EVALUATION OF A TWO-PHASE APPROACH TO SEGMENT SELECTION FOR AREA PROBABILITY SAMPLES LATE IN A DECADE

Leyla Mohadjer, Jill Montaquila, and Erica Sherris, Westat
Jill Montaquila, Westat, 1650 Research Blvd., Rockville, Maryland 20850

KEY WORDS: New construction, measure of size, double sampling, decennial census

1. Introduction

In this paper, we consider a multi-stage area sampling scenario in which both primary sampling units (PSUs) and secondary sampling units (SSUs, or segments) are selected with probabilities proportionate to size (PPS). Further stages of selection may be done within the sampled segments. PPS selection is used in order to yield self-weighting estimates while maintaining control over the sample size. Often, the measure of size (MOS) used for PPS selection of PSUs and segments in household area sample surveys are functions of household or population counts. In large national household surveys, PSUs are typically counties or groups of contiguous counties, and segments typically comprise blocks or groups of contiguous blocks.

Household and population counts from the decennial census are available at both the county and block level. Thus, early in the decade, once the decennial census data have been released, it is possible to get quite accurate MOS for the selection of both PSUs and segments. However, later in the decade, as the decennial census data become outdated (due, for example, to new construction or demolition of dwelling units), MOS based on the census data become inaccurate. Intercensal population estimates are available at the county level and may be used to obtain more accurate MOS for PSU selection late in a decade. However, population and housing estimates for subcounty areas such as blocks are available only from the decennial census.

Probability samples may be classified into two groups: fixed-rate and fixed-size samples. With fixed-rate samples, a fixed, predetermined sampling rate is used and the sample size may vary. With fixed-size samples, a fixed sample size is specified and the sampling rate may vary. Inaccurate segment measures of size will create problems for both of these types of probability samples. With fixed-size samples, large variations in segment sizes result in large variations in probabilities of selection of the ultimate sampling units, which in turn reduces the precision of estimates based on the sample. With fixed-rate samples, large variations in segment sizes results in large variations in the numbers of ultimate sampling units selected in each segment. This may result in sample sizes that are considerably different from anticipated sample sizes; it may also have operational and cost implications.

The sampling of residential building permits for new construction (referred to as "building permit sampling") was introduced as a method to control the deviations in segment size from the expected size. Used for several decades and in many surveys, building permit sampling creates new construction segments, separate from the regular area segments. A frame of building permit data is constructed using data from the Building Permits Survey (BPS) conducted by the U.S. Census Bureau. This frame contains residential dwelling units for which building permits were issued since the last decennial census, and a sample of newly constructed units is selected from this frame. The regular area segments are assigned measures of size based on data from the last decennial census; and during screening newly constructed units are excluded from (screened out of) the regular area segments. Bell et al. (1999) contains a detailed description of the procedures used in permit sampling and issues related to the building permit sampling.

On the whole building permit sampling has been a very effective method of reducing variations in segment sizes, thus allowing for more efficient fieldwork, tighter cost controls, and improved precision of the survey estimates. However, there are a few shortfalls of this approach. Judkins et al. (2000) discuss the costs and benefits of building permit sampling, with consideration to both budgetary issues and statistical issues.

Building permits are not required for the placement of mobile homes. The permit files contain counts of the numbers of units for which building permits were authorized; a small fraction of units for which permits are authorized are never built. An additional concern relates to obtaining accurate measures of size based on the data from the BPS, since some of the data are imputed. Additionally, areas and units that do not require building permits for new construction are not included. According to the Census Bureau (U.S. Bureau of the Census 1994), in 1994, 5 percent of the U.S. population lived in jurisdictions that do not require building permits. This percentage of the population is heavily concentrated in the South Central and Great Plains areas: 26 percent of the population in the East South Central Census division, 17 percent of the population in the West South Central Census division, and 11 percent of the population in the West North Central Census division are contained in these 'nonpermit' jurisdictions, with less than 2 percent of the population of the remaining census divisions in nonpermit jurisdictions. Another concern is that, although a permit is issued, the unit might never have been built.

Additionally, the success of the building permit sampling approach depends upon being able to obtain accurate permit information. If new construction segments are selected, the listers must obtain specific permit information from local building permit offices, this requires the cooperation of the local permit office officials.

An operational concern with building permit sampling is the cost and effort required to screen out new construction in the regular area segments. A related concern is that the success of the method depends in part upon the respondents' knowledge of when the dwelling unit was built. An examination that showed a great deal of variation in estimates of the number of "new construction" units constructed during the 1980s led to concerns about building permit sampling. According to the 2000 decennial census, about 17 percent of housing units were built between 1990 and March 2000. Based on data from one area sample survey conducted by Westat in 1999-2000, about 18 percent of persons reside in housing units built between 1990 and 1999-2000.

In light of these concerns with building permit sampling, a two-phase sampling (or double sampling) approach (Montaquila et al. 1999) was developed and used for segment selection in a national area sample survey conducted by Westat.

With the two-phase approach, a larger sample of segments is selected in the first phase using a MOS updated to reflect expected growth in the segment. Counters are then sent to each Phase 1 segment to obtain counts of the numbers of dwelling units. A new MOS, reflecting the difference between the expected and actual number of dwelling units, is calculated for Phase 2 selection. The final sample of segments is selected from the Phase 1 sample using the updated MOS. Further details of this approach are given in the next section.

2. Description of the Two-Phase Approach

In this section, we describe the calculations used to update the MOS for Phase 1 selection, to compute the MOS for Phase 2 selection, and to determine the size of the Phase 1 sample.

2.1 Selection of the Phase 1 Sample

Let M_{hi} denote the measure of size for segment *i* in PSU *h*, based on data from the most recent decennial census.

Using building permit data, we can obtain estimates of "growth" in places having permit-issuing offices. Let U_{hp} denote the number of units for which building permits were issued since the most recent census in place *p* in PSU *h*. Note that U_{hp} is available only for places with permit offices and is obtained directly from the building permit files compiled by the Building Permits Survey. The number of persons residing in newly constructed units in place *p* may be estimated by $T_{hp}^{[u]} = 2.6U_{hp}$. The estimate of the place-level "growth" ratio is then given by

$$g_{hp} = \frac{T_{hp}^{[o]} + T_{hp}^{[u]}}{T_{hp}^{[o]}}, \qquad (1)$$

where $T_{hp}^{[o]}$ is the total population for place *p* in PSU *h* from the most recent decennial census.

For the first phase of segment selection, the measure of size for each segment is adjusted for the place-level growth (which is used as a proxy for segment-level population changes). That is, the measure of size for segment *i* in PSU *h* in the Phase 1 segment selection is M'_{hi}, where

$$M'_{hi} = M_{hi} g_{hp}, (hi) \in (hp). \qquad (2)$$

Because M'_{hi} is an estimate of the true segment measure of size that is based on place-level data, it is subject to error. The second phase of segment selection (described below) will correct for errors in M'_{hi}. However, in order to ensure that target overall sampling rates or sample sizes can be attained, it is necessary to select a Phase 1 sample of segments that is somewhat larger than the ultimate segment sample size and will enable within-segment target sampling rates or sample sizes to be attained. (See section 2.3.)

2.2 Selection of the Phase 2 Sample

Typically, segments are much smaller than places. Therefore, the estimated place-level growth may differ considerably from the true change in the size of the segment since the last decennial census. In order to obtain more accurate estimates of the true change in the size of the segment, a new "counting" procedure is used for the Phase 1 segments. "Counters" (experienced listers) travel to the Phase 1 segments and count the number of dwelling units in each of the segments. These counts are used to compute measures of size for the Phase 2 segment selection such that the overall probabilities of selection of the segments are accurate.

