Sean Toon

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

10/20/2024

Final Project – Predicting Sunspots with ARIMA

The case and data I've chosen to analyze is the Sunspot data to help predict future sunspots. The CSV file contains records of observed sunspot data, including variables like year, month, date in decimal form, number of sunspots observed, and a mark that might represent significant sunspot activity. My goal is to create a model that helps identify any patterns or connections between earth dates, sunspots, and solar cycles.

The model I'm using to predict future sunspots is ARIMA (AutoRegressive Integrated Moving Average). ARIMA is a great fit for time series data like this because it looks at past values to make predictions about future occurrences. It captures trends and patterns, including the cycles of solar activity. One challenge is tuning the parameters (p, d, q) to make sure the model fits the data correctly, but with some fine-tuning, I expect to generate reliable forecasts. While ARIMA does have limits, like being sensitive to outliers, it's a strong tool for predicting sunspot activity when applied carefully.

A model like this could be useful for astrophysicists and anyone concerned with solar flares or other solar activity. It can help study how space weather affects technology, infrastructure, and human health. For example, the solar storm in May 2024 caused auroras, disrupted drones, delayed agricultural activities, and more (SIDAC, 2024).

I'll be using R to build my model. R is a great tool for data manipulation, modeling, and visualization because it offers a lot of helpful libraries. Here's the process I followed:

1. Collect the data

2. Explore and prepare the data

3. Look for correlations or patterns using graphs, plots, and matrices

4. Train the model

5. Evaluate its performance

6. Improve its performance

First, I downloaded and loaded the "forecast" and "tseries" packages, then loaded the data into an object named "sunspot_data." Next, I edited the column names because they weren't labeled, which I figured would help later.

