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. 19
English only

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

THE IDENTIFICATION OF SPECIAL UNIQUES

Contributed paper

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

Abstract. The paper investigates a method for identifying risky records within a microdata file. The method, *Special Uniques Identification,* involves inferring population uniqueness for a set of key variables from sample uniqueness for a collapsed form of that key variable set. The method is shown to be useful with being considerably higher probability of than the probability of population uniqueness given sample uniqueness on the uncollapsed variables.

I. INTRODUCTION

1. Traditional methods of disclosure risk assessment and control have been based around file level risk metrics, such as population uniqueness and the conditional probability of population uniqueness given sample uniqueness (UUSU ratio). Research into how such metrics behave empirically (for example Bethlehem 1990) has provided some important insights into the key disclosure risk factors. However, more recently, it has been recognised that risk also needs to be analysed at levels other than the whole file (Elliot 2000, Skinner and Holmes 1998, Fienberg and Markov 1998).

2. Elliot et. al. (1998) investigated the effect of geographical detail on disclosure risk in samples of microdata. They found, counterintuitively, that risk levels as measured using the UUSU ratio had a non-monotonic relationship with geographical detail. In other words risk did not increase as expected with greater geographical detail - as exemplified in table 1. [2]

[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).

[2] This finding has subsequently been replicated using a different statistic (the probability of a correct match given a unique match); Elliot (1999).

Table 1: UUSU ratio (expressed as a percentage) by sampling fraction and geographical level for the key age, sex and marital status and a area of population 450,000; from Elliot et al. (1998).					
Level of geographical detail	**% population Uniques**	***sampling fraction***			
		1%	**2%**	**3%**	**4%**
1 (450K)	0.003	0.24	0.74	2.85	6.66
2 (220K)	0.009	0.16	0.39	1.63	4.95
3 (150K)	0.019	0.18	0.37	1.39	4.22
4 (120K)	0.026	0.17	0.33	1.11	3.34
5 (90K)	0.037	0.24	0.45	1.31	3.80
6 (60K)	0.060	0.28	0.50	1.26	3.36
7 (30K)	0.160	0.45	0.79	1.72	3.51

3. This finding led to the development a model with related properties which distinguishes two types of unique records:

- *special uniques* (records which are unique by virtue of some epidemiologically unusual combination of characteristics - for example, a sixteen year old widow)
- *random uniques* (those which are unique as an arbitrary consequence of the coding regime employed); Elliot et al (1998).

4. Conceptually, the model explained the non-monotonicity as follows:

i) Special uniques within the sample are likely to be sample uniques irrespective of geographical detail.
ii) Special uniques within the sample are likely to be sample uniques irrespective of sampling fraction (and therefore they are more likely to be population unique).
iii) Random uniques are likely to be sensitive to both sampling fraction and geographical detail and hence random sample uniques are less likely to be population uniques.
iv) Increasing geographical detail leads to a rapid increase in the number of random uniques within the sample but has smaller impact on the level of population uniques.
v) It follows from i) to iv) that: increasing geographical detail will lead to an increase in random sample uniques but relatively little change in the level of special uniques and since random sample uniques are less likely to be population uniques we would expect that the change in the UUSU ratio would be non-monotonic.

5. This model has other consequences. From point i) and ii) above, one can make some important predictions. Since special uniques are insensitive to geographical detail it can be inferred that persistence of uniqueness through geographical aggregation may be a defining feature of special uniques. Similarly, since special uniques are also insensitive to sampling fraction one might also be able to infer population uniqueness from special uniqueness. Linking these two propositions together yields the following testable proposition:

> **Proposition 1: Records that are sample unique at a coarse level of geographical detail are more likely to be population unique at a finer level of geographical detail than sample uniques at the finer geographical level (as measured by the UUSU ratio).**

6. This is shown graphically in figure 1.

Figure 1: Schematic representation of the greater probability of population uniqueness given special uniqueness.

