DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 1, 2018, pp. 41–54, http://dx.doi.org/10.1515/JOS-2018-0003

Design-Based Estimation with Record-Linked Administrative Files and a Clerical Review Sample

Abel Dasylva[1]

This article looks at the estimation of an association parameter between two variables in a finite population, when the variables are separately recorded in two population registers that are also imperfectly linked. The main problem is the occurrence of linkage errors that include bad links and missing links. A methodology is proposed when clerical-reviews may reliably determine the match status of a record-pair, for example using names, demographic and address information. It features clerical-reviews on a probability sample of pairs and regression estimators that are assisted by a statistical model of comparison outcomes in a pair. Like other regression estimators, this estimator is design-consistent regardless of the model validity. It is also more efficient when the model holds.

Key words: Probabilistic record-linkage; administrative data; clerical-review; mixture-model; probability sample.

1. Introduction

Computerized record-linkage aims at linking records that relate to the same individual or entity, with minimal human intervention. Such records are called matched records. In many cases, this is a challenging task because it must be based on pseudo-identifiers such as names and demographic characteristics, which are non-unique and possibly recorded with spelling variations or typographical errors. These limitations lead to errors that include bad links and missing links.

In general the computerized linkage of two large files comprise of five major steps. First, the linkage variables are parsed and standardized. Second, records in the two files are compared using blocking keys. Only the pairs that agree on some blocking key are subsequently compared more extensively. Third, the linkage variables are extensively compared to produce comparison outcomes. Fourth, a decision is made for each pair. Finally, conflicting linkage decisions are dealt with, such as when linking the same census record to two death records. Linkage methodologies differ according to how linkage decisions are made in the fourth step. In a deterministic linkage, the decision may be based on arbitrary criteria, according to subject matter knowledge. In probabilistic linkage, the decision is based on a linkage weight which is a measure of the similarity between two records. This weight is typically the sum of outcome weights that correspond to the similarity of the different linkage variables. For the final decision, a pair linkage weight is

[1] Statistics Canada, SRID, 100 Tunney's Pasture, Ottawa, Ontario K1A0T6, Canada. Email: abel.dasylva@statcan.gc.ca

compared to one or two thresholds to determine whether it should be linked, reviewed clerically or rejected (Fellegi and Sunter 1969). The overall linkage performance is characterized by the rates of linkage errors, which are determined by the linkage weights and thresholds.

Two files may be linked to study the association between variables that have been recorded separately in each file. For example, consecutive censuses may be linked to create a longitudinal dataset. In this case, the variables of interest measure the same characteristic at different time points. Estimation with an imperfectly linked dataset is challenging because linkage errors must be accurately measured and accounted for (Lahiri and Larsen 2005; Chipperfield et al. 2011; Chambers 2009). The measurement may be based on a statistical model, clerical-reviews or both.

In theory, linkage errors may be estimated from a model, without any human intervention. On one hand, Fellegi and Sunter (1969) have suggested models based on the assumption that the linkage variables are conditionally independent given the match status. However, these models have been quite inaccurate (Belin and Rubin 1995). On the other hand, models that incorporate interactions may lack the identification property, see Kim (1984) and more recently Fienberg et al. (2009). The above difficulties justify the continued use of clerical-reviews or training samples (Belin and Rubin 1995; Howe 1981; Heasman 2014; Gill 2001; Guiver 2011), possibly in conjunction with a statistical model (Larsen and Rubin 2001).

In this work, the problem that consists in estimating an association parameter from an imperfectly linked dataset is framed as a survey sampling problem. In general, survey sampling aims at estimating a finite population parameter without bias, by taking a probability sample, where each population unit has a known and positive inclusion probability. Using such a sample, a population total is estimated without bias with an Horwitz-Thompson (HT) estimator; the sum of sample values weighted by the corresponding reciprocal selection probability. However, the HT estimator may have a large variance, especially when the inclusion probability is not correlated with the variable of interest. A popular alternative is a regression estimator when some auxiliary variables are observed for all population units. The regression estimator is not unbiased but design-consistent, that is, with a bias that is negligibly small in large samples. This estimator also has a smaller variance than the HT estimator, when the variable of interest is a nearly linear function of the auxiliary variables. Regression estimators offer examples of generalized regression estimators (GREG) and calibration estimators that have been thoroughly studied by Särndal et al. (1992) and by Deville and Särndal (1992). These estimators are also referred to as model-assisted estimators because they are inspired by some implicit statistical model; typically a linear model relating the auxiliary variables to the variables of interest. They are efficient when the model holds and less so otherwise. However they remain design-consistent regardless of the model validity (see Särndal et al. 1992, section 6.7, pp. 239).

The proposed problem formulation brings questions that have been already addressed by Särndal et al. (1992) and others, about optimal sampling designs, design-consistent estimators and the efficient use of auxiliary information through statistical models. This body of work is applied to our problem with some adaptation. The resulting estimators are regression estimators, that are built in two steps. First, all record-pairs that satisfy blocking

criteria are used to fit a model for predicting the match status of pairs within the blocks, irrespective of whether they are part of the clerical sample. Second, a regression estimator is fitted based on the clerical data. The described framework also applies when the match status is determined by other means than clerical reviews, for example through limited access to unique identifiers or additional information from a third party.

The following sections are organized as follows. Section 2 presents the notation and background. Section 3 describes model-based estimators in the record-linkage context. Section 4 discusses sampling designs. Section 5 presents simulation results. Section 6 presents the conclusions and future work.

