

Estimating the Count Error in the Australian Census

James Chipperfield¹, James Brown², and Philip Bell³

In many countries, counts of people are a key factor in the allocation of government resources. However, it is well known that errors arise in Census counting of people (e.g., undercoverage due to missing people). Therefore, it is common for national statistical agencies to conduct one or more “audit” surveys that are designed to estimate and remove systematic errors in Census counting. For example, the Australian Bureau of Statistics (ABS) conducts a single audit sample, called the Post Enumeration Survey (PES), shortly after each Australian Population Census. This article describes the estimator used by the ABS to estimate the count of people in Australia. Key features of this estimator are that it is unbiased when there is systematic measurement error in Census counting and when nonresponse to the PES is nonignorable.

Key words: Undercount; capture-recapture; Dual System Estimator.

1. Introduction

In many countries, counts of people are a key factor in the distribution of government resources. However, the (observed) Census counts differ from the (unobserved) true counts for an area because of overcoverage (e.g., a person is counted multiple times or counted once in the wrong location) and undercoverage (e.g., people are missed). In addition, the Census may count individuals on a ‘person present’ basis (de facto) while counting on a ‘usual residents’ basis (de jure) is typically more relevant for government policy. ‘Usual residents’ counts may be significantly smaller than ‘persons present’ counts in certain locations (e.g., tourist areas and city centres) and this will have implications for the type and size of government infrastructure projects.

Methods have been developed to correct for systematic errors in the observed Census counts (for a review see [Belin and Rolph 1994](#)). A classical approach to estimating person counts is the Dual System Estimator (DSE) as developed in [Sekar and Deming \(1949\)](#). In this traditional form, the DSE has been widely used by national statistical agencies, including the US Census Bureau Bureau (see [Xi Chien and Tang 2011](#); [Mule 2008](#); [Griffin and Mule 2008](#); [Alho et al. 1993](#); and [Hogan 1993](#)) and the Office for National Statistics

¹ Australian Bureau of Statistics, Methodology Division, P O Box 10 Belconnen Australian Capital Territory 2616, Australia. Email: james.chipperfield@abs.gov.au

² University of Technology, Sydney, School of Mathematical Sciences, Sydney, New South Wales, Australia. Email: james.brown@uts.edu.au

³ Australian Bureau of Statistics, Methodology Division, Adelaide, South Australia, Australia. Email: philip.bell@abs.gov.au

Acknowledgments: We would like to thank Bill Gross, David Steel, Stephen Carlton and Frank Yu for pointing us in appropriate directions and linking us to relevant literature. Prof James Brown at UTS is part supported by a research grant from the ABS.

(ONS) (see [Brown et al. 1999](#); [Brown et al. 2006](#)). However, in the Australian context, the Australian Bureau of Statistics (ABS) has developed an approach that differs fundamentally from the US and UK approach on perhaps three points.

First, in the US, the approach has developed with two independent ‘audit’ sample surveys of the Census (see [Hogan 1993](#)). In the UK (see [Large et al. 2011](#)), the approach has developed from a single independent coverage survey (audit sample) but with separate adjustments for erroneous enumerations (overcoverage and undercoverage). In contrast, the ABS approach integrates measurement of undercoverage and overcoverage of its Census from a single survey, called the Post Enumeration Survey (PES).

Second, while the Australian Census counts on a ‘persons present’ basis, the PES counts on a ‘usual residents’ basis, so the second count (PES) is not just a repetition of the first count (Census).

Third, as we will see, the ABS approach has also been developed to handle people who are classified differently by the PES and the Census (e.g., a person may be classified as an Aboriginal and Torres Strait Islander by the PES, but not by the Census). The assumption here is that the PES classification is correct. This is perhaps reasonable, since the PES consists of a face-to-face interview conducted by ABS’s professional interviewers, while the Census typically uses self-completion, supported by a large temporary field-force.

In Section 2 we review the general DSE framework, as applied by the US Census Bureau. Section 3 introduces the standard regression estimator of person counts and motivates the more robust estimator that was used in 2011. Section 4 describes a small simulation study of the two estimators in Section 3. Section 5 describes a more realistic and involved simulation study of the two estimators in Section 3. Section 6 contains some concluding remarks.

2. Traditional Approach

In the traditional approach, a population set, U , is defined of people indexed by j . The population can be categorised into H subgroups, defined in terms of characteristics such as age, sex, Aboriginal and Torres Strait Islanders status, country of birth, and geography. Subgroups need not be mutually exclusive. The aim is to estimate the number of people in the h th subgroup $T_h = \sum_{j \in U} t_{jh}$ where $t_{jh} = 1$ if the j th person in the population belongs to subgroup h and $t_{jh} = 0$ otherwise and $h = 1, \dots, H$. The population is counted by the Census. The Census count is denoted for subgroup h by X_h for $h = 1, \dots, H$. While in the traditional DSE approach T_h and X_h are conceptually the same, in the Australian situation they are not (see Section 3).

Consider the situation where following the Census, we conduct an independent PES of the population, typically by sampling dwellings. In the context of the US Census Bureau, the reference date for PES counting would be Census Night and the PES is referred to as the P-sample, as it is a sample of the population ([Hogan 1993, 2003](#)).

The PES will, of course, have nonresponse (undercoverage), but overcoverage can be assumed to be zero because it uses ABS professional interviewers who are familiar with applying rigorous procedures to correctly identify and avoid double-counting usual residents within selected dwellings. For example, these procedures ensure that overseas visitors are identified and discarded from estimation and that each person can only be

selected via a single dwelling. Let the PES responding sample size be n people, and denote the sample set by s . The PES collects t_{ih} where $t_{ih} = 1$ if the i th sample person belongs to subgroup h and $t_{ih} = 0$, $h = 1, \dots, H$, and $i = 1, \dots, n$. After matching the PES to the Census, we can then derive an indicator, m_{ih} , for ‘correct Census enumeration’, where $m_{ih} = 1$ if person i is counted in subgroup h by both the PES and Census, and otherwise $m_{ih} = 0$. Let $w_i = \pi_i^{-1}$, where π_i is the probability, with respect to the PES sample design, that person i was selected.

2.1. Estimating Undercoverage Using a P-Sample

Now, if we impose the assumption that the PES and Census enumerate the populations independently, the classic DSE of T_h without an adjustment for overcoverage (previously used by the US Census Bureau, see [Hogan 1993](#)) is

$$\hat{T}_h = \hat{R}_h \times X_h. \tag{1}$$

where

$$\hat{R}_h = \frac{\sum_{i \in s} w_i t_{ih}}{\sum_{i \in s} w_i m_{ih}}, \tag{2}$$

is an estimate of the probability that an individual is missed by the Census, for all $h = 1, \dots, H$.

This probability, \hat{R}_h , adjusts the total Census count, X_h , for Census undercoverage.

2.2. Estimating Over- and Undercoverage Using E- and P-Samples

For several practical reasons, such as enumerators following up the wrong households when the forms are posted-out, the Census count X_h will have a typically low level of overcoverage. [Table 1](#) in [Large et al. \(2011\)](#) shows that historical levels of overcount have been less than one per cent in Switzerland, Canada, Australia, and the UK; the US being an

Table 1. Description of the simulation scenarios.

Scenario	Error rates*			
	Census misclassification	Census duplication	Census missing	PES missing
1	0.1	LOW	LOW	LOW
2	0.1	MEDIUM	MEDIUM	MEDIUM
3	0.2	LOW	LOW	LOW
4	0.2	MEDIUM	MEDIUM	MEDIUM
5	0.3	MEDIUM	MEDIUM	MEDIUM
6	0.3	HIGH	HIGH	HIGH

*Low scenario = 0.1 in communities and 0.05 outside communities
 *Medium scenario = 0.2 in communities and 0.1 outside communities
 *High scenario = 0.3 in communities and 0.15 outside communities

exception, due to a post-out Census in most areas for several decades. Recently, Census overcount has been increasing; in the 2011 Canadian Census it was 1.85%.

If there is overcoverage in the Census count, X_h , using (1) will be positively biased for T_h . Thus, an additional adjustment is required. In the US context, this involves selecting a sample of Census records and confirming whether the enumerations were correct. This sample is referred to as the E-sample (Hogan 1993). For the k th Census record sampled in the E-sample, we resolve whether the record should have been enumerated ($e_k = 1$) or should not have been enumerated ($e_k = 0$). The DSE of T_h is now

$$\hat{T}_h^{(DSE)} = \hat{R}_h \times X_h \times \hat{E}_h \quad (3)$$

where \hat{R}_h still adjusts for undercoverage, while

$$\hat{E}_h = \frac{\sum_{kh} v_k e_k}{\sum_{kh} v_k} \quad (4)$$

is an estimate of the probability that an enumerated Census record is actually a correct enumeration in subgroup h , \sum_{kh} is the summation of records in the E-sample classified to subgroup h , and v_k is the appropriate sampling weight for the k th record in the E-sample.

As already stated, both (1) and (3) assume that enumerations in Census and PES occur independently, conditional on subgroup. When this independence assumption is not true, the DSE will be biased for T_h (see, for example Wachter and Freedman 2000). Reducing this ‘correlation bias’ is possible by bringing in external information, such as a known sex ratio, as developed in Wolter (1990) and implemented in Bell (2001), or some other additional information (Brown et al. 2006).

In the 2010 Census, the US Census Bureau (see Mule 2008) extended (3) by modelling the probability of correct enumeration and the probability of incorrect enumeration at an individual level (i.e., by fitting a logistic regression to m_i in the P-sample and e_k in the E-sample).

3. The ABS Approach

3.1. Differences Between the Australian and Traditional Approaches

Now we bring in the Australian context, which has additional complications. First, there is only one additional sample - a P-sample (i.e., there is no E-sample). This must be taken into account when estimating overcoverage in the Census.

Second, there are systematic differences between Census counts and PES counts where, for the reasons outlined below, we are interested in the latter. The Census counts people on a ‘person present’ basis, while the PES counts people on a ‘usual residents’ basis. The latter is typically more useful to the government when allocating resources. This means it is quite legitimate for a person to be in one geographic area in the Census and in a different area in the PES. (To facilitate matching a person’s Census and PES records, the PES asks respondents about possible locations for their Census enumeration.) The number of ‘movers’ is expected to be small, given that the Census and the PES are carried out only a

couple of weeks apart. In the US approach, in which the time between Census and survey is longer, an adjustment is made to account for ‘movers’ (Griffin 2000). It is also possible for the Census and the PES to classify a person in different subgroups, even if they are enumerated at the same geographic location. This discrepancy in classification is more noticeable in some subgroups, in particular to a more significant extent in Aboriginal and Torres Strait Islanders status than in others. Since the Census uses self-enumeration, usually with one individual responding for all household members; the PES is more likely to be ‘correct’ in the Australian context, where the ABS utilises its professional field-force for the PES interviews. For the reasons mentioned, it is assumed here that the PES always correctly classifies a person to subgroup. No such assumption is made with regard to census classification. (It is worthwhile to note here that the traditional DSE in Section 2 does not correct for systematic differences in PES and Census classification of people to subgroup).

The observed Census counts $\mathbf{X}^T = (X_1, \dots, X_h, \dots, X_H)$ are calculated by summing over all Census records in each of the H subgroups. We may consider expressing the Census counts by $\mathbf{X} = \sum_{j \in U} \mathbf{x}_j$, where $\mathbf{x}_j^T = (x_{j1}, \dots, x_{jh}, \dots, x_{jH})$, $X_h = \sum_{j \in U} x_{jh}$, and x_{jh} is the number of times person j was counted by the Census in subgroup h for $j \in U$. If person j in the population is missed by the Census (i.e., not counted in any subgroup), then we can notionally set $\mathbf{x}_j = \mathbf{0}$, where $\mathbf{0}$ is an H column vector of zeros. While we can calculate \mathbf{X} , we do not observe \mathbf{x}_j for all j , as this would require that each person in the population is identified and explicitly assigned a value for \mathbf{x} .

