaden

Towards a comprehensive theory and practice of output SDC

Ben Derrick(University of the West of England, UK)
Elizabeth Green(University of the West of England, UK)
Felix Ritchie(University of the West of England, UK)
Paul White(University of the West of England, UK)

e-mail: felix.ritchie@uwe.ac.uk

Abstract

In 2000, the statistical disclosure control of outputs (OSDC) was largely limited to models of table protection developed by and intended for national statistical institutes (NSIs), as a particular branch of general SDC theory. However, in this century OSDC as a field of enquiry has expanded significantly, reflecting the important of secure research environments run by NSIs and others. OSDC is still a relatively under-developed field compared to SDC for tables or microdata. There are a small number of practitioner guides, and some theoretical articles, but this is a diffuse literature.
In the UK, a consortium of universities and data providers is collaborating to provide an integrated analysis of output checking including
- Key theoretical and operational concepts (eg safe statistics, principles-based OSDC)
- A comprehensive listing of statistics, associated risks, and mitigation measures as well as various practical element to support output checking.
A key element of this is a theory-driven classification which enables us to have that comprehensive listing whilst still limiting the dimensionality of OSDC guidelines to a manageable number of rules. This paper explains this model and how it has been co-developed with RDCs and others, and considers whether this provides a sustainable model for future development of the OSDC field.

1 Introduction

Increasingly social scientists are making use of confidential data for research. This has accelerated in the 21st century with the growth of secure environments, referred to as 'safe havens', 'secure data centres', 'research data centres', 'trusted research environments' (TREs) and similar names. These TREs provide standardised secure access to a range of sensitive datasets for research purposes. In OECD countries these are now common as part of the portfolio of research data services offered by National Statistics Institutes (NSIs), and academic groups are also adopting them.

TREs have introduced one substantial change to the way social scientists work. When working with confidential data, researchers are generally unaware of the potential disclosure risk in statistical outputs, as this is not covered in research methods courses (Derrick et al, 2022). However, TREs generally require researchers to submit outputs for a confidentiality review before release (Green et al, 2021). The efficiency of this process relies substantially on the researchers being aware of confidentiality risks and actively aiming to produce non-disclosive outputs (Alves and Ritchie, 2019). Hence, most TREs (Green et al, 2021) provide researchers with some training and/or guidelines in output statistical disclosure control (OSDC). Some organisations that allow downloads have also provided OSDC guidelines eg Eurostat (2015).

The practice of output checking, and the training of researchers and checkers, lags considerably behind other areas of confidential data protection, such as source data anonymisation. For many years, OSDC was limited to models of table protection (frequencies and magnitudes) developed by and intended for national statistical institutes (NSIs). In this century OSDC as a field of enquiry has expanded significantly, largely as a result of the growth of TREs and the need to cover the much wider range of outputs generated by researchers. Nevertheless, general OSDC is still a relatively under-developed field compared to SDC for tables or microdata.

A part of the problem is that the conceptual framework for generalised OSDC is lacking. There are a small number of practitioner guides, and a few theoretical articles, but this is a sparse literature. However, that literature does contains the seeds for a new overarching framework; in particular, the realisation that statistics could be grouped to minimise the need for rules covering every potential output.

In 2023 the UK academic funding council UKRI funded the project SACRO (Semi-automated checking of research outputs; see Green et al, 2023a) to deliver a general-purpose toolkit for automating output checking processes, based on the Eurostat funded pilot ACRO (Green, Ritchie and Smith 2020 and 2021). As part of the project, the team undertook to provide a comprehensive review of SDC theory, integrated with practical guidelines. A key part of the project was to formalise the use of classifications ('statbarns') and push the concept to its limit to minimise the dimensionality problem.

This paper describes the statbarn concept, how it was operationalised, how it simplifies disclosure control processes (both automatic and manual). As of July 2023, this is still a work in progress, so we review the current status and highlight areas where research needs to be done.

2 Generalised OSDC development[1]

Statistical disclosure control (SDC, sometimes called statistical disclosure limitation) is the practice of using statistical analysis to ensure that the use of confidential or sensitive data does not breach the privacy of the data subjects. SDC can be split into 'input SDC' (removing identifying information from the data before analysis is carried out) and 'output SDC' (checking that statistical aggregates do not reveal information).

[1] This short review is based on our own understanding and experience in the last two decades. We would very much appreciate comments from colleagues working in this area as to the accuracy of our representation.

