Time Series Analysis and Forecast

MADSAD

Gustavo Baldaia (up201606386)

1 INTRODUCTION

The study and analysis of a time series is the main purpose of this report, where a selected time series will be tested with different approaches and methodologies in order to verify if there is an accurate and trustful process that can predict and forecast future observations with a high level of effectiveness.

The time series was retrieved from the Federal Reserve Bank of St. Louis economic data repository, and it expresses the total amount in millions of dollars of Alcoholic Beverages sales in the United States of America since January of 2003 until December of 2018 with monthly frequency, totalizing 193 observations. For the purpose of this study, to test the accuracy of the predictions, the last 12 observations were excluded from the analysis.

Before the application of any of the different analysis, an observational analysis of the time series was performed, in order to understand and gather some basic information and knowledge regarding some of its characteristics. Then three approaches were applied, exponential smoothing, decomposition method and the ARIMA model, considering the type and characteristics and all the analysis and tests require from them for then to be possible to estimate the forecast observations from the exponential smoothing and ARIMA model that were compared in the end to verify which obtained the most accurate results.

All the tests and analysis used on this research were performed with the help of the software Rstudio.

2 TIME SERIES STUDY

2.1 OBSERVATIONAL ANALYSIS

From the graphical representation of the data, it is possible to extract some information regarding the time series, that later will be crucial to identify and select the adequate models in this case. The *Figure 1* demonstrates an incremental trend over the years with the total amount of sales more than double from the first few months to the last highest observations.

Another noticeable characteristic of the time series is the clear presence of a seasonal component, as the sales follow a specific pattern every year. About this seasonal component, it is also patent that its variance is

not stable, as it grows over the years. Where the range of the fluctuations over a year is a lot more significant in the last years of the data, when compared to the first few years.

Finally, is also perceived that the time series is non-stationary as the increasing trend, clearly creating a dependency of the mean of the time series with the current period. The behavior of the data is not the same for each period of time t.

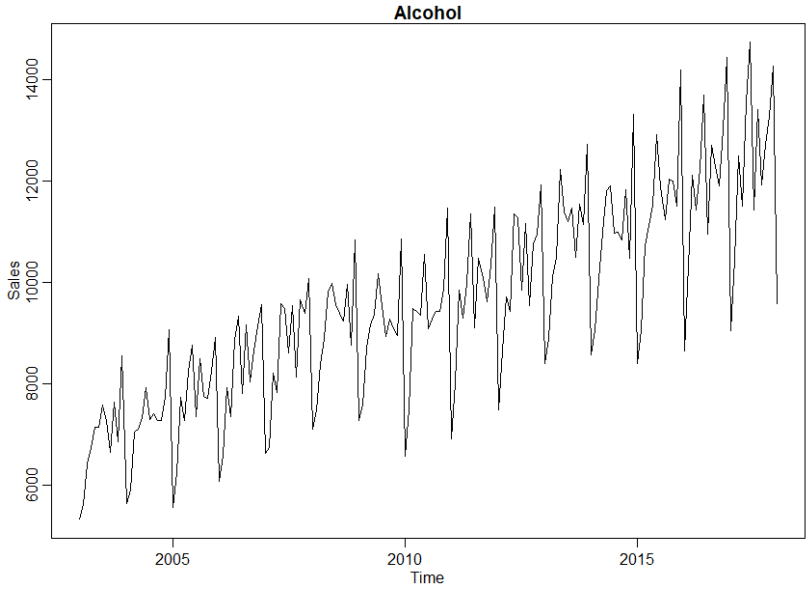

Figure 1 – Graphical representation of the time series

2.2 SMOOTHING METHOD

Exponential smoothing is a forecasting method for univariate data, that is based on the weighted sum of past observations for the determination of future observations. This process is characterized by the moving averages of recent observations, where in simple moving averages methodologies, all the observations have been equally value to determine the next values, exponential methods, decrease exponentially the importance of historical data over time, when estimating new observations.

There are 3 commonly basic variations of exponential smoothing, that are simple exponential smoothing (Brown, 1959); trend-corrected exponential smoothing or Holt's method (Holt, 1957); and Holt–Winters' method (Winters, 1960). The different between these is the incorporation of different components into the model, the first one, the simpler forecast the time series data without considering trends or seasonality, while the Holt's method incorporates a trend component and finally the last one is an extension of the trend-corrected method that additionally incorporates a seasonal component into the analysis of the time series.

As seen before the time series present not only shows some evident seasonality component that seems to grow in terms of its variance over the years, but also an increasing trend of the observations over the course of time, which leads to the necessity of using multiplicative smoothing methods adequate as the multiplicative Holt Winter's method with triple exponential smoothing.

```
Holt-Winters exponential smoothing with trend and multiplicative seasonal component.

Call:
HoltWinters(x = alcohol_ts, seasonal = "m")

Smoothing parameters:
 alpha: 0.092777
 beta : 0.0010976
 gamma: 1

Coefficients:
             [,1]
a   11956.303530
b      26.040462
s1      0.900658
s2      1.072249
s3      0.986742
s4      1.151855
s5      1.244718
s6      0.962912
s7      1.126178
s8      1.004706
s9      1.067449
s10     1.112545
s11     1.198879
s12     0.799913
```

Figure 2 - Holt Winter's multiplicative model

The smoothing parameters assign by the method, α, β and γ, indicate that firstly, by the very low value of alpha, that is related to the moving averages that the past observations also have a crucial role and are take into account to make future predictions. The beta value of almost, suggesting that slop of the trend component (that is the b value) is almost constant over time, and finally the gamma parameter, that is associated with the seasonality with a clean value of 1 indicates that the seasonal component at current observations is based almost entirely in very recent data.

The monthly coefficients evidence the periods where seasonality have a bigger impact on the time series, as the end and bigging of the year, from December to March are the weakest of the year, followed by April and May where the weight of the seasonal component is at very high levels as well in June, September and November.

2.3 DECOMPOSITION METHOD

The decomposition methods, as the name indicates, separate the time series into different components, that are: Trend (T_t), Cycle (C_t), Seasonality (S_t), and random fluctuations (E_t). There are additive and multiplicative

types of decompositions, that are differentiate by the relation between the components that generate the time series. Additive decomposition assumes the time series as a function of the sums of the four components as the multiplicative follows the assumption that the time series is given by the product of its components. As most of the times is very hard to compute the cycle component by itself, it is commonly calculated together with the trend component ($T_t * C_t$).

$$X_t = T_t * C_t * S_t * E_t$$

Again, as has been seen before, due to the increasing magnitude of the seasonal component over time, the multiplicative approach is the most suitable for this time series.

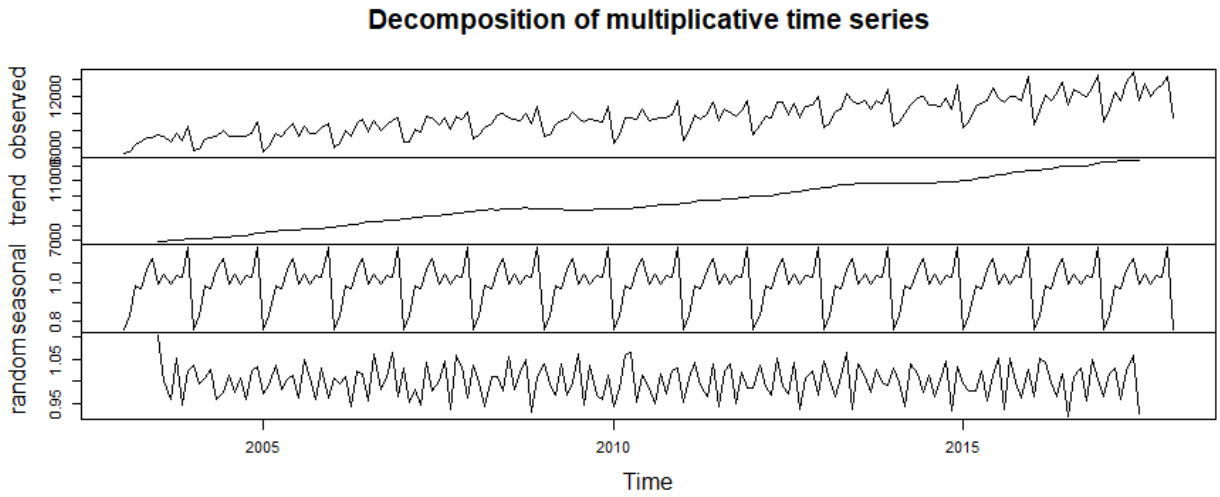

Figure 3 - Multiplicative decomposition method

The results demonstrate once again what have been pointed out before in this report, a presence of an increasing trend of the data, that are adjusted for seasonality, that is almost linear and the existence of an annual seasonal component that is constant and equal over time.

2.4 ARIMA MODEL

The final approach is the development of an ARIMA model for the time series, that stands for AutoRegressive Integrated Moving Average model, that in this case is a SARIMA, as the time series is seasonal. This model links three different aspects of analysis, as the name indicates. First the autoregression, that is the dependency

of a certain observation with a number of lagged observations (p, P). The integrated that is the degree of differencing or subtracting an observation from an observation at a previous time period until make the time series stationary (d, D). Finally, the moving averages, that use the dependency between an observation and a residual error from a lagged moving average (q, Q).

The SARIMA model is defined as $\mathrm{SARIMA}\ (p,d,q)x(P,D,Q)_{12}$ where the lowercased parameters are related to the non-seasonal component and the uppercased parameters relate to the seasonal component of the time series.

Entering the first phase, that is the identification phase, where will be analyzed in depth the time series, to determine the adequate parameters for the model. In this case as have been pointed out the non-stationarity in the variance, it was applied a Box-Cox transformation in order to stabilize the variance.

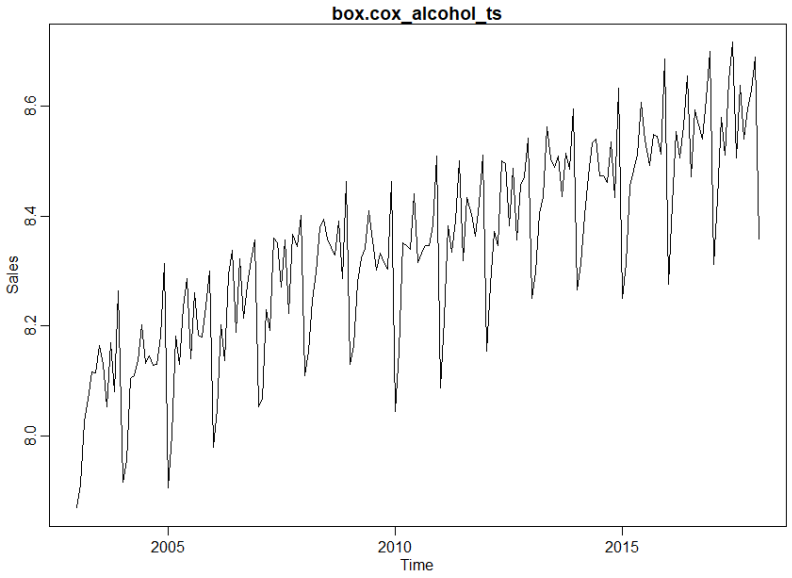

Figure 4 - Graphical representation of the time series with the Box-Cox transformation

After this transformation, with a lambda value of -0.02042464, it is clear that the variance was stabilized over time.

The next step was the identification of the degree of differencing both of the seasonal component and non-seasonal component that would be required to make the time series stationary, as it has concluded from the observational analysis that the time series is non-stationary. To verify that statement it should be analyzed the ACF and PACF graphics of the time series and tested its non-stationarity.

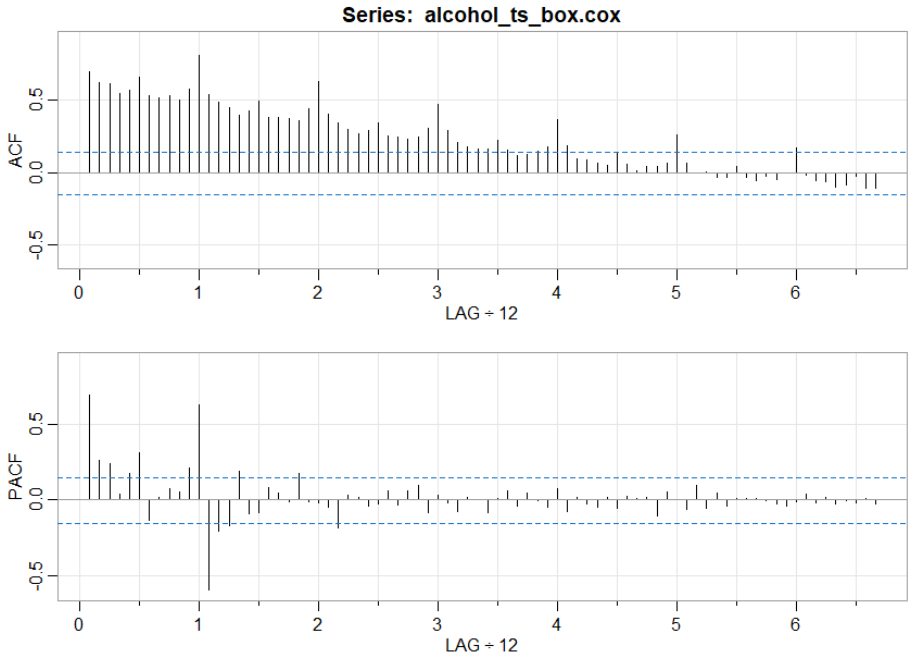

Figure 5 - ACF and PCF graphics with the Box-Cox transformed time series

From the ACF it is possible to notice that it is decreasing the correlation level slowly over the lags, the PACF demonstrates significant correlation levels on the 3 first years, although present a few high peaks after that. This is another strong indication that the time series is non-stationarity as it presents significant levels of correlation for a long period of time.

In order to remove the seasonal non-stationarity and the presence of a unit root, were applied a seasonal difference where it is subtracted a lagged observation, one period apart, from the current observation: $\nabla_{12} X_t = X_t - X_{t-12}$.

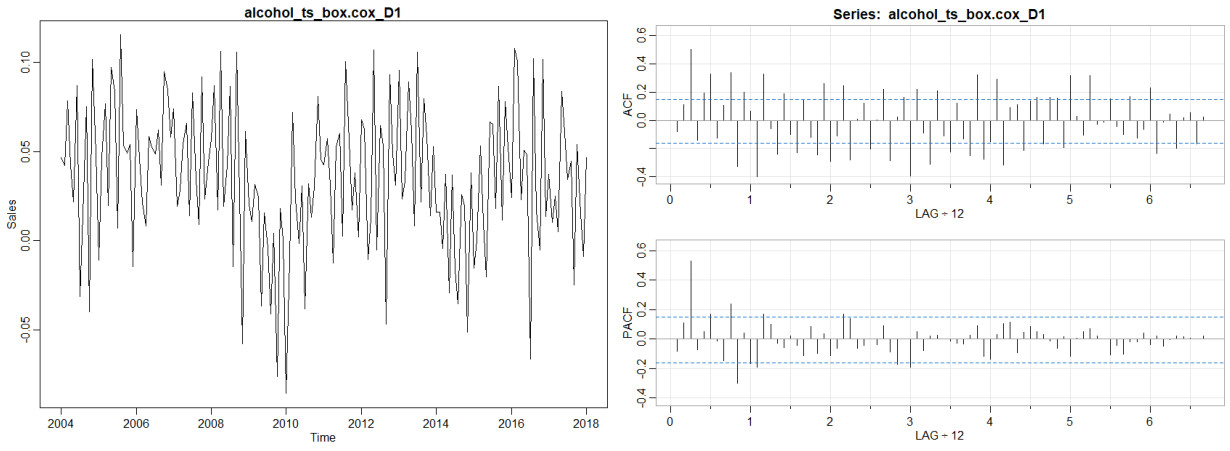

Figure 6 – Graphic of the seasonally difference time series and respective ACF and PAFC

The transformed time series present a more stable plot distribution and the ACF although showing some significant correlation peaks over the time, after the highest peak on the third year the overall correlation levels are lower than the original time series. The PACF also demonstrates the highest correlation peak in year three, from which the significant correlation levels decrease to mostly non statistically significant values.

Before thinking about a non-seasonal differencing, it was needed to verify the stationarity of the one period differenced time series, through some statistical tests as Augmented Dickey-Fuller (ADF) Phillips and Perron test (PP) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS).

The ADF test, that assumes as null hypothesis the non-stationarity of the time series, obtained a t-statistics of -3.9184 that is lower than reference value for a 5% significance level (-1.95) so the H0 is rejected, and the time series is considered stationary.

Phillips and Perron test also assumes as null hypothesis non-stationarity of the time series as it presents more than one unit root. The t-statistic value obtained, of -14.0887 is again lower than the reference value for a 5% significance level (-2.87) so the H0 is rejected, and the time series is considered stationary.

Finally, the KPSS unit root test in contrast to the two tests previously performed, consider as null hypothesis the stationary of the time series. The result obtained is out of the critical region, so in this case the H0 is not rejected, which indicates just like the others test, the stationarity of the time series.

With the stationarity of the time series achieved after a seasonal difference, they are two already known parameters of the ARIMA model that are d=0 and D=1, for the degree of differencing.

For the AR (autoregression) component of the time series, as seen before from the ACF analysis, the highlighted period is year 3, where the correlation achieved its peak value, unregularly decreasing after that. So this suggest an AR(3) model for the non-seasonal component (p). Regarding the seasonal AR component, from the PACF it is possible to conclude that the correlation significance only gets not statistically relevant after year 3, which suggests an AR component between 2 and 3, that will later both be tested to check the most appropriate. The q and Q components related to the moving average are more difficult to identify, so a few values were tested, and different combinations attempted to obtain the best model possible.

Also for additional help in this procedure, a function in Rstudio, that automatically determines the most suitable ARIMA model for a time series (auto.arima) were applied to checked the parameters selection that was later used as a base line to tweak a few values and parameters in order to improve the current model. The ARIMA suggested was a $\text{SARIMA}\ (3,0,0)\text{x}(2,1,2)_{12}$ with drift, that was in line with the previous analysis made to the parameter's selection.

While testing different combinations the Akaike Information criterion was take into account, enhance the models that obtain the lower AIC value, as that indicates a lower loss of information from the model. In the table below can be seen the different best performing models, with the corresponding AIC and the selected one, $\text{SARIMA}\ (3,0,2)\text{x}(2,1,3)_{12}$.

Table 1 – AIC values for different ARIMA models

ARIMA Model	AIC
$\mathbf{SARIMA\ (3,0,0)x(2,1,2)_{12}}$	-720.36
$\mathbf{SARIMA\ (3,0,0)x(3,1,2)_{12}}$	-718.28
$\mathbf{SARIMA\ (3,0,1)x(3,1,2)_{12}}$	-722.29
$\mathbf{SARIMA\ (3,0,2)x(3,1,3)_{12}}$	-733.35
$\mathbf{SARIMA\ (3,0,2)x(2,1,3)_{12}}$	-735.37

The selected model did not only achieve the best AIC value, but also proved to be satisfactory during all the tests performed during the diagnostic checking phase. Firstly, the residuals analysis proved to be very satisfactory ad the ACF and PACF showed very low significance statistically, with almost every value under the significant region, proving a very week correlation between the residual's observations. The Ljung-Box text proved the non-rejection of the null hypothesis, which indicates the randomness of the residuals.

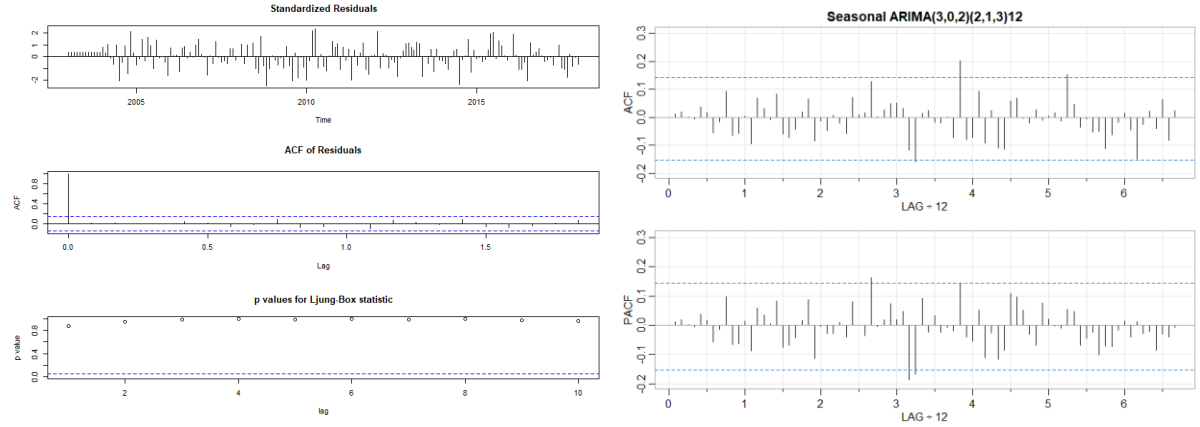

Figure 7 – Diagnostics of the residuals of the $SARIMA\ (3,0,2)x(2,1,3)_{12}$ *and respective ACF and PACF*

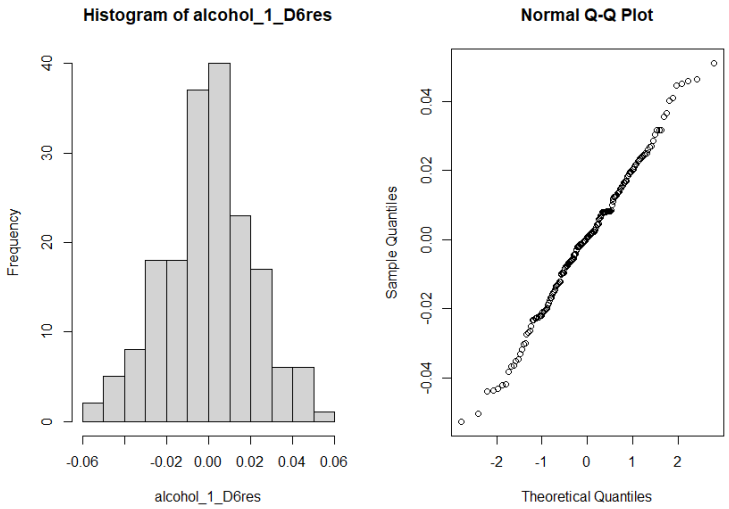

Figure 8 – Histogram and QQ plot of the residuals of the ARIMA model

The histogram and the QQ plot demonstrate an approximation of normal distribution from the residuals, but not fully behaves as a gaussian distribution mainly due to how the tails of the residuals are, where is clear a slight deviation from the data. To further investigate the normality of the residuals, the Shapiro-Wilk Normality Test was applied, from which the null hypothesis was not rejected, reinforcing the normality thesis. Lastly the Jarque–Bera Test was also used to test the normality, that once again leads to the non-rejection of the null hypothesis of the distribution behaving as a normal distribution.

After all the test and analysis, the selected model proved to be satisfactory for the time series, and therefore were the model used to forecast the observations later in this project. The $\text{SARIMA}\ (3,0,2)\text{x}(\mathbf{2},1,3)_{12}$ could be written as:

$$\left(1-0.1055B-0.5534B^2-0.3396B^3\right)\left(1-0.8755B^{12}+0.9987B^{2*12}\right)\left(1-B^{12}\right)X_t$$
$$=\left(1+0.0534B-0.4045B^2\right)\left(1-1.6329B+1.6031B^2-0.7212B^3\right)e_t$$

3 FORECATING

3.1 EXPONENTIAL SMOOTHING

The method was used to forecast the 12 months after the last observation of the time series and were then compared to the real data to understand how accurate the results were.

The forecast by the exponential is shown in the table below with the real observations for each of the months predicted.

Table 2 – Forecast values using the exponential smoothing method

Period	**Forecast : $\widehat{X_t}$**	**Real Observation: X_t**
Feb, 2018	10791.998	10415
Mar, 2018	12875.984	12683
Apr, 2018	11874.876	11919
May, 2018	13891.906	14138
Jun, 2018	15044.290	14583
Jul, 2018	11663.319	12640
Aug, 2018	13670.206	14257
Sep, 2018	12221.874	12396
Oct, 2018	13012.917	13914
Nov, 2018	13591.634	14174
Dec, 2018	14677.571	15504
Jan, 2019	9813.961	10718

Table 3 – Error's statistics from the forecasted observations

Mean Error (ME)	350.8719
Root Mean Square Error (RMSE)	607.9057
Mean Absolute Error (MAE)	522.7506
Mean Percentage Error (MPE)	2.616975
Mean Absolute Percentage Error (MAPE)	4.001068

The results forecasted were closed to the real values, with a mean of 350 million dollars from the real observations, and an overall mean percentage of error of only 2.6%

3.2 ARIMA MODEL

The same 12 new observations were forecasted using the ARIMA model selected for the time series, presented in the table below with the confidence intervals of 95%.

Table 4 - Forecast values using the ARIMA model

Period	Forescast : $\widehat{X_t}$	Real Observation: X_t	Low 95%	High 95%
Feb, 2018	12185.85	10415	10014.96	11115.64
Mar, 2018	11805.58	12683	11557.78	12848.79
Apr, 2018	14302.51	11919	11189.63	12456.17
May, 2018	14246.41	14138	13488.01	15167.26
Jun, 2018	12392.12	14583	13423.59	15120.76
Jul, 2018	13904.47	12640	11650.93	13181.48
Aug, 2018	12223.94	14257	13046.66	14819.92
Sep, 2018	13739.84	12396	11453.34	13047.52
Oct, 2018	13401.98	13914	12847.83	14695.14
Nov, 2018	14811.79	14174	12512.11	14356.53
Dec, 2018	10245	15504	13803.58	15895.26
Jan, 2019	12185.85	10718	9538.508	11004.98

Table 5 - Error's statistics from the forecasted observations

Mean Error (ME)	294.2364
Root Mean Square Error (RMSE)	404.4886
Mean Absolute Error (MAE)	344.2657
Mean Percentage Error (MPE)	2.175942
Mean Absolute Percentage Error (MAPE)	2.586976

The forecasted observations once again did not difference much from the real values, with a mean error of 294.23, a mean percentage of error of almost 2.2%.

4 FINAL CONSIDERATIONS

The three approaches to study and analyze the time series are just a few among all the available and competent methodologies that are suitable for that purpose. In every method apply is either required to shape the model to the time series, or the other way around, and mold the time series for the method selected, and sometimes are both required.

The decomposition method was not used to forecast future observations of the time series, but it played an important role in studying and gathering information that was later used in the other methodologies.

The application of the exponential smoothing and ARIMA model to forecasting, proved that both methods were efficient providing reasonably accurate results, although the ARIMA model outperformed the other method, achieving overall better scores for every statistic associated with the errors obtained from the forecasts. This outcome is expected due to the higher complexity of the ARIMA methods, and the higher adaptation from the model and the time series to each other.

There is still room for improvement and to further explore this time series, with the implementation of different methodologies, or the improvement of the methods used in this project through the realization of more statistically test and graphical analysis.