Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

10-12

Sampling the Maori Population using Proxy Screening, the Electoral Roll and Disproportionate Sampling in the New Zealand Health Survey

Robert Graham Clark and Robert Templeton

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 the Māori Population using Proxy Screening, the Electoral Roll and Disproportionate Sampling in the New Zealand Health Survey

Robert Graham Clark, University of Wollongong, and
Robert Templeton, New Zealand Ministry of Health

1 INTRODUCTION

This chapter describes an instructive example of a hard-to-reach subpopulation: the indigenous Māori population of New Zealand. This population shares some characteristics with others described in earlier chapters: it is relatively rare, over-surveyed, and geographically dispersed, and there is no adequate population frame. There are some unique features as well: Māori are less rare than many indigenous populations, and have a special status in the NZ electoral system, so that the Electoral Roll provides a useful partial frame. A combination of strategies to oversample Māori in the NZ Health Survey is found to work well. A novel approach to setting the large number of design parameters required by this design is described, based on numerical optimization using a training and validation dataset.

The Māori peoples are the indigenous population of New Zealand (NZ), and as such are important for social, political and historical reasons. They have higher rates of poverty and illness than the general population and so are a particular priority in public health planning. For all these reasons, many surveys in NZ aim to over-sample Māori, to give more precise statistics than would be produced by an untargeted survey of the population.

Māori constitute 14% of NZ adults and achieving a higher sample proportion in a household survey requires a combination of imperfect strategies. There is no general population register which can be used as a sampling frame. The Electoral Roll gives a partial frame and electors may indicate Māori descent on the Roll. However, not all Māori choose to do so, and addresses on the Roll are out of date, particularly when an election is not imminent. There is a five yearly Census conducted by Statistics New Zealand, which can be used for area targeting, although Māori are reasonably dispersed across NZ (particularly across the North Island), and populations shift between censuses, particularly at fine area levels. Proxy screening of households is another option, but this tends to under-identify Māori.

All of these strategies have been used in the NZ Health Survey, a multi-stage household interview survey collecting information on health behaviors, use of health services, and current health status. Since May 2011, the survey is conducted continuously with an annual sample size of approximately 14,000 adults and 5,000 children. Prior to this it was run roughly three-yearly. A major goal is to provide accurate and precise statistics on ethnic subpopulations, particularly the Māori and Pacific peoples. This chapter describes the strengths and weaknesses of three tools for sampling the Māori population in the NZ Health Survey.

Section 2 gives some background on the Māori population. Section 3 discusses proxy screening. This was used in the 2006/2007 NZ Health Survey, including a subsample where the proxy information was collected but not used, enabling evaluation of the proxy data against more rigorously collected survey data. Section 4 outlines the use of disproportionate sampling by area based on Census data on Māori and other subpopulations. Section 5 discusses the Electoral Roll, which has been used since May 2011 to sample addresses apparently containing an enrollee with Māori descent, in conjunction with area sampling. Section 6 shows how these contrasting strategies can be combined to reflect the errors and uncertainties attached to each, by using separate training and validation datasets to develop the design. Section 7 is a summary.

2 THE MĀORI POPULATION

Māori are the indigenous people of New Zealand. There were over 560,000 people who identified as belonging to the Māori ethnic group in the 2006 Census of the New Zealand population, representing 15% of the total population.

Māori are a population of particular interest in New Zealand because of the Government's special obligations to them under the Treaty of Waitangi. The Treaty of Waitangi was an agreement entered into by representatives of the Crown and of Māori in 1840, which established British authority in New Zealand (later transferred to the New Zealand Parliament) and which guaranteed Māori full protection of their interests and status, and full citizenship rights. As part of the obligation of the New Zealand Government stemming from this agreement is the need to collect good quality statistical information to inform Māori development and decision-making, and to monitor the effects of government policies and programmes on Māori.

In addition to treaty obligations, understanding the Maori population is important because they are a large group within the New Zealand population with a distinct demographic and social profile. The Māori population is a youthful and growing population. Although there will be more older Māori (as a proportion) in coming years, Māori will continue to have a relatively young population. Fertility rates for Māori women are higher than those for non-Māori, and well above replacement level, contributing to the growing Māori population

In addition to the youthful demographic profile, there exist substantial inequities in terms of outcomes for Māori across areas such as:

- *the social determinants of health*: e.g. education, employment, income, and housing,
- *health risk behaviours*: e.g. tobacco use, nutrition, gambling problems and patterns of alcohol use and
- *long term health conditions*: e.g. diabetes, heart disease and cancer.

Hence monitoring Māori outcomes across a range of social measures is critical to most official social and health surveys undertaken in New Zealand because of both historical obligations but also because Māori are a large distinct sub-population of New Zealand, where substantial disparities exist across a range of social and health outcomes compared to Non-Māori.

There is no comprehensive population register in New Zealand and consequently it has been standard to employ area sampling practises when conducting official social surveys in New Zealand. One reason behind this is the relatively high mobility of the New Zealand population, 55% for the total population (among those aged over five years) changed their place of residence between the 2001 and 2006 Censuses. This is even higher for Maori with sixty percent having moved between those Censuses. This mobility makes constructing and maintaining population registers a challenge.

