Consumer Search with Learning: A Structural Estimation of Gasoline Demand

Xiaosong Wu*
Frank A. Wolak†

November 09, 2018

Abstract

This paper estimates demand for gasoline in the presence of two types of imperfect price information: ex-ante uncertainty about each station's price, and uncertainty about the distribution of all stations' prices. We develop a model in which consumers formulate their prior belief of the current price distribution using the prices observed on past driving trips, and then Bayesian update their beliefs with each new price observed, before deciding whether to purchase gasoline or continue searching for a cheaper price. We estimate this model by utilizing a unique data set of station-level daily gasoline sales and prices, combined with data on the empirical distribution of various traffic flows. Our empirical results suggest that consumers place a relatively high weight on new price observations when formulating their beliefs of the overall price distribution. We find that price distribution uncertainty is the primary component of imperfect price information, and if it were eliminated, consumers could achieve 70 percent of the total savings that could be realized by having perfect price information.

*Job market candidate. John E. Walker Department of Economics, Clemson University, 227 Sirrine Hall, Clemson, SC 29634. Email: xiaosow@clemson.edu.

†Program on Energy and Sustainable Development and Department of Economics, Stanford University, Stanford, CA 94305-6072, Email: wolak@stanford.edu.

We appreciate helpful feedback from Matthew Lewis, Babur De los Santos, Patrick Warren, Christy Zhou, Jonathan Ernest and the participants and faculty in the Clemson IO workshop.

1 Introduction

Many markets are characterized by frequent changes in prices due to rapidly varying market conditions. In these cases, (i) ex-ante uncertainty about each seller's price, and perhaps more importantly, (ii) uncertainty about the distribution of all sellers' prices arise, causing market inefficiencies, including differences in the asking prices among sellers even for homogeneous goods. For example, consumers may not buy from the least expensive seller if it is too costly for them to keep searching for the lowest price or if they are unaware of the existence of this lowest price. When both types of uncertainty are present, consumers engage in costly search not only to realize price offered by sellers (searching), but also to update their beliefs about the current price distribution (learning). While there is rich theoretical and empirical literature that has focused on consumer price search, it has largely ignored the second type of uncertainty by assuming consumers know the current price distribution.[1] Moreover, the few empirical papers that have incorporated both types of uncertainty have not estimated the consumer learning process.[2]

In this study we investigate consumer search with learning behavior in the context of the retail gasoline market. Two salient features, which are depicted in Figure 1, make the retail gasoline market an ideal place for analyzing this phenomenon. First, frequent price changes resulting from a volatile wholesale cost make it difficult for consumers to maintain accurate price information for each station as well as the distribution of these prices.[3] Second,

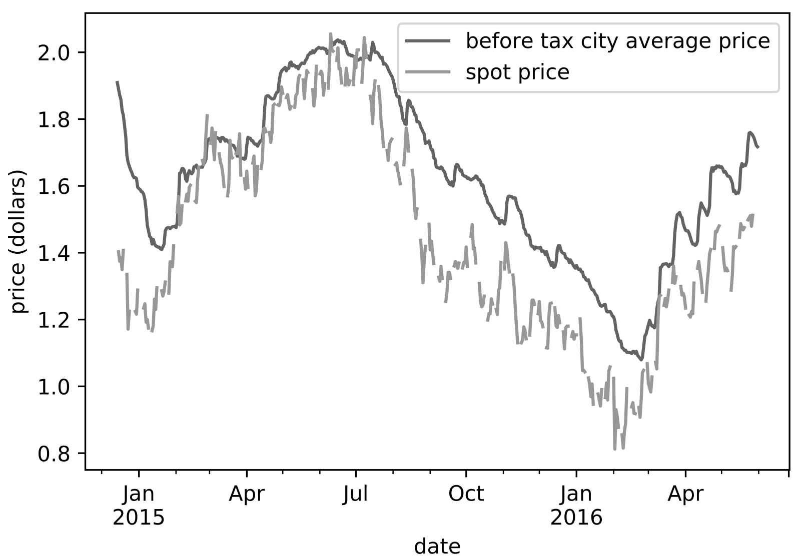
[1]Pioneer works, such as Stigler (1961) and McCall (1970), among others have established the theoretical foundation for models of consumer search and emphasize the importance of search frictions in explaining imperfectly competitive market behavior under the assumption that consumers know the price distribution. Under this key assumption, a large number of empirical papers have estimated the search costs that contribute to sustained price dispersion in various market, for example, Hong and Shum (2006) on online textbooks; Hortaçsu and Syverson (2004) on S&P index funds; Wildenbeest (2011) on supermarkets; Koulayev (2014) on hotel bookings online; Honka (2014) on auto insurance; Moraga-Gonzalez et al. (2015) on automobile markets; De los Santos et al. (2012) and De los Santos (2018) on online books; Nishida and Remer (2018) on retail gasoline markets.

[2]Rothschild (1974) was the first to develop a theoretical model in which consumers search and learn about the price distribution in the process. To our knowledge, Koulayev (2013) and De los Santos et al. (2017) are the only two papers that integrate consumer search with learning in empirical work. However, in both papers, parameters that determine the consumer learning process are not estimated but chosen by the researchers.

[3]Although we focus on how changing market conditions cause uncertainty about the price distribution

the uncertainty of consumers' beliefs about the current price distribution is often used to explain the asymmetric price response to changes in cost (Lewis, 2011). In particular, when consumers do not know the current price distribution, they may form their initial perception based on prices observed during past driving. In this case, when prices and costs decrease, consumers' beliefs of the price distribution tend to be too high, resulting in low expected gains and less searching, which, in turn, causes higher profit margins for the seller. If consumers follow this behavior, then assuming that they know the current price distribution will lead to upward-biased estimates of search costs (either high search costs or low expected gains from this search can contribute to the low levels of search). Consequently, welfare analysis may overstate the benefits of policies that reduce information uncertainty or restrict firm entry in an attempt to lessen the total costs of the search process.

Figure 1: Average City Gasoline Price and Cost



Notes: This figure plots both the average gasoline price in our sample before federal and state taxes are applied and the Gulf Coast regular spot price as a measure of wholesale cost of retail gasoline.

In this paper, we propose a model of search with learning that builds on Rothschild

in this paper, inexperience about the market can also contribute to this uncertainty. Using survey data, Matsumoto and Spence (2016) find that college students who have no prior experience in purchasing textbooks online tend to expect online prices to be higher than what are observed empirically, and consumers with more experience in the marketplace generally have more accurate beliefs about the price distribution, which is consistent with learning.

(1974), with an emphasis on spatial and ex-ante vertical differentiation of sellers. We apply the model to a retail gasoline market where consumers are assumed to formulate their prior beliefs on the current price distribution using the prices observed in the past, and then Bayesian update their beliefs with each new price observed along their predetermined travel routes, before deciding whether to purchase gasoline or continue searching. Consumers maximize their utility from a gasoline purchase by comparing if the realized utility at each station after observing its price is higher than the expected value of the remaining stations on the route and the alternative of waiting to purchase during a future trip. If consumers observe a price higher than expected price at a station, they continue driving since they expect to find lower prices at their remaining options on the route. However, learning influences consumers' purchase decisions because as they sample additional high prices, and the newly obtained price information begins to outweigh their prior beliefs, their posterior beliefs of the actual price distribution are adjusted upward. They are now more likely to make a purchase as they believe that prices at other stations are higher.

To estimate the model, this study uses a novel dataset of daily transactions from each gasoline station in a city from December 14, 2014, to May 31, 2016, and combine it with data on daily prices and an empirical distribution of travel routes. We model a station's daily gasoline sales as an aggregation of purchase decisions made by individuals who are searching and learning along their predetermined travel routes. Our estimation result suggests that consumers update their beliefs of the current price distribution rather quickly with each newly observed price. This trend is consistent with the fact that the distribution of gasoline prices changes regularly, and as a result, prices sampled today are more informative about the current price distribution than past price observations. This modeling of the learning process not only allows us to estimate a more realistic and flexible formation of consumers' beliefs, but also enables us to answer the question of how much consumers would benefit if they knew the current price distribution instead of assuming it is currently the situation. After controlling for the spatial and vertical differentiation of gasoline stations, counterfactual analyses suggest that consumers would save 1.12 cents per gallon of gasoline purchased if uncertainty about the price distribution were removed, and they would save 1.60 cents per gallon if they had perfect information ex-ante about prices. Although the estimated savings are small, they are non-trivial, since gasoline demand at the station level is highly price sensitive, with an estimated elasticity on the order of -9. Moreover, it suggests that uncertainty about the actual price distribution is the primary component of imperfect information. For instance, providing consumers with information about the current price distribution alone can result in a realization of 70 percent of the total saving if they were to know all of the prices.

Our approach is also unique as it uses data on consumers' travel patterns to construct the distribution of the consumers' search sequence of gasoline stations. Most empirical papers that structurally estimate consumer search models using only market-level data assume consumers' search sequence to be random because individuals' search histories are unobserved. However, travel patterns naturally constrain the search order when sellers have physical addresses.[4] Even for sellers without physical addresses, the order of visits can be affected by such constraints as the layout of a web-page.[5] By taking into account traffic data, we are able to improve upon the assumption of random sampling and replace it with observable variation, allowing the model to identify more realistic substitution patterns. While random sampling implies homogeneous substitution patterns between any two sellers, we find that 35 percent of the variation in the estimated elasticity of substitutions is explained by the number of common drivers shared between two stations. In other words, two stations are close competitors if they share a large number of common drivers, regardless of the distance between them. One station's price may not affect another's residual demand if they do not share any common drivers.

Our panel data allow us to estimate separately parameters that govern the consumer

[4] Arbatskaya (2007) develops a theoretical model of ordered consumer search, finding that sellers' prices and profits are affected by the order of search.

[5] As evidenced by Kim et al. (2010), if two products sold online are not being searched together, the cross-price elasticities are numerically zero. There is also growing literature focusing on improving online platform design to reduce search frictions and increase competition (Dinerstein et al., 2018).

learning process and the distribution of the costs for postponing purchase. As volatile wholesale costs change the distribution of prices from day-to-day, variation in consumers' prior beliefs, which are based on past observed prices, is generated. The variation in consumers' prior beliefs across days leads to variation in consumers' purchase behaviors across days that is independent of the distribution of the costs for postponing purchase. The weight consumers place on their prior beliefs determines how quickly their posterior beliefs are updated with the newly observed price information. Our model identifies this weight based on how consumers' purchase behavior today is affected by the information obtained in the past.

