Australian Bureau of Statistics

1352.0.55.096

Research Paper

Sample Design Issues for National Surveys of the Indigenous Population

www.abs.gov.au

New Issue

Research Paper

Sample Design Issues for National Surveys of the Indigenous Population

Geoffrey Brent and Alistair Rogers

Statistical Services Branch

Methodology Advisory Committee

13 June 2008, Canberra

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) THURS 04 SEP 2008

ABS Catalogue no. 1352.0.55.096

Views expressed in this paper are those of the author(s), and do not necessarily represent those of the Australian Bureau of Statistics.
Where quoted, they should be attributed clearly to the author(s).

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 Alistair Rogers, Statistical Services Branch on Canberra (02) 6252 7334 or email <statistical.services@abs.gov.au>.

CONTENTS

The role of the Methodology Advisory Committee (MAC) is to review and direct research into the collection, estimation, dissemination and analytical methodologies associated with ABS statistics. Papers presented to the MAC are often in the early stages of development, and therefore do not represent the considered views of the Australian Bureau of Statistics or the members of the Committee. Readers interested in the subsequent development of a research topic are encouraged to contact either the author or the Australian Bureau of Statistics.

APPENDIXES

SAMPLE DESIGN ISSUES FOR NATIONAL SURVEYS OF THE INDIGENOUS POPULATION

Geoffrey Brent and Alistair Rogers
Statistical Services Branch

ABSTRACT

ABS surveys of the Aboriginal and Torres Strait Islander population are both complex and expensive due to high levels of screening in non-remote areas. The 2008 National Aboriginal and Torres Strait Islander Social Survey is using a new geographical unit, the meshblock. This paper outlines the new meshblock design and investigates the accuracy implications of meshblock-level sampling, concentrating on the effects of migration between Census and survey.

We show that a meshblock-based unbiased design achieves some reduction in screening compared to census collection district based designs used in the past, but this reduction is limited by the timeliness of the frame data. Further reductions in screening require accepting an increased level of undercoverage.

Disclaimer

Cost and sample data are presented for illustrative purposes only and should be treated as approximations. All cost figures given here are based on a simple model used to determine a cost-effective allocation of sample. In practice, costs are then re-estimated based on a more detailed analysis of the sample, including considerations such as interviewer capacity, recruitment, training, and other field management costs. Sample allocations given here have been generated to provide a comparison of certain design options, and are not intended as exact representations of the full NATSISS design.

1. INTRODUCTION

Statistics on the Aboriginal and Torres Strait Islander population are an important part of the ABS' objectives. The ABS produces such statistics from the Census, Demographic population estimates, the Labour force survey, and special household surveys. In 1994 the ABS conducted the National Aboriginal and Torres Strait Islander Survey (NATSIS). This landmark survey was conducted following recommendations from the 1987–1991 Royal Commission into Aboriginal Deaths in Custody that identified the need for regular monitoring of the social conditions of Indigenous people through statistical data collection.

The NATSIS was the first large scale probability-based survey of the Indigenous population conducted in Australia. For some time it was the only existing source of statistically representative data across a wide range of subject matter on the Indigenous population. It remains a key benchmark data source.

Since the conduct of the NATSIS the ABS has committed to a rolling program of similarly sized household surveys of the Aboriginal and Torres Strait Islander population. This includes a six-yearly National Aboriginal and Torres Strait Islander Social Survey (NATSISS) that commenced in 2002, and a six-yearly National Aboriginal and Torres Strait Islander Health Survey (NATSIHS) that commenced in 2004/05. The surveys aim to produce useable statistics at national, state and remoteness levels with a sample size of around 12,000 fully responding persons.

The NATSISS collects data across a broad range of topics including employment, education, health, transport, crime and victimisation, income and social support networks. The demand for social statistical data on the Indigenous population remains high.

The basic sampling approach for special Indigenous surveys has been reasonably bedded down through development of the 2002 NATSISS and the 2004/05 NATSIHS designs. The paper outlines a broad overview of the basic sample design and selection features used in remote Indigenous communities and non-remote areas. The paper then discusses in some detail the incorporation of the newly available 'meshblock' geographic unit into the recently completed 2008 NATSISS sample design.

It is important to recognise that as with any survey the sample design is only one component that determines the success or otherwise of the survey. This is especially the case with surveys of the Indigenous population. There are a myriad of complexities that need to be addressed in all aspects of survey development and field enumeration. Not least of these is effective engagement with the Indigenous population. Without a successful engagement strategy survey quality will be compromised.

2. INDIGENOUS POPULATION: AN OVERVIEW

Indigenous Australians (Aboriginal and/or Torres Strait Islander) are a rare population, comprising about 2.3% of respondents to the 2006 Census and even less at household level. Table 2.1 provides a breakdown of the Indigenous population by state and remoteness.

At regional levels, many Indigenous populations can be summarised as either

- geographically clustered and relatively inaccessible, or
- relatively accessible but geographically disperse.

The relative emphasis of these attributes varies noticeably by state and territory, with the Northern Territory and Victoria representing extremes in distribution.

24% of Indigenous Australians (as compared to less than 3% of the total Australian population) live in ‘remote’ areas (including ‘very remote’, i.e. RA 3–4). For the Northern Territory the remote proportion is 80%. Surveying remote Indigenous populations is time-consuming and expensive due both to the travel involved and conditions on site (e.g. language difficulties, dust and heat interfering with computer-assisted interviewing). Cultural sensitivities may require a specialised approach (e.g. obtaining permission from community leaders before entering the area).

The 76% of Indigenous people who live in non-remote areas (RA 0–2) are also difficult to survey. Here the problem is *finding* Indigenous people. With a few exceptions (e.g. high-Indigenous-density regions of Darwin), non-remote Indigenous populations are scattered sparsely amongst large non-Indigenous populations.

The standard primary sample unit for ABS household surveys is the Census Collection District (CD), with a typical non-remote CD containing around 200 households. CD-level counts of the number of people who self identified as being of Aboriginal and/or Torres Strait Islander origin on Census night 2006 are available. Address information is not available, however. This has major implications for the sampling method: screening.

We define an ‘Indigenous household’ as one that contains at least one Indigenous person and use ‘size’ to indicate the number of Indigenous households in a CD, as a rough indicator of Indigenous population density. For instance, a ‘size-2 CD’ would have two Indigenous households, comprising about 2/200=1% of all households in the CD.

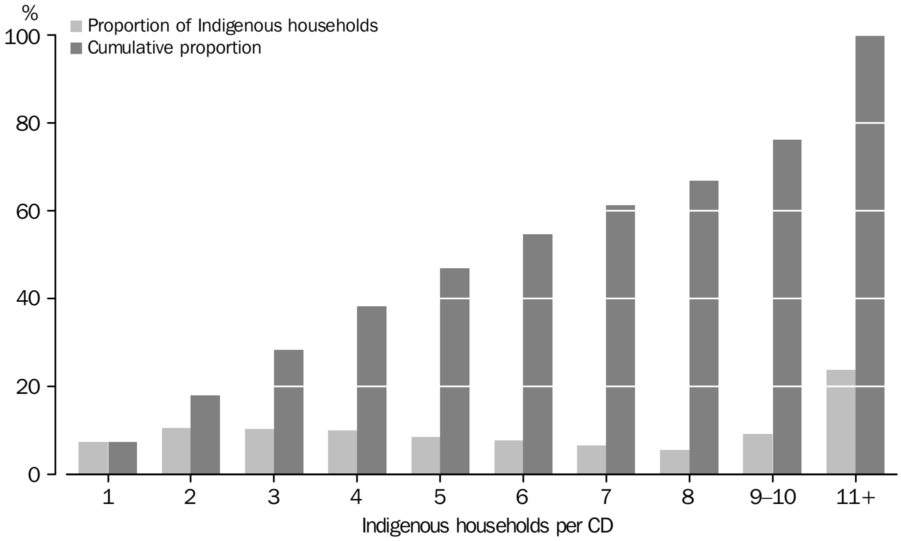
2.1 Australian and Indigenous population, by State and Remoteness area

	Population estimates		*Indigenous population as a proportion of*		
State and Remoteness area	*Total*	*Indigenous*	*Regional population*	*State Indigenous population*	*National Indigenous population*
New South Wales					
0–2	6,502,547	131,029	2.0%	94.9%	28.9%
3–4	36,412	7,038	19.3%	5.1%	1.6%
Total	*6,538,959*	*138,067*	*2.1%*		*30.5%*
Victoria					
0–2	4,921,359	30,011	0.6%	99.9%	6.6%
3	4,622	40	0.9%	0.1%	0.0%
Total	*4,925,981*	*30,051*	*0.6%*		*6.6%*
Queensland					
0–2	3,765,754	98,747	2.6%	77.7%	21.8%
3–4	125,874	28,327	22.5%	22.3%	6.3%
Total	*3,891,628*	*127,074*	*3.3%*		*28.0%*
South Australia					
0–2	1,455,103	20,705	1.4%	81.3%	4.6%
3–4	56,608	4,757	8.4%	18.7%	1.1%
Total	*1,511,711*	*25,462*	*1.7%*		*5.6%*
Western Australia					
0–2	1,823,757	34,132	1.9%	58.4%	7.5%
3–4	129,004	24,349	18.9%	41.6%	5.4%
Total	*1,952,761*	*58,481*	*3.0%*		*12.9%*
Tasmania					
1–2	465,737	16,126	3.5%	96.4%	3.6%
3–4	9,784	595	6.1%	3.6%	0.1%
Total	*475,521*	*16,721*	*3.5%*		*3.7%*
Northern Territory					
NT 2	106,905	10,458	9.8%	19.6%	2.3%
NT 3–4	83,992	43,036	51.2%	80.5%	9.5%
Total	*190,897*	*53,494*	*28.0%*		*11.8%*
Aust. Capital Territory					
Total	*323,328*	*3,845*	*1.2%*		*0.9%*
AUSTRALIA					
RA 0	13,584,496	147,289	1.1%		32.5%
RA 1	3,910,072	99,107	2.5%		21.9%
RA 2	1,869,922	98,657	5.3%		21.8%
RA 3	294,690	39,411	13.4%		8.7%
RA 4	151,606	68,731	45.3%		15.2%
Total	*19,810,786*	*453,195*	*2.3%*		*100.0%*

Source: Census 2006 Community Profiles. Figures and percentages are based on Census counts and exclude migratory and non-responding persons; persons who did not indicate whether they were Indigenous have here been counted as non-Indigenous. Remoteness area (RA) is defined according to the Australian Standard Geographical Classification: RA 0 = Major Urban, RA 1 = Inner Regional, RA 2 = Outer Regional, RA 3 = Remote, RA 4 = Very Remote, RA 5 = Migratory.

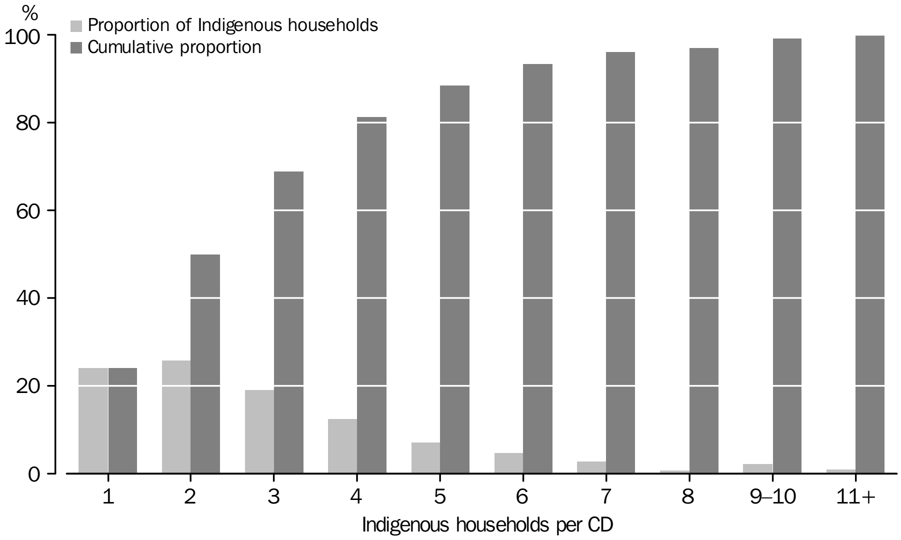
Across Australia, approximately 18% of 'major urban' Indigenous households were found in CDs of size-2 or less:

2.2 Indigenous population distribution, by CD density, Australia RA



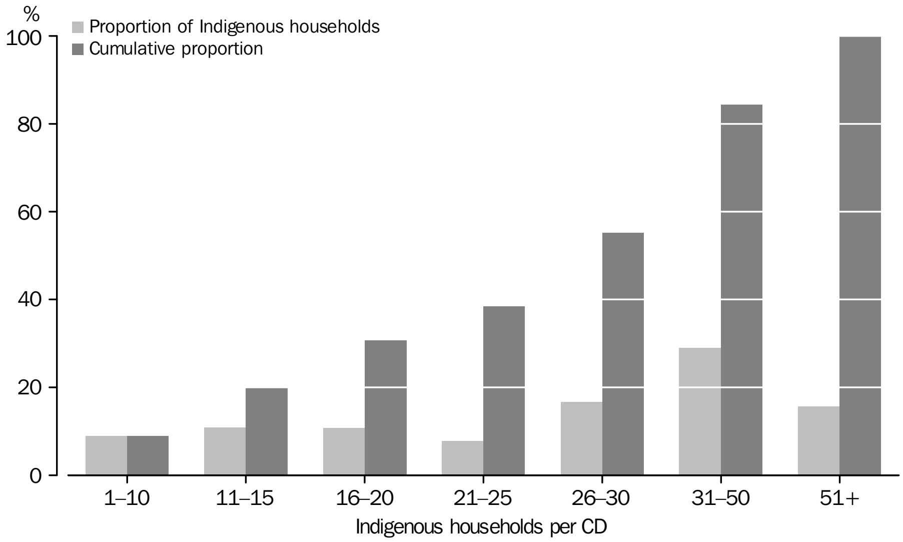
Victoria's Indigenous population is sparser than the national average, with 50% of major urban Indigenous households found in CDs of size 1–2:

2.3 Indigenous population distribution, by CD density, Victoria RA 0



By contrast, in Outer Regional Northern Territory (the least remote category within the Northern Territory, predominantly Darwin), more than 60% of Indigenous households were in size-26+ CDs:

2.4 Indigenous population distribution, by CD density, Northern Territory RA 2



This makes it relatively easy to find Indigenous people in the Northern Territory, but for other states – and especially Victoria – urban areas present a problem. When non-response was factored in, early versions of the NATSISS '08 design for urban Victoria would have required screening 111,589 households in order to interview just 1136 respondents.

The other major issues to consider for sample design are migration and identification. NATSISS enumeration will begin two years after Census '06. We anticipate that in this time, 20–30% of non-remote Indigenous households will have moved. Further, the ABS defines 'Indigenous' status by self-identification. As a result, the numbers and distribution of Indigenous people are affected not only by actual births, deaths, and migration, but also by considerations of identity. In the decade between 1991 and 2001, Indigenous population estimates rose by 62%, due largely to Census respondents' increased awareness of and willingness to acknowledge Indigenous origins. As with migration, this makes it difficult to obtain accurate frame information for the Indigenous population.

3. BASE SAMPLE DESIGN

ABS household surveys use the monthly population survey framework as a 'sampling vehicle'. The standard household form now includes an Aboriginal and Torres Strait Islander identification question, allowing production of estimates for the Indigenous population. However the Indigenous sample take from most household surveys is no more than 1%–2% of the total sample take, making it difficult to produce accurate Indigenous statistics from such data. To support a large scale survey of the Indigenous population a separate sampling framework is required.

3.1 Remote community component

3.1.1 Indigenous community frame

The bulk of the remote Indigenous population resides in Indigenous communities. The ABS' *Community and Housing Infrastructure Needs Survey* (CHINS) is a census of known discrete Indigenous communities with data collected both from administrative Indigenous Housing Organisations and from communities themselves. Recent cycles of the survey have been conducted at the same time as preparation for the census.

Discrete Indigenous communities are defined for the purposes of CHINS as geographic locations, bounded by physical or cadastral (legal) boundaries, that are inhabited or intended to be inhabited predominantly by Indigenous people, with housing or infrastructure that is either owned or managed on a community basis. There are roughly 400 communities and 1300 outstations on the CHINS frame and most of these are in remote areas.

The survey collects data about housing and infrastructure of communities and associated outstations (small groups of dwellings with a wide range of possible uses, e.g. used for hunting purposes through to permanent residences). The population resident in these communities are primarily of Aboriginal and/or Torres Strait Islander origin.

In the early stages of Indigenous survey sampling development the CHINS was considered an obvious source of data for sample frame development if not a frame in itself. There are advantages to using a community based approach compared to a purely census collection district based approach (see Section 3.2) including:

- better representation of Indigenous persons in remote areas, particularly those resident in small communities and outstations;

- the enabling of improved operational procedures to counter the problems with screening CDs in these areas; and
- a mechanism for better monitoring respondent load and implementation of sample overlap control procedures for Indigenous surveys at the community level.

The net outcome from investigations into the use of CHINS data for sampling purposes was the development of the Indigenous Community Frame (ICF). The ICF is not only used for Indigenous surveys but since 2001 has formed a core component of the Monthly Population Survey (MPS) framework to better enable coordination of all ABS survey activity in Indigenous communities.

The ICF is constructed from CHINS and census data. Census collection districts which meet certain criteria are sequestered and for those areas the ICF becomes the primary sampling frame. The criteria for CD inclusion in ICF strata include a predominantly Indigenous population, an assessment that tailored field data collection procedures will be required, and sufficient population to form an MPS stratum in certain areas.

