STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

COMMISSION OF THE EUROPEAN COMMUNITIES

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

Joint ECE/Eurostat Work Session on Statistical Data Confidentiality
(Skopje, The former Yugoslav Republic of Macedonia, 14-16 March 2001)

Working Paper No. 20
English only

Topic II: Impact of new technological developments in software, communications and computing on SDC

DATA INTRUSION SIMULATION: ADVANCES AND A VISION FOR THE FUTURE OF DISCLOSURE CONTROL

Contributed paper

Submitted by the Centre for Census and Survey Research, University of Manchester, U.K.[1]

I. INTRODUCTION

1. The accurate and reliable assessment of statistical disclosure risk is a necessary pre-cursor of the efficient application of disclosure control techniques. Research on statistical disclosure risk assessment with respect to microdata files has two major themes. The first uses the concept of *uniqueness,* examining the level of unique individuals or households within populations and samples (See for example, Dalenius Bethlehem et. al. (1990)). The second attempts to model actual disclosure attempts by matching individual records in a target file with those in an identification file (for example Muller et al(1992), Elliot and Dale(1998)). Both of these approaches are problematic. The uniqueness statistics require population data to calculate them and do not relate directly to with what a data intruder might actually do in attempting to identify individual within a dataset. The matching experiment approaches have a direct relationship but are *ad hoc* with respect of the identification data set chosen.

2. Elliot (2000) has developed a method called *Data Intrusion Simulation* (DIS) for calculating the general risk for a given target file, which does not require population statistics or matching experiments and is grounded in the matching process that the intruder must use in order to identify individuals within an anonymised dataset.

3. Elliot provides empirical evidence that the method gives accurate estimates of matching metrics such as *the probability of a correct match given a unique match* and therefore is a good measure of real disclosure risk.

4. The remainder of this paper is divided into three parts. Part II gives a brief description of the DIS method. Part III delineates a proof that the general DIS method gives an unbiased estimate of the population level of the probability of a correct match given a unique match. Part IV describes some recent work using the DIS special method to estimate the impact of disclosure control methods on the level disclosure risk.

II. THE DIS METHOD

5. The basic principle of the DIS 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

[1] Prepared by Mark Elliot. The work described in this paper was supported by the UK Economic and Social Research Council (grant number R000 22 2852) and through direct funding from the U.S. Bureau of the Census.

file. The method has two computational forms, the *special form*, where the sampling is actually done and the *general form*, where the sampling is not actual done but the equivalent effect is derived using the partition structure of the microdata file and sampling fraction.

The special method

6. The special method follows the following five-step procedure, (a schematic version can be found in Appendix B).

(i) Take a sample microdata file (A) with sampling fraction S.

(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 in B to A' with each record having a probability of being copied back equal to S.

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 differing values 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.

(v) Iterate through stages i-iv until the estimate stabilises.

DIS: The general method

7. 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

8. The critical cells in the above table are those where a unique record is copied back and where one of a sample pair is not. The relative numbers in the cells determine the probability of a correct match given a unique match; pr(cm|um). Given this, it is possible to shortcut the special method[2], since one can derive an estimated probability of a correct match given a unique match from:

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

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

[2] 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.

A justification for the DIS general method

9. Skinner (2000) has generated the following proof that DIS provides an unbiased estimate of the probability of a correct match given a unique match.

Let I denote key value i= 1…K
F_i denote population frequency of
(so F_i =1 if population unique)
f_i denote sample frequency of i
(so f_i =1 if sample unique)

Let I(') denote truth function
So e.g. $I(f_i=1)$ =1 if f_I = 1
=0 if $f_i \neq 1$

$$pr(cm \mid um) = \frac{\sum_{i=1}^{k} I(f_i = 1)}{\sum_{i=1}^{k} F_i \, I(f_i = 1)}$$

DIS general method estimates pr(cm|um) by

$$\hat{pr}(cm \mid um) = \frac{\left[\sum_i I(f_i = 1)\right]\Pi}{\left[\sum_i I(f_i = 1)\right]\Pi + \left[\sum_i 2I(f_i = 2)\right](1-\Pi)}$$

Where Π is the sampling fraction.

