Sean Toon

IT-460-18500-M01 Machine Learning 2024 C-5 (Sep - Oct)

10/17/2024

Milestone Three – Draft of Model Execution

After studying more training models and gathering feedback from my professor and peers, I have decided to use the ARIMA (Autoregressive Integrated Moving Average) model as the best model for predicting sunspots. First, I had to download and load the “forecast” and “tseries” packages, and then of course load the data into an object name “sunspot\_data” using the following code:

A screenshot of a computer program

Description automatically generated

Next, I had to edit the column names of the data because they were not initially labeled, and I figured it would help with clarity further on in the process:

A white background with blue text

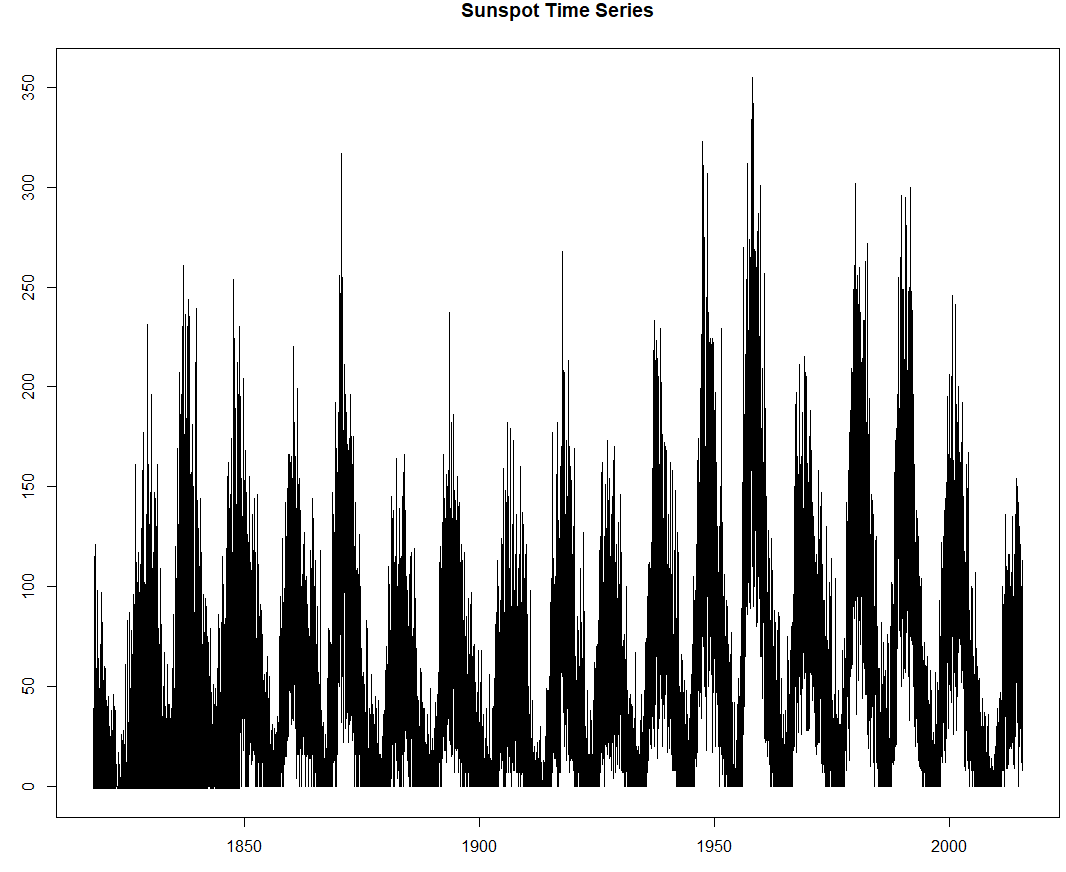
Description automatically generated

As shown, the data of the sunspots includes a year, month, day, fraction year, number of sunspots recorded on the day and a 1 or 0 to determine if the data for that day is definitive or provisional.

Next, I had to convert the data frame into a time series class for it to work for the ARIMA model. The parameters that I used in the conversion are the sunspot number column in my data set, the start of the first observation (in this case the first day recorded in the data) and the frequency of which I want to occur and because sunspots were recorded each day, I set the frequency to 365. Lastly, I created a plot to visualize and confirm that the time series data was created successfully.



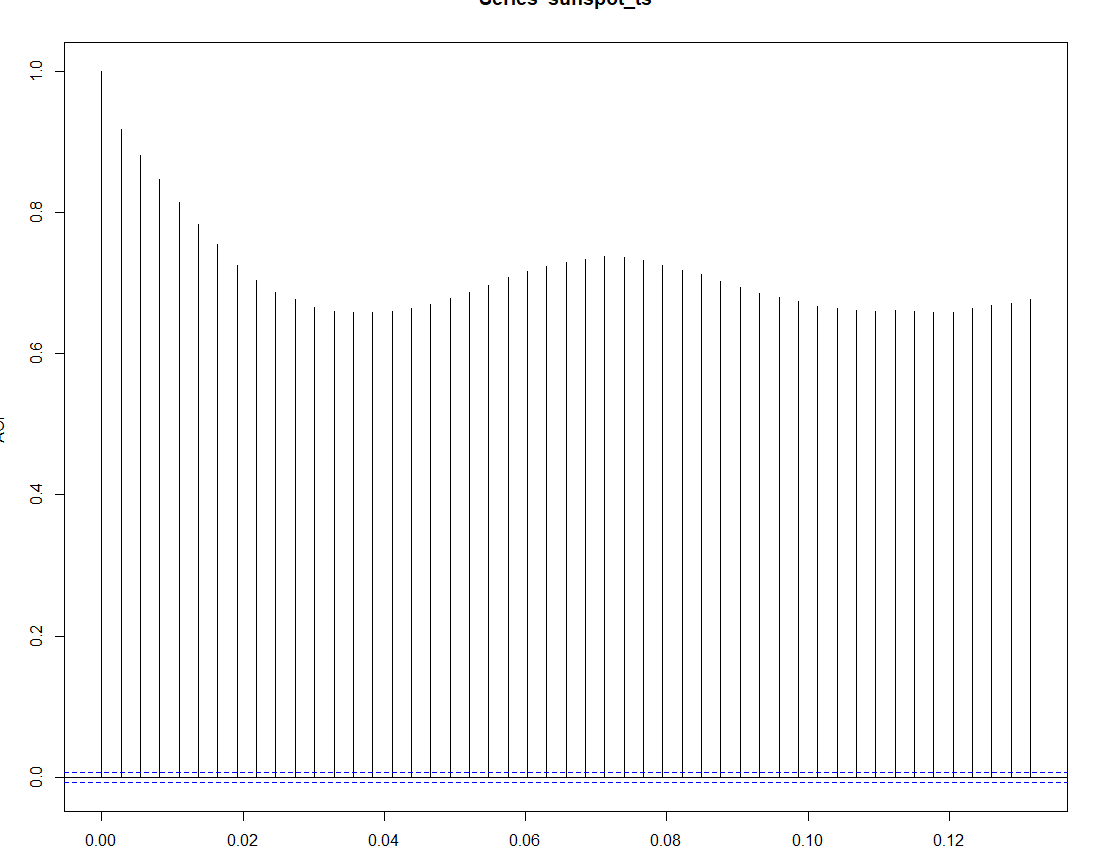
Plot generated:



ARIMA models work best when the data is stationary, so I used three different methods to check the if the data is stationary first the AutoCorrelation method with the following code:

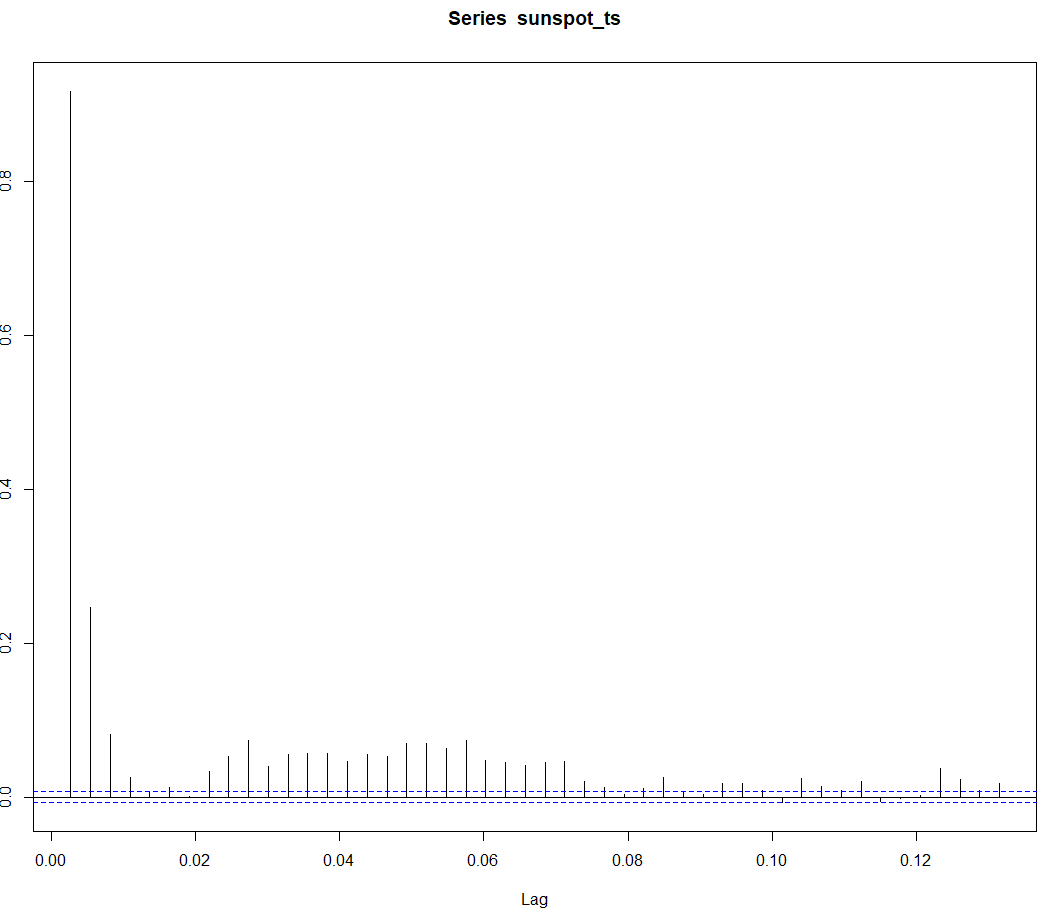


The plot results showed that the data is not stationary because values exist above the blue dotted line:

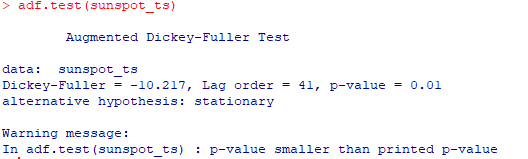


I also used the Partial AutoCorrelation Function to try and help and refine the view and there were still some indicators that the data was not stationary although it looked a little better:

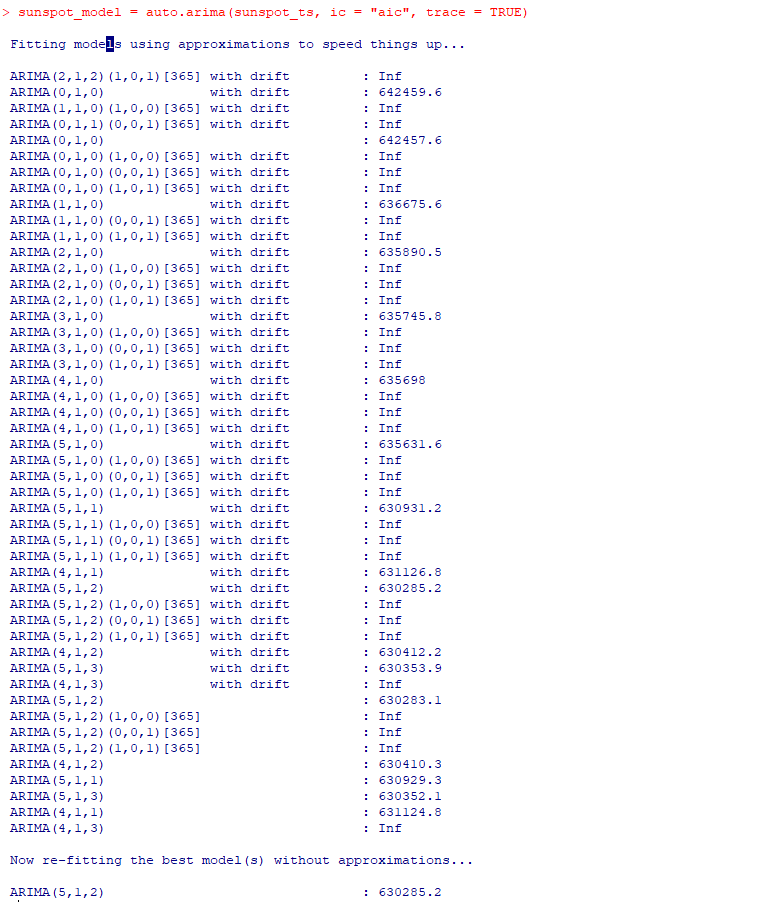




And finally I used the the Augmented Dickey-Fuller test to check with the following code, which then did suggest that the data is stationary because the p-value is smaller that 0.05:



The next step was to fit the model to the ARIMA model, to help choose the best model with the parameters (p,d,q) for the data. The results showed that the best model for my data is ARIMA (5,1,2) because that is the model that had the lowest AIC with a value of 630285.2:



Next, to further validate that the new best model “sunspot\_model” I checked the residuals to see that the data does a better job of staying within the blue dotted lines:



A graph of a graph

Description automatically generated with medium confidence



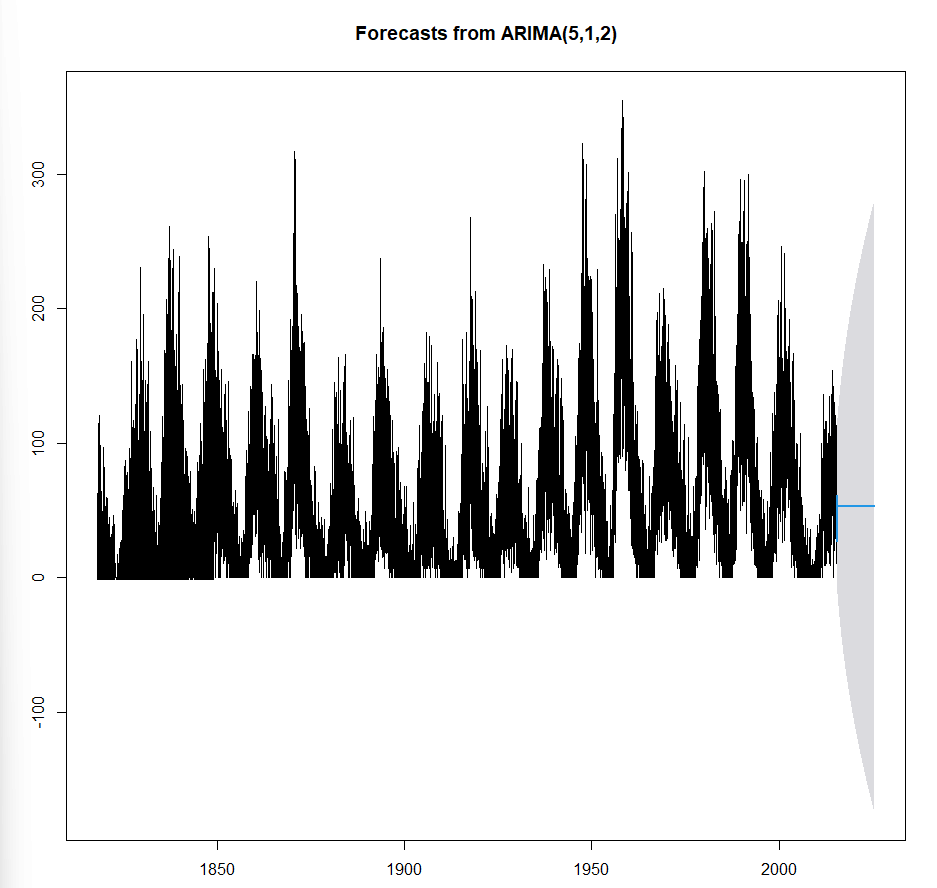
A graph with lines and numbers

Description automatically generated

After finding the best model prediction, I was able to continue with my forecast to help predict the sunspots. I did so by using the forecast function and the best model. The parameters that I used in my forecast model include the best fit model, the level of confidence that I want to achieve (95%) and the how far into the future I want to predict. I chose to forecast ten years to compare my forecasted data to the validation data provided to me in my project guidelines:

Unfortunately, my results did not turn out great. The forecasting plot that I was able to generate showed that prediction intervals increased in range over the next ten years, but the number of actual sunspot predictions stayed stable and were increasingly unpredictable with predictions into the future. With the prediction interval generated, I’m led to believe there is a better model that can be used to predict sunspots.

The graph below shows the historical sunspot data with the black lines, and the forecasted values in blue. The forecast uncertainty is the gray shaded area. When compared to sunspot validation data retrieved from the SIDAC website, I can see that from 2015 to current day the sunspot observations align with the upper bound of the prediction interval of the forecast data quite well.



I tried to improve my model performance in a couple of ways by changing the ARIMA parameters and adding a seasonal component to the function as well. Before changing the values, I had already suspected that the model performance would not change, considering that the auto.arima() automatically chooses the best model. Changing the p value (lag terms) led to overfitting, and lowering the value didn’t improve the forecast accuracy.

After using the ARIMA model to predict sunspot data, I cannot say that it would be a good model to use for organizations (or at least the model I created). The range of unpredictability is too great, especially considering the impact that solar weather can have on humans and earth. Although, I think the model may be a good start, and useful to organizations in the way that it shows the histories and pattern cycles of the sunspots as well as the correlation between past cycles and the predicted forecast.

References

SimpleSPSS (Producer), & . (2020, Mar 30,).*Time Series Analysis-ARIMA Model using R software : A step by step approach.*[Video] YouTube: