Spotify Time Series Analysis

Wenxuan Zhang

12/10/2020

1.Abstract

In this project, we will analyze the average energy index of song tracks on Spotify from 1921-2020. We explore the energy characteristic of songs over 100 years and apply time series techniques such as data transformation, model identification, diagnostic checking and data forecasting to analyze the the trend of the energy of song tracks. We discover that the song tracks are getting increasingly energetic over time in general and derive a time series model that allow us to forecast the average energy index in the next following years.

2. Introduction

In this project, we analyze the average energy index of song tracks on Spotify from 1921-2020. This dataset is derived from *kaggle public dataset*. The contributor of this dataset collects the data from Spotify API, where the level of energy of each song track is rated by the algorithm developed by Spotify developers. Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale.

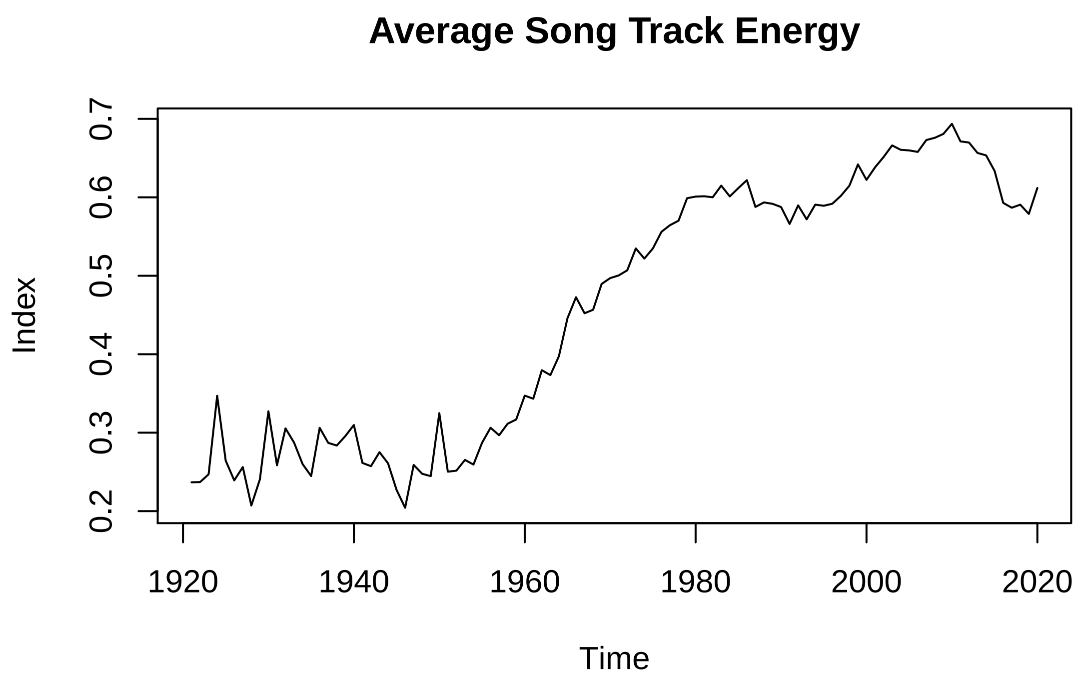
We are interested in analyzing the trend of the level of energy of songs over a large period of time. The analysis is helpful for understanding how music has developed and shaped in the last 100 years. We use various vtime series techniques to detect the trend and seasonality, including graphing histogram, ACF and PACF, as well as spectral analysis. We also performe data transformation, model estimation, and residual analysis for building forecasting model.

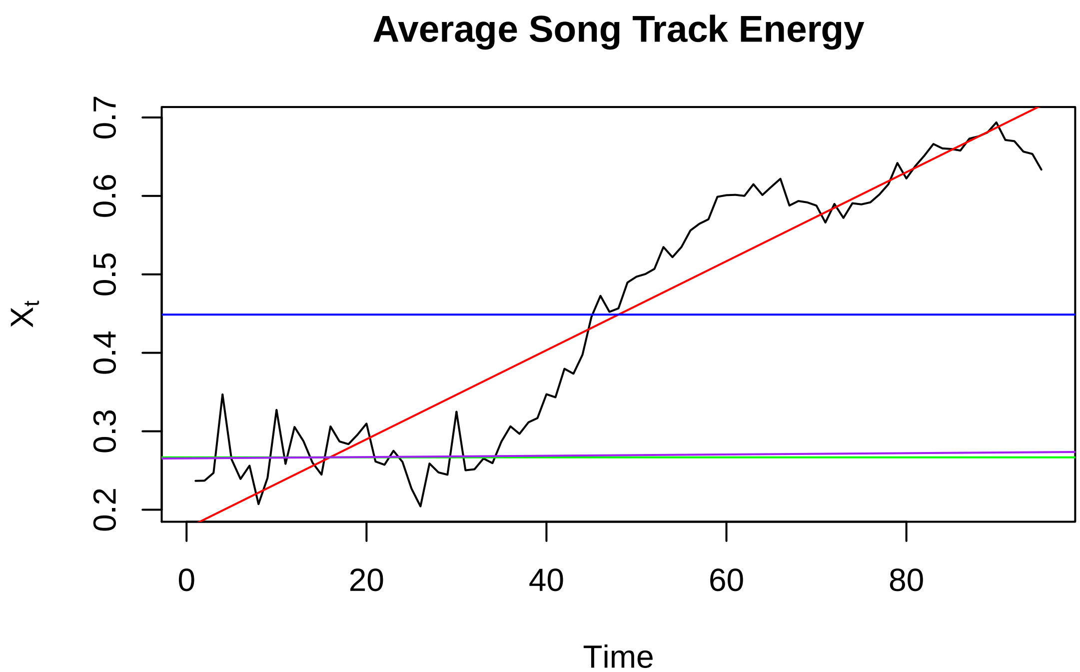
Eventually, we fit the derived model into original dataset to make predictions on the average energy index for incoming years. In general, we are albe to predict the trend of the song track energy and obtain estimates close to the true value.

3. Time Series Analysis

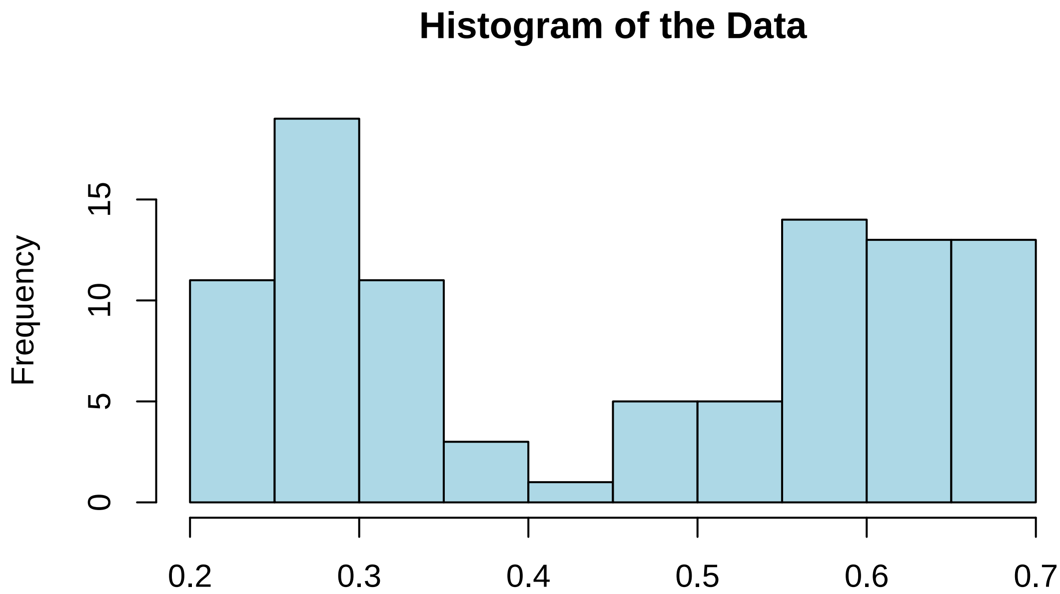
3.1 Exploratory Data Analysis

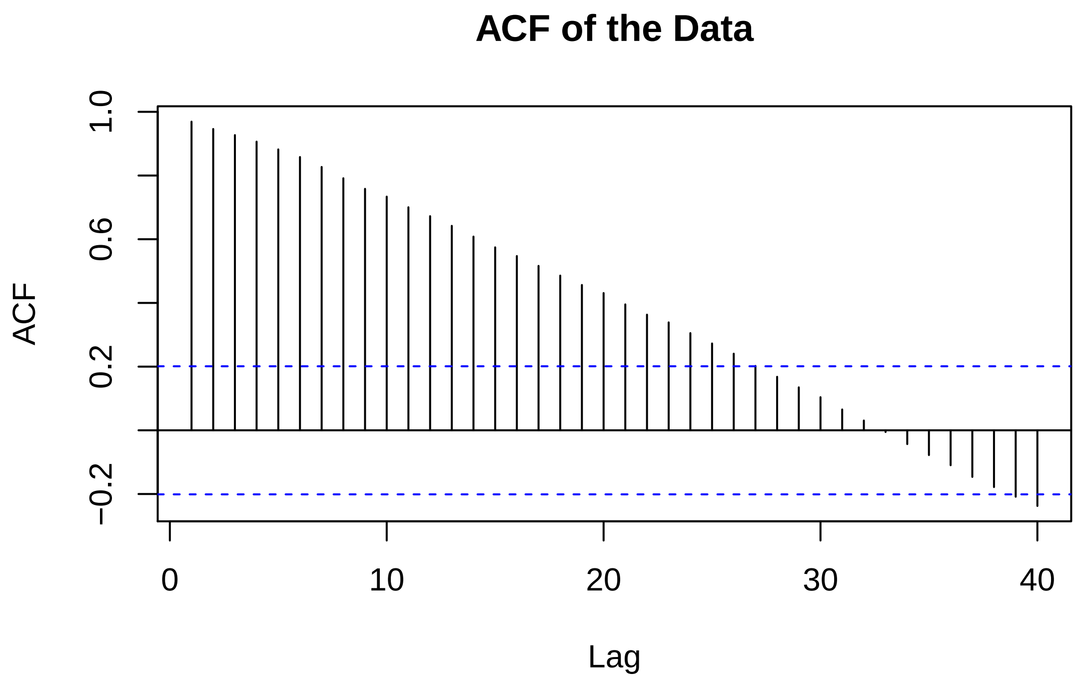
We begin our analysis by plotting the time series and examining the main features of the graph. We plot the average energy index of song tracks over 1921-2020 which gives us 100 observations.





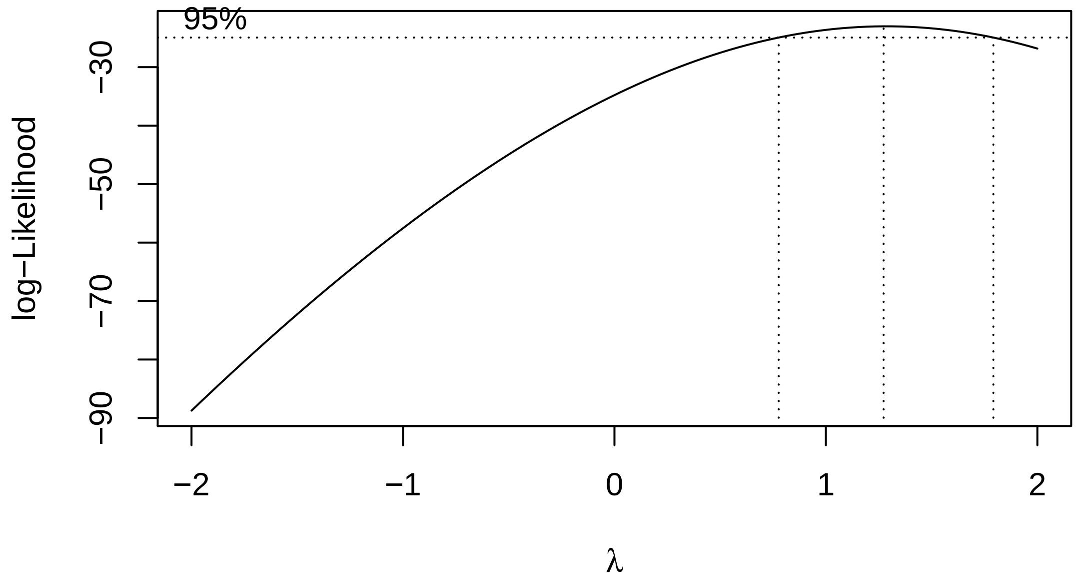
We split our dataset to a trainning dataset with 95 observations and a validation dataset with last 5 observations. Examing the graph from the training set, we see a clear positive linear trend: the average energy index increases over the years. We do not observe any obvious seasonality in the graph. However, the first 30 observation seems to have a more constant mean and smaller variance comparing to the following data.



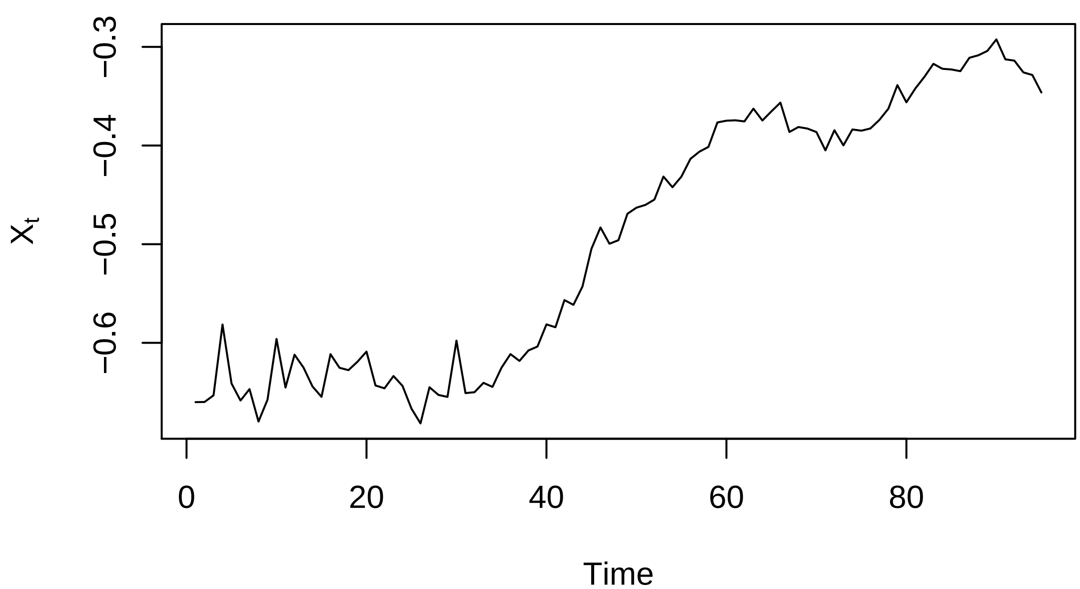


Then, we check the histogram and ACF graph to detect the need for any transformation or differencing. The histogram of the training data is somewhat not symmetric, while the ACF remain large at different lags. Therefore, we decide to use box-cox transformation to further stablize the variance and remove the trend of the data.

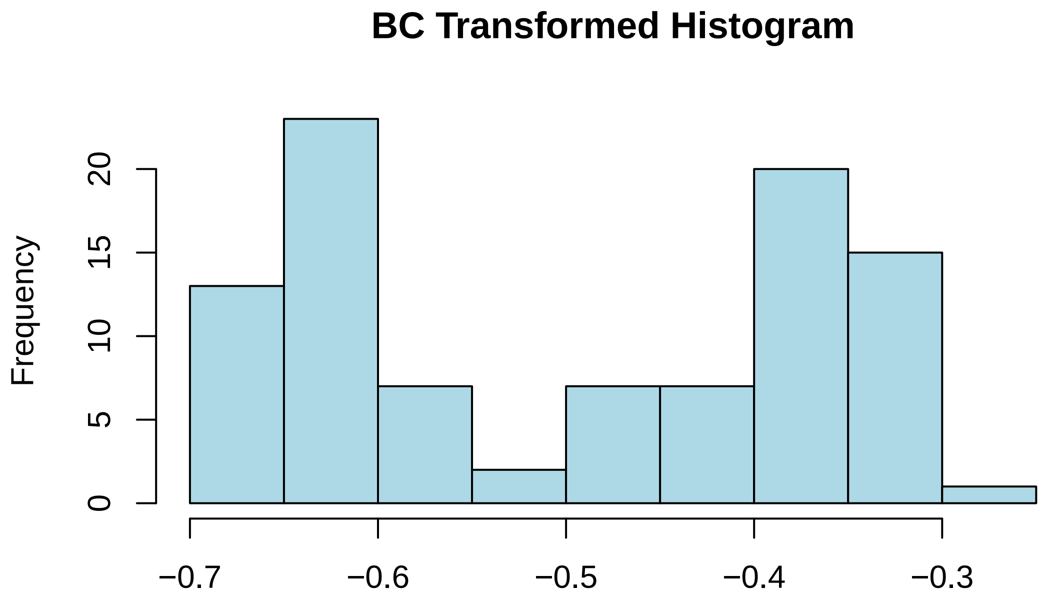
The lambda value we derive from the Box-Cox transformation is 1.27 which is very close to 1, suggesting us keep using the original data.

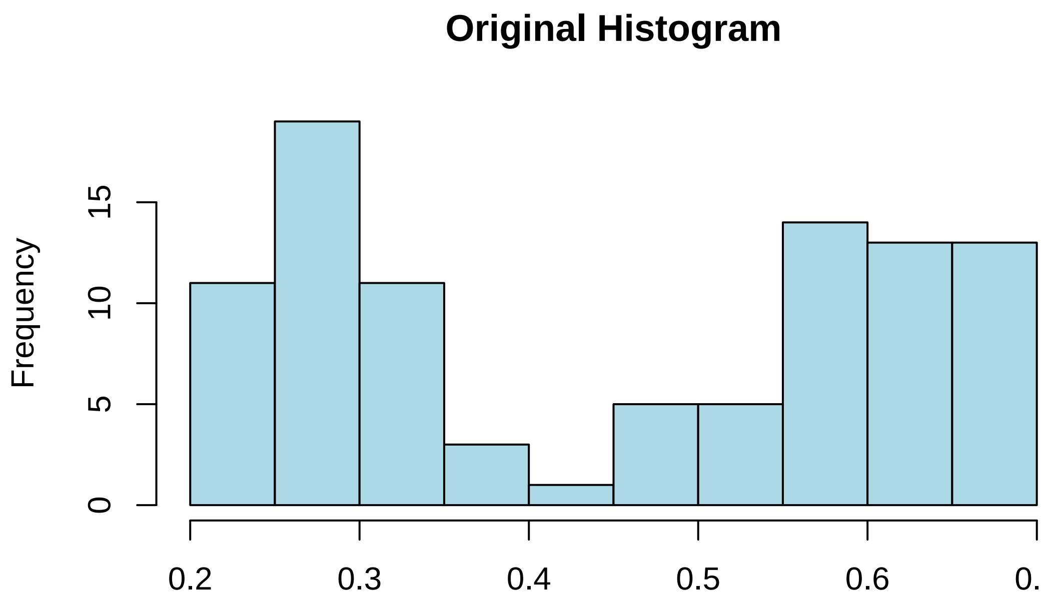


BC Transformed Data

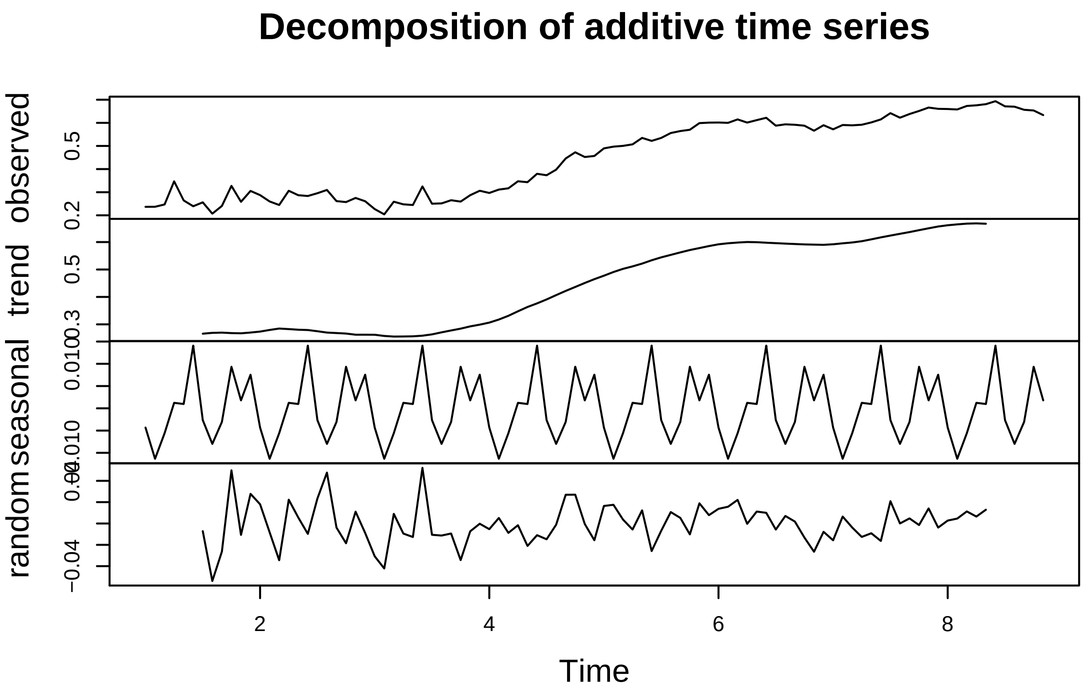


The histogram of the original and transformed data also do not exhibit much difference, and the variance of the original and transformed data is 0.0175 and 0.027 respectively. Since transformation does not lower the variance significantly and 1 is within the confidence interval of lambda, we decide not to transform and use the original data.

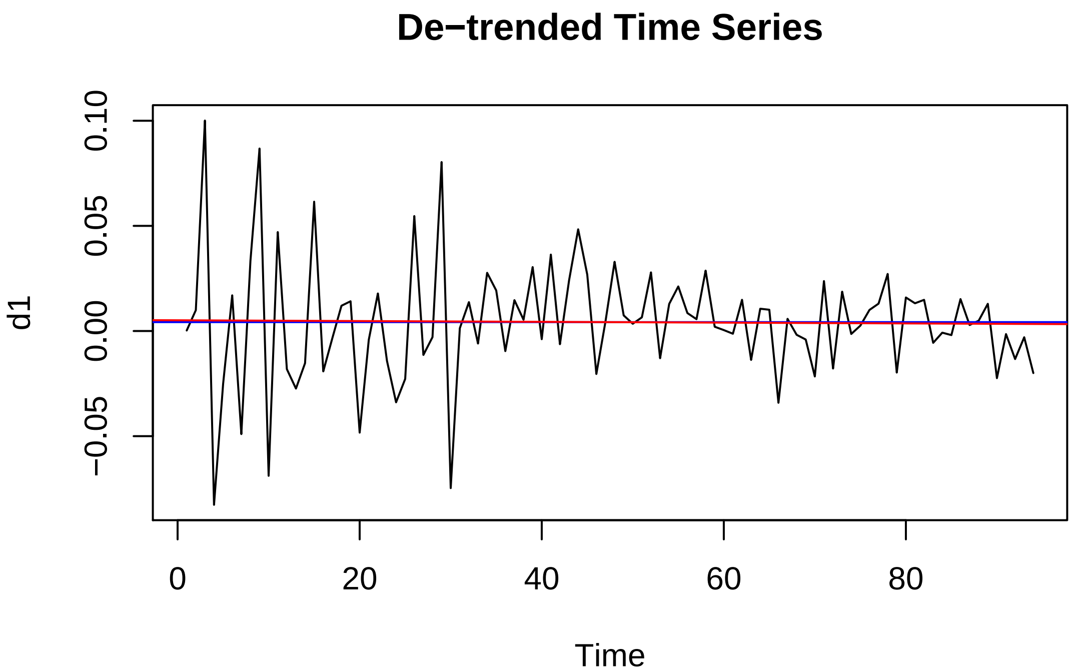




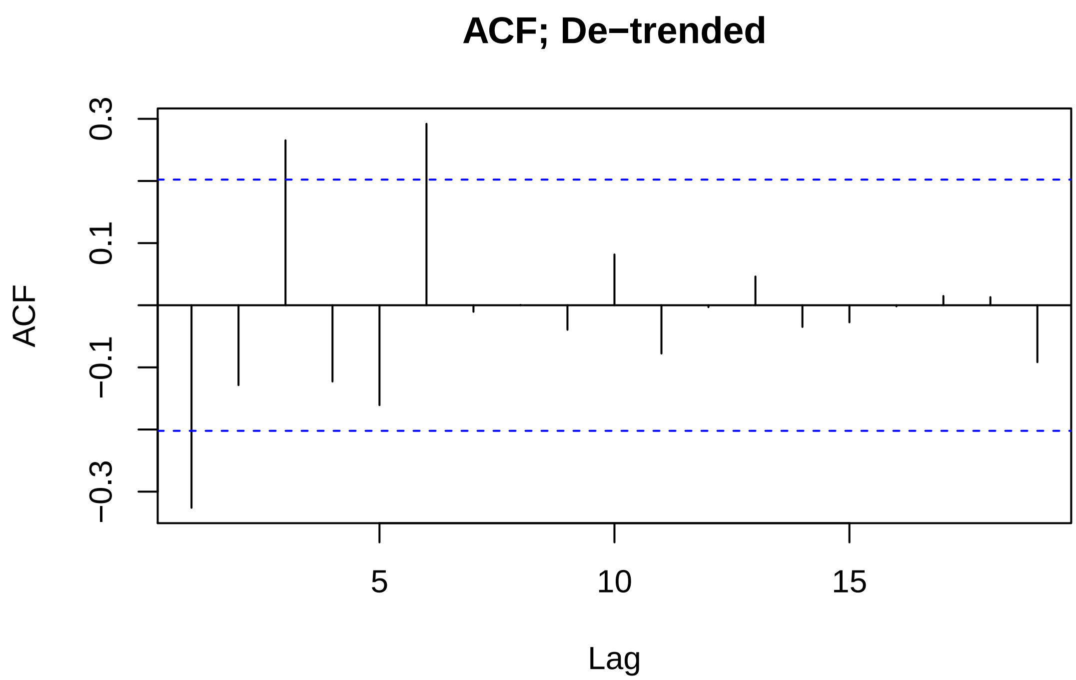
By our previous observation and the decomposition of the data, there is a positive linear trend in our original time series data. To remove the trend, we difference the data and compare the variance before and after.

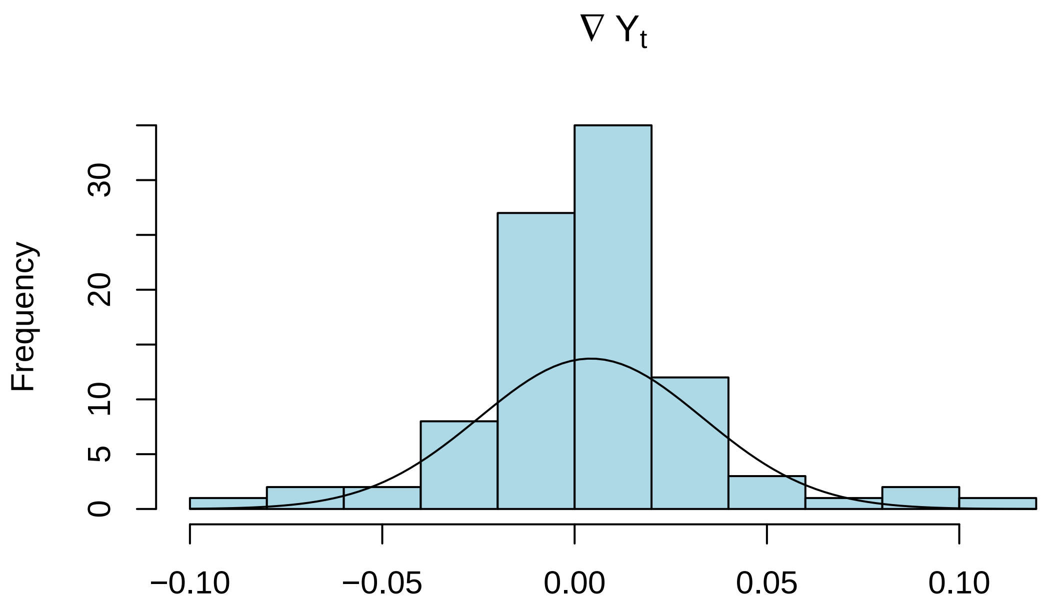


The data differenced at lag 1 yield a variance of 0.00087, which is significantly lower than that of the original data. We plot the de-trended data and the time series exhibit no trend and seasonality.



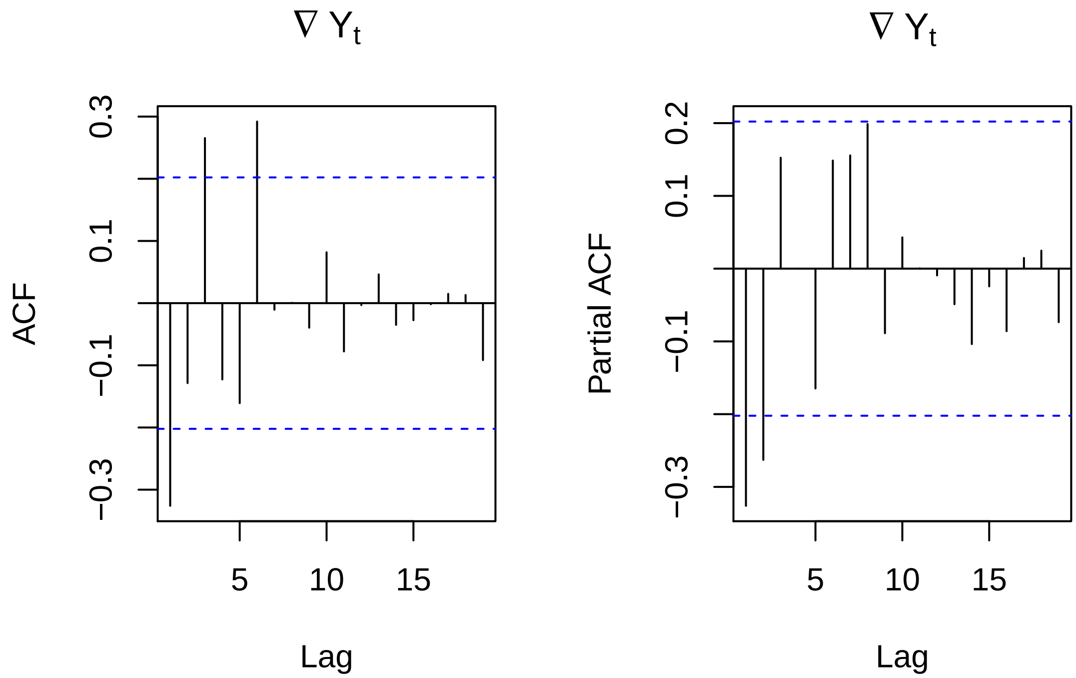
We continoue to check the ACF and histogram of the data. Comparing the original and differenced ACF plot, the ACF of the detrended data decays corresponds to a stationary process. The histogram of the data after differenced at lag 1 also looks symmetric and Gaussian. Therefore, it is appropraite to use our original data differenced at lag 1 to proceed for further model identification.





3.2 Model Identification

We plot the ACF and PACF of our modified data. We find ACF is outside of the confidence interval at lag 1,3,6 and PACF is outside of the confidence interval at lag 1 and 2. Therefore, we suggest ARIMA model with p = 1, 3, 6 and q = 1, 2 to be our candidate models.



3.3 Model Estimation

By computing the AICc value of each candidate models, we get a table of models with each p and q respectively:

```
##      p q      AICc
## [1,] 1 1 -405.0506
## [2,] 3 1 -406.3351
## [3,] 6 1 -409.7722
```

```
##      p q      AICc
## [1,] 1 2 -403.6527
## [2,] 3 2 -407.8522
## [3,] 6 2 -410.0100
```

From the table above, we choose $ARIMA(6,1,1)$ and $ARIMA(6,1,2)$ for further validation, since they have the lowest AICc value. We also check for $ARIMA(3,1,2)$ as it has the lowest AICc value among the rest of the candidates and has less parameters.

First we take a look at $ARIMA(3,1,2)$, we get the coefficients for each parameters after fitting our data into $ARIMA(3,1,2)$. The zero is within the confidence interval of the coefficient of $AR3$, we fix it to be 0 and try the model again. In this way, we get the lowest AICc of $ARIMA(3,1,2)$ to be -409.8522.

```
##
## Call:
## arima(x = d1, order = c(3, 0, 2), method = "ML")
##
## Coefficients:
##           ar1      ar2      ar3     ma1     ma2  intercept
##       -0.6213  -0.7538  -0.0012  0.2696  0.5764     0.0044
## s.e.   0.3851   0.2840   0.1982  0.3663  0.1650     0.0020
##
## sigma^2 estimated as 0.0006477:  log likelihood = 211.41,  aic = -410.82
```

```
##
```

```
## Call:
## arima(x = d1, order = c(3, 0, 2), transform.pars = FALSE, fixed = c(NA, NA,
##     0, NA, NA, NA), method = "ML")
##
## Coefficients:
##           ar1      ar2  ar3     ma1     ma2  intercept
##       -0.6193  -0.7522    0  0.2678  0.5761     0.0044
## s.e.   0.1981   0.1137    0  0.2323  0.1487     0.0020
##
## sigma^2 estimated as 0.0006476:  log likelihood = 211.41,  aic = -412.82
```

```
## [1] -409.8522
```

Then, we perform the same analysis to the model $ARIMA(6,1,1)$ and $ARIMA(6,1,2)$. The lowest AICc we get for $ARIMA(6,1,1)$ is -413.0761; and that of $ARIMA(6,1,2)$ is -413.1787. Comparing the results of the three candidate models, the lowest AICc value of $ARIMA(6,1,1)$ and $ARIMA(6,1,2)$, after fixing some coefficient to be zero, are still very close. However, $ARIMA(3,1,2)$ does not give a lower or close AICc over $ARIMA(6,1,1)$ and $ARIMA(6,1,2)$. Therefore, we proceed $ARIMA(6,1,1)$ and $ARIMA(6,1,2)$ for further examination.

```
##
## Call:
## arima(x = d1, order = c(6, 0, 1), method = "ML")
##
## Coefficients:
##          ar1     ar2     ar3      ar4      ar5     ar6      ma1  intercept
##       0.2433  0.0664  0.1773  -0.1359  -0.0814  0.3621  -0.6102     0.0039
## s.e.  0.1730  0.1137  0.1042   0.1140   0.1113  0.1075   0.1604     0.0026
##
## sigma^2 estimated as 0.0005996:  log likelihood = 214.73,  aic = -413.47
```

```
##
## Call:
## arima(x = d1, order = c(6, 0, 1), transform.pars = FALSE, fixed = c(NA, 0, NA,
##     NA, 0, NA, NA, NA), method = "ML")
##
## Coefficients:
##          ar1  ar2     ar3      ar4  ar5     ar6      ma1  intercept
##       0.2314    0  0.1806  -0.1438    0  0.3437  -0.5834     0.0038
## s.e.  0.1728    0  0.1057   0.1120    0  0.1058   0.1467     0.0026
##
## sigma^2 estimated as 0.0006042:  log likelihood = 214.39,  aic = -416.77
```

```
## [1] -413.0761
```

```
##
## Call:
## arima(x = d1, order = c(6, 0, 2), method = "ML")
##
## Coefficients:
##          ar1      ar2     ar3      ar4      ar5     ar6      ma1     ma2
##       0.5547  -0.1815  0.1031  -0.2577  -0.0066  0.3707  -0.9720  0.4196
## s.e.  0.2594   0.1637  0.1185   0.1409   0.1370  0.1140   0.2783  0.1988
##       intercept
##          0.0038
## s.e.     0.0026
##
```

```
## sigma^2 estimated as 0.00058:  log likelihood = 216.08,  aic = -414.15
```

```
##
## Call:
## arima(x = d1, order = c(6, 0, 2), transform.pars = FALSE, fixed = c(NA, NA,
##     0, NA, 0, NA, NA, NA, NA), method = "ML")
##
## Coefficients:
##          ar1      ar2  ar3      ar4  ar5     ar6      ma1     ma2  intercept
##       0.5342  -0.1713    0  -0.2227    0  0.3637  -0.9717  0.4637     0.0038
## s.e.  0.2633   0.1646    0   0.0980    0  0.0958   0.2999  0.1533     0.0024
##
## sigma^2 estimated as 0.0005848:  log likelihood = 215.66,  aic = -417.32
```

```
## [1] -413.1787
```

At this point, we conclude our candidate model A to be ARIMA(6,1,1):

$$(1 + 0.2314B + 0.1806B^3 - 0.1438B^4 + 0.3437B^6)(1 - B)X_t = (1 - 0.6102B)Z_t$$

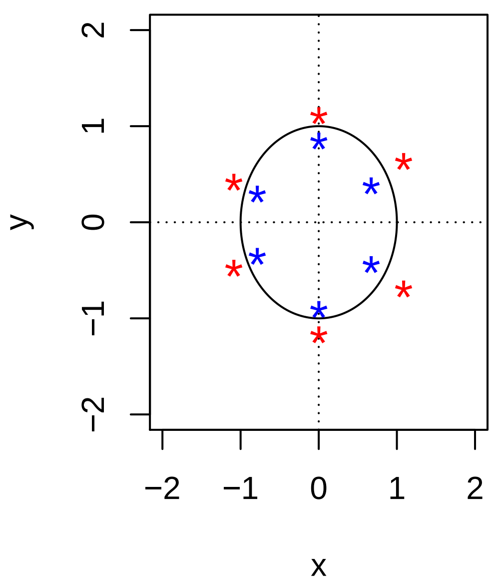
and candidate model B to be ARIMA(6,1,2):

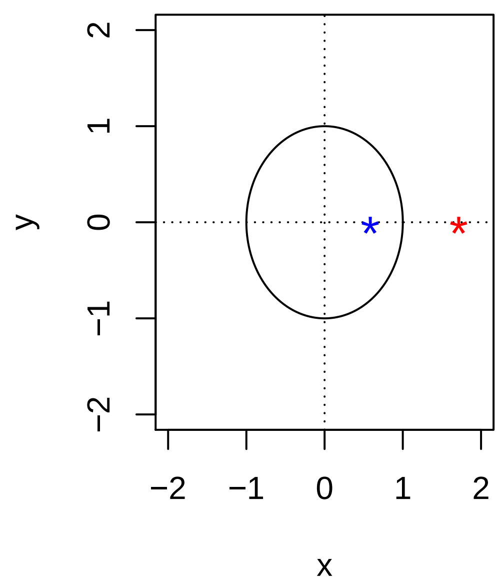
$$(1 + 0.5342B - 0.1713B^2 - 0.2227B^4 + 0.3637B^6)(1 - B)X_t = (1 - 0.9717B + 0.4637B^2)Z_t$$

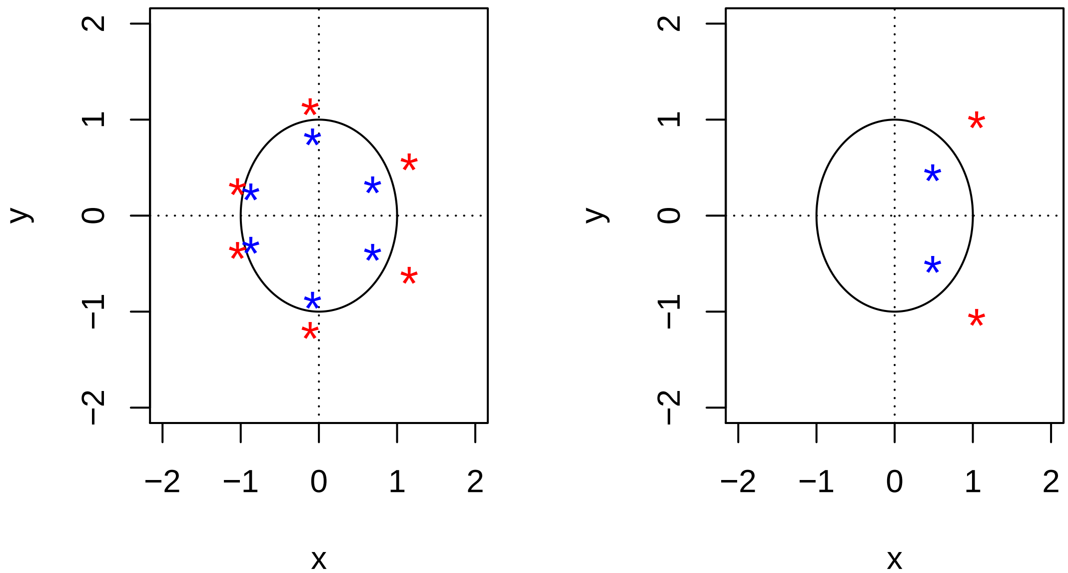
.

3.4 Model Diagnostic

Next step, we check the stationarity and invertibility of both models. We plot the roots of polynomials of both the MA and AR part of the model A. All the roots are outside of the unit circle, which means the model is stationary and invertible. We plot the same graph for model B. The result shows that model B is also stationary and invertible.

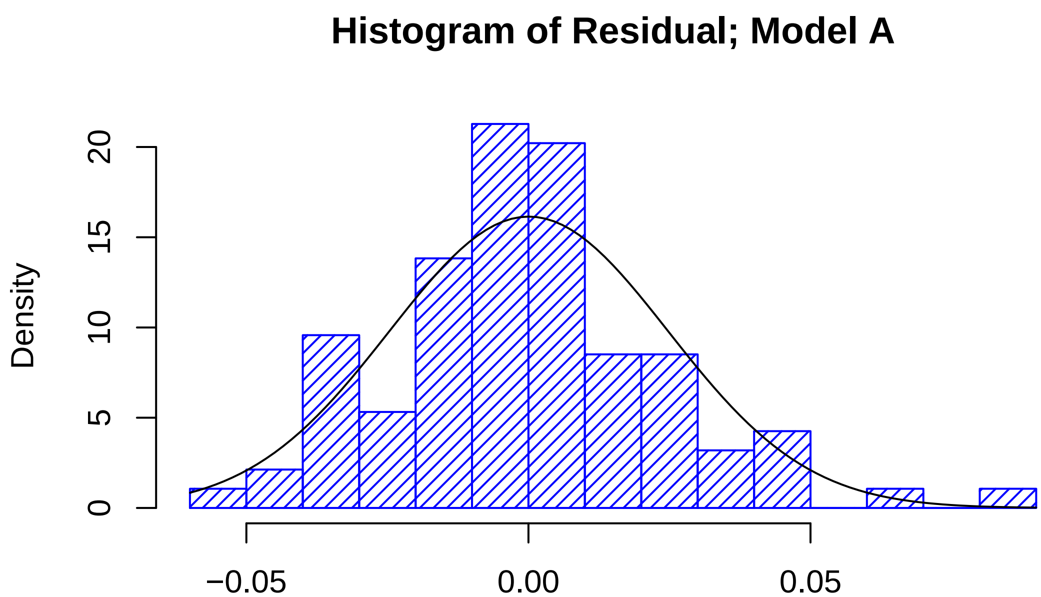


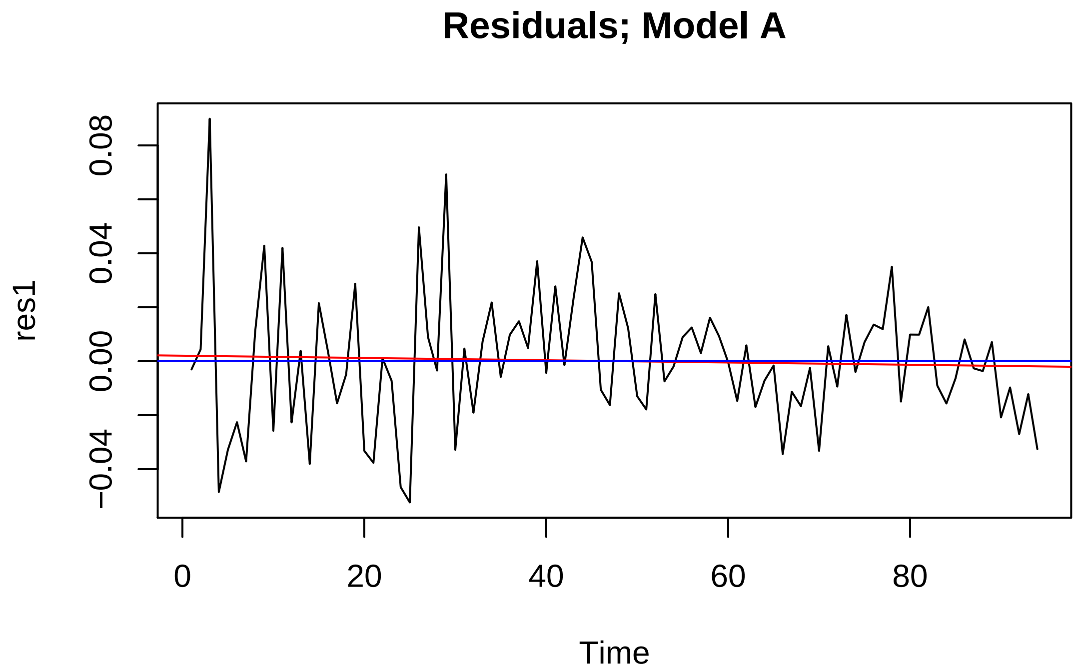


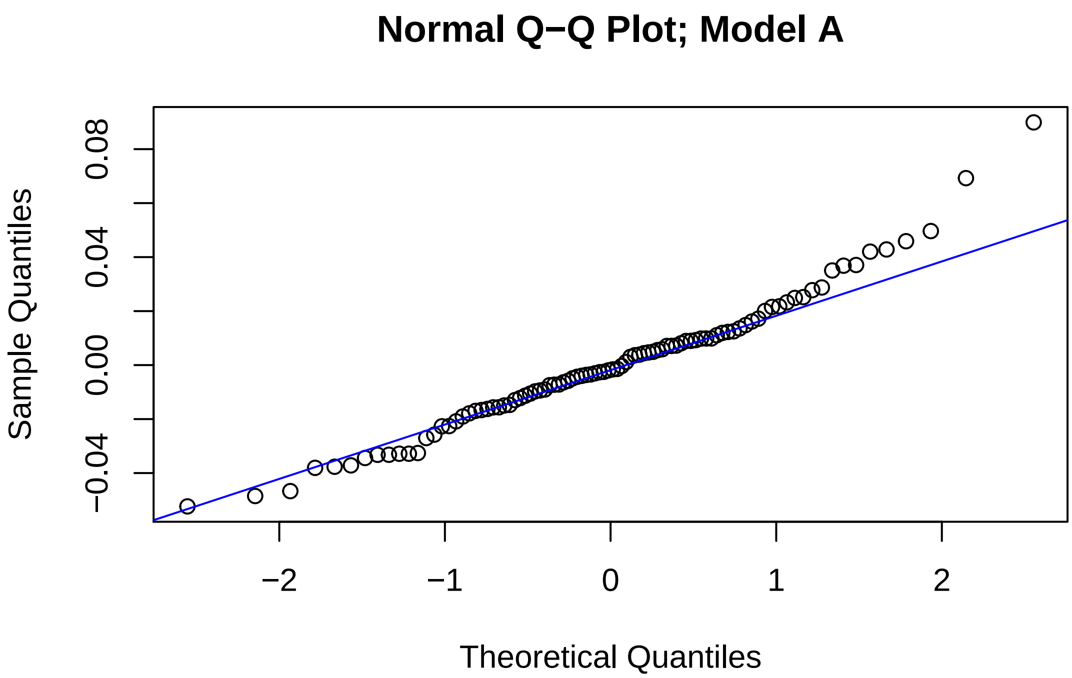


We then and perform more diagnostic checking on the residuals of model A and B. First we look at model A. There is no trend, no seasonality or visible variance change in its residual plot. The histogram is almost Gaussian and the normal Q-Q plot looks good except for 2 data points at the tail. We verify its normality by using Shapiro-Wilk test. And it does pass the test with p-value $0.03914 < 0.05$ at 95% confidence level.

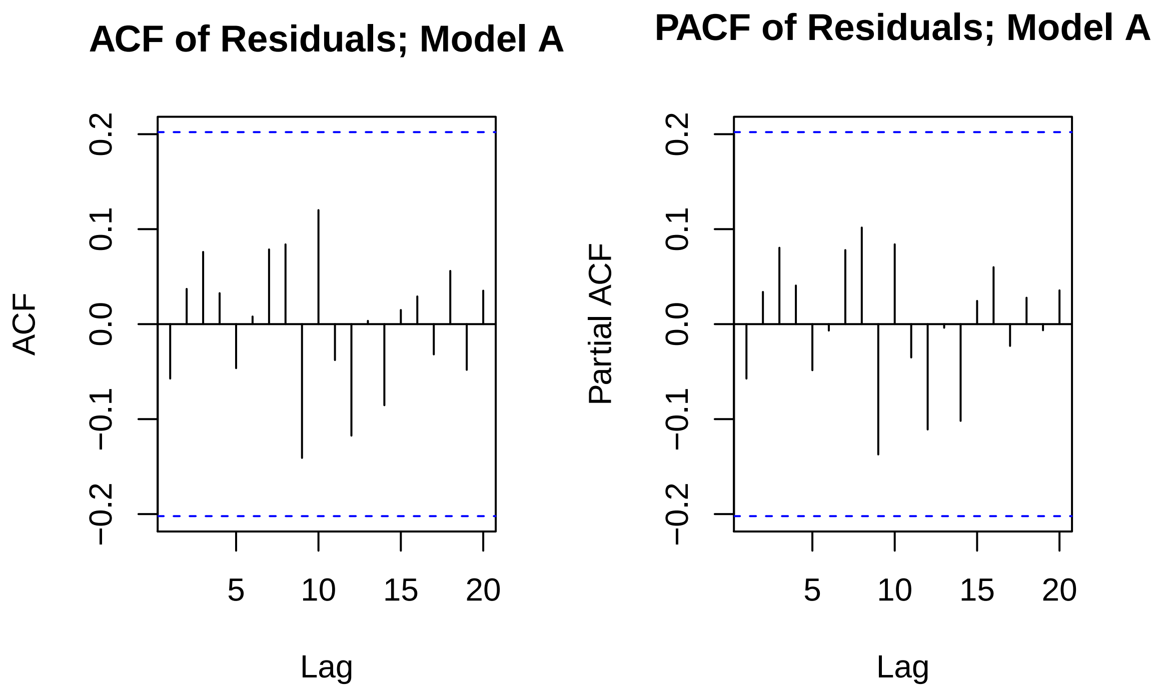
We also perform Box-Pierce test, Box-jung test, and Mcleod-Li test to detect any linear and non-linear correlation between residuals. The model passed three test with p-value: 0.1242, 0.09391, and 0.4296 respectively.







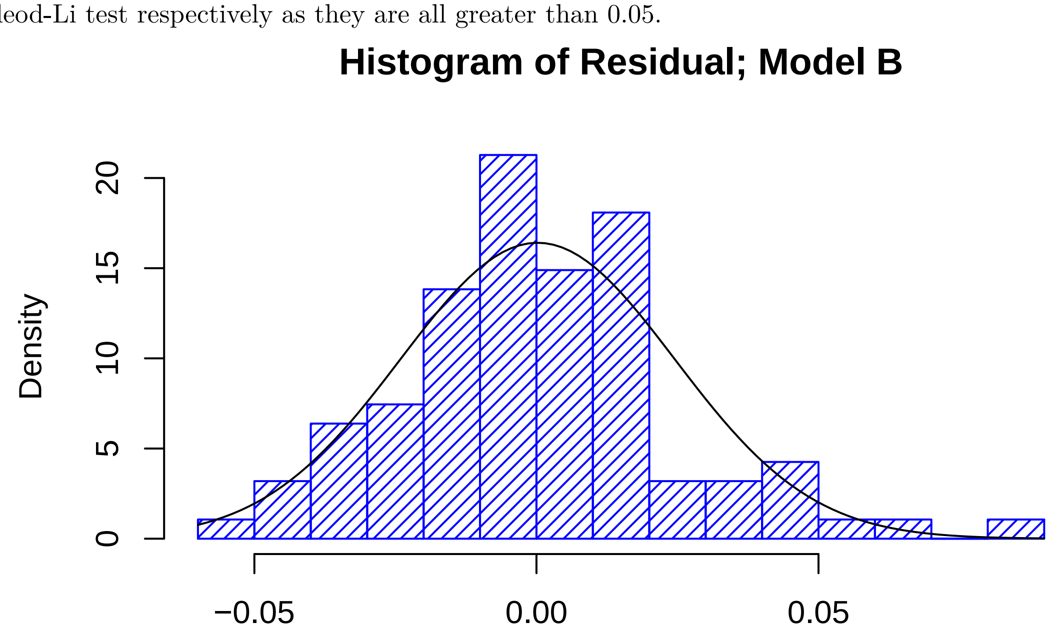
Moreover, we check that all the ACF and PCF of the residuals of model A are within the confidence interval and can be counted as zeros. Fitting residuals to AR(0), we get $\hat{\sigma_z}^2 = 0.0006107$ which means the residuals resemble WN. Even though the other aspect of model A are have good behavior, it does not pass the test for normality of residuals. Thus, model A is not ideal for forecasting.

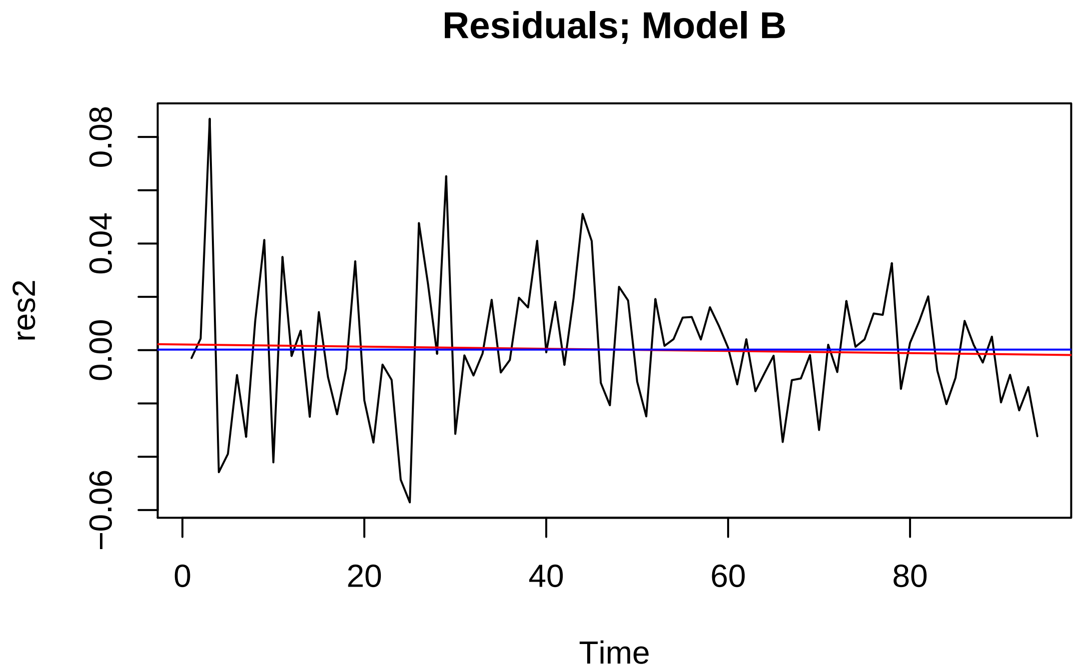


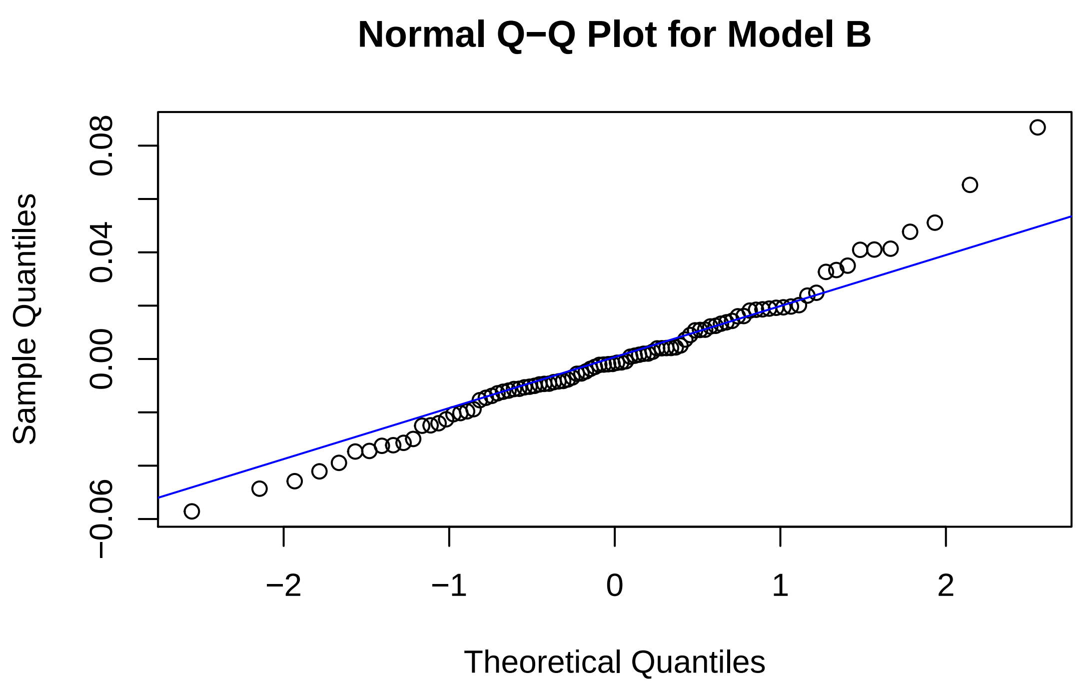

```
##
## Call:
## ar(x = res1, aic = TRUE, order.max = NULL, method = c("yule-walker"))
##
##
## Order selected 0  sigma^2 estimated as  0.0006107
```

We then examine the diagnostic checking results for model B. The plot of residuals for model B does not has trend, seasonality, or sharp change of variance neither. The histogram and normal Q-Q plot looks similar to those of model A, with slightly better shape.

Checking the test results formodel B, it passes Sharpiro-Wilk test with p-value 0.06824, better than model A does. It also passes other three test with p-value 0.3277, 0.2797, 0.3383 for Box-Pierce test, Box-jung test, and Mcleod-Li test respectively as they are all greater than 0.05.







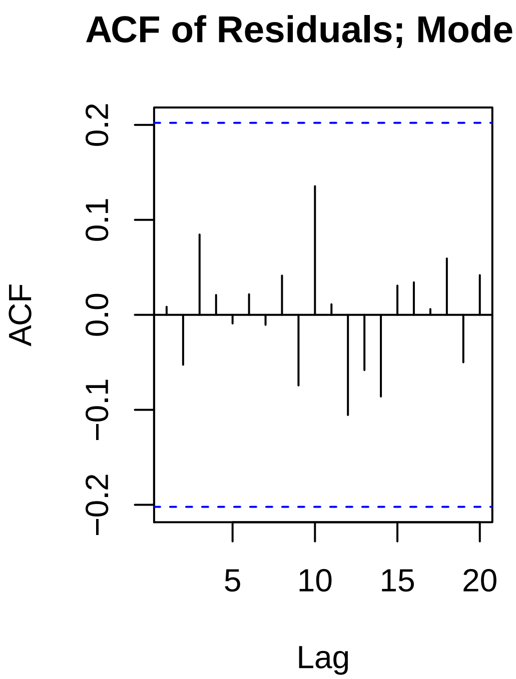

```
##
##  Shapiro-Wilk normality test
##
## data:  res2
## W = 0.97499, p-value = 0.06824
```

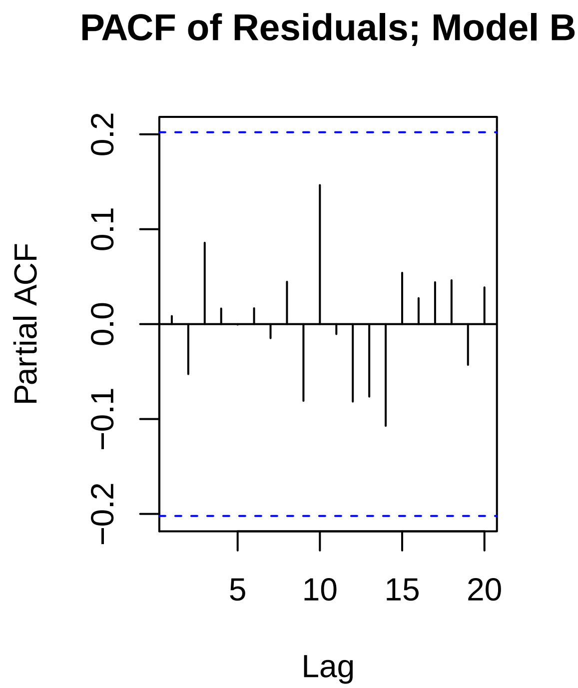
```
##
##  Box-Pierce test
##
## data:  res2
## X-squared = 3.4471, df = 3, p-value = 0.3277
```

```
##
##  Box-Ljung test
##
## data:  res2
## X-squared = 3.8362, df = 3, p-value = 0.2797
```

```
##
##  Box-Ljung test
##
## data:  res2^2
## X-squared = 11.251, df = 10, p-value = 0.3383
```

The ACF and PACF of residuals in model B also fall in the confidence interval which can be treated as zero. And the residuals resember WN with $\hat{\sigma_z}^2 = 0.0005911$.






```
##
## Call:
## ar(x = res2, aic = TRUE, order.max = NULL, method = c("yule-walker"))
##
##
## Order selected 0  sigma^2 estimated as  0.0005911
```

To summarize, model B perform better than model A. Model B has a slightly lower AICc than model A, but the difference is small (0.1); while Model B posses diagnostic tests with a bit higher p-value over all and possesses somewhat better shape in the residual histogram adn normal Q-Q plot. And most importantly, model A does not pass Sharpiro-Wilk test for normality, while model B passes all diagnostic test. Therefore, it is reasonable to determine model A ARIMA(6,1,2):

$$(1 + 0.5342B - 0.1713B^2 - 0.2227B^4 + 0.3637B^6)(1 - B)X_t = (1 - 0.9717B + 0.4637B^2)Z_t$$

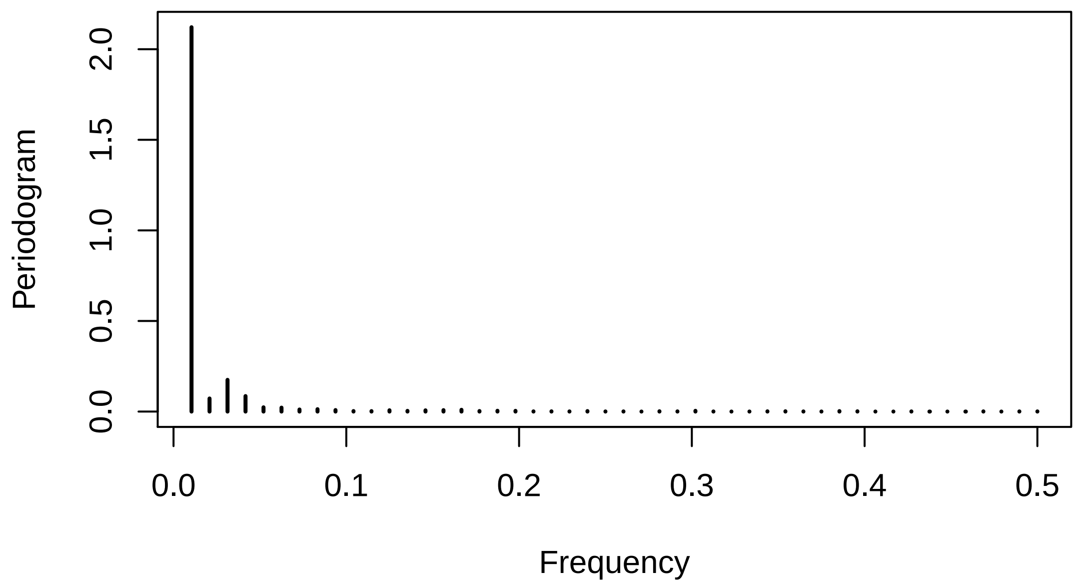
for data forecasting.

4. Spectral Analysis

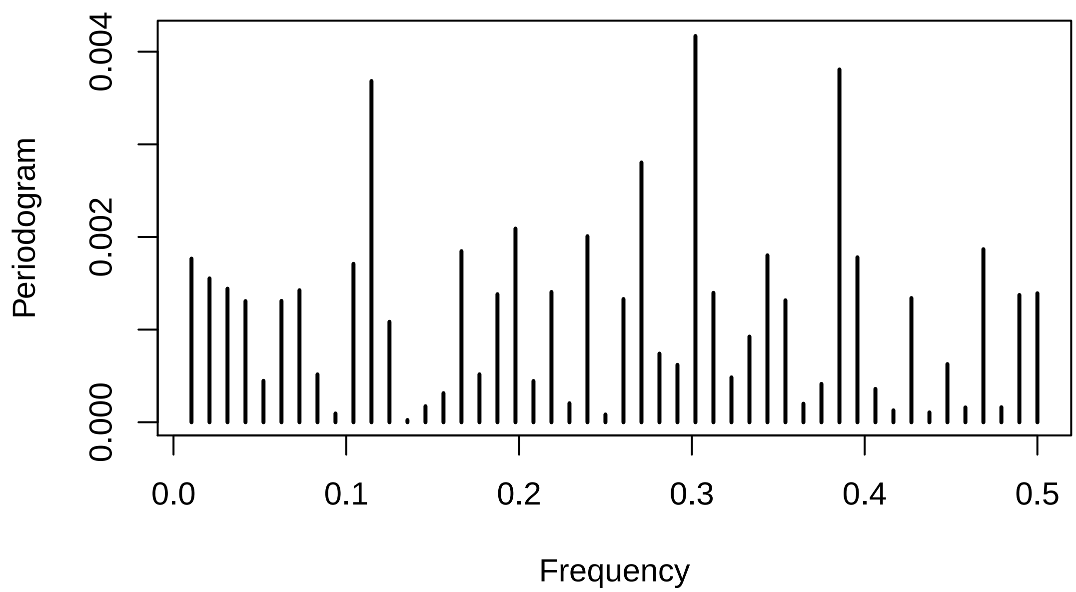
We conclude our data is free of seasonality by examing original data and ACF of differenced data. Before we make final data forecasting, we check the seasonality again with spectral analysis in case there is undetected seasonality within the data.

We plot the periodogram for the data and residuals of our model. and does not detect any frequencies. There seems to have spikes at 0.01 and 0.03, however, the corresponding period indicated by this frequency would be 100 and 33 years, which is not reasonable since we only have data for past 100 years. Therefore, this does not provides us extra insight on the seasonality of the data. The periodogram of the residuals does not have a dominant frequency neither.

```
require(TSA)
TSA::periodogram(df.train)
```




```
TSA::periodogram(res2)
```

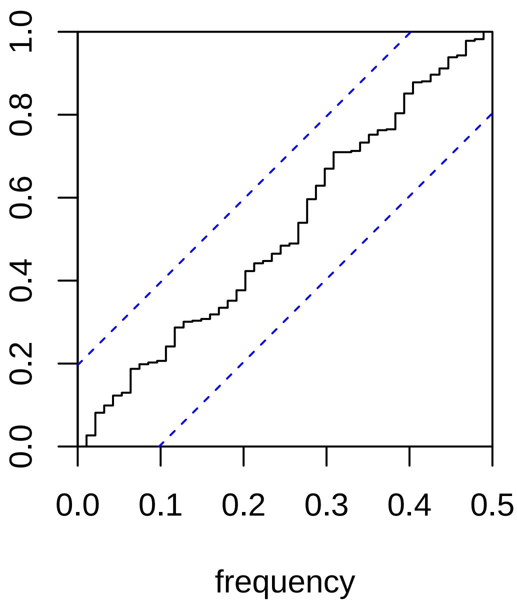


We also apply Fisher's test on the residuals for the presence of hidden periodicities with unspecified frequency.

The result 0.9606491 passes the test, which indicates that no periodicities detected. We then use Komolgorov-Smirnov test for cumulative periodogram of residual.The following graph shows that our residuals passed the test since our test statistics are within the boundaries. These indicate our residual resemble Gaussian white noise resulted from a well fitted model.

```
## [1] 0.9606491
```

```
cpgram(res2, main="")
```

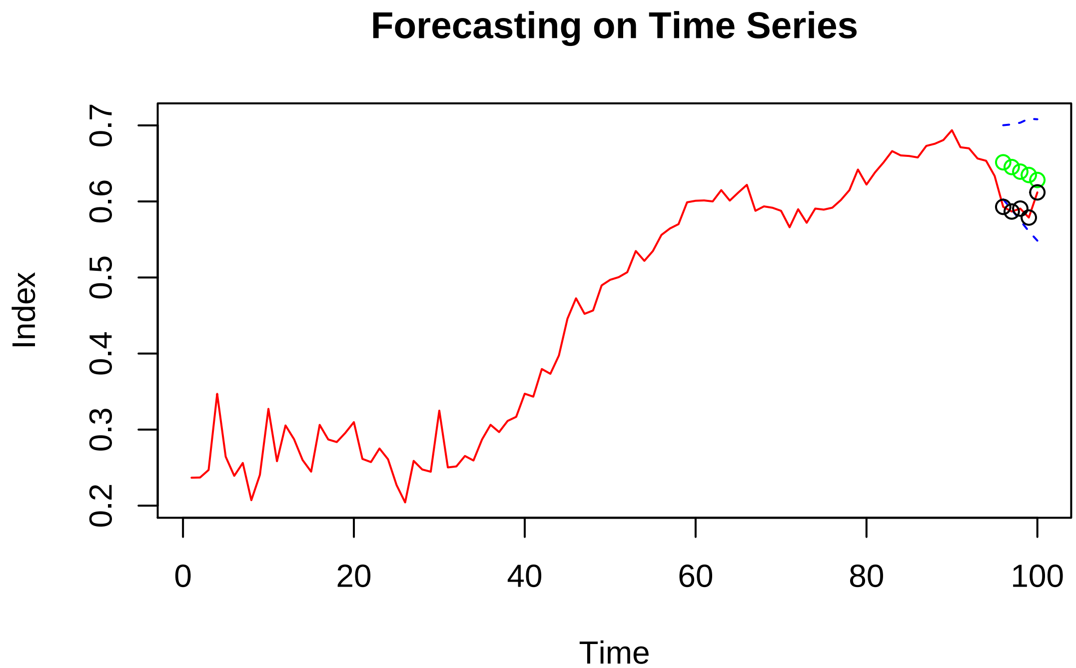


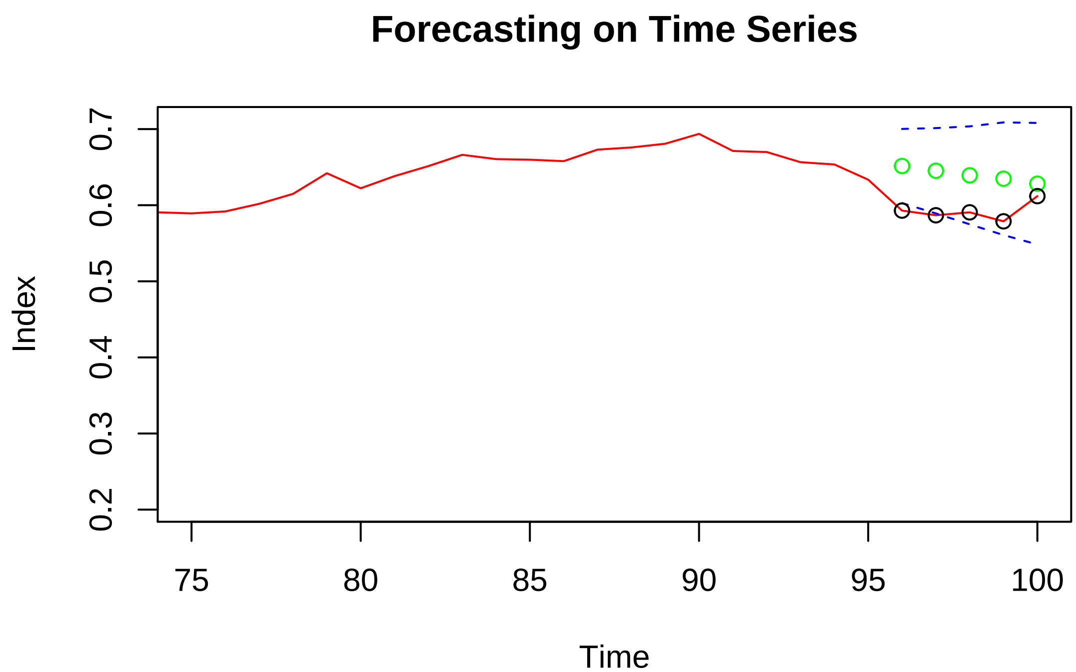
5. Data Forecasting

Based on the model we drived above, we fit our model ARIMA(6,1,2):

$$(1 + 0.5342B - 0.1713B^2 - 0.2227B^4 + 0.3637B^6)(1 - B)X_t = (1 - 0.9717B + 0.4637B^2)Z_t$$

with the training data with 95 observations and make prediction on the next 5 observations. From the forecasting graph, we are able to see that our model correctly predict the negative trend of the engergy index and the prediction is relatively close to the true value. However, our confidence interval fail to catch all the future values.





There are one or two out of the total five data points are outside the interval. The inaccuracy here might be due to the fisrt 20 - 30 observations, which exhibit relatively stationary trend comparing to the following positive linear trend. Those data are conclude from the song track from roughly 1921-1950, where the world is undergoing an unusual time due to the instability of society, such as wars, revolutions... The amount of song track data for those period of time we have is also limited comparing to the data for later years.

To fit a model that could make better prediction on the future value, we might consider exclude those data, however, since we only have 100 observations in total in this dataset, it is not appropriate to do so. Overall, our time series model ARIMA(6,1,2):

$$(1 + 0.5342B - 0.1713B^2 - 0.2227B^4 + 0.3637B^6)(1 - B)X_t = (1 - 0.9717B + 0.4637B^2)Z_t$$

are able to give approximate forecasting the engergy index for future song tracks.

6. Conclusion

In summary, we examine the trend and seasonality and apply transformation and differencing method to the original data. We identify the candidate models by with ACF, PACF, and spectral analysis. We also perform diagnostic checking and determine the final forecasting model. With the derived model. We are able to predict the general trend and approximate value of the energy level of song tracks in the future.