Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Estimating the false negatives due to blocking in record linkage

by Abel Dasylva and Arthur Goussanou

Release date: January 6, 2022

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

© Her Majesty the Queen in Right of Canada as represented by the Minister of Industry, 2022

All rights reserved. Use of this publication is governed by the Statistics Canada Open Licence Agreement.

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, December 2021
Vol. 47, No. 2, pp. 299-311
Statistics Canada, Catalogue No. 12-001-X

Estimating the false negatives due to blocking in record linkage

Abel Dasylva and Arthur Goussanou[1]

Abstract

When linking massive data sets, blocking is used to select a manageable subset of record pairs at the expense of losing a few matched pairs. This loss is an important component of the overall linkage error, because blocking decisions are made early on in the linkage process, with no way to revise them in subsequent steps. Yet, measuring this contribution is still a major challenge because of the need to model all the pairs in the Cartesian product of the sources, not just those satisfying the blocking criteria. Unfortunately, previous error models are of little use because they typically do not meet this requirement. This paper addresses the issue with a new finite mixture model, which dispenses with clerical reviews, training data, or the assumption that the linkage variables are conditionally independent. It applies when applying a standard blocking procedure for the linkage of a file to a register or a census with complete coverage, where both sources are free of duplicate records.

Key Words: Indexing; Massive data sets; Entity resolution; Data integration; Machine learning; Classification.

1. Introduction

Record linkage aims at finding records from the same individual in one or many files (Fellegi and Sunter, 1969; Christen, 2012; Statistics Canada, 2017a). It is different from statistical matching; an imputation method that looks for records from similar individuals (D'Orazio, Di Zio and Scanu, 2006). It has become an important data integration method that includes blocking as an important step. To block is to select a manageable subset of record pairs, which contains most matched pairs, i.e., the pairs with records that come from the same individual. Fellegi and Sunter (1969, Section 3.4) abstractly define blocking as the selection of a subset of the Cartesian product of the two data sources. Herzog, Scheuren and Winkler (2007, page 123, second paragraph) provide a similar definition when they write that "Blocking is a scheme that reduces the number of pairs of records that needs be examined." Christen (2012, page 28, third paragraph) rather uses the term indexing with the same meaning, when he writes that "To reduce the possibly very large number of pairs of records that need to be compared, indexing techniques are commonly applied… These techniques filter out record pairs that are very unlikely to correspond to matches." In this work, the term blocking is used to denote this process that is essential when linking massive data sets that are comprised of millions of records. Indeed the Cartesian product is simply too large. The purpose of blocking is to enforce a trade-off between the computational and memory resources on one hand and the loss of a few matched pairs on the other hand. These matched pairs correspond to false negatives and are an important part of the overall linkage error, if only because the blocking decisions are usually made early on in the linkage process, with no opportunity to change them later. Yet the empirical evidence has been scarce because these false negatives are never reported with few

1. Abel Dasylva, Statistics Canada, 100 Tunney's Pasture, Ottawa, Canada, K1A 0T6. E-mail: abel.dasylva@statcan.gc.ca; Arthur Goussanou, Statistics Canada, 100 Tunney's Pasture, Ottawa, Canada, K1A 0T6. E-mail: arthur.goussanou@statcan.gc.ca.

exceptions, which include the identification of duplicate records in a sampling frame as described by Herzog et al. (2007, Section 12.3). In that rare instance, where the frame comprised of 176,000 business records, the number of matched pairs was estimated at 3,219 of which 3,050 were estimated to be selected by the blocking criteria, i.e. a false negative rate of (3,219 -3,050) / 3,219 = 5.25%, which is not negligible when comparing to the false negative rates reported in various linkage studies reviewed by Bohensky (2016). Nowadays, it is tempting to minimize the blocking false negatives by relaxing the blocking criteria as much as the computing resources permit. After all, these resources are already considerable and ever growing in this age of big data. Yet this may lead to the undesirable situation where the parameters of a probabilistic linkage cannot be estimated because the proportion of matched pairs is too small (Winkler, 2016, Section 2.2.3.2). Thus the issue of the blocking false negatives remains relevant regardless of the available computing resources. However estimating them has been a challenge because of the need to consider all the pairs in the Cartesian product of the two sources, and not just those satisfying the blocking criteria. In that regard, most previous error models are of little use because they do not meet this requirement, including Fellegi and Sunter (1969), Armstrong and Mayda (1993), Thibaudeau (1993), Winkler (1993), Belin and Rubin (1995), Sariyar, Borg and Pommerening (2011), Daggy, Xu, Hui, Gamache and Grannis (2013), and Chipperfield, Hansen and Rossiter (2018). Herzog et al. (2007, Chapter 12.5) have described a capture-recapture technique that does not have this drawback but is impractical because it requires clerical reviews and the conditional independence of some blocking variables.