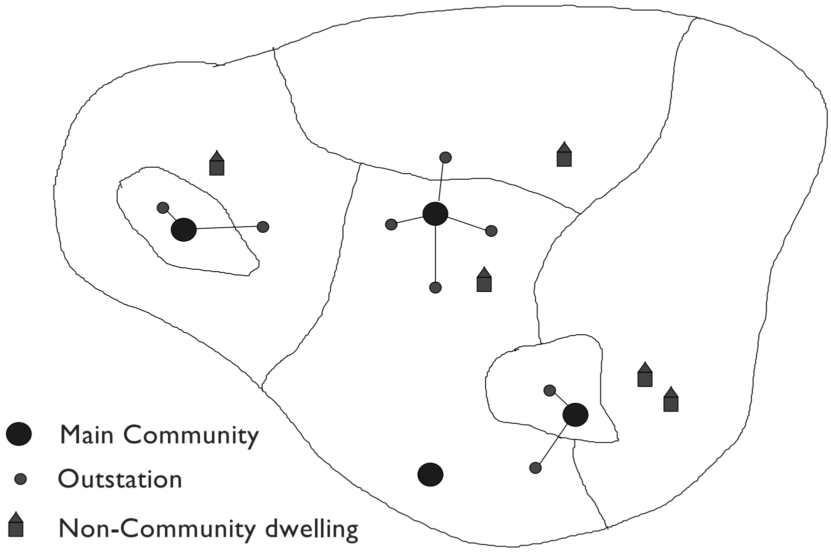
The ICF currently consists of remote communities in Queensland, South Australia, Western Australia and the Northern Territory. There are numerous communities outside these areas however they do not meet the CD inclusion criteria e.g. special enumeration procedures are not necessary, insufficient population to form an MPS stratum, etc..

Communities and outstations in CDs flagged for ICF development are formed into community groups or 'sets' compatible with the preferred selection mechanism. The community groups form the population of primary sampling units from which sample is drawn. The key to set formation is the linking of smaller outstations and communities to larger communities so as to

- represent the population resident in such communities;
- reduce the wide distribution of size measures associated with PSUs; and
- accommodate field enumeration protocol, where contact with a 'main' or 'parent' community is required prior to visiting a selected outstation, in as cost effective a manner as possible.

The diagram below depicts an example of community set formation and possible relationships with CD boundaries.

3.1 Community / outstation links



Non-community dwellings located within ICF CDs, e.g. cattle stations, are incorporated into the frame and selection procedures for the monthly population survey but are excluded from coverage for Indigenous surveys for pragmatic reasons. The level of undercoverage from this exclusion is minimal. From the approximately 1700 communities and outstations on the CHINS frame 331 remote community sets have been created representing two-thirds of the remote Indigenous population.

3.1.2 Remote community sampling

The sample design used for selections from the ICF is a stratified probability proportional to size-3-stage design. Community sets are stratified by state. The stages of selection are

- a random selection of community sets, where a set consists of a 'main' community and associated outstations, with probability proportional to the number of clusters (fixed cluster size) in the main community of the set;
- a random selection of households within the main community using a systematic skip and a random selection of associated outstations with all households enumerated; and
- a random selection of persons within selected households.

Field operational considerations have played a significant role in the determination of the selection method and corresponding sample parameters.

The mechanism of selecting outstations was a major source of development effort in the early stages of ICF and remote community sampling development. The method effectively stratifies community sets into the 'main' community and 'outstations' with the main community being selected with certainty. The reasons for the approach included

- to ensure that necessary contacts and facilitators required to gain permission and visit smaller outstations, could be obtained on the one trip as far as possible;
- minimise the number of outstations selected due to high costs and other difficulties of enumeration; and
- maintain equal dwelling level probabilities of selection within the PSU as far as possible.

The rationale for using a fixed cluster size is to dampen the wide variation in first-stage selection probabilities caused by a highly skewed distribution of numbers of dwellings by community and to maintain a reasonable workload size for individual community visits. For example, in the Northern Territory 48% of the population are resident in 16 large communities out of 644 communities and outstations on the frame.

The choice of cluster size and other parameters such as within household person selection arrangements are also largely determined by operational considerations. Attempts at formal cost variance modelling to optimise these parameters have been made but are limited by input data.

Cost modelling is difficult because of the sheer number of factors that affect travel costs, including availability of interviewers in remote centres such as Alice Springs, cost of car and aircraft hire which can vary depending on which communities are selected, and enumeration timing.

Variance modelling is limited by lack of reliable unit record data from the Census for remote communities and to a lesser extent the quality of dwelling and population counts at community and outstation level. There is also a wide range of levels of intraclass correlation for different items. For example 'employment' has low within and high between community variation, driven no doubt by the presence or absence of Community Development Employment Projects (CDEP) programs within the community, whereas health characteristics are far more similar across communities.

Pragmatic decisions on persons per household and number of dwellings per community have been made. For NATSISS '08 a one adult and one child per household selection method was used, these being deemed the maximum number of respondents that could be interviewed for the given content before the household tires and remaining selected persons 'go missing', noting that 'call back' options are very limited.

The 'cluster size' choice is determined from a mix of

- past surveys' variance outcomes;
- what an interviewer team could reasonably be expected to complete within a one week visit, or the maximum time ABS is prepared to impose on a community, and
- variation in expected dwelling selection probabilities under different cluster size and maximum outstations per community set choices

For NATSISS '08 a target cluster size of 30 dwellings has been used. In many cases this will equate to all community dwellings being selected.

Broad state-level per-community set field costs and design effects are used in final sample allocation in conjunction with non-community design inputs. Community set unit costs and design effects are estimated from previous surveys e.g. for NATSISS '08 in Northern Territory remote community strata a per-community set field cost of approximately $15,200 was assumed with an expected 28 fully responding persons per community set selected and a design effect for person level estimates of 3.0.

3.2 Non-community component

Despite complexities of remote community sampling and challenges in determining design parameters, the non-community component of the sample is where most of the sample design effort of has been spent in recent designs. This is a result of the magnitude of the population in non-community areas, keen user interest in boosting sample in non-remote states such as Victoria, and an overarching aversion to the method of sample selection: screening.

3.2.1 Non-community frame

Census collection districts not covered by the ICF form the frame from which the non-remote community component of the sample is drawn. This includes a reasonable portion of remote areas not covered by the ICF. Key data used at CD level includes the numbers of private dwellings, private dwellings containing at least one Indigenous usual resident as at census night, and numbers of Indigenous persons. CDs containing no Indigenous usual residents (as reported by Census), are excluded from coverage. The estimated level of undercoverage from Indigenous people moving into such areas by the time the survey is conducted is discussed in Section 6. For the 2008 NATSISS the newly available 'meshblock' was a key addition to the non-community frame, also discussed at length in the remainder of the paper.

3.2.2 Non-community sampling

Sample is selected using an area-based stratified probability-proportional-to-size or simple random sampling without replacement (SRSWOR) three-stage design with a screening component. Prior to NATSISS '08 the CD was the primary sampling unit. CDs on the non-community frame are stratified by state, remoteness classification and 'size'. Size refers to the estimated number of Indigenous households in the CD based on the most recent census counts. The stages of sample selection are

1. a random selection of CDs with probability either proportional to the number of 'clusters' (of Indigenous households) in the CD, or fixed at the stratum level;
2. a systematic selection of households within selected CDs, with selected households screened to establish the presence of Indigenous residents; and
3. a random selection of persons within selected households identified to contain Indigenous usual residents.

Details of how sample design parameters for this selection method are determined are given below.

3.2.3 Alternatives to screening

The magnitude of screening required for an Indigenous survey is a real concern due to the operational concerns, not least of which is interviewer morale. Investigations into alternate mechanisms have been made over the years but none of these have been pursued. Kalton and Anderson (1986) outline a range of sampling options in their paper "Sampling Rare Populations". Table 3.2 outlines these and other options investigated.

Several other techniques (e.g. Mitofsky–Waksberg, adaptive sampling) have been developed for sampling rare populations where these populations are clustered at the PSU level. These techniques work on the principle that when an interviewer finds one member of the target population, others are likely to be nearby. However, as discussed in Section 2, urban Indigenous households do not show a great deal of clustering at the CD level, making such techniques inappropriate here.

3.2 Indigenous survey selection mechanisms

Method	*Description*
"Skip then screen"	As described above, the current preferred selection method. It entails a systematic selection of households using a skip through selected PSUs. Selected households are screened to assess presence of Indigenous usual residents via Interviewer asking at door step "Are there any usual residents of this dwelling of Aboriginal and/or Torres Strait Islander origin?". Identified Indigenous households are selected in survey.
"Screen then skip"	All households in selected PSUs are screened. A sampling skip is then run through identified Indigenous households. This approach was adopted in the 1994 NATSIS. The major problem with the method was the loss of Indigenous households between time of identification at screening stage and returning to conduct survey (15% drop out rate).
Quota sampling	A means of reducing the level of screening under either of the above methods, whereby screening ceases after quota of Indigenous households are identified. The method was used in the Indigenous sample supplement to the 1999 Australian Housing Survey. Selected CDs were blocked and randomly selected blocks screened in their entirety before deciding whether to proceed to the next block. Problems included high levels of travel due to the need to progressively complete call-backs and difficulty in estimating initial selection probabilities.
Multiplicity sampling	Also known as network sampling. Where Indigenous households are identified they are asked about all household members and other Indigenous persons with 'linkages' to them, e.g. extended family. The method is primarily used where there is little frame information. Linkages must be clearly specified so selection probabilities can be determined. The method was not seriously considered due to ethical and privacy concerns appropriate for an official government agency.
Administrative lists	The use of existing administrative lists such as Indigenous health clinics, Indigenous public housing lists and Medicare data have been considered to have good potential for use either in a multiple frame sampling context or as a means of search facilitation. The methods have not been totally ruled out but gaining necessary access to address-level data has proved problematic. Lists that have been obtained have had numerous quality issues such as undercoverage, duplication and out of date contact information to the extent that there is no real benefit to reduction in screening.

4. CD-BASED SAMPLE DESIGN

The following outlines the method of determining key design parameters that make up the screened component of the design, using CDs as the primary sampling unit. Subsequent sections discuss improvements on the base method using newly-available meshblock data.

The key design parameters for the screened component of the sampling design include

- stratification
- screening skips
- persons per household selection scheme, and
- sample allocation.

The determination of these parameters is typically an iterative process. The first step is to establish size stratification, screening skips, and the within-household selection scheme. Once these parameters are locked in, stratum level variance contributions for a 'typical' variable of interest are set, unit costs and overheads finalised, remote community strata incorporated, and then allocation scenarios are explored to assess respective tradeoffs between expected accuracy levels for a range of domains of interest within a fixed field enumeration budget.

Previous Indigenous surveys were designed to produce estimates with target levels of accuracy at the state level, with certain state/remoteness level estimates also being of interest. State level accuracy targets have represented a compromise between equal state and territory level accuracy and optimal national level estimates, with the 2008 NATSISS involving a significant user funded boost to the Victorian sample. Given the markedly different make-up of the target population distribution for each state and territory, core sample design parameters have been optimised on a state by state basis.

4.1 Stratification

The determination of size boundaries is the main focus of stratification work. A reasonably fine size stratification is preferred to the broad probability proportional to size approach used in NATSIS '94. Finer stratification provides greater control over expected sample takes, the level of screening, and coverage levels. However, over-stratification is avoided. Strata are formed such that

- there are sufficient CDs on the frame to support an approximate indicative sample allocation; and
- the indicative sample allocation is the equivalent of at least one full cluster expected to be obtained from a CD in the stratum.

Variations in state level population distributions drive corresponding variation in size boundaries. For example, size boundaries adopted for NATSISS '08 in Darwin are <11, 11–15, 16–20, 21–25, 26–30, 31–50 and 51-plus whereas in Melbourne they are 1, 2, 3, ..., 9–10 and 11-plus Indigenous households per CD.

4.2 Screening skips and persons per household

In previous designs, the starting point for determining screening skips and persons-per-household selection arrangements was to treat the sampling scheme as a multi-stage cluster design in a manner similar to the MPS sample design, where the 'cluster' size here represents the number of Indigenous households selected per CD in a given stratum. A standard cost/variance optimisation process is conducted to determine 'optimal' cluster sizes and in turn establish stratum level screening skips.

For each state the objective is to determine optimal cluster sizes and persons per household that minimise cost for notional target variances of person-level estimates of prevalence of a characteristic e.g. minimum prevalence level for which 25% RSE can be achieved.

Assuming use of a Horvitz–Thompson estimator, the resulting variance function is:

$$Var\left(\hat{Y}_s\right) \cong \sum_{h \in s} a_h + b_h \frac{1}{l_h} + c_h \frac{1}{m_h}$$

where s is the state/territory of interest, l_h the number of CDs selected in stratum h, m_h the number of Indigenous households selected in stratum h, and parameters a_h–c_h reflect variance components for each stage of selection under a given within-household sampling scheme. The cluster size q_h is equal to m_h/l_h, i.e. average Indigenous households selected per CD sampled. Details are included in Appendix A, Sections A.1–A.3. Key variance components for each stage of selection were estimated from unit record census data for a range of items such as employment, education status, and income.

Table 4.1 presents variance statistics for Victorian NATSISS strata: approximate Indigenous population, and design effects ('deffs') that indicate the effect of design clustering on variance. These design effects are presented for illustrative purposes only; they are calculated from the hybrid design that will be discussed in Section 8, and as such are slightly different to the deffs that would be achieved with a 'pure CD' design; furthermore, large sampling fractions in some strata mean that deffs will vary with sample size. The first two digits of stratum number indicate state (1–8) and remoteness (0–4); the last two digits indicate number of Indigenous households per CD in the stratum. Appendix E contains data for other states.

4.1 Victorian design effects, by stratum

Victorian stratum	*Persons in stratum*	*Deff*	*Victorian stratum*	*Persons in stratum*	*Deff*
2001	3,345	0.72	2108	451	1.18
2002	3,579	1.79	2109	447	1.11
2003	2,826	1.51	2110	210	1.22
2004	1,956	1.36	2111	784	2.06
2005	1,213	1.38	2116	728	1.21
2006	764	1.37	2121	727	1.50
2007	443	2.11	2201	272	0.55
2008	123	1.02	2202	335	0.98
2009	363	0.88	2203	371	1.05
2011	152	2.54	2204	352	1.58
2101	1,111	0.87	2205	254	1.45
2102	1,517	0.82	2206	224	1.75
2103	1,349	1.18	2207	217	1.21
2104	1,130	1.53	2208	204	2.33
2105	812	1.24	2209	528	1.84
2106	808	1.06	2211	1,116	1.75
2107	600	1.24	2221	999	1.80

A stratum level linear cost model is constructed in the form

$$C_s = \sum_{h \in s} e_h l_h + f_h m_h$$

where l_h is the number of CDs selected in stratum h, e_h represents the average base cost per CD sampled in stratum h (predominantly travel to and from the CD), andf_h represents the average screening and interviewing cost per Indigenous household in stratum h. This linear model is only intended to produce a cost-effective sample allocation, not to provide accurate estimates of true survey costs. Final ABS costings are produced by a separate process that incorporates many aspects of survey cost not included in the linear model (e.g. interviewer capacity and recruitment).

Travel, screening, and interviewing costs are estimated via linear regression analysis of cost, time, travel and screening records from previous survey data at interviewer workload level. Due to limitations in the data, parameters are fitted at broad area type level with the major variation across area types attributable to the travel cost parameter. The resulting stratum-level cost components roughly represent relativities between average per-CD travel costs, screening and interviewing.

Details of the cost model are given in Appendix B, and Appendix F contains State-by-Remoteness area estimates of fixed costs (equal to e_h), screening, and interview costs (which together determinef_h, along with characteristics such as Indigenous population distribution, density, and response rates).

Table 4.2 shows comparisons for Victoria RA 0, Victoria RA 1–2 and Northern Territory RA 2.

4.2 Selected cost coefficients

Region	*Fixed costs per CD or community set*	*Screening costs per household*	*Interview costs per adult*
Victoria RA 0	$661.36	$1.85	$301.35
Victoria RA 1–2	$1,186.96	$2.12	$301.35
Northern Territory RA 2	$212.18	$1.48	$196.45

Note: Cost figures are given for illustrative purposes only and do not represent official ABS costings.

These parameters are assumed to be constant within each state/RA, but their relative significance varies with stratum characteristics. In CDs with low Indigenous population, fixed and screening costs dominate. For instance, in size-1 CDs in Victoria RA 0, screening requirements are approximately 400 households (more than one CD) per fully-responding Indigenous household, and screening costs are more than double interview costs. The total survey cost breaks down to approximately 50% fixed-per-CD, 30% screening, and 20% interviewing.

In CDs with high Indigenous population, interview costs dominate. For size-51+ CDs in Northern Territory RA 2, screening requirements are only nine households per Indigenous household interviewed, and over 90% of costs are due to interviewing.

With cost and variance models formulated, optimisation for q_h is conducted separately under each of the various possible arrangements for selecting adults and children. This process has shown that the more persons interviewed per household the better; for NATSISS '08 up to two adults and two children per household are being selected, this representing the maximum acceptable contact time per household for what is generally a long questionnaire. With persons per household fixed it is shown (Appendix A.3) that optimal 'cluster size' is

$$q_h = \sqrt{\frac{c_h e_h}{b_h f_h}}$$

Outputs from the optimisation process are then used to determine reasonable screening skips, noting skip options are limited to either integer skips or 3/2 i.e. 'screen two miss one'. It has become clear that for the majority of strata a screening skip of 1 will apply and that only in large density CDs would skip be used to achieve an 'optimal' number of Indigenous households per CD. The maximum skip applied in any selected CD for NATSISS '08 is five.

4.3 Allocation

With core parameters set the sample allocation, incorporating remote community strata, is able to proceed. An iterative process is adopted where sample sizes for state and certain state by remoteness classifications are input and expected cost and accuracy levels tradeoffs assessed. Within a key geographic domain of interest such as a state by broad remoteness classification, optimal allocation is used to distribute the sample between strata.

Using a Lagrange-multiplier approach and an approximation given in Appendix A.4.1, it can be shown that when interview costs dominate (e.g. when Indigenous population densities are very high), so that the total sampling cost is approximately proportional to the number of households sampled, and if we ignore the complications of dwelling skip and variable response rate, then optimal allocation is equivalent to proportional allocation, in which every CD and every household in every stratum in the region has the same probability of selection.

