Time Series Analysis of Monthly Retail Trade and Food Services Sales

*Abstract*— This study is concerned with the prediction of billions of dollars worth of retail trade per month using different forecasting models such as ARIMA, ETS, TBATS, HOLT WINTERS', NN and PROPHET. While doing this, since there is a problem of heteroscedasticity in the series, several GARCH models were used to compensate for this shortcoming, and in the end, the model with the highest performance was determined by generating a partial bootstrap forecast.

Keywords—Forecast, SARIMA, garch, ets, holt winters’, bats, nnetar, prophet, volatility.

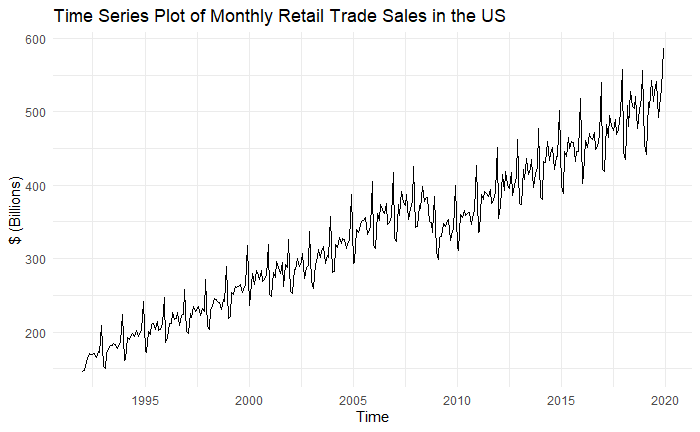
# INTRODUCTION

The web page of the US Census Bureau states, "The Advance Monthly and Monthly Retail Trade Surveys (MARTS and MRTS), the Annual Retail Trade Survey (ARTS), and the Quarterly E-Commerce Report work together to produce the most comprehensive data available on retail economic activity in the United States."

The aim of this work is to comprehend the change in monthly retail trade and food services sales in the US and gain some insight into how to predict the conditional mean of future values based on current and past data.

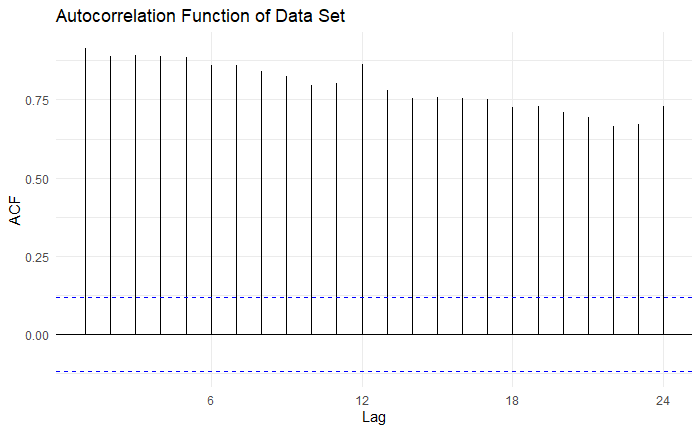
# DATA DESCRITPION AND PREPROCESSING

The data set is taken from The United States Census Bureau [website](https://www.census.gov/econ/currentdata/?programCode=MARTS&startYear=1992&endYear=2022&categories%5b%5d=44X72&dataType=SM&geoLevel=US&adjusted=1&notAdjusted=1&errorData=0) which consistently publishes data related on the American people and economy. The data set includes the estimates of monthly sales for retail trade and food services in the US from the year 1992 to 2019.



