WP.14
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
(Manchester, United Kingdom, 17-19 December 2007)

Topic (i): Microdata

USING TARGETED PERTURBATION OF MICRODATA TO PROTECT AGAINST INTELLIGENT LINKAGE

Supporting Paper

Prepared by Mark Elliot (Centre for Census and Survey Research,
University of Manchester, United Kingdom)

Using Targeted Perturbation of Microdata to Protect Against Intelligent Linkage

Mark Elliot

Centre for Census and Survey Research University of Manchester[1]

1 Introduction

This paper describes linkage experiments using the 2001 individual Sample of Anonymised Records (SARs) from the UK census to the microdata output from the UK Labour Force Survey (LFS) for spring 2001.

The objective of the study was to assess the impact of the statistical disclosure control methods on the used on the 2001 SARs on the ability to link an external dataset and SARs. The Labour Force Survey was selected as the external file because (i) it was of sufficient size to produce a large enough overlap with the SARs and was collected around the census date.

The project follows the tradition of other such studies with official data; e.g. Muller, W; Blien, U.; and Wirth, H. (1992), Elliot and Dale (1998). However, the study here elaborates on that earlier work by examining the impact of a targeted disclosure control technique on the ability of an intruder to attack a dataset by focusing on the high risk records.

2 Method

2.1 Data

Three datasets were used in this study:
The spring 2001 quarter of the standard Labour Force Survey (LFS).
The standard release version of the 2001 individual level SAR (post-SDC SAR).
The pre-SDC version of the 2001 individual level SAR (Pre-SDC SAR).

The 2001 SARS were subject to extensive statistical disclosure risk assessment and targeted control methods. The risk assessment was carried through collaboration of Manchester University and ONS staff using the software system SUDA 1.4. The disclosure control was a mixture of global recoding and local suppress and reimputation based on a variant of the PRAM (post randomisation) method. This involves using a replacement probability matrix for each value for each variable. The result was a file which was regarded as having been sufficiently protected against deliberate disclosure attempts. A version of the SARS without the SDC applied is available in safe settings within ONS. This is in effect the pre-SDC file used here.

2.2 Procedure

The procedure consisted of seven steps:

1) The variables were selected and then the codings of these variables on the different datasets were harmonised.

[1] This work was supported by the Office for National Statistics in the UK, who carried out the verification procedure and provides access to the data under special license.

2) The matching was conducted

3) The SUDA Software was run over the SARs to obtain DIS-SUDA scores for the matches.[2]

4) All the unique matches (one-to-one) were sent to ONS

5) The matches were verified by ONS Census and LFS divisions

6) The matches were returned with an indicator placed on the match file indicating whether the match was true or false.

7) The proportion of correct matches was generated under several different assumptions.

2.3 Variable Selection and Mapping

The first step in the matching was to harmonise the data. This involved selecting the variables that were to be used as the match key and then recoding them on one or both datasets so that the codings were consistent.

Frequent practice when considering disclosure risk assessment is to use standard scenarios for key selection; see Elliot (2006)[3]. However, the variables available on the intersection between the LFS and SARs did not correspond to any of the scenarios in Elliot (2006). Since we were mimicking an attack from an external database Scenario 4a2 was used as a base. However, three of the standard variables within this scenario: "Number of cars in the household", "distance of travel to work" and "workplace" could not be included. To make the scale of the attack represent what was possible under scenario 5a ethnic group and country of birth were added to the scenario, effectively creating a blend of scenarios 4 and 5. To simplify the process only the data for England and Wales and persons in households were included. Non resident students were excluded.

This created a key with the following nine key variables:

Region
Age
Sex
Marital Status
Number of residents in household
Tenure
Primary economic status
Ethnic group
Country of birth

[2] SUDA (Special Uniques Delectation Algorithm) is a method and a software system for detecting risky records within microdata. It does by identifying all minimal sample uniques (sample uniques with no unique supersets) within each record up a user specified size. By using a heuristic to combine all the information for each record with a well established file level probability measure, a per record risk measure is obtained. This enables the records to be sorted in order of risk. The use of this software here enables the simulation of a more sophisticated intruder who is able to take account of the structure of the target file in order to improve his/her confidence in the matches s/he finds.

