Proceedings of Statistics Canada Symposium 2014
Beyond traditional survey taking: adapting to a changing world

Design-based Estimation with Record-Linked Administrative Files

Abel Dasylva[1]

Abstract

Exact record linkage is an essential tool for exploiting administrative files, especially when one is studying the relationships among many variables that are not contained in a single administrative file. It is aimed at identifying pairs of records associated with the same individual or entity. The result is a linked file that may be used to estimate population parameters including totals and ratios. Unfortunately, the linkage process is complex and error-prone because it usually relies on linkage variables that are non-unique and recorded with errors. As a result, the linked file contains linkage errors, including bad links between unrelated records, and missing links between related records. These errors may lead to biased estimators when they are ignored in the estimation process. Previous work in this area has accounted for these errors using assumptions about their distribution. In general, the assumed distribution is in fact a very coarse approximation of the true distribution because the linkage process is inherently complex. Consequently, the resulting estimators may be subject to bias. A new methodological framework, grounded in traditional survey sampling, is proposed for obtaining design-based estimators from linked administrative files. It consists of three steps. First, a probabilistic sample of record-pairs is selected. Second, a manual review is carried out for all sampled pairs. Finally, design-based estimators are computed based on the review results. This methodology leads to estimators with a design-based sampling error, even when the process is solely based on two administrative files. It departs from the previous work that is model-based, and provides more robust estimators. This result is achieved by placing manual reviews at the center of the estimation process. Effectively using manual reviews is crucial because they are a de-facto gold-standard regarding the quality of linkage decisions. The proposed framework may also be applied when estimating from linked administrative and survey data.

Key Words: record-linkage, survey, administrative file, design, estimator, model, auxiliary variable, clerical review, mixture model, Expectation-Maximization (E-M) algorithm, model-assisted.

1. Introduction

1.1 The record-linkage problem

The goal of exact record-linkage, hereafter simply called record-linkage, is linking records that are associated with the same individual or entity. Such records are called *matched* records. Record-linkage is straightforward when the records contain a unique identifier. Without a unique identifier, record-linkage must be based on pseudo-identifiers, which are non-unique and often recorded with variations in different files. For example, in person files, pseudo-identifiers may include the family name, the given name and the birth date. Variations occur due to typographical errors or differences in format for variables that must otherwise record the same information. In this context, record-linkage is prone to errors, including bad links between unrelated records and missing links between related records that accidentally look too different. The record-linkage challenge is essentially the problem of testing the null hypothesis that two given records are related based on the available information. From that view point, missing links are type I errors, while bad links are type II errors. Both types of errors degrade the quality of a linkage and must be controlled based on the observed differences in each record-pair. For small files, this information is easily processed by experienced clerks. However with large files, this solution is too costly for the millions of pairs that may be generated. Instead a computerized solution is required where the linkage decision is automated for most pairs.

[1]Abel Dasylva, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, ON, K1A0T6, Canada (abel.dasylva@statcan.gc.ca);

1.2 Probabilistic record-linkage

Newcombe (1967), and Fellegi and Sunter (1969) proposed probabilistic record-linkage including a theoretical basis for computerized solutions, where clerical reviews are minimized and linkage errors are controlled. In the probabilistic approach, each record-pair is assigned a linkage weight that increases when the records become more similar. The weight is compared to thresholds to determine whether a pair should be linked, reviewed clerically or rejected (Fellegi and Sunter, 1969). The overall linkage performance is determined by the linkage weights and the thresholds.

1.3 Fully automated solutions and their limitations

A probabilistic linkage solution may be fully automated or semi-automated. Newcombe et al. (1983) advocated automated decisions based on their experience with high-quality person files, including names, birth dates or addresses. However the superiority of fully automated solutions has been called into question by the inaccuracy of the estimated error rates and the need to back them up with clerical-reviews (Heasman, 2014) or training samples (Belin and Rubin, 1995). The problem is related to the identifiability of general multinomial mixtures. Kim (1984) has given sufficient conditions for the identifiability of such mixtures. This result applies to record-linkage only in the unrealistic situation where the pairs are partitioned into known clusters that each contains pairs with the same match status. However such a clustering structure does not exist in a practical linkage. Therefore modeling the pairs by a multinomial mixture with interactions may involve untestable assumptions. Such assumptions are also made to deal with nonignorable nonresponse in sample surveys, when the nonresponse affects few units. However, the challenge is greater in record-linkage because any model misspecification affects all the pairs.