The contributions of this paper are threefold. First, it is the first paper to estimate how quickly consumers update their prior beliefs of the unknown price distribution without consumer-level search information. Second, by doing so, we are able to quantify the additional amount of market inefficiency created by the uncertainty of the price distribution, an outcome of consumers searching too much or too little compared to the optimum if they knew the actual price distribution. Third, this paper uses data on consumer travel patterns to simulate their search histories in a consumer search framework, thus relaxing the assumption of random sampling widely adopted by search literature when search histories are not observed. Although this feature is most relevant for the retail gasoline market, as more data about consumers' GPS movements data become available, they can be utilized to formalize how consumers search when their individual search histories have not been observed for a variety of markets.

2 Data

In the absence of micro-data of consumers' search histories of gasoline stations and purchase decisions, we use market-level data to make inferences about consumers' search and learning behavior that leads to gasoline purchase. Our sample consists of 47 gasoline stations in a mid-sized city, with a population of approximately 75,000 in the urbanized area.[6,7] The sample period runs from December 14, 2014, to May 31, 2016, for a total of 535 days, during which time we observe daily price of gasoline at all 47 stations and the daily gasoline transaction volume for 35 of these stations. We combine gasoline data with data on traffic flows for our estimation. Under the assumption that consumers purchase gasoline along their predetermined travel route, daily transactions at a gasoline station, as an outcome of individual purchase decisions, provide information about their searching and learning behavior concerning prices. The empirical distribution of these travel routes are constructed based on the data on traffic flows. In the following subsections, we describe the three main sources of data used for our empirical analysis.

2.1 Gasoline Price Data

We collected per gallon prices of regular unleaded gasoline from two gasoline price gathering websites. The primary source was MapQuest.com, an online web mapping service whose gasoline price data are provided by Oil Price Information Service (OPIS).[8] We visited MapQuest.com daily and web-scraped the prices for every station in the city. However, not every station's price was updated daily by the website. For an average station in the city, a new price was updated on 54 percent of the days.[9] To address the issue of missing prices, we complemented MapQuest.com's data with the price data collected daily from

[6]City name is not disclosed for the protection of the identities of the gas stations.

[7]Since the main focus of the paper is to study consumers' search behavior in the gasoline market, we excluded 14 stations from the city that do not meet the criteria of a station that competes for gasoline sales among the searching consumers. These criteria include a large price board where gasoline prices are distinctively visible and advertised to consumers, a large forecourt, and easy accessible entrance and exit. Almost all excluded gasoline stations are mom-and-pop stations with a small gasoline sales volume and primarily rely on sales inside the convenience store. Seven of eight excluded stations with matching transaction data have gasoline sales less than or at the same level as the smallest volume stations in the sample.

[8]OPIS obtains price information from credit card transactions and direct feeds from gas stations.

[9]The price coverage rate is slightly lower than other studies that directly use OPIS price data. One possible reason is that the city of interest is a mid-sized city with a higher share of low-volume stations than the major cities studied by other papers. Fewer credit card transactions result in fewer price feeds to OPIS.

GasBuddy.com.[10] Unlike MapQuest.com, prices on GasBuddy.com are uploaded by volunteer spotters in the area. To minimize any issues caused by the potential inaccuracy of the prices reported on GasBuddy.com, we filled in only the station-day missing MapQuest.com price with prices from GasBuddy.com.[11] The matching of stations across two data sources was based on the geometric coordinates of the stations, cross-validated with Google Map's geometric coordinates to ensure accuracy.[12] After merging the price data, a station has, on average, missing prices for 9.2 percent of the sample days. We filled in these missing prices with the most recent price observed at that station. The average time lapse for which we imputed price was 1.6 days.[13] In addition to price data, we also obtained information on station characteristics, including name, brand, address and geometric coordinates.

2.2 Gasoline Transaction Data

Daily expenditure data at the station level were obtained from a major financial services provider.[14] These data reflect the total dollar amount of purchases from all debit and credit card users of this provider at each gas station on each day. We obtained two types of expenditure data from the card users for 35 of the 47 stations, pay-at-pump and in-store purchases. However, to eliminate the measurement error caused by non-gasoline transactions, we use only pay-at-pump transactions. We construct a daily measure of the total quantity of gasoline purchased at each station by dividing the total pay-at-pump expenditures by the price of regular unleaded gasoline at the station on that day.[15] We believe that our measure

[10] Gasoline price data collected from MapQuest.com and GasBuddy.com are widely used in the literature on retail gasoline prices, for example, Lewis and Marvel (2011) and Remer (2015).

[11] Atkinson (2008) shows that although occasional error of the published prices may occur on GasBuddy.com, these prices are still able to identify the features of retail gasoline price competition accurately.

[12] A station's name or address cannot be used as a unique identifier for the matching because a station's name is not unique to a station and different websites may use different aliases for a street or highway.

[13] In our structural estimation, missing prices cannot be accommodated, because a station's price affects many stations' sales through the traffic network, and one missing price will result in a large number of lost observations.

[14] The name of the provider as well as the names and locations of stations in the data have been withheld to protect confidentiality.

[15] This calculation introduces potential measurement error, as it overestimates the quantity transacted for mid-grade and premium gasoline, which have higher prices. However, it has been estimated that only 15

of the total quantity of gasoline transacted at each station provides a good representation of the behavior of consumers searching for gasoline.

2.3 Empirical Distribution of Search Routes

As individuals drive along their travel routes, the decision to purchase gasoline at a particular station is affected by the prices observed prior to this station as well as the price expected for the remaining stations on the route. As a result, the search sequence of stations is essential for our analysis of consumers' searching and learning behavior in the market. In this study, we define a search route as a unique set of an ordered sequence of stations visited. In our model of consumer search with learning, an individual's decision is then modeled at the search route level. We constructed an empirical distribution of search routes describing the predicted number of consumers to drive past a specific ordered sequence of stations on an average day.

First, we constructed the empirical distribution of the travel routes comprising two elements: (i) the number of drivers traveling from an origin to a destination, and (ii) the route that drivers took along the street network connecting the two points. For the first element, we obtained the Origin and Destination Table from the state's Department of Transportation,[16] which contains information about the estimated number of drivers traveling from one Traffic Analysis Zone (TAZ) to another TAZ. This Table covers the entire county that includes the city of interest as well as six additional surrounding counties. The entire area is divided into approximately 1800 TAZs, each also associated with census information about the residents living in the zone. Table 1 summarizes the TAZ and the associated demographics of the residents in the county where the city of interest is located. We can see that the TAZs are extremely fine, with 75 percent of the traffic zones occupying an area of less than 1.5 km^2. The larger traffic zones are at the fringe of the county with few residents.

percent of gasoline transactions are mid-grade or premium.

[16]The Origin and Destination Table is an output of the travel demand model constructed by the Department of Transportation to forecast the traffic in year 2020.

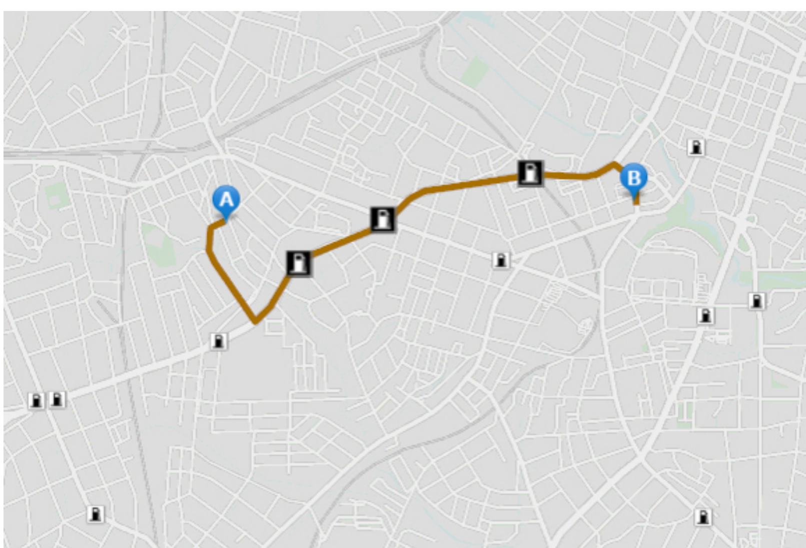
Table 1: Summary Statistics of the Traffic Analysis Zones (TAZs)

	mean	std	25%	50%	75%
Area size (km^2)	2.71	6.64	0.13	0.49	1.49
Population	668.31	920.13	141.50	357.00	733.00
Pop income > 50k	115.62	191.79	11.62	46.46	127.04

Note: The number of Traffic Analysis Zones is around 300. The exact number is not reported for the protection of the identity of the city.

To compute the route that drivers take traveling from an origin TAZ to a destination TAZ, we assume that all drivers take the travel route that minimizes driving time. We selected the centroid of the origin and destination TAZ as the drivers' start and end locations, and for every origin and destination TAZ pair, a single fastest travel route following the street network in the area was generated. We used the ArcGIS Network Analyst package to calculate the fastest travel route and the road network data set was obtained from ArcGIS StreetMap North America.[17,18]

Figure 2: A Travel Route with Stations Passed



Note: This figure represents the road network of a random location. The driving time is less than 5 minutes.

[17] ArcGIS Network Analyst extension: `https://www.esri.com/en-us/arcgis/products/arcgis-network-analyst/overview`. StreetMap North America: `https://www.arcgis.com/home/group.html?id=ddd06a0bde9c45a1b3e786a2b4e695e8#overview`.

[18] To reduce the computation burden, we grouped the TAZs in each of the surrounding counties into 8 clusters of TAZs based on their locations using the K-Mean algorithm.