Let U'_{hi} denote the number of DUs found by counters when counting Phase 1 segment *i* in PSU *h*. The change in the size of Phase 1 segment *i* is estimated based on U'_{hi} as follows:

$$g'_{hi} = \frac{U'_{hi}}{U^{[0]}_{hi}}, \quad (3)$$

where $U^{[0]}_{hi}$ is the number of DUs in segment *i* in PSU *h* at the time of the most recent decennial census.

The measure of size for Phase 2 selection is

$$M^{[2]}_{hi} = \frac{M_{hi} g'_{hi}}{M'_{hi}} = \frac{g'_{hi}}{g_{hp}}, \ (hi) \in (hp) \quad (4)$$

Let U_h denote the set of all segments in PSU *h* on the frame for Phase 1 selection, and let S_{h1} denote the set of segments in PSU *h* selected in Phase 1. The overall probability of selection of segment *i* in PSU *h* (conditional on the sampled PSUs) is

$$p_{hi} = \frac{k_{1h} M'_{hi}}{\sum_{i \in U_h} M'_{hi}} \cdot \frac{k_{2h} M^{[2]}_{hi}}{\sum_{i \in S_{1h}} M^{[2]}_{hi}} = \frac{k_{1h} M'_{hi}}{\sum_{i \in U_h} M'_{hi}} \cdot \frac{k_{2h} M_{hi} g'_{hi} \Big/ M'_{hi}}{\sum_{i \in S_{1h}} M^{[2]}_{hi}} = \frac{k_{1h} k_{2h} M_{hi} g'_{hi}}{\sum_{i \in U_h} M'_{hi} \sum_{i \in S_{1h}} M^{[2]}_{hi}}, \quad (5)$$

where k_{1h} and k_{2h} are the numbers of segments selected in PSU *h* in the Phase 1 and Phase 2 samples, respectively.

2.3 Minimum MOS and Size of the Phase 1 Sample

With two-phase sampling, a key consideration is the size of the phase 1 sample. Here, we will consider this in the context of a fixed-rate sample. A similar derivation applies in the case of a fixed-size sample. Let *r* denote the target overall sampling rate. In many applications different sampling rates will be used for different sampling domains. In those situations subsampling will generally be done at stages subsequent to segment selection, so *r* denotes the maximum overall sampling rate.

For this rate to be attainable, it must be the case that $p_{hi} \geq \frac{r}{\pi_h}$. This condition is equivalent to the condition

$$k_{1h} \geq \frac{r \sum_{i \in U_h} M'_{hi} \sum_{i \in S_{1h}} M^{[2]}_{hi}}{\pi_h k_{2h} M_{hi} g'_{hi}} \forall i \in h. \quad (6)$$

Furthermore, since $M^{[2]}_{hi}$ is the ratio of the actual change in segment MOS to the expected change in segment MOS, on average, this ratio should be equal to 1, so

$$\sum_{i \in S_{1h}} M^{[2]}_{hi} \approx k_{1h}. \quad (7)$$

Thus, expression (6) can be used to obtain an approximate lower bound on the "original" measure of size (i.e., the measure of size based on the previous census data):

$$M_{hi} \geq \frac{r \sum_{i \in U_h} M'_{hi}}{\pi_h k_{2h} g'_{hi}} \forall i \in h. \quad (8)$$

If we let M^* denote the minimum MOS given on the right-hand side of expression (8), it can be shown that the number of segments to be selected in Phase 1 in order to attain the target sampling rate is

$$k_{1h} \geq \frac{M^* \sum_{i \in S_{1h}} d_{hi}}{M_{hi} d_{hi}} \forall i \in h, \quad (9)$$

where $d_{hi} = g'_{hi} \Big/ g_{hp}$.

Thus, the number of Phase 1 segments to be selected is a function of the variation in segment-level growth rates from the place-level growth rates. Although this variation cannot be assessed prior to selecting and listing the Phase 1 sample, we have found some general rules to be useful. In counties with very few places, in which the places are geographically large and potentially diverse, a larger Phase 1 sample should be selected. In counties with a great deal of variation in growth rates from place to place, a larger Phase 1 sample is advisable. In our study the target number of Phase 2 segments was generally around 24. The Phase 1 segment sample sizes ranged from 44 to 145 (see Table 1), with the exception of one PSU that had 395 Phase 1 segments. (This PSU was one with an extremely high rate of new construction.)

