WP. 9
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
(Geneva, Switzerland, 9-11 November 2005)

Topic (ii): Disclosure risk, information loss and usability of data

ASSESSING DISCLOSURE RISK IN MICRODATA USING RECORD-LEVEL MEASURES

Invited Paper

Submitted by the Office for National Statistics and University of Southampton, United Kingdom; Hebrew University, Israel[1]

[1] Prepared by Chris Skinner, Southampton Statistical Sciences Research Institute, University of Southampton and Natalie Shlomo, Southampton Statistical Sciences Research Institute, University of Southampton, Office for National Statistics, Hebrew University.

Assessing disclosure risk in microdata using record-level measures

Chris Skinner[*] and Natalie Shlomo[**]

[*] Southampton Statistical Sciences Research Institute, University of Southampton
[**] Southampton Statistical Sciences Research Institute, University of Southampton, Department of Statistics, Hebrew University, Office for National Statistics

Abstract: We consider the estimation of measures of disclosure risk using Poisson log-linear models. We focus on the question of how to specify the log-linear model. We develop some procedures related to the assessment of over-dispersion. We evaluate these procedures using simulated samples drawn from the 2001 United Kingdom Census and a real dataset being considered for release by the Office for National Statistics. We find that the procedures do indeed help to select models which provide good estimates of disclosure risk measures.

1 Introduction

Assessing disclosure risk for sample-based microdata is a growing challenge for National Statistical Institutes. Most decisions are based on ad-hoc rules, check lists and experience. There is a need to incorporate consistent and high-quality quantitative measures of disclosure risk in order to obtain more objective criteria for releasing microdata to different users. Since data on businesses are rarely released because of the high risk associated with skewed distributions, the focus in this paper is on microdata from social surveys.

Disclosure risk depends on microdata records that are both unique in the sample and in the population on a set of potentially identifying cross-classified key variables (i.e., a key). The key variables are determined according to disclosure risk scenarios. For example, if we are protecting against the risk scenario of matching the microdata to publicly available external files, we would want to choose key variables that are in common between the sources of data. We assume that the key variables are discrete and that there is no measurement error in the way these variables are recorded. Typical key variables include visible and traceable variables: sex, age, ethnicity, religion, place of residence, and occupation. In general, we will be analysing contingency tables spanned by the key variables. These tables contain the sample counts and are typically very large and very sparse.

We consider individual risk measures for each record in the microdata. By targeting only records with high risk, disclosure control techniques can be applied locally and the information loss to the file minimized. One advantage of the probabilistic method for assessing disclosure risk is that individual risk measures can be aggregated to obtain consistent overall global risk measures for the entire file which are useful to Microdata Release Panels in their decision making processes. The assessment and

management of disclosure risk in microdata depends also on the means for disseminating the microdata and the level of protection that is needed. Microdata can be released to on-site data labs, licensed data archives, and public use. Thresholds are set below which the microdata can be released and above which more disclosure control techniques are necessary. The quantitative disclosure risk measures are therefore a necessary tool for ranking files according to their level of disclosure risk.

Let the key define K cells in the contingency table, labeled $k = 1,...,K$. Let the population count in cell k be F_k and the sample count be f_k. We consider the following two global risk measures:

1. Expected number of sample uniques that are population unique, $t_1 = E[\sum_k I(f_k = 1, F_k = 1)]$,
2. Expected number of correct matches for sample uniques to the population, $t_2 = E[\sum_k I(f_k = 1)/F_k]$.

These measures may be expressed as aggregates of record level measures: $t_1 = \sum_{SU} r_{1k}$, $t_2 = \sum_{SU} r_{2k}$, where $r_{1k} = P(F_k = 1 | f_k = 1)$, $r_{2k} = E[1/F_k | f_k = 1]$ with the sums over sample unique cells, denoted SU.

We suppose the f_k are observed but the F_k are unobserved. The measures are estimated using a Poisson model for the f_k, as developed by Bethlehem et al. (1990) and subsequently also used for disclosure risk assessment in the μ-Argus software (Hundepool, 2003; Benedetti et al., 1998; Polettini and Seri, 2003; Rinott, 2003). We shall assume a log-linear model for the underlying means of the Poisson distribution, following Skinner and Holmes (1998) and Elamir and Skinner (2005) and build on their approach by developing methods for the selection of the log-linear model and goodness-of- fit criteria.

In Section 2 we set out the model and its implications for disclosure risk assessment. Section 3 discusses possible criteria for choosing a model. Section 4 covers a model selection algorithm that is implemented taking into account the hierarchical structure of the log-linear models. Section 5 presents examples of how the probabilistic modelling can be implemented on both simulated samples drawn from the 2001 UK Census and a real dataset that is being considered for release by the Office for National Statistics (ONS). Finally, Section 6 contains a discussion and future research.

2 The Poisson Model

Following the earlier notation, a key is defined with cells $k = 1,..., K$. Let $N = \sum F_k$ and $n = \sum f_k$ be the population and sample sizes respectively. Based on natural

assumptions for estimating rare populations we assume for each cell k: $F_k \mid g_k \sim Pois(Ng_k)$ for $g_k > 0$. A sample is drawn by Bernoulli sampling without replacement: $f_k \mid F_k \sim Bin(F_k, p_k)$, where the inclusion probability p_k may vary between cells. It follows that $f_k \mid g_k \sim Pois(Np_k g_k)$ since a thin Poisson variable is again Poisson. Based on these assumptions, we obtain: $F_k \mid f_k \sim Poisson(Ng_k(1-p_k)) + f_k$. Denoting $Ng_k = l_k$, the record level measures may be expressed as:

$$r_{1k} = e^{-l_k(1-p_k)},\ r_{2k} = E(\frac{1}{F_k} \mid f_k = 1) = \frac{1}{l_k(1-p_k)}[1 - e^{-l_k(1-p_k)}].$$

Elamir and Skinner (2005) propose using log-linear modelling to estimate the parameters l_k. Assuming a simple random sampling design where $p_k = p = n/N$ for all cells k, the sample frequencies f_k are independent Poisson distributed with means $u_k = p\,l_k$.

In the Poisson regression log-linear modelling framework, observed counts in a contingency table y_k are independent Poisson distributed given a design vector of regressors $\mathbf{x}_k$ denoting the main effects and interactions of the key variables. The means take the form $u_k = \exp(\mathbf{x}_k' b)$, where b is a parameter vector. The maximum likelihood estimator $\hat{b}$ is obtained by solving the score equations $\sum(y_k - u_k)\mathbf{x}_k = 0$ using iterative proportional fitting (IPF) or methods such as Newton-Raphson. Fitted values are obtained as $\hat{u}_k = \exp(\mathbf{x}_k' \hat{b})$. We calculate $\hat{l}_k = \hat{u}_k / p$ for our estimates of the l_k and plug these estimates into the formulae for the individual and global disclosure risk measures.

3 Criteria for Model Choice

We seek criteria for specifying the vector $\mathbf{x}_k$ in the log-linear model which lead to accurate estimated disclosure risk measures and are robust across different settings. One approach would be to use goodness-of-fit criteria such as Pearson or likelihood-ratio tests. The accuracy of the standard asymptotic approximations involved in the use of these procedures depends on the average cell size n/K being large enough, usually $n/K > 5$ although at least $n/K > 1$. Even this constraint does not hold for the large and sparse contingency tables that are typically used for assessing disclosure risk. For example, the quarterly UK Labour Force Survey individual microdata has 127,200 records in 10,540,000 cells defined by six identifying key variables, and the average cell size is 0.012. Some work on sparse tables (Koehler,

1986) suggests that the Pearson test is preferable to the likelihood ratio test in such circumstances. Nevertheless, our empirical work has suggested that neither of these criteria are very successful in predicting whether the disclosure risk measures will be well estimated and we shall not consider them further in this paper.

Instead, we consider an approach which is motivated more directly by our aim to estimate the disclosure risk measures accurately. Specifically, we seek a criterion for choosing a model which minimises the bias of $\hat{t}_1$ as an estimator of t_1. Treating $\hat{b}$ as fixed, the bias may be expressed as:

$$E(\hat{t}_1 - t_1) = \sum_k p\, l_k e^{-l_k}\{e^{-(\hat{l}_k - l_k)(1-p)} - 1\}$$

$$\approx \sum_k p\, l_k e^{-l_k}\{-(\hat{l}_k - l_k)(1-p) + (\hat{l}_k - l_k)^2(1-p)^2/2\}$$

if the $\hat{l}_k - l_k$ are small. Under the Poisson assumption of equal mean and variance and ignoring estimation error in $\hat{b}$, $(y_k - \hat{u}_k)/p$ and $[(y_k - \hat{u}_k)^2 - y_k]/p^2$ will unbiasedly estimate $(l_k - \hat{l}_k)$ and $(\hat{l}_k - l_k)^2$ respectively. Hence, the bias of $\hat{t}_1$ may be expected to be reduced by minimising the absolute value of:

$$T_1 = \sum_k \frac{(1-p)}{p}\hat{l}_k e^{-\hat{l}_k}\{(y_k - \hat{u}_k)p + [(y_k - \hat{u}_k)^2 - y_k](1-p)/2\}$$

The statistic T_1 is a weighted mean of the $(y_k - \hat{u}_k)$ and the $[(y_k - \hat{u}_k)^2 - y_k]$. A similar expression applies for $\hat{t}_2$ but with different weights. The fitting of the log-linear model by IPF ensures that a weighted mean of the $(y_k - \hat{u}_k)$ is zero so the critical element of T_1 comes from the expression $[(y_k - \hat{u}_k)^2 - y_k]$ (in fact numerical work has shown that the $(y_k - \hat{u}_k)$ term usually only makes a very minor contribution to the value of T_1).

Choosing a model such that a weighted mean of the $[(y_k - \hat{u}_k)^2 - y_k]$ is close to zero may be interpreted as choosing a model which exhibits little under- or over-dispersion. This follows because y_k and $(y_k - \hat{u}_k)^2$ are unbiased estimators of the conditional mean and variance of y_k respectively, again ignoring differences between $\hat{b}$ and b and assuming $u_k = \exp(\mathbf{x}_k' b)$. Hence, an average of $[(y_k - \hat{u}_k)^2 - y_k]$ is a measure of over or under-dispersion. Cameron and Trivedi (1998, p.78) show that the hypothesis that $(y_k - u_k)^2$ and y_k share the same expectation may be tested by first defining $z_k = [(y_k - \hat{u}_k)^2 - y_k]/\hat{u}_k$ and then using OLS to estimate the regression model: $z_k = k + e_k$ where e_k is an error term. Either the estimate $\hat{k}$ or its associated

t-statistic, T_k, for testing $H_0 : k = 0$ may be taken as a measure of over- or under-dispersion. The statistic, T_k, is asymptotically normal and can be used to determine a class of models which do not exhibit significant departures from H_0.

An additional reason for avoiding under-dispersed models is that over-fitting may produce too many zeros on the margins leading to expected cell means being too high for the non-zero cells of the table and disclosure risk measures under- estimated. In contrast, under-fitting in over-dispersed models will produce no zeros on the margins and expected cell means may be too low for the non-zero cells of the table and disclosure risk measures over-estimated. Therefore, the model which manages the random and structural zeros of the contingency table will produce the best estimates for the disclosure risk measures.

4 Model Search Algorithm

We consider a model search algorithm that is similar to the TABU method introduced by Drezner, Marcoulides and Salhi (1999) for variable selection in multiple regression analysis. It is a local search method that depends on a criterion for the selection of the variables, a definition of a neighbourhood for each subset of the variables, a starting solution, a consistent method for moving through the neighbourhood and a stopping criterion. The neighbourhood is defined by adding a variable to the subset, removing a variable from the subset, and swapping variables.

As a starting solution, we begin with the all 2-way interactions log-linear model. This is motivated by the experience that it seems to lead to good estimates of the disclosure risk measures in many empirical experiments that we have undertaken. On the other hand, we have found that the independence log-linear model tends to be over-dispersed and leads to over-estimation of the disclosure risk measures. At the other extreme, the all 3-way interactions model tends to be under-dispersed and leads to under-estimation of the risk measures. Thus we expect a reasonable solution to lie between these extremes.

The algorithm is adapted to take into account the hierarchical structure of the log linear models. If we consider up to 2-way interactions for *k* independent variables, we obtain $k + \binom{k}{2}$ possible variables to examine. The variables, however, are highly dependent on each other. In the hierarchical log-linear modelling framework, for example, a model containing the interaction {*a*b}* means that the expected cell counts are fitted to the sample counts in the 2-way table defined by crossing variables *{a}* and *{b}*. Therefore, including into the model the separate variables {*a}* and {*b}* is redundant. Conversely, if we add into a model an independent term *{a},* we need to remove all interactions that involve that specific term *{a*b}*, *{a*c}*, etc.

When defining the neighbourhood of a chosen model by dropping terms, swapping terms and adding terms, we need to make sure that we are not checking models that were previously examined or will be examined at a later stage of the search.

Starting with the all 2-way interactions model and considering only independent and all 2-way interaction terms, the first round of the algorithm involves dropping each interaction in turn and then swapping in independent terms and removing the relevant interactions involved with the specific term. Note that there are no terms to add in for the first round since these only produce redundant models. For each model in the neighbourhood the goodness-of-fit criteria will determine the most appropriate model for continuing the search and defining the next neighborhood.

5 Practical Implementation

In this Section we present some results on how the risk assessment can be carried out at a National Statistics Institute such as the ONS. We first demonstrate on samples drawn from the 2001 UK Census where we compare the estimated disclosure risk measures with the true disclosure risk measures and check the performance of the model choice criteria. The second example will present an analysis on a real data set that is being considered for release to the UK data archive.

Example 1: Simulated samples from population census

Table 1 presents true and estimated global risk measures for simple random samples of different sizes drawn from two Estimation Areas of the 2001 UK Census (N=944,793). We demonstrate on two keys defined by cross-classifying six traceable and visible key variables. The first key has 412,080 cells (the number of categories is in parenthesis): Estimation Area (2), Sex (2), Age (101), Marital Status (6), Ethnicity (17), Economic Activity (10). The second key has 73,440 cells and is defined as the first key except that age was banded into 18 groupings. We ran two log-linear models: the independence model and the all 2-way interactions model.

Sample Size	Model	True		Estimates		Cameron-Trivedi Test		
		t_1	t_2	$\hat{t}_1$	$\hat{t}_2$	$\hat{k}$	T_k	T_1
Small Key								
4,724	Indep	23	68.2	54.2	126.9	0.5074	8.55	562.24
4,724	2-way			16.0	52.2	-0.0041	-3.62	-11.695
9,448	Indep	39	127.1	99.3	230.2	1.0316	8.58	1,447.20
9,448	2-way			37.8	117.9	-0.0051	-3.91	-30.952
18,896	Indep	75	215.3	174.3	355.7	2.0622	9.56	3,153.22
18,896	2-way			85.5	222.0	0.0059	2.00	16.891
Large Key								
4,724	Indep	80	183.9	197.4	385.1	0.0881	10.58	1,178.91
4,724	2-way			35.9	112.3	-0.0025	-7.96	-16.822
9,448	Indep	159	355.9	386.6	701.2	0.1846	14.42	3,400.76
9,448	2-way			104.9	280.1	-0.0036	-10.32	-59.257
18,896	Indep	263	628.9	672.0	1170.5	0.3865	16.77	7,269.90
18,896	2-way			252.0	591.3	-0.0030	-5.69	-43.594

Table 1. Global Risk Measures on Samples Drawn from the 2001 UK Census

In Table 1, the 2-way interactions model always leads to better estimates than the independence model and this is predicted in all cases by values of $\hat{k}$, T_k and T_1 being closer to 0. For the smaller key, the values of T_k for the 2-way interactions model are close to the critical values for accepting the null hypothesis of equal dispersion. For the larger key, the values of T_k suggest that the 2-way interactions model is over-fitting the data. We continue in our model search based on the large key and the 1% sample (n=9,448). Table 2 presents results of the first round of the neighbourhood search.

Model $t_1 = 159 \quad t_2 = 355.9$	Estimates		Cameron-Trivedi Test		T_1
	$\hat{t}_1$	$\hat{t}_2$	$\hat{k}$	T_k	
Independent	386.6	701.2	0.1846	14.42	3,400.76
All 2-way	104.9	280.1	-0.0036	-10.32	-59.257
Drop {ea*s}	104.6	279.8	-0.0035	-10.15	-58.969
Drop {ea*a}	105.3	281.3	-0.0032	-9.69	-61.684
Drop {ea*m}	103.8	279.1	-0.0034	-10.92	-63.851
Drop {ea*et}	108.7	290.0	-0.0024	-6.09	-58.230
Drop {ea*ec}	105.2	280.0	-0.0035	-10.60	-60.399
Drop {s*a}	104.5	280.7	-0.0033	-9.87	-60.699
Drop {s*m}	105.5	281.8	-0.0032	-8.53	-57.649
Drop {s*et}	105.2	280.3	-0.0035	-10.26	-58.949
Drop {s*ec}	103.2	281.5	-0.0018	-5.18	-64.670

Model $t_1 = 159 \quad t_2 = 355.9$	Estimates		Cameron-Trivedi Test		T_1
	$\hat{t}_1$	$\hat{t}_2$	$\hat{k}$	T_k	
Drop {a*m}	134.0	328.6	0.0071	9.42	-39.178
Drop {a*et}	147.0	346.2	0.0018	1.52	-38.477
Drop {a*ec}	184.7	419.2	0.0316	13.27	543.90
Drop {m*et}	108.7	287.5	-0.0032	-8.56	-59.692
Drop {m*ec}	108.3	284.0	-0.0028	-6.74	-51.510
Drop {et*ec}	132.3	308.2	-0.0015	-2.24	-20.147
In {ea} Out {ea*s}{ea*a}{ea*m}{ea*et}{ea*ec}	109.5	290.6	-0.0020	-5.72	-64.293
In {s} Out {ea*s}{s*a} {s*m}{s*et}{s*ec}	105.0	284.2	-0.0011	-3.13	-64.734
In {a} Out {ea*a}{s*a} {a*m}{a*et}{a*ec}	285.1	576.3	0.0803	18.43	487.31
In {m} Out {ea*m}{s*m}{a*m}{m*et}{m*ec}	134.3	355.5	0.0181	14.05	-62.752
In {et} Out {ea*et}{s*et}{a*et}{m*et}{et*ec}	190.7	396.5	0.0188	3.25	1,155.74
In {ec} Out {ea*ec}{s*ec}{a*ec}{m*ec}{et*ec}	207.7	464.0	0.0457	17.68	117.29

Table 2. Round 1 of the Neighbourhood Search for n=9,448 and K=412,080
(Note: Estimation Area–ea, Sex-s, Age–a, Marital Status–m, Ethnicity–et, and Economic Activity-ec)

From Table 2, removing the {et*ec} interaction provides the minimum value of T_1 and is also defining a model that accepts the null hypothesis of equal dispersion with a small estimate for parameter k. Therefore, we chose this model to continue to the second round of the model search. As mentioned, some models need not be checked in subsequent rounds because of the hierarchical structure of the log linear models. For example, deleting the last interaction {et*ec} means that there is no need to evaluate adding in {et} or {ec} and taking out their relevant interactions since this leads to the same models that were previously checked in round one.

Model $t_1 = 159 \quad t_2 = 355.9$	Estimates		Cameron-Trivedi Test		T_1
	$\hat{t}_1$	$\hat{t}_2$	$\hat{k}$	T_k	
Drop {et*ec}	132.3	308.2	-0.0015	-2.24	-20.147
Drop {ea*s}{et*ec}	132.3	308.2	-0.0015	-2.27	-20.594
Drop {ea*a}{et*ec}	133.4	310.4	-0.0011	-1.65	-14.781
Drop {ea*m}{et*ec}	131.9	307.9	-0.0014	-2.28	-28.398
Drop {ea*et}{et*ec}	139.8	320.8	-0.0002	-0.20	-2.909
Drop {ea*ec}{et*ec}	133.7	309.5	-0.0015	-2.33	-22.478
Drop {s*a}{et*ec}	132.1	309.2	-0.0013	-2.17	-32.570
Drop {s*m}{et*ec}	133.4	310.3	-0.0011	-1.58	-14.389
Drop {s*et}{et*ec}	132.4	308.5	-0.0015	-2.24	-21.111
Drop {s*ec}{et*ec}	130.9	310.3	0.0002	0.35	-38.516
Drop {a*m}{et*ec}	159.7	354.2	0.0091	10.27	-29.537
Drop {a*et}{et*ec}	173.4	370.2	0.0066	2.58	161.58
Drop {a*ec}{et*ec}	208.4	442.5	0.0324	13.38	573.72

Model $t_1 = 159 \quad t_2 = 355.9$	Estimates		Cameron-Trivedi Test		T_1
	$\hat{t}_1$	$\hat{t}_2$	$\hat{k}$	T_k	
Drop {m*et}{et*ec}	137.3	315.8	-0.0011	-1.68	-18.588
Drop {m*ec}{et*ec}	134.0	311.1	-0.0008	-1.10	-12.185
In {ea} Out {et*ec}{ea*s}{ea*a}{ea*m}{ea*et}{ea*ec}	141.3	321.7	0.0002	0.28	0.3363
In {s} Out {et*ec}{ea*s}{s*a}{s*m}{s*et}{s*ec}	132.6	313.0	0.0009	1.36	-37.947
In {a} Out {et*ec}{ea*a}{s*a}{a*m}{a*et}{a*ec}	313.4	596.5	0.0830	19.03	656.94
In {m} Out {et*ec}{ea*m}{s*m}{a*m}{m*et}{m*ec}	166.0	386.5	0.0221	12.64	66.937

Table 3: Round 2 of a Neighbourhood Search for n=9,448 and K=412,080
(Note: Estimation Area–ea, Sex-s, Age–a, Marital Status–m, Ethnicity–et, and Economic Activity-ea)

In Table 3, many models accept the null hypothesis of equal dispersion. The model {In {ea} Out {et*ec}{ea*s}{ea*a}{ea*m}{ea*et}{ea*ec}} has the minimum value of T_1 and also accepts the null hypothesis with a small value T_k. This model is our preferred model. We note that the models that accept the null hypothesis (with $|T_k| \leq 2.4$) are giving good estimated global risk measures compared to the true measures and therefore the global risk measures seem robust to slight deviations in the model.

Per-record risk measures are very important as a means of isolating high-risk sample uniques or groupings of sample uniques (i.e., particular age groups, etc.) for targeting disclosure control methods. In Figure 1 we examine the marginal distribution of the true and estimated per-record risk measures $\hat{r}_{2k}$ for the sample uniques within bands under the preferred model from Table 3. Table 4 presents their joint distribution.

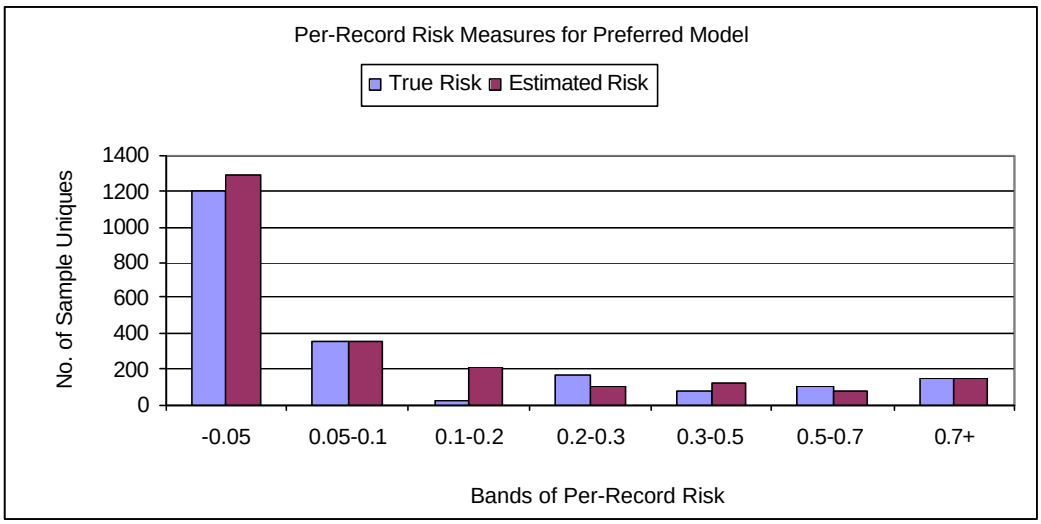

Figure 1: Per-Record Risk Measures for Preferred Model

True Per-Record Risk Measures	Estimated Per-Record Risk Measures			
	0 – 0.3	**0.3 – 0.7**	**0.7 – 1**	**Total**
0 – 0.3	1,838	97	26	1,961
0.3 – 0.7	75	57	52	184
0.7 – 1	45	49	65	159
Total	1,958	203	143	2,304

Table 4: Joint Distribution of Per-Record Risk Measures for Preferred Model: Cramer's V=0.4347

We obtain a good fit between the marginal distributions of the true and estimated per-record risk measures, although for the true high risk sample uniques, only 41% obtain a high estimated risk measure.

Example 2: large UK social survey

In this example, we look at an ONS dataset considered for release to the data archives. The sample size is n=530,013 with a sampling fraction of 0.9%. The microdata underwent disclosure control methods based on recoding key variables and eliminating other identifiable variables. We examined several combinations of key variables:

1) The key variables are: Region (20), Sex(2), Age Bands (45), Marital Status (6), Ethnicity (16) and Economic Activity (23). This resulted in a key of K=3,974,400 out of which 13,954 were sample uniques. The results for the all 2-way interactions log-linear model were the following: The estimated number of population uniques that are sample uniques is $\hat{t}_1 = 440.6$ which is 3.2% of the sample uniques. The expected number of correct matches is $\hat{t}_2 = 1{,}289.5$ which is 9.2% of the sample uniques or 0.2% of the entire sample. The values of the model choice criteria were $T_1 = 310.6$, $\hat{k} = 0.00423$ and $T_k = 7.03$. The slightly high values of T_1 and T_k leads us to expect some over-estimation of the disclosure risk measures.

2) The key is the same as key 1 except for replacing age bands with single years of age (100). This resulted in a key of K=8,832,000 out of which 39,588 sample uniques. This new key increased the disclosure risk. Based on the all 2-way interactions log-linear model, the estimated number of population uniques that are sample uniques is $\hat{t}_1 = 1{,}985.4$ which is 5.0% of the sample uniques. The expected number of correct matches is $\hat{t}_2 = 4{,}779.9$ which is 12.1% of the sample uniques or 0.9% of the entire sample. The model choice criteria were

$T_1 = 568.3$, $\hat{\kappa} = 0.00162$ and $T_k = 5.61$. These are also slightly high values and we expect some over-estimation of the disclosure risk measures.

3) The key is the same as key 1 except for replacing Economic Activity with Occupation (82). This resulted in a key of K=14,169,600 out of which 28,656 sample uniques. Because the key is so large it was necessary to partition the contingency table and carry out the disclosure risk assessment separately on each sub-table. Based on empirical work, it was found that disclosure risk assessment performs best when partitioning the contingency table according to a key variable that is correlated with the other key variables, since the partitioning key variable has an underlying interaction with the other variables. We partitioned the table into two sub-tables according to sex, and within each sub-table carried out an independence log-linear model. After combining the results from the two separate log-linear models, we obtained the following results: the estimated number of population uniques that are sample uniques is $\hat{t}_1 = 1{,}190.1$ which is 4.2% of the sample uniques. The expected number of correct matches is $\hat{t}_2 = 3{,}082.2$ which is 10.8% of the sample uniques or 0.6% of the entire sample. The model choice criteria were $T_1 = 337.0$, $\hat{\kappa} = 0.00021$ and $T_k = 2.43$. These model choice criteria are slightly better than previously obtained based on the other keys.

It is clear that more iterations are needed to determine the recoding of the variables on the final microdata to be released which would manage the disclosure risk while maximising the utility of the data. Also, a model search needs to be carried out in order to obtain a model that indicates acceptance of the null hypothesis of equal dispersion (i.e., the fit of the Poisson Model) and therefore more accurate estimated disclosure risk measures.

6 Discussion

In this paper we have examined the estimation of global and individual disclosure risk measures based on a Poisson log-linear model as developed by Skinner and Holmes (1998) and Elamir and Skinner (2004). We have addressed the implementation of model selection criteria for the large and sparse contingency tables spanned by key variables that are typical in the assessment of disclosure risk in microdata. Empirical results show that the goodness-of-fit criteria do select models that give good estimates for the disclosure risk measures based on the simple random samples that were drawn from the Census. There is a need for further empirical work to assess the impact of the size of the key on the goodness-of-fit criteria and model choice. In addition, since keys can be very large in practice, we need to develop

optimal methods for splitting contingency tables since this has implications for the types of models that can be assessed.

Future work on the Poisson model for disclosure risk assessment will focus on applications for hierarchical datasets and more complex survey designs, and in particular stratified samples with varying probabilities. In addition, variance and confidence intervals need to be developed for the estimated risk measures.

References

Benedetti, R., Capobianchi, A, and Franconi, L. (1998) Individual Risk of Disclosure Using Sampling Design Information.

Bethlehem, J., Keller, W., and Pannekoek, J. (1990) Disclosure Control of Microdata, JASA, Vol. 85.

Elamir, E. and Skinner, C. (2004) Record-level Measures of Disclosure Risk for Survey Microdata, Technical Paper, Southampton Statistical Sciences Research Institute, University of Southampton.

Cameron, A. C. and Trivedi. P.K. (1998), *Regression Analysis of Count Data*, Cambridge University Press, Cambridge.

Drezner, T., Marcoulides, G. and Salhi, S. (1999), TABU Search Model Selection in Multiple Regression Models, Communications in Statistics 28(2).

Hundepool, A., et. al. (2003) Mu-Argus Version 3.1 User's Manual, http://neon.vb.cbs.nl/casc/

Koehler, K.J. (1986) Goodness-of-Fit Tests for Log-Linear Models in Sparse Contingency Tables, Journal of the American Statistical Association, Vol. 81, No. 394 pp. 483-493.

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, http://neon.vb.cbs.nl/casc/

Rinott, Y. (2003) On Models for Statistical Disclosure Risk Estimation, Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Luxemburg, April 7-9, www.unece.org/stats/documents/2003/04/confidentiality/wp.16.e.pdf

Skinner, C. and Holmes, D. (1998), Estimating the Re-identification Risk Per Record in Microdata, JOS, Vol.14, 1998.