DESIGN STRATEGIES FOR NONSEDENTARY POPULATIONS

William D. Kalsbeek, University of North Carolina

1. INTRODUCTION

Migratory segments of the population in many countries continue to be of considerable interest to the survey researcher. These segments (or "nonsedentary populations" as we shall call them in the sequel) occur in many different forms throughout the world. They are, for example, the migrant seasonal farm workers, the transients, the chronic homeless, and the teenage runaways in the United States; the gypsies of Eastern Europe; the street people of Indonesia; and the nomadic tribesmen of East Africa.

This paper examines the problem of sampling nonsedentary populations. It does so by first suggesting that to characterize a dynamic, mobile population such as these may occasionally require a design in which both space and time are sampled. Statistical and cost implications of four space-and-time design strategies are then formulated for the problem of estimating the size of the nonsedentary population. This is followed by an application of these results to the specific problem of sampling migrant seasonal farm workers in the United States. We conclude from this illustration that independently picking a stratified sample of days at each of a sample of migrant camps will be the most cost-efficient design strategy among those considered.

2. PRIOR SAMPLING STRATEGIES

Most prior surveys of nonsedentary populations have enumerated them as they appear at a sample of places where they reside during the study period. For example, Fernandez and Folkman (1975) and Chi (1985) have used labor camps as primary sampling units (PSUs) in multi-stage designs to sample migrant seasonal farm workers. In similarly structured designs, the Ministry of National Planning for Somalia (1981) established watering points as PSUs to sample nomads, and Frankel (1986) used shelters, parks, streets and the like to sample the homeless in Chicago.

The final sample in each of these designs was chosen by identifying population members linked to the PSU at operationally convenient (but not randomly chosen) times during the data collection period. Randomization in these sampling designs is clearly limited to the spatial dimension, which implies that multiple frame linkages (i.e., multiplicity) exist during the study period and therefore must be accounted for in the estimation process. Moreover, the number and complexity of these linkages increase directly with the length of data gathering. Kalsbeek and Cross (1982) have identified the sources of multiplicity in sampling East African nomads, and Kalsbeek (1986) has examined the properties of two alternative design strategies in this context. An extensive literature on multiplicity estimators, beginning with the work of Birnbaum and Sirken (1965) and Sirken (1970), addresses the matter of dealing with the multiple linkage issue.

In addition to the statistical implications mentioned above, extended periods of data collection may cause one difficulty in trying to capture the dynamic quality of nonsedentary individuals. One good example is the population of migrant seasonal farm workers whose size, composition and geographic distribution is known to change dramatically over a 12 month period (Johnston, 1985). Workers move within well-known migratory streams to where the varying seasons among states provide a crop to be picked, occasionally returning to their homeland to visit family and friends. This mobility is important when the measures of interest in this population are tied to an individual's surroundings (e.g., health status, health care availability). A study conducted during the Spring may paint a quite different picture than one done in the Fall.

The designer of a one-time survey of a nonsedentary population is therefore faced with a fundamental dilemma. Should the reference period of the study be shortened to limit or avoid multiplicity, or should the study period be expanded to encompass all of the seasonal variations in behavior? It is the premise of this paper that, in some sense, we can "have our cake and eat it too" by employing a design in which both space and a fully expansive study period are sampled together so that variation along both dimensions can be examined through the sampling process while avoiding the problems of multiple linkage.

3. CONCEPTUAL FRAMEWORK AND PARAMETER

To characterize a nonsedentary population over time let us define a three-dimensional matrix, $\mathbf{X_0}$, defined by the cross-classification of the following:

1) An array made up of L sampling units constituting a frame for sampling spatially,

$$\mathbf{S} = (S_1, S_2, \ldots, S_i, \ldots, S_L)$$

plus the state, S_0, used to denote detachment from this frame (e.g., a migrant worker in transit between jobs or returning home to Mexico for a few months during the off-season);

2) An array of M time units (e.g., days) constituting the time period for the study and from which the sample in time is drawn,

$$\mathbf{T} = (T_1, T_2, \ldots, T_j, \ldots, T_M);$$

3) An array of N eligible members of some population being studied,

$$\mathbf{P} = (P_1, P_2, \ldots, P_k, \ldots, P_N).$$

Entries in $\mathbf{X_0}$ are denoted by X_{ijk}. The three dimensional array consisting of the cross-classification of **S**, **T** and **P** (i.e., excluding S_0 from the frame dimension of $\mathbf{X_0}$) is denoted by $\mathbf{X}$. It is assumed that during any time unit (T_j), every population member (P_k) will be linked to exactly one member of the frame (**S**) or to S_0. In addition, **P** is assumed to be an all-inclusive set of eligible population members during the study period defined by **T**.

