Final Project

Alberto Rodriguez

1/12/2019

Introduction

Mexico went through a significant change in internal governance structure with the constitutional reform on the liberalization of gas prices in 2016. Between February and November of 2017 this liberalization process was implemented across the country, making the price of gasoline to go from a single price nationwide that was determined by the government, to a fully variable pricing system based on both the national and international markets.

For a nation that has deep ties to the national value of oil, this process was met with severe criticism hardly impacting the acting administration of Enrique Peña Nieto (2012-2018) and making it one of the main topics in the presidential campaign of 2018. In this election the liberalization of oil prices was touted as an extremely hurtful measure that would impact households across the country and ended up being on the slogans of the now President Andres Manuel López Obrador.

This brief report aims to create a model, using several Machine Learning Approaches that might predict how much do households spend in energy related expenditures as a way to provide better information on the effects of energy policies, focusing on the effect of gas prices that arose from the energy reform as well as other expenditures. To focus a more specific view, I will try to predict the transport costs of each household given that is one of the most sensible expenditures that is directly related to energy pricing. Such a model would try to provide policy makers a more specific measurements of the impact of their policies.

This model will be created using the National Survey of Household Income and Expenses by the National Institute of Statistics and Geography of Mexico which includes information on more than 74,000 observations of households in the country. This survey includes several variables at the household level, such as socioeconomic level, geographic data and expenditures directly associated with the price of gas such as public and private transportation.

Background

On March 18, 1938, the President of Mexico, Lázaro Cárdenas, issued the Petroleum Expropriation, which consisted of the legal appropriation of the oil operations of 17 foreign companies that controlled the industry. The main reason for the Expropriation was the constant refusal of the international oil companies to improve the wage and labor conditions of the employees of this industry. With this mandate, it was established that the Mexican State would have total control over the production and commercialization of oil in national territory. On June 7, 1938 - the Petróleos Mexicanos (Pemex) parastatal was founded, endowed with the necessary powers to carry out all oil exploration, exploitation, refining and commercialization work in our country (Garcia, 2009).

Since then, oil and fuel have been deeply intertwined with development policies in Mexico. In the 1970's oil production was touted by the government as the solution for the country's development given the price surge of the commodity worldwide, and after the second largest oil field in the world called Cantarell was discovered by PEMEX, the political link to oil prices became stronger with almost a third of the government's budget relying on oil revenues (Wood, 2018).

But by 2008, Cantarell's production had dropped and both petroleum and electricity national companies (PEMEX and CFE) became inefficient enterprises with high debt and dated technology. To try to solve this problem, in 2013, the Congress passed a series of constitutional amendments promoted by former president Enrique Peña Nieto under the "Pact for Mexico"

to open the industry to private and foreign investment for the first time in 75 years and aimed to overcome the legal obstacles that limited his progress, re-configuring the iron-clad constitutional scheme that regulated the use of hydrocarbons since the 1940's (Garcia Rivera, 2015).

However, the open nature of the liberalization process provided an opening for a rise in gas prices in early 2017 called by the press as the gasolinazo. This major crisis provided a platform for the opposition and branded it as one of the worst problems created by the administration with a focus on how it hurt the population. (Nikolewski, 2017). On the presidential campaign of 2018 the now President Lopez Obrador ran on anti-corruption platform and publicly stated that the reform had been a mistake that deeply hurt the people of Mexico. This approach will shed some light on the question of how variables such as the price gas can predict one of the most sensible expenditures of any household that is directly related to gas prices: Transport costs.

Literary Review

Even though the Mexican case is an interesting example to understand how fuel pricing subsidies and regulations affect societies, there are already several studies that have tried to model the effects of fuel prices on other different sectors. These studies range from analyzing the specific changes on the automotive industry, to helping understand the commodity nature of oil and fuel and its position as a good that considered as inelastic.

Starting with the revision of classic economic theory, fuel prices are normally regarded as an inelastic good, meaning that given that the quantity demanded by a purchaser won't change even when the price changes. This inelasticity comes from the fact that there are no substitute goods for it on most markets. However, fuel prices also act as a commodity which means that the good is so ubiquitous to the economy that the price varies as a whole no matter who the producer is. This also means that a change in its price sends ripple effects

throughout the economy. In the case of Mexico, research has found that the change in fuel prices does not affect macro-economic factors such as inflation. Cervantes, López y Montiel used gas prices from 2002 to 2009 to indicate that there is not a significant effect on the level of consumer price index inflation, even though the price of fuel is directly related with several goods an services that are used in the inflation calculation made by the Mexican Central Bank. (Cervantes Jimenez, Lopez Sarabia, & Montiel Alejo, 2011).

There has been also research around how fuel prices affect expenditures on goods that are not as directly related as the fuel auto economy. An example of this is research done by Martin Baughman and Paul Joskow, who analyze how fuel prices affect residential appliance choice in the US. Doing elasticity analysis on the 48 states they found out that relatively small changes in prices can affect consumer appliances buying decisions in the short run (Baughman & Joskow, 1975).

These models focus on measuring the elasticities of various manufacturing sectors according to the price of fuel. The proposed models use a similar approach to measure impact of the change in gas prices and focuses on the household expenditures of the country rather than productive sectors, and providing information on transport costs for the household, providing a direct public policy aspect that could easily be translated into recommendations to key stakeholders and decision-makers.

Data & Wrangling

The datapoints from the National Survey of Household Income and Expenses by the National Institute of Statistics and Geography provide observations on more than 60 expenditure and income sources. These are then cross referenced with specific information to provide as much granularity as possible with information that will be used to control the variation of the theoretical model. To do so, we manually picked specific variables that are associated with the consumption patterns and the general summ of earnings in order to avoid any collinearity

problems in the variables. We also removed form the data any observation other than general household such as multi-familiar households, to also avoid any collinearity problems.

Using the geographic data of each household, I add the national price of fuel on the month that the interview was made to the survey of 2018 to properly control for fuel prices and inflation, with information provided by the newly created Energy Regulation Commission. We then cross-reference the monthly inflation report of the month in which each survey observation of the month was created with information provided by the National Bank of Mexico to provide a nice cross-sectional dataset.

Analysis

After the proper wrangling, I set to establish the basic procedures to ensure that this project is both cross-validated and directly related to the task at hand.

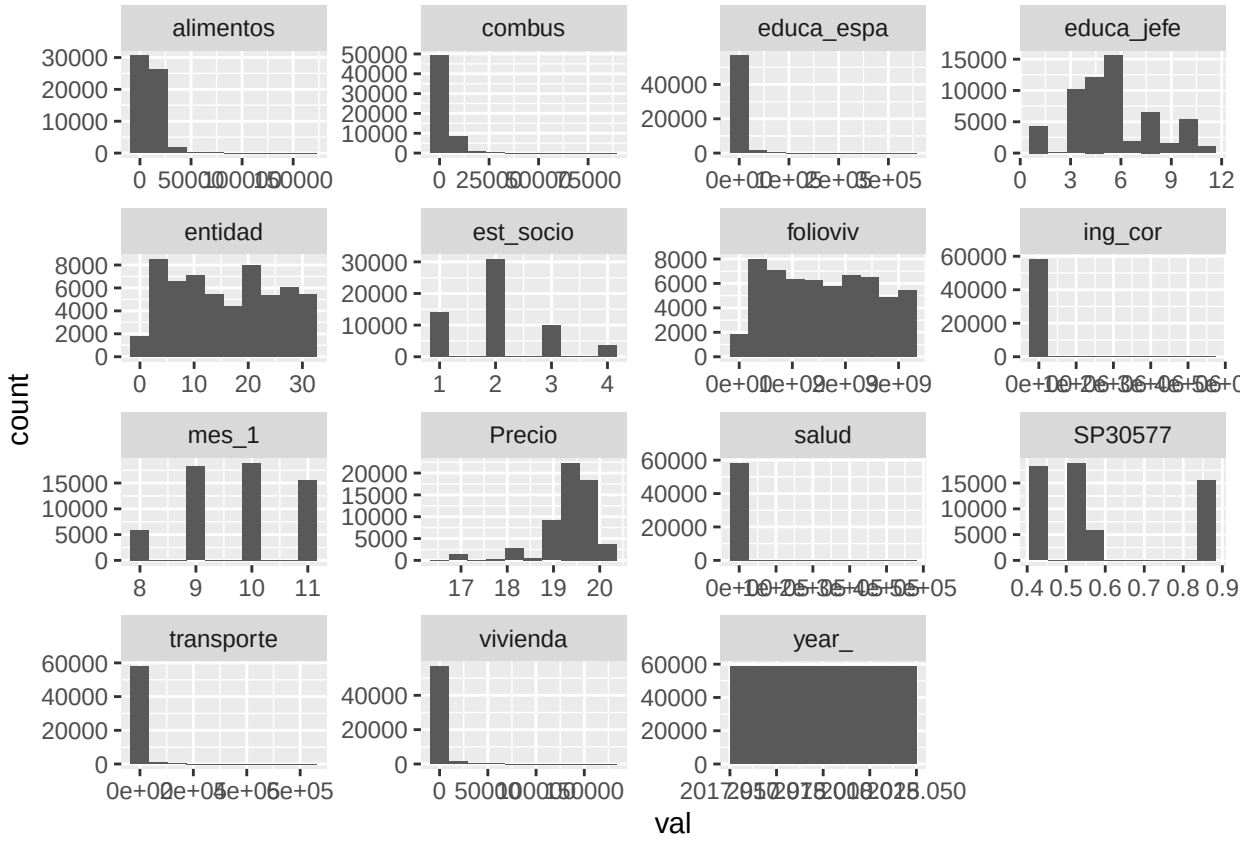
Spliting & Examining the Data

We split the the dataset into 2 groups: a training dataset with 80% of the observations and a test dataset with the other 20%. This is to ensure not to bias the development of my learning model. All of the models are run on the training dataset and will be then used in the testing model.

These summary statistics lets us know that there are no missing values (NAs) after our wrangling and our division, as well as seeing that there are some variables that are character values.

Data Pre-processing

To further understand the variables I'm using in the model, a proper visualization might provide more information:



In this case, we can see several key insights: -Education and socio economic status are coded as categorical variables and have to be treated though dummies to correctly asses the impact on every category they have. -Price of fuel doesn't have as much variation as we expected. -Year was a variable that helped us wrangle the data, but will not be used in the model, neither is the observation number or the state of the observation.

We control for these insights and include them using a common treatment method using the recipes package. This treatment will normalize the observations, as well as create the dummy variables for the two categorical sources. This treatment of the data also provides that the training data and the testing data are handled in the same way, but without observing the changes in order to not bias them.

Cross Validation

Once the Data treatment is done, I aim to focus on providing cross validation through a k-fold cross-validation method that evaluates the model performance on different subsets of the training data. This method randomly split Sthe data set into subsets and reserves one subset and train the model on all other subsets. In this case, we are using 6 folds to provide a robust measurment of the models.

Once this treament is done, we will try three general methods to predict the effects on the cost of transport using the following methods: - Linear Regression - K- Nearest Neighbors - Random Forest

Linear Regression

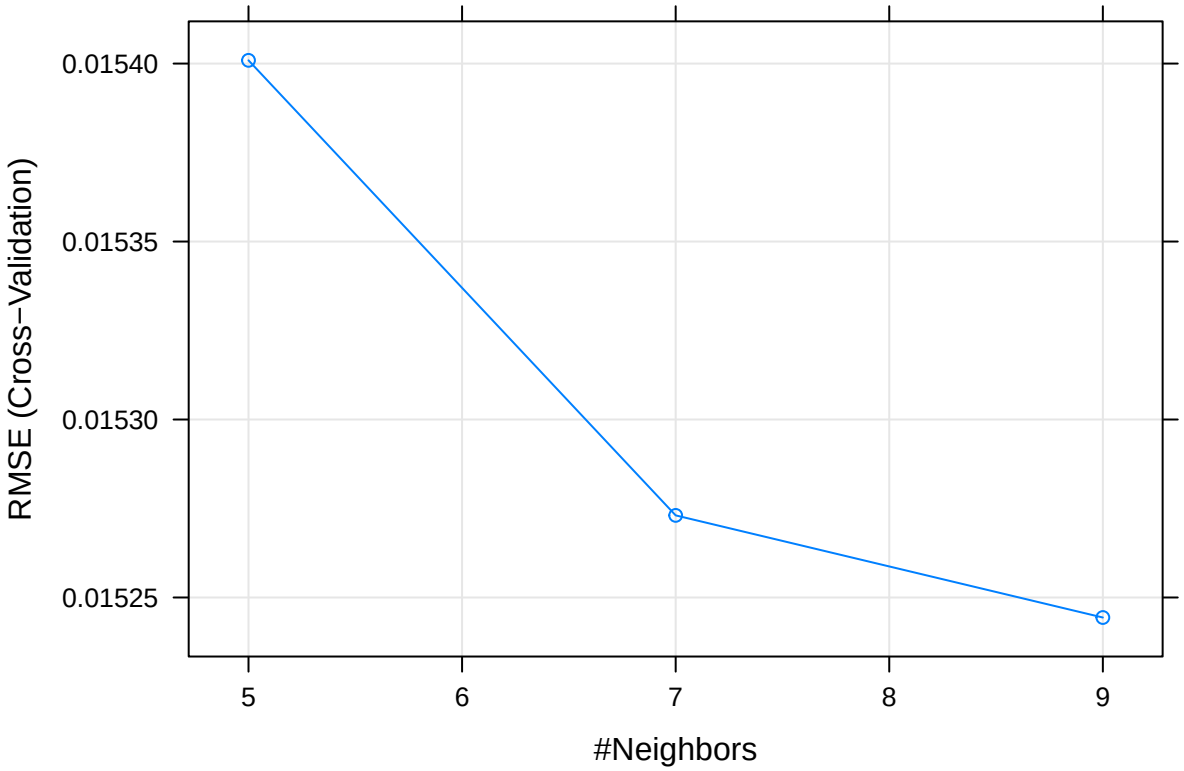
Linear regression searches for a linear relationship between the input variables (all of the covariates such as socioeconomic status, price of gas, inflation etc.) and the single output variable (ain this case the expenditures on transport). This model gives us the following results:

58725 samples with 21 predictors andNo pre-processing RMSE 0.01351838 Rsquared 0.3824566

We can observe a somewhat low R-squared (which represents the proportion of the variance of the transport expenditures that's explained by the other covariates) and a good RSME (That aggregates the magnitudes of the errors in the predictions to measure accuracy).

K-Nearest Neighbors

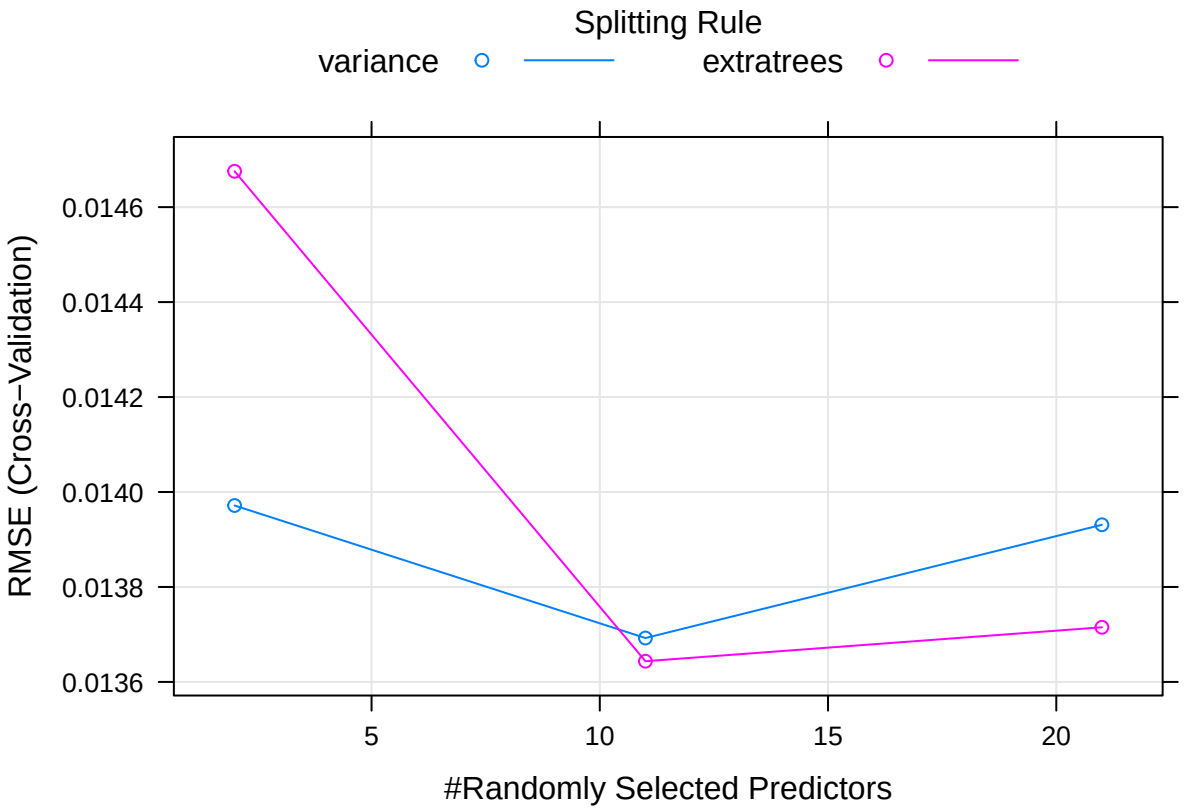
The K-Nearest Neighbors algorithm works on the assumption that similar things exist in close proximity. It calculates the numerical distance on the observations and tries to model a prediction using that information.



We can see in this method a lower R-squared and a poor RSME, even accounting for the ammount of neighbors tested in the data. There weren´t any signifficant changes in the RMSE. This could be because the K-Nearest method is looking for observations close, and given the nature of the data sets and the ammount of dummy variables hinders its predictive power.

Random Forest

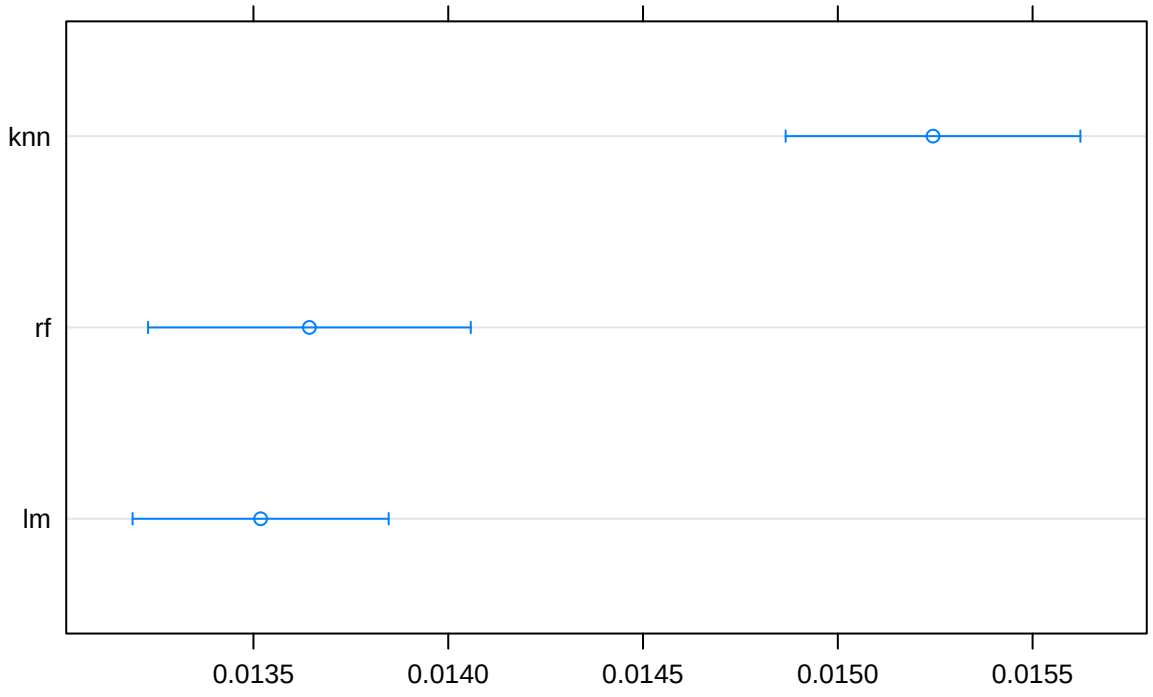
Random forest, uses a large number of individual decision trees to predict the best way of understanding the observations. When it runs a large number of relatively uncorrelated models (trees), it seeks to have a correlation between models and offer an accurate prediction, "protecting each other from their individual errors".

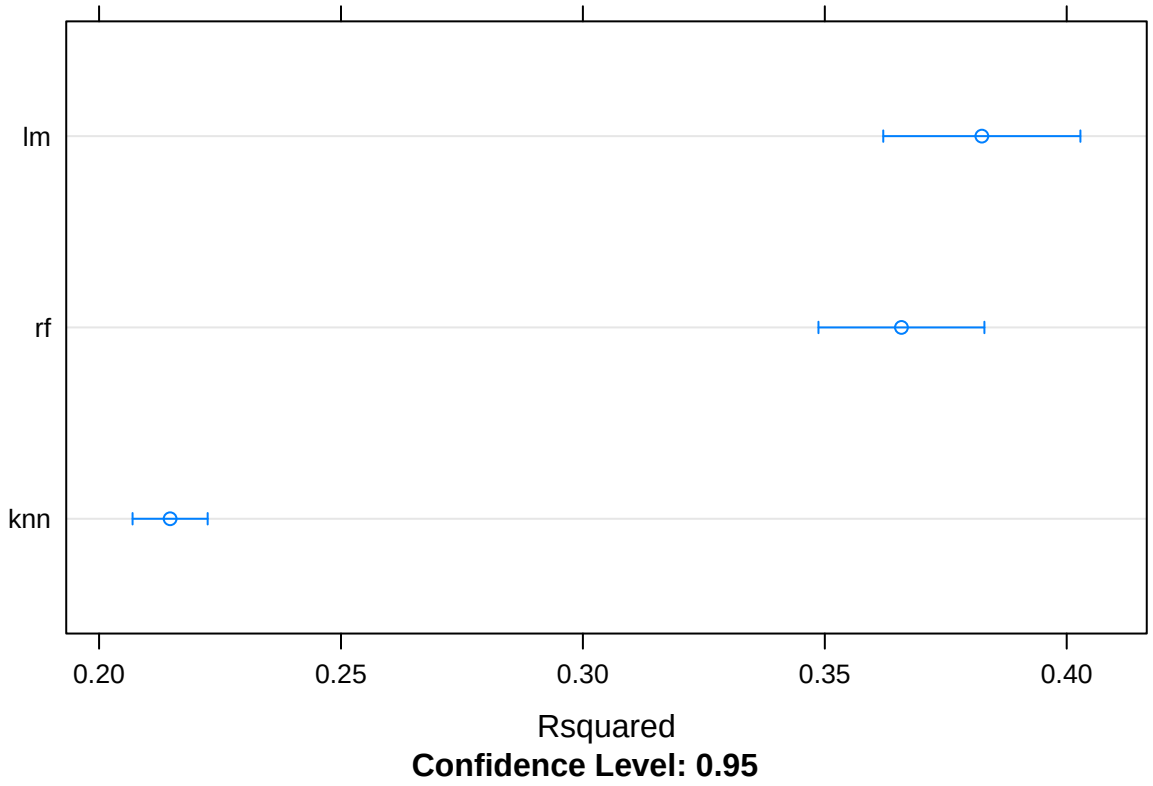


We can see in this method a lower R-squared and a poor RSME as well, even accounting fo the variance of the trees autOmtically selected by the model. This could be because the method is looking for variation with a data sets that has a high ammount of dummy variables, also hindering its prediction possibilities.

Model Comparison

By comparing the previous models, we can clearly see that the model that represents the higher R-squared and the lowest RSME (and therefore the best predictive power) is the linear regression model. This might be because of the nature od the dummies and numerical values used in the model.


knn
rf
lm
0.0135
0.0140
0.0145
0.0150
0.0155
RMSE
Confidence Level: 0.95



Final Results

Predictive Accuracy of the Best Model

```
pred <- predict(mod_lm,newdata = test_data2)

mse = sum(test_data2$transporte-pred^2)/nrow(test_data2)

mse
```

```
## [1] 0.009305205
```

By using the Linera Regression Model on the testing Data we see that the Accuracy of the model declines, seeing that maybe the regression could be boosted showing an RMSE 0.009305205

The model is not good to predict transport cost base on other expenditure patterns, and

might need to be revisited with other data that provide different approaches.

Discussion

All of the three models were not able to provide a good prediction of the transport costs, but do explain that gas prices and socioeconomic status have a direct effect on how Mexicans spend their money to move. On further analyisis a revision of other data sources that provide information on car-wonership or size of the city they live in would boost both the R-Squared and RSME values to better levels.

It is important to note that the results of this analysis are important for current and future energy policy analysis, given that energy security in Mexico is expected to be a worrisome problem (Garcia Perez, 2018), having a strong effect of price changes in trasnport costs could provide a higher incentive to Mexican officials to return to controlled fuel pricing regimes.

Bibliography

Altomonte, H., & Rogat, J. (2004). Politicas de Precios de Combustibles en América del Sur y México, implicaciones económicas y ambientales. Santiago de Chile: División de Recursos Naturales e Infraestructura, CEPAL.

Alvarez, J., & Valencia, F. (2015, February). Made in Mexico: Energy Reform and Manufacturing Growth. (I. M. Fund, Ed.) IMF Working Paper. Retrieved from https://www-elibrary-imf-org.proxy.library.georgetown.edu/doc/IMF001/22386-9781498376426/22386-9781498376426/Other_formats/Source_PDF/22386-9781498333214.pdf

Baughman, M., & Joskow, P. (1975). The Effects of Fuel Prices on Residential Appliance Choice in the United States. Land Economics„ 51(1), 41-49. Retrieved from https://www.jstor.org/stable/3145139

Belisario Dominguez Institute. (2017). Aspectos relevantes sobre la liberalización de los

precios de las gasolinas y el diésel en 2017. Mexico City: Senate of the Republic, Mexico. Retrieved from http://bibliodigitalibd.senado.gob.mx/handle/123456789/1878

Burk, P. J., & Nishitateno, S. (2013). Gasoline prices, gasoline consumption, and new-vehicle fuel economy:. Energy Economics, 36, 363 - 370.

Cervantes Jimenez, M., Lopez Sarabia, P., & Montiel Alejo, J. (2011). Impacto Económico el aumento en el precio de las gasolina en México: un análisis de conitegración y vectores autoregresivos. Estocástica, finanzas y riesgo, 1(2), 63-93.

Espey, M. (1996). Watching the fuel gauge: An international model of automobile fuel economy. Energy Economics, 18, 93-106.

G20 Peer Review Team. (2017). Mexico's efforts to phase out and rationalise its fossil-fuel subsidies. Hamburg: OECD.

Garcia Perez, J. (2018). La seguridad energética en los mercados de gasolina y diesel en México. Madrid: Instituto Español de Estudios Estratégicos.

Garcia Rivera, E. A. (2015). The renewed constitutional Regulation of National Hidrocarbons. Cuestiones Constitucionales - Revista Mexicana de Derecho Constitucional, 53-79.

Garcia, J. C. (2009). La disputa por el petróleo en México. Breve historia de los hidrocarburos. En defensa del petróleo. Ciudad de Mexico: Instituto de Investigaciones Jurídicas de la Universidad Nacional Autónoma de México. Retrieved from https://archivos.juridicas.unam.mx/www/bjv/libros/6/2729/5.pdf

Gobierno de la República México. (2013). Reforma Energética: Resúmen Ejecutivo. Ciudad de México: Presidencia de la República. Retrieved from https://www.gob.mx/cms/uploads/attachment/file/164370/Resumen_de_la_explicacion_de_la_Reforma_Energetica11_1_.pdf

Gramlich, J. (2009). Washington D.C.: Georgetown University.

Gutierrez Rodriguez, R. (2017). The simple arithmetics of the new gasolines and diesel pricing policy. Economía Informa, 404, 40-62.

Jimenez Sanches, J. E. (2018). Impacto del Precio del Diesel y otros insumos en los costos de operación del autotransporte de carga. Sanfandila, Queretaro: Instituto Mexicano del Transporte.

Klier, T., & Linn, J. (2013). Fuel Prices and new vehicle fuel economy - Comparing the United States and Western Europe. Journal of Environmental Economics and Management, 66, 280-200.

Nikolewski, R. (2017, January 31). Boiling over on the border: The reasons behind the gasoline protests in Mexico. Los Angeles Times. Retrieved from https://www.latimes.com/business/la-fi-mexico-gasoline-20170131-story.html

Wood, D. (2018). Mexico's New Energy Reform. Washington D.C.: Mexico Institute Woodrow Wilson International Center for Scholars.