1352.0.55.076

Australian Bureau of Statistics

Research Paper

Exploring Methods for Creating a Longitudinal Census Dataset

New Issue

Research Paper

Exploring Methods for Creating a Longitudinal Census Dataset

Lewis Conn and Glenys Bishop

Analytical Services Branch

Methodology Advisory Committee

18 November 2005, Canberra

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) THU 16 FEB 2006

ABS Catalogue no. 1352.0.55.076
ISBN 0 642 48180 6

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 Lewis Conn, Analytical Services Branch on Canberra (02) 6252 7743 or email <lewis.conn@abs.gov.au>.

CONTENTS

EXPLORING METHODS FOR CREATING A LONGITUDINAL CENSUS DATASET

Lewis Conn and Glenys Bishop

Analytical Services Branch

ABSTRACT

The Australian Bureau of Statistics (ABS) has embarked on a project to create a Statistical Longitudinal Census Dataset (SLCD) by linking records from the 2006 and subsequent Population Censuses. The SLCD will be based upon a 5% sample of the population. Since names and addresses will not be retained, probabilistic matching will be used to link records.

The ABS has investigated methodologies and applications for linking data sets over time. Probabilistic matching was used on ABS data in order to evaluate methodologies and identify particular issues with Australian census data. Results from this paper do not include Date of birth or Meshblock data as it was not collected in Census 2001. This information will improve the linkage provided it is reported well.

This paper outlines previous work in this area, results from similar projects, the methodology proposed for linkage and preliminary results achieved so far.

Key questions from this paper for MAC are:

1. What quality measures should we use to evaluate the best linkage strategy?
2. Is the assumption of conditional independence of linkage variables practical for the constant and changing variables selected?
3. Is the methodology outlined for matching censuses over time feasible to provide a linked dataset for useful analysis?
4. For records that are hard to identify uniquely, should statistical matching be considered to maximise the number of records in the linked dataset?
5. Is there any key data linking methodology/data set that ASB should look at that has not been discussed?

1. INTRODUCTION

1.1 Background

The Census is conducted every five years and attempts to collect information on every individual living in Australia on Census night. It is the largest single collection conducted by the Australian Bureau of Statistics (ABS) and obtains detailed information from all persons. The ABS has been investigating ways of using data linking techniques to create a longitudinal census data set and also the potential to bring other datasets, namely disease registers, births, deaths and immigration data together with the census.

After extensive community consultation the ABS intends to proceed with a Statistical Longitudinal Census Dataset (SLCD) based on a 5% sample of the population and without using Name and Address information. This allows many of the proposed uses of this dataset to be undertaken while maintaining ABS's commitment to privacy and confidentiality of everyone's details. Individual consideration will also be given to bring other data and the SLCD together for statistical purposes. For more details of what the ABS has proposed, readers are referred to the document *Census Data Enhancement Project – Statement of Intention* (ABS, 2005).

The methodology used is 'probabilistic exact matching' (For discussion of our use of the terms see Section 1.3). A key issue discussed in this paper is how to use the linkage framework without corroborating Name and Address information. This is especially relevant in linking records five years apart, as other variables may change during this time including Usual address, Marital status and Labour force status. Between 1996 and 2001, 6.8 million people (42.4%) aged five years and over changed address (see ABS, 2001).

1.2 Our tasks

The ABS has currently been investigating two key issues, namely:

- Develop a method for linking a 5% sample of the 2006 Census with 2011 and subsequent censuses.
- Investigate the feasibility linking administrative datasets to the Census for statistical purposes.

To understand the issues involved, develop the best linkage strategy and obtain estimates of the expected quality of linking we have decided to:

- Test various methods of linking the 2005 Census dress rehearsal with the 2006 Census full population.
- Use Name and Address for quality studies during census processing period.

This information will be used to create within the ABS a core set of knowledge of data linking techniques as well as assisting in deciding what computing and personnel resources are required to create and maintain a data linking unit. This project will also help determine whether the linkage proposal is feasible in its current form.

1.3 Terminology

The classification of 'matching' methods is tricky because there is no common terminology. This paper uses terminology believed to be the least confusing, but is not universally applied throughout the literature.

Whatever terminology is used, matching methods belong to two broad groups:

- *Exact matching* is generally used, in the technical sense, when linking records on two different data sets that are believed to belong to the same unit. The result is a dataset of linked units. Exact matching usually involves large data sets, such as administrative data sources or the Census.
- *Statistical matching* is generally used, in the technical sense, when it involves linking records that do not necessarily belong to the same individual. Rather, the aim is to create a file which accurately represents the population characteristics of both data sets. This is done by modelling the relationship between the variables, using techniques similar to imputation. Statistical matching is usually used when there is little or no overlap between the two datasets (such as linking together two household surveys).

Exact matching methods can be further classified as *deterministic* or *probabilistic*. *Deterministic matching* uses a unique identifier or a set of matching variables, commonly called a matching key. A match is declared when the keys are exactly the same on both data sets. This type of matching can only take place if the data sets have unique, error free identifiers.

In *probabilistic matching* all possible matches or links are evaluated and given a score based on the likelihood of a match. A match is then declared if the link score is higher than some predetermined cut-off. This type of matching would occur when there is partial identifying information, but no unique, error free, identifying key.

In this paper we are considering only *probabilistic matching*.

Note that probabilistic matching does not produce correct matches in all cases, and that it uses statistical techniques to achieve matches. Hence probabilistic matching may be termed statistical matching by some, but for the purposes of this paper we will use the above terminology.

2. PREVIOUS WORK

This section will outline a probabilistic matching framework (Section 2.1) and a summary of related linking work (Section 2.2).

2.1 Theoretical framework for data linkage

Fellegi and Sunter (1969) 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 matching variables and record comparison methods to come up with a single numerical measure of how well two particular records match. This allows ranking of all possible matches and optimal assignment of the match/non-match status.