Input SDC is a very well-established process. It has a large and stable literature, a large evidence base of the efficacy of different measures in different circumstances, and software tools implementing these to de-identify datasets. Research methods courses rarely teach formal de-identification, but researchers are usually given some basic guidance on broad principles.

In contrast, OSDC is a largely unknown quantity. Until 2000, 'output SDC' (had the term been coined then) would have been seen as the need to protect frequency and magnitude tables from inadvertent disclosure. This field had seen some study, and there was a relatively well-established literature, but it remained a specialist area, even for statisticians. We are not aware of research methods courses, then or now, that teach this as a matter of course, with one exception.

The exception is courses in the production of official statistics, which do cover OSDC for tables. Until recently, SDC was very heavily influenced by the needs of national statistics institutes (NSIs), who produce statistical tables and, increasingly, microdata for secondary analysis. These organisations promoted research into relevant SDC, which explains the overwhelming focus on tables for OSDC. The first OSDC papers not focusing on tables appear to be Reznek (2004), Reznek and Riggs (2005) and Corscadden et al (2006), both tacking specific problems.

In 2003 the TRE at the UK Office for National Statistics was set up, and it was run by social science researchers rather than the teams producing official statistics.. The ONS team realised that (a) the literature on tabular OSDC was of limited value in research environments, and (b) the vast majority of research outputs had no guidance at all. As a result, the team began developing guidelines with a research focus. This included an analysis of the principles behind output SDC for research (Ritchie, 2007), and the first statement of 'safe statistics' (Ritchie, 2008).

The concept of 'safe statistics' is key for efficient processing of research outputs. It recognises that certain types of output have no meaningful disclosure risk in any reasonable use. For example, the regression coefficients cannot by themselves reveal an individual value, nor can they be differenced to reveal individual values, nor are they affected by special cases such as single observations in a category (Ritchie, 2019). Of course, it is possible to construct special cases such that the regression is informative about individuals, but these have no meaningful research purpose. For all reasonable purposes, regressions coefficients are non-informative about individuals in all cases[2], and therefore they do not need to undergo output checking.

Ritchie (2016) proposed a method for classifying outputs as safe or unsafe:

- Does the statistic itself pose a risk in the case of low numbers, extreme values or something else which is a legitimate value?
- If the statistic is compared to another with one more observation, does any differencing risk arise?
- Are there are any other reasonable risks to disclosure, specific to this statistic?

If the answer to all three of these is 'no' then the statistic is classified as 'safe'. The innovation in Ritchie (2008) was that the classification should be based upon the *mathematical* characteristics of the statistic, not the statistical ones; in other words, a 'safe' statistic should be safe irrespective of the data it is calculated on.

The ONS guidelines formed the basis for Brandt et al (2010; subsequently re-released, with minor revisions, as Bond et al, 2016). This Eurostat-sponsored project (complementing a second piece on 'traditional' SDC; Hundepool et al, 2010) aimed to provide the first comprehensive guide for researchers and output checkers. The guide covered broad theory, including a discussion of safe statistics; guidelines and 'rules' on specific statistics, grouped into similar types; and suggestions for operationalising good practice, including training. Brandt et al (2010) has been the basis for many of the practice manuals now being produced by NSIs and others for TRE users.

Despite its influence, Brandt et al (2010) has some significant limitations. The most obvious is that the list of statistics covered is not comprehensive but selective, neglecting the interests of the report committee. Thus, it

[2] There are basic rules that can be checked to make sure that the regression is a genuine regression (sufficient degrees of freedom to be clear this is not an equation, regression must not be saturated to ensure this is an estimate and not a table masquerading as a regression) but in genuine situations we would not expect these conditions to occur.

is strong on the measures used by social scientists but has significant gaps relating to health research, for example. The second limitation is that the recommendations are presented 'as is' with little in the way of explanation as to why this came about. A third limitation is that the report is very laconic, offering rules but very little in the way of practical interpretation for researchers or output checkers. Subsequent manuals based on the guide have managed to address some of these; for example, the popular SDAP manual (Griffiths et al, 2019) has both a wider range of statistics, and a commentary for output checkers on how to usefully assess the output.

