Australian Bureau of Statistics

1352.0.55.136

Research Paper

Response Modelling for the 2016 Census Enumeration Model

New Issue

Research Paper

Response Modelling for the 2016 Census Enumeration Model

Julian Whiting and Ross McNaughtan

Statistical Services Branch

Methodology Advisory Committee

29 November 2013, Canberra

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) FRI 05 SEP 2014

ABS Catalogue no. 1352.0.55.136

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 Paul Schubert, Statistical Services Branch on Canberra (02) 6252 6591 or email <statistical.services@abs.gov.au>.

RESPONSE MODELLING FOR THE 2016 CENSUS ENUMERATION MODEL

Julian Whiting and Ross McNaughtan
Statistical Services Branch

QUESTIONS FOR THE COMMITTEE

1. Can the proposed modelling framework be modified so that necessary assumptions will be better informed by data?
2. Can the committee suggest alternative strategies for modelling the 2011 Census data and data from the Census Testing program which will extract more information from these data?
3. Can the committee suggest strategies to inform assumptions concerning differences between respondent behaviour observed in the tests and behaviour which will be observed in the 2016 Census?

CONTENTS

APPENDIXES

RESPONSE MODELLING FOR THE 2016 CENSUS ENUMERATION MODEL

Julian Whiting and Ross McNaughtan
Statistical Services Branch

ABSTRACT

The 2016 Australian Census of Population and Housing is introducing several major changes to the data collection operation which aim to significantly reduce collection costs and improve data quality. The new enumeration model adds complexity to the management of field operations, and the data collection operation needs to be guided by predictions of the field resource requirements across different geographic regions. Fundamental to predicting the field resource requirements are predictions of the response rate within fine geographic regions at the different stages of enumeration. This paper proposes a modelling framework to predict Census response rates for fine geographic regions during different phases of the data collection operation. The modelling task is challenging because the changes to the enumeration model are expected to cause significant changes to the respondent behaviour observed in the 2011 Census. This paper presents strategies for estimating model parameters by combining 2011 Census data with other data and assumptions.

1. INTRODUCTION

1.1 2016 Census enumeration model

The Census of Population and Housing is the largest collection conducted by the Australian Bureau of Statistics (ABS). The Census aims to collect information about all persons in Australia on Census night, and the most important outputs are counts of persons and dwellings in fine geographic regions. A key objective for the Census is achieving high response rates in all regions.

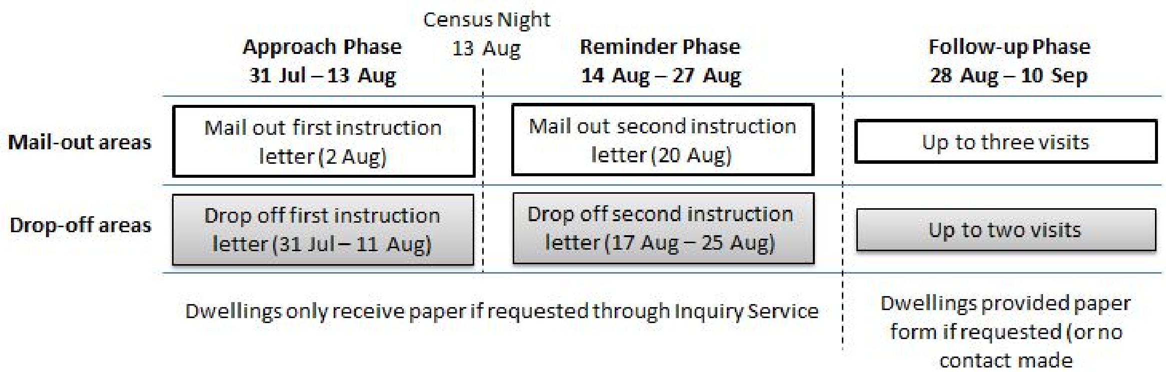
The collection of Census data is an enormous and costly operation requiring a large temporary field force. Important field management tasks include recruitment of staff, staff training, the assignment of workloads to individual staff and monitoring collection progress to ensure response rate targets are attained in all regions.

Until recently, the procedural model for Census data collection was very similar from Census to Census. The traditional enumeration model involved Census collectors being responsible for delivering a Census form to each household, establishing (where possible) the number of persons in the household, collecting completed Census forms after Census night and returning forms to a centralised processing centre. Under this traditional enumeration model each collector was responsible for identifying and collecting forms from all dwellings within their assigned Census Collector Workload (CLW).[1]

The 2016 Census will introduce several changes to the enumeration model which aim to significantly reduce collection costs and improve data quality (ABS, 2012). Figure 1.1 presents the enumeration model tested in the 2013 Census Test, which is indicative of the model planned for the 2016 Census.[2] Some directions and concepts of the 2016 enumeration model are now described.

The collection period will be divided into three broad phases. The first phase covers the period preceding Census night, and during this period all dwellings will receive an instruction letter providing the details needed to respond online. During the second phase dwellings yet to respond will receive a reminder letter. The final phase of collection is the Follow-up phase, during which field officers will repeatedly visit and attempt to make contact with the occupants of dwellings which have not yet responded. Paper Census forms will be provided to dwellings upon request. A 'calling card' and possibly other materials such as a paper form will be left at the dwelling for visits which do not result in contact. In general, it is only during the follow-up phase that field staff will be visiting dwellings to actively seek to make contact.

1.1 Enumeration model adopted for the 2013 Test



Perhaps the most significant change to the enumeration model is that at least half the dwellings in Australia will receive their initial Census instruction letter through the postal system. This mail-out will be undertaken only in areas where the ABS Address Register (the address frame) is high quality, typically in established suburbs in major metropolitan centres. Respondents who complete a paper form will be encouraged to return their completed form by mail, rather than wait for it to be collected by a field officer.

1 Formerly known as a Collection District.
2 The exact details of the model to be adopted in the 2016 Census are still being decided, with variations being the subject of tests in the Census Testing Program.

Field officer workloads will consist of a collection of dwellings located within a field management area called an 'Area Supervisor Workload' (ASW). In the areas where mail-out is not used (referred to as 'drop-off' areas), field officers will be responsible for the enumeration of all dwellings within a pre-defined sub-region contained in an ASW. In the mail-out areas individual field officers will not be responsible for all dwellings within a pre-defined area, but will instead be assigned workloads which are created during the enumeration period.

The multi-phase enumeration model approach is designed to maximise the proportion of dwellings which will respond online with minimal effort from field staff. Significant savings have been estimated compared with the cost of the traditional enumeration model, due to the estimated reduction of effort arising from not visiting dwellings which would respond without a visit. Maximising the level of internet response should also benefit data quality since, for example, item non-response tends to be lower for internet respondents (Statistics Canada, 2012).

1.2 Role of predictive modelling

The focus of this paper is the modelling framework for predicting response behaviour within geographic regions at different phases of the enumeration period. The modelling framework will be subsequently used to estimate staff resource requirements during follow-up. The basis for estimating staff resource requirements is a model for response behaviour during the follow-up phase which relates field officer visits and the amount of response elicited due to the field work.

At the planning stage the predictions of field work are used to determine the number of field staff recruited in each ASW. During the enumeration period, information on the response levels achieved to date can be used to update predictions about the amount of *further* field work needed to attain target levels of response. The updated estimates of further work could result in reallocation of field staff between ASWs.

From the perspective of overall management of field operations, the top-priority quantities to predict are the distribution of the number of dwelling visits needed in each ASW during the follow-up phase. The 'self-response rate' is the measure of the proportion of occupied dwellings requiring zero visits during the follow-up phase. The self-response rate is the focus for predicting resource requirements between ASWs since it is closely related to the total number of dwelling visits required during follow-up.

Predicting response mode distribution

The amount of field work does not directly depend on whether respondents use the internet or paper to respond, so estimating the self-response rate is more important than estimating the internet response rate. Nonetheless, predicting the proportions of response which are by paper and internet is needed for planning numerous aspects of the collection operation. For example, estimates of the demand for paper forms are required to ensure sufficient availability of paper forms and that there are adequate resources for managing the distribution and receipt of forms passing through the postal system. Prediction of the number of paper forms will also inform decisions about resourcing the processing operations which are required for paper responses but not internet responses (e.g. form scanning). Predicting the timing of the responses by the separate modes is also important, since the timing will determine capacity loads which systems and processes will need to accommodate.

Predicting total field effort

The application of the response model predictions to estimate field resource requirements is beyond the scope of this paper. It should be noted that estimating total field resource involves estimating many facets, including:

- the field work for enumeration of persons not in private dwellings;
- the field work to confirm the validity of addresses on the mail-out address frame;
- the field work to confirm whether a dwelling is occupied; and
- the time required to travel between dwellings and complete tasks for dwelling visits

Predicting the number of occupied private dwellings

The scope of this paper is estimating the response behaviour of the population within occupied private dwellings. This paper assumes the number of occupied private dwellings in each ASW is known prior to Census, though in practice determining which dwellings should be covered by the Census and their occupancy status presents many challenges.

