Improvement of Data Quality Assurance in the EIA Weekly Gasoline Prices Survey

Bin Zhang, Paula Mason, Amerine Woodyard, and Benita O'Colmain
Bin Zhang, EIA, 1000 Independence Ave., SW, Washington, DC 20585
Paula Mason, EIA, 1000 Independence Ave., SW, Washington, DC 20585
Amerine Woodyard, EIA, 1000 Independence Ave., SW, Washington, DC 20585
Benita O'Colmain, ICF Macro, 11785 Beltsville Dr., Calverton, MD 20705

Abstract
The EIA weekly survey of retail gasoline prices collects prices using a Computer Assisted Telephone Interview system with interactive data editing embedded to assure data quality. Edit performance statistics, however, showed that the data editing criterion sometimes missed true outliers at one end of the price change distribution and falsely over flagged outliers at the other end of the distribution, especially during times of large price change seen in the last three years. In order to improve the efficiency of the data editing criterion, a new data editing criterion based on price change relative to market change was developed. In addition, a new post-collection data validation procedure for screening price and price change outliers by region and grade that makes use of all respondents collected was also implemented in the survey process to further assure data quality.

Keywords: edit, validation

1. 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. Specifically, retail gasoline pump prices were needed on a weekly basis (or more often) for regular gasoline at the national level that were not only accurate, but could be obtained quickly and inexpensively. The Weekly Motor Gasoline Price Survey (EIA-878) was undertaken to meet this need. 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 of the survey was to collect, process, and release the data to a variety of users, including policy makers and citizens, in a very rapid turnaround mode. Specifically, Monday morning's prices were to be available by the end of the same day. The survey was later iteratively expanded in response to the Clean Air Act and eventually provided estimates for two types of formulation (conventional and reformulated), two additional grades, midgrade and premium (Weir, O'Colmain, and Saavedra, 2007) for the five geographic regions known as Petroleum Administration for Defense Districts (PADDs), three sub-regions (sub-PADDs), nine selected States, and ten selected cities. The selection of those States and cities was specifically intended to provide estimates for at least one State and one city in each PADD and sub-PADD for which gasoline prices are published.

The weekly motor gasoline survey collects retail gasoline prices at the pump from a sample of 800 gasoline stations nationwide. Price collection begins each Monday morning at 9:00 a.m. and is completed by 3:00 p.m. The data are processed and aggregated and released to the public by 5:00 p.m. through the EIA website, email notification, and a telephone hotline. The majority of respondents for the survey are contacted directly via Computer Assisted Telephone Interview (CATI), while other respondents have elected to fax or email their prices to the data collection center. In addition, data collectors download prices from company websites directly for respondents posting prices on their websites regularly. One respondent may report for only one or multiple stations, depending on the size and geographical coverage of the company represented and the reporting preference of the company.

Before the implementation of the improvement described in this paper, quality of the survey data was mainly controlled first through an interactive edit procedure that was embedded in the CATI system. Price change outliers and extreme prices were flagged for verification and investigation. Flagged prices that could not be verified by data interviewers were then followed up in a re-check study by a data collection manager. The manager investigated the data and might contact respondents if necessary. In addition, data quality measures, coefficients of variation, and operations performance measures enabled us to review and monitor data quality and accuracy, and survey execution performance as well.

2. Data Editing

Data editing and data rechecking that further investigates data with unresolved data editing flags are used to assure the quality of the reported data and the final prices that EIA publishes. The data editing procedure, an interactive tool embedded in the CATI system, flags abnormal prices based on a set of predefined criteria during the telephone interview. The criteria in place when this research was conducted were independent of energy market situations and did not change from week to week. The logistics of the data editing criteria were:

- If a price changed more than 3% from the previous week, the respondent was asked to confirm the reported price;
- If a price changed more than 5% from the previous week, the respondent was asked to provide an explanation for the large price change;
- If a price changed more than 12% from the previous week, it was flagged and sent to the recheck study;
- If a price was below $0.50 or over $5.00, or changed more than 24% from the previous week, it was replaced with imputed values automatically during the imputation stage;

All prices with unresolved editing flags were examined in a recheck study report twice, once in the middle of data collection and once after data collection was complete. The data collection manager reviewed, investigated, and contacted respondents, if necessary, to ensure the prices were correct. Editing flags were overridden if the prices were corrected or confirmed by the data collection manager. Non-respondents and those prices with unresolved editing flags from the recheck study were replaced with imputed values in the imputation stage of data processing. Data collected by fax, email, and internet were

entered into the CATI system and, therefore, also went through the same editing and rechecking.

There were only a few prices that failed the last two edits, which were basically for screening extreme values. We also found that most explanations collected for prices flagged by the 5% edit were not very informative. Therefore we decided to drop that edit when we improved the first, the 3% edit.

The 3% edit, which was the number one cause that data received an editing flag, was a range check. Price changes outside of range check intervals were flagged as outliers. The range check interval of each price was an interval with zero as the center and the width of the interval was based on previous week's reported price. Three percent above and three percent below the prices reported in previous week formed the interval. The centers of all intervals remained fixed each week no matter how much the retail gasoline market changed from previous week. As a result, it sometimes caused data quality control problem for missing true outliers and survey execution problem for over flagging data.

Figure 1. Percentages of Data That Changed More Than 3%

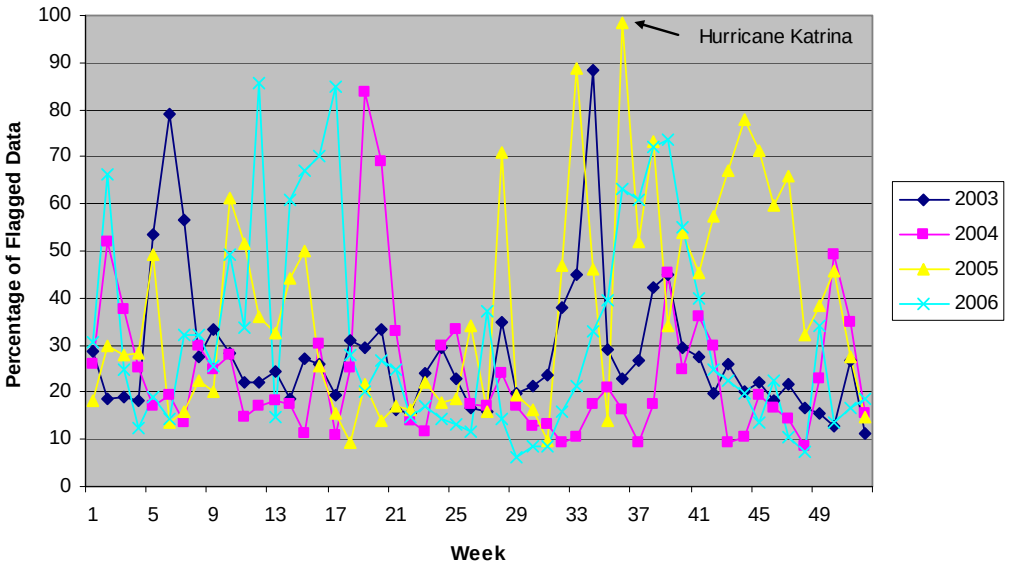
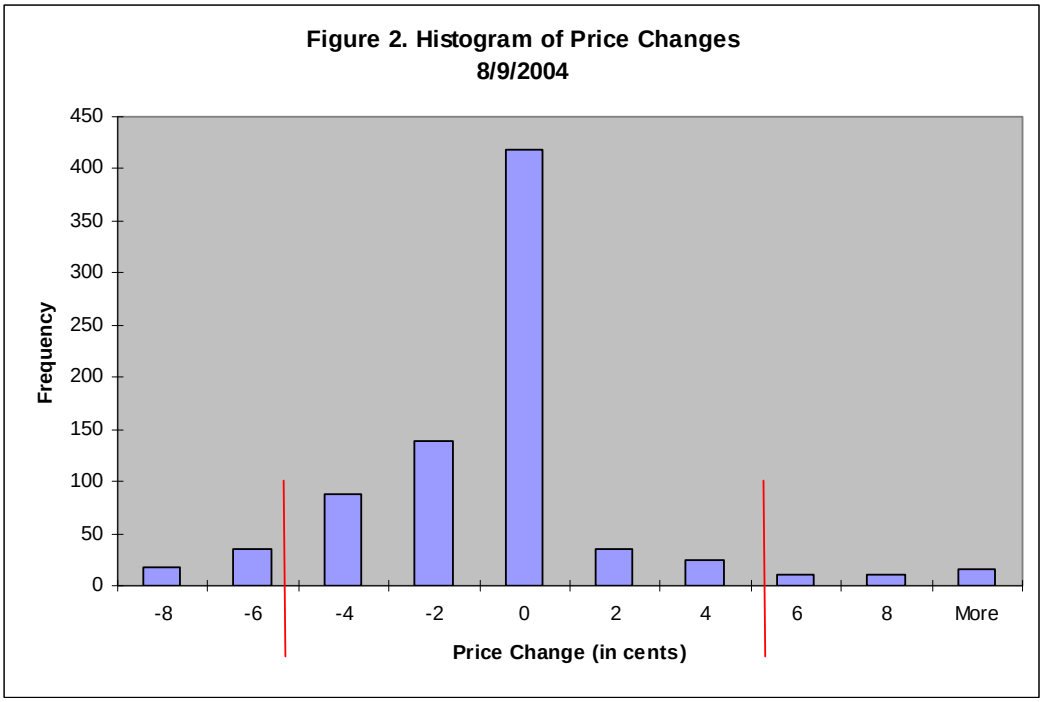
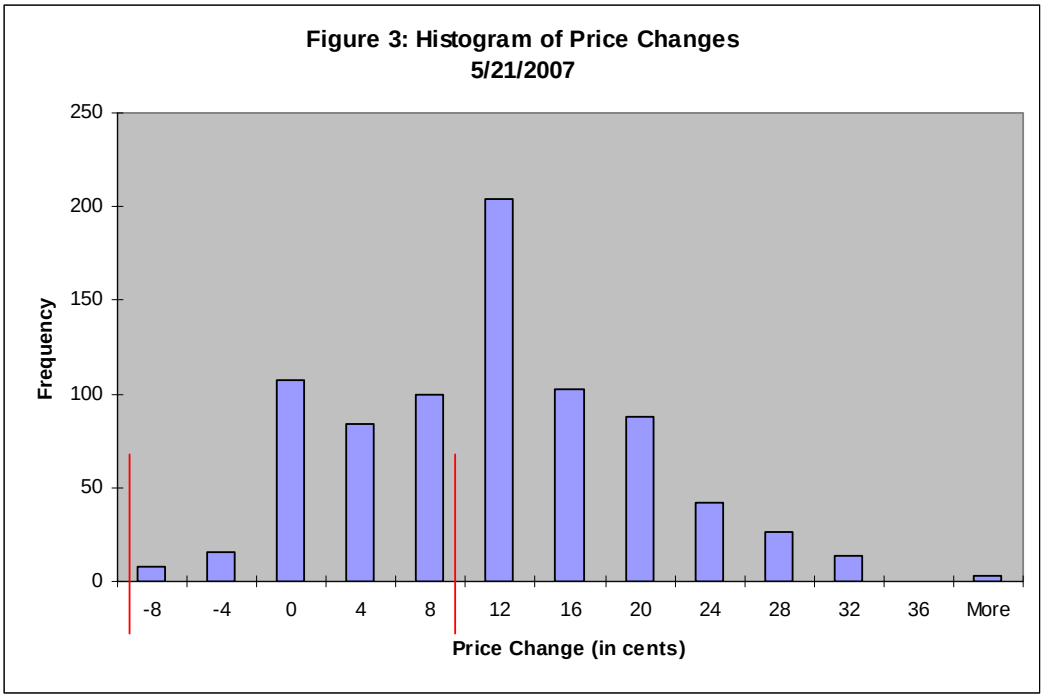


Figure 1 above shows the percentages, by year, of data that changed more than 3% from previous week. In 2003 and 2004, the percentages were below 30% in most weeks, and spikes of over 40% were very rare. In 2005 and 2006, the average percentages were higher than those in 2003 and 2004, and percentages of higher than 40% became quite common. In the week of Hurricane Katrina almost all the prices changed more than 3% from the previous week. In spite of the high percentages of data that failed the 3% edit, the amount of data that needed correction remained at the same level as that of 2003 and 2004. Data quality measures, survey performance measures, and feedbacks from data interviewers all indicated that the 3% edit did not work effectively or efficiently in 2005 and 2006.

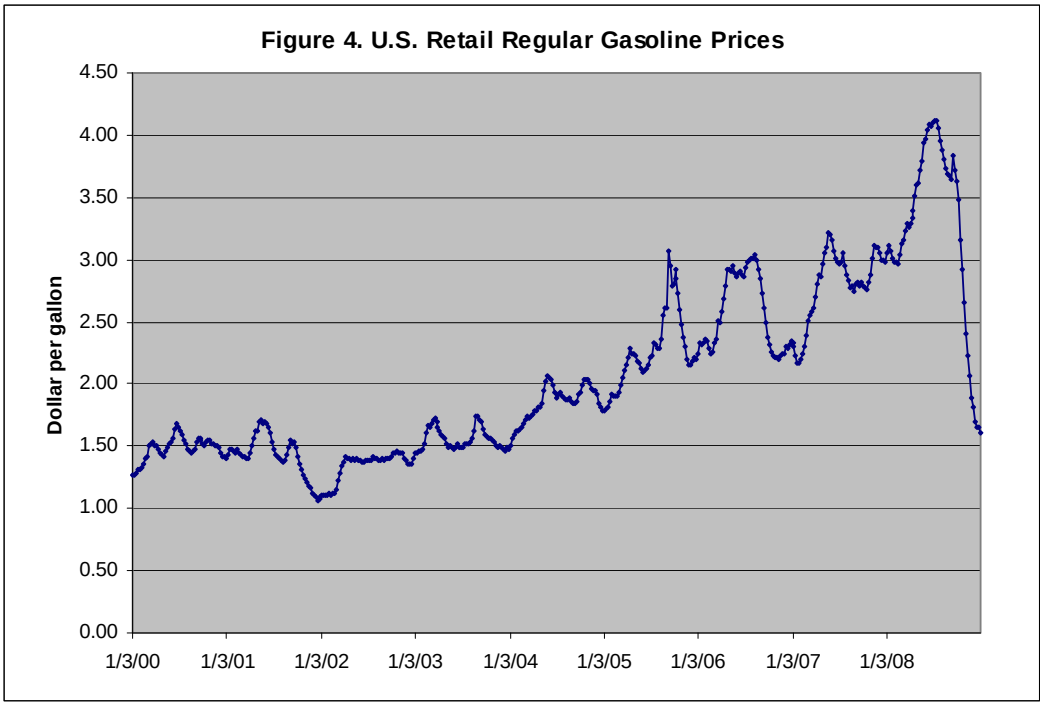
The 3% edit, which flagged prices that changed over 3% from previous week, used intervals, with zero as the centers and 6% of previous week's prices as the widths, to screen price change outliers. It was designed in the early years of the survey when prices were at a relatively low level and price changes from week to week were very small. Under that kind of market condition, the distribution of price change in a particular week was approximately a normal distribution, with mean near zero. The two tails of the distribution was roughly symmetric around zero. Consequently, the price change outliers at the two ends of the distribution were flagged by the 3% edit as outliers. For example, on Monday August 9, 2004, the national average price for regular gasoline decreased 1.1 cents from the previous week's price of $1.888 to $1.877. The center of the distribution of price change was near zero. The 3% data editing edit flagged all prices that either increased or decreased 3% from previous week's reported prices. Those flagged prices, which accounted for 9% of all collected data, were the true price change outliers represented by the two tails shown in figure 2 below.



In a different scenario shown in the data for May 21, 2007 (see figure 3 below), the average price increased a large amount of 11.5 cents from $3.103 to $3.218. Even though the price change distribution was still an approximately normal distribution, the mean of the distributions was no longer near zero. When we still used intervals with zero as fixed centers of range check intervals to screen outliers, it ended up that 54% of the data were flagged by the edit. In fact, many data with average price increases were flagged as outliers, while some true outliers, prices that decreased from previous week, were missed by the edit.

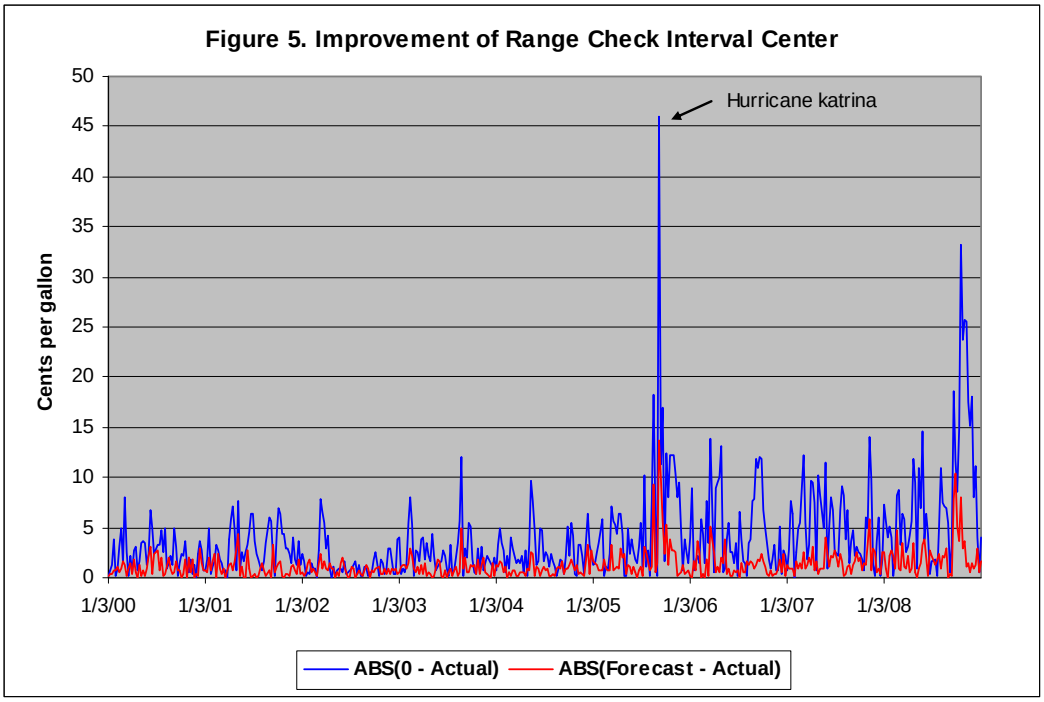


The locations of the centers of the range check intervals were very important to the effectiveness and efficiency of the 3% data edit. In general, it was preferred to have range check interval centers close to the means of actual price changes. As shown in figure 4, before 2005, week to week price changes were small; hence, the centers of the range check intervals, which were always zero, were close to the actual means of the price change distributions. As a result, the data editing rule worked well in that relatively slow moving market. After entering 2005, prices moved up or down much steeper than the changes in earlier years. For many weeks, the week to week average price changes, which were the means of the price change distributions, were far away from the centers of the range check intervals. As a result, the data editing rule had poor effectiveness and efficiency during that fast moving market.



To improve the edit, we first decided to use price change forecasts as centers for the range check intervals instead of a fixed zero. At EIA, we have been using historical series of wholesale price changes to predict retail price changes each week. The forecasts are used to keep management and analysts informed of what to expect before survey estimates become available. The forecasts are available at national and regional levels. A set of forecasts are usually generated on each Friday by using spot prices up to Thursday as what we call pre-weekend heads-up information. The forecasts are then refreshed on the following Monday morning with spot prices up to Friday. Due to the time constraint on survey processing, the set of regional forecasts generated on Friday are used as centers of the new data edit.

The blue line in figure 5 represents the distances between zero, the centers of the range check intervals under the 3% edit, and actual mean price changes. The red line represents the distances between the forecasts, the centers under the new editing criterion, and the actual mean price changes. The average price change forecasts are much closer to the actual average price changes than the fixed centers of zero.



Previously, the width of the range check interval for a price was 6% of its corresponding price of the prior week. Since the range check edit was for screening price change outliers, we found that for setting the width of the rule, it would be better to use price change variation information instead of using the magnitude of previous week prices, even though the two show some correlation. We used the average standard deviations of price changes by region over the period of 2004 through April 2007 to set the widths of the range check intervals under the new edit. It was broken down by region because the average price change standard deviations across all the regions were not at the same level. In particular, the average price change standard deviation for the Midwest is significantly higher than that of other regions.

A new dynamic edit was then finalized such that a price is flagged for verification if its change from the previous week differs from the forecasted regional average change by more than twice the regional average standard deviation of price change from January 2004 through April 2007. The center of a new range check interval is the forecasted price change for a region, and the width of the interval is four times the average standard deviation of price changes for the region.

With the redesign of the data edit, the average percentage of flagged data is reduced to about 13% from about 30% under the 3% edit since the implementation of the new dynamic edit in July 2007 (see figure 6). It is fairly stable and has few big spikes. As a result of the new edit, data quality has been improved because more true outliers are flagged and collection cost has been reduced through the improvement of efficiency as well.

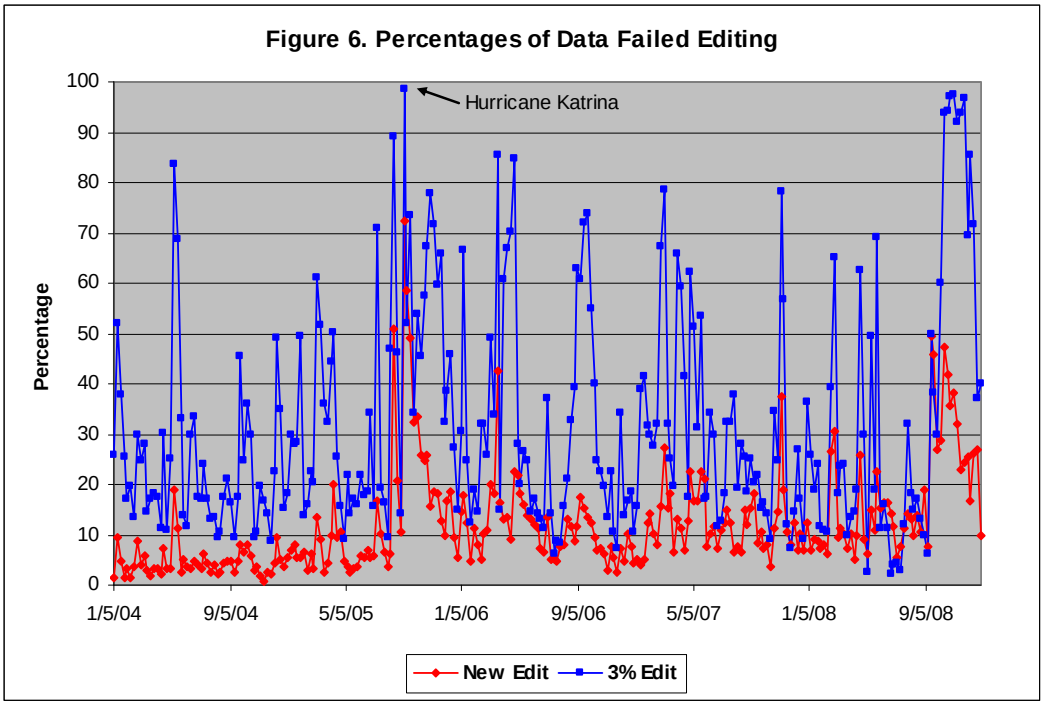


Table 1 below also shows that the average percentages of flagged data are fairly balanced, as desired, among the regions.

Table 1. Average Percentages of Flagged Data under the New Dynamic Edit

Region	July 2007 – December 2008
New England	14.8
Central Atlantic	11.5
Lower Atlantic	14.2
Midwest	14.4
Gulf Coast	16.5
Rocky Mountain	12.9
California	13.1
West Coast less CA	10.0

3. Data Validation

Once data collection is completed after the data recheck procedure, data processing moves on to the imputation stages. The imputation program replaces all non-responses, prices with unresolved editing flags, and prices flagged by the 24% edit with imputed values. An imputed value is calculated by the previous week's value of that respondent adjusted by the average price change of the sampling stratum associated with that respondent. All prices, reported and imputed, are then aggregated to generate the average prices.

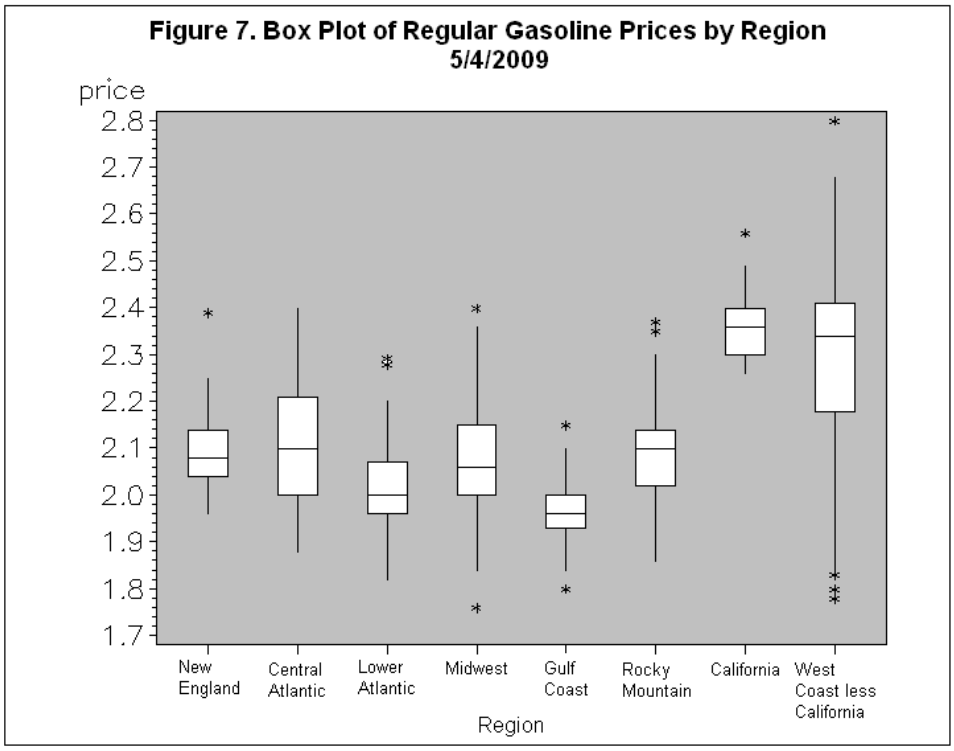
All the data edits during data collection are individual data item based without any involvement of other reported prices of the same week. We developed a post data collection procedure that utilizes box plots and listings for further data validation. Box

plots by grade and region are used to screen price and price change outliers. Listings are used to check those respondents with a price for a lower grade of gasoline being greater than a price for a higher grade of gasoline. Respondents that report the same prices over an extended period are also listed.

Even though the validation procedure and the data editing procedure use different methods to screen questionable reported data, they are not completely independent. With accurate forecasts of price changes as centers of data editing ranges, the data editing procedure should be able to screen most price changes and price outliers. The validation procedure mainly serves as a complementary data assurance procedure. Due to the constraints of processing time, this procedure is executed after the aggregation stage, rather than right after the data recheck study. The validation procedure also includes a feature that estimates the impact of questionable prices identified by this procedure on the aggregates. This feature replaces the prices in question with imputed values, calculates a new set of aggregate prices, and then compares them with the set of aggregate prices that were generated earlier in the aggregation stage of data processing. If the impact exceeds the survey's reprocessing thresholds, the identified prices are further investigated, and the data are re-tabulated using either corrected prices or imputed prices, if necessary. If the impact of the identified prices on aggregate prices is less than the established boundaries for reprocessing, the data are released for publication and the identified prices are investigated afterward for future data quality control. The reprocessing boundaries are:

- Impact greater than or equal to 0.5% at the national level, or
- Impact greater than or equal to 1% at the PADD, State, or city level.

The box plot (figure 7) below is an example generated by the validation program. Many data were flagged as price outliers, but none of them had impact on aggregates that exceeded the reprocessing boundaries. In most situations, the flagged outliers do not have significant impact on the aggregates due to the relative large sample size.



4. Summary

With the improved data edit, the re-check study, and the new validation procedure altogether as data quality assurance, the chance of unverified outliers being included in aggregation for final published data is very slim. It has also made data collection and processing smoother and more efficient. In the future, we should periodically monitor the accuracy of the price change forecasts, follow up the regional standard deviations of price changes, check the percentages of flagged data by region, and adjust the edit criterion accordingly to ensure effective and efficient data editing.

References

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.

P. Weir, B. O'Colmain, and Saavedra, P. J. 2007. "The Evolution of the Weekly Gasoline Price Survey through Changes in Design and Frame"