Description of the Revised Commodities and Services Optimal Sample Design

Onimissi M. Sheidu
U.S. Bureau of Labor Statistics, 2 Massachusetts Ave. NE, Room 3655, Washington, DC 20212 Sheidu.onimissi@bls.gov

Abstract

The Bureau of Labor Statistics (BLS) is responsible for designing and compiling the Consumer Price Index (CPI). The CPI is made up of two large components – Commodities and Services (C&S), which is about 72.5%, and Housing, roughly 27.5%. Twice a year BLS conducts sample rotation for the commodities and services component of the CPI in order to introduce new items into the sample and to select a set of outlets that better reflects current expenditures among outlets. For each rotation the sample of items and outlets are allocated in optimal fashion. This item-outlet optimization requires the estimation of variance and cost components and is modeled such that a sample design can be found that minimizes the sampling variance of the estimator. Recently, there have been updates and revisions to the commodities and services optimal sample design procedures. This paper describes and documents the revised commodities and services optimal sample design.

Key Words: Consumer Price Index, Commodities and Services, Sample Rotation, Sample Design, Sample Allocation, Optimization

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

1. Introduction

Every six months BLS conducts sample rotation, which involves selection of items from a frame of highly stratified groups of items that are mapped to outlets selected from an establishment frame, with the main objective of finding the optimal allocation of resources that minimizes the sampling variance of price change, under some budgetary and operational constraints.

This is conducted in two stages: The first is the item-outlet optimization process, which relies on the use of design models consisting of survey operation costs and sampling variance. At the core of this process is the use of the Item-Outlet Programming Suite (IOOPS), a non-linear program created to aid in conducting the item-outlet rotation process. The program uses the survey operation costs and sampling variance components modeled to solve the design problem by applying a non-linear programming technique to output a solution as design variables. These design variables consist of the number of items to price, and the number of outlets to visit for every item group in each sample

area. The second stage is the use of output derived from the IOOPS processing to allocate the item and outlet sample sizes (hits) with specified sampling constraints.

This paper gives a brief overview of the survey cost and sampling variance models of the item-outlet optimization program suite, how the components are updated, and changes and enhancement made to the allocation process in recent years.

2. Overview of Item-Outlet Sample Design

The CPI is a monthly index of all consumer items calculated for the total U.S. urban population. It is a weighted average constructed from 8,018 basic indexes, which correspond to 38 geographic areas and 211 item categories, known as item strata. An index area represents the most basic geographic area at which price indexes are calculated. Out of the 38 index areas, 31 are selected as certainty units, known as self representing index areas, and 7 are probability units, selected from non-self representing areas. In each item stratum there is one or more narrowly defined categories of goods and services called entry level items (ELIs).

The CPI is made up of two large components – Commodities and Services (C & S), which is about 72.5%, and Housing, roughly 27.5%. BLS conducts sample rotation twice a year for the commodities and services components of CPI, in which about one-eighth of the item and outlet sample in each Primary Sampling Unit (PSU) is reselected to reflect changes in the market place. This constitutes about 25 percent of the item strata per PSU rotated in a given year. Every year, a sampling frame of the item universe is created by combining all four regional item universes for the two most recent years of Consumer Expenditure (CE) survey data. During the rotational periods of the year, using the sampling frame, ELI selections are independently drawn from each stratum for each sample replicate within each index area. This ELIs selection from each stratum is done by a systematic probability proportional to size (PPS) sampling procedure, where each ELI weight is derived from the expenditures reported in the CE.

The Telephone Point-of-Purchase Survey (TPOPS) conducted by Census for BLS serves as the source of outlet sample frames and weights. It provides outlet details and dollar amounts spent on the purchases of groups of items known as Point of Purchase Survey (POPS) categories. A POPS category is a group of commodities and services that are usually sold in the same outlet. A POPS category could contain one or multiple ELIs mapped to it.

Outlet frames, total daily expenditure estimates, and selection probabilities are derived from TPOPS data for each PSU-POPS category-sample replicate. For the purpose of sampling variance estimation, the sample for each self-representing PSU is divided into two or more independent subsets known as replicates. Each replicate-PSU contains a single subset of independently selected ELIs and outlets for all item strata within the PSU.

Like the ELI selection procedure, outlets are selected via a systematic PPS sample from an outlet frame for each PSU-sample replicate for POPS categories corresponding to selected ELIs in item sampling. The selected ELIs are then priced in sample outlets either on monthly, bimonthly, or seasonal basis. There are some items not included in the TPOPS known as non-POPS items. For these items, separate sample designs are constructed.

2.1 Item and Outlet Structures of C&S

The primary objectives of the C&S sample design is to determine the optimal values for all design variables that minimize the sampling variance of price change for the C&S components of CPI under budgetary and operational constraints. That is, to decide how best to allocate item and outlet samples for every item stratum in each replicate - PSU, and POPS category that minimizes the sampling variance of price change measured at the All-US level. To achieve this objective, survey operation costs and sampling variances are modeled to project the decision variables for the sample design. To this effect and because of the complexity of the design, all item strata of C&S are grouped into 13 major groups (MG), and the 38 index areas are classified into 15 PSU groups (PG), see table 2.1, and 2.2 below, for a detailed list MG and PG.

Table 2.1: List of the C&S 13 Item major groups

	Allocation Major Group
1	Food at home – non-meat staples
2	Food at home – meat, poultry, fish
3	Food at home – fruits and vegetables
4	Other food at home + Beverages
5	Food Away from home
6	Household furnishings and operation
7	Utilities
8	Apparel (Less Women Apparel)
9	Transportation less gas and motor oils
10	Women Apparel
11	Medical Care
12	Education & Communication
13	Entertainment-Other C&S

2.2 The Design Problem and Setup

The main task is to determine for all the sample design variables the appropriate values that minimize the variance of price change for the C&S component of the CPI. This problem is solved by implementing the following steps:

- Modeling the variance function that predicts the variance of price change as a function of the sample design variables - comprising the number of ELIs per stratum and the number of outlets per POPS category for every sample replicate-PSU to select;
- Modeling a cost function that predicts the total annual survey operation cost;
- Computing estimate values for all coefficients of the two functions including estimates of outlet sample overlap;
- Determining the optimal item and outlet sample sizes.

This means we have to structure the design problem in a more manageable way to attain a possible solution through the use of non-linear programming (NLP) models. Therefore, we segment item strata into the 13 item major groups, and reclassify the PSUs into 15 PSU groups according to the size and number of replicate panels. This makes the design problem tractable to solving with the Item-Outlet Optimization Program Suite, (see table

2.1 and 2.2). The same item and outlet sample sizes are maintained in all PSUs within the same PSU group.

As such, the first step of IOOPS is to solve a non-linear problem, minimizing the modeled 6-month percentage change standard error for all U.S. Cities in C&S subject to the following constraints:

- The total Cost must not be more than the Budget Ceiling level;
- The number of item selections per item major group for every PSU Group must at least be equal to the number of item strata in that PSU per major group, but not more thane number of the required item selection ceiling;
- The number of outlets selections must be at least 2 per POPS category for every replicate PSU, except for the two apparel groups, item major groups 8 and 10 that must be at least 3;
- And the average quotes for every stratum must be greater than or equals to 9.

This is expressed as,

$$\underset{\{M_{ij},K_{ij}\}}{Minimize} \quad \sigma^2_{total} = \sum_j \sum_k RI^2_{ik}\ \sigma^2_{i,k}$$, subject to the following conditions:

Total Cost ≤ Budget ceiling,
$K_{ij} \geq$ Number of item strata in item group *i*, PSU group *j*,
$K_{ij} \leq$ item selection ceiling for item group *i*, PSU group *j*,

$M_{ij} \geq 2$, *i=1,..., 13, j=1,...,15,*
Average number of quotes per stratum ≥ 9.
RI_{ik} = relative importance of item group *i*, index area *k*, scaled to sum to 1.0 over all items and areas.

In other words, we try to look for the values of both the number of ELI selections per item group-replicate panel within each PSU group, denoted by
{ K_{ij} , *i*=1,...,13, *j* =1,...,15}, and the number of outlet selections per item group-POPS category-replicate within each PSU group, denoted by {M_{ij} , *i*=1,...,13, *j* =1,...,15} that minimize the modeled price change sampling variance.

