CASC PROJECT

Computational Aspects of Statistical Confidentiality

December 2003

TM2 - A Case Study of the Impact of Statistical Disclosure Control on Data Quality in the UK Samples of Anonymised Records

Dr. M. J. Elliot
Dr. K. Purdam

CCSR
University of Manchester
Oxford Road
Manchester M13 9PL
Tel: 0161-275 4721
Fax: 0161-275 4722

CASC Project Deliverable 5D2:

TM2 - A Case Study of the Impact of Statistical Disclosure Control on Data Quality in the UK Samples of Anonymised Records

Dr. M. J. Elliot
Dr. K. Purdam

Dec 2003

CCSR
University of Manchester
Oxford Road
Manchester M13 9PL
Tel: 0161-275 4721
Fax: 0161-275 4722

CONTENTS

1.Introduction

Much of disclosure risk research focuses on the control side of the disclosure issue, asking: "what do we need to do in order to make this data safe?" However, this question is only one side of the problem that a data provider faces in controlling for risk. All risk control methods degrade the data to some extent and therefore reduce the ability of data users to conduct the analyses they need for their legitimate purposes. These effects fall into two categories: ***1. Reduction of analytical completeness.*** Some control methods, typically the recoding of taxonomic schemes into coarser categorisations, mean that analyses that could have been conducted with unrecoded data cannot be done. An example is the use of geographical thresholds in microdata sets leading to smaller administration units being grouped together, preventing researchers within those units from effectively using the dataset. ***2. Loss of analytical validity***. The loss of analytical validity is harder to define, but in some ways more critical because of its insidious nature. Technically, loss of validity can be said to occur when a disclosure control method has changed a dataset to the point where a user reaches a different conclusion from the same analysis.

Discussion of these two issues is at present pre-theoretical. Recent work has attempted to metricise the concepts, see for example Sebe et al (2002), Cox (2003).

However, no principled computational method has been established for the practical assessment of their impact. The development of such a method is vital to improving the efficiency of disclosure control techniques, which are at present haphazard in respect of their analytical consequences. In this research we go some way redressing this lack by categorising the effects on analytical power of several disclosure control techniques and by examining the feasibility of developing methods for measuring the scale of such effects.

2. Methodology

Phase I: Data Selection.

To turn this complex issue into a tractable problem, we have used data available from the 1991 UK census as trial datasets. Specifically we used the 1991 Samples of Anonymised Records (SARs) which are publicly available sets of microdata from the UK Census www.ccsr.ac.uk/sars The SARs contain information on a range of topics including age, gender, ethnicity, household size, household type, employment and health.

The SAR datasets are widely used in research (Li 2004). The use of this particular dataset also enabled the research team to build on work conducted in preparation for the 2001 census surveying the uses made of UK census microdata as well as many years of work analysing disclosure risk with such data.

A typical set of analyses was constructed through a literature review of published analyses using the SARs and through a user survey. These were selected on the basis of providing a good range of variables used and type of analyses conducted.

[1] CCSR, University of Manchester, UK (www.ccsr.ac.uk). This work was conducted under EU grant EU: IST 2000-20569

Phase II: Questionnaire study of the impacts of recoding

An initial survey introduced the research to the authors and asked if they would be willing to assist by re-running of their analyses using data that had been subject to further disclosure control particularly using the software Argus.[2] From these responses and ongoing literature reviews the authors were re-contacted and asked to complete a short questionnaire which interrogated the likely impact of various possible recodes would have on their analyses in relation to their use of the SAR data. A copy of this questionnaire can be found in Appendix 1.

Potential recodes were based upon those that have been suggested for use with the Small Area Microdata from the 2001 census (see Tranmer et al 2004) and were structured in three different forms. (i) one a variable is removed from the dataset, (ii) an near-interval variable is banded (iii) a categorical variable is regrouped. Examples of each are shown below.

(i) **Area** removed from data set but region left in.
(ii) **Age** recoded from single years to Five-year bands.
(iii) **Ethnicity** recoded from 10 to 4 categories:
–a. White
–b. Black
–c. Asian
–d. Other

In all, twenty-nine possible recodes were suggested. For each possible recode Authors were requested to give one of four possible responses:

A. This change **would not affect** the analyses I conducted for this paper.
B. This change **would moderately affect** the analyses that I conducted in this paper.
C. This change **would severely affect** the analyses I conducted in this paper.
D. Other (please indicate the meaning of this in the comments section).

The results of this were collated and are described in section 3.

Phase III: Reanalysis with perturbed data

The decision about how to conduct the perturbation study was one of the more problematic aspects of this work. ARGUS is a disclosure control tool rather an automated disclosure control system. As such it leaves decisions about key variable combinations and parameter selection to the user. As the data we were using was unweighted, the risk model did not apply and therefore we had no means internal to Argus of making decisions on the basis of levels of risk. A further problem was that it was not possible to use full scenario based keys (as developed by Elliot and Dale 1999).[3]

[2] Argus is the EU approved disclosure control software, which has been developed by The CASC consortium.

[3] A further more general problem with ARGUS is that it is not possible to block missing values for use in perturbations. This means that inconsistencies are produced where not applicable categories are used to record suppressions or as Post randomisation categories.

For consistency with other work on this data set (Elliot and Manning 2003), various combinations were produced experimentally and programs extracting the records and variables thus identified as risky were compared with the outputs of the special uniques program. The following variable combination frames appeared to identify risk in a similar way to that program:

A. All individual variables (threshold=4)
B. All pairs of variables (threshold=2)
C. All 3-way combinations under scenarios (threshold=1).

Four SAR datasets were then produced.

File A. Suppression Based File.

On this file the disclosure control was entirely based around suppressions. All three combination levels were used, to determine the suppressions. The default suppression weights were employed.

File B. *PRAM File*

All variables on file PRAMed. The per value change probabilities of PRAM were set to maintain the univariate distributions.[4] For some variables such as age bandwidths were used partly to control the number of inconsistencies.