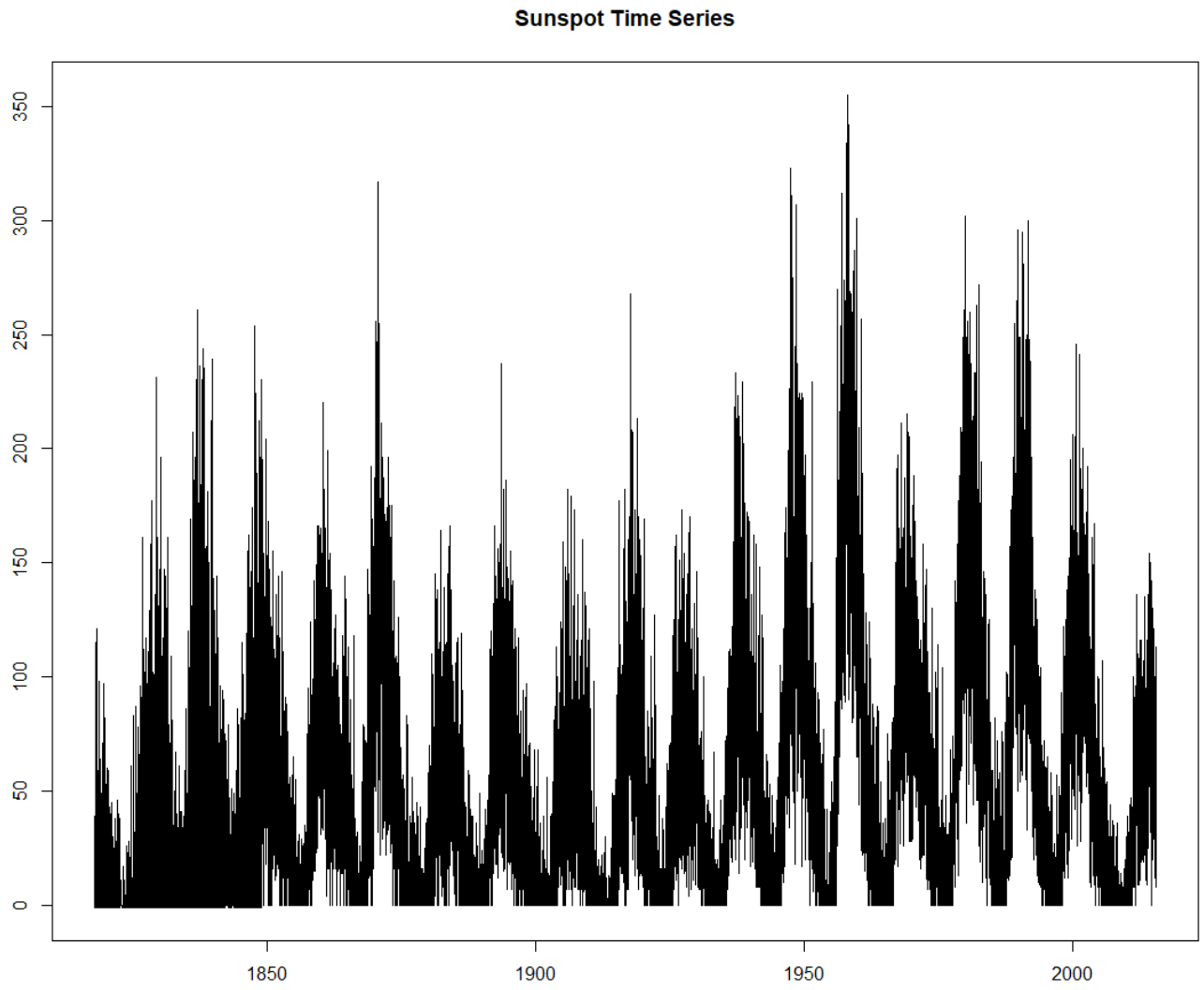
```
> library(forecast)
Registered S3 method overwritten by 'quantmod':
  method            from
  as.zoo.data.frame zoo
> library(tseries)

    'tseries' version: 0.10-58

    'tseries' is a package for time series analysis and computational finance.

    See 'library(help="tseries")' for details.

> sunspot data <- read.csv("C:\\Users\\toons\\Downloads\\ISSN D tot.csv")
```

```
> colnames(sunspot_data) <- c("Year", "Month", "Day", "FracYear", "SN", "OtherCol")
> summary(sunspot_data)
      Year          Month             Day          FracYear          SN           OtherCol
 Min.   :1818   Min.   : 1.000   Min.   : 1.00   Min.   :1818   Min.   : -1.00   Min.   :0.0000
 1st Qu.:1867   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:1867   1st Qu.: 10.00   1st Qu.:1.0000
 Median :1916   Median : 7.000   Median :16.00   Median :1917   Median : 37.00   Median :1.0000
 Mean   :1916   Mean   : 6.516   Mean   :15.73   Mean   :1917   Mean   : 51.31   Mean   :0.9979
 3rd Qu.:1966   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1966   3rd Qu.: 79.00   3rd Qu.:1.0000
 Max.   :2015   Max.   :12.000   Max.   :31.00   Max.   :2015   Max.   :355.00   Max.   :1.0000
```

I had to convert the data frame into a time series object for it to work with ARIMA. I used the sunspot number column, set the start date to the first recorded day, and used a frequency of 365 since sunspots were recorded daily. I then plotted the data to confirm that the time series was created successfully.

```
> sunspot_ts <- ts(sunspot_data$SN, start = c(sunspot_data$Year[1], sunspot_data$Month[1], sunspot_data$Day[1]), frequency = 365)
> plot(sunspot_ts, main="Sunspot Time Series", ylab="Sunspot Number", xlab="Time")
```

Plot generated:

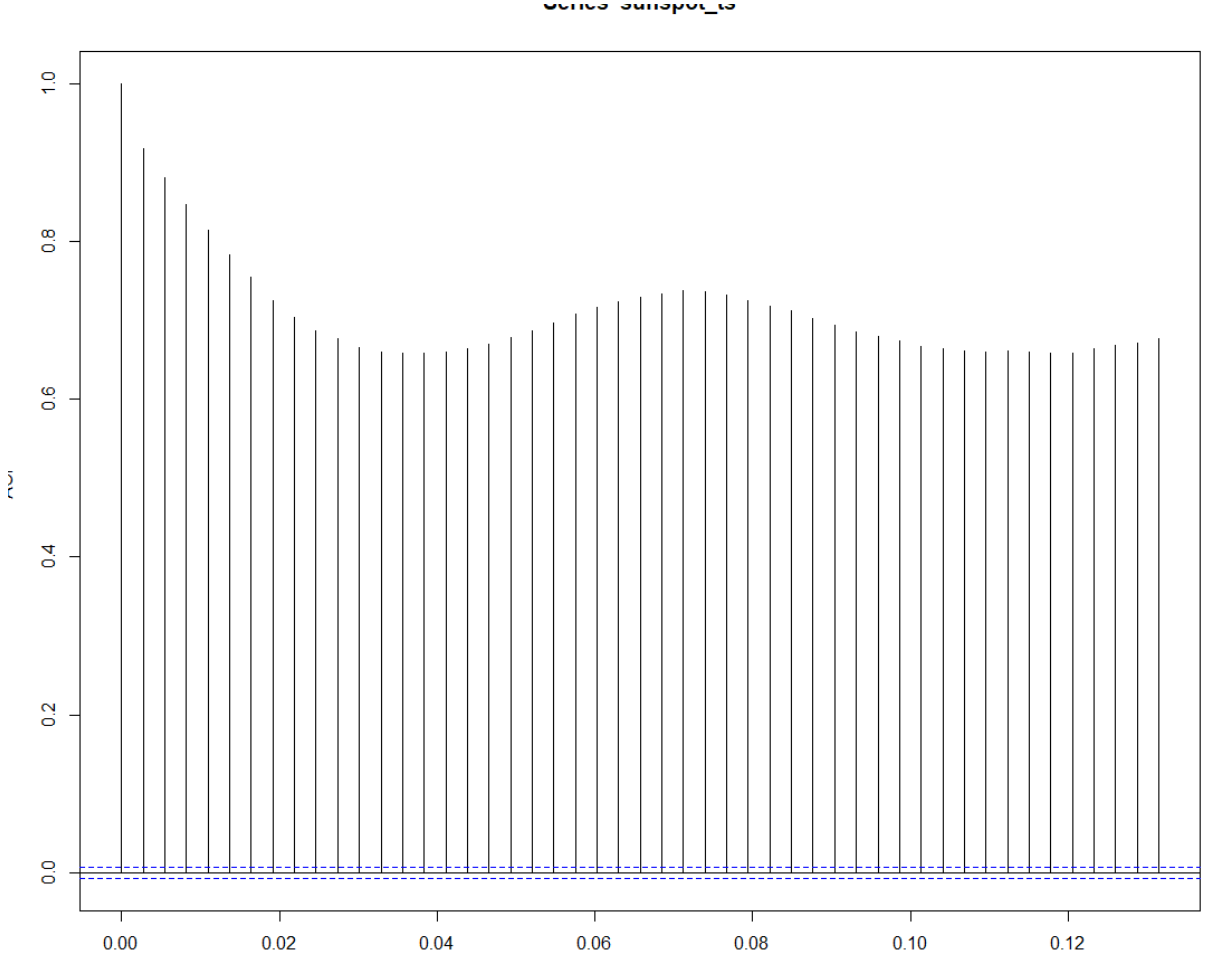


Next, I checked if the data was stationary, which is important for ARIMA models to work well. I used three different methods: AutoCorrelation, Partial AutoCorrelation, and the Augmented Dickey-Fuller (ADF) test. The autocorrelation plot showed that the data wasn't stationary because some values exceeded the blue dotted line. The partial autocorrelation plot gave a slightly better view, but it still suggested the data wasn't stationary. Finally, the ADF test confirmed that the data became stationary when the p-value dropped below 0.05.

AutoCorrelation:

```
> acf(sunspot_ts)
```

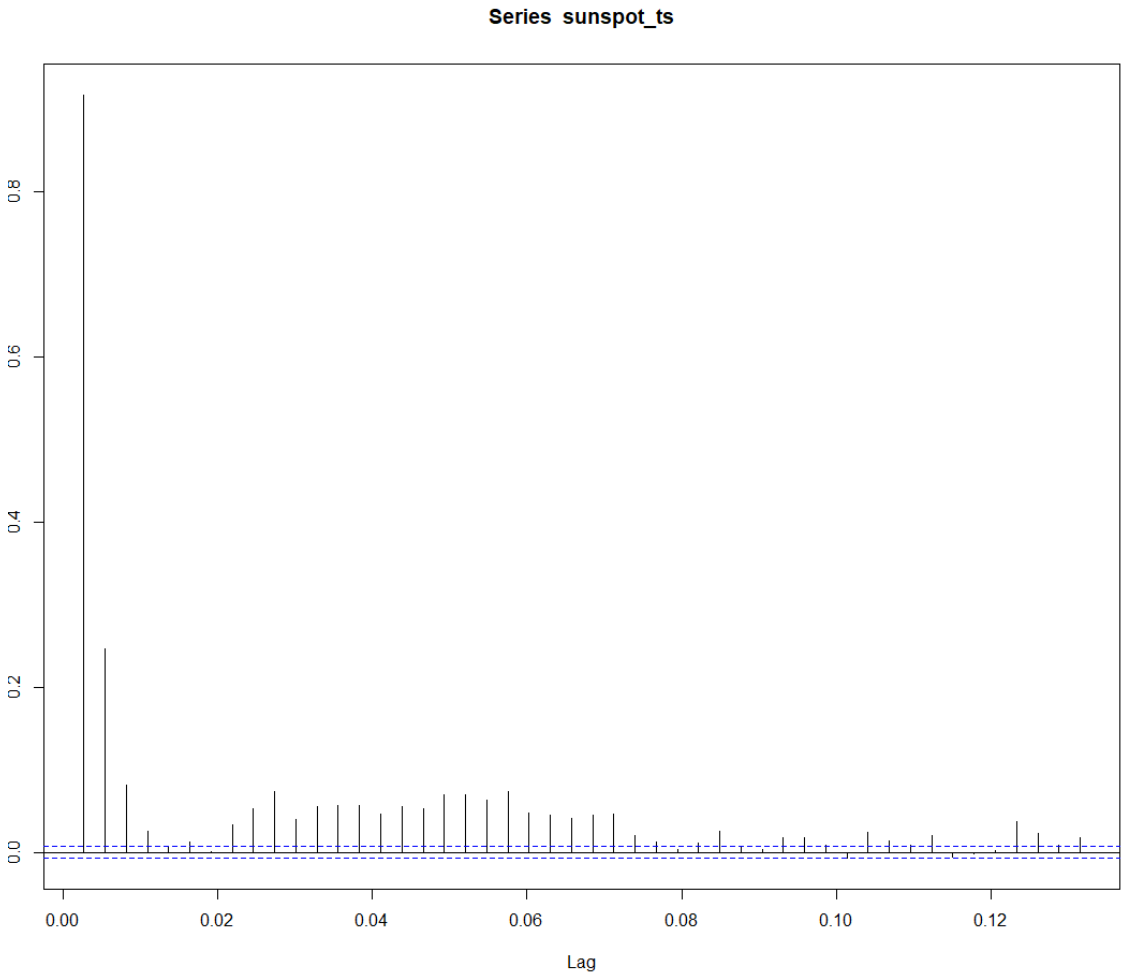
ACF Plot:



Partial AutoCorrelation Function:

```
> pacf(sunspot_ts)
```

Partial AutoCorrelation plot:

