Dissemination of Microdata Files Stemming from Social Surveys Improvements of the Estimation of the Re-identification Risk Measures Based on Log-Linear Models

Daniela Ichim[1]

Servizio di Progettazione e Supporto Metodologico nei Processi di
Produzione Statistica
Istituto Nazionale di Statistica
via Cesare Balbo 16, 00184 Roma Italia
ichim@istat.it

1 Introduction

To face the increasing demand from users, the National Statistical Institutes disseminate more often microdata files. Such dissemination should be constrained to the confidentiality pledge under which a statistical agency collects survey data. To protect the confidentiality of respondents, a statistical disclosure control methodology is generally applied. This methodology may be divided in two main parts. In a first stage, with respect to the assumed disclosure scenario, the risk of disclosure of each unit is assessed/estimated. Then, a masking method is applied in order to guarantee that no confidential information about respondents could be retrieved from the disseminated microdata file. This report addresses only the first problem: the assessment of the risk of disclosure. Moreover, the risk of disclosure is here defined as the risk of re-identification.

After the removal of direct identifiers, e.g. name and address, other indirect identifiers, called key variables, could still allow the re-identification of a unit. Usually, most of the key variables registered in social microdata files are categorical. Particular values taken by variables like place of residence, gender, age, citizenship, and marital status could correspond to a unique person in the population. Therefore, the risk of re-identification for such data is estimated by means of rareness concepts, see for example [7], [4].

This report is divided in three parts. In section 2 the framework used for the estimation of the re-identification risk and its link to the log-linear models is introduced. Some issues related to the estimation of such risk in presence of complex surveys are addressed in Section 3. Finally, a particular improvement of the risk estimation, improvement that might be obtained using some smoothing strategies, is introduced in section XXXX.

2 General framework for the estimation of the risk of disclosure

The microdata files are one of the most important products generally released by the National statistical institutes in order to satisfy the information needs of the users. The dissemination of such files should be performed in full compliance with the regulations pertaining to the privacy of respondents.

2.1 Measures of the Risk of Disclosure

Before microdata file dissemination, the NSI generally make assumptions on the tools an intruder might use in order to breach the confidentiality of respondents. Generally it is assumed that the intruder has access to some external database containing direct identifiers. It is further assumed that the intruder would use the shared variables as comparison variables in a matching experiment. There are many implicit assumptions in this disclosure scenario. Many facets of this scenario were previously discussed in literature, see, for example, [10], [14]. One of the most common approaches adopted by the NSIs is to quantify the risk of disclosure by means of the re-identification risk, that is, the probability of a correct match, see [14]. In this report it will be further assumed that the key variables are all categorical, as it usually happens in the social surveys. Examples of categorical key variables generally registered in the social surveys are: gender, age, place of residence, place of birth, education, occupation, marital status, etc. From the point of view of the external, publicly available, registers, the existence and the registration detail of such variables mostly depend on the particular national setting. Generally, for many of these variables, there are standard international classifications, e.g. occupation (ISCO), educaiton (ISCED), geographical location (NUTS). More details on the European classifications may be found on the Eurostat web-site $http://ec.europa.eu/eurostat/ramon$.

As the units sharing the same values for all the categorical variables have the same re-identification risk, see [7], [10], the key variables are cross-classified forming a contingency table with K cells. Obviously, the re-identification risk depends on both the population and sample frequencies in these cells. Let F_k denote the population frequency and let f_k denote the sample frequency of the k-th cell, $k = 1, \ldots, K$. The usage of only sample frequencies is not sufficient because the risk could be overestimated. The global measure of risk discussed in this report is the number of sample uniques that are also population uniques. Following the approach presented in [14], this measure of risk may be written as:

$$\tau_1 = \sum_{k=1}^{K} \mathbb{I}(F_k = 1, f_k = 1)$$

It should be highlighted that other risk measures could be discussed, for example the expected number of correct re-matches. All the problems/issues discussed for τ_1 in this report, could be extended to other risk measures. This happens because this report is dedicated to the improvement of the underlying log-linear models and not exactly to the improvement of the model that defines the risk of disclosure.

2.2 Estimation of τ_1

τ_1 cannot be directly computed because it depends on the unknown population frequencies F_k. Some modelling assumptions are needed in order to derive an estimable expression of the global risk measure. It is assumed that the population frequencies are independently Poisson distributed with means λ_k, denoted by $Po(\lambda_k)$. In each cell, a Bernoulli sampling scheme is assumed, with selection probability equal to π_k. In this framework, it was proved that, the sample frequencies f_k are also independent following Poisson distributions.