The variance model assumes that the total variance of price change for item group i within index area k is dependent on four sources of variation:

1) PSUs in non-self-representing areas;
2) ELIs within item strata;
3) Outlets, and
4) Residual errors such as occur in item disaggregation.

That is, the U.S. All Cities - C&S level variance:

$$\sigma^2_{total} = \sum_i \sum_k RI^2_{ik}\ \sigma^2_{ik}$$

$$\sigma^2_{j,k} = \sigma^2_{item,\,j,k} + \sigma^2_{outlet,\,j,k} + \sigma^2_{psu,\,j,k} + \sigma^2_{error,\,j,k}.$$

The total annual cost is a function of all the survey operation costs. This is broadly defined as the sum of the initiation costs and repricing costs, and each of these components is a function of the number of unique outlets M'_{ij} and the number of quote $M_{ij} \cdot K_{ij}$.

Therefore, the:

Initiation costs = Initiation Outlet Costs + Initiation Quote Costs;
Repricing Costs = Repricing Outlet Costs + Repricing Quote Costs;

Initiation-related costs are distinguished by response rates, within-outlet times, while repricing outlet and quote-related costs are distinguished by mode of collection (by personal visit, telephone, or web).

The total cost function associated with data collection and processing consists of the total sum of all the outlet-related initiation costs, plus all the quote-related costs of initiation, plus the total outlet-related cost for ongoing pricing, plus the total quote-related cost for ongoing pricing for every item group *i* in PSU group *j*.

Table 2.2: List of C&S 15 PSU Groups with their replicate panels

PSU Group		Number of PSUs	Number of replicate panels
1	New York City	1	4
2	New York City suburbs	2	4
3	Los Angeles City	1	4
4	Los Angeles suburbs	1	2
5	Chicago	1	4
6	Philadelphia and San Francisco	2	4
7	Detroit and Boston	2	4
8	Other large self-representing PSUs	7	14
9	Small self-representing PSUs	12	24
10	Medium-sized PSUs, Census Region 1	8	8
11	Medium-sized PSUs, Census Region 2	10	10
12	Medium-sized PSUs, Census Region 3	22	22
13	Medium-sized PSUs, Census Region 4	6	6
14	Small PSUs, Census Regions 1-4	10	10
15	Anchorage and Honolulu	2	4

3. C&S Sample Rotation Procedure

The sample rotation procedure is conducted in three distinct steps:

1) Processing of item and outlet sample frames with updated components.
2) Item and outlet sample selection and optimization using IOOPS, which yields a Non-linear program (NLP) output file.
3) And the third stage is using the NLP output file to conduct the item and outlet sample size allocation.

Sample rotation is conducted semiannually, and updates are necessary to refresh data for many of the components' variables that are used in the IOOPS. All updates are done based on 24 month worth of data with 6 months of new data being rotated in to replace the oldest 6 months of data. We conduct updates both before and after IOOPS processing and most of the updates are semiannually while the components of variance are updated every three years.

Hence semiannually pre-NLP updates are:
Linear and quadratic coefficients of unique outlet prediction function that are used for both cost and sampling variance components, quote and outlet response rates, quotes and outlet telephone repricing, web repricing, and major group – index area relative importance.
Components of variance are updated triennially.

The post NLP updates conducted semiannually are item strata relative importance, item major group variances, and item stratum variance weighted relative importance. See table 3.1.

Table 3.1: IOOPS Component updates

Pre-NLP		Post NLP
Semiannually	Every 3 years	Semiannually
Coefficients A & B of unique outlet predictor function for sampling frame Non-response rates for both Outlet and quote Telephone repricing rates for both Outlet and quote Web repricing rates for both Outlet and quote Major Groups – Area relative importance	Components of variance (Last updated using 2008-2011)	Within major groups item stratum relative importance Within major groups item stratum variances Item stratum within item group variance weighted relative importances

The above table shows schedule of updates that take place for the item-outlet optimization process during every rotation period.

