

WP. 43
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS** **EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)**

Joint UNECE/Eurostat work session on statistical data confidentiality
(Bilbao, Spain, 2-4 December 2009)

Topic (vii): Risk/benefit analysis and new directions for statistical disclosure limitation

ASSESSING DISCLOSURE RISK UNDER MISCLASSIFICATION FOR MICRODATA

Invited Paper

Prepared by Natalie Shlomo (University of Southampton), United Kingdom

Assessing Disclosure Risk Under Misclassification for Microdata

Natalie Shlomo*

* Southampton Statistical Sciences Research Institute, University of Southampton, Highfield, Southampton, SO17 1BJ, United Kingdom, N.Shlomo @soton.ac.uk

Abstract: Disclosure limitation methods for protecting the confidentiality of respondents in survey microdata often use perturbative techniques which introduce measurement error into the categorical identifying variables. In addition, the data itself will often have measurement errors commonly arising from survey processes. There is a need for valid and practical ways to assess the risk of identification for survey microdata with measurement errors. The risk assessment is based on a common disclosure risk scenario where an intruder seeks to match the microdata with a publicly available external file. In this paper, we examine probabilistic record linkage as a means of assessing disclosure risk and relate it to disclosure risk measures under the probabilistic framework of the Poisson log-linear models.

1 Introduction

Statistical Agencies are obligated to protect the confidentiality of individuals when releasing sample microdata arising from social surveys. The risk assessment is typically based on a disclosure risk scenario where an ‘intruder’ attempts to link the sample microdata to available public data sources through a set of identifying key variables that are common to both sources. The identification of an individual could then be used to obtain more sensitive information and the disclosure of attributes. In order to limit the risk of identification, the statistical agency will implement disclosure limitation methods on the sample microdata, the extent of which depend on the mode of access (eg. safe data settings, special licensing, data archives and public use files). Disclosure limitation methods can be non-perturbative where the information content is reduced without altering the data. These include deleting variables, sub-sampling or recoding and collapsing categories of variables. Perturbative disclosure limitation methods alter the data by introducing forms of misclassification, These include data swapping (Dalenius and Reiss, 1982, Gomatam, Karr and Sanil, 2005), noise addition (Kim, 1986, Fuller, 1993, Brand, 2002) and the extreme case of fully synthetic data where the data released is based on a statistical model (Raghunathan, Reiter, and Rubin, 2003, Reiter, 2005). For more information on these methods see also: Willenborg and De Waal, 2001, Domingo-Ferrer and Torra, 2001.

Before releasing sample microdata, statistical agencies needs to quantify the risk of identification. One method for assessing this risk is to simulate an ‘intruder’ attack by using probabilistic record linkage techniques. One of the first examples was carried out in Spruill,1982 who linked perturbed sample microdata back to the

original sample using distance based matching. In many studies of this type, a conservative assessment of the risk of identification is obtained since it assumes that the ‘intruder’ has access to the original dataset and does not take into account the protection afforded by the sampling. More recent examples use the probabilistic record linkage framework of Fellegi and Sunter (F&S), 1969 (see: Yancy, Winkler and Creecy, 2002, Hawala, Stinson and Abowd, 2005 and Torra, Abowd and Domingo-Ferrer, 2006). The identifying key variables used for matching are typically categorical, such as sex, date of birth, marital status and locality. In the F&S framework, each potential pair is assigned a matching weight as described in Section 2. The matching weights are sorted and appropriate cut-offs determined according to pre-specified type I and type II error bounds. Pairs with high matching weights are considered to be correct matches and pairs with low matching weights are considered to be correct non-matches. Pairs with matching weights between the cut-off thresholds need to undergo clerical review. The matching weights are proxies for the probability of a correct match given an agreement. These probabilities can be used as individual record-level measures of disclosure risk. Global measures of disclosure risk include the proportion of correct matches, the proportion of correct matches to false matches, and one minus the estimated false match rate.

In Skinner, 2008, the probabilistic record linkage framework of F&S is linked to the probabilistic modelling framework for quantifying identification risk based on the notion of population uniqueness (see: Bethlehem, Keller and Pannekoek, 1990, Skinner and Holmes, 1998, Elamir and Skinner, 2006, Skinner and Shlomo, 2008). The probabilistic modelling framework relies on distributional assumptions to draw inference from the sample and estimate population parameters. The individual disclosure risk measure is the expectation of a correct match given a sample unique on the set of key variables. The global measure of disclosure risk is obtained by summing over the sample uniques to derive the expected number of correct matches. Skinner and Shlomo (2007) expanded the original probabilistic modelling framework to include measurement errors in the key variables, either arising naturally through data processing or purposely introduced into the data as a perturbative disclosure limitation method. In this paper we provide empirical evidence of the relationship between the probabilistic record linkage framework of F&S and the probabilistic modelling framework based on population uniqueness taking into account measurement errors.

