Administrative Data Research Network

An ESRC Data Investment

Ensuring the confidentiality of statistical outputs from the ADRN

ADRN Publication

Authors: Philip Lowthian and Felix Ritchie

Series editors: Elaine Mackey & Mark Elliot

Better Knowledge Better Society

Ensuring the confidentiality of statistical outputs from the ADRN

Written by Philip Lowthian and Felix Ritchie

Edited by Elaine Mackey and Mark Elliot

ADRN Publication - May 2017

This guide, 'Ensuring the confidentiality of statistical outputs from the ADRN', was written by Philip Lowthian and Felix Ritchie, edited by Elaine Mackey and Mark Elliot (Administrative Data Service) and published by the Administrative Data Research Network.

Contents

List of tables

1. Introduction

1.1. Description of the ADRN

The Administrative Data Research Network (ADRN) is a UK-wide partnership between academia, government departments and agencies, national statistical authorities, funders and the wider research community to facilitate new economic and social research based on routinely collected government administrative data.

The ADRN is establishing a new legal, secure and efficient pathway for the research community to access de-identified linked administrative datasets.

This will potentially benefit our society by providing a greater evidence base to inform policy.

The ADRN consists of:

- four Administrative Data Research Centres (ADRCs):
  - ADRC England: led by the University of Southampton
  - ADRC Northern Ireland: led by Queen's University Belfast
  - ADRC Scotland: led by the University of Edinburgh
  - ADRC Wales: led by Swansea University
- an overarching Administrative Data Service, which is the co-ordinating body of the Network
- administrative data owners
- the Economic and Social Research Council (the funding body)
- the UK Statistics Authority (chairing the ADRN Board)

The ADRN has commissioned this guide on statistical disclosure control to support the development of knowledge and skills in the subject area.

1.2. Protecting confidentiality in research outputs

The ADRN has a proven safe mechanism to manage research use and outputs of confidential data. However, because the data that the researchers use are sensitive, it is possible that the publication of statistical analyses could inadvertently lead to the disclosure of sensitive information.

Many years of experience in managing research outputs, with very few problems, shows that there is a negligible risk of such a disclosure – but that risk does exist. Outputs are checked by researchers and trained output checkers before they are published. This is called 'output statistical disclosure control', or OSDC. As a result of this, statistical disclosure risk is reduced.

The ADRN generally uses the accepted best practice for controlled research environments in the form of 'principles-based OSDC' or PBOSDC (except ADRC Northern Ireland; see section 2.3.4). This was designed specifically for assessing the disclosure risk that may be associated with research outputs to allow for the complex methods that researchers employ. As OSDC generally goes hand in hand with good analytical practice (for example, having many observations tends to both lower disclosure risk and produce results of greater statistical value), this is a check rather than a restriction on publication. A key element of PBOSDC is that researchers are trained to be fully involved in identifying problematic outputs themselves, rather than relying only on the people who check outputs. This allows the researchers to think in advance about the disclosure risks that their analyses might present.

This guide is intended for

- researchers who need to learn what OSDC is and how that relates to their work
- data owners who are concerned about inadvertent disclosure of results
- the wider public who are interested in how the ADRN's security model (the 'Five Safes') works to encourage research while protecting confidentiality

This is not a comprehensive guide to OSDC: researchers who use the ADRN and therefore have access to the source data will go through an extensive training programme. The aim of this guide is to provide an overview.

2. Statistical disclosure control of outputs

2.1. SDC and output-based SDC

2.1.1. What is SDC?

The ADRN exists to allow researchers to access linked de-identified administrative data which are combined for specific projects for a defined period of time. Many of these contain detailed information which for ethical or legal reasons cannot be released into the public domain. Statistical analyses on the data do not generally, of themselves, produce any disclosure risk, but disclosive outputs may inadvertently be produced. Such outputs, if detailed enough, might allow someone reading the research to discover confidential information about an individual, household or business. This risk is often difficult to measure and depends on a range of factors. The question becomes one of managing this risk: how can it be minimised while also maintaining the usefulness of the output? Statistical disclosure control (SDC) is the approach used in managing the risk. SDC uses statistical techniques to prevent individuals, households or enterprises being identified in published information. SDC aims to make sure data are not disclosive and, importantly, would not be perceived as being disclosive.

