2008

Sampling for Subpopulations in Two-Stage Surveys

Robert Graham Clark
University of Wollongong, rclark@uow.edu.au

Publication Details

This working paper was subsequently revised and published as Clark, R.G. (2009). Sampling of subpopulations in two stage surveys. *Statistics in Medicine*, Vol. 28, Issue 28, pp. 3697-3717, DOI: 10.1002/sim.3723.

Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

23-08

Sampling for Subpopulations in Two-Stage Surveys

Robert G. Clark

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

Sampling for Subpopulations in Two-Stage Surveys

Robert G. Clark[1*]

[1] Centre for Statistical and Survey Methodology, School of Mathematics and Applied Statistics, University of Wollongong, NSW 2522, Australia.

[*] *email:* rclark@uow.edu.au

Summary. Many national household interview surveys aim to produce statistics on small subpopulations, such as specific ethnic groups or the indigenous population of a country. In most countries, there is no reliable frame of the subpopulations of interest, so it is necessary to sample from the general population, which can be very expensive. The most common strategies used in practice for sampling rare subpopulations are the use of a large screening sample, and disproportionate sampling by strata. Optimal sample designs have been derived for the case of one-stage sampling, but most household surveys use two or more stages of selection. This paper develops optimal designs for two-stage sampling, where there is auxiliary information on subpopulation membership for each primary sampling unit. Various alternative designs are evaluated using a simulated population derived from the New Zealand Census.

Keywords: disproportionate sampling; household surveys; multi-stage sampling; sampling rare populations; sample design; screening

1. Introduction

Multi-stage household surveys are used in many countries to measure characteristics of subpopulations of interest, including indigenous populations and specific ethnic groups. For example, Australia, Canada and New Zealand all conduct surveys of their indigenous populations, which comprise 2.4% (Webster et al., 2004), 3.3% (Bowlby et al., 2004) and 12.1% (Clark & Gerritsen, 2006) of the total population of the respective countries. Kalton and Anderson (1986) described a range of strategies for sampling subpopulations, where the aim is to produce statistics about the subpopulation but not national statistics. The strategies which can most readily be applied in practice are:

1. Disproportionate Sampling. The population is divided into strata which have some relationship with subpopulation membership. Strata with a higher proportion of people in the subpopulation are assigned a higher sampling fraction.

2. Screening. A large sample is selected. The first step in the data collection process is to identify whether the selected person or household is a member of the subpopulation. If they are, then a questionnaire or interview is administered. Typically the screening sample needs to be very large to ensure that enough members of the subpopulation are selected. It is therefore crucial to find economical means of identifying subpopulation membership; even so, screening is usually very expensive per eligible respondent achieved.

The most common approach in practice is to use a combination of strategy 1 and 2, and Kalton and Anderson (1986) derived optimal sample designs for the case of one-stage sampling.

Other strategies include multiplicity or network sampling, and snowballing. These approaches require responding subpopulation members to identify other subpopulation members, and to provide sufficient contact details to enable some of these to be contacted. This is sometimes feasible and can result in dramatic improvements in cost-efficiency. However, for many surveys of ethnic or indigenous subpopulations, asking subpopulation members to identify others would be considered offensive, particularly in urban areas where the rate of membership is relatively low. Moreover, ethnicity and indigenous population membership is based on self-identification, and this can vary over time (Simpson & Akinwale, 2007), so that identification of others as indigenous could be unreliable. A recent experiment of network sampling for a survey of Japanese-heritage families in Brazil found that network sampling was much cheaper than probability sampling but subject to substantial biases, so the method was recommended only to give rough indicative results and not when accurate population statistics are needed (McKenzie & Mistiaen, 2008).

Multiple frame surveys are another approach that can sometimes be used to sample subpopulations (see Lohr & Rao, 2000 for a recent discussion). Suppose there is a frame which contains the whole population and a frame which contains many but not necessarily all subpopulation members. A survey of the subpop-

ulation using the first frame would have full coverage but would be inefficient with high standard errors given a fixed budget, because a large screening exercise would be needed. A survey using the second frame would be much more efficient but could have substantial undercoverage bias. A multiple frame approach would combine both frames to give low bias and reasonably low standard errors for the subpopulation. A crucial issue is to avoid double counting of individuals who are on both frames; this can be avoided or adjusted for if respondents who are on both frames can be identified.

Kalton and Anderson (1986) derived formulas for the optimal allocation for sampling a subpopulation, using screening and disproportionate sampling by strata. The aim was to estimate either the prevalence of the subpopulation, or means from the subpopulation. In the latter case, the best allocation for fixed sample size of subpopulation members is to make the sampling fraction for each stratum proportional to the proportion of the stratum who belong to the subpopulation. The best allocation for fixed cost, under a cost model including the cost of screening (i.e. identifying whether a person belongs to the subpopulation) and interviewing, is to make the sampling fraction proportional to

$$\sqrt{\frac{\text{density}}{\text{density} + \text{relative.screening.cost}}}$$

where "density" is the proportion of the strata which belongs to the subpopulation and "relative.screening.cost" is the ratio of the cost of screening a person to the cost of interviewing the person.

The use of the square root means that the sampling fractions are only mildly disproportionate. A greater variation in sampling fractions across the strata often appears to be an attractive strategy, but in fact this is less efficient because it leads to greater than optimal variation in selection weights. In the extreme, an overly disproportionate allocation can actually result in higher standard errors for estimates of the subpopulation compared to equal probability sampling, even though the achieved sample size of the subpopulation may be high (Gray, 2005; Wells, 2005). When the strata are not available for the whole population, a variation on strategy 1 is two-phase sampling, where stratifying variables are collected for a first phase sample. A stratified second phase sample is then selected from within this initial sample.

The optimal allocation derived by Kalton and Anderson (1986) is of great practical importance. It is based on one-stage sampling, but is often applied to multi-stage sampling, in particular the general rule that selection probabilities should be approximately proportional to the square root of the density of the subpopulation in the strata. However, it is not clear how the one-stage method should be extended to two or more stages. One possibility would be to have the same probabilities of selection for the final units as the one-stage optimum. This could be achieved by either giving primary sampling units (PSUs) with higher densities of the subpopulation a higher chance of selection, or by using a higher first phase sampling fraction within selected PSUs with higher densities, or by a combination. Another issue is that the use of a screening process may

not be worthwhile in all PSUs. For example, it is intuitively reasonable to omit this phase in PSUs thought to have low densities of the subpopulation. On the other hand, it might be supposed that PSUs with low densities should be fully or mostly screened, to give the best chance of selecting at least a few subpopulation members in these PSUs.

This paper develops answers to these questions. The general approach is to derive design variances for a two-stage, two-phase design, which is thought to be a reasonable approximation to many of the sample designs used in household interviewer surveys with a focus on estimates for small subpopulations. A simple model is then used to derive the anticipated variance (AV), which is the model expectation of the design variance. The design variance is often regarded as the most appropriate measure of precision after the survey has been conducted, but the AV is easier to work with for sample design purposes, because it depends on fewer unknown parameters, and is easier to estimate using the limited information available in the design stage of a survey. (See Sarndal et al., 1992, ch.12 for a discussion of the use of the AV for sample design, and Clark & Steel, 2007 for a recent example.) A cost model is assumed in terms of the number of PSUs, first phase sample sizes and second phase sample sizes, and an optimal design is derived which minimises the anticipated variance subject to a cost constraint.

Section 2 defines notation and derives the design variance and anticipated variance. Section 3 states the optimal design. Section 4 is a numerical comparison of alternative design strategies using New Zealand census data as an example.

Section 5 contains conclusions.

