

Working Paper No. 22
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 ECE/Eurostat work session on statistical data confidentiality
(Luxembourg, 7-9 April 2003)

Topic (vi): Software tools for statistical disclosure control

USING DIS TO MODIFY THE CLASSIFICATION OF SPECIAL UNIQUES

Invited paper

Submitted by the University of Manchester, United Kingdom¹

¹ Prepared by Mark Elliot (mark.elliott@man.ac.uk) and Anna Manning.

Using DIS to modify the classification of special uniques - draft

Introduction

For some time research into statistical disclosure control for microdata has acknowledged the importance of measuring risk and implementing disclosure control at the record level (Fienberg and Makov 1998). However, it has struggled to obtain a comprehensive empirically grounded metric for measuring risk. This paper describes work that aims to provide such a measure by combining the theoretically and empirically grounded file based DIS metric (Skinner and Elliot 2002) with the effective but ad hoc SUDA method (Elliot et al. 2002) for assessing record level risk. The method essentially works by using SUDA to apportion a grossed up file level risk measure generated using DIS between the records in the sample of microdata.

The paper is divided into four parts. The first two describe the basic DIS and SUDA methods leading to a conclusion about the shortfalls of each and how each may be a remedy for the shortfalls of the other. The third part describes a possible method for combining the metrics and the fourth reports a pilot study showing the efficacy of this approach.

The DIS method

The DIS method has similarities of form to the bootstrapping methods developed by Efron (1979). The basic principle of the method is to remove a small number records from the target microdata file and then copy back some of those records, with each record having a probability of being copied back equal to the sampling fraction of the original microdata file. This creates two files, a new slightly truncated target file and a file of the removed records, which is then matched against the target file. The method has two computational forms, the *special form*, where the sampling, described above, is actually done and the *general form*, where the sampling is not done but the equivalent effect is derived using the partition structure² and sampling fraction of the microdata file.

The special method

The special method follows the following five-step procedure:

- (i) Take a sample microdata file (A) with sampling fraction f .
- (ii) Remove a small random number of records (B) from A, to make a new file (A').
- (iii) Copy back a random number of the records from B to A' with each record having a probability of being copied back equal to f .

The result of this procedure is that B will now represent a fragment of an outside database (an identification file) with an overlap with the A' equivalent to that between the microdata file and an arbitrary identification file with zero data divergence (with no identical values on the matching keys for the same individual).

- (iv) Match B against A'. Generate an estimate of the matching metrics particularly, the probability of a correct match given a unique match, $pr(cm|um)$, between the fragment B and the file A'.
- (v) Iterate through stages i-iv until the estimate stabilises.

² The partition structure of a file is a frequency of frequencies for a given cross-classification. A partition class is a particular cell within the partition structure. So, for example, *uniques* is the partition class where the sample frequency is 1. This has also been referred to as the *equivalence class structure*; Greenberg and Zayatz (1992)

DIS: The general method

A more general method can be derived from the above procedure. Imagine that the removed fragment (B) is just a single record. There are six possible outcomes depending on whether the record is copied back or not and whether it was a unique, in a pair or in a larger partition class.

Table 1: Possible per record outcomes from the DIS general method

record is:	Copied back	not copied back
sample unique	correct unique match	non-match
one of a sample pair	multiple match including correct	false unique match
one of a larger equivalence class	multiple match including correct	false multiple match

The two critical cells in the above table are:

1. where a unique record is copied back - this creates a correct unique match
2. where one of a sample pair is not. - this creates a false unique match

Given this, we can induce that the relative numbers in these two cells determine the probability of a correct match given a unique match; $pr(cm|um)$. Given this, it is possible to shortcut the special method³, since one can derive an estimated probability of a correct match given a unique match from:

$$pr(cm | um) \cong \frac{U * f}{U * f + P * (1 - f)}$$

Where U is the number of sample uniques, P is the number of records in pairs and f is the sampling fraction.

So the general technique provides a simple method for estimating the probability of an intruder being able to match an arbitrary outside file to a given target microdata file. Numerical studies (Elliot, 2000; Skinner and Elliot 2002) have demonstrated empirically that the above method is sound at the file level of risk, in that it provides stable estimates of the population level of matching probabilities. Further, Skinner and Elliot provide proof that estimated $pr(cm|um)$ is an unbiased estimator for the real matching probability.

