1351.0.55.018

Australian Bureau of Statistics

Research Paper

Methodology of Evaluating the Quality of Probabilistic Linking

www.abs.gov.au

New Issue

Research Paper

Methodology of Evaluating the Quality of Probabilistic Linking

Glenys Bishop and Jonathon Khoo

Analytical Services Branch

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) THURS 5 APR 2007

ABS Catalogue no. 1351.0.55.018
ISBN 978 0 64248 290 7

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 Glenys Bishop, Analytical Services Branch on Canberra (02) 6252 5140 or email <analytical.services@abs.gov.au>.

CONTENTS

METHODOLOGY OF EVALUATING THE QUALITY OF PROBABILISTIC LINKING

Glenys Bishop and Jonathon Khoo
Analytical Services Branch

ABSTRACT

The Australian Bureau of Statistics (ABS) will begin the formation of a Statistical Longitudinal Census Data Set (SLCD) by choosing a 5% sample of people from the 2006 population census to be linked probabilistically with subsequent censuses. A long-term aim is to use the power of the rich longitudinal demographic data provided by the SLCD to shed light on a variety of issues which cannot be addressed using cross-sectional data. The SLCD may be further enhanced by probabilistically linking it with births, deaths, immigration settlements or disease registers. This paper gives a brief description of recent developments in data linking at the ABS, outlines the data linking methodology and quality measures we have considered and summarises preliminary results using Census Dress Rehearsal data.

KEY WORDS: data linkage; record linkage; match rate; match accuracy; probabilistic linking.

1. INTRODUCTION

The Australian Bureau of Statistics is conducting a Census Data Enhancement project to add value to the data collected in the five-yearly Census of Population and Housing.

The central feature of the project is the creation of a Statistical Longitudinal Census Data set (SLCD). The SLCD will be based on a 5% sample of the population. The aim is to link records for this sample from each population census by statistical techniques which do not involve the use of name and address. It is intended that the sample will be augmented at each census with a 5% random sample of people who have been born or migrated to Australia since the preceding census.

In addition, the SLCD may be linked with specified administrative datasets for approved statistical projects. Those that have been specified are: birth and death register data, long-term immigration data, and national disease registers.

These decisions were reached after an extensive consultation process involving focus groups, a discussion paper that invited comments (ABS 2005a), the commissioning of an independent Privacy Impact Assessment and the receipt of many submissions from

the public. In August 2005, after consideration of all comments received, the Australian Statistician announced the future of the project, via the Census Data Enhancement – Statement of Intention (ABS 2005b). This work is consistent with the ABS desire to maximise the use of information, for statistical purposes, by finding further uses for its currently available data holdings.

The whole 2006 Census data set may be used for quality studies. During the period of Census processing, names and addresses as well as other variables will be used to link Census data and other selected data sets for these quality studies. This will allow important studies relating to the quality of ABS outputs to be undertaken. Once Census processing is completed, all names and addresses provided by Census respondents will be destroyed. Datasets will not leave the ABS, nor be accessible to anyone other than those ABS officers involved in the quality studies. These linked data sets will be destroyed after use.

2. QUALITY STUDIES

The quality studies that have been proposed for the 2006 census processing period are of two types. The first type is to assess the feasibility and quality of linking without name and address, while the second is to help improve ABS statistical outputs. The six planned studies are shown in table 2.1.

2.1 Planned quality studies

Study	*Data Sets*	*Aim*
Assess feasibility and linkage quality		
Simulated SLCD formation	• Census Dress Rehearsal 2005 • Census 2006	• assess the feasibility of forming the SLCD without names and addresses • make defensible statements about the quality of the linked data
Migrant settlements	• Migrant Settlements since 2000 • Census 2006	• assess the feasibility of a subsequent statistical study to investigate outcomes for immigrants admitted under different entry visas
Improve ABS statistical outputs		
Indigenous mortality	• Deaths since August 2006 • Census 2006	• estimate the under-coverage of reported Indigenous status on death certificates • investigate the use of correction factors for improving estimates of Indigenous mortality
Improve the 2011 Post Enumeration Survey	• Post Enumeration Survey 2006 • Census 2006	• assess the feasibility of replacing the current clerical matching with an automated procedure • widen the search area for people who give vague addresses
Births under-coverage	• Births registered between January 2003 and July 2005 • Census 2006	• quantify the under-coverage of registered births and the impact of delayed registration on population estimates • identify characteristics of parents of unregistered or late-registered children
Labour Force Survey under-coverage	• Labour Force Survey August 2006 • Census 2006	• Use clerical procedures to compare LFS households, etc.