2. Notation and Background

Consider two duplicate-free registers A and B, which contain records about N individuals. Register A contains K linkage variables and the variable of interest x_i for the ith record in A. Register B contains the same linkage variables as A and the variable of interest y_j for the jth record in B. Let U denote the finite population of all N^2 record-pairs in the cartesian product of the two files, that is, of all pairs (i,j) where $1 \leq i,j \leq N$.

For the record-pair (i,j) in the Cartesian product of the two registers, the linkage variables may be compared to produce a K-tuple $\boldsymbol{\gamma}_{ij} = \left(\gamma_{ij}^{(1)}, \ldots, \gamma_{ij}^{(K)}\right)$ of comparison outcomes, also called vector of comparison outcomes. In large files, some linkage variables are also coarsely compared to define blocks that altogether represent a small subset U^* of U and yet contain most matched pairs. The subset U^* of blocked pairs is the union of B disjoint subsets, $U_1^* \ldots U_B^*$, where each subset represents a distinct block. For each pair, this blocking information is also included in the comparison vector $\boldsymbol{\gamma}_{ij}$. The comparison vector $\boldsymbol{\gamma}_{ij}$ provides the basis for linking the records, for example using Fellegi and Sunter (1969) optimal linkage rule. However such a linkage is not required in the proposed estimation methodology.

Let M_{ij} denote the indicator variable that is set to 1 if the pair (i,j) is matched, that is, associated with the same individual. The variable M_{ij} is also called the match status of the pair (i,j). The comparison vector $\boldsymbol{\gamma}_{ij}$ is crucial for making an inference $\widehat{M}_{ij}$ about the unknown match status M_{ij}. The inferred match status $\widehat{M}_{ij}$ can take many forms. For example, it can be set to the conditional or posterior match probability $P\left(M_{ij} = 1 \middle| \boldsymbol{\gamma}_{ij}\right)$ given the comparison vector. It can also be interpreted as the "weight-share" of the pair (i,j), with the meaning of the Generalized Weight Share Method. See Lavallée (2002, chap. 9) for applications of this method to record-linkage.

For finite population inference, the goal is estimating a total of the following form:

$$Z = \sum_{(i,j) \in U} M_{ij} z_{ij} \tag{1}$$

In the above expression, $z_{ij} = f(x_i, y_j)$ and f is some known function.

For model-based inference, assume that the record-generating individuals represent an Independent Identically Distributed (IID) sample according to some distribution or superpopulation depending on a parameter θ. Inference about this parameter may be based on an equation of the form $E\left[S(\theta; x, y)\right] = 0$, where S is a score function (e.g., a log-likelihood), while (x,y) is the observation associated with an individual from the

superpopulation. The parameter θ may be estimated through the following unbiased estimating equation where $z_{ij}(\theta) = S(\theta; x_i, y_j)$.

$$\sum_{(i,j)\in U} M_{ij} z_{ij}(\hat{\theta}) = 0 \tag{2}$$

In both cases, the inferences use the recorded values of the variables in matched pairs, regardless of whether these values are free of nonsampling errors such as typographical errors, measurement errors, etc.

Resources for error-free clerical reviews are available to measure the match status. However they are costly and must be minimized. The clerical sample s has a fixed size. It is split into a blocking stratum U^* and a nonblocking stratum $U \setminus U^*$. Let s^* denote the sample of blocked pairs in the clerical sample. The samples in the different strata are selected independently and their sampling designs are arbitrary. Let π_{ij} denote the first-order sample inclusion probability for the record-pair (i,j).

3. Model-Assisted Estimators

The proposed estimators are regression estimators (Särndal et al. 1992, chap. 6) that have the following general difference form:

$$\hat{Z} = \underbrace{\sum_{(i,j)\in U^*} \widehat{M}_{ij} z_{ij} + \sum_{(i,j)\in s^*} \pi_{ij}^{-1} z_{ij} \left(M_{ij} - \widehat{M}_{ij} \right)}_{(1)} + \underbrace{\sum_{(i,j)\in s\setminus s^*} \pi_{ij}^{-1} M_{ij} z_{ij}}_{(2)} \tag{3}$$

This estimator is the sum of contributions from the two strata. The first contribution exploits the inferred match status to estimate the total over the blocking stratum with a greater precision. The second contribution is simply a Horwitz-Thompson estimator for the total over the nonblocking stratum. The above estimator may be viewed as a calibration estimator (Deville and Särndal 1992), where the estimated total is calibrated to the corresponding total based on inferred match status. It estimates the total with no sampling error and no bias when the following two conditions are met:

i. Perfect blocking criteria selecting all matched pairs.
ii. Perfect inference of the match status, that is, $M_{ij} = \widehat{M}_{ij}$.

The estimator is also unbiased if the inferred status ignores the information of the clerical sample:

$$E\left[\hat{Z}|U\right] = \sum_{(i,j)\in U} M_{ij} z_{ij} \; = \; Z \tag{4}$$

This is the case if $\widehat{M}_{ij}$ is only a function of z_{ij} and γ_{ij}. The inferred status may be set to the conditional match probability given the vector of comparison outcomes and the variables x_i, y_j, that is,

$$\widehat{M}_{ij} = P\left(M_{ij} = 1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}\right) \tag{5}$$

This particular inference strategy would minimize the mean squared error (over the super population) between the predicted total $\sum_{(i,j)\in U^*} \widehat{M}_{ij} z_{ij}$ and the actual total $\sum_{(i,j)\in U^*} M_{ij} z_{ij}$ over the blocking stratum, among all inference strategies where $\widehat{M}_{ij}$ is only a function of x_i, y_j and $\boldsymbol{\gamma}_{ij}$, if the record-pairs were IID. Under a Simple Random Sampling (SRS) design in the blocking stratum, the resulting estimator would also be more efficient than the Horwitz-Thompson estimator, if the pairs were IID.