1.3 Outline of paper

The structure of this paper is as follows. Section 2 briefly reviews recent approaches used by overseas National Statistical Offices to predict respondent behaviour for their Censuses. Section 3 proposes the framework for modelling response to the 2016 Census, and discusses the data available to support the modelling work. Section 4 discusses estimation of the model parameters relating to response prior to the follow-up phase, while Section 5 considers modelling respondent behaviour during the follow-up phase. A summary of the parameters discussed throughout the paper is given in Appendix E.

2. OVERSEAS EXPERIENCES FOR MODELLING CENSUS RESPONSE

Overseas National Statistical Offices have been making similar changes to the ABS in how they collect their Census data, and have accordingly developed models to predict Census response rates under the changed enumeration model. Resource planning decisions for the 2011 Canadian Census were guided by a model predicting response rates, and predictions were regularly updated during the enumeration period to ensure efficient use of resources (Statistics Canada, 2012). The Office of National Statistics (ONS) developed models predicting the self-response rate in fine geographic areas and the rate at which responses will be received during the follow-up period (Townsend, 2011). The U.S. Bureau of the Census (USBC) modelled the return rate data from the 2000 U.S. Census to predict the areas likely to require more follow-up effort for the 2010 U.S. Census (Bates, 2011).

The response prediction models used by Statistics Canada and the ONS are summarised below. They share the approach of separately modelling:

1. the self-response rate (response rate prior to commencement of follow-up); and
2. the number of returns within a region during the follow-up phase based on the amount of field force effort assigned to the region.

2.1 Response model for 2011 Canadian Census

For their 2011 Census Statistics Canada adopted a multi-mode, multi-wave[3] enumeration model similar to that planned for the 2016 Australian Census. Two important differences from the Australian enumeration model are that the follow-up period for the Canadian Census could be extended for a much longer period, and that a shorter Census form enabled Canadian field officers to immediately complete a form upon making contact with a dwelling's occupants.

Statistics Canada's model for self-response was founded on self-response rate predictions at the national level. The model decomposed response within classes defined by the combination of 'Enumeration Group', response wave and response mode. The 'Enumeration Group' classification is a broad three-category grouping of areas which determined the contact strategy adopted for dwellings in the area.

The model for self-response used parameters $\varphi^{g,(w,m)}$ specifying the proportion of dwellings within Enumeration Group g which would respond by mode m due to wave w.

For example, $\varphi^{g,(1,1)} = 0.25$ means 25% of the total occupied addresses in Group g were expected to respond due to wave $w = 1$ via the internet ($m = 1$). The values of $\varphi^{g,(w,m)}$ for waves 2 and 3 assumed the impact of the wave would be the same in

3 A wave is analogous to a 'phase'.

each Enumeration Group. Thus $\varphi^{g,(2,m)}$ and $\varphi^{g,(3,m)}$ were derived as $\varphi^{g,(w,m)} = p_w p_w^{g,(m)}$, where p_w is an overall impact of wave w and $p_w^{g,(m)}$ is the proportion of responses in group g from wave w which are by mode m. The chosen values for these parameters were informed by observations from the 2006 Census, the 2009 Census Test and by judgement.

The primary geographic level of prediction for planning field resource requirements was the 37 Local Census Offices (LCOs).[4] The number of self-responding dwellings for the LCOs was modelled by apportioning the national self-response estimate. The ratios used to apportion the national estimate between the LCOs were a function of: (1) the estimated distribution of dwellings between the LCOs; (2) the self-response rates achieved in the LCO in the 2006 Census; and (3) the distribution of dwellings between the Enumeration Groups within each LCO.

The model for the follow-up period provided weekly predictions of the number of responses received within each LCO. The amount of response achieved in a week was modelled to be proportional to the amount of resources assigned to follow-up. Follow-up effort continued in an LCO until response targets were met.

2.2 Response model for 2011 Census of England and Wales

The 2011 Census of England and Wales was the first Census run by the ONS to mail out questionnaires, but in contrast to the Australian model internet response was not strongly pushed.[5]

A field allocation model was developed to plan the resources required for follow-up to achieve target response levels in fine geographic areas called Lower Layer Super Output Areas (LSOAs). The population of LSOAs is typically between 1,000 and 3,000 persons. One major component of the field allocation model was a model predicting the self-response rate[6] in each LSOA. The basis for this self-response model was a model for *non-response* to the 2001 Census. The logit of estimated LSOA non-response rates were modelled as a linear function of many region variables, some of which could be updated from the values recorded in the 2001 Census (Hopper, 2012). The model was applied to the latest region data and each LSOA was classified into one of five categories of a Hard-to-Count (HTC) classification. Particular attention was given to identifying the most difficult areas and estimating the resource requirements for them. HTC Class 5 ('most difficult-to-count') covered just 2% of the population and Class 4 (second 'most difficult-to-count') covered 8%.

---

4 All LCOs contained dwellings in Enumeration Groups 1 and 2, and most had dwellings in Group 3.

5 The internet take-up rate for the 2011 Census of England and Wales was less than 20%, which is considerably smaller than the 33% internet take-up rate achieved in the 2011 Australian Census.

6 The self-response rate was defined as the rate of response achieved at the commencement of the follow-up period (10 days after Census day)

To produce self-response rate estimates under the conditions of the 2011 Census, the modelled 2001 non-response rates for each HTC class were adjusted in two stages. The first adjustment was to account for the impact of follow-up in 2001 (the impact is the discrepancy between the final response rate and the non-response rate). Analysis of the timing of returns for the 2001 Census found 82% of the returns ultimately received arrived no later than 10 days after Census day, though HTC Class 5 had a noticeably lower rate. So the first adjustment to the modelled 2001 non-response was to multiply them by 0.82 in HTC Classes 1 to 4, and by 0.68 in HTC Class 5. A subsequent blanket adjustment accounted for an expected decrease in the self-response rate between 2001 and 2011 due to the introduction of mailing out Census forms for all dwellings. The value for this adjustment was estimated from an experiment embedded in the 2007 Census Test. The predicted self-response rate for an LSOA was the predicted average self-response rate for its HTC class.

The model used to predict response during the 4½ week follow-up period assumed the number of returns received on a day would be proportional to the estimated number of household contacts two days prior. The number of household contacts per day was derived as the product of the estimated number of household visits achievable with the allocated field staff and assumed contact rates. Contact rates achieved at the third visit for ONS social surveys provided the assumed contact rates at the first Census visit. Social survey contact rates at the fourth visit provided the estimated contact rates at the second Census visit, and so on.

Evaluation of the model predictions showed the achieved response rates prior to follow-up exceeded expectations. This outcome was attributed to applying conservative assumptions and the publicity campaigns and community engagement programs having greater impact than anticipated (ONS, 2012). On a per-dwelling basis, the amount of effort required to elicit response during follow-up was higher than predicted. One reason for this was that the higher self-response rate reduced the anticipated number of ‘easy-to-persuade’ follow-up dwellings. Another reason was overestimation of contact rates during follow-up.

3. FRAMEWORK FOR MODELLING RESPONSE

3.1 Proposed framework for modelling response

The modelling framework divides occupied private dwellings into six classes defined by the response period[7] and response mode preference (table 3.1). The latent dwelling attributes of 'capability to respond promptly' and 'willingness to participate' would explain the different response behaviours underlying the classification. The response class of each dwelling is described by a multinomial distribution, with the parameters $\theta^{(pm)}$ (period $p = A, B$ or C; mode $m = 1$ or 2) specifying the probability of a dwelling belonging to each class. The parameters $\theta^{(pm)}$ can also be interpreted as the expected proportions of dwellings in the population belonging to each class, meaning the self-response rate[8] $\theta^{(AB\cdot)}$ is the sum $\theta^{(AB\cdot)} = \theta^{(A1)} + \theta^{(A2)} + \theta^{(B1)} + \theta^{(B2)}$.

3.1 Theoretical classification of a population by Census response behaviour

		Mode choice	
Period of response		*Internet*	*Paper*
Period *A*:	The time before reminder letters are first received	$\theta^{(A1)}$	$\theta^{(A2)}$
Period *B*:	Time period between when reminder letters are received and commencement of the follow-up phase	$\theta^{(B1)}$	$\theta^{(B2)}$
Period *C*:	The follow-up period – respondents receive various degrees of follow-up prompting during this period	$\theta^{(C1)}$	$\theta^{(C2)}$

The response behaviour of dwellings belonging to these classes is described by separate sub-models, summarised in table 3.2. A sub-model specific to the follow-up phase is needed to describe the relationship between the response during follow-up and the amount of field officer effort to elicit response. Such a model can be used to inform decisions about the timing and frequency of follow-up visits in different geographic regions. To support comparison of alternative follow-up procedures, the modelling framework described in this paper does not assume specific limits on the number dwelling visits during follow-up. Adopting separate models for the periods before and during follow-up is in line with the approaches used by Statistics Canada and the ONS, described in Section 2.

Separate sub-models are proposed for Periods *A* and *B* to distinguish differences in the response time distributions for these periods. Before reminder letters are received the shape of the temporal distribution of response will be affected by a range of factors with impacts difficult to predict from available data. So for Period *A* the predicted temporal distribution will need to be based on judgement (informed to

7 Note these periods are distinct from the phases of the enumeration model described in Section 1.
8 Note this definition of the 'self-response rate' excludes responses received during the Follow-up phase from dwellings which submitted prior to receiving a follow-up visit.

some extent by the timing of returns for 2011 and the Census tests). For the short time window defined by Period B a parametric model for the response time is proposed. Separately modelling these two periods is similar to Statistics Canada's strategy of predicting the impact of each 'wave' prior to follow-up.

3.2 Features of models for three time periods

Period of response	*Model for response time distribution*	*Key inputs which differentiate behaviour between regions*
Period *A*:	Based on empirical distributions from 2011, Tests and judgement	• Region demographics measured in 2011 Census • Assumptions about behaviour under 2016 enumeration model
Period *B*:	Parametric model	
Period *C*	Parametric model	• Follow-up strategy assigned to area • Region demographics measured in 2011 Census

A key goal of the modelling framework is predicting the regional variability in the response behaviour during each response period, and so the multinomial distribution given by the $\theta^{(pm)}$ is specified for each ASW.

The remainder of this section discusses the data sources available to develop the models, noting the limitations of each source.

3.2 2011 Census data

A person-level dataset of respondents to the 2011 Census provides a means to identify characteristics of persons and dwellings associated with prompt response and mode preference under the 2011 enumeration model. While the mode of response was captured for each responding dwelling, the timing of form submission was recorded electronically for online responses only. The number of times a Census field officer visited the dwelling during follow-up is available for each dwelling, but this data item can only be considered a broad indicator of compliance or 'time to respond' under the 2016 enumeration model. The number of visits does not directly measure the amount of prompting needed for dwelling occupants to participate in the Census. Some dwellings which completed their paper form on Census night would have required several visits because of difficulty for the field officer to make contact to collect the completed form.

The changes to the enumeration model will markedly impact both the timing and mode of response, so the 2011 Census data alone cannot provide accurate models for respondent behaviour for the 2016 Census. For example, in 2011 field officers made contact with around 45% of dwellings during delivery, and when there was no contact the field officer left a paper form. Under the 2016 model, during the approach phase there will be no contact with dwelling occupants and minimal paper form dispatch. The enumeration model changes are hoped to approximately double the 33% dwelling internet response rate achieved in 2011.

Because of these limitations the primary roles of the 2011 data in the framework are to:

- identify demographic characteristics in regions associated with compliance and propensity to respond online under the 2011 enumeration model; and
- provide the data about the characteristics of regions to be used as explanatory variables to predict the relative response behaviour of regions.

3.3 Data from 2016 Census Testing Program

The Census Program is undertaking a series of large-scale field tests in the lead-up to the 2016 Census. A key testing objective is measuring how the changes to the enumeration model impact mode choice and timing of response.

Overview of field tests

The enumeration model for the 2013 Test was introduced in Section 1.1. The total number of dwellings sampled for this test was around 20,000 dwellings. The follow-up phase was conducted over a two-week period, during which dwellings which had not responded were to be visited by a field officer up to three times. In practice, many of the non-response dwellings only received one or two visits.

The Major Test in August 2014 will have a sample of around 100,000 dwellings, to be distributed across a range of region types in Australia. The volume of data and spread of sample across different types of regions will be valuable for assessing the extent of variation in response behaviour between geographic regions and our ability to model this variation. The response models should be close to final following the analysis of the 2014 Test data, though there may be some small refinements arising from analysis of data from the final field test, the Census Dress Rehearsal in 2015.

Limitation of test data

Respondent behaviour observed in the tests will differ from the behaviour during the actual Census. Table 3.3(a) classifies the population according to behaviour under the conditions of a test. Since there is no compulsion to respond to the tests this classification includes a class for dwellings which do not respond[9] (the size of this non-response class proportion is $\gamma^{(D\cdot)} = \gamma^{(D1)} + \gamma^{(D2)}$). The test non-respondent dwellings would be distributed across the various response classes in the actual Census. Besides non-response, another reason why behaviour observed under the conditions of the tests will not reflect behaviour during the actual Census is that during tests dwellings are not exposed to the same degree[10] of media attention and public relations exercises. The higher public awareness associated with the Census

9 A non-response class is not included in the classification of Census behaviour because the complete response would be obtained under the Period *C* model which does not limit the amount of follow-up.

10 Some tests may include targeted local media campaigns.

means there is greater willingness to participate in the Census, and so for example, the response proportion during Period *A* would be considerably smaller under the conditions of a test (i.e. $\gamma^{(A\cdot)} < \theta^{(A\cdot)}$).

Responses from a test can be summarised by a set of observed proportions belonging to each class, as shown in table 3.3(b). Theoretically the non-respondents would have a mode preference (given by size of the $\gamma^{(D1)}$ and $\gamma^{(D2)}$), but in a test the mode preference among the non-respondent proportion is not observed.

3.3 (a) Classification of behaviour for Census Test; and (b) observed data from a Census test

(a) Population under test conditions			*(b) Test data*		
	Mode choice			*Mode choice*	
Period of response	*Internet*	*Paper*	*Period of response*	*Internet*	*Paper*
Period *A*	$\gamma^{(A1)}$	$\gamma^{(A2)}$	Period *A*	$b^{(A1)}$	$b^{(A2)}$
Period *B*	$\gamma^{(B1)}$	$\gamma^{(B2)}$	Period *B*	$b^{(B1)}$	$b^{(B2)}$
Period *C*	$\gamma^{(C1)}$	$\gamma^{(C2)}$	Period *C*	$b^{(C1)}$	$b^{(C2)}$
Non respondent	$\gamma^{(D1)}$	$\gamma^{(D2)}$	Non respondent	$b^{(D\cdot)}$	

Due to these limitations, assumptions are needed to translate estimates of the multinomial distribution for the test (described by $\gamma^{(pm)_t}$) to the multinomial distribution for the actual Census (described by $\theta^{(pm)_c}$). These assumptions are specified by conditional probabilities $P\left((PM)_c = (pm)_c \,\middle|\, (PM)_t = (pm)_t\right)$: the probability of a dwelling belonging to Census response class $(pm)_c$ given its response class under test conditions, $(pm)_t$. These assumptions cannot be validated from the outcomes of the testing program.

Although models which predict the $\gamma^{(pm)_t}$ are unsuitable for directly predicting the $\theta^{(pm)_c}$, the tests will provide valuable data for modelling response behaviour in the actual Census. Some example uses of the test data include:

- assessment of the predictive power of 2011 Census data items and other data sources to differentiate 2016 Census response behaviour between geographic regions;
- estimating rates of internet response among the responses received during the different phases of collection;
- estimating of the probability of making contact with dwelling occupants during follow-up;
- measuring of the impact on response of a particular visit number to a dwelling relative to the impact of earlier or subsequent visits.

4. RESPONSE MODELLING PRIOR TO FOLLOW-UP

4.1 Introduction

This section proposes a structure for models describing respondent behaviour prior to the follow-up phase and the strategy for estimating the model parameters using the data sources available. A top-down approach is proposed for modelling the relative response behaviour of individual ASWs across the days of the period prior to follow-up. Estimates for the national response rate at two key time-points are first derived, and then estimates at ASW level[11] and specific points in time are defined as a function of the national estimates.

The starting points for this top-down approach are the national average response probabilities across Periods *A* and *B*. Formally, the two periods are:

- Period *A*: The period between the start of enumeration to when reminder letters are first received (day t_{RL1}). The national response rate for this period is $\theta^{(A\cdot)}$.
- Period *B*: The period between t_{RL1} and the last day of the Reminder phase (day $t_{R\,End}$). The change in the national response rate over this period is $\theta^{(B\cdot)}$.

The expected self-response rate at the national level is $\theta^{(AB\cdot)} = \theta^{(A\cdot)} + \theta^{(B\cdot)}$.

Response probabilities during Periods *A* and *B* at the ASW level ($\theta_a^{(A\cdot)}$ and $\theta_a^{(B\cdot)}$, respectively) are modelled in terms of the *relative* response probabilities between ASWs. Using f_a^p to denote the multiple of the national response probability for period p within ASW a, we have: $\theta_a^{(A\cdot)} = f_a^A\,\theta^{(A\cdot)}$ and $\theta_a^{(B\cdot)} = f_a^B\,\theta^{(B\cdot)}$. The national parameters $\theta^{(A\cdot)}$ and $\theta^{(B\cdot)}$ are effectively scaling factors which convert the relative quantities for ASWs into ASW probability estimates. To ensure the regional self-response estimates cohere with the national estimates, the factors f_a^p must satisfy $\sum_a f_a^p N_a = N$, where N_a is the number of occupied dwellings in ASW a and N is the national count of occupied dwellings.

The top-down modelling strategy is also applied to the *timing* of response within the period prior to follow-up. Response-time distributions $T^p(t)$ specify the cumulative proportion of the Period p responses received by the t-th day of Period p. Separate distributions $T^p(t)$ are defined for Periods *A* and *B*. The model proposes that the same Period *A* distribution $T^A(t)$ applies across all ASWs, meaning on every day each ASW is modelled to attain the same proportion of its Period *A* response. In contrast the distribution $T^B(t)$ will be allowed to vary between ASWs so as to appropriately reflect differences between ASWs for the delivery time of reminder letters.

