Redesign of the Survey of Construction[1]

Bonnie E. Kegan U.S. Census Bureau, Washington, D.C.
James D. Ashley[2] U.S. Postal Service, Office of the Inspector General, Arlington, Virginia

1. Introduction

The Census Bureau conducts the Survey of Construction (SOC) to provide estimates of the number of new privately owned housing units started, under construction and completed; the number of new single-family houses sold and for sale, and other characteristics of new residential housing. Each month, approximately 200 field representatives visit pre-selected permit offices to sample new building permits and locate new residential construction by canvassing all roads in pre-selected areas that do not require building permits. All sampled buildings are followed from start of construction to completion or sale through phone calls or site visits with the builders or owners.

SOC is typically redesigned every ten years after each Decennial Census to account for changes in the population distribution that happen over time. However, the last redesign of SOC was completed in 1984. The redesign following the 1990 Census was not completed due to the conversion from Paper and Pencil Interviewing (PAPI) to Computer Assisted Personal Interviewing (CAPI).

The SOC sample is a three-stage cluster design. In the first stage, a subsample of Primary Sampling Units (PSU) are sampled from the Current Population Survey's (CPS) PSUs. The CPS PSUs are first classified as either self-representing (SR) or nonself-representing (NSR) for SOC. The NSR PSUs are then stratified and selected using a maximum overlap procedure. The independent samples of building permit places (SUP) and non-permit (NP) areas are the second stages of the SOC sample, followed by the selection of building permits in the third stage for SUP. All new residential housing is sampled in selected NP areas.

The purpose of this paper is to describe our redesign of the first stage (SOC PSUs are selected as a subsample of the CPS PSUs). It will discuss the formulation of cost and variance models used to determine optimal allocation, as well as criteria used to identify CPS PSUs as self-representing (SR) or nonself-representing (NSR) for SOC. Finally it will detail the stratification of the NSR PSUs and their selection using overlap methods and will compare the redesigned sample of PSUs to the current sample of PSUs drawn in 1984.

2. Optimal Allocation: Cost and Variance Models

Prior to the stratification of the PSUs for selection into SOC, we had to determine the optimal number of NSR PSUs since this is also the number of strata used in the one PSU per stratum design. Additionally our optimal allocation also specifies the optimal number of permit-issuing places (NSR Places) and nonpermit segments which are needed in later stages of the redesign. The optimal allocation required the formulation of both cost and variance models. The optimal values were then determined by optimizing the proposed cost and variance models for minimum variance with a fixed cost.

The cost function includes the national average cost per month for SOC at three levels: PSU, C_1, place, C_2 and segment, C_3. The total cost C, minus the overhead cost C_0, is set equal to the sum of each marginal cost multiplied by the appropriate number of PSUs (n_1), places (n_2), or segments (n_3.) The cost function is $C - C_0 = C_1 n_1 + C_2 n_2 + C_3 n_3$. In this model the number of SR PSUs and SR places are assumed fixed and the costs associated with them are also fixed. The total variable cost C-C_0 is determined by substituting the current sample sizes into the cost equation. The calculation of the national average monthly cost per unit at each of the three levels is discussed below.

The majority of activities for SOC are reported as interview activity. These activities include "Time and expenses associated with completing survey interviewing assignments; personal visits and telephone interviewing, traveling to and from assignments, call backs, transcribing and editing, and preparing mailings; planning daily interview itinerary; setting appointments; setting up computer to transmit and receive cases; listing of permits." (SOC Regional Office Memorandum 02-03)

The total hours per month and total miles per month were obtained from the Cost and Response Management Network (CARMN) for each of the twelve regional offices for the fiscal years 1999-2002. This provided complete

[1] This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

[2] J. Ashley is currently a statistician at the US Postal Service, Office of the Inspector General. He was at the Census Bureau from 1998-2002. The views expressed in this paper are those of the authors and do not necessarily represent the opinion of the USPS, OIG.

monthly data on mileage and hours for 1999-2001, except total mileage for August 2000. The hours and miles reported are averaged over the three years within each of the twelve months (2000 is excluded from August's average) and then averaged over the twelve months to obtain the average number of hours or miles per month. Since this data includes both NSR and SR PSUs and places within each region, the average number of hours (or miles) per month per PSU (or place) is calculated by first finding the average per month per total number of PSUs (or places) and then by multiplying by the percentage of PSUs (or places) that are NSR in that region in the current sample. The average number of hours (or miles) per NSR-PSU (or NSR place) was then averaged over the twelve regions to find the national average number of hours (or miles) per month per unit. The national averages were then multiplied appropriately by either the average hourly field representative salary ($13.00/hr) or the current mileage stipend ($ 0.365/mi) and then added together to get the total national average per unit cost for PSUs and places as shown in the formula below:

$$C_i = \left[\begin{array}{l} \frac{1}{12} \sum_{\text{region}} \begin{array}{l} (\text{miles/mon /unit i}) \\ *(\$0.365/\text{mi}) \end{array} \\ + \frac{1}{12} \sum_{\text{region}} \begin{array}{l} (\text{hours/mon /unit i}) \\ *(\$13.00/\text{hr}) \end{array} \end{array} \right]$$

(1).

At the PSU level there is an additional monthly cost of recruiting, replacing, and training interviewers that must be added to the total per PSU cost. Estimates of training costs for hiring and training a new interviewer when there is a turnover (i), refresher training (ii), and hiring and training a new interviewer for a new PSU in the sample (iii), were requested from Field Division. The estimate for hiring and training a new interviewer is $7,850, and the national turnover rate is 10%. The monthly cost per PSU, assuming one interviewer per PSU for (i) is $7,850 x (0.10/12) = $65.42. There are an average of 0.20 refresher training sessions per year for about 345 participants. The monthly cost per PSU for (ii) is ($235,250 x 0.20)/(345*12)= $11.36. For introducing a new PSU into the sample the training cost is spread out over 100 months, yielding a $78.50 cost per month per PSU.

Hours and mileage associated with non-permit (NP) segments were obtained from each of the regional offices. Each office was asked to report an estimate of the average number of hours and miles per month spent on NP assignments. They were also asked to report the number of Field Representatives (FRs) with NP assignments and the number of NP assignments that were only canvassed quarterly. Only eight of the twelve regional offices have NP segments. Using the regional office data along with numbers of segments obtained from the current sample a national average monthly cost per segment was obtained similar to the national average cost for PSUs and places. The regional offices reported monthly averages so there was no need to average over a period of several years for each month. However for those offices which have quarterly canvassing, the hours and miles spent during the large months (with quarterly) and small months (without quarterly) canvassing were multiplied by 4 and 8 respectively, and the total divided by twelve to obtain the monthly average hours and miles. Only Dallas and Kansas City offices reported detailed enough information for this. Of the remaining six offices usually none or only one to two segments were canvassed quarterly.

The national average number of hours and miles determined per unit as described in previous sections can be seen in Table 1 below:

Table 1: Time and Mileage Estimates

	Average hours/month	Average miles/month
Per NSR-PSU	17.5342	126.6822
Per NSR-place	3.5349	25.0165
Per NP segment	4.1451	94.6025

After multiplying the averages in Table 1 by the appropriate costs as shown previously in (1) and adding in total training costs the total cost per NSR-PSU is C_1 = $429.46. The total cost per NSR-place is C_2 = $55.08 and the total cost per NP-Segment is C_3 = $88.42. The proposed monthly cost model then becomes:

$$C - C_0 = 429.46\ n_1 + 55.08\ n_2 + 88.42\ n_3 \qquad (2).$$

By substituting the current sample sizes of 119 NSR PSUs, 459 NSR places, and 71 NP segments into (2) for n_1, n_2, and n_3 respectively we found the estimated total monthly variable cost depending on the size and nature of the sample design (C-C_0) to be $82,665.28.

After determining the cost model we then needed to estimate the components of the variance for each stage of sampling. The current method for estimating SOC variances utilizes a modified half sample (MHS) replication method (Thompson, 1998). The VPLX programs currently in place compute estimates of the total variances only, so we adapted the MHS method in order to estimate the variance components to be used in the optimal allocation formulas. We chose to estimate variances for housing starts estimates because housing starts are the most important product of SOC. Our replicate variance components obtained with these decompositions use unbiased estimates. The SOC variance estimate for an unbiased estimate $\hat{\theta}$ can be described as

$$V_{SOC}(\hat{\theta}) = V_{SUP}(\hat{\theta}) + V_{NP}(\hat{\theta}) + 2Cov_{SUP,NP}(\hat{\theta}) \qquad \textbf{(3)}$$

decomposed into four separate variance components

$$V_{SOC}(\hat{\theta}) = V_{\underset{(SUP+NP)}{Between-PSU}}(\hat{\theta}) + V_{\underset{(SUP)}{Between-Place}}(\hat{\theta}) + V_{\underset{(SUP)}{Between-Permit}}(\hat{\theta}) + V_{\underset{(NP)}{Between-Segment}}(\hat{\theta}) \qquad \textbf{(4)}$$

where the first variance component (the between-PSU component) implicitly includes a covariance term.

