Australian Bureau of Statistics

1351.0.55.025

Research Paper

A Linkage Method for the Formation of the Statistical Longitudinal Census Dataset

Research Paper

A Linkage Method for the Formation of the Statistical Longitudinal Census Dataset

Richard Solon and Glenys Bishop

Analytical Services Branch

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) WED 19 AUG 2009

ABS Catalogue no. 1351.0.55.025

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

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Mr Jonathon Khoo, Analytical Services Branch on Canberra (02) 6252 5506 or email <analytical.services@abs.gov.au>.

CONTENTS

A LINKAGE METHOD FOR THE FORMATION OF THE STATISTICAL LONGITUDINAL CENSUS DATASET

Richard Solon and Glenys Bishop
Analytical Services

ABSTRACT

As part of the Census Data Enhancement project, the Australian Bureau of Statistics has conducted a quality study that simulates the formation of the Statistical Longitudinal Census Dataset (SLCD). This simulation has been carried out by linking 2006 Census and 2006 Census Dress Rehearsal data. The linking was carried out both with and without the use of name and address, with the aim that the former would act as a benchmark for the latter when assessing quality. Linking without name and address is the method that will be used for the planned linking of the SLCD sample of the 2006 and 2011 Censuses, forming the first two waves of the SLCD. This paper describes the methods and processes used to simulate the formation of the SLCD.

1. INTRODUCTION

The Australian Bureau of Statistics (ABS) has embarked on the Census Data Enhancement project which aims to add value to data collected in the Census of Population of Housing, henceforth referred to as the Census. The centrepiece of the project is the Statistical Longitudinal Census Dataset (SLCD) which is based on a 5% sample of person records randomly selected from the 2006 Census. Those person records will be brought together with their records from the 2011 and subsequent Censuses. The SLCD will be augmented at each future Census with a 5% random sample of people who have been born or who have migrated to Australia since the preceding Census. As data from subsequent Censuses are added to the Statistical Longitudinal Census Dataset, its value as a resource for longitudinal population studies will increase. Further information about the Census Data Enhancement project is available in ABS publications (2005a, 2005b, 2006a).

These references also describe other aspects of the project, including *quality studies* which involve bringing the full Census dataset together with other specified datasets. If quality studies are performed during the Census processing period before names and addresses have been destroyed, then name and address can be used to assist in bringing records together. One such quality study undertaken using the 2006 Census dataset was the simulated formation of the SLCD. This paper describes the methods and processes used in that quality study.

2. THE SIMULATED FORMATION OF THE SLCD QUALITY STUDY

As a model for the formation of the Statistical Longitudinal Census Dataset, this quality study involved linking 2006 Census records with records from the Census Dress Rehearsal, conducted one year prior to the 2006 Census.

This quality study had two main aims. The first was to develop expertise in linking methods and processes for use when the SLCD is eventually linked. The second was to investigate the quality of the linked dataset and thus make statements about the likely quality of the Statistical Longitudinal Census Dataset itself.

Bishop (2009) reports on the investigation into the quality of the linked dataset. Conn and Bishop (2006) describe the basic linkage methodology that has been used in this quality study. This paper describes the implementation of that methodology, the additions and refinements, and some of the practical considerations that were necessary.

The 2006 Census processing period provided a window of opportunity to link the two datasets using name, address and other variables to create the best quality linked dataset possible. (As noted earlier, names and addresses were destroyed at the conclusion of the Census processing period.) This was termed the Gold Standard linked dataset. Two other standards of linked data, Bronze Standard and Silver Standard, were created as part of this quality study and they are described in detail in Section 4.3.

The Gold Standard, while not perfect, provided a benchmark for assessing the linkage quality of the other two Standards. In particular, the method used to create the Bronze Standard, without using name and address, is the method that is likely to be used to conduct the planned linking of the 2006 wave of the SLCD with the 2011 Census.

3. THE DATA

3.1 Census

The 2006 Census file used for this study consisted of 19,050,146 records excluding overseas visitors and imputed persons. The latter are people known to exist but for whom no Census form was returned and so a statistical method was used to impute their demographic information. See the ABS (2008b) publication.

The vast majority of the Census questionnaires were collected as self-reported hand-written forms, while some respondents chose to answer the questionnaire electronically. Many forms for remote Indigenous communities, and for the homeless population, were completed with the assistance of Census field interviewers. This was supplemented with administrative data from prisons and other institutions. It is important to note that special short forms, used for the homeless population, and administrative data sources supplied less information than that collected using the normal Census questionnaire. Further details are in the ABS (2006b) publication.

3.2 Census Dress Rehearsal

The Census Dress Rehearsal was conducted to test collection and processing procedures. It was held on Tuesday, 9 August 2005 in parts of Sydney, Wagga Wagga and Junee in New South Wales and in parts of Adelaide and Murray Bridge in South Australia. Data were also collected from three remote communities in Western Australia and the Northern Territory. No administrative records were included (ABS, 2006b).

The initial dataset consisted of 78,958 person records. However, some people might have died between the Census Dress Rehearsal and the Census. Therefore, a preliminary linking process using the Dress Rehearsal data and a dataset of Deaths which had occurred between 10 August 2005 and 7 August 2006 was conducted. This process identified 609 deaths among Census Dress Rehearsal respondents. It should be noted that deaths that had not yet been registered and processed could not be included in this exercise.

The Census Dress Rehearsal records, belonging to persons identified as dead, were removed from the Census Dress Rehearsal dataset since there would be no corresponding records expected in the 2006 Census dataset. The resulting Census Dress Rehearsal dataset comprised 78,349 records.

3.3 Data preparation

To make the Census and Census Dress Rehearsal datasets consistent, some variables had to be standardised and adjusted.

3.3.1 Address variables

On each of the Census and Census Dress Rehearsal forms, four addresses were requested (ABS, 2008b):

- the address of the dwelling where the person was on Census night, termed the dwelling address;
- the address where the person usually lives, termed the usual address, if that is different from the dwelling address;
- the address where the person usually lived one year ago, if that was different from the current usual address; and
- the address where the person usually lived five years ago, if that was different from the usual address one year ago.

The usual address one year ago, collected in the Census, and the current usual address, from Census Dress Rehearsal, were used for linking purposes to align them in time. The method of collection of addresses meant that there were often missing values and so the following procedures were adopted. For the Census, if the address one year ago was missing then usual address was used; if that was also missing, then the Census night dwelling address was used. For the Census Dress Rehearsal, if the usual address was missing then the Census Dress Rehearsal night dwelling address was used.

A new building block of statistical and administrative geography that was introduced with the 2006 Census was the mesh block. These are micro-level geographical units for statistics and altogether there are in excess of 300,000 mesh blocks covering the whole of Australia. A mesh block may contain a residential area, an administrative area such as Parliament House or a geographic feature such as a national park. Typically, a residential mesh block contains between 30 and 60 dwellings. A street address can be coded to the appropriate mesh block but since many such addresses will occur in any given mesh block it is not possible to convert a mesh block to an address. Mesh blocks serve as a very useful geographical indicator to be used in linking. Further information is available in ABS publications (2007, 2008a).

Each of the addresses to be used in linking, as outlined above, was coded to the appropriate mesh block.

3.3.2 Name variables

Since names are not retained by the ABS following the Census processing period, they are not normally processed. Special procedures had to be developed for parsing, repair and standardisation so that names could be used as linking variables in the Gold Standard linkage. Parsing was used to remove titles and suffixes from names. Names that are hand-written on forms and then read using optical character recognition often contain errors. Where possible, these can be corrected using an automated method but otherwise a manual method, involving inspection of the original form, is required.

Names from both the Census Dress Rehearsal and the Census were subjected to automatic repair using name repair software and two dictionaries. These dictionaries, corresponding to given names and family names respectively, were obtained from the Australian Electoral Commission. The electoral roll contains only the names of Australian citizens aged 18 years and over and so the given names of some younger people and some recent migrants might not have been included in the dictionary. Therefore, the dictionaries were augmented with additional given names and surnames from births registered between 2003 and 2006 and immigrants arriving in Australia since 2000. List of names of births and immigrants outside of these dates were not available at the time.

