Time Series Analysis On Monthly Rainfall In Andaman-Nicobar Sub-Division Of India, From Year 1901-2015

By, Arindam Adak
M.Sc. Big Data Analytics

Data Description

The data we have in hand consists of monthly & annual rainfall in different Sub-Divisions of India, among which I have chosen to work on Andaman-Nicobar Islands. The dataset has 4117 rows and 15 columns. Here is a glimpse of the dataset at hand

SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANNUAL
ANDAMAN & NICOBAR I	1901	49.2	87.1	29.2	2.3	528.8	517.5	365.1	481.1	332.6	388.5	558.2	33.6	3373.2
ANDAMAN & NICOBAR I	1902	0	159.8	12.2	0	446.1	537.1	228.9	753.7	666.2	197.2	359	160.5	3520.7
ANDAMAN & NICOBAR I	1903	12.7	144	0	1	235.1	479.9	728.4	326.7	339	181.2	284.4	225	2957.4
ANDAMAN & NICOBAR I	1904	9.4	14.7	0	202.4	304.5	495.1	502	160.1	820.4	222.2	308.7	40.1	3079.6
ANDAMAN & NICOBAR I	1905	1.3	0	3.3	26.9	279.5	628.7	368.7	330.5	297	260.7	25.4	344.7	2566.7
ANDAMAN & NICOBAR I	1906	36.6	0	0	0	556.1	733.3	247.7	320.5	164.3	267.8	128.9	79.2	2534.4
ANDAMAN & NICOBAR I	1907	110.7	0	113.3	21.6	616.3	305.2	443.9	377.6	200.4	264.4	648.9	245.6	3347.9

- The no. of subdivisions are 36 for example : Andaman Nicobar , Arunachal Pradesh, Haryana Delhi & Chandigarh, West Bengal etc. The year span for our chosen sub-division Andaman-Nicobar is 1901-2015.
- The are NA values in the columns, that is amount of rainfall in some month of a year is missing. And also there are missing years in between. Data from 1943 to 1945 is not given.
- We use the data for 2015 as test data and rest as train.

Data Cleaning & Preprocessing.

The question is, how do we handle missing values in time series? In principle, we cannot just omit them without breaking the time structure. And breaking it means going away from our paradigm of **equally spaced points in time**. A popular choice is thus replacing the missing value. This can be done with various degrees of sophistication:

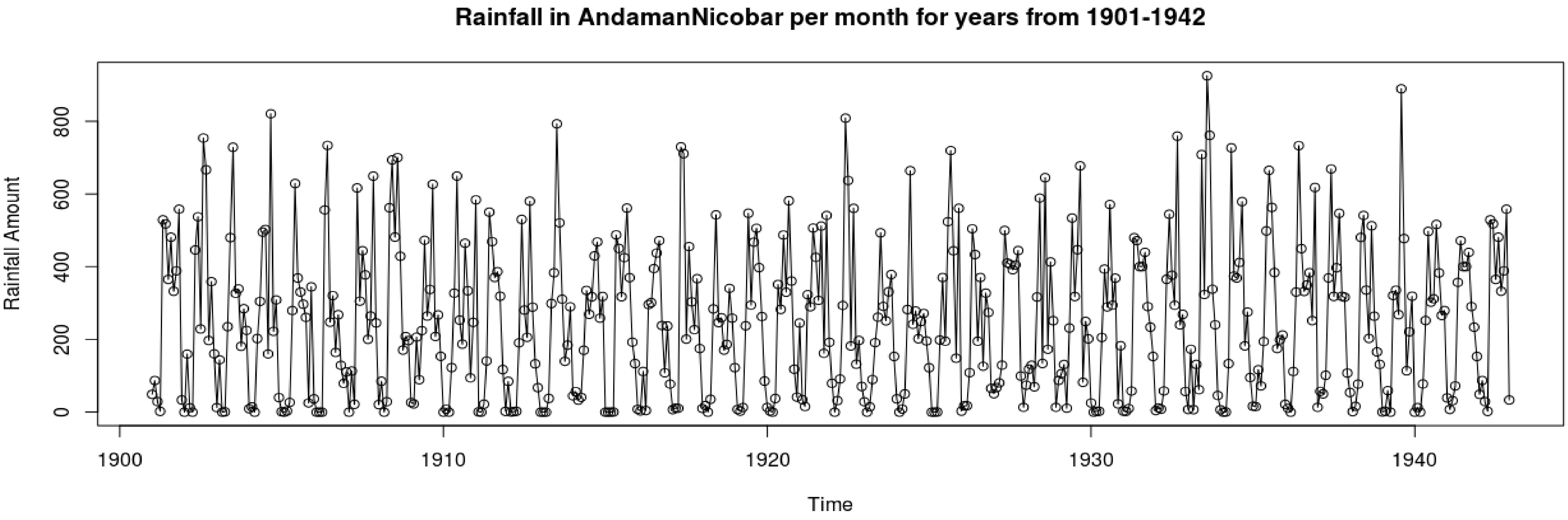
- Replace the NA’s by the mean of that particular column.
- Replace the NA ‘s by the mean of the previous and next 3 data of that column, etc..

Here we choose the first process.

And for the missing years problem I divided the dataset into three parts: one from 1991-2015 so that for our end forecast we have enough data. And then for the rest of them I divided the data into two parts on till the gap ,and the rest.
For example in case of SubDivision : Andaman & Nicobar is divided into
1901-1942, 1946-2000 & 2001-2014.

We will use the data of 2015 for validation.

Subdivision: Andaman & Nicobar: Cluster 1: 1901-42



From the plots in Figure 1 it could be seen that the time series plot displays a wave like pattern an evidence of seasonality and no trend is observed.

Subdivision: Andaman & Nicobar : Cluster 1: 1901-42

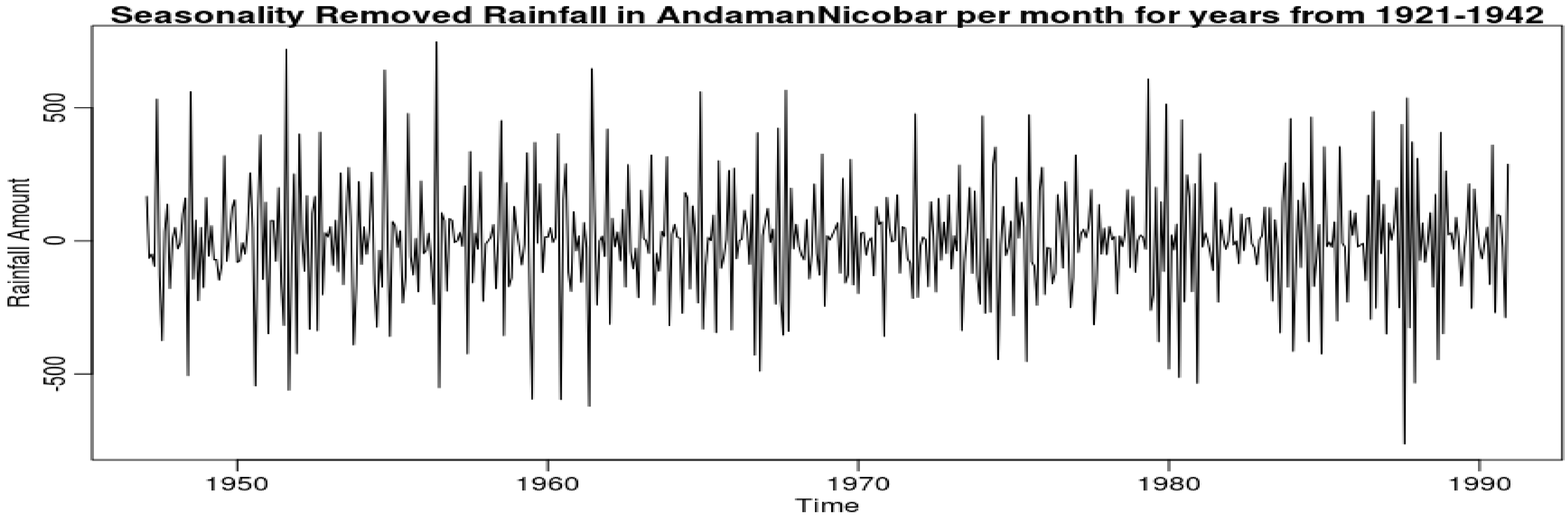
Dickey Fuller Test: the Augmented Dickey-Fuller test with hypothesis H0(Null Hypothesis): The rainfall data has unit root non stationary and H1(Alternative Hypothesis): The rainfall data is stationary.