The above can be written as:

$$pr(cm \mid um) = \frac{\sum_i I(f_i = 1)}{\sum_i f_i I(f_i = 1) + \sum_i (F_i - f_i) I(f_i = 1)}$$

$$\hat{pr}(cm \mid um) = \frac{\sum_i I(f_i = 1)}{\sum_i f_i I(f_i = 1) + \left[\sum_i I(f_i = 2)\right]\frac{2(1-\Pi)}{\Pi}}$$

So pr(cm|um) is a good estimator of pr(cm|um) if :

$$[\sum_i I(f_i = 2)]\frac{2(1-\Pi)}{\Pi}$$

is a good estimator of:

$$\sum_i (F_i - f_i) I(f_i = 1)$$

Assume model:

F_i/λ_i Poisson $(\pi\lambda_i)$, F_i-$f_i|$ λ_i Poisson $((1-\pi)\lambda_i)$

f_i,F_i-f_i conditionally independent given λ_i

Then

$$\mathrm{E}\Big[\sum I(f_i = 2)\Big]\frac{2(1-\Pi)}{\Pi} = \sum \frac{(\Pi\, li)^2}{2}\, e^{-(\Pi\, l_i)}\, \frac{2(1-\Pi)}{\Pi}$$

$$= \sum \Pi(1-\Pi)\, l_I{}^2 e^{-(\Pi li)}$$

$$\mathrm{E}\Big[\sum (F_i - f_i)\, I(f_i = 1)\Big] = \sum (1-\Pi)\, l_i\, (\Pi\, l_i) e^{-(\Pi\, l_i)}$$

$$= \sum \Pi(1-\Pi)\, l_I{}^2\, e^{-(\Pi li)}$$

So $\mathrm{E}\Big[\sum (F_i - f_i)\, I(f_i = 1)\Big]$ is unbiased for $\mathrm{E}\Big[\sum I(f_i = 2)\Big]\frac{2(1-\Pi)}{\Pi}$

So $\hat{pr}(cm\,|\,um)$ is approximately unbiased for $pr(cm\,|\,um)$.

III. USING DIS TO ESTIMATE THE EFFECT OF SDL TECHNIQUES

10. The general DIS method is not able to assess the effect of SDL techniques other than re-coding and in particular it is unable to measure the impact on risk caused by perturbation. This is because perturbation techniques such as blurring or record swapping do not systematically alter the equivalence class structure of a file and therefore the effect they have on estimated pr(cm|um) is arbitrary; whereas, by definition, perturbation SDL techniques will reduce the real levels of this probability. Note, this not a flaw in the general DIS method, as the primary goal in its initial development was to provide an accurate estimate of the worst case level of file-level risk - i.e. assuming that an intruder recovers matches across perturbation and other forms of data divergence. Nevertheless, expanding the DIS method to allow it to estimate the optimum effect on risk of perturbation techniques (and other sources of divergence) is a crucial aspect of its further development and it is this expansion that is reported in this section.

The Revised Special Method

11. In order to circumnavigate the limits of the general method it is necessary to use a revised form of the special method. In this rather than taking iterative random samples each record is removed in turn and then matched directly back onto the source file. Each unique match against the record itself is recorded as a correct unique match and each match against a pair recorded as a false unique match. By summing the number off correct unique matches as *T* and the number of false unique matches as *F*, the resultant formula derives back to that used in the general method

$$pr(cm\,|\,um) \cong \frac{T * f}{T * f + F * (1-f)}$$

where *f* is the sampling fraction. This method was tested using the data described in section 2 and produced identical results to the general method (and so *by induction* has the same statistical backing as the general method).

12. To examine the effect of perturbation on this situation, it is helpful to think about what happens at the record level in terms of the effect of the perturbation on the main constituents of pr(cm|um) uniques and pairs. The situation is open-ended as multiple records will be affected by the SDL technique, however some examples will give a flavour of what is going on:

- When a record(R) is unique and is perturbed this will result in a non-match

- When R is unique and is not perturbed but another record (R2) is perturbed to form a pair with R then (1-*f)* false unique matches will be generated and a correct unique match will be lost.
- When R is in a pair and is perturbed into a unique then a false unique match will be created at a probability of 1 rather than 2*(1-*f)*
- When R is in a pair to which perturbation causes another record to be added to then 2* (1-f) false matches will be lost.

13. As this illustrates the application of an SDL technique is non-linear in respect of its affect on the constituents of pr(cm|um). However, the effects only lead to a slightly different procedure for calculating pr(cm|um) as at the end of the effects of the SDL the there is own one additional component category - an actual false match (as opposed to one that is derived probabilistically from the number of pairs. So, to extend the revised method to allow the testing of SDC techniques the following procedure was adopted.

(i) The SDC technique to be tested was applied to the source file(A) to produce a second perturbed file (B)
(ii) Each record from (B) was then matched against file A.
(iii) Each unique match against the original record was recorded as S correct unique matches (*T*) (where S is the sampling fraction).
(iv) Each unique match against a record other than the original record was recorded a match against a unique false match (*F*).[3]
(v) Each match against a pair was recorded as (*P*).

The formula for calculating pr(cm|um) then becomes:

$$pr(cm \mid um) \cong \frac{T * f}{T * f + P * (1 - f) + F}$$

This will give you a pr(cm|um) value which is adjusted for the perturbation caused by the SDL technique. By comparing this with the pr(cm|um) assessed on the original file using the normal DIS method a figure is obtained for the reduction in risk (at the whole file level) caused by the SDL technique.

14. It should be noted that the scale of the effect is arbitrary with respect a particular SDL manipulation. Thus, by calculating adjusted pr(cm|um) in this way, one obtains a figure which is as accurate for the particular perturbed file as the general method is for an unperturbed file. However, in order to investigate a particular SDL technique in general the procedure needs to be iterated until it produces stable mean values. After experimentation it was decided that each calculation should be iterated until the mean probability over all iterations had stabilised to the third decimal place over the previous the 10 iterations or when 100 iterations had passed. This was partially a computational decision, since the weight of computation required to do this analysis was substantial. Nevertheless, the end product is a stable pattern of results.

Example of Empirical work: The Impact of Record Swapping

15. The empirical work described here uses the 2% individual Sample of anonymised records (SAR) from the 1991British census. The analyses use two forms for key variable selection, The first form, the additional impact analyses, developed by Elliot (2000b), uses a combination of a fundamental key (age, sex, marital status) with each other variable in the dataset. Therefore the figures in the basic key analyses that follow are the means of 41 separate iterative calculations. The second form uses a set of five intrusion scenarios developed by Elliot and Dale (1998).

16. There are a variety of methods of record swapping and it is not possible to provide an all-encompassing assessment of the technique. In the context of the dataset under analysis the record swapping was conducted as follows. A sample of records stratified by SAR area is removed from the SAR. Each of these records is then paired with a record in the remainder of SAR from the same

[3] This will happen when a record in a pair on the original file has been perturbed.

geographical region as the original SAR (Regional geography is a 12 way division of Britain). The pairing is either done randomly or through pairing variables[4]. Four combinations of pairing variables are investigated.

(i) Age
(ii) Age, Sex
(iii) Age, Sex and Marital Status
(iv) Tenure
(v) Tenure and number of residents

Pairing variable set (iii) is the basic key, which is fundamental to the disclosure risk work we have done on individuals, and pairing set (v) is proposed for use in the 2001 UK census (although at a lower level of geographical detail). The other pairing sets are derived from those two. Four Swapping fractions[5] were assessed: 1,2,5 and 10%.

17. Table 2 shows the results for the mean pr(cm|um) values for the additional impact analyses. Table 4 shows the probabilities for the scenario analysis. Tables 3 and 5 shoe the results expressed as a proportion of the mean value for the unperturbed data. Several patterns are quite clear from these results:

(i) The reduction in risk is proportionate to the swapping fraction.
(ii) The reduction in risk is greater on the matched pairings than on the random pairings.
(iii) The greater the differentiation of the pairing variables the greater the reduction of risk.

18. These last two findings are important. The fact that the matched pairing show lower risk pairings is at contradiction with the generally accepted tension between disclosure risk and data quality. Since - by definition - the stronger the match the lower the impact on data quality. It seems likely that the increased number of false causes this effect. When a unique record is swapped out there is a higher probability that a false match will be created when the record is paired. This clearly needs further investigation. However, it is also important to note that the scale of this effect in absolute terms is small particularly in the case of the scenario analyses. Even with 10% data swapping the best risk reduction achieved was from 0.158 to 0.138 which is a 13.3% reduction in the risk probability. This is considerably less than the risk impact of naturally occurring data divergence (Elliot and Dale 1998).

Table 2: mean pr(cm|um) for basic keys by data swapping paring variable set and swapping fraction

	swapping fraction			
pairing variable set	**1%**	**2%**	**5%**	**10%**
Random	0.0261	0.0258	0.0249	0.0230
Age	0.0258	0.0252	0.0243	0.0222
Age, Sex	0.0257	0.0250	0.0244	0.0220
Age, Sex, Marital Status	0.0253	0.0248	0.0238	0.0213
Tenure	0.0260	0.0254	0.0249	0.0227
Tenure/number of residents	0.0259	0.0253	0.0247	0.0225

[4] Where no pair could be found a random record was chosen instead. However, with these small swap keys and large pool of potential swaps, this was an infrequent event.
[5] Note that the swapping fractions are twice the actual sampling factions used in the swap procedure because each sampled record is swapped with an unsampled record.

Table 3: mean pr(cm|um) for basic keys by data swapping paring variable set and swapping fraction expressed as a proportion of that for unperturbed data

	Swapping fraction			
pairing variable set	**1%**	**2%**	**5%**	**10%**
Random	0.989	0.977	0.943	0.871
Age	0.977	0.955	0.921	0.841
Age, Sex	0.974	0.947	0.924	0.833
Age, Sex, Marital Status	0.959	0.940	0.902	0.807
Tenure	0.985	0.964	0.941	0.860
Tenure/number of residents	0.980	0.960	0.933	0.852

Table 4: mean pr(cm|um) for scenario keys by data swapping paring variable set and swapping fraction

	Swapping fraction			
pairing variable set	**1%**	**2%**	**5%**	**10%**
Random	0.157	0.154	0.149	0.142
Age	0.157	0.154	0.149	0.141
Age, Sex	0.156	0.152	0.146	0.138
Age, Sex, Marital Status	0.155	0.153	0.147	0.137
Tenure	0.157	0.153	0.149	0.139
Tenure/number of residents	0.157	0.154	0.149	0.138

Table 5: mean pr(cm|um) for scenario keys by data swapping paring variable set and swapping fraction expressed as a proportion of that for unperturbed data

	swapping fraction			
pairing variable set	**1%**	**2%**	**5%**	**10%**
Random	0.991	0.974	0.944	0.898
Age	0.989	0.975	0.941	0.894
Age, Sex	0.987	0.961	0.922	0.873
Age, Sex, Marital Status	0.980	0.964	0.928	0.867
Tenure	0.992	0.967	0.942	0.880
Tenure/number of residents	0.993	0.973	0.942	0.874

19. To summarise, the method described in this section extends the DIS methodology to allow assesment of the impact of disclosure control techniques. An example using data swapping which indicates that even for fairly high swapping fractions (10%) the risk reduction is relatively small. Work not presented here indicates similar results from masking techniques (see Elliot 2000c).

IV. FUTURE DIRECTIONS - MOVING FROM PASSIVE RISK ASSESSMENT TO ACTIVE RISK DRIVEN FILE CONSTRUCTION: A PROPOSAL FOR THE FUTURE OF DATA RELEASE

20. Research into disclosure risk assessment has produced refinements in the way in which disclosure risk is measured, including the DIS methodology described in this report. These are being taken on board by data providers who are using them to make more sophisticated judgements regarding safe data release.

21. However, risk assessment is still viewed as applied research and so in practical terms the cycle of assessment and refinement is lengthy. The framework is reminiscent of the third generation of computer programming where 'data' and 'analysis' are seen as distinct. Although the results of analyses are used in decision making risk assessment methodology is not itself used in the construction of data sets for release. The existing framework can be seen as:

(i) Data is collected;
(ii) Data is coded and stored in a computer file;
(iii) Decisions are made about release file codes using *ad hoc* rules of thumb (e.g. bivariate cell counts must of at least size X) and risk-proxy thresholds (for example population per geographical unit must be > X);
(iv) Disclosure risk analysis is conducted by researchers;
(v) On the basis of (iv) changes are made to the rules and thresholds used in stage (iii).

22. One of the major problems with this approach is that the thresholds and rules that are applied at stage three are applied across the file, which does not acknowledge the variations in risk patterns. So the threshold set may be too risky in some places and overly cautious in others.

23. A superior framework would be to replace the application of rules and risk-proxy thresholds with a method of file construction that is directly risk assessment driven. So, in this alternative model a risk threshold rather than an arbitrary population threshold determines the level of detail. For example, it may be possible to have more detailed information on ethnic group in ethnically diverse areas than in areas which are predominantly white. This would also be useful analytically since these are the areas in which such detail is most useful.

24. In order for this to be viable a fully-fledged multilevel risk assessment system would need to be constructed. By multilevel, I mean one which takes account of file, record and intermediate levels of data in assessing risk, such a system is close to fruition in the form of DIS system. As DIS is able to assess risk at multiple levels through breaking down variables and examining risk conditional on value constraints as demonstrated by Elliot (2000a) and therefore it is possible to imagine such a system allowing such active analysis.

25. A second aspect of this will be to link the DIS methodology to a record-level risk assessment methodology. A pilot of such a system is under development linking the special uniques identification methodology with the DIS system (see Appendix for a description of the recently conducted validation study). As the pilot analysis shown in the appendix (and in particular Table 20) indicates this type of addition to the DIS system has the potential for a far greater impact on overall risk levels than the methods exhaustive tested in this report.

26. However, even more critical than this is to incorporate into the file generation system a method of assessing the impact on analytical power of the data; this consists of:

- *Analytical completeness:* the ability of data users to conduct the analyses they to with the data set, and
- *Analytical Validity:* the ability of users to obtain the same results from the dataset as they would have from the raw data.

27. This can be seen as a constraint optimisation problem - statistical disclosure risk and data quality demands pulling in opposite directions, with the task of the system being to trade off the two competing demands. However, at some points in the data-quality/disclosure-risk constraint space it may be possible to move in a positive direction with both through appropriate technique selection. Indeed the whole point of an active file construction system is to optimise both disclosure risk and data quality at a higher level than current data release methodology allows.

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,

Dalenius, T. (1986) Finding a needle in a haystack. *Journal of Official statistics*, Vol 2(3), 329-336

Elliot, M. J. (2000a) DIS: A new approach to the measurement of statistical disclosure risk. .*Risk Management: An international Journal. Vol 2 (4)*

Elliot, M. J. (2000b) Data Intrusion Simulation Project. *Final report to the Economic and Social Research Council*

Elliot, M. J. (2000c) Applying the Data Intrusion Simulation methodology to assess the disclosure risk impact of SDL techniques on a microdata file. *Report to the United States Census Bureau.*

Elliot M. J. and Dale (1998). *A disclosure risk matching experiment.* Report to the European Union on disclosure risk for microdata ESP/ 204 62/DGIII; DM1.5.

Muller, W; Blien, U.; Wirth, H. (1992). Disclosure risks of anonymous individual data. Paper presented at the 1st International Seminar for Statistical Disclosure

Skinner, C. J. (2000) *Personal Communication.*