COST MODELS FOR OPTIMUM ALLOCATION IN MULTI-STAGE SAMPLING

William D. Kalsbeek, Ophelia M. Mendoza
and
David V. Budescu[1]

Cost models to determine an optimum allocation of the sample among stages in cluster samples are considered. Results from a proposed cost model, which directly considers the implications of follow-up visits to sample clusters as well as other travel to and from the field by data collectors, are compared with results from existing cost models. The proposed model generally calls for fewer clusters with more elements selected per cluster than the existing models.

1. INTRODUCTION

One of the first issues in designing a multi-stage cluster sample is how to best allocate the sample among stages. In a two-stage design this amounts to deciding on the number of clusters to be selected in the first stage of sampling and the average sample size among selected clusters in the second stage. One normally wishes to choose that allocation of the sample among individual stages which will yield the best possible precision of estimates for the amount of funds available to conduct the survey. In the sequel, we will refer to this issue as the problem of determining an "optimum stage allocation".

The theory of optimum stage allocation requires both a variance and a cost model. The variance model is a mathematical formula for the precision of a survey estimator, written as a function of the sample sizes in each stage and certain measures of the components of unit variance attributable to each stage. Similarly, the cost model is a mathematical formula for the total cost

[1] William D. Kalsbeek and Ophelia M. Mendoza, Department of Biostatistics, University of North Carolina at Chapel Hill, and David V. Budescu, Department of Psychology, University of Haifa.

of conducting the survey, expressed as a function of the same stage-specific sample sizes but also various per-unit costs for each stage of the sampling design.

Variance models for many common multi-stage sampling designs have been produced, when the objective of the survey is to estimate the population means per element (see, for example, Hansen, Hurwitz, and Madow, 1953, and Cochran, 1977). Furthermore, important parameters of these variance models are readily estimable and can often be obtained from published reports. For example, Kish, Groves, and Krotki (1976) present estimates of one such parameter, the intraclass correlation coefficient for several national fertility surveys.

The variance model used in this paper is a simple but common one. Suppose that the sample is selected in two stages from a population consisting of equal-sized clusters. If simple random sampling (with replacement) is used to first select a sample of n clusters and next a sample of m elementary units within selected clusters, then the variance of the estimated population mean per element $\bar{y}$ is simply

$$\text{Var}(\bar{y}) = \sigma^2[1 + \rho(m - 1)]/nm, \tag{1.1}$$

where ρ is the intraclass correlation and σ^2 is the variance among all elementary units in the population. The result of (1.1) may also serve as a reasonable approximation even when clusters are of unequal size and selection procedures other than simple random sampling are used (see Kish 1965, Section 5.4). In this case we may view m as the average within-cluster sample size.

The development of reasonable cost models has received relatively little attention in the survey literature despite the fact that existing models contain parameters of survey cost which, though clearly defined, are difficult to compute. One such parameter is the cost of adding a cluster to the sample. Computing a reasonable measure of this per-unit cost is complicated by the difficulty in determining the impact of data collector travel which depends on such things as the size of the area being covered, the number of

clusters assigned to each data collector, and the pattern of travel followed by the data collector in completing the survey. Some consolation can be derived from the known robustness of optimum stage allocation when imperfect cost measures are used (see Kish, 1976), although nontrivial departures from the best attainable precision may result when severely misinterpreted cost measures are used.

Two well known cost models have been applied to the survey setting in which data collection required a visit to each cluster by a data collector (or in some surveys a team of data collectors). We call the first of these models the simple model in which total non-overhead costs can be expressed as

$$C_0^{(S)} = nC_1^{(S)} + nmC_2^{(S)}, \tag{1.2}$$

where $C_0^{(S)}$ is the total nonoverhead cost, $C_1^{(S)}$ is the average cost of adding a cluster to the sample, and $C_2^{(S)}$ is the average cost of adding an elementary unit to the sample. The simple model, combined with the variance model of (1.1), yields (see Cochran 1977, Section 10.6)

$$m_{opt}^{(S)} = \left\{ \left(\frac{1-\rho}{\rho}\right) \frac{C_1^{(S)}}{C_2^{(S)}} \right\}^{\frac{1}{2}} \tag{1.3}$$

as the optimum value of m.

The costs of travel during data collection often contribute significantly to total survey costs. Data collector travel and accompanying costs may be considered to be of two types. The first is between-cluster travel which refers to movement among clusters during a data collection trip. The second is positioning travel which refers to travel to the first cluster visited from the data collector's home base and then back to the home base from the last cluster visited during the data collection trip. The importance of the second cost model, suggested by Hansen, Hurwitz, and Madow (1953) and called the HHM Model here, is that it isolates between-cluster cost from the rest of the survey's total nonoverhead costs. This is done by assuming that the n clusters

are uniformly arranged in a rectangular geographic area of size A and that associated with each unit of distance travelled is a unit cost (U) consisting of two components: the mileage allowed for travel (e.g., in dollars per mile) and the ratio of hourly wages to the average rate of travel (e.g., in miles per hour).

In many surveys, data collection may require multiple visits to sample clusters. We incorporate the concept of follow-up visits into the HHM model by assuming the data collection is completed in H phases with np^{h-1} clusters being visited in the h-th phase; $0 < p < 1$. The cost of cluster follow-up is determined for the HHM model by summing the between-cluster travel cost over all phases. The HHM model as adapted here thereby takes the form,

$$\underset{\sim}{C_0^{(H)}} = n\underset{\sim}{C_1^{(H)}} + nm\underset{\sim}{C_2^{(H)}} + n^{\frac{1}{2}}\underset{\sim}{C_3^{(H)}} \tag{1.4}$$

where $\underset{\sim}{C_3^{(H)}} = UA^{\frac{1}{2}}(1 - p^{H/2})/(1 - p^{\frac{1}{2}})$ is the cost parameter of the term isolating the cost of between-cluster travel with follow-up visits considered. The cost of adding a cluster ($\underset{\sim}{C_1^{(H)}}$) and the cost of adding an element ($\underset{\sim}{C_2^{(H)}}$) in the HHM model include positioning travel cost but exclude all remaining between-cluster travel costs which are covered by the term, $n^{\frac{1}{2}}\underset{\sim}{C_3^{(H)}}$. The new HHM model, combined with the variance model once again, yields (see Hansen, et al., 1953, Vol. II, Section 6.11)

$$\underset{\sim}{m_{opt}^{(H)}} = \left\{ \left(\frac{1-\rho}{\rho}\right) \frac{\underset{\sim}{C_1^{(H)}} + \underset{\sim}{C_3^{(H)}}/(2n^{\frac{1}{2}})}{\underset{\sim}{C_2^{(H)}}} \right\}^{\frac{1}{2}} \tag{1.5}$$

which must be solved iteratively to determine the optimum value of m.

The intent of this paper is to extend the thinking about cost models used for optimum stage allocation and to produce a new model which more explicitly reflects actual survey costs. In so doing, we develop a cost model which: (1) isolates the increasingly important component of total survey costs due to data collector travel, (2) can easily accommodate follow-up visits to clus-

ters, and (3) can be expressed as a relatively simple function of a number of readily interpretable measures.

2. PROPOSED MODEL

The cost model discussed in this section isolates from other survey costs the cost of both between-cluster and positioning travel for data collectors. This is contrasted by the HHM model where only between-cluster travel costs are isolated and by the simple model where isolation of travel costs does not occur at all. The proposed model can therefore be viewed as an attempt to avoid the difficulty in existing models of having to allocate unisolated travel costs among other per-unit costs, e.g., in the simple model data collector travel costs must be appropriated to $C_1^{(S)}$ and $C_2^{(S)}$. As with the HHM model, assumptions made for the proposed model regarding the location of clusters and the route of between-cluster travel are needed to express the survey's total travel cost as a function of n.

We shall see that assumptions concerning the spatial arrangement of clusters and travel by the data collectors are kept simple and admittedly somewhat naive. Less restrictive and presumably more realistic assumptions could be made, but the effect would be to add prohibitive complexity to the problem. We shall also see that the assumptions made in developing the proposed model allow one to express survey costs in terms of simple, well-known parameters of a survey operation. Thus, optimum stage allocation using the proposed model can be determined by specifying several easily understood measures characterizing a survey protocol.

2.1 Spatial Configuration of Sample Clusters

We now describe the spatial configuration of sample clusters as assumed for the proposed cost model and illustrated in Figure A. The object of the assumed configuration is for the uniformly scattered clusters to be arranged so that distances for reasonable travel routes can be expressed simply as a

function of several readily obtained parameters. One assumes that the expressions will hold true for all possible parameter values.

Suppose that we have a survey population with land area of geographical size A and that the population is divided into t equal and nonoverlapping subareas, each of size A/t and containing v = n/t sample clusters. One data collector is assigned to do the survey work in each subarea, which is shaped as a square with a number of evenly spaced concentric circles contained therein. The data collector's home base, assumed to be one of the clusters in the sample, lies in the center of the subarea in order to assure adequate accessibility to clusters during data collection. The distance from the home base to the outermost circle in each subarea is r. Thus, since the size of each subarea is $4r^2$, we have $r = (A/t)^{\frac{1}{2}}/2$. Moving from the home base in a subarea, the k-th circle (k = 1,...,K) contains 6k clusters. Assuming a multiple of six clusters on each concentric circle allows clusters to be almost uniformly spaced in the subarea, except for the square corners.

2.2 Data Collection Protocol

Using the spatial configuration of clusters just described, we now discuss a protocol for data collection which one might expect to observe in certain kinds of surveys with two or more stages of sampling. Comparison of results from existing cost models is later made within the context of this protocol.

Data collection in a subarea is assumed to require multiple phases of activity since work in most clusters usually involves several visits, some to make arrangements for data collection in the cluster and others to actually collect the data. As mentioned earlier, we let H denote the number of phases required to complete data collection in a subarea. This parameter can also be interpreted as the maximum number of required visits to individual clusters. In the h-th phase of data collection (h = 1,2,...,H), we assume that vp^{h-1} clusters (where $0 \leqslant p \leqslant 1$) are visited in a series of trips before proceeding with the next phase. Each trip involves a visit to ℓ neighboring clusters not previously visited during that phase of data collection. The cluster located in the home base is included in all phases of data collection.

Several assumptions are now made regarding movement of the data collectors among clusters. First the travel route followed in each trip proceeds from that data collector's home base, to each of the ℓ clusters (without backtracking), and then back again to the home base. Second, data collector travel is assumed to proceed in a straight line except between neighboring clusters on a circle where travel follows the arc of the circle. The choice of the arc distance over the straight-line is thought to be feasible since the formula for the former is simpler and since travel in surveys seldom follows a straight line.

Third, movement between two neighboring circles follows the shortest possible straight-line distance. This means that the cluster of departure from one circle and the cluster of destination on a neighboring circle are in line with the home base. The alignment of clusters 7 and 8 in Figure A illustrates this assumption. Fourth, travel within clusters and between data collector subareas is assumed to be negligible and is therefore not specifically isolated in the proposed model.

One final important assumption in the proposed model concerns the problem of the spatial configuration of clusters when $h \geq 1$; i.e., when the number of clusters visited during a phase of data collection is a subset of the v clusters originally selected in the subarea. To retain the simplicity of the concentric circle arrangement through all phases of data collection, we allow the number of concentric circles (K_h) at the h-th phase to vary according to the size of vp^{h-1} while fixing the size of the interviewer subarea at A/t. Thus, we have $K_h = (\alpha_h - 1)/2$, where $\alpha_h = \{1 + \frac{4}{3}(vp^{h-1} - 1)\}^{\frac{1}{2}}$.