Although totals, means, proportions and other more complex time-based parameters can be defined within $\mathbf{X_0}$, we will limit our attention here to the problem of estimating the population's size, N. To formulate this parameter, the entries of $\mathbf{X_0}$ are defined as count variables,

$$X_{ijk} = 1 \quad \text{if } P_k \text{ is linked to } S_i \text{ during } T_j\text{, and}$$
$$\quad\;\; = 0 \quad \text{if otherwise.}$$

The object as stated is to estimate

$$N = \sum_{i=0}^{L} \sum_{j}^{M} \sum_{k}^{N} X_{ijk}/M = L\,\bar{X}_0 ,$$

where $\sum_j \equiv \sum_{j=1}^{M}$, $\sum_k \equiv \sum_{k=1}^{N}$, and

$$\bar{X}_0 = \sum_{i=0}^{L} \sum_{j}^{M} \sum_{k}^{N} X_{ijk}/LM.$$

4. SAMPLING DESIGNS

The four sampling designs described below presume that population members cannot be sampled while in the detached state (S_0) and that the aggregate count of the number of population members linked to the i-th spatial sampling unit as of the j-th temporal sampling unit,

$$X_{ij} = \sum_{k}^{N} X_{ijk} ,$$

can be determined for any combination of spatial and temporal sampling units. Moreover, to simplify formulations for components of the mean square error, with-replacement simple random sampling is done at each of the various selection steps of each design.

We note that sampling jointly in **S** and **T** bears some resemblance to lattice or plane sampling as used in agricultural research (see, for example, Bellhouse, 1977, and Iachan, 1985). There, however, both dimensions are spatial and autocorrelation is expressible within the two-dimensional plane from which the sample is drawn. In the present setting the temporal is the only dimension within which any correlation is likely to exist.

<u>Unrestricted Random Sampling (U):</u>

Each member of the unrestricted random sample (URS) of size l*m (i.e., l* times m) is chosen from the LM members of **S** by **T** is selected by picking one spatial sampling unit at random and then choosing a temporal sampling unit at random to go with it. Subsequent selections are made without regard to prior selections (i.e, <u>with</u> replacement). This design has the advantage of avoiding the negative statistical implications of sampling time clusters, where intra-cluster correlation may be high.

<u>Unstratified 2-Stage (2S):</u>

A URS(l* of L) spatial sampling units is chosen as the primary sample; a URS(m of M) temporal sampling units is independently selected within each sample PSU. This design limits to l* the number of spatial sampling units that must be visited during data collection.

<u>Substratified 2-Stage (2SS):</u>

A URS(l* of L) spatial sampling units is chosen in the first stage; a proportionate stratified URS of size m is chosen from H_i strata formed in the i-th PSU. This design compensates for the losses due to cluster sampling by gains due to stratification in the second stage of sampling. Stratification is <u>not</u> used in the first stage of sampling, although when reasonable predictors of significant between-PSU variation exists, it might be used as well.

<u>Unstratified 2-Way (2W):</u>

The same URS(m of M) sample of temporal sampling units is used for each member of a URS(l* of L) sample of spatial sampling units. This design has the potentially useful feature of having data gathering at the same time points in all selected spatial units. It represents an effort to coordinate the timing of selected time units among spatial units, which in some instances might be useful (e.g., when special preparations are needed for collecting data at each time point).

5. ESTIMATION

Since each of the four designs yields a sample size of l*m as well as equal selection probabilities for each of LM cells in **S**, a common (though biased) estimator of N would be,

$$\hat{N} = L\,[\sum_{i}^{l^*} \sum_{j}^{m} X_{ij}/l^*m] = L\,\bar{x} ,$$

where $\sum_i \equiv \sum_{i=1}^{l^*}$, $\sum_j \equiv \sum_{j=1}^{m}$, and $\bar{x}$ is the overall mean among the measures of X_{ij} taken for the l*m space-x-time units in the sample.

6. RELATIVE BIAS OF $\hat{N}$

The bias of $\hat{N}$ arising from coverage error in the frame **S** is the same for each design and can be obtained by noting that (over all possible samples in space and time)

$$E(\bar{x}) = \sum_{i=1}^{L} \sum_{j}^{M} X_{ij}/LM ,$$

so that the relative bias of $\hat{N}$ is,

$$\text{Rel-Bias}(\hat{N}) \equiv \text{Bias}(\hat{N})/N = - \sum_{j}^{M} X_{0j}/NM , \qquad (1)$$

which means that, relative to N, $\hat{N}$ is a biased underestimate of N by an amount equaling the average proportion of the population in the detached state among time units.

In Eq.(1) we see the first evidence of how the nature of mobility in a nonsedentary population has an impact on the properties of estimators used in conjunction with these four designs. All else constant, we note that a population which tends to be detached from the spatial sampling frame frequently and for longer intervals will tend to present greater problems of underestimation than one where periods of detachment are less frequent and lengthy. Thus $\hat{N}$ will be a greater underestimate for the nomadic tribes of East Africa, with their longer periods away from watering points, than for the migrant seasonal farm workers in the United States periods of travel between jobs are relatively short and returns to their home countries infrequent.

