Full Title: **The Australian Census Longitudinal Dataset: Using Record Linkage to Create a Longitudinal Sample from a Series of Cross-Sections**

Running Title: **The Australian Census Longitudinal Dataset**

James Chipperfield*
Australian Bureau of Statistics
ABS House
45 Benjamin Way
Belconnen ACT 2617
james.chipperfield@abs.gov.au
*corresponding author

James J Brown
Australian Research Council Centre of Excellence for Mathematical & Statistical Frontiers.
School of Mathematical & Physical Sciences

University of Technology Sydney
15 Broadway, Ultimo NSW 2007
James.Brown@uts.edu.au

Nicole Watson
Melbourne Institute of Applied Economic and Social Research
Level 5, FBE Building, University of Melbourne VIC 3010
n.watson@unimelb.edu.au

Abstract

The Australian Bureau of Statistics (ABS) is creating a longitudinal sample, called the Australian Censes Longitudinal Dataset (ACLD), by linking person records across its 5-yearly Census of Population and Housing. This paper proposes a Multi-Panel framework for selecting and weighting records in the ACLD. This framework can be applied more generally to selecting longitudinal samples from a series of cross-sectional administrative files. The proposed

framework avoids some significant limitations of the popular “Top-up” sampling approach to maintaining the cross-sectional and longitudinal representativeness of a sample over time.

.

Key words: survey weighting, sample selection, longitudinal surveys

1) Introduction

The Australian Bureau of Statistics (ABS) selected a 5% random sample of person records from the 2006 Australian Census of Population and Housing (CPH). This sample of records was linked to person records from the 2011 CPH. Since a unique person identifier was not available, a range of linking variables (e.g. date of birth, country of birth, 2006 Mesh Block) were used to match records, where a match is a pair of records that belong to the same person (ABS, 2013). This longitudinal sample of records is referred to as the Australian Census Longitudinal Dataset (ACLD). Analysts can create frequency tables from the ACLD (e.g. transitions in employment status between 2006 and 2011) by accessing the ABS' Table Builder product via its website. A microdata file is available through the ABS network of on-site Data Laboratories in its capital city offices throughout Australia. The intent is to continue linking census records into the future to support longitudinal analysis over different time periods

Since the linkage did not use name and address it is reasonable to expect the presence of linking errors (Felligi and Sunter, 1969). Linkage errors include *incorrect links* and *missed matches*. An incorrect link is a link between two census records that is not a match. Making inference in the presence of incorrect links is a problem that has been considered by Chipperfield & Chambers (2015) and Chipperfield et al (2012). This paper assumes that incorrect links are rare and can be ignored. (Note: ABS (2013) estimates that between 5-10% of links between the 2006 and 2011 Census records are incorrect. The percentage of records with at least one incorrect link will likely increase as we link more censuses.) Given this assumption, from here onwards we note that a link is always a match. A *missed match* occurs if a match is not linked. It is likely that significant

numbers of matches will continue to be missed into the future making it a key issue for the long-term sustainability of the ACLD. (Using anonymised name codes for linking the 2016 CPH with subsequent CPHs is expected to reduce the prevalence of missed matches and the accumulation of incorrect links.) Missed matches can be treated in the same way as non-response in longitudinal surveys, as discussed below.

This paper presents a framework for the sample selection and weighting of the ACLD in the presence of missed matches. More generally, this framework can be used to maintain the longitudinal and cross-sectional representativeness of a sample by selecting records from a series of cross-sectional administrative files. A *representative* sample is a sample that can be treated as a random selection of records from the relevant population. This framework has some similarities with the literature on longitudinal surveys and with the creation of longitudinal samples from administrative files by other National Statistical Offices, but there are some important differences as we now mention.

The United Kingdom (UK) has a well-established framework for undertaking longitudinal studies of its population. These studies are based around a health service data file that contains a record for each person accessing free healthcare in the UK. Because it is continually updated, the health service data provides an up-to-date list that approximates the usual residents of the UK at any point of time. The health services data is used as a population frame from which samples are selected. In the case of a longitudinal study, a random sample of health service records that were active on the 2001 and 2011 Census Nights could be linked to 2001 and 2011 Census records. For many people, up-to-date name and address information can be used to link records (names

were electronically captured by the UK Census in 2001 and 2011 and anonymised for linkage purposes) with very low rates of linkage errors.

Unfortunately, the CPH cannot be used in the same way as the UK health services data to select longitudinal samples. The main reason for this is that the CPH only provides a list of people at a single point in time (Census night) that is not updated thereafter. The CPH, which is conducted every 5 years, is essentially a series of cross-sectional data bases.

In more typical longitudinal surveys (panel survey, rotating panel survey, or cohort study) a sample of people from a cross-sectional population is selected and followed during the life of the survey. Non-response is a key and well-explored issue (Lynn, 2009) and arises because of failure to track respondents, inability to participate (e.g. due to illness or being away from home), and refusal. There are two generic ways to deal with non-response. One is to fit longitudinal models to explicitly account for non-response as discussed in Eideh & Nathan (2009). However, many analysts prefer a weighting approach (see Särndal and Lundström, 2005), in which non-response adjustments are explicitly factored into the weights. This paper focuses on the latter.

Weighting of longitudinal samples needs to consider the range of cross-sectional and longitudinal populations of interest to analysts. For example, the Household, Income and Labour Dynamics in Australia (HILDA) Survey provides cross-sectional weights for each time point, longitudinal weights for a balanced panel from time point t to $t+T$ (for any t and T), and longitudinal weights for a balanced panel between pairs of time points (see Watson, 2012). The HILDA Survey's cross-sectional weights for time point t and longitudinal weights from time

point *t* are calibrated to the cross-sectional population totals at time point *t*. This makes the "static population" assumption (Smith, Lynn, and Elliot, 2009), as per most cohort studies, which is that the longitudinal and cross-sectional populations are the same at time *t* and that deaths and emigration occur in this population over time to *t+T*. Benchmarks for the intersection of the cross-sectional populations from *t* to *t+T* are not available, however analysts make inference about this population if they restrict their analysis to individuals who respond to all time points from time point *t* to *t+T* and apply the balanced panel weight for *t* to *t+T* constructed using the "static population" assumption.

In what follows, sections 2, 3 and 4 review the 2006-11 ACLD's *first release* longitudinal weighting method and propose a longitudinal weighting method for the *second release* of the 2006-11 ACLD. In particular, Section 2 considers estimating the 2006-11 longitudinal population counts. Section 3 discusses the specifics of the first and second release weighting methods, both of which involve calibration to the 2006-11 longitudinal population (and thereby avoid the need to make the "static population" assumption). Section 4 compares estimates based on the two weighting methods. Sections 5 and 6 then consider the long term future of the ACLD. Section 5 proposes a Multi-Panel framework for selecting Census records for inclusion in the ACLD. This framework avoids significant problems associated with "Top-Up" sampling. Section 5 also explains how this framework can be used to select longitudinal samples from a series of cross-sectional administrative files. Section 6 discusses a weighting method under the Multi-Panel framework. Section 7 makes some concluding remarks.