In the automated repair process, a name which was able to be matched to an entry in the dictionary was left unchanged. Where no matching entry was located, the closest dictionary entry was then compared using an approximate string method which would assign a score for similarity of the name to the dictionary entry. A score above some critical value resulted in the name being converted to the dictionary name. Where there were either no strong alternatives, or there were multiple competing alternatives, the original name remained unaltered.

About 80% of names from each of the Census and Census Dress Rehearsal passed through this automated repair with a satisfactory result. The remaining 20% of Census Dress Rehearsal names were repaired manually. For the Census, that left about four million names requiring manual repair. Since this required too many resources, several groups of interest were targeted and their names repaired manually. Targets included those persons whose address one year beforehand fell into one of the Collection Districts used in the Census Dress Rehearsal, immigrants who had arrived since 2000, children under four years old and Indigenous Australians. These groups were of interest in either this quality study or one of the others conducted using the 2006 Census. A list of quality studies is contained in an ABS publication (2006a).

While names were available during the Census processing period, there was an opportunity to explore the possibility of creating hash values by encoding each combination of first name and surname to a numeric value.

Since Census forms are often completed by one person on behalf of other household members, there may be small errors in spelling, or nicknames may be used, but not necessarily consistently between Census Dress Rehearsal and Census. Both of these problems were fixed by converting similar forms of a name and nicknames to the one name, with male and female names treated separately. Then the hash value was derived separately for males and females, from the first four letters of a standardised first name and the first five letters of the surname.

The hash coding algorithm used in this quality study is a strictly many-to-one encoding scheme with 12,005 hash values created. A particular standardised name and surname will always code to the same hash value so that it is a useful linking variable. These hash values were constructed in such a way that any given hash value had a minimum of 1,500 distinct name combinations. Since at least 1,500 distinct names, and often more, received the same hash value, the process was not reversible.

First names and surnames were also converted to a phonetic code using the double metaphone algorithm.

Hash values and phonetic codes were destroyed along with names and addresses at the end of the Census processing period.

3.3.3 Other variables

Many variables are coded as part of Census processing, often to very fine detail. Over the course of a year some respondents may alter descriptions slightly, resulting in different fine-level codes. Therefore the values of some variables were grouped to form broader categories. For instance, country of birth codes and also ancestry codes for various parts of the United Kingdom were combined, as were those of the former Yugoslavia. Religion codes were grouped from the four-digit codes to three digits, the Level of highest qualification was collapsed from a four-digit code to one-digit, and the Field of study of highest post-school qualification from a six-digit code to two-digits.

Another modification was to adjust the Census Dress Rehearsal age by adding one year to it, since Census Dress Rehearsal was held exactly one year before the Census.

3.4 Variables used for linking

The Census Dress Rehearsal and Census forms requested identical information that resulted in the same variables, except that this information was collected one year apart. Some of these variables, such as Date of birth and Country of birth, should have remained constant over time, while others, such as Marital status and Level of highest qualifications, could change legitimately within the space of a year.

While it is easy to count the number of missing or invalid values, it is not possible to determine how many of the valid answers were not actually correct. In practice, there was no certainty that for the same individual, even variables that should have remained constant, such as Date of birth and Country of birth, would have the same values in the two datasets. Such inconsistencies could occur when one household member completes the form on behalf of others. Scanning of handwritten forms also introduced some errors, as did the repair of responses. Further information about Census data quality issues are available in an ABS publication (2008b).

Because there was no unique identifier on both datasets, linking of records was performed using a probabilistic approach. See Conn and Bishop (2006) for more details.

Variables used in linking fell into the following logical groups:

1. Name information: First name, Surname, Hash value, Double metaphone of first name, Double metaphone of surname, Initials;
2. Address information: Street number, Street name, Suburb/locality, Postcode, Mesh block, Collection District, Statistical Local Area and State, using the appropriate address as described in Section 3.3;
3. Personal constant characteristics: Date of birth, Sex, Indigenous status, Country of birth, Mother's birthplace (Australia or overseas), Father's birthplace (Australia or overseas), Ancestry and Year of arrival in Australia; and
4. Personal changeable characteristics: Age, Marital status, Language spoken at home, Highest year of schooling, Level of highest qualification, Field of study of highest post-school qualification, and Religion.

4. THE LINKING PROCESS

4.1 Method

Linking of records between two files involves comparing variable, or field, values of each record from the first file with corresponding variable, or field, values of each record from the second file. The physical position that a variable occupies on an electronic record is called a field and so the term ‘field’ is sometimes used in place of ‘variable’.

Comparison of values of a particular variable from each record-pair produces a numeric field weight. There are various ways in which variable comparisons can be made apart from exact agreement. Examples are approximate string agreement, numerical agreement with tolerances, either absolute or relative. The weights may be altered too. For instance, more weight may be given to agreement on relatively rare values of a variable. A list of different methods of comparison, or comparators, is available in Christen and Churches (2005). Conn and Bishop (2006) also discuss these methods.

The weights for all variables compared are added together to form a record-pair comparison weight. This weight is an indicator of how similar the two records are, the higher the weight the greater the similarity. Since records can belong to more than one record-pair, an algorithm is applied to choose an optimal set of unique record-pairs (Christen and Churches, 2005). The resulting unique record-pairs with weights above a certain cut-off weight are then declared to be links.

4.1.1 Blocking

Comparison of every record from one file with all records from the second file requires a large number of comparisons. For example, comparing every one of the approximately 80,000 records from the Census Dress Rehearsal with all 19 million records in the Census dataset would require (19 million × 80,000 =) 1.52×10^{12} comparisons. Since this would not have been feasible with the hardware and software available, a blocking technique was required, whereby only subsets of records from each dataset would be compared.

Such comparisons are performed within a block determined by a set of blocking variables. For instance, if Sex is used as a blocking variable, then the females from one dataset are compared only with the females from the other dataset, and the males only with the males. Thus blocking on the Sex variable avoids comparing the females of one dataset with the males of the other. If the two datasets have equal numbers of males and females, then the number of comparisons will be halved. In practice, a much greater reduction can be achieved when appropriate blocking variables are chosen.

Although blocking can reduce the number of comparisons significantly, it prevents the comparison of records belonging to the same person if one of them happens to have an erroneous blocking variable value. If a record in the first file belonging to a female had an erroneous Sex value of male , then this record would be excluded from comparison with any of the female records in the second file.

To overcome this problem, multiple passes can be used, each with different blocking variables. The ideal blocking variables are ones that are reasonably accurately reported and processed, consistent in reporting between the datasets being linked and should divide the files into many roughly equal-sized blocks.

4.1.2 Input probabilities

Calculation of the field comparison weights requires two sets of probabilities that can take any value between 0 and 1:

- the probability that the variable values on the two records of a pair agree, if the two records belong to the same individual, i.e. are matched (*m*-probability); and
- the probability that the variable values on the two records of a pair agree, if the two records belong to different individuals, i.e. are unmatched (*u*-probability)

Note that each linking variable (or field) has its own *m*- and *u*-probabilities, and the variable weight given to agreement (or disagreement) depends on the linking variable. For example, most people report their gender consistently at different times and on different data sets; thus the variable, Sex, has an *m*-probability very close to one. In addition, the numbers of males and females are roughly equal and so the *u*-probability is approximately 0.5. On the other hand, consider Street name as a linking variable. An individual may change address between data collections and there are many thousands of possible street names. Therefore, the *m*-probability for the Street name variable is substantially less than one and the *u*-probability is close to zero, because chance agreement on such a large number of possible values is small.

A common method for calculating the *m*- and *u*-probabilities is to assume that their values are constant throughout the whole datasets. In this situation we have used the term *global probabilities*. Global probabilities are most suitable for linking variables that are independent of the blocking variables.

If an individual reports information consistently on one variable, then this increases the probability of consistent reporting on another variable. As mentioned in Section 4.1.1, we use blocking to reduce the number of record-pair comparisons. The linking task, then, is to distinguish true matches from non-matches among all record-pairs which are in the same block.