7. VARIANCE OF $\hat{N}$ AND DESIGN EFFECTS

To formulate the variances of $\hat{N}$ under the four designs, we must first define the following measures of variance:

$$\sigma^2_{BL} = \sum_{i=1}^{L} (\bar{X}_i-\bar{X})^2/L \,, \qquad \text{(Between-Space)}$$

$$\sigma^2_{WL} = \sum_{i=1}^{L} \sum_{j}^{M} (X_{ij}-\bar{X}_i)^2/LM \qquad \text{(Within-Space)}$$

$$\sigma^2_{BM} = \sum_{j}^{M} (X_j-\bar{X})^2/M \qquad \text{(Between-Time)}$$

$$\sigma^2_{WM} = \sum_{i=1}^{L} \sum_{j}^{M} (X_{ij}-\bar{X}_j)^2/LM \qquad \text{(Within-Time)}$$

$$\sigma^2 = \sum_{i=1}^{L} \sum_{j}^{M} (X_{ij}-\bar{X})^2/LM \,, \qquad \text{(Overall)}$$

$$= \sigma^2_{BL} + \sigma^2_{WL} = \sigma^2_{BM} + \sigma^2_{WM} \,,$$

$$\delta_L = \sigma^2_{BL}/\sigma^2 \,, \qquad \text{(Relative Homogenity within Spatial Clusters)}$$

$$\delta_M = \sigma^2_{BM}/\sigma^2 \,, \qquad \text{(Relative Homogenity within Temporal Clusters)}$$

where $\bar{X}_i = \sum_{j}^{M} X_{ij}/M$ and $\bar{X}_j = \sum_{i=1}^{L} X_{ij}/L$. Note that the relative homogeneity measures are comparable but not equivalent to the usual measures of intracluster correlation in that $0 \le \delta_L \le 1$ and $0 \le \delta_M \le 1$.

URS(1*m of LM)

The variance of $\hat{N}$, obtained from $Var(\bar{x})$, is

$$Var_U(\hat{N}) = L^2\sigma^2/1^*m \,. \qquad (2)$$

Unstratified 2-Stage

Once again the variance of $\bar{x}$ is known from the standard 2-stage framework, here with spatial sampling units as PSUs from which a URS(1* of L) are selected, and temporal sampling units as secondary sampling units from which a URS(m of M) are independently selected in each sample PSU.

The design effect for $\hat{N}$ given this 2-stage design, determined as its variance, relative to the variance of the estimator of N for a URS(1*m of LM), will be

$$DEFF_{2S}(\hat{N}) = 1 + \delta_L(m-1) \,. \qquad (3)$$

Substratified 2-Stage

When a proportionate stratified URS of size m replaces the URS(m of M) in the second stage of the 2-stage design above, the variance of $\bar{x}$ can be expressed as,

$$Var(\bar{x}) = \sigma^2_{BL}/1^* + \{\sum_{i=1}^{L} (1-\delta_{iH})\sigma^2_i/L\}/1^*m \,,$$

where $\delta_{iH} = \sigma^2_{iBH}/\sigma^2_i$ measures the effectiveness of substratification in the i-th spatial sampling unit, where H_i substrata are formed,

$$\sigma^2_{iBH} = \sum_{h=1}^{H_i} W_{ih}(\bar{X}_{ih}-\bar{X}_i)^2 \,,$$

is the between-substratum variance in the i-th PSU,

$$\sigma^2_i = \sum_{h=1}^{H_i} \sum_{j}^{M_{ih}}(X_{ihj}-\bar{X}_i)^2/M \,,$$

is the total within-cluster variance for the i-th PSU, $W_{ih}=M_{ih}/M$ is the proportion of time units in the i-th PSU that fall in its h-th substratum, $\bar{X}_{ih} = \sum_{j}^{M_{ih}} X_{ihj}/M_{ih}$, and

$$\bar{X}_i = \sum_{h=1}^{H_i} \sum_{j}^{M_{ih}} X_{ihj}/M = \sum_{h=1}^{H_i} W_{ih}\bar{X}_{ih} \,.$$

When the efficiency of substratification and σ^2_i are uncorrelated among PSUs, then one can express the design effect under the substratified 2-stage design from the $Var(\bar{x})$ given above as

$$DEFF_{2SS}(\hat{N}) = 1 + \delta_L(m-1) - \bar{\delta}_H(1-\delta_L) \,, \qquad (4)$$

where $\bar{\delta}_H = \sum_{i=1}^{L} \delta_{iH}/L$.

Unstratified 2-Way

The variance here is found by reformulating the overall sample mean of the X_{ij}'s as

$$\bar{x} = \sum_{i=1}^{L} \sum_{j}^{M} \Theta_i\Theta_j X_{ij}/1^*m \,,$$

where Θ_i and Θ_j are, respectively, the number of times that the i-th spatial sampling unit and the j-th temporal sampling unit are chosen. From this we obtain

$$\mathrm{Var}(\bar{x}) = \sigma^2_{WL}/m + \sigma^2_{WM}/1^* - \{1^*+m-1\}\sigma_{LM}/1^*m ,$$

where

$$\sigma_{LM} = \sum_{i=1}^{L} \sum_{j}^{M} (X_{ij}-\bar{X}_i)(X_{ij}-\bar{X}_j)$$

is the space-time interaction for the X_{ij}'s. Finally, the design effect for the 2-way design as used to estimate N can be shown as

$$\mathrm{DEFF}_{2W}(\hat{N}) = 1 + \delta_L(m-1) + \delta_M(1^*-1). \quad (5)$$

We note from Eqs.(3)-(5) that, when considering variances alone, the substratified 2-stage design is always preferable to an unstratified 2-stage design which, in turn, is always preferable to the unstratified 2-way design. The ranking of URS(1*m of LM), relative to the substratified 2-stage design,

depends on values of δ_L, $\bar{\delta}_H$, and m. More specifically, $\mathrm{DEFF}_{2SS}<\mathrm{DEFF}_U$ when