3. Evaluation

The double sampling approach was used in 26 PSUs in our study. Within these PSUs there were 2,488 segments selected in the Phase 1 sample, and 617 of those were selected into the Phase 2 sample. For the 2,488 segments selected in the first phase, we would like to determine how efficient the Phase 1 MOS is in light of

changes in population, and whether certain PSU characteristics resulted in a more effective MOS adjustment for Phase 1. As a measure of the efficiency of the Phase 1 MOS, we looked at the reciprocals of the Phase 2 MOS, i.e., $R_{hi} = 1/M_{hi}^{[2]}$. Note, from equation (4), that this reciprocal is the ratio of the place-level growth to the actual change in size of the Phase 1 segment.

Table 1. Sample sizes of Phase 1 and Phase 2 segment selection

Segments selected in Phase 1	Segments selected in Phase 2	Range of place growth ratio
395	24	8.72
110	23	1.45
125	20	0.95
145	24	0.82
110	24	0.78
89	24	0.29
88	24	0.29
45	23	0.27
84	24	0.27
75	24	0.24
80	24	0.23
90	24	0.22
80	24	0.18
125	24	0.18
99	24	0.17
90	24	0.14
60	24	0.11
60	24	0.07
90	24	0.07
74	24	0.05
70	24	0.05
44	24	0.01
95	24	<0.01
55	23	0
65	24	0
45	24	0
2,488	617	

In general, the Phase 1 MOS tended to suggest higher growth than the population warrants. This may be due in part to the fact that the permit data reflect permits issued, but the dwelling units may or may not have been built. Additionally, the permit data enable us to account for growth in the segment, but not demolition. In 72 percent of the segments, R_{hi} had a value greater than 1. Certain PSU characteristics appeared to be related to the percent of segments in the PSU for which the place-level growth was larger than the actual change in size of the Phase 1 segment. For example, 74 percent of segments in PSUs with a low percent elderly had a value of $R_{hi} > 1$, while only 69 percent of segments in PSUs with a high percent elderly had a value of $R_{hi} > 1$ (see Table 2).

Table 2. Number of segments where phase 1 MOS overestimates, underestimates, or is equal to actual MOS

Characteristic of PSU	Phase 1 MOS		
	Over (%)	Under (%)	Equal (%)
Income			
Low	698 (66)	354 (34)	0 (0)
High	1,101 (77)	332 (23)	3 (0)
% Elderly			
Low	1,113 (74)	382 (26)	2 (0)
High	686 (69)	304 (31)	1 (0)
% Black			
Low	686 (66)	346 (33)	3 (0)
High	1,113 (77)	340 (23)	0 (0)
Population			
Low	678 (66)	348 (34)	3 (0)
High	1,121 (77)	328 (23)	0 (0)
Overall	**1,799 (72)**	686 (28)	3 (0)

Similar results can be seen from looking at the median value of R_{hi} by PSU (see Table 3). Looking at percent elderly again, we can see that the median value of R_{hi} is 1.18 for PSUs that have a low percent elderly and 1.06 for those with a high percent elderly. R_{hi} tends to be more variable and less accurate (further from 1) when a PSU has any of the following characteristics: large population; high per capita income; high percent minority; or low percent elderly.