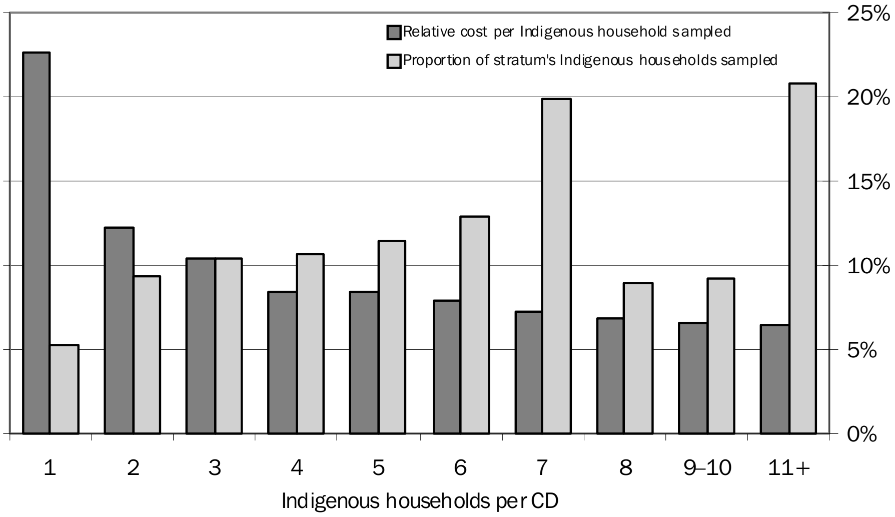
It can similarly be shown that when fixed-per-CD costs dominate, e.g. when densities are very low, optimal selection probabilities are approximately proportional to the square root of density.

In practice, the relative magnitude of fixed/screening and interview costs varies, with interview costs becoming more important in the denser strata. Variations in intra-strata correlations and response rate also affect these relationships. Figures 4.3 and 4.4 show examples of optimal allocations, given as the proportion of each stratum's Indigenous households that are sampled, for Victoria RA 0 and Northern Territory RA 2, along with total (fixed + screen + interview) costs per Indigenous household sampled in each stratum.

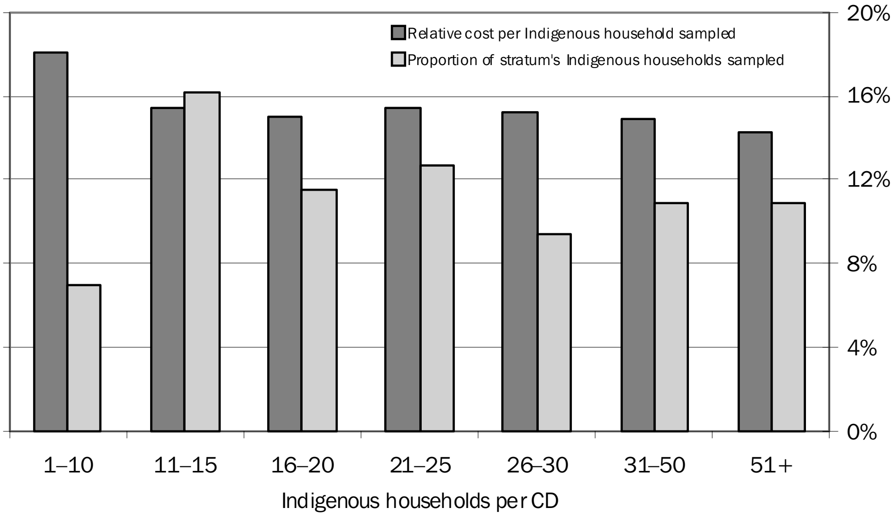
Note: Estimated design effects are relatively large for Victoria strata 7 and 11+, and Northern Territory stratum 11–15, and relatively small for Victoria strata 1 and 8–10, and Northern Territory stratum 1 and 26–30. The allocations for these strata are affected accordingly.

It can be seen that in Victoria RA 0, optimal selection probability increases slightly with density, but overall the differences are small compared to those created by variation in design effects; in Northern Territory RA 2, optimal allocation is close to proportional. Disproportionate screening concentrated on higher-density regions would certainly make it cheaper to achieve a given sample size, but would not be cost-effective in terms of accuracy.

4.3 Optimal sampling proportion and relative cost per Indigenous household sampled, Victoria RA 0, full-CD design



4.4 Optimal sampling proportion and relative cost per Indigenous household sampled, Northern Territory RA 2, full-CD design



5. MESHBLOCKS

Meshblocks were originally created as an output unit to meet the need for more agile geographic census output but have recently become of interest as a basis for sample design. The typical meshblock contains around 30 households (i.e. around 1/7th the size of a CD) and some relevant Census data is available at the meshblock level. This suggests that a meshblock-based Indigenous survey design could offer significantly less screening than a comparable 'pure CD' design – for instance, rather than screen all 200 dwellings to find a single Indigenous household within a size-1 CD, we can identify the meshblock where that household will be and reduce our screening to 30 dwellings.

Early investigations revealed three major problems with using meshblocks:

- Map information for remote meshblocks is not considered reliable enough to use. Therefore CD/community methods for these areas are retained with meshblock data used only for the non-remote portion of the sample.
- Meshblock boundaries do not match CD boundaries. 13% of meshblocks straddle more than one CD, and the worst straddled 22 CDs. This is a big problem for work that requires 'CD-compatibility' (e.g. controlling overlap with previous CD-based surveys) or when trying to estimate meshblock-level variances from CD-level unit record data.
- Meshblocks weren't originally intended as operational units; they are designed for consistency of geography, not workload. Thus, although the typical meshblock contains around 30 households (compared to 200 for a CD), the most populous contains over 1500 (compared to approximately 600 for the biggest CDs). Thus, while use of MBs would reduce expected screening overall, individual workloads would be more variable.

5.1 Split meshblocks

The solution to the boundary and size problems was to break up meshblocks along CD boundaries, creating a new geographical unit: split meshblocks ('SMBs'), with each SMB defined as the intersection of one parent CD with one parent MB. This ensured that SMB boundaries were both CD- and MB-compatible. Average SMB size would be similar to that for MBs (since 87% of MBs already lie entirely within a single CD), and the largest possible SMB would be no worse than its parent CD.

Household counts, both Indigenous and total, were available at the SMB level. This led to exploration of four successive design approaches:

i. Pure SMB-based

This design would use the same structure as the pure CD design, but using split meshblocks as PSUs in place of CDs throughout. This allows maximum use to be made of SMB-level Indigenous population data because selection probabilities can be set separately for each individual SMB. Avoiding size-0 SMBs would reduce screening requirements by 48% overall (62% in Victoria) for fixed sample size. Further, reduced clustering compared to the pure CD design would allow a slight reduction in sample size.

However, the number of PSUs to be sampled would approximately triple. Our linear cost model predicts that this would greatly increase overall survey costs thereby outweighing the savings from screening. It should be noted that this model is drawn from previous CD-based surveys. It is not clear whether it is appropriate to extrapolate this to a design with much smaller, more numerous PSUs. The reduced clustering also means that a larger number of CDs would be at least partially sampled, which could present difficulties for overlap control in any future CD-based surveys.

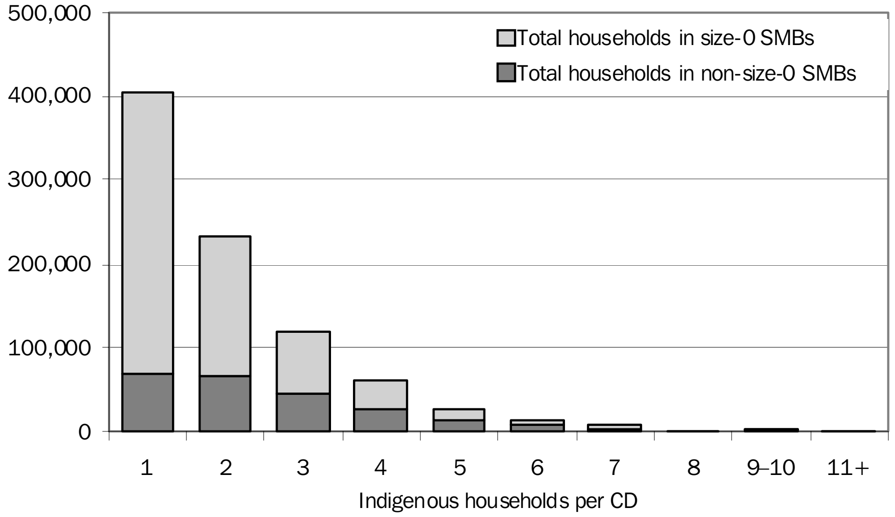
Furthermore, the available auxiliary data is in the form of unit records that only give CD, not MB/SMB, so this approach would require estimating variance characteristics etc. for individual SMBs from those for their parent CDs. This can be done but is likely to increase error in the variance model. For these reasons, the pure SMB-based design was abandoned; a better understanding of costs, combined with SMB-level auxiliary data, might make this approach more viable in the future.

ii. CD-based with SMB-facilitated screening

In this design CDs would be selected as under the pure CD design, but SMBs with no recorded Indigenous households ('size-0') would not be visited. Since most of the households screened in a pure CD design fall within size-0 SMBs, this could potentially provide a huge reduction in screening. For instance, in size-1 CDs in Victoria RA 0, more than 80% of all households (and hence, more than 80% of those screened) are in size-0 SMBs:

This approach represents a compromise between the pure-SMB and pure-CD designs. The elimination of size-0 SMBs produces similar screening savings to the pure-SMB design, while keeping the CD as primary sampling unit means we can use CD-based auxiliary data and paradata (e.g. response rates) with relatively little difficulty. However, post-Census migration means that some 'Census size-0' SMBs will in fact contain Indigenous people by the time NATSISS '08 is conducted, and these people would have no chance of selection. We estimated that the resulting undercoverage would be unacceptably high, up to 20% for urban Victoria (see Section 6). This design was therefore set aside.

5.1 Total households in size-0 and non-size-0 SMBs, by CD indigenous density, Victoria RA 0



iii. CD-based with 'size-0 skip'

In this design CDs would be selected as under the pure CD design. SMBs within selected CDs would then be classified as either 'size-0' or 'non-size-0' (as indicated by Census data). All of the Census non-size-0 SMBs within selected CDs would be selected for screening, and a skip through the Census size-0s in selected CDs would be used to select some of these for screening.

This design would achieve the same coverage as a pure CD design (since any size-0 within a selectable CD has a chance of selection) but would potentially reduce screening requirements (since only a fraction of the size-0 SMBs within selected CDs are chosen for screening). Its weakness is that size-0 SMBs have lower selection probabilities and hence higher weights, reducing design efficiency and requiring the sampling of more Indigenous households to meet accuracy requirements (see Section 7 and Appendix A.5–6 for further discussion). This places a limit on the cost and screening reductions that can be achieved. While this design was considered preferable to the pure CD design, further improvement was needed.

iv. CD-based with 'size-0 skip' and exclusions.

This design represents a compromise between designs (ii) and (iii) above. CDs are selected as under the pure CD design, all 'non-size-0' SMBs within selected CDs are selected, and then the size of the CD (Indigenous household count, as reported by Census) determines whether 'size-0' SMBs within these selected CDs are selected. In CDs with a high Indigenous household count, a skip is used to select some of these 'size-0's, as in design (iii) above. In CDs with a low Indigenous count, size-0 SMBs are excluded from selection.

This design leads to increased undercoverage compared to the pure CD design (because 'size-0's within small CDs have no chance of selection) but less so than design (ii) (because the 'size-0's most likely to have acquired new Indigenous occupants – those in larger CDs – still have a chance of selection). The criteria for size-0 exclusion can be set for each state and remoteness level in order to reduce undercoverage to acceptable levels; potential bias resulting from undercoverage is discussed in Section 8.

Efficiency is improved relative to design (iii) because selection probabilities and weights are less variable, allowing a slightly smaller sample at fixed accuracy targets. Screening requirements are less than for design (iii), due both to the reduction in sample and because *all* 'size-0' SMBs within small CDs have been removed from sample, whereas (iii) only removed some of them. This was chosen as the final NATSISS design.

6. MIGRATION AND MESHBLOCKS

Indigenous households have large rates of migration, meaning that Census-based estimates of CD/SMB size may no longer be current. This may lead to increased variance since selection probabilities are optimised for apparent rather than current PSU sizes. Where PSUs are excluded from selection on the basis of apparent size, i.e. 'size-0s', this can also lead to undercoverage and potential bias. As discussed in Section 5, the presence of Indigenous people in 'size-0' regions is important to choice of design and to design parameters, so we need to estimate the numbers involved.

Census '06 asked respondents whether they had been living at a different address one and five years ago. We classified Indigenous people as 'migrated (*n* years)' if they had been living at a different address *n* years ago. Since moving to an pre-existing Indigenous household doesn't cause undercoverage, we also defined household-level migration: an Indigenous household was defined as 'migrated (*n* years)' if every Indigenous person in the household was also migrated (*n* years). (Note that these migrations may be the result of Indigenous households merging or splitting, as well as simple one-to-one moves.) By interpolating we then estimated household migration rates for a two-year period (Appendix D). For non-remote areas, these are typically 20–30%.

Closer examination of Census data showed that the household migration rates for non-remote CDs of size-5 or smaller are generally similar to the overall figures for all Indigenous households in their state/RA category. For instance, in major urban Victoria 26.9% (interpolated) of Indigenous households had migrated in the two years before Census; within size-1 CDs in this region, the figure was 27.6%.

This indicates that around 27% of CDs that were size-1 as of Census 2006 were size-0 two years previously; by the same token, we might expect that 27% of CDs that are size-1 as of NATSISS 2008 would have been size-0 as of Census 2006 (and hence would have no chance of selection). Note that this relationship is not exact; if Indigenous people move around within a CD or replace others who've left, the CD size could be unchanged even though all its remaining Indigenous occupants are new arrivals.

From this data we can estimate how many Indigenous people might be in 'false size-0' CDs (i.e. those which have Indigenous occupants, but are not acknowledged as such in Census data). Unfortunately, the unit records containing migration data do not include meshblock identifiers. To predict figures for 'false size-0' SMBs, a less direct approach is required.

By treating size changes in SMBs as a Markov process, we will show that within a given region, the population of Indigenous persons in Census size-0 SMBs at the time of NATSISS '08 is approximately equal to the number of 'size-1' SMBs at the time of Census '06, multiplied by the two-year migration rate defined above:

Define $T(x)$ as a matrix of transition probabilities $t_{ij}(x)$, where $t_{ij}(x)$ is the probability that a SMB that contains i Indigenous households at time τ will contain j Indigenous households at time $\tau+x$. (For convenience we will begin indexing from 0, corresponding to size-0 SMBs.) We assume these probabilities are approximately independent of τ.

Note that the sum of transition probabilities from any given size must add to 1:

$$\sum_{j=0}^{\infty} t_{ij}(x) = 1 \quad \forall i, x$$

Define $\vec{d}(\tau) = (d_0(\tau), d_1(\tau), \ldots)'$ as a vector representing the distribution of SMB sizes at time τ, i.e. $d_i(\tau)$ gives the number of size-i SMBs at this time. $\vec{d}(\tau_{Census})$ is known. It can be seen that:

$$E\left(\vec{d}(\tau + x)\right) = T(x) \cdot \vec{d}(\tau)$$

Note also that the expected growth in a size-0 SMB over time x is equal to:

$$\sum_{j=0}^{\infty} j \cdot t_{0j}(x)$$

Given the high rate of churn (as demonstrated by migration rates above), we assume (1) that the distribution of SMB sizes is roughly in equilibrium, i.e.

$$\vec{d}\left(\tau_{Census}\right) \cong T(x) \cdot \vec{d}\left(\tau_{Census}\right)$$

In particular:

$$d_0\left(\tau_{Census}\right) \cong \sum_{i=0}^{\infty} t_{i0}(x)\, d_i\left(\tau_{Census}\right)$$

For $x = 2$ years, we assume (2) that transition probabilities between size-0 and size-2+ (in either direction) are negligible. It then follows that $T(x)$ is approximately equal to:

$$\begin{bmatrix} 1 - t_{01}(x) & t_{10}(x) & 0 & \cdots \\ t_{01}(x) & t_{11}(x) & t_{21}(x) & \cdots \\ 0 & t_{12}(x) & \ddots & \\ \vdots & \vdots & & \end{bmatrix}$$

Therefore:

$$\begin{bmatrix} 1-t_{01}(x) & t_{10}(x) & 0 & \cdots \\ t_{01}(x) & t_{11}(x) & t_{21}(x) & \cdots \\ 0 & t_{12}(x) & \ddots & \\ \vdots & \vdots & & \end{bmatrix} \cdot \vec{d}\left(\tau_{Census}\right) \cong \vec{d}\left(\tau_{Census}\right)$$

i.e.

$$\begin{aligned} d_0\left(\tau_{Census}\right) &\cong d_0\left(\tau_{Census}\right)\left(1-t_{01}\left(x\right)\right)+d_1\left(\tau_{Census}\right)t_{10}\left(x\right) \\ d_1\left(\tau_{Census}\right)t_{10}\left(x\right) &\cong d_0\left(\tau_{Census}\right)t_{01}\left(x\right) \end{aligned}$$

and

$$\begin{aligned} t_{01}\left(x\right) &\cong \frac{d_1\left(\tau_{Census}\right)t_{10}\left(x\right)}{d_0\left(\tau_{Census}\right)} \\ &\cong \frac{d_1\left(\tau_{Census}\right)}{d_0\left(\tau_{Census}\right)}t\left(x\right) \end{aligned}$$

where $t(x)$ is the household migration rate for time x, as defined above.

From this relationship and Census data, we can estimate the proportion of SMBs that will grow from size-0 to size-1 in the time between Census ’06 and NATSISS ’08, and hence estimate how sampling decisions for these SMBs affect variance/ undercoverage.

Using this methodology, we obtain estimates for the proportion of Indigenous households that will be in Census size-0 CDs and SMBs as of NATSISS.

Although these predictions are only approximate, it can clearly be seen that the proportion of Indigenous households that will be in Census ‘size-0’ SMBs is much higher than that for size-0 CDs, because the former population includes the latter. This means that the potential undercoverage associated with excluding Census size-0 SMBs either in a pure-SMB design or under SMB-facilitated screening is much higher than for CDs.

Victoria RA 0 represents a near-limiting case for sparseness: most of the Indigenous population (74% of Indigenous households) live in size-1 SMBs, indicating that when Indigenous households move they generally move to a previously-size-0 SMB. This means that we’d expect the proportion of households in Census size-0 SMBs to be close to the total migration rate. The simple approximation given above is consistent with this, giving an estimate of 20% of Indigenous households in ‘size-0’ SMBs.