2.1.1 The Fellegi–Sunter model of record linkage

The Fellegi–Sunter model assumes there are two files, denoted as A and B. Let a be the elements on A, and b be the elements on B. Then (a,b) are all possible pairs from A and B.

$$A \times B = \{(a,b); a \in A, b \in B\} \tag{2.1}$$

The record pairs are part of two distinct sets. The first is a matched set, denoted by M. This is where the record pair belongs to the same entity. The other group is unmatched, denoted by U, with record pairs which belong to different entities.

$$M = \{(a,b); a = b, a \in A, b \in B\}$$

$$U = \{(a,b); a \neq b, a \in A, b \in B\} \tag{2.2}$$

Each individual or unit has characteristics, some of which are common in both files. The common characteristics are used for record linkage. The characteristics on either file may have errors from reporting, coding or the optical character recognition process. A record pair (a,b) may actually be unmatched ($a \neq b$) but appear identical for the characteristics due to errors or too few common characteristics. Similarly a record pair may actually be matched but appear different due to errors and omissions.

To link the records of two files each possible record pair is compared. In the simple case, the result of a comparison is a set of k binary codes, stored in a comparison vector, γ. For instance, suppose we compare records using three characteristics: Date of birth, Meshblock and Country of birth. If the two records have the same values for Date of birth, then the first element of the comparison vector gets a 1, otherwise a 0. The length of the comparison vector is the number of common characteristics used

for comparison. In our example, there are $k=3$. Let Γ be the comparison space, consisting of all possible realisations of γ. If there are three fields to be linked then there would be 2^3 possible vectors in Γ.

Link status

The vector is a function of the two records being compared and so it can be written $\gamma(a,b)$. We observe $\gamma(a,b)$ for the record pair (a,b) and make a decision whether the record pair is a link, possible link or non-link:

2.1 Possible link status

Decision	*Description*	*Link Status*
A_1	$(a,b) \in M$	positive link
A_3	$(a,b) \in U$	positive non-link
A_2	(a,b) uncertain	possible link

A decision function has the form:

$$d(\gamma) = \left\{P\left(A_1|\gamma\right), P\left(A_2|\gamma\right), P\left(A_3|\gamma\right)\right\} \tag{2.3}$$

and

$$\sum_{i=1}^{3} P\left(A_i|\gamma\right) = 1 \tag{2.4}$$

Select a pair of records (a,b) randomly from $A \times B$ and obtain its comparison vector $\gamma(a,b)$. The conditional probability of γ given that (a,b) belongs to M is

$$m(\gamma) = P\left\{\gamma(a,b)\middle|(a,b) \in M\right\} \tag{2.5}$$

The conditional probability of γ given that (a,b) belongs to U is

$$u(\gamma) = P\left\{\gamma(a,b)\middle|(a,b) \in U\right\} \tag{2.6}$$

Linkage errors

Two types of errors can occur. The first is linking an unmatched comparison. This has probability:

$$P\left(A_1|U\right) = \sum_{\gamma \in \Gamma} u(\gamma).P\left(A_1|\gamma\right) \tag{2.7}$$

The second type of error occurs when a matched comparison is not linked. This has probability

$$P\left(A_3 \middle| M\right) = \sum_{\gamma \in \Gamma} m(\gamma).P\left(A_3 \middle| \gamma\right) \tag{2.8}$$

Linkage Rule

We can order the γ vectors for all record pairs that have been compared into a montonically decreasing sequence. This is done by first removing any γ for which $m(\gamma)=u(\gamma)=0$, then starting the sequence with γ for which $m(\gamma)>0$ and $u(\gamma)=0$. For the remainder calculate $m(\gamma)/u(\gamma)$ and arrange in decreasing order, using subscript i to indicate position in this ordered sequence.

Define a linkage rule $L(\mu,\lambda,\Gamma)$ on the space Γ at levels μ, λ such that $P(A_1|U)=\mu$ (incorrect match) and $P(A_3|M)=\lambda$ (missed match).

Admissible combinations of μ and λ are those for which the two criteria are met simultaneously with some set D of decision functions. The linkage rule, $L(\mu,\lambda,\Gamma)$, will be optimal if $P(A_2|L) \leq P(A_2|L')\ \forall\ L'$ – i.e. the rule minimises the probability of a possible link. An optimal rule is desirable because the possible links must be manually reviewed.

Using the ordered set of comparison vectors, let $u_i = u(\gamma_i)$ and $m_i = m(\gamma_i)$. Form the cumulative sums of u_i and find n such that:

$$\sum_{i=1}^{n-1} u_i < \mu < \sum_{i=1}^{n} u_i \tag{2.9}$$

Calculate the reverse cumulative sums of m_i and find n' such that:

$$\sum_{i=n'}^{N} m_i < \lambda < \sum_{i=n'+1}^{N} m_i \tag{2.10}$$

where N is the number of possible comparison vectors in Γ and $1<n<n'<N$.

For a record pair (a,b) observe the comparison vector γ which occupies position j in the ordered sequence, i.e. observe γ_j. Then take action as follows:

$$\begin{aligned} \text{True link, } A_1 &: \text{ if } j \leq n-1 \\ \text{Possible link, } A_2 &: \text{ if } n < j \leq n'-1 \\ \text{Non-link, } A_3 &: \text{ if } j \geq n'+1 \end{aligned} \tag{2.11}$$

When $j = n$ or n', a random decision is required.

Conditional independence

Fellugi and Sunter suggest the simplifying assumption of conditional independence. Suppose the components of γ can be ordered and grouped into $\gamma=(\gamma^1, \gamma^2, \ldots, \gamma^K)$. There is evidence that Birthplace of individual and Birthplace of both parents are not independent. Hence the Birthplace fields are grouped into γ^1. Similarly γ^2 may contain ancestry responses.

If two records are matched (i.e. belong to the same person) then a disagreement configuration could occur due to errors. The conditional independence assumption says that the chance of an error in the first group of Birthplace responses is independent of the chance of an error in the second group of Ancestry responses. Conversely, if two records belong to different people, then this assumption says the chance of agreement in the group of Birthplace responses is independent of the chance of agreement on Ancestry responses. Note that it is NOT assumed that responses within a group are independent, i.e. Birthplace of individual and parents.

The conditional independence assumption for m and u is written, in a simplified form as (2.12) and (2.13):

$$\begin{aligned} m(\gamma) &= m(\gamma^1)m(\gamma^2)\cdots m(\gamma^K) \\ u(\gamma) &= u(\gamma^1)u(\gamma^2)\cdots u(\gamma^K) \end{aligned} \qquad (2.12)$$

with

$$\begin{aligned} m(\gamma^K) &= P\left\{\gamma^K \middle| (a,b)\in M\right\} \\ u(\gamma^K) &= P\left\{\gamma^K \middle| (a,b)\in U\right\} \end{aligned} \qquad (2.13)$$

This simplifying assumption makes calculations much easier, but can lead to inaccurate results if there is dependence between groups of fields. In practice, the assumption of conditional independence is generally used.

