2012

Minimizing sample overlap with surveys using different geographic units

Kevin Lu
Australian Bureau of Statistics

Minimizing sample overlap with surveys using different geographic units

Abstract

The ABS runs Australia-wide population surveys using area-based multi-stage designs. One challenge for the ABS and other National Statistical Organizations is to avoid returning to areas selected in other recent surveys so that households are not overburdened with multiple surveys, while ensuring areas have the correct unconditional probabilities of selection for the survey to represent all of the country. There is a well-known method to choose primary-stage units in a way that minimizes overlap and leaves the unconditional probabilities of selection unchanged. However, this method cannot simply be applied when the primary-stage units in the current survey are geographically di erent from those used in previous surveys. We develop two extensions to the existing approach for an ABS household survey facing this challenge. The first method uses simulations as part of computing conditional probabilities of selection, while the second uses a weighted average of conditional probabilities applied on the geographic intersections of the previous and current primary-stage units. We show that both methods preserve the unconditional probability of selection, but do not achieve the same levels of overlap.

Keywords

Overlap control, minimizing overlap, sample selection

Publication Details

Lu, Kevin, Minimizing sample overlap with surveys using different geographic units, Proceedings of the Fifth Annual ASEARC Conference - Looking to the future - Programme and Proceedings, 2 - 3 February 2012, University of Wollongong.

Minimizing sample overlap with surveys using different geographic units

Kevin Lu

Australian Bureau of Statistics

Abstract

The ABS runs Australia-wide population surveys using area-based multi-stage designs. One challenge for the ABS and other National Statistical Organizations is to avoid returning to areas selected in other recent surveys so that households are not overburdened with multiple surveys, while ensuring areas have the correct unconditional probabilities of selection for the survey to represent all of the country. There is a well-known method to choose primary-stage units in a way that minimizes overlap and leaves the unconditional probabilities of selection unchanged. However, this method cannot simply be applied when the primary-stage units in the current survey are geographically different from those used in previous surveys. We develop two extensions to the existing approach for an ABS household survey facing this challenge. The first method uses simulations as part of computing conditional probabilities of selection, while the second uses a weighted average of conditional probabilities applied on the geographic intersections of the previous and current primary-stage units. We show that both methods preserve the unconditional probability of selection, but do not achieve the same levels of overlap.

Key words: Overlap control, minimizing overlap, sample selection

1. Introduction

There exists a wide class of overlap control methods in the statistical literature for a variety of situations. An overview of these methods and their limitations can be found in [1]. The ABS is changing to a new geography standard. This means that the geographic areas that are used as primary-stage units (PSU) in the previous surveys will cover different physical regions to the geographic areas used in current and future household surveys. Existing methods generally do not apply to this situation, so we develop some methods to deal with this challenge.

In the design of household surveys, the probability of selection for each PSU is derived assuming that there is no need to avoid areas selected in previous surveys. These probabilities are usually nonzero for all in-scope areas, giving each in-scope person a chance to be selected, so that we do not increase bias by excluding previously selected areas. For this reason and in order to reduce respondent burden, we instead select PSUs using a probability that is conditional on the selection outcome of the previous surveys, such that

1. The unconditional probability of selection is equal to the design probability.
2. Overlap is reduced or minimized when selecting with the conditional probability.

Therefore, we can justify the use of conditional probabilities by seeing that over all possible random samples of the previous surveys, each PSU is selected with the correct design probability. Suppose we want to avoid overlap with m previous surveys. For PSU j, let B_k^j, $k = 1, \ldots, 2^m$, be the selection outcome of PSU j in the previous surveys and $\mathcal{S}_i^j$ be the event that PSU j has been selected in previous survey s_i. We will usually write B_k and $\mathcal{S}_i$ for simplicity. As an example, if there are 2 previous surveys, then $B_1 = \bar{\mathcal{S}}_1 \cap \bar{\mathcal{S}}_2, B_2 = \bar{\mathcal{S}}_1 \cap \mathcal{S}_2, B_3 = \mathcal{S}_1 \cap \bar{\mathcal{S}}_2, B_4 = \mathcal{S}_1 \cap \mathcal{S}_2$. To select PSU j with probability $P(A)$, the first criteria lets us instead select it with probability $P(A|B_k)$, which is any function of B_k that takes values only in $[0, 1]$, satisfying

$$P(A) = \sum_k P(A|B_k)P(B_k). \tag{1}$$

This function should then be chosen to minimize overlap with the previous surveys. In the case where there is only one previous survey, for each PSU, B_1 is the event that the PSU was not previously selected, and B_2 is its complement. If $P(A) \leq P(B_1)$, to minimize overlap, set

