Statistical Challenges in Linking a Retail Gasoline Price Survey with Commercial Data

Maura Bardos, Amerine Woodyard, and Jeramiah Yeksavich
U.S. Energy Information Administration, 1000 Independence Ave., SW, Washington, DC 20585

Abstract[1]

As part of ongoing modernization efforts, the U.S. Energy Information Administration (EIA) is conducting research on utilizing third-party sources to supplement survey data. EIA is uniquely situated since a number of its surveys collect information that is also compiled and sold by commercial vendors. These commercial vendors can provide almost real-time frequency of data that when linked with surveys, have the potential to reduce respondent burden and enhance data products. However, there are statistical challenges including record linkage and evaluation of potential sources of error in commercial sources such as coverage error, specification error, measurement error, and missing data. EIA purchased price data from a commercial vendor for about 110,000 retail gas stations and also created a tool to obtain gas prices via a crowdsourced website. This paper examines the challenges in integrating data from these commercial sources with data collected from the Motor Gasoline Price Survey (EIA-878). EIA conducts this weekly mandatory survey to produce point-in-time gasoline price estimates by geographic area, grade, and formulation from a sample of retail gas stations.

Key words: Record linkage; Commercial Data; GIS; Administrative Records; Establishment Survey; Total Survey Error

1. Introduction

Research on alternative sources of retail gasoline prices was launched in parallel with an overall redesign of the Motor Gasoline Price Survey (EIA-878). As a principal federal statistical agency, EIA holds a responsibility to investigate alternative sources of data and justify the necessity of survey data collection efforts as a part of the Office of Management and Budget clearance process. This research is also part of a broader effort to evaluate surveys across EIA outlined in the Statistical Methodology Improvement Plan (SMIP). The SMIP is a five-year plan to improve the statistical quality of EIA surveys and products through the evaluation and application of rigorous statistical methods.

2. Data and Methods

2.1 EIA-878

The Motor Gasoline Price Survey (EIA-878) is a weekly mandatory survey of a sample of approximately 1,000 retail gasoline stations across the country. The data collected are used to create point-in-time estimates of gasoline prices at the national, regional, and selected state and city levels by grade and formulation, resulting in 276 published price estimates.

[1] The analysis and conclusions contained in this paper are those of the authors and do not represent the official position of the U.S. Energy Information Administration or the U.S. Department of Energy.

Data collection, processing, and dissemination are completed within the same day. EIA defines gasoline price as the station’s pump price (including taxes) as of 8:00 a.m. local time each Monday. This price represents the self-serve price except in areas having only full-serve and the cash price except for stations that only accept credit cards. The majority of respondents comply with the mandatory survey via telephone reporting (CATI, or computer-assisted telephone interviewing); however, other submission methods are also available. Prices are published around 5:00 p.m. ET on Monday, except in the case of government holidays.[2]

The motor gasoline price estimates are published along with on-highway diesel price estimates each Monday, and the data product consistently remains one of the top viewed items on EIA’s website, receiving 3.4 million visits in 2017 and over 3.6 million visits in 2016.

2.2 Oil Price Information Service

The Oil Price Information Service (OPIS) is a commercial source for petroleum pricing and news information. According to company materials, the company collects daily gasoline and diesel prices for nearly 140,000 retail outlets in the United States and Canada (OPIS 2017). OPIS prices are used by a variety of companies including AAA, Google Maps, and MapQuest (OPIS 2017). In March 2013, OPIS acquired GasBuddy.com, a crowdsourcing website which publishes retail motor gasoline and diesel prices (Abcede 2013).

During selected time periods, EIA purchased weekly feeds of gasoline price data from OPIS.[3] Each feed included the station name, physical location, and a unique OPIS identification number. For each gasoline grade (regular, midgrade, and premium), the data file contained the price per gallon, date and time associated with the price, and the source of the information.

OPIS provided a single value for the price without making a distinction between cash or credit prices. OPIS collected the data through either electronic submission directly from credit card transactions or from user submissions. On average, each file consisted of about 110,000 records; however, the exact number varied from feed to feed and from week to week. EIA obtained four feeds with data at 8:00 a.m., 12:00 p.m., 4:00 p.m., and 8:00 p.m. The multiple feeds were used to compare the internal consistency of the data and gain insights into OPIS’ data quality.

2.3 GasBuddy.com

GasBuddy.com is a crowdsourced website that allows users to submit real-time regular, midgrade, premium, and diesel prices. Based on EIA analysis, the site collects prices for about 144,000 outlets in the United States. Users can submit both cash and credit prices to GasBuddy and the information is displayed on the website, tagged by the user’s ID. Some stations also submit price information directly to GasBuddy; these prices are denoted as “GB_Direct” (GasBuddy, “Who/What is GB_Direct”). GasBuddy provides users with definitions for regular, midgrade, and premium, as well as how to capture information on price discounts. In addition to submitting prices, users can also add gas stations to the GasBuddy master list. Users earn points for completing certain activities, such as posting

[2] A more detailed description of the Motor Gasoline Price Survey methodology may be found at https://www.eia.gov/petroleum/gasdiesel/gas_proc-methods.php

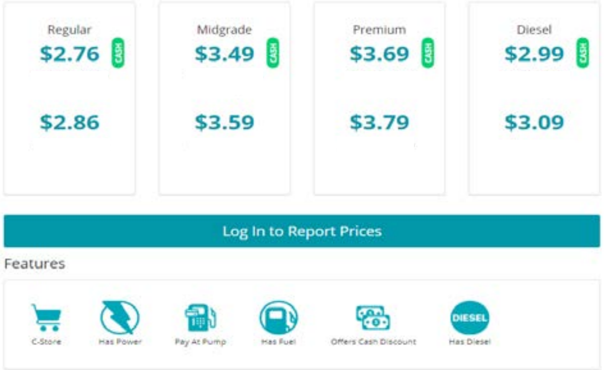
[3] Mondays from May 15, 2017 to July 31, 2017 and May 14, 2018 to May 21, 2018.

or updating a price or participating in a user forum. Points can be redeemed in prize raffles.

GasBuddy makes data for each station available from an associated unique webpage, and we developed a tool to systematically collect price information for stations included in the EIA-878 sample (see Figure 1 for an example). The data set obtained from the tool contained information on the station name and physical location. For each gasoline grade (regular, midgrade, and premium), we collected the price per gallon (cash and credit), date and time submitted, source (user submitted or direct data transmission from the outlet), and gas station features (e.g., open 24/7, has a convenience store).

Gas prices submitted to the website within the last 24 hours include additional information on the exact time of the submission. In contrast, prices reported outside of 24 hours are categorized merely as "1 day old" without additional information about exact time of submission.

Similar to OPIS, we obtained data only during selected time periods to support internal production processes. We did not attempt to obtain gas prices for all stations on GasBuddy.com.



Regular	Midgrade	Premium	Diesel
$2.76 CASH	$3.49 CASH	$3.69 CASH	$2.99 CASH
$2.86	$3.59	$3.79	$3.09

Figure 1: Screenshot of GasBuddy.com website for an individual gas station.

2.4 Linking Methods

A method for linking records between the EIA gas station sample and the OPIS and GasBuddy data sets was needed before any subsequent analyses could proceed. The biggest challenge in linking the three sets of data was the absence of a common primary identification key. The only commonly shared elements were the station names and addresses. We decided not to attempt linking datasets using station names due to differences in station naming conventions, alternative spellings, and station name changes that have occurred since the EIA-878 sample was drawn. We determined address matching would produce better results than simply linking on names. Based on the limited information available to link datasets, we employed a geographic information systems (GIS) approach, specifically nearest neighbor analysis on geolocated stations (illustrated in Figure 2), as a way to link the data sources and avoid the issues described above. Nearest neighbor analysis is a process through which records from separate databases are joined based on the geographical proximity to each other. In short, addresses from two databases are mapped as points and the closet two records from the respective databases are joined.

We geocoded the EIA-878 sample to generate spatial data for station locations. Most of the entries in the EIA-878 sample geocoded on the first attempt with the remainder needing correction in address fields (e.g., PO Boxes preceding street numbers, alternative spellings, etc.). Six stations were not able to be geocoded due to bad or unclear addresses. The OPIS and GasBuddy databases contained existing coordinates, and therefore, they did not need geocoding. For the initial round, we performed nearest neighbor analysis on the EIA-878 sample and OPIS database, and matches between addresses and stations' names were compared to confirm stations were the same. In the first round of matching, approximately 75% of the stations were matched.

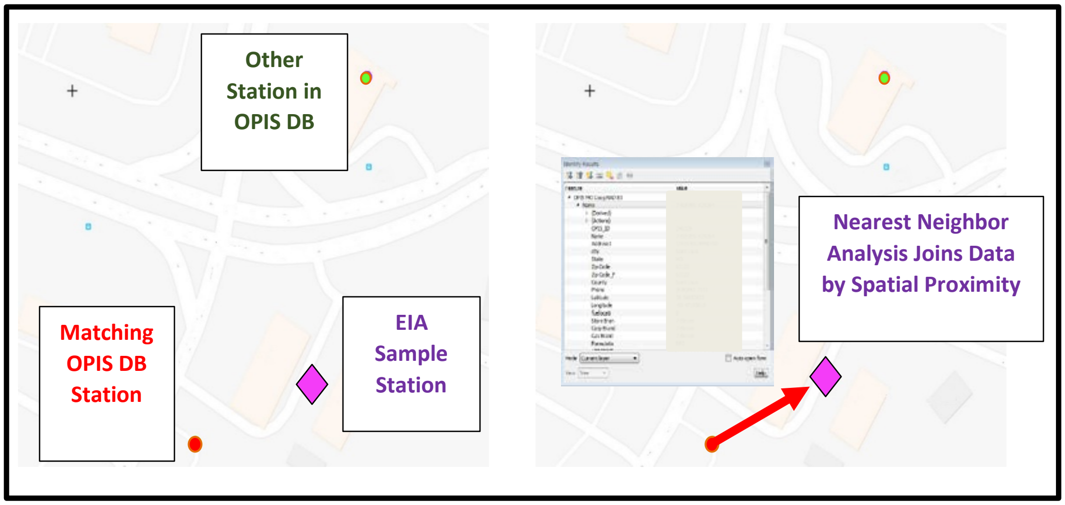


Figure 2: Illustration of the concept of GIS Nearest Neighbor Analysis

After the first round, we checked the remaining unmatched stations to determine reasons for mismatches and resolved them to continue matching. The primary reasons for initial mismatches were:

- Close Station Spacing: Most mismatches occurred around road intersections where multiple gas stations exist and some EIA-878 sample geolocated station points ended up being closer to incorrect stations than the correct gas stations. Once these station location points were corrected, matching was possible.
- Bad Geolocated Points: EIA-878 sample geolocated station points were not near the exact locations of the station and corrections were performed manually so matches could be made.
- Not included in OPIS data feed

Additional rounds of nearest neighbor analysis were performed as new weekly data feeds were received to capture additional stations that did not appear in the initial weekly OPIS data feeds. We also completed the same process with the GasBuddy database for stations. For a few cases, we identified multiple GasBuddy.com websites for the same station; in those cases, staff selected the website with the most up-to-date information. As with the OPIS database, similar results and shortcomings were experienced. In total, we were able to match 98% of EIA-878 sample stations with the OPIS database and 99% of EIA-878 sample stations with the GasBuddy database. Possible reasons for not identifying EIA-878 stations in OPIS and/or GasBuddy files include:

- Non-Physical Station Addresses: EIA-878 records contained mailing addresses that did not match the physical locations of gas stations, such as PO Box addresses.
- Bad Address: EIA-878 records contained some inaccurate addresses that were not

able to be geocoded or the other information did not match the station located at that physical address.

- No OPIS or GasBuddy Record: OPIS and/or GasBuddy did not have a record in the EIA-878 gas station sample set.

3. Applying "Total Error Framework"

To identify, understand, and compartmentalize sources of error in our integrated dataset of retail gasoline prices, we turned to the Total Survey Error (TSE) framework (Deming 1944, Groves 2004, Groves et al. 2009, Biemer 2010). Several scholars have broadened the TSE framework to cover a variety of data sources beyond surveys, including big (e.g., Japec 2015, Biemer 2017) and administrative data (e.g., Wallgreen and Wallgreen 2007). In what follows, we present a case study and apply the total error framework to our dataset. [4]

We outline potential sources of non-sampling errors arising from coverage, nonresponse, specification, measurement, and processing. Since we matched commercial data to a known probability sample survey, the examination of errors focus exclusively on non-sampling errors. Other studies on total error also delve into modeling, estimation, and analytic error, however, these will not be addressed in this research.

3. 1 Coverage

Coverage error occurs when the frame developed for a sample survey is misaligned with the population of interest. In our case, there is no "gold standard" list of retail gas stations available to assess coverage. We instead used a comparative analysis to examine the degree to which survey and commercial datasets under or over cover the population.

Few instances of undercoverage were identified (see Table 1). When we matched the stations in the EIA-878 sample to commercial sources, only a very small percent were not identified, as described previously in the section on linking methodology; however, we did not measure differences between the EIA-878 frame and commercial sources.[5]

Out-of-scope units (e.g., closed stations, card locks, and fleet fueling stations) were identified across all three data sources, however, they only comprised a small percentage. In addition, total counts of retail gasoline stations from OPIS and GasBuddy were also in general alignment with other organizations, including the U.S. Census Bureau and other retail trade groups.

[4] We refer the reader to other resources for a comprehensive review of the total survey error framework in both survey and non-survey context (e.g., Groves and Lyberg 2010, Japec 2015, Biemer 2017).

[5] After this analysis was performed, we revised the EIA-878 sample in May 2018 to better represent the target population, given changes to the universe after the frame was constructed and the sample selected. A more detailed description of the Motor Gasoline Price Survey methodology may be found at https://www.eia.gov/petroleum/gasdiesel/gas_proc-methods.php

Table 1: Evaluation of Potential Coverage Error

	EIA-878	**OPIS**	**GasBuddy**
Under Coverage	98% of sample stations matched in commercial data	Matched 98% EIA-878 sample stations	Matched 99% EIA-878 sample stations
Over Coverage	About 6% of current sample identified as out-of-scope	Out-of-scope units identified	Out-of-scope units and duplicate webpages identified

3.2 Nonresponse

In traditional survey data collection, nonresponse occurs when an eligible sample unit fails to respond to a survey request, either fully or for a specific item.[6] The equivalent concept in commercial datasets is missing data, but we will refer to this as "nonresponse" for simplicity. Since response to the Form EIA-878 is mandatory pursuant to Section 13(b) of the Federal Energy Administration Act of 1974 (Public Law 93-275) for the scientifically selected sample of companies selling gasoline at retail outlets, response rates are typically very high (see Table 2).[7]

In contrast, the commercial data we examined experienced lower response rates (see Table 2). Further, neither commercial data source distinguishes between missing data and whether a station sold the product. This is particularly common for midgrade gasoline, which some retailers do not carry. In contrast, EIA maintains station-level records on grades sold by retailers.

Table 2: Evaluation of Potential Nonresponse Error

	EIA-878	**OPIS**	**GasBuddy**
Nonresponse	Low nonresponse rates (< 10%)	On average, about 6% missing for regular gasoline and about 30% missing for midgrade and premium gasoline	On average, about 22% missing for regular gasoline and about 40% missing for midgrade and premium gasoline

We performed additional exploratory analyses to understand the missing data mechanism in commercial data sources, focusing on GasBuddy. For one week during summer 2018, we collected daily snapshots of GasBuddy data for approximately 1,000 stations that were in the EIA-878 sample and had valid GasBuddy websites. We then developed a dependent variable for price availability, defined as the percent of time the station has a price available. Over the time period observed, 189 stations had regular prices updated 100% of the time, compared to 83 stations for midgrade gasoline and 106 stations for premium gasoline.

As a next step, we gathered additional information on station characteristics that could help explain the variability in price availability by gasoline grade, focusing on gas station features, station utilization, and demographics (see Table 3 for a description of variables and sources).

[6] For the purpose of this discussion, we focus on both unit and item nonresponse.

[7] Form EIA-878 may be found at https://www.eia.gov/survey/form/eia_878/form_a.pdf.

Table 3: List of explanatory variables used in missing data analysis

Variable	Description	Linking	Source
Gas Station Features			
Amenities	Dummy variables indicating station amenities[8]	Station-specific	GasBuddy
Branded gasoline	Dummy variable to indicate whether station sells branded gasoline	Station-specific	OPIS
Price	Reported price during study week	Station-specific	EIA-878
Information source	Dummy variable to indicate whether GasBuddy price was crowdsourced (versus direct submission from the station)	Station-specific	GasBuddy
Utilization			
Annual Sales Volume	Annual sales volume in gallons	Station-specific	EIA-878, schedule B[9]
Demographics			
Population	Log of county population (Table S0101)	County	U.S. Census Bureau, 2012 - 2016, American Community Survey 5-Year Estimates
Gender	Percent male (Table S0101)		
Age	Percent of population age 18 – 65 (Table S0101)		
Race	Percent White alone; Black or African American alone; other (Table B02001)		
Ethnicity	Percent Hispanic (Table B03003)		
Education	Percent of population age 25 and over with high school diploma; bachelor's degree (Table B15003)		
Home ownership	Percent of owner occupied housing units (Table B25003)		
Household Income	Median household income (Table DP03)		
Labor Force Participation	Percent of population in labor force (Table DP03)		
Commute	Percent of workers age 16 and over who drove alone or carpooled (Table DP03)		
Urban	Dummy variable indicating station is in an urban area	Station-specific	2017 Urban Areas Boundary File, U.S. Census Bureau
Smartphone	Percent with one or more smartphone devices (Table B28001)	State	U.S. Census Bureau, 2016, American Community Survey 1-Year Estimates

[8] Gas Station amenities include: convenience store, restroom, pay phone, ATM, air, open 24/7, pay at the pump, has fuel, offers cash discount, has diesel, truck stop, car wash, loyalty discount, has propane, membership required, beer, wine, full service, restaurant, and has power.

[9] Form EIA-878 may be found at https://www.eia.gov/survey/form/eia_878/form_b.pdf.

As a first step, we explored the linear relationship between the dependent and independent variables. Stations that post directly on GasBuddy, sell branded gasoline, have more amenities, and have a higher annual sales volume were associated with having prices available for a larger percentage of the study week across all grades. We also identified a positive correlation between price availability and stations located in urban areas with a larger total population and a larger percentage of the population participating in the workforce. However, none of the pairwise Pearson’s correlation coefficients exceeded 0.4.

We explored several modeling approaches, including linear regression, regression trees, and random forests, to explore the data; however, given the low correlations, we were unable to explain a large portion of the variability in the availability of prices given the available explanatory variables. As a next step, we plan to collect and link data on annual average daily traffic (AADT) and highway type from Federal Highway Administration as this measure will be a more direct mechanism to observe station utilization.

3.3 Specification

In the survey context, specification error is defined as when “the concept implied by the survey question and the concept that should be measured in the survey differ’’ (Biemer and Lyberg 2003). GasBuddy and OPIS definitions of prices may not align with our target concept due to temporal or definitional disagreement and we discuss specific examples of differences in the two subsections that follow.

3.3.1. Timeliness

Timeliness covers two dimensions: frequency that data is collected for each retail gasoline station and at what point is that data then available to an end user. EIA-878 collects retail gasoline prices at a specific point in time across all sampled stations (8:00 a.m. local time) and then releases the data around 5:00 p.m. ET the same day.

In contrast, the time that prices are collected by commercial sources are station specific. As a commercial data vendor, OPIS can extract prices from their internal database at any date and time specified. EIA obtained four feeds with data at 8:00 a.m., 12:00 p.m., 4:00 p.m., and 8:00 p.m. and used the information to examine internal consistency. Internal analysis suggests that prices are generally timely; oldest prices were within 48 hours. GasBuddy displays reported prices in real-time and prices stay online for a maximum of one day.

3.3.1. Payment method

As specified on Form EIA-878 instructions, the reported prices should be the cash pump price for self-serve unleaded gasoline. If the station does not offer self-serve, then respondents are asked to report the cash price for mini-serve (if available) or full serve. If the station does not accept cash, then respondents are asked to report the credit card price.[10] While OPIS does not report payment method, analysis suggests that OPIS provides credit price in most cases. GasBuddy users can submit two different prices (cash and credit) for each grade of gasoline. Purchase mode does matter for price; internal analysis suggests that among stations who offer both a cash and credit price, the average discount is 8¢ for cash payment.

[10] Instructions may be found at https://www.eia.gov/survey/form/eia_878/instructions.pdf.

3.4 Measurement

Mapping the concept of retail gasoline price to a definition provided on a survey form or commercial database is a complex task. Both EIA-878 and GasBuddy provide similar definitions for gasoline price; however, the respondents' ability to map that concept to the survey form or crowdsourced website is unknown. In contrast, OPIS did not provide any documentation on price definition.

Each of the three data sources collect data using different modes as described in Table 4. In addition to mode effects, recall errors may also be present if there is a gap in time between when the respondent observes the price and when it is reported.

Table 4: Evaluation of Potential Measurement Error

	EIA-878	**OPIS**	**GasBuddy**
Data collection method	Probability sample survey with data collected mainly via computer assisted telephone interview (CATI)	Compiled data on universe of stations from credit card swipes, direct submission from stations, and user reports (via GasBuddy)	Reported by GasBuddy.com members and direct submissions from stations for universe of stations
Recall	Holidays: Monday prices collected on Tuesday	Unknown	Recall errors if users complete at a later date (e.g., from a desktop computer)
Gasoline grades collected and definitions	- Regular: 85-87 Octane and E0 - E15 - Midgrade: 88, 89, 90 Octane and E0 – E15 - Premium: 91 Octane or higher and E0 – E15 - Does not collect E20 or E85 - Price excludes discounts	Unknown	- Regular: 85-87 Octane - Midgrade: 89 Octane - Premium: 91-93 Octane - E85 (GasBuddy, "Fuel Types") - For stations in Iowa, Illinois, Nebraska, and Kentucky, E0 should be reported as Regular and E10 should be reported in the Midgrade spot (GasBuddy, "Report Prices") - Price excludes discounts (GasBuddy, "Report Prices")

3.5 Processing

Errors may also occur during the final stages of data collection during processing and adjustment activities. For the EIA-878, EIA follows best practices in survey data collection. CATI data collected are keyed by data collection contractors. The data are then validated using automated checks, and respondents are re-contacted as needed to confirm the data entered in the system. Additional outlier detection and editing programs are also used, and all data are reviewed by subject matter experts.

For the commercial data sources, data processing and editing steps are largely unknown. Correspondence with OPIS indicates price for each station is based on last credit card swipe that occurred before the data were pulled. However, any further steps taken to clean or edit the data are not known. GasBuddy states they use "automated algorithms" to detect "obviously wrong information" and the site also allows users to report other users who submit incorrect information (GasBuddy, "False Prices"). The prevalence of such errors or corrections is not reported.

Any errors associated with creating the blended data set (e.g., geocoding, data linkage) would also be classified as processing and adjustment error; however, no instances of data linkage error have been identified to date.

4. Discussion

Statistical agencies across the federal government have a strong interest and directed effort to investigate the use of alternative and blended data. This report documents one potential application of combining commercial data from OPIS and GasBuddy with traditional survey-based methods used on the EIA-878. We present key findings and recommendations in the sections that follow from this analysis. The findings and recommendations conclude with suggestions for future research efforts to enhance EIA's data collection and publication efforts.

4.1 Assessment of data quality and value in commercial data

We assessed OPIS and GasBuddy on a number of indicators of data quality and value. Both OPIS and GasBuddy exhibited high levels of coverage compared with the current EIA-878 sample. OPIS price feeds were internally consistent in terms of both station information and prices, as well as highly correlated with GasBuddy prices. Analysis also suggests an added value of purchasing data from OPIS rather than obtaining GasBuddy prices via web based methods in terms of missing data.

Neither commercial data source supplied information to explain item or unit nonresponse, such as availability of midgrade/premium gasoline or temporary or permanent station closures. GasBuddy does collect information on power outages or lack of supply, but only under extreme circumstances.[11] Also, neither data source provides transparent information on data collection, processing, and validation.

4.2 Comparability of estimates

OPIS, GasBuddy, and EIA-878 differed on a number of key issues that affected measurement, including price definition, data collection mode, and reference period for the prices. Despite these differences, EIA-878 and OPIS regular gasoline prices were within 5¢ for almost half of the EIA-878 sample (45%). Prices for regular grade gasoline matched exactly for 22% of EIA-878 stations over a 12-week study period from May 15, 2017 to July 31, 2017. Correlations between EIA-878 and commercial data sources were also very strong, particularly for regular gasoline prices.

When comparisons were limited to stations reporting across both EIA-878 and OPIS, national and regional level estimates were within 1¢ and 4¢, respectively, across all grades over the 12-week study period from May 15, 2017 to July 31, 2017. See Figure 3 for a comparison between EIA-878 and OPIS of average unweighted weekly national prices for regular grade gasoline. OPIS prices were generally higher at the national and regional level, likely due to the utilization of credit prices rather than cash prices. City and state-level estimates exhibited larger differences, but these differences are likely due to small sample sizes within the publication cells.

[11] The website displays fuel tracker features at http://tracker.gasbuddy.com

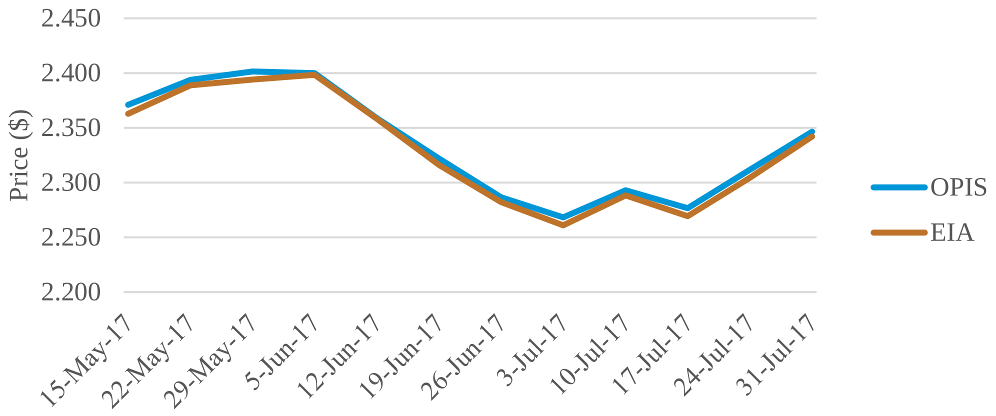


Figure 3: Average weekly national prices between EIA-878 and OPIS data (both unweighted) for regular motor gasoline in dollars per gallon (596 stations)

4.3 Recommendations

This research suggests that there may be various options for incorporating the use of OPIS and/or GasBuddy data with EIA-878. Outlined below are options for incorporating third-party data sources into the EIA-878, starting with the most conservative approach.

- **Validation Tool:** This study demonstrated the ability of third-party data to identify price discrepancies. By using third-party data for validation in real-time, potential reporting errors could be identified and investigated during data collection, improving data quality for EIA-878.
- **Imputation**: The high levels of comparability between survey and commercial data also may support new methods for handling missing data. EIA could identify these stations on the EIA-878 data files and utilize OPIS prices, rather than model-based imputation, in cases of survey nonresponse.
- **Replace weekly survey data collection with third-party sources for selected stations:** To reduce respondent burden, EIA could suspend weekly price data collection from selected stations and instead purchase data from OPIS. This step would involve additional research on cost/trade-off considerations before a decision could be made. However, to compute weekly volume-weighted price estimates, EIA would either need to continue to collect annual volume data from these stations or find a third-party source for motor gasoline volumes.
- **Hybrid approach:** The current EIA-878 sample consists of approximately 1,000 stations. In future iterations, EIA could further expand the sample size and utilize both commercial and survey data collection in tandem, increasing the number of states and cities with published estimates and decreasing sampling variability. In this paradigm, EIA could designate some stations to report via traditional survey-based methods, while relying on third-party data for others. The distinction could be made at random or geographically. If the latter, EIA could publish some areas as "EIA-878" states/cities, and other areas as "third-party" states/cities. Alternatively, EIA could employ a phased design and use third-party data sources as the default and then subsample any stations with missing data using a survey data collection. While EIA could utilize third-party sources for price data, there is no known third-party source for motor gasoline volumes, which are used in deriving both estimates and standard errors. Further research is needed in this area, as well as an understanding of cost implications, legal restrictions on republishing third-party data, and the needs of the data user community.

5. Conclusion

This paper reviews the error properties associated with both survey and commercial data sources of retail gasoline prices and seeks to understand implications for building integrated datasets and producing blended estimates.

Based on the findings, this research suggests a number of promising applications of using commercial price data at the station level from OPIS and GasBuddy to enhance the EIA-878. A conservative approach may include incorporating third-party data into the current EIA-878 validation process, utilizing outside data sources to identify potential reporting errors in real time. Other techniques, such as utilization of commercial data for imputation or replacement of weekly survey data collection with alternative sources for selected respondents, require a heavier reliance on third-party data vendors. Collecting data via commercial sources may also allow EIA to expand our visibility and sample size beyond what is currently feasible using traditional survey data collection methods. Further research is needed to determine if replacement or hybrid approaches are methodologically sound or even feasible when considering costs and benefits.

Acknowledgements

The authors thank the following for their valuable contributions to this research:
Nathan Agbemenyale, Zane Alpher, Shala Brown, Julian Castillo, Carol French, Eliza Goren, Tammy Heppner, David Kinyon, Ruey-Pyng Lu, Renee Miller, Patrick Murray, Kendra Pospychalla, Marcela Rourk, Cecile Sano, Anna Scaramuzza, Lou Schloss, Michael Scott, Nanda Srinivasan, Lawrence Stroud, Andrew Thomson, Ethan Walker-Seim, Dan Walzer, Anna Yakovleva, and Bin Zhang.

References

Abcede, Angel. (March 1, 2013). “OPIS Takes It to Street With GasBuddy Acquisition.” CSP Daily News. Retrieved from http://www.cspdailynews.com/fuels-news-prices-analysis/fuels-news/articles/opis-takes-it-street-gasbuddy-acquisition

Biemer, P, and Lyberg, L. (2003). *Introduction to Survey Quality*. Wiley.

Biemer, P. (2010). “Total survey error: Design, implementation, and evaluation.” *Public Opinion Quarterly*, 74(5): 817-848.

Biemer, P. (2017). “Errors and Inference.” In *Big Data and Social Science: A Practical Guide to Methods and Tools*, edited by Ian Foster, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter, and Julia Lane, 265–97. CRC Press.

Deming, W.E. (1944). “On errors in surveys.” *American Sociological Review* 9(4): 359–369.

GasBuddy (n.d.). “False Prices”. Retrieved from https://help.gasbuddy.com/hc/en-us/articles/206722608-False-Prices

GasBuddy (n.d.). “Report Prices”. Retrieved from https://help.gasbuddy.com/hc/en-us/articles/225454887-Report-Prices

GasBuddy (n.d.). “What Fuel Types Does GasBuddy Support”. Retrieved from https://help.gasbuddy.com/hc/en-us/articles/212869897-What-Fuel-Types-Does-GasBuddy-Support-

GasBuddy (n.d.). “Who/What is GB_Direct”. Retrieved from https://help.gasbuddy.com/hc/en-us/articles/207511357-Who-What-is-GB-Direct-

Groves R.M. (2004). *Survey errors and survey costs.* Second edition. Wiley.

Groves, R.M, Fowler F.J., Couper M.P., Lepkowski J.M., Singer E., and Tourangeau R. (2009). *Survey methodology*. Second edition. Wiley.

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neil, C., and Usher, A. (2015). "Big data in survey research AAPOR task force report." *Public Opinion Quarterly*, 79(4): 839-880.

Oil Price Information Service (OPIS) (2017). "Energy: Pricing and news across the entire fuel supply chain". Retrieved from https://www.opisnet.com/wp-content/uploads/2017/09/OPIS-Brochure..pdf