1352.0.55.087

Australian Bureau of Statistics

Research Paper

An Evaluation of Cube Sampling for ABS Household Surveys

www.abs.gov.au

New Issue

Research Paper

An Evaluation of Cube Sampling for ABS Household Surveys

James Chipperfield

Analytical Services Branch

Methodology Advisory Committee

8 June 2007, Canberra

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) THURS 26 FEB 2009

ABS Catalogue no. 1352.0.55.087

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 James Chipperfield, Analytical Services Branch on Canberra (02) 6252 7301 or email <analytical.services@abs.gov.au>.

CONTENTS

AN EVALUATION OF CUBE SAMPLING FOR ABS HOUSEHOLD SURVEYS

James Chipperfield
Analytical Services Branch

ABSTRACT

The use of design information for the efficient design of surveys has been studied extensively. Well-known methods include stratification and probability proportional to size. These methods are designed to select efficient samples when there is one survey characteristic of interest. Cube sampling aims to select efficient samples when there are multiple characteristics of interest and where a set of design variables could be used for improving the efficiency of the sample design. Cube sampling achieves this efficiency by selecting balanced samples on a set of design variables. A balanced design has the property that the Horvitz–Thompson estimators of total for the set of design variables equal their known totals. This paper presents some exploratory work into measuring the variance reductions in population estimates from Australian Bureau of Statistics' household surveys as a result of selecting a balanced sample of primary selection units. The results in this paper suggest that cube sampling has the potential to provide significant cost savings and therefore that further work in this area should be continued. This paper mentions other issues (e.g. variance estimation and rotation control) that would need to be considered before implementing cube sampling in the ABS.

1. INTRODUCTION

The use of information in the efficient design of surveys has been studied extensively. Well-known methods include stratification and probability proportional to size sampling (Hansen and Hurwitz, 1943). These methods have the potential to select efficient samples when there is one survey characteristic of interest or when the multiple characteristics of interest are highly correlated. More recently Deville and Tille (2004) developed a method, called cube sampling, with the potential to select efficient samples when there are multiple characteristics of interest that are not necessarily well correlated and where a set of design variables could be used for improving the efficiency of the sample. The cube method selects a balanced sample on a set of design variables. A balanced design has the property that the Horvitz–Thompson estimators of total for the set of design variables equal their known totals.

ABS household surveys have a multistage stratified cluster sample design with selections undertaken in typically three stages: Collection District, block and cluster. This paper presents some exploratory work into measuring the impact on the variance of population estimates from ABS household surveys by balancing the sample of CDs on a set of CD-level design variables, obtained from the Census. This paper mentions other issues (e.g. variance estimation and rotation control) that would need to be considered before implementing cube sampling.

The preliminary results in this paper suggest that cube sampling has the potential to provide significant cost savings and therefore that further work in this area should be continued.

Section 2 introduces cube sampling and the Generalised Regression Estimator (GREG). Section 3 gives some background to ABS household surveys. Section 4 measures the potential gains under cube sampling in an empirical study. Section 5 briefly mentions replicated variance estimation under cube sampling. Section 6 mentions some implications of moving to a two-stage sample design, where meshblock is the first stage of selection. Section 7 mentions areas for further investigation.

2. CUBE SAMPLING

2.1 Preliminaries

Consider a finite population U of size N where for each unit i in the population we know a G vector of design variables $\mathbf{z}_i = (z_{i1}, \ldots, z_{ig}, \ldots, z_{iG})$ so that $\mathbf{Z} = \sum_{i=1}^{N} \mathbf{z}_i$ is known. Consider selecting a sample of size n, denoted by s, from U where unit i has a probability of selection given by π_i and units i and j have joint probability of selection given by π_{ij}. Assume that for $i \in s$, the variables $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iK})$' and $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iP})'$ are collected and that $\mathbf{X} = \sum_{i=1}^{N} \mathbf{x}_i$ is known, where $\mathbf{y}_i$ are the characteristics of interest and $\mathbf{x}_i$ are the auxiliary variables.

The Horvitz–Thompson (HT) estimator of $\mathbf{Y} = \sum_{i=1}^{N} \mathbf{y}_i$ is given by

$$\hat{\mathbf{Y}} = (\hat{\mathbf{Y}}_1, \hat{\mathbf{Y}}_2, \ldots, \hat{\mathbf{Y}}_K) = \sum_{i=1}^{n} \pi_i^{-1} \mathbf{y}_i$$

where $$Var(\hat{\mathbf{Y}}_k) = \sum_{i=1}^{n} \sum_{j=1}^{n} (\pi_{ij} - \pi_i \pi_j) \pi_i^{-1} \pi_j^{-1} y_{ik} y_{jk} \ , \ k = 1, \ldots, K$$

The Generalised Regression (GREG) estimator (see Särndal, Swensson and Wretman, 1992, p. 225) of $\mathbf{Y}$ is generally more efficient than the HT estimator because it exploits the information contained in $\mathbf{X}$ and in $\mathbf{x}_i$ for $i \in s$.

The GREG estimator is given by

$$\hat{\mathbf{Y}}_k^{reg} = \sum_{i \in S} \pi_i^{-1} \dot{y}_{ik} + \hat{\mathbf{X}}' \hat{\mathbf{B}}_{y_k}$$

and

$$\dot{y}_{ik} = y_{ik} - \mathbf{x}_i^{\mathrm{T}} \hat{\mathbf{B}}_{y_k}$$

$$\hat{\mathbf{X}} = \sum_{i \in s} \pi_i^{-1} \mathbf{x}_i$$

$$\hat{\mathbf{B}}_{y_k} = \hat{\mathbf{T}}^{-1} \hat{\mathbf{t}}_k$$

with $\hat{\mathbf{T}}^{-1}$ being the generalised inverse of $\hat{\mathbf{T}}$,

$$\hat{\mathbf{T}} = \left(\sum_{i \in s} \pi_i^{-1} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \right)$$

$$\hat{\mathbf{t}}_k = \left(\sum_{i \in s} \pi_i^{-1} \mathbf{x}_i y_{ik} \right)$$

$\hat{\boldsymbol{B}}_{y_k}$ is an estimate of $$\mathbf{B}_{y_k} = \left(\sum_{i \in U} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \right)^{-1} \left(\sum_{i \in U} \mathbf{x}_i y_{ik} \right)$$

2.2 Cube sampling

Deville and Tille (2004) developed a method, called cube sampling, with the potential to select efficient samples when $\mathbf{y}$ is multi-dimensional. The basic feature of cube sampling is that it attempts to balance the sample on $\mathbf{z}$, given an arbitrary set of selection probabilities, π_i , $i = 1, \ldots, N$. A sample design with an arbitrary set of inclusion probabilities π_i is said to be balanced on $\mathbf{z} = (z_1, z_2, \ldots, z_G)$ if and only if $\hat{\mathbf{Z}} = \Sigma_{i \in s}^{n} \mathbf{z}_i \pi_i^{-1} = \mathbf{Z}$ for all possible samples $s \in \Omega$, where Ω is the population of all possible samples under the design. For example, if $\mathbf{z}_i = (1)$ the sample would be balanced under a Simple Random Sample because $\widehat{N} = \Sigma_{i \in s}^{n} \pi_i^{-1} = N$ for all possible samples.

The details of the cube sampling algorithm are mathematically complex but the essential elements are to:

1. Express the balancing equation by $\Sigma_{i \in U}^{N} \mathbf{z}_i^* s_i = \mathbf{Z}$ where $\mathbf{z}_i^* = \mathbf{z}_i \pi_i^{-1}$ and $s_i = 1$ if unit i is in the sample and is 0 otherwise. For a given set of $\mathbf{z}_i^*$s for $i = 1, \ldots, N$ the balancing equation defines a hyper plane Q in $\Re^N$.
2. Define an N dimensional cube by the vector C with i-th element equal to 0 or 1, so that the vertices of the cube denote feasible samples from U.
3. Choose a vertex of C that intersects with Q – that is a feasible sample set that results in a balanced sample. This is done in such a way so as to preserve the selection probabilities. If such an intersection does not exist then the balancing condition is relaxed so that a vertex of C is chosen that minimises the trace of $Var(\hat{\mathbf{Z}})$, which is the variance over cube sampling.

Fixing strata sample sizes can easily be achieved within the balanced sampling framework. Selecting a fixed sample of n_h CDs in stratum h, where $h = 1, \ldots, H$ and H is the number of strata, can be achieved by balancing on a H vector of variables, denoted by $\mathbf{a}_i$ for unit i, where $\mathbf{a}_i$ has h-th element equal to $\delta_{hi}\pi_i$, where $\delta_{hi} = 1$ if unit i is located in stratum h and equals zero otherwise.

In this paper the cube method was implemented by a SAS macro that was obtained from the French National Statistical Office (INSEE).

2.3 Variance under cube sampling and GREG

A good approximation to $Var\left(\hat{\mathbf{Y}}_k^{reg}\right)$ is given by $Var\left(\hat{\mathbf{Y}}_k\right)$ except that y_{ik} is replaced by $\dot{y}_{ik}$ (see Rao, 1997). Under equal probability designs it can be shown that

$$Var\left(\hat{\mathbf{Y}}_k^{reg}\right) = Var\left(\hat{\mathbf{Y}}_k\right)\left(1 - R_{y_k|\mathbf{x}}\right)$$

where $R_{y_k|\mathbf{x}}$ is the proportion of variation in y_k that is explained by $\mathbf{x}$.

Deville and Tille (2005) suggest under the simple situation of an equal probability design with a small sample fraction that the percentage reduction in the variance of the HT estimator by balancing the sample on $\mathbf{z}$ is reasonably approximated by $R_{y_k|\mathbf{z}}$, where $R_{y_k|\mathbf{z}}$ is the proportion of variation in y_k that is explained by $\mathbf{z}$. This illustrates the importance of balancing on a set of variables that are highly correlated with the variables of interest. In fact, in this simple situation, the percentage reduction in the variance is the same whether the sample is balanced on $\mathbf{z}$ or whether it is used as an auxiliary variable in GREG (i.e. $\mathbf{x} = \mathbf{z}$); a disadvantage of the latter is that GREG estimates may become volatile when $\mathbf{x}$ is of high dimension.

It follows from Section 2.1 that the reduction in $Var\left(\hat{\mathbf{Y}}_k^{reg}\right)$ due to balancing on $\mathbf{z}$ is reasonably approximated by $R_{\dot{y}_k|\mathbf{z}}$, defined as the proportion of variation in $\dot{y}_k$ explained by $\mathbf{z}$. This shows that the estimated variance reduction due to cube sampling depends upon the choice of $\mathbf{z}$ and $\mathbf{x}$.

3. BACKGROUND TO THE ABS SITUATION

Two types of household surveys in the ABS are Specials Social Surveys (SSS) and the Labour Force Survey (LFS). A SSS will generally cover one broad subject matter in detail (e.g. health or income and expenditure), occur about every 3–6 years, have enumeration periods that range between three and twelve months and can have a sample size as high as 12 000 dwellings. The LFS collects information about employment and unemployment each month and has a sample size of about 30 000 dwellings.

The LFS and SSS have a multistage stratified cluster sample design with selections undertaken in a number of stages. The sample is stratified by state, dissemination region (about 70 in Australia), and area type. Examples of area type include inner Sydney and Melbourne, rural and very remote. The current design (i.e. 2002–2007) has over 500 strata, some of which can have as small as two CDs selected per stratum. A cluster may be made up of between five and 15 dwellings depending upon the area type and state.

In Self Representing Areas (SRA), defined as geographic areas with a relatively high level of dwelling density and which cover about 80% of the Australian population, there are three stages of selection. The first stage involves selecting Collection Districts (CDs) with probability proportional to the number of clusters (and dwellings) allocated to the CD using systematic sampling. The second stage involves selecting one block from within each selected CD, where the probability of selection is proportional to the number of clusters allocated to the block. The third stage involves selecting one cluster of dwellings from within each selected block with equal probability.

In non-SRA there is an extra stage of selection, called the Primary Sampling Unit (PSU) stage of selection. The PSU stage of selection occurs before the CD stage and is designed so that selected CDs within a selected PSU are geographically clustered so that they can be enumerated by one interviewer in a cost efficient manner.

A new sample of CDs is selected every five years to coincide with the availability of the Census' CD dwelling counts, used as the basis for the selection of CDs. The sample of new CDs is selected without controlling the overlap with the old sample of CDs. Once a new CD is selected, survey operations staff use specific criteria to determine whether its enumeration will result in over-burdening respondents. Overburden may occur if a CD is selected in two subsequent design periods. To avoid overburdening respondents it is sometimes necessary to rotate from the CD that was originally selected to an adjacent CD. This is called a forced rotation. In the 2001 design period approximately 3% of selected CDs were forced to rotate. Identifying CDs for forced

rotation is a labour intensive process when the old and new CD boundaries do not align, which is the case for about 30% of the selected CDs.

After the forced CD rotations are determined, there are an insignificant number of CD rotations during the five year design period. This means a vast majority of selected CDs remain in sample for five years, during which their clusters are systematically enumerated.

A CD on average contains about 250 dwellings. Blocks are typically made up of about 30 dwellings in metropolitan areas. Within selected CDs, blocks boundaries and block counts are often formed by interviewers in the field; because this has been an expensive process to date, the selected clusters for both SSS and LFS are located within the same CD.

In this paper we assume that the LFS and SSS use GREG, where $\mathbf{x}$ is a vector of 560 post stratum indicators defined by age, sex, state and metropolitan/ex-metropolitan. This is an approximation because: in practice SSSs sometimes use a range of additional auxiliary variables in GREG, including variables that indicate the number of adults and children in the household; the LFS also uses an indicator variable for sex in each dissemination region. We also assume that SSS design selects half the number of CDs (and blocks) as the LFS but the number of dwellings per cluster is the same.

4. EMPIRICAL EVALUATION

At the time of writing this paper, CD is the smallest geographical level about which Census data is available. The focus here is to measure the reduction in variance of survey estimates due to balancing the sample of CDs on a set of CD level variables. Balancing the sample at the CD level in SRA potentially reduces the variance due to the CD stage of selection but will have no effect on the variance due to the block and cluster stages of selection. To avoid significant complications, this evaluation considers only SRA and CDs that did not change between the 2001 and 2006 Census (for some discussion on this see Section 7).

Section 4.1 describes the data used in the empirical study. Section 4.2 measures the proportion of total variance that is due to the CD stage of selection. Section 4.3 compares the variance of survey estimates under the cube and current sampling methods when there is a five year delay between when the sample of CDs is selected and when the survey data is collected. Section 4.4 considers how this comparison changes during the five year design period.

4.1 Description of the data

This study used the 1996 and 2001 Census data. The Census variables used in this evaluation were identified as having a similar definition to the most important variables collected by the LFS and SSSs. Due to time constraints the variables in this study were restricted to be only at the person level.

The first step involved deriving 14 dummy variables at the person level covering personal income (seven categories), labour force status (two categories) and education (five categories). The definitions of the variables were the same for the 1996 and 2001 Census data and are listed in table 4.1. The 14 proxy survey variables from the 2001 Census are denoted by $\mathbf{y}$ and the 14 design variables from the 1996 Census are denoted by $\mathbf{z}$. Note that in theory $\mathbf{z}$ and $\mathbf{y}$ do not have to be the same set of variables; in practice a larger set of $\mathbf{z}$ variables are preferred given the multi-purpose nature of the surveys that the sample supports.

The second step involved keeping only those CDs which had the same boundaries in both the 1996 and 2001 Censuses. This represented about 70% of all CDs. This avoids confounding changes in CD characteristics between the 1996 and 2001 with changes to CD boundaries between the 1996 and 2001 Census. (A more rigorous approach would also include 1996 CDs that are subdivided to form multiple 2001 CDs.)

It should be pointed out that CD boundary changes are commonly due to significant dwelling growth. It is likely that areas undergoing such changes are more likely to experience changes in their characteristics over time. This may mean that in these areas the design variables ($\mathbf{z}$) are likely to be less correlated with the proxy survey

variables ($\mathbf{y}$) compared with those areas not experiencing CD changes. Therefore excluding the 30% of CDs that experienced boundary changes from this empirical evaluation *may* lead to a inflated estimate of the gains under cube sampling.

Thirdly, the variables were transformed. We define $\mathbf{z}$ to be the 14 Census data items corresponding to the 1996 Census and $\mathbf{y}$ to be the corresponding variables for the 2001 Census. If we define x_{ij}^{01} to be the auxiliary variable for person i in CD j, the variable of interest for the purpose of balancing is the GREG residual

$$\dot{y}_{ijk} = y_{ijk} - \mathbf{B}_{y_k} x_{ij}^{01}$$

where y_{ijk} is unit i's response to data item k in CD j, $\mathbf{B}_{y_k}$ is defined in Section 1 and the elements of $\mathbf{x}$ are defined at the end of Section 3.

The corresponding variable of interest for CD j is given by

$$\dot{\mathbf{Y}}_{jk} = \sum_{i=1}^{N_j^{01}} \dot{y}_{ijk}$$

where N_j^{01} is the number of people in CD j in 2001.

In order to maximise the correlation between the variable of interest, $\dot{y}_{ijk}$, and the balancing variables empirical evaluation suggests that one should transform the balancing variables $z_{i'jg}$ in the same way that y_{ijk} was transformed to $\dot{y}_{ijk}$ the transformed person level balancing variables are then given by

$$\dot{z}_{i'jg} = z_{i'jg} - \mathbf{B}_{z_g} x_{i'j}^{96}$$

where $z_{i'jg}$ is the response from unit i' in CD j and $\mathbf{B}_{z_g}$ has the same form as $\mathbf{B}_{y_k}$ except that z_g replaces y_k; the corresponding CD level balancing variables are given by

$$\dot{\mathbf{Z}}_{jg} = \sum_{i'=1}^{N_j^{96}} \dot{z}_{i'jg}$$

where N_j^{96} is the number of people in CD j in 1996.

The probability of selecting CD j in the LFS during the 1996 design period is given by

$$\pi_{js} = \frac{C_j^{96}}{k_s}$$

where C_i^{96} is CD j's number of clusters and k_s^{-1} is the sampling fraction in state s.

The probability of selecting CD *j* in a SSS during the 1996 design period is

$$\pi^{*}_{js} = \frac{1}{2}\,\pi_{js}$$

(i.e. half the probability of being selected in the LFS). For the purpose of this evaluation we shall evaluate the alternative design strategies using data from the 2001 Census.

4.2 Potential gains from balancing at the CD level

The total variance of a LFS or SSS estimate $Var\left(\hat{\mathbf{Y}}_{k}^{reg}\right)$, can be broken down into the components due selecting a sample of:

1. CDs probability proportional to size, implemented using a systematic skip
2. Block probability proportional to size and
3. Cluster by simple random sampling.

The total variance is a function of various population parameters (e.g. population size), stratification boundaries, auxiliary data used in estimation, and two design parameters, which are the number of dwellings in a cluster and number of selected CDs. The population parameters, estimated from the Census, and the design parameters for the current design were substituted into the formula and the results are summarised in table 4.1.

Table 4.1 gives the proportion of the total variance due to each stage of selection for the estimates in this study. For example, 22% of the total variance of employment estimates is due to the first stage selection. This 22% also represents the maximum possible gain from balancing the sample at the CD level.

Importantly, if the number of CDs selected was decreased while the sample size was unchanged the percentage of total variance due to the CD stage of selection, and the potential gains due to cube sampling, would also increase. This requires further investigation.

4.1 Percentage of the total variance due to each stage of selection[1,2]

	Stage of selection		
Estimator	*CD*	*Block*	*Cluster*
Employment	22	17	61
Unemployment	8	15	77
Income category 1 (<$0)	5	15	80
Income category 2 ($0–$199)	10	16	74
Income category 3 ($200–$399)	5	16	79
Income category 4 ($400–$599)	6	15	80
Income category 5 ($600–$799)	6	14	80
Income category 6 ($800–$999)	7	14	80
Income category 7 (≥ $1000)	23	15	62
Post graduate	9	13	79
Diploma	6	14	81
Bachelor	15	14	71
Advanced Diploma	5	14	80
Certificate	7	15	79

1 This table assumes that SSS designs select half the number of CDs as the LFS but have the same number of dwellings per cluster

2 This required allocating each dwelling in the 2001 Census to a cluster, block and CD and using standard formula for estimating the components of variance due to each stage of selection.

4.3 Reduction in the first stage variance due to cube sampling after five years

In practice the time gap between the sample selection, using the last Census as a source of design information, and collection of the survey data ranges between about one and six years. The aim of this subsection is to measure the gains due to cube sampling when there is a five year gap between sample selection and collection of the survey data. A time gap of five years was chosen because Census data is available every five years.

Moreover, in this section we compare the first stage variance (FSV) of estimates under both cube sampling and under the current method, where the estimates in question are calculated from data five years after the sample is selected. To do this we selected a large number of samples of CDs that were balanced on 1996 Census CD-level variables and, for these samples, measured the first stage variance of the estimated 2001 Census population totals given in table 4.1. In this way the variables used in the design are obtained from the 1996 Census and the proxy for the survey variables are obtained from the 2001 Census. This process of selecting CDs was repeated for the current method, given by 1. above.

The variables used in this evaluation are:

$\dot{\mathbf{Z}}_j$ with elements $\dot{Z}_{jg}$ that were obtained from the 1996 Census,

$\dot{\mathbf{Y}}_j$ with elements $\dot{Y}_{jk}$ that were obtained from the 2001 Census,

$\pi_{js} = \dfrac{C_j^{96}}{k_s}$, $\pi_{js}^* = \dfrac{1}{2}\pi_{js}$ and C_j^{96}.

An additional set of balancing variables were used to ensure a fixed sample size of m_h CDs in stratum h and state s, where $h = 1, \ldots, H_s$ and H_s is the number of strata in a state s. This set of balancing variables for CD j, denoted by $\mathbf{a}_{sj}$, can be achieved by balancing the sample on the H_s vector of variables $\mathbf{a}_{sj}$, where $\mathbf{a}_{sj}$ has h-th element of $\delta_{shj}\pi_{js}$ (or $\delta_{shj}\pi_{js}^*$ for a SSS), where $\delta_{shj} = 1$ if unit j is located in stratum h and state s and equals zero otherwise. Note: all variables are defined at the CD level.

Consider four scenarios or ways in which the sample of CDs can be selected:

1. Cube sampling for LFS variables

   The balancing variables are the two variables in $\dot{\mathbf{Z}}_j$ corresponding to employment and unemployment at the state level, and $\mathbf{a}_{sj}$. The probabilities for the design are π_{js}.

2. Cube sampling for SSS variables

   The balancing variables are the twelve variables in $\dot{\mathbf{Z}}_j$ corresponding to education and income, and $\mathbf{a}_{sj}$. The probabilities for the design were π_{js}^*.

3. Current method for the LFS

   Systematic selection of CDs that have been sorted in a geographically serpentine fashion and where the probability of selection is π_{js}.

