REPUBLIC OF TURKEY
YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

TRAFFIC FLOW SPEED ESTIMATION BASED ON MULTI-FEATURE SPACES

Enes BİLGİN

MASTER OF SCIENCE THESIS
Department of Computer Engineering
Program of Computer Engineering

Supervisor
Assoc. Prof. M. Amaç GÜVENSAN

Co-supervisor
Assist. Prof. H. İrem TÜRKMEN

August, 2023

REPUBLIC OF TURKEY
YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

TRAFFIC FLOW SPEED ESTIMATION BASED ON MULTI-FEATURE SPACES

A thesis submitted by Enes BİLGİN in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE** is approved by the committee on 03.08.2023 in Department of Computer Engineering, Program of Computer Engineering.

Assoc. Prof. M. Amaç GÜVENSAN
Yildiz Technical University
Supervisor

Assist. Prof. H. İrem TÜRKMEN
Yildiz Technical University
Co-supervisor

Approved By the Examining Committee

Assoc. Prof. M. Amaç GÜVENSAN, Supervisor
Yildiz Technical University _____

Assist. Prof. Göksel BİRİCİK, Member
Yildiz Technical University _____

Prof. Dr. Berk CANBERK, Member
Istanbul Technical University _____

I hereby declare that I have obtained the required legal permissions during data collection and exploitation procedures, that I have made the in-text citations and cited the references properly, that I haven't falsified and/or fabricated research data and results of the study and that I have abided by the principles of the scientific research and ethics during my Thesis Study under the title of Traffic Flow Speed Estimation Based on Multi-Feature Spaces supervised by my supervisor, Assoc. Prof. M. Amaç GÜVENSAN. In the case of a discovery of false statement, I am to acknowledge any legal consequence.

Enes BİLGİN

Signature

This work was supported by the Turkish Scientific and Technological Research Council of Turkey (TUBITAK) under the grant number TUBITAK1001-120E357.

Dedicated to my family

ACKNOWLEDGEMENTS

I want to express my endless thanks to my esteemed advisors, M. Amaç GÜVENSAN and H. İrem TÜRKMEN, who provided their support, knowledge, and valuable ideas throughout the development process of my thesis.

I want to express my endless thanks to my family, who have always been by my side and have contributed significantly to completing my thesis. I would also like to thank my dear friends who guided me throughout my thesis work. I want to thank my brother Burak BİLGİN for his refreshing perspective. Without them, this would not have been possible.

Enes BİLGİN

TABLE OF CONTENTS

LIST OF SYMBOLS viii

LIST OF ABBREVIATIONS ix

LIST OF FIGURES xi

LIST OF TABLES xiv

ABSTRACT xv

ÖZET xvii

1 INTRODUCTION 1

1.1 Purpose of the Thesis . . . 3

1.2 Hypotheses . . . 3

2 LITERATURE REVIEW 5

2.1 Traffic Speed Estimation Problem . . . 5

2.2 The usage of Multi-Feature Spaces . . . 8

3 DATASET 12

3.1 Traffic Speed Data . . . 12

3.2 Spatial Road Connection Data . . . 14

3.3 Accidents and Events Data . . . 16

3.4 Weather Data . . . 16

4 FACTORS AFFECTING TRAFFIC SPEED 18

4.1 Temporal Factors . . . 18

4.2 Meteorological Factors . . . 20

4.3 Spatial Factors . . . 25

4.4 Accidents & Events . . . 29

4.5 Statistical Analysis of Traffic Speed . . . 29

5 SYSTEM DESIGN 33

5.1 System Architecture . . . 33

5.2 Long-Term Traffic Speed Prediction Algorithms 36
5.3 Base Prediction Model (L0) Features 39
5.4 Multi-Feature Space Prediction Network 42
5.5 Post-Prediction Improvements (L3+) 51

6 DEVELOPMENT ENVIRONMENT 53
6.1 Software Resources 53
6.2 Hardware Resources 54
6.3 Software Development Workflows 54

7 EXPERIMENTAL RESULTS 57
7.1 Error Metrics 57
7.2 Base Model Performances 58
7.3 Multi-Feature Space Prediction Model Experiments 62
7.4 Powerful (All-in-one) Model Experiments 81
7.5 Post-Prediction Improvement Results 89
7.6 Time Cost of Traffic Speed Estimation 95

8 CONCLUSION 98

REFERENCES 101

PUBLICATIONS FROM THE THESIS 104

LIST OF SYMBOLS

°C	Celsius
d_i	Coefficient Effect Direction
t_{max}	Maximum Value of Weather Condition
x	Raw Traffic Speed
e_s	Static Effect Rate of Weather Condition
w_t	Weather Condition at Training Day
w_f	Weather Condition at Forecast Time

LIST OF ABBREVIATIONS

AIC	Akaike Information Criterion
ANN	Artificial Neural Network
ARIMA	Auto Regressive Integrated Moving Average
BIC	Bayesian Information Criterion
BPNN	Backpropagation Neural Network
CNN	Convolutional Neural Network
DBN	Deep Belief Network
DOM	Day of Month
DOW	Day of Week
GLT	Greedy Layerwise Training
GPS	Global Positioning System
GRU	Gated Recurrent Unit
HOD	Hour of Day
IBB	Istanbul Buyuksehir Belediyesi (Istanbul Metropolitan Municipality)
LASSO	Least Absolute Shrinkage and Selection Operator
LSTM	Long Short Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MOH	Month of Hour
MOY	Month of Year
RBM	Restricted Boltzmann Machine
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network

RTMS	Remote Traffic Microwave Sensor
SAE	Stacked AutoEncoder
StoS	Square to Sequence Learning
SVR	Support Vector Regression
WOY	Week of Year

LIST OF FIGURES

Figure 1.1 Short-Term Estimation vs. Long-Term Estimation 1
Figure 3.1 RTMS Sensor Instance in Istanbul 12
Figure 3.2 Distribution of 106 Segments over Istanbul 13
Figure 3.3 Google Earth Footage of KMZ File 15
Figure 3.4 Segment Connection Tool (SCT) 15
Figure 4.1 HOD Traffic Speed Analysis 19
Figure 4.2 DOW Traffic Speed Analysis 20
Figure 4.3 MOY Traffic Speed Analysis 20
Figure 4.4 Seasonal Traffic Speed Analysis 21
Figure 4.5 Temperature Effect on Traffic Speed 22
Figure 4.6 Rainfall Effect on Traffic Speed 23
Figure 4.7 Wind Effect on Traffic Speed 23
Figure 4.8 Cloudiness Effect on Traffic Speed 24
Figure 4.9 Visibility Effect on Traffic Speed 25
Figure 4.10 Intersection Speeds Example: Zincirlikuyu 27
Figure 4.11 Intersection Example on Map: Zincirlikuyu 28
Figure 4.12 Hop Distance of Road Segments 28
Figure 4.13 Most Difficult Road Zones to Predict 29
Figure 5.1 System Architecture 35
Figure 5.2 Convolutional Neural Network Layers 37
Figure 5.3 General LSTM Architecture 38
Figure 5.4 Time Data Embedding Schema 40
Figure 5.5 The Overall Aspect of Speed Data Usage Timeline 41
Figure 5.6 Speed Data Input Schema with Sample Speeds 41
Figure 5.7 GRU Network Architecture 42
Figure 5.8 Example Speed Pre-tuning Process - 1 45
Figure 5.9 Example Speed Pre-tuning Process - 2 45
Figure 5.10 Polynomial Regression Models of Rainfall Data with Different Degrees 47
Figure 5.11 Polynomial Regression Models of Temperature Data with Different Degrees 47

Figure 5.12 Polynomial Regression Models of Wind Speed Data with Different Degrees 48
Figure 5.13 Spatial Information Feeding Methods 49
Figure 5.14 Spatial Information Grid Distributed Input Feeding Schema . . . 50
Figure 7.1 MAE: Mean Absolute Error 57
Figure 7.2 MAPE: Mean Absolute Percentage Error 58
Figure 7.3 Base Models HOD Comparison (MAPE) 59
Figure 7.4 Base Models HOD Comparison (MAE) 60
Figure 7.5 Base Models DOW Comparison (MAPE) 60
Figure 7.6 Base Models DOW Comparison (MAE) 61
Figure 7.7 Base Models MOY Comparison (MAPE) 61
Figure 7.8 Base Models MOY Comparison (MAE) 62
Figure 7.9 Temperature-Assisted Model Performance 63
Figure 7.10 Temperature-Assisted Models Performance over Segments . . . 64
Figure 7.11 Top 10 Most Affected Segments - Temperature-Assisted Model 64
Figure 7.12 Rainfall-Assisted Model Performance 65
Figure 7.13 Rainfall-Assisted Models Performance over Segments 66
Figure 7.14 Top 10 Most Affected Segments - Rainfall-Assisted Model . . . 66
Figure 7.15 Wind-Speed-Assisted Model Performance 67
Figure 7.16 Wind-Speed-Assisted Models Performance over Segments . . . 68
Figure 7.17 Top 10 Most Affected Segments - Wind-Speed-Assisted Model . 68
Figure 7.18 Clouds-Assisted Model Performance 69
Figure 7.19 Visibility-Assisted Model Performance 69
Figure 7.20 Weather-Hybrid-Assisted Models HOD Improvement (MAPE) . 71
Figure 7.21 Weather-Hybrid-Assisted Models HOD Improvement (MAE) . . 71
Figure 7.22 Weather-Hybrid-Assisted Models DOW Improvement (MAPE) . 72
Figure 7.23 Weather-Hybrid-Assisted Models DOW Improvement (MAE) . 72
Figure 7.24 Weather-Hybrid-Assisted Models MOY Improvement (MAPE) . 73
Figure 7.25 Weather-Hybrid-Assisted Models MOY Improvement (MAE) . 73
Figure 7.26 Spatial Models HOD Improvement (MAPE) 74
Figure 7.27 Spatial Models HOD Improvement (MAE) 75
Figure 7.28 Spatial Models DOW Improvement (MAPE) 75
Figure 7.29 Spatial Models DOW Improvement (MAE) 76
Figure 7.30 Spatial Models MOY Improvement (MAPE) 76
Figure 7.31 Spatial Models MOY Improvement (MAE) 77
Figure 7.32 Spatial Best Fars Map 77
Figure 7.33 Speed-Groups-Assisted Model HOD Improvement (MAPE) . . 78
Figure 7.34 Speed-Groups-Assisted Model HOD Improvement (MAE) . . . 79
Figure 7.35 Speed-Groups-Assisted Model DOW Improvement (MAPE) . . 79

Figure 7.36 Speed-Groups-Assisted Model DOW Improvement (MAE) . . . 80
Figure 7.37 Speed-Groups-Assisted Model MOY Improvement (MAPE) . . 80
Figure 7.38 Speed-Groups-Assisted Model MOY Improvement (MAE) . . . 81
Figure 7.39 Improvement Distribution of Models over 106 Segments 82
Figure 7.40 Powerful Model HOD Improvements (MAPE) 83
Figure 7.41 Powerful Model HOD Improvements (MAE) 84
Figure 7.42 Powerful Model DOW Improvements (MAPE) 84
Figure 7.43 Powerful Model DOW Improvements (MAE) 85
Figure 7.44 Powerful Model MOY Improvements (MAPE) 85
Figure 7.45 Powerful Model MOY Improvements (MAE) 86
Figure 7.46 Legal Speed Limits in Istanbul 91
Figure 7.47 Powerful+ Model HOD Improvements (MAPE) 92
Figure 7.48 Powerful+ Model HOD Improvements (MAE) 92
Figure 7.49 Powerful+ Model DOW Improvements (MAPE) 93
Figure 7.50 Powerful+ Model DOW Improvements (MAE) 93
Figure 7.51 Powerful+ Model MOY Improvements (MAPE) 94
Figure 7.52 Powerful+ Model MOY Improvements (MAE) 94
Figure 7.53 Time Cost of 1 Segment's 1 Month Model 96

LIST OF TABLES

Table 3.1 Formatted Traffic Speeds 13
Table 3.2 Traffic Speeds After Processing 14
Table 3.3 Accident and Event Types 16
Table 3.4 Weather Conditions Statistics 17
Table 4.1 Abbreviations of Time Periods 18
Table 4.2 106 Segments Statistics and Locations 30
Table 5.1 Prediction Network Default Parameters 42
Table 5.2 Speed Group Bands 43
Table 5.3 Symbol Definitions 45
Table 6.1 Software Preferences 53
Table 6.2 Python Libraries 54
Table 6.3 Hardware Specifications - MSIGP76 54
Table 6.4 Hardware Specifications - HP Z4 Workstation 55
Table 7.1 Base Model Results 58
Table 7.2 Prediction Models Contribution 82
Table 7.3 Final Accuracy Improvements of Segments 86
Table 7.4 Prediction Models Time Costs for 1 Road Segment's 12 Months with GRU 95
Table 7.5 Time Costs of Base Prediction Models Time Costs for 1 Road Segment 96

ABSTRACT

Traffic Flow Speed Estimation Based on Multi-Feature Spaces

Enes BİLGİN

Department of Computer Engineering
Master of Science Thesis

Supervisor: Assoc. Prof. M. Amaç GÜVENSAN
Co-supervisor: Assist. Prof. H. İrem TÜRKMEN

Traffic in cities is a huge problem. More cars and people mean more stress, accidents, and economic losses. It's getting tough to provide transportation solutions. Traffic speed prediction models help manage time in busy cities. Analyzed literature shows that short-term traffic prediction models are 95% accurate for the next two hours. Long-term traffic studies that aim 6 hours and beyond are less common than short-term ones. Short-term traffic forecasting can rely on near-future traffic speed data alone, but long-term forecasting needs more data from both the distant past and the near future. Using multi-feature spaces can help improve long-term estimation. The literature uses ARIMA, CNN, and LSTM techniques. GRU model was selected as the foundational model for the thesis after experiments with polynomial regression, CNN, and LSTM.

We investigated factors affecting traffic and proposed methods to improve estimation. The weather has a varying effect on traffic, so we introduced Pre-Tuning, which uses weather data to adjust past traffic speed during training and prediction. This improved prediction performance by up to 26% in challenging weather. Neighboring roads also influence traffic, so we fed traffic speeds of adjacent roads into the prediction model, increasing performance by up to 50% in some segments. Training models based on traffic speed groups resulted in a 44% improvement in MAPE. We propose a post-improvement structure to utilize unpredictable events. Using multiple techniques according to the segment characteristics can significantly improve traffic speed estimation.

We used data from RTMS sensors provided by Istanbul Metropolitan Municipality to predict traffic speeds for 106 road segments up to a week in advance. Our hybrid model, which considers specific factors for each segment, achieved a 35% improvement in accuracy using a multi-feature space approach. We can predict traffic speeds for up to one week with an error rate of 10% MAPE.

Keywords: Traffic speed, prediction, long-term, multi-feature space, smart city

ÖZET

Çoklu Özellik Uzayına Dayalı Trafik Hızı Tahmini

Enes BİLGİN

Bilgisayar Mühendisliği Anabilim Dalı
Yüksek Lisans Tezi

Danışman: Doç. Dr. M. Amaç GÜVENSAN
Eş-Danışman: Dr. Öğr. Üyesi H. İrem TÜRKMEN

Şehirlerdeki trafik büyük bir sorundur. Daha fazla araç ve insan, daha fazla stres, kaza ve ekonomik kayıpları beraberinde getirir. Ulaşım çözümleri sunmak giderek zorlaşmaktadır. Trafik hızı tahmini modelleri şehirlerde zamanı yönetmede yardımcı olmaktadır. İncelenen literatür örnekleri, kısa vadeli trafik tahmini modellerinin gelecek iki saat için %95 başarılı olduğunu göstermektedir. 6 saat ve daha uzun vadeli tahmin modelleri, kısa vadeli modellerden daha az yaygındır. Kısa vadeli trafik tahmini için yakın gelecek trafik hızı verileri yeterli olurken, uzun vadeli tahminler hem uzak geçmişten hem de yakın gelecekten daha fazla veriye ihtiyaç duyar. Çok özellikli veri uzayları kullanarak uzun vadeli tahminleri iyileştirmek mümkündür. Literatürde, ARIMA, CNN ve LSTM tekniklerini kullanılmıştır. Bu tezde polinomsal regresyon, CNN ve LSTM ile yapılan deneylerin ardından GRU modeli temel model seçilmiştir.

Tez kapsamında trafiği etkileyen faktörler incelenmiş ve tahmini iyileştirmek için yöntemler önerilmiştir. Hava koşulları trafik üzerinde değişken bir etkiye sahiptir, bu nedenle eğitim ve tahmin sırasında hava verilerini kullanarak geçmiş trafik hızını ayarlayan "Pre-Tuning" yöntemi önerilmiştir. Bu yöntem, zorlu hava koşullarında tahmin performansını %26'ya kadar artırmıştır. Komşu yollar da trafik üzerinde etkilidir. Yan yollardaki trafik hızlarını tahmin modeline entegre edilmiş ve bazı segmentlerde %50'ye varan performans artışı sağlanmıştır. Önerdiğimiz trafik hız gruplama yöntemi ile MAPE cinsinden %44'lük bir iyileşme sağlanmıştır. Tahmin edilemeyen olayları tahmin modeline dahil etmek için bir tahmin sonrası iyileştirme yapısını önerilmiştir. Segment özelliklerine göre birden fazla teknik kullanmak,

trafik hızı tahminini önemli ölçüde iyileştirebilir.

İstanbul Büyükşehir Belediyesi'nden alınan RTMS sensör verileri ile 106 yol segmentinde trafik hızlarını bir hafta geleceğe kadar tahmin etmek için bir hibrit model önerilmiştir. Her segmentin karakteristik özelliklerini dikkate alan bu model, çok özellikli veri uzayı yaklaşımını kullanarak %35 oranında performans artışı sunmuştur. Trafik hızları bir hafta sonrasına kadar %10 MAPE ile tahmin edilmiştir.

Anahtar Kelimeler: Trafik hızı, tahmin, uzun vadeli, çoklu özellik uzayı, akıllı şehir

1
INTRODUCTION

Predicting traffic in big cities like Istanbul provides excellent advantages for people and the environment. Optimizing the utilization of urban transportation infrastructure yields improved efficiency across various domains. It is essential for transportation planning and optimization, reducing fuel consumption, carbon footprint, the time spent in traffic, the stress level of society, and increasing public health and productivity. Improved urban traffic planning can enhance the effectiveness of public services. Thus, the available resources will be utilized more effectively to meet the needs of the growing population.

Predictions made up to 2 hours with short-term forecasting models are insufficient for long-term planning. The long-term traffic speed prediction model promises to predict traffic for up to 1 week. To ensure the success of these models, it is necessary to utilize a more significant number of data sources than what is used for short-term forecasting models. As given in Figure 1.1, short-term forecasts can use the recent past to predict the near future. Long-term forecasting models, on the other hand, go further into the past to learn the characteristics of the path. However, more historical traffic speed data is insufficient to sharpen the forecast. At this point, we encounter weather conditions, hourly, daily, weekly, and seasonal road characteristics, regional road-neighborhood relations, accidents, and unplanned events. Using environmental data like these in the prediction model takes the long-term traffic speed prediction model one step ahead.

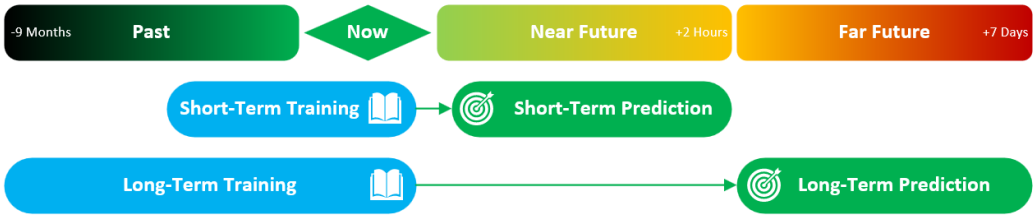


Figure 1.1 Short-Term Estimation vs. Long-Term Estimation

Before improving the forecasting model, finding the base model that gives the best

results is essential. In literature, many examples exist of LSTM, GRU, RNN, CNN, SVR, ARIMA usages [1–5]. In these studies, short-term traffic speed estimations were made. Analysis was performed on Polynomial Regression, CNN LSTM, and GRU models to choose the best-fit base prediction model. Among these models, it was continued with the GRU.

In the literature, studies show that environmental factors dramatically reduce traffic speed. [6–8]. Some analyzes were carried out to see the effect of environmental factors on traffic in this thesis. In the weather category, the effects of temperature, rainfall amount, wind speed, visibility distance, and cloudiness on traffic speed were analyzed. Interaction distances of neighboring road sections were analyzed. After that, a system architecture is designed to use environmental factors and historical traffic speeds in the prediction model.

The weather significantly impacts traffic, with an impact rate of up to 12%. Therefore, it is crucial to use weather data in the model. The real challenge is to feed the weather data into the forecast network. For this, the data normalization method has been tried in the literature [9]. In this way, an 8% increase in performance was achieved. To improve that, the idea of a method called Pre-Tuning has been put forward by analyzing the decreases/increases in traffic speed in harsh weather conditions. With this method, we propose to use the weather data to pre-tune the traffic speed instead of feeding the weather data directly to the forecast model. This means that the speed of traffic serves as an indicator of the current weather conditions.

The traffic speed of a road is directly related to its surroundings. The concentration of traffic in the nearby region is a precursor that it will intensify in the other areas in a short time. At the same time, the characteristics of neighboring roads are similar to each other. The use of nearby roads to improve traffic speed estimation holds potential. It is also possible to improve the forecast using statistical historical traffic speed analysis.

One-year traffic speed estimates were produced for 106 points in Istanbul using the dataset containing 5 minutes resolution traffic speed data from the Istanbul Metropolitan Municipality. The mentioned factors were analyzed individually in the category in which they were effective, and hybrid system architecture was revealed.

1.1 Purpose of the Thesis

The thesis aims to design a system architecture enabling multi-feature spaces in long-term traffic speed estimation. For this purpose, the factors affecting the traffic were examined. Based on the collected statistics, new data usage techniques have been suggested to improve the accuracy of the prediction model. It is believed that the accuracy of long-term traffic speed estimation improves with an increase in the amount of environmental data utilized.

In this study, we compared basic artificial intelligence methods to determine traffic speed estimates for 106 roads in Istanbul. Based on our findings, we analyzed the traffic patterns of different road zones at different times and observed how neighboring zones affected each other. We also examined changes in traffic speed during harsh weather conditions and observed how accidents and incidents impacted overall traffic speed.

1.2 Hypotheses

To effectively tackle the problems within the thesis scope, we have established a definitive set of hypotheses and milestones.

- To make accurate long-term predictions about traffic speed, it is necessary to have access to datasets beyond just traffic speed information.
- Using environmental data improves forecasting performance, especially in difficult-to-predict regions.
- The weather significantly impacts traffic, and feeding this reflection to the traffic speed prediction model will increase the prediction accuracy.
- Using traffic speed information of neighboring roads may not increase performance equally in each segment.
- To improve prediction accuracy, it's best to use a prediction model tailored to the unique characteristics of each road zone.
- Utilizing both LSTM and GRU is crucial for improving the accuracy of long-term traffic speed prediction since traffic data contains time series information.
- To understand the characteristics of a road area, it's best to analyze data from the same hours in the past weeks. However, analyzing a single point in the

past can be misleading. To get a more accurate picture, it's recommended to feed data within +/-15 to 30-minute windows regarding the estimation time.

- Using the post-improvement method for unpredictable situations such as accidents and road maintenance would give precise estimation results.

2
LITERATURE REVIEW

There are many studies in the literature that estimate traffic speed. Many studies estimating traffic flow speed on roads appear to make short-term predictions. It has also been observed that simple methods, such as averaging past traffic speeds, are used. However, traditional methods do not yield successful long-term traffic speed estimation results. Advanced techniques such as deep learning networks and machine learning produce more successful predictions in challenging conditions. It has been seen that methods such as CNN and LSTM are frequently used in the literature.

The use of ancillary data to improve traffic speed estimation is also available in the literature. Studies show the impact of traffic-influencing components, especially weather and environmental factors, on traffic forecasting. This section mentions the models used for traffic speed estimation, the factors affecting the traffic, how these factors are used to improve the traffic estimation, and the success rates.

2.1 Traffic Speed Estimation Problem

When the traffic flow estimation methods in the literature are examined, the ARIMA (Auto Regressive Integrated Moving Average) method, which allows the estimation of time series, is among the most frequently used methods. However, due to the non-linear characteristics of traffic flow, machine learning and deep learning methods have been used frequently in current studies.

Jeong et al. used online support vector machines for regression by blending them with a weighted learning method and made short-term traffic prediction[10]. The researchers processed the average speed, occupancy rate, and volume information taken at 30-second intervals over the vehicle detector by clustering them into five-minute periods. The data required for the tests were taken in two scenarios: normal traffic conditions that did not include an accident or a holiday and abnormal traffic conditions that included one of these situations. For each scenario, 15 days

of data were used for training, while the 16th day's data was used for testing. Mean Absolute Percent Error (MAPE), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) were used to measure system performance. The proposed system produced better results than the Multilayer Perceptron and Locally Weighted Regression methods, compared with an error rate of 5.3% for the first scenario and 11.9% for the second scenario.

Chang et al. made a short-term traffic forecast using the k-Nearest Neighborhood Non-parametric Regression method [2]. Traffic density information was collected at 15-minute intervals over 21 weeks. The results obtained for two days (Friday and Saturday) were used to test the performance of the proposed system, and an average error rate of 7.4% was obtained. Researchers examined the prediction successes obtained for night, morning, afternoon, and evening hours separately, compared their proposed method with ARIMA, and stated that they obtained more successful results.

Studies also use Artificial Neural Networks (ANN) to make short-term traffic forecasts [11]. Kumar et al. calculated the traffic flow by counting the vehicles in the images taken with the cameras they placed on the highway [12]. Data were collected from 11:00 to 13:00 over five days (Monday to Friday). Using the past 45 minutes of data, the researchers estimated the traffic density 15 minutes later. The visual features include the type of vehicle, average vehicle speed, traffic density, day of the week, and time of day. To reach the model with the best performance, the system was trained with different network parameters by trial and error, and the results were compared using MAPE and Mean Absolute Error (MAE). The designed system predicted the traffic density after 15 minutes with an average absolute error of 1.27%.

Huang et al. presented a two-tier deep learning architecture [13]. The proposed architecture uses Deep Belief Network (DBN) in the lower layers for unsupervised feature extraction, while Multitask Regression Layer is used on the top for guided learning. The researchers emphasized that making a separate density estimation for each road without benefiting from the density information of the associated roads is the disadvantage of the existing methods in the literature. To take advantage of the highly correlated nature of the traffic network in density estimation, DBN trained with Greedy Layerwise Training (GLT) and Restricted Boltzmann Machine (RBM) using very little prior knowledge. In the study, a PeMS dataset containing the traffic flow information of the state of California and a dataset containing the entrance and exit density data of the road connection points from China were used. The first ten months of 2011 were used for training, and the last two months were used for

testing. For the density estimation, 50 highways with the highest traffic density were selected, and the data were collected in 15-minute clusters. MAPE and Root Mean Square Error (RMSE) were calculated for performance measurement. While the prediction success (MAPE) of around 90% is achieved in the short term, the success decreases to approximately 87% in the predictions made after 60 minutes. The researchers reported that the proposed system could achieve an average of 75% success in the predictions made 12 hours later.

Another of the studies applying deep learning approaches in traffic flow prediction was done by Lv et al. [14]. The researchers used the Heap Auto-Encoder model in the lower layers to extract the traffic flow features. They added a Logistic Regression layer to the top layer to make predictions. As in [13], the Auto-encoder is trained with Greedy Layer-based training. The PeMS dataset was used to evaluate the system performance. Data collected by 15000 detectors were collected in 5-minute clusters. In the experiments, the data of the working days of the first three months of 2013 were processed. The first two months were used as training data, and the last month as test data. The researchers stated that they included the spatial correlations by using all the paths in the dataset for training and the temporal relationship by using the flow information of the previous r moment to estimate the time t. Three performance metrics, MAE, Relative Absolute Error, and RMSE, were used to measure system performance. They estimated 15, 30, 45, and 60 minutes later with the proposed system. The proposed system has been compared with different algorithms, including BackPropagation Neural Network and Support Vector Machines. Error rates varying between 6.7% and 6.1% were obtained in 15-minute and 60-minute predictions, leaving these algorithms behind.

Recurrent Neural Networks, a branch of Deep Learning approaches, are used to model systems that can be expressed as time series. However, Recurrent Neural Networks are not sufficient to learn longer-term correlations. For this reason, Long Short-Term Memory (LSTM), a subtype of Recurrent Neural Networks, has been developed [15]. The temporal and spatial nature of the traffic flow allows this information to be modeled with Long-Short Term Memory.

Ma et al. used Convolutional Neural Networks and Long Short-Term Memory Networks for short-term traffic flow prediction [4]. The researchers created separate Convolutional Neural Networks and Long Short-Term Memory-based models for each group by dividing their temporal data into four groups with K-Means Clustering. Between December 1, 2015, and November 30, 2016, traffic flow rate data within the borders of Seattle were taken at 5-minute intervals and used for training, while between December 1 and December 10, 2016, it was used for

testing. The researchers used the Average Absolute Percent Error and Standard Error metrics in their predictions for the next 5 minutes. Self-Dependent Walking Average, K-Nearest Neighborhood and Heap Auto-Encoder, Long short-term memory networks, and Gated recurrent unit determined the success of their proposed method. Compared with three different deep learning methods. The proposed approach gave the lowest Average Absolute Percent Error with 7.66%.

In their [1] study, Du et al. aimed to solve the complex and non-linear traffic flow rate estimation problem with the multi-model structure they created using Convolutional Neural Networks and Recurrent Gate Units together. Local features of convolutional networks also enabled recurrent gate units to learn distinctive long-term features. They implemented other basic methods (LSTM, GRU, RNN, CNN, SVR, ARIMA, etc.) as in many studies and compared them with their methods. As a result of the tests, they observed that their multiple models reduced the RMSE values from 10 to 4-5 in short-term traffic flow prediction. A similar study using CNN is provided under [16].

On the other hand, Qu et al. [3] used the Updated Gates and Recurrent Neural Networks structure in their study to estimate traffic flow rates up to 90 minutes. Their study examined how historical data affected success, analyzed the GRU, LSTM, MLP, and SVR methods results, and presented that they achieved slightly better success. They used the PeMS dataset in their studies.

Another study using a combination of Convolutional Neural Networks and Long Short-Term Memory Networks for short-term traffic flow prediction was published in 2018 [5]. Duan et al. used the Xi Taxi dataset, including GPS data between January 1 and May 24, September 1 and November 30, 2016. By dividing the city map into 32*32 parts, they took the entrance and exit traffic flow of each region as input, combined Long Short-Term Memory Networks evaluating temporal properties and Convolutional Neural Networks evaluating spatial correlation. A greedy training approach was used in the training phase so that the proposed method could achieve high performance in a shorter time. They compared the success of their proposed method with the combination of Convolutional Neural Networks, Convolutional Neural Networks, Long Short-Term Memory Networks, and the Linear Model. The method they suggested gave the lowest RMSE value, 11.15%.

2.2 The usage of Multi-Feature Spaces

Most of the studies given in the previous section focus on short and medium-term forecasting, and they model traffic flow using speed or density information. Some

studies conducted in recent years have added parameters such as temperature and humidity depending on weather conditions to their models in addition to speed and density information [17–19]. Some studies like [**ref19**] have researched the effects of multiple co-factors on traffic. Also, [6–8] have already shown that harsh weather conditions dramatically decrease traffic flow speed and the accuracy of traffic forecast methods.

Hu et al. propose a multi-source information fusion-based deep learning framework at [9] for traffic flow prediction that considers various external information, including weather information. They normalized the weather data to the range of [0,1] through the Max-Min operation. Although the effect of the weather is not fully known, 8% improvements were observed with the group of best metrics by feeding from the multi-feature space. Aram and Vilmos propose wavelet transform analysis in [20] to investigate the temporal correlations between traffic flow and weather factors and find relationships. Archana and Sanjay propose a soft spatial and temporal threshold mechanism in [21] to capture the spatial and prolonged impact of weather conditions. They exploit hybrid CNN-LSTM deep learning models. Their experiments show that a deep learning model trained with traffic and rainfall data gives better prediction accuracy than one trained without rainfall data. These models generate 10% and 15% Minimum Absolute Percentage Error (MAPE) values for prediction horizons less than 30 minutes and greater than 30 minutes, respectively.

In [17], Deep Belief Networks and Long Short-Term Memory Networks are used comparatively to examine the effect of different precipitation amounts on traffic flow prediction. The study includes data at a frequency of 2 minutes from Beijing between June and August 2013 and 2014. While the data for June 2013 was used for the optimization of the model parameters, one week of the data for July 2013 was used for testing, and the remaining data was used for training. The four days of the selected test week are rainy. In the study, predictions are made for the next 10 minutes and 30 minutes. In addition to the traffic flow density information, precipitation intensity information was added to the developed traffic forecast models, and the Average Absolute Percent Error in the Long Short-Term Memory model was reduced from 12.84% to 11.69% and from 13.58% to 13.14% in the Deep Belief Networks model. has been dropped.

Dunne et al. have created two different models for the days with and without precipitation, bypassing the precipitation amount and traffic density information through Stationary Wavelet Transform and using Ozlinked Artificial Neural Networks [18]. Hourly data from Dublin in January 2009 are used. A standard

Artificial Neural Network model was used to compare the efficiency of the proposed method. The tests conducted for two different observation points showed that creating two separate models for rainy and non-rainy days increased the prediction success, especially for the days without precipitation. Mean Absolute Percent Error and Square Root Mean Square Error were calculated for performance measurement. The Average Absolute Percent Error Rate for the two observation points of the proposed method was calculated as 8.0% and 9.0%, respectively.

