COMMODITIES AND SERVICES SAMPLE REDESIGN FOR THE 1998 CONSUMER PRICE INDEX REVISION

Sylvia G. Leaver, William H. Johnson, Robert Baskin, Samuel Scarlett, and Robert Morse
United States Bureau of Labor Statistics, 2 Massachusetts Avenue, N.E., Rm. 3655, Washington, D.C. 20212

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

KEY WORDS: Sample design, Optimization, Components of Variance

This paper describes the methods used to allocate data collection resources for the 1998 redesign of the sample for the commodity and services (C&S) component of the U.S. Consumer Price Index. These methods rely on models relating data collection costs and sampling variance of price change to item and outlet selection variables for the sample design. With these models, the optimal allocation of data collection resources to minimize sampling variance of price change, subject to budgetary and operational constraints, can be found using nonlinear programming techniques. This work represents an expansion of models developed for the 1987 CPI sample redesign. Models for sampling variance and costs are given, and solutions to the design problem posed under varying assumptions are discussed. A closing section characterizes the changes in sample allocation from previous designs.

Background

For a full discussion of the Consumer Price Index (CPI), we refer the reader to Chapter 19 of *The BLS Handbook of Methods* (1992). See also Leaver and Valliant (1995) for a more detailed description of the current C&S sample design, selection, and estimation procedures. The CPI is a modified Laspeyres index which is a ratio of the costs of purchasing a set of items of constant quality and quantity in two different time periods. Let $IX(I,t,0)$ denote the index for item aggregate I and month t where month 0 represents the index base or reference period. Then

$$IX(I,t,0) = 100 \cdot \left(\left(CW_t(I) \right) / \left(CW_0(I) \right) \right)$$

$$= 100 \cdot \left(\left(\sum_{i \in I} CW_0(i) \cdot R(i,t,0) \right) \right) / \left(\sum_{i \in I} CW_0(i) \right),$$

where i is summed over all item strata in the item aggregate I, CW_0 represents an estimate of expenditures or cost weight for the base period, and $R(i,t,0)$ denotes the long term relative or estimate of price change from the reference period to time t for stratum i.

In this application, we are concerned with the short term or δ-month price change:

$$PC(I,t,t-\delta) = 100 \cdot \left[\frac{\sum_{i \in I} CW_0(i) \prod_{s=1}^{t} R(i,s,s-1)}{\sum_{i \in I} CW_0(i) \prod_{s=1}^{t-\delta} R(i,s,s-1)} - 1 \right]$$

where $R(i,s,s-1)$ denotes the 1-month relative or estimate of price change from time s-1 to s for stratum i for the geographic area for which the index applies.

An index area is the most basic geographic area for which a price index is published on a monthly, bimonthly, or semiannual basis. There are two types of index areas: self-representing areas, such as New York, which were selected with certainty; and non-self-representing areas, whose sample comprises two or more primary sampling units (PSU's) selected according to a probability sample. The 1998 revised U.S. All Cities CPI will be a weighted average of 34 index area CPI's; 27 for self-representing and 7 for non-self-representing areas. For purposes of variance estimation and operational manageability, the sample for each index area is segmented into two or more subsets called replicate panels.

The commodities and services (C&S) component of the CPI is computed from measurements of price change on a sample of commodities and services, collected from selected outlets in sample cities across the United States. Consumer items are grouped into strata, the most finely defined item classes for which a price index is computed. Each item stratum is composed of one or more narrowly defined classes called entry level items (ELI's). An ELI describes the level of specification for a class of goods with which a data collector enters an outlet for initial pricing.

In CPI sample selection, ELIs are selected from each stratum by a systematic probability proportional to size (pps) procedure, where, in the 1998 revision, each ELI weight will be derived from expenditures reported in the 1993-1995 Consumer Expenditure Surveys. ELI selections are independently drawn for each replicate panel within each PSU.

