CASC PROJECT
CASC Computational Aspects of Statistical Confidentiality
March, 2004

Microdata: Application on GHS

Elsayed Elamir and Chris Skinner
Southampton Statistical Sciences Research Institute,
University of Southampton, SO17 1BJ, U.K.

Deliverable No: 1.2-D5

Contents

1 Introduction

In this report we investigate the statistical properties of disclosure risk measures for microdata at both the file and record level as presented in report 1 using data from the General Household Survey in U.K (GHS). The investigation will focus on statistical properties in terms of bias and variances for different scenarios of key variables and different sample sizes. For example, to what extent did the use of different sets of characteristics affect the calculated risk measures? Another question concerns the effect of the sample size.

2 GHS Data

The General Household Survey is a continuous national survey of people living in private households conducted on an annual basis by the U.K Office for National Statistics. It is a multi-purpose survey, carried out for a number of government departments. It provides information for planning and policy purpose, covering aspects of housing, employment, education, health and social services, transport, population and social security and is also used to monitor progress towards achieving targets. Microdata from the GHS have been released to academic users for many years from the U.K. Data Archive (www.data-archive.ac.uk). The data are currently released by the Economic and Social Data Service- Government (www.esds.ac.uk/government). We obtained GHS microdata for five years from the U.K Data Archive. These years are $1995-1996$; $1996-1997$; $1997-1998$; $1998-1999$. There are records for about 20000 individuals for each year.

3 Identifying variables

The risk measures given in report 1 are calculated assuming that an intruder knew an individual's values for a set of categorical identifying variables available on the GHS. This raises the question of which identifying variables to use. The GHS contains a number of characteristic that could be used as identifiers, covering aspects of housing, employment, education, health and social services, transport, population and social security. The approach is to try several scenarios corresponding to different set of variables known to an intruder and to examine the results. Possible intruder scenarios considered by Dale and Elliot (2001) included the variables, age; sex; marital status; country of birth; ethnic group; long-term limiting illness; primary economic activity; socio-economic; number of cars; central heating; water closet; bath; tenure; number of rooms; housing type, as the most readily available to a potential data intruder and coded at a level judged to be most reliable in terms of matching with the target file. We choose from these variables the following 5 variables:

1. X_1 sex in 2 categories;
2. X_2 marital status in 7 categories;
3. X_3 economic status in 13 categories;
4. X_4 socio-economic group 10 categories;
5. X_5 age in ten-year bands in 8 categories;

We consider two scenarios using 3 and 5 variables.

4 Simulation Set-up

We set up a simulation study as follows:

1. We begin with file unit values X_i for $i = 1, ..., N$.
2. Determine the elements of misclassification matrix $M_{jj^\star}$ for each variable. We use the validation study given by Forks (1994) to determine the elements of misclassification matrix for the study variables.
3. Create a new file (misclassification file) with values $\widetilde{X}_i$ for $i = 1, ..., N$ by applying the matrix $M_{jj^\star}$ independently for each record on the study identifying variables.
4. Draw a simple random sample with sampling fraction π from the population.
5. If we have a unique value we have to make sure that $X_{ki} = \widetilde{X}_{ki}$, $k = 1, ..., m$, for all categories, if so, $\mathrm{I}\left(f_{X_i} = 1, \widetilde{X}_i = X_i\right) = 1$, if not, $\mathrm{I}\left(f_{X_i} = 1, \widetilde{X}_i = X_i\right) = 0$.
6. Compute θ_m, θ_{mm}, $\widehat{\theta}_m$, $\widehat{\theta}_{mm}$ and their variances as given in Report 1.

5 Results for File Level Measures

In the analysis we considered three or five identifying variables and different sample fractions to see how different sets of variables and sample sizes can affect the measures. Also, we studied the effect of misclassification by misclassifying some variables. The misclassification range in some variables from 0.97 in sex to 0.75 in socio-economic group. We use the Census validation survey given by Forks (1994) to determine the misclassification in each variable. The results for θ, θ_m, θ_{mm}, $\widehat{\theta}$, $\widehat{\theta}_m$, $\widehat{\theta}_{mm}$ and their variances given in report 1 are shown in Tables 2, 3, 5, 7, 8, 10, 11, 12, 13. The population size for the years from $1995 - 1999$ was 83000 and for the $1998 - 1999$ was 33000.

These tables shows that

1. The parameters θ, θ_m and θ_{mm} is a sample dependent and is increasing with increase the sample sizes which reflect more risk.

2. In terms of bias, the $\widehat{\theta}$ is a good estimator to θ while $\widehat{\theta}_m$ and $\widehat{\theta}_{mm}$ are good estimators to their parameters θ_m, θ_{mm} in the case of misclassification is high; see, for example, Table 7.

3. On the other hand, in terms of bias by increasing the misclassification and number of misclassified variables the estimators $\widehat{\theta}_m$ and $\widehat{\theta}_{mm}$ are less reliable for estimation their parameters; see, for example, Tables 5 and 11.