Test Statistic	p-values
-17.246	0.01

Decision: Small p-value 0.01 less than 0.05 is in favor of the alternative hypothesis. Thus, strong evidence against the null hypothesis at 5% level of significance.

In order to eliminate the seasonal effect from the time series we will subject the data to a seasonal differencing. Seasonality is yearly , that is 12. After differencing the plot is seen as in the next slide.

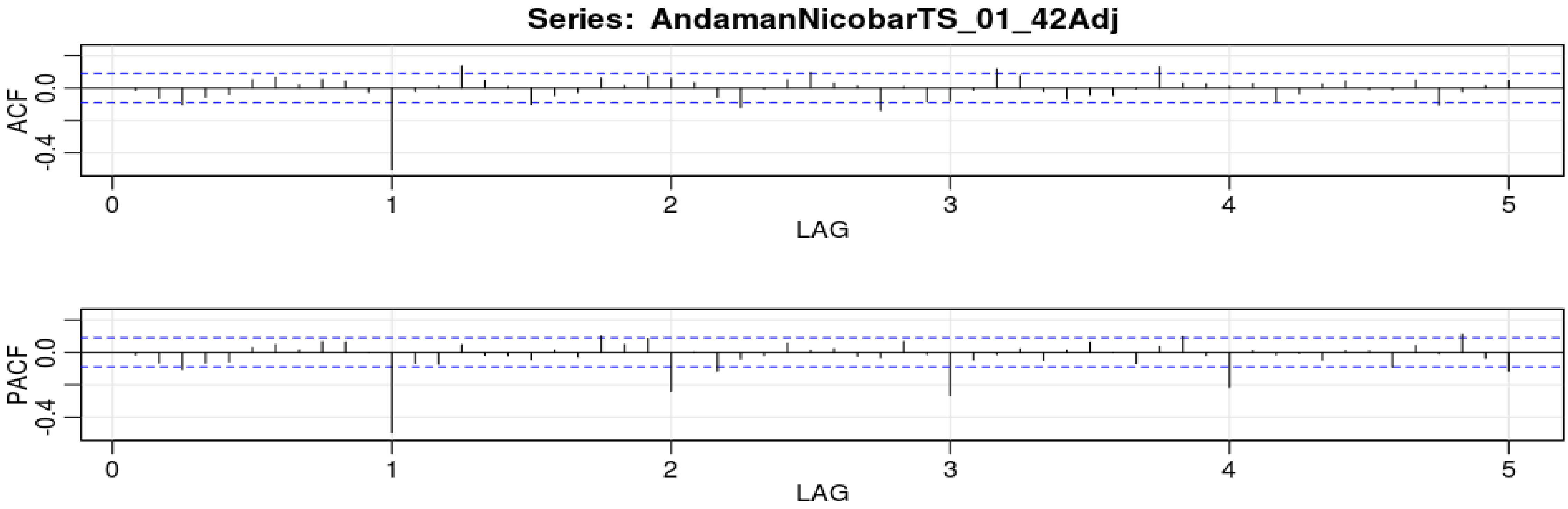
Subdivision: Andaman & Nicobar : Cluster 1: 1901-42



We see that the mean and variance are constant over time after the first order differencing.

Subdivision: Andaman & Nicobar : Cluster 1: 1901-42

A time series is said to be seasonal if there is a sinusoidal or periodic pattern in the series and when this happens the SARIMA model inevitably becomes the choice model. Then the acf plot is shows the following :



Suggests : p = 0, d = 0, q = 0 P = 5, D = 1, Q = 1

Subdivision: Andaman & Nicobar : Cluster 1: 1901-42

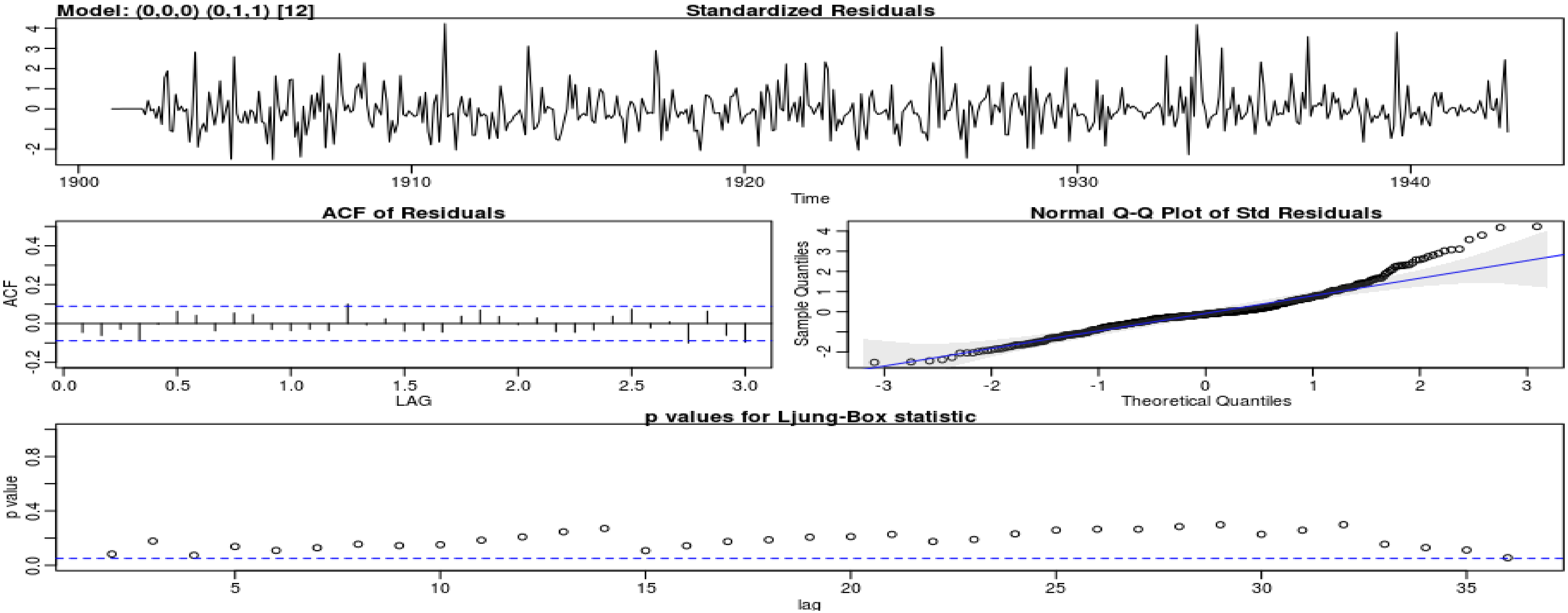
Even though the ACF & PACF plots suggest model with parameters :
p = 0, d = 0, q = 0 P = 5, D = 1, Q = 1 , we see there are too many parameters are to be estimated, thus making the model complex. Hence, we try to make these numbers lesser and make our model much simpler. For that to be done we try with small values of p, q, P, Q and compare the model's AICC values to pick the best one

p	d	q	P	D	Q	AICC
0	0	0	5	1	1	10.67891
0	0	0	1	1	1	10.67931
0	0	0	0	1	0	11.53421
0	0	0	0	1	1	10.67423

Suggests : p = 0, d = 0, q = 0 P = 0, D = 1, Q = 1

Subdivision: Andaman & Nicobar : Cluster 1: 1901-42

Applying SARIMA(0,0,0)(0,1,1) AICC: 10.67423



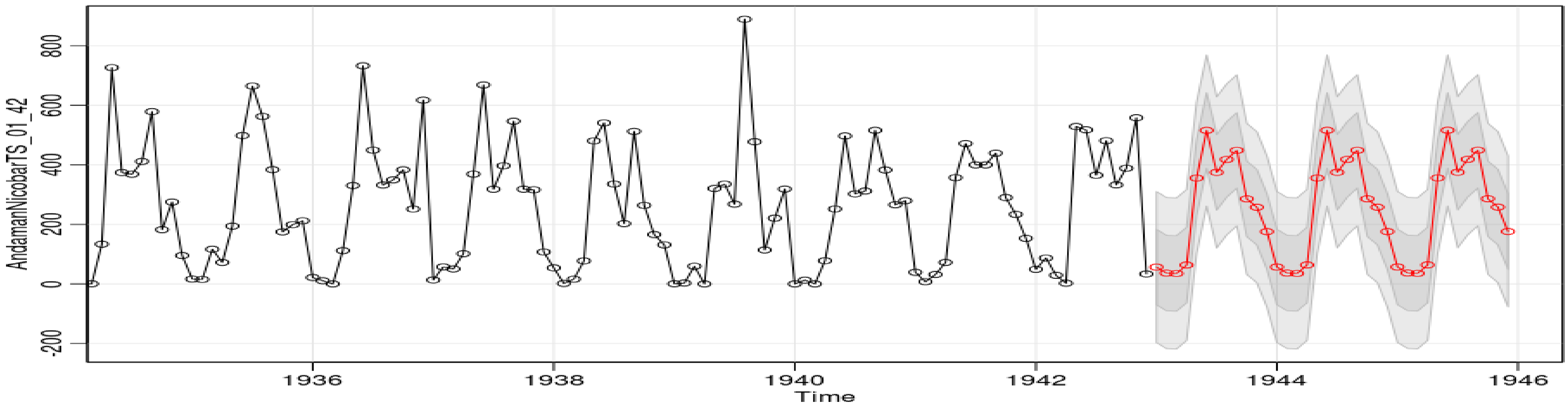
The Ljung-Box test is a dependency test between two variables in this case of the standardized residual .

