USE OF OVERLAP MAXIMIZATION IN THE REDESIGN OF THE NATIONAL COMPENSATION SURVEY

Lawrence R. Ernst, Yoel Izsak, Steven P. Paben
Ernst.Lawrence@bls.gov, Izsak.Yoel@bls.gov, Paben.Steven@bls.gov
Bureau of Labor Statistics, 2 Massachusetts Ave. N.E., Room 3160, Washington, DC 20212-0001

KEY WORDS: sample redesign, overlap maximization, transportation theory

1. Introduction

Following each decennial census in the United States, the Office of Management and Budget (OMB) releases a new set of area definitions. In June 2003 OMB released a set of area definitions that define a set of Core Based Statistical Areas (CBSAs) and designate the remaining geographical units as outside CBSAs counties. The CBSA areas are divided into Metropolitan Statistical Areas (MSAs) and Micropolitan Statistical Areas.

Most national surveys with an area-based design conduct a sample redesign every ten years following the release of the new area definitions by OMB in which a new set of areas are selected. One of these surveys is the National Compensation Survey (NCS) conducted by the Bureau of Labor Statistics (BLS).

The NCS sample is selected using a three-stage stratified design with probability proportionate to employment sampling at each stage. The first stage of sample selection is a probability sample of areas; the second stage is a probability sample of establishments within sampled areas; and the third stage is a probability sample of occupations within sampled areas and establishments.

In surveys that have a multi-stage stratified design with geographic areas as the primary sample units (PSUs), there generally are additional costs incurred with each change of sample PSU. Consequently, some of these surveys, rather then selecting the new sample PSUs independently of the sample PSUs in the previous design, employ a procedure to maximize the expected number of old PSUs retained in the new design. Such a procedure was used in the just completed selection of the new sample PSUs for NCS.

In the case of some surveys, such as the household surveys conducted by the Census Bureau, a key reason for desiring to retain as many old PSUs as possible is to reduce the expenses associated with hiring and training new data collectors. This is a somewhat lesser issue for NCS since much of the data for this survey is collected by Field Economists on travel outside their home areas. Possibly a more important reason for maximizing overlap of PSUs in NCS is that this survey program has a rotating panel design consisting of five rotating sample panels of establishments. Each of the five sample panels is in sample for five years before being replaced by a new panel selected annually from the most current frame. This aspect of the design, together with the fact that annual locality wage publications for as many of the sample areas as possible are among the key NCS products, make it undesirable to change sample PSUs. This is because if the transition from the old to the new sample areas takes place over a normal five year rotation, during at least a portion of that time it may not be possible to produce locality publications for many of the outgoing areas and incoming areas due to insufficient sample size.

There exist many procedures for increasing the expected overlap of sample units, most of which are described in Ernst (1999). They generally are based on the same key principal. That is, such a procedure does not alter the predetermined unconditional selection probabilities for any set of sample units in a new stratum, but conditions each such probability on the set of old sample units in such a manner that the conditional probability of a unit being selected in the new sample is in general greater than its unconditional probability when the unit was in the old sample and less otherwise.

We considered three overlap procedures for use in the NCS redesign. One is the procedure of Causey, Cox, and Ernst (CCE) (1985), which has the key advantage that it yields the true maximum overlap. CCE obtains the optimal overlap by formulating the overlap problem as a transportation problem, a special form of a linear programming problem. Despite this advantage, there are some disadvantages to this procedure that have generally kept it from being used in production, particularly the fact that CCE commonly results in transportation problems that are too large to solve operationally. As a result, we also considered two other procedures, those of Perkins (1970) and Ohlsson (1996), neither of which are difficult to implement operationally. For our particular NCS application, however, we found that the size of the transportation problems in CCE were quite manageable operationally, and we did select our new sample PSUs using it because of the substantially larger expected overlap that it yielded in comparison with the other two procedures considered.

We only overlapped new noncertainty MSAs with old noncertainty MSAs. The other two types of new PSUs, Micropolitan Statistical Areas, and clusters of outside CBSAs counties matched up too poorly with old PSUs for it to be worthwhile to overlap them.

