THE CHALLENGE OF REDESIGNING THE CONSUMER PRICE INDEX AREA SAMPLE

Janet L. Williams and Eugene F. Brown, BLS, and Gary R. Zion, Natl. Inst. of Dental Research
Janet L. Williams, U. S. Bureau of Labor Statistics, 2 Mass Ave NE, Rm 3655, Washington, DC 20212

KEY WORDS: multistage, stratified, controlled selection, overlap

The U. S. Consumer Price Index (CPI) uses a multistage sample design which is revised approximately every ten years. The first stage consists of selecting primary sampling units (PSUs). The selected PSUs are also used in the Consumer Expenditure Survey (CE) and the Consumer Point of Purchase Survey (CPOPS). This paper describes the recently completed PSU selection process for the 1998 CPI Revision and the PC software developed to implement this process.

1. Introduction

Since the steps followed in the selection of the 1998-Revision PSUs are basically the same as those followed for the 1987 Revision, the reader is referred to the paper by Dippo and Jacobs (1983) for a description of that selection and the *BLS Handbook of Methods* (1992) for more extensive background material on the CPI. This paper will highlight the differences in the methodology and the sample design between the 1987 and 1998 Revisions.

For various reasons, crossing survey and agency lines, the new CPI PSUs needed to be selected by April, 1992. CPI PSU boundaries historically have depended on Metropolitan Standard Area (MSA) definitions, which were not completed until December, 1992. The solution to this timing problem and changes in the classification of PSUs for this Revision appear in Section 2.

When planning began for this Revision of the CPI in 1989, one major change envisioned was the publication of a Consumer Price Index for the entire U. S. population (called the total CPI). In contrast, BLS currently publishes the CPI-U (urban CPI) which covers all residents of Census-defined metropolitan areas, as well as residents of urban parts of non-metropolitan areas, and the CPI-W which covers a subset of this population. To accommodate this expanded CPI, a larger number of PSUs needed to be selected throughout the country to cover the population living in rural areas outside metropolitan statistical areas.

Since an increase in the number of selected PSUs entails an increase in the total cost of the CPI, no decision to publish a total CPI had been made in January, 1993, when a file containing final MSA definitions was received by BLS. At that time noncertainty PSU formation and selection had to begin to meet a newly negotiated February, 1993 noncertainty PSU selection deadline. This necessitated a dual strategy in the formation of non-metropolitan PSUs and in the allocation of non-metropolitan sample PSUs to the four Census regions. Section 3 contains details of this strategy.

A mainframe was used in the 1987 selection, but its cost for this project dictated that as much of the selection as possible be done on a PC. All programs for the 1987 PSU selection except that for controlled selection were adapted for the PC. In addition, the formation of non-metropolitan PSUs described in Section 4 was computerized. Section 5 presents research findings on the choice of stratifying variables and characteristics of the final PSU stratifications. The modified controlled selection program developed for this project in 1991 is discussed in Section 6.

2. Classification of PSUs

Early in the planning for the total CPI a decision was made to change the classification of PSUs. Initially PSUs are classified by one of the four Census regions in which they are located. The second classification variable is size with sometimes a third class variable (indicated in parenthesis in the table below).

PSU TYPE	CURRENT	NEW (TOTAL CPI)
self-representing metropolitan	A=MSAs with population> 1.2 million*	A=MSAs with population> 1.5 million*
nonself-representing metropolitan	L=Medium MSAs M=Small MSAs	B=all nonself-representing MSAs
nonself-representing non-metropolitan	R=(Urban only) T=(Rural only) CE only	C=non-metropolitan

*Anchorage and Honolulu are A PSUs with smaller populations.

Having only one class of nonself-representing MSAs rather than the two current classes eliminates the difficulty encountered in the last revision of defining the population boundary between the small and medium sized MSAs. The current need to have two classes of nonself-representing non-metropolitan PSUs is made necessary because pricing for the CPI-U is done only in urban parts of non-metropolitan areas, while household surveys are done for the CE in both urban and rural parts of the non-metropolitan area. This dichotomy is no longer required for the total CPI, since PSUs would be sampled from the total non-metropolitan area.