The ultimate aim of SDC is to maximise the usefulness of the outputs while minimising the risk of disclosure (or the perception of disclosure). This level of risk is never zero, so the aim is to reduce, to an acceptably low level, the possibility that confidential information is released. The definition of an acceptably low level depends on factors such as the age and sensitivity of the output.

2.1.2. What is output-based SDC?

SDC can be applied at different points in the process of producing outputs. 'Input SDC' is the technique of protecting microdata (the individual records held in the data) before researchers start their work. These data can be protected by methods such as removing direct identifiers and making sure combinations of specific variables do not identify (unique) records in the data.

ADRN datasets do have some input SDC applied: researchers never get to see direct identifiers such as name and address. But making these data fully non-disclosive destroys their research value; this defeats the purpose of the project. Instead, ADRN researchers go through extensive certification processes and use secure and tightly governed facilities.

The ADRN then allows unrestricted research on confidential data and checks the statistical results that researchers want to publish. This is called 'output-based SDC', or OSDC.

OSDC is concerned with statistical output such as tables for release into the public domain either directly or via research papers. Outputs are created from confidential data, so a check on the outputs is required. The ADRN can use a variety of techniques to make sure the outputs maintain confidentiality.

In the following examples, we use simple tables to show how disclosive outputs may inadvertently occur in both frequency and magnitude tables.

Example 1 – a frequency table

Frequency tables are straightforward tables of counts (numbers of observations).

Male					
	Good health	**Fair health**	**Bad health**	**Very bad health**	**Total**
White	6	7	3	2	18
Mixed	2	2	3	1	8
Asian	1	0	5	0	6
Black	0	5	0	0	5
Other	0	0	0	1	1
Total	9	14	11	4	38

Table 1: Potential disclosure problems – frequency tables

In a frequency table, there are a number of potential disclosure scenarios, all of which are related to either low counts or unusual distribution of counts.

In example 1:

- Identification disclosure is the act of identifying a specific person or unit in the data. Counts of '1' in the table allow individuals in the dataset to be identified[1]. Individual Attribute disclosure is the inference of information about a unit that is in the table given partial information you already have about that unit. An example of attribute disclosure is shown in the row shaded green. Knowing there is somebody in the 'Other' category in the data discloses this person's health.
- Group (attribute) disclosure occurs when all respondents who have some feature also have some other feature. Group disclosure is shown in the row shaded blue. All people in the 'black' category have fair health.
- Within group (attribute) disclosure occurs when there is one respondent in a single category with all other respondents in a different category. Within group disclosure is shown in the row shaded yellow. The Asian individual in 'Good Health' would know that all others in his group have 'Bad Health'.

Distributions of counts such as that shown in Table 1 do not necessarily prove that the table is risky. Both the data provider and data user may have additional information to help determine if SDC ought to be applied.

Example 2 – a magnitude table

In magnitude tables, each cell value represents the sum (or average) of a value across all respondents belonging to that cell.

Disclosure typically occurs when cells with dominating contributors (for example one large supermarket and a number of corner shops) or cells with a small number of contributors (for example, oil companies) are present. The most likely source of risk is one member of the cell attempting to discover a value relating to a competitor in the same cell. (Frequencies are in brackets in the table shown.)

Local Authority				
Manufacturing turnover in £1000s (frequency)				
	LA1	LA2	LA3	Total
Food products	58 (8)	13 (2)	74 (7)	145 (17)
Textiles	715 (15)	164 (12)	648 (260)	1527 (287)
Paper and paper products	98 (4)	158 (22)	47 (12)	303 (38)
Petrochemicals	18 (7)	50 (9)	64 (1)	132 (17)
Total	889 (34)	385 (45)	833 (280)	2107 (359)

Table 2: Potential disclosure problems – magnitude tables

1 A broad definition of identification disclosure includes self-identification. However with the possible exception of the most sensitive of outputs (e.g. abortion statistics) this is not usually considered to be a major disclosure control issue.

Cells with frequencies of 1 or 2 are disclosive as turnover values for each company in the cell can be worked out exactly. This can be seen in the cells shaded blue.

Cells with dominating contributors can be found by looking at the values which make up the cell. Suppose the cell shaded green, with four contributors giving a total of £98,000, was made up of these values:

Company 1: 48
Company 2: 45
Company 3: 3
Company 4: 2

Company 2 could subtract their contribution from the published total and obtain an estimate of the turnover of Company 1 to within 10% of the true value. This is potentially disclosive and would require further investigation.

2.1.3. Actual versus potential disclosure

Table 1 gives a number of examples of potential disclosure for frequency tables. However, this should not be confused with actual disclosure. Statistical disclosure is scenario-based. This can be summarised by collating the information known about an individual or other unit in an output and seeing what other information can be discovered.

Table 3 (below) considers the row relating to the black males in Table 1. Suppose that the local paper receives a letter about the survey from John Smith identifying himself as one of the black male respondents; the paper publishes Mr Smith's name and address. If Mr Smith is in the table, then we now know his health status; but although he self-identifies as being in the survey, can you reasonably infer with confidence that he is in the table? He might or might not be one of the five black males listed in Table 1, depending on how many black males were sampled altogether, and why the sub-sample in Table 1 was included.

We can consider whether Table 1 is disclosive under several response knowledge scenarios.

Scenario	Number of black male participants in the survey	Who is in Table 1	Disclosive?
1	5	All of the survey sample is present	Probably
2	15	All the survey participants, who are in age band 20-29	Probably
3	15	All the survey participants, who provided usable income information	Probably not
4	15	All survey participants, who have recently visited hospital	Possibly

Table 3: Likelihood of actual disclosure – frequency tables

These scenarios look at different levels of response knowledge and ask whether this knowledge would allow someone to find the respondent in the data with a high level of confidence.

Scenario 1: Mr Smith is definitely one of the five included in the table, and so this is disclosive. You now know his health condition.

Scenario 2: The initial sample is larger, with Table 1 being a subsample with a known age range. Someone who knows Mr Smith might reasonably be expected to know his age to the nearest decade. If you know Mr Smith is in the survey and you know his age range, it is highly likely you now know his health condition.

Scenario 3: The sample is the same size as Scenario 2. The selection criterion is 'is the income information of good enough quality for analysis?' There is doubt as to whether this is disclosive. Even if Mr Smith thinks he supplied income information, this does not mean that the researcher agreed it was accurate. On the other hand, colleagues of Mr Smith might know he is good with figures and therefore likely to have provided good information.

Scenario 4: This requires more thought. Is it reasonable that you know that somebody has been in hospital 'recently'? Mr Smith's line manager might be aware of an absence from work, or a neighbour might have seen an ambulance at his house, or Mr Smith might have had an outpatient appointment that he could fit around work. These hard to define possibilities show the difficulty in determining whether identification is likely.

Table 3 only describes whether Table 1 is 'probably' or 'possibly' disclosive. There is always an element of doubt when attempting to identify an individual or household in a published table, so exact statements are usually avoided. For example, Mr Smith may be in error in stating they have been surveyed, or their details may have been transcribed incorrectly. Suppose that health, rather than being a question in the survey, is actually a value derived from other variables in the survey. In this case, the specific data is not disclosive even if the cells numbers are small, unless detailed information on how the variable is derived is published.

If the variable derivation is straightforward (such as, an income of over £100k defines a predicted health ranking of 'fair'), the data are disclosive, as you now know something more about Mr Smith. In the case of "usable income information supplied", it might be that the researcher supplies details about how the selection was done elsewhere, and so the 'probably not' becomes 'probably' when combined with this additional information.

Of course, Mr Smith might not write to the paper, but he might discuss his participation with friends or colleagues. The aim of these scenario analyses is to think about what anyone looking at your publication would reasonably be able to infer.

In short, it is almost impossible to prove that something is not disclosive. Proving that something is disclosive can be done in specific cases, but low numbers of observations are clearly not sufficient by themselves. Whether something is disclosive or not is a balance of probabilities. The aim is to achieve 'acceptably low risk' – a satisfactory balance between confidentiality and usefulness.

In frequency tables, cells with 1 or 2 contributors are usually assumed to be disclosive for the majority of outputs, as identification and/or attribute disclosure is highly likely. Rows and columns where observations are concentrated in a small number of categories are also problematic.

In magnitude tables, we need to consider how well someone could estimate a true value when one or two responses dominate the total. The following are possible guidelines.

- 5% – for not very sensitive data where an estimate relatively close to the true value is acceptable.
- 15% – for moderately sensitive data.
- 25% – for very sensitive data, a wide range of protection is required, such as business-sensitive data which must be treated especially carefully.

The choice of these values is subjective, and data providers may want special rules for particular cases or datasets. Hence, there is often a need for consultation between the researcher and the output checker.

2.2. Principles-based OSDC

2.2.1. The concept of PBOSDC

When considering how to develop guidelines for output, we need to balance the risks to individuals of being identified in a publication, and the risk to the public good of not being able to use statistical evidence; what we might call the 'confidentiality' and 'usefulness' problems.

For example, in a frequency table such as Table 1, we could consider setting a minimum number of observations in any cell, called a 'threshold', to avoid problems. Deciding what the threshold should be is difficult, as the confidentiality and usefulness problems have opposing requirements:

- confidentiality: a low threshold increases the probability of disclosive cells being published, whereas a high threshold reduces this risk considerably
- usefulness: a low threshold allows most statistical findings to be published; a high threshold is likely to mean that some findings are not published

PBOSDC recognises that this only becomes a problem if you have to obey specific rules without exception. If the threshold rule is treated as a guideline and not a rule, the confidentiality/usefulness relationship can be explored by thinking about the output in context when factors such as variable sensitivity and age of the data are considered.

For example, if the threshold is set at 0, we have essentially decided that the default assumption is that everything can be published. However, the person checking outputs may decide that in some cases the confidentiality risk is unacceptably high and ask for these to be treated as special cases and not released. Alternatively, the threshold might be set at five, which would imply a more cautious approach and more weight placed on confidentiality. The researcher might ask for tables with less than five observations in a table cell to be released for publication, but would need to show that this does not create an unacceptably high confidentiality risk.

This is why this is called 'principles-based': principles-based output checking allows flexible interpretation of the broad standards and considers each analysis in context, consistent with a set of agreed principles. The alternative approach is 'rules based' which is more suited to publications where the context is well known and consistency over time periods is important.

Why is PBOSDC best practice for research environments? Put simply, research outputs generally do not have a high confidentiality risk. For example, researchers typically use frequency tables to describe the general characteristics of the dataset; low cell counts are less likely to occur, or to be essential to the analysis. Hence, the way PBOSDC is implemented in practice is to focus on the confidentiality issue with relatively high threshold limits. Statistical agencies (and textbooks) often use three as the threshold limit; but most controlled environments in the UK using PBOSDC set ten units as the minimum, and in some cases this can be as high as thirty.

These higher limits are acceptable to researchers because of the low impact on their research. More importantly, researchers know they can request that a table with small numbers be released if they can show that there is no significant confidentiality risk. Hence, researchers know that when exceptions matter, they can have a sensible conversation with the output checker and negotiate a solution. This makes researchers happier, and evidence shows that contented researchers will also work actively to maintain data security.
The data providers are also happier. A higher limit means that they can be confident that there is a wide margin of error around regular outputs. At the same time, researchers requesting 'special cases' mean that the output checking is directed towards potentially more risky cases. Overall, security is improved, as limited resources are directed efficiently. A good analogy is with a zoo: you could spend the same amount of effort on cages for rabbits and tigers, but if public safety is your concern, it might be sensible to concentrate on getting that tiger enclosure right. (In statistical terms, the 'tigers' are tables, while the 'rabbits' are regression models.)

Finally, both researchers and output checkers benefit from a much faster approval time. Most researchers soon learn how to make sure their analyses are quickly approved and released; and with a high threshold, most outputs can be approved very quickly because the output checker knows that there is a wide margin for error. This avoids using expensive staff on unproductive activities.

2.2.2. How much risk is there in research outputs?

Unusual values, or 'outliers', may cause confidentiality problems ('is there a billionaire living in your village?'), and researchers produce a much wider range of outputs than the original data providers. For example, researchers often use graphs to give readers a feel for the data.

Output checkers are there to identify such things, and researchers accept that there will be some limitations on their output; after all, this is confidential data being manipulated. Such limitations are reduced by the PBOSDC approach: both researchers and output checkers are trained to look for ways to make difficult data acceptable for publication.

In general, however, statistical outputs produced by researchers are much lower risk than those produced by the data providers. This is for three reasons:

- Researchers choose their own samples

  Researchers often select subsamples of the data in ways which are not obvious to the reader; for example selecting on non-visible variables or characteristics, or by limiting the analysis to high-quality responses. This makes it harder to be certain who is in the data being used by researchers, even if the criteria for the subsample are published, as it requires a detailed knowledge of the data.

- Researchers transform the data

  Researchers spend a lot of time converting the data into different forms: taking logarithms, creating categories, combining multiple variables, generating growth rates, and so on. These reduce the likelihood of the original data being uncovered.

- Researchers carry out a wide range of analyses

  Researchers need access to this sensitive microdata because they are mostly interested in complicated statistical relationships. These tend to have very little disclosure risk because of the complexity of the mathematics behind the statistics, irrespective of the data. For example, many economists will use regression models; life scientists will be interested in odds ratios and survival functions. These have no effective disclosure risk as they are, in effect, summaries of processes within whole populations; they are not about individuals. In general the most valuable outputs produced by analysts fall into the category of 'negligible risk'.

When researchers do produce higher-risk outputs (for example, descriptive tables of the variables) these can be easily checked to see if the data or output has been transformed, or if the subsample is in some way guessable.

In short, research output tends to be much lower risk than the data providers would produce, simply because that need for complexity is what has driven the researchers to get access to this data in the first place.

2.3. Making PBOSDC work in the ADRN

2.3.1. Working with researchers

Mistakes can happen, and it is in everyone's interests to make sure that the risk of this is kept low. This potentially conflicts with both parties' aims of getting good research published so that the public benefit of that research can be realised. So, it is important that output checking is seen by both parties as a positive part of the research process, and the relationship between output checkers and researchers is crucial to PBOSDC working effectively.

Both parties need to understand and agree on what sorts of output are likely to be released. If a researcher continually produces unacceptable outputs which will then be rejected, both parties will be annoyed and frustrated. Equally, if a researcher is worried about rejection and so avoids producing output that might be queried, potentially useful research may be stifled.

For the system to work well, output checkers and researchers must have respect for the other's abilities. If a researcher uses a technique which the checker has not encountered before, the former should help the latter understand the technique so that disclosure risk can be assessed; without that understanding the output cannot be cleared.
A researcher keen for the release of vital results which fail the 'rule of thumb' may try to persuade the output checker to make an exception. The output checker must take the assessment of research value seriously. Equally, the researcher should be aware that the output checker will be responsible if it transpires that the output should not have been released. This collaborative approach is an important part of the training of both researchers and output checkers.

Output checkers and researchers need to be comfortable having conversations with each other about what is acceptable for release. If both understand and respect the other's constraints, the discussion can be useful and the system can work to everyone's advantage. Output checking is necessarily subjective, and the effectiveness of PBOSDC depends on how individuals approach output checking.

2.3.2. Researcher training

As mentioned above, all ADRN researchers go through a face-to-face training programme, organised as part of the application process, to help them produce outputs which are publishable with few or no restrictions. The training has been continuously developed since 2003, and is recognised around the world as best practice. The training is organised jointly with other organisations who manage access to confidential data for research, including the Office for National Statistics, HM Revenue and Customs, the UK Data Service, and the devolved administrations.

This researcher training comes in three forms:

- face-to-face training for researchers using the ADRCs, making them 'ADRN trained researchers'
- online training for researchers working on ADRN projects but who do not use the ADRCs (for example, co-authors, or PhD supervisors)
- online refresher course for 'ADRN researchers' to maintain their training.

These are not available to non-researchers. However, readers can get a feel for the training content: the Eurostat self-study guide[2] provides an online introduction covering many of the same themes as the ADRN training.

Given the importance of the relationship between the researcher and the output checker, a considerable part of the training is spent on exploring the wider concepts of data access. The training can't cover all possible situations that might arise, but it should help individuals understand how to approach a new or difficult situation.

ADRN researchers also have online follow-up training, the plan being that it should be completed on a biennial basis to maintain their 'ADRN researcher' status.

2.3.3. Ensuring consistency and flexibility across the ADRN

ADRN output checkers also go through the same training as researchers. These staff will need to see the clearance issue through the eyes of researchers, just as researchers need to see disclosure control through the eyes of the data providers. Not all researchers fully engage with the security model, so ADRN staff get additional training in difficult conversations.

The ADRN is a network, so groups of output checkers can meet to compare notes and revise training and guidance. The ADRN has also set up an expert group of SDC specialists to provide general oversight of strategy and processes, and to advise on specific problems.

2 See http://ec.europa.eu/eurostat/web/microdata/overview/self-study-material-for-microdata-users

2.3.4. Rules-based output SDC in ADRC Northern Ireland

ADRC Northern Ireland does not use PBOSDC, but instead the simpler rules-based model (RBOSDC). In this model, the 'guidelines' on thresholds, dominance and so on are treated as hard rules: release is automatic depending on whether an output meets the rules or not, and researchers cannot challenge decisions. So, there is less need to train staff or researchers in SDC, and procedures can be automated.