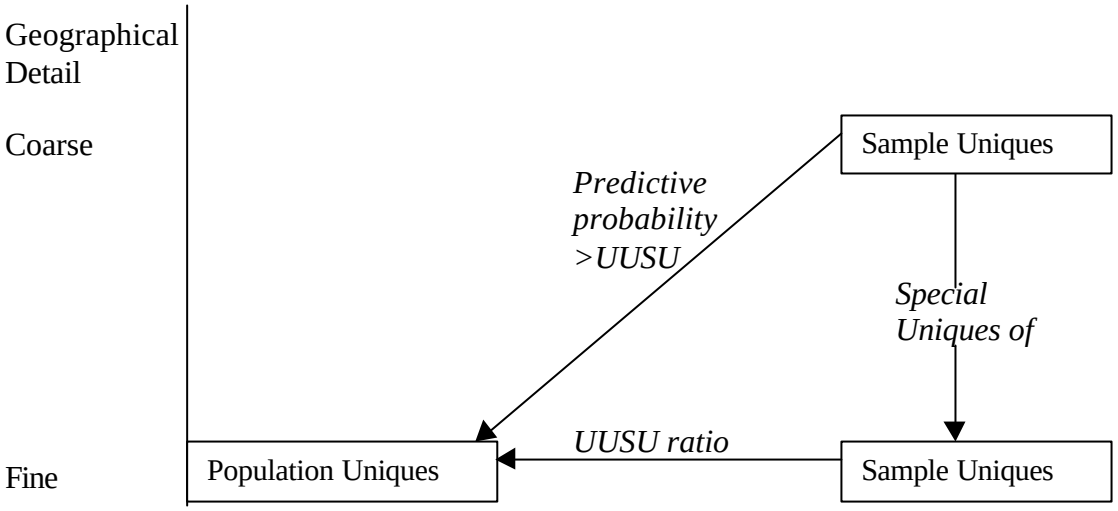

Table 2: Two hypothetical keys showing number of bands		
	Key A	Key B
Age	19	94
Sex	2	2
Marital Status	5	5

7. By induction, it would seem likely that this predictive effect of collapsing geographies would apply to other variables as well. So that sample uniques on key A in table 2 would be more likely to be population uniques on key B than sample uniques on key B.

> **Proposition 2: Records that are unique within a sample using broad variable codings are more likely to be population unique with more detailed variable codings than sample uniques with the more detailed codings.**

8. The remainder of this paper describes empirical work that tests propositions 1 and 2.

II. EMPIRICAL DEMONSTRATION

The Dataset Used

9. In order to perform this analysis population data was needed. ONS supplied anonymised 1991 Census data for seven UK Local Authorities to CCSR under contract.[3] The procedure has been applied to the four largest of these. The results presented here are for one LA with a population of approximately 450,000.

10. The method involved first extracting all parallel, systematic samples from the population data file (which was geographically stratified down to the level of census enumeration district). So the population file was divided into fifty 2% samples, twenty 5% ones, etc.

[3] The data were kept in a secure environment, access was limited to the researcher for the designated parts of the project, and the data has been returned to ONS. For the purpose of the contract, the CCSR was a supplier of services to the Registrar General for England and Wales and the 1991 Census Confidentiality Act applied.

Geographical coding

11. The geographical coding scheme used was that developed by Elliot et al (1998), shown in table 3. The key levels for this study were levels 1 and 7

A note on standard key variables

12. In the description of research that follows, *standard key variable* (SKV) definitions are used. These are a set of variable definitions that have been derived, (through surveys and direct investigations) to reflect the contents of databases and other sources of information (Elliot 2000b).

13. Each SKV has a name consisting of two components: a word, which gives an indication of the variables meaning, and a number, which indicates the number of values that the variable may take. For example, ETHNIC10 is a ten point ethnicity variable.

