WORKLOAD BIAS: A CASE FOR EMPLOYING RANDOMIZATION IN THE SCHEDULING OF FIELD ASSIGNMENTS

Douglas J. Drummond, Research Triangle Institute

1. Introduction

Virtually all survey designs are formulated in response to certain constraints. For the most part, these constraints take the form of restrictions on budget and/or precision levels for estimates of major parameters of interest, which in turn lead to consideration of the underlying response burden and rate, measurement process, instrumentation, capabilities of respondent, etc. This paper considers any such design implemented under the further restriction of an upper bound on the length of the data collection period. Methodology is described involving the use of randomization in the scheduling of field assignments. An analytic structure is thereby furnished for addressing the potential undercoverage bias induced in estimates of study parameters whenever survey efforts are prematurely ended. Major concepts are illustred using material developed by the author for the 1980 Demographic Survey in Bolivia.

2. Background Discussion

The Bolivian Demographic Survey was supported by a stratified three-stage sample design [1,2]. Specifically, area units were selected at each of the first two stages, and a census carried out of all families residing in each penultimate sampling unit. Field efforts were to be carried out over a six week period utilizing thirteen field teams[1] organized into four regions. Within a field region, teams were to move as an integrated unit, enumerating clusters of sites (i.e., primary sampling units (PSUs)) sequenced so as to minimize travel costs. Such a mode of operation was intended to allow the regional supervisors to travel with their group of teams and reach any particular member team within a twenty-four hour period. Given the magnitude of the task (involving the enumeration of a projected 11,000 dwelling units and entailing field efforts in rural areas of the country for the first time) and the political uncertainties in that country at the time (conditions which lead to a military coup prior to the start of field efforts), a strong likelihood existed that field efforts would not be completed in the allotted time.[2] With this in mind, two mechanisms were introduced into the design in the hope of potentially reducing nonresponse bias that would be otherwise present in the study findings as the results of not completing the data collection effort:

1. start-point randomization, by field region, on each travel sequence; and

2. a framework for scheduling workloads within a site in the event that a maximum time-on-site was imposed during the course of the study.

Each will be discussed in turn. In both cases, randomizations were externally imposed (i.e., not applied by individual field teams) and were intended solely to provide a framework for ultimately assessing the magnitude of any realized nonresponse bias, without requiring additional efforts by the field team. It should be recognized at the outset that short timelines and the state of communications in Bolivia all but precluded any intervention in data collection scheduling once the survey began. Moreover, country officials were unwilling a priori to reduce the size of their intended survey to possibly better accommodate such a contingency.

3. Start-Point Randomization Within a Field Region

Discussion of start-point randomization will proceed in four stages:

1. Basic notation and concepts.
2. Formalization of randomization for current survey.
3. Operational considerations.
4. Analytical considerations.

Each will be addressed in turn.

a. Basic Notation and Concepts

Consider field region r which may have m(r) sample primary sampling units (PSUs) and let

$t_j(r)$ = time required to adequately complete field work in the j-th PSU of region r.[3]

Then, conditional on the sample second-stage units (SSUs) in region r and the assigned field team, the workload can be adequately completed in the allotted time period, T(r), provided

$$\sum_{j=1}^{m(r)} t_j(r) \le T(r) \ . \qquad \ldots (1)$$

Otherwise, field efforts in region r will not be completed. In discussing this possibility, suppose the field team elects to survey the sample SSUs in a fixed sequence starting at a pre-determined site.[4] Then, conditional on the sample SSUs, the assigned field team, and the chosen order for surveying sites,

$$\Pr\left\{\begin{matrix}\text{PSU } j \\ \text{adequately} \\ \text{surveyed}\end{matrix}\right\} = \begin{cases} 1 & \text{if } \sum_{i=1}^{j} t_i(r) \le T(r) \\ 0 & \text{if } \sum_{i=1}^{j} t_i(r) > T(r) \ . \end{cases}$$

For any specific schedule for data collection, denote this conditional probability by $p_j(r)$ (j= 1, 2, ..., m(r)). In combination (i.e., encompassing selection probabilities and conditional probabilities for completing the required work),

$$\Pr\left\{\begin{array}{l}\text{data accessible}\\ \text{in PSU } j \text{ of region } r\end{array}\right\} = \pi_j(r)\ p_j(r)$$

where

$\pi_j(r)$ = inclusion probability for PSU j of region r.