In other contexts, \mathbf{x} can define characteristics of people in administrative data. Many countries now use administrative data as either the entire basis for their census or as a major component of their census. Valente (2010) provides a review of different approaches taken by different European countries. In the case of the Netherlands (see Nordholt 2005), \mathbf{x} contains classification errors, and the PES functions somewhat as a quality correction for the administrative data, rather than as a coverage check (Brown and Honchar 2012).

After matching PES and Census records, we can establish x_{ih} , the number of times the i th PES respondent was counted by the Census in subgroup h . Again, $x_{ih} = 0$ if the Census did not count the i th PES respondent in subgroup h (i.e., if the i th PES record could not be matched to any Census record in that subgroup). The variable x_{ih} captures information about overcoverage (if greater than 1) and undercoverage (if zero). This is important since here, unlike in the DSE approach, we only have a single audit sample (i.e., there is a P-sample but no E-sample) to capture information about over- and undercoverage. Integrating the two also recognises that both errors are inherent in the Census, and our target is to recognise this in our estimation. We now have several possible types of coverage outcomes; the main ones being:

- $t_{ih} = 1$, $x_{ih} = 1$, and $x_{ig} = 0$ and $t_{ig} = 0$ for all $g \neq h$
 - the PES and Census counts a person once in the same subgroup,
- $t_{ih} = 1$ and $x_{ih} = 2$
 - the PES and Census counts a person once and twice in the same subgroup, respectively (e.g., duplication),

- $t_{ih} = 1$ and $x_{ig} = 0$ for all g
 - the PES counts a person but the Census did not count them at all,
- $t_{ih} = 1$, $x_{ih} = 0$, $x_{if} = 1$, $x_{ig} = 0$ for all $g \neq h$ or $g \neq f$
 - the PES and Census counts a person in different subgroups. This could be due to Census misclassification or because a person was enumerated by the PES and Census in different geographic locations, and
- $t_{ih} = 1$, $x_{ih} = 1$, $x_{if} = 1$, $x_{ig} = 0$ for all $g \neq h$ or $g \neq f$
 - the PES and Census counts a person once in the same subgroup but the Census also counts the person in a different subgroup.

Third, we know that certain subgroups of the population are more likely to be missed by the PES, even after conditioning on auxiliary information available from the Census. As we see in the next section, this would mean that the standard generalised regression estimator (Subsection 3.2) for the PES in the Australian context would be biased, and so we consider an alternative (Subsection 3.3).

3.2. Generalised Regression Estimator Using a Prediction Model

For simplicity, in the rest of this article we replace t_{ih} with t_i , where $t_i = 1$ if person i is in an arbitrary subgroup of interest (i.e., we drop the h subscript). Similarly, we replace t_{jh} with t_j . Now we are interested in estimating $T = \sum_j t_j$ the usual resident population in an arbitrary subgroup.

Consider the ‘working’ linear prediction model

$$t_j = \mathbf{x}_j^T \boldsymbol{\alpha} + e_j \quad (5)$$

$$E(e_j | \mathbf{x}_j) = 0$$

where the e_j s are independent and identically distributed and $\boldsymbol{\alpha}$ is an H column vector of coefficients that relate to membership of the H subgroups. We call (5) a ‘working’ model because a linear model is not ideal for a binary variable such as t . Nevertheless, we may use this model to motivate the classic generalised regression estimator,

$$\hat{T}^{(GREG)} = \sum_{i \in S} w_i (t_i - \mathbf{x}_i^T \hat{\boldsymbol{\alpha}}) + \mathbf{X}^T \hat{\boldsymbol{\alpha}}$$

where $\hat{\boldsymbol{\alpha}} = (\sum_{i \in S} w_i \mathbf{x}_i \mathbf{x}_i^T)^{-1} (\sum_{i \in S} w_i \mathbf{x}_i^T t_i)$ is the standard ‘survey weighted’ least squares estimator of $\boldsymbol{\alpha}$ (see [Särndal et al. 1992](#)).

If we now allow for nonresponse in the PES, we need to consider whether or not the condition under which $\hat{T}^{(GREG)}$ is asymptotically unbiased is reasonable. Denote the response indicator by I_j , where $I_j = 1$ if person j in the population would respond if selected in the PES and otherwise $I_j = 0$. Now consider the distribution of t given \mathbf{x} in the population,

$$[t_j | \mathbf{x}_j; j = 1, \dots, N] = [t | \mathbf{x}]$$

and the distribution of t given \mathbf{x} in the population of PES respondents,

$$[t_j | \mathbf{x}_j, I_j = 1; j = 1, \dots, N] = [t | \mathbf{x}, I = 1].$$

If these two distributions are equal, we may write

$$[t|x] = [t|x, I = 1]. \quad (6)$$

From (5) and (6) it follows that $\hat{T}^{(GREG)}$ is asymptotically unbiased in the presence of nonresponse (see also Kott and Chang 2010). The condition in (6) means that the distribution of t given x is the same for PES respondents and PES nonrespondents and so we may say that nonresponse is ignorable given x (see Rubin and Little 2002); we may also say that the response, t , and the indicator for response, I , are independent conditional on x and so we may write $[t, I|x] = [I|x][t|x]$.

However, there is strong evidence against (5) or (6) holding in the case of the PES. To illustrate this, consider breaking up the population U into Census respondents ($x_j \neq \mathbf{0}$) and Census nonrespondents ($x_j = \mathbf{0}$). For Census nonrespondents, (5) and (6) become

$$t_j = e_j \quad (7)$$

$$E(e_j|x_j = \mathbf{0}) = 0$$

$$[t|x = \mathbf{0}] = [t|x = \mathbf{0}, I = 1] \quad (8)$$

for all $j \in U$. Equation (7) implies that, for Census nonrespondents, the *unconditional* mean of t in the population and in the population of PES respondents is zero. That is, if a person is missed by the Census, the model expects them to be missed by the PES. This is clearly not an appropriate assumption, since the PES is designed to capture information about Census undercoverage.

Equation (8) is equivalent to the assumption that, within the population of Census nonrespondents, *nonresponse occurs completely at random*; that is, within the population of PES nonrespondents, t , and the nonresponse indicator, I , are unconditionally independent (Rubin and Little 2002). There are at least two reasons why (8) is unlikely in the case of the PES. First, there is strong practical evidence that Aboriginal and Torres Strait Islanders people living in remote communities and people aged 20–29 have a higher rate of being missed by the PES.

