1352.0.55.108

Australian Bureau of Statistics

Research Paper

Sample Design Frameworks for ABS Household Surveys

www.abs.gov.au

New Issue

Research Paper

Sample Design Frameworks for ABS Household Surveys

Alistair Rogers, Justin Lokhorst and Julian Whiting

Statistical Services Branch

Methodology Advisory Committee

18 June 2010, Canberra

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) TUE 23 NOV 2010

ABS Catalogue no. 1352.0.55.108

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Mr Paul Schubert, Statistical Services Branch on Canberra (02) 6252 6591 or email <statistical.services@abs.gov.au>.

CONTENTS

SAMPLE DESIGN FRAMEWORKS FOR ABS HOUSEHOLD SURVEYS

Alistair Rogers, Justin Lokhorst and Julian Whiting
Statistical Services Branch

ABSTRACT

For many years the sampling framework used for ABS household surveys has been based around a single master sample of geographic areas selected after the five-yearly Population Census. The areas in the master sample are divided into finer blocks, with one block providing sample for the Monthly Population Survey (MPS), and a 'parallel' block used for a diverse range of large-scale social surveys, collectively referred to as Special Social Surveys (SSSs). Although the sample design parameters for the master sample have been tailored to the key MPS objectives, there is sufficient flexibility in parallel block sample designs that nearly all SSSs have used the parallel blocks. Growing demand for SSSs to meet a more diverse range of objectives may mean that there will be more surveys for which the parallel block will be less suitable. In addition, changes to the ABS geography standard and new methods and systems for sample preparation and maintenance reduce the cost advantage of tightly coupled samples for the MPS and SSSs. Considering these changes, alternative frameworks may be clearly superior for the future SSS program. This paper compares some alternative sampling frameworks for ABS household surveys, describing key sample design issues which need to be considered.

1. SUMMARY

The current sampling framework used for ABS household surveys has been in use for over 30 years. Underpinning this sampling framework has been a master sample of geographical areas reselected every five years, which provide the sample of dwellings for nearly all household surveys.

The ABS is investigating possible significant changes to the household survey sampling framework as an extension of the routine five-yearly Monthly Population Survey (MPS) sample redesign hinging off the 2011 Census of Population and Housing. This timing allows for implementation of a new framework in the 2012/13 financial year. The basic idea of the potential change to the sampling framework is to allow greater flexibility for the sample designs of individual surveys.

The changes being considered are motivated by numerous factors which are driving or enabling opportunities to modify the current sampling framework so that it supports the survey program in a more cost-efficient manner. The drivers include (but are not limited to):

- changes to the ABS geography classification structure;
- increasing demand for social statistics requiring sophisticated sample designs; and
- the need to find bookable efficiency gains for the organisation.

The enablers include (but are not limited to):

- a rebuild of sample management systems; and
- increasing availability of new technology including mapping and geocoding software.

The purpose of this paper is to describe the range of sample design issues which need to be considered for choosing the future sampling framework. The paper highlights the tension between offering flexibility in the possible designs for a particular survey and enforcing coordination in the sample designs to control program-wide costs and manage sample overlap.

Of the frameworks considered, a decoupled framework – where separate master samples for MPS and Special Social Surveys (SSSs) are created – seems the most logical way to address drivers and enablers for change as well as addressing emerging needs of household survey program.

Does the Committee have advice on whether this is a logical conclusion or are there key statistical aspects which have been overlooked in reaching this conclusion.

Other (specific) considerations for the Committee include:

1. Are there other ways in which the existing parallel sample approach could meet the emerging needs of the household survey program?
2. Geographically, spread and clustering of subpopulations of interest can be mapped and visually assessed. Can the Committee make any recommendations on analytical approaches to determining the required size of a separate SSS master sample, given it potentially needs to meet a range of different subpopulation surveys?
3. Due to lack of detailed data on cost breakdowns for survey enumeration, there is a reluctance to spend too much effort in detailed cost modelling approaches. Given this, can the Committee recommend any simple, yet effective, ways at capturing cost of various sample designs so that simple comparisons across frameworks can be made?
4. How sensitive are low sample fraction sample designs to choice of cluster size and other sample design parameters? What types of analyses could confirm (or otherwise) this behaviour? Are there sample design parameters that should be focussed on more than others?

2. ABS HOUSEHOLD SURVEY PROGRAM – PAST AND PRESENT

2.1 ABS household surveys

The two broad aims of the ABS population statistics program are to:

1. provide data to monitor the well-being of Australians with particular reference to important population subgroups; and
2. support the development and evaluation of Government policies and programs.

The program primarily provides data on social indicators but, particularly through the collection of labour statistics, it also provides data on economic indicators. The population statistics program is supported by the five-yearly Census of Population and Housing, a range of administrative statistics and a program of household surveys (McEwin, 2000).

The number and variety of household surveys conducted by the ABS has steadily increased over time. The ABS began conducting household surveys in 1960, when it first ran the quarterly Labour Force Survey (LFS). Supplementary questions on various topics were added to the LFS questionnaire shortly after, and this practice has continued to the present. The LFS has been conducted monthly since 1978, and today a supplementary topic is appended to the LFS questionnaire in most months. The combination of the LFS and the supplementary survey is referred to as the Monthly Population Survey (MPS).

Since the late 1970s the ABS has conducted stand-alone surveys separate to the MPS. Now referred to as SSSs, these surveys meet the needs for detailed, complex and in-depth data on specific topics and sometimes specific subpopulations. Fundamentally, the questionnaire length and requirement to conduct a personal interview make these surveys unsuitable to be run as a supplement to the LFS.

To provide an indication of the variety of topics and objectives of SSSs, table 2.1 summarises the surveys on the SSS program over the period between 2005 and 2011. The table summarises the geographic level of key objectives and whether population subgroups not defined by geographic boundaries were of particular interest. In practice, the design objectives serve to provide the relative importance of different estimates and are not expressed as rigid requirements for levels of sampling error. The number of sample dwellings for these surveys has typically been less than half the monthly sample of dwellings for the MPS.

2.1 Special Social Surveys program, 2005–2011

Survey	*Geographic design objectives*	*Specific populations of interest*
General Social Survey (GSS)	State and national estimates for wide range of social characteristics	2010 GSS has particular interest in characteristics of population with 'multiple social disadvantage'
Household Expenditure Survey (HES)	National estimates, some broad estimates at State level. Capital city estimates are priority for input into CPI weights.	2009/10 HES had top-up sample of households whose primary source of income is from pensions
National Health Survey (NHS)	State and national estimates for key health indicators	None for 2007/08 survey
Survey of Disability, Ageing and Carers (SDAC)	Sub-State estimates important for key disability indicators	Prevalence rates of disability in general population highest priority, but characteristics of persons with a disability also a key objective
Survey of Education and Training (SET)	State and national estimates for key educational and training indicators	Persons aged 15–64 (persons 65–74 still in or marginally attached to labour force also in process scope)
Survey of Employment Arrangements, Retirement and Superannuation (SEARS)	National and State estimates.	None (though focus of retirement topic is persons aged 45 and over)
Time Use Survey (TUS)	National estimates	None (though time use of persons under pressure and unemployed of high user interest)
Survey of Mental Health and Wellbeing (SMHWB)	Priority on national estimates and estimates by remoteness classification	For 2007 SMHWB persons aged 16–24 and 65–85 were of particular interest
Survey of Income and Housing (SIH)	National estimates, broad estimates at State level	None
National Aboriginal and Torres Strait Islander Social Survey (NATSISS)	Focus on national estimates, though breakdown by remoteness classification of interest	Indigenous population, with objectives specified for estimates relating to children
National Aboriginal and Torres Strait Islander Health Survey (NATSIHS)	National and State, as well as by remoteness class	Indigenous population
Personal safety survey	State-level estimates for broad indicators for females, national estimates for males	Greater focus on females
Survey of Adult Competencies (SAC, previously known as Adult Learning and Life Skills)	National estimates, and State estimates for broad indicators	For 2011 SAC particular interest in 15–19 and 20–24 age groups

Classification of surveys

For the discussion of sampling issues in this paper, it is useful to classify SSSs into three groups according to how their survey objectives relate to subpopulations defined by characteristics other than geographic boundaries:

- General population – survey objectives are cross-sectional estimates across a high proportion of the Australian population.
- Low prevalence – survey has multiple key objectives, some (or possibly all) of which relate to specific subpopulations with low prevalence (e.g. less than 10%).
- Rare population – survey objectives focus on a subpopulation with very low prevalence (e.g. less than 3%), components of which have significant geographic clustering.

Surveys which produce estimates for Australia's Indigenous population are the only current examples of SSSs in the Rare Population class. The Indigenous population comprise around 2.5% of the Australia's population, and the population is highly clustered in some geographic regions while in others it is geographically disperse.

Increasing responsiveness of SSS program

In line with organisation-wide pressure to be more responsive to meet emerging data needs of users, there is increasing flexibility in the requirements and conduct of individual SSSs. The flexibility may involve adopting new methods of data collection, collecting new data items and more broadly satisfying a wider range of objectives. The requirement for a more responsive survey program has meant that:

1. the ABS is finding it difficult to maintain a long-range schedule of static surveys; and
2. it is expected that, in the future, fewer SSSs will share similar characteristics with the MPS.