Second, we further found the identities of the stations on each travel route and the order of visits, by ranking driving distances of the stations to the origin TAZ in ascending order. Figure 2 provides an example of a travel route connecting a start location A and an end location B, including the three stations visited by drivers on this route. The set of an ordered sequence of stations formalizes a search route. Because difference travel routes may go through the same set of stations in the same order, we further collapse the travel routes into a search route level. We obtained the empirical distribution of search routes by summing up the predicted number of drivers driving past the same search route. We then excluded search routes with fewer than 20 drivers traveling per day. As a result, a search route abstracts from any travel time and distance and only contains two elements: (i) a unique set of ordered sequence of stations, and (ii) the number of drivers passing through it along with their demographics.[19]

Table 2 shows the summary statistics for the resulting travel routes and search routes. The median travel time of a travel route is 22 minutes, and a search route involves 5.3 stations and 289 drivers on average. In addition, the average percentage of drivers with an annual income of more than $50,000 is 17.98 percent on a search route.[20]

Even though our data have many advantages, we discuss a few limitations. First, only a limited number of stations are matched with transaction data. In the estimation, although all stations' quantities transacted are predicted by the model, identification is based on the stations with quantity data observed. Due to the limited variation across stations, we can only estimate a small number of station characteristics that contribute to the vertical differentiation of stations. For this reason as well as to protect each retailer's identity, stations are broadly grouped into three categories: major-brand retailers, large-format independent

[19]Among the 47 stations, 2 stations entered the market at different times in the middle of our sample period. We generate 796 search routes before any entry, 810 search routes after the first entry, and 841 search routes after both entries. In the structural estimation, the sample periods are divided into three parts based on the date of entry and the empirical analysis is based on the empirical distribution of search routes in each period respectively.

[20]This number is calculated using the total population whose income is greater than $50,000 divided by the total population in all origin and destination TAZs that comprise a search route.

Table 2: Summary Statistics of Travel Route and Search Route

	count	mean	std	25%	50%	75%
	Travel Routes					
Total minutes	86099	29.80	22.99	13.00	22.12	38.42
	Search Routes					
N. stations	841	5.30	3.01	3.00	5.00	7.00
N. drivers	841	288.72	615.91	43.70	102.39	266.39
Income > 50k (%)	841	17.98	3.76	16.02	17.37	20.53

Note: Approximately 20 percent of the travel routes that have no stations are excluded.

retailers, and other independent retailers.[21] Second, the Origin and Destination Table describes only the travel patterns of local drivers, and it does not include information on the number of outside drivers traveling on the interstate highway, meaning a large proportion of demand at the stations near the exits of the interstate highway is not measured in our data. In addition, as these outside drivers do not observe gasoline prices along their travel route (outside drivers have to deviate from their travel route, i.e., exit the highway, to sample prices), their purchase behavior is different from the focus of this paper. To account for this issue, in our estimation, we use a dummy for each of the stations located at the exit of the interstate highway to adjust for the difference between the quantity predicted in the model and the quantity observed for these stations.

3 Retail Gasoline Market Overview

Before introducing the structural model, it is valuable to discuss the features in the retail gasoline market that motivate our modeling choice. More specifically, we first examine the relationship between the average transaction volumes of the stations and their price reputation and location. Then, we discuss the sources that result in consumers' imperfect

[21]Major branded retailers are branded by the refinery companies that exclusively supply them gasoline. Large-format independent retailers have large convenience stores and many pumps. Other independent retailers are stations that do not have a large-format presence.

price information and highlight that a search with learning model is more appropriate for formalizing consumers' behavior than the standard consumer price search model in this market.

3.1 Relationship Between Average Station Transaction Volume and Station Characteristics

Table 3 describes the distribution of the station average price for all 47 stations in the city and the distribution of the station average quantity transacted for the 35 stations with matching gasoline transaction data over the sample period. The first two rows present the characteristics of all stations in the city. The average gasoline price in our sample is $1.97 per gallon, while the average volume transacted at a station is slightly below 1000 gallons per day. The interquartile range of the station average price is 10 cents per gallon, accounting for approximately 5 percent of the average price.[22] In contrast, considerable heterogeneity is found in the average quantity transacted across stations, with the station at the third quartile being six times the size of the station at the first quartile. The remaining four rows show the distribution for the large-format independent and the major-branded retailers. Comparing the distribution of the stations in each of the two categories with the distribution of all stations in the city, respectively, we can see that there is a negative correlation between stations' average price and average transaction volume, consistent with consumers preferring stations that tend to have lower prices. All large-format independents in our sample have average prices in the lowest quartile, while their quantity transacted is in the highest quartile. On the other hand, the average prices for all the majors are dispersed in the upper three quartiles, with their sales in the lower three quartiles.

In addition to a negative relationship between the average price and the average transaction volume of the stations, Figure 3 depicts a positive correlation between the daily traffic

[22]Interquartile range is calculated by the difference between the 25th percentile and the 75th percentile.

Table 3: Summary Statistics of Station Average Price and Quantity Transacted

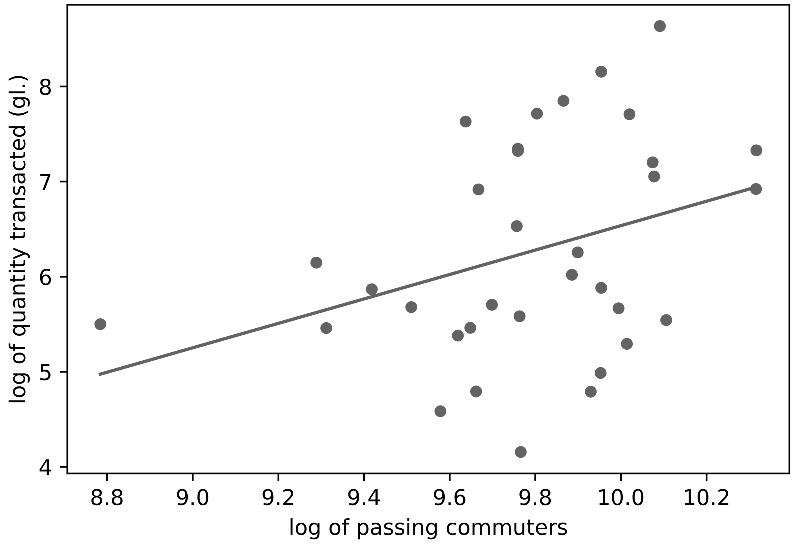
	count	mean	std	25%	50%	75%
	All Gasoline Stations					
St. avg. price ($)	47	1.97	0.07	1.93	1.94	2.03
St. avg. quantity (gl.)	35	958.87	1173.29	240.19	375.05	1428.15
	Large-format Independent Retailers					
St. avg. price ($)	5	1.93	0.01	1.93	1.93	1.93
St. avg. quantity (gl.)	5	2699.12	1677.34	2063.04	2224.84	2239.03
	Major Branded Retailers					
St. avg. price ($)	17	2.02	0.06	1.96	2.03	2.06
St. avg. quantity (gl.)	13	392.26	423.76	120.77	293.09	375.05

volume and the average transaction volume at a station, both measured in logarithms.[23] As this figure suggests, as more drivers pass a station, the sampling probability of this station is higher, which, in turn, leads to higher sales. This positive relationship also confirms the advantage of using traffic data to simulate consumers' search patterns for gas stations over the assumption of random sampling with equal probability. The latter assumption, which is a simplification widely adopted by empirical literature on consumer search when individual search histories and quantity data are not observed (e.g., Hong and Shum 2006; Wildenbeest 2011; and Nishida and Remer 2018), implies that when products are homogeneous and sampled with equal probability, the average sales volumes across stations are expected to be the same. While a simplifying assumption is not supported by observed data, our assumption that consumers search along their predetermined travel route better describes the market structure of the retail gasoline market. It is also important to notice that the relationship between traffic and transaction volume is positive but not perfectly linear, indicating that additional attributes of stations also matter, such as, price reputation and brand quality, among others.

Consumers also pass enough stations to allow them to search without deviating from

[23]One outlier station is excluded from the figure. This station is in a remote area by a lake with a high sales volume, perhaps resulting from the high demand for gas for boats. This is outside the scope of the paper. The average sales volume of this station is controlled for in the structural estimation.

Figure 3: Sales Volume and Traffic Volume



Note: The slope of the linear fitted line is 1.28, significant at 5 percent level.

their travel route as we will assume in the structural model. Figure 4 shows the distribution of the number of stations on drivers' travel routes. More than 50 percent of drivers pass at least three gasoline stations on their trips. The average number of stations passed by a driver is 3.57, a number larger than that searched by consumers before buying an MP3 player or a book from online retailers.[24]

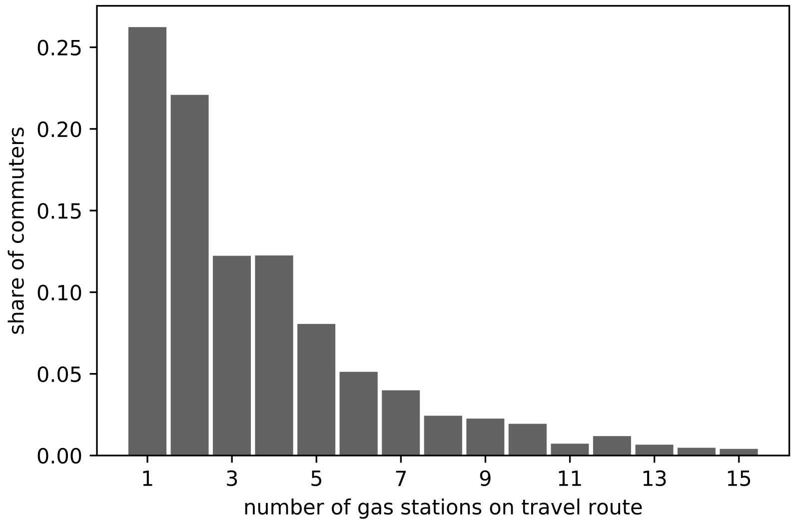
3.2 Two Types of Uncertainty

Frequent price changes in the retail gasoline market make it difficult for consumers to maintain accurate price information. Figure 5 shows that the distribution of the proportion of stations that change their price from the previous day. On an average day, about 30 percent of stations in the sample change their price at least once,[25] resulting in two types of uncertainty in the market: (i) ex-ante uncertainty about the price at each station, and (ii)

[24]Using data on individual online browsing and purchase histories, De los Santos et al. (2017) find consumers visit on average 2.82 online retailers before buying an MP3 player, and De los Santos (2018) finds consumers searched 1.3 online bookstores before purchasing a book.

[25]This number is only a conservative measure of the proportion of stations changing their prices on a day, due to the existence of missing prices. It is likely that price changes happen, but the data did not record them.

