

ANEKANT EDUCATION SOCIETY’s

Tuljaram Chaturchand College of Arts, Science

And Commerce, Baramati

A PROJECT REPORT ON

**“ NETWORK TRAFFIC ANALYTICS USING TIMES SERIES ANALYSIS”**

By

**KOMKAR VAISHNAVI RAHUL**

Under the guidance

Of

**Prof. N. A. Jagtap**

Year 2022-2023

**CERTIFICATE**

This is to certify that the project report entitled **“NETWORK TRAFFIC ANALYTICS USING TIME SERIES ”.**

This is being submitted by KOMKAR VAISHNAVI RAHUL as a partial fulfillment for the award of the degree of the Master of Science (M.Sc. II). This is a record of bonafide work carried out by them under my supervision and guidance.

Prof. N. A. Jagtap Dr. A. S. Jagtap

Project Guide Examiner Head of Department of statistics

PLACE: BARAMATI.

DATE: 11/10/2022

**ACKNOWLEDGEMENT**

This project would not have been possible without the guidance and the help of several individuals who in one way or another contributed and extended their valuable assistance in the preparation and completion of the study. I would like to express my deepest gratitude to my guide **Professor N. A. Jagtap**, Department of Statistics, whose guidance has been of immense help in successfully completing this report.

With great pleasure, I take this opportunity to express my sincere gratitude to my respected teachers, friends and non teaching staff of Department of Statistics, Tuljaram Chaturchand College (Autonomous), Baramati for their valuable advice and encouragement throughout my course. Last but not least, I am indebted to my parents for their unconditional love and support.

**INDEX**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Title** | **Page No.** |
| 1 | Abstract | 5 |
| 2 | Introduction | 6 |
| 3 | Objective | 7 |
| 4 | Methodology | 8 |
| 5 | Time series analysis   * Time plot * Decomposition of time series * Stationarity * ACF and PACF plot * Fitting of ARIMA model * Normality * Ljung-Box test * Forecasting * Accuracy | 9  9  10  11  12-13  14  15  16  16 |
| 6 | Conclusion | 17 |
| 7 | Reference | 18 |

**ABSTRACT**

The network traffic is the amount of data which moves across a network during any given time .network traffic may also be referred to as data traffic or just plain traffic .

In the academic literature,Network are more formally referred to as graph.we use time series prediction models and evaluate whether the model can predict network packets in a certain time,the AR,MA,ARMA and ARIMA model are applied ,we use the data ~500k csv with summary of real network data ,the data set has ~21k rows and covers 10 local IPs were compromised at same point during this period .

This dataset is originally from the https://[www.kaggle.com](https://www.kaggle.com/)/website. This study

explores network trafficdata with time series techniques. The primary aim of this study is to predict future web /network traffic to make decisions for better congestion control

**Keywords**: network traffic, Trend, ARIMA, forecast

**Software used**: MS-Excel, MS-Word, R-Software, ITSM

**INTRODUCTION**

Network traffic is **the amount of data which moves across a network during any given time**. Network traffic may also be referred to as data traffic or just plain traffic. In search engine optimization (SEO), traffic to a network can be characterized as being either direct, organic or paid. Internet traffic is a time series which can be used **to predict the future traffic trends in a computer network**.

Time Series Forecasting is one of the least explored areas and various models are evaluated to improve the accuracy of the forecast. The main focus of the proposal is **to predict future web traffic to make decisions for better congestion control**. Past Values are considered to predict future values.

It is very important to research traffic time-series model in order to describe traffic

behavior. The traditional long time-scale traffic model can only model smoothing

process and some special non-smoothing process. **AR [4]** (Auto Regressive) model,

MA (Moving Average) model, and **ARMA [5]** (Auto Regressive Moving Average)

model can deal with smoothing process. **ARIMA [5]** (Auto Regressive Integrated Moving Average) model and **ARIMA seasonal model [6]** can describe the uniform

non-smoothing process. A large-scale network itself is a complex non-linear system,

and is influenced by many environment factors, which is similar with water-volume

time series that can be decomposed into mutation item, periodic item, trend item,

**random item [7].** So network traffic also can be considered to be the combination of

**periodic item, trend item, random item, and mutation item**, which is very difficult to

describe these traffic characters with a traditional traffic time-series model.

**OBJECTIVE**

* To fit the trend
* Decompose model to check trend and sesonality present or not
* To evaluate whether the model can predict network traffic exactly or not
* To fit the best model.
* Forcast the future prediction

**METHODOLOGY**

This project is based on network traffic ,we analyse the data by using time series analysis. The given data is taken from [www.kaggle.com](http://www.kaggle.com) website ,the data contains ~21k rows and covers 10 workstation IPs over a three month period .

First and foremost step in this analysis is study the data in details. One of the main aims of the study is to predict future web traffic to make decisions for better congestion control. Since the time dependent factors are present in the data, so that the time series technique is adapted to model and Some important terms that are used in time series analysis is as follows.

* Stationary Time Series
* Auto Regressive Integrated Moving Average (ARIMA) model
* Auto Correlation Function(ACF)
* Partial Auto Correlation Function (PACF)
* Forecasting
* Model adequacy
* Normality
* Ljung-Box test
* **EXPLORATORY DATA ANALYSIS (EDA)**

**RCODE**

library(readxl)

> library(forecast)

> library(tseries)

> attach(nwk)

> class(nwk)

[1] "data.frame"

> summary(nwk)

OUTPUT:

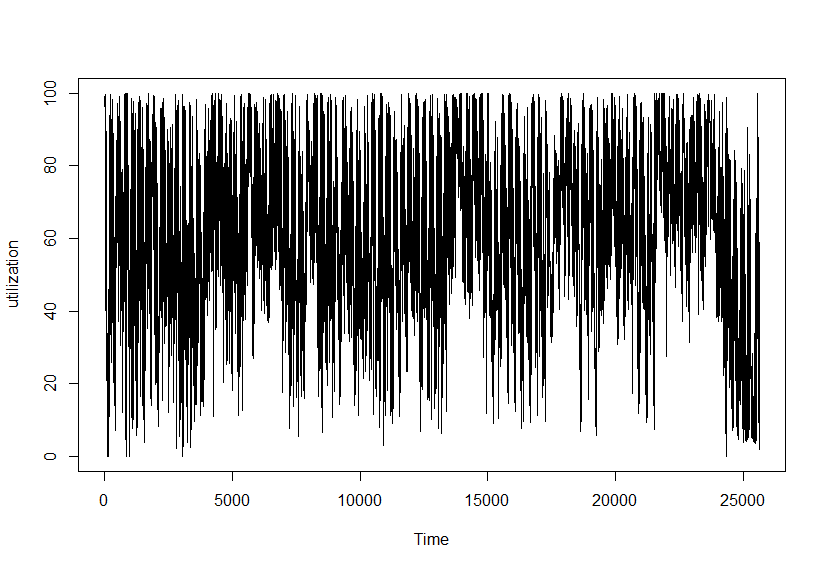
|  |  |  |
| --- | --- | --- |
| Length:25631 | Class :character | Mode :character |
| Mean : 60.37234 | Min. : 0.00005 | Min. : 0.00005 |
| 1st Qu.: 44.87385 | Median : 62.10020 | 3rd Qu.: 77.70215 |

**Time series analysis**

**Time series plot:**

**R CODE**

> ts.plot(utilization)

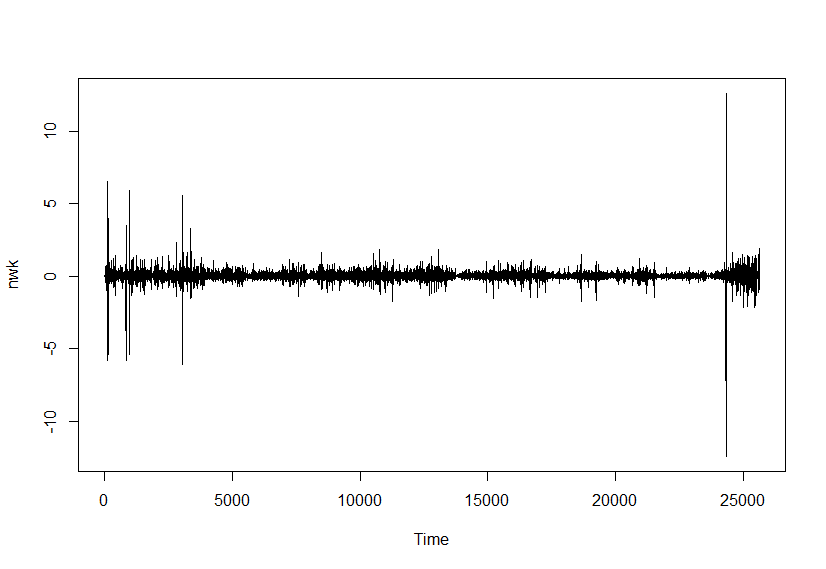


**As data is too noisy ,because it is sampled every 5 minutes ,we will resample it to 1H period**

**RCODE**

> nwk=diff(log(utilization))

> plot.ts(nwk)



Here,we can see that, if we can put a line on zero ,there is constant mean is observed in the given plot.

**Conclusion:**

From time plot we can see here, there is trend in this time series data, as the network traffic is changes overall in time. Therefore this is stationary. And seasonal component may be not present. With this in mind we proceed to the decomposition of the time series.

**Decomposition of time series:-**

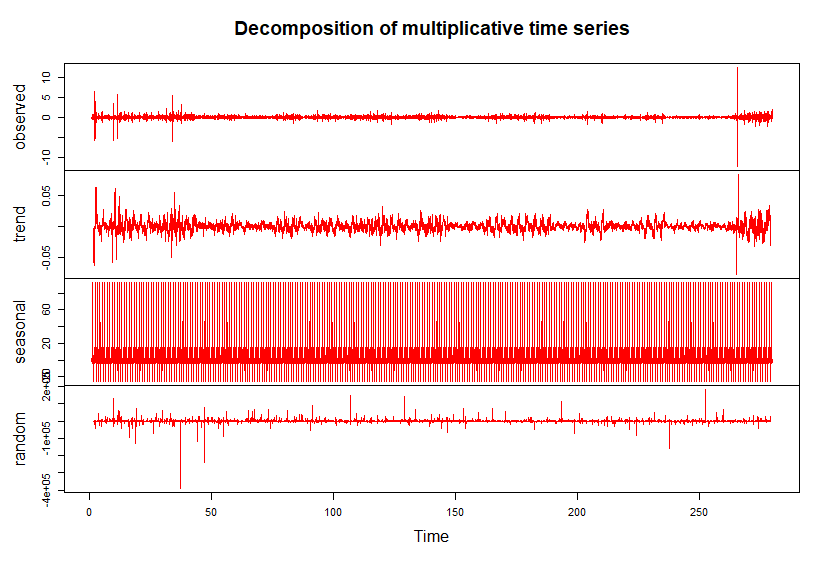
R CODE

> tsdata=ts(nwk$utilization,frequency=92) ~(:. 3 month observation ,daily data)

>tsdata2=diff(log(tsdata))

>d=decompose(tsdata2,type='multiplicative')

>plot(d,col='red',xlab='Time',ylab='NETWORK TRAFFIC')



**Conclusion:-**

From the graph we see that, there is trend in network traffic over time.

There seasonality is absent .

**Stationarity:-**

**Augmented Dickey Fuller Test**

**Hypothesis to be tested:**

H0: The given time series data is non-stationary. V/s

H1: The given time series data is stationary.

Augmented Dickey-Fuller Test

data: nwk

Dickey-Fuller = -36.734, Lag order = 29, p-value = **0.01**

alternative hypothesis: stationary

Here p-value= 0.01<0.05

Therefore, we may rejectt null hypothesis.

**Hence, given time series data is stationary.**

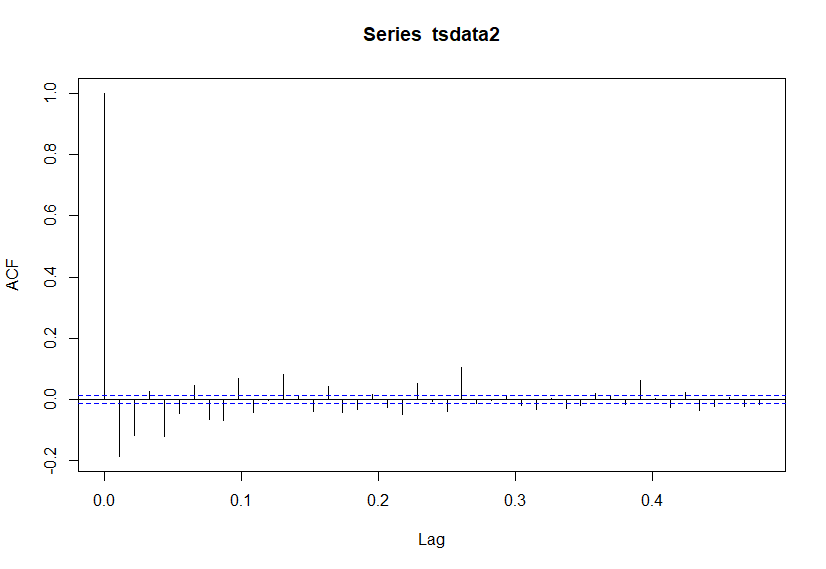
**ACF and PACF plot:-**

**ACF plot:-**

**RCODE**

>acf(tsdata2)

>pacf(tsdata2)



**Conclusion:**

* In ACF graph
* There are several autocorrelations that are significantly zero. Therefore, the time series is random.
* The model has MA (5) order.

**PACF plot:-**

|  |
| --- |
|  |

**Conclusion:**

* From PACF graph, High degree of autocorrelation at lag 0.25.

* The model has AR(5) order.

From ACF and PACF plot, given time series is stationary. i.e. it follows White Noise.

**Fitting of ARIMA model:-**

**RCODE**

> **auto.arima(nwk)**

Series: nwk

ARIMA(4,0,5) with zero mean #**best model has order(4,0,5)**

Coefficients:

ar1 ar2 ar3 ar4 ma1 ma2 ma3 ma4 ma5

0.1532 -0.2039 0.7239 -0.2998 -0.4208 0.0861 -0.8099 0.3403 -0.0251

s.e. 0.2194 0.0607 0.0649 0.1386 0.2194 0.0147 0.0171 0.1762 0.0112

sigma^2 = 0.06405: log likelihood = -1146.82

AIC=2313.63 AICc=2313.64 BIC=2395.15

Now ,we again fit the model by using given order(4,0,5)

**>model=arima(nwk,order=c(4,0,5));model**

Call:

arima(x = nwk, order = c(4, 0, 5))

Coefficients:

ar1 ar2 ar3 ar4 ma1 ma2 ma3 ma4 ma5 intercept

**0.0309 -0.2367 0.6883 -0.2218 -0.2986 0.0863 -0.8035 0.2413 -0.0264 -1e-04**

s.e. 0.3792 0.1024 0.1114 0.2426 0.3791 0.0145 0.0244 0.3079 0.0110 4e-04

sigma^2 estimated as 0.06403: log likelihood = -1147.29, aic = 2316.57

**The given ARIMA model is given by**

**nwk=-0.001+0.0309nwk(t-1)-0.2367nwk(t-2)+0.6883nwk(t-3)-0.2218nwk(t-4)-0.2986(et-1)+0.0863(et-2)-0.8035(et-3)-0.0264(et-4)**

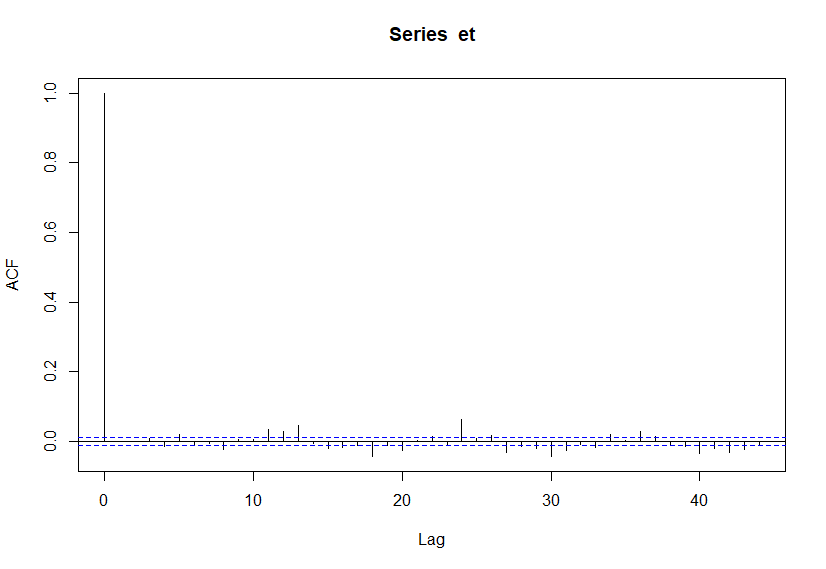
**Conclusion: -** The best model is ARIMA(0,2,1).

**Normality:-**

**To check normality assumptions of the model**

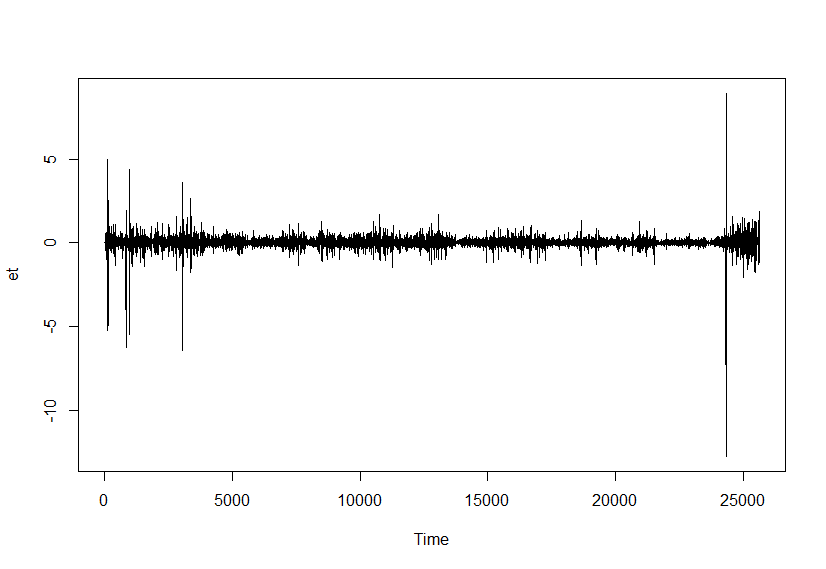
> et=residuals(model)

> acf(et)



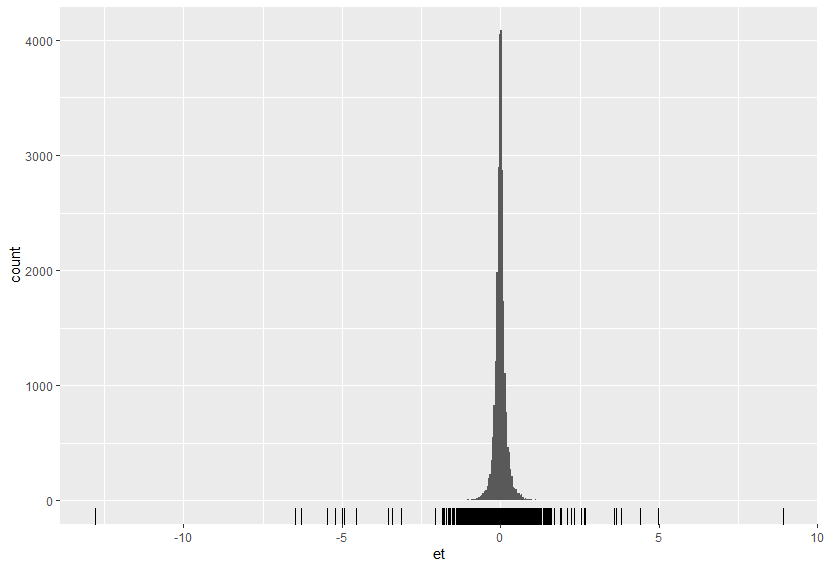
Here we will say that the below pipes are not touching the control limit line upto lag 6 ,so we can say that there is not autocorrelation ;residuals are independent

> plot.ts(et)



Conclusion: If we put straight line at zero ,there is constant mean in residuals at 0

> gghistogram(et)



**According to histogram we can see graph is normally distributed**

**Ljung-Box test:-**

To check data is independently distributed or not. We use Ljung-Box test.

**Hypothesis to be tested:-**

H0: The data are independently distributed. V/s

H1: The data are not independently distributed.

**RCODE** :

> Box.test(et)

Box-Pierce test

data: et

X-squared = 9.282e-05, df = 1, p-value = **0.9923**

**HERE Pvalue >0.05 hence we accept Ho**

**Hence, the data may not independently distributed.**

**Forecasting:-**

f=forecast(model,h=12,level=c(95))

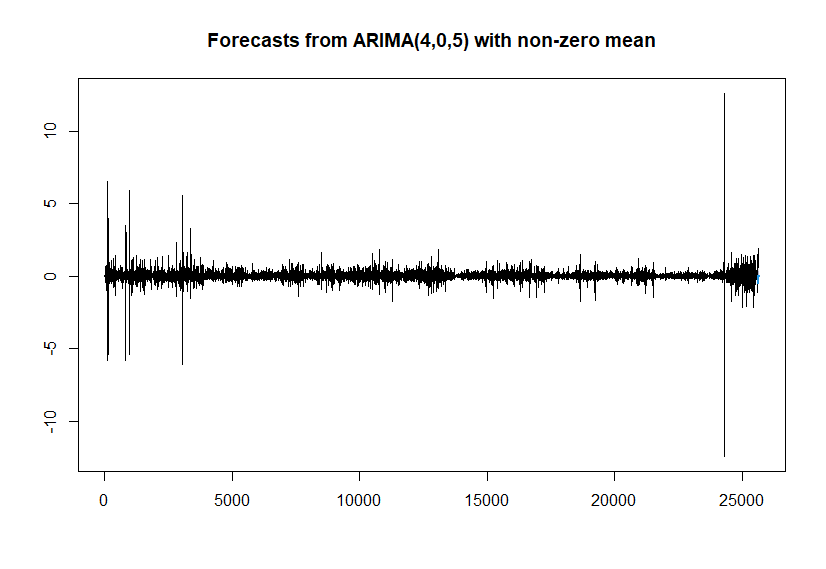
> f

|  |  |
| --- | --- |
| time | Point Forecast |
| 25631 | -0.535333452 |
| 25632 | -0.275199623 |
| 25633 | -0.277270077 |
| 25634 | -0.240020344 |
| 25635 | -0.038140743 |
| 25636 | -0.074252471 |
| 25637 | -0.097052444 |
| 25638 | 0.041485344 |
| 25639 | -0.018454718 |
| 25640 | -0.060784657 |
| 25641 | 0.052505585 |
| 25642 | -0.005948556 |

>

**Conclusion:-** From above table, we get future forecasted values for tax revenue. We can see here the network traffic increases by 5 minutes .

**Forecast Plot:-**

****

**Accuracy**

**accuracy(model)**

|  |
| --- |
| ME RMSE MAE MPE MAPE MASE ACF1 |
| Training set -1.059576e-06 0.2530447 0.1298779 NaN Inf 0.7103595 6.017927e-05 |

**Conclusion**

* From time series plot, trend present in the data.
* From decomposition, we can see only trend is present ,seosonality is absent
* The time series of network traffic is stationary
* From ACF and PACF plot, given time series is stationary.
* The best model is ARIMA(4,0,5) i.e. AR(4) and MA(5) model.
* The forecasted values of tax revenue in India for next 4 years:-

|  |  |
| --- | --- |
| time | Point Forecast |
| 25631 | -0.535333452 |
| 25632 | -0.275199623 |
| 25633 | -0.277270077 |
| 25634 | -0.240020344 |
| 25635 | -0.038140743 |
| 25636 | -0.074252471 |
| 25637 | -0.097052444 |
| 25638 | 0.041485344 |
| 25639 | -0.018454718 |
| 25640 | -0.060784657 |
| 25641 | 0.052505585 |
| 25642 | -0.005948556 |

* Residuals are normally distributed.
* The given time series data is independently distributed.

**Reference**

* [www.google.com](http://www.google.com)
* <https://www.kaggle.com>
* Introduction to Time series and Forecasting (Brockwell Devis)
* Applied Time Series Analysis (Dr. Marcel Dettling)
* Research paper on Time Series Decompose model of Network Traffic