6.1 Estimates of migration into size-0 CDs and SMBs

State and Remoteness area	*Indigenous households in size-0 SMBs*		*Indigenous households in size-0 CDs*	
	Number	*As a fraction*	*Number*	*As a fraction*
New South Wales				
0	2,849	10.4%	382	1.4%
1	1,345	7.1%	82	0.4%
2	424	4.2%	43	0.4%
Victoria				
0	1,510	19.8%	496	6.5%
1	718	14.3%	151	3.0%
2	175	8.2%	35	1.7%
Queensland				
0	2,232	14.6%	205	1.3%
1	1,100	10.3%	60	0.6%
2	523	3.7%	33	0.2%
South Australia				
0	732	13.0%	121	2.1%
1	132	12.1%	26	2.3%
2	139	6.2%	23	1.0%
Western Australia				
0	938	11.7%	147	1.8%
1	250	13.0%	27	1.4%
2	171	5.8%	25	0.9%
Tasmania				
1	304	7.1%	8	0.2%
2	172	4.9%	8	0.2%
Northern Territory				
2	55	1.4%	1	0.0%
Aust. Capital Territory				
0–1	250	13.2%	18	0.9%

7. SAMPLING SIZE-0

In previous Indigenous surveys ABS practice has been to exclude Census size-0 CDs from selection altogether. From the results above, it appears the resulting undercoverage is small; on the order of 2% or less for most areas, at worst approximately 6% for Victoria RA 0.

However, excluding all Census size-0 SMBs would result in much larger undercoverage – on the order of 10% for many areas, up to 20% for major urban Victoria. This risks significant levels of undercoverage bias and it would be difficult to give external users the desired level of confidence in resulting data.

For the substantial user funded sample increase in Victoria capacity limitations made it impractical to survey all Census size-0 SMBs in selected CDs. We therefore explored the possibility of selecting some of these SMBs (along with all non-Census-size-0 SMBs) for selected CDs. We chose to apply a 'size-0 skip' at the SMB level: select CDs, then use a skip of k to select $1/k$ of the Census size-0 SMBs within selected CDs. Interviewer procedures would be consistent for all selected SMBs within a single CD, although selection probabilities (and hence, weights) would differ between size-0s and non-size-0s according to k. The case where size-0 skip equals 1 is equivalent to a 'pure CD' design.

Estimating the effects of this approach on screening requirements was straightforward, since we had counts of Indigenous and total household numbers for each meshblock, and adjusted screening requirements could also be factored into the cost model. Estimating the effects on variance were rather more involved.

Treating cluster size as a variable to be optimised separately from number of clusters would have complicated the analysis further. In most CDs sampled, the expected number of responding Indigenous households is low enough that it is desirable to sample all of them, so a rigorous optimisation of cluster size would not be likely to provide large benefits. For this reason, rather than attempting such an optimisation, we arbitrarily chose a figure of 25 household interviews per CD as a design maximum. For large CDs, a dwelling skip was chosen (separately for each CD) to limit the expected sample to no more than 25 Indigenous households, after taking nonresponse into account. For most CDs, the dwelling skip was 1, meaning that we would screen every dwelling within selected SMBs.

7.1 Effects of size-0 skip on cost and variance

The use of size-0 skip means that Indigenous people within 'size-0' SMBs have lower selection probabilities but are weighted more heavily. This results in an increased variance compared to a pure CD design; an approximation for this variance increase is given in Appendix A.5–6. Size-0 skip also reduces the amount of screening required in each CD, and hence overall costs per CD. These effects are described in Appendix B.

7.2 Setting size-0 skip

In principle, it would be possible to optimise allocations and size-0 skips for each stratum in order to meet regional variance targets at minimal cost. In practice, the NATSISS design was large and complex with several overlapping accuracy requirements, and it was necessary to keep it in a form that could be adjusted without needing to run multivariable optimisation routines every time an aspect of the design needed adjusting.

Therefore, these cost and variance relationships were incorporated into a spreadsheet that allowed users to specify stratum-level size-0 skips and regional sample allocations (state/RA or state/broad RA).

Within each region, the specified sample was then allocated optimally between strata and accuracy data was output for each group of interest, along with cost/operational breakdowns etc.. Sample and skips were then manually adjusted to produce a satisfactory allocation overall.

7.1 Design / size-0 skip options, Victoria RA 0

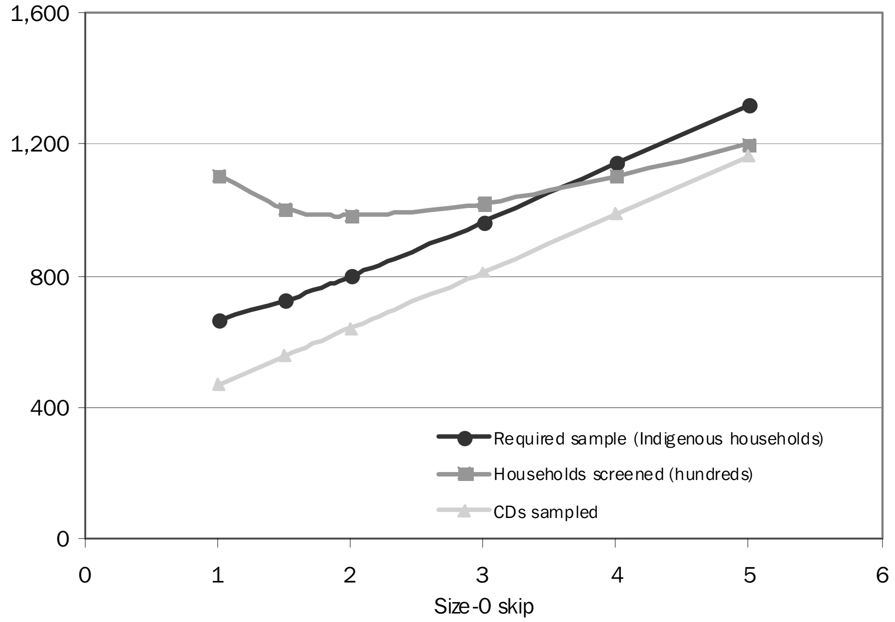


Figure 7.1 shows tradeoffs between skip, households interviewed, CDs sampled and screening for a Victoria RA 0 sample designed to achieve RSE 25% for items with prevalence 5% in children.

By reducing selection probabilities for Indigenous households in size-0 SMBs, we increase their weights, reducing design efficiency. The sample required to meet accuracy targets increases, and the number of CDs that must be sampled increases even more (proportionally) because size-0 skip reduces the number of people sampled per CD. Above skip 2, even the screening requirements begin to increase again due to the higher sample requirements.

In the Northern Territory, size-0 SMBs are relatively uncommon and size-0 skip had little effect on the design. Elsewhere, results were similar to those for urban Victoria: small size-0 skips (1.5 to 2) reduced screening requirements, but larger skips were counterproductive.

The large sample size, especially in Victoria, made it important to reduce screening requirements. The full-CD design would have required screening 111,589 households in Victoria RA 0 alone, in order to meet accuracy requirements. By contrast, estimated screening capacity for the whole of Victoria in 2008 was approximately 70,000 households, and even this figure would have required extensive recruitment and training of interviewers.

Setting size-0 skip at 2 produced a marked reduction in 2008 screening requirements. However, this was still not enough to bring screening in line with interviewer capacity, making further reductions necessary.

8. COMPROMISING: PARTIAL COVERAGE

Having determined that it was infeasible for NATSISS coverage to include all Census size-0 SMBs in non-size-0 CDs, and unacceptable to exclude all of them, we needed to compromise between these two scenarios by excluding *some* size-0 SMBs. We decided to do this by excluding those in each state/RA deemed least likely to contain Indigenous people (i.e. those within the smallest CDs).

Under this approach, total undercoverage is a combination of Indigenous households within Census size-0 CDs (this will be referred to as Component A), and households within non-covered Census size-0 SMBs within Census size-1+ CDs (Component B). We were chiefly interested in Component B, which represents additional undercoverage compared to previous Indigenous surveys.

We know that (size-0 SMBs) = (size-0 CDs) + (size-0 SMBs within non-size-0 CDs). In theory, we could use this relationship to estimate the population of Census size-0 SMBs within non-size-0 CDs, and hence estimate Component B undercoverage. However, both the size-0 SMB and size-0 CD estimates are expected to have large relative errors due to the assumptions made above, so their difference would be subject to even larger relative errors.

Instead, we chose a conservative estimate of Component B undercoverage. We already know that the number of households in Census size-0 SMBs is likely to be several times larger than that within Census size-0 CDs. We therefore ignored the latter, and assumed that the Census size-0 SMB households for each state were distributed entirely within non-size-0 CD strata proportional to strata sizes:

$$N_h^0 \;\cong\; tN_h$$

where N_h^0 is the number of households within size-0 SMBs in stratum h, N_h is the total population in stratum h, and t is the overall state/RA proportion of households within size-0 SMBs. This gives estimates of Indigenous households in Census size-0 SMBs for each individual stratum. (As noted before, this approximation is likely to break down for high-density CDs.)

We then used these estimates to determine how stratum-by-stratum treatment of size-0 SMBs affects total undercoverage. For instance, recall that in Victoria RA 0, 50% of Indigenous households are contained in CDs of size 1–2, and t is approximately 20%. Hence we estimate that if we excluded all Census size-0 SMBs within size 1–2 CDs, Component B undercoverage for Victoria RA 0 would be approximately 50%×20% = 10%. (Recall that this is a conservative estimate; the unusually large Component A undercoverage in this region means that the Component B portion is likely to be somewhat less than this suggests.)

This was still considered too large, so the only remaining option for Victoria RA 0 was to exclude size-0 SMBs within size-1 CDs only, reducing Component B undercoverage to an estimated 4.7%. This might seem like a minimal saving, but in fact the benefits are still quite large because size-1 CDs represent such a large part of the Victorian sample. We compare Victoria RA 0 samples to achieve a fixed accuracy target (RSE 25% for items with prevalence 5% in children) under three different designs: full-CD sampling (i.e. size-0 skip = 1), split-meshblock-assisted sampling with size-0 skip = 2, and split-meshblock-assisted sampling with size-0 skip = 2 and excluding size-0 SMBs within size-1 CDs:

8.1 Sample vs screening for SMB sampling options, Victoria RA 0

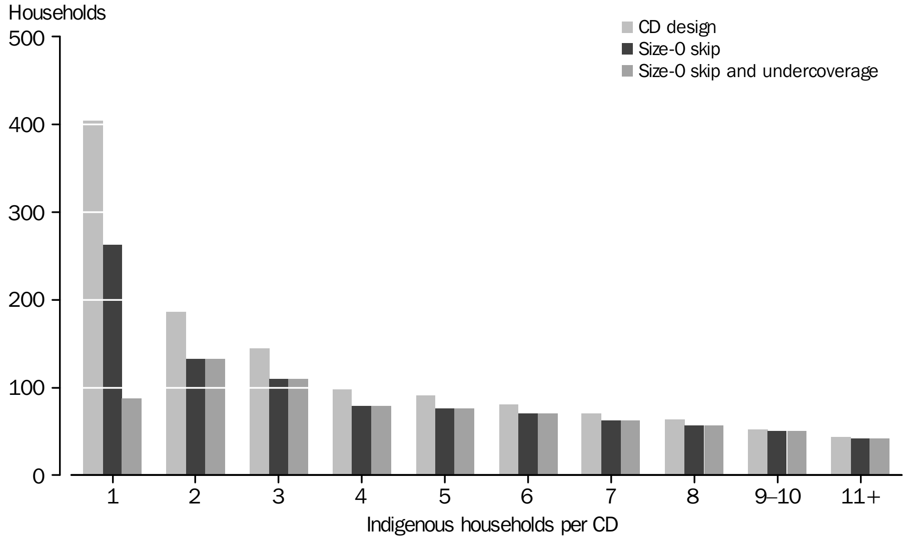
Design option	*Sample required (persons)*	*Screening requirements (households)*
CD design (size-0 skip=1)	1,136	111,589
Size-0 skip=2	1,367	99,468
Size-0 skip=2, and exclude size-0 SMBs in size-1 CDs	1,230	70,692

Observe that excluding these size-0 SMBs allows almost a 30% reduction in screening requirements, on top of that already achieved by introducing the size-0 skip. Part of this is due to a direct reduction in the screening required per household sampled in the size-1 stratum; the other part is because removing high-weight households in size-0 SMBs reduces the design effect in this stratum, allowing sample to be shifted to other strata with higher population densities. Combined with the two-part sample design, this produced a design considered to be operationally practical.

Figure 8.2 shows the effect on screening requirements per Indigenous household found for Victoria RA 0 (recall that here undercoverage applies only to size-1 CDs).

Although screening requirements were highest in Victoria, interviewer capacity had been reduced across Australia. We therefore adopted a similar approach nationwide, aiming for Component B undercoverage < 5% in each state/broad remoteness category and for major-urban and regional Victoria separately (per OFC requirements).

8.2 Households screened per Indigenous household sampled, under different NATSISS designs, Victoria RA 0



8.1 Potential bias due to undercoverage

Undercoverage is likely to create some bias (varying by data item) and the magnitude of this bias needed scrutiny. In general, undercoverage will occur due to people moving into Census size-0 regions (although it can also occur due to changes in self-identification). We therefore assumed that 'undercovered' people have similar characteristics to the migrated (1 year) persons identified in Census unit data.

The true population mean of a characteristic of interest is the weighted average of means for the recently-migrated and non-recently-migrated subpopulations:

$$y_{\text{mean\_true}} = \frac{N_{\text{non-moved}} \times y_{\text{non-moved}} + N_{\text{moved}} \times y_{\text{moved}}}{N_{\text{non-moved}} + N_{\text{moved}}}$$

Under the above assumption, we can simulate undercoverage bias by 'underweighting' the recently-migrated subpopulation, reducing its effective size by 5% of the total population:

$$y_{\text{mean\_under}} = \frac{\begin{array}{l} N_{\text{non-moved}} \times y_{\text{mean\_non-moved}} + \\ \left[N_{\text{moved}} - \left(N_{\text{non-moved}} + N_{\text{moved}}\right) \times 0.05\right] \times y_{\text{mean\_moved}} \end{array}}{\left(N_{\text{non-moved}} + N_{\text{moved}}\right) \times 0.95}$$

Bias and relative bias can be calculated as $(y_{\text{mean\_under}} - y_{\text{mean\_true}})$ and $(y_{\text{mean\_under}} - y_{\text{mean\_true}})/y_{\text{mean\_true}}$ respectively.

Note: Size-0 population estimates and undercoverage were calculated based on two-year migration, but due to time constraints the bias estimates were based on one-year migration. Since the difference between migrants and non-migrants is likely to be greatest for recent migrants, this may lead to some overestimation of bias.

Note also: These estimates were calculated assuming 5% undercoverage of ***persons****, while the undercoverage was actually set as undercoverage of* ***households****. We expect that undercoverage of persons would actually be somewhat less, since recently-migrated households are likely to have less persons, thus this method may somewhat overestimate bias.*

We used this method to calculate absolute and relative bias in prevalence estimates for items of interest including language, housing, and employment status. As might be expected, undercoverage of recently-migrated people skews the sample towards people who own their home (relative bias +2.7% at a national level) or have a mortgage (+1.6%), and away from people who rent (–1.1%). At the national level, the worst biases observed for publication items were of absolute magnitude 0.7% and relative magnitude 2.7%; worst state/RA-level biases were 1.4% and 4.1%.

A few data items may be more closely correlated to migration (in particular, questions about recent migration!) and these would have larger biases than indicated here; the maximum absolute bias in prevalence estimates for such items would be ±5%.

Overall, this was considered an acceptable risk. We expect that the benchmarking process applied to raw data would slightly reduce these biases, since mobility (and hence risk of being 'undercovered') is likely to have some correlation with benchmarking categories.

9. CONCLUSIONS

A meshblock-level approach offers greater geographical precision than a CD-level approach in both frame and selection, but the benefits of this precision are greatly limited by the timeliness of the frame data. The sparse distribution and high migration rate of non-remote Indigenous people make it impossible to achieve accurate results without a large amount of screening, even in areas not believed to have Indigenous occupants. Pragmatic design requires tradeoffs between cost/screening, variance, and undercoverage (bias).

Migration means that a substantial part of the Indigenous population are likely to be living in 'Census size-0' SMBs by the time the survey is run. Excluding these blocks from NATSISS altogether would cause unacceptable levels of undercoverage, creating potential bias. However, ignoring SMB-level data altogether would lead to excessive screening requirements. It is therefore necessary to compromise between the two, reducing the sampling of 'size-0s' without completely eliminating them. There are two ways this may be done:

We may lower the selection probability of 'Census size-0' SMBs without removing them from coverage altogether ('size-0 skip'). This can reduce screening requirements without introducing undercoverage, but the reduction is limited by the effects on accuracy. Large size-0 skips become counterproductive, because higher weights reduce design efficiency, requiring a larger overall sample to maintain accuracy; taken too far, this can actually increase overall screening requirements.

Alternately, we may keep some 'size-0s' in coverage while excluding others altogether. This approach reduces screening without increasing variance, but it causes undercoverage and hence potential bias. Selectively excluding only those SMBs considered least likely to acquire Indigenous migrants improves the tradeoff between these concerns, maximising the screening reductions achieved under a given level of undercoverage.

The final NATSISS design combines these two approaches, first using size-0 skip to reduce screening as far as possible, and then accepting a small degree of undercoverage in exchange for a further large saving in screening.

Given the importance of migration and timeliness to meshblock-based sample design, more detailed information about Indigenous migration patterns would be highly desirable in planning future Indigenous surveys. In particular, since the Census cycle (five years) and the Indigenous survey cycle (six years) are not synchronised, the timeliness of frame data will vary from one survey to the next.

9.1 MAC discussion

Discussion at the Methodology Advisory Committee meeting identified several possible avenues for future work:

1. Create a longitudinal panel of Indigenous Australians, to be tracked between surveys. Once the panel is established, future surveys could use this panel to find interviewees rather than relying on screening. This would also allow longitudinal analysis of results. Drawbacks of such an approach include privacy considerations and the expense/difficulty of tracking individuals.
2. Create a longitudinal panel of dwellings, with a top-up sample gained through screening. While the existing Indigenous occupants of a household may have moved away between surveys, there's a relatively high probability that their successors will also be Indigenous. This approach reduces the need for tracking compared to a panel of individuals, but is more likely to lose people, increasing the need for a top-up sample (presumably acquired via screening) to replenish the panel.
3. Consider basing survey scope and coverage decisions on social conditions, not only Indigenous identification. Often the end users of Indigenous survey data are especially interested in Indigenous households with specific characteristics; these people may have different migration characteristics from the overall Indigenous population, which may make migration-based undercoverage more or less important accordingly.
4. Social interpretation of Indigenous migration. Migration may be between enclaves or large Indigenous households, from an enclave to an 'empty' CD, independent of enclaves, etc.. Differences between these patterns of migration are important when evaluating undercoverage, and a social perspective on migration may help here. How does past/present location influence migration patterns?
5. Consider combining Indigenous surveys with other surveys that can make some use of screened non-Indigenous households.

REFERENCES

Kalton, G. and Anderson, D. (1986) "Sampling Rare Populations",*Journal of Official Statistics*, 149(1), pp. 65–82.

APPENDIXES

A. VARIANCE MODELS FOR THE NON-COMMUNITY SAMPLE

A.1 Variance modelling for basic CD design

Variance modelling is based on Census unit record data, which contains responses for various items of interest (e.g. age, employment status) for Indigenous persons in private dwellings.

For variance-modelling purposes, the basic CD design described in Section 4 is treated as a stratified three-stage design, with fixed-size SRSWOR (Simple Random Sampling Without Replacement) applied at the first and second stages (selection of CDs within stratum, and households within selected CDs). For the purposes of variance modelling, we assume that the Census unit records accurately represent the population of the CD at the time of the survey. (This assumption is likely to be quite inaccurate at the level of an individual CD, but we assume that these effects largely cancel out at the stratum level – the overall population characteristics, distribution etc. will be similar between Census and survey.) We also treat the selection of dwellings as fixed-size SRSWOR, with achieved sample size equal to the expected sample size – i.e. total Indigenous private dwellings present in the CD, multiplied by a 'hit rate' and divided by dwelling skip. (*Rationale*: This simplifies variance calculations, and while it may be inaccurate for a single CD, we expect most of the effects of variable sample size to balance out at a stratum level, especially after benchmarking against demographic data.)

The hit rate is defined as the ratio of fully-responding Indigenous households to the number of Indigenous households nominally screened (according to frame data) within a given area. This hit rate reflects a combination of non-response, refusals, people who have moved away, etc..

From each selected dwelling, in-scope people are then selected for interviewing (usually either one or two persons, depending on the survey and region); this selection is SRSWOR.

The resulting variance for Horvitz–Thompson estimation of response variable totals is:

$$Var\left(\hat{Y}\right) = \sum_{h=1}^{H} Var\left(\hat{Y}_h\right)$$

$$Var\left(\hat{Y}_h\right) \cong \frac{L_h^2}{l_h}\left(1 - \frac{l_h}{L_h}\right)S_h^2 + \frac{L_h}{l_h}\sum_{i=1}^{L_h}\frac{M_{h,i}^2}{m_{h,i}}\left(1 - \frac{m_{h,i}}{M_{h,i}}\right)S_{h,i}^2$$

$$+ \frac{L_h}{l_h}\sum_{i=1}^{L_h}\frac{M_{h,i}}{m_{h,i}}\sum_{j=1}^{M_{h,i}}\frac{N_{h,i,j}^2}{n_{h,i,j}}\left(1 - \frac{n_{h,i,j}}{N_{h,i,j}}\right)S_{h,i,j}^2$$

This represents the sum of variance from the different stages of selection: variance associated with selection of CDs within strata, dwellings within CDs, and persons within dwellings.

Abbreviations:

H = number of strata in region of interest

L_h = number of CDs in stratum *h*

l_h = number of CDs sampled in stratum *h*

M_h = number of in-scope Indigenous dwellings in stratum *h*

m_h = number of Indigenous dwellings sampled in stratum *h*

$M_{h,i}$ = number of dwellings containing at least one Indigenous person in CD *i*, stratum *h*

$m_{h,i}$ = number of Indigenous dwellings that would be interviewed in CD *i*, stratum *h*, if that CD is selected

$N_{h,i,j}$ = number of Indigenous persons in dwelling *j*, CD *i*, stratum *h*.

$n_{h,i,j}$ = selectable number of Indigenous persons in dwelling *j*, CD *i*, stratum *h*; this depends on $N_{h,i,j}$ and the maximum number of persons to be interviewed per household.

Y_h = total of response variable in stratum *h*

$Y_{h,i}$ = total of response variable in CD *i*, stratum *h*

$Y_{h,i,j}$ = total of response variable in dwelling *j*, CD *i*, stratum *h*

S_h^2 = population variance for CD totals within stratum

$$S_h^2 = \frac{1}{\max(1, L_h - 1)}\sum_{i=1}^{L_h}\left(Y_{h,i} - \bar{Y}_h\right)^2 = \frac{1}{\max(1, L_h - 1)}\sum_{i=1}^{L_h}\left(Y_{h,i} - \frac{Y_h}{L_h}\right)^2$$

$S^2_{h,i}$ = population variance for dwelling totals within CD

$$S^2_{h,i} = \frac{1}{\max(1, M_{h,i} - 1)} \sum_{j=1}^{M_{h,i}} \left(Y_{h,i,j} - \bar{Y}_{h,i}\right)^2$$

$$= \frac{1}{\max(1, M_{h,i} - 1)} \sum_{j=1}^{M_{h,i}} \left(Y_{h,i,j} - \frac{Y_{h,i}}{M_{h,i}}\right)^2$$

$S^2_{h,i,j}$ = population variance for person response variables within dwellings

$$S^2_{h,i,j} = \frac{1}{\max(1, N_{h,i,j} - 1)} \sum_{k=1}^{N_{h,i,j}} \left(Y_{h,i,j,k} - \bar{Y}_{h,i,j}\right)^2$$

$$= \frac{1}{\max(1, N_{h,i,j} - 1)} \sum_{k=1}^{N_{h,i,j}} \left(Y_{h,i,j,k} - \frac{Y_{h,i,j}}{N_{h,i,j}}\right)^2$$

Note: In many cases there will only be a single in-scope household within a CD, or a single in-scope person within a dwelling. In such situations, dividing by (population size minus 1) would lead to a zero-divided-by-zero error; the 'max(1,...)' operations above are used to force such expressions to evaluate as zero. (In practice, there will be some variance associated with selecting from a size-1 subpopulation when hit rate means we may not end up selecting anybody in that subpopulation; this effect is ignored here, but may warrant further investigation.)

Note: Some Indigenous designs (e.g. NATSIS) have used probability-proportional-to-size selection of CDs within strata rather than SRSWOR. The exact variance for such designs will not be covered here; it follows a similar form, with contributions due to selection of CDs, dwellings, and persons.

A.2 Adults and children

In practice, surveys distinguish between adults (defined as age 15+) and children. This may be done either by excluding children from scope altogether (in which case, simply replace 'Indigenous person' with 'Indigenous adult' throughout the above) or by stratifying each selected dwelling into adults and children, and selecting separately from each group. NATSISS '08 uses the latter approach, selecting 1–2 Indigenous adults (depending on state and remoteness) and 1–2 Indigenous children per dwelling, allowing separate control for adult and child accuracy.

In this case, the variance expression is modified:

$$
\begin{aligned}
Var\left(\hat{Y}_h\right) \cong{} & \frac{L_h^2}{l_h}\left(1-\frac{l_h}{L_h}\right)S_h^2 + \frac{L_h}{l_h}\sum_{i=1}^{L_h}\frac{M_{h,i}^2}{m_{h,i}}\left(1-\frac{m_{h,i}}{M_{h,i}}\right)S_{h,i}^2 \\
& + \frac{L_h}{l_h}\sum_{i=1}^{L_h}\frac{M_{h,i}}{m_{h,i}}\sum_{j=1}^{M_{h,i}}\frac{N_{adult,h,i,j}^2}{\max\left(1,n_{adult,h,i,j}\right)}\left(1-\frac{n_{adult,h,i,j}}{\max\left(1,N_{adult,h,i,j}\right)}\right)S_{adult,h,i,j}^2 \\
& + \frac{L_h}{l_h}\sum_{i=1}^{L_h}\frac{M_{h,i}}{m_{h,i}}\sum_{j=1}^{M_{h,i}}\frac{N_{child,h,i,j}^2}{\max\left(1,n_{child,h,i,j}\right)}\left(1-\frac{n_{child,h,i,j}}{\max\left(1,N_{child,h,i,j}\right)}\right)S_{child,h,i,j}^2
\end{aligned}
$$

Here $M_{h,i}$ includes any private dwelling with Indigenous occupants.

Note: Many dwellings will have Indigenous adults but no Indigenous children, and some will have Indigenous children but no adults. The 'max(1, ...)' *operations are used to avoid dividing zero by zero in these cases; variances for empty populations (e.g. dwelling variance among adults in a family that has no adults) should be evaluated as 0.*

For an adult-only variable, this variance simplifies to:

$$
\begin{aligned}
Var\left(\hat{Y}_h\right) \cong{} & \frac{L_h^2}{l_h}\left(1-\frac{l_h}{L_h}\right)S_h^2 + \frac{L_h}{l_h}\sum_{i=1}^{L_h}\frac{M_{h,i}^2}{m_{h,i}}\left(1-\frac{m_{h,i}}{M_{h,i}}\right)S_{h,i}^2 \\
& + \frac{L_h}{l_h}\sum_{i=1}^{L_h}\frac{M_{h,i}}{m_{h,i}}\sum_{j=1}^{M_{h,i}}\frac{N_{adult,h,i,j}^2}{\max\left(1,n_{adult,h,i,j}\right)}\left(1-\frac{n_{adult,h,i,j}}{\max\left(1,N_{adult,h,i,j}\right)}\right)S_{adult,h,i.j}^2
\end{aligned}
$$

This is very similar to the expression given in Section A.1 above, but allows for the selection of Indigenous dwellings with no Indigenous adults in them.

In the cases that follow, we will only present variance estimates for adult-only variables (while allowing for no-adult dwellings); N, n, S and Y variables should be assumed to refer to adult populations/totals/etc.. The extension of these results to combined adults and children is straightforward but tedious, so will be omitted.

Note: In practice, variance modelling for child/combined data was hindered by a shortage of input data – while Census unit records contain several useful items for modelling variance in adult-only items (e.g. employment, income, etc.) this was not the case for children.

A.3 Variance modelling with stratum-level cluster size optimisation

Under a fixed-cluster-size approach, dwelling skips are set by CD in order to achieve roughly constant sample size (households interviewed) per CD, i.e.

$$m_{h,i} \cong q_h$$

It then follows that:

$$m_h \cong l_h \, q_h$$

This allows variance to be approximated as a function of $1/l_h$ and $1/m_h$:

$$\begin{aligned}
&Var\left(\hat{Y}_h\right) \\
&\cong \frac{L_h^2}{l_h}\left(1-\frac{l_h}{L_h}\right)S_h^2 + \frac{L_h}{l_h}\sum_{i=1}^{L_h}\frac{M_{h,i}^2}{m_{h,i}}\left(1-\frac{m_{h,i}}{M_{h,i}}\right)S_{h,i}^2 + \\
&\frac{L_h}{l_h}\sum_{i=1}^{L_h}\frac{M_{h,i}}{m_{h,i}}\sum_{j=1}^{M_{h,i}}\frac{N_{h,i,j}^2}{\max(1,n_{h,i,j})}\left(1-\frac{n_{h,i,j}}{\max(1,N_{h,i,j})}\right)S_{h,i,j}^2
\end{aligned}$$

$$\begin{aligned}
&\cong \frac{L_h^2}{l_h}\left(1-\frac{l_h}{L_h}\right)S_h^2 + \frac{L_h}{l_h}\sum_{i=1}^{L_h}\frac{M_{h,i}^2}{q_h}S_{h,i}^2 - \frac{L_h}{l_h}\sum_{i=1}^{L_h}M_{h,i}S_{h,i}^2 + \\
&\frac{L_h}{l_h}\sum_{i=1}^{L_h}\frac{M_{h,i}}{q_h}\sum_{j=1}^{M_{h,i}}\frac{N_{h,i,j}^2}{\max(1,n_{h,i,j})}\left(1-\frac{n_{h,i,j}}{\max(1,N_{h,i,j})}\right)S_{h,i,j}^2
\end{aligned}$$

$$\begin{aligned}
&=\left(\frac{1}{l_h}L_h^2 - L_h\right)S_h^2 + \frac{1}{l_hq_h}L_h\sum_{i=1}^{L_h}M_{h,i}^2S_{h,i}^2 - \frac{1}{l_h}L_h\sum_{i=1}^{L_h}M_{h,i}S_{h,i}^2 + \\
&\frac{1}{l_hq_h}L_h\sum_{i=1}^{L_h}M_{h,i}\sum_{j=1}^{M_{h,i}}\frac{N_{h,i,j}^2}{\max(1,n_{h,i,j})}\left(1-\frac{n_{h,i,j}}{\max(1,N_{h,i,j})}\right)S_{h,i,j}^2
\end{aligned}$$

$$\begin{aligned}
&= -L_hS_h^2 + \frac{1}{l_h}\left(L_h^2S_h^2 - L_h\sum_{i=1}^{L_h}M_{h,i}S_{h,i}^2\right) + \frac{1}{m_h}\left(L_h\sum_{i=1}^{L_h}M_{h,i}^2S_{h,i}^2\right) + \\
&\frac{1}{m_h}\left(L_h\sum_{i=1}^{L_h}M_{h,i}\sum_{j=1}^{M_{h,i}}\frac{N_{h,i,j}^2}{\max(1,n_{h,i,j})}\left(1-\frac{n_{h,i,j}}{\max(1,N_{h,i,j})}\right)S_{h,i,j}^2\right)
\end{aligned}$$

That is,

$$Var\left(\hat{Y}_h\right) \cong a_h + b_h \frac{1}{l_h} + c_h \frac{1}{m_h} \quad \text{or equivalently,} \quad a_h + b_h \frac{q_h}{m_h} + c_h \frac{1}{m_h}$$

where:

$$a_h = -L_h S_h^2$$

$$b_h = L_h^2 S_h^2 - L_h \sum_{i=1}^{L_h} M_{h,i} S_{h,i}^2$$

$$c_h = L_h \sum_{i=1}^{L_h} M_{h,i}^2 S_{h,i}^2 + L_h \sum_{i=1}^{L_h} M_{h,i} \sum_{j=1}^{M_{h,i}} \frac{N_{h,i,j}^2}{\max(1, n_{h,i,j})} \left(1 - \frac{n_{h,i,j}}{\max(1, N_{h,i,j})}\right) S_{h,i,j}^2$$

Under the assumptions above, these three coefficients can all be calculated from the frame. Combined with a linear cost model of the form $\text{cost}_h = e_h l_h + f_h m_h$, Lagrange-multiplier methods can then be used to calculate cost-optimal sample allocation (both number of CDs and number of households within each stratum) for a given variance, or variance-optimal allocation for a given cost. It can easily be shown that these optimal allocations are of the form:

$$l_h = \alpha \sqrt{\frac{b_h}{e_h}} \quad , \quad m_h = \alpha \sqrt{\frac{c_h}{f_h}}$$

for some constant alpha, which can be calculated from terms $a_h - f_h$ and either accuracy or cost requirements. This yields the optimal cluster size:

$$q_h = \frac{m_h}{l_h} = \sqrt{\frac{c_h e_h}{b_h f_h}}$$

Note that this cluster size is *not* dependent on the specific cost/accuracy targets of the survey; altering these targets will alter the number of CDs and households sampled, but their ratio remains constant.

When overall allocations are quantified in terms of persons rather than households, the variance may also be expressed thus:

$$Var\left(\hat{Y}_h\right) \cong a_h + b_h \frac{1}{l_h} + c_h^* \frac{1}{m_h^*}$$

where $c_h^* = c_h \times i_h$ and m_h^* is the number of persons to be sampled in stratum h, and i_h is the expected number of persons sampled per household sampled (determined by household sizes and whether we interview one or two adults per household).

A.4 Variance modelling for CD design with CD-level cluster sizes

It may happen that instead of setting a constant cluster size for all CDs within a stratum, and trying to optimise this cluster size, we prefer to fix the cluster size for each CD. This might happen because cluster size optimisation produces an impractical/implausible result (the linear cost model may become inaccurate if extrapolated too far from our operational data points) or because design/operational considerations indicate fixed CD cluster sizes. (For instance, if SMB-level data on Indigenous households is treated as completely reliable, we might decide to fully enumerate all non-size-0 SMBs within selected CDs, so the CD cluster size is simply the total population of Indigenous households adjusted for expected hit rate.)

In this case, we can treat $\{m_{h,i}\}$ as known constants, and express variance as a linear function of $1/l_h$:

$$\begin{aligned}
&Var\left(\hat{Y}_h\right)\\
&\cong \frac{L_h^2}{l_h}\left(1-\frac{l_h}{L_h}\right)S_h^2 + \frac{L_h}{l_h}\sum_{i=1}^{L_h}\frac{M_{h,i}^2}{m_{h,i}}\left(1-\frac{m_{h,i}}{M_{h,i}}\right)S_{h,i}^2 + \\
&\frac{L_h}{l_h}\sum_{i=1}^{L_h}\frac{M_{h,i}}{m_{h,i}}\sum_{j=1}^{M_{h,i}}\frac{N_{h,i,j}^2}{\max\left(1,n_{h,i,j}\right)}\left(1-\frac{n_{h,i,j}}{\max\left(1,N_{h,i,j}\right)}\right)S_{h,i,j}^2
\end{aligned}$$

$$\begin{aligned}
&=\left(\frac{1}{l_h}L_h^2 - L_h\right)S_h^2 + \frac{1}{l_h}L_h\sum_{i=1}^{L_h}\frac{M_{h,i}^2}{m_{h,i}}\left(1-\frac{m_{h,i}}{M_{h,i}}\right)S_{h,i}^2 + \\
&\frac{1}{l_h}L_h\sum_{i=1}^{L_h}\frac{M_{h,i}}{m_{h,i}}\sum_{j=1}^{M_{h,i}}\frac{N_{h,i,j}^2}{\max\left(1,n_{h,i,j}\right)}\left(1-\frac{n_{h,i,j}}{\max\left(1,N_{h,i,j}\right)}\right)S_{h,i,j}^2
\end{aligned}$$

That is,

$$Var\left(\hat{Y}_h\right) \cong a_h + b_h\frac{1}{l_h}$$

where:

$$\begin{aligned}
a_h &= -L_h S_h^2\\
b_h &= L_h^2 S_h^2 + L_h\sum_{i=1}^{L_h}\frac{M_{h,i}^2}{m_{h,i}}\left(1-\frac{m_{h,i}}{M_{h,i}}\right)S_{h,i}^2 + \\
&\quad L_h\sum_{i=1}^{L_h}\frac{M_{h,i}}{m_{h,i}}\sum_{j=1}^{M_{h,i}}\frac{N_{h,i,j}^2}{\max\left(1,n_{h,i,j}\right)}\left(1-\frac{n_{h,i,j}}{\max\left(1,N_{h,i,j}\right)}\right)S_{h,i,j}^2
\end{aligned}$$

Here, because the cluster size has been fixed for each CD, the cost model is of the form: $Cost_h = d_h + e_h l_h$.

The optimal allocation is then of the form:

$$l_h = \alpha \sqrt{\frac{b_h}{e_h}}$$

A.4.1 Proportionality of b_h

It may be useful to understand how b_h varies with stratum characteristics (number of CDs and number of households). If household characteristics (means and variances) are assumed to be similar in all strata, and we ignore correlation between households in the same CD, it can easily be shown that:

$$S_h^2 \cong \alpha_1 \overline{M}_h$$

where $\overline{M}_h$ is the mean number of Indigenous households for CDs in stratum h. If we further assume that response rates are approximately constant and dwelling skips are equal to 1 everywhere, $M_{h,i}$ becomes simply proportional to $m_{h,i}$.

Under these assumptions, $S_{h,i}^2$ becomes approximately constant, and it can be seen that:

$$L_h \sum_{i=1}^{L_h} \frac{M_{h,i}^2}{m_{h,i}} \left(1 - \frac{m_{h,i}}{M_{h,i}} \right) S_{h,i}^2 \cong \alpha_2 L_h^2 \overline{M}_h$$

$$L_h \sum_{i=1}^{L_h} \frac{M_{h,i}}{m_{h,i}} \sum_{j=1}^{M_{h,i}} \frac{N_{h,i,j}^2}{\max\left(1, n_{h,i,j}\right)} \left(1 - \frac{n_{h,i,j}}{\max\left(1, N_{h,i,j}\right)} \right) S_{h,i,j}^2 \cong \alpha_3 L_h^2 \overline{M}_h$$

Combining these, we find that b_h is approximately proportional to $L_h^2 \overline{M}_h = L_h M_h$.

A.5 Variance modelling for Poisson CD design with size-0 skip

Directly evaluating the effects of a size-0 skip on the designs above, where selection of CDs and of dwellings within CDs is fixed-size SRSWOR, is difficult. We therefore begin by considering a simpler design, where both these stages work by Poisson selection (i.e. selection of units is independent) rather than fixed-sample SRSWOR, and variance associated with the third (within-dwelling) stage of selection is negligible (e.g. if the entire household is sampled.) However, we incorporate the effects of size-0 skip by reducing selection probability (and increasing weights) for households within size-0 SMBs.

We can represent this process thus:

Define the following variables:

k_h = size-0 skip for the stratum

$g_{h,i}$ = dwelling skip for each CD in the stratum

t_h = fraction of stratum IHHs within Census size-0 SMBs

r_h = hit rate (interviewed IHHs / screened IHHs)

$Y'_{h,i,j}$ = total response variable in j-th IHH within non-size-0 SMBs in CD i

$Y^*_{h,i,j}$ = total response variable in j-th IHH within size-0 SMBs in CD i

$Y_{h,i}$ = total response variable in CD i

$M'_{h,i}$ = total Indigenous households in non-size-0 SMBs in CD i

$M^*_{h,i}$ = total Indigenous households in size-0 SMBs in CD i

$M_{h,i}$ = total Indigenous households in CD i

For a given stratum h, define (mutually independent) indicator variables corresponding to CD selection:

$\{\delta_{h,i}\}, i \in (1, L_h)$:

$$\delta_{h,i} = \begin{cases} 1 & \text{with probability } \pi_{h,i} \\ 0 & \text{otherwise} \end{cases}$$

where $\pi_{h,i} = \Pr\{\text{CD } i \text{ selected}\}$ and CD i will be selected if and only if $\delta_{h,i} = 1$.

We will also define indicator variables (mutually independent of one another and of the above) corresponding to dwelling selection *conditional on CD selection* (separately for dwellings in size-0 SMBs and not in size-0 SMBs):

$\{\delta'_{h,i,j}\}, j \in (1, M'_{h,i})$

$$\delta'_{h,i,j} = \begin{cases} 1 & \text{with probability } \pi'_{h,i,j} \\ 0 & \text{otherwise} \end{cases}$$

where
$\pi'_{h,i,j} = \Pr\{j\text{-th IHH in non-size-0 SMBs within CD } i \text{ is selected} \mid \text{CD } i \text{ selected}\}$

and

$\{\delta^*_{h,i,j}\}, j \in (1, M^*_{h,i})$:

$$\delta^*_{h,i,j} = \begin{cases} 1 & \text{with probability } \pi^*_{h,i,j} \\ 0 & \text{otherwise} \end{cases}$$

where $\pi^*_{h,i,j} = \Pr\{j\text{-th IHH in size-0 SMBs within CD } i \text{ is selected} \mid \text{CD } i \text{ selected}\}$

A dwelling is selected only if $\delta_{h,i}\delta'_{h,i,j} = 1$ or $\delta_{h,i}\delta^*_{h,i,j} = 1$ (depending on whether the dwelling is in a size-0 SMB).

Note that as these are defined, it's possible that $\delta'_{h,i,j} = 1$ or $\delta^*_{h,i,j} = 1$ even if $\delta_{h,i} = 0$ – this can be interpreted as representing a case where CD i is not selected, but the j-th dwelling in the appropriate category *would have been selected* had that CD been selected.

Note: Here we are assuming that the true population of the CD matches the Census-identified population, so if we were to screen the entire CD, any disparity would be due to non-response (including non-identification). In practice, some of this disparity is due to emigration, but we'll model it as if those households were present and didn't respond.

In practice, t_h is calculated at a state/RA level (see Section 6 for method) and assumed to be constant across strata within that region. This implies that size-0 SMBs will be more likely to grow to size-1 when they are situated within a CD that already has a large Indigenous population. We consider this to be a realistic assumption, since the same factors that attracted existing Indigenous inhabitants may also attract new ones. The approximation is likely to break down for CDs with very high Indigenous population, where most new arrivals will enter a SMB that already has Indigenous occupants, but outside the Northern Territory such CDs are a relatively minor consideration. Feedback from previous surveys supports the belief that 'size-0' areas are more likely to acquire Indigenous occupants when nearby Indigenous populations are large.

From the skips and hit rate, we can calculate conditional household selection probabilities:

$$\pi'_{b,i,j} = \frac{r_b}{g_{b,i}}$$

$$\pi^*_{b,i,j} = \frac{r_b}{k_b \, g_{b,i}}$$

We will assume that within the stratum, $\left\{Y'_{b,i,j}\right\}$ and $\left\{Y^*_{b,i,j}\right\}$ have approximately equal distributions (i.e. means and variances). The accuracy of this assumption will vary depending on which of many variables of interest we're examining, but for the time being it's a necessary simplification. Following this assumption, define:

μ_b = stratum mean of household totals (i.e. $Y'_{b,i,j}$ and $Y^*_{b,i,j}$).

v_b = stratum population variance for household totals (i.e. $Y'_{b,i,j}$ and $Y^*_{b,i,j}$).

By definition, the Horvitz–Thompson estimator of stratum total under this design is:

$$\begin{aligned}\hat{Y}_b &= \sum_{i=1}^{L_b}\left(\sum_{j'=1}^{M'_{b,i}} \frac{\delta_{b,i}\delta'_{b,i,j}Y'_{b,i,j}}{\pi_{b,i}\,\pi'_{b,i,j}} + \sum_{j^*=1}^{M^*_{b,i}} \frac{\delta_{b,i}\delta^*_{b,i,j}Y^*_{b,i,j}}{\pi_{b,i}\,\pi^*_{b,i,j}}\right) \\ &= \sum_{i=1}^{L_b} \frac{\delta_{b,i}g_{b,i}}{\pi_{b,i}\,r_b}\left(\sum_{j'=1}^{M'_{b,i}} \delta'_{b,i,j}Y'_{b,i,j} + \sum_{j^*=1}^{M^*_{b,i}} k_b\delta^*_{b,i,j}Y^*_{b,i,j}\right)\end{aligned}$$

Therefore:

$$\begin{aligned}Var\left(\hat{Y}_b\right) &= Var\left\{\sum_{i=1}^{L_b} \frac{\delta_{b,i}g_{b,i}}{\pi_{b,i}\,r_b}\left(\sum_{j'=1}^{M'_{b,i}} \delta'_{b,i,j}Y'_{b,i,j} + \sum_{j^*=1}^{M^*_{b,i}} k_b\delta^*_{b,i,j}Y^*_{b,i,j}\right)\right\} \\ &= \sum_{i=1}^{L_b} Var\left\{\frac{\delta_{b,i}g_{b,i}}{\pi_{b,i}\,r_b}\left(\sum_{j'=1}^{M'_{b,i}} \delta'_{b,i,j}Y'_{b,i,j} + \sum_{j^*=1}^{M^*_{b,i}} k_b\delta^*_{b,i,j}Y^*_{b,i,j}\right)\right\}\end{aligned}$$

(by independence of indicator variables for different *i*-values)

$$= \sum_{i=1}^{L_b}\left(\frac{g_{b,i}}{r_b\pi_{b,i}}\right)^2 Var\left\{\delta_{b,i}\left(\sum_{j'=1}^{M'_{b,i}} \delta'_{b,i,j}Y'_{b,i,j} + \sum_{j^*=1}^{M^*_{b,i}} k_b\delta^*_{b,i,j}Y^*_{b,i,j}\right)\right\}$$

$$= \sum_{i=1}^{L_b}\left(\frac{g_{b,i}}{r_b\pi_{b,i}}\right)^2 E\left\{\left(\begin{array}{l}\delta_{b,i}\left(\sum_{j'=1}^{M'_{b,i}} \delta'_{b,i,j}Y'_{b,i,j} + \sum_{j^*=1}^{M^*_{b,i}} k_b\delta^*_{b,i,j}Y^*_{b,i,j}\right) \\ -E\left\{\delta_{b,i}\left(\sum_{j'=1}^{M'_{b,i}} \delta'_{b,i,j}Y'_{b,i,j} + \sum_{j^*=1}^{M^*_{b,i}} k_b\delta^*_{b,i,j}Y^*_{b,i,j}\right)\right\}\end{array}\right)^2\right\}$$

$$= \sum_{i=1}^{L_h}\left(\frac{g_{h,i}}{r_h \pi_{h,i}}\right)^2 E\left\{\left(\begin{array}{l} \delta_{h,i}\left(\displaystyle\sum_{j'=1}^{M'_{h,i}} \delta'_{h,i,j}Y'_{h,i,j} + \sum_{j^*=1}^{M^*_{h,i}} k_h \delta^*_{h,i,j}Y^*_{h,i,j}\right) \\ -E\left\{\delta_{h,i}\right\} E\left\{\displaystyle\sum_{j'=1}^{M'_{h,i}} \delta'_{h,i,j}Y'_{h,i,j} + \sum_{j^*=1}^{M^*_{h,i}} k_h \delta^*_{h,i,j}Y^*_{h,i,j}\right\} \end{array}\right)^2\right\}$$

(using independence again)

Note that:

$$\begin{aligned} & E\left\{\sum_{j'=1}^{M'_{h,i}} \delta'_{h,i,j}Y'_{h,i,j} + \sum_{j^*=1}^{M^*_{h,i}} k_h \delta^*_{h,i,j}Y^*_{h,i,j}\right\} \\ &= \sum_{j'=1}^{M'_{h,i}} E\left\{\delta'_{h,i,j}\right\}Y'_{h,i,j} + \sum_{j^*=1}^{M^*_{h,i}} k_h \, E\left\{\delta^*_{h,i,j}\right\}Y^*_{h,i,j} \\ &= \sum_{j'=1}^{M'_{h,i}} \frac{r_h}{g_{h,i}} Y'_{h,i,j} + \sum_{j^*=1}^{M^*_{h,i}} k_h \frac{r_h}{k_h g_{h,i}} Y^*_{h,i,j} \\ &= \frac{r_h}{g_{h,i}}\left(\sum_{j'=1}^{M'_{h,i}} Y'_{h,i,j} + \sum_{j^*=1}^{M^*_{h,i}} Y^*_{h,i,j}\right) \\ &= \frac{r_h}{g_{h,i}} Y_{h,i} \end{aligned}$$

Therefore:

$$Var\left(\hat{Y}_h\right) = \sum_{i=1}^{L_h}\left(\frac{g_{h,i}}{r_h \pi_{h,i}}\right)^2 E\left\{\left(\begin{array}{l} \delta_{h,i}\left(\displaystyle\sum_{j'=1}^{M'_{h,i}} \delta'_{h,i,j}Y'_{h,i,j} + \sum_{j^*=1}^{M^*_{h,i}} k_h \delta^*_{h,i,j}Y^*_{h,i,j}\right) \\ -\left(\dfrac{\pi_{h,i} r_h}{g_{h,i}} Y_{h,i}\right) \end{array}\right)^2\right\}$$

$$= \sum_{i=1}^{L_h}\left(\frac{g_{h,i}}{r_h \pi_{h,i}}\right)^2 E\left\{\begin{array}{l} {\delta^2}_{h,i}\left(\displaystyle\sum_{j'=1}^{M'_{h,i}} \delta'_{h,i,j}Y'_{h,i,j} + \sum_{j^*=1}^{M^*_{h,i}} k_h \delta^*_{h,i,j}Y^*_{h,i,j}\right)^2 \\ -2\delta_{h,i}\left(\displaystyle\sum_{j'=1}^{M'_{h,i}} \delta'_{h,i,j}Y'_{h,i,j} + \sum_{j^*=1}^{M^*_{h,i}} k_h \delta^*_{h,i,j}Y^*_{h,i,j}\right)\left(\dfrac{\pi_{h,i} r_h}{g_{h,i}} Y_{h,i}\right) \\ +\left(\dfrac{\pi_{h,i} r_h}{g_{h,i}} Y_{h,i}\right)^2 \end{array}\right\}$$

$$= \sum_{i=1}^{L_h} \left(\frac{g_{h,i}}{r_h \pi_{h,i}} \right)^2 \left(\begin{array}{l} E\left\{ \delta_{h,i}^2 \right\} E\left\{ \left(\sum_{j'=1}^{M'_{h,i}} \delta'_{h,i,j} Y'_{h,i,j} + \sum_{j^*=1}^{M^*_{h,i}} k_h \delta^*_{h,i,j} Y^*_{h,i,j} \right)^2 \right\} \\ -2E\left\{ \delta_{h,i} \right\} E\left\{ \sum_{j'=1}^{M'_{h,i}} \delta'_{h,i,j} Y'_{h,i,j} + \sum_{j^*=1}^{M^*_{h,i}} k_h \delta^*_{h,i,j} Y^*_{h,i,j} \right\} \left(\frac{\pi_{h,i}\, r_h}{g_{h,i}} Y_{h,i} \right) \\ + \left(\frac{\pi_{h,i}\, r_h}{g_{h,i}} Y_{h,i} \right)^2 \end{array} \right)$$

(independence again)

$$= \sum_{i=1}^{L_h} \left(\frac{g_{h,i}}{r_h \pi_{h,i}} \right)^2 \left(\begin{array}{l} \pi_{h,i}\, E\left\{ \left(\sum_{j'=1}^{M'_{h,i}} \delta'_{h,i,j} Y'_{h,i,j} + \sum_{j^*=1}^{M^*_{h,i}} k_h \delta^*_{h,i,j} Y^*_{h,i,j} \right)^2 \right\} \\ -2\, \pi_{h,i} \left(\frac{r_h}{g_{h,i}} Y_{h,i} \right) \left(\frac{\pi_{h,i}\, r_h}{g_{h,i}} Y_{h,i} \right) + \left(\frac{\pi_{h,i}\, r_h}{g_{h,i}} Y_{h,i} \right)^2 \end{array} \right)$$

$$= \sum_{i=1}^{L_h} \left(\frac{g_{h,i}}{r_h \pi_{h,i}} \right)^2 \left(\pi_{h,i}\, E\left\{ \left(\sum_{j'=1}^{M'_{h,i}} \delta'_{h,i,j} Y'_{h,i,j} + \sum_{j^*=1}^{M^*_{h,i}} k_h \delta^*_{h,i,j} Y^*_{h,i,j} \right)^2 \right\} - \left(\frac{\pi_{h,i}\, r_h}{g_{h,i}} Y_{h,i} \right)^2 \right)$$

$$= \sum_{i=1}^{L_h} \left(\frac{g_{h,i}}{r_h \pi_{h,i}} \right)^2 \left(\begin{array}{l} \pi_{h,i}\, E^2 \left\{ \sum_{j'=1}^{M'_{h,i}} \delta'_{h,i,j} Y'_{h,i,j} + \sum_{j^*=1}^{M^*_{h,i}} k_h \delta^*_{h,i,j} Y^*_{h,i,j} \right\} \\ +\pi_{h,i}\, Var \left\{ \sum_{j'=1}^{M'_{h,i}} \delta'_{h,i,j} Y'_{h,i,j} + \sum_{j^*=1}^{M^*_{h,i}} k_h \delta^*_{h,i,j} Y^*_{h,i,j} \right\} \\ - \left(\frac{\pi_{h,i}\, r_h}{g_{h,i}} Y_{h,i} \right)^2 \end{array} \right)$$

$$= \sum_{i=1}^{L_h} \left(\frac{g_{h,i}}{r_h\, \pi_{h,i}} \right)^2 \left(\begin{array}{l} \pi_{h,i} \left(\sum_{j'=1}^{M'_{h,i}} Var\left\{ \delta'_{h,i,j} Y'_{h,i,j} \right\} + \sum_{j^*=1}^{M^*_{h,i}} Var\left\{ k_h \delta^*_{h,i,j} Y^*_{h,i,j} \right\} \right) \\ +\pi_{h,i} \left(\frac{r_h}{g_{h,i}} Y_{h,i} \right)^2 - \left(\frac{\pi_{h,i}\, r_h}{g_{h,i}} Y_{h,i} \right)^2 \end{array} \right)$$

(independence again)