Table 3: Levelled geographical coding scheme from Elliot et al (1998)

Level	Label	Approximate population size per unit	Divisions of LA used in this study
1	Large SAR area[4]	450K	Whole LA
2	Medium SAR area	220K	Two
3	Small - medium SAR area.	150K	Three
4	SAR threshold area	120K	Four
5	Parliamentary constituencies	90K	Five
6	medium ward groupings	60K	Seven
7	small ward groupings	30K	Fifteen

14. The following sections also refer to the *basic key*. This is the key consisting of three SKVs: SEX2, AGE94 (age in single years from 0 to 90, grouped from 91-94 and top coded from 95) and MARCON5 marital status (the UK census definition of marital status).

15. The choice of the cross-classifications presented here is partly constrained by the limited range of variables that were 100% coded on the 1991 census file. In fact many different cross-classifications have been tested, all of which yield similar results to those presented here.

Method

16. The following method was used. For each set of key variables and for each of the samples taken from the population file:

(i) The sample uniques (SU) at highest level of geography available (level 1:450K) were identified.

(ii) The number of those that were population uniques (SPU) at the lowest level for which full geographies were available (level 7: 30K) was counted.

(iii) The results were expressed as the across sample mean SPU/ mean PU and these results were compared to the equivalent mean UUSU ratio.

17. Tables 4-6 show the number of sample uniques at level 1 (SU) and the number of these that were population uniques at level 7 (SPU). As it can be seen the results indicate that sample uniqueness at high level of geography is highly predictive of population uniqueness at the lower level when compared to the UUSU ratio at the lower level. These results are typical of those found across the dataset.

[4] The SAR areas in this table refer to the geographical coding for the 2% Individual Sample of Anonymised Records from the 1991 British Census. 120,000 was the threshold for the size of such areas.

Table 4: Identifying special uniques, an example using the basic key +ETHNIC10				
Sample %	UUSU level7	Mean SU (level 1)*	Mean SPU (level 7)*	SPU/SU(%)
2	5.02	404	112	27.70
5	9.65	526	250	42.87
10	16.47	617	362	58.67

Table 5: Identifying special uniques, an example using the basic key +PRIMECON11				
Sample %	UUSU level7	Mean SU (level 4) *	Mean SPU (level 10) *	SPU/SU(%)
2	7.51	703	180	25.57
5	9.72	788	393	43.49
10	14.96	808	480	59.41

Table 6: Identifying special uniques, an example using the basic key +PRIMECON11 and ETHNIC10				
Sample %	UUSU level7	Mean SU (level 1) *	Mean SPU (level 7)*	SPU/SU(%)
2	10.29	1190	397	33.40
5	15.68	1656	822	49.62
10	22.38	2029	1311	64.61

18. This exercise was then repeated with variables other than geography. Holding geography constant but aggregating age to five-year bands produces the results such as those shown in table 7. The general thrust of these results is similar to those found for geographical aggregation. The strength on the effect depends on the degree to which the variables are collapsed.

Table 7: Identifying special uniques, an example using the basic key +ETHNIC10, aggregating on age from single year to 5 year bands				
Sample %	UUSU	Mean SU (AGE19)*	Mean SPU (AGE94)*	SPU/SU(%)
2	6.39	852	153	17.95
5	9.65	975	305	35.28
10	13.28	1148	560	48.78

19. Table 8 indicates that as you aggregate over multiple variables the number of sample uniques reduces but the proportion that is population unique in the unaggregated data increases markedly. This indicates that this multiple aggregation technique is a very effective way of identifying really risky records.

Table 8: Identifying special uniques, an example using the basic key +ETHNIC10, aggregating on age from single year to 5 year bands and geography(level 7 to 1)				
Sample %	UUSU	Mean SU(a19/l1)*	Mean SPU(a94/l7)*	SPU/SU(%)
2	6.39	135	78	57.78
5	9.65	147	96	65.31
10	13.28	174	143	82.18

* These figures are rounded to the nearest integer for presentation purposes SU stands for sample uniques and PU for population uniques

20. In principle the special uniques method allows the identification of the level of risk associated with each record. The more variables in which a unique survives aggregation, the higher the risk, the next stage would be to develop a method of scoring special uniques according the number of single and multiple variable aggregations in which uniqueness is maintained. This would provide a scale of riskiness.