Question: Is the simplifying assumption of conditional independence still practical with ABS census data as there are strong correlations between many data items?

Record pair weight

Any monotone increasing function of $m(\gamma)/u(\gamma)$ can be used as a record weight for the linkage rule. Fellegi and Sunter suggest using logarithms.

Define a *field weight*: $w^k(\gamma^k) = \ln\{(m(\gamma^k)\} - \ln\{ u(\gamma^k)\}$. The sum of the field weights is defined as the *record weight*: $w(\gamma) = w^1 + w^2 + \ldots + w^K$. The *record weight* $w(\gamma)$ is used to decide if a record pair is a link, non-link or possible link with the proviso that if $u(\gamma) = 0$, $w(\gamma) = \infty$ or if $m(\gamma) = 0$, $w(\gamma) = -\infty$.

The additivity of the field weights relies on the conditional independence of the components. The issue with Birthplace of individual and Birthplace of male and female parents agreeing even if the record was not a true link was identified from initial results. An overly high weight was assigned to each field and as a result many duplicate links were found. To address this issue the separate birthplace variables were grouped together into a single birthplace variable containing all information.

2.1.2 Blocking

Jaro (1995) discussed the issue of linkage of large public health data files using a commercial package AUTOMATCH. In the paper he discussed the importance of choosing appropriate blocking strategies to reduce the number of comparisons required. Linking two files with 1,000 records each requires 1,000,000 comparisons of possible pairs, of which at least 999,000 are unmatched pairs.

With the proposed Census data linkage, there will be approximately 1 million records in the 5% sample of Census 2006 to be linked to a larger subset of Census 2011. Implementing appropriate blocking strategies such as only comparing records with the same Collection District, adjusted Age and Sex, significantly reduces the number of comparisons required. However, a large risk with this is that the true matching pair may not be considered if one of the blocking variables differs between the two records. Hence careful choice of variables with high accuracy and low probability of changing over time is required for blocking.

2.1.3 Calculation of weights

Fellegi and Sunter (1969) propose two methods for calculating the weights. The first method assumes prior information is available about the distribution of the comparison characteristics in the *A* and *B* populations and also that the probabilities of different types of errors are known. We do have some of this information. For instance we have been able to determine the proportion of people who write their year last birthday instead of their actual Year of birth.

The second method relies heavily on the independence assumption to estimate the m and u probabilities directly from the input files. For this to work the input files must

be very large and the condition of entry into the file to be linked is uncorrelated to the distribution of comparison characteristics in the population.

Jaro (1989) uses blocking to reduce the number of record pair comparisons and for this reduced set calculates frequencies of the 2^n possible patterns of agreement and disagreement on n fields. An expectation maximisation (EM) algorithm is then used to estimate m and u probabilities.

Tepping (1968) proposes drawing a sample from all record pairs, identifying the matched and unmatched comparisons in this sample and estimate the m and u probabilities directly. It may be feasible for us to try this method during the census processing period when we can use Name, Address and Address-one-year-ago to determine matched and unmatched comparisons between Dress Rehearsal 2005 and Census 2006. Once these have been established we can calculate m and u probabilities using other matching variables.

2.1.4 One-to-one assignment

The theory outlined in Section 2.1.1 assigns record pairs with values greater than a critical value to be a match. However, this can cause problems when a record in the first dataset links to multiple records in the second dataset above this critical value.

Further work has been done by various people including Winkler (1989) and Christen (2004) into this issue. Various algorithms have been developed to account for this by maximising the sum of all the agreement weights through alternative assignment choices. This is called one-to-one assignment.

This methodology appears to give improved results without the need for intensive manual review of records. However, it may lead to incorrect assignment of records and further identifying information is required to verify whether the assigned link is the true match.

2.2 Previous studies with and without Name and Address

This section briefly describes three data linking projects from other agencies. The first study was conducted with Name and Address, while the other two were conducted without Name and Address. A summary of the aims of each study, linkage variables used and their achieved matching quality is included.

2.2.1 UK Longitudinal Study 1971–1991

Linkage Variables: Name, Address, Postcode, Postcode one year ago, Date of birth, Gender and Person type. (NB: UK postcodes are as good as or possibly better than Meshblock.)

The Office for National Statistics (UK) has followed a 1% sample of Census since 1971 (500,000 people) using administrative health records for births, deaths and address changes. With Name and Address, they have higher matching rates than the other two projects mentioned, but the quality of the match rate has decreased over time:

- 91.3% match rate for 1971 to 1981;
- 89.8% match rate for 1981 to 1991;
- 87.8% match rate for 1991 to 2001;
- 99% match rate Census to health records.

They also found certain subgroups of the population (Males, 15–49, Greater London, Unmarried, Not born in England and Wales) were less likely to be linked.

2.2.2 New Zealand Mortality Study

Linkage Variables: Geo-code information – Meshblock (Median pop size 96) and four other levels of area data, Date of birth, Ethnic group, Gender and Country of birth.

Statistics New Zealand investigated socio-economic mortality gradients in New Zealand for a wide range of socio-economic factors. The study attempted to link 41,310 eligible mortality records to 3,373,896 Census records. They used a strategy of first blocking the population into meshblocks, and then linking the residual by other geocode area variables.

This study found that 76.6% of the eligible mortality records were linked with the estimated true matches greater than 95%. However, the percentage of linked records decreased as time between death and census increased: 79.3% of deaths within six months were linked compared with 72.8% of deaths 30–35 months after census. This decline was found to be the greatest for 15–24 year olds (63.0% to 49.0%) and least for

65–74 year olds. They also found that various ethnic groups had lower linkage rates relative to the rest of the population.

2.2.3 Australian Hospital Separations and Aged Care Admission linkages

Linkage Variables: Statistical Local Area, Date of birth, Gender, Date of hospital separation and Date of aged care entry.

The Australian Institute of Health and Welfare (2003) and Karmel (2004) linked hospital separation records to admissions in residential aged care. Analyses were undertaken to establish the quality of matching based on the probability of chance matches on the residential aged care database, the hospital morbidity database, and the linked database, on the basis of various combinations of the proposed linkage variables. Two measures were used:

- the probability of matching by chance due to the distribution of Birth dates; and
- the rate of such chance matches among achieved matches.

