Forecasting an Unknown Dataset: A Time Series Analysis using Seasonal ARIMA and Holt-Winters' Seasonal Method

ECO374 Final Essay

Linhan Zhang 1004734353

August 23, 2021

Introduction

Time series is a series of data whose points are successively taken at equally spaced points in time. A fundamental purpose of time series analysis is to build models which can perform forecasting. Common applications of these methodologies include but are not limited to precipitation, stocks & financial securities, and a country's GDP. Essentially, any quantitative data that can be tracked over time can be analyzed within the framework of time series. This project aims to find a time series model to generate forecasts for the following 12 time periods of the provided dataset. Though the exact purpose of the dataset is unknown various techniques will be employed to better understand the data and determine the final model. In this analysis, Seasonal ARIMA, Holt-Winters' Methods, and Ljung Box Test were used to generate two models. Then, thorough comparisons were conducted to determine the highest quality model. Specifically, AIC and Root Mean Square Error (RMSE) were employed to quantitatively compare the strength of the two models. Ultimately, I forecasted 12 additional periods to the provided dataset.

Methodology

Data

Description of the data (Plots):

The data for this project is a time series dataset consisting of real monthly data which contains 180 observations. To gain a general understanding of the data, a plot of the data was created and illustrated in figure 1 below. Based on this plot, it can be determined that there is an upwards trend and seasonality. The seasonality can be observed through recurring patterns every 12 time periods (months in this case).

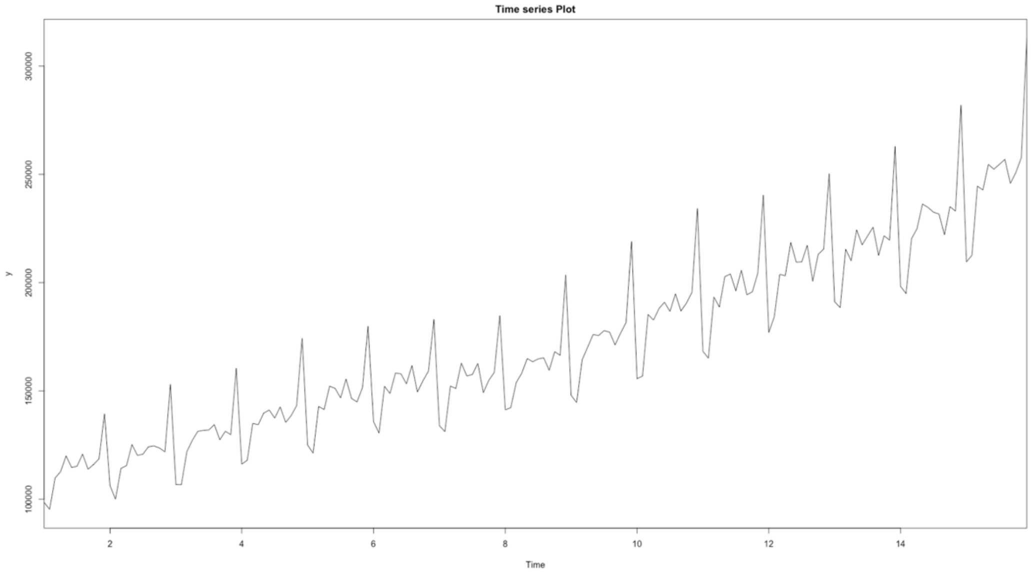

Figure1: Original Plot of the Time series data

Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots were created to determine stationarity. Based on the ACF and PACF plots (Figures 2), there is a strong correlation between data of consecutive time periods (nonstationary). That is, the data are correlated month-to-month (González-Rivera, 2016).

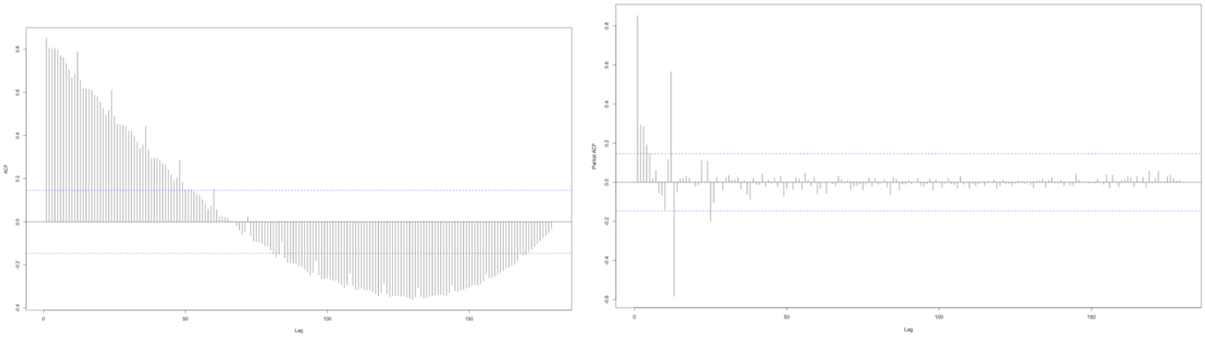

Figure2: ACF and PACF Plots of the Original Time Series

Models

Seasonal ARIMA model

Fundamentally, the ARMA(p, q) model is used for smooth time series; however, often datasets present non-stationary time series. As a result, the difference is taken to develop the ARIMA(p, d, q) model where d represents the order of difference taken by the original data. Seasonal ARIMA model is developed for a time series data when a seasonal variation is observed. It is more appropriate in predicting "long-term trend" and "seasonal effects" (Wang et al., 2013).

I used first order differencing operator $\nabla = (1 - B)$ and seasonal differencing operator $\nabla_{12}= (1 - B^{12})$ where B indicates the delay operator with purpose of shifting the data back for one period (Wang et al., 2013). The Akaike information criterion is an estimator of prediction error and is therefore useful in evaluating the relative quality of models in statistics (González-Rivera, 2016). Essentially, the AIC estimates the strength of a model relative to other models and is thus used as a tool for model selection. Through the comparison of various seasonal ARIMA models, the model with the lowest AIC was chosen. This seasonal ARIMA is written as ARIMA(1,1,2)($0{,}1{,}1)_{12}$ which contains the seasonal part is ARIMA(0,1,1) and the non-seasonal part is ARIMA(1,1,2). According to the table of coefficients (Table1), AIC value for this model is 3215.46 and the Root-Mean Square error (RMSE), a normalized distance between the actual value and predicted value, based on the observed data is 3385.019.

COEFFICIENTS FOR ARIMA(1,1,2)(0,1,2)$_{\mathbf{12}}$

AR1 ($\emptyset$)	0.8097
MA1 ($\boldsymbol{\theta_1}$)	-1.7599
MA2 ($\boldsymbol{\theta_2}$)	0.8701
SMA1 ($\boldsymbol{\Phi}$)	-0.4661
AIC	3215.46
RMSE	3385.019

Table1: Coefficients for ARIMA(1,1,2)(0,1,2)$_{\mathbf{12}}$

Autocorrelation and Stationary Test for Estimated Residuals

Ljung Box Test is used to verify for the absence of serial autocorrelation. Basically, this test examines whether errors are white noise or whether there is more to them. The null hypothesis of this test is that the model does not show lack of fit (the estimated error is stationary). In contrast, the alternative hypothesis is that the model does show lack of fit (the estimated error is non-stationary) (González-Rivera, 2016). According to the result from Ljung Box Test on the estimated error of ARIMA(1,1,2)(0,1,2)$_{12}$ model, the p value is 0.6126 which is larger than the critical value (0.05) (Table2). Therefore, there is not enough information to reject the null hypothesis which means that the model does not show lack of fit.

X-squared	**df**	**P value**
0.2564	1	0.6126

Table2: Ljung Box Test Statistics for ARIMA(1,1,2)(0,1,2)$_{12}$

Holt-Winters' Seasonal Method

The Holt-Winters' seasonal method is made up of the forecast equation and three smoothing equations. Of the three smoothing equations, one is for level l_t, one for trend b_t, and one for seasonal component s_t with the following parameters: α, β and γ. The *m* is used to indicate the frequency of the seasonality. In this case, *m* = 12 since the period is in 12-month intervals (Hyndman & Athanasopoulos, 2021). Here, I used the multiplicative method since the amplitude of the season of the data are increasing over time. The smoothing parameters with this data are: $\alpha = 0.1373,\ \beta = 0.0439,\ \gamma = 0.4785$ which are estimated by minimising RMSE. So, the component form for the multiplicative method is:

$$\hat{y}_{t+h|t=(l_t+hb_t)s_{t+h-m(k+1)}}$$

$$l_t = 0.1373 * \frac{y_t}{s_{t-m}} + 0.8627 * (l_{t-1} + b_{t-1})$$

$$b_t = 0.0439 * (l_t - l_{t-1}) + 0.9561 * b_{t-1}$$

$$s_t = 0.4785 * \frac{y_t}{(l_{t-1} + b_{t-1})} + 0.5215 * s_{t-12}$$

AIC value in this method is 3922.07, and RMSE is 3654.515 (Table3).

AIC	**3840.441**
RMSE	2835.828

Table3: Coefficients for model generate using Holt-Winters' Method

Autocorrelation and Stationary test for the Holt-Winters' Method

Here, I used the Ljung Box Test to determine if the estimated residuals generating from the Holt-Winter's Multiplicative Exponential Smoothing Method is stationary. The table of coefficients shows that the p value for null hypothesis is 0.9053 which is larger than the critical value (0.05) indicating that the residual is stationary (Table4).

X-squared	df	P value
0.014158	1	0.9053

Table4: Ljung Box Test Statistics for model generate using Holt-Winters' Method

Forecasting

Through the comparison of the two models, I decided to use ARIMA(1,1,2)(0,1,2)$_{12}$ to generate the final time series forecast. ARIMA(1,1,2)(0,1,2)$_{12}$ was determined to be the ideal model because it has smaller AIC values indicating better model quality and smaller RMSE value indicating reduced errors.

Based on the coefficients table, the forecasting model formular is (Table1):

$$(1 - 0.8097B)(1 - B)(1 - B^{12})\hat{y_t} = (1 - 0.7599B + 0.8701B)(1 - 0.4661B^{12})\hat{\varepsilon_t}$$

The Table below shows the 12 period forecast by applying ARIMA(1,1,2)(0,1,2)$_{12}$ model (Table5).

T	POINT FORECAST	LOW 80	HIGH 80	LOW 95	HIGH 95
181	235403.7	230900.1	239907.3	228516.1	242291.4
182	237353.6	232844.4	241862.8	230457.4	244249.8
183	267053.8	262494.0	271613.6	260080.1	274027.5
184	267285.8	262607.7	271963.9	260131.3	274440.3
185	280003.2	275136.3	284870.1	272560.0	287446.5
186	277692.3	272574.4	282810.2	269865.2	285519.5
187	278994.3	273574.4	284412.3	270708.2	287280.4
188	281557.0	275803.4	287310.5	272757.7	290356.2
189	270684.1	264571.6	276796.6	261335.9	280032.3
190	278644.6	272159.2	285130.0	268726.1	288563.1

191	282206.8	275341.9	289071.8	271707.8	292705.9
192	332582.6	325336.5	339828.7	321500.7	343664.5

Table5: 12 Periods Forecasting Values

Figure 3 visualizes the model fitted values and the actual values. Also, the forecasting plot shows the forecasting value (Figure 4). Since the actual value is unknown, it is impossible to calculate the forecasting error.

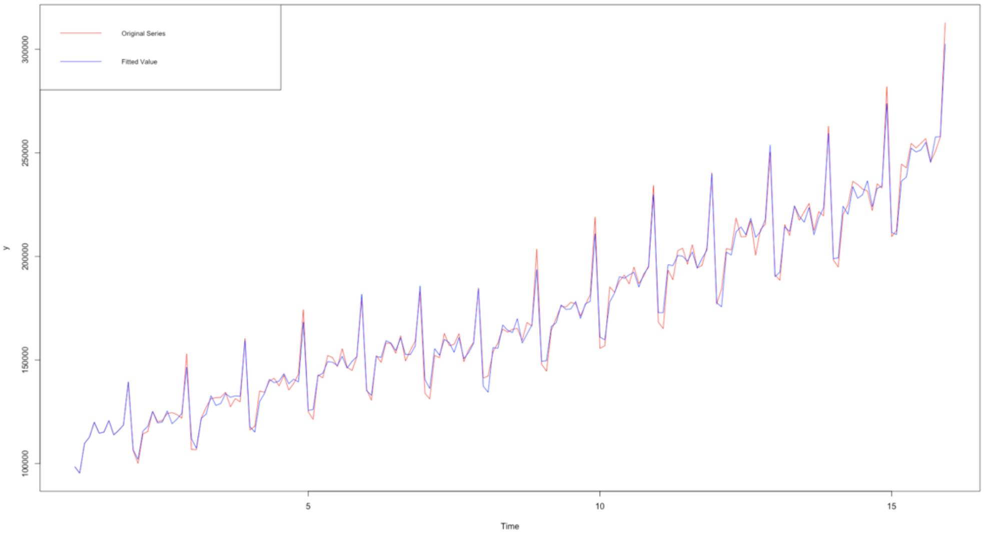

Figure3: Actual Values versus Fitted Values

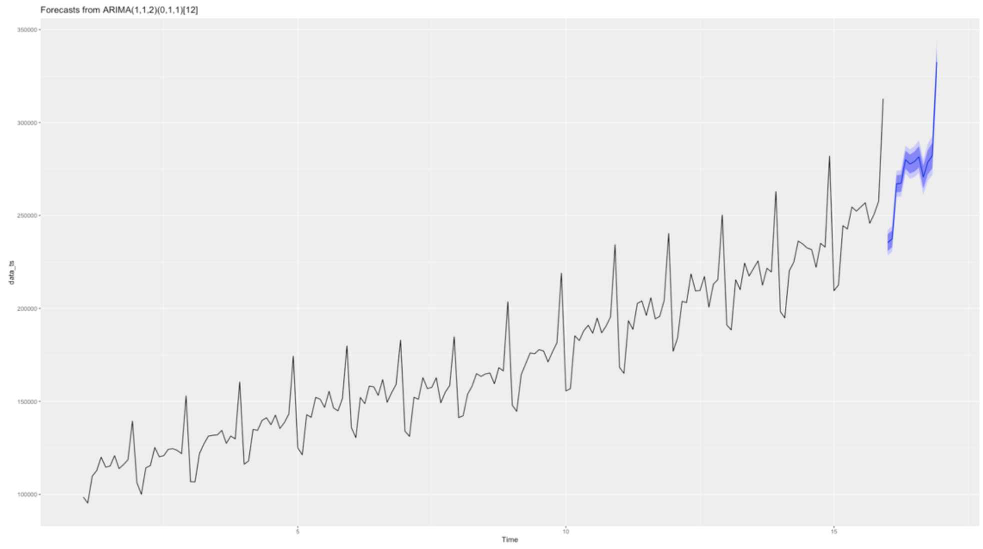

Figure4: 12 Periods Forecasts

Conclusion

In conclusion, this project began with an examination of the unknown data set. It was found that the time series was non-stationary and exhibited trend and strong seasonality. The ultimate purpose was to forecast an additional 12 time periods to the data set. To achieve this, I determined ARIMA(1,1,2)(0,1,2)$_{12}$ model and Holt Winters' Seasonal Method . Then the residuals from both models are stationary which means that they both good fit. Subsequently, AIC was employed along RMSE to determine the highest quality overall model between the two. To this end, the ARIMA(1,1,2)(0,1,2)$_{12}$ model was selected over Holt Winters' Seasonal Method because it had the smallest AIC and RMSE. The forecast was then generated using this Seasonal ARIMA model and presented in the results section. The figures of the projection were also displayed and portrayed a decent fit to the dataset's trend and seasonality.

Reference

González-Rivera Gloria. (2016). *Forecasting for economics and business*. Routledge.

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice*. OTexts.

Wang, S., Feng, J., & Liu, G. (2013). Application of seasonal time series model in the precipitation forecast. *Mathematical and Computer Modelling*, *58*(3-4), 677–683. https://doi.org/10.1016/j.mcm.2011.10.034