2.3 Cost Formulation

Total travel cost in the proposed model is calculated as the product of U and the total distance travelled (D). Formulations for D, expressed alternatively as a function of the cluster workload per data collector (v) and the number of data collector subareas (t), are given below. Although the two formulations

are functionally similar (since $v = n/t$), developing both solutions is thought to be important because either v or t may be specified in designing a survey. Details of the derivations for (2.1) - (2.5) are appended.

Assuming the above data collection protocol, the total distance travelled over all phases, expressed as a function of v, will be

$$D^{(P)} = \delta_3^{(P)} n^{\frac{1}{2}}, \tag{2.1}$$

where

$$\delta_3^{(P)} = (A/v)^{\frac{1}{2}}[\tfrac{4}{3}\{v(1 - p^H)/(1 - p) - H\} + \{1 + (\ell - 1)\pi/2\}\{\sum_{h=1}^{H} \alpha_h + H\}]/2\ell.$$

This leads to a cost model which has the same general form as the HHM model of (1.4) but where the coefficient of the $n^{\frac{1}{2}}$ term is $U\delta_3^{(P)}$ and the optimum value can be obtained from (1.5).

The total distance travelled, obtained as a function of t, can be written as

$$D^{(P)} = \delta_0^{(P)} + n\delta_1^{(P)} + \sum_{h=1}^{H} \alpha_h \delta_4^{(P)}, \tag{2.2}$$

where

$$\delta_0^{(P)} = H(At)^{\frac{1}{2}}\{3(\ell - 1)\pi - 2\}/12\ell,$$

$$\delta_1^{(P)} = 2\{(1 - p^H)/(1 - p)\}(A/t)^{\frac{1}{2}}/3\ell,$$

$$\delta_4^{(P)} = (At)^{\frac{1}{2}}\{(\ell - 1)\pi + 2\}/4\ell.$$

The distance model of (2.2) leads to a cost model of the general form

$$C_0 = nC_1 + nmC_2 + \sum_{h=1}^{H} \alpha_h C_4. \tag{2.3}$$

Obtaining optimum values for n and m from (2.3) is an excessively cumbersome process which can be simplified by substituting a first-order Taylor series

approximation (in n) for α_h, evaluated at t/p^{h-1} for simplicity. By so doing we have

$$\alpha_h \doteq (2p^{h-1}/3t)n + \frac{1}{3}, \tag{2.4}$$

which, when applied to (2.3), reduces the proposed cost model to

$$\underset{\sim}{C}_0^{(P)} = nC_1^{(P)} + nm\underset{\sim}{C}_2^{(P)}, \tag{2.5}$$

where

$$\underset{\sim}{C}_0^{(P)} = \underset{\sim}{C}_0 - U\{\underset{\sim}{\delta}_0^{(P)} + H\underset{\sim}{\delta}_4^{(P)}/3\},$$

$$C_1^{(P)} = \underset{\sim}{C}_1 + U\{\underset{\sim}{\delta}_1^{(P)} + 2\underset{\sim}{\delta}_4^{(P)}(1 - p^H)/3t(1 - p)\},$$

$$\underset{\sim}{C}_2^{(P)} = \underset{\sim}{C}_2.$$

$\underset{\sim}{C}_0$ is the total prespecified nonoverhead cost of the survey, $\underset{\sim}{C}_1$ is the prespecified average cost of adding a cluster to the sample (excluding all costs of data collector travel), and $\underset{\sim}{C}_2$ is the prespecified average cost of adding an element to the sample (excluding, once again, all data collector travel costs). We note from (2.5) that using the approximation for α_h has reduced the proposed model to the form which, except for the three cost parameters, resembles the simple cost model of (1.2). Optimum values of m and n are obtained from (1.3) and by solving for n in (2.5).

3. COMPARISON OF PROPOSED MODEL WITH EXISTING MODELS

In this section we compare results obtained from the proposed cost model (expressed as a function of v) with results from the simple and HHM cost models. We consider the situation where a two-stage survey of the United States is being planned, and the variance model of (1.1) is assumed in all comparisons. Measures used as the basis for comparisons among models are as follows: (1) optimum value of n, (2) optimum value of m, and (3) the variance of the survey estimate given the optimum allocation.

Optimum values of n and m for the simple HHM models are obtained from (1.3)

and (1.5), respectively. To make comparisons with these models more realistic, adjustment factors are calculated to account for those travel costs not specifically isolated by the models. The adjustment procedure is similar to the approach mentioned earlier and suggested by Hansen, et al. (1953, Vol. 1, Section 6.13). To account for positioning travel costs in the HHM model we specify that

$$\underset{\sim}{C}_3^{(H)} = \underset{\sim}{\lambda}^{(H)}\underset{\sim}{C}_1,$$

$$\underset{\sim}{C}_2^{(H)} = \underset{\sim}{\lambda}^{(H)}\underset{\sim}{C}_2,$$

$$\underset{\sim}{C}_3^{(H)} = \underset{\sim}{\lambda}^{(H)}(A)^{\frac{1}{2}}U(1 - p^{H/2})/(1 - p^{\frac{1}{2}}),$$

where

$$\underset{\sim}{\lambda}^{(H)} = \underset{\sim}{C}_0/\{n_{opt}^{(P)}\underset{\sim}{C}_1 + n_{opt}^{(P)}m_{opt}^{(P)}\underset{\sim}{C}_2 + (n_{opt}^{(P)}A)^{\frac{1}{2}}U\} \tag{3.1}$$

is the adjusting factor, $n_{opt}^{(P)}$ is the corresponding optimum value for n under the proposed model, and $m_{opt}^{(P)}$ is the corresponding optimum value for m under the proposed model. Using $\underset{\sim}{\lambda}^{(H)}$ in this way has the effect of assuming that positioning travel costs contribute to each cost parameter of the HHM model by the same relative amount. In similar fashion, we account for all costs of data collector travel in the simple model by setting $\underset{\sim}{C}_1^{(S)} = \underset{\sim}{\lambda}^{(S)}\underset{\sim}{C}_1$ and $\underset{\sim}{C}_2^{(S)} = \underset{\sim}{\lambda}^{(S)}\underset{\sim}{C}_2$, where the adjustment factor is

$$\underset{\sim}{\lambda}^{(S)} = \underset{\sim}{C}_0/(n_{opt}^{(P)}\underset{\sim}{C}_1 + n_{opt}^{(P)}m_{opt}^{(P)}\underset{\sim}{C}_2). \tag{3.2}$$