4. Current method for the SSS

   Same as 3 except that the probability of selection is π_{js}^*.

Under each of these scenarios, 200 independent balanced samples of CDs were selected. The FSV of $\hat{Y}_k^{reg}$ (i.e. the variance due to the CD stage of selection) under scenario m is calculated by

$$FSV\left(\hat{\mathbf{Y}}_{mk}^{reg}\right) = \frac{1}{200}\sum_{r=1}^{200}\left(\hat{\dot{\mathbf{Y}}}_{mkr} - \bar{\dot{\mathbf{Y}}}_{mk}\right)^2 \qquad (3)$$

where $$\hat{\dot{\mathbf{Y}}}_{mkr} = \sum_{j \in s_{mr}} \pi_j^{-1}\dot{\mathbf{Y}}_{jk} \quad , \quad \bar{\dot{\mathbf{Y}}}_{mk} = \frac{1}{200}\sum_r \hat{\dot{\mathbf{Y}}}_{mrk}$$

and s_{mr} is the r-th simulated sample selected under scenario m. (Note: If $m = 2$ or 4 then π_{js} should be replaced by π_{js}^* in (3).)

Tables 4.2 and 4.3 give the percentage reduction in the FSV under scenarios 1 and 2 relative to scenarios 3 and 4 respectively. That is, the percentage reduction in the FSV under scenario 1 and scenario 2 are

$$\left(1-\frac{FSV\left(\hat{\mathbf{Y}}_{1k}^{reg}\right)}{FSV\left(\hat{\mathbf{Y}}_{3k}^{reg}\right)}\right)\% \quad \text{and} \quad \left(1-\frac{FSV\left(\hat{\mathbf{Y}}_{2k}^{reg}\right)}{FSV\left(\hat{\mathbf{Y}}_{4k}^{reg}\right)}\right)\%$$

respectively. For example, the results show that when balancing the sample of CDs on 1996 employment and unemployment characteristics that the FSV of employment and unemployment estimates in 2001 are 74% and 42% lower compared with the current sampling method. This suggests that the characteristics at the CD level are relatively stable over time.

4.2 Percentage reduction in FSV after five years under the cube method

Estimator	*Percentage reduction (%)*
Employment	74
Unemployment	42

4.3 Percentage reduction in FSV after five years under the cube method

Estimator	*Percentage reduction (%)*
Income category 1	22
Income category 2	57
Income category 3	50
Income category 4	30
Income category 5	49
Income category 6	37
Income category 7	71
Post graduate	64
Diploma	29
Bachelor	60
Advanced Diploma	44
Certificate	32

(Aside: even though in practice the CDs selected in a SSS are constrained to be a subset of the CDs selected in the LFS, for simplicity this constraint is not imposed in this simulation: in this simulation the selection of CDs using cube sampling is done independently for the LFS and SSS.)

While tables 4.2 and 4.3 indicate there are significant gains under options 1 and 2, these gains refer to the FSV which is only a small component of the total variance (see table 4.1). Also, while the gains allow a five year gap between when the design variables are measured and when the sample data is collected, in practice this gap ranges from one to six years. (See Section 4.4 for more on this.)

4.4 Reduction in the total variance due to cube sampling over a five year design period

Of interest is the reduction in total variance due to cube sampling over a five year design period. To approximate this assume the reduction in the FSV under cube sampling changes linearly over time from 100% at the time of the 1996 Census to the levels given in tables 4.2 and 4.3 at the time of the 2001 Census. The reduction in the FSV is 100% [1] at the time of the Census because the sample is balanced on the 1996 Census variables which *are* the characteristics of the interest.

Table 4.4 gives the average reduction to the total variance of employment and unemployment estimates due to cube sampling on average during a five year design period. For a range of practical reasons, the five year design period begins one year after the latest Census (i.e. 1997 to 2002, 2003 to 2008, etc.).

To illustrate how table 4.4's figures are derived consider scenario 1, where the reduction in first stage variance for estimates of employment is 100% in 1996 and 74% in 2001. With linear extrapolation, that amounts to an annual decrease of 5.2%. Taking into account that design information is between one and six years out-of-date (or 3.5 years out of date on average), the average reduction in the first stage variance would be about 82% (100% – 3.5 × 5.2%). An 82% reduction in the first stage variance equates to an 18% (82% × 22%) reduction in the total variance, noting from table 4.1 that 23% of the total variance in employment estimates is due to the first stage of selection.

(Note: we could consider non-linear interpolation methods, such as exponential. However it is not easy to justify one interpolation method over another).

4.4 Average percentage reduction in total variance after five years under cube sampling (Scenario 1)

Estimator	*Percentage reduction (%)*
Employment	18
Unemployment	5

1 The reduction in the FSV was equal to 100% after rounding to the nearest percentage point. This means the sample was *almost* but not strictly balanced. If the sample was balanced the reduction in the FSV would have been exactly 100%.

Table 4.5 gives the corresponding figures to table 4.4 except that they relate to SSS estimates. As SSSs may occur at any point during the design period, table 4.5 gives the reduction in the total variance at one year, 3.5 years and six years after the last Census.

4.5 Reduction total variance during the design period (Scenario 2)

	Years after last Census		
Estimate	*1 year*	*3.5 years*	*6 years*
Income category 1	4	2	0
Income category 2	9	7	5
Income category 3	5	3	2
Income category 4	5	3	1
Income category 5	5	4	2
Income category 6	6	4	2
Income category 7	22	18	15
Post graduate	8	7	5
Diploma	5	3	1
Bachelor	14	11	8
Advanced Diploma	4	3	2
Certificate	6	4	1

5. VARIANCE ESTIMATION

Deville and Tille (2005) evaluated several variance estimators by taking single stage cube samples, with sampling fractions ranging from 1/5 to 1/2, from artificially-generated populations. The different variance estimators, which all take the form of the HT variance estimator under Poisson sampling, vary only in their finite population correction factor. Their main conclusion is that the choice of correction factor is important (i.e. it certainly cannot be ignored) and that one of their variance estimators performs well. Accordingly, an explicit approximation to the three stage variance under cube sampling would be given by the sum of the variance due to:

1. the CD stage of selection, approximated by a Poisson variance estimator given in Deville and Tille (2005)
2. the block stage of selection which is a probability proportional to size sample design
3. the cluster stage of selection which is approximated by a simple random sampling design.

The ABS' variance estimation method for household surveys is the Jacknife with 30 replicate groups as described in Kott (1998). The ABS has found that replication methods have been very practical because they only require replicate weights to calculate the variance of a wide variety of statistics, and are simple to implement and maintain in estimation systems.

Variance estimation under cube sampling in the ABS household survey context is an issue that would need to be investigated if cube sampling is to be implemented. Of primary interest is whether any change to the current variance estimation method for household surveys is required and, if so, whether these changes require a completely new variance estimation method.

A theoretical justification for a replicate estimator under cube sampling is that its expectation is equal to (or close to) the explicit three stage variance formula mentioned in the first paragraph of this section. Any evaluation of alternative variance estimators would need to be empirical and should be conducted on Census data or on an artificially-generated population where the true variance could be calculated by the Monte Carlo method.

Next we discuss the pros and cons of the Jacknife and Bootstrap as potential variance estimators under cube sampling.

In the case of the Jacknife, the replicate groups are mutually exclusive groups of CDs that are selected from the main sample in the same way that the main sample is selected from the population. It follows that, under cube sampling, each of the 30

replicate groups would need to be balanced on the same set of variables as the main sample. This could possibly be achieved by simply forming the main sample from the union of 30 non-overlapping balanced replicate samples from the population. However under this approach, it is not clear how to fix the number of CDs selected in the main sample at the stratum level given that strata typically have less than 30 selected CDs.

Another replicate variance estimator is the Bootstrap. Each Bootstrap replicate could be formed by independently selecting balanced samples of CDs of arbitrary size from the main sample using the cube method. Selecting the Bootstrap replicate groups in this way puts no constraint on how the main sample is selected. It would seem therefore that the Bootstrap would be a simpler method to implement compared to the Jacknife.

6. FUTURE DIRECTION: USE OF MESHBLOCKS

6.1 Effect on sample efficiency

By the end of 2007 each dwelling enumerated in the 2006 Census will have been mapped to meshblock, a smaller geographic level than CD. A meshblock typically contains about 50 dwellings and is roughly the same size as a block. For the 2011 design period it is likely that LFS and SSS will become a two stage design, with meshblock and cluster being the first and second stages of selection, respectively. Some consequences are mentioned in the remainder of this section.

Balancing the sample on meshblock-level variables is likely to result in greater gains compared with balancing on CD level variables. Firstly, the correlation between the design and survey variables *could* be higher when these variables are defined at the meshblock level rather than the CD level. Argument for a higher correlation is because meshblock is a much smaller geographic unit than CD and so likely to be more homogenous. This would lead to greater gains measured in tables 4.2 and 4.3.

Secondly, as meshblocks and blocks are roughly of equal size, the percentage of the total variance due to sampling meshblocks would be roughly equal to the percentage of total variance due to selecting CDs and blocks. For example, for estimates of employment the component of variance due to selecting a sample of meshblocks would be 36% (=23% +13%, from table 4.1) instead of 23%; working through the calculations behind table 4.4, but replacing 23% with 36%, means that average gains during the five year design period under cube sampling could increase from 20% (see table 4.2) to 31%.