2) Estimating Longitudinal Population Counts in the (2006-11 ACLD)

Let $U_{(06)}$ and $U_{(11)}$ be the set of usual residents of Australia on Census Night in 2006 and 2011, respectively. The count of people in $U_{(06)}$ and $U_{(11)}$, denoted by $N_{(06)}$ and $N_{(11)}$ respectively, is assumed to be known and is calculated after correcting for counting errors in the CPH. Further, let $U_{(06,11)}$ be the set of usual residents of Australia on both the 2006 and 2011 Census Nights. Let be $N_{(06,11)}$ be the count of people in $U_{(06,11)}$. In this section we consider two ways of estimating $N_{(06,11)}$.

We can estimate $N_{(06,11)}$ by

$\widehat{N}_{(06,11)} = N_{(11)}$ x $P_{(06|11)}$,

where $P_{(06|11)}$ is the proportion of people in the 2011 usual resident population who were also usual residents on the 2006 Census Night. We know from ABS demographic estimates (i.e., the Estimated Residential Population, which is the usual resident count from the Census adjusted for the Census over- and under-coverage) that the number of people aged 5 and older in 2011, $N_{(11)}$, is 20.8 million. The issue is now how to estimate $P_{(06|11)}$ and here we present two approaches.

The Net Overseas Migration (NOM)-based estimate of $N_{(06,11)}$ is

$$\widehat{N}^{(NOM)}_{(06,11)} = N_{(11)} \text{ x} \widehat{P}^{(NOM)}_{(06|11)},$$

where $\hat{P}^{(NOM)}_{(06|11)} = (1 - M_{(06,11)}/N_{(11)})$ and $M_{(06,11)}$ is the NOM between 2006 and 2011, which is calculated as the sum of NOM (the difference between total emigrants and total immigrants in

a given year) for each of the five years. The disadvantages of $\widehat{N}_{(06,11)}^{(NOM)}$ as an estimate of $N_{(06,11)}$ are that it would

i. only be available at the broad levels at which NOM are available (i.e. age, sex, and geography).

ii. incorrectly include:

a. People who were overseas arrivals after the 2006 Census Night and who subsequently left Australia (or died) before the 2011 Census Night;

b. People who left Australia after the 2006 Census Night and then returned to Australia before the 2011 Census Night.

If 10% (20%) of all overseas arrivals between 2006 and 2011 fell into category a., without being balanced by b., $\widehat{N}_{(06-11)}^{(NOM)}$ would be too high by 230,000 (460,000) people.

The Census-based estimate of $N_{(06,11)}$ can be expressed as

$$\widehat{N}_{(06,11)}^{(CEN)} = N_{(11)} \text{ x} \widehat{P}_{(06|11)}^{(CEN)},$$

where $\widehat{P}_{(06|11)}^{(CEN)}$ is the proportion of respondents to the 2011 CPH who, based on the Address 5 Year Ago question, were also usual residents on 2006 Census Night. The estimate of $P_{(06|11)}^{(CEN)}$ was calculated in the following way:

1. Calibrate the initial weights (equal to one) of respondents to the 2011 CPH to $N_{(11)}$, by State, Age, Sex, and Aboriginal and Torres Strait Islanders Status. A respondent is a record with a valid value for State, Age, Sex, Aboriginal and Torres Strait Islanders Status

and Address 5 Years Ago. This 'weighted' 2011 CPH file can be thought of as a hypothetical 2011 CPH with no counting errors.

2. Using the weighted 2011 census file (Step 1), $\hat{P}_{(06|11)}^{(CEN)}$ is the proportion of respondents who, based on their response to the Address 5 Years Ago question, were in scope of the 2006 CPH.

Both $\hat{P}_{(06|11)}^{(CEN)}$ and $\hat{N}_{(06,11)}^{(CEN)}$ can be calculated for any sub-population (e.g. by Age group) defined in terms of 2011 CPH variables.

The Census-based estimator does not suffer from the disadvantages (see i. and ii.) of the NOM-based estimator of $N_{(06,11)}$ and so, on this basis, it is the preferred estimator. Next we make an empirical comparison between the two population estimators.

At the Australia level, $\hat{N}_{(06,11)}^{(CEN)}$ =19.5 million and $\hat{N}_{(06,11)}^{(NOM)} = 18.6$ million- a difference of 1.1 million people. The source of this difference is of course driven by the difference between $\hat{P}_{(06|11)}^{(CEN)}$ and $\hat{P}_{(06|11)}^{(NOM)}$. Table 1 shows that, at the national level, $\hat{P}_{(06|11)}^{(CEN)}$ and $\hat{P}_{(06|11)}^{(NOM)}$ are 93.5% and 89.5% respectively. The most noticeable difference occurs for 25-34 year olds (11.4 percentage points) and for 15-24 year olds (6.4 percentage points). Both $\hat{P}_{(06|11)}^{(CEN)}$ and $\hat{P}_{(06|11)}^{(NOM)}$ indicate that people in the 15-24 and 25-34 age groups in 2011 have the lowest rate of belonging to the 2006-2011 longitudinal population. Conversely, people in the over 65 age group in 2011 have the highest rate of belonging to the 2006-2011 longitudinal population (i.e. 99.3% for $\hat{P}_{(06|11)}^{(CEN)}$ and 98.4% for $\hat{P}_{(06|11)}^{(NOM)}$) .

Given NOM is not available for Aboriginal and Torres Strait Islanders people, it was simply assumed to be zero (i.e. $\hat{P}_{(06-11)}^{(NOM)} = 1$) leading to an NOM-based estimate of 586,000. For the Aboriginal and Torres Strait Islanders' population, $\hat{P}_{(06|11)}^{(CEN)}$=99.6% giving $\hat{N}_{(06,11)}^{(CEN)}$ = 584,000. Across all states and territories, the differences between the Census and NOM-based estimates of the Aboriginal and Torres Strait Islanders population were within 1%.

TABLE 1 ABOUT HERE

As we lengthen the time period associated with a longitudinal population, it will become less like any cross-sectional population. The age of the youngest person in a longitudinal population will increase (e.g. the youngest person in the 2006-11-16 longitudinal population will be 10 years of age). From Table 1 we also see that, compared with the composition of any cross-sectional population, a longitudinal population will have significantly fewer 15-34 year olds and significantly more over 65s. For example, if the values of $\hat{P}_{(06|11)}^{(CEN)}$ in Table 1 are constant into the future, then 96% (= 0.987 x 0.987 x 0.993 x0.993) of 84 year olds in 2026 would belong to the 2006-2026 longitudinal population; in contrast, only 62% (=0.851 x 0.851 x 0.922 x 0.922) of 44 year olds in 2026 would belong to the 2006-2026 longitudinal population. This seems to provide strong evidence that the composition of the 2006 population is different to the composition of the 2006-11 longitudinal population (i.e. $N_{(06)} = N_{(06,11)}$).