In Section 2 we describe the three overlap procedures that we studied. The description assumes that the PSU definitions are identical in the old and new designs. Since that is not the case for NCS, we present in Section 3 two modifications of the overlap procedures that handle the case when the PSU definitions change in the new design. One of these is usable with all overlap procedures. The other is generally usable only with overlap procedures that formulate the overlap problem as a linear programming problem, that is, only CCE among the overlap procedures evaluated. In Section 4, we present an example using a stratum from NCS data that demonstrates the two approaches presented in Section 3 for the CCE procedure along with expected overlap calculations using Ohlsson's, Perkins', and independent selection. In Section 5 we compare, using production NCS data, the overlap obtained

for the three overlap procedures and for independent selection of the new PSUs.

2. The Three Overlap Procedures.

We proceed to describe the three overlap procedures presented in the Introduction. Since both the old and new designs in NCS are one PSU per stratum designs, we limit the description to this specific case. The description of CCE in their paper does not have any restrictions on the number of PSUs per stratum. Ohlsson (1996) is restricted to the one PSU per stratum case, but Ohlsson (1999) generalizes his procedure to any small number of PSUs per stratum. Perkins (1970) is also restricted to the one PSU per stratum case and has not been generalized, but it appears possible to do so.

The following notation will be used for all three procedures. Let *S* be a stratum in the new design consisting of *N* PSUs. (Each such *S* corresponds to a separate overlap problem.) Let $p_i, \pi_i,\ i = 1,...,N,$ denote the probability that the *i*-th PSU in *S* is in the initial and new samples, respectively.

In addition to describing the three overlap procedures, we present in this section expressions for the unconditional probability of overlap, that is the unconditional probability that the new sample PSU in *S* was in the old sample, for these overlap procedures and also for independent selection of the new PSUs.

2.1 CCE Procedure.

Let *I* denote the random set consisting of all PSUs in *S* in the old sample, and let $I_1,....,I_M$ denote all possibilities for *I*. A subset *I** of *S* can be among the I_i only if no more than 1 PSU from each old stratum is in *I**, with the further restriction that 0 PSUs in *I** is the only possibility for any old stratum that does not intersect *S* and 1 PSU in *I** is the only possibility for any old stratum that is a subset of *S*. For $i = 1,...,M,\ j = 1,...,N,$ let $p_i' = P(I = I_i)$; let $\pi_{j|i}$ denote the conditional probability that PSU *j* is selected from *S* as the new sample PSU conditional on $I = I_i$; and let T_i denote the set of initial strata that intersect *S* but do not intersect I_i. In the case when the PSUs in the initial design were selected independently from stratum to stratum, which is the situation for NCS, we then have

$$p_i' = \left(\prod_{k\in I_i} p_k\right)\left(\prod_{T\in T_i}(1-\sum_{k\in(T\cap S)} p_k)\right) \tag{2.1}$$

As for the $\pi_{j|i}$, they are not predetermined but are obtained by first finding the $x_{ij} \geq 0$ that satisfy the following transportation problem. Maximize

$$\sum_{i=1}^{M}\sum_{j=1}^{N} c_{ij}x_{ij} \tag{2.2}$$

subject to the constraints

$$\sum_{j=1}^{N} x_{ij} = p_i',\ \ i = 1,...,M \tag{2.3}$$

$$\sum_{i=1}^{M} x_{ij} = \pi_j,\ \ j = 1,...,N \tag{2.4}$$

where

$$\begin{aligned} c_{ij} &= 1 \text{ if } j \in I_i \\ &= 0 \text{ if } j \notin I_i \end{aligned} \tag{2.5}$$

Here x_{ij} is the joint probability that $I = I_i$ and *j* is the new sample PSU in *S*; (2.2), the objective function, is the probability that the new sample PSU in *S* was also in the old sample; while (2.3) and (2.4) are constraints required by the definitions of the $p_i', \pi_j,$ and x_{ij}.

Once the optimal x_{ij} have been determined by solving the transportation problem, we then have that

$$\pi_{j|i} = x_{ij} / p_i',\ \ j = 1,...,N \tag{2.6}$$

The unconditional probability of overlap for CCE, that is the unconditional probability that the new sample PSU in *S* was in the old sample is simply the value of objective function (2.2) for the optimal x_{ij}.

The CCE procedure has the key advantage that it produces the optimal expected overlap of two samples, when, as in our application, the two samples must be selected sequentially. If the two samples can be selected simultaneously, then a higher expected overlap can be obtained using the procedures of Ernst (1996, 1998), and Ernst and Paben (2002).

There are two major disadvantages to CCE. First it requires that the p_i' must be calculable. This is easily done by using (2.1) if the sampling in the old design had been done independently from stratum to stratum. However, as discussed in Ernst (1999), if the old sample had been obtained by overlapping with a still earlier sample, using CCE or most other overlap procedures, then this independence does not hold and calculation of the p_i' is generally not operationally feasible. Thus CCE generally cannot be used for two consecutive redesigns. Since the old set of sample PSUs for NCS was not selected using an overlap procedure, this was not an issue in our application.

The other drawback to CCE is that the number of variables in the transportation problem (2.2-2.4) can become impractically large. For example, as observed in Ernst (1999), if the PSUs in *S* were in *N* different noncertainty strata in the old design, then the number of x_{ij} would be $N2^N$ if both the old and new designs were 1 PSU per stratum, which would be extremely large for even moderately large *N*. However, for our NCS application the largest number of x_{ij} for any of the strata overlapped was 224.

2.2 Ohlsson's Procedure

Ohlsson (1996) suggests a procedure he calls exponential sampling for one PSU per stratum designs. In exponential sampling, each PSU *i* in the sampling frame of

the old design is independently assigned a permanent random number (PRN), X_i, where X_i is uniformly distributed on the interval (0,1). These same PRNs would then be used when taking all subsequent samples. However, if PRNs were not assigned to each PSU when taking the initial sample, it will still be possible to use Ohlsson's exponential sampling procedure to select a new sample, by retrospectively assigning PRNs to the PSUs, provided the PSUs in the initial sample were selected independently from stratum to stratum. We did not use PRNs in selecting our old sample, so we would need to assign these retrospective PRNs to our PSUs.

Retrospective PRNs are assigned to PSUs as follows. A temporary random number Z_i is independently assigned to each PSU *i* on the frame, where Z_i is uniformly distributed on the interval (0,1). Then, each PSU is assigned a retrospective PRN, based on whether or not it was selected in the old sample. For a PSU that was selected in the old sample (call this PSU *k*), its retrospective PRN is calculated as

$$X_k = 1-(1-Z_k)^{p_k} \tag{2.7}$$

The retrospective PRN for any other PSU *i* in the same old stratum as PSU *k* is given as

$$X_i = 1-(1-Z_k)^{p_i}(1-Z_i) \tag{2.8}$$

Once all of the PSUs in stratum *S* have been given retrospective PRNs, these PRNs are converted into transformed random numbers ξ_i by

$$\xi_i = -\frac{\log(1-X_i)}{\pi_i} \tag{2.9}$$

The PSU with the smallest ξ_i will be selected as the new sample PSU in *S*.

The unconditional probability of overlap for Ohlsson's method, as established in Ohlsson (1996), is

$$\sum_{i=1}^{N} \frac{p_i \pi_i}{\pi_i\left(\sum_{j\in A_i} p_j\right) + p_i\left(\sum_{j\in A_i'} \pi_j\right)} \tag{2.10}$$

where:

A_i is the set of all PSUs in the same old design stratum as *i*, except those PSUs in D_i;

A_i' is the set of all PSUs not in A_i that are in either the same old design or new design stratum as *i*;

and D_i is the set of all PSUs *j* that are in both the same old design stratum and new design stratum as *i*, and which also satisfy $p_i\pi_j > p_j\pi_i$.

Ohlsson's method has a few advantages over other overlap methods. It is relatively simple to implement. Also, another big advantage is that it preserves the independence of PSU selection from stratum to stratum in the new design, unlike most other overlap procedures. As a result of this independence, it can be used for two or more consecutive redesigns. In addition, this independence condition is desirable in variance estimation.

A major disadvantage of Ohlsson's procedure is that it requires the independence of PSU selections between strata in the initial sample, so if the initial sample had been selected using an overlap procedure that doesn't preserve this independence, then Ohlsson's procedure cannot be used at all. Another disadvantage is that, while Ohlsson's procedure will always increase the expected number of retained units over that of independent sampling, it does not produce an optimal overlap.

Ohlsson (1999) generalize the results in Ohlsson (1996) to designs with more than one PSU per stratum.

