Australian
Bureau of
Statistics

1352.0.55.149

Research Paper

Measuring Precision for Deterministic and Probabilistic Record Linkage

New Issue

Research Paper

Measuring Precision for Deterministic and Probabilistic Record Linkage

James Chipperfield, Noel Hansen
and Peter Rossiter

Methodology Transformation Branch

Methodology Advisory Committee

4–5 June 2015, Canberra

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) MON 23 JAN 2017

ABS Catalogue no. 1352.0.55.149

Views expressed in this paper are those of the author(s), and do not necessarily represent those of the Australian Bureau of Statistics.
Where quoted, they should be attributed clearly to the author(s).

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Mr Daniel Elazar, Methodology Transformation Branch, on Canberra (02) 6252 6962 or email <analytical.services@abs.gov.au>.

MEASURING PRECISION FOR DETERMINISTIC AND PROBABILISTIC RECORD LINKAGE

James Chipperfield, Noel Hansen and Peter Rossiter
Methodology Transformation Branch

EXECUTIVE SUMMARY

Record linkage is the act of bringing together records from two files that belong to, or are likely to belong to, the same unit (e.g. person, student, business). Record linkage is an appropriate technique when data sets need to be joined to enhance dimensions such as time and breadth or depth of detail. For example, the Australian Census Longitudinal Database (ACLD), created by linking the 2006 and 2011 Australian Population Censuses, allows longitudinal analysis (ABS, 2013a). Record linkage offers opportunities for new and enhanced statistical output and analysis at relatively low cost.

With these new opportunities comes the associated problem of linkage errors. The prevalence of linkage errors is often difficult to estimate because the errors themselves (e.g. linking records that belong to two different people) may not be detected. Links can be declared deterministically, using a set of pre-defined rules, or probabilistically, where evidence for a link being a match is weighed against the evidence that it is not a match. Both methods are widely used at the ABS. This paper describes methods of estimating the prevalence of linkage errors for deterministic and probabilistic linking. It is envisaged that these methods will be used as part of the quality assurance process for record linkage at the ABS.

First we present some necessary background to record linkage.

A *match* is a pair of records that belong to the same unit. A *non-match* is a pair of records that do not belong to the same unit. The population of interest in record linkage is the complete set of matches. Perfect linkage occurs when all matches are linked and no non-matches are linked. Perfect linkage would be possible if a unique person identifier was available on the files. Perfect linkage of a person's record could be possible with name and address. In many situations, however, name and address are not available and the linking fields that are available do not uniquely identify a unit, are missing or contain errors.

Perfect linkage is typically not possible and linkage errors occur. Linkage errors can have negative consequences for the validity of analysis of the linked file. The two types of linkage errors are *missed records* and *incorrect links*. A missed record is a record that was not linked to any record even though its match exists. Commonly used measures for missed records are the *Link Rate*, which is the number of linked

records divided by the total number of matches that exist, and the *Match Rate*, which is the proportion of all matches that are linked.

The impact of Link Rate on analysis is analogous to the impact of non-response, in the sample survey context, on analysis: the linked records may not be representative of the matches. For example, because some linking variables are not applicable to children (e.g. marital status, highest education attainment, and industry of occupation) we have frequently found that children's records are more likely to be missed than adult records. To minimise the potential for one sub-group to be under represented on the linked file, a reasonable approach is to use as many linking variables as possible to differentiate between matches and non-matches. Explicitly considering linking variables for children's records would be important in this regard. Calibrating the weights of linked records to known population totals, possibly calculated directly from one of the files used in linking, can reduce the bias due to missed records.

A link is either correct (i.e. a match) or incorrect (i.e. a non-match). A commonly used measure of linkage error is *Precision*, which is the proportion of links that are matches. Incorrect links create a type of measurement error and can bias analysis. From detailed studies of linking Census records containing only categorical variables, the analytic conclusions based on a linked file with Precision=95% are often not substantively different to those based on a perfectly linked file. The impacts of this bias and ways to correct it have been studied , but these methods are still new and further advancement in the literature would be required before they are adopted by the ABS.

There is typically a trade-off between Precision and Link Rate: accepting more links typically increases the Link Rate and decreases the Precision. This trade-off has meaning to the extent that an increase in the Link Rate will reduce the potential for bias due to missed records while a decrease in the Precision will increase the potential for bias due to incorrect links. While bias is very difficult to estimate, the trade-off between Precision and Link Rate is still a useful way to compare two competing linking strategies or to decide if a linking strategy is worthwhile at all. This is illustrated later in this paper.

Unfortunately, Precision is not easy to estimate. Even a clerical review of a link cannot always be relied upon to decide if a link is match or non-match. Link Rate is often easy to accurately estimate after files have been linked because the number of links is observed and the total number of matches that exist can usually be accurately approximated. Both Precision and Link Rate are difficult to estimate in the situation where the files' linking variables are known but the files themselves are not available. This paper will describe a framework for estimating Precision and Link Rate. The framework is model-based and does not require clerical review.

The uses of this framework are to estimate:

1. Precision and Link Rate before files are available for linkage. A typical scenario is where a client expresses an interest in funding the ABS to link two files. The ABS would like to estimate, based on limited information, the Precision and Link Rate if it were to proceed with linking the files.
2. Precision during the linking process. This would be useful to refine how linkage is carried out, such as the choice of linking variables.
3. Precision after the files are linked. This would be a useful "quality indicator" for published counts.

ooooo

Record linkage activities at the Australian Bureau of Statistics are strictly carried out in a manner that protects privacy and confidentiality and ensures that there is significant public benefit before a record linkage project is undertaken.

Specifically, the ABS adheres to all relevant legislation and guidelines, including the *Privacy Act 1988* and the *High Level Principles for Data Integration Involving Commonwealth Data for Statistical and Research Purposes*.

The *Census and Statistics Act 1905* and the *Privacy Act 1988* require that all information submitted to, or collected by the ABS remain confidential. All ABS staff, including temporary employees, are legally bound never to release personal information to any individual or organisation outside the ABS. In addition, comprehensive security arrangements are implemented in ABS computer systems. These include use of regularly changed passwords, access controls and audit trails.

Legislative requirements to ensure privacy and secrecy also apply to the outputs of record linkage activities. In accordance with the *Census and Statistics Act 1905*, results must be confidentialised to ensure that they are not likely to enable identification of a particular person or organisation.

QUESTIONS FOR THE COMMITTEE

1. Is it worthwhile considering a completely new estimator of Precision altogether? If not, are there key assumptions in the latent model that we should try to relax to make the estimator more robust?
2. Do the simulation and empirical studies convince you that the proposed estimator of Precision is worthwhile in practice? If not, what other studies should we consider in order to do this?

CONTENTS

The role of the Methodology Advisory Committee (MAC) is to review and direct research into the collection, estimation, dissemination and analytical methodologies associated with ABS statistics. Papers presented to the MAC are often in the early stages of development, and therefore do not represent the considered views of the Australian Bureau of Statistics or the members of the Committee. Readers interested in the subsequent development of a research topic are encouraged to contact either the author or the Australian Bureau of Statistics.

MEASURING PRECISION FOR DETERMINISTIC AND PROBABILISTIC RECORD LINKAGE

James Chipperfield, Noel Hansen and Peter Rossiter
Methodology Transformation Branch

ABSTRACT

It is widely recognised that greater publication, sharing and linking of existing data sources holds considerable potential to increase transparency, improve service delivery, transform policy outcomes and help to drive innovation, productivity and economic growth.

Subject to strict safeguards and where there is significant public benefit, the Australian Bureau of Statistics (ABS) is increasingly making use of record linkage techniques to combine existing sources of data, for the purpose of producing analytical datasets that have enhanced temporal and cross-sectional detail. Frequently this record linkage must be achieved without the benefit of unique or definitive linkage keys, and consequently incorrect links may result. The proportion of links that are correct, or the ‘precision’ of the record linkage, can be difficult to establish when even careful clerical review may fail to resolve whether or not links are correct. Measures of precision are useful for deciding whether to proceed with a record linkage project, for evaluating alternative linking strategies and for establishing quality measures for estimates based on the linked data. This paper proposes an estimator of precision for a linked dataset that has been created by either deterministic (rules-based) or probabilistic record linkage. Both methods are widely used at the ABS. The paper shows that the proposed estimators perform well in simulation, and it is envisaged that the proposed estimator will be part of the ABS’ record linkage tool kit.

1. INTRODUCTION

Record linkage is the act of bringing together records from two files that belong to, or are likely to belong to, the same unit (e.g. person, student, business). Record linkage is an appropriate technique when data sets need to be joined to enhance dimensions such as time and breadth or depth of detail. For example, the Australian Census Longitudinal Database (ACLD), created by linking the 2006 and 2011 Australian Population Censuses, allows longitudinal analysis (ABS, 2013a). Record linkage offers opportunities for new statistical output and analysis at relatively low cost.

With these new opportunities comes the associated problem of linkage errors. The prevalence of linkage errors is often difficult to estimate because the errors themselves (e.g. linking records that belong to two different people) may not be detected. Links can be declared deterministically, using a set of pre-defined rules, or probabilistically, where evidence for a link being a match is weighed against the evidence that it is not a match. Both methods are widely used at the ABS. This paper describes methods of estimating the prevalence of linkage errors for deterministic and probabilistic linking. It is envisaged that these methods will be used as part of the quality assurance process for record linkage at the ABS, as explained in more detail later in this section. First we present some necessary background to record linkage.

A *match* is a pair of records that belong to the same unit. A *non-match* is a pair of records that do not belong to the same unit. The population of interest in record linkage is the complete set of matches. Perfect linkage occurs when all matches are linked and no non-matches are linked. Perfect linkage would be possible if a unique person identifier was available on the files. Perfect linkage of a person's record could be possible with name and address. In many situations, however, name and address are not available and the linking fields that are available do not uniquely identify a unit, are missing or contain errors.

Perfect linkage is typically not possible and linkage errors occur. Linkage errors can have negative consequences for the validity of analysis of the linked file. The two types of linkage errors are *missed records* and *incorrect links*. A missed record is a record that was not linked to any record even though its match exists. Commonly used measures for missed records are the *Link Rate*, which is the number of linked records divided by the total number of matches that exist, and the *Match Rate*, which is the proportion of all matches that are linked. As Link Rate and Match Rate are measures of missed records, for much of the paper we simply mention Link Rate.

The impact of Link Rate on analysis is analogous to the impact of non-response, in the sample survey context, on analysis: the linked records may not be representative of the matches. For example, because some linking variables are not applicable to children (e.g. marital status, highest education attainment, and industry of occupation) we have frequently found that children's records are more likely to be missed than adult records. To minimise the potential for one sub-group to be under represented on the linked file, a reasonable approach is to use as many linking variables as possible to differentiate between matches and non-matches. Explicitly considering linking variables for children's records would be important in this regard. Calibrating the weights (Särndal *et al.*, 1992) of linked records to known population totals, possibly calculated directly from one of the files used in linking, can reduce the bias due to missed records.

A link is either correct (i.e. a match) or incorrect (i.e. a non-match). A commonly used measure of linkage error is *Precision*, which is the proportion of links that are matches. Incorrect links create a type of measurement error and can bias analysis. From detailed studies of linking Census records containing only categorical variables, the analytic conclusions based on a linked file with Precision=95% are often not substantively different to those based on a perfectly linked file. The impacts of this bias and ways to correct it have been studied (see for example Chipperfield and Chambers, 2015 and Chipperfield *et al.*, 2011). These methods are still new and further advancement in the literature would be required before they are adopted by the ABS.

There is typically a trade-off between Precision and Link Rate: accepting more links typically increases the Link Rate and decreases the Precision. This trade-off has meaning to the extent that an increase in the Link Rate will reduce the potential for bias due to missed records while a decrease in the Precision will increase the potential for bias due to incorrect links. While bias is very difficult to estimate, the trade-off between Precision and Link Rate is still a useful way to compare two competing linking strategies or to decide if a linking strategy is worthwhile at all. This is illustrated later in this paper.

Unfortunately, Precision is not easy to estimate. Even a clerical review of a link cannot always be relied upon to decide if a link is match or non-match. Link Rate is often easy to accurately estimate after files have been linked because the number of links is observed and the total number of matches that exist can usually be accurately approximated. Both Precision and Link Rate are difficult to estimate in the situation where the files' linking variables are known but the files themselves are not available. This paper will describe a framework for estimating Precision and Link Rate. The framework is model-based and does not require clerical review.

The uses of this framework are to estimate:

1. Precision and Link Rate before files are available for linkage. A typical scenario is where a client expresses an interest in funding the ABS to link two files. The ABS would like to estimate, based on limited information, the Precision and Link Rate if it were to proceed with linking the files.
2. Precision during the linking process. This would be useful to refine how linkage is carried out, such as the choice of linking variables.
3. Precision after the files are linked. This would be a useful "quality indicator" for published counts.

Section 2 describes the latent model for the probability that matches and non-matches will agree on the value of the linking variables. Section 3 describes two linkage methods, deterministic and probabilistic, used by the ABS and shows they are motivated by the latent model of Section 2. Section 4 describes a method of estimating Precision for deterministic and probabilistic linkage methods that involves simulating the linkage process many times. Section 5 describes an alternative estimator for Precision under deterministic linkage, which is much faster to calculate than the simulation method in Section 4. Section 6 describes the results of a simulation study of the accuracy of the proposed estimators and Section 7 describes the results of an empirical study. The results show that the proposed methods work well. Section 8 makes some conclusions.

2. MODEL FOR COMPARISON OUTCOMES

We consider linking two files, File X containing m records and File Y containing n records, where $m \le n$. Unless otherwise mentioned, we assume that all records on File X have a matching record on File Y. If record i on File X and record j on File Y is a potential link they are referred to as the (i, j)-th record pair, where $i = 1, \ldots, m$ and $j = 1, \ldots, n$. Let there be L linking variables and let $l = 1, \ldots, L$. Denote the random variable for the comparison outcome on L linking variables of a record pair by $\boldsymbol{a} = (a_1, \ldots, a_l, \ldots, a_L)'$, where $a_l = 1$ means the pair agrees on the l-th linking variable, $a_l = 0$ means the pair disagrees on the l-th linking variable, and $a_l = -1$ means one or both of the l-th linking variables in the record pair are missing. The data linker is free to decide what constitutes an agreement or disagreement. In the case of *age*, the data linker may define agreement as either exact or approximate (i.e. within one year). There are a total of $S = 3^L$ comparison outcomes or possible values for $\boldsymbol{a}$.

Now denote the random variable for the comparison outcome of the (i, j)-th record pair by $\boldsymbol{a}_{ij} = (a_{ij1}, \ldots, a_{ijl}, \ldots, a_{ijL})'$, where a_{ijl} gives the comparison outcome (i.e. equals –1, 0 or 1) on the l-th linking variable. If the (i, j)-th record pair agrees on the first two linking variables and disagrees on the third, then $\boldsymbol{a}_{ij} = (1,1,0)'$. To be clear, as the indexes i and j are arbitrary, $i = j$ does not indicate the correct links. We also define the matrix of comparison outcomes for all record pairs by $\mathbf{A} = (\boldsymbol{a}_{11}, \ldots, \boldsymbol{a}_{ij}, \ldots, \boldsymbol{a}_{mn})'$.

Next we describe two latent models for $\mathbf{A}$, where the latent model in Section 2.2 is an extension on the model in Section 2.1. Both these models assume that the rows of $\mathbf{A}$ (i.e. the record pairs) are independently distributed and so we can simply focus on the distribution of $\boldsymbol{a}$.

2.1 Standard approach

Under the standard approach, only the distribution of $\boldsymbol{a}$ is of interest. The distribution of $\boldsymbol{a}$ is assumed to depend upon the latent class it belongs to, where latent classes are indexed by $g = 1, \ldots, G$. To illustrate, it is reasonable to assume that the distribution of $\boldsymbol{a}$ for matched record pairs will be different to that for non-matched record pairs. In particular, one would expect $a_l = 1$ (i.e. agree) with higher probability for matches than for non-matches.

$$\text{Prob}(\boldsymbol{a}) = \sum_g \text{Prob}(\boldsymbol{a} \mid \text{class } g)\,\text{Prob}(\text{class } g) = \sum_g \pi_{\boldsymbol{a}|g}\,\pi_g \qquad \text{(1a)}$$

where $\text{Prob}(\boldsymbol{a} \mid \text{class } g) = \pi_{\boldsymbol{a}|g}$ is the probability that a record pair belonging to the g-th latent class has comparison outcome $\boldsymbol{a}$, and $\text{Prob}(\text{class } g) = \pi_g$ is the probability that a record pair belongs to the g-th latent class and is usually assumed to be approximately known (i.e., in the present set-up, $\pi_1 = 1/n$ and $\pi_2 = (n-1)/n$).

Larsen and Rubin (2001) and Herzog *et al.* (2007) specify $\pi_{\boldsymbol{a}|g}$ by log-linear models with two and three-way interactions. In this paper we make the strong and simplifying assumption that any two different elements of $\boldsymbol{a}$ are conditionally *independent*: the comparison outcome on one linking variable is independent of the comparison outcome on all other linking variables, conditional on the latent class the record pair belongs to. This means we can write

$$\text{Prob}\left(\boldsymbol{a} \mid \text{class}\, g\right) = \pi_{\boldsymbol{a}|g}^{\text{CI}} = \prod_l \text{Prob}\left(a_l \mid \text{class}\, g\right). \tag{1b}$$

We also set $G = 2$, where $g = 1$ indicates the match latent class and $g = 2$ indicates the non-match latent class. Now let $\delta(a_l) = 1$ if $a_l = 1$ and zero otherwise to indicate agreement, and $\phi(a_l) = 1$ if $a_l = 0$ and zero otherwise to indicate disagreement; so that $\delta(a_l) = 0$ and $\phi(a_l) = 0$ indicates the "missing" outcome. Given the independence assumption, we can express $\pi_{\boldsymbol{a}|1}$ by

$$\pi_{\boldsymbol{a}|1}^{\text{CI}} = \prod_{l=1}^{L} M_l^{\delta(a_l)} D_l^{\phi(a_l)} \left(1 - M_l - D_l\right)^{\left[1-\delta(a_l)-\phi(a_l)\right]} \tag{2a}$$

where M_l and D_l are the probabilities that a record pair belonging to the match latent class agrees ($a_l = 1$) and disagrees ($a_l = 0$) on the l-th linking field, and we can express $\pi_{\boldsymbol{a}|2}$ by

$$\pi_{\boldsymbol{a}|2}^{\text{CI}} = \prod_{l=1}^{L} U_l^{\delta(a_l)} R_l^{\phi(a_l)} \left(1 - U_l - R_l\right)^{\left[1-\delta(a_l)-\phi(a_l)\right]} \tag{2b}$$

where U_l and R_l are the probabilities that a record pair belonging to the non-match latent class agrees ($a_l = 1$) and disagrees ($a_l = 0$) on the l-th linking field.

The probability of observing agreement on *all* linking fields under the assumption of conditional independence is $\prod_l M_l$ for a matching record pair, and $\prod_l U_l$ for a non-matching record pair.

Represent the set of parameters (M_l, D_l, U_l, R_l) for $l = 1, \ldots, L$ by $\boldsymbol{\psi}$. An estimate of $\boldsymbol{\psi}$, $\hat{\boldsymbol{\psi}}$, can be obtained using the well-known EM algorithm (for details, see Herzog *et al.*, 2007). The independence assumption is made by some computer packages because it makes estimation of $\boldsymbol{\psi}$ somewhat straightforward.

