Time Series Project

Adams Zequi Mohammed 20131887

May 5th,2021

Section 1 :Introduction: Time Series Analysis and Our Dataset

1.1 Time Series Analysis

A time series, like the name implies, studies the collection of sequence of events happening in a time frame.Time series analysis as a field is used to analyze , model and gain insights into many different phenomena.A few of them are financial time series analysis (eg.stock price modeling and prediction and cost-benefit analysis of net present value of project) ,analyzing and predicting census and population of demographics,analyzing the weather (in terms of temperature predictions,climate predictions and general weather forecasting) and healthcare applications such as insurance premium and benefits calculations.

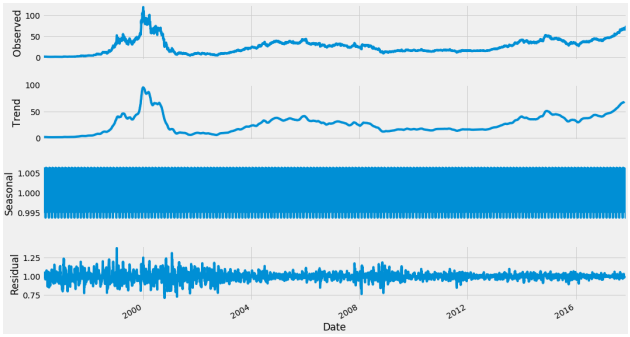


Figure 1: An example of application of Time Series Analysis to finance

1.2 Project Dataset

The data under consideration for this project is the monthly mean value of Bangladesh rainfall.This data spans from 1901 to 1913.The data was gotten from the website link https://www.kaggle.com/yakinrubaiat/bangladesh-weather-dataset. The complete dataset contains data both temperature and rainfall data but for this analysis we use the rainfall dataset compiled in monthly averages.The data analysis will mostly be centered around training and predicting our data into the nearest future.To this end we will split our data into 90%/10% , assess the stationarity of dataset (and apply transformations if necessary),decide on reasonable models for the data , fit them and finally predict and compare with the 10%.

Section 2 : Exploratoory Analysis

2.1 Preliminary Analysis And Stability Tests

To begin our preliminary analysis, we load up our data in R using the *read.csv* function.We notice our data is linear so to simplify it and convert it to a matrix form for analysis.

We then split our database into 90/10 where the 90% is our training dataset and 10% is our test dataset.

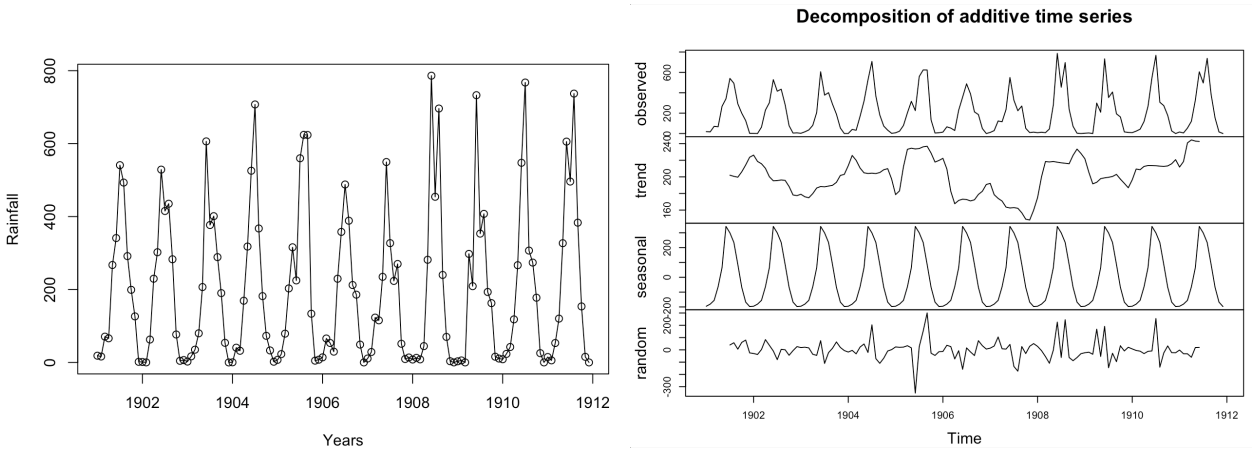


Figure 2: Monthly Average Rainfall Data for Bangladesh and Decomposed form

Figure 2 is the result of plotting our raw data.We try to determine the *short term dynamics* by looking with the naked eye to see if there are any clear patterns in the data and immediately realize that besides the fact that rainfall figures above 600 are outliers , there are hardly any clear patterns so we plot the decompose to check if we can see any clear patterns.We also plotted the data visually to determine if there are any noticeable patterns. We see that there is noticeable pattern of seasonality but no patterns in trend.

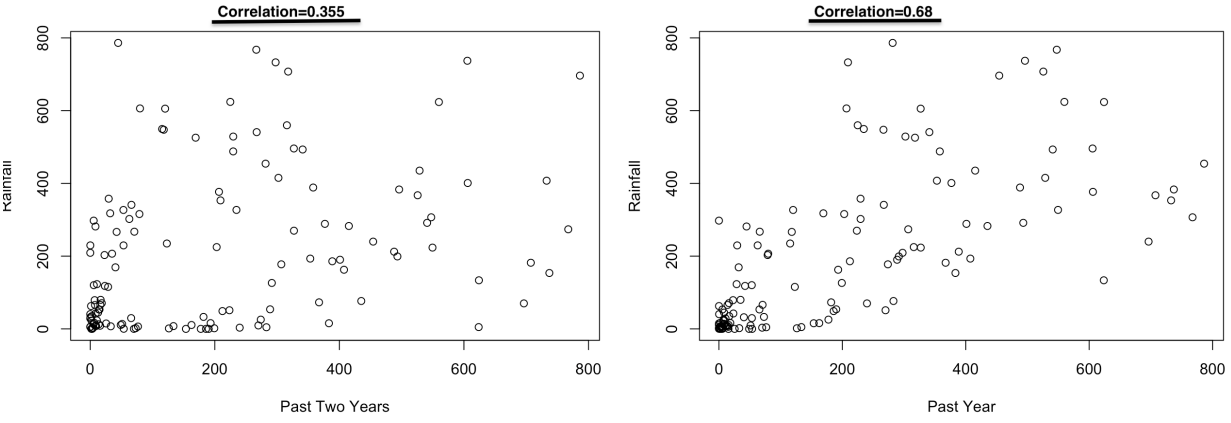
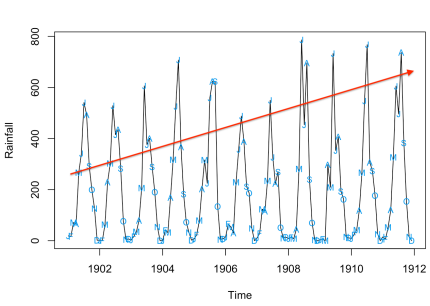
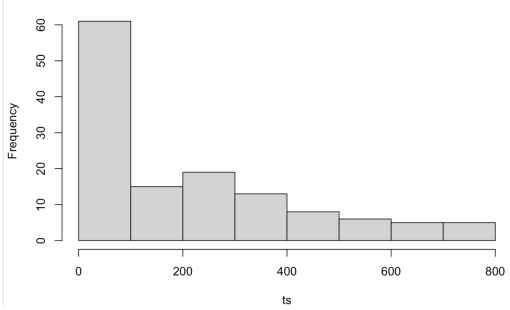


Figure 3: Lag 1 and Lag 2 Correlations of our Raw data

By plotting the lag 1 and 2 correlation of our data, even though the plots are still hazy we can see that the ρ is 0.35 for lag-2 and 0.68 for lag-1(almost twice that of lag-2)

To check for another variable of stationarity , we also plot a seasoned(monthly) filtered plot and boxplot.





season(ts)	rain
January	8.093448
February	21.644157
March	50.912286
April	133.968418
May	268.923636
June	527.819364
July	498.740364
August	461.885636
September	294.660818
October	133.953164
November	30.990403
December	3.836852

From the figure above , we can see that June, July and August recorded the highest rainfall patterns while November and December recorded the lowest.We also have on the far right side, the aggregate figures where we notice that June records the highest mean while December records the lowest.As to the stability,we see that there is a relatively increasing trend of the variance in our data plot indicting a non-stationary distribution and also we see that the histogram plotted doesn't have a gaussian distribution visually which indicates non-stationarity but **we cannot be entirely sure until we do a boxcox transformation to confirm and to concretely stabilize the data**.

2.2 Stability And Transformation

Earlier , we visually tried to determine trends in our data and concluded that while there was an apparent seasonality pattern , there was no apparent trend pattern.To further test the stability of trend of our data , we plotted a steady increasing line in our data visually indicating that our data is non-stable, the shape of the histogram we plotted also didn't look like a gaussian which also strongly indicates a non-stability distribution but we can't be entirely sure so we concluded on using the boxcox transformation.

In this section we use the boxcox function to find an appropriate lambda which tells us what method to use in transformation of our data.The boxcox transformation generally stabilizes the variability of the data and makes it more stable(normally distributed)

The *BoxCox.ar* function in R transforms over a range of lambda values that we provide it and the function finds a log-likelihood based using an Autoregressive (AR) covariance structure. It ultimately finds the most appropriate lambda that transforms the data to a stable series with normal white noise term

By running the boxcox.ar test , we find the most optimal lambda to be 0.4 with a narrow confidence interval of 0.3 *to* 0.4.A lambda of 0.3 indicates that we have to transform th data using a *square root transformation ie.* Y^{lambda} *where* Y *is our data*

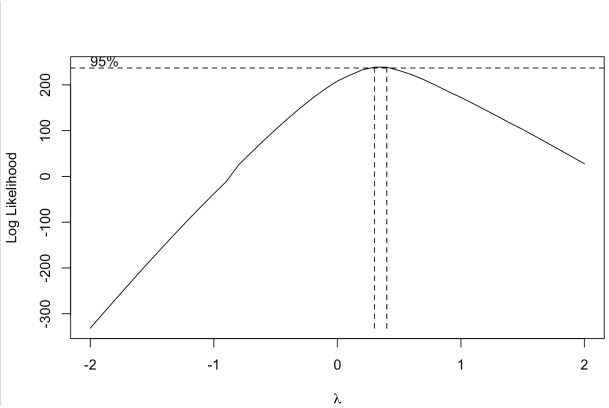
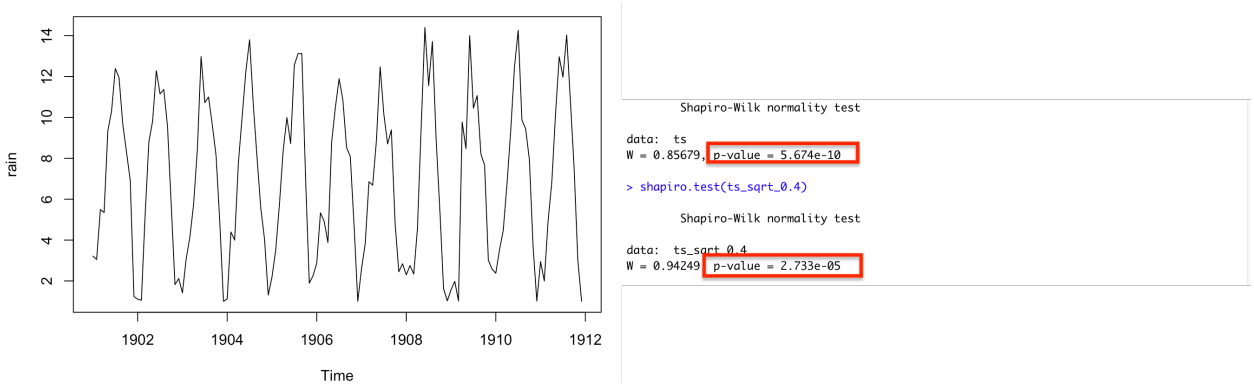


Figure 4: BoxCox.Ar Plot of our data

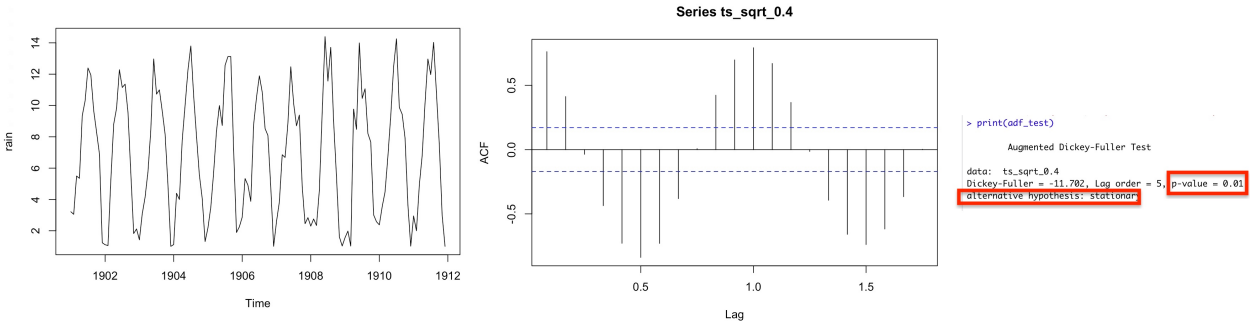


We then transform the data based on our lambda gotten.From Figure above , we see immediatiely that the variance in our data has been reduced. We go further to do a normality test using the $sharpiro - wilk\ normality$ test and find out that the normality of our data set has significantly increased since our p-value increased from $5.674e-10$ to $2.733e-05$.The normality test benchmark is 0.05.

Section 3 : **Stationarity** Tests for Trend and Seasonality

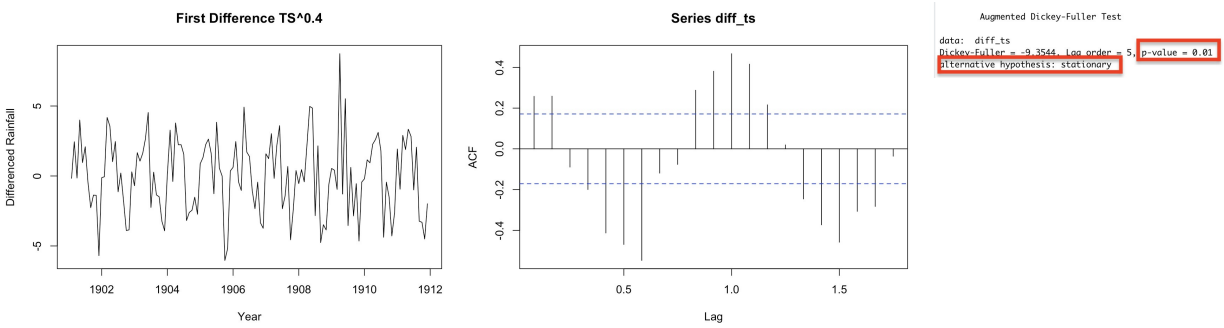
3.1 Stationarity Tests And Trend Differencing

We test for stationarity by plotting our transformed data and trying to visually detect a trend.We can also use an acf to detect a trend and if we detect a trend we would need to difference it further(usually once or twice).



From the Figure above we can see that there is some sort of trend in our data , which calls for differencing.Since we're undecided about if the data carries a trend or not , we use the Augmented Dickey-Fuller test which also gives us a p-value of 0.01 which means that we can reject the null hypothesis for the alternative hypothesis which means stationary.Th adf.test function is subject to errors so we might need to do further analysis to confirm.

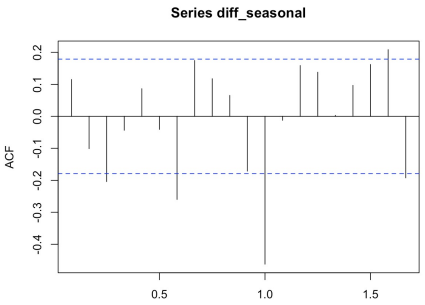
To be sure, we will difference the data to check if it shows lags beyond 95% C.I and if its visually stationary.We will difference it once(for the start).



We can see from the Figure above that our series has been further stationarised while our p-value for the adf is still 0.01 which translates to the fact that we can reject the null hypothesis for the alternative hypothesis that our series is stationary.

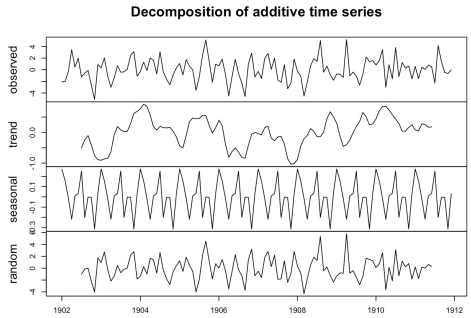
3.2 Seasonality

From Figure 5, we can look for seasonality patterns, we see that the acf is clearly wavy which indicates a monthly seasonality pattern.To startionarise the seasonality ,we do a seasonal difference with lag 12 , indicating 12 month cyclicality.




```
        Augmented Dickey-Fuller Test

data:  diff_seasonal
Dickey-Fuller = -4.992, Lag order = 4, p-value =
0.01
alternative hypothesis: stationary
```



If we look closely at the first(left) plot in the figure above we see that we have stationarised the seasonal pattern in the data as we see a significant spike at lag 1(which is scaled to 12,monthly) but its hard to make conclusions on the decomposed plot(far right) since the time frame of seasonality is small(-0.3 to 0.3).

Carefully analyzing the acf for seasonalised difference,we can see significant autocorrelations that do not look wavy but have significant autocorrelations at lag points that are roughly equal to the period which we term seasonal autocorrelations(Lag 12 which is scaled to 1) so we can confidently say our data is now stationary.Furthermore we used the adf.test where we still got a p-value of p-value = 0.01 so we can reject the null hypothesis and accept the alternative hypothesis which says the series is stationary.

Section 4 : Model Fitting

4.1 Developing our Model

In the previous chapter , we concluded that our data was seasonal and when we differenced for seasonality , we saw that our plot showed significant lags at 1(scaled for 12 months) so we can conclude that because our data is indead seasonal ,we can combine both steps in 3.1 and 3.2 to , also know as seasonal and trend differencing to carry out our model which will be **ARIMA(p,d,q)x(P,D,Q)$_{12}$** model.This model means that the seasonal Arima model has been scaled to a lag 12.

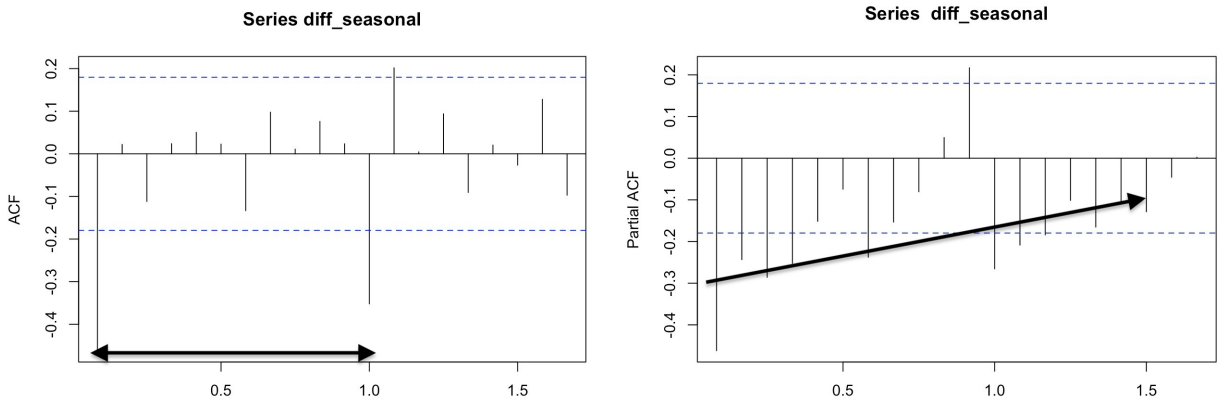


Figure 5: ACF and PACF of seasonally and Trend difference Plots

We deifferenced our data for trend using lag-1 and differenced for seasonality using lag-12.We can see clearly from the ACF plot that there is a significant lag at 1(scaled for 12) indicating that our Q=12 and a significant lag at 1 for q.We can also see that the PACF plot decays with increasing lags while these no significant autocorrelation at lag-1.We can summarise this by writing ARIMA(0,1,1)x(0,1,1)$_{12}$.

GENERAL VARIANTS OF OUR PROPOSED MODEL

MODELS	ARIMA
Proposed Model 1	ARIMA(0,1,1)(0,1,1)
Model 2	ARIMA(1,1,1)(1,1,1)
Model 3	ARIMA(0,1,2)(0,1,2)
Model 4	ARIMA(2,1,2)(2,1,2)

```
   Model.1  Model.2  Model.3  Model.4
1 451.7031 454.1243 454.1243 454.1243
>
```

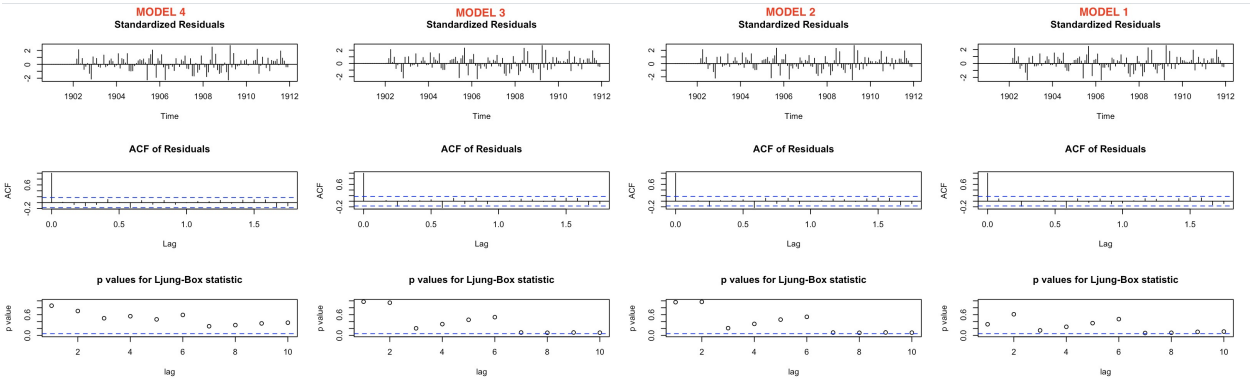
Figure 6: Genral Variant of our model and their corresponding Akaike information Criterion (AIC)

We developed our variant models by modifying our AR and MA components of our ARIMA model by ± 1.We then compared this to their AIC's .In the figure above we see that , the AIC are approximately 451.703,454.1243,454.1243,454.1243 which means that the difference between the smallest and any of the either 3 is 2.423.This means that the most optimal model to pick would be Model-1 since the difference is technically 0.423 more than 2.

Since the approximate difference is 2, we can do more tests to find the most appropriate model since the four models are more or less the same.These tests will be addressed in the next section.

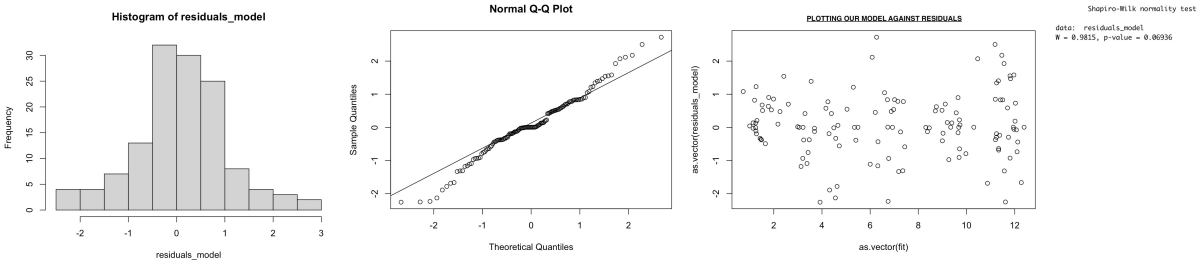
4.2 Choosing our model

From section 4.1 , we noticed that we had 4 models to scrutinize and choose from.We saw from the AIC that model one is barely the most optimal model but as the difference is closer to 2(0.423) than it is to 3, we decided to do more test to develop a concrete model. These tests will test for normality of all four models using different parameters so we pick the most optimal one.



We use the *tsdiag* function to diagnose the normality our four models. From the figure above we can see that our four models are similar in terms of standardized residuals and the ACF.In relation to the standardized residuals , we see that they do not barely exceed -2 to 2 across all 4 models which indicates normality for the the residuals.The ACFs of all 4 models are also only significant at 0 which is virtually insignificant.

The diversion is with the p-values- while model 1,2 and 3 have p values that straddle the 0.05 line, all the p-values of model 4 are greater than 0.05 indicating pure normality so we can pick model-4 as our model to predict the future of our dataset.



To further confirm our models , we plot a number of graphs to check the normality of our chosen model (model-4).We can see that the histogram is visually normal indicating a gausssian distribution.Looking at the QQ plot, we can also see that most of the points fall on our plotted line. When we plotted our residuals against our fitted model , we see an unpatterned scattered formation of dots indicatinng randomness and finally the shapiro-wilk test gives us a p-value of 0.06936 which is greater than 0.05 indicating normality meaning accepting the null hypothesis.

Section 5 : Prediction

5.1 Predicting future values

In the previous chapter we chose our model and further tested it using various parameters.In this chapter , we predict the future using our model

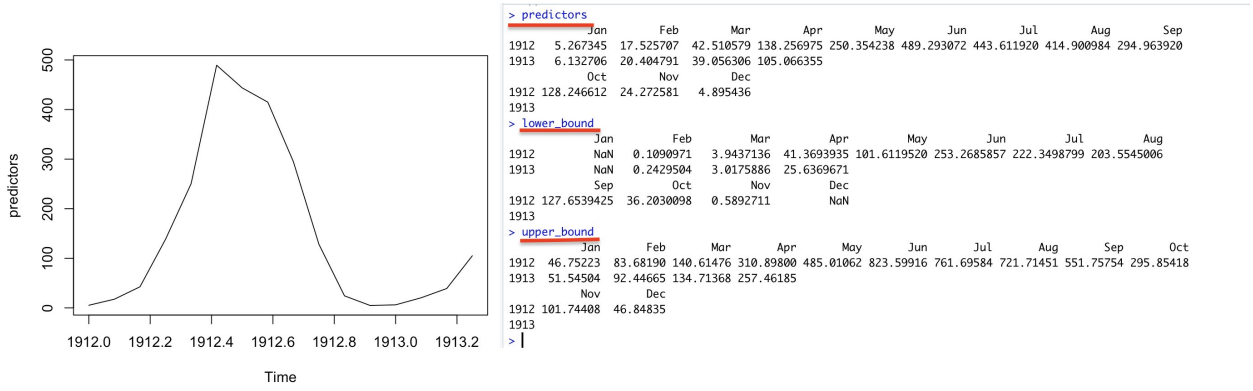


Figure 7: Scaled Predicted Values and Scaled Standeard Errors

We begin by looking at the predicted values and standard errors that we scale to reverse the effect of transforming our data earlier.We then actually place it on our full data to see if it matches with our data.

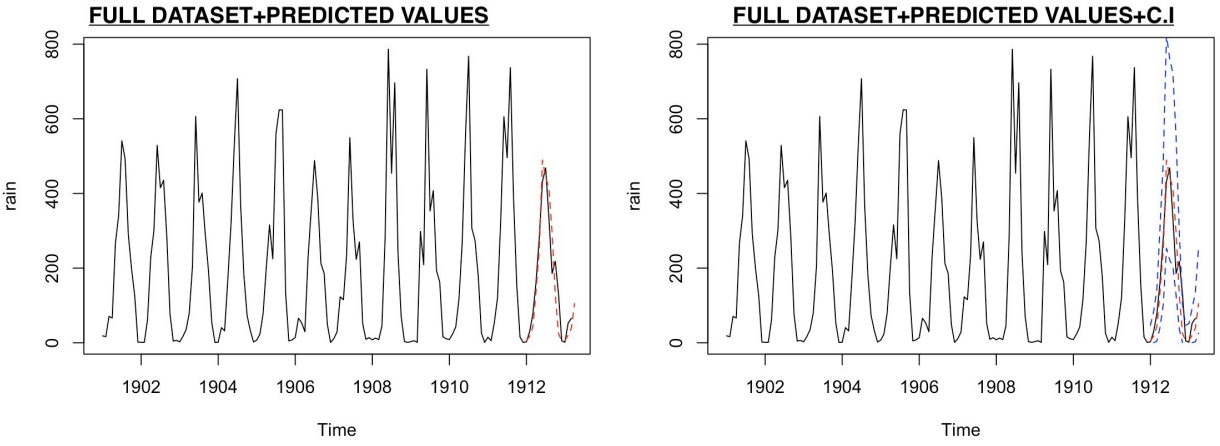


Figure 8: Model Confirmation Plots

To predict our dataset as we see in the above figure ,we use the *predict function* in r using the *n.ahead* to specify number of times ahead.After that, we transform the predicted values we get to scale with our actual dataset and since we used a lamba of 0.4 to stationarise our datset , we scale our predicted value by raising to 2.5.

We use the same method to scale of lower bound and higher bound confidence intervals to plot along with our actual dataset.We plot all these variables after.

Looking at the figure , we can see that our predicted variables fit almost perfectly on our actual dataset meaning we picked the right model.On the second figure , we can see our plotted predicted

values with their confidence intervals which generally stay in the bounndarys of our graph and dont deviate.

Generallly from our graph we can say that this is a successful modelling process.

Section 6 : Conclusions

We began by loading our data, plotting our raw data to see if we can visually make some inferences.We try to determine the short term dynamics by looking with the naked eye to see if there are any clear patterns in the data .After we look at a correlogram to see if we can make visual inferences about the correlation.We then try to check for normality and stabilize our data after trend and seasonal-wise.

Stability is found through getting a lambda using the boxcox and transforming the data.After the data is transformed we check for trend and seasonality and determine our model from there.We do various tests for model determination till we arrive at the best model.

We then fit our chosen model ,use the predict function in r to predict the future values and and combine our predicted values to our full dataset.

We ultimately came to the conclusion that our model fit the data as our predicted values match almost perfectly our test dataset.