Lv et al. also suggested using a combination of [22] Recurrent Neural Networks and Convolutional Neural Networks. Their study ensured that the flow rate information of the sensors coming one after the other on the road is stored on a matrix and feature extraction with the help of convolutional neural networks. On the other hand, since the traffic has a repetitive structure on different days of the week, they used both the traffic's periodicity and the weather and activity information. RMSE error rate could be reduced below 5%. However, this study can also estimate between 5 and 30 minutes.

Lou et al. drew attention to the increased traffic density on highways during the holidays. They used a hybrid method combining Discrete Fourier Transform and Support Vector Regression to estimate traffic specific to these days only. Discrete Fourier Transform was used to model the general characteristic of traffic flow time series and fluctuations out of the general characteristic, and Support Vector Regression was used for prediction. Flow rate data from 1200 data points belonging to 50 days between 2011 and 2015 in the Jiangsu region of China, including National and Religious Holidays, were processed with a 1-hour sampling period. The flow rate of 10 vacation days for 2015 was estimated using 40 vacation days between 2011 and 2014 for education. Mean Absolute Percent Error, Mean Absolute Error, and Square Root of Mean Square Error metrics were used to compare Self-Related Integrated Walking Average, Support Vector Regression, and Empirical Mode Decomposition methods. The researchers determined that their proposed method gave better results than other methods with a root mean square error rate of 9.6%.

It is seen that studies conducted in recent years suggest using spatial and temporal features together, and this contributes positively to success. Wu et al. [23] suggested using the spatial and temporal features of the traffic flow and the weekly daily periodicity feature. One of the rare studies that visualize the features is the study that proposes using Convolutional Neural Networks and Recurrent Neural Networks to correlate spatial and temporal features. This study aims to predict after 45 minutes. As a feature, speed estimation was made for 33 different locations in the study,

which uses the information from the same time period in the past and the traffic flow data for a certain period before and after the relevant time period. The proposed method gave better results than LASSO, BPNN(Backpropagation Neural Network), SAE(Stacked AutoEncoder), DeepST, and StoS(Sequence to Sequence Learning).

Chen et al. have used the Fuzzy Deep Convolutional Neural Networks model to create a system that can accurately predict under these conditions, mentioning that large-scale traffic flow data becomes complex and uncertain due to temporal, spatial, and external factors [24]. For the training and testing of the system, the TaxiBJ dataset containing approximately 13 months of GPS data collected between 2014-2016 within the borders of Beijing was used. The city is considered a grid with i*j pieces to model temporal and spatial dependencies. The traffic flow rate of the moment t for each piece is given as a tensor that includes features such as flow rates of entrances and exits to this piece, holiday status, accidents, and air temperature. Estimation is made for each piece's distance from 30 minutes to 6 hours. The success of the proposed method is compared with the Deep spatiotemporal residual network and Self-Related Integrated Walking Average methods. Metrics of Mean Absolute Error, Square Root of Mean Squared Error, and Mean Relative Error were used. Fuzzy Deep Convolutional Neural Networks outperformed other methods with a ratio of approximately 0.33 Root Mean Squared Error in 30-minute estimation. [25]

3
DATASET

This thesis utilized three different datasets. Traffic speed data is retrieved from Istanbul Metropolitan Municipality (IBB). The interconnections of the road zones are generated with the Segment Connection Tool (SCT) developed to collect this data. The Istanbul Metropolitan Municipality Traffic Branch Directorate provided information regarding accidents and events. The weather information is obtained from the OpenWeather data provider. This section gives specifications of datasets, applied pre-processing techniques, missing data strategies, and analysis related to data.



Figure 3.1 RTMS Sensor Instance in Istanbul

3.1 Traffic Speed Data

Raw traffic speed data consists of 1-minute period speed data collected from 12410 points in 2017, 2018, and 2019. With a data integrity rate of over 99%, it can be stated that the quantity of missing data is minimal. These data were recorded by

RTMS sensors given in Figure 3.1 located at various points in Istanbul. The data was originally in raw SQL format but has been converted to CSV format for easier use with Python scripts and Excel tools. This data formatting has been done for each segment as given in Table 3.1.

Table 3.1 Formatted Traffic Speeds

Date	Speed
01-01-2018 00:00:00	76.00
01-01-2018 00:00:01	77.00
01-01-2018 00:00:02	77.50
01-01-2018 00:00:02	77.50
01-01-2018 00:00:05	78.00
01-01-2018 00:00:06	79.00
01-01-2018 00:00:07	78.50
..	..
01-01-2018 00:02:00	138.00
01-01-2018 00:02:01	139.00
..	..
01-01-2018 00:18:50	7.50
01-01-2018 00:18:51	6.00
01-01-2018 00:18:52	12.00

Road segments also contain direction information. Roads with the same segment ID have outbound and inbound directions called 0 and 1. As the road going and coming can have completely different characteristics due to environmental factors, they are considered separate roads. Within this thesis scope, analysis was conducted on 106 segments, all located on the main roads. The distribution of 106 segments across Istanbul is given in Figure 3.2.



Figure 3.2 Distribution of 106 Segments over Istanbul

3.1.1 Pre-processing of Traffic Speed Data

A value of -1 is used to represent blank data in CSV files. In Istanbul, there is a speed limit of 120 km/h on the main roads within the framework of legal speed limits. According to the authorities, the speed data received from the sensors are expected to be 10 km/h and above due to the hardware limitation of used Remote Traffic Microwave Sensor (RTMS) sensors. Also, the maximum speed limit is stretched to 132 km/h with a tolerance value of 10%. In light of all this information, speed values other than 10 km/h and 132 km/h were accepted as useless information and replaced with -1. After pre-processing, traffic speed look as given in Table 3.2

The raw speed data is in a period of per minute. This is very high-resolution data for a long-term prediction network. Also, a noise-free speed dataset was required to improve prediction stability. To achieve this, the dataset's period was reformatted to 5 minutes by taking the average speed data within a 5-minute time frame. This approach benefited in noise removal and reduced the studied data size by up to 5 times, saving time in training the network.

Table 3.2 Traffic Speeds After Processing

Date	**Speed**
01-01-2018 00:00:00	76.00
01-01-2018 00:00:05	78.50
01-01-2018 00:00:10	78.00
01-01-2018 00:00:15	-1
01-01-2018 00:00:20	-1
01-01-2018 00:00:25	80.24
01-01-2018 00:00:30	82.30
01-01-2018 00:00:35	81.90
01-01-2018 00:00:40	80.00
..	..
01-01-2018 00:18:50	-1
01-01-2018 00:18:55	11.00

3.2 Spatial Road Connection Data

The latitude and longitude information and directions of the segments are received with the traffic speed dataset. The KMZ file, whose roads are shown as drawings on the map as in Figure 3.3, was also provided with the same dataset. This connection information provides visual line data between two segment points. However, this data was not suitable for programmatic usage.

To solve this problem, we mapped the regions behind and after the main road segments using the software we developed, "Segment Connection Tool (SCT)". In



Figure 3.3 Google Earth Footage of KMZ File

addition, it is possible to record the states of the side road connecting to the road and the side road leaving it. The data obtained with this tool has helped us see the effect of spatial data on traffic prediction. Figure 3.4 shows the program's user interface.

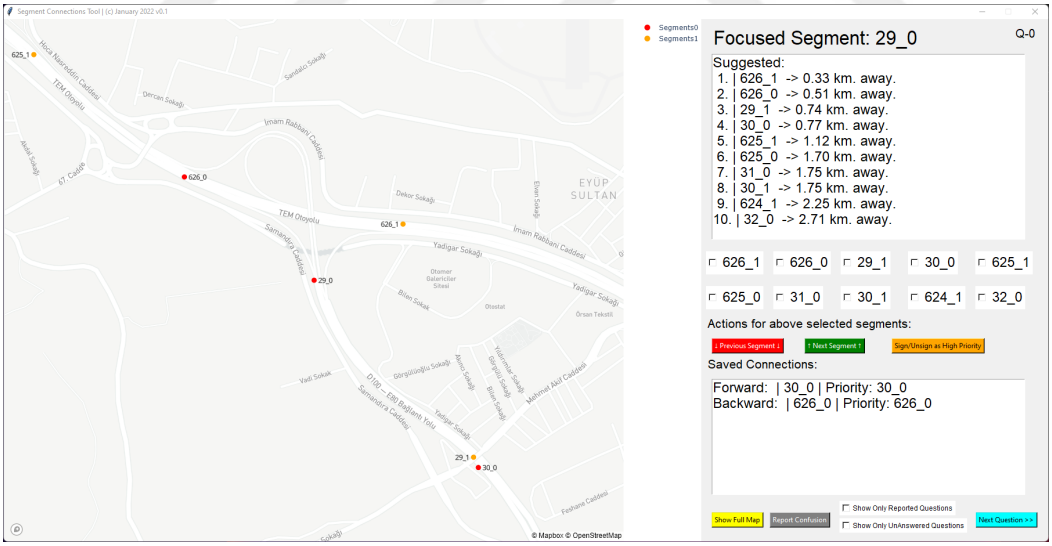


Figure 3.4 Segment Connection Tool (SCT)

The concept of priority roads was introduced to interpret the traffic flow on the main roads correctly. With this software, information can be obtained while determining the road connections, whether a road is a priority or not. While the main road parts, which are the continuation or the past of a road, are considered a priority, the side roads that separate from the main road and the side roads that connect to the main road are considered the second priority. This topic is explained in more detail under Factors Affecting Traffic Speed.

3.3 Accidents and Events Data

Accidents and events data were taken from the archives of Istanbul Metropolitan Municipality. The raw dataset has been taken in CSV file format. The information provided by the dataset regarding accidents and events information is as follows: *Incident type, start date, end date, event response date, latest status of the event, date of record update in the dataset, latitude and longitude information where the event occurred, the status of the accident if the event is an accident.* The event types included in the dataset are given in Table 3.3. The data contains different types of accidents, including chain accidents, damages resulting from accidents, accidents causing loss of life, and accidents resulting in injuries.

Table 3.3 Accident and Event Types

Event Id	Event Description
16	Accident
17	Road Maintenance
18	Road Closed
19	Weather Condition Affect
23	Busy Road
26	Intersection Events
31	Other
32	Vehicle Malfunction
33	Controlled Road Closure
34	Road Construction
35	Landscape Editing
36	Fire
37	Manufacturing Work
38	Infrastructure Work

The received dataset contains the coordinates information that the event affected. However, the segment id was not included directly. The KD-Tree algorithm has been utilized to match the segment and event areas. Thus, the improved dataset established a relationship between road segments and event areas.

In addition to the accident and events dataset, another dataset was created for special days that affect traffic speed. This dataset contains the first and last days of the new year, national holidays such as 23 April, 1 May, 19 May, 15 July, 30 August, and 29 October, and religious holidays declared as public holidays in the relevant year.

3.4 Weather Data

The weather dataset is taken from OpenWeather archive data for 1979 and 2022. The raw data consists of various data such as temperature, rainfall, wind speed,

amount of cloudiness, and visibility. In Table 3.4, the statistics of the weather data received in 2018 are given.

Table 3.4 Weather Conditions Statistics

Condition	Unit	Min.	Avg.	Max.	Std. Dev.	Fill Rate
Temperature	°C	-3.00	16.66	33.30	7.56	100.00%
Rainfall	mm	0.00	0.64	4.53	0.70	100.00%
Wind Speed	m/s	0.00	4.69	16.00	2.16	100.00%
Clouds	%	0.00	37.19	100.00	31.76	100.00%
Visibility	m	250	9549	10000	1330.53	99.27%

The raw dataset received in CSV format has been converted to JSON format for quick and easy access. This data, provided by a weather station, was chosen because it is convenient to work with high occupancy and acceptable consistency.

4
FACTORS AFFECTING TRAFFIC SPEED

Highways reflect the mobility of a living city within itself. In cities such as İstanbul city, where many people live, daily transportation activities are shaped by various factors. These factors can be summarized as temporal factors in short and long periods, special events, holidays, regional events, sports activities, weather events, accidents, road maintenance works, and the density of workplaces and residences in a region. In this section, we will examine the various factors that impact traffic and discuss them under four main categories. We will also interpret any noticeable differences.

4.1 Temporal Factors

Temporal and periodic factors are essential factors based on the routine activities of the population, which can be seen as daily, weekly, monthly, and seasonal. For example, delays in traffic during morning and evening hours and traffic densities depending on the day of the week can be shown. An abbreviation table has been used to systematically examine the temporal factors under Table 4.1.

Table 4.1 Abbreviations of Time Periods

Abbreviation	Definition
MOH	Minute of Hour
MOD	Minute of Day
HOD	Hour of Day
DOW	Day of Week
DOM	Day of Month
DOY	Day of Year
WOY	Week of Year
MOY	Month of Year

Istanbul traffic flows at an average speed of 78 km/h throughout the year. This value rises to over 90 km/h with empty roads at night. Traffic speed typically slows down significantly during working hours and picks up slightly in the middle of the day

but reaches its slowest point in the evening hours. The mass leaving work at the end of the working hours, although it varies regionally, causes traffic to be locked. The average speed graph for the hours of the day (HOD) is given in Figure 4.1.

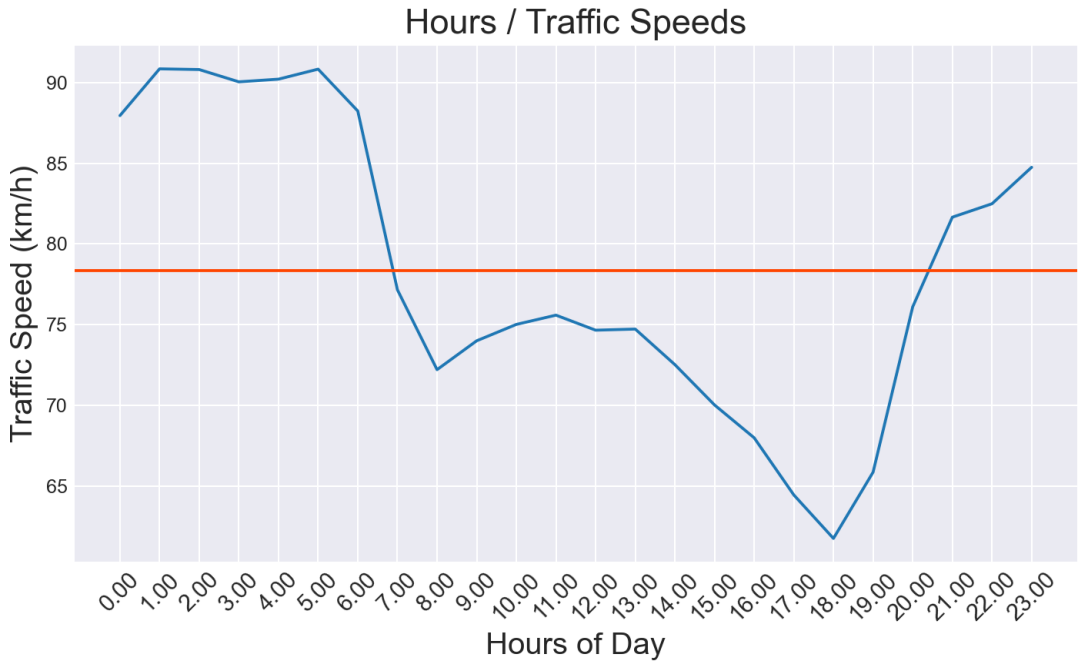


Figure 4.1 HOD Traffic Speed Analysis

Weekends are the fastest days of the week, while Fridays are seen as the day with the slowest average speed because of the increase in inter-city movements on the last day of the week. The effect of the people who prefer to go to work in their automobile on working days is reflected in the traffic speed. The average speed graph for the days of the week (DOW) is given in Figure 4.2.

At various times of the year, there are fluctuations in traffic speed. These changes occur over a long period and are affected by environmental and seasonal conditions and temporal factors. Considering the periodic migration movements of Istanbul, the most active period of the population is before and after the summer vacation. Especially in August, with the dramatic decrease in the number of people staying in Istanbul and the seasonal conditions affecting the traffic flow positively. The highest traffic speed of the year is 83 km/h. With the effect of the return to the city that started in September and the education season beginning in the middle of the month, there is a dramatic decrease in average speed. Average traffic speeds by month (MOY) are given in Figure 4.3, and seasonal traffic speeds in Figure 4.4.

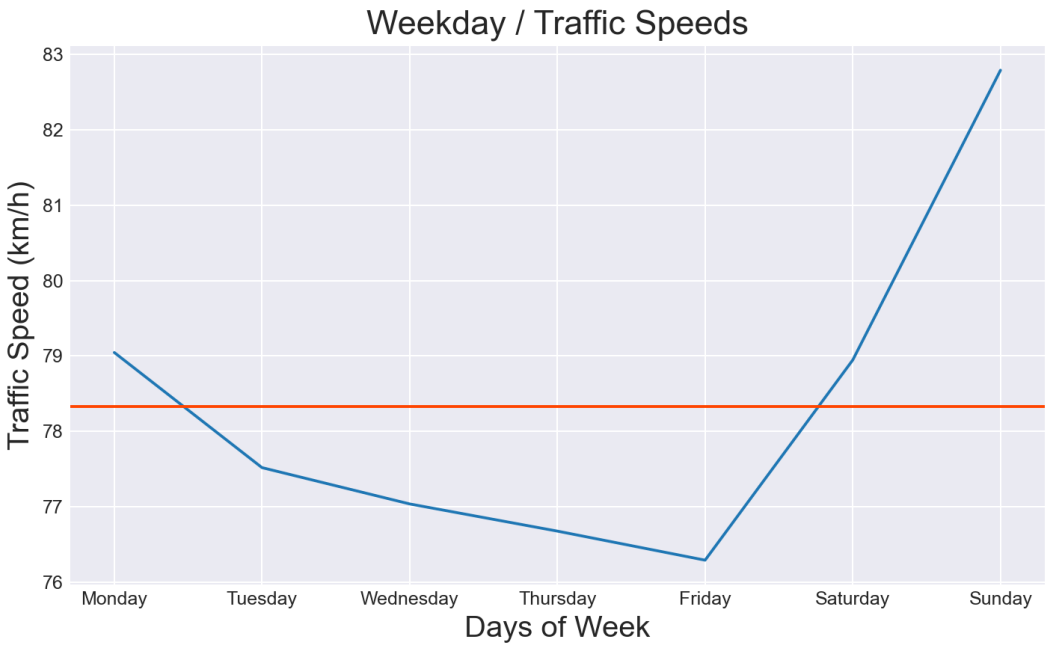


Figure 4.2 DOW Traffic Speed Analysis

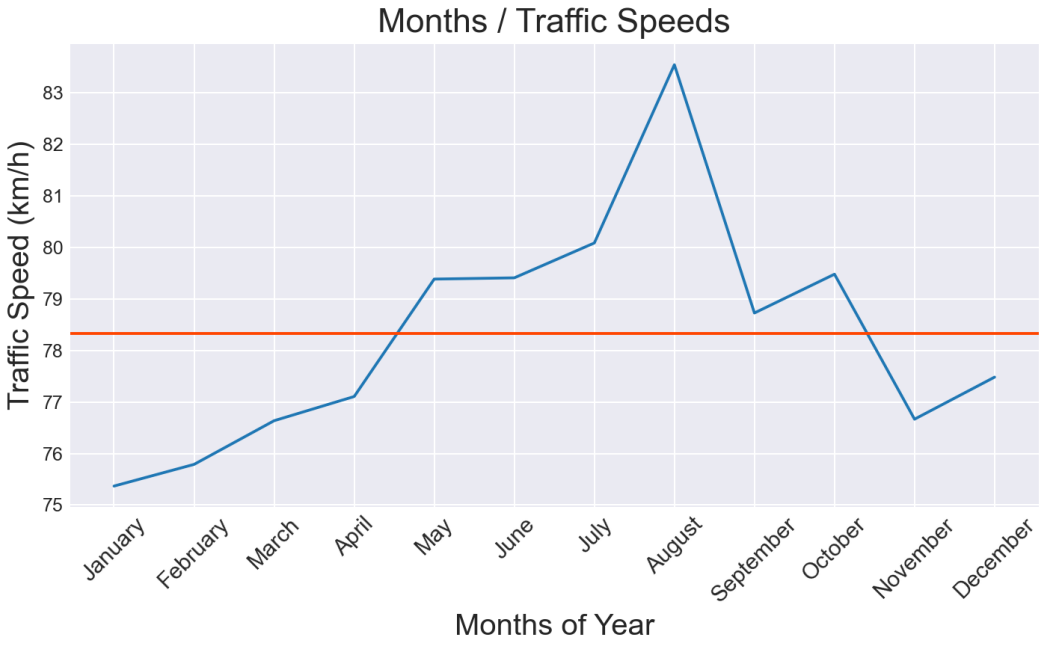


Figure 4.3 MOY Traffic Speed Analysis

4.2 Meteorological Factors

Meteorological factors have direct and indirect effects on traffic speed. The change in the friction coefficient, which affects the adhesion of the vehicles with the road surface, has a direct effect in situations where it endangers traffic safety. Similarly, a severe decrease in the range of vision also causes this effect. However, these effects are not always linear and predictable. The effect of the weather on the traffic speed

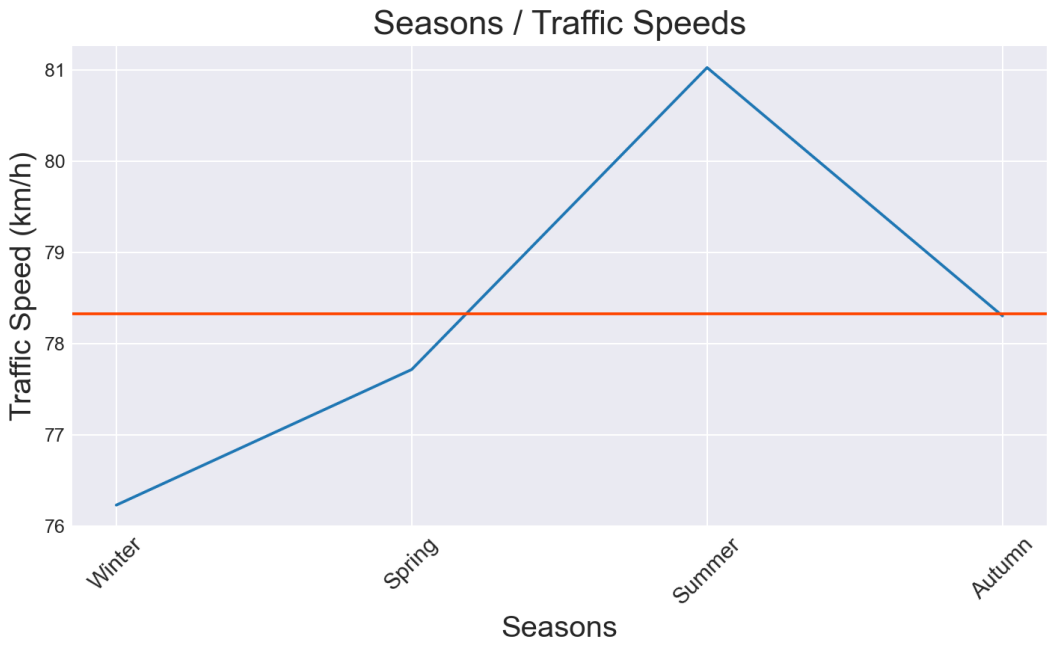


Figure 4.4 Seasonal Traffic Speed Analysis

varies with the combination of tens of physical factors such as the road architecture, the quality of the materials used, the traffic volume that the road can carry, the number of lanes, and the side road connection angles to the slope levels. This section analyses the traffic speed changes of 106 segments in a year according to weather conditions.

4.2.1 Temperature

Temperature is the weather condition that causes the most variable effects on traffic. High and low temperatures also contain seasonal characteristics. As seen in Figure 4.4, traffic speed is at its lowest during winter when the weather conditions are more severe. The temperature-traffic speed relationship in Figure 4.5 shows that the cooled air reduces the traffic speed. This effect is related to the fact that the people of Istanbul use their automobiles more on cold days, and the population density and mobility are more intense in the winter when education continues. During the summer and holidays, when schools are out, the population of Istanbul decreases. This results in a decrease in the use of shuttle vehicles. During the summer months, with lower rainfall and favorable road conditions, the maximum traffic speed is 82 km/h. This situation is especially seen in the temperature range of 20°-25° C.

It is seen that there is a dramatic decrease in average traffic speed as the temperature rises above 26° C. At this point, the fact that the weather conditions reach challenging levels for people increases the use of personal automobiles again. In

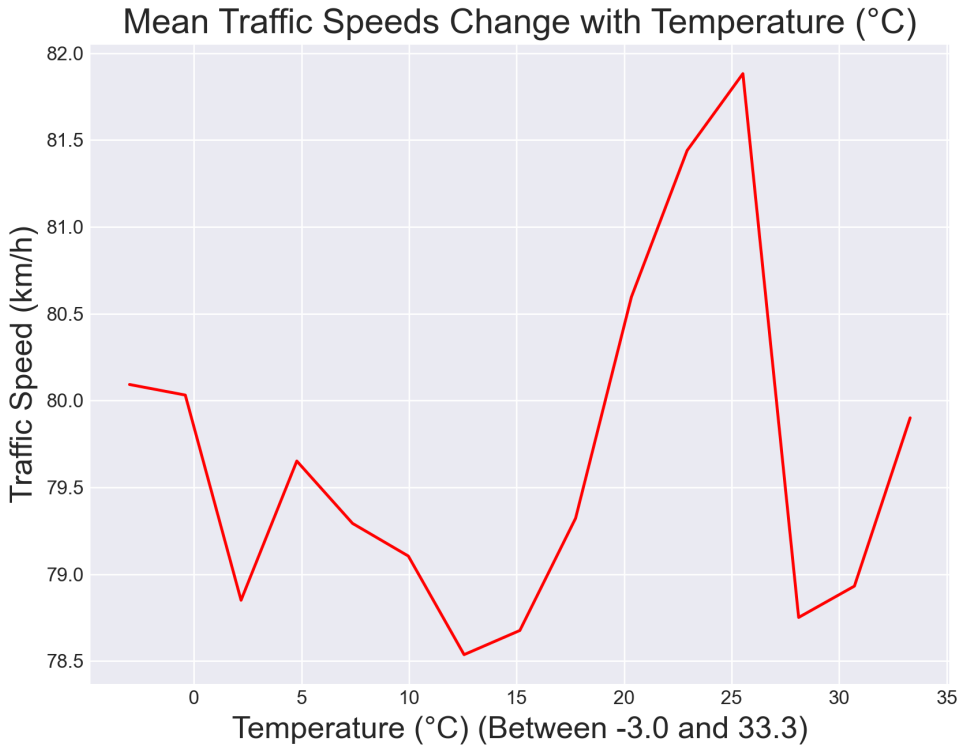


Figure 4.5 Temperature Effect on Traffic Speed

addition, if we consider that the hottest hours of the day are 2-3 hours in the afternoon, the night hours that increase the average traffic speed are not included in this average. However, HOD filtering occurs indirectly on the graph. In light of this information, the decrease in the average traffic speed with temperature values above 26° C makes sense.

4.2.2 Rainfall

Figure 4.6 shows that rainy weather conditions slow the annual average traffic speed by 1-2 km/h. During rainfall up to 3 mm, traffic flow remains relatively smooth and uninterrupted. However, an average speed decrease of up to 10 km/h is observed in rains above 3 mm. This shows that rainy weather doesn't always affect traffic similarly. The amount of rain that falls is essential. The decrease in traffic fluency, increase in accident rates, and decrease in ground friction coefficient in heavy rain conditions show its effect. It is also a factor that cautious drivers drive slower in wet weather.

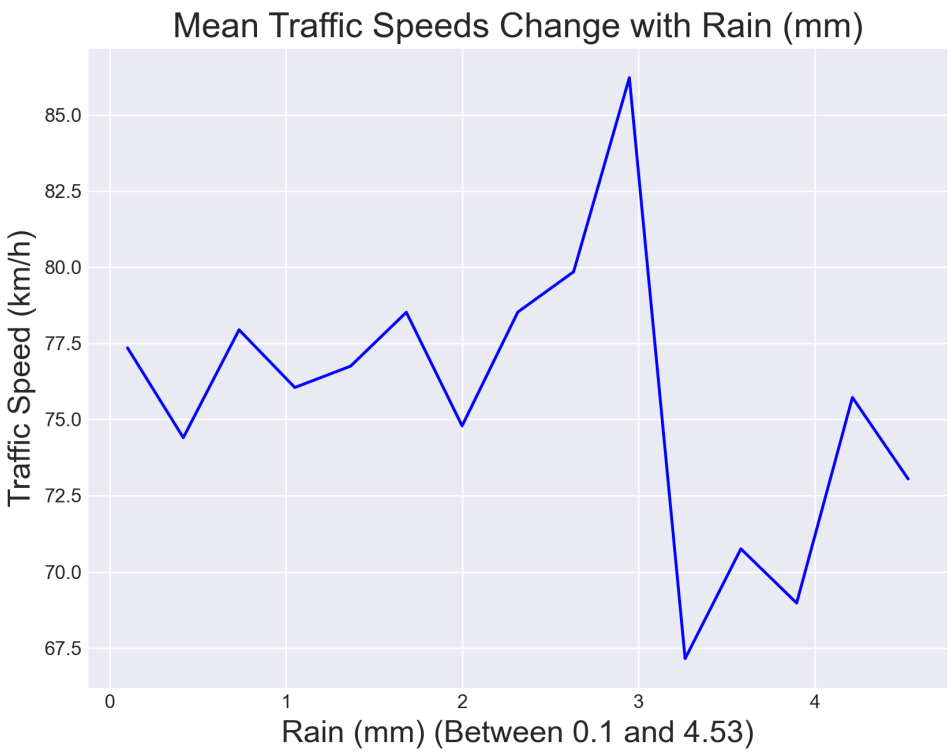


Figure 4.6 Rainfall Effect on Traffic Speed

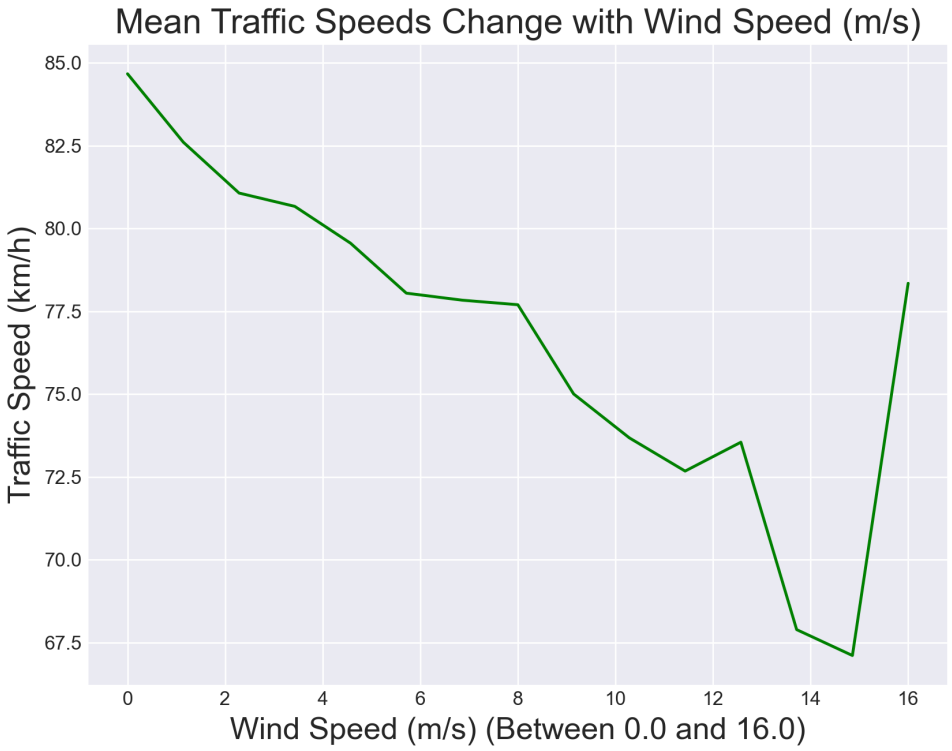


Figure 4.7 Wind Effect on Traffic Speed

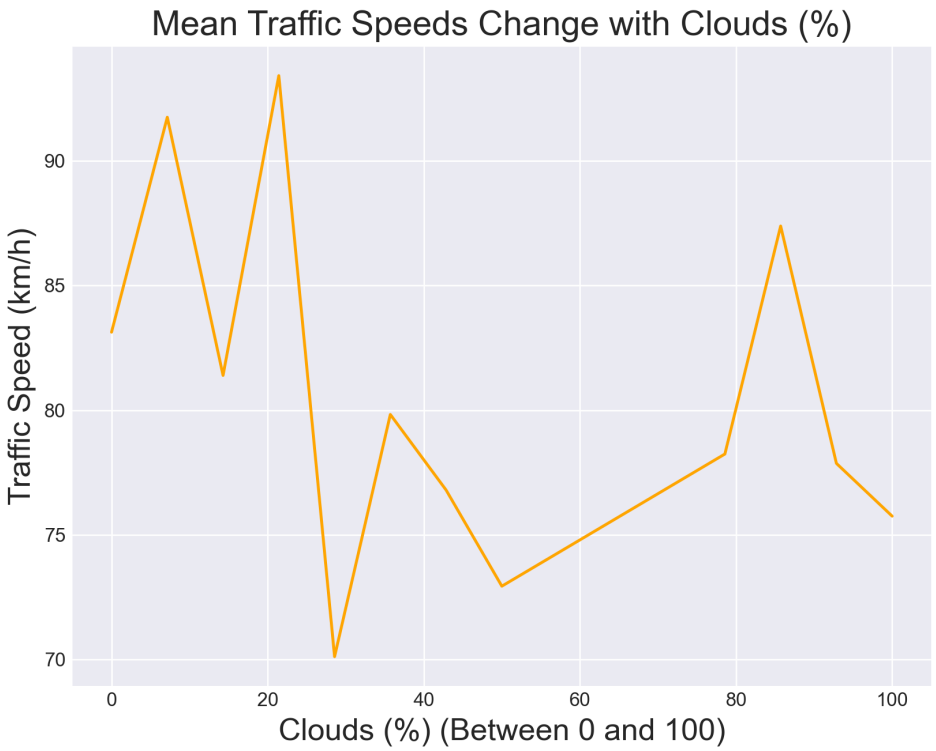


Figure 4.8 Cloudiness Effect on Traffic Speed

4.2.3 Wind Speed

Istanbul is surrounded by two seas from the north and south, with a strait passing through it. As a result of the intertwining of the water and land balance, the winds arising from the day-night temperature difference are always effective. As a result, highways have lateral forces acting on vehicles in motion. The effect of a lateral force on the vehicle moving in the lane makes it challenging to keep the vehicle in the lane and accelerate at high speeds. This is especially effective in large vehicles such as trucks.