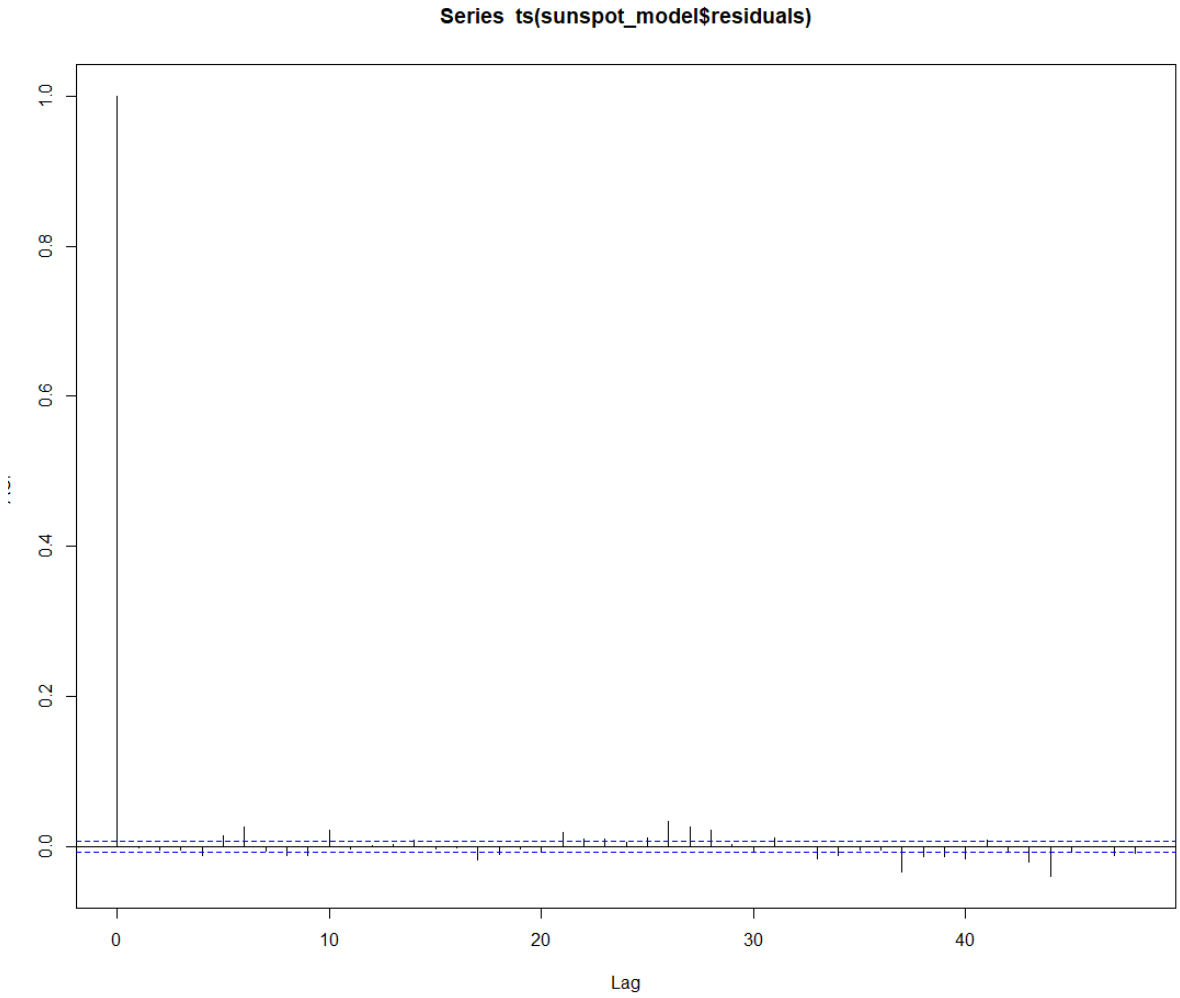


Augmented Dickey-Fuller test:

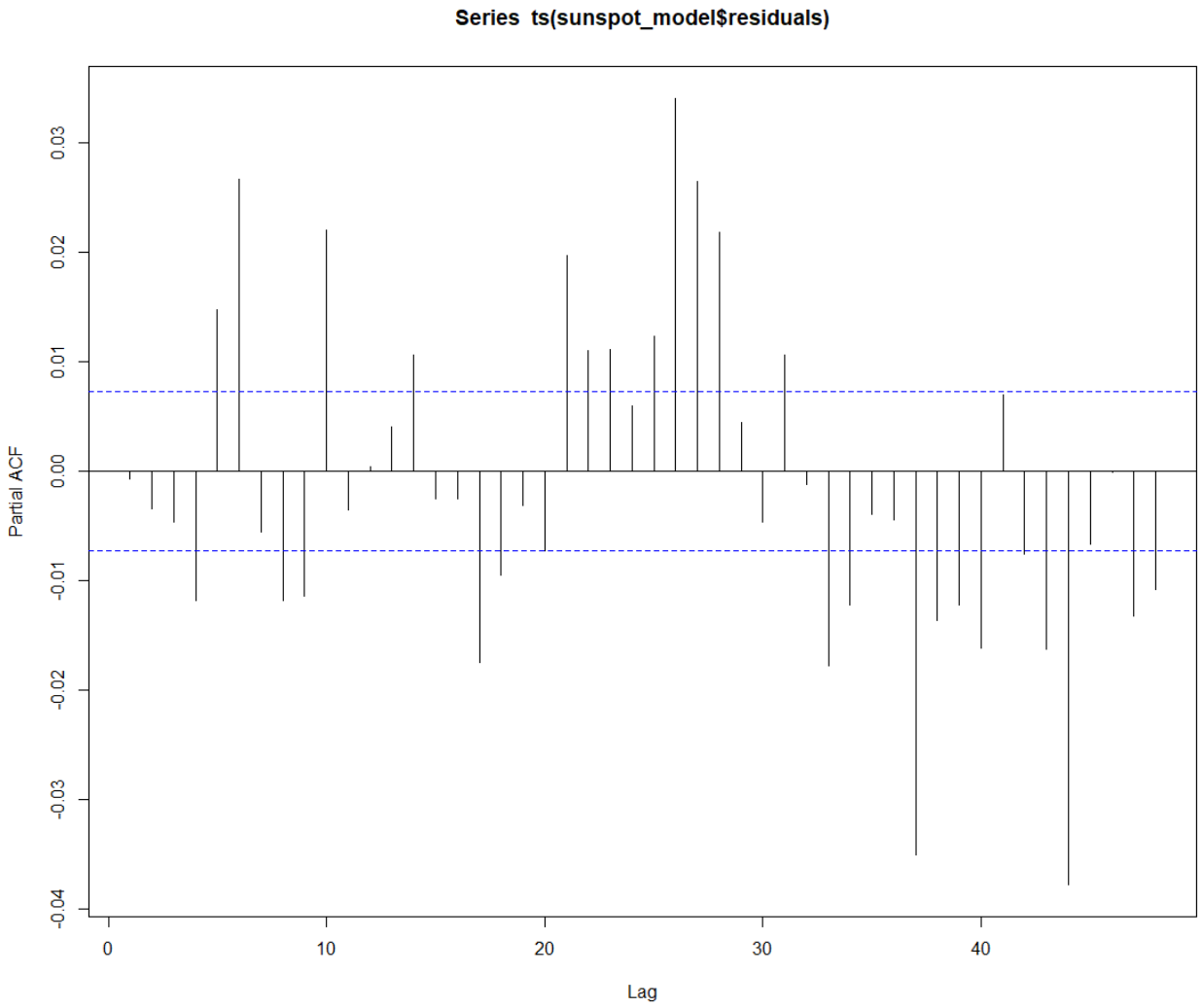
```
> adf.test(sunspot_ts)

        Augmented Dickey-Fuller Test

data:  sunspot_ts
Dickey-Fuller = -10.217, Lag order = 41, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(sunspot_ts) : p-value smaller than printed p-value
```

Once I confirmed stationarity, I fitted the data to an ARIMA model. The auto.arima() function selected the best model for my data, which was ARIMA (5,1,2) because it had the lowest AIC value (630285.2). I then checked the residuals to confirm that the data stayed within the blue dotted lines, further validating the model.

```
> sunspot_model = auto.arima(sunspot_ts, ic = "aic", trace = TRUE)

 Fitting models using approximations to speed things up...

 ARIMA(2,1,2)(1,0,1)[365] with drift         : Inf
 ARIMA(0,1,0)             with drift         : 642459.6
 ARIMA(1,1,0)(1,0,0)[365] with drift         : Inf
 ARIMA(0,1,1)(0,0,1)[365] with drift         : Inf
 ARIMA(0,1,0)                                : 642457.6
 ARIMA(0,1,0)(1,0,0)[365] with drift         : Inf
 ARIMA(0,1,0)(0,0,1)[365] with drift         : Inf
 ARIMA(0,1,0)(1,0,1)[365] with drift         : Inf
 ARIMA(1,1,0)             with drift         : 636675.6
 ARIMA(1,1,0)(0,0,1)[365] with drift         : Inf
 ARIMA(1,1,0)(1,0,1)[365] with drift         : Inf
 ARIMA(2,1,0)             with drift         : 635890.5
 ARIMA(2,1,0)(1,0,0)[365] with drift         : Inf
 ARIMA(2,1,0)(0,0,1)[365] with drift         : Inf
 ARIMA(2,1,0)(1,0,1)[365] with drift         : Inf
 ARIMA(3,1,0)             with drift         : 635745.8
 ARIMA(3,1,0)(1,0,0)[365] with drift         : Inf
 ARIMA(3,1,0)(0,0,1)[365] with drift         : Inf
 ARIMA(3,1,0)(1,0,1)[365] with drift         : Inf
 ARIMA(4,1,0)             with drift         : 635698
 ARIMA(4,1,0)(1,0,0)[365] with drift         : Inf
 ARIMA(4,1,0)(0,0,1)[365] with drift         : Inf
 ARIMA(4,1,0)(1,0,1)[365] with drift         : Inf
 ARIMA(5,1,0)             with drift         : 635631.6
 ARIMA(5,1,0)(1,0,0)[365] with drift         : Inf
 ARIMA(5,1,0)(0,0,1)[365] with drift         : Inf
 ARIMA(5,1,0)(1,0,1)[365] with drift         : Inf
 ARIMA(5,1,1)             with drift         : 630931.2
 ARIMA(5,1,1)(1,0,0)[365] with drift         : Inf
 ARIMA(5,1,1)(0,0,1)[365] with drift         : Inf
 ARIMA(5,1,1)(1,0,1)[365] with drift         : Inf
 ARIMA(4,1,1)             with drift         : 631126.8
 ARIMA(5,1,2)             with drift         : 630285.2
 ARIMA(5,1,2)(1,0,0)[365] with drift         : Inf
 ARIMA(5,1,2)(0,0,1)[365] with drift         : Inf
 ARIMA(5,1,2)(1,0,1)[365] with drift         : Inf
 ARIMA(4,1,2)             with drift         : 630412.2
 ARIMA(5,1,3)             with drift         : 630353.9
 ARIMA(4,1,3)             with drift         : Inf
 ARIMA(5,1,2)                                : 630283.1
 ARIMA(5,1,2)(1,0,0)[365]                    : Inf
 ARIMA(5,1,2)(0,0,1)[365]                    : Inf
 ARIMA(5,1,2)(1,0,1)[365]                    : Inf
 ARIMA(4,1,2)                                : 630410.3
 ARIMA(5,1,1)                                : 630929.3
 ARIMA(5,1,3)                                : 630352.1
 ARIMA(4,1,1)                                : 631124.8
 ARIMA(4,1,3)                                : Inf

 Now re-fitting the best model(s) without approximations...

 ARIMA(5,1,2)                                : 630285.2


> acf(ts(sunspot model$residuals))
```