$$\bar{\delta}_H > \delta_L(m-1)/(1-\delta_L) . \quad (6)$$

8. SURVEY COSTS

Having assessed the primary statistical implications of the four designs, let us now turn our attention to how each design would affect the cost of the survey operation. To do so, we need a model to express these costs. One simple formulation of costs incurred under these designs is the following:

$$C = C_o + \sum_{i=1}^{L} \varphi_i C_{Li} + \sum_{i=1}^{L} \sum_{j}^{M} \varphi_{ij} C_{Mij} , \quad (7)$$

where C_o denotes fixed costs (e.g., instrument development, administration, reporting), C_{Li} represents those costs that are particular to the i-th spatial sampling unit (e.g., solicitation, set-up, certain sampling activities), C_{Mij} denotes the cost of survey activity (e.g., data collection and processing) tied to the j-th time unit in the i-th spatial unit,

φ_i = 1 if ≥1 time units are chosen in the i-th spatial unit;
= 0 if otherwise, and

φ_{ij} = 1 if the j-th time unit in the i-th spatial unit is chosen ≥1 times;
= 0 if otherwise.

Allowing $1-\{1-1/L\}^{1^*} \approx 1^*/L$ and $1-\{1-1/M\}^m \approx m/M$, the expected cost of the survey under the URS(1*m of LM) design will be

$$E_U(C) \approx C_o + 1^*m(\bar{C}_L + \bar{C}_M) , \quad (8)$$

where $\bar{C}_L = \sum_{i=1}^{L} C_{Li}/L$ and $\bar{C}_M = \sum_{i=1}^{L} \sum_{j}^{M} C_{Li}/LM$.

Allowing the same approximations for 1*/L and m/M, once again, the expected costs for the other three designs will be

$$E_{2S}(C) \approx E_{2SS}(C) \approx E_{2W}(C) \approx C_o + 1^*\bar{C}_L + 1^*m\bar{C}_M . \quad (9)$$

We note from Eqs. (8) and (9), that the URS(1*m of LM) design will have non-fixed costs (i.e., excluding C_o) that exceed comparable costs under the other designs by a factor of

$$\text{REL-COST} = 1 + R_{LM}(m-1)/(R_{LM}+m) . \quad (10)$$

where $R_{LM}=\bar{C}_L/\bar{C}_M$ is the ratio of average spatial to temporal unit costs.

9. COST-EFFICIENCY

The overall measure of effectiveness adopted for use in assessing each design (*) in terms of its joint statistical and fiscal impact is

$$\mathrm{CEFF}_*(\hat{N}) = \frac{\{\mathrm{Var}_*(\hat{N})\}^{-1}}{E_*(C)-C_o} ,$$

whose numerator reflects the statistical precision obtainable from the design and denominator accounts for the non-fixed, or variable, component of survey costs over which the designer has some control.

Because our real interest is in comparing the cost-effectiveness of the four designs <u>when 1* and m are the same</u>, we choose to examine CEFF for any given design relative to the measure of CEFF for the URS(1*m of LM) design; i.e., we use

$$\mathrm{RCEFF}_*(\hat{N}) = \mathrm{CEFF}_*(\hat{N})/\mathrm{CEFF}_U(\hat{N}) = \text{REL-COST}/\mathrm{DEFF}_*(\hat{N}) \quad (11)$$

as the final basis for comparison among the four designs.

10. ILLUSTRATION: MIGRANT FARM WORKERS

We choose to illustrate our findings by considering the feasibility of the four designs for estimating the number of migrant seasonal farm workers in the United States during a one year period. Translating the general conceptual framework of the earlier section into this particular setting, the penultimate spatial sampling units are presumed to be migrant camps, although technically other residential areas inhabited by high concentrations of migrant workers would be included to improve sample coverage. The temporal sampling units are individual days, and each value of X_{ij}, the headcount of migrant workers on a specific day at a specific camp, are obtained by a visit to the camp on that day. Days are selected from the growing season and from the off-season, even though enrollment in the camps would be much lower and limited to those migrants with more permanent work in the area.

Because there exists little direct information on the size of other key design parameters, we must rely on quasi-empirical evidence to determine values of

δ_L, δ_M, R_{LM}, $\bar{\delta}_H$, m, and 1*. One key piece of evidence has to do with the pattern (not distribution) of the daily census of each camp (i.e., X_{ij}) from the first to last day of the year, since from this pattern one can obtain σ_i^2, σ_{ih}^2, and ultimately values for $\bar{\delta}_H$ and δ_L. Standardized means and variances for two likely patterns are presented in Figure A.

The "partial square pattern" presumes that the camp is occupied at full capacity for 100ε percent of the growing season, which occurs for 100α percent of the year. This pattern is especially likely in areas where much of the work is done by crews, which arrive together early in the season and finish together near season's end.

The "trapezoid pattern" presumes a peak season occurring for 100β percent of the growing season and a head count that gradually increases to full capacity at the start of the peak growing season and then diminishes in like manner toward the end. This pattern is thought to be common in "home-base" states like California, Texas, and Florida where migrants may return to semi-permanent residences around the peak of the growing season in those states. These residences may also serve as bases of movement out to other states for their growing seasons.