As seen in Figure 4.7, the increase in wind speed decreases the traffic speed nearly linearly. Starting with 85 km/h, the graph goes down to 67 km/h. The dramatic increase in wind speeds above 15 m/s was not evaluated, as it coincided with a region of poor data quality.

4.2.4 Cloudiness

The increase in cloudiness can be expected to affect driving, as it occurs more frequently during rainy seasons and reduces the illuminance level. However, the analysis showed that it has indirect effects compared to other weather conditions. Therefore, it is impossible to reveal the direct impact of increased cloudiness on

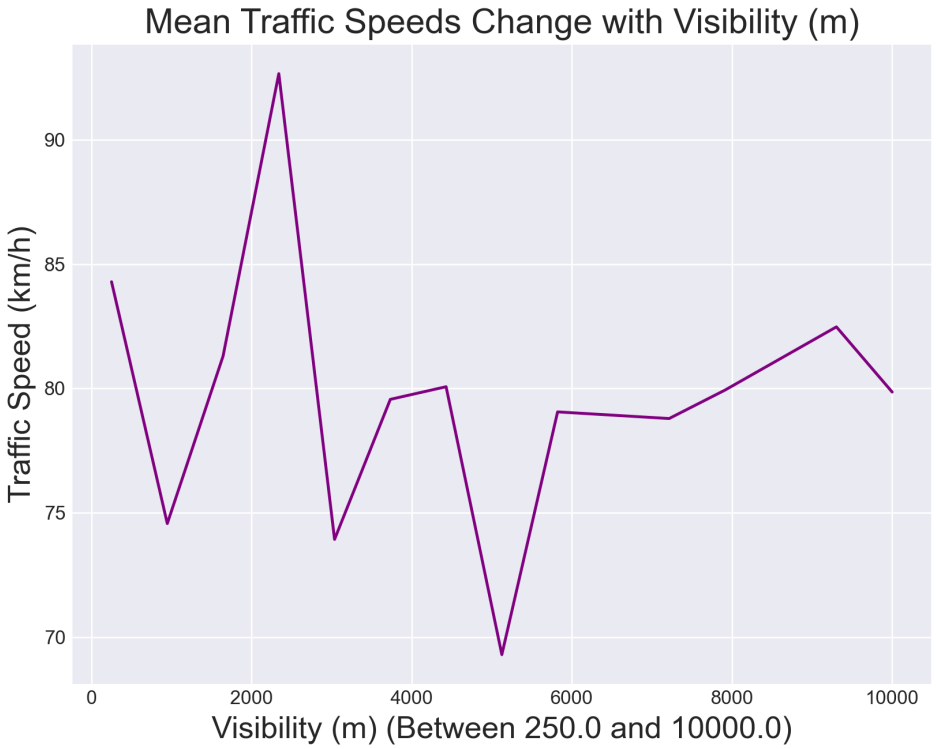


Figure 4.9 Visibility Effect on Traffic Speed

traffic speed. However, in its roughest form, it is possible to say that with the increase in cloudiness, average traffic speed decreases from 90 km/h to 75 km/h. With this analysis given in Figure 4.8, it is possible to say that the relationship between the amount of cloudiness and the traffic speed is not direct.

4.2.5 Visibility

Situations like the fog that result in low visibility below 3000 meters in Istanbul constitute %0.3 of a year. This distance is sufficient to drive safely in the city. The decrease in visibility represents the distance falling with fog rather than the decrease when combined with various precipitation situations. The local speed changes in Figure 4.9 were caused by indirect weather factors, and visibility had no usable direct effect on traffic.

4.3 Spatial Factors

One of the most critical factors affecting traffic speed is the location of the road. Depending on the location, the responsibilities expected from a road are shaped. The road structure must meet some physical conditions to perform its

responsibilities. When these conditions change, traffic speed acts like a chain reaction ahead and behind the road. Adjacent roads affect each other with many factors. Although this effect is purely regional, it depends on several factors. Some of these factors are:

- **Number of road lanes:** Main roads mostly consist of 3 lanes. At the points connecting the two main roads, the connection roads are reduced to 2 lanes. Forcing the main traffic flow to decrease from 3 lanes to 2 lanes increases the density.
- **Speed limits:** Urban traffic speed in Istanbul is legally limited to 80 km/h on main roads.
- **Road engineering level:** On some roads, the bend angle of the road creates a natural speed limit. In such cases, sudden decreases in traffic speed are observed regardless of the number of vehicles. Starting distance of the junction and separation lanes of connecting roads also affects the intersection zone traffic.
- **Viaducts / Tunnels:** Crossing the main road over a bridge or tunnel changes the driving conditions. For example, in the Avrasya tunnel, the traffic speed is fixed at 70 km/h. Limited visibility in tunnels and increased exposure to weather events on bridges affect the driver's decisions.
- **Status of providing connectivity to work, residence, and social activity zones:** In Istanbul, it is possible to divide the regions into sub-regions regarding residence, business center, social activity, sports activity, bus station, and airport. This situation causes mass movements to increase in one direction at certain hours. For example, while traffic is seen in the European direction, which has a high workplace density due to business traffic in the morning, it can be seen that the movement to the Anatolian side, which has a high residential area, increases during rush hour. A similar situation can be summarized as the increase in the number of people going to the stadium areas on match days and the rise in the number of people going to the airport and bus station on holidays.
- **An alternative of the road:** There may be alternative roads. A driver who actively checks the traffic status chooses the most efficient route. For example, there are two bridges and one tunnel alternative for the city center to cross between two continents. Similarly, a driver who wants to go from Kagithane to Esenler can reach it via the Hasdal Connection Road or by

connecting to the E-5 via the main highway. However, when there are no other options but to use a road, the traffic density increases.

An area with traffic jams gives a hint that the traffic will increase at the points it affects. In Figure 4.10, an example of traffic speed decrease with an interval of 15 minutes is given. This situation is like a chain reaction. Each road segment has the potential to influence the preceding road segment. This situation can be compared to the pressure increase in a water pipe compressed at one point. While there is an increase in pressure before the region where the pipe is compressed, there is an increase in the flow rate in the remaining section. The same is valid for traffic flow. But it is directly related to the characteristic of the road. Some roads may have an effect up to the road 3-hops behind them, while some roads may not significantly affect the road before them.

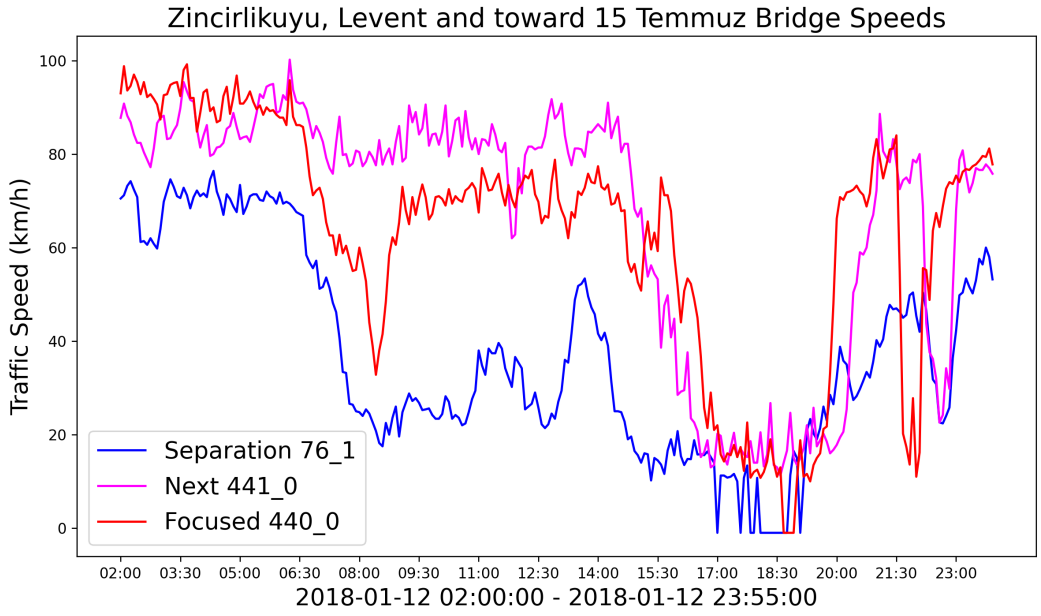


Figure 4.10 Intersection Speeds Example: Zincirlikuyu

In Figure 4.11, the part of the traffic coming from the Zincirlikuyu direction, which departs in the direction of Levent and continues to the 15 Temmuz Bridge, is given. The data on 12/01/2018 representing the one-day traffic speeds of this area are given in the graph in Figure 4.10. It is known that the Levent entrance is busy in the morning hours. On the 76_1 road segment, there is a graph that shows a decrease in speed from 75 km/h to 20 km/h after 7.00. This intensity in morning traffic affects the 440_0 main road segment with a delay of approximately 15 minutes. Despite this, it is noteworthy that there has not been a dramatic decrease in traffic continuing in the direction of the bridge. Based on these data, it can be seen that other roads that provide traffic are also blocked when a side road is blocked. A similar situation



Figure 4.11 Intersection Example on Map: Zincirlikuyu

is also valid for the blockage of the 441_0 point and then the slowdown of the 440_0 segment while going in the direction of Anatolia in the evening. Adjacent road zones have affected each other in a delayed manner. The Zincirlikuyu example reveals the traffic relationship for only one hop distance. As such, there are regions with solid connections to road segments in 1-hop, 2-hop, and 3-hop. The concept of the hop distance of road zones is plotted in Figure 4.12.

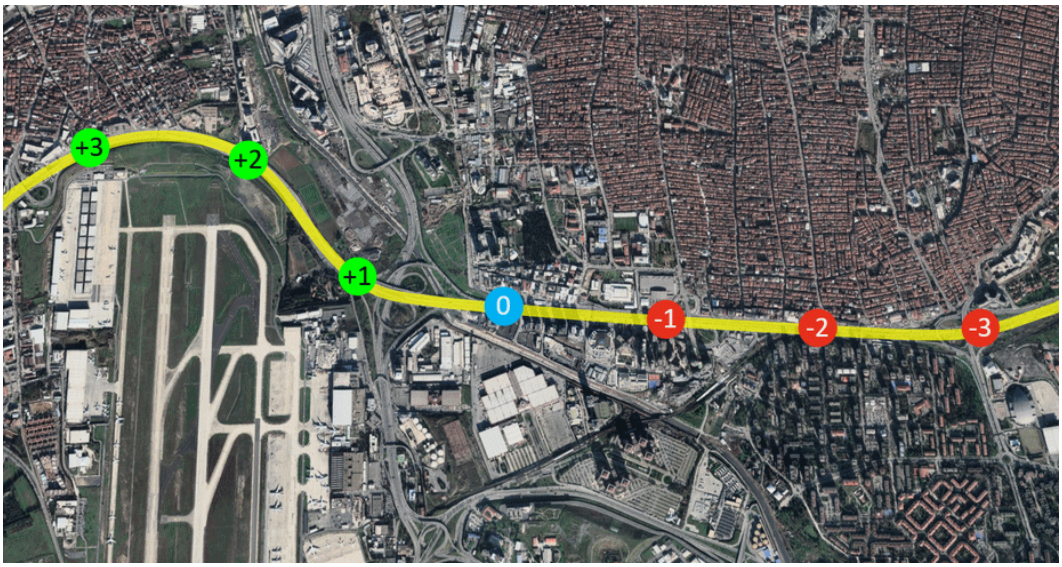


Figure 4.12 Hop Distance of Road Segments

4.4 Accidents & Events

Upon analyzing the accident and events dataset, it was observed that the traffic speed had decreased by 29.3% during the incident. This decrease rate is affected by the specific characteristics of the road segments. The standard deviation value obtained during the incident was measured as 17%. Furthermore, there were significant decreases in traffic speed during the incident in 95% of road areas. Based on these observations, the suggestion was made to reduce the estimated traffic speed during accidents and incidents at specific rates.

When traffic accident times are used in the prediction model's training, it negatively affects its success. When the relevant time intervals were removed from the model training, there was a 0.3% MAPE improvement.

4.5 Statistical Analysis of Traffic Speed

Analyzing historical traffic speed data provides valuable clues to understanding the problem. This section includes the statistical analysis of traffic speed.

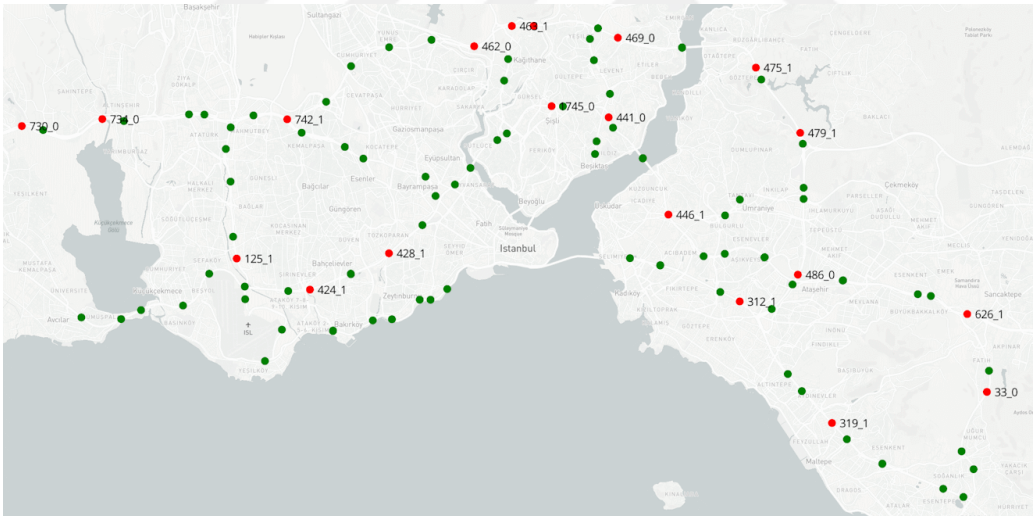


Figure 4.13 Most Difficult Road Zones to Predict

4.5.1 Most Difficult Road Segments to Predict

As the speed of a road segment changes, it becomes harder to estimate. To measure the inconsistency in traffic speeds, the standard deviation was utilized. The literature provides examples of how standard deviation can determine the difficulty level in estimation [26]. Figure 4.13 shows the 20 most difficult-to-predict road segments out of 106.

Table 4.2 106 Segments Statistics and Locations

Segment	County	Mean Speed (km/h)	Standard Deviation ▽
742_1	Bağcılar	74.66	32.23
469_0	Şişli	90.44	32.15
441_0	Şişli	64.89	30.84
312_1	Kadıköy	65.19	30.73
1745_0	Şişli	72.64	30.19
462_0	Eyüp	90.59	29.01
446_1	Üsküdar	62.39	28.16
730_0	Avcılar	84.28	27.81
479_1	Üsküdar	91.16	27.17
424_1	Bakırköy	57.78	26.62
428_1	Zeytinburnu	73.23	26.51
463_1	Kağıthane	93.43	26.11
125_1	Küçükçekmece	75.54	25.95
319_1	Maltepe	71.48	25.78
475_1	Üsküdar	91.1	25.73
33_0	Maltepe	85.15	25.6
486_0	Ataşehir	80.08	25.4
734_0	Başakşehir	76.03	25.27
626_1	Sancaktepe	91.41	25.13
465_0	Kağıthane	86.63	24.7
316_1	Ataşehir	72.26	24.69
426_1	Bakırköy	60.8	24.49
488_0	Ataşehir	67.31	24.35
493_0	Esenler	82.94	24.3
624_1	Ümraniye	82.56	24.03
414_1	Avcılar	64.17	24.02
36_1	Kartal	79.62	23.76
738_0	Başakşehir	77.49	23.75
482_1	Ümraniye	91.89	23.44
36_0	Kartal	81.2	23.38
321_0	Maltepe	68.51	23.23
421_1	Bakırköy	74.18	22.9
510_1	Bakırköy	66.98	22.79
312_0	Kadıköy	73.47	22.71
318_0	Ataşehir	71.56	22.64

Table 4.2 106 Segments Statistics and Locations (continued...)

Segment	County	Mean Speed (km/h)	Standard Deviation ▽
1745_1	Şişli	65.6	22.05
417_1	Avcılar	63.93	22.01
122_1	Bağcılar	78.45	21.78
434_0	Eyüp	72.93	21.6
322_1	Maltepe	72.52	21.24
476_0	Beykoz	93.11	20.51
427_0	Bakırköy	72.19	20.42
486_1	Ataşehir	86.77	20.09
76_1	Beşiktaş	44.35	20.07
620_1	Ümraniye	85.95	20.03
458_0	Sultangazi	100.46	19.81
454_0	Esenler	87.74	19.72
121_0	Bağcılar	81.52	19.7
309_0	Üsküdar	83.4	19.36
74_0	Kağıthane	53.39	18.96
738_1	Başakşehir	74.24	18.92
431_0	Zeytinburnu	72.82	18.88
472_0	Sarıyer	87.65	18.87
617_0	Kadıköy	78.04	18.29
413_0	Avcılar	59.18	18.13
614_1	Üsküdar	92.61	18.12
441_1	Beşiktaş	78.51	18.08
444_0	İstanbul	76.79	17.76
482_0	Üsküdar	96.22	17.72
119_1	Bağcılar	65.59	17.71
603_0	Üsküdar	74.84	17.59
2102_0	Üsküdar	59.46	17.29
424_0	Bakırköy	73.54	17.26
510_0	Bakırköy	71.1	17.11
741_0	Bağcılar	83.07	17.01
621_0	Ümraniye	85.66	16.93
443_1	İstanbul	78.79	16.62
420_0	Avcılar	77.8	16.46
497_0	Eyüp	79.26	16.23
125_0	Bahçelievler	86.64	16.17
490_0	Bağcılar	80.76	15.9

Table 4.2 106 Segments Statistics and Locations (continued...)

Segment	County	Mean Speed (km/h)	Standard Deviation ▽
64_0	Şişli	52.35	15.5
625_0	Sancaktepe	92.32	15.39
416_0	Avcılar	75.72	15.03
459_1	Gaziosmanpaşa	91.88	15.01
480_0	Üsküdar	97.76	14.91
495_1	Eyüp	82.47	14.08
219_1	Kağıthane	53.42	14.04
734_1	Başakşehir	90.97	13.95
455_1	Sultangazi	97.71	13.18
128_0	Bakırköy	67.2	12.58
730_1	Avcılar	92.14	12.55
323_0	Maltepe	79.55	12.53
504_0	Bakırköy	63.54	12.15
216_1	Kağıthane	87.91	12.07
31_1	Sancaktepe	92.15	11.83
436_0	Kağıthane	75.71	11.79
491_1	Esenler	87.02	11.67
614_0	Üsküdar	86.14	11.52
74_1	Kağıthane	64.07	11.29
467_1	Şişli	100.22	11.17
432_1	Eyüp	77.96	11.11
308_1	Üsküdar	81.1	10.94
216_0	Kağıthane	88.44	10.85
603_1	Üsküdar	81.59	10.63
471_1	Sarıyer	100.7	10.37
616_1	Kadıköy	100.82	9.81
514_0	Fatih	78.67	9.39
507_1	Bakırköy	62.19	9.29
1751_0	Zeytinburnu	80.47	9.2
1751_1	Zeytinburnu	79.72	9.05
1779_0	Bakırköy	61.61	8.77
326_1	Kartal	95.14	8.74
64_1	Şişli	54.46	8.54
326_0	Maltepe	85.21	6.9
130_1	Bakırköy	76.87	5.28

5
SYSTEM DESIGN

This section describes the proposed system architecture. Research and analysis studies were carried out on the prediction models to form the basis of the traffic speed prediction model. Then, a standard structure was created to give the traffic speed to the prediction model designed with GRU. After the model architecture was prepared, which can only predict traffic speed, modules representing weather conditions, speed groups, and spatial data features were added. Finally, post-prediction improvement methods are described.

5.1 System Architecture

A system architecture that predicts traffic speed for the long term is designed. The system architecture is given in Figure 5.1. This architecture allows easy incorporation of environmental data to improve traffic speed prediction. The system consists of four primary groups: Model Training, Prediction, Post-Prediction, and Presentation.

5.1.1 Model Training

It is the stage where artificial neural network models are produced. At this stage, archive data is used. First, each data category undergoes preprocessing. Data gaps are cleaned in preprocessing, and minimum and maximum checks are performed. Road segment area and time-specific data are filtered and sent to the Input Generator module. Input Generator is a module used in Model Training and Prediction stages. The task of this module is to convert the archive data into the format required by the prediction model. The data entering the Input Generator module can be in various units. But the output is always in traffic speed (km/h) format. The details of the process performed at this stage are explained in more detail in the Speed Data Utilization and Multi-Feature Space Prediction Network sections. When prediction model training is complete, model sets are saved to storage in h5 format. After that,

the prediction models are ready to make a new forecast.

5.1.2 Prediction

The prediction stage is where the model is run using the most recent weather forecasts and historical traffic speeds. Among the prediction models saved in the storage area in the previous stage, the most valuable one for the relevant road segment is processed. When this phase is completed, traffic predictions with 5 minutes resolution extending up to 1 week into the future are obtained. Choosing the most valuable model ensures that the best possible (promised in the model) estimate is made for the road area. The output results are ready for use but sent to the Post-Prediction stage for further refinement.

5.1.3 Post-Prediction

At this stage, improvements are made to the predictions that have been produced. It is possible to make optimizations specific to accident and event times. Revising forecasts with new resolutions or legal speed limits is offered. For details on Post-Prediction methods, you can review the Post-Prediction Improvements (L3+) section.

5.1.4 Presentation

The presentation phase provides access to long-term traffic speed data that can be useful in various industries. Forecasts generated in this layer are expected to be valuable. However, it should be noted that the presentation layer does not involve any improvements within the scope of the thesis.

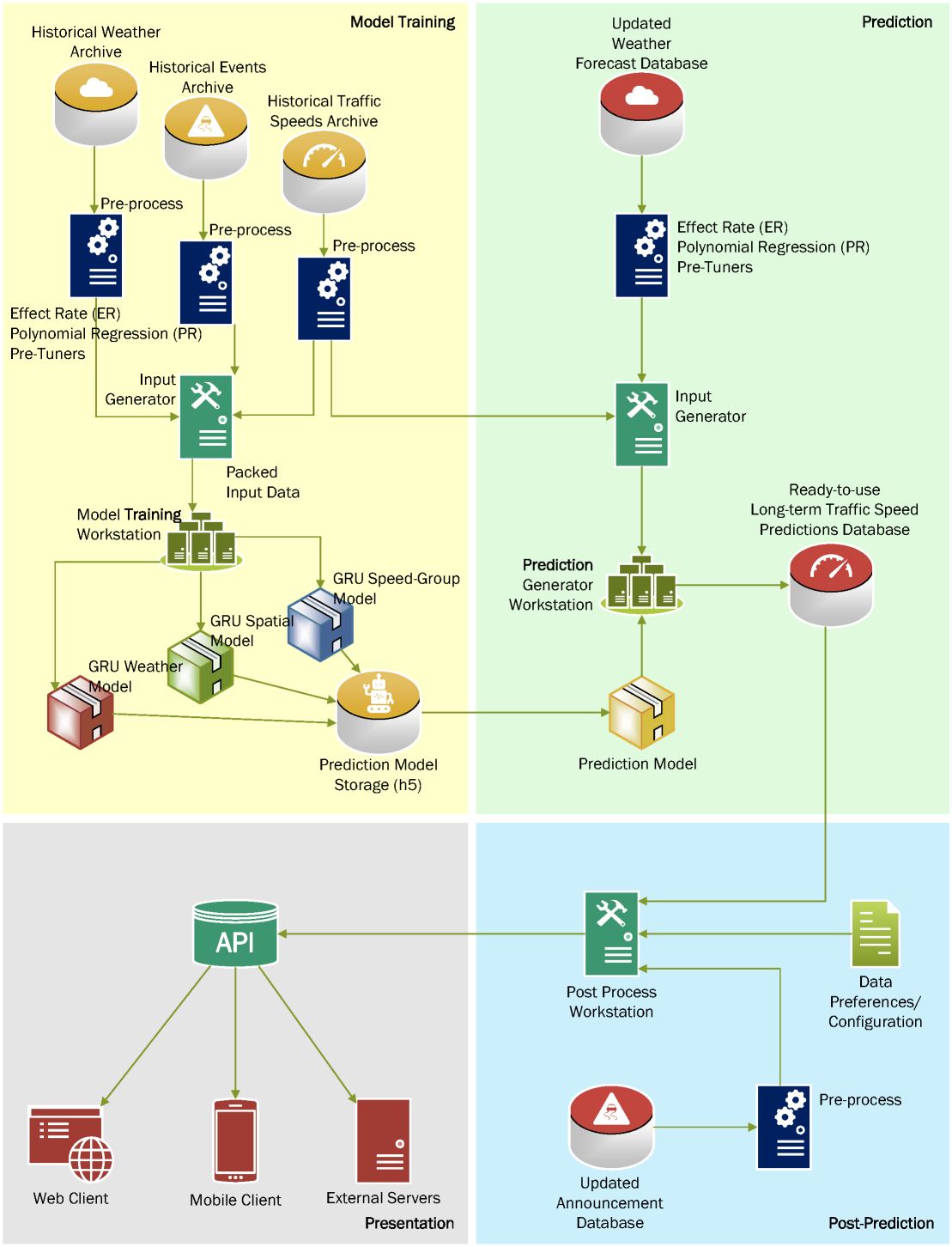


Figure 5.1 System Architecture

5.2 Long-Term Traffic Speed Prediction Algorithms

To construct the predictive model architecture, it is necessary to determine the most successful algorithm. This section details models that utilize different technologies and make predictions solely based on traffic speed and time information. Information on the usage scenarios where the models are strong and the steps taken until reaching the selected model in the final are given.

5.2.1 Polynomial Regression

Observed traffic speeds of a road segment in a day can be imagined as curves with specific breakpoints. By ignoring these micro-level ups and downs, it becomes possible to represent a day's traffic speed with a curve. The number of breakpoints may vary according to the traffic characteristics. It is possible to use n'th order Polynomial Regression to generate the curve. Polynomial regression is a regression analysis in which the relationship between the independent variable x and the dependent variable y is modeled as an nth-degree polynomial in x. Equation 5.1 gives the theoretical formula of Polynomial Regression.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_m x_i^m + \varepsilon_i \ (i = 1, 2, \ldots, n) \tag{5.1}$$

5.2.2 CNN: Convolutional Neural Network

Convolutional Neural Networks (CNN) is an artificial neural network that solves image processing and pattern recognition problems. CNNs use convolution to model the data's local relationships and hierarchical features. In this way, it can analyze large datasets and detect complex patterns. CNNs, unlike traditional neural networks, treat the input data as a multi-channel data structure rather than a flat vector. For example, it renders image data orderly, preserving one or more color channels of each pixel. In Figure 5.2, the layers of the CNN network are given. The basic building block of CNNs is convolution layers. Convolution layers work with the convolution operation applied to the input data with one or more filters (kernel). Each filter consists of learned weights to detect a particular feature (edges, corners, texture, etc.). The convolution operation happens when the filter is shifted over the input, and the dot product operation highlights features in the region of interest. This process captures local features in the input data and creates more complex features. After the convolution layers, activation functions (like ReLU) and pooling layers are used. Activation functions increase the network's capacity

by smoothing the outputs of the convolution layers and providing a non-linear response. Pooling layers, on the other hand, act as size reduction. It also reduces the computational load of the network by reducing the number of parameters and increasing the scalability of features. At the end of the CNNs are the fully connected output layers. Fully-connected layers transform the entire feature map into a single vector, making this vector available for classification or regression. The output layer is adapted to the requirements of the given task. For example, in classification problems, class probabilities are calculated using the softmax function.

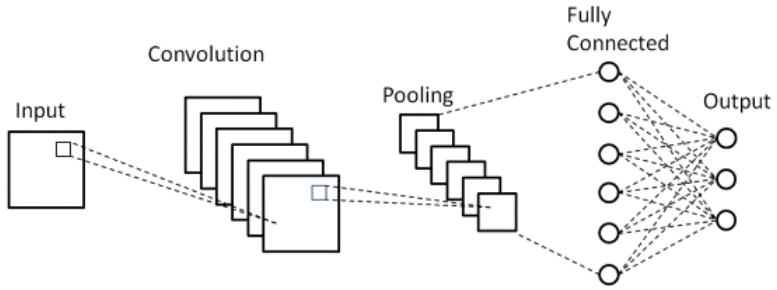


Figure 5.2 Convolutional Neural Network Layers

Considering the neighborhood of road zones and the relationship of time zones in the traffic rapid estimation problem, the CNN method is one of the candidate methods. Considering this point of view, architecture has been designed so that containing the time relation in the vertical and the neighborhood relation in the horizontal is fed to the CNN network. However, since this relationship does not include a continuous pixel change as in the pictures, CNN could not successfully capture the relationship. The traffic speed estimation problem, considered a time series problem, could not be defined successfully using CNN.

5.2.3 LSTM: Long Short-Term Memory

Long Short-Term Memory (LSTM) is particularly effective in processing time series data and solving natural language processing problems. Unlike traditional recursive neural networks (RNN), LSTM can more effectively model long-term dependencies. The LSTM structure is given in Figure 5.3. LSTM consists of special units called memory cells. These cells have three main components that interact with each other: the forget gate f_t, the input gate i_t, and the output gate O_t. These gates control how the memory cell updates and transmits its information. First, the forget gate determines which information the memory cell retains and which it forgets. This gate controls how much of the past information is retained. Next, the input port determines how much new input information it adds to the memory cell. This gate controls how much new entries contribute to the memory

cell. The gate uses a sigmoid activation function and a hyperbolic tangent activation function. New information added to the memory cell updates the contents of the memory cell. This update is calculated using a sigmoid function and a hyperbolic tangent function. Finally, the output port determines how much the updated memory cell must be transmitted to the outside world. This gate is calculated using the hyperbolic tangent function and the sigmoid function. The Sigmoid Function is given in Equation 5.2.

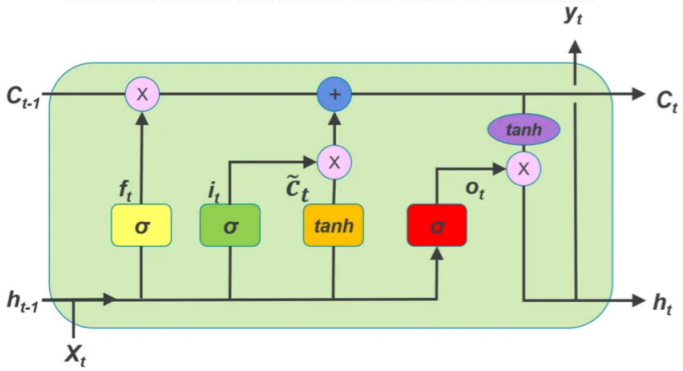


Figure 5.3 General LSTM Architecture

$$S(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1} = 1 - S(-x). \tag{5.2}$$

The primary purpose of LSTM is to model long-term dependencies effectively. LSTM preserves historical information thanks to the forget gate to handle long-term dependencies more effectively. LSTM has yielded successful results in many areas, such as time series estimation, language modeling, text classification, text rendering, language translation, and speech recognition. With this aspect, the chance of success in long-term traffic speed estimation, a time series problem, is considered high.

In addition, variations based on LSTM have been used to solve different problems by changing the structures and functioning of memory cells. For example, a variant of LSTM called GRU (Gated Recurrent Unit) has a more straightforward structure and performs similarly. The GRU is explored in the next section.

5.2.4 GRU: Gated Recurrent Unit

GRU effectively processes time series data and solves natural language processing problems, similar to the long short-term memory (LSTM) method. However,

GRU has a more straightforward structure than LSTM. GRU, unlike LSTM, uses "update" and "reset" gates instead of "forget" and "input" gates. The updated port determines what information will be updated to the memory cell, while the reset port determines what information will be sent to the memory cell. These two gates are calculated using the sigmoid and hyperbolic tangent functions. GRU is designed to model long-term dependencies like LSTM. GRU can effectively handle long-term dependencies while updating the contents of the memory cell, ensuring that historical information is preserved. It can also be trained faster and uses less memory as it contains fewer parameters than LSTM. With GRU, getting more successful results simultaneously by using more parameters than LSTM is possible.

5.3 Base Prediction Model (L0) Features

In this section, the way of feeding the data used in the base model to the network and the default hyperparameters are explained.

5.3.1 Time Data Utilization

The past characteristic of a road segment plays a crucial role in predicting its future. The behavior of the traffic speeds on the days of the week is similar. Based on this observation, a structure was proposed on the input of minute of the day (MOD), day of the week (DOW), day of the month (DOM), and month information in model training.