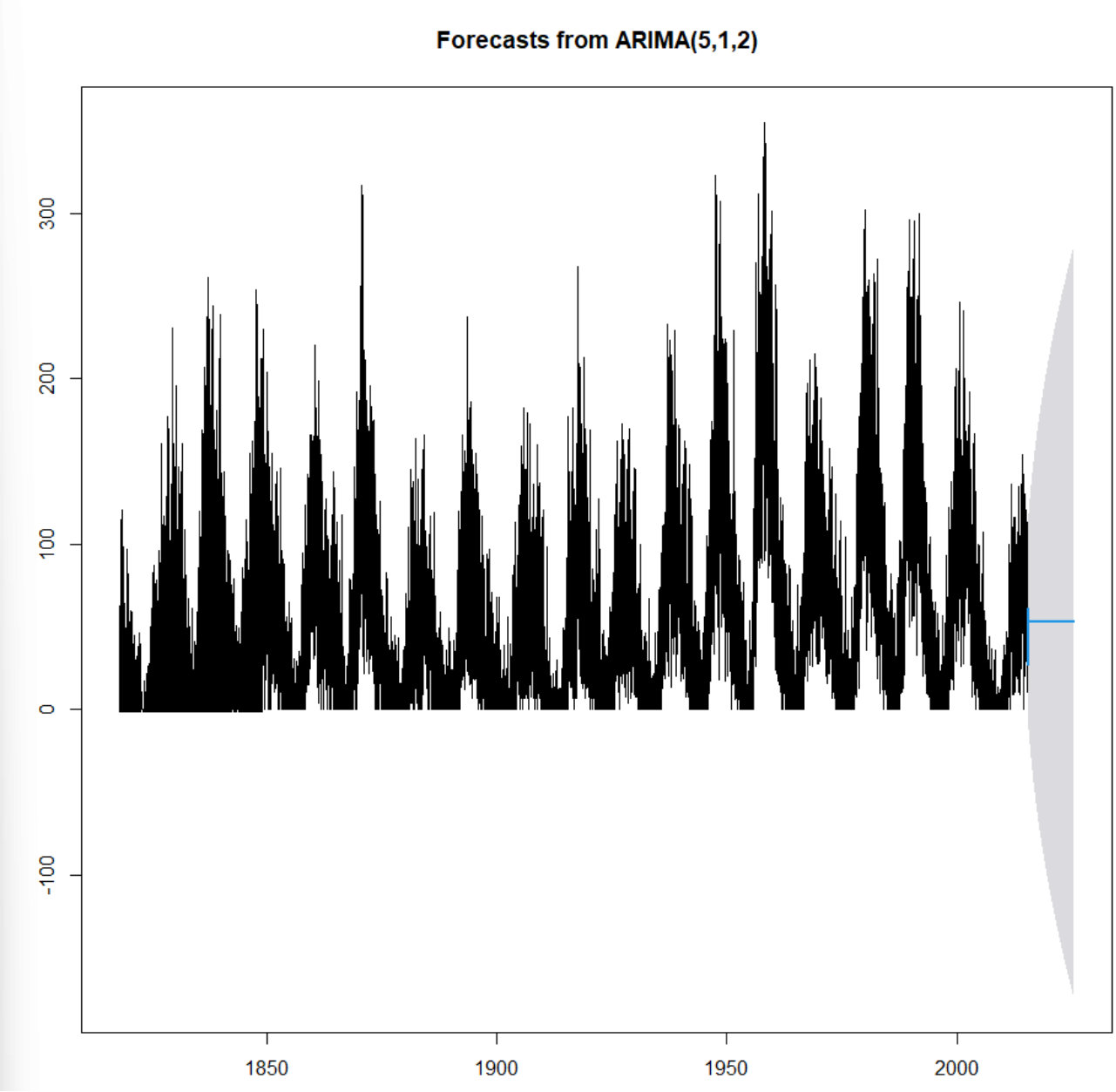

```
> pacf(ts(sunspot_model$residuals))
```



With the best model selected, I used the forecast function to predict sunspot activity. I forecasted ten years ahead, using a 95% confidence level, to compare the forecast to the validation data provided. Unfortunately, the results weren't ideal. While the prediction intervals expanded over the ten-year period, the actual sunspot predictions stayed stable and became increasingly unreliable as they extended into the future. This led me to think that there might be a better model for predicting sunspots.

```
> sunspot_forecast = forecast(sunspot_model, level = c(95), h = 10 * 365)
```

The graph I generated shows the historical sunspot data in black and the forecasted values in blue. The gray-shaded area represents the forecast uncertainty. When I compared my forecast to sunspot validation data from the SIDAC website, I noticed that, from 2015 onward, the observed sunspots aligned well with the upper bound of the prediction interval.



I tried improving my model by tweaking the ARIMA parameters and adding a seasonal component. However, since the auto.arima() function already selects the best model, these changes didn't improve the forecast. Increasing the p-value led to overfitting, while lowering it didn't make much difference in forecast accuracy.

In conclusion, while ARIMA is a good start for predicting sunspots, the model I built might not be suitable for organizations needing precise forecasts, given the range of unpredictability in my results. Still, it could be helpful in showing the historical patterns of sunspots and the correlation between past cycles and future predictions.

References

Paul M. Schwartz.Data Protection Law and The Ethical Use of Analytics., 22 – 27. http://web.archive.org/web/20150218103912/http://www.huntonfiles.com/files/webupload/CIPL_Ethical_Undperinnings_of_Analytics_Paper.pdf

Sunspot Index and Long-term Solar Observations. https://www.sidc.be/SILSO/home

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

Lantz, B. (2019). *Machine Learning with R* (3rd ed.). Packt Publishing. https://mbsdirect.vitalsource.com/books/9781788291552