Figure 4: Share of Drivers By the Number of Gas Stations on Travel Route



uncertainty about the probability distribution of all stations' prices. To further analyze the different sources contributing to these two types of uncertainty, we perform the following basic fixed effect regression,

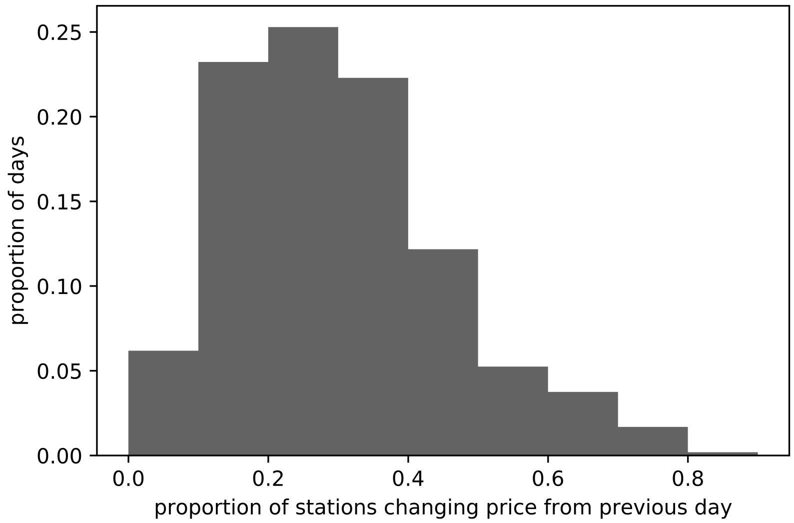
$$p_{j,t} = \sum_{j=1}^{J} \psi_j Station_j + \sum_{t=2}^{T} \phi_t Day_t + \nu_{j,t}, \tag{1}$$

where price at station j on day t, $p_{j,t}$, is a combination of the station fixed effect, ψ_j, day-of-sample fixed effect, ϕ_t, and an idiosyncratic error term, $\nu_{j,t}$.

Not all of the price components cause uncertainty. The average price differences across stations, as measured by the differences in station fixed effects (ψ_j), are probably perceived by consumers. As they repeatedly observe the prices of a number of stations during their everyday driving, they tend to know the identity of higher and lower priced stations on their travel route.

After subtracting the persistent price differences across stations from the observed prices, we obtain the time variant components of the prices. We define time-variant price, $\tilde{p}_{j,t} = \phi_t + \nu_{j,t}$, as a combination of the day-of-sample fixed effect (ϕ_t), which is driven by changes in market conditions (for example, wholesale cost) common to all stations, and the price residual

Figure 5: Proportion of Stations Changing Price From Previous Day



($\nu_{j,t}$), which is a result of the private shock (for example, private cost shocks) specific to a certain station from day to day.[26]

These two components contribute to uncertainty in the market in different ways. First, variation in the price residual ($\nu_{j,t}$) causes station prices to fluctuate relative to one another (e.g. Lewis 2008). More specifically, station specific shocks might result in a lower priced station having a higher price than a higher priced station on a nontrivial number of days (e.g., Chandra and Tappata 2011 find that this situation occurs approximately 15 percent of the time). Consequently, price residuals create the first type of uncertainty by preventing consumers from knowing a station's location in the price distribution prior to search. Price residuals, which reflect the true price differences across stations that are ex-ante unknown to consumers on a day, are the focus of most of the empirical papers that structurally estimate models of consumer search. These studies typically assume that consumers know the distribution of price residuals. In particular, market conditions (reflected in ϕ_t) are either assumed to be constant over time (e.g., Hong and Shum 2006; Wildenbeest 2011; etc.), or the changes are assumed to be known to consumers (e.g., Nishida and Remer 2018 recover

[26]Some search literature consider price residual as a result of firms employing mixed strategy. This paper assumes pure strategy equilibria similar to Hortaçsu and Syverson (2004).

estimates of search costs in the gasoline market assuming that consumers know ϕ_t).

Table 4 summarizes the dispersion of price residuals. The size of the residual price differences across stations, which are ex-ante unknown to consumers on a single day, shows the potential gains when searching for lower prices. Daily price dispersion is calculated by the standard deviation and the range (max minus min) of the price residuals across stations within a day. The first two rows of Table 4 report the city-wide result. On an average day, the standard deviation and range of the price residuals are 0.066 and 0.251 dollars, respectively. However, not all of the stations in the city are easily accessible to each consumer. The dispersion of price residuals of stations along a travel route gives a better picture of the gains of searching that can be exploited by consumers driving on that route. The last two rows present the price dispersion on the search route and at the day level.[27] The average ex-ante unknown price difference across stations on a search route on a day measured in standard deviation (range) equals 0.024 (0.055) dollars. Thus, on average, an individual can realize a maximum gain of 5.5 cents per gallon of gasoline purchase by searching the stations on her travel route.

Table 4: Daily Dispersion of Price Residuals (\$/gal)

	mean	median	min	max	N. obs.
In City					
Standard deviation	0.032	0.032	0.013	0.069	535
Range	0.158	0.147	0.052	0.455	535
On Route					
Standard deviation	0.024	0.019	0.000	0.292	402516
Range	0.055	0.044	0.000	0.455	402516

The second component of price variation and uncertainty, which results from changing market conditions as measured by ϕ_t, is a critical feature in the retail gasoline market, and it creates the second type of uncertainty. As shown by Figure 1, the daily distribution of prices fluctuates frequently and significantly in response to wholesale cost changes. It is unlikely that consumers know the current distribution of time-variant prices with any degree

[27]The calculation is based on search route-day that has at least two price observations.

of certainty.

Consumers' best guess of the distribution of prices today is likely to be based on the prices observed from past trips. Therefore, we assume in the structural model that consumers formulate their prior belief of price distribution on prices they observed while driving in the past. As consumers observe new price, they Bayesian update their prior belief based on the newly acquired price information. The update is not perfect, however, because consumers do not know why prices are changing. When the price observed at a station today is higher than the previous day, consumers do not know if this new price is specific to this station or it reflects changes in the market condition that are common to all stations. The extent to which consumers' prior beliefs of the current price distribution are updated depends on how much a price observed at a station reflects the change in the market condition. For instance, if consumers believe this price change is specific to this station, then they do not adjust their beliefs about the prices at other stations. If, on the other hand, they believe that this price change is common to all stations, then they will adjust their beliefs about the prices at the other stations. In addition, since consumers observe only a limited number of new prices on their travel route, their information about the current price distribution is incomplete. As a result, consumers' prior beliefs always impact beliefs about the current price distribution, but the weight placed on these priors decreases as more new information is obtained.

An imperfect knowledge of the price distribution can have important effects on consumer behavior. For example, this lack of information explains the commonly observed pattern of asymmetric price response to cost changes where cost increases are passed through more quickly than cost decreases, as depicted in Figure 1. Lewis (2011) argues that when cost increases push up prices, consumers' beliefs of the price distribution may be lower than the actual price distribution, causing consumers to increase their search. As a result, price margins go down. On the other hand, when costs and prices fall, the current prices may be lower than consumers expected, meaning they search less, which leads to higher profit margins.

The behaviors of the equilibrium prices are consistent with the implication of a model of consumer search with learning. In the following sections, we propose a model of consumer search with learning, and we then apply the model to data and structurally estimate how consumers form their believes.

4 Model

4.1 Utility

Demand for gasoline is characterized by consumers searching for it from gas stations on their travel routes. We consider a city containing a set, $\mathbb{J}$, of J stations indexed $j = 1, 2, ..., J$. We assume that consumer each demands 10 gallons of gasoline and all share a common utility function. Consumer's indirect utility for gasoline at station j on day t is equal to

$$u_{j,t} = X_j\beta - p_{j,t},$$

where X_j represents station j's non-priced characteristics, and $p_{j,t}$ is the unit cost of gasoline (per gallon price observed multiplied by 10 gallons) at gas station j on day t. Note that the coefficient on the price is normalized to -1, so utilities are expressed in terms of dollar values.

Based on the decomposition of prices in Equation 1, $p_{j,t} = \psi_j + \phi_t + \nu_{j,t}$, we can rewrite consumer's indirect utility as

$$\begin{aligned} u_{j,t} &= X_j\beta - \gamma\psi_j - \phi_t - \nu_{j,t} \\ &= V_j - \tilde{p}_{j,t}, \end{aligned} \tag{2}$$

where ψ_j is station j's average price, ϕ_t represents the day-to-day changes in the city average price, and $\nu_{j,t}$ is the idiosyncratic deviation of station j's price on day t from its own average and the city average. We further partition utility into $V_j = X_j\beta - \gamma\psi_j$, which measures the value of station j's time invariant characteristics, and a time-varying price component $\tilde{p}_{j,t} = \phi_t + \nu_{j,t}$. We use ψ_j as a measure of a station's reputation of being a high or low priced station, and we allow consumers to respond to the differences in price reputation across stations separately from changes in time-variant prices. The difference in response is measured by $\gamma - 1$.

The partition of the utility function is motivated by the features of the retail gasoline market. Repeated observations of a number of stations and frequent purchases of gasoline at these stations allow consumers to become aware of the characteristics of stations that are constant over time, such as the location, brand, and price reputation. V_j then represents the part of utility known to consumers prior to search. In contrast, time-variant prices, representing the changes in prices over time and across stations, are unknown to consumers, as discussed in the previous section.

Our approach allowing for ex-ante differentiation of products relaxes the assumption adopted by papers on search using market level data that products are ex-ante homogeneous (e.g. Hortaçsu and Syverson 2004). Kim et al. (2010) and Moraga-González et al. (2015) allow ex-ante differentiation of products, however, in their papers the sampling process is endogenous. In particular, consumers search the alternatives that yield the highest expected utilities first based on the characteristics observed. In our model, the search order of stations is exogenous as it is predetermined by consumers' travel routes.

As consumers drive along their travel route, they observe prices to realize the time-variant price at each station. Consumers then Bayesian update their belief of the current distribution of time-variant prices with each newly observed price. In the following section, we detail the consumer learning process.