Of prime importance is the quality study to assess the feasibility of forming the SLCD by linking a 5% sample of the 2006 Census with the 2011 Census without names and addresses. This study will use data from the Census Dress Rehearsal conducted one year before the 2006 Census and comprising approximately 80,000 persons. These will be linked to the 2006 Census. It is proposed to link the two data sets both with and without name and address. It is essential that we can make defensible statements about the quality of linkage for studies of this type.

The second type of study is proposed for linking the 2006 Census with registered deaths occurring between August 2006 and June 2007 using name, address and other variables. This forms part of a strategy for improving the coverage of reporting Indigenous deaths. The quality of Indigenous mortality data is recognised as having limitations. Current estimates indicate that while most deaths are registered, only 60% of Indigenous deaths are identified as being Indigenous. Indigenous deaths are critical for the estimation of Indigenous mortality rates and for use in the calculation of Indigenous population estimates. The latter are used to develop assumptions for compiling Indigenous population projections. With this level of under-coverage, gaining an accurate estimate of the population is more difficult.

3. BRIEF SUMMARY OF DATA LINKING METHODOLOGY

The classification of linking methods is tricky because there is no common terminology. This paper uses the term probabilistic linking where the aim is to link records from two different data sets that are believed to belong to the same unit or person. This type of linking is used when there is partial identifying information, but no unique, error free, identifying key.

The ABS uses the model of Fellegi and Sunter (1969) as the basis for its data linking. They proposed a probabilistic framework to link records together using several fields. A key feature of this framework is the ability to handle a variety of linking variables and record comparison methods and come up with a single numerical measure of how well two particular records link. This allows ranking of all possible links and optimal assignment of the link/non-link status.

3.1 Data linking process

The process of linking records from two files A and B has been implemented using a general framework. The steps are standardisation, blocking and searching, record pair comparisons and a decision model.

The contents of two data sets need to be standardised to allow comparison between the different data sources. Standardisation may involve removing inconsistencies and parsing text fields such as name and address.

Blocking reduces the number of comparisons needed by only comparing record pairs where links are more likely to be found. Both files are divided into the same blocks and records within a particular block on file A are compared only with records within the same block on file B. Thus the opportunity to link a pair of matched records will be missed if blocking variable values for an entity's records disagree between the two files. Blocking variables are most effective when they break up the population into small groups of similar size. For example, comparing 100 small groups with 5 people in each only requires 2,500 comparisons whereas 10 groups with 50 people in each group requires 25,000 comparisons.

During the comparison stage, each linkage field for a record pair is compared and the result is a binary code, γ=1 (agree), 0 (disagree) or missing. This code is converted to a field weight. Calculation of field weights depends on obtaining two probabilities for each linkage variable, the first being the probability that the variable values agree if two records belong to the same entity, while the second is the probability that the variable values agree if the two records belong to different entities. These are called m and u probabilities, respectively, i.e.

$$m = \Pr\{\gamma = 1 \mid \text{records belong to same entity}\} \quad (1)$$

and

$$u = \Pr\{\gamma = 1 \mid \text{records belong to different entities}\} \quad (2)$$

Field weights can be modified for a number of circumstances. The comparison options we are looking at include:

- Exact match (e.g. sex). The field either matches or it does not and no adjustment is made to the field weight.
- Exact match (e.g. country of birth) but the weight is modified so that rarer values are given higher weights than more common values when they agree.
- Approximate string comparisons (e.g. name). Allows the weight to depend on the number of characters that are different, allowing for misspellings, poor handwriting, etc.. Winkler (1990) describes this method.
- Numerical difference (e.g. date of birth). Allows the weight to depend on how far apart values for day, month and year are.
- Geographical difference (e.g. address or mesh block). Can use spatial information to calculate distance between fields.

In the decision model, the assumption of conditional independence enables us to calculate the final record-pair comparison weight for each record pair by summing the individual field weights. Finally, a decision rule based on cut-offs determines whether the record pair is linked, not linked or considered further as a possible link.

3.2 Data linking strategy

The strategy for linking consists of determining which variables to use for blocking and which to use for linking. Where possible the combination of blocking variables should have many values that are approximately uniformly distributed, be accurately reported and have few missing values. This is not always possible and so it is usual to have more than one pass using a different set of blocking variables on each pass. Setting up the blocking variables for each of several passes is termed the blocking strategy.