11 The method described to derive the ASW estimates from the national estimates could be applied to produce estimates for geographic regions finer or broader than ASW.

Denoting $R_a(t)$ as the proportion of occupied private dwellings in ASW a which have responded by time t :

$$R_a(t) = \begin{cases} T^A(t)\,\theta_a^{(A\cdot)} & \text{if } t \le t_{RL1} \\ \theta_a^{(A\cdot)} + T_a^B(t - t_{RL1})\,\theta_a^{(B\cdot)} & \text{if } t_{RL1} \le t \le t_{R\,End} \end{cases}$$

By definition, $R_a(t_{RL1}) = \theta_a^{(A\cdot)}$ and $R_a(t_{REnd}) = \theta_a^{(A\cdot)} + \theta_a^{(B\cdot)}$.

This section is structured as follows. Sections 4.2 to 4.4 concern modelling for Period *A*: Section 4.2 discusses estimation of the national response rate, Section 4.3 presents strategies to model ASW variation and Section 4.4 discusses the temporal distribution of responses during Period *A*. Section 4.5 discusses modelling all aspects of response behaviour during Period *B*, and Section 4.6 concerns modelling the response distribution between the response modes.

4.2 Estimating the national Period *A* response probability

The challenge for estimating the national response probability for Period *A*, $\theta^{(A\cdot)}$, is that the 2011 Census does not provide an analogous measure. Under the proposed 2016 enumeration model, by the end of Period *A* (around one week after Census night), no face-to-face contact with dwellings will have occurred. In contrast, at the corresponding time point in the 2011 Census follow-up had commenced, and some collectors had made contact with dwelling occupants at the time of delivery.

The Period *A* response rate observed in a test, $b^{(A\cdot)}$, will almost certainly under-estimate $\theta^{(A\cdot)}$ due the differences between the conditions of the tests and the actual Census. Nonetheless the distribution of the test sample across the classes in figure 3.2(b) is useful for suggesting upper and lower bounds for $\theta^{(A\cdot)}$. Deriving these bounds requires speculating a plausible range for the probability a dwelling in the *test* response class $(pm)_t$ would respond during Period *A* in the Census. This probability is denoted as $P\left(A\cdot_c \mid (pm)_t\right)$, $(p = A, B, C, D\ ;\ m = 1, 2)$. For example, we would expect practically all dwellings which responded during Period A in a test to also respond during Period *A* in the actual Census (i.e. $P\left(A\cdot_c \mid A1_t\right) \approx 1$ and $P\left(A\cdot_c \mid A2_t\right) \approx 1$). A hypothetical set of test results and probability ranges are shown in table 4.1.

4.1 Hypothetical example for estimating bounds on $\theta^{(A\cdot)}$ using test data and assumptions

	$(pm)_t$						
	*A*1	*A*2	*B*1	*B*2	*C*1	*C*2	*D*•
Test value for $b^{(pm)_t}$	0.30	0.05	0.05	0.02	0.10	0.10	0.38
Range for $P\left(A\cdot_c \mid (pm)_t\right)$	(0.95,1)	(0.95,1)	(0.5,0.7)	(0.5,0.7)	(0.4,0.7)	(0.4,0.7)	(0.2,0.6)

For these ranges a simple calculation shows the bound for $\theta^{(A\cdot)}$ is (0.52, 0.77). Ultimately, the point estimate of $\theta^{(A\cdot)}$ chosen within bounds derived in this fashion will be based on judgement on how the public will react to the new enumeration model.

4.3 Modelling regional variation for Period *A* response

This section presents two possible strategies for modelling ASW variation in Period *A* response behaviour. The first method derives models for response mode and response time outcomes in the 2011 Census, while the second method fits a model to the response mode and response time outcomes for the tests. Analysis of the 2013 test data shows further work is needed to derive a model which combines response behaviour information from both the 2011 Census and the tests.

Modelling using 2011 behaviour

Although the 2011 Census data does not provide any measures which are direct analogues for response during Period *A* under the 2016 enumeration model, the 2011 Census data should be useful to identify demographic characteristics related to the probability a dwelling will respond with limited prompting under the new enumeration model. It seems reasonable to assume that demographic characteristics positively associated with internet response or response with minimal follow-up in 2011 will be positively associated with the probability of response during Period *A* in 2016. We can derive models for internet response and response with no follow-up which use 2011 region demographics $X_{1,a},\ldots,X_{l,a}$ as explanatory variables. For example, we could fit a model for the 2011 internet response rate in 2011 Census Collector Workloads[12] (CLWs) w, I_w^{2011}:

$$I_w^{2011} = \alpha_0 + \alpha_1 X_{1,w} + \cdots + \alpha_l X_{l,w} + \varepsilon_a \ .$$

By applying the model to demographic characteristics of the 2016 ASWs a, we have:

$$f_a^A \propto \alpha_0 + \alpha_1 X_{1,a} + \cdots + \alpha_l X_{l,a} \ .$$

Linear regression models have been fit to 2011 Census data aggregated to the CLW geography to assess the quality of predictions for the response behaviour of regions under the 2011 enumeration model. Separate models were fit for the outcome variables ‘proportion of dwellings in the CLW which responded by internet’ and ‘proportion of dwellings in the CLW which responded without any follow-up visits’. Most of the predictor variables were either proportions of dwellings or proportions of persons in the CLW with a particular characteristic, as measured in the 2011 Census.

12 The average size of a CLW was approximately 250 dwellings.

The quality of fit was much higher for the model for the internet response proportion. Further details are given in Appendix *A*.

Response data from the 2013 Test have been analysed to test the assumption that demographic characteristics positively associated with internet response or response with minimal follow-up in 2011 are also positively associated with the probability of self-response for the 2016 enumeration model. The analysis suggested 2011 modelled predictions for the internet response rate or 'no follow-up rate' in 2011 are not proportional to the Period *A* response rate under the new enumeration model. The proportion of returns which were by internet had stronger correlation with ASW demographics than did the proportion of dwellings which respond prior to follow-up.

Care is required in interpreting these analyses performed on 2013 Test data. Firstly, the analyses were based on the small amount of sample data which could be used.[13] Secondly, the high non-response to the test also clouds the results. An illustration of the impact of the non-response is that the 2011 internet response predictions were more strongly correlated to the proportion of the test sample which responded by paper than the proportion of the test sample which responded by internet. If there was complete response, the correlations would be the same (since these rates would sum to 1 in each ASW).

The results of the Major Test in 2014, which will have a much larger sample spread across a broader cross-section of areas, will better indicate the relationship between the modelled estimates from 2011 data and self-response under the new enumeration model. It is worth noting the sample of areas in 2013 did not include areas with characteristics likely to be associated with very low self-response. The sample for the 2014 Test will include areas with characteristics expected to be associated with low internet response and higher follow-up, thus facilitating analysis to identify whether low predicted internet response or other area characteristics are associated with low response during Period *A*.

It should be noted there is scope for data items other than those collected in the 2011 Census to be used as predictors. An example is an indicator of whether the majority of dwellings in the region have broadband access (e.g. through National Broadband Network (NBN) connectivity).

---

13 Data from the mail-out areas of the tests was disregarded because issues which arose for this component likely impacted on respondent behaviour in the early weeks of enumeration of the test.

Modelling response behaviour in tests directly

Modelling the self-response rates in the tests would appear a more direct strategy for modelling the relative response probabilities of ASWs during Period A. The proportion of response during Period A in a test can be modelled by linear regression, again using the 2011 Census data as predictors:

$$\gamma_a^{(A\cdot)} = \delta_0 + \delta_1 X_{1,a} + \cdots + \delta_p X_{l,a} + \varepsilon_a \ .$$

There are some obvious drawbacks of modelling behaviour observed in a test. Firstly, the tests provide much fewer data points than the 2011 Census, so a model derived from test data cannot capture the behaviour of demographic groups with low or no representation in the test samples. Another concern is that there are likely differences between the population characteristics related to self-response for the non-compulsory tests and the actual Census. The regression predicts the response rate parameter $\gamma_a^{(A\cdot)}$, which concerns behaviour <u>under the conditions of a test</u>. As discussed in Section 3.3, the parameters describing behaviour we can predict from the tests, $\gamma_a^{(pm)}$, are different from the parameters of interest, $\theta_a^{(pm)}$, which describe behaviour under Census conditions.

Considering the differences between the $\gamma_a^{(pm)}$ and $\theta_a^{(pm)}$, we need a better strategy than using the $\hat{\gamma}_a^{(A\cdot)}$ as proxies for $\hat{\theta}_a^{(A\cdot)}$ for estimating the relativities between the $\theta_a^{(A\cdot)}$. Using the set of estimates $\hat{\gamma}_a^{(pm)}$ ($p = A, B, C, D$; $m = 1, 2$) should improve the estimation of the relativities. Along with the set of regression estimates $\hat{\gamma}_a^{(pm)}$, assumptions about differences in behaviour under test and Census conditions are required, and these assumptions are in the form of conditional probabilities, $P\left((pm)_c \,\middle|\, (pm)_t\right)$. These probabilities would be incorporated into an algorithm which redistributes the probabilities $\hat{\gamma}_a^{(pm)}$ across classes to give estimates $\hat{\theta}_a^{(pm)}$. Table 4.3 presents a hypothetical example. Relativities between the $\hat{\theta}_a^{(A\cdot)}$ derived in this way provide the factors f_a^A estimating ASW variation for $\theta_a^{(A\cdot)}$, as described in Section 4.1.