In this work, a new solution is described, which requires neither. It is based on an extension of the model by Blakely and Salmond (2002) for situations where the records are heterogeneous and the underlying finite population is large. The solution is first developed in the ideal setting where two duplicate-free sources are linked, including a file and a register or a census with complete coverage, such that the decision to keep a pair in the blocks solely depends on its two records, as with standard blocking procedures (see Christen, 2012, Chapter 4.1). Yet, it is of interest in practical settings where both sources have few duplicate records and the census has near complete coverage, such as the linkage of tax records to the Canadian Census to replace income questions (Statistics Canada, 2017b), or a cohort study where mortality records are linked to a census (Blakely and Salmond, 2002).

The following sections are organized as follows. Section 2 presents the assumptions, notations and terminology. Section 3 explains why the distribution of neighbours provides important error information. Section 4 describes the proposed mixture model. Section 5 presents the expectation-maximization procedure. Section 6 describes the empirical study. Section 7 presents the conclusions and future work.

2. Definitions, notations and assumptions

Matched records: In record linkage, like in other automated classification problems, a clear distinction must be made between the nature of the entities to classify (whether two records are actually from the

same entity) and the decisions made (whether the records are *deemed* from the same entity) according to the observations on these entities (the level of agreement between the records). However there is no consensus on the terms used to refer to these key concepts because record linkage is a multidisciplinary field, at the intersection of statistics, epidemiology and computer science. Indeed, in the first paragraph of their abstract, Fellegi and Sunter (1969) writes that "A mathematical model is developed to provide a theoretical framework for a computer-oriented solution to the problem of recognizing those records in two files which represent identical persons, objects or events (said to be matched)." Thus they refer to whether two given records belong to the same entity. In their book, Herzog, Scheuren and Winkler (2007, page 83, last paragraph) use the term "true match" for the same concept. Yet in the computer science literature, the word "matched" has an entirely different meaning. It refers to the classification decision; the best example being given by Christen (2012) in his book entitled "Data matching". In his book, Newcombe (1988, page 105, second paragraph) also laments the lack of consensus on the meaning of the word "matched" when he writes that "This word is variously used in the literature on record linkage. In this book, however, it is given no special technical meaning and merely implies a pairing of records on the basis of some stated similarity (or dissimilarity)."

In what follows, the term "matched" is used according to the definition given by Fellegi and Sunter (1969) to refer to records from the same entity that may be a person, business, household, etc. It is also applied to a pair with the meaning that the constituent records are matched. Two records are called *unmatched* if they come from different entities.

Finite population and data sources: For the problem at hand consider a large finite population that comprises of N individuals and a recording process such that records from different individuals are mutually independent with independent recording errors. Let m denote the file size, which is assumed to be a random variable such that $m \leq N$ and $m \to \infty$ when $N \to \infty$ (e.g. $m = O(N)$). Let V denote the set of possible record values in either data source, and let v_i denote record i from the file where $v_i \in V$ by definition. For simplicity V is assumed to be finite even if it is usually very large. To further simplify, assume that the two data sources are actually free of duplicate records and that the register has no undercoverage. In other words, each record from the file corresponds to exactly one record from the same individual in the register. Each record is also assumed complete, i.e. without missing values.