4. In terms variance, the variance estimators in few cases are good but in most other cases are not for the estimators $\widehat{\theta}$, $\widehat{\theta}_m$ and $\widehat{\theta}_{mm}$ although $\widehat{\theta}$ is less effected.

5. This makes us ask if we are interested in estimators which gives less bias or gives less variance or both of them.

6. If we just interested in bias the results show that $\widehat{\theta}$ might be used for helping in taking decision of releasing microdata.

	$\pi = 0.02$	$\pi = 0.05$	$\pi = 0.10$
Risk Measure, θ	0.01256	0.03627	0.05243
Estimator, $\widehat{\theta}$	0.01328	0.03714	0.05316
Bias, $\widehat{\theta} - \theta$	0.00072	0.00087	0.00073
standard error(θ)	0.00480	0.00736	0.00823
standard error($\widehat{\theta}$)	0.00234	0.00833	0.01451
standard error($\widehat{\theta}$-θ)	0.00645	0.00845	0.00734

Table 1: Results for GHS 1995−1999 using three variables sex (2), economic status (13) and marital status (7) and there is no misclassification.

	$\pi = 0.02$	$\pi = 0.05$	$\pi = 0.10$
Risk measure, θ_m	0.01141	0.02374	0.03406
Estimator, $\widehat{\theta}_m$	0.01196	0.02334	0.03422
Bias, $\widehat{\theta}_m - \theta_m$	0.00054	−.00040	0.00016
standard error(θ_m)	0.00271	0.01503	0.00407
standard error($\widehat{\theta}_m$)	0.00374	0.00504	0.00291
standard error($\widehat{\theta}_m$-θ_m)	0.00300	0.01232	0.01943
Risk Measure, θ_{mm}	0.01283	0.03327	0.03946
Estimator, $\widehat{\theta}_{mm}$	0.01793	0.04681	0.03842
Bias, $\widehat{\theta}_{mm} - \theta_{mm}$	0.00510	0.01353	0.00104
standard error(θ_{mm})	0.00312	0.01132	0.00511
standard error($\widehat{\theta}_{mm}$)	0.01078	0.01542	0.01921
standard error($\widehat{\theta}_{mm}$-θ_{mm})	0.01152	0.02057	0.03223

Table 2: Results for GHS 1995−1999 using three variables sex (2), economic status (13) and marital status (7) and misclassification in one variable, economic status(85%).

	$\pi = 0.02$	$\pi = 0.05$	$\pi = 0.10$
Risk measure, θ_m	0.00829	0.01631	0.02134
Estimator, $\widehat{\theta}_m$	0.00846	0.01598	0.02216
Bias, $\widehat{\theta}_m - \theta_m$	0.00016	−.00033	0.00082
standard error(θ_m)	0.00322	0.01404	0.01230
standard error($\widehat{\theta}_m$)	0.00136	0.00241	0.00146
standard error($\widehat{\theta}_m$-θ_m)	0.00299	0.01189	0.01097
Risk Measure, θ_{mm}	0.0059	0.01357	0.02479
Estimator, $\widehat{\theta}_{mm}$	0.01920	0.03638	0.05114
Bias, $\widehat{\theta}_{mm} - \theta_{mm}$	0.01324	0.02267	0.02635
standard error(θ_{mm})	0.00155	0.00159	0.00928
standard error($\widehat{\theta}_{mm}$)	0.00314	0.00675	0.03646
standard error($\widehat{\theta}_{mm}$-θ_{mm})	0.00404	0.00622	0.04328

Table 3: Results for GHS 1995 − 1999 three variables sex (2), economic status (13) and marital status (7) and misclassification in two variables, economic status (85%), marital status (93%).

	$\pi = 0.02$	$\pi = 0.05$	$\pi = 0.10$
Risk Measure, θ	0.02586	0.05337	0.09307
Estimator, $\widehat{\theta}$	0.02781	0.05599	0.09482
Bias, $\widehat{\theta} - \theta$	0.00195	0.00262	0.00175
standard error(θ)	0.00162	0.00310	0.00413
standard error($\widehat{\theta}$)	0.00460	0.00537	0.00793
standard error($\widehat{\theta}$-θ)	0.00503	0.00642	0.00731

Table 4: Results for GHS $1995 - 1999$ five variables sex (2), economic status (13), marital status (7), socio-economic group (10) and age (8) and there is no misclassification.

	$\pi = 0.02$	$\pi = 0.05$	$\pi = 0.10$
Risk measure, θ_m	0.01019	0.01906	0.03494
Estimator, $\widehat{\theta}_m$	0.02311	0.03038	0.04312
Bias, $\widehat{\theta}_m - \theta_m$	0.01292	0.01132	0.00818
standard error(θ_m)	0.00171	0.00337	0.00409
standard error($\widehat{\theta}_m$)	0.00168	0.00238	0.00362
standard error($\widehat{\theta}_m$-θ_m)	0.00193	0.00390	0.00427
Risk Measure, θ_{mm}	0.00778	0.01583	0.02899
Estimator, $\widehat{\theta}_{mm}$	0.02836	0.03699	0.06403
Bias, $\widehat{\theta}_{mm} - \theta_{mm}$	0.02058	0.02116	0.03504
standard error(θ_{mm})	0.00071	0.00152	0.00228
standard error($\widehat{\theta}_{mm}$)	0.00345	0.00403	0.00594
standard error($\widehat{\theta}_{mm}$-θ_{mm})	0.00336	0.00396	0.00544

Table 5: Results for GHS $1995 - 1999$ five variables sex (2), economic status (13), marital status (7), socio-economic group (10) and age (8) and misclassification in one variable, socio-economic group (75%).

	$\pi = 0.02$	$\pi = 0.05$	$\pi = 0.10$
Risk Measure, θ	0.04091	0.08750	0.15594
Estimator, $\widehat{\theta}$	0.04307	0.08786	0.15517
Bias, $\widehat{\theta} - \theta$	0.00215	0.00035	$-.00076$
standard error(θ)	0.00307	0.00575	0.00790
standard error($\widehat{\theta}$)	0.00696	0.011430	0.017677
standard error($\widehat{\theta}$-θ)	0.00799	0.013655	0.01956

Table 6: Results for GHS $1998 - 1999$ using five variables sex (2), marital status (7), economic status (13), soci-economic status (10) and age (6) and misclassification in five variables, sex (0.97), marital status (0.90), economic status (0.90), soci-economic status (0.90) and age (0.90).

	$\pi = 0.02$	$\pi = 0.05$	$\pi = 0.10$
Risk measure, θ_m	0.02878	0.05064	0.07888
Estimator, $\widehat{\theta}_m$	0.01935	0.04401	0.08201
Bias, $\widehat{\theta}_m - \theta_m$	$-.00942$	$-.00663$	0.00312
standard error(θ_m)	0.00333	0.00638	0.00852
standard error($\widehat{\theta}_m$)	0.00138	0.00280	0.00479
standard error($\widehat{\theta}_m$-θ_m)	0.00385	0.00751	0.00936
Risk Measure, θ_{mm}	0.02355	0.05130	0.09277
Estimator, $\widehat{\theta}_{mm}$	0.02755	0.05621	0.09875
Bias, $\widehat{\theta}_{mm} - \theta_{mm}$	0.00399	0.00491	0.00597
standard error(θ_{mm})	0.00250	0.00485	0.00657
standard error($\widehat{\theta}_{mm}$)	0.00445	0.00731	0.01120
standard error($\widehat{\theta}_{mm}$-θ_{mm})	0.00568	0.00952	0.01328

Table 7: Results for GHS $1998 - 1999$ using five variables sex (2), marital status (7), economic status (13), soci-economic status (10) and age (8) and misclassification in five variables, sex (0.97), marital status (0.90), economic status (0.90), soci-economic status (0.90) and age (0.90).

	$\pi = 0.02$	$\pi = 0.05$	$\pi = 0.10$
Risk measure, θ_m	0.01887	0.03151	0.04648
Estimator, $\widehat{\theta}_m$	0.00781	0.02708	0.04201
Bias, $\widehat{\theta}_m - \theta_m$	$-.01106$	$-.00443$	$-.00447$
standard error(θ_m)	0.00228	0.00391	0.00592
standard error($\widehat{\theta}_m$)	0.00020	0.00052	0.00096
standard error($\widehat{\theta}_m$-θ_m)	0.00230	0.00395	0.00594
Risk Measure, θ_{mm}	0.01257	0.02726	0.05006
Estimator, $\widehat{\theta}_{mm}$	0.01301	0.02911	0.05084
Bias, $\widehat{\theta}_{mm} - \theta_{mm}$	0.00043	0.00185	0.00077
standard error(θ_{mm})	0.00166	0.00334	0.00483
standard error($\widehat{\theta}_{mm}$)	0.00176	0.00371	0.00579
standard error($\widehat{\theta}_{mm}$-θ_{mm})	0.00244	0.00491	0.00736

Table 8: Results for GHS $1998 - 1999$ using five variables sex (2), marital status (7), economic status (13), so-economic status (10) and age (8) and misclassification in five variables, sex (0.80), marital status (0.80), economic status (0.80), socio-economic status (0.80) and age (0.80).

	$\pi = 0.02$	$\pi = 0.05$	$\pi = 0.10$
Risk Measure, θ	0.01979	0.04178	0.07875
Estimator, $\widehat{\theta}$	0.02125	0.04519	0.08032
Bias, $\widehat{\theta} - \theta$	0.00146	0.00340	0.00157
standard error(θ)	0.00350	0.00633	0.01736
standard error($\widehat{\theta}$)	0.0109	0.04039	0.03391
standard error($\widehat{\theta}$-θ)	0.01177	0.04177	0.04167

Table 9: Results for GHS $1998 - 1999$ using three variables sex (2), marital status (7) and soci-economic status (10) and there is no misclassification.

	$\pi = 0.02$	$\pi = 0.05$	$\pi = 0.10$
Risk measure, θ_m	0.005838	0.00669	0.00809
Estimator, $\widehat{\theta}_m$	0.00461	0.00567	0.00611
Bias, $\widehat{\theta}_m - \theta_m$	$-.00123$	$-.00101$	$-.00197$
standard error(θ_m)	0.00163	0.00167	0.00275
standard error($\widehat{\theta}_m$)	0.00073	0.00067	0.00065
standard error($\widehat{\theta}_m$-θ_m)	0.00163	0.00179	0.00284
Risk Measure, θ_{mm}	0.00999	0.01934	0.03779
Estimator, $\widehat{\theta}_{mm}$	0.01112	0.02493	0.04112
Bias, $\widehat{\theta}_{mm} - \theta_{mm}$	0.00113	0.00558	0.00333
standard error(θ_{mm})	0.00279	0.00539	0.013136
standard error($\widehat{\theta}_{mm}$)	0.00565	0.02068	0.01736
standard error($\widehat{\theta}_{mm}$-θ_{mm})	0.00632	0.02149	0.02402

Table 10: Results for GHS $1998 - 1999$ using three variables sex (2), marital status (7) and soci-economic status (10) and misclassification in three variables, sex (0.80), marital status (0.80) and soci-economic status (0.80).

	$\pi = 0.02$	$\pi = 0.05$	$\pi = 0.10$
Risk measure, θ_m	0.01066	0.01718	0.02449
Estimator, $\widehat{\theta}_m$	0.00870	0.01174	0.01340
Bias, $\widehat{\theta}_m - \theta_m$	$-.00196$	$-.00544$	$-.01108$
standard error(θ_m)	0.00248	0.00438	0.00963
standard error($\widehat{\theta}_m$)	0.00168	0.00182	0.00241
standard error($\widehat{\theta}_m$-θ_m)	0.00273	0.00477	0.00998
Risk Measure, θ_{mm}	0.01432	0.03034	0.06057
Estimator, $\widehat{\theta}_{mm}$	0.01504	0.03228	0.06711
Bias, $\widehat{\theta}_{mm} - \theta_{mm}$	0.00716	0.00193	0.00653
standard error(θ_{mm})	0.00328	0.00783	0.01845
standard error($\widehat{\theta}_{mm}$)	0.00631	0.01446	0.03006
standard error($\widehat{\theta}_{mm}$-θ_{mm})	0.00704	0.01734	0.03585

Table 11: Results for $1998 - 1999$ using three variables sex (2), marital status (7) and soci-economic status (10) and misclassification in three variables, sex (0.90), marital status (0.90) and soci-economic status (0.90).

	$\pi = 0.02$	$\pi = 0.05$	$\pi = 0.10$
Risk measure, θ_m	0.01646	0.02995	0.04597
Estimator, $\widehat{\theta}_m$	0.01495	0.02373	0.03128
Bias, $\widehat{\theta}_m - \theta_m$	$-.00150$	$-.00621$	$-.01469$
standard error(θ_m)	0.00301	0.00752	0.01624
standard error($\widehat{\theta}_m$)	0.00505	0.00611	0.01181
standard error($\widehat{\theta}_m$-θ_m)	0.00649	0.00998	0.01864
Risk Measure, θ_{mm}	0.01779	0.03789	0.07373
Estimator, $\widehat{\theta}_{mm}$	0.01846	0.03948	0.07882
Bias, $\widehat{\theta}_{mm} - \theta_{mm}$	0.00067	0.00158	0.00508
standard error(θ_{mm})	0.00334	0.00866	0.01892
standard error($\widehat{\theta}_{mm}$)	0.00835	0.01548	0.03494
standard error($\widehat{\theta}_{mm}$-θ_{mm})	0.00985	0.01831	0.04080

Table 12: Results for 1998−1999 using three variables sex (2), marital status (7) and soci-economic status (10) and misclassification in one variable, soci-economic status (0.90).

	$\pi = 0.02$	$\pi = 0.05$	$\pi = 0.10$
Risk measure, θ_m	0.01302	0.02174	0.03003
Estimator, $\widehat{\theta}_m$	0.01116	0.01557	0.01840
Bias, $\widehat{\theta}_m - \theta_m$	$-.00185$	$-.00617$	$-.01161$
standard error(θ_m)	0.00302	0.00652	0.01649
standard error($\widehat{\theta}_m$)	0.00307	0.00415	0.00799
standard error($\widehat{\theta}_m$-θ_m)	0.00436	0.00756	0.01760
Risk Measure, θ_{mm}	0.01506	0.03238	0.06059
Estimator, $\widehat{\theta}_{mm}$	0.01582	0.03620	0.07347
Bias, $\widehat{\theta}_{mm} - \theta_{mm}$	0.00076	0.00382	0.01287
standard error(θ_{mm})	0.00324	0.00733	0.01775
standard error($\widehat{\theta}_{mm}$)	0.00777	0.01482	0.03456
standard error($\widehat{\theta}_{mm}$-θ_{mm})	0.00863	0.01730	0.04001

Table 13: Results for GHS 1998 − 1999 using three variables sex (2), marital status (7) and soci-economic status (10) and misclassification in one variable, soci-economic status (0.80).

6 Record Level Measures

6.1 Study Set-up

In this section we seek to evaluate the properties of the $\widehat{\theta}_j$ empirically using an artificial finite population. We wish to avoid basing our evaluation on any single assumed model and hence cannot simply compare the values of $\widehat{\theta}_j$ with 'true values' θ_j, since the latter are defined with respect to a model. We therefore adopt two alternative approaches. First, we study the relation between $\widehat{\theta}_j$ and the empirical proportion of population uniques among sample unique units. Second, we study the relation between the average value of $\widehat{\theta}_j$ and the average value of $1/F_j$ within subgroups. For $\widehat{\theta}_j$ to be a useful measure, we expect a strong positive relationship in the first case and a strong positive relationship, with approximate equality between the two averages, in the second case.

As a basis for studying these relationships, we constructed an artificial population file by combining data for two years (1996,1997) from the U.K. General Household Survey, resulting in records on $N = 33142$ individuals. Following consideration of possible intruder scenarios by Dale and Elliot (2001), we used the following five variables: sex; marital status; economic status; socio-economic group; age in ten-years band. We evaluated the estimated measures of disclosure risk for two simple random samples from this population, one of size $n = 2500$ ($\pi = 0.075$) and one of size $n = 5000$ ($\pi = 0.15$).

6.2 Results

The numbers of sample uniques were $n_1 = 370$ in the first sample and $n_1 = 495$ in the second sample; see, Tables 14 and 15. The four file-level measures of risk were:

- sample 1 ($n = 2500$) : $\Pr(PU) = 0.024, \Pr(PU|SU) = 0.159$, $\theta_U = 0.115, \theta_s = 0.313$;
- sample 2 ($n = 5000$) : $\Pr(PU) = 0.026, \Pr(PU|SU) = 0.262$, $\theta_U = 0.210, \theta_s = 0.443$.

As expected, we find $\Pr(PU) \leq \Pr(PU|SU) \leq \theta_s$ and $\theta_U \leq \theta_s$ for both samples so that θ_s is the most conservative measure.

We next compute values of $\widehat{\theta}_j$ for each of the sample unique cases in each sample. We first assume fixed λ_j and compute $\widehat{\theta}_j$ using iterative proportional fitting, for the following two specifications of Poisson model:

- Model 1: a log-linear model including all main effects;
- Model 2 : a log-linear model including also all two-factor interactions.

Tables 16, 17, 18 and 19 show the distributions of $\widehat{\theta}_j$ across sample unique cases for these two models for both samples. For the first sample ($n = 2500$), we find the mean values of $\widehat{\theta}_j$ to be 0.442 and 0.296 for Models 1 and 2 respectively, compared with the 'expected' mean $\theta_s = 0.313$. For the second sample ($n = 5000$) we find mean values of $\widehat{\theta}_j$ of 0.513 and 0.435 for the two models, compared with $\theta_s = 0.443$. The correspondence with θ_s seems rather better for Model 2. (This suggests a means of estimating θ_s to augment the simpler approach to estimating θ_U discussed by Skinner and Elliot (2002)). In all cases θ_U understates substantially the average record-level measure.

The five divisions of the range $[0, 1]$ for $\widehat{\theta}_j$ in Tables 16 and 17 define subsets of sample uniques with similar values of $\widehat{\theta}_j$. For each of these subsets, the proportion of population unique cases are presented in these tables. As in Skinner and Holmes (1998), we find that the $\widehat{\theta}_j$ are useful for deciding whether a sample unique case is population unique, with Model 2 providing better discrimination. For the first sample, it is more likely than not that a sample unique is population unique if $\widehat{\theta}_j > 0.8$ for Model 2, but not for Model 1. The ability to detect population uniques with high probability is even stronger for the second sample.

Tables 18 and 19 give the results when λ_j is random and follows a gamma distribution, as discussed in report 1. We find similar results to the model with no overdispersion, with no evidence of improved discrimination for the model with random effects.

We next study the relationship between the mean of $\widehat{\theta}_j$ and the mean of $1/F_j$ within the 40 (=2+7+13+10+8) subgroups defined by the univariate categories of the five key variables for sample unique records for each of the two samples. Tables 20 and 21 gives the results for the main effects and all two-way interaction models for $\pi = 0.075$ and 0.15. Given the lack of evidence of improved performance using random effects, we only consider the model with λ_j fixed. We find, as expected, a strong relationship between the mean of the $\widehat{\theta}_j$ and the mean of the values $1/F_j$. The two means are

$N = 33142$, $\pi = .075$	$n = 2500$	Key values= 14560	k. v.= 5
F_j	Freq.	f_j	Freq.
0	12481	0	13911
1	707	1	370
2	319	2	96
3	191	3	45
4	118	4	27
5	90	5	22
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\sum F\mathrm{I}\,(f=1) = 3202$	$N_1 = 707$	$n_1 = 370$, $n_2 = 96$	$N_1,\ n_1 = 59$
File risk measure	θ	$\widehat{\theta}_j$	$P\,(\mathrm{pu} \mid \mathrm{su})$
	0.115	0.135	0.159

Table 14: Results of general household survey (96 and 97) using five key variables (k.v.): Sex (2), Marital status (7), Economic status (13), Socio-economic group (10) and Age (8) with $\pi = 0.075$.

$N = 33142$, $\pi = 0.15$	$n = 5000$	Key values= 14560	K. v.= 5
F_j	Freq.	f_j	Freq.
0	12481	0	13580
1	707	1	495
2	319	2	166
3	191	3	71
4	118	4	46
5	90	5	37
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\sum F\mathrm{I}\,(f=1) = 2348$	$N_1 = 707$	$n_1 = 495$, $n_2 = 166$	$N_1,\ n_1 = 130$
	θ	$\widehat{\theta}_j$	$P\,(\mathrm{pu} \mid \mathrm{su})$
	0.210	0.208	0.262

Table 15: Results of general household survey (96 and 97) using five key variables (K.v.): Sex (2), Marital status (7), Economic status (13), Socio-economic group (10) and Age (8) with $\pi = 0.15$

		Model 1		Model 2
$\widehat{\theta}_j$	Freq.	Prop. Pop. Unique	Freq.	Prop. Pop. Unique
0−	84	0.07	113	0.07
0.20−	61	0.11	68	0.08
0.40−	88	0.13	78	0.09
0.60−	79	0.19	67	0.18
0.80 − 1	58	0.33	44	0.59
Total	370		370	

Table 16: Frequency distributions of $\widehat{\theta}_j$ (Freq.) and Proportions of population unique records (Prop. Pop. Unique) for models 1 and 2 with no overdispersion and $n = 2500$.

		Model 1		Model 2
$\widehat{\theta}_j$	Freq.	Prop. Pop. Unique	Freq.	Prop. Pop. Unique
0−	79	0.05	105	0.06
0.20−	64	0.08	86	0.06
0.40−	85	0.15	79	0.10
0.60−	87	0.22	59	0.27
0.80 − 1	55	0.34	41	0.58
Total	370		370	

Table 17: Frequency distributions of $\widehat{\theta}_j$ (Freq.) and Proportions of population unique records (Prop. Pop. Unique) for models 1 and 2 with overdispersion and $n = 2500$.

		Model 1		Model 2
$\widehat{\theta}_j$	Freq.	Prop. Pop. Unique	Freq.	Prop. Pop. Unique
0−	110	0.11	137	0.07
0.20−	94	0.11	92	0.08
0.40−	98	0.12	88	0.14
0.60−	92	0.42	76	0.49
0.80 − 1	101	0.55	92	0.70
Total	495		495	

Table 18: Frequency distributions of $\widehat{\theta}_j$ (Freq.) and Proportions of population unique records (Prop. Pop. Unique) for models 1 and 2 with no overdispersion and $n = 5000$.

		Model 1		Model 2
$\widehat{\theta}_j$	n_1	Prop. Pop. Unique	n_1	Prop. Pop. Unique
0–	88	0.09	114	0.08
0.20–	123	0.17	146	0.20
0.40–	102	0.23	111	0.23
0.60–	99	0.32	83	0.45
$0.80-1$	83	0.54	41	0.71
Total	495		495	

Table 19: Frequency distributions of $\widehat{\theta}_j$ (Freq.) and Proportions of population unique records (Prop. Pop. Unique) for models 1 and 2 with overdispersion and $n = 5000$.

broadly similar for all the subgroups h, except for some cases where the size of the subgroup, n_h, is small. The correlation coefficients between the two means are 0.76 and 0.82 for the two models with $\pi = 0.075$ and 0.75 and 0.96 for the models with $\pi = 0.15$. It is clearly preferable to include the two-way interaction in the model.

Regression curves, obtained using the loess method (locally weighted regression scatter plot smoothing; see, Cleveland (1979) and Bowman and Azzalini (1997)) are displayed in Figure 1 for the data in Tables 20 and 21. They confirm the strong linear relationship between the mean of $\widehat{\theta}_j$ and the mean of $1/F_j$, especially for the model including two-way interactions.

7 Conclusion

Skinner and Elliot (2002) argued in favor of measuring disclosure risk at the file level by the probability that an observed match is correct rather than by the probability of population uniqueness. We have shown how the record-level measure of disclosure risk of Skinner and Holmes (1998), defined in terms of the probability of population uniqueness, may be extended in a parallel way to misclassification and a record-level measure of the probability that an observed match is correct. Both measures depend on the specification of a log-linear model for an assumed set of key variables. In an empirical evaluation of different versions of the new record-level measure using real survey data, we found evidence of discrimination by the measure between records of different levels of risk, in particular records which are very likely to be population unique could be identified by consideration of records with high values of the measure. We found no evidence, however, that allowance for overdispersion via the inclusion of random effects

Variable	Subpop. h	Samp. Unique n_{1h}	Mean of $\widehat{\theta}_j$ Model 1	Model 2	mean of F_j^{-1}
Sex	1	202	0.431	0.300	0.333
	2	168	0.455	0.292	0.288
Marital	3	108	0.291	0.224	0.266
status	4	40	0.522	0.385	0.349
	5	94	0.401	0.296	0.294
	6	31	0.393	0.221	0.202
	7	62	0.593	0.347	0.351
	8	33	0.690	0.410	0.465
	9	2	0.988	0.932	1
Economic	10	104	0.197	0.214	0.206
status	11	7	0.926	0.245	0.541
	12	3	0.921	0.167	0.541
	13	33	0.506	0.327	0.308
	14	33	0.577	0.425	0.472
	15	34	0.610	0.362	0.345
	16	58	0.316	0.281	0.308
	17	38	0.597	0.269	0.235
	18	6	0.831	0.386	0.478
	19	14	0.806	0.467	0.587
	20	8	0.256	0.405	0.441
	21	2	0.977	0.661	0.75
	22	30	0.293	0.213	0.182
Socioeco.	23	26	0.349	0.325	0.338
group	24	40	0.388	0.291	0.380
	25	42	0.434	0.290	0.286
	26	46	0.374	0.287	0.310
	27	58	0.405	0.259	0.253
	28	73	0.496	0.267	0.285
	29	42	0.524	0.369	0.327
	30	8	0.256	0.405	0.441
	31	29	0.561	0.338	0.340
	32	6	0.602	0.208	0.444
Age	33	26	0.634	0.272	0.361
	34	28	0.627	0.283	0.315
	35	60	0.463	0.297	0.274
	36	72	0.403	0.288	0.292
	37	64	0.437	0.294	0.344
	38	40	0.426	0.311	0.315
	39	50	0.449	0.332	0.311
	40	30	0.531	0.431	0.409

Table 20: means of $\widehat{\theta}_j$ and $1/F_j$ across forty subpopulations (subpop.) defined by Sex (2), Marital status (7), Economic status (13), Socio-economic group (10) and Age (8) for sample unique records with models 1 and 2 and $n = 2500$.

Variable	Suppop. h	Samp. Unique n_{1h}	Mean of $\widehat{\theta}_j$ Model 1	Model 2	Mean of F_j^{-1}
Sex	1	231	0.510	0.448	0.442
	2	264	0.515	0.423	0.443
Marital status	3	119	0.352	0.355	0.379
	4	59	0.619	0.538	0.500
	5	123	0.468	0.405	0.408
	6	53	0.407	0.364	0.361
	7	80	0.628	0.476	0.488
	8	55	0.732	0.532	0.538
	9	6	0.958	0.817	0.875
Economic status	10	125	0.267	0.315	0.338
	11	5	0.958	0.500	0.475
	12	7	0.975	0.653	0.671
	13	55	0.583	0.420	0.421
	14	42	0.668	0.468	0.471
	15	56	0.656	0.490	0.465
	16	60	0.373	0.407	0.402
	17	65	0.551	0.440	0.452
	18	8	0.878	0.728	0.783
	19	32	0.840	0.688	0.674
	20	14	0.212	0.594	0.470
	21	1	0.976	0.933	1
	22	25	0.610	0.521	0.502
Socioeco. group	23	28	0.500	0.500	0.547
	24	54	0.410	0.440	0.461
	25	51	0.496	0.405	0.409
	26	72	0.45	0.429	0.418
	27	70	0.485	0.421	0.430
	28	89	0.550	0.416	0.399
	29	59	0.610	0.438	0.422
	30	14	0.212	0.594	0.470
	31	50	0.656	0.480	0.506
	32	8	0.675	0.641	0.629
Age	33	33	0.721	0.480	0.506
	34	55	0.636	0.470	0.465
	35	83	0.508	0.431	0.443
	36	100	0.514	0.397	0.406
	37	72	0.479	0.424	0.401
	38	68	0.505	0.476	0.517
	39	49	0.535	0.473	0.464
	40	35	0.196	0.358	0.324

Table 21: means of $\widehat{\theta}_j$ and $1/F_j$ across forty subpopulations (subpop.) defined by Sex (2), Marital status (7), Economic status (13), Socio-economic group (10) and Age (8) for sample unique records with models 1 and 2 and $n = 5000$.

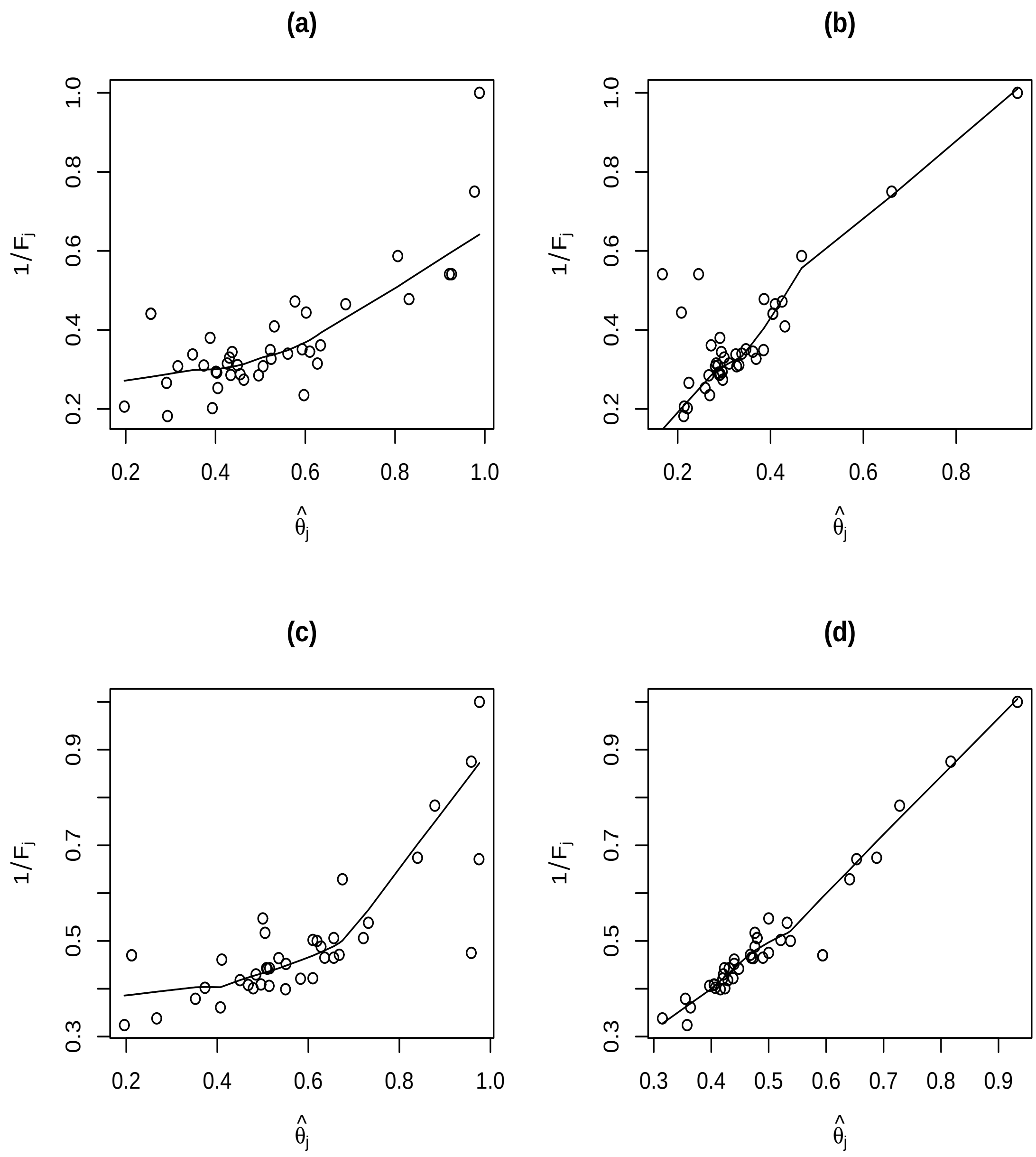

Figure 1: Scatter plot of mean of estimated measure of risk $\widehat{\theta}_j$ and mean of $1/F_j$ and loess curves with smoother span 2/3 for (a) Model 1 with $n = 2500$, (b) Model 2 with $n = 2500$, (c) Model 1 with $n = 5000$, (d) Model 2 with $n = 5000$.

in the model improved its performance. The measure obtained under the simpler model with no random effects was validated by comparing its average value in forty subpopulations with the 'true' population quantity it was estimating and the relationship was found to be very good for a model including only one and two-way interactions. This measure is much easier to compute, requiring only the fitting of a standard log-linear model, than the measure proposed by Skinner and Holmes (1998), which additionally required numerical integration. In summary, we suggest for use in practice the measure obtained from Poisson model for a log-linear model with main effects and two-way interactions. We are currently exploring the robustness of the measure to model choice and whether any improvements can be obtained through the use of higher-order interactions and model selection techniques.

The measure obtained from Poisson and Poisson-gamma models ignores any error in estimating the parameters β of the log-linear model by $\widehat{\beta}$. In principle, if the true measure is taken as the posterior probability of a correct match from a Bayesian perspective and if uncertainty about β can be represented in an appropriate way (this may need to take account of the complexity of the survey sampling scheme) then this uncertainty could be integrated out, perhaps using a simulation-based approach. We have not pursued this possibility, however, and suspect that it is more important initially to explore the dependence of the measure on model specification. The measures which accommodate for misclassification need more study and investigations.

References

Bowman, A. and A. Azzalini (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations.* Clarendon: Oxford.

Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association 74*, 829–836.

Dale, A. and M. Elliot (2001). Proposals for 2001 samples of anonymized records: An assessment of disclosure risk. *Journal of Royal Statistical Society, Ser. A 164*, 427–447.

Forks, K. (1994). The general household survey. *Report of the 1991 census. Unpublished.*

Skinner, C. and M. Elliot (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society, Series B 64*, 855–867.

Skinner, C. and D. Holmes (1998). Estimating the re-identification risk per record in microdata. *Journal of Official Statistics 14*, 361–372.