Stop-Based Time Series Traffic Change Analysis Using Data from Istanbul Metropolitan Municipality (IMM)

Kaan Çolakoğlu
Computer Science
Galatasaray University
Istanbul,Turkey
kaan.colakoglu@ogr.gsu.edu.tr

Abstract—**Traffic congestion is an important urban problem that affects travel time, air quality and the general well-being of the public. This study analyses and forecasts traffic density patterns in the Beşiktaş coastal region of Istanbul, using hourly vehicle count data collected in September 2024. The research determines peak traffic hours, examines the differences between weekdays and weekends, and investigates the relationship between vehicle density and average speed.**

Descriptive analysis shows that traffic density peaks during the morning (07:00-09:00) and evening (16:00-19:00) hours, with higher congestion observed on weekdays compared to weekends. Ordinary Least Squares (OLS) regression analysis suggests a weak inverse relationship between traffic density and average speed, though the low R^2 value indicates limited explanatory power for this model, highlighting the need for additional factors or variables to better explain the observed variation.

Three models - Seasonal ARIMA (SARIMA), Facebook Prophet and Random Forest - are implemented and evaluated to predict traffic patterns. The results show that Random Forest provides the most accurate short-term predictions, while SARIMA and Prophet effectively capture seasonal trends in traffic density.

This study provides valuable insights into urban traffic behaviour and a comparative evaluation of forecasting methods. The results can be used as a basis for the development of efficient traffic management strategies and decision-making processes in congested areas.

I. Introduction

In this document we will describe and analyse our dataset for predicting the possible traffic hours.

A. Overview

Today, the rapidly growing population and the number of vehicles in cities have increased the importance of urban planning and traffic management. In this project, traffic density is analyzed by focusing on the number of vehicles. The aim of the study is to understand how the variations in the number of vehicles differ by days of the week and hours of the day.

In this analysis, one month of traffic data from the Istanbul Metropolitan Municipality covering a specific region is used. These data are analyzed to understand the traffic flow within the city and to provide a basis for future traffic management decisions.

B. Motivation

This study has been done to understand the dynamics of traffic congestion and use this information for more optimized transportation planning. In big cities, traffic congestion causes a lot of problems from wasting time to polluting the environment. Especially in the rush hours like morning and evening, increasing commute times affect the quality of life of individuals and societies.

Data selected from geohash of sxk9kp is for understanding the reasons of the problems in regional level. This regions analysis will not only help to understand today's conditions and will also form a basis for traffic prediction models and traffic management systems. The results of this study will also help to determine the measures for solving traffic problems.

This study aims to effectively examine how traffic congestion changes by day of the week, how it fluctuates in hours of the day, is there a pattern and if there is what is this pattern? With these obtained results, the development of sustainable transportation solutions and traffic management strategies will be predicted.

C. Research Questions

1) How does traffic density change at different times of the day?
   - How does vehicle density change on certain days of the week (e.g. weekdays and weekends)?
2) Is there a significant difference between weekday and weekend traffic density?
   - Can the traffic situation be categorically classified using vehicle density? Are the results of this categorization consistent with daily or hourly analysis?
3) What is the relationship between density and speed ?
   - Is there a statistically significant relationship between the number of vehicles and average speed?
4) How accurately can future traffic density be predicted using time series forecasting methods?
   - Which model (SARIMA, Prophet, Random Forest) performs better in short-term and long-term traffic forecasts?

D. Literature Review

Several studies have explored traffic forecasting using various time-series and machine learning models. Ghosh et al. [1] utilized a Bayesian SARIMA model for short-term traffic flow forecasting, highlighting its ability to handle extreme fluctuations and provide flexible confidence intervals. Similarly, Kumar and Vanajakshi [4] applied Seasonal ARIMA (SARIMA) for traffic prediction, achieving acceptable accuracy with limited data, making it suitable for Intelligent Transportation Systems (ITS) applications.

In the domain of machine learning, Liu and Wu [2] employed a Random Forest-based model to predict urban traffic congestion, achieving an accuracy of 87.5%. Han and Shi [5] extended this approach by utilizing real-time data for Random Forest predictions, reducing dependence on historical data.

Hybrid models have also been investigated to improve prediction accuracy. Wang et al. [6] proposed a hybrid approach combining SARIMA, Random Forest, and other methods to enhance traffic volume prediction at busy road junctions. Hong et al. [7] demonstrated the effectiveness of integrating SVR with ant colony optimization techniques for forecasting complex traffic patterns.

Despite these advancements, limited research has been conducted on regional-level traffic analysis with comprehensive descriptive analytics. This study bridges this gap by focusing on traffic density in the Beşiktaş region, combining detailed descriptive analysis with model performance evaluation.

E. Contributions

The contributions of this study can be summarised as follows:

- A region-specific analysis was conducted, focusing on the coastal area of Beşiktaş in Istanbul, addressing local traffic congestion issues rather than city-wide or generalised models.
- Comprehensive descriptive analysis was performed, including visualizations such as box plots and heat maps to reveal daily and weekly traffic patterns.
- A comparative evaluation of three forecasting models - SARIMA, Prophet and Random Forest - was performed, highlighting their respective strengths in short-term and seasonal forecasting.
- The study provides insights into the applicability of predictive models for regional traffic management in Turkey, demonstrating the utility of both statistical and machine learning methods.
- This work serves as a foundation for future research, including the development of hybrid models or deep learning approaches for more accurate and scalable traffic forecasting.

II. Method

A. Dataset

1) Story/Overview: The data set used in this study was obtained from the open data portal of Istanbul Metropolitan Municipality (https://data.ibb.gov.tr) The data includes hourly traffic information such as vehicle density, average speed, maximum speed and minimum speed for one month (September 2024) in different parts of Istanbul. The data is collected from traffic sensors in geographically defined areas (in geohash format).

The focal point of the analysis is the Beşiktaş coastal area. This region was chosen because it has a high traffic density in general and the majority of the geohash covers roads. In order to analyze the different patterns in traffic density more clearly, data from this region was studied.

2) Attributes: The data set includes the following columns:

- DATE TIME: The hourly time zone in which the traffic measurement was made.
- GEOHASH: The geographic code of the region where the data was collected. In this study only the geohash sxk9kp was selected.
- NUMBER OF VEHICLES: the number of vehicles recorded in the corresponding time period.
- AVERAGE SPEED: average speed (in km/h) of the vehicles registered in the corresponding time period.
- LONGITUDE and LATITUDE: Geographical coordinates of the area where the data was collected.
- MINIMUM SPEED and MAXIMUM SPEED: the minimum and maximum speed, respectively, of the vehicles recorded in the corresponding time period.