3) Weighting the Longitudinal Sample (2006-11 ACLD)

As mentioned, a 5% cross-sectional sample, $s_{(06)}$, of $n_{(06)}$ records was selected from the 2006 CPH and then linked to all 2011 CPH records. Denote the resulting set of 2006-2011 linked records by $s_{(06,11)}$ and the number of linked records by $n_{(06,11)}$. This section describes the first and second release methods of weighting records in $s_{(06,11)}$ to represent the population $U_{(06,11)}$. One point of difference between the two methods is that the first release weights uses NOM-based benchmarks while the second release weight uses the Census-based benchmarks (see Section 2). Another point of difference is that, when making an adjustment for missed matches, the second release weight does not make the implicit assumption that the 2006-2011 population is the same as the 2006 population.

We observe that $n_{(06,11)} < n_{(06)}$ because of:

a. deaths between 2006 and 2011
b. emigration between 2006 and 2011
c. 2006 CPH records that could have been correctly linked but were not linked at all (i.e. missed matches).

Records in a. and b. are not in scope of the longitudinal usual resident population. However, records in c. are in scope and so would ideally have been linked. Records in c. can be treated as non-response. Therefore a bias in the sample would arise if the rate of missed matches was relatively high in some subpopulations. For example, children may have a higher rate of missed matches because some variables that are useful for linking adults (e.g. educational attainment, occupation and marital status) are not useful for linking children. It is therefore important for the weighting method to address the problem of missed matches.

Let $\widehat{\boldsymbol{N}}_{(06,11)} = (\widehat{N}_{1,(06,11)}, \ldots \widehat{N}_{h,(06,11)}, \ldots, \widehat{N}_{H,(06,11)})$ be a vector of sub-population counts, where $\widehat{N}_{h,(06,11)}$ is the estimated number of people in the *h*th sub-population who belong to $U_{(06,11)}$. The estimate, $\widehat{N}_{h,(06,11)}$, may be the Census or NOM-based estimator (see Section 2). Let $\boldsymbol{z}_{i(11)} = (\boldsymbol{z}_{i1,(11)}, \ldots \boldsymbol{z}_{ih,(11)}, \ldots, \boldsymbol{z}_{iH,(11)})$ denote a set of covariates on the linked file, where $\boldsymbol{z}_{ih,(11)} = 1$ if the *i*th record in $s_{(06,11)}$ belongs to sub-population h in 2011 and is zero otherwise.

We are interested in an estimator of $Y_{(06,11)} = \sum_{i \in U_{(06,11)}} y_i$ that is given by

$$\hat{Y}_{(06,11)} = \sum_{i \in s_{(06,11)}} y_i \widetilde{w}_{i,(06,11)},$$

where the longitudinal weight for the *i*th linked record in $s_{(06,11)}$ can be expressed by

$$\widetilde{w}_{i,(06,11)} = w_{(06)} \times l_{i,(06,11)}^{-1} \times g_{i(06,11)}, \qquad (1)$$

the initial weight is $w_{(06)} = 20 = 1/(5\%)$, $l_{i,(06,11)}$ is the probability that the *i*th record was linked between 2006 and 2011 and is a function of covariates $\boldsymbol{x}_i$, and

$$g_{i,(06,11)} = 1 + \left(\widehat{\boldsymbol{N}}_{(06,11)} - \sum_{i \in s_{(06,11)}} \tilde{q}_{i,(06,11)} \boldsymbol{z}_{i(11)}\right) \left(\sum_{i \in s_{(06,11)}} \tilde{q}_{i,(06,11)} \boldsymbol{z}'_{i(11)} \boldsymbol{z}_{i(11)}\right)^{-1},$$

where $\tilde{q}_{i,(06,11)} = w_{(06)} \times l_{i,(06,11)}^{-1}$. The approach taken by (1) is discussed in Särndal & Lundström (2005) in the traditional survey sampling context. In the sampling context, the motivation for (1) is to reduce the error due to sampling from the population and to reduce

potential bias from non-response, as noted previously. The motivation for (1) is the same here, where 'non-response' arises due to missed matches.

The purpose of $l_{i,(06,11)}(\boldsymbol{x}_i)$ is to correct for over and under-representation of the linked sample due to missed matches, where $\boldsymbol{x}$ are binary census covariates for Sex, Age Group, Country of Birth, English Proficiency, Language spoken, Religion, Occupation, Marital Status, Qualification, Degree of Remoteness and whether or not the person was an inter-state migrant between 2006 and 2011. In doing so, we allow the probability of a missed match to vary across subpopulations. The last of these variables was included in case inter-state migrants were more likely to be missed, despite the fact that the Address 5 Year Ago question was used in linking. For details of exactly how the implicit assumption made about missed matches depends upon $\boldsymbol{x}$ we refer the reader to Haziza and Lesage (2016). (We remind the reader that, as mentioned in Section 1, we continue to assume that incorrect links are rare and so can be ignored.)

The purpose of $g_{i(06,11)}$ in (1) is to correct for the 2006 and 2011 Census over- or under-coverage of the usual resident population. The term $g_{i(06,11)}$ adjusts the sample of Census records, which clearly reflects any Census over- or under-coverage, to align with the population totals $\widehat{\boldsymbol{N}}_{(06,11)}$, which are net of Census over- or under-coverage errors. The covariates $\boldsymbol{z}_{(11)}$ include binary variables for Age Groups (10 year ranges), Sex, State, Aboriginal and Torres Strait Islanders Status and whether or not the person was an inter-state migrant between 2006 and 2011.

One difference between the first and second release weights is that the counts $\widehat{\boldsymbol{N}}_{(06,11)}$ in (1) were estimated by the NOM and Census approach, respectively (see Section 2). Another difference in the weights is due to how $l_{i,(06,11)}$ was defined. For example, in the first release, $l_{i,(06,11)}$ was the probability that a 2006 record was linked to a 2011 record whereas in the second release it was the probability that a 2011 record was linked to a 2006 record. As we discuss in detail below, the latter was preferred because it could be estimated more precisely with the information collected by the Census.

In the case of the first release weight, $l_{i,(06,11)}$ was specified to be the probability that the *i*th record $s_{(06)}$ was linked. Accordingly $l_{i,(06,11)}$ was predicted by fitting a logistic regression model using $s_{(06)}$, where the outcome variable took the value 1 if the record was linked and 0 otherwise, and the covariates $\boldsymbol{x}$ were derived from the 2006 CPH. This adjustment for missed matches would be correct under the assumption that the population between 2006 and 2011 did not change, as it assumes a match exists for all records in $s_{(06)}$ or, equivalently, that there were no deaths or migration since 2006 (see cases a. and b. above). This assumption, as discussed in Section 2, is hard to justify.

To motivate the estimation of $l_{i,(06,11)}$ for the second release weighting, define the Match Rate as the number of links divided by the expected total number of matches in the sample. An estimate of the Match Rate for the 06-11 ACLD in sub-population *h* of the 2006-11 population is R_h=20$m_h/\widehat{N}^{(CEN)}_{(06,11)h}$, where m_h is the number of linked records in sub-population *h* and $\widehat{N}^{(CEN)}_{(06,11)h}$ (see Section 2 for details) is the estimated count of people in the 2006-11 longitudinal population

who belong to sub-population *h*, and '20' is the inverse of the sample fraction for $s_{(06)}$. Next we estimate the Match Rate across various sub-populations.

For records with Age ≥ 15 in 2006 the overall Match Rate was 84%. Across a range of sub-populations, the lowest Match Rate was 58% for Aboriginal and Torres Strait Islanders people. The sub-population with the highest Match Rate of 96% was for people with the same 2006 and 2011 Census address. In contrast, people with different 2006 and 2011 Census addresses had a Match Rate of 66%.

For records with Age < 15 in 2006 the overall Match Rate was 88%. Across a range of sub-populations, the lowest Match Rate was 26% for Birth Place =" Southern and Central Asia". The sub-population with the highest Match Rate was 94% for people with Country of Birth= "Oceania and Antarctica (Excluding Aboriginal and Torres Strait Islanders people"). Again, the sub-population with the highest Match Rate of 98% was for people with the same 2006 and 2011 Census address compared with 74% for people with different 2006 and 2011 Census addresses. Table 2A and 2B give match rates for select sub-populations.

TABLE 2 ABOUT HERE

Since the Match Rate varies widely across sub-populations, there is strong evidence that weighting will improve the representativeness of these sub-populations on the linked file. In the second release weighting method, $l_{i,(06,11)}$ is the probability that the *i*th record would be linked given it belonged to the 2006-11 longitudinal population (i.e. given it is a match). This

probability was predicted using a logistic model fitted to the counts $\{(m_{(11)k}, \widehat{N}^{(CEN)}_{(06,11)k}, \boldsymbol{x}_{(11)k})$: k=1,...K) where k indexes the covariate patterns, $\boldsymbol{x}_{(11)k}$ is the kth covariate pattern in $\boldsymbol{x}$ on the 2011 CPH, $m_{(11)k}$ is the number of linked records with covariate pattern k, and $\widehat{N}^{(CEN)}_{(06,11)k}$ is calculated using the Census estimator in Section 3. In this way, the second release adjustment for missed matches does not make the assumption that the 2006 population is the same as the 2006-2011 population (unlike the first release weight adjustment).

Next we discuss some of the results of fitting the logistic models to predict $l_{i,(06,11)}$ for the first and second release weights. Across all coefficients, the odds for the first release model range from 0.09 for Remoteness =”Missing” to 1.94 for Post School Qualification =”Post Graduate”; in contrast, the odds for the second release coefficients range from 0.73 for Remoteness=”Missing” to 1.2 for Marital Status=”N/A”. The fact that the odds for the first and second release model are very different shows that assuming the composition of 2006 population and the 2006-2011 populations are the same has a significant impact on the ACLD weights.

To calculate the sample variance of $\hat{Y}_{(06,11)}$, we may use the Group Jacknife variance estimator. This estimator requires that each record in $s_{(06,11)}$ is randomly allocated to one of G replicate groups such that each replicate group contains the same (or approximately the same) number of records. (Here we use G=30.) If we index the replicate groups by g = 1,...,G, the Jacknife variance estimator is

$$\mathrm{V_J}\left(\hat{Y}_{(06,11)}\right) = G/(G-1)\sum_{\mathrm{g}}\left(\hat{Y}_{(06,11)}(g) - \hat{Y}_{(06,11)}\right)^2,$$

where $\hat{Y}_{(06,11)}(g)$ is calculated in the same way as $\hat{Y}_{(06,11)}$ except that it is based on $s_{(06,11)}$ after excluding records in the *g*th group. The Jacknife is suitable here since it is unbiased for the variance of a function of sample means, where the sample is selected by simple random sampling (see Shao and Tu, 1995). Even though $\widehat{\boldsymbol{N}}_{(06,11)}$ is an estimate (and therefore has a "hat") it is treated as fixed in variance estimation. In general this assumption will lead to a slight under-estimate of the variance. The degree of under-estimation will be only slight because $\widehat{\boldsymbol{N}}_{(06,11)}$ is based on about 20 million Census records- at least 20 times more than the number of sampled records in $s_{(06,11)}$ that are used to calculate $\hat{Y}_{(06,11)}$.

4) Comparing the Outputs from the Two Longitudinal Weighting Approaches

To compare the difference between the first release and second release weighting approaches, consider two examples of transitions between 2006 and 2011. Table 3 shows that for self-reported Aboriginal and Torres Strait Islanders status there is very little difference between estimates based on the second and first release weights. There is a slight decrease in the marginal proportion reporting Aboriginal and Torres Strait Islanders Status in 2011 under the second release weights. This is likely due to the fact that the estimate of the number of non-Indigenous people in the 2006-11 longitudinal population is larger (by about 1.1 million) under the second release weights.

Table 4 shows that the estimates of marital status transition probabilities using the first and second release weights are reasonably similar. An exception is for the probability of being *Never Married* in 2011 given being *N/A* in 2006, where the estimate for the first and second release

weights are 32.4% and 35.6%, respectively. This difference makes sense since this estimate is made up of people aged between 15-20 years in 2011, a sub-population that is assigned more weight under the second release weights (see Table 1).

TABLES 3 AND 4 ABOUT HERE

5) Framework for ACLD Sample Selection from Future Censuses

We now turn to the issue of selecting records from future CPHs so as to maintain longitudinal and cross-sectional representatives of the ACLD into the future. We discuss two options: "Top-Up" and "Multi-Panel". These two competing (conceptual) frameworks will be assessed against the following:

a. Will it give a representative cross-sectional sample of every census?
b. Will it give a representative longitudinal sample across censuses?
c. Will it be sustainable? This means that any reduction in the cross-sectional and longitudinal representativeness of the sample over time due to missed matches is minimised.
d. Will it be conceptually straight-forward?
e. Will it be practical for the ABS to implement?

The Multi-Panel framework selects a 5% sample of records from each CPH (i.e. the 2006 CPH, the 2011 CPH and so on). Each of these samples is selected according to the same "Rule". Subject to CPH non-response, over- or under-coverage, scope rules, and reporting errors, this Rule will select the same set of people from each CPH (i.e. if a person is selected in the 2006

Panel we would expect that they would also be selected in the 2011 Panel, assuming they remain in scope of the CPH). To explain, applying the selection Rule involves two steps:

(a) constructing a person ‘variable’ on the CPH that is not likely to change over time;

(b) randomly selecting a sample of values that the ‘variable' may take; records with one of these selected values are in turn selected.

Options for the construction of this ‘variable’ are discussed later in this subsection for the case of sampling from an administrative file. (Note it is currently standard practice for the ABS not to publically release how it selects Census records for the ACLD).

Each 5% sample of CPH records is the beginning of a new panel of the ACLD. For example, the 5% sample of 2006 CPH records is the beginning of the 2006 Panel of the ACLD- it is created by linking the 2006 sample of CPH records to all 2011 CPH records; the resulting 2006-2011 linked file would then be linked to all 2016 Census records, and so on. Similarly, a 5% sample of 2011 Census records would be the beginning of the 2011 Panel of the ACLD. The Multi-Panel Framework is illustrated in Figure 1.

In answer to the respective questions mentioned above:

a. The records selected in the 2006 Panel of the ACLD are a representative sample of the 2006 CPH cross-section. The same comment applies to future panels (e.g. 2011). This means analysis of any cross-section of the CPH will be unbiased.

b. If there were no missed matches, the 2006 Panel of the ACLD would be a representative sample of the longitudinal population from 2006 to any future time. The same comment applies to future panels (e.g. 2011 and 2016 panels).
c. Perhaps the most important feature of the Multi-Panel approach is that it limits the accumulation of bias due to linkage error (e.g. missed matches). For example the 2016 Panel is not biased by missed matches between 2006, 2011 and 2016 CPH records.
d. It is straight-forward for an analyst to choose which panel of the ACLD best meets their needs. For example, if interest is in transitions between 2011 and 2016 or between 2011 and 2021, the 2011 Panel is the most appropriate panel.
e. Each panel of the ACLD would be linked to the latest CPH. As illustrated in Figure 1, the 2021 CPH would have to be linked to the 2006, 2011 and 2016 Panels. Since, due to selection Rule, records in the 2006 and 2011 panels must also be in the 2016 Panel, many of the links in the different panels would also be in common. This makes the linkage exercise manageable. In fact, operationally the 2021 CHP is just linked to the 2016 cross-section as the other ongoing panels are contained within it.

FIGURE 1 ABOUT HERE

We now consider the "Top-Up" Framework, as used by the HILDA Survey. The "Top-up" framework for the ACLD would involve:

1. Selecting a 5% sample of 2006 CPH records.
2. Linking the 2006 sample of CPH records to all 2011 CPH records resulting in a 2006-2011 linked file.

3. From each successive CPH, starting with the 2011 CPH, selecting a "top-up" sample in order to maintain the cross-sectional and longitudinal representativeness of the sample.
4. Linking this combined sample to the 2016 CHP records, and so on.

Under the Top-Up framework, the ACLD at 2016 would consist of records selected in the 5% sample of 2006 CPH records and records selected in the 2011 and 2016 CPH top-up samples. The aim of the 2016 top-up sample would be to ensure that the ACLD is representative of the CPH 2016 cross-section. In order to do this, the 2016 top-up sample would need to be a representative sample of people in the 2016 CPH who are born after the 2011 Census Night or who arrived in Australia after the 2011 Census Night but before the 2016 Census Night. However, since we cannot accurately identify people in this second group in order to sample from them (we only know from the Census when an individual *first* arrived in Australia), top-up sampling will likely lead to sample bias that will accumulate over time. For example, if an analyst is interested in transitions between 2016 and 2021 any accumulation of this sample bias arising from the 2011 and 2016 top-up samples will be present in the 2016 cross-sectional sample.

While it is likely that the Multi-Panel framework requires more resources to link records, selecting top-up samples would be technically complex, require a substantial amount of time to develop and justify, and is unlikely to be effective in maintaining a representative longitudinal or cross sectional sample over time. The over-riding benefit of the Multi-Panel framework over the Top-Up framework is that, by constantly taking a cross-sectional sample at the time of each

CPH, the need for top-up samples and its associated problems are avoided. For these reasons the ABS is planning to implement the Multi-Panel approach for the ACLD.

While we have explained Multi-Panel framework in the context of the ACLD, it also applies generally to longitudinal surveys whose sample is selected from an administrative file, as we now explain. Under the Multi-Panel framework, a typical longitudinal study (e.g. the Longitudinal Survey of Australian Children) of a certain population would apply a Rule to select, at each time t =1, .., T, a sample of records from an administrative file (e.g. Medicare) that covers the cross-sectional population. The person 'variable' (discussed above) could be constructed from certain combinations of: digits from a unique person identifier; day and month of birth; or letters in name. Thus each cross-sectional sample at time t would be representative of the cross-sectional population at time t; and if a person is selected they would also be selected at each time point that they remain on the administrative file. Each cross-sectional sample at time t will contain two groups of people: (1) people who were selected in the survey prior to time t and who could be tracked over time in the usual way for a longitudinal sample (e.g. via mobile phone numbers or via contact details on the administrative file); (2) people who were not selected prior to time t and who are contacted for the first time via their contact details on the administrative file. This, of course, assumes that the person 'variable' (for example birth day and month in the UK Cohort studies) is stable over time (i.e., free from real changes or errors) and that duplicate records for an individual are minimal.

6) Cross-sectional and Longitudinal Weighting under the Multi-Panel Framework

We now consider the general problem of weighting under the Multi-Panel framework discussed in Section 5. It is an extension of the estimator in Section 3, with weights given by (1), to multiple time points and multiple panels.

Panel *t* of the ACLD can be considered to be a single file made up of all records in the set $s_{(t)}$. Each record $i \in s_{(t)}$ would have:

- a cross-sectional weight, denoted by $\widetilde{w}_{i,(t)}$ for use in making inference about the population $U_{(t)}$;
- a set of longitudinal weights $\widetilde{w}_{i,(t,t+1),}\widetilde{w}_{i,(t,t+1,t+2)}, \ldots \widetilde{w}_{i,(t,t+1,\ldots,t+T)}$ for making inference about the populations $U_{(t,t+1)}, U_{(t,t+1,t+2)}, \ldots, U_{(t,t+1,\ldots,t+T)}$, respectively, where consistent with earlier notation, $U_{(t,t+1\ldots,t+T)} = U_{(t)} \cap U_{(t+1)} \ldots \cap U_{(T+t)}$ is the set of usual residents from time *t* to *t+T*. It makes sense to set $\widetilde{w}_{i,(t,t+1,\ldots,t+T)} = 0$ if record $i \notin s_{(t,t+1,\ldots\, t+T)}$ (i.e. if a record was selected in $s_{(t)}$ but was not linked at all the time points from time *t* to time *t+T*). This is equivalent to an analysis of completers in a standard longitudinal survey.

The various weights are summarised in Table 5. The choice of weight is straight-forward. If there is interest in transitions between 2006 (*t*=1) and 2016 (*t*=3) then the population of interest is $U_{(1,2,3)}$ and the appropriate weight is $\widetilde{w}_{i,(1,2,3)}$. The remainder of this section defines the cross-sectional and longitudinal weights introduced above.

TABLE 5 ABOUT HERE

6.1 Cross Sectional Weights

Consistent with earlier notation, let $N_{(t)}$ be the count of usual residents of Australia on Census Night at time *t* and denote the set of records belonging to this population by $U_{(t)}$. A 5% cross-sectional sample, $s_{(t)}$, of size $n_{(t)}$ will be selected from $U_{(t)}$ at each time point *t*. This set $s_{(t)}$ contains records that begin panel *t*. Here we describes the cross-sectional weight for records in $s_{(t)}$ that are designed to make inference about $U_{(t)}$.

The initial cross-sectional weight for records in $s_{(t)}$ is $w_{(t)} = 20$. This initial weight is calibrated to known cross-sectional counts $\boldsymbol{N}_{(t)} = (N_{1,(t)}, \ldots N_{h,(t)}, \ldots, N_{H,(t)})$ for *H* sub-populations at time *t*. For records in $s_{(t)}$ define $\boldsymbol{z}_{i(t)} = (\boldsymbol{z}_{i1,(t)}, \ldots \boldsymbol{z}_{ih,(t)}, \ldots, \boldsymbol{z}_{iH,(t)})'$, where $\boldsymbol{z}_{ih,(t)} = 1$ if the *i*th record belongs to sub-population *h* at time *t* and is zero otherwise. The cross-sectional weight for the *i*th record in $s_{(t)}$ is

$$\widetilde{w}_{i,(t)} = w_{(t)} \times g_{i,(t)},$$

where $g_{i,(t)} = 1 + \left(\boldsymbol{N}_{(t)} - \widehat{\boldsymbol{N}}_{(t)}\right)\left(\sum_{i \in s_{(t)}} \boldsymbol{z}'_{it}\boldsymbol{z}_{it}\right)^{-1}$ is principally designed to correct for counting errors in the CPH at time *t* , and $\widehat{\boldsymbol{N}}_{(t)} = \sum_{i \in s_{(t)}} w_{(t)}\boldsymbol{z}_{i(t)}$ is an estimate of $\boldsymbol{N}_{(t)}$. (This correction occurs because in reality $\boldsymbol{N}_{(t)}$ has been adjusted for census coverage errors using the post-enumeration survey.) The estimator of $Y_{(t)} = \sum_{i \in U_{(t)}} y_{it}$, is $\hat{Y}_{(t)} = \sum_{i \in s_{(t)}} y_{it}\widetilde{w}_{i,(t)}$ and its variance can be estimated using the Group Jacknife, as discussed earlier.

6.2 Estimating the Longitudinal Population Size

Consistent with earlier notation let $N_{(t,t+1\ldots,t+T)}$ be the number of people in the set $U_{(t,t+1\ldots,t+T)}$. Similarly, let $U_{h,(t,t+1\ldots,t+T)} = U_{h,(t)} \cap U_{h,(t+1)} \ldots \cap U_{h,(T+t)}$ be the set of usual residents in sub-population *h* from time *t* to *t*+*T* and let $N_{h,(t,t+1\ldots,t+T)}$ be the number of people in the set $U_{h,(t,t+1\ldots,t+T)}$. Here we consider the estimator

$$\widehat{\boldsymbol{N}}_{(t,t+1\ldots,t+T)} = (\widehat{N}_{1,(t,t+1\ldots,t+T)}, \ldots \widehat{N}_{h,(t,t+1\ldots,t+T)}, \ldots, \widehat{N}_{H,(t,t+1\ldots,t+T)}),$$

where $\widehat{N}_{h,(t,t+1\ldots,t+T)}$ is the estimate of $N_{h,(t,t+1\ldots,t+T)}$.

The probability that record *i* is a usual resident at time *t*, given that they were a usual resident at time *t* +1 and belongs to sub-population *h*, can be estimated by

$$\hat{P}_{h,(t|t+1)} = \frac{\widehat{N}_{h,(t,t+1)}}{N_{h,(t+1)}}$$

Note $P_{h,(06|11)}$ was estimated in Section 2 and given in Table 1 in the case of *t*=2006, *t*+1=2011, and *h* denoting Age Group). An estimate of $\widehat{N}_{h,(t,t+1\ldots,t+T)}$ is then

$$\widehat{N}_{h,(t,t+1\ldots,t+T)} = N_{h,(t+T)} \prod_{r=1}^{T} \times \hat{P}_{h,(t+r-1|t+r)}.$$

This estimator assumes that whether or not a person is a usual resident at time *t* only depends upon sub-population *h* at time *t*+1.

6.3 Weighting to the Longitudinal Population

As mentioned, $s_{(t)}$ contains the records that are selected at time *t* (referred to as Panel *t*). Records belonging to Panel *t* after *T* time points consists of records in $s_{(t)}$ that are linked to the

CPH at times t+1, t+2, …, and t+T. Denote this set of records by $s_{(t,t+1,\ldots,t+T)}$ and denote the number of records in $s_{(t,t+1,\ldots,t+T)}$ by $n_{(t,t+1,\ldots,t+T)}$. Next we consider the longitudinal weight for records in $s_{(t,t+1,\ldots,t+T)}$ that are designed to make inference about $U_{(t,t+1\ldots,t+T)}$. This is a generalisation of (1).

The weight for the ith record in $s_{(t,t+1)}$ is given by $\widetilde{w}_{i,(t,t+1)} = \widetilde{w}_{i,(t)} \times l_{i,(t,t+1)}^{-1} \times g_{i,(t,t+1)}$ and is described by (1) for the case where t=2006. For $T \geq 2$ the weight for the ith record in $s_{(t,t+1\ldots,t+T)}$ is given by

$$\widetilde{w}_{i,(t,t+1\ldots,t+T)} = \widetilde{w}_{i,(t,t+1\ldots,t+T-1)} \times l_{i,(t+T-1,t+T)}^{-1} \times g_{i,(t,t+1\ldots,t+T)} \qquad (2)$$

$$= \widetilde{w}_{i,(t)} \left(\prod_{r=1}^{T} l_{i,(t+r-1,t+r)}^{-1} \times g_{i,(t,t+1\ldots,t+r)} \right)$$

where $l_{i,(t+r-1,t+r)}$ is the probability that the ith record was linked between time t+r-1 and t+r and is allowed to depend upon covariates $\boldsymbol{x}_{i,(t+r)}$ available at time t+r. The term $g_{i,(t,t+1\ldots,t+r)}$ is designed to account for coverage errors in the CPH at time t+r and is given by

$$g_{i,(t,t+1\ldots,t+r)} = 1 + \left(\widehat{\boldsymbol{N}}_{(t,t+1\ldots,t+r)} - \sum\nolimits_{i \in s_{(t,t+1,\ldots,t+r)}} \tilde{q}_{i,(t,t+1\ldots,t+r)} \boldsymbol{z}_{i(t+r)} \right)$$

$$\left(\sum\nolimits_{i \in s_{(t,t+1,\ldots,t+r)}} \tilde{q}_{i,(t,t+1\ldots,t+r)} \boldsymbol{z}'_{i(t+r)} \boldsymbol{z}_{i(t+r)} \right)^{-1},$$

where $\tilde{q}_{i,(t,t+1\ldots,t+r)} = \widetilde{w}_{i,(t,t+1\ldots,t+r-1)} \times l^{-1}_{i,(t+r-1,t+r)}$ is the weight at time t+r-1 adjusted by the probability that a match is made between time t+r-1 and t+r, and $\widehat{\boldsymbol{N}}_{(t,t+1\ldots,t+r)}$ is estimated from Section 6.2.

This treatment of sequential linkage errors is similar to sequential modelling of response propensities for drop-out in standard longitudinal surveys, where the probability of responding at time t+T is the product of a set of conditional response probabilities at each time point (see for example Veiga, Smith & Brown (2014) following the approach of Lepkowski (1989)). This allows the most flexible use of information for modeling non-response at each time point. However, we do assume that non-response at time t (i.e. missed matches between t and t-1) and non-response at any other time (i.e. missed match between any other two time points) are conditionally independent.

To be clear how we use the longitudinal weights, define the vector of outcomes for a variable y across censuses to be $\boldsymbol{y}_{i,(t,t+1\ldots,t+T)} = (y_{i(t)}, y_{i,(t+1)}, \ldots, y_{i,(t+T)})$ and let $\theta_{i,(t,t+1\ldots,t+T)} = \theta(\boldsymbol{y}_{i,(t,t+1\ldots,t+T)})$ for some function $\theta(.)$. For example, $\theta_{i,(t,t+1\ldots,t+T)}$ may indicate employment status for record i at time t+T given a particular employment history from time t to time t+T-1. The estimator of $\theta_{(t,t+1\ldots,t+T)} = \sum_{i\in U_{(t,\ldots,t+T)}} \theta_{i,(t,t+1\ldots,t+T)}$ is

$$\hat{\theta}_{(t,t+1\ldots,t+T)} = \sum_{i\in s_{(t,\ldots,t+T)}} \theta_{i,(t,t+1\ldots,t+T)} \widetilde{w}_{i,(t,t+1\ldots,t+T)}.$$

The Jackknife variance estimator of $\hat{\theta}_{(t,t+1\ldots,T)}$ given by

$$\mathrm{V_J}\big(\hat{\theta}_{(t,t+1\ldots,t+T)}\big) = \mathrm{G}/(\mathrm{G}-1)\sum_{\mathrm{g}}\big(\hat{\theta}_{(t,t+1\ldots,t+T)}(g) - \hat{\theta}_{(t,t+1\ldots,t+T)}\big)^2,$$

where $\hat{\theta}_{(t,t+1\ldots,t+T)}(g)$ is calculated in the same way as $\hat{\theta}_{(t,t+1\ldots,t+T)}$ except that it is based on the set of records in $s_{(t,t+1,\ldots,t+T)}$ after excluding the gth group (see earlier discussion about Jacknife groups in section 3).

7) Summary and Discussion

The development of the ACLD is an exciting edition to the census outputs produced by the ABS. But across censuses it is important that the representativeness of the linked sample is maintained as well as possible. This paper proposes a Multi-Panel framework for the ACLD. The framework can be applied to select a longitudinal sample of records from cross-sectional administrative files that are available over the life of the longitudinal sample. The framework avoids some of the limitations of the popular "Top-up" sampling approach that aims to maintain both the cross-sectional and longitudinal representativeness of a single sample that is extended over time.

REFERENCES

ABS (2013) Information Paper: Australian Census Longitudinal Dataset, Methodology and Quality Assessment, 2006-2011. ABS Cat.No. 2080.5. Canberra: ABS.

Chipperfield, J. O. and Chambers, R. L. (2015) Using the Bootstrap to Analyse Binary Variables with Probabilistically-Linked Data, *Special Issue of The Journal of Official Statistics* (Accepted)

Chipperfield, J., Bishop, G. R. and Campbell, P (2011), Maximum Likelihood Estimation for Contingency Tables and Logistic Regression with Incorrectly Linked Data, *Survey Methodology*, **37**, No. 1, pp. 13-24.

Eideh, A. and Nathan G. (2009) Joint Treatment of Nonignorable Dropout and Informative Sampling for Longitudinal Survey Data. In Lynn, P. (Ed.), *Methodology of Longitudinal Surveys* (pp. 251-264). Chichester: Wiley.

Fellegi, I.P., and Sunter, A.B. (1969), A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

Haziza, D. and Lesage, E. (2016), A Discussion of Weighting Procedures for Unit Nonresponse. *The Journal of Official Statistics* 32, pp. 129–145,

Lepkowski, J. M. (1989) The treatment of wave nonresponse in panel surveys. In D. Kasprzyk, G. Duncan, G. Kalton and M. Singh (Eds.), *Panel Surveys* (Ch. 5). New York: Wiley.

Lynn, P. (2009) Methods for Longitudinal Surveys. In Lynn, P. (Ed.), *Methodology of Longitudinal Surveys* (pp. 1-19). Chichester: Wiley.

Rendtel, U. and Harms, T. (2009) Weighting and Calibration for Household Panels. In Lynn, P. (Ed.), *Methodology of Longitudinal Surveys* (pp. 265-286). Chichester: Wiley.

Särndal, C. E., Swensson, B. and Wretman J. (1992) *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Särndal, C. E. and Lundström, S. (2005) *Estimation in Surveys with Nonresponse*. Chichester: Wiley.

Shao, J., and Tu, D. (1995). The Jackknife and Bootstrap. New York: SpringerVerlag.

Sikkel D. and Joop Hox, E. de L. (2009) Using Auxiliary Data for Adjustment in Longitudinal Research. In Lynn, P. (Ed.), *Methodology of Longitudinal Surveys* (pp. 141-155). Chichester: Wiley.

Smith P., Lynn, P. and Elliot, D. (2009) Sample Design for Longitudinal Surveys. In Lynn, P. (Ed.), *Methodology of Longitudinal Surveys* (pp. 21-33). Chichester: Wiley.

Veiga, Alinne, Smith, Peter W. F. and Brown, James J. (2014) The use of sample weights in multivariate multilevel models with an application to income data collected by using a rotating panel survey. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, **63**, 65-84.

Watson, N. (2012) Longitudinal and Cross-sectional Weighting Methodology for the HILDA Survey. *HILDA Project Technical Paper Series*, **No. 2/12**. Melbourne: The University of Melbourne.

Figure 1: Example of Multi-Panel selection-Census Files used to create the 2006, 2011, 2016 and 2021 ACLD Panels

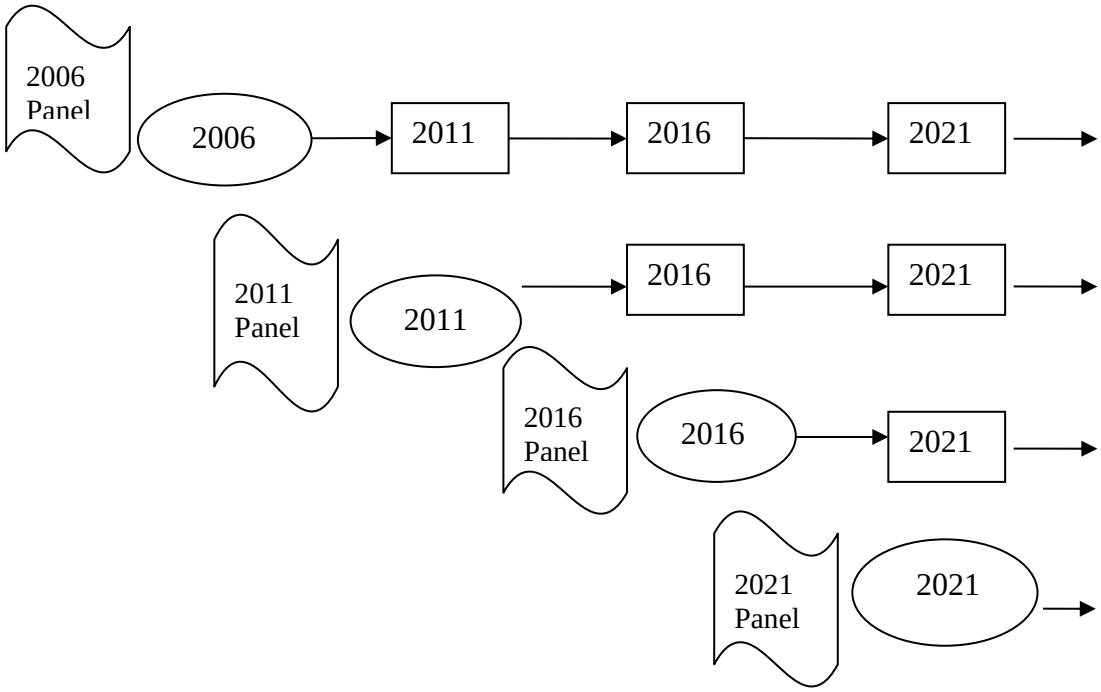

TABLE 1

Comparison of NOM and Census Approaches to Estimating the 2006-11 Longitudinal Population

Age Group in 2011	**Net Overseas Arrivals** $\hat{P}_{(06\|11)}^{(NOM)}$	**Census** $\hat{P}_{(06\|11)}^{(CEN)}$	**Difference** $(\hat{P}_{(06\|11)}^{(NOM)}\text{-}\hat{P}_{(06-11)}^{(CEN)})$
5-14	0.930	0.941	-0.011
15-24	0.844	0.908	-0.064
25-34	0.737	0.851	-0.114
35-44	0.884	0.922	-0.038
45-54	0.943	0.965	-0.022
55-64	0.965	0.982	-0.017
65-74	0.978	0.987	-0.009
75-84	0.984	0.993	-0.009
Total	0.895	0.935	-0.040

TABLE 2 Match Rate for Records

	Age >= 15 in 2006 (%)	**Age < 15 in 2006 (%)**
Birthplace		
Americas	63	40
Indigenous Australian	58	71
North Africa & Middle East	61	40
North-East Asia	55	36
North-West Europe	81	39
Oceania & Antartica (Non-Indigenous)	80	94
South-East Asia	66	30
Southern & Central Asia	48	26
Southern & Eastern Europe	86	39
Sub Saharan Africa	62	32
Missing	80	72
Mobility between 2006 to 2011		
Living at same address	96	98
Moved address	66	74
All	84	88

TABLE 3

Aboriginal and Torres Strait Islanders Status Transitions

a) Weighted Percentages (%) based on Second Release Weights			
Aboriginal and Torres Strait Islanders in 2006	Aboriginal and Torres Strait Islanders in 2011		*Total*
	Yes	No	
Yes	93.72	6.27	*100.00*
No	0.36	99.64	*100.00*
Total	*2.99*	*97.01*	*19.5 million*
b) Weighted Percentages (%) based on First Release Weights			
Aboriginal and Torres Strait Islanders in 2006	Aboriginal and Torres Strait Islanders in 2011		*Total*
	Yes	No	
Yes	93.06	6.94	*100.00*
No	0.35	99.65	*100.00*
Total	*3.15*	*96.85*	*18.6 million*

TABLE 4

Marital Status Transitions:

a) Weighted Percentages (%) based on Second Release Weights							
Marital Status (2006)	Marital Status (2011)						
	Never Married	Widowed	Divorced	Separated	Married	N/A	Total
Never Married	81.4	0.2	0.8	0.9	16.5	0.0	100
Widowed	1.0	92.8	2.0	0.3	3.8	0.0	100
Divorced	2.2	2.0	81.0	1.5	13.0	0.0	100
Separated	3.5	3.6	33.2	41.6	17.9	0.0	100
Married	0.8	2.8	2.4	3.1	90.7	0.0	100
N/A	35.6	0.0	0.0	0.0	0.0	64.2	100
b) Weighted Percentages (%) based on First Release Weights							
Marital Status (2006)	Marital Status (2011)						
	Never Married	Widowed	Divorced	Separated	Married	N/A	
Never Married	82.5	0.2	0.8	0.8	15.5	0.0	100
Widowed	1.0	92.6	2.0	0.3	4.0	0.0	100
Divorced	2.3	2.1	80.1	1.5	13.8	0.0	100
Separated	3.8	3.8	32.1	41.2	19.0	0.0	100
Married	0.8	2.8	2.1	2.7	91.6	0.0	100
N/A	32.4	0.0	0.0	0.0	0.1	67.4	100

TABLE 5:

Weights available under the ACLD until 2021

Panel Year	Panel Number (t)	Wave (Time Points)			
		***t*=1 (2006)**	***t*=2 (2011)**	***t*=3 (2016)**	***t*=4 (2021)**
2006	**1**	$\widetilde{w}_{i,(1)}$	$\widetilde{w}_{i,(1,2)}$	$\widetilde{w}_{i,(1,2,3)}$	$\widetilde{w}_{i,(1,2,3,4)}$
2011	**2**		$\widetilde{w}_{i,(2)}$	$\widetilde{w}_{i,(2,3)}$	$\widetilde{w}_{i,(2,3,4)}$
2016	**3**			$\widetilde{w}_{i,(3)}$	$\widetilde{w}_{i,(3,4)}$
2021	**4**				$\widetilde{w}_{i,(4)}$