The Evolution of the Weekly Gasoline Price Survey through Changes in Design and Frame

Paula Weir, Benita J. O’Colmain, Pedro J. Saavedra

Weir, Energy Information Administration; Saavedra and O’Colmain, Macro International Inc., Paula.Weir@eia.doe.gov

Abstract

The Energy Information Administration (EIA) weekly survey of gasoline prices produces estimates of gasoline pump prices at the national and regional levels, as well as estimates for several states and cities, two formulations and three grades of gasoline. This survey originated as a response to the First Gulf War and was later expanded to monitor the effects of the Clean Air Act on the price of gasoline. The original design was a two-phase sample, with a monthly survey’s sample of gasoline refiners and resellers serving as the first phase, and the individual gasoline stations owned by the monthly respondents serving as the second. Subsequent expansions to produce estimates at lower geographic levels required the use of two cycles of the monthly survey sample and the Census County Business Patterns database. The most recent expansion of state and city estimates, however, required a new approach to the sampling frame and sample design. The new design is a two-stage sample where the first stage is an area sample in which counties are selected and the second stage is the sample of stations from the selected counties. The design made use of a newly constructed outlet level frame.. However, because the old corporate level frame and the new outlet level frame approached the population differently, the potential existed for discontinuity in the survey estimates. In order to attenuate the discontinuity, the sample drawn under the new design was combined with a subsample of the old design. Variance estimates of the new sample were obtained taking into account the procedure that combined the two samples and frames.

Survey Background and History

The August 1990 Iraqi invasion of Kuwait and the resulting rise in gasoline prices led to a need for more frequent monitoring of motor gasoline prices by an independent source. Pump price information was needed on a weekly basis (or more often) for unleaded, regular gasoline at the national level that was not only accurate, but could be obtained quickly and inexpensively. A survey to collect the data needed to be operational within a week**.** Commercial data were found to be insufficient in satisfying the timeliness, coverage and independence requirements. As a result of these limitations, it was decided that the EIA would conduct its own weekly price survey that would produce a national estimate of the price of regular gasoline at the pump.

The Weekly Motor Gasoline Price Survey was initially a survey of retail motor gasoline outlet prices drawn from the sample that was already in place for the monthly price survey of petroleum products. The new survey was intended to monitor consumer prices during the Persian Gulf War in 1990 and 1991 (Saavedra and Weir, 1991). The principal objective was to collect, process, and release the data to a variety of users, including policy makers and citizens, in a very rapid turn around mode. Specifically, Monday morning’s prices should be available by the end of the same day. A two-phase sample was used, with a subset of the monthly survey respondents as the first phase of the sample (Saavedra, 1988). In particular, monthly survey respondents who sold gasoline through retail outlets were sampled with probabilities proportional to sample weighted sales’ volumes in each state, as reported in the monthly survey. For the second phase, one or more gasoline stations in a given reported state were sampled, according to the size of each company selected in the first phase. The appeal of the design was that it: 1) permitted the use of a simple average as the price estimator, thereby simplifying the development of an over-night survey processing system; 2) allowed for quick implementation of the sample because of the ongoing monthly contact with those companies.

The survey was later iteratively expanded in response to the Clean Air Act and, eventually, estimates for Conventional, Oxygenated, Reformulated and OPRG gasoline for the five geographic regions known as Petroleum Administration for Defense Districts (PADDs), three sub-regions or sub-PADDs, and the State of California were added, as well as estimates for midgrade and premium grades of gasoline. These expansions in detail required an increase in sample size and presented the difficulty that the first-phase sample from the monthly survey was not sufficiently large to provide the allocations needed to meet the targeted accuracy. As a result, two survey rotation cycles of the monthly survey (current and previous samples) were combined in order to form Phase 1 (Weir and Saavedra, 1998).

More recently, EIA conducted a two-part expansion of the weekly gasoline price survey. As part of the first expansion, the sample was augmented minimally to allow release of average prices for 5 states (one in each PADD) and 6 cities, in addition to the regional and U.S. average prices previously released. The first expansion also included the collapsing of formulation designations into just two types. The newly combined categories were then backcast to provide the data users a historical continuous database

The second part of the expansion required a complete redesign of the weekly price survey (Saavedra et. al., 2002). In this expansion, several more cities and states were added to the required estimates rendering a two-phase design no longer viable for the estimates required. The previous expansion was performed for the most part in cities and states that already satisfied allocations that would provide the required CVs. For this expansion, the monthly survey, which collected company/state level prices, was no longer sufficiently large, nor could it target the sample at the city level. As a result, it was determined that a new frame was needed in order to develop a sample that would meet the necessary requirements and be flexible in meeting future requirements. In response to this need, the EIA developed a retail gasoline outlet frame and designed a new version of the weekly gasoline sample. The new sample design entailed using a totally new outlet level frame that only overlapped with the old company level frame when a company reporting retail gasoline sales from the monthly survey was represented by retail gasoline outlets on the new frame. However, identifying and measuring this overlap was virtually impossible, increasing the possibility for discontinuity between comparable price estimates produced from the old and new designs.

The Frame Development

Nine potential sources of nationwide lists of gasoline outlets with names and addresses were evaluated to determine the basis for the sampling frame. The primary focus of the evaluation was comprehensiveness of the data, although other factors such as availability of volumetric information, frequency of updates to the database, and expense in terms of database costs and level of effort needed to convert the data into a useable format were also considered (See Table 1).

Comprehensiveness was assessed based on the total number of outlets contained in each database, and, also by conducting a "zip code" test for each database. Twelve zip codes were randomly selected, six from urban areas and six from rural areas. Each vendor was then asked to provide a count of the number of retail outlets in those zip code areas. Comprehensiveness of the final sampling frame was assessed at the regional level using two different sources: 1) Regional estimates of retail gasoline stations published in the May 2002 issue of the National Petroleum News, and 2) 1999 U.S. Commerce Department County Business Patterns (CBP) estimates.

Table 1. Criteria for Evaluation of Available Data Sources

Main Issue	Variable Evaluated
Availability of Volumetric Data	Availability of SIC codes
	Availability of a variable that denotes volume of gasoline sold
Coverage	Comprehensiveness
	Absence of systematic bias
Format of the Data	Availability of a useable flat file
Expense	Cost
	Level of effort

The advantages and disadvantages of each of the initial nine data sources based on these criteria are presented in Table 2. The combination of all factors was considered in the selection of the source to be used for the sampling frame, however comprehensiveness and expense were weighted most heavily in the choice of the final data source.

The results of the zip code test to assess coverage in randomly selected urban and rural areas for each of the data sources are shown in Table 3, as well as the counts of outlets for each zip code obtained from the U.S. Census CBP database.

Table 2. Advantages and Disadvantages of Available Data Sources

Data Source	Advantages	Disadvantages
Dun & Bradstreet	• Flat file available • More comprehensive • Primary and secondary SIC codes available • Low level of effort required • No known systematic bias exists	• Expensive • Variable that denotes volume of gasoline sold not available
Survey Sampling, Inc.	• Flat file available • Inexpensive to obtain list of 3,000 records • Primary and secondary SIC codes available • Low level of effort required	• Expensive to obtain entire database • Inexpensive option requires that SSI selects sample records on our behalf • Less comprehensive • Variable that denotes volume of gasoline sold not available • Systematic bias exists for stations with P.O. box only, no phone number, no annual report
OPIS	• Flat file available • Medium expense • More comprehensive • Low level of effort required	• SIC codes not available; categorical data used instead • Variable that denotes volume of gasoline sold not available • Systematic bias may exist for stations that do not allow credit card usage
New Image Marketing	• Flat file available • More comprehensive • Variable that denotes volume of gasoline sold available (number of fueling positions) • Low level of effort required	• Expensive • SIC codes not available; categorical data used instead • Systematic bias exists as data are not collected for rural areas
Reference USA	• Inexpensive • Primary and secondary SIC codes available • No known systematic bias exists	• Flat file not available • Less comprehensive • Variable that denotes volume of gasoline sold not available • High level of effort required
Powerfinder	• Inexpensive • Primary and secondary SIC codes available • No known systematic bias exists	• Flat file not available • Less comprehensive • Variable that denotes volume of gasoline sold not available • High level of effort required
InfoUSA	• Flat file available • Medium expense • Primary and secondary SIC codes available • Low level of effort required • No known systematic bias exists	• Less comprehensive • Variable that denotes volume of gasoline sold not available
Acxiom/ USA Data	• Flat file available • Medium expense to obtain entire file • Inexpensive to obtain list of 3,000 records • Low level of effort required	• Inexpensive option requires that Acxiom/USA Data randomly selects sample records on our behalf • Less comprehensive • Secondary SIC code not available as selection criterion • Variable that denotes volume of gasoline sold not available • Systematic bias exists to the extent that secondary SIC codes are not available as selection criterion
Internet Yellow Pages	• Free	• Flat file not available • Less comprehensive • SIC codes not available • Variable that denotes volume of gasoline sold not available

		• High level of effort required

Table 3. Zip Code Comparison for Available Data Sources

	Urban						Rural						Total number of stations in all 12 zip codes	Total number of stations availabl e from source
City	New York City	Chicago	Houston	Denver	Los Angeles	San Francisco	Theresa	Walnut Grove	Lostant	Queen City	Fountain	Norwalk		
State	NY	IL	TX	CO	CA	CA	NY	MN	IL	TX	CO	CA		
Zip Code	10033	60615	77058	80226	90057	94112	13691	56180	61334	75572	80817	90650		
Dun & Bradstreet	2	6	9	7	1	9	3	1	0	3	1	15	57	108,073
Survey Sampling Inc.	3	6	8	11	1	9	3	0	0	2	2	17	62	78,674
OPIS	2	6	5	10	0	8	1	2	0	5	4	20	63	114,861
New Image Marketing	5	8	10	26	1	22	NA	NA	NA	NA	4	24	100	86,000
Reference USA	5	5	3	13	1	13	0	0	0	1	2	20	63	81,392
Powerfinder	5	5	11	6	2	11	0	1	0	2	4	3	50	78,651
InfoUSA	4	5	3	13	1	11	0	0	0	0	2	20	59	69,209
Acxiom/ USA Data	3	6	7	16	1	7	1	0	0	2	1	14	58	63,341
Yellow Pages	4	5	4	11	0	12	0	0	0	0	1	20	57	NA
Census	**5**	**4**	**6**	**17**	**1**	**13**	**2**	**1**	**0**	**3**	**3**	**25**	**80**	**121,082**

Based on the evaluation of all factors and the zip code comparison for comprehensiveness, the Oil Price Information Service (OPIS) database of gasoline outlets was chosen as the basis for the sampling frame. However, the OPIS list excluded gasoline outlets owned by hypermarkets, including mass retailers such as supermarkets, discount retailers, and warehouse clubs, which had begun selling gasoline at their locations in the last few years. To adjust for this known area of under coverage, lists of hypermarket-branded gasoline outlets were obtained from nine top hypermarket companies and were used to supplement the OPIS database. One source, New Image Marketing (NIM), which had limited market coverage but other variables, was used for analytic purposes to determine the relationship between the type of station, sales volume, and number of pumps. Table 4 presents the regional and national comparison of retail gasoline outlet counts from the National Petroleum News and the U.S. Census Bureau with those from the selected and augmented sampling frame, OPIS augmented by the hypermarket data. The National Petroleum News data suggests there were just over 170,000 outlets in the year 2002.

Table 4. Number of Outlets by Region and Percentage of National Total

		Number			Percent		
PADD	**Region**	**NPN[1]**	**CBP[2]**	**OPIS+HM[3]**	**NPN**	**CBP**	**OPIS+HM**
1	East Coast	61,842	44,560	42,912	36.2%	36.8%	37.0%
2	Midwest	50,731	37,216	35,787	29.7%	30.7%	30.9%
3	South Central	34,490	20,492	20,968	20.2%	16.9%	18.1%
4	Mountain	7,177	4,911	4,971	4.2%	4.1%	4.3%
5	West Coast	16,438	13,903	11,292	9.6%	11.5%	9.7%

[1] Source: National Petroleum News, Volume 94, Number 5 (May 2002), p. 6 and previous years of Market Facts (mid-July). Underlying survey is conducted in the first quarter of each year.

[2] CBP 1999 numbers are from the U.S. Department of Commerce and only include a fraction of outlets selling gasoline due to definition restrictions, i.e. outlet must have employees, payroll and more than half of outlet income must be from gasoline sales.

[3] National total includes 114,861 stations from OPIS and 1,069 stations from the aggregated hypermarket (HM) data file.

		Number			Percent		
PADD	Region	NPN[1]	CBP[2]	OPIS+HM[3]	NPN	CBP	OPIS+HM
	National Total	170,678	121,082	115,930	100.0%	100.0%	100.0%

The Sample Design

In addition to the newly constructed outlet frame, other data sources were used for sampling and weighting purposes. These included the counts of gasoline stations per county obtained from the Census Bureau's County Business Patterns (CBP) database, unpublished EIA data from the monthly survey used to assign a percentage of sales by grade to each station within a state, the NIM data, and the current weekly gasoline survey prices used for estimating unit variances for various sampling cells. A comparison of the OPIS data set and the CBP indicated 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. To handle this issue the maximum of the two numbers was assumed as the number of stations in a county. This meant that if a county had 20 stations listed in OPIS, 24 in the CBP and 1 station is allocated to the county, it was assumed that the one station was sampled from 24, even though only 20 were available in the frame.

The new sample design was driven by the definitions of Publication Cells and Sampling Cells. A Publication Cell is one defined by a PADD, state, city and attainment status. Hence, PADD 2 reformulated gasoline is a publication cell. Another example of a publication cell is New York State, conventional gasoline. Sampling cells are the smallest units whose borders are defined by publication cells, but for which estimates are not published. Sampling cells contribute to publication cells. For example, the part of New York State where reformulated gasoline is required, but is not in New York City would be a sampling cell, contributing to the state estimate.

Sample sizes were determined using the historical survey data, and auxiliary data where necessary to obtain the desired coefficients of variation for all three grades of gasoline and for totals for each publication cell. These sample sizes were first converted into an allocation for each sampling cell, made up of counties, and then transformed into allocations (possibly fractional) for each county. More specifically, using the unit variance in each sampling cell from the historical weekly prices, Chromy's Allocation Algorithm (Chromy, 1987, Zayatz and Sigman, 1995) was employed to calculate allocations for each sampling cell. A minimum of five stations was assigned to each sampling cell. Using the maximum outlet count described above, each county's proportion of outlets in the sample cell was calculated. Each county's proportion was then multiplied by the allocation for the sample cell to derive the county allocation. The county allocation helped to mitigate the potential frames coverage issues mentioned above. However, because these allocations were not necessarily integers, only the integer part of the allocation was assigned to the county, and counties were selected (with replacement) with probabilities proportional to the fractional part of the allocation. Thus, if a county had an allocation of 2.3 stations, the sampling procedure assigned it at least two stations, and the third station was assigned with a probability of .3.

Once integer allocations were assigned for each county, stations were selected from each sampling cell in two stages. For the first stage, a Goodman-Kish PPS sampling method was used, ordering counties within states by the number of stations in that state. Counties were selected with probabilities proportional to the number of stations that did business in the county, based on the CBP, allowing the possibility of more than one station in the county if the formula for the probability of selection of the county yields a number greater than one. For the second stage, stations were randomly selected from the selected counties using the new outlet frame, which had been mapped to county designations.

The Chromy algorithm yielded an allocation of 875 stations. Table A1 in the Appendix presents the allocations for the new sample, and a comparison between the number of stations reported by CPB and the number on the constructed frame for each of the 38 sampling cells in the design. It should be noted that some of the sampling cells (e.g. cities contained in only one State, the state of Minnesota, the reformulated part of PAFDD 1C and the conventional part of New York State) are also publication cells, while others are not. However, every publication cell can be expressed as the union of sampling cells (for example, combining cells 7, 9, and 10 yields New York State, and combining cells 30 to 38 yields PADD 5.

For the purpose of estimating average prices for sampling cells, sample weights were constructed using information obtained from the sampled outlet during initiation on the number of gas pumps at the outlet. The number of pumps served as a proxy for sales' volume based on the research conducted using the NIM data. These weights were applied to the reported weekly outlet prices to obtain averages for the various grades within the sampling cells. An outlet's proportion of sales by grade was assigned using the monthly survey volume data based on the State where the outlet was located. In other words, the weights were used to estimate total volume and the proportion of sales in the state assigned the appropriate amount to each of the three grades. Similarly, the sampling cells were combined to form publication cells using the cell's volume from the monthly survey.

One advantage of the above approach is that should early variance estimates indicate that standard errors are too large for some publication cells, the sampling cells forming the publication cells can be examined, and the sample can be augmented in one or more sampling cells.

Implementation and Data Continuity

As part of the process used each time that a new sample is implemented, prices were collected initially from both the new and the old samples. This insures continuity as the overlapping data are used to study and smooth any discontinuities that might result with frame, sample, and respondent changes. A comparison of prices was made between the old sample and the new sample for the 200 publication cells that the two surveys had in common after the necessary formulation cells were combined and the new areas included. The comparison indicated that a number of cells were unacceptably discrepant. Further examination revealed wide differences in coverage between the sample cycles. The old cycle contained a number of stations owned by refiners or large resellers not present in the new cycle, while the new cycle contained mostly smaller independent outlets. As shown in Table 5, for conventional gasoline, the average differences appeared consistent across grade (with the exception of the Gulf Coast) but varied among the regions with the largest differences occurring in the Lower Atlantic and Central Atlantic regions.

Table 5. Average Regional Price Differences for Overlap-Weeks, Conventional Gasoline

	Grade		
Region	**Regular**	**Midgrade**	**Premium**
New England	$0.015	$0.018	$0.014
Central Atlantic	$0.021	$0.039	$0.039
Lower Atlantic	$0.035	$0.027	$0.035
Midwest	$0.022	$0.019	$0.020
Gulf Coast	$0.008	$0.008	$0.020
Rocky Mountain	$0.011	$0.012	$0.010
West Coast	$0.020	$0.019	$0.006
U.S.	$0.016	$0.015	$0.013

Likewise, the price differences at the city level were larger (not previously targeted or published) and were consistent across grade with the exception of premium in Houston, which deviated less than the other grades, contrary to the regional Gulf Coast difference, as shown in Table 6. The largest city differences occurred in Los Angeles.

Table 6. Average City Price Differences for Overlap-Weeks

	Grade		
City	**Regular**	**Midgrade**	**Premium**
Chicago	$0.028	$0.029	$0.020
Denver	$0.013	$0.006	$0.015
Houston	$0.021	$0.017	$0.005
Los Angeles	$0.030	$0.034	$0.030
New York City	$0.009	$0.012	$0.010
San Francisco	$0.013	$0.019	$0.023

This same pattern was exhibited at the state level, varying from state to state, consistent across grades with the exception of premium grade in Texas, as well as Minnesota midgrade, as shown in Table 7. The largest differences occurred in California.

Table 7. Average State Price Differences for Overlap-Weeks

	Grade		
State	**Regular**	**Midgrade**	**Premium**
California	$0.036	$0.038	$0.035
Colorado	$0.017	$0.024	$0.020
Minnesota	$0.015	$0.025	$0.015
New York	$0.021	$0.037	$0.034
Texas	$0.006	$0.006	$0.020

Three of these states (California, New York, and Texas) sell both reformulated and conventional gasoline, which might have been a contributing factor to discontinuity as a result of the change in the formulation weights. In fact, the combined formulation effect was further demonstrated as shown in Table 8, which presents average regional price differences for all formulations, as compared to Table 5, which presents only conventional formulation. The differences almost doubled in the New England and West Coast regions, but were cut in half for the Central Atlantic region. These regions contain significant clusters of reformulated gasoline. The other regions contain either only small amounts of reformulated gasoline, or none at all.

Table 8. Average Regional Price Differences for Overlap-Weeks, All Formulations

	Grade		
Region	**Regular**	**Midgrade**	**Premium**
New England	$0.036	$0.033	$0.035
Central Atlantic	$0.005	$0.022	$0.024
Lower Atlantic	$0.036	$0.029	$0.034
Midwest	$0.020	$0.020	$0.019
Gulf Coast	$0.006	$0.006	$0.014
Rocky Mtn.	$0.011	$0.012	$0.010
West Coast	$0.039	$0.039	$0.032
U.S.	$0.019	$0.022	$0.018

The effect of reformulated gasoline in the regions was further demonstrated by examining the regional price differences for reformulated gasoline. In all regions and grades, with the exception of all grades in the Central Atlantic region, and premium grade for the Lower Atlantic, Midwest, and Gulf Coast, the differences are even larger, as shown in Table 9.

Table 9. Average Regional Price Differences for Overlap-Weeks, Reformulated

	Grade		
Region	**Regular**	**Midgrade**	**Premium**
New England	$0.053	$0.051	$0.051
Central Atlantic	$0.002	$0.020	$0.020
Lower Atlantic	$0.043	$0.036	$0.009
Midwest	$0.024	$0.040	$0.016
Gulf Coast	$0.019	$0.020	$0.005
West Coast	$0.029	$0.032	$0.028
U.S.	$0.030	$0.034	$0.026

To further understand these price differences, a bootstrap was used to estimate standard errors for each sample using data from the third overlap week. For each bootstrap sample, the number of stations allocated to a cell was sampled with replacement from the cell, and the estimates calculated. The resulting standard errors and CVs were in accordance with what the design had targeted. A z- test was then used to test the differences between the samples for significance, where z was calculated as:

$$z = M_{new} - M_{old})/(s^2_{old}+s^2_{new})^{1/2}$$

In 46 of the 200 publication cells common to the old and the new sample, the value of z was greater than 1.96 (i.e., the probability of the differences happening by chance was less that .05). The median difference across the cells was 1.8 cents, but varied across the overlap weeks as shown in Table 10. The number of cells for each overlap week that were significantly different between the two samples, and the number of cells for which the coefficient of variation was less than .01 are also shown in Table 10.

The price difference for one week and the standard deviation by grade for the 38 sampling cells are shown in Appendix Table A2.

Table 10. Median Absolute Difference and CVs for 200 Common Cells

Week	Median Diff.	Old, Median CV	New, Median CV	# signif. Diff. P < .05	# cells old CV < .01	# cells new CV < .01
Week 1 Overlap	$0.019	0.007	0.007	37	163	158
Week 2 Overlap	$0.027	0.007	0.007	72	163	161
Week 3 Overlap	$0.018	0.006	0.007	46	172	152

Coverage Issues Identified after Sample Initiation

Due to the large discontinuities in price estimates evidenced in some areas after the new sample was initiated, the sampling frame was examined further to determine whether coverage issues might account for some of the price differences. A review was conducted to determine whether stations from the old sample in six cities with published prices were represented on the sampling frame. Additionally, the sampling frame was analyzed to determine whether stations owned by the largest five companies (according to sales volume) in each state were included.

Several strategies were used to determine whether companies and specific outlets were represented on the frame. Companies were identified primarily by name and brand of gasoline sales. When searching for individual stations, the file was sorted on various variables (e.g., station name, state, city, address, phone number) and then visually examined for the street address of the station in question. But it was often difficult to determine exactly which outlets and how many were owned and operated by a particular company due to differences in naming conventions for the individual stations and the inability to determine whether some stations sold branded gasoline but were independently owned. Ownership status for only the sampled stations was ascertained during the initial phone contact with each station.

Results of searching the frame for each outlet in the old sample that contributed to the six previously published city prices indicated that stations were omitted from the frame with rates ranging from 10.8 percent in Denver to 58.8 percent in Los Angeles. Of the stations used to generate the Chicago estimate, 14.3 percent of stations were not found on the sampling frame. In Houston 36.6 percent were omitted from the frame, in New York, 21.2 percent were not found on the frame, and 31.0 percent of stations in San Francisco were missing from the frame. In Los Angeles, a single company was identified that owned 70 percent of the missing stations. Additional review revealed that none of more than 350 Los Angeles area stations from this company were included on the sampling frame. It was later learned that this company did not allow credit card transactions which was the reason for not appearing on the OPIS outlet frame. Among the 21.2 percent of New York stations not found on the frame, more than half were from one major company that is a top seller in eight states, including three published States.

Results of the frame search for the top five companies in sales volumes for each state indicated that at least six of these companies were not included on the sampling frame. Additionally, it was difficult to discern whether one of the largest five companies in Washington, D.C. was included on the frame. These seven companies were top volume sellers across

15 states and the District of Columbia, including two states with published prices, Florida and Texas. An additional eight of the top five companies in at least one State were under-sampled, suggesting these companies were under-covered on the frame. These eight companies were among the top five companies across seventeen unpublished States and seven published States. Additionally, two of the top five sellers were under-represented in seven States, including three published States. Finally, in New York, a published State, three of the top five companies appeared to be under-covered.

After targeted matching of the old and new samples and the new frame, it was determined that the price differences were primarily attributable to incomplete frame coverage (see O'Colmain, Churchill, Weir and Saavedra). These coverage issues impacted the city, state, regional and U.S. level estimates.

Revising the Sample

It appeared from this analysis that the two samples represented different populations of outlets, neither representing the complete population. As a result, the alternative of combining the two samples was considered. This not only was expected to reduce the discontinuity, but also was expected to produce more accurate estimates as the coverage of the combined frame would be greater than that of either frame by itself. The following steps were taken to draw the composite sample.

1) The outlets in the old sample were classified into the new sample's sampling cells.
2) Within each sampling cell, the largest weight was identified, and all the outlets in the cell were subsampled with probabilities proportional to the ratio of the weight in the old sample to the maximum weight in the old sample for outlets in that cell. This step had the effect of making the old sample stations in each new sample-sampling cell have the same probability of selection. Thus, if in a given cell some of the stations had a probability of .02, and some a probability of .06, we would select all the stations with .06 probability and one third of the stations with .02 probability. This would give all the stations in the cell the same probability of selection (.06).
3) New unit variance estimates were obtained for each sampling cell using all stations in the cell from the old or new sample. The Chromy algorithm was then rerun and new allocations obtained.
4) If a sufficient number of outlets were available from each sample, fifty percent of the allocations were filled by old sample outlets and fifty percent by new sample outlets.
5) For some of the new publication cells (e.g. Miami and Cleveland) an insufficient number of outlets existed in the old sample, so the bulk of the outlets in the revised sample came from the new sample.
6) Outlets from the old sample were presumed to represent larger companies and outlets, so the number of pumps used as the estimation weight was set to the maximum.
7) To smooth the transition from the old sample estimates to the composite sample estimates, estimates for the third overlap week were recalculated using the composite sample. The ratios of the composite sample estimates to the old sample estimates for that week were then applied to the first non-overlap week of the composite sample. The following week the ratios were adjusted so that r'= (2r+1)/3, and the following week adjusted again to r''= (r+2)/3. This procedure minimized the discontinuity in the estimates due to the change in samples.
8) Bootstraps were conducted for the composite sample, preserving the proportion of old and new sample companies per stratum for every bootstrap sample.

The estimates that were generated for the third overlap week using the composite sample were compared to the old sample estimates for that week. This comparison showed:

- 165 of 200 cell differences were reduced
- The mean difference between the new and old sample of 1.74 cents was reduced to a 0.05-cent difference between the composite sample and the old sample
- The median difference between the new and old sample of 1.8 cents was reduced to a 0.0-cent difference between the composite sample and the old sample.

Table 11 shows the differences in regional prices for reformulated gasoline for the third overlap week for the originally selected new sample and the composite sample.

The median price, standard error, and coefficient of variation, across all cells for the composite sample, for the last overlap week and the two weeks after smoothing was completed are shown in Table 12. These composite sample estimates and their errors satisfied the original sample requirements.

Variance Estimation for the New Sample

Estimating variance for the new sample required the use of a bootstrap, where the outlets that came from the old frame and the ones that came from the new frame were each selected with replacement in the same number per stratum as in the actual sample. The one exception was a stratum with only one old sample outlet, where the two samples were bootstrapped as if they came from one source. Appendix A presents the resulting standard errors and CVs for the month.

Table 11. Overlap Week 3, Regional Price Differences, Reformulated, New vs. Composite

Region	Sample	Regular	Midgrade	Premium
New England	new	0.058	0.056	0.051
	comp	0.015	0.012	0.011
Central Atlantic	new	0.004	0.020	0.018
	comp	0.014	0.003	0.009
Lower Atlantic	new	0.040	0.033	0.006
	comp	0.027	0.019	0.005
Midwest	new	0.008	0.020	0.003
	comp	0.001	0.007	0.005
Gulf Coast	new	0.023	0.022	0.004
	comp	0.015	0.008	0.002
West Coast	new	0.033	0.033	0.023
	comp	0.019	0.018	0.018
U.S.	new	0.028	0.031	0.021
	comp	0.000	0.003	0.007

Table 12. Composite Sample Median Price, SE and CV (all cells)

Week	Price	SE	CV
Overlap Week 3 (5/19)	$1.583	$0.010	0.006
7/14	$1.603	$0.009	0.006
7/21	$1.617	$0.009	0.006

Summary

In order to meet formulation changes in the industry reflecting regulatory requirements, as well as, increased user requirements for more city and state level estimates of gasoline pump prices, the sample design of the gasoline price survey has evolved to meet these changing requirements. Originally, the survey requirement was to produce a timely U.S. price of regular gasoline in the first Gulf war. This requirement was later expanded to other grades and then to multiple formulations to monitor the effects of regulatory requirements stemming from the Clean Air Act. More recent requirements have centered on producing lower geographic level prices, select cities and states. In order to accommodate this last change in requirements, a more targeted gasoline outlet frame was constructed. The initial list of outlets from OPIS was purchased after examining and comparing nine sources, but was immediately augmented to cover hypermarket outlets not adequately represented on the list. This frame appeared to be fairly comprehensive, containing 115,930 gasoline stations after augmentation for hypermarket outlets in contrast to the 121,082 gasoline stations reported by the Census Bureau's *County Business Patterns.*

After selecting the outlet sample, using first an area selection of counties and then selecting outlets within the counties, the two samples were overlapped and estimates were compared. The differences between the two sample estimates over a three-week period motivated the selection of a composite sample. The composite sample, including stations from both the old and new sample, mitigated the effects of the new frame's under-coverage of some of the larger gasoline sellers. The composite sample produced a data series more continuous with the old series and reduced the CVs while still satisfying the new requirement for accurate gasoline price estimates to be produced for an extended number of states and cities, three grades, and two formulations.

References

O'Colmain, B., T. Churchill, P. J. Saavedra, and P. Weir. (2003). "Non-sampling Error in the New Cycle of the Weekly EIA Gasoline Price Survey." Presented at the Joint Statistical Meeting, American Statistical Association, San Francisco, CA, Section on Survey Research Methods.

Saavedra, P. J. (1988). "Linking multiple stratifications: Two petroleum surveys", *1998 Proceedings of the American Statistical Association,* Section on Survey Research Methods [CD-ROM], Alexandria, VA: American Statistical Association, pp. 777–781.

Saavedra, P. J., and P. Weir. (1991). "A Telephone Survey of Gasoline Retailers Drawn as a Subsample of a National Survey", *1991 Proceedings of the American Statistical Association,* Section on Survey Research Methods [CD-ROM], Alexandria, VA: American Statistical Association.

Saavedra, P. J., P. Weir, B. O'Colmain, T. Churchill, and E. Carlton. (2002). "The New Design for the EIA-878 Gasoline Price Survey", *2002 Proceedings of the American Statistical Association,* Section on Survey Research Methods [CD-ROM], Alexandria, VA: American Statistical Association.

Weir, P. and P. J. Saavedra. (1998). "Two Multiple-Phase Surveys that Combine Overlapping Sample Cycles at Phase 1", *1998 Proceedings of the American Statistical Association,* Section on Survey Research Methods [CD-ROM], Alexandria, VA: American Statistical Association, pp. 443-447.

Weir, P. , Saavedra, P.J. , O'Colmain, B. and Churchill T., "Quality of Estimates in the New Cycle of the Weekly EIA Gasoline Price Survey", Presented at the Joint Statistical Meeting, American Statistical Association, San Francisco, CA, Section on Survey Research Methods.

APPENDIX A

Table A1: Allocation and Population of Cells by CPB and Frame

#	Cell Description	New Sample*	CPB Stations	Frame Stations
1	Boston in MA	8	2,101	1,686
2	Boston not in MA	5	373	412
3	NYC in PADD 1A	5	702	746
4	Mass. Not in Boston	5	365	351
5	Rest of conventional PADD 1A	26	1,548	1,590
6	Rest of reformulated PADD 1A	5	1,021	948
7	NYC in New York State	14	2,787	2,221
8	NYC in rest of PADD 1B	8	2,204	1,993
9	Conventional NY State	18	2,836	2,272
10	Rest of reformulated NY State	5	126	145
11	Rest of conventional PADD 1B	6	3,977	4,027
12	Rest of reformulated PADD 1B	17	3,591	3,499
13	Miami	13	658	626
14	Rest of Florida	20	5,915	6,024
15	Rest of conventional PADD 1C	29	14,848	14,996
16	Reformulated 1C	66	1,508	1,376
17	Chicago	31	2,255	2,125
18	Cleveland	13	785	729
19	Minnesota	24	2,553	2,669
20	Rest of Ohio	19	3,571	3,207
21	Rest of conventional PADD 2	93	26,130	25,357
22	Rest of reformulated PADD 2	33	1,922	1,700
23	Houston	38	2,014	1,998
24	Conventional Texas	20	6,806	7,040
25	Rest of reformulated Texas	21	1,792	1,910
26	Rest of PADD 3	30	9,880	10,020
27	Denver	18	707	587
28	Rest of Colorado	17	1,042	1,044
29	Rest of PADD 4	59	2,521	2,739
30	Los Angeles	19	3,729	2,736
31	Alaska	5	238	203
32	Rest of California	30	3,457	2,548
33	Hawaii	7	334	231
34	Rest of Washington State	16	1,340	1,337
35	Rest of conventional PADD 5	41	2,637	2,333
36	Rest of reformulated PADD 5	5	834	816
37	San Francisco	50	1,251	992
38	Seattle	36	724	697

*Prior to revision

Table A2. Price Difference and Standard Deviations for One Week by Sample Stratum

Cell Description	Diff. Regular	Diff. Midgrade	Diff. Premium	SD Regular	SD Midgrade	SD Premium
Boston in MA	-0.002	-0.004	-0.014	0.049	0.051	0.057
Boston not in MA	0.020	0.020	-0.003	0.049	0.049	0.049
NYC in PADD 1A	-0.015	-0.023	-0.054	0.060	0.078	0.077
Mass. Not in Boston	0.008	0.006	0.008	0.077	0.087	0.081
Rest of conventional PADD 1A	-0.011	-0.009	-0.006	0.075	0.076	0.072
Rest of reformulated PADD 1A	-0.009	-0.015	-0.005	0.052	0.056	0.058
NYC in New York State	0.067	0.077	0.098	0.064	0.062	0.071
NYC in rest of PADD 1B	0.001	0.007	-0.005	0.038	0.047	0.047
Conventional NY State	0.013	0.006	0.000	0.030	0.030	0.032
Rest of reformulated NY State	0.001	0.021	0.011	0.048	0.050	0.050
Rest of conventional PADD 1B	0.065	0.077	0.086	0.066	0.072	0.074
Rest of reformulated PADD 1B	-0.004	0.017	0.028	0.110	0.115	0.126
Miami	NA	NA	NA	0.056	0.066	0.060
Rest of Florida	0.032	0.039	0.048	0.068	0.072	0.073
Rest of conventional PADD 1C	0.043	0.041	0.046	0.070	0.070	0.072
Reformulated 1C	-0.013	-0.008	-0.034	0.089	0.095	0.094
Chicago	-0.049	-0.040	-0.047	0.077	0.080	0.082
Cleveland	0.068	0.066	0.066	0.084	0.095	0.093
Minnesota	0.012	-0.002	-0.006	0.045	0.053	0.053
Rest of Ohio	0.062	0.092	0.088	0.069	0.072	0.077
Rest of conventional PADD 2	0.033	0.034	0.042	0.073	0.076	0.088
Rest of reformulated PADD 2	0.060	0.070	0.023	0.111	0.114	0.126
Houston	0.025	0.014	0.004	0.044	0.046	0.047
Conventional Texas	0.000	0.012	-0.001	0.059	0.065	0.079
Rest of reformulated Texas	0.038	0.045	0.034	0.049	0.052	0.055
Rest of PADD 3	0.046	0.018	0.009	0.058	0.063	0.063
Denver	-0.028	0.001	0.002	0.067	0.068	0.067
Rest of Colorado	0.026	-0.001	0.032	0.085	0.084	0.081
Rest of PADD 4	0.040	0.029	0.034	0.065	0.064	0.070
Los Angeles	0.041	0.040	0.040	0.047	0.050	0.047
Alaska	0.019	0.085	0.005	0.071	0.063	0.056
Rest of California	0.003	-0.003	-0.005	0.071	0.069	0.072
Hawaii	0.108	0.116	0.129	0.117	0.122	0.137
Rest of Washington State	0.021	0.014	0.011	0.040	0.047	0.049
Rest of conventional PADD 5	0.041	0.053	0.040	0.092	0.090	0.096
Rest of reformulated PADD 5	0.033	0.029	0.023	0.036	0.039	0.037
San Francisco	0.004	0.003	-0.003	0.070	0.073	0.074
Seattle	0.075	0.045	0.034	0.078	0.070	0.071