[3] Scenarios 4a2 and 5 are shown in appendix A

2.3.1 Harmonising the key variables

As with most matching studies there then followed a complicated process of variable harmonisation; this required manipulation of all three datasets and resulted in the following
Age (95 categories for the pre SDC SARS-LFS match, 44 categories for the post-SDC SARs-LFS match)
Sex (2 Categories)
Marital Status (5 Categories)
Region of residence. (11 Categories)
Number of Residents (7 categories)
Primary economic status (9 categories)
Country of birth (14 Categories)
Ethnic group (15 categories)
Tenure (5 categories)

2.4 Matching

Once the variable harmonisation had been conducted, the matching process was relatively straightforward. In principle, we could have used fuzzy matching methods – to allow for data divergence – however, the number of direct one to one matches was very large on both files and therefore this was deemed to cause an unnecessary administrative burden at the match verification stage. Therefore, a simple combine and sort algorithm was used for the matching.

In all there were 6085 one to one matches between the pre-SDC SAR and the LFS and 3130 one to one matches between the released SAR and the LFS. The SAR id and LFS identifier fields for the matched records were combined into a file and these were sent to ONS for verification.

2.5 Match verification

A problem occurred when at the verification stage as for a significant number of matches there was no address linkable to the LFS identifying variables. This affected 1602 matches (26.32%) against the pre-SDC SARS file and 895 matches (28.95%) against the post-SDC SARS file.

There was also some interaction between some of the matching variables and whether a name and address has been found. Ethnic group in particular was related to whether a name and address could be found with 22.6% of white people and 34.2% of non-white people on the post-SDC file not being found ($V=0.126$, $p<0.0005$). On the pre-SDC match file the figures were 23.6% and 33.5% respectively ($V=0.101$, $p<0.0005$).

There was some apparent interaction with the SUDA score indicating that there was a higher probability of a name and address not being found if a record had a high DIS-SUDA score(>0.3). However, this was only slight and not statistical significant ($V=0.005$, $p=0.773$).

Although there is some cause for concern about the number of cases for which it was not possible to verify or not whether a match was true and the interaction between that and some of the match key variables, it was decided that the non significant association with the SUDA score meant that it was acceptable to proceed with the analysis discounting the matches for which there was no name an address was found.

3 Results

The headline results are given in tables 1 and 2. Overall there were 3130 matches between the post-SDC SARS file and the LFS and 6085 between the pre-SDC SARS file and the LFS. After discounting the matches for which no name and address was found these figures dropped to 2235 and 4483 respectively. The difference between these two totals bears discussion in itself. In effect the disclosure control has reduced the number of matches (true and false) by approximately half. This result is probably mostly due to the global recoding particularly of the age variable; as we might expect perturbation to produce new false matches as well as disguising true existing ones.

In terms of the absolute number of matches the SDC process can be said to have had a significant impact with 123 correct matches pre-SDC dropping to 51 post-SDC, nearly a 60% reduction.

In terms of the match rates we can observe that the raw match rates on both pairs of files are low: 2.74% in the matching with the pre-SDC SARs and 2.28% on the post-SDC SARs. This indicates that the SDC has had some effect in the gross match rates. However, even the match rate for the pre-SDC SAR is low and these are the sorts of figures that should be of little concern to ONS. An intruder faced with that sort of success probability would be in effect swamped by false matches.

		Frequency	Percent	Valid Percent
Valid	False match between SAR and LFS	2184	69.78	97.72
	Correct Match between SAR and LFS	51	1.63	2.28
	Total	2235	71.41	100.00
Missing	No name and address from LFS file	895	28.59	
Total		3130	100.00	

Table 1: Match Indicator for matches between post-SDC SAR and the LFS

		Frequency	Percent	Valid Percent
Valid	False match between SAR and LFS	4360	71.65	97.26
	Correct Match between SAR and LFS	123	2.02	2.74
	Total	4483	73.67	100.00
Missing	No name and address from LFS file	1602	26.33	
Total		6085	100.00	

Table 2: Match Indicator for matches between pre-SDC SAR and the LFS

However, this is not the whole picture. In the second stage of the experiment SUDA was run over the SARs file, this represents how an intelligent intruder would use knowledge about the structure of attack file to focus on the higher probability matches. Table 3 shows the frequency of true and false matches for various 7 bands of DIS-SUDA scores for the pre-SDC match file. There appears to be increasing risk as the DIS-SUDA score increases.