4.2 Consumer Learning

Let $\boldsymbol{\rho} = (\rho_1, \rho_2, ..., \rho_H)$ be a set of all possible values of time-variant prices. These values are sorted in increasing order, with ρ_1 being the lowest possible value and ρ_H being the highest. The probability of sampling each value is given by a vector $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_H)$. Consumers consider these probabilities to be random variables that are distributed according to a Dirichlet distribution, following Rothschild (1974), Koulayev (2013) and De los Santos et al. (2017). We assume consumers' prior beliefs for the probabilities follows

$$f(\boldsymbol{\pi}|\boldsymbol{a}) = \frac{\Gamma(\sum_{h=1}^{H} a_h)}{\prod_{h=1}^{H} \Gamma(a_h)} \prod_{h=1}^{H} {\pi_h}^{a_h - 1}, \quad \sum \pi_h = 1, \quad \pi_h > 0, \tag{3}$$

where Γ is the gamma function and $\boldsymbol{a} = (a_1, a_2, a_H)$ is a vector of concentration parameters, which can be re-written as $\boldsymbol{a} = \alpha\boldsymbol{\omega}$, where $\boldsymbol{\omega}$ is a vector of the expected value of the probabilities of prior beliefs

$$\boldsymbol{\omega} = E(\boldsymbol{\pi}|\boldsymbol{a}) = (\frac{a_1}{\alpha}, \frac{a_2}{\alpha}, ..., \frac{a_H}{\alpha}), \quad \alpha = \sum a_h, a_h > 0, \tag{4}$$

and α is a scale that is often interpreted as the weight on the initial prior belief (prior sample size). The ratio a_h/α represents the expected likelihood of value ρ_h being sampled.

The prior belief is updated as consumers observe each price on their travel route. Since the Dirichlet distribution is the conjugate prior of the multinomial distribution, the posterior distribution will be Dirichlet as well. Let s_k be the set of k stations that consumers observe on their travel route. $\boldsymbol{N}_{s_k} = (n_1, n_2, ..., n_H)$ then represents the number of times each time-variant price has been sampled during the search of k stations. The posterior expected value of the sampling probabilities is

$$E(\boldsymbol{\pi}|\alpha\boldsymbol{\omega}, \boldsymbol{N}_{s_k}) = (\frac{a_1 + n_1}{\alpha + \sum n_h}, \frac{a_2 + n_2}{\alpha + \sum n_h}, ..., \frac{a_H + n_H}{\alpha + \sum n_h}). \tag{5}$$

Consumers' posterior beliefs of the price distribution depends on three aspects, the shape of the prior belief, presented by $\boldsymbol{\omega}$; its weight presented by α; and the newly obtained information, $\boldsymbol{N}_{s_k}$. According to the learning process, the expected likelihood of observing ρ_h, $\frac{a_h + n_h}{\alpha + \sum_h^H n_h}$, increases every time it is observed. The speed of consumer belief updates is inversely related to α. More specifically, the more confident consumers are in their prior beliefs, the larger the α and the slower the beliefs are updated with each new observation. As more prices are sampled from the current distribution, the posterior belief converges to the actual price distribution.

4.3 Ordered Search

Consumers are assumed to search the stations along their travel route. In practice, drivers commonly encounter plenty of options without having to go out of their way. The average driver in our data observes 3.57 stations on her travel route. Therefore, any deviation from the travel route, including recall (driving back to a previously passed station) is assumed to be too costly. As a result, consumers' search of gasoline stations is sequential and ordered, as they know ex-ante the predetermined order of visits of differentiated stations. In this section, we discuss how a consumer makes a decision to purchase gasoline along the route. Then in the following section, we aggregate all driver decisions to obtain the total volume of gasoline transacted at a station.

A search route s is defined as a set of sequentially ordered stations that a consumer will drive past on her trip, with $\mathbb{S}$ as the set of all search routes. At each station on the search route, a consumer updates her belief of the price distribution with each new price observed before deciding whether to purchase gasoline at this station or go to the next one. As the travel route is predetermined, observing the price of the next station is assumed to be costless. The consumer maximizes her utility of a gasoline purchase by comparing the realized utility at each station with the expected value of the remaining options, which is characterized as reservation utility.

Let there be K stations, with k representing the location of consumer i on search route s, and $k = 1$ being the first station and $k = K$ the last. Let $s(k)$ return the station index j for the kth element in s, with $s(k) \in \mathbb{J}$.

Consumer i's decision at the kth station can be formalized as follows. We begin by dividing the search route into two parts, $s_k \cup \bar{s}_k$, with s_k being the set of stations already visited by consumer i, and $\bar{s}_k$ containing the remaining stations on the route. Then, based on the Equation 5, her belief about the current distribution of time-variant price after observing prices at the first k stations is $E(\boldsymbol{\pi}|\alpha\boldsymbol{\omega}, \boldsymbol{N}_{s_k})$. Conditional on her belief, she derives the reservation utility at the current kth station using backward induction from the outside option.

If she does not buy from the final station, she is allowed the alternative of waiting to purchase gasoline on a future trip. However, she needs to pay a cost of postponing purchase c_i to access the alternative. This c_i will be higher for those running out of gas or who need to purchase now for some other reasons, and lower for those who are seeking to purchase gas but not under pressure to do so immediately. Based on the observation that a driver regularly travel different routes on different occasions, the outside option is simplified as sampling a station randomly from the set of all stations, $\mathbb{J}$. Denote $Z_x(\cdot)$ as the reservation utility that represents the expected value of having x stations remaining on the route. Thus, at the last station where there are no stations remaining, the reservation utility $Z_0(\alpha\boldsymbol{\omega}, \boldsymbol{N}_{s_k}, \boldsymbol{V}_\emptyset, \boldsymbol{V}_\mathbb{J}, c_i)$ solves the following equation,

$$c_i = \sum_{j=1}^{J} \lambda_j \sum_{\rho_h < V_j - Z_0} (V_j - \rho_h - Z_0) E(\pi_h|\alpha\boldsymbol{\omega}, \boldsymbol{N}_{s_k}), \tag{6}$$

where λ_j is the probability of sampling station j on future trips, $\boldsymbol{V}_\emptyset$ is a vector of the values of the station attributes remaining on the route, and $\boldsymbol{V}_\mathbb{J}$ is a vector of values of station attributes outside of the route. Later in the empirical exercise, we assume equal probability of sampling a station on an individual's future trips, $\lambda_j = 1/J$. The right hand side of Equation 6 describes the expected gains of choosing to buy gasoline during future trips, if the utility at the last station is at Z_0. By equating expected gains with the cost of postponing purchase c_i, Z_0 summarizes the expected value of the outside option, conditional on the driver's belief at the kth station. Thus, if the realized utility at the last station is higher than Z_0, the cost of postponing purchase is higher than the expected gain from postponement, and the driver will make a purchase at the final station, and vice versa.

The calculation of reservation utility at the second to last station is then

$$\begin{aligned} Z_1(\alpha\boldsymbol{\omega}, \boldsymbol{N}_{s_k}, \boldsymbol{V}_{\bar{s}_{K-1}}, \boldsymbol{V}_{\mathbb{J}}, c_i) &= \sum_h max\{Z_0, V_{s(K)} - \rho_h\} E(\pi_h|\alpha\boldsymbol{\omega}, \boldsymbol{N}_{s_k}), \\ &= Z_0 + \sum_{\rho_h < V_{s(K)} - Z_0} (V_{s(K)} - \rho_h - Z_0) E(\pi_h|\alpha\boldsymbol{\omega}, \boldsymbol{N}_{s_k}). \end{aligned} \tag{7}$$

Equation 7 shows that at the final station (Kth) station, consumer i will make a purchase only if the realized utility is greater than the reservation utility $u_{s(K)} = V_{s(K)} - \tilde{p}_{s(K)} > Z_0(\alpha\boldsymbol{\omega}, \boldsymbol{N}_{s_k}, \boldsymbol{V}_{\emptyset}, \boldsymbol{V}_{\mathbb{J}}, c_i)$, conditional on her belief at the kth station; otherwise, she will choose the outside option Z_0. Therefore, the reservation utility at the second to last station, Z_1, is higher than the reservation utility at the last station Z_0 by the amount of expected gain from search at the last station.

As consumer i continues to backward solving this recursive relationship for reservation utility for each station from the end of the route to her current location, the reservation utility at the kth station is

$$\begin{aligned} &Z_{K-k}(\alpha\boldsymbol{\omega}, \boldsymbol{N}_{s_k}, \boldsymbol{V}_{\bar{s}_k}, \boldsymbol{V}_{\mathbb{J}}, c_i) \\ =&Z_{K-k-1} + \sum_{\rho_h < V_{s(k+1)} - Z_{K-k-1}} (V_{s(k+1)} - \rho_h - Z_{K-k-1}) E(\pi_h|\alpha\boldsymbol{\omega}, \boldsymbol{N}_{s_k}). \end{aligned} \tag{8}$$

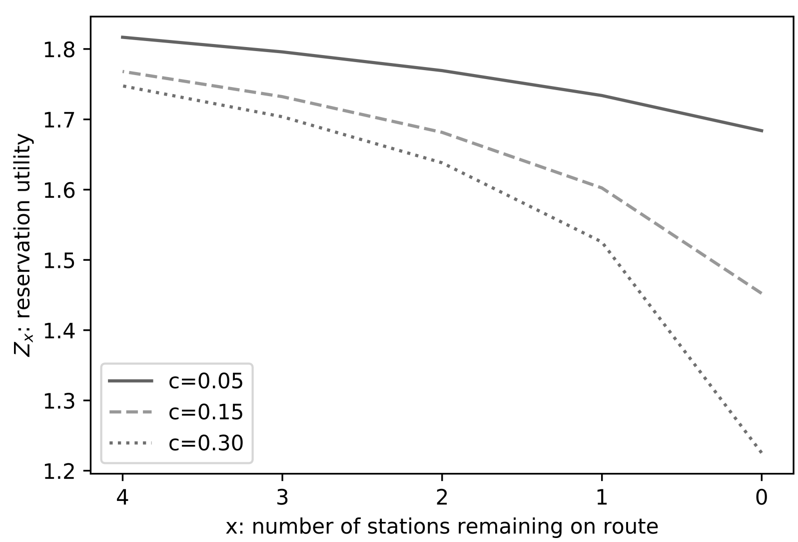
The transition of the reservation utility at each station has the following properties.

Proposition 1: Conditional on consumer i's belief, reservation utility Z_k decreases in c_i for any $k = 0, 1, ..., K-1$.

