1352.0.55.120

Australian Bureau of Statistics

Research Paper

Using the EM Algorithm to Estimate the Parameters of the Fellegi-Sunter Model for Data Linking

www.abs.gov.au

Research Paper

Using the EM Algorithm to Estimate the Parameters of the Fellegi-Sunter Model for Data Linking

Carrie Samuels

Analytical Services Branch

Methodology Advisory Committee

18 November 2011, Canberra

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) FRI 03 FEB 2012

ABS Catalogue no. 1352.0.55.120

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Dr Phillip Gould, Analytical Services Branch on Canberra (02) 6252 6307 or email <analytical.services@abs.gov.au>.

USING THE EM ALGORITHM TO ESTIMATE THE PARAMETERS OF THE FELLEGI–SUNTER MODEL FOR DATA LINKING

Carrie Samuels
Analytical Services Branch

QUESTIONS FOR THE COMMITTEE

1. Is the Committee aware of other methods not described in this paper that may be suitable for estimating m and u probabilities?
2. Can the Committee identify a way of modifying the algorithm to calculate m and u probabilities for use with 'method two' of handling missing data?
3. Can the Committee suggest other theoretical or practical issues which may explain Winkler's observation of incorrect convergence when $p < 0.05$?
4. What other investigations would the Committee recommend we undertake to confirm that the EM algorithm is suitable for use in our production environment?

CONTENTS

USING THE EM ALGORITHM TO ESTIMATE THE PARAMETERS OF THE FELLEGI–SUNTER MODEL FOR DATA LINKING

Carrie Samuels
Analytical Services Branch

ABSTRACT

Data linking is the act of linking two or more data files to bring together records which belong to the same individual. Data linking is performed at the Australian Bureau of Statistics (ABS) under the banner of the Census Data Enhancement Project, and involves linking Census data to administrative data sets. This data linking is done under the framework of the Fellegi–Sunter model. The parameters of this model need to be estimated for each linkage project. Previously the ABS has used training data to estimate these parameters, but there are limitations and drawbacks to this method. The use of the Expectation–Maximisation (EM) algorithm to estimate the parameters of the Fellegi–Sunter model is well established in the literature. This paper reviews and consolidates the existing research into using the EM algorithm for this purpose. It also documents the results of empirical work to investigate the behaviour of the algorithm on synthetic data sets where the true match status of the records is known.

1. INTRODUCTION

Data linking is the act of linking two or more data files to bring together records which belong to the same individual. The motivation behind data linking is generally to obtain a richer data set by increasing the number of fields available for analysis. In 2005, the Australian Bureau of Statistics (ABS) released a statement of intention to perform data linking using 2006 Census data. This included the proposed creation of a Statistical Longitudinal Census Dataset (SLCD) to be formed by linking a 5% sample of 2006 Census records to records from subsequent Censuses.

This data linking is done using the Fellegi–Sunter model. The theoretical background to this model is discussed in Section 2. In this section we also present some extensions to the model that have been subsequently developed. The key parameters for the Fellegi–Sunter model are known as m and u probabilities, and it is their estimation which is this paper's primary concern. In Section 2 we also give an intuitive discussion of the importance of these parameters and their interpretation.

The focus of this paper is the use of the Expectation Maximisation (EM) algorithm to estimate m and u probabilities. However before we present the details of this application of the EM algorithm, in Section 3 we outline a range of other methods

which can be used to estimate m and u probabilities. In this section we also discuss the estimation methods which were used for the 2006 linking projects. It becomes clear in this discussion that the EM algorithm has several practical advantages over the methods used previously. The ABS originally intended to use the EM algorithm for the 2006 linking projects but there were implementation difficulties and the work had to be abandoned.

In Section 4 we proceed to a detailed theoretical treatment of the application of the EM algorithm to this problem. We also discuss practical issues which arise in this application. The biggest of these issues is the treatment of missing data, and we propose three different ways to treat missing data within the model. Section 4 also includes a discussion of the use of the EM algorithm for data linking by other statistical agencies, which gives precedent for our current investigations.

Section 5 documents the results of empirical investigations into the behaviour of the algorithm on synthetic data sets. This synthetic data was created for our research into data linking methods, and was designed to have similar properties to real Census data. The advantage of testing with this synthetic data is that we know which record pairs are true matches, which allows us to evaluate the accuracy of the estimated parameters.

In Section 6 we summarise the conclusions of this work, and also outline future research questions which our team will need to address before introducing the EM algorithm into our data linking production environment.

The overarching purpose of this paper is twofold. The first aim is to review and consolidate the existing research around using the EM for data linking purposes, and to identify potential avenues for future research. The second aim is to use empirical investigations to inform a decision on whether this work is worth pursuing in practice. The use of the EM algorithm is likely to form a significant part of our future data linking work program, and this paper is intended to lay the foundations for that work.

2. THE FELLEGI–SUNTER METHOD

Fellegi and Sunter (1969) formalised what has become known as the Fellegi–Sunter method for data linking. We give a brief outline and discussion of the method here; the reader is referred to their paper for the full details. We also discuss some extensions to their method which have been subsequently developed.

2.1 Outline of the method

The Fellegi–Sunter method is used to link two files belonging to two populations A and B, with elements or individuals a and b respectively. In the original notation, the records corresponding to the members of A and B are denoted $\alpha(a)$ and $\beta(b)$ respectively. As it is assumed that the two populations have some elements in common, the set of all ordered pairs $A \times B$ can be partitioned into two sets:

$$M = \{(a,b) : a = b, a \in A, b \in B\} ,$$
$$U = \{(a,b) : a \neq b, a \in A, b \in B\} .$$

These are called the matched (M) and unmatched (U) sets respectively. In this paper however we will use the term 'non-matched' rather than 'unmatched'. The aim is to decide whether an observed record pair $(\alpha(a), \beta(b))$ corresponds to a pair (a,b) belonging to the matched set or the non-matched set. To simplify the notation, we refer to individual record pairs as r_j and we refer to them as belonging to M or U, although technically it is the pairs of individuals the record pairs correspond to which are actually the elements of the sets. Note that as defined previously, the terms 'matched' and 'non-matched' are used to refer to the true status of record pairs, which is generally unknown. The terms 'linked' and 'non-linked' are used to refer to the status assigned to record pairs by the probabilistic linking process.

The key piece of notation is the comparison vector γ, which we define piecewise as

$$\gamma_i^j = \begin{cases} 1 & \text{if field } i \text{ is identical on both of the records of record pair } r_j \\ 0 & \text{otherwise} , \end{cases}$$

where $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, N$, that is there are n fields and N record pairs.

Then the comparison vector for the j-th record pair is defined

$$\boldsymbol{\gamma}^j = \left[\gamma_1^j, \gamma_2^j, \ldots, \gamma_n^j \right]$$

and the comparison vector over all the record pairs is defined

$$\boldsymbol{\gamma} = \left[\boldsymbol{\gamma}^1, \boldsymbol{\gamma}^2, \ldots, \boldsymbol{\gamma}^N \right].$$

Fellegi and Sunter show that the optimum linkage rule is given by assigning each record pair a weight

$$w_j = \frac{P\left[\boldsymbol{\gamma}^j \mid r_j \in M\right]}{P\left[\boldsymbol{\gamma}^j \mid r_j \in U\right]}.$$

Two cut-offs are then determined such that record pairs with weights above the high cut-off are assigned as links, record pairs with weights below the low cut-off are assigned as non-links, and record pairs with weights in between the two cut-offs are designated possible links and sent for clerical review to ultimately determine whether they are links.

To reduce the number of parameters in the model and simplify the computation, they introduce the following conditional independence assumption:

$$P\left[\boldsymbol{\gamma}^j \mid r_j \in M\right] = \prod_{i=1}^{n} m_i^{\gamma_i^j} \left(1-m_i\right)^{1-\gamma_i^j},$$

$$P\left[\boldsymbol{\gamma}^j \mid r_j \in U\right] = \prod_{i=1}^{n} u_i^{\gamma_i^j} \left(1-u_i\right)^{1-\gamma_i^j}.$$

These m_i and u_i are called the m and u probabilities respectively, and are given by

$$m_i = P\left[\gamma_i^j = 1 \mid r_j \in M\right] \text{ and } u_i = P\left[\gamma_i^j = 1 \mid r_j \in U\right].$$

These are the conditional probabilities of observing agreement on field i for a random record pair j, given the record is a match or a non-match respectively.

The conditional independence assumption is equivalent to the following two statements.

1. If two records belong to the same unit, the chance of disagreement on one field is independent of the chance of disagreement on another field.
2. If two records belong to different units, the chance of agreement on one field is independent of the chance of agreement on another field.

This assumption is strong, and while it may be violated in practice, the decision rule remains an effective classifier nonetheless.

Fellegi and Sunter state that it is computationally convenient to work with the base 2 logarithm of the weights, and so under the conditional independence assumption the weight for the j-th record pair is given by

$$W_j = \sum_{i=1}^{n} W_{j,i},$$

where $W_{j,i}$ is the contribution from field i to the overall weight, given by

$$W_{j,i} = \log_2 \left(\frac{m_i^{\gamma_i^j} \left(1-m_i\right)^{1-\gamma_i^j}}{u_i^{\gamma_i^j} \left(1-u_i\right)^{1-\gamma_i^j}} \right).$$

In practice we will always have $m_i \geq u_i$, and so agreement on a field will contribute positively to the overall weight, and disagreement on a field will contribute negatively. Higher m probabilities will give larger absolute weights (positive for agreement and negative for disagreement), and so will lower u probabilities.

2.2 Extensions to the method

We now discuss some extensions to the theory presented in the original Fellegi–Sunter paper. While these ideas have occurred elsewhere in the literature, we are not aware of a definitive reference for them. The material in this part is largely based on our own data linking manual (Australian Bureau of Statistics, 2011).

In practice, it is not computationally feasible to calculate weights for every single record pair. Instead a blocking strategy is used to identify a smaller set of plausible links for evaluation. For example blocking on date of birth means that only those record pairs that agree on full date of birth are evaluated. In practice, multiple linking passes can be made using different blocking strategies so that records with missing or incorrect data for the blocking field still have a chance to be compared to their matches.

The weighting method outlined above is still applicable after blocking, but m and u probabilities now need to be calculated conditional on agreement on the blocking fields. That is, the probabilities are now defined as

$$m_i = P\left[\gamma_i^j = 1 \,\middle|\, r_j \in M, \boldsymbol{\gamma}_b^j = \mathbf{1} \right] \quad ,$$
$$u_i = P\left[\gamma_i^j = 1 \,\middle|\, r_j \in U, \boldsymbol{\gamma}_b^j = \mathbf{1} \right] ,$$

where $\boldsymbol{\gamma}_b^j$ is the subset of the j-th comparison vector which refers to the blocking fields, and $\mathbf{1}$ is a vector of ones. We henceforth refer to such probabilities as 'blocking- dependent' m and u probabilities, in contrast with the 'global' probabilities described earlier. In practice, blocking-dependent m probabilities tend to be marginally higher than their corresponding global m probabilities. This will be the case if records with missing data on blocking fields are likely to have missing data on other fields as well. This effect is more pronounced when several blocking fields are used simultaneously. Also, if initials are used for blocking, the blocking-dependent m probabilities for first name and surname may be higher than the global m probabilities, because they are now conditional on part of each name agreeing.

The relationship between blocking-dependent u probabilities and their global counterparts is less predictable. When the linking field has a predictable relationship with the blocking field, this can affect the blocking-dependent u probability. For instance, the blocking dependent u probability for surname when blocking on mesh block (a very fine geographic classification) is higher than its global u probability, because the presence of multiple family members with the same surname within a mesh block increases the chance of agreement on surname in non-matched pairs. In fact, when using any geographical variable for blocking, non-matched pairs may also be more likely to agree on age, and even educational level, because of the tendency for people of similar age and education to live in the same neighbourhood.

A further issue that arises is how to assign weights when the record pair is missing data for a field on one or both of the records. This is a key issue for us in practice, because all the data sets we use for linking feature missing data to various extents. Despite this, the issue is not widely addressed in the published literature. There are three possible approaches we have identified. A crude approach that requires no extension to the framework is to treat missingness as disagreement, that is assign $\gamma_i^j = 0$ if field i is missing on one or both records. However this approach results in negative field weights when there is missing data, which is perhaps an undesirable penalty. Conceptually it may be desirable to assign a zero weight for field i when it is missing on one or both records. More formally, this is an assumption of *non-informative* missingness; that is to say observing a missing field on one or both records of a record pair adds no information that the record pair is a match, or that it is a non-match.

The second method for handling missing data assigns zero weight in line with this reasoning. This is achieved within the present framework by re-defining m and u probabilities conditional on field i being non-missing on both records:

$$m_i = P\left[\gamma_i^j = 1 \,|\, r_j \in M, \text{ field } i \text{ non-missing on both records of } r_j\right],$$
$$u_i = P\left[\gamma_i^j = 1 \,|\, r_j \in U, \text{field } i \text{ non-missing on both records of } r_j\right].$$

Under this framework, the contribution to the weight from field i is now given by

$$W_{j,i} = \begin{cases} 0 & \text{if field } i \text{ is missing on one or both records of } r_j \\ \log_2\left(\dfrac{m_i^{\gamma_i^j}\left(1-m_i\right)^{1-\gamma_i^j}}{u_i^{\gamma_i^j}\left(1-u_i\right)^{1-\gamma_i^j}}\right) & \text{otherwise.} \end{cases}$$

This was the approach which was used for most 2006 linking projects (Conn and Bishop, 2006).

A third approach is to adopt a three comparison value approach. In this approach, rather than introducing an assumption about the missing data, the missingness is accounted for explicitly. This allows missing data to give information on whether or not two record pairs are match. Under this framework, the components of γ are no longer binary. Instead we have

$$\gamma_i^j = \begin{cases} -1 & \text{if field } i \text{ is missing on one or both of the records of record pair } r_j \\ 1 & \text{if field } i \text{ is identical on both of the records of record pair } r_j \\ 0 & \text{otherwise.} \end{cases}$$

The previous definitions extend straightforwardly to this framework. The conditional independence assumption is now

$$P\left[\boldsymbol{\gamma}^j \mid r_j \in M\right] = \prod_{i=1}^{n} m_{a,i}^{I\left[\gamma_i^j=1\right]} m_{d,i}^{I\left[\gamma_i^j=0\right]} m_{m,i}^{I\left[\gamma_i^j=-1\right]} ,$$
$$P\left[\boldsymbol{\gamma}^j \mid r_j \in U\right] = \prod_{i=1}^{n} u_{a,i}^{I\left[\gamma_i^j=1\right]} u_{d,i}^{I\left[\gamma_i^j=0\right]} u_{m,i}^{I\left[\gamma_i^j=-1\right]} .$$

The subscripts a,d,m denote the probabilities relate to agreement, disagreement, or missingness. Under the previous definition, the m and u probabilities related to agreement, with the disagreement probabilities given by $1-m$ and $1-u$ respectively. Now that the outcome is no longer binary, we define probabilities for each of the three comparison values explicitly, although note that since they are constrained to sum to 1 it is strictly only necessary to explicitly define two.

The definitions are now:

$$m_{a,i} = P\left[\gamma_i^j = 1 \mid r_j \in M\right] \quad , \quad u_{a,i} = P\left[\gamma_i^j = 1 \mid r_j \in U\right] ,$$
$$m_{d,i} = P\left[\gamma_i^j = 0 \mid r_j \in M\right] \quad , \quad u_{d,i} = P\left[\gamma_i^j = 0 \mid r_j \in U\right] ,$$
$$m_{m,i} = P\left[\gamma_i^j = -1 \mid r_j \in M\right] \quad , \quad u_{m,i} = P\left[\gamma_i^j = -1 \mid r_j \in U\right] ,$$
$$1 = m_{a,i} + m_{d,i} + m_{m,i} \quad , \quad 1 = u_{a,i} + u_{d,i} + u_{m,i} ,$$

so the contribution to the weight from field i under this framework is given by

$$W_{j,i} = \log_2\left(\frac{m_{a,i}^{I\left[\gamma_i^j=1\right]} m_{d,i}^{I\left[\gamma_i^j=0\right]} m_{m,i}^{I\left[\gamma_i^j=-1\right]}}{u_{a,i}^{I\left[\gamma_i^j=1\right]} u_{d,i}^{I\left[\gamma_i^j=0\right]} u_{m,i}^{I\left[\gamma_i^j=-1\right]}}\right).$$

Note that under this framework a zero weight for missing values on field i will be given if $m_{m,i} = u_{m,i}$. As is clear from the definition of these probabilities, this will occur if missingness is non-informative. Note however that even if this occurs, the approach is still distinct from method two. In both cases zero weight is given for missingness, but method two achieves this by conditioning on having non-missing data, rather than explicitly defining $m_{m,i}$ and $u_{m,i}$.

If method three is used and zero weights for missingness are desired, the condition $m_{m,i} = u_{m,i}$ can be imposed as a constraint. As per our earlier comments, this imposes the assumption of non-informative missingness. This approach was taken in the 2006 Deaths–Census linking project (Wright, 2010). Imposing this constraint also makes this approach easier to implement, for reasons which will be discussed in Section 4.3.

Appendix A contains a simple example which illustrates the three approaches and how they differ in practice. An important point to note is that the m probabilities for *agreement* are the same under method 3 as under method 1, as are the u probabilities for agreement. The m probability for disagreement under method 1, $1-m_i$ is equal to the sum of the m probabilities for disagreement and missing under method 3, i.e. $1-m_i = m_{d,i} + m_{m,i} = 1-m_{a,i}$, and likewise for the u probabilities, i.e. $1-u_i = u_{d,i} + u_{m,i} = 1-u_{a,i}$.

2.3 Interpretation of *m* and *u* probabilities

It is worth commenting on the interpretation of the m and u probabilities. Recall the m probability for a field is the probability a record pair which is a match will agree on that field. m probabilities depend both on the quality of the data sets, and the tendency for the value of the field to change over time. An m probability of 1 for a particular linking field means that if a record pair is a match, the records must agree on that field. In practice, errors in the data may result in disagreement on a field even though the pair is a match. Such errors can be caused by mis-reporting, mistakes in the optical character recognition process, or other data entry and processing errors. Similarly, if a field value can change over time, as can marital status for example, then records may disagree on the field even if they are a match.

Recall also that the u probability for a field is the probability that a record pair which is not a match will agree on that field. This is effectively the probability of chance agreement on a field. The u probability for a field depends on the distribution of the responses to the field within the population. For example, the u probability for sex is approximately 0.5, since there is around a 50% chance that two randomly selected people will be of the same sex.

The m and u probabilities are unknown parameters of the model, which need to be estimated in order to use the model for data linking. To know the m and u probabilities exactly we would need to know the exact membership of the M and U sets. Of course the data linking process itself is aimed at determining the M set, so this becomes a circular argument. Suffice to say for now that estimating m and u probabilities is a crucial and non-trivial part of the data linking process. The next section outlines the existing methods for doing this.

3. METHODS FOR ESTIMATING *m* AND *u* PROBABILITIES

It is in fact straightforward to obtain good estimates of u probabilities from the data to be linked. In this section we outline how this can be done. The m probabilities are more difficult to estimate in practice. We discuss three methods that can be used to estimate m probabilities. The first method involves the use of an existing linked data set as training data. Variations on this method were used by the ABS for all 2006 census linking projects. The second method is the 'iterative refinement' method which involves iteratively estimating m probabilities by re-linking the same data sets. The third method is the use of the EM algorithm to find maximum likelihood estimates of the m and u probabilities. This method is the focus of this paper. It was first investigated in the lead-up to the 2006 linking projects, but there were implementation problems and the work had to be abandoned.

