

The redesign of the CPI geographic sample

The selection of new geographic sampling areas ensures that the 1998 revised Consumer Price Index is representative of current demographics

Janet L. Williams

The most basic element of the Consumer Price Index (CPI) decennial revision program is the selection of new CPI samples. The selection of geographic areas is the first stage of the CPI's multistage sample design. In subsequent stages, BLS analysts select the outlets (places where area residents make purchases), goods and services (items purchased), and residents' housing units.

Historically, the Bureau of Labor Statistics has used the Office of Management and Budget's (OMB) definition of Metropolitan Areas first to determine the geographic boundary between the metropolitan and nonmetropolitan areas of the United States for the CPI,¹ and second to divide the metropolitan United States into geographic sections called primary sampling units (hereafter, called sampling units). However, there are five sampling units within the metropolitan area that are not OMB-designated Metropolitan Areas.² In the nonmetropolitan area (a total of 77 percent of U. S. land), BLS forms nonmetropolitan sampling units. In general, a sampling unit is delineated by county borders (with some exceptions in New England), and can comprise several counties.

Currently, BLS publishes the Consumer Price Index for All Urban Consumers (CPI-U) which covers residents of the metropolitan area, as well as residents of urban parts of the nonmetropolitan area.³ Based on the 1990 census, 87 percent of the U.S. population is included in the CPI-U definition. In 1989, when planning began for the 1998 revision of the CPI, one major change envisioned was to publish a Total U.S.

Population Consumer Price Index, the CPI-T. To accommodate this expanded CPI-T, a larger number of sampling units needed to be selected throughout the country to represent the previously unrepresented population.

However, an increase in the number of selected sampling units entails an increase in the total cost of the CPI. When the sampling unit selection process was scheduled to begin in 1993, no decision to publish the CPI-T had been made. To meet the deadline for sampling unit selection, BLS decided to use a dual strategy when forming nonmetropolitan sampling units and determining how many sampling units to select from each of the four census regions.

This article describes the area selection process for the 1998 CPI revision. The basic steps in the geographic area selection process are:

- Determine sample classification variables
- Construct sampling units
- Classify sampling units by population size and allocate sample
- Determine stratification variables within each region's size class and stratify nonself-representing sampling units
- Select sampling units for the CPI geographic sample

These steps are basically the same as those followed for the 1987 revision. This article highlights how the 1998 revision methodology and the final sample design differ from the previous revision.⁴

Janet L. Williams is a branch chief in the Division of Price Statistical Methods, Bureau of Labor Statistics.

Determine sample classification variables

In both the 1987 and 1998 sample designs, sampling units were classified first by location, based on one of the four census regions: Northeast, Midwest, South, and West. In the 1987 design, population size, the second classification variable, had four classes; whereas in the 1998 design it has three.

For the metropolitan area, the population size class variable is used to designate self-representing sampling units (areas which have a large enough population to be selected for the sample with certainty) and nonself-representing sampling units (areas which are randomly selected to represent themselves as well as other metropolitan areas not selected for the sample). Both the 1987 and the 1998 designs have one size class for self-representing metropolitan sampling units (A size class). The 1987 design used two size classes for nonself-representing metropolitan sampling units by drawing a distinction between medium (B size class) and small (C size class) nonself-representing metropolitan sampling units, and the population boundary depended on the census region of the sampling unit.⁵ These two population-size classes were combined in the 1998 design. The decision to have just one population class (designated as B/C) of nonself-representing metropolitan sampling units eliminated the difficulty of defining the population boundary between small and medium metropolitan sampling units, as encountered in the 1987 revision. (See exhibit 1.)

