Linking a Retail Gasoline Price Survey with Commercial Data

For
FCSM Research and Policy Conference
March 9, 2018 | Washington, DC

By
Maura Bardos, Amerine Woodyard, and Jeramiah Yeksavich
Energy Information Administration

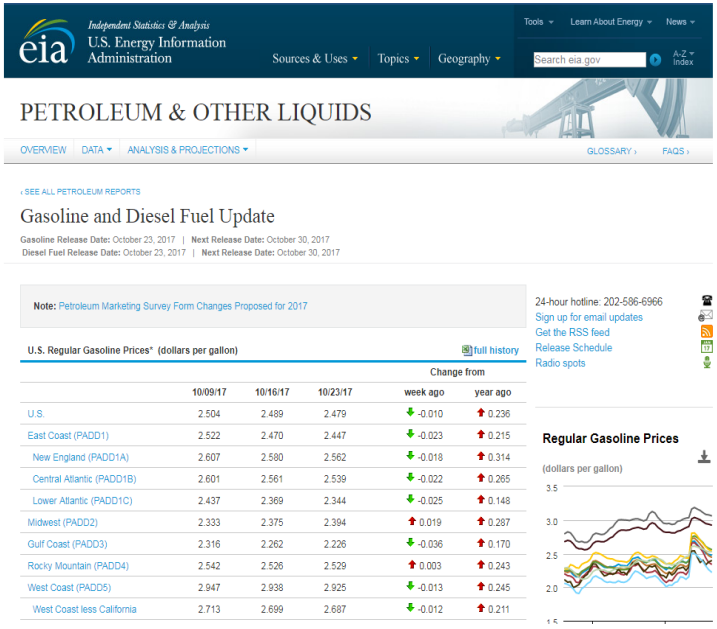
Overview

- Motivation
- Data sources
  - Motor Gasoline Price Survey (EIA-878)
  - Oil Price Information Service (OPIS)
  - GasBuddy.com
- Methods and findings
- Approaches to incorporate third-party data
- Challenges and Future Work

Motivation

- Responsibility of federal statistical agencies to investigate alternative sources of data to:
  - Increase time and cost efficiencies
  - Reduce respondent burden
- Statistical Methodology Improvement Plan
- Larger context of research priorities for federal statistical agencies
  - CNSTAT reports on Multiple Data Sources
  - Commission on Evidence Based Policy Making

Data Source: Motor Gasoline Price Survey (EIA-878)

