WP.13
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

MICRODATA RISK ASSESSMENT IN AN NSI CONTEXT

Supporting Paper

Prepared by Jane Longhurst and Paul Vickers (Office for National Statistics, United Kingdom)

Microdata risk assessment in an NSI context

Jane Longhurst* and Paul Vickers*

*Office for National Statistics, Segensworth Road, Fareham, UK, Jane.Longhurst@ons.gov.uk/ Paul.Vickers@ons.gov.uk

1. Introduction

National Statistics Institutes (NSIs) collect and publish a wide range of economic and social data. Making confidentiality commitments and keeping these commitments is an important factor in maintaining trust between data providers and the NSI. Official statistics are generally released in the form of tables and microdata (individual level records) and the demand from policy makers and researchers for more detailed data and innovative ways to supply the data is continuing to increase. This increased demand increases the potential disclosure risks and this places greater pressure on NSIs to develop sophisticated methods to identify the level of risk posed by any release and to minimise this risk where it is deemed too great.

Traditionally NSIs have adopted microdata risk assessment procedures based on checklist criteria, ad-hoc rules and simple data-based summary measures. More recently there has been a recognised demand for quantitative disclosure risk measures in order to gain more objective criteria for release. This paper will focus on methods that estimate disclosure risk measures for microdata based on the population that make use of statistical models to estimate risk from the sample information in particular a probabilistic disclosure risk approach based on the Poisson Distribution and log-linear models (Elamir and Skinner (2006)). The scope is to investigate methodological aspects and practical issues related to the implementation of the method in particular for anonymised social survey microdata released under license.

In order to provide context for this work Section 2 gives an overview of the microdata release process at the Office for National Statistics (ONS) in the UK. Sections 3 and 4 describe in detail the research that has been carried out to investigate the feasibility of using the method based on the Poisson Distribution and log-linear models and to develop it further in an NSI context. The conclusion is drawn that the method could be used to assess the risk of a microdata file at the individual level but that further research is required to assess the risk at the household level.

2. Background to microdata release at the ONS

ONS has a long history of releasing microdata from its surveys. The release of all microdata from the ONS is approved by the Microdata Release Panel (MRP). Approval of release will depend on a number of factors described by Jackson and Longhurst (2005) and summarised in the following sections.

2.1 Legal and Policy Issues

The sixth United Nations Fundamental Principle of Official Statistics states:
Individual data collected by statistical agencies for statistical compilation, whether or not they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes.

Access for research purposes is an exception to this default position. As such it needs to be a properly planned, managed, authorised and publicly acceptable activity. For research access to be lawful requires all these to be addressed, because compliance with the law is not just a matter of following legislation, but requires compliance with the laws for public administrative and generic information and common law at both the national and European Union (EU) level.

2.2 Risk Assessment

It is important that any disclosure risk assessment is carried within the legal and policy frameworks. For microdata, disclosure risk occurs when there is a possibility that an individual can be re-identified using information contained in the file, and on the basis of that, confidential information is obtained. Microdata is released only after taking out directly identifying variables, such as names, addresses, and identity numbers. However, other variables in the microdata can be used as indirect identifying variables such as gender, age, occupation, place of residence, country of birth, family structure. This combination of indirectly identifying variables is defined as a key and provides the basis for identification of a respondent and hence the disclosure risk.

In terms of disclosure risk assessment an intruder is someone who deliberately or inadvertently determines confidential information about a respondent from a dataset or attempts to do so. To assess the disclosure risk, one firsts need to make realistic assumptions about what an intruder might know about respondents and what information will be available to them to match against the microdata and potentially make an identification and disclosure. These assumptions are known as disclosure risk scenarios. ONS considers various disclosure risk scenarios in carrying out its disclosure risk assessment. The scenarios cover topics such as political attacks, private and public database cross match, journalists, local search and inquisitive neighbours, Elliot and Dale (1998).

These disclosure risk scenarios can be used to define the key variables within a microdata set. ONS has developed a checklist that can be used to focus on these variables. Responses to the questions on the checklist from data suppliers allow a disclosure risk assessment to be made. This assessment is generally made using subjective judgements and precedents. Sections 3 and 4 describe work to provide a quantitative risk assessment to support these subjective judgements.

2.3 Risk Management

The outcome of the risk assessment determines whether further measures need to be carried out or put in place to allow the data to be released. The MRP recognises all microdata releases are not risk free and aims to release microdata in a way that minimises this risk. It uses a combination of disclosure control methods (mostly recoding) and licence agreements or safe settings to control how researchers and policy makers use the data and present the results of any analysis.

3. Quantitative risk assessment

3.1 Introduction

The focus of this paper is the use of quantitative disclosure risk measures and how they can be used for risk assessment at the ONS for social survey microdata released under licence. Social surveys are typically samples of households where the

characteristics of the population are not fully known. When the population is unknown there is a need to rely on models or heuristics in order to quantify the disclosure risk.

Previous work (Shlomo and Barton (2006)) has been undertaken to compare the performance of three different methods; the Special Uniques Detection Algorithm (SUDA) (Elliot et al (2005)), the probabilistic disclosure risk method based on the Poisson Distribution and log-linear models (Elamir and Skinner (2006)) and the method based on the Negative Binomial Distribution (Polettini and Seri (2003)) which is embedded in the Computational Aspects of Statistical Confidentiality (CASC) project software Mu-Argus (CASC (2004)). An evaluation of these three methods has been carried out for a range of different sample sizes but limited key sizes, mostly 6 variables. For the examples considered the results show that the Poisson model with log-linear modelling performs the best but is more complex than the other methods and requires more computing time and intervention in a model search algorithm.

The scope of this paper is to investigate the practical issues related to the implementation of the probabilistic disclosure risk method based on the Poisson Distribution and log-linear models. In particular addressing the performance and feasibility of this method for larger key sizes. The method can be used to estimate risk for individual records and globally for a whole file. The focus here is on file level measures of risk that can be used within a microdata release procedure. Consideration will also be given to the feasibility of estimating disclosure risk for hierarchical microdata sets, e.g. individuals within households.

To introduce the basic measure of identification risk, suppose a key has *K* cells and each cell $k = 1,\ldots K$ is the cross product of the categories of the identifying variables. Let F_k and f_k be the population and sample counts in cell *k* respectively. The aim of quantitative disclosure risk assessment methods for microdata is to estimate an individual per-record disclosure risk measure that is formulated as $\frac{1}{F_k}$, that is the probability that a record in the microdata and a record in the population having the same values of identifying key variables will be correctly matched. Since the uniques in the population, $F_k = 1$, are the dominant factor in the disclosure risk measures, this measure is focused on the case when $f_k = 1$, i.e. for sample unique cells. This leads to the following record-level risk measure: $r_k = E\left[1/F_k \,|\, f_k = 1\right]$

3.2 The Poisson Model and Log-Linear Modelling

As described in Bethlehem et al (1990) consider models where the F_k are realisations of independent Poisson random variables with means λ_k ($k = 1,\ldots\ldots K$), $F_k \sim P(\lambda_k)$. Assume that the sample is drawn by Bernoulli sampling with common inclusion probability π so that $f_k \sim P(\mu_k)$ where $\mu_k = \pi\lambda_k$. The record level measure can be expressed as $r_k = E\left[1/F_k \,|\, f_k = 1\right] = \frac{1}{\lambda_k(1-\pi)}(1 - e^{-\lambda_k(1-\pi)}) = h(\lambda_k)$

This measure depends on unknown λ_k. In order to 'borrow strength' between cells suppose the μ_k are related via the log linear model $\log \mu_k = x_k' \beta$ where x_k is a design vector denoting the main effects and interactions of the model for the key variables. Using standard procedures, such as iterative proportional fitting, this model is fitted to the sample data to obtain the maximum likelihood estimates for the vector β and the fitted values $\hat{\mu}_k = \exp(x_k' \hat{\beta})$ are calculated. The estimate for $\hat{\lambda}_k$ is then substituted in the formula for r_k which can be aggregated across sample uniques to obtain the following file level measure estimate:

$$\hat{\tau} = \sum_{SU} \hat{r}_k = \sum_{SU} \hat{E}\left[1/F_k \,|\, f_k = 1\right] = \sum_k I(f_k = 1) h(\hat{\lambda}_k)$$

the expected number of correct matches for sample uniques, where $SU = \{k : f_k = 1\}$. Such an approach has been described in Skinner and Holmes (1998) and Elamir and Skinner (2006).

A key issue with this method is that inference may be sensitive to the adequacy of the specification of the log linear model. Skinner and Shlomo (2006) develop criteria for assessing whether the vector x_k may be expected to lead to accurate estimated risk measures. Standard approaches such as Pearson or likelihood-ratio tests or Akaike's Information Criterion are discounted since they are not appropriate for the large and sparse tables considered in this application. In this analysis the minimum error test statistic $\frac{\hat{B}}{\sqrt{v}}$ is used, as defined in Skinner and Shlomo (2006). It has an approximate standard normal distribution under the hypothesis that the expected value of $\hat{B}$ is zero. A positive value under 1.96 accepts the fit of the log linear model for obtaining a good disclosure risk measure.

3.3 Method

The data used for this analysis was taken from the 2001 UK Census. Following the method used by Shlomo and Barton (2006) five different samples were drawn from the Census data for England and Wales covering 52 million people within 22 million private households (communal establishments were excluded). The different samples simulate typical sample sizes for different ONS social surveys and the samples were drawn using a similar design as standard social surveys. Clustered samples are not simulated since it is assumed that this aspect of sample design would not affect the risk measurements. Some ONS social surveys sample disproportionately across geographies, this is not simulated here. Table 1 describes the five different samples.

Sample	Sampling Fraction	Number of households in the sample	Number of persons in the sample
A	0.000323	7,000	16,651
B	0.001062	23,000	54,560
C	0.002308	50,000	119,618
D	0.006924	150,000	357,888
E	0.010155	220,000	524,399

Table 1: Samples used in the analysis

The Poisson model with log-linear modelling is used to estimate $\hat{\tau}$ for the different sample sizes and using different key sizes. Since the samples have been drawn from Census data, the true risk measure, τ, can be calculated and used to evaluate the performance of the method.

Following Skinner and Shlomo (2006) a forward search algorithm is used, starting from simpler models and adding interaction terms until the specification is judged to be adequate. As in Shlomo and Barton (2006) the analysis is based on the private database cross match scenario, Elliot and Dale (1998). As a first stage the analysis is restricted to 6 variables with categories typically used in ONS social survey microdata releases: region (11), age (96), sex (2), number of residents (7), marital status (6), number of cars (5). An 8 variable key is also considered, this covers the 6 variable key plus number of earners (5) and number of dependent children (5). The assumption is made that there are no discrepancies in the values of the key variables between the microdata and the intruder's data.

3.4 Results

Table 2 displays the results (final model, estimated and true risk and the minimum error test statistic) for the five samples for the 6 variable key where K = 443,520, replicating the results in Shlomo and Barton (2006). A detailed breakdown of the model search for Sample C is included in the Appendix as an example. For all samples the estimated risk is close to the true risk, in all cases within 10%. The breakdown of the model search shows that as outlined in Skinner and Shlomo (2006) large negative values for the test statistic (overfitting) lead to an underestimate of the true risk and large positive values (underfitting) lead to an overestimate. When the test statistic for a model is small this can lead to either over or under estimation. Each model takes a few minutes to run, some intervention is required through the model search algorithm to select the interactions for inclusion and so models with more interactions take longer to run. As the sample size increases, the global risk estimate increases and the model becomes more complex.

Sample	**Final Model**	**True Risk (τ)**	**Estimated Risk ($\hat{\tau}$)**	**Minimum error test statistic ($\frac{\hat{B}}{\sqrt{v}}$)**
A	All 2-way interactions	83.0	81.4	0.14
B	All 2-way interactions + {age, residents, mstatus}	220.3	204.5	1.13
C	All 2-way interactions + {age, mstatus, cars} +{age, residents, mstatus}	446.6	410.6	1.65
D	All 2-way interactions + {age, mstatus, cars} +{age, residents, mstatus} +{region, age, residents}	1193.8	1182.4	1.74
E	All 2-way interactions + {age, residents, mstatus} +{age, residents, cars} +{age, mstatus, cars} +{region,	1701.8	1569.7	1.03

	age, residents} +{age, sex, mstatus} +{region, residents, mstatus}			

Table 2: Results for 6 variable key

The modelling exercise was repeated for the 8 variable key where K = 11,088,000 but problems were encountered related to computing times. Even running just the all 2-way interaction model takes approximately 40 hours[1]. The different samples take approximately the same time, run time is dependent on key size and number of variables, although as shown in Table 2 sample size tends to affect model complexity which is related to the time taken to complete the model search. It can be calculated that the next stage in the model search, i.e. testing each 3-way interaction term for inclusion in the model would take about 2000hrs or over 33 days! This would not be practical. The next section of the paper describes an approach to improving the time taken to fit the log-linear models.

3.4.1 Partitioning

The approach considered here to reduce the time taken for the model search and model fitting procedures involves partitioning the data into smaller datasets. Models are fitted to each of these subsamples to obtain risk estimates which are then aggregated over all the subsamples to obtain to final risk estimate for the whole file. Since this approach will reduce the size of the sample and key size (if partitioning is based on a key variable) it should result in simpler models which will be easier and faster to fit. Two issues are considered; which variables should be used to partition the sample and how many subsamples should the sample be divided into.

Partitioning by a particular key variable assumes an underlying interaction with that variable in the model in each of the sub-tables. The Cramer's V statistic was used to measure the strength of association between pairs of the key variables in Sample C. The strongest association is between age and marital status (0.434). Sample C was partitioned into different subsamples based on different key variables and the estimated risk measures for the 6-variable key were compared, Table 3.

Partition variable	**Subsamples**	**Average size of partition**	**Key size (K)**	**Estimated Risk ($\hat{\tau}$)**
Age	4 (24 years in each)	29905	110,880	457.1
Region	4 (2 or 3 regions in each)	29905	120,960 or 80,640	503.6
Marital status	4 (1 or 2 marital status categories in each)	29905	147,840 or 73,920	404.3
Age	2 (48 years in each)	59809	221,760	480.2
Sex	2 (male/female)	59809	221,760	501.7

Table 3: Results for partitioning Sample C, 6-variable key, τ = 446.6

The results show that different models are selected for different subsamples, as expected these tend to be simpler models (since the sample size of them is smaller). For some subsamples the true risk is overestimated for others it is under estimated. As

[1] Run times are obviously very dependent on the PC used. All results were obtained on a standard ONS PC; 2.8 GIG processor, 512 MEG memory, Windows XP, SAS Version 8.2.

expected partitioning by age produces the best risk estimate when splitting into 4 or 2 subsamples. The results are robust, the risk estimates for each subsample are good and the overall risk estimates for partitioning by age are more accurate than the estimate with no partitioning (see Table 2).

Now consider the best number of subsamples to implement when partitioning by age. In general as sample sizes get smaller models become simpler. Ideally one would want to find the case when the all 2-way interaction model or the independence model fits. These cases are quicker and easier to run since they require no stepwise procedure and therefore no user intervention in the modelling process. This approach is tested using the 8 variable key across all samples. The all 2-way interaction model is fitted to all subsamples. Different sized subsamples were selected depending on the overall sample size. Table 4 summarises the partitioning results across all samples, displaying the partition that produced the best risk estimate.

Sample	**No. of sub samples**	**Average size of subsample**	**Key size (K)**	**True Risk (τ)**	**Estimated Risk ($\hat{\tau}$)**	**Range for $\frac{\hat{B}_2}{\sqrt{v}}$**	**Time taken**
A	4 age bands	4163	2,772,000	406.0	414.0	[-0.22,17.5]	16 hrs
B	32 age bands	1705	346,500	1284.6	1231.9	[-0.7,7.1]	5 hrs 14 mins
C	32 age bands	3738	346,500	2693.4	2918.9	[-0.9,8.2]	5 hrs
D	32 age bands	11,184	346,500	7703.5	9052.3	[-1.0,22.1]	5 hrs 25 mins
E	32 age bands	16,388	346,500	11111.6	13224.5	[-0.2,20.4]	5 hrs 50 mins

Table 4: Partitioning results, 8-variable key, all 2-way interactions models

The results show that for all samples (other than Sample A) the best results are obtained using 32 partitions. In general as the size of the partition decreases the fit of the all 2-way interaction model improves and hence the accuracy of the risk estimate increases. The results for Sample A and B have shown that if sample sizes become too small (average around 2000 individuals) the all 2-way interaction model starts to overfit the data and leads to an underestimation of the final risk. Overall as the sample sizes increase the error in the risk estimate increases. Ideally one would implement more than 32 partitions for the larger samples (D and E) in order to improve the fit of the all 2-way interaction model and increase the accuracy of the risk estimate. The results suggest that an ideal subsample size for this 8-variable key is 2000-4000 individuals.

4. Measuring risk for hierarchical files

4.1 Introduction

Many social surveys are hierarchical in nature, allowing groups of individuals to be recognised within the file, the most common case being households. If it is possible to

link individuals within the released microdata file then it is important to take into account this dependence when measuring disclosure risk. Within the software Mu-Argus (CENEX (2006)) household risk is defined as the probability that at least one individual in the household is identified and is computed from the individual risk of the household members. An alternative scenario considered here would be that the intruder matches directly at the household level, where a single record represents a household characterised by the identifying variables of all household members. This approach was adopted for the disclosure risk assessment for the Household Sample of Anonymised Records (SARs) (CCSR (2005)) from the 2001 UK Census. Different approaches relate to the assumed availability of hierarchical external databases. Initial observations from the CAPRI (Confidentiality and Privacy Group) Data Monitoring Service (CCSR (2004)) indicate that some hierarchical household information is available in many datasets (particularly when more than one adult lives in the household).

4.2 Method

When measuring risk at the household level the key variables will need to be modified to incorporate information on all the individuals within the household as well as some household level variables, Elliot (2005). Here analysis is restricted to a key with basic household information for all members of the household, e.g. age, sex and preliminary results are outlined for 2 person household only.

Before fitting the model the microdata file needs to be modified. The file is split by household size and one record is created for each household that contains information on all members of the household. The key is then constructed using variables on each member of the household and household level variables. Care needs to be taken in constructing this key in terms of ordering the members within the household. It is possible that two households could be identical, but not recognised as such if they are sorted in different orders.

4.3 Results

Table 5 details the results of the stepwise modelling procedure for all 2 person households in Sample A (of which there are 2414) using a 4-variable key constructed using age and sex of both household members, K = 36,864. The household members are ordered within the key by age and then sex. The minimum error test statistic for the all 2-way interaction model is negative, providing evidence of overfitting and as expected the true risk is underestimated. The forward stepwise procedure is used to add 2-way interaction terms to the independence model. As 2-way terms are added to the model the test statistic decreases (indicating a better fit) but the risk estimate moves further away from the truth. No model can be found with a test statistic that is positive and less than 1.96. A similar pattern of results is observed for sample B. For sample C, D and E the all 2-way interaction model produces the best estimate of risk but this is not reflected in the test statistic. The results show that the modelling procedure is not as effective in this case as it was for modelling at the individual level.

Constructing keys at the household level and in particular the ordering of household members will introduce dependencies into the variables in the key which could affect the validity of fitting a log-linear model to the data. Consider two person households and a simple key of age and sex for both household members where the household

members are ordered by age and then sex. The age of the second household member will always be less than or equal to the age of the first household member.
The format of the household level key introduces structural zeros into the key. In order to reduce these problems the proposal is made to order the key by sex and then age rather than age and then sex. This will not totally overcome the problem interdependencies between the two age variables will still exist but there should be less structural zeros forced into the key by the ordering procedure.

Another factor that may be affecting the results seen here is the number of variables in the key. A 4-variable key has been investigated whereas previous analysis investigated 6 and 8 variable keys. Further analysis should be carried out with larger keys to investigate whether including more variables in the key improves the performance of the models. In particular household type should be considered as a key variable and potentially used as a partitioning variable.

Model	**Estimated Risk ($\hat{\tau}$)**	**Minimum error test statistic ($\frac{\hat{B}}{\sqrt{v}}$)**
Independence	9.39	21.5
All 2-way interaction	1.57	-1.2
Independence + {sex, sex2}	11.75	8.1
Independence + {sex, sex2} + {age, sex2}	11.5	4.6
Independence + {sex, sex2} + {age, sex2} + {age, sex}	12.2	3.4
Independence + {sex, sex2} + {age, sex2} + {age, sex} + {age2, sex2}	1.47	-1.7

Table 5: Results for Sample A, 2 person households, 4-variable key, τ = 4.8

5. Conclusions

There is a strong, widespread and increasing demand for National Statistics Institutes (NSIs) to release microdata files. These data sets are a vital resource for key research and thus it is important to make the microdata as detailed as possible while protecting the confidentiality of the information provided by the respondents. This paper has investigated issues concerned with the practical implementation of a method for quantitatively assessing disclosure risk for microdata based on the Poisson Distribution and log-linear models.

The results have shown that it is feasible to assess the risk of a microdata file at the individual level for a 6 and 8-variable key and that the results are robust. Quantitative file level measures of risk can be used within the microdata release approval process alongside current (more subjective) measures such as the checklist. Splitting the risk assessment by subpopulations reduces computational demands for the 8-variable key. The results show that final risk estimates are more accurate when partitions are determined by a key variable that is most correlated with the other key variables, here age. In general as sample sizes get smaller the best fitting log-linear models become simpler. The recommendation is made that the all 2-way interaction model is taken for each subsample. This will necessarily impact on the quality of the final risk estimate

but will avoid lengthy model search procedures. The results suggest that an ideal subsample size for the 8-variable key is 2000-4000 individuals.

Assessing disclosure risk for larger keys is possible with partitioning but the time taken to carry out the modelling is anticipated to take days rather than hours on a standard ONS PC. Future work should consider the likely availability of external databases with more than 8 or 10 identifying variables in order to gauge whether risk assessments are required for these larger keys. In addition the availability of hierarchical external databases needs to inform the hierarchical disclosure risk scenarios.

The preliminary results outlined here have indicated that as currently implemented the modelling procedure is not as effective for estimating risk at the household level as it is at the individual level. Further analysis is required to overcome the interdependencies in the key variables introduced by the hierarchical structure. Further analysis is also required to investigate larger keys and alternative scenarios.

6. Acknowledgements
A special thank you to Natalie Shlomo and Jeremy Barton for their help in preparing the samples and initial SAS programmes used in the empirical work and to Natalie Shlomo and Chris Skinner for their general support.

References

Bethlehem, J., Keller, W., and Pannekoek, J. (1990) *'Disclosure Control of Micro-data'*. Journal of the American Statistical Association 85 No. 49 (1990) 38-45
CASC website (2004), *Computational Aspects of Statistical Confidentiality*, www.neon.vb.cbs.nl/casc/default.htm
CCSR. (2004) *'ONS Disclosure control report – Special licence Household SAR'*. www.ccsr.ac.uk/sars/guide/2001/disclosure.html
CCSR. (2005) *'A scoping study for the establishment of a data monitoring service'*. www.ccsr.ac.uk/research/datamonitor.htm
CENEX website (2006), *Centre of Excellence for Statistical Disclosure Control*, www.neon.vb.cbs.nl/cenex/
Elliot, M. J., and Dale, A. (1998) *'Disclosure Risk for Microdata'*, Report to the European Union ESP/204 62/DG III.
Elliot, M. J. and Dale, A. (1999) *'Scenarios of Attack: The data intruder's perspective on statistical disclosure risk'*. Invited paper for special edition of Netherlands Official Statistics.
Elamir, E. and Skinner, C. (2006) *'Record-Level Measures of Disclosure Risk for Survey Micro-data'*. Journal of Official Statistics 22 525-539(2006)
Elliot, M. (2005) *'Assessment of Disclosure Risk for Hierarchical Microdata Files'*, ONS Report, Confidentiality and Privacy Group, Cathie March Centre, University of Manchester, 2005.
Elliot, M., Manning, A., Mayes, K., Gurd, J. and Bane, M. (2005) *'SUDA: A Program for Detecting Special Uniques'*. Monographs of Official Statistics, UNECE and Eurostat Work Session on Statistical Data Confidentiality Geneva (November 2005).
Jackson, P. and Longhurst, J. (2005*) 'Providing access to data and making microdata safe, experiences of the ONS'*, Monographs of Official Statistics, UNECE and Eurostat Work Session on Statistical Data Confidentiality Geneva..

Polettini, S and Seri, G. (2003) *'Guidelines for the protection of social micro-data using individual risk methodology - Application within mu-argus version 3.2'*, CASC Project Deliverable No. 1.2-D3.
Shlomo, N. and Barton, J. (2006) *'Comparison of Methods for Estimating Disclosure Risk Measures for Microdata at the UK Office for National Statistics'*, Privacy in Statistical Databases, CENEX-SDC Project International Conference, PSD 2005 proceedings.
Skinner, C. J. and Holmes, D. (1998) *'Estimating the Re-identification Risk Per Record in Micro-data.'* Journal of Official Statistics 14 No. 4, 361-372.
Skinner, C. J. and Shlomo, N. (2006) *'Assessing Identification Risk in Survey Micro-data Using Log-Linear Models'*. www.eprints.soton.ac.uk/41842

Appendix

Model	**Estimated Risk ($\hat{\tau}$)**	**Minimum error test statistic ($\frac{\hat{B}}{\sqrt{v}}$)**
All 2-way interaction	578.4	8.0
All 3-way interaction	271.8	-2.4
All 2-way interaction + {age, mstatus, cars}	516.3	4.1
All 2-way interaction + {age, mstatus, cars} + {age, residents, mstatus}	410.6	1.65

Table A1: Detail of model search for 6-variable key, sample C, τ = 446.6