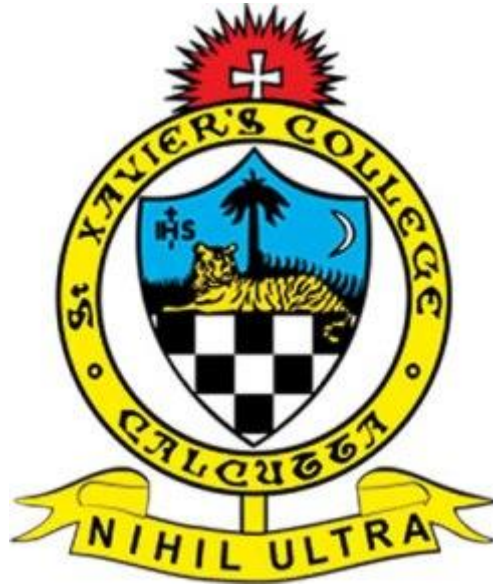


**DEPARTMENT OF STATISTICS**  
**ST. XAVIER'S COLLEGE (AUTONOMOUS), KOLKATA**



**Name: Akash Roy**

**Roll No. : 468**

**Registration Number: A01-1112-0738-22**

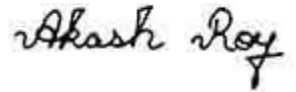
**Supervisor: Prof. Rahul Roy**

**Session : 2022-2025**

**Title: Mobile Share in Device Market: An ARIMA based forecasting approach**

## **DECLARATION**

I affirm that I have identified all my sources and that no part of my dissertation paper uses unacknowledged materials.



---

Akash Roy

Department of Statistics

St. Xavier's College (Autonomous), Kolkata

Date: April,2025

## **ACKNOWLEDGEMENT**

I hereby extend my sincere appreciation to Father Principal Rev. Dr. Dominic Savio, S.J., and the Department of Statistics for offering me the distinguished opportunity to work on this dissertation. Above all, I owe a debt of gratitude to my mentor, Prof. Rahul Roy, who has showered constant support, intelligent criticism, and invaluable advice, that have been crucial to my thesis. Additionally, I owe special thanks to all my department professors who provided me with the statistical foundation required for the conduction of this research. Further, I shall always be grateful to my family and peers for their constant shower of support and motivation. Their unaltered encouragement, and critical interest in my work, gave me the boost of confidence and tenacity which were needed to accomplish the task. Lastly, I'm highly appreciative of everyone who has extended their direct or indirect contributions to making this dissertation project a success.

Akash Roy

St. Xavier's College (Autonomous), Kolkata

---

## **CONTENTS**

<u><b>SL. No.</b></u>	<u><b>Topic</b></u>	<u><b>Page No.</b></u>
<b>1</b>	Introduction	5
<b>2</b>	Literature Review	6
<b>3</b>	Keywords and Terminologies	7
<b>4</b>	Residual Measures	9
<b>5</b>	Data Description	10
<b>6</b>	Analysis	11
	Data Preparation	11
	Check for presence of Autocorrelation	11
	Estimating difference parameter “d”	11
	Estimating MA parameter “q”	13
	Estimating AR parameter “p”	15
	Choosing the best model	16
<b>7</b>	Fitting the model	17
<b>8</b>	Residual Analysis	18
<b>9</b>	Forecasting	20
<b>10</b>	Conclusion	21
<b>11</b>	Future Scope	21
<b>12</b>	Bibliography	22

## **INTRODUCTION**

From being considered as a luxury suited only for a limited section of population, to being an actual necessity now, the market domain of mobile phones in India, has undergone significant transformation, ushering in the age of rapid technological growth and progress. Owing to several factors including falling prices, cost-effective plans, and widespread availability of internet facilities, mobile phones became increasingly popular in the late 2000s, especially from brands like Nokia, Samsung, and Apple. As time progressed, the country witnessed the rise of quite a few domestic brands like Micromax, Lava that offered even more budget-friendly options to the greater population. With time, Xiaomi, Realme, and Vivo entered the global smartphone market, and captured consumers with their lucrative deals and prices. The Indian market is still evolving, incorporating every new-age technology that cater to the modern needs of the consumers and business.

### **Why is it necessary to forecast the share of mobile in device market?**

Life in modern age is primarily driven by a single force – transformation. Be it in technology or domestic market, transformation is inevitable. And hence, to be in tandem with the changing needs, businesses, legislators, investors, and people in general must be aware of the proportion mobile devices occupy in the device market. Being now considered a necessity, it is imperative to understand the trajectory of the mobile market to take informed and wise decisions.

Share forecasting of mobile device market is especially important for the businesses who need to have an account of the market leaders and their competitors. By gauging the profits and losses of various companies through share trends, businesses can make necessary alterations in their strategy, product marketing, or device specifications, to remain ahead in the fast-paced competitive market. This forecasting strategy has been successfully used by several companies like Apple and Samsung to maintain their predominance in the market by incorporating the latest trends and cost-effective options to the masses.

Another necessity of forecasting the share of mobile in device market is resource allocation and the spotting of newer opportunities. Unlike a segmented marketplace where there are multiple contenders holding roughly equal market shares, an informed business company restricts the expansion of new entrants. It does not only inform the companies of their rival's strengths and weaknesses, profits and losses, but also gives a window to learn from the market trends and modify the products for better reach. Besides being aware of the market changes, forecasting the shares also helps in predicting the succeeding events and changes that might take place in the market.

---

## **LITERATURE REVIEW**

ARIMA models have been widely used to forecast observations in various sectors including health and technology. One such article by **Raydonal Osipina , Joao A.M. Gondim, Victor Leiva and Cecilia Castro (2023)** throws light on the fact that ARIMA models have been extensively applied to forecast trends in cases to inform policy decisions and make healthcare resource allocation. Of the models, ARIMA has been investigated to generate short-term predictions. While ARIMA models have been found less effective for highly dynamic, multifaceted scenarios, they remain a useful resource for short-term forecasting. Whereas ARIMA models have shown forecast potential, the shortfalls in long-term forecasts make it worth investigating further. Hybrid models fusing machine learning with ARIMA might improve the accuracy of forecasting by extracting non-linear relationships and external factors. A study by **Xu Ye (2010)** shows that compared to other prediction models, including exponential smoothing or machine learning, the ARIMA model is a straightforward but consistent prediction tool for mobile adoption. This research's conclusions are applicable for telecommunication companies and policymakers to use in strategic planning and the development of infrastructure.

---

## **KEYWORDS AND TERMINOLOGIES**

**Time Series Data:** Let  $(\Omega, F, P)$  be a probability space. Let  $T$  be an index set (subset of real line). Let  $X_t(w)$  be a real valued function defined on  $T \times \Omega$  with  $t \in T$  and  $w \in \Omega$ . For fixed  $t=t_0$ ,  $X_{t_0}(w)$  is a random variable defined on  $(\Omega, F, P)$ . Thus  $(X_t(w), t \in T)$  is a collection of random variables. For fixed  $w=w_0$ ,  $X_t(w_0)$  is a real valued function of  $t$ . The collection of all realisations (as  $w$  varies in  $\Omega$ ) is referred to as an ensemble. A time series is a single realisation from the ensemble of all possible ensembles.

**Moving Average Process :** Let  $\{Z_t\}$  be a purely random process with mean 0 and variance  $\sigma_z^2$ , then  $\{X_t\}$  is said to be a moving average process of order  $q$  (MA( $q$ )) if

$$X_t = \beta_0 Z_t + \beta_1 Z_{t-1} + \dots + \beta_q Z_{t-q}$$

where  $\beta_i$ 's are the model parameters.

**Auto Regressive Process :** Let  $\{Z_t\}$  be a purely random process with mean 0 and variance  $\sigma_z^2$ , then  $\{X_t\}$  is said to be an auto regressive process of order  $p$  if

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \dots + \alpha_p X_{t-p} + Z_t$$

where  $\alpha_i$ 's are the model parameters known as partial auto correlation coefficient. The above model is like a multiple regression model. Here  $X_t$  is regressed on its past values instead of separate predictors.

**Differencing :** Differencing is a method employed to transform a time series into a stationary one by calculating the difference between consecutive values. This technique helps eliminate trends and seasonal effects, making the data more suitable for analysis and modelling.

**ARIMA (Auto Regressive Integrated Moving Average) :** It is a combination of auto regressive, differencing and moving average process used for time series analysis and forecasting. It has 3 parameters: 'p' for the order of AR process, 'd' for the number of differencing required to make the time series stationary and 'q' for the order of MA process. Let  $\{Z_t\}$  be a purely random process with mean 0 and variance  $\sigma_z^2$ , then a process  $\{X_t\}$  is said to be ARIMA( $p, d, q$ ) if

$$X_t = \mu + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \dots + \alpha_p X_{t-p} + Z_t - \beta_1 Z_{t-1} - \dots - \beta_q Z_{t-q}$$

**Durbin Watson Test:** It is a statistical test to detect the presence of autocorrelation at lag 1 in the residuals and is valid under certain assumptions. Let  $e_t$  be the residual given by  $e_t = \rho e_{t-1} + v_t$

We are to test  $H_0: \rho=0$  ag  $H_1: \rho \neq 0$

An appropriate test statistic is given by  $D = \frac{\sum_{t=2}^T (e_t - e_{t-1})^2}{\sum_{t=1}^T e_t^2}$ .

The test rule is given by: if  $D$  is close to 2 then we accept  $H_0$  i.e. there is no autocorrelation between the residuals whereas values close to 0 and 4 indicates positive and negative autocorrelation respectively.

Augmented Dickey Fuller Test: It is a statistical test used to determine whether a time series is stationary or not. The ADF test considers the possibility of a linear trend in the time series and can also handle autoregressive (AR) terms in the model. The test involves estimating a regression model of the form:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \theta_1 \Delta y_{t-1} + \theta_2 \Delta y_{t-2} + \dots + \theta_p \Delta y_{t-p} + \varepsilon_t$$

where  $\Delta y_t$  is the first difference of the time series,  $t$  is a linear trend term,  $y_{t-1}$  is the lagged value of the time series, and  $\theta_1, \theta_2, \dots, \theta_p$  are the AR coefficients. The null hypothesis of the ADF test is that there is a unit root in the time series, indicating that the time series is nonstationary. The alternative hypothesis is that the time series is stationary. The ADF test statistic is based on the t-statistic of the coefficient  $\beta$ . If the absolute value of the test statistic is greater than the critical value at a given significance level, then the null hypothesis is rejected and the time series is considered stationary. If the absolute value of the test statistic is less than the critical value, then the null hypothesis cannot be rejected, indicating that the time series is non-stationary.

Run Test: A run is defined as an uninterrupted sequence of one symbol or attribute. It is a non-parametric test to verify whether the observations are random or not. The test procedure is given as follows:

We are to test  $H_0$  : Residuals are random against  $H_1$ : not  $H_0$

Let  $N$  = total number of observations

$N_1$  = No. of positive residuals

$N_2$  = No. of negative residuals

$R$  = No. of runs

Assumption:  $N_1 > 10$  and  $N_2 > 10$

An appropriate test statistic is given by:  $T = \frac{R - E(R|H_0)}{se(R|H_0)}$  where  $E(R|H_0)$  and  $se(R|H_0)$  are the expectation and standard deviation of the test statistic under the null hypothesis.

Under  $H_0$ ,  $T \sim N(0,1)$

Test Rule: Reject  $H_0$  at  $\alpha$  level of significance iff  $|T_{obs}| > \tau_{\alpha/2}$  where  $\tau_{\alpha/2}$  is the upper  $\alpha/2$  % point of a  $N(0,1)$  distribution.

---



## RESIDUAL MEASURES

We first define the forecasting error as  $e_t = x_t - F_t$  where  $x_t$  is the observation of the actual time series at time  $t$ ,  $F_t$  is the forecast at time  $t$  and  $e_t$  is the error due to forecast at time  $t$ . Let there be  $n$  observations.

The following are the residual measures used in this dissertation to check the efficacy for various models. The measures are:

- Akaike Information Criterion (AIC) : AIC is a statistical measure used to evaluate and compare models by considering both their fitness and their complexity. It helps in selecting the most suitable model for forecasting by balancing accuracy and simplicity. The formula is given by :  $AIC = 2k - 2(\log\text{-likelihood})$ , where  $k$  is the number of model parameters

- Mean Absolute Error (MAE): It is a commonly used measure. Regardless of positive or negative error, it reports the absolute magnitude of error committed.

$$MAE = \frac{1}{n} \sum_t |e_t|$$

- Root Mean Square Error(RMSE) : RMSE is given by  $\sqrt{\sum_t \frac{(x_t - \bar{x}_t)^2}{n}}$  where  $\bar{x}_t = \frac{1}{n} \sum_t x_t$

- Mean Percentage Error (MPE) : This gives the average percentage deviation of the forecast from the actual series.

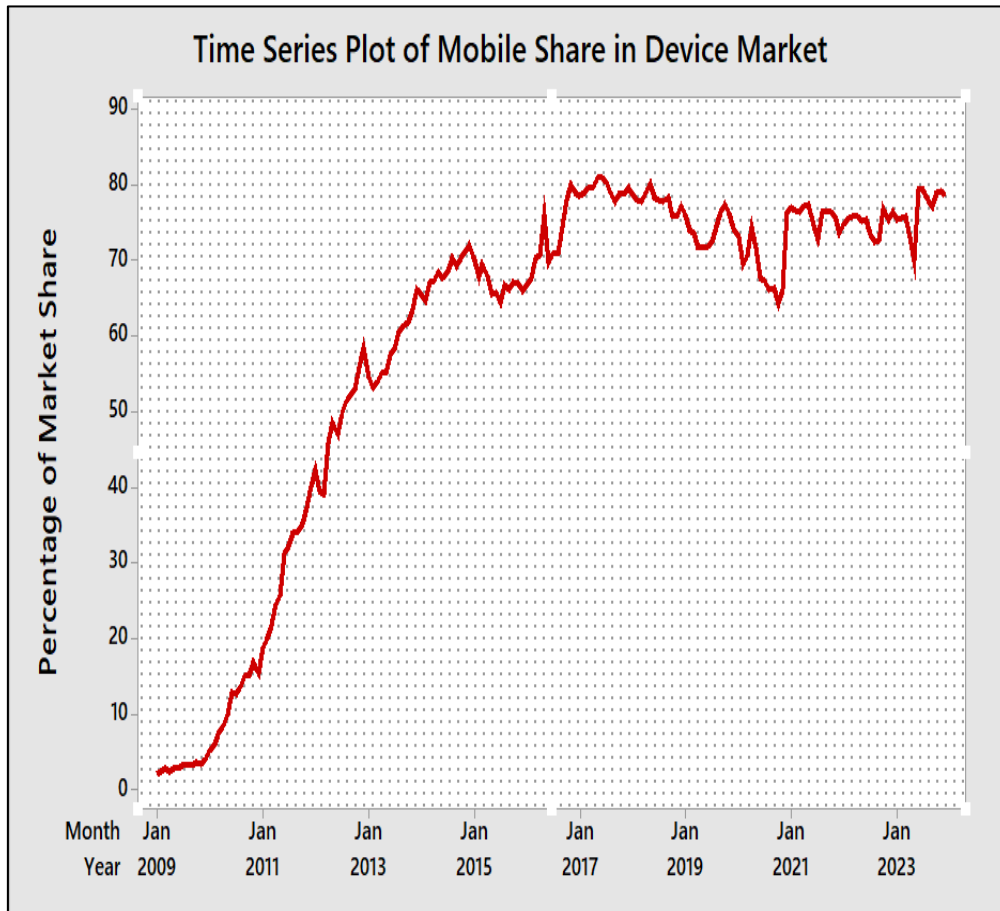
$$MPE = \frac{1}{n} \sum_{t=1}^n \frac{e_t}{x_t} \times 100$$

- Mean Absolute Percentage Error(MAPE) : This gives the average absolute percentage deviation of the forecast from the actual series.

$$MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{e_t}{x_t} \right| \times 100$$

## DATA DESCRIPTION

The data used in this project is monthly data on the share of mobile in the electronic device market expressed as percentage of market share rounded to 2 decimal places from 2009 to 2023. Let  $X_t$  denote the market share of mobile corresponding to time point  $t$ . The variable  $t$  takes integer values from 1 to 180. We obtain the time-series plot by plotting  $X_t$  values against  $t$  values.



*The time-series plot*

From the plot, we observe the presence of an evolutive pattern dominated by an upward trend. There is no presence of seasonality. There seems to be a presence of cyclic patterns, though not well pronounced and increases with increasing trend. Thus, the components seem to be interdependent, and a multiplicative model would be appropriate for such a time series data.

## ANALYSIS

### Data Preparation:

We shall divide the dataset into two parts – training and testing dataset. They are as follows:

Training dataset: Data from January 2009 to December 2021.

Testing dataset: Data from January 2022 to December 2023.

We shall be fitting an appropriate model on the training data and measure its efficacy by comparing it with the testing dataset.

### Test for Presence of Autocorrelation: (Durbin Watson Test)

We shall resort to the Durbin Watson Test to check for the autocorrelation between the residuals. From R , we get:

```
lag Autocorrelation D-W Statistic p-value
1      0.9783991    0.02163236      0
Alternative hypothesis: rho != 0
```

As discussed earlier, the value of the Durbin Watson d-statistic being 0.021632 which is close to 0 indicates the presence of a positive autocorrelation. It is further confirmed by the p value which is less than our desired level of significance (0.05). Hence, there is a presence of autocorrelation between the observations which makes our dataset fit for time series analysis.

### Estimating the value of difference parameter ‘d’:

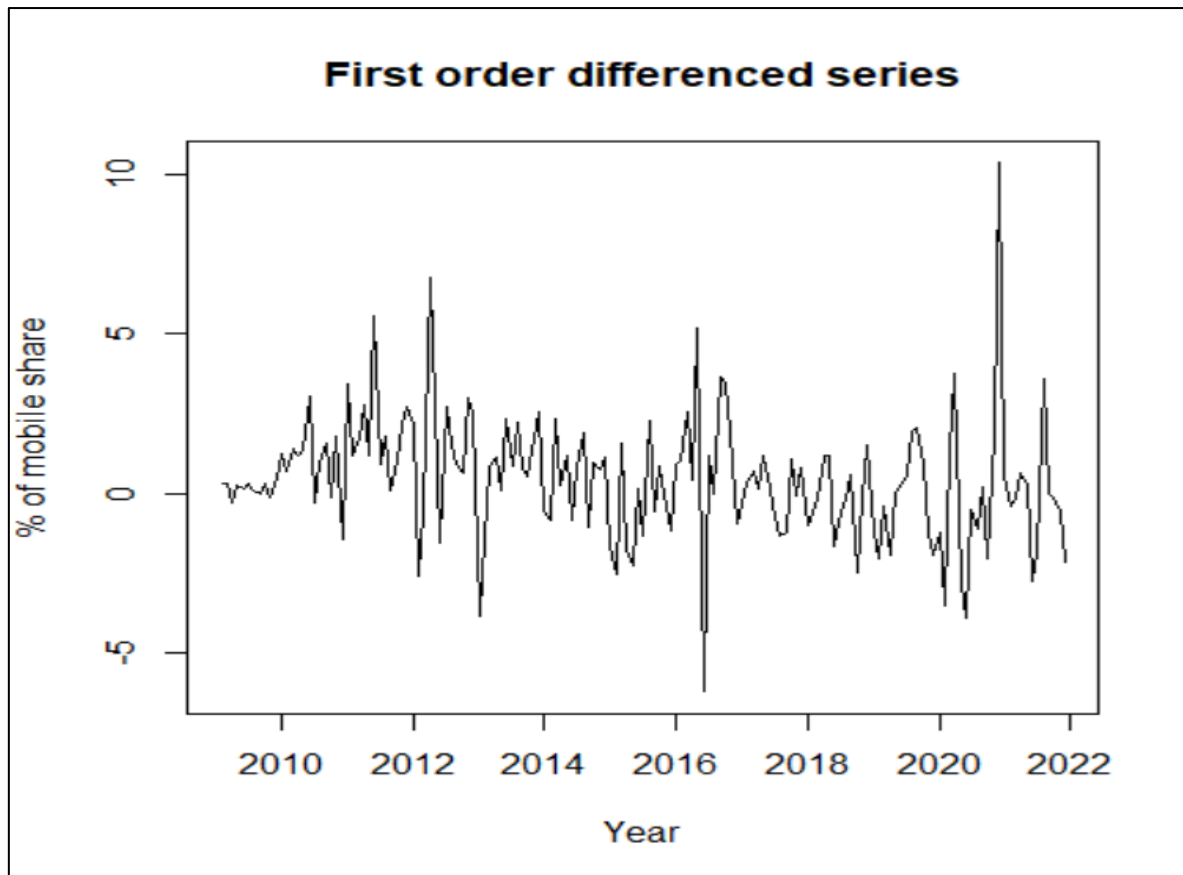
We shall resort to the Augmented Dickey Fuller test (which has been discussed earlier) to check for the stationarity in the dataset. At first , we applied the test on the original training dataset and got the following result using R.

```
Augmented Dickey-Fuller Test

data: train_data
Dickey-Fuller = -1.1233, Lag order = 5, p-value = 0.9159
alternative hypothesis: stationary
```

The p value is higher than our desired level of significance, which is 0.05 , hence we accept  $H_0$  indicating a non-stationary time series. It is also in compliant with our time series plot from which we observe that the statistical properties like mean , variance and covariance changes with time.

To make it stationary , we compute the first and second order differenced series given as follows:

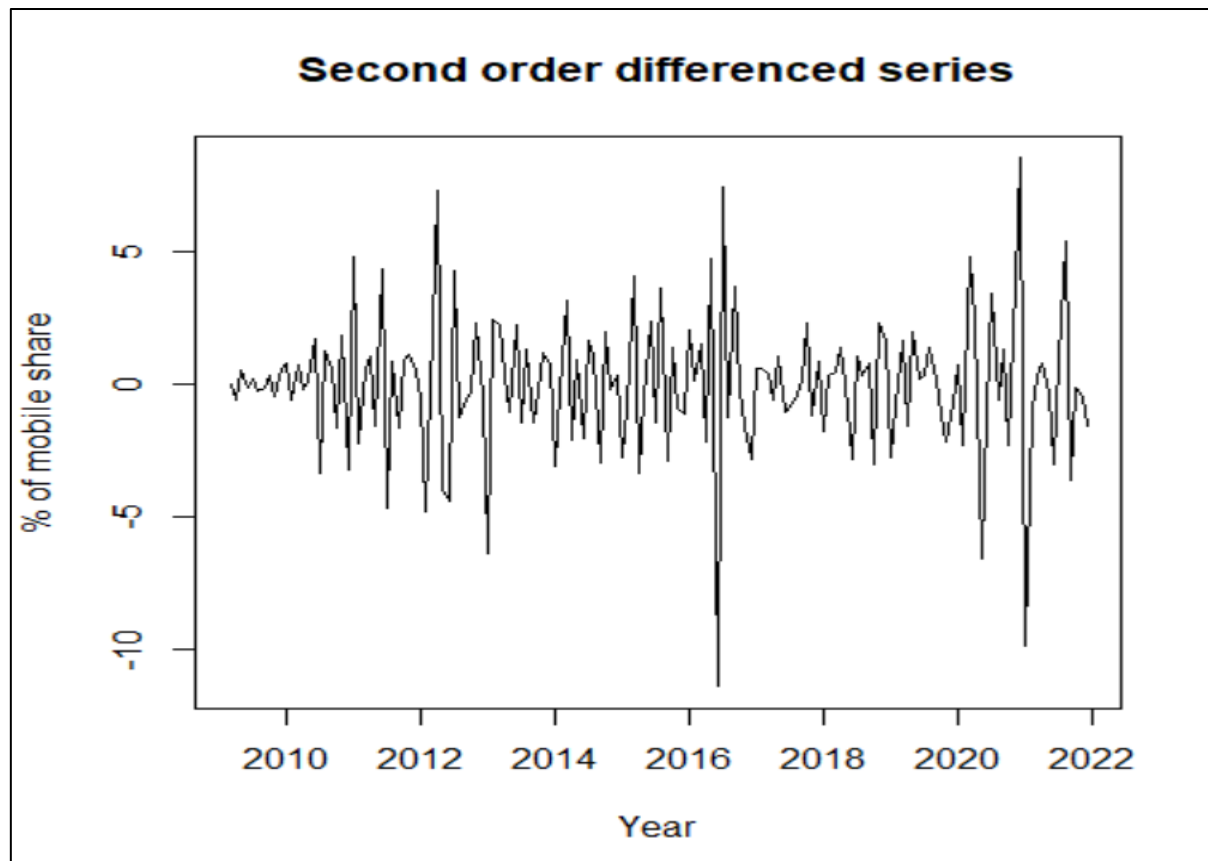


On applying ADF test on the above differenced series , we get the following results:

```
Augmented Dickey-Fuller Test
data: dl
Dickey-Fuller = -5.5753, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary
```

As the p-value is less than our desired level of significance (0.05) , we shall reject  $H_0$  (i.e. non stationarity). Hence, the first differenced series is stationary.

Further , the time series plot of the second order differenced series is given by:



#### Augmented Dickey-Fuller Test

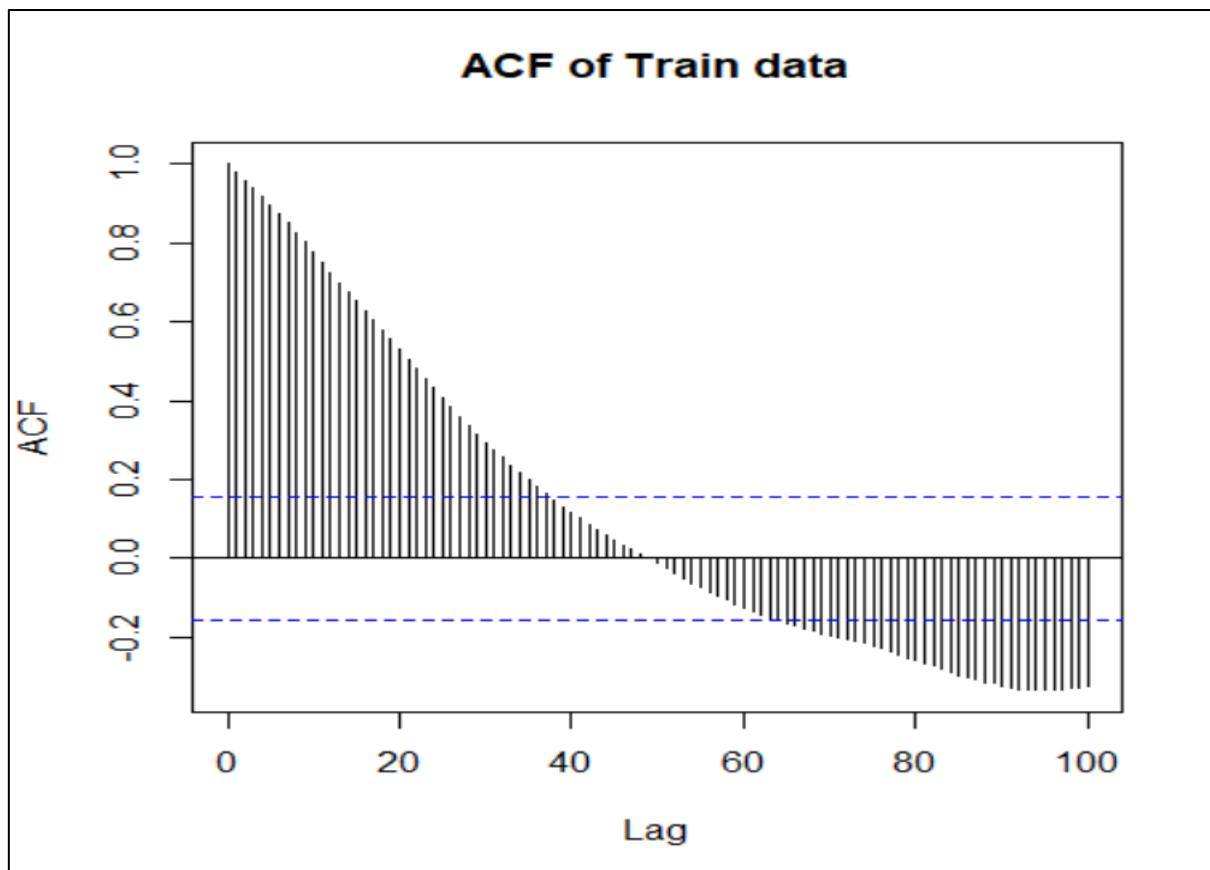
```
data: d2
Dickey-Fuller = -8.7942, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary
```

As the p-value is less than our desired level of significance (0.05) , we shall reject  $H_0$  (i.e. non stationarity). Hence, the second differenced series is also stationary.

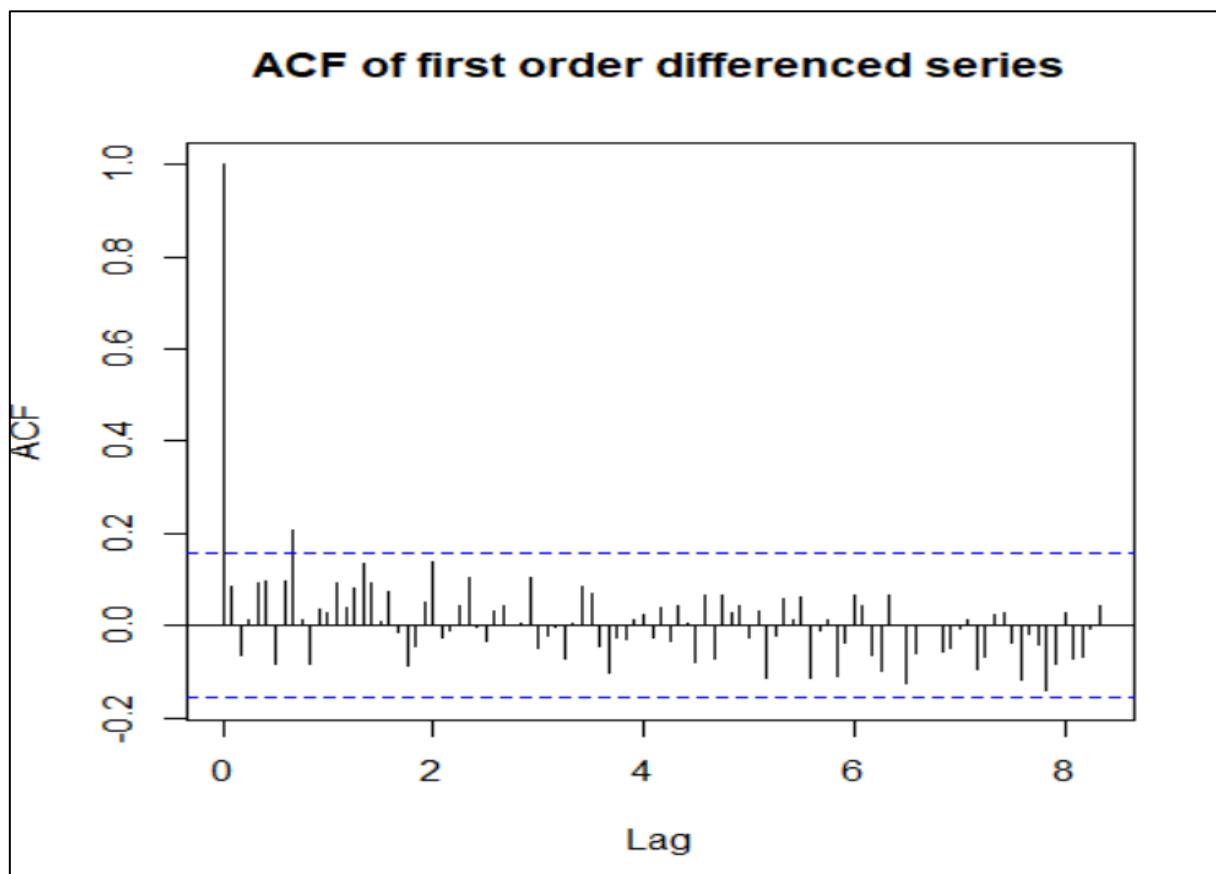
Hence , we shall be considering both 1 and 2 as the probable value of  $d$  and proceed with our study.

#### Estimating the MA parameter 'q':

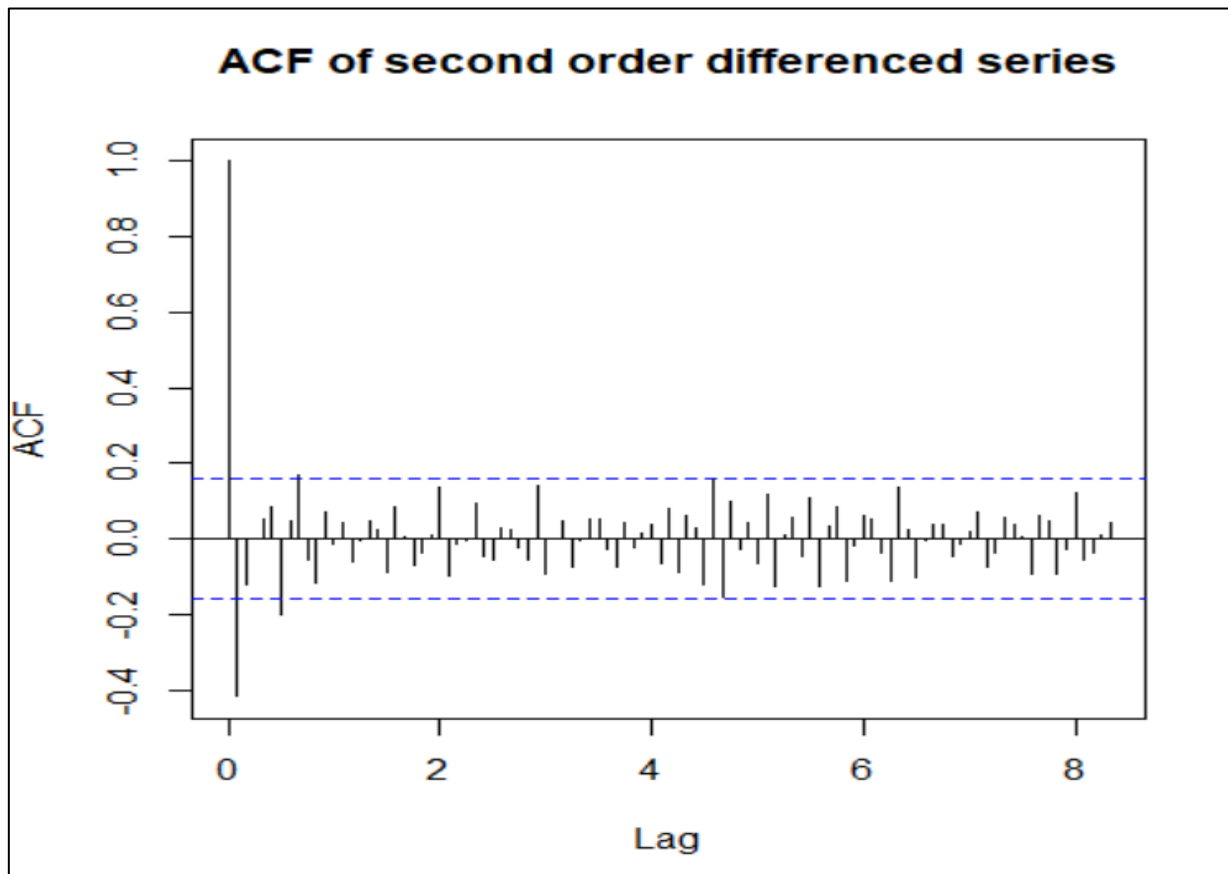
To get an idea of the MA parameter 'q' , we plot the ACF plot of the training dataset.



We cannot find the exact value of 'q' from the ACF plot hence we plot the ACF of the differenced series. The ACF plot of the first order differenced series is given below:



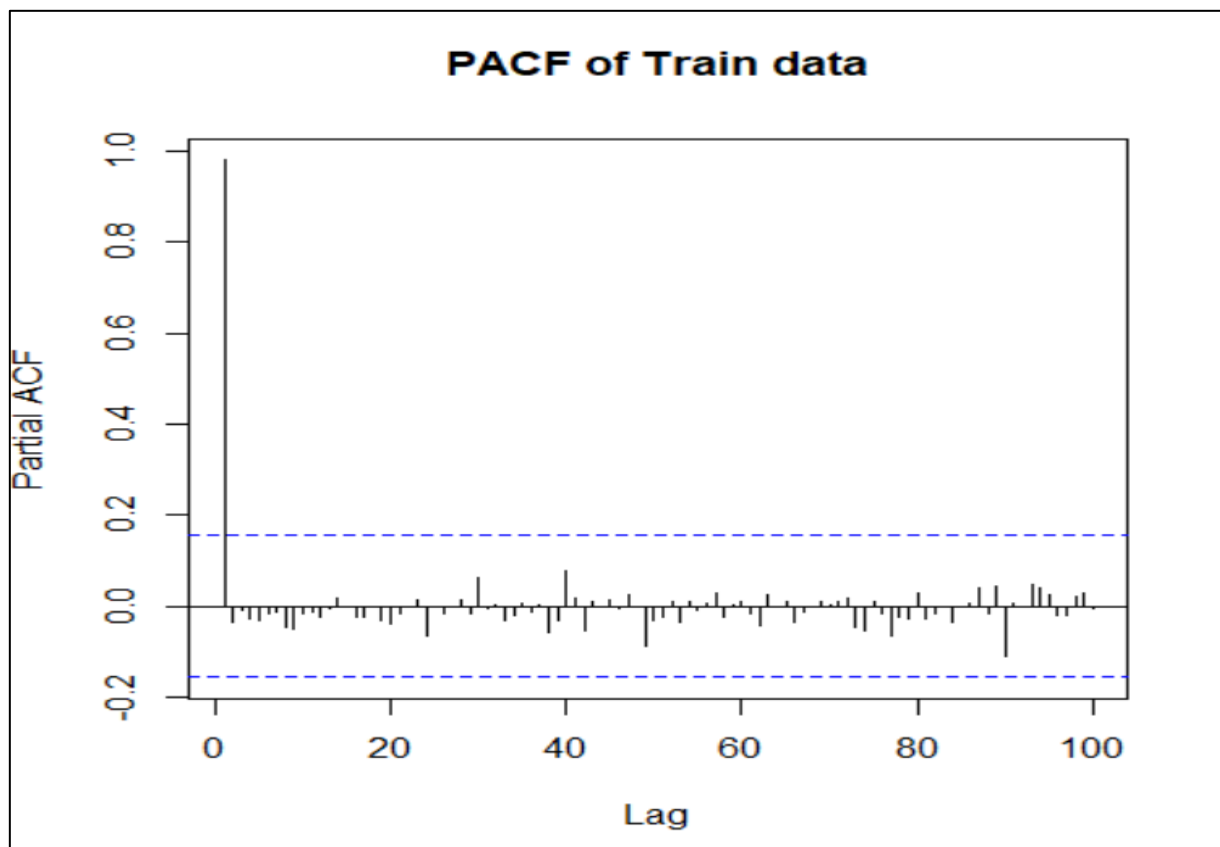
The ACF of the second order differenced series is given below:



Hence , we shall consider both 0 and 1 as the probable value of  $q$  and proceed with our study.

#### Estimating the AR parameter 'p':

To estimate the AR parameter  $p$  , we need to plot the PACF (i.e. partial auto correlation function) plot of the training data.



In the PACF plot, we observe that there is a significant spike only at 0 and cuts off after that. Hence, our AR parameter  $p$  for the model is 0.

Hence, we can have four combinations of parameters for the ARIMA model, namely  $(0,1,0)$ ,  $(0,1,1)$ ,  $(0,2,0)$ ,  $(0,2,1)$ .

### **Choosing the best set of model parameters:**

For each of the above-mentioned combinations of the model parameters, we compute different measures for goodness of fit.

Model	AIC	MAE	RMSE	MPE	MAPE
<b>(0,1,1)</b>	659.07	1.4081	1.9955	1.8301	3.7365
<b>(0,2,0)</b>	740.88	1.8426	2.6473	0.0185	4.4827
<b>(0,1,0)</b>	659.95	1.4364	2.0143	2.0744	3.8391
<b>(0,2,1)</b>	<b>650.99</b>	<b>1.3310</b>	<b>1.9484</b>	<b>0.3603</b>	<b>3.244</b>

### **Observation:**

We shall first observe the AIC of each of the model combinations. As a thumb rule, we claim that model to be better which has the least AIC. We observe that, the AIC of  $(0,2,1)$ ,  $(0,1,1)$  and  $(0,1,0)$  are close and smaller than that of  $(0,2,0)$ . Hence, we discard the parameter combination  $(0,2,0)$ .

Next, we check the other measures of the remaining three parameter combinations. We observe that MAE, RMSE, MPE and MAPE are least for the parameter combination  $(0,2,1)$ . Hence it seems that ARIMA(0,2,1) is the most appropriate model for our dataset.

---



## FITTING THE MODEL

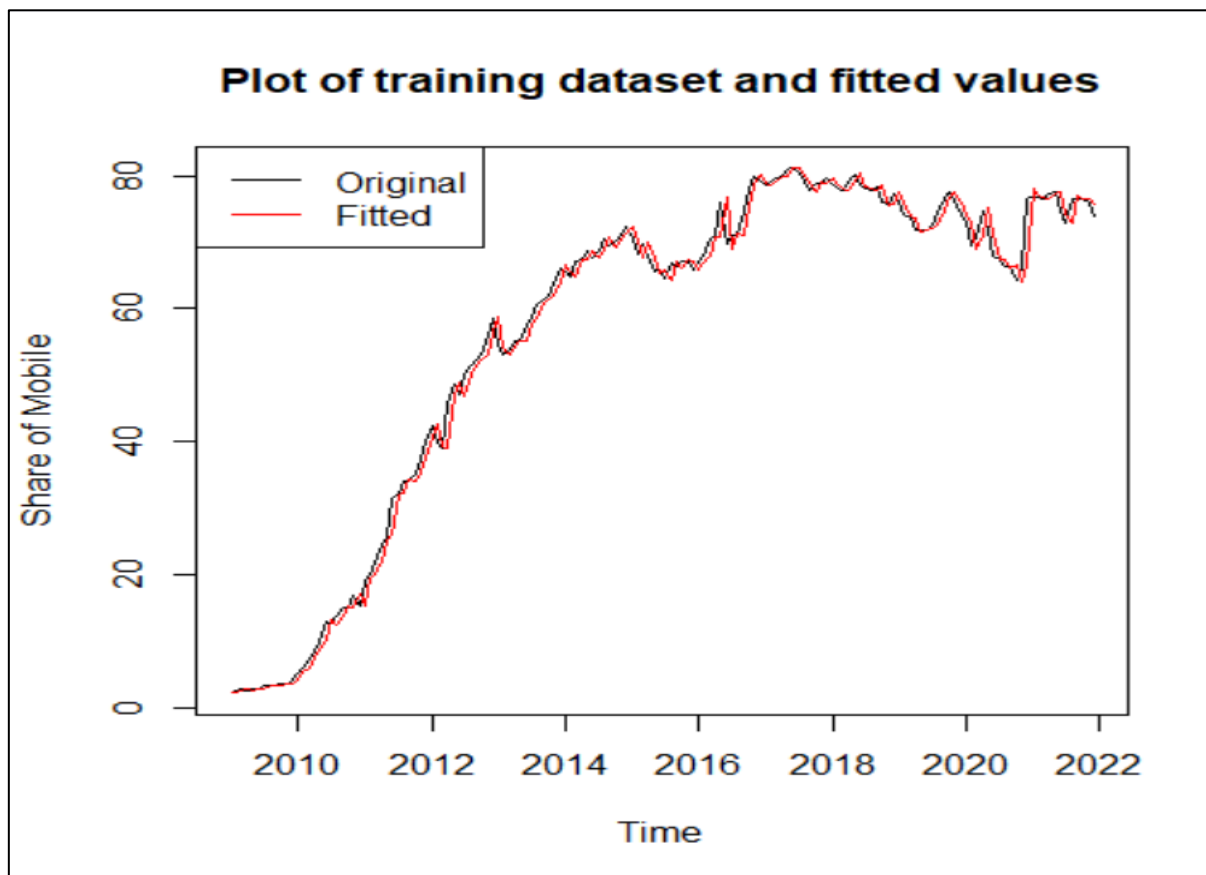
We fit ARIMA(0,2,1) to our data which is given by:

$$X_t = \mu + Z_t - \beta_1 Z_{t-1}$$

Where  $\mu$  is the mean term and  $\beta_1$  is the MA parameter.

Based on our training dataset, the estimate of  $\beta_1$  denoted by  $\hat{\beta}_1$  is given as -0.9590 and the estimate of  $\mu$  denoted by  $\hat{\mu}$  is given as -0.009  $\approx$  0. Hence, the intercept term becomes insignificant as it is almost equal to 0.

Further, we shall plot the actual observed values from the training data and the fitted values overlaid on the same time series plot to check the efficacy of our model. The plot is given below:



### Comment:

From the above graph, we observe that the fitted model is very close to the original data. Our fitted model closely follows the original data indicating a good fit. Also, we must note that the above graph is based on the training dataset and the judgement of goodness of the fit is merely based on the training dataset.

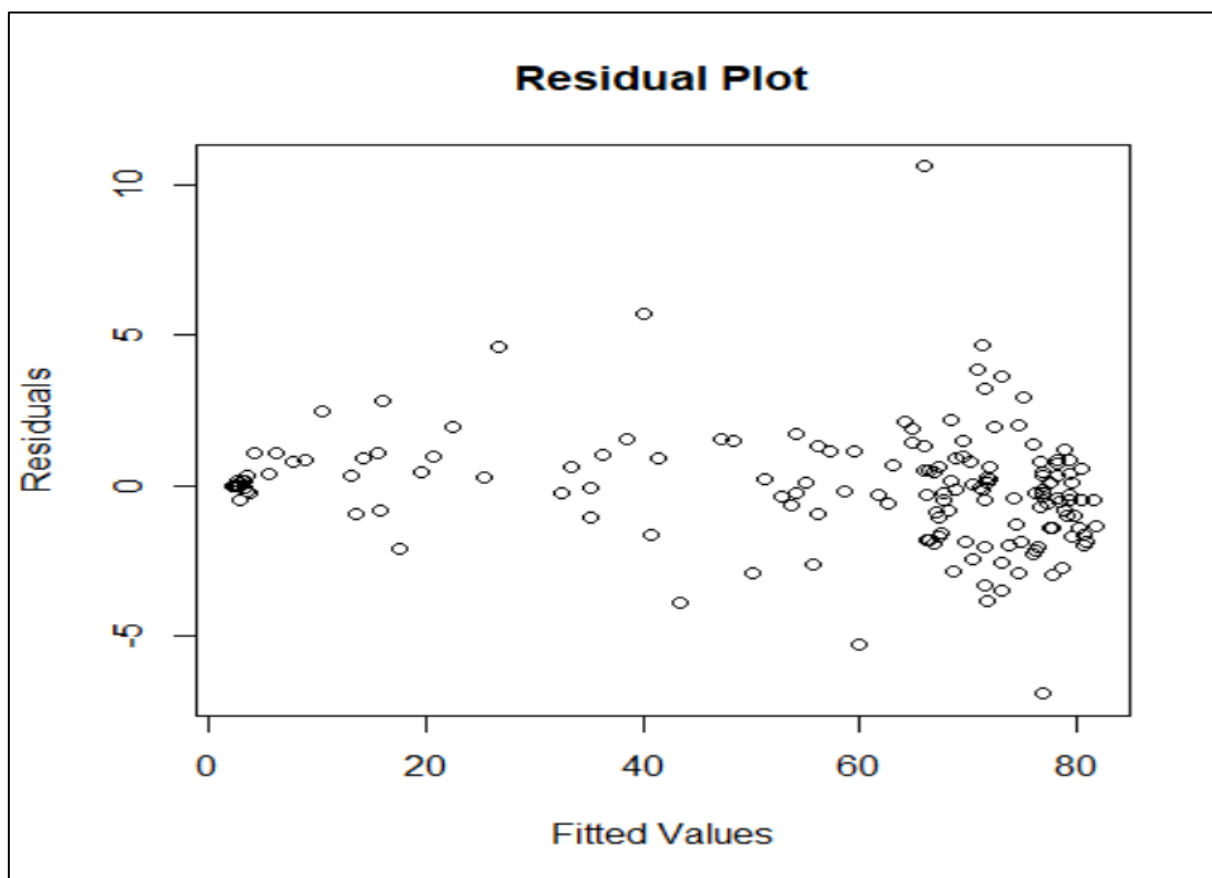
---

## RESIDUAL ANALYSIS

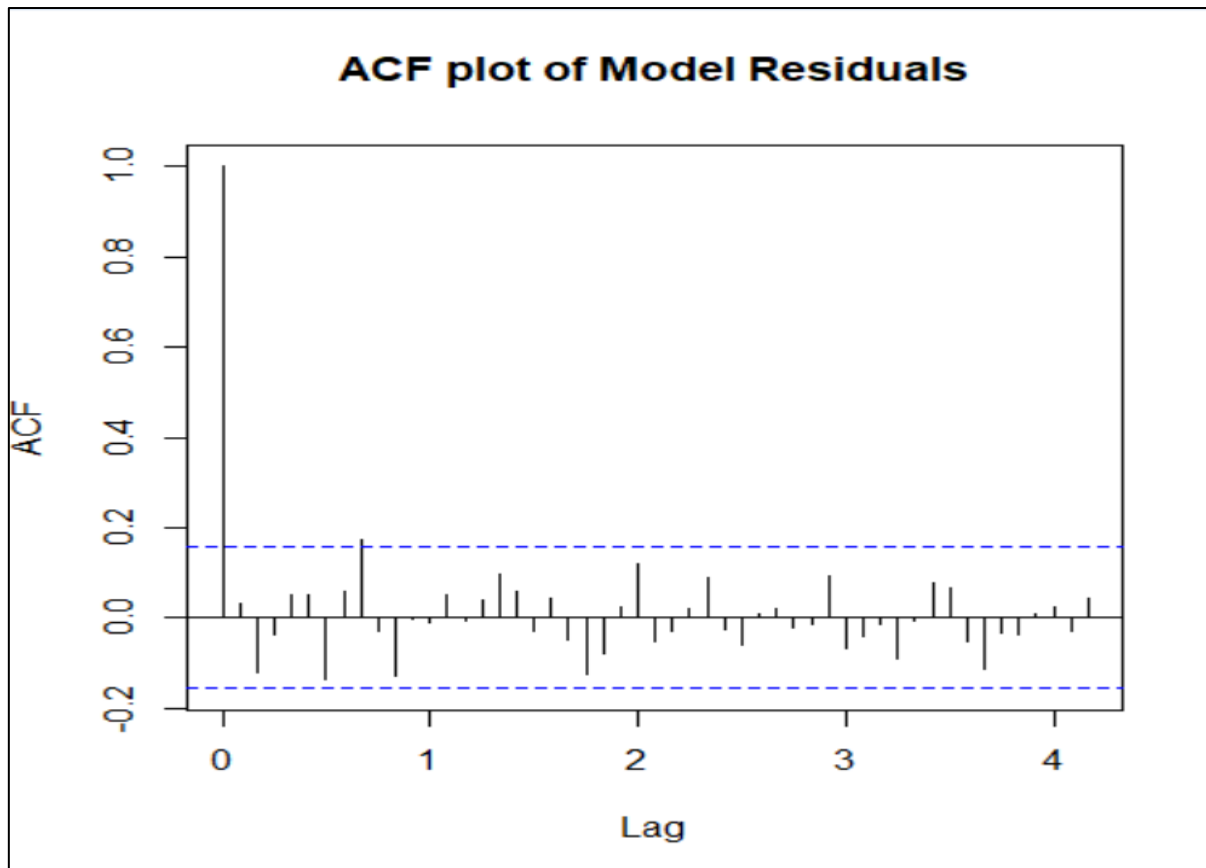
### What is the need for residual analysis?

Residual analysis in time series forecasting is essential for evaluating the accuracy and reliability of a model. By examining the residuals, which are the differences between actual and predicted values, analysts can assess whether a model captures the underlying patterns in the data effectively. Proper residual analysis helps identify issues such as autocorrelation, non-stationarity, or heteroscedasticity, ensuring that the chosen model is appropriate for making accurate forecasts.

To check whether the residuals are random or not, we obtain the residual plot by plotting the residuals against the fitted values.



Comment: The residuals seem to be randomly scattered i.e. there is no presence of non-random nature which indicates a good fit. There is a slight clustering at two points which is insignificant.



Comment: From the ACF plot , we observe that there is only a single spike at 0 which indicates a purely random process.

Further , we can verify the randomness in the residuals using RUN TEST.

#### Test for Randomness of Residuals:

We shall apply RUN test on the residuals to check for any non-randomness in their behavior. Using R , we obtain,

```
data: res
statistic = 0.48194, runs = 82, n1 = 78, n2 = 78, n = 156, p-value =
0.6298
alternative hypothesis: nonrandomness
```

As the p-value is much higher than the desired level of significance (0.05), we accept  $H_0$  ,i.e. the residuals are random in nature.

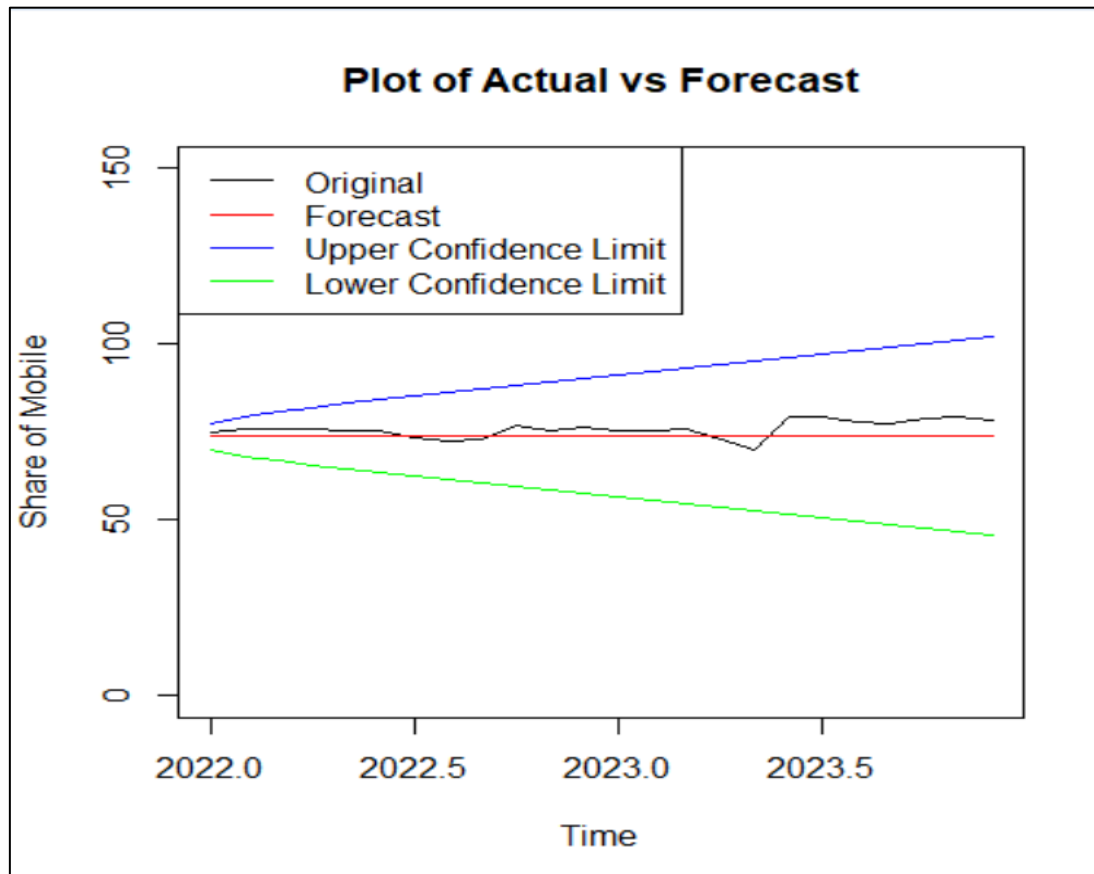
Now , as we have confirmed that our fitted model is appropriate for our dataset , we can proceed to forecasting the share of mobile in device market and compare it with our testing dataset.

## FORECASTING

To test the efficacy of our model , we shall compare our fitted model with the test dataset. Hence , we shall forecast the share of mobile in the device market for the year 2022 and 2023 using the forecast equation as given below :

$$\hat{x}_t = \hat{\beta}_1 x_{t-1}$$

The forecasted values are obtained using the forecast function in R. Next , we plot the test data, and the forecasted values overlaid on the same time series plot as given below:



Comment: The forecasted values seem to be very close to the test data which lies well between the upper and lower 95% confidence limits of our forecast.

Further , we shall compute the MAPE values of our forecast , which comes out to be **2.64**, which indicates a good fit for our ARIMA model.

## **CONCLUSION**

We have determined and set a proper model to forecast the mobile market share in the device market for a short-term future horizon.

ARIMA is an effective and stable method for investigating how past values affect current trends in a time series. Additionally, when it comes to forecasting, ARIMA proves to be superior to techniques like Simple Exponential Smoothing, courtesy to its greater flexibility in representing different time series components and structures, hence being a more trustworthy option for precise predictions.

## **FUTURE SCOPE**

The ARIMA model is inherently linear, hence it finds it difficult to tackle non-linear and dynamic relationships or sudden events capable of having dramatic impacts on the observations.

Such a weakness appears in the actual vs. fitted values plot where deviations from patterns are indicated. Cases where predictions drastically differ from patterns, other statistical methods such as Monte Carlo simulation can be used to produce good forecasts.

Moreover, further work could investigate embedding a seasonal element within the ARIMA function in R.

Although market share of mobile in device market are not amenable to simple seasonality, using external economic indicators via ARIMA with Regressors can address this problem.

This method combines ARIMA and multiple regression so that appropriate economic indicators that can affect the movement of these market shares can be included. Improving ARIMA models in this fashion can have a significant impact on improving forecasting performance.

---

## **BIBLIOGRAPHY**

The information used in this dissertation has been acquired from the following books and websites:

- “An Overview of Forecast Analysis with ARIMA Models during the COVID-19 Pandemic: Methodology and Case Study in Brazil” by Raydonal Osipina , Joao A.M. Gondim, Victor Leiva and Cecilia Castro published on July 12, 2023.
  - “The Application of ARIMA Model in Chinese Mobile User Prediction” by Xu Ye (Marketing Department, Alcatel Lucent Shanghai Bell Company Limited, China) published on 2010 IEEE International Conference on Granular Computing.
  - Introduction to Time Series and Forecasting by Peter J Brockwell , Richard A. Davies , published by Springer Publications
  - Fundamental of Statistics (Volume II) by A.M. Gun , M.K. Gupta and B. Dasgupta
  - <https://www.kaggle.com/datasets/michau96/device-market-in-india-over-last-15-years>
  - <https://otexts.com/fpp2/arima.html>
  - <https://www.geeksforgeeks.org/time-series-analysis-in-r>
-