Unfortunately there is no data from the testing program of the 2011 Australian Census which can suggest appropriate values for the $P\left((pm)_c \,\middle|\, (pm)_t\right)$. Assumptions about how the 'test non-respondent' class would behave in the actual Census could be informed by analysis of differences between the demographic distributions of the test non-respondents and the overall population. If it was found, for example, that dwellings with demographics associated with higher 2011 internet response were over-represented in the Test non-respondent class, we would expect test non-respondents to be more likely to be internet respondents for the actual Census.

4.3 Illustration of algorithm to derive $\hat{\theta}_a^{(pm)}$ from modelled test predictions $\hat{\gamma}_a^{(pm)}$

Test response class (pm)	$\hat{\gamma}_a^{(pm)}$	$\hat{\theta}_a^{(pm)}$, *Step 1*	$\hat{\theta}_a^{(pm)}$, *Step 2*	$\hat{\theta}_a^{(pm)}$, *Step 3*	$\hat{\theta}_a^{(pm)}$, *Step 4*
*A*1	30.0%	36.8%	36.8%	44.6%	46.7%
*A*2	5.0%	7.9%	7.9%	7.9%	9.9%
*B*1	5.0%	5.0%	7.5%	7.5%	7.5%
*B*2	1.0%	1.0%	1.5%	1.5%	1.5%
*C*1	15.0%	28.6%	26.1%	18.3%	18.3%
*C*2	15.0%	20.7%	20.2%	20.2%	16.2%
D	29.0%				
Total	100%	100%	100%	100%	100%

Example algorithm:
1. Redistribute one third of the non-respondents to Class *A*, two thirds to class *C*, maintaining the mode distribution for the ASW (i.e. $P\left(A\cdot_c \mid D\cdot_t\right)=1/3$, $P\left(C\cdot_c \mid D\cdot_t\right)=2/3$).
2. Increase the Class *B* rates by 50%, drawing the increase from the Class *C* proportions of the same mode.
3. Reduce Class *C*1 by 30%, assigning all to Class *A*1.
4. Reduce Class *C*2 by 20%, assigning half to Class *A*1 and half to Class *A*2.

4.4 Estimating the temporal distribution during Period *A*

The daily distribution of returns prior to the commencement of the follow-up phase in the 2016 Census will likely be very different to the distribution observed in 2011. An indication of the impact of the new enumeration model on the temporal distribution is discussed in Appendix B, which contrasts the timing of internet returns for the 2011 Census and the 2013 Test.

The temporal distribution $T^A(t)$ specifies the cumulative proportion of the $\theta_a^{(A\cdot)}$ responses received by day t of Period *A*. During Period *A* the daily distribution of returns will be heavily dependent on a range of factors including the timing of when dwellings receive their invitation to respond to Census, public reaction to the push for internet response, the timing of Census public relation messages and how these messages encourage *when* forms should be submitted. The impact of the combination of these factors on the distribution of daily returns cannot be ascertained from the test data. Given the issues above, the distribution $T^A(t)$ will need to be based on judgement, combining behaviour observed in the 2011 Australian Census, the tests and possibly the 2011 Canadian Census. The uncertainty surrounding the quality of predictions early in the enumeration period should not have a large impact on operational decision making.[14]

14 Predictions very early in the enumeration period would unlikely be of much practical use since operational managers would not be looking to modify their plans early in the enumeration period. More generally, highly disaggregated early predictions would could overwhelm or serve as a distraction.

4.5 Estimating the parameters for Period *B*

Modelling the response time distribution during Period B

It is proposed to apply a parametric model to describe the response time of dwellings which require no further prompting after the receipt of the targeted reminder letter. The modelling strategy follows Highland (2007), who uses the exponential and Weibull distributions to model the response time of Census internet responses subsequent to a 'stimulating event' to encourage response. The exponential distribution model for response time can be motivated by queuing theory: each dwelling can be thought to put the task 'complete Census form' into a queue of tasks for the dwelling. Under certain assumptions about how the queue of tasks operates, the time for tasks to leave the queue is described by the exponential distribution.

Highland's analysis assessed the exponential model for internet response data from the 2006 Canadian Census and tests conducted in Canada and the United States. The distribution of responses was analysed after stimulating events which not only included explicit reminder prompts but also Census Day and the beginning of a weekend (when households may have more time). The response profile model for the entire enumeration period consists of a sequence of exponential distributions, with each one associated with a particular stimulating event.

There are two important differences between Highland's application of the exponential distribution and its application described here. Firstly, the distribution is used only for Period *B* and during follow-up period (discussed in Section 5). Secondly, in our application the exponential distribution describes the response time distribution among the expected number of dwellings requiring a particular degree of prompting in order to respond.[15] Our application also accounts for the fact that a particular prompt (such as a reminder letter) may occur on different days for different dwellings.

The exponential response-time model for dwellings which require no further prompting after receiving the reminder letter (at day t_{RL}) is:

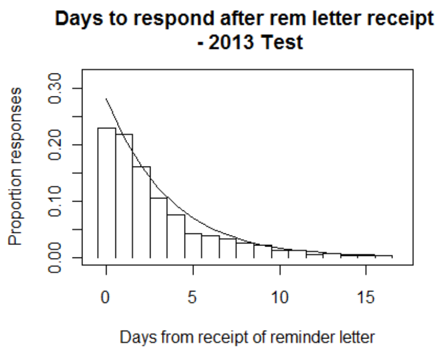
$$f(t - t_{RL}) = \frac{1}{\lambda_{RL}} e^{-\frac{(t - t_{RL})}{\lambda_{RL}}} \quad \text{for day } t \geq t_{RL} . \qquad (1)$$

The mean parameter (the 'decay rate') λ_{RL} is the average response time for dwellings which require no further prompting after receiving the reminder letter.[16]

15 In contrast Highland directly models the total responses received on a particular day.
16 Another interpretation for the mean parameter is the time by which 63.2% respond.

Figure 4.2 shows the exponential model fits very well to the response time of the 2013 Test dwellings which responded online following the receipt of the reminder letter and did not receive a follow-up visit.[17] The vertical axis shows the proportion of this collection of dwellings which responded a particular number of days after receipt of the reminder letter. The estimate of λ_{RL} is 3.56, and for this estimate 77% of responses after the reminder letter prompt are received within four days of the prompt. This suggests the impact of the reminder letter is fairly immediate, and so commencing follow-up soon after reminder letters are dispatched is efficient.

4.2 Response time distribution for dwellings which responded after reminder letter receipt ('Day 0' includes the day of visit and the day following the visit)



The mean parameter λ_{RL} of the exponential distribution fitted in figure 4.2 is a biased estimate for the sample group "test internet respondents who require no further prompt after receiving the reminder letter". This is because there would have been test participants who did not respond by the time of receiving a visit, *but would have responded* without further prompting if given longer. The response times for such respondents are effectively censored response time observations. An iterative procedure could be used to derive a less-biased estimate of λ_{RL} for the response time distribution for the specific group of interest. Further details are given in Appendix C.

Estimation of response time distribution, $T_a^B(t)$

The temporal distribution for the proportion of response in ASW a during Period B, $T_a^B(t)$, will depend on the timing of despatch and delivery of the reminder letters within the ASW. For ASWs in which letters are mailed out, dwellings will typically receive their letter on the same day. In these ASWs the exponential distribution will provide the relative proportion of responses on each day of Period B. In ASWs where reminder letters are delivered by field officers the receipt of letters will be across a small number of days. In these regions estimation of $T_a^B(t)$ will require estimates of the proportion of dwellings to receive their reminder letter on each day of the delivery period and aggregating the distributions for each of the delivery days.

17 Paper returns could not be included in this analysis because the submission date of paper returns in the 2013 Test was not available.

Estimation of Period B response probabilities

It remains to estimate the ASW-level response probabilities during Period B, $\theta_a^{(B\cdot)} = f_a^B \, \theta^{(B\cdot)}$. The national average response probability achieved during Period B, $\theta^{(B\cdot)}$, could be estimated using the method described in Section 4.2 to estimate $\theta^{(A\cdot)}$. Data from the 2014 Test will be analysed to investigate whether there are demographic predictors of response probability during Period B. Of particular interest is whether there is a correlation between an ASW's proportion of returns received during Periods A and B. If such a correlation exists, the factor for the relative response probability in ASW a during Period B, f_a^B, could be modelled as a function of the relative response probability during Period A, f_a^A.

4.6 Estimating distribution of response by mode

So far this section has described estimating the probability of response within Periods A and B irrespective of the mode of response. Denote π_a^p for the probability that a response during Period p ($p = A, B$) in ASW a is by internet. We then have $\theta_a^{(p1)} = f_a^p \, \pi_a^p \, \theta^{(p\cdot)}$.