In the 1987 sample design, an additional class variable—urban or rural nonmetropolitan—was required because the geographic areas selected for the CPI-U were also used in the Consumer Expenditure Survey. The definition of “population” in the Consumer Expenditure Survey includes the total nonmetropolitan population—urban and rural—compared with the CPI-U population definition, which includes only the urban parts of the nonmetropolitan area. In the 1987 design, in order to support the expenditure survey’s total population definition and the more restrictive CPI-U definition, the sample design in the expenditure survey required two nonmetropolitan classes—urban and rural nonmetropolitan. The nonmetropolitan area for the 1987 design was first divided into urban and rural areas. Then the urban area was divided into urban sampling units, which were sampled simultaneously for the CPI and the expenditure survey. Subsequently, the rural area was divided into rural sampling units from which the rest of the sampling units for the expenditure survey were chosen. The map in exhibit 2 illustrates the size of the nonmetropolitan land area.⁶ In the 1998 design, this dichotomy was not required, because nonmetropolitan sampling units were sampled from the total nonmetropolitan area, based on the CPI-T population definition. If a decision was made not to publish a CPI-T after the selection of the

nonmetropolitan CPI-T sampling units, urban parts of a subsample of these units would become the nonmetropolitan CPI-U sampling units. However, the selection of CPI-U sampling units and the proportion of the CPI-U population they represent is based on the CPI-T sampling unit selection.

Construct sampling units

For the 1998 revision, the nonmetropolitan sampling units were formed from counties (or from minor civil divisions in Hawaii and in all six New England States). To create a potential sampling unit containing some urban consumer units,⁷ 5,000 urban consumer units were necessary per sampling unit, while 5,000 rural consumer units were needed if the potential sampling unit contained no urban consumer units. This sampling unit population size is required in order to have enough consumer units to support the various household surveys using this design—the Consumer Expenditure Survey, the Continuing Point-of-Purchase Survey, and the CPI Housing Survey—without unduly burdening respondents. All counties in the sampling units had to be contiguous, and a reasonable attempt was made to stay within State boundaries. In some areas, it was impossible to find contiguous counties with either more than 5,000 urban consumer units or more than 5,000 rural consumer units with no urban consumer units. In these cases, BLS eventually formed some sampling units containing some urban consumer units (but not 5,000 of them) and with at least 5,000 total consumer units. For example, the combination of Lake and Cook counties in northeastern Minnesota contains 6,353 consumer units, but only 1,665 urban units. If the CPI-T was abandoned, and the urban part of one of these sampling units was selected for the CPI-U, BLS planned to add urban parts of neighboring sampling units in the same stratum to be used only for the CPI Housing Survey sample.⁸ (Details on stratifying sampling units into classes are discussed later in this article.)

ATLAS-GIS (geographic information system) mapping software, which drew computer maps overlaid with the relevant census population data, was employed in this sampling unit formation. This software also was used to derive the sampling unit location variables—longitude and latitude—employed in sampling unit stratification.

Classify units by population; allocate sample

Census region and population size are used to partition all of the sampling units into a total of 12 classes—the four census regions and three population-size classes within each region. The CPI’s sample allocation consists of determining how many sampling units will be sampled from each of these 12 size classes. The combination of sampling unit classification and sample allocation is an iterative process that is con-

Exhibit 1. Size classifications of sampling units in CPI and Consumer Expenditure Surveys, 1987, 1996, and 1998

Sampling unit	1987 CPI-U and Consumer Expenditure Survey ¹		1996 Consumer Expenditure Survey (CPI-T)		1998 revision CPI-U ²	
	Class	Definition	Class	Definition	Class	Definition
Self-representing metropolitan	A	Metropolitan Areas with 1980 population greater than 1.2 million ³	A	Metropolitan Areas with 1990 population greater than 1.5 million ³	A	Metropolitan Areas with 1990 population greater than 1.5 million ³
Nonself-representing metropolitan	B	Medium Metropolitan Areas ⁴	B/C	Metropolitan Areas with 1990 population of 1.5 million or less	B/C	Metropolitan Areas with 1990 population of 1.5 million or less
	C	Small Metropolitan Areas ⁴				
Nonmetropolitan	D	(Urban only)	Y and Z	Represent total nonmetropolitan population	D	Represent urban nonmetropolitan population
	T	(Rural only) Consumer Expenditure Survey only				

¹ Current class B publication indexes include prices from the class B sampling units and Honolulu, while the current class C publication indexes include prices from the class C sampling units and Anchorage.

