The Impact on Data Quality of the Transition to Clean Burning On-Highway Diesel

Paula Weir[1], Benita O'Colmain[2], Pedro Saavedra[2]
[1]Energy Information Administration, Washington, DC, 20585
[2]Macro International Inc.

Abstract

The U.S. Environmental Protection Agency has required that beginning in June 2006 refiners and importers of petroleum must ensure that at least 80 percent of the volume of on-highway diesel fuel they supply be ultra-low sulfur diesel (ULSD). By December 2010, all on-highway diesel is required to be ULSD. Between 2006 and 2010, both ULSD and low sulfur diesel (LSD) may be offered for sale at retail locations outside of California, with some diesel fuel outlets carrying both fuels and others choosing to sell only one or the other. Until January 2007, EIA has collected the price of on-highway diesel fuel without distinguishing the sulfur level. This paper describes how the weekly diesel price survey was modified to account for the transition to ULSD. Evaluations of the variance using a bootstrap method and sensitivity analysis to explore the impact of alternate assumptions are presented.

Keywords: variance; bootstrap; on-highway; ULSD; price

1. Background

Each Monday morning, the Weekly Retail On-Highway Diesel Price Survey, the EIA-888, collects prices of diesel fuel. These data are collected primarily through computer assisted telephone interview (CATI), but also by email, fax and company web site from approximately 350 retail diesel outlets including truck-stops and service stations across the country. Average prices are calculated for the U.S. and regions of the country and released by 5 P.M. on the day of collection through EIA's web site, list serve, and wireless service. These data are widely used by shippers and truckers to determine fuel surcharges as negotiated privately by the shipper and the trucking companies.

The sample design for the weekly diesel price survey is a two-phase design in which units in a monthly survey sample who report diesel fuel volume sales through company-operated outlets constitute the first phase. For the second phase, a sub-sample of those units is selected using probability proportional to size (PPS) with the annual state sales volumes from the monthly survey divided by the unit's probability of selection in the monthly survey as the normalized measure of size for the company-state unit. Within the second phase is a second stage to identify the actual outlets reporting for the company. This identification was done by contacting the sampled companies and asking them to provide the names, addresses and telephone numbers for the number of outlets in each State that the company-state unit was sampled.

Sample allocations were calculated using the average standard errors across reporting periods for the previous year of weekly diesel fuel survey prices for each of the cells. An average sample size was then determined for each cell by the formula: $n' = (e/t)^2 n$, where t was the target standard error, n was the previous sample size for the cell, e was the average of the previous sample's standard errors, and n' was the new sample cell allocation.

In addition, a second allocation based on proportional representation within the next larger aggregation cell to which the original cell would contribute was also obtained. For example, the PADD IB cell contributes to the PADD I cell. The maximum of the two allocations for each cell was then designated as the cell allocation.

The price estimates were obtained through simple averages at the sampling cell level. For publication cells that constituted a combination of sampling cells, the volume of the diesel sold as collected by an EIA monthly survey was used to weight the cells and obtain publication cell prices.

As of June 1, 2006, the Environmental Protection Agency required refiners and importers to ensure that at least 80 percent of the volume of on-highway diesel fuel they produce or import was ultra-low sulfur diesel (ULSD), defined as diesel fuel with less than 15 parts-per-million sulfur. Furthermore, EPA required that by September 1, 2006, diesel fuel classified as ULSD must reach distribution and marketing points downstream from refineries (July 15, 2006 in California). As a result, it was expected that ULSD would be available at many retail locations by October 15, 2006 (September 1, 2006 in California). However, the transition to ULSD at all retail locations was not required to be completed until December 1, 2010. As a result, diesel fuel classified as low sulfur diesel fuel (LSD), i.e., diesel fuel with sulfur content between 15 and 500 parts-per-million, could still

be sold at retail locations outside of California until December 1, 2010.

Historically, EIA had collected the price of on-highway diesel fuel without distinguishing the sulfur level. In order to measure the price impact through the transition to ULSD, the weekly on-highway diesel price survey was adjusted to collect diesel prices for LSD and ULSD separately. The sample size and set of respondents were preserved but the survey instrument was modified to ask the current sample of outlets which fuels they sold and the price of each sold. Preliminary research and discussions with key respondents had revealed that most outlets would switch from LSD to ULSD in accordance with what was provided by their supplier, but some outlets could be expected to sell both types of diesel for some period of time. It was therefore necessary to determine how to construct average prices for both fuel types given the current weight construct, and to define criteria for determining which prices were of high enough quality to be released.

2. Determination of Weights

2.1 The "Don't Knows" and Amount Sold Problems

Preliminary discussions with industry associations provided an early warning that respondents at individual outlets might not be aware of which type of diesel fuel was sold at that location. As a result, interviewers were instructed to first ask respondents which type of diesel fuel they sold. If the respondent said "don't know", the interviewers further prompted the respondent to identify the fuel type by checking the label on the pump. Federal regulations require the labeling of all diesel fuel pumps to specify the type of fuel dispensed. Respondents who further insisted they did not know the fuel type and could not leave their booth to read the label on the pump were instructed to check the pump label before the next Monday's price collection. These respondents were classified as "don't knows" but their prices still had to contribute to either the ULSD or LSD average price. In addition, there was no information available in the industry as to the percentage of sales that could be expected to be ULSD or LSD at any level—U.S., region, company, or outlet. These two unknowns, the diesel type for the "don't knows" and the proportion of diesel being sold by fuel type, were necessary for allocating the current weights to estimate average prices by diesel type while preserving the weight construct for the combined type.

2.2 The Assumptions Made and a Data Driven Algorithm for Weighting

Given that there was only company level volume information but no outlet diesel volume information available, the design of the diesel sample originally assumed that each retail outlet represented the same volume of sales as any other outlet in the same sampling cell. As a result, if every outlet in the sample sold only one fuel type, the proportion of outlets in the sample that reported ULSD would be a good estimate of the proportion of the volume of diesel that was ULSD and the same for LSD. If, for example, a sampling cell had 40 stations in the sample, and 30 said they sold ULSD and 10 said they sold LSD, the combined average diesel price could be obtained by multiplying the ULSD average price by .75 and the LSD price by .25 and adding the two numbers, or more simply, by averaging the prices of all 40 stations. Unfortunately though, the simplified example does not account for stations that would report selling both kinds of diesel and stations that would say they do not know what kind of diesel they sell. The original sample weight for the outlets selling both types would have to be allocated between the two types.

Given the economic cost to an outlet for the transition from LSD to ULSD, it was expected that once an outlet selling only LSD shifted to only ULSD, it was very unlikely to sell LSD again. Therefore, for data collection, when a single fuel-type seller reported selling ULSD, the CATI interviewers no longer asked which type of fuel the outlet sold. Respondents reporting that they sold LSD, and respondents reporting they sold both types, continued to be asked each week which type of fuel they sold that week. Respondents reporting "don't know" were assumed to sell LSD for average calculations but continued to be asked each week which type of fuel they sold. However, this price classification did not eliminate the need to determine the proper allocation of diesel volume for outlets selling both types of diesel within the construct of the weights previously used at the all-type of diesel level nor provide an allocation of the aggregate volume weights to derive higher level aggregates. For the outlet's selling two types of diesel, two different base alternatives were considered. These included:

1) Assume that stations that sell both types of diesel sell the same proportions as the proportions sold for their sampling cell by one-type outlets.
2) Assume that the stations that sell both types of diesel sell an equal amount of each.

For the weighting algorithm, the sampling cells were defined as the geographic areas released (PADDs, subPADDs, California and PADD 5 excluding California) indexed by j=1, 2 … m. Each of the sampling cells had associated with it the combined diesel volume, W_j, based on annualized diesel volume data reported on the EIA monthly survey, "Monthly Report of Prime Supplier

Sales of Petroleum Products Sold for Local Consumption" (September 2005 to August 2006). If it was not known which type of diesel was sold, it was assumed that the outlet only sold LSD, but the station was flagged and not used in computing the proportions for allocating weights for outlets that sold both types.

Indexing station i in cell j by ij where i= 1 to n_j (the number of stations selected for cell j), the following terms were defined:

u_{ij} =1 if the station sells ULSD; 0 otherwise.
v_{ij} =1 if the station sells LSD; 0 otherwise.
f_{ij} =1 if the station sells only one type; 0 otherwise.
x_{ij} = the price of ULSD in station ij; 0 if it sells no ULSD.
y_{ij} = the price of LSD in station ij; 0 if it sells no LSD

Let $q_j = \Sigma\ (u_{ij}\ f_{ij}) / \Sigma\ f_{ij,}$ where q_j = the imputed proportion ultra low diesel sales for stations that claim to sell both types of diesel under the first base alternative.

Now defining $u'_{ij} = u_{ij}$ if the station does not sell both types of diesel, and $u'_{ij} = q_j$ if it sells both, and likewise defining $v'_{ij} = v_{ij}$ if the station does sell both and v'_{ij} =1- q_j if it does, the algorithm to impute a station's proportion of ULSD sales is complete.

Note that the algorithm is designed to impute what proportion of a station's sales is ULSD assuming that it will be the same as the proportion of the cell's sales. However, there are two situations where this algorithm could be problematic. First, it is possible that every station in a cell sold both LSD and ULSD. In that case, one can simply let q_j =.5. Second, it is possible that in some cell all the stations that sell LSD sell both (or all that sell ULSD sell both). In those cases the algorithm would lead to information contrary to what the station has reported, since the proportion of stations selling only one type that sell LSD would be 0, but the station would have specifically reported selling both. One solution to this problem would be to force q_j to be between .1 and .9 (for example).

Another solution would be to set u'_{ij} and v'_{ij} automatically to .5 when the outlet reports both. This alternative was the solution chosen. For estimation, note that $u'_{ij} + v'_{ij}$ =1 for all outlets reporting, preserving the equal outlet weighting for combined diesel. Cell level volume, prices and revenues for ULSD are computed as:

$$(VU)_j = W_j \Sigma\ (u'_{ij}\ ) / \Sigma\ (u'_{ij} + v'_{ij})$$
$$(PU)_j = \Sigma\ (u'_{ij}\ x_{ij}\ ) / \Sigma\ (u'_{ij})$$
$$(RU)_j = (PU)_j * (VU)_j$$

where, W_j is the total volume of diesel sales for cell j from the monthly survey, $(VU)_j$ is the volume for ULSD for cell j, $(PU)_j$ is the price for ULSD for cell j, and $(RU)_j$ is the revenue for ULSD for cell j. $(VL)_{j,}$ $(PL)_j$, and $(RL)_j$ are defined analogously for LSD.

The combined diesel price at a station is then defined by $z_{ij}=u'_{ij}x_{ij} + v'_{ij}y_{ij.}$ It is easy to see that if one takes the average of the z_{ij} and multiplies it by W_j one gets $(RU)_j$+ $(RL)_j$. Thus, one can calculate z_{ij} for each outlet and get the average for the cell , or calculate the ratio of the sum of the ULSD and LSD revenues to the sum of the ULSD and LSD volumes, i.e.,

$(PC)_j = [(RU)_j+(RL)_j] / [(VU)_j+(VL)_j]$ for the same result.

The separate volumes and revenues for the two kinds of diesel are then used to form national estimates or estimates for PADDs that are composed of sampling cells. The noteworthy point in estimation is that the entire estimate depends on the data as it is reported each week. The proportion of stations by diesel type will change from week to week as the transition is implemented.

2.3 The Impact of Alternative Assumptions

Starting with the January 16, 2007 collection, prices were collected for ULSD and LSD separately. However, for the first three weeks of collection, only the combined diesel prices were released. During those three weeks, attempts were made to resolve as many "don't knows" as possible, the number of outlets selling each type of fuel was examined, and the quality of the data was monitored.

One of the issues examined through a sensitivity analysis was the degree to which the assumption that if a station sells both types of diesel, it sells 50% of each. To examine this, the price effect for one week was compared under three scenarios: 50% sold for each type of diesel, 90% ULSD/10% LSD, and lastly, 90% LSD/10% ULSD sold. As it turned out, there were not that many outlets that reported selling both types of diesel, so it was not considered likely that there would be a large difference in price under the three assumptions, but some differences in the estimates did arise. Table 1 contains the results for one week. From this it can be seen that the combined diesel average price did not change under the alternative scenarios for any area, except for a change of .1 cents in Padd 1C. The price for LSD changed by .7 cents in one direction and .5 cents in the other at the U.S. level, and as much as 2.1 cents in one direction and 1.0 cents in the other for PADD 3. The national price for ULSD, in comparison, changed by .3 cents in one direction and .4 cents in the other. PADDs 2 and 4 had the largest changes at .6 cents in one direction and .7 cents in the other. The conclusion was, therefore, that the assumption

Table 1. Effect on Average Prices of Different Assumptions

	Combined Diesel			LSD			ULSD		
	Price using .5	**Change using .9**	**Change using .1**	**Price using .5**	**Change using .9**	**Change using .1**	**Price using .5**	**Change using .9**	**Change using .1**
U.S.	$2.849	$0.000	$0.000	$2.809	$0.007	-$0.005	$2.859	-$0.003	$0.004
PADD 1	$2.853	$0.000	$0.000	$2.818	$0.000	$0.000	$2.866	-$0.002	$0.001
PADD 1A	$2.942	$0.000	$0.000	---------	-------	--------	$2.942	$0.000	$0.000
PADD 1B	$2.936	$0.000	$0.000	$2.963	-$0.013	$0.008	$2.933	$0.002	-$0.002
PADD 1C	$2.810	-$0.001	$0.000	$2.798	$0.004	-$0.003	$2.815	-$0.002	$0.003
PADD 2	$2.822	$0.000	$0.000	$2.796	$0.011	-$0.008	$2.832	-$0.006	$0.007
PADD 3	$2.785	$0.000	$0.000	$2.747	$0.021	-$0.010	$2.787	-$0.002	$0.002
PADD 4	$2.955	$0.000	$0.000	$2.862	$0.000	$0.000	$2.962	-$0.006	$0.007
PADD 5	$2.987	$0.000	$0.000	$2.904	$0.002	-$0.001	$2.996	-$0.002	$0.002
PADD 5 (w/o CA)	$2.918	$0.000	$0.000	$2.904	$0.002	-$0.001	$2.921	-$0.001	$0.001
CA	$3.090	$0.000	$0.000	---------	--------	--------	$3.090	$0.000	$0.000

that had been made for outlets that report both types of diesel setting u'_{ij} and v'_{ij} automatically to .5 results in a slight difference in the estimates, but the scenarios tested constituted extreme cases. In addition, these differences should be viewed with respect to the Coefficients of Variation for each of the cells. It should also be noted that the proportions of LSD and ULSD sold by an outlet depends on both the outlet's supplier and the demand by the consumers. A more extreme effect, however, could be expected if the assumption regarding outlets that report not knowing (currently assumed to be selling low) was incorrect.

2.4 Coefficients of Variation

A bootstrap approach was used to obtain price variances for each diesel type and the combined diesel, and for each publication cell. If a station had a weight of 2 in the original sample the company had been sampled twice at the second phase, but they had only one station in that state, so the outlet was counted twice. The bootstrap treated the station as if it were two different stations. Table 2 shows the Coefficient of Variation (CV) for each average price targeted for release. Every CV is under one percent for every publication cell for the combined price, with the national CV being 0.24% and the highest CV occurring in PADD 1A at 0.61%. For ULSD, the national CV is 0.25% and the highest occurred in PADD1B at only .63%, still below 1%. The CVs for LSD were much larger, with the national CV being 0.45% and the highest being PADD 4 at 2.99%. As the sales of LSD decrease, we expect the CVs for ULSD to approach the CVs for combined diesel, and the CVs for LSD to increase.

Table 2. CVs by Diesel Type (%)

	ULSD	**LSD**	**Combined Diesel**
U.S.	0.25	0.45	0.24
PADD 1	0.33	0.68	0.29
PADD 1A	0.61	---------	0.61
PADD 1B	0.63	1.40	0.58
PADD 1C	0.40	0.69	0.37
PADD 2	0.60	0.72	0.50
PADD 3	0.52	0.84	0.50
PADD 4	0.39	2.99	0.40
PADD 5	0.38	1.06	0.35
PADD 5 (less CA)	0.56	1.06	0.50
CA	0.42	---------	0.42

2.5 Implementation Issues

The initiation of the new survey and the implementation of the aggregate price calculations resulted in some practical issues that needed to be resolved prior to the first week of data collection. It was intended that the type of diesel being sold for each station, either LSD or ULSD would be ascertained on the first day of data collection during the Computer Assisted Telephone Interview (CATI). However, many of the station's prices are obtained via fax, e-mail, or from the company's web site; therefore, these respondents are not routinely contacted directly by phone during the weekly data collection. Thus, it was necessary to identify the appropriate corporate contact to provide this information prior to the

first day of data collection. Over a short period of time prior to the first Monday of data collection, corporate contacts were called and asked to provide the type of fuel sold for each of their reporting locations. From these non-CATI calls it was determined that: 126 locations sold ULSD only, 71 locations sold LSD only, and 23 stations sold both types.

The second issue for implementation was regarding nonresponse imputation. Aggregate regional prices are calculated using both reported prices and prices imputed for nonrepsonse. Imputation for the diesel survey consists of applying the calculated aggregate regional price change from the prior week to the current week to an individual station's reported price for the prior week. For the first week of data collection, the type of fuel being sold in the prior week was only known for those stations with prices obtained via fax, e-mail or their web site. For imputation purposes, it was assumed that the prior week's fuel type was the same as the current week for all remaining stations where the fuel type was not known. For stations reporting sales of both ULSD and LSD, the prior week's price was used for imputation for both types of fuel. Similarly, if a station switched during the course of the survey from selling only one type of diesel, to selling both, the prior week's price was used for imputation for both fuel types.

In addition to the assumptions made at the station-level for imputation, it was also necessary to derive the aggregate price change to apply to the individual station level prices for the first week of data collection. To determine the aggregate price change for each type of fuel, only those stations with known fuel types contributed to the prior week's ULSD and LSD aggregate regional price while all stations with reported prices contributed to the current week's ULSD and LSD aggregate regional prices.

3. Measures of Data Quality and Establishing Criteria for Data Release

In addition to timeliness, data quality for the weekly diesel survey is measured in terms of response rates and the accuracy of reported prices. Response rates are tracked and monitored throughout the data collection process each week with target levels set in the 98 to 100 percent range. Because one contact may report for more than one outlet, response rates are monitored at the contact level and at the outlet level. Prices for non-reporting outlets at survey closure are imputed unless it is known that the station is shut down.

Accuracy of the reported prices is checked throughout the data collection process. A system of price edits and rechecks are built into the CATI interview as well as during the processing of the data. Historical price and range checks are conducted during the CATI interview at the individual outlet level. A second set of outlet-level edit checks is conducted during survey processing post collection and an attempt is made to re-contact the respondent to verify any prices that fail the processing edits.

In addition to the outlet level edit checks, the accuracy of reported prices is also monitored at the aggregate level. Standard errors and coefficients of variation (CV) are computed and published each week along with aggregate prices. The original survey sample was designed to meet target CV levels of one percent for published aggregate prices.

With the need to publish prices for both ULSD and LSD, it became necessary to monitor response rates and CVs for each fuel type separately, as well as overall. The split between ULSD and LSD effectively reduced sample sizes for each fuel type, potentially yielding higher CVs. Beginning with the first week of data collection, survey response rates and CVs for ULSD and LSD were calculated separately and used to evaluate the impact of the reduced sample size on price variability by region.

Both the number of outlets reporting for each fuel type and the resulting CV for the aggregate prices were used as criteria for determining whether an aggregate price for each fuel type should be released. These measures were monitored for several weeks to allow stabilization before the decision was made to release separate ULSD and LSD prices. The minimum acceptable criteria for release of aggregate prices by fuel type were established as follows: 1) a minimum of 10 outlets must contribute to the aggregate price, and 2) the CV for the aggregate price must be lower than two percent. All aggregate prices for ULSD met these criteria, but only the U.S. and PADD 1 prices met the criteria for LSD consistently for those weeks. In addition, the softer criteria of how long an aggregate could be published, as well as potential inaccuracy not measured by the CVs (incorrectly classified don't knows and the impact of the assumption used for two type sellers) were considered. In particular, aggregates with CVs that were growing quickly and/or the number of outlets were quickly decreasing, and aggregates with a high number of don't knows were further examined from the customers' perspective. As a result, all ULSD prices were deemed releasable, but only the U.S. and PADD 1 prices for LSD were releasable. The survey results for LSD are continually monitored each week relative to these data quality standards for continued release of the US and PADD 1 prices. At some point during the completion of the transition to ULSD, LSD prices will no longer be releasable from a quality viewpoint.

4. Summary and Future Work

In measuring the price impact of the transition to ULSD, the sample design and sample size was not changed. It was not known prior to collection of prices by type, what the effect would be on the variances, and the resulting sample sizes that would be required to produce a 1% CV for all areas and types. Similarly, it was not known which outlets would be selling which products at which point in time. Given this, it was not considered to be cost effective to either re-design the sample or increase the sample sizes. The resulting CVs show that LSD sample sizes would have required a large increase, particularly over an extensive part of the transition. The use of the reported fuel type by the outlets to drive the allocation of the volume weight for aggregating sampling cells, and the allocation of the outlet weight for outlets selling two types of diesel equally to the type of diesel, allowed the historic sample weights to be preserved for historic data continuity. Sensitivity testing of the equal allocation for those outlets selling both types showed the impact to be slight, particularly for those areas and types for which the prices satisfied the quality criteria.

Future work is expected to center on the construction of an outlet level sampling frame, and sample design that takes full advantage of the outlet level frame. This work is expected to be independent of the transition to ULSD.

5. Bibliography

Weir, P. and Saavedra, P. (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.

EIA diesel price web site,
http://tonto.eia.doe.gov/oog/info/wohdp/diesel.asp