21. Notwithstanding the above, it is possible to obtain an idea of the impact of controlling special uniques would have on file level risk metrics. Using the probability of a correct match given a unique match - pr(cm|um) - (Elliot 2000) and making the assumption that manipulating the special uniques would reduce the number of correct matches by the number of population uniques within the specials, one can obtain a crude idea of the scale of their impact. Table 9 shows how this works. Using the special uniques identified in Table 7 the reduction in pr(cm|um) are between 40-60% depending on sample size. The smallest sampling fractions show the greatest impact but they also have the largest proportion of affected records; in table 8, 1.6% of the 2% samples are classed as specials where as only 0.4% of the 10% samples are.

Table 9: The impact of controlling special uniques, an example using the basic key +ETHNIC10, aggregating on age from single year to 5 year bands and geography(level 7 to 1)

Sample %	Mean SU	Mean pr(cm|um) before controlling specials	Mean pr(cm|um) after controlling specials	%after/before
2	135	0.028	0.012	42.8%
5	147	0.067	0.033	49.2%
10	174	0.118	0.071	60.1%

22. With greater precision through the sort of scaling discussed above the number of affected records could be reduced and/or the scale of the impact could be increased. However the results here are very encouraging.

Future Directions

23. The empirical work reported in this paper indicates using a sample of simple keys that the concept of special uniqueness is useful. However, at present in order to fully utilise the concept in disclosure control, a method is needed that fully identifies all special uniques within a dataset. The combinatorial explosion involved in considering all variable combinations makes the solution of this problem with traditional linear programming techniques infeasible.

24. A second issue is one of grading. As presented, special/random uniqueness is a dichotomy. However, clearly this is an expository and methodological convenience. In reality 'specialness' must be a matter of degree and therefore the possibility of accurately measuring degrees of specialness should be investigated.

25. In order to deal with these issues, a multi-disciplinary team has been formed with the aim of using expertise in both Disclosure Control and Data Mining.[5] In particular we will be investigating the possibility of using Data Mining Techniques to effectively control the search space of key variable combinations. Data mining is a generic term for software tools that aim to find useful and implicit patterns from very large datasets.

26. In related work, data mining tools have already been used to discover unusual records, in a process known as *outlier detection.* For example, the JAM system identifies attribute patterns consistent with known fraudulent behaviour for financial institutions; Stolfo et. al. (1997). Breunig et. al. (1999) has shown it is possible to distinguish between local and global outliers.

[5] The team consists of the author and Dr. A Manning of the Centre for Novel Computing at the University of Manchester.

27. In a specific sense then, we propose to build on this work by designing and implementing a search algorithm that will guarantee to find all special uniques in anonymised microdata. The method will have two parts, the first involving a search strategy for identifying a set of potentially 'risky' records and the second for grading the risk of each member of this set.

28. Looking still further into the future we envisage that this system will be linked to the DIS system for file (and intermediate) level risk evaluation; Elliot (2000) and then tied to a specific methodology for dealing with records of various degrees of riskiness. This will enable precise targeting of disclosure control and consequent increase in both data quality and security.

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,

Breunig S., Kriegel H.-P., Ng R., Sander J., (1999) 'Identifying Local Outliers', Proceedings of the 3rd European Conference on Principles of Knowledge Discovery (PKDD'99), Prague, Czech Republic.

Elliot, M. J. (1999). New approaches to Statistical Disclosure Risk. *Paper presented to International Conference of the Royal Statistical Society*, Warwick, UK, May 1999.

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

Elliot, M. J. (2000b) Standard Key Variable Definitions version 3.0; Centre for Census and Survey Research internal working document.

Elliot, M. J., 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 53-58.

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

Stolfo S. J, Prodromidis A. L., Selepis S. T, Lee W., Fan D. and Chan P. K. (1997), 'JAM: Java Agents for Meta-learning over Distributed Databases, *Proceedings of KDD-97*