The majority of the sample frames and weights used in outlet selection are derived from the Current Point of Purchase Survey (CPOPS), conducted by the Census Bureau for the BLS. This survey provides the names and addresses of outlets and dollar amounts of purchases, for item classes known as CPOPS categories. A CPOPS category is a class of items which are normally sold in the same kind of outlet. Each ELI belongs to only one CPOPS category. Outlet frames and selection weights are derived from CPOPS survey data for each PSU-CPOPS category-replicate panel.

In outlet selection, outlets are selected by systematic pps from frames for each PSU-replicate panel for CPOPS categories corresponding to ELIs selected in item sampling. Selected items are then priced in sample outlets on a monthly, bimonthly, or seasonal basis.

History

Hansen, Hurwitz,and Madow (1953), Kish (1965), and Cochran (1977) present several examples of sample design optimization via cost and error modeling. Groves (1990) discusses sample design for social surveys.

Cost and sampling error models were first formulated for the C&S sample design for the 1978 CPI Revision (Westat, 1974). Item classes comprised two categories - food, and other goods and services, and sample size allocation were made for six PSU classes. Selection of the sample design implemented in that revision was based on evaluation of a number of alternative designs. The 1987 CPI Revision redesign (Leaver, et al., 1987) expanded on this approach, refining models for eight item groups and ten PSU classes. This implementation relied on detailed use of administrative records and modeled estimates for cost and variance function estimates. Solution methods used nonlinear programming techniques to identify local minimizers of a modeled relative variance function, under varying assumptions of annual inflation and price change interval. For another BLS survey, Valliant and Gentle (1994) developed a generalized system for constrained optimization of a two-stage stratified sample design implemented on a UNIX platform, with a weighted summed relative variance objective function.

The approach taken in this application generally follows that taken for the 1987 CPIR. Sampling variance, rather than relative variance is minimized in this application. Data collection and processing cost models were revised; costs were derived from administrative records and a time and travel study of CPI data collection. The size of the nonlinear programming problem solved was expanded, and detailed distribution of item-outlet sampling resources used stratum-level variance estimates not previously available for sample design allocation.

The Design Problem

The primary objective of the C&S sample redesign for the 1998 CPI revision was to determine values for all sample design variables which would minimize the sampling variance of price change for the C&S portion of the CPI. The variances for 2, 6, and 12-month price changes were all examined to determine a final allocation. Sample design variables for the C&S component were the number of ELI's to select in each item stratum and the number of outlets to select per CPOPS category-replicate panel in each sample PSU. The number of PSU's , the number of replicate panels per PSU, and the classification of ELI's into item strata were previously determined (Williams et al., 1993; Lane , 1996).

Certain simplifying assumptions were made to render the problem tractable. Newly revised item strata were divided into eleven item groups: food at home, food away from home and alcoholic beverages, household furnishings and operations, fuels and utilities, apparel, transportation less motor fuel, motor fuel, medical care, education and communications, recreation, and other commodities and services. The 87 PSU's were divided into 16 groups according to size and number of replicate panels. It was assumed that the same outlet sample sizes would apply to all PSU's within the same PSU group. It was also assumed that the same item selection sizes would apply across all PSU's . This reduced the allocation problem to one of determining the values of the design variables $\{ K_i , i=1,...,11\}$, the number of ELI selections per replicate panel by item group and $\{ M_{ij} , i=1,...,11, j =1,...,16 \}$, the number of outlet selections per CPOPS category per replicate by item group, which would minimize a modeled price change sampling variance, subject to additional allocation and cost constraints.

The variance of price change for all C&S items was modeled as a function of the design variables, as were total annual data collection and processing costs. Nonlinear programming methods were then used to determine optimal values for the design values under various cost, variance, and sample share constraints. Detailed descriptions of these activities follow.

