Does Location Matter? A Case-study of the Influence of Geography in the Measurement of Gasoline Price Inflation November 2019

David Popko[1], Ilmo Sung[2]

[1]Bureau of Labor Statistics, 2 Massachusetts Ave NE, Washington, DC, 20212

[2]Bureau of Labor Statistics, 2 Massachusetts Ave NE, Washington, DC, 20212

Disclaimer: The opinions expressed in this paper are those of the authors and do not represent the policy of the Bureau of Labor Statistics.

Introduction

The objective of the Consumer Price Index (CPI) is to measure the change in cost of living experienced by the average urban consumer residing in the United States. Currently, the outlet sample in this measurement is selected via probability sampling proportional to average daily expenditure on items within a core-based-statistical-area (CBSA). This process yields a large variety of outlets in the CPI sample, but the outlets selected are based exclusively on the expenditures reported and household sampling weights in CPI's Telephone-Point-of-Purchase Survey (TPOPS) survey.

Using data collected from GasBuddy.com[1], we attempt to model the explanatory variables of price change in order to identify possible stratification variables in outlet selection. Focusing on the Washington-Arlington-Alexandria, DC-VA-MD-WV CBSA, we calculate one-month price changes for each gas station in the GasBuddy sample. We then apply various statistical techniques to assess the significance of a variety of independent variables constructed using driving distances between stations, population density, income, and housing prices while controlling for various geographic flags. Finally, we construct indexes from the GasBuddy sample using a proposed stratified methodology, and compare them with indexes constructed using the traditional CPI methodology.

1. Data

1.1 Gasbuddy

[1] All data collection and use of GasBuddy's name for research purposes such as this was done with the written permission of GasBuddy LLC

Gasbuddy is a company that provides gasoline payment services and real time gasoline price information for over 140,000 stations in the United States and Canada. They collect prices through partnerships with station owners, credit card companies, and reporting by a network of approximately 2.5 million application users. Users enter prices at a station while at the pump, in exchange for entry in a daily drawing of a $100 prepaid gas card; they can also report via the *Pay with Gasbuddy* feature of the app, for a discount on the fill-up.

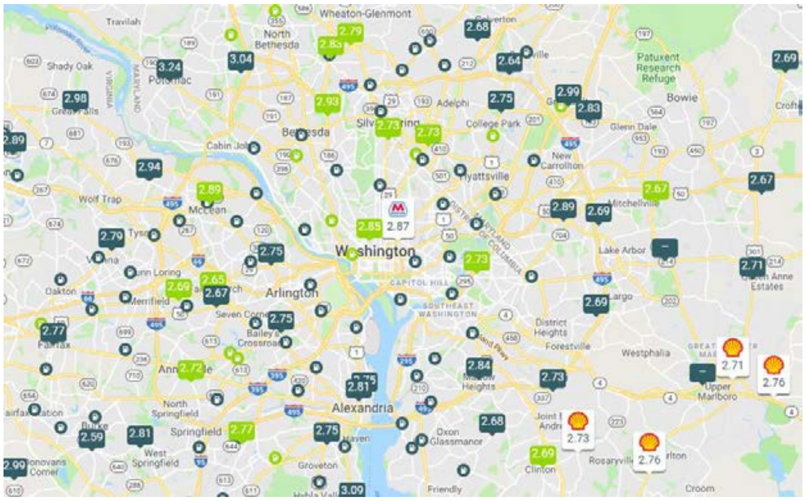


Figure 1: Rendering of the GasBuddy.com map of gasoline prices, showing all reported stations in the Washington DC metro area.

1.2 Attributes

Each row in the GasBuddy dataset contains the reported gas price, posting time, and information about the station. During data collection from October 2017 to June 2019, the CPI collected four separate datasets each day, for regular unleaded, midgrade, premium, and diesel fuel. Each row contains a price observation for a given station, denoted by a unique *stationID,* and fuel grade. The data also contain additional geographic info on the station, such as address, zip-code, and geographic coordinates, as well as information on the user that recorded the observation. Using this coordinate level information, we calculated the distance from each station to its nearest neighbor as an independent variable for our analysis. Finally, each observation is timestamped according to its time of upload to GasBuddy; any observation existing in the GasBuddy database for more than 48 hours is automatically removed.

1.3 Collection Methodology

With the company's permission, the Bureau of Labor Statistics (BLS) research team has been collecting daily motor fuel price data for regular, midgrade, premium, and diesel from their website since June 2016. The automated data collection we implemented is differently from a regular web scraping in which one captures information on a web browser or html source files, but rather a method to pull prices and information of gas stations directly from the company's database, which is indirectly open to the public in a series of web communications. This data-mining method is based on information obtained from communications (network packets) between users and the company's servers, which were captured through reverse-engineering mobile apps or web applications. This source of web scraped data provides us with the opportunity to measure the daily price fluctuations that consumers encounter at the gas pump.

Employing the above methodology, BLS collects prices from approximately 90,000 stations daily (approximately 1,450 in the Washington DC metro area, our realm of interest). By comparison, the CPI collects approximately 4,000 gasoline price observations from some 1,300 stations monthly. This cornucopia of data allows for a more granular view of price movements across a geographical area.

