Australian Bureau of Statistics

1352.0.55.106

Research Paper

Small Area Estimation with Simulated Samples from the Population Census

New Issue

Research Paper

Small Area Estimation with Simulated Samples from the Population Census

Chris Davies, Daniel Elazar and Noel Hansen

Analytical Services Branch

Methodology Advisory Committee

6 November 2009, Canberra

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) MON 18 JUL 2011

ABS Catalogue no. 1352.0.55.106

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Mr Daniel Elazar, Analytical Services Branch on Canberra (02) 6252 6962 or email <analytical.services@abs.gov.au>.

SMALL AREA ESTIMATION WITH SIMULATED SAMPLES FROM THE POPULATION CENSUS

Chris Davies, Daniel Elazar and Noel Hansen
Analytical Services

QUESTIONS FOR THE COMMITTEE

1. Does MAC broadly agree with the approach taken here, given the limitations mentioned?
2. Does MAC have any comments about improving the quality of small area estimates for indigenous / remote areas due to no sample and poor quality auxiliary information?
3. How does MAC suggest we best adjust for the observed “design informativeness” that we have referred to as parametric estimation bias?
4. Given how unreliable the SAEs and their estimated RRMSEs are for LGAs with small average sample sizes, which of the following options would MAC recommend to address this issue:
   - use a spatial SAE model approach,
   - not publish estimates for areas with small sample sizes,
   - put a spline smoother through the plot of RRMSEs by SAEs and use the predicted RRMSEs for release purposes, or
   - another possibility MAC could suggest?
5. What prospect does MAC think there is of designing ABS surveys to meet the needs of both survey publications and SAE, under current cost constraints?
6. To what extent would MAC expect gains to be realised from using a parametric bootstrap simulation approach, as opposed to this design-based simulation?

CONTENTS

SMALL AREA ESTIMATION WITH SIMULATED SAMPLES FROM THE POPULATION CENSUS

Chris Davies, Daniel Elazar and Noel Hansen
Analytical Services

ABSTRACT

Small area estimation involves producing estimates for small geographical regions for which direct survey estimates are statistically unreliable. This is achieved by constructing a model involving auxiliary variables as well as survey data and using it to predict for all units not surveyed. Analytical Services Branch has been researching and applying small area estimation (SAE) techniques since 2003, including recently evaluating small area estimates (SAEs) of labour force status at the local government area (LGA) level. The primary quality measures for SAEs are their estimated relative root mean squared errors (RRMSEs). This paper describes an investigation into the quality of the SAEs and estimated RRMSEs. This investigation concluded that the small area estimates of labour force status are generally of reasonable quality. Exceptions occur for local government areas with low average sample sizes due to being in remote parts of Australia or having small populations. The major cause of bias in the estimates is the difference between the parameter estimates in models fitted to the whole population and those in models fitted to samples, with bias due to the model choice being the secondary cause. RRMSE estimates are generally conservative but can greatly underestimate the mean squared error for some local government areas with small average sample sizes.

1. INTRODUCTION

Small area estimation involves producing estimates for small geographical regions for which direct survey estimates are statistically unreliable. The Australian Bureau of Statistics (ABS) designs surveys to produce estimates using direct survey estimators for large geographic regions with enough sample size for these estimators to be reliable. In recent years there has been a growing demand for various estimates to be produced in smaller geographical regions. The direct survey estimates for these small regions are considered to be too unreliable. One way to reliably estimate for these small regions is to produce model-based estimates which borrow strength from administrative and Census data and other types of auxiliary variables. A well fitting and parsimonious model is fitted to the survey data and is used to predict the response for all units not surveyed. These model-based estimators may produce estimates with lower error than the direct survey estimates.

The Analytical Services Branch of the ABS has been researching and applying small area estimation (SAE) techniques since 2003. Standard SAE methodologies as covered by Rao (2003) and specifically Saei and Chambers (2005) have been used. Over this time, a wealth of technical knowledge and experience with SAE applications has been accumulated. Applications have included:

- experimental local government area (LGA) level disability estimates for National Disability Administrators (now the Disability Policy & Research Working Group) using the Survey of Disability and Aged Carers 2003 and Census 2001;
- supporting other small area practitioners throughout the ABS, such as small area estimates (SAEs) of Health for Australians and Indigenous people and SAEs of Water Usage.

Experimental SAE work previously conducted on labour force estimates used a generalised linear mixed model (GLMM) applied to the 'labour force status' variable of the Labour Force Survey (LFS). A range of quality measures, including the bias, coverage and additivity tests of Brown *et al.* (2001), are used to assess the quality of the SAEs produced, however the primary quality measures for SAEs are their estimated relative root mean square errors (RRMSEs).

In all the small area applications conducted so far at the ABS, there has been a strong desire to allay persistent concerns about the reliability and accuracy of the estimated SAEs. Internal stakeholders and users of small area output want reassurance that the quality measures and estimated measures of accuracy are reliable. Therefore, the principal motivation for this investigation was to evaluate estimates and quality measures used in these applications, assessing the extent of any bias which may be observed.

There are two recurring issues that previous SAE applications work has raised.

1. Most of our SAE applications have used LGA as the small area. Many LGAs, especially those in the more remote areas, have very small or no sample. These LGAs tend to have the most volatile estimates of RRMSE; this is supported by the small area literature (see for example, Saei and Chambers (2005)). Options that have been suggested to deal with this problem include:
    - the use of spatial modelling approaches
    - consider only publishing SAEs for areas with sufficiently large sample size.
2. There has been the issue of design informativeness; that is, to what extent are our SAEs biased because we have not taken full account of the survey design in our small area model estimation.

Finding workable solutions to these issues involves significant effort. In order to justify the expenditure of resources, the impact on SAEs needs to be assessed.

This paper describes an investigation into the quality of the SAEs and their estimated RRMSEs. Under normal circumstances the SAEs cannot be compared with the true population quantities as this requires a census to have been undertaken. However the ABS Census of Population and Housing (Census) collects information about the labour force status of individuals for the entire population on Census night. Therefore, in this study we were able to repeatedly sample from the Census and compare SAEs from these repeated samples with the known population values, to investigate:

- the differences between the true population values and the average estimates based on repeated samples, and
- the variability of the SAEs calculated from repeated samples drawn from Census data, under the Labour Force Survey sample design.

This study also compares the estimated model-based RRMSEs with the design-based RRMSE of the estimates. A similar study by Bleuer *et al.* (2007) estimated design-based RRMSEs from Monte Carlo simulations. They found the model-based analytic RRMSE estimates did not follow the RRMSE estimates calculated from design-based simulations.

In this study, a design-based simulation, using Census labour force data as a proxy for labour force status as collected under the LFS, has been used to understand how the established labour force estimation methodology performs, using the models chosen and evaluated using LFS data[1].

1 Parametric bootstrap is an alternative approach where a superpopulation model approach is used. This has implications for the way design informativeness has been evaluated and the accuracy of the estimated RRMSEs.

The objective of this simulation study is to identify where differences exist between the true population value and:

- the estimate based on a model fitted to the whole census, and
- the average estimate based on repeated samples.

These differences can indicate the level of design informativeness as well as errors due to the model fitted being unable to capture the variation in the response based on the covariates available. Although considerable work has been done in SAE, particularly with the use of parametric bootstrap simulations to evaluate the properties of estimators (Hall and Maiti, 2006), little work has been done using a design-based simulation involving real data and a sample design involving out-of-sample areas.

This investigation found that the current experimental small area estimates of labour force status are generally of reasonable quality. Exceptions occur for local government areas with low average sample sizes due to being in remote parts of Australia or having small populations. The major cause of bias in the estimates is the difference between the parameter estimates in models fitted to the whole population and those in models fitted to samples, with bias due to the model choice being the secondary cause. The bias in the parameter estimates was worst for remote and very remote areas, and fitting a separate model for these areas reduced the bias but did not remove it. RRMSE estimates are generally conservative but can greatly underestimate the mean squared error for some local government areas with small average sample sizes.

The remaining sections of this paper are as follows:

- Section 2 of the paper details the methodology used, including:
  - the current SAE methodology,
  - the methodologies used to assess the quality of the SAE model, the SAEs themselves and their estimated RRMSEs, and
  - the limitations of these methodologies;
- Section 3 describes the results of quality assessments for
  - the model,
  - the estimates, and
  - the estimated RRMSEs;
- Section 4 describes further work that could be completed following on from this investigation; and
- Section 5 concludes the paper.

2. METHODOLOGY

2.1 Sample design simulation and estimation methodology

The following sections describe the sample selection methodology for this investigation, the current SAE methodology used, including the model for predicting labour force status, and also issues with the data used.

2.1.1 Sample design simulation methodology

The purpose of this project was to assess the quality of SAEs and their associated RRMSEs derived from a generalised linear mixed model (GLMM), when applied to LFS data. This was achieved by simulating 1000 samples of the LFS from the known population of the 2001 Census of Population and Housing and applying a GLMM to each sample to produce a set of SAEs and RRMSEs for each sample. The design of the samples closely mimicked the actual design of the LFS to make the simulation as realistic as possible. The distribution of the SAEs across samples was then compared with the known values from the population. This assessment included analysing the accuracy as well as the precision of the SAEs. The distribution of the estimated RRMSE across samples was also compared with the “design-based RRMSE” we obtain from the SAEs, as defined in equation (2) in Section 2.2.2. Once again this included investigating the accuracy and the precision of the model RRMSEs when compared with the design-based RRMSE of the SAEs. The details of the assessment methodologies can be found in Section 2.2.

This design-based simulation assumes the population values of the Census are fixed, and that the variation between samples is a result of the different samples taken. Alternatively we could have used a parametric bootstrap approach where a superpopulation model is assumed. The approach taken was partly determined by the limited resources available, but also by doubts regarding the benefits a superpopulation model would provide above the design-based approach.

One thousand samples were selected from the Census under the LFS design, which is a multi-stage clustered design that selects a sample of about 0.24% of the population of Australia. A geographical frame of Census collection districts (CD) is used, with selected CDs divided into blocks. Selected blocks are then divided into clusters of dwellings, with the dwellings within a cluster being a systematic sample throughout the block. The LFS has a self-representing sample design such that one in every k_s dwellings is chosen in state / territory s, that is, the sampling fraction for state s is $\frac{1}{k_s}$. k_s is known as the state skip for state s.

A sample selection program used to undertake the variance modelling for the 2006 Monthly Population Survey (MPS) redesign was used to select the 1000 samples. More information about the LFS can be found in the LFS Sample Design documentation (ABS, 2002). Note that this is not the most current LFS Sample Design document, however it details the relevant design which was used in the variance modelling sample selection program for the 2006 MPS redesign.

For each of the 1000 samples, the following process was then followed. One of the k_s possible samples was randomly chosen from each state / territory, with equal probability, and these samples were combined to form a whole sample for Australia.

2.1.2 SAE methodology

The following logistic random effects regression model was used to predict each of the three labour force statuses (employed, unemployed and not in the labour force (NILF)):

For area d, and age–sex class c,

$$y_{c(d)} \sim Bin\left(n_{c(d)}, p_{c(d)}\right)$$

$$\log\left(\frac{p_{c(d)}}{1-p_{c(d)}}\right)=\beta_0+\beta_1 x_{c(d)1}+\ldots+\beta_p x_{c(d)p}+u_d$$

and $$u_d \sim N(0,\phi)$$

where

d = 1, …, 644 are local government areas (LGA) described below, and

c = 1, ..., 10 are age–sex classes for the age groups 15–24, 25–34, 35–44, 45–54 and 55–64 years for each sex.

Also

$y_{c(d)}$ is the observed labour force status count of $n_{c(d)}$ sampled persons within age–sex class c of area d;

$p_{c(d)}$ is the probability of having the particular labour force status within age–sex class c of area d;

$x_{c(d)1}, \ldots, x_{c(d)p}$ are the explanatory variables chosen by a stepwise selection algorithm for each separate labour force status;

$\beta_0, \ldots, \beta_p$ are the fixed effects of the intercept and the coefficients of the explanatory variables; and

u_d is the random effect for area d with variance ϕ across areas.

Each of the models included the explanatory variables state, remoteness (based on the 2001 Australian Standard Geographical Classification (ASGC) remoteness classification) in three groups, age–sex class in ten groups, and full payment social welfare benefits from Centrelink in the ten age–sex groups. The specific models for each of the labour force statuses also included these variables:

- For employed: Socio-economic index for areas (SEIFA) in four groups, household type in five groups and unemployment benefits, meaning Newstart Allowance or Youth Allowance (Other) payments.
- For unemployed: Unemployment benefits as well as an interaction between these benefits and remoteness.
- For NILF: SEIFA in four groups and household type in five groups.

For a full description of the explanatory variables see Appendix A. These models were selected using a stepwise selection algorithm with *SAS PROC LOGISTIC*, when applied to LFS data from August 2006. For each of the models a group of coefficients, such as the group of coefficients corresponding to the covariate state, was included in the model if at least one of the coefficients within the group was significant. Specific interaction effects were chosen as candidates for the model, based on assumptions about which were most likely to be significant, however there was not a comprehensive check of all interaction effects.