The Sampling Variance Function

For the purposes of the allocation problem, we write the All U.S. City Average C&S price change estimator as

$$PC(\cdot,\cdot,t,t-\delta) = \sum_i \sum_k RI_{i,k} w_k \, PC(i,k,t,t-\delta),$$

where $PC(i,k,t,t-\delta)$ is the estimated price change from time t-δ to t for item group i and index area k, $RI_{i,k}$ is the relative importance of item group i in index area k, and w_k is the 1990 Census population weight of index area k. Deriving a component form of the variance of this price change estimator, accounting for the stages of sampling described above, would be extremely difficult. Rather than this direct route, we have taken a more indirect, modeling approach described below. Four sources of variation were modeled: PSU selection, item selection, outlet selection, and other sources, such as sampling within the outlet.

The variance function for the CPI revision was modeled for index areas. Each self-representing PSU is a single index area. Non-self-representing PSU's were selected to represent 7 index areas, whose sample consisted of 2 to 22 PSU's. The variance model assumes that the total variance of price change for item group i within index area k can be expressed as a sum of four components:

$$\sigma^2_{i,k} = \sigma^2_{psu,i,k} + \sigma^2_{eli,i,k} + \sigma^2_{outlet,i,k} + \sigma^2_{error,i,k}$$

where

$\sigma^2_{i,k}$	is the total variance of price change for item group i in index area k,
$\sigma^2_{psu,i,k}$	is the component of variance due to sampling PSU's in non-self-representing areas, 0 for self-representing areas,
$\sigma^2_{eli,i,k}$	is the component of variance due to sampling of ELI's within item strata,
$\sigma^2_{outlet,i,k}$	is the component of variance due to sampling of outlets, and
$\sigma^2_{error,i,k}$	is a residual component of variance attributable to other aspects of the sampling process, including the final stage of within-outlet item selection, called disaggregation .

We assume that the variance of price change of an individual sampled unit or quote has the same structure:

$$\sigma^2_{unit,i,k} = \sigma^2_{unit,psu,i,k} + \sigma^2_{unit,eli,i,k} + \sigma^2_{unit,outlet,i,k} + \sigma^2_{unit,error,i,k}, \text{ where}$$

$\sigma^2_{unit,i,k}$	is the total variance of price change of an individual sampled unit or quote for item i in area k,
$\sigma^2_{unit,psu,i,k}$	is the component of unit variance due to sampling PSU's in non-self-representing areas,
$\sigma^2_{unit,eli,i,k}$	is the component of unit variance due to sampling of ELI's within item strata,
$\sigma^2_{unit,outlet,i,k}$	is the component of unit variance due to sampling of outlets, and
$\sigma^2_{unit,error,i,k}$	is the corresponding residual component of unit variance.

It follows that each component of $\sigma^2_{i,k}$ can be written in terms of its corresponding unit variance components:

$$\sigma^2_{i,k} = \sigma^2_{unit,psu,i,k} / N'_k + (\sigma^2_{unit,eli,i,k} / (N_k H_k K_i)) NCi + \sigma^2_{unit,outlet,i,k} / (N_k H_k M'_{i,k} P_i) + \sigma^2_{unit,error,i,k} / (N_k H_k K_i M_{i,k}) \quad \text{where}$$

N_k	is the number of PSU's in index area k,
N'_k	is the number of non-self-representing PSU's in the index area,
H_k	is the number of replicate panels per PSU in the index area,
$M'_{i,k}$	is the number of unique in-scope outlets selected per PSU-replicate
P_i	is the number of CPOPS categories in item group i, and
NC_i	is the percent of strata in item group i which are non-certainty strata.