$$P(A|B_2) = 0 \tag{2}$$

so that we cannot select PSUs that were previously used, this implies

$$P(A|B_1) = \frac{P(A)}{P(B_1)}. \tag{3}$$

Otherwise $P(B_1) < P(A)$ and we should set

$$P(A|B_1) = 1 \tag{4}$$

so that PSUs that have not been used are selected with certainty, this leaves

$$P(A|B_2) = \frac{P(A) - P(B_1)}{P(B_2)}. \tag{5}$$

Note that the 2 cases are needed to ensure the conditional probabilities are in $[0, 1]$. This method can be generalized when there are multiple previous surveys for which we want to avoid, see [2] for the relevant formulas.

This is the standard overlap control method applied in ABS household surveys. We develop two methods of overlap control that extends the approach described to deal with a change in geography, while satisfying both outlined criteria.

2. Simulation Method

The standard method cannot be simply applied when PSUs have geographically changed, since current PSUs were not directly selected in previous surveys. A straightforward fix is to replace the notion of selected used in the definition of the events B_k with touched and call these events B'_k. We define a current PSU as being *touched* if it intersects with at least one previously selected PSU. The unconditional probability is still preserved when B_k is replaced with B'_k since these events form a partition.

This reduces the problem to calculating the probabilities B'_k to be used in Equations (2)-(5) or its generalizations, if there are multiple previous surveys. Depending on the selection scheme used, these probabilities may not be easy to find analytically. For example, in Figure 1, B'_3 for the current PSU C1 may be the event that it is touched in the previous survey s_1, but not in s_2. This is equivalent to at least one of the previous PSUs P1, P2, P3 being selected in s_1 and none of them being selected in s_2, that is $B'^{C1}_3 = \mathcal{S}^{P1}_1 \cup \mathcal{S}^{P2}_1 \cup \mathcal{S}^{P3}_1 \cap \bar{\mathcal{S}}^{P1}_2 \cap \bar{\mathcal{S}}^{P2}_2 \cap \bar{\mathcal{S}}^{P3}_2$.

Suppose we want to avoid overlap with m previous surveys $s_1, \ldots, s_m$. Typically, each survey has selected previous PSUs with probabilities that are conditional on the surveys that have come before it. We assume that s_1 was able to select with design probabilities. For each current PSU, the probabilities $P(B'_k)$ can be estimated in the following way.

1. Simulate a selection for s_1.
2. Find the conditional probabilities of selection for s_2 given the selection simulated for s_1.
3. Using these conditional probabilities simulate a selection of s_2.

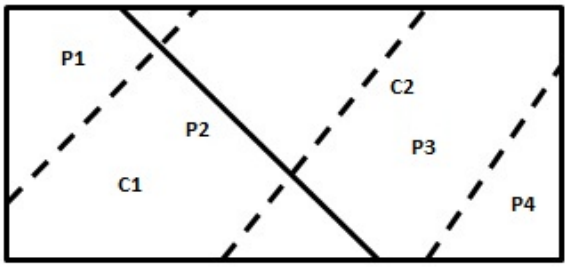

Figure 1: An area divided into 2 current PSUs (C1, C2) by the solid line and 4 previous PSUs (P1, P2, P3, P4) by the dotted lines.

4. Continue to select all previous surveys in this way, where the conditional probability of selection for s_i is dependent on the selection simulated for $s_1, \ldots, s_{i-1}$.
5. Repeat step 1 to 4 a large number of times.

The probabilities $P(B'_k)$ are estimated by the relative frequencies obtained in the simulation.

3. Weighted Average Method

In this section, we give an alternative way to extend the standard overlap control method. Define a *split PSU* as the geographic area formed by the intersection of a previous PSU with a current PSU. In Figure 1, there are 7 split PSUs.

For each split PSU i in current PSU j, use Equations (2)-(5) to compute the conditional probability of selection, π_i, where $P(A)$ is set to the unconditional probability of selecting the current PSU containing split PSU i, and $P(B_k)$ is defined in Section 1. If there are multiple past surveys, then the probabilities $P(B_k)$ may be estimated using simulations similar to the procedure outlined in Section 2.

Define the conditional probability of selecting PSU j in the current survey as the weighted average of the conditional probabilities

$$\pi^*_j = \sum_{i \in S_j} w_i \pi_i,$$

where

$$w_i \in [0, 1] \text{ for all } i \in S_j \quad \text{and} \quad \sum_{i \in S_j} w_i = 1, \tag{6}$$

and S_j is the set of split PSUs in current PSU j.