Linking variables may also change with each pass as they should not be too highly correlated with the blocking variables. For instance, if year of birth is used for blocking, then age will not be a useful linking variable.

4. METHODS FOR INVESTIGATING LINKAGE QUALITY

Some terminology needs to be introduced for this discussion. Define match status as the true status of a record pair. A match means that the records belong to the same entity (such as a person); a non-match means that the records belong to different entities. Define link status as the status assigned from a record linkage procedure: record pairs can be assigned as links or non-links.

Thus record pairs can be classified into one of four groups, as shown in table 4.2.

4.2 Classification of matches and links

		Match status (True)		
		Matches	Non-matches	
Link status (assigned by computer)	Links	True Links (matches that are linked)	False Links (non-matches that are linked)	Total Links
	Non-links	Missed Links (matches that are not linked)	True Non-links (non-matches that are not linked)	Total Non-links
		Total Matches	Total Non-matches	Total record pairs

Two statistics used in information retrieval are precision and recall. In the data linking context, these can be termed *match accuracy* and *match rate*, respectively and are defined as:

$$match\ accuracy = \frac{\#\ true\ links}{total\ links} \tag{3}$$

and

$$match\ rate = \frac{\#\ true\ links}{total\ matches} \tag{4}$$

Usually only total links and total non-links are known. The challenge of a quality measure is to determine the match status for each record pair so that the unknown cell totals can be estimated.

Several quality measures and methods for estimating them have been suggested. Fellegi and Sunter (1969) used two quality measures, μ and λ to determine the accuracy of linkage. The first, μ, is the probability that a record pair is incorrectly assigned to *link*. This would be estimated by the ratio of *false links* to *total non-matches* in table 4.2. The second, λ, is the probability that a record pair is incorrectly assigned to *non-link*. This would be estimated by the ratio of *missed links* to *total matches* in table 4.2. It is also equal to (1 – *match rate*). Calculation of these

quantities depends on obtaining the m and u probabilities for each linkage variable, as defined in section 3.1. Conditional independence of the linking variables is assumed in order to allow multiplication of these probabilities. This method does not allow for approximate string comparison.

Belin and Rubin (1995) observed that the above method tends to underestimate the true error rate because the conditional independence assumption does not usually hold. They suggest, as an alternative, calculating the "false match rate", denoted by ϕ and equal to (1 – *match accuracy*). The distribution of linkage weights is considered to be a mixture of two distributions, one for the matched pairs and one for the unmatched pairs. The two component distributions are estimated by fitting transformed normal curves to the record-pair weights. The transformation parameters are estimated using training data. While the method is robust to independence assumptions and asymptotic standard errors can be computed, clerical review is required to obtain parameter estimates and the distributional assumptions may not be valid. Winkler (2006) notes that this method only works well in situations where the curves can be well separated. This will generally not be the case when linking without names and addresses.

Winglee, Valliant and Scheuren (2005) suggest an approach, Simrate, for calculating μ and λ, defined above. They use a preliminary record linkage to obtain frequencies for matched pairs and a sample of unmatched pairs to obtain frequencies for non-matched pairs, thus estimating m and u probabilities. These two sets of the relative frequencies obtained are used to simulate record pairs corresponding to the match and non-match groups. This resulting simulated multinomial distribution is then used to set appropriate cut-offs so that the error rates of μ and λ can be calculated. The advantage of this method is that it can allow for dependencies between linkage variables and also approximate string comparisons.

Karmel (2004) and Blakely, Salmond and Woodward (1999) have taken a rather different approach by calculating the probability of a chance link for some given assumptions. Because probabilities can be calculated for different blocking and linking strategies, they can be used to determine an appropriate strategy. The method involves theoretical calculations and data are only used to estimate distributional parameters or verify assumptions. However, the calculations become very complicated as the number of linking variables is increased.

At the ABS, we have also considered three heuristic methods. These are estimating the expected number of matches, creating a benchmark file in a separate linking process, and counting the number of duplicate links. These methods are described briefly below.

In many data linking applications it is possible to estimate the expected number of matches. This expected number of matches can then be compared to the total

number of links obtained. This is a simple methodology and enables a quick assessment of linkage quality. However, it will depend on the assumptions used to estimate the expected number of matches. Nor is there any indication of the more complex quality measures discussed above such as match accuracy and false match rate.