Some columns in the dataset (e.g. MINIMUM SPEED, MAXIMUM SPEED, LONGITUDE and LATITUDE) were removed as they were unnecessary for the analysis.

The AVERAGE SPEED column was excluded from most analyses to prioritize vehicle count, as its inclusion offered limited additional value; the observed relationship between average speed and vehicle density was weak suggesting that vehicle count alone better aligns with the study's focus.

Only the following columns were used in the analysis:

- DATE TIME
- GEOHASH
- NUMBER OF VEHICLES

3) Data Cleaning and Preprocessing: The dataset included traffic data from across Istanbul and in its initial version contained around 1.7 million records and 2460 unique geohashes. However, for the analysis, we focused on a specific region (Beşiktaş coastal area) and followed the steps below:

1) Zone Filtering:
   - The data was selected only from the Besiktas coastal region corresponding to geohash code sxk9kp.
   - After this process, the dataset was reduced to 720 records.
2) Removal of Unnecessary Columns:
   - The MINIMUM SPEED, MAXIMUM SPEED, LONGITUDE and LATITUDE columns are excluded as they are not directly used for the analysis.
3) Historical Filtering:

- Registration on September 1 (Monday) and September 30 (Sunday) in order to maintain the weekly order and for a clearer analysis has been removed.
- After this process, the total number of records was updated to 672.

4) Defining Time Zones:
   - For more detailed analysis, traffic time zones are categorized as follows:
     - Morning Peak Traffic: 07:00 - 09:00
     - Midday Dip Traffic: 09:00 - 16:00
     - Evening Peak Traffic: 16:00 - 19:00
     - Late Night or Early Morning Traffic: 19:00 - 07:00
   - These categories have been used to more clearly analyze the different patterns in traffic density and speed.

B. Method for Answering Reseach Questions

1) Correlation Between Average Speed and Vehicle Count: In this analysis, OLS regression model is used to examine the relationship between average speed (AVERAGE_SPEED) and number of vehicles (NUMBER_OF_VEHICLES). The purpose of the model is to evaluate the effect of the increase in the number of vehicles on average speed. The regression analysis model is constructed as follows:

$$\text{AVERAGE\_SPEED} = \beta_0 + \beta_1 \times \text{NUMBER\_OF\_VEHICLES} + \epsilon \quad (1)$$

In this model:

1)
- β_0 represents the constant term.
- β_1 represents the effect of the number of vehicles on the average speed.
- ϵ represents the error term.

2) Using the SARIMA model: The SARIMA (Seasonal Autoregressive Integrated Moving Average) model is used to model *trend* and *seasonality* in time series analysis. In this study, the SARIMA model is preferred to analyse the daily (24-hour) seasonal cycle in traffic density data.

a) Testing the stationarity of time series: The Augmented Dickey-Fuller (ADF) test was applied to determine whether the time series is stationary. As a result of the ADF test, the series is stationary and there is no need for differencing.

b) Determining seasonality: Autocorrelation function (ACF) and partial autocorrelation function (PACF) plots were used to confirm the seasonal structure. As a result, a 24 hour periodic cycle was identified in the series.

c) Parameter selection: The parameters for the SARIMA model were determined on the basis of the ACF/PACF analyses and the AutoARIMA method:

- **ARIMA (p, d, q)**: (1, 0, 1)
- **Seasonal ARIMA (P, D, Q, S)**: (1, 0, 0, 24)

Where p is the AR component, d is the degree of differencing, q is the MA component, P, D, Q are the seasonal ARIMA parameters and S is the seasonal period (24 hours).

d) Model training: The SARIMA model was trained on the training portion of the dataset, and its performance was evaluated on the test dataset using the *rolling forecast* method. Forecasts were generated for the testing phase.

e) Residual Analysis: Residual analysis is used to evaluate the performance of the model. The randomness of the residuals and their conformity to the normal distribution were checked using the following tests and visualisations:

- Ljung-Box test (autocorrelation check of residuals)
- Histogram and Q-Q plot (conformity to normal distribution)

3) Using the Prophet model: Prophet is a flexible parametric model for the modelling of trend and seasonal effects in time series analysis. In this study, the Prophet model is used to analyse daily and weekly seasonal effects on traffic density and to make short-term forecasts. The flexible nature of the model allows efficient modelling of trend turning points and seasonal cycles.

a) Model Components: The Prophet model has three main components:

- **Trend**: The component that models the general trend of the data as linear (*linear*) or logistic (*logistic*).
- **Seasonality**: Covers daily, weekly and yearly cycles.
- **Residuals**: Random components that cannot be included in the model.

These components played a key role in Prophet's estimation of traffic density.

b) Parameter Selection and Fitting: All parameters used in the Prophet model have been carefully selected to best model the characteristics of the time series. The parameters and their settings are described in detail below:

c) Model Training: The Prophet model is trained on a **80%** part of the dataset. Daily and weekly seasonal effects are taken into account during model training. The model has been used to make 48 hour short-term forecasts on the test data.

d) Residual Analysis: Residual analysis was performed to assess model accuracy. The following methods were used to check whether the residuals were randomly distributed (white noise):

- **Result plots**: Histogram and Q-Q plots were generated to visualise the randomness of the forecast errors.
- **Normal Distribution Test**: Checked that the residuals were normal distributed.
- **growth**: The trend type *linear* is chosen because there is no capacity limit in the traffic data analysed.

- **changepoint_prior_scale**: This parameter, which controls the flexibility of the trend change points, is set to **0.001**. A lower value makes the model less sensitive to trend changes.
- **changepoint_range**: Specifies the percentage of the date range for which change points can be modelled. The default value **0.8** is used in this study.
- **seasonality_prior_scale**: Controls for the elasticity of seasonal effects. In this study we set **20.0** to allow the model to better capture seasonal variations.
- **daily_seasonality**: Daily seasonality is enabled (**True**).
- **weekly_seasonality**: Weekly seasonality is enabled
- **yearly_seasonality**: Annual seasonality is disabled (**False**) as the traffic data is for only one month.
- **seasonality_mode**: The effect of seasonality on the trend is selected as *additive*.
- **uncertainty_samples**: The number of samples used to assess the uncertainty of the forecast is set to **1000**.
- **interval_width**: The width of the prediction interval is set by default to **0.8** (80% confidence interval).