Null Hypothesis: Independent Alternative Hypothesis: Not Independent.

p-values>0.05(mostly), we don't have enough statistical evidence to reject the null hypothesis. So we can not assume that your values are dependent. Thus we proceed to modelling and forecasting.

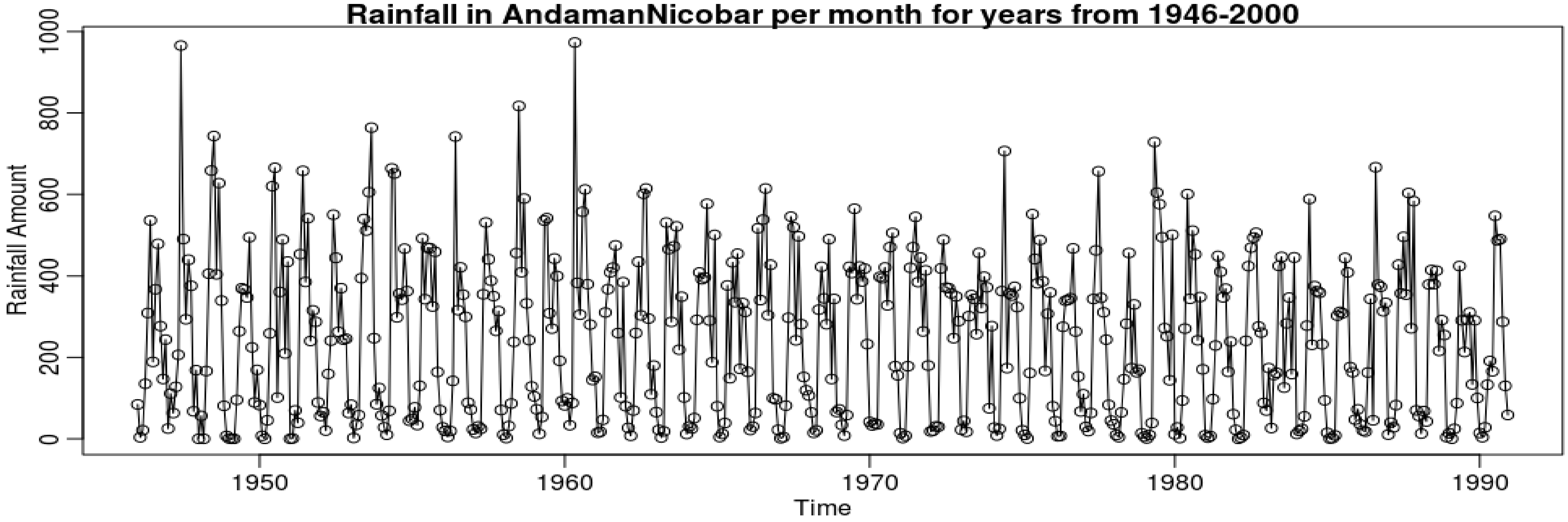
Subdivision: Andaman & Nicobar : Cluster 1: 1901-42

Forecasting the missing Three Year 1943 to 1945:



Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1943	56.51	36.58	35.22	63.46	355.64	515.86	374.69	418.34	448.80	285.78	257.68	175.76
1944	56.72	36.80	35.44	63.68	355.86	516.07	374.91	418.55	449.01	285.99	257.90	175.97
1945	56.94	37.01	35.65	63.89	356.07	516.28	375.12	418.76	449.23	286.20	258.11	176.19

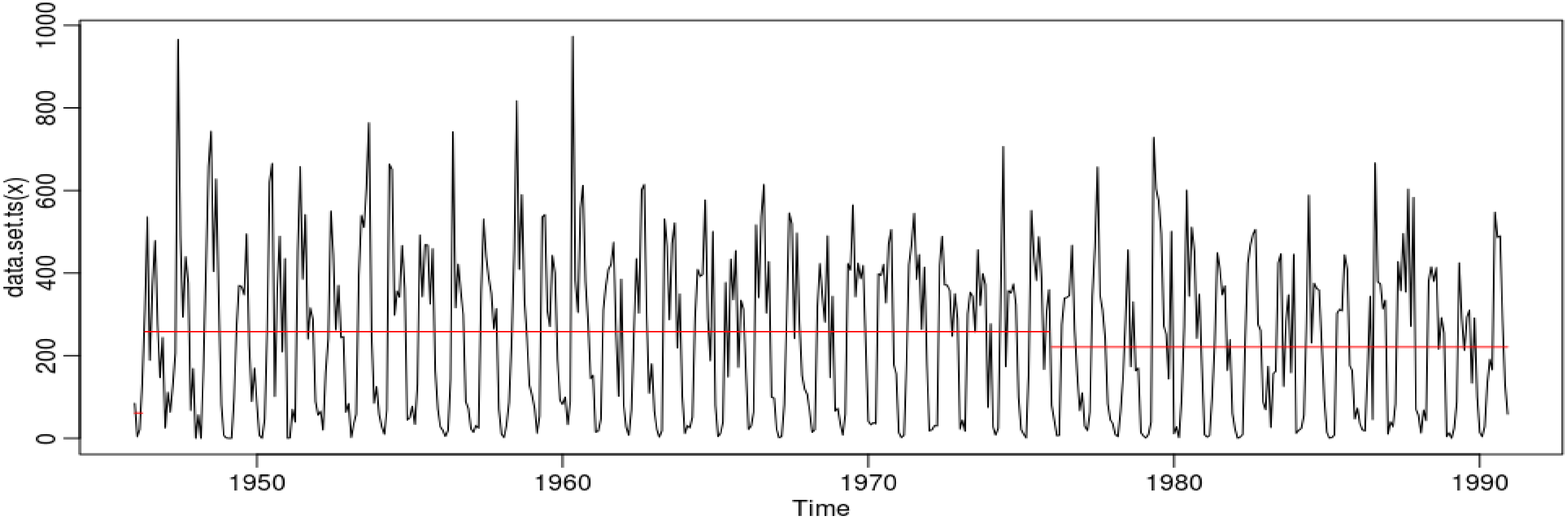
Subdivision: Andaman & Nicobar : Cluster 2: 1946-00



Here again we see a wave like pattern confirming seasonality, of 12 months. But a interesting thing we notice that there is a certain difference in the overall rainfall amount. To check that I try to find out if there are any change points in between (using cpt.mean function of R)

Subdivision: Andaman & Nicobar : Cluster 2: 1946-00

Change Points



Shows that there is a significant difference in mean after year 1976. So we divide the remaining years into two parts 1946-76 and 1977-14.

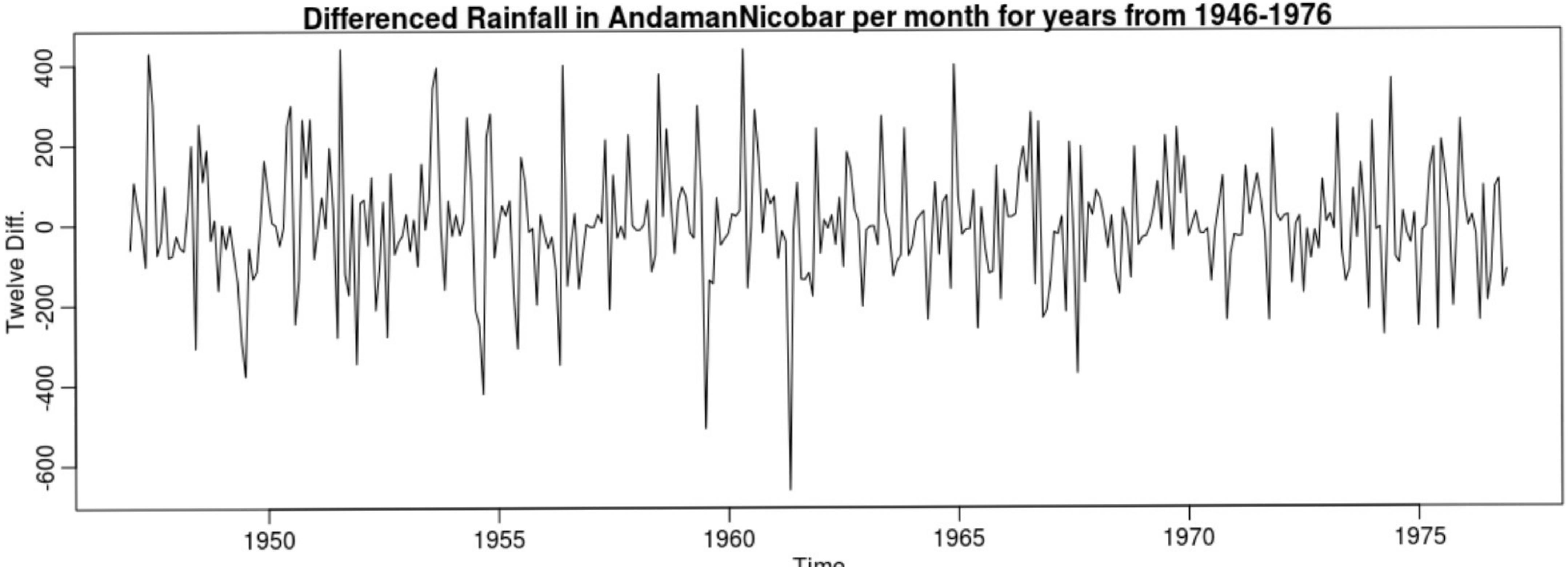
Subdivision: Andaman & Nicobar : Cluster 2: 1946-76

Dickey Fuller Test: the Augmented Dickey-Fuller test [11] with hypothesis H0: The rainfall data has unit root non stationary and H1: The rainfall data is stationary.

Test Statistic	p-values
-14.051	0.01

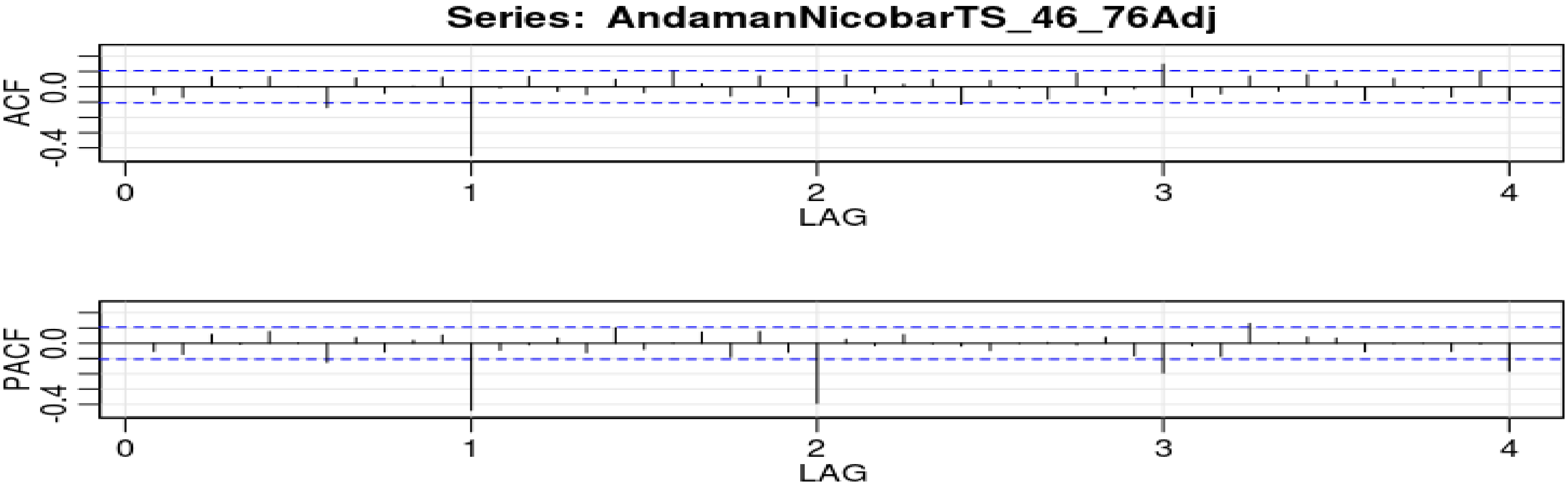
Decision: Small p-value 0.01 less than 0.05 is in favor of the alternative hypothesis. Thus, strong evidence against the null hypothesis at 5% level of significance.

In order to eliminate the seasonal effect from the time series we will subject the data to a seasonal differencing. Seasonality is yearly , that is 12. After differencing the plot is seen in this figure that mean and variance are constant over time..



Subdivision: Andaman & Nicobar : Cluster 2: 1946-76

ACF and PACF plots showing : p = 0, q = 0 , d =0 , P = 4, D =1, Q =2



Subdivision: Andaman & Nicobar : Cluster 2: 1946-76

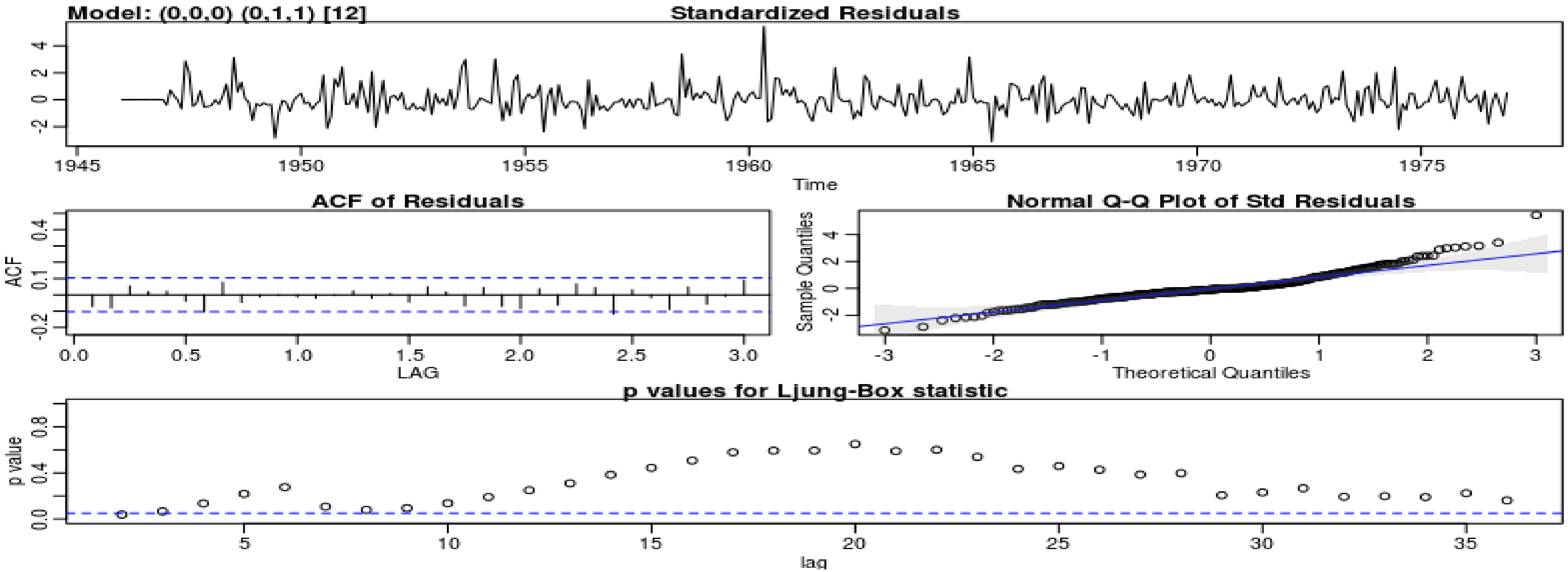
Even though the ACF & PACF plots suggest model with parameters :
p = 0, d = 0, q = 0 P = 4, D = 1, Q = 2 , we see there are too many parameters are to be estimated, thus making the model complex. Hence, we try to make these numbers lesser and make our model much simpler. For that to be done we try with small values of p, q, P, Q and compare the model's AICC values to pick the best one

p	d	q	P	D	Q	AICC
0	0	0	4	1	2	10.45321
0	0	0	1	1	1	10.44316
0	0	0	0	1	0	11.43521
0	0	0	0	1	1	10.43423

Suggests : p = 0, d = 0, q = 0 P = 0, D = 1, Q = 1

Subdivision: Andaman & Nicobar : Cluster 2: 1946-76

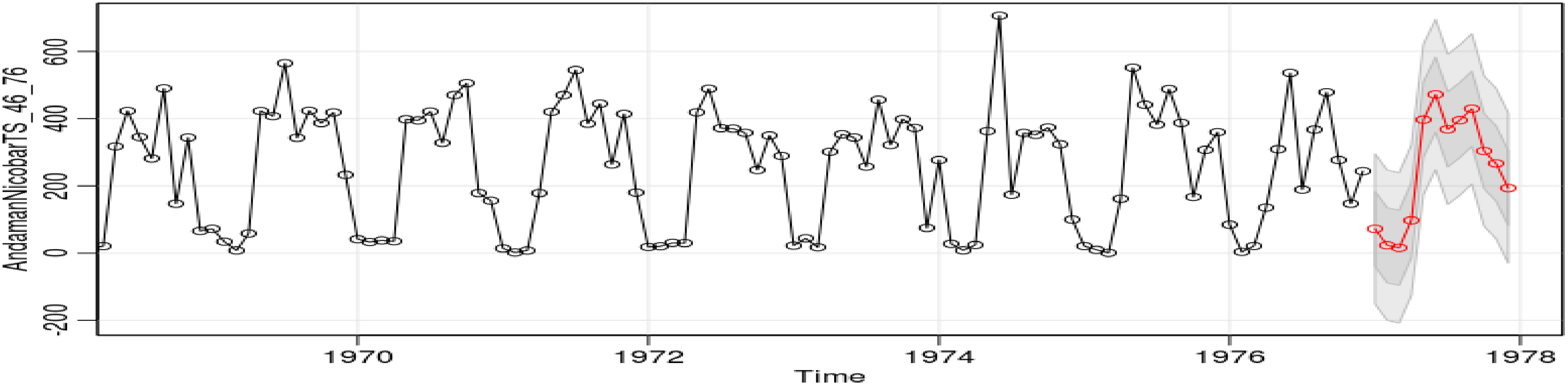
Applying SARIMA(0,0,0)(0,1,1) AICC: 10.43423



The Ljung-Box test is a dependency test between two variables in this case of the standardized residual .
Null Hypothesis: Independent Alternative Hypothesis: Not Independent.
p-values>0.05(mostly), we don't have enough statistical evidence to reject the null hypothesis. So we can not assume that your values are dependent. Thus we proceed to modelling and forecasting.

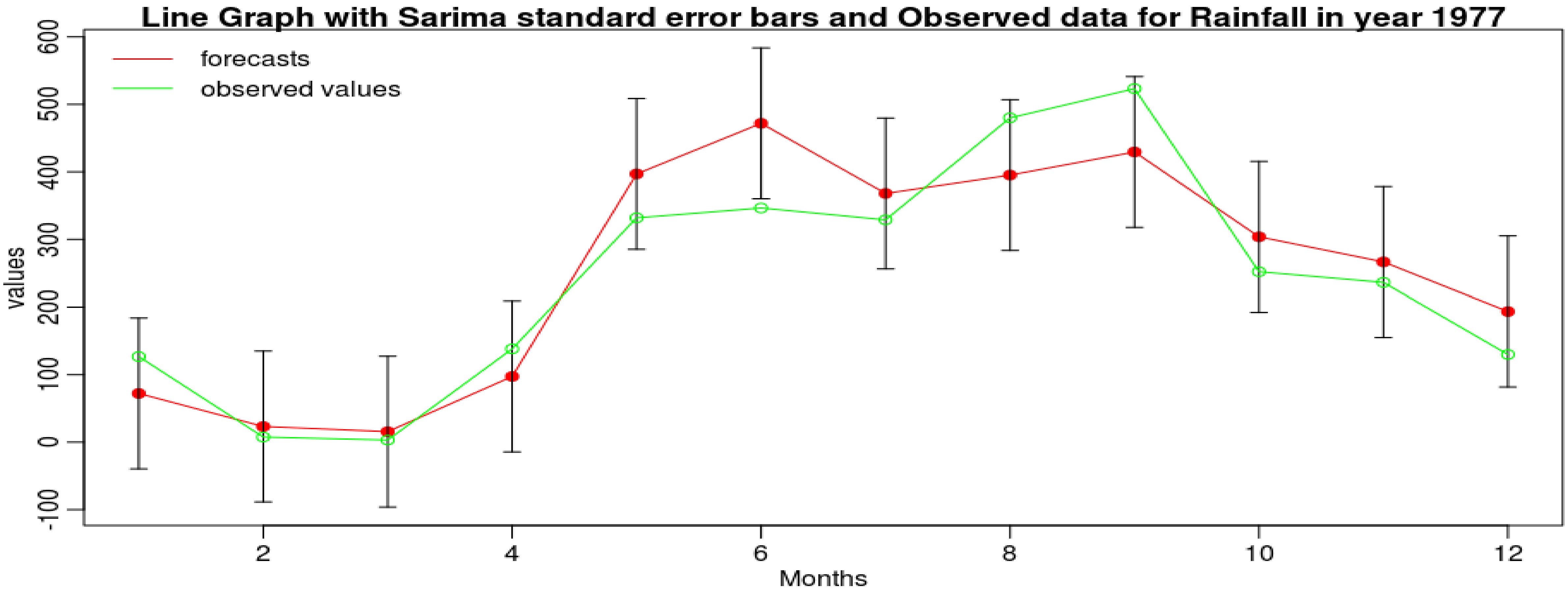
Subdivision: Andaman & Nicobar : Cluster 2: 1946-90

Choosing the above model forecasting year 1977 and RMSE calculating compared to the exixting data



RMSE = 79.63782

Subdivision: Andaman & Nicobar : Cluster 2: 1946-76



Thus we can say that our forecasting was not that bad.

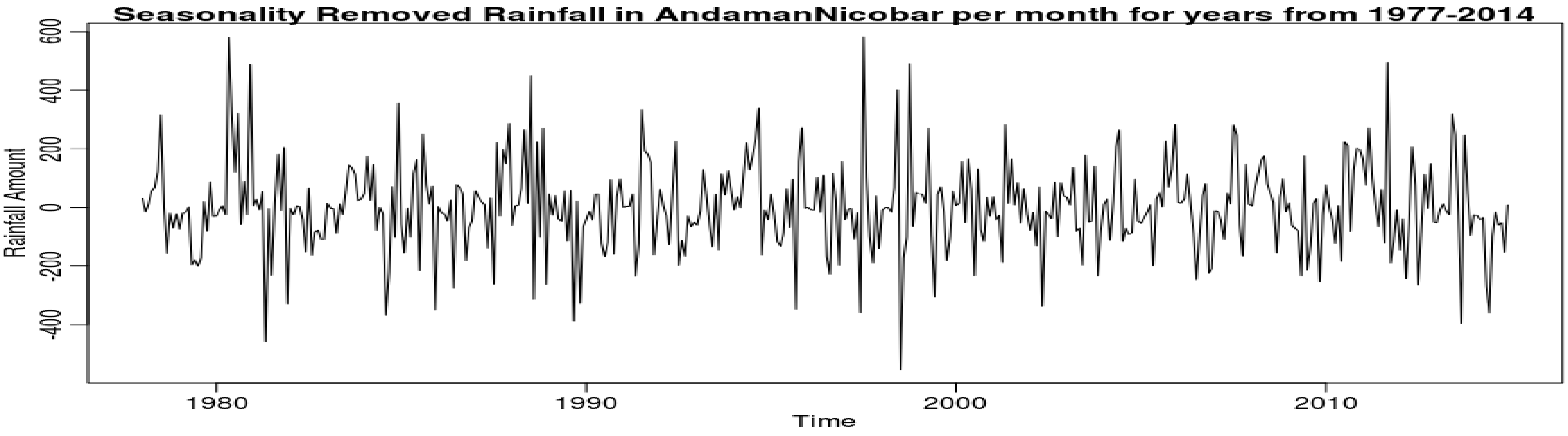
**Subdivision: Andaman & Nicobar : ** Cluster 2: 1977-14

Dickey Fuller Test: the Augmented Dickey-Fuller test [11] with hypothesis H0: The rainfall data has unit root non stationary and H1: The rainfall data is stationary.

Test Statistic	p-values
-13.596	0.01

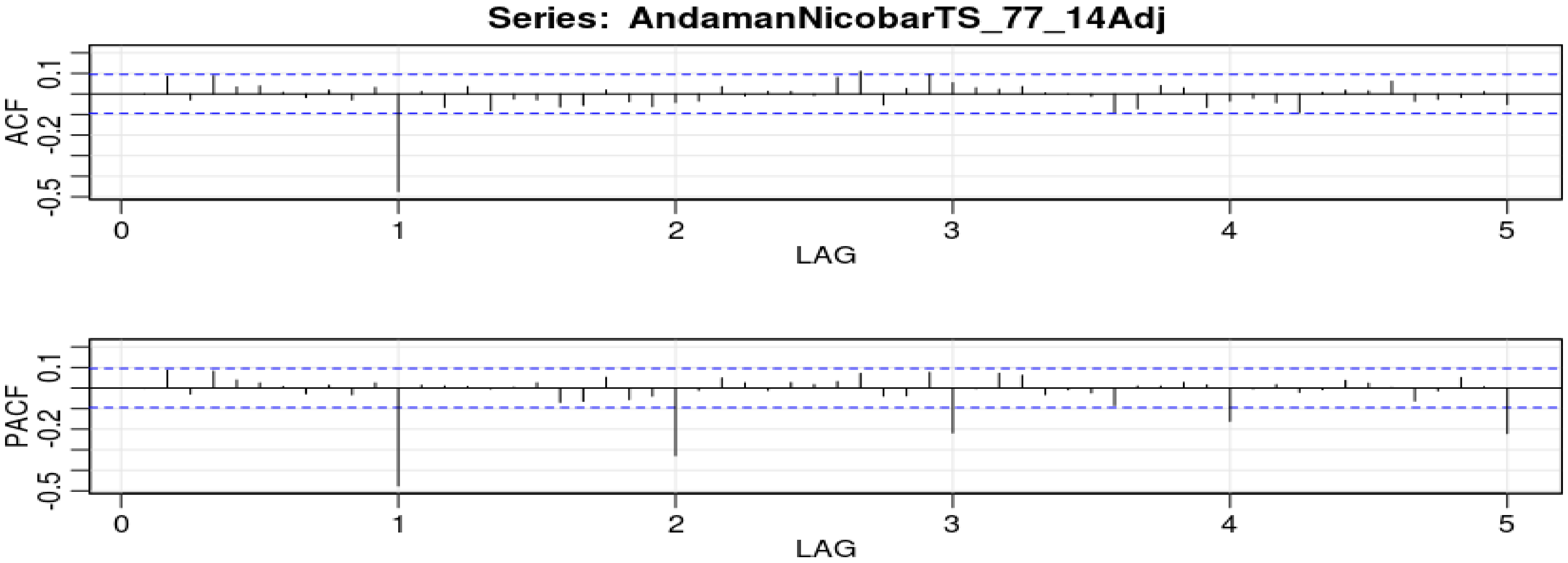
Decision: Small p-value 0.01 less than 0.05 is in favor of the alternative hypothesis. Thus, strong evidence against the null hypothesis at 5% level of significance.

In order to eliminate the seasonal effect from the time series we will subject the data to a seasonal differencing. Seasonality is yearly , that is 12. After differencing the plot is seen in this figure that mean and variance are constant over time..



Subdivision: Andaman & Nicobar : Cluster 2: 1977-2014

ACF and PACF plots showing : p = 0, q = 0 , d =0 , P = 5, D =1, Q =1



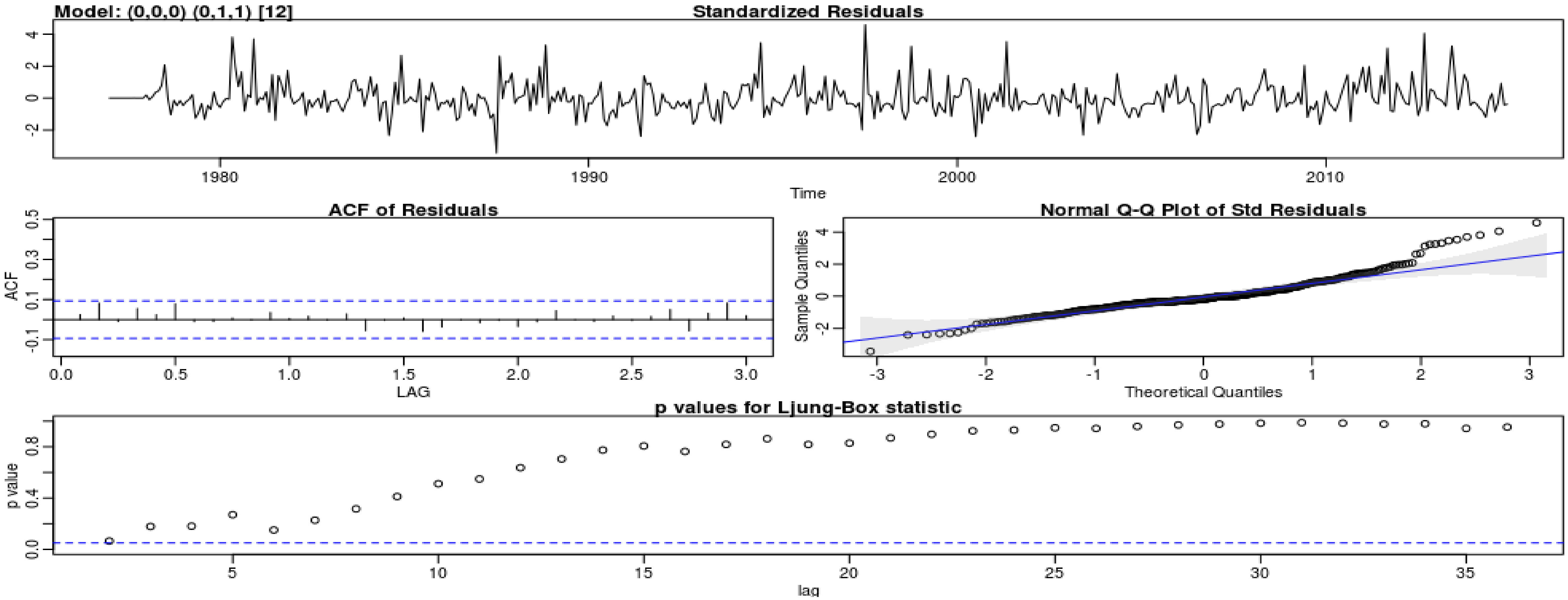
Subdivision: Andaman & Nicobar : Cluster 2: 1977-2014

Even though the ACF & PACF plots suggest model with parameters :
p = 0, d = 0, q = 0 P = 5, D = 1, Q = 1 , we see there are too many parameters are to be estimated, thus making the model complex. Hence, we try to make these numbers lesser and make our model much simpler. For that to be done we try with small values of p, q, P, Q and compare the model's AICC values to pick the best one

p	d	q	P	D	Q	AICC
0	0	0	5	1	1	10.37
0	0	0	1	1	1	10.36
0	0	0	0	1	0	11.37
0	0	0	0	1	1	10.35