6.2 Rotation of meshblock samples

For SSSs and the LFS there would be a significant number of meshblock rotations during a five year design period. A balanced sample of meshblocks will not continue to be balanced after meshblock rotations, unless the meshblock rotations are carefully managed. There are at least three ways in which this issue could be addressed.

The first option is to continue with the current rotation method. That is, meshblocks which have been exhausted are rotated to the neighbouring meshblock. This could be justified if we assumed that the out-going and incoming meshblocks, which would be geographically adjacent, had similar values of the design variables. In this way the sample would be *closely balanced* after meshblock rotation.

The second option is to create a new first stage selection unit by combining two or three meshblocks, so that it has a sufficient number of dwellings to last a five year design period without a rotation.

The third option is to select multiple non-overlapping balanced sub-samples of meshblocks for the five year design period (for details, see Tille and Favre, 2004). Assume for example there are 24 sub-samples in total, eight are live at any time point during the design, and that a selection can rotate between three meshblocks during the period. Maintaining a balanced sample after a meshblock rotation could be achieved by replacing one of the eight live sub-samples with one of the (16) unused sub-samples, noting that the union of any number of balanced samples is also balanced as long as the probabilities of selection are equal for each sub-sample.

Note that under the third option, the constraints on forming the Jacknife replicate groups would make the selection of the main sample impractical. To ensure the Jacknife replicates are balanced at all times during the design period the 30 replicate groups in each of the 24 sub-samples would need to be balanced. This would require selecting 720 non-overlapping balanced samples.

7. AREAS FOR FURTHER INVESTIGATION

Given the potential efficiency gains from cube sampling we recommend that the method be further investigated with a view for adoption in the 2011 Census-based sample redesign.

Methodological issues for further investigation include:

1. Measuring the gains for household and family level estimates (e.g. number of single parent households) items under cube sampling.
2. Choice of design variables. A large set of design variables will select a sample that best meets the varied requirements of the household survey program.
3. Deciding the level at which the sample is to be balanced. For example should the sample be balanced within each labour force dissemination region or simply at the state level?
4. Identifying the optimal design parameters (number of CDs selected and the number of dwellings in a cluster) under cube sampling and measuring the resulting variance. Intuitively, the number of selected CDs would be much lower under the cube method when compared with the current method.
5. Measuring the impact that differences in the data item definitions of the Census, on the one hand, and SSSs and the LFS, on the other hand, have on the gains due to cube sampling. For example, the Census collects only categorical variables whereas some surveys collect continuous variables (e.g. income in dollars).
6. Measuring the gains due to cube sampling in non-SRA. In non-SRA it may be possible to balance the sample at the PSU level thereby reducing the variance due to the PSU stage of selection. However, it may not be possible to select a balanced sample of CDs from within each selected PSU given the high sampling fractions sometimes involved (e.g. it would be clearly impossible to select a balanced sample of three CDs from a PSU with only four CDs).
7. Measuring the gains due to cube sampling after forced rotation of CDs. After forced rotation the sample will no longer be balanced. This would marginally reduce the gains for cube sampling measured in this report.
8. Developing a replicate variance estimator under cube sampling. This includes considering the variance estimators given Deville and Tille (2005) for complex designs used by ABS household surveys.
9. Maintaining a balanced sample of meshblocks during the design period, even after meshblock rotations.
10. Include CDs that change significantly between the 2001 and 2006 Censuses in the evaluation.

A more practical issue is the delay to implementing the 2011 design as a result of introducing cube sampling. The 2006 design required only CD dwelling counts to select a sample of CDs. However, if cube sampling is implemented in 2011 then a range of data items (e.g. income and employment) would also be required before the sample of CDs or meshblocks is selected. Any delay further increases the time gap between when the design data and survey data are collected. As mentioned in this paper, the current time gap ranges between one and six years – if there is a further delay of one year under cube sampling this time gap would become two to seven years. Obviously such a delay would reduce the gains due to cube sampling.

REFERENCES

Cochran, W.G. (1963) *Sampling Techniques*, John Wiley and Sons.

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

Deville, J. and Tille, Y. (2005) "Variance Approximation under Balanced Sampling", *Journal of Statistical Planning and Inference*, 128, pp. 569–591.

Kott, P. (1998) "Using the Delete-a-Group Jacknife Variance Estimation in Practice", American Statistical Association, *Proceedings of the Survey Research Methods Section*, pp. 763–768.

Särndal, C.E., Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*, SpringerVerlag, New York.

Tille, Y. and Favre, A. (2004) "Coordination, Combination and Extension of Balanced Samples", *Biometrika*, 91, pp. 913–927.

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