The pre-eminence of area based sampling means the geographic distribution of Māori is important in terms of understanding some of the special issues involved in collecting Māori social and health statistics. Although the majority of Māori live in the North Island of New Zealand (87%) and in urban areas, the key is: that Māori are relatively well spread across all parts of the country. In fact 82% of Maori live in areas (meshblocks) where they are a minority (a meshblock is a standard geographic unit used in the New Zealand Census and in household surveys.). This makes surveying them directly them in an area-based face-to-face approach more costly because dwellings without Māori living in them can only be excluded after a cost-incurring doorstep screening exercise.

There are, however, some population list resources that can be useful, including the electoral roll. As well as being eligible for the general electoral roll, Māori have the option of voting in one of 7 Māori electorates. Hence the electoral registering process includes a declaration of one's ethnic ancestry and this information is stored, regardless of whether the person opts to be on the Māori roll or the General roll. An electronic version of the roll is available to those doing scientific and health related research. The electoral roll on its own, covers about 80% of the Māori population. It is not always completely up-to-date, with more push for greater coverage and up-to-date information in the lead up to elections. For these reasons the roll on its own is not considered an ideal frame. An approach, described later, is to combine the roll information with an area frame approach, to create an effective sampling frame with good coverage properties.

3 PROXY SCREENING FOR MĀORI AND OTHER SUBPOPULATIONS

The 06/07 NZ Health Survey used a proxy screening tool to oversample Māori, Pacific and Asian people. The first stage of selection was a sample of meshblocks (small areas containing on average about 100 people), stratified by District Health Board (21 broad regions) based on Census data on ethnic and total population sizes. The second stage of selection was of dwellings within meshblocks. This was divided into two parts: a core and a booster sample. In the core sample, one adult and one child

(if any) was selected from each household. In the booster sample, screening questions were first asked of any adult contact in the household, regarding the number of adults and children in the household and their ethnicities (Māori, Pacific, Asian or other, with multiple identification possible). One eligible adult and one eligible child (if any) was then selected, with eligible meaning Māori, Pacific or Asian according to the proxy screener. The final sample consisted of adults and children selected via either the core or booster avenues.

To enable probabilities of selection in the pooled sample to be calculated, the proxy screening questions were also asked of the core dwellings. For more information on this question of the use of a screening tool on a core and a booster sample, see also Wells (1998). The use of both a core and booster sample means that under-identification in the proxy screener does not result in bias, only in increased standard errors, because all people have a chance of selection in the core sample (and hence in the pooled sample), and the probability of selection in the pooled sample can be calculated for each respondent.

For respondents in the core sample, we have both the survey report of ethnicity and the proxy screener results. Thus we can identify how many Maori, Pacific and Asian people were missed by the screener. This is useful to assess the efficiency of the booster sample. It also allows us to evaluate the undercoverage that would result if we were to omit the core sample and rely wholly on the screening tool to survey Māori or other ethnicities.

Table 1 shows the number of adults cross-tabulated by their screener and their survey identification as Māori / non-Māori. The major discrepancy between the survey and screener is that 20.5% of Māori (according to the survey) fail to be identified in the screener. There is very little over-identification of Māori in the screener.

Table 1: Screener and Survey Classification of Maori Status for Adults in the Core Sample

unweighted count *proportion within screener result* *proportion within survey result*		Survey Result (Gold Standard)	
		non-Māori	Māori
Proxy Screener Result	non-Māori	7747 96.8% 98.9%	256 3.2% 20.5%
	Māori	84 7.8% 1.1%	992 92.2% 79.5%

Table 2 shows the rates of under-identification of Maori, Pacific and Asian adults, broken down by single adult vs multi-adult households. Weighted rates, reflecting the unequal probability nature of the sample design, are shown in brackets. It is

clear that the screener does worse at identifying Māori than Pacific or Asian adults, with misclassification rates over 20%, as opposed to 10-13%. Surprisingly, the identification of Māori is nearly as poor for single adult households as for multi-adult households. Clearly it is not proxy reporting of ethnicity that is the problem, given that 18% of Māori adults living alone do not identify as Māori in the screener. The situation is different for Pacific and Asian respondents, who are correctly identified much more often in single adult households.

No definite explanation of this under-reporting of Māori in the screener has been identified. Perhaps some respondents correctly intuit that responding as Māori in the screener may increase their chances of being selected for the main survey, because of the fact that Māori are an over-surveyed group. Or the fact that the first contact involves a very brief ethnicity question may be off-putting to Māori respondents.

Table 2: Unweighted (weighted) Under-Identification Rates (%) of Proxy Screener by Type of Household

	Single Adult Households	Multiple Adult Households	All Households
Māori [a]	17.8 (18.3)	21.5 (22.1)	20.5 (21.7)
Māori [b]	17.5 (18.0)	20.9 (21.5)	20.0 (21.1)
Pacific [c]	14.9 (15.4)	8.6 (7.6)	9.8 (0.0)
Asian [d]	17.5 (16.5)	11.7 (10.6)	12.5 (10.9)
Māori, Pacific or Asian [e]	17.0 (17.2)	16.3 (15.3)	16.4 (15.4)

a: proportion of respondents where proxy screener indicates non-Māori, but survey indicates Māori.
b: proportion of respondents where proxy screener indicates non-eligible (not Māori, Pacific or Asian), but respondent reports as Māori in the survey
c: proportion of respondents where proxy screener indicates non-eligible (not Māori, Pacific or Asian), but respondent reports as Pacific in the survey
d: proportion of respondents where proxy screener indicates non-eligible (not Māori, Pacific or Asian), but respondent reports as Asian in the survey
e: proportion of respondents where proxy screener indicates non-eligible (not Māori, Pacific or Asian), but survey shows otherwise

Suppose a survey is to be conducted of Māori only, using the screener to enhance the design. What are the consequences of about 20% of this population being missed in the screener?