We must acknowledge the synthetic nature of the adjustment factors, $\underset{\sim}{\lambda}^{(H)}$ and $\underset{\sim}{\lambda}^{(S)}$, used for our comparisons. In each case the adjustment factor is a function of the optimum values of n and m obtained from the corresponding proposed model. In reality, these factors would be calculated for the HHM and simple models by estimating the proportion of the survey's budget not spent on those travel costs left unaccounted for by the model. One might suspect that this estimated proportion would, at best, amount to a rough approximation which would probably differ from the adjustments produced from (3.1) and

(3.2). Thus, we suspect that using these factors may contribute to making the simple and HHM models seem more comparable to the proposed model than they in fact are.

3.1 Assumed Parameter Values

Producing the findings of the comparison study required several numerical values for the various statistical and cost parameters of the models. First, we consider national surveys in the United States, A=3,042,265 square miles, the land area of the United States, excluding Alaska and Hawaii. We also arbitrarily set $\underset{\sim}{C}_0$ = \$500,000, the total nonoverhead cost of the survey, and U = \$0.45, the unit cost per mile travelled. The latter figure is obtained by assuming a mileage allowance of \$0.25 per mile, an interviewer salary of \$6.00 per hour, and an average travel rate of 30 miles per hour. All combinations of the following groups of parameters are considered in our comparisons:

$(\underset{\sim}{C}_1, v)$: (\$50, 20); (\$250, 5)

$(\underset{\sim}{C}_2, \rho, H)$: (\$10, 0.3, 5); (\$25, 0.8, 20)

ℓ: 1; 2

ρ: 0.05; 0.15

Parameters were grouped in this manner since many of the combinations resulting from individual parameters were thought to be unrealistic.

The parameters $\underset{\sim}{C}_1$ and v are grouped together to indicate the degree of difficulty that data collectors would have in setting up and maintaining participation among clusters in the survey. For example, in a one-time survey or the first installment of an ongoing survey, one might expect to find cluster set-up costs to be high and the set-up activities to be sufficiently burdensome so that the average number of clusters assigned per data collector would of necessity be low. Thus, for present purposes we designate $\underset{\sim}{C}_1$ = \$250 and v = 5 to indicate cluster set-up and maintenance which is "difficult". Activities such as obtaining endorsements, making initial visits to solicit cooperation, and constructing the frame for selecting the second stage would all

contribute toward the determination of these values. We designate $\underset{\sim}{C}_1$ = \$50 and v = 20 to indicate cluster set-up and maintenance activities which are "easy". This situation might be observed in surveys in which set-up activities are relatively simple. One example would be a subsequent installment of the ongoing survey while another would be a survey in which arrangements can be made by mail or telephone. The parameters $\underset{\sim}{C}_2$, p, and H are used to jointly indicate the level of difficulty in the data collection protocol. When $\underset{\sim}{C}_2$ = \$10, p = 0.3, and H = 5, the average number of times a cluster will be visited is 1.4 and data collection is assumed to be "easy". This may occur, for example, in a survey where the protocol requires only that a small amount of readily accessible data be extracted for each element in a cluster. When less accessible data are extracted or when follow-up of selected elements is required, data collection might be called "difficult" in which case we assume that $\underset{\sim}{C}_2$ = \$25, p = 0.8, and H = 20, thus implying that the average number of times a cluster will be visited is 5.6.

The parameter indicating the number of clusters visited per trip (ℓ) assumes the values 1 or 2 in these comparisons. Allowing $\ell \geq 2$ is thought to be unrealistic in national surveys since distances would preclude visiting a large number of clusters on a single trip. Two moderate values of intraclass correlation (ρ) are assumed.

3.2 Findings

Tables 1-3 contain the results of the comparison study involving the proposed model and the versions of the simple model and of the HHM model where $\underset{\sim}{\lambda}^{(S)}$ and $\underset{\sim}{\lambda}^{(H)}$ are applied, respectively. Optimum values of n and m, as determined under the proposed model, are presented in Table 1. As expected, optimum values of n tend to be lower when cluster set-up and maintenance is difficult, and optimum values of m tend to be lower when data collection is difficult.

The major focus of the comparison study is the difference between optimum results under the proposed model and comparable results under the simple and HHM models. Optimum results for the proposed and simple models are compared

in Table 2 in which one notes that differences are generally substantial. Optimum values for n under the proposed model are found to be between 2.4 and 60.0 percent lower than under the simple model, while optimum values for m are between 7.6 and 198.2 percent higher under the proposed model. These large differences are thought to be attributable to the ability of the proposed model to isolate between-cluster and positioning travel costs. This results in greater per-cluster costs and a smaller optimum number of sample clusters. The greatest differences in optimum variances, computed by applying the optimum values of n and m to (1.1), occur in surveys with easy cluster set-up and maintenance and difficult data collection. One might speculate that the magnitude of these variance differences is largely due to the relatively heavy cluster workload (i.e., $v = 20$) assumable when cluster set-up and maintenance is deemed easy. However, when this workload is lightened (i.e., $v = 5$) and considered with the same combination of parameters, the relative difference among optimum variances is reduced but remains substantial at 11-16 percent, as opposed to the 18-27 percent figures presented in Table 2.