The message here is more clearly seen in table 4, where we consider an intruder who uses various confidence thresholds. Using the pre-SDC match file the sophisticated intruder would be able to achieve matching rates of up to 22% using medium threshold matches. These are rates where an NSI might start to be concerned.

DIS-SUDA Band	False matches	Correct matches	% correct	Total
0->0.1	3900	81	2.03	3981
0.1->0.2	218	8	3.54	226
0.2->0.3	111	11	9.02	122
0.3->0.4	77	9	10.47	86
0.4->0.5	33	8	19.51	41
0.5->0.6	10	3	23.08	13
0.6->0.7	1	1	50.00	2
>0.7	10	2	16.67	12
Total	4360	123	2.74	4483

Table 3: Correct and incorrect matches by bands of DIS-SUDA scores for the pre-SDC SAR.

DIS-SUDA Threshold	False Matches	Correct matches	%correct
>0	4360	123	2.7
>0.1	460	42	8.4
>0.2	242	34	12.3
>0.3	131	23	14.9
>0.4	54	14	20.6
>0.5	21	6	22.2
>0.6	11	3	21.4
>0.7	10	2	16.7

Table 4: Match rates achieved by a hypothetical intruder using various thresholds of confidence as measured by the DIS-SUDA score. For the pre-SDC SAR

So what is the impact of the SDC process on these results? Table 5 shows the frequency of true and false matches for various bands of DIS-SUDA scores for the post-SDC match file. Again the match rates do seem to increase as the DIS SUDA rates get higher but as Table 6 shows the impact is nowhere near as marked as with the pre-SDC file. This is undoubtedly mostly the effect of the perturbation, which specifically targeted these high risk records (although not using the particular combination of key variables that were used in this experiment).

DIS-SUDA Band	False matches	Correct matches	% Correct	Total
0->0.1	1577	28	1.74	1605
0.1->0.2	394	8	1.99	402
0.2->0.3	85	2	2.30	87
0.3->0.4	52	8	13.33	60

0.4->0.5	45	2	4.26	47
0.5->0.6	14	2	12.50	16
0.6->0.7	6	0	0.00	6
>0.7	11	1	8.33	12
Total	2184	51	2.28	2235

Table 5: Correct and incorrect matches by bands of DIS-SUDA scores for the post-SDC SAR.

DIS-SUDA Threshold	False Matches	Correct matches	%correct
>0	2184	51	2.3
>0.1	607	23	3.7
>0.2	213	15	6.6
>0.3	128	13	9.2
>0.4	76	5	6.2
>0.5	31	3	8.8
>0.6	17	1	5.6
>0.7	11	1	8.3

Table 6: Match rates achieved by a hypothetical intruder using various thresholds of confidence as measured by the DIS-SUDA score.

It is also noteworthy that the correct matching rate become non-monotonic with respect of the DIS-SUDA score above the threshold of 0.2, this again is the result of the targeting of the high risk records and in essence there is no value in a hypothetical intruder - in this example – operating at a confidence threshold greater than 0.2. To put it another way, the SDC that was employed on the 2001 SARs appears to seriously undermine intrusion attempts based on the fishing method of attack.

So from the point of view of the SARs the results look reassuring. Whether the residual correct matching rates (2.3% overall and less than 10% on the broad fishing type of attack) fall within the boundary of "undue effort" criterion is something for ONS to decide.

4 Discussion

There are various caveats that need to be placed on the interpretation of these results. The first concerns data divergence. The two datasets concerned here are both produced by ONS. Therefore, in general we would expect that there would be lower rates of data divergence, than either would have with a non-ONS dataset. This will tend to inflate the match rates compared to what an intruder would achieve. Against this a very particular form of data divergence would have been present in this case, arising from the data collection. The SARs were generated from the census which was collected on one day in April 2001, whereas the LFS file we used was collected from March to May 2001. This has an impact on the data divergence of the age variable in particular. Someone whose birthday fell between the census date and the collection of their LFS data would have a different age recorded on the two dataset. This form of data divergence affects the pre-SDC

SARs (which is coded in single years) more than the post-SDC SARs (which is not). The impact of this on match rates is difficult estimate exactly but it would be reasonable to assume that approximately one eighth of records would be affected.