Table 3. Statistics on the medians of R_{hi} by PSU

	Mean	Range
Number of Families (1990)		
Low	1.08	0.19
High	1.16	0.48
Per Capita Income (1989)		
Low	1.07	0.18
High	1.17	0.47
Percent White/Other (1996)		
Low	1.13	0.50
High	1.11	0.31
Percent Black (1996)		
Low	1.09	0.20
High	1.15	0.50
Percent Hispanic (1996)		
Low	1.10	0.35
High	1.14	0.48
Percent Elderly (1996)		
Low	1.18	0.48
High	1.06	0.14

We ran a regression analysis to determine whether there were PSU characteristics that predicted the deviation of R_{hi} from 1. With 9 PSU characteristics considered in the model, the only significant predictor of distance of R_{hi} from 1 was percent elderly (p-value = .0002), where an increase in percent elderly indicated a decrease in the efficiency of the Phase 1 MOS.

While most of the Phase 1 segment MOS fairly accurately represented the desired MOS based on current population (2,116 segments had a value of R_{hi} between 0.5 and 1.5), there were several extreme ratios. Values of R_{hi} ranged from 0.08 to 60.99. Table 4 shows the number of segments, and the number of PSUs containing segments, that have extreme ratio values.

Table 4. Number of segments with extreme values of R_{hi}

Characteristics of ratio	Number of segments	Number of PSUs with segments
$R_{hi} < 0.15$	5	2
$R_{hi} < 0.25$	20	9
$R_{hi} < 0.50$	69	17
$R_{hi} > 1.50$	296	24
$R_{hi} > 2.50$	52	9
$R_{hi} > 3.00$	36	7

***** The data are based on 2,481 segments in 26 PSUs with valid values of the ratio.

4. Conclusions and Recommendations

An examination of the cost of two-phase sampling indicated that, on average, the cost of counting in a PSU is approximately equal to the cost of listing. Thus, in terms of pre-field activities, the two-phase procedure is approximately double the cost of building permit sampling. However, this additional cost is somewhat offset by the cost savings during data collection due to the reduced travel (within PSUs) for interviewers and the reduced amount of screening out of households.

The primary advantage of two-phase sampling, relative to permit sampling, is the improved coverage of persons residing in newly constructed units.

At this time, it appears a hybrid approach to segment selection—with permit sampling used in some PSUs and two-phase sampling used in others—is advisable. In that case, permit sampling would be recommended for PSUs with very low rates of new construction and demolition, with very good permit coverage (i.e., very few mobile homes and no areas that are exempt from permits), and—if known—good access to building permit records. Two-phase segment selection would be advisable in PSUs with high rates of new construction or with poor permit coverage.

When two-phase sampling is used, ideally, a very large sample of segments should be selected in the first phase so that any unusual segment growth can be incorporated into the final segment sample. The size of the Phase 1 sample is a function of how variable the rate of growth is among segments in a PSU. If the rate of growth is exactly uniform among all segments then the Phase 1 sample will be equal to the final segment sample. However, if the rate of growth is very variable, then it will be necessary to select many more segments within the PSU. Based on the findings of this evaluation, after accounting for variation in growth rates within the PSU, the accuracy of the Phase 1 MOS is related to the percent elderly in the PSU; the lower the percent elderly, the less accurate the Phase 1 MOS. Thus, in stands with low percent elderly, it is advisable to select a larger number of segments; in stands with high percent elderly, it is possible to select relatively fewer segments.

5. References

Bell, B., Mohadjer, L., Montaquila, J., and Rizzo, L. (1999). Creating a frame of newly constructed units for household surveys. *Proceedings of the Survey Research Methods Section*. Washington, DC: American Statistical Association, 306-310.

Judkins, D., Cadell, D., and Sczerba, K. (2000). Costs and benefits of a permit sample late in the decade. *Proceedings of the Survey Research Methods Section*. Washington, DC: American Statistical Association, 671-676.

Montaquila, J.M., Bell, B., Mohadjer, L., and Rizzo, L. (1999). A methodology for sampling households late in a decade. *Proceedings of the Survey Research Methods Section.* Washington, DC: American Statistical Association, 311-315.

U.S. Bureau of the Census (1994). *Current Construction Reports—Housing Units Authorized by Building Permits: Annual 1994*, Publication C-40/94-A, U.S. Government Printing Office, Washington, DC.