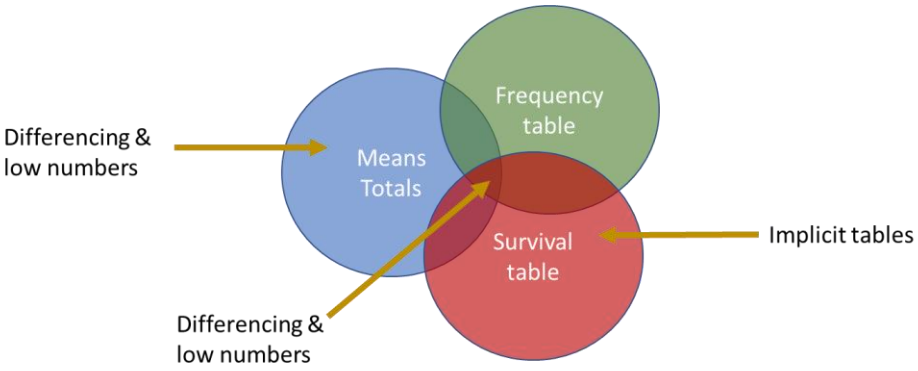
However, the major limitation of Brandt et al (2010) is that there is no overall integrating conceptual framework. The guide reduces the range of rules somewhat by grouping statistics, but these are as likely to be on whether they are commonly put together, rather than on their disclosure characteristics. Moreover, the structure of the guide implies that any additional statistics will need to have their own rules added, rather than being seen as variations on existing ones. Other manuals follow this (implicit) approach as well, listing outputs and associated rules as if they were separate entities. The implications of safe statistics and the grouping approach used in Brandt et al (2010) have not been followed through. We consider this now.

3 Conceptual foundations of an integrated approach

Analysts use a great range of statistical techniques in their models. Devising statistical rules for all of these separately is not feasible. However, it is possible to combine statistics into groups based not on statistical relation but on *common disclosure risks and solutions*. For example:

- means and totals are identical in terms of the disclosure risk for all practical purposes
- means and frequencies generate the same risks of low numbers and potential for differencing
- means have the potential for dominance
- survival tables are frequencies but they also generate an implicit secondary table

So a grouping would put means, totals, frequency tables and survival tables into three different disclosure groups:



Everything in the groups should have the same risks and solutions. For example, suppression, rounding or noise addition are valid solutions to disclosure risks in both frequency and survival tables, but on the latter they need to be implemented in a different way to allow for the monotonic relationship between cells.

The advantages of this approach are both statistical and operational:

- Fewer rules/cases for researchers and output checkers to learn
- More consistent treatment of outputs
- Clearer distinctions between outputs
- Easier to develop the theoretical basis for any guidance
- Easier to update guidance when it changes (which it does)
- Adding new statistics is now a case of 'what category does it fall into?' rather than 'what rules are needed?'
- Output checker (and researcher) training can focus on the risky classes rather than trying to cover all cases

Because classification is used in this field in many different ways, we refer to the groupings as 'statistical barns' or 'statbarns'[3].

The real value of this comes from finding that, in terms of disclosure characteristics, the minimum number of statbarns is fairly small. To a researcher, estimation of a hazard model bears little analytical relation to a quantile regression; but they pose the same disclosure risks: that is, no meaningful risk in any reasonable use, and so the only test needed is to make sure that this a genuine research use. In the case of estimated models, the tests are always

- Are there sufficient residual degrees of freedom (ie making sure this a model not an equation)?
- Is the model saturated (explanatory factors all categorical and all fully interacted ie making sure this is not a table masquerading as an estimate)?

And just like that, a large and essential part of research output is consigned to the box 'nothing to see here'.

4 The SACRO classification model

As it currently stands, the SACRO models contains fourteen statbarns:

	Barn	Example	Class	Status
1	Frequencies	Frequency tables	Unsafe	Very well understood
2	Statistical hypothesis tests	t-stats, p-stats, f-stats	Safe	Provisional
3	Correlation coefficients	Regression coefficients	Safe	Confirmed
4	Position	Median, quartiles, min, max	Unsafe	Provisional
5	Shape	s.d., skewness, kurtosis	Safe	Provisional
6	Linear aggregations	Means, totals	Unsafe	Very well understood
7	Mode	n/a	Safe	Confirmed
8	Smooth distributions	Kernel density functions	Safe	Provisional
9	Concentration ratios	Herfindahl index	Safe	Provisional
10	Calculated ratios	Odds & risk ratios	Unsafe	Provisional
11	Implicit tables	Hazard/survival tables	Unsafe	Provisional
12	Linked/multi-level tables	Nested categorical data	?	No knowledge
13	Clusters	Cluster analysis	?	No knowledge
14	Gini/lorenz curves	n/a	?	No knowledge

It is clear that some of these statbarns cover a very large number of cases ('correlation coefficients' cover linear and non-linear regression, ANOVA, ANCOVA, pairwise correlation etc). In contrast, the disclosure risks of the mode are unlike any other statistic, and so it merits its own class. This shows the importance of identifying exactly what are the disclosure characteristics of a particular statistic.