Firstly, if we were to only apply the full survey when the screener indicated a Māori respondent, we would under-cover the full population by about 20%. The covered sub-population of Māori are apparently slightly less healthy than the full Māori adult population, with obesity and smoking rates a few percentage points higher. They are also more concentrated in the most deprived quintile of meshblocks in NZ (42.6% vs 38.9%).

Secondly, a two-phase design could be used, where some adults would be sampled even when the screener indicated they are non-Māori. To enable a simple rough evaluation, suppose that we can take a simple random of adults, rather than using a complex multi-stage design. Further suppose the cost of applying the screener is 0.3 times of the cost of applying the subsequent full interview (this was confirmed as broadly reasonable by the survey company conducted the 06/07 survey). This abjectly fails Deming's (1977) rule, also quoted in Kalton and Anderson (1986), that the ratio of second to first phase costs needs to be at least 6:1 and preferably 40:1 or more. However, application of the approximate formulas (1) and (2) on page 70 of Kalton and Anderson (1986) suggests that the screener still provides useful gains. Using the proportions from Table 22.1 and assuming that the cost of applying the full survey is the same for Māori and non-Māori, we have, in the notation of Kalton and Anderson, P=0.14, P_1=0.92, W_1=0.88 and c=1. Formula (2) tells us the best relative rate of sampling of adults who screen as Māori to people who do not: k=5.2. Formula (1) then gives the variance for estimating means for the Māori population relative to equal probability sampling: R=0.55. However, we really want to know the efficiency relative to equal probability sampling where the screener is not even applied. This means that the per-respondent cost becomes 1.3 times cheaper, so the relevant relative efficiency is 0.55*1.3 which equals 0.71. Thus, the use of the screening tool improves the cost-efficiency of the survey by 29%, which is not too bad.

In summary, household screening for Māori resulted in a substantial undercount (21%) and small overcount (1%), even for single adult households. This was less accurate than screening for other ethnicities, perhaps because Māori are over-surveyed. At least some fraction of adults not screening as Māori should still be surveyed, to avoid undercoverage bias. The use of the screener can improve the variances of estimated population means for Māori by a ballpark 30%, in surveys where Māori are the only priority.

4 DISPROPORTIONATE SAMPLING BY AREA

Statistics New Zealand conducts a five yearly Census which includes an ethnicity question. Population sizes by Māori and other ethnicities are available for each meshblock. It makes sense to use this data to improve the sample size of Māori, by assigning higher selection probability to areas where more Māori live. Disproportionate sampling can increase the achieved sample size of Māori for fixed cost. However, the unequal selection probabilities need to be corrected for by appropriate use of survey weights, otherwise Māori statistics would over-represent Māori living in higher density areas (such as Auckland) at the expense of the substantial number of Māori living in other parts of New Zealand. The resultant increased variation in survey weights tends to lead to higher standard errors, partially undoing the benefit of disproportionate sampling.

Kalton and Anderson (1986) derived results on the best allocation of sample sizes to strata, to optimally balance the sample size of a subpopulation and the variability of the estimation weights. Assuming that the only aim is to estimate means for the

subpopulation, and also assuming simple one-stage stratified sampling, as well as some simplifying assumptions, they found that the probability of selection in stratum h should be proportional to $\sqrt{\varphi_h /(R+\varphi_h)}$ where φ_h is density (i.e. the proportion of the population in stratum h who belong to the subpopulation), and *R* is the cost of identifying whether a sampled unit is in the subpopulation relative to the cost of fully surveying the unit.

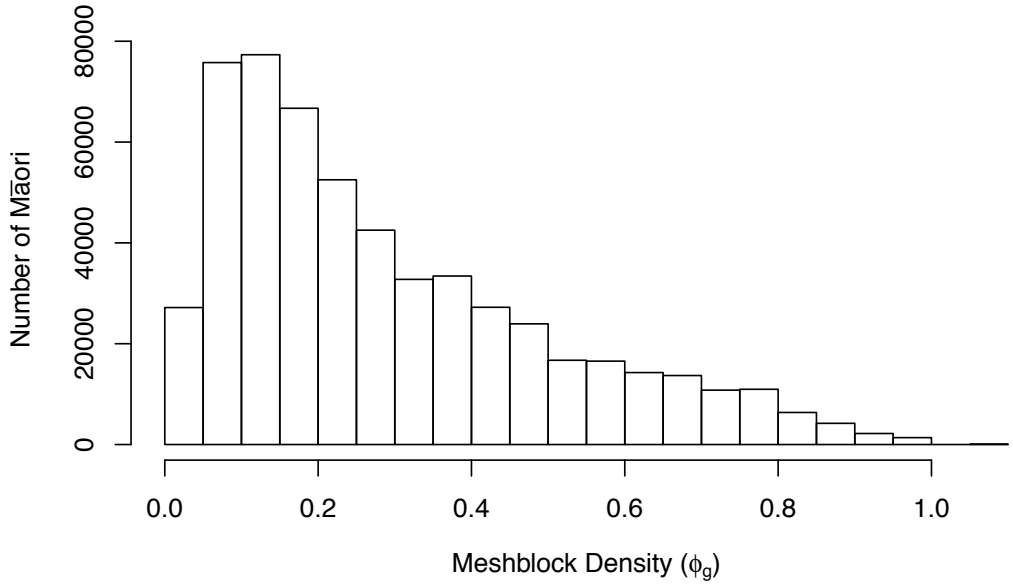
Densities are available for all meshblocks in NZ, so it makes sense to use this information in sampling. However, there are too many meshblocks (approximately 40,000) for them to be feasible strata. Instead, the NZ Health Survey uses multi-stage sampling, with meshblock as the primary sampling unit, followed by households, followed by one adult and one child per selected household. Clark (2009) extended Kalton and Anderson's optimal allocation to multi-stage sampling. Under simplifying assumptions, the best design is to assign a final person probability of selection proportional to $\sqrt{\varphi_g /(R+\varphi_g)}$ where φ_g is the density for meshblock g. In single-stage stratified sampling, the probabilities of selection fully specify the design. In multi-stage sampling, these probabilities of selection could be implemented by assigning higher selection probabilities to high density meshblocks, or by assigning higher sampling fractions within these meshblocks, or by a combination. Clark (2009) found that in most cases, an approximately optimal strategy is to

- select meshblocks with probability proportional to $N_g\sqrt{\varphi_g /(R+\varphi_g)}$, where N_g is the total population for meshblock g; and to
- use a fixed sample size (including both Māori and non- Māori) within each selected meshblock.

This is a modification of a standard self-weighting design. If the subpopulation is relatively rare, then $N_g\sqrt{\varphi_g /(R+\varphi_g)}$ may be replaced by $N_g\sqrt{\varphi_g}$.

The preceding strategies will only be of much use if Māori are geographically clustered. If the density varies little across meshblock, then the above design will be close to equal probability sampling. Figure 1 is a histogram showing the number of Māori living in meshblocks with various densities. The figure shows that there is some but not dramatic concentration of Māori people in meshblocks. Similar calculations show that the median value of the meshblock density across all Māori in NZ is 23%, and only 17% of Māori live in meshblocks where they are the majority.

Figure 1: Histogram of Māori Meshblock Densities (φ_g)



The optimal designs of Kalton and Anderson (1986) and Clark (2009) also assume that design data is perfectly accurate. In reality, census data will be out of date to some extent. Changes between census dates are likely to be greater at the meshblock level than for broader regions. For this reason, the 2006/2007 NZ Health Survey design calculated probabilities of selection using densities defined for each District Health Board, a broad region containing on average about 200,000 people.

Table 3 shows the relative efficiency achieved by using a design based on Clark (2009) with density defined at various geographic levels. The table was calculated using matched meshblock data from the 2001 and 2006 NZ Censuses. The first column shows the efficiency based on using 2001 NZ Census data calculated under the assumption that this data is perfectly accurate at the time of surveying. The second column shows the efficiency of a design based on 2001 data if the actual meshblock counts were as in the 2006 Census. The first column shows that the best design when the Census data is perfectly accurate uses meshblock level data, giving a reduction in variance of 23% (efficiency of 0.77) compared to equal probability sampling, for fixed cost. However, the second column shows that when the efficiency is more realistically evaluated using 2006 Census data, the efficiency deteriorates to 0.88. In fact, allocating probabilities of selection based on the area unit densities is slightly superior.

Table 3: Approximate Relative Efficiency (compared to equal probability sampling) of Various Designs Meshblock Selection Probability Proportional to $N_g\sqrt{\hat{\varphi}_g}$ where φ_g estimated using 2001 Census Data at Various Levels of Aggregation

Level of Aggregation of 2001 Census Data	Average Total Population (all ages) per area (2001 Census)	Efficiency Estimated using 2001 Census Data	Efficiency Estimated using 2006 Census Data
Meshblock[1]	110	0.766	0.876
Area Unit	2,200	0.857	0.870
Territorial Authority	53,000	0.921	0.923
District Health Board	180,000	0.937	0.937

1. 0.01 added to $\hat{\varphi}_g$ to avoid assigning zero probability of selection to any meshblocks.

5 USING THE ELECTORAL ROLL

Screening for Māori respondents on the doorstep is expensive and inefficient, as discussed in Section 3. In many countries, there is little other choice, because there are no adequate frames of ethnic and other subpopulations. However, New Zealand Māori have the opportunity to indicate their descent, and can choose whether to vote in a general electorate or a Māori electorate. As a result, the Electoral Roll constitutes a partial frame of Māori.

Figure 2 is an extract from the first page of the NZ voting enrolment form. A question just before Section B of the form is "are you a NZ Māori or a descendant of a NZ Māori?" The list of enrollees answering yes to this question has been used as a supplementary sampling frame of Māori in the NZ Health Survey since April 2012. This frame is certainly not perfect, in particular:

i. Address and other information may be somewhat out of date, particularly when there has not been an election recently.
ii. Self-identification of Māori may be different on the Roll than in the survey, leading to over or under-coverage.
iii. The survey aims to cover the whole population, so non-Māori also need to be given appropriate chance of selection in the survey.

