A NEW COST MODEL FOR OPTIMUM ALLOCATION IN TWO-STAGE SAMPLING

William D. Kalsbeek and Ophelia M. Mendoza, University of North Carolina at Chapel Hill
David V. Budescu, Research Triangle Institute

1. INTRODUCTION

Determining an optimum stage allocation requires assumptions about the variance of survey estimators and about thė nature of survey costs. These assumptions manifest themselves in the form of two models, each expressed as a mathematical function of certain parameters for which estimates are required to calculate the optimum results. The variance model can be derived explicitly, depending on the type of multi-stage design and the population value being estimated from the sample. Virtually all standard sampling texts present the most commonly assumed variance models for the estimator of the population mean per element (see especially Hansen, Hurwitz, and Madow, 1953). Furthermore, estimates of the important parameters of the variance model are easily estimable and can be obtained from published reports. For example, Kish, Groves, and Krótki (1976) present such estimates for several national fertility surveys.

The model reflecting survey costs, on the other hand, is largely dependent on how one views the survey protocol and the amount of complexity one allows in its formulation. We wish for the cost model to reflect important components of the budget, allocated to such things as sampling, travel, interviewing, processing, and analysis. Like the variance model, the cost model is expressed as a function of sample sizes for each stage of selection in the sample. However, unlike the variance model which can be mathematically derived given the statistical implications of the sampling design, identification of the functional form of the cost model is a less rigorous process.

In addition to the matter of choosing an appropriate functional form for a cost model, one is faced with the problem of obtaining good estimates of unit costs, the parameters of the model that is chosen. Even with a clear understanding of what is sought, it may be difficult to calculate, for example, a reasonable measure of the average cost per PSU when it must include that portion of travel costs which depends on the number of PSU's which will be in the sample. The problem here is that the number of sample PSU's is unknown at the time when this unit cost is needed.

We believe that the ideal cost model has the following three characteristics. First, it must realistically represent the way in which costs are incurred in an actual survey operation. Second, the formulation should be simple enough so that the optimum solution is tractable. Third, unit costs which constitute the parameters of the cost model should be sufficiently straightforward in interpretation so that they can be easily understood by operations staff to develop useful estimates for calculating optimum allocations.

Our primary intent here is to introduce a new cost model which we believe is closer to the ideal model than those existing in the literature. We focus our attention on the problem of estimating a population mean per element from an important class of national health surveys using two-stage sampling designs. We demonstrate that substantially different optimum allocations and minimum variances may result when choosing the proposed model over existing models.

There exists several health surveys to which the findings of this paper apply. One example is the Study of Efficacy of Nonsocomial Infectious Control (SENIC) which used hospitals as PSUs, and medical records as secondary sampling units or elements. The National Survey of Nursing Homes (NSNH) used nursing homes as PSUs and patients and employees as elements. In the National Longitudinal Study (NLS), a sample of schools were selected in the first stage, and a sample of students from each school in the sample were selected in the second stage.

2. VARIANCE MODEL

A two-stage variance model, which will be used in conjunction with the cost models compared in this paper, is now presented. We assume that the survey population consists of N PSUs each of size M. A simple random sample of n PSUs, selected without replacement, is chosen in the first stage. Within each selected PSU a simple random sample of m elements, selected without replacement, is chosen in the second stage. Following the notation of Cochran (1977), the variance of the estimated population mean per element $(\bar{y})$ is simply

$$\mathrm{Var}(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_1^2 + \left(\frac{1}{m} - \frac{1}{M}\right) S_2^2/n$$

$$\doteq \left(\frac{1}{nm} - \frac{1}{MN}\right) S^2\{1 + \rho(m-1)\} \quad , \qquad (2.1)$$

where ρ is the intraclass correlation coefficient, S^2 is the overall element variance in the population, and S^2_1 and S^2_2 are between-PSU and within-PSU variance components, respectively.

3. EXISTING COST MODELS

3.1 Simple Model

Assuming that the survey setting is one in which two-stage sampling is used and data collection requires a visit to each PSU by an interviewer or interviewing team, the simplest model for survey costs, excluding overhead expenses, takes the form

$$C_0^{(S)} = nC_1^{(S)} + nmC_2^{(S)} \qquad (3.1)$$

where $C_0^{(S)}$ is the total nonoverhead cost, $C_1^{(S)}$ is the average cost of adding a PSU to the sample, and $C_2^{(S)}$ is the average cost of adding an element to the sample. The latter two parameters are assumed to cover all survey costs, including the costs of interviewer travel.

With the model of (3.1) and the variance model of (2.1), the optimum value of m will be (see Cochran 1977, Section 10.6)

$$m_{opt}^{(S)} = \left\{ \frac{S_2^2}{S_1^2 - S_2^2/M} \; \frac{C_1^{(S)}}{C_2^{(S)}} \right\}^{1/2}$$

$$\doteq \left\{ \left(\frac{1-\rho}{\rho}\right) \frac{C_1^{(S)}}{C_2^{(S)}} \right\}^{1/2} \qquad (3.2)$$

where $n_{opt}^{(S)}$ is obtained by solving for n in (3.1). The result of (3.2) has been widely used, yet the simple model from which it is derived fails to isolate any components of cost due to interviewer travel occurring during data collection. Instead, a decision is needed as to how one might reasonably appropriate these travel costs to $C_1^{(S)}$ and $C_2^{(S)}$ in (3.1).

3.2 HHM MODEL

In the ensuing discussion we consider interviewer travel and accompanying costs to be of two types. We refer to interviewer movement among PSUs during data collection as between-PSU travel. Since most interviewers or interviewing teams operate from a home base, some amount of travel for each data collection trip is required to the first PSU from the home base and then back to the home base from the last PSU covered in the trip. This second type of travel is called positioning travel.

The importance of the cost model suggested by Hansen, et al., (1953) is that it isolates between -PSU travel costs from the rest of the survey's total nonoverhead costs. This is done by assuming a particular configuration of PSUs in the population. Suppose that n PSUs are uniformly arranged in a rectangular survey population whose geographic area is of size A (see Figure A). Then the vertical or horizontal distance (d) between neighboring PSUs is exactly $(A/n)^{\frac{1}{2}}$. Furthermore, if travel between PSUs is in a straight line, and if the sequence of travel corresponds to the numbering in Figure A, then subject to $n \doteq n-1$, the total distance travelled is approximately $nd=(An)^{\frac{1}{2}}$. A spatial arrangement of PSUs, like this rectangular configuration, serves to enable one to express the number of PSUs as a function of the area of the survey population.

Associated with each unit of distance travelled is a unit cost (U) which is the sum of two components: the mileage allowance for travel (e.g., dollars per mile) and the ratio of the hourly wages to the average rate of travel (e.g., miles per hour). This leads then to a cost model, we call the HHM model, which takes the form

$$C_0^{(H)} = nC_1^{(H)} + nmC_2^{(H)} + (n)^{\frac{1}{2}}C_3^{(H)} \quad , \qquad (3.3)$$

where $C_3^{(H)} = U\sqrt{A}$ is the cost parameter of the term isolating between-PSU travel costs. The cost of adding a PSU $(C_1^{(H)})$ and the cost of adding an element $(C_2^{(H)}$ in the HHM model include positioning travel costs but exclude all remaining between-PSU travel costs which are covered by the term, $(n)^{\frac{1}{2}} C_3^{(H)}$. Applying (3.3) to the simple variance model, the optimum value of m is (see Hansen, et al. 1953, Vol. 11, Section 6.11)

$$m_{opt}^{(H)} \doteq \left\{ \left(\frac{1-\rho}{\rho}\right) \frac{C_1^{(H)} + C_3^{(H)}/2(n)^{\frac{1}{2}}}{C_2^{(H)}} \right\}^{\frac{1}{2}} \quad , \qquad (3.4)$$

which has an iterative solution involving the dummy variable $q = 2(n)^{\frac{1}{2}}$ so that $n_{opt}^{(H)} = q_{opt}^2/4$, where q_{opt} is the value of q on the final iteration before pre-established convergence criteria are met.