2.3 Perkins' procedure

Let I_j', $j=1,...,J$, denote the *J* old strata that intersect *S*. Perkins' method requires that it first be determined from which I_j' the new PSU is to be selected. To do this, we first calculate a probability y_j for each I_j', by summing π_i over all PSUs in I_j'. Then the selection among the I_j' is made with probability proportional to y_j.

Once a subgroup I_j' has been selected, the new sample PSU for S is selected from the set of PSUs in I_j', in the following manner:

If a PSU *k* in I_j' were selected in the old sample, the new sample PSU for *S* will be selected from among the PSUs in I_j' with the following probabilities conditional on *j* and *k*:

$$\pi_{k|jk} = \min\left\{\frac{\pi_k}{y_j p_k}, 1\right\} \tag{2.11}$$

$$\pi_{i|jk} = \left(1-\min\left\{\frac{\pi_k}{y_j p_k}, 1\right\}\right)\frac{\max\{\pi_i - y_j p_i, 0\}}{\sum_{\ell\in I_j'}\max\{\pi_\ell - y_j p_\ell, 0\}},\ i\neq k \tag{2.12}$$

If none of the PSUs in I_j' were selected in the old sample, the new sample PSU for S will be selected from among the units *i* in I_j' with a probability proportional to

$$\max\{\pi_i - y_j p_i, 0\} \tag{2.13}$$

Perkins does not provide an algorithm for computing the unconditional probability of overlap, but we present one, that is

$$\sum_{j=1}^{J}\sum_{k\in I_j'}\min\{y_j p_k, \pi_k\} \tag{2.14}$$

Note that (2.14) follows immediately from (2.11).

Similar to Ohlsson's procedure, Perkins' method is simple to implement, but will not yield an optimal overlap. However, unlike Ohlsson's procedure, Perkins' procedure does not select the PSUs in the new design independently from stratum to stratum. Perkins's procedure does have an advantage over both CCE and Ohlsson's procedure in that Perkins' procedure does not require that the PSUs in the old

design to have been selected independently from stratum to stratum, and consequently Perkins's procedure can be used to overlap samples in two or more consecutive redesigns, even if the PSUs had not been selected independently in the initial design.

2.4 Independent Selection

If the new PSUs are selected independently of the old PSUs then the probability of overlap for *S* is

$$\sum_{i=1}^{N} p_i \pi_i \qquad (2.15)$$

3. PSU Definition Changes

The 2003 area definitions released by OMB, are substantially different from the 1993 area definitions. The changes can range from the relatively mundane, such as the addition or deletion of a single county from an area definition, to the considerable, such as dividing an old area definition into multiple new areas. In addition, we are only overlapping noncertainty MSAs in the new design with noncertainty MSAs in the old design and our PSU frames for the two designs are limited to such PSUs. These facts required us to define what we considered an overlapped PSU between the old and new design to be. Additionally, the description of the three overlap methods as presented in Section 2 assumes that the PSU definitions are identical in the old and new designs. This is obviously not the case here. We tried two different approaches to handling such a complication. The first approach assumes a one-to-one correspondence between the PSUs in the old and new design, with the correspondence defining what constitutes an overlapped PSU. The second approach treats as a partial success an outcome for which some but not all counties in a new PSU were in the old sample, with the magnitude of the partial success depending on the proportion of the employment in the new PSU that are in counties that were in the old sample.

3.1 One-to-One Approach

This approach assumes each new area corresponds to one and only one old area and each old area to exactly one new area. This correspondence is found by the following short algorithm, which follows the procedure developed by Ernst and Ikeda (1992). All mentions of employment refer to the average frame employment for 2002, which is the measure of size used for PSU selection in the new design.

For simplicity, we refer in this subsection to the old design PSUs being overlapped as MSAs and to the new design PSUs being overlapped as CBSAs, although it must be understood that is only the MSAs among the new design PSUs that are being overlapped, not all CBSAs. Furthermore in both the old and new designs, the matching is limited to noncertainty PSUs.

1. CBSAs are sorted by employment, in descending order.
2. Within each CBSA, its component counties are grouped by the MSA that they belonged to in the old design, which creates one or more "mini-MSAs" in each CBSA.
3. We then look at each CBSA in order. The "mini-MSA" that makes up the largest proportion of employment in a CBSA will be matched to this CBSA. So, its full MSA will have a one-to-one correspondence to this CBSA for the purposes of the overlap process. If an MSA that would be selected as the match to a CBSA has already been matched to a previous CBSA, then the MSA with the next largest proportion of the CBSA's employment will be matched to this CBSA instead. If there are no MSAs to match to a CBSA that have not already been matched to another CBSA, then that CBSA will be matched to a "dummy MSA" for overlap purposes. A dummy MSA is assigned an old selection probability of 0 and can be assigned to any old stratum.
4. After all CBSAs have been matched, if any MSA remains unmatched to a CBSA, it will be matched to a "dummy CBSA" for overlap purposes. The dummy CBSA is assigned a new selection probability of 0 and can be assigned to any new stratum.

For example, the Greenville-Spartanburg-Anderson, SC MSA in the old design has been split into three metropolitan areas and a micropolitan area in the new design: Greenville, SC CBSA; Spartanburg, SC CBSA, Anderson, SC CBSA, and the micropolitan Gaffney, SC CBSA. Only the CBSA that makes up the largest proportion of the employment in the old MSA would be matched using this approach. That is, the Greenville, SC CBSA at approximately 59% of the old MSA employment would be the matched area. The other two metropolitan CBSAs would have "dummy MSAs" created and have an old selection probability of zero. The micropolitan area would not be taken into consideration, since we have limited the overlap process to only noncertainty metropolitan areas.

3.2 Partial Success Method

The general idea in this approach is that if PSU *j* was selected from stratum *S* to be the new sample PSU and some but not all of the counties in PSU *j* were in sampled PSUs in the old design, that is a partial overlap, then the outcome would be considered a partial success, with the proportion of success depending on the ratio of the new employment in the counties in PSU *j* that were in the old sample to the total new employment in PSU *j*. This is handled in the CCE approach by modifying the c_{ij} so they no longer are restricted to being either 0 or 1 in the case of one PSU per stratum designs. Similar modifications can be done for other overlap procedures that use linear programming, but we are not aware of how to modify overlap procedures that do not use linear programming, such as Ohlsson's and Perkins', to handle partial overlaps. Consequently, we present the modifications to account for partial success only for the CCE procedure.

The modifications are as follows. S, π_i, N are as in Section 2. S' is the set of PSUs in the old design that intersect at least one PSU in S. N' is the number of PSUs

in S' and p_i, $i = 1,..., N'$ is the probability that the *i*-th old PSU in S' was in the old sample. I is now the random set denoting the set of PSUs in sample in the old design that intersect at least one PSU in S, with $I_1,....,I_M$ denoting all possibilities for I and $p_i' = P(I = I_i)$. For $k = 1,..., N'$, $j = 1,..., N$, let f_{kj} denote the ratio of the new employment in the intersection of the *k*-th PSU in S' and the *j*-th PSU in S to the total new employment in the *j*-th PSU in S and let

$$c_{ij} = \sum_{k \in I_i} f_{kj}, i = 1,..., M, \; j = 1,..., N \qquad (3.1)$$

The transportation problem remains as given by (2.2-2.4).

For our particular application, since we are only overlapping noncertainty metropolitan areas in the new design with noncertainty metropolitan areas in the old design, we restrict S' to PSUs that were noncertainty metropolitan in the old design and in the calculation of f_{kj} the employment of PSU *j* is restricted to that portion of the PSU that was in a noncertainty metropolitan area in the old design. If this portion is empty for PSU *j*, then none of the f_{kj}, $k = 1,..., N'$ are defined and we simply set $c_{ij} = 0$, $i = 1,..., M$, that is, there is no measure of overlap success for PSU *j*. Consequently, we want to base the expected overlap for a stratum with such PSUs only on the measure of success for the subset, designated *S**, of *S* consisting of all PSUs for which at least a part of the PSU was in a noncertainty metropolitan PSU in the old design. Consequently, for such a stratum the expected overlap instead of being the maximum value of (2.2) is that value divided by $\sum_{j \in S^*} \pi_j$.

In a sense CCE as presented in Section 2 is a particular case of CCE presented here for which $S' = S, N' = N$; $f_{kj} = 1$ if $k = j$, and $f_{kj} = 0$ if $k \neq j$; and hence $c_{ij} = 1$ if $j \in I_i$, $c_{ij} = 0$ if $j \notin I_i$.