1.4 Additional Data Sets

To bolster the geographical richness of our dataset, we synthesized it with a number of additional data sets. Using 2010 census estimates, the population density of a given station's surrounding zip-code, as the median estimated home value of a station's surrounding zip-code, taken from Zillow. We also mapped stations' zip-codes to their respective county FIPS codes, grouped gasoline brands according to the TOP TIER designation. TOP TIER signifies whether the gasoline brand is treated with TOP TIER Detergent[2]; most "premium" gasoline brands possess the TOP TIER designation.

2. Methodology

2.1 Data Preparation

Our scope of study is the set all GasBuddy observations collected in the Washington-Arlington-Alexandria, DC-VA-MD-WV CBSA between the months of October 2017 and June 2019. As we are primarily concerned with month-over-month price changes, we derive our dependent variable – one-month percentage change in price, or *price relative* – in the following manner: (1) divide each month in our period of study into three "pricing periods." Observations recorded from the 1st to the 10th of the month are assigned pricing period 1, from the 11th to the 20th pricing period 2, and from the 21st to the end of the month pricing period 3. This pricing period approach mimics CPI data collection methodology, and increases the number of pairwise comparisons (price changes) threefold. (2) For each station, calculate an arithmetic average of all prices in a given pricing for each fuel grade. This, by assumption, is the station's price. (3) Our dependent variable, *price relative*, is defined by dividing each station's price in a given months pricing period, by the station's price in the same pricing period of the previous month.

[2] A list of all gasoline brands that retain the TOP TIER designation as of 9/5/2016 -- https://toptiergas.com/licensed-brands/

$$price\_relative_{t,p,s,g} = \frac{price_{t,p,s,g}}{price_{t-1,p,s,g}}$$

$$Where\ t = Month, p = 1,2,3, s = Station\ ID, g = Fuel\ Grade$$

Stations without any recorded observations in a given pricing period were assigned the average price relative across all stations in their respective county as an imputed value for the dependent variable.

2.2 Fixed Effects Linear Regression

We rely primarily on a Fixed-Effects Multivariate Linear Regression model as our primary method of analysis. Controlling for the fixed effects of month (pricing period) of observation, and fuel grade, we estimate a linear regression of of our dependent variable *price_relative* on the following independent variables:

- *log(prev_price)* – log of the price level of a given station in the previous month (*month t-1*)
- *log(population_density)* – log of the population density of the zip-code in which a station is located
- *top_tier* – binary variable denoting whether a station's gasoline is treated with TOP TIER Detergent
- *county*
- *nearest_station* – distance in meters from a given station to its nearest neighbor
- *log(home_value_2018)* – log of the median home value of the zip-code in which a station is located

2.3 Random Forest Validation

In an effort to validate the results of our fixed effects model, we also estimate a random forest model regression model; this allows us to observe the feature importance of our independent variables, and to determine whether both models considered the same variables most significant. We retain the same independent variables in estimating this other model, with one minor change. As we are only able to estimate our random forest regression over numerical data, we replace the County variable with a state level geographic variable – rather, state-tax on gasoline, a numerical value. This still captures a degree of geographic variation that is more digestible to these machine learning methods; the loss of granularity is acceptable, as this technique is only used in a validation capacity.

3. Results

3.1 Model Results

Our initial fixed-effects regression model found all of the above parameters significant in determining price change variation. The model also appeared to be highly explanatory, with an adjusted R-square value of 0.6714, though much of this variation is attributable to the time fixed effect. Previous month price level at a given station was found to be most significant among our independent variables in explaining price change variation for that station.

Estimating a random forest regression model for each month – and subsequently averaging the results – we confirmed the findings of our regression model. R-squared values for 10-fold cross validated models ranged from 0.45 to 0.65, and were veered higher in more volatile months. The random forest model also confirmed the finding that price level observed in the previous month was most significant in explaining price change in the current month; alone, it accounted for almost 40% feature importance in the model on average.

After determining that each of our independent variables were significant in explaining gasoline price variation, we sought to construct counterfactual CPI gasoline indexes; for each variable, we stratified our station index sample, with the constraint that the sample frequency distribution for said variable match that for the population. In a subsequent regression of station price level on our other independent variables, we found that each was quite significant in explaining *price level* as well as price change. As such, we considered each factor a proxy determinant of price level and price change, and excluded price level from this stratification exercise.