To avoid incurring extra costs, PSU definitions were needed by the Census Bureau in April, 1992 so that the CE and CPOPS household samples could be drawn with samples for Census' other household surveys. To partially meet this deadline and limit extra charges, a decision was made to select only the self-representing CPI PSUs by this date. BLS would use Census' best prediction of which counties would be included in final MSA definitions. Census would then choose household samples within the widest possible geographic area which might be included in the self-representing PSUs. Later when the MSA definitions were finalized and the remaining total CPI PSUs were selected, Census would select household samples in the nonself-representing PSUs with the additional costs being incurred by BLS. In addition, Census would drop sample from counties not included in the final MSA-defined self-representing PSUs or not needed because the proportional to size allocation had decreased.

The initial population boundary for being self-representing was determined at this point with only approximate populations for the potential self-representing PSUs as Total Population / number of halfsample equivalents = 248,709,873 / 138 = 1,802,245, where a halfsample equivalent consists of approximately 1100 item pricings and 138 was the number of halfsample equivalents for which the budget for a total CPI was being planned.

Possible SR Population Boundary (in Millions)	Number of Current SR PSUs No Longer SR	Number of New SR PSUs added
1.8	5	1
1.5	2	2
1.2	1	7

The consequences of this boundary, as well as two other boundaries, to the current set of self-representing PSUs (excluding Honolulu and Anchorage) are shown in the table above.

BLS wants as many as possible of the current self-representing PSUs to remain self-representing, because the only selected PSUs for which indexes are individually published are the self-representing PSUs. In order to balance this desire with the mandate to manage field costs by not introducing too many new self-representing PSUs, the boundary was set at 1.5 million with Honolulu and Anchorage remaining self-representing because of their uniqueness. When final MSA definitions were in hand only one additional PSU, instead of the two in the table above, was made self-representing.

3. Design Strategy

Consideration was given to many design strategies to make sure that the final design for the CE and possible total CPI had sample PSUs allocated in as proportional to population size manner as possible, while still being adaptable to a CPI-U. The selected strategy was to allocate nonself-representing metropolitan and non-metropolitan PSUs to the eight Census-region-by-size classes proportional to their total populations. On the basis of preliminary MSA populations, it was also decided to have the same metropolitan PSUs in both designs. However, to prepare for a possible urban only CPI, the selected non-metropolitan PSUs would be divided into two classes, the C PSUs, in which prices would be collected, and the D PSUs, which would be priced only if a total CPI was published. Both C and D PSUs would be surveyed for the CE. Eighteen PSUs would be D PSUs which would represent the rural non-metropolitan population and be allocated to regions on the basis of the regions' rural non-metropolitan population.

After all PSUs were selected, the selected non-metropolitan PSUs with no urban population would become D PSUs. From the remaining selected non-metropolitan PSUs, PSUs would be selected by region to become C PSUs with probability proportional to the urban population of their strata. The remaining non-metropolitan PSUs would be designated D PSUs. In addition, the PSU's percent urban population would be used as a stratifying variable to ensure that the PSUs in each stratum were as alike as possible on this variable.

On the basis of final MSA definitions, one newly-added A PSU no longer qualified by population

size to be self-representing and some minor modifications had to be made in the original halfsample equivalent allocation to the A PSUs. Sixty-four halfsample equivalents were allocated to the A PSUs which contain 46% of the total population and 53% of the CPI-U population. Since each nonself-representing PSU contains only one halfsample equivalent, there were then seventy-four nonself-representing PSUs for a total CPI and fifty-six for an urban only CPI.

For monthly publication of a Census region-by-size class index and for variance calculation, it is necessary to have at least four and an even number of PSUs in a B or C (C and D combined for a total CPI) Census region class, since most items in these PSUs are priced bimonthly. After the proportional to population size PSU sample allocation to region-by-size classes was completed, only the C urban CPI for the Northeast and West could not be published, This is the case for the current index for the R PSUs in these areas. However, for a total CPI, a combined C and D index would be published in every region. Because the Boston PSU has absorbed almost all of the previously non-metropolitan urban population in the Northeast, the Northeast did not qualify to have even one C PSU in a proportional to urban population allocation. A comparison of the present allocation of halfsample equivalents with the new allocation shows an increase in halfsample allocation to the South and West.

4. PSU Formation

Counties were used to form the non-metropolitan PSUs (minor civil divisions were used in Hawaii and all six New England states). In PSUs with urban consumer units, 5000 urban consumer units were necessary per PSU, while 5000 total consumer units were necessary in PSUs without any urban consumer units. All counties in PSUs had to be contiguous, and a reasonable attempt was made not to cross state boundaries. In some areas, there were PSUs in which it was impossible to find contiguous counties with more than 5000 urban consumer units (an example is Lake and Cook counties in Minnesota). These counties were treated as counties with no urban consumer units. ATLAS-GIS mapping software and an overlay of the relevant Census population data were employed in this PSU formation.

5. Stratification

The procedure and implementation of stratification, with a few basic changes, is described in the paper by Dippo and Jacobs (1983). Research done on the selection of variables for the stratification using first 1980 Census data and then, after its receipt at the end of 1991, 1990 Census data shortened the time needed for the final stratification and ultimately for the PSU selection. This research is described in Sections 5.1-5.2. Section 5.3 discusses the actual stratification used in this revision.

5.1 Stratification Variables

Several linear models of PSU-level price change were developed to determine the best variables to use in stratifying the nonself-representing PSUs. Most independent variables in these models were computed from a county data file derived by the Census Bureau from 1990 Census data. Also considered were latitude and longitude variables provided by an ATLAS-GIS database. (ATLAS-GIS is a geographic information system which automatically calculates the latitude and longitude of a geographic area's centroid. This was how each PSU's latitude and longitude were found.).

The table below exhibits percent R^2 values for four competing models of PSU price change of various durations in the A PSUs only (excluding Anchorage and Honolulu, which are outliers due to the fact that they are in a different location and are also demographically different):

Interval of	Model			
Price Change	Geo	Best 4	7 Var	11 Var
6 month	40.23	45.01	34.28	47.69
1 year	28.66	21.28	21.07	28.89
2 year	46.26	39.75	30.22	65.38
3 year	53.01	34.66	24.73	66.31
4 year	63.01	50.86	44.91	79.15
5 year	68.97	62.30	53.37	83.71

The geographic model is a four variable model with independent variables normalized longitude, the square of normalized longitude, normalized latitude, and percent urban consumer units (LONG, LONG2, LAT, and PERURB). LONG2 is used because both the East and West coasts tended to have high price change during the past few years, while the middle part of the country had smaller price change. PERURB is used in case the CPI remains an urban one, in which situation the non-metropolitan strata should contain PSUs with as uniform a percent urban population as possible.

The three other comparison models, which use only Census variables, are the best 4 Census variable model, the 7 variable model created ten years ago, and an 11 variable model which was developed in 1991 using 1980 Census data. Note

that the geographic model does better than all models except the 11 variable model, which was the best Census variable model. Taking into account that the latter model uses 11 variables and the geographic model employs only four, the geographic model was selected as the best.

The table below shows the variables used in the 7 and 11 variable models, along with the percent R^2 obtained when each census variable was regressed against the set of variables in the geographic model, using county data for the 48 contiguous states.

CENSUS VARIABLE	7 VAR MODEL	11 VAR MODEL	% R^2
% fuel oil heated HUs	X	X	81.34
% gas heated HUs		X	70.47
Mean contract rent		X	54.01
% electric heated HUs	X	X	47.20
% two or more wage earner CUs		X	39.82
% black CUs	X	X	39.09
% owner occupied HUs	X		36.59
% white CUs		X	35.44
Mean gas bill/HU with gas bill		X	33.97
% wage and clerical CUs		X	33.13
Mean wage & salary income/CU	X		33.12
% CUs with retired persons	X		10.42
Mean interest & dividend income/CU	X		7.83
% HUs with electric bill		X	7.52
Mean family size		X	7.18

Notice that heating variables tend to be modelled very well by the set of independent variables in the geographic model. Only the last six census variables in the above list have less than one-third of their variation explained by the set of variables in the geographic model.

5.2 Overlap

Finally, expected overlap, which is the expected number of old PSUs in the new design, can be computed once the stratification has been completed. Several stratifications using the variables in these models with various weights on the variables were completed. The following table exhibits the expected overlap found in these stratifications:

REGION	7 VAR=	7 VAR≠	GEO	MIXED
Northeast	3.89	4.70	4.60	4.60
Midwest	3.44	3.78	2.91	2.91
South	10.17	10.30	7.96	10.17
West	2.57	2.66	2.75	2.75
U. S.	20.07	21.44	18.22	20.43
Range	18-22	19-23	15-19	18-22

As shown in the table's second column, the stratification using the seven variables from 10 years ago along with their weights gave the largest and, thus, the most desirable expected overlap. The first column is the overlap expected when using the same variables with equal weights. The third column is the expected overlap when stratifying with the new geographic model variables with equal weights.

The variables in the geographic model were used for stratification purposes in the Northeast, West, and Midwest B PSUs, and also for all of the non-metropolitan PSUs. The seven variables (with equal weights) used ten years ago were employed to stratify the South B PSUs, since too much overlap would have been lost otherwise. This stratification leads to the expected overlap given in the last column.

There are several advantages to using the four geographic variables for stratification. The variables will not change very much over time. This will lead to much better overlap values in the next revision, as the stratifications will be basically the same. In addition, the only variables needed from the Census Bureau will be urban and rural population. In the future, it is not clear that the Census Bureau will continue to collect the variables used in past stratifications. Finally, the stratification with these variables is much easier to understand. It basically groups PSUs together that are close together geographically.

5.3 Stratification Results

The stratification program for the 1987 Revision was written in PL/I for the mainframe. It was converted to C and run on a PC first for the B PSUs in each of the four Census regions and then for the non-metropolitan PSUs in each Census region. It was found to be easier to run this program under OS/2 for the Midwest and South non-metropolitan PSUs due to RAM memory problems under DOS. For each of the eight Census region-by-size classes of PSUs (B and non-metropolitan), twenty stratifications were

completed. In each class, the stratification selected had the smallest sum of between-PSU-within-strata variances over all stratifying variables.

The distribution of the number of PSUs in each final B stratum in each region is fairly uniform with strata containing two PSUs being made up of either two formerly L-sized PSUs or a formerly A-sized PSU and a formerly M-sized PSU. The B strata containing the larger number of PSUs are made up entirely of formerly M-sized PSUs. The expected total overlap among the B PSUs ranges between 19 and 23. The amount of overlap in the new sample among the new nonself-representing PSUs is 21.

6. Keyfitzing and Controlled Selection

The SAS program used in the 1987 PSU selection on the mainframe to Keyfitz PSU probabilities to increase overlap was adapted for the PC and run only on the B strata. The non-metropolitan PSUs, which for this revision contain both urban and rural population, were sufficiently different from the former R and T PSUs to make it difficult to develop criteria for judging which ones were in the current sample. In addition, these PSUs are small in population and respondent burden in them is heavy. Hence, the non-metropolitan PSU probabilities were not Keyfitzed.

The modified controlled selection program was developed for this revision and written in C for the PC. It was run on a file containing data from all nonself-representing PSUs in a region. The first control is that one PSU is to be selected per stratum. The other controls are on the state in which the PSU is located and on the PSU's overlap status (1=in current sample; 0=not in current sample). The input data is put into a file like the following example, which is the first "no solution" three-dimensional problem given by Lin (1992):

PSU	Stratum	State	Overlap	Probability
1	1	1	0	.5
2	1	2	1	.5
3	2	1	1	.5
4	2	2	0	.5

First the program checks that the expected number of PSUs to be selected from each stratum is one. Then it computes expected overlap and the expected number of times each state will occurr in a stratified sample containing one PSU per stratum. From these expected values it calculates the upper and lower control limits for each state and for overlap and the probabilities of these limits. Then the program finds an admissible pattern, i.e., a stratified sample which satisfies the controls, by generating an initial pattern, and then manipulating it until it becomes admissible. The iterative process consists of the following steps:

1. Randomly select one PSU per stratum, without regard to the PSU probabilities. A PSU whose probability has already been exhausted by its appearance in previously generated patterns will not be selected.

2. Impose an admissible value of overlap on the pattern.

3. Impose an admissible value on the representation of each state. If this cannot be done without violating the results of step 2, then keep looping through steps 2 and 3 until acceptable values are found.

4. Assign to the pattern a probability, which is the minimum of the probabilities of the PSUs contained in the pattern and of the control limits satisfied by the pattern.

5. By the amount found in step 4, reduce the probabilities of the selected PSUs and of the control limits satisfied by the pattern .

6. Reduce the remaining probability, which is set equal to 1 for the first iteration, by the amount found in step 4.

7. If the remaining probability is zero, then stop because a solution of the controlled selection problem has been found. Otherwise, generate another pattern by repeating steps 1-6 with the new probabilities found in the previous steps 5-6.

At the beginning of the program's execution, admissible patterns are generated quickly if there are any. The program has two modes of operation. In the interactive mode the program stops when prompted by the user, usually when the program is slow at generating another pattern. In the unattended mode the user can specify a maximum time which should be allowed to elapse between the generation of a pattern and the abandonment of the solution attempt. The program always stores the initial set of patterns and replaces it by any subsequent set of patterns with a smaller remaining probability. This process is called modified controlled selection to contrast it with controlled selection in which a solution is required to have a remaining probability of zero.

When the previous example, for which there are no admissible patterns, was run with a 2 second limit, one saw immediately there was no solution. However, if the overlap status of PSUs 3 and 4 are reversed there is an exact solution to the problem, which this program finds instantaneously; namely, Pattern 1 contains PSUs 1 and 4 and Pattern 2 contains PSUs 2 and 3 with each pattern having probability .5. A third example, which can be gotten from the first example by changing the probabilities of PSU #1 - 4 to .25, .75, .45, and .55, respectively, has two patterns which satisfy the control limits, but a hefty remaining probability of .5. In addition, PSU #1 does not appear in either pattern. This type of solution is not allowed by those running the modified controlled selection program.

The following table contains the results of each region's controlled selection:

REGION	# PSUs SELECTED	# PATTERNS	# PSUs	REMAINING PROB x 10^{-5}	TIME USED
Northeast	12	121	136	1.0	2 mins.
Midwest	18	384	409	2.9	2 mins.
South	34	463	535	307.4	12 hrs.
West	10	150	167	2.0	2 mins.

Each pattern in the set, generated for each region, satisfies the control limits of the problem; however, this set has a nonzero remaining probability. This means that when the region's final PSU sample is selected by choosing one of the patterns from the generated pattern set, the final probability that a particular PSU is in the sample may not be equal to that PSU's input probability. In every region except the South these probabilities agree to at least four decimal places. In the South they agree to at least two decimal places. As can be seen in the table, the remaining probabilities are of comparable sizes in all regions except the South, for which the program was run overnight because of the large number of southern PSUs.

7. Sample Introduction

The selected PSUs will be phased into the CE Survey beginning in November, 1995; however, sampling for CPOPS and pricing for the CPI in these PSUs will be scheduled after the plans and funding of the next revision are approved. It is probable that a CPI-T will not be introduced in the next Revision. However, since the data from the CE survey conducted in these new PSUs will be used to calculate weights for the CPI in the CPI revision after that of 1998, the weights for a CPI-T will be easy to calculate if needed.

8. Acknowledgements

The authors would like to thank everyone in the CPI Survey Research and Analysis Branch of Prices SMD for their help and support on this project, especially Bob Baskin who contributed his computer expertise and insightful analysis leading to the choice of region-type variables for the PSU stratification. Also, our thanks go to John Marcoot for his encouragement and historic perspective and all readers of this paper for their careful reading and for their helpful comments. Finally, we thank Rick Valliant and Stuart Scott for their help in converting this paper's first version to its present form.

9. References

Bureau of Labor Statistics, *BLS Handbook of Methods*, U.S. Government Printing Office, 1992, 176-235.

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.

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