To provide a starting point to model π_a^p, we assume the propensity to respond online is the same for the test self-respondents and those who will self-respond in the actual Census. This assumption can alternatively be described in terms of the non-respondents to the tests: those who do not respond to the test but self-respond in the Census have the same propensity to respond online as the self-respondents in the test. Considering that the proportions π_a^p could be quite close to 1, a logistic regression model would be more appropriate than a linear regression model:

$$\log\left(\frac{\pi_a^p}{1-\pi_a^p}\right) = \beta_0 + \beta_1 X_{1,a} + \cdots + \beta_l X_{l,a} + \varepsilon_a \,.$$

The model's parameters would be estimated by fitting to the ASW observed rates of internet response in the tests. Again the explanatory variables $X_{1,a}, \ldots, X_{l,a}$ are l demographic characteristics of the ASW (as measured in the 2011 Census).

The results of the Canadian Census could be helpful to indicate the rate of internet response which should be expected. In the regions of Canada subjected to an enumeration model similar to the 2016 Australian model, online returns accounted for more than 85% of responses received before the follow-up period. If the model fitted to the internet response proportions in the test does not provide internet response rates expected for the actual Census, a scaling adjustment could be applied so that the distribution of modelled propensities better aligns with expectations. The adjustment could be as simple as applying a multiplicative factor to the modelled estimates.

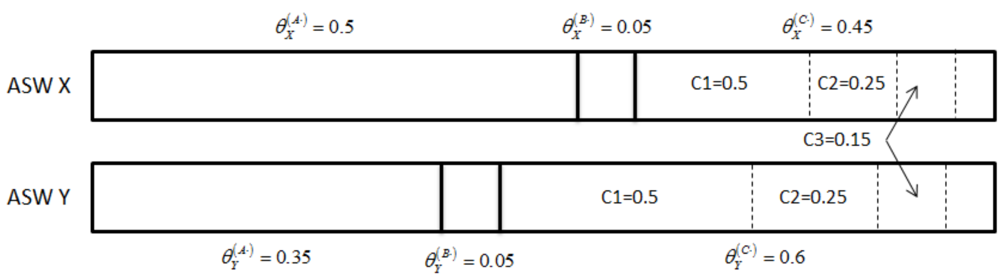
5. MODELLING RESPONSE DURING FOLLOW-UP

5.1 Form of model for follow-up phase

The response model for the follow-up phase, Period C, fits into the overall framework presented in Section 3. The basis of the model for response during Period C is that dwellings will require varying numbers of field officer visits prior to responding. Visits which result in contact with dwelling occupants are much more likely to result in a response, so much of the variation in the number of visits required for dwellings to respond would be attributable to the number of visits required to make first contact. The model for the response time of a dwelling during follow-up has two components. The first component is a probability model describing the number of visits needed in order for the dwelling to respond. The second component is an exponential distribution which provides the response time after the necessary number of visits has been made. The model fits well to the data from the 2013 Test.

The probability model for the number of visits required prior to a dwelling responding provides an estimate of the proportion of Period C dwellings which will respond after a minimum of k visits. This proportion, denoted as $|Ck|$, is relative to the proportion of occupied private dwellings outstanding after Periods A and B. The definition of $|Ck|$ is illustrated in figure 5.1, which compares two ASWs with differing proportions of dwellings responding during Periods A and B. For each $k = 1,\ldots,3$, $|Ck|$ in ASW X relative to $\theta_X^{(C\cdot)}$ is the same as $|Ck|$ in ASW Y relative to $\theta_Y^{(C\cdot)}$.

5.1 Illustration of the definition of $|Ck|$ within two populations with a different distribution of dwellings responding during Periods *A*, *B* and *C*.



The geometric distribution is used to model the required number of visits K for a dwelling:

$$K \sim \text{Geometric}\,(\nu), \text{ which gives } |Ck| = \nu(1-\nu)^{k-1}.$$

Only a single parameter v_a is needed to determine $|Ck_a|$ for each ASW. The proportion of Period C dwellings in ASW a which respond on day t of the follow-up period, r_a^C, is given by:

$$r_a^C(t) = \begin{cases} \dfrac{|C1|}{\lambda_{C1}} & \text{if } t = 0\,; \\ \alpha_{k,a}\dfrac{1}{\lambda_{Ck}}e^{-\frac{(t-t_k)}{\lambda_{Ck}}} & \text{if } t > 0,\ \ k = \min\limits_{j} \text{ such that } t \geq t_j\,. \end{cases}$$

where:

- k is the visit number of the most recent visit at day t under the follow-up strategy in ASW a. Visit $k = 1$ is assumed to occur on day 0.
- t_k denotes the day on which visit k occurs under the follow-up strategy.
- $\alpha_{k,a}$ is a scaling parameter, and is the parameter through which the response time distribution is linked to the model specifying the required number of visits. $\alpha_{k,a}$ approaches $|Ck_a|$ as the time gap between visits increases. See Appendix C.2 for more details.
- λ_{Ck} is the mean of the exponential density function for response following the k-th visit (assumed to be the same across all ASWs a).

Implications of model on overall response-time distribution

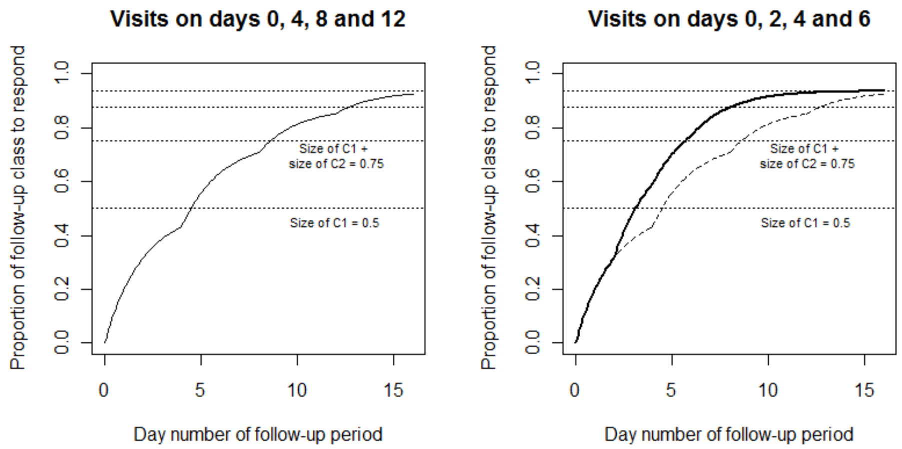
The overall response-time distribution is heavily dependent on the frequency of the visits specified by the follow-up strategy. To illustrate, figure 5.2 compares the overall response-time distribution for follow-up strategies when dwellings are visited every second and every fourth day.[18] In this example it is assumed $v = 0.5$. The strategy which visits dwellings every second day involves significantly more second visits to dwellings which require only one visit to respond. The cost associated with a faster return rate is the additional visits made to dwellings which would have responded without further prompting.

This model illustration highlights how estimates of both the λ_{Ck} and v should inform the follow-up strategy. In the above example the strategy which visits dwellings every second day achieves a 94% follow-up response rate around five days earlier. However, this strategy would require significantly more resources for two reasons: (1) more visits need to be conducted in total (due to more visits to dwellings which would have responded without further prompting), and (2) a larger field force would be needed to conduct the required visits in the shorter time period.

---

18 To simplify the presentation, it is assumed all dwellings receive their k^{th} visit on the same day.

5.2 Illustration of relationship between follow-up strategy and response-time distribution



5.2 Parameter estimation

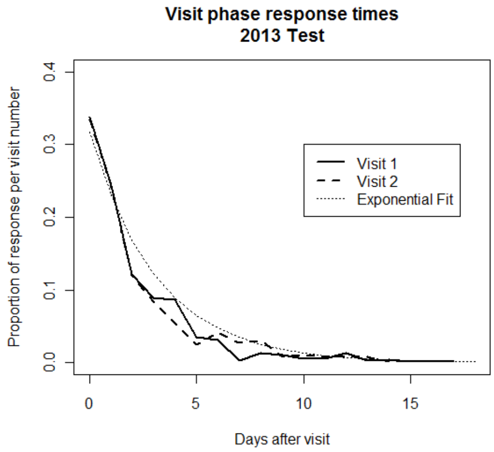
The proportion of dwellings in ASW a which have not responded by the start of follow-up, $\theta_a^{(C\cdot)}$, is estimated by $\hat{\theta}_a^{(C\cdot)} = 1 - \hat{\theta}_a^{(A\cdot)} - \hat{\theta}_a^{(B\cdot)}$. The response time distribution for responses during Period C is determined by the follow-up strategy in ASW a, the exponential mean parameters λ_{Ck} and the relative sizes of the Ck_a, denoted $|Ck_a|$.

Estimation of exponential mean parameters

An exponential distribution fits well to the response-time distributions after the first and second visits made to internet respondents in the 2013 Test (figure 5.3). There was no significant difference between the estimates of the mean parameters λ_{C1} and λ_{C2} fit separately to the distributions relating to the first and second visits. The similarity of the distributions following the first and second visits is consistent with the analysis in Highland (2007), which identified a similar shape for the response-time distributions across a variety of 'stimulating events' during an enumeration period.

The above analysis will be repeated for data from the 2014 Test. If the results from the 2014 Test are similar to those observed in the 2013 Test, a consistent exponential mean parameter will be used after each visit number (i.e. $\lambda_{Ck} = \lambda_C$). Estimation of λ_C should apply the iterative procedure of Appendix C.1 so that the estimate accounts for the response times of relevant 'censored' dwellings in the test. It could be argued that it is reasonable to assume the 2014 Test estimate of λ_C will be appropriate to apply for modelling behaviour in the actual Census. Given that a dwelling is committed to respond after the k-th visit, the factors which determine the time needed to submit a response after a visit are the same for when the response is for a voluntary test or the actual Census.

5.3 Response time distributions for Visit phase respondents in the 2013 Test



Estimation of geometric distribution parameter

There is empirical evidence supporting the geometric distribution for the required number of visits for a dwelling. The geometric model implies an exponential-like decay for $|Ck|$ and hence a constant value for $|Ck|$ relative to the proportion of follow-up dwellings which require at least k visits. For example,

$$|C1| = v = \frac{|C2|}{1-|C1|} = \frac{|C3|}{1-|C1|-|C2|}, \text{ etc.} \qquad (2)$$

Highland's analysis of internet responses to the 2006 Canadian Census showed the response impact of successive events decayed in approximately exponential fashion. In the 2011 Australian Census there was an approximate exponential decay in the amount of returns generated by visit numbers 2, 3 and 4 among dwellings which received at least one visit during follow-up (see Appendix D for further details).

The parameter v will be estimated from the Period C respondents in the 2014 Test. The 2014 Test data will be analysed to determine the improvement in model fit if v is allowed to vary between geographic regions. Regions would be expected to display different response behaviour during follow-up if their Period C respondents vary significantly with respect to their willingness to participate and capability to participate unassisted. For example, the first visit could have above-average impact on response in regions containing a high proportion of dwellings with 'high willingness' but 'low capability to respond unassisted'. For these dwellings, the offer of a paper form when the field officer visits is likely have an above-average impact on eliciting response.

Identifying predictors to classify ASWs by ‘willingness to participate’ could be useful for choosing the value of ν_a for each ASW. A model for non-response or high levels of follow-up in the 2011 Census could be used to classify ASWs by their average ‘willingness’ to respond, based on their 2011 Census demographic characteristics. ASWs would be assigned into a small number of ‘Hard-to-Count’ classes, and a single value for ν_a is estimated for each class. A smaller value for ν_a would be assigned to the regions in the most difficult ‘Hard-to-Count’ class. This strategy is similar to that applied by the ONS, and avoids estimating the ν parameter for individual ASWs.

6. CONCLUSION

This paper has presented a Census response modelling framework focused on supporting prediction of response behaviour for fine geographic regions. Analysis of data from the 2013 Census test shows some aspects of response behaviour can be modelled better than others. For example, the 2013 data showed the proportion of returns which were by internet had stronger correlation with ASW demographics than did the proportion of dwellings which respond prior to follow-up. An encouraging outcome from the analyses of data from the 2013 test was the quality of fit of the exponential distribution for describing the response time distribution for responses which follow targeted prompts. The fit provides confidence in using the exponential response time model to inform decisions about follow-up procedures.

The paper has highlighted the need for assumptions about response behaviour which are difficult to validate from the data available. Under the proposed framework, some assumptions are described in terms of how aggregate behaviour observed in the tests will differ between the environments of a test and the actual Census.

The paper has noted further modelling work which will be undertaken using data from the Major Test in August 2014. Areas for further work which have not been addressed in this paper include:

- quantifying uncertainty of the various predictions;
- incorporating the response rate impact of localised campaigns; and
- updating predictions as 'live' response rate data becomes available during enumeration.

REFERENCES

Australian Bureau of Statistics (2012) *Census of Population and Housing: Consultation on Content and Procedures, 2016*, cat. no. 2007.0, ABS, Canberra.

Bates, N. and Mulry, M. (2011) "Using a Geographic Segmentation to Understand, Predict and Plan for Census and Survey Mail Nonresponse", *Journal of Official Statistics*, 27(4), pp. 601–618.

Highland, F. (2007) "Time Series Analysis of Census Internet Response", *Proceedings of the ASA Survey Research Methods Section*, American Statistical Association, pp. 2804–2810.

Hopper, N.A. (2012) "Predicting Patterns of Household Non-Response in the 2011 Census", *Survey Methodology Bulletin*, 69, pp. 9–22.

Office of National Statistics (2012) *2011 Census Field Operations*, ONS 2011 Census Evaluation Report, ONS, London.

Statistics Canada (2012) *Internet Data Collection in the Canadian Census of Population*, Conference of European Statisticians, 13 March 2012.

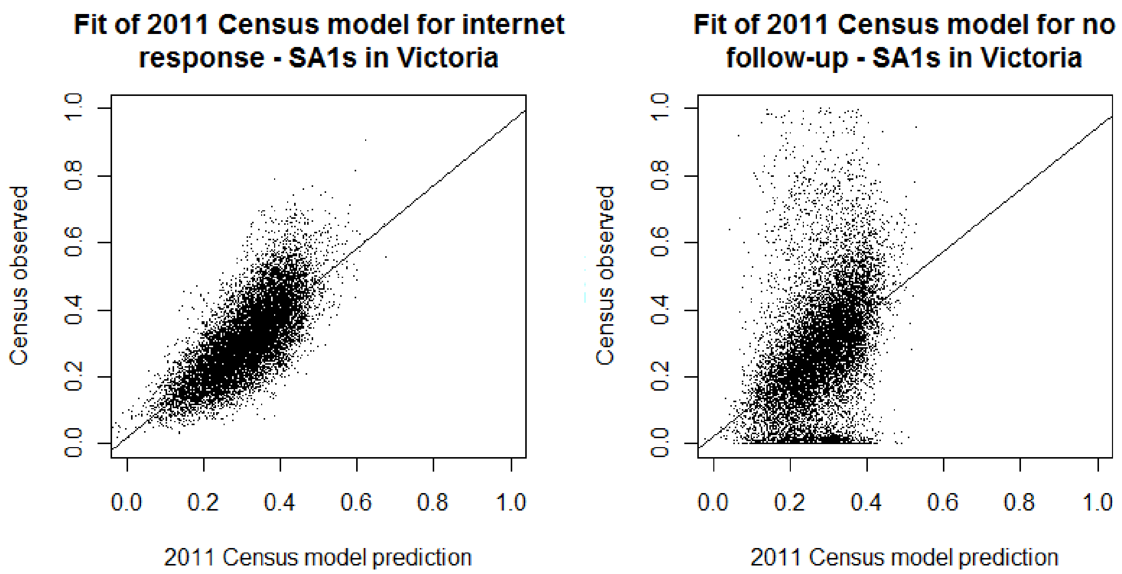
Townsend, N. (2011) "2011 Census Field Design: Getting the Numbers Right", *Population Trends*, 143, pp. 22–31.

APPENDIXES

A. MODELS FIT TO 2011 CENSUS DATA

Linear regression models were fit to 2011 Census data aggregated to CLW level for the outcome variables 'proportion of dwellings in the region which responded by internet' and 'proportion of dwellings in the region which responded without any follow-up visits'. Since the models were fit by linear regression, the model parameters can be applied to produce predictions for larger or smaller geographic aggregations than the CLW level. All of the explanatory variables are listed in table A.2. The fit of the model predictions to the rates actually observed in Victoria in the 2011 Census for the Statistical Area Level 1 geographic classification is shown in figure A.1. Further work is required to investigate a potential quality issue with count variable for the number of follow-up visits. This issue could explain the large number of SA1s with very low and very high proportions of dwellings with zero follow-up visits.

A.1 Fit of 2011 Census model predictions for a region's 2011 internet response rate (left) and 2011 proportion for which no follow-up is required (right) and (R^2 values are 0.55 and 0.18)



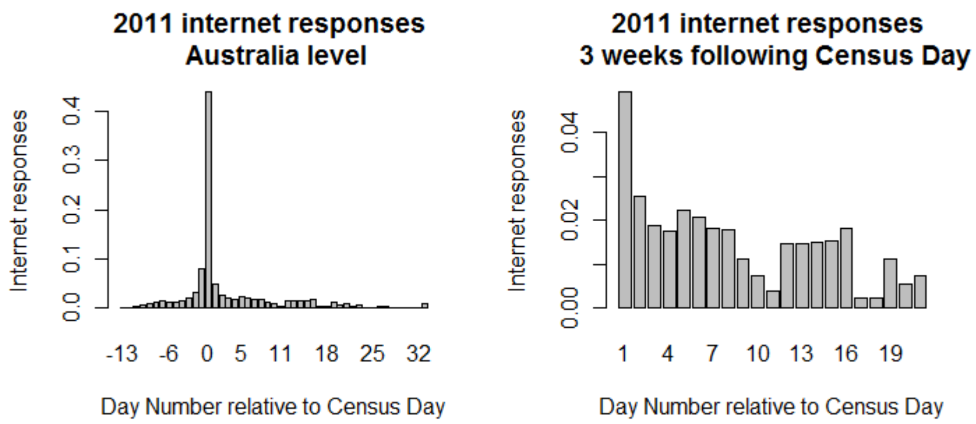
A.2 Explanatory variables for models predicting a region's 2011 Census internet response proportion and proportion of dwellings requiring no follow-up