File C. *Control File*

No perturbation test to see if the author was able to replicate results with original data.

File D. *Combined Pram and suppressions.*

Suppressions were applied to the PRAM FILE C, with only level A and B combinations being used.

These datasets were then sent out to the authors who were asked to comment on the effects on their analyses. See questionnaires in Appendix 2. Where authors were unable to continue to participate in the study their analysis was re-run where possible by the research team.

3. Results

3.1 Recoding Questionnaire

Twenty-three authors returned their recoding questionnaires. All authors suggested that their analyses would be affected by one or more of the potential recodes. Figure 1 breaks down the number of recodes impacting upon analyses by author. As it can be seen nearly a third were affected by thirteen or more recodes.

[4] See De Woolf at all 1998 for description of the PRAM method.

Figure 1 Number of recodes impacting on analyses per author.

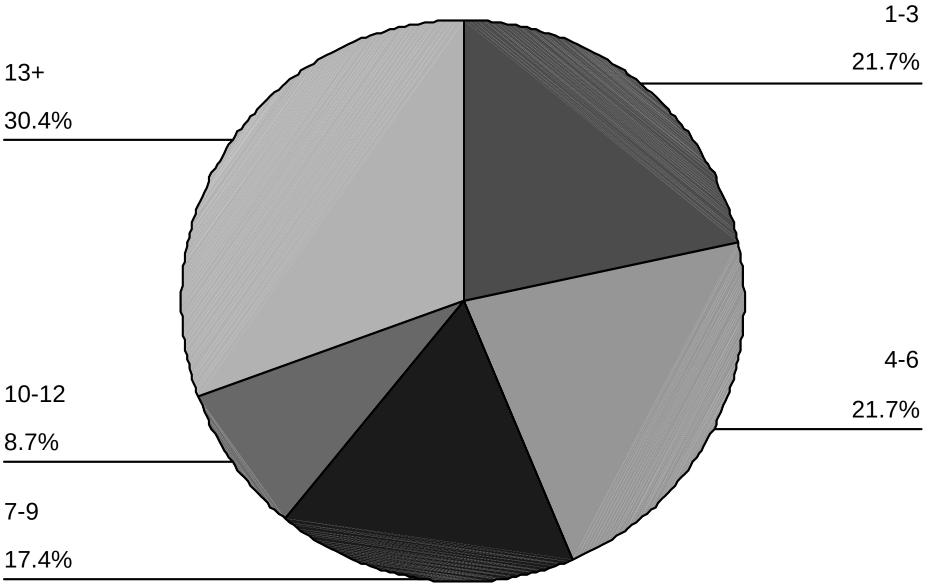


Figure 2 Number of recodes severely impacting on analyses per author.

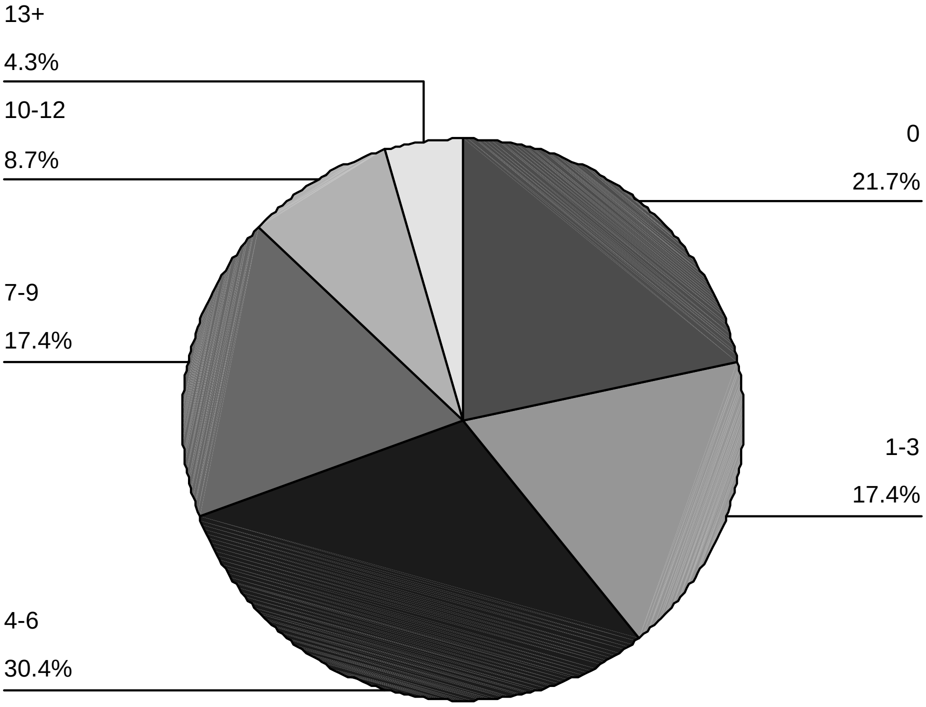


Figure 2 shows the number of recodes **severely** affecting analyses again broken down by author. About a fifth of authors say that their analyses would not be severely affected by any of the suggested recodes, and over 60% say they would be severely affected by four or more.

Turning now to specific variables. Two variables that are often considered for recoding are age and geographical detail. These two variables have a large number of categories and are universal keys in that they tend to be included in all scenario keys (Elliot and Dale 1999). Therefore recoding age and geographical detail is likely to have a strong disclosure risk impact (both as measured and implicit).

Figure 3 shows the impact of recoding age into ten-year bands. This indicates that about a third would be severely affected by the recode. The milder recode into five-year bands has considerably smaller effect with 60% of authors responding that the recode would not affect their analyses (see Figure 4). This indicates the potential value of consultation exercises along the lines of this study. The considerably milder impact of the five-year banding suggests that the playoff between usability and disclosure risk is complex. This is further illustrated by comparing the two possible recodes of geographical detail. Figure 6 shows the impact of recoding 278 areas area to 4 countries (England, Scotland, Wales and Northern Ireland). Compare this to figure 5, which shows the impact of recoding to the twelve standard UK regions.

Both recodes have a strong effect on usability, but in fact there is little difference between them, which indicates that retaining the extra detail that region provides very little in the way of usability, this itself is interesting because many UK microdata files are coded to regional level. A broader study of the use of this particular variable would be necessary before any firm conclusion could be drawn, however.

Figure 3 Percentage of authors giving to each category of response to whether recoding age into ten-year bands would affect their analyses.

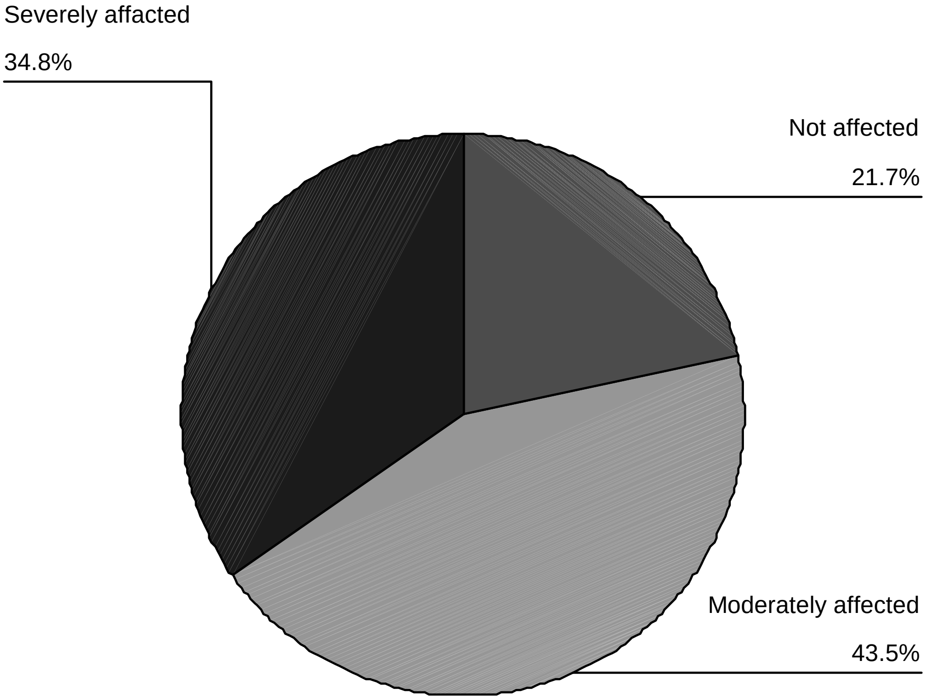


Figure 4. Percentage of authors giving each category of response to whether recoding age into five-year bands would affect their analyses.

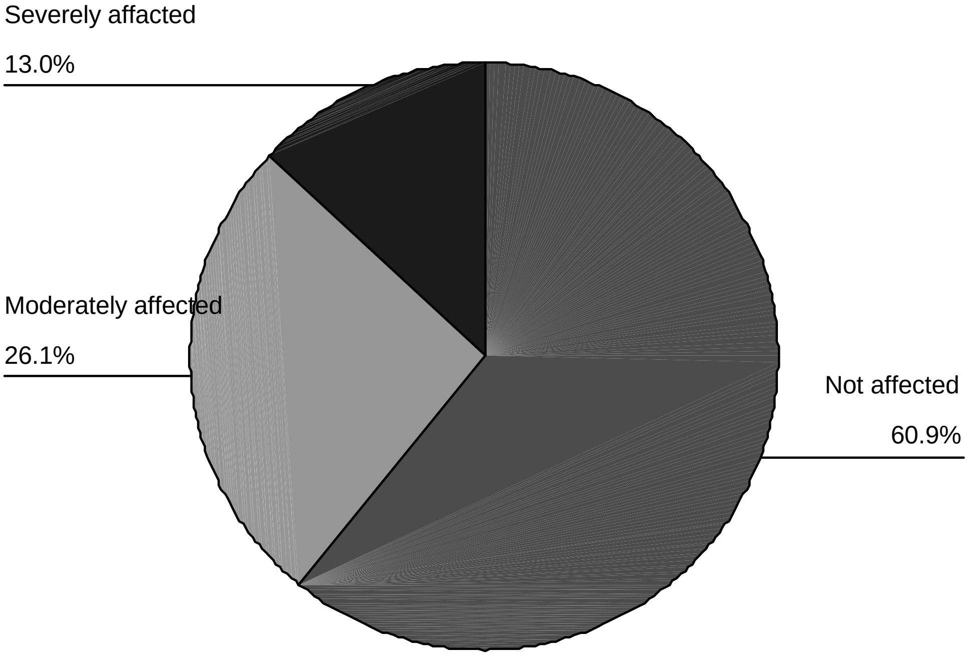


Figure 5. Percentage of authors giving each category of response to whether removing area but retaining region would affect their analyses.

