# A Detailed Comparative Analysis Towards Longer Terms Traffic Loads Forecasting with Autoregressive Integrated Moving Average (ARIMA/SARIMA) Models to Improve Transportation Services

BORIS PENGILI[1], DIMITRIOS A. KARRAS[2,3,a]

[1] EPOKA University, Tirana, ALBANIA
[2] National and Kapodistrian University of Athens, GREECE
[3] EPOKA university, Comp. Engineering Dept., Tirana, ALBANIA
[a]0000-0002-2759-8482

*Abstract:-* Traffic flow forecast will be a great help in life of big cities citizens, especially in a time where road system cannot handle rush hours loads without constraining citizen to wait in traffic line and the population is in growth. The dataset used was captured by loop detectors that evaluate speeds length, time etc. In this paper, ARIMA and SARIMA were the models built for the prediction of the number of cars consisting of the traffic load. Although this study has been conducted for Tirana city, the methodology and discussions should be relevant to any big city. First, we transformed the time series to stationary with log scale transformation. After that, we found the right parameters for our models. Then we compared the results of two models: ARIMA (which we built with auto-Arima) and SARIMA, where Arima had the best outcome for the given dataset. The results were very satisfactory and with the Arima model we can make accurate forecasts for at least 3 months, showing that not only short-term forecasts are possible but even longer-term traffic load forecasting might be viable.

*Key-Words:* - Time series, Forecasting, ARIMA, SARIMA, Traffic Load forecasting

Received: April 29, 2024. Revised: November 20, 2024. Accepted: December 27, 2024. Published: February 24, 2025.

**List of Abbreviations**

| | |
|---|---|
| **ANN** | Artificial neural networks |
| **AR** | Auto regression |
| **ARIMA** | Autoregressive integrated moving average |
| **ED** | Euclidean Distance |
| **MA** | Moving average |
| **MAD** | Mean absolute deviation |
| **MAE** | Mean absolute error |
| **MAPE** | Absolute percent error |
| **MSE** | Mean square error |
| **RMSE** | Root mean square error |
| **SARIMA** | Seasonal autoregressive integrated moving average |

# 1 Introduction

## 1.1 Background of the study

With thousands of people moving at the same hours of day, one of the Albania main concerns is that its road system cannot handle rush hours loads without constraining citizen to wait in line for many hours. In fact, this problem exists in many other countries, but here in Albania the traffic congestion is worse than most other countries because the roads are not in such a good condition.

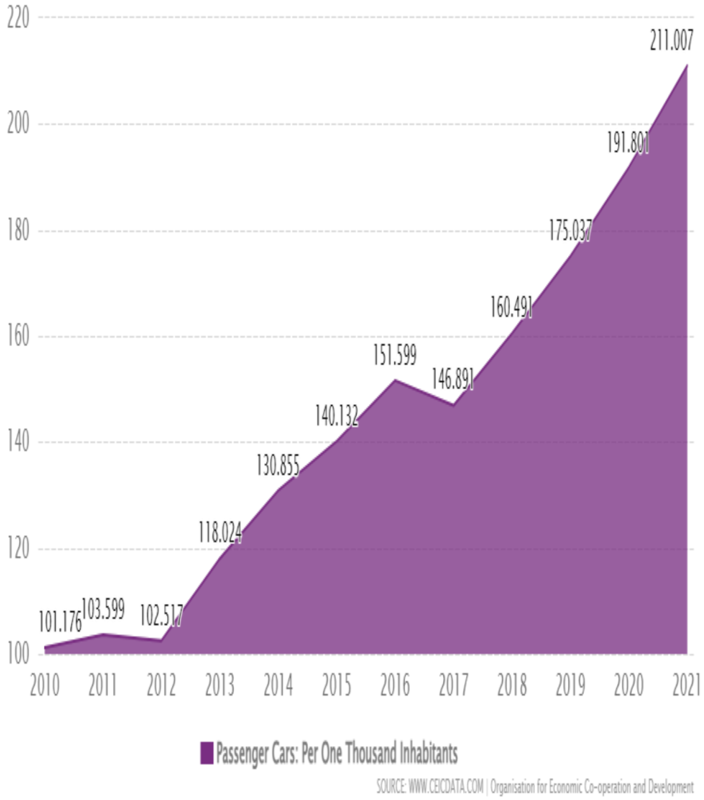
For most of Tirana and Durres residents, the highway "Tirana-Durres" is the only viable way of going to work, school and medical and commercial centers. Unfortunately, the current state of highway does not adequately meet the needs of population movements in terms of speed, quality and comfort. Rush hour is not the main problem, but rather an alternate way to our basic mobility obstacle, which is that most people must move at the same time every day. Why? Because all our operations in school systems and economy need that people work, travel to school, and even fulfill some small business about the same time so they can have the opportunity to collaborate with each other. That simple demand cannot be fulfilled without disturbing our society

In Albania, the greater parts of citizens intend to move during rush hours using private vehicles. The main reason for this is that the public transport in Albania has many problems and it is not comfortable or faster than the private vehicles. Also, another reason can be that private vehicles can help people to do multiple tasks. According to statistics, even though household incomes are decreasing, more and more people try to buy their own cars and trucks.

Several, but few studies exist in the literature for short term vehicles traffic load forecasting. [1]-[3] However, in this paper we attempt to show that longer term forecasts are possible too, using ARIMA and SARIMA modelling. We have to say that such a study is the first time is published, to the best of our knowledge, especially concerning Southern European regions.

## 1.2 Problem statement

The balance of transportation system can also be influenced by city's population or economic activity. If the graphic of the population in city is in growth, as in Tirana, this will soon mean by more cars generated by the newcomer population. This result is more obvious because Albania's vehicle has been increasing faster day by day. Open Data Albania has conducted a survey on the number of road vehicles in Albania. (Website 1) The data are based on INSTAT statistics, the Ministry of Public Works and Transport, and the World Bank. They are classified as vehicles: automobiles, minibuses and. Meanwhile, the classification of road vehicles, besides vehicles, includes road tractors, motorcycles and trailers.



SOURCE: WWW.CEICDATA.COM | Organisation for Economic Co-operation and Development

**Figure 1-1 Vehicles for 1000 people (retrieved from [https://www.ceicdata.com/en/albania/motor-vehicles-statistics-non-oecd-member-annual/passenger-cars-per-one-thousand-inhabitants](https://www.ceicdata.com/en/albania/motor-vehicles-statistics-non-oecd-member-annual/passenger-cars-per-one-thousand-inhabitants) )**

Based on the statistics, in Albania there were 121 vehicles for 1,000 inhabitants for year in 2010 and they were doubled during 2021. Saying in another way, it turns out there are more than 294,729 cars, 7,032 microbuses and 83,413 trucks. The total of road transport vehicles (including road tractors, motorcycles and trailers) is more than 500000 vehicles. Compared with 2000, the total number of road vehicles has increased by 126%. Historically, the number of vehicles per 1,000 inhabitants has been steadily increasing.

Changes in the economy also affect regional congestion. At the late 1990s, the graphic of traffic in the San Francisco increased extremely because of the internet and the telecommunications boom. Anyways, after the American economic “bubble”, it was noticed a decrease in congestion even though there was no major change in population. So, the congestion can also be an economic downturn.

## 1.3 Cope with the mobility challenge

There are some ways that Albania can try to avoid the rush hours in highway, but some of them are not so applicable for Albania government and some others have no great interest for Albanian’s citizens.

Charging tolls to enter the highway. Governments can oblige citizens to pay money to enter on highway. If tolls have the right price that can be affordable for the people, the number of vehicles on highway during rush hour could decrease enough so that vehicles could not wait in long line. That would allow people to travel and to run their errands faster and to spend less on fuel.

Another reason to charge tolls is the maintenance of the road. The highway “Tirane-Durres” is not in a great condition, and this can cause accidents and incidents, which many specialists believe is one of the main reasons for traffic congestion. Year after year, the graphic of number of accidents is always in growth and the accidents could have caused only more traffic and longer line of cars.

The government has proposed to charge tolls, but most people disagree for two reasons. Tolls would only help the wealthier citizens and considering the economy of Albania, most people would reject this solution, partly because they will think that this is idea is more a disadvantage than an advantage. The second reason is that Albanians will believe that these tolls are just like another tax, because they have already paid for this road once with their taxes. For both these reasons, this is not a solution that will be in a great interest for Albanians citizens.

Expanding road capacity from two lanes to three lanes. The second solution is to build another lane and to expand road capacity, so all drivers can drive faster and don’t have to wait in traffic. But this solution is impossible and too expensive. Albania’s government would have to add the new lane by cutting down trees and by destroying many businesses and houses and then by compensating. Except that you have to take into consideration that it will take a long time to widen the road, and the highway will be closed, and it will cause more traffic. Expanding road capacity is a great idea, but the government cannot afford such a high price.

Improve public transport. There are many ways that public transport can be made better. First of all, it is needed new buses and minibuses, because the actual ones are too old, and they don’t have air conditioning. Secondly the tickets for buses need to be cheaper, so that the citizens can have more interest in travelling by bus than in their own vehicles. Another improvement would be to make the interchange between Tirane-Durres easier for public transport. For the moment, the bus stations for these cities are in the suburbs of the city and it will be easier for the citizens if the bus stations are in the center. Also giving priority to public transport at traffic signals is an excellent solution to arrive at destination faster. Anyways, even though all these changes are provided, it will be difficult for the government to make all these changes.

The last solution for the Albania’s traffic congestion to remain in control is to provide a decision support system. So, to choose the optimal control option, a fine view of the ongoing and future situation of the traffic needs to be provided to the traffic operator. For the moment, the evaluation and forecast of the traffic situation is principally built on observation and tactic judgment: traffic operators completely create their point of view of the traffic situation at moment and in the future based on their background with the job. The use of an application that can forecast better the traffic situation would be a great help to traffic control.

### 1.4 Methodology of the study

A time-series ARIMA model was used to evaluate the growth and decrease of traffic congestion of Albania through time-series data. This model can help to forecast how the traffic is going to continue on the Tirana-Durres highway. The study will be based on the data that are generated by Inductive loop detectors that can detect each vehicle that passes over them. It measures vehicle speed, length etc. Python and libraries like matplotlib, pandas were used for the analysis. Also, in this study is used to see the benefits of auto- Arima over building the Arima itself and to evaluate which model is better for our time series ARIMA or SARIMA.

### 1.5 Objectives of the Study

The first objective is that based on the current data το develop a prediction model that is beneficial for the traffic management in the Tirana. Later, this model can be seen as guideline for a forecast tool. The main objective of the course is to develop modelling methodology and tools to generalize these outcomes to the majority of cities in the world too.

The second objective is to further examine the best performance model for prediction tool.

### 1.6 Justification of the Study

The study would serve how to cope with the mobility challenge in Tirana. The road system cannot handle rush hours loads without constraining citizens to wait traffic and while the population in Tirana is growing, the traffic congestion will get worse and worse. Considering the negative effects of traffic like delays, just in case time, fuel consumption and pollution, finding a fix to rush hour could result in a great improvement in the quality of life in Tirana.

### 1.7 Organization of the study

This paper is split into five sections. In the actual chapter, it is presented the background of the study and the problem statement as well as objectives and justification of studies. The last thing discussed is how the paper is organized.

Section two discusses the literature review showing the work done by the other authors in this area. After that is a summary of literature review.

Section three is about the methodology of the study. Firstly, it discusses the Arima model and the basic terms. Then it shows the way that we are going to follow to build the model Arima and Sarima.

Section four is the most important one. It presents the data, and it shows analysis of time series. After that it shows how the models are built, finding the best one and the forecast of the traffic.

The last section is about the summary and the conclusion. In the summary is explained about the whole results and in the conclusion, it discusses if the objectives are fulfilled.

## 2 Literature Review

### 2.1 Literature review of related research

Firstly, Sukruti Vaghasia at Windsor, Ontario, Canada published in 2018 [4] aimed to develop an intelligent transport system to be adequate to outcome a realistic outcome for prediction of traffic line in both the short and long terms. However, primarily the timed series was transformed to time stationary because of the integrally data. For this reason, fuzzy rules were helpful to split the dataset based of factors such as weather condition, season of a year and rush hours. In conclusion he discussed that he achieved more accurate forecast using Arima instead of the neural network forecast. Anyways he used the fuzzed rules with binary values to acquit the presence of features instead of getting the real values. For example, the effect of rain is captured only if it rained or not instead of getting in calculation the amount of rain. He achieved very good results with Arima with 83% short term forecast accuracy for one day. For the long term he compared hybrid Arima and Arima had result around 98%, while Arima result to 82% without special features of information. Finally, he mentioned that his dataset was only for one year and more data is needed to test his algorithms.

Secondly, C. Narendra Babu and B. Eswara Reddy from JNT University College of Engineering, Anantapuramu, India published in 2015 [5] tried different available model of Arima to predict for both one-step ahead and multi-step head. Their dataset was on Traffic Internet Data, and they organized their data to 30 and 60 minutes. Their results were based using the using the mean absolute error and mean square error measurement. All the models that were studies were ARIMA-ANN. The ANN means that they were capable of building models with data that are nonlinear. As a conclusion, the MA filter-based hybrid ARIMA-ANN was the best model compared to the other models, in terms of both MAE and MSE.

Also, Priyanka Rani at Technical University Kapurthala- Jalandhar Highway mentioned [6] the problem that the road infrastructure can't handle any more with the rapid increase of traffic congestion. He emphasized that after his literature review other datasets are used offline, while his work was based on sensor data. So, he suggested a new way of forecasting by combining K Nearest Neighbor techniques and Euclidean Distance. The new model has been tested to forecast traffic flow. He evaluated the traffic flow and took into consideration only appropriate data from the dataset. After that he analyzed the effectiveness and performance of forecast techniques by using K Nearest Neighbor and ED to pay attention to new points in the traffic flow. He suggested a K Nearest Neighbor and Euclidean Distance for traffic prediction method. Another contribution to mention in his paper is the effectiveness of data mining techniques that can help to deal with real concerns. In the last 10 years, analyzation and prediction of system activity has resulted in main investigation in different fields of PC systems. The suggested model used hybrid way of K Nearest Neighbor and ED to accomplish forecast accuracy, prediction accuracy that can have better a result by reduction in mean absolute error and mean square error measurement. In conclusion he said that he had better result as compared to existing techniques without KEARIMA distance.

R. M. Reza et al. from the University of North Carolina at Charlotte focused on his paper [7] on an algorithm of (ARIMA) model based on traffic flow data from neighboring links in predicting short-term travel time through a way due to an accident. His datasets were built from 181 “Vehicle Accident” type incidents that happened along ~19-mile freeway segment in the city of Charlotte, North Carolina. The spatial-temporal context-based data was organized with the ARIMA model using pre-whitening of the cross-correlation function alongside lagged regression. His outcome showed that the movement times for successive roads are highly linked and, both segments of a target way on a corridor will influence the current traffic flow. The mean absolute percent error (MAPE) and mean absolute deviation (MAD) of the developed model for every road that he proved resulted in less than 10%. So, the outcome achieved from the validation of models at both segment- and corridor-level showed that the estimated effect of incidents on travel time is almost equal to and close to the real-time observations. MAPE and MAD computed from validation data (26 55 incidents) are observed to be less than 10% for 95% of samples indicating an accurate estimation of the effect of incidents on travel time using the proposed method.

Billy M. Williams tried [8] to predict the vehicular traffic flow with the Sarima model. His dataset is from two away location, one in the United Kingdom and another at United States. Another objective of his paper was to compare fitter Arima model with three heuristic predicting method that are the random walk forecast, the historical average forecast and a deviation from the historical average forecast. His Sarima model was used to the series of running 15 min averages and for this reason his Sarima prediction was calculated using only terms evaluated only on the original 15 minutes interval time series. In his discussion of finding, he discusses that Sarima had way better results than all others heuristic forecast methods. Also, he emphasized that seasonal Arima disapproves theoretical motivation to investigate high level nonlinear mapping approaches, such as neural networks. Finally, he recommends that seasonal Arima can be a good starting point to predict the traffic forecast.

Another paper that studied short term traffic prediction is from Guoqiang Yu and Changshui Zhang from Tsinghua University [9]. They suggested that ordinary changes model is not the right way to check the pattern changes in the traffic flow and so they need to add a sigmoid function to have better results. The data are obtained from Traffic Management of Beijing. Their data set contains twenty-six days, from which twenty were used for the training set and the others one for test set. Their intention was to compare their model with three other models that are Random Walk, Historical Average and Informed Historical Average. The outcome was based on the mean root square error (MRSE) and the mean absolute relative error (MARE). The outcome showed that their model Arima outperformed all the other models with at least 20 percent. In conclusion they discussed that for each different pattern they used a different model to describe each of them, because they had different characteristics. Anyways they mentioned that there is room for more future work. Firstly, they set the parameters manually and in practice this will take a huge amount of time. It is needed to find a way to set the parameters automatically. Then he asked what are going to be the results if they used another model. Finally, he said that the data they used was only from one site, and more data is needed to check the performance of the model.

### 2.2 Summary of literature review, the research gap and contribution of this paper

Based on the literature review, we noticed that Arima is one of the best models used for prediction. Many authors compared different models like Random Walk, Historical Average Informed Historical Average etc., but Arima have the best results for traffic flow prediction. They discussed the fact that different approaches of this model have been used and the outcomes are satisfying. The problem is to find the right parameters for your data. Also, many authors mentioned the fact that the data was not sufficient. All previous research for traffic loads forecasting was focused on short-term forecasting, due, also, to data insufficiency. In the herein research we extend previous studies in an effort to model longer terms forecasting of traffic loads, carefully optimizing ARIMA and SARIMA models.

## 3 Methodology

### 3.1 Time Series Prediction

A time series is a sequence that is calculated through the same amount of time. The time series can be of different types. It can begin with seconds (data collected by sensors), two minutes (calls to a call center), daily (stock market), weekly (sales to a store), monthly (business balance), annual (annual budget) etc.

Of course, the main step in time series prediction is to forecast how time series is going to continue in future. As you can think yourself, forecasting can be very valuable. Forecasting techniques can be applicable in every field of life. It may be important from the simplest things like if there is traffic on the road so you can avoid it to the most important ones like business planning. Normally, in any forecast, there may be a small percentage of error, but it should be minimized as much as it can because it may have a very serious cost for the business.

There are two types of time series prediction: Univariate Time Series Forecasting and Multi Variate Time Series Forecasting. The first one uses only the previous values of the time series, while the second one, different from the other, uses predictors other than time series to predict the future time series.

This paper is focused more on Arima modeling and its various types. Arima or differently Autoregressive Integrated Moving Average Univariate Time Series Forecasting and so it can just use the information of the previous value to predict the next one [12-15].

### 3.2 Basic Definitions of Time Series

1)Time series is a sequence of monitoring data over a specific time zone. There are two types of times series. The first one is time series discrete that can be monitored over a defined time zone daily, weekly, monthly etc. The second type is the one that is observed in every instance… for example like traffic in highway. It can be used in many different fields of work like weather forecasting, signal processing, astronomy etc.

2) Time series analysis is the method to analyze time series to obtain meaningful stats and to find out the components of time series (seasonality, cyclical, trend or noise) and to analyze why components happens. Basically, it tries to understand what happens with time series over a period so it can predict future values based on the real ones. There are many time series models that can make prediction of data like FIGARCH, CGARCH, but we going to focus more on Arima model by Box-Jenkins and Tao.

3) Time series graph is simple observation of data that is displayed in the graph. It consists of x-scale that represents time scale, and y-scale that represents data scale. The graph makes it easier for us to detect trends, seasonality etc.

#### 3.2.1 Time Series Components

There are four main components that are critical for time series. These four components are: Seasonality, cyclical. Trend, and Noise or Irregularity. Each of them can be important in every change of time series.

The trend can be the main reason for the long-term behavior of the time series. It can be caused by many factors like socio-economic, politic, price-inflation etc. These changes can affect drastically in the growth or the decline of the graph. For example, in our case, people don't need to pay money to use the highway, but if it happens the opposite, they are going to find an alternative road for their purpose. So, this type of tendency is going to continue for a long time.

Seasonal variation is short term movement that can affect time series for a short period of time, but that it happens almost every time in specific period of the year. For example, it is observed that after the school season, there is a decline in the demand for rental houses. Also, it happens the opposite when it is the peak holiday season in August. Anyways they are going to take into consideration if the data is only for a short period like monthly or weekly, and not for a year.

Cyclical variations are an irregular rising and falling in number or amount that exist in data for a specific reason that are not for a fixed period. For example, in my case this could happen if there was the need to fix road faults for a period. This could cause a decline in the number of cars at the end of day because there was going to be traffic, and people were going to avoid passing on that road.

Irregular Fluctuations are sudden changes that can happen that are not included in the trend, or in seasonal variation or in cyclical variations and have no explanations. They are going to have a drastic change in time series, but they are not going to repeat over time, and it may happen only once. For example, they can be accidental in nature like fire, earthquakes etc. During my observation for one year, nothing like this happened.

## 3.3 Data Preprocessing

Before we transform data into stationary, we first need to transform it into under stable data. This is what it is called data preprocessing. Our data sometimes can be incomplete, inconsistent or lacking some behaviors that can produce errors in our results. A simple example can be just casting data from string to integer to calculate the sum or the format of data or data-time.

## 3.4 Stationarity in time series

Stationary in the time series means that the properties of time don't change over time. This doesn't mean that there is going to be only a linear line on the graph of the time series, but the way of growth and the decline of the graph remain constant. But why is important stationary in time series? It is important because it is going to be easier to analyze time series and we can find a value that captures the rate of these growth and decline. When it is time stationary, we have more odds to achieve prediction, because these changes are predictable. Also, another great reason is that we can use mathematical formulas for analysis and prediction.

### 3.4.1 Tests to check if time series is stationary or not

Anyways, the first thing to do is to find out if the time series is stationary or not. There are two ways that we can prove this.

1) Rolling Statistics, that is a visualize technique. we can find out with my eyes if the plot moving avg varies with time. But why do we need a visualization technique when we can have better a statistic one? Because we can find out immediately about the trend that have time series, that will be hard to detect. Standard deviation has drastically changed when they detect a trend or there is a surprise growth or decline of the graph in time series. When the data is stationary, even the graph of rolling mean and rolling std remains normal.

2) Augmented Dickey–Fuller test precise technique and it is based on statistics. Differently it is called unit root test. There are many unit root tests that can give us different results to find out if our time series is stationary, but Augmented Dickey- Fuller is the most famous one. It can show us how much the trend has affected our time series. This unit root test is based on the null hypothesis. If the null hypothesis(H0) is accepted, it means that there is a unit root and data is not stationary, otherwise if it is rejected, it shows that time series is stationary. The results include critical values, test statistics and the p-value. There are two conditions to say that our data is stationary. The first one is that the critical values have to be higher than the test statistic. The other condition is that p-value has to have lower values as possible based on the null hypothesis. If the p-value is greater than 0.05, it means that it has a unit root and it is not stationary, otherwise if it is lower than 0.05, it rejects H0 and it doesn't have unit root. If these 2 conditions are not fulfilled, we have to turn the time series into stationary so we can proceed further in our work.

### 3.4.2 Transform time series to stationary data

But what to do if the time series is not stationary? We are going to transform them into stationery. First of all, the main reason that data is not stationary is because of the trend that is mentioned before. To avoid trend, we can fit some type of curve to the data and then model the residuals from that fit. There are many algorithms that can help to achieve stationary, but we are going to use three of the most famous ones:

a) Log Scale Transformation is more used to deal with skewed data. Normally data in research is so skewed that it can't be made any prediction. So, it needs some transformation. The log scale is the most famous one. Some of the most used logarithm x based on 10 log, x based on 2 log, (lne) etc. The overall idea is to make exponential decline or growth ($y = a \exp(bx)$) in a linear one $\ln y = \ln a + bx$. Anyways it can be used with 0 values or negative ones.

b) Exponential Decay Transformation- A quantity is subject to exponential decay if it decreases at a rate proportional to its current value. Symbolically, this process can be expressed by the following differential equation, where N is the quantity and λ (lambda) is a positive rate called the exponential decay constant: $dN/dt=-\lambda N$

The solution to this equation (see derivation below) is: $N(t)=N0*e-\lambda t$

where $N(t)$ is the quantity at time t, and $N0 = N(0)$ is the initial quantity, i.e. the quantity at time $t = 0$.

c) Time Shift Transformation–- there are many techniques to use with Exponential Decay Transformation, but we are going to use the simple one. So, let's say if my time series is something like y0, y1, y2, y3,y4,y5…..yn, the shifted transformation is going to be something like this null, y0, y1,y2,y3,y4,y5….yn, so every index is going to be shift one by the right. So, the time shift transformation is going to be ynull,(y1-y0),(y2-y1),(y3-y2),(y4-y3),(y5-y4)….(yn-y1)

These are going to be very easy to implement with the help of python.

## 3.5 Arima models

Arima or otherwise Auto-Regressive Integrated Moving Average is a model that is used for prediction of time series that are stationary based on the previous values. The reason that has to be stationary is that Arima is a linear regression model that uses its own period of time from one event to another to predictors. It's the combination of two other Model that are AR, that stands for Auto-Regressive and Ma that's stand for Moving Average. These models have the best results when the predictors have no connection with each other and work independently. There are two types of Arima: Non-Seasonal Arima and Seasonal Arima. Firstly, let's focus on the Non-Seasonal Arima. The Arima model has 3 important parameters:

a) p – defines the order of the Auto Regressive Model

b) q – defines the order of Moving Average Model

c) d – parameters that it is needed to make the time stationary (it remains 0 when the time series is stationary or otherwise the smallest value possible to make stationary)

Before finding out how to find the best values for our 3 parameters, first we have to understand what Auto Arima and Moving Average are. Justan. Auto Arima model means that Yt has no connection with no other and works independently in specific period. Let's see the main function of the Yt:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} +..+ \beta_p Y_{t-p} + \epsilon_1$$

Where t is the period of the series, beta is the parameter of lags and alpha is the intercept coefficient, that it is estimated by Auto Arima.

Like Auto Arima, even Moving Average Model has its own dependency, and it only depends on its own lagged prediction error. The function of the MA model is:

$$Y_t = \alpha + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} +..+ \phi_q \epsilon_{t-q}$$

Where the error parameters are the error of AR model on its own period. The error of Et and E(t-1) correspond for these equations:

$$Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} +..+ \beta_0 Y_0 + \epsilon_t$$

$$Y_{t-1} = \beta_1 Y_{t-2} + \beta_2 Y_{t-3} +..+ \beta_0 Y_0 + \epsilon_{t-1}$$

There were Auto Arima model and Moving Average model respectively. Anyways we want the combination of both these models to produce the Arima Model and its own equation. The equation of Arima is:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}$$

In other words, one could say that the Arima models are:

**Predicted Yt = Constant + Linear combination Lags of Y (upto p lags) + Linear Combination of Lagged forecast errors (upto q lags)**

Finally, now we can find out what values can be the parameters q, p and d. Let's start defining them

### 3.5.1 Defining parameter *d* in the Arima model

As we mentioned before, the purpose of d is to make time series stationery. But we need to be sure to avoid over-difference in our data. The reason for this is even though my time series can be stationary, it can affect other parameters of the model, and it isn't going to give the right result. So, how are we going to find the right value for the d? The right value for the order differencing is the lowest value of differencing to achieve near-stationary series which roams nearly defined mean, and the autocorrelation function (ACF) gets 0 as quickest as possible.

Generally, the time series needs either further differencing to achieve stationary or it is the opposite case, and the time series is over-differenced. The further differencing is needed when correlation between the elements of a series and others from the same series separated from them by a given interval are greater than zero for many numbers of lags. Differently, if the lag of autocorrelation is less than zero, it means that time-series is over-different. If there is doubt between two values of d, then choose the one that results in the least standard deviation. Anyways, we don't have to forget that firstly we have to check if the data are stationary, because if they are stationary, we don't the order of differencing and the d is going to be 0.

### 3.5.2 Defining parameter Ar term p in the Arima

To find out the required value of the p, we are going to use the help of the Partial Correlation (PACF). But firstly, we are going to explain PACF. Partial Correlation means the connection between the series and its own period, removing other lags that can intervene. The equation of PACF is:

$$Yt = \alpha0 + \alpha1 Y{t-1} + \alpha2 Y{t-2} + \alpha3 Y{t-3}$$

So, if Y t is the current series and Y t-1 is the lag 1 of Y, then the partial autocorrelation of lag 3 (Y_t-3) is the coefficient $\alpha_3$ of Y_t-3 in the above equation. Now how is going to help this to find the Ar term? Any relationship between the elements of a series in each interval in series that are stationary can be put right by adding enough Auto Regressive values. So, to begin with we take the value of AR parameter to be equal as the numbers that pass the significance limit in the PACF graph.

### 3.5.3 Defining parameter MA term in the Arim model

As we said before, the Moving Average Model has its own dependency, and it only depends on its own lagged prediction error. And like we did with the help of PACF to find the right parameters for the Auto Aggressive model, the same way ACF graph is going to help to find values of MA parameter. The PACF is going to show the number of Moving Average terms that are needed to remove relationship between the elements of a series in a given interval in series that are stationary.

### 3.6 Auto-Arima model

The auto.Arima() function in python uses the Hyndman-Khandakar algorithm, which tries to find out the best combination of minimal values of AIC, MLE and unit root tests. There are many arguments of the auto.arima() function that gives different behavior to the algorithm, but this is the default one:

1) The parameter d is between 0 and 2 and it is defined by repeated KPSS tests

2) Terms p and q are defined by finding minimal values of AIC after finding out parameter. Auto.arima() does not try out every combination of Auto Arima term p and Moving Average q but try a stepwise way to travel through the model space.

3) There are four default models: Arima(0,d,0), Arima(1,d,0), Arima(0,d,1) and Arima(2,d,2). If d is equal to two, a constant is defined, and it is included in the algorithm. If d is less than one, then another model is added: Arima(0,d,0) with no constants.

4) The most successful model (with the best minimization of the AIC) fitted in the previous step is defined as the current model.

5) From the current model are considered different variations: increase or decrease p and q of the current model by 1. Now the most successful model from these different variations or the current model is considered the new current model.

6) Repeat the third step until no minimization of AIC can be found.

### 3.7 Modeling Arima

The general procedure to build an Arima model are these:

a) Load the dataset

b) Preprocessing: removing all null data, creating timestamps etc

c) Stationary: In order that we can go further on with modeling, our time series need to be stationary. This includes defining time series components like (trend, seasonality etc.) and trying to avoid them.

d) Defining d: For making time series stationary, the d parameter is necessary. If the time series is stationary, d is equal to 0.

e) ACF and PACF graphs: Plot the ACF and PACF graphs to define the p and q value.

f) Choose the best module: Build the module AR, MA and ARIMA. Compare the Rss value and choose the one with lowest RSS.

g) Prediction: Predict the future values.

With the help of auto-arima, it takes cares from step c to e.

### 3.8 Arima vs Auto Arima

Why should we build the model of Arima ourselves when auto Arima can do much better? Yes, in many cases, auto Arima function can be much better and faster. Anyways, in practice, to sense check findings of an automated selection process against specified criteria and it can be modified if necessary. For example, it can be more productive if we can replace (2,1,0) with a better AIC value (1,1,0), if the second model makes intuitive sense and it is going to be better for the business.

### 3.9 Train, Validation and Test Set

But how we are going to test my results? We are going to split my dataset into 3main components that are train, validation and test. Each of them has its own importance in prediction. We are going to explain each of them. The train set is the sample of the dataset that is going to train the model. The model (in our case Arima) is going to learn and see from the test set that we are going to specify. The validation test is used to evaluate the model fit on training set while taking in the consideration the parameters that are defined before that the learning process start. Anyway, if the model has no parameters or few parameters, the validation test can well be ignored. The test set is a sample of data that is going to check if the prediction based on the test set is correct. So, test set is only used after the train set is created and after that it provides evaluation about the forecast.

The dataset split ration is the main concern when we are going to train and test the data. It is based on 2 things. First, how big is the dataset and the model that we are taking into consideration. Some models need considerable size of data to train, so in this case it is better to give as much as possible to the train data. Usually, people first split their data into the train set and test set. In most cases, the best practice is between 70% -80% for training data and 20-30% for the test data. After that, they choose a Y percent of their train data to be real train dataset and the remaining (100-Y)% is going to be the validate set. This is called Cross Validation, but there are other ways that you can choose the validation set. The benefit of Cross Validation is that tries to avoid overfitting of data, and the most popular method of the Cross Validation is called K-fold Cross Validation.

### 3.10 Arima Model Results

After defining parameters (p,q,d) and splitting time series in train, validation and test set, we can build Arima model. There are 2 ways that we can check our results. The first one is visual. we can see when the forecast graph can fall over the actual graph that is the test set. From the last point of the train set until the last point that the forecast graph and actual graph combine it can be shown the total days that the prediction is correct based on the x-axis that shows time. Also, in my help come the lower and the upper limit that specify the range of values in prediction.

The other one is Arima model Results. Here we can check what goes wrong with my model. It shows us a lot of information. Let's check one by one.

-*Dep.variable*- corresponds for the main column that we are going to predict. In my case it's going to be the number of vehicles that pass on the highway in one day.

- *Model* - corresponds for values of the parameters(p,q,d) that we choose for my Arima model.

- *Method* - can be one of these three: css-mle, mle, css. They're about the loglikelihood to maximize. Each of them has their own way to get the maximum conditional sum of squares likelihood and the results are as starting values for the calculation of the exact likelihood thanks Kalman filter.

- *Date&Time*– Date and time that Arima model is generating.

- *Sample*–the dataset that is used for training the data.

- *No. Observations & Log likehood*– number of days that we are going to take into consideration and how we are going to express them as a function of statistical parameters.

- *s.d innovations* - the standard deviation of the model's residuals/innovations.

- *Aic&Bic&Hqic* - are penalized-likelihood criteria. They are sometimes used for choosing the best predictor subsets in regression and often used for comparing non nested models, which ordinary statistical tests cannot do. The AIC or BIC for a model is usually written in the form [-2logL + kp], where L is the likelihood function, p is the number of parameters in the model, and k is 2 for AIC and log(n) for BIC.(Websites 6)

- const – it shows each parameter that takes part in the building of Arima model.

- Coeff – are the final values that takes our parameters. They are the numbers by which the results of the terms are calculated in Arima model.

- *std error* – or differently the SE Coeff defines the lack of consistency between parameters when you try to achieve the results taking the sample again and again. The std error must be as small as possible, so the prediction can be accurate.

- *P value* – the most important feature in the Arima Model Results. It shows which parameter is not right. The idea is based on the null hypothesis that we mentioned before. Each value of the parameter should be lower than 0.05, so it can be against the null hypothesis. If the parameter is greater than 0.05, it means that is not significant and you should remove it or change the value of it.

- *[0.025 0.975]* – measure the ratio between the Coeff. and the std error.

### 3.11 Residuals Plots in Arima

Residual plots it's going to help us to find the difference between forecast and the true values. There are four main components that are included in residual plots: standard residual, histogram plus estimated density, normal Q Q and the correlogram. Now let's see how what's mean each of them:

-Standard residual - the residual errors should rise and fall irregularly in the mean of zero. In case that this doesn't happen, and it fluctuates around another number, it means that the prediction is biased.

- Histogram plus estimated density – shows the distribution and to be a normal distribution, should have mean 0.

- Normal Q Q – the dots of prediction data should fall over the red line. In case there is any significant deviation, would suggest that the distribution is skewed.

- Correlogram – or differently the Acf plots will if the auto errors are auto correlated or not. If there exists any autocorrelation, it means that there is a residual error that is not defined in the model.

Once the results of the residuals plots seem to be a good fit, we can go further on with our forecast and show the final results.

### 3.12 SARIMA

Often it can be found seasonality in time series, that repeats every n times in observation. In order to find a solution, let's say there is an extension of the Arima model that is called Sarima (Seasonal Autoregressive

Integrated Moving Average Model). So, the difference between Arima and Sarima is the addition of seasonal error components. Like we said before, the first thing that Arima does is to verify if the time series is stationary and in case there is no stationary, it has to transform into a stationary series. This is very important to get the right values of the coefficients.

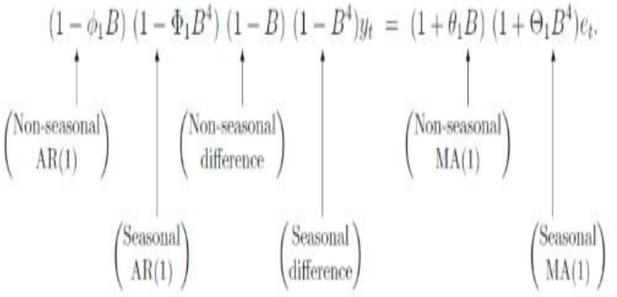
The same logic is used in the Sarima too. The purpose of Sarima is yet to make the time series to have a stationary behavior, so that is going to have a correct prediction. we have to mention that Arima model can work as good as Sarima in case these two conditions are fulfilled:

a) the dataset is enough large to find the correct number of parameters.

b) to take the risk of suggesting a complicated error structure.

Sarima model has 7 terms. The first three are the same of Arima model, that's the reason that we said before that it is called as an extension of Arima. The other four parameters are seasonal autoregressive component, the seasonal difference, the seasonal moving average component and the length of the season, defines the seasonality. Finally, the SARIMA model is like this (p,d,q)(P,D,Q)[S]. Since it is difficult to write the formula out directly, a backshift operator makes it easier to describe.

we are taking the case when the formula is (1,1,1)(1,1,1)[4]:



The B parameter is a helpful notational device when we have to work with time series lag. That is By(t)=y(t-1). Again, like we did with the Arima, we are going to use the ACF/PACF plot to define all of SARIMA parameters.

### 3.12.1 ACF/PACF plot for SARIMA

Like we said before, the order d is 0 if there isn't no trend or seasonality. If there exists any trend or the AIC values are very high, the d must be one or greater than one. If it is one, it means that there is a constant trend or otherwise it has a various trend. The p value is equal to the first lag that is across the significant level. The same logic is used and with q value. When the first lag across the significant level in the PACF plot, it is the value of q.

The above part was to identify the Arima parameters. Now we must find the seasonal parameters. S is equal with the highest lag in the ACF plot. D can be 0 or 1. It is one when exist a seasonal stable over time, otherwise if there is unstable pattern the d is 0. P is equal to one or greater than one, if in the ACF plot the lag S is negative. If it is positive, P = 0. The same happens with Q. If the lag S is negative in the ACF plot, Q is one or greater than one, else it is 0.

If you are going to do a manual analysis of the time series, it is going to be a long task especially if you have a large dataset. It is logical to use an automation selection that is called grid search. In case if we set the limit of the value of our parameters too high, it is going to take a long time that our grid search to complete the task. Also, it exists the risk of overfitting the training data.

So, to avoid the problem of overfitting the training data and the long task, it is applied the parsimony principle that the sum of all parameters of SARIMA model(p,q,d,P,Q,D) except S to be less than 6. Another way of solution is to set each parameter 2,1,0 using AIC with each combination.

We have to remember that grid search can't give always the best model. To achieve the best model, we must experiment with the values based on ACF/PACF plot.

### 3.12.2 Build SARIMA model

As we describe with Arima model, we are going to use Sarima model results to check if our parameters are the right one. The report is similar to Arima model description. The most important feature is going to be p<|Z|. The values of parameters should be less than 0.05 to not accept the null hypothesis. After my model has passed the test, we can create a visualization so we can compare my prediction to the actual results.

# 4 Data Presentation and Analysis

## 4.1 Dataset Details

The application “Transport” shows statistics and generates reports for traffic on national roads. This dataset contains real data for 17 national roads such as Tirane-Durres, Elbasan-Tirane etc. Each table is connected to one road. For exampleTirane-Durres.txt is related to both Tirane to Durres and Durres to Tirane. The separator is |. This text data has been used to generate 4 kinds of reports: daily, weekly, month and yearly. It shows statics like the length of vehicle, the maximal speed, minimal speed, has traffic or not etc. Each table contains the same data that Id, date, speed, type, point, length, direction, traffic. One file of one road contains 875639 road. For the whole dataset, the number of instances stands over 3000000 and are all real data.

I am going to focus only on one table “Tirane-Durres” for because it is the largest one (around 520 Mg) and it is the only one that has data for one year. All the others have a maximum of only 6 months, and we can’t have successful results based on this amount data.

**Table 4-2 Explanation of Attributes of table**

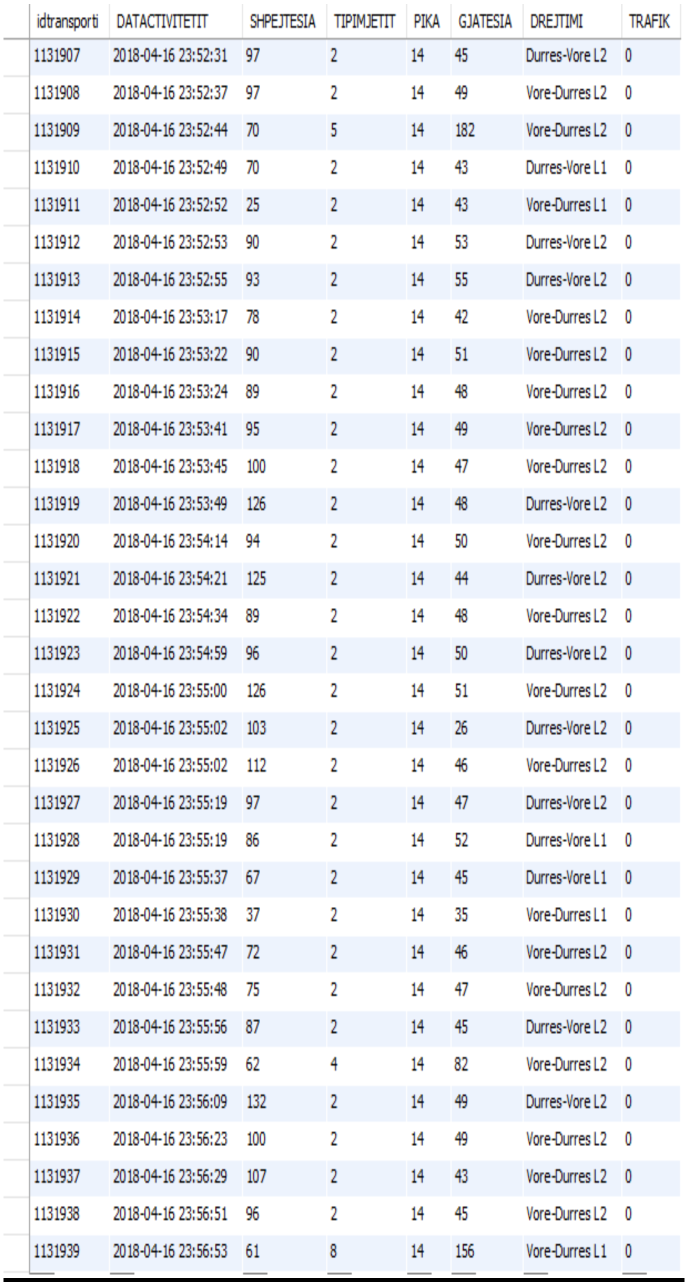
| Attribute | Explanation |
|---|---|
| Id | Primary key/Numerical data |
| Date | Day,month,year,<br>Hour,minutes,seconds |
| Speed | Integer/speed of vehicle |
| Type | Integer / car, motorcycle, bus etc |
| Point | Integer / which road it is |
| Length | Integer/ length of vehicle |
| Direction | String |
| Traffic | Integer(0 or 1)/ has traffic or not |

**Table 4-1 Tables of dataset Transport**

| Road | Rows | Columns |
|---|---|---|
| Tirane- Durres | 12159580 | Id ,date, speed ,type, point, length, direction, traffic |
| Tirane – Elbasan | 8875639 | Id ,date, speed ,type, point, length, direction, traffic |
| Vlore – Levan | 7855987 | Id ,date, speed ,type, point, length, direction, traffic |
| Burrel-Klos | 5502012 | Id ,date, speed ,type, point, length, direction, traffic |
| Elbasan –Librazhd | 6856980 | Id ,date, speed ,type, point, length, direction, traffic |
| Elbasan-Rrogozhin | 7689344 | Id ,date, speed ,type, point, length, direction, traffic |
| Vore-FshKruje | 8768034 | Id ,date, speed ,type, point, length, direction, traffic |
| FshKruje-Milot | 6554345 | Id ,date, speed ,type, point, length, direction, traffic |
| Levan-Tepelene | 7603239 | Id ,date, speed ,type, point, length, direction, traffic |
| Shkozet-Plepa | 8579748 | Id ,date, speed ,type, point, length, direction, traffic |
| Rrogozhine-Lushnje | 6885983 | Id ,date, speed ,type, point, length, direction, traffic |

**Table 4-3 Characteristic of tables**

| | | | | | |
|---|---|---|---|---|---|
| **Data Set Characteristics:** | Text | **Number of Columns:** | 8 | **Area:** | Vek S4 Detector |
| **Attribute Characteristics:** | Real | **Number of Attributes:** | 3000000 | **Date Donated** | 2018-04-19 |
| **Associated Tasks:** | Data Integration | **Missing Values?** | N/A | **Number of Cars Daily:** | 30000 |



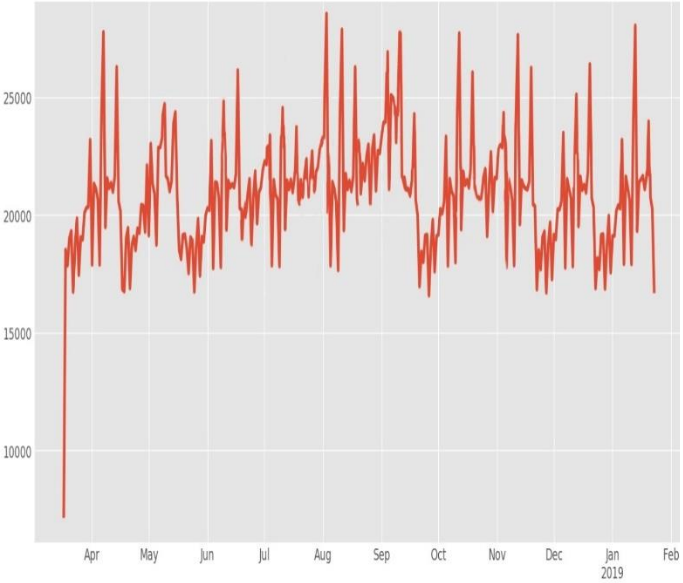
| idtransporti | DATACTIVITETIT | SHPEJTESIA | TIPIMJETIT | PIKA | GJATESIA | DREJTIMI | TRAFIK |
|---|---|---|---|---|---|---|---|
| 1131907 | 2018-04-16 23:52:31 | 97 | 2 | 14 | 45 | Durres-Vore L2 | 0 |
| 1131908 | 2018-04-16 23:52:37 | 97 | 2 | 14 | 49 | Vore-Durres L2 | 0 |
| 1131909 | 2018-04-16 23:52:44 | 70 | 5 | 14 | 182 | Vore-Durres L2 | 0 |
| 1131910 | 2018-04-16 23:52:49 | 70 | 2 | 14 | 43 | Durres-Vore L1 | 0 |
| 1131911 | 2018-04-16 23:52:52 | 25 | 2 | 14 | 43 | Vore-Durres L1 | 0 |
| 1131912 | 2018-04-16 23:52:53 | 90 | 2 | 14 | 53 | Durres-Vore L2 | 0 |
| 1131913 | 2018-04-16 23:52:55 | 93 | 2 | 14 | 55 | Durres-Vore L2 | 0 |
| 1131914 | 2018-04-16 23:53:17 | 78 | 2 | 14 | 42 | Vore-Durres L2 | 0 |
| 1131915 | 2018-04-16 23:53:22 | 90 | 2 | 14 | 51 | Vore-Durres L2 | 0 |
| 1131916 | 2018-04-16 23:53:24 | 89 | 2 | 14 | 48 | Vore-Durres L2 | 0 |
| 1131917 | 2018-04-16 23:53:41 | 95 | 2 | 14 | 49 | Vore-Durres L2 | 0 |
| 1131918 | 2018-04-16 23:53:45 | 100 | 2 | 14 | 47 | Vore-Durres L2 | 0 |
| 1131919 | 2018-04-16 23:53:49 | 126 | 2 | 14 | 48 | Durres-Vore L2 | 0 |
| 1131920 | 2018-04-16 23:54:14 | 94 | 2 | 14 | 50 | Vore-Durres L2 | 0 |
| 1131921 | 2018-04-16 23:54:21 | 125 | 2 | 14 | 44 | Durres-Vore L2 | 0 |
| 1131922 | 2018-04-16 23:54:34 | 89 | 2 | 14 | 48 | Vore-Durres L2 | 0 |
| 1131923 | 2018-04-16 23:54:59 | 96 | 2 | 14 | 50 | Durres-Vore L2 | 0 |
| 1131924 | 2018-04-16 23:55:00 | 126 | 2 | 14 | 51 | Vore-Durres L2 | 0 |
| 1131925 | 2018-04-16 23:55:02 | 103 | 2 | 14 | 26 | Durres-Vore L2 | 0 |
| 1131926 | 2018-04-16 23:55:02 | 112 | 2 | 14 | 46 | Vore-Durres L2 | 0 |
| 1131927 | 2018-04-16 23:55:19 | 97 | 2 | 14 | 47 | Durres-Vore L2 | 0 |
| 1131928 | 2018-04-16 23:55:19 | 86 | 2 | 14 | 52 | Durres-Vore L1 | 0 |
| 1131929 | 2018-04-16 23:55:37 | 67 | 2 | 14 | 45 | Durres-Vore L1 | 0 |
| 1131930 | 2018-04-16 23:55:38 | 37 | 2 | 14 | 35 | Vore-Durres L1 | 0 |
| 1131931 | 2018-04-16 23:55:47 | 72 | 2 | 14 | 46 | Vore-Durres L2 | 0 |
| 1131932 | 2018-04-16 23:55:48 | 75 | 2 | 14 | 47 | Vore-Durres L2 | 0 |
| 1131933 | 2018-04-16 23:55:56 | 87 | 2 | 14 | 45 | Durres-Vore L2 | 0 |
| 1131934 | 2018-04-16 23:55:59 | 62 | 4 | 14 | 82 | Vore-Durres L2 | 0 |
| 1131935 | 2018-04-16 23:56:09 | 132 | 2 | 14 | 49 | Durres-Vore L2 | 0 |
| 1131936 | 2018-04-16 23:56:23 | 100 | 2 | 14 | 49 | Vore-Durres L2 | 0 |
| 1131937 | 2018-04-16 23:56:29 | 107 | 2 | 14 | 43 | Vore-Durres L2 | 0 |
| 1131938 | 2018-04-16 23:56:51 | 96 | 2 | 14 | 45 | Vore-Durres L2 | 0 |
| 1131939 | 2018-04-16 23:56:53 | 61 | 8 | 14 | 156 | Vore-Durres L1 | 0 |

**Figure 4-1 Example of Tirana-Durres Table**

## 4.2 Data Preprocessing

So, like we mentioned above, we focused only on the table Tirana-Durres. For my time series prediction, we don't need data from both directions and for this reason we dropped all data that contains string "Tirana-Durres" in column Tirana-Durres. With the help of library pandas this was very easy. The second step is that we remove all data that has 1 in the column "Trafik". If there was traffic on that part of the road, it wasn't going to give us a valuable result to calculate the sum of total car passed in the end of day. After we searched in the whole dataset all the rows that has 1 in column "Trafik", it showed a total of 38 hours of traffic. We decided to replace them with random data from the dataset. The third step was to check if there was any data null or the system didn't work for more than 30 minutes, but there wasn't any missing data. After that we used the column traffic for different intentions. We changed from 0 to 1 for normal cars, from 0 to 3 for trucks and 0 for motorcycles. In this way it was easy for us to calculate the total sum of cars passed by during the day. Finally, we set the DataActivitet as primary index because all others calculation are going to be based on the DataActivitet.

**Figure 4-2 Plot of Tirana - Durres graphic**



We can notice that the peak of traffic is in August because it is holiday time. Also, we can notice that in July there is a decline in the graph because perhaps the schools are over and there is an increase again when schools begin.

## 4.3 Transforming time series to stationary

**Figure 4-3 Rolling Mean & Standard Deviation**

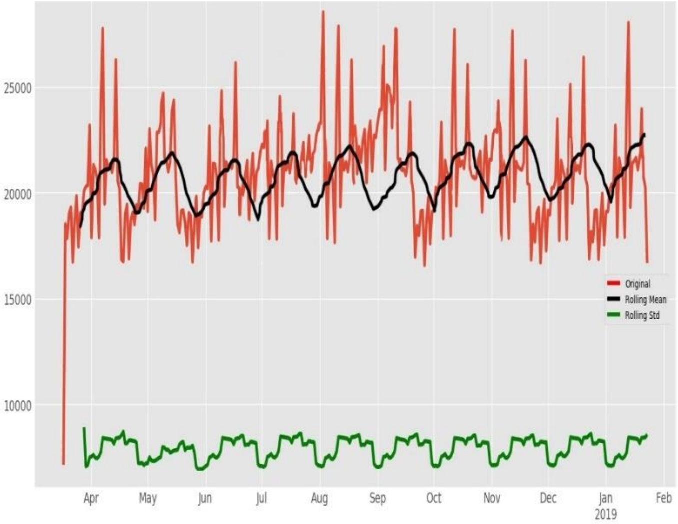
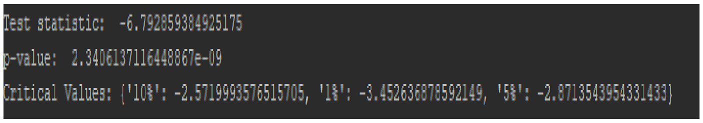


**Figure 4-4 Dickey Fuller Test**



```
Test statistic:  -6.792859384925175
p-value:  2.340613711644886‌7e-09
Critical Values: {'10%': -2.5719993576515705, '1%': -3.452636878592149, '5%': -2.8713543954331433}
```

After we did the rolling mean standard deviation, we can see that there is almost a stationary in my data time series. We can't be sure of this without doing dickey-fuller test. The results are that test static is lower than all of critical values. As is obvious from the picture above the test static is nearly – 6.7 and our critical values are: 10% = 2.57199, 5%: = 2.87134 and 1% = -3.4526. So, the first condition is fulfilled, but the p- value is higher than 0.5. The null hypothesis says that the p-value should be lower than 0.5 to reject the hypothesis and to determine that the time series is stationary. Anyways this result is good, and we are going to try to make it even better.

**Figure 4-5 Exponential Decay Transformation**

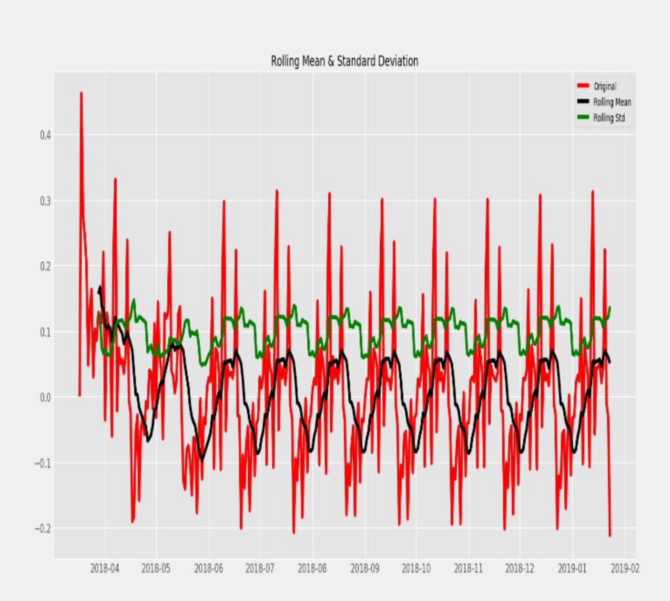
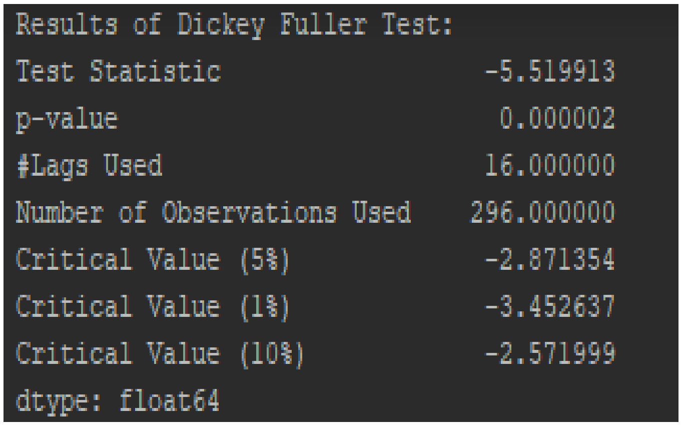


**Figure 4-6 Exponential Decay - Dickey Full Text**



```
Results of Dickey Fuller Test:
Test Statistic                 -5.519913
p-value                         0.000002
#Lags Used                     16.000000
Number of Observations Used   296.000000
Critical Value (5%)            -2.871354
Critical Value (1%)            -3.452637
Critical Value (10%)           -2.571999
dtype: float64
```

We can observe clearly that my data is stationary. we can say the same thing for the rolling mean and rolling Std that follow the same pattern as the original data. The p-value is lower than 0.05 and test static higher than the critical value. We can say that my time series is now stationary.

**Figure 4-7 Time Shift Transformation**

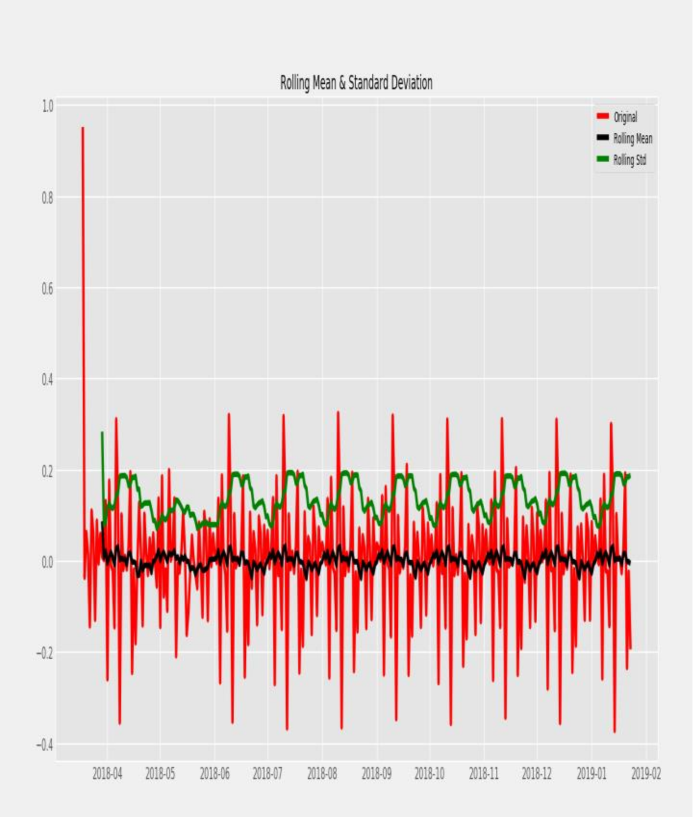
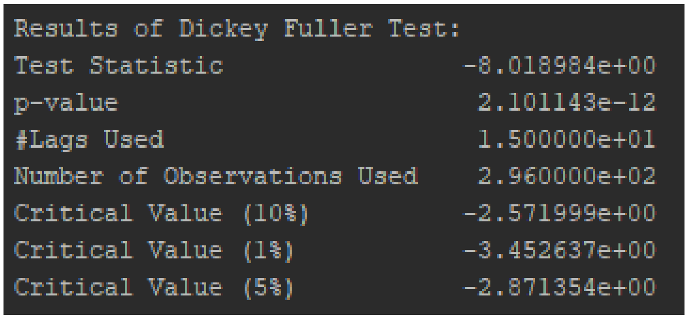


**Figure 4-8 Time Shift - Dickey Full Text**



```
Results of Dickey Fuller Test:
Test Statistic                -8.018984e+00
p-value                        2.101143e-12
#Lags Used                     1.500000e+01
Number of Observations Used    2.960000e+02
Critical Value (10%)          -2.571999e+00
Critical Value (1%)           -3.452637e+00
Critical Value (5%)           -2.871354e+00
```

We can clearly see that rolling mean and rolling std doesn't follow the same pattern as the original. Also, the p-value doesn't have any effect for good. So, the time shift transformation didn't have any good results.

**Figure 4-9 Log Scale Transformation**

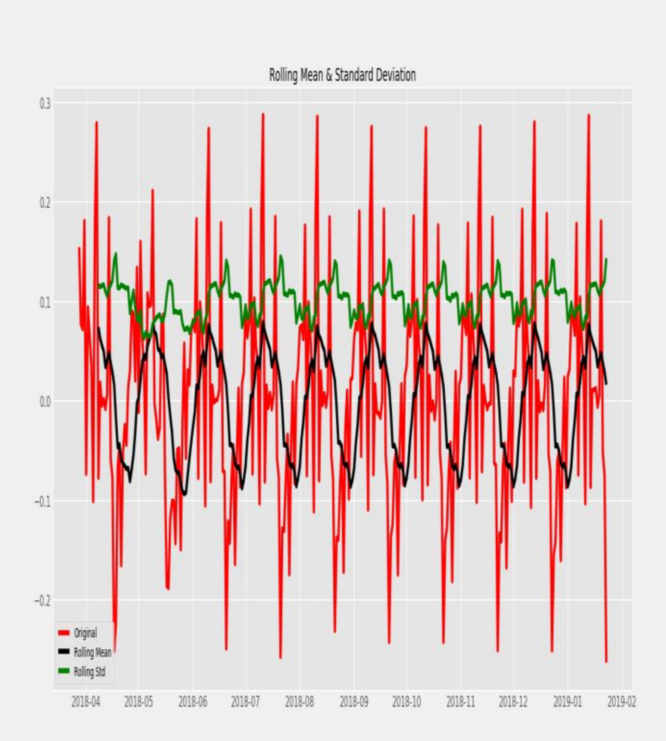
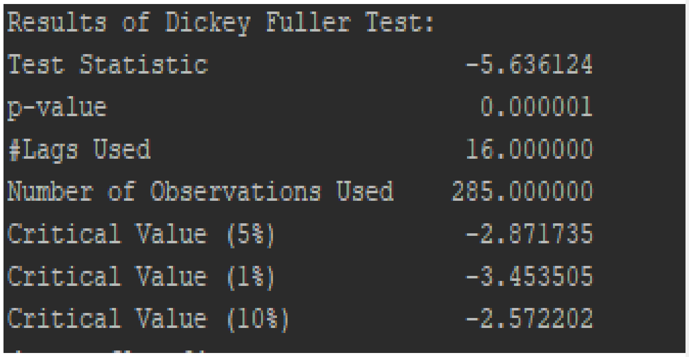


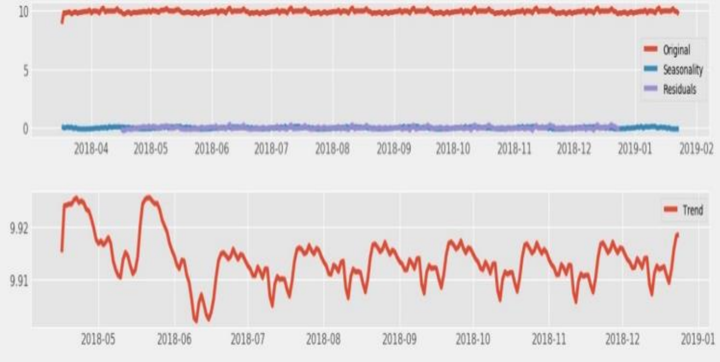
**Figure 4-10 Log Scale - Dickey Fuller Test**



```
Results of Dickey Fuller Test:
Test Statistic                  -5.636124
p-value                          0.000001
#Lags Used                      16.000000
Number of Observations Used    285.000000
Critical Value (5%)             -2.871735
Critical Value (1%)             -3.453505
Critical Value (10%)            -2.572202
```

We observe that the p- value decreased from 2.3 to 0.0001. This is very good because the p-value is smaller than 0.05 and the test statistic remains greater than critical value. It is almost the same with exponential decay transformation, but it is slightly better. To be sure that the log scale is the right one, we are going to define each component of time series (seasonality, cyclic variation and trend).

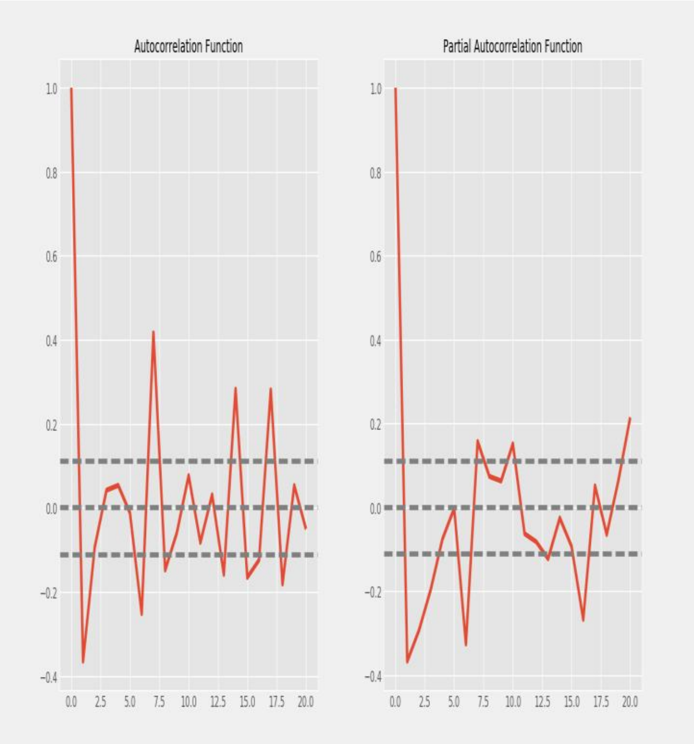
**Figure 4-11Components of time series**



We can observe that everything is normal. Trend remains almost in the same parameters between 9.91 and 9.92. So, we can continue further on with our log scale transformation.

## 4.4 Estimate parameters of Arima model

**Figure 4-12 Acf and Pacf plots**



As we said, the reason that we plot the ACF and the PACF is to find the parameters of Auto Regressive Model and Moving Average Model. To check what value is AR model we have to see the point where the graphic is approach more at y= 0. In my case it is between x=1 and x=2. We are going to test both cases to see the best result. The same thing we can say even with Partial Auto Correlation Function. It is between x=1 and x=2. So even AR model and Ma model have values 1 and 2. The Arima model is AR+I+MA. Anyways before building Arima model, we are going to see each model alone. Each model is going to give the value of RSS. The lower the value, the better the module.

**Figure 4-13 AR Module**

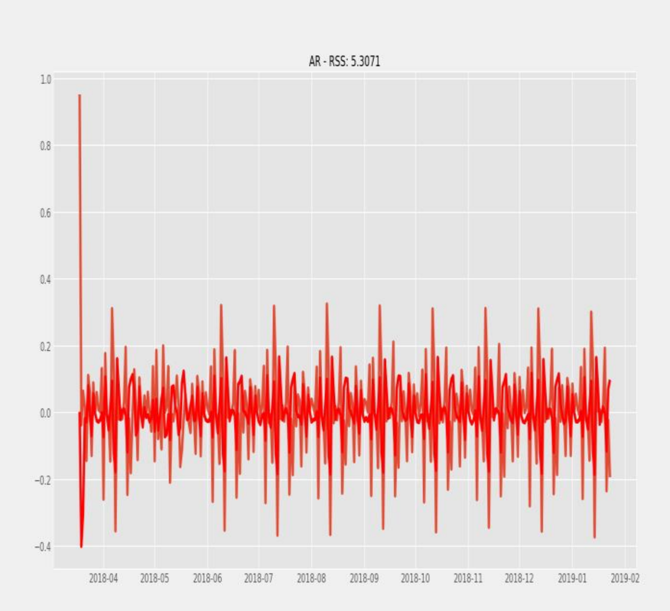


**Figure 4-14 MA module**

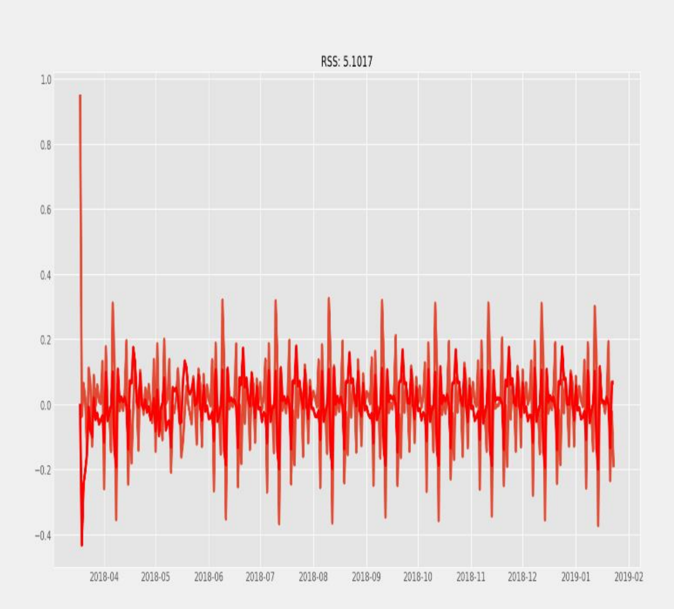
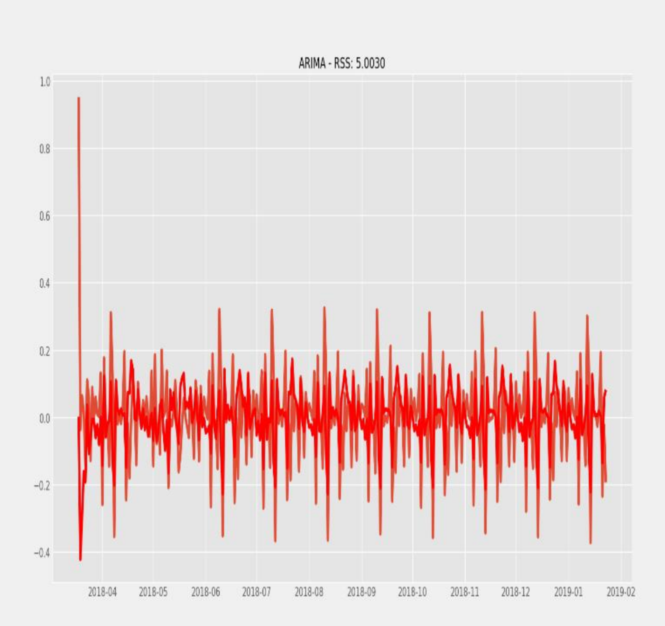


**Figure 4-15 Arima module**



The values of RSS for each model are AR = 5.3701, MA = 5.1017 and ARIMA =5.0030. we see that Rss value has decreased and the Arima module has the lowest value of them all, showing that Arima is the best module and the one that we need to go on.

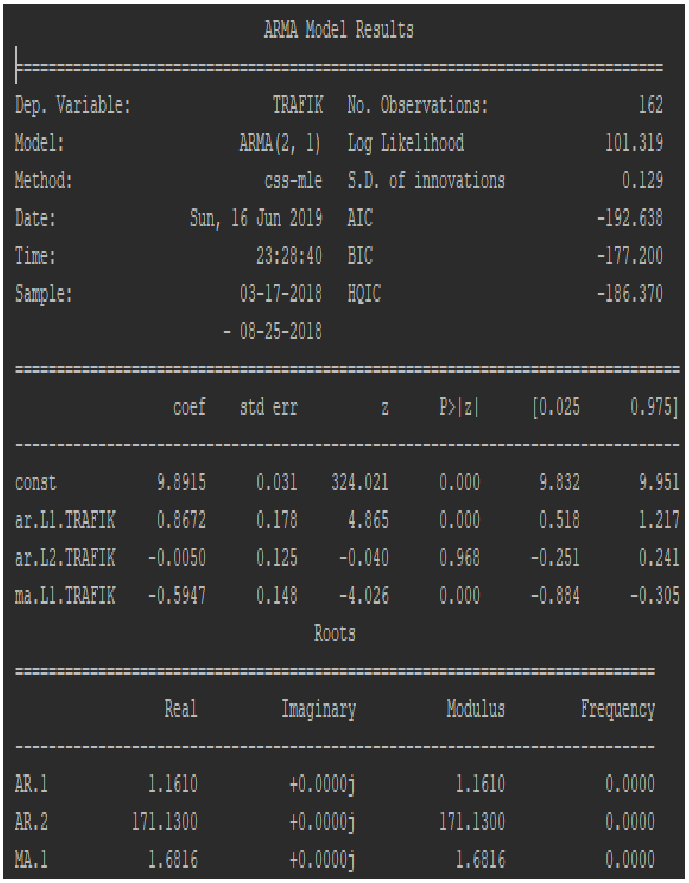
## 4.5 Training & Test Set

Generally training and test set is split between 70-30 or 80-20. Anyways we split my data set nearly 60% for the training set and 40% for the test set. The reason that we did this is because we thought that 2 months or 3 months is a short time to evaluate my prediction with the real data. Also, we let the alpha argument of the forecast() to the default that is 0.05 or 95% confidence interval. An alpha of 0.05 means there exists only 5% chance that real value can’t be in the range.

Firstly, we are going to build the Arima model based on my ACF and PACF plot. we are going to try many parameters and then Iwill compare my results with Auto Arima. Finally, we will try even with Sarima to achieve some results.

## 4.6 Building Arima models

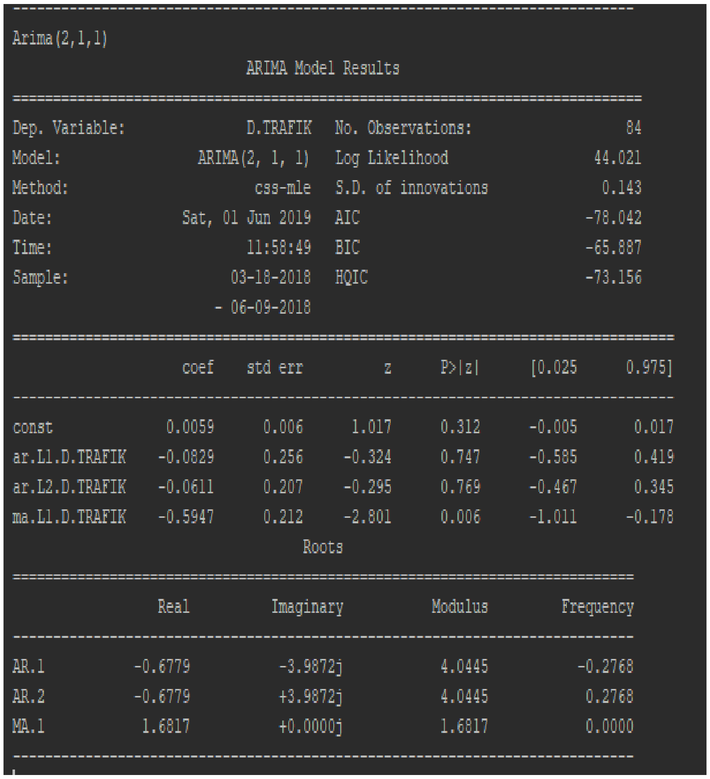
**Figure 4-16 Arma(2,0,1)**



```
                              ARMA Model Results
==============================================================================
Dep. Variable:                 TRAFIK   No. Observations:                  162
Model:                     ARMA(2, 1)   Log Likelihood                 101.319
Method:                       css-mle   S.D. of innovations              0.129
Date:                Sun, 16 Jun 2019   AIC                           -192.638
Time:                        23:28:40   BIC                           -177.200
Sample:                    03-17-2018   HQIC                          -186.370
                         - 08-25-2018
================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const            9.8915      0.031    324.021      0.000       9.832       9.951
ar.L1.TRAFIK     0.8672      0.178      4.865      0.000       0.518       1.217
ar.L2.TRAFIK    -0.0050      0.125     -0.040      0.968      -0.251       0.241
ma.L1.TRAFIK    -0.5947      0.148     -4.026      0.000      -0.884      -0.305
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.1610           +0.0000j            1.1610            0.0000
AR.2          171.1300           +0.0000j          171.1300            0.0000
MA.1            1.6816           +0.0000j            1.6816            0.0000
```

So, let's start with p=2 and q=1 from PACF plot. Like we said we turned the data to their original form. Firstly, we tried supposing that data are stationary, and we set d equal to 0. The values of 2 parameters are correct, but the overall coefficient is larger than 0.05 and it accepts the null hypothesis. Another thing that goes wrong is the value of AIC that is very large. This is not a good sign. Anyway, these results are not so bad, but we can make it better.

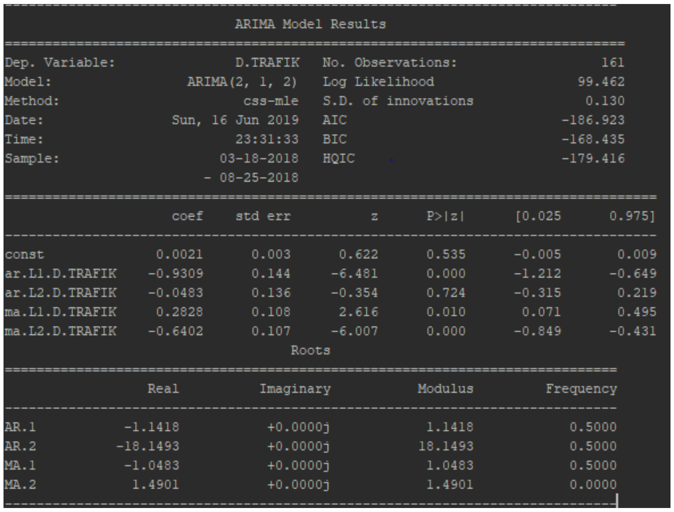
**Figure 4-17Arima(2,1,1)**



```
-----------------------------------------------------------------------------
Arima(2,1,1)
                             ARIMA Model Results
==============================================================================
Dep. Variable:               D.TRAFIK   No. Observations:                   84
Model:                 ARIMA(2, 1, 1)   Log Likelihood                  44.021
Method:                       css-mle   S.D. of innovations              0.143
Date:                Sat, 01 Jun 2019   AIC                            -78.042
Time:                        11:58:49   BIC                            -65.887
Sample:                    03-18-2018   HQIC                           -73.156
                         - 06-09-2018
==================================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------
const              0.0059      0.006      1.017      0.312      -0.005       0.017
ar.L1.D.TRAFIK    -0.0829      0.256     -0.324      0.747      -0.585       0.419
ar.L2.D.TRAFIK    -0.0611      0.207     -0.295      0.769      -0.467       0.345
ma.L1.D.TRAFIK    -0.5947      0.212     -2.801      0.006      -1.011      -0.178
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1           -0.6779           -3.9872j            4.0445           -0.2768
AR.2           -0.6779           +3.9872j            4.0445            0.2768
MA.1            1.6817           +0.0000j            1.6817            0.0000
-----------------------------------------------------------------------------
```

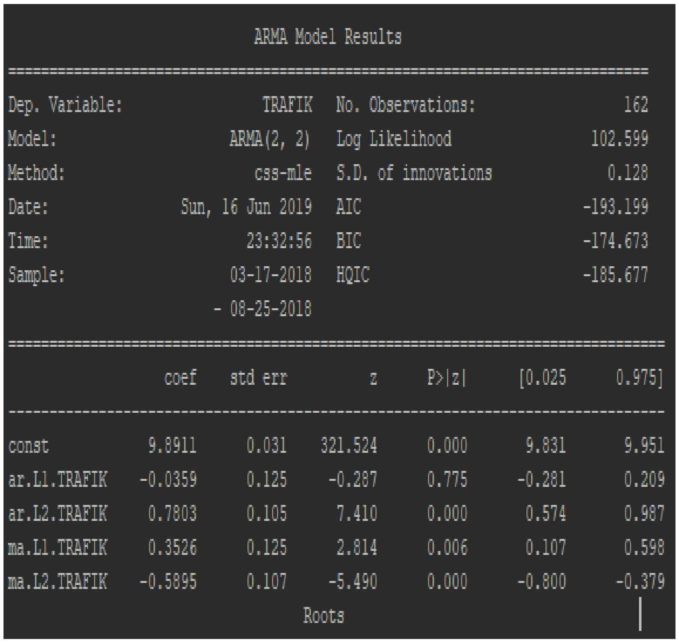
Differently from the previous case, we keep the same value of p and q, but we just change the order of differencing d = 1. We can see that the differencing order set equal to 1 just makes the case worse. Three of the parameters have a value larger than 0.05 and the AIC has still a large value. It just results that our PACF and ACF plot for (2,1) are not so correct and we have just to try other values to find the best one, that fits the model.

**Figure 4-18 Arima (2,1,2)**



```
-----------------------------------------------------------------------------
                             ARIMA Model Results
==============================================================================
Dep. Variable:               D.TRAFIK   No. Observations:                  161
Model:                 ARIMA(2, 1, 2)   Log Likelihood                  99.462
Method:                       css-mle   S.D. of innovations              0.130
Date:                Sun, 16 Jun 2019   AIC                           -186.923
Time:                        23:31:33   BIC                           -168.435
Sample:                    03-18-2018   HQIC                          -179.416
                         - 08-25-2018
==================================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------
const              0.0021      0.003      0.622      0.535      -0.005       0.009
ar.L1.D.TRAFIK    -0.9309      0.144     -6.481      0.000      -1.212      -0.649
ar.L2.D.TRAFIK    -0.0483      0.136     -0.354      0.724      -0.315       0.219
ma.L1.D.TRAFIK     0.2828      0.108      2.616      0.010       0.071       0.495
ma.L2.D.TRAFIK    -0.6402      0.107     -6.007      0.000      -0.849      -0.431
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1           -1.1418           +0.0000j            1.1418            0.5000
AR.2          -18.1493           +0.0000j           18.1493            0.5000
MA.1           -1.0483           +0.0000j            1.0483            0.5000
MA.2            1.4901           +0.0000j            1.4901            0.0000
-----------------------------------------------------------------------------
```

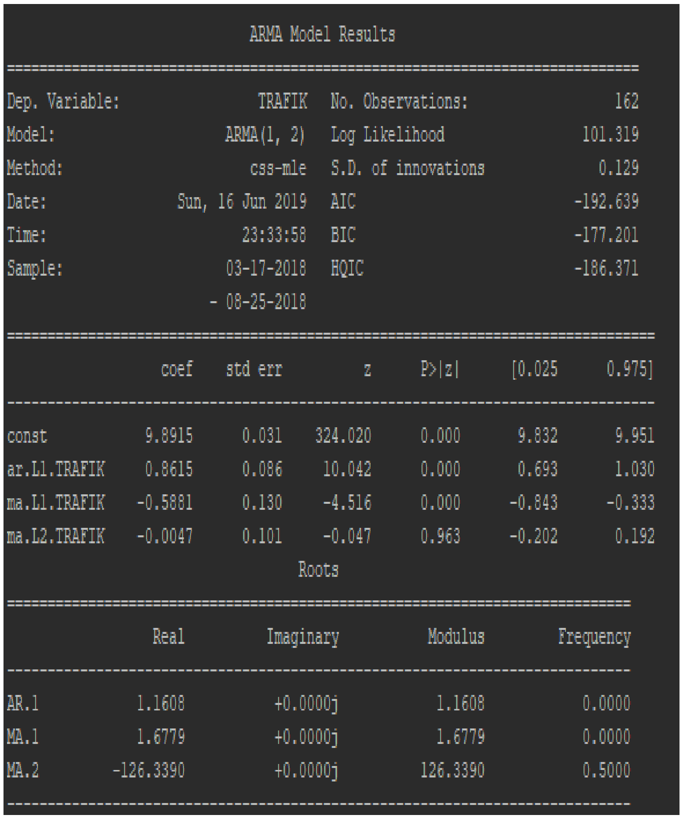
**Figure 4-19 Arima(2,0,2)**



```
                           ARMA Model Results
==============================================================================
Dep. Variable:                 TRAFIK   No. Observations:                  162
Model:                     ARMA(2, 2)   Log Likelihood                 102.599
Method:                       css-mle   S.D. of innovations              0.128
Date:                Sun, 16 Jun 2019   AIC                           -193.199
Time:                        23:32:56   BIC                           -174.673
Sample:                    03-17-2018   HQIC                          -185.677
                         - 08-25-2018
================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const            9.8911      0.031    321.524      0.000       9.831       9.951
ar.L1.TRAFIK    -0.0359      0.125     -0.287      0.775      -0.281       0.209
ar.L2.TRAFIK     0.7803      0.105      7.410      0.000       0.574       0.987
ma.L1.TRAFIK     0.3526      0.125      2.814      0.006       0.107       0.598
ma.L2.TRAFIK    -0.5895      0.107     -5.490      0.000      -0.800      -0.379
                                    Roots
```

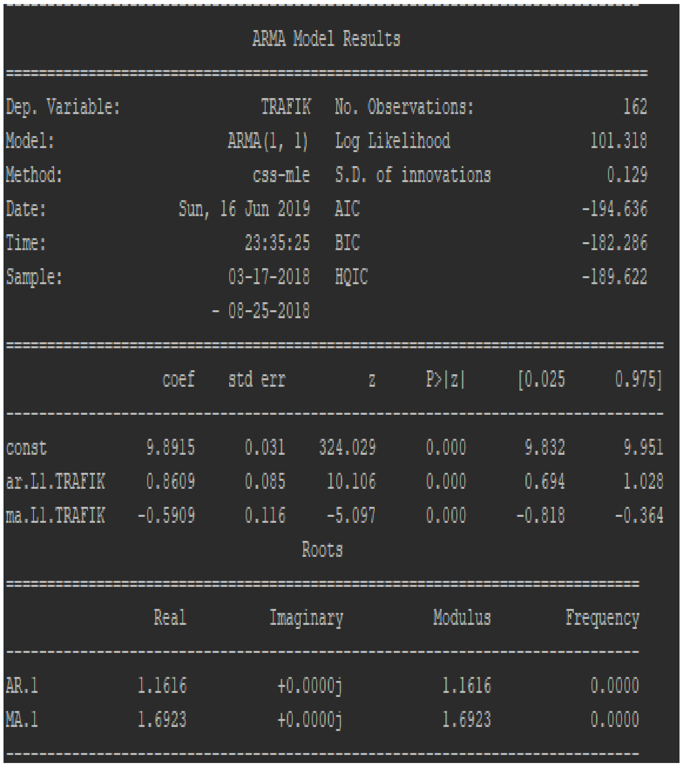
Yet it results that the outcome values still are not good. The value of p is still higher than 0.05. Let’s try with other values.

**Figure 4-20 Arima(1,0,2)**



```
                           ARMA Model Results
==============================================================================
Dep. Variable:                 TRAFIK   No. Observations:                  162
Model:                     ARMA(1, 2)   Log Likelihood                 101.319
Method:                       css-mle   S.D. of innovations              0.129
Date:                Sun, 16 Jun 2019   AIC                           -192.639
Time:                        23:33:58   BIC                           -177.201
Sample:                    03-17-2018   HQIC                          -186.371
                         - 08-25-2018
================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const            9.8915      0.031    324.020      0.000       9.832       9.951
ar.L1.TRAFIK     0.8615      0.086     10.042      0.000       0.693       1.030
ma.L1.TRAFIK    -0.5881      0.130     -4.516      0.000      -0.843      -0.333
ma.L2.TRAFIK    -0.0047      0.101     -0.047      0.963      -0.202       0.192
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.1608           +0.0000j            1.1608            0.0000
MA.1            1.6779           +0.0000j            1.6779            0.0000
MA.2         -126.3390           +0.0000j          126.3390            0.5000
-----------------------------------------------------------------------------
```

**Figure 4-21Arima (1,0,1)**

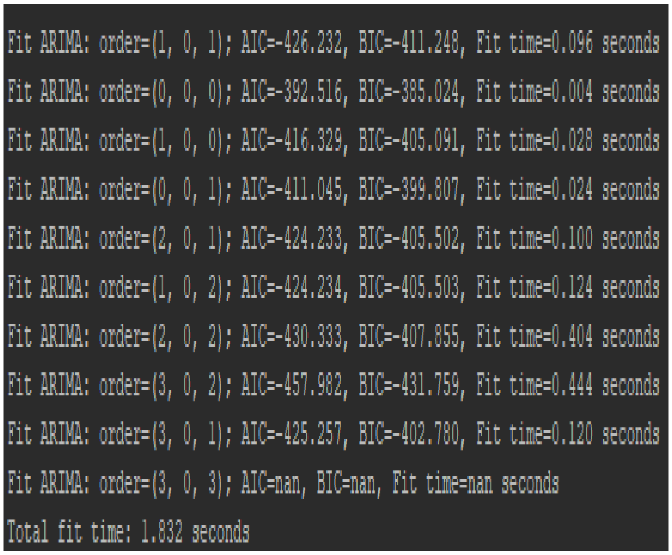


```
                           ARMA Model Results
==============================================================================
Dep. Variable:                 TRAFIK   No. Observations:                  162
Model:                     ARMA(1, 1)   Log Likelihood                 101.318
Method:                       css-mle   S.D. of innovations              0.129
Date:                Sun, 16 Jun 2019   AIC                           -194.636
Time:                        23:35:25   BIC                           -182.286
Sample:                    03-17-2018   HQIC                          -189.622
                         - 08-25-2018
================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const            9.8915      0.031    324.029      0.000       9.832       9.951
ar.L1.TRAFIK     0.8609      0.085     10.106      0.000       0.694       1.028
ma.L1.TRAFIK    -0.5909      0.116     -5.097      0.000      -0.818      -0.364
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.1616           +0.0000j            1.1616            0.0000
MA.1            1.6923           +0.0000j            1.6923            0.0000
-----------------------------------------------------------------------------
```

Finally, we can see that for p = 1 and q = 1 all values are less than 0 and Aic value is good. Now we are going to try with auto-Arima to check if there is any other value possible.

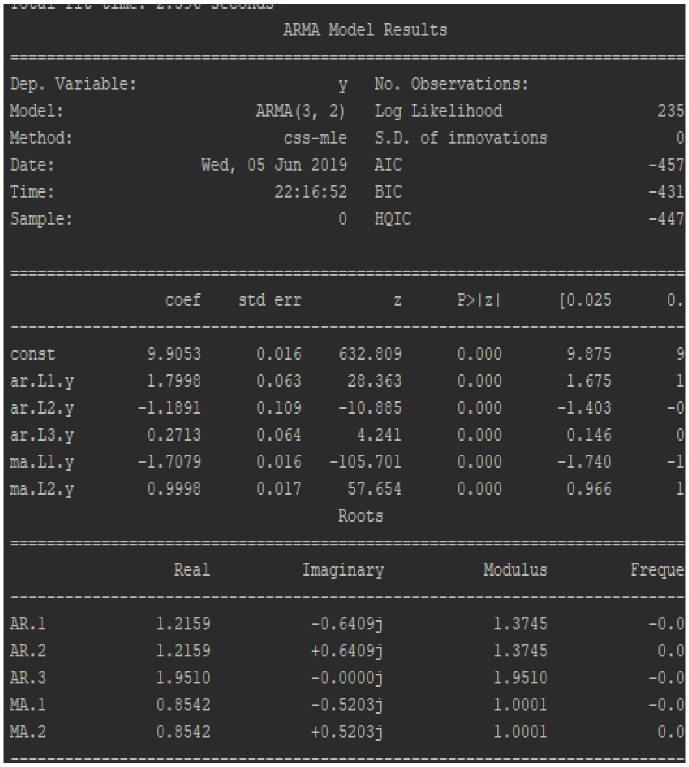
## 4.7 Auto Arima

**Figure 4-22 Auto-Arima checking all possibilities**



```
Fit ARIMA: order=(1, 0, 1); AIC=-426.232, BIC=-411.248, Fit time=0.096 seconds
Fit ARIMA: order=(0, 0, 0); AIC=-392.516, BIC=-385.024, Fit time=0.004 seconds
Fit ARIMA: order=(1, 0, 0); AIC=-416.329, BIC=-405.091, Fit time=0.028 seconds
Fit ARIMA: order=(0, 0, 1); AIC=-411.045, BIC=-399.807, Fit time=0.024 seconds
Fit ARIMA: order=(2, 0, 1); AIC=-424.233, BIC=-405.502, Fit time=0.100 seconds
Fit ARIMA: order=(1, 0, 2); AIC=-424.234, BIC=-405.503, Fit time=0.124 seconds
Fit ARIMA: order=(2, 0, 2); AIC=-430.333, BIC=-407.855, Fit time=0.404 seconds
Fit ARIMA: order=(3, 0, 2); AIC=-457.982, BIC=-431.759, Fit time=0.444 seconds
Fit ARIMA: order=(3, 0, 1); AIC=-425.257, BIC=-402.780, Fit time=0.120 seconds
Fit ARIMA: order=(3, 0, 3); AIC=nan, BIC=nan, Fit time=nan seconds
Total fit time: 1.832 seconds
```

### Figure 4-23 Arma(3,2)



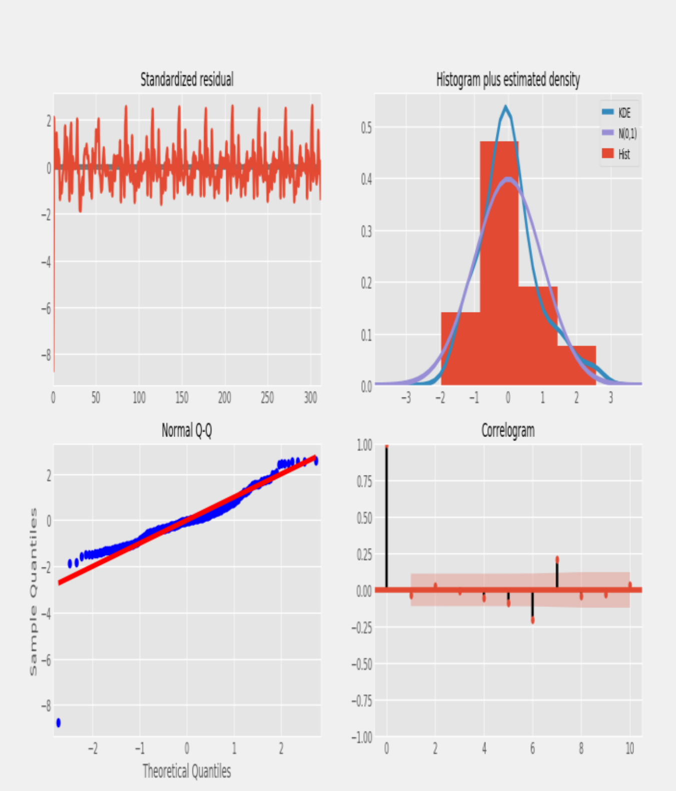
```
Total fit time: 2.930 seconds
                              ARMA Model Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  
Model:                     ARMA(3, 2)   Log Likelihood                  235
Method:                       css-mle   S.D. of innovations               0
Date:                Wed, 05 Jun 2019   AIC                            -457
Time:                        22:16:52   BIC                            -431
Sample:                             0   HQIC                           -447

==============================================================================
                 coef    std err          z      P>|z|      [0.025       0.
------------------------------------------------------------------------------
const          9.9053      0.016    632.809      0.000       9.875        9
ar.L1.y        1.7998      0.063     28.363      0.000       1.675        1
ar.L2.y       -1.1891      0.109    -10.885      0.000      -1.403       -0
ar.L3.y        0.2713      0.064      4.241      0.000       0.146        0
ma.L1.y       -1.7079      0.016   -105.701      0.000      -1.740       -1
ma.L2.y        0.9998      0.017     57.654      0.000       0.966        1
                                    Roots
==============================================================================
                  Real          Imaginary           Modulus         Freque
------------------------------------------------------------------------------
AR.1            1.2159           -0.6409j            1.3745           -0.0
AR.2            1.2159           +0.6409j            1.3745            0.0
AR.3            1.9510           -0.0000j            1.9510           -0.0
MA.1            0.8542           -0.5203j            1.0001           -0.0
MA.2            0.8542           +0.5203j            1.0001            0.0
------------------------------------------------------------------------------
```

Auto Arima checks all possible combinations of parameters to find the best one with the lowest AIC. The total time for auto Arima to finish this process is 1.832 seconds. The best model is Arma (3,2). we can see that all values of P are less than 0.05. It means that refuses the null hypothesis. Now we can check with the residual plots to be sure if it is the right model.

## 4.8 Interpret Residuals Plots in Arima

### Figure 4-24 Residuals Plots in Arima



There are four figures in the residual plots. Let’s interpret one by one:

-*Standardized residual* - the residuals rise and fall irregularly in the mean of zero. It doesn’t fluctuate in any another number, and it means that the prediction is not based.

- *Normal Q Q* – the dots of prediction fall over red line. There isn’t any significant deviation and the distribution is alright.

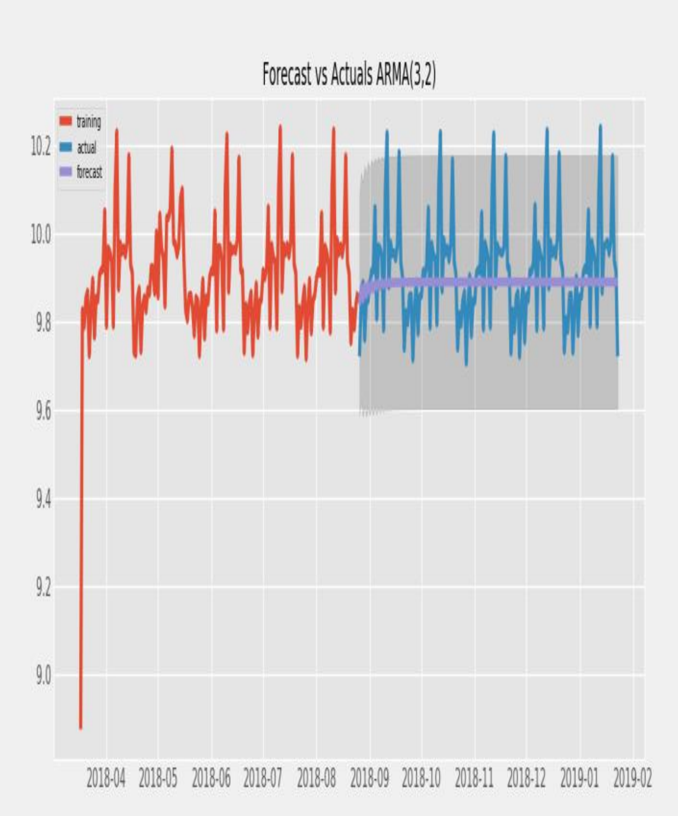
- *Histogram plus* estimated density – it shows that it has a normal distribution, and it has a mean 0.

- *Correlogram* – shows that the residual errors are not auto correlated. If there exists any autocorrelation, it means that there is a residual error that is not defined in the model.

It’s a good model. Now we can build our forecast, and we can evaluate it.

## 4.9 Forecasting

Figure 4-25 Forecast vs Actuals



As we mentioned before, we set the train set to 60% and the test set to 40%. The train set is from 18 March to September, while the test set is from September to January. So, we can evaluate my prediction for 5 months. We can see that the actual values are inside lower and upper series and forecast falls over the actual one. To do a better evaluation it is needed a bigger train set, so it can evaluate better if there exists any trend or seasonality. Anyways, we can say that for this train set we can predict the values of traffic for 5 months.

## 4.10 SARIMA

The first thing to do to build the Sarima model is to turn the time series into stationary. We achieved this while we were building the Arima model. So, we are now going to try to find all parameters of Sarima models.

Figure 4-26 ACF chart

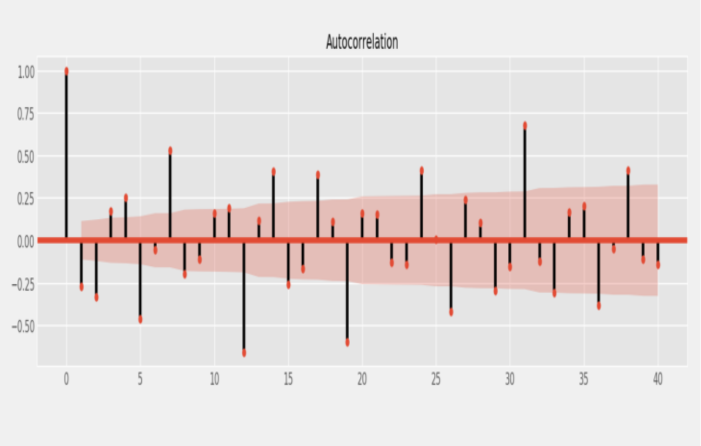
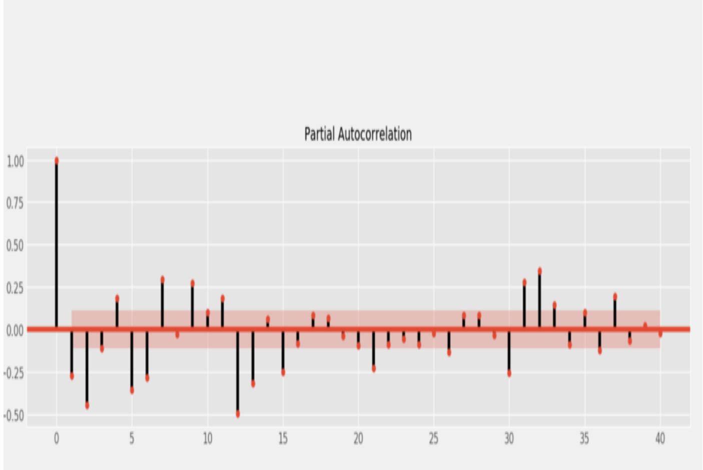
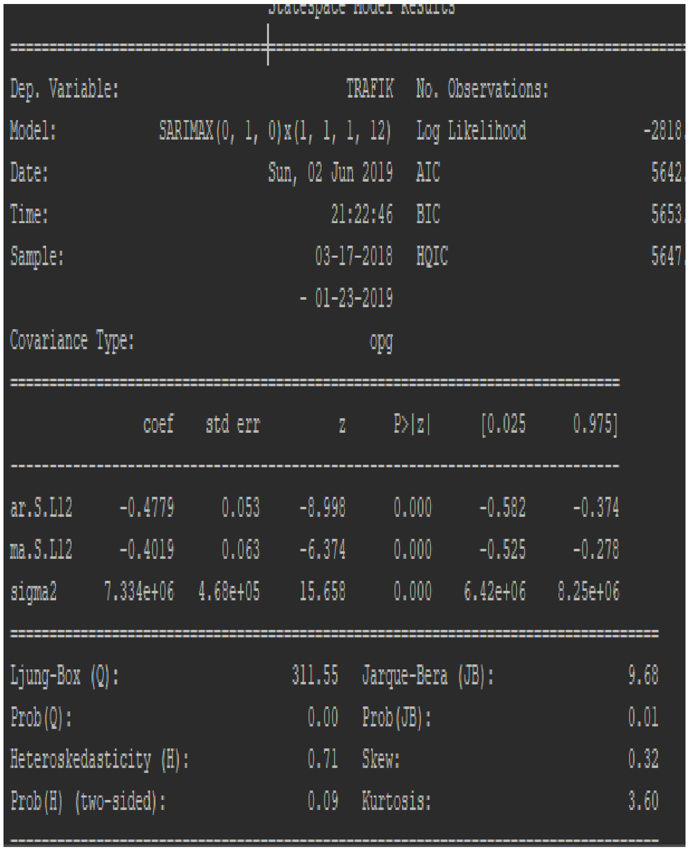


Figure 4-27 PACF charts



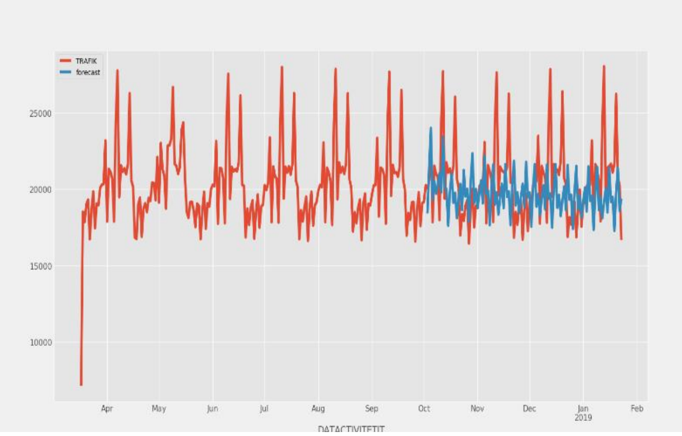
The highest value of the lag at the autocorrelation plot is at value 12, so the last parameter S = 12. Because S is negative at both parameters, the values of P and Q are going to be at 1. For the other parameters, we are going to use different combinations based on AIC and the p value.

**Figure 4-28 Sarima model**



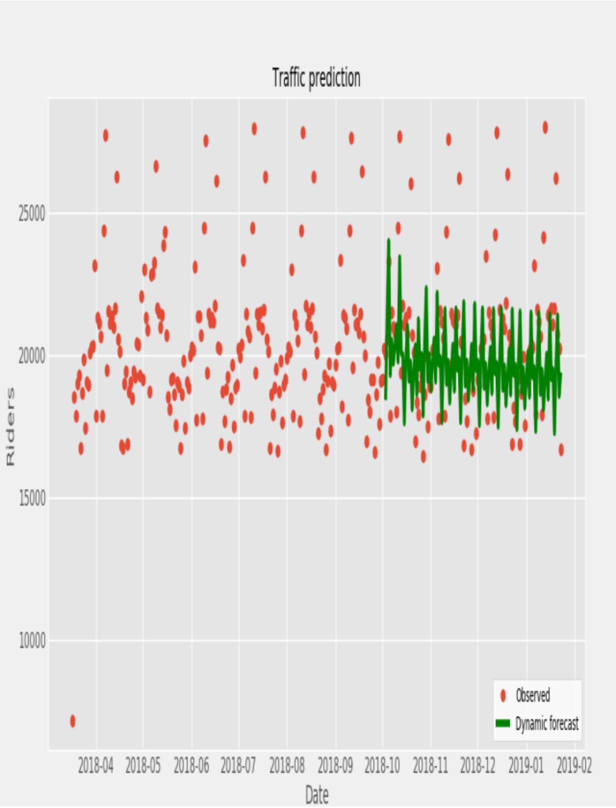
```
                                 Statespace Model Results
==========================================================================================
Dep. Variable:                          TRAFIK   No. Observations:
Model:            SARIMAX(0, 1, 0)x(1, 1, 1, 12)   Log Likelihood                 -2818.
Date:                         Sun, 02 Jun 2019   AIC                             5642.
Time:                                 21:22:46   BIC                             5653.
Sample:                             03-17-2018   HQIC                            5647.
                                  - 01-23-2019
Covariance Type:                           opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.S.L12      -0.4779      0.053     -8.998      0.000      -0.582      -0.374
ma.S.L12      -0.4019      0.063     -6.374      0.000      -0.525      -0.278
sigma2      7.334e+06   4.68e+05     15.658      0.000    6.42e+06    8.25e+06
===================================================================================
Ljung-Box (Q):                      311.55   Jarque-Bera (JB):                  9.68
Prob(Q):                              0.00   Prob(JB):                          0.01
Heteroskedasticity (H):               0.71   Skew:                              0.32
Prob(H) (two-sided):                  0.09   Kurtosis:                          3.60
===================================================================================
```

**Figure 4-29 Sarima forecast**



SARIMA (0,1,0) x(1,1,1,12) is the best model from what we tried. The p values are all below 0.05, but the value of AIC is very high. Anyway, we can say that overall, this is a good model. We can say this after we see our forecast values fit well in with the actual values for 5 months from October till January. Below we can see better our traffic prediction

**Figure 4-30 Sarima dynamic forecast**



# 5 Summary and Conclusion

## 5.1 Summary

The intention of this paper was to try to predict the traffic that passes in “Tirana-Durres” highway with ARIMA model. The first thing that we did was preprocess the data to make them in under stable for the program. The second step was to transform data into stationery. We used three algorithms: Time Shift Transformation, Exponential Decay Transformation and Log Scale Transformation. Only the first one didn’t manage to transform time series to stationery. Between Exponential Decay Transformation and Log Scale, we chose the second one because it had a lower value of P. After that we built the ACF and PACF for ARIMA model to define the p and q parameter. Both values for p and q were between 1 and 2. The next step was to compare three model AR model, MA model and Arima model based on RSS. The Arima model had the lowest value. Before finding out the best model, we split the dataset to train set and test set. Then we tried a different combination of Arima model with these parameters, but we didn't achieve any result. The values of p were never lower than 0.05 and the AIC's value was too high. So, we attempted with auto-Arima to find the right model and it was the ARMA model (3,2). The model was good, and the forecast was right for 5 months. After that we build the Sarima model using the same way as Arima model. We plot the ACF and PACF graph and we got the value of parameters for the Sarima model. After attempting some combination of parameters, we found that Sarima (0,1,0) (1,1,1)(12) was the best one. Even though the AIC value was very high, the p value was 0 and the prediction was correct for 5 months.

*References*

[1] Chaianong, Aksornchan, Christian Winzer, and Mario Gellrich. "Impacts of traffic data on short-term residential load forecasting before and during the COVID-19 pandemic." Energy Strategy Reviews 43 (2022): 100895.

[2] Wang, Haizhong, et al. "A novel work zone short-term vehicle-type specific traffic speed prediction model through the hybrid EMD–ARIMA framework." Transportmetrica B: Transport Dynamics 4.3 (2016): 159-186.

[3] Guo, Jinquan, Hongwen He, and Chao Sun. "ARIMA-based road gradient and vehicle velocity prediction for hybrid electric vehicle energy management." IEEE Transactions on Vehicular Technology 68.6 (2019): 5309-5320.

[4] Sukruti Vaghasia et al. (2018) University of Windsor, Ontorio, Canada. Retrieved from scholar.uwindsor.ca https://scholar.uwindsor.ca/cgi/viewcontent.cgi?article=8609&context=etd

[5] C. Narendra Babu and B. Eswara Reddy, (2015) from Department of Computer Science and Engineering, JNT University College of Engineering, Anantapuramu, India. Retrieved from https://www.itl.waw.pl/czasopisma/JTIT/2015/1/67.pdf

[6] Priyanka Rani (July 2018) Computer Science & Engineering I.K.G Punjab Technical University Kapurthala- Jalandhar Highway from https://pdfs.semanticscholar.org/3d31/6307e13a7c179e19fba5fbfd823013a57211.pdf

[7] Reza, R.M & Pulugurtha, Srinivas & Duddu, Venkata Ramana. (2015). "ARIMA Model for Forecasting Short-Term Travel Time due to Incidents in Spatio-Temporal Context", Conference: TRB 94th Annual MeetingAt: Washington D.C

[8] Billy M. Williams, M.ASCE1 and Lester A. Hoel (2013), F.ASCE2 from North Carolina, retrieved from https://www.researchgate.net/publication/234164011_Modeling_and_Forecasting_Vehicular_Traffic_Flow_as_a_Seasonal_ARIMA_Process_Theoretical_Basis_and_Empirical_Results

[9] Guoqiang Yu and Changshui Zhang, from Virginia Polytechnic Institute and State University from https://www.researchgate.net/publication/4087561_Switching_ARIMA_model_based_forecasting_for_traffic_flow

[10] Alghamdi, T., Elgazzar, K., Bayoumi, M., Sharaf, T., & Shah, S. (2019, June). Forecasting traffic congestion using ARIMA modeling. In *2019 15th international wireless communications & mobile computing conference (IWCMC)* (pp. 1227-1232). IEEE.

[11] Acun, F., & Gol, E. A. (2021, June). Traffic prediction on large scale traffic networks using ARIMA and K-means. In *2021 29th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE.

[12] Arlt, J., & Trcka, P. (2021). Automatic SARIMA modeling and forecast accuracy. *Communications in Statistics-Simulation and Computation*, *50*(10), 2949-2970.

[13] Shumway, Robert H., David S. Stoffer, Robert H. Shumway, and David S. Stoffer. "ARIMA models." *Time series analysis and its applications: with R examples* (2017): 75-163.

[14] Adineh, A. H., Narimani, Z., & Satapathy, S. C. (2020). Importance of data preprocessing in time series prediction using SARIMA: A case study. *International Journal of Knowledge-based and Intelligent Engineering Systems*, *24*(4), 331-342.

[15] Rabbani, M. B. A., Musarat, M. A., Alaloul, W. S., Rabbani, M. S., Maqsoom, A., Ayub, S., ... & Altaf, M. (2021). A comparison between seasonal autoregressive integrated moving average (SARIMA) and exponential smoothing (ES) based on time series model for forecasting road accidents. *Arabian Journal for Science and Engineering*, *46*(11), 11113-11138.

### Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

### Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

### Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.