Proposition 2: Conditional on consumer i's belief, the reservation utility decreases as the number of stations remaining on the route decreases. $Z_k \leq Z_{k+1}$ for any $k = 0, 1, ..., K-1$.

Figure 6 summarizes the two propositions. The three curves in the graph depict the transition of the reservation utility as the number of stations remaining on the route changes, given three different costs of postponing purchase, assuming that the consumers' beliefs of the price distribution are uniformly distributed from 0 to 1 and stations are homogeneous with attribute value $V = 2$. First, as the postponement cost is higher, the value of purchase on future trips becomes lower, as shown by the curves when the number of stations remaining is zero. Consequently, a higher postponement cost means drivers require a lower threshold at all points along the route (*Proposition 1*). A driver who has an almost empty gas tank is more likely to accept a lower utility (higher price) at any station along her route than a driver who still has plenty of gas left in the tank. Second, as the number of stations remaining on the route goes down, the reservation utility goes down (*Proposition 2*). More specifically, when consumers come to the end of their routes, they are more likely to accept lower utilities (higher prices) as they have fewer options remaining. In addition, *Proposition 2* also suggests that, if there is a station with a high value of attributes on the route, the reservation utility is higher at any station before that station, suggesting that a driver will buy from a station only if the realized utility is higher than the expected utility at any station remaining on the route.

Figure 6: Transition of Reservation Utility



Note: This figure assumes that consumers' belief of the price distribution follows a uniform distribution with support from 0 to 1. Stations are homogeneous with attribute values equal to 2.

We further illustrate how the distribution of the cost of postponing a purchase determines the proportion of consumers who choose each option on the route. At the kth station on the route, after realization of the utility u_k and updating the belief of the price distribution $E(\boldsymbol{\pi}|\alpha\boldsymbol{\omega}, \boldsymbol{N}_{s_k})$, we obtain the critical cutoff of postponement cost c_k^* that equates the reservation utility at the kth station with u_k. Because Z_{K-k} is a decreasing function in c, according to *Proposition 1*, its inverse is

$$c_k^* = Z_{K-k}^{-1}(u_k|\alpha\boldsymbol{\omega}, \boldsymbol{N}_{s_k}, \boldsymbol{V}_{\bar{s}_k}, \boldsymbol{V}_{\mathbb{J}}). \tag{9}$$

Therefore, consumers who have search cost $c > c_k^*$ will purchase at the kth station as the realized utility u_k is higher than the expected value of the remaining options, given that they have not purchased from any previously visited stations. Let G denote the distribution of the postponement costs c_i for consumers who drive on route s. Then the proportion of consumers who purchase from the kth station on the route is

$$q_k = \begin{cases} G(c_k^{**}) - G(c_k^*) & \text{if } c_k^{**} > c_k^* \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

where $c_k^{**} = min\{c_g^* \text{ for any } g < k\}$. As consumers with a cost of postponing purchase higher than c_k^{**} would have already bought from a station before the kth station, only when $c_k^{**} > c_k^*$, would the kth station have positive gasoline sales on that particular search route.

This concludes the formalization of the individual level purchase decision as a result of searching and leaning on each search route. In the next section, we explain the aggregation from individual to station level, and the use of daily price and quantity data at each station, combined with traffic data for estimating the model.

5 Estimation

Our general estimation approach is as follows. We begin by aggregating a station's share of sales, as described in Equation 10, across all routes the station is on to obtain the model predicted share of sales at the market level, which is the level of observation of our quantity data.

The observed market share of station, j, on day, t, $Q_{j,t}$, is defined as the share of total drivers who purchase gas at the station on that day. The number of drivers who purchase at a station is measured by the observed daily quantity of gasoline transacted at the station divided by 10 gallons, based on the assumption that a driver only buys 10 gallons of gas at each purchase, while the total number of drivers are obtained from the traffic data.[28]

Since the majority of drivers on a search route have fairly full tanks and are unlikely to even consider purchasing gas, we allow cost of postponing purchase till a future trip to have a probability mass of $1 - \eta_t$ at zero. These drivers never purchase because their expected value of future purchase, Z_0, equals the highest utility (lowest price) possible from a gasoline purchase. The remaining η_t share of drivers have a positive probability of purchasing gasoline on their search routes. Note that η_t is allowed to vary over time to account for the changes in overall demand (frequency of purchase) for gasoline. For example, during summer seasons drivers purchase more frequently, meaning the proportion of drivers with a positive probability of purchasing gasoline also goes up. We assume consumers' costs for postponing purchase follow a log-normal distribution and associate these costs to the demographics of drivers who drive along a search route as follows,

$$lnc_{i,s} = C + \mu_c Y_S + \zeta_i^c, \tag{11}$$

where C is a constant, Y_S is the measures of the demographics of drivers on a route, including the proportion of drivers with incomes greater than \$50,000 dollars, and a stochastic term ζ_i^c

[28]A unit of gasoline purchase of 10 gallons is a scaler chosen for the convenience of interpreting the estimation results.

that follows $N(0,\sigma_c^2)$ that is common to all routes. Therefore, Equation 10, which describes the share of drivers on search route s who purchase from the kth station, can be rewritten as

$$q_{k|s} = \begin{cases} \eta_t(\Phi(\frac{lnc_k^{**} - C - \mu_c Y_S}{\sigma_c}) - \Phi(\frac{lnc_k^{*} - C - \mu_c Y_S}{\sigma_c})) & \text{if } c_k^{**} > c_k^{*} \\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

where $\Phi(\cdot)$ is standard normal CDF. Let $s^{-1}(j)$ returns the location index k of station j on s. The share of total drivers who buy from station j at time t is,

$$\hat{Q}_{j,t}(\boldsymbol{\theta}) = \frac{1}{\sum_{s\in\mathbb{S}} W_S} \sum_{s\in\mathbb{S}} q_{s^{-1}(j)|s,t} W_S \mathbb{1}(j \in s), \tag{13}$$

where W_s is the number of drivers on route s obtained from the traffic data. The set of parameters to be estimated is given by $\boldsymbol{\theta} = (\alpha, \beta, \gamma, C, \mu_c, \sigma_c, \eta_t)$, where α is the weight on consumers' prior beliefs, β is the value of station attributes such as brand, γ is the sensitivity to price reputation, and C, μ_c and σ_c determine the shape of the postponement cost distribution.

Next, we assume that the model predicted market share differs from its observed value at each station by a additive independent normal error, $\epsilon_{j,t} = Q_{j,t} - \hat{Q}_{j,t}(\boldsymbol{\theta})$, with $\epsilon_{j,t} \sim N(0,\sigma)$, and subsequently construct the likelihood function to be maximized as follows,

$$\prod_t \prod_j \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Q_{j,t} - \hat{Q}_{j,t}(\boldsymbol{\theta}))^2}{2\sigma^2}}. \tag{14}$$

The error term $\epsilon_{j,t}$ represents an aggregate level measurement (prediction) error, similar to Bresnahan (1987), Hortaçsu and Syverson (2004), Kim et al. (2010), etc. Several sources can lead to measurement error. First, inaccuracy of the gasoline prices and the imputation of missing prices can directly cause measurement error in the observed market share at a station since the quantity of gasoline transacted at a station is calculated by dividing the total value of transactions by the price. Second, the model predicted market shares are based on the estimated daily average traffic flows that do not reflect temporary road repair or congestion. True daily market shares may be influenced by such temporary changes.

Consumers' prior beliefs about the price distribution on day t before observing any prices on their search route is assumed to be a result of continuous learning of the price distribution based on the observed prices on past trips, with older price observations discounted more heavily. Consumers are assumed to have a homogeneous prior belief and updating process, described as follows,

$$\boldsymbol{\omega_t} = E(\boldsymbol{\pi}|\alpha\boldsymbol{\omega}_{t-1}, \boldsymbol{N}_{t-1,\mathbb{J}}/J), \tag{15}$$

where $\boldsymbol{\omega_{t-1}}$ is consumers' prior beliefs on day $t-1$, and $\boldsymbol{N}_{t-1,\mathbb{J}}/J$ is the vector of the number of times each price is being observed in the city on day $t-1$ divided by the total number of prices observed, which is also the empirical density distribution of the city price distribution. One can think of this process as an exponential moving average of the city price distribution. As discussed in Section 3, because consumer information about price distribution on any given day is imperfect and incomplete, the prices observed in the past always carry some weight in the consumers' beliefs. This weight, measured by α, governs the speed of the learning process.

Identification of the different parameters in the model relies on different sources of variation. First, consumer preferences, β and γ, for product attributes, including brand and price reputation, are identified by observed variation in average market shares across stations. Second, the parameters, C, μ_c and σ_c that govern the distribution of the costs of postponing purchase are identified by (i) how observed variation in market share is related to the variation in station average utilities (normalized to dollars) across stations, conditional on the variation in the number of drivers who go past each station and (ii) how variation in a station's market share over time relates to changes in day-to-day prices. More specifically, the first variation shows how likely it is for drivers assigned to a route with bad stations (those with high price reputations) to postpone purchase until they are assigned a route with better options on average, while the second variation represents intertemporal substitution reflected in a station's market share. Third, the weight on the prior belief α is identified by the variation of the difference between consumers' prior beliefs and the current price distribution over time.

6 Results

6.1 Parameter Estimates

The estimated parameters are shown in Table 5. The parameter estimate of the weight on the prior belief of the price distribution is 5.31. More specifically, consumers' initial prior belief, which is formulated by past price observations during everyday driving, is as informative about the current price distribution as if consumers had already observed approximately five prices today. When less weight is placed on prior beliefs, consumers update their beliefs more quickly with each new price observation. For example, if an individual believes that the average price today is \$2.00, a prior weight of 5 suggests that if she observes a price of \$2.12, she will update her belief of the average price to \$2.02. However, if the prior weight is 11, her posterior belief of the average price is \$2.01. It is important to note that consumers do not update their belief of the expected price to be equal to the price observed at a particular station because they do not know whether the change is specific to this station or is common to all stations. The estimated rate of updating is much faster than what is generally assumed by literature. In the two empirical papers that apply models of search with learning (Koulayev 2013 and De los Santos et al. 2017), the weight on the prior belief is not estimated but rather chosen by the researchers to be the number of product-retailer combinations. This results in a much higher weight on the prior. (For example, Koulayev 2013 set α to be ranging from 24 to 82). Our estimate of a relatively fast learning process is consistent with the fact that the distribution of gasoline prices changes regularly in response to wholesale cost volatility. As a result, past price observations carry only limited information, while prices sampled today are much more informative about the current distribution of prices.

