Individual Disclosure Risk Measures Based on Log-Linear Models

Ichim, Daniela and Foschi, Flavio
Istat, Division for Users' needs, Integration and Territory
P.zza Indipendenza, 4
00185 Roma,Italy
E-mail: ichim@istat.it, foschi@istat.it

Abstract

Dissemination of microdata files should be constrained to the confidentiality pledge under which a statistical agency collects survey data. To protect the confidentiality of respondents, statistical agencies perform a two-stage statistical disclosure control procedure. In the first stage, with respect to a disclosure scenario, the risk of disclosure of each unit is estimated.

After the removal of direct identifiers, e.g. name and address, other indirect identifiers, called key variables, could still allow the disclosure of some confidential information about a unit. Usually, most of the key variables registered in social microdata files are categorical. An important problem in statistical disclosure control (SDC) is the estimation of the (number of) sample uniques that are also population uniques, i.e. units at risk of disclosure. In this paper, extensions of the Poisson-log-linear model to estimate a disclosure risk measure in contingency tables are presented. The main contribution is the development of smoothing strategies based on a penalized likelihood approach and on graphical log-linear models decomposition. Results of several tests performed on Italian 2001 census data will be presented.

1 Introduction

To face the increasing demand from users, the National Statistical Institutes (NSI) 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 (SDC) methodology is generally applied. This methodology may be divided in two main parts. In a first stage, with respect to an adopted disclosure scenario, the risk of disclosure of each unit is assessed/estimated. Then, a masking method is applied to guarantee that no confidential information about respondents could be retrieved from the disseminated microdata file. This paper addresses only the first problem: the disclosure risk assessment. 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, Elamir & Skinner (2006) and Franconi & Polettini (2004). In this work it is assumed that the key variables are all categorical.

This paper is divided in three parts. In section 2 the framework used for the re-identification risk estimation and its link to the log-linear models is introduced. In section 3 an estimation methodology based on graphical log-linear models decomposition is described. The advantages of such methodology are twofold. First, it would allow the disclosure risk estimation in presence of a large sparse contingency tables. Second, the decomposition of the likelihood estimates would allow the application of different penalization (smoothing) strategies, as for the approaches described in Ichim (2008) or Rinott & Shlomo (2007). The penalized likelihood decomposition methodology was applied to simulated data from the 2001 Italian census. The simulation setting and several results are illustrated in section 4. Finally, in section 5, some conclusions are drawn and further developments are indicated.

2 Measures of Disclosure Risk

To evaluate the disclosure risk, the NSIs generally make assumptions on the tools an intruder might use to breach the confidentiality of respondents. It is usually assumed that the intruder may access 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. The implicit assumptions of this disclosure scenario were previously discussed in literature, see, for example, Polettini (2003) and Skinner & Holmes (1998). The NSIs commonly quantify the disclosure risk by means of the re-identification risk, that is, the probability of a correct match, see Skinner & Holmes (1998).

As the units sharing the same values for all the categorical variables have the same re-identification risk, the key variables are cross-classified; a contingency table with K cells is then derived. Obviously, the re-identification risk depends on both the population and sample frequencies of these cells. Let F_k denote the population frequency and let f_k denote the sample frequency of the k-th cell, $k = 1, \ldots, K$. In this paper the units at risk of disclosure are the sample uniques that are also population uniques. Following the approach described in Skinner & Holmes (1998), a global risk measure may be written as:

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

τ_1 cannot be directly computed because it depends on the unknown population frequencies F_k. It is generally assumed that the population frequencies are independently Poisson distributed with means λ_k. In each cell, a Bernoulli sampling scheme is assumed, with selection probability equal to π_k. It follows that the sample frequencies f_k are also independent following Poisson distributions, see Skinner & Holmes (1998).

Then an estimation of τ_1 may be expressed as in (1).

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

$\hat{\tau_1}$ depends on both the sampling fractions, π_k and the expected cells frequencies. Moreover, for simplicity, it is commonly assumed that $\pi_k = \pi, k = 1, \ldots, K$.

To estimate τ_1 the relationships between the expected cell frequencies are generally modelled by means of a log-linear model including the desired main effects and interactions, $\log(\mu_k) = \boldsymbol{x}_k^{'}\boldsymbol{\beta}$. The estimates are then computed by maximizing the relevant part of the log-likelihood function $\ell(\boldsymbol{\beta}) = \sum(f_k \log(\mu_k) - \mu_k)$. Iterative algorithms like iterative proportional fitting (IPF) or Newton-Raphson may be used to maximize the likelihood $\ell(\boldsymbol{\beta})$.

3 Smoothing Contingency Tables

For large sparse tables, the likelihood could get maximized on the boundary of the parameter space and too many cells estimates might be zero. In practical applications this means that it is often impossible to obtain the maximum likelihood estimates (MLE). Two possible solutions are the table redesign or the addition of a flattening constant. Both solutions have their drawbacks either because they do not solve the given dissemination problem or because the sample size is artificially increased. For example, in Agresti & Yang (1987) and Fienberg & Holland (1972) more details on these methods are given.

A valid alternative could be the usage of parsimonious models. Anyway, in the risk estimation framework, see Rinott & Shlomo (2007), it was observed that when a simple (independence) log-linear model is used, the estimation of μ_k would be based on information from all the cells having in

common even a single characteristic. In Rinott & Shlomo (2007) and Ichim (2008), two smoothing tables approaches based on local neighborhoods were proposed.

In this paper, an alternative smoothing methodology is illustrated. As neighbourhoods of cells are not involved, this approach should equally apply to any kind of categorical variables, i.e. to both nominal and ordinal variables. The main idea is to exploit the link between graphical models and log-linear models. Obviously, marginal tables have smaller dimensions and, simultaneously, they are less sparse. Consequently, maximum likelihood estimates of log-linear models of higher interactions order could be more easily computed. Then, these MLE estimates could be combined together through the graphical log-linear models decomposition formula. As a by-product of this methodology, also the computational burden should be greatly reduced. The marginal tables to be used in such an estimation procedure are given by a proper decomposition of the corresponding graphical log-linear model.

3.1 Graphical Log-Linear Models

In this section it is discussed how to combine information from ordinary marginals to estimate the original data. The general notations and definitions in Edwards (2000) are adopted here. Only a very brief introduction of the main concepts used in this work is given.

3.1.1 Graphical Models

The concept of conditional independence plays a key role in graphical model theory and it is relevant for factorizing a probability function of several variables as the product of some marginals. An *undirected graph* $G = \{V, E\}$ is a set of vertices (the variables) V and a set of edges (the links between the variables, without any direction) E. Two vertices x and y are *adjacent*, $x \sim y$, if they are linked by means of an edge $[x, y] \in E$. A graph is *complete* if there is an edge between each pair of vertices. A *subgraph* $G_z = \{Z, F\}$ induced by Z is the graph whose edges $F \in E$ have all vertices in $Z \in V$. A subset $U \in V$ is a *clique* if it induces a subgraph maximally complete, that is if $U \subset W \in V$, then W is not complete. A *path* is a sequence of vertices linked by edges. A *path* $x_1, x_2, \ldots, x_n, x_1$ defines an *n-cycle*. If the vertices of an *n-cycle* are distinct and $x_i \sim x_j$ only if $|i - j| \in \{1, n-1\}$ then the *n-cycle* is *chordless*. A graph is *triangulated* if it has no *chordless* cycle of length greater than three. Given three subsets $X, Y, S \in V$, S *separates* X and Y if all paths between X and Y intersect S. A statistical model M for the observed data is a family of probability laws $\{f(\theta) : \theta \in \Theta\}$ where Θ is the of parameters; when all members of M share some conditional independence features, then those features can be represented by an *undirected graphical model*. Denoting by f, g, h the joint probability of random variables (X, Y, S), (Y, S) and (X, S) respectively, X and Y are conditionally independent given S, $X \perp\!\!\!\perp Y \mid S$ if $f(x, y, s) = h(x, s)\, g(y, s)$. Moreover, a conditional independence relation satisfies the following:

