STAT 5550 Final Project

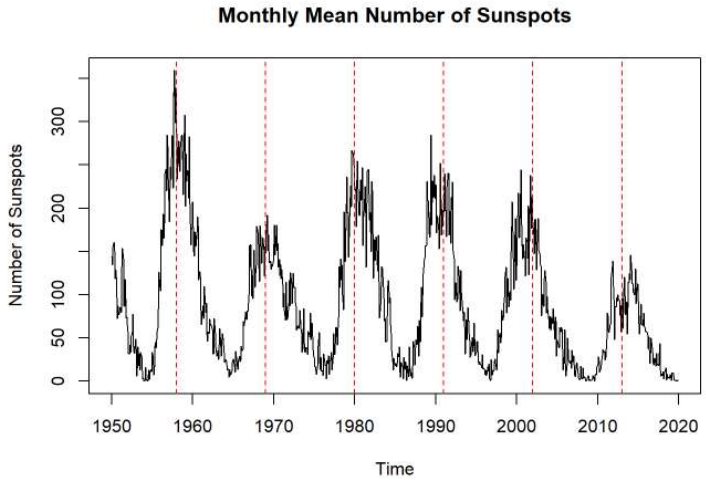
Sheldon Kappel

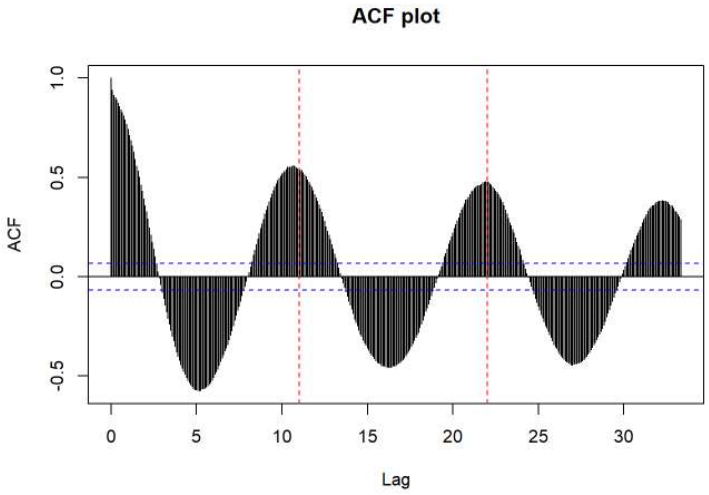
Introduction:

For this project, I chose to examine the fluctuation of the average total number of sunspots per month. Sunspots are areas of lower temperature on the surface of the Sun caused by undulations in the solar magnetic field. These magnetic oscillations create disturbances in the convection of the Sun's gaseous exterior, creating a zone that is on average one third cooler than the surrounding gases. This difference in temperature creates a dark discoloration in contrast to the rest of the Sun. The number of Sunspots waxes and wanes in accordance with an approximately eleven-year solar cycle, in which time the Sun's magnetic field flips polarity. In this project, I attempt to accurately model this phenomena, using data recorded by the Solar Influences Data Analysis Center, the solar physics research department of the Royal Observatory of Belgium.

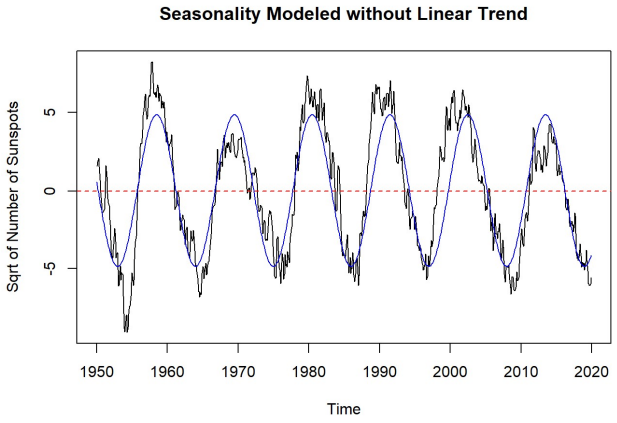
Models for Trend and Seasonality

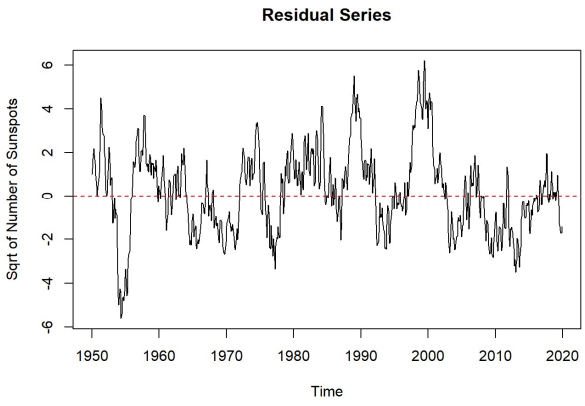
The original dataset began in January 1749 and ended in December 2019. For this assignment, I instead chose to start my examination at January 1950. Below is a plot of the data along with its ACF. The dashed red lines are placed at eleven-year intervals.



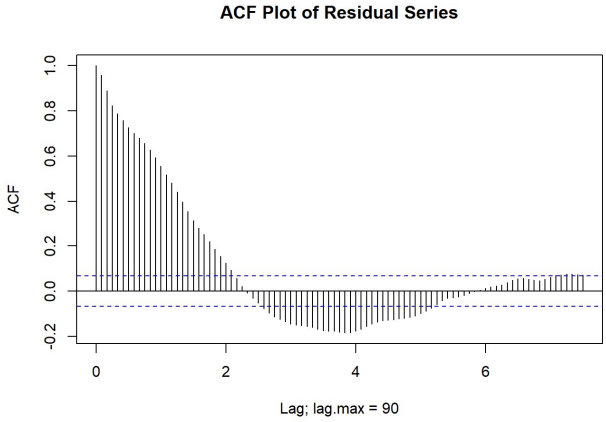


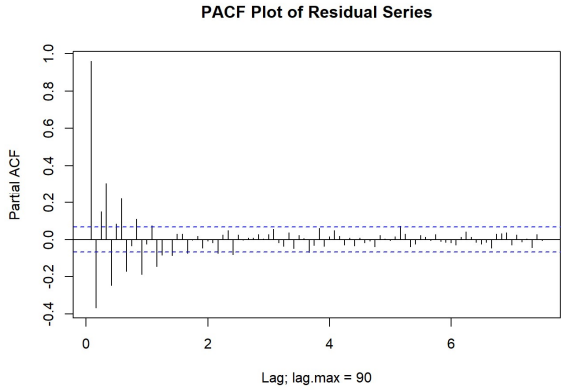
Looking at these plots, there obviously exists some non-stationarity, mainly due to the easily visible seasonality but also because of a slight downward trend in this subset of the data. There exists an additional issue—large differences in the maxima of the cycles, thus resulting in a non-constant variance. To help mitigate this, I performed a square root transformation of the data. A plot of the seasonality modeled on this transformation is shown below as well as the resulting series.

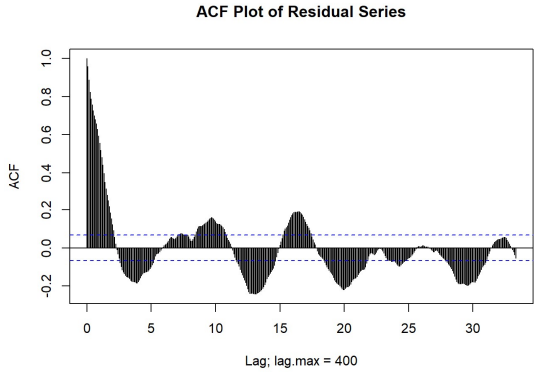


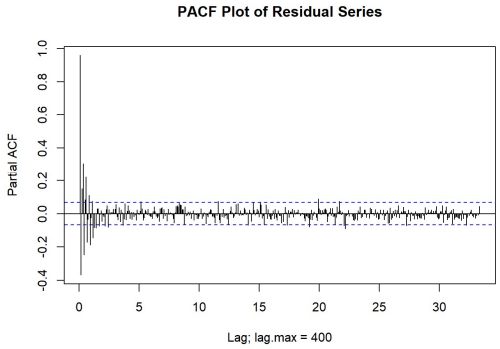


Examining the plots below, it is apparent that the ACF function never really "cuts off," thus removing the possibility of solely an MA model for the residual data. Looking at the PACF plot, I deduced that the pacf cuts off after lag 16 and tails off after lag 14. This led me to consider two models: AR(16) ann ARMA(14,14).



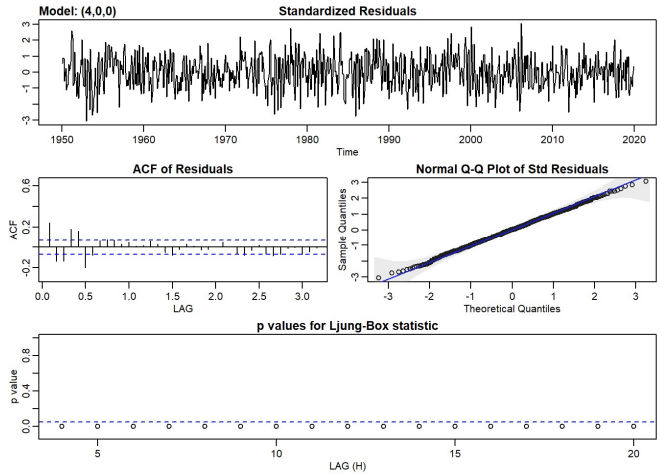






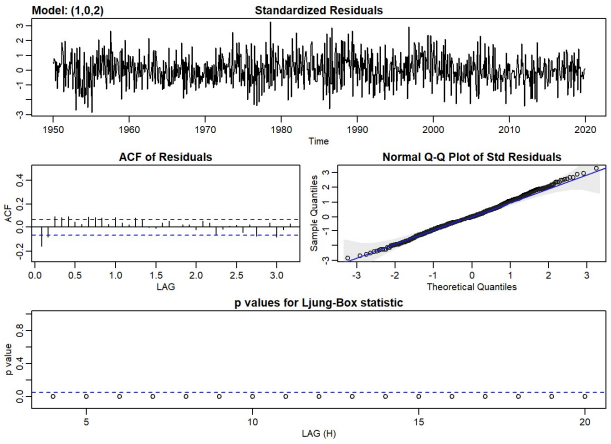
For the AR(16) model, I found that only $\hat{\phi}_1$, $\hat{\phi}_2$, and $\hat{\phi}_4$ were significant at the 95% level. I thus considered the AR(4) subset model where $\hat{\phi}_2 = 0$. This gives a residual series model of

$$y_t = 1.1317y_{t-1} - .5552y_{t-3} + .3892y_{t-4} + w_t, \quad \text{with} \quad w_t \sim iid\, N(0, .2379) \quad \text{and} \quad AIC_c = 1.415$$



For the ARMA(14,14) model, I found that only $\hat{\phi}_1$, $\hat{\theta}_1$, and $\hat{\theta}_2$ were significant at the 95% level. I thus considered the ARMA(1,2) model. This gives a residual series model of

$$y_t = .792y_{t-1} + w_t + .9991w_{t-1} + w_{t-2}, \quad \text{with} \quad w_t \sim iid\ N(0,.1809) \quad \text{and} \quad AIC_c = 1.156$$



Both models' residuals fit the Q-Q plot well. However, both models raise causes of concern since their p-values for the Ljung-Box statistic test lie below the threshold of significance and their ACF's do not immediately taper to zero, thus casting doubt to the claim that the residuals are white noise.

With this noted, I personally would go with the ARMA(1,2) model, because not only is the corrected AIC score lower for it, but also because its ACF values jut out less and decrease more smoothly at higher lags than the AR(4) model.

All together, the model I formed is given by,

$$sqrt(x_t) = 109.2597 - .0507t + 1.27sin(2\pi * t/11) + 4.685cos(2\pi * t/11) + y_t, \quad y_t = .792y_{t-1} + w_t + .9991w_{t-1} + w_{t-2} \quad \text{with} \quad w_t \sim iid\ N(0,.1809)$$

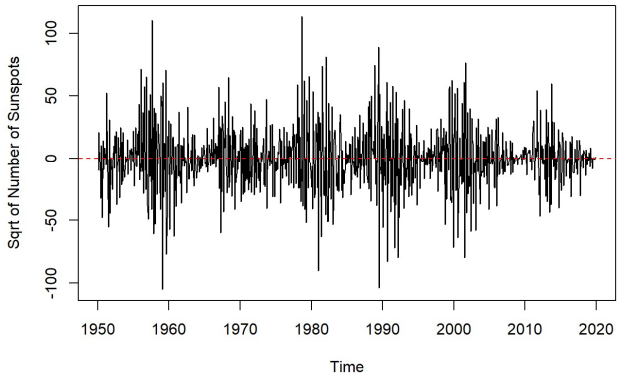
The original dataset extends all the way to December 2019. Currently it is April 2020, meaning January, February, and March can be forecasted and compared against the actual recorded average monthly sunspots for the past three months. These "actual" values were recorded by the same Belgian observatory.

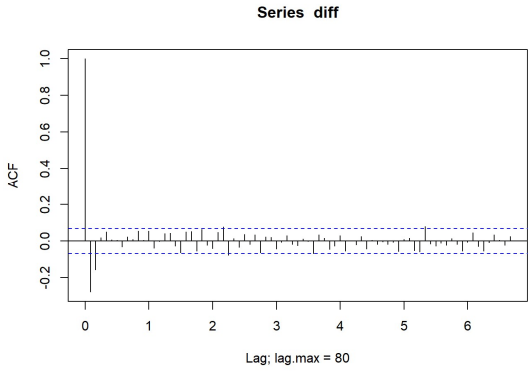
Month	Prediction	95% Prediction Interval	Actual
January	3.037	(.8236, 6.6445)	4.4
February	4.619	(.0947, 13.9292)	5.6
March	6.018	(.2114, 23.3448)	7.1

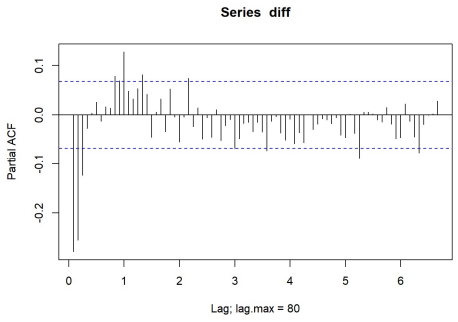
Although the model has a tendency to under-predict, all of the actual values are within the 95% prediction interval.

Sarima Modeling

A single seasonal difference is taken and seen to create a stationary series.

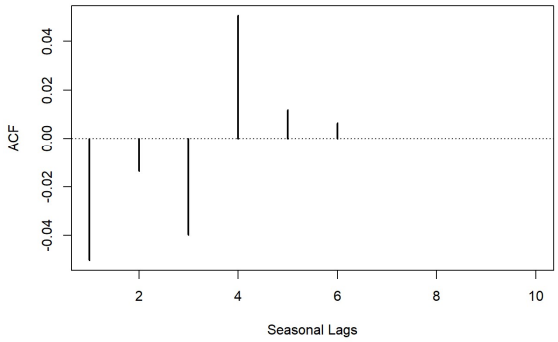


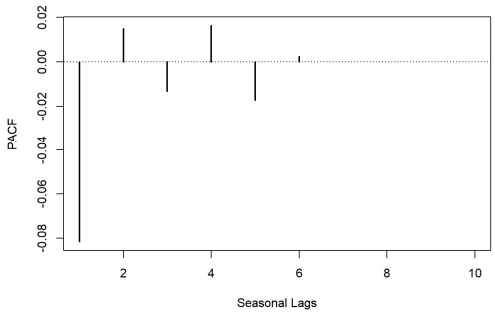




The ACF appeared to tail off after lag 1 and cut off after lag 2. The PACF could be said to tail off after lag 1 and lag 2. This suggested a model for the non-seasonal component in the ballpark of ARMA(2,2) or MA(2).

The ACF and PACF with just the seasonal lags are displayed below. (Note: I refer to a "season" as one of the eleven years comprising the solar cycle.)

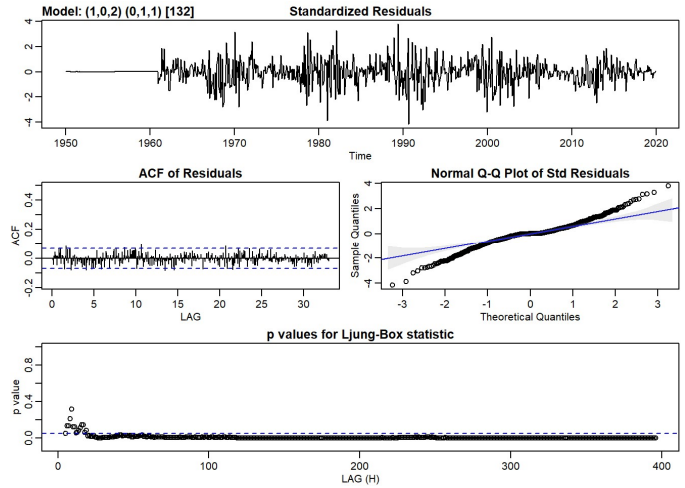




The PACF tails off after seasonal lag 1. One could say that the ACF tails off after seasonal lag 4, but this could also be because of the lack of data past the seasonal lag 6. A model of the ARMA(4,1) was attempted for the seasonal component.

I unfortunately had issues in R with the period being so large of 11 years * 12 months = 132 months. I was unable to bypass this issue, and so I had to simplify my model down just to get it to run without taking 10 minutes. I ultimately settled for an ARMA(1,0,2) x $(0,1,1)_{132}$ model given below.

$$(1-.9809B)x_t = (1-B^{132})(1-.4642B-.1712B^2)w_t \quad \text{with} \quad w_t \sim iid\ N(0,518.8)$$



Although the ACF of the residuals of this model are not too bad, the Q-Q plot and the p-value display a sharp contrast from the desired white noise requirements. This is evidence that the model is not adequate. To show this, a prediction table is provided.

Month	Prediction	95% Prediction Interval	Actual
January	.4376	(0, 48.659)	4.4
February	0	(0, 53.0645)	5.6
March	12.8016	(0, 69.4402)	7.1

Model Comparison / Conclusions

Although neither model produced desired Ljung-Box p-values, the first model vastly outshines the second. It fit the Q-Q plot well and the ACF of the residuals were sufficient. Also, all of the actual values for the number of sunspots in the past three months were close to the predicted value and well within the 95% prediction intervals generated by the first model. In contrast, the second model's diagnostics warranted much concern and the difference between the actual and predicted values were much greater. The second model also had all of the actual values well within the corresponding prediction intervals, but the predictions generated had large standard errors, thus producing extremely wide intervals.

The final model chosen to model the average number of sunspots per month from 1950 to 2019 is

$$sqrt(x_t) = 109.2597 - .0507t + 1.27sin(2\pi * t/11) + 4.685cos(2\pi * t/11) + y_t, \quad y_t = .792y_{t-1} + w_t + .9991w_{t-1} + w_{t-2} \quad with \quad w_t \sim \text{iid } N(0, .1809)$$