2. Data and methods

Data

The analysis presented in this paper is based on data from the 2006 Census of Population and Housing. The census is designed to collect information on every person in Australia with the main aim of obtaining a count of the number of people at a given point in time. This count is then used to allocate the number of seats in Federal and State parliaments, as well as financial grants to various levels of government. At the same time, a large amount of information is collected on the characteristics of those counted in the census which is used for both administrative and research purposes. There were 19 855 288 individuals counted in the 2006 Census, of which 455 030 were identified as being Indigenous. Because information is collected on such large numbers of people, it is possible to obtain information on very specific population subgroups when analysing the census. That is, it is not only possible to analyse differences between the Indigenous and non-Indigenous population, but also to analyse variation within the Indigenous population.

To look at the socioeconomic correlates of particular life events, it is necessary to be able to control for variation in characteristics across individuals. To facilitate such types of analysis, the ABS provides a 5% CSF of occupied private dwellings and individuals in non-private dwellings which can be interrogated online via the Remote Access Data Laboratory. This CSF has information on 1 002 793 respondents, of which 22 437 were identified as being Indigenous; 913 262 were identified as being non-Indigenous; 56 935 did not have their Indigenous status stated; and 10,159 were overseas visitors. The latter two groups were excluded from the remainder of the analysis. Although the 5% CSF available for analysis is ostensibly a random sample, there are a number of reasons why results from the analysis may not reflect the true population values. Firstly, it may be the case that through chance the sample has different characteristics to the total population. Although this sampling error is controlled for, to a certain extent, through the use of standard errors and hypothesis tests (as outlined in the next section), the issue cannot be discounted entirely. The second set of reasons, non-sampling error, is more difficult to control for and can arise because of the way the census itself is collected and processed, or if the selection of the 5% CSF is non-random. We discuss each of these possibilities below.

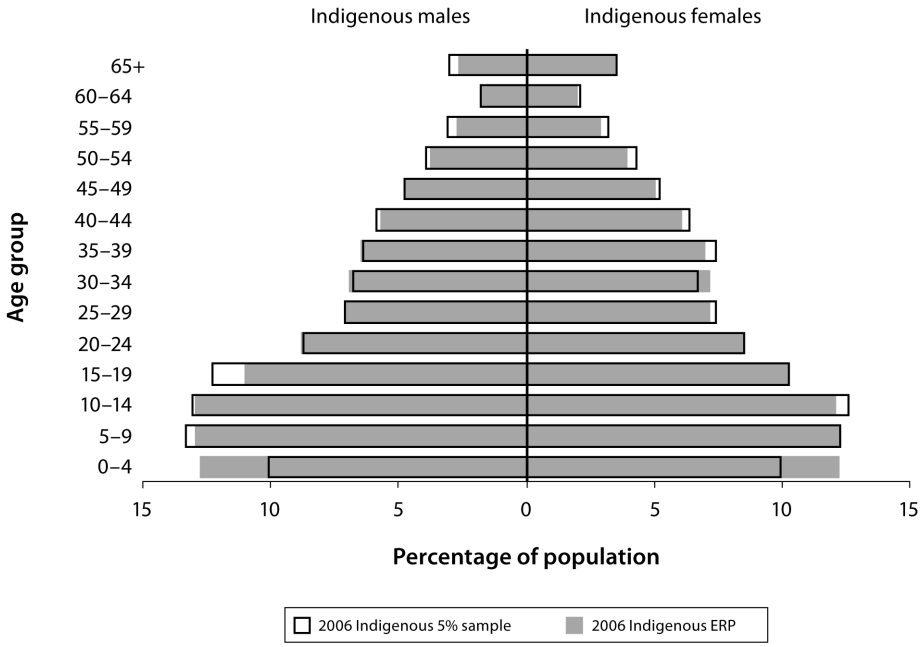
Although the aim of the census is to collect accurate information on every person in Australia, in reality this aim is not achieved. A number of people are missed, and others are counted twice. The ABS adjusts for the net undercount (the difference between those who were missed and those who were counted twice), as well as Australians who were temporarily overseas, by creating an estimated resident population (ERP). In doing so, the 19 855 288 individuals counted in the 2006 Census were adjusted upwards to create a population estimate of 20 701 448 as of 30 June 2006 (ABS 2008b).

The ABS also calculates a separate ERP for the Indigenous population, with the 455 030 people who identified as being Indigenous according to the 2006 Census count increased to an Indigenous ERP of 517 174 (ABS 2008b). The implied adjustment factor of 13.7 per cent is much higher than the adjustment factor for the total population (4.2 per cent). This is due in part to the fact that some of those who did not have their Indigenous status stated on the census were allocated to the Indigenous ERP. However, the main reason that the Indigenous population estimates were adjusted by a greater percentage than for the total population is that a higher proportion of Indigenous Australians were missed from the census altogether (ABS 2008b). That is, if those collected on the census are viewed as a sample from the true population, then unfortunately this sample is not random. If the only non-random aspect of census undercount was a person's Indigenous status, then it would be possible to adjust the estimates from the census accordingly. However, as noted by a number of authors in Morphy (2007), there are certain characteristics within the Indigenous population that make individuals more or less likely to be missed from the count than other Indigenous Australians. Specifically, individuals who live in remote Australia and in particular those who are highly mobile are more likely to be missed from the census than individuals in non-remote Australia.

Leaving aside the issue of census undercoverage, there is also the possibility that there are systematic aspects of the way in which the 5% CSF is chosen that might lead to further non-sampling error. For example, the number of people in occupied private dwellings that was included in the census was restricted to eight usual residents. Extra persons in households with more than eight usual residents were randomly removed from the sample. While this only resulted in a total of 1 283 person records being removed, it is likely that those removed were disproportionately Indigenous, due to the fact that Indigenous Australians are much more likely to live in large households (Biddle 2008). In aggregate, the size of the Indigenous sample is not substantially lower than would be expected. Indigenous Australians made up 2.5 per cent of the total ERP and 2.4 per cent of the 5% CSF. However, Figures 2.1 and 2.2 show that there is an uneven sample loss across the age distribution. Fig. 2.1 replicates the previous

age pyramid, with the distribution of the Indigenous ERP in the grey bars and the distribution of the Indigenous component of the 5% CSF in the hollow bars. Fig. 2.2 has a similar figure for the non-Indigenous population.

Fig. 2.1 Age distribution of the Indigenous ERP and the Indigenous component of the 5% CSF

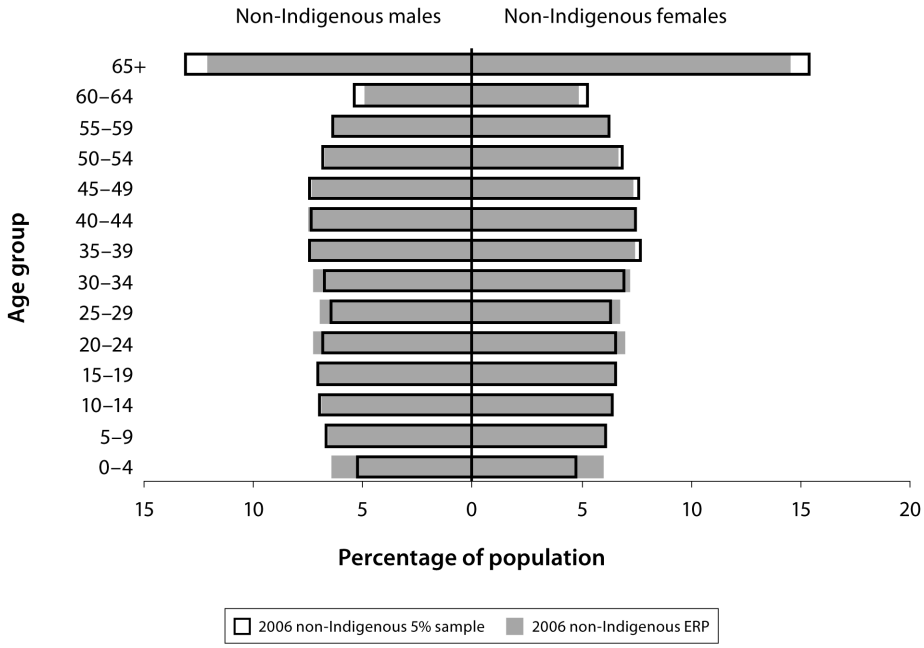


Source: Customised calculations using the 2006 5% CSF, ABS Census of Population and Housing

The percentage of the Indigenous and non-Indigenous samples that are in the majority of the age groups match up reasonably well with the equivalent percentage of the ERP. There were, however, a few exceptions, with a smaller percentage of the sample aged 0–4 years the most notable. This was particularly the case for the Indigenous population, with 10.0 per cent of the Indigenous component of the 5% CSF aged 0–4 years compared to 12.5 per cent of the Indigenous ERP. Counterbalancing the underrepresentation of 0–4 year-olds in the sample, there were slightly more Indigenous males aged 15–19 years and non-Indigenous males and females aged 65 years and over.

Ultimately, the effect of non-sampling error is mitigated to a certain extent by controlling for a number of the observable characteristics that impact on census capture in the model. However, there may still be a residual bias from unobservable characteristics that needs to be kept in mind when interpreting the results.

Fig. 2.2 Age distribution of the non-Indigenous ERP and the non-Indigenous component of the 5% CSF



Source: Customised calculations using the 2006 5% CSF, ABS Census of Population and Housing

Model specification

The analysis presented in this paper is structured around a number of thematic topics. Within each of these topics, a number of key dependent variables are chosen and the demographic and socioeconomic factors that are associated with them estimated. Table 2.1 outlines the topics and respective dependent variables. All but one of these dependent variables are binary in that they indicate whether or not a person has that particular characteristic. The exception, the number of children ever born (for females), is a count variable capped at eight children.

A separate set of explanatory variables is used for each of the dependent variables. These are chosen based on a review of the available literature whilst taking into account the limitations that arise from using census data. These specifications are outlined at the start of the relevant sections. However, reflecting our focus on the Indigenous lifecourse, there is a common structure for each of the dependent variables, including a flexible age structure, the use of a number of model specifications, and the inclusion of a separate estimation for the Indigenous population.

Table 2.1 Topics and dependent variables analysed

Topic	Dependent variable
Fertility and family formation	In a registered or de facto marriage
	In a registered marriage (for those in a registered or de facto marriage)
	Number of children ever born (for females)
	Provided unpaid child care
Migration and mobility	Changed place of usual residence between 2001 and 2006
	Away from place of usual residence on census night
Education participation	Participating in education
	Attending a non-government school (for infants, primary and secondary school students)
Employment	Employed
	Employed part-time (for those employed)
	Employed as a Manager or Professional (for those employed)
	Undertook voluntary work for an organisation or group in the last 12 months
	Undertook at least 5 hours of unpaid domestic work in the last week
Housing	Lives in a dwelling that is owned or being purchased
	Lives in a community rental dwelling (those in a rented dwelling)
	Lives in a dwelling with more than one person per bedroom
Health	Has a 'core activity' need for assistance
Childhood outcomes	Lives in a single-parent family
	Lives in a household without anyone employed
	Lives in a household where no-one has completed Year 12

For each dependent variable, a minimum of three specifications is used. In general, the first specification includes a dummy variable for whether or not the person is Indigenous; a dummy variable for whether or not the person is female; and a set of dummy variables that indicates whether the person is in a particular five-year age cohort ranging from 0–4 years to 50–54 years, with the final age cohort aged 55 years and over. The final set of variables in Model 1 is the same age dummies for females only. The purpose of this model is to test whether there is any difference between Indigenous and non-Indigenous Australians in the probability of a person having the particular characteristic after controlling for age and sex.

The second specification adds a number of additional geographic, demographic and socioeconomic characteristics to the explanatory variables in Model 1. All the dependent variables contain a dummy variable indicating that the person's usual residence is in a major city,[1] as well as a set of dummy variables indicating

1 Unfortunately, the standard Accessibility/Remoteness Index of Australia is not available for analysis using the 5% CSF. An approximation is used, with the following cities or regions included: Sydney, Newcastle/

whether the person lives in one of the seven States or Territories other than New South Wales. The remaining explanatory variables included in Model 2 vary by dependent variable and are outlined at the start of the relevant section.

The specification for Model 2 is outlined in the following equation, with Model 1 found by setting β_5, β_6 and β_7 to zero:

$$P(y_i = 1) = f\begin{pmatrix} \beta_0 + \beta_1 Indig_i + \beta_2 Fem_i + \beta_3 Age_i + \beta_4 Age_i Fem_i + \\ \beta_5 Majcit_i + \beta_6 State_i + \beta_7 X_i \end{pmatrix}$$

In the above equation, $Indig_i$, Fem_i and $Majcit_i$ are binary variables indicating the Indigenous status, sex and location of usual residence of individual i, whereas Age_i, $State_i$ and X_i are vectors of characteristics. The coefficients β_1, β_2 and β_5 as well as the vectors of coefficients β_3, β_4, β_6 and β_7 indicate the size and direction of the association between the dependent and explanatory variables. The constant term (β_0) determines the probability of the *base case* individual having that particular characteristic. Apart from being male, non-Indigenous and living outside a major city, this base case is different for different dependent variables and defined in the relevant sections.

There are two ultimate purposes of Model 2. Firstly, to test whether there is still variation across the lifecourse in the probability of having that particular characteristic after controlling for geography and other socioeconomic outcomes. Secondly, to test whether there is still a significant difference between the Indigenous and non-Indigenous population after controlling for these characteristics in addition to variation across the lifecourse.

The third specification that is estimated for all dependent variables is designed to identify whether the patterns across the lifecourse hold for the Indigenous population in isolation. Model 3 is therefore estimated on the Indigenous sample only and has β_1 set to zero.

A fourth specification (Model 4) is also estimated for a number of the dependent variables. This specification is estimated for the Indigenous population only and includes an additional variable for whether or not the individual lives in a mixed Indigenous and non-Indigenous household. This variable is not included in Model 3, as it would make it more difficult to compare the estimated marginal effects with results from Model 2, and it is necessary to restrict the sample to those who live in private dwellings only. For this reason, certain dependent variables that can only take on one particular value for those in a non-private dwelling (for example temporary mobility) do not have a Model 4 estimated for

Hunter, Wollongong/Illawarra, Melbourne, Brisbane, Gold Coast, Sunshine Coast, Adelaide, Perth, and the Australian Capital Territory. It was not possible to separately identify Hobart from the rest of Tasmania or Darwin from the rest of the Northern Territory.

them. The following table summarises the four estimated models. There is some variation depending on the particular dependent variable. This is discussed further in the relevant chapter in this volume.

Table 2.2 Summary of estimated models

Model	Population	Included variables
Model 1	Indigenous and non-Indigenous	Indigenous status, sex, five-year age group, sex interacted with five-year age group
Model 2	Indigenous and non-Indigenous	Same as Model 1, plus individual socioeconomic variables
Model 3	Indigenous population only	Same as Model 2, but without Indigenous status
Model 4	Indigenous population living in a private dwelling	Same as Model 3, plus whether or not the household includes non-Indigenous usual residents

Because the majority of the dependent variables are binary (that is you either do or do not have the characteristic), the assumed functional form of the majority of the models is the standard probit. The parameters of the models, that is the β coefficients and their standard errors, are estimated using the maximum likelihood estimation procedure. The only exception to this is the analysis of the number of children ever born. As this dependent variable is constructed using count data, the Poisson model is used (after testing for and rejecting over-dispersion).

It is worth pointing out at this stage the limitations of the analysis presented in this monograph and what it is not trying to achieve. Firstly, in the absence of longitudinal data, it is not possible to analyse how the outcomes of individual Indigenous Australians have changed through time. Rather, comparisons are restricted to the average outcomes of Indigenous or non-Indigenous Australians aged 30–34 years in 2006 with those aged 35–39 years (for example), after controlling for other observed characteristics. A further limitation of using cross-sectional data is that it is not possible to identify causal relationships between the explanatory and dependent variables. That is, it is possible to identify whether or not living in a major city is associated with a higher or lower probability of being married (for example), but not possible to show whether geographic location has a direct impact. It may be a situation of reverse causality, with those who are married being more or less likely to live in a particular area. Alternatively, there may be a third unobserved variable that influences them both.

Because of these limitations, no attempt is made in the analysis to model the dependent variables as a system of equations. It would not be possible with cross-sectional census data to identify instrumental variables that are correlated

with the endogenous explanatory variables but not with the dependent variable. Rather, each model is estimated as a single equation, with the potential bias in the estimated standard errors kept in mind when interpreting the robustness of the results. Furthermore, although theoretical justifications are used for each of the model specifications, variables that were used as a dependent variable in a particular estimation are still considered as explanatory variables in other equations.