1.4 Semi-automated solutions

Most practical solutions are semi-automated and ultimately rely on clerical reviews or training samples. In these solutions, a small sample of pairs is reviewed clerically while all remaining pairs are automatically linked according to parameters. The parameters may be estimated from the same sample. The Fellegi-Sunter optimal decision rule (Fellegi and Sunter, 1969) is a good example of a semi-automated solution because it requires clerical reviews in the *grey zone*, between the thresholds. G-LINK (previously called GRLS), Statistics Canada generalized record linkage system, provides another example, where linkage weights and thresholds are estimated in a manual and iterative manner (Howe, 1981). Both Gill (2001) and Guiver (2011) have recommended using clerical reviews for setting the weight thresholds. Guiver (2011) also notes that *spot-checking*, a common procedure for setting the thresholds, does not have a sound statistical basis. He instead recommends the review of a probabilistic sample of pairs. Larsen and Rubin (2001) have described another iterative solution incorporating clerical reviews to assist with the selection of a model for the pairs. Belin and Rubin (1995) have also relied on a training sample for estimating the rate of false matches.

1.5 Design-based estimation with linked data

This study looks at the related issues of design-consistent estimation and sampling with clerical samples. It includes using auxiliary information based on comparison outcomes and models of their distribution in pairs. Särndal et al. (1992) have already addressed the general problem of efficient and design-consistent estimation with auxiliary variables, models and related estimators. They have proposed regression estimators that are design-consistent and make the best possible use of the available auxiliary data if the assumed model holds. Applying these results to the record-linkage setting has required some adaptation to take advantage of new possibilities. The resulting estimators are regression estimators, which incorporate the auxiliary variables through a nonlinear function. They are built in two parts. In the first part, 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. In the second part, a regression estimator is computed and the potential inaccuracies of the model are corrected based on the clerical data. The described framework also applies when the match status is determined by other means than clerical reviews, e.g. through limited access to unique identifiers or additional information from a third party.

The following sections are organized as follows. Section 2 presents the model and notation. 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. Model and notation

2.1 Registers and finite population of record-pairs

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 i-th record in A. Register B contains the same linkage variables as A and the variable of interest y_j for the j-th record in B. Let U denote the finite population of all N^2 record-pairs in the cartesian product of the two files, i.e. of all pairs (i,j) where $1 \leq i,j \leq N$.

2.2 Comparison outcomes and blocks

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 γ_{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.

2.3 Pair match status

Let M_{ij} denote the indicator variable that is set to 1 if the pair (i,j) is matched, i.e. 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.

2.4 Inference problems

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

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

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

Inferences about a superpopulation assume an Independent Identical Distribution (IID) of the variables of interest x_i and y_j in matched pairs. They target a superpopulation parameter θ which satisfies a score equation of the form $E\left[S\left(\theta; x_i, y_j\right)\right] = 0$, where S is a score function (e.g. a log-likelihood) and the expectation is taken with respect to the superpopulation. The parameter θ may be estimated through the following unbiased estimating equation where $z_{ij}(\theta) = S\left(\theta; x_i, y_j\right)$.

$$\sum_{(i,j) \in U} M_{ij} z_{ij}\left(\hat{\theta}\right) \quad = \quad 0$$

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.

2.5 The clerical sample

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 $U \backslash U^*$. 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.

3. Model-assisted estimators

3.1 A general form

The proposed estimators have the following general difference form:

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

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

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

This the case if $\widehat{M}_{ij}$ is only a function of z_{ij} and γ_{ij}.

3.2 Inferring the match status

The inferred status may be set to the conditional match probability given the vector of comparison outcomes and the variables x_i, y_j, i.e.

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

This particular inference strategy would minimize the mean squared error (over the superpopulation) 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 $\tau = 0,1$:

$$P\left(x_i, y_j, \gamma_{ij} \middle| M_{ij} = \tau\right) \quad = \quad P\left(x_i, y_j \middle| M_{ij} = \tau\right) \prod_{k=1}^{K} P\left(\gamma_{ij}^{(k)} \middle| M_{ij} = \tau\right)$$

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 $P\left(\gamma_{ij}^{(k)} \middle| M_{ij} = 1\right)$, and the marginal u-probabilities $P\left(x_i, y_j \middle| M_{ij} = 0\right)$ and $P\left(\gamma_{ij}^{(k)} \middle| M_{ij} = 0\right)$, 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:

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

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 $\widehat{M}_{ij} =$

$P(M_{ij}=1|\boldsymbol{\gamma}_{ij})$ is used instead. The estimator is also less efficient if the linkage variables are correlated but their conditional independence is assumed.

Let $P(M_{ij}=1|x_i,y_j,\boldsymbol{\gamma}_{ij};\widehat{\boldsymbol{\psi}})$ 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 correct this bias, the match status may be inferred using a linear function $\beta_0+\beta_1 P(M_{ij}=1|x_i,y_j,\boldsymbol{\gamma}_{ij};\widehat{\boldsymbol{\psi}})$ 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:

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

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

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

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

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

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

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

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

$$E[M_{ij}|x_i,y_j,\gamma_{ij}] \quad = \quad \beta_0+\beta_1 P(M_{ij}=1|x_i,y_j,\gamma_{ij};\widehat{\boldsymbol{\psi}})$$

$$var(M_{ij}|x_i,y_j,\gamma_{ij}) \quad \propto \quad P(M_{ij}=1|z_{ij},\gamma_{ij};\widehat{\boldsymbol{\psi}})[1-P(M_{ij}=1|x_i,y_j,\gamma_{ij};\widehat{\boldsymbol{\psi}})]$$

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

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

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 classical regression estimators (Särndal et al., 1992), 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 $var(\hat{Z}|U)$ of the model-assisted estimator is the sum of two terms:

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

The first term may be approximately minimized by a Neyman allocation where the pairs are stratified according to the model-based conditional variance $var\left(z_{ij}(M_{ij} - \widehat{M}_{ij})\middle| x_i, y_j, \boldsymbol{\gamma}_{ij}\right)$. Indeed, suppose known conditional match probabilities given by $P\left(M_{ij} = 1\middle| x_i, y_j, \boldsymbol{\gamma}_{ij}\right)$ and the best possible inference $\widehat{M}_{ij} = P\left(M_{ij} = 1\middle| x_i, y_j, \boldsymbol{\gamma}_{ij}\right)$ for each blocked pair. Suppose that the pairs are stratified based on (x_i, y_j), or some accurate discrete approximation of them, and $\boldsymbol{\gamma}_{ij}$. The variance in the stratum of $\left(x_i, y_j, \boldsymbol{\gamma}_{ij}\right)$ would be given by $z_{ij}^2 \widehat{M}_{ij}(1 - \widehat{M}_{ij})$, if the pairs were IID. 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 the estimate $z_{ij}^2 \widehat{M}_{ij}(1 - \widehat{M}_{ij})$ of the conditional error variance.

5. Simulations

Two simulation scenarios were used to evaluate the different estimators and measure their sensitivity to the underlying assumptions. Both scenarios deal with a one-to-one linkage between two registers. In each register the records are partitioned into perfect blocks of equal size. Consequently two matched records always fall within the same block.

5.1 First scenario

In the first scenario, the linkage variables are binary, IID and with the same distribution of typographical errors. This distribution is given by the following transition probabilities:

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

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

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, i.e. to estimate the frenquencies of the different cells in a two-way contingency table. In this case $z_{ij} = I\left(\left(x_i, y_j\right) = (x, y)\right)$ 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 $\widehat{M}_{ij} = P\left(M_{ij} = 1\middle| x_i, y_j, \boldsymbol{\gamma}_{ij}; \widehat{\boldsymbol{\psi}}\right)$ and a third estimator (hereafter simply referred to as 3rd estimator) using the inference $\widehat{M}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 P\left(M_{ij} = 1\middle| x_i, y_j, \boldsymbol{\gamma}_{ij}; \widehat{\boldsymbol{\psi}}\right)$.

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.

The first scenario is selected to evaluate the two model-assisted estimators in the best case, where the correct model is used for the distribution of outcomes. This situation should maximize their relative advantage over the naïve H-T estimator.

5.2 Second scenario

The second scenario is identical to the first scenario, except for the correlation of 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 $P\left(M_{ij}=1\middle|x_i,y_j,\boldsymbol{\gamma}_{ij};\widehat{\boldsymbol{\psi}}\right)$ is estimated under the assumption of conditional independence among all linkage variables.

The second scenario is selected to evaluate the two model-assisted estimators, when the distribution of outcomes is misspecified. This is a less favourable but more realistic situation. In this case, their relative advantage should be reduced when compared to the H-T estimator, which is expected to perform as in the first scenario.

5.3 Results

Simulation parameters other than those already described are given in Table 5-1. The results of the simulation are found in Table 5-2. In each x-y cell, the 1st column of numbers is $E\left[\left(1-\hat{Z}/Z\right)^2\middle|U\right]^{1/2}$, which is the square-root of the relative mean-squared error, respectively for the H-T estimator, the 2nd and 3rd estimators, for the first sample design. In the same x-y cell, the 2nd column contains the values of the same estimators under the second sample design.

Table 5-1
Simulation parameters

Parameter	***Value***
α	0.1
K	7
N	10,000
Number of blocks	1,000
Block size	10
$P\left(\zeta_i^{(k)}=1\right)$	0.5

Parameter	***value***
$P(\xi_i=1)^*$	0.5
$P\left(\zeta_i^{(k)}=1\middle\|\xi_i=0\right)^*$	0.3
$P\left(\zeta_i^{(k)}=1\middle\|\xi_i=1\right)^*$	0.7
$P(x=1)$	0.5
$P(y=1\|x=0)$	0.4
$P(y=1\|x=1)$	0.7

Parameter	***Value***
E-M iterations	100
x-y stratum sample size	1,000
Substrata per x-y stratum	10
Number of samples	100

* used in the 2nd scenario

In the first scenario, and the first sample design, the 3rd estimator is by far superior to the 2nd estimator, which is itself superior to the H-T estimator. The large performance difference between the two model-assisted estimators is surprising given the use of a correct parametric model for the pairs. With the second sample design, all the estimators perform equally well and much better than under the first design. This good performance is due to the interest for a single variable. Hence it is possible optimize the substrata sample allocation for this particular variable. It is harder to reap this benefit when many variables are of equal interest.

In the second scenario, the results show a similar trend except for two notable differences. First, the model-assisted estimators perform worse than in the first scenario, under the same designs. Second, the degradation is performance is greatest under the first sample design, for the 2nd estimator using the inference $\widehat{M}_{ij}=P\left(M_{ij}=1\middle|x_i,y_j,\boldsymbol{\gamma}_{ij};\widehat{\boldsymbol{\psi}}\right)$, where the assumed model is not adjusted based on the clerical sample. In this case, this estimator actually performs much worse than the H-T estimator. Under the second optimized sample design, the degradation in performance is barely noticeable for both model-assisted estimators. This second scenario also clearly establishes the 3rd estimator as a practical solution, because its performance remains good even when the mixture model is misspecified.

In both scenarios, the H-T performance is markedly better under the second sampling design than under the first sampling design. Indeed the performance goes from abysmal to stellar between the first and the second sample designs, including in the second scenario where the underlying model is misspecified. This stark difference further underscores the importance of using auxiliary variables that leverage the comparison outcomes.

The performance of the model-assisted estimators is quite sensitive to the estimated mixture proportion from the E-M algorithm. Unlike the m- and u- probabilities, this parameter converges quite slowly to its Maximum Likelihood Estimate, hence the need for a large number (100) of E-M iterations. Previous studies, which focused on the m- and u- probabilities, have overlooked this critical issue when the conditional match probability must be estimated. It is yet another manifestation of the linear E-M convergence rate that has been lamented in previous studies. The problem is resolved by using a Newton-Raphson numerical procedure, where the convergence rate is quadratic.

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.

A companion report provides a larger set of simulation results as well as further details about the supporting methodology (Dasylva, 2015).

Table 5-2
Square-root of the relative mean squared error (%) for the different estimators and scenarios

1st scenario (%)				
	Y=0		*Y=1*	
X=0	(91)	(5)	(96)	(7)
	(61)	(5)	(83)	(7)
	(14)	(5)	(24)	(7)
X=1	(95)	(7)	(96)	(6)
	(95)	(7)	(69)	(6)
	(32)	(7)	(20)	(6)

2nd scenario (%)				
	Y=0		*Y=1*	
X=0	(96)	(7)	(96)	(8)
	(92)	(7)	(148)	(8)
	(16)	(7)	(44)	(8)
X=1	(95)	(8)	(97)	(5)
	(147)	(7)	(94)	(5)
	(42)	(7)	(17)	(5)

6. Conclusions and future work

6.1 Importance of auxiliary variables, models and clerical reviews in estimation

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.

6.2 Remaining challenges

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.

References

Belin, T.R. and Rubin, D.B. (1995). "A Method for calibrating false-match rates in record linkage", *Journal of the American Statistical Assocation*, 90, pp. 694-707.

Chipperfield, J. O., Glenys R.B. and Campbell, P. (2011), "Maximum likelihood estimation for contingency tables and logistic regression with incorrectly linked data", *Survey Methodology*, 37, pp. 13-24.

Dasylva, A. (2015), "Design-based Estimation with Record-Linked Administrative Files", unpublished report, Canada: Statistics Canada.

Dempster, A., Laird, N., and Rubin, D. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm*", Journal of the Royal Statistical Society Series B*, 39, pp. 1-38.

Fellegi, I.P., and Sunter, A.B. (1969), "A Theory of Record Linkage", *JASA*, 64, pp. 1183-1210.

Gill, L. (2001), "Methods for Automatic Record Matching and Linkage and their Use in National Statistics", *National Statistics Methodological Series*, 25.

Guiver, T. (2011), "Sampling-Based Clerical Review Methods in Probabilistic Linking", unpublished report, Australia: Australia Bureau of Statistics.

Heasman, D. (2014), "Sampling a matching project to establish the linking quality", *Survey Methodology Bulletin*, Office of National Statistics, 72, pp. 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, pp. 327-340.

Jaro, M. A. (1989), "Advances in record linkage methodology to matching the 1985 census of Tampa, Florida", *JASA*, 84, pp. 414-420.

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

Kim, B.S. (1994), "Studies of multinomial mixture models", PhD thesis, University of North Carolina , Chapel Hill.

Larsen, M., and Rubin, D. (2001), "Iterated automated record linkage using mixture models", *JASA*, 96, pp. 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. (1967), "Record linking: the design of efficient systems for linking records into individual and family histories", *Amercian Journal of Human Genetics*, 19, no. 3, Part I.

Newcombe, H.B., Smith, M.E., and Howe, G.R. (1983), "Reliability of computerized versus manual death searches in a study of the health of eldorado uranium workers", *Computers in Biology and Medecine*, 13, pp. 157-169.

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

Schnell, R., Bachteler, T., and Reiher, J. (2009), "Privacy-preserving Record Linkage using Bloom Filters", BMC 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", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 667-671.