2.2 Frequency-based approach

When modelling the comparison outcome $\boldsymbol{a}$ in Section 2.1, the specific values of the linking variables were not of interest. For example, in the case of the linking variable Country of Birth, the model made no distinction between a record pair agreeing on the value "Australia" and a record pair agreeing on the value "Iceland". However, the value on which there is agreement is information that can be used to improve the discrimination between matches and non-matches. In the Australian context for example, a record pair agreeing on Country of Birth ="Australia" provides less evidence for a match than a record pair agreeing on Country of Birth ="Iceland", because records with Country of Birth ="Iceland" occur less frequently. Next we extend (1) and (2) to the so-called *frequency-based* approach, which models the agreement outcome as well as the values of the linking variables upon which there are agreement.

For simplicity in notation, let each linking variable have R (non-missing) possible values given by $r = 1, \ldots, R$. For example, in the case of the linking variable Country of Birth, $r = 1$ may correspond to "Australia" , $r = 2$ may correspond to "Iceland", and so on. If the l-th linking variable for both records in a pair equals r then let the variable $\upsilon_l = r$ and otherwise let $\upsilon_l = 0$. For a record pair we can then define $\boldsymbol{\upsilon} = (\upsilon_1, \ldots, \upsilon_l, \ldots, \upsilon_L)$, which gives the values of the linking variables on which there was agreement.

Under the frequency-based approach, interest is in the distribution of $\boldsymbol{a}$ and $\boldsymbol{\upsilon}$. We consider the distribution of $\boldsymbol{a}$ and $\boldsymbol{\upsilon}$ conditional upon the latent class to which the record pair belongs, which we write as $\pi_{\boldsymbol{a},\boldsymbol{\upsilon}|g}$. The joint distribution of $\boldsymbol{a}$ and $\boldsymbol{\upsilon}$ can be expressed by

$$\begin{aligned}\text{Prob}(\boldsymbol{a},\boldsymbol{\upsilon}) &= \textstyle\sum_g \pi_{\boldsymbol{a},\boldsymbol{\upsilon}|g}\, \pi_g \\ &= \textstyle\sum_g \pi_{\boldsymbol{a}|g}\, \pi_{\boldsymbol{\upsilon}|\boldsymbol{a},g}\, \pi_g \\ &= \textstyle\sum_g \pi_{\boldsymbol{a}|g} \prod_l \left(\pi^{(l)}_{\upsilon_l|g}\right)^{\delta(a_l)} \pi_g\,, \end{aligned} \qquad (3)$$

where $\pi^{(l)}_{\upsilon_l|g}$ is the probability of observing agreement on the value υ_l for record pairs belonging to the g-th latent class that agree on the l-th linking variable, and the simplification $\pi_{\boldsymbol{\upsilon}|\boldsymbol{a},g} = \prod_l \left(\pi^{(l)}_{\upsilon_l|g}\right)^{\delta(a_l)}$ follows from the assumption that the elements of $\boldsymbol{\upsilon}$ are independently distributed conditional on the record pair's agreement outcome and latent class. Since we are willing to make the independence assumption of Section 2.1, we set $\pi_{\boldsymbol{a}|g} = \pi^{\text{CI}}_{\boldsymbol{a}|g}$ in (3). All that remains to evaluate (3) is to evaluate $\pi^{(l)}_{\upsilon_l|g}$.

Now let f_{lr} be the number of records on File X with l-th linking variable equal to r so that if there were no missing values then $m = \sum_r f_{lr}$ for all l. Similarly let h_{lr} be the number of records on File Y with l-th linking variable equal to r so that in the case of no missing values $n = \sum_r h_{lr}$ for all l. Now assume that whether or not an error occurs in the value of a linking variable is completely independent of the linking variable itself (e.g. the prevalence of errors in the variable Country of Birth does not depend upon whether a person was born in Australian or Iceland). This means if $\upsilon_l = r$ we can write

$$\pi_{r|1}^{(l)} = M_{lr} = \frac{f_{lr}}{m}$$

and

$$\pi_{r|2}^{(l)} = U_{lr} = \frac{(h_{lr} - 1) f_{lr}}{\sum_s (h_{ls} - 1) f_{ls}},$$

which is the proportion of non-matches agreeing on the l-th linking variable that agree on the value r .

Now we can express $\pi_{\boldsymbol{a},\boldsymbol{\upsilon}|1}$ (i.e. for matches) by

$$\pi_{\boldsymbol{a},\boldsymbol{\upsilon}|1}^{\text{CI}} = \pi_{\boldsymbol{a}|1}^{\text{CI}} \left\{ \prod\nolimits_l M_{l\upsilon_l}^{\delta(a_l)} \right\} \tag{4a}$$

and we can express $\pi_{\boldsymbol{a},\boldsymbol{\upsilon}|2}$ (i.e. for non-matches) by

$$\pi_{\boldsymbol{a},\boldsymbol{\upsilon}|2}^{\text{CI}} = \pi_{\boldsymbol{a}|2}^{\text{CI}} \left\{ \prod\nolimits_l U_{l\upsilon_l}^{\delta(a_l)} \right\}. \tag{4b}$$

The probability of observing agreement on *all* linking fields under the assumption of conditional independence for the frequency-based approach may be expressed as $\prod_l M_l M_{l\upsilon_l}$ for a matching record pair, and $\prod_l U_l U_{l\upsilon_l}$ for a non-matching record pair.

The extra effort required for the frequency-based approach would appear to be minimal since the extra counts required (i.e. m , n , h_{lr} and f_{lr}) are observed.

3. LINKING PROCEDURES

The ABS uses deterministic and probabilistic record linkage. Section 3.1 describes probabilistic linkage and Section 3.2 describes deterministic linkage. Section 3 also shows how the latent models of Section 2 motivate these linking methods.

3.1 Probabilistic linkage

In probabilistic linking (see Herzog, Scheuren and Winkler, 2007 and Winkler, 2001 and 2005) each record pair (i.e. a possible link) is given a weight based on the likelihood that they are a match. The weight takes into account the evidence that the record pair is a match (i.e. agreement on linkage variables) and the evidence that the record pair is a non-match (i.e. disagreement on linking variables). Naturally, linking variables vary in how much evidence they provide in this regard.

Fellegi and Sunter (1969) suggest ranking the S comparison outcomes by the weight $W_{\boldsymbol{a}} = \pi^{\text{CI}}_{\boldsymbol{a}|1} \big/ \pi^{\text{CI}}_{\boldsymbol{a}|2}$, where $W_{\boldsymbol{a}} > W_{\boldsymbol{a}'}$ means that the comparison outcome $\boldsymbol{a}$ is more likely to be a match than comparison outcome $\boldsymbol{a}'$. The estimate of the weight for the (i, j)-th record pair with comparison outcome $\boldsymbol{a}_{ij}$, is $\hat{W}_{ij} = \sum_l \hat{w}_{ijl}$, where

$$\hat{w}_{ijl} = \begin{cases} \ln\left(\hat{M}_l / \hat{U}_l\right) & \text{if } a_{ijl} = 1 \\ \ln\left(\hat{D}_l / \hat{R}_l\right) & \text{if } a_{ijl} = 0 \\ \ln\left[\left(1 - \hat{M}_l - \hat{D}_l\right) \big/ \left(1 - \hat{U}_l - \hat{R}_l\right)\right] & \text{if } a_{ijl} = -1 \end{cases}$$

Under the frequency based approach, the ranking would instead be based on the weight $W_{\boldsymbol{a},\boldsymbol{\upsilon}} = \pi^{\text{CI}}_{\boldsymbol{a},\boldsymbol{\upsilon}|1} \big/ \pi^{\text{CI}}_{\boldsymbol{a},\boldsymbol{\upsilon}|2}$, where $W_{\boldsymbol{a},\boldsymbol{\upsilon}} > W_{\boldsymbol{a}',\boldsymbol{\upsilon}'}$ means that outcome $[\boldsymbol{a},\boldsymbol{\upsilon}]$ is more likely to be a match than comparison outcome $[\boldsymbol{a}',\boldsymbol{\upsilon}']$. Again let the weight for the (i, j)-th record pair be denoted by $\hat{W}_{ij}$.

A typical linking algorithm involves maximising the sum of the weights for linked pairs, subject to the constraints that links must be 1–1, links can only be formed from record pairs belonging to the same 'block' on File X and Y, and that links must have a weight greater than a threshold or cut-off value (see Herzog *et al.*, 2007). A cut-off value of c means that the comparison outcome, $\boldsymbol{a}$, can be declared a link only if $W_{ij} > c$.

To illustrate, once all weights $\hat{W}_{ij}$ are calculated, a simple 1–1 probabilistic linkage algorithm is described below:

1. List all record pairs, sorted by their weight, $\hat{W}_{ij}$, from highest to lowest;
2. The first record pair in the ordered list is linked if it has a weight greater than the cut-off value;

3. Record pairs containing either of the records linked in step 2 are removed from list;

4. Return to step 2 until no more records can be linked.

It is obvious from at least this particular linking algorithm that the ranking of the record pairs by weight, rather than the value of weights themselves, determine the links. The ABS currently uses the package *Febrl* (Christen and Churches, 2005) to probabilitistically link records. *Febrl* uses a more sophisticated linking algorithm than that described above.

3.2 Deterministic linkage

Deterministic linking concatenates the values of a set of linking variables into what is called a 'linking key'. For example, the Australian Institute of Health and Welfare developed a 14 digit linking key comprised of the second, third and fifth letters of a person's last name, the second and third letters from a person's first name, date of birth and sex (AIHW, 2013). A link is then declared between two records if the value of the records' linking keys is equal and is unique on each individual file.

If a small number of accurately-coded and discriminating linking variables are available (e.g. small area geography and birth date), deterministic linkage is an ideal approach. The main advantage of deterministic linkage over probabilistic linkage is its simplicity – no specialised linkage software is required and it is simple to implement and explain.