We estimated $V_{\underset{(SUP)}{Between-Place}}(\hat{\theta})$ and $V_{\underset{(SUP)}{Between-Permit}}(\hat{\theta})$ from the SUP data, and estimated $V_{\underset{(NP)}{Between-Segment}}(\hat{\theta})$ from the NP data. We obtained our between-PSU variance estimates by subtracting these directly-estimated variance components from $V_{SOC}(\hat{\theta})$. Thompson (1998) describes the procedure for obtaining the estimate of $V_{SOC}(\hat{\theta})$. We used a slightly modified version of that procedure for variance component estimation, applying coefficients of 1.5 and 0.5 in all replicates. We obtained $V_{\underset{(SUP)}{Between-Permit}}(\hat{\theta})$ directly with MHS replication. We then obtained $V_{\underset{(SUP)}{Between-Place}}(\hat{\theta})$ indirectly by estimating $V_{\underset{(SUP)}{Between-Place}}(\hat{\theta}) + V_{\underset{(SUP)}{Between-Permit}}(\hat{\theta})$ with replication and then subtracting the $V_{\underset{(SUP)}{Between-Permit}}(\hat{\theta})$ term. Since we treat the NP survey as a two-stage sample, the variance of $\hat{\theta}$, the unbiased (expansion) estimate of housing starts is given by:

$$V_{NP}(\hat{\theta}) = V_{\underset{(NP)}{Between-PSU}}(\hat{\theta}) + V_{\underset{(NP)}{Between-Segment}}(\hat{\theta}) \qquad \textbf{(5)}.$$

Again we obtained the between-segment component using MHS replication methods. The discussion above only briefly describes the estimation procedure we used for variance components. The details on the exact assignments of places, permits or segments to Hadamard matrices and the programs used to determine the variance component estimates can be found in detail in Ashley, Thompson (2002).

The levels and percentages of the individual variance components are quite variable from month to month, as expected. The previous redesign (and the production variance estimates) use six-month averages of variance estimates because of their "instability" from month to month. Individual monthly estimates of the total and between variances were so variable we used a 10 – month (2000, Jan-Aug, Nov, Dec) average of these estimates presented in Table 2 for allocation. We were unable to calculate variance estimates for September 2000 because of corrupted input data files. We excluded the housing starts variance component estimates for October 2000 because the estimated between-PSU variance was negative. The negative variance estimate resulted from an unusually high number of permits issued in two separate places assigned to the same panel in the same place pair, inflating the between-place variance estimate.

Table 2: Variance Component Estimates

	Variance	**Percent**
Between PSU	9,525,359	25.13%
Between Place	17,599,926	46.43%
Between Permit	8,315,813	21.94%
Between Segment	2,617,159	6.90%
Total	**37,902,935**	**100%**

Using the estimates of cost and variance determined by the methods described above we can minimize the variance at a fixed cost. This is equivalent to minimizing the product

$$S^2\left(C - C_o\right) = \left(\sum_{i=1}^{3} \frac{S_i^2}{n_i}\right)\left(\sum_{i=1}^{3} C_i n_i\right) \tag{6}$$

By applying the Cauchy-Schwartz inequality,

$$S^2\left(C - C_o\right) = \left(\sum_{i=1}^{3} \frac{S_i^2}{n_i}\right)\left(\sum_{i=1}^{3} C_i n_i\right) \geq \left(\sum_{i=1}^{3} S_i \sqrt{C_i}\right)^2 \tag{7}$$

it can be seen that the minimum values occur when

$$\frac{S_i}{\sqrt{n_i}} \propto \sqrt{C_i n_i} \qquad or, \qquad n_i = \lambda \frac{S_i}{\sqrt{C_i}}$$

λ is a constant to be determined from the following condition:

If C is fixed, then

$$C - C_o = \lambda \sum_{i=1}^{3} S_i \sqrt{c_i} \quad or, \quad \lambda = \frac{C - C_o}{\sum_{i=1}^{3} S_i \sqrt{c_i}}$$

Then, the optimal values of n_i will be

$$n_i(opt) = \frac{S_i(C - C_0)}{\sqrt{C_i} \sum_{i=1}^{3} (S_i \sqrt{C_i})} \qquad i = 1,\ 2,\ 3 \tag{8}$$

and the minimum variance is

$$S_{\min}^2 = \frac{\left(\sum_{i=1}^{3} (S_i \sqrt{C_i})\right)^2}{(C - C_0)}\ . \tag{9}$$

The S_i in (8) and (9) are calculated by multiplying the variances in Table 2 which are $\frac{S_i^2}{n_i}$, by the current sample sizes n_i and then taking the square root. Assuming a fixed sampling rate of 1 in 50 permits allows the between permit variance to be dropped from the variance model in (4). The resulting model predicts the portion of the SOC variance that is a function of n_1, n_2, and n_3. We will refer to this as the total design variance. Therefore the cost and variance models given by equations (2) and (4) respectively, with a fixed sampling rate will yield the optimal values of 90 NSR-PSUs, 671 NSR-Places and 80 Segments. The corresponding minimum total design variance is 26,962,631. Compared to the current sample, this allocation calls for a 24% reduction in the number of NSR-PSUs and a 46% increase in the number of NSR-Places. The optimal number of segments is similar (13% increase) to the current sample. For reasons of consistency we decided to maintain the current number of 169 PSUs and to keep the number of NSR-PSUs close to the current number of 119. The total number of permit places will be maintained around 900, with 671 NSR and the remaining designated as self-representing (SR). The total of non-permit segments sampled will be equal to the optimal number.

3. Self-Representing PSU Definitions and Stratification of NSR-PSUs

Unlike the current SOC sample, the 2004 design is not required to produce monthly estimates of selected Metropolitan Statistical Areas (MSAs). For this reason we did not combine any PSUs that CPS systematically split across state boundaries. The SOC frame of PSUs is equivalent to the set of sampled CPS PSUs. We obtained the following information for each PSU:

- Annual permit activity for 1998-2002 from BPS Annual Survey of all permit issuing places
- A weighted average of annual permit activity using weights of 1.0,1.2, 1.3, and 1.5 for the years 1999-2002 respectively
- CPS measure of size (non-institutional population age 16 and over)
- SR/NSR code from the CPS sample.

A self-representing PSU met the criteria that the PSU was self-representing in the CPS sample and one of the following:

- Weighted average of annual permit activity greater than 8000
- CPS measure of size was greater than or equal to the 95th percentile.

These criteria identified 48 PSUs to be SOC self-representing PSUs. States not represented by SR PSUs that were represented by the current SOC SR PSUs are Alaska, Delaware, DC, Hawaii, Kansas, Kentucky, Louisiana, Oklahoma, and Utah.

Similar to the current SOC design, the 2004 redesign will require one PSU to be selected from each stratum. There were 820 PSUs in the CPS Sample, of which 48 will be defined as SR for SOC as described previously. The remaining 772 NSR PSUs will be grouped into 121 strata within each of the nine Census divisions. PSUs were classified as metropolitan (met) if the majority of counties within the PSU were in metropolitan CBSAs, defined as of February 2003. Otherwise PSUs were classified as non-metropolitan (non-met). The PSUs were stratified so that they were similar in terms of metropolitan classification, weighted average of 1999-2002 permit activity, and total PSU population (non-institutional population age 16 and over based on 2000 Census). First PSUs were grouped by division and then separated into class intervals based on the weighted average of permit activity. Cumulative square root frequencies were calculated for each class interval and used to determine initial stratification boundaries. Once initial stratification boundaries were set, PSUs were grouped based on met/non-met classification and then population was used to further stratify within the pre-set boundaries. Table 3 shows the number of SR and NSR strata within the nine divisions and also shows the number of met and non-met NSR strata.

4. Selection of PSUs with Maximum Overlap

In many of the overlap methods available it is required that selection probabilities for the new sample be conditioned on the set of units in the intersection of the initial strata and the set of all possible old samples. Strata and PSU definitions are allowed to change. The sampling from stratum to stratum in the old design can be independent or dependent. In designs where dependency exists, the method may require knowledge of the joint selection probabilities in the old sample of sets of PSUs that are in the same stratum in the new design. However, Ernst's method for Maximizing Overlap when Information is Incomplete (1986) provides a way to maximize the overlap without knowledge of the joint probabilities of selection for all sets of old PSUs. This is a particularly important feature for SOC.

In the 1970 design the CPS strata were stratified into superstrata and one stratum selected from each superstratum. The SOC PSU was the sampled CPS PSU in the selected stratum. In the 1980 redesign the CPS sample PSUs, with some collapsing, were used as the SOC frame. The PSUs were then stratified into SOC strata and a sample of one PSU per stratum was taken while maximizing the overlap with the 1970 sample. In 1990 no new sample was taken. It is also important to note that our sample is taken from the CPS sample PSUs, which themselves are reselected with overlap of the previous sample every ten years. Due to the fact that we have sampled with overlap methods previously, the selection of our PSUs is not independent from stratum to stratum and thus we cannot calculate the joint probabilities of selection. However the joint probabilities are not required by Ernst's 1986 method.

Use of Ernst's method required identifying the 1980 SOC frame of PSUs and reconstruction of the 1980 SOC strata. The CPS masterfile provided the list of the 1980 CPS sample of PSUs which was sorted into the original 169 strata with the aid of a paper listing from 1980 of the PSUs within each stratum. Each stratum was then named according to its region and whether they were SR or NSR. For example, a stratum that was SR in the Midwest would have a name beginning MWSR followed by a number, an NSR stratum in the Midwest would just begin with MW. Numbers were assigned within regions separately for SR and NSR Strata. Thirteen PSUs on the masterfile remained unassigned to a stratum after using the paper listing. Each stratum was checked to see if all PSUs listed on the paper listing were actually present in the stratum. In several cases it was found that a PSU was not present, meaning that

Table 3: PSU Stratification Results

REG	DIV	SR	MET NSR	NON MET NSR	TOTAL STRATA
NE	**1**	1	5	3	9
	2	6	7	3	16
MW	**3**	6	13	6	25
	4	2	9	7	18
S	**5**	14	15	6	35
	6	1	7	3	11
	7	5	10	3	18
W	**8**	3	8	7	18
	9	10	6	3	19
TOTAL		**48**	**80**	**41**	**169**

it was not on the CPS masterfile. In some cases the missing PSU number was very close to the number of one of the 13 unassigned PSUs on the CPS masterfile. In these cases the PSU was assigned to the stratum with the missing PSU. The other unassigned PSUs were assigned to strata by locating the county of the PSU and the neighboring counties in an atlas, and then identifying the stratum containing those counties. The PSU was assigned to the stratum containing its nearest neighboring counties. There were no unassigned PSUs in the northeast region. The 1980 SOC Sample PSUs were then identified to verify that the stratum were constructed correctly, with only one sample PSU in each stratum. In 1980 several PSUs were combined and considered as a single PSU for SOC sampling purposes although they may have retained their original CPS PSU number. Taking these situations into account the strata reconstruction was found to be correct.

Ernst's method requires knowledge of the unconditional probabilities of selection (UPOS) in the current sample of each PSU in order to determine the conditional probabilities that will yield maximum overlap. In 1980 the SOC UPOS for each PSU is the ratio of the PSU civilian non-institutional 16+ population (CNP16+) to the total stratum population. The CPS masterfile included the measure of size used by CPS in 1980 that was equivalent to the CNP16+. However the PSU population totals were found to be slightly different than the four PSU populations that were reported in the current SOC documentation (SOC Technical Paper). We then obtained county level Census population data from 1980 for CNP16+. After aggregating this data to the PSU level, we again found the totals to be different from the 1980 documentation. The differences can be seen in Table 4.

Table 4: Differences in Population Data

1980 PSU Number	SOC Technical Paper	CPS Masterfile	Census Population Data
303	322,186	322,064	324,994
304	298,391	298,283	303,876
305	398,740	398,565	403,023
343	312,322	312,133	315,400

We then compared the UPOS calculated using both the masterfile and census data populations and found that the differences range in absolute value from 0 to 0.7253. Comparing these UPOS to the ones in the SOC Technical Paper we found that the UPOS calculated from the masterfile more closely matched the ones in the technical paper. This can be seen in Table 5.

Table 5: Differences in Unconditional Probabilities of Selection

1980 PSU Number	SOC Technical Paper	CPS Masterfile	Census Population Data
303	0.2420	0.24196	0.24122
304	0.2241	0.22410	0.22555
305	0.2994	0.29944	0.29914
343	0.2345	0.23450	0.23410

In addition to the differences discussed above the Census data was missing two counties that the masterfile covered. As a result of this comparison, we decided to use the CPS masterfile CNP16+ data in our calculations of the UPOS to be used in the maximum overlap procedure.