Then an estimation of the global risk measure τ_1 may be derived as in [14]. The estimate is given in the equation (1).

$$\hat{\tau_1} = \sum_{k=1}^{K} \exp(-\mu_k(1-\pi_k)/\pi_k), \quad \mu_k = \pi_k \lambda_k \tag{1}$$

The estimate of τ_1 depends on both the sampling fractions, π_k and the expected cells frequencies. It should be observed that the summation should be done on the sample uniques. Moreover, for simplicity, it is generally assumed that the sampling fractions are equal across the cells of the contingency table, $\pi_k = \pi, k = 1, \ldots, K$.

This formula by itself is still not sufficient for a practical risk computation because it depends on the unknown $\mu_k, k = 1, \ldots, K$ parameters. One might assume that the parameters $\mu_k, k = 1, \ldots, K$ are independent and identically distributed, but, in absence of a reference value, it would be impossible to decide which frequencies are high and which frequencies are low. Generally the relationships between the expected cell frequencies are modelled by means of a log-linear model including the desired main effects and interactions, see equation (2).

$$\log(\mu_k) = \boldsymbol{x}_k^{'}\boldsymbol{\beta} \tag{2}$$

The estimates are then computed by maximizing the relevant part of the log-likelihood function $\mathcal{L}(\boldsymbol{\beta}) = \sum(f_k \log(\mu_k) - \mu_k)$. Iterative algorithms like iterative proportional fitting (IPF) or Newton-Raphson may be used for solving this optimisation problem (maximizing the likelihood).

3 On Using the Sample Weights in the Log-linear Models

The dependence of τ_1 on the sampling fraction already indicates that the survey design has an important role even in the statistical disclosure control framework. The surveys conducted by the National Statistical Institutes generally have a very complex structure. These surveys involve, for example, stratification and post-stratification, multistage sampling and clustering of units. A direct way to account for the influence of these factors is to associate to each unit a weight. These weights are generally calibrated in order to maintain some known population totals. The weights are used for the estimation of the desired population characteristics, but also for the variance estimation. Generally speaking, the weights play an important role in survey data analysis.

Besides the information on the survey strategy, the weights are probably the most important numerical (quantifiable) variable that accounts for the survey design. The question is: how could we introduce these weights in the estimation of τ_1, the global measure of risk of disclosure?

3.1 Ignore the Sampling Weights

The easiest solution is, of course, to ignore the weights. This is a good solution when the weights depend **only** on the independent variables $\boldsymbol{x}_k^{'}$ in the log-linear model (2). On the contrary, when the weights depend on other variables, the estimation would be biased and the standard errors computation would be incorrect. If the stratification is the only problem, it is generally recommended to include the stratification variables in the model, as explanatory variables. In the statistical disclosure control framework, this would mean that the stratification variables should be all considered as key variables. First, it should be observed that this is not a very natural requirement because the disclosure scenario is not defined by means of the stratification variables. Second, this strategy would surely imply an overprotection because the associated contingency table would not defined at the appropriate hierarchical level.

3.2 Weighted Frequencies

The second solution is to analyse the table derived from the weighted frequencies, see [11]. First the weighted frequencies f_k^w are calculated, $f_k^w = \sum_{i \in J_k} w_i$, where J_k denotes the k-th cell and w_i is the sampling weight of the i-th unit in the sample. Then the contingency table is analyzed as if it were an unweighted table. This is equivalent to the maximisation of the pseudo-likelihood function. This time the parameter estimates would be unbiased but the standard errors computation would still be incorrect. The same statement is true for other measures of goodness of fit and the statistical tests might be misleading. This happens because the assumption of independence would be violated when comparing the weighted frequencies with the expected frequencies.

3.3 Log-rate Models

The third solution is a mixture of the previous two ones. That is, the parameter estimates should be based on the weighted frequencies while the standard errors estimates should be based on the unweighted frequencies, see [2]. This could be achieved by means of an offset variable depending on weights. The derived model is the log-rate model first introduced by Haberman and it is given in equation .

$$\log(\mu_k) = \log(z_k) + \boldsymbol{x}_k^{'}\boldsymbol{\beta} \tag{3}$$

where $z^k = 1/w^k$ is the inverse of the average cell weight $w^k = f_k^w/f_k$.