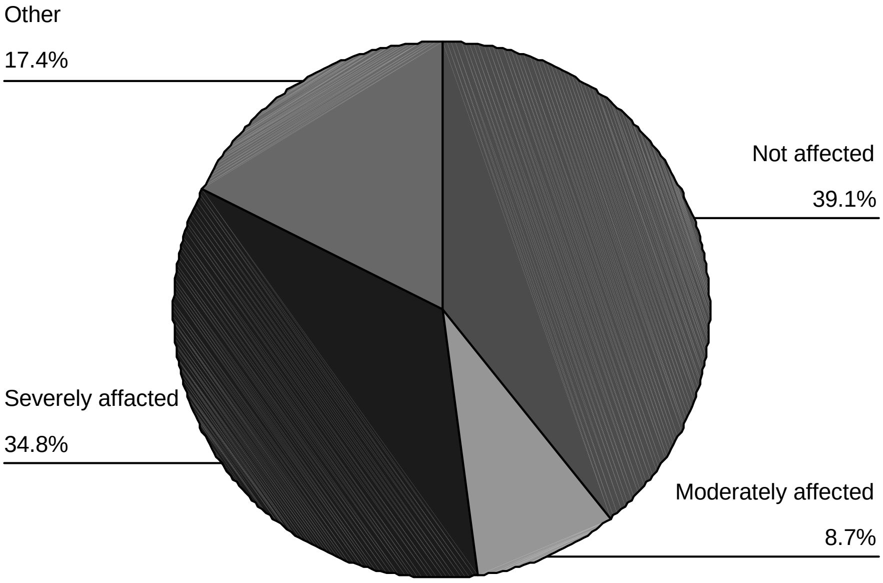
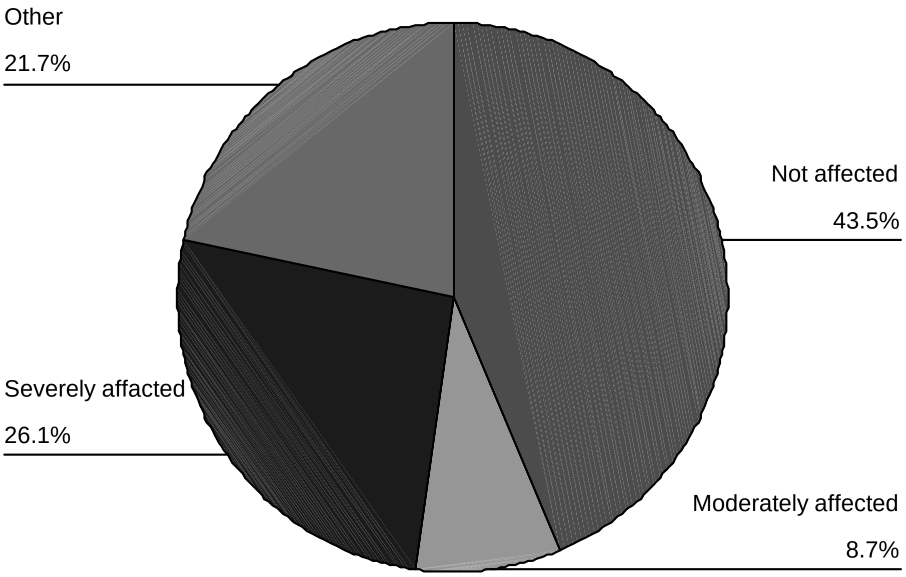


Figure 6. Percentage of authors giving each category of response to whether removing area and region but retaining country would affect their analyses.



Two other variables, which might be considered as risky are the skewed variables ethnic group and country of birth. Skewed variables give rise to risky records although not necessarily to high file-level risk (Elliot and Manning 2003).

Recoding ethnic group to four bands from ten, by consolidating minor categories has a dichotomous effect with about half of the authors saying the recode would have no effect and most of the reminder indicating that it would severely affect their analysis (see figure 7), this grouping basically divides the papers into whether ethnicity was a major part of the analysis or not, indicating that the finer detail was vital where the dataset was being used for its ethnicity variable.

A similar pattern can be observed with recoding Country of birth to two categories (UK/other) from 42, as indicated in figure 8. The majority of authors' analyses would not be affected by this recode but where they were the effect tended to be severe. Recoding to four categories (England, Other UK, Europe, Other) rather than two doesn't really help with the results in figure 9, similar to figure 8. As with the geographical level it is the really detailed coding which gives the data its utility, the differences between degrees of course coding are relatively minor.

Of the remaining recodes many appear to have only a small impact on the authors analyses. Table 1 shows the category breakdown for each recode; only the recodes for tenure, socio-economic group, family type, and economic status appear to have a marked effect.

Figure 7. Percentage of authors giving each category of response to whether recoding ethnicity from ten categories to four would affect their analyses.

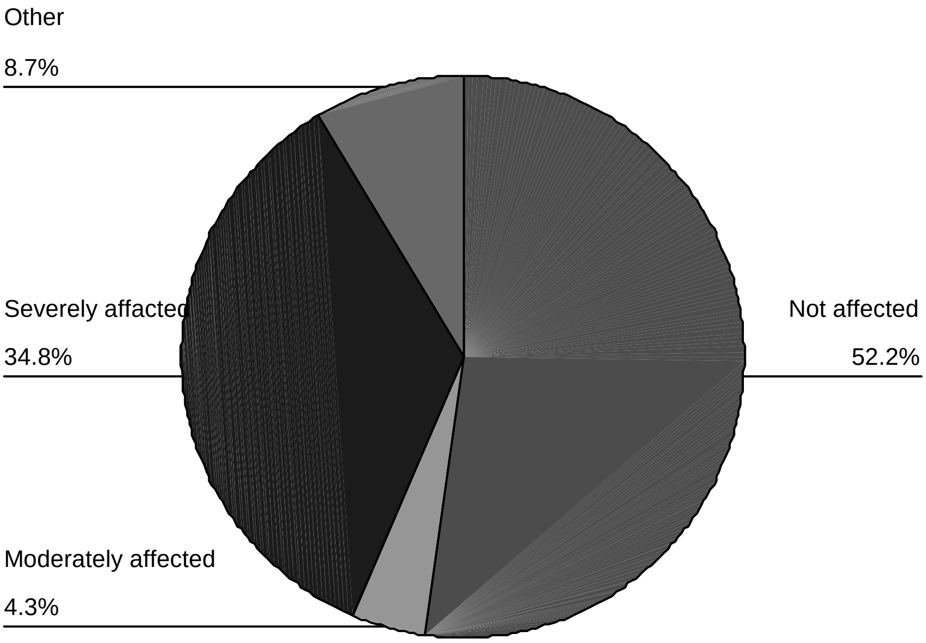


Figure 8. Percentage of authors giving each category of response to whether recoding country of birth from 42 categories to two would affect their analyses