Given that record-pairs are only compared if they agree on the blocking variable values, a linking variable's *m*-probability should be conditional on agreement on the blocking variable values. This block-specific probability is generally higher than the corresponding global *m*-probability, since it is calculated over matched pairs with higher data quality, i.e. those which agree on the blocking variable values. During the linkage process, if an arbitrary record-pair disagrees on the linking variable, then a block-specific *m*-probability penalises this disagreement more heavily than a global *m*-probability would. Intuitively, the block-specific *m*-probability recognises that a matched pair's agreement on the blocking variables has decreased the probability of a reporting error in the linking variable.

A linking variable's *u*-probability refers to the probability of chance agreement on that variable for a record-pair belonging to two different individuals. As mentioned previously, the linkage process only considers record-pairs which agree on the values of the blocking variables. A chance agreement on the blocking variable can substantially increase the probability of agreement on a linking variable. Therefore, it may be appropriate to use a block-specific *u*-probability rather than a global *u*-probability.

4.2 Software and hardware

The data linking software chosen was Febrl (Christen and Churches, 2005). Febrl 0.3 was released under an open source licence and was modified significantly in the ABS. The main changes were in improving the speed of access to records and in adding provision for clerical review and acceptance sampling.

Other data linking software, namely BigMatch, provided to the ABS by the US Bureau of Census (Yancey, 2002), was used for one stage of the Gold Standard linking process. It reduced the larger of the two files, the Census file, by keeping only the records which were most likely to be matches for the remaining Census Dress Rehearsal records.

Hardware was designed to cater for the memory-intensive requirements of Febrl and slow processing. For the four quality studies conducted using Febrl and one that did not use Febrl, the hardware consisted of a server with four 2.8 GHz dual core AMD Opteron processors, 64 GB RAM and a 250 GB hard disk, running a 64-bit Windows 2003 Server operating system. Even in such computing environment, a typical Febrl run linking 80,000 records to 20 million records, was taking between 2 and 10 hours to complete and was using between 5 and 50 GB of memory, depending on the linking strategy applied.

4.3 Linkages

Three standards of linking were performed between the 2006 Census and the Census Dress Rehearsal:

- Gold Standard was produced using name, address, mesh block and other variables and was created to serve as a benchmark for comparison;
- Bronze Standard was produced using mesh block and other variables and was created because this method is the one likely to be used for the planned linking of the 2006 wave of the SLCD with the 2011 Census; and
- Silver Standard was produced using hash value, mesh block and other variables and was created for exploratory purposes.

For each of the Bronze and Silver Standards, four separate linked datasets were created by using different record-pair comparison weight cut-offs when determining which pairs to declare as links. These cut-offs are referred to as High, Medium, Low and Very Low.

5. IMPLEMENTATION IN THIS QUALITY STUDY

5.1 Blocking

Test runs showed that the processing time for a linking pass was affected not only by the number of record-pair comparisons but also by the number of linking variables and the field size of those variables. After some experimentation it was decided that, to achieve a run overnight (i.e. in 13 to 14 hours), the number of comparisons should be restricted to less than 100 million. Blocking variables were selected to comply with this limit.

We adopted the approach of using a fine level of blocking on the first pass to make as many high quality links as possible, followed by coarser blocking on subsequent passes. At each pass there are fewer unlinked records left for consideration and so it is feasible to have blocking variables that divide the remaining records into broader groups. Section 6 of this paper shows how the number of Census Dress Rehearsal records under consideration for linking decreased at each pass.

A program for analysing block sizes and numbers of record-pair comparisons with different combinations of blocking variables was developed and used to help determine the most suitable sets of blocking variables.

A major concern was the number of record-pair comparisons that would have to be performed in a pass. Table 5.1 shows details of the blocking variables that were eventually selected.

For each blocking variable or combination of blocking variables, table 5.1 shows there were a few more blocks created in the Census Dress Rehearsal than appear in the combined Census Dress Rehearsal and Census datasets. One reason for this is that some people from the Census Dress Rehearsal did not complete a Census form. Bishop (2009) discusses various reasons for this. If the value of their blocking variable is rare enough, nobody else might record this value on the Census. For Mesh block, it is more likely that insufficient address information was given so that a dump code was used and these could not be used for linking purposes. For Sex and Date of birth blocks, some people may have entered their date of birth on the Census Dress Rehearsal but only reported their age on the Census.

Mesh block was used for the first pass in each linking standard since it was available for use in all of them, was reasonably accurately reported, divided the files into fairly uniformly sized blocks and had among the least number of record-pair comparisons of all blocking variables considered.

5.1 Details of blocking variables used in the Gold, Silver and Bronze Standards

Blocking variables		*CDR*	*Census*	*Combined CDR & Census*
All Standards				
Mesh block	Number of blocks	2,914	231,844	2,898
	Mean number of records per block	25	76	3,021
	Median number of records per block	2	75	151
	SD of number of records per block	41	49	13,426
	Record-pair comparisons			8,649,095
Sex & Date of birth	Number of blocks	41,682	75,237	41,675
	Mean number of records per block	2	238	528
	Median number of records per block	1	281	373
	SD of number of records per block	1	129	372
	Record-pair comparisons			21,984,650
Gold Standard				
Fn1, DM_Sn & Sex	Number of blocks	39,734	520,679	39,486
	Mean number of records per block	2	35	1,575
	Median number of records per block	1	3	141
	SD of number of records per block	2	166	8,522
	Record-pair comparisons			62,200,514
Silver Standard				
Hash value & Sex	Number of blocks	22,786	24,014	22,786
	Mean number of records per block	3	770	2,773
	Median number of records per block	3	752	2,304
	SD of number of records per block	2	199	2,108
	Record-pair comparisons			63,177,371

Explanation of variables used in blocking strategies:

Fn1 = Initial of first name

Fn2 = First two letters of first name

DM_Fn = Double-metaphone of first name

Sn1 = Initial of surname

Sn2 = First two letters of surname

DM_Sn = Double-metaphone of surname

A second major concern was to minimise the number of records in the Census Dress Rehearsal dataset that would not be compared with any Census record at all. One important step in ensuring this is to minimise the overlap of combinations of blocking variables between passes. For instance if Mesh block were missing from a record because Street name was missing then a second pass, blocking by Street name, would not enable that record to be compared in the second pass either.

The combination of Sex and Date of birth was used as a second pass for all standards of linking because it was reasonably accurately reported, divided the files into fairly uniformly sized blocks, was not as fine as Mesh block and was also independent of it.

Table 5.2 shows the number of records on each dataset that were not compared with any record on the other dataset because of missing blocking variable values under the blocking strategies that were finally adopted. After the first pass using Mesh block as the blocking variable, 4,707 Census Dress Rehearsal records had not been compared with any Census record. If the Sex and Date of birth combination were used, 6,557 records would not have been compared. However, all but 842 of these would have been compared in the first pass. A third pass was used in each of the Gold and Silver Standards and by that stage only 239 Census Dress Rehearsal records had not been compared with any Census record.

Only two passes were used for the Bronze Standard whereas a third pass was used for each of the Gold and Silver Standards because of the extra name information available. To maximise the opportunity of comparing record-pairs, blocking using name information did not use any form of geography or age.

5.2 Number of records that do not fall into valid blocks in the Gold Standard

	CDR		*Census*	
Pass	*Unblocked records*	*Cumulative*	*Unblocked records*	*Cumulative*
All standards				
1: Mesh block	4,707	4,707	1,577,229	1,577,229
2: Sex & Date of birth	6,557	842	1,289,844	112,996
Gold standard				
3: Fn1, DM_Sn & Sex	1,134	239	744,246	16,602
Silver standard				
3: Hash value & Sex	1,129	239	743,875	16,584

Terms are explained at the bottom of table 5.1.

Several candidate strategies were considered and subjected to analysis. Table 5.3 summarises the results. Initial of first name, Double metaphone of surname and Sex (highlighted in the table) were chosen for blocking because the number of record-pair comparisons was large enough to provide useful differentiation, but not so large that the linking run would take an unacceptable time to complete. The Silver Standard third pass used the Hash value and Sex for blocking, as Hash value was the only extra variable available.