In section 2 we introduce the notation and theory of the two frameworks for disclosure risk assessment: the F&S probabilistic record linkage framework and the probabilistic modelling framework based on sample uniques. We also provide examples that link the two frameworks as set out in Skinner, 2008. Section 3 presents an empirical study based on datasets from the UK 2001 Census. We first assume the perspective of the statistical agency where a perturbative method of disclosure limitation has been applied to the data and therefore misclassification

probabilities and population parameters are known. We apply both the record linkage and probabilistic modelling framework for assessing the risk of identification and compare results. We also demonstrate how an intruder might estimate population parameters through log-linear modelling in the probabilistic modelling framework or use the EM algorithm to estimate matching parameters in the F&S record linkage framework. We conclude in Section 4 with a discussion.

2 Notation and Theory

In this section we describe the F&S probabilistic record linkage framework and the probabilistic modelling framework based on notions of population uniqueness taking into account misclassification. We demonstrate the relationship between the two frameworks.

2.1 Fellegi and Sunter Probabilistic Record Linkage

Using the notation of Skinner, 2008, let \tilde{X}_a denote the value of the vector of cross-classified identifying key variables for unit a in the microdata ($a \in s_1$) and X_b the corresponding value for unit b in the external database ($b \in s_2$). Note that s_2 can be the population P or any subset $s_2 \subset P$. The different notation of X allows for different values of the two vectors due to natural misclassification in the data or an application of a perturbative disclosure limitation method to the sample microdata file. Denote this misclassification matrix by:

$$P(\tilde{X}_a = k | X_a = j) = \theta_{kj} \quad (1)$$

Based on the F&S theory of record linkage, a comparison vector $\gamma(\tilde{X}_a, X_b)$ is calculated for pairs of units $(a, b) \in s_1 \times s_2$ where the function $\gamma(\cdot, \cdot)$ takes values in a finite comparison space Γ . For the disclosure risk scenario we assume that the intruder uses the comparison vector to identify pairs of units which contain the same unit $(a, a) \in s_1 \times s_2$. Typically the intruder will use a combination of exact matching and probabilistic matching by considering only pairs that are blocked through an exact match on some subset $\tilde{s} \subset s_1 \times s_2$. The intruder seeks to partition the set of pairs in \tilde{s} into a set of matches: $M = \{(a, b) \in \tilde{s} | a \in s_1, b \in s_2, a = b\}$ and non-matches: $U = \{(a, b) \in \tilde{s} | a \in s_1, b \in s_2, a \neq b\}$. The approach by F&S is to define the likelihood ratio $m(\gamma)/u(\gamma)$ as the matching weight where: $m(\gamma) = P(\gamma(\tilde{X}_a, X_b) = \gamma | (a, b) \in M)$ and $u(\gamma) = P(\gamma(\tilde{X}_a, X_b) = \gamma | (a, b) \in U)$. We denote $m(\gamma)$ as the m -probability and $u(\gamma)$ as the u -probability. The higher values of the likelihood ratio are more likely to belong to M and the lower values of the likelihood ratio are more likely to belong to U . In addition, under the assumption of independence the m -probability and the u -probability can be split into individual

components for each separate key variable. Let $p = P((a,b) \in M)$ the probability that the pair is in M . The probability of a correct match $p_{M|\gamma} = P((a,b) \in M | \gamma(\tilde{X}_a, X_b))$ can be calculated using Bayes Theorem:

$$p_{M|\gamma} = m(\gamma)p / [m(\gamma)p + u(\gamma)(1-p)]. \quad (2)$$

An intruder might estimate the matching parameters $m(\gamma), u(\gamma)$ and p by linking the released microdata to an external file containing all or subsets of the population. The parameters can be estimated using the EM algorithm which is an iterative maximum likelihood estimation procedure for incomplete data. Based on the estimation of the parameters, we can calculate the probability of a correct match given an agreement $p_{M|\gamma}$ by the Bayes Theorem. The EM algorithm is described next:

Consider the case where an intruder identifies a sample unique with $\tilde{X}_a = j$ and matches it to all possible pairs in an arbitrary subset s_2 of the population P . The EM algorithm estimates the m and u probabilities and the proportion of correct matches p as follows:

Denote γ_q^a as the 1,0 agreement indicator for the a 'th record pair and the q 'th key variable ($q = 1, \dots, Q$) included in the vector \tilde{X}_a . The complete data is $\{\gamma^a, g\}$ where $\gamma^a = (\gamma_1^a, \gamma_2^a, \dots, \gamma_Q^a)$ is the agreement vector across Q key variables for the a 'th record pair. The vector g is the unknown indicator vector, $\{g_{am}, g_{au}\}$ where $g_{am} = 1$ if the record pair a is the same person (set M) and $g_{au} = 1$ if the record pair a does not belong to the same person (set U). The estimates of g_{am} and g_{au} are the conditional probabilities of being in set M or U given the observed data for the a 'th pair. Let \hat{p} be the estimated proportion of record pairs in set M . Applying Bayes Theorem, the estimates for the a 'th pair (assuming conditional independence of the key variables) are:

$$\hat{g}_{am} = \frac{\hat{p} \prod_{q=1}^Q m_q^{\gamma_q^a} (1-m_q)^{1-\gamma_q^a}}{\hat{p} \prod_{q=1}^Q m_q^{\gamma_q^a} (1-m_q)^{1-\gamma_q^a} + (1-\hat{p}) \prod_{q=1}^Q u_q^{\gamma_q^a} (1-u_q)^{1-\gamma_q^a}}$$

and

$$\hat{g}_{au} = \frac{(1-\hat{p}) \prod_{q=1}^Q u_q^{\gamma_q^a} (1-u_q)^{1-\gamma_q^a}}{\hat{p} \prod_{q=1}^Q m_q^{\gamma_q^a} (1-m_q)^{1-\gamma_q^a} + (1-\hat{p}) \prod_{q=1}^Q u_q^{\gamma_q^a} (1-u_q)^{1-\gamma_q^a}} \quad (3)$$

With the estimated conditional probabilities given the data from the E-step, the M-step of the EM algorithm is used to update the conditional m and u probabilities and the estimated proportion of correct matches p as follows:

$$\hat{m}(\gamma_q) = \sum_{i=1}^R \hat{g}_{am} \gamma_{qi}^a / \sum_{i=1}^R \hat{g}_{am} \quad \text{and} \quad \hat{u}(\gamma_q) = \sum_{i=1}^R \hat{g}_{au} \gamma_{qi}^a / \sum_{i=1}^R \hat{g}_{au} \quad \text{where } R \text{ is the total number of record pairs. The update of the estimate for the proportion of record pairs in set } M \text{ is: } \hat{p} = \sum_{i=1}^R \hat{g}_{am} / R.$$

2.2 Probabilistic Modelling for Measuring Identification Risk

The probabilistic modelling framework for estimating the risk of identification is based on theory which uses models for categorical key variables. Let $f = \{f_j\}$ denote a q -way frequency table, which is a sample from a population table $F = \{F_j\}$, where $j = (j_1, \dots, j_q)$ indicates a cell and f_j and F_j denote the frequency in the sample and in the population cell j , respectively. Denote by n and N the sample and population size, respectively and the number of cells by J . We assume that the q attributes in the table are categorical identifying key variables. Disclosure risk arises from small cells, and in particular when $f_j = F_j = 1$ (sample and population uniques). We focus on a global disclosure risk measure based on sample uniques: $\tau = \sum_j I(f_j = 1)1/F_j$. This measure is the expected number of correct matches if each sample unique is matched to a randomly chosen individual from the same population cell. We consider the case that f is known, and F is an unknown parameter and the quantity τ should be estimated. An estimate of τ is:

$$\hat{\tau} = \sum_j I(f_j = 1) \hat{E}[1/F_j | f_j = 1] \quad (4)$$

where \hat{E} denotes an estimate of the expectation. The formula in (4) is naïve in the sense that it ignores the possibility of misclassification. A common assumption in the frequency table literature is $F_j \sim \text{Poisson}(\lambda_j)$, independently, where $\sum_j F_j = N$ is a random parameter. Binomial (or Poisson) sampling from F_j means that $f_j | F_j \sim \text{Bin}(F_j, \pi_j)$ independently, where π_j is the sampling fraction in cell j . By standard calculations we then have:

$$f_j \sim \text{Poisson}(\lambda_j \pi_j) \quad \text{and} \quad F_j | f_j \sim f_j + \text{Poisson}(\lambda_j(1 - \pi_j)), \quad (5)$$

where $F_j | f_j$ are conditionally independent.

We take the approach as developed in Skinner and Holmes, 1998 Elamir and Skinner, 2006 and Skinner and Shlomo, 2009 and use log linear models to estimate population parameters and estimate identification risk. The sample counts $\{f_j\}$ are

used to fit a log-linear model: $\log \mu_j = x'_j \beta$ where $\mu_j = \lambda_j \pi_j$ in order to obtain estimates for the parameters: $\hat{\lambda}_j = \hat{\mu}_j / \pi_j$. Under simple random sample and $\pi_j = \pi$ for all j , the maximum likelihood (MLE) estimator $\hat{\beta}$ may be obtained by solving the score equations: $\sum_j [f_j - \exp(x'_j \beta)] x_j = 0$. Under a complex survey design and differential weights, a pseudo-likelihood approach can be taken (Rao and Thomas, 2003) for estimating $\hat{\beta}$. Using the second part of (5), the expected individual disclosure risk measures for cell j is defined by:

$$E_{\lambda_j} (1/F_j | f_j = 1) = [1 - e^{-\lambda_j(1-\pi)}] / [\lambda_j(1-\pi)]. \quad (6)$$

Plugging $\hat{\lambda}_j$ for λ_j in (6) leads to the desired estimates $\hat{E}_{\hat{\lambda}_j} [1/F_j | f_j = 1]$ and then to $\hat{\tau}$ of (4).

The probabilistic modelling approach does not consider the case of misclassification naturally arising in surveys or purposely introduced into the data as a disclosure limitation method. Skinner and Shlomo (2007) defined disclosure risk measures that take into account misclassification. The identification risk in this case is defined as:

$$\Pr(A = B | \tilde{f}_j = 1) = [\theta_{jj} / (1 - \pi\theta_{jj})] / [\sum_k F_k \theta_{jk} / (1 - \pi\theta_{jk})] \quad (7)$$

and it follows that $\Pr(A = B | \tilde{f}_j = 1) \leq 1/F_j$ with equality holding if there is no misclassification. The extent to which the left hand side of this inequality is less than the right hand side measures the impact of misclassification on disclosure risk. If the sampling fraction is small we can approximate (7) by:

$$\Pr(A = B | \tilde{f}_j = 1) \approx \theta_{jj} / (\sum_k F_k \theta_{jk})$$

Moreover, if the population size is large, we have approximately $\sum_k F_k \theta_{jk} \approx \tilde{F}_j$, where \tilde{F}_j is the number of units in the population which would have $\tilde{X} = j$ if they were included in the microdata (with misclassification). Hence a simple approximate expression for the risk, natural for many social surveys, is

$$\Pr(A = B | \tilde{f}_j = 1) \approx \theta_{jj} / \tilde{F}_j \quad (8)$$

Note that the approximations in (8) does not depend upon θ_{jk} for $j \neq k$ and so knowledge of these probabilities is not required in the estimation of risk if ‘acceptable’ estimates of θ_{jj} and \tilde{F}_j are available. The definition of risk in (7) applies to a specific record. The aggregated measure across sample unique records, defined from (7) is

$$\tau_\theta = \sum_{j \in SU} [\theta_{jj} / (1 - \pi\theta_{jj})] / [\sum_k F_k \theta_{jk} / (1 - \pi\theta_{jk})] \quad (9)$$

where SU is the set of key variable values which are sample unique. Similar to the case with no misclassification, this measure may be interpreted as the expected number of correct matches among sample uniques.

Since the values of F_j or \tilde{F}_j appearing in (7) through (8) are unknown, we need to estimate them. We do suppose that the values of θ_{jk} are known, especially when a statistical agency purposely perturbs the data as a disclosure limitation method. Expression (8) provides a simple way to extend the log-linear modelling approach described above provided θ_{jj} is known. Since the $\tilde{f}_j, j=1, \dots, J$ represent the available data, all that is required is to ignore the misclassification and estimate $1/\tilde{F}_j$ from the $\tilde{f}_j, j=1, \dots, J$ by fitting a log-linear model to the $\tilde{f}_j, j=1, \dots, J$ following the same criteria as before. This results in an estimate $\hat{E}(1/\tilde{F}_j | \tilde{f}_j = 1)$ based on the assumptions of the Poisson distribution for the population and sample counts. These estimates should be multiplied by θ_{jj} values and summed if aggregate measures of the form in (9) are needed.