2. Theory on Two-Stage, Two-Phase Sampling for Subpopulations

2.1 Notation and Assumed Design

Primary sampling units (PSUs) (generally small geographic areas) are denoted by g. The set of population PSUs is U_I (of size M) and the first stage sample of PSUs is s_I (of size m). Final units (people) are denoted by i, and the set of all units is U. The set of units in PSU g is U_g.

Subscript A will be used for members of the subpopulation, and subscript B for others. For example, N_A is the total number of units in the subpopulation. The density of the subpopulation in PSU g is denoted $\phi_g = N_{gA}/N_g$.

The sample design is assumed to be as follows. PSUs are selected by Poisson sampling with probabilities π_g. A simple random sample without replacement (SRSWOR) s'_g (of n'_g units) is selected from each selected PSU g. Screening information is collected from units in s'_g, to enable them to be accurately divided into members of the subpopulation and others (s'_{gA} and s'_{gB}, of sizes n'_{gA} and n'_{gB} respectively). It is then assumed that all subpopulation members are selected, and a SRSWOR s_{gB} is selected from s'_{gB}. The final sample in subpopulation g is denoted $s_g = s'_{gA} \bigcup s_{gB}$. Let n_{gB} be the size of s_{gB} and define $f_{gB} = n_{gB}/n'_{gB}$. It is assumed that n'_g and f_{gB} are defined for each PSU g in the population, independent of sampling. The probability of selection for a unit i in stratum h

of PSU g is therefore $\pi_i = \pi_g \frac{n'_g}{N_g}$ if $i \in U_A$ and $\pi_i = \pi_g \frac{n'_g}{N_g} f_{gB}$ if $i \in U_B$.

This design is not intended to exactly cover every design used in practice. Often PSUs would be selected by stratified sampling, or unequal probability sampling rather than Poisson sampling. There may be an intervening stage of selection between PSUs and the final units, for example households may be selected from each PSU and then individuals within households. The design we have assumed is intended to be simple enough to allow optimal designs to be derived, while still capturing the essence of the problem of sampling subpopulations. This will lead us to guidelines which survey designers can then adapt to suit their specific situation.

2.2 Estimation

It is assumed that the generalized regression estimator (e.g. Sarndal et al., 1992) will be used. The variable of interest for unit i is y_i. The aim is to estimate $Y = \sum_{i \in U} y_i$. Typically there is some auxiliary information about the whole population which can be used to enhance estimation of Y. Let $\boldsymbol{x}_i$ be the set of auxiliary variables for unit i, and let $\boldsymbol{X} = \sum_{i \in U} \boldsymbol{x}_i$. The estimator of Y is

$$\hat{Y}_r = \sum_{i \in s} \pi_i^{-1} \left(y_i - \boldsymbol{b}^T \boldsymbol{x}_i \right) + \boldsymbol{b}^T \boldsymbol{X} \tag{1}$$

where

$$\boldsymbol{b} = \left(\sum_{i \in s} \pi_i^{-1} c_i \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} \sum_{i \in s} \pi_i^{-1} c_i \boldsymbol{x}_i y_i$$

is a weighted least squares regression coefficient of y_i on $\boldsymbol{x}_i$,

The estimator $\hat{Y}_r$ is approximately equal to

$$\tilde{Y}_r = Y + \sum_{i \in s} \pi_i^{-1} e_i$$

where $e_i = y_i - \boldsymbol{B}^T \boldsymbol{x}_i$ and

$$\boldsymbol{B} = \left(\sum_{i \in U} c_i \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} \sum_{i \in U} c_i \boldsymbol{x}_i y_i$$

is a population weighted least squares regression coefficient of $\{y_i\}$ on $\{\boldsymbol{x_i}\}$.

We will assume that the weights c_i used in the calculation of regression parameters have the property that $c_i = \boldsymbol{\lambda^T x_i}$ for all $i \in U$, for some vector $\boldsymbol{\lambda}$. In this case, the population mean, $\bar{E}$, of $\{e_i\}$, is zero (this can be shown using the same argument as in Sarndal et al., 1992, Result 6.5.1, p. 231). This condition simplifies a number of our results, and would usually be satisfied in practice. For example it is true if the ratio estimator is used, or if $c_i = 1$ and the auxiliary variables include an element equal to 1 for all i.

We write $e_{g1} = \sum_{i \in U_g} e_i$ for the cluster totals of e_i and $\bar{e}_g = N_g^{-1} e_{g1}$ for the cluster means. The variance for cluster g is $S_g^2 = (N_g - 1)^{-1} \sum_{i \in U_g} (e_i - \bar{e}_g)^2$.

It is further assumed that subpopulation totals are part of the benchmark information, so that $\bar{E}_A = \bar{E}_B = 0$. Annual demographic benchmarks are available for the Māori population in New Zealand (Statistics New Zealand, 2008). In Australia, experimental demographic benchmarks for the Indigenous population have been compiled for a single time point in 2006 and an ongoing time series is planned (ABS, 2008). Statistics Canada has produced projections of the

Aboriginal populations of Canada but not ongoing estimates (Statistics Canada, 2005). In countries without ongoing demographic benchmarks for the subpopulation of interest, approximate benchmarks could be compiled by combining census data and demographic benchmarks for the whole population. If no subpopulation benchmarks can be compiled, the variance expressions in this paper will understate the true variance, although the proposed designs may still be reasonably efficient, particularly if the aim is to estimate means or rates for the subpopulation rather than totals.

2.3 Design Variance for Estimator of Total

Using straightforward but tedious manipulations, the design variance is:

$$\begin{aligned}
var_p\left[\tilde{Y}_r\right] &= var_p\left\{E_p\left[\tilde{Y}_r|s_I\right]\right\} + E_p\left\{var_p\left[\tilde{Y}_r|s_I\right]\right\} \\
&= var_p\left\{\sum_{g\in s_I}\pi_g^{-1}E_g\right\} \\
&\quad + E_p\left\{var_p\left[E_p\left[\tilde{Y}_r|s_I,s'\right]|s_I\right] + E_p\left[var_p\left[\tilde{Y}_r|s_I,s'\right]|s_I\right]\right\} \\
&= \ldots \\
&\approx \sum_{g\in U_I}\pi_g^{-1}\left(E_g^2 - N_gS_g^2\right) + \sum_{g\in U_I}\left(\pi_g n'_g\right)^{-1}N_g^2\left(S_g^2 - (1-\phi_g)\,S_{gB}^2\right) \\
&\quad + \sum_{g\in U_I}\left(\pi_g n'_g f_{gB}\right)^{-1}N_g^2\,(1-\phi_g)\,S_{gB}^2 + const \qquad (2)
\end{aligned}$$

where "const" refers to terms which do not depend on π_g, n'_g or f_{gB}. See Clark et al., 2008, Section 2.2.3 for details of this derivation.

For subpopulation totals, the estimator is identical but with Y_i replaced by

$$Y_{iA} = \begin{cases} Y_i & \text{if} \quad i \in U_A \\ 0 & \text{if} \quad i \notin U_A. \end{cases}$$

This is approximately equal to

$$\tilde{Y}_{rA} = \sum_{i \in s} \pi_i^{-1} E_{iA}$$

where

$$E_{iA} = \begin{cases} E_i & \text{if} \quad i \in U_A \\ 0 & \text{if} \quad i \notin U_A \end{cases}$$

because of the fact that $\bar{E}_A = \bar{E}_B = 0$. The variance is therefore given by substituting E_{iA} in place of E_i in all the terms in (2):

$$\begin{aligned} var_p\left[\tilde{Y}_{rA}\right] &\approx \sum_{g \in U_I} \pi_g^{-1}\left(E_{gA}^2 - N_g \tilde{S}_{gA}^2\right) + \sum_{g \in U_I}\left(\pi_g n_g'\right)^{-1} N_g^2\left(\tilde{S}_{gA}^2 - 0\right) \\ &\quad + \sum_{g \in U_I}\left(\pi_g n_g' f_{gh}\right)^{-1} N_g^2 (1-\phi_g)\, 0 + const \\ &= \sum_{g \in U_I} \pi_g^{-1}\left(E_{gA}^2 - N_g \tilde{S}_{gA}^2\right) + \sum_{g \in U_I}\left(\pi_g n_g'\right)^{-1} N_g^2 \tilde{S}_{gA}^2 + const \end{aligned} \quad (3)$$

where

$$\begin{aligned} E_{gA} &= \sum_{i \in U_{gA}} E_i \\ \tilde{S}_{gA}^2 &= (N_g - 1)^{-1}\left\{\sum_{i \in U_g} E_{iA}^2 - N_g^{-1}\left(\sum_{i \in U_g} E_{iA}\right)^2\right\} \end{aligned}$$

2.4 Anticipated Variance

The following model will be assumed:

$$\begin{cases} E_M\left[E_i\right] = 0 \\ var_M\left[E_i\right] = \sigma^2 \\ cov_M\left[E_i, E_j\right] = \rho\sigma^2 (i \neq j; ij \in U_g) \\ cov_M\left[E_i, E_j\right] = 0 (i \in U_{g1}, j \in U_{g2}, g1 \neq g2) \end{cases} \quad (4)$$

The expectations of the terms in (2) are:

$$\begin{aligned} E_M\left[E_g^2\right] &= var_M\left[E_g\right] = \sigma^2 N_g\left(1 + (N_g - 1)\rho\right) \\ E_M\left[S_g^2\right] &= E_M\left[S_{gB}^2\right] = \sigma^2(1-\rho). \end{aligned}$$

and so $E_M\left[S_g^2-(1-\phi_g)\,S_{gB}^2\right]=\sigma^2\,(1-\rho)\,\phi_g$. It follows that

$$\begin{aligned}
AV\left[\tilde{Y}_r\right] &= E_M var_p\left[\tilde{Y}_r\right] \\
&\approx \sigma^2\sum_{g\in U_I}\pi_g^{-1}\left\{N_g\left(1+(N_g-1)\,\rho\right)-N_g(1-\rho)\right\} \\
&\quad +\sigma^2(1-\rho)\sum_{g\in U_I}\left(\pi_g n_g'\right)^{-1}N_g^2\phi_g \\
&\quad +\sigma^2(1-\rho)\sum_{g\in U_I}\left(\pi_g n_g' f_{gB}\right)^{-1}N_g^2\,(1-\phi_g)+const \\
&= \sigma^2\rho\sum_{g\in U_I}\pi_g^{-1}N_g^2+\sigma^2(1-\rho)\sum_{g\in U_I}\left(\pi_g n_g'\right)^{-1}N_g^2\phi_g \\
&\quad +\sigma^2(1-\rho)\sum_{g\in U_I}\left(\pi_g n_g' f_{gB}\right)^{-1}N_g^2\,(1-\phi_g)+const. \qquad (5)
\end{aligned}$$

Similarly, taking the model expectation of (3) gives

$$AV\left[\tilde{Y}_{rA}\right] \approx \rho\sigma^2\sum_{g\in U_I}\pi_g^{-1}N_g\phi_g\left\{N_g\phi_g-1+\phi_g\right\}+\sigma^2\sum_{g\in U_I}\left(\pi_g n_g'\right)^{-1}N_g^2\phi_g\left(1-\phi_g\rho\right).$$

where the approximation is based on dropping terms of order N_g^{-1} relative to the remainder of the expression. Suppose we further assume that $N_g\phi_g-1+\phi_g\approx N_g\phi_g$. The assumption is reasonably accurate for larger values of ϕ_g, but not for smaller values (e.g. N_{gA} equal to 3 or less). However we will still make this approximation because it substantially simplifies the derivation of the optimal design. Designs based on the approximation will be evaluated empirically in Section 4. The approximation implies:

$$AV\left[\tilde{Y}_{rA}\right]\approx\rho\sigma^2\sum_{g\in U_I}\pi_g^{-1}N_g^2\phi_g^2+\sigma^2\sum_{g\in U_I}\left(\pi_g n_g'\right)^{-1}N_g^2\phi_g\left(1-\phi_g\rho\right). \qquad (6)$$

3. Optimal Allocations

We model the cost as:

$$C = C_0 + C_1 m + C_2 n' + C_3 n$$

where C_0 are fixed costs, C_1 is the cost per PSU in sample (e.g. travel and blocklisting costs), C_2 is the cost per screening interview (including time needed to contact the household or person, callbacks, and the time in collecting the screening data on whether or not the person is in A) and C_3 is the cost per full interview. The expected cost is

$$\begin{aligned} C_E &= E\left[C_0 + C_1 m + C_2 n' + C_3 n\right] \\ &= C_0 + C_1 \sum_{g \in U_I} \pi_g + \sum_{g \in U_I} \left(\pi_g n'_g\right)\left(C_2 + C_3 \phi_g\right) + \sum_{g \in U_I} \left(\pi_g n'_g f_{gB}\right) C_3 \left(1 - \phi_g\right) \end{aligned}$$

Suppose that the aim is to estimate a linear combination of the AV of the subpopulation estimator and the AV of the population estimator with respect to π_g, n'_g and f_{gB}, subject to a cost constraint

$$C_E = C_f. \tag{7}$$

Let P be a value between 0 and 1. The objective measure is defined by:

$$\begin{aligned} F &= (1-P) AV\left[\tilde{Y}_r\right] + P AV\left[\tilde{Y}_{rA}\right] \\ &\approx \rho \sum_{g \in U_I} \pi_g^{-1} N_g^2 \left(P\phi_g^2 + 1 - P\right) + \sum_{g \in U_I} \left(\pi_g n'_g\right)^{-1} N_g^2 \phi_g \left\{1 - \rho\left(P\phi_g + 1 - P\right)\right\} \\ &\quad + (1-P)(1-\rho) \sum_{g \in U_I} \left(\pi_g n'_g f_{gB}\right)^{-1} N_g^2 \left(1 - \phi_g\right) \end{aligned} \tag{8}$$

(excluding constants, and substituting $\sigma^2 = 1$, as this does not affect the optimal design). If only the subpopulation was of interest, then P would be set to 1, and if only the total population was of interest, P would be set to 0.

Theorem 1 states the optimal sample design for F.

Theorem 1: Optimal Design

Let U_I^a be the set of PSUs g satisfying

$$P\phi_g \geq \frac{C_2}{C_3}(1-P)(1-\rho) \tag{9}$$

and let U_I^b contain other PSUs. It is assumed that $\rho\phi_g << 1$. The values of π_g, n'_g and f_{gB} which minimise F in (8) subject to $C_E = C_f$ and $f_{gB} \leq 1$ are

$$\left.\begin{array}{rcl} \pi_g & \propto & N_g\sqrt{\frac{\rho\left(P\phi_g^2+1-P\right)}{C_1}} \\ n'_g & = & \begin{cases} \sqrt{\frac{\phi_g(1-\rho(1-P))}{\rho\left(P\phi_g^2+1-P\right)}\frac{C_1}{C_2+C_3\phi_g}} & \text{if} \quad g \in U_I^a \\ \sqrt{\frac{(1-P)(1-\rho)+P\phi_g}{\rho\left(P\phi_g^2+1-P\right)}\frac{C_1}{C_2+C_3}} & \text{if} \quad g \in U_I^b \end{cases} \\ f_{gB} & = & \begin{cases} \sqrt{\frac{(1-P)(1-\rho)}{(1-\rho(1-P))}\frac{C_2+C_3\phi_g}{C_3\phi_g}} & \text{if} \quad g \in U_I^a \\ 1 & \text{if} \quad g \in U_I^b \end{cases} \end{array}\right\} \tag{10}$$

Proof: See Appendix.