Variable	*Description*
State	Indicator for each State (8 classes)
Remoteness	Indicator for area remoteness class (3 classes)
Age	Proportion of persons in region aged under specified age cut-off (8 classes)
Average HH size	Average number of persons per occupied private dwelling
School completion	Proportion of persons who completed high school
Non-standard dwelling	Proportion of dwellings which are in caravan parks or camping grounds marinas, home estates or retirement villages
Income class	Proportion of households belonging to a particular household income class (9 classes)
ATSI proportion	Proportion of Aboriginal or Torres Strait Islander peoples
Dwellings in high-rise	Proportion of dwellings in buildings with at least three storeys
Private dwelling proportion	Proportion of dwellings which are private dwellings
Language not English	Proportion of persons who speak a language other than English at home
Female proportion	Proportion of persons who are female
Lone person dwellings	Proportion of households which are a single-person household
Elderly person households	Proportion of households in which all persons are aged at least 65

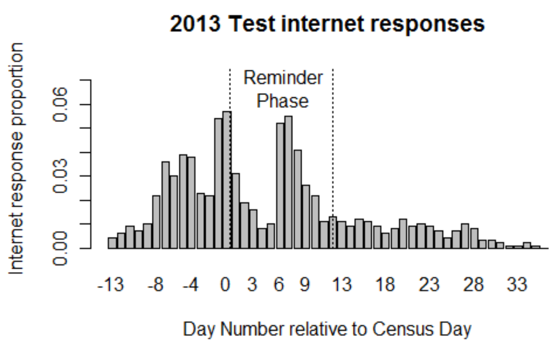
B. TIMING OF INTERNET RESPONSES

Figures B.1 and B.2 contrast the daily response counts of internet returns for the 2011 Australian Census and the 2013 Test. The right plot in figure B.1 presents only the data for the three weeks following the 2011 Census Day. The new 'phase-based' enumeration model is expected to produce a result substantially different to that observed in 2011. However, the distribution observed in the 2013 Test is not indicative of the distribution expected for the 2016 Census, as the public relations exercises and media coverage associated with the actual Census should result in a much higher proportion of responses being received on Census Day and the days immediately surrounding.

B.1 Timing of internet responses for the 2011 Australian Census



B.2: Timing of internet responses for the 2013 Test (Percentages are proportion of total internet response)



C. PARAMETER ESTIMATION

C.1 Iterative procedure to estimate λ_{RL} from test data

We would like to use data from the tests to estimate the exponential mean parameter λ_{RL} in the response time for the sample group, "internet respondents who require no further prompt after receiving the follow-up letter". Some test dwellings belonging to this hypothetical group are censored observations because they *would have responded* without further prompting if given more time. Ignoring such dwellings biases the estimate of λ_{RL} for the sample group of interest, so an iterative procedure which attempts to estimate the number of censored observations is proposed.

The value of λ_{RL} estimated from the fit to just the pre-visit respondents (denoted $\lambda_{RL}^{(0)}$) provides initial estimates of the number of dwellings in the test sample censored at each day after receipt of the reminder letter. Refitting the distribution with the censored dwellings counts added provides an updated estimate, $\lambda_{RL}^{(1)}$. Next, $\lambda_{RL}^{(1)}$ could be used to revise the estimates of the number of test sample dwellings censored, and several iterations could be performed. Accounting for the censored observations for the 2013 Test data has only a minor impact on the estimate of λ_{RL}. Since the fitted value of λ_{RL} gives a sharp decay rate, the time gap between the reminder letter and the first visit meant most dwellings which would have responded without further prompt had done so by the time of the first visit.

C.2 Scaling parameter for response distribution during follow-up

In Section 5.1, the model for the proportion of the follow-up responses received on day t of the follow-up phase was given by:

$$r^C(t) = \begin{cases} \dfrac{|C1|}{\lambda_{C1}} & \text{if } t = 0\,; \\ \alpha_k \dfrac{1}{\lambda_{Ck}} e^{-\frac{(t-t_k)}{\lambda_{Ck}}} & \text{if } t > 0,\ \ k = \min\limits_j \ \text{such that } t \geq t_j\,. \end{cases}$$

where:

- k is the visit number of the most recent visit at day t under the follow-up strategy, assuming $k = 1$ at day 0;
- t_k denotes the day on which visit k occurred;
- λ_{Ck} is the mean of exponential density function for response following visit k;
- α_k is a scaling parameter.

Recall that $|Ck|$ denotes the relative size of subclass Ck within the Period C dwellings. The scaling parameter α_k can be approximated by relative size of $|Ck|$ if the follow-up strategy is such that visit number $k+1$ occurs after most responses are received from dwellings in the subclass Ck. If the follow-up strategy involves conducting visits with high frequency, α_k should be inflated so that it accounts for outstanding dwellings which would have responded prior to visit k without further prompting, if given more time before their subsequent visits.

The parameter α_k can be written as:

$$\alpha_k = \begin{cases} |C1| & \text{for } k = 1; \\ |C1| + \left(\sum_{j<k} |Cj| - \sum_{s<t_k} r^C(s) \right) & \text{for } k > 1. \end{cases}$$

D. DISTRIBUTION OF 2011 CENSUS RESPONSE BY FOLLOW-UP VISITS

Table D.1 presents the distribution of responses received in the 2011 Australian Census by the number of collector visits (dwellings requiring no visits are not shown). The right-most column expresses the number of responses from dwellings with k visits as a proportion of responses from dwellings which had at least k visits. It is a measure of the relative impact of each visit since it shows the amount of response at a visit number as a proportion of the outstanding dwellings which responded. The data suggests visits 2, 3 and 4 had a similar impact, each visit number approximately halving the number of the responses which were ultimately received during follow-up.

From these data, it seems reasonable to assume the relative sizes of the subclasses C_k in the follow-up model of Section 5 decay in an exponential fashion (and hence have the relationship given in (2) in Section 5.2). There is a downward trend in the proportions in these data, and a conservative approach to estimating required field effort would be to assume such a trend also applies for the 2016 Census. In this case, the decay in the $|C_k|$ would be slower than exponential.

D.1 2011 Census responses received by number of collector visits

Number of collector visits, k	*Number of responses as a proportionof total responses from dwellings with at least one visit during follow-up*	*Number of responses as a proportion of responses from dwellings which received at least k visits during follow-up*
1	0.69	0.69
2	0.18	0.58
3	0.07	0.50
4	0.03	0.47
5+	0.03	NA
Total	1.00	

E. SUMMARY INFORMATION

E.1 Summary of time periods for modelling

Time period	*Description*	*Section discussed*
A	Time period before reminder letters are first received	4.1–4.4, 4.6
B	Time period between when reminder letters are first received and the last day of the Reminder phase	4.5, 4.6
C	Time period after last day of Reminder phase	5.1–5.2

E.2 Summary of parameters

Parameter	*Description*	*Section discussed*
$\theta^{(A\cdot)}$	Probability of a dwelling responding during Period *A* (at the national level). Also can be interpreted as the response rate at the end of Period *A*.	4.2
$\theta_a^{(A\cdot)}$	Probability of a dwelling in ASW a responding during Period *A*.	4.3
$\theta^{(B\cdot)}$	Probability of a dwelling responding during Period *B* (at the national level). Also can be interpreted as the change in the national response rate during Period *B*.	4.5
$\theta_a^{(B\cdot)}$	Probability of a dwelling in ASW a responding during Period *B*.	4.5
$\theta^{(C\cdot)}$	Probability of a dwelling responding during Period *C* (at the national level). Also can be interpreted as the change in response rate during Period *C*. Can be derived by sum of $\theta_a^{(C\cdot)}$.	–
$\theta_a^{(C\cdot)}$	Probability of a dwelling in ASW a responding during Period *C*. Derived from complement of $\left(\theta_a^{(A\cdot)} + \theta_a^{(B\cdot)}\right)$.	5.2
$T^A(t)$	Cumulative proportion of Period *A* response attained at day t of Period *A*. Same distribution applies to all ASWs.	4.4
$T^B(t)$	Cumulative proportion of Period *B* response which is attained at day *t* of Period *B* (at national level). Not estimated explicitly (could be derived from $T_a^B(t)$ and $\theta_a^{(B\cdot)}$).	–
$T_a^B(t)$	Cumulative proportion of Period *B* response in ASW a which is attained at day t of Period *B*.	4.5
λ_{RL}	The mean parameter of the exponential distribution describing response time of Period *B* respondents (response time measured from the day of receipt of reminder letter).	4.5
λ_{Ck}	The mean parameter of the exponential distribution describing response time of Period *C* respondents requiring k visits (response time measured from time of visit k).	5.1–5.2
ν	Parameter for Geometric distribution, specifying the probability a dwelling will respond after receiving visit k, given it has not already responded.	5.1–5.2
$\vert Ck_a \vert$	Proportion of the Period *C* respondents in ASW a requiring at least k visits to respond.	5.1–5.2

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.

LIBRARY A range of ABS publications are available from public and tertiary libraries Australia wide. Contact your nearest library to determine whether it has the ABS statistics you require, or visit our website for a list of libraries.

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, or purchase a hard copy publication. 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