2.3 The Relationship Between Frameworks

Skinner 2008 relates the F&S record linkage framework to the probabilistic modelling framework by providing the following examples:

Example 1: Assume no misclassification has occurred, i.e. $\tilde{X}_a = X_a$ in both the population (P) and the sample (s) and that the true match status is known by the agency. Assume that sample (s) was drawn by simple random sampling from the population P . We can calculate the contingency table in Table 1 for each $X_a = j$ in the realized sample where the rows are a binary agreement/disagreement on the comparison vector: $\gamma(X_a, X_b)$ for pairs $(a, b) \in s \times P$ and the columns the matching status.

Table 1: Contingency table of binary agreement status and match status for $X_a = j$ with no misclassification

	Non-match	Match	Total
Disagree	$n(N-1) - f_j(F_j-1)$	$n - f_j$	$Nn - f_j F_j$
Agree	$f_j(F_j-1)$	f_j	$f_j F_j$
Total	$n(N-1)$	n	Nn

From Table 1, we calculate directly $p_{M|\gamma} = 1/F_j$. We also obtain: $m(\gamma) = f_j/n$, $u(\gamma) = f_j(F_j - 1)/n(N - 1)$ and the probability of a correct match: $p = 1/N$. Small F_j therefore results in a high probability of a correct match given an agreement in the comparison vector.

Example 2: In continuation of Example 1, assume now that the microdata has undergone misclassification (either as a result of errors or purposely perturbed for disclosure limitation). Denote \tilde{f}_j the observed misclassified sample counts with $\tilde{X}_a = j$ derived by $\tilde{f}_j = \theta_{jj}f_j + \sum_{k \neq j} \theta_{jk}f_k$. We calculate the contingency table on the realized misclassified sample in Table 2 for each $\tilde{X}_a = j$ where the rows are a binary agreement/disagreement on the comparison vector: $\gamma(\tilde{X}_a, X_b)$ for pairs $(a, b) \in s \times P$ and the columns the matching status.

Table 2: Contingency table of binary agreement status and match status for $\tilde{X}_a = j$ with misclassification

	Non-match	Match	Total
Disagree	$Nn - n - \tilde{f}_j F_j + \theta_{jj} f_j$	$n - \theta_{jj} f_j$	$Nn - \tilde{f}_j F_j$
Agree	$\tilde{f}_j F_j - \theta_{jj} f_j$	$\theta_{jj} f_j$	$\tilde{f}_j F_j$
Total	$Nn - n$	n	Nn

From Table 2, we can calculate directly $p_{M|\gamma} = \theta_{jj} f_j / \tilde{f}_j F_j \approx \theta_{jj} / \tilde{\pi}_j \approx \theta_{jj} / \tilde{F}_j$ where \tilde{F}_j is the number of units in the population (P) with $\tilde{X}_a = j$ (imagining that the misclassification takes place before the sampling). We also obtain: $m(\gamma) = \theta_{jj} f_j / n$, $u(\gamma) = (\tilde{f}_j F_j - \theta_{jj} f_j) / n(N - 1)$ and the probability of a correct match: $p = 1/N$. The expression for $p_{M|\gamma}$ is similar to expression (8) derived from the probabilistic modelling framework. Skinner, 2008 also shows that the derivation of the probability of a correct match given an agreement holds for any subset of the population which may be selected arbitrarily.

3 Empirical Study

In this section, we provide empirical evidence based on real datasets of the connection between probabilistic record linkage according to F&S and the probabilistic modelling framework for calculating the risk of identification. We start

from the perspective of the statistical agency where it is assumed that the misclassification matrix is known either because the data was purposely perturbed by the agency for disclosure limitation or a study was carried out to assess error rates in various stages of the data processing. We begin with assuming that population counts are known and hence the agency can calculate the necessary parameters to measure identification risk in both frameworks for the comparison. We also consider the case where an intruder would need to estimate the necessary parameters to identify high risk records and examine the proximity of estimated individual per-record disclosure risk measures to true disclosure risk measures in both frameworks.

3.1. Preparation of the Data

We use the method of data swapping on an extract of individuals from the 2001 UK Census to compare the F&S framework and the probabilistic modelling framework. The population includes $N=1,468,255$ individuals and we draw a 1% Bernoulli Sample ($n=14,683$). There are six key variables for the risk assessment: Local Authority (LAD) (11), sex (2), age group (24), marital status (6), ethnicity (17) and economic activity (10) where the numbers of categories of each variable are in parenthesis ($J=538,560$). We implement a random data swap by drawing a 20% sub-sample in each of the LADs. In each of the sub-samples, half of the individuals are flagged. For each flagged individual, an unflagged individual is randomly chosen within the sub-sample and their LAD variables swapped, on condition that the individual chosen was not previously selected for swapping and that the two individual do not have the same LAD, i.e. no individual is selected twice for producing a pair.

The misclassification matrix θ for the data swapping design of LAD can be expressed in terms of the 11 by 11 misclassification matrix defined by:

(1) On the diagonal: $\theta_{jj} = 0.8$

(2) Off the diagonal: $\theta_{jk} = 0.2[n_k / \sum_{l \neq j} n_l]$ where n_k is the number of records in the sample in LAD k .

The number of sample uniques on the misclassified sample is 2,853.

3.2 Identification Risk Based on Probabilistic Modelling

Since we know the misclassification matrix θ and the true population counts F_j in this study, we can assess the performance of the naïve risk measure in (6) and under misclassification in (7) based on the probabilistic modelling framework. Table 3 presents global disclosure risk measures for our sample, which are obtained by summing individual risk measures across the sample uniques. The first row of Table 3 shows the true disclosure risk τ in terms of the expected number of correct matches in the data before the misclassification. The second row contains the

estimated naive disclosure risk measure of (4) ignoring the misclassification. The third row in Table 3 contains the true disclosure risk τ_θ in (9) taking into account the misclassification and the final measure the estimated disclosure risk measure under misclassification $\hat{\tau}_\theta$ defined by summing $\theta_{jj} \hat{E}(1/\tilde{F}_j | \tilde{f}_j = 1)$ across sample uniques. As can be seen, estimation of global disclosure risk measures follow closely true disclosure risk measures and are generally accurate (see Skinner and Shlomo 2008 for a discussion on model selection and goodness of fit criteria for estimating the risk of identification using log-linear modelling).

Table 3: Global risk measure on sample uniques for the 20% random data swap in the probabilistic modelling framework

Global Risk Measure	Expected correct matches out of sample uniques
True risk measure τ in original sample	358.1
Estimated naïve risk measure (4) ignoring misclassification	358.6
Risk measure (8) under misclassification τ_θ	298.9
Estimated risk measure under misclassification $\hat{\tau}_\theta$	286.8

The individual per-record risk measures as presented in (7) are more difficult to estimate accurately by estimates: $\theta_{jj} \hat{E}(1/\tilde{F}_j | \tilde{f}_j = 1)$. Figure 1 compares the individual per-record estimated risk measures $\theta_{jj} \hat{E}(1/\tilde{F}_j | \tilde{f}_j = 1)$ on the X-axis with the individual risk measure (7) on the Y-axis assuming the misclassification matrix is known. Equation (7) also assumes that population counts are known. The figure is presented on the logarithmic scale.

Figure 1 confirms that on average, the global disclosure risk measure in (9) is estimated accurately with the graph being symmetrical about the equality diagonal. The individual per-record risk measures however vary and the estimation is less accurate. From the perspective of an intruder who might use log-linear modelling to identify high risk individuals, it would be difficult to ascertain exactly which of the individuals are population uniques.

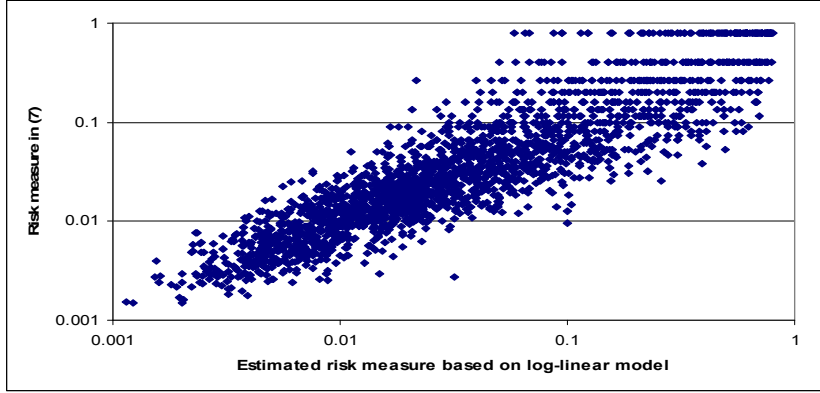


Figure 1: Graph of individual risk measure in (7) and estimated individual risk measure: $\theta_{jj} \hat{E}(1/\tilde{F}_j | \tilde{f}_j = 1)$ (logarithmic scale)

We turn now to the F&S probabilistic record linkage framework. For our record linkage experiment we block on all key variables that match exactly and calculate the probability of a correct match given an agreement on the perturbed LAD. We focus only on sample uniques in order to compare to the probabilistic modelling framework. All possible pairs between the population dataset and the 2,853 sample uniques result in a dataset with 1,534,293 possible pairs. Table 4 presents the counts based on all possible pairs according to the true match status and the agreement pattern in LAD.

Table 4: Frequency counts for all sample uniques

	Non-match	Match	Total
Disagree LAD	1,388,069	619	1,388,688
Agree LAD	143,321	2,234	145,555
Total	1,531,390	2,853	1,534,293

From Table 4, $m(\gamma) = 0.78$, $u(\gamma) = 0.09$ and the probability of a correct match $p = 0.002$. Note that the m -probability is approximately the same as the overall non-misclassification rate (the diagonal of the misclassification matrix θ). The u -probability represents the proportion of random agreements on LAD. On average, the probability of a correct match given an agreement on LAD is: $p_{M|\gamma} = 0.015$ or 1.5%.

To assess the probability of a correct match given an agreement $p_{M|\gamma}$ for each individual $\gamma(\tilde{X}_a, X_b) = j$ separately, we carry out the record linkage procedure for each j . In Figure 2, we compare these probabilities to $\theta_{jj} / \tilde{F}_j$ in (8) based on the probabilistic modelling framework for each j . The individual disclosure risk measures in (8) follow closely the probabilities of a correct match given an agreement from the F&S framework. Summing over the $p_{M|\gamma}$ across the sample uniques that agree on LAD, we obtain the global disclosure risk measure of 289.5.

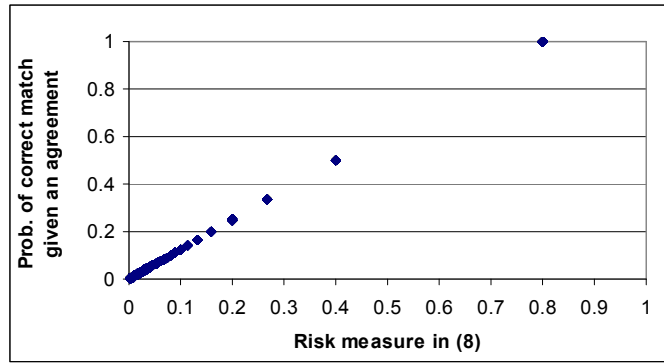


Figure 2: Plot of $p_{M|\gamma}$ against $\theta_{jj} / \tilde{F}_j$ in (8) for each $\gamma(\tilde{X}_a, X_b) = j$

Turning to the estimation of $p_{M|\gamma}$, we demonstrate the procedure of the EM algorithm using data from one particular $\tilde{X}_a = j$ in the real dataset as shown in Table 5.

Table 5: Data for EM Algorithm

	Non-match	Match	Total
Disagree LAD	2,283	1	2,284
Agree LAD	48	2	50
Total	2,331	3	2,334

The true parameters from the table $m(\gamma) = 0.667$, $u(\gamma) = 0.021$, the probability of a correct match $p = 0.0013$ and $\hat{p}_{M|\gamma} = 2/50 = 0.040$. We initiate the EM algorithm with the m -probability of approximately the error rate 0.75, the u -probability of 0.02 which is the percent of random agreements and the overall probability of a correct match 0.002. Convergence in the EM algorithm means that the sum of squared change of the estimates $\hat{m}(\gamma)$ and $\hat{u}(\gamma)$ is less than 0.0000001. The estimation of the

EM algorithm resulted in: $\hat{m}(\gamma) = 0.726$, $\hat{u}(\gamma) = 0.020$, and $\hat{p} = 0.0015$. From here, we obtain: $\hat{p}_{M|\gamma} = \frac{0.0015(0.726)}{0.0015(0.726) + (1 - 0.0015)(0.020)} = 0.052$.

As can be seen, it is difficult to estimate the parameters exactly using the EM algorithm. Generally, the EM algorithm will estimate parameters more accurately when there is a large number of pairs and a relatively large number of correct matches (approximately over 5%).

4 Discussion

In this paper, we have provided empirical evidence of the connection between the F&S record linkage framework to the probabilistic modelling framework for estimating identification risk based on the notion of population uniqueness as discussed in Skinner, 2008. We have seen that statistical agencies are able to estimate accurate global disclosure risk measures that can be used to assess optimal disclosure limitation methods in a risk-utility framework assuming that the probability of not being misclassified is known.

Individual per-record disclosure risk measures are more difficult to estimate without knowing true population parameters in both frameworks. The estimation is carried out through the use of log linear modelling for the probabilistic modelling framework or the EM algorithm for the F&S record linkage framework. The results show that from the perspective of the intruder, it is difficult to identify high risk sample uniques, and in particular when they are population uniques, due to the variability of the estimation of the risk measures.

References

- Bethlehem, J., Keller, W., and Pannekoek, J. (1990). Disclosure Control of Microdata. *Journal of the American Statistical Association* 85, pp 38-45.
- Brand, R. (2002). Micro-data Protection Through Noise Addition. In *Inference Control in Statistical Databases* (ed. J. Domingo-Ferrer), New York: Springer, pp 97-116.
- Dalenius, T. and Reiss, S.P. (1982). Data Swapping: a Technique for Disclosure Control. *Journal of Statistical Planning and Inference*, pp 73-85.
- Domingo-Ferrer, J. and Torra, V. (2001). A Quantitative Comparison of Disclosure Control Methods for Microdata, in (P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz, eds.) *Confidentiality, Disclosure Control and Data Access: Theory and Practical Applications*. Amsterdam, The Netherlands: North Holland, pp 111-145.

- Elamir, E. And Skinner, C.J. (2006). Record-Level Measures of Disclosure Risk for Survey Micro-data. *Journal of Official Statistics*, 22, pp 525-539.
- Fellegi, I. and Sunter, A. (1969). A Theory for Record Linkage.. *Journal of the American Statistical Association*, 64, pp 1183-1210.
- Fuller, W. (1993). Masking Procedures for Micro-data Disclosure Limitation. *Journal of Official Statistics*, 9, pp 383-406.
- Hawala, S., Stinson, M., Abowd, J. (2005). Disclosure Risk Assessment Through Record Linkage. In: *Proceedings of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Geneva*.
- Kim, J.J. (1986). A Method for Limiting Disclosure in Micro-data Based on Random Noise and Transformation. American Statistical Association, *Proceedings of the Section on Survey Research Methods*, pp 370-374.
- Raghunathan, T.E., Reiter, J. and Rubin, D. (2003). Multiple Imputation for Statistical Disclosure Limitation. *Journal of Official Statistics*, 19, No. 1, pp 1-16.
- Rao, J.N.K and Thomas, D.R. (2003). Analysis of Categorical Response Data from Complex Surveys: an Appraisal and Update in (Chambers, R.L. and Skinner, C.J. eds.) *Analysis of Survey Data*, UK: Wiley and sons, pp 85-108.
- Reiter, J.P. (2005), Releasing Multiply Imputed, Synthetic Public-Use Microdata: An Illustration and Empirical Study. *Journal of the Royal Statistical Society, A*, Vol.168, No.1, pp 185-205.
- Skinner, C.J. (2008). Assessing Disclosure Risk for Record Linkage in (Domingo-Ferrer, J. And Saygin, Y., eds.) *Privacy in Statistical Databases, Lecture Notes in Computer Science 5262*, Berlin: Springer, pp 166-176.
- Skinner, C. and Holmes, D. (1998). Estimating the Re-identification Risk per Record in Microdata. *Journal of Official Statistics*, 14, pp 361-372.
- Skinner, C.J. and Shlomo, N. (2007) Assessing the Disclosure Protection Provided by Misclassification and Record Swapping. *56th Session of the International Statistical Institute Invited Paper*, Lisbon 2007.
- Skinner, C.J. and Shlomo, N. (2008). Assessing Identification Risk in Survey Microdata Using Log-linear Models. *Journal of the American Statistical Association*, Vol. 103, Number 483, pp 989-1001.
- Spruill, N.L. (1982). Measures of Confidentiality. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pp 260-265.
- Torra, V., Abowd, J.M. and Domingo-Ferrer, J. (2006). Using Mahalanobis distance-based record linkage for disclosure risk assessment in (in (Domingo-Ferrer, J. And Franconi, L. eds.) *Privacy in Statistical Databases, Lecture Notes in Compute Science, 4302*. Berlin: Springer pp 233-242.
- Willenborg, L.C.R.J. and De Waal, T. (2001). *Elements of Statistical Disclosure Control in Practice*. Lecture Notes in Statistic 155. New York: Springer-Verlag.
- Yancey, W.E., Winkler, W.E. and Creecy, R.H. (2002). Disclosure Risk Assessment in Perturbation Micro-Data Protection in (Domingo-Ferrer, J., ed.) *Inference Control in Statistical Databases*, New York: Springer, pp 135-151.