Although the partial success method does not appear to be usable with the other two overlap procedures considered, it can be used to measure the magnitude of success for independent selection, that is

$$\sum_{k=1}^{N'} \sum_{j=1}^{N} f_{kj} p_k \pi_j \qquad (3.2)$$

As in the case for CCE, if for some PSU in *S* no portion of the PSU was in a noncertainty metropolitan area in the old design, then (3.1) would be divided by $\sum_{j \in S^*} \pi_j$.

4. An Example Using NCS Data

We present an example of each procedure for a particular new stratum in NCS. Table 1 shows the new stratum with its three PSUs and their given new and old selection probabilities.

Table 1. New Stratum with Corresponding New and Old Selection Probabilities

j	CBSA Name	π_j	p_i
1	Oshkosh-Neenah, WI	.181	0
2	Saginaw-Saginaw Township, MI	.188	.244
3	Madison, WI	.631	.300

New PSU 1 had no possibility of selection in the old design using the one-to-one approach. This is because PSU 1 was previously part of the Appleton-Oshkosh-Neenah, WI MSA, with Appleton making up the largest proportion of the old MSA employment. Appleton is in a different stratum in the new design. This leaves PSU 1 without a corresponding match in the old design. Additionally, PSU 2 was in a different stratum from the other PSUs in Table 1 in the old design. PSU 2 was also split from a larger old PSU, the Saginaw-Bay City- Midland, MI MSA, in the old design. However PSU 2, unlike PSU 1, does make up the largest proportion of the old MSA employment. Using the one-to-one approach, there are four possibilities of selection in the old design presented in Table 2.

Table 2. Possible Sets of Old Sample PSUs and Corresponding p_i' for One-to-One Approach

i	Possible Sets of Old Sample PSUs	p_i'
1	{2}	.171
2	{3}	.227
3	{2,3}	.073
4	Ø	.529

Using the CCE procedure with the possibilities as shown in Table 2, the c_{ij}'s become as given in Table 3.

Table 3. Values of c_{ij} for One-to-One Approach

	j		
i	1	2	3
1	0	1	0
2	0	0	1
3	0	1	1
4	0	0	0

On maximizing (2.2) subject to (2.3) and (2.4) with the given π_j's, p_i''s, and c_{ij}'s in Tables 1, 2, and 3, respectively, an optimal set of x_{ij}'s, given in Table 4, is found.

Table 4. Values of x_{ij} That Maximize (2.2) for the One-To-One Approach

	j		
i	1	2	3
1	.000	.171	.000
2	.000	.000	.227
3	.000	.017	.056
4	.181	.000	.348

The maximum value of the objective function, that is the value of (2.2) corresponding to the x_{ij}'s in Table 4, is .471. In contrast, if the new sample for this stratum was selected independently of the initial sample we would have an expected overlap probability of only .235.

For both Ohlsson's and Perkins', it is also possible to calculate the expected overlap directly. For the example above, we obtain an expected overlap of .374 for Ohlsson and .235 for Perkins from (2.10) and (2.14), respectively. In this case, Perkins does no better than the independent selection. This will always be the case when all of the PSUs in the new design were in different strata in the old design.

The conditional probabilities of selection for CCE given that I_i is the set of initial sample PSUs in *S* can be obtained by dividing the *i*-th row of Table 4 by p_i'. Table 5 shows the conditional probabilities of selection for the one-to-one approach.

Table 5. Conditional Probabilities of Selection $\pi_{j|i}$ for the One-to-One Approach

	j		
i	1	2	3
1	.000	1.00	.000
2	.000	.000	1.00
3	.000	.232	.768
4	.342	.000	.658

For possible outcomes 1 and 2 the conditional probability of selection in the new sample would be certainty for a PSU that was in the old sample. Possible outcomes 3 and 4 do not have a certainty conditional probability. Possible outcome 3 occurs when PSUs 2 and 3 were selected in the old design within this stratum. Possible outcome 4 occurs when none of the PSUs were in the old design. For possible outcome 3, the procedure tries to select either PSU 2 or 3 and is able to do so. For possible outcome 4, the null set, the procedure divides up the remaining probability in order to meet the constraints. As for most of the other new strata, the only initial outcome for which the new sample PSU is not always in the old sample is the null set.