To deal with these issues, the 2011-2014 Health Survey sample is selected as follows:

- A sample of meshblocks, called the "area component", is selected from the whole of New Zealand, with probabilities of selection based on 2006 Census total population and population by ethnicity. 215 meshblocks were selected in this way for every quarter of enumeration.
- Addresses where at least one person indicated Māori descent were selected from the Electoral Roll. These addresses were grouped into meshblocks. A sample of meshblocks, called the "roll component" was selected with probability proportional to the number of addresses. This sample was forced

to be non-overlapping with the area component. 100 meshblocks were selected in this way for every quarter.

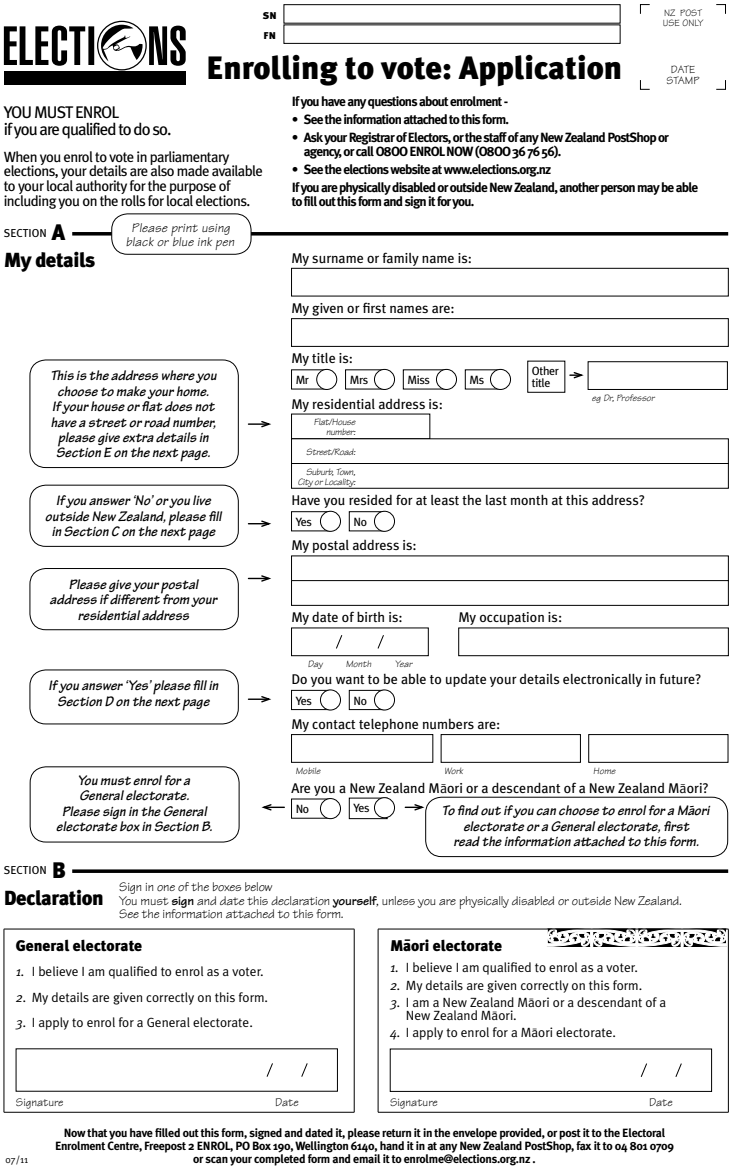
- A sample of households was selected in the area component by taking a systematic sample from each selected meshblock in this component. One adult and one child was selected and surveyed from each selected household.
- A sample of addresses was selected from each meshblock in the roll component. One adult and one child was selected and surveyed from each selected address, without reference to their ethnicity. Electoral Roll data was not used at this stage – the selected adult was not necessarily the same as the Māori enrollee, and the Māori enrollee might even have moved. This was done to enable a probability of selection to be calculated for every respondent, to avoid (for example) under-representation of people who have moved without updating their electoral enrolment.
- Ethnicity (including Māori identification) was collected as part of the survey, and was used in calculating statistics for Māori and other subpopulations.

The sample contained 215 meshblocks in the area component and 100 in the roll component in each quarter of enumeration. Approximately 14% of approached households were selected via the Roll component. The next section will discuss how the relative sizes of the area and roll components, as well as the method of disproportionate sampling, were chosen.

It turned out that 52% of adult respondents in the roll component were Māori, compared to only 14% of adults in the whole population (estimated using data from September quarter 2011). It is clear that the Roll is able to substantially increase the rate of Māori in sample, without incurring screening costs. The Roll as used here is far from perfect though, with nearly half of all selected adults being non-Māori. Weighted estimates showed that approximately 68% of the adult Māori population live in Māori addresses (i.e. addressed where at least one enrollee has indicated Māori descent).

To further clarify the potential gain from the use of the Māori Roll, suppose that we were to stratify the population of all adults by Māori vs non-Māori address, and then select an optimally allocated sample. The relative efficiency for Māori statistics would then be 0.73, compared to equal probability sampling, i.e. a 27% gain. While we might hope for greater gains in efficiency, this is a much greater gain than from either disproportionate sampling by area or from proxy screening.

