Forecasting & Analytics: Gas Prices in Brazil

Data Science & Analytics, University of Oklahoma
ISE/DSA 5133: Energy Analytics
Dr. Talayeh Razzaghi

Anna Christensen & Sonaxy Mohanty

December 13th, 2022

Contents

1 Introduction **3**

2 Project Description **4**
2.1 Approach . . . 4
2.2 Dataset . . . 4

3 Project Tasks **4**
3.1 Data Pre-Processing . . . 4
3.2 Hold-out Validation . . . 5
3.3 Brazillian Gas Prices Trend and Seasonality . . . 5
3.4 Decompose Time Series . . . 9
3.5 Transformation . . . 11
3.6 Fitting Models . . . 11
3.6.1 Benchmark & Exponential Smoothing Models . . . 11
3.6.2 ARIMA Model . . . 11
3.6.2.1 Checking Seasonal & Order Differences . . . 11
3.6.2.2 Differenced Data Plots . . . 12
3.6.2.3 PACF & ACF plots for Centro Oeste . . . 13
3.6.2.4 Notations . . . 14
3.7 Evaluating Models . . . 14
3.7.1 Forecast Plot . . . 14
3.7.2 Accuracy Measures . . . 15
3.7.3 Residual Diagnostics from the best method . . . 15
3.7.3.1 Histograms . . . 15
3.7.3.2 Time series plot . . . 16
3.7.3.3 ACF plots of residuals . . . 17
3.7.3.4 Ljung-Box test . . . 17
3.8 Additional Analysis . . . 17
3.8.1 Checking for & Removing Outliers . . . 17
3.8.2 Seasonally-Adjusted Data versus Trend-Cycle component and Actual Data . . . 19
3.8.3 Neural Network with Best Model . . . 19
3.8.4 Accuracy Measures on an average . . . 20

4 Conclusion **20**

5 References **21**
5.1 Dataset & R References 21
5.2 Works Cited 21

1 Introduction

As large-scale collections of energy data are made available to the public, forecasting this data becomes a new challenge and concern for companies and entities that are affected by the uncertain nature of this data. This data can be particularly difficult to forecast given its complexity due to the numerous independent and confounding factors that affect the response values. This project focuses on energy price data obtained from the National Agency of Petroleum, Natural Gas, and Biofuels in Brazil to develop improved forecasting models that can predict price data in the future.

We hypothesize that an ARIMA or ETS model will provide an improvement in forecasting the data when compared to simple benchmark methods. We hope that developing these models would allow for price data to be predicted with more accuracy, which would allow entities in the energy industry to better plan for future price and distribution strategies.

Brazil is a unique country in its approach to satisfying the demand for vehicle fuels. In response to uncertainty and embargoes dictated by the Organization of the Petroleum Exporting Countries (OPEC), Brazil has been pushing for an innovative biofuels program since the 1970s (Stecker, 2013). As a result of their investments and the introduction of Flex Fuel vehicles, Brazil has a booming production of sugarcane ethanol, and about 95 percent of cars in Brazil are flex-fuel vehicles. Brazil has also been experiencing a crude oil boom for the last decade, with discovery and production occurring rapidly. As a result, Brazil has an energy mix that includes cheap and abundant ethanol and gasoline.

One significant event in energy markets that occurred during the period contained in the dataset is the global oil price drop of 2014. From June to December 2014, oil prices dropped about 40% for the Brent benchmark, which is an international index of crude oil prices (Samuelson, 2014). The main cause of this steep price decrease was simple supply and demand: the world was producing a large supply of oil for too little demand. Since consumers' needs for petroleum products are rigid in the short term, a surplus (or shortage) of oil can cause significant and sudden price swings. In this particular instance, the price drop lasted until about 2016, when the global supply and demand balance shifted again. It is expected that these major changes in oil prices will affect the gasoline price data that we are considering. The price of oil is one of the major factors in the prices of petroleum-derived products like gasoline and diesel, so these prices should also decrease significantly during this time.

ARIMA models and their derivations have become very common as a simple yet effective way to forecast energy demand and price data. Research by Zhang and Zhao (2022) proposed an application of this method to forecasting gasoline data by performing decomposition on the dataset and fitting a relatively simple model to each component. When aggregated, these simple models produce powerful and accurate forecasting results. For the application discussed in the paper, X11 decomposition was performed, and a linear regression model was used for the trend component since this component is relatively stable with little fluctuation. A SARIMA model was fitted to the seasonal component of the data and can also be fitted to a periodic component. The authors suggest using a neural network model for the uncertainty component if needed since it is generally more complex to forecast. This method is relatively easy to understand and apply, but the results from the application in the paper are impressive, as the MAPE for all models was less than 1%.

For more extensive studies, machine learning models have been found to produce extremely accurate models for gasoline time series data, which is generally uncertain and unstable. One paper that attempted to create new methods of forecasting gasoline consumption found that a random forest machine learning algorithm, used for both classification and regression, produced a stable and robust model that could most accurately forecast gasoline consumption (Ceylan et al., 2022). The random forest method generates large numbers of decision trees from random subsets of the training set of data. This model had a MAPE value of 11.529%, which is a very good fit for this type of data. Machine learning models such as the random forest model could be considered in the future analysis of this data to produce more advanced and accurate models.

2 Project Description

2.1 Approach

To build the model, several tasks are undertaken, which are summarized in the four-phase process displayed in Figure 1.

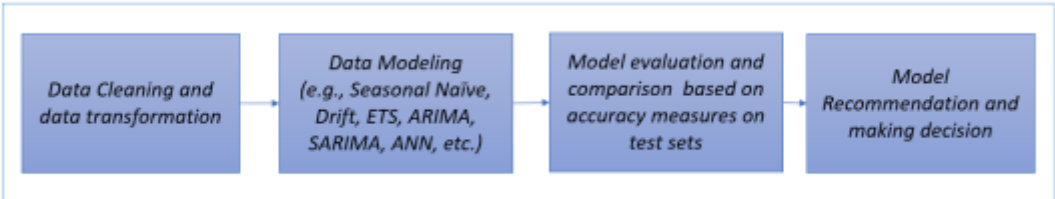


Figure 1: Our approach

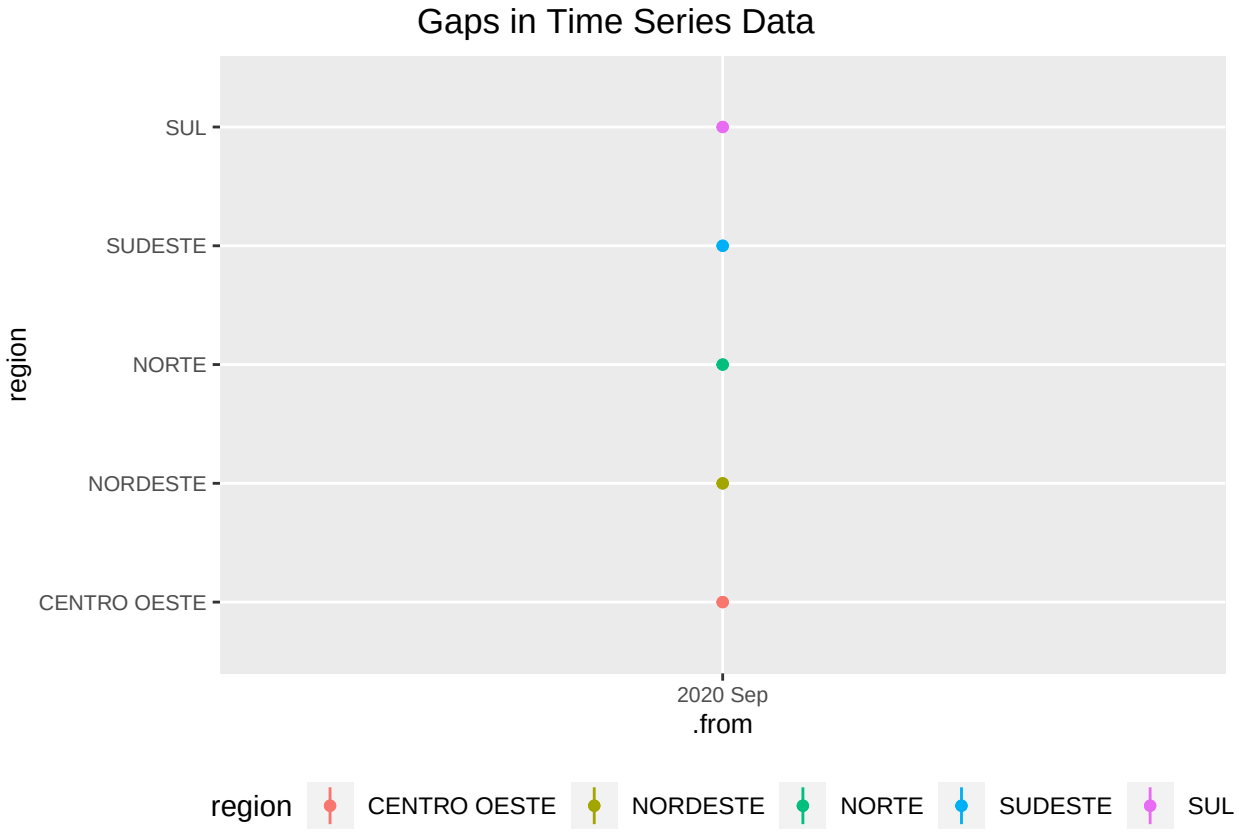
2.2 Dataset

The dataset contains metrics from weekly reports of gasoline, diesel, and other fuels used in transportation. This data focuses on the pricing of fuels in Brazil and is obtained from a Brazilian government agency, The National Agency of Petroleum, Natural Gas, and Biofuels (ANP in Portuguese). The observations in the dataset range from 2004 to May 2021. The data set includes 120,823 observations. The variables contained in the data are the mean resale price (price per liter, or per 13 kilograms, or per cubic meter), the minimum resale price (price per liter, or per 13 kilograms, or per cubic meter), the number of gas stations analyzed, and the standard deviation. The data points are grouped by product, region, and state. The link to the dataset can be found below:
https://www.kaggle.com/matheusfreitag/gas-prices-in-brazil

3 Project Tasks

3.1 Data Pre-Processing

Since the .tsv data file is in Portuguese, we have to change the column names to English for ease of understanding each attribute. The numeric attributes related to the gas price have many different units, which needed to be normalized. Then a tsibble object is created with month as the index and region as the key, since we focused mainly on analyzing and forecasting the gas price of a region in Brazil. There are some gaps in the time series data, as seen in the below plot, that were taken care of.

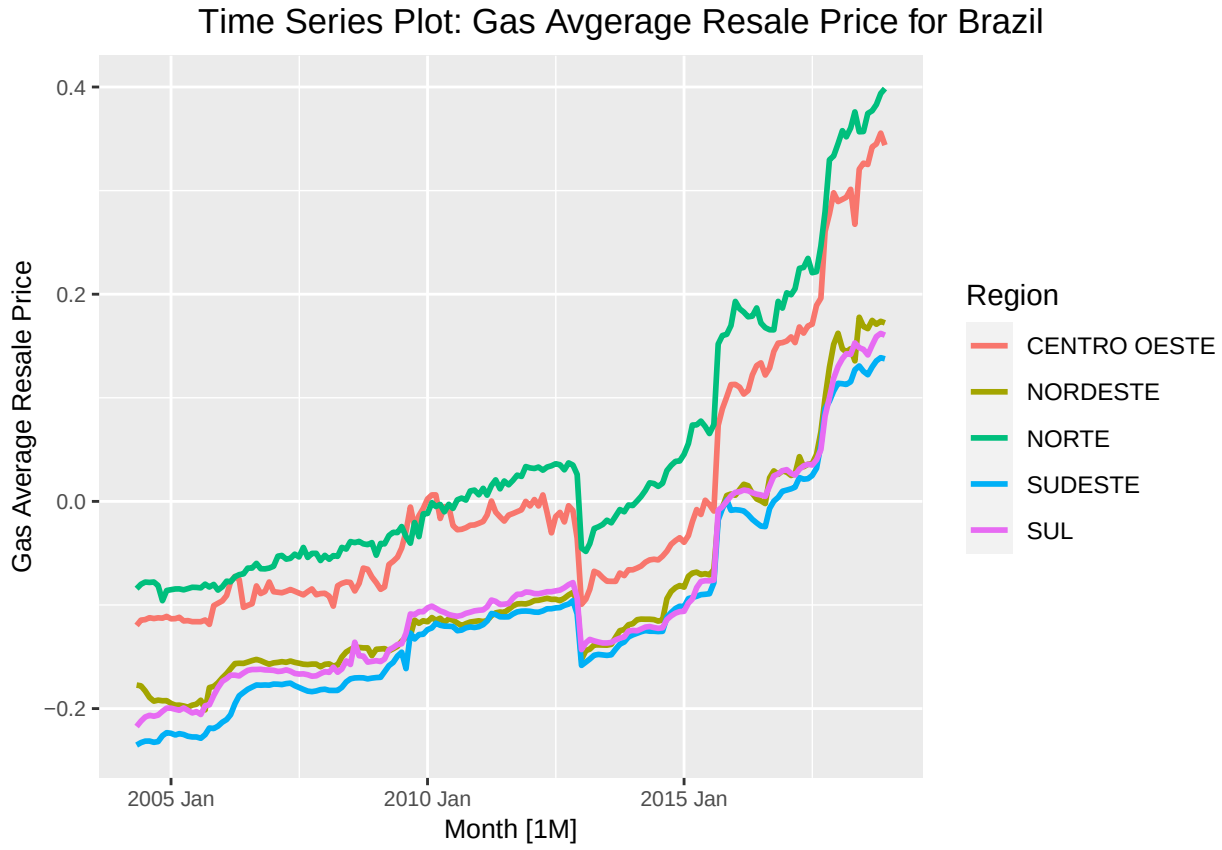


3.2 Hold-out Validation

The collected data were divided into a training dataset (2004 May to 2018 Dec) and a testing dataset (2019).

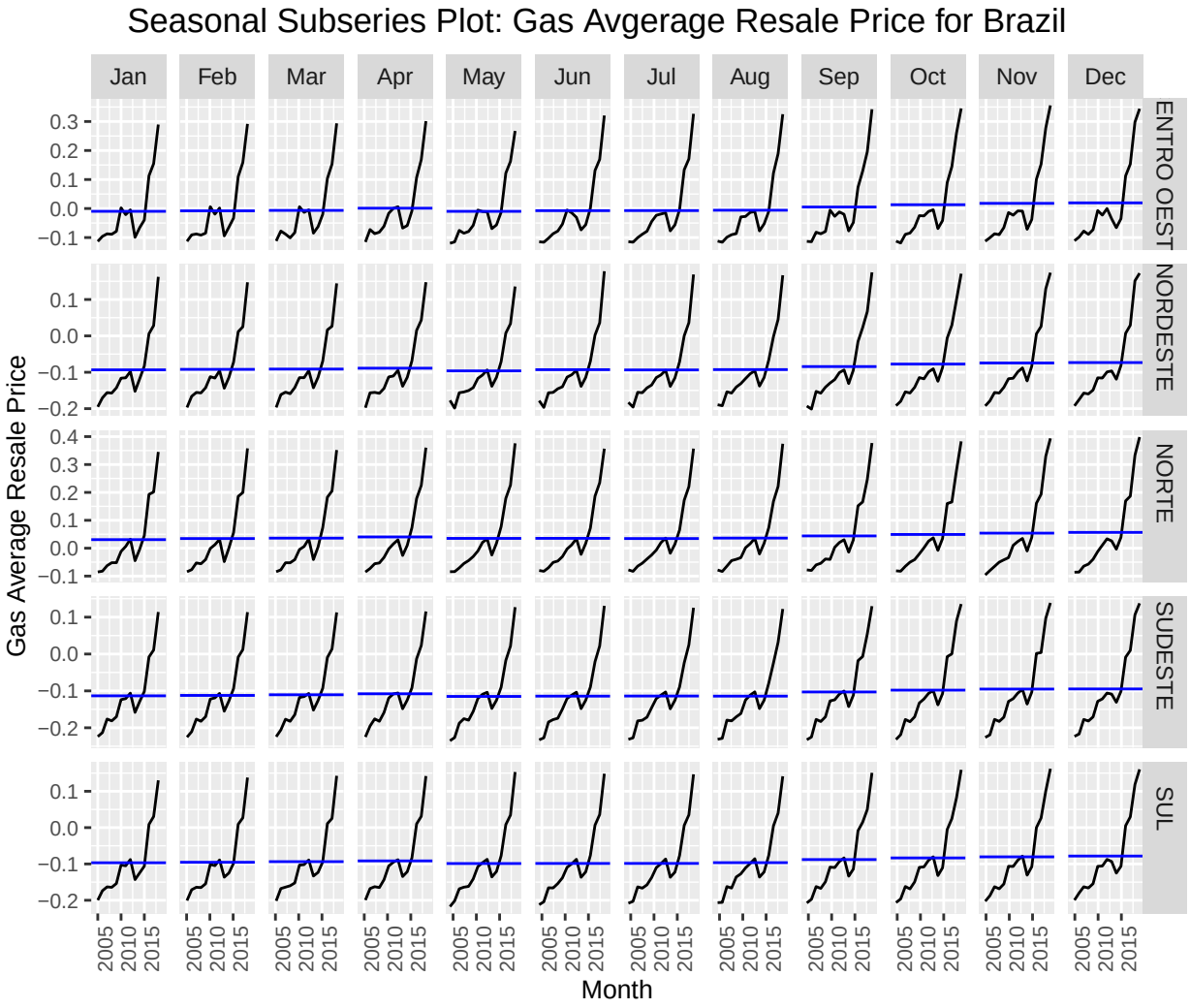
3.3 Brazillian Gas Prices Trend and Seasonality

The gas price has an upward trend. It also appears that there may be monthly seasonality present in the data, but more detail is needed to see this. No cyclical pattern is present. For all regions, the period from about 2013–2015 is an outlier. There was a sharp decline in the data around 2014, but the data has since risen steeply back to the historical trend around 2016.

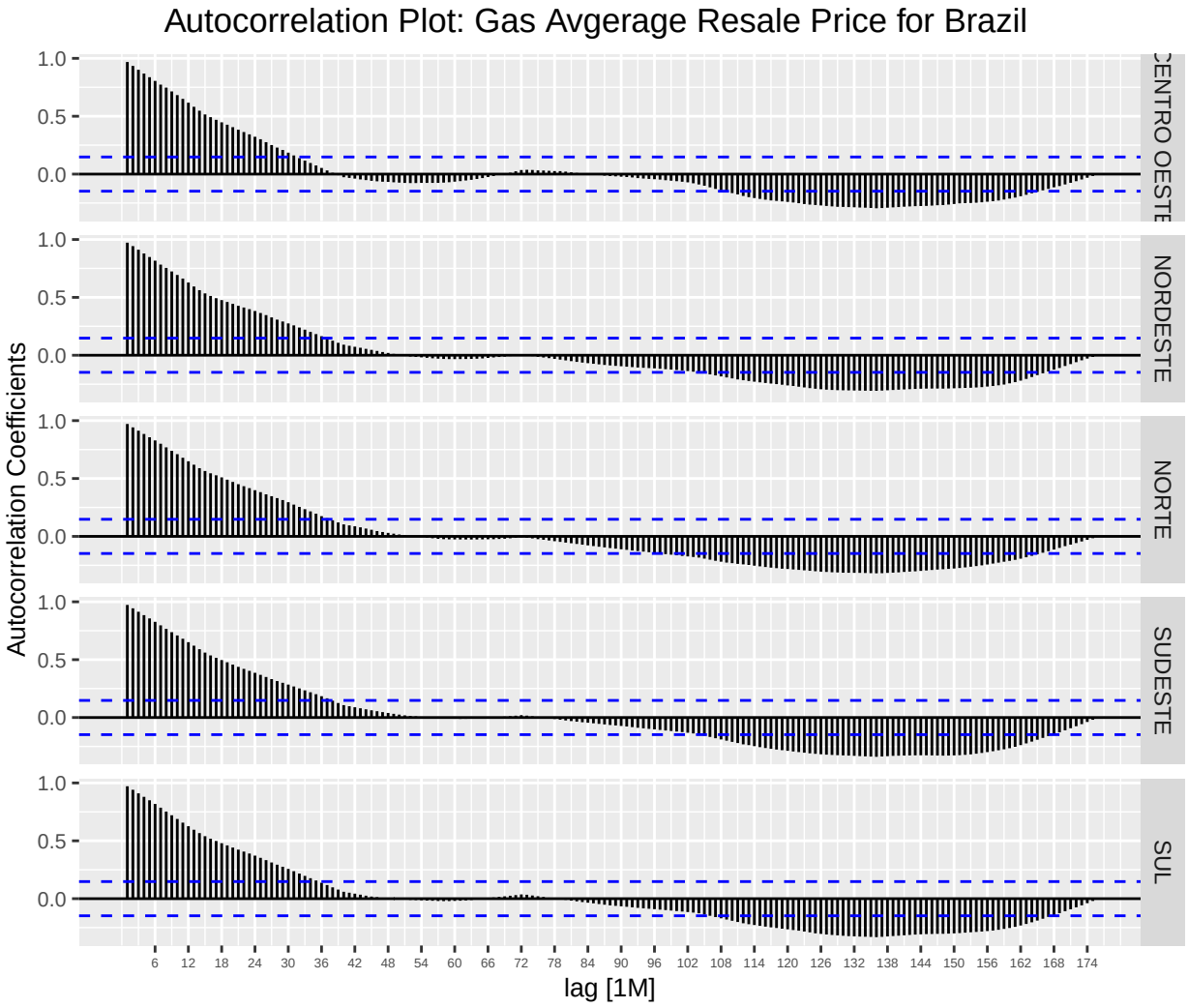


The seasonal plot shows the subtle yearly seasonality in the data. There is an increase in the data each year beginning around August. Each year follows a similar pattern and the upward trend causes increases in the data each year.

We can see the seasonal pattern with the increase at the end of the year and the upward trend in the subseries plot.

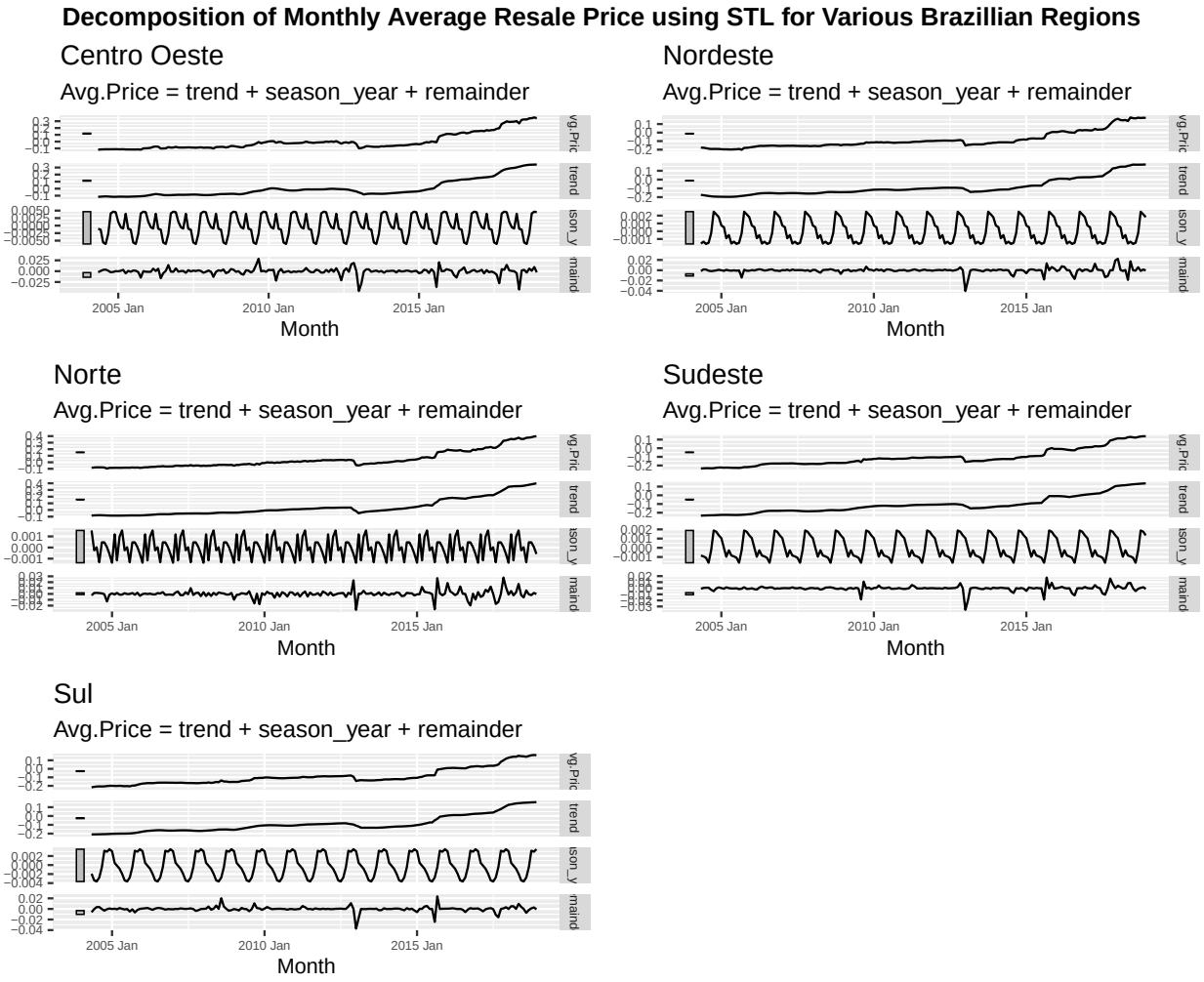


The ACF plot shows that significant autocorrelation is present in the data, so the time series does not resemble white noise. Strong positive autocorrelation is present in the first three years, and the ACF value decreases with each lag. Interestingly, there is significant negative autocorrelation with data further back as well, with the peak around a lag of 11-12 years back.



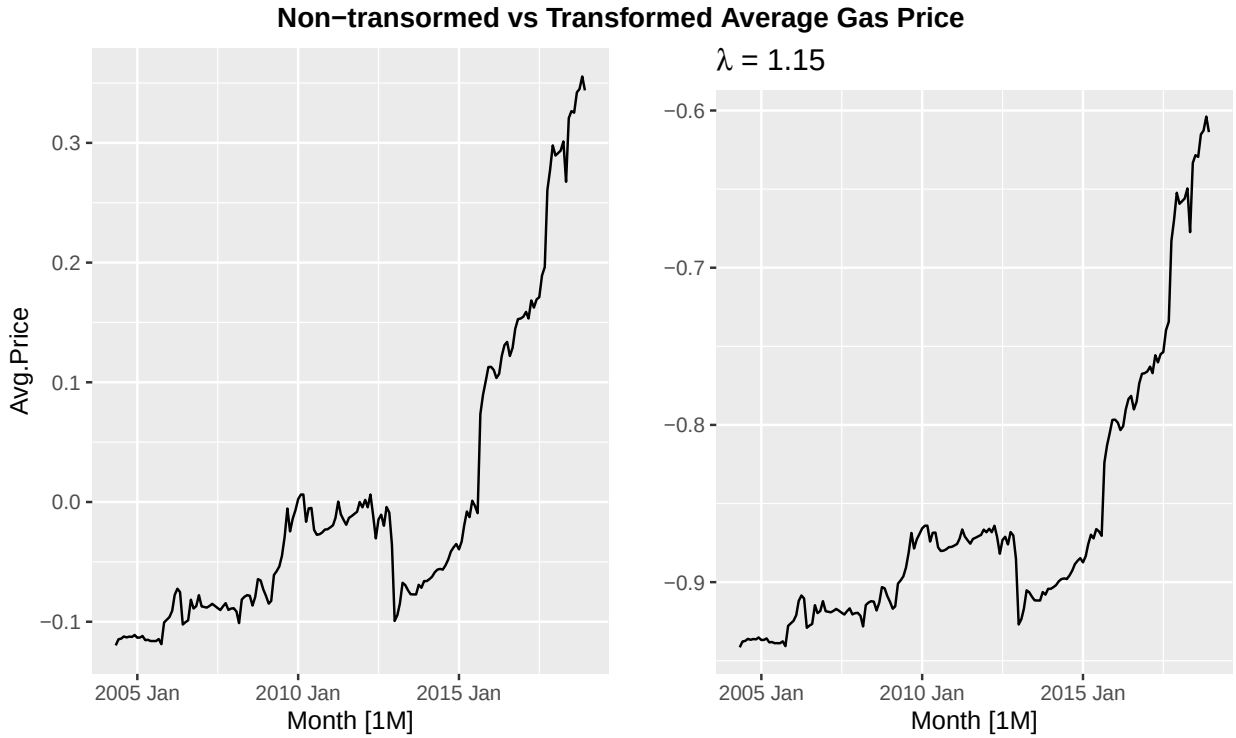
3.4 Decompose Time Series

STL decomposition is performed because STL is versatile and robust for decomposing time series as compared to SEATS and X11 cannot. The decomposition breaks the data down into several components by region. The data show an upward trend with a change in trend around 2014-2016. The monthly seasonality component is shown in more detail for each region. The scale shows that, while it has a smaller effect, there is a monthly seasonal pattern present. Several outliers are also present in the data. Some regions have a smaller error component with fewer outliers than others. For example, the series for the Norte region shows a far lower smooth error component than the others, suggesting a high level of unexplained uncertainty and variability. In contrast, the Nordeste, Sudeste, and Sul regions have a smoother error component with less unexplained variability. One outlier that is present in every region is the extremely low value around 2014, when there is a steep drop in the whole data series.



This plot shows the identified trend and seasonal component plotted over the actual data. We see that the trend explains most of the values seen in the actual data.

3.5 Transformation



A box-cox transformation is applied to the data to see if it improves the data variability. When comparing the transformed data with the actual data, we see that the transformation has very little effect on the shape of the data. A log transformation yielded similarly insignificant results. Therefore, we have concluded that a transformation is not necessary.

3.6 Fitting Models

3.6.1 Benchmark & Exponential Smoothing Models

Several benchmark methods have been fitted to the training set - Mean, Naive, Seasonal Naive, Drift, alongwith Exponential Smoothing (ETS) model.

3.6.2 ARIMA Model

Due to the limitations of the ARIMA model, we are choosing one region, *Centro Oeste*, to focus on fitting the model to. We will select optimal parameters for this region, fit the ARIMA model, and compare it to the benchmark models that were previously fitted.

3.6.2.1 Checking Seasonal & Order Differences Since the time series data for gas prices is not stationary, differencing needs to be performed to stabilize its mean.

Table 1: Seasonal Differences for Different Regions

Region	nsdiffs
CENTRO OESTE	0
NORDESTE	0
NORTE	0
SUDESTE	0

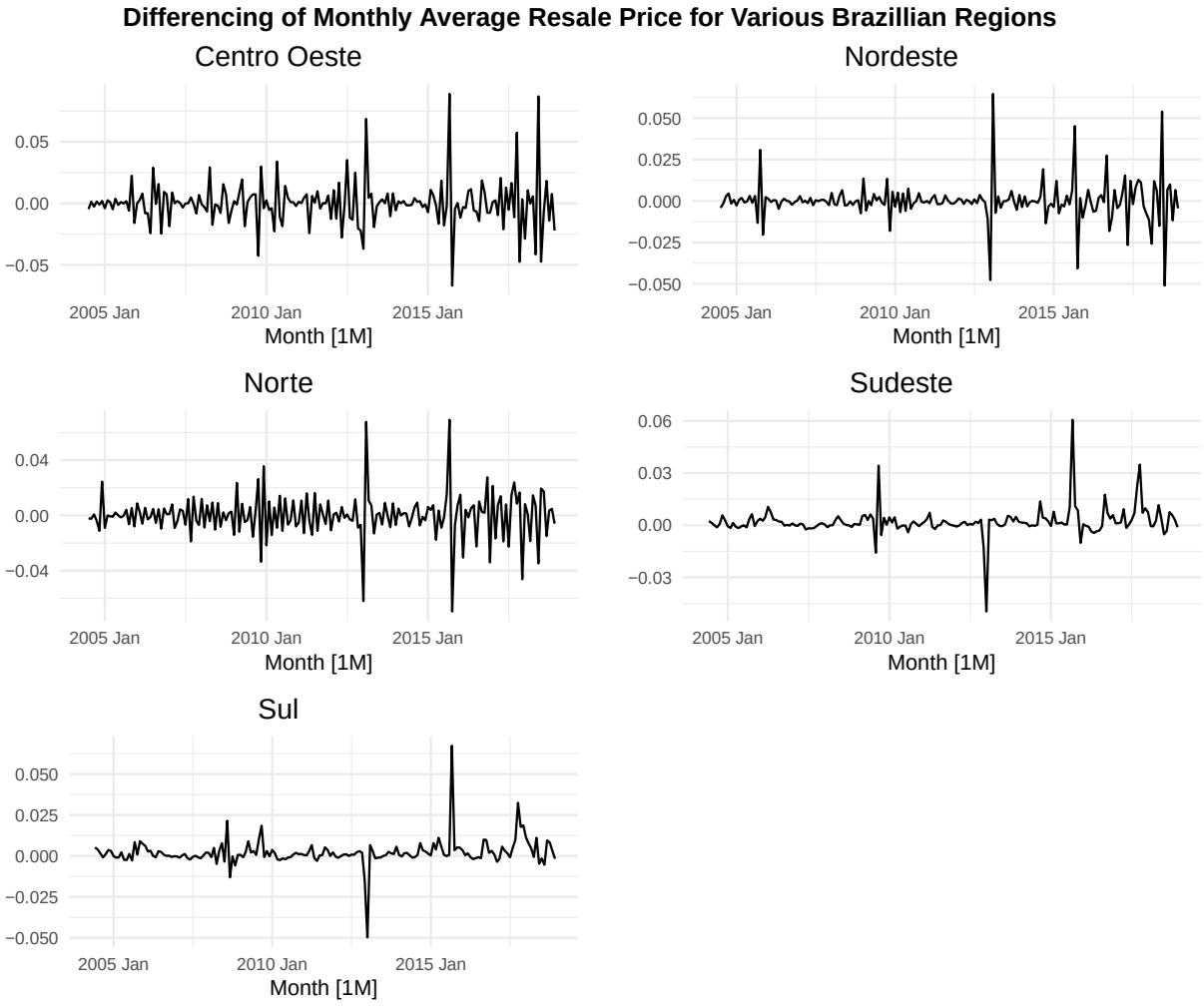
Region	nsdiffs
SUL	0

Table 2: Order Differences for Different Regions

Region	ndiffs
CENTRO OESTE	2
NORDESTE	2
NORTE	2
SUDESTE	1
SUL	1

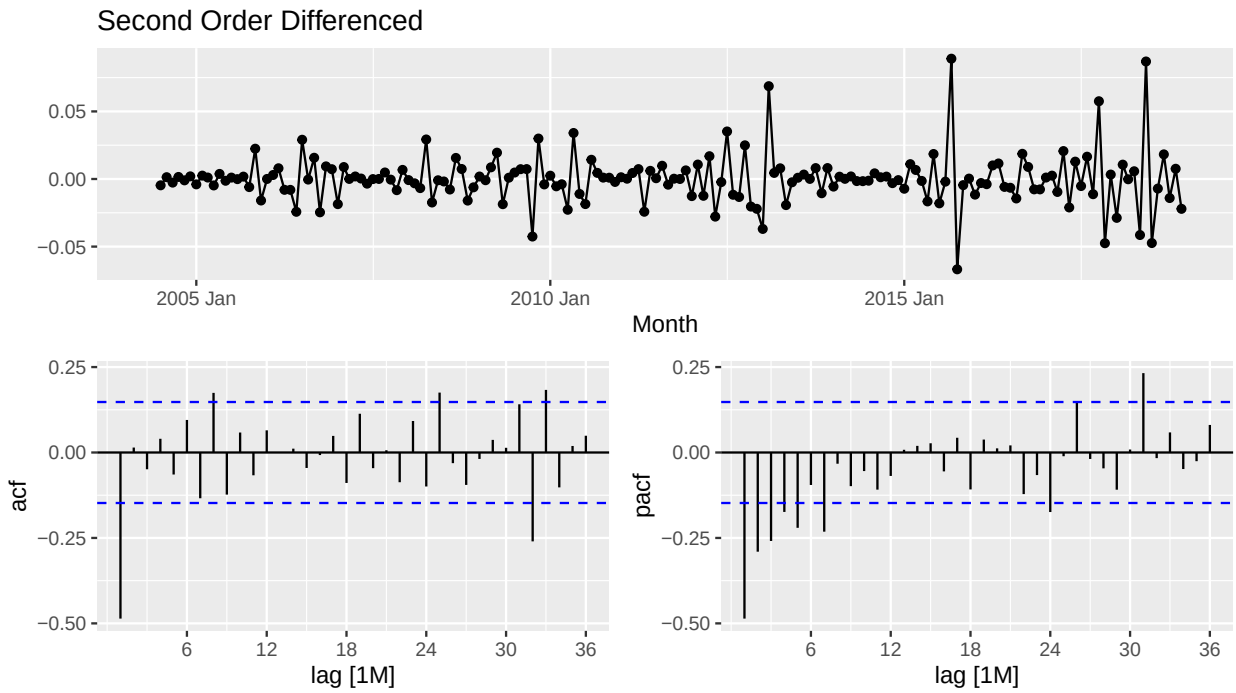
When using the KPSS test to conduct the Unit Root test on the time series data, we have obtained results for the number of differences and seasonal differences needed to obtain stationary data. The test results show that we don't need any seasonal differences, since the seasonal component is not strong enough to have a significant effect. The results also show that for the Centro Oeste, Nordeste, and Norte regions, 2 differences are needed, and for the Sudeste and Sul regions, only 1 difference is needed.

3.6.2.2 Differenced Data Plots Next, we will plot the differenced data by region to ensure that the results are stationary.



The differenced data sets now show a stationary series for each region.

3.6.2.3 PACF & ACF plots for Centro Oeste



The results of the ACF plot shows a significant autocorrelation at lag 1. The PACF plot shows a significant autocorrelation at lag 1 that exponentially decays for the next 4 periods. We will use these results to inform the ARIMA model that we will test.

Table 3: Different ARIMA Models Tested

.model	sigma2	log_lik	AIC	AICc	BIC
Arima	0.0001808	502.0721	-1000.1442	-1000.0740	-993.8260
Arima221	0.0001829	502.0908	-996.1816	-995.9449	-983.5454
Arima321	0.0001831	502.5136	-995.0271	-994.6700	-979.2318

The auto-selected $ARIMA(0, 2, 1)$ model performs better than the ARIMA models with parameters that we selected to test, according to the three measures that we have chosen to evaluate the models: AIC, AICc, and BIC. The next best model that we created was a 3, 2, 1 ARIMA model. We will move forward with the auto ARIMA model and compare this model to the benchmarks we have fitted.

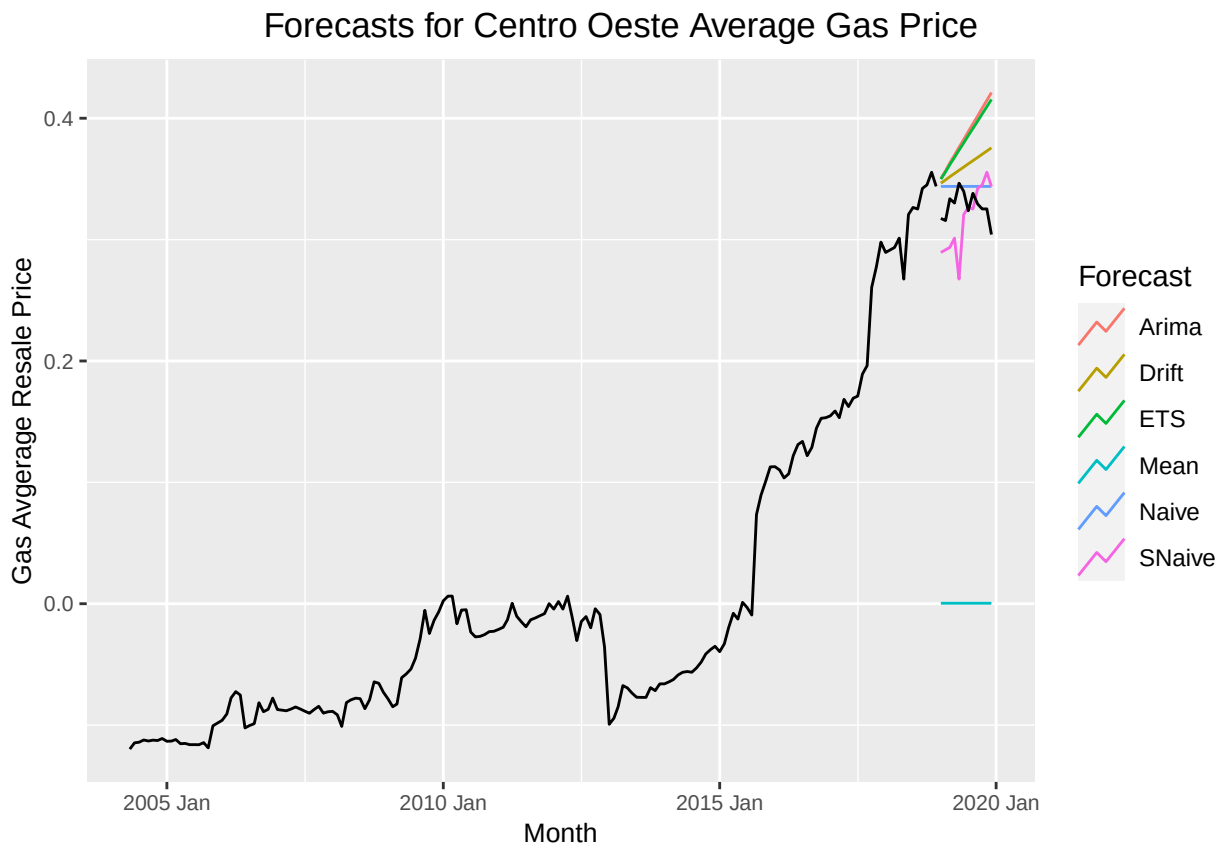
3.6.2.4 Notations Second-order differencing `without backshift operation` for $ARIMA(0, 2, 1)model$: $y'_t = \theta_1 \epsilon_{t-1} + \epsilon_t$ where y'_t is second-order differenced series

Second-order differencing `with backshift operation` for $ARIMA(0, 2, 1)model$:
$(1 - B)^2 y_t = (1 + \theta_1 B)\epsilon_t$

3.7 Evaluating Models

3.7.1 Forecast Plot

From the plot, the Naive and Seasonal Naive appear to be the methods that perform the best.



3.7.2 Accuracy Measures

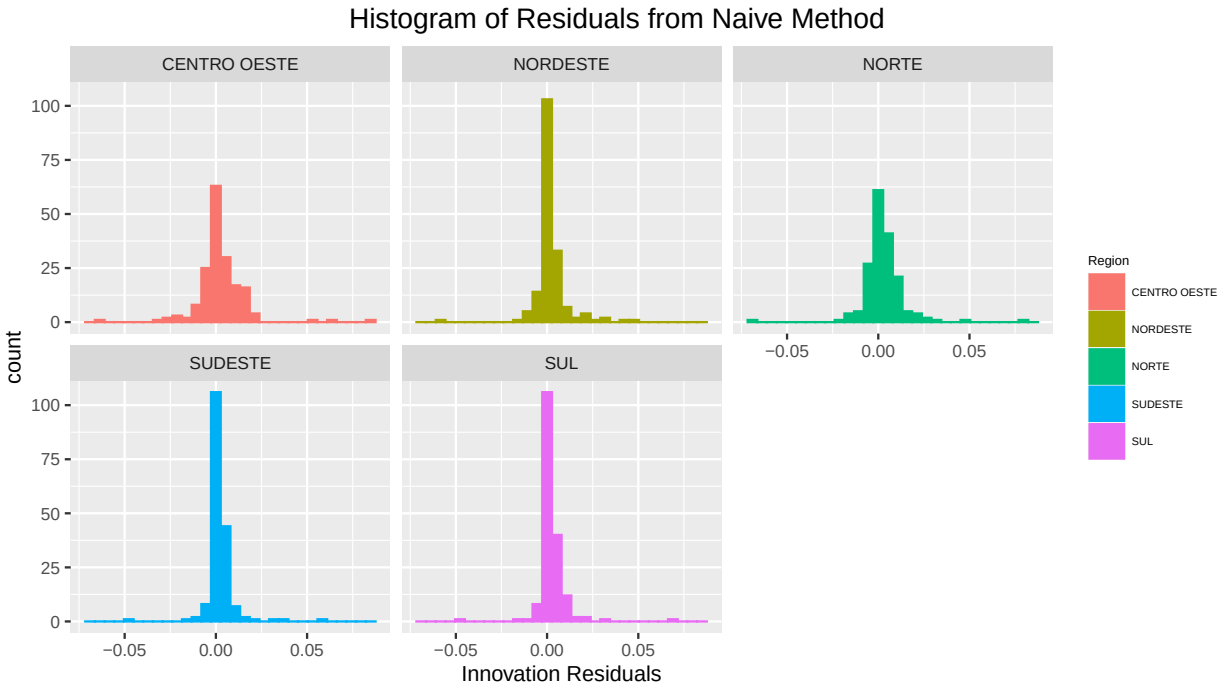
The accuracy measures confirm that the Naive and then Sesonal Naive methods did perform the best. The ETS model is also more accurate than the ARIMA model, though neither is the most accurate.

Table 4: Accuracy Measures of Models

.model	RMSE	MAE	MAPE	MASE
Naive	0.0198276	0.0168645	5.265532	0.3784485
SNaive	0.0337012	0.0281085	8.550631	0.6307715
Drift	0.0370823	0.0336489	10.422766	0.7551020
ETS	0.0606647	0.0552372	17.045827	1.2395552
Arima	0.0640640	0.0583007	17.985690	1.3083027
Mean	0.3272411	0.3270528	99.883829	7.3392560

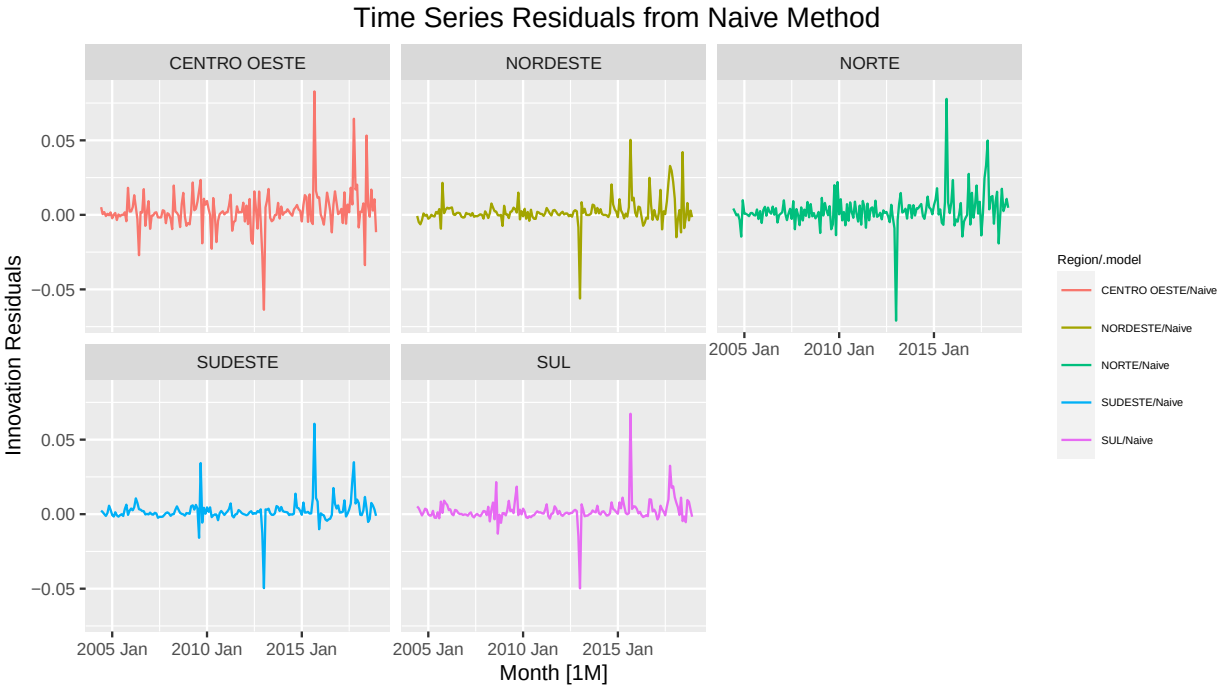
3.7.3 Residual Diagnostics from the best method

3.7.3.1 Histograms



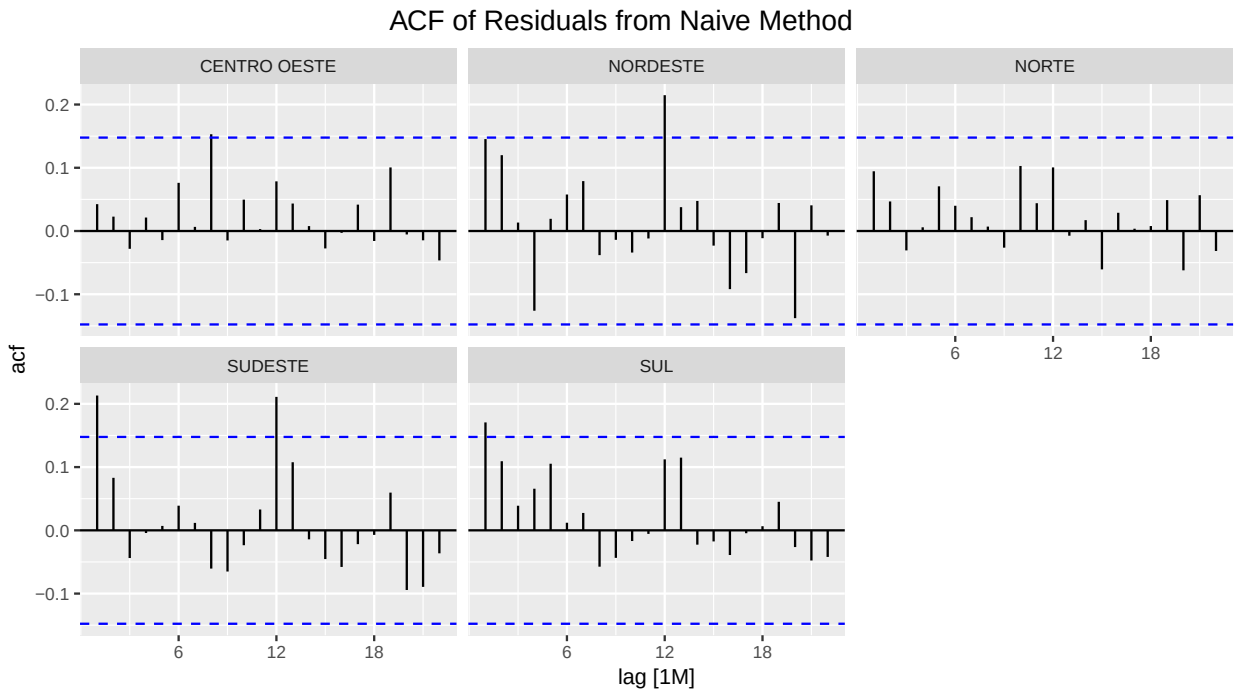
The residuals all appear to be normally distributed and centered around zero. The residuals from the centro oeste and norte regions in particular closely approximate a normal distribution.

3.7.3.2 Time series plot



The residual plots show that there may be some unexplained variation in the model in the more recent years, as the variation in the resiudal plots increases at the end. There appears to be several unusual observations in recent years that also could explain why there are more errors with more recent data.

3.7.3.3 ACF plots of residuals



The ACF plot shows no significant autocorrelation in the residuals for the centro oeste region.

3.7.3.4 Ljung-Box test The Ljung Box test results are not signficant for the centro oeste region, so we can conclude that the residuals are indistinguishable from a white noise series.

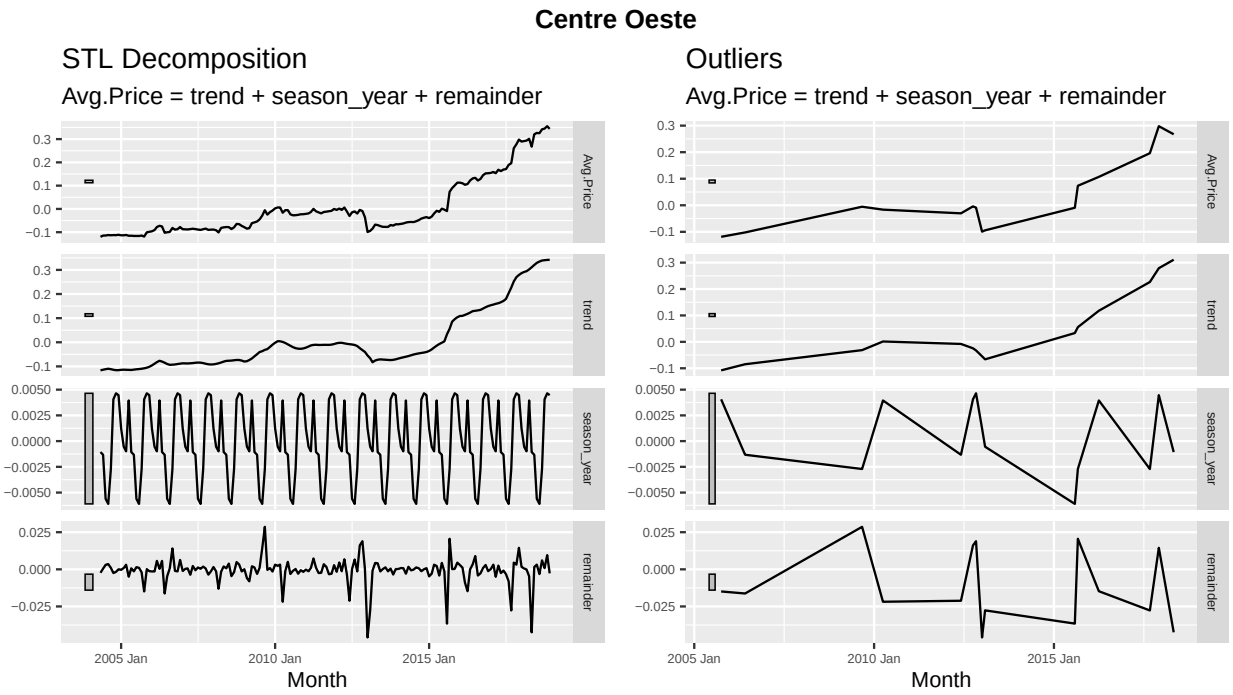
Table 5: Ljung-Box Test for Detecting White Noise

Region	.model	lb_stat	lb_pvalue
CENTRO OESTE	Naive	6.593474	0.7631848
NORDESTE	Naive	11.594769	0.3130919
NORTE	Naive	5.569727	0.8500257
SUDESTE	Naive	11.557918	0.3157319
SUL	Naive	11.584791	0.3138052

3.8 Additional Analysis

3.8.1 Checking for & Removing Outliers

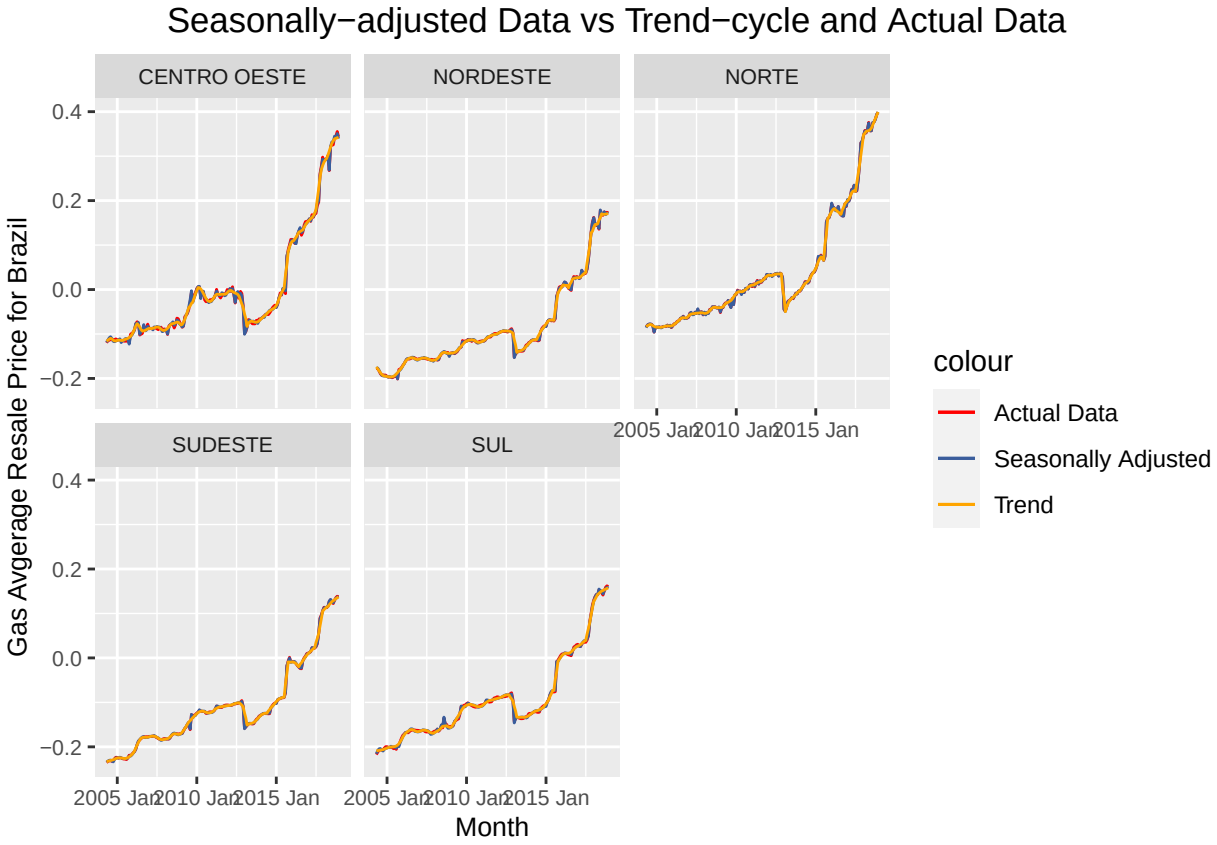
Since we were not satisfied with the results from our forecasting models, we decided to look for outliers and see if removing the outliers would improve our model.



Even after removing the outliers, we saw that the Naive model performed the best in forecating the gas prices for the test set.

3.8.2 Seasonally-Adjusted Data versus Trend-Cycle component and Actual Data

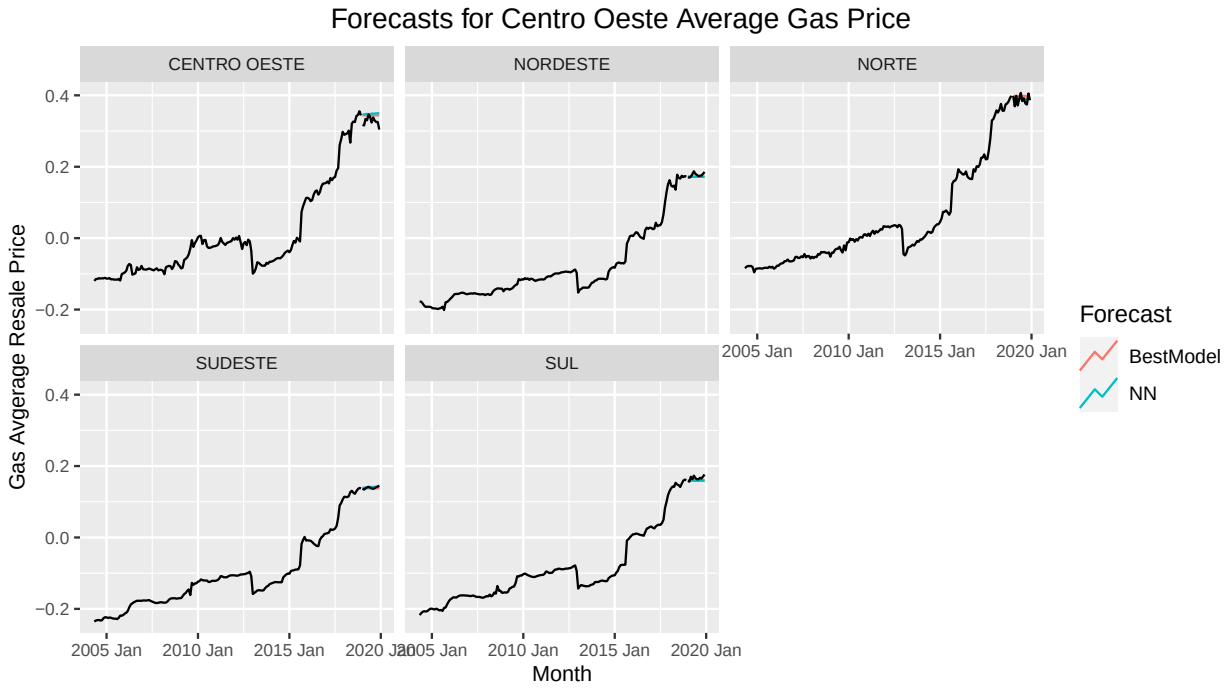
We have plotted the seasonal and trend-cycle components with the actual data to understand each component



more.

3.8.3 Neural Network with Best Model

We have applied a neural network to compare with the best-performing model from our earlier analysis, Naive Model.



3.8.4 Accuracy Measures on an average

Table 6: Accuracy Measures on an Average

.model	RMSE	MAPE	MASE	MAE
BestModel	0.0109625	3.531220	0.2290829	0.0090083
NN	0.0111878	3.790205	0.2420825	0.0095316

We can see that there is not much significant difference in the RMSE or MAPE value for both the models.

4 Conclusion

In conclusion, we have discovered that a simpler model actually performs better and is more accurate for this data set than an ARIMA or ETS model. A possible explanation for this result is that the trend component is the only significant component that we have identified in the model apart from the error component. In the absence of a strong seasonal trend or additional component, a simple Naive method performed the best among the models we tested. The Naive model had a MAPE value of 5.26%.

One way we could improve upon these results is by considering the hierarchical grouping of the data in our forecasting model. For example, the data has been filtered by region, with several states that are aggregated into a single region. There are also several classifications of products that have been aggregated for the model. A bottom-up forecasting approach, beginning at the state and product classification level and then aggregating upwards, could produce a more accurate and nuanced model.

Another consideration for the future is applying more complex models. Two models that have gained popularity in forecasting energy data are the Grey model and the Prophet model. Fitting these models is a far more advanced process than the methods applied in this project, but would likely produce a very accurate forecast. As discussed in the literature review, machine learning models like the random forest model have also shown great success in forecasting energy data. Future studies on this data could focus on applying

any one of these methods to obtain more accurate and robust forecasting models for energy prices and other applications.

5 References

5.1 Dataset & R References

- https://www.kaggle.com/datasets/matheusfreitag/gas-prices-in-brazil
- https://community.rstudio.com/t/help-with-tsibble/72885/5
- https://stackoverflow.com/questions/71914704/override-using-groups-argument
- https://cran.r-project.org/web/packages/tsibble/vignettes/implicit-na.html

5.2 Works Cited

Ceylan, Z., Akbulut, D., & Baytürk, E. (2022). Forecasting gasoline consumption using machine learning algorithms during COVID-19 pandemic. Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, 1–19. https://doi.org/10.1080/15567036.2021.2024919 Samuelson, R. J. (2014, December 3). Key facts about the Great Oil Crash of 2014. The Washington Post. Retrieved December 8, 2022, from https://www.washingtonpost.com/opinions/robert-samuelson-key-facts-about-the-great-oil-crash-of-2014/2014/12/03/a1e2fd94-7b0f-11e4-b821-503cc7efed9e_story.html Stecker, T. (2013, October 17). How the oil embargo sparked energy independence - in Brazil. Scientific American. Retrieved December 8, 2022, from https://www.scientificamerican.com/article/how-the-oil-embargo-sparked-energy-independence-in-brazil/ Zhang, J., & Zhao, J. (2022). Trend- and periodicity-trait-driven gasoline demand forecasting. Energies, 15(10). https://doi.org/10.3390/en15103553