5.3.1.1 Feeding Time Data as Vectors

Time data is fed into the network with 7-dimensional vectors of the day of the week and 31-dimensional vectors of the day of the month. The minute information of the data with a data resolution of 5 minutes is fed with the vector to represent the minute of the day. Since the size of this vector will be 12-period regions in 24 hours, it is fed with 288-dimensional vectors. An instance is given in Figure 5.4

In the experiments carried out during the design process, it has been observed that feeding the time information to the network as data with the embedding method has a limited effect on the prediction. Therefore, this technique was not used in the base model that was decided upon.

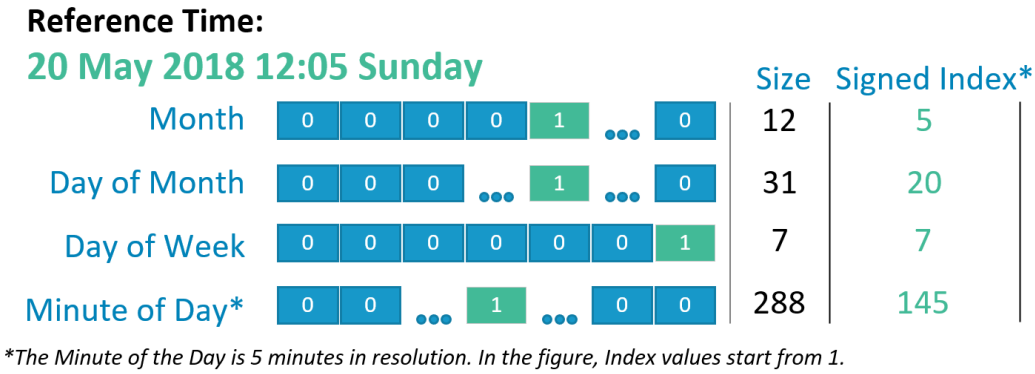


Figure 5.4 Time Data Embedding Schema

5.3.1.2 Indirect Usage of Time Data

Using the traffic speed data corresponding to the periods as input removes the requirement of directly feeding the time information to the forecast model. For instance, to predict a Monday, looking at the history of Mondays is enough to train the network in the characteristic that Mondays have. Similarly, for around 17:00 on a Monday to be predicted, using the same time interval of the previous weeks provides accurate time information to the prediction model during training and testing. A time interval between half an hour before and after 17:00 on Mondays provides valuable data. By applying this approach back to 4 to 6 weeks, the stability of the forecast model is also increased.

5.3.2 Speed Data Utilization

Traffic speed shows ups and downs periodically. To catch these harmonic changes, examining their periodic characteristics and evaluating them with time together is necessary. The traffic speed of a road segment that has shown specific characteristics in the past will be substantially similar in the future. Although environmental information mentioned in the following sections will strengthen the estimation, Past traffic speeds are the primary data source. Therefore, correctly interpreting the speed and time relationship, which is the basis of the base model, directly affects the result.

In the drawing given in Figure 5.5, an input chart that goes back four weeks and spans 30 minutes backward and forward is given. An example of the traffic speed input structure is given in Figure 5.6. This designed input structure has been used for input in prediction models with various technologies and developments. The values of 4 weeks and 30 minutes on this chart are variable, and six weeks and 15 minutes are also used in different scenarios. This architecture aims to take the speed frame in a time interval instead of taking the speed values in the past time series as

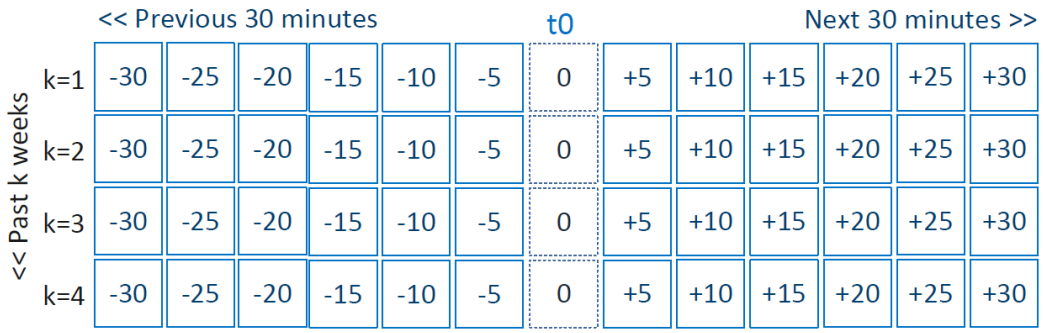


Figure 5.5 The Overall Aspect of Speed Data Usage Timeline

just one point and thus enable the prediction model to learn more data.



Figure 5.6 Speed Data Input Schema with Sample Speeds

5.3.3 Default Hyper Parameters of Network

Experiments were run iteratively within the scope of the thesis. In these experiments, a default parameter set was determined to see the effect of the newly added feature. The hyperparameters given in Table 5.1 are the values used for the base GRU model. Since all needs cannot be met with a single parameter set, the idea of two parameter sets named Paramset-A and Paramset-B has been proposed. In experiments with Paramset-A, the model has a training set of 9 months, while experiments with Paramset-B learn from 6 months of historical data. This difference allows Paramset-A to learn more, while Paramset-B learns from closer months (from recent data). Paramset-B is less affected by the exceptional data due to the difference in the 4-week and 6-week retrograde points. Also, the window size read from the previous week is larger in Paramset-A, which similarly increases the data stability. There have been experiments that have utilized both sets of parameters. A GRU long-term traffic speed prediction model architecture that goes back to 4 weeks and uses a 15-minute window is given in Figure 5.7.

A custom function 5.3 using MAE and MAPE is used as the Loss Function. Models trained with this optimization function were used because they produced more



Figure 5.7 GRU Network Architecture

Table 5.1 Prediction Network Default Parameters

Parameter	Paramset-A Value	Paramset-B Value
Number of Past Weeks	4	6
Time Window size	-30 to +30 mins	-15 to +15 mins
Training data (Previous months)	9	6
Estimation Resolution	5 mins	
Testing data (Months)	12	
Number of models (Per segment)	12 (model per month, 1 year predicted)	
Number of road segments	106	
Prediction Horizon	Up to 1 week (Long-Term Estimation)	
Missing data strategy	Discard whole instance	
Regularizer	L2	
Optimizer	Adam	
Loss Function	Multiple Loss Function (with MAE and MAPE)	
Number of Epochs	Max 1500 (with early stopping)	

successful results. In the equation, the weight of MAPE is set to 3, while a weight of 1 is defined for MAE. The value expressed as s_r represents the actual velocity value, and s_p represents the estimated velocity value.

$$[3 * mape(s_r, s_p) + mae(s_r, s_p)]/4 \tag{5.3}$$

5.4 Multi-Feature Space Prediction Network

Many factors affect traffic. These factors need to be correctly included in the prediction network. Factors whose effects can be ignored in short-term forecasting systems become even more effective when long-term forecasting is desired. This

section mentions speed categories, weather conditions, spatial (regional) location effects, incidents and accidents that affect traffic speed, and time factors. In the sub-headings, how these factors are fed as input to the network is given in detail.

5.4.1 Speed Grouping Layer (L1)

It is possible to divide the road segments into slow and fast-flowing times and handle them individually. However, simply categorizing a road segment as fast or slow will not be enough. Based on this idea, we created the idea of producing models in four-speed categories for each segment. In its simplest definition, it became possible to produce predictions by using the speed category group to which the average speed of the segment belongs.

For example, let's say we want to predict the traffic speed of a segment for the next week. We can know the average traffic speed of this segment over the past four weeks. We can assign it to a speed group by looking at historical traffic speed averages. In this example, let's assume the average speed is 75 km/h and categorize it as Group-3, shown in Table 5.2. The group-3 model can produce successful predictions between 60 km/h and 80 km/h. We can use Group-3 to generate forecasts and get more accurate forecasts. This example was for the Test phase.

Table 5.2 Speed Group Bands

Speed Band (km/h)	**Group-1 Very Slow**		**Group-2 Slow**		**Group-3 Fast**		**Group-4 Very Fast**	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
110 - 120							x	x
100 - 110							x	x
90 - 100							x	x
80 - 90					x		x	x
70 - 80					x	x	x	
60 - 70			x		x	x		
50 - 60	x		x	x	x			
40 - 50	x	x	x					
30 - 40	x	x						
20 - 30	x	x						
10 - 20	x	x						

We have a strategy applied during the training phase of the model. Specifying a speed group with narrow ranges to train the model prevents the speed group from generating predictions outside of the ranges where it is assertive. However, we know that the predictive value may not fall within the range for which the speed group is

designed. To avoid this, the speed bands used in the training were superimposed on each other. For example, Group-3 is a model that can produce successful predictions between 60 km/h and 80 km/h, while data between 50 km/h and 90 km/h were used in its training. In addition, while determining which groups the data to be used in the training phase can be used to train, label (target) data were used instead of historical averages.

This method produces up to 4 forecast models to make a monthly forecast for a road segment. This means that up to 48 models can be produced in one year. However, if no data can be distributed to all four speed groups in a month, no models may be produced for some speed groups in some months. It is possible to generate 12 to 48 models in this scenario, and this number is dependent on the road segment data.

5.4.2 Weather Layer (L1)

Various weather conditions have slowing effects on traffic. Especially in harsh conditions, traffic slows down noticeably. Depending on the quality and quantity of weather conditions, these effects are variable and do not always show linearity. Some weather conditions may cause traffic to accelerate up to a particular value and slow down at later values. Effect-Rate (ER) and Polynomial-Regression (PR) have been proposed to accurately evaluate effects with such breakpoints. Afterward, the preprocessing method was put forward to feed this traffic speed-weather characteristic of the forecast model. This section explains how the weather is incorporated into long-term traffic speed forecasting.

5.4.2.1 Effect Rate Approach

In our first method, we propose to exploit an impact factor, i.e., tuning rate, to transform the traffic speed based on weather conditions for the given time. ER Pre-tuner function of raw traffic speed is calculated in Equation 5.4 where the variables are given in Table 5.3.

$$t(x) = x(1 + d_i e_s(w_f - w_t)/t_{max}) \tag{5.4}$$

The pre-tuner function $t(x)$ was utilized to calculate the possible acceleration or deceleration, which should be applied where the weather parameters diverge from normal conditions. As presented in experiments with various weather conditions, not all weather conditions affect traffic similarly. Therefore, an effect rate e_s coefficient is included in the system to control the tuning rate $t(x)$. The effect rate e

Table 5.3 Symbol Definitions

Definiton	Symbol
Weather Condition at Training Days	w_t
Weather Condition at Forecast Time	w_f
Raw Traffic Speed	x
ER Pre-tuner Function	t(x)
Weather Effect (Positive/Negative)	$d_i[+1/-1]$
Maximum Value of Weather Condition	t_{max}
Static Effect Rate of Weather Condition	$e_s[0.01, .., 0.12]$

can take a value between 1% and 12%, which is obtained depending on the average speed changes due to the weather condition. The function takes the difference of weather conditions on training day and forecast time values and proportions them into t_{max} value. This approach is linear and is intended for use in weather conditions that affect traffic flow linearly. Calculation examples are given in Figure 5.8 and Figure 5.9. Weather forecasts are used for the time to be estimated.

Figure 5.8 Example Speed Pre-tuning Process - 1

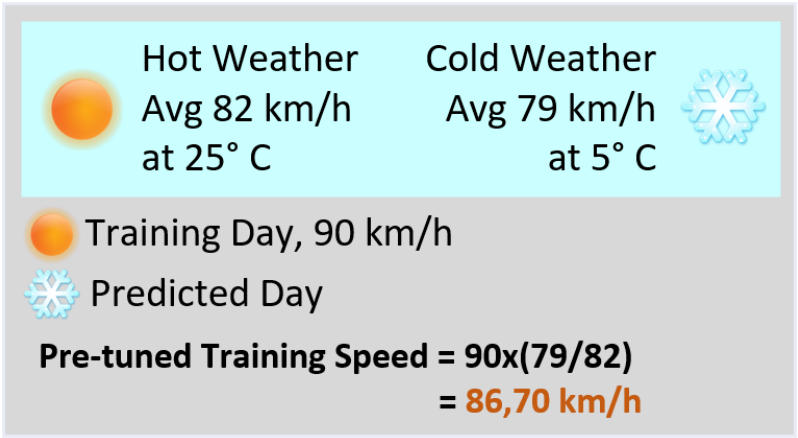
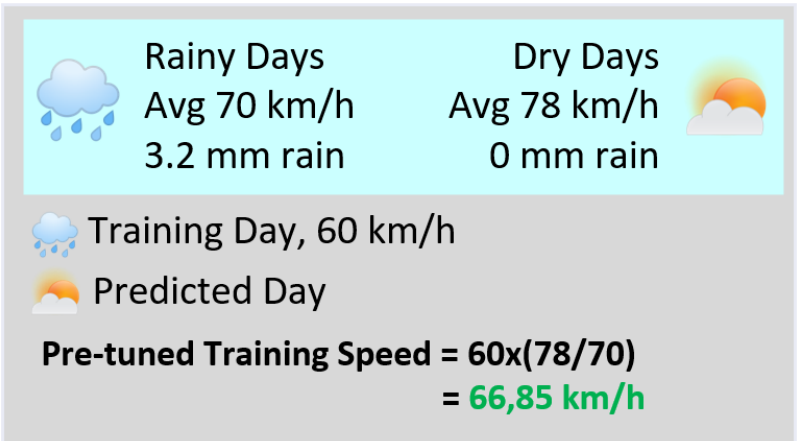


Figure 5.9 Example Speed Pre-tuning Process - 2



5.4.2.2 Polynomial Regression Approach

All weather conditions may not have a linear effect on traffic. In such cases, the impact on traffic speed must be adjusted dynamically depending on the weather conditions. Thus, we propose exploiting the polynomial regression-based method, which calculates the breakdown experienced at specific values. Polynomial regression aims at finding the degree of the polynomial and its coefficients for the given distribution of continuous data. Polynomial regression $p(x)$ function takes x value representing raw weather input and returns the average speed value of that weather condition. After that, when we put values on $t(x)$ in Equation 5.6, we get the required ratio of speed values and use it to pre-tune the corresponding speed.

$$p(x) = c_0 + c_1 x + c_2 x^2 ... c_n x^n \tag{5.5}$$

$$t(x) = x(p(w_f)/p(w_t)) \tag{5.6}$$

When the results of the regression curves according to degrees and the distribution of the real data are compared, curves with an appropriate degree that suits the given points best could be selected easily. Following this outcome, the preferred regression curve for each weather characteristic is unique. Figure 5.10, Figure 5.11, and Figure 5.12 demonstrate these characteristics of varying events.

Equation 5.5 and Equation 5.6 enable us to find the correct number of breakdowns (n). This method allows us to fit a polynomial regression model exploiting weather events that affect traffic characteristics. Test results showed that the 7$^{\text{th}}$-order and 9$^{\text{th}}$-order polynomial regression models gave the most successful results. At this point, speed data distributed to specific groups with lower and upper limits could also be used, but this would cause the outlier values between groups to be reflected in the model sharply. This would lead to a decrease in prediction performance. Since each weather parameter has a unique relationship with traffic characteristics, the preferred regression model for each is chosen individually.

5.4.2.3 Hybrid Approach

Weather-based models are expected to improve performance when severe weather conditions occur. The most potent models must be used together to use the suitable improvement model at the right time. In this hybrid weather-supported model, precipitation information, wind speed information, and temperature information;

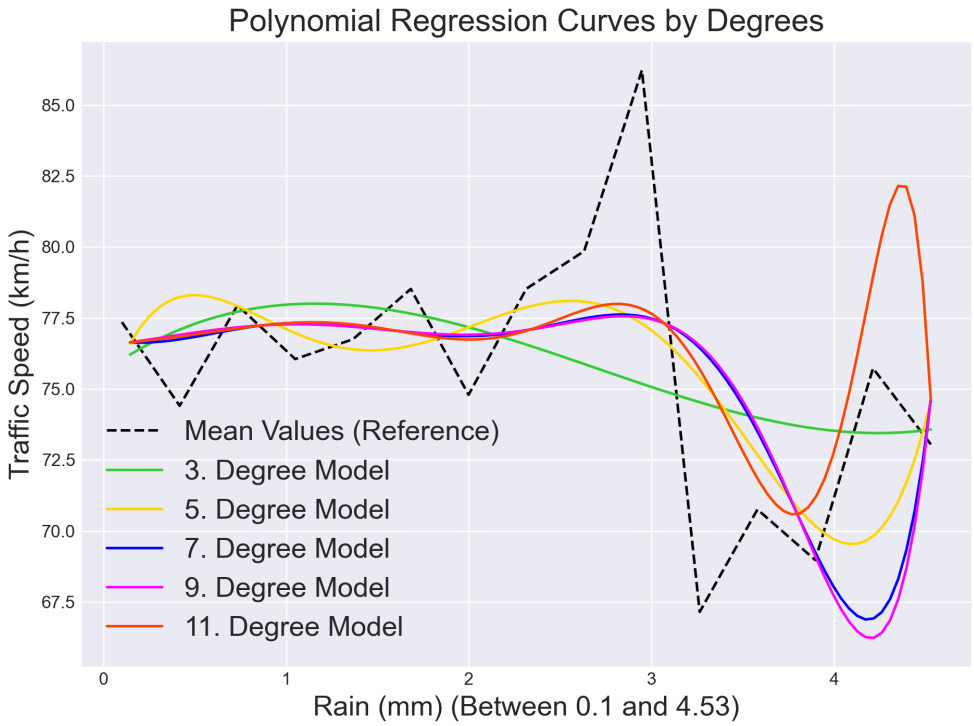


Figure 5.10 Polynomial Regression Models of Rainfall Data with Different Degrees

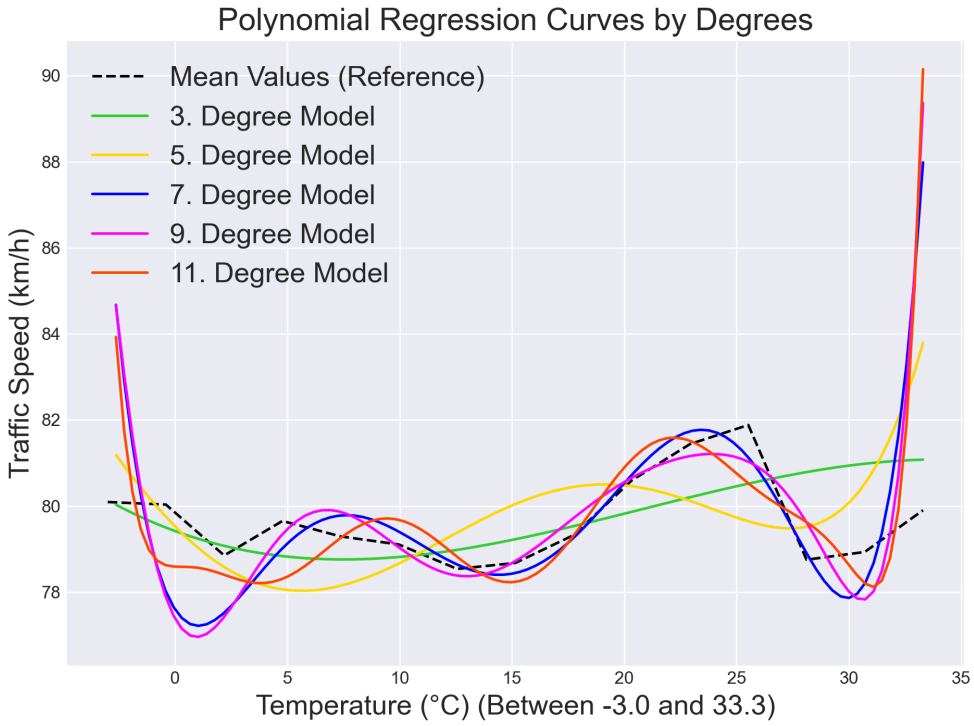


Figure 5.11 Polynomial Regression Models of Temperature Data with Different Degrees

are used together by a rule-based algorithm. Among the three weather conditions,

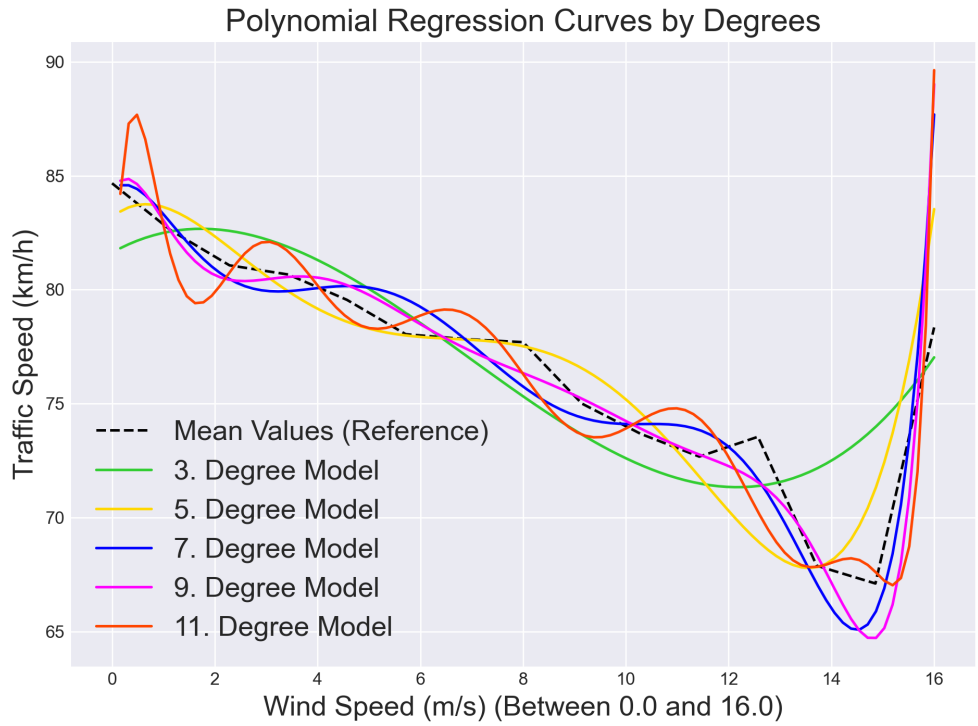


Figure 5.12 Polynomial Regression Models of Wind Speed Data with Different Degrees

the condition that is used the least and affects the traffic the most is the precipitation information. As seen in Figure 4.6 in the dataset analysis section, there is a decrease in traffic speed in cases with 1.5 mm or more precipitation. Based on this information, forecasts are taken from the precipitation-based model when this condition is met. The second-rank model is the one that makes use of wind speed. In cases where the wind speed is between 11 m/s and 15 m/s, the effect seen in Figure 4.7 occurs on the traffic speed. Finally, information is obtained from the model of temperature data, which effectively reflects seasonal characteristics throughout the year. Since the temperature model is the most wide-spectrum data that can be used among the other three models, it is placed at the lowest priority in the algorithm. As seen in Figure 4.5, temperature affected traffic speed to varying degrees throughout the year. The amount of cloudiness and visibility information examined were not used in the hybrid model as they did not provide sufficient improvement gains.

5.4.3 Spatial Layer (L1)

Using neighboring segment velocity data adds a new dimension to currently used data. A data matrix of scatter over the past k weeks and historical data goes up to 3, 5, and 7, depending on the number of neighbors included. In Figure 5.13, input methods use adjacent segment velocities at one hop distance. In addition to the

focused segment, the velocity data of the one-behind and one-ahead segments are used.

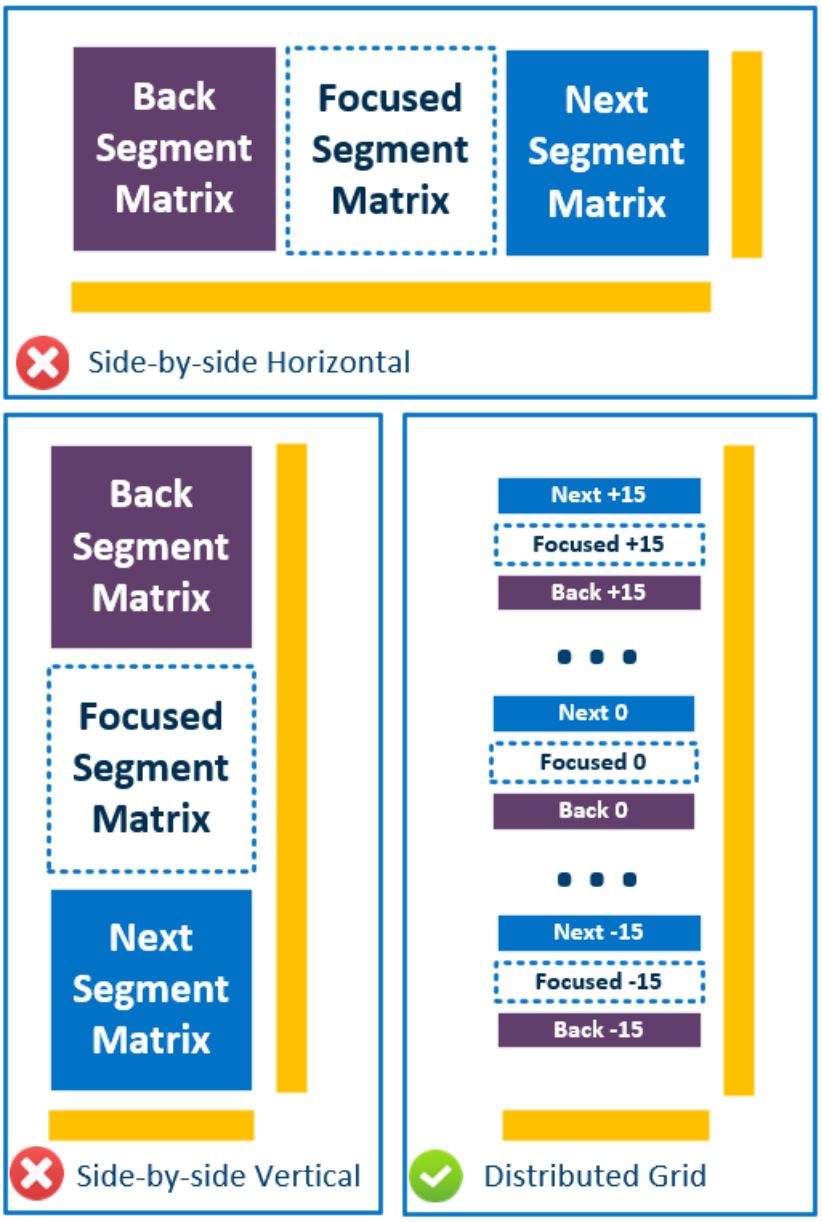


Figure 5.13 Spatial Information Feeding Methods

The first thing that comes to mind in this diagram is putting the three matrices side by side or under the other. However, juxtaposition approaches do not establish a neighborhood relationship between the past weeks of the segments. Therefore, the prediction model cannot establish a connection between the past speeds of the segments. The method detailed in Figure 5.14 is proposed to describe this relationship. In order, past distributions of the minute grouping (+15, +10 .. 0 .. -10, -15) were made on past speeds, which continued in the form of minute distributions. In this way, the traffic speeds of all segments for the appropriate time are positioned to be close to each other. Thus, the prediction model can evaluate the region and

make sense of temporal groups.

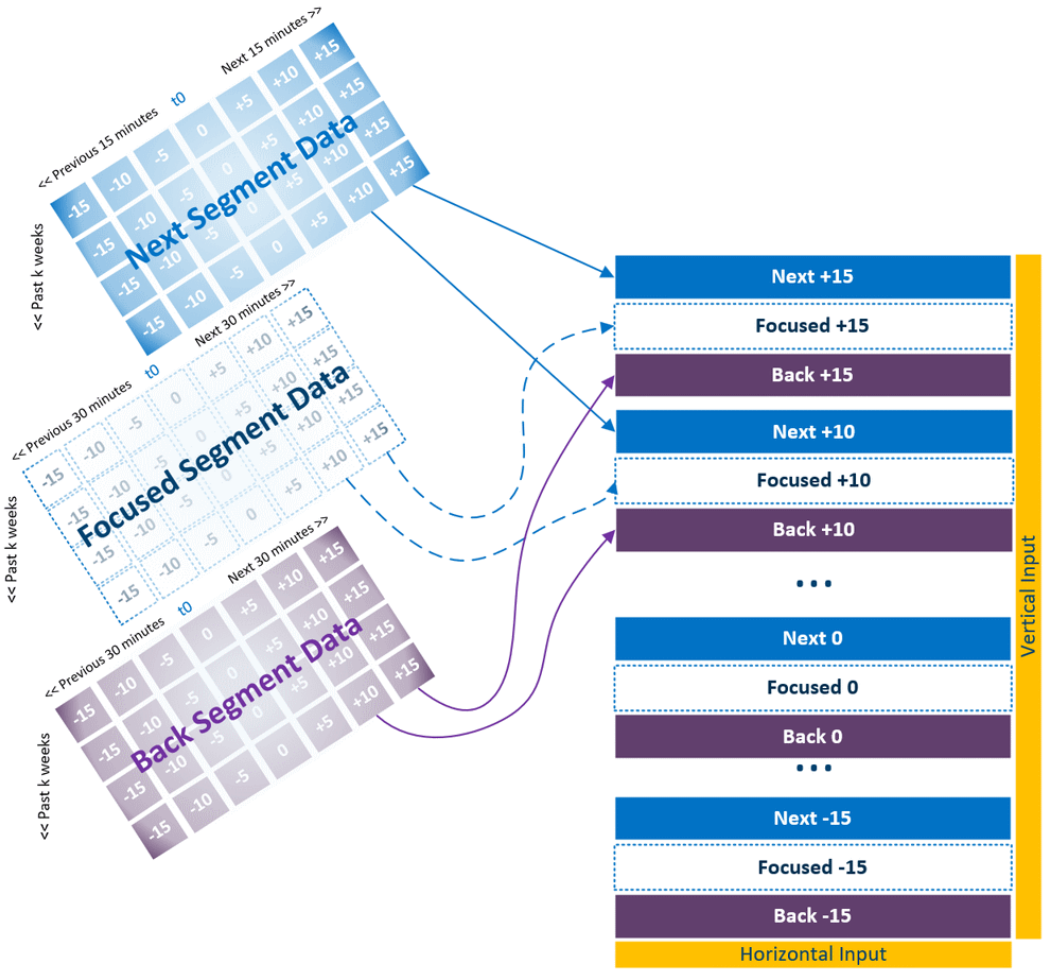


Figure 5.14 Spatial Information Grid Distributed Input Feeding Schema

5.4.4 Couple-Integrated Hybrid Models (L2)

Models that provided an improvement over the base model were named L1. When both improvement methods are used together, more advanced models called L2 are presented. Although it seems possible to use two technics independently under normal conditions, some technics may conflict with each other due to their nature. We focused on three technics: speed groups, weather, and spatial data usage. L2 models were implemented with the combinations of these models as C(3,2).

5.4.4.1 Conflicted Technics and Solutions

The weather model included in the system as a preprocessing method was the least conflicting method with other methods. Receiving data from a singleton weather source prevented data overlapping in the usage of spatial data. In addition, pre-tuning according to the weather conditions before the speed groups distribute

the seasonal characteristics to the correct speed groups. With all these advantages, the weather contributed to the system as the most independent method.

The main difference in the speed groups model is that four models are produced in 1 month. This affects the final output. Therefore, the question of choosing the correct speed-group model arises. For this, the historical speed averages of the segment to be predicted are taken into account. The problem has been in the use of spatial data. The contribution of the speed mean of the neighboring segments may lead to a decrease in success in the L2 model, which includes a spatial and speed group. For this reason, it was considered to use only the historical speed data of the relevant segment in the selection of the speed group model.

5.4.5 Powerful (All-in-one) Model (L3)

In the level 3 layer, three elements were used together to increase the performance in traffic speed estimation. It was decided to use the most successful models of each segment within the scope of the Powerful model. Thus, it has been ensured that the features that provide the highest performance regionally contribute to the performance by using the minimum processing power.

5.5 Post-Prediction Improvements (L3+)

The performance increases provided using post-prediction improvements named Level 3+ have been examined. Transactions made in this layer do not directly affect the traffic speed estimation. Structures that increase prediction performance by tuning depending on the environmental conditions of the predictions produced by the traffic speed prediction module are in this scope.

5.5.1 Accidents & Events Speed Adjuster

It is a probabilistic process to predict that a traffic accident will occur in a specific time interval on the road. Modifying the traffic speed estimation assuming an accident will reduce the estimation performance on flowing traffic. Since calculating the probability of a traffic accident occurring is not the main subject of this thesis, it has been evaluated how the traffic speed will be affected in the traffic accidents that occur. Using the accident and traffic incidents dataset, it was analyzed how the region was affected by the incident using the average traffic speed of the past four weeks. In light of this information, it has been seen that the coefficient of influence of each road segment from the events is different. Tuning was performed

on the traffic speed estimation using the accident and traffic incident information kept in the segment-specific time interval format in the dataset.

5.5.2 Legal Speed Limits Adjuster

It is vital for any end user who will use this data that the artificial intelligence model, which makes traffic speed estimation, produces predictions within the legal traffic speed limits. It is known that the 80 km/h speed limit valid in Istanbul is valid on the main roads where the traffic speed is estimated. Within the scope of the legal traffic speed limits, this speed limit can only go up to 88% km/h with a 10% tolerance in case of overtaking. Limiting this speed to 80 km/h would have been an option in preprocessing the dataset studied. However, this limitation was not carried out in the directly studied dataset in order not to affect the decision mechanism of the trained forecasting model. Instead, if needed, revising the estimate by post-processing the estimated data was deemed appropriate.