Thus the sampling variance of price change for the All U.S. City Average C&S index is

$$\sigma^2_{TOTAL} = \sum_i \sum_k RI^2_{i,k} w^2_k \sigma^2_{i,k}.$$

The Cost Function

The total annual cost of the C&S portion of the CPI includes costs of initiation data collection and processing, personal visit and telephone pricing, and pricing data processing, each of which were developed in terms of outlet and quote related costs. For PSU group j and item group i, outlet related costs for initiation are:

$$CI_O(M_{ij},K_j) = 0.2 N_i \cdot H_i \cdot (C_{O,j} + C'_{O,j}) \cdot (a_{ij} M^2_{ij} + b_{ij} M_{ij} + c_{ij}) \cdot P_j$$

where

$CI_O(M_{ij},K_i)$	is the outlet-related initiation cost for item group i in PSU group j

N_j	is the number of PSU's in group j,
H_j	is the number of replicates per PSU in PSU group j,
$C_{O,i}$	is the initiation cost per outlet for item group i,
$C'_{O,i}$	is the initiation processing cost per outlet for item group i,
P_i	is the number of POPS categories in item group i.

and $(a_{ij} M_{ij}^2 + b_{ij} M_{ij} + c_{ij})$ is an overlap function used to predict the number of unique sample outlets, accounting for the overlap of elements in the outlet sample within and between item groups for a replicate panel. The number 0.2 accounts for the rotation or reinitiation of the outlet sample in one-fifth of the sample PSU's each year.

Quote related initiation costs are:

$$CI_Q(M_{ij}, K_i) = 0.2 N_j H_j \cdot WODSI_i \cdot C_{Q,i} \cdot M_{ij} \cdot K_i \cdot NR_i$$

where

$CI_Q(M_{ij}, K_i)$	is the quote-related cost of initiation for item group i in PSU group j,
$WODSI_i$	is a seasonal items initiation factor for item group I,
$C_{Q,i}$	is the initiation cost per quote for item group i, and
NR_i	is the outlet initiation survival rate for item group i.

Note that the expected number of quotes per PSU-replicate panel- item group is estimated by the product of the designated outlet sample size and the number of item stratum selections, $M_{ij} \cdot K_i$.

The costs of ongoing price data collection and processing were also developed as both outlet and quote related costs. For PSU group j and item group i, outlet related costs for ongoing pricing are:

$$CP_O(M_{ij}, K_i) = MB_{ij} \cdot N_j \cdot H_j \cdot NR_i \cdot (a_{ij} M_{ij}^2 + b_{ij} M_{ij} + c_{ij}) \cdot P_i \cdot [(C_{PV,O,i} + C_{PV,T,i}) \cdot (1 - R_{T,O,i}) + C_{T,O,i} \cdot R_{T,O,i} + C_{P,O,i}]$$

where

$CP_O(M_{ij}, K_i)$	is the total outlet-related cost for ongoing pricing,
$C_{PV,O,i}$	is the cost for a personal visit for pricing per outlet for item group i,
$C_{PV,T,i}$	is the travel cost for a personal visit for pricing per outlet for item group i,
$R_{T,O,i}$	is the proportion of outlets priced by telephone for item group i,
$C_{T,O,i}$	is the per outlet cost for telephone collection,
$C_{P,Q}$	is the per outlet cost for processing ongoing pricing data, and
MB_{ij}	is a factor to adjust for the monthly/ bimonthly mix of outlets and quotes by PSU and major product group.

Quote related costs for ongoing pricing are:

$$CP_Q(M_{ij}, K_i) = MB_{ij} \cdot N_j \cdot H_j \cdot M_{ij} \cdot K_i \cdot WODSR_i \cdot [C_{PV,Q,i} \cdot (1 - R_{T,Q,i}) + C_{T,Q,i} \cdot R_{T,Q,i}],$$ where

$CP_O(M_{ij}, K_i)$	is the total outlet-related cost for ongoing pricing for item group i in PSU group j,
$CP_Q(M_{ij}, K_i)$	is the total quote-related cost for ongoing pricing,
$C_{PV,Q,i}$	is the per quote cost for a personal visit for pricing,
$R_{T,Q,i}$	is the proportion of telephone collected quotes for item group i,
$C_{T,Q,i}$	is the per quote cost for telephone collection for item group i, and
$WODSR_i$	is a seasonal items ongoing pricing factor for item group i.

