Predictive Models for Data Traffic

1st Abdelrahman Yasser Zakaria  
202002033  
a.zakaria@nu.edu.eg

4th Ziad Tarek Fahem  
211001879  
z.tarek2179@nu.edu.eg 2nd Mohamed Osama Shosha  
202002755  
m.abouzeidshosha@nu.edu.eg

5th Omar Ayman  
202000537  
O.Ayman@nu.edu.eg 3rd SeifEldien Galal  
202000155  
s.hatem@nu.edu.eg

# **Introduction**

In a world where there is a growing amount of data traffic volumes, the demand for an efficient and accurate way to predict data traffic has never been more needed. Data traffic prediction is crucial in many ways that different telecommunication companies, streaming services, and internet companies can benefit from it to manage, plan, and optimize their networks to accommodate the predicted data traffic by studying its time series behavior to enhance the services they provide for the user, reduce additional costs that may occur, and reduce traffic congestion. To showcase the important of such predictions, one of the leading streaming services companies Netflix uses complex prediction algorithms in real-time to be able to anticipate the users’ data traffic on the services to ensure high-quality streaming experiences for their users during peak times.

In recent years, due to the advancements in the field of machine learning and data analysis, various approaches have been proposed to predict future data traffic. The different approaches that haven been proposed leverages different time series prediction algorithms. These approaches include approaches that are based on neural network like feedforward neural networks and approaches that are based statistical models like autoregressive integrated moving average (ARIMA). The ARIMA model is widely used because of its ability to describe autocorrelations in time series data.

In this paper we are aiming to build and use a statistical based model, autoregressive integrated moving average (ARIMA), to predict the future download data traffic of Youtube’s streaming service of the France city Lille using the dataset from the NetMob23 challenge. This dataset compromises of a large amount of mobile data traffic of various mobile applications across 20 different France cities from 16/03/2019 till 31/05/2019. In this paper we only utilized 1 week worth of data traffic from the dataset.

The rest of this paper is organized as follows. Section 2 will present and overview the related studies that have been done before. Section 3 will present our methodology that we took and go over the work we have done step by step. Section 4 will showcase the results that we produced. Finally, section 5 will present the conclusion remarks of the paper.

# **Related Studies**

This examination of the literature looks at how mobile data traffic forecasting is currently doing, with an emphasis on models and techniques that can be used to analyze and anticipate traffic trends in the context of Industry 4.0. We examine current methods for anticipating technological disruptiveness, evaluate the benefits and drawbacks of different forecasting strategies, and explore the uses and restrictions of the widely used ARIMA model. Additionally, we investigate the application of spatiotemporal models to mobile data flow.

This review attempts to lay the groundwork for future studies that will create more thorough and precise techniques for predicting mobile data traffic in this quickly changing technological environment by noting the gaps and difficulties in the current research.

An explosion of data has resulted from our growing reliance on internet services and mobile devices, providing researchers with an invaluable resource to study network dynamics and human behavior. The NetMob23 dataset, a collection of high-resolution, service-level mobile data traffic data gathered across 20 French metropolitan areas, is one well-known dataset used for this purpose. This dataset, which was made available by Martínez-Durive et al. [1], offers a rare chance to investigate spatiotemporal trends in mobile data consumption and evaluate how disruptive technologies affect network capacity.

**The NetMob23 dataset is superior to earlier datasets in several ways:**

Traffic at the service level: In contrast to previous datasets that mostly concentrated on call detail records (CDRs), NetMob23 records comprehensive data regarding the data traffic produced by 68 well-known mobile services. This makes it possible for researchers to examine how certain programs are consumed, offering insightful information on user behaviour and network usage.

High geographical resolution (data mapped to 100x100 m2 grids) and temporal resolution (15 minutes) are provided by the dataset, which allows for the comprehensive capture of traffic fluctuations over time. A more sophisticated knowledge of how traffic patterns change throughout various time periods and geographical regions is made possible by this degree of granularity.

Coverage of a developed nation: Unlike datasets that are mostly focused on underdeveloped nations, NetMob23 focuses on France, a developed nation, providing a new viewpoint. This makes it possible to compare and gain understanding of the distinctive features of mobile traffic in developed nations.

**The authors use several analyses to show how rich the dataset is:**

**Anomaly Detection:** The dataset highlights the significance of considering real-world disturbances when evaluating network data by revealing a variety of anomalies, such as network failures impacting regions or services.

**Time Patterns:** Different applications show peak usage at periods of the day or week, reflecting normal user behavior. The analysis indicates diverse temporal patterns in mobile service consumption.

**Spatial Patterns:** The authors show notable differences in the spatial distribution of traffic, with some applications concentrating parts of a city, which reflects the features of the urban environment and the activities of its users.

The authors also stress the importance of merging the NetMob23 dataset with information from other sources, including socioeconomic indicators, administrative boundaries, and telecommunication data. An even more comprehensive understanding of the variables affecting mobile data traffic is provided by this integrated method.

With its emphasis on service-level traffic, high spatiotemporal resolution, and coverage of a developed nation, the NetMob23 dataset is a useful tool for scholars delving into the intricate dynamics of mobile data traffic.

**Several important conclusions are drawn from the author’s analysis:**

Unique Spatiotemporal Trends: The use of mobile services follows distinct temporal patterns, with peak usage hours that differ depending on the application and represent user habits and behavior. Additionally, the dataset shows notable regional differences in traffic distribution that are impacted by user behavior and urban factors.

**Impact of Real-World Disruptions:** When evaluating network data, it is crucial to consider real-world disruptions. This is highlighted by the existence of anomalies in the dataset, such as network outages that impact regions or services.

**Potential for Integrated Analysis:** An integrated understanding of the factors impacting mobile data traffic can be achieved by combining NetMob23 with other data sources, such as administrative boundaries, socio-economic indicators, and telecommunication data.

These insights highlight how important spatiotemporal data analysis is to comprehending mobile data traffic and its effects on network management. Future study examining the effects of disruptive technologies, urban planning, and user behaviour on mobile network demand is made possible by the comprehensive data offered by NetMob23 and the possibility of combining other data sources.

An extensive study on traffic congestion forecasting using the ARIMA (Autoregressive Integrated Moving Average) model is presented by Alghamdi et al. [2]. Their study sheds light on the difficulties and possibilities involved in interpreting short-term, non-stationary, and irregularly dispersed traffic data. The authors provide an example of how to efficiently preprocess and get ready such data for ARIMA modeling so that traffic flow patterns can be accurately predicted.

The study focuses on an actual dataset that was gathered from the Caltrans performance measurement system, which records hourly traffic flow observations throughout California's freeway network. The writers discuss a number of important topics regarding the study of traffic data:

**Non-stationary Data:** The authors use differencing techniques to convert the data into a stationary series so that it may be used for ARIMA modeling, acknowledging that real-time traffic data frequently displays non-stationarity.

**Non-normal Distribution:** To verify that the data satisfies the requirements of ARIMA modeling, the authors address the problem of non-normal data distribution by using statistical techniques such the Box-Cox transformation and unit root testing.

To enhance model performance, the study highlights the significance of precisely adjusting the ARIMA parameters (p, d, and q) using a combination of ACF and PACF analysis, the Bayesian Information Criterion (BIC), and Akaike's Information Criteria (AIC).

Key Results and Conclusions:

**Model Selection:** By evaluating the results of various model parameter combinations, the authors show how useful it is to use both AIC and BIC to determine the best ARIMA model.

**Model Performance:** The analysis shows that the ARIMA (2,1,3) model performs better in terms of accuracy, as evidenced by its lowest Mean Absolute Scaled Error (MASE) value. By accurately capturing the underlying patterns in the traffic data, this model generates projections of future traffic flow that are precise.

**Residual Analysis:** To further confirm the accuracy and dependability of the model, the authors stress the significance of examining the residuals to make sure they are normally distributed, uncorrelated, and have a zero mean.

The study concludes that the ARIMA model is a potent tool for short-term traffic forecasting, especially for non-stationary and non-normally distributed data, when used appropriately and fine-tuned. For transportation engineers and academics looking to enhance traffic management and prediction skills, their research offers a comprehensive manual on how to preprocess, choose parameters, and assess ARIMA models.

The authors are aware of the shortcomings of conventional ARIMA modeling, which frequently has trouble projecting traffic patterns correctly in the face of disruptive technologies. They use a disruptive formula that includes four essential components to overcome this, introducing a judging approach:

**Market Time (TTM):** This variable evaluates the rate at which new technologies are embraced and put into practice, considering elements such as user behavior potential and ease of integration.

**Cost:** Adoption rates and total demand are influenced by the expenses associated with deploying and utilizing new technologies.

**PEST Analysis (Political, Economic, societal, Technological):** This approach assesses the external factors—such as governmental policies, prevailing economic conditions, societal trends, and technological advancements-that impact the adoption of technology.

**Market Share:** A new technology's market share indicates how well-liked it is and how much room it must develop in the future, which affects how much data is used overall.

**Key Results and Conclusions:**

Increased Accuracy: The authors show that the ARIMA model considerably increases the accuracy of mobile data traffic forecasting when paired with the disruptive formula, especially when compared to the traditional ARIMA model alone. The reduced error rates seen for both 3G and 4G traffic estimates clearly demonstrate this.

Cost Influence: The study emphasizes how important a factor cost is in causing disruptive traffic. This accentuates how important it is to think about how new technologies will affect user behavior and how much they will cost.   
Future Implications: According to the study, combining statistical and judging methods—especially when taking disruptive variables into account—allows for a more accurate prediction of mobile data traffic, which can be useful for Industry 4.0 network planning and resource allocation.

The methodology employed by the authors offers a useful framework for comprehending the intricate relationship between mobile data traffic and technical interruptions. Their study emphasizes the necessity of a more thorough approach to forecasting that takes into consideration judgmental elements along with statistical models to account for Industry 4.0's dynamic and frequently unpredictable nature.

**Related Works Conclusion:**

To maximize network performance and resource allocation in the era of Industry 4.0, precise projections of mobile data traffic are crucial, as this literature study examined the state of the field today. Three well-known research offer insightful information about this intricate field:

The NetMob23 dataset was presented by Martínez-Durive et al. [1] as a valuable resource for high spatiotemporal resolution analysis of service-level mobile data traffic patterns. Their findings highlight the significance of considering both spatial and temporal variables, in addition to the effects of disruptive technology and actual events on network demand.

The usefulness of the ARIMA model for short-term traffic forecasting was shown by Alghamdi et al. [2], especially for non-stationary and non-normally distributed data. For the benefit of transportation engineers and academics, their study offers a thorough manual on how to preprocess, choose parameters, and assess ARIMA models.

To solve the problem of traffic prediction in the context of Industry 4.0, Arifin and Habibie [3] combined the ARIMA model with a "disruptive formula" that takes the impact of new technologies into consideration. Their work emphasizes that to fully represent the dynamic and frequently unpredictable nature of Industry 4.0; judgmental elements must be considered in addition to statistical models.

All this research points to the necessity of approaching mobile data traffic forecasting from multiple angles. Future studies should concentrate on: Creating more complex models that consider the temporal as well as spatial components of traffic, taking disruptive technology into account.   
investigating how to improve forecast accuracy by integrating data from various sources, such as socioeconomic factors and user behavior data. Tackling the problem of traffic pattern forecasting while taking into consideration the special traits and network demand consequences of new technologies such as 5G and the Internet of Things.

Researchers and network operators may better negotiate the complexity of Industry 4.0 and make sure that networks are ready to manage the growing needs of a connected world by adopting a comprehensive and creative approach to mobile data traffic forecasting.

# **Methodology**

## **Dataset:**

The dataset that we used is the NetMob23 dataset. The NetMob23 dataset compromises of a large amount of mobile data traffic taken at 15-minutes intervals of various mobile applications across 20 different France cities from 16/03/2019 till 31/05/2019 which one can choose from.

In our case we only utilized 2 weeks (16/03/2019 to 29/03/2019) worth of download data traffic of the in the city Lille for streaming service Youtube.