1. $X \perp\!\!\!\perp Y \mid S \Leftrightarrow Y \perp\!\!\!\perp X \mid S$
2. $X \perp\!\!\!\perp (Y \cup Z) \mid S \Rightarrow X \perp\!\!\!\perp Y \mid S,\ X \perp\!\!\!\perp Z \mid S$
3. $X \perp\!\!\!\perp (Y \cup Z) \mid S \Rightarrow X \perp\!\!\!\perp Y \mid (Z \cup S)$
4. $X \perp\!\!\!\perp Y \mid S \cap X \perp\!\!\!\perp Z \mid (Y \cup S) \Rightarrow X \perp\!\!\!\perp (Y \cup Z) \mid S$
5. $X \perp\!\!\!\perp Y \mid (Z \cup S) \cap X \perp\!\!\!\perp Z \mid (Y \cup S) \Rightarrow X \perp\!\!\!\perp (Y \cup Z) \mid S$ if $f(x, y, s, z)$ is strictly positive.

The *Markov properties*, i.e. the connections between *separation* in graph theory and *conditional independence* are important to our aims. The *global Markov property* is referred to subsets: if the subset S separates X and Y then, given S, X and Y are conditionally independent. A strictly positive

probability law satisfying the *global Markov property* w.r.t a G graph is said *G-Markov*. Hence, defining C the set of *cliques* in G; a *G-Markov* probability law can be factorized as $f = \prod_{c \in C} h_c(x_c)$. Following Lauritzen (1996), a triple (X, Y, S) of disjoint subsets of V is a (weak) decomposition of a graph G if $V = X \cup Y \cup S$ and

1. S is the *separator* of X and Y,
2. S is *complete*.

If X and Y are both non empty, the decomposition is *proper.* Due to the decomposition, $G_{X \cup S}$ and $G_{Y \cup S}$ are subgraphs of G. A fundamental statement is that a graph is decomposable if it is *complete* or if there exists a *proper decomposition* into *decomposable* subgraphs. It is important to stress that for a *proper* decomposition, subgraphs have fewer vertices than G. It was demonstrated that a graph is decomposable if it is *triangulated*. So, if a graph is decomposable, it can be recursively decomposed into its undecomposable subgraphs. This implies the possibility of greatly simplifying calculations: if the joint probability law f is *G-Markov* and the triple (X, Y, S) is its *proper* decomposition, then $f = \frac{f_{X \cup S} f_{Y \cup S}}{f_S}$. Moreover, if C and S are respectively the set of *cliques* and *separators* of G and f is *G-Markov,* it holds

$$f = \frac{\prod_{c \in C} f_c(x_c)}{\prod_{s \in S} f_s(x_s)} \tag{2}$$

Equation (2), the decomposition formula, shows how a joint probability law may be expressed as a function of particular subsets of its marginals. The conditional independence relations are usually assessed starting from the whole graph and recursively deleting *edges* between *vertices* showing weak associations. The association relevance is evaluated by means of hypotheses tests. In this paper, the approach described in Dahinden *et al.* (2009) is adopted; it is based on the use of *random forests* to determine a *decomposable graph* conditionally to a given maximal *cliques* dimension.

A log-linear model is graphical if the interaction terms are exactly the maximal cliques of the corresponding interaction graph, and it is decomposable if it is graphical and if the interaction graph is decomposable as well.

The similarity between the interaction terms of a decomposable log-linear model and the marginals (of the original contingency table) with the same factors is not superficial. The model can be built using only the marginals whose attributes are specified by the interaction terms, so that the global log linear model can be obtained by combining a set of estimates from *cliques* and *separators*. Consequently, the expected cell values for a graphical decomposable log-linear model are given by (2), see Lauritzen (1996). Finally, it should be noted that, by assuming the log-linear model decomposition, a kind of smoothing is already considered.

3.1.2 Log Linear Models and Penalized Likelihoods

For count data, a Poisson distribution is often assumed, so that $Y_i \sim Poisson(\mu_i)$ and $f(y_i; \mu_i) = exp[y_i log(\mu_i) - \mu_i - log(y_i!)]$.

Given $E(Y_i) = \mu_i$ with $\mu_i = \eta(\theta_i)$, if there is a transformation such that $g(\mu_i) = \boldsymbol{x}_i^t \boldsymbol{\beta}$, where $\boldsymbol{x}_i$ is a vector of explanatory variables related to the i^{th}observation, $\boldsymbol{\beta}$ is a set of parameters and g is a monotone and differentiable function, then a generalized linear model is defined. Considering the log likelihood function, estimates of $\boldsymbol{\beta}$ parameters mqy be computed.