The total cost function associated with data collection and processing for C&S, summed over all item groups and PSU groups, is then given by:

$$C_{Total} = \sum_{i,j} [CI_O(M_{ij}, K_i) + CI_Q(M_{ij}, K_i) + CP_O(M_{ij}, K_i) + CP_Q(M_{ij}, K_i)]$$

Thus, the sample design problem can be expressed as the nonlinear programming problem:

Minimize $\sigma^2_{Total}(\{K_i\}, \{M_{ij}\})$ subject to:

$C_{Total} \leq$ *Total expenditure limit*

$K_i \geq$ *Number of item strata in item group i,*

$K_i \leq$ *Maximum number of item hits for item group i,*

$M_{ij} \geq 2$, $i=1,..., 11, j=1,...,16$

Model coefficients

Estimates of components of the cost function were developed using agency administrative records. Fiscal year 1994 data was used to obtain a total cost per outlet to initiate, and then data provided by the field office produced a per hour cost of initiation. Outlet unit costs and quote unit costs of initiation, by item group, were derived by taking these per outlet and per hour costs and combining them with data obtained from a data collection time and travel study conducted in 1987. Travel costs per quote, by item group, were estimated

by using an overall travel cost per outlet and again comparing it to data from the 1987 time and travel study.

Pricing costs were figured in a similar manner. Distinctions between personal visit and telephone collection of data were made based upon information from the field office and from an analysis conducted within the Prices Statistical Methods Division. Outlet initiation survival rates and quote and outlet retention rates for each item group were developed from field initiation records and ongoing pricing records.

"Overlap" functions were modeled to project the number of unique outlets realized in sample selection as a function of designated sample size. These were obtained by modeling the number of unique outlets obtained in simulations of sampling procedures for each PSU and item group, using CPOPS sampling frames for 1991-1994.

Components of price change variance were computed using 3-way analysis of variance estimation methods and C&S price micro-data collected in 1993-94 (Baskin and Johnson, 1995). Component estimates were developed for 2-, 6-, and 12-month price change for the 11 item groups for 4 Census regions. Total price change sampling variances were also estimated from CPI production index data for the same time period, price change lags, item groups, and 4 Census regions using stratified random groups implemented in VPLX (Fay, 1993). These component estimates were then transformed into unit level components by multiplying them by their degrees of freedom. Average unit total variance and components of variance estimates were then computed by averaging the sums of unit components of variance across regions and months.

Problem Solution

A sequential unconstrained minimization technique, implementing in the nonlinear code Symbolic Factorable SUMT (Ghaemi and McCormick, 1979) was used to find a local minimum to the design problem. Solutions were found using variance data for 2-, 6-, and 12-month price change components of variance estimates. For each item group, the number of item selections was bounded below by the number of strata in the item group and above by a ceiling of 133% of the item group's previous item sample allocation.

Only minor differences were observed between the problem solutions found for differing pricing intervals. Unit variance estimates for apparel and food at home were so remarkably larger than those for other item groups that they dominated the sample share in these solutions. These solutions represented an unacceptable increase in projected price change variance for other item groups. The problem was resolved, additionally constraining the distribution of the food at home and apparel samples across PSU groups to approximately 19% and 14% of total costs , respectively. A separate allocation was also performed for motor fuels for which average prices are published monthly. The remaining sample resources were then allocated among the other eight item groups, using the same modeling methodology.

Item hits were then distributed among item strata within each item group, with consideration given to differences in relative importance, stratum level price change variance estimates, and response rates among the item strata within each item group, as well as special problems identified by commodity analysts and field staff. Similarly, designated outlet sample sizes were adjusted among the various CPOPS categories in item groups to manage variation in expected response rates and respondent burden.

Although major revisions in the CPI occur every 10 years, incremental revisions can occur each year in PSU's where item-outlet samples are rotated. The table below characterizes the revision sample design, contrasting it with the design implemented in the previous four years' sample rotations. In general, the sample design shifted resources in many item groups from sampling many outlets to fewer outlets, with more item selections per outlet. This is due primarily to the large residual component of price change sampling variance estimated for most item groups. This component was regarded as negligible in earlier estimation (Leaver, et al., 1987).

Acknowledgments

The authors would like to thank Janet Williams and David Swanson and Richard Valliant for their careful reading of this paper and their helpful comments. The authors would also like to thank James Branscome, Kirk Hagemeier, Mary Fuxa, and Ken Archer for their insight in CPI sample selection. The authors also would like to thank Janet Williams and Brian Hedges for their support on this project.

References

Baskin, R.M., and Johnson, W. H. (1995), "Estimation of Variance Components for the U.S. Consumer Price Index", *Proceedings of the Survey Research Methods Section, American Statistical Association.*

Cochran, W. G. (1977). *Sampling Techniques,* Wiley, New York.

Fay, R. (1993) VPLX: *Variance Estimates for Complex Samples*, Washington, D.C.: U.S. Bureau of the Census.

Comparison of Modeled Sampling Error and the Distribution of Sample Resources between Current and Revision C&S Design

Item Group	Root Mean 12- Month PC Variance 9401-12	% Change in Modeled PC SE from Current to Revision	Share, Total Costs, Current Design	Share, Total Costs, Revision Design	92-94 CE Relative Importance
Total, All Items less Rent and Owners' Equivalent Rent	.1671	-7.55	100.0	100.0	100.0
Food at home	.2084	+ 7.38	26.8	19.5	13.1
Food away + Alcoholic Beverages	.1295	+11.97	4.2	2.7	9.3
Household Furnishings & Operations	.4029	-12.93	9.2	12.7	10.2
Fuels and Utilities	.3108	+5.17	4.6	4.0	6.8
Apparel & Upkeep	1.6800	-12.56	11.6	14.3	8.1
Transportation less Motor Fuels	.1678	-4.23	9.8	11.1	19.3
Motor Fuels	.2551	+10.78	1.5	0.8	4.3
Medical Care	.2300	+2.33	12.1	10.4	7.7
Education & Communication	.3913	-11.43	10.6	13.0	6.7
Recreation	.4033	-17.10	4.1	6.1	8.4
Other C&S	.4367	+1.21	5.6	5.6	5.9

Ghaemi A. and McCormick, G. P. (1979) "Factorable SUMT: What Is It? How is It Used?" Technical Paper Serial T-402, Washington, D.C., the George Washington University, Institute for Management Science and Engineering.

Groves, Robert (1990) *Cost and Error Modeling in Social Science Surveys*, Wiley, New York.

Hansen, Morris G., Hurwitz, William N., and Madow, William G. (1953) *Sample Survey Methods and Theory*, Wiley, New York.

Kish, Leslie (1965) *Survey Sampling*, Wiley, New York.

Lane, Walter (1996) "Changing the CPI Item Structure," *Monthly Labor Review,* U.S. Bureau of Labor Statistics, to appear.

Leaver, S.,Weber, W., Cohen, M., and Archer, K. (1987) "Item-Outlet Sample Redesign for the 1987 U.S. Consumer Price Index Revision," *Proceedings of the 46th Session, International Statistical Institute.* Tokyo, Vol. 3, pp. 173-185.

Leaver, S. and Valliant, R. (1995) "Chapter 28: Statistical Problems in Estimating the U.S. Consumer Price Index," *Business Survey Methods*, Brenda G. Cox, et al., editors, Wiley, New York

Valliant, R., and Gentle, J. (1994), "An Application Of Mathematical Programming to Sample Allocation," *Proceedings of the Section on Survey Research Methods*, Washington DC: American Statistical Association, 683-688.

Westat, Inc. (1974). "Proposals for and Evaluation of Alternative Designs for Allocation of CPI Pricing Efforts to Items, Outlets, and within Outlets," CPIR-WS-4, Rockville, Maryland.

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