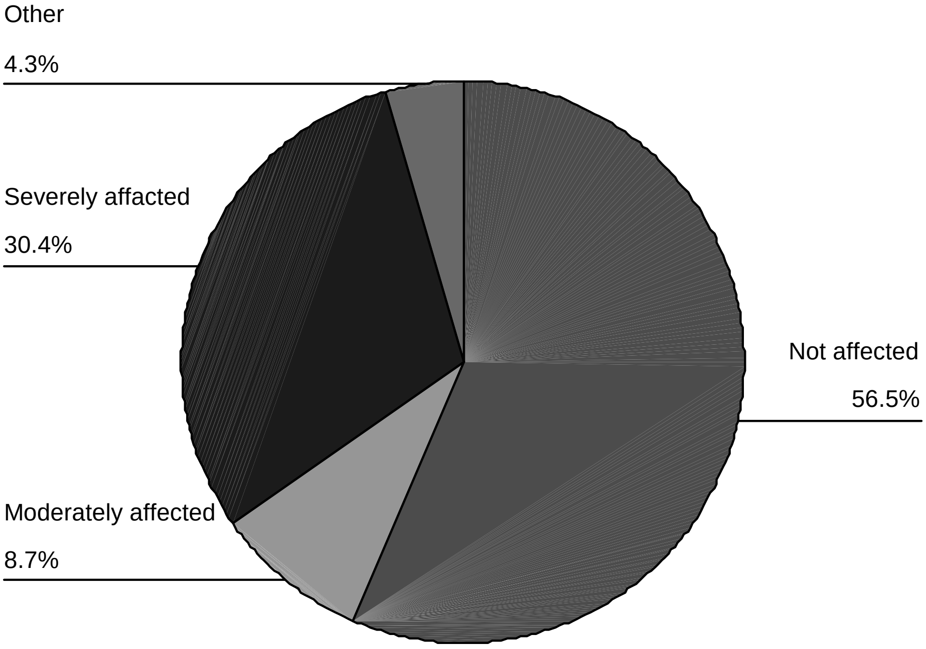


Figure 9. Percentage of authors giving each category of response to whether recoding country of birth from 42 categories to four would affect their analyses.

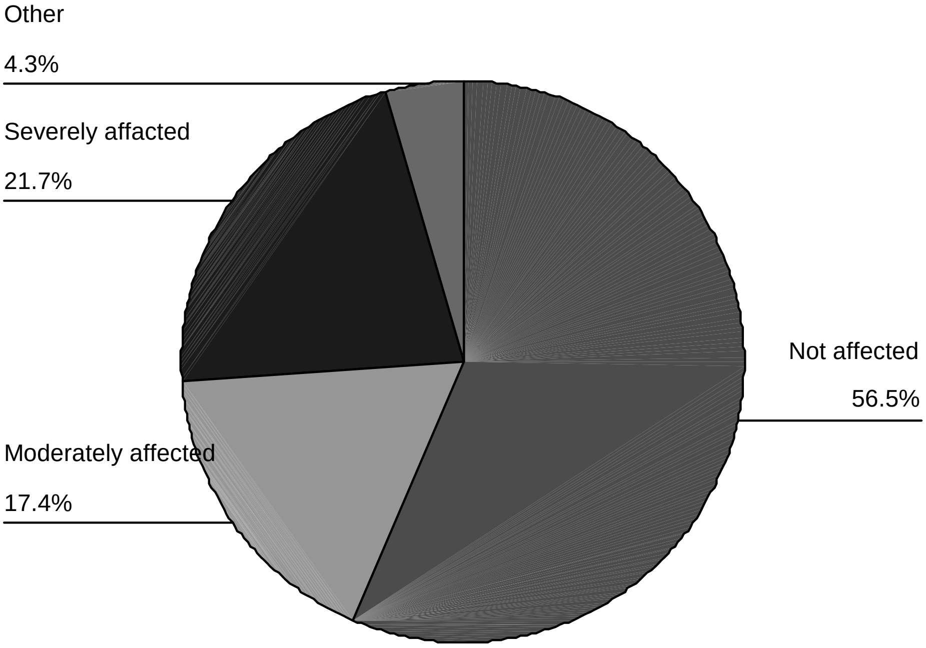


Table 1: Proportion of respondents indicating each category of impact for all twenty nine recodes and the resultant utility index for the data after recode							
Variable	**From**	**To**	**None**	**Moderate**	**Severe**	**Other**	**Utility index**
Age	Single years	Five year bands	60.9	26.1	13.0	0.0	74
Age	Single years	Ten year bands	21.7	43.5	34.8	0.0	43
Area	278 areas	12 regions	39.1	8.7	34.8	17.4	52
Area	278 areas	4 countries	43.5	8.7	26.1	21.7	59
Country of birth	42 categories	2 categories	56.5	17.4	21.7	4.3	67
Country of birth	42 categories	4 categories	56.5	8.7	30.4	4.3	63
Ethnic group	10 categories	4 categories	52.2	4.3	34.8	8.7	59
Distance of move	14 categories	3 categories	78.3	8.7	13.0	0.0	83
Distance to work	9 categories	5 categories	91.3	0.0	8.7	0.0	91
Primary economic status	10 categories	4 categories	56.5	13.0	30.4	0.0	63
Secondary economic status	8 categories	Omit	82.6	8.7	8.7	0.0	87
Family type	8 categories	3 categories	52.2	17.4	30.4	0.0	61
Work hours	Single hours	4 bands	91.3	4.3	4.3	0.0	93
Work hours	Single hours	Top coded at 50	91.3	4.3	4.3	0.0	93
Industry	61 categories	9 categories	82.6	17.4	0.0	0.0	91
Marital status	5 categories	3 categories	65.2	17.4	17.4	0.0	74
Occupation	73 categories	9 categories	69.6	26.1	4.3	0.0	83
Number of highest qualification	3 categories	Omit	73.9	8.7	17.4	0.0	78
Level of highest qualification	3 categories	2 categories	73.9	13.0	13.0	0.0	80

Subject of highest qualification	35 categories	Omit	87.0	8.7	4.3	0.0	91
Relationship to household head	8 categories	4 categories	78.3	8.7	13.0	0.0	83
Socio-economic group	17 categories	Omit	52.2	13.0	30.4	4.3	61
Term time address	4 categories	Omit	95.7	0.0	4.3	0.0	96
Method of transport to work	10 categories	5 categories	95.7	0.0	4.3	0.0	96
Work place	5 categories	Omit	87.0	4.3	8.7	0.0	89
Number of cars in household	4 categories	3 categories	82.6	8.7	4.3	4.3	89
Dwelling space type	14 categories	5 categories	91.3	4.3	4.3	0.0	93
Number of residents per room	5 categories	3 categories	87.0	4.3	8.7	0.0	89
Tenure	10 categories	3 categories	60.9	4.3	30.4	4.3	65

The table also includes a utility index. This is simply derived from %none+ (%moderate + %other)/2 and gives a useful overall indicator of the impact of the recode.