However, as the method stands the DIS provides only a *file level* measure of disclosure risk. This has been very useful in for example comparing a proposed data file with an existing one (Tranmer et al 2001) or examining the impacts of disclosure control techniques (Elliot 2001) Recently, it has been recognised that risk varies across any given data file and so also needs to be analysed at levels other than the whole file (Elliot 2000, Skinner and Holmes 1998, Fienberg and Markov 1998) and in particular at the level of individual records and for DIS to be of wider use it must be integrated into a method of record level risk assessment.

The SUDA method

The phrase “special uniqueness” was first used by Elliot et al (1998) to describe records within a dataset which were unique by virtue of processing a demographically unusual combination of characteristics as opposed to merely happening to be unique because of the sampling fraction and variable codings employed

³ As later discussion will show the special method is still necessary to, for example, assess the impact of a disclosure control method that does not systematically alter the partition structure of a file.

(random uniques). The initial technical definition of special uniques was a sample unique which maintained its uniqueness despite collapsing of the geographical component of its key variables.

Elliot and Manning (2001a) have shown that special uniques using this definition are far more likely to be population unique than are more likely to be population unique than random sample uniques. Elliot and Manning also showed that the same principle could also be applied to variables other than geography and that persistence of uniqueness through aggregation of any variable indicated ‘specialness’. This in turn has led to a re-definition of special uniqueness. In the final definition a special unique is: *a sample unique on key variable set K which is also sample unique on variable set k where k is a subset of K*. In fact this technical definition is one part of a broader definition of the distinguishing features of special Uniques, Table 2 shows the full set.

Property	Special Uniques	Random Uniques
<i>Cognitive</i>		
Spontaneously recognisable	Maybe	No
Appear Unusual	Maybe	No
<i>Data - Analytical</i>		
Sensitive to removal of key variables	No, except if one of the contributing variables	Yes
Sensitive to changes in sampling fraction	No	Yes
Sensitive to changes in geographical detail	No	Yes
Uniqueness arises from	Small number of variables	Large Number of Variables

Elliot, et al (2002) have developed a high performance computing methodology (called SUDA – the Special Uniques Detection Algorithm) to enable comprehensive cross file analyses of special uniques, potentially circumnavigating the issue of key variable choice. SUDA works by identifying every unique pattern (combination of attributes) in every record and rating records according to the number and size of unique patterns that they have.¹ The smaller the number of variables the more risky it is; this is implicit in the final definition of special uniqueness. SUDA weights the size of a unique pattern according to the number of unique supersets produced from that pattern.

Elliot, et al. show that SUDA is very good at identifying population uniques within a sample and has a strong relationship with the population level equivalence class for any given record/key.⁴

Although the concept is undoubtedly useful in itself⁵, as a measure of disclosure risk it has one serious flaw, the SUDA output metric is *ad hoc* with respect to a particular analysis on a particular file, with a particular combination of key variables. This means it is very difficult to interpret what a particular SUDA output score means in terms of the underlying risk and it is impossible to compare two scores produced by different analysis. In order for SUDA to be useful it needs to be calibrated against reliable consistent measure, which can be calculated alongside the SUDA metrics.

⁴ Work is also in progress to examine the psychological aspect of the special uniqueness concept.

⁵ Indeed, as the special uniqueness metric is essentially a distance metric – i.e. a measure of unusualness is a measure of distance from the ‘centre’ of some sociometric space we can view it as a measure of demographic individually and therefore the distribution of it within a population as a measure of demographic diversity. Work is in progress examining the properties of this distribution.

The integrated DIS-SUDA method

Since DIS produces a well-calibrated, consistent metric that needs to be integrated into the record level and SUDA produces a record level risk metric that needs calibrating, it would make sense to examine if the two methods can be combined. Since DIS produces a calibrated metric, the logical way to proceed is to use DIS to calibrate SUDA. There are a variety of ways to do this, the method reported here makes most sense computationally, although from a statistical point of view the quiriness of dealing with a small number of probabilities greater than 1 means it may be sub-optimal.

