# # Introduction {-}

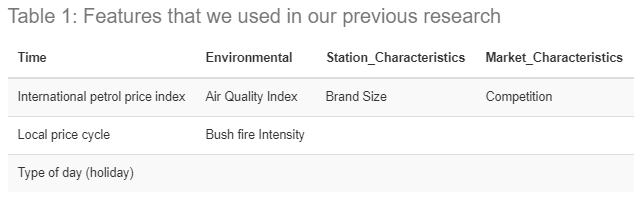
According to the Australian Bureau of Statistics, around 15% of household spending is destined for transportation [@absMainFeaturesAverage2017]. Considering that 57% of people use cars as their primary means of transport [@absWorkingPopulationProfile2016], a high proportion of the household budget is spent on petrol. This makes people extremely sensitive to changes in petrol prices, which can vary mostly due to international variables. However, according to our previous research, petrol stations have significative differences in prices across Greater Sydney, where competition seems to be one of the factors that influence these local variations.

Nevertheless, how competition affects petrol prices is not constant across different petrol stations clusters. Therefore, it could be affected by market characteristics and even possible collusion between retailers. According to @byrneLearningCoordinateStudy2015, this last factor was present in Perth petrol market, but there was no evidence to call it formal collusion, just tacit.

The propose of this research is to understand how competition affects retail petrol prices in Greater Sydney using spatial analysis, so we can capture how it differs across different locations. The outcome of this analysis may be helpful for organizations, like the ACCC, whose aim is to promote competition and fair trading. Therefore, these results could help them to focus their investigations on specific risk areas.

# # Background {-}

In our previous analysis, we aimed to evaluate a series of factors that affect petrol prices for three years of data. We used variables that capture variability across time (seasonality and trend), environmental status, stations attributes and market characteristics. Table \@ref(tab:prevvariables) shows the details of the variables that we used.



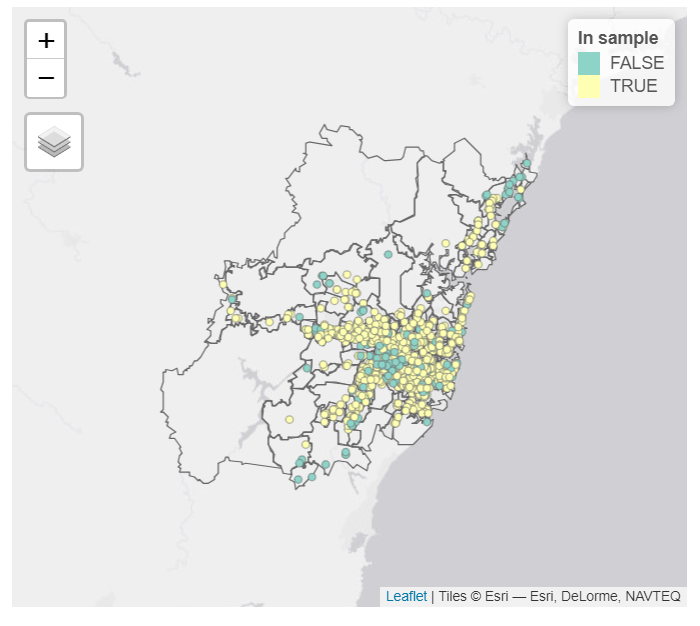
There were several limitations in that research that could have influenced the results. Firstly, we summarised the data of same-brand stations in each suburb; therefore, local variation was underestimated. Additionally, we utilized linear regression, which assumes that the observations are not autocorrelated through space and time, which is counterintuitive because consecutive days and close stations have similar prices. On the other hand, the competition variable did only indicated if a rival brand was present in the suburb, not considering other closest stations in different suburbs or the increase in competition when more than one brand was present in the area.

Considering those reasons, it was not possible to use our previous analysis to understand the effect of competition on petrol prices. Therefore, if the assumption of spatial autocorrelation is correct, spatial analysis and the geographically weighted regression (GWR) appears as a good alternative.

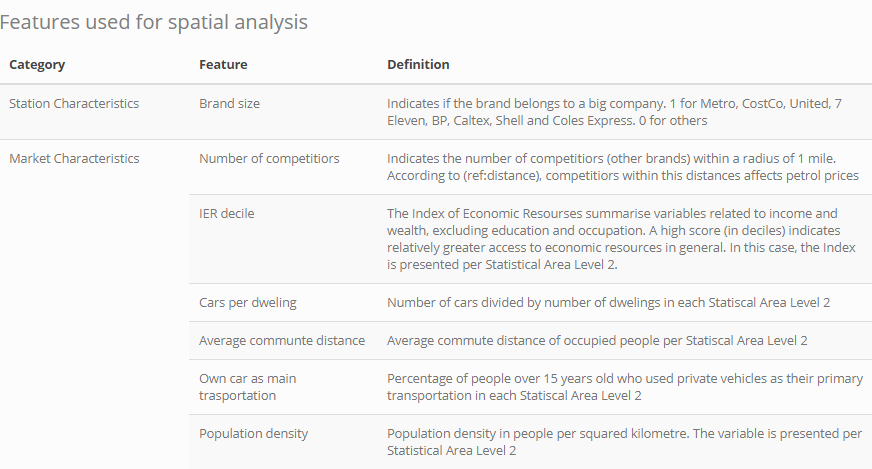
# # Data Understanding {-}

To start the spatial analysis is necessary to have data without temporal autocorrelation. For this reason, I changed the data and variables that we used in our first research. In regards to the data, only the day that represented better the population of petrol stations was selected to remove temporal autocorrelation. In this case, November 29th of 2019 had the prices for 87% of Greater Sydney petrol stations. Moreover, the fuel E10 was selected because it was the one with more observations.

The next map shows the petrol stations in Greater Sydney with at least one observation between November and December of 2019, where the yellow dots correspond to the stations of our sample (November 29th of 2019). It can be noticed that the distribution of the sample is similar to the population of petrol stations.

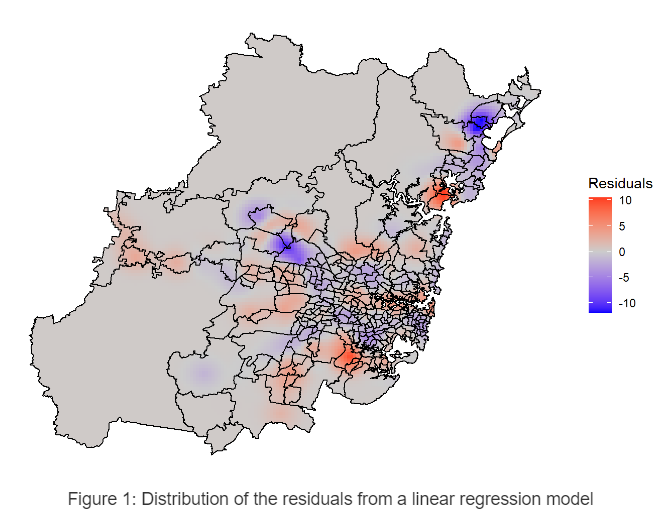


The features that I used for the analysis could be divided into two groups, petrol stations characteristics and market characteristics. It is worth mentioning that the number of competitors is the only variable that is capturing competition. Table \@ref(tab:currentvariables) shows a brief description of each variable.

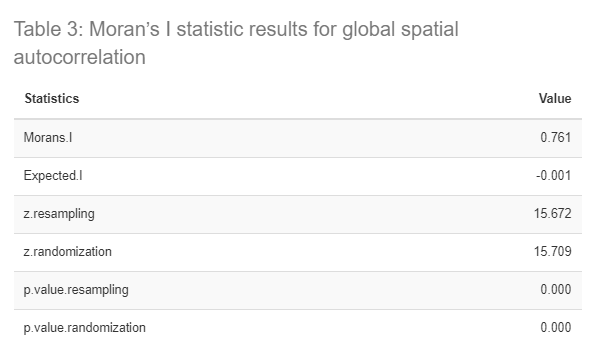


## ## Spatial Autocorrelation {-}

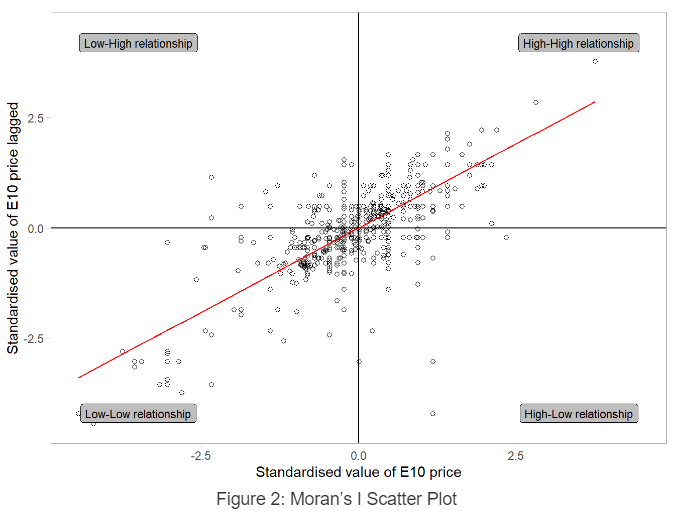
Spatial autocorrelation measures how close observations (or neighbours) influence a variable, in this case, E10 price. When spatial autocorrelation is present, the residuals from a linear regression model, tend to have a spatial dependency. Figure \@ref(fig:residualsplot) shows the distribution of the residuals from linear regression (with E10 price as the independent variable) in Greater Sydney, where we can see that similar residual values are grouped. In this case, I used Kriging interpolation to fill the gaps between petrol stations.



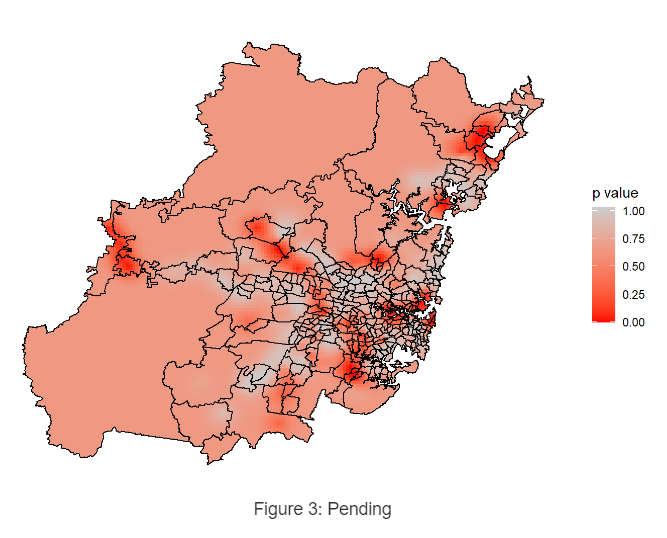
Additionally, a statistical test was used to prove spatial autocorrelation. The statistic that I used is called Moran's I, which evaluates whether the pattern expressed is clustered, dispersed, or random [@esriHowSpatialAutocorrelation2020]. The hypothesis test related to this statistic has random spatial distribution as the null hypothesis and presence of spatial clusters as the alternative hypothesis. Table \@ref(tab:globalmorantest) shows a p-value close to zero; therefore, the null hypothesis is rejected, and we can assume that there is global spatial autocorrelation at the 0.05 significance level.



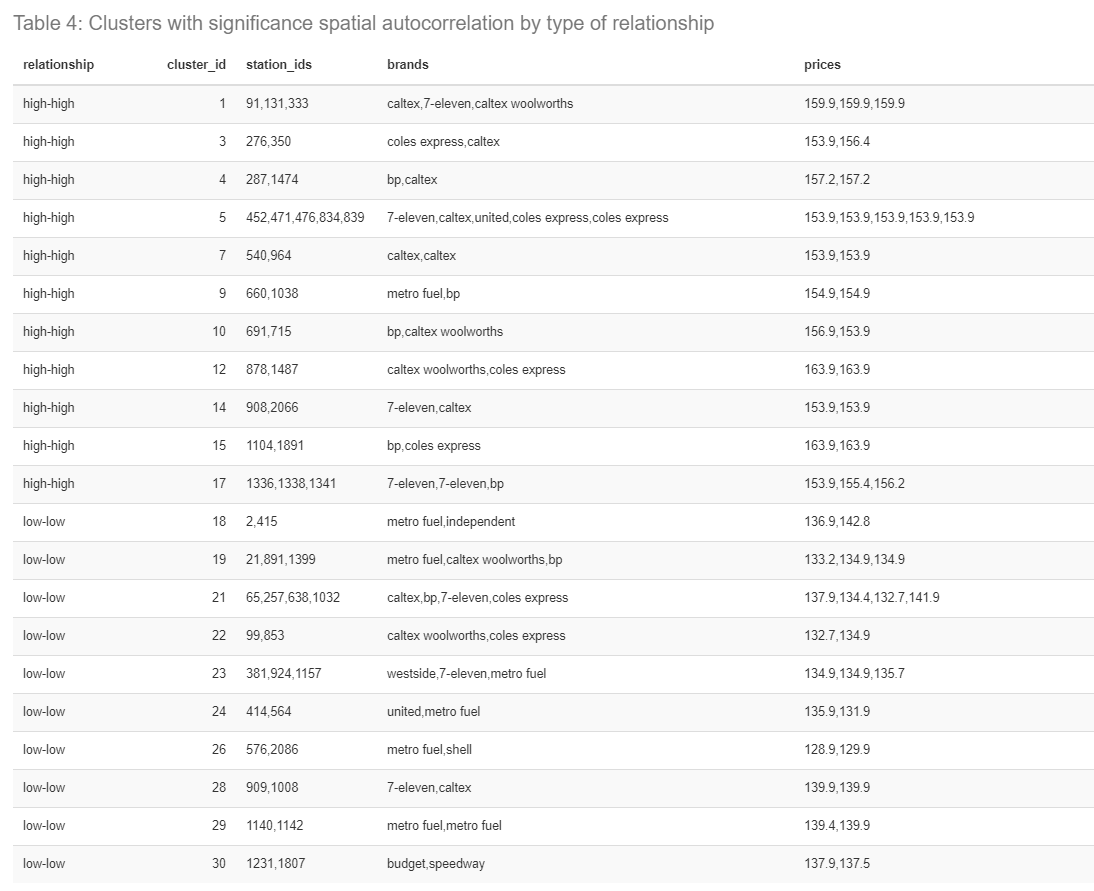
Moran's I statistic can also be calculated locally, so we can see where these clusters formed. Figure \@ref(fig:moranplot) shows how each observation is related to their neighbours; therefore, the diagonal line corresponds to a positive spatial autocorrelation. However, there are points separated from the diagonal line, which means that the relationship between neighbours is different. Four types of relationships are shown in each quadrant of the Figure, where the most important for this research corresponds to "high-high". This relationship means that where the petrol station has a price over the average, their neighbours also have it.



Nevertheless, there is no spatial autocorrelation in all locations; therefore, not all relationship are significative. Figure \@ref(fig:localmorantest) shows the areas where Moran's I statistic enable us to reject the null hypothesis, which means that we can assume that there is spatial autocorrelation (in intense red).



According to @lundbergCartelDetectionMoran2017 the Moran's I test for spatial correlation could be used to detect potential price coordination, which in this case means clusters of petrol stations with high prices, not being affected by the nature of competition. Table \@ref(tab:clusters) shows all the clusters with significative spatial autocorrelation and their type of relationship. For this task, only the clusters with relationship "high-high" are considered as possible collusion due to "low-low" is a signal of high competition. It is worth noticing that some of the clusters have only stations of the same company; therefore, it can not be considered as possible collusion.



# # Modelling {-}

The next step consists of understanding how each of the features exposed in Table \@ref(tab: currentvariables) affect E10 price in the clusters. We discovered that spatial autocorrelation was present in the data, which means that some E10 prices are highly correlated with its neighbours, but not related to prices of distant stations. This means that features have a different effect on price depending on the location; therefore, a linear regression model would not be able to capture the local variations.

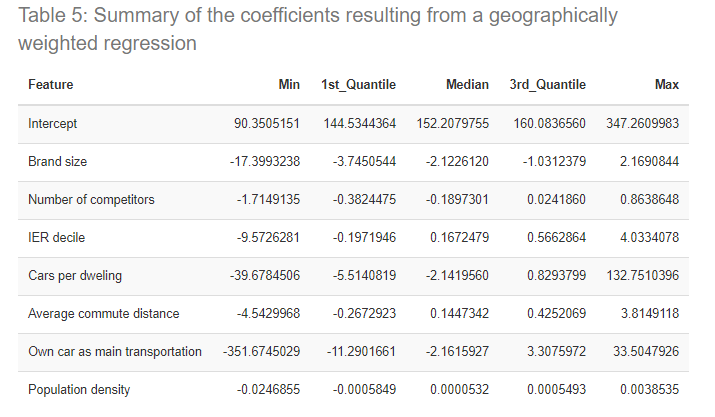
On the other hand, the geographically weighted regression (GWR) can capture spatial variation in the coefficients. This exploratory technique work in a similar way as linear regression; however, it assigns different weights to each observation, which values are given based on distance. In other words, data points more close to the regression point are more important in the regression than far points [@fotheringhamSageHandbookSpatial2009]. Therefore, the weights and the resulting coefficients change every time the regression points change.

The weighted function could be categorized into two types; fixed or adaptative. In this case, I used an adaptative weight function; therefore, the weights depend on the density of data points in the area and not only on a fixed distance. In this dataset that means that in more dense regions (regarding to petrol stations) the distance to high weighted observations will be lower than in areas with few data points.

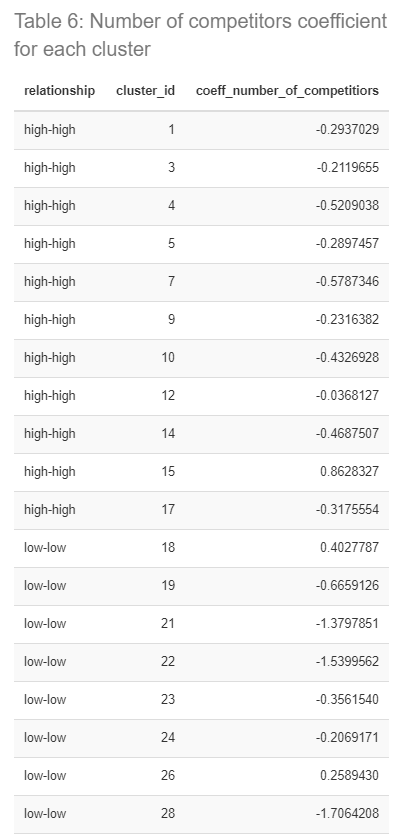
## ## Results {-}

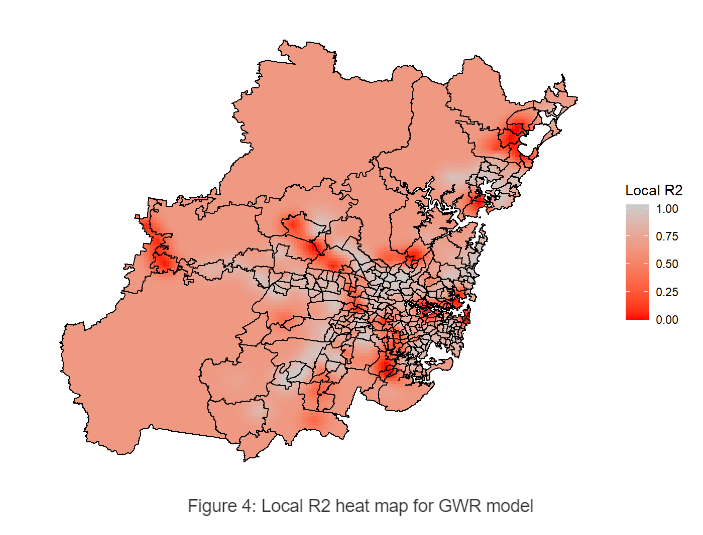
A summary with the coefficient of the geographically weighted regression with adaptative weight function is shown in Table \@ref(tab:gwr). In this Table we can see some measure of the distribution of the coefficients, which vary across space. Due to the objective is to find information that can expose signs of collusion, special attention is given to the variable "Number of competitors".

According to Table \@ref(tab:gwr) the coefficient of "Number of competitors" range from -1.71 to 0.86, which means that in some places more competitors decrease petrol prices and in others, they increase it.



In Table \@ref(tab:clusters) we saw the clusters that are more suspicious in case of collusion; therefore, I only did an analysis of coefficients of those clusters. Table \@ref(tab:clustercoeff) shows the coefficient of "Number of competitors" for each cluster\_id, where no significant difference between "high-high" and "low-low" relationship was found.



In regards to how the model explains E10 prices, the Quasi-global R^2^ (which indicates how much variation of the prices is capture by the model) corresponds to `r globalr2`, which is far superior compared to the R^2^ obtained from a linear regression (0.21). However, due to spatial dependency, the R^2^ from a GWR is not constant through space. For this reason, the local R^2^ is calculated, so we can see it in each location (see Figure \@ref(fig:localr2)). It is worth noticing that the areas with significance spatial autocorrelation are also the areas with less local R^2^.

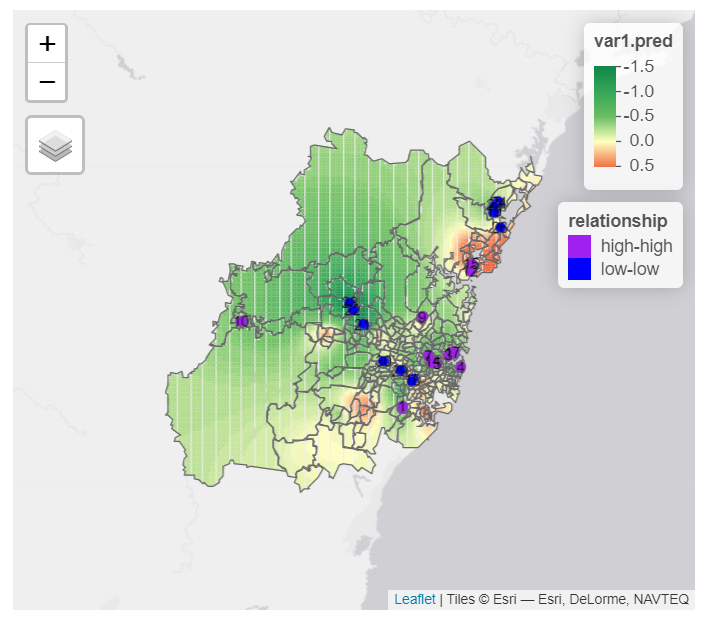
## ## Interpreting results {-} (300)

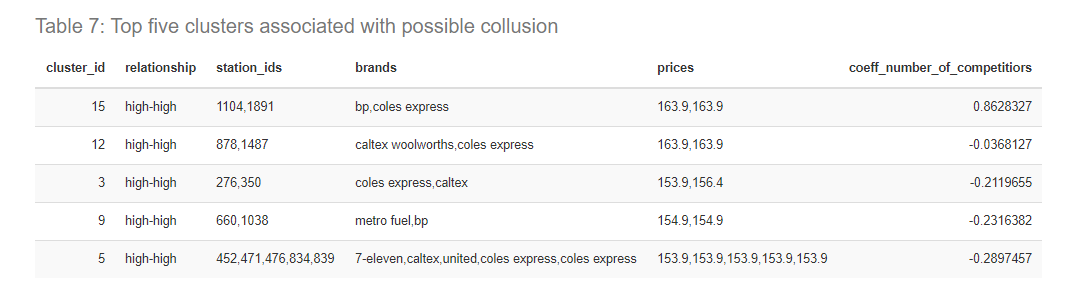
The focus of this research is to find clusters where collusion could be present; therefore, the first step was to find areas where E10 prices were spatial dependent. Using the Moran's I statistic twenty-one clusters were found, which correspond to two main categories, "high-high" and "low-low". While "low-low" category could indicate lower prices due to competition, "high-high" could reflect price coordination used to achieve higher prices.

The second step consisted in understanding how the number of close competitors affects petrol prices in the clusters previously described, which is capture by the coefficient of the variable "Number of competitors". The coefficient of the feature was divided into three categories, where the presence of one of the last two could reflect signs of collusion:

* Coefficient < 0: The market has not reached an equilibrium; therefore, when a new competitor enters, the prices decrease.
* Coefficient > 0: The market has reached an equilibrium; however when a new competitor enters, prices increases. This could be a signal of price coordination.
* Coefficient ≈ 0: The market has reached an equilibrium, which could be natural or artificial. This means that when a new competitor enters, the prices stay the same. The equilibrium is artificial when it delivers greater profits than those available in a static Nash equilibrium [@byrneLearningCoordinateStudy2015].

The following map shows the clusters divided by categories and a heat-map with the value of the coefficient of "Number of competitors". Special attention has to be given to "high-high" clusters (in purple) in zones where the coefficient is close or more than zero (yellow and red zones) because they are the clusters more suspicious of having unfair trading practices. Table xxx shows the top five clusters associated with possible collusion.





# # Conclusion {-}

This research presented how to use the Moran's I statistic in conjunction with the geographically weighted regression to obtain petrol station clusters with possible price coordination. However, this methodology must not be used as the only tool to detect collusion, I only recommend it as a reference for resources allocation, so fair-trading agencies could know where to start further investigations. Moreover, any possible collusion detected will be difficult to verify because there is no explicit communication and coordination between companies, as stated in @byrneLearningCoordinateStudy2015 research. This type of coordination, called tacit collusion, is easier to develop in situations where competitors' prices are publicly available, which is the case here with platforms like (Fuel Watch)[<https://www.fuelcheck.nsw.gov.au/>].

On the other hand, due to the model simplicity, the results could be out of touch with the market reality. There are three main limitations in the methodology used, where the main one corresponds to the lack of temporal analysis. In this case, we only selected one day worth of data, which is not enough to reflect price coordination between petrol stations due to collision must be consistent over time. Moreover, according to @wheelerMulticollinearityCorrelationLocal2005 multicollinearity in GWR is higher than in a global regression model, which means that the interpretation of the variable coefficient could lead to misleading conclusions. Lastly, we used a distance of 1 mile to define competitors; however, that distances could vary through space, being less in dense areas and higher in remote places.

Nevertheless, this methodology could be improved using a spatio-temporal analysis, which can capture consistent price coordination in particular areas. Moreover, features could be improved, so an index of competition could be created for each petrol stations, considering its location and particularities. Finally, multicollinearity diagnostics tools must be used to detect and understand its effect on the GWR parameters.