It is interesting to compare this index to the disclosure risk impact of the recodes. Using the additional impact method with the data intrusion simulation system of disclosure risk analysis (Skinner and Elliot 2002) one can express the impact (DRI) of each recode on the probability of a correct match given a unique match for a base key plus the recoded variable. The DRI figure is effectively the residual disclosure risk after recoding expressed as proportion of the original risk level. Examples for the central recodes are shown in table 2. The impact of each recode is to reduce the risk as measured by DIS, by between 10 and 50% for these small keys including that variable.[5]

Table 2: A DIS analysis showing the probability of a correct match given a unique match of the SARs using a base key (basic = age94, sex2, marital status5) + a selection of other variables before and after recoding

Key	Recoded variable	Categories bef>aft	SARS	Recoded	Impact
Area,Age,sex,mstatus,Ocupation	OCCUPATION	74->10	0.055	0.025	0.459
Area,Age,sex,mstatus,Industry	INDUSTRY	63->10	0.049	0.026	0.524
Area,Age,sex,mstatus,hours	HOURS	73->50	0.044	0.038	0.864
Area,Age,sex,mstatus,cobirth	COBIRTH	42->2	0.041	0.038	0.927
Area,Age,sex,mstatus,primecon	PRIMECON	10->4	0.028	0.021	0.766
Area,Age,sex,mstatus,tenure	TENURE	10->3	0.028	0.022	0.802
Area,Age,sex,mstatus,ethnic	ETHNIC	10->4	0.023	0.020	0.870
Area,Age,sex,mstatus,primecon	Age	93->10	0.028	0.020	0.726
Area,Age,sex,mstatus,primecon	Age	93->20	0.028	0.021	0.753
Region,Age,sex,mstatus,primecon	Geography	273->12	0.028	0.020	0.711

Expressing this as ratio of the utility index gives a useful indicator of the disclosure risk value of the recode against its utility cost. This information is given in table 3. Clearly these figures are not general as they are derived from one small, *ad hoc* study. However, they do serve as indicators of the form of the relationship between utility and disclosure risk costs and also demonstrate a method for analysing this. In terms of these results it is clear that recoding the variable "industry" has a much better cost benefit ratio than say country of birth. The interpretation of these data should be

[5] Other evidence (Dale and Elliot 2000) shows that as the size of the key increases the impact of the recode decreases. So these figures could be viewed as an upper estimate of the benefit of each recode.

conducted carefully (even if we had a much larger selection of research as an input). For example, these simple ratios do not take account of say the importance of a key variable to a would-be data intruder (Elliot and Dale 1999), a factor that would need to be considered when assessing the relative values of say geography and other variables. Clearly further work is needed.

Table 3: Relationship between utility index and disclosure risk impact

Variable	**From**	**To**	**Utility index (UI)**	**Disclosure risk impact (DRI)**	**UI/ (DRI*100)**
Age	Single years	Five year bands	74	0.75	0.98
Age	Single years	Ten year bands	43	0.73	0.59
Area	278 areas	12 regions	52	0.71	0.73
Country of birth	42 categories	2 categories	67	0.93	0.72
Ethnic group	10 categories	4 categories	59	0.87	0.68
Primary economic status	10 categories	4 categories	63	0.77	0.82
Work hours	Single hours	Top coded at 50	93	0.86	1.08
Industry	61 categories	9 categories	91	0.52	1.74
Occupation	73 categories	9 categories	83	0.46	1.81
Tenure	10 categories	3 categories	65	0.80	0.81

3.2 Reanalysis of perturbed data

Unfortunately the amount of work required to replicate analyses four times meant that many of the original researchers were unable to reproduce their work for part two of the study. In order to incorporate those studies the authors have replicated the studies themselves. This is clearly a less than perfect solution, since it introduces a new variability into the interpretation however as the study is illustrative only, it is probably adequate for current purposes. So far ten studies have been replicated in this way. Of the original twenty-three many were excluded either because they were impossible to replicate from the original paper (usually because the procedure was unclear), or the interpretation was too complex to carry out with out the original researcher's intervention, some further analyses are in progress. For each of the recoded files the authors/researchers were required to give a four-point estimation of the impact of the perturbation on their analysis:

> Overall would you say that the results in your paper and the interpretation of the results were: [please tick one]
>
> Unaffected
> Moderately Affected
> Severely Affected
> Other

In no cases was "other" category used, leaving a three-category measure. The frequencies for the four files are shown in table 4. A severe effect indicated that the results of analyses were sufficient different that many of the conclusions were affected. Moderate affects tended to indicate a change in emphasis rather than a completely different finding, whereas no effect indicates that the figures may have been slightly different but the overall pattern was not, indicating the same conclusion would be consistently drawn.

Table 4: Author/Researcher Description of effect of perturbations by suppression method used. Ten example studies.

		Affect of perturbation		
File	**Perturbation method**	**None**	**Moderate**	**Severe**
A	Suppressions	5	5	0
B	PRAM	2	7	1
C	None	10	0	0
D	Both	1	5	4

The example studies here obvious represent a small selection and therefore for no firm conclusions can be drawn however it is indicative that the perturbations applied by the ARGUS system can have a significant impact on the outcome of analyses conducted using them. Again more research is needed.

4. Conclusions

This research has allowed an empirical investigation of the feasibility of assessing the impact of disclosure control techniques on analytical power an initial categorisation of the effect on those analyses of the application of SDC methods has been developed.

The work is being taken forward to consider the plausibility of generalised metrics of analytical power, which will then be assessable for their relationship with disclosure risk impact. Further research is necessary to look at the relationship in detail.

References

Dale, A. and Elliot, M. J., (2001) 'Proposals for the 2001 SARs: an assessment of disclosure risk.' *Journal of the Royal Statistical Society, Series A*; 164(3), 1-21

De Wolf, P.P Gouweleeuw, J., Kooiman, P. and Willenborg, L. (1998) Reflections on PRAM *Proceedings of the conference on Statistical data protection 98*, Lisbon, March 1998

Elliot, M. J., and Dale, A. (1999) 'Scenarios of Attack: The Data Intruder's Perspective on Statistical Disclosure Risk.' *Netherlands Official Statistics*. Spring 1999. pp 6-10.

Li, Y. (2004) Samples of Anonymised Records from the UK Censuses: A Unique Source for Social Research. Forthcoming in *Sociology* 2004.

Sebe, F. Domingo-Ferrer, J., Mateo-Sanz, J. M. and Torra, V.(2002) in J. Domingo-Ferrer(ed) *Inference Control in Statistical Databases*. pp 163-171 Springer-Verlag Berlin.

Skinner, C. J. and Elliot, M. J. (2002). 'A measure of disclosure risk for microdata', Journal of the Royal Statistical Society Series B, 64(4) pp 855-867.

Tranmer, M, Fieldhouse E., Elliot, M. J., Dale A., and Brown, M. (2004) ‘Proposals for Small Area Microdata’. To appear in *Journal of the Royal Statistical Society, Series A.*

Papers used in Study

Ballard, R. (1996), "The Pakistanis: stability and introspection," in *Ethnicity in the 1991 Census: The ethnic minority populations of Great Britain*, Vol. 2, C. Peach, ed. London: HMSO, 121-49.

Borooah, V. (1999) "Is there a penalty to being a Catholic in Northern Ireland? An Econometric Analysis of the Relationship between religious belief and occupational success"," *European Journal of Political Economy* 15,2, 163-92

Boyle, P. (1998) "Migration and housing tenure in South East England," *Environment & Planning A* 30, 855-66.

Champion, T. (1996), "Internal migration and ethnicity in Britain," in *Social Geography and Ethnicity in Britain*, Ethnicity in the 1991 Census ed., Vol. 3, P. Ratcliffe, ed. London: HMSO, 135-73.

Dale, A., M. Williams, and B. Dodgeon (1996), *Housing deprivation and social change, A report based on the analysis of individual level census data for 1971, 1981 and 1991 drawn from the Longitudinal Study and the Samples of Anonymised Records*, ONS, LS Series No. 8 ed. London: HMSO.

Drinkwater, S. and O. O'Leary (1997) "Unemployment in Wales: Does language matter?," *Regional Studies* 31, 6, 583-91.

Duncan, S. and R. Edwards (1997) "Lone mothers and paid work: rational economic man or gendered moral rationalities," *Feminist Economics* 3,2, 29-61.

Eade, J., T. Vamplew, and C. Peach (1996), "The Bangladeshis; the encapsulated community," in *The ethnic minority populations of Britain*, Ethnicity in the 1991 Census ed., Vol. 2, C. Peach, ed. London: HMSO, 150-60.

Gardiner, C. and R. Hill (1996) "Analysis of access to cars from the 1991 UK Census Samples of Anonymised Records: a case study of the Elderly Population of Sheffield," *Urban Studies* 33, 69-281.

Gardiner, C. and R. Hill (1997) "Cycling on the Journey to Work: Analysis of Socio-Economic Variables from the 1991 Population Census of Samples of Anonymised Records," *Planning Practice and Research* 12, 251-61.

Gould, M. and K. Jone (2000), "Multilevel modelling of limiting long-term illness using the 1991 Individual SAR for Great Britain," in *Analyzing Census Microdata*, E. Dale, E. Fieldhouse, and C. Holdsworth, eds. London: Edward Arnold, 205-12.

Green, A. (1997), "Patterns of ethnic minority employment in the context of industrial and occupational growth and decline," in *Employment, Education &*

Housing Among Ethnic Minorities in GB, Vol. 4, V. Karn, ed. London: HMSO, 67-90.

Hakim, C. (1994) "A century of change in occupational segregation 1891-1991," *Journal of Historical Sociology* 7,4, 435-54.

King, D. and D. Bolsdon (1998) "Using the SARs to add policy value to household projections," *Environment & Planning A* 30,5, 867-80.

Leventhal, B. (1994), *Case Study Examples to Demonstrate the use of Samples of Anonymised Records in Marketing Analysis, CMU Occasional Paper No. 5.* Manchester: CMU, University of Manchester.

Murphy, M. (1996), "Household and Family Structure among Ethnic Minority Groups in Britain," in *Ethnicity in the 1991 Census. Volume 1 Demographic Characteristics of the Ethnic Minority Populations*, Vol. 1, D. Coleman and J. Salt, eds. London: HMSO, 213-42.

Murphy, M., K. Glaser, and E. Grundy (1997) "Marital status and long-term illness in Great Britain," *Journal of Marriage and the Family* 59,1, 156-64.

Owen, D. (1996) "Black-Other: the melting pot," in *The ethnic minority populations of Great Britain, Ethnicity in the 1991 Census*, Vol. 2, C. Peach, ed. London: HMSO, 66-94.

Owen, D. (1996), "The Other - Asians: the salad bowl," in *The ethnic minority populations of Great Britain, Ethnicity in the 1991 Census*, Vol. 2, C. Peach, ed. London: HMSO, 181-205.

Phillips, D. (1997), "The Housing Position of Ethnic Minority Owners," in *Employment, Education & Housing Among Ethnic Minority Populations of Britain*, V. Karn, ed. London: HMSO.