5.3 Possible blocking strategies considered for Pass 3 in the Gold Standard

Blocking strategy	*Maximum block size*	*Number of record-pair comparisons*
Fn2, Sn2 and Sex	1,911,300	149,505,494
Fn2, Sn1 and Sex	17,213,994	819,040,819
Fn1, Sn2 and Sex	4,921,459	495,898,568
Fn1, DM_Sn and Sex	**471,408**	**62,200,514**
DM_Fn, DM_Sn and Sex	38,976	5,893,808
DM_Sn and DM_Sn	97,846	8,407,549
Fn2, DM_Sn and Sex	121,286	18,801,646
Fn2 and DM_Sn	274,896	32,483,072

Terms are explained at the bottom of table 5.1

5.2 Input probabilities

As explained in Section 4.1, m- and u-probabilities are required as inputs for each linking variable. The methods used for calculating these quantities were different for the Gold Standard and the other two standards.

5.2.1 Calculation of input probabilities for Gold Standard

Following each Census, a Post Enumeration Survey (PES) is conducted for the purpose of estimating under-count and over-count in the Census. During this process for the 2006 Census, 79,824 PES records were each clerically matched to a Census record. This matched set provided a training set of matched pairs, from which block-specific m-probabilities were calculated for the Gold Standard linkage.

For a given blocking pass, the number of PES and Census matches which agreed on the values of the blocking variables were counted. For each linking variable within that pass, the number of matches which agreed on the values of both the blocking variables and the linking variable were also counted. The ratio of the latter number to the former gave an estimate of the block-specific m-probability for that linking variable.

The training matched data had some shortcomings for the purposes of this quality study:

- The PES and Census were conducted approximately one month apart whereas the Census Dress Rehearsal and Census were one year apart so that there are more likely to be changes between Census Dress Rehearsal and Census than between the Census and Post Enumeration Survey.

- The Post Enumeration Survey data were collected by interviewer and were therefore of higher quality than either the Census or Census Dress Rehearsal, which were mainly collected by hand-written self-completion forms. Notable exceptions to this are electronic forms and interviewer-collected Indigenous forms in remote areas.

Because of these shortcomings the m-probability estimates obtained from the training data were expected to be too high. On the basis of a simulation of pass 1 using an in-house implementation of Simrate (Winglee, Valliant and Scheuren, 2005), the m-probabilities calculated from the training set were down-weighted by a factor of 0.9 for address variables and 0.95 for non-address variables.

To estimate u-probabilities for a particular pass of the Gold Standard linkage, all record-pairs which agreed on blocking variable values for that pass were considered. For a given linking variable, the number of agreements on the linking variable were counted and then the m-probability was used to estimate the number of these agreements that came from matched pairs. The remaining number of agreements were then assumed to come from the unmatched pairs. In a similar way, the relevant m-probabilities were used to obtain an estimate of the number of unmatched pairs that agreed on the values of blocking variables. The ratio of the former number to the latter provided an estimate of the block-specific u-probability for that linking variable.

The Post Enumeration Survey collected a limited set of variables so that m-probabilities could not be calculated directly for all linking variables from the training data. For each of the extra linking variables, a proxy Post Enumeration Survey variable, that might have similar rates of agreement and disagreement, was nominated. Thus, the block-specific m-probability for the linking variable was estimated as the block-specific m-probability for the proxy variable. For example, Language spoken at home was not collected on the Post Enumeration Survey, but its m-probability was estimated via the proxy, Birthplace. The block-specific u-probability was estimated as described earlier.

Provided that the number of linking variable agreements was sufficiently larger than the expected number of matches, the above method for estimating u-probabilities was stable against errors in the estimates of the m-probabilities. The method directly took account of any peculiarities of the Census Dress Rehearsal and Census data. For example, pass 1 used Mesh block as a blocking variable and Street number as a linking variable. Among the unmatched pairs which agreed on Mesh block, the probability of further chance agreement on Street number depended on the Census Dress Rehearsal sample design. This was directly accounted for by the estimation method and, since there were so many chance agreements on Street number, the estimate of u-probability was robust against errors in the estimate of m-probability for Street number.

On the other hand, in pass 1 the majority of agreements on First name came from matches. Therefore, the estimate of *u*-probability for the First name variable was quite sensitive to errors in the estimates of the global *m*-probability for Mesh block, and the block-specific *m*-probability for First name.

5.2.2 Calculation of input probabilities for Silver and Bronze Standards

The Silver Standard linkage used no address information apart from mesh block. As described in Section 3.3, a hash value between 1 and 12,005 was created, but no other name information was used. The Bronze standard linkage used Mesh block but not Hash value.

To obtain *u*-probability estimates that were insensitive to errors in the *m*-probability estimates, and to complete the Silver and Bronze standard linkages within the limited time available, global *m*- and *u*-probabilities, which were easier to calculate than block-specific probabilities, were used.

The Gold Standard was linked from the same datasets that were to be used for Silver and Bronze Standards and was to serve as a benchmark for them. Therefore Gold Standard links were used to estimate *m*-probabilities directly for all linking variables. Each global *u*-probability was calculated as the probability that two different people randomly selected from the Census agreed on the values of that variable.

5.3 Linking runs

5.3.1 General

Although a linking run can involve just one linking pass, a run typically consisted of a number of passes to combat the problem of missed links caused by blocking.

Each linking pass consisted of the following steps:

- determining a set of blocking and linking variables;
- setting comparator types for the linking variables;
- calculating *m*- and *u*-probabilities for the linking variables;
- setting initial weight cut-offs;
- conducting the linking by running Febrl software;
- using an acceptance sampling method implemented in the ABS version of Febrl to determine final cut-offs;
- confirming final cut-offs using other programs that were developed for that purpose;
- re-running the linking in Febrl using the final cut-offs; and
- conducting Clerical Review.

These steps were repeated for each pass. Linking results from all passes were then combined to form the final links file which consisted of two variables, a person identifier from the Census Dress Rehearsal file and the corresponding person identifier from the Census file. This file of identifiers could then be used to construct a linked data set consisting of variables required for analyses from each of the respective files.

Altogether nine datasets were produced: one Gold Standard, four Silver Standard and four Bronze Standard.

Blocking strategies and the resulting passes were common to all three Gold, Silver and Bronze Standards as much as possible. These are shown in table 5.4. The first pass used geography, the second personal constant characteristics and the third pass used name information. In the Bronze Standards, there was no third pass as name information was not used.

5.4 Comparison of blocking strategies by Standard

Pass	*Gold Standard*	*Silver Standard*	*Bronze Standard*
1	Mesh block	Mesh block	Mesh block
2	Date of birth & Sex	Date of birth & Sex	Date of birth & Sex
3	First name initial, Double metaphone of surname & Sex	Hash value	NO THIRD PASS

5.3.2 Gold Standard passes

Table 5.5 shows the comparators that were used for the Gold Standard in each of the three passes described in table 5.4. Linking variable m- and u-probabilities were specific to the blocking strategy and were calculated before each pass and these are also shown in table 5.5.

The Gold Standard links were to be regarded as true matches, i.e. belonging to the same person, for the purpose of quality assessment. It was therefore important to conduct an exhaustive linking process, including diverse strategies and manual clerical reviews, to capture the maximum number of actual matches.

After three passes, about 13% of the Census Dress Rehearsal records were still not linked. Since the Gold Standard was to serve as a benchmark, it was important to link as many of the true matches as possible. Thus further linking strategies were applied.