4) Using the Random Forest Model: Random Forest is a supervised learning algorithm used in machine learning, based on an ensemble of decision trees. Random Forest is suitable for both regression and classification problems. In this study, the Random Forest model is used to predict and classify traffic density.

a) Why Random Forest is preferred: Random Forest was chosen because of its ability to model complex and non-linear relationships. It is very good at understanding the dynamic nature of a variable such as the density of traffic. Also,

- Resistant to overfitting.
- Calculates the importance level of features, helping to understand the effect of variables.
- Provides a flexible structure with different hyperparameters.

b) Features used: The Random Forest model is applied to time-based traffic density data. The main features used to estimate traffic density are:

- **hour**: The hourly information is used to model daily variations in density.
- **dayofweek**: The day of the week is used to model density differences between weekdays and weekends.

c) Parameter selection and tuning: Hyperparameter optimisation was performed to improve the performance of the Random Forest model and to avoid overfitting. The following parameters and values have been used:

- **n_estimators**: Number of decision trees, set to 200 in this study.
- **max_depth**: Maximum depth of the trees. It is set to 10.
- **min_samples_split**: Minimum number of samples required to split a node. Set to 10.
- **min_samples_leaf**: The minimum number of samples required for leaf nodes. Set to 1.
- **random_state**: Used to control randomness. Set to 42.

Hyperparameter optimisation was performed using the **GridSearchCV** method. This method tries different parameter combinations and selects the settings that give the best results.

d) Training and testing the model: The Random Forest model was trained and tested with the following steps:

a) The data set is divided into **train** and **test** (80% training, 20% test).
b) The model is trained on the training data.
c) The prediction performance of the model is evaluated on the test data.

e) Feature importance: Random Forest provides *feature importance* calculations to measure the impact of the features used on the model. In this study, the degree of significance between the variables *hour* and *dayofweek* was found to be as follows:

- **hour** 77%
- **dayofweek**: 23%

These results suggest that the hour variable is more important in determining traffic density.

5) Performance metrics: The forecasting performance of all models is evaluated using the following metrics:

- **Mean Absolute Error (MAE):** Measures the average magnitude of errors in predictions, without considering their direction.
- **Root Mean Squared Error (RMSE):** Highlights larger errors by squaring them, providing a more sensitive metric to outliers.
- **R-squared ($\mathbf{R^2}$):** Indicates the proportion of variance in the dependent variable that is predictable from the independent variable(s).

These metrics were chosen to comprehensively evaluate both the accuracy and explanatory power of the forecasting models.

III. Results

A. Descriptive Analysis

1) Average Vehicle Count by the Day of the Week

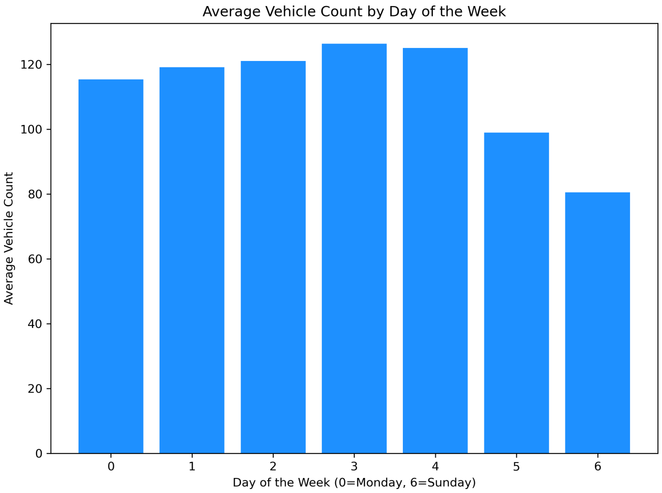

Fig. 1. Average Vehicle/Hour Count by Day of the Week

(Figure 1) analyzes the vehicle density by day of the week. According to the data, vehicle density is higher on weekdays compared to weekends. Especially Thursday stands out as the busiest day of the week. The average number of vehicles for this day is 126.35 vehicles/hour and the median value is 136 vehicles/hour. Thursday is followed by Friday with an average of 125.11 vehicles/hour. On weekends, vehicle density drops significantly, reaching its lowest levels of the week with an average of 98.94 vehicles/hour on Saturday and 80.52 vehicles/hour on Sunday.

2) Average Vehicle Count by Hour of the Day

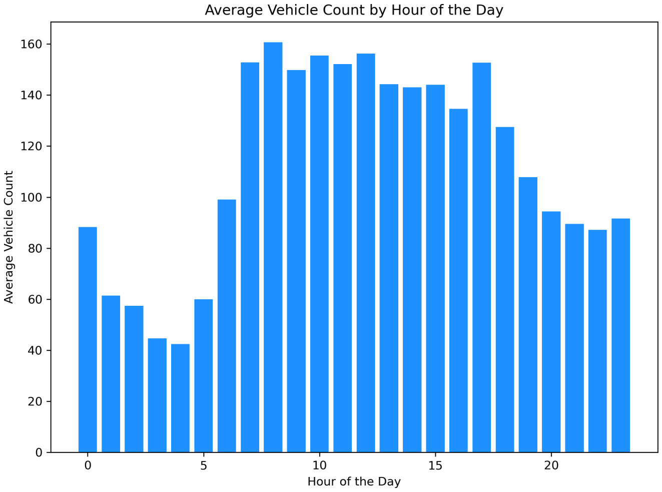

Fig. 2. Average Vehicle/Day Count by Hour of the Day

(Figure 2) shows vehicle density by time of day started at low levels at night and peaked in the morning. At midnight (00:00), the average number of vehicles was 88.32 vehicles/hour. At 08:00 in the morning, vehicle density reached its highest level with an average of 160.64 vehicles/hour.

In the afternoon, vehicle density stabilized, averaging 144 vehicles/hour around 15:00. In the evening, the density decreased again, reaching an average of 107.86 vehicles/hour at 19:00 and 91.57 vehicles/hour at 23:00.