Moreover, $E(Y_i) = \mu_i = exp\left(\boldsymbol{x}_i^t \boldsymbol{\beta}\right)$ and, consequently, $log(\mu_i) = \boldsymbol{x}_i^t \boldsymbol{\beta}$

Contingency data, if there are no constraints on the $Y_i's$, can be interpreted as independent random variables having joint probability distribution $f(\boldsymbol{y}; \boldsymbol{\mu}) \propto \prod e^{-\mu_i} \mu_i^{y_i}$.

Constraints about the total number of observation or more fixed marginals imply the use of multinomial and product multinomial models respectively. However, it is possible to show that MLE are invariant w.r.t those probability distributions providing parameters representing fixed marginals totals. Hence, for estimation purposes, the Poisson model could be always used. Log-linear models are generally used to study associations between variables; hypotheses tests can be performed by comparing goodness of fit statistics between a general model and a restricted one. In SDC framework the focus is on the population uniques identification (units at risk of disclosure). In this work, disclosure risk derived from other small frequencies is ignored. Ideally, fitted values should produce a partition of the sample uniques set into two subsets containing population (true) uniques and false uniques.

In high dimensional data, a constraint on the log likelihood parameters could be imposed:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= argmax\left[\ell\left(\boldsymbol{\beta}\right)\right] \\ & subject \quad to \quad \sum |\boldsymbol{\beta}|^d < s \end{aligned} \tag{3}$$

This penalization approach is generally used to facilitate the computation (existence) of the inverse matrices involved in MLE derivation. It should be stressed that this penalization might be seen as a smoothing technique as well. Indeed, as $\boldsymbol{\beta} \to 0$, the expected cell frequencies tend to a mean value. To solve the constrained optimization problem (3), the Lagrangian function is maximized $\hat{\boldsymbol{\beta}} = argmax\left[\ell\left(\boldsymbol{\beta}\right) - \lambda \sum |\boldsymbol{\beta}|^d\right]$.

In other words, $\hat{\boldsymbol{\beta}}'s$ maximize the penalized log likelihood, being λ the sensitivity of the objective function. $\lambda = 0$ means no penalization while an increasing value of λ shrinks coefficients $\boldsymbol{\beta}$ to zero. Two popular choices for d are $d = 1$ and $d = 2$, leading to the lasso ridge estimators, respectively. The latter was adopted in this work because the L_2 penalization decreases coefficients with different intensities. The λ estimation is a complex task; it may be performed by means of cross validation. In a leave-one-out cross validation framework, for each value in a set of λ candidates, the profile log likelihood of each observation is calculated using the parameter estimates obtained from the rest of the sample. So, the optimal λ belonging to the considered set of candidates is the one which maximizes the joint cross validated profile log likelihood.

4 Experiments

In this section a comparison between the generalized log-linear models and the decompostion-based models was performed is described. Two disclosure scenarios were adopted, both defined by the following key variables:*Province, Gender, Age, Citizenship, Marital status, Education, Type of activity, Occupational condition, Type of contract, Full-time or part-time contract*, and *Position*. In the experiments reported here the household hierarchy was ignored. Moreover, only 5 provinces Trieste (238000 residents), Trento (470000 residents), Torino (2147000 residents), Palermo (1229000 residents) and Potenza (391000 residents) were selected to illustrate the obtained results.

In order to perform the simulations and to assess the properties of the above fitting methods, samples from the Italian 2001 census were drawn by means of a simple random sampling scheme. From the microdata census file, records corresponding to any missing value in any of the key variables were eliminated. Consequently, the missing values problem was not dealt with in this work. Moreover, the structural zeros were not considered, either (the census cells having zero value frequencies were defined as structural zeros). For each province, 500 stratified samples were drawn from the census data. The stratification variables were *Gender* and *Age* (14 categories), simulating a real sampling scheme as in use at Istat. For each province, three different sampling fractions, i.e. $\pi = 0.002$, $\pi = 0.006$ and $\pi = 0.01$, were used. Then, the weights were computed, by means of calibration estimators, see Deville & Sarndal (1992), in order to preserve the population totals in each strata.