A survey program which is more fluid also presents numerous operational challenges (e.g. managing an interviewer panel) as well as challenges providing cost-efficient samples across the survey program.

There have been two recent examples of this increased flexibility whereby new objectives focusing on particular subpopulations have been added to existing surveys. The 2009 Household Expenditure Survey (HES) selected a top-up sample of households for which the pension or other government benefit was the principal source of income. The second example is the 2010 General Social Survey (GSS), for which informing about persons with multiple social disadvantage was added as a key survey objective. In both examples, the existing survey can be thought to have moved from the 'General population' to the 'Low prevalence' category.

It appears within the organisation that more demand for this type of subpopulation targeting is emerging, although formal assessment of this demand still needs to be undertaken.

2.2 Enumeration modes

A fundamental difference between MPS and SSSs is the mode of enumeration. Sample dwellings are in the MPS sample for eight consecutive months, with enumeration in the first month by computer-assisted face-to-face interview and enumeration in subsequent months typically by computer-assisted telephone interview. The LFS enumerates all in-scope persons, with any responsible adult in a household able to report on behalf of other household members. In contrast the most effective collection mode for most SSSs is face-to-face personal interview. This is due to the length and complexity of the questionnaires involved, and also sometimes the sensitivity of the survey topic.

The differences in enumeration procedures mean MPS and SSSs have quite different enumeration costs, with the enumeration cost per sampled person substantially higher for SSSs. The longer interview time for SSSs not only implies greater effort required for interviewing, but will also typically result in interviewers needing to make several visits to each sampled area. The differences in the cost structure for MPS and SSSs is one reason why the most cost-efficient sampling strategy for MPS will not necessarily be cost efficient for SSSs.

2.3 Integrated survey designs

Practically all SSSs have been run as fundamentally stand-alone surveys. Aside from the collection of core demographic information, there has been limited sharing of questionnaire modules between surveys, and there has not been a common data collection phase shared between surveys. There are a range of possible alternative models the ABS could adopt for collecting survey data. One alternative which provides a degree of integration is to create an 'omnibus' survey vehicle. The vehicle would provide a continuous regular sample of dwellings each month and surveys topics are assigned to this sample in a coordinated fashion. Adopting such a vehicle presents a range of issues for sample design as well as survey coordination.

At some point in the future the ABS survey program could be structured around one or more integrated survey vehicles. The transition would take some time because it would require development of new infrastructure and new processes for planning and survey management. Given the lengthy transition time, this paper assumes that over the next few years the SSS survey program will continue to be dominated by stand-alone surveys. An initiative which could possibly be implemented in the short-to-medium term is the establishment of an omnibus-type survey vehicle separate from the MPS and large-scale SSSs. Such a survey vehicle could be designed specifically to cater for collecting data on multiple independent short survey topics.

3. MOTIVATIONS FOR REVIEWING THE SAMPLING FRAMEWORK

3.1 Sampling framework

The focus of this paper is the future sampling strategy for SSSs from a program-wide perspective, and attention is restricted to sampling strategies involving multi-stage sampling from an area-based frame. A formal specification of the sampling strategy for a survey is the *sample design*, which specifies the selection probability of each possible sample. This paper will summarise sample designs from the area-based frame in terms of the following key parameters:

- definition of sampling units at each sampling stage;
- selection probabilities of sampling units at each stage;
- selection algorithm at each stage;
- number of dwellings within the finest area selected; and
- sample rotation scheme (for repeating surveys).

The program-wide sampling strategy is described by a sampling framework, which defines the suite of possible sample designs available across a program. Implementing coordination between samples selected across the survey program provides cost efficiencies from a program-wide perspective and simplifies management of operations and respondent burden. The price of coordination is restriction on the range of possible sample designs for individual surveys, which could compromise the ability to produce efficient designs for individual surveys. The overall challenge of developing a sampling framework is achieving the optimal balance between offering flexibility in the range of possible designs and controlling program-wide costs.

3.2 Parallel block design

Over the past 30 years the conceptual sampling framework supporting the MPS and SSSs has remained relatively unchanged. The characteristic feature of the framework is the selection of a master sample of Census Collector Districts (CDs), within which separate finer areas provide parallel samples for MPS and SSSs. A new population master sample of CDs has been selected every five years following the Census.

To ensure cost-efficient operations across Australia, sampling procedures are different in the densely and sparsely settled areas of Australia. This categorisation of areas has been derived from an 'area type' class assigned to CDs based on characteristics such as dwelling density, population growth and geographic remoteness. The area type class is used in the stratification of the area frame, and the strata are grouped into two classes herein referred to as 'dense' and 'sparse' strata.

In the dense strata, the master sample of CDs is selected in a single sampling stage. The sampling frame is a list of CDs divided into strata based on their geographic location and area type. A systematic sample of CDs is selected within each stratum, where the CD list within strata is ordered in serpentine fashion in an attempt to achieve even geographic spread. For the sparse strata the CDs are first grouped into larger areas, and a sample of these larger regions is selected at the first stage. A sample of CDs is selected within these First-Stage Units (FSUs), thereby controlling the geographic spread of CDs so that viable interviewer workloads can be formed.

Two subsequent stages of selection are conducted to provide the sample for individual surveys. Each selected CD is divided into a number of smaller areas called blocks, where block boundaries are marked by geographic features (e.g. roads, rivers). One block is then randomly selected for the MPS, and a second 'parallel' block is subsequently assigned to provide sample for the SSSs. The dwellings within each block are listed and systematic sampling is used to divide the dwelling list into groups of dwellings referred to as 'clusters'. The monthly MPS sample includes one cluster from each MPS block, while a standard SSS selects at most one cluster from a subsample of the parallel blocks.

The master sample of CDs is selected with probability proportional to the number of clusters in the CD, and similarly blocks are selected with probability proportional to the number of clusters in the block. The number of clusters assigned to CDs and blocks is determined by Census dwelling counts and the desired number of dwellings per cluster (the cluster size).

The master sample properties such as stratum sample size and cluster size are driven by LFS requirements. A typical sample redesign process will reassess these requirements including the relativity of sample error for key LFS estimates at state and national level, as well as enumeration costs within different strata. Optimisation is performed to determine the optimal cluster size within each strata, which reflects the balance between cost and variance tradeoffs for each stratum. Sample allocation is performed according to minimising cost whilst meeting sample error requirements and ensuring that there is Equal Probability of Selection (EPS) within state, thereby controlling somewhat for estimates at sub-state levels.

This approach works well for LFS because it is tailored around requirements and process of collecting data for LFS, but will be suboptimal for SSSs because of vastly differing enumeration cost structures and mode of enumeration. In particular, this process does not ensure that certain subpopulations of interest to SSSs are oversampled.

3.3 Motivation behind modifying the sampling framework

There are several drivers that are motivating changes to the sampling framework that are described below, including: changes to the ABS geographic standard, the evolving budget balance between MPS and SSSs. Other drivers mentioned through this paper include: need for more sophisticated SSS sample design and the need for finding bookable efficiency gains.

There are also several enablers that are motivating changes to the sampling framework, namely: technological improvements in sample preparation procedures, the advent of cube sampling and redevelopment of internal sample management systems. These are described below.

These motivators impact on the merit of the parallel block framework relative to alternatives. Section 3.4 analyses the properties of the parallel block framework and how the changes relate to these framework properties.

New geographic standard

The ABS will progressively replace its current geographic framework with the new *Australian Statistical Geography Standard* (ASGS) from July 2011 (ABS website). The ASGS has been developed so that all regions used by the ABS to output data can be constructed by aggregating mesh blocks, the finest geographic region in the ASGS. The definition of the mesh block boundaries have been influenced by a range of factors, including the need for mesh block aggregations to align with the desired geographic areas for output needs (Australian Bureau of Statistics, 2005).

One branch of geographic regions used to output data are the ABS structures, which are regions defined and maintained by the ABS. The ABS structures are organised hierarchically by levels created for the release of particular ABS statistics. The finest level of this hierarchy is the mesh block. There are approximately 347,000 mesh blocks covering the whole of Australia, with the majority of residential mesh blocks containing between 30 and 60 households.

From the perspective of producing an area-based sample design, the most important change in the geography standard is that mesh blocks are significantly smaller than CDs (the smallest building block in the current geography standard).

For SSSs, not only are mesh blocks a convenient size to be an FSU, the availability of Census characteristics for finer geographic areas provides opportunities for improving the efficiency of SSS sample designs. In particular, using a finer FSU from the area frame will enable more efficient targeting of areas with higher prevalence of subpopulations of interest.

For MPS, the ability to build up FSUs to the minimum required size will reduce sample wastage and increase efficiency of first stage selections will also be an advantage. Possible candidates for this are Statistical Area 1 (SA1), Census Collector Workloads (CCWs) or a custom FSU built from 3–4 existing mesh blocks tailored around geographic proximity and labour force characteristics.