Special Case: Estimating the Population Total Only

Consider the special case when national estimates are the only priority, so that $P = 0$. It is clear that (9) is never satisfied, so that every PSU belongs to

U_I^b. The optimal design comes from substituting $P = 0$ into (10):

$$\left.\begin{array}{rcl} \pi_g & = & \lambda N_g \sqrt{\frac{\rho}{C_1}} \\ n'_g & = & \sqrt{\frac{(1-\rho)}{\rho}\frac{C_1}{C_2+C_3}} \\ f_{gB} & = & 1 \end{array}\right\} \tag{11}$$

This is the standard optimal two-stage design for estimating a population total, see for example Hansen et al. (1953).

Special Case: Estimating the Subpopulation Total Only

Another special case of interest is when only the subpopulation estimates are important, and national estimates are irrrelevant, so that $P = 1$. In this case, (9) is always satisfied, and the optimal design is:

$$\left.\begin{array}{rcl} \pi_g & = & \lambda N_g \phi_g / \sqrt{\rho / C_1} \\ n'_g & = & \sqrt{\frac{\phi_g^{-1}}{\rho}\frac{C_1}{C_2+C_3\phi_g}} \\ f_{gB} & = & 0 \end{array}\right\} \tag{12}$$

The probability of selection for units in the subpopulation is proportional to $\sqrt{\frac{\phi_g}{C_2+C_3\phi_g}}$, the same as in Kalton and Anderson (1986). However, the optimal design achieves this in a surprising way. The probability of selection of PSUs is proportional to ϕ_g, which means targeting high density PSUs more aggressively than the square root of the density. However, the sample sizes within PSUs are *inversely* related to the density ϕ_g.

It is of interest to see the optimal design when screening is free. In this case, $C_2 = 0$ and the design becomes:

$$\left.\begin{array}{rcl} \pi_g & = & \lambda N_g \phi_g / \sqrt{\rho / C_1} \propto N_{gA} \\ n'_g & = & \phi_g^{-1} \sqrt{\frac{C_1}{C_3 \rho}} \\ f_{gB} & = & 0 \end{array}\right\}$$

which implies

$$n_{gA} = \phi_g n'_g = \sqrt{\frac{C_1}{C_3 \rho}}$$

which is very close to the standard two-stage self-weighting design, treating A as the whole population of interest. This makes sense because in this scenario, there is no cost from identifying the subpopulation membership for everyone in U.

Special Case: A Particular Compromise Allocation

Power allocations are a method of allocating sample to strata where the precision of both stratum and national estimates are important (Bankier, 1988). An exponent between 0 and 1 defines the relative priority of stratum versus national precision. Suppose that we define "strata" to be A and B, that is subpopulation members and non-members. Suppose that the exponent is 0.5 indicating that national and subpopulation estimates are of equal importance in some sense. Further suppose that the stratum population means are all equal, and that the importance measure X_h for stratum h in Bankier's notation is set to the population size for the stratum. Then the objective criteria in Bankier's expression (2.1) is approximately equivalent to our F defined in (8) with P set to $P = 1/\left(1 + \bar{\phi}\right)$, where $\bar{\phi} = N_A/N$.

In this case, the cutoff for subsampling becomes $\phi_g \geq \frac{1-P}{P}\frac{C_2}{C_3} = \phi\frac{C_2}{C_3}$, i.e. $\frac{\phi_g}{\phi} \geq \frac{C_2}{C_3}$. For example, suppose $\frac{C_2}{C_3} = 0.3$ and $\bar{\phi} = 0.11$ (as for Māori in New Zealand). Then the cutoff is $\phi_g \geq 0.11 \times 0.3 = 0.033$. So subsampling would be used in PSUs where the proportion of Māori is 3.3% or more.

This definition of P gives a reasonable first attempt at a good design for both national and subpopulation estimates, and will be used in the numerical study in Section 4. In practice, the relative priority of national and subpopulation estimates is likely to be a difficult decision. P would normally be chosen by providing a number of options to the survey owner or sponsor, or an advisory group of survey users. A value of P would be chosen after perusing a table of the standard errors for national and subpopulation estimates that would be expected from each option for P.

An Alternative Within-PSU Selection Method

An alternative way of selecting the sample within each PSU g is to:

- Select a main sample of $n_{g(main)}$ units. All of these units are then interviewed.
- Select an oversample of $n_{g(over)}$ units. Of these units, only those in the subpopulation, A, are interviewed.

Wells (1998) referred to a number of studies in New Zealand and the United States which have used this approach, and developed and compared weighting methods for the design. The approach is similar to the design we have assumed, so that our results could also be applied to this case. The correspondence between $\left(n'_g, f_{gB}\right)$ and $\left(n_{g(core)}, n_{g(boost)}\right)$ is

$$\begin{aligned} n'_g &= n_{g(main)} + n_{g(over)} \\ f_{gB} &= n_{g(main)} / \left(n_{g(main)} + n_{g(over)}\right) \end{aligned}$$

4. Numerical Study

4.1 Simulated Data

Nine alternative sample designs were compared empirically using two binary variables simulated using 2001 New Zealand (NZ) Census meshblock data. Meshblocks are a small geographic area containing on average 94 adults (15 and over). Virtually the whole population was used, giving a total of 32,168 meshblocks after deleting four meshblocks containing less than 5 people and one in "areas outside Territorial Authority". PSUs were defined to be meshblocks as this is the case in many NZ surveys including the NZ Health Survey. The subpopulation A was defined to be Māori adults, and B consisted of other adults. The approach was to use the census values of N_g and N_{gA}, and to simulate two Y-variables for the population.

The total population size for these meshblocks was approximately 3.01 million, of whom approximately 325,000 were Māori (10.8%). Figure 1 shows the distribution of the density of Māori adults, ϕ_g. It can be seen that there are relatively few PSUs with a high density, making it difficult to geographically target the sample in order to efficiently over-sample Māori. In fact, only 28% of Māori adults live in PSUs where Māori comprise more than 30% of the PSU population, and only 10% of Māori live in PSUs where Māori are a majority.

The two Y-variables were simulated from the beta-binomial distribution, to

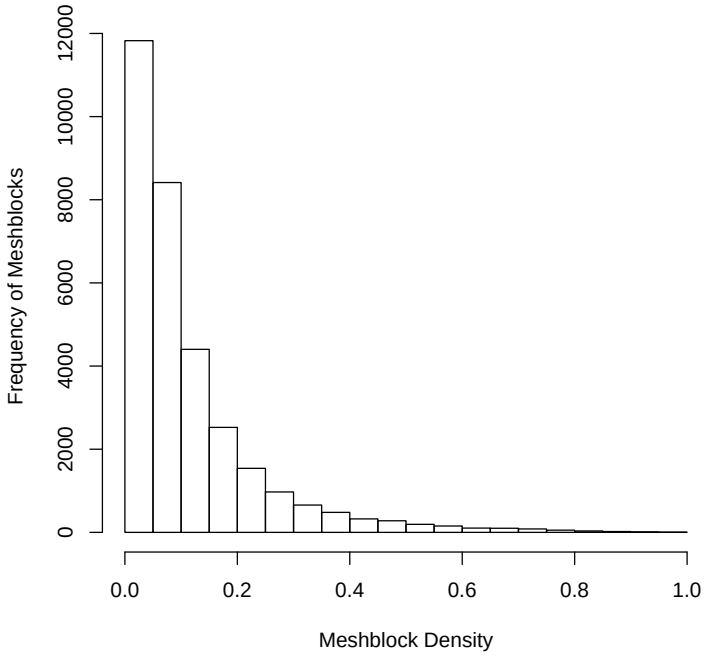

Figure 1: Distribution of Meshblock Density of Māori Adults (ϕ_g)

give correlated data within PSUs. This distribution is defined by:

$$P_i \sim Beta\,(\alpha, \beta)$$

$$Y_k \sim binomial\,(1, P_i) \text{ conditional on } P_i$$

where Y_k is the value of the binary variable for person k in PSU i. The data was generated such that $P\,[Y_k = 1] = 0.5$ for all k, and the intra-PSU correlation, ρ, was 0.025 and 0.1 for variables 1 and 2, respectively. This was achieved by setting $\alpha = \beta = 0.5\,(\rho^{-1} - 1)$ in the Beta distribution. The model implies that

$$Y_{gA} \sim binomial\,(N_{gA}, P_i)$$

$$Y_{gB} \sim binomial\,(N_{gB}, P_i)$$

conditional on P_i. Values of Y_{gA}, Y_{gB}, were generated from this model for each PSU g. The values of S^2_{gA}, S^2_{gB} and $\tilde{S}^2_{gA}$ were calculated using these values, as for binary data these quantities depend in a simple way on Y_{gA}, Y_{gB}, N_{gA} and N_{gB}.

For each design, the design variances of the estimators for the population and subpopulation totals were calculated using formulas (2) and (3), respectively, so the evaluation avoided the approximations and modelling assumptions made in Sections 2.4 and 3. It was assumed that the generalized regression estimator is used with the only auxiliary variable being membership of the subpopulation, A or B. Thus $E_k = Y_k - \bar{Y}_A$ for $k \in U_A$ and $E_k = Y_k - \bar{Y}_B$ for $k \in U_B$, where $\bar{Y}_A = Y_A/N_A$ is the mean of the variable over all people in subpopulation A and $\bar{Y}_B = Y_B/N_B$ is the mean over all people in B. E_{gA} and E_{gB} were calculated accordingly.

4.2 Designs Considered

(1) Approximately optimal design for the population total from (11).

(2) Approximately optimal design for the subpopulation total from (12). This design results in only members of the subpopulation being interviewed. PSUs containing no Māori adults have zero chance of selection.

(3) Target at First Stage Only. Set $\pi_g \propto N_g\sqrt{C_1\phi_g/\left(C_2 + C_3\phi_g\right)}$, $n'_g = \sqrt{\frac{1-\rho}{\rho}\frac{C_1\bar{\phi}}{C_2+C_3\bar{\phi}}}$ and $f_{gB} = 0$. This results in the person probabilities of selection being the same as in design 2, but the targeting all occurs at the first stage of selection.

(4) Target at First Stage, and also at Second Stage using a Rough Rule. This is the same as design 3, except that PSUs where $\phi_g \geq \bar{\phi}$ have their screening sample size (n'_g) increased by 20% and other PSUs have their screening sample size decreased by 20%. This option was included because it often seems preferable in practice to make the most of a selected PSU with high ϕ_g by topping up its screening sample size.

(5) Target at Second Stage only: $\pi_g \propto N_g$; $n'_g = \sqrt{\frac{\phi_g(1-\rho)C_1\left(C_2+C_3\bar{\phi}\right)}{\rho\bar{\phi}(C_2+C_3\phi_g)}}$; $f_{gB} = 0$. This results in the person probabilities of selection being the same as in design 2 and 3, but the targeting all occurs at the second stage of selection.

(6) This is the approximately optimal design for the combined criterion from (10), with $P = \left(1+\bar{\phi}\right)^{-1}$ (Carroll Allocation).

(7) This is a simplified compromise design. The design has π_g the same as design (2), but with ϕ_g replaced by $\tilde{\phi}_g = \left(\phi_g + \bar{\phi}\right)/2$, and f_{gB} is equal to 0.5 for all g.

(8) This is another simplified compromise design, with $\pi_g \propto N_g\sqrt{P\phi_g^2 + 1 - P}$, $n'_g = 20$ and $f_{gB} = 0.5$.

All of the designs were based on an intra-MB correlation of $\rho = 0.025$, and a cost model with $C_1 = 2$, $C_2 = 0.3$ and $C_3 = 1$. All of the designs were normalized to cost 10,000 units according to this cost model. The values of n'_g were rounded to the nearest whole number 1 or higher, and truncated so that $n'_g \leq N_g$ in all

cases.

4.3 Results

Table 1 shows the design variances for each design for variables 1 and 2, expressed as relative standard errors. The combined criterion $F = (1-P)var\left[\tilde{Y}_r\right] + Pvar\left[\tilde{Y}_{rA}\right]$ is also shown for each variable (divided by 1e7 for readability). Note that designs 2, 3, 4 and 5 give no chance of selection to non-members of the subpopulation, so the RSEs for the overall population and the combined criteria are not shown.

For variable 1, design 2 (the approximately optimal design) is the best for the purpose of estimating the subpopulation total Y_A. For variable 2, design 3 (target at first stage only) performs slightly better. This may be because the intra-MB correlation for variable 2 was considerably higher than that assumed in design 2 ($\rho = 0.025$ vs $\rho = 0.1$). Designs 4 and 5 have RSEs around 5-10% higher. This is equivalent to the sample size increasing by 10%-20% for the same precision, so this is a substantial inefficiency. It can be concluded that the best approach is to use the optimal design; if not, to target at the first stage only. Targeting at the second stage only is less efficient, and ad hoc targeting at both stages is less efficient again.

The approximately optimal compromise design, design 6, substantially reduced the combined criterion relative to design 1, not surprisingly. Two simplified versions, designs 7 and 8, performed worse than design 6.

The designs derived in Section 2 assumed knowledge of ρ and the cost pa-

Table 1: Comparison of Design Variances of Alternative Designs

Design	Variable 1(ICC=0.025)			Variable 2(ICC=0.1)		
	RSE(%) Overall	RSE(%) Subpopulation	Combined Criterion	RSE(%) Overall	RSE(%) Subpopulation	Combined Criterion
Design 1	1.35	3.85	7.58	1.63	4.08	9.82
Design 2	n/a	2.01	n/a	n/a	2.30	n/a
Design 3	n/a	2.12	n/a	n/a	2.30	n/a
Design 4	n/a	2.21	n/a	n/a	2.37	n/a
Design 5	n/a	2.33	n/a	n/a	2.50	n/a
Design 6	1.43	3.10	6.81	1.74	3.35	9.35
Design 7	1.52	3.18	7.55	1.72	3.27	9.06
Design 8	1.52	2.94	7.19	1.91	3.30	10.69

rameters C_1/C_3 and C_2/C_3. In reality neither would be known perfectly. The intraclass correlation would need to be estimated by judgement, or using data from a pilot survey or a previous census or survey. Also, the design would be based on just one value of ρ, but many variables would be collected, with a different ρ for each. The cost parameters are also difficult to acquire and would usually be based on the judgement of a statistician or survey manager, or on cost data recorded in past surveys, which would be subject to many errors. Thus it is important to understand how the various designs perform if the assumed value of ρ or the cost model is incorrect.

Table 2 shows how the designs based on $\rho = 0.025$ perform for estimating the subpopulation total when the true ρ varies. The cost model $C_1 = 2$, $C_2 = 0.3$ and $C_1 = 1$ is assumed to be correct. The last row of the table shows the performance of Design 2 (the approximately optimal design for subpopulation estimates) when the correct ρ is used rather than $\rho = 0.025$. The table was calculated by simulating data based on the desired value of ρ using the method described earlier in this section. The table shows that Design 2 based on $\rho =$

0.025 is substantially better than Design 1 based on $\rho = 0.025$, for every value of the true ρ considered. Thus Design 2 is an effective method of targeting a subpopulation compared to not targeting at all, even if the assumed ρ is incorrect.

Comparing Design 2 assuming $\rho = 0.025$ to Design 2 using the true ρ, shows that the former is very efficient even if the true ρ departs moderately from 0.025 (in the range 0.01 - 0.05). There is moderate inefficiency at $\rho = 0.1$, increasing as ρ becomes larger. For $\rho = 0$, the best design is much more efficient than assuming $\rho = 0.025$. However, the design based on $\rho = 0$ would not be a sensible option in practice, because it completely enumerates every person in the selected PSUs, which would be highly nonrobust to small departures from $\rho = 0$. Design 2 is also superior to Designs 3, 4 and 5 for ρ between 0 and 0.1.

The conclusion is that Design 2 is a reasonably efficient design compared to not targeting, and compared to other designs for subpopulations, even if ρ is not known precisely.

Table 2: RSE(%) for Subpopulation Estimator of Designs assuming $\rho = 0.025$ and $C_2 = 0.3$ for Various Values of ρ

Design	ρ							
	0	0.01	0.025	0.05	0.1	0.2	0.5	1
Design 1	3.80	3.82	3.85	3.91	4.08	4.33	5.08	6.05
Design 2	1.93	1.96	2.01	2.10	2.30	2.62	3.42	4.40
Design 3	2.07	2.09	2.12	2.17	2.30	2.50	3.07	3.78
Design 4	2.17	2.18	2.21	2.26	2.37	2.56	3.08	3.75
Design 5	2.29	2.31	2.33	2.38	2.50	2.69	3.22	3.91
Design 2 with Correct ρ and C_2	2.42	1.98	2.01	2.09	2.22	2.37	2.68	2.98

Table 3 shows how well designs based on $\rho = 0.025$ and the cost model $C_1 = 2$, $C_2 = 0.3$ and $C_1 = 1$ perform when the true cost parameters vary. The assumed

value of ρ is assumed to be correct. The main focus of this article is on efficient targeting and screening for subpopulations, rather than on optimal clustering, so only the cost parameter C_2 associated with screening is varied, and not the cost parameter associated with travel and blocklisting, C_1. If the assumed value of C_2 is incorrect, the effect will be that Designs 1-5 will no longer be of equal total cost, so that comparing the RSEs from these designs would be misleading. To enable a fair comparison of the designs, the values of π_g were multiplied by a constant factor so that each design is of equal cost under the correct cost function. The table shows that Design 2 is the best of Designs 1-5 for every value of C_2 considered. Substantial gains over Design 1 (no targeting for subpopulation) are seen in every case. Design 2 based on the correct cost model is at best slightly more efficient than Design 2 in every case. Design 2 does become substantially less precise if C_2 is much larger than 0.3, however the other 5 designs in the table are affected just as much. The conclusion is that Design 2 is an efficient strategy even if there is uncertainty about the correct cost function.

Table 3: RSE(%) for Subpopulation Estimator of Designs assuming $\rho = 0.025$ and $C_2 = 0.3$ for Various Values of C_2

Design	C_2							
	0	0.1	0.2	0.3	0.4	0.5	0.75	1
Design 1	3.46	3.59	3.72	3.85	3.97	4.09	4.37	4.64
Design 2	1.29	1.57	1.80	2.01	2.20	2.38	2.76	3.10
Design 3	1.49	1.73	1.93	2.12	2.29	2.45	2.81	3.13
Design 4	1.60	1.82	2.03	2.21	2.38	2.53	2.89	3.21
Design 5	1.78	1.98	2.16	2.33	2.49	2.64	2.97	3.28
Design 2 with Correct ρ and C_2	1.27	1.57	1.81	2.01	2.20	2.37	2.74	3.07

4.4 Comparison of Designs 2 and 6

Table 1 suggested that design 2 is the most appropriate method when only the subpopulation is a priority. Design 6 was the best design when both the subpopulation and population totals are important. It is of interest to understand the behaviour of designs 2 and 6, in terms of how the PSU and within-PSU sampling rates changes according to the proportion of the PSU belonging to A. Figure 2 shows how these designs behave for meshblocks of size 100. (Similar behaviour would be expected for other meshblocks but the figure restricts to size 100 for clearer presentation.) The figure is based on the same values of ρ, C_1, C_2 and C_3 assumed in Table 1.

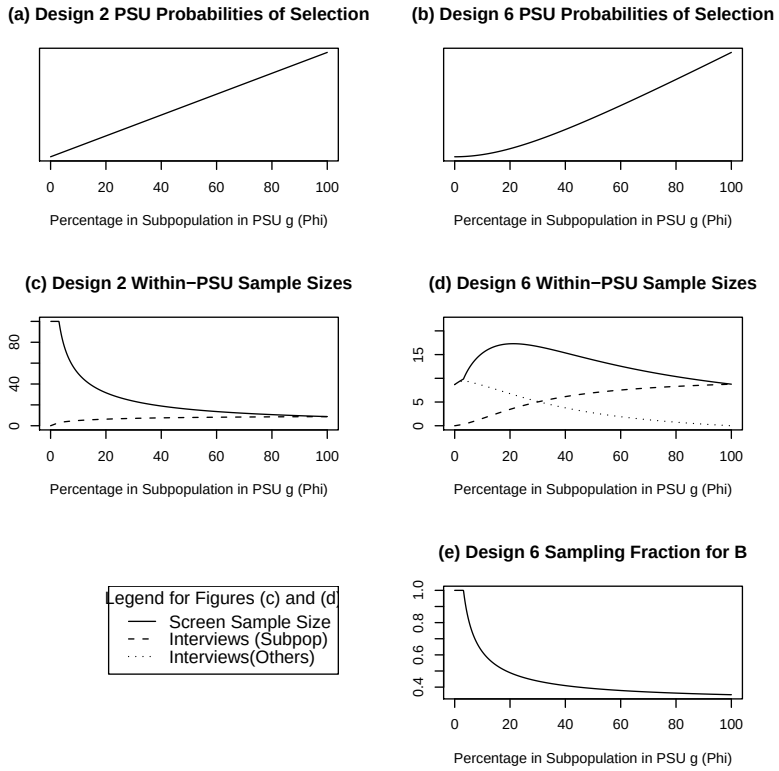

Figure 2: Comparison of Designs 2 and 6 for PSUs containing 100 People

Figure 2(a) shows how the PSU probability of selection, π_g, depends on the proportion of the PSU belonging to A, ϕ_g, for design 2. There is a straight line relationship through the origin for this design. Figure 2(b) shows the same plot for design 6. The values of π_g increase more slowly as ϕ_g increases, compared to design 2, indicating that targeting of high density areas is less aggressive for design 6.

Figure 2(c) shows how the within-PSU sample sizes vary according to ϕ_g. The solid line is the within-PSU screening sample size, n'_g. It can be seen that the screening sample size actually decreases with the density. For ϕ_g less than about 5%, the whole PSU is screened (however, if $\phi_g = 0$, then $\pi_g = 0$ and the PSU is never selected). The dashed line shows the expected sample size belonging to A, $n_{gA} = \phi_g n'_g$. This sample size increases from 0 at $\phi_g = 0$ to about 9 at $\phi_g = 1$.

Figure 2(d) shows similar information for the compromise design, Design 6. The screening sample size is increasing with ϕ_g up to a maximum of about 18 at $\phi_g \approx 20\%$, then gently decreasing. This is contrast to Design 2 where very large screening sample sizes were seen for smaller values of ϕ_g. The expected sample size of A, n_{gA}, performs similarly to Design 2. Figure (d) also shows the expected sample size of non-subpopulation members, n_{gB}. This decreases from about 9 to 0 as ϕ_g increases from 0 to 1. This is not surprising as it means that the sample size from B increases as the proportion of the PSU belonging to B increases.