Selection of our sample with maximum overlap using Ernst's method involved identifying all possible ways in which the new SOC sample could overlap with the current SOC sample. In the twenty years since the 1980 redesign, CPS PSU definitions have changed. Counties that were grouped together in a current PSU may now be contained in one or more PSUs in the 2000 CPS sample. Since the primary purpose of maximizing overlap in our sample is to reduce costs by retaining areas already covered by SOC field representatives, we decided to define an old and a new PSU as overlapping if they had at least one county in common.

The first step was to create input files for each NSR stratum in the new design that specified the cost function to maximize and the restraints for the linear programming problem defined in Ernst's paper (1986) by equations 3.2 through 3.5 and 5.2. This involved identifying all of the 1980 strata that contained PSUs having at least one county in common with the PSUs contained in the new stratum. During this identification we determined the number of PSUs in the new stratum, the number of old strata with overlapping PSUs, and the number of PSUs within each of the old strata. Unconditional probabilities of selection were defined for PSUs in both designs. The old UPOS were based on CNP16+ as previously discussed, while the new UPOS are based on the weighted average of permit activity. Not all of the PSUs in either the old or new strata had counties in common. If an old PSU did not overlap with any of the PSUs in the new stratum than its UPOS in the new design was set to zero. Likewise if a new PSU did not overlap with any PSUs in the old strata, than its UPOS in the old design was set to zero. All such PSUs are considered as a single "dummy" PSU in the individual old or new strata during the overlap calculations.

A linear program written in Fortran (Fagan) uses the restraints specified for the given cost function to maximize the cost function for each stratum through linear optimization methods. The solution is then used to calculate the conditional probabilities of selection, as given in Ernst's paper, for each PSU in the new design. This requires a list of all possible old samples from the old strata having PSUs that overlap with PSUs in the new stratum. Once the conditional probabilities are assigned, the actual old sample combination is identified and the probabilities corresponding to that outcome were used to select the new sample of one PSU from each new stratum. Sample selection was accomplished using the cumulative conditional probability and a random number. Three of the 121 strata did not have any PSUs that overlapped with old PSUs, so the sample PSUs from those strata were selected by probability proportional to size (PPS) methods using the cumulative weighted average of permit activity and a random number. The PSU was selected if the random number was equal to or less than the cumulative weighted average. The unconditional probabilities of selection for all PSUs is the PSU weighted average of permit activity divided by the cumulative weighted average in the stratum. The PSU weights are then calculated as the product of the inverse of the CPS unconditional probability of selection and the inverse of the unconditional probability of selection in SOC.

5. Evaluation of New PSU Sample

In our overlap procedure a new PSU overlapped with an old PSU if at least one county in the new PSU was present in the OLD PSU. Thus it was possible for a single new PSU to overlap with several PSUs in the old design. Figure 1 illustrates the overlap of a new PSU with two old PSUs. Note that the new PSU has five counties total, two that overlap with one old PSU, two that overlap with a second old PSU, and one county that has no overlap. Overlap can be counted in two ways, either as the number of unique old PSUs or the number of unique new PSUs that fit the overlap criteria. It should be noted that these methods yield different results as illustrated in Figure 1. If we counted old PSUs we would have 2 overlapped PSUs, but if we counted new PSUs we would get only 1 overlapped PSU. Since the goal of overlap is to retain interviewers from old PSUs it seems more appropriate to count the number of old PSUs that are overlapped by new PSUs. Table 6 presents the counts and percentages for retained old PSUs compared to the overlap from a PPS sample taken from the set of CPS PSUs based on a cumulative weighted average of 1999-2002 yearly permit activity. This information is shown to support the assumption that the more complicated selection of the sample with overlap will save money. Note that the overlapped sample has 25 more PSUS in common with the old sample, resulting in a savings of $196,250 since cost per non-overlapped PSU is $7,850. The activity coverage of the overlapped sample was also compared to the PPS sample. The results (Table 7) show that the two sampling methods have similar percent coverage of total US permit activity.