5.5.3 Prediction Resolution Changer

When the traffic speed estimation is made for the long term, estimating high resolution is not a desirable feature in the first place. It is preferred if prediction performance can be improved with an acceptably lower resolution. In a real-world usage scenario, a user will prefer to plan with a precision of 15 minutes, which will be more reliable, instead of planning with a precision of 5 minutes in the user's travel plan for the next week. With this point of view, we changed the estimation resolution from 5 minutes to 15 minutes in our model, where we make long-term traffic speed estimation for up to a week. We produced new forecasts for the 15-minute time interval using the forecasts before and after the 5-minute forecast. We have designed this as a module so that the predictions that come out of the prediction model with high resolution can be modified in case of need and to increase the success.

6
DEVELOPMENT ENVIRONMENT

Within the thesis scope, hardware with parallel processing power was used to process big data. General-accepted software and reliable hardware were preferred to minimize long model training times and sustain the software development process. This section explains the technical details of hardware and software resources. The difficulties experienced in this process, the model training processes, and the experience of software development workflows are also explained in this section.

6.1 Software Resources

The software has been used for various needs such as data analysis, artificial intelligence model training, recording experiment results, storing and backing up big data, statistical evaluation of the results, producing graphics for visual presentation and report preparation, and using map images. Ready-to-use software is given in Table 6.1. In addition, more than 20 Parser and Analyzer software has been developed with Python, and the test results have been automated and evaluated. Table 6.2 gives all libraries used in this scope.

Table 6.1 Software Preferences

Specification	**Definition**
Main Programing Language	Python 3.9.10
Coding IDE	Jetbrains Pycharm 2023.1.1
GPU Acceleration	CUDA 11.3
Schema Drawing Tool	Microsoft Visio
Satellite Images and Maps	Google Earth
Text and CSV Edit Tool	Notepad++

Table 6.2 Python Libraries

Library	Definition
Numpy	Provides high-level mathematical functions.
Scipy	Provides scientific computing and technical computing.
Pandas	Provides tools for data manipulation and analysis.
Tensorflow	Provides tools for machine learning and artificial intelligence.
Keras	Provides tools for high-level deep learning with tensorflow.
Anytree	Provides lightweight tools for Python tree-data-structure.
Plotly Go	Provides graph and map drawing tools for Python.
Tkinter	GUI toolkit
Pillow	Python Imaging Library for saving images and maps.
Base Libraries	Basic functionality for read/write, asynchronous tasks etc.

6.2 Hardware Resources

The thesis uses MSI brand GP76 at 6.3 as a personal computer. A laptop computer is preferred, especially in short-term experiments and analysis phases. Long-running experiments have been completed on the HP Z4 Workstation at 6.4. CUDA software has been activated on both devices to utilize two powerful graphics cards.

Table 6.3 Hardware Specifications - MSIGP76

Specification	Definition
Brand / Model	MSI GP76 Leopard 11UG-205XTR (Laptop)
Processor (CPU)	Intel® Core™ i7-11800H, up to 4.60 GHz
CPU Cores	8 Cores, 16 Threads
CPU Caches	24M L3 Cache
GPU	NVIDIA GeForce RTX 3070 8 GB Mobile
CUDA Cores	5120 Cores, 160 Tensor Cores, 40 RT Cores
Memory (RAM)	32 GB DDR4 3200 Mhz (2 Channels)
Storage	2 TB Samsung PCIe 4.0 NVMe™ M.2 SSD
Operating System (OS)	Windows 11 Education
Acceleration Technologies	CUDA, OneAPI, AVX, AVX2, AVX512

6.3 Software Development Workflows

The forecasting model development process needed a development infrastructure with interactive features. Software development was carried out with modular approaches in the Python development environment to achieve this. Developments continued with an agile approach, adhering to SOLID principles. As a result of planned software development studies, results were produced within acceptable time limits by giving minimum impact to previous studies. This section contains detailed information on solutions to encountered errors and performance improvements.

Table 6.4 Hardware Specifications - HP Z4 Workstation

Specification	Definition
Brand / Model	HP Z4 Workstation (Remote Accessible)
Processor (CPU)	Intel® Xeon™ W-2255, up to 3.70 GHz
CPU Cores	10 Cores, 20 Threads
CPU Caches	19M L3 Cache
GPU-0	NVIDIA RTX A5000 24 GB
GPU-1	NVIDIA RTX A5000 24 GB
CUDA Cores	8192 Cores, 256 Tensor Cores, 64 RT Cores
Memory (RAM)	128 GB DDR4 2933 Mhz (2 Channels)
Storage	500 GB Sata SSD + 1 TB Sata HDD
Operating System (OS)	Windows Server 2019
Acceleration Technologies	CUDA, AVX, AVX2

The long training times are the most challenging aspect of working with AI prediction models. Even a tiny mistake can cost days, as the detection of an error occurs after the training of the entire model is completed. The best solution to prevent this situation is to shorten the training periods. Since we work with multiple road segments, dividing a large job into sub-small parts is possible. A development has been made to automatically divide the work into two to work on 2 Nvidia RTX A5000 graphics cards that we use in our development process. Since two graphics cards can work independently, it has directly accelerated the work by two times. Even in this case, training a prediction model can take over two days.

Parallel processing power is needed to train neural networks. Training of artificial neural networks cannot be completed in a reasonable time using computer processors that offer serial processing power. Therefore, Nvidia RTX A5000 graphics cards with CUDA and Tensor cores are preferred. Some trials have been made using the Intel OneDNN Library to evaluate the alternatives to the graphics card and to see the performance of the artificial neural network training made with the processor. According to the results, it was decided to continue with the graphics card. The performance values obtained can be examined in the Experimental Results section.

Model training is wasted in case of power outages or hardware failures. To avoid this, a UPS Electric source was used. In addition, writing to disk and backing up of completed models have been automated. The system can continue from where it left off the next time it is online. MSI laptop and Apple Macbook Pro M1 were used as an alternative when the graphics card usage schedule was stuck. This required creating a file-sharing system that allows devices to access shared storage. Traffic speed data and weather data are served over a device, and a structure has

been established in which all devices will write only the results they produce in the common area due to the operations performed. With this approach, artificial neural network training times are shortened by using distributed system architecture.

The most significant factor that extended the model training in the experiments with the weather conditions was using the weather data by calculating one by one. As a solution, a cache structure was built on the memory, and once access was provided, weather data was obtained from this cache, if available. With this approach, experimental times are shortened by approximately 40%. The use of data that needs to be calculated repeatedly by bringing it from the database is one of the biggest obstacles to performance. Depending on the volume of data, a choice has to be made between caching and solid state drive or dynamic memory. Retrieving data using a well-organized micro-service network in a real-world use case is also possible. However, even in this scenario, choosing file-based caching structures will provide significant performance gains over database usage.

7
EXPERIMENTAL RESULTS

In this section, the analyzes of the experiments performed are given. We conducted multiple experiments using a vast amount of data with various features to enhance the estimation of traffic speed. In the first part, the concepts of MAE and MAPE that we use for error measurement are explained. In the second part, the performances of the base prediction models are evaluated. Then, the performances of the weather, spatial data, and speed grouping techniques that are part of the multi-feature space are evaluated. Finally, the results of the Powerful model approach have been presented.

7.1 Error Metrics

Two error metrics, MAE and MAPE, were used to express the numeric errors in the experiments. Error calculations were made over 5-minute period estimates in all experiments.

7.1.1 MAE: Mean Absolute Error

MAE is calculated as the sum of absolute errors divided by the sample size. With this aspect, in cases where MAE is used, the error value grows only as much as the difference. The magnitudes and ratios of the two values compared are ineffective. Thesis-specific use of MAE is beneficial in seeing places with a fixed s km/h estimation error. Figure 7.1 gives the mathematical equation for calculating MAE.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}.$$

Figure 7.1 MAE: Mean Absolute Error

7.1.2 MAPE: Mean Absolute Percentage Error

MAPE is calculated as the absolute value of this ratio is summed for every forecasted point in time and divided by the number of fitted points n in Figure 7.2. Thus, the error in the estimation can be seen as a percentage. MAPE is often used to show the error rates of prediction models. However, some parts need attention when examined through sample values. When we use MAPE, the absolute value between the values is used, and their ratio becomes essential. When considering this, a notable contrast will exist between the MAPE value computed for speeds of 10-16 km/h and the MAPE value calculated for speeds of 70-76 km/h. This means that mistakes in estimating at lower speeds will seem more significant.

$$\text{MAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

Figure 7.2 MAPE: Mean Absolute Percentage Error

7.2 Base Model Performances

This section gives the results of the base experiments with Polynomial Regression, CNN, LSTM, and GRU models. In the experiments with base models, no data was used except the traffic speed. These experiments aim to select the best model that captures the traffic pattern. In Table 7.1, MAE and MAPE values of 4 models are given.

Table 7.1 Base Model Results

Model	Avg. MAE	Avg. MAPE▽
Polynomial Regression	8.67	16.51%
CNN	7.27	13.81%
LSTM	7.09	13.45%
GRU	6.89	13.23%

The hourly results obtained in the experiments with four models are given in Figure 7.3 and Figure 7.4, the results according to the day of the week are shown in Figure 7.5 and Figure 7.6, and the monthly results are given in Figure 7.7 and Figure 7.8. Night hours are the hours when the traffic flows fast and steadily. It is possible to predict nighttime without a need for high-resolution forecasting. Among these experiments, the model expected to make the lowest accrued estimation is the Polynomial Regression. The results were acceptable, particularly from 00:00 to 4:00.

Similarly, it is shown that there are improvements in the forecast between 09.00-12.00 when the traffic flow becomes stable during the day. Unfortunately, Polynomial Regression does not provide reliable predictions for traffic situations after 16.00. The performance levels are not acceptable. It produced 5% worse results than its closest competitor, CNN. These results show that the Polynomial Regression method provides acceptable results for easily predictable times but produces poor results when the traffic flow slows down and becomes uncertain.

It is known that the CNN model is highly susceptible to noise. CNN is powerful in capturing regional relationships in the fed data. In this way, CNN had similar results to LSTM most of the day except during rush hours. The LSTM and GRU model produced the best results for most of the day. This is because LSTM and GRU are less affected by noise and are more successful in estimating time series. Although both models are widely used in deep learning, GRU is known for having a lighter architecture. Compared to LSTM, GRU has been shown to have slightly better results.

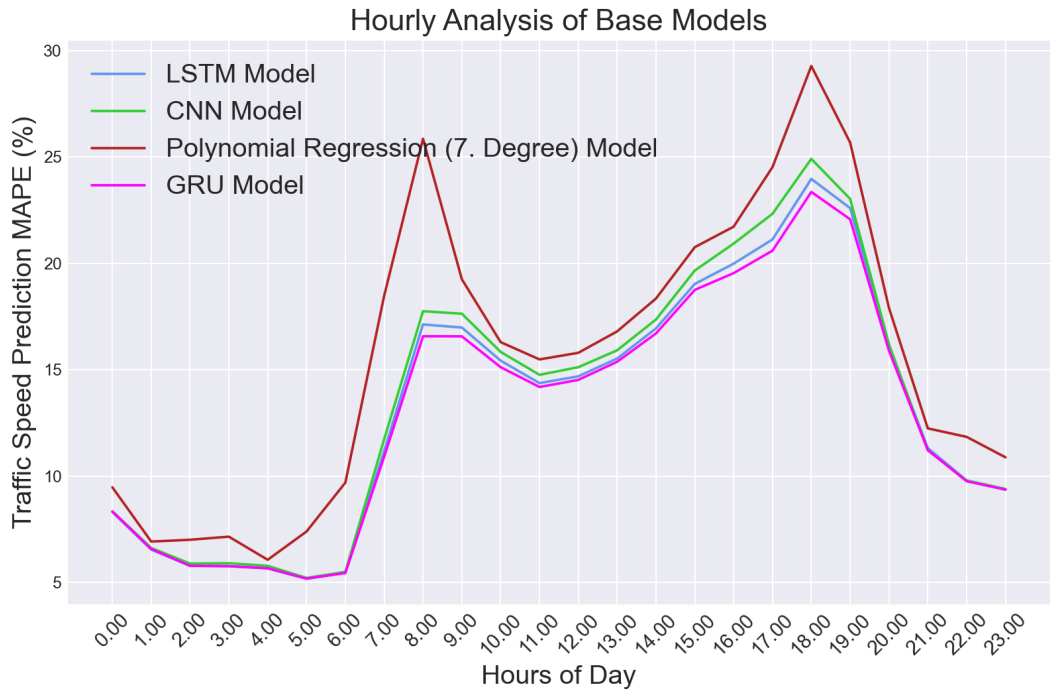


Figure 7.3 Base Models HOD Comparison (MAPE)

The days of the week are characteristically separated from each other. Sunday is the easiest day to predict for all models, while Friday is the most challenging. Typically on Sundays, the traffic moves swiftly and consistently. When all days are examined, it is seen that the GRU model delivers the best estimate by far. Although the LSTM forecasting model produced forecasts close to the GRU, it lagged behind every day of the week.

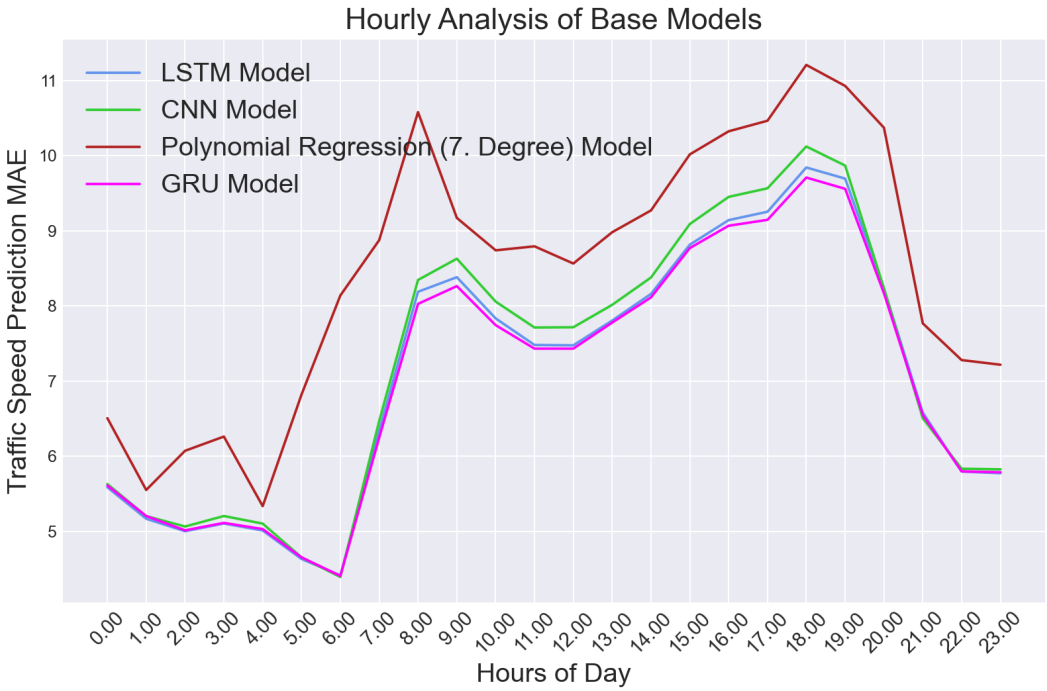


Figure 7.4 Base Models HOD Comparison (MAE)

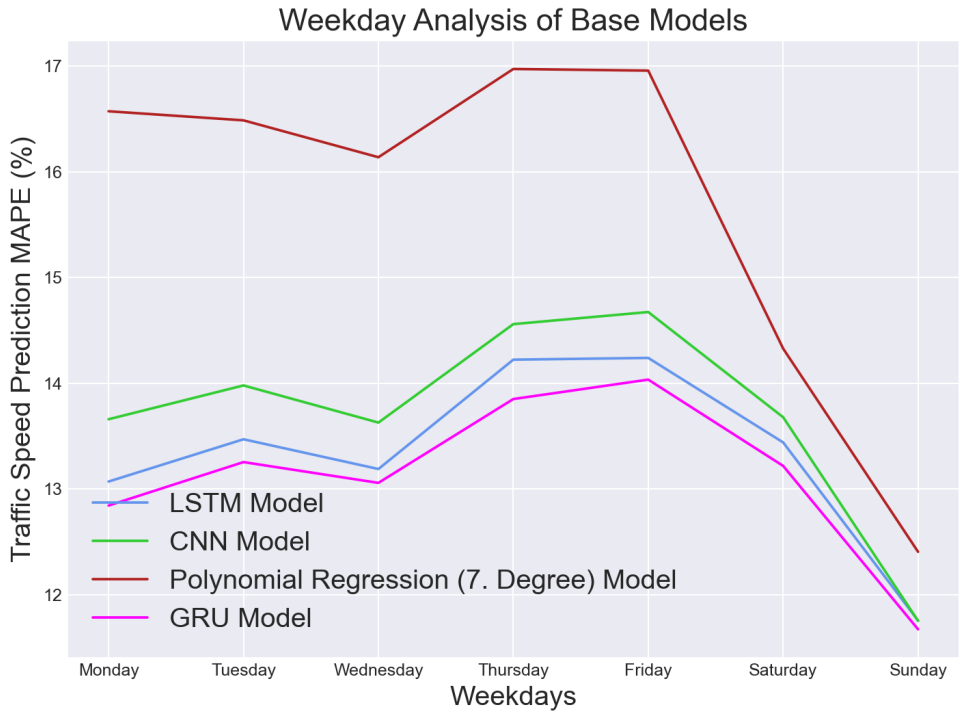


Figure 7.5 Base Models DOW Comparison (MAPE)

When the monthly results are analyzed, it is seen that the models had difficulties in September. This situation can be attributed to the characteristic change of the population arising from the migration movements in the transition from August to September. When the results were examined, GRU gave the best results for each

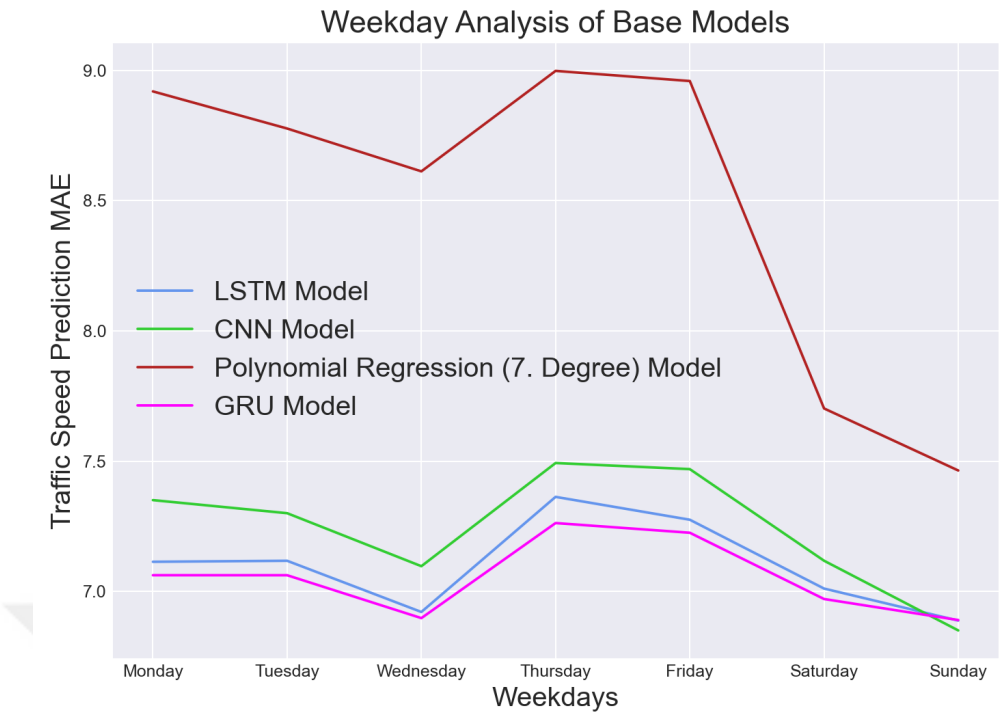


Figure 7.6 Base Models DOW Comparison (MAE)

month. The LSTM model produced results close to the GRU model except for September. The most significant difference of the GRU model was the 1.4% more successful results in September.

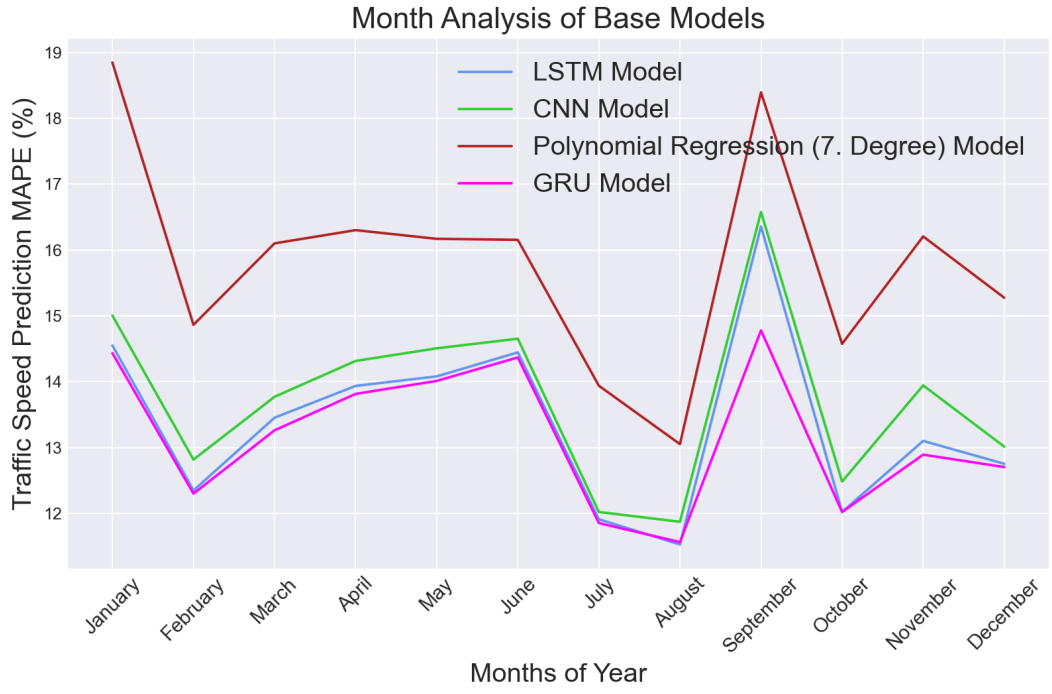


Figure 7.7 Base Models MOY Comparison (MAPE)

When MAPE and MAE results were analyzed in the same category graphs, we

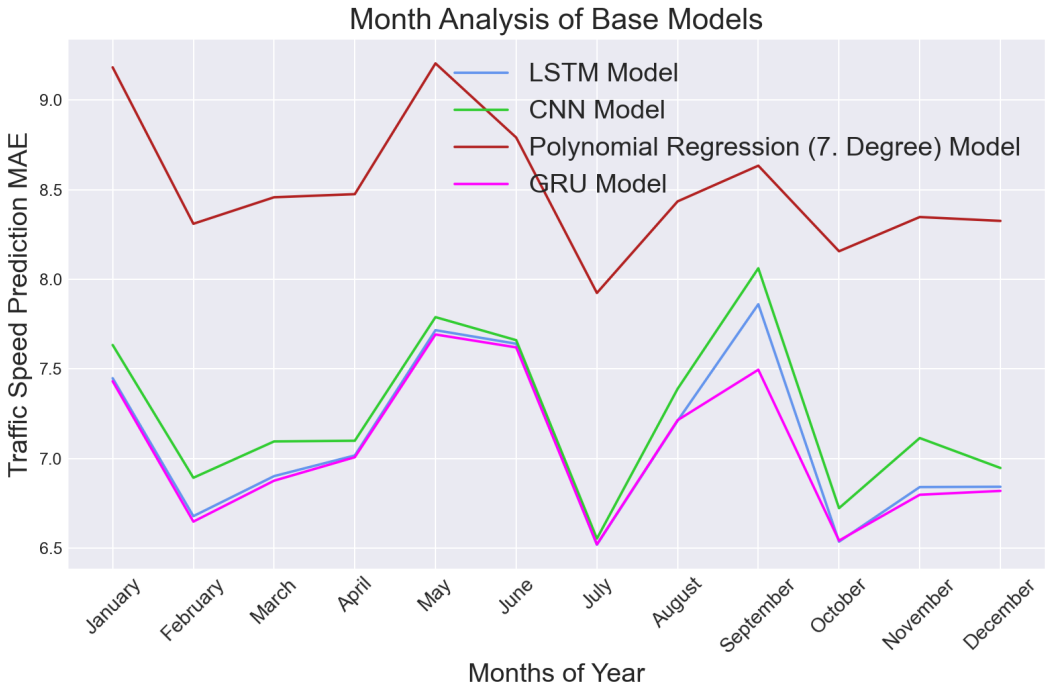


Figure 7.8 Base Models MOY Comparison (MAE)

observed some different results. In the graphs, when the fast traffic periods are examined (such as night hours and summer months), it is seen that MAE results are dramatically worse in these regions. MAPE results are better compared to MAE. This is because the MAPE (calculated by the dividing method) shows the errors of high speeds lower.

7.3 Multi-Feature Space Prediction Model Experiments

To evaluate the effectiveness of multi-feature-based methods, various experiments were conducted by integrating new techniques into the base model. This section delves into analyzing the impact of weather conditions, speed grouping, and neighboring roads on the forecasting results.

7.3.1 Weather-Assisted Models

Weather models aim to improve traffic speed prediction when the relevant weather condition is challenging. Therefore, the analysis was made on condition-specific quantities. Celsius for temperature, mm for precipitation, m/s for wind speed, percentage for cloudiness, and m for visibility distance are used.

This section examines the models trained for temperature, rainfall, wind speed, cloudy weather, and visibility. In this analysis, Effect Rate (ER Tuner) and

Polynomial Regression (PR Tuner) results are given under the same graph. Thus, it is shown which of the two approaches is more powerful.

7.3.1.1 Temperature-Assisted Model

In this model, temperature data is used to assist the prediction success. Temperature data reflect seasonal characteristics. In this aspect, the temperature can train the estimation network for speed differences between seasons. When the graph in Figure 7.9 is examined, it is seen that there is an improvement in estimation at all temperature levels. It has been observed that the success rate of forecasting in cold weather has improved from 13.5% MAPE to 12.8% MAPE. The performance increase is maximum in scenarios where the temperature drops below 15 °C. It has been observed that the performance increase gradually decreases from 30 °C to 35 °C. This is because temperatures above 30°C are only seen during the day and in summer. The limited data in this region limits the increase in the performance of the prediction.

Test results of ER pre-tuner showed that the maximum performance was obtained with an effect rate of 9%. PR Pre-tuner, on the other hand, gave the most successful results at the 9th degree among all polynomial regression degrees. The results obtained with the PR Pre-tuner are promising as it increases performance over a wider period of time.

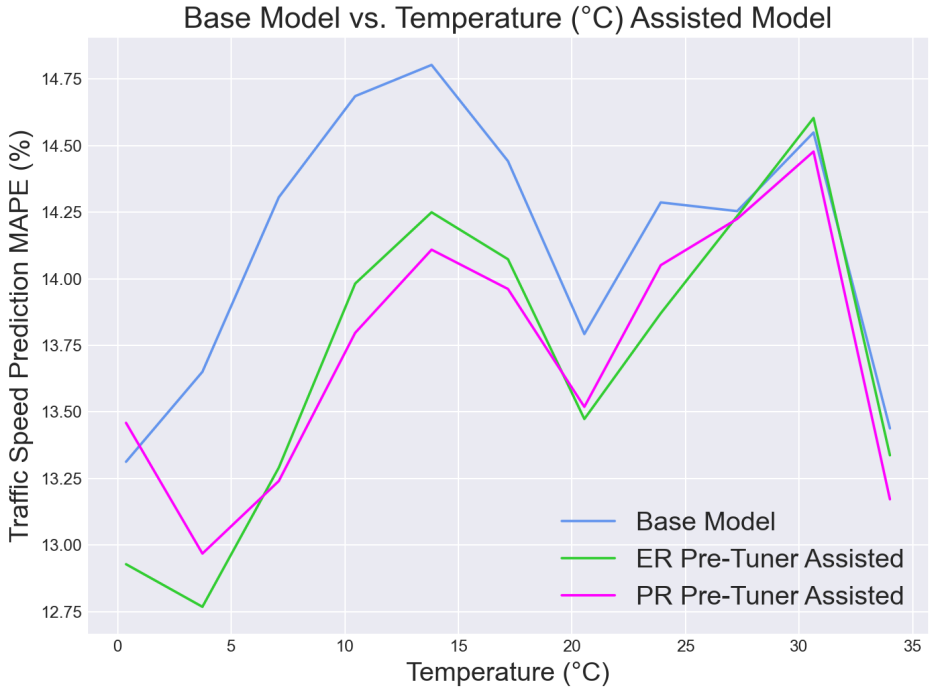


Figure 7.9 Temperature-Assisted Model Performance

Figure 7.10 and Figure 7.11 show the road regions where the temperature model provides the most performance increase. Unlike other weather conditions, there is no significant difference in forecast improvements on a segment basis, as temperature increases performance throughout the year. Although in small quantities, success in all segments improves the general average. When the map is examined, it is seen that the segments with the highest performance increase spread throughout Istanbul.

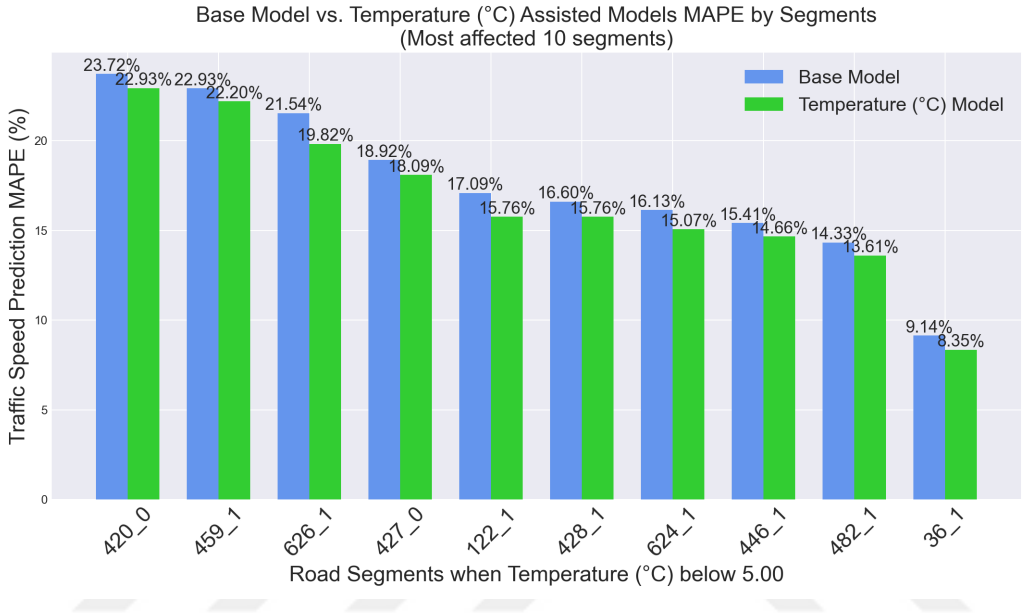


Figure 7.10 Temperature-Assisted Models Performance over Segments

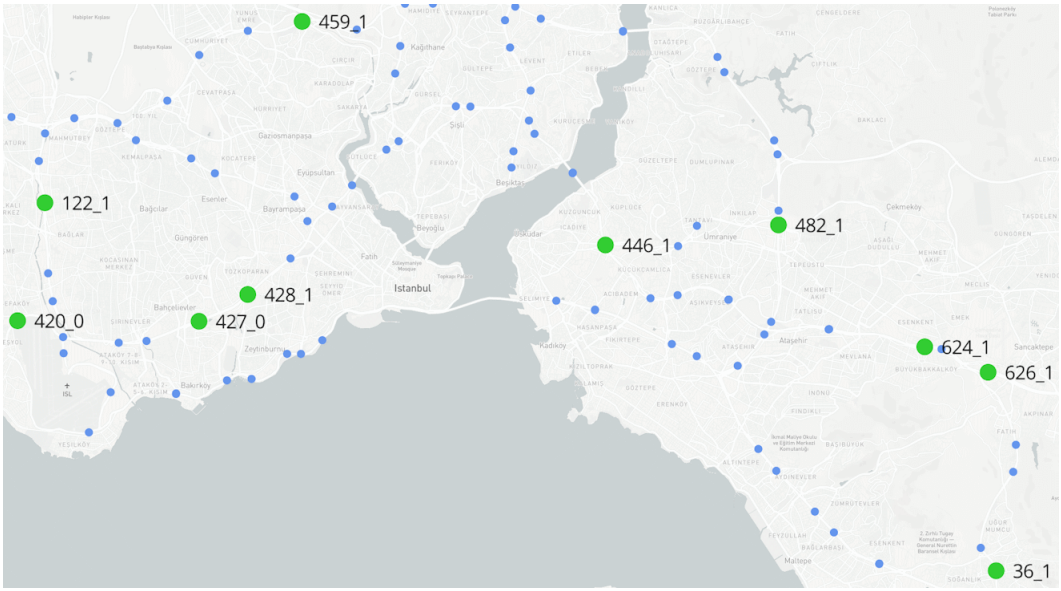


Figure 7.11 Top 10 Most Affected Segments - Temperature-Assisted Model

7.3.1.2 Rainfall-Assisted Model

In society, it is commonly known that traffic slows down during rainy weather. In this model, rainfall data is used to assist the prediction success. It was observed that the forecast performance increased from a 22% MAPE to 18% MAPE, mainly when precipitation levels were at 3 millimeters or higher.

Both pre-tuning methods improve the estimation accuracy and increase the performance of the models in which the data is used in training. Accordingly, a continuous improvement in forecast performance has been observed, from the absence of precipitation to the maximum amount of rainfall seen in Figure 7.12. Test results of ER pre-tuner showed that the maximum performance was obtained with an effect rate of 12%. PR Pre-tuner, on the other hand, gave the most successful results at the 9th degree among all polynomial regression degrees. The ER Pre-tuner results are promising as they provide better results in case of heavy rain, starting from 1 mm and above.

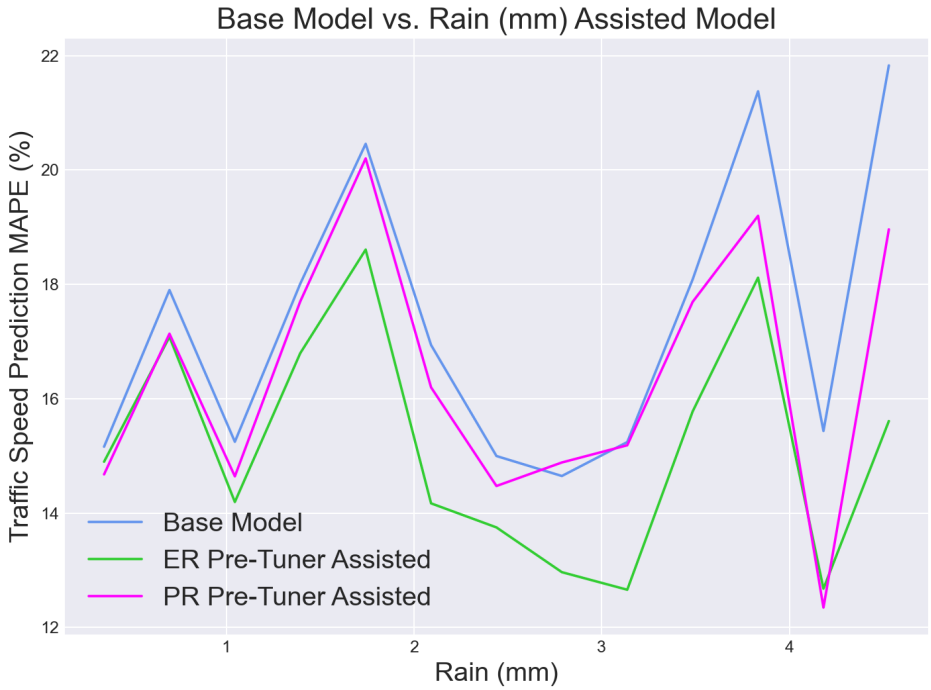


Figure 7.12 Rainfall-Assisted Model Performance

The road segments with the most varying forecast performance at precipitation levels of 3.5 millimeters and above are given in Figure 7.13. The use of the precipitation data has improved even in segments with more than 100% MAPE ratio. This indicates that one reason for poorly predicted road segments is precipitation. Similarly, the success rate of a segment that can be predicted with a MAPE value of 13.70% is improved up to 6.95% MAPE. It is seen that the biggest problem in segments like this is the changing characteristics during precipitation.

Istanbul is a city that receives precipitation throughout the year. The road segments with increased performance in the precipitation data estimations are given in Figure 7.14. When the map is examined, the forecast improvements, especially in the road areas close to the Bosphorus bridges, are noteworthy. However, forecasts made in various parts of the city have been improved using precipitation data.

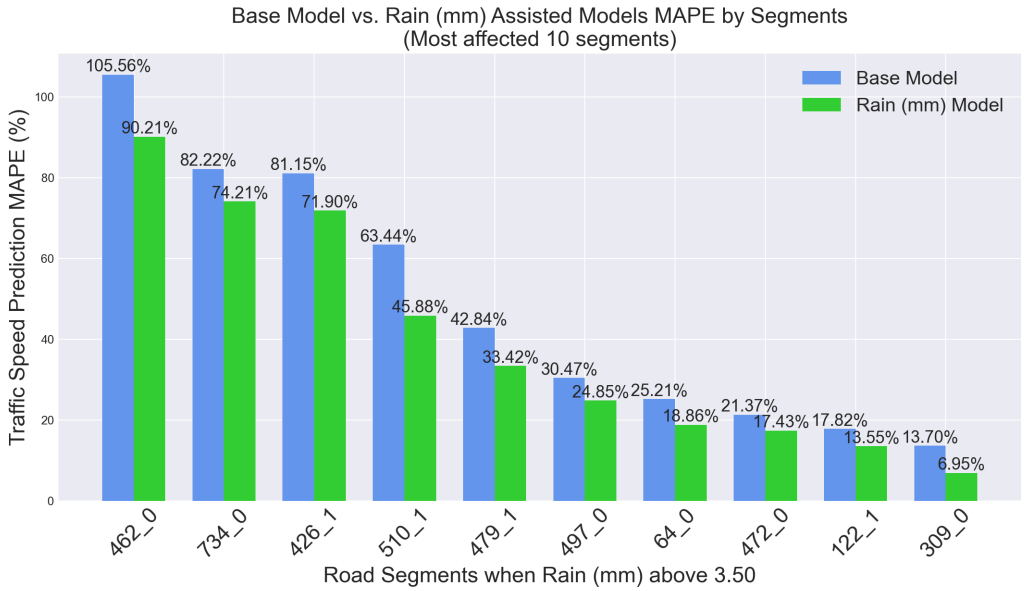


Figure 7.13 Rainfall-Assisted Models Performance over Segments

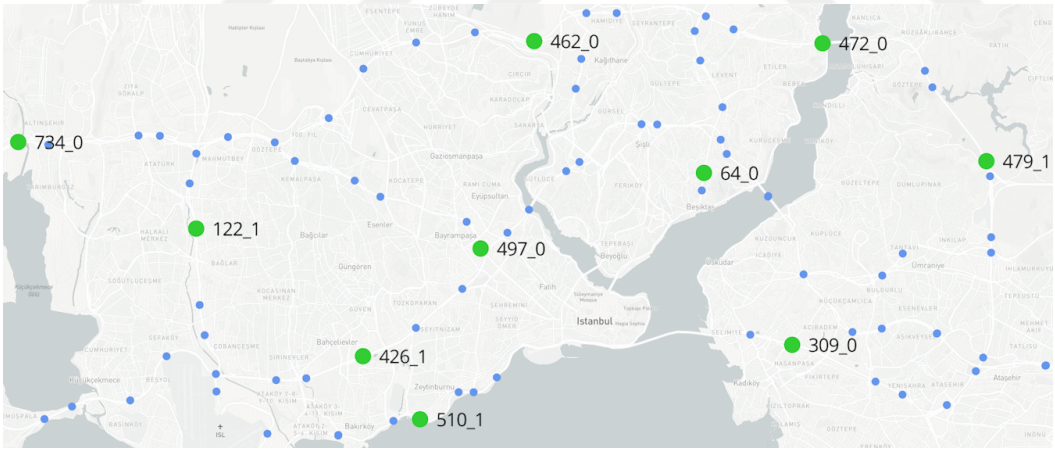


Figure 7.14 Top 10 Most Affected Segments - Rainfall-Assisted Model

7.3.1.3 Wind-Speed-Assisted Model

Istanbul is a city with a sea passing through the middle and surrounded by the sea on both sides. The temperature differences between land and water cause winds coming from various directions. The wind on highways indirectly affects traffic speed. Side winds have a severe effect, especially in large vehicles with closed bodies.

In this model, wind speed data is utilized to improve the accuracy of predictions. The graph in Figure 7.15 shows that traffic speed prediction accuracy significantly improves when wind speeds exceed 11 m/s, from 20% MAPE to 16% MAPE.

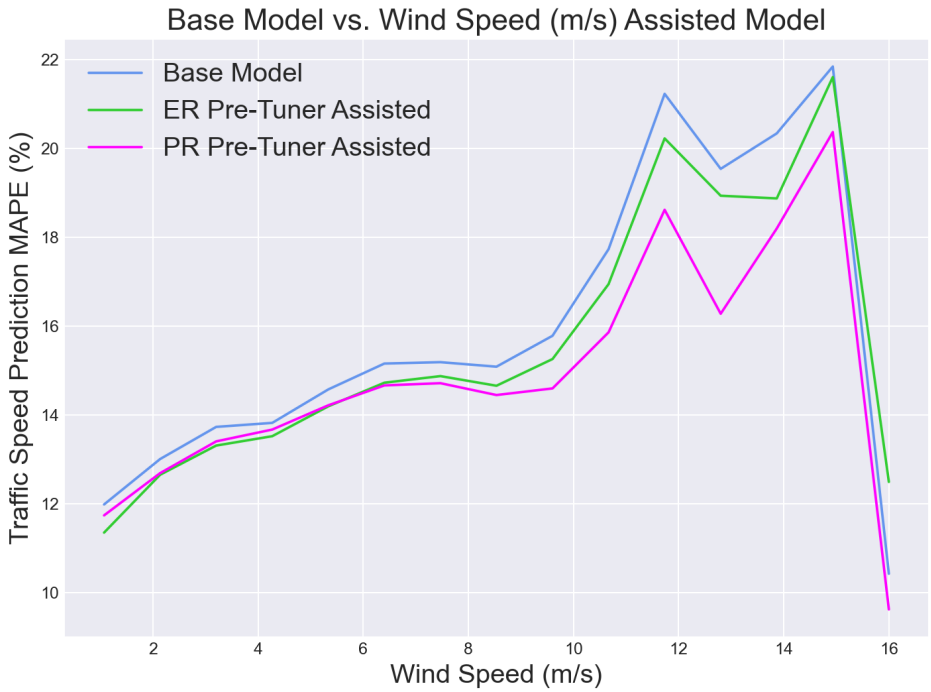


Figure 7.15 Wind-Speed-Assisted Model Performance

In winds above 8 m/s, success increases were observed between 1% and 2% on the segments in Figure 7.16. When the map given in Figure 7.17 is examined, it is seen that the prediction success of the segments located in the southwestern part of Istanbul is increased by using wind speed. Notably, the segments with the maximum prediction performance increase are part of a road extending in the North-South alignment. This road is more open to Winds from the west and east. Based on this information, it can be said that the Istanbul Basın Ekspres Road is one of the highways most affected by the wind.

The ER pre-tuner test results revealed that the highest performance was achieved with an effect rate of 1%. Meanwhile, the PR Pre-tuner produced the most favorable outcomes at the 9th degree compared to all other polynomial regression degrees. These results show that the effect of wind speed on traffic speed is not more than 1%. Although it contributes little besides temperature and precipitation data, the contribution of the PR Tuner method is valuable in terms of system architecture in stormy weather.

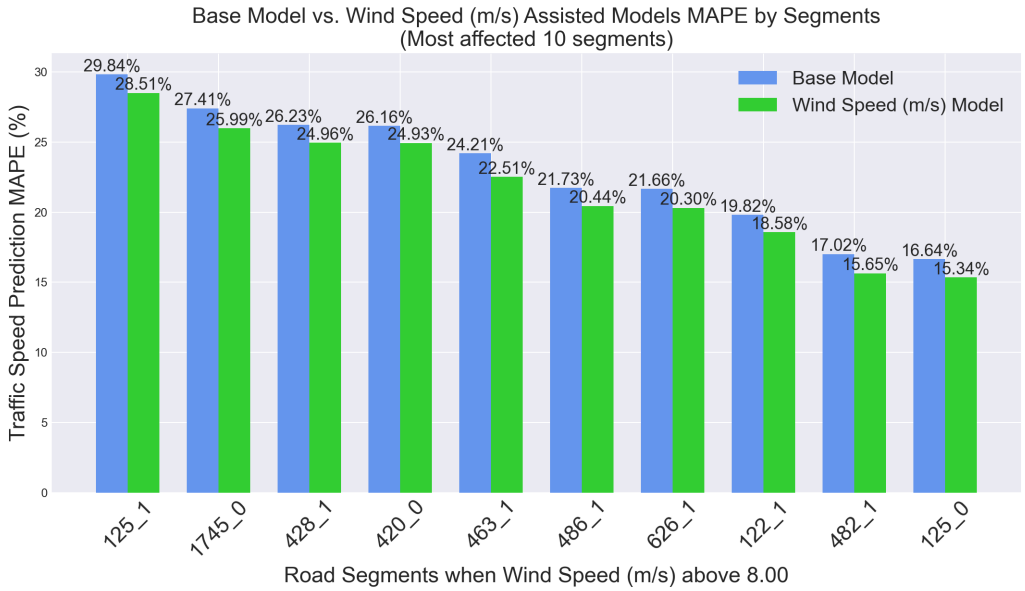


Figure 7.16 Wind-Speed-Assisted Models Performance over Segments

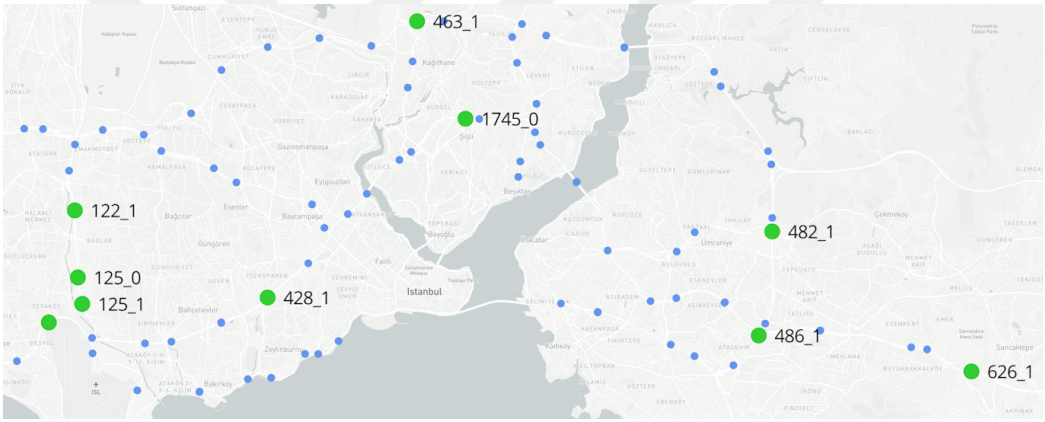


Figure 7.17 Top 10 Most Affected Segments - Wind-Speed-Assisted Model

7.3.1.4 Clouds-Assisted Model

Minimal performance improvement has been observed in forecast improvements using the amount of Clouds in Istanbul. In the ER pre-tuner results given in Figure 7.18, the effect rate is 6%. PR Pre-tuner results are shown over the 9th degree. When we analyzed the graph, it is noticeable that there is a performance increase of 2.5% between 10% and 20% clouds cases but not in others. The amount of clouds alone is not a weather condition that can affect traffic. So, evaluating this change directly with the amount of cloudiness is not healthy. Based on this information, the cloudiness data was given the lowest priority among the weather data used in the hybrid model to predict traffic speed.

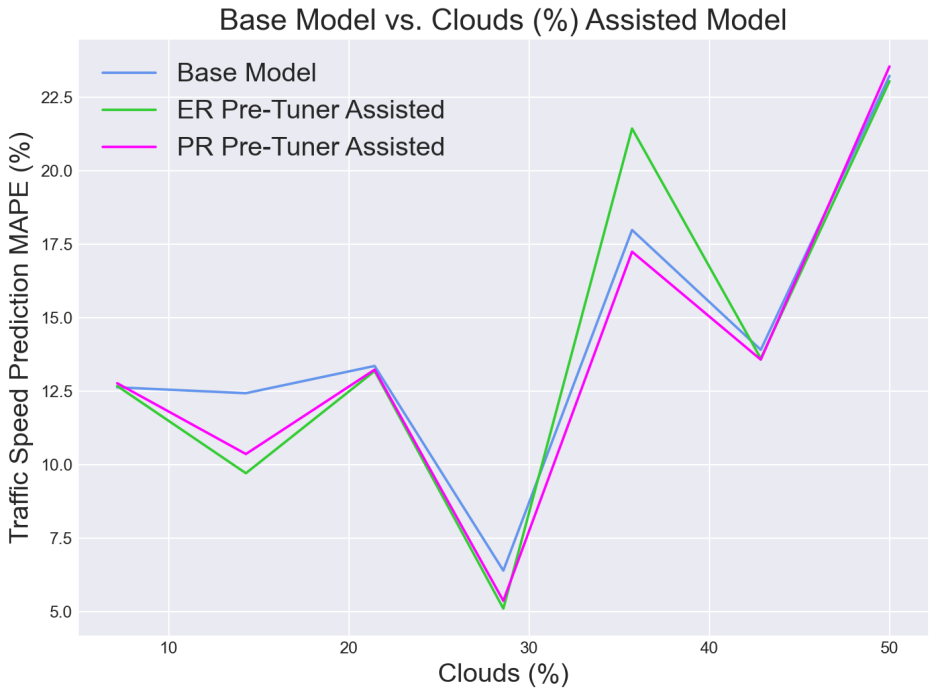


Figure 7.18 Clouds-Assisted Model Performance

7.3.1.5 Visibility-Assisted Model

Figure 7.19 shows estimation improvements using visibility distance data. Based on the graph, it appears that the ER tuner has an effect rate value of 3%, while the PR tuner has a polynomial regression degree of 9.

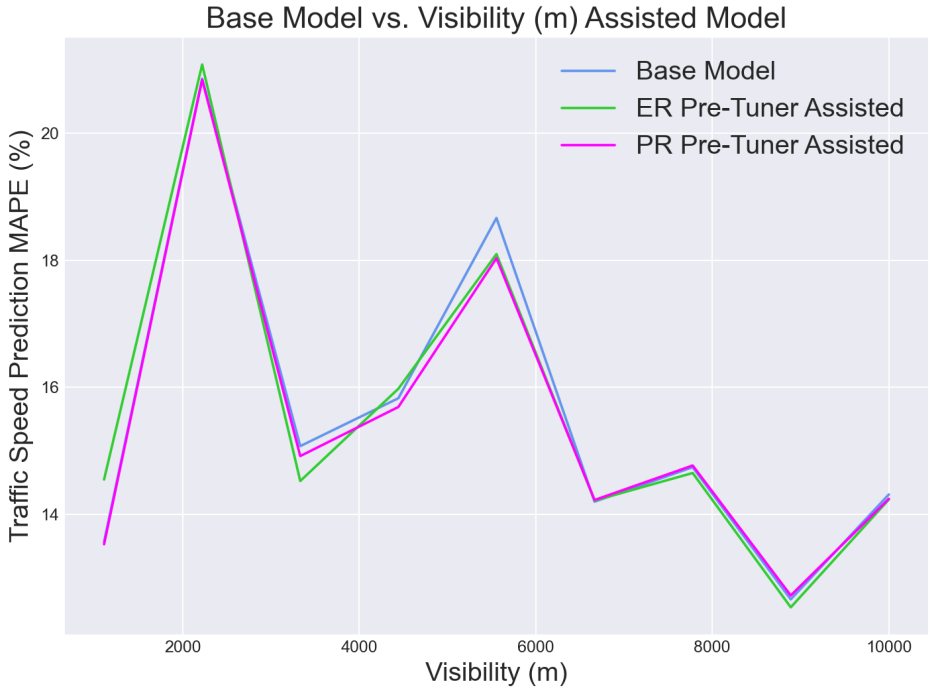


Figure 7.19 Visibility-Assisted Model Performance

Occasionally, fog is visible in the mornings in Istanbul. However, the frequency of this happening is shallow. When the figure is examined, it is seen that there is an improvement in the scenario where the visibility is above 5000 meters and below 6000 meters. The increase in overall performance was limited because the number of cases where the view distance was reduced enough to affect traffic was insufficient. Considering this infrequent use case, it was decided not to use the sight range (visibility) in the hybrid prediction model.

7.3.1.6 Hybrid Model

The hybrid weather model uses the most valuable information from 3 different weather conditions to increase prediction performance on a broader scope. In this section, we present the model's performance results, which demonstrate an improvement in forecast success through the use of the most valuable weather data.

Improvement was made at all times of the year using a hybrid weather-assisted traffic speed forecasting model. Significant improvements were made, especially in weather conditions where performance dropped dramatically, and more successful results were achieved in all months of the year using temperature data. Throughout the year, there was an improvement in the MAPE value from the base model's average of 13.23% to 13.06%.

Figure 7.20 and Figure 7.21 show the performance increases according to the hours of the day. When we examine the performances during the hours of the day, it is shown that the performance increase is higher, especially at night hours and after work hours.

Model performances are given in Figure 7.22 and Figure 7.23 according to the days of the week. Studies have demonstrated that there is a consistent rise in performance every day of the week as weather conditions impact traffic regardless of the time.

Model performances are given in Figure 7.24 and Figure 7.25 according to the months of the year. This approach not only provides better estimation but also largely equalizes the conditional differences among the training data (months). The best example of this is seen in the monthly performance chart. As the forecast improved in the spring, it also increased performance in September and December, which could not be predicted well due to the seasonal transition.

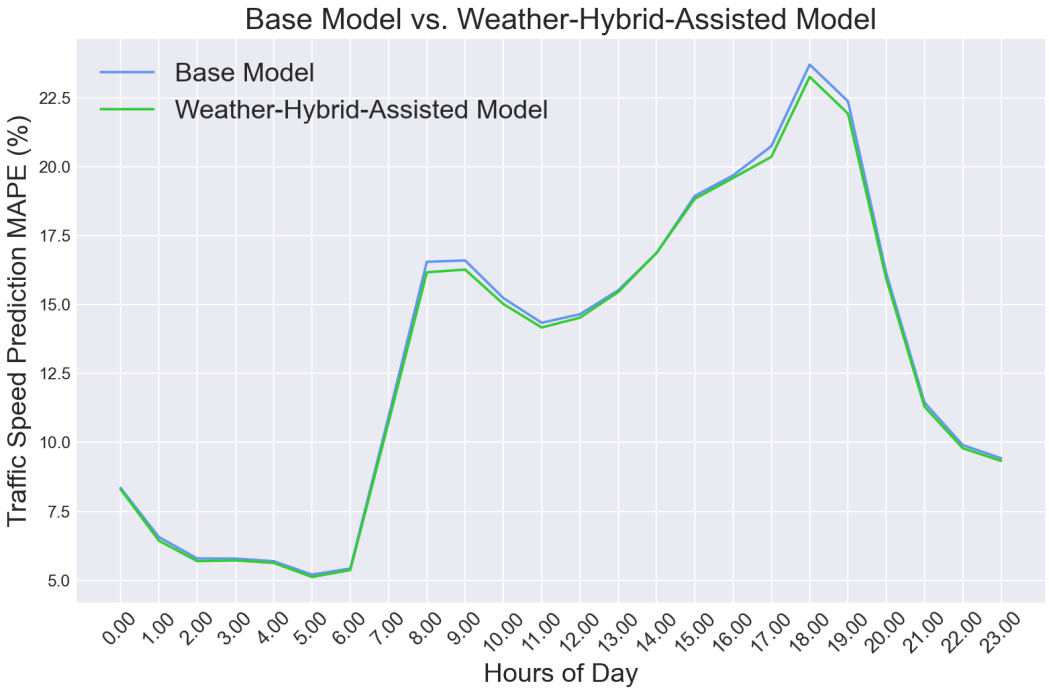


Figure 7.20 Weather-Hybrid-Assisted Models HOD Improvement (MAPE)

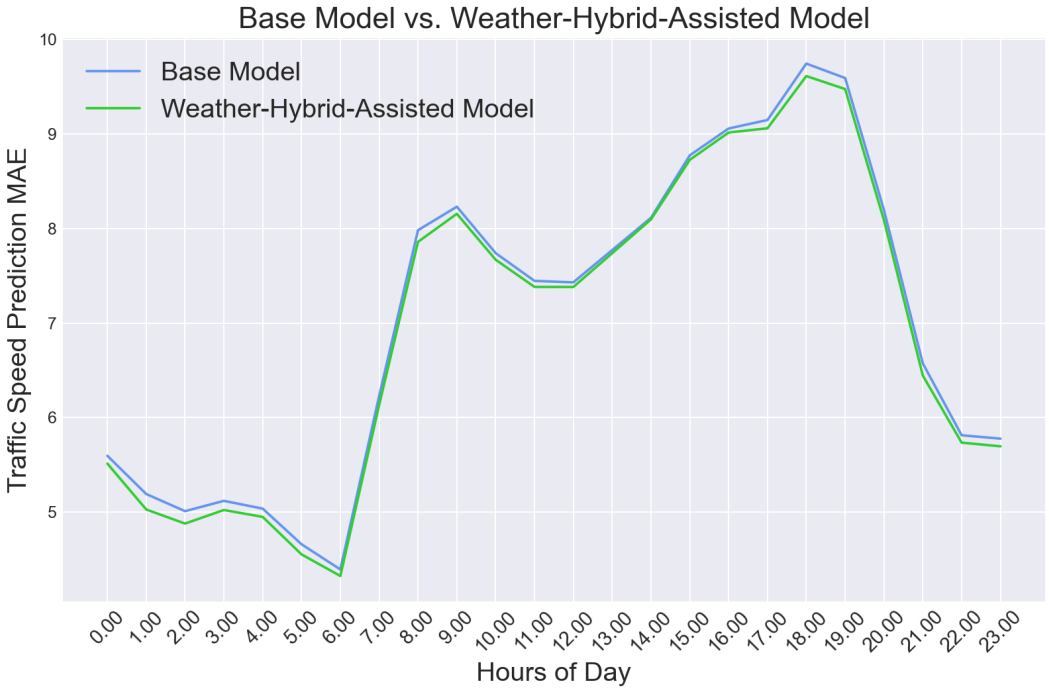


Figure 7.21 Weather-Hybrid-Assisted Models HOD Improvement (MAE)

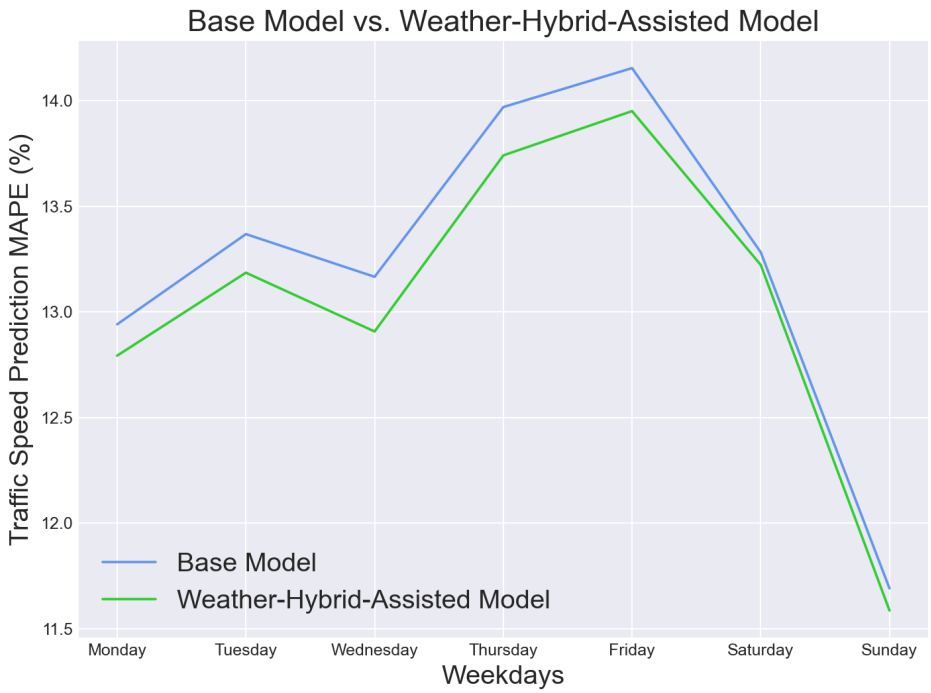


Figure 7.22 Weather-Hybrid-Assisted Models DOW Improvement (MAPE)

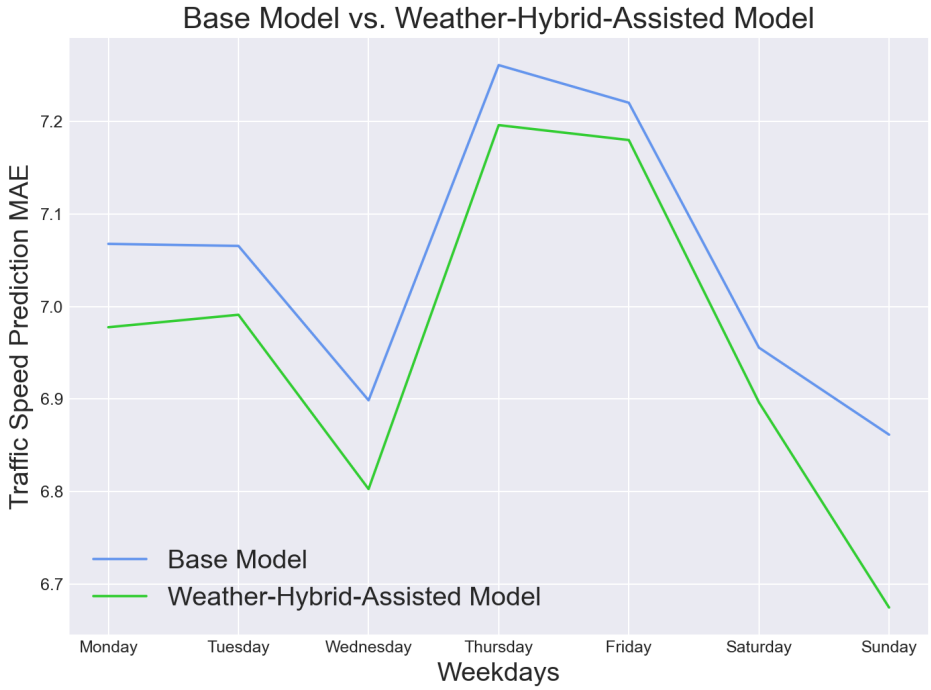


Figure 7.23 Weather-Hybrid-Assisted Models DOW Improvement (MAE)

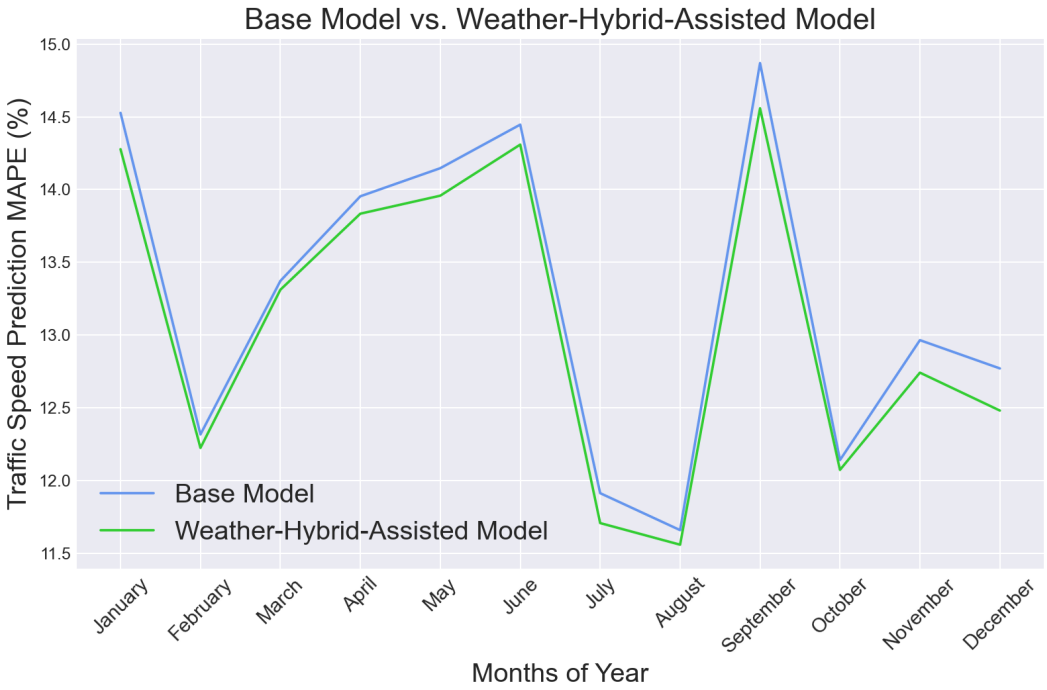


Figure 7.24 Weather-Hybrid-Assisted Models MOY Improvement (MAPE)

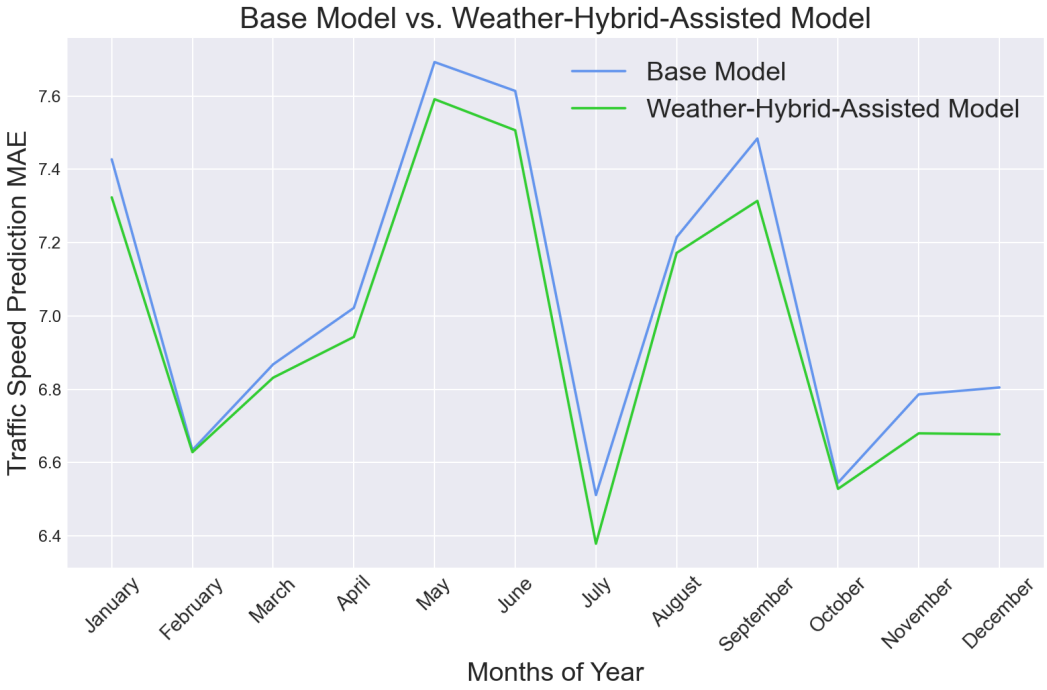


Figure 7.25 Weather-Hybrid-Assisted Models MOY Improvement (MAE)

7.3.2 Spatial-Assisted Models

This section gives the results of the estimations made using spatial data. Figure 7.26 and Figure 7.27 show the performance increases according to the hours of the day. One of the most basic scenarios we have seen during traffic speed estimation studies is that a normally flowing road is exposed to traffic jams in the morning and evening hours. When the figure is examined, it is seen that the forecasts made between 16.00 and 19:00 can be improved by up to 1.3%. This situation shows that the amount of exposure to the environment of the road areas is higher, especially during commuting hours. Traffic that flows at night and during the middle of the day is less affected by the previous and next road segments. Considering that a normally flowing road starts to condense from the junction points towards evening hours, we can say that the road segments ahead gradually affect the traffic.

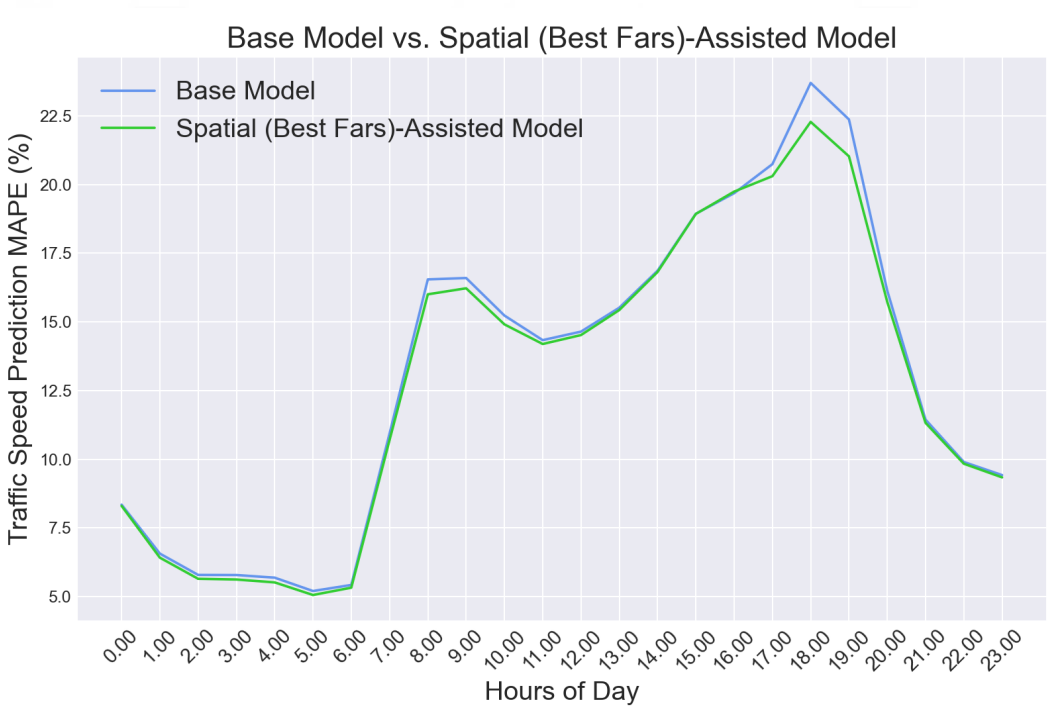


Figure 7.26 Spatial Models HOD Improvement (MAPE)

Model performances are given in Figure 7.28 and Figure 7.29 according to the days of the week. It is seen that there is an improvement in the MAPE value at the level of 0.5% every weekday. On weekends, gains are limited. These results show that using spatial data improves performance in heavy traffic situations.

Model performances are given in Figure 7.30 and Figure 7.31 according to the months of the year. The graph shows the forecast performances are higher, especially in the year's first half. Especially in May, the performance increase is remarkable.

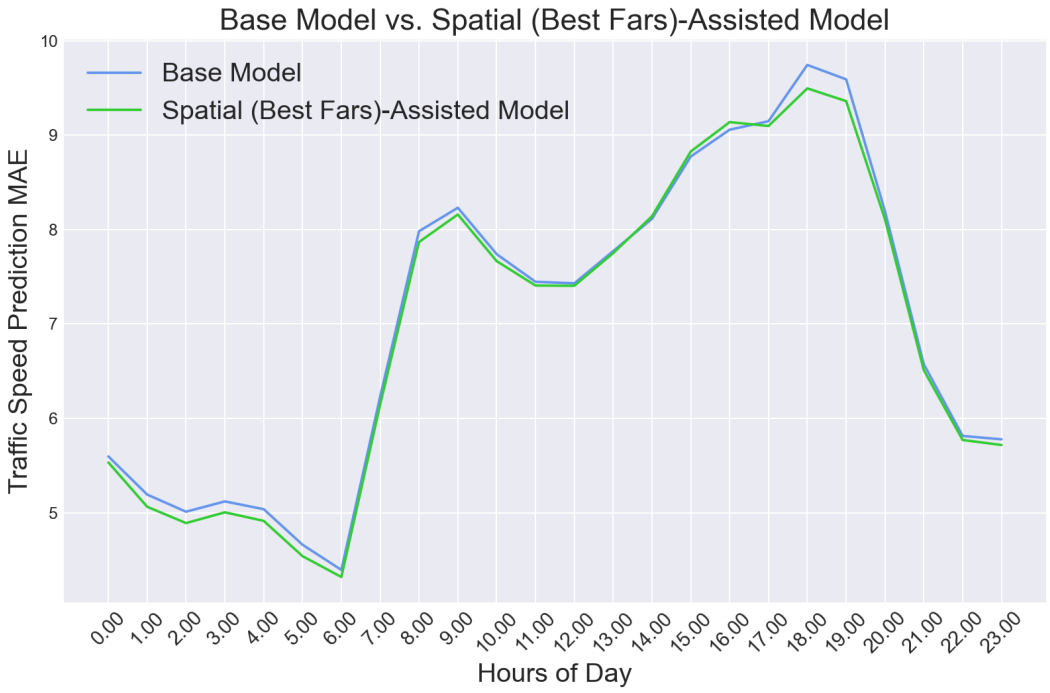


Figure 7.27 Spatial Models HOD Improvement (MAE)

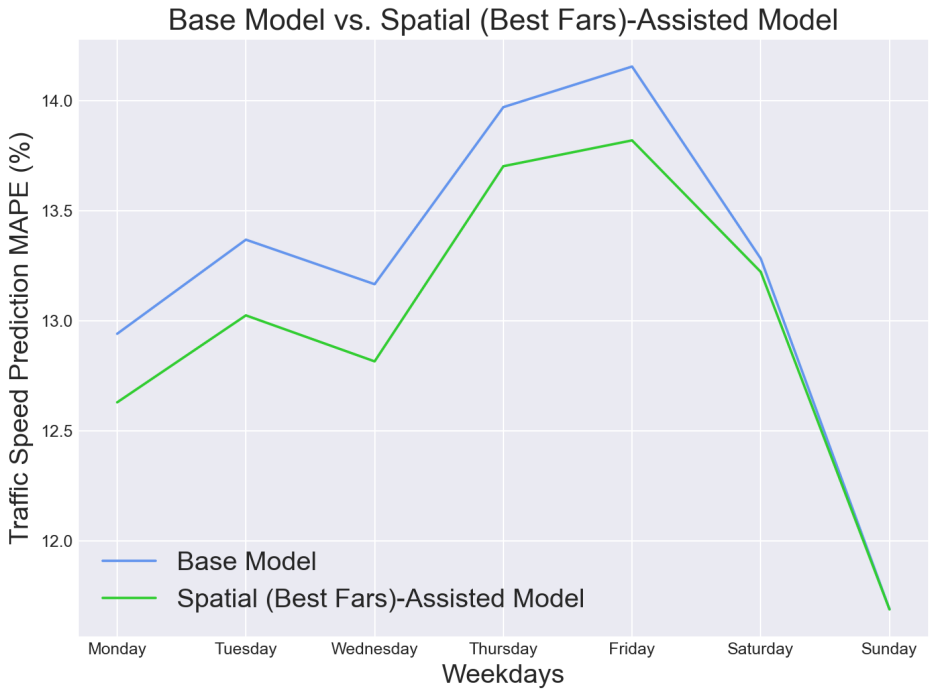


Figure 7.28 Spatial Models DOW Improvement (MAPE)

We ran experiments by feeding information up to 1, 2, and 3 neighboring distances since we think the influence coefficients from their surroundings are not the same for all road segments. From the results obtained, we developed a hybrid model using the neighbor distance where each segment succeeds. The map in Figure 7.32 shares which road segment area gives the most successful results at what

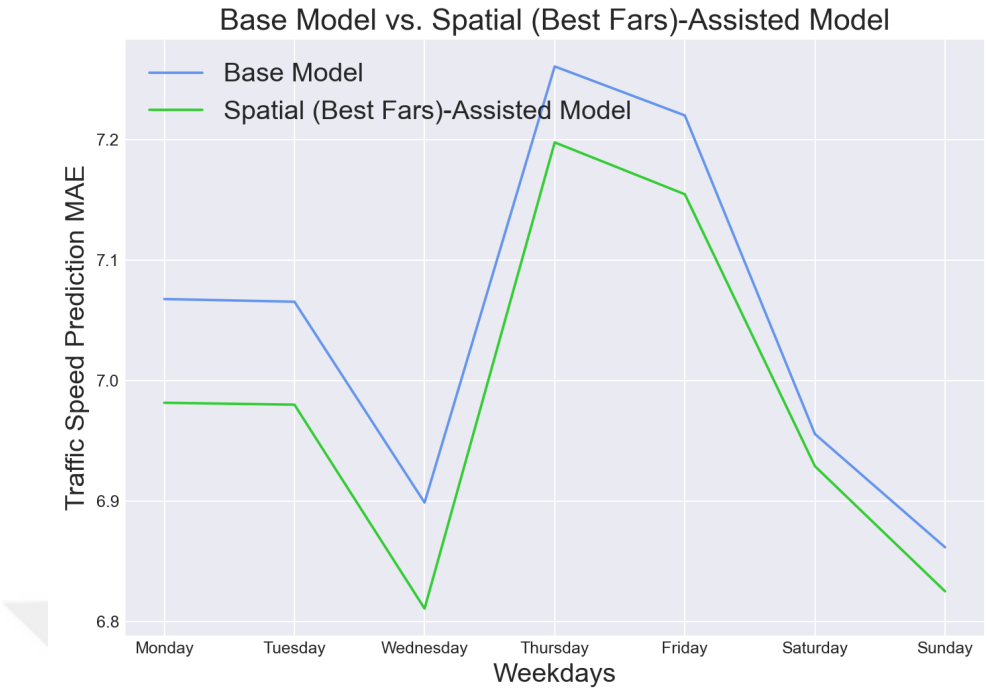


Figure 7.29 Spatial Models DOW Improvement (MAE)

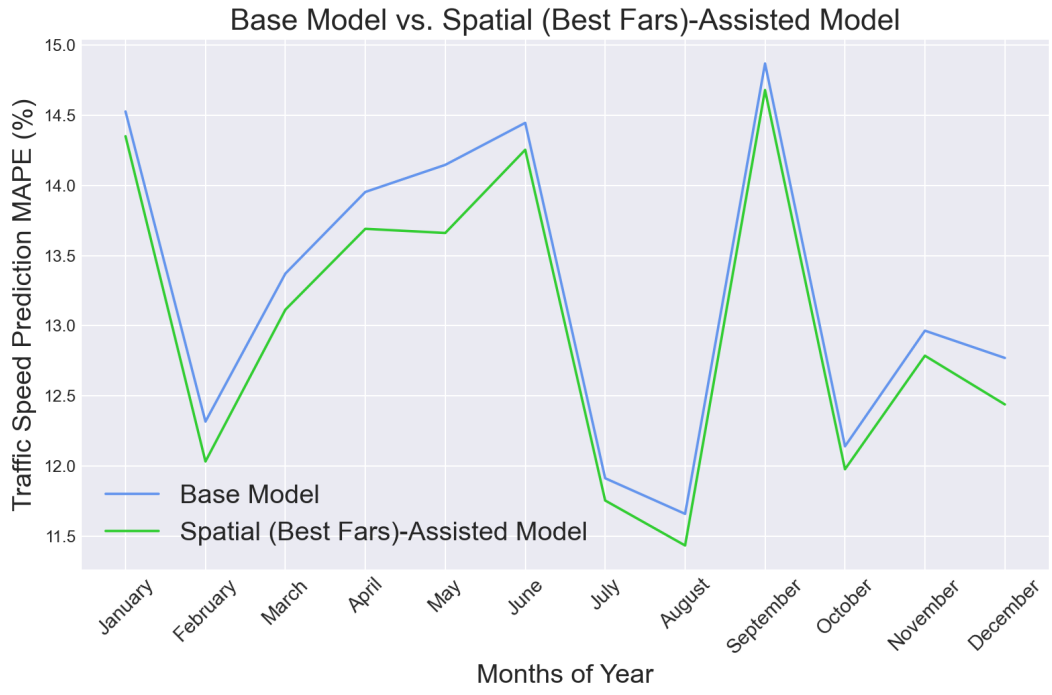


Figure 7.30 Spatial Models MOY Improvement (MAPE)

distance from the neighbor. When the map is examined, the rate of being affected by neighboring segments increases, especially in regions that host the junction and serve as a transfer road during rush hour. In areas far from road junctions, the rate of being affected by neighboring segments decreases noticeably. For example, in the areas before and after the Fatih Sultan Mehmet Bridge, the rate of being affected by

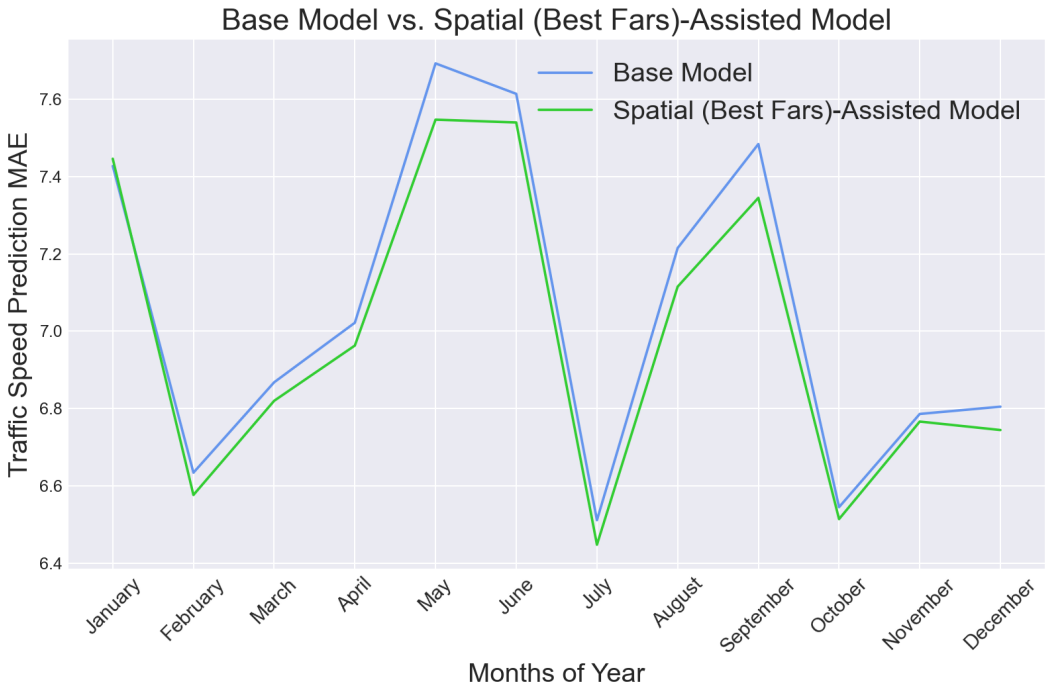


Figure 7.31 Spatial Models MOY Improvement (MAE)

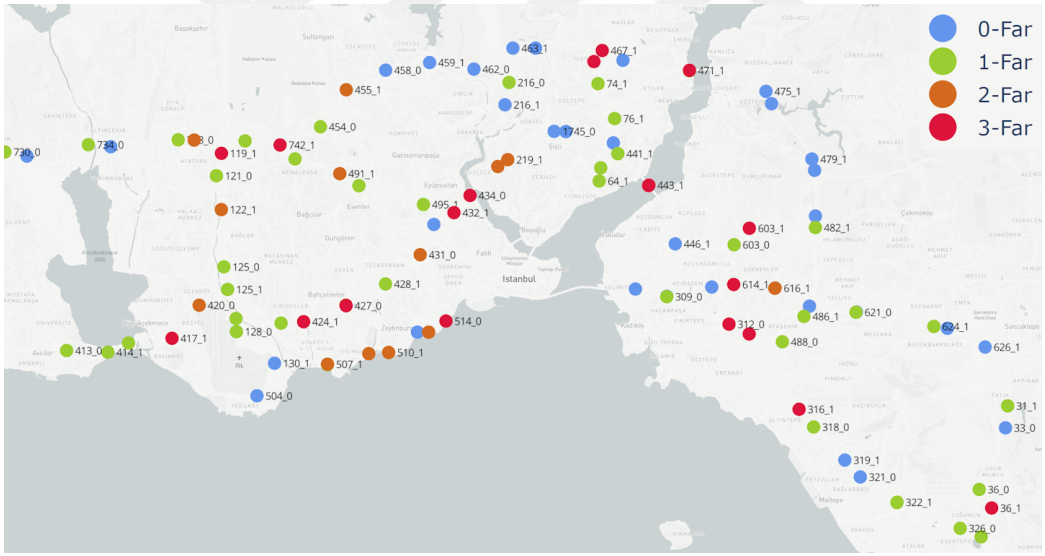


Figure 7.32 Spatial Best Fars Map

the neighboring segments is shallow. In the Levent region, it was observed that the predictions made using the information of the neighboring segments at a distance of 1 unit were more successful. This shows that feeding the regions within the traffic domain to the prediction model increases the overall performance.

7.3.3 Speed-Group-Assisted Models

Model performances are given in Figure 7.33 and Figure 7.34 according to the hours of the day. When the graph is examined, it is seen that the speed grouping technique increases performance during rush hour and daytime hours. This result shows that training and estimating each speed category improves the model prediction quality.

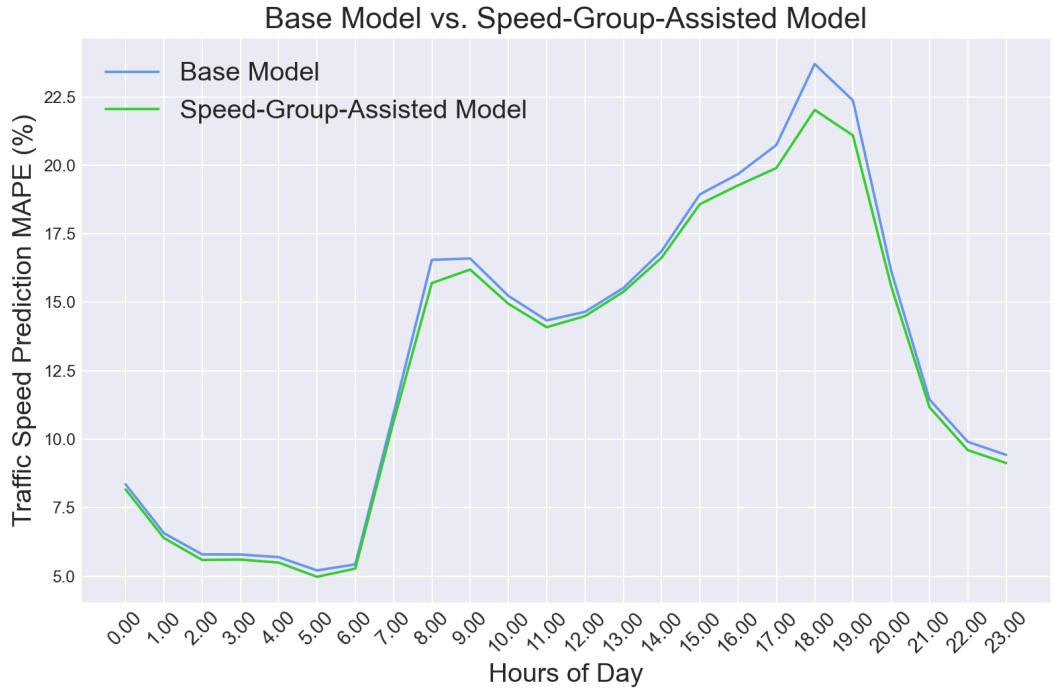


Figure 7.33 Speed-Groups-Assisted Model HOD Improvement (MAPE)

Model performances are given in Figure 7.35 and Figure 7.36 according to the days of the week. When the graph is examined, it is shown that this method achieves forecast improvements of up to 1% every day of the week. When the chart is scrutinized, it is shown that the increase in performance on Thursday is higher than Friday in terms of MAE. Fridays of the week host multiple speed categories that change rapidly in themselves. This may reduce the performance when a traffic speed value outside the traffic speed category decided according to the past traffic speed information is realized. Also, Fridays are the slowest day of the week. This situation affects Friday especially. Generally speaking, it is evident that categorizing traffic speeds leads to more efficient results during weekdays.

Model performances are given in Figure 7.37 and Figure 7.38 according to the months of the year. September is the worst predicted month of the year, with the base model. The reason for inaccurate traffic predictions is that the historical data is from the summer months, and traffic patterns change quickly. In the estimations made using the traffic speed category technique, the average estimation error rate in September has been improved from 16.4% to 14.3%. This is a huge step up for the

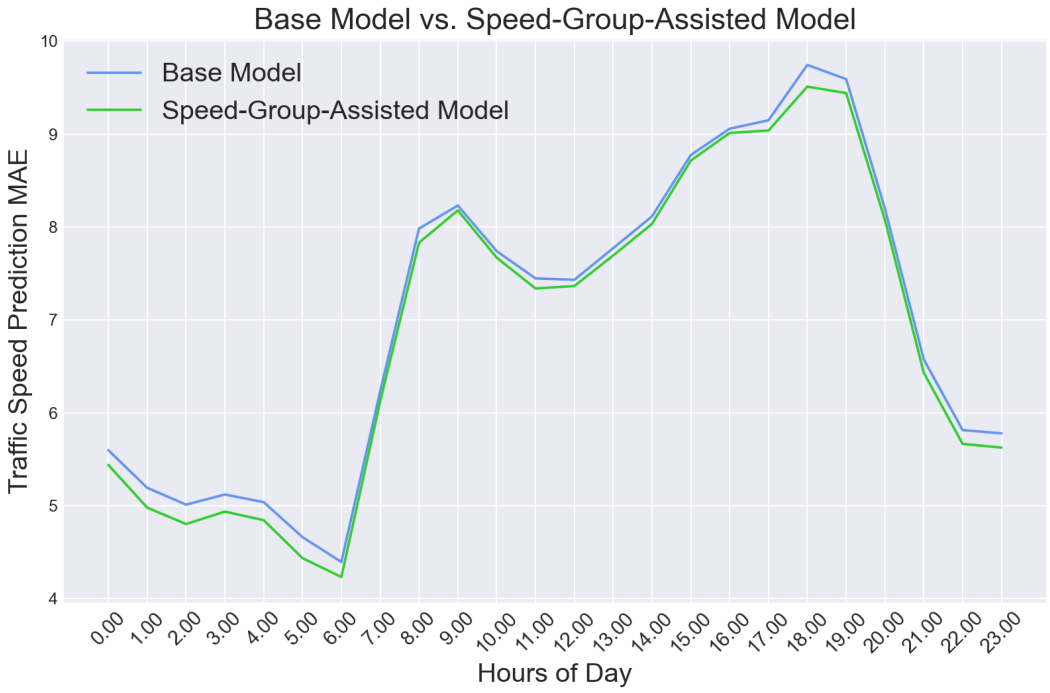


Figure 7.34 Speed-Groups-Assisted Model HOD Improvement (MAE)

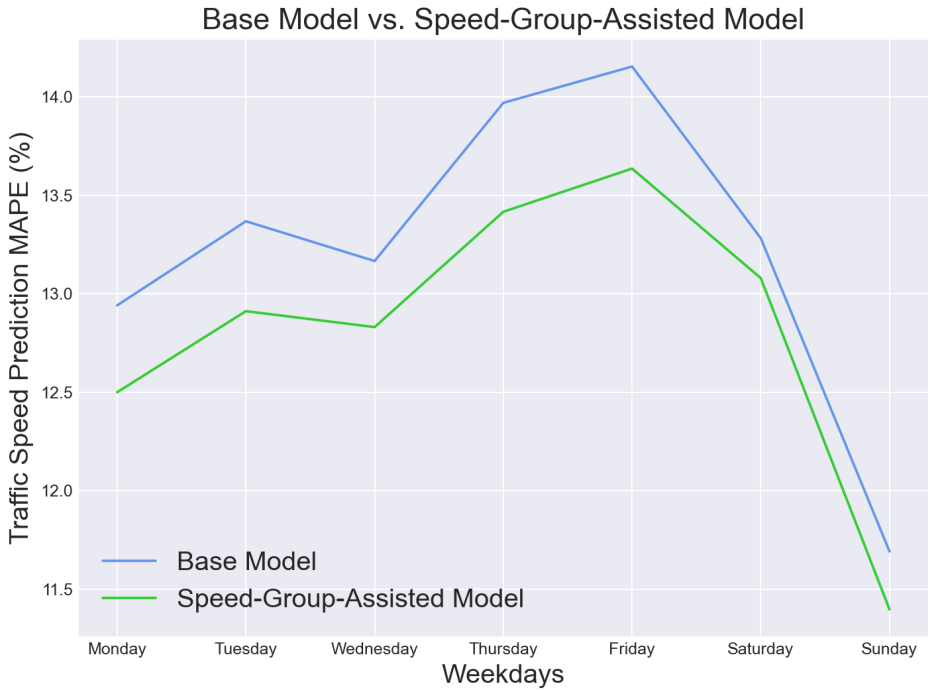


Figure 7.35 Speed-Groups-Assisted Model DOW Improvement (MAPE)

most challenging month.

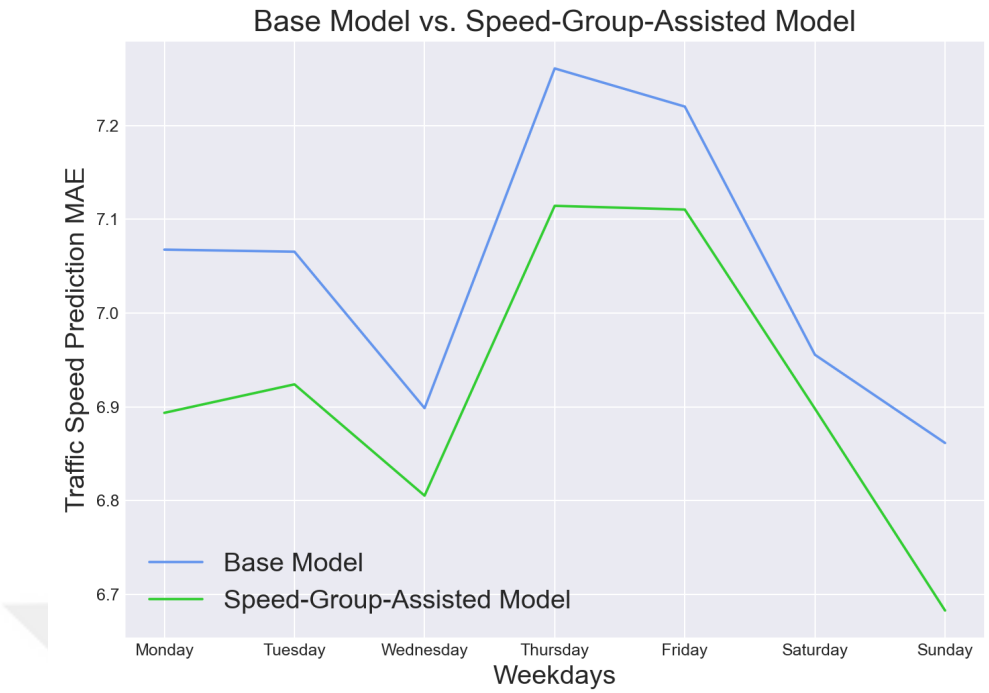


Figure 7.36 Speed-Groups-Assisted Model DOW Improvement (MAE)

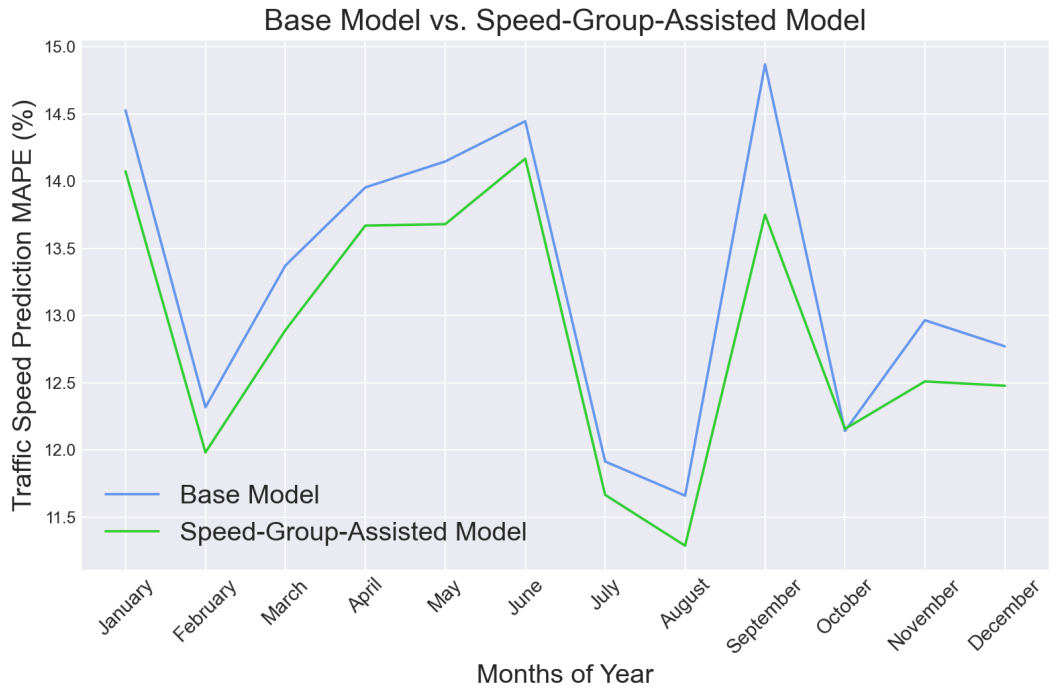


Figure 7.37 Speed-Groups-Assisted Model MOY Improvement (MAPE)

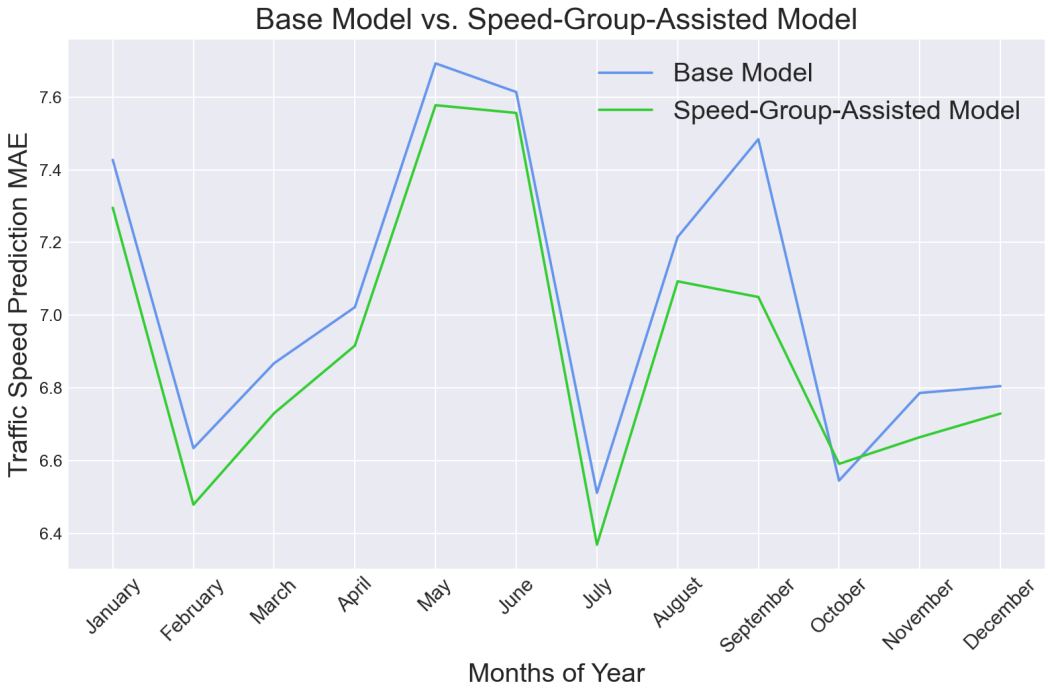


Figure 7.38 Speed-Groups-Assisted Model MOY Improvement (MAE)

7.4 Powerful (All-in-one) Model Experiments

Based on the experiments, the approach for enhancing the accuracy of segment prediction varies. One approach may be effective for a particular segment, while another may work best for a different segment. In order to improve performance for whole situations, a powerful model was created. This section will analyze the performance improvements achieved using the powerful model compared to the base model.