We next consider the partial success approach for the new stratum in Table 1. There are three noncertainty metropolitan areas in the old design that intersect at least one of the three PSUs in the new stratum, namely Appleton-Oshkosh-Neenah, Saginaw-Bay City-Midland, and Madison, which we label old PSUs 1,2, and 3 respectively. In this case, each new PSU was completely in a single noncertainty MSA in the old design. Consequently,

$$f_{11} = f_{22} = f_{33} = 1, \quad f_{kj} = 0 \text{ for all other } k, j \qquad (4.1)$$

Using the CCE procedure with the partial success approach, an overlap of new PSU 1 with old PSU 1 can now occur and $p_1 = .218$. Additionally, there are six possibilities of selection in the old design presented in Table 6. PSUs 1 and 3 were in the same old stratum. Therefore, they cannot be in the old sample together.

Table 6. Possible Sets of Old Sample PSUs and Corresponding p_i' for Partial Success Approach

i	Possible Sets of Old Sample PSUs	p_i'
1	{1}	.165
2	{2}	.118
3	{3}	.227
4	{1,2}	.053
5	{2,3}	.073
6	Ø	.364

The c_{ij}'s, obtained from (3.1), (4.1), would become as given in Table 7.

Table 7. Values of c_{ij} for Partial Success Approach

	j		
i	1	2	3
1	1	0	0
2	0	1	0
3	0	0	1
4	1	1	0
5	0	1	1
6	0	0	0

On maximizing (2.2) subject to (2.3) and (2.4) with the given π_j's, p_i''s, and modified c_{ij}'s for the partial success approach. We obtain the following set of x_{ij}'s in Table 8.

Table 8. Values of x_{ij} That Maximize (2.2) for the Partial Success Approach

	j		
i	1	2	3
1	.165	.000	.000
2	.000	.118	.000
3	.000	.000	.227
4	.016	.037	.000
5	.000	.033	.040
6	.000	.000	.364

The maximum value of the objective function is .636. In contrast, if the new sample for this stratum was selected independently of the initial sample we would have an expected overlap probability of only .274. Recall we do not know how to calculate an overlap for the partial success approach using Ohlsson's or Perkins'.

The conditional probabilities of selection given that I_i is the set of initial sample PSUs in *S* can be obtained by dividing the *i*-th row of Table 8 by p_i'. Table 9 shows the conditional probabilities of selection for the partial success approach.

Table 9. Conditional Probabilities of Selection for the Partial Success Approach

	j		
i	1	2	3
1	1.00	.000	.000
2	.000	1.00	.000
3	.000	.000	1.00
4	.313	.687	.000
5	.000	.459	.541
6	.000	.000	1.00

The addition of old PSU 1 as a possibility for inclusion in the initial outcome greatly affected the outcome of this stratum in the partial success approach.

5. Results Using NCS Data

First, we compare the unconditional overlap using the one-to-one approach for each method. Table 8 shows the expected proportion of new PSUs overlapped and the expected number of overlapped new PSUs using the one-to-one approach. The expected proportion of PSUs overlapped is calculated by first determining the expected overlap for each new stratum and then calculating the arithmetic average over all strata. The expected number of PSUs overlapped is simply the expected overlap for each new stratum summed over all new strata. There are 60 noncertainty metropolitan area strata in the new design.

Table 10. Expected Proportion and Number of PSUs Overlapped PSUs Using 1-to-1 Approach

Method	Expected Proportion Overlapped	Expected Number Overlapped
CCE	.479	28.7
Ohlsson's	.418	25.1
Perkins'	.310	18.6
Independent	.186	11.2

Since the CCE method produces an optimal expected overlap for each stratum, it obviously has the highest expected overlap. All three methods produce a higher expected overlap than selecting the PSUs in the new design independently. The proportion of the expected overlap may look low, but that has to do with the redesign of the NCS and the fact that we now have more noncertainty metropolitan strata. In actuality, there were only 32 PSUs in the new design that could be overlapped with the old design since there are only 32 noncertainty metropolitan areas in the old sample that are noncertainty metropolitan areas in the new design. If we calculate the expected overlap conditional on this set of 32 old sample PSUs, we obtain 30.8 as the expected number of PSUs using the CCE procedure, compared to 10.6 as the expected number of PSUs using independent sampling. Only two of these 32 old sample PSUs not retained with conditional certainty using the CCE procedure, one was due to the fact that there were two old sample PSUs in one new stratum, so both could not be retained. So, in actuality the CCE procedure almost performed perfectly.