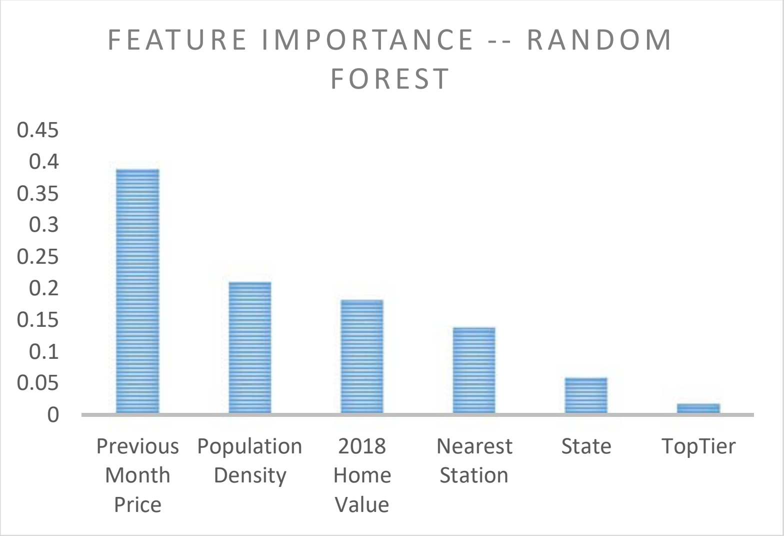
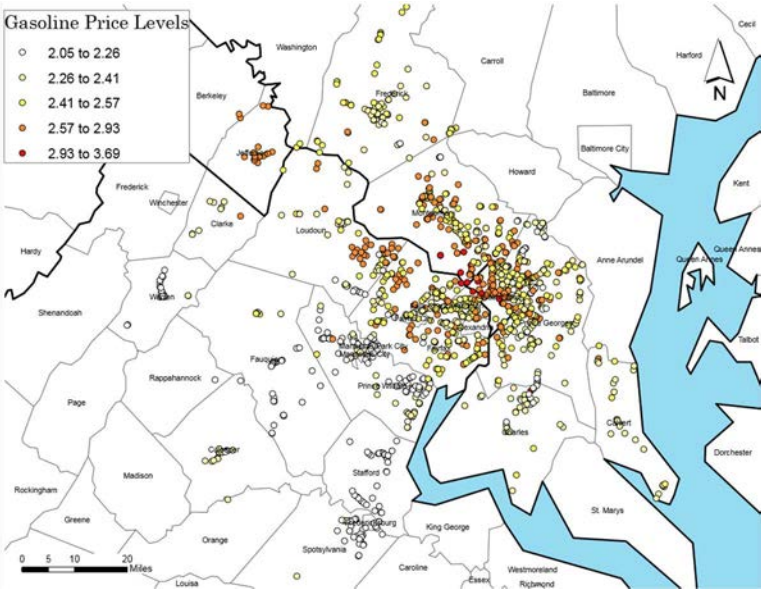


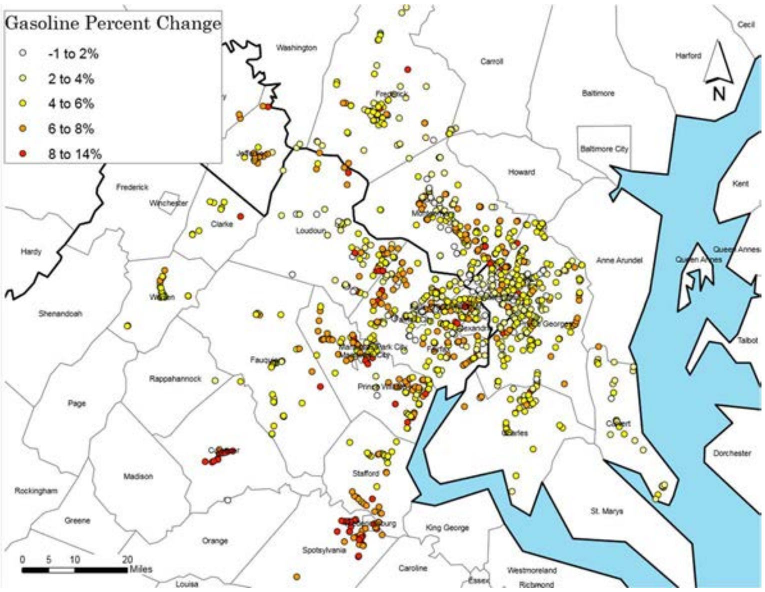
Figure 2: Feature importance of each variable in our random forest model, used to validate the findings of our fixed effects model.

3.2 An Explanation of Price Level as a Determinant of Price Change

We attempt to provide a possible economic explanation as to why price-level might be significant in explaining price change. Consider an example of two gasoline stations; a low-priced station, and a high-priced station. The data suggest that the low priced owner may operate under a market paradigm more closely resembling one of perfect competition; he is a price-taker, and sets prices in accordance with the cost of his inputs. Conversely, we assume that the high priced owner operates under a monopolistically competitive market framework; because her product has a degree of product differentiation – in the form of branding (and associated brand loyalty), more ideal location, etc. – she is able to set prices in accordance with her input costs, as well as the value ascribed to her differentiated product. In this sense, the high priced owner is a price-maker. As such, when a shock occurs in each of the station's input costs, the low-priced station owner faces a higher degree of exposure; this additional exposure to input cost shocks results in greater volatility among lower priced stations[3].



[3] We offer a more technical explanation of these concepts in appendix 2.



Figures 3 and 4: Arithmetic mean gasoline price level and percent change in April 2018 across the Washington, DC metro area.

3.3 Index Sampling Stratification

Our objective is not only to observe the determining factors of price change variation in an area, but also to examine whether controlling for said factors in an index methodology has a significant effect on the indexes produced. In this instance, we account for these factors in the area of sample selection; for each variable found to have significant effect on price change in the model, we create counterfactual indexes with a sample stratified on each variable. Among categorical variables – i.e. county, and TOP TIER branding – we assign a sample stratum for each category in accordance with its representation among the whole population of stations. For numerical variables – population density and median home value of a stations zip-code, as well as distance from a station to its nearest neighbor – we divide the population of stations into quantiles, and include an equal number of observations from each quantile in the sample.