Suggests : p = 0, d = 0, q = 0 P = 0, D = 1, Q = 1

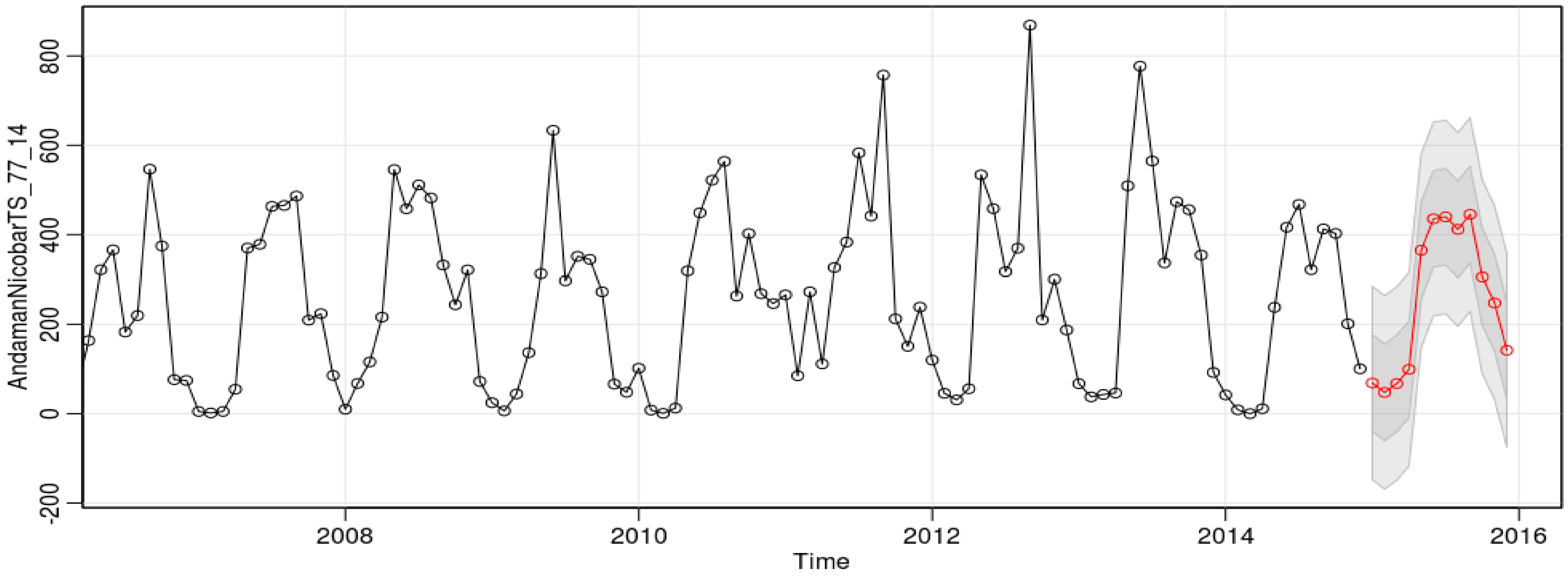
Subdivision: Andaman & Nicobar : Cluster 3: 1977-2014



Everything looks good.

Subdivision: Andaman & Nicobar : Cluster 3: 1977-14

Now we forecast 2015 based on this data and using SARIMA model (0,0,0),(0,1,1) and calculate the RMSE



RMSE = 61.80561

Subdivision: Andaman & Nicobar : Cluster 2: 1977-14

Thus we can say that our forecasting was not that bad.

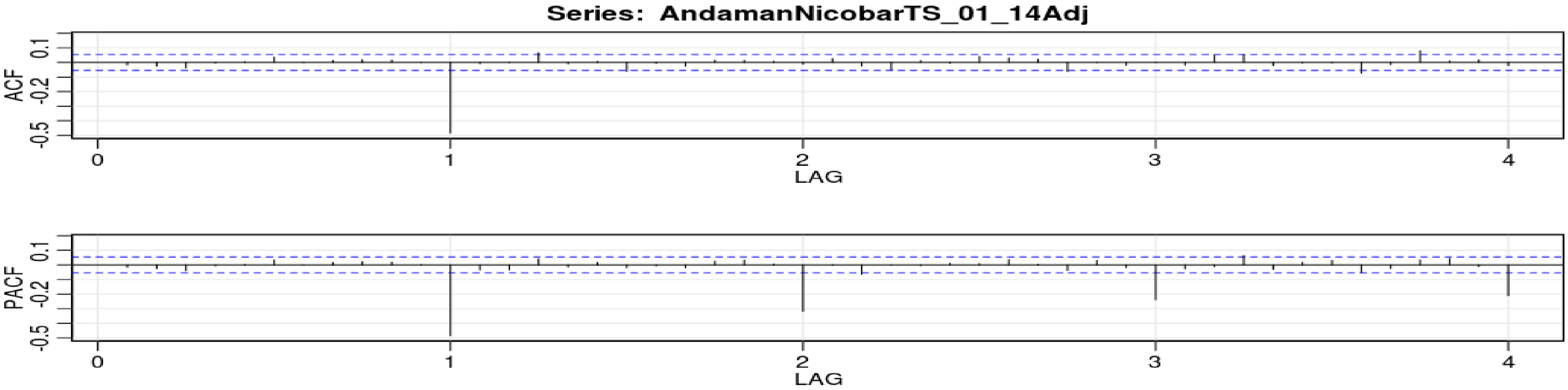
Subdivision: Andaman & Nicobar : Using the whole dataset to predict for the year 2015

Now we use the whole dataset with the predicted values for years 1943-45 to forecast for years 2015

Dickey Fuller test for Stationarity. The following result showsstationarity.

Test Statistic	p-values
-8.8413	0.01

ACF and PACF plots showing : p = 0, q = 0 , d =0 , P = 4, D =1, Q =1



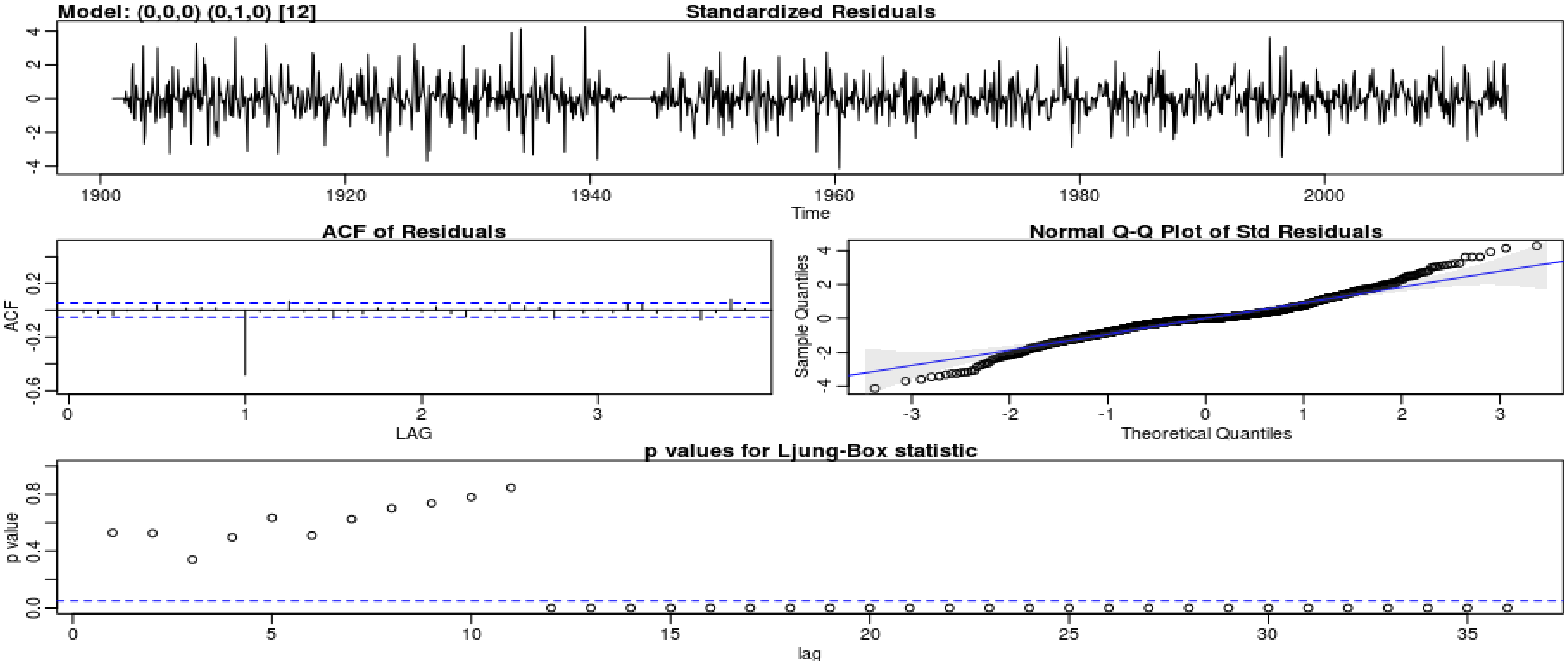
Subdivision: Andaman & Nicobar : Using the whole dataset to predict for the year 2015

Even though the ACF & PACF plots suggest model with parameters :
p = 0, d = 0, q = 0   P = 4, D = 1, Q = 1 , we see there are too many parameters are to be estimated, thus making the model complex. Hence, we try to make these numbers lesser and make our model much simpler. For that to be done we try with small values of p, q, P, Q and compare the model's AICC values to pick the best one.

p	d	q	P	D	Q	AICC
0	0	0	4	1	1	10.50964
0	0	0	1	1	1	10.50504
0	0	0	0	1	0	11.15895
0	0	0	0	1	1	10.50382

Suggests : p = 0, d = 0, q = 0   P = 0, D = 1, Q = 1

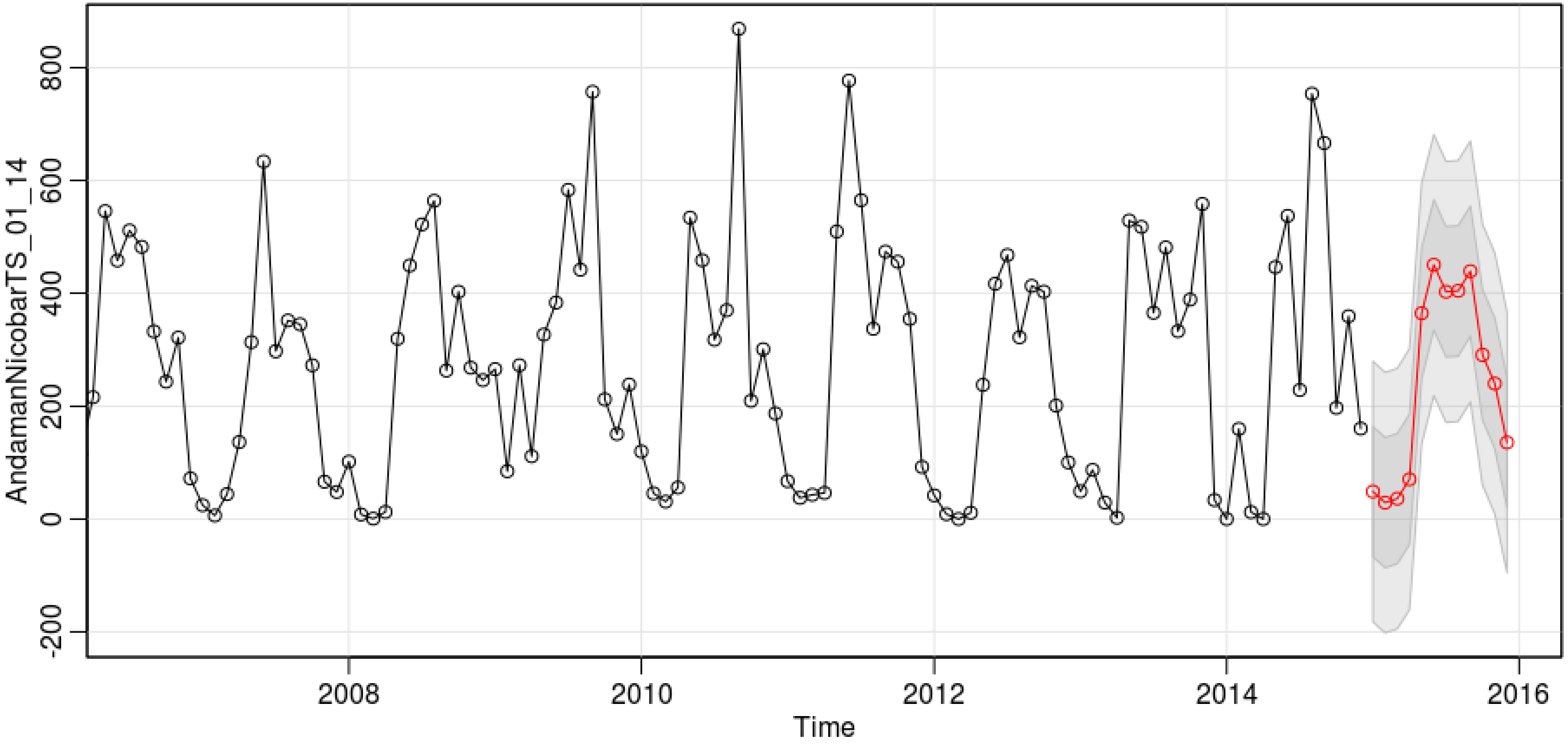
Subdivision: Andaman & Nicobar : Using the whole dataset to predict for the year 2015



Everything looks good.

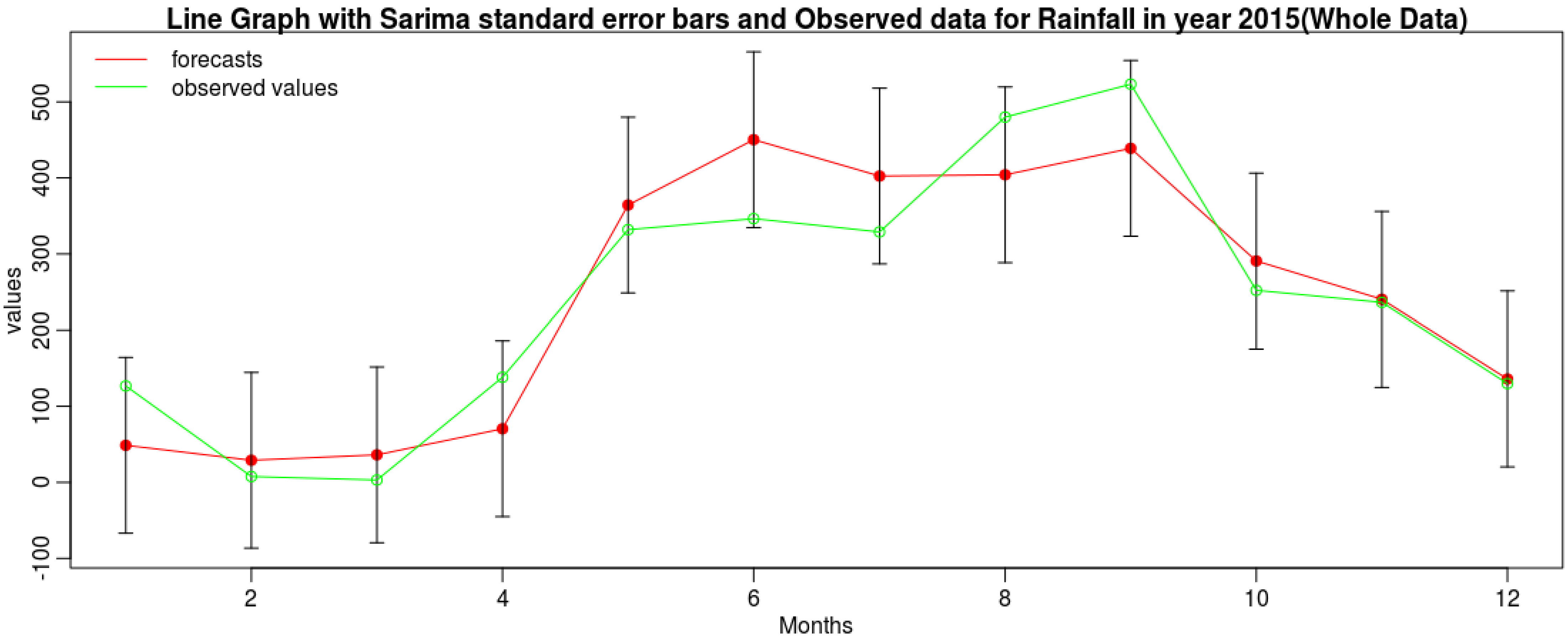
Subdivision: Andaman & Nicobar : Using the whole dataset to predict for the year 2015

Now we forecast 2015 based on this data and using SARIMA model (0,0,0),(0,1,1) over the whole data and calculate the RMSE



RMSE = 60.49766

Subdivision: Andaman & Nicobar : Using the whole dataset to predict for the year 2015



Thus we can say that our forecasting was not that bad. And using the whole data makes the forecast better as the RMSE comes lowerand all the actual values are in between the standard errors.

Subdivision: Andaman & Nicobar : Alternate way of testing if the forecast is good enough

Using Student's T-test:

Hypothesis:

H0 : Mean of observed values = Mean of forecasted values

H1 : Mean of observed values not equal to mean of forecasted values

Year	Mean of Observed	Mean of Forecasted	Test Statistic	Degrees of freedom	95% lower C.I	95% upper C.I	p-value
1977	199.0167	252.792	0.81057	22	-83.81025	199.0167	0.4263
2015	242.0417	256.2803	0.21059	22	-125.9851	154.4624	0.8351
2015(Using whole data)	242.0417	242.6088	0.0082207	22	-142.5145	143.6488	0.9935

Thus, we see in each of the cases the p-values are greater than 0.05 thus, we cannot reject the null hypothesis with 95% confidence. Thus we can say they are equal, i.e. mean of the observed values and predicted values are "equal". Hence we can say our forecasting is quite good.

References:

(For Using T-test)
American Journal of Mathematics and Statistics 2015, 5(2): 82-87
DOI: 10.5923/j.ajms.20150502.05
SARIMA Modelling
A. C. Akpanta 1,* , I. E. Okorie 1 , N. N. Okoye 21
Department of Statistics, Abia State University, Uturu, Nigeria
Agromet Unit, National Root Crops Research Institute Umudike, Nigeria

THANK YOU