WP. 20
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Bilbao, Spain, 2-4 December 2009)

Topic (iii): Research data centres and virtual labs

GUIDELINES FOR THE CHECKING OF OUTPUT IN RESEARCH DATA CENTRES

Invited Paper

Prepared by Jan Mol, Statistics Netherlands

Guidelines for the checking of output in Research Data Centres

Jan Mol*

* Centre for Policy Related Statistics, Statistics Netherlands, PO box 24500, 2490 HA The Hague, The Netherlands, jmol@cbs.nl

Abstract: This paper briefly describes the (preliminary) results of an ESSnet project on guidelines for output checking. After explaining the difficulty of output checking in a Research Data Centre, two models for checking output are presented: the rules of thumb model and the principles-based model. All types of output are categorized, and for each category the rule of thumb and the underlying principles are discussed. Only a glimpse at these rules and principles is given, since the ESSnet project is not yet completely finished.

1 Preface

In this paper the preliminary results of the project 'Guidelines on output checking' will be presented. This project is carried out by a consortium of European National Statistical Institutes (NSIs) as a so called ESSnet project (ESS stands for European Statistical System). The consortium consists of the statistical institutes of Germany, Italy, the United Kingdom and the Netherlands. The project will finish at the end of 2009. After that, the full results of the project can be found at http://neon.vb.cbs.nl/casc.

The project deals with the confidentiality of the results of micro data research by non-NSI researchers. A lot of NSIs provide access to confidential micro data to external researchers. This access is provided within the safe environment of a Research Data Centre (RDC) at the NSI. All results that researcher want to take out of this safe environment, to include in their publication, are called output. This output needs to be checked to ensure that no confidential data is released to the public domain. Most NSIs have developed their own practices for output checking.

There are a number of reasons to start a project to define guidelines and best practices of output checking:

1. By sharing experiences and defining best practices, experienced countries can learn from each other.
2. Countries wanting to set up a RDC and having little or no experience in output checking, can follow these best practices and be confident that the released output will be safe.
3. A move towards cross-border data access can clearly be seen within the ESS. Cross-border data access is only possible in an efficient way, if one NSI can check output that is based on data from several other NSIs as well. Agreement between NSIs on the output checking practice is crucial for this.

2 The difficulties of output checking

2.1 The nature of an RDC

To understand the difficulties of output checking, one needs to understand the nature of an RDC.

Most NSIs realise that the full potential of their micro data can never be extracted by just their own staff and in their own official publications. The shift from survey-based to register-based statistics has led to a rapid increase in the amount of available micro data. At the same time, IT developments have made it possible to analyse these very large datasets. But a lot of NSIs simply do not have the resources available to make full use of these available datasets. Luckily a very large number of qualified researchers in the world of academics and policy-makers are willing to share in this work. NSIs therefore provide researchers access to their micro data files, usually in a RDC. All the NSIs then have to do is find the right balance between enabling these researchers to do their work while always ensuring the confidentiality of the data.

Essentially an RDC is a safe environment were accredited researchers can access the most detailed original micro data to perform any research that they desire. This makes output checking in a RDC totally different from disclosure control of official publications of an NSI. The official publications are of a well defined form (usually a table) and the intruder scenarios are limited in number. Whereas the output of a RDC can be anything! Researchers twist, transform and link the original data in different and complex new ways. This makes it very difficult to come up with a set of rules for output checking that cover everything. As one expert vividly states it: designing rules for output checking is like designing cages for a zoo – that will keep the animals both contained *and alive* - without knowing which animals will be kept in the cages.

2.2 Two types of errors

Output checking is all about balancing disclosure risk with the value added for society of micro data research. The perfect way to check output minimises the disclosure risk while maximising the use of the data sets. So phrased differently, a less than perfect output check can lead to two types of errors:

1. Confidentiality errors: unsafe (disclosive) output is released
2. Inefficiency errors: safe output is not released

Consider for instance a simple threshold rule for the cell count of tabular output. If this threshold is set too high, this will lead to inefficiency errors: the output will indeed be safe but it will contain less information than could have been obtained from the data file. The researcher might have had to group together the actual populations that he was interested in to meet the high threshold.
On the other hand, setting the threshold too low, leads to high disclosure risks.

3 The rules and guidelines

3.1 Two different models: “principles-based” and “rules of thumb”

Two different models for output checking will be presented in this paper. The first is called the principle-based model. This model keeps the chances on both confidentiality and inefficiency errors as small as possible. The other is called the rules of thumb model. For this model, the focus is on preventing confidentiality errors and not caring too much about inefficiency errors.

3.2 Principles-based model

The principles-based model centres on a good collaboration between researchers and RDC staff. Because this model also tries to prevent inefficiency errors, it can not use simple rules for checking output. The reason is that simple rules can never take into account all the complexity of research output. To give maximum flexibility to the researcher, no output is ruled in or out in advance. All output needs to be considered in its entire context before deciding on its safety. For instance, a table that contains very small cell counts (maybe even some cell counts of 1) isn’t necessarily unsafe. If, for instance, the original data was transformed beforehand, the information that the ‘risky’ cells disclose, might not be traceable to the individual. What is needed is a clear understanding of the governing principles behind disclosure control. Both the researcher and the RDC staff therefore need training in disclosure control.

In the principles-based model all different forms of output are classified as either safe or unsafe. This classification is done solely on the functional form of the output, not on the data itself. For instance, a regression analyses is considered a safe output, irrespective of the underlying data.