5.5 Gold Standard blocking and linking variables, Passes 1–3

	Linking variables			
Blocking variables	*Variable*	*Comparator*	*m-probability*	*u-probability*
		PASS 1		
Mesh block	First name	Approximate string	0.6400	0.0019
	Surname	Approximate string	0.7471	0.0136
	Day of birth	Exact string	0.8339	0.0288
	Month of birth	Exact string	0.8508	0.0723
	Year of birth	Exact string	0.8483	0.0132
	Sex	Exact string	0.9443	0.5005
	Indigenous status	Exact string	0.9289	0.9253
	Country of birth	Exact string	0.9096	0.4358
	Street number	Exact string	0.8666	0.1062
	Street name	Approximate string	0.8974	0.4222
		PASS 2		
Date of birth & Sex	First name	Approximate string	0.6390	0.0046
	Surname	Approximate string	0.7532	0.0006
	Marital status	Exact string	0.7352	0.4439
	Indigenous status	Exact string	0.9296	0.9161
	Country of birth	Exact string	0.9125	0.5155
	Street number	Exact string	0.7600	0.0106
	Suburb	Approximate string	0.7793	0.0009
		PASS 3		
First name initial, Double metaphone of surname & Sex	First name	Approximate string	0.7108	0.0454
	Surname	Approximate string	0.8881	0.2843
	Day of birth	Exact string	0.8420	0.0283
	Month of birth	Exact string	0.8583	0.0723
	Year of birth	Exact string	0.8558	0.0119
	Marital status	Exact string	0.7260	0.2453
	Indigenous status	Exact string	0.9299	0.9117
	Country of birth	Exact string	0.9140	0.5530
	Street number	Exact string	0.7642	0.0109
	Street name	Approximate string	0.8070	0.0004

BigMatch (Yancey, 2002) was used on the remaining unlinked records. Unlike Febrl, BigMatch does not link records, but instead reduces the larger of the two files by keeping only probable matches from the larger file. It also allows for multiple blocking to be applied simultaneously in one run. Six blocking strategies were applied: Double metaphone of first name, Double metaphone of surname, Day of birth, Month of birth, Year of birth and Country of birth.

Applying BigMatch substantially reduced the size of the Census file and so no blocking was necessary in pass 4. A dummy variable was used as a blocking variable in pass 4 because Febrl requires that at least one blocking variable must be specified.

After pass 4, there was still a reasonably large number of unlinked CDR records, and so the last pass, pass 5, used an expanded set of blocking and linking variables, to capture additional links. Table 5.6 shows the blocking and linking variables used in passes 4 and 5, together with comparators and *m*- and *u*-probabilities used for each linking variable.

The linked record-pairs from all the passes were then combined into one Gold standard links file.

5.6 Gold Standard blocking and linking variables, Passes 4–5

	Linking variables			
Blocking variables	*Variable*	*Comparator*	*m-probability*	*u-probability*
		PASS 4		
No blocking variable	First name	Approximate string	0.6288	0.0013
	Surname	Approximate string	0.7402	0.0004
	Day of birth	Exact string	0.8311	0.0283
	Month of birth	Exact string	0.8508	0.0726
	Year of birth	Exact string	0.8480	0.0107
	Sex	Exact string	0.9444	0.5002
	Indigenous status	Exact string	0.9273	0.9209
	Country of birth	Exact string	0.9081	0.5390
	Street number	Exact string	0.7560	0.0109
	Suburb	Approximate string	0.7775	0.0005
		PASS 5		
Age & State	First name	Approximate string	0.6288	0.0013
	Surname	Approximate string	0.7402	0.0004
	Day of birth	Exact string	0.8311	0.0283
	Month of birth	Exact string	0.8508	0.0726
	Year of birth	Exact string	0.8480	0.0107
	Sex	Exact string	0.9444	0.5002
	Indigenous status	Exact string	0.9273	0.9209
	Country of birth	Exact string	0.9081	0.5390
	Language spoken	Approximate string	0.9113	0.6581
	Religion	Exact string	0.7056	0.1528
	Level of highest qualification	Exact string	0.2446	0.0326
	Field of study of highest qualification	Exact string	0.2569	0.0153
	Highest year of schooling	Approx. numeric (±1 year)	0.6290	0.2308
	Postcode	Approximate string	0.9992	0.8847

5.3.3 Gold Standard clerical review

For each pass of the Gold Standard linkage, two cut-offs were set conservatively. Record-pairs with weights above the upper cut-off were assigned as links, record-pairs with weights below the lower cut-off as non-links, and record-pairs with weights between the cut-offs as possible links. Possible links were then subjected to clerical review to be classified as links or non-links.

The clerical review module of Febrl, developed within the ABS, displays two records under consideration, side by side. A reviewer needs to look through the variable values of the records and make a decision as to whether the records belong to the same individual. The reviewer either accepts the record-pair as a link or rejects it and it becomes a non-link. Then the next record-pair is displayed.

Record-pairs that were assigned link status were removed from the datasets before the next pass. Therefore clerical review had to be completed after each pass before the next pass was undertaken.

Clerical review is a time and labour intensive stage of the data linking process and it can be subjective and prone to errors. It was therefore important to plan the approach and to ensure that resources were managed properly.

Almost 11,000 record-pairs from the first pass were clerically reviewed and these results were later audited. It was found that about 200 record-pairs were incorrectly rejected. These record-pairs were subsequently reassigned as links.

However, the important lesson was that methods were developed for ensuring consistency among the four clerical reviewers. A sample of clerical review pairs (over a wide range of record-pair comparison weights) were selected and the group of reviewers discussed whether each record-pair was likely to be a match or non-match. They considered whether legitimate change and feasible reporting or processing errors could have led to the fields disagreeing. User fatigue was minimised by having code references accessible without needing to use hands or even move the head. Common codes, such as those for United Kingdom, New Zealand, Italy and Vietnam birthplace were highlighted. These are two examples of an array of procedures adopted to ensure consistency.

The amount of clerical review was reduced by using an acceptance sampling scheme, such as is used in industrial and other applications. See, for example, Montgomery (2005) or Juran and Godfrey (1999). The record-pairs were ordered by record-pair comparison weight and divided into batches of equal weight ranges. A sample of record-pairs was selected at random without replacement from a batch in the clerical review range. The record-pairs in the sample were examined clerically and each was assigned a link or non-link status. The number of links in the sample was compared to the two threshold values, defined by the acceptance sampling scheme, and one of three possible actions was taken:

- If the number of links observed in the sample was less than the lower sample threshold all the record-pairs in the batch were assigned as confirmed non-links (except for any sampled record-pairs actually identified as links).
- If the number of links observed in the sample was greater than the upper sample threshold all the record-pairs in the batch were assigned as confirmed links (except for any sampled record-pairs actually identified as non-links).
- If the number of links observed in the sample was between or equal to the thresholds all the non-sampled record-pairs in the batch were sent to clerical review.

Altogether in the five passes, more than 23,000 record-pairs were clerically reviewed.

5.3.4 Bronze Standard passes

The Bronze Standard did not use name or address, and so First name, Surname, Street number, Street name and Suburb were not used in linking. However other geographic components, namely Mesh block, Postcode, Statistical Local Area, Collection District and State were used. The reduced number of variables available for linking resulted in only two passes.

Table 5.7 shows the blocking and linking variables used in the two passes for the Bronze Standard, together with the comparator and *m*- and *u*-probability for each linking variable.

The *u*-probability that was used for Indigenous status was subsequently found to be incorrect; it should have been 0.921 (rather than 0.0921). Furthermore, since Indigenous persons only constitute approximately 3% of the population, a frequency-based weight would have been more appropriate. A later run using both of these corrections and a Low cut-off was performed. Whereas for the original run the Indigenous status variable disagreement weight was –4.54, the corrected run had a disagreement weight of –1.02. Furthermore the original run had a weight of 3.38 for agreement on Indigenous status, regardless of the actual value; but the corrected run gave a weight of 0.04 to agreement on a value of non-Indigenous and 5.06 to agreement on a value of Indigenous. This resulted in some small improvements in linking Indigenous people from remote areas. Unfortunately, for operational reasons it was not possible to incorporate these new findings in the quality assessment of the Bronze Standard.

5.7 Bronze Standard blocking and linking variables

	Linking variable			
Blocking variables	*Variable*	*Comparator*	*m-probability*	*u-probability*
		PASS 1		
Mesh block	Day of birth	Approx. numeric (± 2 days)	0.8821	0.0283
	Month of birth	Approx. numeric (± 1 month)	0.8896	0.0726
	Year of birth	Approx. numeric (± 1 year)	0.8804	0.0107
	Sex	Exact string	0.9966	0.5002
	Marital status	Exact string	0.7544	0.2370
	Indigenous status	Exact string	0.9610	0.0921
	Country of birth	Exact string	0.9179	0.5399
	Year of arrival	Approx. numeric (± 2 years)	0.2152	0.0009
	Language spoken	Approximate string	0.9113	0.6581
	Religion	Exact string	0.7056	0.1528
	Level of highest qualification	Exact string	0.2446	0.0326
	Field of study of highest qualification	Exact string	0.2569	0.0153
	Highest level of schooling	Approx. numeric (± 1 year)	0.6290	0.2308
		PASS 2		
Date of birth & Sex	Marital status	Exact string	0.7544	0.2370
	Indigenous status	Exact string	0.9610	0.0921
	Country of birth	Exact string	0.9179	0.5399
	Year of arrival	Approx. numeric (± 2 years)	0.2152	0.0009
	Language spoken	Approximate string	0.9113	0.6581
	Religion	Exact string	0.7056	0.1528
	Level of highest qualification	Exact string	0.2446	0.0326
	Field of study of highest qualification	Exact string	0.2569	0.0153
	Highest level of schooling	Approx. numeric (± 1 year)	0.6290	0.2308
	Mesh block	Exact string	0.8166	0.0000

5.3.5 Setting cut-offs in the Bronze Standard

Without name and address, full clerical review has limited use in resolving links. However, the method of acceptance sampling discussed in 5.3.3, was adapted for use in setting a single cut-off.

The adapted method consisted of ordering the record-pairs according to comparison weight in descending order and dividing the pairs into batches with equal weight ranges. From each batch, a random sample of record-pairs was selected and clerically reviewed to estimate the proportions of matches and non-matches. Multiplying these proportions by the batch size gave estimates of the numbers of matches and non-matches in the whole batch. Starting at the maximum weight batch and working down, the estimated numbers of matches (or correct links) and non-matches (or incorrect links) were accumulated if each successive batch were assigned as links.

Inspection of the cumulative graphs made it possible to decide where a single cut-off should be positioned to trade off missing correct links against including incorrect links.

As was stated in Section 4.3, four separate Bronze Standard linked datasets were created using different cut-offs, High, Medium, Low and Very Low. After the first pass, each of these cut-offs was set using the adapted method described above and four separate datasets were created. The High cut-off set had the smallest number of links and the most unlinked records from each dataset to go through to the second pass. The Very Low cut-off set, on the other hand, had the most links and the fewest unlinked records to go through to the second pass. Medium and Low cut-offs were between these.

The weight cut-off levels were kept consistent for each of the four sets over their two passes. For example, if the cut-off was high in pass 1 for Bronze High Standard, it remained high in pass 2, if it was very low in pass 1 of Bronze Very Low Standard, it also remained very low in pass 2, and so on. Thus in total, only one first pass was conducted but four second passes, one for each cut-off level.

The linked record-pairs above the high cut-off on the first pass and the additional linked pairs from above the high cut-off second pass were combined to form the Bronze Standard High linked dataset. Similarly the other three Bronze Standard levels were formed.

5.3.6 Silver Standard passes

The first two passes of the Silver Standard used the same blocking variables as those in the Gold and Bronze Standards. The third pass was blocked by Hash value and Sex. This pass was designed to facilitate further investigations into the suitability of using Hash value as a blocking variable. The blocking and linking variables together with input probabilities and the comparators used are shown in table 5.8.

The incorrect u-probability for Indigenous status was also used in creating the Silver Standard. No steps were taken to repeat the linking with the correct value substituted, as was done for the Bronze Standard, as, by that stage, hash values had been deleted.

5.8 Silver Standard blocking and linking variables

	Linking variable			
Blocking variables	*Variable*	*Comparator*	*m-probability*	*u-probability*
		PASS 1		
Mesh block	Hash value	Exact string	0.7381	0.0001
	Day of birth	Approx. numeric (± 2 days)	0.8821	0.0283
	Month of birth	Approx. numeric (± 1 month)	0.8896	0.0726
	Year of birth	Approx. numeric (± 2 years)	0.8804	0.0107
	Sex	Exact string	0.9966	0.5002
	Marital status	Exact string	0.7544	0.2370
	Indigenous status	Exact string	0.9610	0.0921
	Country of birth	Exact string	0.9179	0.5399
	Year of arrival	Approx. numeric (± 2 years)	0.2152	0.0009
	Language spoken	Approximate string	0.9113	0.6591
	Religion	Exact string	0.7056	0.1528
	Level of highest qualification	Exact string	0.2446	0.0326
	Field of study of highest qualification	Exact string	0.2569	0.0153
	Highest level of schooling	Approx. numeric (± 1 year)	0.6290	0.2308
		PASS 2		
Date of birth & Sex	Hash value	Exact string	0.7381	0.0001
	Marital status	Exact string	0.7544	0.2370
	Indigenous status	Exact string	0.9610	0.0921
	Country of birth	Exact string	0.9179	0.5399
	Year of arrival	Approx. numeric (± 2 years)	0.2152	0.0009
	Language spoken	Approximate string	0.9113	0.6591
	Religion	Exact string	0.7056	0.1528
	Level of highest qualification	Exact string	0.2446	0.0326
	Field of study of highest qualification	Exact string	0.2569	0.0153
	Highest level of schooling	Approx. numeric (± 1 year)	0.6290	0.2308
	Mesh block	Exact string	0.8166	0.0000
		PASS 3		
Hash value & Sex	Day of birth	Approx. numeric (± 2 days)	0.8821	0.0283
	Month of birth	Approx. numeric (± 1 month)	0.8896	0.0726
	Year of birth	Approx. numeric (± 2 years)	0.8804	0.0107
	Marital status	Exact string	0.7544	0.2370
	Indigenous status	Exact string	0.9610	0.0921
	Country of birth	Exact string	0.9179	0.5399
	Year of arrival	Approx. numeric (± 2 years)	0.2152	0.0009
	Language spoken	Approximate string	0.9113	0.6591
	Religion	Exact string	0.7056	0.1528
	Level of highest qualification	Exact string	0.2446	0.0326
	Field of study of highest qualification	Exact string	0.2569	0.0153
	Highest level of schooling	Approx. numeric (± 1 year)	0.6290	0.2308
	Mesh block	Exact string	0.8166	0.0000

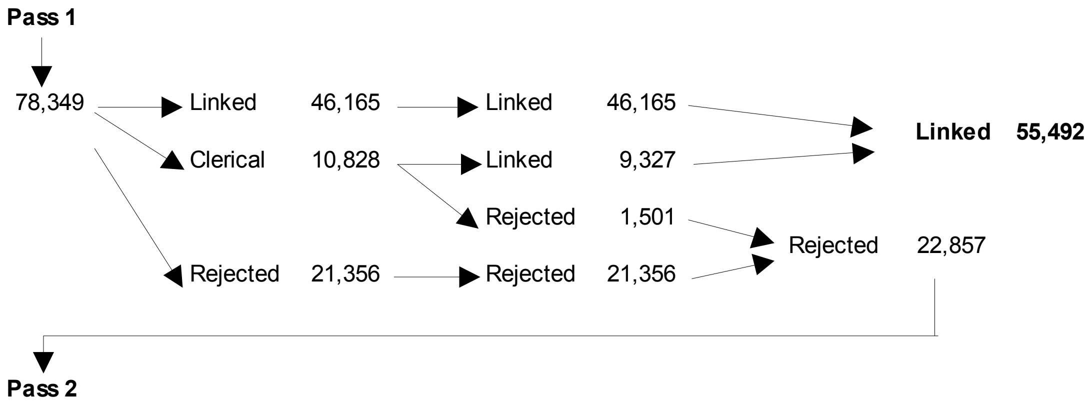
5.3.7 Setting cut-offs in the Silver Standard

The same method described in Section 5.3.5 for the Bronze Standard, was used to set cut-offs for the Silver Standard. In this case, however, there were three passes. Thus for the High level, record-pairs with weights above the high cut-off were linked and other records went through to the second pass where a high cut-off was set again and those not linked on this pass went through to the third pass to once again be subjected to a high cut-off. The linked record-pairs from the three passes were then combined to form the Silver Standard High level linked dataset. Similarly, each of the other three Silver Standards, namely Medium, Low and Very Low, were formed.

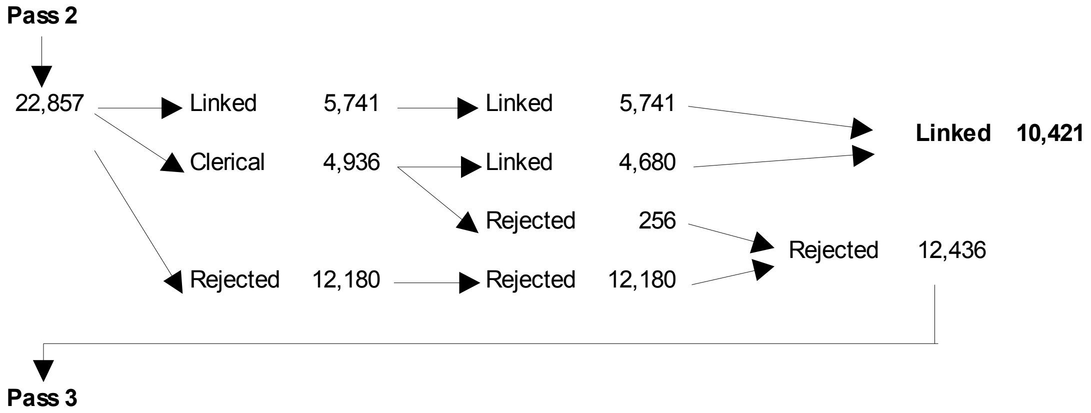
6. RESULTS

6.1 Gold Standard

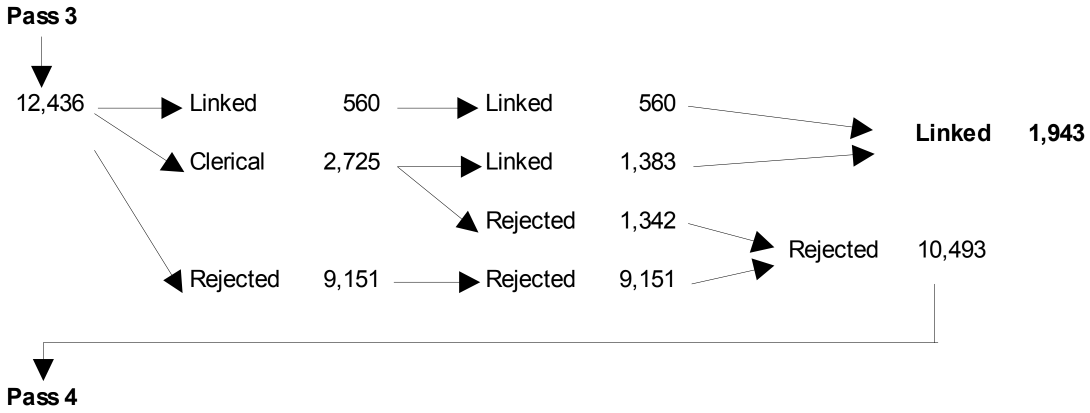
Pass 1 resulted in 55,492 links; 10,828 record-pairs were reviewed clerically, resulting in 9,327 confirmations and 1,501 rejections.



Pass 2 gave an additional 10,421 links, bringing the total number of links to 65,913. Clerical review of 4,936 record-pairs confirmed 4,680 pairs as links.

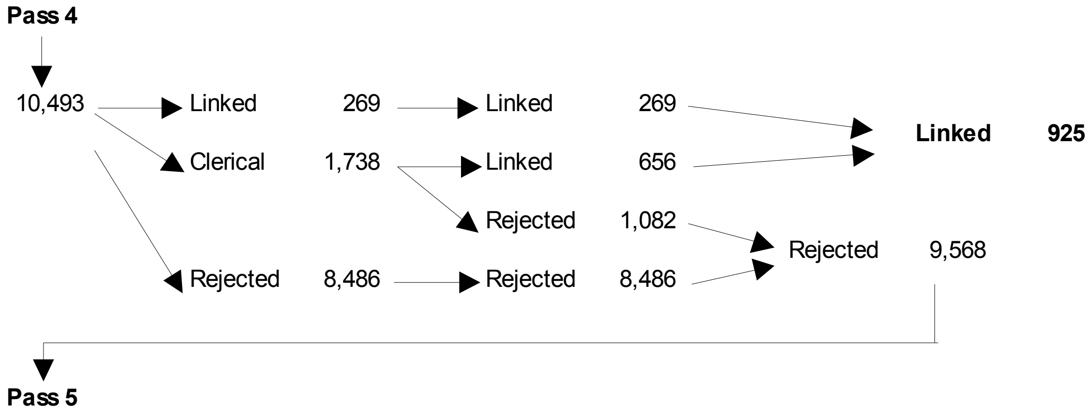


Pass 3 produced a further 1,943 links. Clerical review was conducted on 2,725 record-pairs, of which 1,383 were confirmed as links and 1,342 were rejected.

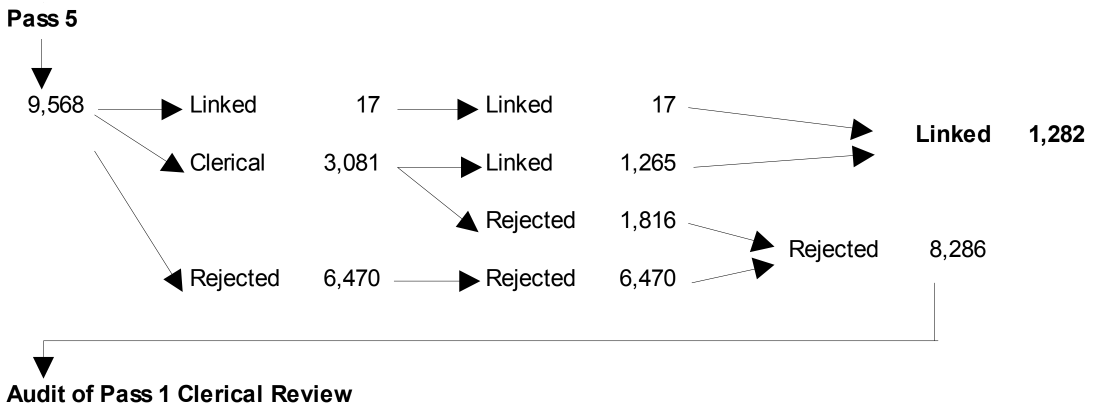


After three passes, 10,493 of the Census Dress Rehearsal records from the total 78,349, or 13%, were still not linked. Records that were already linked in the first three passes were removed from the Census Dress Rehearsal and from the Census files, leaving 10,493 in the Census Dress Rehearsal and 18,982,290 in Census. The Census file was reduced by BigMatch to 2,747 records.

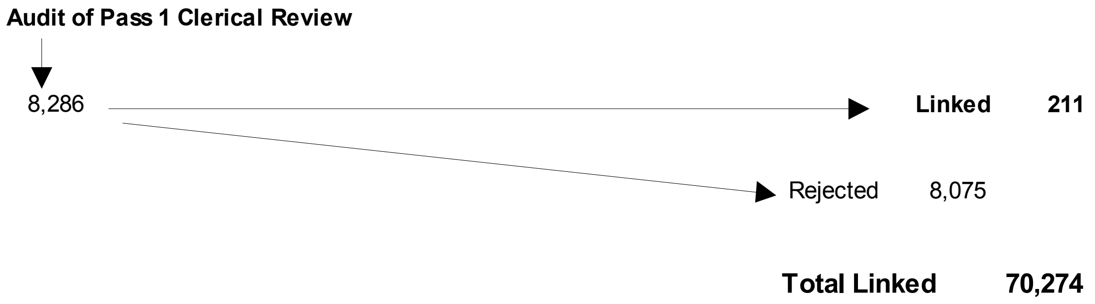
In pass 4, the remaining 10,493 Census Dress Rehearsal records were compared with these 2,747 Census records. This comparison produced another 269 links, and 1,738 record-pairs that were clerically reviewed, of which 656 were confirmed as links. In total, 925 additional links were found, bringing the total number of links to 68,781.