The effects of the number of clusters visited per trip (ℓ) and the intraclass correlation (ρ) are also readily apparent in Table 2. Larger differences appear when $\ell = 1$ than when $\ell = 2$. This effect can be attributable to the greater importance that travel costs would play when only a single cluster can be visited per trip to the field. Furthermore, when $\rho = 0.05$, relative differences for n and m are somewhat greater than when $\rho = 0.15$; however these differences are an artifact due in part to the iterative approach which is used to obtain $m_{opt}^{(P)}$. From (1.3) and (1.5) we would expect relative differences on optimum values of m to be identical.

The relative differences between the proposed and HHM models presented in Table 3 remain notable but are generally smaller than the differences reported in Table 2. We suspect that the greater similarity between results under the proposed and HHM models can be attributable to the fact that the HHM model represents a more realistic reflection of survey costs than does the simple model. However, as with comparisons involving the simple model, optimum values of n are smaller and optimum values of m are higher under the proposed model in Table 3. These comparisons also reveal once again that the largest

differences in optimum variance occur in surveys with easy cluster set-up and maintenance and difficult data collection. Variance differences in other instances are negligible.

3.3 Discussion

We have proposed a cost model where the important component of travel during data collection can be completely set apart to improve one's ability to accurately reflect survey costs in determining an optimum stage allocation. In addition, a study designed to compare optimum results of this proposed model with two existing cost models has indicated substantial differences. However, aside from these differences, perhaps the most important practical implication of the proposed cost model is that the optimum stage allocation can be produced by specifying measures which are intuitively simple. These measures are of two types: fiscal and nonfiscal characteristics of the survey design. The required fiscal characteristics (i.e., $\underset{\sim}{C}_0$, $\underset{\sim}{C}_1$, and $\underset{\sim}{C}_2$) can be determined by estimating the costs of certain components of the survey. For example, we might determine $\underset{\sim}{C}_1$ from a recent similar survey as the average per-cluster cost of choosing the sample of clusters, soliciting among clusters for participation in the survey (excluding travel costs), and constructing the sampling frame for sampling units within selected clusters. The required nonfiscal characteristics of the survey (i.e., A, v or t, p, H, ℓ, and ρ) can be obtained as factual information from prior surveys. For example, knowledge of the maximum and average number of visits required per cluster in a recent similar survey would determine p and H.

We conclude by briefly examining the robustness and artificiality of the proposed cost model. Robustness is considered on the one hand by determining (from stated assumptions) the types of surveys for which the model is likely to be useful. Assumptions of the model imply that the sample points are clustered rather than randomly scattered in the population and that during data collection a group of these clusters is assigned to each data collector. This arrangement of sample points and data collection assignments will occur in certain types of household and institutional samples. An example of one such

arrangement is the National Survey of Nursing Homes (see National Center for Health Statistics, 1968) which is selected in two stages with nursing homes designated as clusters.

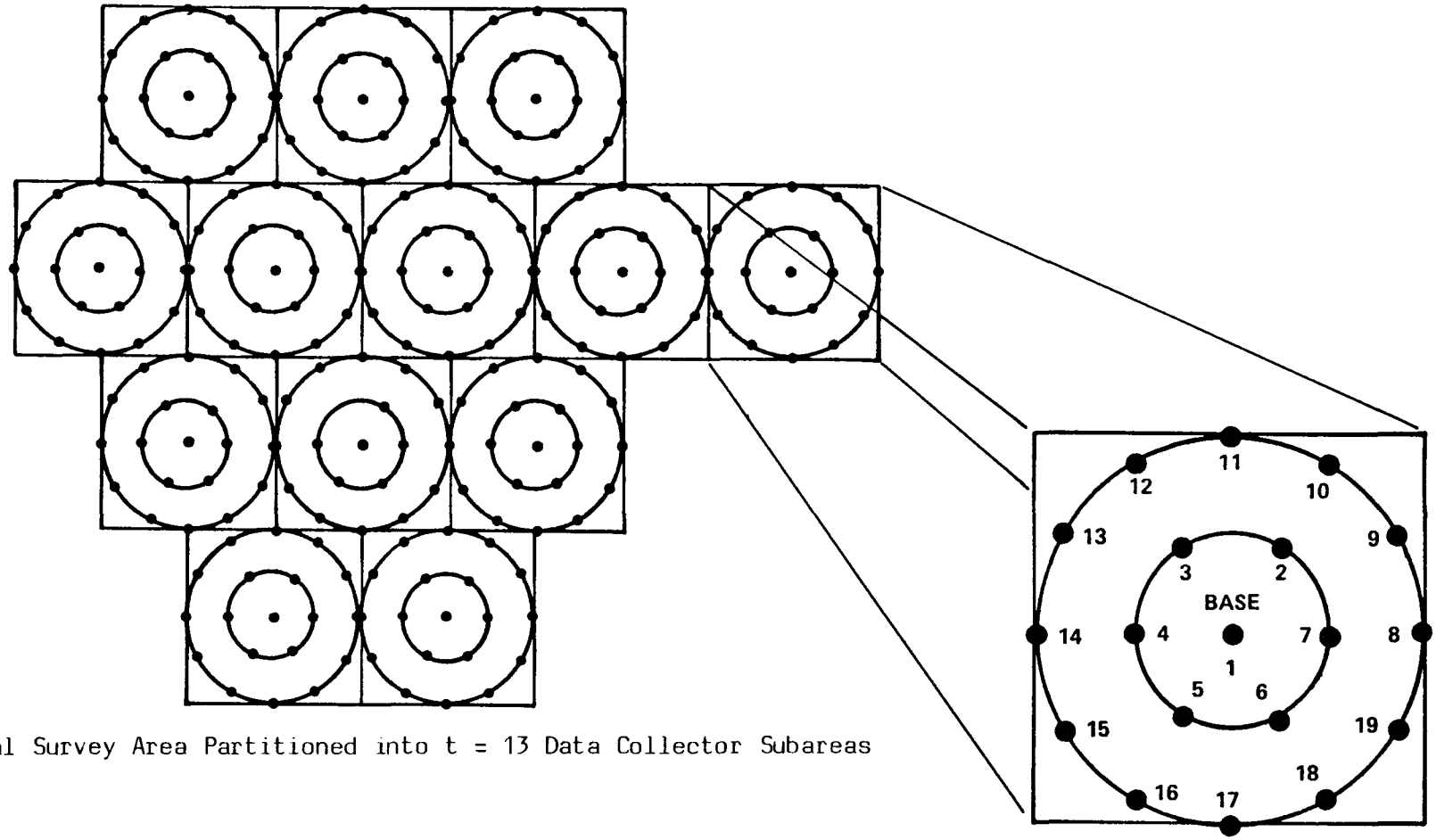
The arrangement might also appear in household surveys where the sample is chosen in two or more stages and where data collectors complete interviews within several small area segments (see, for example, the Virginia Health Survey conducted by the Statistical Sciences Group, Research Triangle Institute, 1978). A household sample chosen in three or more stages can be accommodated by treating A as the size of the land area occupied just by selected primary sampling units (PSU's) and then considering sampling units from the second or subsequent stages to be the clusters that follow a concentric configuration within each data collector subarea (i.e., consider Figure A with t scattered rather than contiguous subareas). Procedurally, one would substitute $t\bar{A}$ for A in (2.2), where $\bar{A}$ is the average land area to be covered by each data collector in the planned survey. Given this adaptation, it is important to note that the number of sample PSU's would be prespecified and thereby not optimized, that n in the cost and variance models would be the number of sample clusters (i.e., not PSU's), and that m would be the average number of elementary units per cluster. Treating the number of sample PSU's to be fixed and then determining the optimum allocation for subsequent stages would be reasonable for certain surveys where the ultimate sample is chosen from a sample of PSU's which is used repeatedly for different surveys. The variance model of (1.1) may have to be modified to reflect the additional sampling stages (see Hansen, et al., 1953, Vol. II, Section 6.9). Some institutional samples selected in three or more stages (e.g., the Hospital Discharge Survey of the National Center for Health Statistics, 1970) could be considered for the multi-stage adaptation as well. However, the proposed model would be less practical for those surveys where cluster sizes are so large that each data collector is assigned only one or two clusters or where selected clusters are not likely to be uniformly scattered about within subareas.

Another facet of the robustness issue is the generalizability of the findings. Clearly, any conclusions drawn from our findings must be limited by the parameter values we have assumed. Rather than using values from existing sur-

veys in which case inferences would be limited to those surveys, our strategy was to create several prototype surveys based upon combinations of unit costs and other parameters thought to reflect current survey practice. Values used to create the prototype were often taken directly or inferred from recent surveys known to the authors.

Finally, a certain degree of impracticality is the price one pays to keep things simple since realism and simplicity seem to be indirectly related in building survey cost models. Thus, while the intent of our research has been to find a more realistic yet simple model, we must acknowledge a substantial amount of remaining artificiality in our assumptions. For example, clusters are more likely to be randomly scattered than to exist as multiples of six lying on concentric circles. Moreover, travel between neighboring clusters would follow winding, circuitous routes rather than arcs or straight lines, and return visits to clusters would have more haphazard schedules than well-established phases of follow-up with the number of clusters per phase decreasing each time by a factor of p. While the proposed model reflects the orderliness which one hopes for in most survey field operations, it, like other existing models, fails to capture the unpredictability of things which tends to blend into the orderliness. Stochastic events can be used to create unpredictability but adding them tends to complicate the model to the point of being less useful mathematically. Until more realistic assumptions can be tied to simplicity, we are faced with the need to settle for cost models which fall short of the realism we seek.

Figure A. Proposed Spatial Configuration of Clusters in a Survey Population



Total Survey Area Partitioned into t = 13 Data Collector Subareas

Spatial Arrangement of Clusters
in Each Data Collector Subarea

TABLE 1. Optimum Values for n and m Under the Proposed Cost Model
(A = 3,042,265 square miles; $\underset{\sim}{C}_0$ = \$500,000)

Prototype Survey	Parameters: Cluster set-up and maintenance	Parameters: Data collection	Parameters: ℓ	Parameters: ρ	Optimum Values: $n^{(P)}_{\underset{\sim}{opt}}$	Optimum Values: $m^{(\rho)}_{\underset{\sim}{opt}}$
[1]	Easy	Easy	1	0.05	1673	14.1
[2]				0.15	2319	7.4
[3]			2	0.05	1910	13.1
[4]				0.15	2669	6.9
[5]		Difficult	1	0.05	385	18.4
[6]				0.15	518	9.4
[7]			2	0.05	489	16.0
[8]				0.15	675	8.2
[9]	Difficult	Easy	1	0.05	871	23.6
[10]				0.15	1095	12.8
[11]			2	0.05	847	23.9
[12]				0.15	1065	12.9
[13]		Difficult	1	0.05	426	19.0
[14]				0.15	560	10.1
[15]			2	0.05	378	20.2
[16]				0.15	493	10.6

Cluster set-up and maintenance $(\underset{\sim}{C}_1, v)$ { Easy (\$50, 20); Difficult (\$250, 5) }

Data collection $(\underset{\sim}{C}_2, p, H)$ { Easy (\$10, 0.3, 5); Difficult (\$25, 0.8, 20) }

TABLE 2. Relative Differences Between the Proposed Model and the Simple Model

($\underset{\sim}{C}_0$ = \$500,000)

Prototype Survey	Parameters: Cluster set-up and maintenance	Parameters: Data collection	ℓ	ρ	Relative difference: proposed vs simple model (in percent): n_{opt}	m_{opt}	Optimum Variance
[1]	Easy	Easy	1	0.05	-22.7	44.3	3.1
[2]				0.15	-16.6	38.5	2.7
[3]			2	0.05	-18.5	34.4	2.0
[4]				0.15	-13.3	29.7	1.7
[5]		Difficult	1	0.05	-60.0	198.2	24.4
[6]				0.15	-52.9	179.3	26.9
[7]			2	0.05	-54.6	159.2	18.1
[8]				0.15	-47.2	142.3	19.6
[9]	Difficult	Easy	1	0.05	-3.8	8.4	0.2
[10]				0.15	-2.4	7.6	0.1
[11]			2	0.05	-4.3	9.7	0.2
[12]				0.15	-2.7	8.7	0.2
[13]		Difficult	1	0.05	-18.0	37.8	2.6
[14]				0.15	-12.4	33.6	2.1
[15]			2	0.05	-21.2	46.2	3.6
[16]				0.15	-15.1	41.3	3.0

Relative difference is computed as the measure under the proposed model minus the measure under the simple model divided by the measure under the simple model, and multiplied by 100.

Cluster set-up and maintenance ($\underset{\sim}{C}_1$, v)

- Easy (\$50, 20)
- Difficult (\$250, 5)

Data collection ($\underset{\sim}{C}_2$, p, h)

- Easy (\$10, 0.3, 5)
- Difficult (\$25. 0.8, 20)

TABLE 3. Relative Differences Between the Proposed Model and the HHM Model

($\underset{\sim}{C}_0$ = \$500,000)

Prototype Survey	Parameters: Cluster set-up and maintenance	Parameters: Data collection	ℓ	ρ	Relative difference: proposed vs HHM model (in percent): n_{opt}	m_{opt}	Optimum Variance
[1]	Easy	Easy	1	0.05	-12.5	22.9	0.9
[2]				0.15	-8.9	20.2	0.8
[3]			2	0.05	-8.7	15.2	0.4
[4]				0.15	-6.0	13.4	0.4
[5]		Difficult	1	0.05	-22.8	49.6	3.5
[6]				0.15	-17.5	46.2	3.3
[7]			2	0.05	-17.2	34.2	1.8
[8]				0.15	-13.0	31.7	1.7
[9]	Difficult	Easy	1	0.05	-1.3	2.9	0.0+
[10]				0.15	-0.8	2.6	-0.0
[11]			2	0.05	-1.8	4.0	0.0+
[12]				0.15	-1.1	3.6	0.0+
[13]		Difficult	1	0.05	-4.0	7.9	0.2
[14]				0.15	-2.6	7.2	0.1
[15]			2	0.05	-6.5	13.5	0.4
[16]				0.15	-4.4	12.2	0.3

Relative difference is computed as the measure under the proposed model minus the measure under the HHM model, divided by the measure under the HHM model, and multiplied by 100.

Cluster set-up and maintenance ($\underset{\sim}{C}_1$, v)

- Easy (\$50, 20)
- Difficult (\$250, 5)

Data collection ($\underset{\sim}{C}_2$, p, h)

- Easy (\$10, 0.3, 5)
- Difficult (\$25. 0.8, 20)

APPENDIX

Details of the derivations for (2.1) - (2.5) in the text are presented here. Using the assumed spatial configuration of clusters and data collection protocol for the proposed model as discussed in Sections 2.1 and 2.2, respectively, the total distance travelled ($D^{(P)}_{\sim}$) is first expressed as a function of the number of sample clusters assigned to each data collector (v). Given the configuration of clusters as illustrated in Figure A, note that the positioning and between-cluster travel distances for each data collector during the h-th phase of data collection are $\{12r/K_h\ell\} \sum_{k=1}^{K_h} k^2$ and $\{2\pi r(\ell - 1)/K_h\ell\} \sum_{k=1}^{K_h} k$, respectively. Summing these two distances, recalling that $K_h = (\alpha_h - 1)/2$, where $\alpha_h = \{1 + \frac{4}{3}(vp^{h-1} - 1)\}^{\frac{1}{2}}$, and multiplying times the number of data collectors (t), we have the total positioning and between-cluster travel distance for the h-th phase expressed as:

$$D_h = rt[2K_h(K_h + 1)(2K_h + 1) + (\ell - 1)\pi K_h(K_h + 1)]/K_h\ell$$

$$= rt[(\alpha_h - 1)^2 + \{6 + (\ell - 1)\pi\}(\alpha_h - 1)/2 + 2 + (\ell - 1)\pi]/\ell. \qquad (A.1)$$

Noting that $\sum_{h=1}^{H} p^{h-1} = (1 - p^H)/(1 - p)$, we sum D_h over all phases to obtain:

$$D^{(P)}_{\sim} = \sum_{h=1}^{H} D_h = (rt/\ell) \sum_{h=1}^{H} [\{1 + \tfrac{4}{3}(vp^{h-1} - 1)\} + \alpha_h + \{\alpha_h + 1\}\{(\ell - 1)\pi/2\}]$$

$$= rt[\tfrac{4}{3}\{v(1 - p^H)/(1 - p) - H\} + \{1 + (\ell - 1)\pi/2\}\{\sum_{h=1}^{H} \alpha_h + H\}]/\ell. \qquad (A.2)$$

Recalling that $r = (A/t)^{\frac{1}{2}}/2$ and $t = n/v$ and substituting these identities into (A.2) leads to (2.1).