Rees, P. (1992) "Resources for Research: The 1991 Census of Population," *Environment & Planning A* 24, 1371-80.

Senior, M. L. (1998) "Area variations in self-perceived limiting long-term illness in Britain, 1991: Is the Welsh experience exceptional," *Regional Studies* 32,3, 265-80.

Shouls, S., P. Congdon, and S. Curtis (1996) "Modelling inequality in reported long term illness in the UK: combining individual and area characteristics," *Epidemiology and Community Health* 50,3, 366-76

Simpson, L. (2000), "Small area estimation using census microdata in the UK," in *Analyzing Census Microdata*, A. Dale, E. Fieldhouse, and C. Holdsworth, eds. London: Edward Arnold, 217-21

Tranmer, M. and D. G. Steel (1998) "Using census data to investigate the causes of the ecological fallacy," *Environment & Planning A* 30,5, 817-32.

Williamson, P., M. Birkin, and P. Rees (1998) "The estimation of population microdata by using data from small area statistics and samples of annoymised records," *Environment & Planning A* 30, 5, 785-816.

Appendix 1- Disclosure Control Impact Questionnaire

Data Quality and Disclosure Control Measures Study

Case Study Data Set: 1991 SARs

Study……………………………………………………………………………………………

Please look at the enclosed list of possible changes to the 1991 2% Individual. For each change tick one of the possible responses below.

E. This change **would not affect** the analyses I conducted for this paper.
F. This change **would moderately affect** the analyses that I conducted in this paper.
G. This change **would severely affect** the analyses I conducted in this paper.
H. Other (please indicate the meaning of this in the comments section).

Note: you should consider each change in isolation, imagine when responding that only this particular change is being made. Also, it is important that you only consider the effect on the analyses that you have conducted for the above paper, you not responding regarding your perception as to the general impact of the suggested change.

You may feel that individual changes would have less impact than combinations of changes. We would value your comments on this and any other issues that might affect the meaning of your responses.

Change No.					Change No.				
	A	B	C	D		A	B	C	D
1					16				
2					17				
3					18				
4					19				
5					20				
6					21				
7					22				
8					23				
9					24				
10					25				
11					26				
12					27				
13					28				
14					29				
15									

Comments (Please Continue on separate Sheet if necessary):

Possible Changes to SAR Variables

1) **Age** recoded from single years to Five-year bands.
2) **Age** recoded from single years to Ten-year bands.
3) **Area** removed from the data set but country left in.
4) **Area** and country removed from data set but region left in.
5) **Country of Birth** recoded from 42 categories to 4:
    a. England
    b. Other UK
    c. Europe
    d. Other
6) **Country of Birth** recoded from 42 to two categories:
    a.
    b. UK
    c. NON UK
7) **Ethnicity** recoded from 10 to 4 categories:
    a. White
    b. Black
    c. Asian
    d. Other
8) **Distance of move** recoded to three categories:
    a. 0-9km
    b. 10+ km
    c. Outside GB
9) **Distance to work** recoded as five categories:
    a. Working at home
    b. 0-2km
    c. 3-4km
    d. 5-9km
    e. 10km+
10) **Primary Economic Status** recoded as four categories
    a. Employed
    b. Unemployed/On govt Scheme
    c. Student
    d. Inactive (sick/retired/other)
11) **Secondary Economic Status** Omitted
12) **Family Type** recoded to three categories:
    a. Couple no dependent children
    b. Couple with dependent children
    c. Lone Parent Family
13) **Usual hours of work** recoded to 4 categories:
    a. 0-16
    b. 17-30
    c. 30-40
    d. 41+
14) **Usual hours of work** topcoded at 50
15) **Industry** replaced by single digit standard industrial classification
16) **Marital Status** recoded to three categories
    a. Single
    b. Married
    c. Previously Married
17) **Occupation** replaced by SOCMAJOR

18) **Number of highest qualifications** Omitted

19) **Level of Highest Qualifications** recoded to
    a. First degree or higher
    b. Other 18+ qualification

20) **Subject of Highest Qualification** Omitted

21) **Relationship to household head** recoded to four categories:
    a. Household head
    b. Spouse/Cohabitee
    c. Son/Daughter of Household Head or Spouse/Cohabitee
    d. Other

22) **SEGroup** Omitted

23) **Term-time address** Omitted

24) **Method of Transport to work** recoded to 5 categories
    a. Car
    b. Public Transport
    c. Bike
    d. Foot
    e. Other

25) **Workplace** Omitted

26) **Number of Cars** recoded to three categories
    a. 0
    b. 1
    c. 2+

27) **Household Dwelling Space Type** recoded to five categories
    a. Detached
    b. Semi-detached
    c. Terrace
    d. Flat/Flatlet
    e. Other

28) **Number of Residents per room** recoded to three categories
    f. Up to 0.5
    g. 0.5 - 0.75
    h. Over 0.75

29) **Tenure** recoded to 3 Categories:
    i. Owner-Occupier
    j. Rented Privately
    k. Rented Social

Appendix 2 - Perturbed Files Questionnaire

For each of the four files that you have been sent, having rerun the analyses that you conducted on the 1991 SARs, indicate the level of impact if any that the disclosure control applied to that file has had on results (and the interpretation of the results) by ticking the appropriate response to the question. We would also be grateful for any comments/qualifications you may want to make.

[a] File - indivA
Overall would you say that the results in your paper and the interpretation of the results were: [please tick one]

Unaffected
Moderately Affected
Severely Affected
Other

Comments:

[b] File - indivB
Overall would you say that the results in your paper and the interpretation of the results were: [please tick one]

Unaffected
Moderately Affected
Severely Affected
Other

Comments:

[c] File - indivC
Overall would you say that the results in your paper and the interpretation of the results were: [please tick one]

Unaffected
Moderately Affected
Severely Affected
Other

Comments:

[d] File indivD

Overall would you say that the results in your paper and the interpretation of the results were: [please tick one]

Unaffected
Moderately Affected
Severely Affected
Other

Comments: