Gas Price Micro Markets

Kai Lukowiak

December 16, 2018

Abstract

Gas station prices are highly dependent on competition. Due to the nature of gasoline markets this competition is very local. Industry does not segment markets beyond the city level, potentially leaving profits on the table.

Stations are also able to set their own prices to some extent, introducing varibility into pricing. I use this variation to try and identify submarkts.

This paper investigates whether or not there are submarkets within cities using un-supervised clustering to first define markets that move together and then a KNN learning approach to see if there is a geographic component to these similar stations.

I found that this technique outperformed naive benchmarks but that the effect was not as strong as expected; there was significant geogrpahic overlay between different clusters. This finding suggests that the marginal consumer is willing to travel beyond a small, distinct area for gas. Alternatively, consumers may often travel accross many such areas in their drives and thus be free to choose the best prices.

1 Introduction

Gas price markets are heavily discussed in economic literature. Any basic econ course will use gas stations as an example of perfect competition (where prices fall to the marginal cost). However, as the figure below indicates, this is not totally constant.

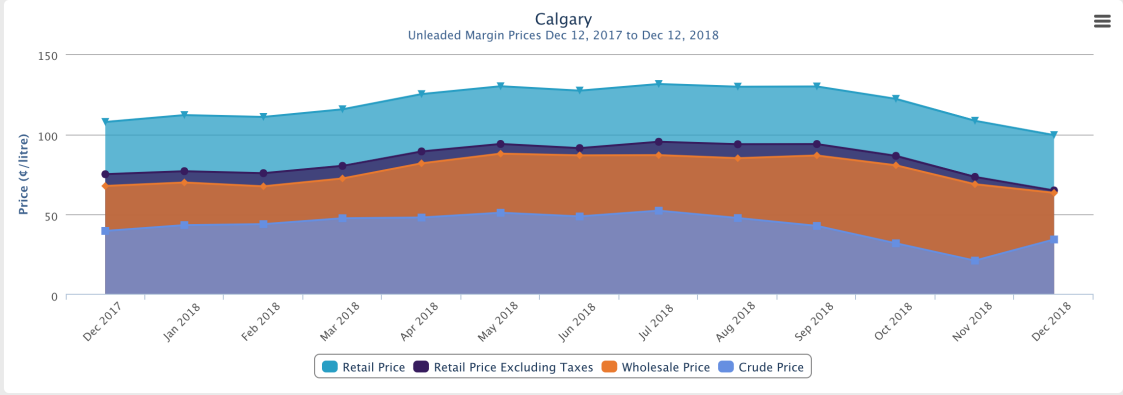

Figure 1: Gas Price Margins Over Time

Further, stations in these markets all buy at a similar price, but don't necessarily sell at the same price [1].

Gas prices are also heavily studied in academic literature. Many of these papers study the effect of near by gas stations[2]. These results did find a proximity component to gas station pricing but found this effect to be limited to around 1 km. Many of these studies are also linear in nature, using variants of OLS. These are useful and interesting. However, they do not effectively model the location of these sub-markets.

The constraint of linearity that is imposed means that we cannot see which areas are more constrained, or are actually sub markets. This is especially important because gas stations tend to be relatively evenly distributed. Thus, while near gas stations may influence prices, all stations are lain out evenly so if one station is not effected by another, there could be an intermediary station which would transfer the price.

There has also been investigation done on non-linear pricing areas, however, it focused on commuter routes. [3] While it did find a correlation, this analysis does not transfer easily into areas not on distinct commuter routes.

It is also possible that these spatial regression studies under estimate the importance of geography because the effect of one nearby prices in a near by but distinct market would lessen the spatial

[1] https://www.jstor.org/stable/1830441?seq=1#page_scan_tab_contents

[2] https://pdfs.semanticscholar.org/e00c/9ab1295963efe24aeeb75dba78f404af1b83.pdf

[3] https://www.jstor.org/stable/20111978?seq=1#page_scan_tab_contents

component of the regression.

My paper instead looks at how stations with similar pricing strategies are clustered geographically. This does not put a linear constraint on locations and, unfortunately does not estimate spatial effects. Thus, it is less interesting from a causal inference point of view, but might be interesting to industry and anti-trust/competition policy, where decisions on relevant markets are more important.

2 Materials and Methods

2.1 Materials

2.1.1 Price Scraping

Gasoline price data was scraped from GasBuddy.com every 6 hours. I used AWS, selenium, headless chromium, pandas, beautiful soup, and mysql to scrape and store the data. I had several issues implementing this with my automation chron jobs failing. I also had issues finding a consistent way to scrape GasBuddy because of the sites dynamic nature and variable add placement.

While there are holes due to these issues in my data, I don't believe they introduce a systemic bias due to the random nature of their occurrence.

2.1.2 Geo-location

Location data was scraped from Google Maps' API using unique address from the GasBuddy scrape.

2.1.3 Rack and Taxes

Tax information was taken from Wikipedia. [4] Rack prices were downloaded from nrcan.gc.ca. [5] These were applied to the raw gas prices to get the margins.

Data on margins is important because the rack price is one of the most important factors in determining pump prices. Correcting for this makes our analysis more robust to swings in the price that are out of retailers hands.

2.2 Methods

2.2.1 Data Manipulation

The scraped data was joined with the location information from Google Maps. It was then cleaned to remove duplicate price points. For example if a certain station had was scraped one hour after it's price was updated on GasBuddy and again seven hours after, the duplicate, later data was removed.

With only the earliest input data still available, the data was transformed into a wide format with each column being a site and each row representing a point in time. This dataframe was then reindexed to hourly intervals using pandas's re index function. Missing values were then filled in with the nearest available data point. This was limited to 48 hours to ensure that 'stale' prices were not included.

I also found the daily averages by city and subtracted this from the margins. This was an attempt to further remove trends in the pricing data that might effect clustering.

This clean data was then fed into R. Missing values were removed from both rows and columns in a way that minimized data loss. First columns with exceedingly high amounts of missing values were removed. This greatly reduced the number of missing values in the rows, however, due to scraping problems data towards the end of the scraping interval contained more missing values. This was

[4] `https://en.wikipedia.org/wiki/Motor_fuel_taxes_in_Canada`

[5] `http://www2.nrcan.gc.ca/eneene/sources/pripri/wholesale_bycity_e.cfm?priceYear=2017&productID=9&locationID=8,10,6,9,2,3&frequency=D#priceGraph`

then removed. Overall, this process lead to approximately 600 sites having enough continuous observations. A more naive approach would have eliminated all but 83 sites, or the majority of the days.

The wide time series data was then turned into a time series matrix using R's eXtensible Time Series (xts).[6]

2.2.2 Time Series Clustering

The time series matrix was then fed into R's TScluster. [7] Two clustering methods were used. Both raw margins and margins that had been adjusted for the daily average price for the city were tested.

2.2.3 Correlation

The first was a simple time series correlation. Correlation was chosen because it effectively identifies the relationship between two time series. I correlated the margins of all stations over time with each other to create a large, square matrix.

Correlation is defined as:

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{(x-\bar{x})^2}\sqrt{(y-\bar{y})^2}}$$

When computing the distance function of a correlation the following formula is used to create a distance matrix:

$$d = 1 - |r|$$

Correlation has benefits and drawbacks. It can be thought of as the relation between x and y. This might seem ideal, but existences of a trend can greatly overstate the relationship. This problem is partially controlled for because of our use of margins instead of prices. This fairly effectively

[6] `https://cran.r-project.org/web/packages/xts/xts.pdf`

[7] `https://cran.r-project.org/web/packages/TSclust/index.html`

de-means the variable. However, within cities, there can be trends of high and low margin. In fact we see this play out in the results section where correlation was more efficient at differentiating cities and the adjusted data was superior within cities. (See figure 2 which illustrates the effect of correlation on two different random walk time series, one with a trend).

2.2.4 Euclidean Distance

Euclidean distance measures the distance between two points. In 2-dimensional space such as a time series the formula takes the form:

$$D = \sqrt{(x_i - y_i)^2}$$

This will not overstate the effects of a trend on the results however, there are downsides to using this approach. Most significantly, this approach would penalize sites that effected each other, but due to branding or convenience had a sustained price differential.

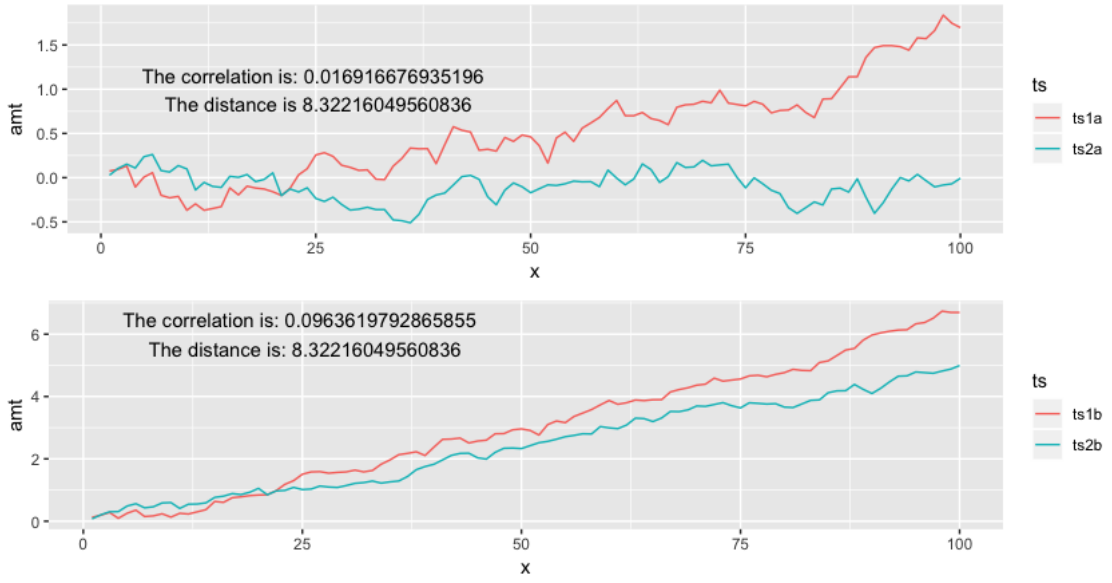

Figure 2: Correlation vs. Euclidean Distance

2.2.5 Other Distances

There are a multitude of other distance metrics available to differentiate time series data (see the package TSclust). The most notable of these are Dynamic Time Warping (DTW) and Frechet Distance. Both these distance metrics attempt to remove issues with the lag between different time series and are popular with voice recognition algorithms because they can cluster curves even if the period of them is different.

- DWT links points along a curve with the closest ordered match. Thus, a sin and cosine wave would have zero distance because the algorithm would adjust for the difference in period.
- Frechet Distance is often analogized as the distance between two paths held together by a leash. Thus, the distance can be thought of as the maximum length of a leash needed to walk a dog on a certain path.

I decided against exploring these options for two reasons. First, I doubted the performance would be very high because the way they mask time dependence means that relevant market fluctuations could be obscured. For example, if two stations increase their prices in a very close time frame, this makes them more likely to be related. If the time difference between price changes is greater, the likely hood they are related decreases. Both DWT and Frechet could ignore this.

The other reason I did not model these distance metrics is that they are very computationally expensive. Unfortunately, running the code on my computer took a prohibitively long time and after all the time scraping, I did not have room in my budget for high powered computational AWS instances. Further research in this area would be very interesting.

2.2.6 Clustering

R's hierarchical clustering function was applied to the distance matricies in order to group sites with their closest counterparts. This new cluster is then averaged and clustered with the next closets site or cluster.

As the algorithm proceeds, the height of the graph increases until there are only two clusters left. The final cluster is not modelled because it would contain all sites and is thus trivial.

The figure below shows the results of this on a subset of our data.

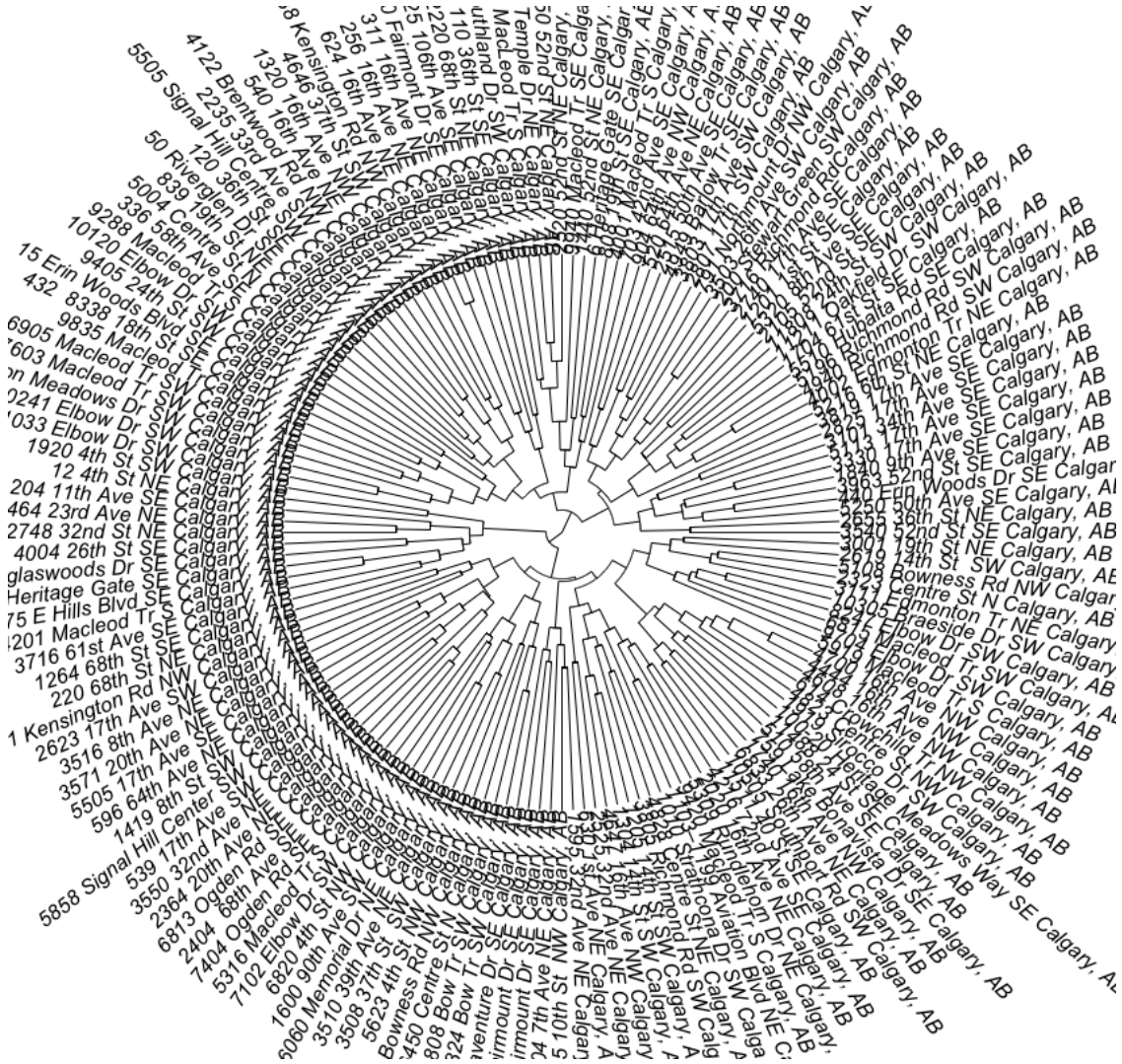

Figure 3: Adjusted Correlation Cluster in Calgary

To obtain labelled data, I arbitrarily chose the number of clusters. This in effect cuts the diagram at a certain height, creating numbered sub groups.

These groups were then applied to the relevant addresses and turned into factor variables. These labeles were then used to see if there was a spatial component to the data.

2.2.7 KNN

K^{th} Nearest Neighbours is an algorithm that uses labels of the K nearest data points to vote on an data point. This can be applied to N-dimensional space, although performance becomes poor when N becomes larger.

The earth, when comparing a small enough geographic area, can be considered flat. Thus latitude and longitude can be used to find the distance between two points.

Generally the best practice for KNN is to scale each variable to zero mean and standard deviation of one. This is because KNN is susceptible to overstate the importance of variables with a large numeric value. In OLS, if a feature is scaled by x the corresponding value of the model is scaled by $\frac{1}{x}$ and other variables are unaffected, as is the accuracy of the model.

With KNN, distances are calculated based on Euclidean distance and are effected by scale. For example, using age and income to classify people into different groups would lead to income being overstated because it would range from 0 to over $100,000. Therefore, the variables need to be scaled to the same distribution.

In this case, scaling the latitude and longitude is unnecessary because they are of the same magnitude. Leaving the coordinates in raw form also allows a model looking only at one city to be compared to a model using all of Western Canada without worrying that the scaling factor would be totally different.

Because no city had fewer than 40 gas stations and the max allowable value for K in my algorithm was 20, no different site would be compared to sites in another city.

KNN is an interesting way to evaluate labelled data because we can see how much near by stations lead predict the classification of a new station. This algorithm also is useful because for any geographic point, we can guess which cluster it will be in. This is very useful for industrial planning of new sites.

2.2.8 Specifics of KNN

I used R's `caret` package to run the KNN algorithm with 10 fold cross validation with the model evaluated on accuracy. This cross validation was used to select the optimal K, with values ranging from 1 to 20.

Once the model was selected, it was evaluated on the testing values, composed of 30% of the entire dataset.

2.2.9 Model Evaluation

Classification models are usually evaluated based on their accuracy (percent correct classification) as well as a confusion matrix which shows the true values and the results of classification.

This gives a good idea of how the model is performing, but does not the next best alternative. This paper's thesis is that there are sub markets within cities. The null hypothesis is that there are none. If there are no geographic sub markets, one would assume that the most common cluster in a city would be the best predictor of a site's class, instead of the KNN cluster class. I therefore evaluated each classification using the modal class for the city.

3 Results

3.1 Overall Performance

The results of this experiment are mixed. There is clear evidence that clustering works, especially on the stations that are very near to each other. As you can see in the figure 4 many of the sites that are on the same road and near by, have the lowest cluster levels. This is common across all cities and for all clustering methods. However, somewhat surprisingly, the results for a city level cluster, even for the un-adjusted dataset, did not match each city very well, refer to table 1.

This result is particularly surprising because most companies operate on the idea that cities are more similar. However, our algorithm did a good job at matching very close clusters, but only a

mediocre job clustering cities.

It is possible that the relationship between cities is overstated in industry. This results will be discussed further, but it poses a problem because I sought to find sub markets, and I have issues finding intermediary markets.

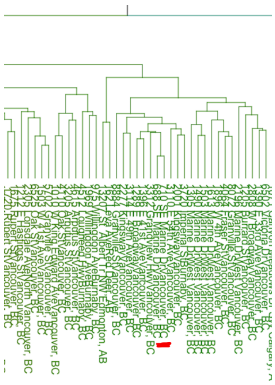

Figure 4: Cluster Example

The follow table demonstrates that there is a definite relation on the city level.

3.2 Visual Analysis Single Market

While lacking a robust analytical theory, visual analysis of the potential clusters is important because it helps us see if there are potential clusters and evaluate different models.

Later I will use a more robust analysis of the different models, but it is important to understand and visualize what the data actually looks like.

This image is exemplary of the issues of hierarchical clustering. Except for the adjusted correlated cluster, most of Calgary is labelled as only one class. This is problematic because when we look at Table 1, we see that Calgary does not necessarily conform to a single cluster when compared to other cities.

The one graph that show some interesting results is the adjusted correlation of two clusters for Calgary. Even in this plot, there are several sites that are clustered as class 1 but well within an area predominantly of class two.

Table 1: Raw Correlation Cluster by City

area name	Cluster	Count by City
vancouver	$Cluster_1$	72
victoria	$Cluster_2$	38
calgary	$Cluster_3$	25
vancouver	$Cluster_4$	34
red deer	$Cluster_2$	32
calgary	$Cluster_5$	131
edmonton	$Cluster_5$	136
red deer	$Cluster_3$	5
edmonton	$Cluster_3$	33
kelowna	$Cluster_6$	32
vancouver	$Cluster_6$	1
red deer	$Cluster_6$	2

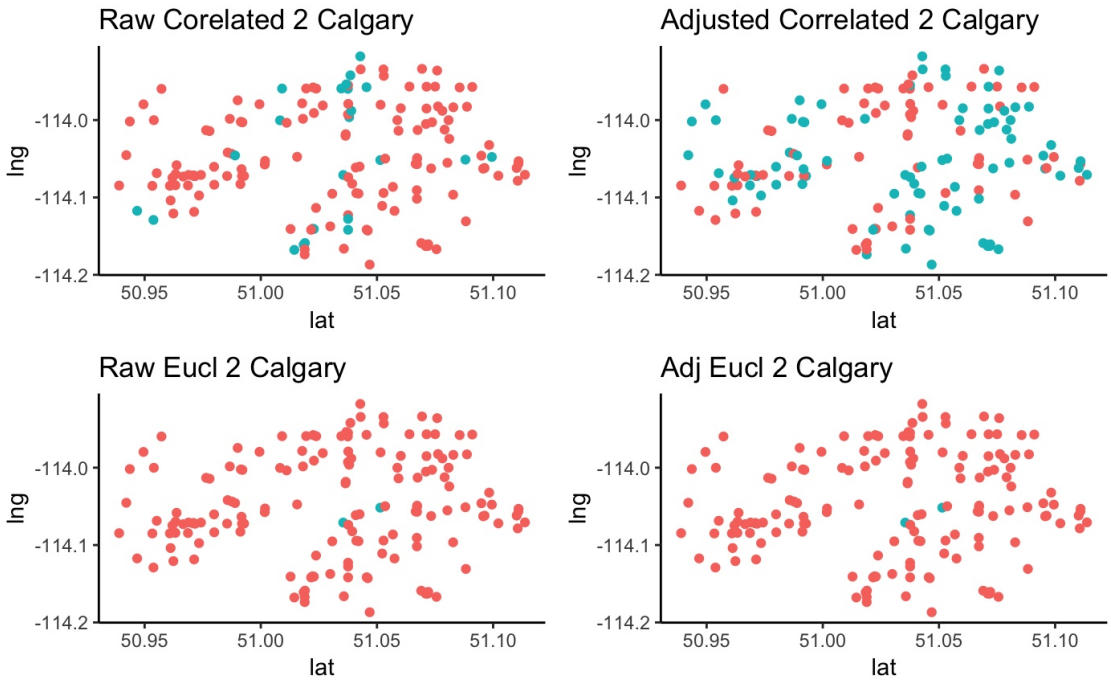

Figure 5: Different Methods of Correlation for Calgary

If we increase the number of clusters, we get more promising results.

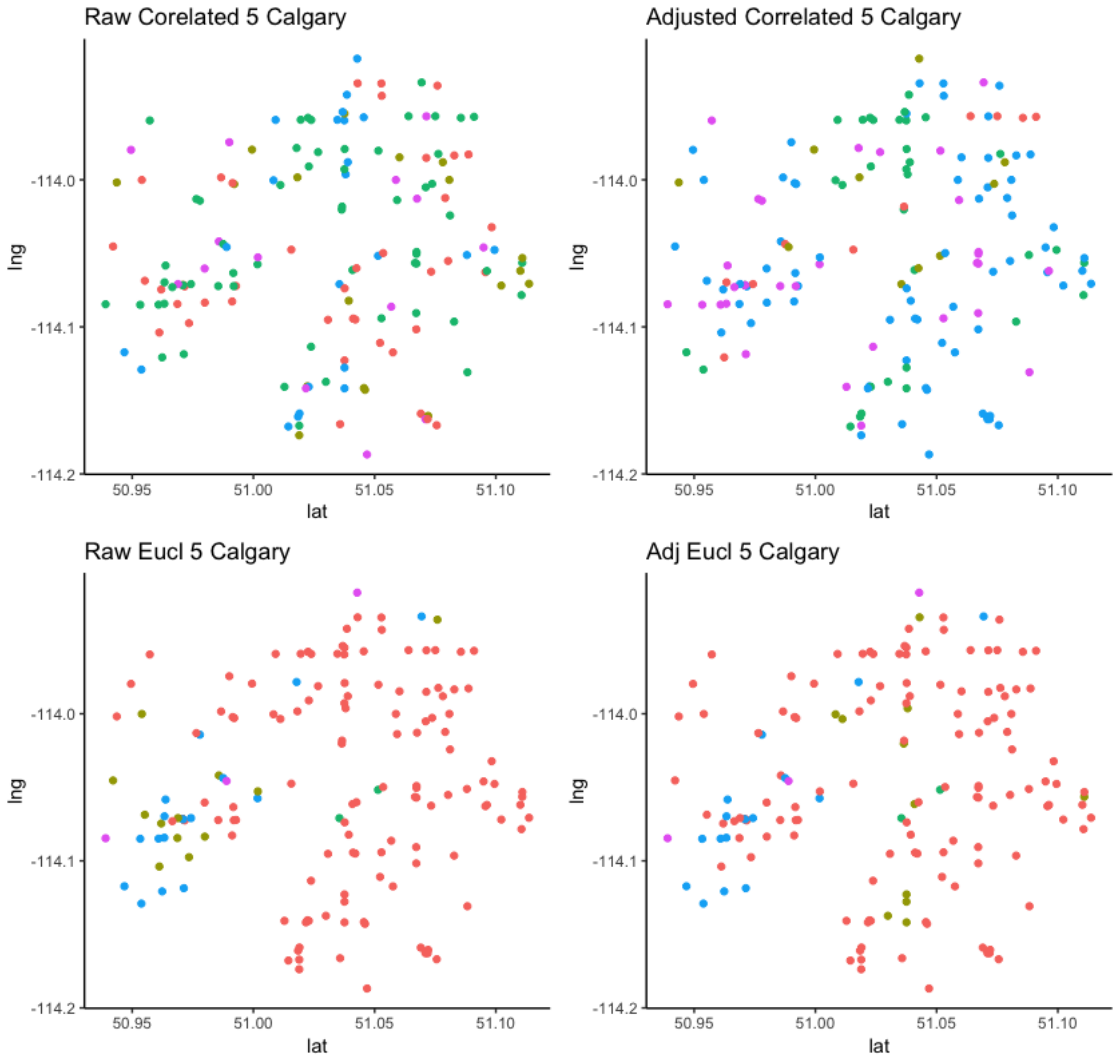

Figure 6: Different Methods of Correlation for Calgary 5

6 give more options for clusters. There also seems to be a distinct area in the south west of Calgary that differs from the rest. It's also interesting that Euclidean distance seems to outperform the correlation based on the fact that there are more tightly spaced groups of one particular cluster.

However, while there are areas that seem distinct, they are not as distinct as one would expect if there were separate sub markets.

3.3 Visual Analysis Multi Market

The results for multi market graphs are, somewhat surprisingly, more segmented.

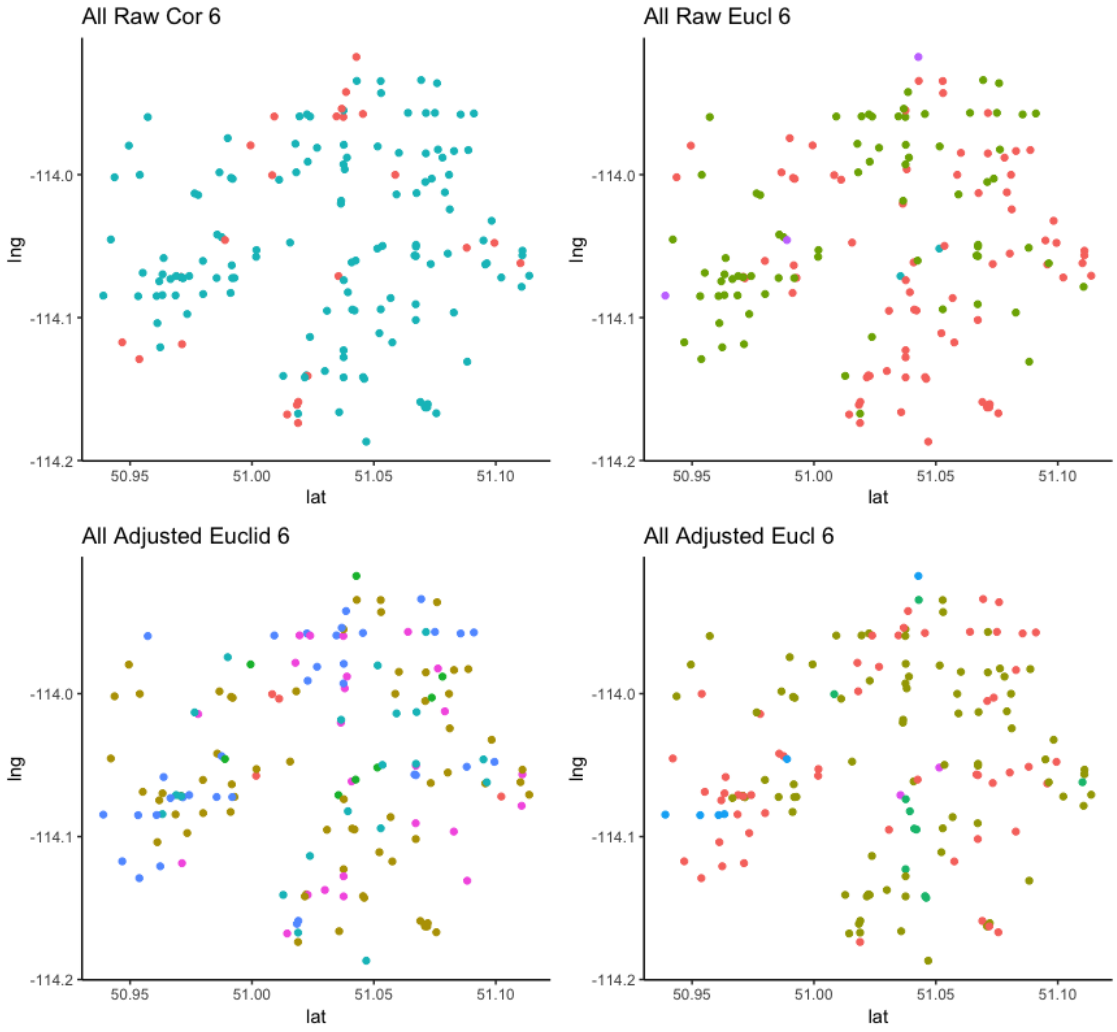

Figure 7: Western Canada 6 Calgary

When we apply six clusters to all of western Canada we get more interesting results. While there looks like there are some interesting clusters, each area still seems to have a few outliers that make me believe that there are no totally separated segments.

3.4 KNN Analysis

So far this analysis has amounted to little more than visual data analysis. While this can help us see groups, it lacks rigour and is susceptible to human bias.

Table 2: KNN Classification Results for Euclidean Raw 6

actuals / preds	Cluster_1	Cluster_2	Cluster_3	Cluster_4	Cluster_5	Cluster_6
Cluster_1	32	0	0	0	0	0
Cluster_2	0	20	0	0	1	0
Cluster_3	0	0	29	18	2	1
Cluster_4	0	0	17	28	0	0
Cluster_5	0	0	0	0	10	0
Cluster_6	0	0	1	0	0	0

```
  Accuracy : 0.761
    95% CI : (0.687, 0.825)
No Information Rate : 0.2956
P-Value [Acc > NIR] : < 2.2e-16
```

We see that for several clusters, the spatial clustering accuracy is quite high. The overall accuracy for this is 0.761. Naively, we could compare this to $\frac{1}{6} = 0.1666667$ and be quite impressed with the accuracy. But the next best alternative is to simply compare how using the most common value by city would compare.

The most common clusters by city are given in table 3 If we compare these clusters in figure 3 we get the null hypothesis accuracy.

This gives an accuracy of 0.6326531 which is less than the 95% CI for the model. Therefore, while the model might not perform quite as well as expected, there is evidence for a spatial component.

Table 3: Most Common Cluster by City

area name	cluster
calgary	$Cluster_4$
edmonton	$Cluster_3$
kelowna	$Cluster_5$
red deer	$Cluster_2$
vancouver	$Cluster_1$
victoria	$Cluster_2$

Table 4: Most Common by City Cross Table

raw_{cor6}	$Cluster_1$	$Cluster_2$	$Cluster_3$	$Cluster_4$	$Cluster_5$	$Custer_6$
$Cluster_1$	73	0	0	0	0	0
$Cluster_2$	0	70	0	0	0	0
$Cluster_3$	0	0	166	156	0	0
$Cluster_4$	34	0	0	0	0	0
$Cluster_5$	0	3	0	0	32	0
$Cluster_6$	0	4	1	0	0	0

4 Discussion

Existing literature, along with economic theory suggests that proximity increases the correlation between gasoline prices. My KNN analysis seems to suggest that this could be a way to 'see' these sub markets.

There are several ways I think this analysis could be improved if my time and computational budget were increased.

4.1 Issues with Analysis

Data collection was not as granular. Scraping every six hours did not give enough information on quick, reactionary changes, although the main problem is that GasBuddy data is not updated with great precision. This means that competitive actions between stations could be missed.

A more robust data source would make this a lot more interesting.

It is also possible that there has not been enough price exploration in these sub markets for them to be seen. Imagine a hypothetical scenario with two large gasoline retailers in a city. If they do not differentiate their prices by sub markets my analysis would not be able to pick out potential markets because the entire city would face the same price regime. Thus, depending on how much autonomy each stations has, there may not be enough competitive variability, even if there was the potential for it.

4.2 Further Research

I see three main areas of further research in this area (asides from improving the data). The first is exploring different time series similarity methods such as limited DTW or Freshet distance. The second is to explore different cluster methods, beyond the cut tree method. Finally, using a controlled experiment to investigate sub markets would offer the gold standard to understanding if sub markets exist.

4.2.1 Similarity Methods

It is possible to set limits on DTW so that finding areas completely out of sync are not clustered together. This was computationally expensive to investigate but deserves further research. Freshet distance also bears more investigation.

While I cannot say for sure if these are valuable, looking into them is an important step to rule out possible model miss-specification.

4.2.2 Clustering Methods

In this paper, I used a cut tree method which draws a line at a certain height and uses these clusters. This has several downfalls. Specifically, it can lead to unbalanced number of classes, lots of one and few of another. It also make groups have more variability then might be desired.

Using data entropy to variable tree depths could help adjust for this.

In figure 8 the black horizontal line is the level chosen to achieve 6 clusters in our analysis. The blue step function show a possible cut off that would keep clusters of a similar size and reduce entropy.

4.2.3 Controlled Experiment

This analysis relied on sites adjusting to the competitive market as a source of variability. If a gasoline firm desired to implement this, it should assign, at random, stations to increase or decrease their price from what it would otherwise have been. This is a far better way of investigating the question. It would also be useful to have volume data to see if other nearby stations reacted to this over what would be expected.

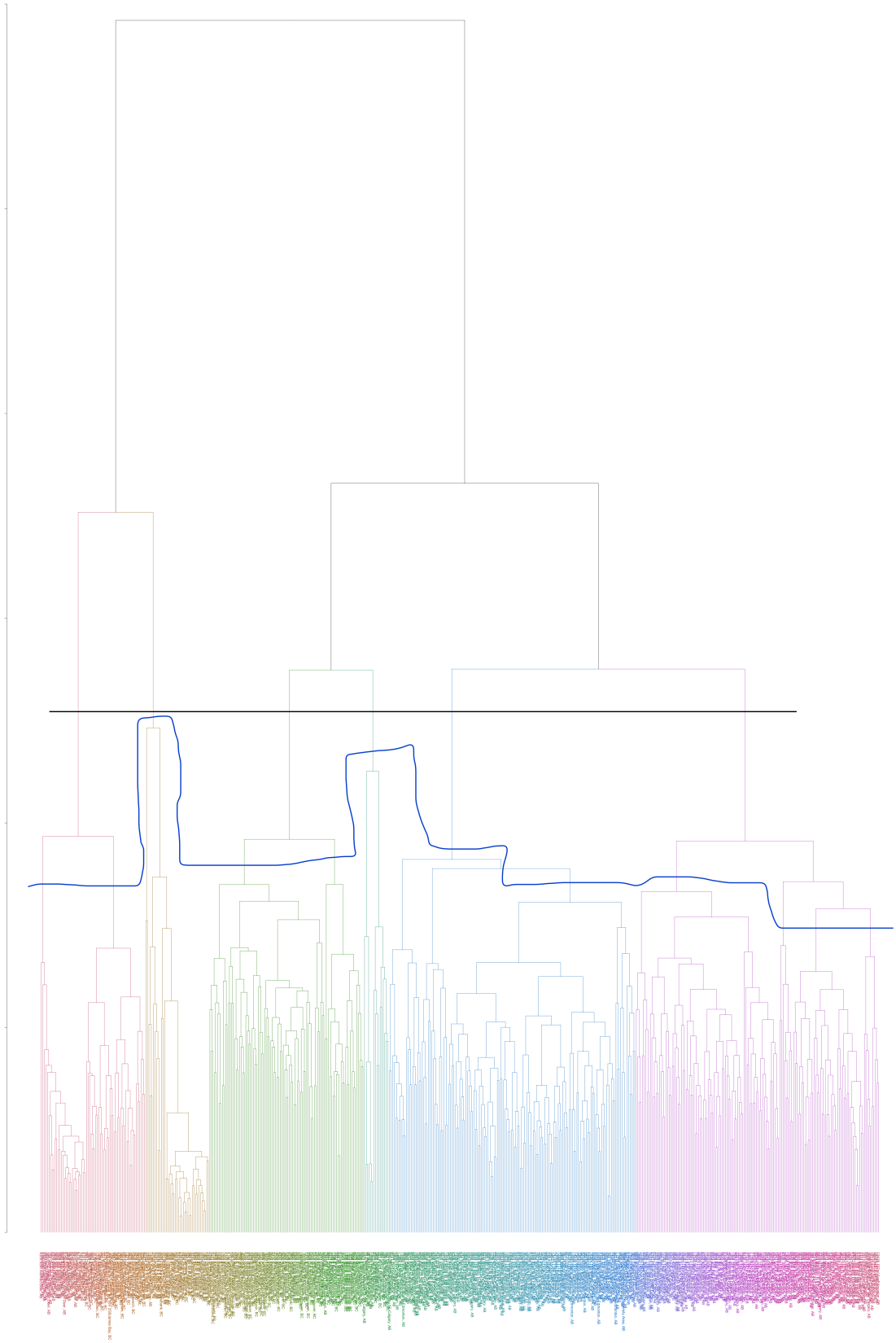

Figure 8: Different Cut Tree Methods

5 Conclusion

It appears that stations very close together to each other cluster well. It also seems that there are modal groups in each city, but that each cluster does not conform to different cities, even when using the non-adjusted margins. This is somewhat surprising because most retailers assume that city level markets are distinct. These results suggest that there is more inter market-variation.

This result suggests that there is a local relationship, but this does not completely transfer to the city level. Looking at the results between the most local level and the city level, there appears to be a spatial relationship, or sub markets.

While these results did not conform to my expectations, there are potential interesting results or clusters. These should be investigated further.