Two comments regarding the HHM model are needed here. First, the costs of positioning travel (i.e., travel costs to and from the PSUs from the interviewer's home base) are not directly accommodated by the HHM model since these travel distances cannot be easily expressed in a simple mathematical form. The most reasonable alternative is to incorporate positioning travel costs into $C_1^{(H)}$, $C_2^{(H)}$, and $C_3^{(H)}$. Hansen, et al. (1953) have suggested a procedure whereby an adjustment factor is computed for the cost parameter of each of the three components in the model. The adjustment factor for each component requires an estimate of the portion of total costs due to positioning travel. A form of this adjustment procedure is used in the comparison study of cost models discussed later. Second, data collection requiring one or more follow-up visits to PSUs can be incorporated into the HHM model by assuming that work is completed in several phases. All n PSUs are visited in the first phase before beginning the second phase, wherein only a subset of the PSUs are visited. Continually smaller subsets are included in subsequent phases until all work is completed in the final phase. For present purposes, we assume that np^{h-1} (where $0<p<1$) PSUs will be visited in the h-th phase. Given these assumptions, the effect of PSU followup can be determined for the HHM model by summing between-PSU travel costs over all phases, whereupon we have $C_3^{(H)} = U(A)^{\frac{1}{2}}(1-p^{H/2})/(1-p^{\frac{1}{2}})$.

4. PROPOSED MODEL

4.1 SPATIAL CONFIGURATION OF PSUs

We now describe in greater detail the proposed spatial configuration of PSUs as illustrated in Figure B. Suppose that we have a survey population with land area of geographical size A and that the population is divided into t nonoverlapping subareas, each of size A/t and containing v = n/t PSUs. One interviewer is assigned to do the data collection work in each subarea, which is shaped as a square with a number of evenly spaced concentric circles contained therein. The interviewer's home base, assumed to be one of the PSUs in the sample, lies in the center of the subarea in order to assure adequate accessibility to PSUs 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 outward in a subarea, the k-th circle contains 6k PSUs. Assuming a multiple of six PSUs on each concentric circle allows PSUs to be almost uniformly spaced in the subarea, except for the square corners.

4.2 DATA COLLECTION PROTOCOL

Using the spatial configuration of PSUs just described, we now discuss a protocol for data collection which one might expect to observe in the two-stage national surveys mentioned in Section 1. 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 PSUs usually involves several visits, some to make arrangements for data collection in the PSU and others to actually collect the data. 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 PSUs. In the h-th phase of data collection (h=1,2,...,H), we assume that vp^{h-1} PSUs (where $0<p<1$ are visited in a series of trips before proceeding with the next phase. Each trip involves a visit to ℓ neighboring PSUs not previously visited during that phase of data collection. The PSU located in the home base is included in all phases of data collection.

Several assumptions are now made regarding movement of the interviewers among PSUs. First, the travel route followed in each trip proceeds from the interview's home base, to each of the ℓ PSUs (without backtracking), and then back again to the home base. Second, interviewer travel is assumed to proceed in a straight line except between neighboring PSUs on a circle where travel follows the arc of the circle. We believe that the choice of the arc distance over the straight-line distance is 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 PSU of departure from one circle and the PSU of destination on a neighboring circle are in line with the home base. The alignment of PSUs 7 and 8 in Figure C illustrates this assumption. Fourth, travel within PSUs and between interviewer 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 PSUs when $h>1$; i.e., when the number of PSUs visited during a phase of data collection is a subset of the v PSUs 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 $\alpha_h = \{1+\frac{4}{3}(vp^{h-1}-1)\}^{\frac{1}{2}}$.

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

$$D^{(P)} = \delta_3^{(P)} (n)^{\frac{1}{2}} \quad . \qquad (4.1)$$

where

$$\delta_3^{(P)}=(A/v)^{\frac{1}{2}}\left[\frac{4}{3}\ \{v(1-p^H)/(1-p)-H\}+\{1+(\ell-1)\pi/2\}\ \{\sum_1^H \alpha_h + H\}\right]/2\ell \quad .$$

This leads to a cost model which has the same general form as the HHM model of (3.3) but where the coefficient of the $(n)^{\frac{1}{2}}$ term is $U\delta_3^{(P)}$ and the optimum value can be obtained from (3.4).

5. 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. Two general types of two-stage sample surveys are considered: "small" surveys of local areas like cities or counties and "large" surveys of states or small countries. The case where the land area (A) of the survey population is the size of the United States is also considered. In all comparisons, the variance model of (2.1) is assumed. Measures used as the basis for comparisons among models are as follows: (1) optimum value of n, (2) optimum value of m, and (3) expected variance of the survey estimate given the optimum allocation.

Optimum values of n and m for the simple and HHM models are obtained from (3.2) and (3.4), 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 discussed 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 $C_1^{(H)} = \lambda^{(H)} C_1$; $C_2^{(H)} = \lambda^{(H)} C_2$; and $C_3^{(H)} = \lambda^{(H)} (A)^{\frac{1}{2}} U(1-p^{H/2})/(1-p^{\frac{1}{2}})$, where $\lambda^{(H)} =$

$$C_0/\{n_{opt}^{(P)}\ C_1 + n_{opt}^{(P)}\ m_{opt}^{(P)}\ C_2 + n_{opt}^{(P)}\ A)^{\frac{1}{2}}\ U\} \qquad (5.1)$$

is the adjustment 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 $\lambda^{(H)}$ 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 interviewer travel costs in the simple model by setting $C_1^{(S)} = \lambda^{(S)}C_1$ and $C_2^{(S)} = \lambda^{(S)}C_2$, where the adjustment factor is

$$\lambda^{(S)} = C_0/(n_{opt}^{(P)}\ C_1 + n_{opt}^{(P)}\ m_{opt}^{(P)}\ C_2). \qquad (5.2)$$

We must acknowledge a certain degree of artificiality in the adjustment factors, $\lambda^{(H)}$ and $\lambda^{(S)}$, 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 (5.1) and (5.2).

Calculating the results of the comparison study requires several numerical values for the various statistical and cost parameters of the models. The assumed parameter values are shown in Table 1. Values of A and the total nonoverhead survey cost (C_0) are considered together since it seems reasonable to assume that the total nonoverhead cost and the geographic area will be directly related. The assumed values of the different parameters are based on previous experience with surveys and published survey reports.

The results of the comparisons are presented in Tables 2 and 3. In relatively low-budget surveys conducted in areas of relatively small geographic size, the results from the proposed cost model are quite similar to results obtained from existing models. In particular, the proposed model yields variances which are within 1% of comparable variances from the HHM and simple models. On the other hand, substantial relative differences, approaching 20 to 30 percent, in the optimum allocations of n and m may occur in large surveys. Substantially greater differences (i.e., approaching 10 percent for variances, 50 percent for n, and 115 percent for m) may occur in large-budget surveys of the entire United States.

In conclusion, while the proposed cost model is not a complete remedy for the problem of optimum stage allocation, we do suggest that it possesses several advantages over existing models, as follows:

1. Results can be obtained with relatively simple input specifications. The model directly accounts for all interviewer travel costs, thus simplifying unit cost computations for survey operations staff.

2. In relatively small scale surveys, results from the proposed model are comparable to results from existing models in which cost parameters have been adjusted to accommodate interviewer travel costs. Variances of estimates from optimum allocations obtained from the proposed model differ only slightly from variances obtained from the HHM and simple models.

3. The concentric home-base orientation of the spatial configuration in the proposed model enhances the realism of results. For example, in many present-day surveys data collection is decentralized by hiring local interviewers to complete data collection in an assigned area surrounding the location of their individual residences. This data collection protocol is realistically represented by any of the proposed travel models.

4. Other parameters of different realistic data collection protocols can be directly accommodated by the proposed model. For example, the number of PSU's visited per day can be specified and, as with the HHM model, follow-up can be considered. The simple and HHM models, on the other hand, have limited flexibility and require often complex and abstractly defined cost parameters.

REFERENCES

Cochran, William G. (1977), Sampling Techniques, 3rd Edition, New York: John Wiley and Sons.

Hansen, Morris H.; Hurwitz, William N.; and Madow, William G. (1953), Sample Survey Methods and Theory, vols. I and II, New York: John Wiley and Sons.

Kish, Leslie; Groves, Robert M.; and Królki, Karol P. (1976), "Sampling Errors in Fertility Surveys," World Fertility Survey, Occasional Paper no. 17, London.

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. Public Health Service. Washington. U.S. Government Printing Office.

Quade, D. et al. The SENIC Sampling Process. American Journal of Epidemiology, 1980, III, 486-502.

Westat, Inc. (1972), "Sample Design for the Selection of the Sample of Schools with Twelfth-Graders for a Longitudinal Study," Westat Inc: Rockville, Maryland, June 1972.

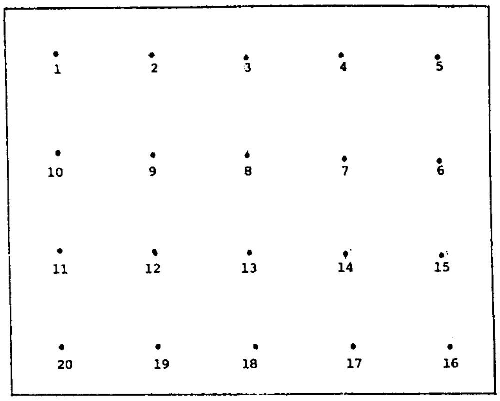

Figure A. Spatial Arrangement of PSUs in HHM Model

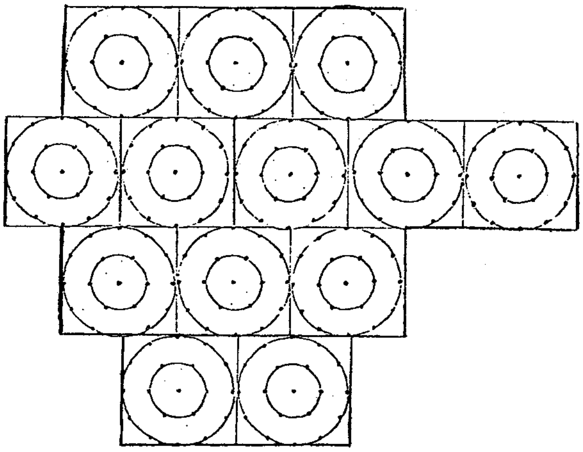

Figure B. Illustration of Survey Population With t = 13 Interviewer Areas

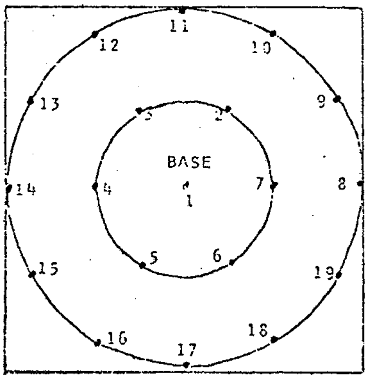

Figure C. Spatial Arrangement of PSUs in Each Interviewer Area

Table 1. Assumed Parameter Values

	Small Survey		Large Survey	
Parameter	Low	High	Low	High
(A^a, C_0)	(40, $5,000)	(400, $50,000)	(20,000, $50,000)	(120,000[b], $500,000)
C_1	$25	$75	$50	$250
C_2	$5	$15	$10	$25
U	$0.42		$0.35	
ρ	0.02	0.10		0.40
v	5	20	5	25
t	2	10	5	50
(p,H)	(.5, 4)	(.7, 8)	(.5, 4)	(.7, 8)

[a] in square miles

[b] This figure is increased to 3,042,265 sq. mi., the land area of continental United States in a special comparison.