Table 1: Key variables and their number of factors.

Variable	Exp I	Exp II
Gender	2	2
Age	4	4
Citizenship	2	2
Marital status	3	3
Education	3	3
Type of activity	5	5
Occupational condition		4
Type of contract		3
Full-time or part-time contract		2
Position		4

Province, Age, Citizenship, Marital status, Education, Type of activity were considered as key variables in a first experiment (Exp I). In order to test the proposed approach in a more complex setting, the number of key variables was set up to 10. The chosen key variables were *Province, Gender, Age, Citizenship, Marital status, Education, Type of activity, Occupational condition, Type of contract, Full-time or part-time contract*, and *Position*. The number of modalities of each categorical variable is shown in table 1. In practice, such aggregated (recoded) variables, might be hardly be considered as defining an optimal microdata release because of their reduced analytical utility. Anyway, in special releases like public use files (without any signed agreement), these recodings might still be useful. This statement holds especially for the second experiment, where the reduced number of modalities is mitigated by the number of key variables. The experiments I and II were replicated for each province. In experiment II, only 300 samples were used, for each province and for each sampling fraction.

The first obvious problem is given by the number of sampling zeros. Indeed, for example in Exp I, even if the key variables were aggregated, the minimum percentage of zero cells in a sample equals 40%, corresponding to the largest province (Torino) and to the largest sampling fraction ($\pi = 0.01$). On the opposite side, the maximum percentage of zero cells, corresponding to Trieste, ($\pi = 0.002$), equals 86 %. These large percentages of zero cells are due to the small, but realistic, sampling fractions used and to the number of key variables. Especially in Exp II, the number of key variables might be considered quite large for practical SDC problems. In this work, a Poisson distribution was always assumed for the cells frequencies. In future works, zero-inflated Poisson models will be taken into account too.

The generalized log-linear models were applied to each sample in the following manner. Firstly, the independence model (*glm.ind*), was used. Secondly, using a stepwise search, see Jenrich & Sampson (1968), the model (*glm.step*) minimizing the AIC criterion was found. The model search space was defined by the independence model (the minimum) and the model derived from the interactions up to an order equal to the number of key variables minus 1 (the maximum). Thus, only the saturated model was eliminated from the model search space. Finally, for the best model identified by the stepwise procedure, a penalized approach was implemented. The penalization was constructed as an L_2-shrinkage aiming at diminishing the overfitting effects. The resulting model was called *glm.pen*. For the cross-validation, the state space of the penalization parameter, λ, was set equal to the interval (0.1, 10). To estimate the parameters, an iterative weighted least squares method was always applied.

Then, 4 versions of decomposition-based log-linear models were considered. In order to apply any decomposition-based model, for each sample, the corresponding graphical log-linear model was

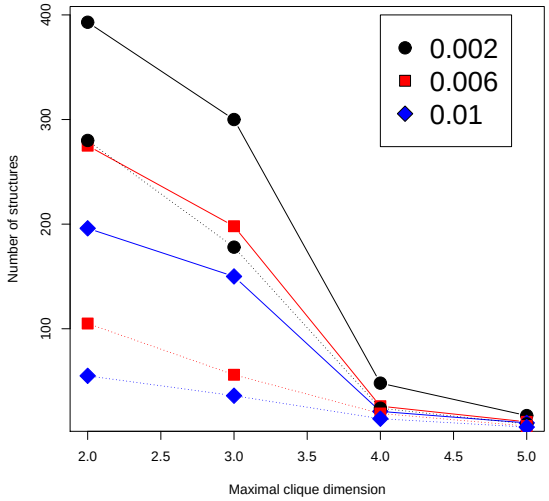

Figure 1: Number of structures for Trieste (solid lines) and Trento (dotted lines).

identified. That is, for each sample, the best decomposition was estimated, conditionally to the allowed maximal clique dimension, i.e. 3, 4 or 5. Given the maximal clique dimension, the best decomposition was defined as the one for which the variables belonging to the same clique are highly associated, while variables belonging to different cliques have a weak association. The search of the marginal tables (structures of the log-linear models) was performed using a random forest approach, as described in Dahinden et. al. 2009. Generally, as exemplified, the number of different decompositions decreases as either the province dimension (number of residents), the maximal clique dimension or sampling fraction increases.

Once the decompositions are found, the first decomposition-based method (*Lau.sat*) was derived from the utilization of the saturated model on each clique and separator. It should be noted this method already generates a weak smoothing induced by the application of the decomposition formula itself. The second decomposition-based method (*Lau.ind*) was derived by assuming the independence log-linear model on each clique and separator. Thirdly, a stepwise model search procedure (*Lau.step*) was used, equivalent to the one applied to the generalized log-linear models. These stepwise procedure allows the increase of model complexity, i.e. the interactions order. The stepwise search was applied independently for each clique and separator. Finally, the fourth decomposition-based method (*Lau.pen*) penalizes the best model identified by the stepwise procedure. Here the smoothing is twofold, due to the graphical model decomposition and to the parameters shrinking. As for the stepwise procedure, this penalization is applied to each clique and separator independently. For the cross-validation, the state space of the penalization parameter, λ, was set equal to the interval (0.1, 10).

In Figure 2, the results obtained in Exp I are shown, for both generalized log-linear models and decomposition-based models. Only the results obtained for province Trento and π=0.006 are illustrated since the other provinces and sampling fractions show similar trends. In figure 2, for each drawn sample, the mean of false uniques squared residuals (x-axis) was plotted against the mean of true uniques squared residuals (y -axis). The false uniques are defined as sample uniques that are not unique cases in the population (census data). The true uniques are unique cases in both sample and population. Where the case, the maximal clique dimensions are indicated in the title of subplots. For each graph, the red line indicates the bisector line. As it can be observed, there is only a weak difference among the three versions of generalized log-linear models. This is due to the fact that, for large sparse contingency tables, only low order interaction log-linear models are really

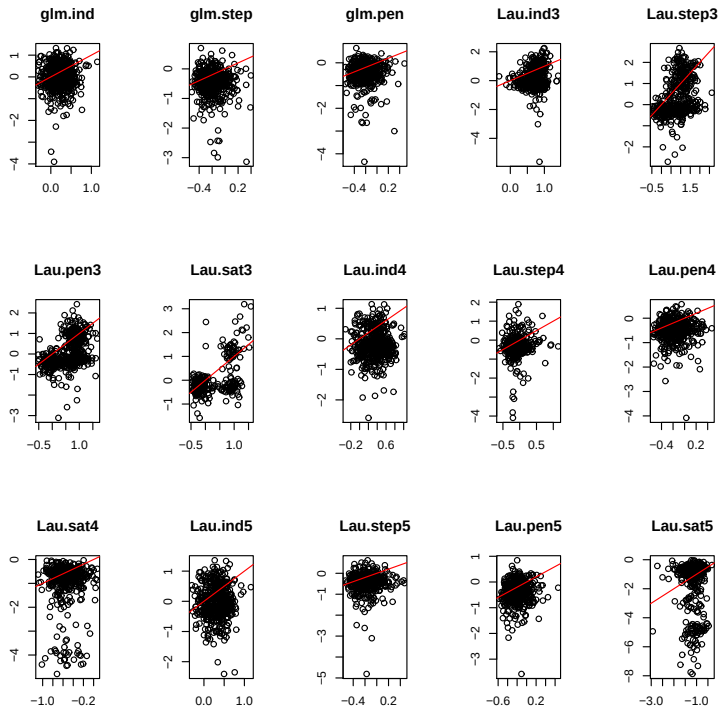

Figure 2: Exp I (6 key variables). Comparison of fitting methods. Trento, $\pi = 0.006$. x-axis - mean of sample uniques squared residuals, y-axis - mean of true uniques squared residuals. Logarithmic scale.

feasible; hence the impact of the stepwise search or penalization procedures is negligible. In figure 2, it may be also observed that as the maximal clique dimension increases, the results obtained using the decomposition-based methods approach those obtained the generalized log-linear models. This is due to the fact that as the maximal clique dimension increases, the marginal tables tend to the complete table, thus reducing the decomposition effect. Finally, it may be noticed that the decomposition-based method with maximal clique dimension equal to 3 discriminates best between the two types of uniques, i.e. true and false uniques. Indeed, the mean of squared residuals corresponding to true uniques is almost always lower than the mean of squared residuals corresponding to false uniques. It follows that an effective discrimination criteria or threshold might be set. The identification of such criteria will be subject of future work. It should be also noted that the penalized decomposition-based method discriminates better than the other decomposition-based methods. In Figure 3, only for the penalized decomposition method, a comparison between three provinces and three sampling fractions is illustrated. As expected, as the sampling fraction or province dimension (number of residents) increase, the true uniques might be better identified. These aspects are obviously related to the (marginal) contingency table sparsity.

In Figure 4, some results obtained in Exp II are shown. The meaning of x and y axes is the same as in Figure 2. Again, the mean of squared residuals corresponding to true uniques is almost always lower than the mean of squared residuals corresponding to false uniques; hence a discrimination between true and false uniques should be possible. In this experiment, the improvement of the penalized decomposition-based method over the other two methods is even more obvious. Moreover, it should be emphasized that the maximal clique dimension is still equal to 3, supporting the potential of the decomposition-based smoothing methods. The main advantage is given by the usage of log-linear models in disclosure scenarios defined by a medium-large number of key variables should be possible using the above described properties of the decomposable log-linear models.

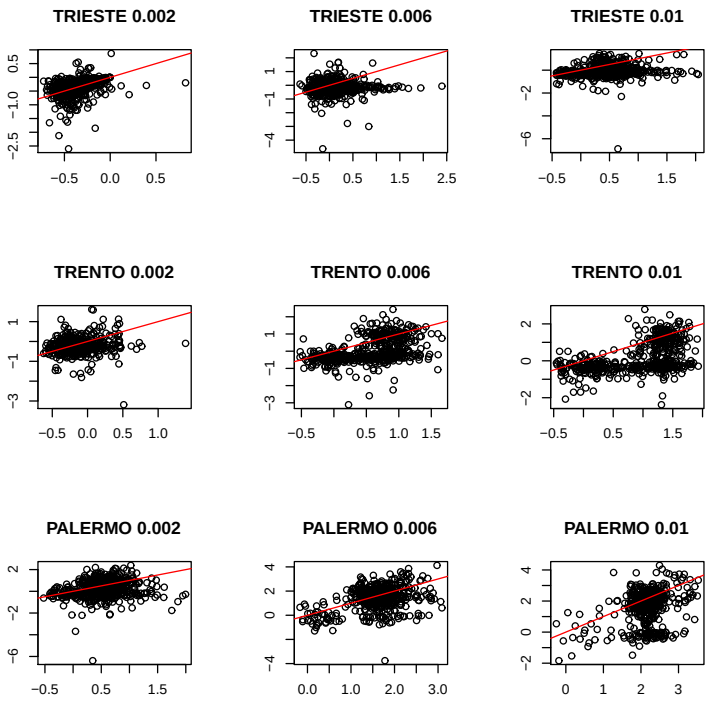

Figure 3: Exp I (6 key variables). Comparison of the penalized decomposition-based fitting methods for different provinces and sampling fractions, maximal clique dimension = 3. x-axis - mean of sample uniques squared residuals, y-axis - mean of true uniques squared residuals; logarithmic scale.

5 Conclusions

In this paper an approach to the estimation of the individual disclosure risk was presented. The main feature of the presented strategy is that it allows disclosure risk estimation through log-linear models even in presence of large sparse contingency tables. The idea is to estimate the best log-linear model decomposition, to fit log-linear models on marginal tables, i.e. clique and separators, and, finally, to put the results together by means of the decomposition formula. Another advantage of the proposed methodology is given by the possibility to use more complex log-linear models on the marginal tables; this is due to the fact that marginal tables are smaller than the full contingency table and, at the same time, they are also less sparse. Additionally, with respect to generalized log-linear models, the computational burden is greatly reduced.