To express $D^{(P)}_{\sim}$ as a function of the number of data collectors (t), first note that we must use $\alpha_h = \{1 + \frac{4}{3}(np^{h-1}/t - 1)\}^{\frac{1}{2}}$ as opposed to the earlier expression for α_h. Using the new expression complicates things a bit since α_h is

now a function of both t and the number of sample clusters (n), which is one of the parameters to be optimized. Using the new expression for α_h and recalling once again that $r = (A/t)^{\frac{1}{2}}/2$, a bit of algebra allows us to recast (A.1) as:

$$\begin{aligned} D_h &= rt[\alpha_h^2 + \{1 + (\ell - 1)\pi/2\}\alpha_h + (\ell - 1)\pi/2]/\ell \\ &= rt[\{3(\ell - 1)\pi - 2\}/6 + (\tfrac{4}{3} p^{h-1}/t)n + \{1 + (\ell - 1)\pi/2\}\alpha_h]/\ell \\ &= (At)^{\frac{1}{2}}\{3(\ell - 1)\pi - 2\}/12\ell + n\{2p^{h-1}(A/t)^{\frac{1}{2}}/3\ell\} \\ &\quad + \alpha_h(At)^{\frac{1}{2}}\{(\ell - 1)\pi + 2\}/4\ell. \end{aligned} \tag{A.3}$$

Summing D_h from (A.3) over all phases leads us to the total distance given in (2.2),

$$D^{(P)} = \delta_0^{(P)} + n\delta_1^{(P)} + \sum_{h=1}^{H} \alpha_h \delta_4^{(P)}, \tag{A.4}$$

where

$$\delta_0^{(P)} = H(At)^{\frac{1}{2}}\{3(\ell - 1)\pi - 2\}/12\ell,$$

$$\delta_1^{(P)} = 2\{(1 - p^H)/(1 - p)\}(A/t)^{\frac{1}{2}}/3\ell,$$

$$\delta_4^{(P)} = (At)^{\frac{1}{2}}\{(\ell - 1)\pi + 2\}/4\ell.$$

The total travel distance given by (A.4) leads to an overall survey cost model given by:

$$C_0 = nC_1 + nmC_2 + U\delta_0^{(P)} + Un\delta_1^{(P)} + U\sum_{h=1}^{H}\{1 + \frac{4}{3}(np^{h-1}/t - 1)\}^{\frac{1}{2}}\delta_4^{(P)}. \tag{A.5}$$

Where C_0 is the total prespecified nonoverhead cost of the survey, C_1 is the prespecified average cost of adding a cluster to the sample (excluding all costs of data collector travel), and C_2 is the prespecified average cost of

adding an element to the sample (excluding, once again, all data collector travel costs).

Using the cost model given by (A.5) to obtain optimum values for n and m is disadvantageous because the final righthand term of (A.5) is a complex function of n. To circumvent this difficulty we suggest substituting a first-order Taylor series approximation in n for $\alpha_h = \{1 + \frac{4}{3}(np^{h-1}/t - 1)\}^{\frac{1}{2}}$, which is arbitrarily evaluated at t/p^{h-1} to simplify the approximation. By so doing we have

$$\alpha_h = f(n) = \{1 + \frac{4}{3}(np^{h-1}/t - 1)\}^{\frac{1}{2}} \doteq f(t/p^{h-1}) + f'(t/p^{h-1})(n - t/p^{h-1})$$

where f'(.) is the first partial derivative of f(.) with respect to n. Since $f(t/p^{h-1}) = 1$ and $f'(t/p^{h-1}) = 2p^{h-1}/3t$, we have

$$\alpha_h \doteq (2p^{h-1}/3t)n + \frac{1}{3} \qquad (A.6)$$

which is a linear function of n. Applying the approximation of (A.6) to (A.5) reduces the proposed model to the form,

$$\underset{\sim}{C}_0^{(P)} = n\underset{\sim}{C}_1^{(P)} + nm\underset{\sim}{C}_2^{(P)}, \qquad (A.7)$$

where

$$\underset{\sim}{C}_0^{(P)} = \underset{\sim}{C}_0 - U\{\underset{\sim}{\delta}_0^{(P)} + H\underset{\sim}{\delta}_4^{(P)}/3\},$$

$$\underset{\sim}{C}_1^{(P)} = \underset{\sim}{C}_1 + U\{\underset{\sim}{\delta}_1^{(P)} + 2\underset{\sim}{\delta}_4^{(P)}(1 - p^H)/3t(1 - p)\},$$

and

$$\underset{\sim}{C}_2^{(P)} = \underset{\sim}{C}_2.$$

The result of (A.7) corresponds to (2.5) in the main text.

REFERENCES

[1] Cochran, W.G. (1977). Sampling Techniques. Third Edition. New York: John Wiley and Sons.

[2] Hansen, M.H., Hurwitz, W.H., and Madow, W.G. (1953). Sample Survey Methods and Theory. Vols. I and II. New York: John Wiley and Sons.

[3] Kish, L., Groves, R.M., and Krotki, K.P. (1976). Sampling Errors in Fertility Surveys. World Fertility Survey Occasional Paper No. 17. London. World Fertility Survey.

[4] Kish, L. (1976). "Optima and Proxima in Linear Sample Designs". Journal of Royal Statistical Society, Series A. 139: 80-95.

[5] National Center for Health Statistics (1968). "Design and Methodology for a National Survey of Nursing Homes". Vital and Health Statistics. PHS Pub. No. 1000, Series 1, No. 7. Washington: U.S. Government Printing Office.

[6] National Center for Health Statistics (1970). "Development of the Design of the NCHS Hospital Discharge Survey". Vital and Health Statistics. PHS Pub. No. 1000, Series 2, No. 39. Washington: U.S. Government Printing Office.

[7] Statistical Sciences Group (1978). Virginia Health Survey: Volume I --- Methodological Report, Report No. RTI/1546/00-00F, Research Triangle Institute, Research Triangle Park, North Carolina.

[8] U.S. Bureau of the Census (1978). The Current Population Survey: Design and Methodology. Technical Paper No. 40. Washington: U.S. Government Printing Office.