A different model was subsequently used to predict unemployed for LGAs in remote and very remote areas (remoteness classification 3), for reasons described in Section 3.2.1, and was only used for producing figure 3.18 in that section. To determine which covariates to include in the model for remoteness classification 3 LGAs, a stepwise selection procedure was run using the step function in R, which is based on reducing Akaike's (1974) 'An Information Criterion' (AIC) at each step. Step can be used in conjunction with most model packages and in our case was used for our GLMM. For estimating the GLMMs, maximum likelihood and numerical integration via Gauss–Hermite quadrature was used in the R package 'glmmML' as opposed to the penalised quasi-likelihood with restricted maximum likelihood used for the rest of the estimation in this paper. The space of covariates the model could possibly choose included the main effects of state, age in five groups, sex, SEIFA in four groups, household type in five groups, proportion of indigenous, unemployment benefits and full payment social welfare benefits from Centrelink. Also included in the space of covariates was the interaction effect between age and sex, as well as all other two-way interactions between all of the effects mentioned so far, including the interaction effect of age with sex. However some of these effects could not be included due to the co-linearity they possess when applied to data for remoteness classification 3 LGAs.

A parsimonious model was chosen for the remote classification 3 areas by using the covariates which provided the greatest reduction in AIC without increasing the number of parameters too greatly. This model contained the following covariates: age, sex, state, proportion indigenous, unemployment benefits as well as the interactions between age and unemployment benefits, proportion indigenous and unemployment benefits, proportion indigenous and sex, and, state and unemployment benefits. However some of the parameters within these effects had to be collapsed with other parameters due to the co-linearity they possessed when applied to the small sample sizes of remote LGAs. For a full description of the explanatory variables in this model for remote classification 3 areas see Appendix A.

Maximum Penalised Quasi-Likelihood (MPQL) with Restricted Maximum Likelihood was used to estimate the model parameters, as was done by Saei and Chambers (2003).

In the case that $n_{c(d)}=0$ for $c = 1,...,10$, in a particular area d, the estimate of the random effect, $\hat{u}_d$, is defined to be zero.

Once the parameters have been estimated the small area estimator of labour force status count in area d, $\hat{\theta}_d$, is constructed as follows:

$$\hat{\theta}_d=\sum_{c=1}^{10}\left(y_{c(d)}+\hat{p}_{c(d)}\left(N_{c(d)}-n_{c(d)}\right)\right)$$

where $\hat{p}_{c(d)}$ is the estimate of $p_{c(d)}$ and $N_{c(d)}$ is the known population size. In the case that $n_{c(d)}=0$ within age–sex class c of area d, the observed labour force status count $y_{c(d)}=0$.

These models have been validated in a number of ways including by checking that the coefficient estimates make sense in terms of size and direction, adjusting for multi-collinearity in the covariates, checking the residuals for influential points and for overdispersion as well as using other goodness-of-fit tests. These validations are covered in internal reports which are available on request.

As the estimates of employed, unemployed and NILF are estimated independently, additivity to the population size $N_{c(d)}$ is not guaranteed. However, we have found from previous work that the sums of the three estimates are generally close to the respective population sizes. A multinomial model for producing SAEs of labour force, that ensures coherence, has also been investigated and is described in Scealy (2010).

Also calculated is a Saei–Chambers (2003) model-based RRMSE estimate, $\widehat{RRMSE}_d$, for each small area estimate for area d. This RRMSE estimator takes account of the errors in estimating the model fixed effects, the area level effects, the variation in the response variable and additionally the error in estimating the variance component parameter, ϕ.

The RRMSE estimates cannot be said to unconditionally account for model misspecification, so the level of faith one has in the RRMSE estimates is, among other things, predicated upon the wise choice of the most appropriate fitting model. In addition, it is known that the MPQL estimation approach, like approximate ML for GLMMs, is known to seriously underestimate the variance component ϕ when the true value is large, i.e. when the fixed effects in the model explain only a small proportion of the variation between areas (Pawitan, 2001). Stability in the small area estimates and their associated estimates of MSE can also be diminished when the number of small areas is small and hence the variance component cannot be estimated with sufficient precision.

This estimation methodology was applied to each of the 1000 samples obtained by the methods described in Section 2.1.1, While the parameters used in each model are the same for all samples, their values are re-estimated each time the models are applied to a sample.

2.1.3 Data issues

Due to differences in the questions asked and the mode of collection, the Census labour force variable is different to the labour force status variable collected from the LFS, which adheres more closely to the International Labour Organisation's concepts and definitions. Despite this, we believe that the properties of the GLMM-derived SAEs and RRMSEs based on Census labour force status will provide useful information about the accuracy of the SAEs and RRMSEs of labour force status collected from the LFS. This is because the labour force status definitions of the two collections are similar. Moreover the estimates of the model coefficients based on Census data and LFS data are generally of the same sign.

From the Census data, people aged between 15 and 64 were in scope and international visitors were excluded, as is done in the LFS.

If a person's labour force status was not stated on the census (3.16% of the in-scope population), a single multinomial observation was placed into that LGA by age–sex category with probabilities based on the proportions of the three statuses in that category.

LGA boundaries were defined according to the Australian Standard Geographical Classification (ASGC) 2001. Exceptions were made for the ACT and Brisbane. The ACT was divided into eight statistical subdivisions (SSDs) and Brisbane was divided into nine statistical region sectors (SRSs) following the definitions in the ASGC 2001. These two areas were divided because their large population sizes made them too influential in the model estimation process. To make referencing easy, the eight SSDs in ACT and nine SRSs in Brisbane will be referred to as separate LGAs in the remainder of this paper.

CD of usual residence was used to assign the Census unit records to LGAs. However, samples were selected based on CD of Census enumeration. Thus in sample' and 'out-of-sample' LGAs were defined using CD of Census enumeration not CD of usual residence. Therefore sometimes an 'out-of-sample' LGA had small amounts of sample selected from it, because some people who usually reside there were selected in an 'in sample' LGA, where they were enumerated in the Census. A difference between the CD of Census enumeration and an adequately described CD of usual residence occurred for a relatively small number of units (2.65% of the in-scope population) so the effect of this is expected to be minor.

If CD of usual residence was not stated (0.08% of the in-scope population), or was inadequately described (0.65% of the in-scope population), the CD of enumeration was used instead to assign the units to LGAs.

The Census population totals of the LGA by age–sex classes were used, rather than the Estimated Resident Population (ERP) values which would normally be used for the LFS, as we were treating the Census data as the entire population. In doing so, we ignored the net undercount of the 2001 Census of 1.8% of the Australian population (ABS, 2003).

2.2 Quality assessment methodology

The following sections describe the methods used to assess the quality of the estimated model parameters, the small area count estimates and the RRMSE estimates.

2.2.1 Model parameters

To compare the estimated coefficients from the sample models with the corresponding Census model coefficient estimate, we calculated their coverage proportion. This was done using the confidence intervals for the 1000 estimates of the coefficients obtained from sample models and those for the corresponding estimates of the coefficients from the Census model. Approximately 95% of the sample “95% confidence intervals” should overlap with the Census “95% confidence interval”, with departure from this possibly indicating the presence of bias[2] in the coefficient estimates for the sample models. The coverage proportion was calculated as the number of sample “95% confidence intervals” overlapping with the Census “95% confidence interval”, including using the coverage adjustment for multiple confidence intervals from Brown *et al.*'s (2001) paper on SAE quality diagnostics. This coverage adjustment was necessary to ensure a nominal 95 percent overlap, as the

2 Bias refers to any difference between the centre of the distribution of the sample model estimates and the Census model estimate. This includes possibly measuring the centre of the distribution with the median, and the fact that the Census coefficient estimate may not necessarily be the true value of the sample coefficient estimates.

degree of overlap between two independent 95% confidence intervals for the same quantity will be higher than 95%.

The quality of the estimates of random effects, $\hat{u}_d$, was determined by their mean error (ME) and standard deviation (SD), given by:

$$ME\_\hat{u}_d = \bar{\hat{u}}_d^{(.)} - u_d$$

and
$$SD\_\hat{u}_d = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}\left(\hat{u}_d^{(i)} - \bar{\hat{u}}_d^{(.)}\right)^2},$$

where

$\hat{u}_d^{(i)}$ is the random effect estimate for area d from the model fitted to sample i,

$\bar{\hat{u}}_d^{(.)} = \frac{1}{1000}\sum_{i=1}^{1000}\hat{u}_d^{(i)}$ is the average random effect estimate across the samples, and

u_d is the random effect estimate from the model fitted to the Census population.

Note that the u_d are not necessarily the true random effect values; however, under this design-based simulation they are appropriate target values for the sample model estimates $\hat{u}_d^{(i)}$. The mean error and standard deviation can be combined into a root mean square error (RMSE) of the random effect estimates, calculated as follows:

$$RMSE\_\hat{u}_d = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}\left(\hat{u}_d^{(i)} - u_d\right)^2}.$$

A relative measure was not necessary as the random effects are approximately normally distributed about zero. The root MSE is related to the mean error and the SD through the formula:

$$\left(RMSE\_\hat{u}_d\right)^2 = \left(ME\_\hat{u}_d\right)^2 + \left(SD\_\hat{u}_d\right)^2. \qquad (1)$$

2.2.2 Small area estimates of count

To assess the quality of the small area estimator of labour force status count in area d, $\hat{\theta}_d$, we considered their relative mean error (RME) and relative standard deviation (RSD) which are given by:

$$RME\_\hat{\theta}_d = \frac{\bar{\hat{\theta}}_d^{(.)} - \theta_d}{\theta_d},$$

and

$$RSD\_\hat{\theta}_d = \frac{\sqrt{\frac{1}{1000}\sum_{i=1}^{1000}\left(\hat{\theta}_d^{(i)} - \bar{\hat{\theta}}_d^{(.)}\right)^2}}{\theta_d},$$

where

$\hat{\theta}_d^{(i)}$ is the small area estimate of labour force status count in area d from sample i,

$\bar{\hat{\theta}}_d^{(.)} = \frac{1}{1000}\sum_{i=1}^{1000}\hat{\theta}_d^{(i)}$ is the average of the sample estimates, and

θ_d is the true labour force status count in area d from the Census population.

It was necessary to use relative measures of the standard deviation and mean error because there is great variation in the population sizes of LGAs.

These quantities can be combined into the key quantity used to assess the quality of $\hat{\theta}_d$, the design-based relative root mean squared error (RRMSE) of the sample estimates. This quality measure is given by:

$$RRMSE\_\hat{\theta}_d = \frac{\sqrt{\frac{1}{1000}\sum_{i=1}^{1000}\left(\hat{\theta}_d^{(i)} - \theta_d\right)^2}}{\theta_d}. \tag{2}$$

From now on when RRMSEs are referred to, except where specified, we are referring to these design-based RRMSEs as opposed to model-based estimates of RRMSEs. The RRMSE is related to the relative standard deviation and the relative mean error through the formula:

$$\left(RRMSE\_\hat{\theta}_d\right)^2 = \left(RME\_\hat{\theta}_d\right)^2 + \left(RSD\_\hat{\theta}_d\right)^2. \tag{3}$$

As is described in Section 3.2.1 we were interested in decomposing the relative mean error into error due to technical bias and error due to parametric estimation bias. Here, technical bias refers to a difference between the expected prediction of the model fitted to the population and the true population value. That is, technical bias is the bias resulting from the model fitted being unable to capture the variation in the response based on the covariates available, possibly due to non-linearity in the relationship with the covariates. This is different to error due to model misspecification, as we do not know the true model. In this case, the population is the Census data and we measured technical bias with the relative technical bias (RTB) of $\hat{\theta}_d$:

$$RTB\_\hat{\theta}_d = \frac{\hat{\theta}_d^C - \theta_d}{\theta_d} ,$$

where $\hat{\theta}_d^C$ is the estimate of θ_d solely based on the model fitted to the Census population. That is, the model fitted to the Census population is used to predict θ_d without any sample information about the labour force status response variable.

On the other hand, we define parametric estimation bias to be the difference between predictions obtained from models fitted and applied to the samples and those predictions obtained from models fitted to the Census population but applied to the samples. We are using the term parametric estimation bias rather than design informativeness as, once again we do not know the true population model. Here we measured it with the relative parametric estimation bias (RPEB) of $\hat{\theta}_d$:

$$RPEB\_\hat{\theta}_d = \frac{\bar{\hat{\theta}}_d^{(.)} - \hat{\theta}_d^C}{\theta_d} .$$

Using $\hat{\theta}_d^C$ was effectively the same as using the average of the estimates when the Census model is applied to the samples, as the sampling fraction of the LFS is very small. Therefore the model used basically determined the estimates produced. It was necessary again to use relative measures of the technical bias and the parametric estimation bias because of the variation in the population sizes of LGAs.