This approach is well-suited to statistical organisations producing tabular outputs because it provides a consistent approach across many similar outputs and production units. For example, Eurostat's current SDC training formally recommends RBOSDC. This was also the dominant model for research centres until the last decade, and is still used in some countries, although it is generally being phased out in favour of some form of PBOSDC.

Note that the same general SDC principles apply to PBOSDC and RBOSDC. The difference is in the implementation, in particular the role of context-sensitive flexibility and researcher feedback. Hence, our staff in Northern Ireland go through the same SDC training as other ADRN staff.

3. Resources

3.1. Websites

General information on the ADRN can be found at adrn.ac.uk, including other guides in this series and educational material.

The UK Data Service (www.ukdataservice.ac.uk) has an extensive range of documentation about data access, management and security.

The Five Safes website www.fivesafes.org contains guidance and further information on all aspects of confidential data management, including practical advice for output checkers and useful metaphors for non-specialists. In September 2015 the UK Data Service organised a workshop on the confidential data management using the Five Safes framework. All the presentations are available on the UKDS website at www.ukdataservice.ac.uk/news-and-events/eventsitem/?id=4058.

Statistical agencies are a source of extensive expertise in data access, management and confidentiality. In the UK, the Office for National Statistics' methodology advice page is https://www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/disclosurecontrol. This is where developments in guidance can be found.

Wider information from the Government Statistical Service is available at https://gss.civilservice.gov.uk/statistics/methodology-2/statistical-disclosure-control/.

Eurostat guidelines on microdata access at the European level are at http://ec.europa.eu/eurostat/web/microdata, and a self-study guide is at http://ec.europa.eu/eurostat/web/microdata/overview/self-study-material-for-microdata-users.

4. Summary

Experience in many countries has shown that researcher analysis is one of the safest uses of confidential data, and public benefit can be generated by allowing access to such data. The ADRN is designed to make sure the security of the data is managed according to best practice throughout the data lifecycle.

This guidance document has looked at the final part of that chain: can we make sure researchers analysing data for the public good do not accidentally disclose any confidential information in their statistical findings? **The answer is a resounding yes**.

The ADRN builds on the experience of similar facilities in the UK and around the world. It has adopted principles-based output statistical disclosure control (PBOSDC), which was designed specifically for research environments like those of the ADRN. If disclosure control is to be applied to any output, the researcher will play an active role in deciding, for example, whether to redesign the table to protect sensitive cells or to apply suppression or another disclosure control technique. There should be no surprises in either outputs presented by the researcher or decisions reached by the output checker if there is an ongoing dialogue between the two throughout the process.

A key element of PBOSDC is the researcher's engagement in both the ethos and practices of the ADRN. This is necessary for PBOSDC to work well, but more widely it encourages all parties to learn about and work with each other. This collaborative approach between the researcher, the output checker and the data provider promotes the smooth and secure operation of the ADRN.

5. Further Reading

5.1. Non-statistical papers for a general audience

The key paper on designing OSDC for research is *Statistical disclosure control in a research environment* (Ritchie, 2007. mimeo, Office for National Statistics).

Principles- versus rules-based OSDC in remote access environments (Ritchie and Elliott, 2015) discusses the rationale for the ADRN's approach to OSDC.

The importance of developing positive relationships between researchers, output checkers and data providers can be found in *Effective researcher management* (Desai and Ritchie, 2010) and Addressing the human factor (Ritchie and Welpton, 2014).

An overview of the confidentiality protection strategy used by the ADRN can be found in *Five Safes: designing data access for research* (Desai et al, 2016).

Readers interested in how SDC sits within the a wider framework of anonymisation and particularly its relationship with data protection legislation should consult the *Anonymisation Decision Making Framework* by Elliot et al (2016).

5.2. Statistical papers

A detailed review of SDC, plus some comments on OSDC, can be found in *Statistical Disclosure Control* (Hundepool et al, 2012). For a wider review of the statistical confidentiality field see Duncan et al (2011). For an up to date review of the SDC field see the biannual *Privacy in Statistical databases publication* by Domingo-Ferrer and colleagues and published by Springer.

Guidance for output checkers can be found in Brandt et al (2010). Specific rules for linear regression analysis, along with a review of the role of subjectivity in disclosure control, are covered in *Operationalising safe statistics: the case of linear regression* (Ritchie, 2014).

6. References

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* https://tinyurl.com/jr9wbh4 [accessed 28/2/2017].

DESAI T., RITCHIE F. and WELPTON R. (2016) "The Five Safes: designing data access for research", *Working papers in Economics no. 1601, University of the West of England, Bristol*. https://tinyurl.com/he7wyb5 [accessed 28/2/2017].

DESAI T. and RITCHIE F. (2010) "Effective researcher management", Centre for Economic Performance London School of Economics and Political Science. https://tinyurl.com/hkw9xx9 [accessed 28/2/2017].

DUNCAN G. T., ELLIOT M. J., and SALAZAR-GONZÁLEZ J. J. (2011) *Statistical Confidentiality*. New York: Springer.

ELLIOT M. J., MACKEY E., O'HARA K. M. and TUDOR C. (2016) *The Anonymisation Decision Making Framework*. Manchester: UKAN publications

HUNDEPOOL A., DOMINGO-FERRER J., FRANCONI L., GIESSING S., NORDHOLT E. S., SPICER K. and DE WOLF P. P. (2012) *Statistical Disclosure Control*. London: John Wiley & Sons.

RITCHIE F. and Elliot ELLIOT M. (2015) "Principles- versus rules-based output statistical disclosure control in remote access environments", *IASSIST Quarterly* v39 pp5-13 https://tinyurl.com/z7wdj8m [accessed 28/2/2017].

RITCHIE F. (2014) "Operationalising safe statistics: the case of linear regression", *Working papers in Economics no. 1410, University of the West of England, Bristol*. https://tinyurl.com/hxjbxzm [accessed 28/2/2017].

RITCHIE F. and WELPTON R. (2014) "Addressing the human factor in data access: incentive compatibility, legitimacy and cost-effectiveness in public data resources". *Working papers in Economics 1413, University of the West of England, Bristol.* https://tinyurl.com/hdfw7jg [accessed 28/2/2017].

RITCHIE F. (2011) Statistical disclosure detection and control in a research environment https://tinyurl.com/ho92ocv [accessed 28/2/2017].

7. Glossary

The following definitions are used in the document.

Confidential: Information which is not expected or allowed to be put in the public domain

Data provider: The organisation which collected the data and makes it available for research use through the ADRCs

De-identified: Microdata with direct identifiers such as name and address removed

Disclosure: The release of confidential information to unauthorised parties

Five Safes: A framework for identifying sources of risk in when accessing data (the Five Safes are Projects, People, Settings, Data, Outputs)

Frequency table: A table of counts of the number of individuals in different categories; for example, GP registrations tabulated by local authority

Identification: The association of a cell in a table or a microdata record with a member of the population

Magnitude table: A table containing totals or averages of a particular response across all individuals in the table cells; for example, average income tabulated by family size

Microdata: Record level data, each row represents responses relating to that particular individual, household or business

OSDC (Output Statistical Disclosure Control): The protection of outputs which are to be circulated outside a research environment

Output checker: Individual at one of the ADRCs who reviews outputs before release from the secure environment

PBOSDC (Principles Based Output Statistical Disclosure Control): OSDC where the emphasis is on researchers and output checkers working together to ensure that outputs are checked quickly and efficiently

Statistical Disclosure Control: The prevention of the release (and the perception of release) of information which may lead to the identification of a statistical unit or attribute associated with that unit

8. Appendix 1: A 'belt and braces' approach to confidentiality

Output checking is only one of the ways that the ADRN maintains confidentiality. The ADRN uses the 'Five Safes' framework to evaluate risks when handling confidential data.

The Five Safes are:

- safe projects: ensuring that the use of the data is lawful, ethical and creates a public benefit
- safe people: ensuring that users of the data know their responsibilities and obligations and have the tools to use data appropriately
- safe settings: making the data available in was which maintain confidentiality but do not unnecessarily restrict researchers' freedom
- safe data: removing unnecessary identifying information from data files before making them available to researchers
- safe outputs: checking outputs using PBOSDC

Hence, output checking should be seen as the last link in the chain of protection, there to guard against accidents despite all the other checks in place.

ADRN PUBLICATION
University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ
T. +44 (0) 1206 87 2976 E. help@adrn.ac.uk W. adrn.ac.uk

Funded by

E·S·R·C
ECONOMIC & SOCIAL RESEARCH COUNCIL