Coverage of the intended target populations then encompasses three dimensions:

1. All penultimate sampling units be given a positive chance of being selected for the survey.

2. All target population members associated with a sample penultimate unit be given a positive chance of being selected into the survey (assuming field work attempted in the sample SSU).

3. Workload assignment and scheduling guarantee that every sample SSU have a positive chance of being surveyed.

In the event that data collection continues until all sample SSUs are adequately completed, sample designs exhibiting coverage (i.e., properties 1 and 2) will translate into operational designs exhibiting coverage (apart from any nonresponse allowable under "adequately complete" survey operations). Under deterministic scheduling and violations of condition (3), however, target population coverage will not be realized, and what will be termed "workload nonresponse bias" will be introduced into the survey findings.[5] Clearly, this source of nonresponse bias can be avoided by guaranteeing the third dimension of population coverage via workload randomization. That is, conditional on the sample SSUs and field team, to ensure that

$$p_j(r) > 0 \qquad j = 1,2,\ldots,m(r) \qquad \ldots \quad (2)$$

by deviating from a deterministic scheduling of the realized workload. Doing this in a cost-effective fashion while recognizing that any randomization serves solely as a back-up mechanism (i.e., design assumption is that work can be done and only unforeseen difficulties will detract from this goal) is the topic of the next subsection.

b. Developing a Randomization Scheme for the Bolivian Demographic Survey

Scheduling decisions for the Bolivian Demographic Survey were made with respect to all field teams under the jurisdiction of a regional supervisor (and not independently for each team) and by necessity involved clusters of field sites.[6] Moreover, any randomization imposed on the workload scheduling was not permitted to greatly distract from the otherwise efficient scheduling of field sites.[7] With this in mind, regional supervisors were asked to provide a closed-path sequence for surveying clusters of field sites within their region so as to minimize their overall travel costs (i.e., both time and money) between cluster sites. Figure 1 depicts the conceptual structure.

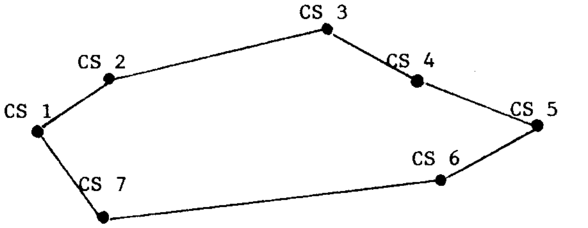

Figure 1: Minimum travel cost sequence for enumerating cluster sites (CS) in a hypothetical field region.

Notationally, let

$$t_{i+}(r) = \sum_{j=1}^{m(r)} t_j(r)\ I_{CS_i}(j)$$

where

$$I_{CS_i}(j) = \begin{cases} 1 & \text{if PSU } j \text{ belongs to the } i\text{-th site} \\ 0 & \text{otherwise} \end{cases}$$

be the time required to adequately survey the i-th cluster site in region r. We wish to address the possiblity that field work may not be completed in a given region, where

$$\Pr\left\{\begin{array}{l}\text{incomplete field work}\\ \text{in region } r\end{array}\right\} = \Pr\ \{(t_{++}(r) > T(r)\}$$

with

$$t_{++}(r) = \sum_{i=1}^{N(r)} t_{i+}(r) ,$$

and

N(r) = number of cluster sites formed in region r.