2.2.3 Model RRMSEs

To be able to assess the quality of the predicted Saei–Chambers (2003) model-based RRMSE estimate in sample *i* for area *d*, $\widehat{RRMSE}_d^{(i)}$, we defined $RRMSE_d = RRMSE\_\hat{\theta}_d$.

That is, we considered the design-based RRMSE calculated using the small area estimates, $RRMSE\_\hat{\theta}_d$ defined in (2), to be the target value of the model predicted RRMSE values, $RRMSE_d$. Even though the design-based RRMSE is not the true RRMSE of $\hat{\theta}_d$, which is given by

$$\frac{\sqrt{E\left[\left(\hat{\theta}_d^{(i)} - \theta_d\right)^2\right]}}{\theta_d},$$

we have assumed that as the design-based RRMSE is based on 1000 samples it is a reasonable approximation to the true RRMSE of $\hat{\theta}_d$.

We have also made the assumption that the design-based RRMSE of the SAEs is a reasonable target for the model-based RRMSEs, despite the interpretations of the design-based RRMSE and the model-based RRMSE being different. This assumption is reasonable as, in theory, the model-based RRMSE estimates should estimate the model expectation of the design-based RRMSE (Bleuer *et al.*, 2007).

As we did for the small area estimates, we considered a relative mean error, $RME\_\widehat{RRMSE}_d$ and a relative standard deviation, $RSD\_\widehat{RRMSE}_d$ of the model RRMSE values. These were calculated as follows:

$$RME\_\widehat{RRMSE}_d = \frac{\overline{\widehat{RRMSE}}_d^{(.)} - RRMSE_d}{RRMSE_d},$$

and

$$RSD\_\widehat{RRMSE}_d = \frac{\sqrt{\frac{1}{1000}\sum_{i=1}^{1000}\left(\widehat{RRMSE}_d^{(i)} - \overline{\widehat{RRMSE}}_d^{(.)}\right)^2}}{RRMSE_d},$$

where

$$\overline{\widehat{RRMSE}}_d^{(.)} = \frac{1}{1000}\sum_{i=1}^{1000}\widehat{RRMSE}_d^{(i)}.$$

We were again able to combine the relative mean error and the relative standard deviation of the model RRMSEs into the key quality measure, the RRMSE of the model predicted RRMSEs for each area *d*:

$$RRMSE\_\widehat{RRMSE}_d = \frac{\sqrt{\frac{1}{1000}\sum_{i=1}^{1000}\left(\widehat{RRMSE}_d^{(i)} - RRMSE_d\right)^2}}{RRMSE_d}.$$

2.3 Limitations of the methodology

- There are differences between the labour force status variable collected in the LFS and that collected in the Census, hence the conclusions from the simulation may not apply to SAEs of the LFS.
- There are some LFS design variables, such as area type, a classification of areas with 15 levels from inner city Sydney/Melbourne to sparse and indigenous areas, which were not included in the models used as it could lead to over-parameterization.
- There are possibly some explanatory variables, which explain a large amount of the variation in the response variables, that were not available. It is also possible that the administrative covariates may not have been of the best quality. For example the unemployment benefit counts included those Indigenous Australians who are employed on Community Development Employment Projects.
- The original model selection for the three models did not include all possible two way interactions, and therefore some important interactions may be missing.
- The random effects, u_d, were assumed to be independent between areas.
- The LFS design cannot be perfectly emulated without the geographic information used to divide CDs into blocks and then into clusters.
- This was a design-based simulation which assumed the population was fixed, which was different to the model-based assumptions of the SAE predictions.
- The design-based RRMSE of the small area estimates was assumed to be a reasonable target value for the model-based RRMSE estimates.
- Adjustments were made for those not stating their labour force status or their CD of usual residence on the Census.

3. RESULTS

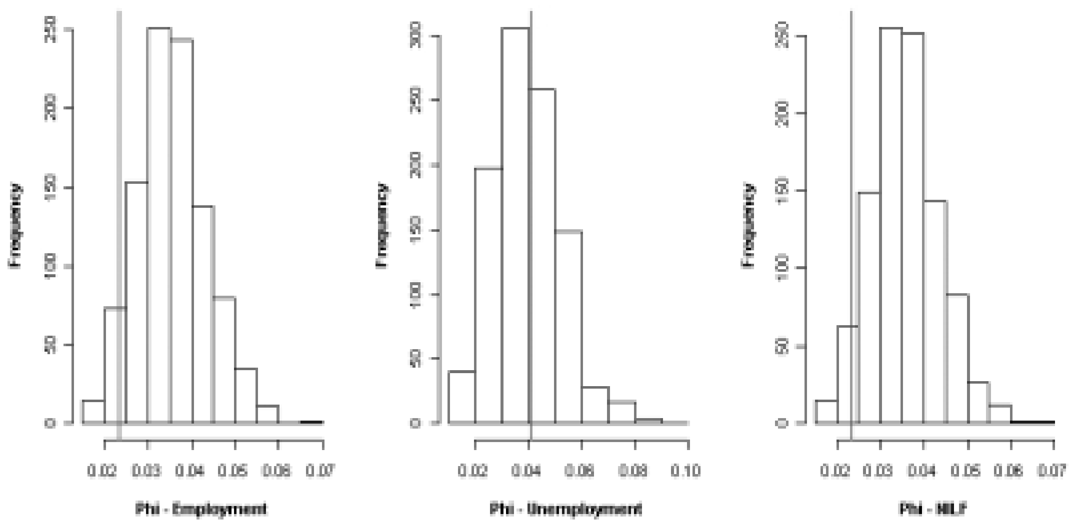
3.1 Model parameters

The following sections describe the results of the assessments of the parameters of the GLMM, including the random effect variance phi, the model coefficient estimates and the random effect estimates.

3.1.1 Random effect variance – Phi

The sample estimates of phi for employed and NILF had medians of 0.0352 and 0.0353, which were both greater than the Census model estimates of 0.0235 and 0.0233 respectively. Whereas for unemployed the simulations had a median of 0.0387 that was less than the Census model estimate of 0.0407. This bias can be seen in figure 3.1.

3.1 Histograms of the phi values for the 1000 samples for each of the three labour force statuses. The Census phi value for the model is shown as a vertical line.



3.1.2 Model coefficients

As described in Section 2.1.1 the models were selected on the basis of applying them to LFS data from August 2006. When these models were applied to the Census data, all of the covariates remained significant, including those categorical variables corresponding to a group of coefficients.

The majority of the coefficients from the sample models were similar to the Census model value. We calculated the coverage proportion of the Census model value by the sample values as described in Section 2.2.1 above. The coverage proportions for all coefficients and all models can be found in Appendix B.

Table 3.2 shows the number of coefficients with coverage proportion of the Census value less than 0.9, the number of coefficients with coverage proportion less than 0.8, and the total number of coefficients.

3.2 Number of model coefficients with various levels of coverage for each of the models

Model	*Model coefficients with coverage less than 90%*	*Model coefficients with coverage less than 80%*	*Total number of model coefficients*
Employed	6	1	37
Unemployed	8	2	32
NILF	7	5	36
Unemployed in remoteness classification 3	0	0	19

The two unemployed model coefficients with coverage less than 80% were remoteness classification 3 and the interaction effect between remoteness classification 3 and unemployment benefits. This bias of coefficients specific for remoteness classification 3 LGAs is part of the cause of the positive median of the relative parametric estimation bias for remoteness classification 3 areas described in Section 3.2.1. As mentioned above, a different model was subsequently used to predict unemployed for LGAs in remote and very remote areas, for reasons described in Section 3.2.1. The new unemployed model fitted to just the remoteness classification 3 areas had no coefficients with coverage less than 90%, and the two covariates with lowest coverage of their coefficients both involved the proportion indigenous covariate. For a detailed description of model coefficients with the lowest coverage for each of the models see Appendix B.

The bias in these coefficients, for the original models and the unemployed model fitted to remoteness classification 3 areas, suggests there are some design parameters not being accounted for in the model. For example, area type, which gives the degree of clustering in the multi-stage model, is not included in the models. This may be the cause of the difference in the estimates of coefficients for the samples and the estimate for the entire census population. When only the remoteness classification 3 areas were considered, for predicting unemployed, large undercoverage was only observed for coefficients involving the proportion indigenous covariate. This indicates that there is possibly something about the design of the LFS samples, in remote and very remote areas, which resulted in estimates of coefficients involving the proportion indigenous covariate being different to the estimate for the entire census population. This is possibly caused by either of the two reasons mentioned above.

3.1.3 Random effects – u_d

The estimated random effects, $\hat{u}_d$, allow the different small areas being modelled to have different intercepts. Our current methodology assumes the random effects are independent. However this assumption may not be reasonable, as neighbouring LGAs are more likely to be similar than distant LGAs. To investigate the spatial relationship between the random effects, we can plot them on a map of the LGAs in Australia. Figure 3.3 shows the random effects for the unemployed model when applied to the Census data. The LGAs are shaded based on the quintile of their random effect.

3.3 Unemployed model random effects when applied to the entire Census population

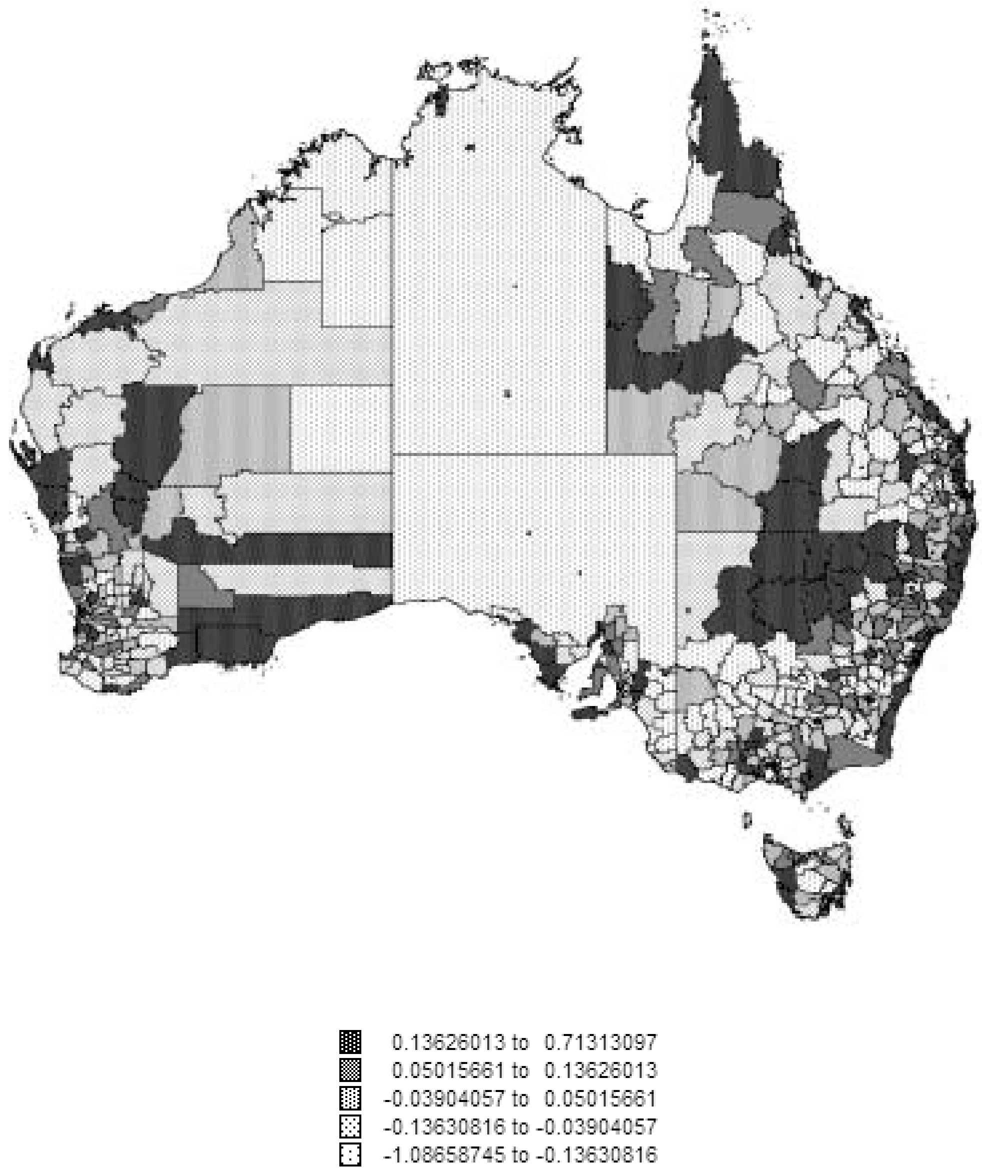


Figure 3.3 has been done with the shrunken random effect estimates rather than with the unshrunken estimates due to the time constraints related to the production of this paper. Although it might appear from this plot that the random effects are clustered, further work carried out at the ABS showed there was no significant spatial autocorrelation in the random effects.