Using this model, the estimates of the parameters would depend on the weighted frequencies, while the standard errors will depend on the unweighted frequencies. Both parameter estimation and goodness-of-fit tests will be correct. Consequently, the log-rate model should be preferred to the other two choices when the contingency tables derive from complex survey data.

An interesting feature of this model is the natural way to deal with the structural zeros. The model may be rewritten in a multiplicative form

$$\mu_k = z_k \mathrm{exp}(\boldsymbol{x}_k^{'}\boldsymbol{\beta})$$

and the z_k may be set equal to zero for all the structural zero cells. This formulation is important especially from a practical point of view. It is known that the large contingency tables derived from the social surveys conducted by the National Statistical Institutes generally contain a large number of structural zeros cells.

3.4 Experiments

In order to assess the properties of this later model, a simulation experiment was performed. From the Italian 2001 census data, the variables *Province*, *Gender*, *Age* (14 categories), *Marital Status* (6 categories) and *Education* (6 categories) were selected. Variables *Gender* and *Age* were used as stratification variables. For each province, a stratified simple random sampling scheme was used. The weights were computed in order to preserve the population totals, by *Gender* and *Age*.

A second experiment was conducted using the Labour Force Survey 2001 data. For this survey a two-stage stratified sampling was used and the applied stratification technique involved also the dimension of municipalities, a variable that could be hardly seen as a key variable in practical disclosure scenarios.

In table 1 there are the results obtained using the 4 selected variables as key variables, so included in the log-linear model. As it may be observed, ignoring the weights doesn't seem to produce good results. The same effect was observed using the pseudo-likelihood approach. The log-rate model produces a little bit better results. For the smallest sampling fractions, an overestimation is observed while for the larger sampling fractions an underestimation may be seen. Of course, more testing should be performed in order to assess the properties of these estimators.

As it can be observed in table 1, when one of the stratifying variables is not a key variable, the situation worsens, with all the models. But again the log-rate models seem to give more reliable results. As it was previously stated, each time it is possible to include the stratification variables in the set of key variable (in the disclosure scenario), it is recommended to do it. Anyway, the forced inclusion of the stratification variables in the disclosure scenario makes the contingency

Table 1. τ_1 estimation when *Age* is a key variable. SaUn = number of sample uniques, NoW = unweighted model, LR = log-rate model, I = independence model, S = saturated model.

