Redesigning the Consumer Price Index Area Sample

William H. Johnson, Owen J. Shoemaker, and Yeon W. Rhee
U. S. Bureau of Labor Statistics, 2 Mass Ave NE, Room 3655, Washington, DC 20212

KEY WORDS: multistage, stratified, controlled selection, overlap

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

This paper describes the PSU selection process for the next CPI Revision. The U. S. Consumer Price Index (CPI) employs a multistage sample design that has been revised every ten years. The first stage consists of selecting primary sampling units (PSUs) which are formed from Metropolitan or Micropolitan Core Based Statistical Areas (CBSAs) based on preliminary definitions by the Office of Management and Budget.

The PSU selection process for the next CPI Revision is quite similar to the process of selecting the sample for the 1998 CPI Revision (see Williams et al). The biggest difference has been the use of variance models of six-month index change for the Commodities and Services part of the CPI-U in determining the set of certainty PSUs and the distribution of non-certainty PSUs across Census region by size class combinations. Alternative methodologies for stratifying PSUs prior to selection were considered and work on modeling CPI-U change since 1992 influenced the selection of stratifying variables. All of the programs involved in the work on selecting the 1998 CPI Revision PSU sample were updated or rewritten.

The process of selecting the PSU sample involves six steps:

1) Determine the PSUs selected with certainty
2) Determine the number of non-certainty PSUs and their distribution across regions
3) Stratify the non-certainty PSUs
4) Use Keyfitzing to improve expected overlap
5) Use controlled selection to generate a set of sampling patterns and weights
6) Select a sample of PSUs

Determining Certainty PSUs

The first step in the process of selecting the PSU sample is to determine which PSUs are certainty PSUs. In order to determine the certainty PSUs it was necessary to determine the possible certainty PSUs. The most likely certainty PSUs are those which are already certainty PSUs in the existing CPI area sample. However with the shift to CBSA based definitions it became necessary to determine what the new definitions of the current certainty PSUs are likely to be. The certainty cities were mapped along with preliminary CBSA definitions. It was assumed that a CBSA would either be entirely included or entirely excluded from these areas. In cases where a CBSA was partially contained in a current certainty PSU, the probability of the outside counties being in the final definition given to BLS by the Census Bureau was examined as part of the assessment of whether to include or exclude the CBSA.

After the expected definitions of the current certainty cities were decided, the remaining possible certainty cities were the remaining individual metropolitan CBSAs. The largest metropolitan CBSAs outside of the current certainty cities were determined and considered for inclusion in the list of new certainty PSUs.

Next it was necessary to determine the criteria for PSUs to be selected with certainty. There were several possible options. The entire CPI-U population to be represented is the total population contained in all metropolitan and micropolitan CBSAs. This population is 257,010,167.

The options considered included:

1) 1,500,000 – the population cutoff used previously for determining certainty cities
2) 1,680,000 – a population cutoff that wouldn't cause the loss of any current certainty cities
3) 1,800,000 – a population cutoff which was considered for use previously
4) 2,141,751 – the population cutoff obtained by using 120 half sample equivalents (HSEs) to represent the total population of 257,010,167
5) 2,570102 – the population cutoff obtained by using 100 HSEs to represent the population of 257,010,167
6) 4,283,501 – the population cutoff obtained by using 60 HSEs to represent the population of 257,010,167

A half sample equivalent is a unit of sample size. Each certainty city will receive at least two HSEs and each selected non-certainty city will receive one HSE.

The option of using 1,500,000 as a population cutoff for determining certainty cities was dropped as it would add too many certainty cities to be affordable. Each certainty city must have enough sample for their individual city CPIs to be publishable on at least a semi-annual basis.

This makes the certainty cities much more costly than non-certainty cities.

The decision as to which set of cities should be selected with certainty required information so one could compare the various possible sets of certainty PSUs. In order to compare the various options, the model used for optimizing the CPI Commodities and Services sample was generalized. (see Leaver et al) This model attempts to select outlet and item sample sizes for groups of PSUs which will produce the lowest variance given the available budget for travel and data collection. The model was generalized by allowing the number of non-certainty PSUs in each non-self representing index area to be a variable that the optimization program could optimize over. This created the need for an additional constraint though as the number of non-certainty PSUs was determined by the total number of HSEs minus the number of HSEs used by the certainty PSUs.