Balance of budget for MPS versus SSSs

As the size of ABS household survey program has grown over time, so too has the relative size of the SSS component of the ABS household survey program budget. As a result there is increasing justification to limit compromises to the sample design efficiency of SSSs.

Technological advances affecting sample preparation procedures

Technological advances are providing opportunities for cheaper processes for creating dwelling lists and updating them prior to selection. The advances include the availability of a more complete Geo-coded National Address File (GNAF) and satellite-imagery software. The accuracy and currency of these technologies may be sufficient in some areas of Australia for them to be used to create the dwelling lists within mesh blocks. If these sources were updated frequently, they could even be used for updating dwelling lists to identify new and demolished dwellings.

Adopting the Cube Method for MPS selection

The Cube Method (Deville and Tillé, 2004) is a selection algorithm which can select balanced or approximately balanced samples. A balanced sample has the property that for a set of design variables, estimates weighted by the selection probability weights will reproduce known population totals. An initial study into the potential gains in sampling efficiency of the Cube Method for ABS household surveys recommended further investigation (Chipperfield, 2007). This study found that by balancing the sample of CDs on Census indicators of labour force status, for estimation of employment there are moderate gains in sampling efficiency relative to the current systematic sampling procedure. The study also indicated that from a program-wide perspective there would only be marginal benefit from using the Cube Method to select a master sample of areas providing sample for many SSSs.

Adoption of the Cube Method has been recommended for consideration to the labour subject-matter area as a relatively low-risk strategy to improve sampling efficiency of the MPS. There are several details on method application and implementation which still need to be addressed before a final decision is made on its use.

New sample management systems

The systems used by the sampling operations area for maintaining data on dwelling selections and providing selections are being redeveloped. Besides reflecting the new geography standard, the new systems will provide greater flexibility than the current systems for the method of selecting a sample of dwellings from the set of areas in a master sample.

3.4 Properties of the parallel block framework

The parallel block framework has been a cost-efficient sampling framework under the existing environment and available statistical infrastructure and systems. The following discussion describes why this framework has been efficient, as well as how changes described in the previous section impact on the relative merits of the parallel block framework.

Sample preparation costs

The initial steps of sample preparation has involved creating block boundaries within each selected CD and producing lists of dwellings in each selected block (typically referred to as 'blocklisting'). Until recently, blocklisting has required the expensive exercise of visiting each selected CD. The parallel block framework has minimised the cost of defining block boundaries within CDs by not selecting CDs exclusively for the SSS sample.

The introduction of the new geography standard should reduce the costs of sample preparation procedures. If mesh blocks were used as the first-stage sampling unit the second stage of selection would be eradicated and no effort should be required to define the block boundaries. Even if the FSU is an aggregation of mesh blocks, mesh block boundaries can define the regions within which dwelling lists are produced. Listing and updating dwelling lists will also become cheaper since new technology will enable these procedures to be conducted in the office. The cost reduction of these sample preparation tasks reduces the incentive in densely population areas for tight geographic coupling of the areas selected to provide MPS and SSS sample.

Interviewer workload formation

As SSSs are enumerated by face-to-face personal interviews, interviewers must travel to each selected dwelling, and often several visits are required for call backs. Particularly in areas with low dwelling density where high travel can be required, sample designs for SSSs should control travel costs by enabling efficient formation of interviewer workloads. A workload is a collection of dwellings a single interviewer is required to enumerate in a fixed period of time. Just as important as the spread of dwellings within a workload for a survey is coordination of workloads between MPS and SSSs.

Interviewers for SSSs will typically also have a monthly workload for MPS, with overall interviewing managed by devoting one fortnight in the month to MPS enumeration and the remaining time to SSS enumeration.

The parallel block framework ensures all dwellings selected for SSSs have close proximity to MPS dwellings. In sparse areas the location of interviewers recruited onto the interviewer panel is driven by the location of areas selected for MPS, so in these areas interviewers will typically be located near the parallel block selections. The additional stage of selection prior to selecting the CDs in sparse strata ensures workloads in these areas are viable.

Respondent load

The ABS seeks to minimise the load it places on individual respondents, since too much load could impact participant cooperation, leading to reduced response rates and data quality. The historical policy guideline has been that a dwelling should not be selected more than once for the MPS or a SSS within a five-year period. Implementing such a policy requires careful coordination of the area-based sample selection across surveys and maintaining sample usage information.

Management of sample overlap is straightforward under the parallel block framework. Assigning separate blocks for MPS and SSSs surveys avoids overlap between MPS and SSSs, while defining clusters by systematic sampling within blocks neatly controls management of sample usage and appropriately deals with growth in the selected areas.

Flexibility for sample design

The sample design parameters underlying the MPS and SSS parallel samples are chosen to optimise the sampling efficiency for meeting the key objectives of the LFS, as described earlier. Three key sample design parameters and their properties are:

- stratum cluster sizes – the cluster sizes provide optimal sampling efficiency according to models for the cost of MPS enumeration and the variance of labour force status estimates under the multi-stage survey design.
- dwelling selection probabilities – for the MPS situation of selecting one cluster per selected block, dwellings within State and Territory have equal probability of selection.
- sample allocation across States – based on desired relativities between the RSEs of State and Territory estimates for unemployment and employment.

Each of these parameters can be varied to some extent when producing a sample design for an individual SSS using the master sample of parallel blocks (the flexibility provided by the parallel sample is discussed in detail in Section 5.3). However, the limited number of blocks in fixed locations available for selection can be problematic for sample designs for surveys which have objectives focused on specific subpopulations. These limitations were apparent for the 2009 HES and 2010 GSS surveys mentioned earlier, as both adopted sample designs which selected sample outside of the parallel block.

4. DIRECTIONS OF OTHER NATIONAL STATISTICAL INSTITUTIONS

4.1 Introduction

Many other National Statistical Institutes (NSIs) conduct household surveys collecting information on a similar range of topics to the ABS. They face similar challenges to the ABS for extending the value of their survey program to meet increasing data needs in an environment in which resources are becoming more scarce. In responding to the need to make their survey program more efficient, a common move across several NSIs has been introducing greater integration of surveys. For example:

- The Brazilian Institute of Geography and Statistics (IBGE) is currently undertaking a project, named Integrated System of Household Surveys (ISHS), which will result in greater coordination between the agency's household surveys. Full implementation of the ISHS will begin in 2011 (Hypolito and Quintslr, 2009).
- The Office of National Statistics (ONS) in the UK have in recent years been undertaking an integration process to merge their major continuous household surveys into an Integrated Household Survey (IHS) (Smith, 2009).
- Statistics New Zealand (SNZ) are looking to consolidate their household sampling strategy by developing three survey vehicles built around distinct themes (Minshall and Bycroft, 2009).

Multi-stage sampling from an area-based frame remains the predominant household sampling approach used by NSIs. The pressure to reduce collection costs is seeing increasing exploration into cheaper alternatives to area-based sampling and enumeration by personal interview. Although these alternatives are providing promise, most agencies are keeping with area-based sampling for the time being. A program-wide sampling strategy which efficiently uses an interviewer panel and save sample preparation costs continues to be a key goal across NSIs.

The proposed sampling frameworks underlying the future household survey programs of IBGE and SNZ are summarised in Sections 4.2 and 4.3. Section 4.4 describes the existing sampling strategy used by Statistics Canada, and Section 4.5 discusses how sampling approaches used by other NSIs are relevant to the ABS and reasons why they may not be applicable in the Australian context.

4.2 Brazilian Institute of Geography and Statistics (IBGE)

Survey coordination principles

The survey program will include two ongoing surveys serving multiple purposes. One of these surveys has labour and income as its core topics, while income and consumption are the core topics of the other. Surveys can be run as supplementary modules to these continuous surveys, or in the case of specialised topics be run in the ISHS framework independent of the two ongoing surveys.

Sample frame

The ISHS will select a master sample of areas (Census sectors) to provide the sampling infrastructure for practically all surveys. Sampling outside of the master sample could be used to meet a specific data need. The sampling method to select outside of the master sample is still being considered.

Sampling strategy

The master sample of Census sectors will be selected within strata using Probability Proportional to Size (PPS) sampling, with number of households used as the size measure. The stratification will be quite extensive, incorporating political and administrative divisions, an urban/rural classification of the geographic area and an income classification of the areas.

Individual surveys will use two stages of sampling to obtain a sample of households from the master sample: a sample of sectors is selected from the master sample, followed by selection of a sample of households within the selected sectors. In the case of the two ongoing surveys, the labour-focused survey will use all sectors in the master sample while the other will select a subsample of sectors. For the initial set of surveys selected from the ISHS, the subsamples of Census sectors will not be tailored for the specific survey topics.

The sampling method to manage coordination of households within selected areas has not been decided. One possibility being studied is Sequential Poisson Sampling (Ohlsson, 1998), a Permanent Random Number (PRN) method.

4.3 Statistics New Zealand

Survey program coordination

SNZ are proposing to integrate household survey content into three survey vehicles, thus moving away from 'stand-alone' surveys which require new samples from the area frame. Each survey vehicle would be organised around a theme and based around three existing surveys: Household Labour Force Survey (HLFS), Household Economic Survey (HES) and the General Social Survey (GSS).