For design 2, the sampling fraction for non-subpopulation members identified in the screen is $f_{gB} = 0$. For design 6, f_{gB} is non-zero since the overall population

total is of interest, as well as the subpopulation total. Figure 2(e) illustrates this. The value of f_{gB} is decreasing with ϕ_g. Equivalently, as the number of members of B increases, the fraction selected decreases. For values of ϕ_g less than 3%, f_{gB} is equal to 1, so everyone in the screening sample is selected regardless of whether they are in A or B, i.e. the screening information is not used for subsampling and so the screening process could be omitted.

5. Conclusions

If only the subpopulation total is of interest, then the design should be such that the person probabilities of selection are proportional to $\sqrt{\phi_g / (C_1 + C_2\phi_g)}$ as recommended by Kalton and Anderson (1986) for one-stage sampling. This should be achieved by "over-targeting" at the first stage (with PSU probabilities of selection proportional to the density ϕ_g), and "under-targeting" at the second stage (with the screening sample size decreasing with ϕ_g, proportional to $1/\sqrt{\phi_g (C_1 + C_2\phi_g)}$).

If subpopulation and population totals are both of interest, then the optimal design derived in Section 2, or an approximation to it, should be used. This design also over-targets at the first stage and under-targets at the second stage. PSUs containing a relatively small proportion in the subpopulation (in our example, $\phi_g \leq 3\%$) should have a subsampling rate of 1 for non-subpopulation members found in the screen. That is, for low density PSUs, it is not worth using a two-phase sampling procedure.

We have assumed that the screening process can identify subpopulation membership without error. In practice, the initial identification may turn out to be wrong once the full interview is conducted. It has also been assumed that the PSU densities ϕ_g are without error for the whole population of PSUs. In practice these are likely to be based on census counts and will be somewhat out of date. Clark et al. (2008) found that the first issue was significant at least in one scenario, but that the second appeared to be less important. Future research should extend our designs to the case of imperfect screening.

Acknowledgements: This reseach was part of a Statistics New Zealand Official Statistics Research Fund project. The author thanks Mike Doherty, Robert Templeton, Sarah Gerritsen and Larry Hill for some stimulating discussions on this topic. Li-Chung Zhang reviewed Clark et al. (2008) and his comments improved this paper.

References

ABS. (2008). (Tech. Rep. No. 3238.0.55.001). Australian Bureau of Statistics. Available from `http://www.abs.gov.au/ausstats/abs@.nsf/mf/3238.0.55.001`

Bankier, M. D. (1988). Power allocations: determining sample sizes for subnational areas. *The American Statistician*, *42*(3), 174–177.

Bowlby, G., Denis, J., Langlet, E., & Malo, D. (2004). Aboriginal Data Initiative - Survey Component. In *Proceedings of Statistics Canada Symposium: Innovative Methods for Surveying Difficult to Reach Populations.* Statistics Canada Catalogue Number 11-522-XIE.

Clark, R., Forbes, A., Templeton, R., & Doherty, M. (2008). Sampling for subpopulations in household surveys with application to Māori and Pacific sampling. *The Official Statistics System*. (accepted for publication)

Clark, R., & Gerritsen, S. (2006). Sampling the Maori population in the

2006/2007 New Zealand Health Survey. In *Proceedings of Statistics Canada Symposium: Methodological Issues in Population Health.* Statistics Canada Catalogue Number 11-522-XIE.

Clark, R., & Steel, D. (2000). Optimum allocation to strata and stages with simple additional constraints. *Journal of the Royal Statistical Society Series D: The Statistician*, *49*, 197–207.

Clark, R., & Steel, D. (2007). Sampling within households in household surveys. *Journal of the Royal Statistical Society Series A*, *170*(1), 63–82.

Gray, A. (2005). Strategies for New Zealand household surveys which oversample Māori and Pasifika. In *Proceedings of the International Association for Official Statistics Meeting, Wellington New Zealand.* Available from `http://isi.cbs.nl/iaos/` (paper 16.2)

Hansen, M., Hurwitz, W., & Madow, W. (1953). *Sample Survey Methods and Theory Vol.1 and 2.* New York: Wiley.

Kalton, G., & Anderson, D. W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society Series A*, *149*(1), 65–82.

Lohr, S. L., & Rao, J. (2000). Inference from dual frame surveys. *Journal of the American Statistical Association*, *95*(449), 271–280.

McKenzie, D. J., & Mistiaen, J. (2008). Surveying migrant households: a comparison of census-based, snowball, and intercept point surveys. *Journal of the Royal Statistical Society Series A: Statistics in Society*. (Accepted for publication)

Sarndal, C., Swensson, B., & Wretman, J. (1992). *Model Assisted Survey Sampling.* New York: Springer-Verlag.

Simpson, L., & Akinwale, B. (2007). Quantifying stability and change in ethnic group. *Journal of Official Statistics*, *23*(2), 185–208.

Statistics Canada. (2005). *Projections of the Aboriginal Populations, Canada, Provinces and Territories 2001 to 2017* (Tech. Rep. No. 91-547-XIE). Statistics Canada. Available from `http://www.statcan.ca/english/freepub/91-547-XIE/91-547-XIE2005001.pdf`

Statistics New Zealand. (2008). *Māori Population Estimates Tables.* Available from `http://www.stats.govt.nz/tables/maori-popn-est-tables.htm`

Webster, A., Rogers, A., & Black, D. (2004). Surveying Aboriginal and Torres Strait Islander Peoples: Strategies and Methodologies of the Australian Bureau of Statistics. In *Proceedings of Statistics Canada Symposium: Innovative Methods for Surveying Difficult to Reach Populations.* Statistics Canada Catalogue Number 11-522-XIE.

Wells, E. J. (1998). Oversampling through households or other clusters: comparison of methods for weighting the oversampled elements. *Australian and New Zealand Journal of Statistics*, *40*(3), 269–277.

Wells, E. J. (2005). Targeting and screening as strategies for oversam-

pling (invited paper). In *Proceedings of the International Association for Official Statistics Meeting, Wellington New Zealand.* Available from `http://isi.cbs.nl/iaos/` (paper 16.3)

Appendix: Proof of Theorem 1

Stationary Point

Write $\theta_g = \pi_g n'_g$ and $\theta_{gB} = \pi_g n'_g f_{gB} = \theta_g f_{gB}$. Then C_E is linear in π_g, θ_g and θ_{gB} and F is linear in π_g^{-1}, θ_g^{-1} and θ_{gB}^{-1}. To be valid, solutions must satisfy $\theta_{gB} \leq \theta_g$ corresponding to $f_{gB} \leq 1$. (Other inequality constraints apply, but this turns out to be the most important because it is active in most cases, because this occurs whenever one or both of the first phase strata are fully enumerated in the second phase.) This is a special case of the Neymann allocation problem, and solutions must lie on either a stationary point or on a boundary. Stationary points are given by

$$\begin{aligned}
\pi_{g1} &= \lambda\sqrt{\frac{\rho N_g^2\left(P\phi_g^2+1-P\right)}{C_1}} = \lambda N_g\sqrt{\frac{\rho\left(P\phi_g^2+1-P\right)}{C_1}} \\
\theta_g &= \lambda\sqrt{\frac{N_g^2\phi_g\left(1-\rho\left(P\phi_g+1-P\right)\right)}{C_2+C_3\phi_g}} = \lambda N_g\sqrt{\frac{\phi_g\left(1-\rho\left(P\phi_g+1-P\right)\right)}{C_2+C_3\phi_g}} \\
\theta_{gB} &= \lambda\sqrt{\frac{(1-P)N_g^2(1-\rho)\left(1-\phi_g\right)}{C_3\left(1-\phi_g\right)}} = \lambda N_g\sqrt{\frac{(1-P)(1-\rho)}{C_3}}
\end{aligned}$$