If an output class is considered safe, this means that this type of output will normally be released. Only in exceptional cases can an NSI decide that the submitted output will not be released. For instance, a regression were all explanatory variables are binary is basically just a table, and these can potentially be disclosive. The exceptional cases should be well defined and limited in number.

If an output class is considered unsafe, this means that this type of output will generally not be released, unless the researcher can demonstrate that it is non-disclosive. To be able to do this, the researchers needs a good understanding of the principles of disclosure control and the specific data he is working with. Therefore, training of researchers is essential. Note that the burden of proof to convince the NSI to release output of an unsafe type, rests with the researcher.

The principles-based model has the obvious advantage that it leaves a maximum amount of flexibility to the researcher. Data files will therefore be used to their fullest extent. However, the model also has some possible drawbacks:

- The model relies on serious training of NSI staff and researchers. Researchers need to be willing to invest their time and effort on this topic, which is not one they naturally take an interest to.
- The model spreads the responsibility for clearing an output. In a rules-based model, the responsibility lies with the people that design the rules. In this principles-based model, the responsibility lies with each individual checker. There are no strict rules to follow and each checker has to make his own decision on clearing the output, based on his experience and understanding of the underlying principles.

To circumvent these drawbacks, an alternative model is presented: the rules of thumb model.

3.3 Rules of thumb model

In this model the main focus is on preventing confidentiality errors. Some inefficiency errors are taken for granted. This typically leads to very strict rules. The chance that an output, that passes these rules, is non-disclosive is very high. The advantage is that the rules can be applied more or less automatically by both researchers and staff members with only limited knowledge of disclosure control.

It is important to stress the fact that, although the rules of thumb are set very strict, this is no 100 % guarantee that all output that pass these rules is indeed non-disclosive. There is a very small chance that a disclosive output slips through. This is because the rules are rigid and do not take the full context of the output into consideration.

The rules of thumb model is useful for a number of situations:

- Naïve researchers whose output is usually far away from the cutting edge of disclosure control (for instance policy makers who just want some tabular output with limited detail)
- Inexperienced NSIs starting up a RDC. In this case, both users and RDC staff could have too little experience to be able to work with the principles-based model. The rules of thumb model provides them with a starting point that ensures a maximum of safety. In using the rules of thumb model, they build up experience along the way. At some point in time they might feel confident enough to set up the principles based model and open up the way to clearing more complex output.
- Automatic disclosure control methods for RDCs. This will mainly be useful for more controlled types of data access like remote execution. In remote execution, researchers write their scripts on dummy datasets. They then sent the finished script to the RDC, where a staff member (or an automated system) runs it on the full datasets. The results are then returned to the researcher.
- Even for RDCs using the principles based model, the rules of thumb are usually the starting point when checking any particular output. Using the rules of thumb, attention is quickly focused to the parts of the output that breach these rules. These parts can then be considered more carefully with the full principles-based model to decide whether they can be released or not.

3.4 Rules and principles for different classes of output

Possible types of output have been grouped into classes, based on their functional form. Each class is then considered safe or unsafe, as described above for the principles-based model. For each class a rule of thumb and the underlying principles were defined. The extensive overview of all classes, rules and principles is not given in this paper, due to the fact that the project group has not yet finalised their results. Only some examples are given below.

Output class	Rule of thumb	Principles-based model
Regressions	Released	Released
Frequency tables	• Unweighted cell count >= 10 • No cell contains more than 90 % of the row/column total	Any table could be released, taking into account the following issues: • whether the data is itself disclosive (whether it has been transformed; level of detail) • whether firms making up the data could be identified • threshold and dominance results • whether the rank ordering of contributors is known • choice of the cell units • sample choice • weighting
Minima/ maxima	Not released	Released when non disclosive

3.5 Guidelines on organisational and procedural aspects

So far, the paper has dealt with the actual rules for output checking. Apart from these rules, an efficient output checking process of high quality is only possible when some organisational and procedural issues are adequately dealt with. The project also defined practical guidelines for each of these issues, split into a minimum requirement and a best practice. Meeting the minimum requirements enables an RDC to manage the security of outputs effectively. The best practices include operational recommendations and represent the target that an RDC should aim for.

As before, this paper will not go into all the details of these guidelines since they are not yet finalised. But to give the reader an impression, some of them are briefly mentioned below:

- Responsibility for outputs: NSIs should make it clear that the researcher is fully responsible for the content and quality of his outputs. The NSI's only concern is the disclosure risk of the output, nothing else.
- Number of checkers: the minimum requirement is that each output is checked by one person. The best practice is that each output is checked by two persons, to increase quality and objectivity. If one of these two works at the RDC and the other one at the subject-matter department, each will bring his specific knowledge to the checking process. In this case, the final decision on clearing the output rests with the RDC.
- Number/size of outputs. As a minimum standard the NSI should have a policy that it can refuse an output on the grounds of volume of quantity, irrespective of its content. As a best practice, a barrier for submitting large or many outputs is installed. This barrier can be a penalty in money or time. In other words, a fee could be charged or the output can be shifted to the back of the queue, which enhances the time to get it cleared.

4 Concluding remarks

This paper discussed the preliminary results of the ESSnet project 'Guidelines on output checking'. The project will finish at the end of 2009. After that, the full results of the project can be found at http://neon.vb.cbs.nl/casc.

The project has been a first effort to capture the fuzzy process of output checking in harmonised rules. But more work still needs to be done to develop an efficient output checking process for the newly arising international data access infrastructures.

References

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