The New Design for the EIA-878 Gasoline Price Survey

Pedro J. Saavedra, Paula Weir, Benita O'Colmain, Tracy Churchill and Ewa Carlton
Pedro J. Saavedra, ORC Macro, 11785 Beltsville Dr., Calverton, MD 20705

Keywords: area sample, weighting, gasoline, allocations

The EIA-878 is a survey of motor gasoline outlet prices that produces estimates of national and regional level prices, as well as separate estimates for several states and cities, two formulations and three grades of gasoline. Up to recently, this survey has used a monthly survey of resellers and refiners as phase I of a multi-phase sample, subsampling the sample units of the monthly survey that report the specific outlet sales category. A new design extends coverage to independent stations, and targets additional states and cities. The design is an area sample that uses data from several sources, allocating stations to counties and sampling stations from the selected counties. Weights make use of the number of places at which gas may be pumped as a proxy for volume and the proportion of gas by grade in the state. Several data sources are used for analytic purposes and to obtain allocations and size measures.

1. Introduction

The August 1990 Iraqi invasion of Kuwait and the resulting rise in gasoline prices led to a need for monitoring more frequently motor gasoline prices. The requirement was to have available an unleaded, regular gasoline pump price at the national level that was not only accurate, but could be obtained quickly and inexpensively. The survey (which was to become the EIA-878) needed to be up and going within a week. The Lundberg survey was considered inadequate since it only collected price every other Friday. The American Automobile Association also runs a weekly price survey, conducted by the Computer Petroleum Corporation (CPC). This survey sampled 1250 service stations out of the Yellow Pages frame of 5,000 stations with a one-fourth rotation each week. It was designed to represent main travel corridors and vacation/resort areas rather than to produce a representative estimate of all outlets.

The EIA-878 Motor Gasoline Price Survey was initially a survey of retail motor gasoline outlet prices drawn from the sample of EIA-782 respondents and was meant to monitor consumer prices during the Persian Gulf War in 1990/91 (Saavedra and Weir, 1991). A principal objective was collecting, processing, and releasing the data to a variety of users in a very rapid turn around mode. A two-phase sample was used, with the EIA-782 (Saavedra, 1988) as Phase I of the sample. Gasoline resellers and refiners who sold through outlets were chosen with probabilities proportional to weighted volume in each state and selected for the sample. From each reseller/refiner selected for the sample in a given state, one or more gasoline stations were sampled. The design permitted the use of a simple average as the main price estimator. The only estimate was a national estimate of the price of regular gasoline at the pump.

The survey was expanded to be responsive to the Clean Air Act and eventually estimates for Conventional, Oxygenated, Reformulated and OPRG gasoline for all Petroleum Allocation Defense Districts (PADDs) and sub-PADDs for the State of California were added, as well as estimates for midgrade and premium gasoline. This required an increase in sample size and presented the difficulty that the Phase I sample from the EIA-782 was not sufficiently large. As a result two cycles of the EIA-782 were combined in order to form Phase I for the survey (Weir and Saavedra, 1998). When estimates for additional states and cities were required, the two-phase design was not sufficient for this survey.

More recently, EIA has decided to conduct a two-part expansion of the EIA-878 Motor Gasoline Price Survey. As part of the first expansion, the sample was recently augmented to allow release of average prices for 5 states and 6 cities, in addition to the regional and U.S. average prices previously released. The completed first expansion included:

- Discontinuing the publication of oxygenated and OPRG gasoline prices with customer notification
- Publishing prices for six cities and five States (one in each PADD); these cities are New York, Chicago, Houston, Denver, and San Francisco and Los Angeles; the States are New York, Minnesota, Texas, Colorado, and California
- Examining the retail gasoline pricing behavior in the midwest/PADD 2
- Modifying the software that receives the outlet and aggregate data files from the collection

system and produces reports for the web and the WPSR to accommodate the changes to formulation groups and geographic areas

- Redesigning the web site for the dissemination of the gasoline prices
- Creating the historical data base back to 1990 with backcast prices back to 1995. Backcast prices were used in the historic database for any new series' price. Previously published prices, however, were not replaced by backcast prices

The EIA-878 is being redesigned for a second expansion. Several cities and states were added to the required estimates. It was decided that a two-Phase design was no longer viable for this survey. One reason is that the EIA-782 is no longer sufficiently large, and another is that the two-Phase design restricted the sample to gasoline stations that were owned by resellers or refiners, and did not include independents.

The Energy Information Administration (EIA) of the U.S. Department of Energy through its contractors developed a gasoline outlet frame and designed a new version of the EIA-878 gasoline sample. In preparation for drawing a sample of gasoline outlets for the EIA-878 motor gasoline survey expansion, the coverage, available information, and data file format of 10 potential sources from which to obtain a list of gasoline outlets were reviewed. Eventually several sources of data were identified as possible frames or auxiliary data bases.

2. Basic Design

The design is driven by the definitions of Publication Cells and Sampling Cells. A Publication Cell is one defined by a PADD, state, city and attainment status area (the latter being restricted to reformulated and conventional gasoline). Hence, New York State reformulated gasoline is a publication cell. So are New York City, conventional gasoline in PADD 1A (New England) and all of the United States. Sampling cells are the smallest units whose borders are defined by publication cells. Thus, the part of New York State where reformulated gasoline is required, but is not in New York City would be a sampling cell.

One slight deviation from this concept pertains to counties where reformulated gasoline is required by part of the county. Conceptually, the county may be split between sampling cells. In practice it may be difficult to establish the location of individual stations with respect to attainment boundaries before the sample is drawn. Allocation will thus be made to the county, with the possibility of reclassifying some stations after a more thorough determination.

The steps for drawing the sample may be described as follows:

1) Using the current survey, and possibly some auxiliary data determine the sample size needed to obtain the desired estimates (price CVS for all three grades of gasoline and for totals) for each publication cell.
2) Convert Step 1 into an allocation for each sampling cell.
3) Transform the allocations into allocations (possibly fractional) for each county.
4) Select gasoline stations as allocated per county.

3. Data Sources

The data sources ORC Macro will use for sampling and weighting are as follows:

1) A primary data source which provides a flat file with a list of stations that covers the nation and is representative within each county. In other words, it is less important that the same percentage of stations in each county be included in the list than that the stations within each county not be biased toward the urban or the rural areas of the county. The recommended source of such a file is OPIS.

2) The counts of the number of gasoline stations per county obtained from the Census Bureau's County Business Patterns database.

3) A supplementary data source where stations from individual counties can be looked up if necessary (e.g. Switchboard or other online yellow pages directory).

4) A file identifying the sampling cells by county.

5) A source from which to estimate total volume of gasoline sales per sampling cell. Possible sources or estimators include:
    a) Total number of stations from the County Business Patterns database.
    b) Total volume from EIA-782 combined with the County Business Patterns database.
    c) Census data, possibly using the number of households with cars.

6) A method of assigning a percentage of sales by grade to each station. Here the EIA-782AB or the EIA-782C percentages for the state where the station is located seems appropriate.

7) An analytic file that can help determine the relationship between price, volume and number of fueling positions at gasoline outlets. A total of 1,000 stations from 20 markets has already been obtained from NIM.

8) A source of prices suitable for estimating unit variances for various sampling cells. The current EIA-878 would be most appropriate. The file NIM can serve as an auxiliary file.

New Image Marketing (NIM) is the only source that can provide volumetric data. However, the cost of obtaining the entire NIM database is high and coverage is limited to urban areas. Of all the databases we reviewed, NIM is the only database with a known bias that affects coverage. On the other hand, Census data can be relied upon to provide an accurate estimate of the number of gasoline stations that exist across the country. County level Census data will, therefore, be useful in drawing a sample.

4. Steps in Design, Allocation and Draw

A comparison of the OPIS data set and the CBP indicates a very close level of agreement in terms of the number of stations per county, with the exception of the state of California, where the CBP indicates a larger number of stations. One way of handling this issue is to assume the maximum of the two numbers as the number of stations in a county. This means that if a county has 20 stations listed in OPIS, 24 in the CBP and 1 station is allocated to the county, we will assume that the one station is sampled from 24, even though only 20 are available in the frame.

We estimated the unit variance in each sampling cell using the current EIA-878. Then we used the Chromy Allocation Algorithm (Chromy, 1987, Zayatz and Sigman, 1995) to calculate allocations for each sampling cell. In one instance we subdivided a sampling cell to take into account the heterogeneous character of the cell (which included Alaska and Hawaii) Initially the target was set to a coefficient of variation of one percent, but this proved to be unnecessarily large. In addition, there was some concern regarding underestimation of some of the variances. As a result, the target was set for .4 for the United States, .55 for PADDs and U.S. formulations, .70 for sub-PADDS and the PADD formulations, .85 for cities and states, and 1.0 for the remaining cells (e.g. state and sub-PADD formulations). A minimum of five stations was assigned to each cell. Having assigned allocations to the sampling cells, the integer part of the allocation of each county will be assigned to the county, then counties will be selected with probabilities proportional to the fractional part of the allocation. Thus if a county has an allocation of 2.3 stations, the sampling procedure will assign it at least two stations, and the third with a probability of .3. A Goodman-Kish PPS sampling method was used, ordering counties within states by number of stations.

Once integer allocations were drawn for each county, the proper number of stations were randomly selected for the OPIS file within each county. It is, of course, possible that a county be selected in which the OPIS file does not list any stations, but the where the CPB does. In such a case the first step was to be to use a yellow pages search engine to try to locate a station in that county. If that failed, then a new station from the same state and sampling cell would be drawn randomly. As it happened, this was not required for the initial sample.

It is, naturally, possible that some stations will have ceased to exist and some will refuse participation. We will assume for the time being a constant number of stations per county. Refusals and out-of-scopes will be replaced by another station in the county. After two replacements fail to yield a responding station, the subsequent ones will be drawn from the entire sampling cell.

After the stations have been drawn, we need to calculate weights for the stations. The strategy will be to assign volume to sampling cells and to sampled stations, then to assign revenues to the sampling cell and finally to combine sampling cells to assign prices for each publication cell. In order to assign volumes to cells we start with the EIA782 retail gasoline estimates by state for 2001. We will use several variables such as population, number of stations and proportion of families with cars to predict volume at a county level.

Within cells, however, different stations are likely to sell different proportions of gasoline of each grade, and different volumes. The proportion of gasoline of each grade will be imputed from the state totals. The total volume will be imputed by asking the stations how many cars can be filling up at the same time. An analysis using the New Images Marketing data base indicated a curvilinear relationship, and we will impute the median volume for the number of cars that can fill up at once. It is assumed that this figure will

be easier for the station to provide than their actual volume.

Using this approach we can obtain a volume and revenue for each sampling cell, and hence a price for each publication cell.

5.Results

The Chromy algorithm yielded an allocation of 875 stations. Table 1 presents the allocations, populations and sampling fractions for each of the 38 sampling cells in the design. One of the advantages of this design is that one can measure the variances early on and augment individual cells as necessary. The following table describes the allocations resulting from the Chromy algorithm.

Table 1

Number	Allocation	Population	Cell Description	Fraction
1	8	2,101	Boston in MA	0.38%
2	5	414	Boston not in MA	1.21%
3	5	746	NYC in PADD 1A	0.67%
4	5	378	Mass. Not in Boston	1.32%
5	26	1,687	Rest of conventional PADD 1A	1.54%
6	5	1,054	Rest of reformulated PADD 1A	0.47%
7	14	2,802	NYC in New York State	0.50%
8	8	2,204	NYC in rest of PADD 1B	0.36%
9	18	2,847	Conventional NY State	0.63%
10	5	150	Rest of reformulated NY State	3.33%
11	6	4,226	Rest of conventional PADD 1B	0.14%
12	17	3,693	Rest of reformulated PADD 1B	0.46%
13	13	658	Miami	1.98%
14	20	6,381	Rest of Florida	0.31%
15	29	16,256	Rest of conventional PADD 1C	0.18%
16	66	1,571	Reformulated 1C	4.20%
17	31	2,298	Chicago	1.35%
18	13	789	Cleveland	1.65%
19	24	2,797	Minnesotta	0.86%
20	19	3,677	Rest of Ohio	0.52%
21	93	27,811	Rest of conventional PADD 2	0.33%
22	33	1,933	Rest of reformulated PADD 2	1.71%
23	38	2,078	Houston	1.83%
24	20	7,542	Conventional Texas	0.27%
25	21	1,911	Rest of reformulated Texas	1.10%
26	30	10,786	Rest of PADD 3	0.28%
27	18	710	Denver	2.54%
28	17	1,121	Rest of Colorado	1.52%
29	59	2,904	Rest of PADD 4	2.03%
30	19	3,729	Los Angeles	0.51%
31	30	3,471	Rest of California	0.86%
32	16	1,441	Rest of Washington State	1.11%
33	41	2,710	Rest of conventional PADD 5	1.51%
34	5	834	Rest of reformulated PADD 5	0.60%
35	50	1,251	San Francisco	4.00%
36	36	739	Seattle	4.87%
37	5	254	Alaska	1.97%
38	7	334	Hawaii	2.10%
Total	875	128,288		0.68%

An analysis of the NIM data yielded the following median and quartile volumes for each number of pumps at the station:

pumps	n	Q1	median	Q3
1	10	15,000	6,000	5,000
2	104	30,000	20,000	15,000
4	268	60,000	45,000	40,000
6	126	90,000	75,000	65,000
8	336	125,000	105,000	90,000
10	52	160,000	140,000	120,000
12	89	200,000	165,000	150,000
16	15	500,000	230,000	180,000

Finally, a comparison of the OPIS and CBP databases (adding an auxiliary file of hypermarkets to the OPIS) indicated that the greatest undercoverage from the OPIS database was in the State of California, and a future examination of this phenomenon is planned.

References

Chromy, J. 1987 Design Optimization with Multiple Objectives, Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 194199

Saavedra, P. J., and P. Weir. 1991. A telephone survey of gasoline retailers drawn as a subsample of a national survey. Joint Meeting of the American Statistical Association, Atlanta, GA.

Weir, P., and P. J. Saavedra. 1998. Two multiphase petroleum price surveys that combine cycles at phase I. Presented at the Joint Statistical Meetings, American Statistical Association, August, Dallas.

Zayatz, L. and Sigman, R. 1995 Chromy_Gen: general-Purpose Program for Multivariate Allocation of Stratified samples Using Chromy=s Algorithm, Economic Statistical Methods Report series ESM-9502, June 1995, Bureau of the Census.