The conditional match probability may be estimated under the assumption of IID pairs according to a two-component mixture distribution, where the different comparison outcomes and the variables x_i, y_j are assumed conditionally independent given the match status, where $\tau = 0, 1$:

$$P\left(x_i, y_j, \boldsymbol{\gamma}_{ij} \middle| M_{ij} = \tau\right) \;=\; P\left(x_i, y_j \middle| M_{ij} = \tau\right) \prod_{k=1}^{K} P\left(\gamma_{ij}^{(k)} \middle| M_{ij} = \tau\right) \tag{6}$$

The parameters $\boldsymbol{\psi}$ of this mixture include the mixing proportion $\lambda = P\left(M_{ij} = 1\right)$, the marginal m-probabilities $P\left(x_i, y_j \middle| M_{ij} = 1\right)$ and $P(\gamma_{ij}^{(k)} | M_{ij} = 1)$, and the marginal u-probabilities $P\left(x_i, y_j \middle| M_{ij} = 0\right)$ and $P(\gamma_{ij}^{(k)} | M_{ij} = 0)$, under the assumption of IID pairs. They may be estimated with an Expectation-Maximization (E-M) algorithm. See Jaro (1989) or Winkler (1988) for applications of E-M to record-linkage, and Dempster et al. (1977) for a general reference on E-M. An important feature of this mixture model is the use of x_i and y_j as additional linkage variables. The mixture model becomes simpler when the variables x_i and y_j are highly correlated with the linkage variables. In this case x_i and y_j bring no new information about the match status, given $\boldsymbol{\gamma}_{ij}$. Mathematically, this is expressed by the conditional independence of $\left(x_i, y_j\right)$ and the match status given the comparison outcomes:

$$P\left(M_{ij} = 1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}\right) \;=\; P\left(M_{ij} = 1 \middle| \boldsymbol{\gamma}_{ij}\right) \tag{7}$$

The inference strategy may be inefficient if the assumed mixture model does not hold. For example, this problem may occur if the couple $\left(x_i, y_j\right)$ contains additional information about the match status, but the inference $\widehat{M}_{ij} = P\left(M_{ij} = 1 \middle| \boldsymbol{\gamma}_{ij}\right)$ is used instead. The estimator is also less efficient if the linkage variables are correlated but their conditional independence is assumed.

Let $P\left(M_{ij} = 1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}; \hat{\boldsymbol{\psi}}\right)$ denote a preliminary estimate of the conditional match probability according to the mixture model. This estimate is computed in the E-Step of the E-M algorithm and it does not use the clerical results. In most cases, this mixture model will estimate the conditional match probability with some bias even if it accounts for some of the interactions among the different variables. To adjust for this bias, the match status may be inferred using a linear function $\beta_0 + \beta_1 P\left(M_{ij} = 1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}; \hat{\boldsymbol{\psi}}\right)$ of the estimated conditional probability, where the regression coefficients β_0 and β_1 are estimated from the clerical sample. In this case, the inferred match status is computed as follows:

$$\widehat{M}_{ij} \;=\; \hat{\beta}_0 + \hat{\beta}_1 P\left(M_{ij} = 1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}; \hat{\boldsymbol{\psi}}\right) \tag{8}$$

A special case is when a ratio estimator estimates the total over the blocking stratum. That is,

$$\hat{Z} = \frac{\sum_{(i,j)\in U^*} z_{ij} P\left(M_{ij}=1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}; \hat{\boldsymbol{\psi}}\right)}{\sum_{(i,j)\in s^*} \pi_{ij}^{-1} z_{ij} P\left(M_{ij}=1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}; \hat{\boldsymbol{\psi}}\right)} \sum_{(i,j)\in s^*} \pi_{ij}^{-1} z_{ij} M_{ij} + \sum_{(i,j)\in s\setminus s^*} \pi_{ij}^{-1} M_{ij} z_{ij} \tag{9}$$

In this case $\hat{\beta}_0 = 0$ and $\hat{\beta}_1$ is computed as follows:

$$\hat{\beta}_1 = \frac{\sum_{(i,j)\in U^*} z_{ij} P\left(M_{ij}=1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}; \hat{\boldsymbol{\psi}}\right)}{\sum_{(i,j)\in s^*} \pi_{ij}^{-1} z_{ij} P\left(M_{ij}=1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}; \hat{\boldsymbol{\psi}}\right)} \tag{10}$$

The estimator can also be written in terms of uniform g-weights $\left[g_{ij}\right]_{ij}$, where $g_{ij} = \hat{\beta}_1$:

$$\hat{Z} = \sum_{(i,j)\in s^*} g_{ij} \pi_{ij}^{-1} z_{ij} M_{ij} + \sum_{(i,j)\in s\setminus s^*} \pi_{ij}^{-1} M_{ij} z_{ij} \tag{11}$$

The following model provides the basis for better weighted least squares estimators:

$$E\left[M_{ij} \middle| x_i, y_j, \gamma_{ij}\right] = \beta_0 + \beta_1 P\left(M_{ij}=1 \middle| x_i, y_j, \gamma_{ij}; \hat{\boldsymbol{\psi}}\right) \tag{12}$$

$$var\left(M_{ij} \middle| x_i, y_j, \gamma_{ij}\right) \propto P\left(M_{ij}=1 \middle| z_{ij}, \gamma_{ij}; \hat{\boldsymbol{\psi}}\right)\left[1 - P\left(M_{ij}=1 \middle| x_i, y_j, \gamma_{ij}; \hat{\boldsymbol{\psi}}\right)\right] \tag{13}$$

In this case, the estimated regression coefficients minimize the following quadratic function:

$$Q\left(\beta_0, \beta_1; \hat{\boldsymbol{\psi}}\right) = \sum_{(i,j)\in s^*} \frac{\pi_{ij}^{-1}\left[M_{ij} - \beta_0 + \beta_1 P\left(M_{ij}=1 \middle| x_i, y_j, \gamma_{ij}; \hat{\boldsymbol{\psi}}\right)\right]^2}{P\left(M_{ij}=1 \middle| x_i, y_j, \gamma_{ij}; \hat{\boldsymbol{\psi}}\right)\left[1 - P\left(M_{ij}=1 \middle| x_i, y_j, \gamma_{ij}; \hat{\boldsymbol{\psi}}\right)\right]} \tag{14}$$

The resulting estimator may be written in terms of nonuniform g-weights incorporating the inferred match status. This estimator is improved by fine tuning the variance structure with Generalized Estimating Equations (Jiang 2007).

The proposed estimators are no longer unbiased because the clerical review results are used to make inferences about the pairs match status. However, like other regression estimators (Särndal et al. 1992, Result 6.6.1, pp. 235, section, 6.7, pp. 238), they are design-consistent regardless of the assumed models.

4. Sampling Design

Model-based stratified sampling has been used to approximately minimize the variance of regression estimators (Särndal et al. 1992). In this design, the strata are defined by the variance of the error in the assumed linear model. This strategy also applies to the current

context where a single total is estimated. To be specific, the design-based variance $var\left(\hat{Z}|U\right)$ of the model-assisted estimator is the sum of two terms:

$$\begin{aligned} var(\hat{Z}|U) \;=\; & var\left(\sum_{(i,j)\in s^*} \pi_{ij}^{-1} z_{ij}\left(M_{ij} - \widehat{M}_{ij}\right)\middle| U\right) \\ & + var\left(\sum_{(i,j)\in s\setminus s^*} \pi_{ij}^{-1} M_{ij} z_{ij}\middle| U\right) \end{aligned} \tag{15}$$

The first term is approximately minimized by a Neyman allocation where the pairs are stratified according to the model-based conditional variance of the error $e_{ij} = z_{ij}\left(M_{ij} - \widehat{M}_{ij}\right)$, that is $var\left(e_{ij}\middle|x_i, y_j, \boldsymbol{\gamma}_{ij}\right)$. This conditional variance is given by the following expression.

$$\begin{aligned} var\left(e_{ij}\middle|x_i, y_j, \boldsymbol{\gamma}_{ij}\right) \;&=\; var\left(z_{ij}\left(M_{ij} - \widehat{M}_{ij}\right)\middle|x_i, y_j, \boldsymbol{\gamma}_{ij}\right) \\ &=\; z_{ij}^2 var\left(M_{ij} - \widehat{M}_{ij}\middle|x_i, y_j, \boldsymbol{\gamma}_{ij}\right) \\ &=\; z_{ij}^2\Big(var\left(M_{ij}\middle|x_i, y_j, \boldsymbol{\gamma}_{ij}\right) \\ &\quad + \left[\widehat{M}_{ij} - P\left(M_{ij} = 1\middle|x_i, y_j, \boldsymbol{\gamma}_{ij}\right)\right]^2\Big) \\ &=\; z_{ij}^2\Big(P\left(M_{ij} = 1\middle|x_i, y_j, \boldsymbol{\gamma}_{ij}\right)\left[1 - P\left(M_{ij} = 1\middle|x_i, y_j, \boldsymbol{\gamma}_{ij}\right)\right] \\ &\quad + \left[\widehat{M}_{ij} - P\left(M_{ij} = 1\middle|x_i, y_j, \boldsymbol{\gamma}_{ij}\right)\right]^2\Big) \end{aligned} \tag{16}$$

With known conditional match probabilities $P\left(M_{ij} = 1\middle|x_i, y_j, \boldsymbol{\gamma}_{ij}\right)$ and the best possible inference $\widehat{M}_{ij} = P\left(M_{ij} = 1\middle|x_i, y_j, \boldsymbol{\gamma}_{ij}\right)$ we have

$$var\left(e_{ij}\middle|x_i, y_j, \boldsymbol{\gamma}_{ij}\right) \;=\; z_{ij}^2 \widehat{M}_{ij}\left(1 - \widehat{M}_{ij}\right) \tag{17}$$

Suppose that the pairs are stratified based on $\boldsymbol{\gamma}_{ij}$ and $\left(x_i, y_j\right)$ or some fine discrete approximation if these variables are continuous. Note that by design, in such as stratum, the pairs have the same $z_{ij} = z$ value and an identical conditional match probability $P\left(M_{ij} = 1\middle|x_i, y_j, \boldsymbol{\gamma}_{ij}\right) = p$. Thus they are identically distributed according to $zBernoulli\left(p\right)$. If these pairs were independent, the variance of the errors e_{ij} would be well approximated by the common variance $var\left(e_{ij}\middle|x_i, y_j, \boldsymbol{\gamma}_{ij}\right) = z^2 p\left(1 - p\right)$, based on the Law of Large Numbers (LLN). In the corresponding Neyman allocation, the sample size is proportional to the stratum variance. An estimator with the same minimum variance is obtained via a Neyman allocation, where the strata are based on $z_{ij}^2 \widehat{M}_{ij}\left(1 - \widehat{M}_{ij}\right)$ the estimated conditional error variance. The resulting allocation is no longer optimal when the conditional match probability $P\left(M_{ij} = 1\middle|x_i, y_j, \boldsymbol{\gamma}_{ij}\right)$ is estimated with some bias. Let $\hat{p}$ denote the corresponding stratum estimate. In this case the stratum variance is increased to $z^2 p\left(1 - p\right) + \left(\hat{p} - p\right)^2$.

5. Simulations

The proposed estimators are evaluated in six scenarios that consider different factors, including the discriminating power of the linkage variables, the sample size, the model for the distribution of linkage errors, clerical errors, and the correlation among the pairs. All the scenarios consider a one-to-one linkage between two registers. In each register the records are partitioned into perfect blocks of equal sizes. Consequently two matched records always fall within the same block.

The different scenarios account for different features of practical linkages.

Scenario 1 emulates the process by which administrative records may be generated from a finite population of individuals, with correlations among pairs that are within the same block. It considers seven binary linkage variables that have conditionally independent typographical errors with a common distribution. This distribution is given by the following transition probabilities:

$$P\left(c_i^{(k)}, c_j^{(k)} \middle| \zeta_i^{(k)}, M_{ij} = 1\right) = P\left(c_i^{(k)} \middle| \zeta_i^{(k)}\right) P\left(c_j^{(k)} \middle| \zeta_i^{(k)}\right) \tag{18}$$

$$P\left(c_i^{(k)} \middle| \zeta_i^{(k)}\right) = (1-\alpha) I\left(c_i^{(k)} = \zeta_i^{(k)}\right) + \alpha I\left(c_i^{(k)} \neq \zeta_i^{(k)}\right) \tag{19}$$

where α is the probability of a recording error.

In the above expressions, $c_i^{(k)}$ is the k-th linkage variable for record i in register A, $\zeta_i^{(k)}$ is the latent true (i.e., free of recording errors) value of the variable for the associated individual, with $c_j^{(k)}$ and $\zeta_j^{(k)}$ denoting the corresponding variables in register B. Note that, by definition $\zeta_i^{(k)} = \zeta_j^{(k)}$ in a matched pair (i,j). For each record i, the latent variables $\zeta_i^{(k)}$ are IID. The comparison outcomes are based on exact comparisons with $\gamma_{ij}^{(k)} = I\left(c_i^{(k)} = c_j^{(k)}\right)$.

The variables of interest x_i and y_j are also binary and mutually independent of the linkage variables in each register, and each matched pair. The files are linked to study the joint distribution of these two variables, that is, to estimate the frequencies of the different cells in a two-way contingency table. In this case $z_{ij} = I\left(\left(x_i, y_j\right) = \left(x, y\right)\right)$ where $x, y = 0, 1$. This setup is similar to that described by Chipperfield et al. (2011). However, the goal here is finite population inference on a single finite population.

From the finite population, different IID samples are drawn using one of two designs. For each resulting sample, three estimators are computed for the number of matched pairs in each cell of the two-way contingency table. They include the H-T estimator, a second model-assisted estimator (hereafter simply referred to as 2nd estimator) using the inference $\widehat{\boldsymbol{M}}_{ij} = P\left(M_{ij} = 1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}; \hat{\boldsymbol{\psi}}\right)$ and a third estimator (hereafter simply referred to as 3rd estimator) using the inference $\widehat{\boldsymbol{M}}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 P\left(M_{ij} = 1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}; \hat{\boldsymbol{\psi}}\right)$.

The first sample design is stratified according to the x-y value pairs. In each stratum, a fixed size SRS sample is drawn. The second sample design is also stratified based on the x-y value pairs, but it uses substrata, which are based on the conditional variance of the prediction error. Each x-y stratum has the same number of substrata but the boundaries are selected to obtain nearly equal substrata sizes, after the pairs are sorted according to their conditional variance in each stratum. Consequently, substrata boundaries may differ

from an x-y stratum to the next. The same x-y stratum sample size is used as in the first design. However in the second sample design, this sample size is allocated optimally among the substrata using a Neyman allocation, where the estimated variance of a substratum is estimated as the mean conditional error variance over all the corresponding pairs. A substratum sample allocation is further constrained to have at least two units and not to exceed the substratum size.

Scenario 1 is the baseline scenario. It evaluates the two model-assisted estimators in the best case, with the correct model for the comparison outcomes. This situation maximizes their relative advantage over the naïve H-T estimator. Scenarios 2 through 5 are built after Scenario 1, that is, with correlated pairs. However they each incorporate a slight modification. Scenario 2 considers linkage variables with more typographical errors and hence less discriminating power than in Scenario 1. Scenario 3 considers a smaller (1,000 pairs instead of 4,000 pairs) clerical-review sample. Scenario 4 considers linkage variables that are not conditionally independent by correlating the latent variables $\zeta_i^{(k)}$. This correlation is produced by generating the $\zeta_i^{(k)}$'s according to a mixture model with conditional independence based on a binary latent class ξ_i. However the estimated conditional match probability $P\left(M_{ij} = 1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}; \hat{\boldsymbol{\psi}}\right)$ is estimated under the assumption of conditional independence among all linkage variables. Scenario 5 considers clerical errors.

Scenario 6 considers agreement frequencies for variables such as names and birthdate that have been used for linking high quality person files. The specific frequencies are based on an example provided by Newcombe (1988, Table 5.1). Unlike the other scenarios, Scenario 6 considers pairs with IID and conditionally independent comparison vectors.

The simulation parameters are as follows. All scenarios are based on $N = 10,000$ individuals, 1,000 blocks, a block size of 10, $K = 7$ linkage variables, $P(x = 1) = 0.5$, $P\left(y = 1|x = 0\right) = 0.4$, $P\left(y = 1|x = 1\right) = 0.7$, 10 substrata per x-y stratum, 100 E-M iteration and 100 repetitions.

The x-y stratum sample size is set to 1,000 for all scenarios except for Scenario 3 (smaller clerical sample), where it is set to 100. The conditional agreement probabilities are uniform across the linkage variables in Scenarios 1 through 5. However, they vary across these scenarios. For Scenarios 1 and 3 through 5, the conditional probability of agreement is 0.98 for a matched pair and 0.5 for an unmatched pair. For Scenario 2, these conditional probabilities are respectively 0.82 and 0.5. For Scenario 6, the conditional agreement probabilities are given in Table 1. The remaining parameters only apply to Scenarios 1 through 5 and are set as follows. The parameter α is set to 0.1 for Scenarios 1 through 5. For the intrinsic variables, the probability $P\left(\zeta_i^{(k)} = 1\right)$ is uniformly set to 0.5. For the recording errors, the probability $P\left(c_i^{(k)} = 1 \middle| \zeta_i^{(k)} = 0\right)$ is set to 0.01 except for Scenario 2 (weaker linkage variables), where it is set to 0.1. As for the probability $P\left(c_i^{(k)} = 1 \middle| \zeta_i^{(k)} = 1\right)$ is set to 0.99 except for Scenario 2, where it is set to 0.9. Scenario 4 (misspecified case) involves the additional parameters that are set as follows. The probability $P(\xi_i = 1)$ is set to 0.5, while the conditional probabilities $P\left(\zeta_i^{(k)} = 1|\xi_i = 0\right)$ and $P\left(\zeta_i^{(k)} = 1|\xi_i = 1\right)$ are respectively set to 0.3 and 0.7.

For cell (0,0), the results for the H-T estimator and the second estimator are shown in the box plots of Figures 1 and 2, and in Tables 2 and 3. The box plots show the five-number summary of the relative bias for the estimated cell count. In these figures, the horizontal axis

Table 1. Agreement frequencies for Scenario 6 based on Newcombe (1988, Table 5.1).

	Agreement probability	
Linkage variable	Matched	Unmatched
Surname	0.965	0.001
First name	0.79	0.009
Middle initial	0.888	0.075
Year of birth	0.773	0.011
Month of birth	0.933	0.083
Day of birth	0.851	0.033
Province/country of birth	0.981	0.117

indicates the estimator (1 for the H-T estimator, or 2 for the second estimator), the sampling design (1 or 2) and the scenario (1 through 3 in Figure 1, and 4 through 6 in Figure 2). For example, in Figure 1, 2.1.1 corresponds to the box plot for the second estimator under the first scenario and the first design. As for Tables 2 and 3 they show the average bias and CV of the estimated count for cell (0,0). The results for the other cells are not shown because they are similar to those of cell (0,0). As for the third estimator, the corresponding results are not shown because they are similar to those of the second estimator.

For Scenario 1 (our baseline), all three estimators have a very small relative bias, with no clear advantage for the H-T estimator under either sampling design. However the model-assisted estimator halves the CV of the H-T estimator, under the first sampling design. The gain in precision becomes negligible under the second sampling design. This is expected because the model information is already exploited through the stratification, which also benefits the H-T estimator.

The results for Scenario 2 show a worse performance for the model-assisted estimator, when the linkage variables are less discriminating. Indeed, the corresponding absolute relative bias is larger than that of the H-T estimator, under either sampling design. As for

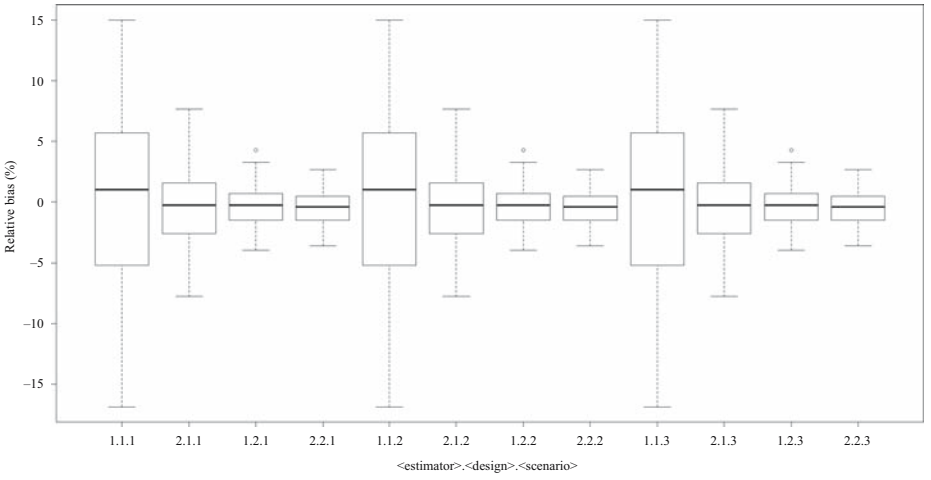

Fig. 1. Box plots of the relative bias for cell (0,0) in Scenarios 1 through 3. Estimator 1 is the HT estimator. Estimator 2 is the 2nd estimator.

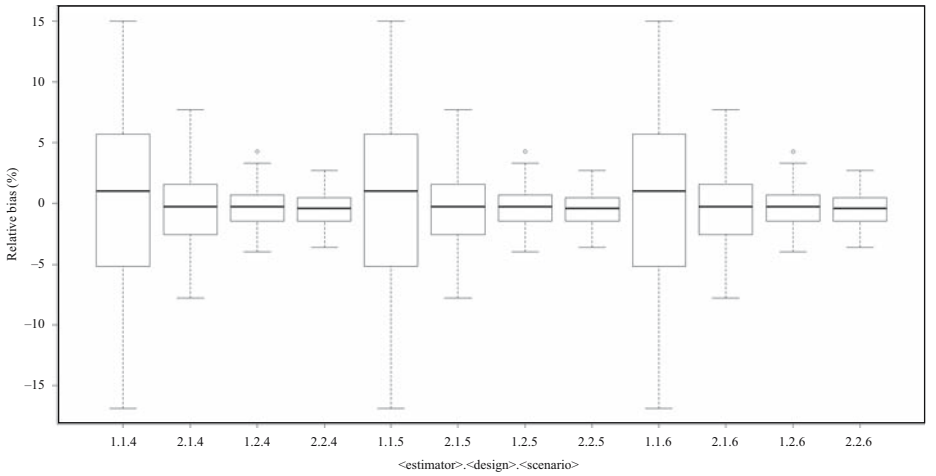

Fig. 2. Box plots of the relative bias for cell (0,0) in Scenarios 4 through 6. Estimator 1 is the HT estimator. Estimator 2 is the 2nd estimator.

the expected gain in precision under the first sampling design, it is dramatically smaller than in Scenario 1. Under the second design, the gain is negligible.

The results for Scenario 3 show the same trends as in Scenario 1, with similar gains in precision for the model-assisted estimator. Intuitively the use of a model partially makes up for the reduced sample size.

For Scenario 4, where the model is misspecified, both the H-T estimator and the model-assisted estimator have a small relative bias, under either design. For the model-assisted estimator, the gain in precision is slightly reduced compared to Scenario 1.

In Scenario 5, with clerical errors, Table 2 shows that the relative bias of all the estimators is significantly increased compared to Scenario 1. However, under the first sampling design, the model-assisted estimators offer a significant advantage over the HT estimator, even if this advantage is smaller than in Scenario 1. Under the second design, this gain in precision vanishes and all the estimators have much less precision than in the first sampling design.

Table 2. Relative bias and CV for cell (0,0) for Scenarios 1 through 3.

Scenario	Design	Estimator	Relative bias (%)	CV (%)
1	1	1	−0.12	7.52
		2	0.45	3.33
	2	1	0.34	1.52
		2	0.48	1.36
2	1	1	0.77	7.62
		2	0.94	6.43
	2	1	−0.17	5.67
		2	−0.29	5.44
3	1	1	0.18	25.18
		2	0.11	12.57
	2	1	0.32	6.79
		2	−0.04	6.37

Table 3. Relative bias and CV for cell (0,0) for Scenarios 4 through 6.

Scenario	Design	Estimator	Relative bias (%)	CV (%)
4	1	1	1.21	7.71
		2	0.62	4.22
	2	1	0.25	2.40
		2	0.21	2.29
5	1	1	−4.94	8.25
		2	−5.25	3.66
	2	1	−6.31	14.84
		2	−6.23	14.79
6	1	1	−0.79	7.40
		2	−0.10	0.48
	2	1	−0.01	0.82
		2	0.01	0.12

In Scenario 6, the model-assisted estimator greatly outperforms the H-T estimator both regarding the bias and the precision, under either sampling design. The gain in precision is also dramatically larger than in the other scenarios. This is because in Scenario 6, the linkage variables collectively provide much more discrimination than in the previous scenarios. The combination of this discrimination with a correct statistical model produces the observed gains.

Overall, the model-assisted estimators offer the best performance when the following three conditions are met:

i. The linkage variables provide a high discrimination.
ii. The clerical-reviews are very reliable.
iii. The assumed statistical model is correct.

Of the above three conditions, the reliability of the clerical-review is the most critical one as it may be expected.

The simulation results also shed some light on the choice of the sampling design. In all scenarios without clerical errors, the precision is much greater under the second sampling design, where the pairs are stratified according to the estimated conditional match probability. This result further underscores the importance of using auxiliary variables that leverage the comparison outcomes.

Although this work considers a one-to-one linkage, this assumption does not play a major role in the estimation procedure. Hence the proposed methodology also applies to an incomplete linkage so long as the clerical reviews remain error-free. However the resulting model-assisted estimators may be less efficient if the unmatchable records greatly differ in distribution from the other records. Then the pairs outcomes are better modeled by a three component mixture including two classes of unmatched pairs. In this case, specifying a good model may be more challenging.

6. Conclusions and Future Work

This study casts the problem of design-based estimation with linked administrative files in the classical survey methodology framework. It also proposes a new estimation

methodology based on model-assisted estimators and sampling-designs that are evaluated through simulations. The simulations clearly demonstrate the equal importance of auxiliary variables based on the linking variables and high quality clerical reviews. Specifying good models is also important for the efficiency of the resulting estimators. However using the correct model is not required, because, like previous model-assisted estimators (Särndal et al. 1992), the proposed estimators remain design consistent even when the model is misspecified.

There are two potential issues with clerical reviews including the quality of the supporting information and the quality of the review process. Meaningful clerical reviews are obviously impossible unless the supporting information is sufficient and reliable. Even when it is the case, many questions remain about the quality of the review process and ways to objectively measure it. Indeed there are few studies on this subject, beyond that by Newcombe et al. (1983). Furthermore, such studies may be hard to replicate, either because they have not disclosed important methodological details, or because their results are heavily dependent on the used datasets that are unavailable. A second challenge is the development of anonymization techniques. They prevent clerical reviews and adversely impact the linking efficacy. Solutions based on privacy-preserving record linkage are being actively researched to address these problems (Schnell et al. 2009). However, in situations where clerical reviews have been effective (e.g., with available names, birthdates and addresses in the original files), it is still unclear whether these solutions offer competitive privacy-preserving alternatives to clerical reviews. A third challenge concerns missing values in the linked files. The problem arises because clerical reviews are expensive, such that it is desirable to avoid sampling pairs where some variables of interest are missing. Such missing variables represent an unusual form of item nonresponse, because it is known prior to sample selection. Devising solutions for an optimal sample selection represents a new and promising avenue of research.

7. References

Belin, T.R. and D.B. Rubin. 1995. "A Method for Calibrating False-Match Rates in Record Linkage." *Journal of the American Statistical Association* 90: 694–707. Doi: http://dx.doi.org/10.2307/2291082.

Chambers, R. 2009. "Regression Analysis of Probability-Linked Data." *Official Statistics Research Series*, vol. 4.

Chipperfield, J.O., G.R. Bishop, and P. Campbell. 2011. "Maximum Likelihood Estimation for Contingency Tables and Logistic Regression with Incorrectly Linked Data." *Survey Methodology* 37: 13–24.

Dempster, A., N. Laird, and D. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society Series B* 39: 1–38. Available at: http://www.jstor.org/stable/2984875 (accessed November 2017).

Deville, J.-C. and C.-E. Särndal. 1992. "Calibration Estimators in Survey Sampling". *Journal of the American Statistical Association* 37: 376–382.

Fellegi, I.P. and A.B. Sunter. 1969. "A Theory of Record Linkage." *Journal of the American Statistical Association* 64: 1183–1210.

Fienberg, S., P. Hersh, A. Rinaldo, and Y. Zhou. 2009. "Maximum Likelihood in Latent Class Models for Contingency Table Data." In *Algebraic and Geometric Methods in Statistics*, edited by Paolo Giblisco, Eva Riccomagno, Maria Piera Rogantin, and Henry P. Wynn, 7–62. New York: Cambridge University Press.

Gill, L. 2001. *Methods for Automatic Record Matching and Linkage and their Use in National Statistics*. London: Office of National Statistics.

Guiver, T. 2011. *Sampling-Based Clerical Review Methods in Probabilistic Linking*. Canberra: Australia Bureau of Statistics.

Heasman, D. 2014. "Sampling a Matching Project to Establish the Linking Quality." *Survey Methodology Bulletin* 72: 1–16.

Howe, G.R. 1981. "A Generalized Iterative Record Linkage Computer System for Use in Medical Follow-Up Studies." *Computers and Biomedical Research* 14: 327–340.

Jaro, M.A. 1989. "Advances in Record Linkage Methodology to Matching the 1985 Census of Tampa, Florida." *Journal of the American Statistical Association* 84: 414–420.

Jiang, J. 2007. *Linear and Generalized Linear Mixed Models and Their Applications*. New York: Springer.

Kim, B.S. 1984. *Studies of Multinomial Mixture Models*. PhD thesis, Chapel Hill: University of North Carolina.

Lahiri, P. and D. Larsen. 2005. "Regression Analysis with Linked Data." *Journal of the American Statistical Association* 100: 222–227. Available at: http://www.jstor.org/stable/27590532 (accessed November 14, 2017).

Larsen, M. and D. Rubin. 2001. "Iterated Automated Record Linkage Using Mixture Models." *Journal of the American Statistical Association* 96: 32–41.

Lavallée, P. 2002. *Le Sondage indirect ou la méthode du partage des poids*. Bruxelles: Éditions de l'Université de Bruxelles.

Newcombe, H.B., M.E. Smith, and G.R. Howe. 1983. "Reliability of Computerized Versus Manual Death Searches in a Study of the Health of Eldorado Uranium Workers." *Computers in Biology and Medicine* 13: 157–169.

Newcombe, H. 1988. *Handbook of Record Linkage*. New-York: Oxford Medical Publications.

Särndal, C.-E., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. New-York: Springer.

Schnell, R., T. Bachteler, and J. Reiher. 2009. "Privacy-Preserving Record Linkage using Bloom Filters". *BioMed Central Medical Informatics and Decision Making*, 9.

Winkler, W.E. 1988. "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage". In Proceedings of the Section on Survey Research Methods: American Statistical Association, August 22–25, 1988, New Orleans, Louisiana. 667–671.

Received July 2015
Revised October 2017
Accepted October 2017