Province	π	SaUn	True	NoW(I)	LR (I)	NoW(S)	LR(S)
Asti	0.015	246	6	0.22	17.31	0.70	23.04
Asti	0.150	316	41	0.58	12.42	4.02	30.43
Asti	0.277	307	73	0.73	7.84	5.30	32.46
Asti LFS	0.005	86	4	0.00	12.71	0.00	18.38
Biella	0.017	215	3	0.18	9.45	1.09	19.11
Biella	0.167	291	38	0.80	10.73	5.22	29.79
Biella	0.308	279	73	1.20	7.33	6.64	24.95
Biella LFS	0.005	127	7	0.00	13.02	0.03	21.27
Cuneo	0.006	201	2	0.00	7.87	0.06	12.23
Cuneo	0.056	288	14	0.00	11.07	0.09	18.93
Cuneo	0.105	270	23	0.01	7.08	0.38	19.17
Cuneo LFS	0.003	108	10	0.00	17.47	0.00	23.66
Ferrara	0.009	204	2	0.01	8.38	0.21	15.27
Ferrara	0.090	259	23	0.02	8.65	1.98	23.66
Ferrara	0.166	259	34	0.04	6.08	2.13	21.35
Ferrara LFS	0.003	122	4	0.00	22.24	0.36	26.08
Frosinone	0.006	217	1	0.00	9.97	0.04	15.15
Frosinone	0.064	301	23	0.01	14.13	0.50	22.79
Frosinone	0.119	275	32	0.01	4.31	1.09	19.54
Frosinone LFS	0.003	82	6	0.00	12.35	0.11	15.35
Latina	0.006	241	4	0.00	6.43	0.02	10.51
Latina	0.064	278	11	0.00	11.18	0.26	20.99
Latina	0.118	307	34	0.00	6.29	1.73	19.24
Latina LFS	0.003	103	7	0.00	12.69	0.04	18.38
Novara	0.009	214	1	0.00	10.71	0.09	15.81
Novara	0.091	308	30	0.02	4.85	0.87	24.24
Novara	0.168	290	44	0.02	4.57	1.10	22.67
Novara LFS	0.003	112	8	0.00	14.20	0.35	20.95
Parma	0.008	222	3	0.00	11.09	0.10	18.13
Parma	0.079	273	19	0.00	5.38	1.81	18.22
Parma	0.146	270	34	0.01	5.00	2.20	17.01
Parma LFS	0.003	113	6	0.00	17.51	0.04	21.51
Ravenna	0.009	237	4	0.01	13.84	0.30	20.48
Ravenna	0.089	292	17	0.01	9.14	1.44	24.03
Ravenna	0.165	306	43	0.04	4.13	2.02	19.56
Ravenna LFS	0.003	116	7	0.00	18.36	0.02	28.12
Rieti	0.021	195	7	0.35	10.64	1.07	20.74
Rieti	0.211	288	71	1.47	12.88	6.54	38.65
Rieti	0.391	301	116	2.98	10.19	16.32	45.92
Rieti LFS	0.007	82	2	0.02	7.46	0.51	13.23
Rimini	0.011	225	1	0.02	12.58	0.45	16.91
Rimini	0.114	299	30	0.05	8.78	2.09	33.62
Rimini	0.212	283	58	0.12	5.34	4.20	29.14
Rimini LFS	0.004	86	3	0.00	11.66	0.00	16.38
Verbano	0.020	193	1	0.29	8.08	0.72	16.96
Verbano	0.195	289	50	0.74	16.24	6.11	44.19
Verbano	0.362	296	113	2.21	8.57	9.11	34.65
Verbano LFS	0.006	107	8	0.00	22.42	0.41	26.60
Vercelli	0.018	225	9	0.28	8.58	1.43	14.80
Vercelli	0.176	279	60	0.89	10.13	6.34	30.93
Vercelli	0.327	283	80	1.04	7.19	7.10	28.52
Vercelli LFS	0.005	111	3	0.00	18.95	0.15	23.14
Viterbo	0.006	203	4	0.02	10.51	0.05	11.76
Viterbo	0.064	305	11	0.26	20.99	0.14	8.21
Viterbo	0.118	315	34	1.73	19.24	0.21	6.34
Viterbo LFS	0.003	91	7	0.04	18.38	0.00	14.41

Table 2. τ_1 estimation when *Age* is not a key variable. SaUn = number of sample uniques, NoW = unweighted model, LR = log-rate model, I = independence model, S = saturated model.

Province	π	SaUn	True	NoW(I)	LR (I)	NoW(S)	LR(S)
Asti	0.015	15	0	0.00	0.10	0.00	0.20
Asti	0.150	24	1	0.00	0.00	0.00	0.09
Asti	0.277	11	0	0.00	0.00	0.00	0.00
Asti LFS	0.005	13	1	0.00	3.73	0.00	3.08
Biella	0.017	27	0	0.00	3.01	0.00	2.24
Biella	0.166	14	1	0.00	0.00	0.00	0.01
Biella	0.309	17	3	0.00	0.00	0.00	0.00
Biella LFS	0.005	9	1	0.00	1.96	0.00	2.38
Ferrara	0.009	29	0	0.00	1.89	0.00	1.45
Ferrara	0.089	16	0	0.00	0.00	0.00	0.00
Ferrara	0.166	13	1	0.00	0.00	0.00	0.00
Ferrara LFS	0.003	13	0	0.00	5.68	0.00	5.41
Latina	0.006	28	0	0.00	1.62	0.00	2.24
Latina	0.064	25	0	0.00	0.00	0.00	0.36
Latina	0.118	12	0	0.00	0.00	0.00	0.00
Latina LFS	0.003	15	0	0.00	2.95	0.00	2.94
Vercelli	0.018	29	0	0.00	3.04	0.00	2.64
Vercelli	0.176	18	0	0.00	0.00	0.00	0.01
Vercelli	0.327	17	1	0.00	0.00	0.00	0.00
Vercelli LFS	0.005	12	2	0.00	2.55	0.00	2.31

table more sparse. And the tables we have to work with in practical situations are already sparse enough to create us serious problems.