The act of creating the list is itself a useful exercise, forcing one to consider what are the meaningful differences. For example, mean and median are often grouped together in OSDC guidelines, but they have quite different characteristics. On the other hand, maxima and minima are often dealt with on their own but they can be considered as a special case of percentiles. This means that we no longer need separately rules for 'structural' end points (such as 0% or 100% in a proportion variable) but can apply general percentile rules.

This list is likely to undergo change over time. Even in the development process, the list changed as more statistics were deemed to be of the same type, and others demand a new type. The process of identifying risks

[3] The term originally came from an analogy with a farmer trying to organise her livestock, but as a neologism it has the advantage of being unambiguous

and defining OSDC guidelines for each class is crucial, as this is usually the point at which it becomes clear whether a new type is needed or not. It may also be the case that trying to identify a minimal set is counter-productive. As noted, formally maxima/minima can be treated as percentiles; but in terms of communication of risk to researchers, it may be sensible to separate them again. Finally, we have created some categories as, at the moment, we don't have enough information to be comfortable that they fit an existing category. Category 12 "linked/multiple tables" is an example – it seems like these should be covered by frequency tables, but we suspect there are nuances which need to be explored, and so creating it as a separate category show the need for more understanding.

The coverage of OSDC theory is decidedly patchy. The 'status' column has four values:

Very well understood	This disclosure issues, things to be checked and protection mechanisms have been comprehensively studied and there is a consensus
Confirmed	These have not been so well studied (conclusions rest on one or two papers) but we are confident that the conclusions and guidance are robust, well-founded and comprehensive
Provisional	We have confidence in our conclusions but this is based on extrapolation from other types, and from our own understanding; there is substantial further work to be done (for example, on the impact of extreme values) before the classification can be confirmed
No knowledge	While we may have suspicion of how these should be seen, basic analysis has not been carried out

At present, the focus is to get the 'provisional' status raised to 'confirmed'.

The list above is provisional and was devised by the SACRO team based at the University of the West of England, Bristol. SACRO's network of output checkers was consulted as to whether this was a sensible approach in general; the response was positive, but expected: earlier evidence-gathering sessions had already indicated a desire for simplification of the current OSDC landscape. The initial categories seemed both sensible and comprehensive, although these are likely to be modified as they develop in practice.

Of more concern to the output checkers was how they (and researchers) would easily check the guidelines for statistics. This is achieved by a look-up table, linking statistics to the appropriate statbarn, from which the corresponding checks, problems and solutions could be found:

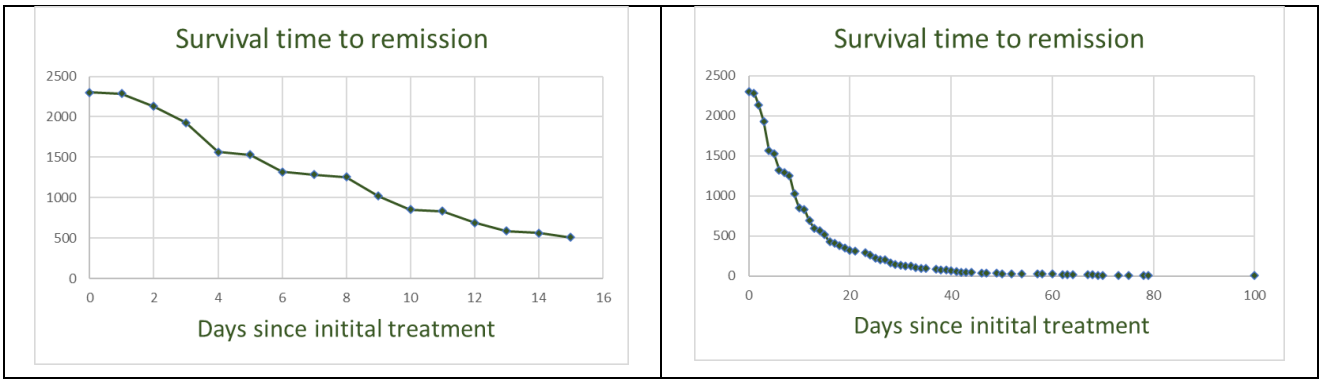
Statistic	Class	Family	Family description	Checks	Solutions (if requ	Safe or Unsafe
Magnitude table	6	Linear aggregations	This class covers sums, counts of observations, mean etc. These are stats that provide a snapshot of the data's characteristics that don't relate to any one data point and instead are calculated with many or all of the points in a data set.	Potential problems: Low counts Differencing Dominance	Solutions: Suppression Rounding Noise Outlier removal	Unsafe (Medium)
Histogram	1	Frequencies	This class covers frequencies ie counts of things, either in tables (most common), in certain graphs such as histograms or bar charts, or single as in a description of the number of survey participants	Potential problems: Low counts Differencing Class disclosure	Solutions: Suppression Rounding Noise	Unsafe (High)
Frequencies	1	Frequencies	This class covers frequencies ie counts of things, either in tables (most common), in certain graphs such as histograms or bar charts, or single as in a description of the number of survey participants	Potential problems: Low counts Differencing Class disclosure	Solutions: Suppression Rounding Noise	Unsafe (High)
Linear regression coefficients	3	Correlation coefficients	Correlation coefficients are statistical measures that quantify the relationship between two or more variables. This includes measures such as Pearson's r, Spearmans's rank correlation coefficient (ρ) and Kendall's rank correlation coefficient (τ).	Potential problems: Low d.o.fs Saturated models	Solutions: No meaningful mitigation	safe
Mean	6	Linear aggregations	This class covers sums, counts of observations, mean etc. These are stats that provide a snapshot of the data's characteristics that don't relate to any one data point and instead are calculated with many or all of the points in a data set.	Potential problems: Low counts Differencing Dominance	Solutions: Suppression Rounding Noise Outlier removal	Unsafe (Medium)
R-squared	2	Statistical hypothesis tests	Statistical tests are used to make inferences and observe differences in data. These involve a wide range of tests including T-tests (tests of difference between groups), P-values (tests of probability observing a value or more extreme value), F-tests (analysis of variance in more than	Potential problems: Low d.o.fs	Solutions: No meaningful mitigation	safe

This will be created as a searchable file, but the output tools being developed by the SACRO project (Green et al, 2023) intend to incorporate this in the user front end. Researchers and output checkers should be able to click on a link to see more information about the output, drawn from the statbarn classification. In the initial project this will only include basic data such as that shown above, but in future it may be useful to expand the information on each classification. This highlights the advantage of classification: the SACRO coders only need to know the statbarn code and then can draw all this information from a finite set of outputs.

5 Graphical outputs

Graphs do not present new issues. In theory, every graph can be represented as a table in some way, and so the above rules could be applied. To take an obvious example, a pie chart or a histogram are clearly just one-way tabulations, whereas a waterfall graph is a two-way table. As a counter example, a kernel density estimate could be represented as a mathematical form, but in practice is almost always show graphically. In practice, we need separate rules because (a) the quantity of information differs, and (b) precision is likely to be lower in a graph.

Consider the Kaplan-Meier graph, which is simply a survival table re-presented, usually in proportional form (we assume that counts and proportions are equally disclosive as the total from which the proportion is calculated is likely to be published somewhere). Survival tables are classed as 'unsafe but very low risk' because, even in the case of a unit being identified, the personal information content in the survival table is negligible. Griffiths et al (2019) suggest that the underlying survival table should be supplied along with the graph, but this can cause more problems:



In the left-hand graph, the source table would have 15 steps and be checkable by a human. But that table would have precise numbers easily readable, whereas getting the exact figures from the graph depends on the way that the image was produced (and even then, some laborious analysis). In the right-hand diagram, a survival table with 100 rows in it is much harder to assess accurately, whereas identifying individual data points from the image has become harder.

The above graphs are presented as numbers. Formally Kaplan-Meier graphs should show the survival rate rather than numbers (ie 0%-100%). In theory this makes graphs slightly more disclosive than the survival table: tables are likely to limit the number of decimal points shown, whereas the full decimal value may be used in creating the graph points.