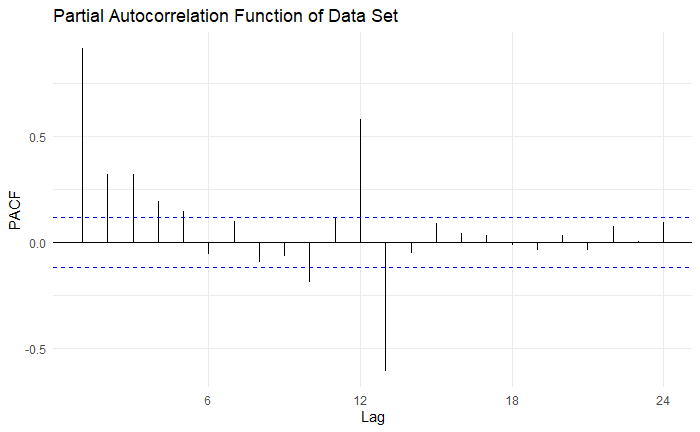
***Graph 1****: Time Series Plot of Data Set*

We see that there is a steady increase from 2004 to 2008, but retail trade declined around the second quarter of 2008 and did not start to climb back until the last quarter of 2009. The fall in retail trade due to the stock market crash on September 29, 2008, partially reflects the economic recession and causes a structural break in the series. In addition, it is found out that the average increase rate is 13.619 billion dollars per year. The time series plot clearly shows that the series is not stationary as it has a stochastic trend with some indicative ups and downs. Moreover, the series displays the seasonality of monthly retail trade and does not seem to be stationary in variance, though it's better to check with a few tests to be able to reach a firm statement.



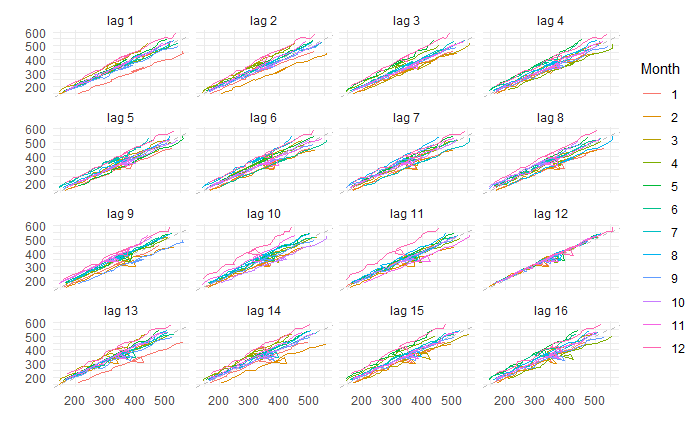
***Graph 2****: ACF Plot of Data Set*

All values shown are significantly far from the White Noise band, and the only pattern is perhaps a slow linear decrease with increasing lag. It is seen that there is a slow linear decay at seasonal lags. As in the time series plot of the data, the ACF plot also points out a non-stationary process, so we do not need to interpret the PACF plot.



***Graph 3****: PACF Plot of Data Set*

It can be said that the PACF is cut off after the first lag, which is a more noticeable spike. However, we cannot interpret this plot for now, as we have had previous findings that the process is not stationary.

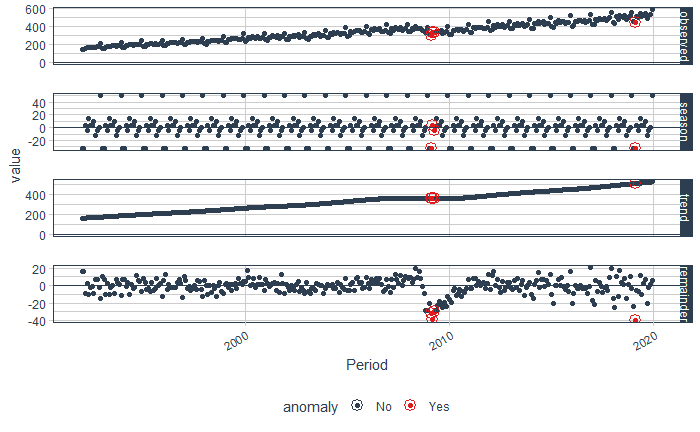


***Graph 4****: Lagged Scatterplots*

While examining the Graph 4 that provides the bivariate plot for each level of lag (1-16 lag), at the right-most 12 lag we can see that the relationship appears the strongest for it, thus supporting our seasonality pattern which appears in the ACF plot as well.

In addition, the mean values for each month, which also reveal the underlying seasonality over time, were examined and observed that it is the lowest in the first quarter, shows an increasing trend over the year, and reaches its highest point in December.

After performing a detailed exploratory data analysis, we check the anomalies in all our data and clean the series from the anomalies it has. In doing this, we use stl decomposition method and look at every component that our time series data exhibits. The 18 identified anomalies were removed and replaced with interpolated values.



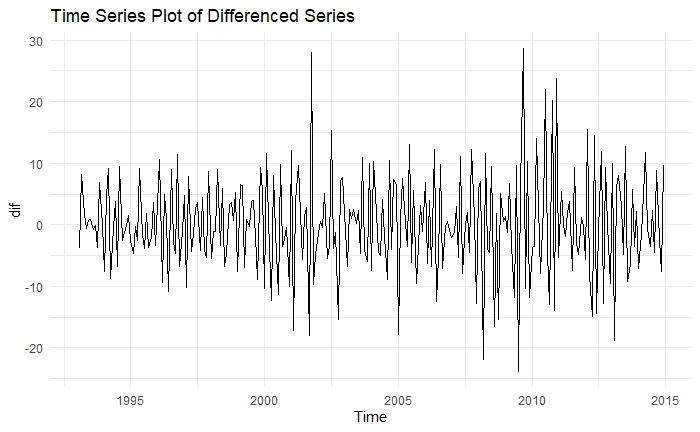
***Graph 5****: Anomaly Detection Plot*

As a next step, we divide our data set into two parts which are train and test sets being necessary for model validation. Here we use 80% of the data as train and the remaining as test set. We keep the last 60 observations for the test set.

To identify the structure of time series data we need to satisfy the most vital assumption known as stationarity. When the process is in statistical equilibrium, we can apply statistical-based modeling. With this purpose, we generate specific lambda values concerning the yule-walker method. According to the outputs, we can use either the 0.4144289th power of the series or do y^05 transformations seeing the range of λ is very close to 0.5. However, after transforming the series since there is no significant change in the shape of the time series plot, it is decided not to use any variance-stabilizing transformation at this stage.

In this project, non-seasonally adjusted monthly data is used; therefore, for the non-stationary checks, first both KPSS and PP tests are applied and seen that our series is not stationary, which also supports mentioned statements while examining the time series and ACF plots. Then, for seasonal unit root, HEGY and Canova-Hansen tests are required to obtain more realistic results, as we have sufficient evidence to conclude that the series has a stochastic trend and shows a seasonal pattern.

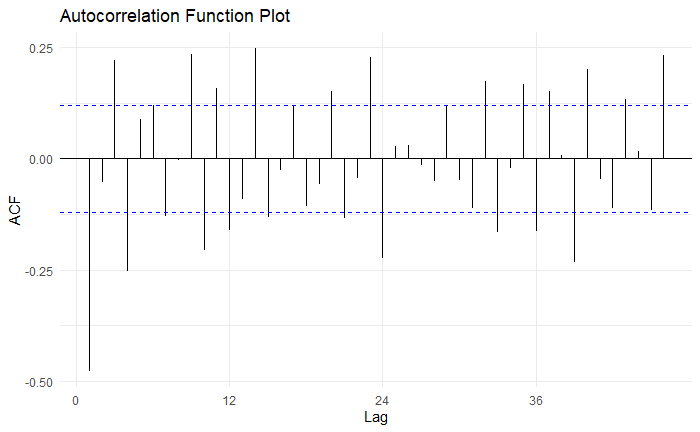
Although the output of our tests states the series becomes stationary after taking only one regular difference, it is judged appropriate since taking the additional seasonal difference does not cause over-differencing and makes the mean constant around zero. Therefore, after taking both regular and seasonal differences, we make sure that we have removed the trend and the series becomes stationary in the mean by checking the unit root test results again.