4. Item and Outlet Sample Size Allocation Procedure

The item-outlet optimization process is roughly structured into two sequential steps: The item-outlet sample selection and processing stage, and the item-outlet sample size allocation.

During the first step, the nonlinear programming suite is used to conduct item-outlet sample selections and sampling data processing. This yields an output file known as NLP file. The NLP file is made up of unrounded numbers of item and outlet selections, which are defined at the item major group – PSU group level. These numbers are converted to an integer number of item stratum selections for each item stratum – PSU combination, and number of outlet selections per TPOPS category for every PSU/ half sample.
These item stratum selections are then used in the calculation of item-outlet sample sizes.

The method for allocating the final sample sizes for items is slightly different than the method for outlets.

4.1 Calculated Statistics for Item Stratum

For the allocation of the item stratum size, three main statistics are calculated, which are as follows:

1). Item stratum variance weighted relative importance, also known as, NewRI

The Item stratum variance weighted relative importance, is calculated based on 24 months of monthly data on item stratum level standard errors, relative importance and number of quotes. This method helps to reflect the monthly variability in the number of quotes.

It is calculated as the *24 months average of the product of 12-month SE, RI, and square root of* quotes, *all divided by the sum of such over item strata in all IOOPS major groups*100*. This is expressed as

$$NewRI_{IS} = 100 * \frac{\frac{1}{24} * \sum_{t=t0}^{t=t0+23} SE12_{t,IS} * RI_{t,IS} * \sqrt{quotes_{t,IS}}}{\sum_{IS' \in IOOPMG} \frac{1}{24} * \sum_{t=t0}^{t=t0+23} SE12_{t,IS'} * RI_{t,IS'} * \sqrt{quotes_{t,IS'}}}$$

Where,
SE12 is the item stratum (*IS*) 12-month standard error;
RI is the item stratum (*IS*) Relative importance calculated for every item-stratum in each area by IOOPS major group category, using the item stratum cost-weight;
IOOPMG is the IOOPS item stratum major group.

2). Initial sample size, which we called Base Hits

The Initial sample size or Base Hits is an integer value of the above calculated variance weighted relative importance multiplied by the number of item selections, $K_{ij,}$ derived from NLP output file. This is more or less a proportional allocation for the number of item selections among the item strata, depending on the average of the products of the relative importance (RI) and the standard error, with weighted standard error inflated by the quote count of the month. Thus, the base number of selections for the item stratum is equal to:

$$Base\ \ Hits_{is} = Integer(Number\ of\ item\ stratum\ selections, K_{ij} * NewRI_{is}).$$

3). Remainder

The Remainder, *R* is the unassigned number of item selections calculated as number of item selections minus the sum of Base Hits across all Item Strata in the IOOPS major

groups. In another words, this is the remaining number of hits after the base number of hits is determined for every item stratum. Hence, it is calculated as

$$\text{Re}\,mainder_{IOOPSMG} =$$

$$Number\ of\ item\ stratum\ selections, K_{ij} * \sum_{IS \in IOOPS} BaseHits_{is}\ ,$$

Where,

$\sum_{IS \in IOOPS} BaseHits_{is}$ - is the sum of the item hits of all item strata within IOOPS Major Group.

4.2 Calculated Statistics for Outlet

The process of allocating outlets is straightforward. Two main statistics are calculated, the Base Hits, and the Remainder.

The NLP program gives values of the number of outlet selections, M_{ij} per replicate - PSU per /POPS category. For each IOOPS Major Group, the number of POPS categories is multiplied by the number of outlet selections. Each POPS category is assigned a base number of hits,

Base Hits=Floor $\left(\frac{\text{Floor(Number of pops categories* M)}}{\text{Number of pops categories}}\right)$, M= M_{ij} .

Hence, the remainder of the number of outlet hits to be allocated is calculated as:

$R =$
$Floor(Number\ of\ pops\ categories * M) - BaseHits *$
$Number\ of\ pops\ categories.$

4.3 Item-Outlet Sample Size Allocation

Sample size allocation process for item and outlet sample is more or less the same except in the assignment of the remainder. The determination of which item strata get a remainder of 1 or 0 is based on their variance weighted relative importance ranked in descending order, while in outlets, the remainder of 1 or 0 is a random assignment through their POPS categories. In general,

Sample Size = Base Hits + Remainder [1 or 0], where additional constraints may apply.

The process is summarized as follows:

For Item Strata:

- Initially all item strata within an item group are allocated at least one selection;
- The number of selections for IOOPS Apparel groups (groups 8 and 10) are limited to 9, and all other IOOPS major groups are limited to maximum of 6;
- Item strata with special constraints such as HB02, TB01, and some others are selected accordingly, by making use of the manual adjustment lists, see table 4.1.
- With the remainder of the item stratum hits, the item strata are sorted by their weighted variance relative importances, NewRI, and the top item strata are given one extra selection until the remainder is zero.

Thus, the number of item stratum selections is Base Hits or Base Hits plus one.

While for outlets, the POPS categories are assigned random number and then sorted based on this number. The first *POPS* categories (the categories ranked 1) receive Base

Hits plus one hit, and the remaining POPS categories get Base Hits plus zero as the number of outlet selections.

Some items and outlets with fixed sample sizes are classified as a group with manual adjustment, and are held steady for every rotation. In effect, Items with fixed allocations are removed from the NLP model used in IOOPS. See table 4.1, showing an example of a list of some of the items with fixed allocations

Table 4.1: Example of Item Strata for manual adjustment with fixed item hits

Item Stratum	POPS	Title	Frame Source	Item Hit (Maximum)
HB02	H02	Lodging away from home	TPOPS	3
TA01	T01	New vehicles	TPOPS	3
TB01	T05	Regular unleaded gasoline	TPOPS	3
TB02	T05	Automotive diesel fuel	TPOPS	2
TF01	T92	Vehicle registration, licenses	Non-POPS - by PSU	2
TG01	T93	Air fares	Non-POPS - by PSU	1
TG02	T94	Buses	Non-POPS - by PSU	1
TG02	T95	Trains	Non-POPS - by PSU	1

5. Changes and Enhancements to IOOPS

In recent years C&S sample design has undergone several changes and enhancements. These are:

1) Alternative item and area groupings have been put in place. Item major groups have been reorganized, splitting the apparel group (major group 8) into two groups - major group 8 (Men and children apparel), and reclassifying women apparel into a separate item major group of its own, under major group 10. This is in place of Gasoline and Motor Fuels, which has been moved to the manual adjustment POPS category. Item and outlet sample sizes such as this are held steady and manually adjusted as deemed necessary.
2) Web data collection costs are now accounted for separately. Before this change, it was accounted for as part of the telephone repricing cost. Therefore, additional parameters were created in NLP model in IOOPS to account for this change.
3) We have rewritten some program in SAS Procedure known as “Proc Optmodel.”
4) Explicit procedure is put in place for converting NLP output into integer sample sizes, IOOPS Format module was rewritten to provide final unrounded optimizations.
5) Some item and outlet sample sizes are now held fixed and therefore grouped into a separate category as a manual adjustment group.
6) Frequent updates of design components and parameters are now put to practice as the norm to take care of seasonal or monthly variability.

5.1 Changes Underway

There is continuous effort being made to improve and enhance C&S sample design.
We are currently studying the impact of upcoming changes such as the New Estimation system and the New Area Sample scheduled for 2018

6. Reference

Bureau of Labor Statistics (2003), BLS Handbook of methods, Washington, D.C http://www.bls.gov/opub/hom/home.htm

Johnson, William H., Leaver, S.G., and Benson, T.S. (1999) "Modeling the Realized Outlet Sample for the Commodities and Services Component of the
U.S. Consumer Price Index," *Proceedings of the Section on Government Statistics, American Statistical Association,* pp 304-308.

Lane, Walter (1996) "Changing the CPI Item Structure," U.S. Bureau of Labor Statistics http://www.stats.bls.gov/mlr/cpiwl001.htm.

Leaver, Sylvia G., Solk, D.T (August, 2005), Handling Program Constraints in the Sample Design for the Commodities and Services Component of the U.S. Consumer Price Index.

Kish, Leslie (1995), Survey Sampling, Wiley, New York.