Next, we compare the unconditional probabilities using the partial success approach. Table 9 shows the expected proportion of new PSUs overlapped and the expected number of overlapped new PSUs using the partial success approach.

Table 11. Expected Proportion and Number of PSUs Overlapped PSUs Using the Partial Success Approach

Method	Expected Proportion Overlapped	Expected Number Overlapped
CCE	.547	32.8
Independent	.206	12.4

The average expected proportion of PSUs overlapped with CCE was .547 compared to the .206 for independent selection. Due to the fact that more than one PSU in the new design can be matched to the same PSU in the old design, there are more potential overlaps using the partial success approach than the one-to-one approach.

Since the one-to-one and partial success approach are so different, it is not appropriate to compare the average expected overlaps between the two approaches. Therefore, we are left to decide between the two methods by using the approach that we feel is better suited to NCS. Our concern with using the partial success approach arises from cases where a PSU in the design is split in the new design. For example, as mentioned earlier in section 3 the old Greenville-Spartanburg-Anderson, SC MSA has been divided into three new PSUs. The three new PSUs are all in different strata in the new design. With the partial success approach, if the MSA was in sample in the old design, the overlap program would attempt to select all three of the corresponding CBSAs in the new design, increasing the probability that all three PSUs are in the sample together. We do not consider it a desirable outcome to have two or more (in this case three) adjacent CBSAs in sample together. However, using the one-to-one approach, the Greenville, SC CBSA was the corresponding match for the old MSA. Meanwhile, the other two CBSAs were left without a corresponding match, and, therefore matched to dummy PSUs. Consequently, when the Greenville-Spartanburg-Anderson, SC MSA was in the old sample, the program attempts to select Greensville, SC CBSA in the new sample, but not the other two new PSUs, thereby avoiding the problem of increasing the probability of selecting two or more adjacent CBSAs.

In selecting the actual NCS sample, we decided to use the CCE procedure with the one-to-one approach. We ended up selecting 31 of the 32 old noncertainty PSUs in the new design. Whereas, if we had selected the new sample independently, the expected number of the 32 old noncertainty PSUs selected in the new sample would have been 10.6. This selection retained approximately 86% of the employment from the old 32 sampled PSUs. The percentage of employment retained is somewhat less than 31/32 primarily due to employment loss in the PSUs that were split in the new design.

6. References

Causey, B. D., Cox, L. H., and Ernst, L. R. (1985). Applications of Transportation Theory to Statistical Problems. Journal of the American Statistical Association, 80, 903-909.

Ernst, L. R. and Ikeda, M. (1992). Modification of the Reduced-Size Transportation Problem for Maximizing Overlap When Primary Sampling Units Are Redefined in the New Design. Bureau of the Census, Statistical Research Division, Technical Note Series, No. TN-92/01.

Ernst. L. R. (1996). Maximizing the Overlap of Sample Units for Two Designs with Simultaneous Selection. Journal of Official Statistics, 12, 33-45.

Ernst, L. R. (1998). Maximizing and Minimizing Overlap When Selecting a Large Number of Units per Stratum Simultaneously for Two Designs. Journal of Official Statistics, 14, 297-314.

Ernst, L. R. (1999). The Maximization and Minimization of Sample Overlap Problems: A Half Century of Results. International Statistical Institute, Proceedings, Invited Papers, IASS Topics, 168-182.

Ernst, L. R. and Paben, S. P. (2002). Maximizing and Minimizing Overlap When Selecting Any Number of Units per Stratum Simultaneously for Two Designs with Different Stratifications. Journal of Official Statistics, 18, 185-202.

Ohlsson, E. (1996). Methods for PPS Size One Sample Coordination. Institute of Actuarial Mathematics and Mathematical Statistics, Stockholm University, No. 194.

Ohlsson, E. (1999). Comparison of PRN Techniques for Small Sample Size PPS Sample Coordination. Institute of Actuarial Mathematics and Mathematical Statistics, Stockholm University, No. 210.

Perkins, W. (1970). 1970 CPS Redesign: Proposed Method for Deriving Sample PSU Selection Probabilities Within 1970 NSR Strata. Memo to Joseph Waksberg, U.S. Bureau of the Census.

Any opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics.