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

Abstract. This paper provides an overview of how the Office for National Statistics (ONS) in the UK is tackling the problem of balancing the need to provide users with access to microdata and the need to protect confidentiality. A key stage in the microdata release process is risk assessment. A significant amount of research has been undertaken into methods for estimating disclosure risk measures for microdata. This paper will address the issues concerned with the practical implementation of such methods at the ONS focusing on the probabilistic disclosure risk method based on the Poisson Distribution and log-linear models (Elamir and Skinner, 2006).

Keywords. Microdata, disclosure risk assessment, log linear modelling

1. Introduction

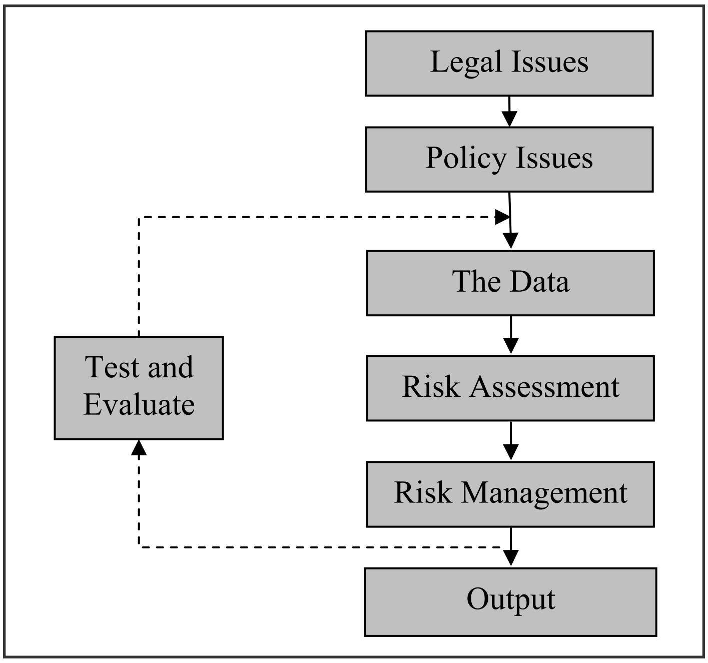
National Statistics Institutes (NSIs) collect and publish a wide range of economic and social data to support policy development and research. This data has been collected from censuses, surveys and administrative data sources. Making confidentiality commitments and keeping these commitments is an important factor in maintaining trust between data providers (businesses and people) and the NSI. It helps preserve the high quality of the information provided.

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. Striking the right balance between removing risk whilst retaining enough detail in the dataset to underpin robust research and policy making is a key challenge facing all Statistical Disclosure Control practitioners.

At the Office for National Statistics (ONS), microdata is only released after its Microdata Release Panel has considered a range of factors. One factor is risk assessment and this paper describes a method being developed by ONS to provide a quantitative risk assessment. Section 2 gives an overview of the MRP process and the factors it considers. Sections 3 and 4 describe in detail the research that has been carried out to develop microdata disclosure risk assessment. 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 and currently has datasets deposited at the UKDA covering around 13 different surveys. The release of all microdata from the ONS is approved by the Microdata Release Panel (MRP). The panel consists of senior officials who have experience releasing microdata across the full range of National Statistics outputs. Approval of release will depend on a number of factors. These are covered by the following framework which is described in detail by Jackson and Longhurst (2005) and is based on the Risk - Utility decision problem approach developed by Duncan, et al. (2001).



The first two stages determine the need for any confidentiality protection whilst stages three to five relate to the risk utility approach required.

2.1 Legal and Policy Issues

Data obtained for statistics is confidential. This is the common feature of all good statistical law, and also of common law for public administration. It is a UN Fundamental Principle for Official Statistics. 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.

- Planning.
  Research data access is something that ideally ought to be enabled through either consent or sufficient advance information to be able to assume consent through participation. Respondents ought to be aware of this valuable use of their data, and be supportive of it. Information needs to be provided, and advance leaflets and confidentiality pledges ought to refer to research access to data. It requires some forethought, so that the benefits of such access can be properly presented to respondents. Failure to plan properly could either result in research access being a breach of data protection law or common law, or an inability to support research.
- Management.
  Research access, being an exception to a zero risk activity (non-disclosure), is a priori a risk activity. Risk must be managed. Where a public authority makes a decision without properly managing risk it is at risk of Judicial Review or some other sanction.
- Authority.
  Legislation may provide the lawful authority to provide research access to data. This authority may be specific, in the sectorial legislation for a particular collection or it may be general. In the UK there are several pieces of legislation that provide powers to release and share data. These include the Census Act (1920), the Statistics of Trade Act (1947), the Population and Statistics Act (1960). Where they are specified, disclosures other than as specified may give rise to a criminal offence.
- Publicly acceptable.
  It is a general obligation on public authorities to carry out their functions in a manner that is compatible with the reasonable expectations of the public they serve. It is appropriate for

statistics offices to give visibility to access to data to ensure the expectation of the public is that such research access will take place. As this public expectation becomes embedded, so more respondents will participate when given information about appropriate research data access. It is an iterative process, and as such it is essential that the NSI does not step ahead of what its public considers acceptable. Nor, of course, should it lag behind and hinder the social, environmental and economic research that the public can legitimately expect to be conducted to better inform public policy and political debate. A NSI could be found to be failing under its administrative law if it is found under scrutiny (whether parliamentary, judicial, government, or peer review under the EU Code) to be out of step with public expectation.

2.2 Risk Assessment

It is important that any disclosure risk assessment is carried out in line with the legal and policy frameworks outlined above. 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. A combination of identifying variables is defined as a key and provides the basis for identification of a respondent and hence the disclosure risk. The disclosure risk is a function of such identifying variables/keys both in the sample and the population.

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. Keeping aware of the types of database available for matching against is an important factor in developing and maintaining disclosure risk scenarios (see Elliot and Dale (1998), Elliot and Purdam (2002), Purdam, Mackey and Elliot (2004)).

These disclosure risk scenarios can be used to define the identifying variables within a microdata set. ONS has developed a checklist that can be used to focus on these identifying variables when considering the risk associated with any particular request for microdata. This is summarised below. For each item on the list the data supplier must provide the necessary information, justify why the level of detail is needed and any analysis of associated disclosure risk that has been carried out:

- level of geography
- details of ethnicity coding
- details of occupational coding
- information on any other visible or traceable variables released in the dataset
- sampling fraction of the survey (and the data released)
- sampling design of the survey (eg details of cluster sampling etc)
- inclusion of hierarchical (eg individual/household) or longitudinal variables
- any measures used to assess and treat outliers in the data
- any assurances given to respondents before their consent was obtained
- information on previous releases of the same or similar data

Responses to these questions 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.

It aims to release microdata in a way that minimises this risk. It uses a combination of disclosure control methods and licence agreements to control how researchers and policy makers use the data and present the results of any analysis.

Disclosure control methods considered for microdata include adding random noise to continuous variables, changing values of categorical variables, recoding or suppression of the data. The method used will depend on the type of data being released. In ONS the majority of microdata files released are protected using recoding.

The second vehicle used to manage the risk of microdata is to release the data under licence arrangements or in a safe setting. These are described briefly below:

- End-User Licence. This is the most commonly used licence and is applied to general microdata products that are generated by all ONS social surveys. It has only limited control over the use and the users and the dataset itself is not seen as having a high confidentiality risk.

- Special Licence. Some users require more detail than can be provided using an end-user licence. The special licence places more controls on the user and how the data can be held and are applied to more detailed microdata products.

Both the End User Licence and the special licence require published outputs to meet the Confidentiality Guarantee of the National Statistics Code of Practice.

- ONS also provides access to identified or identifiable data in its microdata laboratory. Researchers using the laboratory are supervised and all output is checked for disclosure before it is removed from the laboratory.

3. Quantitative risk assessment

3.1 Introduction

Section 2 has provided an overview of a framework for protecting and providing access to microdata. A key stage in this framework is assessing disclosure risk. As described the current risk assessment procedure for microdata files being released by the ONS is based on a checklist criteria (informed by disclosure risk scenarios), subjective judgement and past experience. More recently there has been a recognised demand for quantitative disclosure risk measures for microdata in order to gain more objective criteria for release. The focus of this paper is how this can be achieved for ONS social survey microdata released under an End-User licence. Social surveys are typically samples of households and therefore the characteristics of the population are not fully known. The disclosure risk is a function of both the population and the sample, and in particular depends on records that are unique on a set of identifying key variables. When the population is unknown there is a need to rely on models or heuristics in order to quantify the disclosure risk. A significant amount of research has been undertaken into methods for estimating quantitative disclosure risk measures for microdata.

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 at the ONS for social survey microdata files. In particular addressing the performance and feasibility of this method for larger key sizes, specifically investigating the practicality of splitting the risk assessment by subpopulations in order to reduce computational demands. An objective will be to balance the quality of the risk estimates against the simplicity of the method and ease of implementation. 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. The utility of individual level measures are recognised particularly for targeting disclosure control methods. At present research has been restricted to estimating risk for individual respondents. 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,….*K* is the cross product of the categories of the identifying variables. Let F_k be the population count in cell *k* of the multi-way contingency table formed by cross-classifying the key variables. Let the corresponding sample count be f_k. The modelling approach investigated here implies that the population and sample size (denoted by *N* and *n,* respectively) are random and that $\sum_{k=1}^{K} F_k = N$ and $\sum_{k=1}^{K} f_k = n$. Throughout this report, *n* or *N* refers to their expectations which are estimated by their natural estimators: the actual sample size and population size.

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:

$$r_k = E\left[1/F_k \middle| f_k = 1\right]$$

This individual risk measure can be aggregated across sample uniques to obtain a global measure for the entire file:

$$\tau = \sum_{SU} r_k = \sum_{SU} E\left[1/F_k \middle| f_k = 1\right]$$

the expected number of correct matches for sample uniques, where $SU = \{k : f_k = 1\}$.

For this measure the problem of risk assessment becomes one of statistical inference since the f_k are observed but the F_k are not.

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,……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 $\mu_k = \pi\lambda_k$ and $f_k | F_k \sim Bin(F_k, \pi)$. Therefore the record level measure can be expressed as

$$r_k = E\left[1/F_k \middle| f_k = 1\right] = \frac{1}{\lambda_k(1-\pi)}(1 - e^{-\lambda_k(1-\pi)}) = h(\lambda_k)$$

This measure depends on unknown λ_k for cells where the observed counts f_k are just one. The assumption was made that the sample frequencies, f_k are independently Poisson distributed with mean μ_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 $\hat{\lambda}_k$ equal to $\frac{\hat{\mu}_k}{\pi}$ is then substituted in the formula for r_k which can be aggregated across sample uniques to obtain the following file level measure estimate:

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

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, where

$$\hat{B} = \sum_k \hat{a}(f_k - \hat{\mu}_k) + \hat{b}\left[(f_k - \hat{\mu}_k)^2 - f_k\right], \text{ and}$$

$$v = \sum_k \hat{a}^2 \hat{\mu}_k + 2\hat{b}^2 \hat{\mu}_k^2 \text{, where}$$

$$\hat{a} = -\hat{\lambda}_k \exp(-\hat{\mu}_k) h'(\hat{\lambda}_k)$$

and

$$\hat{b} = \frac{\hat{\lambda}_k}{2\pi} \exp(-\hat{\mu}_k) h''(\hat{\lambda}_k).$$

$\hat{B} / \sqrt{v}$ 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

Every 10 years since 1801, the UK has set aside one day for the census whereby information is obtained on every member of the population. It is the only survey which provides a detailed picture of the entire population, it covers everyone at the same point in time and asks the same core questions everywhere. 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. In the first stage households were sorted according to geography and systematically sampled with equal probability. In the second stage all individuals in the sampled household were selected. Clustered samples, i.e. by sampling postcode sectors and then households within these sectors, 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**	**No. households in the sample**	**No. 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 $\hat{r}_k$ for each unique individual in each of the samples for a particular key. Summing these record-level measures across the sample uniques results in the file level measure $\hat{\tau}$, an estimate of the expected number of correct matches for sample uniques. This approach is repeated 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.

The log-linear models are fitted using Iterative Proportional Fitting (IPF) programmed in SAS since log-linear model fitting procedures in standard statistical software will often not cope with the large numbers of variables and cells that are used here. Initially five iterations are used for the IPF procedure. The estimation method deals automatically with zero marginal counts corresponding to a given model, for example because of impossible combinations of key variables values (structural zeros), by setting the fitted values for cells falling in these margins to zero.

Since $\hat{B}/\sqrt{v}$ was primarily derived as a means to detect underfitting (and empirical results suggest that it is more effective for this purpose than for detecting overfitting) a forward search algorithm is used, starting from simpler models and adding interaction terms until the specification is judged to be adequate. Empirical experiments (Skinner and Shlomo (2006)) have shown that the independence log-linear model tends to underfit and at the other extreme the all 3-way interaction model tends to overfit. As a practical approach the suggestion is made to first compute the minimum error test for the independence model and the all 2-way interactions model. If the latter shows no sign of underfitting (negative test statistic) then start with the independence model and use forward stepwise procedure to add in 2-way interaction terms to the independence model. If all 2-way interaction model does show signs of underfitting (large positive test statistic) then use a forward stepwise procedure adding 3-way interaction terms to the all 2-way interaction model. Following Shlomo and Barton (2006) a model is accepted when the value of the test statistic, $\hat{B}/\sqrt{v}$, is less than 1.96.

The sample design deviates from the underlying assumption of a simple random sample, it is possible that a complex sampling scheme could invalidate the assumption that the f_k are Poisson distributed. Under a clustered sample of households, dependencies will be introduced into the data in particular for cells defined by household level key variables, e.g. region, number of residents, number of cars, however other individual level key variables will cut across clusters thus reducing these dependencies. Skinner and Shlomo (2006) suspect that sampling error effects are less important than the impact of model choice and Shlomo and Barton (2006) report that the deviation from a simple random sample is less noticeable when aggregating the individual risk estimates to obtain the file level measures.

As in Shlomo and Barton (2006) the analysis is based on the private database cross match scenario. This has been identified by the ONS as the most important scenario, where an intruder aims to match a released microdata file with an external database. As defined in Elliot and Dale (1998) this scenario has a 12 variable key, however previous analysis (Shlomo and Barton (2006)) has mostly focused on a 6 variable key. 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. In all cases here the true risk is underestimated for the final model. 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, age, residents} + {age, sex, mstatus} + {region, residents, mstatus}	1701.8	1569.7	1.03

Table 2: Results for 6 variable key

The interpretation of the final risk measure is the expected number of correct matches for sample uniques. If an intruder matched sample A to a population database (assuming there are no discrepancies in the values of the key variables) the estimated expected number of correct matches (for sample uniques) would be 81.4. One could present this figure relative to sample size or the number of sample uniques. The size of Sample A is 16651 of which 9705 are sample uniques. The final reported risk measures could therefore be:

- 81.4 correct matches (based on sample uniques)
- 0.5% of sample correctly matched (based on sample uniques)
- 0.8% of sample uniques correctly matched

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. The run time also depends on the number of times the SAS programme loops through the raking procedure during the IPF, this depends on the number of variables and the number of interations. 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 and hence two approaches that could speed up the modelling process are considered. Firstly whether it is possible to reduce the number of times the programme iterates through the IPF. Secondly consideration is given to partitioning the data (reducing sample and key size) and estimating the risk for subpopulations. This approach should result in simpler models which will be easier and faster to fit. Necessarily these approaches will involve a trade off between practicality and the accuracy of the risk estimates, i.e. is it possible to fit simpler models that do not necessarily fit the data as well but still produce reasonable estimates of risk.

3.4.1 IPF Iterations

In order to investigate the number of iterations used in the IPF procedure consider the 6-variable key for samples A, B and C and models selected using 5 iterations as described in Table 2. The model fitting procedure is repeated for 1 to 7 iterations. The results for Sample B are displayed graphically in Figure 1 and in a table in the Appendix.

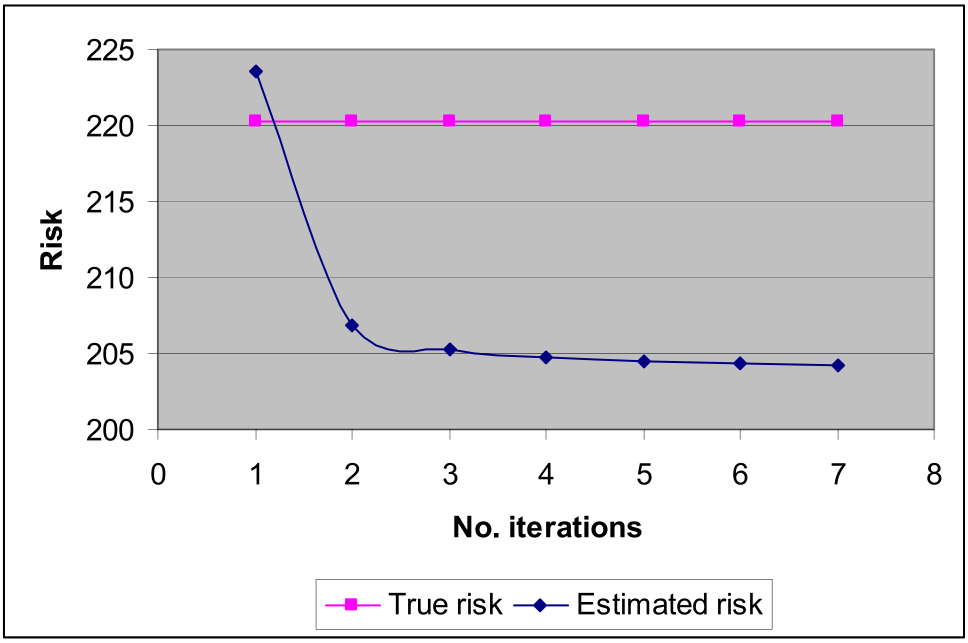


Figure 1: Results for Sample B, 6-variable key, all 2-way interaction model + {age, residents, mstatus}

The results for all 3 samples show that after 2 iterations the total number of iterations used in the IPF procedure does not significantly affect the error in the risk estimate. The conclusion is drawn that it is acceptable to only use 2 iterations for the IPF procedure. This will significantly reduce the time needed to fit the models.

The results described above have assumed a fixed model. However, in practice using a different number of iterations in the IPF may lead to a slightly different model selection and hence different risk estimates. For example for Sample B, after 3 iterations the best fitting model is actually all 2-way

[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.

interaction + {age, residents, cars} which has a slightly better test statistic than the model considered above and leads to a risk estimate of 223.9 rather than 205.2.

3.4.2 Partitioning

The second 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 partitions to obtain risk estimates which are then aggregated over all the partitions 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 partitions 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. If a survey uses a stratified sample then the stratification variables will have to be included in the key and should be considered as partitioning variables. Here the samples were not stratified and so it seems sensible to choose the variable which is the most correlated with other variables in the key. 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). There is also some association between age and number of residents (0.276) and age and number of cars (0.183). Similar levels of association are observed between number of residents and marital status (0.285) and number of residents and number of cars (0.207). It seems that age would be a good candidate for a partitioning variable.

In order to test this hypothesis consider partitioning Sample C into 4 subsets based on age, region, and marital status and compare the estimated risk measures for the 6-variable key. Also consider partitioning Sample C into 2 subsets based on age and gender. Table 3 displays the results.

Partition variable	**Partitions**	**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

A detailed breakdown of results by partition for 4 age partitions is included in the Appendix. The results show that different models are selected for different partitions, as expected these tend to be simpler models (since the sample size of them is smaller). For some paritions the true risk is overestimated for others it is under estimated. As expected partitioning by age produces the best risk estimate when splitting into 4 or 2 partitions. The results are robust, the risk estimates for each partition are good and the overall risk estimates for partitioning by age are more accurate than the estimate with no partitioning (see Table 2). The conclusion is drawn that age is the best variable to partition by.

At present partitions were formed by grouping an equal number of categories of the key variable by order, e.g. ages 1 to 24 then 25 to 48 etc. Simpler models (which may be quicker to fit) may be obtained by forming partitions based on more natural groupings of the categories, e.g. 1 to 15, 16 to 19, 20 to 30 etc.

Now consider the best number of partitions to implement when partitioning by age. Table 4 compares the results for 2, 4 8 and 16 age bands. Detailed breakdowns by partition are provided in the appendix for 8 age bands.

No. of partitions	**Average**	**Key size (K)**	**Estimated**

	size of partition		Risk ($\hat{\tau}$)
2 age bands	59809	221,760	480.2
4 age bands	29905	110,880	457.1
8 age bands	14952	55,440	434.9
16 age bands	7476	27,720	457.3

Table 4: Results for Sample C, 6-variable key, partitioned into 2, 4, 8 and 16 age bands, τ = 446.6

The results in Table 4 are robust, even for very small samples the risk estimates are reasonable. The results show that the accuracy of the total risk estimate is good (within 3%) whether partitioning is carried out for 4, 8 or 16 age bands. The choice for the best number of partitions would therefore depend on time taken to carry out the modelling exercise. In general as sample sizes get smaller models become simpler, e.g. independence model or independence model plus some 2-way interactions has a better fit than the 2-way interaction model or the 2-way interaction model plus some 3-way interactions. 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. Consider the case when the all 2-way interaction model is taken for each partition regardless of the minimum error test statistic. For 16 age bands this would result in an estimate of 443.4, some over and some under estimation cancels out across the partitions. For 8 age bands this would result in an estimate of 467.5 (still within 5% of the true risk).

This approach is tested further using the 8 variable key. Following the results from Section 3.4.1 the all 2-way interaction model is fitted to partitions of Sample C based on 8 and 16 bands using 2 iterations in the IPF procedure. Table 5 displays the overall results including the time taken to carry out all of the modelling (aggregated over the partitions). The results by partition are shown for 8 age bands in the Appendix.

No. of partitions	Average size of partition	Key size (K)	Estimated Risk ($\hat{\tau}$)	Time taken
8 age bands	14952	1,386,000	3278.7	8 hrs
16 age bands	7476	693,000	3134.8	2 hrs 40mins
32 age bands	3738	346,500	2918.9	5 hrs

Table 5: Results for Sample C, 8-variable key, partitioned into 8/16 age bands, all 2-way interactions models, τ = 2693.4

The results are not as accurate as for the 6-variable key. The risk is overestimated in all cases. Partitioning into 8 age bands results in an estimate which is within 22% of the truth. Using 16 age-bands is quicker and more accurate (within 17%). 32 age bands takes longer to model but results in the most accurate estimate of the true risk, within 9%. For the 8 band partitioning the minimum test statistics range from 2.5 to 22.5 indicating that the all 2-way interactions model is underfitting the data for each partition and hence the overall risk is overestimated. For 16 bands the range is -0.7 to 13.9, again this leads to overestimation. For 32 bands the range is -0.9 to 8.2. Both results are quicker and more accurate than running a 2-way interaction model on the full dataset without any partitions.

Similar modelling exercises were carried out for the other samples where different sized partitions were selected depending on the overall sample size. Table 6 summarises the partitioning results across all samples, displaying the partition that produced the best risk estimate.

Sample	No. of partitions	Average size of partition	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 6: 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. For Sample A the best result is for 4 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. Alternatively one could form equal sized partitions by implementing different groupings of age. The results suggest that an ideal partition size for this 8-variable key is 2000-4000 individuals.

The practicalities of running a 10 variable key defined as the 8 variable key plus tenure (6), primary economic status (14) are also considered. This would mean that K = 931,392,000. Even with 96 age partitions this would mean K = 9,702,000 for each partition. Based on previous results the whole model would take 32 days to run (just using 2-way interaction and 2 iterations within the IPF). It may be possible to reduce running times by using a more powerful PC. This modelling exercise would only need to be carried out once for each social survey and only repeated if the design of the microdata file altered significantly. In addition running times could be shortened by reducing the number of categories for each variable, however this will determined by the design of the microdata file. It does not seem practical to run a 10-variable key.

The ONS will need to consider whether it is reasonable to measure risk based on an 8 variable key. Is it likely that an intruder would have access to more than 8 identifying variables in an external database? Is it possible to set benchmarks against agreed risk thresholds using 8 variables? Is it possible to combine risk estimates from different 8-variable keys?

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. Limited research has been carried out on this problem. A handbook produced as part of the Eurostat funded Centre of Excellence (CENEX) in Statistical Disclosure Control (CENEX (2006)) includes a discussion of household or hierarchical disclosure risk scenarios and how the Mu-Argus risk methodology estimates household risk. Rather than estimating household risk directly it 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. This approach does not consider the increased risk from combining the characteristics of the individuals in the household. The assumed scenario is that the intruder matches the microdata file with the external database at the individual level (which includes household characteristics). An alternative scenario 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 alternative approach was adopted for the disclosure risk assessment for the Household Sample of Anonymised Records (SARs) (CCSR (2005)) from the 2001 UK Census and in the assessment of the UK 1991 Census Household SAR using the SUDA method, (Elliot (2005)). The reason for these two different approaches is related to the assumed availability of hierarchical external databases. Initial observations from the CAPRI (Confidentiality and Privacy Group) Data Monitoring Service (CCSR (2004)) indicate that detailed information on children is rarely collected outside of the benefits system so that some information needed by an intruder for a hierarchical matching attack on a household file is not available. However, it is noted that some hierarchical household information is available in many datasets (particularly when more than one adult lives in the household) and the data environment is constantly changing. Given increasing trends in data collection and availability it is not possible to state with certainty that full hierarchical datasets would not become available in the future. The assumption made here is that such an external database is available.

The disclosure risk scenario will involve an intruder identifying a household based on the individual and household information in the microdata file. Studies have shown that this significantly affects the risk, where based only on the age and sex of the individuals any household with more than 6 people has a very high probability of being unique in the population (CCSR (2005)).

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. Following Elliot (2005) three different types of keys can be considered using variables from the private database scenario:

1. Basic household information for all members of the household, e.g. age, sex.
2. As above but includes some household structure information, e.g. geography, number of cars.
3. Detailed information about one member of the household, e.g. age, sex, marital status, primary economic status, plus some basic information about the other household members, e.g. broad age bands, sex.

Here analysis is restricted to the first scenario and preliminary results are outlined.

Before fitting the model the microdata file needs to be modified. The file is split by household size and separate models are fitted to each different sub-file (almost like a partition). For each household one record is created 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. 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. Certain age combinations will be more likely, i.e. both members having similar age (partners) or one member being older than the other (parent and child), whereas other combinations will be rare or impossible, e.g. two under 16 year olds. The construction of the household key introduces many structural zeros which may or may not be modelled effectively.

The modelling approach is the same as was used at the individual level. Note, at the individual level the sample design deviated from the assumption of a simple random sampling since all individuals were sampled within each household. However, when considering modelling at the household level the simple random assumption holds.

4.3 Results

The results here are restricted to 2 person households. Table 7 describes the files for the different sample sizes.

Sample	**No. 2 person households in the sample**
A	2414
B	7651
C	16715
D	50365
E	73888

Table 7: Samples used for 2 person household analysis

Table 8 details the results of the stepwise modelling procedure for Sample A using scenario 1, 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. Two possible reasons are proposed.

The format of the household level key introduces dependencies between the key variables and many structural zeros, particularly for the ages of the household members. The inclusion of the 2-way interaction term between the two age variables will ensure that for any combination of these two variables with a zero marginal count in the sample, all fitted cells will be zero. Hence, structural zeros will be accounted for in the model. However, other random zeros will also be fitted and the model could overfit the data. This is observed for Sample A and B where the all 2-way interaction model overfits the data. However, this is not observed for the other samples. In order to reduce the problems associated with the interdependencies between the two age variables 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. One combination of sex for the two household members will become a structural zero determined by the ordering, however the 2-way interaction between the two sex variables could be forced into the model to take account of this. Including this 2-way interaction term should not cause overfitting or impact on the degrees of freedom in the model as much as the 2-way interaction term between the age variables would since there are only 2 categories of sex. Interdependencies between the two age variables will still exist but there should be less structural zeros forced into the key by the ordering procedure.

The modelling exercise described above should be repeated with the same samples and same key variables but with the household members ordered within the key by sex and then age.

Another factor that may be affecting the results seen here is the number of variables in the key. Here 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. Creating partitions for particular households, e.g. 2 adults or 1 adult and 1 child, will introduce restrictions to the age categories, simplifying the keys and hence reducing the number of structural zeros.

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	-1.16
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	-1.69

Table 8: 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. However, this objective conflicts with the obligation that NSIs have to protect the confidentiality of the information provided by the respondents. A framework for protecting and providing access to microdata has been proposed. A key stage in this framework is assessing disclosure risk. 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.

The time taken to estimate disclosure risk for the 8-variable key can be shortened by reducing the number of iterations employed when fitting the log-linear models using IPF. It is acceptable to only use 2 iterations and this does not significantly affect the error in the final risk estimate.

Splitting the risk assessment by subpopulations or partitions also 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. Further work could investigate how the modelling procedure could be simplified by forming partitions by more natural groupings of age categories, e.g. 1 to 15, 16 to 19, etc.

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 partition. This will necessarily impact on the quality of the final risk estimate but will avoid lengthy model search procedures. For small microdata samples (17,000 individuals) the recommendation is made to split the file into 4 partitions (average 4000 individuals). For larger samples (50,000-500,000 individuals) more partitions are recommended aiming again for an average of around 2000-4000 individuals in each partition.

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. Other methods, in particular SUDA, can handle larger keys but this is a heuristic method rather than a method which is based on well-defined statistical theory. 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

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.*

Elliot, M. J. and Purdam, K. (2002) 'An evaluation of the availability of public data sources which could be used for identification purposes – A Europe wide perpsective'. *CASC report,* www.neon.vb.cbs.nl/casc/

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).*

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.*

Purdam, K., Mackey, E. and Elliot, M. J. (2004) 'The Regulation of the Personal: Individual Data Use and Identity in the UK', *Policy Studies, Oxfordshire.*

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 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

No. iterations	**Estimated Risk ($\hat{\tau}$)**	**Minimum error test statistic ($\frac{\hat{B}}{\sqrt{v}}$)**	**Min correction factor**	**Max correction factor**
1	223.6	0.79	0	4.7
2	206.8	0.77	0.4	2
3	205.2	0.97	0.6	1.5
4	204.8	1.06	0.9	1.3
5	204.5	1.13	0.9	1.2
6	204.3	1.19	0.9	1.1
7	204.2	1.24	0.9	1.1

Table A2: Results for Sample B, 6-variable key, all 2-way interaction model + {age, residents, mstatus}, by number of iterations, $\tau = 220.3$

Partition	**No. persons in the partition**	**Final Model**	**True Risk (τ)**	**Estimated Risk ($\hat{\tau}$)**	**Minimum error test statistic** ($\frac{\hat{B}}{\sqrt{v}}$)
1 to 24 years	36711	All 2-way interactions	63.5	68.4	0.04
25 to 48 years	40539	All 2-way interactions	145.8	152.8	0.59
49 to 72 years	29948	All 2-way interactions	143.1	167.8	0.38
Over 72 years old	12420	All 2-way interactions + {age, residents, cars} + {region, age, mstatus}	94.3	68.1	0.74
		Total	446.6	457.1	

Table A3: Results for Sample C, 6-variable key, partitioned into 4 age bands, K = 110,880

Partition	**No. persons in the partition**	**Final Model**	**True Risk (τ)**	**Estimated Risk ($\hat{\tau}$)**	**Minimum error test statistic** ($\frac{\hat{B}}{\sqrt{v}}$)
1 to 12 years	18216	Independent + {age, mstatus} + {age, residents}	14.3	16.5	0.6
13 to 24 years	18495	Independent + {residents, cars} + {age, mstatus}	49.2	48.8	1.8
25 to 36 years	21781	Independent + {mstatus, cars} + {region, mstatus} + {age, mstatus}	69.5	66.3	1.06
37 to 48 years	18758	All 2-way interactions	76.3	78.1	1.4
49 to 60 years	16939	All 2-way interactions	74.1	79.7	0.08
61 to 72 years	13009	Independent + {residents, cars} + {age, mstatus}	69.0	56.2	1.8
73 to 84 years	8941	All 2-way interactions + {region, mstatus, cars}	60.9	55.9	1.04
Over 85 years old	3479	All 2-way interactions + {age, sex, residents}	33.4	33.4	0.29
		Total	446.6	434.9	

Table A4: Results for Sample C, 6-variable key, partitioned into 8 age bands, K = 55,440

Partition	**No. persons in the partition**	**Final Model**	**True Risk (τ)**	**Estimated Risk ($\hat{\tau}$)**	**Minimum error test statistic** ($\frac{\hat{B}}{\sqrt{v}}$)
1 to 12 years	18216	All 2-way interaction	153.5	181	10.3
13 to 24 years	18495	All 2-way interaction	347.6	376.9	22.5
25 to 36 years	21781	All 2-way interaction	550.0	646.8	7.13
37 to 48 years	18758	All 2-way interaction	602.4	739.1	3.8
49 to 60 years	16939	All 2-way interaction	483.2	623.2	2.5
61 to 72 years	13009	All 2-way interaction	302.8	396.6	4.3
73 to 84 years	8941	All 2-way interaction	172.3	226.5	7.7
Over 85 years old	3479	All 2-way interaction	81.7	88.9	4.6
		Total	2693.4	3278.7	

Table A5: Results for Sample C, 8-variable key, partitioned into 8 age bands, K = 1,386,000