Each vehicle would have capacity for selected dwellings to receive core demographic questions, the main interview, rotating topic modules and supplementary topics. While the core demographic and main interview questions remain constant over time, the rotating topic modules will cover content repeated at regular intervals and supplementary topics provide capacity for ad hoc topics. This framework is an extension of the current situation in which the HLFS is a vehicle with capacity for supplementary topics to be added. Surveys outside of the three vehicles would only be conducted in cases when particular subpopulations are to be targeted and an alternative frame provides for efficient targeting.

Sample frame

The samples for the three vehicles would be selected from an area-based frame in which the FSU are fine geographic regions. These FSUs on average contain around 60–70 dwellings and built up from mesh blocks, the basic geographic building blocks in New Zealand's statistical system.

Sampling strategy

The sample for the three vehicles would be provided by different FSUs. The cost benefit of ensuring close geographic proximity of the FSUs of the three vehicles is currently being investigated. It is likely the FSUs for each vehicle would be selected using a consistent stratification similar to the current stratification. Currently, at the broadest level the frame is stratified by geographic region and an urban/rural indicator, and the next levels are breakdowns by high/low Maori, followed by high/low Pacific, followed by high/low Asian. The final level of stratification is based on socio-economic variables, however this finest level has not been used for sample allocation. The high Maori and high Pacific strata are examples of strata which have been oversampled in the past.

If current practice is continued, the sample of FSUs for each vehicle would be selected by simple random sampling with a PRN scheme used to avoid overlap of FSUs. Within each selected FSU the dwellings are divided into clusters by systematic sampling.

4.4 Statistics Canada

In recent years Statistics Canada have been exploring options to replace their current household survey sampling strategy because the current strategy can no longer meet client demand for household surveys (Gambino *et al.*, 2007). An early strategy considered was to create a very large master sample of dwellings, say 300,000, for which basic information would be collected (Tambay *et al.*, 2009). Efficient sample designs targeting specific populations could then be developed for individual surveys. The approach was not feasible because of the excessive cost of creating the master sample.

The following discussion describes the existing sampling framework used by Statistics Canada, which is heavily based on the area sample selected for its LFS. It is interesting to analyse the framework from the point of view that Statistics Canada and the ABS have similar survey programs, statistical systems and infrastructure.

Survey coordination principles

The area sample selected for the LFS provides sample for practically all of Statistics Canada's household surveys.

Sampling frames

The first stage of selection of the LFS sample is a sample of geographic areas which contain an average of around 200 occupied dwellings. These FSUs are formed by combining Census blocks. It should be noted Statistics Canada have also used random digit dialling and a sample of telephone numbers from telephone lists to provide sample for a small number of its household surveys.

Sampling strategy

A group of dwelling clusters is created for each selected FSU by selecting systematic samples over the complete FSU. The clusters are the ultimate sampling unit, making the LFS sample design essentially a two-stage design. Other surveys besides the LFS draw sample from the area master sample frame by selecting a defined set of available clusters in the selected FSUs. Unlike Australia, there is not an explicit sample of non-LFS clusters contained within a geographic area created at a sampling stage.

There are three ways the LFS area sample has been used by surveys other than LFS. Two of these methods are also used by the ABS: adding supplementary modules to the end of the LFS questionnaire and using clusters in the FSU which have been reserved for non-LFS surveys. The other method is contacting households one or more months (but within 24 months) after their six months in the LFS. An advantage of this approach is that the data collected from the LFS can be used to screen respondents, though this approach cannot provide large sample for rare populations.

The sample design parameters for the LFS area sample are driven by LFS objectives, which are specified for much finer geographic regions than in Australia. Strata are defined within each region, including special strata for rare populations (high-income, immigrant and Aboriginal). These special strata help control selection of areas with high prevalence of these populations. Dwellings in the same region have equal chance of selection, so oversampling of rare populations is determined by the sample allocation to the regions within which these special strata are created.

4.5 Potential applicability of approaches in the Australian context

There are numerous reasons why the most appropriate sampling framework used by one NSI may not be suitable for another, including:

- size and structure of the survey program;
- detail, currency and quality of data to form the area-based frame;
- size and extent of interviewer panel;
- geographic spread of populations of interest; and
- requirements for interviewer workloads.

The sampling frameworks proposed by IBGE and SNZ are based around survey programs involving multiple omnibus-type vehicles (introduced in Section 2.3). These survey programs are more integrated than the type of survey program this paper is assuming the ABS will use in the short term.

The theme-based multi-vehicle approach in the form proposed by SNZ would have drawbacks for the ABS survey program. Firstly, the multi-vehicle framework is based around a small number of on-going surveys. The LFS is currently the only ABS household survey run on an on-going basis developed for a particular theme. (It should be noted that in recent years the ABS has run an on-going survey called the Multi-purpose Houshold survey (MPHS), a form of omnibus suvey which is not developed around a specific theme.) There may not be demand from other survey areas for obtaining a small number of core data items on an on-going basis. Secondly, the wider range of surveys conducted in Australia compared with New Zealand makes it more difficult to structure surveys around a small set of themes. In order to collect data on specific topics with the same detail as current, the rotating module components of the vehicle would dominate the vehicle. With limited content which remains unchanged, the value of establishing a continuous collection vehicle is diminished. In some cases a vehicle may need to support diverse topics, which has the potential drawback that the sample design underlying the vehicle may not suit all topics.

The current ABS sampling framework is similar to the IBGE approach in that a single master sample of areas is selected to support LFS and other surveys. Reasons why the ABS is exploring alternatives to this approach were discussed in Sections 3.3 and 3.4. Three particular reasons why a single master sample of areas has more appeal in the Brazilian context are:

- the cost of listing and drawing of maps would be relatively more expensive;
- areas are more diverse so interviewer familiarity of selected areas is more important; and
- the finest geographic unit in Brazil contains sufficient dwellings to provide sample for both LFS and other surveys, so sharing them for both is an efficient use of resources.

5. POTENTIAL SAMPLING FRAMEWORKS

5.1 Introduction

This section outlines possible sampling frameworks for the future ABS household survey program. The new sample designs underlying the framework will use data from the Census in August 2011, and implementation will likely occur during 2013.

To limit the range of frameworks considered, two assumptions have been made. Firstly, the survey framework will not be centred on an omnibus-type survey vehicle described in Section 2.3. Secondly, the framework will not significantly compromise MPS efficiency in order to improve the efficiency of the SSS program. The size and on-going nature of the MPS justifies implementing a sample design focused on its specific objectives. The MPS sample design is almost identical under all frameworks discussed here, so sampling for SSSs is the primary issue covered.

Three alternative frameworks are presented, labelled 'Parallel Samples', 'Dual Master Samples' and 'Free Access to Areas'. It is important to note the master samples used by the first two frameworks do not support the complete ABS household survey program. This is because the proposed master samples cannot support sample designs for surveys focused on a rare and clustered population like the Indigenous population. Table 5.1 presents the range of surveys supported by the master samples proposed for 'Parallel Samples' and 'Dual Master Samples' frameworks, using the survey classification of Section 2.1. When a survey cannot be supported by an established master sample, a specialised sample design which selects from the area-based frame would be used.

5.1 Surveys supported by 'Parallel Samples' and 'Dual Master Samples'

	Parallel Samples	*Dual Master Samples*
General population surveys	All	All
Low prevalence surveys	Some	Most
Rare population surveys	None	None

The following discussion of frameworks only concern sampling of private dwellings and dwellings outside of discrete Indigenous communities. The complete MPS sample includes sample from non-private dwelling establishments and discrete Indigenous communities, but responses from them comprise less than 5% of the total sample. The majority of SSSs do not sample from non-private dwellings.

Section 5.2 describes the changes and innovations planned for the MPS design (almost the same across the frameworks), while the alternative frameworks for the selection of SSSs are described in Sections 5.3 to 5.5.

5.2 MPS sample design

It is assumed the sample design for the MPS sample will retain all the key properties of recent MPS sample designs:

1. final stage of selection is a systematic sample of a cluster of dwellings located in a fine geographic region (typically containing around 30–50 dwellings);
2. stratification of area frame based on geographic location and area type classification of the area;
3. equal probability of selection of dwellings within each State or Territory;
4. cluster sizes chosen to maximise cost-efficiency for estimation of employment and unemployment and vary across area type;
5. additional selection stage in sparse areas to ensure viable workloads can be formed;
6. separate sample frame for selecting non-private dwelling establishments;
7. areas containing Indigenous communities assigned to separate strata.

A necessary change impacting on the implementation of the MPS sample design is that the FSUs and 'blocks' will need to be based around geographic units defined in the new geographic standard. A potential change is using the Cube Method to select the sample of MPS FSUs.

Some of the impacts of these changes are described below. Another impact which is dependent on the choice of framework is the size of the FSU, and this is discussed in Section 5.3.

Sample frame

The geographic units on the MPS sample frame will be built up from mesh blocks. These aggregated units may or may not be units in the ASGS ABS structure hierarchy. Mesh blocks would typically be of suitable size to fill the role of the blocks in the current framework.

In the dense strata the frame of FSUs are areas comprised of enough mesh blocks to last the five-year sample design period. Assuming a cluster size of six and an average mesh block size of 25–50 dwellings, the number of clusters per block would be expected to range between four and eight. Since eight clusters are needed to last a five-year period, most FSUs would need to comprise at least two mesh blocks. In line with current practice, in the sparse strata the FSUs on the frame would cover a much larger area and contain more dwellings.

Stratification

The stratification would be based around the cross-classification of two variables:

- the geographic LFS dissemination regions (defined via the ASGS); and
- 'area type' variable.

Stages of selection

The stages of selection in the dense strata would follow the stages used for the current MPS design (Section 3.2). The first stage sample in dense strata and second stage sample in sparse strata create an MPS master sample of areas. The master sample provides clusters of dwellings for the five-year life of the MPS sample design.

Sample rotation

A single cluster within each FSU contributes to the initial sample of clusters comprising the MPS sample. Beyond the initial sample, selection is controlled by the 'rotating panel' design. After a cluster has been in the MPS sample for the required period it is replaced by the next cluster in order. An overall ordering of clusters is established by ordering the clusters within a block and the blocks within an FSU. Provided the current cluster is not the last cluster in the block, the next cluster in order is the next cluster within the current block; otherwise the next cluster is the first cluster in the next block in the FSU.

An alternative to this rotation strategy would be to use all clusters in the initially-selected block prior to rotation into the next block in the list. The merits and disadvantages of this 'Cluster Re-use Rotation' (CRR) strategy have been previously investigated. The study recommended against CRR due to its potential statistical impact resulting from dwellings in larger blocks having higher chance of selection.

5.3 Parallel sample framework

This framework creates parallel master samples of areas for MPS and SSSs which are contained within small geographic areas. The concepts and properties of this framework were discussed in Section 3.2. The introduction of the ASGS and the accompanying improvements to sample management systems build on the sophistication of sample designs which could be produced for individual SSSs.

Size of First-Stage Units

A key property of the framework is the FSUs would need to comprise at least four mesh blocks in most cases, since the FSUs need to contain sufficient clusters for both MPS and SSSs. It is desirable for the FSUs to be as small as possible. This is because there would be less area 'used up' by the master sample, and if it were adopted the

Cube Method of selection would provide greater benefit. However designing FSUs which are small increases the risk of exhausting the dwellings in the area and needing to use another FSU to replace it. Rotation into a new FSU should have smaller cost implications than in the past (due to cheaper sample preparation processes), so the statistical risks are of greater concern.

SSS sample design options

Although the sample design parameters underlying the SSS master sample would be chosen to meet MPS objectives, the framework provides considerable flexibility for SSS sample designs. Over the past five years this flexibility has been exploited to produce increasingly sophisticated sample designs from the parallel sample. A major reason for this flexibility is the smaller sample sizes of SSSs compared with MPS.

For the MPS design (one cluster selected in each MPS block), the selection probability of a dwelling is

$$P\left(\text{Dwelling selected}\right) = \frac{C}{K_s}\frac{c_b}{C}\frac{1}{c_b} = \frac{1}{K_s}$$

where

C is the number of clusters in the FSU;
c_b is the number of clusters in the block; and
K_s is the inverse sampling fraction of clusters selected in State s.

The two basic parameters for producing a sample design from the parallel block master sample are:

- the proportion of master sample blocks selected, denoted f; and
- the number or fraction of clusters selected per block, denoted g.

These parameters could be controlled for the strata defined within States. Incorporating these parameters, a dwelling's selection probability becomes

$$P\left(\text{Dwelling selected}\right) = f\frac{C}{K_s}\frac{c_b}{C}\frac{g}{c_b} = \frac{fg}{K_s}$$

Together these parameters control the dwelling selection probabilities (and hence total dwelling sample size), while the g parameter controls the level of clustering. Historically most SSSs have adopted the MPS cluster size and sampled no more than half the blocks (i.e. $g = 1, f \leq \frac{1}{2}$). A notable exception has been income surveys, for which visiting all blocks and halving the cluster size has been found to be a more efficient approach (i.e. $g \approx \frac{1}{2}, f = 1$). Most SSS designs have used the same cluster size adjustment g across all strata, but it has been common for surveys to vary f in order to achieve a State-level sample allocation different to the MPS.

Historically the cluster sizes used with the parallel sample have been constrained to basic fractions *g* of cluster adopted by the MPS. If the redeveloped sample management systems allowed more dynamic updating of dwelling lists in systems, it would be possible to use any cluster fraction. However, selecting fractions of the established sample clusters has the drawback of creating partially-used clusters. Existing sample management processes determine that non-sampled dwellings in partially-used clusters cannot be used for future surveys, meaning that they represent sample 'wastage' and increase the likelihood of block rotation.

Another design strategy besides adjusting *f* and *g* is selecting an unequal probability sample of blocks within strata (e.g. use selection probabilities proportional to some size measure particular to each block). The availability of Census data at mesh block level makes this a more attractive option than in the past, when the size measure has needed to have been based on the properties of the parallel block's CD. An example of a design from the parallel block which used unequal probability sampling is the 2005 Personal Safety Survey (PSS). In this case, the motivation for the unequal probability design was to mitigate the risk of extreme travel costs in ex-metropolitan areas due to a reduced interviewer panel. The parallel blocks in these areas were arranged into workloads and workloads assigned to available interviewers prior to sample selection. A sample of workloads was then selected with each workload's selection probability a function of the distance from its assigned interviewer.

5.4 Dual master sample framework

The foundation for this sampling framework are separate master samples of geographic areas for MPS and for SSSs. The motivation for selecting a master sample tailored for SSSs is that the SSS master sample can be designed to better cater for the needs of the SSS survey program. There would be reduced likelihood of needing to select a sample of areas outside of the established master sample, thereby reducing survey-wide program costs. In the simplest application of this framework there would be a single master sample catering for SSSs (assumed below), but the framework could be extended to include separate master samples catering for different groups of similar SSSs.

The key element to this sampling framework is that a SSS master sample would consist of a much larger number of FSUs than provided by the parallel sample. Individual SSSs can select a subsample of these blocks depending on their specific needs and there would be oversampling of blocks with high prevalence of subpopulations of interest to SSSs.

Selection of master sample: Stratification

The FSUs for the SSS master sample would be selected from an area frame stratified by one or more variables correlated with common populations of interest across the SSS survey program. The strata containing areas with high prevalence of common populations of interest would be oversampled, thereby enabling the master sample to better support sample designs requiring oversampling to meet their objectives.

Sampling units for SSS master sample

The first stage of selection and link between the MPS and SSS master samples would differ between the dense and sparse strata.

- In dense strata, the FSUs would typically be individual mesh blocks (though mesh blocks with few dwellings would need to be amalgamated). The sample of these FSUs defines the SSS master sample of areas in the dense strata.
- In sparse strata, it is expected that all selected areas on the SSS master sample frame would be contained within the (larger) sparse FSUs selected for MPS. A separate selection exercise of SSS FSUs in sparse strata would not be needed. A sample of finer areas would be obtained within each selected FSU, and this sample of finer areas would constitute the SSS master sample of areas in sparse strata.

The size of the SSS FSUs in dense strata would be much smaller than the size of CDs, which will reduce the amount of area to be avoided by future samples. The SSS FSUs must not cut across boundaries defining the FSUs on the MPS frame (otherwise it will be impossible to compute the probability these areas are selected in the MPS master sample).

Similar to the MPS design, the number of clusters in each FSU would be assigned prior to selection. The number of clusters is based on the number of dwellings and a cluster size considered optimal for SSSs across the SSS program.

Stages of selection

In dense strata, the selection process for a SSS could be as follows:

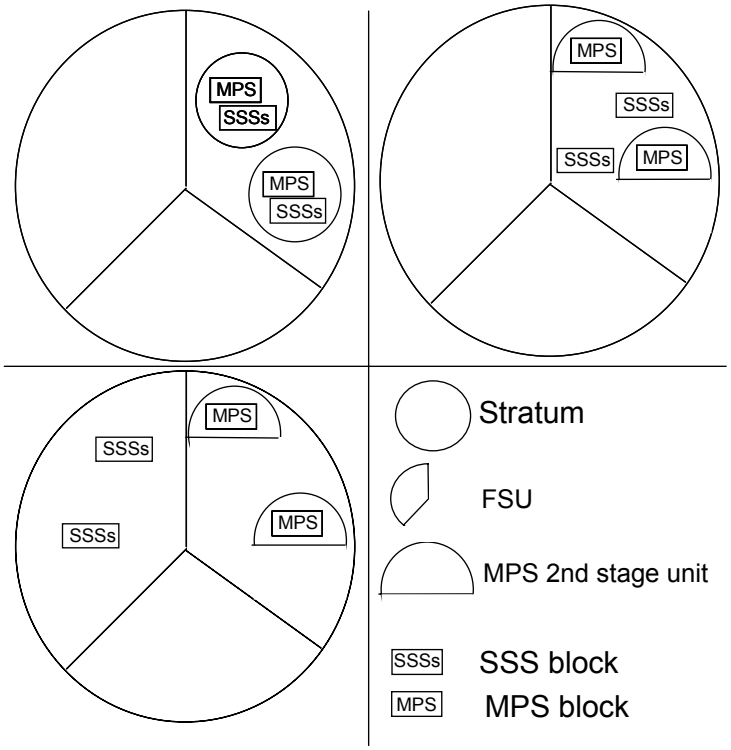
- First stage: The sample of FSUs selected for the SSS master sample would be selected using PPS sampling within strata. The size measure for the PPS selection would be the number of clusters in the FSU, thereby ensuring clusters within strata have equal probability of selection. It would be desirable for the selection algorithm to control the sample size at this stage.

- Second stage: A simple random sample of FSUs from the master sample would be selected within each stratum. Sampling fractions would differ across strata to provide a sample of FSUs from the master sample to meet the specific objectives of the survey involved. The sampling algorithm is dependent on the design adopted. Potential sample design options are discussed further below.
- Third stage: One cluster (or multiples or fractions thereof) of dwellings is randomly selected from the survey's selected FSUs.

In sparse strata, it is expected the first stage of selection is common to the MPS and SSS master samples. The subsequent stages follow the three stages described above for the dense strata, except that the selection of units described under 'First stage' are constrained to lie within the selected FSUs.

Further analysis is required to determine the most cost-effective strategy for the relationship between the MPS and SSS master samples in the sparse strata. Figure 5.2 provides a schematic representation of some alternatives. Considering that in sparse areas the location of interviewers recruited onto the interviewer panel is driven by the location of areas selected for the MPS, Option 3 (bottom left) would be expected to be the most expensive option as it involves travel to different FSUs for MPS and SSSs. Option 1 (top left) would be expected to be the cheapest, although Option 2 (top right) may be just as economical since the location of blocks within an FSU may have little impact on average total travel.

5.2 Alternative representations of the relationship between MPS and SSS master samples



Managing sample overlap

To avoid sample overlap, the SSS master sample would be selected conditional on the areas selected for the MPS master sample and the master sample currently in use. The areas to avoid will be given zero chance of selection, and the remaining areas will be selected with a conditional probability which preserves the desired unconditional probability. The method is discussed in more detail in Section 6.5.

Design options using SSS master sample

Similar to the Parallel Samples framework, the basic alternative sample designs for individual SSSs are parameterised in terms of the proportion of master sample blocks and the number or fraction of clusters in the selected blocks. The major difference for this framework is accounting for the disproportionate sampling across strata in the SSS master sample. Following the notation of Section 5.3 and denoting K_h^{-1} as the sampling fraction of clusters for selection of the master sample in stratum h, the probability of selection of a dwelling is

$$P\left(\text{Dwelling selected}\right) = f_h \frac{C}{K_h} \frac{g_h}{C} = \frac{f_h g_h}{K_h}$$

The proportion of blocks selected in stratum h, f_h, would be chosen to achieve the desired selection probabilities within strata for the particular survey. Depending on the subpopulations of interest in the survey objectives, different surveys could have very different f_h and within a survey the f_h could vary significantly across strata. For example, a survey requiring an equal-probability design would choose the f_h to 'undo' the disproportionate sampling resulting from the differential K_h^{-1}.

As for the Parallel Samples framework, more dynamic updating of dwelling lists in the redeveloped systems would allow individual surveys to adopt any cluster fractions g_h. Cluster fractions close to but not equal to 1 would be avoided since the marginal efficiency gain from not using exactly a whole cluster is likely to be outweighed by the costs arising from partially-used clusters.

More complex design alternatives which go beyond modifying the proportion of blocks and cluster sizes (via f_h and g_h) would also be possible. The master sample blocks could be stratified further, or within strata blocks could be selected with unequal probabilities proportional to some size measure. The feasibility of such approaches will depend on the fineness of the master sample strata.

5.5 Free Access to Areas framework

This sampling framework creates a master sample of geographic areas for the MPS, but no master sample is established for SSSs. Individual SSSs would select a sample of mesh blocks from the mesh block frame according to a sample design tailored to provide a sample which will best meet their objectives. The following description assumes maximum flexibility, but in practice some constraints could be imposed to sacrifice some flexibility in order to significantly simplify implementation (particularly for sample management).

Frame and stratification

Surveys would select from a common area-based frame. Selection units would typically be mesh blocks, but mesh blocks with a small number of dwellings would be amalgamated. Stratification of the frame would be the key sample design tool, with each SSS able to adopt a different stratification of the frame. Flexibility in the stratification has the advantage that emerging changes in enumeration cost structures (e.g. advent of internet collection) could be dynamically incorporated into the stratification of sample designs midway through the five-year sample design period.

Stages of sampling

In dense strata sampling for individual SSSs would be conducted in two stages, since the FSUs would be sufficiently small that a list of dwellings can be created for them directly. In areas with low dwelling density, the first stage may select regions much larger than mesh blocks so that more cost-efficient workloads are formed. It would be possible for surveys to use different geographic boundaries to determine where an additional selection stage is required.

Since SSS sample usage would be spread thinly across a larger number of FSUs, compared with the master sample frameworks there is less motivation to establish a fixed number of clusters per FSU. An alternative within-FSU sampling strategy is for each survey to select a simple random sample of a specific number of dwellings rather than fractions of pre-defined clusters. A drawback of this approach is greater variability in dwelling selection probabilities due to the effects of growth within the FSU between the Census and time of selecting a survey sample. A PRN method could be used to manage sample usage and overlap of dwellings within an FSU.

Sample design options

Developing a sample design for each SSS would first involve determining a stratification and sample allocation of the FSUs on the area frame, as well as a suitable cluster size for each stratum. The sample selected for previous surveys would not restrict design parameters for individual SSS designs.

Managing sample overlap

If each survey selects its own sample of areas, avoiding overlap between selected areas on the frame could become impossible in the smaller States and Territories. Exhaustion of areas is most likely in the Northern Territory, where the sampling fraction of dwellings for the current MPS design is 1/54. Assuming SSSs have a sampling fraction of approximately half the MPS fraction, and the SSS FSUs contain an average of six clusters, there would be sufficient distinct FSUs for 18 SSSs. In practice the number of SSSs which could select a distinct sample of FSUs would be much smaller than 18: overlap with MPS FSUs must be avoided, some FSUs would contain more than six clusters and higher sampling fractions would be used in some areas.

The selection method would need to allow selection of common areas and carefully manage sample usage within areas to ensure no overlap between the dwellings selected for different surveys. Research is required to evaluate possible selection methods which can preserve the desired selection probabilities when selecting partially-used areas and avoiding overlap with previously selected dwellings. The method would need to track and account for the probability of selection of each FSU on the frame for each survey conducted within the time period for which overlap must be avoided. A particular concern is how the widespread usage of areas could compromise selection the next MPS master sample five years later.

6. COMPARISON OF FRAMEWORKS

6.1 Introduction

This section compares the alternative frameworks with respect to key properties which affect the framework choice. The properties to consider cover not only the cost efficiency from a program-wide perspective, but also the exposure to operational and statistical risks. Considering all the properties together, an overall assessment is made on the most suitable sampling framework for the future.

6.2 Framework agility to support varied objectives

Depending on the survey, restriction to select a survey sample from a master sample may or may not inhibit the ability to produce an efficient sample design. For surveys producing general cross-sectional population estimates, an optimal 'stand-alone' sample design would not be much more cost-efficient than a sample design based on a master samples like those described in Section 5. In contrast, for surveys with key objectives focusing on particular subpopulations, restriction to a master sample of areas can severely compromise the ability to produce an efficient sample design. In the face of such restrictions, the required sample could either be obtained from the master sample in an inefficient manner (e.g. by selecting more blocks with more extensive use of screening), or sample could be selected outside of the established master sample. Both of these alternatives raise the overall cost.

Flexibility provided by master samples

For the Free Access to Areas framework, restrictions on the areas sampled only arise from the efforts to manage overlap. Comparing the two master sample frameworks, a master sample developed specifically for SSSs offers greater flexibility than a master sample parallel to the MPS sample because:

- the number of clusters per stratum in the master sample can be made much larger than the number of clusters required for MPS; and
- a higher proportion of blocks can be selected in the areas which the survey program is expected to require oversampling.

The limit on the number of blocks in the parallel sample is not a hard limit, as exhausted blocks would be replaced by an adjacent block. However this strategy compromises the accuracy of the selection probability of this new selection. The number of blocks in the parallel master sample could be increased at cost to efficiency of the MPS. More blocks could be made available by designing larger FSUs so that the parallel master sample provides several SSS blocks per FSU. If the Cube Method were

adopted, a drawback of larger FSUs is that some of the efficiency gain from using balanced sampling for MPS would be eroded.

Although a master sample selected specifically for SSSs obviously provides greater flexibility than a parallel master sample, it is less clear whether the additional flexibility will result in lower overall program-wide costs. The benefit of having a custom SSS master sample ultimately depends on the number of surveys with objectives which align with the design of the SSS master sample but not the parallel master sample. This benefit is explored below in the context of how a SSS master sample and a parallel master sample could support surveys requiring disproportionate sampling.

Disproportionate sampling

There are several possible strategies to disproportionately sample population members. A simple approach is the subsampling of persons within households, but this is only effective for characteristics with high heterogeneity across household members (as is the case for sex). A more generally-applicable strategy is stratifying the area frame into strata based on the subpopulation prevalence in the areas. There are some simple theoretical guidelines on the appropriate extent of disproportionate sampling.

These theoretical guidelines are described using the following notation. If the stratum h counts for subpopulation members and total population are M_h and N_h respectively, the prevalence of the subpopulation in stratum h is $P_h = M_h/N_h$. Denote $A_h = M_h/M$ as the proportion of the subpopulation in stratum h, and $W_h = N_h/N$ as the proportion of all persons in stratum h. Finally let c be the ratio of the data collection cost for a subpopulation member to the cost of screening (or sampling) non-members.

Under simple random sampling the variance of a sample mean for the subpopulation is minimised (for fixed total budget) by selecting subpopulation members with probability proportional to

$$\sqrt{\frac{P_h}{P_h\left(c-1\right)+1}} \tag{6.1}$$

(Kalton, 2009). This also assumes variances are constant across strata. The screening methods used at the ABS are relatively expensive, so the value of c tends towards 1. In this case the optimal unit sampling fraction of subpopulation members in stratum h tends towards being proportional to $\sqrt{P_h}$.

Clark *et al.* (2009) extend Kalton's results to multi-stage sampling designs and surveys for which estimates of the general population and subpopulations are both of interest. The recommended strategy for achieving the person-level sampling fractions of (6.1) is to 'over-target' at the FSU stage of selection by selecting FSUs with probabilities

closer to proportional to P_h. Smaller cluster sizes and screening would then be used in the FSUs with higher sampling fractions. In situations when survey objectives involve both the subpopulation of interest and non-members, the unit selection probabilities for subpopulation members should approximate (6.1) with over-targeting of high-prevalence FSUs.

There may be occasions when it is not possible to use screening to identify a subpopulation of interest. For example the population of persons suffering 'multiple social disadvantage' in the 2010 GSS is very difficult to identify by simple screening. In this situation the screening cost and main data collection cost are equivalent, so the optimal unit sampling fractions for subpopulation members are proportional to $\sqrt{P_h}$.

Kalton concludes disproportionate allocation only yields substantial gains in efficiency if the following conditions hold:

1. the subpopulation has much higher prevalence in the oversampled strata;
2. the oversampled strata account for a high proportion of the subpopulation; and
3. the cost of the interview per subpopulation member relative to identification cost is not high.

The key message is that the potential gain from disproportionate sampling critically depends on the spread of the subpopulation. Table 6.1 provides a simplified summary of when disproportionate sampling is likely to be of value in the ABS context of relatively high screening costs.

6.1 Application of disproportionalt sampling to ABS surveys

	Clustered population	*Evenly spread population*
Low prevalence	Substantial benefit	Minimal benefit
Moderate prevalence	Some benefit if highly clustered	Minimal benefit

Possible application for SSS Master Sample

There is currently high user demand for social statistics on persons with low socio-economic status. Assuming this demand continues, the future survey program can be expected to include surveys requiring oversampling of this subpopulation. The following example shows this population has moderate geographic clustering, so there would be benefit from designing a SSS master sample around this subpopulation.

6.2 Prevalence of multiple social disadvantage in South Australian mesh blocks

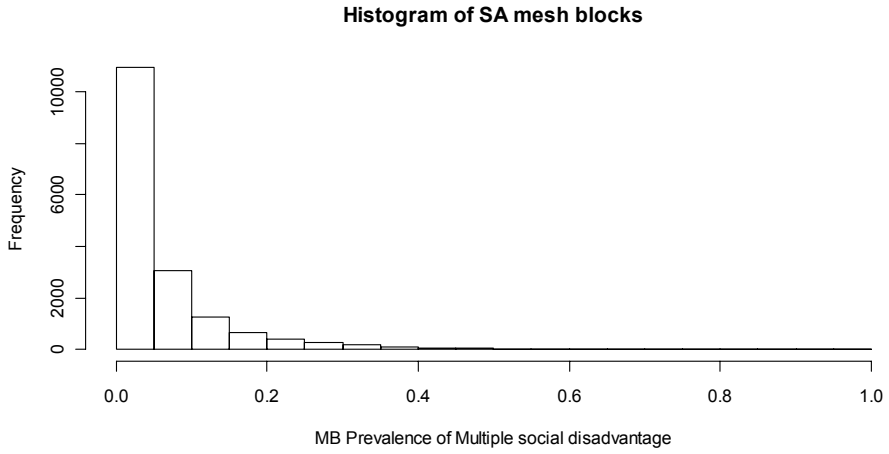


Figure 6.2 is a histogram of mesh blocks in South Australia classified according to prevalence of persons with a derived 'multiple social disadvantage' (MSD) characteristic. This indicator, developed for the 2010 GSS sample design, has been found to be an effective indicator for identifying persons who would report 'disadvantaged' answers for questions on financial, social networks, community involvement and access to service questions. The geographic clustering of low socio-economic status would be similar to the clustering of the MSD indicator.

Table 6.3 shows a possible stratification of the mesh blocks based on their MSD prevalence with relative sampling fractions for sample allocations when sampling fractions are proportional to $\sqrt{P_h}$ and P_h. These relative sampling fractions are defined as the ratio of these allocations to the allocation using equal sampling fractions (proportional to W_h). Under the 'proportional to $\sqrt{P_h}$' allocation, the high prevalence strata have person sampling fractions which are between two and three times the overall sampling fraction. Following the recommendations of Clark, the optimal FSU sampling fractions would be closer to the 'proportional to P_h' allocation.

6.3 Possible stratification of mesh blocks based on MSD prevalence

Range of P_h for stratum	W_h	A_h	P_h	$\sqrt{P_h}$	*% sample for $\sqrt{P_h}$ allocation*	*Relative sampling fraction for*	
						$\sqrt{P_h}$ allocation	*P_h allocation*
[0.00 – 0.05]	0.66	0.19	0.02	0.13	41%	0.6	0.3
(0.05 – 0.10]	0.17	0.21	0.07	0.27	22%	1.3	1.3
(0.10 – 0.15]	0.07	0.15	0.12	0.35	12%	1.7	2.2
(0.15 – 0.25]	0.04	0.11	0.17	0.42	8%	2.1	3.1
> 0.25	0.06	0.34	0.31	0.55	17%	2.7	5.5
	1.00	1.00			100%		

Since the parallel sample is an equal-probability sample, the maximum relative sampling fraction of FSUs from it is the inverse of the overall proportion of parallel sample FSUs selected for the survey. For many SSSs, this maximum relative sampling fraction would be between two and three. Clearly the parallel sample could not support the desired levels oversampling of FSUs for the population considered in this example. Either the parallel sample would be used inefficiently or sample outside of the parallel sample would be selected.

The design of the custom SSS master sample would need to balance the needs of the complete survey program. The master sample requires capacity to cater for the many surveys requiring an approximately equal-probability sample, so the relative allocation to the low-prevalence strata would be larger than given by the 'proportional to $\sqrt{P_b}$' allocation in table 6.3. A possible strategy for developing a SSS master sample allocation would be to start with a basic equal-probability allocation and then increase the allocation to the high-prevalence strata to ensure surveys requiring oversampling can be supported.

Expected benefit of SSS master sample

The potential benefit of designing a master sample specifically for SSSs is well illustrated by the way it can be designed to support disproportionate stratification. However, moving to a SSS master sample framework will only provide significant program-wide savings relative to a parallel master sample framework if:

- there are several surveys in the program with objectives relating to low-prevalence clustered subpopulations and;
- the variables used for disproportionate stratification make a suitable proxy for identifying the subpopulations of interest for multiple surveys in the program.

Both conditions would be satisfied if there is continued demand for data on persons with low socio-economic status (and the master sample is designed for this).

6.3 Costs of sample preparation, maintenance and interview travel

Increasing coordination between the areas sampled across the survey program reduces the costs of sample preparation and maintenance. The average level of interviewer travel per sample unit could also be reduced through increased coordination.

Sample preparation costs

Assuming mesh blocks are used as the blocks within which dwelling lists are created, the cost of sample preparation and maintenance over a five-year period is roughly proportional to the number of mesh blocks used over that period. The master sample frameworks would likely require a similar number of blocks over a five-year period.

Although the custom SSS master sample would likely initially select more blocks than a parallel master sample, the parallel sample framework could more frequently select areas outside of its master sample. The Free Access to Areas framework would use several times more blocks than either master sample approach.

The option to conduct sample preparation activities from the office may mean sample preparation costs become a much smaller proportion of the survey program cost. The operations area needs to be consulted to assess costs under alternative scenarios for the proportion of blocks for which sample preparation activities can be undertaken from the office.

Interviewer travel

The average cost of interviewer travel per sampled dwelling is difficult to compare for the three frameworks. Interviewer travel distance depends on the location of interviewers in relation to the sample and the breakdown of total travel by between-block travel versus travel to visit blocks from the interviewer's home.