If a benchmark file is created using a separate linking process, a subsequent linking output can be compared to the benchmark file to determine the quality of a linkage. The benchmark could be obtained in a number of different ways. In the ABS, we have experimented with a benchmark created by linkage using name and address for comparison with linkage strategies that do not use name and address. We have also obtained a benchmark file linked clerically to evaluate a file created by automated probabilistic linking. It may also be possible to use one probabilistic linking strategy to evaluate another.

Finally, we have investigated the quality of a linking run by counting the number of duplicate links. A duplicate link occurs when a record from one file, usually the smaller file, is linked to more than one record on the other file. In most of our linkage applications, it is theoretically impossible for this to happen and so linkage software includes an algorithm that forces 1:1 linkage when there are several likely candidate matches for a given record. By turning this facility off, we can allow duplicate links to occur, but the presence of a duplicate link indicates that there has been at least one incorrect link. The number of duplicates can be counted for different values of record comparison weight. This can be used to assist in the construction of appropriate cut-off weights.

5. APPLICATION OF SELECTED METHODS

The various methods described in the last section will be useful for different situations. In our early investigations, we have used data from the Census Dress Rehearsal (CDR) which was conducted one year before the Census proper and responses were received from approximately 81,000 people. This sample was not randomly selected and so many of the findings cannot be generalised to the Census. We have used two other data sets to link with CDR data. One consists of approximately 700,000 births registered between January 2003 and September 2005. The other is the dress rehearsal of the PES in which approximately 3300 observations were collected from 2100 dwellings. We restricted our attention to 2893 individual responses obtained from people who were usual residents of the dwelling surveyed and were at home on CDR night.

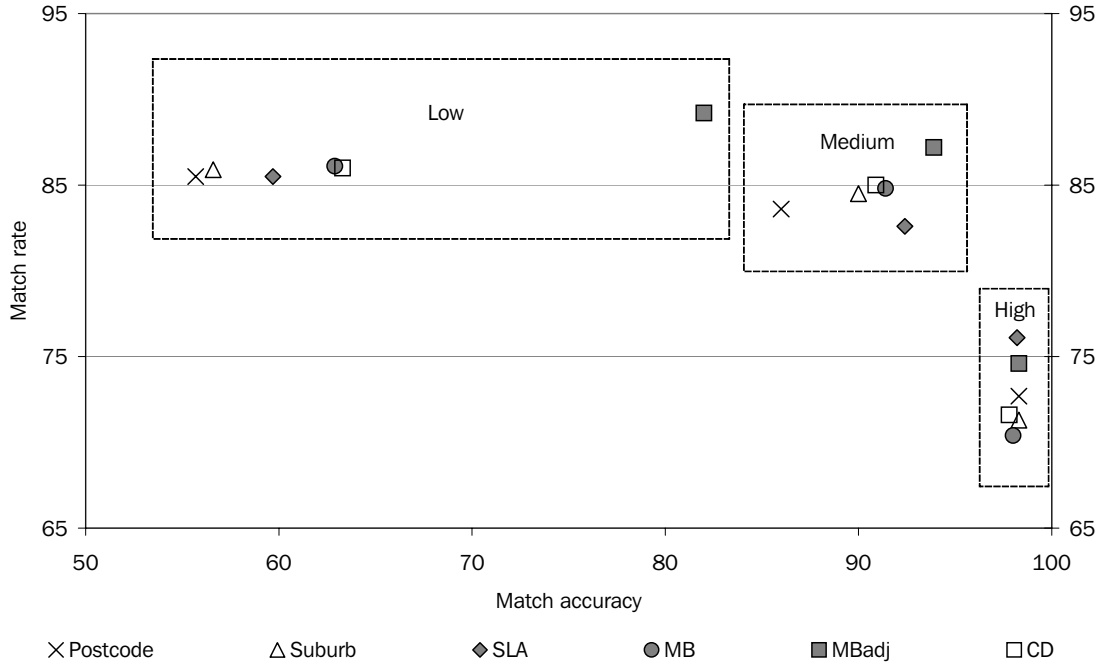
The method of estimating the expected number of matches has been used when linking birth registrations with CDR. The expected links consisted of 2747 Australian-born children aged 2 years and under on the night the CDR was conducted, who are present on the CDR data set. By including name and address among the linking variables, we were able to create a linked set of 2198 records. This constituted 80% of the expected number of links. The linkage percentage was the highest for those born in 2003 (85.2% of maximum expected links) and the lowest for those born in 2005 (74% of maximum expected links). These figures seem reasonable given that it is known that there is a delay in registering births.

When linking the CDR and the restricted PES DR data sets we had available to us a file that had been linked manually. This file was linked to the CDR file using names and addresses and probabilistic linking. The results were compared with the manually linked file. There was agreement on 87% of the linked pairs. On further examination, it was possible to resolve some of the differences and eventually obtain a file that could be regarded as a benchmark. Then linking runs were conducted without names and addresses and the results compared with this benchmark file to assess the match accuracy and match rate. In this way we have been able to compare the effectiveness of various geographic linking variables when address is not available.

We have also used this benchmark technique for assessing geographic linking variables with the data set obtained from linking births to CDR using names and addresses. First, the linkage was run using names and addresses to obtain a best estimate of matched pairs. Then a separate linkage was run for each geographic variable, but without names and addresses. Three cut-offs, low, medium, and high, were used so that all record pairs with weights above the selected cut-off were assigned as links. The linked pairs were compared with the matched pairs to obtain match rate and match accuracy for each cut-off and for each geographic variable.

The results are shown in figure 5.1. The general trend shown by each geographical variable is that as the match accuracy increases the match rate decreases. There is usually some point when a large decrease in match rate accompanies only a small increase in match accuracy. It is interesting to compare two forms of mesh block (MB and MBadj). A mesh block is a small geographic area representing about 50 dwellings and determined by man-made structures, such as roads, and natural features, such as rivers. About 11,000 mesh block codes were missing from the CDR file. The first linkage, indicated as MB, included these 11,000 units with missing mesh block codes whereas the linkage labelled MBadj excludes them. The use of mesh blocks gives a much higher match rate and match accuracy than other geographic variables, if mesh blocks are present.

5.1 Match rate vs match accuracy for Birth to Census Dress Rehearsal linkage without names and addresses for low, medium and high upper cut-offs



6. FUTURE PLANS

We have found some heuristic measures provide useful guidance in deciding whether our results look reasonable. However, in order to publish findings of quality studies we will need to make defensible statements about the likely linkage quality.

Generally the methods described above require some knowledge of the true status of match and non-match. The CDR and PES DR manual matching gives us a sound basis for obtaining frequencies of agreement for linkage variables among matched pairs. The frequencies obtained can then be fed into the Fellegi and Sunter method and also the Simrate approach. Thus our plan is to try both these approaches on the PES DR to CDR linking.

Because the linkage variables available for the PES are a useful subset of those available for CDR to Census linking, we can use the frequencies obtained in the former for the latter.

ACKNOWLEDGEMENTS

The authors wish to thank Dr Ewa Orzechowska-Fischer, Ms Charity Liaw, Mr Lewis Conn, Mr Jianke Li and Mr Reshen Soorinarain Dodhy for all their work in linking data and applying various quality measures.

REFERENCES

Australian Bureau of Statistics (2005a) *Enhancing the Population Census: Developing a Longitudinal View*, ABS cat. no. 2060.0, ABS, Canberra.

Australian Bureau of Statistics (2005b) *Census Data Enhancement – Statement of Intention*, available on the ABS web site, www.abs.gov.au.

Australian Bureau of Statistics (2006), Census Data Enhancement Project: an Update, ABS cat. no. 2062.0, ABS, Canberra.

Belin, T.B. and Rubin, D.B. (1995) "A Method for Calibrating False Match Rates in Record Linkage", *Journal of the American Statistical Association*, 90(430), pp. 694–707.

Blakely T., Salmond, C. and Woodward, A. (1999) "Anonymous record linkage of 1991 census records and 1991–94 mortality records: The New Zealand Census–Mortality Study (NZCMS Technical Report No. 1)", *Public Health Monograph Series*, No. 4, ISSN 1173–6844.

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

Karmel, R. (2004) *Linking Hospital morbidity and residential aged care data: examining matching due to chance*, AIHW Cat. No. AGE 40, AIHW, Canberra.

Winglee, M., Valliant, R. and Scheuren, F. (2005) "A Case Study in Record Linkage", *Survey Methodology*, 31(1), pp. 3–11.

Winkler, W.E. (1990) "String Comparator Metrics and Enhanced Decision Rules in the Fellegi–Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 354–359.

Winkler, W.E. (2006) "Overview of Record Linkage and Current Research Directions", *US Bureau of the Census Research Report*, No. 2006–02.

FOR MORE INFORMATION . . .

INTERNET	**www.abs.gov.au** the ABS web site 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 web site 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 web site, 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 web site can be downloaded free of charge.

WEB ADDRESS www.abs.gov.au

2000001558928
ISBN 9780642482907

RRP $11.00