Computationally, the method works as follows:

1. Run DIS over target database
 - Let $T = pr(cm|um) * SU$
2. For each record
 - Let $PS = LN(SUDA\ score) / \sum LN(SUDA\ score)$
 - Let $R = T * PS$
3. Let $T = 0$
4. For each record where $DS > 1$
 - Let $T = T + R - 1$
 - Let $DS = 1$
5. For each record (where $DS < 1$)
 - Let $PS' = LN(SUDA\ score) / \sum LN(SUDA\ score)$
 - Let $DS = DS + T * PS'$
6. Repeat steps 3 through 6 until $T = 0$

In effect the grossed up DIS score is divided up between the records in the sample in proportion to the natural log of their SUDA output score. Any records that have a score that is greater than 1 has its score reset to 1 and the difference is divided amongst the remaining records (again in proportion to their score). The effect is to produce an output metric that is a record level probability any given match being correct.

Numerical Study

If the DIS-SUDA algorithm works, the output metric should show a strong relationship with the underlying matching probability. Ignoring the effect of data divergence this can be defined as the reciprocal of the population equivalence class for the given key/record.

The following pilot study was conducted to examine this.

1991 GB census data was used for twelve variables and two local authority files for which we have population data these are. The population equivalence class for the twelve variable key was recorded against each record and its reciprocal calculated. The reciprocal of the population equivalence class for a given record/key is the effective underlying matching probability for that record (on the assumption of zero data divergence). The cross-file mean of this figure is effectively what DIS estimates.

Fifty parallel, stratified 2% samples⁶ were drawn. For each sample the DIS score was calculated and the DIS-SUDA score was calculated for each record using the method described in the previous section. The correlation between the reciprocal of the equivalence class was calculated.

⁶ The files we used are stratified down to the household level. ONS supplied these anonymised 1991 Census data for UK Local Authorities us under contract. The data are kept in a secure environment, access is limited to members of the project team for the designated parts of the project, and the data will be returned to ONS when the work is

Table 3 shows the relationship between the equivalence class and the grouped DIS-SUDA score. As it can be seen there a clear relationship between the two. DIS-SUDA score of one are highly predictive of population uniques (i.e. records with a matching probability of 1)

(Table 3 goes about here)

Conclusions

Further work is in progress investigating the properties of this output metric and examining other was of combining DIS and SUDA. However, this work is highly promising in that it indicates that an accurate matching probability based record level risk metric is achievable.

References

Bethlehem, J. G., Keller, W. J., and Pannekoek, J. (1990) Disclosure control of microdata, *Journal of the American Statistical Association* Vol 85, 38-45.

Efron, B. (1979) Bootstrap methods: Another Look at the Jack-knife. *Annals of Statistics*, Vol 7 1-26.

Elliot, M. J. DIS: "A new approach to the measurement of statistical disclosure risk." *International Journal of Risk Management* 2(4) (2000): 39-48.

Elliot, M. J. (2001) "Data intrusion Simulation: Advances and a vision for the future of disclosure control." Paper presented to the 2nd UNECE work session on statistical data confidentiality; Skopje March 2001.

Elliot, M. J., Manning, A. M. and Ford, R. W. (2002). 'A Computational Algorithm for Handling the Special Uniques Problem'. (To appear) *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems* 5(10), pp 493-509.

Elliot, Skinner, C. J., and Dale, A. (1998) 'Special Uniques, Random Uniques and Sticky Populations: Some Counterintuitive Effects of Geographical Detail on Disclosure Risk.' *Research in Official Statistics*; 1(2) pp 53-68

Fienberg, S. E. and Makov, U. E. (1998), 'Confidentiality Uniqueness and Disclosure Limitation for Categorical Data', *Journal of Official Statistics* 14(4). pp 361-372.

Skinner, C. J. and Elliot, M.J. (2001) 'A Measure of Disclosure Risk for Microdata'. CCSR occasional paper 23, *Journal of the Royal Statistical Society Series B*. 64(4) pp 855-867.

Skinner C. J. and Holmes D. J. (1998), 'Estimating the Re-identification Risk per Record', *Journal of Official Statistics* 14(4). pp 361-372.

Tranmer, M, Fieldhouse E., Elliot, M. J., Dale A., and Brown, M. (forthcoming 2001) 'Proposals for Small Area Microdata'. Accepted subject to revisions *Journal of the Royal Statistical Society, Series A*.