Figure 1: Overlap of Old and New PSUs

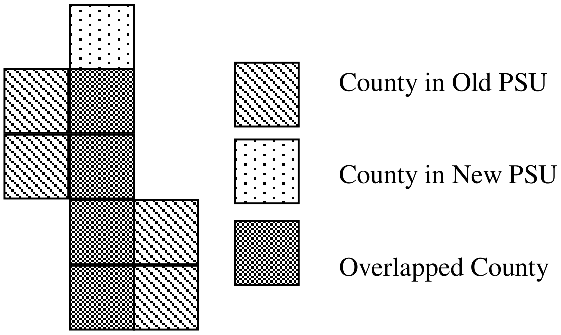

Table 6: Percentage of Overlap in Old and New Samples

	# of SR PSUs/ % of SR	# of NSR PSUs/ % of NSR	Total # of PSUs/ % of Total
Unique Old PSUs (74 SR, 123 NSR)	55 74.3%	47 38.2%	102 51.8%
PPS Sample Unique Old PSUs (74 SR, 123 NSR)	55 74.3%	22 27.3%	77 39.1%

Table 7: Permit Activity Coverage PPS, Overlapped and Current Samples

	2000 **# of permits/ % of US**	**2001** **# of permits/ % of US**	**2002** **# of permits/ % of US**
OVERLAP	878725 55.2%	898217 54.8%	950312 54.2%
PPS	873608 54.8%	893425 54.6%	938325 53.6%
Current	614339 38.6%	618934 37.8%	657650 37.6%
Total US Permit Activity	1591837	1637616	1750518

Table 7 also shows that the permit activity coverage of the current sample is about 16% less than the coverage in the PPS and overlapped samples, justifying the need for an updated sample.

6. Conclusions

The selection of SOC PSUs is the first stage of a multistage sample design. First, cost and variance models were constructed to determine the optimal number of sample PSUs that minimize variance at a fixed cost. The minimum total variance was determined to be 26,962,631 at a total monthly variable cost of $82,665.28. This cost is almost three times the fixed variable cost from the 1984 design, which is not surprising since the cost per PSU, place or NP area has increased from 2 to 5 times the 1984 cost. The optimal number of PSUs was determined to be 90 NSR-PSUs. The actual number of NSR-PSUs selected in the new sample was 121 as there are 119 NSR-PSUs in the current sample and there was a consensus that the total number of PSUs in the current sample could be maintained, despite the decrease specified by the optimal allocation. Forty-eight SR PSUs were designated based on high permit activity or a large 16+ population. The remaining NSR-PSUs were stratified within nine census divisions into 121 strata based on similar permit activity, metropolitan classification, and population. One PSU was selected per stratum using maximum overlap methods. Unconditional probabilities of selection used in the overlap methods were based on permit activity, instead of on population as in 1984. The purpose of the overlap methods was to reduce costs by retaining old PSUs where interviewers are already in place, since the cost of hiring and training new interviewers is high. The overlap methods produced a PSU sample with about 52% of old PSUs retained. This is a greater overlap than would be obtained by using a simpler method such as a PPS sample of PSUs which only produced 39% overlap in this case. It was also determined that the new sample has a larger percent coverage of US permit activity then the current sample. The new sample does not include any PSUs in the states of Alaska, Delaware or DC. These states were represented by SR PSUs in 1984, but were only represented by NSR PSUs in the new design which were not selected during the sampling process.

Future stages of the SOC sample redesign involve sampling building permit places and non-permit areas within each of the SOC sample PSUs. In the third stage, the selection of permits, variable sampling rates will need to be established for the sampling of permits within permit places. All new residential construction will be sampled with certainty in non-permit areas.

7. Acknowledgements

The authors wish to thank Katherine J. Thompson, Larry Ernst, Jim Fagan, and Bruce Williams for sharing their knowledge and providing their assistance throughout this sample redesign.

References

Ashley, James D. and Thompson, Katherine J. (2002) *Estimating Survey of Construction Variance Components for Housing Starts.* Washington, D.C. U.S. Bureau of the Census. (Internal Memorandum, available from the Manufacturing and Construction Division, Construction Programs Methodology Branch)

Ernst, Lawrence R. (1986) *Maximizing the Overlap Between Surveys When Information is Incomplete.* European Journal of Operational Research 27, 192-200.

Thompson, Katherine J. (1988) *Evaluation of Modified Half-Sample Replication for Estimating Variance for the Survey of Construction (SOC).* Washington, D.C. U.S. Bureau of the Census. (Technical Report #ESM-9801, available from the Economic Statistical Methods and Programming Division).