(e.g. Clark & Steel, 2000) where λ is such that (7) is satisfied. By assumption, $\rho\phi_g << 1$ and so:

$$\theta_g \approx \lambda N_g\sqrt{\frac{\phi_g\left(1-\rho(1-P)\right)}{C_2+C_3\phi_g}}.$$

The corresponding values of n'_g and f_{gB} are

$$n'_g = \theta_g/\pi_g = \sqrt{\frac{\phi_g\left(1-\rho(1-P)\right)}{C_2+C_3\phi_g}\Big/\frac{\rho\left(P\phi_g^2+1-P\right)}{C_1}}$$

$$
\begin{aligned}
&= \sqrt{\frac{\phi_g\left(1-\rho(1-P)\right)}{\rho\left(P\phi_g^2+1-P\right)}\frac{C_1}{C_2+C_3\phi_g}} \\
f_{gB} &= \theta_{gB}/\theta_g = \sqrt{\frac{(1-P)(1-\rho)}{C_3}/\frac{\phi_g\left(1-\rho(1-P)\right)}{C_2+C_3\phi_g}} \\
&= \sqrt{\frac{(1-P)(1-\rho)}{(1-\rho(1-P))}\frac{C_2+C_3\phi_g}{C_3\phi_g}}
\end{aligned}
$$

Solution when Stationary Point is Invalid

The stationary point is only a valid solution if $f_{gB} \leq 1$. This is true when:

$$
\begin{aligned}
& f_{gB} \leq 1 \\
\Leftrightarrow\ & \frac{(1-P)(1-\rho)}{(1-\rho(1-P))}\frac{C_2+C_3\phi_g}{C_3\phi_g} \leq 1 \\
\Leftrightarrow\ & (1-P)(1-\rho)\left(C_2+C_3\phi_g\right) \leq \left(1-\rho(1-P)\right)C_3\phi_g \\
\Leftrightarrow\ & \phi_g C_3\left(1+P\rho-P-\rho\right)+C_2(1-P)(1-\rho) \leq \phi_g C_3\left(1-\rho+\rho P\right) \\
\Leftrightarrow\ & -\phi_g C_3 P + C_2(1-P)(1-\rho) \leq 0 \\
\Leftrightarrow\ & P\phi_g \geq \frac{C_2}{C_3}(1-P)(1-\rho)
\end{aligned}
$$

which is condition (9). The set of PSUs g satisfying (9) has been defined as U_I^a, with U_I^b containing other PSUs. For PSUs in U_I^b, the stationary point is not valid, so the optimum must be on the boundary $f_{gB} = 1$, or equivalently, $\theta_{gB} = \theta_g$. So F and C_E become:

$$
\begin{aligned}
F \approx\ & \rho\sum_{g\in U_I}\pi_g^{-1}N_g^2\left(P\phi_g^2+1-P\right)+\sum_{g\in U_I}\theta_g^{-1}N_g^2\phi_g\left\{1-\rho(1-P)\right\} \\
& +(1-P)(1-\rho)\sum_{g\in U_I}\theta_{gB}^{-1}N_g^2\left(1-\phi_g\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \rho \sum_{g\in U_I} \pi_g^{-1} N_g^2 \left(P\phi_g^2 + 1 - P\right) + \sum_{g\in U_I^a} \theta_g^{-1} N_g^2 \phi_g \left\{1-\rho(1-P)\right\} \\
&\quad + \sum_{g\in U_I^b} \theta_g^{-1} N_g^2 \left\{\phi_g \left(1-\rho(1-P)\right) + (1-P)(1-\rho)\left(1-\phi_g\right)\right\} \\
&\quad + (1-P)(1-\rho) \sum_{g\in U_I^a} \theta_{gB}^{-1} N_g^2 \left(1-\phi_g\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \rho \sum_{g\in U_I} \pi_g^{-1} N_g^2 \left(P\phi_g^2 + 1 - P\right) + \sum_{g\in U_I^a} \theta_g^{-1} N_g^2 \phi_g \left\{1-\rho(1-P)\right\} \\
&\quad + \sum_{g\in U_I^b} \theta_g^{-1} N_g^2 \left\{(1-P)(1-\rho) + P\phi_g\right\} \\
&\quad + (1-P)(1-\rho) \sum_{g\in U_I^a} \theta_{gB}^{-1} N_g^2 \left(1-\phi_g\right)
\end{aligned}
$$

$$
\begin{aligned}
C_E &= C_0 + C_1 \sum_{g\in U_I} \pi_g + \sum_{g\in U_I} \theta_g \left(C_2 + C_3\phi_g\right) + \sum_{g\in U_I} \theta_{gB} C_3 \left(1-\phi_g\right) \\
&= C_0 + C_1 \sum_{g\in U_I} \pi_g + \sum_{g\in U_I^a} \theta_g \left(C_2 + C_3\phi_g\right) + \sum_{g\in U_I^b} \theta_g \left(C_2 + C_3\phi_g + C_3\left(1-\phi_g\right)\right) \\
&\quad + \sum_{g\in U_I^a} \theta_{gB} C_3 \left(1-\phi_g\right) \\
&= C_0 + C_1 \sum_{g\in U_I} \pi_g + \sum_{g\in U_I^a} \theta_g \left(C_2 + C_3\phi_g\right) + \sum_{g\in U_I^b} \theta_g \left(C_2 + C_3\right) + \sum_{g\in U_I^a} \theta_{gB} C_3 \left(1-\phi_g\right)
\end{aligned}
$$

The optimal design is then

$$
\begin{aligned}
\pi_g &= \lambda \sqrt{\frac{\rho N_g^2 \left(P\phi_g^2 + 1 - P\right)}{C_1}} = \lambda N_g \sqrt{\frac{\rho \left(P\phi_g^2 + 1 - P\right)}{C_1}} \\
\theta_g &= \begin{cases} \lambda N_g \sqrt{\frac{\phi_g(1-\rho(1-P))}{C_2 + C_3\phi_g}} & \text{if} \quad g \in U_1^a \\ \lambda N_g \sqrt{\frac{(1-P)(1-\rho) + P\phi_g}{C_2 + C_3}} & \text{if} \quad g \in U_1^b \end{cases} \\
\theta_{gB} &= \lambda \sqrt{\frac{(1-P) N_g^2 (1-\rho)\left(1-\phi_g\right)}{C_3\left(1-\phi_g\right)}} = \lambda N_g \sqrt{\frac{(1-P)(1-\rho)}{C_3}}
\end{aligned}
$$

The corresponding values of n'_g and f_{gB} are

$$
n'_g = \theta_g / \pi_g
$$

$$
\begin{aligned}
&= \begin{cases} \sqrt{\frac{\phi_g(1-\rho(1-P))}{\rho\left(P\phi_g^2+1-P\right)}\frac{C_1}{C_2+C_3\phi_g}} & \text{if} \quad g \in U_1^a \\ \sqrt{\frac{(1-P)(1-\rho)+P\phi_g}{\rho\left(P\phi_g^2+1-P\right)}\frac{C_1}{C_2+C_3}} & \text{if} \quad g \in U_1^b \end{cases} \\
f_{gB} &= \theta_{gB}/\theta_g \\
&= \begin{cases} \sqrt{\frac{(1-P)(1-\rho)}{(1-\rho(1-P))}\frac{C_2+C_3\phi_g}{C_3\phi_g}} & \text{if} \quad g \in U_1^a \\ 1 & \text{if} \quad g \in U_1^b \end{cases}
\end{aligned}
$$