To illustrate, under the deterministic approach, record i on File X will be linked to record j on File Y only if:

1. the (i, j)-th record pair agrees on *all* linking variables that make up the linking key. That is, all elements of $\boldsymbol{a}_{ij}$ take the value of 1; and

2. the value of the linking key for the (i, j)-th record pair is unique on both Files. This condition can be expressed solely in terms of the $\boldsymbol{a}_{ij}$ s but we omit it here.

3.3 Multiple passes

Often, a strategy for record linkage is comprised of a sequence of linking passes indexed by $t = 1, \ldots, T$: records linked in pass t are not eligible to be linked in pass t^* if $t^* > t$. The first pass may aim to link people who have not changed address by using address information as linking variables, and the second pass may aim to link people who have changed address by omitting address information from the linking variables. A linkage strategy may use probabilistic and deterministic record linkage in different passes. For example, the creation of the Australian Census Longitudinal Dataset (ACLD) involved deterministic linkage in passes 1 and 2 and probabilistic linkage in passes 3–12 (ABS, 2013).

Each deterministic pass can only link record pairs with a specific comparison outcome (e.g. agreement on all linking variables used in the link key) and this can often mean few links are made in a pass. By contrast, a probabilistic pass can link record pairs with a range of different comparison outcomes, as long as the associated weight is greater than the cut-off. For this reason, most probabilistic linkage strategies at the ABS usually require fewer than five passes while deterministic linkage strategies may involve hundreds of passes.

4. ESTIMATING THE PRECISION BY SIMULATION

Deriving an analytic estimator of Precision for probabilistic linkage is difficult because it often uses complex non-linear search algorithms under a 1–1 constraint to declare links. Other complexities such as missing linking variables, multiple linking passes, and the possibility that a record on File X has no match on File Y create even more complexity. In fact, Chipperfield and Chambers (2015) show that the analytic estimates of Precision in Lahiri and Larsen (2005) are poor for 1–1 probabilistic linkage.

The basic idea here is to simulate the linkage process, whether it is deterministic or probabilistic, many times in order to estimate Precision. The key step is to simulate the comparison outcome, $\mathbf{A}$. When linking with multiple passes, $\mathbf{A}$ would contain the blocking and linking variables used in all passes.

Denote the b-th simulated version of $\mathbf{A}$ by $\mathbf{A}(b) = \left(\boldsymbol{a}_{11}(b), \ldots, \boldsymbol{a}_{ik}(b), \ldots, \boldsymbol{a}_{mn}(b)\right)$, where $\boldsymbol{a}_{ik}(b) = \left(a_{ik1}(b), \ldots, a_{ikl}(b), \ldots, a_{ikL}(b)\right)'$, $a_{ikl}(b)$ is the b-th simulated comparison on the l-th linking field for the (i,k)-th record pair, and $\boldsymbol{a}_{ii}(b)$ is the b-th simulated comparison outcome for the i-th record on File X with its matching record on File Y.

In order to generate $\mathbf{A}(b)$ under the independence assumption of Section 2.1, the $a_{ikl}(b)$ are independently calculated over i, k and l in the following way:

For $i = k$ (i.e. for record pairs that are a match)

$$a_{iil}(b) = \begin{cases} 1 & \text{with probability } \hat{M}_l , \\ 0 & \text{with probability } \hat{D}_l , \\ -1 & \text{with probability } \left(1 - \hat{M}_l - \hat{D}_l\right). \end{cases}$$

If $i \neq k$ (i.e. for record pairs that are non-matches)

$$a_{ikl}(b) = \begin{cases} 1 & \text{with probability } \hat{U}_l , \\ 0 & \text{with probability } \hat{R}_l , \\ -1 & \text{with probability } \left(1 - \hat{U}_l - \hat{R}_l\right). \end{cases}$$

Now, the specific steps in estimating Precision are:

1. repeat steps 1 a., b., and c. a total of B times:

   a. Generate $\mathbf{A}(b)$ as described above with the set of parameters $\hat{\boldsymbol{\psi}} = (\hat{M}_l, \hat{D}_l, \hat{U}_l, \hat{R}_l)$ for $l = 1, \ldots, L$ (see Section 2).

   b. Link File X and File Y using exactly the same process (i.e. same deterministic or probabilistic linking algorithm) that was initially used to link File X and File Y but using $\mathbf{A}(b)$ instead of $\mathbf{A}$.

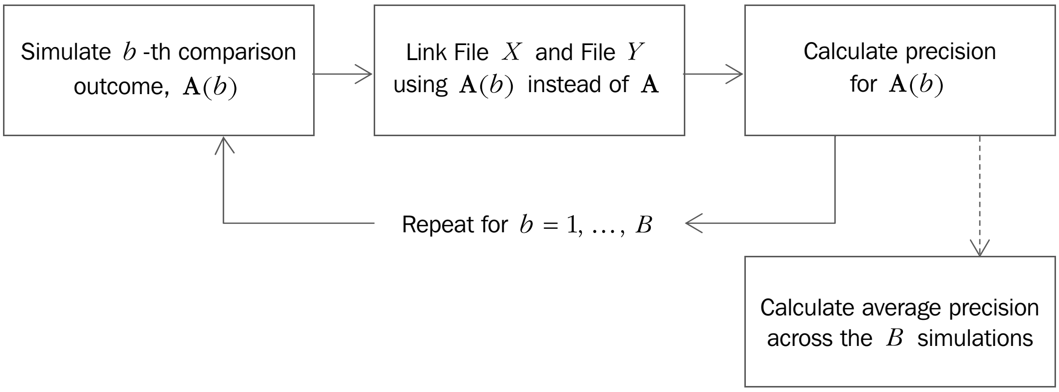
   c. Define $c(b)$ and $n(b)$ to be the number of links that are matches and the number of links, respectively, in the b-th simulation. Calculate the Precision for the b-th simulation by $P(b) = c(b)/n(b)$.

2. Estimate Precision by $\hat{P}_{\text{Prob}} = \sum_{b=1}^{B} P(b)\big/B$.

If there is interest in the Link Rate (see point 1. in Section 1) then in Step 1c. calculate the Link Rate for the b-th simulation by $LR(b) = n(b)/m$ and in Step 2 estimate the Link Rate by $\widehat{LR} = \sum_{b=1}^{B} LR(b)\big/B$. It is straightforward to incorporate into the simulation that a known proportion of records on File X do not have a corresponding match on File Y: only comparison outcomes for non-match record pairs would be generated for this proportion of records. It would also be straightforward to estimate the variability of $\hat{P}_{\text{Prob}}$ using the Bootstrap technique, but we do not consider this issue here.

Figure 4.1 gives an overview of the simulation procedure for probabilistic and deterministic linkage procedures. The figure shows that the key ingredient is the simulation of the comparison matrix, $\mathbf{A}$.

4.1 Overview of simulating the deterministic and probabilistic linkage process



Simulate b -th comparison outcome, $\mathbf{A}(b)$
Link File X and File Y using $\mathbf{A}(b)$ instead of $\mathbf{A}$
Calculate precision for $\mathbf{A}(b)$
Repeat for $b = 1, \ldots, B$
Calculate average precision across the B simulations

The extension to the frequency-based approach of Section 2.2 is now described.

Let $\boldsymbol{\upsilon}_{ik} = \left(\upsilon_{ik1}, \ldots, \upsilon_{ikl}, \ldots, \upsilon_{ikL}\right)$ where υ_{ikl} is the unobserved value for υ_l for the (i,k)-th record pair, so that $i = k$ denotes matches. (υ_{ikl} is unobserved because the k index requires knowledge of matches, which are unobserved.) Let $\boldsymbol{\Delta}$ be the $mn \times L$ matrix with rows given by the $\boldsymbol{\upsilon}_{ik}$. Under the frequency-based approach we now need to simulate $\Delta(b)$ given $\mathrm{A}(b)$, where $\boldsymbol{\Delta}(b)$ has rows given by $\boldsymbol{\upsilon}_{ik}(b) = \left(\upsilon_{ik1}(b), \ldots, \upsilon_{ikl}(b), \ldots, \upsilon_{ikL}(b)\right)$, and $\upsilon_{ikl}(b)$ is the b-th simulated value of υ_{ikl}. Under (4a) and (4b), the $\upsilon_{ikl}(b)$ are independently simulated over i, k and l in the following way:

For $i = k$ (i.e. for record pairs that are a match):

> If $a_{iil}(b) = 1$ then $\upsilon_{iil} = r$ with probability $\hat{M}_{lr}$ for $r = 1, \ldots, R$.
>
> If $a_{iil}(b) = 0$ then $\upsilon_{iil} = 0$.

For $i \neq k$ (i.e. for record pairs that are non-matches):

> If $a_{ikl}(b) = 1$ then $\upsilon_{ikl} = r$ with probability $\hat{U}_{lr}$ for $r = 1, \ldots, R$.
>
> If $a_{ikl}(b) = 0$ then $\upsilon_{ikl} = 0$.

The steps involved in estimating Precision under the frequency-based approach of Section 2.2 are exactly the same as described above except that both $\mathrm{A}(b)$ and $\Delta(b)$ are simulated (see Step 1a) and then used in the b-th simulated linkage process (see Step 1b).

5. A QUICK WAY OF ESTIMATING PRECISION FOR DETERMINISTIC LINKING

This section describes an algebraic estimator of Precision for deterministic linkage. The deterministic linkage process is reasonably straightforward suggesting that an algebraic estimator is possible. The main benefit of an algebraic approach over the simulation approach of Section 4 is that it is significantly quicker to calculate since the computation associated with simulating $\mathbf{A}$ can be time-consuming.

As was the case in Section 4, the estimate of Precision here is based on the independence assumption and the latent model of Section 2.1. In the simple case of a single pass it is shown in Appendix A.1 that an estimate of Precision for deterministic linkage on $\boldsymbol{a}$ is

$$\hat{P} = \frac{W(1-Q)}{(1-W)(n-1)Q + W(1-Q)} \tag{5}$$

where $W = M_1 \times \ldots \times M_l \times \ldots \times M_L$ and $Q = U_1 \times \ldots \times U_l \times \ldots \times U_L$.

We assumed in (5) that all records on File X have a match on File Y. We may know that only a proportion, D, of records on File X have a matching record on File Y. An estimate of the precision in this case is (see Appendix A.2)

$$\hat{P}_{\text{Dup}} = \frac{W(1-Q)D}{(1-W)(n-1)QD + W(1-Q)D + nQ(1-Q)(1-D)} . \tag{6}$$

In the case of multiple, say T, passes it is not hard to see that Precision depends upon the order in which linking keys are assigned to passes. For example, if all matches were found in pass 1, then no matches can be found in pass 2.

Since each of the T passes has a distinct linking key, we can also index the linking keys by t (e.g. the second linking key is by definition used in pass $t = 2$). Looking at each linking key in isolation, let N_t be the number of links that could be made by the t-th linking key and let P_t be the Precision of these links, where P_t can be estimated by $\hat{P}_t$, which is given by (5) but with $\boldsymbol{a}$ defined by the t-th linking key. While $N_t\hat{P}_t$ is the estimated number of matches that could be made by the t-th linking key, it is a naive estimate of the number of matches made in pass t, because it does not take into account that some or even all of the matches that could have been made by the linking key in the t-th pass were already made in previous passes. Again, it does not take into account if all matches were made in pass 1, then no matches can be made in pass 2.

Now define $\mathscr{L}_t$ to be the number of links made in pass t and let $\mathscr{L}_{kt}$ be the number of links that could have been made by the linking key used in the t-th pass but were in fact made in pass k, where $k = 1, \ldots, t$. For example, consider a linking strategy with eight linking passes and associated linking keys $(T = 8)$. If there are 50 and 100 links that could be made by the linking key used in the first and eighth pass, respectively $(N_1 = 50, N_8 = 100)$ and if 50 of these records are linked in pass 1 and 50 are linked in pass 8, then $\left(\mathscr{L}_{18}, \ldots, \mathscr{L}_{58}, \ldots, \mathscr{L}_{88}\right) = \left(50, 0, 0, 0, 0, 0, 0, 50\right)$.

Let (MP_t) be the estimated Marginal Precision for records linked in pass t.

For pass 1, $(MP_1) = \hat{P}$. For pass 2 onwards, $(MP_t) = \left\{N_t \hat{P}_t - C_{t-1}\right\} / \mathscr{L}_t$, where

$$C_{t-1} = \sum\nolimits_{k=1}^{t-1} \mathscr{L}_{k(t-1)}(MP_{t-1})$$

is the estimated number of matches that could have been made by the linking key used in pass t but were in fact made using a linking key in a previous pass (i.e. pass $1, \ldots, t-1$). In some cases we needed to impose bounds on (MP_t) so that it remained between 0 and $\hat{P}_t$. The cumulative Precision for the linked file after T passes is estimated by

$$\hat{P}_{(\mathrm{T})} = \frac{\sum_t \mathscr{L}_t(MP_t)}{\sum_t \mathscr{L}_t} . \tag{7}$$

The estimators (5) and (6) are estimators of Precision under the standard approach of Section 2.1. The corresponding estimators to (5) and (6) under the frequency-based approach of Section 2.2 are given by (8) and (9). For a link agreeing on $\boldsymbol{\upsilon}$, the frequency-based estimators of Precision are

$$\hat{P}^{\mathrm{Freq}} = \frac{W^*\left(1-Q^*\right)}{\left(1-W\right)\left(n-1\right)Q^* + W^*\left(1-Q^*\right)} , \tag{8}$$

$$\hat{P}^{\mathrm{Freq}}_{\mathrm{Dup}} = \frac{W^*\left(1-Q^*\right)D}{\left(1-W\right)\left(n-1\right)Q^*D + W^*\left(1-Q^*\right)D + nQ^*\left(1-Q^*\right)\left(1-D\right)} \tag{9}$$

where W^* is the same as W except that M_l is replaced by $M_l M_{lr}$ and Q^* is the same as Q except that U_l is replaced by $U_l U_{lr}$ for all l and r (see Appendix A.3 for proof). We found that frequency-based estimates of Precision, on a pass-by-pass basis, were much more accurate than the standard estimates (see figure 6.5 in Section 6). This makes sense because, for example, knowing the specific country of birth on which a link agrees can be very useful information when predicting the probability that the link is a match. However, when looking at the overall average Precision, the gains from the frequency-based approach were marginal.

6. SIMULATION STUDY

This simulation study shows that the estimators of Precision for deterministic and probabilistic linkage are accurate when the underlying model for the comparison outcomes is known.

6.1 The data

File Y comprises 400,000 records. The blocking and linking variables are listed in tables 6.1 and 6.2. For a record, the value of each variable is generated independently (e.g. the value for Eye colour is independent of Birth year). With the exception of Country of birth (COB), each value of a variable is equally likely (e.g. each eye colour is equally likely). For COB, 76% of records are assigned "Born in Australia" and the remaining 24% of records are randomly assigned one of about 300 country codes, where the probability of being assigned a particular code is equal to the proportion of those people in the 2006 Census with that country code. Another version of COB, called COBO, was created where COBO=COB except that for people born in Australia, COBO was set to "missing".

File X is a random subsample of 50,000 records from File Y. Initially, each record on File X has the same value as its matching record on File Y. Some values on File X may be changed in order to simulate errors in linking fields. If a value for a variable on File X is flagged to be changed, its replacement value was either chosen completely at random from records on File Y or it was set to 'missing'. Table 6.1 summarises the frequency with which these errors occur. For example, the probability of a matching and non-matching pair agreeing on Birthday is 0.8924 and 0.0025, respectively, while the probability that a record pair (matches and non-matches) has a 'missing' Birthday is 0.0892.

6.1 Construction of synthetic data

		Match latent class			*Non-match latent class*		
	Variables	*agree*	*disagree*	*missing*	*agree*	*disagree*	*missing*
MB	Mesh Block	0.9216	0.0784	0.0000	0.0002	0.9998	0.0000
SA1	SA1	0.9504	0.0496	0.0000	0.0010	0.9990	0.0000
BDAY	Birthday	0.8924	0.0184	0.0892	0.0025	0.9083	0.0892
BYEAR	Birth year	0.9400	0.0500	0.0100	0.0180	0.9720	0.0100
	Birth year (±1)	0.9880	0.0020	0.0100	0.0540	0.9360	0.0100
EYE	Eye colour	0.8000	0.1000	0.1000	0.1800	0.7200	0.1000
SEX	Sex	0.9990	0.0010	0.0000	0.5000	0.5000	0.0000
COB	Country of birth	0.9700	0.0100	0.0200	0.5575	0.4225	0.0200
COBO—COB	(recode)	0.2350	0.0050	0.7600	0.0033	0.2367	0.7600

6.2 Probabilistic linkage

Table 6.2 shows four alternative probabilistic linkage strategies, where 'Link' indicates linking variables and 'Block' denotes blocking variables. We see across each of the four strategies that Run 2 has by far the greatest number of blocks, containing on average 2.6 records on File X and 18 records on File Y. For each Run, File X and File Y were probabilistically linked using the computer package *Febrl* at a range of cut-off values and where ψ was obtained from table 6.1 and assumed to be known. Figure 6.3 plots the resulting Precision (*actual*) and Link Rate (*actual*) for a range of cut-off values. The Link Rate and Precision were then estimated at a range of cut-off values using the simulation process of Section 4 (standard approach of Section 2.1), using the parameters ψ from table 6.1 with $B = 10$.

6.2 Blocking and linking strategies

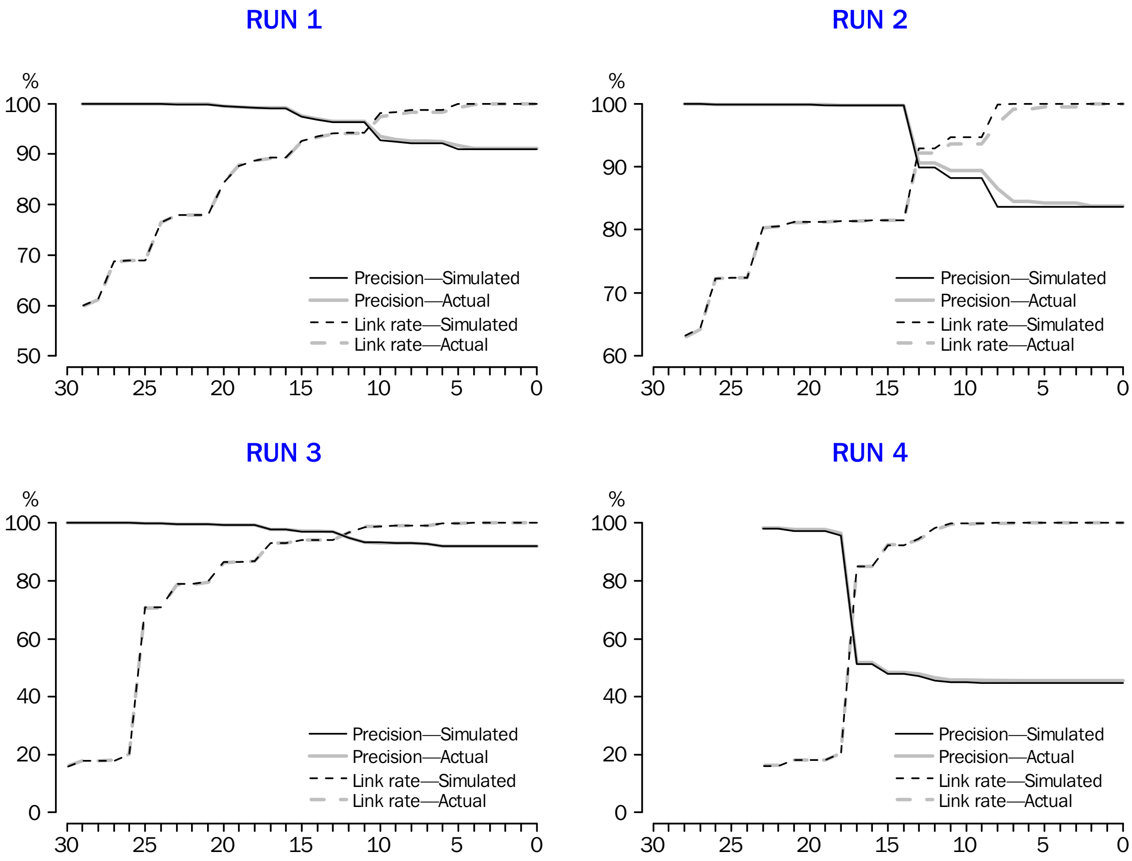
		RUN 1	*RUN 2*	*RUN 3*	*RUN 4*
MB	Mesh Block	Block	—	—	—
SA1	Statistical Area 1	—	Link	Block	—
BDAY	Birthday	Link	Block	Link	Block
BYEAR	Birth year	Link	Block	Link (±1)	Link
EYE	Eye colour	Link	Link	Link	Link
SEX	Sex	—	Link	Block	Block
COB	Country of birth	Link	Link	—	—
COBO—COB	Country of birth (recode)	—	—	Link	Link
	Number of Blocks	5,000	18,488	2,000	732
	Average File A block size	10	2.6	25	67
	Average File B block size	80	18	200	502

Block = Blocking field; Link = Linking field

Figure 6.3 shows that as we lower the cut-off, the actual Link Rate monotonically increases to close to 100% and the actual Precision monotonically decreases from 100%.

The use of geographical blocking fields in Run 1 and Run 3 ensures that precision remains high, even for low cut-off weights, although matches that disagree on the blocking variables are necessarily excluded. From Run 2 it is apparent that precision falls substantially for those links that disagree on SA1. In Run 4, no geographical variables are used – either as blocking or linking variables – and consequently very few high precision links are found.

6.3 Probabilistic linkage: Link Rate and Precision, by weight cutoff



RUN 1

RUN 2

RUN 3

RUN 4

The highly-skewed Country of Birth linking field is inherently inconsistent with the simplified assumptions that underlie both the probabilistic linking strategy and the standard model used to simulate precision. This accounts for the poorer performance of the simulation-based estimators at lower cut-off weights, especially in Runs 1 and 2. The use of the recoded Country of Birth variable in Runs 3 and 4 has the effect of reducing this distortion, while also raising the precision of links for the 24% of the population that were not born in Australia.

Of the four blocking and linking strategies, Run 3 is inclusive of the highest proportion of links and displays the most favourable trade-off between precision and link rate – the Link Rate and Precision lines intersect at about 97%, corresponding to a cut-off weight of about 11. In addition, the simulation-based estimators of Precision and Link Rate under the standard approach are highly accurate– the estimates sit on top of the actual or true values across the range of cut-off values.

This simulation shows that Precision and Link Rate can be estimated accurately using the simulation approach when the assumption of conditional independence holds and ψ is known.

6.3 Deterministic linkage

We considered a linkage strategy with 500 deterministic passes, where the linking key in each pass was created by combining between two and six of the variables in table 6.1. The linking key with the highest Precision, as estimated by equation (5), was assigned to pass 1, the linking key with the second highest Precision was assigned to pass 2, and so on. Figure 6.4 shows that the quick estimate of cumulative Precision (see equation (7)) using the frequency-based approach is close to the actual or true value.

Figure 6.4 shows that the Link Rate and Precision lines intersect at about 96%, which is slightly less than the 97% under the best of the probabilistic linking strategies (Run 3). Figure 6.4 shows that the estimates of Precision are very close to the true Precision.

6.4 Deterministic linkage: Link Rate and Precision, by the number of links

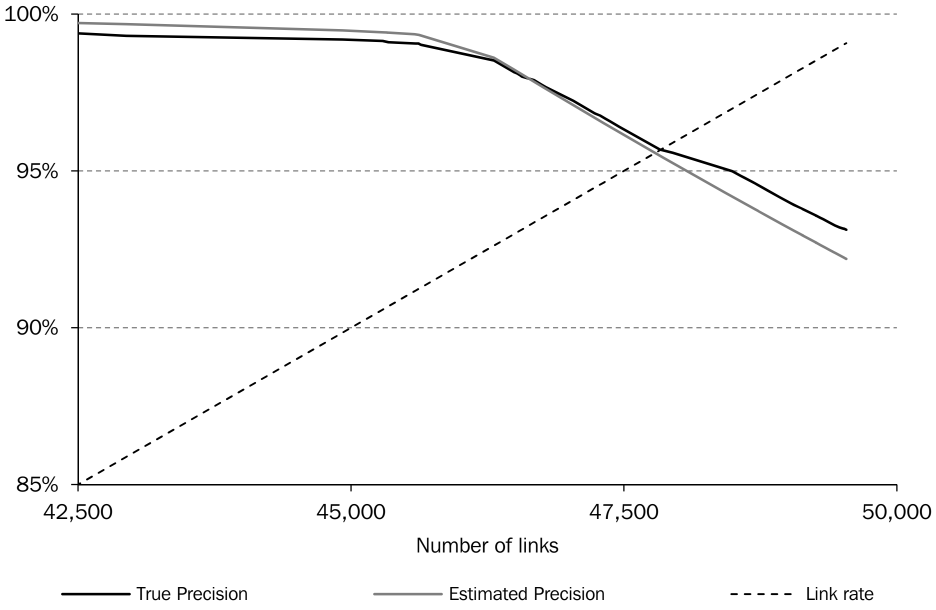
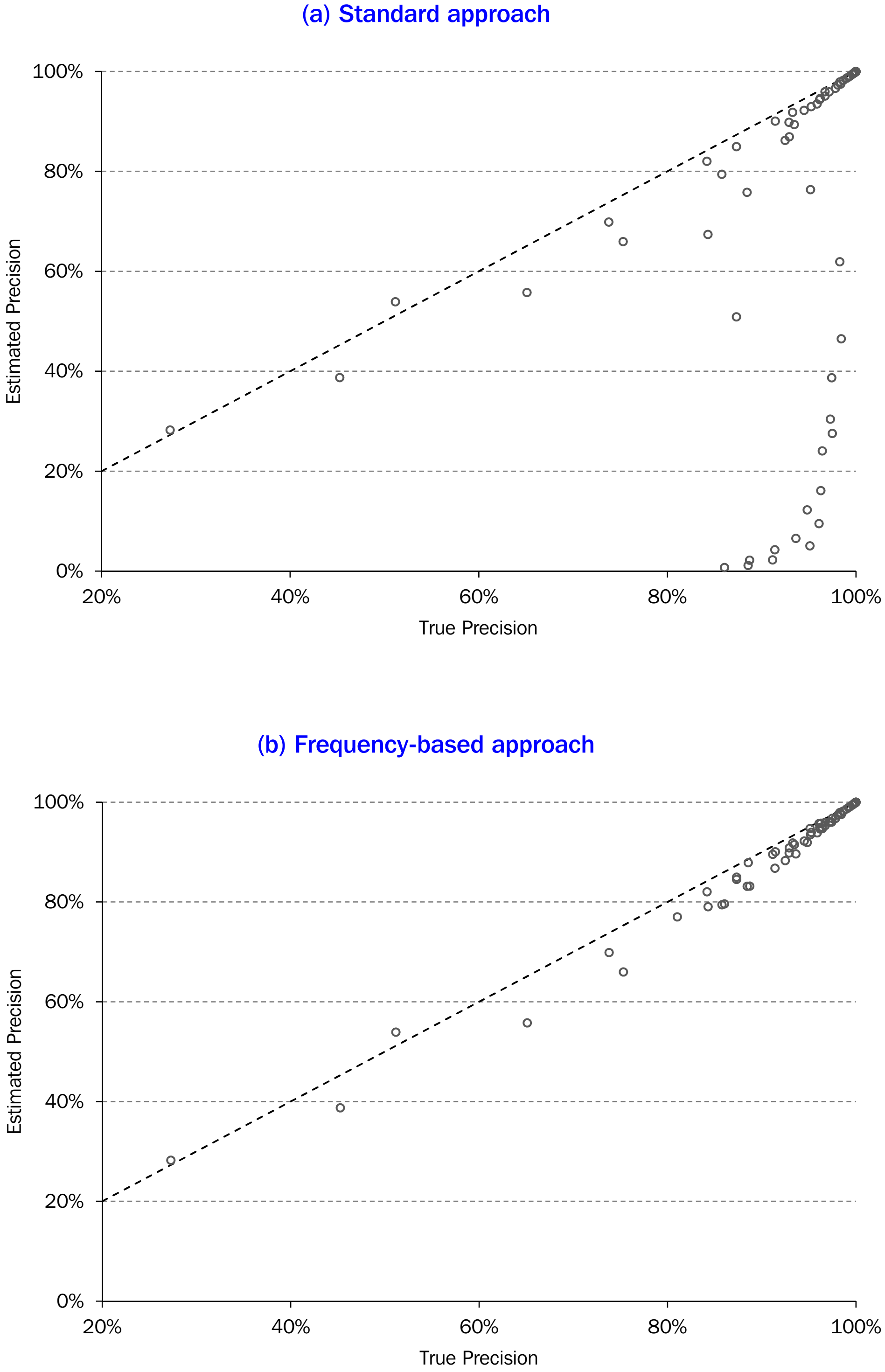


Figure 6.5 compares the standard (Section 2.1) and frequency-based (Section 2.2) approaches to estimating precision for the 500 deterministic linkage keys. Figure 6.5(a) shows that for some linkage keys, the estimates (circles in figure) of precision under the standard approach (equation (5)) are noticeably lower than the true precision, given by the red line. Figure 6.5(b) shows that this issue appears to be resolved under the frequency-based approach, using equation (8). Our investigations suggest that if only an overall estimate of precision is of interest, then the standard and frequency-based approaches appear to perform equally well.

6.5 Comparison of the standard and frequency-based approaches to estimating precision for deterministic linkage with different linkage keys


(a) Standard approach

(b) Frequency-based approach

7. EMPIRICAL STUDY

The Census of Population and Housing was linked to deaths registered during 2011–12 following the Census reference night. The primary aim of the linkage was to assess the consistency of Indigenous Status reported in death registrations and in Census data. This is an important input into the compilation of life tables and life expectancy estimates for Aboriginal and Torres Strait Islander people (ABS, 2013b; ABS, 2016).

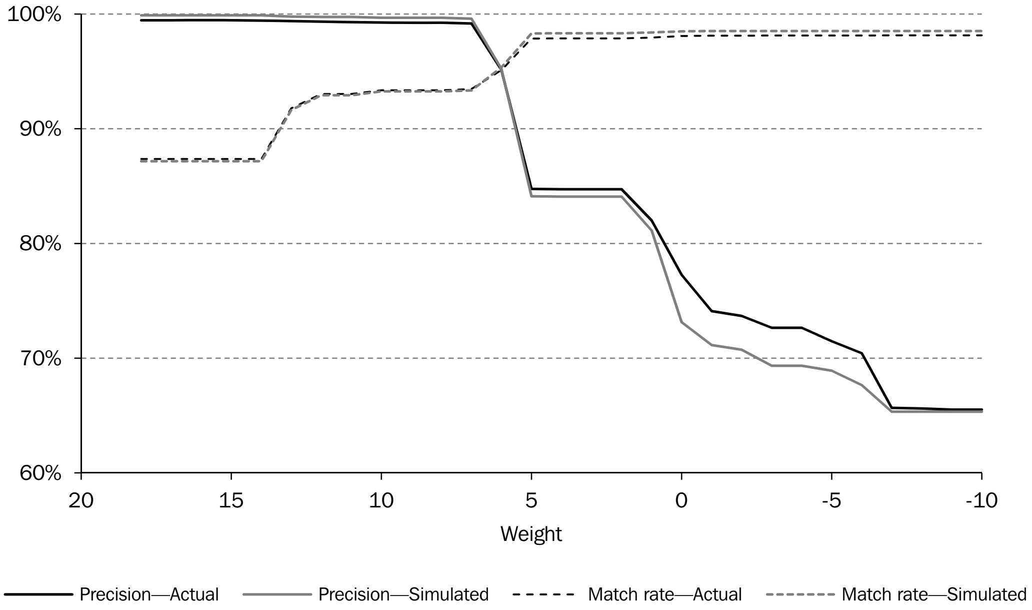
The linkage was performed in two ways. The Bronze probabilistic and deterministic linkage strategies used *Mesh Block*, *Birthday*, *Birth Year*, *Sex*, *Marital Status* and *Country of Birth*. The Gold linkage used all the Bronze linkage variables as well as name and address. We assume here for the purpose of this investigation that the Gold linkage is perfect (i.e. all matches are linked and no non-matches are linked). Under this assumption the True Precision and Link Rate of the Bronze linkage strategy can be calculated.

Sections 7.1 and 7.2 measure the accuracy of the estimates of Precision for probabilistic and deterministic linkage with the Bronze files, respectively. Both estimators in these sections make the independence assumption described in Sections 2.1 and 2.2 and the parameter, ψ, was estimated using a standard EM algorithm (for details see Herzog, 2007). Section 7.3 uses the Gold file to comment on the validity of the independence assumption made by the estimators in Sections 7.1 and 7.2.

7.1 Probabilistic linkage: Death Registrations to Census Linkage

The Bronze linking strategy used *Mesh Block* as a blocking variable and *Birthday*, *Birth Year*, *Sex*, *Marital Status*, *Country of Birth*. The weights were calculated under the independence assumption, and records were linked using 1–1 assignment at a range of different cut-off values. The Precision and Match Rate for the Bronze linked file were estimated using the method proposed in Section 4 under the standard approach, for a range of cut-off values. Figure 7.1 shows that, in general, the estimates of Precision and Match Rate track the true values very well. At its worst, at the cut-off value of 0, the estimates of Precision were about 4% higher than the true value. This is a surprisingly good result.

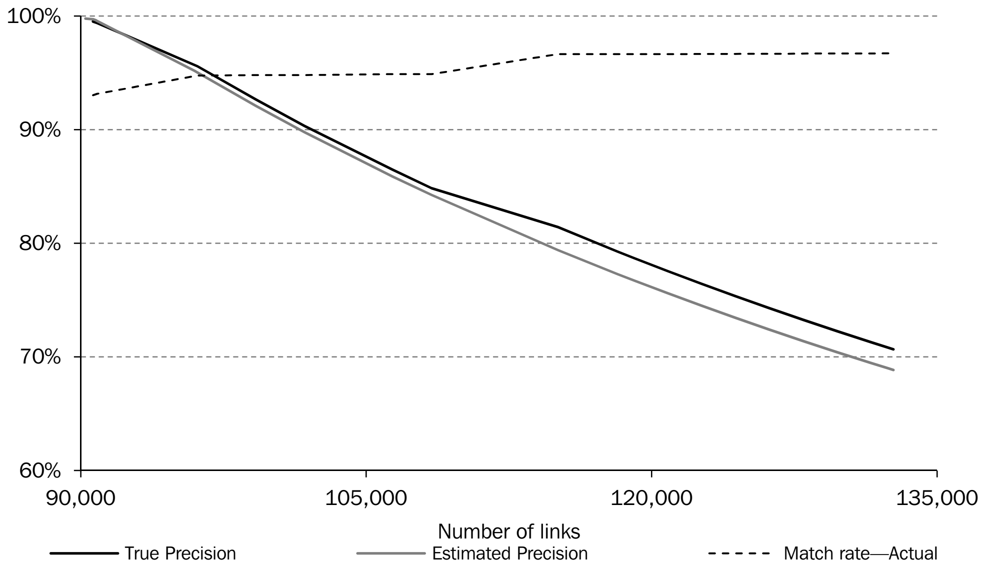
7.1 Probabilistic linkage: Link Rate and Precision, by weight cutoff for the Death Registrations to Census Linkage



7.2 Deterministic linkage: Death Registrations to Census Linkage

The deterministic approach used 32 passes, based on a different combinations of the linking variables available on the Bronze files. The estimates of Precision under the frequency approach are summarised in figure 7.2. Again, the estimate of Precision tracks the true Precision very well. Though the details are not provided, the estimates under the standard approach were only marginally less accurate than under the frequency-based estimates.

7.2 Deterministic linkage: Precision, for Death Registrations to Census Linkage



7.3 Validity of the Independence assumption

The estimates of Precision in Section 7.1 and 7.2 make the independence assumption, which implicitly fits log-linear models with only 1-way effects to the comparison outcomes for matches (see equation (2a)) and non-matches (see equation (2b)). To test the appropriateness of (2a) in particular, we fitted a log-linear model with all two-way and three-way interactions to the agreement outcomes for matches, as identified by the Gold file. Based on the Likelihood Ratio test, the two-way and three-way model was preferred to the one-way model. Also, some of the two and three-way interaction terms were larger and more statistically significant compared with some of the one-way (or main) effects.

The proposed estimators of Precision perform very well despite the fact that they are explicitly based on the independence assumption that is shown to be strongly violated. This suggests that the proposed estimators are robust against violations of the independence assumption.

8. SUMMARY AND CONCLUSIONS

This paper develops and evaluates an estimator of precision for probabilistic and deterministic linkage strategies. These estimators are based on the strong assumption that comparison outcomes on the different linking variables are independent. Nevertheless these estimators perform very well in an empirical study and in simulations, when this assumption is violated. This suggests that the estimators are robust against moderate violations of these assumptions. It is envisaged that these methods will be used as part of the quality assurance process for record linkage at the Australian Bureau of Statistics.

REFERENCES

Australian Bureau of Statistics (2013a) *Australian Census Longitudinal Dataset, Methodology and Quality Assessment, 2006-2011*, Information Paper, cat. no. 2080.5, ABS, Canberra.
< http://www.abs.gov.au/ausstats/abs@.nsf/mf/2080.5 >

—— (2013b) *Death Registrations to Census Linkage Project - Methodology and Quality Assessment, 2011-2012*, Information Paper, cat. no. 3302.0.55.004, ABS, Canberra.
< http://www.abs.gov.au/ausstats/abs@.nsf/mf/3302.0.55.004 >

—— (2016) "Death Registrations to Census Linkage Project - A Linked Dataset for Analysis", *Methodology Research Papers*, cat. no. 1351.0.55.058, ABS, Canberra.
< http://www.abs.gov.au/ausstats/abs@.nsf/mf/1351.0.55.058 >

Australian Institute of Health and Welfare (2013) *Statistical Linkage Key 581 Cluster*, website content from METeOR Metadata Online Registry, AIHW, Canberra.
< http://meteor.aihw.gov.au/content/index.phtml/itemId/349510 >

Chipperfield, J.O.; Bishop, G. and Campbell, P. (2011) "Maximum Likelihood Estimation for Contingency Tables and Logistic Regression with Incorrectly Linked Data", *Survey Methodology*, 37(1), pp. 13–24.

Chipperfield, J.O. and Chambers, R.L. (2015, to appear) "Using the Bootstrap to Analyse Binary Data Obtained Via Probabilistic Linkage", *Journal of Official Statistics*.

Christen, P. and Churches, T. (2005) "Febrl – Freely Extensible Biomedical Record Linkage", Release 0.3.1.
< http://cs.anu.edu.au/~Peter.Christen/Febrl/febrl-0.3/febrldoc-0.3/contents.html >

Fellegi, I.P. and Sunter, A.B. (1969) "A Theory for Record Linkage". *Journal of the American Statistical Association*, 64(328), pp. 1183–1210.

Herzog, T.N.; Scheuren, F.J. and Winkler, W.E. (2007) *Data Quality and Record Linkage Techniques*, Springer, New York.

Lahiri, P. and Larsen, M.D. (2005) "Regression Analysis with Linked Data", *Journal of the American Statistical Association*, 100(469), pp. 222–230.

Larsen,M.D. and Rubin, D.B. (2001) "Iterative Automated Record Linkage Using Mixture Models", *Journal of the American Statistical Association*, 96(453), pp. 32–41.

Särndal, C.; Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*, Springler-Verlag, New York.

Winkler, W.E. (2001), "Record Linkage Software and Methods for Merging Administrative Lists", *Statistical Research Report Series,* No. RR2001/03, U.S. Bureau of the Census.
< https://www.census.gov/srd/papers/pdf/rr2001-03.pdf >

Winkler, W.E. (2005) "Approximate String Comparator Search Strategies for Very Large Administrative Lists", *Statistical Research Report Series,* No. RRS2005/02, U.S. Bureau of the Census.
< https://www.census.gov/srd/papers/pdf/rrs2005-02.pdf >

< All URLs last viewed on 20 January 2017 >

APPENDIX

A.1 Estimating precision from a single pass under standard approach

We know,

$$\begin{aligned}
&\Pr\left(\text{All linking fields for the } (i,j)\text{-th pair agree} \mid (i,j)\text{-th pair is the match}\right)\\
&= M_1 \times \ldots \times M_l \times \ldots \times M_L\\
&= W\,.
\end{aligned}$$

$$\begin{aligned}
&\Pr\left(\text{All linking fields for the } (i,j)\text{-th pair agree} \mid (i,j)\text{-th pair is not the match}\right)\\
&= U_1 \times \ldots \times U_l \times \ldots \times U_L\\
&= Q\,.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\Pr\left(\text{Link} \mid \text{Link is the match}\right)\\
&= \Pr\left(\text{Only one record pair agrees on all linking fields and this pair is the match}\right)\\
&= \Pr\left(\text{Match record pair agrees on all linking fields}\right)\\
&\qquad \times \Pr\left(\text{No non-match record pair agrees on all linking fields}\right)\\
&= W \times \left(1-Q\right)^{(n-1)}.
\end{aligned}$$

and that

$$\begin{aligned}
&\Pr\left(\text{Link} \mid \text{Link is a non-match}\right)\\
&= \Pr\left(\text{Only one record pair agrees on all linking fields and this pair is not the match}\right)\\
&= \left[1-\Pr\left(\text{Match record pair agrees on all linking fields}\right)\right]\\
&\qquad \times \ \Pr\left(\text{Only one non-match record pair agrees on all linking fields}\right)\\
&= \left(1-W\right) \times \left(n-1\right) Q \left(1-Q\right)^{(n-2)}.
\end{aligned}$$

From Bayes Theorem, the Precision is given by

$$\begin{aligned}
\Pr\left(\text{Match} \mid \text{Link}\right) &= \frac{\Pr\left(\text{Link} \mid \text{Link is the match}\right)}{\left[\Pr\left(\text{Link} \mid \text{Link is the match}\right) + \Pr\left(\text{Link} \mid \text{Link is a non-match}\right)\right]}\\
&= \frac{W \times \left(1-Q\right)}{\left[W \times \left(1-Q\right) + \left(1-W\right) \times \left(n-1\right) Q\right]}
\end{aligned}$$

A.2 Estimating precision when only a proportion, D, of matches exist

Let D be the proportion of records on File X that have a corresponding match on File Y. We assume *that the process of identifying records without a match is completely random.*

First,

$$\begin{aligned}
&\Pr\left(\text{Link}\,\middle|\,\text{Link is the match}\right)\\
&= \Pr\begin{pmatrix}\text{Only one record pair agrees on all linking}\\ \text{fields and this record pair is the match}\end{pmatrix}\\
&= \Pr\left(\text{Match record pair agrees on all linking fields}\right)\\
&\qquad \times\ \Pr\left(\text{No non-match record pair exists agrees on all linking fields}\right)\\
&\qquad \times\ \Pr\left(\text{Matching record exists}\right)\\
&= W\times\left(1-Q\right)^{(n-1)}\times D\,.
\end{aligned}$$

Second,

$$\begin{aligned}
&\Pr\left(\text{Link}\,\middle|\,\text{Link is a non-match}\right)\\
&= \Pr\begin{pmatrix}\text{Only one record pair agrees on all linking}\\ \text{fields and this record pair is not the match}\end{pmatrix}\\
&= \Pr\left(\begin{matrix}\text{Only one pair agrees on all linking}\\ \text{fields and this pair is not the match}\end{matrix}\,\middle|\,\text{matching record exists}\right)\\
&\qquad \times\ \Pr\left(\text{matching record exists}\right)\\
&\quad +\ \Pr\left(\begin{matrix}\text{Only one pair agrees on all linking}\\ \text{fields and this pair is not the match}\end{matrix}\,\middle|\,\text{matching record does not exist}\right)\\
&\qquad \times\ \Pr\left(\text{matching record does not exist}\right)\\
&= \left(1-W\right)\times\left(n-1\right)Q\left(1-Q\right)^{(n-2)}\times D\ +\ nQ\left(1-Q\right)^{(n-1)}\times\left(1-D\right).
\end{aligned}$$

Again using Bayes Theorem, an estimate of the Precision is

$$\frac{W\times\left(1-Q\right)\times D}{W\times\left(1-Q\right)\times D+\left(1-W\right)\times\left(n-1\right)Q\times D\ +\ nQ\left(1-Q\right)\times\left(1-D\right)}\ .$$

A.3 Estimating precision from a single pass under frequency-based approach

We know from (4) that,

$$\begin{aligned}
&\Pr\left(\text{Linking fields for the } (i,j)\text{-th pair agree on } \boldsymbol{\upsilon} \mid (i,j)\text{-th pair is the match}\right)\\
&= M_1^* \times \ldots \times M_l^* \times \ldots \times M_L^*\\
&= W^* .
\end{aligned}$$

$$\begin{aligned}
&\Pr\left(\text{Linking fields for the } (i,j)\text{-th pair agree on } \boldsymbol{\upsilon} \mid (i,j)\text{-th pair is not the match}\right)\\
&= U_1^* \times \ldots \times U_l^* \times \ldots \times M_L^*\\
&= Q^* .
\end{aligned}$$

It follows that,

$$\begin{aligned}
&\Pr\left(\text{Link agrees on } \boldsymbol{\upsilon} \mid \text{Link is the match}\right)\\
&= \Pr\left(\text{Only one record pair agrees on } \boldsymbol{\upsilon} \text{ and this pair is the match}\right)\\
&= \Pr\left(\text{Match record pair agrees on } \boldsymbol{\upsilon}\right) \times \Pr\left(\text{no non-match record pair agrees on } \boldsymbol{\upsilon}\right)\\
&= W^* \times \left(1 - Q^*\right)^{(n-1)} .
\end{aligned}$$

$$\begin{aligned}
&\Pr\left(\text{Link agrees on } \boldsymbol{\upsilon} \mid \text{Link is a non-match}\right)\\
&= \Pr\left(\text{Only one record pair agrees on } \boldsymbol{\upsilon} \text{ and this record pair is not the match}\right)\\
&= \left[1 - \Pr\left(\text{matched record pair agrees on all linking fields}\right)\right]\\
&\qquad \times \ \Pr\left(\text{only one non-match record pair agrees on } \boldsymbol{\upsilon}\right)\\
&= \left(1 - W\right) \times \left(n-1\right) Q^* \left(1 - Q^*\right)^{(n-2)} .
\end{aligned}$$

(noting there is no "*" superscript on the above *W*).

From Bayes Theorem, the Precision is given by

$$\begin{aligned}
\text{Prob}\left(\text{Match} \mid \text{Link agrees on } \boldsymbol{\upsilon}\right) &= \frac{\text{Prob}\left(\text{Link agrees on } \boldsymbol{\upsilon} \mid \text{Link is the match}\right)}{\left[\begin{array}{l}\text{Prob}\left(\text{Link agrees on } \boldsymbol{\upsilon} \mid \text{Link is the match}\right) + \\ \text{Prob}\left(\text{Link agrees on } \boldsymbol{\upsilon} \mid \text{Link is a non-match}\right)\end{array}\right]}\\
&= \frac{W^* \times \left(1 - Q^*\right)}{\left[W^* \times \left(1 - Q^*\right) + \left(1 - W^*\right) \times \left(n-1\right) Q^*\right]}
\end{aligned}$$

This proves (8). The proof of (9) is very similar to the steps taken in (6) and (8) and so we omit the details here.

FOR MORE INFORMATION . . .

INTERNET **www.abs.gov.au** The ABS website is the best place for data from our publications and information about the ABS.

LIBRARY A range of ABS publications are available from public and tertiary libraries Australia wide. Contact your nearest library to determine whether it has the ABS statistics you require, or visit our website for a list of libraries.

INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free
of charge from our website, or purchase a hard copy publication. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice.

PHONE 1300 135 070

EMAIL client.services@abs.gov.au

FAX 1300 135 211

POST Client Services, ABS, GPO Box 796, Sydney NSW 2001

FREE ACCESS TO STATISTICS

All statistics on the ABS website can be downloaded free of charge.

WEB ADDRESS www.abs.gov.au