As described in Section 2.2.1, to determine the quality of the estimates of the random effects, $\hat{u}_d$, we considered their root MSE as well as their mean error and standard deviation (SD). Figure 3.4 shows the root MSE of the employed random effects against the average sample sizes, $\bar{n}_d^{(.)} = \frac{1}{1000}\sum_{i=1}^{1000} n_d^{(i)}$, where $n_d^{(i)}$ is the sample size from area d in sample i. From figure 3.4 we observe that the random effect estimates for employed were quite reliable for LGAs with average sample sizes greater than 100, where the root MSEs were less than 0.25, and decreased as the average sample size increased. The LGA 'Unincorporated NT' was the single outlier with a large root MSE despite its large average sample size of 305. Although the random effect estimates for employed were quite reliable for LGAs with average sample sizes greater than 100, the estimates for LGAs with small average sample sizes could be unreliable, with some LGAs having root MSEs over 0.5. This was similarly the case for unemployed and NILF random effects estimates.

3.4 Root MSEs of random effects of employed against average sample size

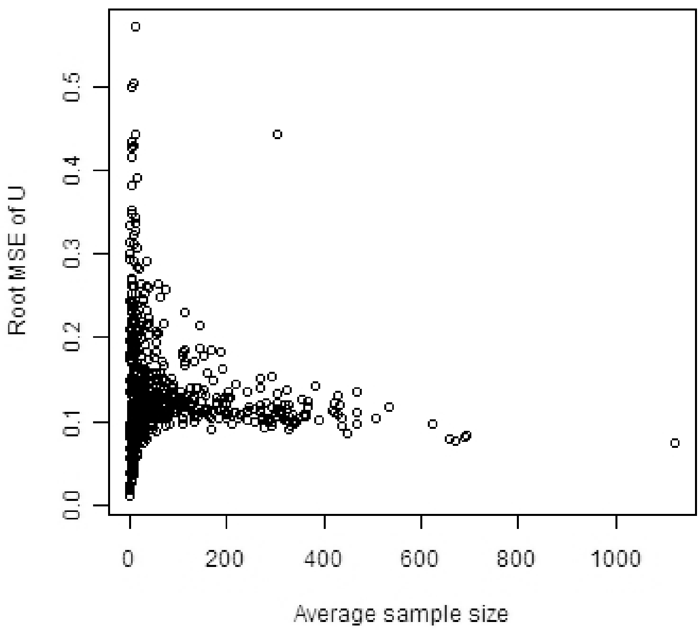
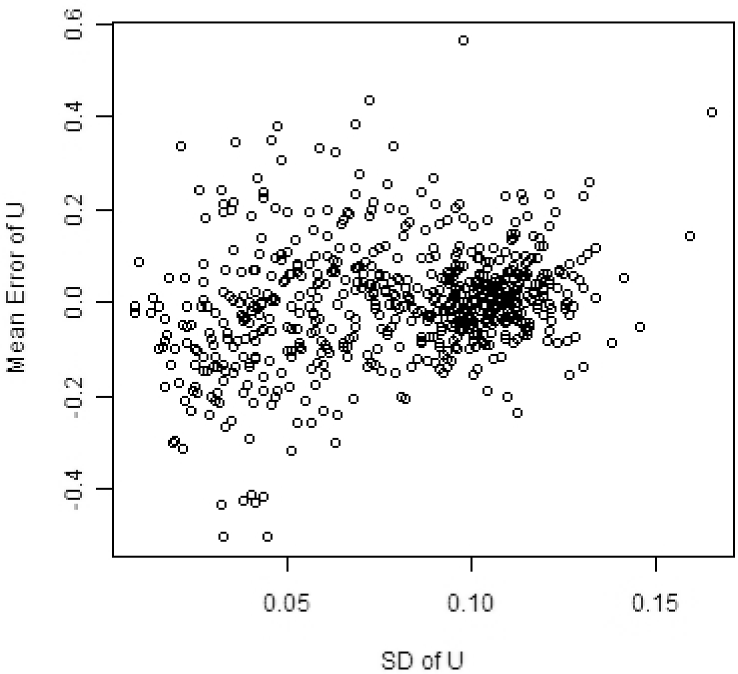


Figure 3.5 shows the mean error against the SD of the employed random effects. From figure 3.5 we can see that the mean error was more variable than the SD, and the mean error obtained values larger in absolute value than the SD did. Therefore due to the relationship between the mean error, the SD and the root MSE, given in (1) , the largest root MSE values of the LGAs with small average sample sizes were caused by LGAs having large mean errors of their random effect. A large number of LGAs had random effects with SDs around 0.1, which resulted in the root MSE being above this level for all LGAs except those with very small average sample sizes. This was again similar for unemployed and NILF random effects estimates.

3.5 Mean error against standard deviation of random effects of employed



3.2 Small area estimates

The following sections describe the results of the assessments of the SAEs and their estimated RRMSEs.

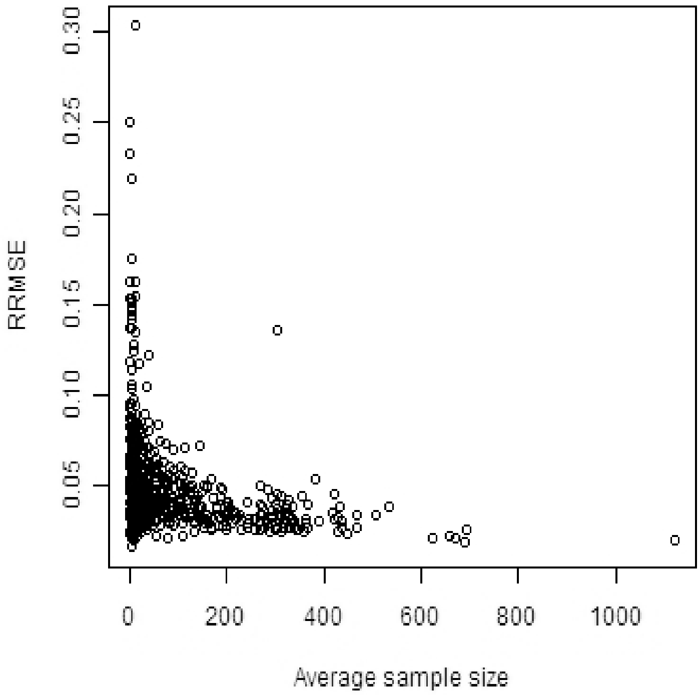
3.2.1 Small area count results

As described in Section 2.2.2, to assess the quality of the small area estimator of labour force status count in area d, $\hat{\theta}_d$, we considered the RRMSE of the sample estimates as well as the relative standard deviation and the relative mean error (RME) of those estimates. The RRMSE values for employed and unemployed are shown in figures 3.6 and 3.7 respectively, against the average sample size of each LGA.

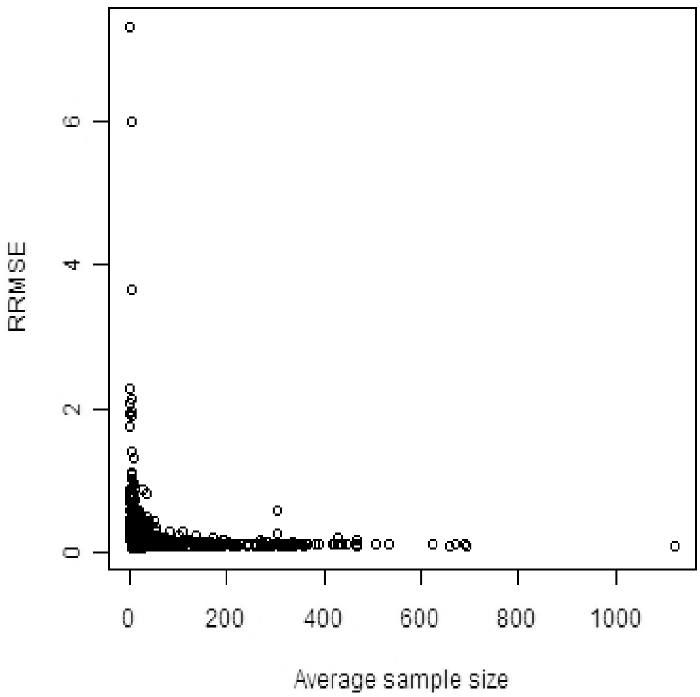
From these plots we can see that the RRMSE decreases as the average sample size increases. We also observe that the RRMSE values for employed were generally small, with only one greater than 25%, whereas for unemployed they were much larger. However even for unemployed, almost all LGAs with average sample size greater than 50 had RRMSEs less than 25%. An RRMSE of 25% means that roughly 95% of the time we would expect the estimates to be within plus or minus 50% of the original value. The pattern of RRMSEs decreasing as the average sample size increases was also observed for NILF, with RRMSEs slightly larger than those for employed. Therefore in general the small area estimates were of reasonable quality for LGAs with considerable average sample sizes, however they could be unreliable for LGAs with small average sample sizes, especially when predicting rare responses such as unemployment. This suggests that for those LGAs with small average sample sizes, the current SAE methodology cannot be used to give reliable predictions for rare responses such as unemployment. It is difficult to determine whether the small sample size was the cause of the volatility, whether it was the remoteness which is typical of those areas, or whether it was the poorer quality auxiliary information for indigenous people that typically have higher proportions in those areas. There was a single outlier, the LGA 'Unincorporated NT', with a much larger RRMSE than for other LGAs with average sample sizes similar to its 305.

Plots of the relative mean errors against the relative standard deviations for employed and unemployed are shown in figures 3.8 and 3.9 respectively. From these plots we find that the relative mean error was more variable than the relative standard deviation and obtained values of larger absolute value. This indicates that the largest RRMSEs, for the LGAs with small average sample sizes in figures 3.6 and 3.7, were due to LGAs with large relative mean errors. Once again the plot for NILF was similar to that of employed, but with slightly larger values. The large relative mean errors of some LGAs were due to large technical bias or parametric estimation bias values for those LGAs and the reasons for these biases are described subsequently. Also to note is that although the distribution of relative mean errors was roughly symmetrical for employed, in figure 3.8, it was skewed positively for unemployed, in figure 3.9. This is due to some LGAs with very low population counts of unemployed people, which skew the relative measure positively. This is a result of the model being unable to predict the very small counts of unemployed which exist in some small or remote LGAs. The median of the unemployed relative mean error distribution was 0.0415 indicating a positive bias in the relative mean error distribution. This is in contrast to the median of the employed relative mean error distribution of –0.00446. The distribution of relative mean error for NILF was similar to that of unemployed, but with less extreme values. The distribution for NILF had a slight positive skew and a median of 0.00838.

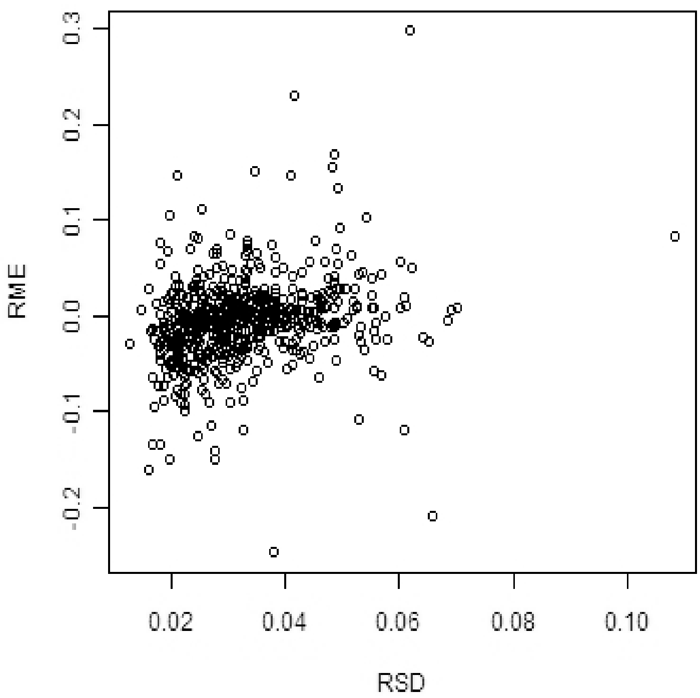
3.6 RRMSEs of small area estimates of employed against the average sample size in each of the LGAs



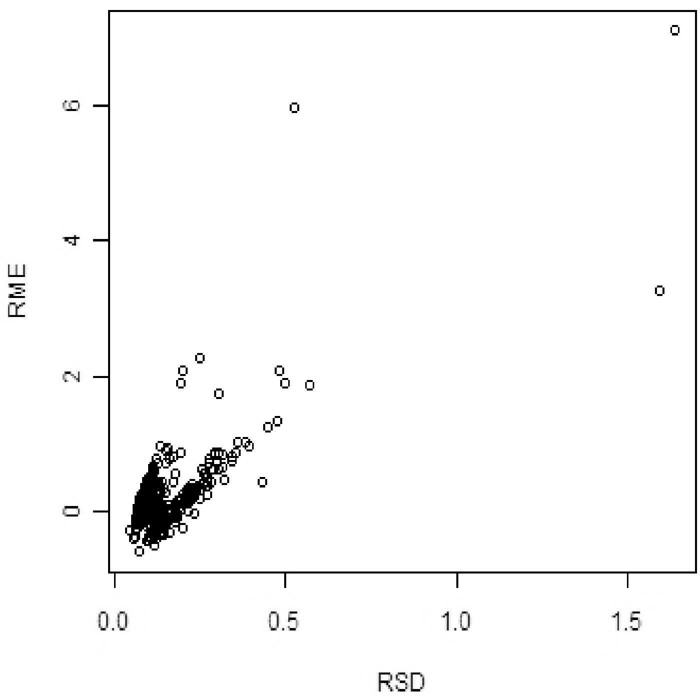
3.7 RRMSEs of small area estimates of unemployed against the average sample size in each of the LGAs



3.8 Relative mean errors of estimates of employed against their relative standard deviations

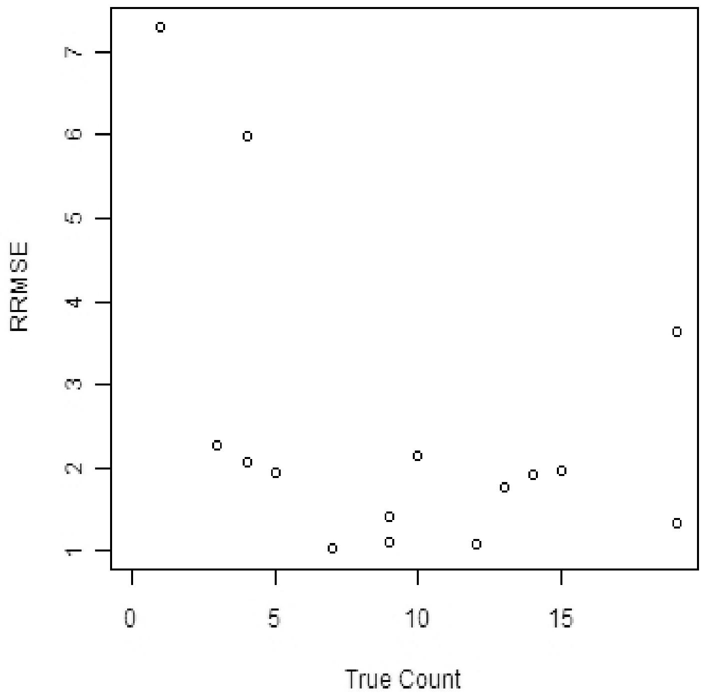


3.9 Relative mean errors of estimates of unemployed against their relative standard deviations



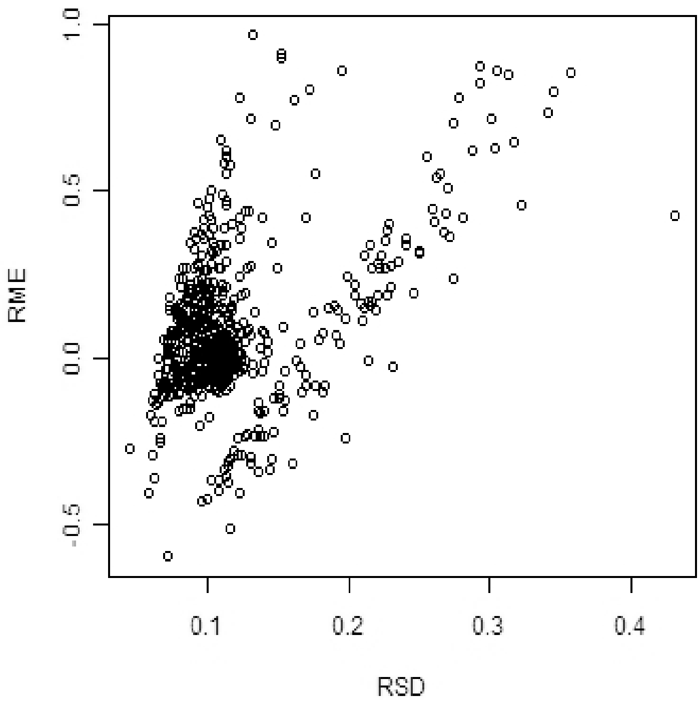
To verify that the largest RRMSEs of the unemployed estimates, shown in figure 3.7, were for LGAs with very low population counts of unemployed, the RRMSEs greater than 1 were plotted against their true unemployed count in figure 3.10. From figure 3.10, we can see that the largest RRMSEs were for LGAs with true unemployed counts less than 20. Therefore, due to the relationship among RRMSE, relative mean error and relative standard deviation in (3), the largest relative mean errors and relative standard deviations were also caused by very low population counts of unemployed.

3.10 RRMSEs of estimates of unemployed against their true unemployed count, for LGAs with RRMSEs greater than 1

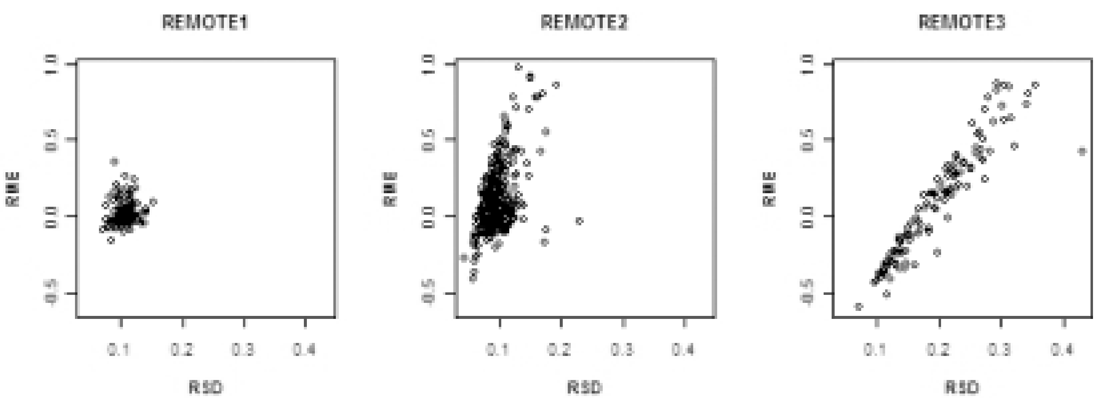


The relationship between unemployed relative mean errors and relative standard deviations, for LGAs with unemployed RRMSEs less than 1, is shown in figure 3.11. A feature to note from figure 3.11 is that two groups of LGAs look as though they could be separated by a straight line. The distinction between these two groups is clear when the LGAs within each remoteness classification are plotted separately, as is done in figure 3.12. Almost all LGAs in remote or very remote areas had larger relative standard deviations than LGAs in non-remote areas of similar relative mean error. This increased variability was most likely due to the highly variable coefficient of the interaction effect between unemployment benefits and remoteness classification 3. This coefficient was highly variable across the 1000 samples due to the small sample present in remoteness classification 3 areas. Despite its variability this coefficient was significant and was originally included in the model to reduce the positive bias of the relative mean errors for unemployed.

3.11 Relative mean errors of estimates of unemployed against their relative standard deviations, for LGAs with RRMSEs less than 1



3.12 Relative mean errors of estimates of unemployed against their relative standard deviations, for LGAs with RRMSEs less than 1, by remoteness



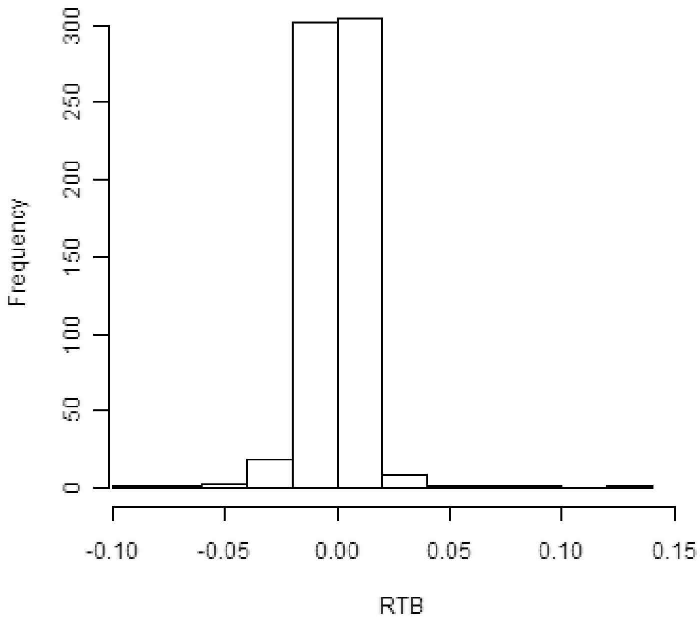
The plots in figure 3.12 also illustrate that whilst the centre of the unemployed relative mean error distribution was positively biased for all LGAs, this was possibly related to remoteness. This was confirmed, as the median unemployed relative mean error for LGAs in remoteness classification 1 (major cities) was 0.00523. Whereas for remoteness classifications 2 (inner and outer regional areas) and 3, the medians of 0.0593 and 0.154 respectively were much larger. The large positive bias of the relative mean errors for LGAs in remoteness classification 3 areas was possibly due to the coefficients of remoteness classification 3 and the interaction effect between unemployment benefits and remoteness classification 3.

As was described in the model coefficients results in Section 3.1.2, the sample estimates of these coefficients were both biased from the coefficient estimate when the model was applied to the Census data. However the positive bias of the relative mean errors for remoteness classifications 2 and 3 had been considerably reduced for this model when compared to our previous model for unemployed, which did not contain the covariates for remoteness or the interaction effect of remoteness and unemployment benefits.

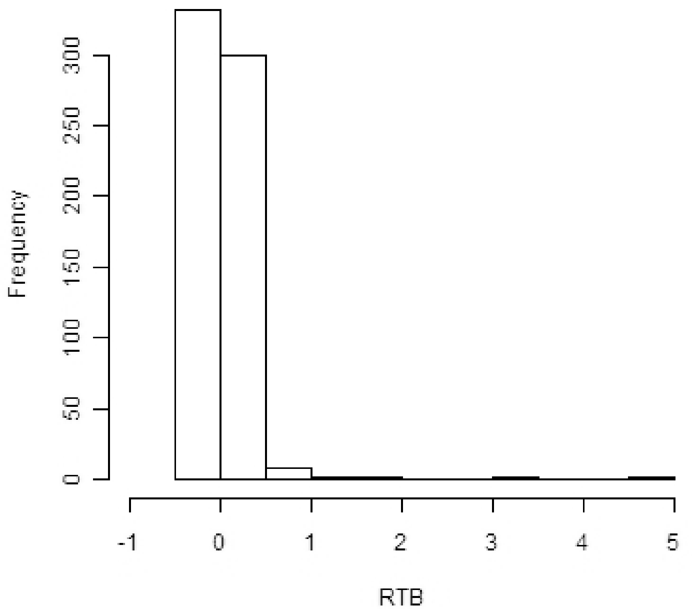
As the relative mean error was the major component of the largest RRMSEs, we were interested in decomposing the relative mean error into error due to technical bias and error due to parametric estimation bias, as defined in Section 2.2.2. Shown in figures 3.13 and 3.14 are histograms of the relative technical bias for employed and unemployed. These plots show that the technical bias was relatively small for the majority of LGAs; however, there were a few LGAs with large relative technical bias values. Values were again larger for unemployed than for employed. NILF was once again similar, with values slightly larger than those for employed. It was again apparent that there were some large relative technical bias values for unemployed due to very low counts of unemployed in the Census. These occurred for small LGAs or for those in remote areas. Both distributions were centred on zero, with the medians of $-6.20\text{x}10^{-6}$ for employed and $-1.11\text{x}10^{-3}$ for unemployed being being very close to zero.

Shown in figures 3.15 and 3.16 are histograms of the relative parametric estimation bias for employed and unemployed. These plots show the relative parametric estimation bias had more values of large magnitude than the relative technical bias. For instance 16.6% of relative parametric estimation bias values had magnitude greater than 0.25 whereas only 5.12% of technical bias values had such magnitudes. Also to note is that although the distribution of relative parametric estimation bias was roughly symmetrical for employed, it was skewed positively for unemployed. The distribution of relative parametric estimation bias was however roughly symmetrical for NILF. Furthermore, the medians of the unemployed and NILF relative parametric estimation biases of 0.043 and 0.009 respectively, were greater than zero. This means that for the majority of LGAs, the bias in the parameter estimates caused unemployed and NILF estimates to be larger than they would be if predicted by a model based on the entire population. This is in contrast to the median of the distribution for employed of –0.00440, which was less than zero. As described in Section 3.1.2 the bias in the parameter estimates may be because there are some design parameters not being accounted for in the model. For example, area type, which gives the degree of clustering in the multi-stage model, is not included in the models.

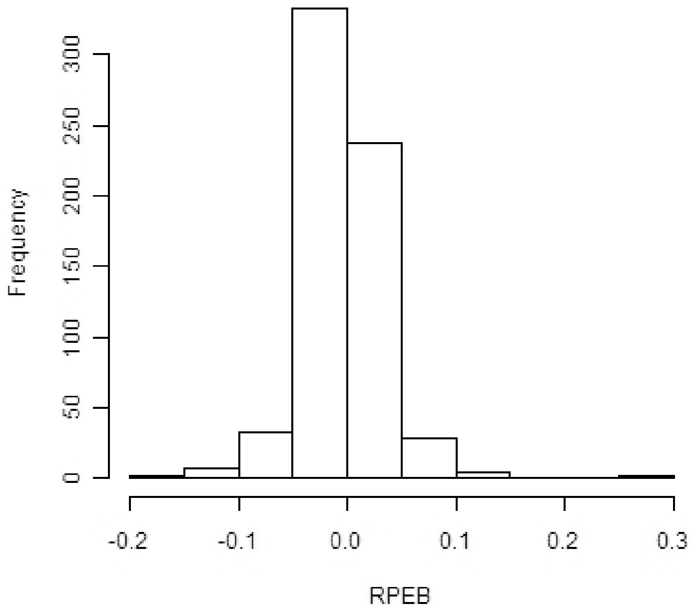
3.13 Relative technical bias for employed



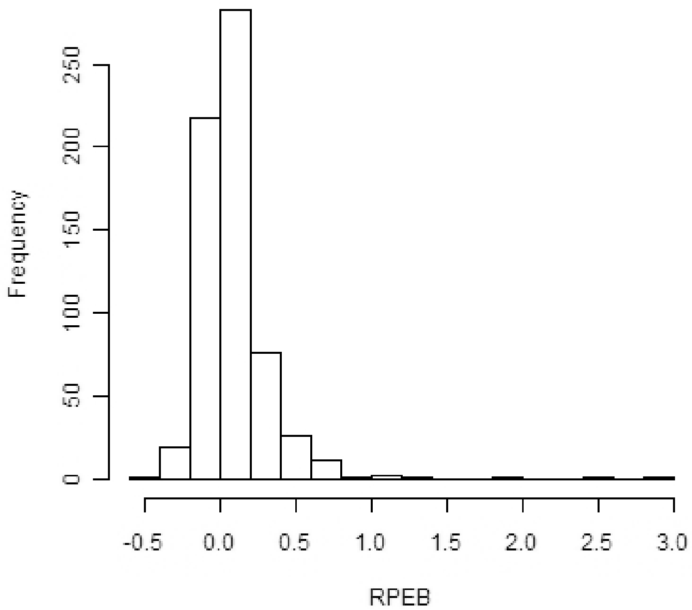
3.14 Relative technical bias for unemployed



3.15 Relative parametric estimation bias for employed

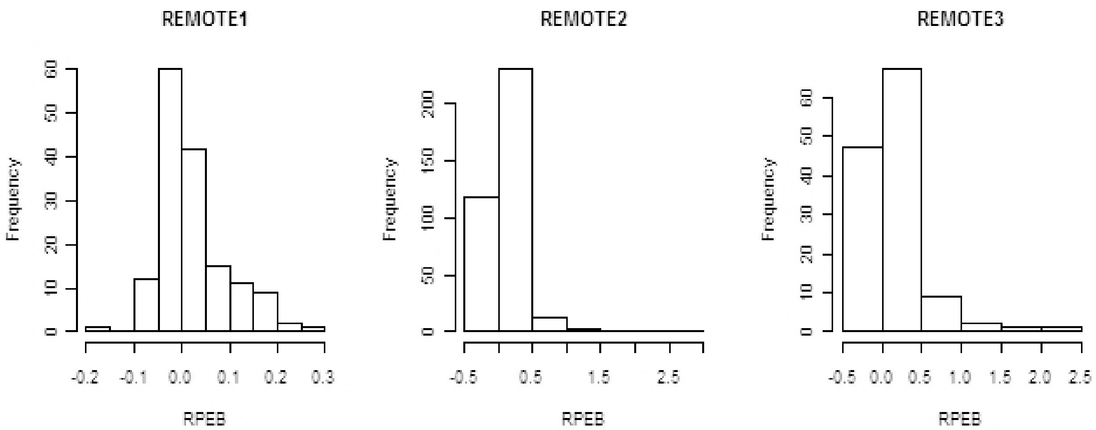


3.16 Relative parametric estimation bias for unemployed



The large positive median of the unemployed relative parametric estimation bias values may have been related to remoteness, as was the case for the relative mean errors. To determine if this was the case, the relative parametric estimation bias values were plotted by remoteness and these plots are shown in figure 3.17. The plots in figure 3.17 illustrate that the distribution of relative parametric estimation bias for major cities was centred on zero, whereas the distributions for regional and remote areas were centred above zero. This was confirmed as the medians for remoteness classifications 2 and 3 were 0.059 and 0.123 respectively, whereas the median for remoteness classification 1 of 0.004 was much closer to zero. As was the case for the relative mean errors, we suspect the positive median of the relative parametric estimation bias values for remoteness classification 3 areas to be due to the biased estimates of the coefficients of remoteness classification 3 and the interaction effect between unemployment benefits and remoteness classification 3.

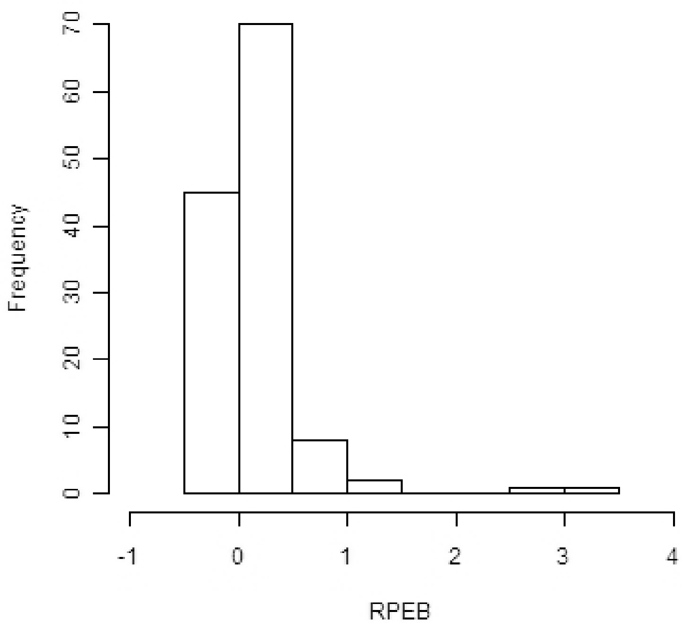
3.17 Relative parametric estimation bias for unemployed, by remoteness



As the positive median of the unemployed relative parametric estimation bias was largest for LGAs with remoteness classification 3, we decided to investigate whether this could be reduced by fitting a separate unemployed model to areas with remoteness classification 3. This could have reduced the bias if the estimates of the parameters for LGAs in remoteness classification 3 areas were different to those for LGAs in non-remote areas. Currently, if this was the case, it would result in bias for the remote areas because their data would be dwarfed by data from the major cities and regional areas.

With the new unemployed model fitted to remoteness classification 3 LGAs, the median of the relative mean error distribution was reduced from 0.154 for the old model, to 0.070. Figure 3.18 shows the distribution of the relative parametric estimation bias for the new unemployed estimates. Although the distribution appears similar to that for remoteness classification 3 areas with the old model shown in figure 3.17 above, the median was reduced from 0.123 to 0.0826. This median of 0.0826 was however still far from zero. Therefore fitting a new model for remoteness classification 3 LGAs reduced the amount of bias resulting from estimating the parameters using sample data. However, it is still the case that for a majority of LGAs, the bias in the parameter estimates from models fitted to samples caused unemployed estimates to be larger than if predicted by a model fitted to the population of remoteness classification 3 LGAs. That is, even when only remoteness classification 3 areas are considered, the design of the LFS results in a positive bias of unemployed estimates, for a majority of LGAs.

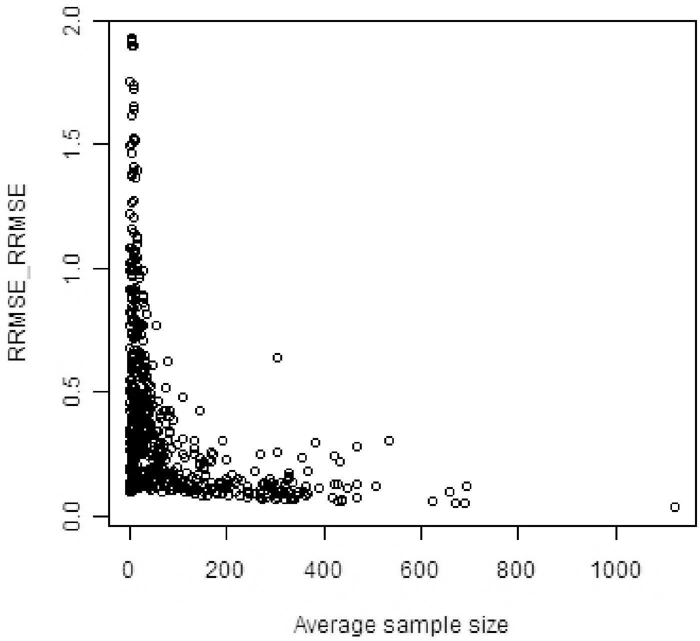
3.18 Relative parametric estimation bias for unemployed, for remoteness classification 3 LGAs with the new model fitted



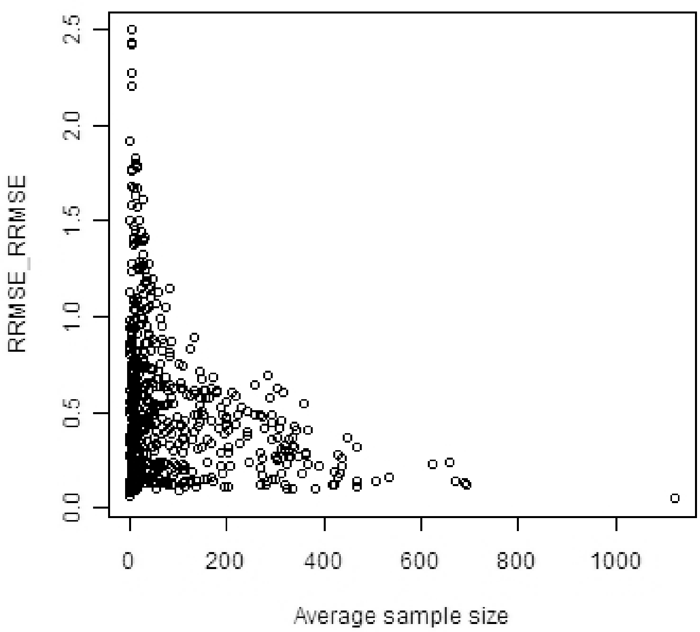
3.2.2 RRMSE results

As described in Section 2.2.3, to assess the quality of the predicted model-based RRMSEs we considered the design-based RRMSE calculated using the small area estimates, $RRMSE\_\hat{\theta}_d$ to be the target value of the model predicted RRMSE values. This allowed us to calculate RRMSEs of the model predicted RRMSEs as well as relative mean errors and relative standard deviation of the model RRMSE values. The RRMSEs of the model RRMSEs for employed and unemployed are shown in figures 3.19 and 3.20 against the average sample size of their LGAs. From these plots, we see that the RRMSEs again decrease as the average sample size increases, as was the case in figures 3.6 and 3.7 for the RRMSEs of the count estimates. The RRMSEs of the model RRMSEs were in general larger than the RRMSEs for the small area estimates, indicating the model RRMSE estimator is more volatile than the small area count estimator. The values for unemployed were in this case only slightly larger than those for employed, with the NILF values being similar. This indicates that when compared with the count estimator, the RRMSE estimator is less affected by the rarity of unemployment. There was a single outlier for the employed model, shown in figure 3.19, again for the LGA 'Unincorporated NT', with a much larger RRMSE of model RRMSEs than for other LGAs with similar average sample sizes to its 305.

3.19 RRMSE of model RRMSEs against average sample size for employed



3.20 RRMSE of model RRMSEs against average sample size for unemployed



Quartiles of the relative mean errors and relative standard deviations of the model RRMSEs for employed are also shown in table 3.21. The values for unemployed and NILF were again similar. The relative mean error of the model RRMSEs was again more variable than the relative standard deviation, and obtained values of larger absolute value. Therefore the relative mean error was the component which contributed to the largest $RRMSE\_\widehat{RRMSE}_d$ values. The distribution of relative mean errors was also positively skewed and centred above zero for employed. Similarly this was the case for unemployed and NILF. The proportion of relative mean errors greater than zero was 68% for employed and unemployed, and 71% for NILF. This means that for the majority of LGAs, the RRMSE estimates were conservative. However there were some LGAs with greatly optimistic RRMSE estimates, such as those in the first quartile of relative mean error values. The conservative nature of the RRMSE estimates may be due to a combination of:

- the parametric estimation bias,
- bias due to non-linearity in the relationship with the covariates, and
- overestimated variance between samples (i.e. the sampling variation implicitly estimated under the binomial assumption of the model overstates the actual sampling variation as measured by the design-based RRMSE).

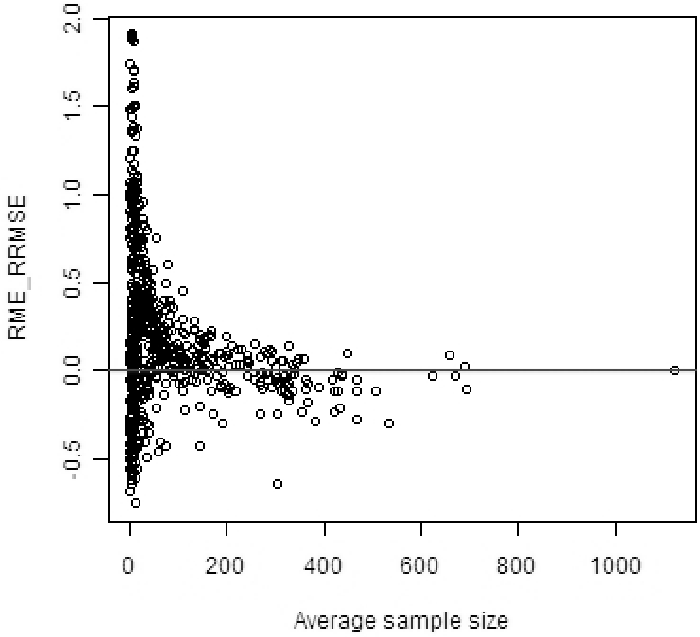
3.21 Summary statistics of relative mean errors and relative standard deviations of model RRMSEs for employed

	Minimum	*Q1*	*Median*	*Mean*	*Q3*	*Maximum*
$RME\_\widehat{RRMSE}_d$	–0.749	–0.050	0.122	0.228	0.409	1.910
$RSD\_\widehat{RRMSE}_d$	0.026	0.084	0.118	0.127	0.159	0.354

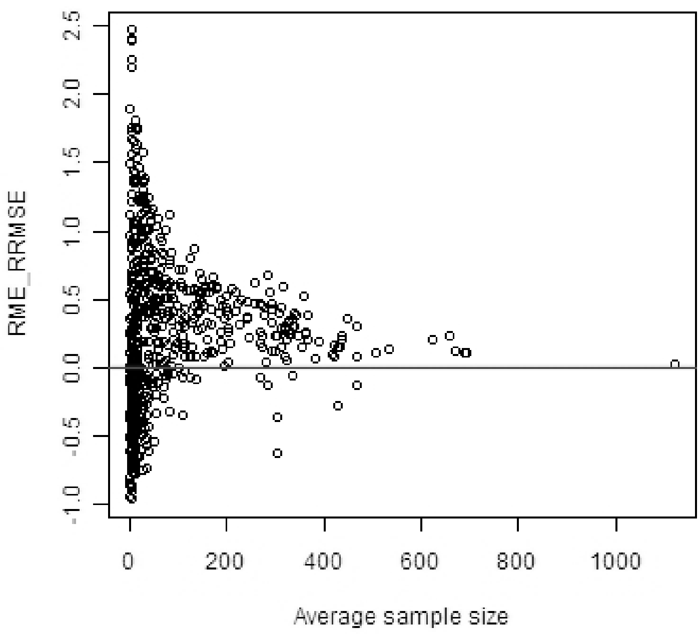
To further analyse the distribution of the relative mean errors of the model RRMSEs, we plotted them against the average sample size in figures 3.22 and 3.23, for employed and unemployed respectively. The plot for NILF was again similar.

From viewing figures 3.22 and 3.23, we can see that almost all of the optimistic RRMSE estimates occurred for LGAs with very small average sample sizes. For unemployed in particular there were very few optimistic RRMSE estimates for LGAs with average sample sizes greater than 100. Therefore, although RRMSE estimates were unreliable for remote or small LGAs with small sample sizes, they generally were accurate or conservative for the remainder of the LGAs of Australia. This suggests that the current RRMSE estimator should not be used for LGAs with small average sample sizes, whereas it is suitable for LGAs with reasonable average sample sizes.

3.22 Relative mean errors of model RRMSEs against average sample size for employed



3.23 Relative mean errors of model RRMSEs against average sample size for unemployed



4. FURTHER WORK

A spatial SAE model approach could be used to possibly improve the quality of the SAEs and RRMSEs for areas with small average sample sizes. If there is spatial clustering of random effects in figure 3.3, a spatial model could improve the estimates by allowing the estimator for a particular area to borrow strength from areas around it.

To attempt to remove more of the parametric estimation bias resulting from the design of the LFS samples, we could investigate including area type in the models of labour force status. This may be successful as area type is a variable used in the design of the LFS samples, although over-parameterization may be a risk. An alternative approach could be to use the methods of Pfeffermann and Sverchkov (2007) to adjust for the parametric estimation bias.

Further simulation studies could also be undertaken to investigate the optimal sample design for the output of both publication survey estimates and SAEs. For example a simple random sample could be taken from across Australia to determine whether this resolved the parametric estimation bias. Alternatively, equal size samples could be taken from all LGAs across Australia to determine how much this improved the quality of SAEs.

Another possibility for further work is to generate entire Census populations from models fitted to the Census data, from which LFS samples could then be taken. This type of parametric bootstrap approach would more appropriately suit the model-based assumptions used for the prediction of SAEs using GLMMs than this design-based simulation. This would allow us to calculate a true measure of design informativeness as well as model misspecification.

5. CONCLUSION

Current experimental small area estimates of labour force status are generally of reasonable quality. However, this simulation investigation has identified a number of important issues worthy of further investigation for future small area applications:

- Local government areas with small average sample sizes, generally in remote parts of Australia or with small populations, can have volatile small area estimates . The remote LGAs have the additional confounding factor of high proportions of indigenous persons, for which our auxiliary information may not be as strong as for the rest of the population.
- The parametric estimation bias of model parameters based on sample data is the major cause of differences between sample based SAEs and the true population value, with technical bias being a secondary cause.
- The parametric estimation bias was worse for remote and very remote areas, and fitting a separate model for these areas reduced the bias but did not remove it.
- Mean squared error estimates are generally conservative, however can greatly underestimate the mean squared error for some local government areas with small average sample sizes.
- Many of the model coefficient estimates have undercoverage of the Census value, resulting in the parametric estimation bias of the estimates.
- As is the case for the count and RRMSE estimates, the random effect estimates are reliable for LGAs with reasonable average sample sizes, however the estimates for LGAs with small average sample sizes can be unreliable.

With this greater awareness of the quality of the experimental SAEs and their estimated RRMSEs, we can better understand and describe the quality of SAEs for future applications. The ABS needs to decide how this knowledge influences our future production of SAEs. For example whether the ABS will not publish estimates for areas with small sample sizes, of if it will choose small areas to have a minimum population size. Alternatively the ABS could consider changing the design of surveys from which SAEs are desired to have large enough sample sizes from all chosen small areas. Additionally, in terms of future production of SAEs, we need to be aware of the parametric estimation bias we have observed and should attempt to account for the survey design in the analysis in the model selection or through other methods.

REFERENCES

Akaike, H. (1974) "A New Look at the Statistical Model Identification", *IEEE Transactions on Automatic Control*, 19 (6), pp. 716–723.

Australian Bureau of Statistics (2002) *Information Paper: Labour Force Survey Sample Design*, cat. no. 6269.0, ABS, Canberra.

Australian Bureau of Statistics (2003) *2001 Net Undercount*, Demography Working Paper, 2003/2, cat. no. 3134.0, ABS, Canberra.

Bleuer, S.; Godbout, S. and Morin, Y. (2007) "Evaluation of Small Domain Estimators for the Survey of Employment and Hours", *SSC Annual Meeting June 2007: Proceedings of the Survey Methods Section*, Statistics Canada, Ottawa.

Brown, G.; Chambers, R.; Heady, P. and Heasman, D. (2001) "Evaluation of Small Area Estimation Methods – An Application to Unemployment Estimates from the UK LFS", *Statistics Canada International Symposium Series – Proceedings*, cat. no. 11-522-XIE.

Department of Education, Employment and Workplace Relations (2008) *Small Area Labour Markets – December Quarter 2008*, website accessed 15 October 2009. <http://www.workplace.gov.au/workplace/Publications/LabourMarketAnalysis/SmallAreaLabourMarkets-Australia.htm>

Hall, P. and Maiti, T. (2006) "On Parametric Bootstrap Methods for Small Area Prediction", *Journal of the Royal Statistical Society (Series B)*, 68(2), pp. 221–238.

Pawitan, P. (2001) *In all Likelihood: Statistical Modelling and Inference using Likelihood*, Oxford University Press, New York.

Pfeffermann, D. and Sverchkov, M. (2007) "Small Area Estimation under Informative Probability sampling of Areas and within the Selected Areas", *S3RI Methodology Working Papers*, M07/06 Southampton, UK, Southampton Statistical Sciences Research Institute.

Purcell , N.J. and Kish, L. (1979) "Estimation for Small Domains", *Biometrics*, 35, pp. 265–384

Rao, J.N.K. (2003) *Small Area Estimation*, Wiley, New York.

Saei, A. and Chambers, R. (2003) "Small Area Estimation Under Linear and Generalised Linear Mixed Models With Time and Area Effects", *S3RI Methodology Working Papers*, M03/15 Southampton, UK, Southampton Statistical Sciences Research Institute.

Saei, A. and Chambers, R. (2005) "Out of Sample Estimation for Small Areas Using Area Level Data", *S3RI Methodology Working Papers*, M05/11 Southampton, UK, Southampton Statistical Sciences Research Institute.

Scealy, J. (2010) "Small Area Estimation Using a Multinomial Logit Mixed Model with Category Specific Random Effects", *Methodology Research Papers*, cat. no. 1351.0.55.029, Australian Bureau of Statistics, Canberra.

ACKNOWLEDGEMENTS

We would like to thank DEEWR for the use of their administrative data which was necessary to complete this analysis. We would also like to thank Jonathon Khoo, Frank Yu and Benjamin Ashley from the ABS for their comments on earlier drafts of this paper.

APPENDIXES

A. VARIABLE DEFINITIONS

Variable definitions for the original models of employed, unemployed and NILF.

All models contain the following covariates:

AS1	= 1 if class consists entirely of 15 to 24 year old males, 0 otherwise. This variable is the base case of the AS classes and is not explicitly in the model.
AS2	= 1 if class consists entirely of 15 to 24 year old females, 0 otherwise.
AS3	= 1 if class consists entirely of 25 to 34 year old males, 0 otherwise.
AS4	= 1 if class consists entirely of 25 to 34 year old females, 0 otherwise.
AS5	= 1 if class consists entirely of 35 to 44 year old males, 0 otherwise.
AS6	= 1 if class consists entirely of 35 to 44 year old females, 0 otherwise.
AS7	= 1 if class consists entirely of 45 to 54 year old males, 0 otherwise.
AS8	= 1 if class consists entirely of 45 to 54 year old females, 0 otherwise.
AS9	= 1 if class consists entirely of 55 to 64 year old males, 0 otherwise.
AS10	= 1 if class consists entirely of 55 to 64 year old females, 0 otherwise.

NSW	= 1 if LGA is located in New South Wales, 0 otherwise. This variable is the base case of the state variables and is not explicitly in the model.
QLD	= 1 if LGA is located in Queensland, 0 otherwise.
VIC	= 1 if LGA is located in Victoria, 0 otherwise.
SA	= 1 if LGA is located in South Australia, 0 otherwise.
WA	= 1 if LGA is located in Western Australia, 0 otherwise.
TAS	= 1 if LGA is located in Tasmania, 0 otherwise.
NT	= 1 if LGA is located in the Northern Territory, 0 otherwise.
ACT	= 1 if LGA is located in the Australian Capital Territory, 0 otherwise.

REMOTE1 = 1 if LGA is located in a major city, 0 otherwise. This variable is the base case of the remoteness classifications and is not explicitly in the model.

REMOTE2 = 1 if LGA is located in a non remote area (inner regional or outer regional Australia), 0 otherwise.

REMOTE3 = 1 if LGA is located in a remote area (remote or very remote Australia), 0 otherwise.

FULL_PAY = Proportion of population in class receiving full payment of YLS, AUS, DSP, ABY, PPP, PPS, CAR, WFD, PTA, WDA, WFA, NMA or SPL. This variable is not used in either of the four models. Instead it is used to create 10 interaction variables ASPAY1 to ASPAY10 by multiplying it by each of AS1 to AS10. This implies that the effect that PAY has on probability is different for each age–sex class.

The models for employed and NILF additionally contain the following covariates:

HHT0 = Proportion of population in another dwelling type such as special dwelling, visitors only or mixed household. This variable is the base case of the household types and is not explicitly in the model.

HHT1 = Proportion of population in class that lives in dwelling consisting of married couple only or married couple with at least one child aged 15 or over.

HHT2 = Proportion of population in class that lives in dwelling consisting of married couple with children all aged 0 to 14.

HHT3 = Proportion of population in class that lives in dwelling consisting of one person only or one person with at least one child aged 15 or over.

HHT4 = Proportion of population in class that lives in dwelling consisting of one person with children all aged 0 to 14.

SEIFA1 = 1 if LGA has a SEIFA Advantage–Disadvantage score in the top 25% of all LGAs in Australia, 0 otherwise. This variable is the base case of the SEIFA variables and is not explicitly in the model.

SEIFA2 = 1 if LGA has a SEIFA Advantage–Disadvantage score in the second 25% of all LGAs in Australia, 0 otherwise.

SEIFA3 = 1 if LGA has a SEIFA Advantage–Disadvantage score in the third 25% of all LGAs in Australia, 0 otherwise.

SEIFA4 = 1 if LGA has a SEIFA Advantage–Disadvantage score in the bottom 25% of all LGAs in Australia, 0 otherwise.

The models for employed and unemployed additionally contain the following covariate:

FULL_NSA_YAO = Proportion of population in class receiving full payment of Newstart Allowance or Youth Allowance (Other), which we refer to as unemployment benefits.

The model for unemployed also contains the following interaction effect between remoteness and unemployment benefits:

FULL_NSA_YAO_REMOTE1 = FULL_NSA_YAO value if LGA is located in a major city, 0 otherwise. This variable is the base case of the remoteness and unemployment benefits interaction effect and is not explicitly in the model.

FULL_NSA_YAO_REMOTE2 = FULL_NSA_YAO value if LGA is located in a non remote area (inner regional or outer regional Australia), 0 otherwise.

FULL_NSA_YAO_REMOTE3 = NSA_YAO value if LGA is located in a remote area (remote or very remote Australia), 0 otherwise.

Variable definitions for the model of unemployed for remoteness classification 3 areas:

Age1or5 = 1 if class consists entirely of 15 to 24 year olds or 55 to 64 year olds, 0 otherwise. This variable is the base case of the age classes and is not explicitly in the model.

Age2 = 1 if class consists entirely of 25 to 34 year olds, 0 otherwise.

Age3 = 1 if class consists entirely of 35 to 44 year olds, 0 otherwise.

Age4 = 1 if class consists entirely of 45 to 54 year olds, 0 otherwise.

Male = 1 if class consists entirely of males, 0 otherwise.

Otherstates = 1 if LGA is located in New South Wales, Victoria, Tasmania or the Australian Capital Territory, 0 otherwise. This variable is the base case of the state variables and is not explicitly in the model.

QLD = 1 if LGA is located in Queensland, 0 otherwise.

SA = 1 if LGA is located in South Australia, 0 otherwise.

WA = 1 if LGA is located in Western Australia, 0 otherwise.

NT = 1 if LGA is located in the Northern Territory, 0 otherwise.

full_nsa_yao = Proportion of population in class receiving full payment of Newstart Allowance or Youth Allowance (Other), which we refer to as unemployment benefits.

Indig2_01 = Proportion of population in class that are indigenous.

indig2_01_full_nsa_yao = Product of the proportion of population in class receiving unemployment benefits and the proportion of population in class that are indigenous.

Male_indig2_01 = Proportion of population in class that are indigenous if class consists entirely of males, 0 otherwise.

Otherstates_full_nsa_yao = full_nsa_yao if LGA is located in New South Wales, Victoria, South Australia, Tasmania or the Australian Capital Territory, 0 otherwise. This variable is the base case of the state_full_nsa_yao variables and is not explicitly in the model.

QLD_full_nsa_yao = full_nsa_yao if LGA is located in Queensland, 0 otherwise.

WA_full_nsa_yao = full_nsa_yao if LGA is located in Western Australia, 0 otherwise.

NT_full_nsa_yao = full_nsa_yao if LGA is located in the Northern Territory, 0 otherwise.

Age1or5_full_nsa_yao = full_nsa_yao if class consists entirely of 15 to 24 year olds or 55 to 64 year olds, 0 otherwise. This variable is the base case of the age_full_nsa_yao classes and is not explicitly in the model.

Age2_full_nsa_yao = full_nsa_yao if class consists entirely of 25 to 34 year olds, 0 otherwise.

Age3_full_nsa_yao = full_nsa_yao if class consists entirely of 35 to 44 year olds, 0 otherwise.

Age4_full_nsa_yao = full_nsa_yao if class consists entirely of 45 to 54 year olds, 0 otherwise.

B. COVERAGE PROBABILITIES

The tables show the variables, Census model coefficients and standard deviations (SD) and coverage proportions for the 1000 sample coefficient "95% confidence intervals" and the Census coefficient "95% confidence interval".

B.1 Employed model

Variable	*Coefficient*	*Coefficient SD*	*Coverage proportion*
Intercept	0.507	0.023	0.941
AS2	0.101	0.009	0.955
AS3	1.772	0.010	0.882
AS4	0.877	0.009	0.919
AS5	2.101	0.012	0.862
AS6	0.802	0.011	0.927
AS7	1.737	0.009	0.902
AS8	0.755	0.008	0.920
AS9	0.533	0.009	0.885
AS10	–0.419	0.009	0.898
VIC	–0.012	0.022	0.968
QLD	0.074	0.019	0.973
SA	0.074	0.023	0.968
WA	0.004	0.019	0.965
TAS	0.017	0.033	0.950
NT	–0.004	0.058	0.922
ACT	0.176	0.060	1.000
REMOTE2	0.055	0.037	0.960
REMOTE3	0.130	0.021	0.842
ASPAY1	–2.640	0.063	0.934
ASPAY2	–2.302	0.049	0.938
ASPAY3	–10.202	0.140	0.904
ASPAY4	–3.910	0.056	0.934
ASPAY5	–11.541	0.101	0.916
ASPAY6	–3.347	0.052	0.869
ASPAY7	–9.877	0.081	0.940
ASPAY8	–4.628	0.048	0.922
ASPAY9	–4.127	0.036	0.957
ASPAY10	–4.572	0.044	0.926
HHT1	0.015	0.020	0.922
HHT2	0.888	0.033	0.810
HHT3	–0.461	0.024	0.894
HHT4	–0.723	0.088	0.960
SEIFA2	–0.053	0.021	0.970
SEIFA3	–0.052	0.022	0.956
SEIFA4	–0.165	0.023	0.888
FULL_NSA_YAO	–2.054	0.068	0.462

B.2 Original unemployed model

Variable	*Coefficient*	*Coefficient SD*	*Coverage proportion*
Intercept	–2.69	0.03	0.875
AS2	–0.21	0.02	0.962
AS3	–0.54	0.01	0.938
AS4	–0.89	0.02	0.952
AS5	–0.83	0.01	0.949
AS6	–0.99	0.02	0.953
AS7	–0.87	0.02	0.959
AS8	–1.24	0.02	0.950
AS9	–0.83	0.02	0.943
AS10	–1.77	0.03	0.953
VIC	–0.04	0.03	0.969
QLD	–0.03	0.03	0.863
SA	–0.06	0.03	0.967
WA	0.02	0.03	0.954
TAS	0.09	0.04	0.860
NT	–0.31	0.08	0.928
ACT	–0.10	0.08	0.989
REMOTE2	0.01	0.02	0.956
REMOTE3	–0.21	0.03	0.779
ASPAY1	–0.22	0.11	0.903
ASPAY2	0.65	0.08	0.910
ASPAY3	1.66	0.21	0.934
ASPAY4	2.45	0.08	0.919
ASPAY5	3.34	0.16	0.933
ASPAY6	2.54	0.10	0.865
ASPAY7	2.62	0.14	0.931
ASPAY8	2.28	0.11	0.906
ASPAY9	0.67	0.08	0.898
ASPAY10	0.85	0.15	0.929
FULL_NSA_YAO	8.31	0.14	0.857
FULL_NSA_YAO_REMOTE2	–0.96	0.11	0.961
FULL_NSA_YAO_REMOTE3	–1.75	0.29	0.540

B.3 NILF Model

Variable	*Coefficient*	*Coefficient SD*	*Coverage proportion*
Intercept	–0.85	0.02	0.758
AS2	–0.07	0.01	0.934
AS3	–2.05	0.01	0.911
AS4	–0.81	0.01	0.929
AS5	–2.35	0.01	0.898
AS6	–0.75	0.01	0.927
AS7	–1.87	0.01	0.899
AS8	–0.64	0.01	0.910
AS9	–0.45	0.01	0.847
AS10	0.62	0.01	0.879
VIC	0.02	0.02	0.966
QLD	–0.06	0.02	0.971
SA	–0.06	0.02	0.979
WA	0.01	0.02	0.955
TAS	–0.05	0.03	0.978
NT	0.10	0.06	0.971
ACT	–0.16	0.06	0.999
REMOTE2	–0.08	0.04	0.976
REMOTE3	–0.12	0.02	0.834
ASPAY1	2.50	0.07	0.935
ASPAY2	2.32	0.05	0.931
ASPAY3	10.42	0.16	0.947
ASPAY4	3.89	0.06	0.951
ASPAY5	13.08	0.11	0.941
ASPAY6	3.61	0.05	0.934
ASPAY7	11.50	0.09	0.935
ASPAY8	5.08	0.05	0.955
ASPAY9	4.51	0.04	0.904
ASPAY10	4.84	0.04	0.916
HHT1	0.15	0.02	0.672
HHT2	–1.09	0.03	0.800
HHT3	0.67	0.03	0.748
HHT4	0.82	0.09	0.955
SEIFA2	0.02	0.02	0.935
SEIFA3	0.03	0.02	0.777
SEIFA4	0.11	0.02	0.962

B.4 New unemployed model for remoteness classification 3

Variable	*Coefficient*	*Coefficient SD*	*Coverage proportion*
Intercept	–3.44	0.11	0.955
Age2	–0.01	0.04	0.947
Age3	–0.24	0.04	0.940
Age4	–0.34	0.05	0.946
Male	0.06	0.03	0.955
QLD	–0.51	0.13	0.949
SA	–0.16	0.14	0.954
WA	–0.33	0.12	0.943
NT	–0.73	0.20	0.958
full_nsa_yao	13.15	0.63	0.945
indig2_01	1.36	0.17	0.926
indig2_01_full_nsa_yao	–17.74	1.57	0.927
Male_indig2_01	0.40	0.11	0.940
QLD_full_nsa_yao	2.64	1.12	0.960
WA_full_nsa_yao	2.90	0.76	0.970
NT_full_nsa_yao	0.43	0.90	0.953
Age2_full_nsa_yao	–2.80	0.48	0.949
Age3_full_nsa_yao	–0.92	0.55	0.952
Age4_full_nsa_yao	0.14	0.85	0.959

The following sections describe the model coefficients with the lowest coverage for each of the models.

Employed

The single employed model coefficient with coverage less than 80% was for the coefficient of unemployment benefits, which was biased to have a larger negative coefficient in samples than for the Census. The median of the sample coefficient estimates was –3.65 whereas the Census coefficient estimate was –2.05.

Original unemployed model

The two unemployed model coefficients with coverage less than 80% were remoteness classification 3 and the interaction effect between remoteness classification 3 and unemployment benefits. The interaction effect had a median of the sample estimates of –6.77, whereas for the Census the estimate was –1.75. On the other hand, the remoteness classification 3 coefficient had a median of sample estimates of 0.0187, which was greater than the Census model coefficient of –0.208. This bias of coefficients specific for remoteness classification 3 LGAs is part of the cause of the positive median of the relative parametric estimation bias for remoteness classification 3 areas described in Section 3.2.1.

NILF – Persons Not in the Labour Force

The five coefficients of the NILF model with coverage less than 80% were three household type variables which were biased negatively in samples, and SEIFA classification 3 and the intercept which were biased positively. These biases were similar to those described for employed and unemployed, in terms of the medians of the sample estimate distributions being different to the Census estimate.

New unemployed model for remoteness classification 3 LGAs

As mentioned above, a different model was subsequently used to predict unemployed for LGAs in remote and very remote areas, for reasons described in Section 3.2.1. The two covariates with lowest coverage of their coefficients, in this new model fitted to just the remoteness classification 3 areas, both involved proportion indigenous, the main effect as well as the interaction effects of proportion indigenous with unemployment benefits. The proportion indigenous effect, with a median of the sample estimates of 0.462, was biased to have a smaller coefficient in samples than for the Census, where the estimate was 1.35. The interaction effect between proportion indigenous and unemployment benefits had a median of sample estimates of –19.7, which was more negative than the Census model coefficient of –17.7.

FOR MORE INFORMATION . . .

INTERNET	**www.abs.gov.au** the ABS website is the best place for data from our publications and information about the ABS.

INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our website. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice.

PHONE	1300 135 070
EMAIL	client.services@abs.gov.au
FAX	1300 135 211
POST	Client Services, ABS, GPO Box 796, Sydney NSW 2001

FREE ACCESS TO STATISTICS

All statistics on the ABS website can be downloaded free of charge.

WEB ADDRESS	www.abs.gov.au