Values for δ_L:

The following two key assumptions are made in arriving at our estimate of δ_L=0.51:

(1) All camps follow a partial square pattern with ε=1 with an average peak enrollment per camp of 75 persons, and an average growing season of 9 months (Johnston, 1985); and

(2) The distribution of the peak enrollment of all camps is asymmetrically triangular with a range of from 0 to 200 and a mean of 75.

The values δ_L = 0.4, 0.5 and 0.6 are subsequently used in this illustration, since the actual figure could be higher or lower than the computed value, depending on the frequency and shape of patterns, other than the square pattern, that would appear.

Values for δ_M:

Here it is thought that δ_M must be quite small, since variation in the average aggregate enrollment in camps $(\bar{X}_j)$ over time is likely to be small. This reasoning implicitly assumes that the total number of migrants in the detached state (S_0) will not vary much from one day to the next in the course of a year. The values δ_M = 0.01, 0.03 and 0.05 are therefore used.

Values of $\bar{\delta}_H$:

The likely effectiveness of substratification in an individual camp (δ_{iH}) will depend on the camp's enrollment pattern during the year and on the number and definition of substrata. For the partial square pattern (see Figure A) where there are two substrata, one covering the length of the growing season and the other spanning the rest of the year,

$$\delta_{iH} = 1 - \{(1-\varepsilon)/(1-\alpha\varepsilon)\} \; . \tag{12}$$

(Although using the actual periods of maximum enrollment to define substrata would be preferable, it is unlikely that this information would be available at the time of sampling.) In camps where enrollment follows a trapezoidal pattern and three substrata are formed, one covering the peak period, a second the off-peak portion of the growing season, and the third the rest of the year,

$$\delta_{iH} = \frac{3\{(3\beta+1)-\alpha(\beta+1)^2\}}{4(2\beta+1)-3\alpha(\beta+1)^2} \; . \tag{13}$$

Table A presents δ_{iH} for various values of α, β and ε. Since the majority of camps following the partial square pattern will be in non-"home-base" states where most growing seasons are 6-9 months long, we see that δ_{iH} is likely to exceed 0.5 there. Camps in home-base states have seasons nearly year-round which would imply δ_{iH} between 0.3 and 0.5. Finally, assuming that 60-80 percent of camps follow the partial square pattern, it seems plausible that the overall measure of the substratification efficiency $(\bar{\delta}_H)$ might comfortably be encompassed by the values 0.5, 0.7 and 0.9.

Values of R_{LM}:

If survey cost is expressed in person-days of effort and one can reasonably assume that it takes roughly one person-day of effort to visit a camp on a selected day and to process the measure of X_{ij} through analysis, then $R_{LM}=\bar{C}_L$, the average number of person-days needed to add a camp to the sample.

The size of $\bar{C}_L$ will depend on several things. First, since the sample of camps is likely to be chosen through some multi-stage process, part of the unit cost for camp will depend on the amount of effort expended in developing lists of existing migrant camps to be used as sampling frames. Prior experience has shown that frame construction can be very costly if the object is to achieve high coverage rates for these frames. A second determinant of $\bar{C}_L$, related to the first, is the number of such frames to be constructed, which would depend on the allocation of the spatial sample among the stages identified for selection. Clearly, $\bar{C}_L$ would vary directly as the number of such frames to construct. Finally, $\bar{C}_L$ would be directly affected by the targeted response rate for selected camp as well as the levels of effort expended in training, supervision, quality control and the like.

Given that R_{LM} could be high or low, depending on the priorities of the study, the values used in our illustration reflect this uncertainty. Values of R_{LM} ranging from five to 50 are used for subsequent computations.

Values of m:

The values chosen for m in the illustration were based on the cost-times-variance optimum values of m that would arise from the three non-URS designs and the model for the variable components of their corresponding expected costs as presented in Eq. (9). Optimum values for these designs can be determined for the unstratified 2-stage design (widely used) as

$$m_{2S}^{(opt)} \doteq [(1-\delta_L)R_{LM}/\delta_L]^{1/2}, \quad (14)$$

for the substratified 2-stage design as

$$m_{2SS}^{(opt)} \doteq [(1-\bar{\delta}_H)(1-\delta_L)R_{LM}/\delta_L]^{1/2}, \quad (15)$$

and for the unstratified 2-way design as

$$m_{2W}^{(opt)} \doteq [\{(1-\delta_L)+\delta_M(1^*-1)\}R_{LM}/\delta_L]^{1/2}. \quad (16)$$

Optimum values of m for both versions of the 2-stage design are presented in Table B, where we note that $4 \leq m \leq 8$ is generally optimum for the unstratified 2-stage design and that $2 \leq m \leq 4$ is often preferable for the substratified 2-stage design. Optimum values of m, subject to the availability of 5,000 person-days of variable costs and C_M = one person-day, are presented for the unstratified 2-way design in Table C. There one notes that $10 \leq m \leq 18$ generally covers the range of optimum values.

The values, m = 2, 6 and 10, used in later computations are intended to represent the values most likely to be effective under each of the designs other than URS(1*m of LM).

Findings on Relative Cost-Efficiency:

Table D presents values of the cost-efficiency of each design relative to URS (i.e., RCEFF). Values less than 1 indicate that the URS(1*m of LM) design is more cost-efficient by our criteria, while those greater than 1 point to the referent design being preferable to URS. Relative superiority and inferiority among the non-URS designs can also be gauged using these entries.

Several potentially useful findings can be inferred from Table D. First, the substratified 2-stage design is generally the most cost-efficient among the four designs considered. Its preference is due mainly to its lower variance than the other non-URS designs, with which it shares notably lower non-fixed costs than the URS design. As expected, its strongest showing overall occurs when m is relatively small. The unstratified 2-stage is preferable to URS in the majority of instances, implying that substratification is not neccessarily needed to counteract the substantial variance increase due to cluster sampling with large δ_L. Second, the two 2-stage designs are most similar in preference and substantially superior to the unstratified 2-way design when R_{LM} is low. The overall last-place showing of the 2-way design is largely due to the size of 1* which amplifies its design effect, even with relatively small values of δ_M. The 2-way design is most competitive with the URS design when both R_{LM} and m are relatively large. The superiority of the 2-way design in this case is atttributable to its relatively moderate design effect combined with its substantial cost savings over the URS design. Finally, the only notable instance where the URS design does well is when R_{LM} is lowest and m is highest among observed values.

Discussion:

Findings in Table D generally portray the substratified 2-stage design as the one of choice among the four considered when estimating population size, given cost models where equal variable costs for the non-URS designs are much lower than comparable costs for the URS design. One must then wonder if and how these findings might be altered for other design settings. For example, how might the comparison of the two 2-stage designs been altered if the cost of stratification had been allowed to increase the variable costs of the substratified 2-stage design? Findings not presented revealed that the stratified design is still generally preferred over the unstratified design. Another facet of the assumptions of this study that must be examined is the effect of the design used to choose the sample of 1* spatial sampling units. What if, as would be expected, the sample of migrant camps in the illustration were to be chosen by a complex multi-stage process rather than URS? Here the implications are less clear-cut, although we suspect that changes in the absolute sizes of RCEFF are not likely to be great since each design would experience similar increases in both variance and cost. These claims are of course conjectural and must be substantiated by empirical data to allow one to better choose among the option given to sample migrant seasonal farm workers and other nonsedentary populations.

REFERENCES

Bellhouse, D.R. (1977). "Optimal Designs for Sampling in Two Dimensions," Biometrika, 64:606-611.

Birnbaum, Z.W. and Sirken, M.G. (1965). "Design of Sample Surveys to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimates," Vital and Health Statistics, Series 2 - Number 11, National Center for Health Statistics, U.S. Government Printing Office, Washington, DC.

Chi, P.S.K. (1985) "A Note on Sampling Migrant Farm Workers," Rural Sociologist, 5:3:158-162.

Fernandez, R.R. and Folkman, D.V. (1975). "Survey of Vocational Educational Needs of Migrants in Wisconsin," University of Wisconsin, Extension Center for Community Leadership Development, Madison, WI.

Iachan, R. (1985). "Plane Sampling," Statistics and Probability Letters, 3:151-159.

Johnston, H.L. (1985). Health for the Nation's Harvesters, National Migrant Workers Council, Inc., Farmington Hills, MI.

Kalsbeek, W.D. amd Cross, A.R. (1982). "Problems in Sampling Nomadic Populations," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 398-402.

Kalsbeek, W.D. (1986). "Nomad Sampling: An Analytic Study of Alternative Design Strategies," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 164-169.

Kish, L. (1965). Survey Sampling, John Wiley and Sons, New York.

Ministry of National Planning (1981). Report on the Demograghic Survey of Banadir, Bay, and Lower Shebelle Regions of Somalia. 1980-81, Mogadishu. Summarized in Summary Series No. 4, International Population Laboratories, Chapel Hill, NC.

Sirken, M.G. (1970). "Household Surveys with Multiplicity," Journal of the American Statistical Association, 65:257-266.

TABLE A: VALUES OF DELTA-SUB-iH FOR THE PARTIAL SQUARE AND TRAPEZOID PATTERNS IN SAMPLING MIGRANT FARM WORKERS

PROPORTION OF SEASON IN RESIDENCE (EPSILON)	PROPORTION OF THE YEAR SPANNED BY THE GROWING SEASON (ALPHA)=	NUMBER OF MONTHS IN THE GROWING SEASON: 3	4	5	6	7	8	9	10	11	12
		0.25	0.33	0.42	0.50	0.58	0.67	0.75	0.83	0.92	1.00
PARTIAL SQUARE PATTERN/ 2-STRATUM SEASON CONFIGURATION:											
0.75		0.69	0.67	0.64	0.60	0.56	0.50	0.43	0.33	0.20	0.00
0.80		0.75	0.73	0.70	0.67	0.63	0.57	0.50	0.40	0.25	0.00
0.85		0.81	0.79	0.77	0.74	0.70	0.65	0.59	0.49	0.32	0.00
0.90		0.87	0.86	0.84	0.82	0.79	0.75	0.69	0.60	0.43	0.00
0.95		0.93	0.93	0.92	0.90	0.89	0.86	0.83	0.76	0.61	0.00
1.00		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	----
TRAPEZOID PATTERN/3-STRATUM CONFIGURATION ASSUMING A PEAK OF 2 MONTHS IN ALL STATES: (I.E., ALPHA x BETA = 1/6 IN ALL CAMPS)											
----		0.95	0.91	0.87	0.83	0.79	0.74	0.68	0.60	0.49	0.33
	BETA=	0.67	0.50	0.40	0.33	0.29	0.25	0.22	0.20	0.18	0.17

NOTE: DELTA-SUB-iH, RANGING BETWEEN 0 AND 1, MEASURES THE EFFECTIVENESS OF SUBSTRATIFICATION in the i-TH CAMP.

TABLE B: OPTIMUM NUMBER OF DAYS (m) TO SELECT PER CAMP IN SAMPLING MIGRANT FARM WORKERS BY A 2-STAGE DESIGN WITH AND WITHOUT SUBSTRATIFICATION

RATIO: CAMP UNIT COST DIVIDED BY TIME UNIT COST (R-SUB-LM)		UNSTRATIFIED TIME SAMPLING			STRATIFIED TIME SAMPLING								
	DELTA-BAR-SUB-H=	0.0			0.5			0.7			0.9		
	DELTA-SUB-L=	0.4	0.5	0.6	0.4	0.5	0.6	0.4	0.5	0.6	0.4	0.5	0.6
5		3	2	2	2	2	1	2	1	1	1	1	1
10		4	3	3	3	2	2	2	2	1	1	1	1
15		5	4	3	3	3	2	3	2	2	1	1	1
20		5	4	4	4	3	3	3	2	2	2	1	1
25		6	5	4	4	4	3	3	3	2	2	2	1
30		7	5	4	5	4	3	4	3	2	2	2	1
35		7	6	5	5	4	3	4	3	3	2	2	2
40		8	6	5	5	4	4	4	3	3	2	2	2
45		8	7	5	6	5	4	5	4	3	3	2	2
50		9	7	6	6	5	4	5	4	3	3	2	2

NOTE: DELTA-BAR-SUB-H MEASURES THE EFFECTIVENESS OF SUBSTRATIFICATION IN TIME; DELTA-SUB-L MEASURES THE WITHIN-CAMP HOMOGENEITY AMONG DAYS.

TABLE C: OPTIMUM NUMBER OF DAYS (m) TO SELECT PER CAMP IN SAMPLING MIGRANT FARM WORKERS BY AN UNSTRATIFIED 2-WAY DESIGN IN A SURVEY WITH 5,000 PERSON-DAYS AVAILABLE FOR NON-FIXED COST ACTIVITY

RATIO: CAMP UNIT COST DIVIDED BY TIME UNIT COST (R_{LM})										
	δ_M =	0.01			0.03			0.05		
	δ_L =	0.4	0.5	0.6	0.4	0.5	0.6	0.4	0.5	0.6
5		8	7	6	11	10	9	13	12	11
10		9	8	7	13	12	11	16	15	14
15		10	9	8	15	13	12	18	16	15
20		11	9	8	15	14	13	19	17	16
25		11	10	9	16	15	13	20	18	16
30		12	10	9	17	15	14	20	18	17
35		12	11	9	17	16	14	21	19	17
40		12	11	10	18	16	14	21	19	18
45		13	11	10	18	16	15	22	20	18
50		13	11	10	19	17	15	22	20	18

NOTE: (1) δ_M MEASURES THE WITHIN-DAY HOMOGENEITY AMONG CAMPS; δ_L MEASURES THE WITHIN-CAMP HOMOGENEITY AMONG DAYS;

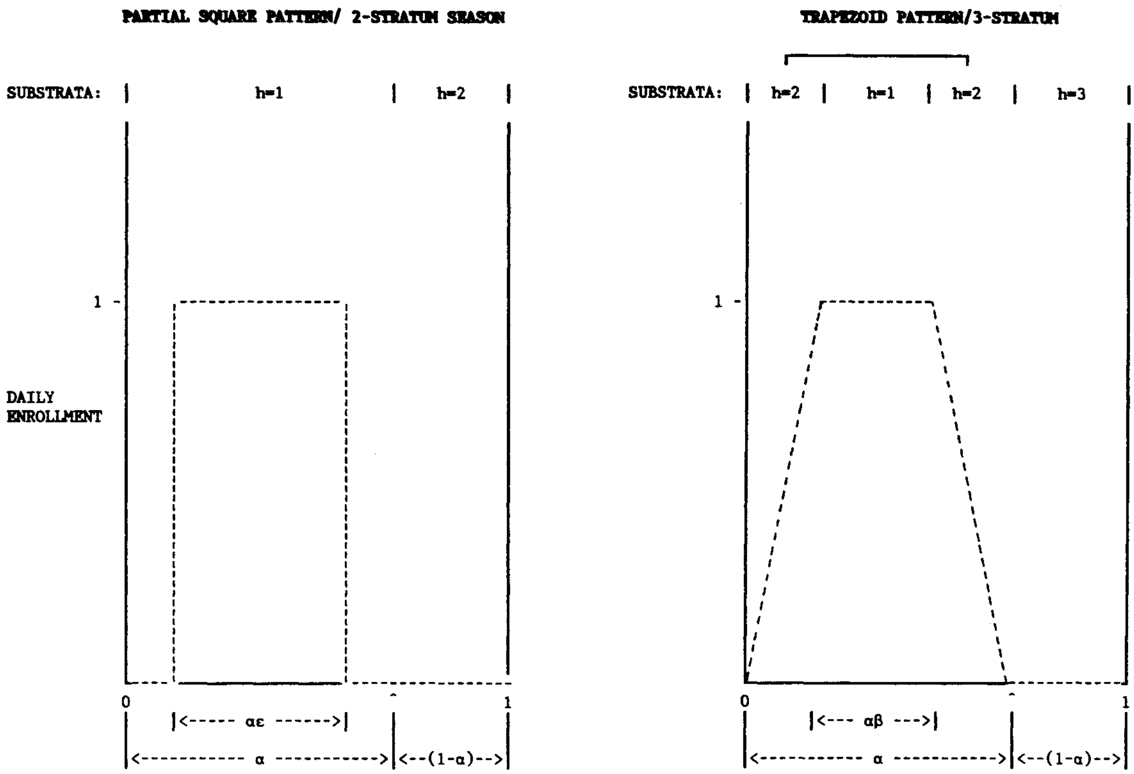
(2) THE AVERAGE UNIT COST AMONG ALL SELECTED DAYS IS ONE PERSON-DAY OF SALARY, I.E., $\bar{C}_M$=1.