***Graph 6****: Time Series Plot of Differenced Data Set*

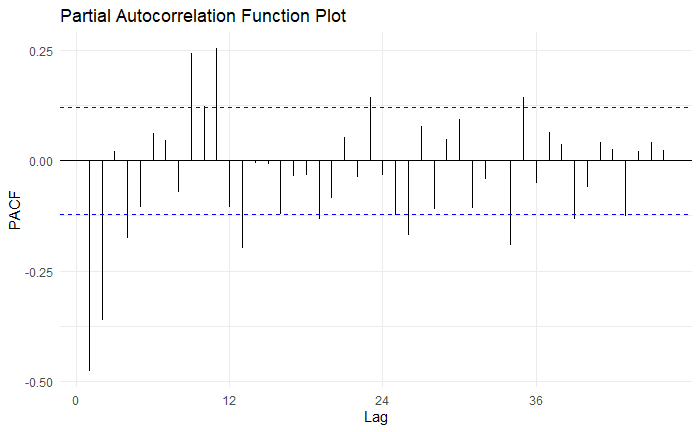
# MODEL SUGGESTION

Unfortunately, identifying a proper SARIMA model in our case is somehow challenging because we encounter long-term memory in our series, and determining the order of the process by looking at ACF and PACF plots makes some confusion. Besides, we cannot consider additional methods to suggest a model since our series has a seasonality component.



***Graph 7****: ACF Plot of Stationary Data Set*

It seems that the ACF still shows slow decay, normally by looking at this plot, the MA order of process should have been identified.

******

***Graph 8****: PACF Plot of Stationary Data Set*

On the other hand, the PACF of the process cuts off after lag 2 that gives the order of AR model. Accordingly, the suggested models are SARIMA(2,1,2)x(2,1,1)[12] and SARIMA(4,1,1)x(2,1,2)[12].

# MODELLING AND DIAGNOSTIC CHECKING

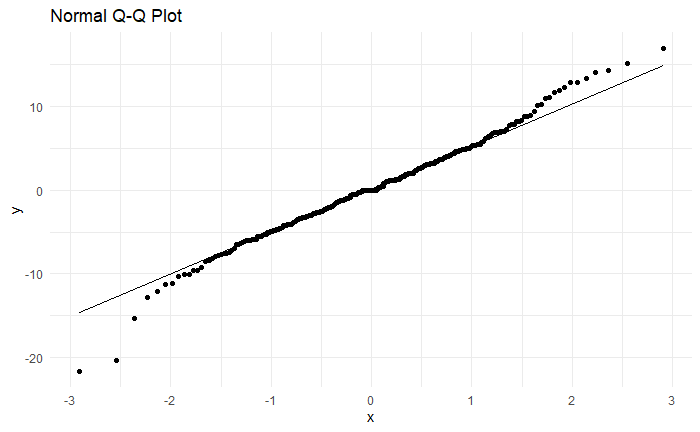
After deciding the order of the possible models, we run MLE and estimate the parameters. By comparing the information criteria of several models, we choose the best-fitted one whose coefficient estimates are all highly significant and then keep on further diagnostic checks of the model. Hereby, SARIMA(2,1,2)x(2,1,1)[12] performed the best and found the most appropriate one in this respect.

***Table 1:*** *Summary of Model*

|  |
| --- |
| 1. Model 1 |
| 1. Coefficients:   ar1 ar2 ma1 ma2 sar1 sar2 sma1 |
| 1. -1.025 -0.315 0.371 -0.371 0.184 -0.189 -0.622 |
| 1. s.e:   0.108 0.102 0.107 0.110 0.107 0.072 0.088 |
| 1. sigma^2 estimated as 34.54: log likelihood=-838.83 2. AIC= 1693.33 AICc=1694.22 BIC=1722.23 |

On the residuals, we perform portmanteau lack of fit test to identify if there is any dependence structure and then we also check the validity of some assumptions.

First, let us find out whether the residuals are normally distributed or not by drawing the Q-Q plot.

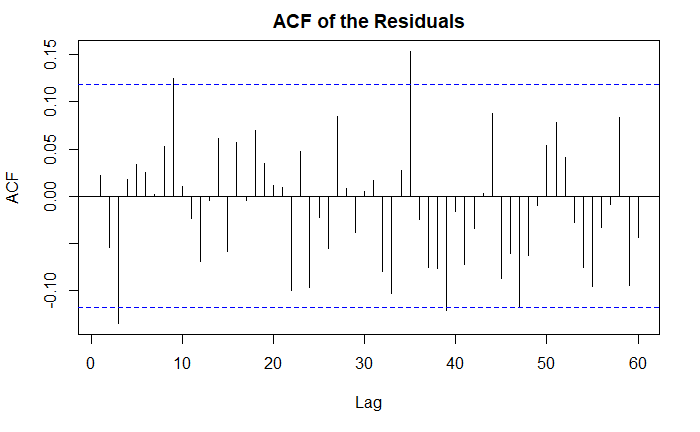


***Graph 9****: QQ plot of the standard residuals*

The QQ Plot shows that some residuals of the model do not lie on the 45-degree straight line which indicates residuals may not follow a normal distribution. To be sure about non-normality, Shapiro-Wilk and Jarque-Bera should be considered too. In our case, both tests give the same result, Jarque-Bera (p<0.05) and Shapiro-Wilk (p<0.05) suggest that errors are not distributed normally. In other words, we do not have enough evidence to claim that we have residuals with normal distribution.

Another point we might need to pay attention to here is that even when we used transformed data to ensure the assumption of normality, we could not achieve the desired result, so we can say that the main reason for this is the non-constant variance problem that will be discussed later in the paper.

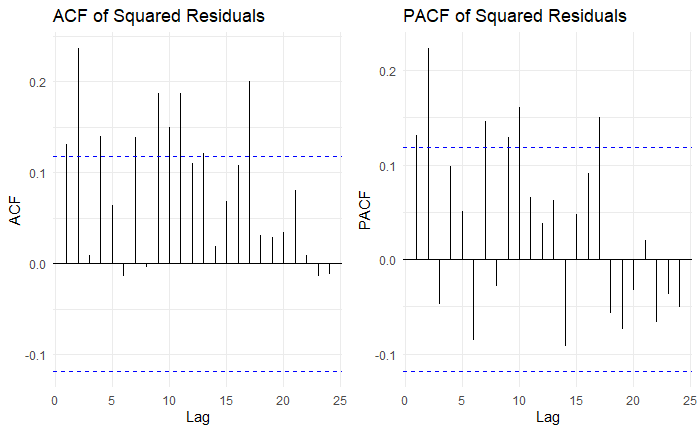
Second, we perform Breusch-Godfrey, Ljung-Box, and Box\_Pierce tests for possible autocorrelation in residual series. For detecting serial correlation, the ACF plot of the standardized residuals from the fitted model is also useful. If all spikes are in the White Noise band, we can say that the residuals are uncorrelated.



***Graph 10****: ACF Plot of the standard residuals*

In the ACF plot, almost all spikes are in the White Noise band, other than some strange behavior, the plot does not suggest any major irregularities with the model. To be sure, we should apply the mentioned formal tests. Briefly, all tests suggest with 95% confident that the residuals of the model are uncorrelated (p>0.05). Hence, we can conclude that the series does not show statistically significant evidence of nonzero autocorrelation in the residuals.

The third assumption check is the heteroscedasticity of the residuals. To test this assumption, we apply White’s test and Bresuch-Pagan test and look at the ACF and PACF plots of the squared residuals.



***Graph 11****: ACF and PACF of the squared residuals*

From the above plots, we can see that squared residuals are not in the 95% White Noise band and they display some significant autocorrelations and hence provide some evidence that the residuals are not independently and identically distributed. Before going any further, let us consider the outputs of formal tests. The errors are heteroscedastic according to both the Studentized Breusch-Pagan test and Engle’s ARCH test (p<0.05). Thus, we state that the error variance is not constant by time, and it should be modeled.

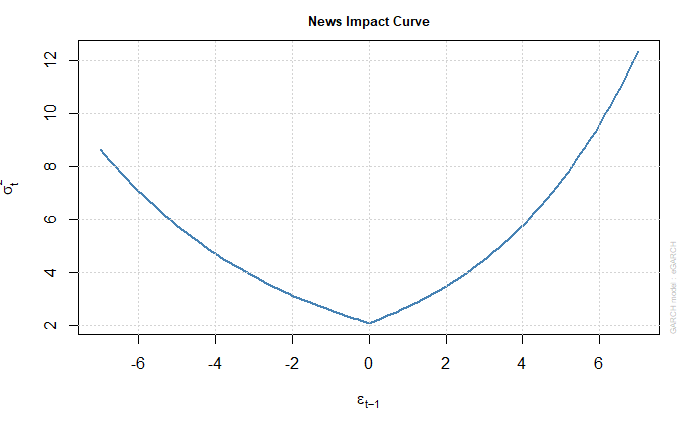
For coping with the heteroscedasticity problem, except knowing how to model the conditional variance structure of a time series, it is also nice to be aware of the conditional variance of a financial asset's return, which is one of the widespread features of financial time series, is often adopted as a measure of the risk of the assets.

Here we could not manage the use of the GARCH method for the series since it includes seasonal components. However, at this phase, we decided to work on the same data set with a seasonally adjusted just for the GARCH-type model. Every step made up until now has also been applied to the seasonally adjusted data set. Fortunately, this is how we get the best-performing model.

***Table 2:*** *Summary of GARCH Model*

|  |
| --- |
| Conditional Variance Dynamics  -----------------------------------  GARCH Model : eGARCH(2,2)  Mean Model : ARFIMA(2,0,2)  Distribution : norm |
| LogLikelihood : -615.8873 |
| Information Criteria  ------------------------------------  Akaike 4.5499  Bayes 4.7073  Shibata 4.5463  Hannan-Quinn 4.6131 |
| Weighted Ljung-Box Test on Standardized Residuals  ------------------------------------  statistic p-value  Lag[1] 0.376 5.397e-01  Lag[2\*(p+q)+(p+q)-1][11] 9.291 1.149e-06  Lag[4\*(p+q)+(p+q)-1][19] 17.653 3.383e-03  d.o.f=4 |
| Weighted Ljung-Box Test on Standardized Squared Residuals  ------------------------------------  statistic p-value  Lag[1] 0.007757 0.9298  Lag[2\*(p+q)+(p+q)-1][11] 1.884165 0.9674  Lag[4\*(p+q)+(p+q)-1][19] 3.142338 0.9937 |
| Adjusted Pearson Goodness-of-Fit Test:  ------------------------------------  group statistic p-value(g-1)  1 20 22.84 0.2444  2 30 29.87 0.4205  3 40 32.12 0.7746  4 50 45.38 0.6208 |
| Sign Bias Test  ------------------------------------  t-value prob  Sign Bias 0.03194765 0.9745373  Negative Sign Bias 0.82503079 0.410079  Positive Sign Bias 1.33524532 0.1829166  Joint Effect 3.44609150 0.3278215 |

The summary provides the Ljung Box Test and the result shows that residuals have autocorrelation but squared residuals do not. The Adjusted Pearson Goodness-of-Fit Test: is above 0.05, so we are able to obtain asymmetric distribution.



***Graph 11****: New Impact Curve*

When the effect of news is taken into consideration, Sign Bias Test is used to test leverage effect in the standardized residuals. We fail to reject the null hypothesis which states that no significant negative and positive reaction shocks. Furthermore, from the news impact curve, we can see that asymmetries are present in response to positive and negative shocks.

***Table 3:*** *The forecasting performance of GARCH Models*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ME | RMSE | MAE | MAPE | ACF1 |
| **eGARCH** | **-5.51** | **6.672** | **5.644** | **1.174** | **0.682** |
| gjrGARCH | 13.279 | 17.382 | 13.677 | 2.745 | 0.917 |
| sGARCH | 9.343 | 14.357 | 10.940 | 2.192 | 0.915 |

The forecast values obtained from eGARCH model outperforms the other trials. Later on, after performing minimum MSE forecast for the stochastic model, we also apply ets, Holt Winters’, prophet, tbats and nnetar. The best exponential smoothing model for the series is given below.

***Table 4:*** *Summary of ETS Model*

|  |
| --- |
| ETS(M,M,M) |
| Smoothing parameters: |
| alpha = 0.1996 |
| beta = 0.0573 |
| gamma = 0.2155 |
| phi = 0.9687 |
|  |
| Initial states: |
| l = 159.1284 |
| b = 1.0058 |
| s = 1.1938 1.0066 1.0015 09691 1.0243 1.0074 |
| 1.0157 1.0281 0.9878 0.9925 0.8837 0.8894 |
|  |
| sigma: 0.0192 |
|  |
| AIC AICc BIC |
| 2516.554 2519.216 2581.721 |

As illustrated in Table 4, we have exponential smoothing model having multiplicative error, multiplicative trend and multiplicative seasonality. Hence, ETS(M,M,M) is fitted and afterwards the residuals of the model checked by Shapiro-Wilk test, and it is seen that we have residuals with normal distribution. (p>0.05).

After exponential smoothing and Holt Winters’ models, TBATS model is fitted to the series. The model details are given below.

***Table 4:*** *Summary of BATS Model*

|  |
| --- |
| BATS (0.421, {3,1}, -, {12} |
|  |
| Call: tbats(y = AirPassengers) |
|  |
| Parameters |
| Lambda: 0.421019 |
| Alpha: 0.04405455 |
| Gamma Values: 0.1140967 |
| Sigma: 0.208863 |
| AIC: 2546.755 |

By using Shapiro-Wilk test the residuals of the TBATS model are checked and seen that they do not follow normal distribution (p<0.05).

The next one is the Neural Network model where past observations are considered as input variables. The model details are represented as follows.

***Table 5:*** *Summary of NNETAR Model*

|  |
| --- |
| Nnetar model |
| ## Series: df.t |
| ## Model: NNAR(2,1,2) |
| ## Call: nnetar(y = df.t) |
| ## Average of 20 networks.each of which is a 3-2-1 network with 11 weights |
| ## options were - linear output units decay=0.08 |
| ## sigma^2 estimated as 96.47 |

The model is NNAR(2,1,2) and due to seasonality, p = 1 is set. Whether the residuals of the created model with normal distribution are examined by Shapiro-Wilk test and it is found that the residuals are not normally distributed (p<0.05).

Lastly, we fit prophet model to the series and then check the residuals of the model, we see that they are not normally distributed with respect to Shapiro-Wilk test (p<0.05).

After fitting the models, we obtain the accuracy measures of all models on both train and test sets. The results are presented in Table 6 for train tests and in Table 7 for test sets.

***Table 6:*** *The train accuracy of models*

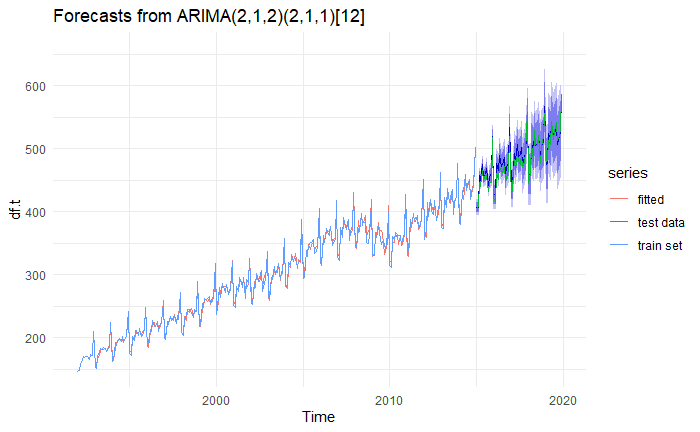
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | ME | RMSE | MAE | MAPE | MASE | ACF1 |
| **SARIMA** | **0.191** | **5.660** | **4.289** | **1.409** | **0.297** | **-0.001** |
| ETS | 0.388 | 5.742 | 4.491 | 1.492 | 0.311 | -0.005 |
| HW | -0.00 | 5.771 | 4.443 | 1.472 | 0.308 | -0.034 |
| BATS | 0.482 | 5.678 | 4.407 | 1.506 | 0.306 | -0.031 |
| PROPHET | 0.001 | 6.386 | 5.100 | 1.812 | 0.324 | 0.108 |
| NNETAR | 0.001 | 9.822 | 7.140 | 2.295 | 0.495 | 0.574 |

***Table 7:*** *The forecasting performance of models*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | ME | RMSE | MAE | MAPE | MASE | ACF1 |
| **SARIMA** | **-6.13** | **9.939** | **8.377** | **1.781** | **0.581** | **-0.00** |
| ETS | 6.936 | 18.843 | 14.798 | 2.984 | 1.026 | 0.777 |
| HW | -14.96 | 17.551 | 15.081 | 3.092 | 1.046 | 0.032 |
| BATS | 7.001 | 15.043 | 11.538 | 2.318 | 0.799 | 0.683 |
| PROPHET | -10.09 | 13.207 | 11.213 | 2.401 | 0.698 | -0.117 |
| NNETAR | -4.594 | 16.713 | 13.390 | 2.760 | 0.984 | 0.056 |

The SARIMA model performed very well for all measurements compared to the other model trials. The root mean squared errors and mean absolute percentage errors are somewhat similar between the models, however the SARIMA model outperformed the other methods, therefore we can conclude that the SARIMA model prevailed on the MASE metric with the best forecasting performance. Also, it is observed that prophet has the second-best forecasting performance when compared to other models. Lastly, ets model has the lowest forecasting accuracy as shown in the tables above.

The best forecasting performance of the model can also be analyzed from the following plot.



***Graph 10****: Forecast Plot of SARIMA Model*

# DISCUSSION AND CONCLUSION

In this paper, after cleaning the data set and splitting into train-test, the stationarity of the series checked; time series, ACF&PACF plots, and KPSS, PP, HEGY and Canova-Hansen tests are examined, it is seen that this condition cannot be met because there is a stochastic trend and seasonality in the series. To solve these problem, both regular and seasonal differences were applied. After making the process as stationary as possible, the best model with fewer and more significant parameters was selected by using the AIC comparison and t statistics from the determined temporal models.

Using the best model obtained, diagnostic checks are performed on the residuals. At this point, the problem of non-normal errors arose, and box-cox transformation was applied, however it could not be a permanent solution as the series had non-constant variability. In addition, the existence of volatility, the most time-consuming part of our project, was verified using visual inspection tools and formal tests. At this stage by modeling eGARCH model on seasonally adjusted series, we achieved some good results*.* Overall, we have learned to work on time series analysis within the scope of the project, for further work we should especially focus on finding superior methods relative to available alternative modelling strategies.

# REFERENCES

|  |
| --- |
| Baillie R.,, Chipulu, M., (2009). Modelling long memory and structural breaks in conditional variances: An adaptive FIGARCH approach. Journal of Economic Dynamics & Control, 33, 1577–1592.  Ghulam A. (2013). EGARCH, GJR-GARCH, TGARCH, AVGARCH, NGARCH, IGARCH and APARCH Models for Pathogens at Marine Recreational Sites. Journal of Statistical and Econometric Methods, 42(3), 57-73.  Karlsson, O.. (2016). Volatility forecasting under structural breaks. Uppsala University, Retrieved from <https://www.diva-portal.org/smash/get/diva2:957411/FULLTEXT01.pdf> |