In addition, the relative importances for each index area and group of items had to be recalculated for each scenario. The populations used for calculating the population relative importances were from the 2000 Census. The cost weights used for calculating the relative importance of groups of items were from the 1999 Consumer Expenditure survey. An index area as used in this paper is either a certainty city or a Census region by size class combination. There are four Census regions: Northeast, Midwest, South, and West. There are two size classes corresponding to metropolitan and micropolitan CBSAs. Note that some micropolitan CBSAs are part of the current certainty cities and thus their population should be included with the certainty city and not with the non-self representing index area covering micropolitan CBSAs in the Census region in which the PSU resides..

Some additional options were explored. Even though we currently allocate one HSE to each non-certainty PSU, there was interest in what would happen if two HSEs were allocated to each non-certainty PSU. It would be expected to roughly halve the number of non-certainty PSUs, but the effect on variance was less obvious. Also, there was concern that the grossly uneven relative importances of the index areas may have a negative impact on sample allocation and on the variance of the all U.S. – all items CPI-U. Thus an option was explored where the largest Census region, the South, was broken apart using Census divisions. The South was divided into two index areas, one being the South Atlantic division and the other index area being composed of the East South Central and West South Central divisions. New variance components for the optimization model were calculated for the new index areas.

The optimization model yielded a result with non-integer numbers of PSUs in each non-self representing index area. These values were rounded to even integers in such a way that the total number of HSEs added up correctly. The optimization model was then rerun using these fixed numbers of PSUs to provide results that could be compared with results from other scenarios. The information was used in determining what the set of certainty PSUs would be.

The list of certainty PSUs is not yet public information and can't be included in this paper. Some of the results that were found can be discussed. In comparing the allocation of one vs. two HSEs to each non-certainty PSU, it was found that allocating two HSEs to each non-certainty PSU increased the modeled standard error of six month CPI change for C&S by an average of 13.6% across the scenarios. This was primarily due to the large contribution of the between PSU component of variance in non-self representing index areas. This was surprising given that the PSU components of variance are so small compared to other components of variance. However the much smaller divisor of the PSU component of variance as compared to other components allowed it to have a greater contribution to the total variance. In all cases the PSU component of variance ended up contributing more than 50% of the total variance for all of the index areas representing metropolitan CBSAs.

Dividing the South based on Census divisions also ended up increasing the total variance. It appears based upon the model used that it is preferable to have fewer and larger index areas with larger PSU samples than to have a larger number of smaller index areas. This is again a result of the large contribution of the between PSU component of variance of non-self representing index areas.

Once the decision was made on a set of certainty PSUs, the number of PSUs in each non-self representing index area was also determined based on the output of the optimization program from the chosen scenario. The chosen design did shift towards having more PSUs in the West region and slightly fewer elsewhere. There are more of what are called G-size PSUs as the population they cover has grown greatly in relative importance between 1990 and 2000. For the 1998 CPI Revision sample, the C PSUs were the urban part of areas outside of metropolitan statistical areas. The C PSUs now represent the micropolitan CBSA population, excluding those CBSAs which are part of a certainty PSU. Having the CPI-U population be the total population in CBSAs resulted in an increase in the total percent of the U.S. population covered by the CPI-U.

Stratifying Non-Certainty PSUs

Non-certainty PSUs are grouped together into strata and one PSU is selected from each stratum. (see Dippo et al) It is desirable that the PSUs within a stratum be homogeneous. The first task was to determine by what measure the PSUs should be homogeneous.

It the early 1990's, work was done on modeling CPI-U change for certainty PSUs by variables we had available from Census as well as geographic variables. None of these models were especially promising. However, for the 1998 CPI Revision, a four variable model using normalized latitude, normalized longitude, normalized latitude squared, and percent urban was chosen for use in three out of four Census regions and a model consisting of seven Census variables was chosen for the South region. Once a model was chosen, the strata were formed so as to be as homogeneous as possible with respect to these variables, subject to the restriction that strata should have roughly equal population. (see Williams et al)

This research was updated by examining the predictive power of these models for more recent time periods as well as examining their value in modeling CPI-U change for non-self representing PSUs and for modeling changes in the housing index. The chosen models have performed worse since they were originally researched and no other really good models have been found. Thus the chosen model this time was simply the four variable model from before with normalized longitude squared included for the purpose of symmetry.

Given the relatively weak predictive power of the chosen model, two other options were also examined: Using no stratification and a purely geographic stratification.

With no stratification, the PSUs would be drawn from each region by size class without replacement and with probability proportional to expenditure. This was done for simulation purposes with SAS PROC SURVEY SELECT.

The purely geographic stratification was based on Peano ordering the PSUs based on the median latitude and longitude of the centroids of the counties composing the PSUs. Examples of Peano curves can be found at http://www.contrib.andrew.cmu.edu/~malin/java/PeanoHilbert.html. The Peano curve for a $2^N \times 2^N$ grid is based on a recursive N-shaped pattern. In each region by size class combination, the points representing the PSUs were placed on a $2^{20} \times 2^{20}$ grid. The calculation of an ordering value is based on interleaving the digits of the binary representations of the coordinates of the PSUs. Once the PSUs are ordered, the ordered list of PSUs in each region by size class is cut into the appropriate number of strata. The cut points are made so that the population in each stratum is roughly the same. It was also attempted to make the cut points such that when there was a large jump in the calculated ordering value between two points then the two points would fall in different strata. This purely geographic stratification ended up producing strata which looked like rectangular stripes.

In order to cluster PSUs to be similar according to the five variable model discussed above, a program using a hill climbing algorithm by Friedman and Rubin was used. This program first rescales all of the variables so that they are of roughly equal importance. It does this by calculating an unstratified population weighted sum of squares for each of the variables and then multiplies the values of the variables by ten divided by the square root of the sum of squares:

$$V'_{i,j} = V_{i,j} * \frac{10}{\sqrt{Total\ Population * \sum_j PSUpop_j * (V_{i,j} - \overline{V}_{i,j})^2}}$$

where

$V_{i,j}$ is the value of the ith variable for the jth PSU

$PSUpop_j$ is the population of the jth PSU

$$Total\ Population = \sum_j PSUpop_j$$

$$\overline{V}_{i,j} = \sum_j \frac{PSUpop_j}{Total\ Population} * V_{i,j}$$

The program then attempts to minimize the stratified total sums of squares

$$\sum_{i,s} Stratumpop_s * \sum_{j \in s} PSUpop_j * \left(V'_{i,j} - \sum_{k \in s} \frac{PSUpop_k}{Stratumpop_s} * V'_{i,k} \right)^2$$

given the total number of strata, which is an input to the program. This program repeats the minimization procedure to form strata in each Census region by size class. The program is constrained on the size of the strata, and these constraints were estimated using the minimum and maximum stratum populations from the geographic stratification and adjusting them by 10%.

Keyfitzing to increase overlap

Given our budgetary limitations, it is generally desirable to keep as many of our current PSUs in the next sample as possible.

The first step was to determine what is meant by an overlap PSU. Given the considerable changes in definitions of the PSUs it is possible that part of a PSU might currently be in the CPI sample but not other parts. The preliminary definition was that 30% of the counties or 30% of the 2000 population of a PSU currently be covered by the CPI sample. This was complicated by the fact that counties are composed of Minor Civil Divisions (MCDs)

in the Northeast region. Current CPI PSUs in the Northeast are defined at the MCD level, while the new PSUs are defined at the county level. It was decided that a county composed of MCDs was overlap if at least 5% of its 2000 population was overlap. A PSU composed of MCDs is considered overlap as long as 30% of the counties are overlap and at least one of those counties has at least 30% of its 2000 population being overlap based on MCDs.

The inherited Keyfitzing procedure attempts to increase the likelihood of selecting PSUs which are overlap, or which have a greater relative importance in 2000 than in 1990. Some changes in the program had to be made due to the massive redefinition of PSUs. The Keyfitzing procedure operates at the level of the intersection of a new stratum with a stratum for the 1998 CPI Revision PSU sample. Due to redefinitions, there are many cases where only part of a new PSU lies within one of these intersections. Thus the PSUs were broken in pieces for the purpose of Keyfitzing and then the pieces were added together to give the total new probability of selection of a PSU.

The procedure works as follows:
For each Region X City Size X New Stratum$_i$ X Old Stratum$_j$ calculate the new probability of the PSU k or the part of PSU k being selected:

$$Newoldpr_{i,j,k} = \frac{newprob_k}{\sum_{l \in new_i \cap old_j} newprob_l}$$

where $newprob_k$ is the probability of selection of the intersection of PSU k with new stratum i and old stratum j.

There are several possible cases:
a) The intersection is empty so there are no PSUs to consider
b) The intersection is a single PSU k. Then the Keyfitzed probability is $Keyfitz_k = newprob_k$
c) There is no PSU in the intersection which was selected in the old sample:
For each PSU k in the intersection assign the Keyfitz probability as
If $Newoldpr_{i,j,k} \le oldprob_k$ then $Keyfitz_k = 0$
If $Newoldpr_{i,j,k} > oldprob_k$ then

$$Keyfitz_k = \frac{Newoldpr_{i,j,k} - oldprob_k}{\sum_{l \in new_i \cap old_j} \max(Newoldpr_{i,j,l} - oldprob_l, 0)} * \sum_{l \in new_i \cap old_j} newprob_l$$

Here $oldprob_k$ is the probability of selection of PSU k intersected with new stratum i and old stratum j based on 1990 populations.

d) A PSU s was selected in the old sample and at least partially resides in the intersection:
If $Newoldpr_{i,j,s} \ge oldprob_s$ then

$$Keyfitz_s = \sum_{l \in new_i \cap old_j} newprob_l$$

$Keyfitz_k = 0$ for all other PSUs k within the intersection.
Here the new and old probabilities are based on the old PSU definition for PSU s intersected with new stratum i and old stratum j. The Keyfitz probability for new PSUs within the intersection of new stratum i and old stratum j is calculated by determining the percentage of 2000 population of the old PSU s resides within each of the new PSUs.
e) A PSU s was selected in the old sample and at least partially resides in the intersection:
If $Newoldpr_{i,j,s} < oldprob_s$ then

$$Keyfitz_s = \frac{newprob_s}{oldprob_s} \sum_{l \in new_i \cap old_j} newprob_l$$

If k is a PSU in the intersection other than s, then
if $Newoldpr_{i,j,k} \le oldprob_k$ then $Keyfitz_k = 0$
else if $Newoldpr_{i,j,k} > oldprob_k$ then

$$Keyfitz_k = \left(\sum_{l \in new_i \cap old_j} newprob_l \right) * \left(1 - \frac{newprob_s}{oldprob_s} \right) * \frac{Newoldpr_{i,j,k} - oldprob_k}{\sum_{l \in new_i \cap old_j} \max(Newoldpr_{i,j,l} - oldprob_l, 0)}$$

After this procedure has been done for each intersection of new and old strata then the PSUs are reaggregated and their total probabilities of selection are determined.

The selection of a stratification was made on the basis of the total expected number of overlap PSUs. It turned out that the stratifications with the highest overlap were from the clustering procedure using normalized latitude, normalized longitude, normalized latitude squared, normalized longitude squared, and percent of population which is urban. As the clustering procedure had been run multiple times, there was usually more than one stratification to choose from in each Census region by size class. It turned out that having a lower total sums of squares did not equate with having higher expected overlap.

The following table summarizes the expected number of overlap PSUs for the various options examined, both pre- and post-Keyfitzing:

Region – City size	#overlap PSUs No stratification	#overlap PSUs Peano ordering	#overlap PSUs clustering program
X100	1.07	1.01	0.98
X200	5.00	4.56	4.30
X300	5.06	5.01	4.96
X499	1.45	1.37	1.40
C100	0.10	0.05	0.05
C200	0.24	0.23	0.22
C300	0.27	0.30	0.30
C400	0.35	0.34	0.34
X000	12.58	11.95	11.64

Region – City size	#overlap PSUs Peano ordering after Keyfitzing	#overlap PSUs clustering program after Keyfitzing
X100	2.43	2.82
X200	6.56	8.40
X300	6.44	7.59
X499	3.10	4.46
C100	0.05	0.05
C200	0.23	0.22
C300	0.30	0.30
C400	0.34	0.34
X000	18.53	23.27

Controlled selection of PSUs

It is hoped that the number of overlap PSUs selected is not much less than the expected number of overlap PSUs. Thus a procedure called controlled selection was used. A program used to do the controlled selection for the 1998 CPI Revision PSU sample could not be successfully compiled and run in our current computing environment. An alternative called PC Consel (see Lin) was investigated. We had some success with this program, however in the South region it would not give a solution as apparently no exact solution to the controlled selection problem exists. Thus a new SAS IML program was written in order to handle the controlled selection problem.

The following is a description of the controlled selection problem:

Create a 3-dimensional grid of stratum x state x overlap status. Sum the probabilities of selection of the PSUs in each cell. A pattern describes an entire sample. In each cell it has either a zero (select zero PSUs from this cell) or one (select one PSU from this cell). The controlled selection problem is to find a set of patterns P_i, $i = 1,...,n$ with probabilities of selection p_i such that $\sum_{i=1}^{n} p_i * P_i(x, y, z) = C(x, y, z)$, where $P_i(x, y, z)$ is the value of zero or one for the ith pattern for stratum x, state y, and overlap status z and $C(x, y, z)$ is the sum of probabilities of selection of PSUs in the cell for stratum x, state y, and overlap status z.

In addition there are constraints with respect to the number of PSUs selected per state and per overlap status. These constraints are imposed on each individual pattern. Let $S_i = \sum_x \sum_z C(x, i, z)$ be the total probability of PSUs in state i. Let $\underline{S}_i = floor(S_i)$ be the integer part of S_i. Then each pattern must contain either $\underline{S}_i$ or $\underline{S}_i + 1$ PSUs in state i. The sum of probabilities of patterns having $\underline{S}_i$ PSUs is $1 - (S_i - \underline{S}_i)$ and the sum of probabilities of patterns having $\underline{S}_i + 1$ PSUs is $S_i - \underline{S}_i$.

Let $O = \sum_x \sum_y C(x, y, 1)$ be the sum of probabilities of selection of overlap PSUs across all strata and states. Let $\underline{O} = floor(O)$ be the integer part of O. Then each pattern must select $\underline{O}$ or $\underline{O} + 1$ overlap PSUs. The sum of probabilities of patterns with $\underline{O}$ overlap PSUs is $1 - (O - \underline{O})$ and the sum of probabilities or patterns with $\underline{O} + 1$ overlap PSUs is $O - \underline{O}$.

The above constraints on the set of patterns comprises the controlled selection problem. Once this problem is solved, a pattern is selected based on the probabilities of the patterns. If there is more than one PSU corresponding to a cell with a value of one, then a single PSU is selected with probability proportional to its probability of selection within its stratum.

Note that there isn't necessarily a solution for the controlled selection problem. If there is no exact solution, then it is desirable to have a partial set of patterns P_i, $i = 1,..., n$ which have a sum of probabilities as close to one as possible.

The program randomly generates patterns by selecting a value of zero or one in each cell of the pattern using the probability in that cell. The program then verifies that the pattern meets the state and overlap constraints. If the pattern violates any constraints then the pattern is discarded and a new pattern is generated. If the pattern meets the state and overlap constraints then the pattern is kept and it is assigned a probability. The probability assigned to the pattern is the smallest remaining probability in any cell where a PSU was selected or the

smallest remaining probability of the state and overlap controls met:

Let $\min cell_i = \min_{x,y,z} C(x,y,z) * P_i(x,y,z)$

For each state i, the associated probability with the constraint is $1-(S_i - \underline{S}_i)$ if $\underline{S}_i$ PSUs are selected and $S_i - \underline{S}_i$ if $\underline{S}_i + 1$.

For the overlap constraint, the associated probability is $1-(O-\underline{O})$ if O overlap PSUs are selected in the pattern and $O-\underline{O}$ if O+1 overlap PSUs are selected.

The probability assigned to the pattern is the minimum of the cell probabilities, the state constraint probabilities, and the overlap constraint probability.

Once the pattern has a probability, that probability is deducted from each cell where a PSU was selected as well as from the state and overlap constraints met. For example, if the pattern probability is 0.2 and the number of PSUs in a state with 2.4 expected PSUs is 2, then the 0.6 probability initially assigned to selecting 2 instead of 3 PSUs in that state would be reduced to 0.4.

The new problem with the probabilities subtracted now goes through the same procedure until all probability is exhausted.

The way the patterns are constructed and the probabilities assigned, the sum of probabilities of patterns where a given PSU is selected will add up to the probability of the given PSU being selected. In addition, the probabilities associated with the state and overlap constraints will add up properly.

References:

Dippo, Cathryn S., and Jacobs, Curtis A., "Area Sample Redesign for the Consumer Price Index," *Proceedings of the Survey Research Methods Section*, American Statistical Association, 1983, 118-123.

Leaver, Sylvia G., Johnson, William, Shoemaker, Owen, and Benson, Thomas S., (1999) "Sample Redesign for the Introduction of the Telephone Point of Purchase Survey Frames In the Commodities and Services Component of the U.S. Consumer Price Index ," *Proceedings of the Section on Government Statistics, American Statistical Association,* 1999, 292-297.

Lin, Ting-Kwong, "Some Improvements on an Algorithm for Controlled Selection," *Proceedings of the Survey Research Methods Section*, American Statistical Association, 1992, 407-410.

Williams, J.L., Brown, E.F., Zion, G.R., "The Challenge of Redesigning the Consumer Price Index Area Sample," Proceedings of the Survey Research Methods Section , American Statistical Association (Vol. 1), 1993, 200-205.