TABLE D: RELATIVE COST-EFFICIENCY (RCEFF) FOR 2-STAGE AND 2-WAY DESIGNS COMPARED TO A URS(1*m of LM) DESIGN

SAMPLE SIZES		RATIO OF UNIT COST	WITHIN-CAMP HOMOGENEITY	SUBSTRATIFIED 2-STAGE			UNSTRATIFIED 2-STAGE	UNSTRATIFIED 2-WAY		
				DELTA-BAR-SUB-H				DELTA-SUB-M		
DAYS (m)	CAMPS (1)	(R-SUB-LM)	(DELTA-SUB-L)	0.5	0.7	0.9	---	0.01	0.03	0.05
2	714	5	0.4	1.56	1.75	1.99	1.22	0.20	0.08	0.05
			0.5	1.37	1.49	1.63	1.14	0.20	0.07	0.05
			0.6	1.22	1.30	1.38	1.07	0.20	0.07	0.05
2	185	25	0.4	1.75	1.97	2.24	1.38	0.59	0.28	0.18
			0.5	1.54	1.67	1.83	1.28	0.58	0.27	0.18
			0.6	1.38	1.46	1.55	1.20	0.56	0.27	0.18
2	96	50	0.4	1.78	2.00	2.28	1.40	0.83	0.46	0.32
			0.5	1.57	1.71	1.87	1.31	0.80	0.45	0.31
			0.6	1.40	1.49	1.58	1.23	0.77	0.44	0.31
6	455	5	0.4	1.21	1.27	1.33	1.09	0.43	0.20	0.13
			0.5	1.01	1.04	1.07	0.94	0.41	0.19	0.13
			0.6	0.86	0.88	0.90	0.82	0.38	0.19	0.12
6	161	25	0.4	1.86	1.95	2.05	1.68	1.09	0.64	0.46
			0.5	1.55	1.60	1.65	1.44	0.99	0.61	0.44
			0.6	1.32	1.35	1.38	1.26	0.90	0.57	0.42
6	89	50	0.4	2.02	2.12	2.22	1.82	1.41	0.97	0.74
			0.5	1.68	1.73	1.79	1.56	1.25	0.89	0.69
			0.6	1.44	1.47	1.50	1.37	1.12	0.82	0.65
10	333	5	0.4	0.93	0.96	0.99	0.87	0.50	0.27	0.19
			0.5	0.76	0.78	0.79	0.73	0.45	0.26	0.18
			0.6	0.65	0.65	0.66	0.63	0.41	0.24	0.17
10	143	25	0.4	1.73	1.78	1.83	1.61	1.23	0.84	0.64
			0.5	1.41	1.44	1.47	1.35	1.07	0.76	0.59
			0.6	1.20	1.21	1.23	1.16	0.95	0.70	0.55
10	83	50	0.4	1.98	2.03	2.09	1.85	1.57	1.20	0.98
			0.5	1.62	1.65	1.68	1.55	1.34	1.07	0.88
			0.6	1.37	1.39	1.41	1.33	1.18	0.96	0.81

NOTE: COMPUTATIONS FOR 1* ASSUME A TOTAL OF 5,000 PERSON-DAYS IN NON-FIXED COSTS FOR THE SURVEY.

FIGURE A: STANDARDIZED MEANS AND VARIANCES OF DAILY ENROLLMENT PATTERNS FOR MIGRANT CAMPS



MEAN = $\alpha\varepsilon$

VARIANCE = $\alpha\varepsilon(1-\alpha\varepsilon)$

MEAN = $\alpha(1+\beta)/2$

VARIANCE = $\alpha[4(2\beta+1)-3\alpha(1+\beta)^2]/12$

NOTES:

(1) STANDARDIZATION IS TO UNIT LENGTH AND UNIT MAXIMUM ENROLLMENT. MEANS FOR A CAMP WITH A MAXIMUM ENROLLMENT OF X^* CAN BE OBTAINED BY MULTIPLYING THE STANDARDIZED MEAN BY X^*. VARIANCES CAN BE OBTAINED BY MULTIPLYING BY X^{*2}.

(2) MEANS AND VARIANCES ARE UNAFFECTED BY DEPARTURES FROM SYMMETRY OF THE INTERVAL OF PEAK ENROLLMENT WITHIN THE GROWING SEASON.