3.4 Index Calculation Methodology

We attempt to adhere as closely as possible to CPI elementary index calculation methodology. There are approximately 50 stations reporting prices in the CPI gasoline sample for the Washington DC metro area. For each of these stations, we observe a station level price relative $R_{s,g}$, for each fuel grade; this is defined as the price observed at the station in the current period divided by that in the previous period. In accordance with CPI methodology, we calculate our area indexes using a simplified geometric means approach;

that is, the price relative R for the DC metro area a at-large, from period *t-1* to t , is equal to the following:

$$R_{[t,t-1],a} = \prod_{s,g\in a} (\frac{P_{s,g,t}}{P_{s,g,t-1}})^{\frac{W_{s,g}}{\Sigma W}}$$

Where:

$$P_{s,g,t} = observed\ price\ of\ fuel\ grade\ g\ at\ station\ s\ at\ time\ t$$

And

$$W_{s,g} = weighted\ expenditure\ for\ grade\ g\ at\ station\ s$$

Weighted expenditures are derived using estimates of the total daily expenditure on an item in a given area as previously[4] reported in the TPOPS survey; in this case, gasoline in the DC metro area. For each observation's, we divide this total daily expenditure by the number of observations and then adjust it according to the proportion of expenditure on each item subgroup – in this case, individual fuel grade – in relation to total expenditure on the entire item.

3.5 Index Results: Baseline

We calculate a baseline index using the methodology above, across all stations in the CPI gasoline sample for Washington DC; after matching the CPI sample with observations in the GasBuddy dataset, we substitute CPI collected prices with their respective GasBuddy prices. This index serves as a basis of comparison with all subsequent stratified indexes. By calculating this baseline index we ensure a *ceteris paribus* comparison; the only difference between calculation the baseline and respective stratified indexes is the sampling methodology.

[4] The TPOPS was discontinued in 2019.

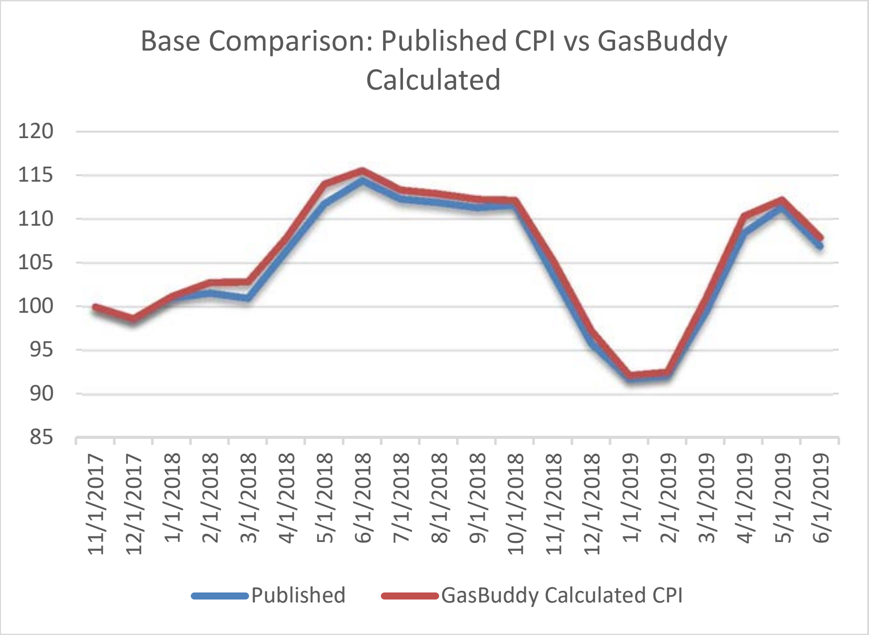


Figure 5: Comparison of the GasBuddy calculated CPI for Gasoline in DC with published data

For each of our 5 stratification criteria, 1000 stratified samples of 50 stations – the approximate size of the CPI sample – are selected from the GasBuddy population of stations in DC. Each of these samples is used to calculate monthly index relatives across our period of observation, and the mean relative across each month is used in calculation of the stratified index. Finally, a paired t-test between the baseline and stratified methodology is used to determine whether the stratified index is significantly different.

3.6 TOP TIER Stratified

Our analysis found that the CPI has more[5] TOP TIER branded stations in the DC Metro area than the GasBuddy sample. While approximately 65% of GasBuddy stations in the DC metro area TOP TIER branded, they account for approximately 80% of the CPI station sample.

[5] See the Appendix for comparisons of frequency distributions across all stratification variables.

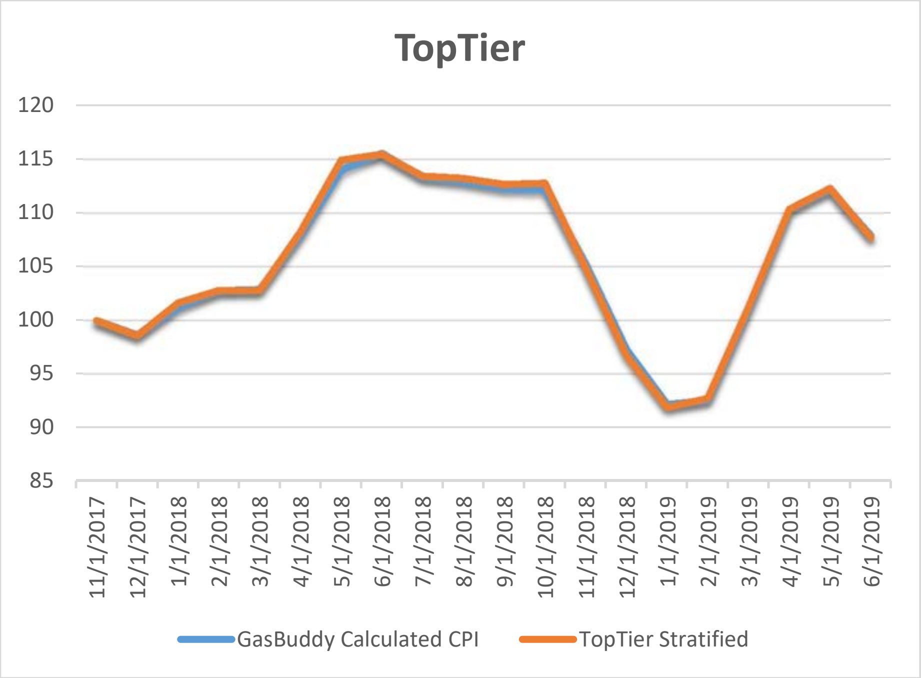


Figure 6: Comparison of the GasBuddy calculated baseline CPI with that stratified on TOP TIER brand status.

A comparison against the baseline result using a paired-t test failed to reject the null hypothesis that relatives calculated using a TOP TIER stratified sample were not significantly different from the baseline. Stratifying on brand, as represented by TOP TIER was not shown to have a significant impact on relatives.

3.6 County Level Stratification

The CPI sample has more stations populous counties near the center of the DC metro area, and fewer stations in counties that are further away, even when such counties contain significant proportions of the population of stations. Two outlying counties, for example, comprise approximately 10% of total GasBuddy stations in the DC metro, but there are no stations from these counties is in the CPI sample. Our stratified sample ensures at least one station in each county is selected. Under this stratification, we select a minimum of one station from each county, and the rest are allocated proportionally.

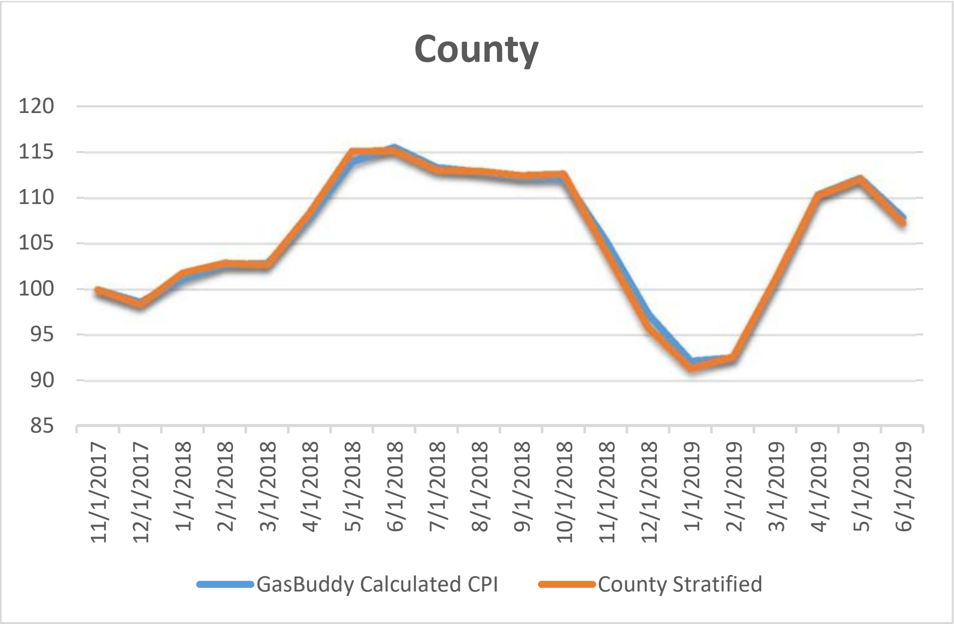


Figure 7: Comparison of the GasBuddy calculated baseline CPI with that stratified on County.

A comparison against the baseline result using a paired-t test failed to reject the null hypothesis that relatives calculated using a county stratified sample were not significantly different from the baseline. Stratifying on county was not shown to have a significant impact on relatives.

3.6 Population Density Stratification

When considering population density, the CPI sample appeared to only differ markedly from its GasBuddy counterpart in zip-codes with lower population densities. High-population density areas showed roughly equal representation across samples. On the other hand, the GasBuddy sample had a higher concentration of stations in low population density areas than its CPI counterpart.

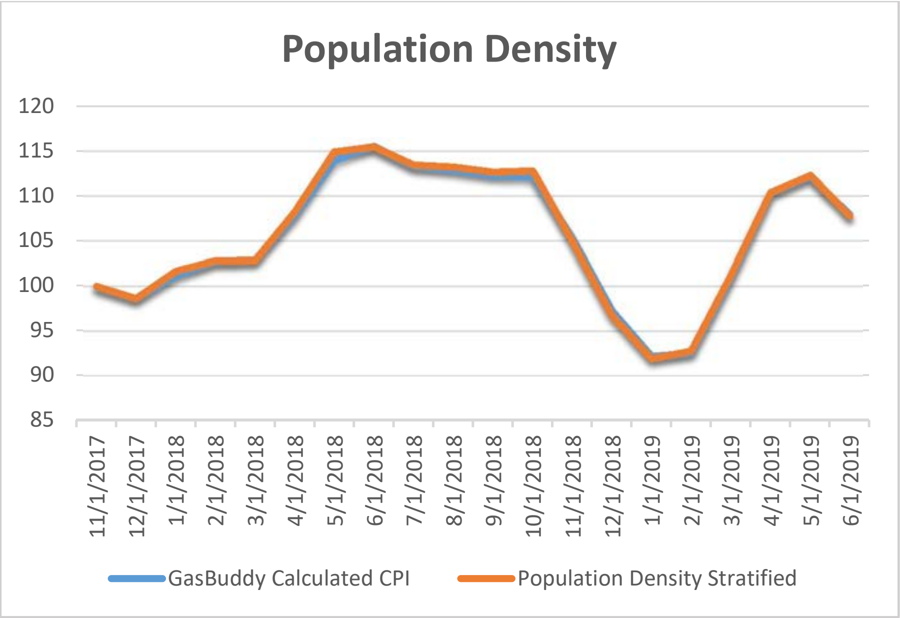


Figure 8: Comparison of the GasBuddy calculated baseline CPI with that stratified on Population Density of the station's zip-code

A comparison against the baseline result using a paired-t test failed to reject the null hypothesis that relatives calculated using a population density stratified sample were not significantly different from the baseline. Stratifying samples on population density was not shown to have a significant impact on relatives.

3.7 Nearest Station Stratification

When comparing them on the basis of station isolation – distance from its nearest neighbor – the CPI and GasBuddy samples were largely convergent. A comparison against the baseline result using a paired-t test failed to reject the null hypothesis that relatives calculated using a station-distance stratified sample were not significantly different from the baseline. Stratifying samples on the distance from a station to its nearest neighbor was not shown to have a significant impact on relatives.

3.8 Home Value Stratification

The CPI sample was has relatively more stations in zip-codes with high home values, and relatively fewer stations in zip-codes with low home values. For example, stations in zip-codes with median home values in the interval $750,000-$1,000,000 occurred more than twice as frequently in the CPI sample then among the GasBuddy sample; conversely, stations in zip-codes with median home values between $100,000-$300,000 occurred only half as frequently in the CPI when compared with the GasBuddy sample.

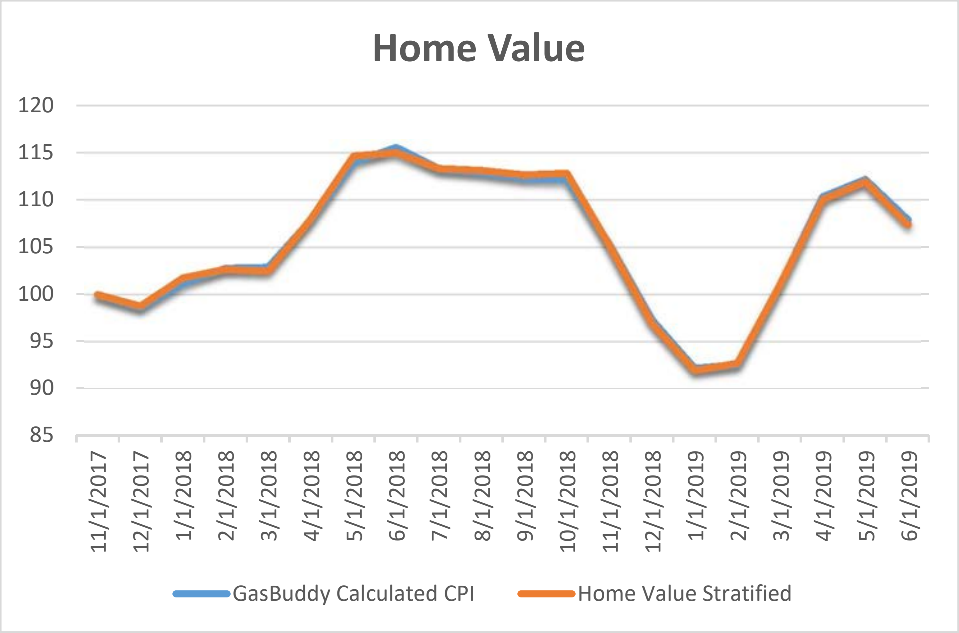


Figure 9: Comparison of the GasBuddy calculated baseline CPI with that stratified on Median Home Value of the station's zip-code

A comparison against the baseline result using a paired-t test failed to reject the null hypothesis that relatives calculated using a home value stratified sample were not significantly different from the baseline. Stratifying samples on home value was not shown to have a significant impact on relatives.'