For example, the analysis of sparse tables is related to the possible non-existence of the maximum likelihood estimates. More precisely, sometimes the parameter estimates take on values plus or minus infinity. In such cases, the iterative proportional fitting or the Newton-Raphson methods may even fail to converge. Several issues

4 On using smoothing solutions

If the table to be analysed is sparse, the first possible solution is the table redesign. Some variables could be recoded or even omitted, but this doesn't solve the specific dissemination problem. That is, in the statistical disclosure control framework, recoding should be applied as a protection method, not because the table is sparse.

A second possible solution would be the usage of a flattening constant. To each cell or only to the empty cells, a constant is added. Various values were proposed in literature, e.g. $1, 0.5, \sqrt{n}/K$, etc. All these flattening constants have the same main drawback: the sample size is artificially increased and the introduced total count might dominate the estimates.

The third solution is the usage of parsimonious models. Without such parsimonious representation for large sparse tables, the likelihood could get maximized on the boundary of the parameter space and too many cells estimates might be zero. Anyway, in the risk estimation framework, see [12], it was observed that when a simple log-linear model is used, the estimation of μ would be based in information from all the cells having in common even a single characteristic.

4.1 Local neighbourhoods

In the statistical disclosure control framework it was proposed to find a compromise between the model and the quantity of information used: complicate a little bit the model, but use only the information from the neighbouring cells, see [12]. Of course, the neighbourhoods may be defined only for ordinal variables. Consequently, it is supposed that a distance between cells may be defined, namely $d(k^{'}, k)$.

This approach is base on the assumption that in a certain neighbourhood, the $\log(\boldsymbol{\mu})$ may be approximated by a polynomial, i.e. $\log(\boldsymbol{\mu}) = \left[\beta_0 + \ldots + \beta_t d(k^{'}, k)^t\right]$. Then, it was proposed to maximize, for each cell, the local likelihood function:

$$\mathcal{LL}(\boldsymbol{\beta}) = \sum_{k^{'} \in N^k} \left[f_{k'} \left[\beta_0 + \ldots + \beta_t d(k^{'}, k)^t\right] - \exp\left(\beta_0 + \ldots + \beta_t d(k^{'}, k)^t\right)\right] \quad (4)$$

where N^k denotes the considered neighborhood of the k-th cell.

In [12] several choices of N^k and d are presented, taking into account their possible multi-dimensionality.

It should be noted that the total number of parameters to be estimated is proportional to the degree of the polynomial and to the number of cells. In some situations, this increased number of parameters might generate some problems in data analysis.

With respect to the standard local polynomial regression framework, see [5], the kernel function used in equation (4),

$$W(u) = \begin{cases} 1,\, u \in (-1,1) \\ 0,\, otherwise \end{cases}$$

is not a continuous function. This might cause problems to the asymptotic properties of the estimators. Indeed, W is not a proper smoothing function; it ensures only a truncated local polynomial regression. Moreover, if the kernel function W is used, all the cells in N^k would contribute in equal manner to a cell parameter estimation.

4.2 Measuring the smoothness

Even if it is not obvious from the above formulas, the main idea in the previous proposal is that the sample uniques with small values neighbouring cells are more likely population uniques. This idea could be further generalized. If smoothness is assumed, the neighbouring cells should have similar values. Note that the local polynomials of the previous approach provide a kind of smoothness, too.

Now the problem is how to maximize the likelihood and to obtain this smoothness at the same time. Or, how could we quantify the smoothness and introduce it into the modelling procedure.

Let's see what smoothness means in practice. Let us take for example a simple 2 by 2 table.

Table 3. Example of a 2 x 2 table

a	b
c	x

If a, b and c have similar values, small or large, it doesn't matter, and if smoothness is assumed, the fourth value x should take more or less the same value. This means that the cross-ratio $\theta = \frac{ax}{bc}$ is approximately 1. We know from the analysis of contingency tables that the values of θ close to 1 represent the independence of the two categorical variables, while values of θ farther from 1, represent stronger levels of association, that is, no independence.

In the statistical disclosure control framework, the cross-ratio θ is a number that may be associated to the independence, hence to smoothness. In other words, this independence criteria could be a possible way to quantify the smoothness. All we have to do is to introduce θ in the modelling procedure.