3) Boxplot of Vehicle Count by Hour of the Day

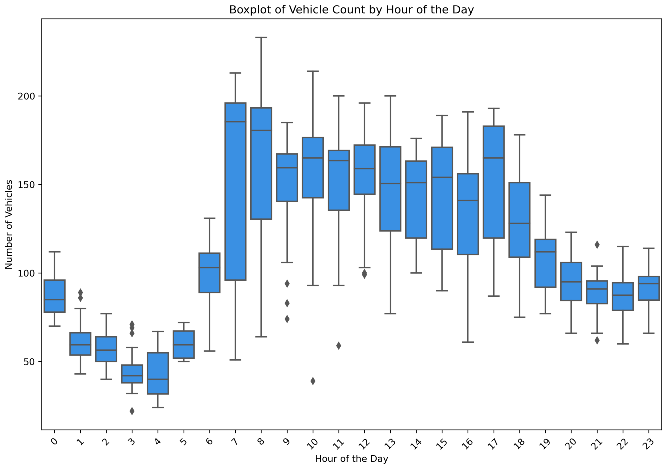

Fig. 3. Boxplot of Vehicle Count by Hour of the Day

(Figure 3) shows how vehicle density changes at different times of the day. The graph includes minimum, maximum, median and quartile values as well as outliers for each time zone. In the morning hours (06:00 - 09:00), vehicle density increases significantly. Especially between 07:00 and 08:00, the peak in density is recorded. At 07:00, the median number of vehicles is around 180 vehicles/hour, while the upper limit reaches 200-213 vehicles/hour. This indicates that the morning commute is the peak hour. However, there are a few outliers during these hours.

During the midday hours (10:00 - 16:00), congestion becomes more stable, with the median number of vehicles generally ranging between 150-165 vehicles/hour. During these hours, a few outliers below the lower limit are observed.

In the evening hours (17:00 - 19:00), the density increases again, reaching a level close to the morning peak at 17:00, with a median value of 165 vehicles/hour. However, at 18:00, the density drops slightly, with a median of 128 vehicles/hour, indicating some variability during this period. This variability indicates that heavy traffic is more unpredictable in the evening hours. In addition, some outliers below the lower bound at 19:00 indicate that traffic is less intense in certain areas than expected.

During the night hours (21:00 - 23:00), vehicle density is generally low, with a median value ranging between 50-90 vehicles/hour. Traffic is more consistent during these hours. However, a few outliers were also observed during the night. For example, at 22:00 and 23:00 these values are below the lower limit, indicating that there are fewer vehicles in traffic than expected.

This graph shows that vehicle density is higher during the morning and evening peak hours and that this density changes in a predicable manner. However, outliers at some times of the day indicate that the density may vary in unusual circumstances. This is important data to be taken into account in the planning and management of traffic flow.

4) Boxplot of Vehicle Count by Day of the Week

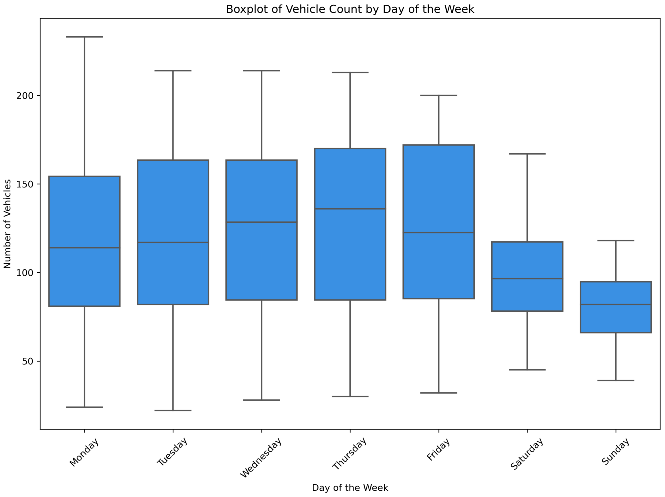

Fig. 4. Boxplot of Vehicle Count by Day of the Week

(Figure 4) shows the distribution of vehicle density by day of the week. For each day, the minimum, maximum, median, and quartile values, as well as outliers, are visualized. It is noticeable that on weekdays (Monday-Friday), vehicle density is generally higher and distributed more widely. Thursday stands out as the busiest day, with a median value of 136 vehicles/hour. Friday follows with a median value of 122.5 vehicles/hour, indicating it is the second-busiest day of the week.

On weekend days (Saturday and Sunday), vehicle density drops significantly. On Saturday, the median value is 96.5 vehicles/hour, while on Sunday it drops to 82.0 vehicles/hour. This shows that work and school traffic on weekdays gives way to a lower density on weekends. Outliers are also noticeable. For the weekend, such values are lower, and the boxplot shows a narrower range, indicating a less variable traffic density.

This boxplot clearly shows the difference in vehicle density between weekdays and weekends. While weekdays exhibit a wider distribution of traffic density, weekend traffic is more regular and low density. In particular, Thursday shows the peak in terms of density, followed by Friday, while Sunday shows the lowest level of traffic.

5) Total Vehicle Count Per Day

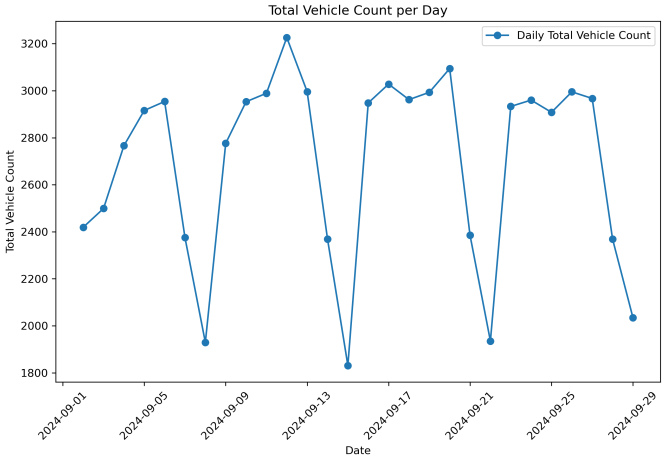

Fig. 5. Total Vehicle Count per Day

(Figure 5) shows the daily variation of the total number of vehicles per day. When the graph is analyzed, it is noticeable that there is a weekly cycle. It is observed that the total number of vehicles increases at the beginning of each week and decreases at the end of the week.

During weekdays, the total number of vehicles generally stays around 3000 vehicles/day. By the middle of the week (especially on Wednesdays and Thursdays), the vehicle density peaks. However, the graph shows that the total number of vehicles steadily decreases each weekend, dropping to around 2000 vehicles/day.

This cycle illustrates how the daily vehicle density changes between weekdays and weekends. There is a steady increase during each week followed by a significant decrease at the end of the week.

6) Vehicle Count on Weekdays vs Weekends

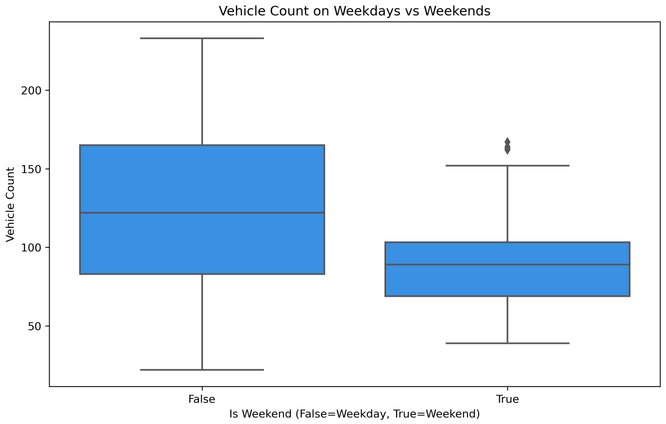

Fig. 6. Vehicle Count on Weekdays vs Weekends

(Figure 6) compares the vehicle density on weekdays and weekend days using boxplot. The graph clearly shows that vehicle density is higher on weekdays than on weekends. The number of vehicles on weekdays (Monday-Friday) has a wider distribution, with a median value of 122 vehicles/hour and a generally higher range compared to weekends. The maximum values reach 233 vehicles/hour, while the minimum values are 22 vehicles/hour.

On weekend days (Saturday-Sunday), vehicle density is lower and more stable. The median value is 89 vehicles/hour, and the maximum value is 167 vehicles/hour. However, a few outliers were observed for the weekend, indicating that in extreme cases, weekend traffic density may be higher than expected.

On weekend days (Saturday-Sunday), on the other hand, vehicle density is lower and more stable. The median value is around 100 vehicles/hour and the maximum value is around 150 vehicles/hour. However, a few outliers were observed for the weekend. These outliers indicate that in extreme cases, weekend traffic density may be higher than expected.

This boxplot clearly shows the difference in vehicle density between weekdays and weekends. On weekdays, vehicle density is generally higher and more variable, whereas on weekends, a lower and more stable traffic density is observed.

7) Boxplot of Vehicle Count by Time of the Day

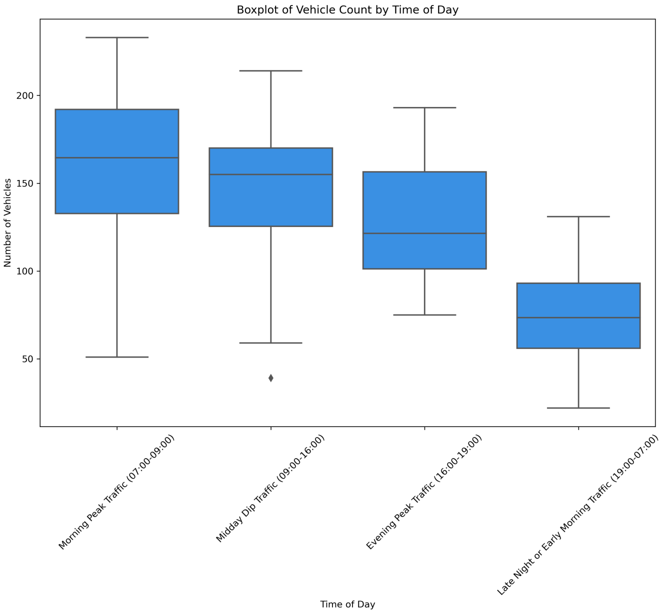

Fig. 7. Boxplot of Vehicle Count by Time of Day

The graph (Figure 7) shows how vehicle density changes during certain time periods of the day. This graph was created based on the peak hour intervals identified in the previous analysis (morning rush hour, midday fall, evening rush hour and night/early morning). The aim is to verify whether these time periods really have their own specific density characteristics.

The morning peak traffic (07:00-09:00) is characterized by the highest vehicle density. The median number of vehicles is 184 vehicles/hour, and the maximum density reaches 233 vehicles/hour. However, the lower limit of congestion in the morning hours is around 51 vehicles/hour, indicating that congestion decreases in certain areas or in the early hours.

Midday (09:00-16:00) is the time of day when congestion decreases significantly compared to the morning and evening peaks. The median value during this period is 158 vehicles/hour. The maximum density reaches 214 vehicles/hour, while the lower limit is 39 vehicles/hour. This decrease shows the calm period between morning and evening traffic.

The evening peak traffic (16:00-19:00) is similar to the morning peak in terms of density levels. The median value is 142 vehicles/hour, and the maximum density is 193 vehicles/hour. Evening traffic is more widely distributed than morning traffic.

The night and early morning hours (19:00-07:00) are the times of lowest traffic density. During this period, the median value is 78 vehicles/hour, and the maximum density is 144 vehicles/hour. The lower limit is recorded at 22 vehicles/hour. These results confirm that the night and early morning hours are generally quiet in terms of traffic.

This graph supports the peak hour patterns found in the previous analyses. It is clear that vehicle density peaks in the morning and evening hours, while decreasing in the midday hours. In addition, the night and early morning hours stand out as a period of low traffic density and less variability.

8) Heatmap of Vehicle Count by Hour and Day

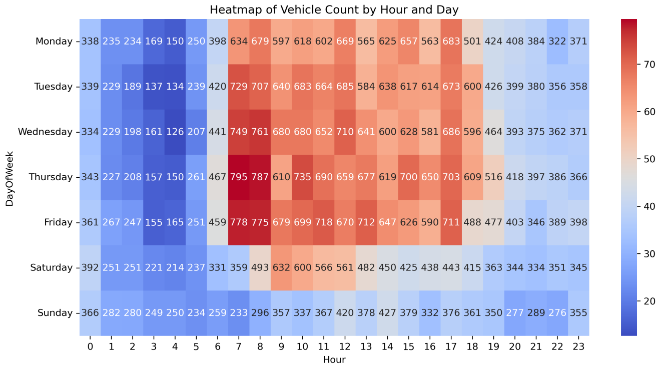

Fig. 8. Heatmap of Vehicle Count by Hour and Day

The graph (Figure 8) shows the vehicle density by day of the week and time of day in the form of a heatmap. The map provides a powerful tool to visualize the hours and days during which vehicle density increases or decreases. Density is represented by red tones for higher levels and blue tones for lower levels.

According to the graph, on weekdays (Monday-Friday), vehicle density peaks between 07:00-09:00 in the morning and between 16:00-18:00 in the evening. On Thursday at 07:00 in the morning, the vehicle density is highest at approximately 795 vehicles/hour. Similarly, the density also increases in the evening hours, reaching a peak of 703 vehicles/hour at 17:00 on Thursday. This clearly shows that the traffic intensifies during weekdays due to work and school hours.

On weekends (Saturday and Sunday), traffic density drops significantly compared to weekdays. On Saturday, there is a significant congestion during the morning hours (i.e. 09:00-12:00), but this congestion is low compared to the weekday peak. On Sunday, congestion remains relatively constant and low during the day. This suggests that weekends are characterized by lower traffic volumes and a regular traffic flow.

This analysis confirms that vehicle density is focused on the morning and evening peaks on weekdays, indicating a more consistent and lower traffic density on weekends. It also reveals that traffic density in the morning and evening hours may vary on different days of the week.

9) OLS Regression Analysis Results

Ordinary Least Squares (OLS) regression analysis was applied to examine the relationship between vehicle density (NUMBER_OF_VEHICLES) and average speed (AVERAGE_SPEED). This analysis was performed to explain the inverse relationship between speed and vehicle density. The general statistics of the regression model are presented below:

- **R-squared (R^2)**: 0.195
- **F-statistic**: 162.4
- **p-value (F-statistic)**: 1.87e-33
- **Intercept (constant)**: 27.37
- **Slope (NUMBER_OF_VEHICLES)**: -0.0498

Interpretation of Results According to the regression analysis results, there is a statistically significant decrease in average speed as vehicle density increases (*p ¡ 0.05*). However, the model can explain only 19.5% of the variance in speed (R^2 *= 0.195*), suggesting that other factors likely contribute to variations in average speed beyond vehicle density.

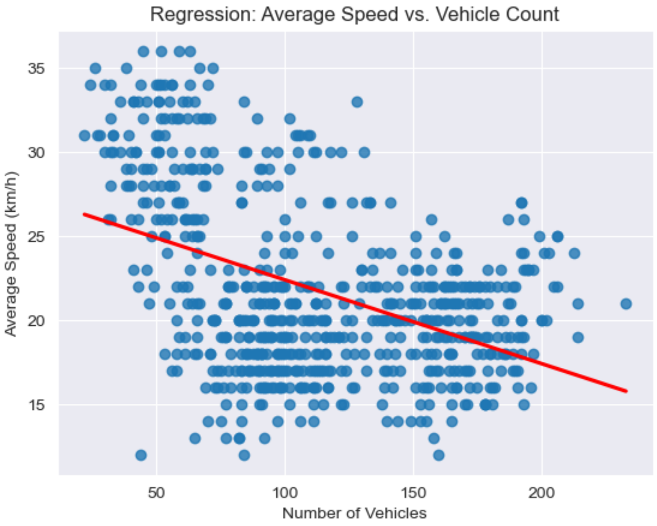

Fig. 9. OLS Regression: Average Speed vs Vehicle Count

Constraints and Evaluation: This analysis examines the direct relationship between speed and density without impacting the predictive performance of time series models. While an inverse relationship is observed, the low explanatory power suggests other factors significantly influence average speed, highlighting the need for further investigation.

B. Analysis for answering research questions

This section compares the performance results of SARIMA, Prophet and Random Forest models. The predictive performance of each model is supported by performance metrics and visualisations. These analyses have been calculated on the basis of the results of SARIMA, Prophet and Random Forest models.

1) Results of SARIMA:

a) Prediction Results: The SARIMA model is tested using the *rolling forecast* method and the forecasting performance is determined according to the dynamic structure of the time series. Based on the results of the residual analysis, the errors of the model are randomly distributed, which confirms that the model is reliable.

- **Mean Absolute Error (MAE)**: 15.81
- **Root Mean Squared Error (RMSE)**: 20.44
- **R-squared (R^2)**: 0.80

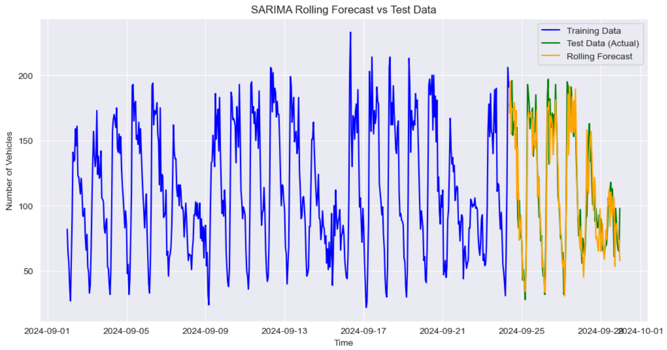

Fig. 10. SARIMA Rolling Forecast vs Test Data

b) Residual Analysis: Figure 11 shows that the SARIMA model's residuals are centered around zero, indicating no significant bias in the model. However, the variability in residuals suggests that some patterns in the data may remain unexplained.

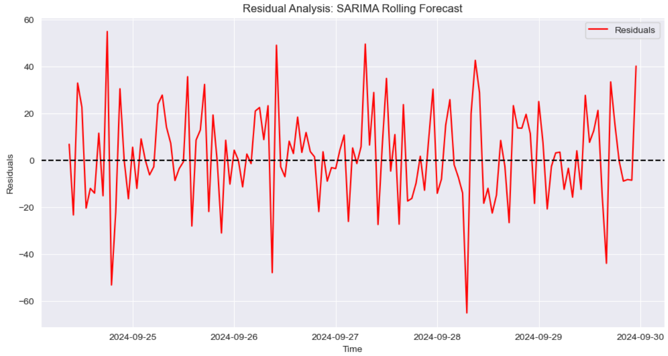

Fig. 11. SARIMA Residual Analysis

(Figure 12) shows that the residuals are approximately symmetrically distributed around zero, suggesting a reasonable model fit. However, the slight deviation from a perfect bell shape indicates some non-normality in the residuals.

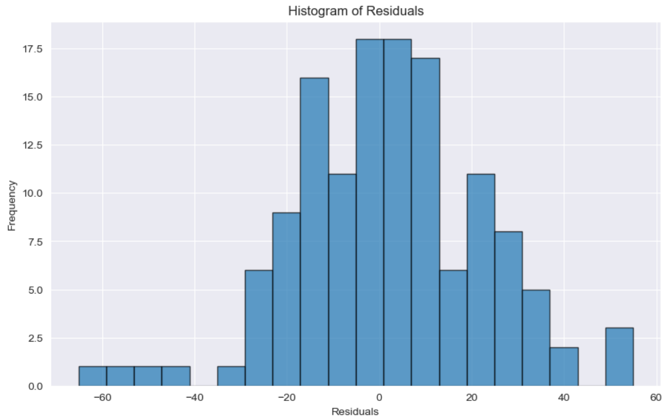

Fig. 12. Histogram of SARIMA Residuals

(Figure 13) demonstrates that the residuals align closely with the diagonal line, indicating approximate normality. Minor deviations in the tails suggest the presence of some outliers or unexplained variability.

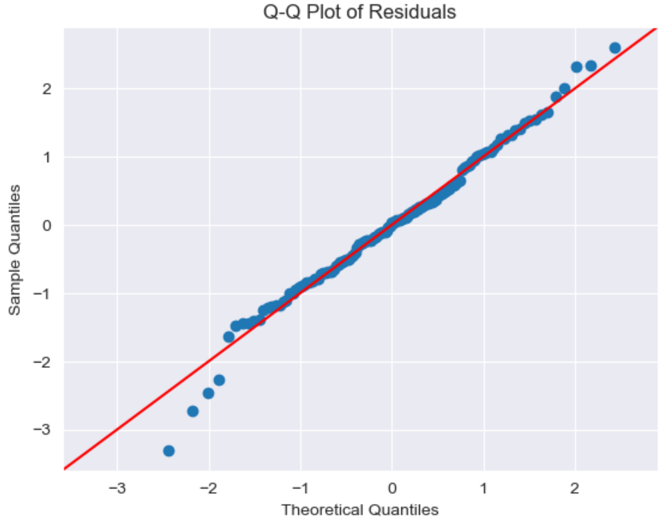

Fig. 13. Q-Q Plot of SARIMA Residuals

2) Results of the Prophet Model: The Prophet model is applied to model daily and weekly seasonal effects. The forecasting performance of the model is measured by the following metrics:

- **Mean Absolute Error (MAE)**: 18.91
- **Root Mean Squared Error (RMSE)**: 23.06
- **R-squared ($\mathbf{R^2}$)**: 0.75

a) Prediction Results: The Prophet model demonstrated a similar accuracy to the SARIMA model in short-term (48 hours) forecasts. However, the uncertainty ranges widened in long-term forecasts.

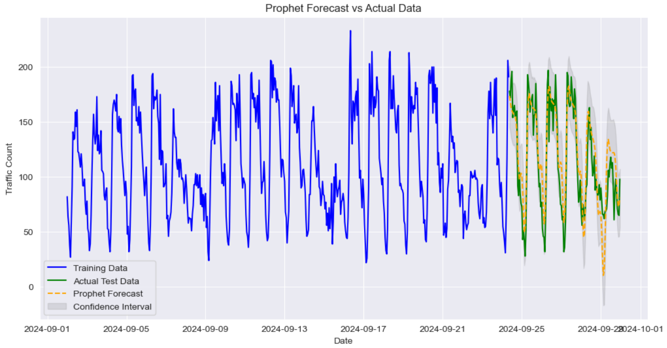

Fig. 14. Prophet Forecast vs Actual Data

b) Residual Analysis: (Figure 15) shows that the residuals are centered around zero, indicating no systematic bias in the Prophet model's predictions. However, some variability in the residuals suggests that not all patterns in the data are captured.

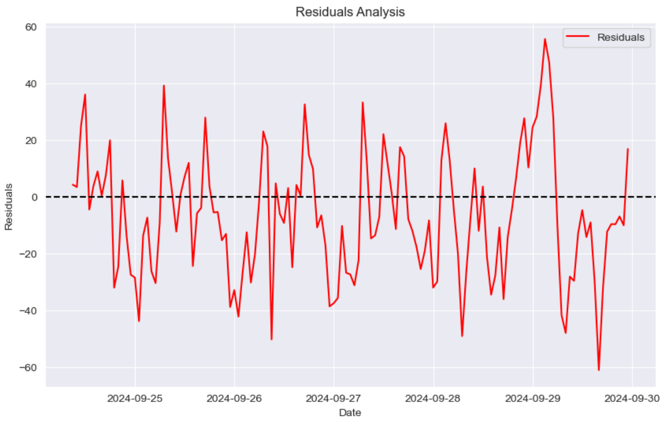

Fig. 15. Prophet Residual Analysis

(Figure 16) The histogram displays a roughly symmetric distribution of residuals around zero, suggesting a reasonable model fit. However, the distribution deviates slightly from perfect normality, as evidenced by minor skewness in the tails.

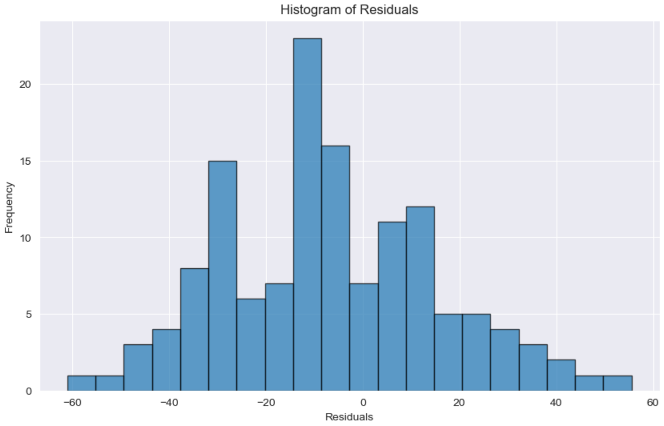

Fig. 16. Histogram of Prophet Residuals

(Figure 17) demonstrates that the residuals follow the diagonal line closely, indicating that the residuals are approximately normally distributed. Some deviations are observed in the upper tail, suggesting potential outliers or unexplained variance in predictions.