Given the low information content in any data point, even if relating to one person, producing survival tables alongside graphs seems to increase risk rather than reducing it. Hence, the current guidance from SACRO is that Kaplan-Meier graphs should be released subject to the researcher confirming that each step and the end point meets thresholds

The objective for the SACRO guide is that it will show the statbarns that each graph falls into (which in itself might lead to additional statbarns being defined, as in the case of kernel densities), but will concentrate on the practical assessment; in particular, how graphical representation adjusts the perspective on what is discoverable

from an output. Again, this is the value of the grouping – we can see what we should be looking for in the output.

6 Conclusion

As the use of confidential microdata for research rises, so does the need for efficient and effective OSDC. OSDC for research has made considerable advances in this century, but guidelines have tended to develop on an ad hoc basis as new statistical queries are raised. The strategic approach being taken by SACRO and described in this paper attempts to provide a longer-term solution to the problem.

The idea of grouping statistics was first raised in Ritchie (2008) partly as a response to proliferation of OSDC rules emerging from research use of the ONS TRE. While the safe-unsafe classification is crude, it highlights how applying a structure can significantly improve operational as well as statistical outcomes. Classification also changes the way we think about outputs. When Brandt et al (2010) was written, the implication is that additional statistics would require new rules. In the statbarn model, risk assessment for a new statistic should be a matter of deciding whether it fits into an existing category. If it does, then no further work is needed. If not, then a new category is added, but this should be a rare event.

The statbarn approach is part of the development of a wider set of operational guidelines aiming to bring consistency between theory and practice to output checking.

7 References

Alves, K., & Ritchie, F. (2020). Runners, repeaters, strangers and aliens: Operationalising efficient output disclosure control. Statistical Journal of the IAOS, 36(4), 1281-1293.

Brandt M., Franconi L., Guerke C., Hundepool A., Lucarelli M., Mol J., Ritchie F., Seri G. and Welpton R. (2010), Guidelines for the checking of output based on microdata research, Final report of ESSnet sub-group on output SDC

Bond S., Brandt M., de Wolf P-P (2015) Guidelines for Output Checking. Eurostat.

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

Derrick, B., Green, E., Ritchie, F., Smith J. & White, P. (2022, April). Disclosure protection: a systemic gap in statistical training?. Paper presented at Scottish Economic Society Annual Conference 2022: Special session 'Protecting confidentiality in social science research outputs', Glasgow

Eurostat (2015) Self-study material for Microdata users. Eurostat.

Green, E., Ritchie, F., Tava, F., Ashford, W., & Ferrer Breda, P. (2021, July). The present and future of confidential microdata access: Post-workshop report.

Green, E., Ritchie, F., & Smith, J. (2020). Understanding output checking. Luxembourg: European Commission (Eurostat - Methodology Directorate)

Green, E., Ritchie, F., & Smith, J. (2021). Automatic Checking of Research Outputs (ACRO): A tool for dynamic disclosure checks. ESS Statistical Working Papers, 2021 Edition

Griffiths E., Greci C., Kotrotsios Y., Parker S., Scott J., Welpton R., Wolters A. and Woods C. (2019) Handbook on Statistical Disclosure Control for Outputs. Safe Data Access Professionals Working Group.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Schulte Nord-holt, E., Seri, G. and De Wolf, P. (2010). Handbook on Statistical Disclosure Control. ESSNet SDC.

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

Reznek A. and Riggs T. (2005) "Disclosure Risks in Releasing Output Based on Regression Residuals" ASA 2004 Proceedings, Section on Government Statistics and Section on Social Statistics pp1397-1404

Ritchie F. (2007) Statistical disclosure control in a research environment, mimeo, Office for National Statistics; available as WISERD Data Resources Paper No. 6

Ritchie F. (2008) “Disclosure detection in research environments in practice”, in Work session on statistical data confidentiality 2007; Eurostat; pp399-406

Ritchie, F. (2014). Operationalising ‘safe statistics’: The case of linear regression. UWE Working Papers in Economics no 14/10. Bristol

Ritchie, F. (2019). Analyzing the disclosure risk of regression coefficients. Transactions on data privacy, 12(2), 145-173

Smith J., Preen R., Ritchie F., Green E., Stokes P., & Bacon S. (2023) SACRO: Semi-Automated Checking Of Research Outputs. Paper prepared for the 2023 UNECE/Eurostat Workshop on Statistical Data Confidentiality,September.