Reminding that we should maximise the likelihood and find a smooth solution at the same time. This could be done by pushing the likelihood maximisation towards a smooth solution. In the maximisation problem, we could penalise for missed smoothness or, equivalently, we could penalise for the missed independence. Maximizing the penalized likelihood given in equation (5) would guarantee a smooth solution.

$$\mathcal{PL}(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) - A \sum_{i=1}^{I-1} \sum_{j=1}^{J-1} \left[\log \left(\frac{\mu_{i,j}\mu_{i+1,j+1}}{\mu_{i,j+1}\mu_{i+1,j}} \right) \right]^2 \tag{5}$$

where I is the number of rows and J is the number of columns. A is a penalty constant whose values might be chosen according to some statistical criteria discussed in [13], for example.

The function $\mathcal{PL}$ penalizes for missed local independence in the reduced 2x2 tables since $\mathcal{PL}$ takes smaller values when the cross-ratios are much greater (smaller) than 1.

4.3 Properties of the Penalized Likelihood Function

There are several theoretical advantages of this penalized likelihood approach.

First the existence, uniqueness and consistency of the estimators were proved under general conditions, see [13].

Second, the number of parameters to be estimated is greatly reduced. This means that the number of degrees of freedom is much more controlled with respect to the local neighboring approach introduced in [12].

Third the penalized likelihood could be extended to multidimensional tables. It is sufficient to find the expression of independence in such multidimensional tables. And this expression if generally known from the contingency tables theory. Just as an example, a a 3-dimensional table, the penalty should be related to the quantity $\log\left(\frac{\mu_{ijm}\mu_{i+1jm}\mu_{ij+1m}\mu_{ijm+1}}{\mu_{i+1j+1m+1}\mu_{i+1j+1m}\mu_{i+1jm+1}\mu_{ij+1m+1}}\right)$. The computational properties of this extension should be anyway tested in real cases with tables of thousands of cells.

The penalized likelihood could also be integrated with the log-rate models. The penalized likelihood approach is only an estimation method, so, in principle, it could be integrated with whatever model. The advantage of using the penalised likelihood is given only by its smoothing properties.

The last, but surely not the least, the penalised likelihood approach could be extended to non-ordinal key variables. The large (full) table could be divided in many 2-way reduced tables. For these reduced tables, the independence could be simply expressed by the cross-ratio. Then the penalty should be expressed in terms of independence in the reduced tables. It should be observed that it is not necessary to assume the smoothness, hence independence, for all the reduced 2-way tables. Only for a subset of such reduced tables the smoothness property might be assumed. For ordinal variables, the properties of the penalized estimators were studied, [13]. The same should be done for non-ordinal variables, too.

4.4 Numerical example

Just to illustrate you what kind of results might be obtained using this approach, a small numerical example was generated. The generated table is given in table 4. The cells (5, 2) and (7, 7) have both a value equal to 1. The neighbors of the cell (5, 2) have large values; instead, the neighbors of the cell (7, 7) have small values.

Table 4. Simulated contingency table.

5	4	3	3	5	1	2	5
1	2	3	3	6	4	2	5
5	4	8	4	4	4	11	4
15	8	8	6	5	6	4	3
10	1	11	2	4	4	3	9
8	7	9	3	2	1	2	1
8	2	4	5	7	2	1	1
6	4	3	7	1	1	2	1

The results obtained using the classical and the penalized likelihood approaches are shown in tables 5 and 6 respectively. In both cases an independence model was fitted. The parameter A in equation (5) was determined by means of the expectation-maximisation algorithm described in [13]. The results obtained using the penalised likelihood approach seem more adequate to the statistical disclosure control smoothing problems.

Table 5. Results of the maximum likelihood estimation approach.

6.5	5.9	5.3	4.8	4.3	3.9	3.5	3.2
6.4	5.8	5.2	4.7	4.3	3.9	3.5	3.1
6.4	5.7	5.2	4.7	4.2	3.8	3.4	3.1
6.3	5.6	5.1	4.6	4.1	3.7	3.4	3.0
6.2	5.6	5.0	4.5	4.1	3.7	3.3	3.0
6.1	5.5	4.9	4.5	4.0	3.6	3.3	3.0
6.0	5.4	4.9	4.4	4.0	3.6	3.2	2.9
5.9	5.3	4.8	4.3	3.9	3.5	3.2	2.9

When smoothness is assumed, tables 5 and 6 indicate that the maximisation of a penalized likelihood function might be a valid methodology for the analysis of contingency tables.

5 Conclusion

Two problems related to the estimation of a global risk measure were addressed. First the analysis of contingency tables derived from complex surveys was dis-

Table 6. Results of the penalized maximum likelihood estimation approach.

5.0	4.0	3.1	3.2	5.0	2.1	1.6	5.0
1.8	1.8	3.1	3.2	6.0	3.9	2.9	4.7
4.9	4.6	8.0	4.4	4.4	4.1	11.1	4.0
15.1	8.2	8.0	6.0	5.0	6.0	4.0	3.4
10.1	3.2	5.6	2.2	4.5	4.0	3.2	8.2
8.4	5.8	10.1	3.2	2.3	1.7	1.7	1.3
8.1	2.2	4.4	5.0	6.8	1.7	1.7	1.2
6.0	4.1	3.2	7.0	1.2	1.3	2.2	1.6

cussed. A log-rate model using the weights as an offset variable was presented. Unbiasness and validity of goodness of fit tests are two significant characteristics of these models. Promising results are obtained when real data were fitted using this modeling procedure. The performed tests show that the introduction of the survey design information into the risk modelling procedure might be a valid approach for improving the estimation of the risk of disclosure.

In the statistical disclosure control framework, table smoothness is particularly important since the estimation of any risk of re-identification measure might be performed by borrowing information from the neighboring cells. Moreover, the relationships between the cross-classifying variables might determine the number of sample uniques that are also population uniques. A penalized likelihood approach was proposed to deal with the smoothness characteristic of the tables. The penalty function was expressed in terms of independence constraints. Several possible extensions of this approach were illustrated. The theoretical flexibility is proved by the fact that the penalised likelihood approach might be extended to multidimensional tables and to non-ordinal categorical key variables. Anyway, the theoretical properties should be studied followed by a testing/validation using real surveys.

Acknowledgments. Istat is not responsible for any views or results presented. The author was supported by the Network of Excellence in the European Statistical System in the field of Statistical Disclosure Control (ESSnet-SDC). A previous version of this paper was published in the Procedings of the Conference *Privacy in Statistical Databases 2008.*

References

1. Agresti, A.: Categorical Data Analysis. Wiley, New York (1990)
2. Clogg, C.C., Eliason, S.R.: Some Common Problems in Log-Linear Analysis. Sociological Methods and Research **16** (1987) 8–44
3. Elamir, E.A.H.: Analysis of Re-Identification Risk Based on Log-Linear Models. In: Domingo-Ferrer, J. and Torra, V. (eds.) PSD 2004, LNCS, vol. 3050, pp. 273-281, Springer Heidelberg (2004)
4. Elamir, E., Skinner, C.: Record level measures of disclosure risk for survey microdata. Journal of Official Statistics, **22(3)** (2006) 525–539.

5. Fan, J., Gijbels, I.: Local Polynomial Modelling and its Applications. Chapman & Hall, London (1996)
6. Fienberg, S.E., Holland, P.W.: On the Choice of Flattening Constants for Estimating Multinomial Probabilities. Journal of Multivariate Analysis **2** (1972) 127–134
7. Franconi, L. Polettini, S.: Individual risk estimation in μ-ARGUS: a review. In Domingo-Ferrer, J. and Torra, V. (eds.), PSD 2004, LNCS, vol. 3050, pp. 262-272, Springer Heidelberg (2004).
8. Haberman, S. J.: Analysis of Qualitative Data, Vol 2. New Developments. : Academic Press, New York (1979)
9. Lohr, S. L.: Sampling: Design and Analysis. Duxbury Press (1999)
10. Polettini, S.: Some Remarks on the Individual Risk Methodology. Monographs of Official Statistics. Work Session on Statistical Data Confidentiality. European Comission (2003)
11. Rao, J. N. K., Thomas, D. R.: The Analysis of Cross-Classified Categorical Data from Complex Surveys. Sociological Methodology **18** (1988) 213–269
12. Rinott, Y., Shlomo, N.: A Smoothing Model for Sample Disclosure Risk Estimation. IMS Lecture NotesMonograph Series Complex Datasets and Inverse Problems: Tomography, Networks and Beyond Vol. 54 (2007) 161-171
13. Simonoff, J. S.: A Penalty Function Approach to Smoothing Large Sparse Contingency Tables. The Annals of Statistics **11** (1983) 208–218
14. Skinner, C., Holmes, D.: Estimating The Re-Identification Risk per Record in Microdata. J. Official Statistics **14** (1998) 361–372
15. Willenborg, L., De Waal, T.: Elements of Disclosure Control. Lecture Notes in Statistics, vol. 155, Springer, Berlin (2001)