Note that:

$$Var\left(\delta^{'}_{h,i,j}\right) = \pi^{'}_{h,i,j}\left(1-\pi^{'}_{h,i,j}\right) = \frac{r_h}{g_{h,i}}\left(\frac{g_{h,i}-r_h}{g_{h,i}}\right)$$

$$Var\left(\delta^{*}_{h,i,j}\right) = \pi^{*}_{h,i,j}\left(1-\pi^{*}_{h,i,j}\right) = \frac{r_h}{k_h g_{h,i}}\left(\frac{k_h g_{h,i}-r_h}{k_h g_{h,i}}\right)$$

Therefore:

$$Var\left(\hat{Y}_h\right) = \sum_{i=1}^{L_h}\left(\frac{g_{h,i}}{r_h\pi_{h,i}}\right)^2 \left(\begin{array}{l} \pi_{h,i}\left(\begin{array}{l} \displaystyle\sum_{j'=1}^{M^{'}_{h,i}} Y^{'2}_{h,i,j}\frac{r_h}{g_{h,i}}\left(\frac{g_{h,i}-r_h}{g_{h,i}}\right)+ \\ \displaystyle\sum_{j^*=1}^{M^{*}_{h,i}} k_h^2 Y^{*2}_{h,i,j}\frac{r_h}{k_h g_{h,i}}\left(\frac{k_h g_{h,i}-r_h}{k_h g_{h,i}}\right)\end{array}\right) \\ +\pi_{h,i}\left(\dfrac{r_h}{g_{h,i}}Y_{h,i}\right)^2-\left(\dfrac{\pi_{h,i}r_h}{g_{h,i}}Y_{h,i}\right)^2 \end{array}\right)$$

$$= \sum_{i=1}^{L_h}\left(\frac{1}{r_h\pi_{h,i}}\right)\left(\begin{array}{l} \displaystyle\sum_{j'=1}^{M^{'}_{h,i}} Y^{'2}_{h,i,j}\left(g_{h,i}-r_h\right)+\sum_{j^*=1}^{M^{*}_{h,i}} Y^{*2}_{h,i,j}\left(k_h g_{h,i}-r_h\right) \\ +\left(r_h Y_{h,i}\right)^2-\pi_{h,i}\left(r_h Y_{h,i}\right)^2 \end{array}\right)$$

$$= \sum_{i=1}^{L_h}\left(\frac{1}{r_h\pi_{h,i}}\right)\left(\begin{array}{l} \displaystyle\sum_{j'=1}^{M^{'}_{h,i}} Y^{'2}_{h,i,j}\left(g_{h,i}-r_h\right)+\sum_{j^*=1}^{M^{*}_{h,i}} Y^{*2}_{h,i,j}\left(g_{h,i}-r_h\right) \\ \displaystyle+\sum_{j^*=1}^{M^{*}_{h,i}} Y^{*2}_{h,i,j}\left(k_h-1\right)g_{h,i}+\left(r_h Y_{h,i}\right)^2-\pi_{h,i}\left(r_h Y_{h,i}\right)^2 \end{array}\right)$$

$$= \sum_{i=1}^{L_h}\left(\frac{1}{r_h\pi_{h,i}}\right)\left(\begin{array}{l} \displaystyle\left(g_{h,i}-r_h\right)\sum_{j=1}^{M_{h,i}} Y^{2}_{h,i,j}+\sum_{j^*=1}^{M^{*}_{h,i}} Y^{*2}_{h,i,j}\left(k_h-1\right)g_{h,i} \\ +\left(1-\pi_{h,i}\right)\left(r_h Y_{h,i}\right)^2 \end{array}\right)$$

$$= \sum_{i=1}^{L_h}\left(\frac{g_{h,i}-r_h}{r_h\pi_{h,i}}\right)\left(\left(1-\pi_{h,i}\right)\left(r_h Y_{h,i}\right)^2+\sum_{j=1}^{M_{h,i}} Y^{2}_{h,i,j}\right)+\left(k_h-1\right)\sum_{i=1}^{L_h}\left(\frac{g_{h,i}}{r_h\pi_{h,i}}\sum_{j^*=1}^{M^{*}_{h,i}} Y^{*2}_{h,i,j}\right)$$

$$\cong \sum_{i=1}^{L_h}\left(\frac{g_{h,i}-r_h}{r_h\pi_{h,i}}\right)\left(\left(1-\pi_{h,i}\right)\left(r_h Y_{h,i}\right)^2+\sum_{j=1}^{M_{h,i}} Y^{2}_{h,i,j}\right)+\left(k_h-1\right)\sum_{i=1}^{L_h}\left(\frac{g_{h,i}}{r_h\pi_{h,i}}M^{*}_{h,i}\left(\mu_h^2+\nu_h\right)\right)$$

$$\cong \sum_{i=1}^{L_h}\left(\frac{g_{h,i}-r_h}{r_h\pi_{h,i}}\right)\left(\left(1-\pi_{h,i}\right)\left(r_h Y_{h,i}\right)^2+\sum_{j=1}^{M_{h,i}} Y^{2}_{h,i,j}\right)+\left(k_h-1\right)\sum_{i=1}^{L_h}\left(\frac{g_{h,i}}{r_h\pi_{h,i}}M_{h,i}\,t_h\left(\mu_h^2+\nu_h\right)\right)$$

When size-0 skip equals 1, the second term vanishes. If CD selection probabilities are equal within a stratum, i.e. $\pi_{h,i} = \pi_{h,1}$, we can therefore estimate the increased variance due to size-0 skip as:

$$Var\left(\hat{Y}_h\right)_{\text{Poisson, stage-3 variance zero, size-0 skip=k}} - Var\left(\hat{Y}_h\right)_{\text{Poisson, stage-3 variance zero, size-0 skip=1}}$$

$$\cong \left(k_h - 1\right)\frac{t_h\left(\mu_h^2 + \nu_h\right)}{\pi_{h,1} r_h}\sum_{i=1}^{L_h}\left(g_{h,i} M_{h,i}\right)$$

A.6 Variance modelling for SMB-facilitated CD design with size-0 skip

When size-0 skip equals 1, the SMB-facilitated CD design is equivalent to the CD design discussed in Section 4 above:

$$Var\left(\hat{Y}_h\right)_{\text{SRSWOR, stage-3 variance non-zero, size-0 skip=1}} \cong a_h + b_{1,h}\frac{1}{l_h}$$

where

$$\begin{aligned}
a_h &= -L_h S_h^2 \\
b_{1,h} &= L_h^2 S_h^2 + L_h \sum_{i=1}^{L_k} \frac{M_{h,i}^2}{m_{h,i}} \left(1 - \frac{m_{h,i}}{M_{h,i}}\right) S_{h,i}^2 \\
&\quad + L_h \sum_{i=1}^{L_h} \frac{M_{h,i}}{m_{h,i}} \sum_{j=1}^{M_{h,i}} \frac{N_{h,i,j}^2}{\max(1, n_{h,i,j})} \left(1 - \frac{n_{h,i,j}}{\max(1, N_{h,i,j})}\right) S_{h,i,j}^2
\end{aligned}$$

We know that $Var(\hat{Y}_h)_{\text{Poisson, stage-3 variance zero, size-0 skip=k}}$ may not be a good estimator of $Var(\hat{Y}_h)_{\text{SRSWOR, stage-3 variance non-zero, size-0 skip=k}}$; the former is affected by variable sample size, while the latter is affected by selection within households.

However, if we assume that interactions between these effects and the effect of size-0 skip are relatively small, we can use the result for the Poisson design to provide a first-order estimate of how size-0 skip affects the SRSWOR design:

$$\begin{aligned}
&Var\left(\hat{Y}_h\right)_{\text{SRSWOR, stage-3 variance non-zero, size-0 skip=k}} - Var\left(\hat{Y}_h\right)_{\text{SRSWOR, stage-3 variance non-zero, size-0 skip=1}} \\
&\cong Var\left(\hat{Y}_h\right)_{\text{Poisson, stage-3 variance zero, size-0 skip=k}} - Var\left(\hat{Y}_h\right)_{\text{Poisson, stage-3 variance zero, size-0 skip=1}}
\end{aligned}$$

Therefore,

$$\begin{aligned}
&Var\left(\hat{Y}_h\right)_{\text{SRSWOR, stage-3 variance non-zero, size-0 skip=k}} \\
&\cong a_h + \frac{b_{1,h}}{l_h} + (k_h - 1)\frac{t_h(\mu_h^2 + \nu_h)}{\pi_{h,1} r_h} \sum_{i=1}^{L_h} (g_{h,i} M_{h,i}) \\
&= a_h + \frac{b_{1,h}}{l_h} + (k_h - 1)\frac{t_h(\mu_h^2 + \nu_h)}{l_h r_h} \sum_{i=1}^{L_h} (g_{h,i} M_{h,i}) \\
&= a_h + \frac{b_h}{l_h}
\end{aligned}$$

where

$$\begin{aligned}
b_h &= b_{1,h} + (k_h - 1)\, b_{2,h} \\
b_{2,h} &= \frac{t_h(\mu_h^2 + \nu_h)}{r_h} \sum_{i=1}^{L_h} (g_{h,i} M_{h,i})
\end{aligned}$$

With this modification, we can estimate the variance for a given allocation, and quickly find the optimal allocation under specified size-0 and dwelling skips. The size-0 skips could in theory be optimised along with the allocation; in practice they were manually adjusted due to the complexity of the accuracy, cost, and operational requirements involved.

Note: The NATSISS design used an earlier version of this size-0 skip adjustment, in which $g_{h,i}$ was assumed to be 1. Since CDs with a large enough Indigenous population to have a dwelling skip are rare, and have relatively few size-0 SMBs, the effects of this simplification are unlikely to be significant.

Note also: Late in the design process, it was decided that it would be undesirable to combine size-0 skip and dwelling skip within the same CDs – therefore, after variance modelling had been completed and CD-level selection probabilities were set, all CDs large enough to have a dwelling skip greater than 1 were given a size-0 skip of 1 when selecting SMBs. Since remote CDs already had size-0 skip set to 1, this change only affected size-0 SMBs within non-remote CDs that had a large Indigenous population; such CDs are rare and generally have few size-0 SMBs, so the resulting increase in accuracy and sample size is expected to be negligible.

A.6.1 Variance when size-0 SMBs are excluded

For variance purposes, excluding size-0 SMBs from sampling in certain strata is equivalent to a full-CD design, using the same dwelling skips, where only persons in non-size-0 SMBs are treated as 'in-scope'. Under the assumption that households in size-0 SMBs have similar characteristics to other households in the same strata, this can be addressed by the methods given in Section 4. Replacing CD household count with $M_{h,i}(1-t_h)$ and sample count with $M_{h,i}(1-t_h)r_h/g_{h,i}$ and scaling the variance of CD totals from S_h^2 to $S_h^2(1-t_h)^2$ (assuming that undercount scales totals for each CD by approximately $(1-t)$) leads to a modified version of the b-coefficient:

$$\begin{aligned} b'_{1,h} &= L_h^2 S_h^2 (1-t_h)^2 + L_h \sum_{i=1}^{L_k} \frac{g_{h,i} M_{h,i} (1-t_h)}{r_h} \left(1 - \frac{r_h}{g_{h,i}}\right) S_{h,i}^2 \\ &+ (1-t_h) L_h \sum_{i=1}^{L_h} \frac{g_{h,i}}{r_h} \sum_{j=1}^{M_{h,i}} \frac{N_{h,i,j}^2}{\max(1, n_{h,i,j})} \left(1 - \frac{n_{h,i,j}}{\max(1, N_{h,i,j})}\right) S_{h,i,j}^2 \end{aligned}$$

This modified b-coefficient is smaller than that for a full-CD design covering size-0s (it has been scaled by a factor of $(1-t_h)^2$ in the first term, and a factor of $(1-t_h)$ in the other two; note that in the full-CD design, $m_{h,i} = M_{h,i} r_h / g_{h,i}$, reducing variance for stratum estimates accordingly.

However, the total for response variables among in-scope persons also decreases, being scaled by a factor of $(1-t_h)$, meaning that a lower variance is required to achieve the same RSE.

Due to an oversight this was not implemented in the NATSISS sample design; instead, response variable totals were left unchanged and b_h was calculated as for the size-0 skip design without any reduction for the smaller sample and surveyable population. For an individual stratum, we expect that the resulting proportional error in RSE would be on the order of:

$$\frac{RSE_{est} - RSE_{true}}{RSE_{true}} = \frac{\dfrac{SE_{est}}{Y_{est}} - \dfrac{SE_{true}}{Y_{true}}}{\dfrac{SE_{true}}{Y_{true}}}$$

$$= \frac{\dfrac{\sqrt{Var_{est}}}{Y_{est}} - \dfrac{\sqrt{Var_{true}}}{Y_{true}}}{\dfrac{\sqrt{Var_{true}}}{Y_{true}}}$$

$$\cong \frac{\dfrac{\sqrt{Var_{est}}}{Y_{est}} - \dfrac{\sqrt{(1-t_h)Var_{est}}}{(1-t_h)\,Y_{est}}}{\dfrac{\sqrt{(1-t_h)Var_{est}}}{(1-t_h)\,Y_{est}}}$$

Examination of the data indicates that in general:

$$\|a_h\| \ll \frac{b_h}{l_h}, \frac{b_h^{'}}{l_h}$$

Therefore,

$$Var\left(\hat{Y}\right) \cong \frac{b_h^{'}}{l_h}$$

Therefore, substituting in the correct and erroneous values of b_h':

$$\frac{RSE_{est} - RSE_{true}}{RSE_{true}} \cong \frac{\dfrac{\sqrt{b_{h,err}^{'} / l_h}}{Y_{est}} - \dfrac{\sqrt{b_h^{'} / l_h}}{(1-t_h)Y_{est}}}{\dfrac{\sqrt{b_h^{'} / l_h}}{(1-t_h)Y_{est}}}$$

$$\cong \frac{\dfrac{\sqrt{b_{h,err}^{'}}}{Y_{est}} - \dfrac{\sqrt{(1-t_h)\, b_{h,err}^{'}}}{(1-t_h)Y_{est}}}{\dfrac{\sqrt{(1-t_h)\, b_{h,err}^{'}}}{(1-t_h)Y_{est}}}$$

$$= \frac{(1-t_h) - \sqrt{(1-t_h)}}{\sqrt{(1-t_h)}}$$

$$= \sqrt{(1-t_h)} - 1$$

$$\cong -\frac{t_h}{2}$$

For the highest t_h (~20% for urban Victoria), this would translate to approximately 10% proportional error in predicted RSE for strata where size-0 SMBs are excluded. Note that within each output region (e.g. state/RA), only some strata exclude size-0 SMBs, so the overall resulting error at these levels will be significantly less than 10%; given the general level of uncertainty in predicting RSEs, this is not likely to be a serious concern, but should be corrected in future.

B. COST MODELS FOR THE NON-COMMUNITY SAMPLE

NATSISS cost modelling was based on the cost model used for IHS 04/05. Within a region (generally state/broad RA), sample-related cost is assumed to follow a linear model of the form:

$$Cost = \chi^{adult} I^{adult} + \chi^{child} I^{child} + \psi S + \theta T + \phi l + \varepsilon$$

Definition of terms:

I^{adult} = number of adults interviewed

I^{child} = number of children interviewed

S = number of households screened

T = number of kilometres travelled

l = number of PSUs sampled

m = number of Indigenous households interviewed

The values of χ^{adult}, χ^{child}, ψ, θ and ϕ are estimated from operational data (payment systems, records from previous surveys, etc.) and modified as appropriate for inflation etc.; we also assume that T is proportional to l and estimate typical values of

$$\bar{T} \cong \frac{T}{l}$$

(i.e. the distance travelled per PSU sampled). Response data from previous surveys is used to estimate stratum-level hit rates r_h (defined as in A.1). This hit rate varies by area type, with IHS 04/05 recording an overall national level of around 65%. The deficiency includes genuine non-response (~ 15%), undercoverage of persons who migrate to size-0 CDs (causing a net reduction in the population of identified non-size-0 CDs, ~ 5%), and an unknown component (~ 15%).

Frame data (i.e. the most recent Census) provides estimates of household totals for each CD:

$M^{I}_{h,i,1}$ = number of Indigenous households reported as of Census (by definition, all are in Census non-size-0 SMBs).

$M^{T}_{h,i,1}$ = total households (Indigenous and non-Indigenous) reported in Census non-size-0 SMBs.

$M^{T}_{h,i,0}$ = total households (by definition, all non-Indigenous) reported in Census size-0 SMBs.

$M^{T}_{h,i}$ = total households within the CD.

Recalling that k_h is the size-0 skip (set to 1 for simple CD designs), $g_{h,i}$ is the CD's dwelling skip, and fraction t_h have migrated to size-0 SMBs between Census and survey, we can estimate the total number of Indigenous households interviewed in each CD, if that CD is selected:

$$m_{h,i} = \frac{r_h}{g_{h,i}}(1-t_h)M^I_{h,i,1} + \frac{r_h}{g_{h,i}k_h}t_h M^I_{h,i,1} = \frac{r_h M^I_{h,i,1}}{g_{h,i}}\left(1-t_h+\frac{t_h}{k_h}\right)$$

We can also estimate the total number of households screened in that CD:

$$m^{screened}_{h,i} = \frac{1}{g_{h,i}}\left(M^T_{h,i,1}+\frac{M^T_{h,i,0}}{k_h}\right)$$

This lets us estimate the number of households screened per Indigenous household interviewed for each stratum:

$$\frac{m^{screened}_{h,i}}{m_h} = \frac{\sum_{i=1}^{L_h}\frac{1}{g_{h,i}}\left(M^T_{h,i,1}+\frac{M^T_{h,i,0}}{k_h}\right)}{\sum_{i=1}^{L_h}\frac{r_h M^I_{h,i,1}}{g_{h,i}}\left(1-t_h+\frac{t_h}{k_h}\right)}$$

Frame data also gives us numbers of Indigenous adults and children in each household. Combined with the interviewing rule for the region (i.e. whether we will select one or two of each), this allows us to calculate the average numbers of adults and children interviewed per Indigenous household interviewed, for each stratum ($\overline{I}_h^{adult}$ and $\overline{I}_h^{child}$).

We can then estimate sample-related costs in each stratum as a linear function of CDs and households sampled:

$$\begin{aligned} Cost_h &\cong \chi_h^{adult} I_h^{adult} + \chi_h^{child} I_h^{child} + \psi_h S_h + \theta_h T_h + \phi_h l_h \\ &= \chi_h^{adult}\overline{I}_h^{adult} m_h + \chi_h^{child}\overline{I}_h^{child} m_h + \psi_h m_h \frac{\sum_{i=1}^{L_h}\frac{1}{g_{h,i}}\left(M^T_{h,i,1}+\frac{M^T_{h,i,0}}{k_h}\right)}{\sum_{i=1}^{L_h}\frac{r_h M^I_{h,i,1}}{g_{h,i}}\left(1-t_h+\frac{t_h}{k_h}\right)} + \theta_h \overline{T}_h l_h + \phi_h l_h \end{aligned}$$

B.1 Cost modelling for cluster size optimisation

Section 3 above deals with an approach in which size-0 skip is 1 and dwelling skips are set to achieve an approximately uniform cluster size q_h for all CDs within each stratum h, i.e.

$$g_{h,i} \cong \frac{r_h M^I_{h,i,1}}{q_h}.$$

Under these conditions, the above cost formula simplifies:

$$\begin{aligned}
Cost_h &\cong \chi_h^{adult} \bar{I}_h^{adult} m_h + \chi_h^{child} \bar{I}_h^{child} m_h + \psi_h m_h \frac{\sum_{i=1}^{L_h} \frac{1}{g_{h,i}} \left(M^T_{h,i,1} + M^T_{h,i,0} \right)}{\sum_{i=1}^{L_h} \frac{r_h M^I_{h,i,1}}{g_{h,i}}} + \theta_h \bar{T} l_h + \phi_h l_h \\
&\cong \chi_h^{adult} \bar{I}_h^{adult} m_h + \chi_h^{child} \bar{I}_h^{child} m_h + \psi_h m_h \frac{\sum_{i=1}^{L_h} \frac{q_h}{r_h M^I_{h,i,1}} M^T_{h,i}}{\sum_{i=1}^{L_h} \frac{q_h r_h M^I_{h,i,1}}{r_h M^I_{h,i,1}}} + \theta_h \bar{T} l_h + \phi_h l_h \\
&= \chi_h^{adult} \bar{I}_h^{adult} m_h + \chi_h^{child} \bar{I}_h^{child} m_h + \psi_h m_h \frac{\frac{1}{r_h} \sum_{i=1}^{L_h} \frac{M^T_{h,i}}{M^I_{h,i,1}}}{L_h} + \theta_h \bar{T} l_h + \phi_h l_h \\
&= l_h \left(\theta_h \bar{T} + \phi_h \right) + m_h \left(\chi_h^{adult} \bar{I}_h^{adult} + \chi_h^{child} \bar{I}_h^{child} + \psi_h \frac{1}{r_h L_h} \sum_{i=1}^{L_h} \frac{M^T_{h,i}}{M^I_{h,i,1}} \right) \\
&= l_h e_h + m_h f_h
\end{aligned}$$

where:

$$\begin{aligned}
e_h &= \left(\theta_h \bar{T} + \phi_h \right) \\
f_h &= \chi_h^{adult} \bar{I}_h^{adult} + \chi_h^{child} \bar{I}_h^{child} + \psi_h \frac{1}{r_h L_h} \sum_{i=1}^{L_h} \frac{M^T_{h,i}}{M^I_{h,i,1}}
\end{aligned}$$

B.2 Cost modelling for size-0 skip

In the size-0 skip designs discussed here, dwelling skips $g_{h,i}$ are treated as known (equal to 1 for most CDs). For a given value of k_h, we can estimate m_h as a function of l_h:

$$\begin{aligned}
m_h &= \sum_{i=1}^{L_h} \delta_{h,i} \frac{r_h M^I_{h,i,1}}{g_{h,i}} \left(1 - t_h + \frac{t_h}{k_h}\right) \\
&\cong E\left\{ \sum_{i=1}^{L_h} \delta_{h,i} \frac{r_h M^I_{h,i,1}}{g_{h,i}} \left(1 - t_h + \frac{t_h}{k_h}\right) \right\} \\
&= \sum_{i=1}^{L_h} E\left(\delta_{h,i}\right) \frac{r_h M^I_{h,i,1}}{g_{h,i}} \left(1 - t_h + \frac{t_h}{k_h}\right) \\
&= \sum_{i=1}^{L_h} \frac{l_h}{L_h} \frac{r_h M^I_{h,i,1}}{g_{h,i}} \left(1 - t_h + \frac{t_h}{k_h}\right) \\
&= \frac{l_h r_h}{L_h} \left(1 - t_h + \frac{t_h}{k_h}\right) \sum_{i=1}^{L_h} \frac{M^I_{h,i,1}}{g_{h,i}}
\end{aligned}$$

The cost function then becomes:

$$Cost_h \cong l_h \, e_h$$

where

$$\begin{aligned}
e_h = {} & \theta_h \overline{T} + \phi_h + \frac{r_h}{L_h} \left(1 - t_h + \frac{t_h}{k_h}\right) \left(\sum_{i=1}^{L_h} \frac{r_h M^I_{h,i,1}}{g_{h,i}} \right) \times \\
& \left(\chi_h^{adult} \overline{I}_h^{adult} + \chi_h^{child} \overline{I}_h^{child} + \psi_h \frac{\displaystyle\sum_{i=1}^{L_h} \frac{1}{g_{h,i}} \left(M^T_{h,i,1} + \frac{M^T_{h,i,0}}{k_h} \right)}{\displaystyle\sum_{i=1}^{L_h} \frac{r_h M^I_{h,i,1}}{g_{h,i}} \left(1 - t_h + \frac{t_h}{k_h}\right)} \right)
\end{aligned}$$

C. COST/VARIANCE MODELS FOR THE COMMUNITY SAMPLE

The NATSISS design includes five 'community' strata: one each in South Australia, Western Australia, and the Northern Territory, and two in Queensland, where communities are divided into Torres Strait Islands and non-TSI communities. Within each stratum, communities are formed into 'sets' of approximately 30 persons.

Due to the limitations of available frame data for communities, variance modelling for these strata uses a simple design-effect approach:

$$\begin{aligned}
Var\left(\hat{Y}_h\right) &\cong \frac{N_h^2 p(1-p)}{n_h} deff_h \\
&= \frac{1}{l_h} \frac{N_h^2 p(1-p)}{\overline{n}_h} deff_h \\
&= \frac{1}{l_h} b_h \ , \\
b_h &= \frac{N_h^2 p(1-p)}{\overline{n}_h} deff_h
\end{aligned}$$

where l_h here represents the number of community sets sampled, n_h is the number of persons sampled, $\overline{n}_h$ is the mean number of persons sampled per community set, N_h is the total number of persons in the stratum, and p is the prevalence of the variable of interest. Deffs are estimated based on previous surveys, and in practice p was approximated by the remote non-community prevalence.

Average community sampling costs for each stratum are estimated based on travel requirements, etc.. For purposes of allocation, these are then converted to average costs per person interviewed. (In practice, per-person costs are difficult to predict accurately, both because selection effectively rounds the specified number of people interviewed to an integer number of community sets, and because costs may vary widely between different communities in the same stratum. For instance, some communities may be reached by road, while others require chartering boats or aircraft.)

D. HOUSEHOLD TWO-YEAR MIGRATION RATES

This table gives the estimated proportion of Indigenous households, as recorded by Census '06, who had migrated to their Census address within the previous two years (i.e. households in which none of the Indigenous people living there on Census night had been living there two years previously). For further discussion, see Section 6.

D.1 Migration rates, by state and remoteness area

State and Remoteness area	*Two-year Indigenous household migration rate*
New South Wales	
0	25.5%
1	27.4%
2	22.9%
3	20.0%
4	14.4%
Victoria	
0	26.9%
1	27.4%
2	25.3%
3	15.6%
Queensland	
0	33.1%
1	33.8%
2	26.7%
3	20.7%
4	10.3%
South Australia	
0	25.9%
1	27.5%
2	24.0%
3	27.2%
4	10.2%
Western Australia	
0	27.3%
1	31.6%
2	28.7%
3	23.5%
4	12.6%
Tasmania	
1	26.2%
2	22.8%
3	23.1%
4	19.4%
Northern Territory	
2	24.7%
3	19.8%
4	4.5%
Australian Capital Territory	
0–1	28.1%

E. DESIGN EFFECTS BY STRATUM

This table gives the population (Indigenous persons) and design effects ('deffs') for each NATSISS stratum. These design effects are based on variances calculated for the final sample allocation, using the approximation given in A.6 (see Appendix C for community strata). In some strata the sampling fractions are large enough to significantly affect the deffs shown.

Non-community strata are identified by a four-digit code indicating state (first digit), remoteness (second digit), and minimum number of Indigenous households per CD in stratum (third and fourth digits). For instance, stratum 1021 comprises CDs in New South Wales (state 1), major urban (RA 0), with 21–25 Indigenous households (upper boundary indicated by the next stratum listed, 1026). Some states have a single community stratum, and Queensland also has a separate Torres Strait Islander community stratum.

E.1 Design effects by stratum, New South Wales

New South Wales stratum	*Persons in stratum*	*Deff*
1001	2,752	0.75
1002	4,256	1.32
1003	4,558	1.54
1004	4,710	1.19
1005	4,410	1.32
1006	4,363	1.35
1007	4,112	1.28
1008	3,060	1.07
1009	3,489	1.80
1010	2,615	1.61
1011	8,933	1.76
1016	5,285	1.61
1021	2,928	1.72
1026	4,532	1.97
1101	750	0.69
1102	1,416	0.98
1103	1,955	1.96
1104	2,382	1.47
1105	2,396	1.30
1106	2,469	1.52
1107	2,022	1.22
1108	2,187	1.68
1109	2,253	1.50
1110	2,027	1.91
1111	8,235	1.93
1116	4,905	1.92
1121	2,738	2.27
1126	2,733	1.75
1131	4,894	2.59
1151	4,488	2.35
1201	644	5.36
1202	855	1.57
1203	1,110	1.52
1204	918	1.58
1205	1,170	1.13
1206	4,331	1.69
1211	3,221	2.25
1216	2,522	1.58
1221	3,584	1.79
1231	3,082	2.31
1251	4,928	3.37
1305	550	2.81
1311	1,007	1.60
1331	4,603	2.56
1421	868	1.36

E.2 Design effects by stratum, Victoria

Victorian stratum	*Persons in stratum*	*Deff*	*Victorian stratum*	*Persons in stratum*	*Deff*
2001	3,345	0.72	2108	451	1.18
2002	3,579	1.79	2109	447	1.11
2003	2,826	1.51	2110	210	1.22
2004	1,956	1.36	2111	784	2.06
2005	1,213	1.38	2116	728	1.21
2006	764	1.37	2121	727	1.50
2007	443	2.11	2201	272	0.55
2008	123	1.02	2202	335	0.98
2009	363	0.88	2203	371	1.05
2011	152	2.54	2204	352	1.58
2101	1,111	0.87	2205	254	1.45
2102	1,517	0.82	2206	224	1.75
2103	1,349	1.18	2207	217	1.21
2104	1,130	1.53	2208	204	2.33
2105	812	1.24	2209	528	1.84
2106	808	1.06	2211	1,116	1.75
2107	600	1.24	2221	999	1.80

E.3 Design effects by stratum, Queensland

Queensland stratum	*Persons in stratum*	*Deff*	*Queensland stratum*	*Persons in stratum*	*Deff*
3001	1,269	1.16	3121	3,296	2.17
3002	2,370	1.06	3201	276	0.99
3003	3,083	1.26	3202	502	1.48
3004	3,414	1.40	3203	781	1.13
3005	3,309	1.08	3204	811	1.70
3006	3,292	1.53	3205	879	1.63
3007	2,678	1.36	3206	1,222	1.42
3008	2,724	1.39	3207	928	1.36
3009	1,980	1.83	3208	1,210	1.44
3010	1,915	1.43	3209	1,095	2.21
3011	6,235	2.09	3210	1,168	1.84
3016	2,528	1.74	3211	5,166	1.69
3021	2,859	1.91	3216	5,421	2.15
3101	1,788	5.62	3221	7,365	2.39
3102	804	0.91	3231	3,630	1.98
3103	1,216	1.32	3241	5,035	3.31
3104	1,401	1.21	3301	175	0.71
3105	1,756	1.73	3303	339	1.10
3106	1,575	1.13	3305	742	1.26
3107	1,699	1.53	3311	2,096	2.21
3108	1,869	1.49	3326	3,567	4.32
3109	1,535	2.21	3405	927	1.21
3110	1,271	2.15	3421	2,469	6.77
3111	5,777	1.66	Community, non-TSI	14,352	2.50
3116	2,741	2.25	Community, TSI	8,423	1.40

E.4 Design effects by stratum, South Australia

South Australian stratum	*Persons in stratum*	*Deff*	*South Australian stratum*	*Persons in stratum*	*Deff*
4001	881	1.08	4201	201	0.99
4002	1,477	1.18	4202	289	3.28
4003	1,560	0.92	4203	395	1.17
4004	1,552	1.44	4204	189	0.81
4005	1,480	2.66	4205	239	1.19
4006	1,271	2.00	4206	372	0.99
4007	818	1.61	4207	267	0.99
4008	1,019	1.31	4208	193	2.06
4009	575	2.47	4209	373	0.50
4010	590	1.31	4211	639	1.82
4011	1,283	1.77	4216	319	0.65
4016	314	0.75	4221	979	2.29
4101	179	0.53	4231	734	1.87
4102	330	1.10	4241	515	2.39
4103	236	1.29	4305	320	1.93
4104	275	1.24	4311	571	2.12
4105	220	0.85	4403	304	1.49
4106	614	1.31	4411	968	1.97
4111	443	1.42	Community	2,469	2.50

E.5 Design effects by stratum, Western Australia

Western Australian stratum	*Persons in stratum*	*Deff*	*Western Australian stratum*	*Persons in stratum*	*Deff*
5001	1,184	6.35	5111	1,555	3.07
5002	1,579	2.35	5201	261	1.84
5003	1,776	1.44	5202	217	1.86
5004	1,864	1.10	5203	253	0.69
5005	1,508	1.35	5204	306	1.33
5006	1,568	2.48	5205	271	0.88
5007	1,558	2.42	5206	1,247	1.49
5008	1,606	1.95	5211	1,334	1.39
5009	1,700	2.25	5216	2,588	1.42
5010	1,259	2.01	5231	2,305	3.50
5011	3,674	2.25	5305	792	1.78
5016	2,319	1.81	5311	1,414	1.89
5101	208	1.63	5321	2,787	2.53
5102	429	1.84	5351	3,092	2.18
5103	413	1.17	5405	742	2.41
5104	371	2.76	5411	1,158	2.74
5105	373	1.57	5431	2,878	3.30
5106	1,368	1.69	Community	10,403	2.30

E.6 Design effects by stratum, Tasmania

Tasmanian stratum	*Persons in stratum*	*Deff*	*Tasmanian stratum*	*Persons in stratum*	*Deff*
6101	50	0.59	6202	157	0.75
6102	217	0.91	6203	187	2.20
6103	360	1.32	6204	277	0.99
6104	335	1.01	6205	392	0.74
6105	461	1.48	6206	409	1.06
6106	594	0.79	6207	438	0.93
6107	602	1.35	6208	405	0.99
6108	486	0.87	6209	437	1.01
6109	505	0.95	6210	318	0.92
6110	576	1.30	6211	1,406	1.29
6111	1,858	1.38	6216	1,114	1.55
6116	1,245	1.07	6221	600	1.14
6121	743	2.03	6226	357	1.34
6126	323	1.48	6231	241	1.56
6131	754	1.88	6236	434	0.91
6201	70	0.60	6305	365	0.96

E.7 Design effects by stratum, Northern Territory

Northern Territory stratum	*Persons in stratum*	*Deff*	*Northern Territory stratum*	*Persons in stratum*	*Deff*
7201	880	1.52	7251	1,678	1.37
7211	1,077	2.14	7311	494	1.54
7216	1,044	1.59	7321	2,287	2.54
7221	877	1.52	7341	2,385	2.25
7226	1,734	1.55	7421	2,063	5.13
7231	3,153	1.67	Community	34,937	3.00

E.8 Design effects by stratum, Australian Capital Territory

Australian Capital Territory stratum	*Persons in stratum*	*Deff*	*Australian Capital Territory stratum*	*Persons in stratum*	*Deff*
8001	111	0.40	8007	372	3.68
8002	372	1.33	8008	383	1.48
8003	369	1.58	8009	181	2.67
8004	514	1.58	8010	250	0.82
8005	478	1.46	8011	267	2.09
8006	550	1.83			

F. COST COEFFICIENTS BY STATE AND REMOTENESS

This table gives estimated fixed-per-CD, screening, and adult interviewing costs for non-community CDs within each state and remoteness classification, as used for the cost model in Appendix B.

F.1 Cost coefficients, by state and remoteness

State and Remoteness area	*Fixed costs per CD or community set*	*Screening costs per household*	*Interview costs per adult*
New South Wales			
0	$718.49	$1.34	$226.35
1–2	$1,335.38	$1.54	$226.35
3–4	$780.76	$2.35	$226.35
Victoria			
0	$661.36	$1.85	$301.35
1–2	$1,186.96	$2.12	$301.35
Queensland			
0	$426.50	$1.52	$204.24
1–2	$512.45	$1.74	$204.24
3–4	$210.27	$2.65	$204.24
Community	$14,705.00	–	–
South Australia			
0	$524.69	$1.28	$223.90
1–2	$1,242.16	$1.46	$223.90
3–4	$2,053.92	$2.23	$223.90
Community	$13,400.00	–	–
Western Australia			
0	$888.59	$1.08	$189.85
1–2	$345.63	$1.23	$189.85
3–4	$750.77	$1.88	$189.85
Community	$14,705.00	–	–
Tasmania			
1–2	$1,216.39	$0.92	$141.08
3	$970.40	$1.41	$141.08
Northern Australia			
2	$212.18	$1.48	$196.45
3–4	$121.07	$2.26	$196.45
Community	$15,205.00	–	–
Australian Capital Territory			
0–1	$395.16	$1.14	$173.23

Note: Cost figures are given for illustrative purposes only and do not represent official ABS costings.

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

Produced by the Australian Bureau of Statistics

1352.0.55.096 • RESEARCH PAPER: SAMPLE DESIGN ISSUES FOR NATIONAL SURVEYS OF THE INDIGENOUS POPULATION (METHODOLOGY ADVISORY COMMITTEE) • June 2008