Table 5: Estimation Results

	Coeff.	Std. Err.
Learning		
Prior Weight (α)	5.3055	(1.4703)
Seller Attributes		
Price Reputation (γ)	-1.7872	(0.3093)
Majors (β)	-0.2419	(0.1749)
Large-format Independents (β)	2.1853	(0.1429)
Log of Postponement Cost		
Constant (C)	-2.7456	(0.1936)
Income $>$ 50k (μ_c)	0.0773	(0.0801)
SD (σ_c)	4.8452	(0.2194)

Note: Bootstrapped standard errors are in parentheses. The coefficient estimates are for an one time purchase of 10 gallons of gasoline. Day of sample fixed effects and month of sample fixed effects are included in η_t to capture the change in the daily overall demand for gasoline. Estimates of η_t are not reported.

The estimated value of station attributes, which are ex-ante known to consumers, are compared to their response to day-to-day price changes, which are unknown prior to search. More specifically, the coefficient on the known and the persistent price reputation across stations is estimated to be -1.79, a value significantly different from -1, the response to time-variant price changes. This suggests that when consumers are driving, they are likely to postpone gasoline purchase until they reach a station that tends to be less expensive even if it means forgoing lower prices offered by a station that tends to be more expensive. If a station's price goes up by two cents (approximately 1 percent of average gasoline price, \$1.97, in the sample) in a day, which is ex-ante unknown to consumers and only revealed to those who drive by this station, it would see an average decrease in market share of 9.49 percent. On the other hand, if a station increases its persistent price reputation by two cents, which is known to all consumers ex-ante, an average station would experience a 24.65 percent drop in its market share. Table 6 summaries the results. This large difference in the change in market share is driven by the fact that consumers are almost twice as responsive to a change in price reputation than to a change in time-variant price. In addition, change in price reputation results in an ax-ante change in consumers' purchase probability at the station, while a time-variant change results only in a change in the purchase probability ex-post.

Table 6: Average Percentage Change in a Station's Market Share Under Two Types of Price Change

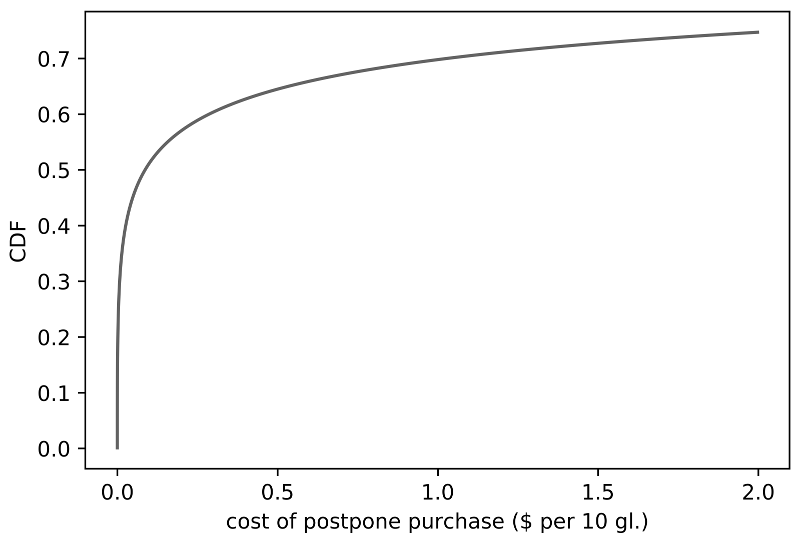
	mean	std	25%	50%	75%
Transitory Price (%)	-9.49	7.41	-11.56	-7.78	-4.85
Price Reputation (%)	-24.65	17.00	-31.79	-22.20	-11.07

Note: The stations near the interstate highway and the lake are excluded from the calculation.

The coefficient on the major-branded retailers is negative, however not significantly different from zero, which is the mean value of the other independent retailers. This result does not provide evidence that consumers are willing to pay a premium for branded gasoline. However, consumers are willing to pay about $2.19 premium for 10 gallons of gas at the large-format independent retailers, probably because they have standardized and consistent amenities, with large islands, plenty of pumps, and large, clean convenience stores, as well as a brand image of having a low price, in addition to their individual station price reputations.

Based on the postponement cost estimates, for drivers who have a positive probability of purchasing gasoline, the median cost of postponing purchase to a future trip is 6.4 cents per 10 gallons purchased, and 25 percent of these drivers have postponement costs higher than $2.09. Figure 7 depicts the estimated postponement cost CDF. These numbers indicate that the majority of drivers start seeking gasoline well in advance of their tank being empty. As they do not act as if they are pressured to buy gas immediately, they can postpone their purchase until they drive on a route with low priced stations. However, approximately 25 percent of the drivers seeking gasoline are willing to pay 20 cents per gallon more to purchase gasoline on the current route, due, for example, to an almost empty gas tank. Although the coefficient on income is positive, it is not statistically nor economically different from zero. One would expect stations on routes with a higher proportion of high-income drivers to have higher sales volumes even if prices are higher because such drivers are less price sensitive and have higher postponement costs. However, this expectation is not supported by the data, perhaps because the variation in the proportion of residents with annual income higher than $50,000 across traffic zones is too small for identification (the standard deviation is 3.76 percentage points).

Figure 7: CDF of the Cost of Postponing Purchase



6.2 Cross-Price Elasticities

In this subsection, we evaluate the importance of spatial differentiation in explaining substitution patterns across stations by analyzing the impact of the distance and the number of common drivers shared between stations on the estimated elasticity of substitution. Based on the parameter estimates shown in Table 5, we obtain the estimated cross-price elasticity for a pair of stations by calculating the percentage change in target station i's market share due to an ex-ante unknown one percent increase in the price of the other station, j, $\varepsilon_{i,j} = \%\Delta\hat{Q}_i/\%\Delta p_j$. Given that a majority of station pairs have virtually zero elasticities (it is not surprising that two stations located at the opposite ends of the city do not compete closely against each other as they do not share much common traffic), this analysis focuses only on pairs of stations with a cross-price elasticity greater than 0.01, a value exhibited by less than 20 percent of the total possible station pairs.

Table 7 summaries the distribution of the estimated cross-price elasticities and two measures of spatial differentiation between stations. We can see that even after the exclusion of station pairs that do not compete closely against each other, the distribution of cross-price elasticities is right-skewed. The average cross elasticity is 1.73, with only 10 percent of the pairs having a cross elasticity greater than 2.16. In our restricted sample, the average common traffic share between two stations, which is defined as the proportion of the target station i's traffic that also goes past station j, is 24 percent, and the average driving distance is 3.36 miles.

Table 7: Summary Statistics of Elasticity of Substitution and Measures of Spatial Differentiation Between Stations

	count	mean	std	min	10%	50%	90%	max
Cross Price Elasticity	284.0	1.73	8.85	0.01	0.02	0.15	2.16	127.91
Share of Common Traffic	284.0	0.24	0.28	0.00	0.01	0.11	0.72	1.00
Driving Distance (*miles*)	284.0	3.36	2.08	0.05	0.91	3.11	6.28	10.51

Note: This table includes only station pairs with a cross-price elasticity greater than 0.01, a value exhibited by fewer than 20 percent of the total station pairs.

Table 8 reports the results of a regression of the logarithm of the estimated cross-price elasticity on the measures of the spatial differentiation of station pairs. As expected, these regression results indicate that cross-price elasticity between two stations increases in the proportion of common traffic and declines as the driving distance between them increases. A comparison of the results in Columns (1) and (2) suggest that the proportion of common traffic explains a more significant fraction of the variation in substitution patterns between stations (35 percent) than driving distance. Based on the coefficient estimates in Column (3), holding the driving distance between the two stations constant, a doubling of the common traffic share will increase the elasticity of substitution by 57 percent.

Table 8: Regression Results of Estimated Elasticity of Substitution on Measures of Spatial Differentiation Between Stations

	$ln(\hat{\varepsilon}_{i,j})$		
	(1)	(2)	(3)
$Intercept$	-0.6585**	-0.0166	0.0546
	(0.2451)	(0.2237)	(0.2426)
$ln(Distance)_{ij}$	-1.0342***		-0.3803**
	(0.1585)		(0.1419)
$ln(CommonTraffic)_{ij}$		0.6977***	0.5730***
		(0.0667)	(0.0647)
N	284	284	284
R2	0.22	0.35	0.37

Note: Standard errors, shown in parentheses, are clustered at the target station level. $CommonTraffic_{ij}$ measures the proportion of station i's traffic that also goes past station j. ***, **, and * represent significance at the 1%, 5% and 10% level respectively.

7 Counterfactual

In this section, we study the value of price information to consumers using our estimated demand system. More specifically, we calculate the net change in consumers' surplus after stepwise removal of two types of information uncertainty, (i) uncertainty about the price distribution, and (ii) uncertainty about the price at each station. To do so, we simulate consumers' purchase decisions and compute the aggregated value of these choices across hypothetical cases with different amounts of information uncertainty, while holding the supply side constant.

We begin our analysis with the current state in the presence of both types of uncertainty. The total value for all consumers seeking gasoline on day t under dual uncertainty, $TV_{du,t}$ is calculated as,

$$TV_{du,t} = \sum_{s\in\mathbb{S}} \eta_t W_s \sum_{k\in s} u_{k,t} q_k(\alpha\boldsymbol{\omega}_t, \boldsymbol{N}_{s_k,t}) + \sum_{s\in\mathbb{S}} \eta_t W_s \sum_{k\in s} \int_0^{c^{**}_{du,s}} Z_0(c|f_t) g(c) dc. \qquad (16)$$

$\sum_{s\in\mathbb{S}} \eta_t W_s \sum_{k\in s} u_{k,t} q_k(\alpha\boldsymbol{\omega}_t, \boldsymbol{N}_{s_k,t})$ measures the total realized utility for consumers who have made the purchase decision conditional on their beliefs at each station on each search route, and $\sum_{s\in\mathbb{S}} \eta_t W_s \sum_{k\in s} \int_0^{c_s^{**}} Z_0(c|f_t)g(c)dc$ computes the total expected value for consumers who have chosen to postpone their purchases to future trips, where $Z_0(c|f_t)$ is the value of the outside option evaluated at the current price distribution measured by the empirical price distribution, f_t. $c^{**}_{du,st} = min\{c^*_{du,k}$ for all $k \in s\}$ is the critical cut-off postponement cost on search route s under both types of uncertainty, with consumers whose $c > c^{**}_{du,st}$ choosing to purchase on the route, and consumers whose $c \leq c^{**}_{du,st}$ choosing to postpone purchase to future trips, and $c^*_{du,k}$, solved by Equation 9, is the critical cut-off postponement cost for each station on the route, conditional on consumers' beliefs.

The resulting inefficiencies that occur to consumers under both types of uncertainty are illustrated by the following situations. First, when prices decline, the observed prices are likely to be lower than consumers' beliefs about them. As a result, some consumers may choose to purchase gas immediately at an early station on a route, forgoing options later in the route or future trips that may generate larger savings. Second, when prices increase, the observed prices are likely to be higher than consumer's beliefs of prices. As a result, some consumers may choose to continue searching for less expensive stations or postpone their purchases, eventually having to settle down on more expensive options before learning that the price increases are global.

To evaluate the magnitude of these inefficiencies caused by information uncertainty, we calculate the total value for all consumers under the counterfactual scenario where there is only price uncertainty and the consumers know the distribution of current prices. This assumption is the one typically adopted in the literature on consumer search,

$$TV_{pu,t} = \sum_{s\in\mathbb{S}} \eta_t W_s \sum_{k\in s} u_{k,t} q_k(f_t) + \sum_{s\in\mathbb{S}} \eta_t W_s \sum_{k\in s} \int_0^{c^{**}_{pu,s}} Z_0(c|f_t)g(c)dc, \qquad (17)$$

where $c^{**}_{pu,s}$ is the critical cut-off postponement cost for consumers who are indifferent to purchasing on the current route or postponing their purchases.

Finally, we compute the total value for all consumers under the counterfactual scenario where they know the current prices at all stations. Consequently, consumers choose whether to buy from the station that provides the highest utility on a route or to postpone their purchases to a future trip. The total value under perfect information is,

$$TV_{pi,t} = \sum_{s \in \mathbb{S}} \eta_t W_s \sum_{k \in s} u^*_{st}(1 - G(c^{**}_{pi,s})) + \sum_{s \in \mathbb{S}} \eta_t W_s \sum_{k \in s} \int_0^{c^{**}_{pi,s}} Z_0(c|f_t) g(c) dc, \qquad (18)$$

where $u^*_{st} = max\{u_{kt}$ for all $k \in s\}$ and $c^{**}_{pi,s}$ is the critical cut-off postponement cost for consumers who are indifferent between purchasing from the station with highest utility on the route or postponing their purchases.

Based on our computation, we find that if consumers know about the current distribution of prices, their net surplus increases by an average of \$210.30 on a day in the city, calculated by $\frac{1}{T}\sum_{t=1}^{T}(TV_{pu,t} - TV_{du,t})$. This increased benefit is approximately 0.56 percent of the total value of transactions on a day in the city predicted by the model. It is equivalent to an average of 1.12 cents in savings per gallon when consumers are fully informed about the current price distribution.

Under the second counterfactual scenario, where consumers have perfect information about today's price offered by each station, their net surplus would increase by an average of \$299.90, which is equivalent to an average of 1.60 cents in savings per gallon. This estimated benefit of reduced information uncertainty seems to be small compared to the observed price dispersion in the city (summarized in Table 4) because our estimated model takes into account the limited number of stations on consumers' travel routes and the costs of accessing more stations not on the current travel route. More specifically, the costs of postponing the purchase to future trips, independent of the amount of information consumers have, prevent all consumers from purchasing at the cheapest station in the city, a situation that would result in significant benefits of having perfect information. In addition, the estimated benefit of information is net of the ex-ante known attributes of the stations, including price reputation and brand qualities. As a result, our simulation predicts 90.42 percent of consumers who purchased gasoline on a route have done so from the station that generates the highest utility, while 82.69 percent of consumers have bought from the station with the lowest price under both types of uncertainty.

8 Conclusion

In this paper, we developed a model to study consumer behavior for searching for the best offer while learning about the distribution of prices in the context of the retail gasoline market. Using a novel panel dataset of daily station-level gasoline sales and prices, combined with data on the empirical distribution of drivers, we estimated the parameters that govern consumers' learning, utility, and cost of substitution from current purchase to future purchases. We are the first to estimate the learning process in a consumer search framework. We found that consumers place a relatively high weight on newly observed prices when formulating their beliefs of the overall price distribution, a founding consistent with the observation that prices sampled today are more informative about the current price distribution than past price observations in a market of volatile prices.

Our modeling framework has many important advantages. First, we model consumers' search sequences of gasoline stations based on their travel patterns. Consequently, search sequences of stations are simulated based on the empirical distribution of traffic flows, allowing us to estimate consumer search and learning behavior using only aggregated station-level data. In addition, the model generates realistic substitution patterns that depend on the amount of common traffic shared between two stations. Second, the model allows for ax-ante vertical differentiation of stations and is not affected by dimensionality when integrating out individuals' search histories. We find that consumers respond to the ex-ante known differentiation of stations, including brand and price reputation by more than the actual dollar amount of the price difference. Although we do not explicitly model risk aversion in our framework, this result likely captures the value of a reduction in uncertainty, which a station or a brand provides to consumers through a consistently lower price. Third, we allow for uncertainty in consumers' beliefs about the current price distribution. Our counterfactual results suggest that information on the actual price distribution benefits consumers by an amount that is equivalent to a 1.12 cents in savings per gallon of gasoline purchased. On the other hand, perfect consumer information ex-ante about the price at each station generates an average of 1.60 cents of savings per gallon. Although the suggested benefit of reducing information uncertainty in the retail gasoline market seems small, it is non-trivial. It is also important to note that this result is obtained after controlling for the cost of substitution across spatially-differentiated stations, as well as the known average price difference across stations. Analysis that fails to take into account these two factors, for example if everyone buys from the lowest priced station in the city, would overestimate the benefit of increasing the amount of information.

There are several avenues for further development of our model, the most important being to model the supply side decision. By including the pricing decision of the stations, we would be able to analyze the amount of equilibrium price dispersion in the retail gasoline market caused by information uncertainty and spatial differentiation. In addition, our modeling framework can be applied to multiple cities and be used to empirically analyze the link between the speed of learning and the degree of asymmetric price response.

References

Arbatskaya, M. (2007). Ordered search. *The RAND Journal of Economics 38*(1), 119–126.

Atkinson, B. (2008). On retail gasoline pricing websites: potential sample selection biases and their implications for empirical research. *Review of Industrial Organization 33*(2), 161–175.

Bresnahan, T. F. (1987). Competition and collusion in the american automobile industry: The 1955 price war. *The Journal of Industrial Economics*, 457–482.

Chandra, A. and M. Tappata (2011). Consumer search and dynamic price dispersion: an application to gasoline markets. *The RAND Journal of Economics 42*(4), 681–704.

De los Santos, B. (2018). Consumer search on the internet. *International Journal of Industrial Organization 58*, 66–105.

De los Santos, B., A. Hortaçsu, and M. R. Wildenbeest (2012). Testing models of consumer search using data on web browsing and purchasing behavior. *The American Economic Review 102*(6), 2955–2980.

De los Santos, B., A. Hortaçsu, and M. R. Wildenbeest (2017). Search with learning for differentiated products: Evidence from e-commerce. *Journal of Business & Economic Statistics*, 1–16.

Dinerstein, M., L. Einav, J. Levin, and N. Sundaresan (2018). Consumer price search and platform design in internet commerce. *American Economic Review 108*(7), 1820–59.

Hong, H. and M. Shum (2006). Using price distributions to estimate search costs. *The RAND Journal of Economics 37*(2), 257–275.

Honka, E. (2014). Quantifying search and switching costs in the us auto insurance industry. *The RAND Journal of Economics 45*(4), 847–884.

Hortaçsu, A. and C. Syverson (2004). Product differentiation, search costs, and competition in the mutual fund industry: A case study of s&p 500 index funds. *The Quarterly Journal of Economics 119*(2), 403–456.

Kim, J. B., P. Albuquerque, and B. J. Bronnenberg (2010). Online demand under limited consumer search. *Marketing science 29*(6), 1001–1023.

Koulayev, S. (2013). Search with dirichlet priors: estimation and implications for consumer demand. *Journal of Business & Economic Statistics 31*(2), 226–239.

Koulayev, S. (2014). Search for differentiated products: identification and estimation. *The RAND Journal of Economics 45*(3), 553–575.

Lewis, M. S. (2008). Price dispersion and competition with differentiated sellers. *The Journal of Industrial Economics 56*(3), 654–678.

Lewis, M. S. (2011). Asymmetric price adjustment and consumer search: An examination of the retail gasoline market. *Journal of Economics & Management Strategy 20*(2), 409–449.

Lewis, M. S. and H. P. Marvel (2011). When do consumers search? *The Journal of Industrial Economics 59*(3), 457–483.

Matsumoto, B. and F. Spence (2016). Price beliefs and experience: Do consumers beliefs converge to empirical distributions with repeated purchases? *Journal of Economic Behavior & Organization 126*, 243–254.

McCall, J. J. (1970). Economics of information and job search. *The Quarterly Journal of Economics*, 113–126.

Moraga-Gonzalez, J., Z. Sándor, and M. Wildenbeest (2015). Consumer search and prices in the automobile market.

Moraga-González, J. L., Z. Sándor, and M. R. Wildenbeest (2015). Consumer search and prices in the automobile market.

Nishida, M. and M. Remer (2018). The determinants and consequences of search cost heterogeneity: Evidence from local gasoline markets. *Journal of Marketing Research 55*(3), 305–320.

Remer, M. (2015). An empirical investigation of the determinants of asymmetric pricing. *International Journal of Industrial Organization 42*, 46–56.

Rothschild, M. (1974). Searching for the lowest price when the distribution of prices is unknown. *Journal of Political Economy 82*(4), 689–711.

Stigler, G. J. (1961). The economics of information. *Journal of political economy 69*(3), 213–225.

Wildenbeest, M. R. (2011). An empirical model of search with vertically differentiated products. *The RAND Journal of Economics 42*(4), 729–757.