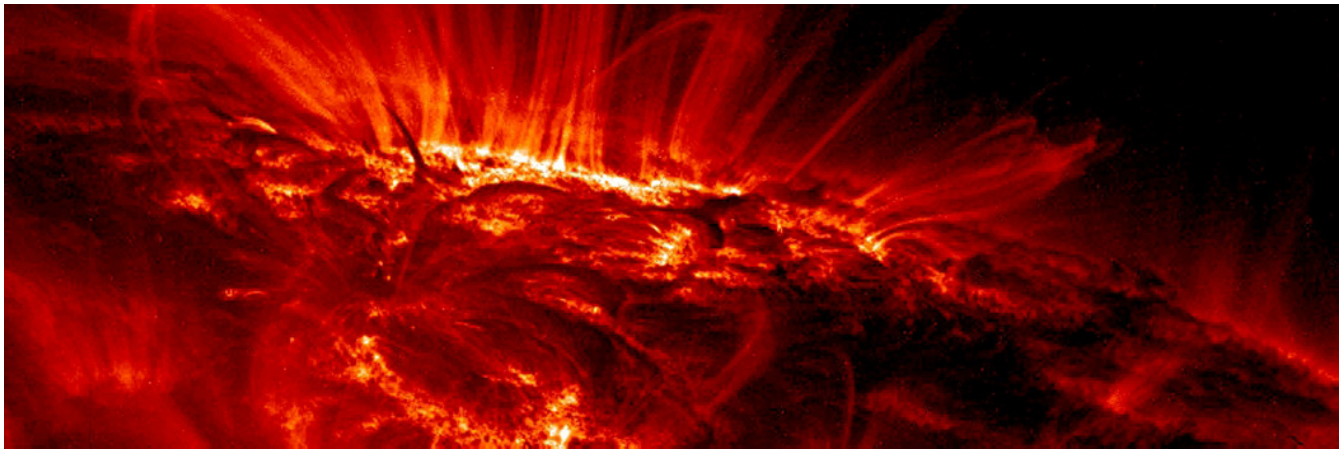
Solar Flare Time Series Analysis

Alex Haas, Areeb Khan Shabih, Ivan Gvozdanovic

03 May, 2022

Contents



Introduction

With the constant output of energy, and thus data, the Sun represents an ideal object of study for a data scientist. In particular, studying X-ray flux emitted by the Sun can prove to be beneficial to the prolonged safety of our civilization. Solar flares are large eruptions of electromagnetic radiation from the Sun lasting from minutes to hours. They usually take place in active regions, which are areas on the Sun marked by the presence of strong magnetic fields. These areas are typically associated with sunspot groups. As these magnetic fields evolve, they can reach a point of instability and release energy in a variety of forms. These include electromagnetic radiation, which are observed as solar flares.

Although, the Earths magnetic field protects us from solar flares, space-earth radio communication and electronic equipment on-board different space probes and satellites can be effected or completely destroyed. Therefore, in the this report, we set out to study the temporal structure of the the X-ray flux data, its predictive power as well as whether it is possible to predict solar flares.

Project Objectives

Given the X-ray flux time series, we want to understand these questions:

- Explore and deduce whether the X-ray flux data contains seasonal components.
- Analyze the structure of the data and explore the relationship between X-ray intensity build up and occurrence of solar flares.
- Finally, explore the predictive power of the data at hand and try predicting a future solar flare using a optimally trained model.

Data Preparation & Processing

Data Selection

We used the the NASA GOES Satellite data from 1999 to 2019. We downloaded the data from source. Then, we took advantage of a custom python script made to scrap data from the source based on the type and granularity of aggregation. The script allowed us to aggregate the data based on any level of aggregation i.e. hourly, daily, monthly, weekly etc. and provided us with the flexibility to take the maximum, minimum, average or any percentile of aggregated data.

There were a total of 12 CSV files per year and each file contained X-ray flux intensities recorded by GOES every 1 minute.

Data Aggregation Methodolgy:

Data aggregation was important because 20 years worth of minute wise data creates a lot of noise.R makes it practically impossible to do spectral analysis on minute wise data to explore underlying cyclical trends. We used hourly and daily aggregation to generate two different datasets for achieving different objectives. The idea was to pick top X-ray intensities in the aggregation window because it was in line with the Solar Flare analysis task. Solar Flares happen when X-Ray flux intensifies and breaches a certain threshold. It is however important to note that there can be blips for a very small period of time in the X-ray intensity too due to events other than Solar Flares. It is therefore of extreme importance to account for these other events and distinguish them from Solar Flares. Taking maximum over an aggregated window can result in getting such observations which were most likely driven by other factors as Solar Flares are rare events. The type of aggregation used was percentile based. We selected the percentile based on our definition Solar Flares in terms of X-Ray Intensities. How did we end up selecting the right percentile will be explained below but before that we need to define Solar Flares in terms of X-Ray intensities,

Whenever the X-ray Intensities breach a certain threshold for a minimum of ten minutes or more, we treat this event as one Solar Flare. The Solar Flare can last for few minutes (greater than ten), hours, days or even weeks depending on the solar activity.

The key to select percentile is to set it corresponding to ten minutes based on the aggregation window. As we have minute-wise data, for hourly aggregation, we chose the percentile which corresponded to the 10th highest observation i.e. in hourly data if we see the threshold being breached at the tenth highest observation then it implies that the threshold was breached for ten minutes or more and hence solar flare occurred. The corresponding percentile for hourly aggregation is given by,

$$Percentile = \frac{(60 - 10) * 100}{60} = 83.33$$

Thus 83.33 percentile corresponds to the Solar Flare definition for hourly aggregation.

Generalizing the formula for selection of percentile,

$$Percentile = \frac{(Minutes\ in\ Aggregation\ Window - Minimum\ Minutes\ Solar\ Flare) * 100}{Minutes\ in\ Aggregation\ Window}$$

Similarly, for daily data we input following parameters in the formula,

Minutes in Aggregation Window = 1440

Minimum Minutes Solar Flare = 10

$$Percentile = \frac{(1440 - 10) * 100}{1440} = 96.5$$

In addition, we can also convert the raw intensities into flare categories by using the below table (from Stanford Solar Center) to provide another perspective of the data. Finally, since all of the intensities in the dataset are smaller number $< 10^{-6}$, we take log transformation on the time series for numerical stability.

It is important to note here that the threshold mentioned above for X-Ray intensities to be characterized as Solar Flares is 10^{-5}.

Data Processing & Pre-liminary Time Series Analysis

After aggregation, the hourly and daily data is loaded.

Hourly Data

First loading the hourly data,

Populating Data for Missing Years

The data for 2007, 2008 and 2009 was missing. In order to impute this data, we took the mean and standard deviation of X-Ray intensities of surrounding years and generated a sequence of random numbers from the corresponding normal distribution. We generated 26280 random observations for 3 years missing hourly data.

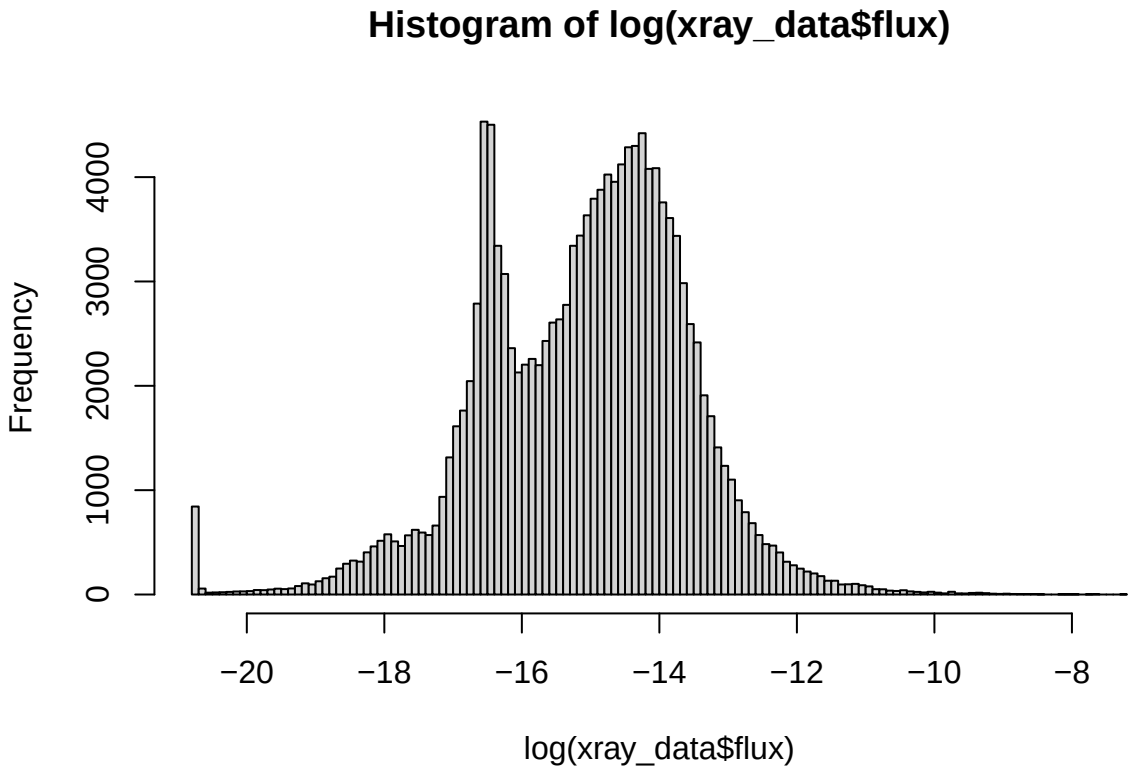
```
## The total number of observation in time series is: 141219
```

```
##
## The number of NAs in the dataset is: 0
```

Dataset Characterisitcs

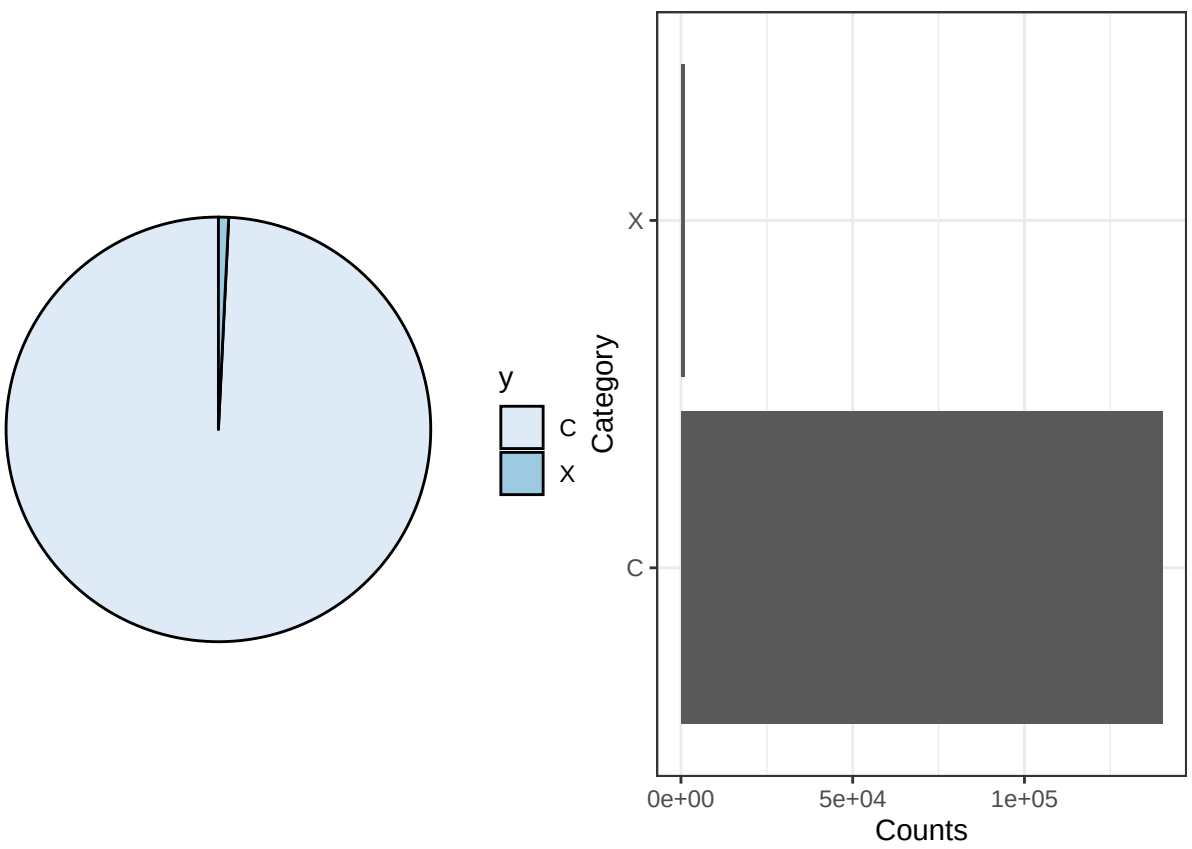
Plotting Histogram and getting descriptive Statistics for the dataset,

```
##       flux
##  Min.   :1.000e-09
##  1st Qu.:9.270e-08
##  Median :3.350e-07
##  Mean   :8.954e-07
##  3rd Qu.:7.890e-07
##  Max.   :7.178e-04
```



Here we have plotted the historgram of natural log of X-ray intensities because it was hard to visualize very small values of raw data. We can see that the distribution is multi-modal.

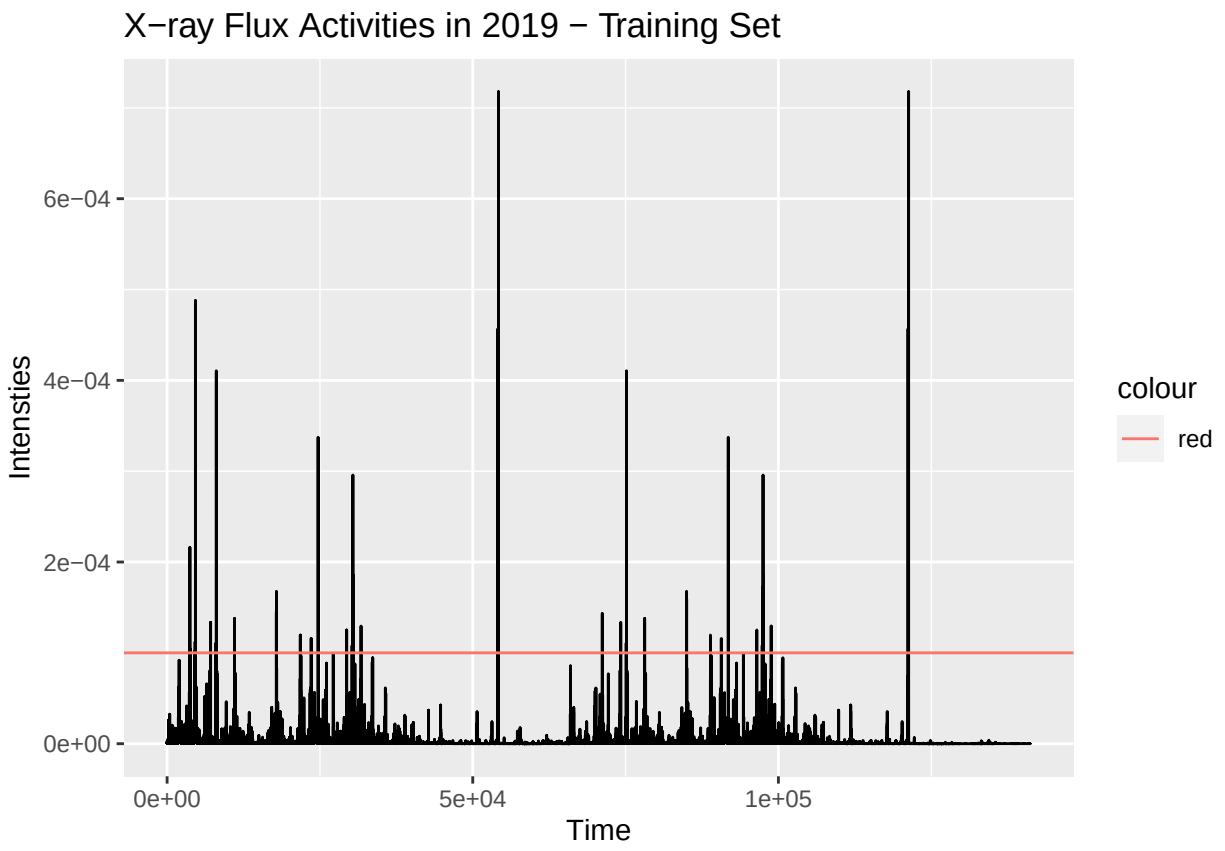
Now checking the distribution of Solar Flare vs Non-Solar Flare events.

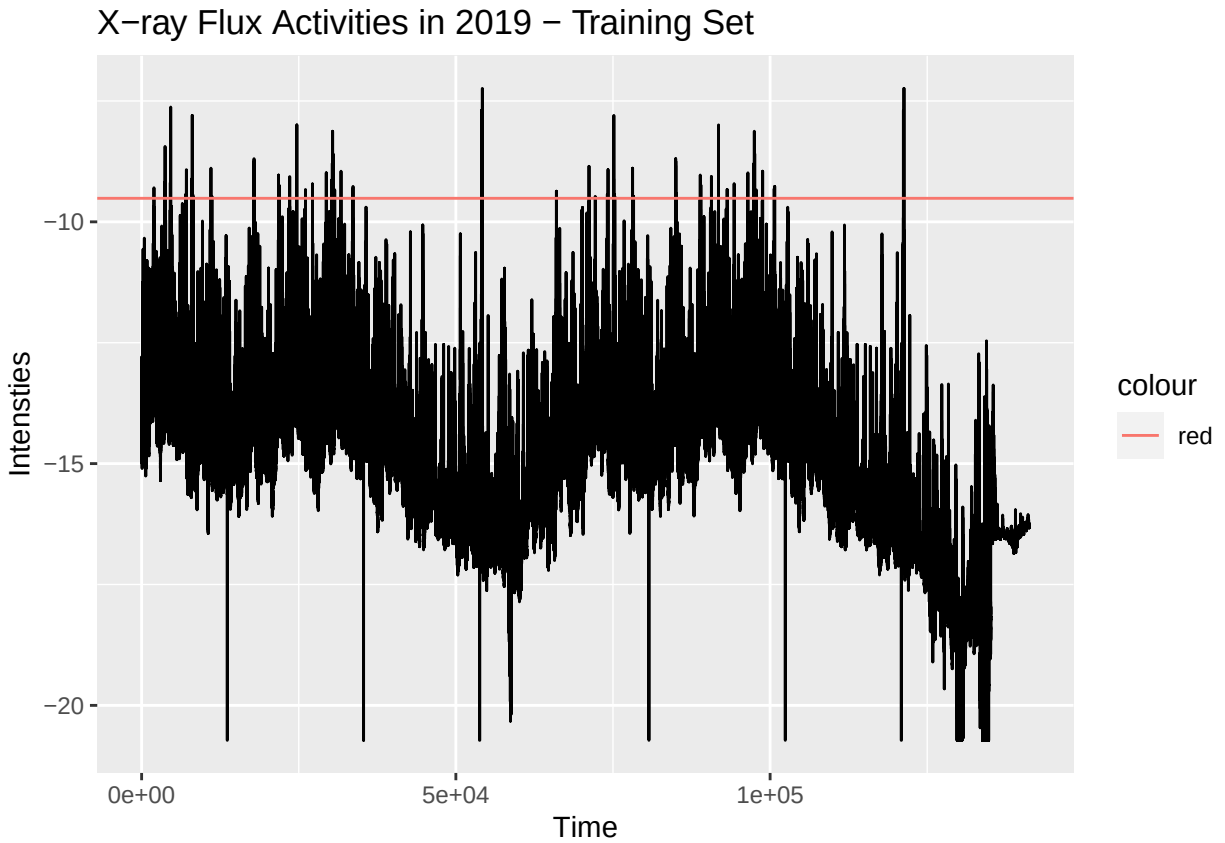


It is evident that there are very few solar flare events. This observation is in line with the research and reality.

Checking for Stationarity

Plotting the Xray data and log of Xray data to check for stationarity,





Visually inspecting the time series $\{y_1, y_2, \ldots, y_{141219}\}$ hints towards non-stationarity.

Performing Augmented Dickey Fuller Test to check for stationarity,

```
##
```

```
##  Augmented Dickey-Fuller Test
##
## data:  xray_data$flux
## Dickey-Fuller = -33.234, Lag order = 52, p-value = 0.01
## alternative hypothesis: stationary
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  log(xray_data$flux)
## Dickey-Fuller = -12.414, Lag order = 52, p-value = 0.01
## alternative hypothesis: stationary
```

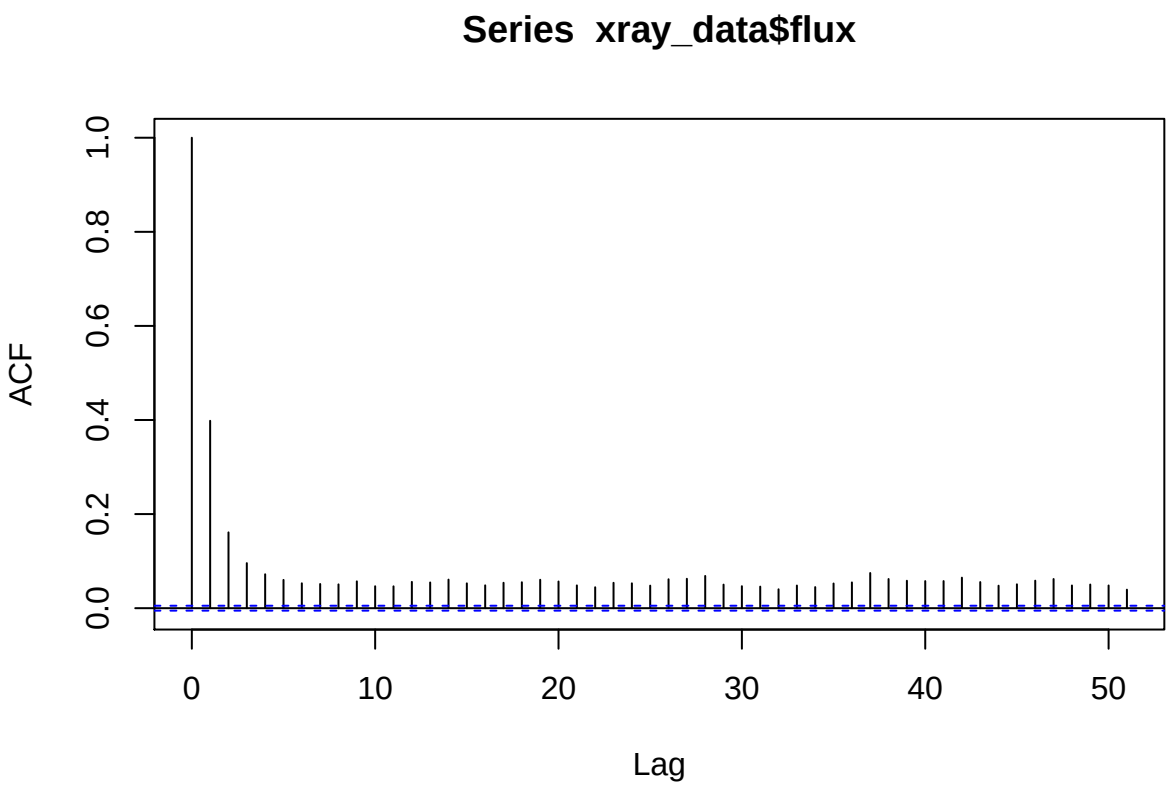
As per our research, the results of Augmented Dickey Fuller test can be slightly misleading. Hence double checking for presence of trend by fitting a Linear Regression model is a good idea,

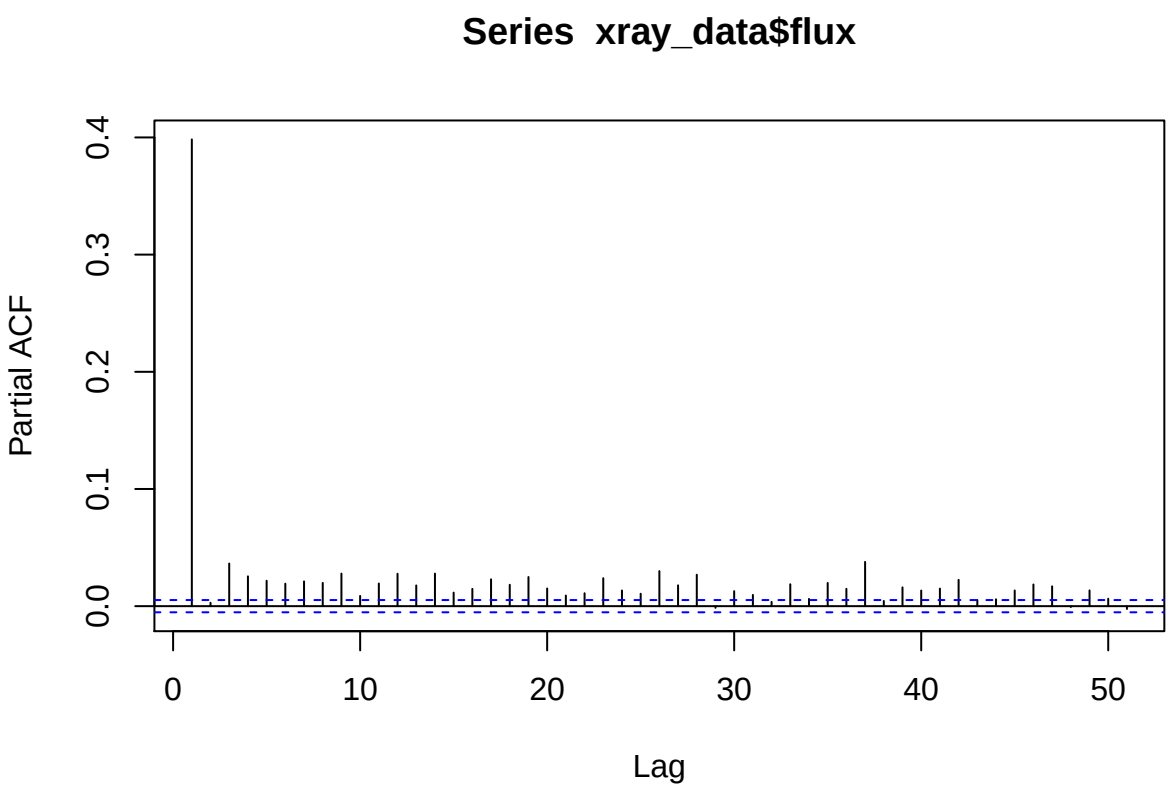
```
##
## Call:
## lm(formula = xray_data$flux ~ seq, data = xray_data)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.350e-06 -7.900e-07 -4.100e-07 -1.600e-07  7.173e-04
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.462e-06  3.013e-08   48.51   <2e-16 ***
## seq         -8.020e-12  3.696e-13  -21.70   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.662e-06 on 141217 degrees of freedom
## Multiple R-squared:  0.003324,   Adjusted R-squared:  0.003317
## F-statistic:   471 on 1 and 141217 DF,  p-value: < 2.2e-16
```

The seq variable estimate indicate that there is a slight trend. Hence we will de-trend the data by taking the first difference.

Exploring the temporal structure

Plotting the ACF and PACF plots of hourly data,



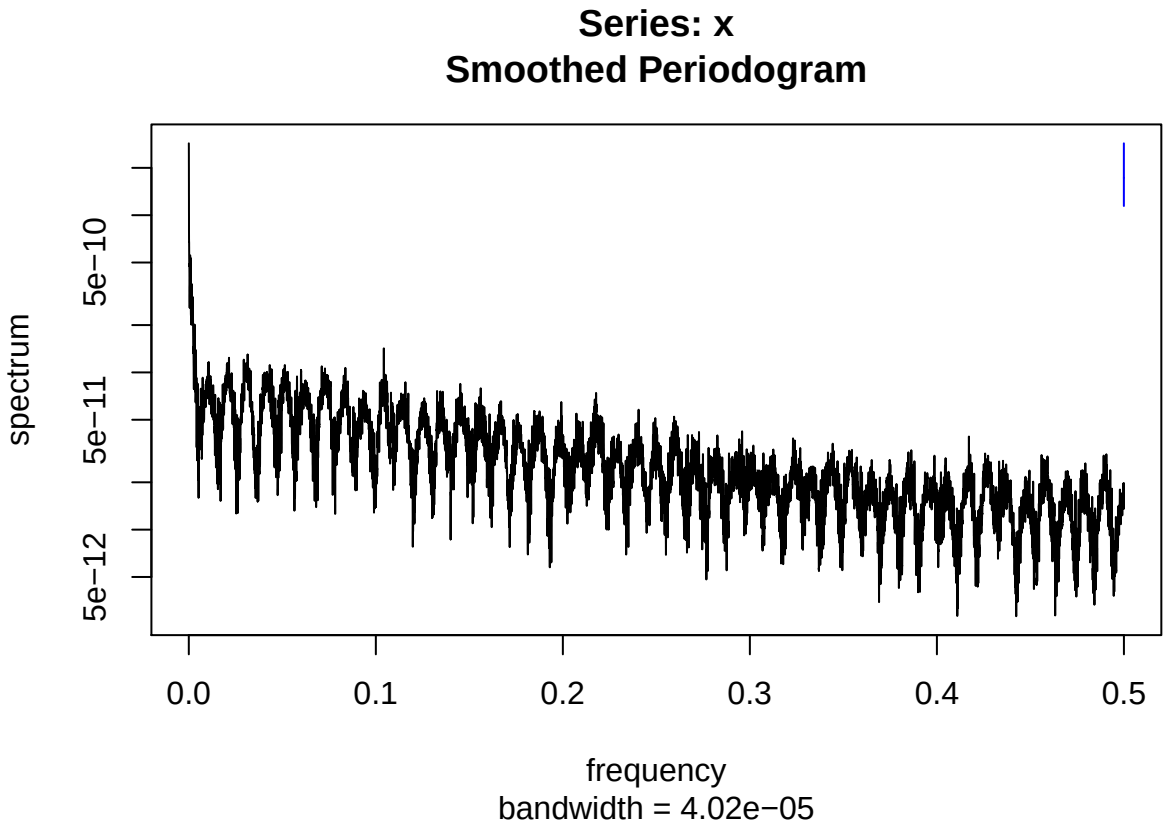


Takeaways:

- Nothing decisive can be said about the MA component. Just by looking at the ACF, q seems be 2 or 3 because after that there is a sharp decline in ACF.

- The order of autoregressive component seems to be 1 or 3.
- There appear to be some seasonal patterns in the graphs but the nature of cyclical trends will be explored during spectrum analysis.

Spectrum Analysis



Takeaways:

We are performing spectral analysis on hourly data. As per research, the solar flares are expected to happen every 11 years. This gives us a hint that using hourly data we will not be able to characterize different periodic components corresponding to solar activity. The spectrogram is expected to be multi-modal and the seasonality is expected to be at-least annual or every few years. This is also evident in the spectrogram above that as the frequency approaches zero the power approaches a very high number. There are no significant cyclical trends after every few hours.

Although not useful to explore seasonal patterns but still this hourly data can be used to explore build-up to a solar flare. This problem will be investigated later in the project.

Daily Data

Now loading and analyzing the monthly data,

Populating Data for Missing Years

The data for 2007, 2008 and 2009 was missing. In order to impute this data, we took the mean and standard deviation of X-Ray intensities of surrounding years and generated a sequence of random numbers from the corresponding normal distribution. We generated 1095 random observations for 3 years missing daily data.

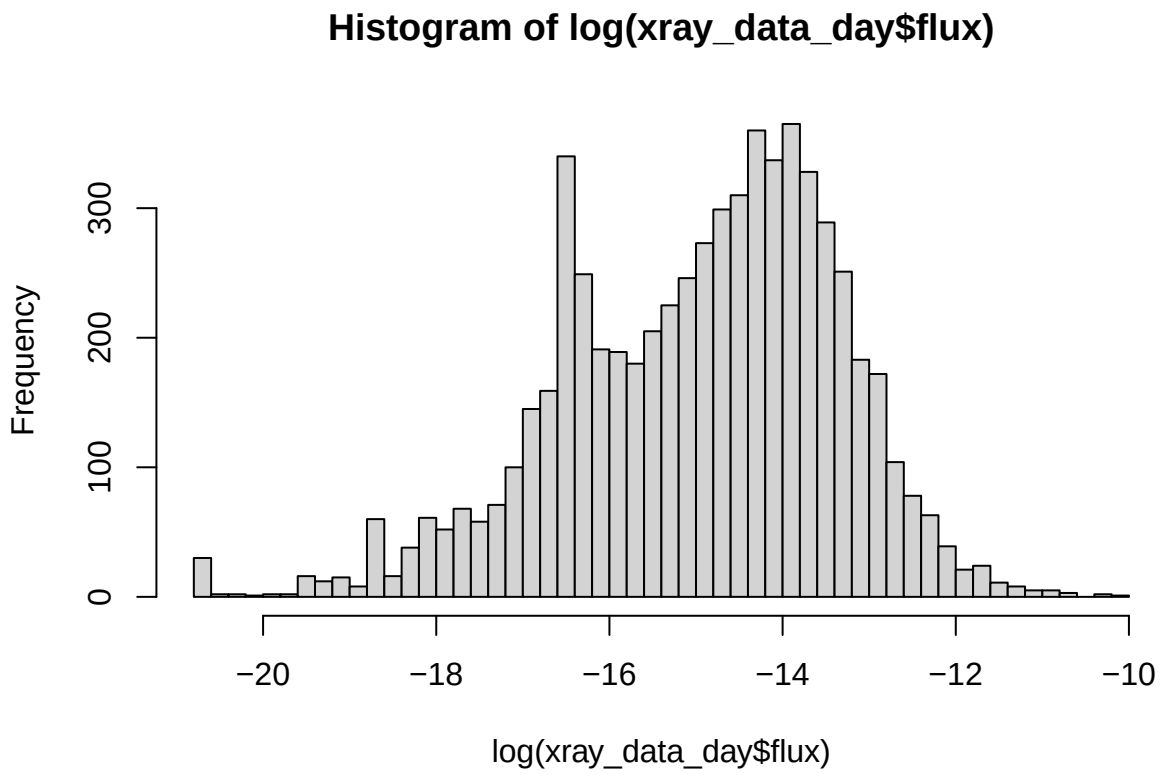
```
## The total number of observation in time series is: 6274
```

```
##
## The number of NAs in the dataset is: 0
```

Dataset Characterisitcs

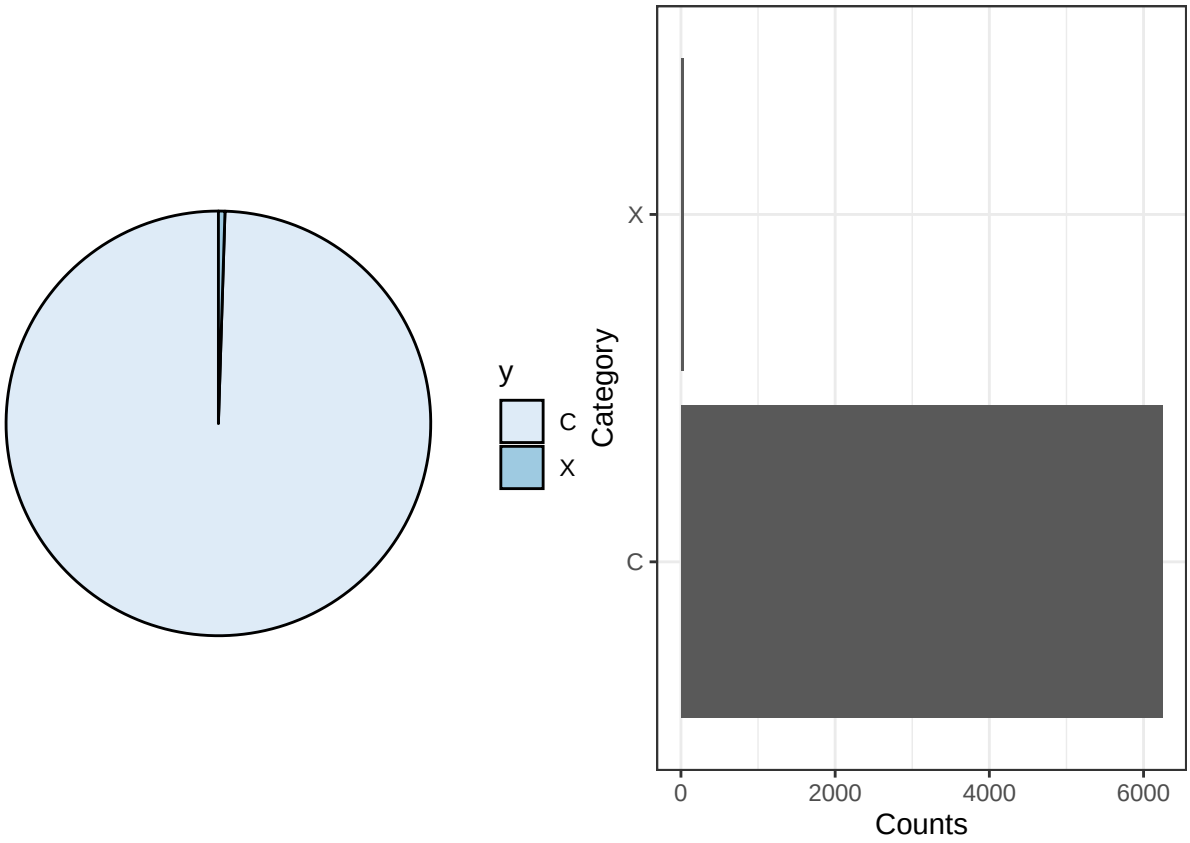
Plotting Histogram and getting descriptive Statistics for the dataset,

```
##       flux
##  Min.   :1.000e-09
##  1st Qu.:9.790e-08
##  Median :4.103e-07
##  Mean   :8.713e-07
##  3rd Qu.:1.030e-06
##  Max.   :4.298e-05
```



Like hourly data, the distribution of daily data appears to be multi-modal as well.

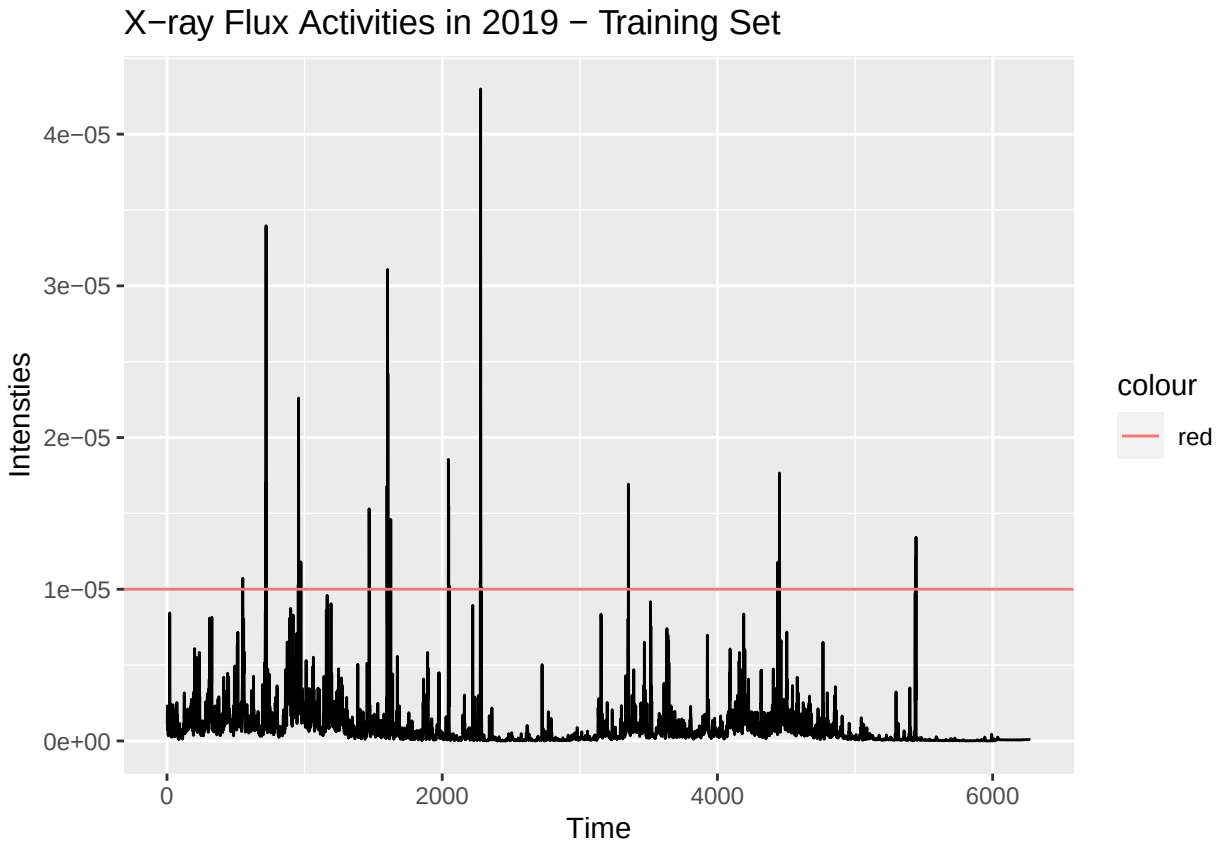
Now checking the distribution of Solar Flare vs Non-Solar Flare events in daily data.

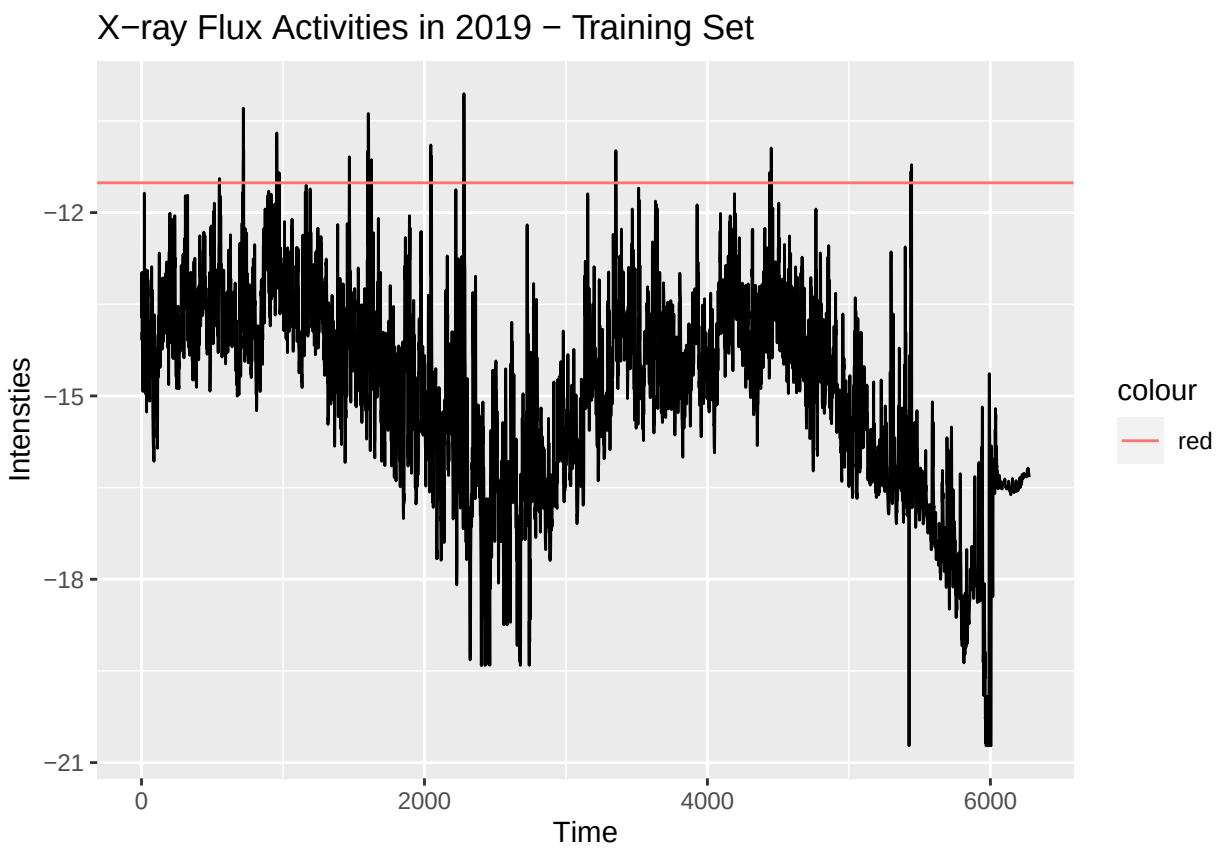


It is evident that like hourly data, in daily data as well there are very few solar flare events.

Checking for Stationarity

Plotting the Xray daily data and log of Xray daily data to check for stationarity,





Visually inspecting the time series $\{y_1, y_2, \ldots, y_{6274}\}$ hints towards non-stationarity.

Performing Augmented Dickey Fuller Test to check for stationarity,

```
##
##  Augmented Dickey-Fuller Test
##
## data:  xray_data_day$flux
## Dickey-Fuller = -13.176, Lag order = 18, p-value = 0.01
## alternative hypothesis: stationary
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  log(xray_data_day$flux)
## Dickey-Fuller = -6.1362, Lag order = 18, p-value = 0.01
## alternative hypothesis: stationary
```

As per our research, the results of Augmented Dickey Fuller test can be slightly misleading. Hence double checking for presence of trend by fitting a Linear Regression model is a good idea,

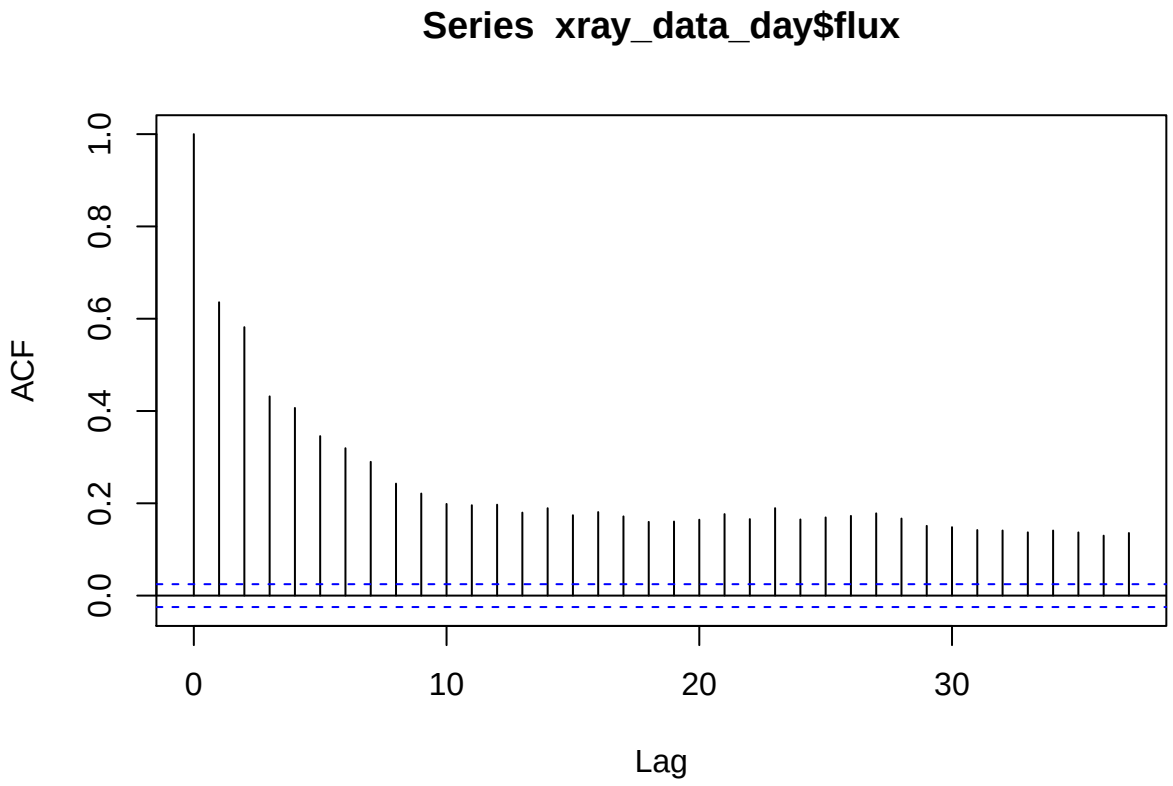
```
##
## Call:
## lm(formula = xray_data_day$flux ~ seq, data = xray_data_day)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.492e-06 -7.430e-07 -2.460e-07  7.800e-08  4.190e-05
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

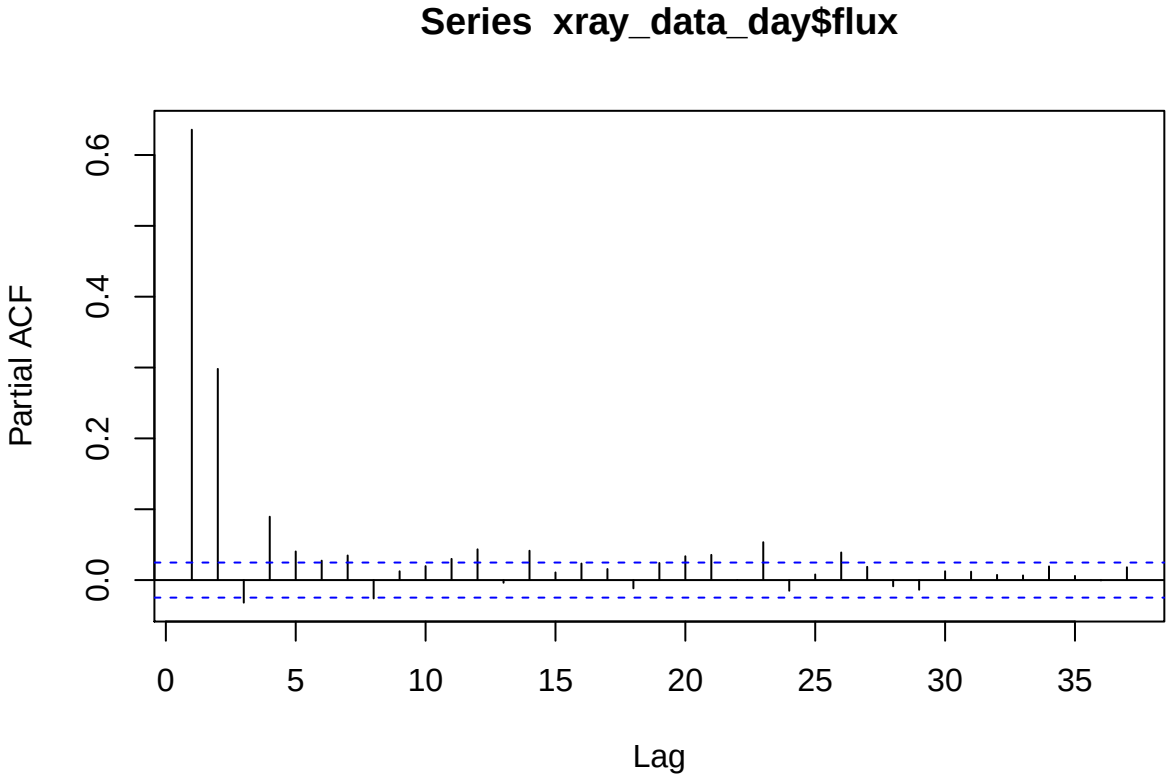
```
## (Intercept)  1.618e-06  4.056e-08   39.90   <2e-16 ***
## seq         -2.381e-10  1.120e-11  -21.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.606e-06 on 6272 degrees of freedom
## Multiple R-squared:  0.06723,    Adjusted R-squared:  0.06708
## F-statistic: 452.1 on 1 and 6272 DF,  p-value: < 2.2e-16
```

The seq variable estimate indicate that there is a slight trend. Hence we will de-trend the data by taking the first difference.

Exploring the temporal structure

Plotting the ACF and PACF plots of hourly data,

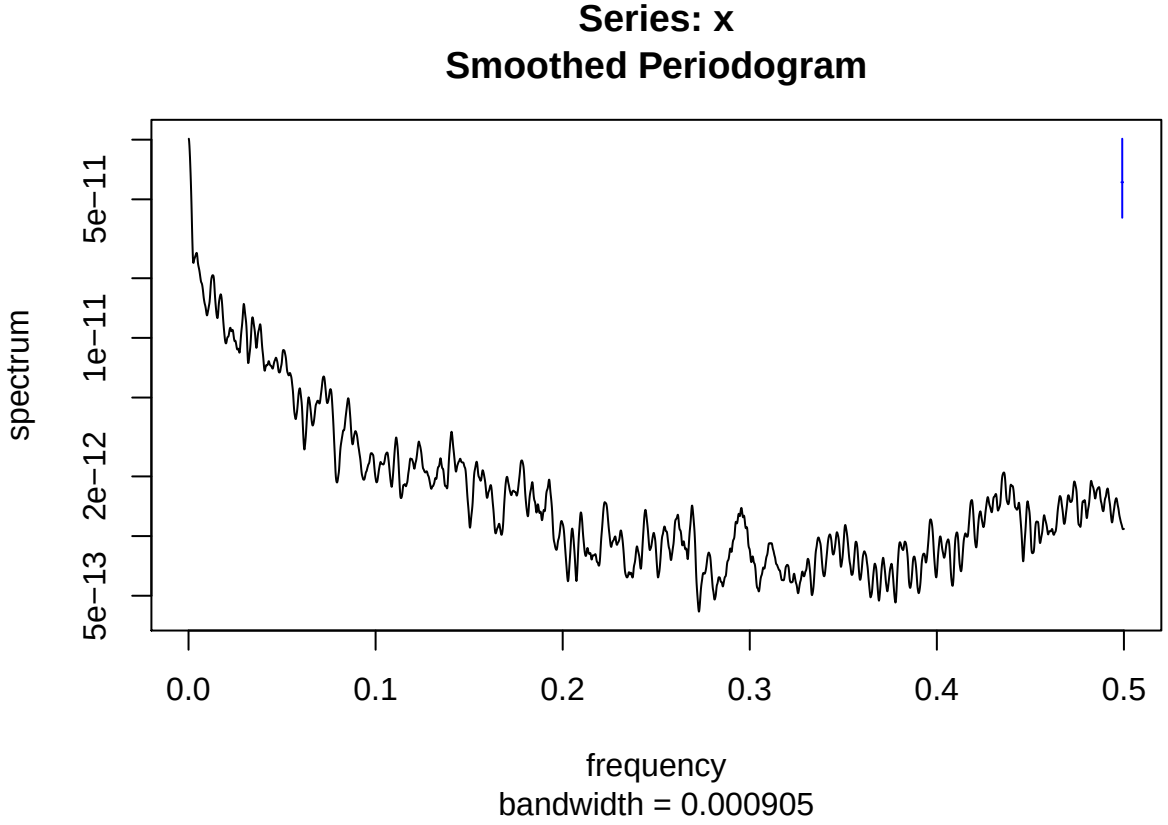




Takeaways:

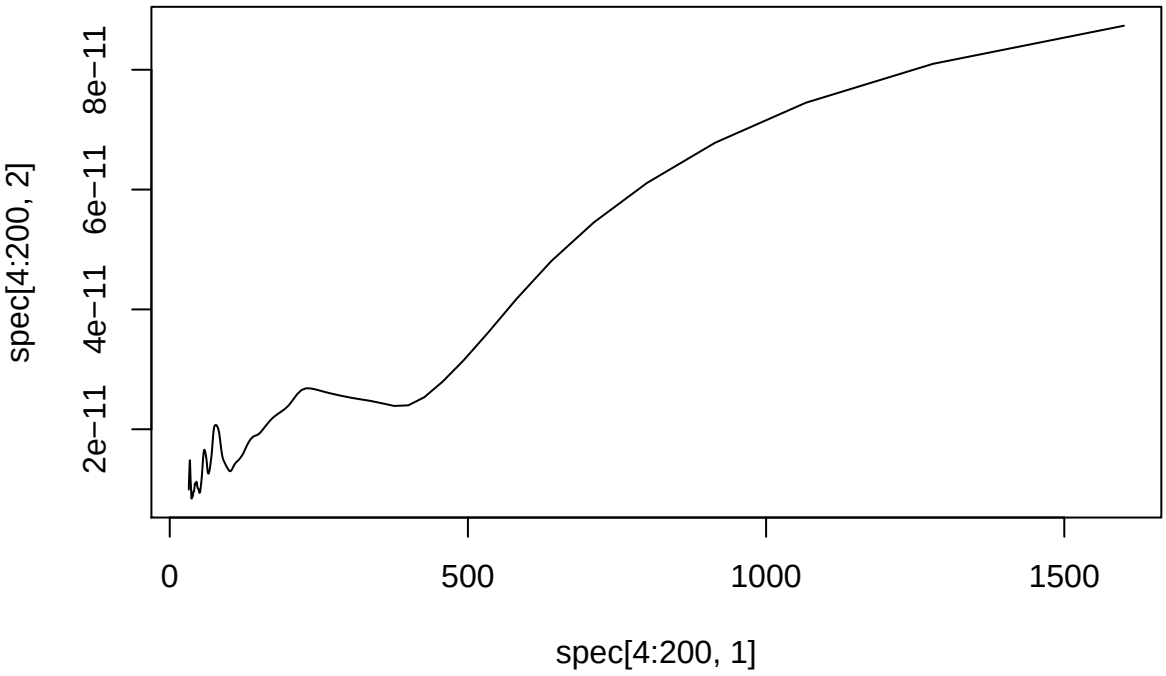
- Nothing decisive can be said about the MA component.
- The order of autoregressive component seems to be 2.
- Seasonality is not very evident from the ACF and PACF plots. Conducting spectral analysis will tell us more about the seasonal components present in daily X-Ray data.

Spectrum Analysis



This is evident in the spectrogram above that as the frequency approaches zero the power approaches a very high number.

In order to get a better view of different frequency components let us convert frequency to the period (in number of days) and plot it vs power. We will plot a subset of the frequencies to have a better view of different seasonal components in play.



Takeaways:

- There is a peak at roughly 1 year period. Meaning that there are hidden annual seasonal trends associated with x-ray intensities.
- The power increases a lot as the number of years increases indicating that most of the seasonal trends follow a multi year cycle. This is in line with the fact that Solar Flare repeats after every 11 years.
- Due to the granularity of the data and some constraints in the R spectrum function, the exact peaks can't be pin-pointed in the spectrum plot.

Research Questions

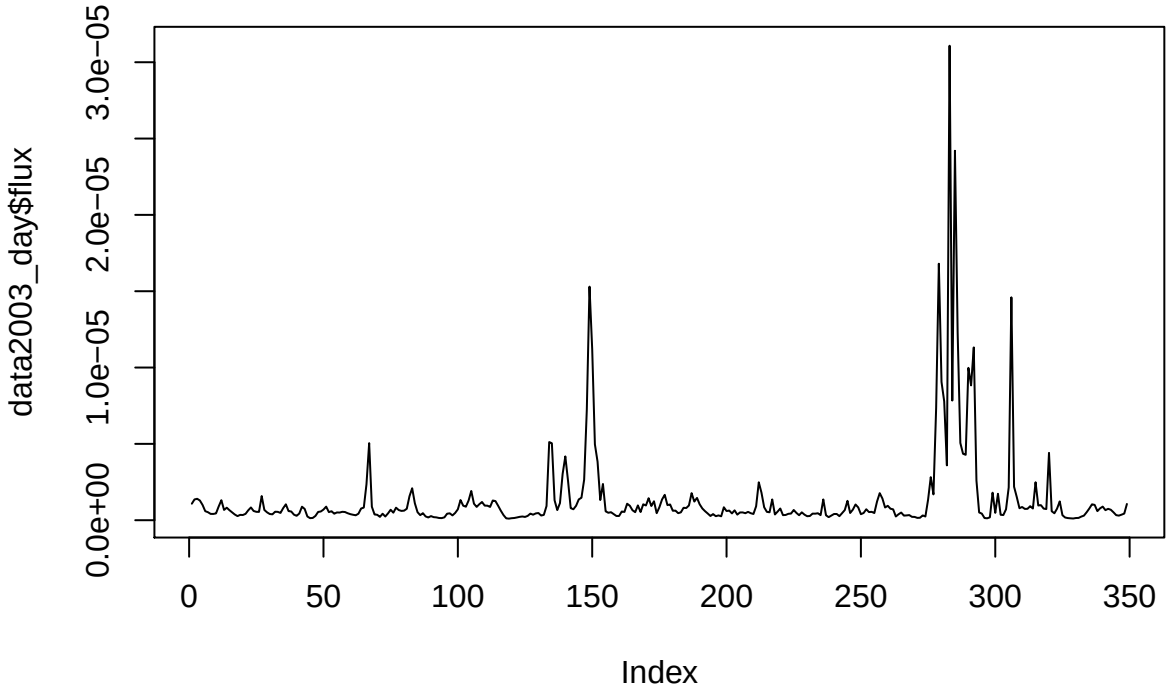
Question 1&2: Exploring seasonality and build up of X-ray Intensity

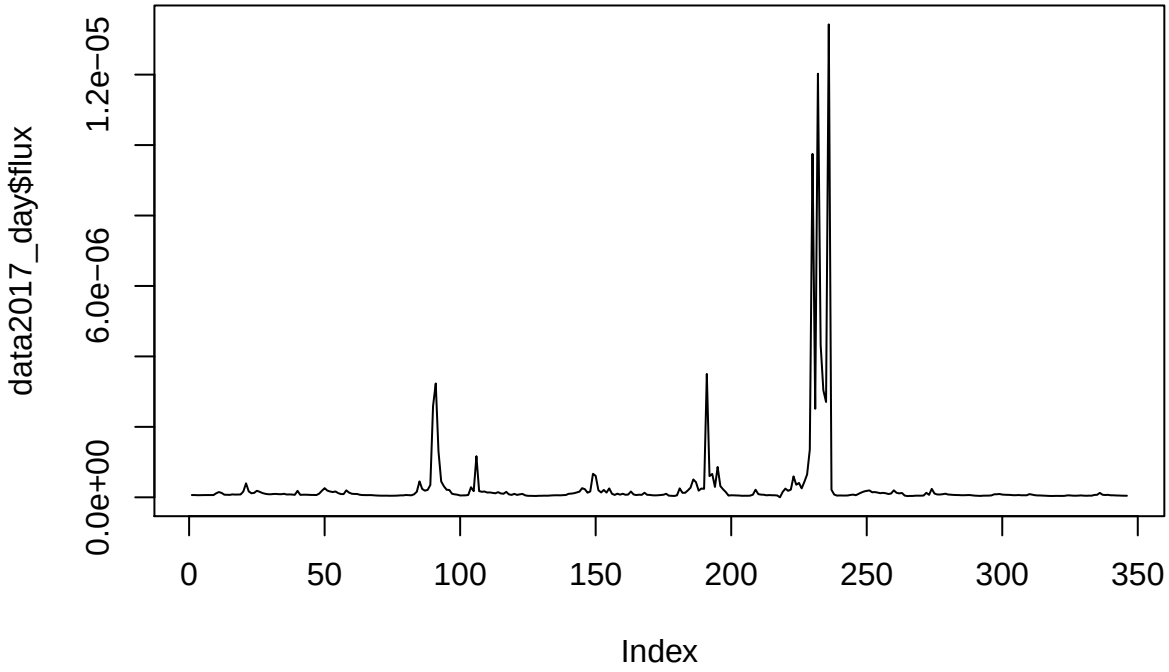
As mentioned in the previous sections, although the **PACF** gave us a somewhat decisive answer for our parameter, this was not the case with the **ACF**. Hence, in this section we will explore seasonality further. This time around, instead of looking at the whole data we will focus on different years separately and try find some common traits in the years where solar flares happened versus the years where no solar flares occurred.

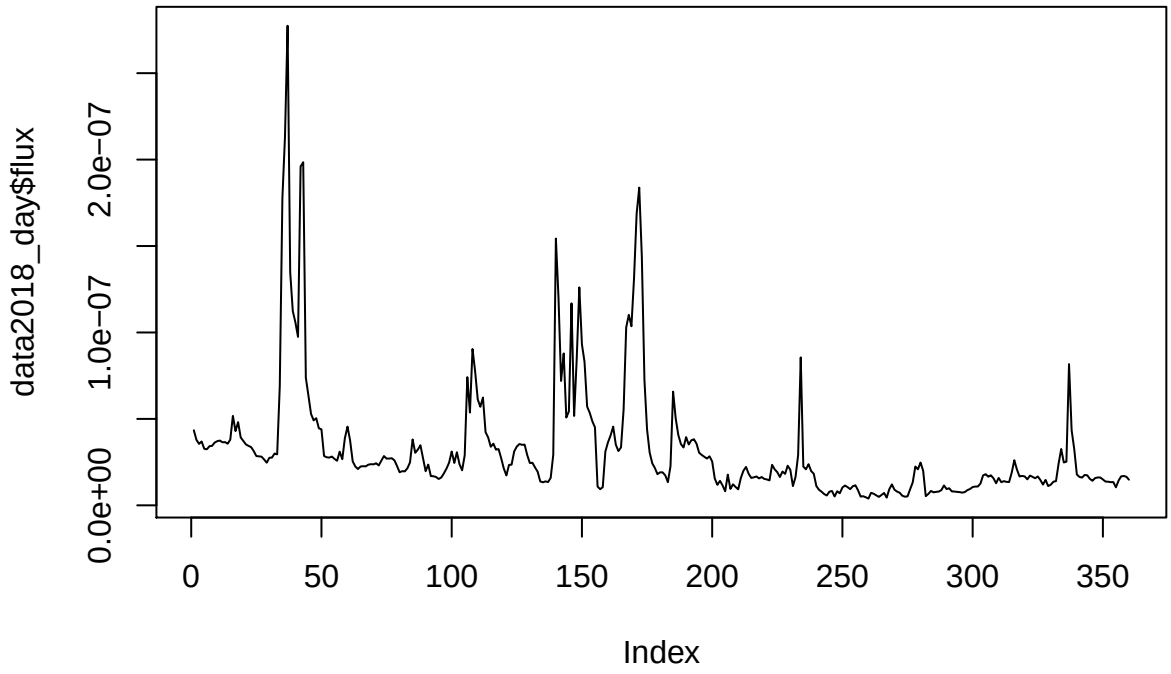
Our initial idea was the following:

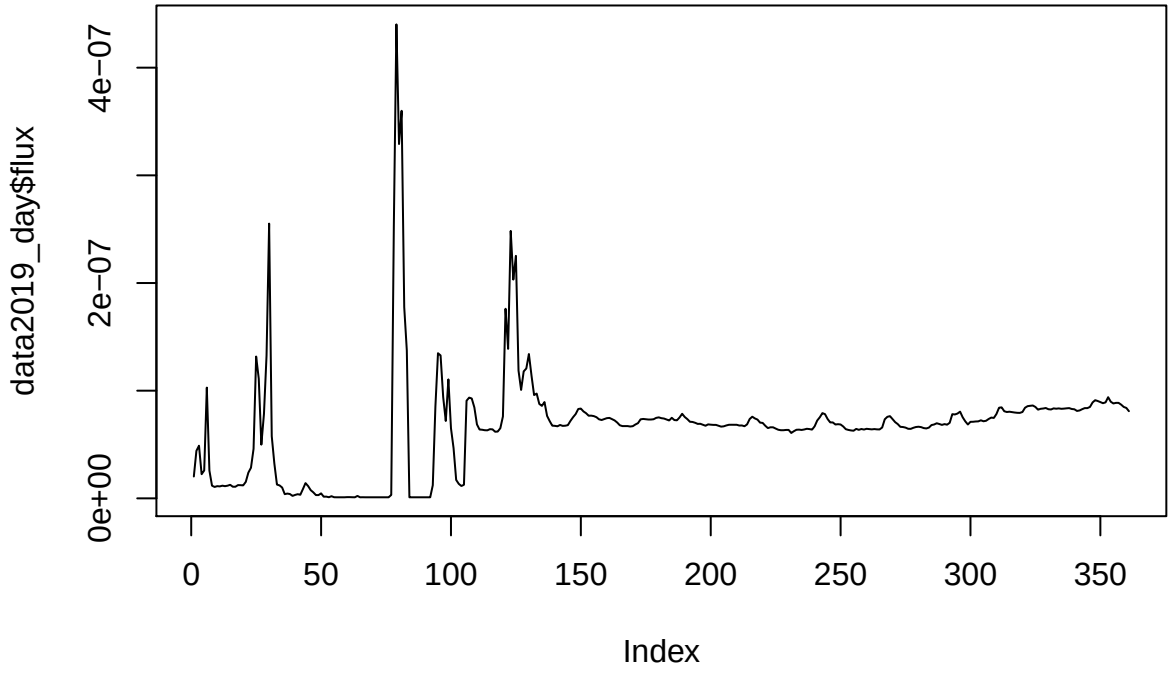
We first split the Sun's state into 2 categories. Category 1 consists of periods when the sun is emitting lower intensity x-rays ("buzzing"). Naturally Category 2 refers to the periods when the sun produces higher intensity x-rays (blips) and hence generates solar flares.

We want to see if in the years (or parts of years) that fall into Category 1, we can observe shorter seasonal periods which are more significant that the longer ones. On the other hand, in the years (or parts of years) that fall into Category 2, we want to see whether the longer seasonal periods are more significant. For example, lets plot years 2003,2017 vs 2018,2019 (first two had a solar flare, the second two did not).


data2003_day$flux
0.0e+00
1.0e−05
2.0e−05
3.0e−05
0
50
100
150
200
250
300
350
Index


data2017_day$flux
0.0e+00
6.0e−06
1.2e−05
0
50
100
150
200
250
300
350
Index



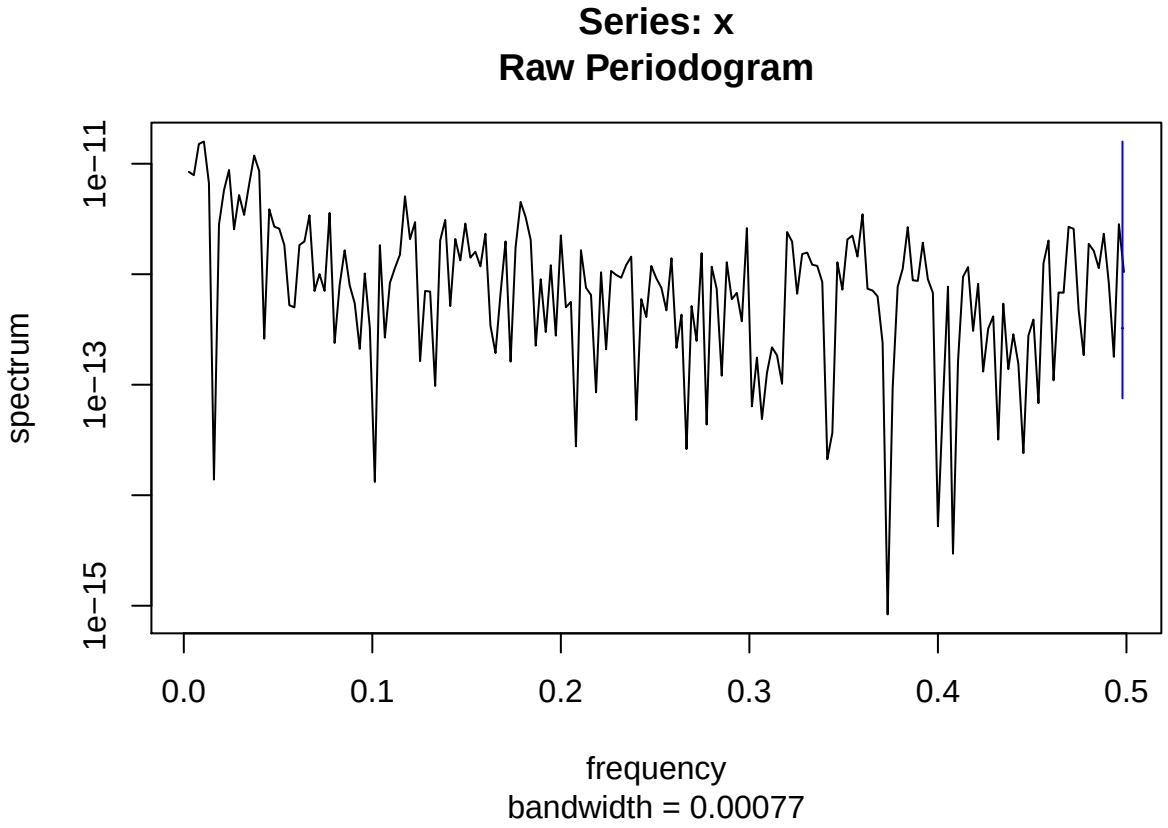


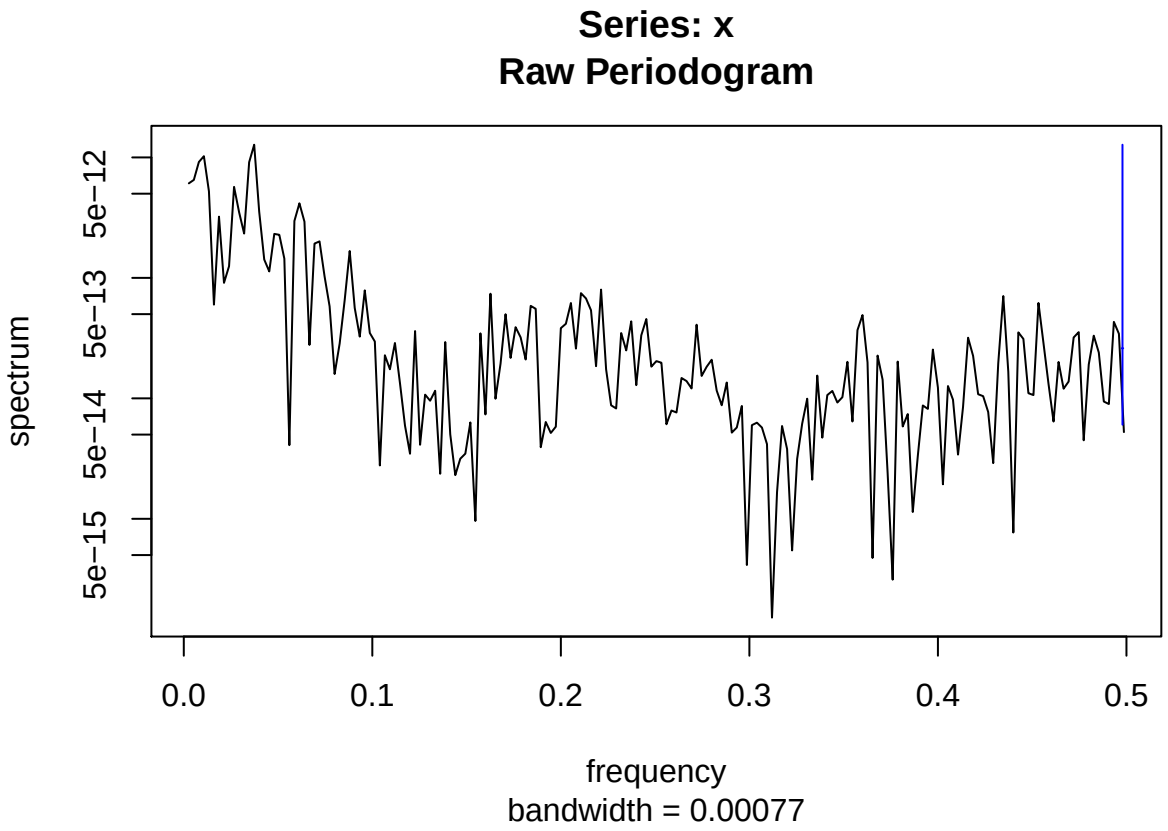
In the first two graphs, we can see about 2-3 blips before the solar flare occurres. These blips are separated in intervals of approximately 50 up to 150 days. Therefore, we are expecting that frequencies of around 50 days and 100-150 days would be more significant that shorter ones.

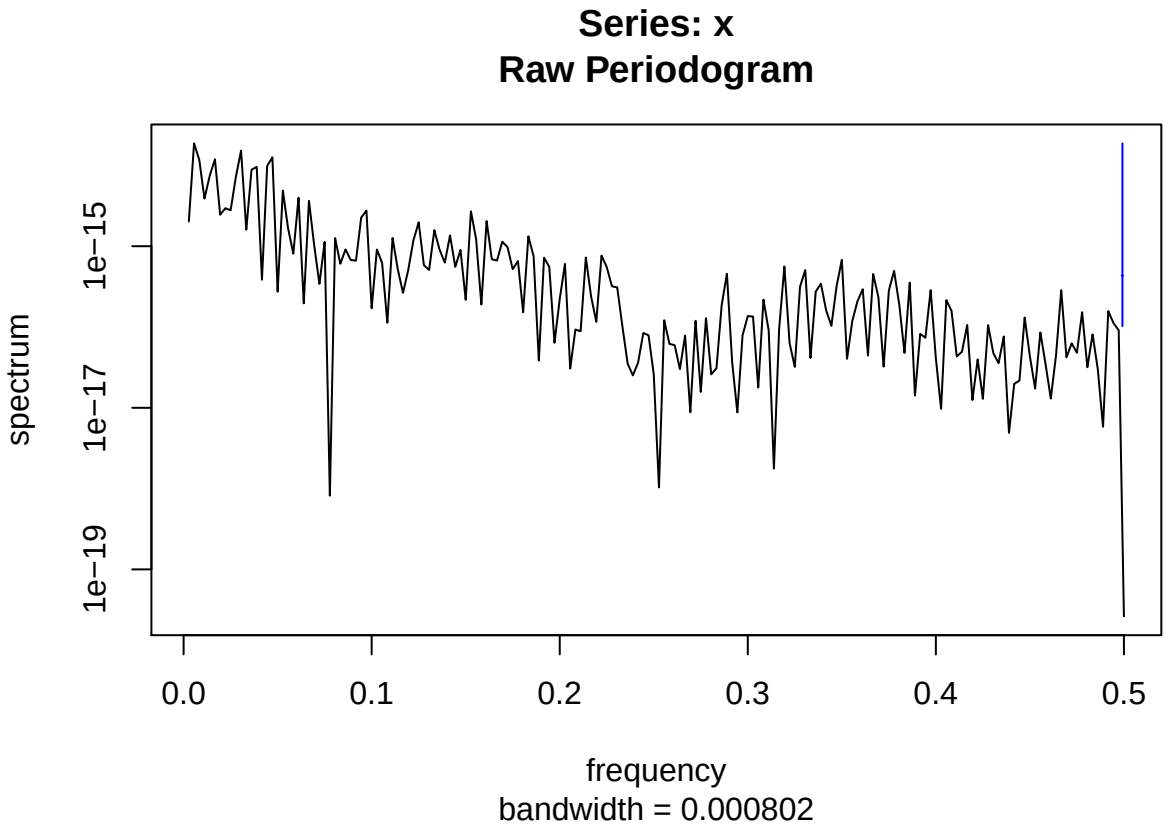
On the other hand, if we look at the y-scale in the third and fourth graph we can see that these blips are

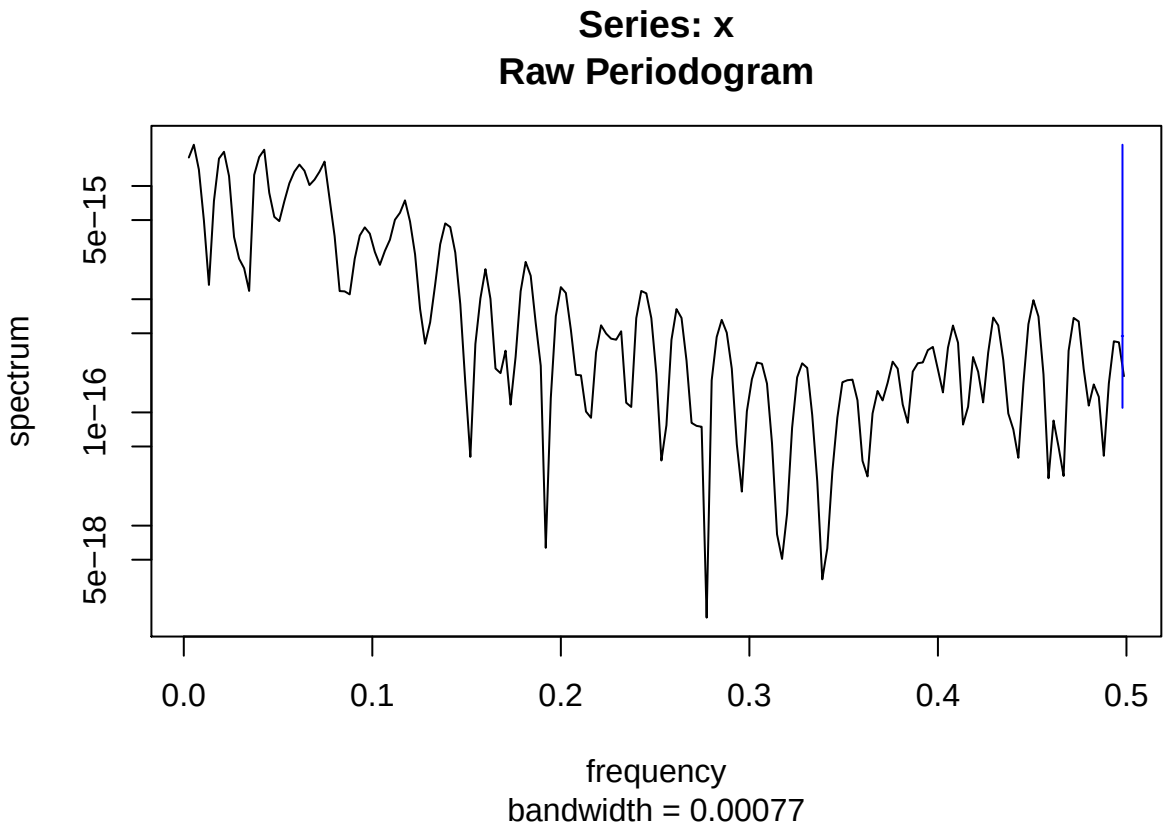
insignificant and too small in magnitude. However, in years such as 2018,2019, we are expecting low intensity x-rays ("buzzing") and there is no seasonality corresponding to the larger build-up of blips before a solar flare happens.

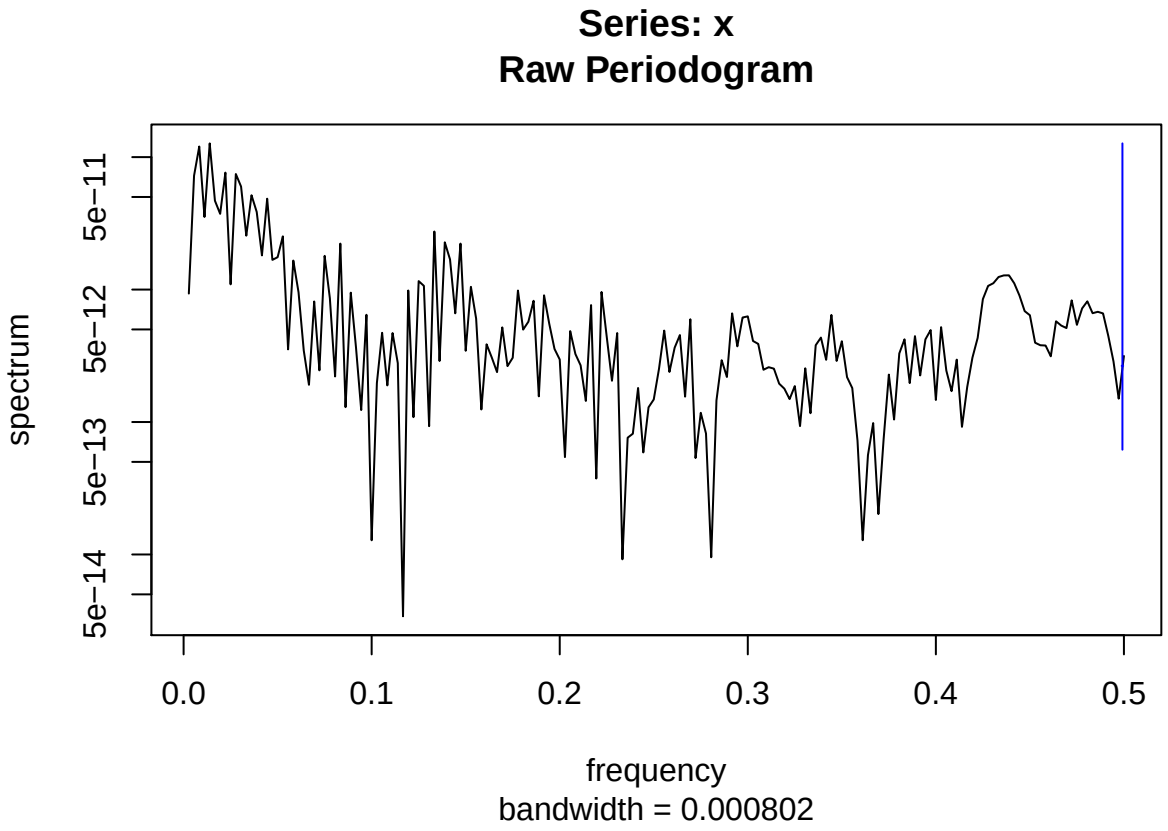
In order to verify this claim, we plot the spectrogram of years: 2002,2004,2018,2019,2003,2006,2011,2017. Note that the first 4 years listed did not have solar flares where as the last 4 did.

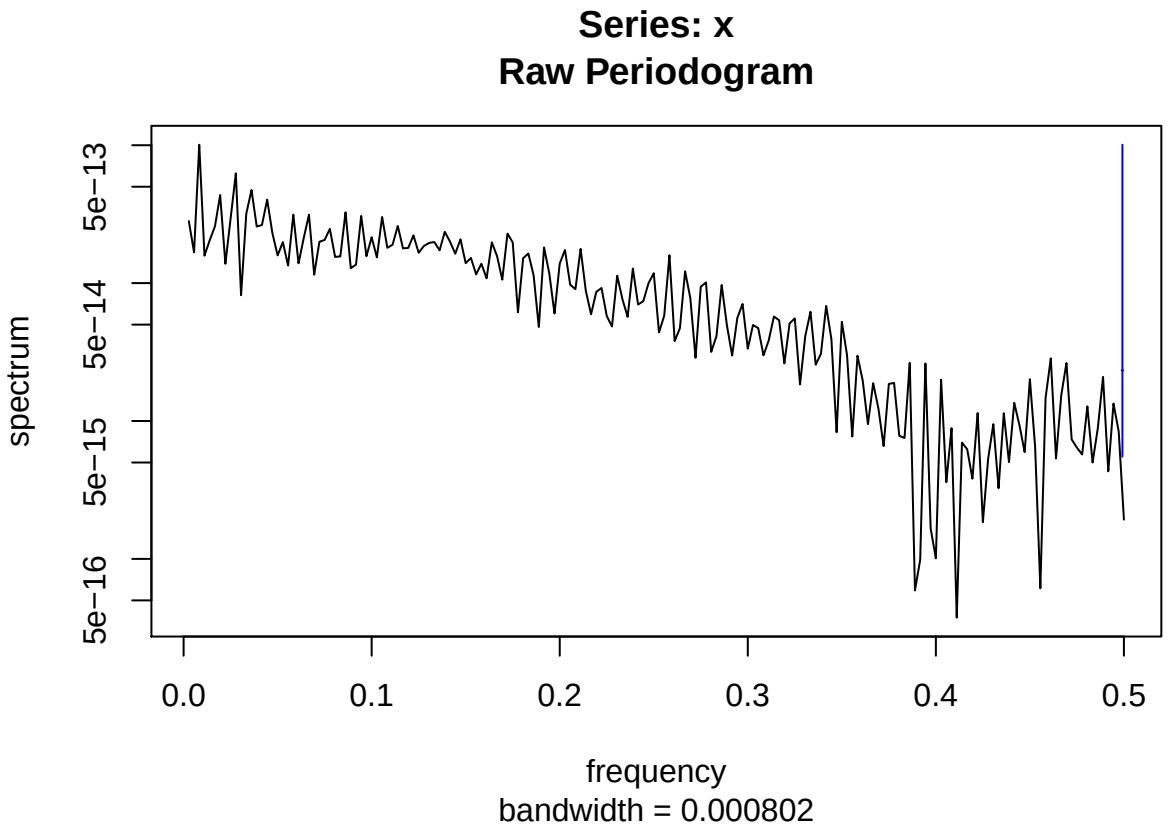


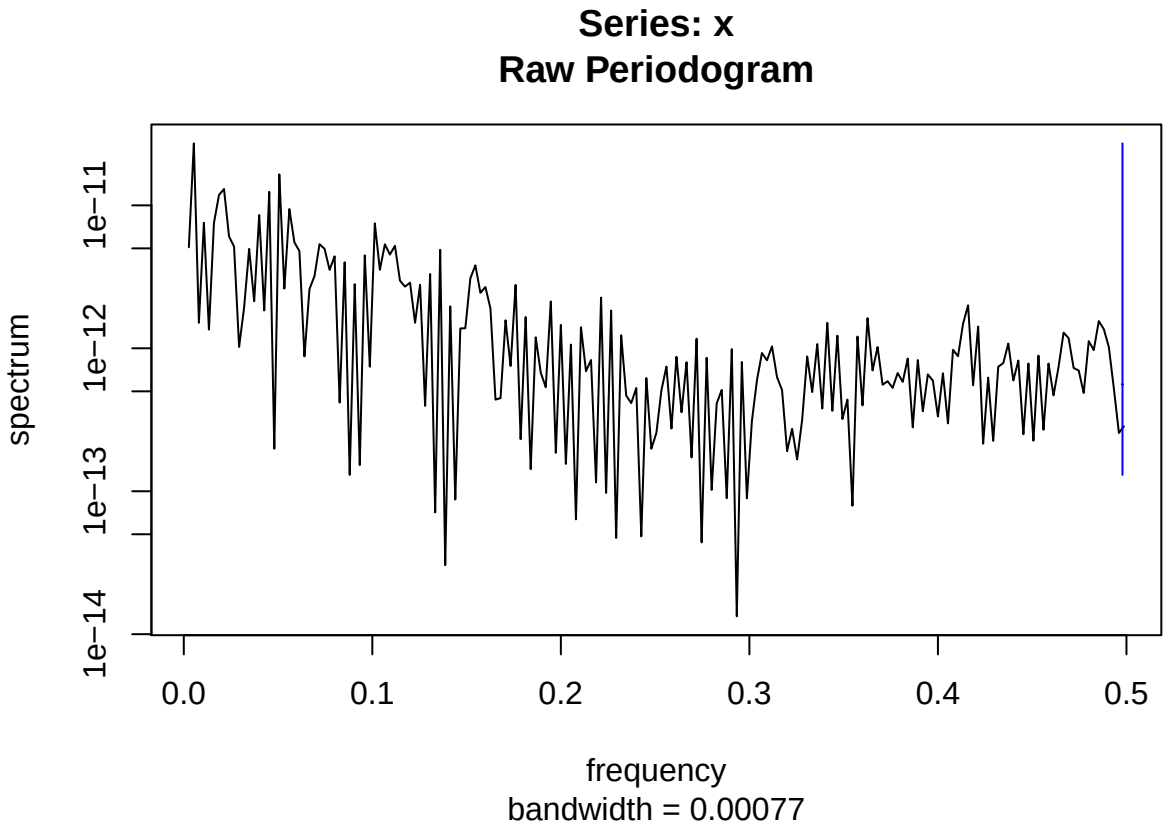

Series: x
Raw Periodogram
spectrum
5e−12
5e−13
5e−14
5e−15
0.0
0.1
0.2
0.3
0.4
0.5
frequency
bandwidth = 0.00077

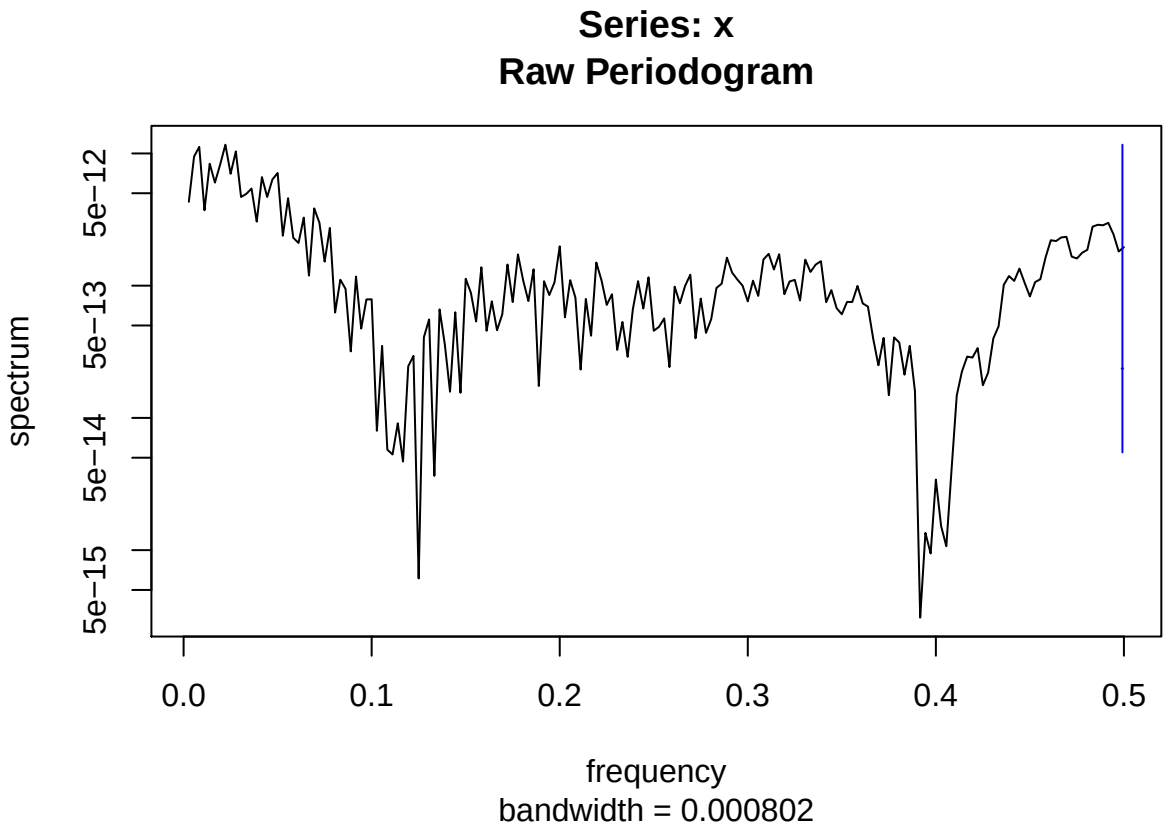

Series: x
Raw Periodogram
spectrum
1e−15
1e−17
1e−19
0.0
0.1
0.2
0.3
0.4
0.5
frequency
bandwidth = 0.000802

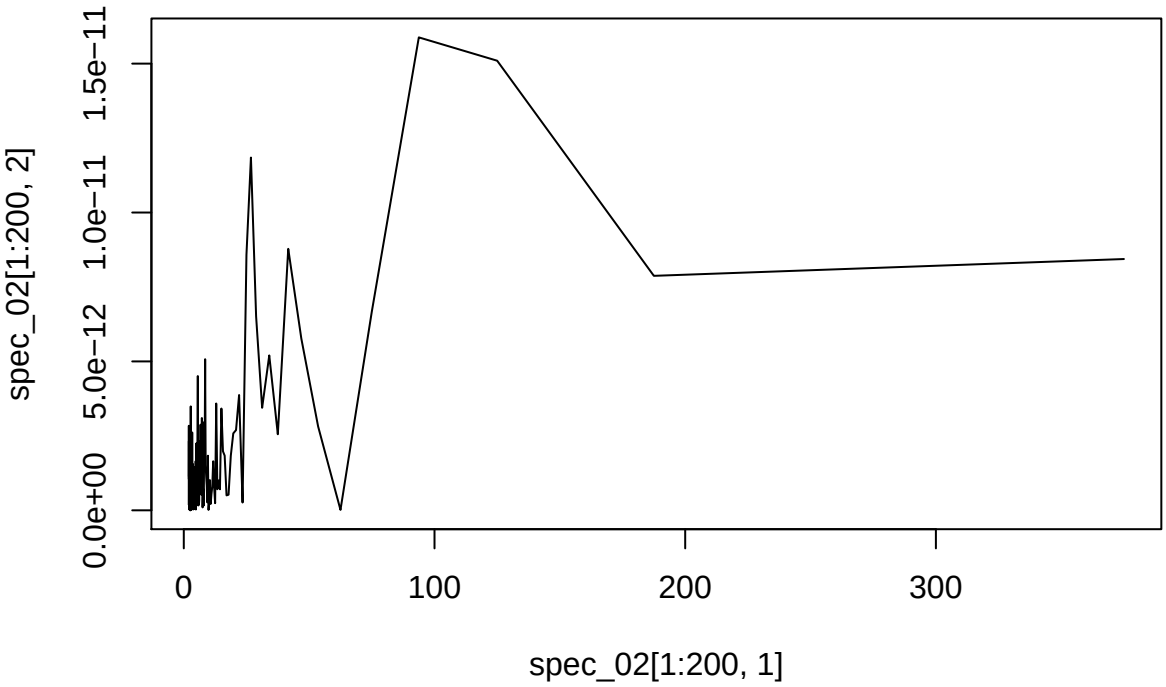

Series: x
Raw Periodogram
spectrum
5e−15
1e−16
5e−18
0.0
0.1
0.2
0.3
0.4
0.5
frequency
bandwidth = 0.00077

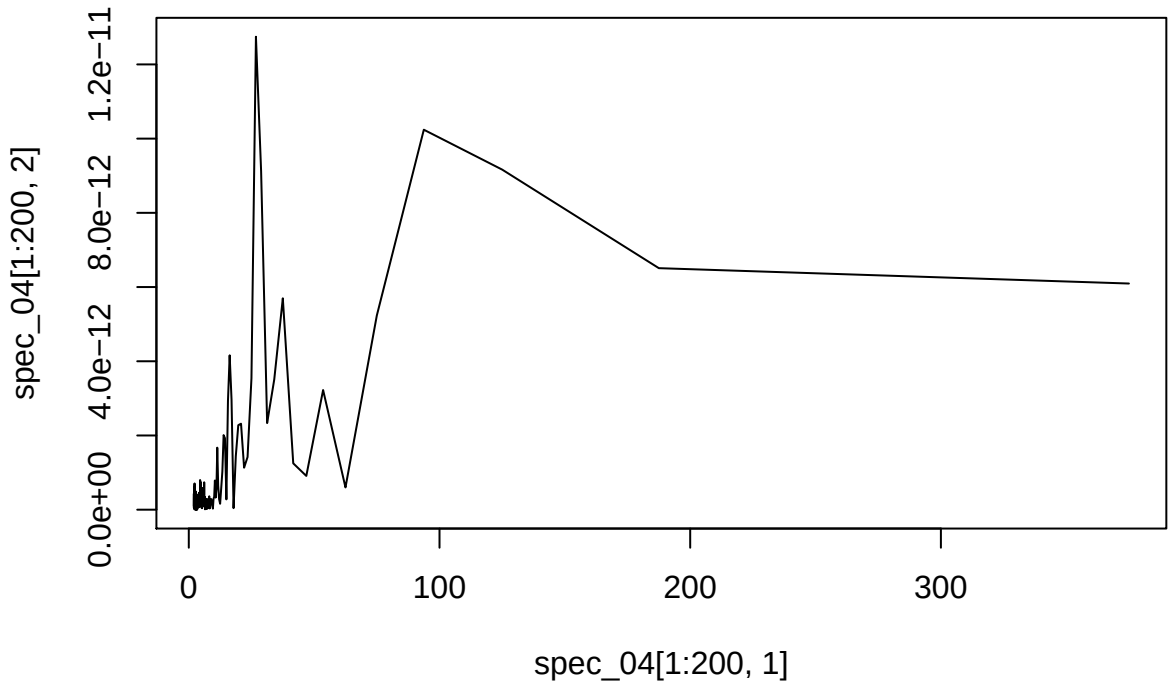

Series: x
Raw Periodogram
spectrum
5e−11
5e−12
5e−13
5e−14
0.0
0.1
0.2
0.3
0.4
0.5
frequency
bandwidth = 0.000802

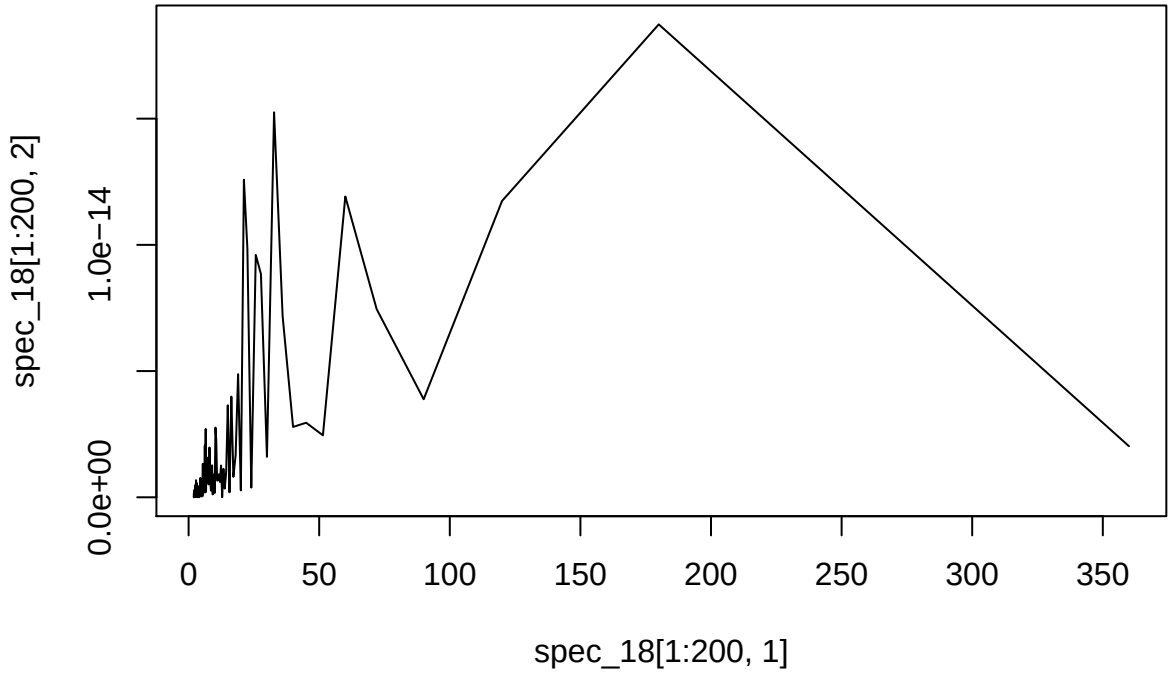

Series: x
Raw Periodogram
spectrum
5e−13
5e−14
5e−15
5e−16
0.0
0.1
0.2
0.3
0.4
0.5
frequency
bandwidth = 0.000802

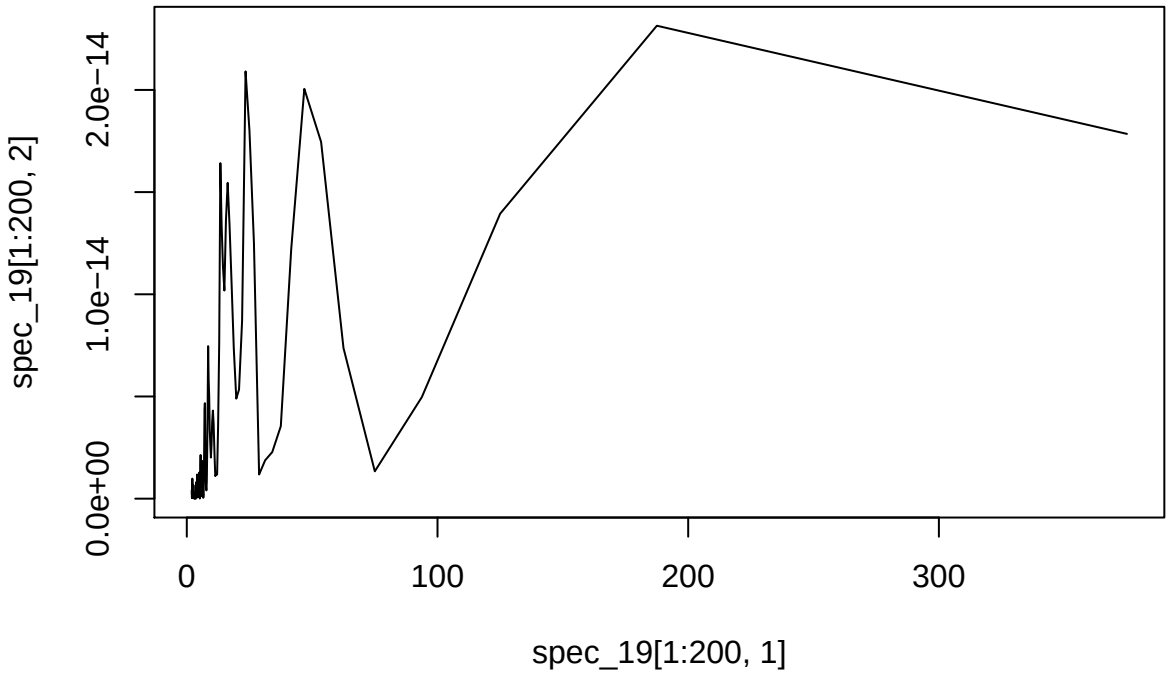

Series: x
Raw Periodogram
spectrum
1e−11
1e−12
1e−13
1e−14
0.0
0.1
0.2
0.3
0.4
0.5
frequency
bandwidth = 0.00077

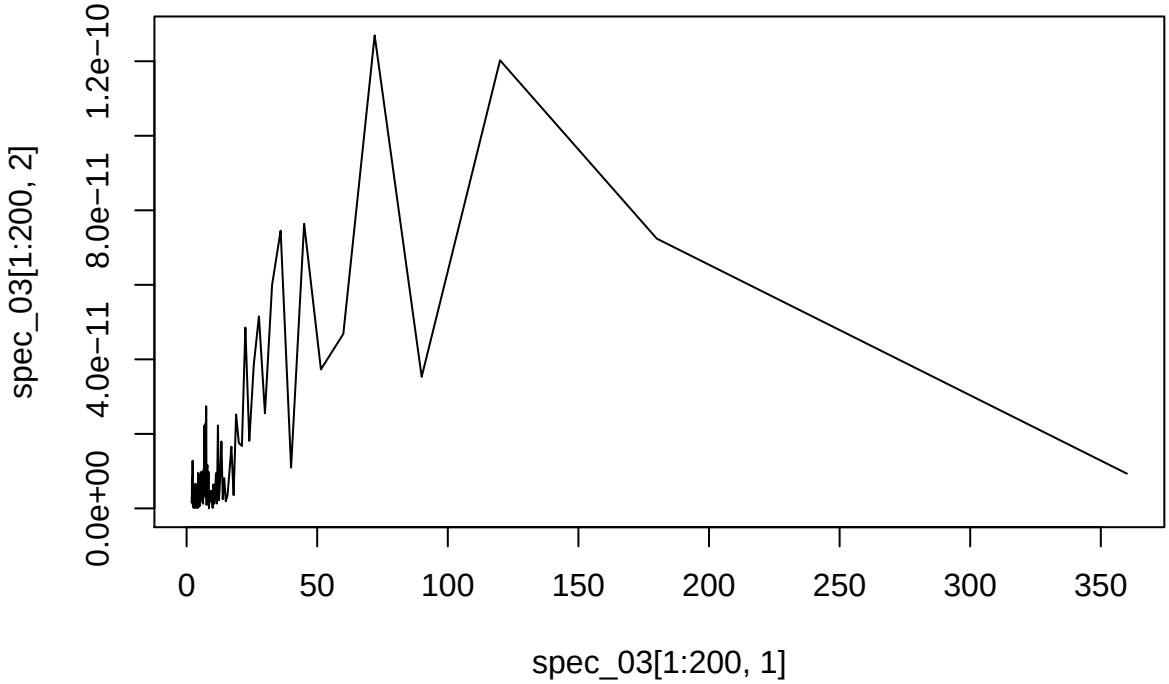

Series: x
Raw Periodogram
spectrum
5e−12
5e−13
5e−14
5e−15
0.0
0.1
0.2
0.3
0.4
0.5
frequency
bandwidth = 0.000802

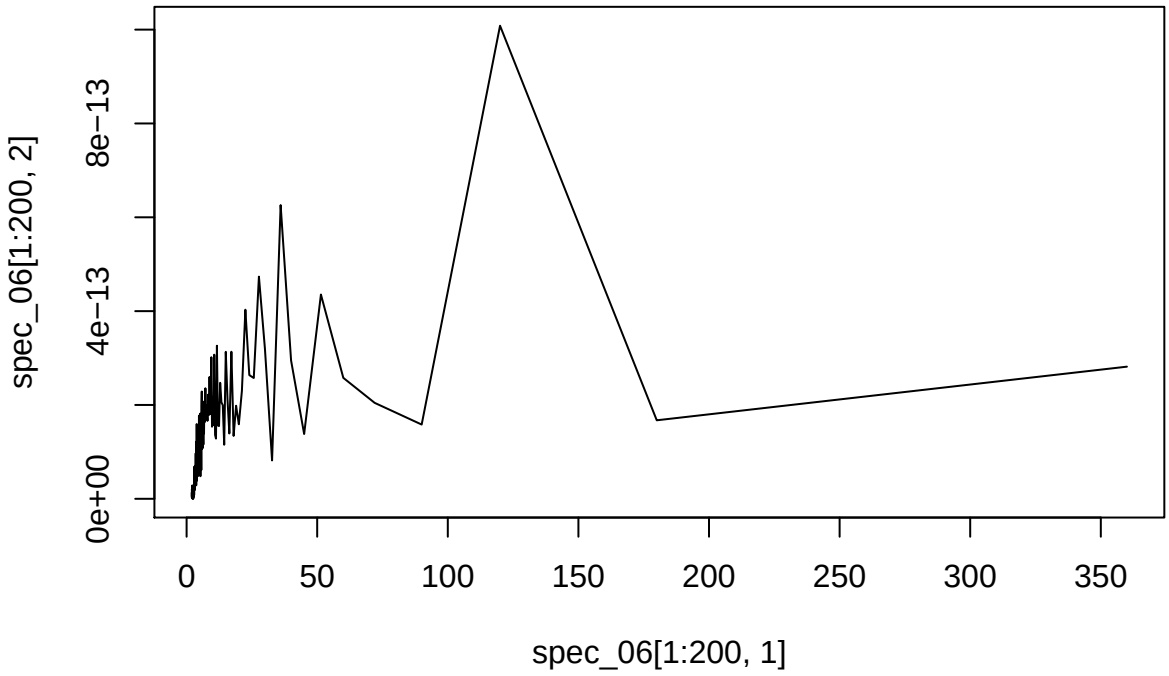

spec_02[1:200, 2]
0.0e+00
5.0e−12
1.0e−11
1.5e−11
0
100
200
300
spec_02[1:200, 1]

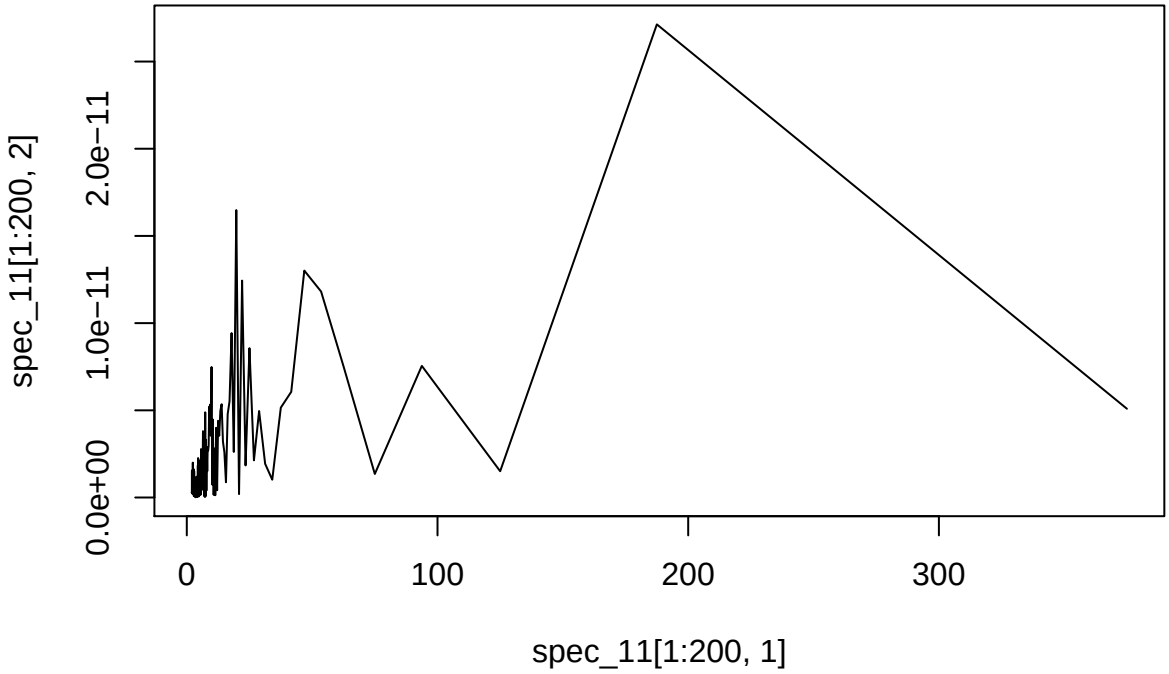

spec_04[1:200, 2]
0.0e+00
4.0e−12
8.0e−12
1.2e−11
0
100
200
300
spec_04[1:200, 1]

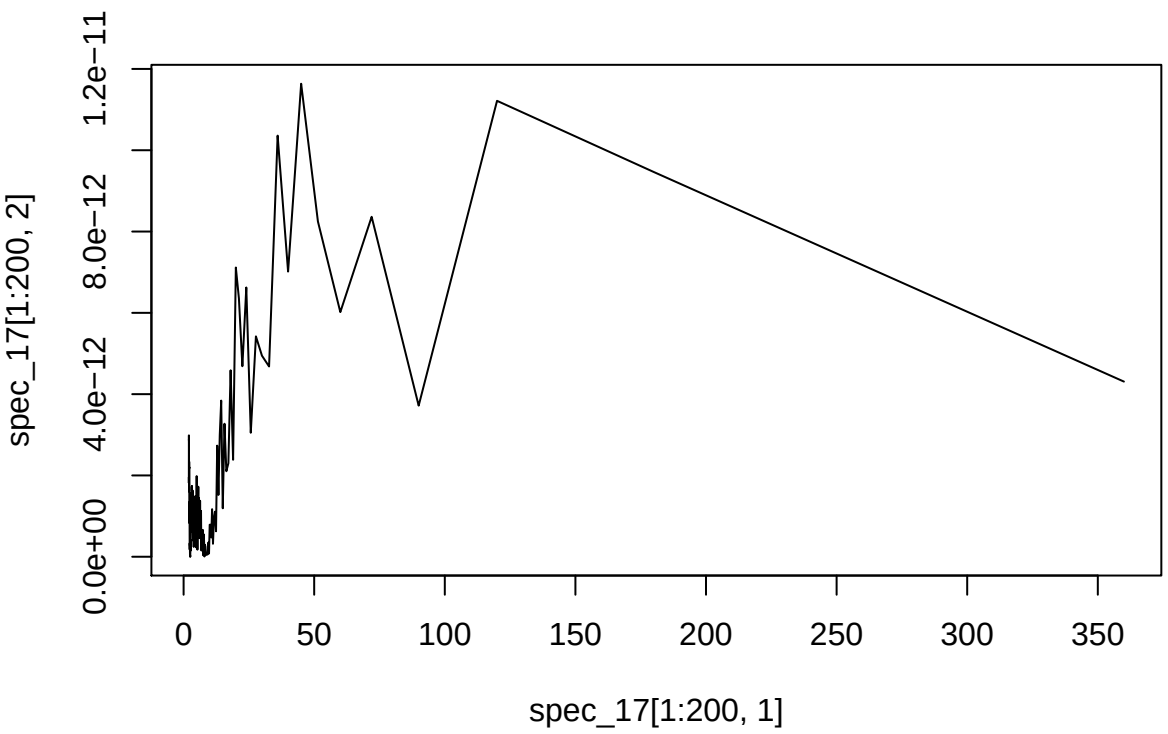

spec_18[1:200, 2]
0.0e+00
1.0e−14
0
50
100
150
200
250
300
350
spec_18[1:200, 1]


spec_19[1:200, 2]
0.0e+00
1.0e−14
2.0e−14
0
100
200
300
spec_19[1:200, 1]


spec_03[1:200, 2]
0.0e+00
4.0e−11
8.0e−11
1.2e−10
0
50
100
150
200
250
300
350
spec_03[1:200, 1]


spec_06[1:200, 2]
0e+00
4e−13
8e−13
0
50
100
150
200
250
300
350
spec_06[1:200, 1]


spec_11[1:200, 2]
0.0e+00
1.0e−11
2.0e−11
0
100
200
300
spec_11[1:200, 1]



In the first 4 graphs, we can see that frequencies which are less than 50 days are indeed significant. Contrary, graphs 5,6,7,8 have frequencies around 50+ days which are more significant than the shorter frequencies.

It is important to note that all years share some sort of seasonality which repeats every 100 days and goes up to around 200 days as well. Moreover, all years share very short frequencies as well, due to the low frequency Sun state which dominates in terms of duration.

Takeaways:

- Pinpointing exact seasonal components of the X-ray flux data is currently not achievable. We are able to precisely pinpoint common seasonal components in the years in which solar flare occured versus the years where there were no solar flares. Whether these blips are connected to the frequencies we are seeing in the spectrographs is yet to be determined.
- One drawback from this approach is that the spectrographs assume that frequencies happen through out the year. So, if we find a frequency of 50 days to be relevant to the build-up blips we are seeing in the raw data; there is not clear answer when those 50 day-occurring blips start and when they lead to the "final" biggest blip which we define as a solar flare.

Question 3: Fitting and Forecasting using an ARMA model

We will fit the model to the daily data and try to see whether our model is a good fit and helps us to forecast or not. As the data was not stationary, we will use the first order difference data while fitting the model,

Model Selection:

Method 1: Fitting Basic ARMA Model From the PACF plot we saw that the optimal order of the model for daily coefficient p is 2. We couldn't determine the order of q. First fitting a basic model without seasonality. Finding optimal AIC through grid search.

	MA0	MA1	MA2	MA3	MA4	MA5
AR0	9909.56	9760.72	9740.84	9709.51	9702.15	9659.42
AR1	9778.39	9435.93	9413.79	9414.13	9416.11	9408.15
AR2	9758.62	9412.94	9413.39	9413.57	9415.54	9409.93
AR3	9728.18	9413.92	9413.59	9415.77	9415.42	9417.38
AR4	9725.74	9415.78	9415.47	9412.24	9403.11	9404.60
AR5	9718.76	9408.27	9398.68	9411.27	9404.59	9406.57

As per AIC Log Likelihood test based on Wilk's approximation, the optimal model is AR2 and MA1 with an AIC of 9421.

```
##
## Call:
## arima(x = log(xray_data_day$flux), order = c(2, 1, 1))
##
## Coefficients:
##          ar1     ar2      ma1
##       0.7718  0.0639  -0.9842
## s.e.  0.0129  0.0128   0.0027
##
## sigma^2 estimated as 0.2622:  log likelihood = -4702.47,  aic = 9412.94
##
## Training set error measures:
##                        ME      RMSE       MAE         MPE     MAPE      MASE
## Training set -0.004165812 0.5119774 0.3592255 -0.09015376 2.478642 0.9949701
##                      ACF1
## Training set 0.001536339
```

Before conducting any further analysis, we want to see whether adding annual components in the model makes an improvement the AIC or not.

Method 2: Fitting ARMA Model with Fourier Transforms as Exogenous Regressors From the spectrogram we saw that there was a frequency corresponding to an annual cycle. There were also other dominant frequency components corresponding to multi-year cycles. Hence, we will need to come up with a way which accounts for all these different frequency components while fitting the model. By design, SARMA model only allows us to define one periodicity parameter. The models can't account for multiple periodicities. Also, the SARMA model doesn't allow us to use long periods. We want to model periodicity corresponding to multiple years from the daily data which is impossible to do so using simple SARMA model.

In order to account for these problems, we need to create a new time series object from the existing one. The new object is divided on the basis of defined frequency such that one period contains the number of observations defined in frequency. We can define the maximum period of interest using frequency parameter in ts function. We then apply Fourier transforms to the time series and get cosine and sine components corresponding to the maximum period and smaller periods contained within and use those as exogenous regressors.

We have a data worth 20 years so let us try exploring seasonality for 6 years and 3 years. In order to do so, we set K = 2.

Let us use the optimal ARMA model obtained in method 1 and use fourier transform with K = 2 as exogenous regressor,

```
##
## Call:
## arima(x = ops.ts, order = c(2, 1, 1), xreg = fourier(ops.ts, K = 2))
##
## Coefficients:
##          ar1     ar2      ma1  S1-2190  C1-2190  S2-2190  C2-2190
```

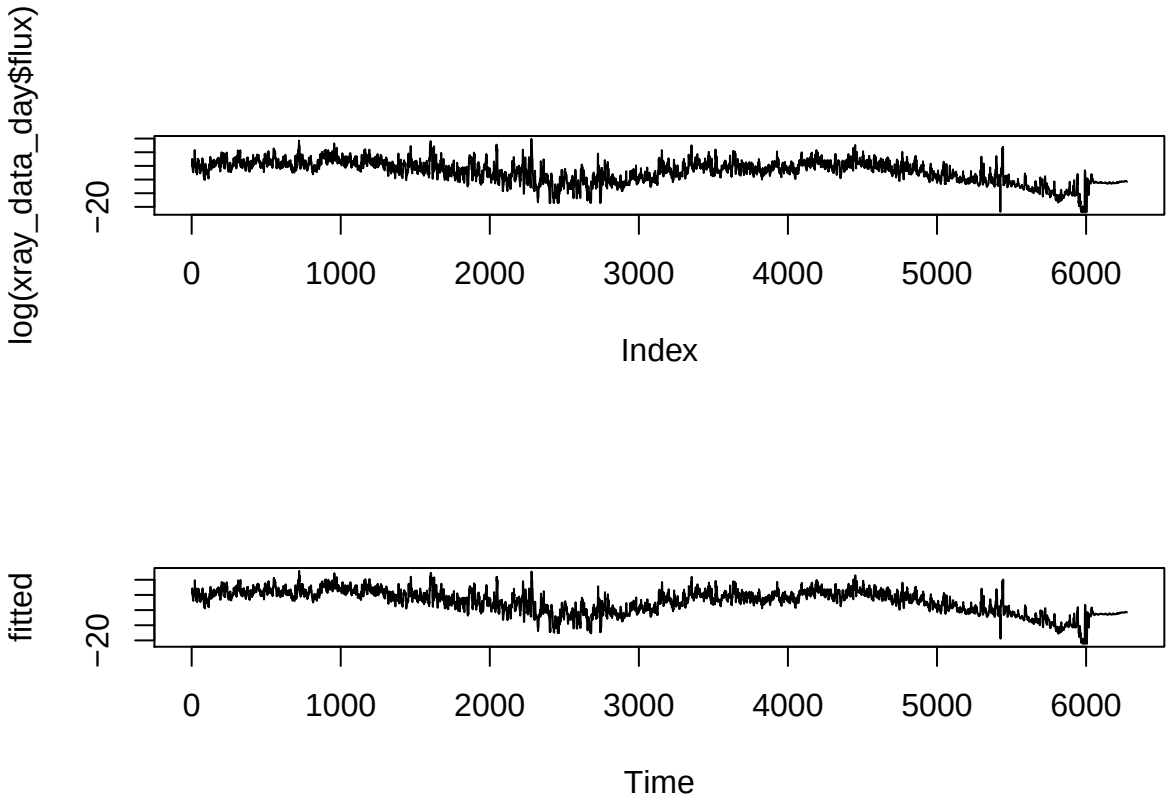
```
##       0.7733  0.0645  -0.9856  -0.0233   0.0834  -0.1034   0.2827
## s.e.  0.0129  0.0128   0.0025   0.2990   0.2877   0.1540   0.1545
##
## sigma^2 estimated as 0.262:  log likelihood = -4700.59,  aic = 9417.19
##
## Training set error measures:
##                         ME      RMSE       MAE         MPE     MAPE      MASE
## Training set -0.003930947 0.5118219 0.3589379 -0.09229453 2.476908 0.9941734
##                      ACF1
## Training set 0.0007316442
```

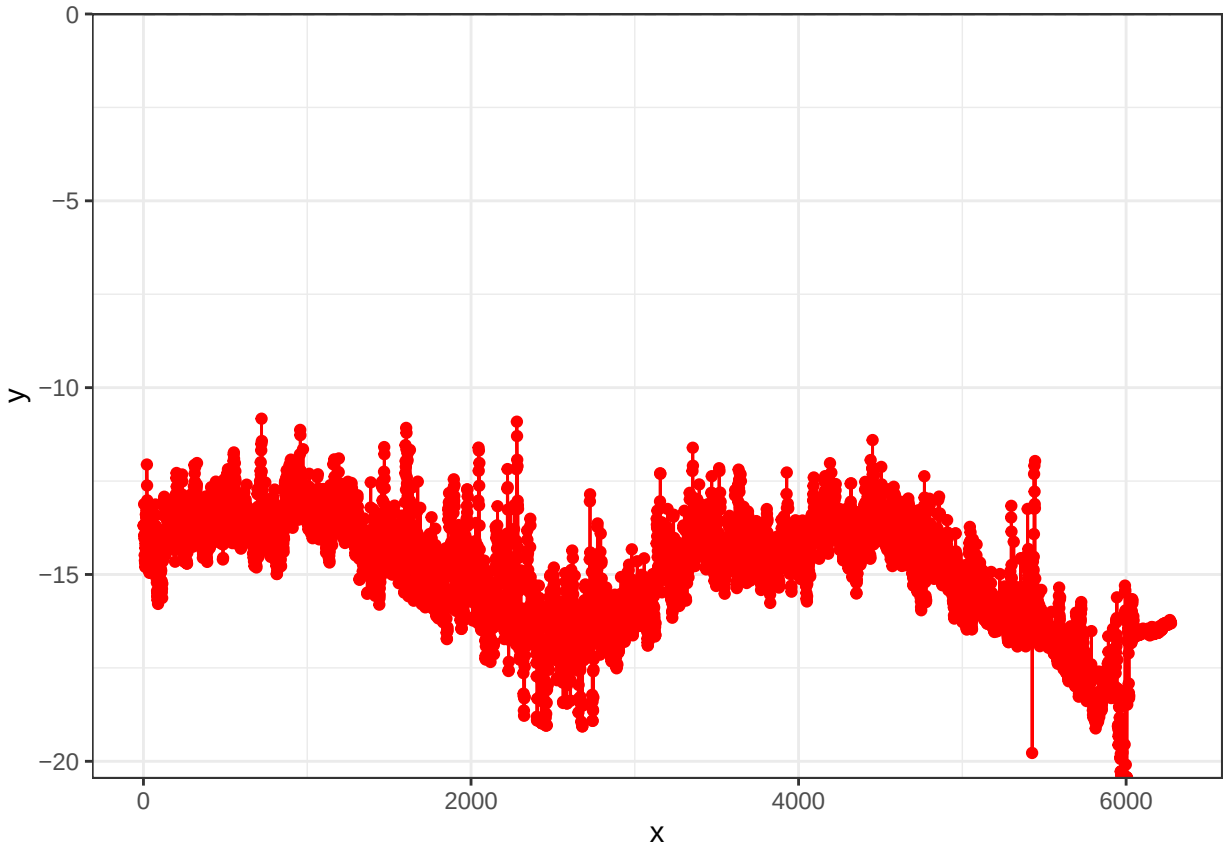
We see that the AIC got worse than method 1. Hence, out of 2 models we choose the model obtained by method 1. We will not conduct any other analysis for this model as we are choosing model 1 to be our final model.

Model Fitting and Goodness of Fit

Fitted vs Actual Values

Plotting fitted values vs actual values,

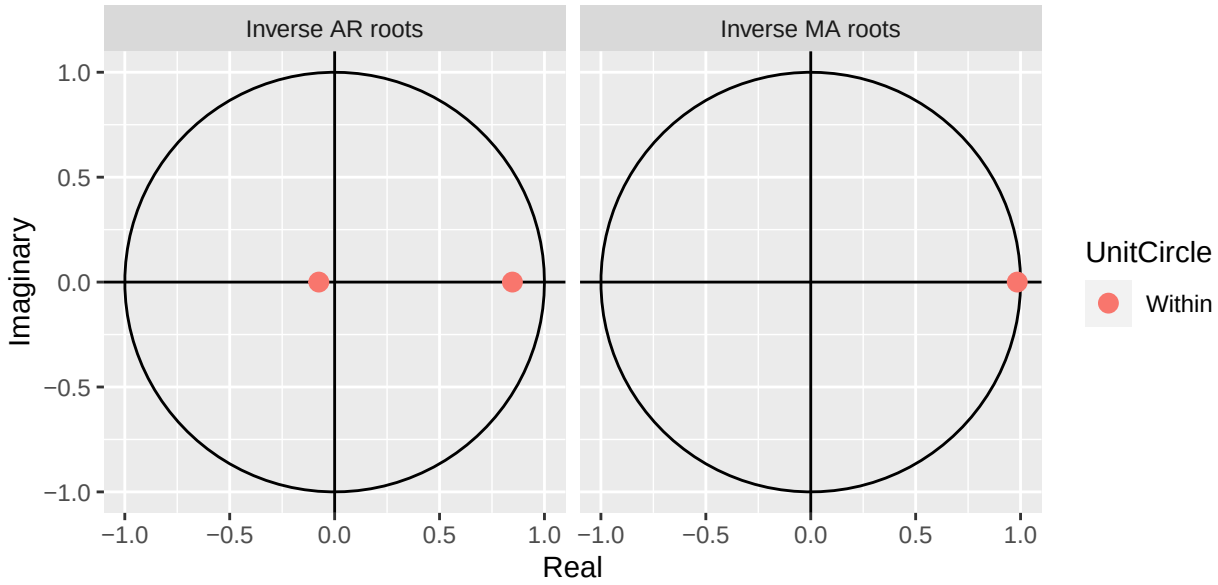




We can see that the model fits the data perfectly. This can lead to overfitting and pose problems in forecasting but that is something to be seen later.

Checking for Causality and Invertibility

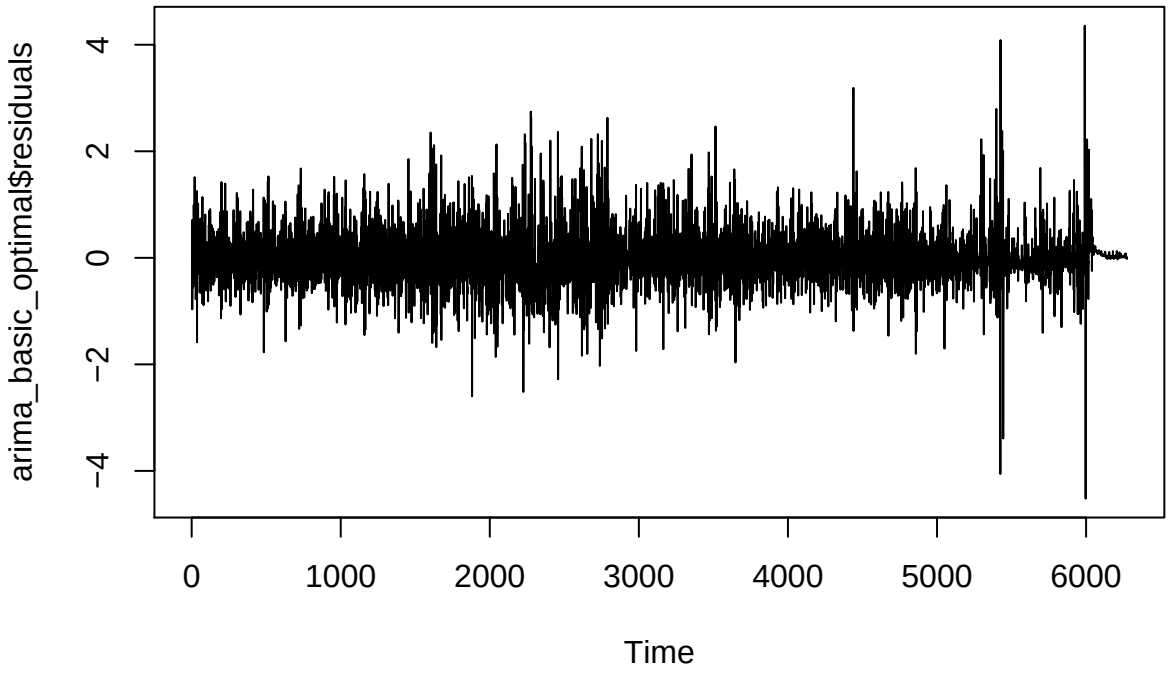
The following plot is to visualize the inverse roots of ARMA polys. As such, for invertible and causal models, the inverse roots should be inside the unit circle.

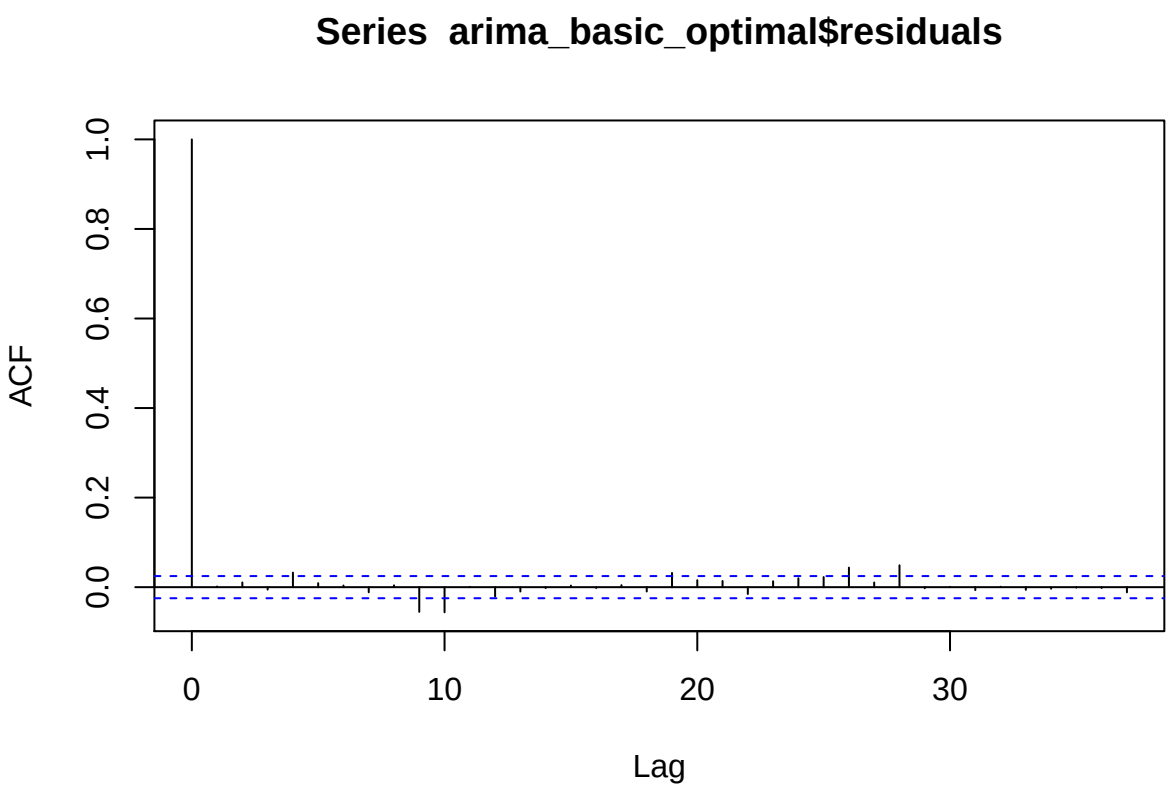


The AR roots are outside the unit circle but the MA roots are at the verge of non-invertibility. This is a bit concerning and might be a sign of model miss-specification. Let us explore the fit further by analyzing residuals.

Exploring Residuals

Looking at th e RMSE and MAE values, the model seems like a reasonable fit to the data. Checking ACF of residuals to see if there is still any unexplained temporal component left





The residuals resemble white noise and there doesn't appear to be any problem.

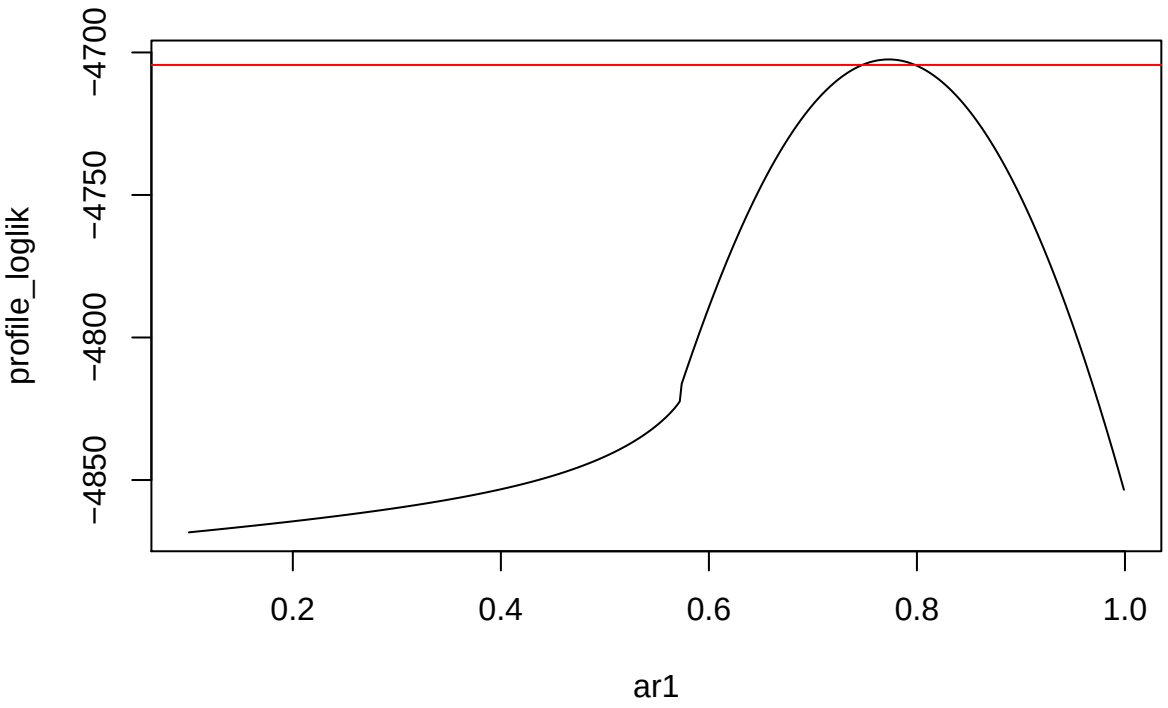
Checking 95% Confidence Intervals for the for Models Coefficients

Before forecasting, we need to check the confidence interval of the coefficients to see whether the model will be able to make reliable predictions or not.

Here are the Confidence Intervals derived from the Fisher Information method.

	2.5%	97.5%
AR1	0.745816	0.796384
AR2	0.039512	0.089688
MA1	-0.989492	-0.978908

Let us validate these confidence intervals through profile likelihood method,



We can see the points where the horizontal line intersects the profile log likelihood curve on both sides represent confidence intervals for AR1. This confidence interval is more or less the same as Fisher Confidence Interval. Hence Fisher Method's confidence interval is also reliable and this might be because we are dealing with a large dataset.

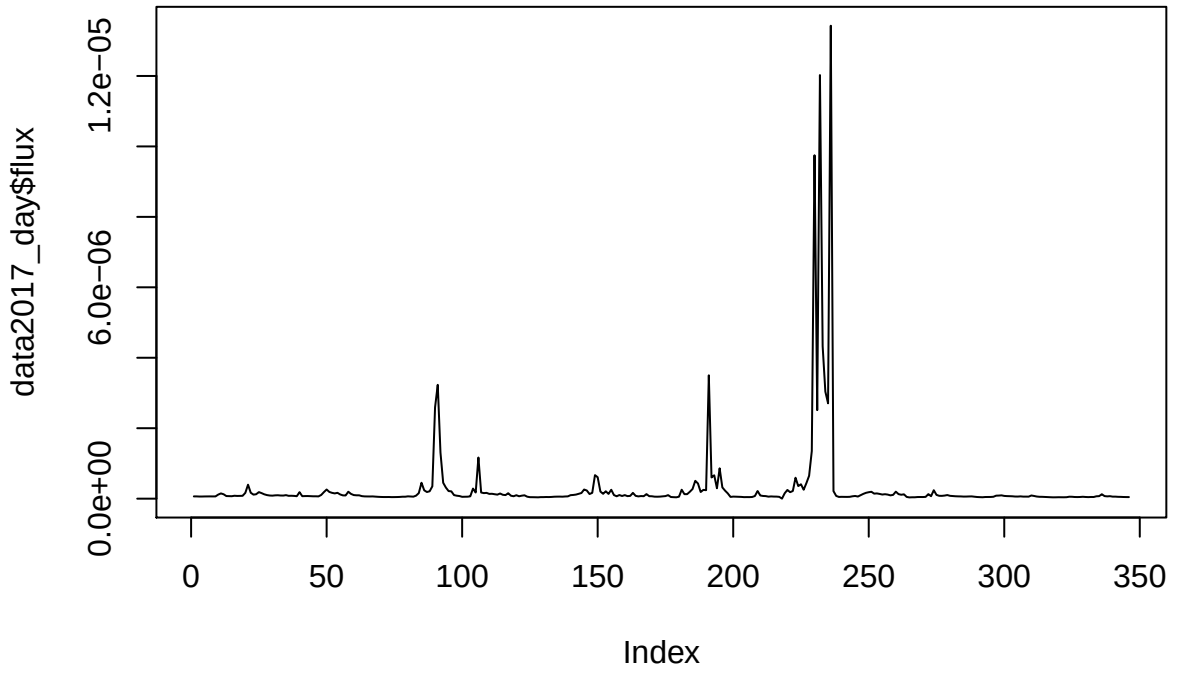
Takeaway:

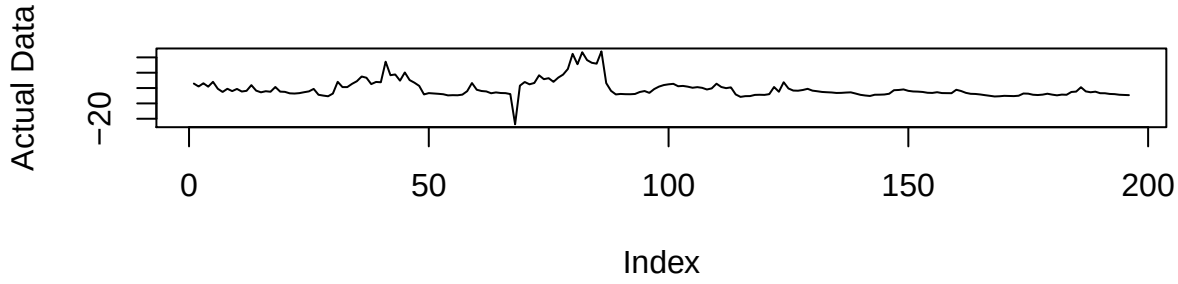
The confidence intervals are not too wide and hence the model can be used for forecasting.

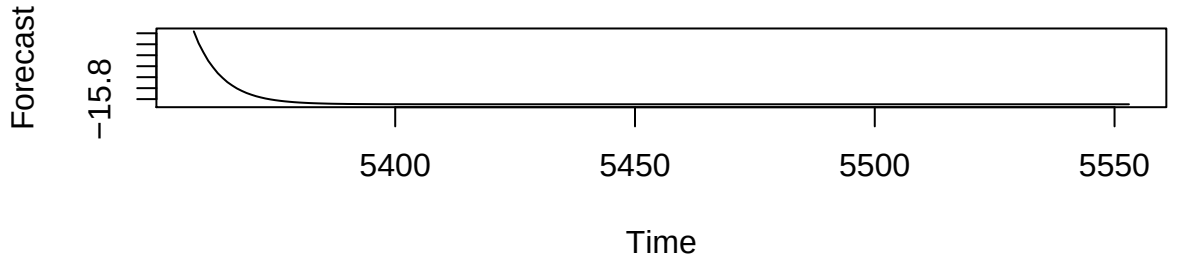
Forecasting the Solar Flares

In order to forecast, we divide our data into training and testing dataset.

For training, we use data from 1999 till mid 2017 and then we forecast for the rest of 2017. The intuition behind this division is that Solar Flares were observed in the second half of 2017. We want to test whether our model is able to predict sudden increase in X-ray intensities corresponding to Solar Flares successfully or not.







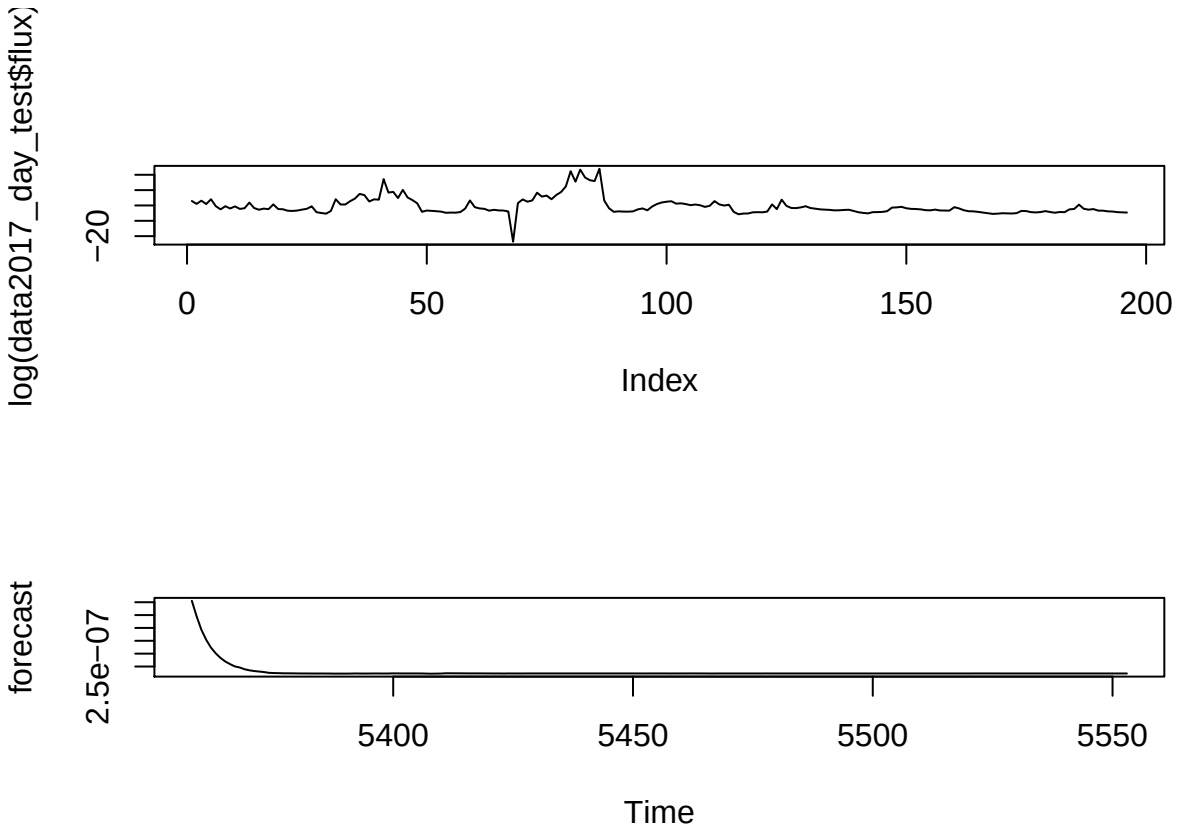

```
## The MAE of forecast is: 211.8505

##
```

```
## The MSE of forecast is: 44880.67
```

This indicates that our model is very poor for forecasting and this is probably because of the overfitting we observed earlier. It lacks predictive power and can't be used for predicting solar flares.

One hypothesis is that our model might not be able forecast because we have not taken seasonal effects into account at all. We can try training different models with varying periodicities of short intervals just to account for short term seasonal components and then forecast individually for all models. The final forecast will be the average of all forecasts.




```
## The MAE of forecast is: 196
```

```
##
## The MSE of forecast is: 38416
```

There is a slight improvement in MAE and MSE but this method also results in poor forecasts.

Conclusion:

- Solar Flares are very rare events and happens once after every few years. The ARMA models are not powerful enough to predict such rare events
- The X-ray Intensities time series under investigation contain multiple significant frequency components. Since using ARMA models, we are unable to model these frequency components simultaneously hence we are unable to capture the true nature of the data through models and end up getting unreliable forecasts.
- It might be possible that non-linear models fit the data better and since ARMA belongs to a linear class of models hence we are not able to get reliable predictions.

Project Novelties:

- The first novelty was the extensive data mining process and scraping 20 years worth X-Ray intensity data from NASA's website. We wrote a data processing script in which we were able to define the method of aggregation as well as the aggregation granularity.
- We were able to come up with a unique way to correlate Solar Flare's definition with the percentile based aggregation such that they key information isn't lost while aggregating the data.
- We also researched on the build up of Solar Flare and its seasonality.
- We selected the optimal ARMA model on the basis of log likelihood ratio based AIC test.

- We used Fourier transform of the time series analysis as exogenous regressors in the model to account for multiple seasonalities.

Future Work:

- Using State Space models to predict and characterize Solar Flares
- Using volatility as exogenous regressor to account for variability in the X-ray intensities and seeing whether it aids in the prediction of Solar Flares or not
- Combining this model with Convolutional Neural Network models trained on Collection of Sun Images with Solar Flare/Non Solar Flare label.

Reference:

[1] Robert H., David S. Stoffer.Time Series Analysis and Its Applications: With R Examples. Springer, 3rd ed, 2011.

[2] Chen et al, Identifying Solar Flare Precursors Using Time Series of SDO/HMI Images and SHARP Parameters, AGU, 2019

[3] Vlad Landa et al, Low Dimensional Convolutional Neural Network For Solar Flares GOES Time Series Classification, arXiv, 2021.

[4] Sunspots and Solar Flares, https://spaceplace.nasa.gov/solar-activity/en/

[5] What are the different types, or classes, of flares?, http://solar-center.stanford.edu/SID/activities/flare.html

[6] 2019 GOES X-ray flux data, https://satdat.ngdc.noaa.gov/sem/goes/data/avg/2019/

[7] 2019 Sunpot Count data, http://www.sidc.be/silso/datafiles