Table 2. Comparison of Proposed Model With HHM and Simple Models

	Cost in Dollars						Follow-up		Percent Relative Difference[d]		
Area[b] (A)	Total[c] (C_0)	PSU (C_1)	Element (C_2)	PSU Workload (V)	Intraclass Correlation (ρ)	PSU's Per Trip (ℓ)	Proportion (p)	Phases (H)	n_{opt}	m_{opt}	Variance
					Small Two-Stage Sample Survey:						
Versus HHM Model:											
0.40	50	25	15	20	.02	1	.70	8	-3.0	3.6	0.07
0.40	50	25	5	20	.02	1	.70	8	-2.2	2.9	0.07
0.40	50	25	15	20	.10	1	.70	8	-1.8	2.6	0.06
0.04	5	25	5	20	.10	1	.50	4	-0.8	1.4	0.04
0.04	5	25	5	20	.10	2	.50	4	-0.5	0.8	0.04
Versus Simple Model:											
0.40	50	25	15	20	.02	1	.70	8	-4.9	6.0	0.11
0.40	50	25	5	20	.02	1	.70	8	-3.6	4.8	0.10
0.40	50	25	15	20	.10	1	.70	8	-3.0	4.3	0.09
0.40	50	25	15	20	.02	2	.70	8	-3.7	4.5	0.07
0.40	50	25	5	20	.10	1	.70	8	-2.1	3.6	0.06
					Large Two-Stage Sample Survey:						
Versus HHM Model:											
20	50	50	25	25	0.02	1	0.7	8	-11.7	16.2	0.30
20	50	50	10	25	0.02	1	0.7	8	-9.3	13.9	0.28
20	50	50	25	25	0.10	1	0.7	8	-7.5	12.3	0.26
120	500	50	10	25	0.02	1	0.7	8	-7.5	10.8	0.26
20	50	50	10	25	0.10	1	0.7	8	-5.7	10.9	0.23
Versus Simple Model:											
20	50	50	25	25	0.02	1	0.7	8	-18.5	27.3	0.82
20	50	50	10	25	0.02	1	0.7	8	-14.9	21.1	0.81
20	50	50	25	25	0.10	1	0.7	8	-12.1	20.3	0.76
20	50	50	10	25	0.10	1	0.7	8	-9.3	17.8	0.70
20	50	50	25	25	0.40	1	0.7	8	-7.0	16.1	0.56

[a] The five largest absolute relative differences of variance are presented for each comparison and type of survey.

[b] In thousands of square miles.

[c] In thousands of dollars

[d] Percent Relative Difference = $\left[\frac{\text{Proposed - Comparison}}{\text{Comparison}}\right] \times 100.$

Table 3. Supplementary Comparison of Proposed Models With HHM and Simple Models for National Samples of the United States (A = 3,042,265)[a]

Cost in Dollars			PSU	Intraclass	PSU's	Follow-up		Percent Relative Difference[c]		
Total	PSU	Element	Workload	Correlation	Per Trip	Proportion	Phases			
(C_0)[b]	(C_1)	(C_2)	(V)	(ρ)	(ℓ)	(p)	(H)	n_{opt}	m_{opt}	Variance
Versus HHM Model:										
500	50	10	25	0.02	1	0.7	8	-23.8	45.3	2.8
500	50	25	25	0.10	1	0.7	8	-20.3	42.1	2.8
500	50	25	25	0.02	1	0.7	8	-27.7	49.7	2.7
500	50	10	25	0.10	1	0.7	8	-16.3	39.0	2.6
500	50	25	25	0.40	1	0.7	8	-12.7	36.7	2.2
500	50	10	25	0.40	1	0.7	8	-9.4	35.0	1.8
500	50	25	25	0.02	1	0.5	4	-22.3	35.8	1.5
500	50	10	25	0.02	1	0.5	4	-18.6	31.6	1.4
500	50	25	25	0.10	1	0.5	4	-15.6	28.8	1.4
500	50	10	25	0.10	1	0.5	4	-12.2	26.1	1.3
Versus Simple Model:										
500	50	25	25	0.10	1	0.7	8	-38.4	91.5	9.6
500	50	10	25	0.10	1	0.7	8	-32.4	83.6	9.3
500	50	10	25	0.02	1	0.7	8	-43.3	100.5	9.3
500	50	25	25	0.02	1	0.7	8	-48.7	114.2	8.6
500	50	25	25	0.40	1	0.7	8	-26.6	78.1	8.5
500	50	10	25	0.40	1	0.7	8	-20.8	74.0	7.2
500	50	25	25	0.10	2	0.7	8	-30.7	65.2	5.6
500	50	10	25	0.02	2	0.7	8	-35.4	72.3	5.6
500	50	25	25	0.02	2	0.7	8	-40.9	83.2	5.4
500	50	10	25	0.10	2	0.7	8	-25.2	58.8	5.3

[a] The ten largest absolute relative differences of variance are presented for each comparison.

[b] In thousands of dollars.

[c] Percent Relative Difference $= \left[\frac{\text{Proposed - Comparison}}{\text{Comparison}}\right] \times 100$