Second, it is reasonable to suppose that whether a person responds to the PES may be correlated in some way to whether the person responds to the Census. For example, people may avoid the PES interviewer specifically because they do not want to own-up to being a Census nonrespondent. This would lead to the PES sample having an unrepresentatively high rate of completed Census forms. This creates the ‘correlation bias’. In dual sampling literature this correlation is sometimes assumed to be negligible after conditioning on an appropriate set of covariates. However, for Census nonrespondents, there are effectively no covariates (x) on which to condition.

In order to reduce any impact of the PES on the Census response (another potential form of ‘correlation bias’), the PES is conducted four weeks after Census night. Census records matched with PES records (i.e., x in this article) are those that were received before the date that PES field operations began, Census forms that were returned by mail or by Internet after this date are essentially ignored. In addition, data collected by the PES and Census are processed independently.

In short, for the reasons outlined above, nonresponse is *not* ignorable given x . Given $\hat{T}^{(GREG)}$ is biased in this case, next, we consider an alternative.

3.3. Using a Two Stage Prediction Model

Now consider a vector \mathbf{z} that is comprised of variables on people selected in the PES. Accordingly, \mathbf{z} may be a function of \mathbf{x} or t . Now, instead of (6) consider assuming

$$[t_j | \mathbf{z}_j, I_j = 1] = [t_j | \mathbf{z}_j] \text{ for } j = 1, \dots, N \quad (9)$$

$$[\mathbf{x}_j | \mathbf{z}_j, I_j = 1] = [\mathbf{x}_j | \mathbf{z}_j] \text{ for } j = 1, \dots, N \quad (10)$$

Assumption (9) is that the distribution of $t | \mathbf{z}$ in the population and in the population of PES respondents is the same. Assumption (10) is that the distribution of $\mathbf{x} | \mathbf{z}$ in the population and in the population of PES respondents is the same. If \mathbf{z} is a function of only \mathbf{x} , then (9) and (10) collapse to (6). In an attempt to overcome the failings of (6), we allow \mathbf{z} to be a function of t . By allowing \mathbf{z} to be a function t we allow the indicator for nonresponse, I , to depend upon the response value itself (see Little and Rubin 2002). In this case, nonresponse is said to be nonignorable given \mathbf{x} .

To make the ideas more concrete, consider the underlying working model

$$t_j = \mathbf{z}_j^T \boldsymbol{\theta} + \varepsilon_{1j} \quad (11)$$

$$\mathbf{x}_j = \mathbf{z}_j^T \boldsymbol{\gamma} + \varepsilon_{2j},$$

$$E(\varepsilon_{1j} | \mathbf{z}_j) = 0$$

$$E(\varepsilon_{2j} | \mathbf{z}_j) = \mathbf{0}$$

where ε_{1j} and ε_{2j} are independent over j . Using (11) in a regression of t on \mathbf{x} , it follows that $\boldsymbol{\alpha}$ and e_j from (5) can be expressed by $\boldsymbol{\alpha}^* = \boldsymbol{\gamma}^{-1} \boldsymbol{\theta}$ and $e_j^* = \varepsilon_{1j} - \varepsilon_{2j} \boldsymbol{\gamma}^{-1} \boldsymbol{\theta}$. We may then re-consider the working model of (5) and write

$$t_j = \mathbf{x}_j^T \boldsymbol{\alpha}^* + e_j^* \quad (12)$$

$$E(e_j^* | \mathbf{x}_j) = 0$$

We know from (9), (10), and (11) that (12) holds in the population and in the population of PES responders. So while (5) and (12) are both linear regressions with the same dependent and independent variables, only (12) holds in the population.

Now we may use (12) in a standard generalised regression estimator. Accordingly, unbiased estimates of $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\alpha}^*$ are

$$\tilde{\boldsymbol{\theta}} = \left(\sum_{i \in s} w_i \mathbf{z}_i \mathbf{z}_i^T \right)^{-1} \left(\sum_{i \in s} w_i \mathbf{z}_i t_i \right),$$

$$\tilde{\boldsymbol{\gamma}} = \left(\sum_{i \in s} w_i \mathbf{z}_i \mathbf{z}_i^T \right)^{-1} \left(\sum_{i \in s} w_i \mathbf{z}_i \mathbf{x}_i \right)$$

and

$$\tilde{\boldsymbol{\alpha}}^* = \tilde{\boldsymbol{\gamma}}^{-1} \tilde{\boldsymbol{\theta}} = \left(\sum_{i \in s} w_i \mathbf{z}_i \mathbf{x}_i^T \right)^{-1} \left(\sum_{i \in s} w_i \mathbf{z}_i t_i \right)$$

and the classical regression estimator of T under (12) is

$$\hat{T}^{(PREG)} = \sum_{i \in s} w_i (t_i - \mathbf{x}_i^T \tilde{\boldsymbol{\alpha}}^*) + \mathbf{X}^T \tilde{\boldsymbol{\alpha}}^* \quad (13)$$

Since under (12), we know that $E(T) = \mathbf{X}^T \boldsymbol{\alpha}^*$ and $E(\sum_{i \in s} w_i e_j^*) = 0$ it follows that, since $\tilde{\boldsymbol{\alpha}}^*$ is unbiased for $\boldsymbol{\alpha}^*$, $E(\hat{T}^{(PREG)}) = T$. Kott and Chang (2010) show that (13) is unbiased and note that for $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\gamma}}^{-1}$ to exist the inverse of $\sum_{i \in s} w_i \mathbf{z}_i \mathbf{x}_i^T$ must exist and so constrain the dimension of \mathbf{x} and \mathbf{z} to be the same. Since (13) is essentially a function of means, the jackknife can be used to give an asymptotically unbiased estimate of the sampling variance of $\hat{T}^{(PREG)}$ in large samples.

If we let the predicted value of \mathbf{x}_i be $\tilde{\mathbf{x}}_i = \mathbf{z}_i^T \hat{\boldsymbol{\gamma}}$ it is easy to show that an alternative expression for (13) is $\hat{T}^{(PREG)} = \sum_{i \in s} w_i g_i t_i$, where $g_i = 1 + (\mathbf{X} - \sum_{i \in s} w_i \tilde{\mathbf{x}}_i)^T (\sum_{i \in s} w_i \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T)^{-1} \tilde{\mathbf{x}}_i$ - the same as $\hat{T}^{(GREG)}$ but with $\tilde{\mathbf{x}}$ replacing \mathbf{x} . In other words, the Predicted value of \mathbf{x} is used in an otherwise standard REGression estimator (PREG). In other words, the weight adjustment g_i depends upon \mathbf{z}_i (via $\tilde{\mathbf{x}}_i$) instead of \mathbf{x}_i and so the weight adjustment does not depend on whether or not person i is a Census respondent. This is analogous to classic DSE, where membership of a subgroup defines the PES nonresponse adjustment for both the Census responders and nonrespondents.

In 2011, the ABS application of \mathbf{x} and \mathbf{z} contained the number of times a person was counted on a ‘persons present’ basis (collected by Census) and ‘usual –residents’ basis (collected by the PES) in each subgroup, respectively, where subgroup was defined by cross-classifying age, sex, Aboriginal and Torres Strait Islanders status collected by the PES. The dimension of \mathbf{x} and \mathbf{z} was about $H = 450$ (i.e., 450 subgroups indicators). Defining \mathbf{z} completely in terms of PES counts allowed for nonignorable nonresponse in the PES.

The proposed estimator of (13) relies on assumptions similar to the DSE:

- Assumptions (9) and (10) correspond to the assumption of *independence* between the PES and the Census. If the independence assumption was violated then the distributions in the sample and populations would not be the same, as required by (9) and (10);
- The assumption of *perfect matching* between the PES and the Census ensures the values of \mathbf{x} that are assigned to PES respondents are correct;
- The *closed population* assumption is implicit in the definition of the population set U . We assume that everyone who responds to the Census must be in U and all people in U have a chance of being selected by the PES. The closed population assumption will be violated (and (13) will be biased) if a Census respondent *does not* correctly identify as an overseas visitor and leaves Australia before the PES; and
- The *homogeneity of response* assumption is implicit in the response model of (11).

Chang and Kott (2008) consider a response propensity model, rather than a prediction model, such as (11), to justify (13). We may suppose that $p_j = E(I_j) = 1/f(\mathbf{z}_j^T \boldsymbol{\theta})$, where $f(\cdot)$ is some appropriate nonlinear function and $\boldsymbol{\theta}$ is a matrix of coefficients. Chang and Kott (2008) show that under this response propensity model (and certain conditions), an unbiased estimator of T is

$$\hat{T}^{(RESP)} = \sum_{i \in s} w_i f(\mathbf{z}_i^T \boldsymbol{\omega}) t_i$$

where ω is an H vector (same dimension as \mathbf{z}) of constants that satisfies $X = \sum_{i \in s} w_i f(\mathbf{z}_i^T \omega) \mathbf{x}_i$. Chang and Kott (2008) show that the estimators $\hat{T}^{(RESP)}$ and $\hat{T}^{(PREG)}$ have the same form if $f(\mathbf{z}_i^T \omega) = 1/(1 + \mathbf{z}_i^T \omega)$. Here, it may be more appropriate to consider a response propensity model rather than a linear model, such as (5), for a binary variable such as t . This will be the subject of future work.

4. Simulation to Demonstrate the PREG

This section describes a simple but illustrative simulation study of the PREG and GREG. Given that we are working in a situation in which there is no E-sample, the DSE was not evaluated. In this simulation, the population is made up of four subgroups of interest (i.e., there are four different ways of defining t). The subgroup population totals of interest are 5000 (Subgroup 1: non-Aboriginal and Torres Strait Islanders males), 2000 (Subgroup 2: Aboriginal and Torres Strait Islanders males), 2000 (Subgroup 3: non-Aboriginal and Torres Strait Islanders females), and 1000 (Subgroup 4: Aboriginal and Torres Strait Islanders females). The proportion of people in subgroup 1, 2, 3, 4 living in Aboriginal communities is 0.2, 0.8, 0.3, and 0.7, respectively. Using these proportions, each person in a subgroup is randomly assigned a value for c , where $c = 1$ if a person lives in an Aboriginal community and otherwise $c = 0$. In reality, Census and PES coverage of Aboriginal communities is potentially more difficult due to their remoteness and the wide geographic area that they cover.

The population is counted by the Census. Each person in the population is assigned a value for $\mathbf{x} = (x_1, x_2, x_3, x_4)$ where x_{hj} is the number of times the person was counted in subgroup h by the Census. The type of errors in the Census counting include misclassification (records in subgroup $h = 2$ are misclassified to subgroup $h = 1$), duplication (person counted twice), and missing (person not counted). Consistent with earlier notation, if a person is missed, then $\mathbf{x} = \mathbf{0}$, if a person in subgroup $h = 1$ is duplicated in the subgroup, then the first element of \mathbf{x} is 2 and all other elements are zero. The probability of these Census counting errors occurring depends only on c . Table 1 gives these probabilities for a range of scenarios. For example, in Scenario 1 the Census rate of misclassification, duplication and missing for people living in Aboriginal communities was 0.1.

A simulated PES sample of size 1,000 people was selected by SRSWOR from the population. From the sample the variable c and subgroup membership variables z_{hi} were collected, where $z_{hi} = 1$ if person i belonged to population h and is otherwise zero. However, the PES did miss people. As in the Census, the probability of the PES missing a person depended upon c . Table 1 gives these probabilities for a range of scenarios. For example, in Scenario 1 the probability of missing a person living within and outside an Aboriginal community was 0.1 and 0.05, respectively. Since it is difficult, in practice, to measure the probability of missing a person when the ‘missing’ mechanism itself is nonignorable, the range of scenarios aims to explore a wide scope of possibilities (rather than be motivated by a particular case study).

The PREG and GREG used the same definition of \mathbf{x} . The following estimators were considered:

PREG1: Equation (13) with $\mathbf{z} = \mathbf{1}$. This assumes that the PES provides a representative sample of t and of \mathbf{x} from the population.

- PREG2:

Equation (13) with $\mathbf{z} = (1, r)$, where $r = z_2 + z_4$ is an Aboriginal and Torres Strait Islanders status indicator. If it is suspected that PES nonresponse depends on Aboriginal and Torres Strait Islanders status alone then there would be substantive reasons for defining \mathbf{z} in this way. In this situation r is a good proxy for c , which drives the PES nonresponse mechanism.
- PREG3:

Equation (13) with $\mathbf{z} = (1, c)$. This is the estimator that would be used if the mechanism generating the count errors in Census and PES were known.
- GREG:

Generalised regression estimator with auxiliary \mathbf{x} . This estimator assumes that the distribution of $\mathbf{t}|\mathbf{x}$ in the PES and in the population is the same.

The results are presented in Table 2 in terms of Relative Bias of $\hat{T} = (\hat{T} - T)/T$. Since the dimension of \mathbf{x} and \mathbf{z} were not the same, we calculated the generalised inverse of $(\sum_{i \in s} w_i \mathbf{z}_i \mathbf{x}_i^T)$ to calculate the PREG.

Table 2 shows that the GREG is consistently biased, but that reduced the bias in the observed Census counts. The table shows that PREG1 is biased because it is not well specified. PREG2 assumes PES nonresponse depends on ‘Aboriginal and Torres Strait

Table 2. Bias (%) in population estimates for various scenarios.

Scenario	GREG				PREG1			
	Subgroup							
	1	2	3	4	1	2	3	4
1	5.6	9.8	6.0	9.0	−0.9	2.2	−0.8	2.5
2	11.0	20.0	12.0	17.8	−1.8	5.5	−2.3	4.7
3	5.3	11.0	6.0	8.8	−0.6	1.5	−1.9	1.2
4	10.7	21.9	11.6	17.6	−2.5	5.0	−2.5	4.3
5	10.6	22.1	12.0	18.5	−2.1	4.9	−1.8	3.5
6	15.5	34.9	17.5	26.9	−3.8	8.2	3.9	5.8

Scenario	PREG2				PREG3			
	Subgroup							
	1	2	3	4	1	2	3	4
1	−0.4	0.3	0.5	−0.1	−0.3	−0.3	0.1	0.5
2	−0.0	−0.1	−0.4	−3.1	−0.1	−0.1	0.3	0.3
3	−0.4	0.6	1.5	0.5	−0.2	−1.2	−0.7	−0.2
4	−0.1	−0.3	−0.3	0.4	−0.8	1.5	0.7	1.7
5	0.2	0.6	1.2	1.8	0.1	−0.3	−0.4	0.0
6	−0.9	−0.1	−0.0	1.3	0.0	−1.3	−0.6	0.5

Census counts				
Scenario	Subgroup			
	1	2	3	4
1	−10.2	26.5	0.3	2.0
2	−9.1	29.0	0.4	−1.5
3	−20.7	53.7	−0.2	−0.6
4	−29.5	80.8	−0.5	−1.9
5	−21.3	60.7	0.85	1.4
6	−27.5	84.5	0.9	0.3

Islanders’ status. Interestingly, PREG 2 only has a small bias because ‘Aboriginal and Torres Strait Islanders’ and ‘Community’ are well-correlated. As expected, PREG3 is unbiased in all scenarios, as it correctly assumes that PES nonresponse is due to ‘community’.

5. Set-Up of the Simulation

Subsection 5.1 creates a synthetic version of the population, U , and simulates Census counting of the population. Subsection 5.2 simulates PES sampling from the synthetic population. Subsection 5.3 evaluates the GREG and PREG. The aim was for the simulation to be realistic.

5.1. Simulating the Population

Records of the 2001 Census define the synthetic population U and subgroup t . The different subgroup totals of interest, T , are given in [Tables 3 and 4](#) (e.g., in the last row of [Table 3](#), t is defined as the membership indicator for the Australian Capital City and T is the population total of the Australian Capital City).

Define \mathbf{z} to be a $H = 449$ vector of 2001 Census variables given by Aboriginal and Torres Strait Islanders status and the cross-classification of region, sex, and age. This defines \mathbf{z}_j for all $j \in U$. Given that t and \mathbf{z} are defined in terms of the same source of data (i.e., the 2001 Census), their values are consistent.

In the simulation, \mathbf{x} has the same categories as \mathbf{z} . However, to allow for errors in the Census counting, \mathbf{x} and \mathbf{z} can be different. To do this, we used the 2001 PES to model the following three probabilities, conditional on a range of dwelling and person-level covariates (including covariates in \mathbf{z}) from the 2001 Census:

- $p^{(1)}$ = the probability that a person was counted correctly by the Census (i.e., Census and PES classification is the same),
- $p^{(2)}$ = the probability that a person is misclassified by the Census (i.e., Census and PES classification is *not* the same), and
- $p^{(3)}$ = the probability that a person is missed (i.e., not counted) by the Census.

Table 3. Relative Bias (RB), Relative Root Mean Squared Error (RRMSE) and Relative Standard Error at State and National Level.

Subgroup	RB (%)		RSE (%)		RRMSE (%)	
	PREG	GREG	PREG	GREG	PREG	GREG
Australia	−0.03	−0.02	0.10	0.10	0.11	0.10
New South Wales	−0.03	−0.01	0.21	0.21	0.22	0.21
Victoria	−0.03	−0.02	0.21	0.20	0.21	0.20
Queensland	−0.03	−0.01	0.25	0.25	0.25	0.25
South Australia	−0.04	−0.04	0.28	0.27	0.28	0.28
Western Australia	−0.06	−0.05	0.30	0.29	0.30	0.30
Tasmania	−0.07	−0.07	0.40	0.39	0.41	0.40
Northern Territory	−0.11	−0.08	1.00	0.97	1.01	0.98
Australian Capital Territory	−0.04	−0.05	0.53	0.51	0.53	0.52

Table 4. Mean squared error and bias for other subgroups.

	MSE (relative to PR)		Bias (%)	
	PREG	GREG	PREG	GREG
Male, age 0–19	100	94	−0.04	0.00
Male, age 20–29	100	92	−0.03	0.24
Male, age 30–59	100	89	−0.03	0.05
Male, age 60+	100	93	−0.03	0.27
Female, age 0–19	100	97	−0.02	−0.10
Female, age 20–29	100	91	0.06	0.02
Female, age 30–59	100	104	−0.01	−0.17
Female, age 60+	100	91	0.04	0.06
Not Aboriginal and Torres Strait Islanders	100	111	−0.02	−0.07
Aboriginal and Torres Strait Islanders	100	139	0.08	2.80
Born in Australia	100	100	−0.01	0.00
Born overseas	100	99	0.00	0.02
Not married	100	100	−0.02	0.00
Married	100	97	−0.02	−0.01

Each person in the simulated population was assigned these three probabilities under the model, giving $(p_j^{(1)}, p_j^{(2)}, p_j^{(3)})$ for all $j \in U$. Then the value of \mathbf{x}_j for population record j was equal to:

- \mathbf{z}_j with probability $p_j^{(1)}$.
- \mathbf{z}_j^* with probability $p_j^{(2)}$, where \mathbf{z}_j^* is same as \mathbf{z}_j but changed to an ‘adjacent’ category in one dimension (e.g., if \mathbf{z}_j indicates a male aged 20–25, then \mathbf{z}_j^* may indicate a male aged 26–30, where the dimension that is changed is age and the ‘adjacent’ categories are 20–25 and 26–30).
- $\mathbf{0}$ with probability $p_j^{(3)}$
- $2\mathbf{z}_j$ with probability $1 - p_j^{(1)} - p_j^{(2)} - p_j^{(3)}$ (i.e., person j is counted twice).

Finally, all people in the simulated population were assigned a value for η_j , the probability that person j would respond to the PES. The important point here is that η_j was allowed to be a function of \mathbf{z} . It is worth noting here that Aboriginal and Torres Strait Islanders status had a strong influence on the propensity to respond. (The coefficients in the propensity model were obtained from the logistic regression of 2001 Census response propensity using the 2001 PES. One difference between the variables in the models used to obtain $p_j^{(3)}$ and η_j is that the latter included additional dwelling and person-level covariates, such as whether born outside of Australia, dwelling type, and marital status).

5.2. Simulating the PES Samples

Repeated PES samples of size 90,000 people were drawn from the synthetic population. The simulated PES sampling scheme was designed to mimic the actual PES sampling scheme. The first stage of this sampling scheme divides the Census Collector’s Districts

(CDs) into strata, and chooses a sample of these CDs with probability proportional to the number of dwellings in the CDs. The second stage divides the CDs into blocks, and selects a block at random. Finally, a cluster of dwellings is selected within each selected block by skipping through the list of dwellings. The skip lengths are such that each dwelling has an equal probability of selection in a state. Once the PES sample was selected from the synthetic population, each selected person was randomly assigned to be a PES nonrespondent with the probability $1-\eta_j$. The variable t was collected from PES respondents.

Given that the propensity to respond to the PES depends upon on \mathbf{z} (and not \mathbf{x}), nonresponse is nonignorable given \mathbf{x} . If the propensity to respond was based only on \mathbf{x} , nonresponse would be ignorable given \mathbf{x} . The nonresponse rate was simulated to be about 94% (of those who were contacted by the 2011 PES, 94% responded).

5.3. Evaluation of Alternative Estimators Using Simulation

The estimators in the simulation are:

- GREG: Generalised regression estimator, $\hat{T}^{(GREG)}$ where \mathbf{x} is defined in Subsection 5.1.
- PREG: The proposed regression estimator, $\hat{T}^{(PREG)}$ where \mathbf{x} and \mathbf{z} are defined in Subsection 5.1 This definition of \mathbf{x} and \mathbf{z} was used in estimation for the actual 2011 PES.

For each of 1,000 simulated PES samples, we calculated the

- Relative Bias (RB) $(\hat{T} - T)/T$
- Relative Standard Error (RSE) of $\hat{T} = \sqrt{\widehat{\text{Var}}(\hat{T})/T}$
- Relative Root Mean Squared Error (RRMSE) of $\hat{T} = \sqrt{\widehat{\text{Var}}(\hat{T} - T)/T}$

for the GREG and PREG estimators, where $\widehat{\text{Var}}(\hat{T})$ is calculated using the group jackknife. Table 3 gives the average RB, RSE and RRMSE over these 1,000 simulations. Table 3 shows that GREG and PREG perform equally well. However, Table 4 shows that PREG outperforms GREG for estimates of the Aboriginal and Torres Strait Islanders population. This is driven by the fact that the Aboriginal and Torres Strait Islanders population was simulated to have significant and nonignorable influence on PES nonresponse status. While the results are not shown here, the coverage rates of PREG were close to their nominal level of 95%.

6. Discussion

The development of the ABS's Census coverage estimation strategy for Australia has been driven by Census counting on a 'persons present' basis and PES counting on a 'usual residents' basis, Census misclassification, and by nonignorable nonresponse in the PES.

The estimator proposed here also has the potential to aid population estimation when using an imperfect administrative source as the basis, rather than a traditional Census. In such a situation, the estimation strategy has to deal with individuals having some characteristics poorly defined and being in the wrong locations in the administrative data, as well as having people completely missing from the administrative data.

This concept of a survey for quality assessment of administrative data has been proposed in [Brown and Honchar \(2012\)](#), where they suggest the ABS PREG estimator as an approach.

Completely erroneous returns was part of the reason for the E-sample, as adopted by the US Census. In part because there is no E-sample, the proposed estimator is biased if there are completely erroneous returns. More generally, it is biased if people who cannot be counted by the PES can be counted by the Census. Although there is no evidence of an erroneous returns problem in the Australian context, it does mean that temporary residents should be excluded (i.e., excluded from **X**). In the context of estimating people counts from an administrative list (instead of a Census), this means that individuals who have emigrated, but erroneously remain on the list, need to be removed using other approaches, such as evidence of activity within the system prior to estimation.

As discussed, the PREG relies on the same assumptions as classic DSE, including independence between the Census and PES. [Gerritse et al. \(2015\)](#) explicitly explored issues of dependence with two lists. More generally, within the classic capture-recapture framework work has been done exploring the use of multiple systems to tackle the issue of dependence. A recent review is given by [Baffour et al. \(2013\)](#); while [Zhang \(2015\)](#) makes an interesting contribution to using multiple administrative sources with a survey to deal with completely erroneous returns. Clearly, looking at how the ABS PREG can fit into this multiple system situation is important future work, as many countries now using a traditional Census are looking to use multiple administrative sources as an alternative.

7. References

- Alho, J.M., M.H. Mulry, K. Wurdeman, and J. Kim. 1993. "Estimating Heterogeneity in the Probabilities of Enumeration for Dual-System Estimation." *Journal of the American Statistical Association* 88: 1130–1136.
- Baffour, B., J.J. Brown, and P.W.F. Smith. 2013. "An Investigation of Triple System Estimators in Censuses." *Statistical Journal of the IAOS* 1: 53–68. Doi: <http://dx.doi.org/10.3233/SJI-130760>.
- Belin, T.R. and J.E. Rolph. 1994. "Can We Reach Consensus on Census Adjustment?" *Statistical Science* 9: 486–508. Doi: <http://dx.doi.org/10.1214/ss/1177010261>.
- Bell, W. 2001. *ESCAP II: Estimation of Correlation Bias in 2000 A.C.E. Estimates Using Revised Demographic Analysis Results*. Report No. 10 to the Executive Steering Committee for A.C.E. Policy II, U.S. Bureau of the Census.
- Brown, J., O. Abbott, and I. Diamond. 2006. "Dependence in the 2001 One-Number Census Project." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 169: 883–902. Available at: URL: <http://www.jstor.org/stable/3877405>.
- Brown, J.J., I.D. Diamond, R.L. Chambers, L.J. Buckner, and A.D. Teague. 1999. "A Methodological Strategy for a One-Number Census in the UK." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 162: 247–267. Available at: URL: <http://www.jstor.org/stable/2680581>.

- Brown, J.J. and O. Honchar. 2012. "Design and Estimation of Surveys to Measure Data Quality Aspects of Administrative Data." *Lithuanian Journal of Statistics* 51: 5–16.
- Chang, T. and P.S. Kott. 2008. "Using Calibration Weighting to Adjust for Nonresponse Under a Plausible Model." *Biometrika* 95: 557–571. Available at: <http://www.jstor.org/stable/20441486>.
- Gerritse, S.C., P.G.M. van der Heijden, and B.F.M. Bakker. 2015. "Sensitivity of Population Size Estimation for Violating Parametric Assumptions in Log-linear Models." *Journal of Official Statistics* 31: 357–379. Doi: <http://dx.doi.org/10.1515/jos-2015-0022>.
- Griffin, R.A. 2000. "Accuracy and Coverage Evaluation: Dual System Estimation." DSSD Census 2000 Procedures and Operations Memorandum Series, Q-20, US Census Bureau.
- Griffin, R.A. and T. Mule. 2008. DSSD 2010 Census Coverage Measurement Memorandum #2010-E-20. Available at: http://www.census.gov/coverage_measurement/post_enumeration_surveys/2010_results.html (accessed December 2015).
- Hogan, H. 1993. "The 1990 Post Enumeration Survey: Operations and Results." *Journal of the American Statistical Association* 88: 1047–1060.
- Hogan, H. 2003. "The Accuracy and Coverage Evaluation: Theory and Design." *Survey Methodology* 29: 129–138.
- Kott, P.S. and T. Chang. 2010. "Using Calibration Weighting to Adjust for Nonignorable Unit Nonresponse." *Journal of the American Statistical Association* 105: 1265–1275. Doi: <http://dx.doi.org/10.1198/jasa.2010.tm09016>.
- Large, A., J.J. Brown, O. Abbott, and A. Taylor. 2011. "Estimating and Correcting for Over-count in the 2011 Census." *Survey Methodology Bulletin* 69: 35–48.
- Mule, T. 2008. DSSD 2010 Census Coverage Measurement Memorandum #2010-E-18. Available at: https://www.census.gov/coverage_measurement/pdfs/2010-E-18.pdf (accessed January 2017).
- Nordholt, E.S. 2005. "The Dutch Virtual Census 2001: A New Approach by Combining Different Sources." *Statistical Journal of the United Nations Economic Commission for Europe* 22: 25–37.
- Rubin, D.B. and R.J.A. Little. 2002. *Statistical Analysis of Missing Data (2nd Edition)*. New Jersey: John Wiley and Sons.
- Särndal, C.-E., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. Berlin: Springer Verlag.
- Sekar, C.C. and W.E. Deming. 1949. "On a Method of Estimating Birth and Death Rates and the Extent of Registration." *Journal of the American Statistical Association* 44: 101–115. Doi: <http://dx.doi.org/10.2307/2280353>.
- Valente, P. 2010. "Census Taking in Europe: How Are Populations Counted in 2010?" *Population & Societies* 467: 1–4.
- Wachter, K.W. and D.A. Freedman. 2000. "The Fifth Cell: Correlation Bias in U.S. Census Adjustment." *Evaluation Review* 24: 191–211. Doi: <http://dx.doi.org/10.1177/0193841X0002400202>.
- Wolter, K.M. 1990. "Capture–Recapture Estimation in the Presence of a Known Sex Ratio." *Biometrics* 46: 157–162. Doi: <http://dx.doi.org/10.2307/2531638>.

- Xi Chien, S. and C.Y. Tang. 2011. “Properties of Census Dual System Population Size Estimators.” *International Statistical Review* 79: 336–361. Doi: <http://dx.doi.org/10.1111/j.1751-5823.2011.00150.x>.
- Zhang, Li-Chun. 2015. “On Modelling Register Coverage Errors.” *Journal of Official Statistics* 31: 381–396. Doi: <http://dx.doi.org/10.1515/jos-2015-0023>.

Received February 2015

Revised September 2016

Accepted October 2016