² The basic publication index names and composition for the 1998 revision are shown in Appendix 2, (pages 70–77). The West B/C class

index will include all B/C sampling units in the West, along with Honolulu and Anchorage.

³ Anchorage and Honolulu are class A sampling units with smaller populations.

⁴ For the 1987 revision, classes B and C population size boundaries vary by census region.

strained by budget as well as index continuity and publication considerations which are discussed below.

Classifying metropolitan sampling units. After sampling units are formed, BLS determines the population boundary between the size of self-representing and nonself-representing metropolitan sampling units. This process is subject to both budget constraints and CPI users' needs. Sampling units included in the current 1987 design are efficient in terms of program costs and users' needs. Continuing sampling units are less expensive to resample because trained data collection staff are already available in these areas. CPI users want the current class A (self-representing) sampling units to remain as they are because published indexes are available for most of these areas individually.⁹ To balance this desire with the mandate to keep data collection costs under control by limiting the number of new sampling units, BLS classified all sampling units with populations greater than 1.5 million as class A (self-representing) units for the 1998 revision.¹⁰ Honolulu and Anchorage remain class A sampling units because their geographic locations make price

change in these consumer markets unique. The self-representing sampling units form 4 of the 12 regional size classes and include 31 sampling units. All Metropolitan Areas not included in the class A sampling units were classified as class B/C (nonself-representing metropolitan) and all nonmetropolitan sampling units were classified as either class Y or class Z. Exhibit 1 contrasts the 1987 size classifications for sampling units in the CPI and expenditure survey with those in the 1998 revised CPI-U and the 1996 total population Consumer Expenditure Survey. The budget for the 1998 revised CPI required that the sample size remain the same as the current one. This meant that there would be 74 nonself-representing sampling units chosen, with 18 of them not priced for a CPI-U, but only surveyed for consumer expenditure data.

Dual strategy for sample allocation. BLS considered many sample allocation strategies to make sure that the final sample allocation for the Consumer Expenditure Survey and the proposed CPI-T had regional size class samples that were as proportional to population size as possible, while still being