In general, $t_{++}(r)$ will vary according to many factors (e.g., PSU allocation to region; reallocation of housing unit sample size to PSUs based on realized sample of PSUs and adjustments for growth in the housing unit population; travel time to and from SSU; skill level of field team, etc). Conditional on the observed SSU sample, assigned field team, and workload schedule, the completion of field efforts reduces to a deterministic event. For example, if cluster sites are sequentially surveyed in their natural order, then conditional on this ordering and field team,

$$\Pr\left\{\begin{matrix}\text{CS j in}\\ \text{region r}\\ \text{adequately}\\ \text{surveyed}\end{matrix}\right\} = \begin{cases} 0 \quad \text{if} \quad \sum_{i=1}^{j} t_{i+}(r) > T(r) \\ \\ 1 \quad \text{if} \quad \sum_{i=1}^{j} t_{i+}(r) \le T(r). \end{cases}$$

More generally, denote these conditional cluster site probabilities by $p_{j+}(r)$ $(j = 1,2,\ldots,N(r))$ for a specified underlying schedule of the workload in region r. We would like to impose a randomization scheme that guarantees

$$\text{(i)} \quad \Pr\left\{\begin{matrix}\text{CS j in region r}\\ \text{adequately surveyed}\end{matrix}\right\} > 0$$

for all $j = 1,2,\ldots, N(r)$; and

(ii) retains minimum field cost sequencing of cluster sites.[8]

That is, we would like to select the start-point cluster site for field efforts at random (direction of sequence is inferred by Figure 1 but could be selected at random as well) so that every cluster site has a positive chance of being surveyed during the data collection period, regardless of whether all PSUs can be surveyed in the time allotted. Notationally,

$$\Pr\left\{\begin{matrix}\text{CS j in region r}\\ \text{adequately surveyed}\end{matrix}\right\} = \sum_{i \varepsilon S_j} \gamma_i(r)\, I_{A_j(r)}(i) \qquad \ldots (3)$$

where

$$I_{A_j(r)}(i) = \begin{cases} 1 \quad \text{if} \quad \sum_{\substack{k \varepsilon [i,j] \\ \text{in } S_j}} t_{k+}(r) \le T(r) \\ \\ 0 \quad \text{otherwise} \quad , \end{cases}$$

and

$$S_j = (j+1,\ j+2,\ \ldots,\ N(r),\ 1,2,\ldots,\ j-1,\ j)' .$$

Clearly then, if

$$\left.\begin{array}{l} \text{(i)} \quad \Pr\left\{\begin{matrix}\text{CS i selected as}\\ \text{start-point}\end{matrix}\right\} > 0 \quad ; \\ \\ \text{and} \\ \text{(ii)} \ t_{j+}(r) \le T(r) \quad , \quad j = 1,2,\ldots,N(r) \end{array}\right\} \quad \ldots (4)$$

then this is sufficient, but not necessary, to guarantee that

$$\Pr\left\{\begin{matrix}\text{CS j in region r}\\ \text{adequately surveyed}\end{matrix}\right\} > 0 \quad \text{(as required)} \quad .$$

In the case that survey efforts can be completed in the required period of time, conditional on the sample SSUs and field staff,

$$\Pr\left\{\begin{matrix}\text{CS j in region r}\\ \text{adequately surveyed}\end{matrix}\right\} = 1$$

regardless of what randomization was employed. If conditional on the sample SSUs and field staff the survey efforts cannot be completed, however, the conditional probabilities for being surveyed depend, by (3), on the parameters.

$$\left\{ \left(\Pr\left\{\begin{matrix}\text{CS j in region r}\\ \text{selected as start-}\\ \text{point}\end{matrix}\right\},\ t_{j+}(r)\right) : 1 \le j \le N(r). \right\}$$

Unfortunately, the time-on-site information (i.e., $t_{j+}(r)$: $j = 1,2,\ldots N(r)$), will only be known for surveyed clusters. Three options are available:[9]

(1) Determine all missing values for time-on-site after completion of survey.

(2) Estimate missing time-on-site parameters based on observed data (possibly under pooling of data from all regions).