Figure 2: First Page of NZ Voting Enrolment Form



ELECTIONS

SN
FN

NZ POST USE ONLY

DATE STAMP

Enrolling to vote: Application

YOU MUST ENROL
if you are qualified to do so.

When you enrol to vote in parliamentary elections, your details are also made available to your local authority for the purpose of including you on the rolls for local elections.

If you have any questions about enrolment -

- **See the information attached to this form.**
- **Ask your Registrar of Electors, or the staff of any New Zealand PostShop or agency, or call 0800 ENROL NOW (0800 36 76 56).**
- **See the elections website at www.elections.org.nz**

If you are physically disabled or outside New Zealand, another person may be able to fill out this form and sign it for you.

SECTION **A**

Please print using black or blue ink pen

My details

My surname or family name is:

My given or first names are:

My title is:

Mr ○ Mrs ○ Miss ○ Ms ○ Other title →

eg Dr, Professor

This is the address where you choose to make your home. If your house or flat does not have a street or road number, please give extra details in Section E on the next page. →

My residential address is:

Flat/House number:

Street/Road:

Suburb, Town, City or Locality:

If you answer 'No' or you live outside New Zealand, please fill in Section C on the next page →

Have you resided for at least the last month at this address?

Yes ○ No ○

Please give your postal address if different from your residential address →

My postal address is:

My date of birth is: / /

Day Month Year

My occupation is:

If you answer 'Yes' please fill in Section D on the next page →

Do you want to be able to update your details electronically in future?

Yes ○ No ○

My contact telephone numbers are:

Mobile *Work* *Home*

You must enrol for a General electorate. Please sign in the General electorate box in Section B. ←

Are you a New Zealand Māori or a descendant of a New Zealand Māori?

No ○ Yes ○ →

To find out if you can choose to enrol for a Māori electorate or a General electorate, first read the information attached to this form.

SECTION **B**

Declaration

Sign in one of the boxes below
*You must **sign** and date this declaration **yourself**, unless you are physically disabled or outside New Zealand.*
See the information attached to this form.

General electorate

1. I believe I am qualified to enrol as a voter.
2. My details are given correctly on this form.
3. I apply to enrol for a General electorate.

/ /

Signature Date

Māori electorate

1. I believe I am qualified to enrol as a voter.
2. My details are given correctly on this form.
3. I am a New Zealand Māori or a descendant of a New Zealand Māori.
4. I apply to enrol for a Māori electorate.

/ /

Signature Date

Now that you have filled out this form, signed and dated it, please return it in the envelope provided, or post it to the Electoral Enrolment Centre, Freepost 2 ENROL, PO Box 190, Wellington 6140, hand it in at any New Zealand PostShop, fax it to 04 801 0709 or scan your completed form and email it to enrolme@elections.org.nz .

07/11 ROE 1

6 COMBINING SAMPLING STRATEGIES EFFICIENTLY

6.1 Overview

The previous three sections cover three strategies for sampling the Māori population. It is not at all clear how to put these together in practice, in a way that reflects the out-of-datedness and different ethnicity definitions of the census area data, the undercoverage and overcoverage of the electoral roll, and the under-identification of Māori by the proxy screener. This section will describe a methodology for simultaneously making these and other design decisions while reflecting the imperfections of the design information. The approach was used to design the 2011-2014 NZ Health Survey sample.

Firstly, we will express the design in terms of 12 design parameters that need to be set. Then we will discuss how to estimate the variances that will result from any given set of values for these parameters, using 2001 Census data and 2006/2007 NZ Health Survey data. Finally, the design parameters are numerically optimized.

6.2 Expressing the Design in Terms of 15 Design Parameters

Section 4 suggested that to optimize for a given subpopulation, meshblock probabilities of selection should be proportional to $N_g\sqrt{\varphi_g}$, with a fixed number of households to be selected from each selected meshblock. This means that household probabilities of selection are proportional to $\sqrt{\varphi_g}$. However, this design is approximately optimal when a single subpopulation is of interest, and when the densities φ_g are known perfectly. In reality, in the NZ Health Survey, the Māori, Pacific and Asian populations are all important, although Māori statistics are given the highest priority. National all-ethnicity estimates are also important. Moreover, the densities are not known perfectly. Table 3 showed that Census densities at the broader area unit level appear to give better results than meshblock densities, when 5-year-old Census data is used. Even better results might be achievable by using an appropriate mix of meshblock, area unit and district health board densities.

Such a mix can be given by making meshblock probabilities of selection in the area component of the sample proportional to the population size N_g multiplied by a targeting factor f_g:

$$
\begin{aligned}
f_g = {} & w_1\sqrt{\text{Maori MB density}} + w_2\sqrt{\text{Maori AU density}} + w_3\sqrt{\text{Maori DHB density}} \\
& + w_4\sqrt{\text{Pacific MB density}} + w_5\sqrt{\text{Pacific AU density}} + w_6\sqrt{\text{Pacific DHB density}} \\
& + w_7\sqrt{\text{Asian MB density}} + w_8\sqrt{\text{Asian AU density}} + w_9\sqrt{\text{Asian DHB density}} \\
& + w_{10}\times 1
\end{aligned}
\qquad (1)
$$

where the parameters w_1, …, w_{10} are non-negative weights which sum to 1. The final parameter w_{10} is there to make sure that no probabilities of selection are too close to zero in meshblocks with few subpopulation members. A fixed sample size within meshblocks of 20 was assumed. (The within-meshblock sample size could have been treated as another design parameter to be optimized, but for simplicity was set independently. Intra-class correlations for many health variables are low, and the value of 20 was chosen mainly to be less than the meshblock size for the great majority of meshblocks.)

A further parameter, p_{screen}, is needed to define the use of the proxy screener for ethnicity. It is assumed that the proxy screener is applied to the 20 selected households in each meshblock in the area component. Of these households, $20p_{screen}$ are defined to be the booster households, and one adult and one child of eligible ethnicity according to the screener (Māori, Pacific or Asian) (if any) is selected. The remaining $20(1-p_{screen})$ households are defined to be core households. One adult and one child is selected from each household regardless of their screening results.

Finally, a parameter p_{roll} defines the use of the Electoral Roll. The complete sample consists of a roll component and an area component. The roll component is selected as described in section 5, such that a proportion p_{roll} of the combined sample is in the roll component, with the remainder coming from the area component.

6.3 Estimating the Variance for any Given Set of Values for the Design Parameters

To choose values for the 15 design parameters, we want to be able to estimate the variances that would be achieved for any given set of values. The estimation should reflect that the 2006 Census data is 5 years old when the continuous survey commences in 2011. It should also reflect the imperfections of the Electoral Roll and the proxy screening tool.

To achieve this, for any given set of design parameters, a hypothetical design will be constructed using the Electoral Roll and the 2001 Census data. The probabilities of selection for this design will then be calculated for every respondent in the 2006/2007 NZ Health Survey. This sample data will then be used to estimate the variance that will be achieved for estimates for the total population, and the Māori, Pacific and Asian sub-populations. The discrepancy between ethnicity as recorded by the 06/07 survey and the design data from the 2001 Census and other sources will enable a realistic assessment of any set of design parameters.

We want an estimator for the variance that will be achieved from the hypothetical new design of an estimated prevalence for Māori and other subpopulations. A commonly used approximation is:

$$
\begin{aligned}
\hat{\mathrm{var}}\left(\hat{P}_{sub}\right) &\approx P(1-P)n^{-1}_{sub(hypothetical)}\left(1+c^2_{w(hypothetical)}\right) \\
&= P(1-P)n^{-1}_{sub(hypothetical)}\frac{\sum_{s_{sub\,(hypothetical)}}\pi_i^{-2}/n_{sub(hypothetical)}}{\left(\sum_{s_{sub\,(hypothetical)}}\pi_i^{-1}/n_{sub(hypothetical)}\right)^2}. \qquad (2) \\
&= P(1-P)\left(\sum_{s_{sub\,(hypothetical)}}\pi_i^{-2}\right)\left(\sum_{s_{sub\,(hypothetical)}}\pi_i^{-1}\right)^{-2}
\end{aligned}
$$

where P is the population prevalence. See for example Kish(1992) and Gabler et al (…). We can't apply (2) as is, because the hypothetical sample has not actually been selected. Instead, the 06/07 survey data is used. We can calculate the probability of selection π_i in the hypothetical sample for every unit in the 06/07, using the 2001 Census to obtain densities by ethnicity which lead to the meshblock selection probabilities via (1), and also using an Electoral Roll extract from 2006. The 06/07 survey data file contains a weight w_i which reflects the design used to select this sample. We can then estimate the right hand side of (2) by replacing sums over $s_{sub(hypothetical)}$ with sums over $s_{sub(06/07)}$ weighted by $w_i\pi_i$ to reflect the difference between the 06/07 design and the hypothetical design:

(3) $$\hat{\text{var}}\left(\hat{P}_{sub}\right) = P(1-P)\left(\sum_{s_{sub}(06/07)} w_i \pi_i^{-1}\right)\left(\sum_{s_{sub}(06/07)} w_i\right)^{-2}.$$

The variance estimator (3) uses the 06/07 sample data to estimate the variance that will be achieved by the hypothetical new design defined by any given set of design parameters.

The crucial feature of estimator (3) is that the design data used to calculate π_i in (3) is based on out-of-date Census and roll information, as well as the somewhat error-prone proxy screening data from the 06/07 survey. In contrast, the ethnicity used to define the subpopulation sample $s_{sub(06/07)}$ is based on the gold-standard survey-collected ethnicity. Hypothetical designs will be penalized to some extent when there are differences between the survey ethnicity and the design ethnicity.

6.4 Optimising the Design Parameters

The objective criterion for the survey was defined to be

$$F = SE\left(\hat{P}_{Maori}\right) + SE\left(\hat{P}_{Pacific}\right) + SE\left(\hat{P}_{Asian}\right)$$

where $\hat{P}_{subpop}$ is an estimated prevalence for a given subpopulation, with the true prevalences assumed to equal 0.2. The standard errors were estimated as described in 6.3. This objective criterion was defined in consultation with the Ministry of Health, by tabulating estimates standard errors for designs based various weighted standard error criteria. Ministry staff chose a criterion reflecting their priorities for Māori, Pacific, Asian and national statistics. Standard errors for national prevalences were given no weight in the final criteria, because national standard errors were considered to be low enough even without being explicitly reflected in F.

The estimated objective criterion F was then coded as a function of w_1, …, w_{10} , p_{roll} and p_{screen} in the R statistical software environment (R Development Core Team, 2012). This function was then minimized using the *optim* function in R. Table 4 summarises the optimal designs. Each option shows the result of an optimization with some or no design parameters constrained to equal 0. Option 1 is equal probability sampling and is included for comparison purposes. Option 2 is the unconstrained optimal design. Option 3 constrains p_{screen} to equal 0. Ministry of Health felt that the screener could give a poor first impression to respondents, and the table shows that the objective criteria F is not too much worse when it is omitted. The final option, 4, sets various other parameters to zero for simplicity, based on those parameters which were close to 0 in Option 3. An option similar to this one was implemented in 2011.

Some notable features from the designs in Table 4:

- The optimal design gives almost no weight to Māori densities in the area targeting. The optimization process revealed that the Roll should be used instead to oversample this population – this would not have been apparent otherwise.

- Approximately 14% of the total sample should be selected using the Roll, with the remaining 86% coming from the area component. Perhaps not coincidentally, 14% is also the proportion of the adult population who are Māori.
- Omitting the screen increases the Māori SE by 4% (from 0.97% to 1.01%) and increases F by 9% (0.77 to 0.84), while reducing the national SE. This was considered to be an acceptable price to improve the initial contact process.
- Options 2, 3 and 4 give some weight to MB densities, but more to AU densities, and almost no weight to the DHB data.

Typical approaches to sample design would have based area targeting on meshblock densities only, since this would be optimal if the Census design data was perfectly accurate and up to date. The approach described in this section allows the limitations of the design data to be taken account of, so that broader level area data are used in combination with meshblock data for a more robust design. To borrow some terminology from the machine learning literature (e.g. Hastie et al 2009, chapter 7), the objective criteria (3) makes use of separate training and validation datasets. The design probabilities π_i are optimal according to the 2001 Census data (the training dataset), but are instead evaluated in (3) using the 2006/2007 survey data (the validation dataset). If these two datasets agreed on ethnicity, the numerically optimized design would have weight attached to meshblock densities but none attached to area unit and district health board densities. The discrepancies between the training and validation datasets enable a realistic evaluation of the strengths and weaknesses of the design data, and the numerically optimized design reflects this. The approach here is based on a general statistical learning methodology for sample design developed in Clark (2012).

Table 4: Numerically Optimized Designs (all cost equivalent assuming 1 cost unit for each full interview and 0.3 for each household contact)

Design Parameter	Interpretation	Option 1: Equal Probability Sample	Option 2: Unconstrained Optimal Design	Option 3: No Screener	Option 4: Simplified Design with Selected Design Parameters Zeroed
w_1	Māori MB weight	**0**	0.00	0.01	**0.00**
w_2	Māori AU weight	**0**	0.05	0.03	**0.00**
w_3	Māori DHB weight	**0**	0.00	0.00	**0.00**
w_4	Pacific MB weight	**0**	0.33	0.29	0.31
w_5	Pacific AU weight	**0**	0.26	0.34	0.37
w_6	Pacific DHB weight	**0**	0.01	0.01	**0.00**
w_7	Asian MB weight	**0**	0.05	0.11	0.09
w_8	Asian AU weight	**0**	0.23	0.18	0.20
w_9	Asian DHB weight	**0**	0.03	0.00	**0.00**
w_{10}	weight attached to "1"	**1**	0.05	0.02	0.03
p_{roll}	proportion of total households selected via Roll sample	**0**	0.09	0.14	0.14
p_{screen}	proportion of area sample households where screen is applied	**0**	0.61	**0.00**	**0.00**
Properties of Optimal Design					
SE (%) Māori		1.18	0.97	1.01	1.01
SE (%) Pacific		1.90	1.32	1.46	1.46
SE (%) Asian		1.41	1.17	1.31	1.31
SE (%) National		0.41	0.55	0.47	0.47
F relative to option 1		1.00	0.77	0.84	0.84

7 SUMMARY

A number of tools have been used to improve Māori statistics in the NZ Health Survey. Screening based on proxy household ethnicity was applied in the 2006/2007 survey, but abandoned in the 2011-2013 design, because of under-identification of about 20% of Māori, and because of the potential to give a poor impression of the survey. Unequal probability sampling of areas can improve the precision of Māori statistics as well as other populations of interest, such as Pacific and Asian people. The improvements are worthwhile but modest, due to datedness of the census data and the dispersion of the populations of interest. The use of the electoral roll in a dual frame design gave the most significant improvements in the efficient sampling of the Māori population. A statistical learning methodology was effective in simultaneously optimizing a large number of design parameters while reflecting the imperfections of the design data.

REFERENCES

Clark, R. G. (2009). Sampling of subpopulations in two-stage surveys. Statistics in Medicine, 28 (29), 3697-3717.

Clark (2012). Statistical Learning in Sample Design. University of Wollongong Centre for Statistical and Survey Methodology Working Paper. www.cssm.uow.edu.au .

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer. Available from http://www-stat.stanford.edu/~tibs/ElemStatLearn/

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

Ministry of Health (2011). The New Zealand Health Survey Sample Design Years 1-3 (2011-2013). http://www.health.govt.nz/publication/new-zealand-health-survey-sample-design-years-1-3-2011-2013

R Development Core Team. (2012). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from http://www.R-project.org/ (ISBN 3-900051-07-0)