Some preliminary results were illustrated. Real data stemming from the Italian 2001 census was used. In the simulations performed, realistic sampling fractions and sampling schemes were applied. The improvement of the penalized decomposition method over the other methods was discussed. The penalization is a smoothing method and it was shown that, by preventing overfitting problems, it generates an improvement itself. From the empirical results obtained so far, it may be deduced that, the decomposition-based methods may be successfully applied even in presence of a large number of key categorical variables. Finally, it should be stressed that the proposed methodology does not make any assumption on the type of categorical key variables, i.e. it works independently on ordinal and nominal categorical key variables.

More testing and simulations will be performed in order to assess the properties of the decomposition-based methods in the SDC framework. Some variations will be implemented and tested as well. For example, one might use the same decomposition, possibly derived directly from the entire census data, for all the samples. Alternatively, model averaging strategies could be used.

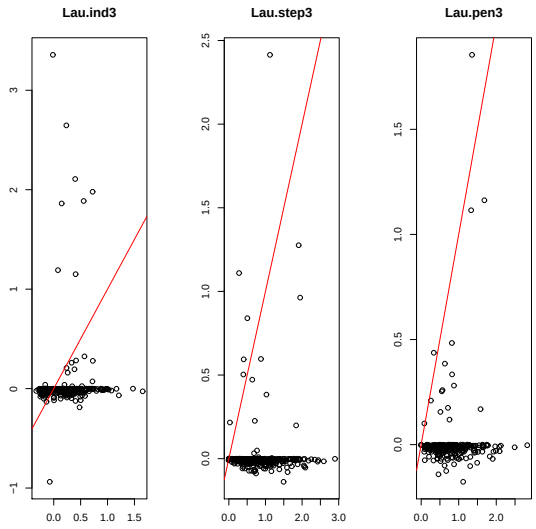

Figure 4: Exp II (10 key variables). Comparison of decomposition-based fitting methods. Trento, $\pi = 0.006$, maximal clique dimension = 3. x-axis - mean of sample uniques squared residuals, y-axis - mean of true uniques squared residuals; logarithmic scale.

The decomposition-based strategy should be further improved. First, an effective discrimination criteria between true and false uniques should be derived. Second, some reflections should be made on the possibility to take into account calibration weights and other complex surveys features.

REFERENCES (RÉFERENCES)

Agresti, A., Yang, M.: An Empirical Investigation of Some Effects of Sparness in Contingency Tables. Computational Statistics and Data Analysis **5** (1987) 9–21.

Dahinden, C., Kalisch, M., Bühlmann, P.: Decomposition and Model Selection for Large Contingency Tables. Biometrical Journal (2009).

Deville, J.C., Särndal, C.B., Calibration Estimators in Survey Sampling, Journal of the American Statistical Association, **87** (1992), 376–382.

Edwards, D.: Introduction to Graphical Modelling. Springer, 2000.

Elamir, E., Skinner, C.: Record Level Measures of Disclosure Risk for Survey Microdata. Journal of Official Statistics, **22(3)** (2006) 525–539.

Fienberg, S.E., Holland, P.W.: On the Choice of Flattening Constants for Estimating Multinomial Probabilities. Journal of Multivariate Analysis **2** (1972) 127–134.

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, Springer Heidelberg (2004), 262–272.

Jennrich, R. I., Sampson, P. F.: Application of Stepwise Regression to Non-Linear Estimation. Technometrics, 10(1)(1968) 63–72.

Ichim, D.: Extensions of the Re-identification Risk Measures Based on Log-linear Models. In Domingo-Ferrer, J. and Saygm, Y. (eds.), PSD 2008, LNCS, vol. 5262, Springer-Verlag Berlin Heidelberg, 203-212.

Lauritzen, S. L.: Graphical Models. Oxford Science Publications, 1996.

Polettini, S.: Some Remarks on the Individual Risk Methodology. Monographs of Official Statistics. Work Session on Statistical Data Confidentiality. European Commission (2003).

Rinott, Y., Shlomo, N.: A Smoothing Model for Sample Disclosure Risk Estimation. IMS Lecture Notes-Monograph Series Complex Datasets and Inverse Problems: Tomography, Networks and Beyond. **54** (2007) 161-171.

Skinner, C., Holmes, D.: Estimating The Re-Identification Risk per Record in Microdata. Journal of Official Statistics **14** (1998) 361–372.