3.1 Estimating *u* probabilities from the data sets to be linked

Estimating u probabilities directly from the data sets to be linked is a straightforward process. Since u is the probability that a non-matched record pair agrees on the linking field, both global and blocking-dependent u probabilities can be estimated as follows:

$$\hat{u}_{i,\text{global}} = \frac{\text{number of record pairs that agree on linking field } i}{\text{number of record pairs}},$$

$$\hat{u}_{i,\text{block}} = \frac{\text{number of record pairs that agree on linking variable } i \text{ and blocking fields}}{\text{number of record pairs that agree on blocking fields}}.$$

For large comparison spaces it is sufficient in practice to calculate these using a large random sample of record pairs, rather than the full comparison space. These estimates will be biased upwards due to the presence of matches, hence a key assumption is that the number of matches greatly exceeds the number of non-matches, and so this bias is negligible. However, for a tight blocking strategy this assumption may not hold because the number of matches may be a non-negligible proportion of the record pairs. In this case, it is possible to make a straightforward adjustment to account for the presence of matches:

$$\hat{u}_{i,\text{global}} = \frac{\text{number of record pairs that agree on linking field } i - \hat{N}_{\text{match}}\ \hat{m}_{i,\text{global}}}{\text{number of record pairs} - \hat{N}_{\text{match}}},$$

$$\hat{u}_{i,\text{block}} = \frac{\left(\begin{array}{l}\text{number of record pairs that agree on linking variable } i \\ \text{and blocking fields } - \hat{N}_{\text{match,block}}\ \hat{m}_{i,\text{global}}\end{array}\right)}{\text{number of record pairs that agree on blocking fields} - \hat{N}_{\text{match,block}}}.$$

Here $\hat{N}_{\text{match}}$ is the estimated number of matches, and $\hat{N}_{\text{match,block}}$ is the estimated number of matches which agree on blocking fields. The term $\hat{N}_{\text{match}}\,\hat{m}_{i,\text{global}}$ is used in the numerator because we only need to remove the number of matches that agree on the linking field. The assumption behind this adjustment is that reasonable estimates $\hat{N}_{\text{match}}$, $m_{i,\text{global}}$ or $\hat{N}_{\text{match,block}}$, $\hat{m}_{i,\text{block}}$ (whichever is applicable) are available. Note that an upper bound on $\hat{N}_{\text{match}}$ is given by assuming all records of the smaller file will have a match in the larger file.

These approaches were used to estimate u probabilities for all 2006 Census linking projects. Further details are given by Solon and Bishop (2009), and Wright, Bishop and Ayre (2009).

3.2 Estimating *m* probabilities from training data

One method of estimating m probabilities is to use an existing linked data set as 'training data' to estimate m probabilities for the data sets being linked. Using this training data, both global and blocking dependent m probabilities can be calculated as follows:

$$\hat{m}_{i,\text{global}} = \frac{\text{number of matched pairs that agree on linking field } i}{\text{number of matched pairs}},$$

$$\hat{m}_{i,\text{block}} = \frac{\text{number of matched pairs that agree on linking field } i \text{ and blocking fields}}{\text{number of matched pairs that agree on blocking fields}}.$$

There are two assumptions which underpin the accuracy of this method. The first assumption is that the training data is free of errors. Two types of error can arise in data linking: missed links, where a match that exists is not linked, and false links, where a non-match is linked. If either of these errors occur in the training data this may lead to biased estimates of m probabilities. From this perspective, training data which has been clerically linked is ideal. The second assumption is that the data to be linked is similar to the training data, with respect to data quality. This assumption arises because m probabilities are largely determined by data quality. Training data was used to calculate m probabilities for all the 2006 Census linking projects. Following are some examples of the different ways training data was used for these projects and the practical considerations which arose.

The ABS conducts a post enumeration survey (PES) to measure undercount in the Census. For previous censuses, the PES records were matched to Census records by an entirely clerical process. In 2011, data linking techniques will for the first time be used to augment the process and reduce the clerical burden. The 2006 PES–Census linked file was used to estimate m probabilities for Migrants–Census, and CDR–Census linkages. An advantage of using the PES–Census linkage as training data was

that the well-established clerical process meant the file was known to have minimal errors. There were however several practical disadvantages to this approach. Firstly, not all linking fields for the various linkage projects were collected on the PES. These fields were assigned the m-probabilities of other similar fields as proxies. For example, the linking field 'Language Spoken at Home' was not collected on the PES, but its m probability was estimated via the proxy PES field 'Birthplace'. Another practical disadvantage came about because the PES has relatively high data quality since it is conducted by interviewers. This quality was not representative of the quality of data sets for other linking projects. Thus m-probabilities calculated from PES–Census clerical pairs were biased upwards for such linking projects. Additionally, because only a month passes between Census and PES, the m probabilities for fields which can change over time, for example marital status, were also biased upwards for linkages involving data sets with a longer time lapse between collections. Recognising this, the m probabilities calculated from the PES–Census linkage for the CDR–Census linkage were adjusted downwards according to the results of a simulation study. This simulation approach was discussed by Liaw (2007), and further developed by Hardy (2008).

The PES was not the only source of training data for 2006 linking projects. The 2006 linked data from the Deaths–CDR linkage was used to estimate m probabilities for the Deaths–Census linkage. For some projects, linked data from the first stage of the linking project was used as training data to estimate m probabilities for later stages. The high quality linked CDR–Census dataset formed using name and address (in addition to other linking fields) was used to estimate m probabilities for CDR–Census linking without name and address. For the migrants–Census linkage, linked data from the first three blocking passes was used to estimate m probabilities for the subsequent two blocking passes. Both these scenarios overcame some of the disadvantages of using PES–Census as training data. Deaths–CDR is a very similar linkage to Deaths–Census, and so we would expect the m probabilities for the two to be similar. There is a caveat to this however, in that the CDR is not performed on a random sample of the population, and so may not be representative of the full Census. In the case of CDR–Census, using linked data from earlier blocking passes meant that the training data was representative of the data to be linked, albeit with some bias because the training data contained the more easily matched record pairs. However the first passes of the CDR–Census linkage were still trained using PES–Census data, and the Deaths–CDR linkage was trained using PESDR–CDR data, so these projects still had dependencies on PES. To summarise, in 2006 all the linking projects relied either directly or indirectly on the PES–Census as training data. We consider it worth investigating methods for estimating m probabilities that do not have this reliance.

3.3 Estimating *m* probabilities using an iterative refinement procedure

Newcombe (1988) outlines an 'iterative refinement' procedure for estimating m probabilities. While his treatment is not in the context of the Fellegi–Sunter method, the concept is transferable. The idea is to first perform the linking using 'best guess' m and u probabilities which may, for instance be obtained from another linking project. After the usual clerical review process, this linked data is used to construct improved estimates of m and u probabilities, and the linking is performed again using these new parameters. In theory this process could be repeated several times, but Newcombe suggests a single re-linking is sufficient in practice. In this sense it can be viewed as a special case of using training data, where the first linked file is used to train the second linkage. The underlying assumption here is that the first linked file is of sufficient quality to be used as training data. The major drawback to this method is the need to perform two linkages, including two clerical review processes, to get the one linked file. In particular, to use this method in conjunction with multiple blocking passes would likely be unacceptably time-consuming.

3.4 Estimating *m* and *u* probabilities using the EM algorithm

The obvious drawback of using training data is that it requires the existence of an appropriate file which has already been linked. Clerical matching, as has been historically done for PES–Census, or an iterative refinement approach, are two ways to create such a file, but they are both resource-intensive. Another linked file may be used as training data, but this will give biased results if the data to be linked does not have similar characteristics.

An alternative approach, which uses only the data to be linked, and does not require clerical resources is to use the EM algorithm to find the maximum likelihood estimates of the m and u probabilities. This method is the focus of this paper, and we discuss the underlying theory in detail in the next section.

> Question for the Committee: Is the Committee aware of other methods not described in this paper that may be suitable for estimating m and u probabilities?

4. THE EM ALGORITHM

The EM Algorithm, introduced by Dempster, Laird and Rubin (1977) has become the standard method for maximum likelihood estimation where the model depends on unobserved latent variables. Winkler (1988) formulated the application of the EM algorithm to estimate m and u probabilities for data linking. The first part of this section outlines the technical details of this formulation. The second part is a less technical discussion which aims to impart a more intuitive understanding to the reader. The last two parts constitute a literature review of practical issues which arise when using the EM algorithm to estimate m and u probabilities and of the current use of the EM algorithm by other data linking practitioners.

We do not discuss the underlying theory of the EM algorithm in detail; the best reference remains Dempster *et al.* (1977). However, following is a bare bones explanation to help the unfamiliar reader follow the remainder of this section. The basic idea of the EM algorithm is to solve an incomplete data problem by associating it with a complete data problem with a tractable solution. The EM algorithm maximises the complete data likelihood by iterating between two steps – the Expectation (E) step, and the Maximisation (M) step. The first E step is to calculate the expected value of the complete data likelihood with respect to the unobserved data, conditional on the observed data, and a guess of the maximum likelihood estimates (MLEs) $\boldsymbol{\theta}^m$. This is known as the Q function in the literature, and is a function of the parameters $\boldsymbol{\theta}$. The M step is to maximise the Q function over $\boldsymbol{\theta}$ to obtain a new value for $\boldsymbol{\theta}^m$, which is then used to perform another E step, and so on. The algorithm iterates between the E and M steps until the value of $\boldsymbol{\theta}^m$ converges.

4.1 Technical overview

To be able to use the EM algorithm, we need to specify the complete data likelihood. In this section we continue to use the notation introduced in Section 2.1, but there is also some further notation we need to introduce. For convenience, we denote the vectors of m and u probabilities for all n fields by

$$\boldsymbol{m} = \left[m_1, m_2, \ldots, m_n\right] ,$$
$$\boldsymbol{u} = \left[u_1, u_2, \ldots, u_n\right] .$$

We denote the proportion of all record pairs which belong to the matched set M by p,

i.e. $$p = P(r_j \in M) = \frac{|M|}{N} ,$$

where $|M|$ is the number of record pairs in the set M.

We also define a class indicator function

$$g_j = \begin{cases} 1 & \text{if } r_j \in M \\ 0 & \text{if } r_j \in U \end{cases}.$$

This is an unobserved function indicating whether or not a particular record pair r_j is a a match. For convenience we write $\boldsymbol{g} = [g_1, g_2, \ldots, g_N]$.

The complete data vector is $< \boldsymbol{g}, \boldsymbol{\gamma} >$. For the moment we will assume there is no missing data and the entries of $\boldsymbol{\gamma}$ are binary. In Section 4.3 we extend the framework to deal with missing data, as foreshadowed in Section 2.2. In the data linking context, the agreement vector $\boldsymbol{\gamma}$ is observed, but the class indicator $\boldsymbol{g}$ is unobserved. The complete data likelihood is given by

$$L(\boldsymbol{m}, \boldsymbol{u}, p \mid \boldsymbol{g}, \boldsymbol{\gamma}) = \prod_{j=1}^{N} \left(p \cdot P\left[\boldsymbol{\gamma}^j \mid r_j \in M \right] \right)^{g_j} \left((1-p) P\left[\boldsymbol{\gamma}^j \mid r_j \in U \right] \right)^{1-g_j}.$$

Hence under the conditional independence assumption, the log-likelihood is

$$\begin{aligned} l(\boldsymbol{m}, \boldsymbol{u}, p \mid \boldsymbol{g}, \boldsymbol{\gamma}) = & \sum_{j=1}^{N} g_j \log\left(p \cdot \prod_{i=1}^{n} m_i^{\gamma_i^j} (1-m_i)^{1-\gamma_i^j} \right) \\ & + \sum_{j=1}^{N} (1-g_j) \log\left((1-p) \prod_{i=1}^{n} u_i^{\gamma_i^j} (1-u_i)^{1-\gamma_i^j} \right). \end{aligned}$$

Now the E-step is to calculate

$$\begin{aligned} Q(\boldsymbol{m}, \boldsymbol{u}, p \mid \hat{\boldsymbol{m}}, \hat{\boldsymbol{u}}, \hat{p}) &= E_{g|\gamma}\left[l(\boldsymbol{m}, \boldsymbol{u}, p \mid \boldsymbol{g}, \boldsymbol{\gamma}) \mid \boldsymbol{\gamma}; \hat{\boldsymbol{m}}, \hat{\boldsymbol{u}}, \hat{p} \right] \\ &= \sum_{j=1}^{N} \hat{g}_j \left(\log(p) + \sum_{i=1}^{n} \left[\gamma_i^j \log(m_i) + (1-\gamma_i^j) \log(1-m_i) \right] \right) \\ &\quad + \sum_{j=1}^{N} (1-\hat{g}_j) \left(\log(1-p) + \sum_{i=1}^{n} \left[\gamma_i^j \log(u_i) + (1-\gamma_i^j) \log(1-u_i) \right] \right), \end{aligned}$$

where

$$\hat{g}_j = E\left[g_j \mid \boldsymbol{\gamma}; \hat{\boldsymbol{m}}, \hat{\boldsymbol{u}}, \hat{p} \right] = \frac{\hat{p} \prod_{i=1}^{n} \hat{m}_i^{\gamma_i^j} (1-\hat{m}_i)^{1-\gamma_i^j}}{\hat{p} \prod_{i=1}^{n} \hat{m}_i^{\gamma_i^j} (1-\hat{m}_i)^{1-\gamma_i^j} + (1-\hat{p}) \prod_{i=1}^{n} \hat{u}_i^{\gamma_i^j} (1-\hat{u}_i)^{1-\gamma_i^j}}.$$

The derivation of this expression for $\hat{g}_j$ is as follows:

$$
\begin{aligned}
E\left[g_j \mid \gamma\right] &= 0 \cdot P\left[g_j = 0 \mid \gamma\right] + 1 \cdot P\left[g_j = 1 \mid \gamma\right] \quad \text{(definition of expectation)} \\
&= P\left[g_j = 1 \mid \gamma\right] \\
&= P\left[r_j \in M \mid \gamma\right] \quad \text{(definition of } g_j\text{)} \\
&= \frac{P\left[r_j \in M\right] P\left[\gamma \mid r_j \in M\right]}{P\left[r_j \in M\right] P\left[\gamma \mid r_j \in M\right] + P\left[r_j \in U\right] P\left[\gamma \mid r_j \in U\right]} \quad \text{(i)} \\
&= \frac{p \prod_{i=1}^{n} m_i^{\gamma_i^j} \left(1 - m_i\right)^{1-\gamma_i^j}}{p \prod_{i=1}^{n} m_i^{\gamma_i^j} \left(1 - m_i\right)^{1-\gamma_i^j} + \left(1-p\right) \prod_{i=1}^{n} u_i^{\gamma_i^j} \left(1 - u_i\right)^{1-\gamma_i^j}} \quad \text{(ii)}
\end{aligned}
$$

Here (i) follows from Bayes' theorem for a binary partition, and (ii) follows from the conditional independence assumption.

The M-step is to maximise the Q function over $<\boldsymbol{m}, \boldsymbol{u}, p>$ to obtain new estimates of $<\hat{\boldsymbol{m}}, \hat{\boldsymbol{u}}, \hat{p}>$. By equating partial derivatives of the Q function to zero, it can be shown that these estimates are

$$
\hat{m}_i = \frac{\sum_{j=1}^{N} \hat{g}_j \gamma_i^j}{\sum_{j=1}^{N} \hat{g}_j} = \frac{\sum_{j=1}^{2^n} \hat{g}_j \gamma_i^j f\left(\gamma^j\right)}{\sum_{j=1}^{2^n} \hat{g}_j f\left(\gamma^j\right)},
$$

$$
\hat{u}_i = \frac{\sum_{j=1}^{N} \left(1 - \hat{g}_j\right) \gamma_i^j}{\sum_{j=1}^{N} \left(1 - \hat{g}_j\right)} = \frac{\sum_{j=1}^{2^n} \left(1 - \hat{g}_j\right) \gamma_i^j f\left(\gamma^j\right)}{\sum_{j=1}^{2^n} \left(1 - \hat{g}_j\right) f\left(\gamma^j\right)},
$$

$$
\hat{p} = \frac{\sum_{j=1}^{N} \hat{g}_j}{N} = \frac{\sum_{j=1}^{2^n} \hat{g}_j f\left(\gamma^j\right)}{\sum_{j=1}^{2^n} f\left(\gamma^j\right)}.
$$

Here $f(\gamma^j)$ is the number of times the pattern γ^j occurs in the N record pairs. This simplification occurs because there are only 2^n distinct patterns of agreement that can be expressed by the vector γ^j (since it has length n and each component is binary). Since in practice we will have $2^N \ll N$, significant computational savings may be possible by using this expression.

In summary, the following four steps constitute the application of the EM algorithm to this problem.

- Pick starting values for $<\hat{\boldsymbol{m}}, \hat{\boldsymbol{u}}, \hat{p}>$.
- Using these, calculate $\hat{\boldsymbol{g}}$ (The E-step).
- Use this $\hat{\boldsymbol{g}}$ to calculate new values of $<\hat{\boldsymbol{m}}, \hat{\boldsymbol{u}}, \hat{p}>$ (The M-step).
- Iterate between steps 2 and 3 until the values of $<\hat{\boldsymbol{m}}, \hat{\boldsymbol{u}}, \hat{p}>$ converge.

Note that while we have not explicitly discussed blocking in the section, the same theory applies if the comparison space has been blocked. The only difference is now the estimated m and u probabilities will be blocking-dependent, rather than global probabilities.

4.2 Intuitive explanation

As mentioned previously, the basic idea of the EM algorithm is to solve an incomplete data problem by associating it with a complete data problem with a tractable solution. In this case, the complete data problem is where the class indicator, i.e. the match status of the record pairs is known. As we have said before, the motivation for data linking itself is the fact that the class indicator is unknown. Although we will never know the true value in practice, we *can* calculate the expected value, and then use this to estimate the maximum likelihood estimates.

Technical details aside, it is useful to have a more intuitive understanding of what the algorithm is doing. The following is an explanation of the steps of the algorithm in words.

1. Pick a starting value for the MLEs.
2. Assuming these are the 'true' MLEs, calculate the expected value of the class membership indicator.
3. Assuming this is the 'true' class membership, calculate new, improved MLEs.
4. Use these improved MLEs to re-calculate the expected value of the class membership indicator, and use this to recalculate the MLEs, and so on. Iterate until the values of the MLEs converge.

The use of the expected value of the class indicator is perhaps the aspect of the problem about which it is difficult to form an intuition. Each g_j is either 0 or 1, however the $\hat{g}_j$ take values in the interval [0,1]. Here, values of $\hat{g}_j$ closer to 1 mean it is more likely that record pair j belongs to the matched set, and values of $\hat{g}_j$ closer to 0 mean it is more likely that record pair j belongs to the non-matched set. It is useful to observe that if $\boldsymbol{g}$ is known, the MLEs as given in Section 4.1 simplify as you would expect. That is, $\hat{m}_i$ simplifies to the proportion of matches that agree on

field i, $\hat{u}_i$ simplifies to the proportion of non-matches that agree on field i, and $\hat{p}$ simplifies to the proportion of all record pairs which are matches. It is also useful to note the following properties of $\hat{g}_j$.

- If any $u_i = 0$ and the corresponding $\gamma_i^j = 1$ then $E(g \mid \gamma) = 1$. The interpretation is that if agreement on a field cannot occur by chance, any record pairs which agree on this field must be matches.
- For fixed values of u, higher values of p give values of $E(g \mid \gamma)$ closer to 1. For the trivial case $p = 1$, we have $E(g \mid \gamma) = 1$.
- For fixed values of p, lower values of $\prod_{i=1}^{n} u_i^{\gamma_i^j} \left(1 - u_i\right)^{1-\gamma_i^j}$, i.e. lower probability of observing an agreement pattern by chance, gives values of $E(g \mid \gamma)$ closer to 1. When p is small, $\prod_{i=1}^{n} u_i^{\gamma_i^j} \left(1 - u_i\right)^{1-\gamma_i^j}$ needs to be small also to achieve $E(g \mid \gamma)$ close to 1.

A useful intuitive way to think about this is that the EM algorithm implicitly tries to divide the records into the M and U sets. The key assumption is that the two latent classes which the EM algorithm identifies do actually correspond to the M and U sets. Pragmatically, it is important to realise that record pairs which agree on several fields end up implicitly assigned as matches, and record pairs which disagree on several fields end up implicitly assigned as non-matches. Beyond this, the EM algorithm does not have a 'magic' ability to divine match status, and will not surpass the accuracy of a well-trained clerical reviewer. The hope is that it will achieve comparable accuracy of parameter estimates, with substantially increased efficiency.

Note that up until this point we have only been concerned with obtaining estimates of m and u probabilities for input into our data linking software. However as we have discussed the EM algorithm also gives us an estimate $\hat{g}_i$ of the probability that record pair i is a match. This raises the possibility of using these $\hat{g}_i$ to directly drive the linking. In fact, Larsen and Rubin (2001) successfully implemented this method on datasets from the United States Bureau of the Census. Currently our research priority is to obtain better estimates of m and u probabilities for input into our current data linking software and practices, and so we do not pursue this direction further in this paper. It is however worth noting as a potential avenue for future research.

4.3 Treatment of missing data

The presentation of the material in Section 4.1 is consistent with the original presentation by Winkler (1988). Recall in Section 2.2 we discussed three options for extending this theory to account for missing data when assigning weights to record pairs. Option 1 is to treat missingness as disagreement, option 2 is to assign a zero weight when one or both fields are missing, and option 3 is to use the three comparison value approach. It is necessary that the definition of the parameters as estimated by the EM algorithm aligns with the definition of the parameters that is used to calculate weights. Option 1 works within Winkler's existing framework, but as we have discussed it is the least desirable option, and we have not used it previously in practice. Ideally we would like the EM algorithm to be used to estimate parameters for both options 2 and 3, in line with our current practices.

Yancey (2007) gives passing mention to the use of the EM algorithm with what we have named option 3. However we cannot find a reference to the technical details, and so we have derived the following independently, following on from the definitions we gave in Section 2.2.

Under the conditional independence assumption, the log-likelihood is now

$$l(\boldsymbol{m},\boldsymbol{u},p\,|\,\boldsymbol{g},\boldsymbol{\gamma}) = \sum_{j=1}^{N} g_j \log\left(p\cdot\prod_{i=1}^{n} m_{a,i}^{I\left[\gamma_i^j=1\right]} m_{d,i}^{I\left[\gamma_i^j=0\right]} m_{m,i}^{I\left[\gamma_i^j=-1\right]} \right) + \sum_{j=1}^{N} (1-g_j)\log\left((1-p)\prod_{i=1}^{n} u_{a,i}^{I\left[\gamma_i^j=1\right]} u_{d,i}^{I\left[\gamma_i^j=0\right]} u_{m,i}^{I\left[\gamma_i^j=-1\right]} \right),$$

and the conditional expectation of g_j is

$$E\left[g_j\,|\,\gamma\right] = \frac{p\prod_{i=1}^{n} m_{a,i}^{I\left[\gamma_i^j=1\right]} m_{d,i}^{I\left[\gamma_i^j=0\right]} m_{m,i}^{I\left[\gamma_i^j=-1\right]}}{p\prod_{i=1}^{n} m_{a,i}^{I\left[\gamma_i^j=1\right]} m_{d,i}^{I\left[\gamma_i^j=0\right]} m_{m,i}^{I\left[\gamma_i^j=-1\right]} + (1-p)\prod_{i=1}^{n} u_{a,i}^{I\left[\gamma_i^j=1\right]} u_{d,i}^{I\left[\gamma_i^j=0\right]} u_{m,i}^{I\left[\gamma_i^j=-1\right]}}.$$

Hence the Q function becomes

$$Q(\boldsymbol{m},\boldsymbol{u},p\,|\,\hat{\boldsymbol{m}},\hat{\boldsymbol{u}},\hat{p}) = E_{g|\gamma}\left[l\left(\boldsymbol{m},\boldsymbol{u},p\,|\,\boldsymbol{\gamma},\boldsymbol{g}\right)|\,\gamma;\hat{\boldsymbol{m}},\hat{\boldsymbol{u}},\hat{p}\right]$$
$$= \sum_{j=1}^{N}\hat{g}_j\left(\begin{array}{l}\log(p)+\sum_{i=1}^{n}\left[\begin{array}{l}I\left[\gamma_i^j=1\right]\log\left(m_{a,i}\right)+\\ I\left[\gamma_i^j=0\right]\log\left(m_{d,i}\right)+\\ I\left[\gamma_i^j=-1\right]\log\left(m_{m,i}\right)\end{array}\right]\end{array}\right)$$
$$+\sum_{j=1}^{N}\left(1-\hat{g}_j\right)\left(\begin{array}{l}\log(1-p)+\sum_{i=1}^{n}\left[\begin{array}{l}I\left[\gamma_i^j=1\right]\log\left(u_{a,i}\right)+\\ I\left[\gamma_i^j=0\right]\log\left(u_{d,i}\right)+\\ I\left[\gamma_i^j=-1\right]\log\left(u_{m,i}\right)\end{array}\right]\end{array}\right).$$

By using the method of Lagrange multipliers to maximise the Q function subject to the constraints $m_{a,i}+m_{d,i}+m_{m,i}=1$ and $u_{a,i}+u_{d,i}+u_{m,i}=1$ it is straightforward to show that the maximum likelihood estimates are

$$\hat{m}_{a,i}=\frac{\sum_{j=1}^{N}\hat{g}_j I\left[\gamma_i^j=1\right]}{\sum_{j=1}^{N}\hat{g}_j},\quad \hat{m}_{d,i}=\frac{\sum_{j=1}^{N}\hat{g}_j I\left[\gamma_i^j=0\right]}{\sum_{j=1}^{N}\hat{g}_j},\quad \hat{m}_{m,i}=\frac{\sum_{j=1}^{N}\hat{g}_j I\left[\gamma_i^j=-1\right]}{\sum_{j=1}^{N}\hat{g}_j},$$

$$\hat{u}_{a,i}=\frac{\sum_{j=1}^{N}\left(1-\hat{g}_j\right)I\left[\gamma_i^j=1\right]}{\sum_{j=1}^{N}\left(1-\hat{g}_j\right)},\quad \hat{u}_{d,i}=\frac{\sum_{j=1}^{N}\left(1-\hat{g}_j\right)I\left[\gamma_i^j=0\right]}{\sum_{j=1}^{N}\left(1-\hat{g}_j\right)},$$

$$\hat{u}_{m,i}=\frac{\sum_{j=1}^{N}\left(1-\hat{g}_j\right)I\left[\gamma_i^j=-1\right]}{\sum_{j=1}^{N}\left(1-\hat{g}_j\right)},\quad \hat{p}=\frac{\sum_{j=1}^{N}\hat{g}_j}{N}.$$

These results are all analogous to the binary agreement case, and in fact results for arbitrarily many agreement states continue to follow analogously. A drawback of introducing more agreement states is that it introduces more parameters to be estimated and more possible patterns of agreement $\boldsymbol{\gamma}^j$. For example with eight fields, there are $2^8=256$ patterns of agreement for a two-state comparison, but $3^8=6561$ patterns for a three- state comparison. Note it would also be possible to consider three agreement states for some fields and two for others.

Recall however that in Section 2.2, we mentioned that introducing the constraint $m_{m,i} = u_{m,i}$ results in zero weight contribution when field i is missing. Note that it would be possible to estimate $u_{m,i}$ from the data to be linked, as described in Section 3.1, and then set $m_{m,i} = u_{m,i}$. With this constraint imposed, we could implement a three-state comparison space without the extensions to the EM framework just discussed, since we would only need to estimate $m_{a,i}$ and $u_{a,i}$.

It is not however obvious to us how to modify the framework to accommodate the use of option 2. When estimating using training data, it is possible to first delete all record pairs that have field i missing on one or both records, and estimate the required probabilities from the remaining record pairs. However, an analogous deletion process is not possible when using the EM algorithm as it considers a joint distribution taking into account all linking fields simultaneously. One inexact approach would be to delete all record pairs that have missing data on *any* of the linking fields before using the standard binary outcome EM algorithm. We would expect that this would result in biased parameter estimates because the record pairs with lower data quality would not be used in the estimation.

> Question for the Committee: Can the Committee identify a way of modifying the algorithm to calculate m and u probabilities for use with 'method two' of handling missing data?

Winkler (personal communication, 24 May, 2005) suggested two further options for handling missing data.

1. Impute a fixed value for γ_i^j. Imputing $\gamma_i^j = 0$ corresponds to treating missingness as disagreement, which we have discussed previously. However Winkler also suggests imputing $\gamma_i^j = 0.5$.

2. Impute $E(\gamma_i^j)$ in the E-step.

The implication of imputing a value of γ_i^j other than 0 or 1, which could occur in both of these scenarios, is that both the agreement and disagreement terms contribute to the likelihood. For instance imputing $\gamma_i^j = 0.5$ introduces the terms $\sqrt{m_i(1-m_i)}$ and $\sqrt{u_i(1-u_i)}$, effectively hedging bets about the true agreement status.

Although these options allow missing data to be accounted for in the binary agreement state EM algorithm, it is not clear how he intended these ideas to actually fit in with weight calculation for missing data.

4.4 Other practical issues

A literature search has uncovered some other practical issues which arise when using the EM algorithm in this context. Two of these issues relate to the partitioning of the record pairs into the matched and non-matched sets. Yancey (2002) states that if the proportion of matches p is below 0.05 then the algorithm will likely fail to identify M and U accurately, and thus the resulting estimates $<\hat{\boldsymbol{m}},\hat{\boldsymbol{u}},\hat{p}>$ will be meaningless. Yancey (2002) also discussed an issue that arises when linking individual level data where there are multiple persons per household, and address fields are used for linking. In this situation, the EM algorithm may identify classes representing pairs at the same address, and pairs at different addresses. We investigate this phenomenon in Section 5.3. Yancey suggests using a three-class EM algorithm to overcome this problem. In this case the three classes correspond to matches, non-matches at the same household, and non-matches at different households. The set U is now the union of the latter two sets, so the results can still be used within the existing framework. Our code is not currently set up to implement this extension, but it is a possible topic for future investigation.

As we discussed in Section 2.1, the Fellegi–Sunter model is underpinned by the conditional independence assumption. It is worth noting that some research has gone into using EM type algorithms to fit extended models which can account for dependencies between different fields (Winkler, 1989a, 1992, 1993). Winkler explored the use of convex constraints, and fitting selected interaction terms, to extend the models. He showed these models lead to improved decision rules in some circumstances, but the choice of convex constraint or interaction terms was highly dependent on the data sets being linked. This work is not widely cited in subsequent data linking literature, and for this reason, as well as the complexity of the method, we consider this work out of scope for this research paper. That is, we only consider fitting the Fellegi–Sunter model. Further investigation into models which allow dependencies is a potential topic of future research.

4.5 Current usage

Although the purpose of this paper is to determine the suitability of the EM algorithm for data linking within the ABS, it is useful to discuss the use of this method in the wider data linking community. Winkler, who authored the original paper formulating the EM algorithm for data linking, and several subsequent related papers, is a researcher at the United States Bureau of the Census (USBC). It seems from their published papers that the EM algorithm is a key part of their data linkage processes. The USBC have in-house software for data linking, which they will give out on request, although they provide very limited support for external users. It is unknown to us if non USBC researchers have implemented the EM algorithm for data linking projects using the USBC software.

The Italian National Institute of Statistics (Istat) have also developed their own data linking software RELAIS (Record Linkage At Istat). This software is publicly available on their website, and it has an implementation of the EM algorithm as the standard way to estimate the m and u probabilities (Cibella *et al.*, 2007).

It is worth noting that not all statistical agencies which perform data linking use the EM algorithm. Statistics Canada previously used in-house software called GRLS (generalised record linkage system). GRLS required the user to provide m and u probabilities as inputs; it did not have functionality to estimate the probabilities using the EM algorithm or otherwise (Statistics Canada, 2001). Online references indicate this software has seen further development, and is now known as 'G-Link', but it seems the methodology implemented in the software has remained the same (Chevrette, 2011).

Statistics New Zealand do not have their own data linking software, but they estimate m probabilities using an iterative refinement method of the type we described in Section 3.3 (Statistics New Zealand, 2006).

There exists other software which implements the EM algorithm for data linking. We will not attempt to list all such software exhaustively; the point to note is that the use of the method is well established elsewhere.

5. EMPIRICAL INVESTIGATIONS

In this section we discuss some empirical investigations we have conducted to explore the practical feasibility of using the EM algorithm as part of our data linking processes. The work discussed in this section is limited to some basic investigations using synthetic data sets. In the first parts of this section we describe the properties of the synthetic data that was used for testing, and the code we have used to implement the EM algorithm. We then discuss the tests that were performed, and present the results.

5.1 Synthetic data sets

All the tests outlined in this section were done using synthetic data files that were developed for use in data linking research, specifically geared towards PES–Census linking. We used two files for testing, one of which is a subset of the other, so that all its records have matches on the other file. The smaller file has 3,000 records and the larger has 24,000.

These data files contain name and address fields, as well as date of birth, sex, country of birth and marital status. They are simulated to replicate the 'person within household' nature of the true PES and Census data files – that is every person within a household is included on the files. The files are simulated to have realistic demographic properties. In particular, they were created with the following properties.

- Realistic age-by-sex distributions for Australian-born and overseas-born subpopulations.
- Realistic proportions of overseas-born (by country of birth) and Indigenous subpopulations, by state.
- Realistic inclusion of overseas-born and Indigenous individuals within households.
- Realistic patterns of marital status within households.
- Realistic distributions of surnames within the population.
- Realistic patterns of surnames within households.
- Realistic distributions of male and female first and middle names.
- Realistic correlations between common and uncommon first names and surnames.

In order to give true m-probabilities less than 1, the files also have randomly generated errors. The nature of these errors is outlined in Appendix B. The rates of these errors are different for each field, and in some cases for different responses for a field. The error rates are also different for each of the two files. Table 5.1 lists the fields we are using for testing, and their error rate on both files.

5.1 Synthetic data fields and their error rates

Field	Abbreviation	Condition	File A error rate	File B error rate
First name	FNAME		0.0200	0.0500
Surname	SURNAME		0.0200	0.0500
Street name	STREET		0.0137	0.0122
First initial	IFN		0.0200	0.0500
Last initial	ISN		0.0200	0.0500
Day of birth	BDAY		0.0500	0.0800
Month of birth	BMONTH		0.0500	0.0800
Year of birth	BYEAR		0.0200	0.0400
Country of birth	COB	if Australia	0.0010	0.0010
		otherwise	0.0250	0.0400
Sex	SEX		0.0010	0.0010
Marital status	MST	If under 15	0	0
		otherwise	0.0150	0.0200

5.2 Computing environment

Our implementation of the EM algorithm is written in SAS. This code was originally written for the 2006 linking projects, but there were implementation problems and as mentioned previously this work was abandoned. When we revisited this work recently, another team member was able to make extremely significant efficiency improvements to the code.[1] Rather than taking hours to run, and crashing when used with realistically large comparison spaces, the improved code now runs in a matter of minutes. It is these improvements to the code which have made these investigations possible this time around.

5.3 Empirical tests on synthetic data

Since we know the true matches for the synthetic data files, we can calculate the true m and u probabilities exactly. This means that we can evaluate the accuracy of the estimates obtained using the EM algorithm against the true values. One of the issues identified in Section 4.4 was that if the true proportion of matches is low, the algorithm may not converge to a solution corresponding to the true M and U sets. By varying the blocking strategy from none to very tight, we can investigate the accuracy of the EM estimates as the true proportion of matches is varied.

The following tests were conducted on the two synthetic data sets using the EM SAS macro. The starting value $\hat{m}_i = 0.9$ was used for each field. Global u probabilities as calculated using the methods described in Section 3.1 were used as starting values for the (generally blocking-dependent) $\hat{u}_i$. A realistic starting value for $\hat{p}$ was calculated as the ratio of the number of records on the smaller file to the total number of record

[1] Damien Melksham was the team member responsible for these improvements.

pairs in the blocked comparison space. The absolute difference in the parameter estimates required for convergence was 0.0001.

Table 5.2 shows the results for the EM algorithm applied to the synthetic data sets with no blocking.

5.2 EM estimates without blocking

Fields	m	$\hat{m}$	u	$\hat{u}$
IFN	0.925	0.074	0.068	0.068
ISN	0.926	0.066	0.061	0.061
BDAY	0.871	0.038	0.032	0.032
BMONTH	0.877	0.089	0.083	0.083
BYEAR	0.940	0.999	0.012	0.006
COB	0.985	0.688	0.591	0.591
SEX	0.997	0.506	0.499	0.499
MST	0.973	0.999	0.304	0.298

Blocking:	None
p =	4.17 x 10^{-5}
$\hat{p}$ =	6.52 x 10^{-3}
Comparisons:	72,000,000
Iterations:	28

The first thing to note is that the algorithm has not converged to the correct solution. Given Winkler's comments that the algorithm can fail to identify the M and U sets correctly when the true proportion of matches is below 0.05, this result is not surprising, since in this case we have $p = 4.17 \times 10^{-5}$. Although the algorithm has not reached the correct solution, some interesting insights can still be gained from examining the results. The estimated m probabilities for year of birth and marital status are almost 1, which indicates that rather than identifying the match set, the algorithm has identified the set of pairs which agree on both year of birth and marital status. The remaining m probabilities are not much greater than the true u probabilities, indicating that there is little better than chance agreement on these fields. Note that the estimated u probabilities for all fields except year of birth and marital status are accurate, but this is because the proportion of true matches in the 'non-match' set is low, despite the algorithm having identified this set incorrectly, because there is no blocking. The estimated u probabilities for year of birth and marital status are too low however, because non-matches that randomly agree on both have been included in the 'matched' set.

Having observed these results, we now seek to determine what level of blocking is required before the algorithm converges correctly on this data. Blocking by sex gives $p = 8.33 \times 10^{-5}$, still very low, and the algorithm converges to essentially the same

solution as it did with no blocking. Blocking by marital status and sex gives $p = 2.66 \times 10^{-4}$, still two orders of magnitude below Winkler's suggested cut-off of 0.05. However, this time the algorithm does converge to the nearly correct solution. These results are given in table 5.3.

5.3 EM algorithm estimates blocking on sex and marital status

Fields	m	$\hat{m}$	u	$\hat{u}$
IFN	0.923	0.906	0.073	0.073
ISN	0.927	0.920	0.061	0.061
BDAY	0.869	0.852	0.032	0.032
BMONTH	0.876	0.862	0.083	0.083
BYEAR	0.941	0.943	0.027	0.027
COB	0.986	0.988	0.574	0.574
SEX	NA	NA	NA	NA
MST	NA	NA	NA	NA

Blocking:	SEX , MST
p =	2.66 x 10^{-4}
$\hat{p}$ =	2.79 x 10^{-4}
Comparisons:	10,944,654
Iterations:	19

Blocking by first initial gives $p = 6.15 \times 10^{-4}$, a larger p again, and yet this time the algorithm does not converge accurately. These results are given in table 5.4.

5.4 EM algorithm estimates blocking on first name initial

Fields	m	$\hat{m}$	u	$\hat{u}$
IFN	NA	NA	NA	NA
ISN	0.926	0.131	0.061	0.061
BDAY	0.870	0.099	0.032	0.032
BMONTH	0.876	0.151	0.083	0.083
BYEAR	0.940	0.991	0.013	0.006
COB	0.985	0.719	0.593	0.592
SEX	0.997	0.581	0.537	0.537
MST	0.973	0.999	0.304	0.299

Blocking:	IFN
p =	6.15 x 10^{-4}
$\hat{p}$ =	6.94 x 10^{-3}
Comparisons:	4,877,929
Iterations:	51

In fact here we see again the same behaviour as when there was no blocking. The algorithm has identified those record pairs which agree on year of birth and marital status as being the match set. Recall however that the previous scenario with a lower proportion of true matches, but blocked on marital status, converged accurately. An interesting insight can be gained here by considering the relationship between the fields 'year of birth' and 'marital status'. Marital status has four levels: never married, currently married, no longer married, and not applicable – 15 years or under. Because all children have the same marital status, the probability of records agreeing on year of birth given that they agree on 'not applicable' marital status is substantially different to the probability of records agreeing on year of birth given that they agree on another marital status. This is a violation of the conditional independence assumption. Recall this assumption is part of the underlying Fellegi–Sunter model, and so is applied to the likelihood when fitting the model using the EM algorithm. This result seems to illustrate that this violation is making it harder for the algorithm to converge to the correct M and U sets.

To investigate this further, table 5.5 gives the results from running the algorithm without any blocking, but this time removing marital status, and using only the seven remaining linking fields.

5.5 EM algorithm estimates without blocking, marital status removed

Fields	m	$\hat{m}$	u	$\hat{u}$
IFN	0.925	0.914	0.068	0.068
ISN	0.926	0.920	0.061	0.061
BDAY	0.871	0.853	0.032	0.032
BMONTH	0.877	0.864	0.083	0.083
BYEAR	0.940	0.924	0.012	0.012
COB	0.985	0.986	0.591	0.591
SEX	0.997	0.995	0.499	0.499
MST	—	—	—	—

Blocking:	None
p =	4.17×10^{-5}
$\hat{p}$ =	4.43×10^{-5}
Comparisons:	72,000,000
Iterations:	24

It is now the case, even with no blocking, that the algorithm has converged accurately. It is also the case that for various finer blocking strategies, the algorithm converges accurately when marital status is omitted, although we do not show the results here logical follow up investigation is to see if there is a level of blocking at which the algorithm does converge accurately when marital status is included.

Table 5.6 shows results obtained from blocking on first and second initial. Note here we have $p = 0.0087$.

5.6 EM algorithm estimates blocking on first and surname initials

Fields	m	$\hat{m}$	u	$\hat{u}$
IFN	NA	NA	NA	NA
ISN	NA	NA	NA	NA
BDAY	0.872	0.793	0.032	0.032
BMONTH	0.873	0.813	0.084	0.083
BYEAR	0.940	0.944	0.012	0.011
COB	0.986	0.981	0.594	0.593
SEX	0.997	0.986	0.537	0.536
MST	0.974	0.977	0.304	0.304

Blocking:	IFN , ISN
p =	8.71×10^{-3}
$\hat{p}$ =	9.68×10^{-3}
Comparisons:	297,633
Iterations:	15

We can see here that the estimates are reasonably accurate. While the estimated m probabilities for day of birth and month of birth are somewhat off, the overall results are not wildly wrong as they have been in previous examples. With tighter blocking, we see further improvement. Table 5.7 shows results obtained from blocking on collection district (CD). This gives $p = 0.0429$, which is close to Winkler's suggested cut-off of 0.05.

5.7 EM algorithm estimates blocking on collection district

Fields	m	$\hat{m}$	u	$\hat{u}$
IFN	0.925	0.907	0.069	0.068
ISN	0.926	0.924	0.133	0.132
BDAY	0.871	0.863	0.034	0.033
BMONTH	0.877	0.870	0.083	0.082
BYEAR	0.940	0.939	0.014	0.013
COB	0.985	0.985	0.620	0.620
SEX	0.997	0.983	0.488	0.488
MST	0.974	0.976	0.313	0.312

Blocking:	CD
p =	0.0429
$\hat{p}$ =	0.0424
Comparisons:	72,996
Iterations:	6

The accuracy of these results has improved relative to the previous example. The biggest discrepancy is for the first name initial m probability, which has a difference of 0.18 between the true and estimated values. It would seem then, that even with conditional independence violations, the algorithm can still converge correctly for these values of p.

In fact, it is interesting to note that despite Winkler's cautioning about problems when p is below 0.05, we have still been able to achieve correct convergence with much smaller values of p in some situations. It would be useful to perform further investigations into the behaviour of the algorithm for small values of p, using real-world data, rather than synthetic data.

> Question for the Committee: Can the Committee suggest other theoretical or practical issues which may explain Winkler's observation of incorrect convergence when $p < 0.05$.

Recall that in Section 4.4 we discussed that when address fields and surname are used for linking data representing multiple persons within households, the EM algorithm may converge to an incorrect solution. Specifically, it may identify the set of all pairs belonging to the same household, that is those agreeing on surname and address information, as being true matches. We call this the 'household problem'. Our tests so far have not used full surname or address fields, so we would not expect to have observed this. The following tests attempt to replicate this known problem. For the previous tests we have only been using initials as linking fields, and no address fields. For these tests we proceed to using first name, surname, and street name as linking fields. For these tests we block on CD throughout, thus holding p constant.

The results from the first such test are shown in table 5.8. These results show we have produced the convergence problems that we anticipated. The estimated proportion of true matches is twice as high as the true one, and while the estimated m probabilities for surname and street are close to correct (indeed both are slightly high), with exception of country of birth, the others sit between 0.3 and 0.4, indicating that while the match set does contain a significant proportion of true matches, many non-matches have also been identified as part of the match set. These results are consistent with the hypothesis that the algorithm has identified record pairs belonging to the same household as the match set.

5.8 EM algorithm estimates blocking on collection district, including household fields

Fields	m	$\hat{m}$	u	$\hat{u}$
FNAME	0.926	0.344	0.004	0.004
SURNAME	0.925	0.941	0.076	0.007
STREET	0.975	0.988	0.232	0.172
BDAY	0.871	0.356	0.034	0.032
BMONTH	0.878	0.395	0.083	0.081
BYEAR	0.941	0.371	0.014	0.012
COB	0.985	0.886	0.619	0.602

Blocking:	CD
$p =$	0.041
$\hat{p} =$	0.111
Comparisons:	71,404
Iterations:	15

The question now becomes what can we do to overcome this. Firstly, note here that sex and marital status were not used as linking fields. While not originally intentional, this serves to illustrate an interesting point. Table 5.9 shows the results for the same linking, but with sex, and then marital status added as linking fields.

5.9(a) EM algorithm estimates as for table 5.8 with sex included

Fields	m	$\hat{m}$	u	$\hat{u}$
FNAME	0.926	0.355	0.004	0.004
SURNAME	0.925	0.942	0.076	0.010
STREET	0.975	0.988	0.232	0.175
BDAY	0.871	0.366	0.034	0.032
BMONTH	0.878	0.405	0.083	0.080
BYEAR	0.941	0.382	0.014	0.012
COB	0.985	0.897	0.619	0.602
SEX	0.998	0.629	0.487	0.494
MST	0.974	—	0.313	—

Blocking:	CD
$p =$	0.041
$\hat{p} =$	0.108
Comparisons:	71,404
Iterations:	68

We can see from this table that with sex added as a linking field the algorithm still converges incorrectly, although the estimated proportion of true matches is slightly lower, and some m probabilities are slightly higher. However when marital status is also added, the algorithm now converges to the correct solution, although some of the estimated m probabilities are slightly low. These results would suggest that the household problem can be overcome by introducing more linking fields that are person-level rather than household-level.

5.9(b) EM algorithm estimates as for table 5.8 with marital status and sex included

Fields	m	$\hat{m}$	u	$\hat{u}$
FNAME	0.926	0.895	0.004	0.004
SURNAME	0.925	0.925	0.076	0.075
STREET	0.975	0.976	0.232	0.231
BDAY	0.871	0.862	0.034	0.033
BMONTH	0.878	0.871	0.083	0.082
BYEAR	0.941	0.935	0.014	0.013
COB	0.985	0.985	0.619	0.619
SEX	0.998	0.981	0.487	0.488
MST	0.974	0.974	0.313	0.312

Blocking:	CD
$p =$	0.041
$\hat{p} =$	0.042
Comparisons:	71,404
Iterations:	15

There is however another issue at play here. Recall that in all the tests so far, we have been using the global u probabilities, which can be calculated from the data itself, as the starting values for the blocking-dependent u probabilities in the algorithm. Up until now this approach has worked well for us, however in this case an added complication is introduced because we are blocking by CD. The global u probabilities for surname and street are 0.0045 and 0.0036 respectively. However the corresponding blocking-dependent u probabilities when blocking on CD are 0.076 and 0.232. In starting the algorithm with the much lower global probabilities, too much weight is given to agreement on these fields. Table 5.10 gives results for the same original test shown in table 5.8, but this time with starting values of 0.10 and 0.25 for the surname and street u probabilities respectively.

We can see now that this simple change of starting values has caused the EM algorithm to converge to the correct solution. While we have verified that the household problem does occur with our data, we have also identified two possible measures to fix the problem.

5.10 EM algorithm estimates as for table 5.8 with some starting values changed

Fields	m	$\hat{m}$	u	$\hat{u}$
FNAME	0.926	0.883	0.004	0.004
SURNAME	0.925	0.928	0.076	0.074
STREET	0.975	0.977	0.232	0.230
BDAY	0.871	0.856	0.034	0.032
BMONTH	0.878	0.868	0.083	0.082
BYEAR	0.941	0.927	0.014	0.013
COB	0.985	0.985	0.619	0.618

Blocking:	CD
$p =$	0.041
$\hat{p} =$	0.043
Comparisons:	71,404
Iterations:	7

6. CONCLUSIONS AND FUTURE DIRECTIONS

The results of the tests discussed in Section 5.3 indicate that the EM algorithm has good potential to be implemented as part of our data linking 'production environment'. We have demonstrated the algorithm produces accurate parameter estimates for our synthetic data sets. We observed that with a sufficiently high proportion of true matches the algorithm converged to the correct solution even in the presence of conditional independence violations. Furthermore, we observed that when the conditionally dependent field was removed from consideration, that the algorithm converged to the correct solution even with an extremely small proportion of true matches. It would seem that although Winkler has cautioned the algorithm may not converge correctly when p is much smaller than 0.05, in some cases it is possible to achieve convergence with a much smaller value of p.

Our results so far are promising, but there is more work which will need to be done before we can confidently adopt the EM algorithm as part of our production environment. Since our testing thus far has been restricted to synthetic data it will be important to perform further empirical investigations on the real Census and administrative data sets which will be used for the linking projects, when they become available to us. While the synthetic data was created as a realistic representation of Census data, we acknowledge the importance of repeating these investigations on real data to account for features not captured by the synthetic data sets. In particular we want to investigate the convergence of the algorithm on these data sets with different blocking strategies, thus varying the proportion of true matches. In doing this we will try and replicate the convergence for very small values of p that we were able to achieve with the synthetic data.

When testing on real data, the true values of the parameters will be unknown, and so the exact accuracy of the results will be indeterminable. However the results of Section 5.3 give us some informal diagnostic observations we can make to determine if the algorithm has converged to the correct solution. Observing estimated m probabilities close in magnitude to estimated u probabilities is a strong indicator that the algorithm has converged incorrectly. Observing estimated m probabilities very close to 1 for a subset of fields indicates the algorithm has identified the match set as containing all record pairs which agree on those few fields. Additionally, using prior knowledge and the results from previous linking projects, it will be possible to judge whether the estimates seem correct. Generally we would expect m probabilities to be greater than 0.9, with exceptions only occurring in the case of poor data quality or where field values can change over time. We would also be able to check the EM algorithm estimates against the m and u probabilities calculated from the linked files from 2006 linking projects. Furthermore, since u probabilities can be estimated accurately from the data to be linked, as discussed in Section 3.1, it is possible to

check the estimated u probabilities from the EM algorithm against the estimated u probabilities from this method.

Before we can put the EM algorithm into production we will need to make a decision on how to treat missing data. Recall we discussed three different options for dealing with missing data in Sections 2.2 and 4.3. Our SAS code currently implements 'option one', that is treating missingness as disagreement. The question of whether it is possible to modify the framework of the EM algorithm to work with 'option two' of conditioning on fields not being missing has been posed as a question to the committee. We are currently in the process of modifying our SAS code to calculate estimates for 'option three' – the three comparison value approach. It is likely these modifications will cause the code to run more slowly, though to what extent remains to be seen.

The most practical option would seem to be obtaining an estimate $\hat{u}_m$ from the data to be linked, setting $\hat{m}_m = \hat{u}_m$, and then using our existing code to obtain the estimates $\hat{m}_a$ and $\hat{u}_a$. The advantage of this approach is that it results in the assignment of a zero weight when one or both fields are missing, in line with current practice. Another advantage is it would require no changes to the existing code. We will need to test this approach once the real data is available. Note also that once we modify the code to calculate $\hat{m}_m$ and $\hat{u}_m$ explicitly, this will allow us to check the assumption that $\hat{m}_m = \hat{u}_m$, which would validate this approach.

It would also be useful for our future investigations to develop our SAS code to be able to implement the three class EM algorithm, which we discussed in Section 4.4. This would allow us to investigate whether the data we will be linking displays the 'household problem' identified by Yancey.

> Question for the Committee: What other investigations would the Committee recommend we undertake to confirm that the EM algorithm is suitable for use in our production environment?

Another potential modification to the code which we have not yet discussed, is the possibility of introducing constraints to the algorithm. For instance, there exists an upper bound on the proportion of true matches p, given when every record on the smaller file has a match on the larger file. An estimate $\hat{p}$ greater than this upper bound indicates the EM algorithm has not converged to the correct solution. In future we may investigate whether it is possible to improve the performance of the algorithm by incorporating such constraints.

Another high priority research topic for us is the use of frequency based weights within the data linking framework. The intuition behind frequency based weights is that agreement on a rare response for a field should be stronger evidence for a match than agreement on a common response. For example, agreement on Australia as country of birth is not very informative when linking Australian data sets, since about 70% of Australian residents were born here. However agreement on any other particular country of birth is much less likely to occur by chance, and therefore should be more informative. While the basic Fellegi–Sunter model does not account for this, Conn and Bishop (2006) discussed a possible extension. Further work on this is required, including the investigation of the use of the EM algorithm with frequency based weights.

Finally, note that while the focus of this paper is on methods to accurately estimate m and u probabilities, the extent to which the Fellegi–Sunter decision rule is affected by inaccurate parameter estimates remains an open question . So far we have not identified any papers which investigate this question in any depth, although Winkler and Thibaudeau (1993) suggest that linking outcomes can be materially affected by different estimation procedures for m and u probabilities. We expect to conduct a sensitivity analysis to determine the impact that mis-specified m and u probabilities have on the actual linking process.

ACKNOWLEDGEMENTS

I gratefully acknowledge several people who provided inputs for this paper. Damien Melksham made significant improvements to the SAS code used to implement the EM algorithm. Peter Rossiter created the synthetic data used for the empirical investigations and Section 5.1 draws from his documentation of this data. Noel Hansen helped devise and run the empirical investigations. Additionally, I received many helpful suggestions at various stages of drafting from Glenys Bishop, Phillip Gould, Gokay Saher, Paul Campbell, Sean Buttsworth, Peter Rossiter, and Guangyu Zhang.

REFERENCES

Australian Bureau of Statistics (2011) *Data Linking Manual*, Internal publication, Canberra.

Chevrette, A. (2011) *G-Link: A Probabilistic Record Linkage System*, System Engineering Division, Statistics Canada, Ottowa.

Cibella, N.; Fortini, M.; Scannapieco, M.; Tosco, L. and Tuoto, T. (2007) *RELAIS: Don't Get Lost in a Record Linkage Project*, Italian National Institute of Statistics, Rome.

Conn, L. and Bishop, G. (2006) "Exploring Methods for Creating a Longitudinal Census Dataset", *Methodology Advisory Committee Papers*, cat. no. 1352.0.55.076, Australian Bureau of Statistics, Canberra.

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

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

Hardy, M. (2008) *Calculation of m and u probabilities for Deaths–Census Linkage*, Internal report, Australian Bureau of Statistics, Canberra.

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

Liaw, C. (2007) "A Simulation-Based Method of Assessing the Quality of Linked Datasets", Unpublished *Methodology Advisory Committee Paper*, November 2007, Australian Bureau of Statistics, Canberra.

Solon, R. and Bishop, G. (2009) "A Linkage Method for the Formation of the Statistical Longitudinal Census Dataset", *Methodology Research Papers,* cat. no. 1351.0.55.025, Australian Bureau of Statistics, Canberra.

Statistics Canada (2001) *Generalised Record Linkage System Concepts*, System Development Division, Ottowa.

Statistics New Zealand (2006) *Data Integration Manual*, Statistical Methods, Wellington.

Winkler, W. (1988) "Using the EM Algorithm for Weight Computation in the Fellegi–Sunter Model of Record Linkage", *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, pp. 667–671.

Winkler, W. (1989a) "Near Automatic Weight Computation in the Fellegi–Sunter Model of Record Linkage", *Proceedings of the Fifth Census Bureau Annual Research Conference*, pp. 145–155.

Winkler, W. (1992) "Comparative Analysis of Record Linkage Decision Rules", *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, pp. 829–834.

Winkler, W. (1993) "Improved Decision Rules in the Fellegi–Sunter Model of Record Linkage", *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, pp. 274–279.

Winkler, W. and Thibaudeau, Y. (1991) *An Application of the Fellegi–Sunter Model of Record Linkage to the 1990 U.S. Decennial Census*, U.S. Bureau of the Census, Statistical Research Division Report.

Wright, J. (2010) "Linking Census Records to Death Registrations", *Methodology Research Papers*, cat. no. 1351.0.55.030, Australian Bureau of Statistics, Canberra.

Wright, J.; Bishop, G. and Ayre, T. (2009) "Assessing the Quality of Linked Migrant Settlement Records to Census Data", *Methodology Research Papers*, cat. no. 1351.0.55.027, Australian Bureau of Statistics, Canberra.

Yancey, W. (2002) "Improving EM Algorithm Estimates for Record Linkage Parameters", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 3835–3840.

Yancey, W. (2007) *Record Linkage: Theory and Practice*, United States Census Bureau.

APPENDIXES

A. CALCULATING WEIGHTS WITH MISSING DATA

The following toy example illustrates how the three different methods discussed in Section 2.2 for calculating weights with missing data work in practice.

These two files contain a unique identifier, and the common field 'sex'. Observe that of the 10 matches, 8 pairs agree on sex, 1 pair disagrees on sex, and 1 pair is missing sex on one file. The remaining 190 possible pairs are all non-matches, of those it can be seen that 73 pairs agree on sex, 80 pairs disagree on sex, and 37 pairs are missing sex on one or both files.

FILE A	
IDENT	SEX
P01	M
P02	F
P03	M
P04	F
P05	M
P06	F
P07	M
P08	F
P09	.
P10	F

FILE B	
IDENT	SEX
P01	M
P02	F
P03	M
P04	M
P05	M
P06	F
P07	M
P08	F
P09	M
P10	F
P11	M
P12	F
P13	.
P14	F
P15	M
P16	F
P17	M
P18	F
P19	.
P20	F

Using these figures we can calculate weights under the three different methods:

Method 1	
Probabilities	
m	0.800
$1-m$	0.200
u	0.384
$1-u$	0.616
Weights	
Agreement	1.058
Non-agreement	–1.622

Method 2	
Probabilities	
m	0.889
$1-m$	0.111
u	0.477
$1-u$	0.523
Weights	
Agreement	0.898
Disagreement	–2.234
Missing	0.000

Method 3	
Probabilities	
m_a	0.800
m_d	0.100
m_m	0.100
u_a	0.384
u_d	0.421
u_m	0.195
Weights	
Agreement	1.058
Disagreement	–2.074
Missing	–0.962

B. ERROR GENERATION ON SYNTHETIC DATA

The following table details the nature of the errors introduced into the synthetic data files.

Field	*Nature of error introduced*	
FNAME	As for IFN, and a transposition of two letters within the name	
SURNAME	As for ISN, and a transposition of two letters within the name	
STREET	Another letter appended to the name	
IFN	A letter randomly selected from the other 25 letters of the alphabet	
ISN	A letter randomly selected from the other 25 letters of the alphabet	
BDAY	A day randomly selected from the other 30 possible days	
BMONTH	A month randomly selected from the other 11 possible months	
BYEAR	Y0 = BYEAR + round(AGE x N(0,0.0225)) Y1 = Year formed by transposing third and fourth digits of BYEAR Choose randomly between Y0 and Y1; If new BYEAR < 1906 or new BYEAR > 2006 or new BYEAR = old BYEAR, then instead choose new BYEAR randomly from (BYEAR+1, BYEAR –1)	
COB2		
	11 (Australia):	An alternative COB selected with probability proportional to the observed frequency of occurrence in the complete data.
	Other COB:	50% probability of 11 (Australia); and 50% probability of an alternative COB selected with probability proportional to the observed frequency of occurrence in the complete data. If the assigned COB is the same as the original, then replace by 11 (Australia).
SEX	Opposite sex	
MST		
	0 (15 & under):	No error introduced
	1 (Never married):	Random choice between 2 & 3
	2 (No longer married):	Random choice between 1 & 3
	3 (currently in a registered marriage):	Random choice between 1 & 2

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.

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. 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