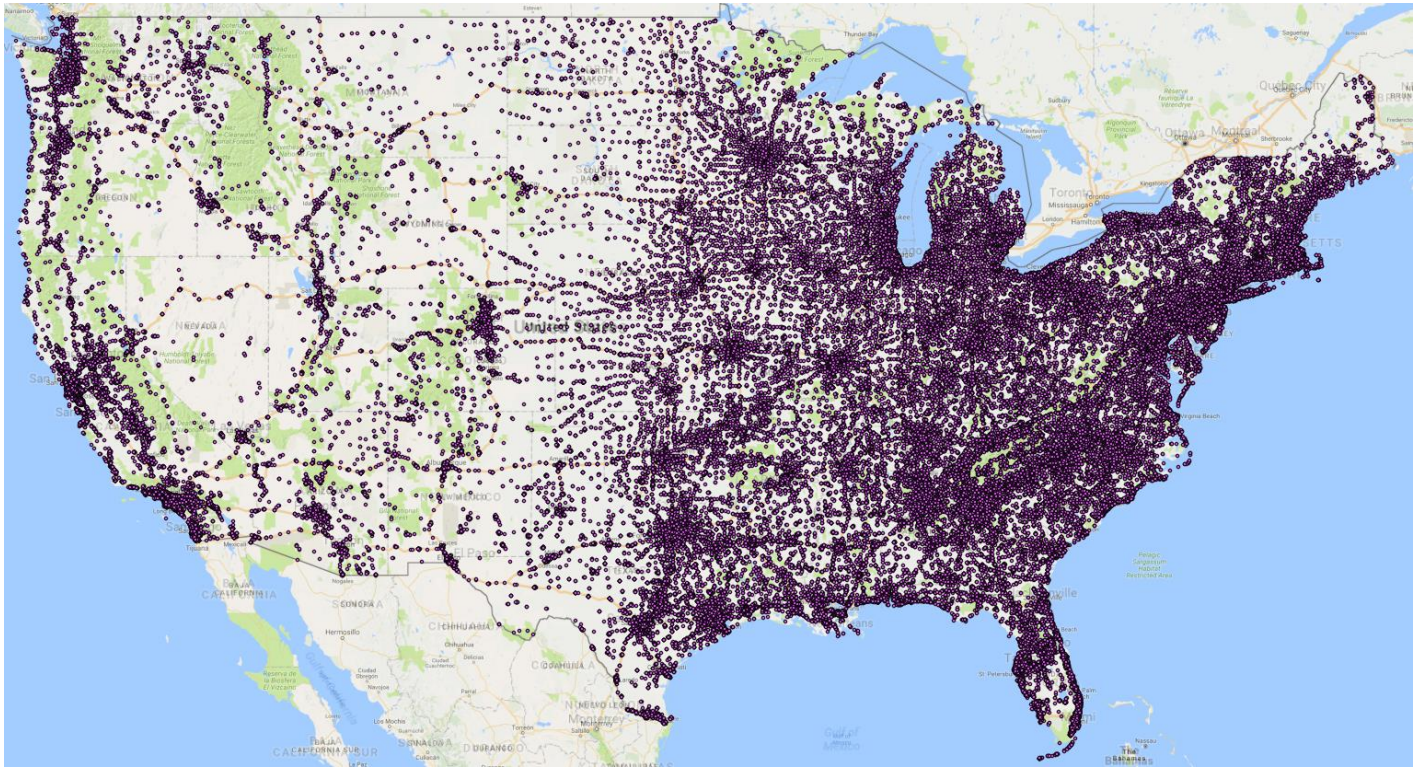


- Weekly mandatory CIPSEA survey of approximately 800 retail gasoline stations across the country.
- "Gasoline price" definition: Cash price per gallon (including taxes) as of 8:00 a.m. local time each Monday
  - Regular, midgrade, and premium gasoline.
- Mode: Mostly CATI; however, other modes also available
- Same day data collection, processing, and dissemination
- Estimates are produced for 276 publication cells
  - Nation, regions, 10 cities, 9 states
  - Regular, midgrade, and premium
  - Conventional and reformulated gasoline

Motor Gasoline Price Survey (EIA-878) Methodology

- This presentation covers a small slice of the overall EIA-878 redesign
- Investigated many potential frame sources, but ultimately decided to use a list purchased from OPIS (~130,000 retail gasoline outlets)
  - Process described in detail at last year's FCSM conference
- Selected stratified systematic sample with total sample size of 1,000
  - Stratified by retail gasoline outlet type: big-box or non-big-box
  - Allocated to sampling strata and outlet type in proportion to the number of outlets in the sampling strata and approximate sales volume per outlet
- Measure of Size: Annual Sales Volumes
  - No known third-party source
  - Survey to collect this information is currently in the field
- Volume weighted price estimates

Gas Station Locations from OPIS Database



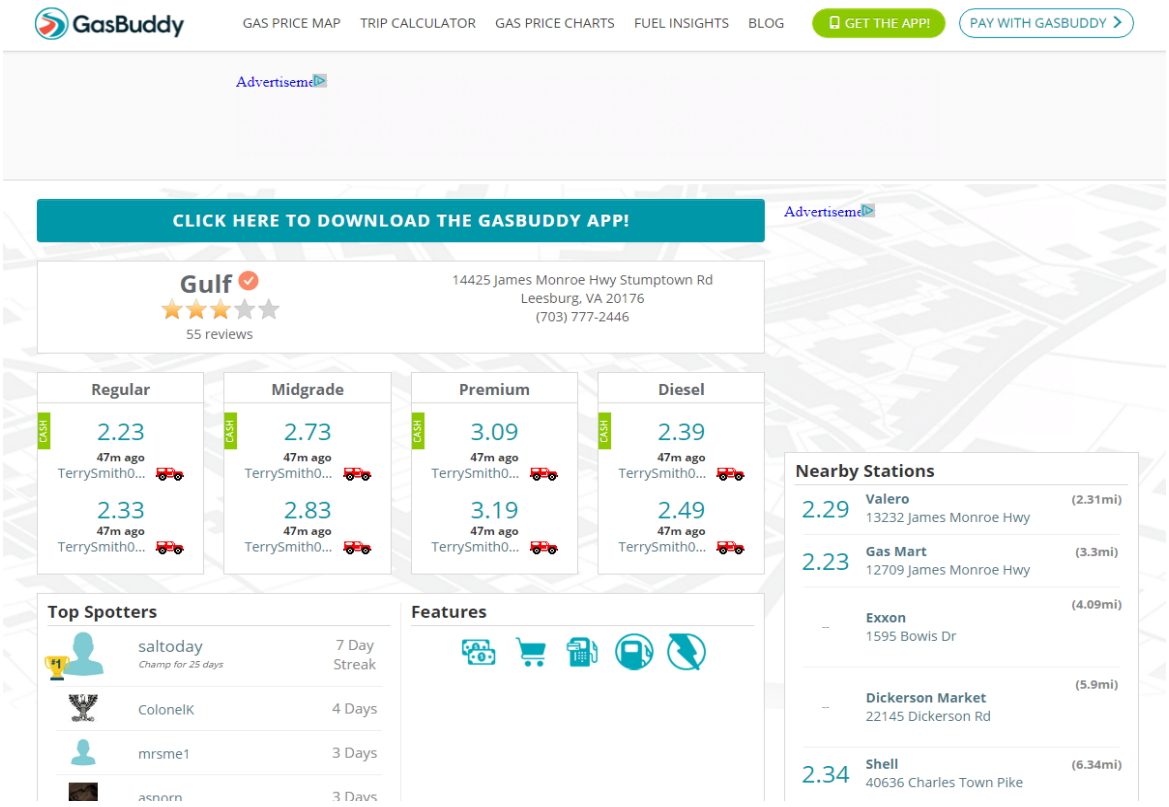
Data Source: Oil Price Information Service (OPIS)

- Commercial source of daily gasoline prices for retail outlets in the U.S. and Canada
- From May 15, 2017 to July 31, 2017, EIA purchased weekly feeds of gasoline price data
- Each feed included the following for regular, midgrade, and premium:
  - Price per gallon (single value, no distinction between cash or credit)
  - Date and time associated with the price
  - Source: Electronic or User Submitted
- Multiple feeds available on Monday: 8:00 a.m., 12:00 p.m., 4:00 p.m., and 8:00 p.m.
- On average, each file consisted of about 110,000 records
  - 76% (83,910 stations) appeared consistently each week

Data Source: GasBuddy.com

- Crowdsourced website, prices for about 144,000 U.S. outlets
- Owned by OPIS
- Prices updated in real-time
- Members earn points updating a gas prices; redeemed for prizes in raffles.
- “Automated algorithms” used to detect “obviously wrong information.” Users can also report incorrect information
- EIA inter-office collaboration to build additional functionality into a web-based data collection tool for one Monday (July 31, 2017)
  - Cash and credit price for regular, midgrade, premium
  - Date and time submitted
  - Source (user submitted or direct data transmission from the outlet)

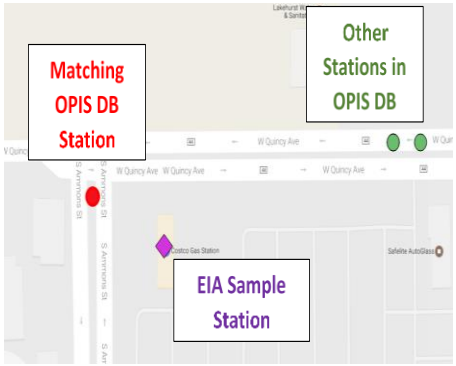
Data Source: GasBuddy.com

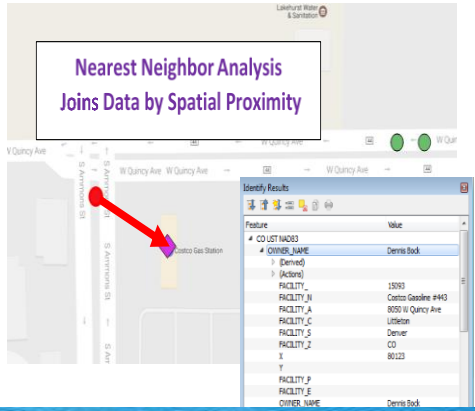


Methods

- Data Linkage
  - Challenge: No common ID between EIA-878, OPIS, and GasBuddy
  - Solution: Deterministic address matching using geographic information systems (GIS) approach and nearest neighbor analysis
  - Results:
    - OPIS: matched 785 EIA-878 sample stations (98%)
    - GasBuddy: matched 798 EIA-878 sample stations (99%)

- Analytical Datasets
  - Longitudinal: Stations with nonmissing EIA-878 and OPIS data over 12 weeks
  - Point-in-Time: Stations with nonmissing EIA-878, OPIS, and GasBuddy data for July 31, 2017



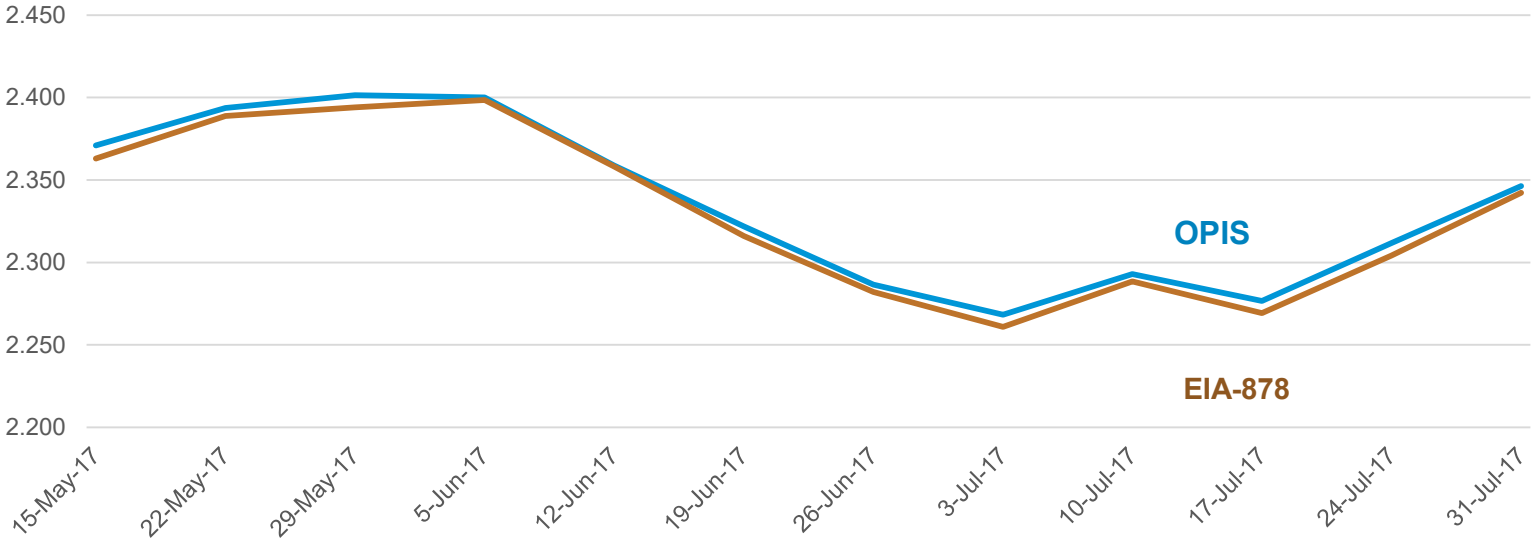


Data Quality

	OPIS	GasBuddy
Data collection method	Compiles data from universe of stations from - credit card swipes - direct submission from stations - user reports (via GasBuddy)	Compiles data from universe of stations from - User reports
Coverage	Matched 785 EIA-878 sample stations (98%)	Matched 798 EIA-878 sample stations (99%)
Timeliness	- File can be delivered at any date/time specified. - From analysis of the weekly feeds, the oldest prices were as of 12:00 a.m. Sunday.	- Prices are updated in real-time - Prices stay online for a maximum of one day
Price Definition	- Unknown	- Octane for regular/mid/pre aligns with EIA - Provides similar instructions to EIA-878 for exclusion of price discounts, inclusion of taxes in price reports
Cash vs. Credit	Analysis suggests that OPIS provides credit price	Users can submit two prices: cash and credit
Data editing and processing	- Unknown	- "Automated algorithms" detects "obviously wrong information" - Users can correct submit incorrect information
Missing Data Rates	- Regular: 6% - Midgrade and premium: ~30%	- Regular: 22% - Midgrade and premium: ~40%

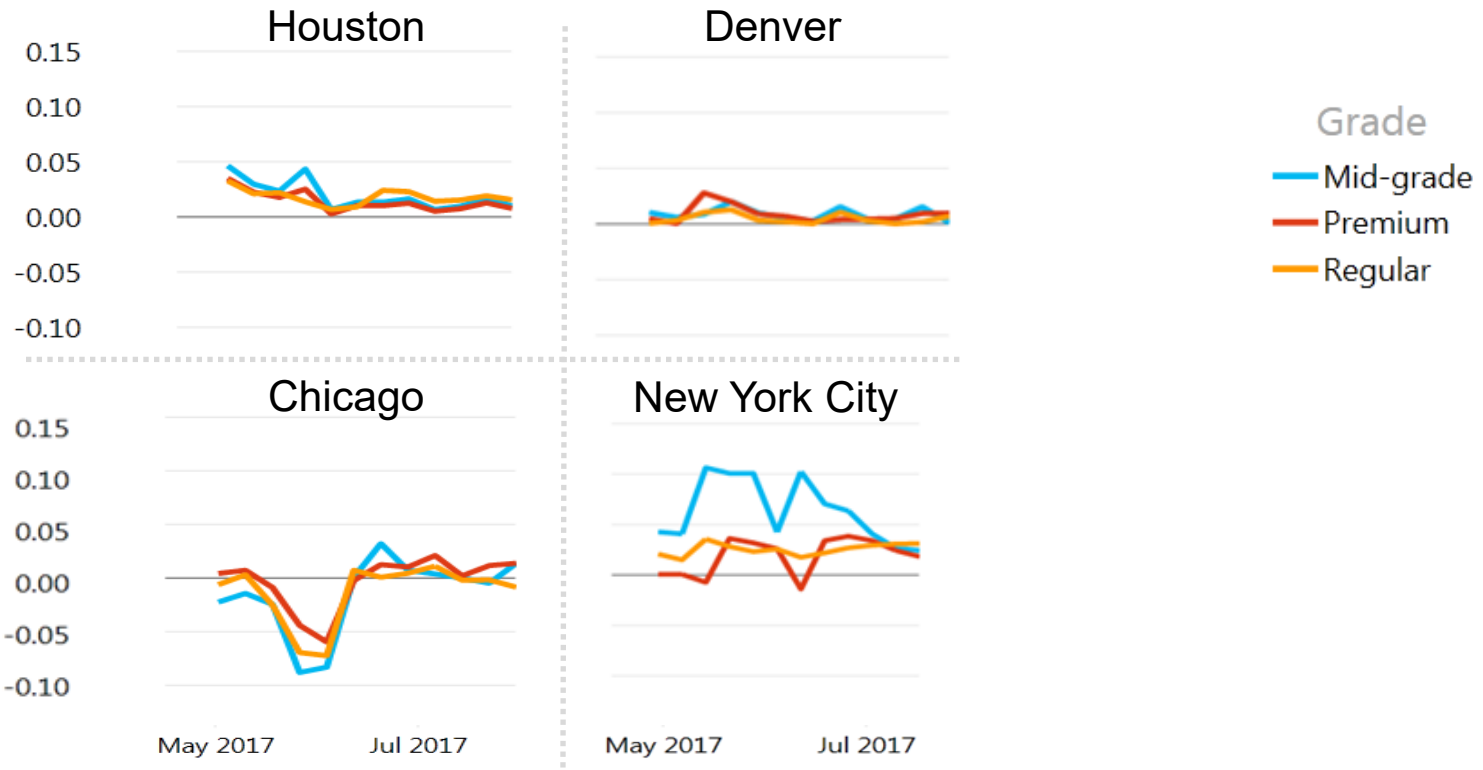
Survey and third-party data yield estimates for the U.S. within 1¢ across the 12-week study period

U.S. regular gasoline price
dollars per gallon



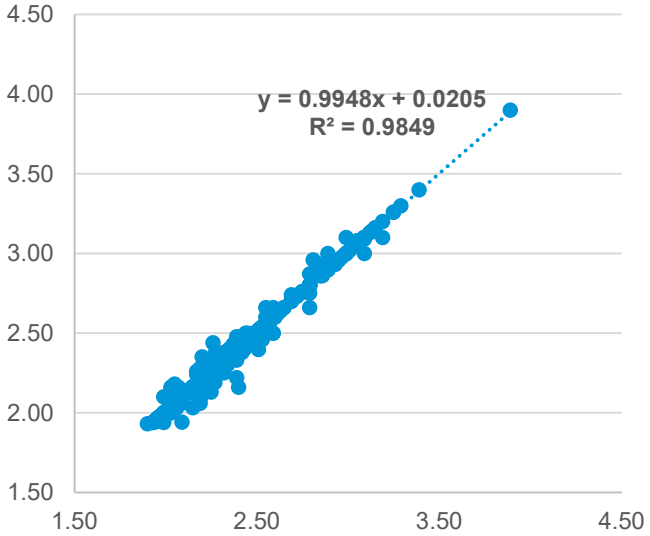
Source: EIA-878, OPIS
Note: Unweighted average calculated from 596 stations with non-missing data over the 12-week study period

Larger differences are identified at the city level, however, none exceed 10¢ for any city or grade type



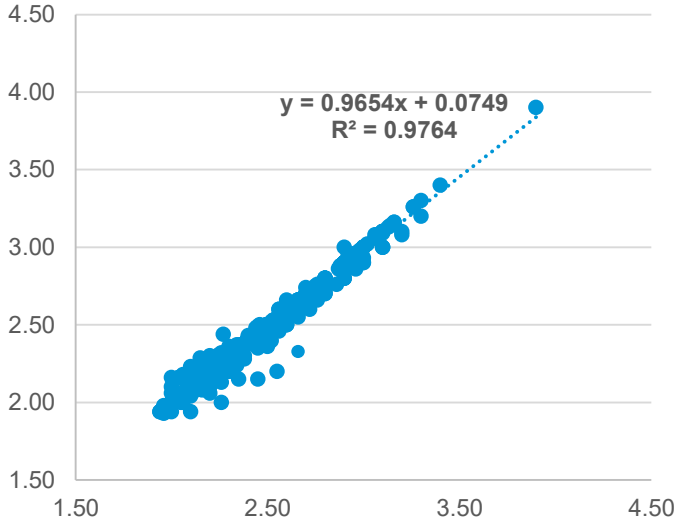
Regular gasoline prices are highly correlated at the station level between EIA, OPIS, and GasBuddy, but large differences for stations were identified

EIA-878 regular gasoline price versus GasBuddy regular gasoline price, July 31, 2017
dollars per gallon



Source: EIA-878; GasBuddy.com
Note: 616 stations with non-missing data for regular gasoline price July 31, 2017

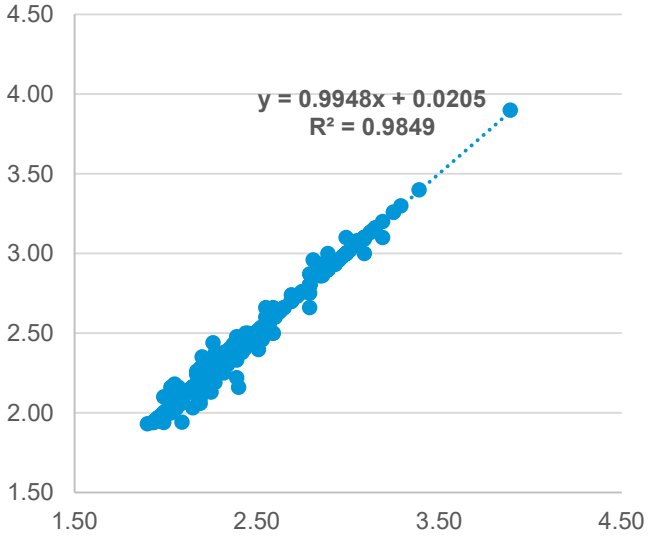
EIA-878 regular gasoline price versus OPIS regular gasoline price, July 31, 2017
dollars per gallon



Source: EIA-878; GasBuddy.com
Note: 742 stations with non-missing data for regular gasoline price July 31, 2017

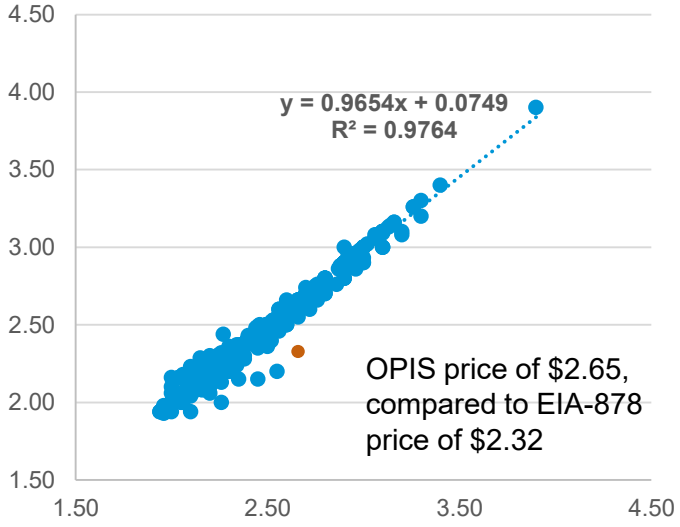
Regular gasoline prices are highly correlated at the station level between EIA, OPIS, and GasBuddy, but large differences for stations were identified

EIA-878 regular gasoline price versus GasBuddy regular gasoline price, July 31, 2017
dollars per gallon



Source: EIA-878; GasBuddy.com
Note: 616 stations with non-missing data for regular gasoline price July 31, 2017

EIA-878 regular gasoline price versus OPIS regular gasoline price, July 31, 2017
dollars per gallon

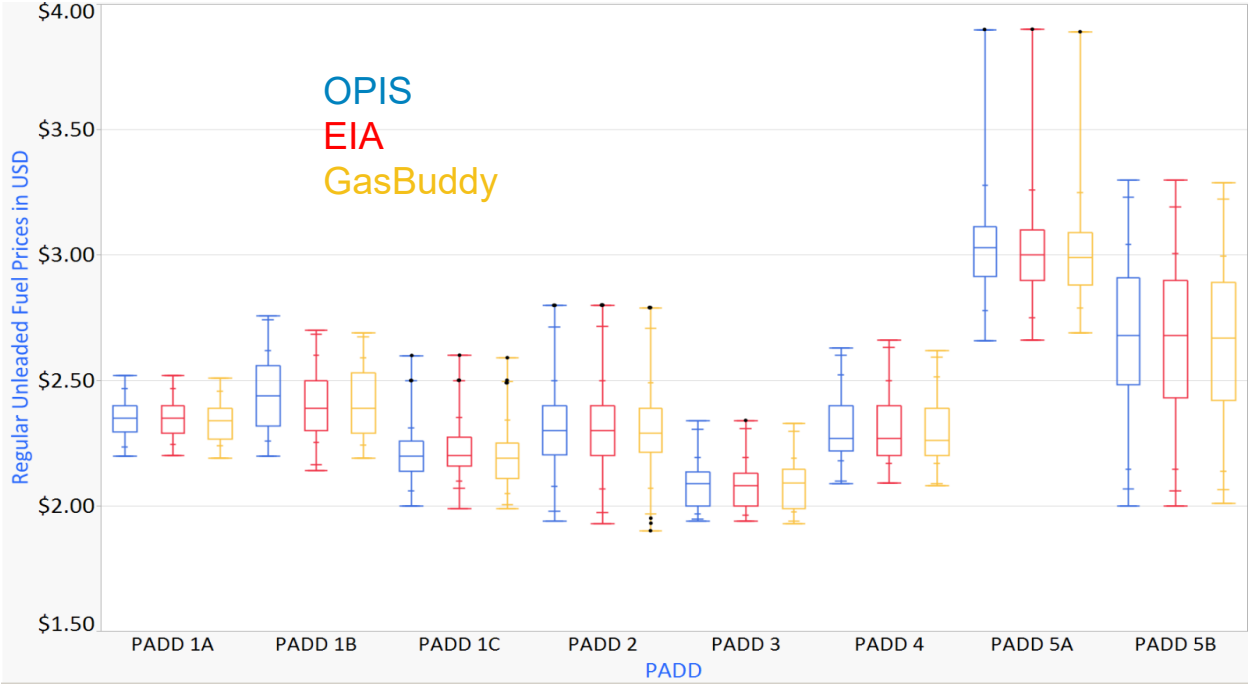


Source: EIA-878; GasBuddy.com
Note: 742 stations with non-missing data for regular gasoline price July 31, 2017

OPIS, EIA, and GasBuddy capture the distribution of prices for PADDs

Distribution regular gasoline prices by published PADD and data source

dollars per gallon

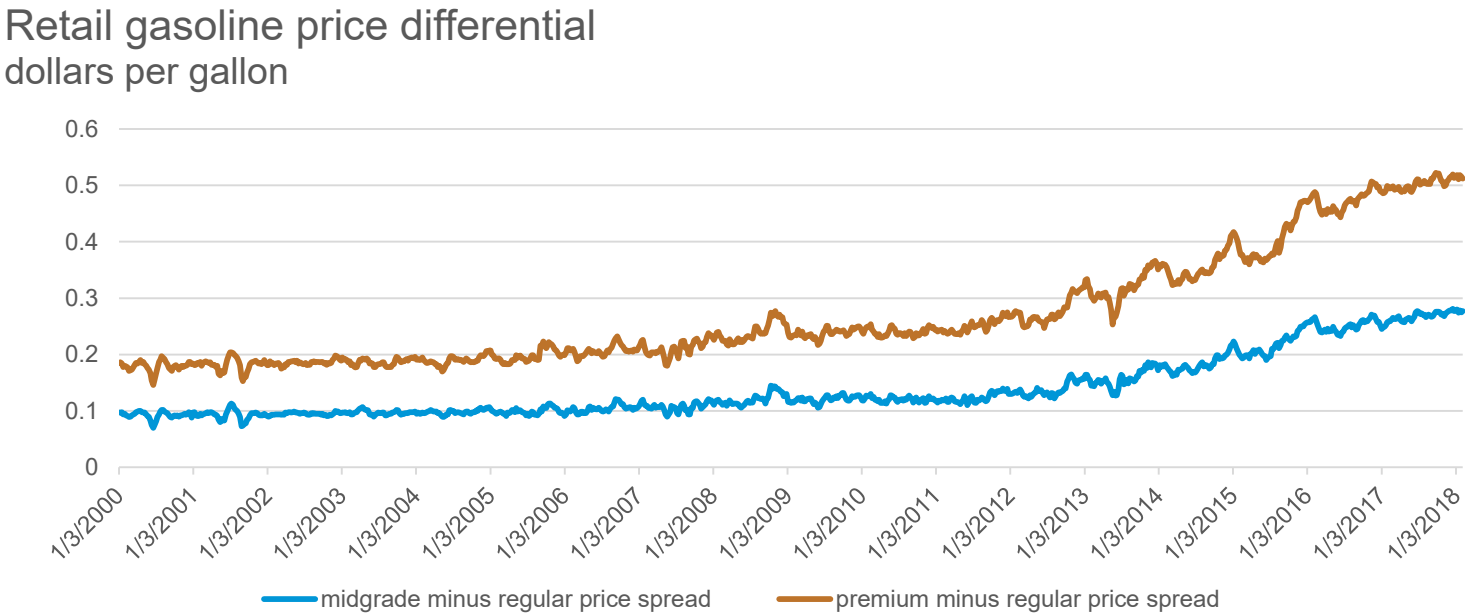


Source: EIA-878, OPIS, and GasBuddy.com

Note: Unweighted average calculated from stations with non-missing data on July 31, 2017

Model-based approach

Research question: Can we take advantage of the relationship between gasoline grades to produced model-based estimates?



Model-based approach

- Developed model to predict midgrade and premium prices at the outlet level based on the following covariates:
  - Price of regular gasoline from EIA-878
  - Economic Conditions: Index of Consumer Sentiment, Inflation rate, median household income, labor force
  - Geography: State, population
  - Petroleum Industry: Gasoline formulation, proximity to refinery, crude oil spot prices, number of grades sold at the station
  - Transportation: Vehicle miles traveled, registered vehicles, use of public transportation

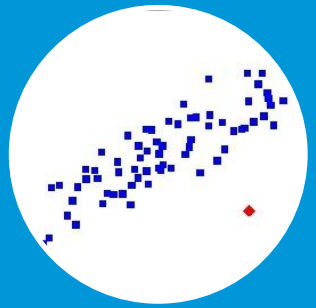
- Regression tree approach
  - Used Rpart package in R
  - Randomly divided one week of data into test and training to develop and evaluate the model (~800 observations in each weekly dataset)

Model-based approach

- The cumulative R^2 values for the midgrade and premium gasoline models were 0.85 and 0.81
  - As expected, the price of regular gasoline was selected as the first split
  - Additional auxiliary variables also offered predictive power, in particular those related to transportation and economic conditions

- Results are promising, but more work to do
  - Validate for robustness over a number of weeks and understand model differences
  - Compare results to linear model
  - Utilize third-party datasets with many, many more observations (~110,000)

Approaches to incorporate third-party data



Data Validation

- Develop automated tools to identify potential outliers in real-time

Imputation

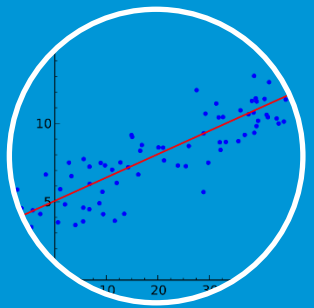
- Rely on third-party sources for survey nonresponse, rather than model based imputation

Third-party data for select stations

- Designate stations with third-party price reports similar to EIA-878 as 'third-party' stations
- Survey data collection for the remainder

Hybrid Approaches

- Increase the sample size and small area estimates
  - Two-Phase design
  - Randomly assign data source
- But, volumes needed



Model-Based Approach

- Obtain data for regular gasoline and model other grades

Less Use of Third-Party Data More

Ongoing Work and its Challenges

Current Path Forward…

- Utilize third-party data for validation and imputation

…and its challenges

- Data Access and Availability
  - Subscriptions are costly, and at this time, cost prohibitive
  - Web-sourced data changes over time
  - Even if we can access data, necessary data elements may be missing
- Transparent methodology for data users
- Building automated tools into production environment

Future Work

- Expand the study period and continue to learn about third-party sources
  - Understand internal consistency of GasBuddy data and how it tracks with EIA-878 over time
  - Extend OPIS subscription
  - Supply disruptions (Hurricanes Maria and Harvey)
- Continue initial modeling efforts
  - Use third-party data as source for regular prices (~110,000 stations)
  - Identify other variables with predictive power
- Explore options to publish more local information
  - Currently publish for 10 cities/9 states, but third-party data may allow us to expand

Acknowledgements

Nanda Srinivasan	Tammy Heppner
Nathan Agbemenyale	Andrew Thomson
Maura Bardos	Dan Walzer
Ruey-Pyng Lu	Kendra Pospychalla
Lawrence Stroud	Anna Scaramuzza
Anna Yakovleva	David Kinyon
Renee Miller	Bin Zhang
Marcela Rourk	Shala Brown
Cecile Sano	Julian Castillo
Lou Schloss	Eliza Goren
Michael Scott	Ethan Walker-Seim