Blocking strategies: When linking two large data sources, blocking is used to eliminate the vast majority of pairs with records from different individuals, while keeping all the other pairs and expanding few computing resources. Yet some pairs with records from the same individual are inevitably lost in the process. Christen (2012, Chapter 4.4) has reviewed a variety of blocking procedures including the simplest strategy, where a pair is selected if the records agree perfectly on a single key. Such a procedure is often assumed in the published literature on the analysis of linked data (Chambers and Kim, 2016; Han and Lahiri, 2018). It selects a subset of pairs based on the union of Cartesian products across disjoint post-strata that are also called blocks. In practice, a refinement of this approach is used where a pair is kept if the records agree perfectly on at least one key among many. As a result, the subset of selected pairs is no

longer the union of Cartesian products across disjoint post-strata. In what follows we shall not be concerned with such details but with our ability to accurately estimate the loss resulting from the blocking procedure, when linking a file to a register or census, where both sources have few duplicate records and the register or census has little undercoverage. Perfect examples of such studies are provided by the linkage of tax records to the Canadian Census (Statistics Canada, 2017b) or by a cohort study with mortality records linked to a census (Blakely and Salmond, 2002).

In what follows, it is assumed that the decision to keep a pair only depends on its constituent records. i.e. the blocking decision is equivalent to a mathematical map from $V \times V$ into $\{0,1\}$. This includes a large class of blocking procedures, including standard blocking procedures (Christen, 2012, Section 4.4). Yet it excludes blocking strategies that use some form of clustering such as canopy clustering (Christen, 2012, Section 4.8).

Errors: When applying blocking criteria, two kinds of errors may arise including *false negatives* and *false positives*. A false negative occurs if a matched pair is rejected by the blocking criteria. A false positive occurs if an unmatched pair is accepted by the blocking criteria. These errors are measured by the *false negative rate* (FNR) and the *false positive rate* (FPR), where the former is the proportion of matched pairs that are rejected, and the latter is the proportion of unmatched pairs that are accepted.

When designing the blocking criteria one may minimize the false positive rate while keeping the false negative rate below a threshold (e.g. 1%). Since there are usually many more unmatched pairs than matched pairs in the blocks, this roughly corresponds to minimizing the number of pairs in the blocks while keeping the proportion of lost matched pairs below the said threshold. Of course, the implementation of such a strategy requires the accurate estimation of both error rates. The false positive rate is often much easier to estimate than the false negative rate. Indeed, let B denote the total number of pairs accepted by the blocking criteria. Since the false positive rate isno less than $(B-m)/m(N-1)$ and no more than $B/m(N-1)$, it is well approximated by B/mN if $B >> m$. This estimator is related to the *reduction ratio* that is defined as $1-B/mN$ (Christen, 2012, Chapter 7.3). Estimating the false negatives is a much harder problem. Fortunately the concept of *neighbour* provides valuable insights.

3. Neighbours and errors

When examining the potential errors, it helps to look at how many register records form an accepted (by the blocking criteria) pair with a given file record. In what follows, these records are called *neighbours* of the file record, and their number is denoted by n_i for record i on the file. The empirical n_i distribution provides much information about the errors because in the current setting, each file record would have exactly one neighbour if the blocking strategy were error-free, i.e. no false negatives or false positives. Note that the satisfaction of this condition does not imply the absence of errors. As an example, consider the situation shown in Figure 3.1, where the two sources are registers of a population with $N = 6$ individuals, such that individual i is associated with record i in each register, i.e. the record pair (i, i) is matched for $i = 1, \ldots, 6$. Before looking at the n_i's, it is known that the number of false negatives is either

0 or 1, while the number of false positives is between 0 and 5, for each record in the first register. However the n_i's provide additional error information. Indeed, when $n_i = 0$ (e.g. with record 2), it is known with certainty that there is a false negative but no false positives. When $n_i = 1$, one of two cases may occur, including a first case (as with record 5) where the neighbour is the matched record such that there are no errors, and a second case (as with record 3) where the neighbour is an unmatched record such that there are two errors including a false negative and a false positive. In summary, when $n_i = 1$, the number of false negatives is 0 or 1, while the number of false positives is also 0 or 1. Thus there is no additional information about the false negatives since it was known to be 0 or 1 prior to looking at n_i. However, much information is gained about the false positives, since it was known to be in a wider interval (0 to 5) before looking at n_i. This observation confirms the relative ease with which the false positives may be estimated.

Figure 3.1 Two registers with six individuals.

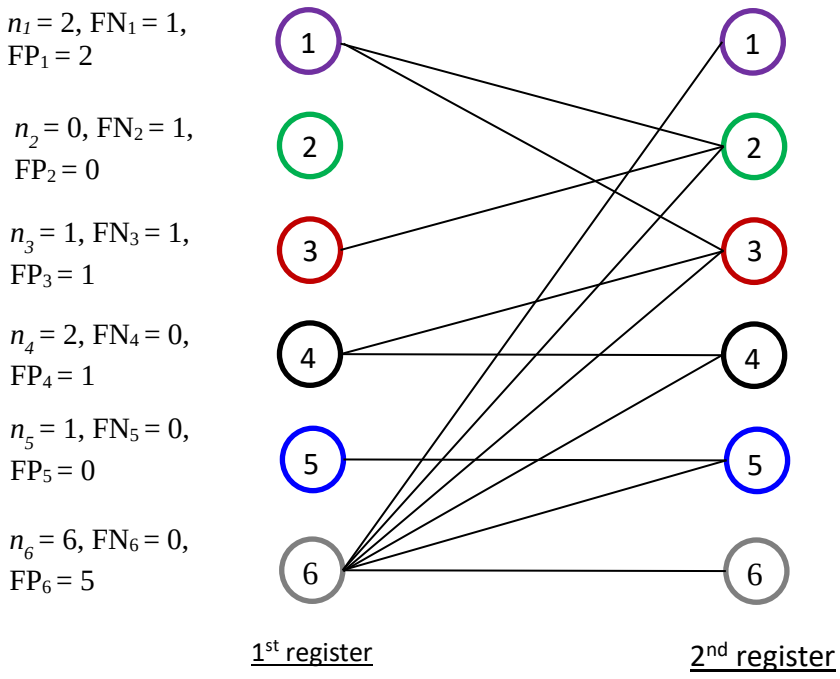

Table 3.1 summarizes the general connection between the number of neighbours and the linkage errors at a given record, in the current setting where each file record is matched with exactly one register record.

Table 3.1
Neighbours and errors

Neighbours (n_i)	False negatives	False positives	Full error information (yes/no)
0	1	0	Yes
1	0 or 1	0 or 1	No
$[2, N-1]$	0 or 1	$n_i - 1$ or n_i	No
N	0	$N-1$	Yes

The above table clearly demonstrates that the number of neighbours provides much error information, including in the case $n_i = 0$ and the more unlikely case $n_i = N$, where this information is complete. When the blocking decision about two records depends on no other record, with a high probability when m and N are large, some file records are bound to have no neighbour. In theory one could design the blocking criteria to ensure a positive n_i for each file record, but this would violate the assumption made in Section 2 that the blocking decision about two records depends on no other record. Under this assumption, the number of neighbours does provide valuable error information, but some uncertainty remains in the case $1 \le n_i \le N-1$, where a statistical model is needed to predict the errors based on n_i.

4. Finite mixture model

The linkage of the two sources is of interest when it is a viable option even N is very large. To capture the essence of such situations, the following two regularity conditions are assumed.

a. Two matched records are neighbours with a probability that is bounded away from 0 regardless of N.
b. Two unmatched records are *accidental* neighbours with a probability of $O(1/N)$.

These assumptions imply that each record has a bounded expected number of neighbours and that $O(N)$ pairs (instead of $O(mN)$ pairs and even $O(N^2)$ pairs if $m = O(N)$) are selected by the blocking criteria. They also imply that there is enough linkage information to identify matched records with a success probability, which is bounded away from zero, regardless of the population size. The above assumptions further imply a particular limiting distribution for the number of neighbours n_i. Indeed, let $n_i = n_{i|M} + n_{i|U}$, where $n_{i|M}$ is the number of matched neighbours and $n_{i|U}$ is the number of unmatched neighbours. Note that these latter variables are not directly observed expect when $n_i = 0$ or $n_i = N$ (see Table 3.1). They are also conditionally independent given v_i and such that $n_{i|M} \mid v_i \sim \text{Bernoulli}(p(v_i))$, $n_{i|U} \mid v_i \sim \text{Binomial}(N-1, \lambda(v_i)/(N-1))$, if an unmatched record is a neighbour with the probability $\lambda(v_i)/(N-1)$ independently of the other unmatched records. When the functions $p(.)$ and $\lambda(.)$ do not depend on N and N is large, we have $n_{i|U} \mid v_i \approx \text{Poisson}(\lambda(v_i))$ (Billingsley, 1995), where $\approx$ means approximately distributed as. Hence, $n_i \mid v_i \approx \text{Bernoulli}(p(v_i)) * \text{Poisson}(\lambda(v_i))$, where $*$ is the convolution operator. Note that, in general, the functions $p(.)$ and $\lambda(.)$ are unknown high-dimensional parameters. To simplify, further assume that $(p(.), \lambda(.))$ is (well approximated by) a piecewise constant function with G levels, such that we have the finite mixture model $n_i \sim \sum_{g=1}^{G} \alpha_g (\text{Bernoulli}(p_g) * \text{Poisson}(\lambda_g))$ holds approximately. When G is fixed, the unknown model parameters are given by the vector $\psi = \left[(\alpha_g, p_g, \lambda_g)\right]_{1 \le g \le G}$ that may be estimated with the Expectation-Maximization (EM) procedure in the next section.