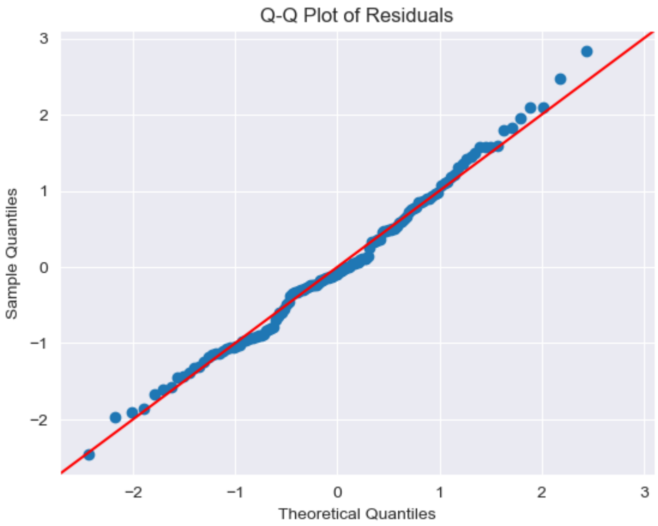

Fig. 17. Q-Q Plot of Prophet Residuals

3) Results of Random Forest: The Random Forest model was applied to predict vehicle density and model hourly variations. The performance results of the model are as the below:

- **Mean Absolute Error (MAE)**: 11.54
- **Root Mean Squared Error (RMSE)**: 15.80
- **R-squared (R^2)**: 0.86

a) Feature Importance: The importance level of *hour* and *dayofweek* variables used in the model was calculated. Time information (*hour*) was found to be the most effective variable in determining traffic density.

b) Feature Importance : The *feature importance* values of the Random Forest model were calculated and the results were found as below:

- **hour**: 77%
- **dayofweek**: 23%

4) Comparative Analysis: The forecasting performances of SARIMA, Prophet and Random Forest models are summarized in the table below:

TABLE I
MODEL PERFORMANCE COMPARISON

Model	**MAE**	**RMSE**	**R^2**
SARIMA	15.81	20.44	0.80
Prophet	18.91	23.06	0.75
Random Forest	11.54	15.80	0.86

From this comparison, the Random Forest model performed the best in short-term forecasting, providing the lowest MAE and RMSE values. However, the SARIMA model showed strong accuracy in modeling seasonal cycles and the Prophet model handled seasonal effects with similar accuracy.

IV. Conclusion

In this study, several statistical and machine learning methods were used to analyse traffic density and its variation by time and day for a specific area of Istanbul. The data set consisted of hourly vehicle counts over a period of one month. Below are the main findings and results of the study:

A. Summary of Results

The descriptive analysis shows that traffic density varies significantly according to the time of day and day of the week:

- Vehicle density peaks in the morning (07:00-09:00) and in the evening (16:00-19:00), confirming the impact of rush hours.
- Traffic density is significantly higher on weekdays than on weekends. Thursday is the busiest day and Sunday the least congested.
- There is evidence of an inverse relationship between vehicle density and average speed, as indicated by OLS regression analysis. However, the low explanatory power($R^2 = 0.195$) suggests that this relationship is weak and other factors may play a significant role in influencing average speed.

The performance comparison of the three forecasting models showed the following results:

- **Random Forest** achieved the lowest MAE (11.54), RMSE (15.80) and R^2 (0.86), making it most suitable for short-term traffic prediction.
- The **SARIMA** and **Prophet** models were more suitable for long term trend analysis as they effectively captured the seasonal patterns in traffic density. SARIMA achieved an R^2 score of 0.80, while Prophet achieved 0.75.

Residual analysis for the SARIMA and Prophet models confirmed that the errors were randomly distributed, indicating no major biases in the predictions of these models.

B. Future Work

While the current study provides valuable insights into analysing and predicting traffic density, there are several areas for further improvement:

- **Inclusion of additional features:**Including weather conditions, holidays, and roadwork information could enhance prediction accuracy.
- **Model generalisation:** Extension of the analysis to other regions and time periods would be helpful in the assessment of the robustness and generalisability of the models.
- **Deep Learning Models:** Research into advanced techniques such as long term memory (LSTM) and convolutional neural networks (CNN) could further improve forecasting performance.
- **Real-time forecasting:** Integrating the models into a real-time traffic management system could provide actionable insights for urban planning.

Overall, this study highlights the importance of understanding traffic patterns in order to effectively manage traffic in urban areas. The findings can be used as a basis for developing future traffic forecasting systems, contributing to more sustainable and efficient traffic management strategies.

References

[1] S. Ghosh, B. Basu, and M. O'Mahony, "Bayesian Time-Series Model for Short-Term Traffic Flow Forecasting," *ASCE Journal of Transportation Engineering*, vol. 133, no. 3, pp. 180–189, 2007.

[2] H. Liu and J. Wu, "Prediction of Road Traffic Congestion Based on Random Forest," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 2, pp. 377–385, 2017.

[3] M. Zarei, R. Zarei, and A. Sattari, "Road Traffic Prediction Using Context-Aware Random Forest Based on Volatility Nature of Traffic Flows," in *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*, 2013, pp. 175–180.

[4] P. Kumar and L. Vanajakshi, "Short-Term Traffic Flow Prediction Using Seasonal ARIMA Model," *IEEE Intelligent Transportation Systems Magazine*, vol. 7, no. 2, pp. 45–55, 2015.

[5] X. Han and X. Shi, "Online Traffic Congestion Prediction Based on Random Forest," in *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*, 2015, pp. 102–107.

[6] X. Wang and Y. Liu, "Traffic Volume Prediction on Busy Road Junctions," *Transportation Research Part C: Emerging Technologies*, vol. 95, pp. 21–36, 2018.

[7] W. Hong, "Traffic Flow Forecasting by Seasonal SVR with Chaotic Simulated Annealing Algorithm," *International Journal of Computational Intelligence Systems*, vol. 4, no. 4, pp. 568–576, 2011.

[8] Istanbul Metropolitan Municipality, "Hourly Traffic Density Data Set," available at: https://data.ibb.gov.tr/dataset/hourly-traffic-density-data-set/resource/914cb0b9-d941-4408-98eb-f378519c26f4Istanbul Metropolitan Municipality Data Portal, [Accessed: Month Day, Year].