Australian Bureau of Statistics

Information Paper

Australian Census Longitudinal Dataset: Methodology and Quality Assessment

Australia

2006-2011

2080.5

www.abs.gov.au

ABS Catalogue No. 2080.5

Produced by the Australian Bureau of Statistics.

INQUIRIES

For further information about these and related statistics, contact the National Information and Referral Service on 1300 135 070.

CONTENTS

SECTION 1 – INTRODUCTION

The Australian Census Longitudinal Dataset (ACLD) uses data from the Census of Population and Housing to build a rich longitudinal picture of Australian society. This product was formerly known as the Statistical Longitudinal Census Dataset (SLCD). In this first release, a sample of almost one million records from the 2006 Census (Wave 1) was brought together with corresponding records from the 2011 Census (Wave 2) to form the largest longitudinal dataset in Australia. As subsequent Censuses are added to the ACLD, its value as a resource for longitudinal population studies will increase.

This paper describes the background and rationale for the ACLD, the data linkage methodology used and an assessment of its quality.

1.1 OVERVIEW

In 2005, the ABS embarked on a project to enhance the value of Census data by bringing it together with other datasets, both ABS and non-ABS, to leverage more information from the combination of datasets than would be available from the individual datasets separately. The ACLD was proposed as an enduring longitudinal dataset constructed through the linking of records from successive Censuses.

As part of the development phase, a quality study was undertaken in which data from the 2005 Census Dress rehearsal were linked to data from the 2006 Census. This quality study concluded that the linkage methodology was feasible and that the expected quality of the linked data file would be sufficient for longitudinal analysis. For more information see, *Assessing the Likely Quality of the Statistical Longitudinal Census Dataset* (cat. no. 1351.0.55.026).

As a result of the positive assessment from this quality study, a 5% random sample (979,661 records) was selected from the 2006 Census to comprise Wave 1 of the ACLD. This sample was then brought together with data from the 2011 Census using data linkage techniques, resulting in a linked data file consisting of 800,759 records.

Data linkage is typically undertaken using probabilistic and/or deterministic methods, both of which were used in the ACLD project:

- Probabilistic: linkage is based on the level of overall agreement on a set of variables common to the two datasets. This approach allows links to be assigned in spite of missing or inconsistent information, providing there is enough agreement on other variables to offset any disagreement.

- Deterministic: linkage involves assigning record pairs across two datasets that match exactly or closely on common variables. This type of linkage is most applicable where the records from different sources consistently report sufficient information and can be an efficient process for conducting linkage.

In addition, the ABS refers to three types of linkage which are based on the variables used. These can be broadly grouped in order of linkage quality:

- Gold: linking using name, address and personal characteristics such as age and sex
- Silver: linking using an encrypted, non-identifiable numeric version of name and personal characteristics
- Bronze: linking using only personal characteristics

Bronze linkage with both deterministic and probabilistic components was used to combine the 2006 Census sample and the 2011 Census. This method was selected based on the type of information available for linkage and the results from the quality study that linked the 2005 Census Dress Rehearsal and the 2006 Census. The quality study had investigated the relative suitability of Gold, Silver and Bronze methods and concluded that, whilst linkage using name and address information would provide a high quality match, a Bronze linkage would still yield a dataset of sufficient quality for longitudinal analysis. This study also identified that use of a non-identifying, grouped numeric code (hash code) based on name (Silver linkage) could also improve the quality and efficiency of the linkage process in the future. For more information see, *Assessing the Likely Quality of the Statistical Longitudinal Census Dataset* (cat. no. 1351.0.55.026).

At each Census, the ACLD will be augmented with a sample of children who have been born and immigrants who have arrived in Australia since the previous Census, to maintain the size of the longitudinal dataset.

For many individuals the linkage process will have accurately matched their 2006 Census record with the corresponding record from the 2011 Census. In some cases, the link will represent different people who share a number of characteristics in common. Some inaccuracy in the linkage will not generally affect statistical conclusions drawn from the linked data, although care should be taken in the interpretation of results. For more information see the 'Data linking methodology' chapter.

1.2 BENEFITS OF THE ALCD

Each five-yearly Census provides a rich set of information about Australian people and households at a point in time. The Census provides information on characteristics such as age, sex, Indigenous status, country of birth and year of arrival; together with topics such as family structure, education and qualifications, work - including hours worked, occupation and industry - income and housing, and presence of a severe or profound disability. It is able to provide a rich picture of social and economic conditions at a particular point in time, and how these conditions are changing over time and across population groups.

The ACLD adds the ability to study changing patterns in social and economic conditions at the individual level, gives insight into the pathways that tend to lead to particular outcomes, and how these pathways vary for different population groups. It aims to help in the development of strategies to promote positive pathways and avoid negative ones, and assist policy makers in assessing both the social and financial benefits of related intervention strategies.

The ACLD is accessible through the ABS Survey TableBuilder, which is an online tool, enabling continuous access to data. In built confidentiality routines in TableBuilder ensure that no information that is likely to enable identification of an individual or household is released. For more information about access to the ACLD in ABS Survey TableBuilder see, *Microdata: Australian Census Longitudinal Dataset, 2006-2011* (cat. no. 2080.0).

The data linking process used to create the ACLD included a series of steps which can be generalised into the following:

- standardisation of data
- blocking
- record pair comparison
- decision model

2.1 STANDARDISATION

Before records on two datasets are compared, the contents of each need to be as consistent as possible to facilitate comparison. This process is known as 'standardisation' and includes a number of steps such as verification, recoding and re-formatting fields, and parsing text fields (i.e. separating text fields into their components). Additionally, some fields require substantial repair prior to standardisation.

Some variables, such as age, differ between the two datasets in a predictable way, and an adjustment is required to account for this difference. Some variables are coded differently at different points in time, and concordances may be necessary to create variables which align on the two datasets. Variables may also be recoded or aggregated in order to obtain a more robust form of the variable. Standardisation takes place in conjunction with a broader evaluation of the dataset, in which potential linking variables are identified.

The standardisation procedure for the ACLD linkage project involved coding imputed and invalid values for selected variables to a common missing value. These variables include day of birth, month of birth, year of birth, age, sex, year of arrival and marital status. Standardisation for hierarchical fields involved collapsing at higher levels of aggregation to minimise disagreement when linking records which may have had a small intercensal change or to allow for potential differences in the coding of the variable. This allows for records to agree using broader categories rather than disagree on specific information that may have changed over time or be reported and/or coded inconsistently. An example of this is country of birth. Whereas in 2006 the respondent may have been coded to ‘Northern Europe’, in 2011 they may have reported a specific country such as ‘England’ or ‘Norway’. If left in its original state, a comparison between 'Northern Europe' and 'England' would not agree, even though one is a sub-category of the other. Variables grouped in this manner included country of birth, occupation, field and level of qualification, language spoken and religion.

2.2 BLOCKING

Once data files have been standardised, record pairs (consisting of one record from each file) can be compared to see whether they are likely to be a match, i.e. belong to the same person. However, if the files are even moderately large, comparing every record on File A with every record on File B is computationally infeasible. Blocking reduces the number of comparisons by only comparing record pairs where matches are likely to be found – namely, records which agree on a set of blocking variables. Blocking variables are selected based on their reliability and discriminatory power. For instance, sex is partially useful as it is typically well reported, however it is minimally informative as it only divides datasets into two blocks, and is thus used in conjunction with other variables.

The process of blocking reduces the computational intensity of data linking. However, comparing only records that agree on a particular set of blocking variables means a record will not be compared with its match if it has missing, invalid or legitimately different information on a blocking variable. To mitigate this, the linking process is repeated a number of times ('passes'), using a range of different blocking strategies. For example, on the first pass, a block using a low level of geography (Mesh Block) was used to capture the majority of 2006 Census records that had matching information with their corresponding 2011 Census record. This approach meant, however, that those persons that did not have the same Mesh Block were not compared and conversely a potentially better match may exist in another Mesh Block. To mitigate this, a conservative approach to identify and confirm links was adopted in the early passes. This ensured that records which failed to link in the first pass proceeded to the next pass, in which a different set of blocking variables was used. Each pass used a different combination of blocking and linking variables to ensure each record pair had the highest possible chance of being linked. The blocking variables used for each pass are outlined in section 2.4.

2.3 RECORD PAIR COMPARISON

There were two different linking methods utilised in the linkage of the ACLD, deterministic and probabilistic. Deterministic linkage methods were used in the first two passes to identify matches that had high quality linking information. Probabilistic linking was then used in subsequent passes.

2.3.1 DETERMINISTIC LINKING

Deterministic data linkage, also known as rule-based linkage, involves assigning record pairs across two datasets that match exactly or closely on common variables. This type of linkage is most applicable where the records from different sources consistently report sufficient information to efficiently identify links. It is less applicable in instances where there are problems with data quality, or where there are limited characteristics.

Initially, a deterministic linkage method was used to identify matches that contained high quality linking information. This involved combining selected personal and demographic characteristics into a single field (e.g. Sex, age, Mesh Block), which was then used to merge the two files together, identifying each record pair.

2.3.2 PROBABILISTIC LINKING

Probabilistic linking allows links to be assigned in spite of missing or inconsistent information, providing there is enough agreement on other variables to offset any disagreement. In probabilistic data linkage, records from two datasets are compared and brought together using several variables common to each dataset (Fellegi & Sunter, 1969). A key feature of the methodology is the ability to handle a variety of linking variables and record comparison methods to produce a single numerical measure of how well two particular records match. This allows ranking of all possible links and optimal assignment of the link or non-link status (Solon and Bishop, 2009).

Within a blocking pass, records on the two files which agree on the specified blocking variables are compared on a set of linking fields. Each linking field has associated field weights, which are calculated prior to comparison. Field weights indicate the amount of information (agreement, disagreement, or missing values) a linking field provides about whether the records belong to the same or a different person (true match status). Field weights are based on two probabilities associated with each linking field: first, the probability that the field values agree on a record pair given that the two records belong to the same person (match); and second, the probability that the field values agree on a

record pair given the two records belong to different persons (unmatch). These are called m and u probabilities (or match and unmatch probabilities) and are defined as:

$$m = P(\text{fields agree} \mid \text{records belong to the same entity})$$

$$u = P(\text{fields agree} \mid \text{records belong to different entities})$$

Given that the m and u probabilities require knowledge of the true match status of record pairs, they cannot be known exactly, but rather must be estimated. For estimating m and u probabilities for the ACLD, the ABS used the Expectation Maximisation (EM) algorithm (see Samuels, 2012). In some instances the EM algorithm is deemed unsuitable, or fails to converge on an estimate, and in such cases m and u probabilities are based on those of similar linking projects. Note that m and u probabilities are calculated for each pass, conditional on agreement on the specified blocking fields, as all records compared will agree on blocking variables.

Match (m) and unmatch (u) probabilities are then converted to agreement and disagreement field weights. They are as follows:

$$\text{Agree} = \log_2\left(\frac{m}{u}\right)$$

$$\text{Disagree} = \log_2\left(\frac{1-m}{1-u}\right)$$

These equations give rise to a number of intuitive properties of the Fellegi–Sunter framework. First, in practice, agreement weights are always positive and disagreement weights are always negative. Second, the magnitude of the agreement weight is driven primarily by the likelihood of chance agreement. That is, a low probability of two random people agreeing on a field (for example, Date of Birth) will result in a large agreement weight being applied when two records do agree.

The magnitude of the disagreement weight is driven by the stability and reliability of a variable. That is, if a variable is well-reported and stable over time (for example, Sex) then disagreement on the variable will yield a large negative weight. For each record pair comparison, the field weights from each linking field are summed to form an overall record pair comparison weight or 'linkage weight'.

Before calculating m and u probabilities for some variables it is first necessary to define what constitutes agreement. Typical comparison functions include:

- Exact match (e.g. Sex). Agreement occurs only when the two field values are identical. This criterion is used for most linking fields.
- Logical movement (e.g. Highest year of schooling). Agreement occurs when the two numerical field values are identical, with interpolated weights attributed as the field values increase/decrease in a pre-specified direction. A pair may be defined to agree if their field values differ by an amount less than or equal to a specified maximum difference.
- Numeric difference (e.g. Age). A pair may be defined to agree if their field values differ by an amount less than or equal to a specified maximum difference.

For further details on comparison functions used for probabilistic linkage, see Christen & Churches (2005).

Alternatively, near or partial agreement may be factored into the linking process by converting m and u probabilities to weights. For example, a person's age on equivalent records will frequently be an exact match, and the m and u probabilities are calculated based on this definition. During linkage, however, a partial agreement weight was given for ages within two years difference to cater for persons who may have understated their age in 2006 and overstated it in 2011 or vice versa.

Blocking fields, linking fields, comparator types, and m and u probabilities are used as input parameters for the linking software. Records which agree on the blocking variable(s) are compared on all linking fields.

2.4 BLOCKING AND LINKING STRATEGY USED IN THE ACLD

After examining the information available and the quality of the data, a blocking and linking strategy for the ACLD was developed. This was based on the quality study previously undertaken by the ABS (*Assessing the Likely Quality of the Statistical Longitudinal Census Dataset* (cat. no. 1351.0.55.026)). The final strategy employed for linking the 2006 Census sample to the 2011 Census followed on from this work, but contained tighter quality controls and a more comprehensive approach than the two pass approach used in the original quality study. A key feature of the enhanced strategy was the identification and specific targeting of key sub-populations and trying to maximise their opportunity to be linked.

The main features of the enhanced blocking strategy used for the ACLD were:

- more linking runs/passes
- more restrictive blocks in order to maximise linkage quality
- combination of deterministic and probabilistic linkage techniques
- combination of clerical review and decision rules for filtering
- targeted sub-populations in linking and clerical review, such as children, and Aboriginal and Torres Strait Islander people, to improve the quality of these sub-populations
- using information about other members of the household to assist in linkage and clerical review in order to improve the overall quality of the final linked file.

Another feature of the new approach was the ability to be responsive to any under-represented sub-populations in the current linkage file. Accordingly, at the end of each pass, the remaining unlinked records were analysed to determine the best approach to be undertaken for the next pass. This allowed for blocking fields to be customised in order to broaden the search for the remaining unlinked records and for tolerances to be relaxed for some linking fields in later passes of the linkage.

Table 1 displays the blocking and linking fields applied in this linking project for each pass.

TABLE 1 - BLOCKING AND LINKING FIELDS, By pass number and method

	LINKING METHOD											
	Deterministic		Probabilistic									
PASS NUMBER	1	2	3	4	5	6	7	8	9	10(a)	11	12
CENSUS FIELDS												
Personal information												
Age	B	B	L	L	L	L	L	L	L	L	L	L
Sex	B	B	B	B	B	B	B	B	L	L	L	L
Day and Month of Birth	B	B	B	B	B	. .	L	L	L	L	L	L
Indigenous status	B	B	B	B	B	L	B	. .	L	L	L	L
Birthplace	. .	. .	L	L	L	L	. .	. .	L	L	L	. .
Year of Arrival	. .	. .	L	L	L	. .	. .	. .	L	L	L	. .
Marital status	. .	. .	L	L	L	. .	L	L	. .	. .	. .	. .
Level of Qualification	. .	. .	L	L	L	. .	. .	. .	L	L	. .	. .
Field of Qualification	. .	. .	L	L	L	. .	L	L	L	L	L	. .
Highest year of Schooling	. .	. .	L	L	L	. .	. .	. .	L	L	. .	. .
Occupation	. .	. .	. .	. .	. .	. .	L	L	. .	. .	. .	. .
Religion	. .	. .	L	L	L	. .	. .	. .	. .	. .	. .	. .
Language spoken	. .	. .	L	L	L	. .	. .	. .	L	L	. .	. .
Aged less than 15 block	. .	B	. .	. .	. .	B	. .	. .	. .	. .	. .	. .
Household information												
Mothers Age	. .	L	. .	. .	. .	L	. .	. .	. .	. .	. .	L
Mother's Day and Month of Birth	. .	L	. .	. .	. .	L	. .	. .	. .	. .	. .	L
Fathers Age	. .	. .	. .	. .	. .	. .	. .	. .	. .	. .	. .	L
Father's Day and Month of Birth	. .	. .	. .	. .	. .	. .	. .	. .	. .	. .	. .	L
Family ID block	. .	. .	. .	. .	. .	. .	. .	. .	. .	. .	B	. .
Geographic information												
Mesh Block	B	. .	B	. .	. .	. .	B	. .	. .	B	. .	. .
SA1	. .	. .	. .	. .	. .	. .	. .	B	B	. .	. .	. .
SA2	. .	B	. .	B	. .	. .	. .	. .	. .	. .	. .	. .
SA4	. .	. .	. .	. .	B	B	. .	. .	. .	. .	. .	B

(a) The results of Pass 10 were used to identify the blocking field to be used in Pass 11. As a result, there were no records output from Pass 10.

2.5 DECISION MODEL

It is important to note that even where the original data is of very high quality, the information on equivalent records may not be identical across all the blocking and linking variables. For this reason, several 'passes' are used to optimise the opportunity for equivalent records to be linked, with different combinations of blocking and linking variables for each pass. Records that were not linked on one pass are included in the pool of possible links for the next pass.

In deterministic linking, an exact match is required on each of the variables specified in the blocking and linking strategy (see Table 1). Using this approach, links were only accepted if a single record pair was identified. Where a record was included in more than one possible pair, it was returned to the pool of unlinked records for subsequent passes.

In probabilistic linking, once record pairs are generated, a decision rule determines whether the record pair is linked, not linked or considered further as a possible link. The first phase of this process is automated, in which a record is assigned to its best possible pairing. This process is known as one-to-one assignment. Ideally (and often true in practice) each record has a single, obvious best pairing, which is its true match.

Probabilistic linking projects in the ABS have typically used an auction algorithm to assign optimally one record on the first dataset to one record on the second dataset. The auction algorithm maximises the sum of all the record pair comparison weights through alternative assignment choices, such that if a record A1 on File A links well to records B1 and B2 on File B, but record A2 links well to B2 only, the auction algorithm will assign A1 to B1 and A2 to B2, to maximise the overall comparison weights for all record pairs.

The second phase of the probabilistic decision rule stage takes the output of one-to-one assignment and decides which pairs should be retained as links, and which should be rejected as non-links. This is done by defining cut-off weights against which record pair comparison weights are evaluated. The simplest decision rule uses a single cut-off such that all record pairs with a weight greater than or equal to the cut-off are assigned as links, and all those pairs with a weight less than the cut-off are assigned as non-links. In order to establish the cut-off value, a sample of the record pairs are clerically reviewed. This provides the opportunity to ascertain the level of quality at each link weight and enables an estimate of the number of false links.

A more sophisticated decision rule employs lower and upper cut-off weights. Record pairs with a link weight above the upper cut-off are declared links while those with a weight below the lower cut-off are declared non-links. The record pairs with weights between the upper and lower cut-off weights are not automatically assigned a status, but designated for clerical review where all records within the upper and lower cut-off are reviewed and a judgement about the link status is made manually for each record pair.

As clerical review is a time and labour intensive element of data linkage projects, not all record pairs can be individually reviewed to determine their match status. While it is critical to examine a selection of record pairs manually to assess the quality of the automated linkage process and prepare for the next pass, it is also important to optimise the resource load so as to achieve the best value for effort.

For the ACLD project, a sample of record pairs were clerically reviewed to set a single cut-off in each pass. The single cut-off weight was set at a point where the review showed this was adequate to assign a high proportion of links with high accuracy. In this case, no further clerical review would be performed and unlinked records proceeded to the next pass.

There are some limitations with using a single cut-off. For example, adults for whom a wide range of Census characteristics (such as occupation or educational attainment) is collected will generally have a higher linkage weight than children (for whom there is limited information). Thus, an adult record pair could be positioned above, and a child record pair below the clerical cut-off, even when the adult link is false and the child link is true. Linkage weights for children may also give a false estimate of quality when compared with adult records at the same linkage weight. This could lead to an output bias and an under-representation of children in the final output file. The same considerations apply to adult records that have different amounts of missing information, for example, a record pair of an employed person (for whom there is a range of employment-related information) compared with a record pair for someone who was not in the labour force. To mitigate these factors, specifically targeted passes were conducted throughout the process as well as separate record sampling to determine accurate cut-offs for various sub-populations.

In clerical review, each sampled record pair was manually inspected to resolve its match status. A clerical reviewer is often able to use information which cannot be captured in the automated comparison process, such as common transcription errors (e.g. 1 and 7) or transposed information, such as the day of birth reported as the month or vice versa.

Along with the linking fields, supplementary information was also used to confirm a match. This included:

- Non-linking fields such as Ancestry.
- Frequency counts of personal characteristics, i.e. the number of people with the same date of birth in the same Mesh Block.
- Displaying the dates of birth and ages of other members within the household.
- Providing information on other members in the household that were listed on the Census form as temporarily absent.

These supplementary fields helped to clarify difficult decisions, especially on record pairs belonging to children, allowing for greater insight into whether a record pair was an actual match or just contained similar demographic and personal characteristics for two different individuals.

3. LINKAGE RESULTS

At the completion of the linkage process, 800,759 (82%) out of the 979,661 records from the 2006 Census sample (Wave 1) were linked to a 2011 Census (Wave 2) record to create the linked ACLD. This linkage rate was consistent with results from other Bronze linkage projects using the 2006 and 2011 Census.

All results presented in this publication (unless identified in the relevant table) are based on characteristics from the 2006 Census sample and have been confidentialised to prevent the identification of individuals.

Table 2 displays the linkage rate for a range of sub-populations.

TABLE 2 - LINKAGE RATES, By selected characteristics

	2006 Census sample (no.)	*ACLD* (no.)	*Linkage rate* (%)
Sex			
Male	480 285	390 487	81.3
Female	499 372	410 274	82.2
Age group (years)			
0-14	194 017	170 834	88.1
15-19	66 247	51 220	77.3
20-24	66 512	49 327	74.2
25-29	62 249	48 642	78.1
30-39	140 271	117 655	83.9
40-49	142 911	123 946	86.7
50-59	126 285	108 962	86.3
60-69	86 385	71 906	83.2
70-74	31 004	23 678	76.4
75 and over	63 784	34 586	54.2
Indigenous status			
Non-Indigenous	942 253	775 419	82.3
Aboriginal	19 697	13 340	67.7
Torres Strait Islander	1 449	923	63.7
Both Aboriginal and Torres Strait Islander	839	543	64.7
Not stated	15 416	10 530	68.3
State/Territory of usual residence			
New South Wales	323 136	263 369	81.5
Victoria	244 095	203 668	83.4
Queensland	192 606	154 013	80.0
South Australia	75 481	62 239	82.5
Western Australia	95 795	77 921	81.3
Tasmania	23 787	19 583	82.3
Northern Territory	8 469	6 226	73.5
Australian Capital Territory	16 186	13 680	84.5
Remote areas			
Major Cities	669 274	552 339	82.5
Inner Regional	195 401	159 611	81.7
Outer Regional	92 396	73 122	79.1
Remote	13 989	10 533	75.3
Very Remote	6 546	4 602	70.3
No Usual Address	2 029	539	26.6
Total(a)(b)(c)	**979 661**	**800 759**	**81.7**

(a) Data presented in the table have been perturbed. As a result, the sum of individual categories may not align with totals.

(b) Includes Other Territories.

(c) Includes Migratory areas.

The linkage rates that were achieved for the ACLD were relatively consistent across most sub-populations and were in line with expected results. Compared with the national average of 82%, the sub-populations which achieved the highest linkage rates were persons:

- aged 0 to 14 years (88%), followed by 40 to 49 years (87%) and 50 to 59 years (86%)
- of non-Indigenous origin (82%)
- who usually lived in the ACT (85%) and Victoria (83%)
- who usually lived in Major cities (83%).

The sub-populations which achieved the lowest linkage rates were persons:

- aged 20-24 years (74%) and 75 years and over (54%)
- of Aboriginal (68%), Torres Strait Islander (64%) or both Aboriginal and Torres Strait Islander origin (65%)
- who usually lived in the Northern Territory (74%)
- who usually lived in remote (75%) and very remote areas (70%) or who had no usual address in 2006 (27%).

Traditionally, the Census Post Enumeration Survey (PES) has shown that the Census has higher rates of undercount for people of Aboriginal and/or Torres Strait Islander origin, those aged between 20 and 29 and for those in the Northern Territory. As expected, the lower ACLD linkage rates broadly aligned with the same groups that experience higher levels of undercount in the Census. One additional group that had lower linkage rates were persons aged 75 and over at the time of the 2006 Census who, due to age, had an increased risk of death over the ensuing five years. Further information on Census undercount can be found in *Census of Population and Housing - Details of Undercount, 2011* (cat. no. 2940.0)

Further data cubes, demonstrating the linkage rates for various sub-populations are available as an attachment to this Information paper.

3.1 LINKAGE ACCURACY

The following quality measures were calculated for the ACLD and indicate a good level of overall quality:

1. The linkage rate, that is the proportion of the 2006 Census sample records linked to a 2011 Census record, including the number of true matches and false links.
2. The consistency of reporting of common information between record pairs.

3.1.1 LINKAGE RATES, TRUE AND FALSE LINKS

Not all record pairs assigned as links in a data linkage exercise are a match, that is, a record pair belonging to the same individual. While the methodology is designed to ensure that the vast majority of links are true, some are nevertheless false. The linkage strategy used for the ACLD was designed to achieve both a high number of links and to ensure a high level of accuracy to enable longitudinal research. Accordingly, the strategy was restrictive and conservative, especially in the early passes.

Analysis from the results of clerical review was conducted to determine the quality of the linkage process and estimate the number of true links in the linked ACLD file. This process involved calculating the proportion of rejected record pairs at each linkage weight and determining the amount of false links this would represent in the final output file. Table 3 provides a summary from the results of clerical review, including an estimate of the number of false links accepted in each pass. Due to the nature of deterministic linking and the way in which linked records were retained, no false links were identified in passes 1 and 2. While it is assumed that all links assigned in these passes were true, as they contained consistent information across all key linking fields, in reality there may have been a small but un-quantifiable number of false links.

TABLE 3 - LINKAGE RESULTS, By pass number

Pass number(a) (no.)	*Links created* (no.)	*Sampled in clerical review* (no.)	*Links assigned* (no.)	*Total false links* (no.)	*False link rate* (%)
1	559 182	30	544 925	0.0	0.0
2	131 575	30	10 919	0.0	0.0
3	11 131	240	10 489	997	9.5
4	182 285	400	62 570	9 929	15.9
5	212 071	400	87 248	17 274	19.8
6	57 713	345	18 988	1 832	9.6
7	10 489	206	1 723	237	13.7
8	10 156	120	159	29	18.4
9	236 180	411	50 007	10 712	21.4
11	133 555	201	9 827	1 051	10.7
12	29 911	200	3 903	731	18.7
Total(b)	**1 574 248**	**2 583**	**800 759**	**42 792**	**5.3**

(a) The results of Pass 10 were used to identify the blocking field to be used in Pass 11. As a result, there were no records output from Pass 10.

(b) Data presented in the table have been perturbed. As a result, the sum of individual categories may not align with totals.

The combined clerical review results indicate that the number of false links in the final ACLD file could be as low as 5%. By including a tolerance around these results and assuming a small false link rate for the deterministic passes, the false link rate for the ACLD is estimated to be about 5–10%. The passes that contained the highest proportion of false links were Pass 9 (21.4%), where family information was used to try and resolve unlinked records, and Pass 5 (19.8%), which used a broad geography (SA4) as the blocking field. Whilst this is only an approximate estimate, it does give an indication of the high level of overall quality examined through reviewing a sample of over 2,500 record pairs.

The linkage rate of 82% with a false link rate of 5% was broadly consistent with, or better than, other ABS Census linkage projects which did not use name and address as linkage variables (see *Assessing the Likely Quality of the Statistical Longitudinal Census Dataset* (cat. no. 1351.0.55.026)).

The conservative and restrictive nature of the blocking and linking strategy helped to minimise the number of estimated false links throughout the linkage process accompanied by quality controls that were implemented during clerical review.

About two-thirds (68%) of all links were achieved in the first pass of the project, which used a deterministic linking methodology to identify and filter matches. In Pass 1, a tight geographic and demographic restriction was implemented to maximise the amount of high quality links assigned and to limit the amount of alternative comparisons required. Using this approach, links were only accepted if a single record pair was identified.

3.1.2 CONSISTENCY OF COMMON INFORMATION ON RECORD PAIRS

In data linkage projects, geographic boundaries function as blocking variables that restrict the search for record pairs. They are also used as linking variables, and when combined with other linking fields such as age, sex and date of birth, provide a high level of uniqueness, and reduce the likelihood of linking to an incorrect record.

Table 4 displays the number of records that had consistent information and is grouped by the consistency of the record pairs across varying levels of geography.

TABLE 4 - CONSISTENCY OF LINKED RECORDS, By geography and selected linking fields

	Consistency of key linkage fields(a)(b)	
	(no.)	(%)
Mesh Block		
Age exact, Mesh Block, Sex, DOB Day and Month agree	**552 714**	**69.0**
Age exact, Mesh Block, Sex agree	**41 135**	**5.1**
Age +/- 2 years, Mesh Block, Sex agree	**77 98**	**1.0**
SA2		
Age +/- 2 years, SA2, Sex, DOB Day and Month agree	**84 265**	**10.5**
Age +/- 2 years, SA2, Sex agree	**26 739**	**3.3**
SA4		
Age +/- 2 years, SA4, Sex, DOB Day and Month agree	**66 623**	**8.3**
Total records included	**779 274**	**97.3**
Total records linked	**800 759**	**100**

(a) Only includes records that agree on all key linking fields.

(b) Categories are mutually exclusive. Records that agree in each category are excluded from subsequent categories.

Just over 97% of all records that were matched in the ACLD linkage process agreed on small to medium levels of geographic area combined with other key linking fields, such as age, sex and date of birth. While the number of consistent fields can give a strong indication of likely linkage quality, other factors should be taken into account, for example, the expected number of people in a geographic area that are likely to share a characteristic by chance. A tolerance of plus or minus two years was used at certain parts of the linkage process to cater for persons who may have understated their age in 2006 and overstated it in 2011 or vice versa.

By contrast, record pairs may have inconsistent information and yet be a true link. Inconsistent information may be recorded for the same person in different Censuses due to a range of factors, including:

- Transcription errors in the Census, where the wrong category is selected or the information is transposed, such as the day the person was born being reported in the month instead of as the day field.

- Data capture errors, where the Census form is scanned using Optical Character Recognition software and certain characters may be mis-classified, such as a 1 captured as a 7 or a 3 as an 8.
- Reporting errors, where information is given for the wrong member of the household (e.g. person 1's information is reported for person 3) or where the person completing the Census estimates information that they do not know (e.g. about a fellow group household member).
- Information that was not stated by the respondent and has been imputed as part of Census processing (such as age or sex).
- A different person fills out the Census form at the different time points and interprets the questions differently.

3.1.2.1 Consistent reporting of Indigenous status

Consistency of Indigenous status is a special case, since the change in reporting over time is both a potential indicator of linkage quality, and is of analytical interest.

Results from the 2011 Census observed an unexpected increase in persons who identified as being of Aboriginal and/or Torres Strait Islander origin. This was due, in part, to improvements in Census collection practices that resulted in a more complete enumeration of the Aboriginal and Torres Strait Islander population in 2011 than in 2006. In addition, a significant contributor to this increase, was a change in the propensity of people to identify as being of Aboriginal and/or Torres Strait Islander origin in 2011 compared with 2006 (see *Census of Population and Housing: Understanding the Increase in Aboriginal and Torres Strait Islander Counts, 2006-2011* (cat. no. 2077.0)).

While there was a group of people in the ACLD who were identified as non-Indigenous in 2006 and of Aboriginal and/or Torres Strait Islander origin in 2011, this group was relatively small and was counterbalanced by an almost equally sized group who reported the opposite. This pattern of change is different to that expected, given the increasing propensity of people to identify their Aboriginal and Torres Strait Islander origin observed at the aggregate level in the entire 2011 Census.
Throughout the linkage process, Indigenous status was used as a blocking and linking variable. Whilst this would have only made a small contribution to the linkage weight, this may have increased the likelihood of assigning a link to a record pair that contained consistent information for Indigenous status. Record pairs that contained inconsistent information for Indigenous status still had a good chance of being linked, however, providing there was sufficient additional information available for linking.
Differences in the reporting of Indigenous status between 2006 and 2011 on the ACLD may be due to a range of reasons. These include:

- people deliberately identifying their Indigenous origin differently at the two time points
- false links, where similar but not identical persons have been linked
- data capture errors, where multiple boxes may have been selected
- a different person filling out the Census form at each period of time and interpreting the question on Indigenous status differently
- transcription errors in the Census, where the wrong category is selected by accident.

Table 5 shows the reporting of Indigenous status for the linked records on the ACLD, across the 2006 and 2011 Censuses. Further data cubes, demonstrating a more detailed breakdown, by remoteness areas, are provided as an attachment to this Information paper.

TABLE 5 - CONSISTENCY OF INDIGENOUS STATUS FOR LINKED RECORDS, 2006 and 2011

	2011 INDIGENOUS STATUS			
	Non-Indigenous (no.)	*Aboriginal and/or Torres Strait Islander* (no.)	*Not stated* (no.)	*Total* (no.)
2006 INDIGENOUS STATUS				
Non-Indigenous	766 851	1 697	6 868	775 419
Aboriginal and/or Torres Strait Islander	1 367	13 274	165	14 802
Not stated	9 729	226	575	10 530
Total(a)	**777 946**	**15 205**	**7 609**	**800 759**

(a) Data presented in the table have been perturbed. As a result, the sum of individual categories may not align with totals.

3.2 CHARACTERISTICS OF LINKED AND UNLINKED 2006 CENSUS SAMPLE

The random sample selected from the 2006 Census was intended to be representative of the Australian population, including by age, sex and jurisdiction as well as other characteristics such as Indigenous status, occupation and country of birth.

Table 6 shows the distribution of key populations across the 2006 Census, the 2006 Census sample and the ACLD.

TABLE 6 - SELECTED CHARACTERISTICS, By 2006 Census, ACLD sample, ACLD Linked file

	2006 Census		2006 Census sample		ACLD Original results		ACLD Weighted results(a)	
	(no.)	(%)	(no.)	(%)	(no.)	(%)	(no.)	(%)
Sex								
Male	9 896 500	49.3	480 285	49.0	390 487	48.8	9 193 092	49.4
Female	10 165 146	50.7	499 372	51.0	410 274	51.2	9 432 201	50.6
State/Territory of usual residence								
NSW	6 549 174	32.6	323 136	33.0	263 369	32.9	6 093 946	32.7
Vic.	4 932 422	24.6	244 095	24.9	203 668	25.4	4 624 754	24.8
Qld.	3 904 531	19.5	192 606	19.7	154 013	19.2	3 635 806	19.5
SA	1 514 340	7.5	75 481	7.7	62 239	7.8	1 445 720	7.8
WA	1 959 088	9.8	95 795	9.8	77 921	9.7	1 858 559	10.0
Tas.	476 481	2.4	23 787	2.4	19 583	2.4	465 052	2.5
NT	192 899	1.0	8 469	0.9	6 226	0.8	179 713	1.0
ACT	324 034	1.6	16 186	1.7	13 680	1.7	319 439	1.7
Age group (years)								
0-9	2 579 496	12.9	127 331	13.0	114 298	14.3	2 551 435	13.7
10-19	2 756 102	13.7	132 937	13.6	107 761	13.5	2 541 580	13.6
20-29	2 684 371	13.4	128 760	13.1	97 973	12.2	2 348 176	12.6
30-39	2 893 058	14.4	140 271	14.3	117 655	14.7	2 800 102	15.0
40-49	2 942 353	14.7	142 911	14.6	123 946	15.5	2 868 488	15.4
50-59	2 574 589	12.8	126 285	12.9	108 962	13.6	2 473 152	13.3
60-69	1 733 297	8.6	86 385	8.8	71 906	9.0	1 640 218	8.8
70-79	1 168 675	5.8	58 277	5.9	42 262	5.3	993 893	5.3
80 and over	729 705	3.6	36 502	3.7	16 002	2.0	407 967	2.2
Indigenous status								
Non-Indigenous	18 266 814	91.1	942 253	96.2	775 419	96.8	17 806 585	95.6
	455 027	2.3	21,985	2.2	14,802	1.8	561 088	3.0
Aboriginal and/or Torres Strait Islander								
Aboriginal	407 700	2.0	19 697	2.0	13 340	1.7	507 554	2.7
Torres Strait Islander	29 515	0.1	1 449	0.1	923	0.1	32 876	0.2
Both Aboriginal and Torres Strait Islander	17 812	0.1	839	0.1	543	0.1	20 805	0.1
Not stated	1 133 449	5.6	15 416	1.6	10 530	1.3	257 343	1.4
Overseas visitor	206 357	1.0	0	0.0	0	0.0	0	0.0
Total(b)(c)(d)	**20 061 646**	**100**	**979 661**	**100**	**800 759**	**100**	**18 625 130**	**100**

(a) For more information on weighting see chapter 3.4.

(b) Data presented in the table have been perturbed. As a result, the sum of individual categories may not align with totals.

(c) Includes Other Territories.

(d) Includes Migratory areas.

The distribution of the ACLD file by sub-population was generally well aligned with both the 2006 Census sample and the entire 2006 Census. When looking at the relative difference between these proportions, however, some differences are more clearly observed.

Compared with the entire 2006 Census, the linked ACLD contains relatively more records for people aged 0-9 years, and to a lesser extent those aged 40-49 years, 50-59 years and 60-69 years. By contrast, the ACLD contains relatively fewer records for people aged 20-29 years and 80 years and over. There is also relatively fewer people of Aboriginal and Torres Strait Islander origin in the ACLD, than the entire 2006 Census (1.8% compared with 2.3%). The corresponding weighted estimate, however, represents 3.0% of the total population, which is attributed to benchmarking the 2006 sample to the Aboriginal and Torres Strait Islander population in 2011 and therefore to the higher level of identification observed in the 2011 Census than in 2006 (see section 3.4).

In general, the distribution of weighted counts for the linked ACLD file is close to that of the entire 2006 Census, but it is not designed to produce counts corresponding to the population in 2006. Rather, the weighted population is that of people who were in scope of both the 2006 and 2011 Censuses (see section 3.4). Thus, for example, the lower proportion of older people in the linked file, even after weighting, reflects that impact of deaths on the 2006 sample that occurred between 2006 and 2011.

Further data cubes, demonstrating more detailed population distributions, are provided as an attachment to this Information paper.

3.3 REASONS FOR UNLINKED RECORDS

There are two main reasons why records from the 2006 Census sample were not linked to a 2011 Census record:

1. Records belonging to the same individual were present in the 2006 Census sample and the 2011 Census but these records failed to be linked because they contained missing or inconsistent information.
2. There was no 2011 Census record corresponding to the 2006 Census sample because the person was not counted in the Census.

3.3.1 MISSING AND/OR INCONSISTENT INFORMATION

In these cases, the true match was present in the pool of all record pairs but it was not identified because there was a high level of inconsistency between information on the 2006 Census sample and the 2011 Census record, or key linking fields were missing altogether. The reasons for the match being missed can be categorised into the following groups:

- The missing or inconsistent information did not allow the record pair to be compared in the same blocking categories and could not be linked.
- The record pair did not contain enough common information to distinguish the match from other potential record pairs.
- The record pair was linked, but was attributed a low link weight as it contained a lot of missing or inconsistent information and was positioned below the cut-off identified in sample clerical review.
- The record pair was subjected to clerical review, but the high level of inconsistency did not enable it to be deemed a link.

Accurate address coding was crucial in narrowing the search and differentiating between true and false links. It was a particular challenge for persons who had moved, since linkage was then dependent on the information supplied in 2011 about the person's address in 2006. Processing for the 2011 Census involved coding for address five years ago to a fine level of geography, ideally Mesh Block. This was not always possible, either due to the insufficient detail of address information supplied or because by 2011, Census respondents may not have accurately remembered their address on Census Night in 2006.

3.3.2 NO 2011 CENSUS RECORD

A person included in the 2006 Census sample may have had no equivalent 2011 Census record because they were no longer in scope for the Census due to migration from Australia, or death between 2006 and 2011, or they may simply have been missed in the Census.

According to mortality data compiled by the ABS from data supplied by the Registrars of Births, Deaths and Marriages, about 700,000 people died in Australia between 2006 and 2011. If 5% of these people were represented in the 2006 sample, then it could be expected that up to 35,000 people could not have been linked due to death between 2006 and 2011. Similarly, migration data shows that just over one million people left Australia as permanent emigrants over the same period, potentially resulting in up to 50,000 people from 2006 Census sample being unlikely to have a corresponding 2011 Census record.

Due to the size and complexity of the Census, it is inevitable that some people are missed and some are counted more than once. It is for this reason that the Census Post Enumeration Survey (PES) is run shortly after each Census, to provide an independent measure of Census coverage. The PES determines how many people should have been counted in the Census, how many were missed (undercount), and how many were counted more than once (overcount). It also provides information on the characteristics of those in the population who have been missed or overcounted.

The net undercount rate for the 2011 Census was 1.7%, with a higher rate for Aboriginal and Torres Strait Islander people than for the non-Indigenous population (see *Census of Population and Housing - Details of Undercount, 2011* (cat. no. 2940.0)) Thus, roughly, 15,000 people from the 2006 Census sample could have been missed in the 2011 Census. This estimate is a starting point only and does not take into account the likelihood of people being missed in successive Censuses.

When taking into account all of these factors, it is estimated that over half of the unlinked 2006 Census sample (100,000 out of the 180,000 unlinked records) would not have a corresponding record in the 2011 Census. This would indicate that the initial linkage rate of 82% could be representative of up to 91% of the population that actually had an opportunity to be linked.

3.4 WEIGHTING

Weighting is the process of adjusting a sample to infer results for the relevant population. To do this, a 'weight' is allocated to each sample unit - in this case, persons. The weight can be considered an indication of how many people in the relevant population are represented by each person in the sample. Weights were created for linked records in the ACLD to enable longitudinal population estimates to be produced. Cross-sectional population estimates for 2006 and 2011 are available from each Census.

The ACLD began as a random sample of 5% of the Australian population in 2006. As such, each person in the sample should represent about 20 people in the population. Between Censuses, however, the in scope population changes as people die or move overseas. In addition, Census net undercount and data quality can affect the capacity to link equivalent records across waves. The ACLD weighting process benchmarked the linked ACLD records to the population that was in scope of both the 2006 and 2011 Censuses. The weights were based on four components: the design weight, undercoverage adjustment, missed link adjustment and population benchmarking.

The original population benchmark was the 2011 Estimated Resident Population (ERP). The 2011 ERP was chosen over the 2006 ERP as the baseline population as it is more recent. The ERP was than adjusted to exclude births and overseas arrivals that had occurred between 2006 and 2011.

Weights were benchmarked to the following population groups:

- state by age (ten year groups), by sex, by mobility (interstate arrivals benchmarked separately)
- Indigenous status by state

Note that the ERP by Indigenous status for the period 2006 - 2011 is currently being revised in view of a higher than expected intercensal increase in the number of Aboriginal and Torres Strait Islander persons (see *Census of Population and Housing: Understanding the Increase in Aboriginal and Torres Strait Islander Counts, 2006-2011* (cat. no. 2077)). As a result, weights for the ACLD will be reviewed when this data becomes available.

The initial weights have a mean value of 23.3 and range between 4 and 168. Higher weights are associated with people of Aboriginal and Torres Strait Islander origin and people who moved interstate between 2006 and 2011. For more information see the Appendix.

The ABS plans to continue to build the ACLD by combining it with successive Censuses and, where feasible, linking the ACLD to other datasets. The sample will also be augmented so that it remains representative of the Australian population.

4.1 SAMPLE AUGMENTATION

Once the linkage phase to each Census is complete, the ACLD will be augmented with a sample of children who have been born and immigrants who have arrived in Australia since the previous Census. The first release of the ACLD in TableBuilder will not contain any records from the augmentation as these will be added to the file as part of the linkage to the 2016 Census.

4.2 FUTURE LINKAGE

Following the completion of processing for the 2016 Census, the ACLD (including the augmentation to maintain a 5% sample), will be enhanced further, by linking to the 2016 Census. The main difference for this linkage will be the use of a Silver linking method. Under this approach, personal and demographic information will be used for linkage, but the process will also utilise an encrypted, numeric version of name which is non-identifiable to an individual and is grouped with several other name combinations. It is anticipated that the encrypted numeric code will improve the number and quality of records linked.

The findings from the Information paper: *Assessing the Likely Quality of the Statistical Longitudinal Census Dataset* (cat. no. 1351.0.55.026) indicate that a Silver linkage method could potentially achieve between a 5 and 10 percentage point increase in the linkage rate. These estimates are indicative only as the time period between the Census Dress rehearsal and the Census was only one year, whereas the five year time gap between Censuses adds to the complexity of linking records for the ACLD.

4.3. ACCESS TO THE ACLD

The first release of the ACLD is accessible online through ABS TableBuilder, where clients can build, customise, save and export their own tables and graphs. In this product, confidentiality methods are applied to the data prior to output to ensure that no information that is likely to enable identification of an individual or household will be released.

Both the ACLD and TableBuilder are new innovations from the ABS and feedback is welcomed on their usability, accessibility and relevance. Comments can be sent to <data.integration@abs.gov.au>.

For more information, or to access the ACLD, see *Microdata: Australian Census Longitudinal Dataset, 2006-2011* (cat. no. 2080.0).

REFERENCES

Australian Bureau of Statistics (2013*), Census of Population and Housing: Understanding the Increase in Aboriginal and Torres Strait Islander Counts, 2006-2011*, cat. no. 2077.0.

(2013) *Death registrations to Census linkage project - Key Findings for Aboriginal and Torres Strait Islander peoples, 2011-12*, cat. no. 3302.0.55.005.

(2013) *Life tables for Aboriginal and Torres Strait Islander Australians, 2010-2012,* cat. no. 3302.0.55.003.

(2013) *Microdata: Australian Census Longitudinal Dataset, 2006-2011*, cat. no. 2080.0.

(2012) *Census of Population and Housing - Details of Undercount, 2011,* cat. No. 2940.0.

(2010) *Census Data Enhancement: An Update*, cat. no. 2062.0.

(2009) *Assessing the Likely Quality of the Statistical Longitudinal Census Dataset*, cat. no. 1351.0.55.026.

(2009) *Research Paper: A Linkage Method for the Formation of the Statistical Longitudinal Census Dataset*, August 2009, cat. no. 1351.0.55.025.

Australian Institute of Health and Welfare and Australian Bureau of Statistics (2012). *National best practice guidelines for data linkage activities relating to Aboriginal and Torres Strait Islander people,* cat. no. IHW 74. AIHW, Canberra.

Christen, P. and Churches, T. (2005) *Febrl 0.3 Documentation*, (last viewed on 8 April 2013), <http://cs.anu.edu.au/~Peter.Christen/Febrl/febrl-0.3/febrldoc-0.3/>

Christen, P., Churches, T., and Hegland, M. (2004) "Febrl – A Parallel Open Source Data Linkage System", *Proceedings of the 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia*, pp. 638-647.

Cross Portfolio Statistical Integration Committee (2010), *High Level Principles for Data Integration Involving Commonwealth Data for Statistical and Research Purposes*, CPSIC, Canberra.

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.

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

Samuels, C. (2012) "Using the EM Algorithm to Estimate the Parameters of the Fellegi–Sunter Model for Data Linking", *Methodology Advisory Committee Papers*, cat. no. 1352.0.55.120, Australian Bureau of Statistics, Canberra.

APPENDIX: WEIGHTING THE ACLD

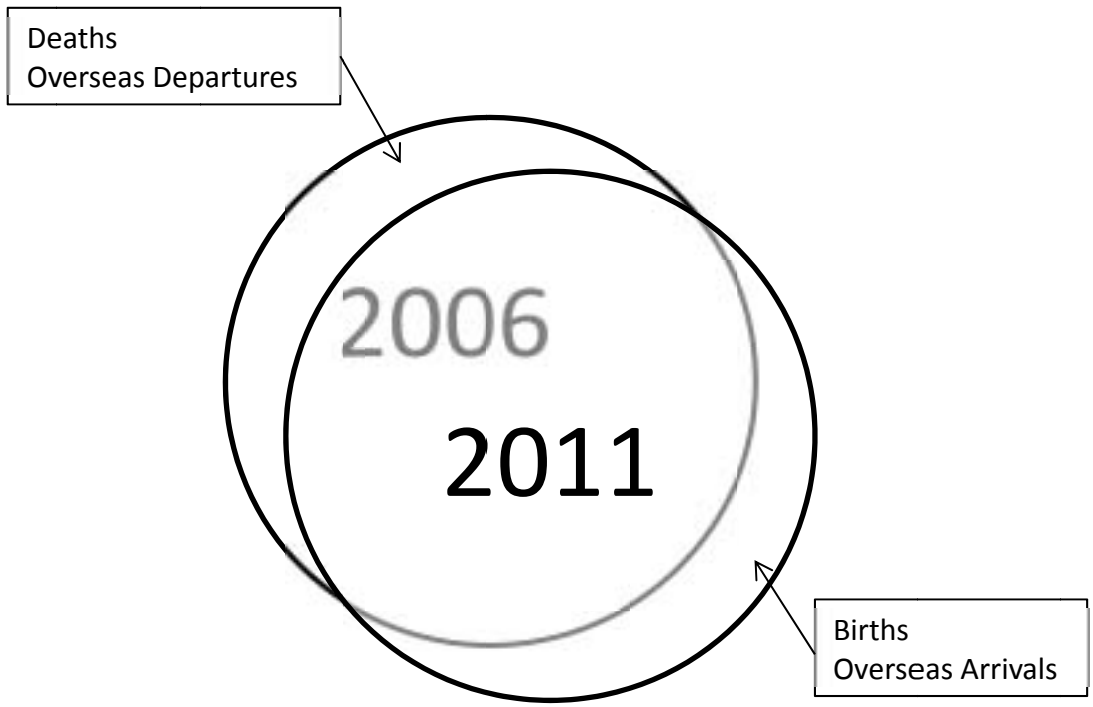
PURPOSE OF WEIGHTING

Weighting is the process of applying factors to a sample to infer results and calculate estimates for the population. To do this, a 'weight' is allocated to each enumerated person. This is a value which indicates how many population units are represented by the sample unit. The purpose of weights is to allow the data user to estimate the number of people in the population with particular characteristics, based on the sample.

The Australian Census Longitudinal Dataset (ACLD) is designed to measure change in Australian society over time, with the first two waves being from the 2006 and 2011 Census. For the first issue of the ACLD, a longitudinal weight has been developed which allows the weighted sample to represent all persons who were in scope of both the 2006 and 2011 Census. As shown in Figure 1, the population in scope or the longitudinal population is the overlap between the two Censuses (the shaded region). To estimate this population, the 2011 Estimated Resident Population (ERP) was reduced by the number of births and overseas arrivals that occurred since the 2006 Census. It could also have been estimated by reducing the 2006 population by the number of deaths and departing migrants between 2006 and 2011. However, as 2011 ERP is more recent, it was used as the starting point.

Cross-sectional estimates for the 2006 and 2011 populations can, and should, be extracted from 2006 or 2011 Census data rather than the ACLD.

FIGURE 1: IN SCOPE POPULATION FOR THE AUSTRALIAN CENSUS LONGITUDINAL DATASET, 2006-2011



DESCRIPTION OF WEIGHTING PROCESS FOR LONGITUDINAL WEIGHTS

Positive weights were calculated for each linked record on the ACLD, that is each 2006 sample record that was successfully linked to a 2011 Census record. No weights were calculated for the unlinked records. The resulting weights in the ACLD are a measure of how many population units each person represents, taking into account both the likelihood that the person was linked, and the general composition of the in scope population. The weights consist of four components. The first two components address the likelihood of a person being selected for the 2006 sample through a sample design and undercoverage adjustment. The third and fourth components adjust the weight on the basis of the longitudinal population in scope of both Censuses, by adjusting for missed links and benchmarking to the relevant population and sub-populations that were at risk of being underrepresented otherwise. The following describes each component in depth.

DESIGN WEIGHT

For a sample survey, the design weight needs to take into account any differential likelihood of selection on the basis of survey design. Given that the 2006 sample of the ACLD was taken from a population Census and the design utilised random selection, the design weight for the ACLD is quite simply the inverse of the probability of selection. Given that the probability of selection is 1 in 20 (5%), the design weight is:

$$W_1 = 20$$

UNDER COVERAGE ADJUSTMENT

In order to represent the full 2006 Estimated Resident Population (ERP), the design weight was then adjusted for the small proportion of people who were in scope for the 2006 Census but did not complete a Census form in 2006. While this proportion varies substantially between demographic groups, the 2006 Census net undercount proportion of 2.7% was used for simplicity. This resulted in an undercoverage adjusted weight of:

$$W_2 = W_1 \times (1 \div 1 - 0.027) = 20.555$$

MISSED LINK ADJUSTMENT

The aim of this component was to account for missed links, that is, 2006 sample records that had corresponding 2011 Census records, but were not linked. No attempt was made to correct for false links. The missed link adjusted weight is the product of the undercoverage adjusted weight and the inverse of the estimated propensity to link.

$$W_3 = W_2 \times (inverse\ of\ the\ estimated\ propensity\ to\ link)$$

The propensity to link was estimated using a logistic regression model that was applied to the 2006 sample, with the response variable being the link status. The logistic regression model describes a relationship between a 2006 sample record's propensity to link and its values for a range of 2006 Census variables such as Indigenous status, marital status, country of birth, language spoken at home and English proficiency, labour force participation and occupation, educational attainment, mobility (whether moved in the preceding year) and remoteness. The estimated propensity to link varied considerably between records.

Two separate models were applied to the 2006 sample. The first model was applied to people under the age of 15 years on 2006 Census night. This model excluded the variables that were not applicable to people under 15 years of age, such as marital status. The second model was applied to the remainder of the sample (persons aged 15 years or over in 2006).

Each model was initially estimated using a training dataset, which consisted of 75% of the respective records. For each model, an out of sample Hosmer-Lemeshow type of analysis was applied to the remaining 25% of the records to determine the estimated propensity ranges for which each model provided a poor fit. For the model applied to the sample that was aged under 15 years, the model significantly underestimated the linkage rates where the estimated propensities were less than 0.65. To improve the estimated propensities, all links for people aged under 15 years on 2006 Census night with estimated propensities less than 0.65 had their estimated propensities set to 0.65. Similarly, all links for people aged 15 years or over on 2006 Census night with estimated propensities less than 0.61 had their estimated propensities set to 0.61.

The missed link adjustment carries the assumptions that the ACLD contains no false links and that all records in the 2006 sample that weren't linked, did have a corresponding 2011 Census record. As with many linked datasets, both of these assumptions are invalid for the ACLD. The violation of these assumptions results in the missed link adjustment correcting not only for missed links, but also for the records in the 2006 sample that weren't linked because they didn't have a 2011 Census record. Therefore the missed link adjustment erroneously corrects also for persons that died between the 2006 and 2011 Census nights, persons that moved overseas between the 2006 and 2011 Census nights and (of less concern because it is an objective of the calibration component) persons that were living in Australia on 2011 Census night but weren't counted. Furthermore, records that are less likely to be linked are expected intuitively to be more likely to be linked incorrectly. Giving these links a higher missed link adjusted weight can increase the influence of false links in the ACLD. The calibration component remediates the over-representation of persons who have died or moved overseas to some extent.

Odds ratios and accompanying Wald confidence intervals for the predictor variables for the first model (for persons aged under 15 years in 2006) are contained in Table A.1. A comparison group is selected for each characteristic, and the odds ratio for the other categories represents the ratio of the odds of being linked in contrast to the comparison group. For instance, Table A.1 shows the odds ratios by age group in 2006. Those aged 8-13 years were less likely to be linked than those aged 0-7 years (the comparison group), but more likely than those aged 14 years. Conversely, the odds ratios for school type in 2006 show that persons attending Catholic schools were more likely to be linked than those attending government school (the comparison group).

TABLE A.1 - ODDS RATIOS FROM THE LOGISTIC REGRESSION MODEL, Persons aged under 15 years, 2006

		95% CONFIDENCE LIMITS		
Selected characteristics	*Odds ratio*	*Low limit*	*Upper limit*	*no.*
Age group				
0-7 years (comparison group)	1.000	. .	. .	101 476
8-13 years	0.651	0.626	0.676	79 293
14 years	0.426	0.403	0.450	13 246
Country of Birth				
Oceania and Antarctica (non-Indigenous persons) (comparison group)	1.000	. .	. .	172 907
Aboriginal and Torres Strait Islander persons	0.449	0.424	0.476	8 246
North-West Europe	0.531	0.481	0.586	2 600
Southern and Eastern Europe	0.587	0.465	0.743	495
North Africa and the Middle East	0.528	0.451	0.619	1 012
South-East Asia	0.620	0.538	0.715	1 436
North-East Asia	0.451	0.386	0.527	1 192
Southern and Central Asia	0.623	0.522	0.745	1 041
Americas	0.414	0.347	0.493	716
Sub-Saharan Africa	0.790	0.675	0.924	1 168
Missing(a)	0.575	0.523	0.632	3 189
Language				
English (comparison group)	1.000	. .	. .	164 431
Other North European	0.617	0.514	0.741	790
Southern European	0.744	0.673	0.823	3 659
Eastern European	0.819	0.713	0.941	2 089
Southwest and Central Asian	0.945	0.843	1.058	4 805
Southern Asian	1.367	1.153	1.622	2 275
Southeast Asian	0.879	0.782	0.988	3 881
Eastern Asian	0.971	0.872	1.081	4 710
Australian Indigenous Languages	1.391	1.118	1.729	650
Other	0.466	0.410	0.531	1 477
Missing(a)	0.983	0.859	1.125	5 238
English Proficiency				
Very Well (comparison group)	1.000	. .	. .	179 427
Well	0.849	0.775	0.929	4 981
Not Well	0.798	0.715	0.890	2 985
Not at All	0.765	0.671	0.871	2 330
Missing	0.852	0.738	0.983	4 294

. . Not applicable

(a) Includes Supplementary codes.

Source: Australian Census Longitudinal Dataset 2006-2011

TABLE A.1 - ODDS RATIOS FROM THE LOGISTIC REGRESSION MODEL, Persons aged under 15 years, 2006 *continued*

		95% CONFIDENCE LIMITS		
Selected characteristics	*Odds ratio*	*Low limit*	*Upper limit*	*no.*
Mobility				
Same usual address one year ago (comparison group)	1.000	. .	. .	160 427
Different usual address one year ago	0.519	0.501	0.536	30 442
Missing	0.524	0.477	0.576	3 154
School sector				
Government (comparison group)	1.000	. .	. .	78 096
Catholic	1.266	1.204	1.330	24 086
Other Non-Government	1.072	1.014	1.134	14 611
Missing/Other(b)	0.813	0.781	0.846	77 222
Remoteness				
Major City (comparison group)	1.000	. .	. .	128 850
Inner Regional	0.840	0.809	0.871	40 240
Outer Regional	0.684	0.654	0.715	19 846
Remote	0.571	0.520	0.628	3 266
Very Remote	0.614	0.538	0.701	1 648
Other(c)	0.447	0.318	0.627	164

. . Not applicable

(b) Includes other school sector and pre-school

(c) Includes Migratory, Offshore and Shipping Zones and No usual address

Source: Australian Census Longitudinal Dataset 2006-2011

Odds ratios and accompanying Wald confidence intervals for the predictor variables for the second model (for persons aged 15 years or over in 2006) are contained in Table A.2. A wider variety of variables were available for this age group. There are some differences between the two models. For instance, English speaking proficiency appears to have a detrimental impact on the propensity to link for persons aged under 15 years, but no clear impact for those aged 15 years or over.

TABLE A.2 - ODDS RATIOS FROM LOGISTIC REGRESSION MODEL, persons aged 15 years or over, 2006

		95% CONFIDENCE LIMITS		
Selected characteristics	*Odds ratio*	*Low limit*	*Upper limit*	*no.*
Country of Birth				
Oceania and Antarctica (non-Indigenous persons) (comparison group)	1.000	. .	. .	557 076
Aboriginal and Torres Strait Islander persons	0.563	0.541	0.586	13 741
North-West Europe	0.75	0.734	0.766	67 836
Southern and Eastern Europe	0.837	0.806	0.869	36 573
North Africa and the Middle East	0.772	0.724	0.823	11 663
South-East Asia	0.730	0.695	0.767	26 628
North-East Asia	0.548	0.515	0.584	18 435
Southern and Central Asia	0.839	0.787	0.894	12 622
Americas	0.515	0.488	0.543	8 383
Sub-Saharan Africa	0.764	0.723	0.809	8 708
Missing	0.795	0.768	0.822	23 980
Language				
English (comparison group)	1.000	. .	. .	640 446
Other Northern European	0.722	0.682	0.765	7 336
Southern European	0.916	0.885	0.948	35 926
Eastern European	0.796	0.760	0.835	17 496
Southwest and Central Asian	0.853	0.801	0.909	14 771
Southern Asian	0.875	0.806	0.950	10 266
Southeast Asian	0.901	0.849	0.956	17 267
Eastern Asian	0.820	0.773	0.870	24 864
Australian Indigenous Languages	1.871	1.638	2.137	1 401
Other	0.407	0.379	0.438	3 680
Missing	0.686	0.639	0.737	12 201
English Proficiency				
Very Well (comparison group)	1.000	. .	. .	714 344
Well	1.057	1.024	1.091	38 298
Not Well	1.138	1.093	1.186	19 108
Not at All	0.932	0.862	1.007	3 278
Not stated	1.012	0.937	1.093	10 611
Mobility				
Same usual address one year ago (comparison group)	1.000	. .	. .	642 109
Different usual address one year ago	0.520	0.512	0.527	129 782
Missing	0.577	0.552	0.603	13 742

. . Not applicable

Source: Australian Census Longitudinal Dataset, 2006-2011

TABLE A.2 - ODDS RATIOS FROM LOGISTIC REGRESSION MODEL, persons aged 15 years or over, 2006 *continued*

		95% CONFIDENCE LIMITS		
Selected characteristics	*Odds ratio*	*Low limit*	*Upper limit*	*no.*
Remoteness				
Major City (comparison group)	1.000	. .	. .	540 431
Inner Regional	0.909	0.895	0.923	155 158
Outer Regional	0.798	0.782	0.814	72 549
Remote	0.639	0.609	0.669	10 721
Very Remote	0.536	0.501	0.572	4 902
Other(b)	0.094	0.084	0.105	1 894
Registered marital status				
Married (comparison group)	1.000	. .	. .	393 344
Separated	0.468	0.454	0.483	24 208
Divorced	0.577	0.564	0.589	64 060
Widowed	0.379	0.370	0.387	46 980
Never Married	0.543	0.534	0.551	257 053
Highest year of school completed				
Year 12 (comparison group)	1.000	. .	. .	347 382
Year 11	1.121	1.097	1.146	82 239
Year 10	1.191	1.171	1.212	187 565
Year 9 or below(a)	0.874	0.858	0.891	122 886
Not stated	0.628	0.610	0.647	45 562
Labour force status and Occupation				
Not in the labour force (comparison group)	1.000	. .	. .	272 133
Unemployed	1.087	1.054	1.121	25 661
Employed				
Professional	1.724	1.679	1.770	93 407
Manager	1.494	1.455	1.534	61 709
Technicians and trades	1.527	1.488	1.567	67 989
Community and personal service	1.430	1.390	1.471	41 932
Clerical and administrative	1.873	1.826	1.922	70 225
Sales workers	1.552	1.510	1.595	46 283
Machinery operators and drivers	1.367	1.324	1.410	31 422
Labourers	1.362	1.329	1.397	49 376
Employed, occupation not stated	1.091	1.032	1.153	7 922
Not stated	0.839	0.807	0.872	17 573

. . Not applicable

(a) Includes persons who did not go to school.

(b) Includes Migratory, Offshore and Shipping Zones and No usual address

Source: Australian Census Longitudinal Dataset, 2006-2011

TABLE A.2 - ODDS RATIOS FROM LOGISTIC REGRESSION MODEL, persons aged 15 years or over, 2006 *continued*

		95% CONFIDENCE LIMITS		
Selected characteristics	*Odds ratio*	*Low limit*	*Upper limit*	*no.*
Level of non-school qualification				
No post-school qualification (comparison group)	**1.000**	**. .**	**. .**	**388 254**
Postgraduate Degree	**1.391**	**1.334**	**1.451**	**21 363**
Graduate Diploma and Graduate Certificate	**1.940**	**1.823**	**2.066**	**11 868**
Bachelor Degree	**1.807**	**1.763**	**1.853**	**94 628**
Advanced Diploma and Diploma	**1.584**	**1.542**	**1.626**	**58 854**
Certificate	**1.864**	**1.828**	**1.900**	**138 647**
Level of non-school qualification not stated or inadequately described	**0.925**	**0.904**	**0.947**	**72 033**
Student status				
Not studying or studying part time (comparison group)	**1.000**	**. .**	**. .**	**685 704**
Full time student	**1.200**	**1.174**	**1.226**	**71 221**
Not stated	**0.812**	**0.786**	**0.839**	**28 716**

. . Not applicable

Source: Australian Census Longitudinal Dataset, 2006-2011

CALIBRATION TO KNOWN POPULATION TOTALS

The missed link adjusted weight was calibrated so that the resulting weighted counts of the ACLD links would be equal to estimates of the longitudinal population size at the national and selected sub-national levels. For the ACLD, weights were calibrated to two sets of benchmarks simultaneously using a 'raking' tool. This is a program which was developed to determine record level weights using iterative horizontal and vertical passes through the unit records until a satisfactory set of weights are converged upon. To mitigate against the possibility of the tool producing calibrated weights that were less than one, lower bounds for the calibrated weights were set to 20% of the missed link adjusted weight. Upper bounds were not necessary because extremely high weights were not produced.

The first set of benchmarks comprise state/territory, by interstate migration, by sex, by ten year age group population benchmarks. There were two interstate migration groups, with the first group consisting of the population that resided in the given state/territory on 30 June 2006 and 30 June 2011, and the second group consisting of the population that resided in the given state/territory on 30 June 2011 but were in a different state/territory on 30 June 2006 (i.e. interstate arrivals). The interstate migration groups served to correct for the lower linkage rates among people who moved interstate between 30 June 2006 and 30 June 2011. The second set of benchmarks comprised Indigenous status (according to the 2011 Census) by state/territory.

Note that the ERP by Indigenous status for the period 2006 - 2011 is currently being revised in view of a higher than expected intercensal increase in the number of Aboriginal and Torres Strait Islander persons (see *Census of Population and Housing: Understanding the Increase in Aboriginal and Torres Strait Islander Counts*, 2006-2011, ABS cat. no. 2077.0). As a result, weights for the ACLD will be reviewed when this data becomes available.

The first set of benchmarks were estimated by first dividing the 30 June 2011 ERP into the state/territory, by interstate migration, by sex, by ten year age groups and then subtracting the number of overseas arrivals between 30 June 2006 and 30 June 2011 from each of the groups. Births between 30 June 2006 and 30 June 2011 were automatically excluded because the youngest age group consisted of those aged 5-14 years on 30 June 2011. Groups that had very small ERPs were merged together. For example, the male ERP for those aged 75 to 84 years and those aged 85 years or over on 30 June 2011 who resided in the Northern Territory during the intercensal period were summed. As a result, the first set of benchmarks comprises 275 age by sex by state/territory groups.

These benchmarks are displayed in Table A.3.

TABLE A.3: BENCHMARKS OF THE LONGITUDINAL POPULATION, By state/territory, interstate migration status, sex and age, 2006-2011

SAME STATE/TERRITORY 2006 AND 2011			INTERSTATE ARRIVALS		
Age group (years)	*Males* (no.)	*Females* (no.)	*Age group* (years)	*Males* (no.)	*Females* (no.)
NEW SOUTH WALES					
5-14	400 217	377 186	5-14	27 114	25 824
15-24	375 838	358 689	15-24	32 410	33 108
25-34	298 798	303 881	25-34	58 229	56 585
35-44	399 107	419 910	35-44	36 421	34 076
45-54	436 653	451 411	45-54	20 721	19 089
55-64	383 051	386 488	55-64	16 116	16 039
65-74	260 573	267 604	65-74	8 796	8 030
75-84	148 054	184 237	75-84	3 232	3 315
85 or over	46 742	89 666	85 or over	957	2 111
VICTORIA					
5-14	293 457	278 736	5-14	19 935	19 470
15-24	293 685	284 251	15-24	26 554	27 042
25-34	251 013	253 892	25-34	46 619	47 420
35-44	318 557	335 260	35-44	28 823	27 454
45-54	330 277	345 319	45-54	15 808	14 476
55-64	285 499	295 800	55-64	11 026	11 337
65-74	192 330	201 673	65-74	5 503	5 173
75-84	110 940	139 196	75-84	2 194	2 414
85 or over	34 859	66 440	85 or over	752	1 511

Source: adjusted 2011 Estimated Resident Population

TABLE A.3: BENCHMARKS OF THE LONGITUDINAL POPULATION, By state/territory, interstate migration status, sex and age, 2006-2011 *continued*

SAME STATE/TERRITORY 2006 AND 2011			INTERSTATE ARRIVALS		
Age group (years)	*Males* (no.)	*Females* (no.)	*Age group* (years)	*Males* (no.)	*Females* (no.)
		QUEENSLAND			
5-14	242 570	230 766	5-14	33 372	31 600
15-24	230 975	223 291	15-24	38 001	39 376
25-34	175 661	181 449	25-34	58 638	54 444
35-44	235 954	248 810	35-44	41 256	38 129
45-54	253 252	265 598	45-54	25 371	23 461
55-64	226 449	227 398	55-64	18 067	17 434
65-74	152 418	152 725	65-74	8 928	7 643
75-84	77 295	92 815	75-84	3 079	3 377
85 or over	23 290	42 614	85 or over	1 007	2 204
		SOUTH AUSTRALIA			
5-14	83 980	80 417	5-14	7 552	7 167
15-24	90 740	85 131	15-24	8 565	9 117
25-34	75 882	72 981	25-34	13 782	13 171
35-44	90 517	90 871	35-44	9 435	9 215
45-54	102 486	105 325	45-54	6 312	5 911
55-64	92 646	96 117	55-64	4 773	4 512
65-74	62 766	67 417	65-74	2 279	1 882
75-84	37 710	47 404	75 or over	922	1 194
85 or over	12 584	24 669			
		WESTERN AUSTRALIA			
5-14	119 638	115 820	5-14	11 813	11 326
15-24	125 121	120 334	15-24	13 894	13 257
25-34	97 067	99 008	25-34	33 410	24 585
35-44	125 149	128 262	35-44	17 519	14 944
45-54	136 019	140 324	45-54	10 267	8 637
55-64	119 256	120 800	55-64	6 145	5 169
65-74	74 713	75 856	65-74	2 102	1 702
75-84	39 089	48 564	75 or over	753	1 064
85 or over	11 514	22 013			

Source: adjusted 2011 Estimated Resident Population

TABLE A.3: BENCHMARKS OF THE LONGITUDINAL POPULATION, By state/territory, interstate migration status, sex and age, 2006-2011 *continued*

SAME STATE/TERRITORY 2006 AND 2011			INTERSTATE ARRIVALS		
Age group (years)	*Males* (no.)	*Females* (no.)	*Age group* (years)	*Males* (no.)	*Females* (no.)
		TASMANIA			
5-14	27 963	25 968	5-14	4 193	3 987
15-24	28 402	25 856	15-24	3 789	4 277
25-34	19 788	19 587	25-34	6 532	6 882
35-44	25 882	27 131	35-44	5 186	5 301
45-54	31 718	32 470	45-54	3 746	3 862
55-64	30 159	29 824	55-64	3 460	3 585
65-74	20 900	20 977	65-74	1 873	1 630
75-84	11 243	13 710	75 or over	638	804
85 or over	3 439	6 434			
		NORTHERN TERRITORY			
5-14	11 523	10 716	5-14	4 955	4 684
15-24	9 123	7 858	15-24	8 072	7 126
25-34	2 011	4 557	25-34	14 309	11 643
35-44	8 451	9 768	35-44	6 823	5 832
45-54	9 968	10 064	45-54	4 362	3 863
55-64	8 592	7 310	55-64	2 836	2 403
65-74	4 147	3 367	65 or over	(a)1 801	
75 or over	1 465	1 586			
		AUSTRALIAN CAPITAL TERRITORY			
5-14	14 539	13 902	5-14	5 323	5 077
15-24	16 672	15 417	15-24	9 211	9 154
25-34	9 776	9 951	25-34	15 579	14 755
35-44	14 942	16 162	35-44	8 359	7 787
45-54	17 341	19 609	45-54	4 437	3 876
55-64	16 017	17 140	55-64	2 293	2 291
65-74	9 593	10 236	65-74	900	937
75-84	4 751	5 937	75 or over	637	1 111
85 or over	1 388	2 548			

(a) Males and females were placed into a single group for this benchmark category

Source: adjusted 2011 Estimated Resident Population

After setting these benchmarks, the data was assessed for how well sub-populations were represented. Aboriginal and Torres Strait Islander persons were underrepresented at this stage, partly owing to intercensal growth in this sub-population (see *Census of Population and Housing: Understanding the Increase in Aboriginal and Torres Strait Islander Counts, 2006-2011*, ABS cat. no. 2077.0). At the time of publication, finalised ERP data by Indigenous status for 2006 was unavailable - this data is due for release early in 2014. As a result, the second set of benchmarks was

estimated by applying the rate of growth for the Aboriginal and Torres Strait Islander population from 2006-2011 from previous projections (*Experimental Estimates and Projections, Aboriginal and Torres Strait Islander Australians, 1991 to 2021*, ABS cat. no. 3238.0) to the 2011 ERP for Aboriginal and Torres Strait Islander persons (using the B series of projections) and removing deaths and interstate departures between 2006 and 2011. Overseas departures were not estimated or removed.

The Indigenous status benchmark groups comprised 17 state/territory by Indigenous status groups, where Indigenous status was either 'Aboriginal/Torres Strait Islander' or 'Not Aboriginal/Torres Strait Islander' (including both non-Indigenous and not stated). Due to the small population size in Other Territories, this benchmark was not disaggregated by Indigenous status.

The benchmarks by Indigenous status are displayed in Table A.4.

TABLE A.4: BENCHMARKS OF THE LONGITUDINAL POPULATION, By state/territory and Indigenous status, 2011

State/Territory	*Aboriginal and Torres Strait Islander persons* (no.)	*Other persons(a)* (no.)
New South Wales	181 529	5 808 749
Victoria	41 805	4 582 890
Queensland	164 462	3 564 255
South Australia	33 001	1 392 431
Western Australia	78 036	1 817 098
Tasmania	21 078	440 118
Northern Territory	60 841	128 374
Australian Capital Territory	5 431	302 217

(a) Includes non-Indigenous persons and persons who did not state an Indigenous status in 2011.

Source: adjusted 2011 Estimated Resident Population

The mean weight for selected characteristics gives an indication of how much the weight has been increased or reduced from the initial probability of selection (which would give a weight of 20) in order to address missed links and Census undercount. Table A.5 shows that the mean weight for the linked records is 23.3 - that is, each person in the linked dataset generally represents just over 23 persons in the population. The largest weight was 168 and the smallest was 4. The mean weight was higher for Aboriginal and Torres Strait Islander persons and for people who had moved, particularly interstate.

TABLE A.5: DESCRIPTIVE STATISTICS FOR WEIGHTS, By selected characteristics, 2011

	no.	*Minimum Weight*	*Maximum Weight*	*Mean Weight*	*Standard Deviation*	*Median Weight*
Sex						
Male	390 467	4.3	168.1	23.5	8.1	21.4
Female	410 293	4.3	129.8	23	7.3	20.9
Age group (years)						
0-14	114 222	12.1	104.4	22.3	6.6	20.4
15-24	109 254	8.4	129.8	23.6	8.4	21.2
25-34	98 277	4.3	168.1	23.9	12.9	20.1
35-44	117 135	10.5	98.5	23.8	8.1	21.6
45-54	123 606	13.7	116.1	23.2	6.3	21.5
55-64	108 853	17	109.3	22.7	5.3	21.3
65-74	71 782	17.2	91.9	22.8	4.6	21.6
75-84	41 998	16.2	91.9	23.5	4.4	22.3
85 or over	15 630	15.9	118.4	25.6	5.5	25.5
Indigenous status						
Aboriginal and/or Torres Strait Islander	15 205	5.9	168.1	38.6	11	36.8
Other(a)	785 552	4.3	132.1	23	7.3	21.1
State/Territory (Usual Residence)						
NSW	260 720	16.7	107.2	23	7.1	21.1
Vic.	203 623	17	118.4	22.7	6.5	21
Qld	156 438	16.3	106.3	23.8	8.2	21.3
SA	61 904	17.3	103.8	23	6.9	21.1
WA	78 373	17.5	114.8	24.2	8.5	21.8
Tas.	19 726	16.2	114.9	23.4	7.9	20.9
NT	6 213	4.3	168.1	30.5	23.6	19.5
ACT	13 705	11.1	88	22.4	11.8	18.4
Other Territories	62	27.5	91.1	45.7	19.3	39.4

(a) Includes non-Indigenous persons and persons who did not state an Indigenous status in 2011.

Source: Australian Census Longitudinal Dataset 2006-2011

TABLE A.5: DESCRIPTIVE STATISTICS FOR WEIGHTS, By selected characteristics, 2011 *continued*

	no.	*Minimum Weight*	*Maximum Weight*	*Mean Weight*	*Standard Deviation*	*Median Weight*
Interstate arrivals						
State/Territory (Usual Residence 2011)						
NSW	6 654	34.4	107.2	60.4	11.5	59.5
Vic.	5 512	38.4	118.4	56.9	9.4	55.4
Qld	8 409	35.9	106.3	52.9	10.6	49.9
SA	1 871	40.4	103.8	56.6	10.2	53.5
WA	2 931	36.4	114.8	60.2	10.2	57.6
Tas.	1 255	29.2	114.9	47.5	13.7	42.7
NT	1 165	47	168.1	67.5	18.7	62.4
ACT	1 846	31.5	88	49.6	10.6	49.4
Other Territories	14	74.5	91.1	80.8	5.4	79.6
Moved last year						
State/Territory (Usual Residence 2011)						
NSW	26 476	12.8	132.1	25.9	12.9	21.4
Vic.	21 090	8.9	118.4	25.6	11.8	21.4
Qld	21 829	4.6	116.1	27.5	13.1	21.9
SA	6 130	5.1	101.5	27	12.3	22.2
WA	9 712	15.4	114.8	28.3	14.1	22.4
Tas.	2 215	16.2	104.7	27.9	13.6	22.2
NT	1 034	4.3	168.1	46.1	30.2	50.4
ACT	1 912	11.2	88.1	31	17.3	21.9
Other Territories	9	10.7	91.1	81.9	30.9	76.3
Total Persons	**800 759**	**4.3**	**168.1**	**23.3**	**7.7**	**21.2**

Source: Australian Census Longitudinal Dataset 2006-2011

FOR MORE INFORMATION . . .

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.

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

FAX 1300 135 211

EMAIL client.services@abs.gov.au

PHONE 1300 135 070

FREE ACCESS TO STATISTICS

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

WEB ADDRESS www.abs.gov.au