We now show that selecting current PSU j with probability π^*_j preserves the unconditional probability. We can treat the conditional probability of selecting the split PSUs as a random variable $\pi_i = P(A|B_k)$ with probability $P(B_k)$. Using Equation (1), $E(\pi_i) = P(A)$. Thus, π^*_j is also a random variable. Its realizations, the probabilities used to select current PSU j, are in $[0, 1]$

because of (6). Using the Law of Total Probability, the unconditional probability of selecting current PSU j is

$$E(\pi_j^*) = \sum_{i \in S_j} w_i E(\pi_i) = P(A),$$

since the weights add up to 1.

3.1. Choosing weights

For each current PSU, selecting with any weighted average of conditional probabilities does not change the unconditional probability, as long as the weights are fixed and satisfy (6). We now discuss a few ways to choose these weights.

For simplicity, we can choose equal weights within each current PSU, changing the weighted average into a simple average.

We can also optimize the weights to minimize overlap with the selection of previous surveys that has actually occurred. This can be done by solving the linear program (LP)

$$\begin{aligned} \min_{w} \quad & \sum_{j \in U} O_j \left(\sum_{i \in S_j} w_i \pi_i \right) \\ \text{subject to} \quad & \sum_{i \in S_j} w_i = 1 \quad \text{for all } j \in U \\ & 0 \le w_i \le 1 \quad \text{for all } i \\ & \sum_{j \in U} N_j \left(\sum_{i \in S_j} w_i \pi_i \right) = n, \end{aligned} \tag{7}$$

where U is the set of all current PSUs, O_j is some measure of overlap in current PSU j, N_j is the number of ultimate selection units in current PSU j, and n is the required number of ultimate selection units. The equality constraint fixes the expected sample size. This is advantageous as all other overlap control methods mentioned here changes the expected sample size in general.

While LP (7) is quite easy to solve, using these weights will *not* preserve the unconditional probability of selection, since the weights are no longer fixed, they will change depending on the selection of the previous survey that has occurred.

This can be overcome by minimizing the expected overlap over all possible selections of the previous surveys. Therefore we want to find fixed weights w_i that solves

$$\begin{aligned} \min_{w} \quad & E\left(\sum_{j \in U} O_j(B) \left(\sum_{i \in S_j} w_i \pi_i(B) \right) \right) \\ \text{subject to} \quad & \sum_{i \in S_j} w_i = 1 \quad \text{for all } j \in U \\ & 0 \le w_i \le 1 \quad \text{for all } i \\ & E\left(\sum_{j \in U} N_j \left(\sum_{i \in S_j} w_i \pi_i(B) \right) \right) = n, \end{aligned} \tag{8}$$

where the expectation is taken over B, the sample selected in all previous surveys. Note that O_j and π_i depends on B. Therefore, solving this problem may be very computationally demanding, since there are usually an extremely large number of possible samples for the previous surveys.

3.2. Solvability of the LPs

There is no guarantee that either LPs have solutions. However, they are both LP minimization problems bounded from below by 0, so there exists an optimal solution if the feasible set is nonempty. We derive a condition for solvability, so that if there is no solution, the LP can be altered to have a solution. Assume that the sample size constraints of the LPs have been rewritten in the form

$$\sum_{j \in U} N_j \left(\sum_{i \in S_j} w_i \alpha_i \right) = n.$$

We show that

$$\sum_{j \in U} N_j \min_{i \in S_j}(\alpha_i) \le n \le \sum_{j \in U} N_j \max_{j \in S_j}(\alpha_i) \tag{9}$$

is a necessary and sufficient condition for the existence of an optimal solution.

Using (6), it can be shown that

$$\min_{i \in S_j}(\alpha_i) \le \sum_{i \in S_j} w_i \alpha_i \le \max_{i \in S_j}(\alpha_i),$$

which implies (9).

We now use (9) to construct a feasible point to show that it is sufficient. Let

$$x_j = \sum_{i \in S_j} w_i \alpha_i.$$

Consider the continuous function f defined on $X = \left\{ \mathbf{x} \in \mathbb{R}^{|U|} \middle| \min_{i \in S_j}(\alpha_i) \le x_j \le \max_{i \in S_j}(\alpha_i), \text{ for all } j \right\}$ by

$$f(\mathbf{x}) = \sum_{j \in U} N_j x_j.$$

Using (9) with the Intermediate Value Theorem, there exists a $\mathbf{c} \in X$ such that $f(\mathbf{c}) = n$. So for all j,

$$\sum_{i \in S_j} w_i \alpha_i = c_j \in \left[\min_{i \in S_j} \alpha_i, \max_{i \in S_j} \alpha_i \right]. \tag{10}$$

It is trivial to choose w_i to satisfy (6) and Equation (10). Therefore, a feasible solution exists and (9) is sufficient.

In the case where the LP has no solution, replace the n with the nearest bound given in (9).

4. Comparing the methods for a household survey

We applied the simulation and weighted average methods to two strata used in an ABS household survey. In the Low Usage strata 14.4% of previous PSUs has been selected and we want to select 56.6 out of 2248 current PSUs. The High Usage strata have proportionally higher usage, 84.6% of previous PSUs has been selected and we want to select 3.4 out of 23 current PSUs. The unconditional probabilities of selection within the strata are equal. The previous survey selected PSUs with probability proportional to cluster size. We computed the conditional probability of selection, π_j^*, with both the simulation method and the weighted average method. The expected amount of overlap is then given by

$$\sum_{j \in U} O_j \pi_j^*,$$

where O_j is the estimated number of households in the overlapping areas of current PSU j.

When there is no overlap control, the expected number of dwellings in overlapping areas is 1227.8 in the Low Usage strata and 500.0 in the High Usage strata. The results summarized in Tables 1 and 2 give the percentage reduction in expected overlap from the case with no overlap control.

Strata	Reduction in overlap	Expected number of current PSUs selected
Low Usage	100%	54.2
High Usage	39.7%	3.1

Table 1: Overlap results for the simulation method.

Strata	Reduction in overlap	Expected number of current PSUs selected
Low Usage	72.4%	54.9
High Usage	47.9%	3.0

Table 2: Overlap results for the weighted average method using a simple average.

These results indicate that the simulation method performs far better at avoiding overlap in strata where there is a small amount of previous PSU usage, whereas the weighted average method may be better when there is a large amount of previous PSU usage. Note that the expected number of current PSUs selected has changed from the design values.

4.1. Explaining the results

For a current PSU, let $E_i, i = 1, \ldots, n$ be the event of selecting its ith split PSU in the previous survey. Suppose that $P(E_1), \ldots, P(E_n), P(E_1 \cup \ldots \cup E_n)$ are sufficiently large, the current PSU has been completely selected and we only want to avoid a single previous survey, then the conditional probability of selection using the simulation method is

$$\frac{P(A) - P(\overline{E_1 \cup \ldots \cup E_n})}{P(E_1 \cup \ldots \cup E_n)} \leq P(A),$$

which approaches its upper bound as $P(E_1 \cup \ldots \cup E_n)$ approaches 1.

However, with the weighted average method, the conditional probability is

$$\sum_{i \in S_j} w_i \frac{P(A) - P(\bar{E}_i)}{P(E_i)} \leq P(A),$$

which approaches the same upper bound under the much stronger condition that $P(E_i)$ approaches 1 for all i with nonzero weight. So this current PSU does not as readily contribute to overlap when using the simulation method.

When the probabilities and joint probabilities of selection in the previous survey and the number of previously selected PSUs are all sufficiently large, the expected amount of overlap using the simulation method will be at least be that achieved under the weighted average method. Heuristically, the simulation method performs worse in these situations because it does not take into account the amount of overlap that would result from selecting a current PSU, only whether it was touched. The weighted average method incorporates information about the selection outcome of split PSUs and the overlap to which it contributes.

Conversely, the simulation method tends to perform far better than the weighted average method when the number of previously used PSUs is sufficiently small, because it minimizes the occurrence of selecting current PSUs that have been touched.

5. Conclusion

We have proposed two methods for avoiding overlap with previous surveys under changes in the geographic areas used as PSUs. Both methods preserve the unconditional probability of selection. However, the simulation method generally reduces overlap more than the weighted average method which is only better when the probability of selection in the previous survey and the amount of previous PSU usage is large.

Further work could investigate ways to choose optimal weights for the weighted average method, in particular finding simplifications to LP (8) so that it can be feasibly implemented and explicit conditions under which one method is better than the other.

References

[1] L. R. Ernst, The maximization and minimization of sample overlap problems: A half century of results, Proceedings of the International Statistical Institute (1999) 168–182.

[2] S. Chowdhury, A. Chu, S. Kaufman, Minimizing overlap in nces surveys, Proceedings of the Survey Research Methods Section, American Statistical Association (2000) 174–179.