Pass 5 resulted in only 17 extra links and in 3,081 record-pairs that were reviewed clerically, 1,265 were confirmed as links. In total, 1,282 new links were found in pass 5, bringing the total number of links to 70,063.



In the five passes, 23,308 record-pairs were clerically reviewed altogether. The audit of the clerical review results from the first pass found 211 record-pairs that were incorrectly rejected; these were reassigned as links.



The final number of links was 70,274 of a possible 78,349. An analysis of 8,075 (78,349–70,274) records from the Census Dress Rehearsal, that were not linked, showed that variables that were critical to linking, names (first name and surname), date of birth or address (street number, street name, suburb and mesh block), were missing in a large number of these records:

- 488 (6%) records had either first name or surname missing;
- 1,211 (15%) records had mesh block missing;
- 1,934 (24%) had either all or a part of date of birth missing;
- street name or suburb was missing in 1,160 (14%) records;
- street number was missing in 1,423 (18%) records;
- about 38% or 3,082 records had some of these linking variables missing in various combinations; and
- the remaining 62% or 4,993 records had values present, but it is not known if the values were actually valid.

The numbers and percentages of unlinked Census records (19,050,146 – 70,274 = 18,979,872) with these linking variables missing were:

- 739,489 (4%) records had either first name or surname missing;
- 1,572,560 (8%) records had mesh block missing;
- 1,250,433 (7%) had either all or a part of date of birth missing;
- street name or suburb was missing in 778,935 (4%) records;
- street number was missing in 1,076,851 (6%) records;
- about 18% or 3,423,096 records had some of these linking variables missing in various combinations; and
- the remaining 82% or 15,731,006 records had values present in these variables, but it is not known if the values were actually valid.

6.2 Bronze Standard

Four Bronze Standard linked datasets were created: High, Medium, Low and Very Low. As described in Section 5.3.5, one cut-off was set for each level in pass 1 and then pass 2 was run separately for each level, with a different cut-off.

The two record-pair comparison weight cut-offs, one for each of the two passes for each standard were:

- 28 in pass 1 and 24.5 in pass 2, for High;
- 21 in pass 1 and 19.5 in pass 2, for Medium;
- 18 in pass 1 and 17.5 in pass 2, for Low; and
- 13 in pass 1 and 14 in pass 2, for Very Low.

Linking record-pairs with weights above the relevant cut-off in each pass and level produced the numbers of links shown in table 6.1 for each level of the Bronze Standard.

6.1 Number of links achieved in Bronze Standard

		Bronze Standard			
Pass	*Blocking variables*	*High*	*Medium*	*Low*	*Very Low*
1	Mesh block	19,827	44,715	49,889	53,851
2	Date of birth & Sex	14,773	5,170	2,087	3,939
Total		34,600	49,885	51,976	57,790

6.3 Silver Standard

As for the Bronze, four Silver Standard datasets were created: High, Medium, Low and Very Low. As described in Section 5.3.7, one cut-off was set for each level in pass 1 and then passes 2 were run separately for each level, with a cut-off specific to the level at each pass.

The record-pair comparison weight cut-offs, for each level were:

- 29, 23 and 28, for High level;
- 23, 18 and 25, for Medium level;
- 20, 15 and 22, for Low level; and
- 16, 11 and 15.86, for Very Low level (double-digit precision for the cut-off was used to include a group of record-pairs with weights of 15.86 but exclude the next group of record-pairs with weights of 14.87).

Linking record-pairs with weights above the relevant cut-off in each pass and level produced the numbers of links shown in table 6.2 for each level of the Bronze Standard.

6.2 Number of links achieved in Silver Standard

		Silver Standard			
Pass	*Blocking variables*	*High*	*Medium*	*Low*	*Very Low*
1	Mesh block	41,219	48,152	52,196	54,271
2	Date of birth & Sex	10,287	11,516	10,211	10,963
3	Hash code & Sex	1,725	699	637	1,380
Total		53,231	60,367	63,044	66,614

7. CONCLUSIONS

This paper has described the methods and processes used to simulate the formation of the Statistical Longitudinal Census Dataset. There were several issues for which solutions had to be found during this quality study.

One major concern was to obtain software that was well-documented, provided a variety of comparators, was transparent in its operation and allowed us to make changes to suit our purposes. Once we had found suitable linking software, the next concern was to access hardware on which the linking could be performed in a reasonable time. The solutions found were reported in Section 4.2.

Finding suitable methods for estimating *m*- and *u*-probabilities was another important task and different solutions were found for the Gold Standard and the other two standards, as indicated in Section 5.2.

While the actual linking was the main focus, data preparation – a vital step if the linking is to succeed – was much more time consuming. As discussed in Section 3.3, data preparation included name repair, both automated and manual, mesh block coding of addresses, both automated and manual, name standardisation, derivation of variables and broader recoding of finely coded variables. These steps required staff to develop an understanding of Census data and also of automated coders as well as becoming proficient in the use of editing software.

As part of the linking process, the clerical review step (described in Section 5) presented a number of problems. First, the software had to be modified to enable clerical review to be performed in batches. Second, a method using the statistical techniques of acceptance sampling was developed to reduce the clerical review workload to an acceptable level. Finally, staff conducting clerical review had to be trained in a protocol for determining whether a record-pair was a match, to ensure consistency among reviewers.

The clerical review process was adapted to help in the setting of cut-offs, also described in Section 5. Initially these were set quite high, and after some initial analyses it became clear that such high cut-offs were too stringent. To obtain a reasonable number of linked pairs in the Bronze and Silver Standards, it was necessary to lower the cut-offs. To help assess the optimal cut-off, several were used to create a four linked datasets for each of the Bronze and Silver Standards.

It is beyond the scope of this paper to report on quality issues arising from the different cut-offs and different linking standards. Bishop (2009) makes a full quality assessment of the linked data, including determining that linking the SLCD without name and address is feasible, although with caveats.

REFERENCES

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

—— (2005b) *Census Data Enhancement – Statement of Intention*, (last viewed on 5 August 2009) <http://www.abs.gov.au/websitedbs/D3110124.NSF/f5c7b8fb229cf017ca256973001fecec/5812a287d6a2e78fca2571ee001a7a49!OpenDocument>

—— (2006a) *Census Data Enhancement Project: An Update*, Information Paper, cat. no. 2062.0, ABS, Canberra.

—— (2006b) *How Australia Conducts a Census*, Information Paper, cat. no. 2903.0, ABS, Canberra.

—— (2007) *Review of the Australian Standard Geographical Classification, 2007*, Information Paper, cat. no. 1216.0.55.001, ABS, Canberra.

—— (2008a) *Mesh Blocks Digital Boundaries, Australia, 2006*, Information Paper, cat. no. 1209.0.55.002, ABS, Canberra.

—— (2008b) *Census Dictionary, 2006*, cat. no. 2901.0, ABS, Canberra.

Bishop, G. (2009) "Assessing the Likely Quality of the Statistical Longitudinal Census Dataset", *Methodology Research Papers*, cat. No. 1351.0.55.026, Australian Bureau of Statistics, Canberra.

Christen, P. and Churches, T. (2005) *Febrl 0.3 Documentation*, (last viewed on 5 August 2009) <http://cs.anu.edu.au/~Peter.Christen/Febrl/febrl-0.3/febrldoc-0.3/>

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

Juran, J.M. and Godfrey, A.B. (1999) *Quality Control Handbook*, 5th Edition, McGraw-Hill, New York.

Montgomery, D.C. (2005) *Introduction to Statistical Quality Control*, 5th Edition, John Wiley & Sons, Hoboken, New Jersey.

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

Yancey, W.E. (2002) *BigMatch: A Program for Extracting Probable Matches from a Large File for Record Linkage*, Research Report Series (Computing #2002-01), Statistical Research Division, U.S. Bureau of the Census, Washington D.C. 20233.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the linking and associated work performed by a team of methodologists, technologists and social statisticians and in particular Tenniel Guiver, Charity Liaw, Matthew Hardy and Zhuo Dai.

The authors would like to thank Tenniel Guiver, Peter Rossiter, Pramod Adikari and Alan Wong for valuable comments in the preparation of this paper.

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

Produced by the Australian Bureau of Statistics