The resulting dataset used is compromised of 1344 observations that have been obtain in 15-minutes intervals of 2 weeks of the traffic and have 2 attributes “TileID” and “Traffic” for the Lille city.

## **Preprocessing:**

### **Excluding data outside the city:**

To make sure we only include traffic data of the city, we used the a geojson file of the Lille city which is provided by the dataset providers. This geojson file contains some geographical information about the city such as city coordinates and they also include all the tiles ids of the tiles inside the city.

To exclude tiles data outside the city we made use of this geojson file. We went through our dataset and searched for each tile id in the dataset inside this geojson file if it exists then its inside our city and if it doesn’t the tile will be dropped from the dataset.

After doing this we found that all the tile IDs inside our dataset is indeed inside the city boundaries.

### **Checking for missing data:**

To check for missing data, we processed the dataset to check for an NA or 0 valued data. If a datapoint is found to be valued NA or 0 it will be printed.

After processing the dataset we found no missing data.

### **Checking for outliers:**

To check for outliers in the dataset, we applied 2 techniques boxplots and z-score.

In the first technique we applied boxplots. To detect outliers using this technique we plotted the traffic data and if any outlier is present it the data point will be plotted outside the box plot. However, we found no outliers using this technique.

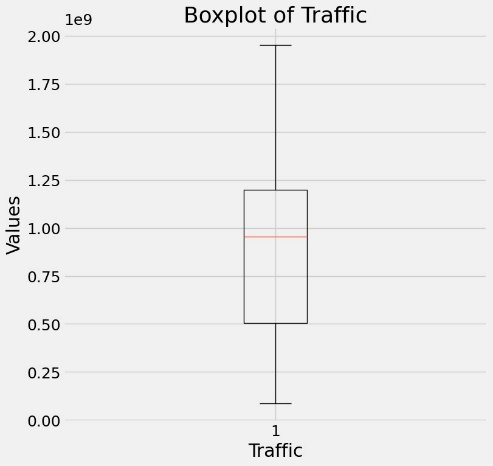


Figure 1: Boxplot.

In the second technique we used z-score where datapoints with z-scores greater than 2 and less than -2 are considered outliers. However, in the z-score technique we found 1 outlier and we replace traffic value for this outlier by traffic value of previous Timestamp.

### **Data Transformation:**

Our dataset consists mainly of a column “TilesID” and many columns each representing a timestamp. This format of is not practical to be furtherly operated on and we need to transform it into a more practical format.

To transform the dataset we first decided on a practical format to be easily used in our project. We decided transforming it to a basic time series table.

Firstly, we summed all the traffic data regardless of the tile id with its timestamp since in our project we are not focusing on the where the traffic is coming from inside the city. Secondly, Instead of having a column for each time stamp we created 2 new columns which are called timestamp and traffic. The timestamp column will hold all the timestamps and the traffic column will hold all the traffic corresponding to each timestamp.

## **Exploratory Data Analysis (EDA):**

Before continuing to the building of the ARIMA model, we first have to apply some exploratory data analysis on the dataset to obtain insights.

### **Traffic pattern across the week:**

To explore activity levels across the week, we plotted a time series plot of the traffic across the week.

The plot, shown in figure 1, showcased that the highest traffic can be noticed at the end of the 20th of March in our week. Furthermore, the plot showcased that the traffic is pretty consistent across the week.

A graph showing the time and time

Description automatically generated with medium confidence

Figure 2: Traffic across the week.

### **Traffic Median & Mean across the week:**

To view the median and mean traffic for each day in the week we plotted them as shown in figure 2.

A graph of a line graph

Description automatically generated with medium confidence

Figure 3: Traffic Mean & Median

### **Traffic density:**

To view the density of the traffic, we made a density plot for the traffic as shown in figure 3. The plot shows significant clustering around two values, one around 0.3 billion and another around 1.0 billion. The higher peak around 1.0 billion suggests that this traffic value is more common than the one around 0.3 billion.

A graph with a blue line

Description automatically generated

Figure 4: Traffic Density

### **Hourly Traffic Trends:**

To view the hourly trend of all the days in our week, we plotted a heatmap. The heatmap showcases the hour of each day and its traffic heat. The lighter the cell the more traffic and the darker the cell the less the traffic.

The plot gave us some insight into the hourly trend of the traffic. We found out that after working hours (5 PM) that there are more traffic than before working hours. This showcases that people usually use Youtube mostly after finishing work or school.

A chart of heatmap

Description automatically generated

Figure 5: Hourly Traffic Trends Heatmap

### **Time series Components:**

In general, most time series can be decomposed into three major components. The first is seasonality, which describes the periodic signal in your time series. The second component is a trend, which describes whether the time series is decreasing, constant, or increasing over time. Finally, the third component is noise, which describes the unexplained variance and volatility of your time series.

#### **Seasonality Component:**

The seasonality component describes the periodic signal in the time series. A daily seasonal pattern might exist due to influences like working hours and sleeping hours such as seen in figure 5.

A graph showing the value of a traffic time series

Description automatically generated

Figure 6: Seasonal Component Plot

#### **Trend Component:**

The trend component shows whether the time series is decreasing, constant, or increasing over time. For instance, in figure 6 we observed that the traffic increased significantly on March 20th, which might be due to a special event on that day.

A graph showing a line

Description automatically generated

Figure 7: Trend Component Plot

#### **Noise Component:**

The noise, or residual component, represents the random, irregular influences that cannot be attributed to either trend or seasonality. For instance, this can be seen on March 17th in figure 7, where there are only irregular influences present.

A graph showing a graph

Description automatically generated with medium confidence

Figure 8: Noise Component Plot

## **ARIMA Model Building:**

Among many methods used to estimate data traffic, ARIMA is the most used technique for studying traffic by predicting the variation of traffic data over the time domain. Some studies introduced the ARIMA approach for time series analysis, given that the traffic data is normally distributed. First we check whether our data is stationary or not. Next, in order to make predictions, we estimate the ARIMA model's parameters. The choice of these parameters has a significant impact on the model's performance, as the next subsections will demonstrate.

### **Autoregressive integrated moving average**

We use the standard ARIMA to directly capture autocorrelation [20] in order to model time series. It can produce powerful mathematics and statistical theory as its foundation.   
  
intervals of prediction. The ARIMA model is represented by three parameters (AR, I, and MA) that have a statistically significant impact on the model accuracy: (p, d, q) stand for the moving average window size, difference order, and autoregressive, respectively. In order to determine these three parameters, we fit the ARIMA model to the various series after applying differentiation lag-1 for a moving trend or seasonal differencing. Equation (1) is used to find ARMA (p, q):

Equation 1

The common technique for figuring out these characteristics involves looking at the correlation and partial correlation charts in addition to visually inspecting the time series to look for trends. Because the values of our dependent traffic flow variable, measured at time T, are influenced by the traffic flow values in the past, disregarding lags has a negative impact on the model's implementation on non-normal distribution data.

We split the data into training and testing



Figure 9: Data Train & Test Split

There are many ways to test stationary, one of them with eyes, and others are more formal using statistical tests. There are also ways to transform non-stationary time series into stationary ones. We’ll address both of these in this subsection and then we will be ready to start modelling.

The most common test for identifying whether a time series is non-stationary is the augmented Dicky-Fuller test. This is a statistical test, where the null hypothesis is that your time series is non-stationary due to trends. We can implement the augmented Dicky-Fuller test using stats models. First, we import the adfuller function as shown, then we can run it on our data.

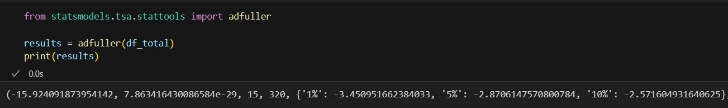


Figure 10: Dicky-Fuller Test Results

The last item in the tuple is a dictionary. This stores the critical values of the test statistic which equate to different p-values. In this case, if we wanted a p-value of 0.05 (5%) or below, our test statistic needed to be below -2.8706.The zeroth element is the test statistic, in this case, it is -15.924. Based on this result, we are sure that the time series is stationary.

### **Using ACF and PACF to find the best model parameters**

The ACF can be defined as the correlation between a time series and itself with n lags. So ACF(I) is the correlation between the time series and a one-step lagged version of itself. An ACF(2) is the correlation between the time series and a two-steps lagged version of itself and so on.

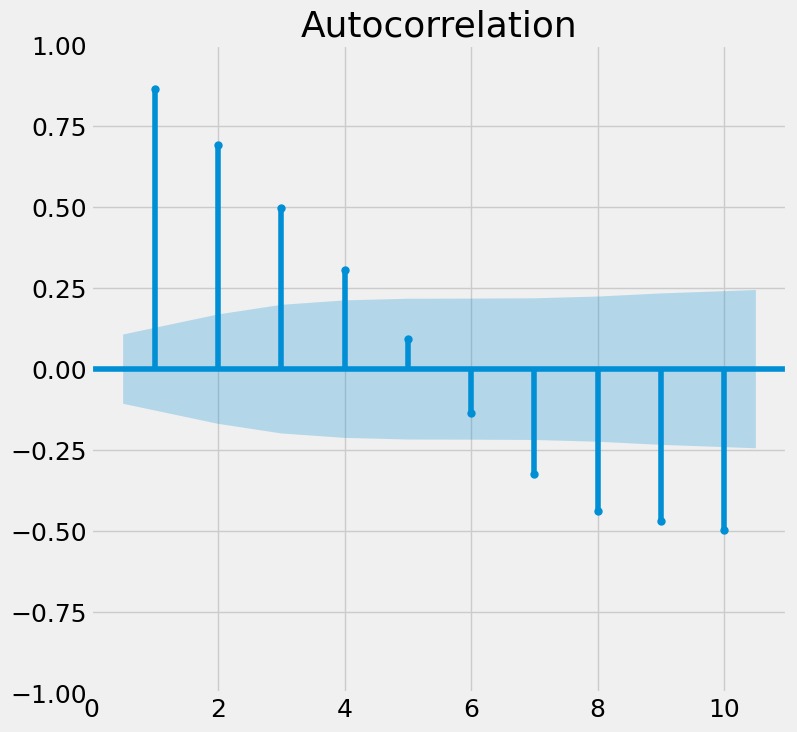


Figure 11: ACF Plot

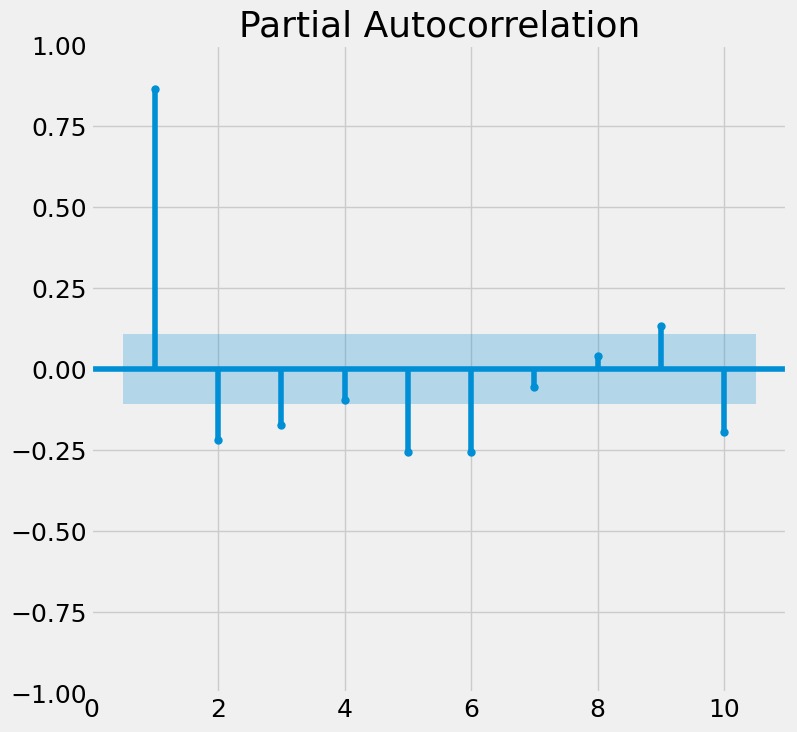


Figure 12: PACF Plot

However, in our case we couldn’t determine the best parameters using ACF and PACF. Now to achieve the highest accuracy of our model, we made a trial and error function and stored the values of the ARIMA parameters in a list that corresponds with the AIC and BIC values. Then we converted this list to a data frame and sort them based on AIC and BIC. This function shows that ARIMA (5,0,7) gave smallest AIC and BIC and it is the recommended model parameters.

# **Results**

### **The model diagnostic**

The next step is to diagnose `the model to know whether the model is behaving well or not. To diagnose the model we will focus on the residuals of the training data. The residuals are the difference between the model’s one-step-ahead predictions and the real values of the time series. We plot some diagnostics graphs

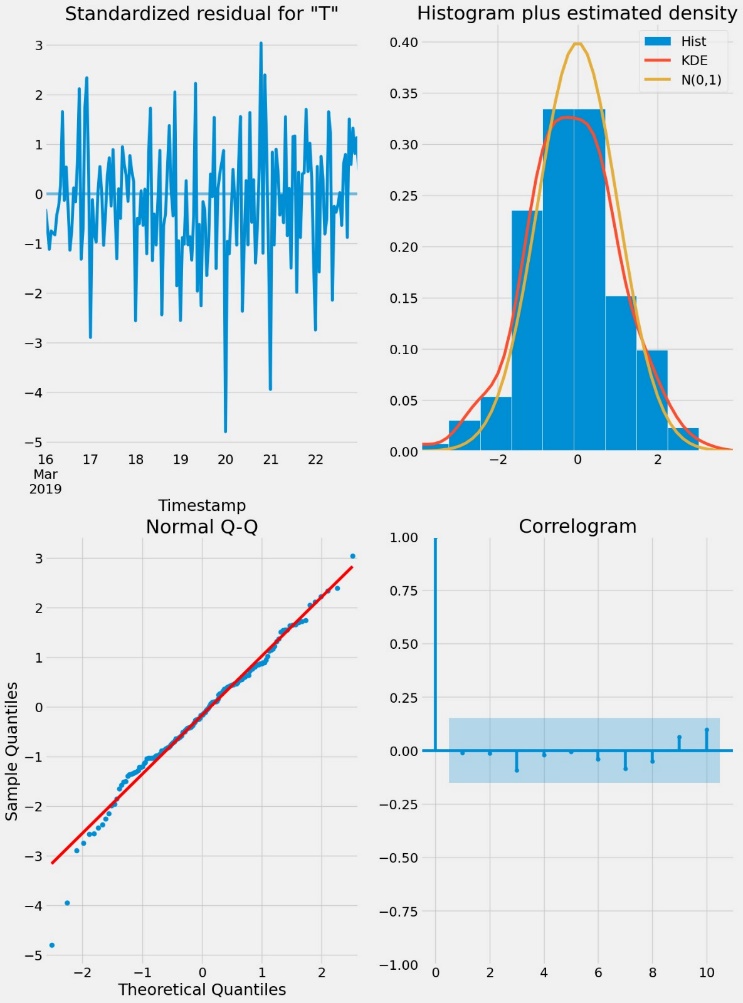


Figure 13: Residual Diagnosis

There are four plots in the residual’s diagnostic plots:

1-Standardized residuals plot: The top left plot shows one-step-ahead standardized residuals. If our model is working correctly, there should be no obvious pattern in the residuals. This is shown here in this case.

2-Histogram plus estimated density plot: This plot shows the distribution of the residuals. The histogram shows us the measured distribution; the orange line shows a smoothed version of this histogram, and the green line shows a normal distribution. If the model is good these two lines should be the same. Here there are small differences between them, which indicate that our model is doing well .

3-Normal Q-Q plot: The Q-Q plot compare the distribution of the residuals to the normal distribution. If the distribution of the residuals is normal, then all the points should lie along the red line, except for some values at the end. Correlogram plot:

4-The correlogram plot is the ACF plot of the residuals rather than the data. 95% of the correlations for lag greater than zero should not be significant (within the blue shades). If there is a significant correlation in the residuals, it means that there is information in the data that was not captured by the model.

### **Making Predictions:**

We used get\_forecast() method to predict the traffic of the next day "2019-03-23" and we visualize these predictions

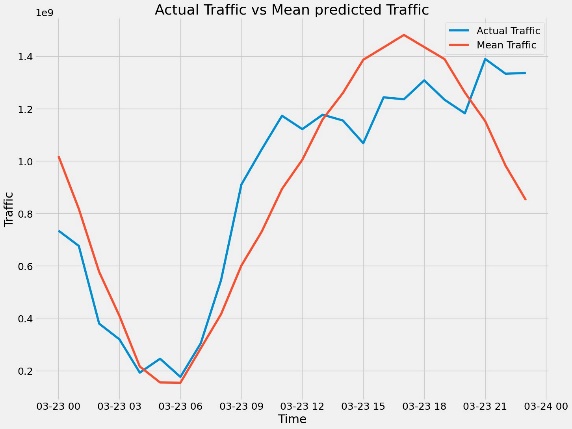


Figure 14: Mean predicted traffic vs actual traffic.

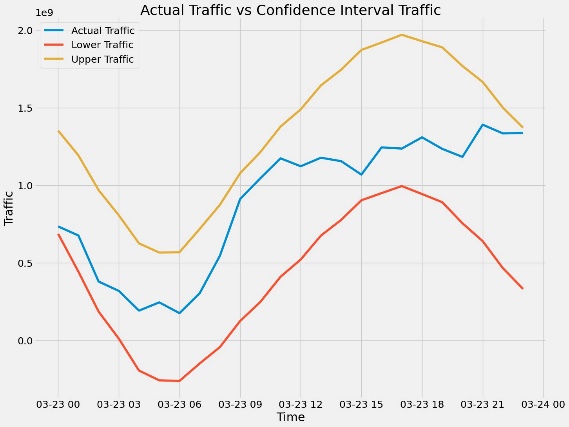


Figure 15: Confidence interval traffic vs Actual Traffic

we see that actual traffic always within the confidence interval.

|  |  |  |  |
| --- | --- | --- | --- |
| Data | R2 Score | RMS | MAE |
| Train | 0.85 | 162113034.88 | 123526858.92 |
| Test | 0.82 | 173435537.94 | 137383754.76 |

Table 1: Model Evaluation

# **Conclusion**

In conclusion, in this paper we used a subset of the NetMob23 dataset to apply some exploratory data analysis and build a predictive model based on ARIMA to forecast the traffic. The subset of the NetMob23 contains a week (16/03/2019 to 22/03/2019) worth of the traffic data for the streaming service Youtube in the France city Lille. We first applied some preprocessing steps on this subset of the dataset to make it ready to be further explored and applied in our ARIMA model. After preprocessing the subset of the dataset we applied some exploratory data analysis techniques on it and acquired some insights. Finally, using the dataset we successfully built an ARIMA model to be able to predict the future data traffic.

##### **References**

1. Orlando E. Martínez-Durive, undefined., et al, "The NetMob23 Dataset: A High-resolution Multi-region Service-level Mobile Data Traffic Cartography," 2023.
2. Alghamdi, Taghreed & Elgazzar, Khalid & Bayoumi, Magdi & Sharaf, Taysseer & Shah, Sumit. (2019). Forecasting Traffic Congestion Using ARIMA Modeling. 1227-1232. 10.1109/IWCMC.2019.8766698.
3. Arifin, Ajib & Habibie, Muhammad. (2020). The prediction of mobile data traffic based on the ARIMA model and disruptive formula in industry 4.0: A case study in Jakarta, Indonesia. TELKOMNIKA (Telecommunication Computing Electronics and Control). 18. 907. 10.12928/telkomnika.v18i2.12989
4. Hanyu Yang, Xutao Li, Wenhao Qiang, Yuhan Zhao, Wei Zhang, & Chang Tang (2021). A network traffic forecasting method based on SA optimized ARIMA–BP neural network. Computer Networks, 193, 108102.
5. Alghamdi, T., et al, "Forecasting Traffic Congestion Using ARIMA Modeling," in 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC), 2019, pp. 1227-1232.