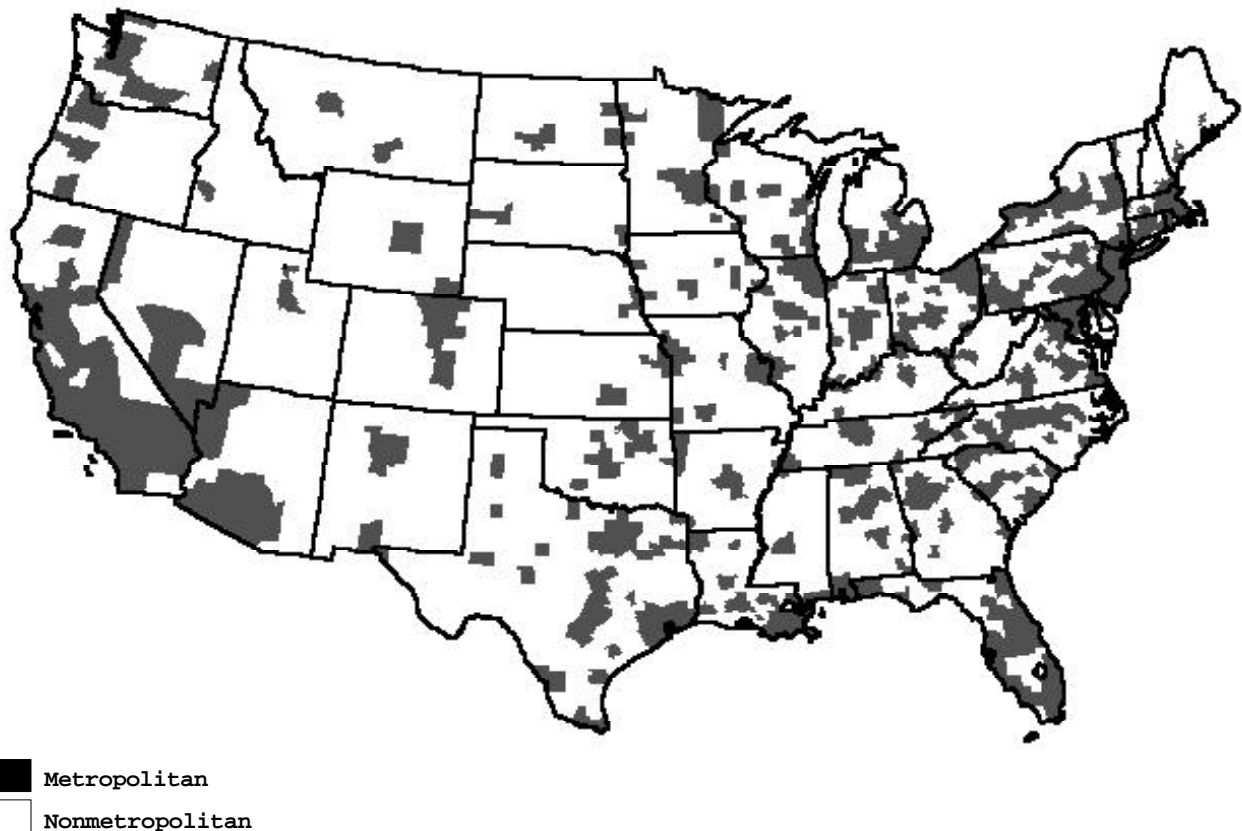
adaptable to a CPI-U. The selected strategy first declared that the CPI-U and expenditure survey would have the same selected class A and class B/C sampling units. The next step was to allocate the number of sample nonself-representing metropolitan and nonmetropolitan sampling units (74) to the remaining eight regional size classes, proportional to their total populations. (For example, the number allocated to the West B/C size class should be approximately equal to the population in the West B/C size sampling units times 74 divided by the population in nonself-representing sampling units.) The CPI-U and the expenditure survey each contain 46 nonself-representing metropolitan sampling units.

To prepare for the possibility of producing an urban-only CPI, BLS adopted the strategy of classifying all nonmetropolitan sampling units into one size class and of selecting 28 nonmetropolitan units. If, after the selection, it was decided that the CPI would use the CPI-U population definition rather than the CPI-T definition, the selected nonmetropolitan sampling units would be divided into two classes, class Y

and class Z. The CPI-U would use urban parts of 10 of the 28 selected nonmetropolitan sampling units to represent the urban nonmetropolitan population; these urban parts would be designated as D sample units in the CPI-U. The 10 sample units of which these 10 are parts are called Y sample units. The expenditure survey would use these Y sample units and the remaining 18 nonmetropolitan sample units (called Z sample units) to represent the total nonmetropolitan population.

The method used to classify the selected sampling units as class Y or Z was iterative. First, the chosen nonmetropolitan sampling units with no urban population would become Z sample units. Then, from the remaining selected nonmetropolitan sampling units, a total of 10 would be chosen to be classified as Y sampling units with probability proportional to the urban population of their strata. This selection was performed in each region, based on the number of nonmetropolitan sampling units allocated to each region. This is illustrated in table 1 in the row labeled D (Y for the

Exhibit 2. Metropolitan and nonmetropolitan areas in the contiguous United States, December 1992



NOTE: Honolulu and Anchorage (Metropolitan Areas) are not shown.

Table 1. Regional distribution of selected sample units, 1998 revision

Size class	Total	Northeast	Midwest	South	West
Total, CPI-U	87	14	22	33	18
A	31	6	8	7	10
B/C	46	8	10	22	6
D (Y for CES)	10	0	4	4	2
Total, CES	105	18	26	41	20
Z (CES only)	18	4	4	8	2

NOTE: CES = Consumer Expenditure Survey.

expenditure survey). Finally, the remaining nonmetropolitan sample units would also be classified as Z units. In addition, the sampling unit's percent urban population would be used as a stratifying variable to ensure that the units in each stratum were as alike as possible on this variable. The number of sample Z units in each region was determined by the region's rural nonmetropolitan population.

With the exception of food and energy items, the CPI collects prices in most sampling units¹¹ every other month; this is known as bimonthly pricing. Bimonthly pricing makes it necessary to pair each selected nonself-representing metropolitan and nonmetropolitan sampling unit priced in odd months with a sampling unit in the same regional size class priced in even months, so that each region's monthly B/C and D size class indexes represent approximately the same size populations. Thus, each region's B/C and D size class must have an even number of sampled units. Index publication requires calculation of index variances. (See "Publication strategy for the 1998 revised Consumer Price Index," pages 26–30.) Variance calculation of a particular region's B/C and D size class index also requires that sampling units in that size class be paired with each other (each pair is called a replicate) and that there are at least two replicates in that nonself-representing size class.¹² Thus, index publication requires that each published nonself-representing regional size-class index area has an even number of sampling units, amounting to at least four.

Table 1 presents the proportional-to-population size sample allocation to the regional size classes for the 1998 geographic area design. The 31 class A sampling units in table 1 represent 46 percent of the total population and 53 percent of the CPI-U population. Also of note is the fact that there are 74 nonself-representing sampling units for a CPI-T and 56 for a CPI-U.

Comparing the sampling unit allocation in table 1 with the publication requirements (mentioned earlier), we see that the nonmetropolitan CPI-U indexes (size class D) for the Northeast and West will not be published when the 1998 area design is used to produce the January 1998 index. (Currently, no Northeast or West nonmetropolitan urban indexes are pub-

lished.) These regional size classes do not meet publication requirements, which require a minimum of four sampling units. However, for a total CPI, a combined Y and Z class (nonmetropolitan) index could have been published in every region. Because the Boston sampling unit has absorbed almost all of the previously nonmetropolitan urban population in the Northeast, that region did not qualify to have even one selected D sampling unit.

Stratify sampling units into classes

The next phase of the sampling unit selection for the CPI-T was to stratify (group) the units in each region's size class (for example, South B/C) into strata (groups) of similar sampling units based on their scores on several stratifying variables. The number of strata is the same as the number of sampling units to be selected because one sampling unit is chosen from each stratum. Each class A sampling unit is in a stratum by itself; thus the name, self-representing. Selection of the stratifying variables to stratify a region's B/C and D size classes was based on linear regression modeling of 1987 through 1992 price change for various time intervals. The independent variables used in this modeling were subsets of 1990 census and geographic sampling unit variables. How well CPI price change was explained by these models was measured by percent R^2 .¹³ Table 2 exhibits percent R^2 values for three competing models of sampling unit price change of various time lags. Data used were from current class A sample units, excluding Anchorage and Honolulu. (Anchorage and Honolulu sample units are statistical outliers because they are geographically removed from the contiguous United States and also demographically different.)

The geographic model consists of four independent variables: normalized (centered and scaled by the range) longitude, the square of normalized longitude, normalized latitude, and percent urban. The two other comparison models, which use census variables, are the 7-variable model which contains the seven variables of the 1987 revision stratification¹⁴ along with percent urban, and an 11-variable model. Note that the R^2 values for the geographic model are larger

Table 2. Percent price change variance explained by models

Interval of price change	Geographical (4-variable) model	7-variable model	11-variable model
6 months	40.23	34.28	47.69
1 year	28.66	21.07	28.89
2 years	46.26	30.22	65.38
3 years	53.01	24.73	66.31
4 years	63.01	44.91	79.15
5 years	68.97	53.37	83.71

Table 3. Percent variance of some census variables in the 11-variable model, explained by the variables in the geographic model

Census variable	Percent variance explained
Percent fuel oil heated housing units	81.34
Percent gas heated housing units	70.47
Mean contract rent	54.01
Percent electric heated housing units	47.20
Percent two or more wage earner consumer units	39.82
Percent black consumer units	39.09

than those for the 7-variable model and smaller than those of the 11-variable model. Taking into account that the latter model uses 11 variables and the geographic model employs just 4, the geographic model was judged best because it was simpler and understandable. The independent variables used in it will be available for future revisions. The reason the 4-variable geographic model performed so well is attributed to the model's high explanatory power for selected variables within the 11-variable model. For example, table 3 contains the 6 of these 11 variables with the largest percent R² obtained when each census variable was modeled by the set of variables in the geographic model. County 1990 census data for the 48 contiguous States were used in this analysis.

Another consideration when choosing stratification variables is the resulting expected overlap (the expected number of old sampling units in the new design). The 1987 geographic sample contained 45 sampling units that were eligible for reselection as part of the new sample of 46 B/C sampling units. Of these, two (Buffalo and New Orleans) were former class A sampling units that were no longer self-representing in the new design. Subject to the requirement of obtaining a statistically representative sample, choosing a stratification that will increase the expected number of reselected sampling units avoids unnecessary training and other personnel costs. Because one sampling unit is selected from each stratum, the expected overlap can be computed once the stratification has been completed. Several stratifications of the metropolitan nonself-representing regional sampling units were completed using the variables in these models with various weights on the variables.¹⁵ Table 4 exhibits the expected numbers of overlap sampling units found in the best of these stratifications using approximate definitions of Metropolitan Areas.

As shown in the third column (7-variable/unequal) of table 4, the stratification using the seven 1987 revision variables along with their 1987 weights and percent urban with a weight of 1 gave the largest and, thus, the most desirable expected overlap.¹⁶ The second column of the table (7-variable/equal) is the overlap expected when using the same variables with equal weights. The fourth column (geographical/equal) is the

expected overlap when stratifying with the geographic variables with equal weights. The last column (mixed/equal) shows the results of a mixed stratification scheme with equal weights.

The last row in table 4 shows the range of the possible number of overlap class B/C sampling units for each set of (weighted) stratifying variables. Note that after stratification, BLS "Keyfitted"¹⁷ each sampling unit's probability of selection from a B/C stratum to improve the possibility that a current sampling unit in the stratum would be reselected, while reflecting shifts in sampling unit populations between censuses. For example, if a 1998 revision stratum contains the same sampling units as a 1987 stratum and a current sampling unit in that stratum has a probability of selection (1990 sampling unit population divided by 1990 stratum population) which is greater than or equal to its 1987 probability of selection, then its Keyfitted probability of being selected from that stratum is 1 and it is selected with certainty.

The final solution was to use the variables in the geographic model for stratification of the B/C sampling units in the Northeast, West, and Midwest, and also for all of the nonmetropolitan sampling units. The seven variables (with equal weights) used for the previous revision along with percent urban were employed to stratify the South B/C sampling units, because too much overlap would have been lost otherwise. This is the mixed stratification and expected overlap in the last column of table 4.

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, a complete change in stratifying variables will eventually have to be made because census 2000 will probably not collect data necessary to construct the 1987 variables, but the geographic variables will definitely be available for the next CPI revision from the ATLAS-GIS software. The program used to do the stratifications is a modified version of the Friedman-Rubin¹⁸ clustering algorithm which puts sampling units in the same strata based

Table 4. Expected overlap using various stratifying variables with equal and unequal weights for class B/C sampling units, by region

Region	7-variable/ equal	7-variable/ unequal	Geographical/ equal	Mixed/ equal
United States	20.07	21.44	18.22	20.43
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. range	18-22	19-23	15-19	18-22

on their similarities on the stratification variables, while keeping the population sizes of the strata approximately equal.

Stratification results. For each of the eight census regional size classes of nonself-representing sampling units (B/C and nonmetropolitan), 20 stratifications were completed. In each class, the final stratification was characterized by possessing the smallest sum of between sampling unit within strata variances over all stratifying variables. This number measures how close the sampling units in each strata are with regard to their values on the stratifying variables.

The distribution of the number of sampling units in each final regional B/C stratum is fairly uniform with strata containing two sampling units being made up of either two formerly B-sized sampling units or a formerly A-sized sampling unit and a formerly C-sized sampling unit. The B/C strata containing the larger number of sampling units are made up entirely of formerly C-sized sampling units. The expected total overlap among the B/C

sampling units ranges between 19 and 23.

Select sampling units

A program was used to select one sampling unit per stratum so that the selected CPI-T sampling units are well distributed over the States and that there are many current sampling units among the newly selected ones. When the decision to publish only a CPI-U was made, the previously outlined strategy was implemented. This resulted in designating selected nonmetropolitan areas as Z sampling units which had urban population in their strata. To account for the Z strata urban population in the CPI-U publication indexes, each selected Z sampling unit containing urban population was paired with a chosen geographically close D sample unit (B/C sample unit in the Northeast) in the same region. The urban stratum population of each Z sampling unit was then added to the stratum population of its paired CPI-U sampling unit to calculate the CPI-U population represented by each D (B/C in the Northeast) sample unit in the pair. These

Exhibit 3. Class size of selected CPI-U primary sampling units in the contiguous United States, 1998



Class size
 ■ A
 ■ B / C
 ■ D

NOTE: Honolulu and Anchorage (A class size primary sampling units) are not shown.

population numbers are used to calculate the percent of index population shown in Appendix 2 (pages 70–77).

Of the 46 final B/C strata, 32 contained at least one sampling unit from the current sample. A current sampling unit was selected in 21 of these 32 strata; that is, the amount of overlap in the new CPI-U nonself-representing sample is 21 sampling units. The map in exhibit 3 shows all counties contained in the contiguous U. S. (Honolulu and Anchorage are not shown) CPI-U sample by size class.

Footnotes

ACKNOWLEDGMENT: The author thanks John Greenlees, Marybeth Tschetter, and members of the CPI Survey Research and Analysis Branch of the Prices Statistical Methods Division who contributed to the final version of this article and appendix 2. In particular, David Swanson coordinated the final editing of the electronic versions of both this article and appendix 2, and William Johnson created the printed and electronic versions of this article's maps.

¹ Each of the decennial census-based Metropolitan Areas is either a Metropolitan Statistical Area, Primary Metropolitan Statistical Area, or Consolidated Metropolitan Statistical Area. For more information, see the Statistical Policy Office of the Office of Management and Budget (OMB) Attachments to OMB Bulletin No. 93–05, *Metropolitan Areas 1992, Lists I–IV*. The CPI metropolitan area includes all OMB-designated Metropolitan Areas.

² The five sampling units in the metropolitan area that are not OMB-designated Metropolitan Areas are the Los Angeles suburbs, CA, sampling unit, the three sampling units that together form the New York-Northern New Jersey-Long Island, NY-NJ-CT-PA publication area, and the Washington, DC-MD-VA-WV sampling unit. (See appendix 2, pages 70–77.)

³ BLS also publishes the CPI-W, which covers urban wage earners and clerical workers.

⁴ A more detailed description of the current and 1998 revision area sample selection is contained in Cathryn S. Dippo and Curtis A. Jacobs, "Area Sample Redesign for the Consumer Price Index," *Proceedings of the Survey Research Methods Section* (American Statistical Association, 1983), pp. 118–23; and J. L. Williams, E. F. Brown, and G. R. Zion, "The Challenge of Redesigning the Consumer Price Index Area Sample," *Proceedings of the Survey Research Methods Section*, vol. 1 (American Statistical Association, 1993), pp. 200–05.

⁵ In 1987, the census region population boundaries between C and B sampling unit population sizes were (in thousands): Northeast–500, Midwest–360, South–450, and West–330.

⁶ This map shows the contiguous U.S. metropolitan area. Anchorage and Honolulu are the only Metropolitan Areas not shown.

⁷ A consumer unit consists of one of the following: (1) all members of a particular housing unit who are related by blood, marriage, adoption, or some other legal arrangement, such as foster children; (2) two or more unrelated persons living together who pool their income to make joint expenditure decisions; or (3) a person living alone or sharing a household with others, or living as a roomer in a private home, lodging house, or in permanent living quarters in a hotel or motel, but who is financially independent and is not included in (2). A student living in university-sponsored housing is included in the sample as a separate consumer unit.

Appendix 2 (pages 70–77) shows the names of sampling units selected for the 1998 revised CPI-U and counties contained therein. The sample contains 36 new sampling units: 1 in class A (Phoenix), 25 in class B/C and 10 in class D. Prices from these 36 sampling units will be introduced into CPI index calculations with the release of the January 1998 index. The appendix also gives the percent of the CPI-U population represented by each selected sampling unit along with its pricing cycle. □

⁸ Four sampling units of this type are in the sample—two in the Midwest and two in the South.

⁹ All current A sampling units are published except those which are part of A101 (New York-Northern New Jersey-Long Island, NY-NJ-CT-PA) and A421 (Los Angeles-Riverside-Orange County, CA). These are published together as A101 and A421, respectively. The Office of Management and Budget calls A101 and A421 Consolidated Metropolitan Statistical Areas.

¹⁰ This decision classified two current A sampling units, Buffalo and New Orleans, as B/C sampling units. In addition, Phoenix, a 1987 class A sampling unit, which was dropped in 1988 due to budget cuts, is a new class A sampling unit. However, a Phoenix index will not be published individually.

¹¹ For the 1998 revision, prices will be collected monthly in just three A areas—A101, A421, and A207 (the New York, Los Angeles, and Chicago Consolidated Metropolitan Statistical Areas).

¹² For information on replicates and how they are used in CPI variance calculation, see Sylvia Leaver and Richard Valliant, "Chapter 28: Statistical Problems in Estimating the U.S. Consumer Price Index," *Business Survey Methods* (New York, John Wiley & Sons, Inc., 1993).

¹³ Values of R^2 always increase as more independent variables are added to a model.

¹⁴ The 1987 stratifying variables were: mean interest and dividend income per consumer unit, mean consumer unit wage and salary income, percent housing units heated by electricity, percent housing units heated by fuel oil, percent owner occupied housing units, percent black consumer units, and percent consumer units with a retired person.

¹⁵ The weights used for the 1987 stratification were 0.5 on each of the non-income variables and 1 on each of the two income variables. A variable's weight is used as a multiplier of a statistic calculated to judge how close every stratum's sampling units are on this particular variable. These products are then summed over all of the stratifying variables. The resulting number is used to judge how good a particular weighted stratification is. The smaller the number, the better the stratification. See Dippo and Jacobs, footnote 4.

¹⁶ See footnote 15.

¹⁷ See Dippo and Jacobs, footnote 4, for more information on this technique.

¹⁸ See D. Kostonich, D. Judkins, R. Singh, and M. Schautz, "Modification of Friedman-Rubin's Clustering Algorithm for Use in Stratified PPS Sampling," *Proceedings of the Survey Research Methods Section* (American Statistical Association, 1981), pp. 285–90.