A second point to be born in mind in interpreting these results is that we have simulated only one sort of attack, a database cross match of common variables. A sophisticated, determined intruder could use this initial match to pursue further information about potential matches.

Given that, in a real data intrusion, identification information would be present on the attack file, the intruder could go through the list of potential matches and try to extend the match keys for those cases, by attempting to gather further information about the individuals (which would confirm or deny the matches). When we are talking about 6000 possible matches that would obviously be quite onerous but by focusing on the matches with high SUDA scores this might enable the intruder to differentiate the false from true matches[4] quite effectively. This again emphasises the importance of the perturbation in masking the high risk matches.

The third caveat is that an intruder is unlikely to have access to a dataset which has exactly the same coverage to that of the SARs. A different rate of coverage would change the absolute matching rates and would also affect the information available form the data structure of the attack file and therefore the confidence of an intruder in any given match. An attack file with a significantly larger coverage than the SARs 3% would represent a much increased risk.

A final point is that the SUDA algorithm is only one method for assessing record level disclosure risk. Evidence suggests that others such as that of Skinner and Holmes(1998) maybe superior although Shlomo and Barton (2006) found that the SUDA metric produced similar correlations to that of Skinner and Holmes' model based approach, so this is not likely to have affected the overall pattern of results greatly.

Overall then, by comparing the results for the pre and post-SDC SARs files it has been possible to estimate the effect of the SDC process. Generally, the recoding appears to have substantially reduced the total number of matches and the perturbation has effectively masked the higher risk records on the SARS. Looking at the results of linking the pre-SDC SARS with the LFS we can say that high risk records on the LFS do not seem to be sufficiently protected.

There are three processes which would cause variation in the results obtained here; (i) differences in the data divergence rates between the files used in the simulation and that used in an actual intrusion attempt (ii) An intruder using secondary intrusion techniques to confirm matches (or not) (iii) an intruder using an attack file with different coverage than the LFS. Nevertheless, the headline result of this study is that the targeted disclosure control used with the 2001 SARS appears to be very effective in protecting against an attack based on identifying high risk records.

[4] Of course, extending the key would also increase the data divergence, so the intruder would be need to be operating at a high degree of sophistication.

References

Elliot, M. J. and Manning A (2004) A Special Uniques Analysis of the Labour Force Survey. Report to the Office for National Statistics. January 2004.
Elliot, M.J. (2006) Scenario Keys version 3, CAPRI group working paper.
Elliot, M. J. and Dale A (1998) Disclosure Risk for Microdata. Report to the European Union ESP/ 204 62/DG III.
Muller, W; Blien, U.; and Wirth, H. (1992). Disclosure risks of anonymous individual data. Paper presented at the 1st International Seminar for Statistical Disclosure. Dublin 1992.
Skinner, C. J. and Holmes, D. J.(1998) Estimating the Re-identification Risk per Record in Microdata. *Journal of Official Statistics.* Vol. 14 (4) 361-372.

Appendix A: Description of Intrusion Scenarios used as bases for this study.

Scenario IND4a2: Private Database Cross Match

This scenario is based upon an analysis of the information commonly available in restricted access databases, a slightly extended version of 4a with additional, less common variables. Typical variables are:

Age
Sex
Marital status
Number of dependent children
Workplace (typically a geographical identifier)
Distance of journey to work
Number of earners
Tenure
Number of Cars
SOCmajor
Primary economic status

Attacker Profile: Person with access to restricted access dataset or hacker able to obtain such access.

Scenario IND5a: Worker using information about colleagues.

This scenario is based upon a study of what people commonly know about people with whom they worked. Typically this includes considerable detail on economic characteristics, basic physical characteristics and some very crude information about personal circumstances. Typical variables are:

Age
Sex
Ethnic Group
Occupation
Workplace
Distance of journey to work
Industry
Hours
Economic Status
Long Term Illness
Number of residents in Household.

Attacker profile: Anyone working preferably in a large organisation.