Key findings were that for small regions (<10,000 people), the maximum rate of linkage between non-matching records was 2% and for larger regions this rose to be greater than 3%.

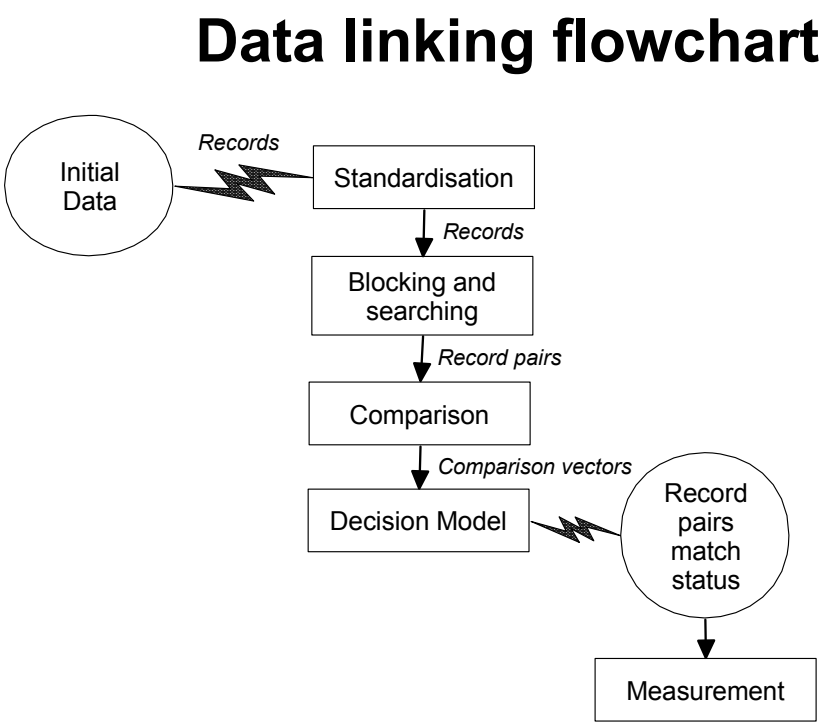
Question: Should we develop these concepts used in the AIHW study to be used as a measure of quality for linking census data?

3. METHODOLOGY

3.1 Process for linking datasets

From the methodology of linkage records described in Section 2.1, the process of linking records has been implemented using the following general framework. It is also the process used by the data linkage software packages being considered.

3.1 Data linking flowchart, reproduced from Gu, et al. (2003)



The following is a brief description of each step.

3.1.1 Standardisation

The contents of two datasets need to be standardised to allow comparison between two different data sources. Standardisation may involve removing inconsistencies and parsing text fields such as Name and Address. With census data, the processes already exist to clean and standardise input data to a high standard.

3.1.2 Blocking and searching

As outlined in Section 2.1.2, blocking reduces the number of comparisons needed to be made by only comparing record pairs where links are more likely to be found. In test runs of relatively small files, test Census 2004 (2,586 people) were linked to

people in the same Collection Districts in Census 2001 (38,117 people). If every single record pair was considered there would be over 95 million possible comparisons. However, by using a blocking strategy of Collection District (CD), Age and Sex on this test run we can reduce this to just 12,032 comparisons. With larger census data sets, blocking is critical to make the record comparison process efficient and feasible given time and computational constraints.

Blocking variables are most effective when they break up the population into small groups. Comparing 100 small groups with 5 people in each only requires $100 \times 5^2 = 2{,}500$ comparisons whereas 10 groups with 50 people in each group requires $10 \times 50^2 = 25{,}000$ comparisons. Hence blocking by CD or Date of birth variables are particularly effective as they break the population into small groups.

One problem with blocking is the opportunity to link a pair of matched records may be missed if blocking variable values disagree. Approximately 10% of Australians move each year and 42% of Australians moved between the 1996 Census and the 2001 Census. If the Address five years ago has not been written down or coded correctly, then using Collection District as a blocking variable may cause more people who have moved to be missed as the record pair will not even be compared.

Most software packages allow for different blocking strategies to be used. Some allow a second pass through the data using another blocking strategy while others allow for multiple blocking strategies to be implemented simultaneously.

3.1.3 Comparison

Record pairs from files *A* and *B* are only compared when the blocking fields agree. From the Fellegi–Sunter methodology in Section 2.1.1, each field is compared and the result is a set of binary codes, stored in a comparison vector, γ, which can be 1, 0 or missing.

This idea has been expanded further to give a *field weight* for each linking variable depending on how well they agree between the datasets. The comparison options we have looked at include:

- Exact match (e.g. Sex). The field either matches or it doesn't.
- Key difference tolerated (e.g. Name). Allows the weight to depend on number of characters that are different, allowing for misspellings, poor handwriting, etc..
- Numerical difference (e.g. Date of birth). Allows the weight to depend on how far apart values (day, month, year) are.
- Geographical difference (e.g. Address or Meshblock). Can use spatial information to calculate distance between fields.

3.1.4 Decision model

The decision model takes the comparison vector γ and assigns the *field weight* based on the results from the comparison model. Then the final *record weight* for each record pair is calculated by summing up the individual *field weights*. Finally, a decision rule based on cut-offs determines whether the record pair is a link, non-link or possible link.

Field weights

The *field weights* are calculated using a technique similar to that discussed in Section 2.1.3. For our linkage runs we have used the m and u probabilities derived directly from the data after blocking as discussed by Jaro (1989). For this, the m and u probabilities are defined slightly differently from Section 2.

$$m_k = P\left(\text{Field } k \text{ agree} \middle| (a,b) \in M\right) \tag{3.1}$$

$$u_k = P\left(\text{Field } k \text{ agree} \middle| (a,b) \in U\right) \tag{3.2}$$

The *field weight* w_k for each field k is calculated as:

$$w_k = \begin{cases} \log\left(\dfrac{m_k}{u_k}\right) & \text{, Agree} \\ \log\left(\dfrac{1-m_k}{1-u_k}\right) & \text{, Disagree} \\ 0 & \text{, Missing} \end{cases} \tag{3.3}$$

The *field weight* is always positive when the fields agree and negative when the fields disagree. By allowing the m and u values to vary this allows some fields to be given additional weight relative to other fields to reflect that some fields have lower error rates (Sex relative to Occupation) or are better at identifying records (e.g. Date of birth relative to Sex).

The field comparison is critical in developing an optimal linkage strategy for ABS data. Most studies have used Names and Date of birth or Address to be their linkage variables. As we only have Date of birth or Age and some geographic information (Meshblock), we must rely on other census variables to find unique records and link them (over time).

However, with more variables it is hard to know which variables to give more weight to, especially when most records have the same value such as Birthplace of individual

where 73% of individuals were born in Australia in 2001. Hence, for those records Birthplace does not differentiate well. However, for people born overseas, especially from smaller less common countries, Birthplace identifies extremely well.

To take this effect into account we use a frequency-dependent *field weight* calculation for most census variables. The frequency of each field value (i) is listed in a frequency table, which is used to derive the frequency probability of this entry

$$P(N_{ik}) = \frac{\delta_{ik}}{\sum_{i=1}^{I} \delta_{ik}} \tag{3.4}$$

where

i = each value within a field k

I = number of unique values within a field

δ_{ik} = frequency of value i in field k

$P(N_{ik})$ = relative frequency probability of value i in field k.

The frequency field weight w_k is then calculated using the following formula

$$w_k^{freq} = \begin{cases} \log\left(\dfrac{1}{P(N_{ik})}\right) & \text{, Agree} \\ \log\left(\dfrac{1-m_k}{1-u_k}\right) & \text{, Disagree} \end{cases} \tag{3.5}$$

Question: With the frequency adjusted weight there is nothing to adjust the field's agreement weight relative to other fields, as in equation (3.3). Can we incorporate the two methods for calculating weights together?

Record weights and cut-offs

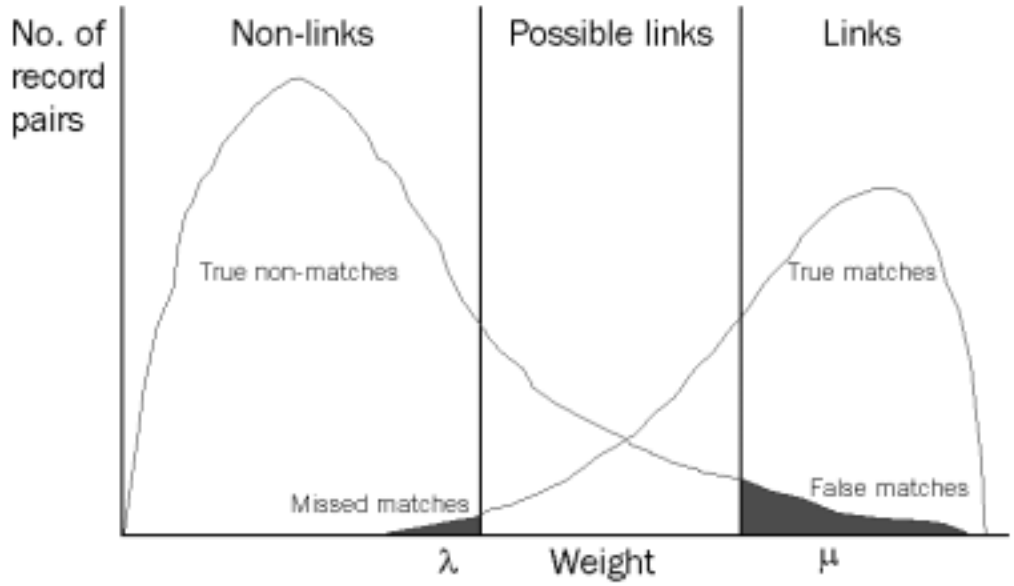
The *record weight* (W) for each record pair is calculated by summing the field weights for all fields.

$$W = \sum_k w_k \tag{3.6}$$

A decision rule is then used based on upper and lower cut-offs to determine whether the record pair is a link, non-link or possible link.

Figure 3.2 shows a typical distribution of record weights for true match and non-match records. This shows the trade off between assigning cut-offs to maximise the number of true matches assigned as links while minimising the record pairs for manual review, missed matches (λ) and false matches (μ).

3.2 Theoretical distribution of final weights and assignments for a typical matching exercise, Jaro (1989)



This methodology outlined above involves assigning weight cut-offs such that any weight above the upper cut-off is assigned a link and any weight below another cut-off is assigned a non link and any weight in between the cut-offs is a possible link.

Christen (2002) has implemented a one-to-one linking algorithm based on 'auction theory' developed by Bertsekas (1992). This procedure outputs the unique links above a cut-off and then uses the 'auction algorithm' for records that have multiple links. This maximises the overall weight of all record pair agreements and reduces the probability of records being incorrectly assigned.

3.1.5 Measurement

The main task here is to assess the quality of the links by identifying how many matching record pairs have been missed, and how many unmatched record pairs have been linked (decision model has declared a link, but it is actually a non-link). To date, the true links have not been available, but this analysis is planned in future work. During census processing period, which lasts for approximately one year after the census is run, we will have additional personal information to establish the true links for quality measures.

As the true links have not been available, the focus to date has been on analysing the number of unique and duplicate matches, the distribution of the assigned and not-assigned matches and comparison of results with previous runs to identify what fields differentiate records accurately.

Further research is required to analyse the probability of incorrect matches as has been used in other studies outlined in Section 2.2.

Question: What quality measures should we use to evaluate the best linkage strategy?

4. RESULTS

This section outlines the preliminary studies undertaken so far and specific issues with the data linkage work.

The census dataset is a rich dataset with lots of information. However, many variables are inconsistently reported or may change over time. Hence we have divided the variables being considered into three classes that we expect to be constant over time and variables that we expect to change over time either predictably or unpredictably. Predictably changing is where we can model some expect changes from one value to another (e.g. Marital status can go from 'never married' to 'married', but not the other way).

4.1 Constant and changing person level Census variables

Constant variables	*Predictable change over time*	*Unpredictable change over time*
CD of usual residence	Marital status	Labour force status
Sex	Australian citizenship	Religious affiliation
Age (adjusted by years)	Highest qualification completed	Area of study of highest non-school qualification
Birthplace of individual	Highest level of primary or secondary school completed	Number of people in the household
Birthplace of female parent		Age of eldest person in the household
Birthplace of male parent		Age of youngest person in the household
Indigenous status*		
Year of arrival		
Ancestry*		

* For these variables, people's perception may change over time

Question: Is the assumption of independence on matching variables valid for these variables chosen? Are there any strategies we can use to minimise any multicollinearity?

4.1 Census 2001 to itself

From the list of constant and changing variables we analysed how well they uniquely identified people and find any particular groups that were hard to identify. For this study, we used the entire Australian population in Census 2001.

4.2 Percentage of records uniquely identified

Model	*Constant variables*	*Changing variables*	*% Uniquely identified*
1. Blocking variables only	CD Age Gender		6.7%
2. Selected constant variables	CD Age Gender Birthplace of individuals and parents		44.8%
3. All constant variables	CD Age Gender Indigenous status Birthplace of individuals and parents Year of arrival Ancestry		76.1%
4. Constant and predictably changing over time	CD Age Gender Indigenous status Birthplace of individuals and parents Year of arrival Ancestry	Marital status Citizenship Highest qualification Highest level of schooling	89.3%
5. All variables included	CD Age Gender Indigenous status Birthplace of individuals and parents Year of arrival Ancestry	Marital status Citizenship Highest qualification Highest level of schooling Labour force status Religious affiliation Area of study Number of people in household Eldest in household Youngest in household	98.4%

From this analysis, we found that a combination of constant and changing variables was required to uniquely identify above 80% of the population. Populations that were hard to match were identified and implications are further discussed in Section 4.1.1.

Using the constant variables, we were only able to uniquely identify 76.1% of the population. Including variables changing predictably over time increased this to 79.3% and the addition of unpredictably changing variables increased this further to 98.4%.

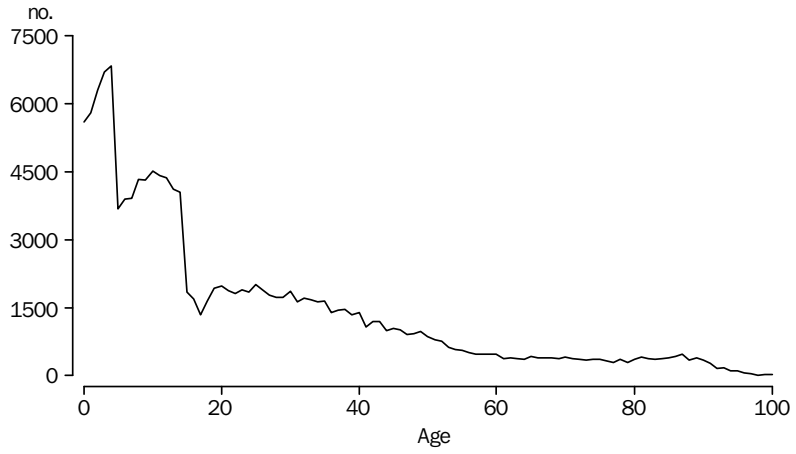
This analysis is based on just the 2001 data, and gives a likely upper limit assuming there is no error or change in status over time. The results for the 2006 Census should be higher providing Date of birth and Meshblock are well reported. There are also other variables available such as Number of issue for females and Family type.

However, many variables will change over a five year period making it difficult to link records and a large risk is that Address five years ago is not accurately reported making it difficult to follow records over time.

4.1.1 Groups difficult to match

Of the 1.6% of the population that was not uniquely identified, further analysis was carried out on this group looking at their common characteristics. Most of these records were attributable to not stated / inadequately described geographic fields. However, there were some definite groups of people in the data. These included people in prisons, defence bases, indigenous children (possibly in missions), and younger Australians, who were born in Australia, have Australian ancestry and are non-indigenous.

4.3 Three or more people with the same characteristics by Age from Census 2001



People under the age of 15 were difficult to uniquely identify as various characteristics such as education, Marital status and occupation are not applicable. Younger people were also likely to have similar characteristics such as 'Studying' and 'Never married'. A summary of where there were three or more people with the exact same characteristics by Age is shown in figure 4.3.

Indigenous children were also of concern as there were up to 29 children with the exact same information. They were found to be in private dwellings, but all had identical ages, language, religion and previous address characteristics. If this is not addressed it may have implications for indigenous linkage studies.

Finally, the defence and prison populations had many duplicates (up to 58 people with the same characteristics). While these populations will most likely be excluded in the 5% sample, there is an issue that we may not be able to uniquely identify someone who is in the sample, but may reside in one of these establishments on Census night in 2011.

4.1.2 Implications for linkage

While this data does not address the issue of fields changing over time or being misreported, it does give us an indication of an upper limit to the percentage of unique matches we could get using the various linkage strategies. It also assists in the selection of variables that can be used to uniquely identify records to match with. Finally, it has also highlighted particular groups that we will have trouble linking.

> Question: For records that are hard to identify uniquely, should statistical matching be considered to maximise the number of records in the linked dataset?

4.2 Census 2001 to Ecensus 2004

This section summarises the key results of linking methods of censuses data across time, using freely extensible biomedical record linkage (Febrl) software.

4.2.1 Data

In this analysis, we attempted to link respondents of the 2004 Ecensus who were at their main residence, had not moved in the last year or arrived in Australia after 2001 (2,586 observations) to respondents of the 2001 Census who lived in the same Collection Districts as the Ecensus (38,167 observations). These data sets were chosen as we expect to match most observations on the Ecensus to the 2001 Census while keeping the data sets considered a reasonable size to test many different linkage strategies. We expect that 22% of the 2,586 observations will have moved into those areas between 2001 and 2003, which will not be able to be linked. However, we expect to be able to link the remaining 78% (2,017 observations).

4.2.2 Blocking strategy

The blocking strategy used for these test runs was Collection District, Age and Gender. This reduced the number of comparisons required from over 98 million (2,586 × 38,167) to only 12,032 comparisons. This also gave marked improvements in time taken to compare records.

4.4 Comparison time of different blocking strategies

Blocking variables	Linkage time
Collection District	11 minutes, 9 seconds
Collection District, Sex	4 minutes, 48 seconds
Collection District, Sex, Age	4 seconds

4.2.3 Comparison

Many different comparison models and strategies were tested. Using exact match agree/disagree weights as defined in equation (3.3), the results showed that one record on the Ecensus linked to multiple records on the Census with high weights. On closer inspection this was found to occur for multiple fields agreeing, such as Ancestry, Country of birth and Indigenous status. Hence the frequency-dependent calculation as outlined in equation (3.5) was used.

Using CD, Age and Sex as blocking variables, two comparison strategies are included in this report:

A. *Constant variables only*: Year of arrival, Birthplace of parents with non-frequency-dependent field comparison functions and Birthplace of individual, Ancestry and Indigenous status as frequency-dependent.

B. *Constant and changing variables*: Year of arrival, Labour force status, Number in household, Eldest and Youngest in household were used as linking variables with non-frequency-dependent field comparison functions. Birthplace of individual and parents, Indigenous status, Ancestry, Marital status, Highest level of schooling, Religion, Qualification level and Qualification field.

4.2.4 Decision model

No decision rule as outlined in Section 2.1 or Section 3.1.4 was used at this preliminary stage, except for an arbitrary cut-off of zero for weights. Any record pairs with weight less than zero were declared non-links. To focus on how the linkage is working and assist in gaining knowledge of how to set cut-offs, the results focus on the distribution of linkage pairs for various choices of weights.

4.2.5 Examination of final output

From the final agreement weights, the number of unique, duplicate matches and missed links for any record weight cut-off greater than zero were calculated.

For this analysis, a record pair is defined as a link if its record weight is greater than the weight cut-off; otherwise, we define it as a non-link. In addition, we define the following quantities:

- The number of unique matches – the number of records on the smaller dataset which link only one record on the larger dataset for a given weight cut-off.
- the number of duplicate matches – the number of links which cannot be true links (with respect to the smaller dataset) for a given weight cut-off. If, for example, record 6 on data *A* links records 101, 105 and 107 on data *B*, record 6 contributes two to the number of duplicate links.
- the number of missed links – the number of records on the smaller dataset which match no records on the larger dataset for a given weight cut-off.

4.5 Model A uniques, duplicates and misses for a given weight cut-off

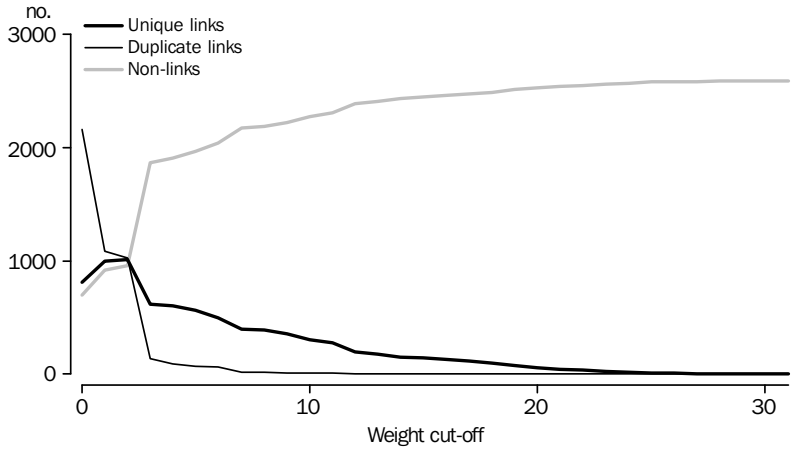
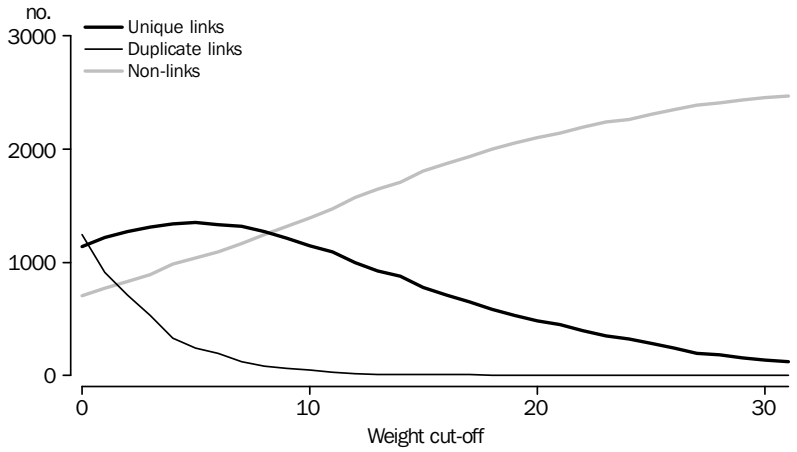


Figure 4.5 shows that as the weight cut-off increases, the number of missed links increases as you expect. A key issue for census data is the high number of duplicates with low weights. These are records that are hard to distinguish because they are Australian born, non-indigenous, with Australian or English ancestry. By using frequency-dependent weights, these duplicate records have a small weight meaning other variables have a greater explanatory power. The maximum number of unique links is just over 1,000 links out of the possible 2,586 at a low weight cut-off of 2 and it can be seen there are virtually no duplicates at a cut-off of 7 or above. To improve the linking we need to use variables that change over time.

4.6 Model B uniques, duplicates and misses for a given weight cut-off



Model B shown in figure 4.6 is the best model achieved so far by the linkage team. At a cut-off of 5, the number of unique matches is maximised with 1,349 uniques and only 243 duplicate links.

Using the one-to-one auction algorithm described in 3.1.4 to automatically assign links and a minimum cut-off of zero, 1,832 links were allocated (70% of records). However, some records only had a very small positive weight and there is no easy method of evaluating the quality of these links without further information.

4.2.6 Analysis of results

Our results so far have shown that using only constant variables does not differentiate between records sufficiently. With full census data we hope this can be improved with the ability to use Date of birth which will be captured on the 2006 Census. However, the Date of birth field has an option to put Date of birth OR Age, so these results are indicative of what we may be able to achieve for people who have not moved and only write down Age in years.

These graphs relate to figure 3.2 shown earlier. As the weight increases the number of records linked decreases. A good sign of quality is when we have high positive links for unique record pairs – i.e. there is one good link for each record on file *A* to file *B*. When there are duplicates this indicates there is a greater chance of allocating a false link, as we are not as definite on which possible link is the best.

For this study we have excluded people who have moved one year ago or are not at their usual residence. However, for forming the Statistical Longitudinal Census Dataset (SLCD), we will not have these constraints. We expect the linkage quality to be lower for this population as people cannot accurately remember where they were five years ago.

5. SUMMARY AND FUTURE WORK

5.1 Summary

As there is no unique identifier or simple rule to link census records together, deterministic matching is infeasible. Hence probabilistic matching is required and we have demonstrated that it needs to match on a range of constant and changing variables.

A key issue identified early on is that most linkage that is done worldwide either has a unique identifier such as Medicare number or uses Name and either Address or Date of birth as linking variables. As ABS does not keep unique identification numbers and will not be retaining Name and Address information for privacy reasons, there is a reduction in the expected quality of matching and no way from the available information to confirm that the two records linked together are actually a match. The methodology has focussed on identifying key alternative variables that can uniquely identify records and that are constant over time. Initial findings show that the constant variables do not uniquely identify records adequately and hence variables that change over time must be considered. This may have an impact on the quality of matches for people whose circumstances and location changes dramatically, which may affect results for some analysis.

The preliminary results without Date of birth show that the best proportion of uniquely identified census records is 98.4%. From preliminary studies linking census data over time we have been able to link 72% of records successfully. However, including Date of birth information and Meshblock information will hopefully reduce the reliance on alternative variables and improve matching rates. This will not be applicable for all people however.

5.2 Future work

Dress Rehearsal 2005 to Census 2006

The main quality assessment of census linkage techniques consists of linking the Census dress rehearsal 2005 and Census 2006 during census processing period. This process will allow us to test various linkage strategies, develop quality statements and test systems used for linking datasets over time. Linkages being considered include creating a 'true link' file using Name and Address and using it to compare linkage runs without Name and Address. Using the 'true link' file we will be able to calculate true links, false links and missed links and evaluate the performance of various assignment algorithms.

Births and deaths to census data

We plan to link births data to census data primarily for quality assessment, but also to develop systems and methods for bringing together deaths and the SLCD recognising it would be in the absence of name and address information. This will improve the quality of the output by reducing the number of records recorded as either a missed or false link.

Other proposals for quality studies using migration data and a disease register are being considered.

Question: Is there any key data linking methodology/data set that ASB should look at that have not been discussed?

ACKNOWLEDGEMENTS

The authors would like to thank the valuable work and contributions of Marion McEwin, Jonathan Khoo, Jeffrey Wright, Jianke Li and Charity Liaw.

BIBLIOGRAPHY

Australian Bureau of Statistics (2001) *Census of Population and Housing: Population Growth and Distribution, Australia*, cat. no. 2035.0, ABS, Canberra.

—— (2005) *Census of Population and Housing: Census Data Enhancement Project – Statement of Intention*, ABS, Canberra. [http://www.abs.gov.au/websitedbs/D3110124.NSF/0/F49CD4A018C93E93CA2570600079BF0E?Open]

Australian Institute of Health and Welfare (2003) *Interface between hospital and residential aged care: Feasibility study on linking hospital morbidity and residential aged care data*. [http://www.aihw.gov.au/publications/age/ibhrac/]

Bertsekas, D. (1992) "Auction algorithms for network flow problems: A tutorial introduction", *Computational Optimization and Applications*, 1, pp. 7–66.

Blakely, T. (2001) *Socio-Economic Factors and Mortality Among 25–64 Year Olds. The New Zealand Census–Mortality Study*, Doctoral Thesis, University of Otago.

Blakely, T., Salmond, C. and Woodward A. (1999) *Anonymous record linkage of 1991 census records and 1991–94 mortality records*, NZCMS Technical Report No. 1, Department of Public Health, Wellington School of Medicine, University of Otago, Wellington.

Blakely, T., Salmond, C. and Woodward, A. (2000) "Anonymous linkage of New Zealand mortality and census data", *Australian and New Zealand Journal of Public Health*, 24(1), pp. 92–95.

Christen, P. and Churches, T. (2002) *Febrl – Freely extensible biomedical record linkage*, ANU Computer Science Technical Report TR-CS-02-05, Australian National University, Canberra. [http://datamining.anu.edu.au/linkage.html]

Christen, P., Churches, T. and Hegland, M. (2004) "Febrl - A parallel open source data linkage system", *Proceedings of the 8th PAKDD'04 (Pacific–Asia Conference on Knowledge Discovery and Data Mining)*, Springer Lecture Notes in Artificial Intelligence (3056), Sydney.

Christen, P. and Goiser, K. (2005) *Quality and complexity measures of data linkage and deduplication*, ANU Computer Science Technical Report TR-CS-02-05, Australian National University, Canberra. [http://datamining.anu.edu.au/linkage.html]

Churches, T., Christen, P., Lim, K. and Zhu, J.X. (2002) "Preparation of name and address data for record linkage using hidden Markov models", *BMC Medical Informatics and Decision Making*, 2, p. 9. [http://www.biomedcentral.com/1472–6947/2/9]

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

Gu, L., Baxter, R., Vickers, D. and Rainsford, C. (2003) "Record linkage: Current Practice and Future Directions", *CMIS Technical Report No. 03/83*, CSIRO Mathematical and Information Sciences, Australia. [http://datamining.csiro.au]

Jaro, M.A. (1989) "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida", *Journal of the American Statistical Association*, 89, pp. 414–420.

—— (1995) "Probabilistic linkage of large public health data files", *Statistics in Medicine*, 14, pp. 491–498.

Karmel, R. (2004) *Linking hospital morbidity and residential aged care data: Examining chance matching*. [http://www.aihw.gov.au/publications/index.cfm/title/10065]

Winkler, W.E. (1988) "Using the EM algorithm for weight computation in the Fellegi–Sunter model of record linkage", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1988, pp. 667–671. [http://www.census.gov/srd/papers/pdf/rr2000–05.pdf]

—— (1989) "Frequency-based matching in the Fellegi–Sunter model of record linkage", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1989, pp. 778–783. [http://www.census.gov/srd/papers/pdf/rr2000–06.pdf]

—— (1994) *Advanced Methods for Record Linkage*, Technical Report, Statistical Research Division, U.S. Bureau of the Census, Washington, DC. [http://www.census.gov/srd/papers/pdf/rr94–5.pdf]

—— (1999) "The state of record linkage and current research problems", *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 1999, pp. 73–80. [http://www.census.gov/srd/papers/pdf/rr99–04.pdf]

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 ABS statistics can be downloaded free of charge from the ABS web site.

WEB ADDRESS www.abs.gov.au

2000001524251
ISBN 0 642 48180 6

RRP $11.00

1352.0.55.076 • RESEARCH PAPER: EXPLORING METHODS FOR CREATING A LONGITUDINAL CENSUS DATASET (METHODOLOGY ADVISORY COMMITTEE) • November 2005