(3) Determine probability of being surveyed (conditional on sample SSUs in region, optimum sequencing, and field staff) conditional on number of cluster sites in which data collection efforts were completed (say k(r)).

$$\text{i.e., } \Pr \begin{matrix}\text{cluster j}\\ \text{adequately}\\ \text{surveyed}\end{matrix} \doteq \sum_{\substack{\ell:\ j-\ell < k(r) \\ \text{in } S_j}} \gamma_\ell(r).$$

Clearly, Option 1 exists only in theory. Further, a general preference can be expressed for Option 2 (if possible), leaving Option 3 to be implemented only if inadequate data is furnished to support the underlying modelling effort. Certainly Option 3 merely assumes that k(r) has a degenerate distribution whereas Option 2 allows for the possibility that k(r) may depend on the start-point. Predicting the missing time-on-site parameters allows the distribution of k(r) to be estimated (i.e., summary of $\hat{k}(r)$-values over possible start-points) which in turn allows the required $\{p_{j+}(r) \ : \ j = 1,2,\ldots,N(r)\}$ to be approximated. Even without further information from the field, data on sample PSUs (SSUs) relevant to the task at hand may already exist. For example, in the Bolivian Demographic Survey, information was available on the sampling frame concerning:

1. Estimated number of housing units (households, residents) by PSU (SSU).

2. Presence/absence of urban population center(s) in PSU (if applicable).

3. Estimated proportion of rural housing units associated with population centers in PSU (if applicable).

Moreover, little effort would be required to determine:

4. Land area of PSU (SSU).

5. Extent and quality of road system in PSU (SSU).

6. Major geographic barriers in PSU (SSU).

7. Subjective assessment specific to PSU (SSU) of:

   (i) prevalence of housing units speaking only an Indian dialect;

   (ii) willingness of households to cooperate (based on anticipated strength of community leader, knowledge about whether residents might be expected to be off at market or harvesting sugar cane crop, etc.).

Furthermore, design efforts attempted to make the housing unit sample size in each PSU equal (by urban-rural) so that differences in the time-on-site parameters (by urban-rural) should primarily be due to these types of factors under consideration and variations within urban-rural components should ideally be small (supporting the simplicity of Option 3).[10]

c. Operational Considerations

To operationalize the notion of a randomized start-point in each region, regional supervisors were asked to review the sample PSUs and guesstimate expected workload (by PSU). Cluster sites of approximately equal workloads[11] were then formed by combining neighboring sample SSUs (equivalently, by combining PSUs since only one SSU per sample PSU was selected in the survey). Travel considerations subsequently dictated the efficient sequencing of cluster sites, leaving only the selection of the start-point to finalize field scheduling, which might reasonably be done at random or proportional to the expected workloads.

It is probably worth noting that apart from an initial misunderstanding that allowed for completely randomized scheduling, few reservations were expressed by country officials relative to their willingness/ability to operate within the specified guidelines. In retrospect, however, greater attention should have been paid to field documentation requirements, particularly with respect to any realized deviations from the proposed scheduling of cluster sites.

d. Analytical Considerations

In the event that field work cannot be completed in a region, partial data will be more clustered than would otherwise be desired. Furthermore, adjusting for this nonresponse in the previous manner will cause unequal weighting effects to be introduced into an otherwise self-weighting design like that intended for the Bolivian Demographic Survey. Finally, the design will exhibit a reduced capability to approximate the precision of resulting estimates. In this latter regard, it should be clear the the conditions in (4) do nothing to address the need for pairwise positive joint realization probabilities for sample PSUs (i.e., obvious extension of (2)). Indeed, start-pint randomization intentionally minimizes the chance of realizing same by not deviating from the minimum cost travel sequence. Clearly, however, second-order estimability could be retained (i.e., assuming sample design provided for same) by permitting greater disruptions to the minimum-cost travel sequence. For example, the extreme case of same occurs when field assignments are randomly sequenced. As in all design work, costs must be traded-off against capabilities. In the case of the Bolivian Demographic Survey where only one PSU/stratum was selected at the first-stage of the design the precision of estimated totals was to be approximated by collapsing adjacent strata to form pseudo-replicates. In the event of premature termination of the survey, it was intended to form such pseudo-replicates among the realized PSUs (i.e., design strata). In situations where the design provided for direct replication, field work could of course be sequenced according to same, and the precision approximated using the number of replicates completed during implementation of the design.