In metropolitan areas where more interviewers are available and smaller distances are travelled, it is unlikely the proximity between SSS and MPS selections would significantly influence interviewer travel. Therefore in these areas average interviewer travel cost should be similar under all three frameworks. In non-metropolitan areas interviewer location is more important. Since the location of interviewers is driven by the location of MPS selections, samples selected from a parallel sample would be expected to require the least travel on average. There is risk of very high interviewer travel costs in non-metropolitan areas for a sample selection approach which does not coordinate with the MPS selections. For this reason the SSS master sample approach proposed in Section 5.4 coordinates selections with MPS in sparse areas.

6.4 Operational risks

Operational risks are errors in the implementation of the sample design. The errors could mean the obtained sample of dwellings does not match the intended sample, potentially affecting the ability to meet survey objectives and compromising management of sample overlap. Errors detected early could be corrected, but the effort to make the corrections will raise the cost.

The parallel sample framework involves least operational risk because it requires least implementation change from the current and selects fewer areas from the area frame. Selecting and maintaining a SSS master sample increases the exposure to risks. Interviewers would visit the master sample blocks and surrounding areas less often, and new infrastructure is required to implement conditional selection of the SSS master sample areas. The Free Access to Areas framework has greatest operational risk because of the number of blocks used and requiring maintenance.

6.5 Statistical risks

Managing sample overlap and bias in selection probabilities

Managing sample overlap in household surveys requires a strategy to select non-overlapping samples of dwellings while seeking to preserve the pre-specified selection probabilities of each survey's design. There is extensive literature on methods for controlling overlap for business surveys. A potential complication for preserving the desired selection probabilities in household survey sampling is that there are multiple stages of selection.

The problem is simplified somewhat for the master sample frameworks, since overlap management can be separated into two problems: (1) avoiding overlap between the areas selected at the first stage of sampling and (2) controlling overlap between dwellings selected within the first-stage units. As noted previously, under the Free Access to Areas framework the selection method would need to select areas in which some dwellings have been selected for prior surveys. Research would need to be undertaken to explore the properties of methods when it is impossible to avoid overlap between areas. The remainder of this section considers the problem of avoiding overlap between samples of areas selected from the area frame.

The conditional selection approach introduced in Section 5.4 can be used to avoid overlap between the areas selected for a custom SSS master sample and the areas selected for current and previous MPS master samples. The method assigns selection probabilities to the FSUs which are conditional on the selections of previous samples. The first step for selecting a sample of FSUs is identifying the FSUs which must be avoided. These are given zero chance of selection, and the selection probabilities of the remaining areas available for selection (conditional probabilities) are adjusted to preserve the unconditional selection probabilities. This process has been used for ABS Indigenous surveys to avoid selecting Indigenous communities selected for the MPS sample.

To illustrate, consider the simple case in which we require a sample of FSUs for a custom SSS Master Sample (SSSMS) which does not overlap with the sample of areas selected for the MPS Master Sample (MPSMS). The overall probability of selection for an FSU in SSSMS is

$$P\left(\text{selected in SSSMS}\right) = P\left(\text{selected in MPSMS} \middle| \text{not selected in MPSMS}\right) \times P\left(\text{not selected in MPSMS}\right)$$

(This assumes the probability of non-selection in MPSMS is greater than the probability of selection in SSSMS, which will be true for the small sampling fractions used for FSU selection.)

An analogy can be drawn with two phase sampling, with selection of the sample for MPSMS being the first phase and the second phase sample selected from those areas not selected for SSSMS. The method can be extended to avoiding overlap between multiple FSU samples arising from different selection processes. This requires recording the selection probabilities for each sample so that the probability of not being selected in any prior sample can be computed. The method does not exactly preserve the desired unconditional selection probabilities when the sample designs fix the sample size. However the bias would be negligible when applied to selecting a small number of master samples of FSUs.

Impact on LFS

If the Cube Method were used to select MPS FSUs, adopting the larger FSUs necessary for the parallel sample framework would reduce the MPS efficiency benefit from using the Cube Method.

6.6 Overall assessment

There is a very strong case for the household sampling framework to include a master sample of areas which can provide sample for the majority of SSSs. Adopting a master sample reduces the sample preparation and maintenance costs incurred each time a new set of areas is selected from the area frame. In addition the requirement to select from the area frame for every survey adds complexity in managing sample overlap for the selection process. The Free Access to Areas framework would be the most cost-efficient framework only if, for every survey, the improvement in sampling efficiency realised through not being restricted to a master sample outweighs the costs noted above. This would not be the case for surveys with objectives focusing on the general population; a sample selected from a master sample should be almost as efficient as a 'stand-alone' sample selected from the complete area frame.

It remains to determine the extent to which the master sample for SSSs is decoupled from the MPS master sample, how the master sample is selected and how it would be used. The key motivation for selecting a SSS master sample decoupled from the MPS selections is to enable the master sample to cater for a wider range of survey objectives and hence reduce the likelihood of selecting a sample of areas outside the master sample. If there is continued interest in data for persons with low socio-economic status across subject matter areas, for example, a SSS master sample designed to cater for this demand would reduce the program-wide sampling costs. The master sample would disproportionately sample FSUs in strata which distinguish areas by prevalence of low socio-economic status.

Even if user demand changes and the oversampled areas do not get used, overall costs may not be much more than the costs if a parallel sample framework were used. To mitigate against the lower than expected use of blocks in the oversampled strata, it would be wise not to undertake sample preparation work for all blocks in the oversampled strata upon sample selection. Some sample preparation could be performed as the need arises. The main potential drawback of moving away from the parallel master sample is the increased operational risk arising from implementing a separate large-scale master sample selection. The likelihood and size of the operational risks do not appear to outweigh the benefits of the increased flexibility provided by a separate SSS master sample and increased efficiency of the MPS design through applying the Cube Method to select smaller FSUs.

6.7 Further work

The next two broad steps for developing the household sampling framework are to:

1. Work with operations area to formally assess operational costs in the new environment and confirm the above expectations about costs under the three frameworks.
2. Clarify with survey program managers expectations for future surveys to help guide the sample design for the SSS master sample, in particular the most appropriate variables to use in stratification and disproportionate sampling.

The following stages of development are the specific sample designs of the master samples. The process for developing the sample designs is expected to be similar to the process used previously for master sample designs. One aspect which may be different is devoting less effort for choosing cluster sizes in each area type, given that:

- enumeration cost does not have a simple relationship with the key sample design parameters, and there is not a clear approach for developing an accurate cost model;
- the quality of cost data may not allow accurate modelling of costs; and
- overall cost-efficiency is insensitive to the cluster size values close about the true optimal value

REFERENCES

Australian Bureau of Statistics (2005) *Information Paper: Draft Mesh Blocks, Australia*, cat. no. 1209.0.55.001, ABS, Canberra.

Chipperfield, J. (2007) "An Evaluation of Cube Sampling for ABS Household Surveys", *Methodology Advisory Committee Papers*, cat. no. 1352.0.55.087, Australian Bureau of Statistics, Canberra.

Clark, R.G.; Doherty, M.; Forbes, A. and Templeton, R. (2009) "Sampling for Subpopulations in Household Surveys with Applications to Maori and Pacific Sampling", *Official Statistics Research Series,* 4 (last viewed 7 June 2010): <http://www.stats.govt.nz/sitecore/content/statisphere/Home/official-statistics-research/series>

Deville, J-C. and Tillé, Y. (2004) "Efficient Balanced Sampling: The Cube Method", *Biometrika*, 91(4), pp. 893–912.

Gambino, J.; Tambay, J-L. and Laflamme, G. (2007) "Statistics Canada's New Household Survey Strategy", *Proceedings of the Survey Methods Section, SSC Annual Meeting, June 2007,* Statistics Canada.

Hypólito, E.B. and Quintslr, M.M.M. (2009) *Development on an Integrated System of Household Surveys: The Brazilian Experience*, Paper presented to the 57th Session of the International Statistical Institute, Invited Paper Meetings, Durban.

Kalton, G. (2009) "Methods for Oversampling Rare Subpopulations in Social Surveys", *Survey Methodology*, 35(2), pp. 125–141.

McEwin, M. (2000) *The New ABS Household Survey Program,* Paper presented to the 10th Biennial Conference of the Australian Population Association, Melbourne.

Minshall, G. and Bycroft, C. (2009) *Moving Towards an Integrated Sample Approach for Statistics New Zealand Social Surveys*, Paper presented to the 57th Session of the International Statistical Institute, Contributed Paper Meetings, Durban.

Ohlsson, E. (1998) "Sequential Poisson Sampling", *Journal of Official Statistics*, 14(2), pp. 149–162.

Smith, P. (2009) *Survey Harmonisation in Official Household Surveys in the United Kingdom*, Paper presented to the 57th Session of the International Statistical Institute, Invited Paper Meetings, Durban.

Tambay, J-L.; Laflamme, G. and Gambino, J. (2009) *The Canadian Experience in Creating a Master Sample*, Paper presented to the 57th Session of the International Statistical Institute, Invited Paper Meetings, Durban.

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