3.9 A Possible Explanation of the Results

None of the indexes constructed using stratified samples were found to be significant from the index using the baseline CPI sample; this result was surprising. As each of our models proved significant in terms of explanatory power and individual variable significance, we expected a more significant differences in the resultant indexes. We attempted to explain this discrepancy by revising our initial fixed-effects model. Rather than controlling for the fixed-effect of each 10 day period of observation, we substituted the change in the price of WTI crude oil over the first 5 days of said period. As all observations experienced the same change in the spot price of crude oil in any given time period. The Energy Information Administration (EIA) estimates that crude oil accounts for 52% of the price of gasoline[6]. As such, its observed prices movement serves as an excellent quasi fixed effect (though observable contemporaneously), and also adds additional info to the model; it allows for

[6] As of July 2019, Energy Information Administration, Gasoline and Diesel Fuel Update -- https://www.eia.gov/petroleum/gasdiesel/

us to observe the significance of parameters in our initial model *relative* to a characteristic that is known to be highly significant.

The resultant regression model behaved as we expected; while the overall significance of our model was diminished – a result of removing the time fixed effect – the oil price change variable was found to be more significant and of much higher magnitude than any of the other parameters.

We confirmed this result by recalculating our random forest model with the inclusion of oil price change. In a similar fashion to the regression model, this inclusion dwarfed the respective feature importances of all the other parameters, accounting for over 82% of price change variation; whereas previous month's gasoline price accounted almost 40% of price change variation in the initial model, it only accounted for 14% in the revised mode. The importance of other variables were also similarly diminished.

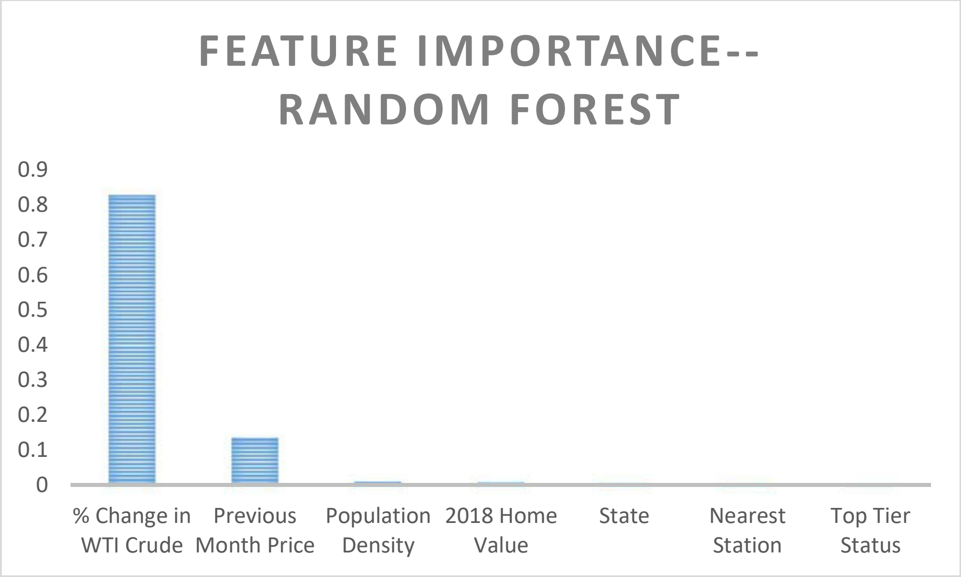


Figure 10: Feature importance of each variable in our random forest model, with the inclusion of percent change in crude oil for each period of observation.

4. Conclusions

Our objective was to determine the sources of variation in gasoline prices in a single metropolitan area. To achieve this, we examined price movements across the DC metropolitan area from November 2017 through June 2019. Using a fixed effects model – controlling for period of observation and fuel grade – we examined the effect of geography (e.g., county, population density, median home value), and individual station characteristics (brand, distance to a stations nearest neighbor, previous month price level at a station) on individual station price movements in a given month. We found that all of the above considerations were significant in explaining price change variation, most significantly a stations previous month price; namely, higher priced gasoline stations tended to experience less volatile price movements. Nevertheless, the non-price determinants of price change also were determinants of price level, prompting us to exclude price level from subsequent index analysis.

For each of our significant independent variables, we constructed counterfactual indexes using each variable as a stratification criterion; we constrained each of our counterfactual samples so that they would match the distribution of the variable across the station population. Indexes constructed using these stratification showed no significant differences when compared with the baseline CPI index. Finally, to understand this, we re-ran our model including observed change in the price of oil as an additional independent variable. This inclusion dwarfed the significance and explanatory power of the other variables, and allowed us to reach a final conclusion: that it is possible to observe *cross-sectional* - variation in gas price movements at a given point in time, but the significance of these determinants are conditional on the observed price movement of crude oil – the underlying commodity from which it is derived. Gas prices move together, in tandem with the price movement of their inputs; only minor cross-sectional variations due to geography, brand, or other station characteristics can be explained at a given point in time.

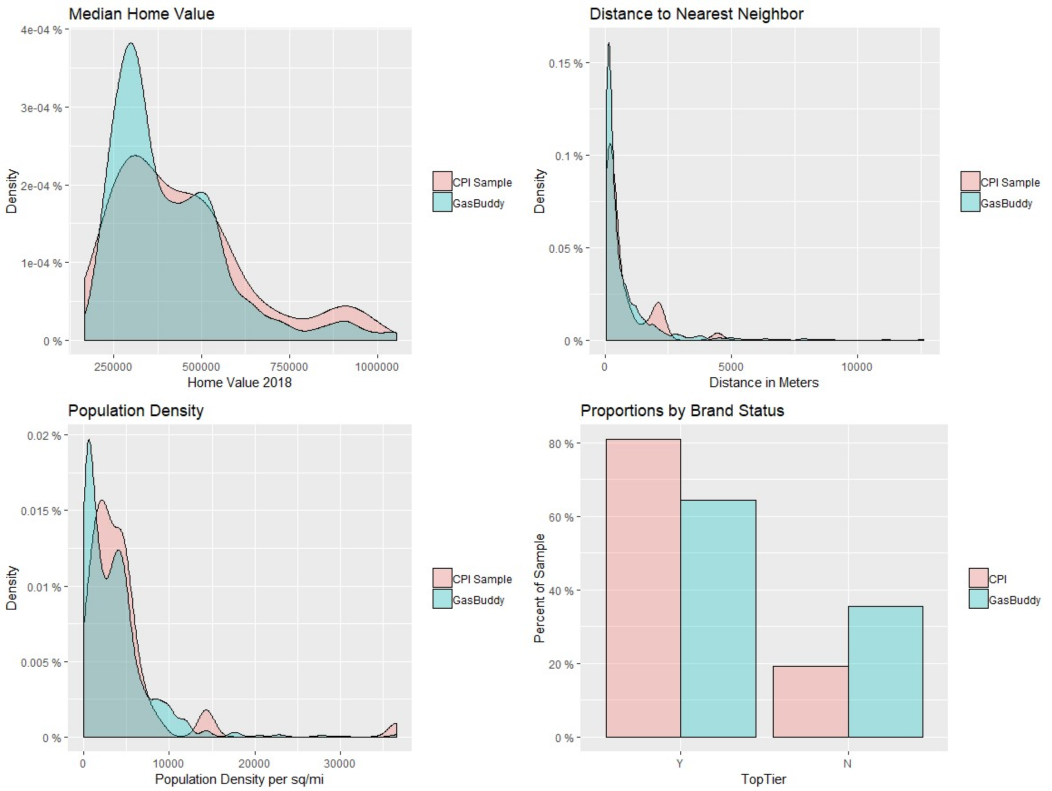
References

Bieler, John. Niedergall, Sarah. Popko, David. Sung, Ilmo. *Alternative Data for CPI Motor Fuels: Creating Gasoline and Diesel Indexes using GasBuddy Data*. 2019

U.S. Bureau of Labor Statistics. *BLS Handbook of Methods, Chapter 17. The Consumer Price Index,* pages 18-19. 18 April 2019

Liaw, Andy. Wiener, Matthew. *Classification and Regression by randomForest.* December 2002

Appendix 1: Frequency distribution comparison of the CPI sample and GasBuddy station population.



Appendix 2:

Here we offer a more technical explanation of the concepts in section 3.2. Once again beginning with our two station model, we assume that the low-priced station operates under a near perfectly competitive market structure; he is a price taker, his price equaling, and fluctuating with his input costs m – i.e. crude oil:

$$P = m$$

Conversely, the high-priced station owner operates under a paradigm of monopolistic competition; by assumption, her station may exhibit a degree of product differentiation – in the form of better location, branding, etc. This differentiation justifies the discrepancy in price, and allows her to operate under a paradigm of monopolistic competition. Unlike the owner of our low-priced station, who faces a horizontal (i.e. perfectly elastic) demand curve, the high-priced station owner faces a downward sloping demand curve, and is allowed to set her own prices. Under this assumption, her profit equals the following:

$$\pi = PQ - mQ - F$$

Where:

$$P = price, Q = quantity, m = input\ cost, F = fixed\ cost$$

Her price is defined by the downward sloping demand curve, parameterized by *a* and *b*:

$$P = a - bQ$$

To obtain the profit maximizing quantity *Q,* we take the derivative of the profit function with respect to quantity, and set it to 0:

$$\max_{Q}\ (a - bQ)Q - mQ - F = \max_{Q} \pi(Q)$$

$$0 = \frac{d\pi}{dQ} = a - 2bQ - m$$

Allowing us to realize our profit maximizing quantity:

$$Q = \frac{a - m}{2b}$$

Substituting this back into the original price function, defined by the demand curve, we see that the high-priced station owner sets her prices according to the following:

$$P = a - bQ = \frac{a}{2b} + \frac{m}{2}$$

Where *a* and *b* are assumed fixed.

Under this paradigm, we observe that, in the case of the low priced station owner, 100 percent of shocks input costs are passed on to the consumer; on the other hand, input costs comprise a smaller portion of the high priced stations total price, and fluctuations exhibit a dampened effect. This offers a possible explanation as to why high priced gasoline stations exhibit lower volatility price movements in the subsequent month.