The connection between the error rates and model parameters is made by first noting that the FNR and FPR definitions imply

$$\begin{aligned} \text{FNR} &= \frac{1}{m}\sum_{i=1}^{m}(1-n_{i|M}), \\ (N-1)\,\text{FPR} &= \frac{1}{m}\sum_{i=1}^{m} n_{i|U}. \end{aligned} \tag{4.1}$$

When $m = N$ almost surely, the above equations imply that

$$\begin{aligned} E[\text{FNR}] &= 1 - E[n_{i|M}], \\ &= 1 - E[p(v_i)], \\ (N-1)\,E[\text{FPR}] &= E[n_{i|U}] \\ &= E[\lambda(v_i)], \end{aligned} \tag{4.2}$$

where $E[p(v_i)] = \sum_{g=1}^{G} \alpha_g p_g$ and $E[\lambda(v_i)] = \sum_{g=1}^{G} \alpha_g \lambda_g$ with the finite mixture model. When m is random and such that

$$\begin{aligned} \frac{1}{m}\sum_{i=1}^{m} n_{i|M} &\xrightarrow{p} E[n_{i|M}], \\ \frac{1}{m}\sum_{i=1}^{m} n_{i|U} &\xrightarrow{p} E[n_{i|U}], \end{aligned} \tag{4.3}$$

as $N \to \infty$, the error rates and the model parameters are related as follows

$$\begin{aligned} \text{FNR} &\xrightarrow{p} 1 - E[p(v_i)], \\ (N-1)\,\text{FPR} &\xrightarrow{p} E[\lambda(v_i)]. \end{aligned} \tag{4.4}$$

5. Estimation procedure

The model parameters may be estimated by maximizing the *composite* likelihood (Varin, Reid and Firth, 2011) of the sample $n_1, \ldots, n_m$. For brevity, this composite likelihood is subsequently called likelihood. To develop the EM procedure (Dempster, Laird and Rubin, 1977) it is convenient to first derive the maximum likelihood (ML) equations for the *complete data*, which are comprised of the latent variables $n_{i|M}$, $n_{i|U}$ and $(c_{i1}, \ldots, c_{iG})$ for each i; c_{ig} being the indicator that record i is from class g.

After some algebra, the ML equations for the complete data are as follows.

$$\begin{aligned} \hat{p}_g &= \frac{\sum_{i=1}^{m} c_{ig} n_{i|M}}{\sum_{i=1}^{m} c_{ig}} \\ \hat{\lambda}_g &= \frac{\sum_{i=1}^{m} c_{ig} n_{i|U}}{\sum_{i=1}^{m} c_{ig}} \\ \hat{\alpha}_g &= \frac{1}{m}\sum_{i=1}^{m} c_{ig}. \end{aligned} \tag{5.1}$$

Consequently the ML equations for the observed data (the n_i's) are as follows.

$$\begin{aligned} \hat{p}_g &= \frac{\sum_{i=1}^{m} E[c_{ig} n_{i|M} \mid n_i; \psi]}{\sum_{i=1}^{m} E[c_{ig} \mid n_i; \psi]} \\ \hat{\lambda}_g &= \frac{\sum_{i=1}^{m} E[c_{ig} n_{i|U} \mid n_i; \psi]}{\sum_{i=1}^{m} E[c_{ig} \mid n_i; \psi]} \\ \hat{\alpha}_g &= \frac{1}{m} \sum_{i=1}^{m} E[c_{ig} \mid n_i; \psi]. \end{aligned} \tag{5.2}$$

The EM procedure alternates between the M-step given by Equation (5.2) and the E-step equations in Appendix A.

The above procedure may produce consistent point estimators even if it treats the sample $n_1, \ldots, n_m$ as if it were independent and identically distributed. However this is likely to generate some bias when estimating the variance and the critical levels of hypothesis tests.

6. Empirical study

The empirical study is based on staffing data from the Public Service Resourcing System (PSRS), which is used by applicants to the federal public service in Canada. A given user may open many accounts and apply to many jobs using the same account; each account being associated with a distinct email address. To fulfill its mandate, the Public Service Commission needs to identify all accounts from a given applicant. However this is a challenge because there is no unique identifier except for a minority of applicants. Instead, for most records, the linkage must be based on the given name, the surname and the partial birthdate, which are available for all records. The partial birthdate is comprised of the day and month of birth along with the last digit of the birth year.

The empirical study is based on a subset of 126,330 records selected from the PSRS data since 2006. The selection is based on the following criteria.

- A nonmissing unique identifier.
- Nonmissing given name, surname and partial birthdate.
- Two records for each selected value of the unique identifier.

The selected records represent 63,155 distinct values of the identifier and so many distinct individuals, with two matched records per individual. These records are split into two complete and duplicate-free registers that are linked with the following blocking criteria, and without the unique identifier. A pair is selected if the partial birthdate is the same and the SOUNDEX code (Herzog et al., 2007, Chapter 11) is the same for the given name or the surname. The expected error rates are estimated with the model and compared with the actual values based on the unique identifier.

In Figure 6.1, the histogram shows that the vast majority or records have exactly one neighbour. However 1,659 records have no neighbour, while five records have five neighbours; the maximum number of neighbours of any record.

Figure 6.1 Histogram of the number of neighbours.

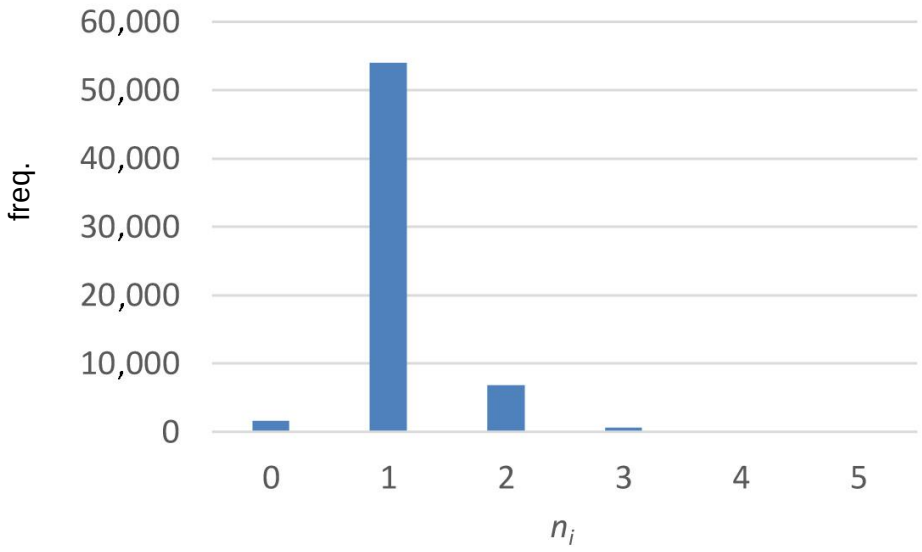

Table 6.1 cross-classifies the records by their number of neighbours and linkage errors, in agreement with Table 3.1.

Table 6.1
Number of neighbours and errors

Neighbours (n_i)	False negatives	False positives	Freq.
0	1	0	1,659
1	1	1	116
1	0	0	53,835
2	1	2	8
2	0	1	6,867
3	1	3	1
3	0	2	602
4	0	3	62
5	0	4	5

The confusion matrix is as follows.