4. Framework for Scheduling Workloads Within a Cluster Site

The randomizations proposed in the previous section refer to within cluster site affairs only insofar as

1. The time-on-site to "adequately complete" the associated survey effort;

and

2. The requirement that every target population member be given a positive chance of being included in the survey.

The intent of the current section will be to examine these notions in the context of the proposed survey. Each topic will be addressed in turn.

a. Defining of "adequately complete"

In spite of efforts expended to obtain a usable questionnaire from every household in each sample SSU, household nonresponse will undoubtedly occur in the survey. (e.g., household may be off harvesting sugar cane, on vacation, or merely at a distant market and not available). As such, it is pointless to impose a field procedure that calls for a field team to remain on-site until all questionnaires are completed without exception. Furthermore, the projected workloads for the study were such that only minor deviations from the anticipated time-one-site could be accommodated without seriously affecting the ability of field teams to survey all sites in their region. To address this concern, consideration was given to requiring that field teams not exceed a pre-specified maximum number of days at a given site.

To estimate the need for such a procedure, regional supervisors were asked to review information available on each sample SSU in their region and arrive at some estimate of the minimum and maximum time-on-site that might be required (most likely estimate of time on site was earlier used to form cluster sites of approximately equal workloads). Once in-hand, it was proposed to develop a heuristic rule which would have attempted to obtain at least partial data on on minimum of, say 90% of all PSUs (by urban-rural) under an assumption of maximum projected time-on-site requirements. The reasoning underlying this rule was simply that of attempting to maximize the geographic dispersion of whatever partial data was realized under the current survey. Two weaknesses of such a scheme were recognized at the outset:

1. A field team might prematurely terminate field work at a cluster site under the maximum time-on-site rule only to complete all sites with time to spare (and hence requiring the team(s) to return to an earlier area as time permits).

2. The true time-on-site for cluster site j of region r, $t_{j+}(r)$, may be in excess of the maximum time allowable, and hence no guarantee can be made that all target population members will be surveyed at such a site.

The first concern is a fact of life and underscores the importance of carefully choosing the maximum time-on-site parameter for the study. The second concern is magnified by the potential for field staff to ignore "difficult" interviews whenever it is known that survey operations will cease after a given period of time, causing their problem to no longer exist. Unfortunately, the nonresponse biases introduced by such systematic exclusions linger on to plague the ultimate analysis and interpretation of study findings. This reality will be the topic of discussion in the next subsection.

b. Coverage of Intended Target Population

Adherence to a strict probability design requires that every household in each sampled SSU be given a positive chance of being interviewed for the survey. Under a design calling for a maximum time-on-site, however, this property may not always be realized in the field implementation of the study. To address this concern, sample SSUs could be partitioned into sub-units having

(i) well-defined separating boundaries; and

(ii) maximum projected workloads smaller than the ceiling placed on time-on-site for this sample SSU.

A probability scheme could then be applied to these sub-areas to arrive at the ultimate schedule for surveying the sample SSU (i.e., order in which sub-areas would be surveyed). For the purposes of the current survey, two considerations were dominant:

a. Desire to have any sub-area randomizations carried out external to the field teams; and

b. Information on expected workloads below the segment level were not available during the design phase.

In light of these realities, the decision was made to schedule workloads in sample SSUs via a random permutation of the component segments.[12] Alternatively, regional supervisors could have been used to either once and for all reduce to workload via sub-area subsampling, or by imposing an alternative randomized workload schedule on these sub-areas.

5. Concluding Remarks

This paper has attempted to make a case for the potential use of randomization in formulating field assignments. Methodology developed was in response to a design situation in which there was little ability to make coordinated adjustments during the actual conduct of the survey and in which there were substantial uncertainties associated with such conduct (e.g., political instability; response rate variability; logistical difficulties; abilities of interview staff, etc). As such, it was intended to merely introduce the notion of field assignment randomization as a potential mechanism for addressing nonresponse bias caused through premature termination of the survey. Accordingly, no attempt has been made to impose formal cost/total error modelling to arrive at an optimal degree of randomization for a given level of nonresponse. Further, methodology was purposely chosen to be simple and unimposing, giving rise to a minimum of resistance from the operational staff of the host country. Unfortunately, a military coup forced the abandonment of this approach in the current survey, further emphasising the true need for such procedures in similar studies.

More importantly, methodology presented is easily seen to be applicable to a much larger class of designs involving sequential stopping rules which are a function of sample attributes (e.g., cost, precision, time). Such designs are not uncommon in the statistical literature (e.g., random digit dialing telephone surveys; sample selection in waves to better estimate underlying sample size parameters involving screening and response rates, etc.). Moreover, factors such as escalating survey costs, rapid turnaround requirements, and the ability to quickly convert field data to computer readable form should further increase interest in such methods. In this regard, it is perhaps of particular note that the methodology presented herein has attempted to model the dynamics of the stopping rule in a complex survey environment and to take partial account of same in the subsequent weighting of the survey data. Ideally, of course, one would like to go even further and compute estimates of precision which reflect variation induced by the stopping rule.

BIBLIOGRAPHY

[1] Drummond, Douglas J. [1980]. "Design and Selection of the First-Stage Sample Units in Support of the Bolivian Demographic Survey". Trip Report No. 37 submitted to Population Laboratories, University of North Carolina, Chapel Hill, N.C.

[2] Drummond, Douglas J. [1980]. "Design and Selection of Second - and Third-Stage Sample Units in Support of the Bolivian Demographic Survey. Trip Report No. 44 submitted to the Population Laboratories, University of North Carolina, Chapel Hill, N.C.

REFERENCES

[1] Each field team consisted of four interviewers and a supervisor.

[2] Planning for the proposed Census of Agriculture precluded any extension of the data collection period.

[3] In practice, $t_j(r)$ could vary under repeated observation on the same SSU by the same field team but will be taken as fixed for the purposes of this discussion. Moreover, "adequately complete" refers to some pre-determined rule for terminating the field effort at every site (includes case where field team stays on site until all data collection is complete).

[4] Without loss of generality, we shall assume that PSUs are sequentially numbered according to such a schedule.

[5] Workload bias should ideally be defined in the absence of any other sources of nonresponse (i.e., if reasonable attempt made to survey target population member then individual responds). In practice, workload is a catalyst for other sources of nonresponse, a subtlety not addressed in this paper.

[6] Cluster sites were formed by looking at the geographic distribution of sample SSUs and attempting to group "close" SSUs so as to realize equal parcels of work.

[7] The extent to which randomization is allowed to distort the otherwise efficient sequencing of field sites should be a function of the "cost" and potential for workload bias relative to incremental costs associated with the distorted schedule. For the current survey, a decision was made to accommodate the efficient sequencing of sites while preserving some residual ability to reduce workload bias.

[8] Certainly this requirement could be relaxed in order to trade-off field costs with nonresponse bias and/or second-order estimability

[9] Strong consideration might also be given to merely re-weighting an intended self-weighting sample via post-stratification ratio adjustment to known totals when such data are available.

[10] This assumes that cluster sites are formed so as to represent equal expected workloads (which may not solely depend on the housing unit sample size). The use of Option 2 would be mandatory whenever large workload differences are known to exist between cluster sites.

[11] This may not always be possible but is highly desirable. Documentation should be maintained on the estimated workloads for use in implementing Option 2 of the previous subsection.

[12] Each penultimate sampling unit consisted of a cluster of contiguous land areas defined to be "segments" according to the 1976 Bolivian Census of Population.