Table 6.2
Confusion matrix

	Link	Non-link	Total
Matched	61,371	1,784	63,155
Unmatched	8,412	3.99E9	3.99E9
Total	69,783	3.988E9	3.989E9

From this matrix, $\text{FNR} = 1,784 / 63,155 = 0.0282$ and $\text{FPR} = 8,412 / 3.99\text{E}9 = 2.11\text{E} - 6.$ Both measures may be viewed as the estimators $\hat{E}[\text{FNR}]$ and $\hat{E}[\text{FPR}]$ of their respective expectations. Since the false negative rate is the summation of independent and identically distributed random variables, its variance may be estimated by

$$\hat{\text{var}}(\text{FNR}) = \frac{1}{N(N-1)} \sum_{i=1}^{N} (1 - n_{i|M} - \text{FNR})^2,$$

based on the latent variables $n_{1|M}, \ldots, n_{m|M}$, which are not directly observed in practice. As a result, the estimated FNR variance is $\hat{\text{var}}(\text{FNR}) = 4.35\text{E} - 7.$ This means the estimated standard error $\hat{\text{SE}}(\hat{E}[\text{FNR}]) = 6.6\text{E} - 4$ for the estimator $\hat{E}[\text{FNR}]$, and the 95% normal confidence interval $\hat{E}[\text{FNR}] \mp z_{\alpha/2} \hat{\text{SE}}(\hat{E}[\text{FNR}]) = (2.82\text{E} - 2 \mp 1.3\text{E} - 3)$ for the expected FNR, where $\alpha = 0.05$ and $z_{\alpha/2} = 1.96$. The corresponding 99% confidence interval is $(2.82\text{E} - 2 \mp 1.71\text{E} - 3)$. Estimating the FPR variance is more difficult because the FPR involves a second order U statistic (Hoeffding, 1948; Lee, 1990). As a matter of fact, Table 6.1 does not give enough information to estimate this statistic. Estimating the variance of the model-based estimators is also challenging because the n_i's are correlated. All the point estimates are given in Table 6.3, where the first row gives the actual FNR and FPR.

Table 6.3
Point estimates

		$\hat{E}$[FNR]	$\hat{E}$[FPR]
Unique id		0.0282	2.11E-6
Model	G = 1	0.0301	2.14E-6
	G = 2	0.0298	2.13E-6
	G = 3	0.0303	2.14E-6

The results show that the model based estimates are very close to the actual FNR and FPR when using one, two or three classes. For the false negative rate, the relative error is $100 \times |0.0303 - 0.0282| / 0.0282 = 7.45\%,$ while this relative error is $100 \times |2.11 - 2.14| / 2.11 = 1.42\%$ for the false positive rate. The small relative errors are encouraging regarding the accuracy of the proposed estimators, even if the model estimates of the expected FNR lie slightly outside the 95% confidence interval. However, the estimate belongs to the 99% confidence interval when using two classes. Choosing two classes seems optimal because the resulting estimate has the smallest relative error with respect to the actual FNR.

7. Conclusions and future work

A new finite mixture has been proposed for estimating the false negatives due to a standard blocking procedure, when linking a file to a register or a census with complete coverage, when both sources are free

of duplicate records. An empirical study with social data gives encouraging results. Yet future work must address the issues of variance estimation and statistical inference about the number of classes. Extensions are also required to account for undercoverage and duplicate records.

Disclaimer

The content of this paper represents the authors' opinions and not necessarily those of Statistics Canada. It describes theoretical methods that might not reflect those currently implemented by the Agency.

Acknowledgements

The authors express their gratitude towards Dr. Jonnagada Rao for his insight and towards the Public Service Commission for access to the data.

Appendix A

For the E-step, the equations are as follows.

$$
\begin{aligned}
P(n_i \mid c_{ig}=1) &= I(n_i=0)(1-p_g)e^{-\lambda_g} + I(n_i>0)\left(p_g+(1-p_g)\frac{\lambda_g}{n_i}\right)\frac{e^{-\lambda_g}\lambda_g^{n_i-1}}{(n_i-1)!} \\
P(c_{ig}=1 \mid n_i) &= \frac{\alpha_g P(n_i \mid c_{ig}=1)}{\sum_{g'=1}^{G}\alpha_{g'}P(n_i \mid c_{ig'}=1)} \\
P(n_{i|M}=1 \mid n_i, c_{ig}=1) &= \frac{p_g n_i}{p_g n_i + (1-p_g)\lambda_g} \\
P(n_{i|U}=n_i \mid n_i, c_{ig}=1) &= I(n_i=0) + I(n_i>0)\frac{(1-p_g)\lambda_g}{p_g n_i + (1-p_g)\lambda_g} \\
P(n_{i|U}=n_i-1 \mid n_i, c_{ig}=1) &= \frac{p_g n_i}{p_g n_i + (1-p_g)\lambda_g}
\end{aligned}
$$

and

$$
\begin{aligned}
E[c_{ig}n_{i|M} \mid n_i] &= P(c_{ig}=1 \mid n_i)\,P(n_{i|M}=1 \mid n_i, c_{ig}=1) \\
E[c_{ig}n_{i|U} \mid n_i] &= P(c_{ig}=1 \mid n_i)\,E[n_{i|U} \mid n_i, c_{ig}=1] \\
E[n_{i|U} \mid n_i, c_{ig}=1] &= \left(\frac{p_g(n_i-1)+(1-p_g)\lambda_g}{p_g n_i + (1-p_g)\lambda_g}\right)n_i.
\end{aligned}
$$

References

Armstrong, M., and Mayda, J. (1993). Model-based estimation of record linkage error rates. *Survey Methodology*, 19, 2, 137-147. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1993002/article/14459-eng.pdf.

Belin, T., and Rubin, D. (1995). A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*, 90, 694-707.

Billingsley, P. (1995). *Probability and Measure*, Wiley.

Blakely, T., and Salmond, C. (2002). Probabilistic record linkage and a method to calculate the positive predicted value. *Journal of Epidemiology*, 31, 1246-1252.

Bohensky, M. (2016). Bias in data linkage studies. In *Methodological Developments in Data Linkage*, (Eds., K. Harron, H. Goldstein and C. Dibben), Chichester: Wiley, 63-82.

Chambers, R., and Kim, G. (2016). Secondary analysis of linked data. In *Methodological Developments in Data Linkage*, (Eds., K. Harron, H. Goldstein and C. Dibben), Chichester: Wiley, 83-108.

Chipperfield, J., Hansen, N. and Rossiter, P. (2018). Estimating precision and recall for deterministic and probabilistic record linkage. *International Statistical Review*, 86, 219-236.

Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution and Duplicate Detection*, Springer.

Daggy, J., Xu, H., Hui, S., Gamache, R. and Grannis, S. (2013). A practical approach for incorporating dependence among fields in probabilistic record linkage. *BMC Medical Informatics and Decision Making*, 13, 1-8.

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, 1-38.

D'Orazio, M., Di Zio, M. and Scanu, M. (2006). *Statistical Matching, Theory and Practice*, Wiley.

Fellegi, I., and Sunter, A. (1969). A theory of record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

Han, Y., and Lahiri, P. (2018). Statistical analysis with linked data. *International Statistical Review*, 87, special issue, 139-157, 1C19 doi:10.1111/insr.12295.

Herzog, T., Scheuren, F. and Winkler, W. (2007). *Data Quality and Record Linkage Techniques*, Springer.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distributions. *Annals of Statistics*, 19, 293-325.

Lee, A.J. (1990). *U-Statistics: Theory and Practice*, Marcel Dekker.

Newcombe, H. (1988). *Handbook of Record Linkage*, Oxford University Press.

Sariyar, M., Borg, A. and Pommerening, K. (2011). Controlling false match rates in record linkage using extreme value theory. *Journal of Biomedical Informatics*, 44, 648-654.

Statistics Canada (2017a). *Record Linkage Project Process Model*, Catalogue No. 12-605-X.

Statistics Canada (2017b). *Income Reference Guide, Census of Population, 2016*, Catalogue No. 98-500-X2016004.

Thibaudeau, Y. (1993). The discrimination power of dependency structures in record linkage. *Survey Methodology*, 19, 1, 31-38. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1993001/article/14477-eng.pdf.

Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21, 5-42.

Winkler, W. (1993). Improved decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the 1993 Joint Statistical Meetings*, American Statistical Association, 274-279.

Winkler, W.E. (2016). Probabilistic linkage. In *Methodological Developments in Data Linkage*, (Eds., K. Harron, H. Goldstein and C. Dibben), Chichester: Wiley, 7-35.