7.4.1 Model Selection

Figure 7.39 shows the scatter plot of the improvement models contributing to the Powerful model. The models that gave the most successful results on the segments were determined by using the MAPE values of the experiments. In this selection, the characteristics of the road zone up to the moment of estimation were taken into account. Table 7.2 shows the average MAPE and MAE values obtained from the experiments, along with the additive ratios of the powerful model. When the results were examined, it was shown that the improvement models with the triple combination had a total contribution rate of up to 38%. Remarkably, the contribution rate of the base model is just one of the results. This result shows that performance increase is provided for almost each road segment. It is shown from the results that the parameter set of data feeding to the model is as effective

as the estimation model. It is shown that the road regions that benefit more from the speed grouping technique get more successful results with parameter set B. Similarly, parameter set A shines in road areas more affected by weather conditions. This reflects the importance of window size in capturing detail, which is the main difference in parameter sets.

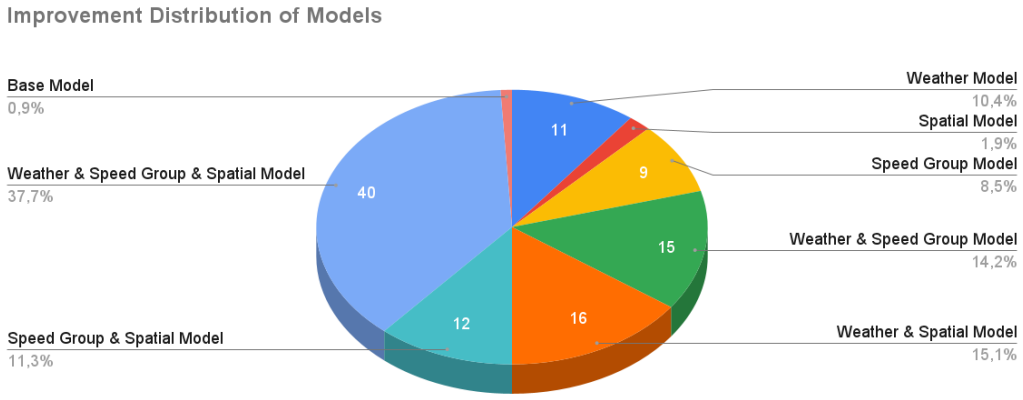


Figure 7.39 Improvement Distribution of Models over 106 Segments

The MAPE value, originally 13.23% in the base model, was increased to 12.64% MAPE after the improvements were implemented. When the performance increase is analyzed with MAE, progress has been made from 7.04 km/h to 6.89 km/h. In the next section, the performance of the Powerful model is examined in detail.

Table 7.2 Prediction Models Contribution

Model Definition	**Paramset**	**MAPE**	**MAE**	**Imp.Seg.Rate ∇**
Wtr. & Sp. Group & Spatial	B	12.84%	6.99	24.5%
Wtr. & Sp. Group	B	12.89%	6.93	14.2%
Wtr. & Sp. Group & Spatial	A	12.87%	6.99	13.2%
Weather	A	13.06%	6.96	10.4%
Speed Group & Spatial	B	12.86%	7.00	11.3%
Speed Group	B	12.91%	6.94	8.5%
Weather & Spatial	B	13.03%	6.98	8.5%
Weather & Spatial	A	12.97%	6.96	6.6%
Spatial	A	12.99%	6.98	1.9%
Base	A	13.23%	7.04	0.9%

7.4.2 Time Slot Performances

In Figure 7.40 and Figure 7.41, the change graphs of the performance gains of the Powerful model compared to some models are given according to the hours of the day. When the charts are examined, it is shown that more successful forecasts

are produced at the rate of at least 0.2% at every hour of the day. It has been observed that the prediction performance improvements peak, especially during rush hour, when the prediction performance drops too much. Over 3% improvement in prediction performances was seen at 6 PM. Thanks to this method, the problem of traffic forecasting, which is particularly challenging during rush hours, has been addressed and improved.

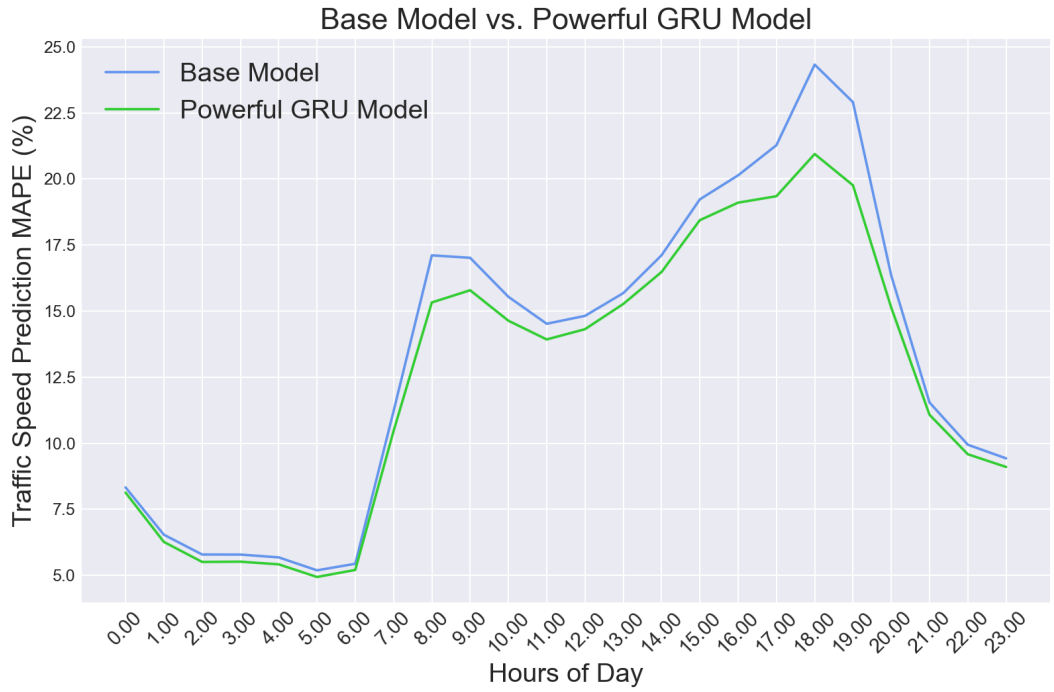


Figure 7.40 Powerful Model HOD Improvements (MAPE)

Figure 7.42 and Figure 7.43 show the predicted performance gains of the Powerful model by days of the week. The actual performance gains occur on weekdays when traffic is even harder to predict.

Figure 7.44 and Figure 7.45 shows the prediction performance gains of the Powerful model by months of the year. As a result, the increase in forecast performance on a percentage basis in September was at the maximum level. When the results given in MAE are examined, it is seen that September can be predicted with even fewer errors than August. The fact that the traffic speed is lower in September significantly affects the formation of this situation. As a result, the amount of errors made seems much higher when evaluated in terms of MAPE.

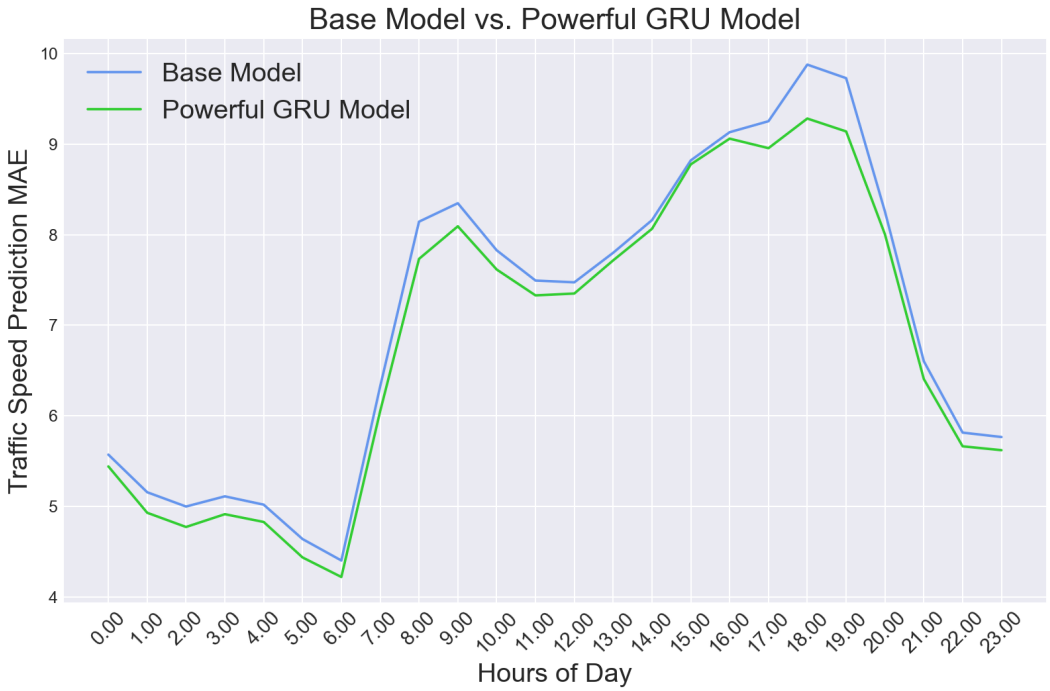


Figure 7.41 Powerful Model HOD Improvements (MAE)

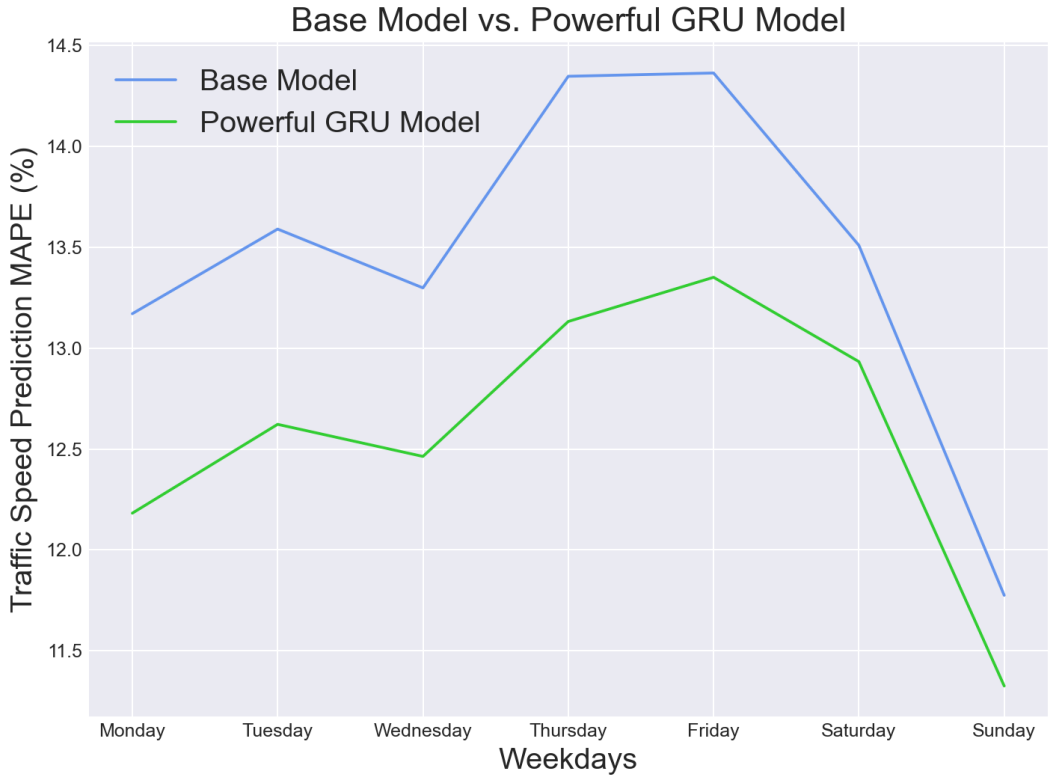


Figure 7.42 Powerful Model DOW Improvements (MAPE)

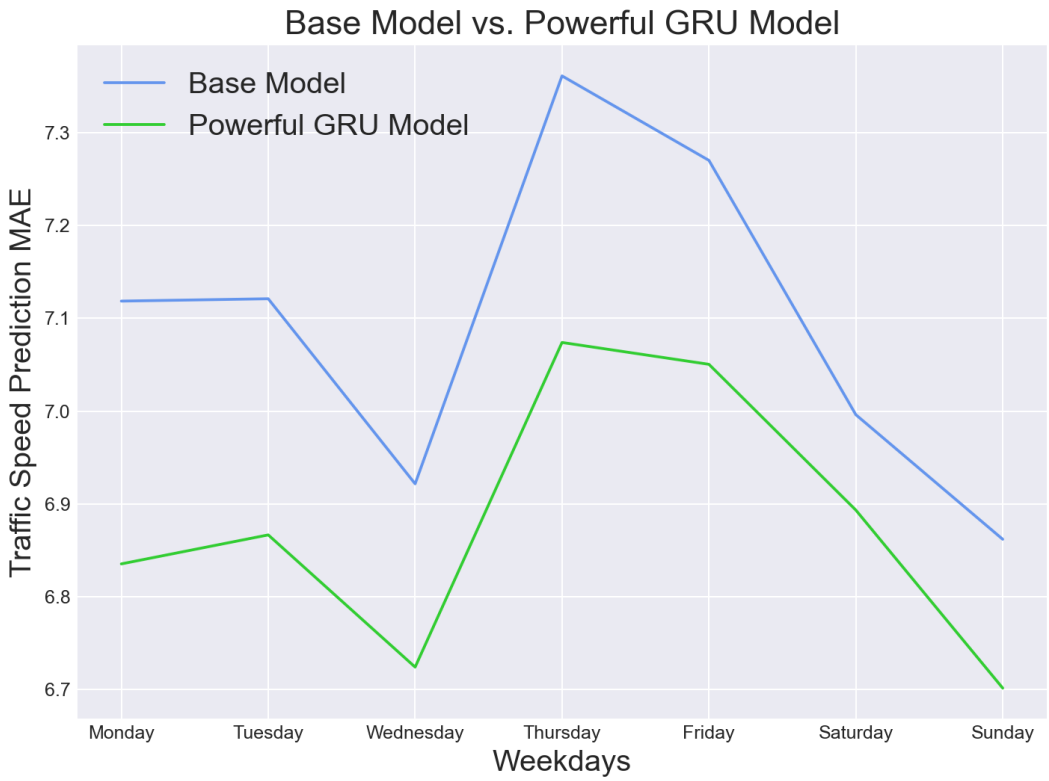


Figure 7.43 Powerful Model DOW Improvements (MAE)

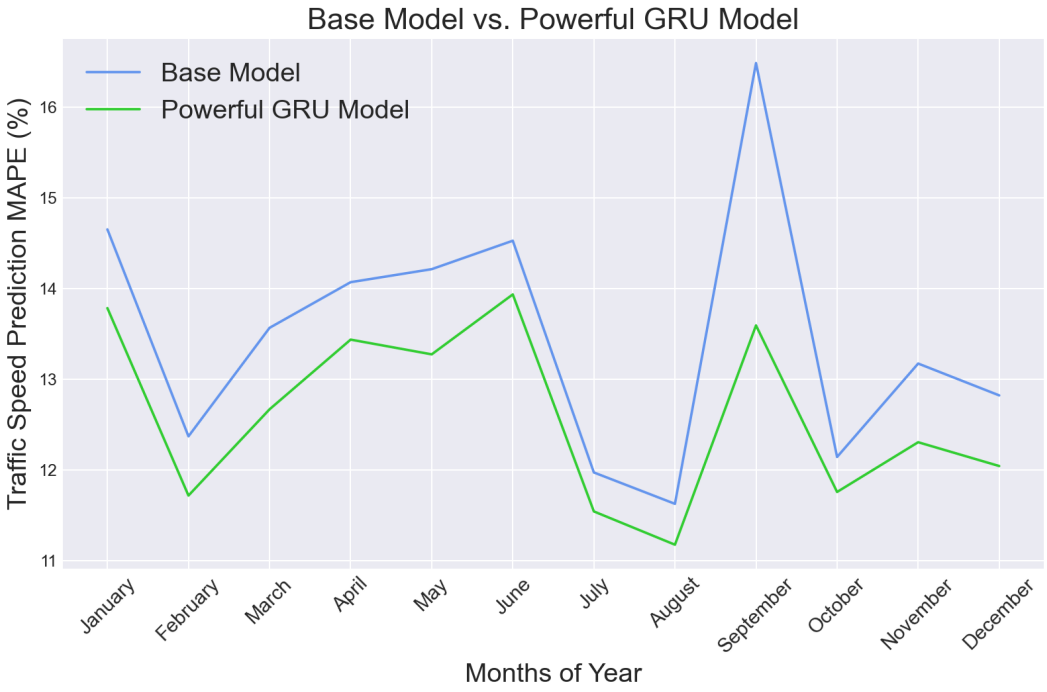


Figure 7.44 Powerful Model MOY Improvements (MAPE)

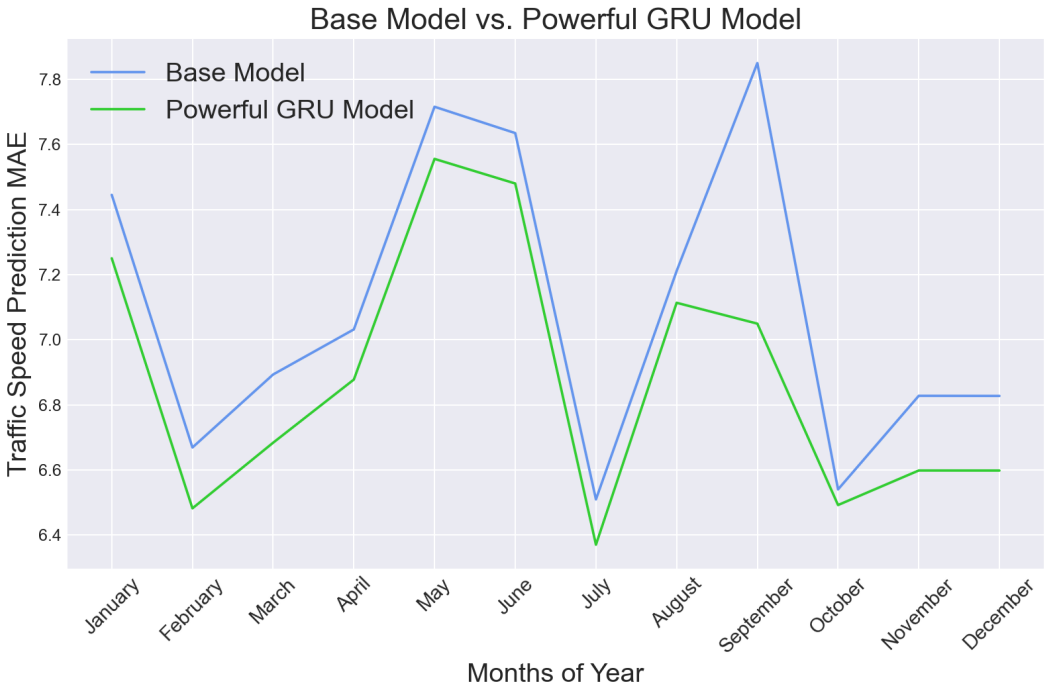


Figure 7.45 Powerful Model MOY Improvements (MAE)

7.4.3 Road Segment Performances

This section analyzes the gains achieved based on road segments with the Powerful model. In Table 7.3, MAE and MAPE values obtained from the base model, MAE and MAPE values obtained from the Powerful model, and the accuracy improvements between the two models are given for each segment. The values in the table are ordered according to the highest MAPE gain. This table is important because it shows the performance achieved by the multi-feature space in the format of the road regions.

Table 7.3 Final Accuracy Improvements of Segments

Segment	Base MAE	Base MAPE	Powerful MAE	Powerful MAPE	Improved MAE	Improved MAPE ∇
614_1	6.60	11.20%	5.02	6.39%	1.58	4.81
122_1	8.14	15.49%	7.10	12.25%	1.04	3.24
64_0	6.82	17.90%	6.06	15.11%	0.76	2.79
446_1	8.48	21.77%	8.35	19.60%	0.13	2.17
119_1	6.42	12.94%	5.95	10.84%	0.47	2.10
443_1	7.74	12.85%	7.02	10.88%	0.72	1.97
491_1	5.78	8.54%	4.95	6.58%	0.83	1.96
316_1	10.61	25.17%	10.39	23.25%	0.22	1.91

Table 7.3 Final Accuracy Improvements of Segments (continued...)

Segment	Base MAE	Base MAPE	Powerful MAE	Powerful MAPE	Improved MAE	Improved MAPE ∇
424_1	7.77	17.38%	7.55	15.62%	0.21	1.76
74_0	7.62	21.15%	7.30	19.41%	0.32	1.74
510_1	7.16	16.40%	7.14	14.77%	0.02	1.64
421_1	7.10	16.86%	7.33	15.24%	-0.23	1.61
469_0	8.57	18.21%	8.16	16.64%	0.41	1.58
742_1	8.20	22.78%	7.97	21.27%	0.23	1.51
738_0	10.90	22.02%	10.64	20.57%	0.26	1.44
426_1	8.75	22.63%	8.53	21.27%	0.22	1.36
432_1	4.45	8.37%	4.10	7.04%	0.35	1.34
428_1	8.56	20.18%	8.59	18.84%	-0.03	1.34
312_1	7.88	17.00%	7.68	15.79%	0.20	1.21
309_0	5.91	10.52%	5.74	9.40%	0.18	1.12
319_1	9.53	19.65%	9.30	18.56%	0.23	1.09
482_1	8.12	15.42%	8.11	14.42%	0.01	1.00
514_0	4.03	6.69%	3.64	5.69%	0.40	1.00
463_1	8.97	17.58%	8.97	16.62%	0.00	0.96
64_1	4.04	9.12%	3.80	8.17%	0.24	0.95
444_0	7.47	12.75%	7.18	11.81%	0.30	0.94
125_1	9.72	19.17%	9.42	18.24%	0.30	0.94
312_0	6.66	13.08%	6.47	12.19%	0.19	0.89
441_1	7.46	20.18%	7.35	19.31%	0.11	0.86
434_0	8.98	17.66%	8.74	16.79%	0.24	0.86
467_1	5.30	7.21%	5.00	6.36%	0.30	0.85
462_0	12.59	24.27%	12.32	23.43%	0.27	0.84
734_0	10.34	21.30%	10.12	20.48%	0.22	0.81
624_1	9.19	15.26%	8.83	14.46%	0.36	0.80
36_1	5.60	7.33%	5.01	6.55%	0.59	0.79
626_1	9.69	16.74%	9.59	16.01%	0.10	0.73
420_0	7.49	18.26%	7.46	17.53%	0.04	0.73
625_0	6.59	11.46%	6.21	10.74%	0.38	0.73
441_0	8.73	21.42%	8.55	20.72%	0.18	0.70
603_1	5.33	8.28%	5.01	7.63%	0.32	0.65
482_0	6.15	9.89%	5.87	9.24%	0.27	0.65
74_1	4.51	9.51%	4.34	8.89%	0.16	0.61
454_0	8.99	14.89%	8.76	14.29%	0.23	0.59

Table 7.3 Final Accuracy Improvements of Segments (continued...)

Segment	Base MAE	Base MAPE	Powerful MAE	Powerful MAPE	Improved MAE	Improved MAPE ∇
486_1	8.35	15.88%	8.13	15.30%	0.22	0.59
427_0	8.55	18.04%	8.39	17.47%	0.16	0.57
480_0	6.26	8.89%	6.08	8.32%	0.18	0.57
414_1	10.31	23.74%	10.27	23.19%	0.05	0.55
431_0	8.06	18.14%	8.03	17.60%	0.03	0.54
417_1	8.41	18.91%	8.24	18.38%	0.17	0.53
486_0	9.71	16.50%	9.52	16.00%	0.19	0.50
125_0	6.23	11.44%	6.12	10.94%	0.12	0.49
322_1	7.82	13.86%	7.58	13.37%	0.25	0.49
318_0	8.13	15.78%	7.95	15.29%	0.18	0.49
436_0	5.27	8.61%	5.14	8.12%	0.13	0.48
31_1	6.19	7.10%	5.86	6.63%	0.34	0.48
472_0	8.19	11.55%	7.89	11.08%	0.30	0.47
76_1	5.43	15.29%	5.40	14.83%	0.04	0.46
475_1	8.69	14.83%	8.45	14.38%	0.24	0.44
323_0	4.67	6.19%	4.37	5.78%	0.30	0.41
479_1	8.89	15.66%	8.63	15.25%	0.25	0.41
121_0	7.55	13.31%	7.43	12.92%	0.11	0.39
621_0	6.59	10.21%	6.41	9.82%	0.17	0.39
424_0	6.95	12.90%	6.81	12.51%	0.14	0.38
1745_1	8.93	18.67%	8.90	18.29%	0.03	0.38
738_1	8.88	15.72%	8.68	15.36%	0.20	0.36
219_1	6.40	15.16%	6.36	14.80%	0.04	0.36
495_1	5.29	8.44%	5.07	8.08%	0.22	0.36
33_0	6.24	7.91%	5.98	7.56%	0.26	0.34
321_0	7.68	14.95%	7.53	14.61%	0.15	0.34
510_0	4.74	6.86%	4.50	6.52%	0.23	0.34
413_0	7.35	15.67%	7.20	15.33%	0.14	0.34
416_0	6.28	12.84%	6.17	12.50%	0.11	0.34
36_0	5.24	6.39%	4.94	6.06%	0.30	0.33
326_0	3.81	4.51%	3.54	4.20%	0.27	0.30
476_0	7.55	10.97%	7.39	10.68%	0.15	0.29
620_1	7.28	10.16%	7.09	9.88%	0.19	0.28
216_1	4.94	6.81%	4.77	6.53%	0.17	0.28
741_0	7.28	13.00%	7.14	12.72%	0.14	0.28

Table 7.3 Final Accuracy Improvements of Segments (continued...)

Segment	Base MAE	Base MAPE	Powerful MAE	Powerful MAPE	Improved MAE	Improved MAPE ∇
130_1	3.23	4.36%	3.02	4.09%	0.21	0.27
617_0	7.02	12.09%	6.99	11.83%	0.04	0.26
1751_1	3.92	6.11%	3.76	5.85%	0.16	0.26
128_0	5.58	10.41%	5.49	10.14%	0.08	0.26
1751_0	4.30	5.76%	4.11	5.52%	0.18	0.25
493_0	9.46	16.35%	9.35	16.11%	0.11	0.24
616_1	4.38	4.58%	4.20	4.34%	0.18	0.23
458_0	8.49	13.59%	8.30	13.37%	0.18	0.22
455_1	6.81	10.51%	6.63	10.29%	0.18	0.22
459_1	6.59	9.95%	6.43	9.74%	0.15	0.21
507_1	3.38	7.43%	3.30	7.23%	0.08	0.20
730_0	11.50	20.11%	11.50	19.91%	0.00	0.20
216_0	4.99	6.48%	4.86	6.29%	0.13	0.19
497_0	6.26	11.81%	6.20	11.62%	0.05	0.19
308_1	5.04	7.25%	4.94	7.06%	0.10	0.19
465_0	9.15	15.81%	9.05	15.63%	0.10	0.18
730_1	7.29	9.65%	7.17	9.48%	0.13	0.17
471_1	5.03	5.54%	4.90	5.39%	0.14	0.16
490_0	6.89	12.37%	6.80	12.22%	0.09	0.16
734_1	6.81	10.27%	6.67	10.13%	0.14	0.14
603_0	6.65	11.72%	6.51	11.60%	0.14	0.13
1779_0	3.34	6.24%	3.28	6.11%	0.07	0.12
504_0	4.58	10.15%	4.61	10.03%	-0.03	0.12
326_1	4.35	4.95%	4.26	4.84%	0.09	0.11
614_0	4.82	7.03%	4.77	6.92%	0.05	0.10
1745_0	8.31	22.84%	8.27	22.82%	0.04	0.03
488_0	10.50	20.09%	10.43	20.08%	0.07	0.00
2102_0	3.05	5.07%	3.05	5.07%	0.00	0.00

7.5 Post-Prediction Improvement Results

In this section, the performance increases obtained by using post-prediction improvements are examined. Since these modules can be included in the results obtained later, the performance increase provided by each module was examined

first; Then, cumulative performance increases are mentioned.

7.5.1 Accidents & Events Speed Adjuster

The process aims to reduce the failure of traffic speed estimation during a traffic event. 9869 accidents and traffic events, which were seen as significant in the dataset, were used. The experiments carried out using the dataset that included information on traffic accidents and road maintenance studies showed improvements in success rates. Before the module's operation, the MAPE value for accident predictions was 55.45%, but it was improved to 42.62%. Due to various external factors, predicting the impact of a traffic accident on traffic speed is quite challenging. However, the performance boost of more than 12% MAPE is a significant improvement that can enhance traffic forecasting in the case of an accident. Also, in traffic speed estimations made throughout the year, it was observed that the overall estimation performance was improved from 12.64% MAPE to 12.62% MAPE.

7.5.2 Legal Speed Limits Adjuster

The estimated value and actual speed limit in areas where traffic speed is forecasted have been recently updated to a maximum of 80 km/h as given in Figure 7.46. Thus, 12.64% MAPE was improved up to 9.45% just by applying the legal speed limit. When the exact change is analyzed in terms of MAE, it is seen that the error rate has been enhanced from 6.89 km/h to 4.05 km/h. Such a significant decrease in the error rate indicates that the prediction performance decreased due to exceeding the traffic speeds that should not be exceeded legally. With this module, the proportion of road zones that are predicted faster than they were, decreases from 48.32% to 25.85%. From an end-user perspective, estimating road zones as slower than the actual value is ideal, as the forecasts are expected to give the maximum times that will prevail in the worst-case scenario. When evaluated from this point of view, the reliability of the estimates is doubled.

7.5.3 Prediction Resolution Changer

We generated more than 11 million predictions with 5-minute resolutions for one year for 106 road zones. When we reduce it to 15 minutes resolution, instead of giving a prediction output with 5 minutes resolution until a week later, the number of times we estimate decreases to 3.2 million. With the resolution change, the error margin from 12.64% has been reduced to 11.28%. When the margin of error is analyzed as MAE, the margin of error, which is 6.89 km/h, has been reduced to



Figure 7.46 Legal Speed Limits in Istanbul

6.21 km/h. With this approach, the effect of instantaneous data noise on the margin of error has decreased, and the performance ratio has increased. In addition, the disk space allocated for traffic speed estimation has been reduced from 444 MB to 275 MB. Although disk space usage may seem insignificant in the short term, it has become possible to save 38% for the needs that will increase in the long term.

7.5.4 Post-Prediction Improvements Overall Contribution

The performance increase of the post-prediction improvements applied to the results produced by the traffic speed estimation model improved the MAPE level from 12.64% of the base Powerful model to 8.63%. When the error results are analyzed as MAE, the error rate of 6.89 km/h in the base model has been improved to 3.80 km/h. Traffic speed prediction models are as powerful as network-fed data. Real-world needs and factors change dynamically. Using the post-prediction improvement structure makes it possible to touch the predictions made by robust artificial neural networks in a customized way. Detailed temporal analyses are given in Figure 7.47, Figure 7.48, Figure 7.49, Figure 7.50, Figure 7.51, and Figure 7.52.

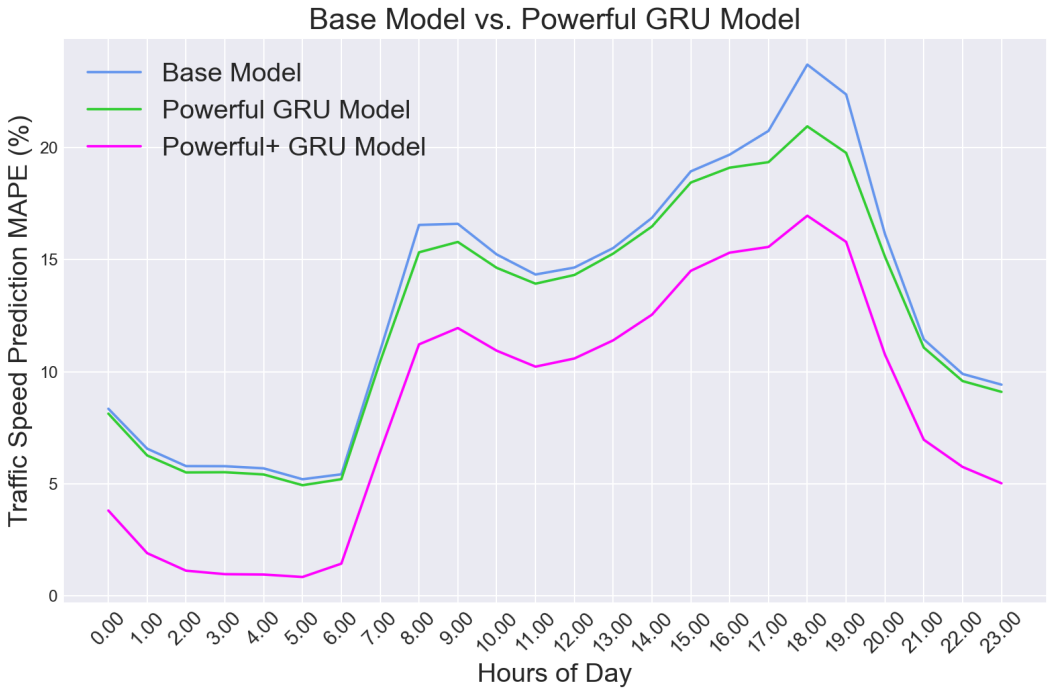


Figure 7.47 Powerful+ Model HOD Improvements (MAPE)

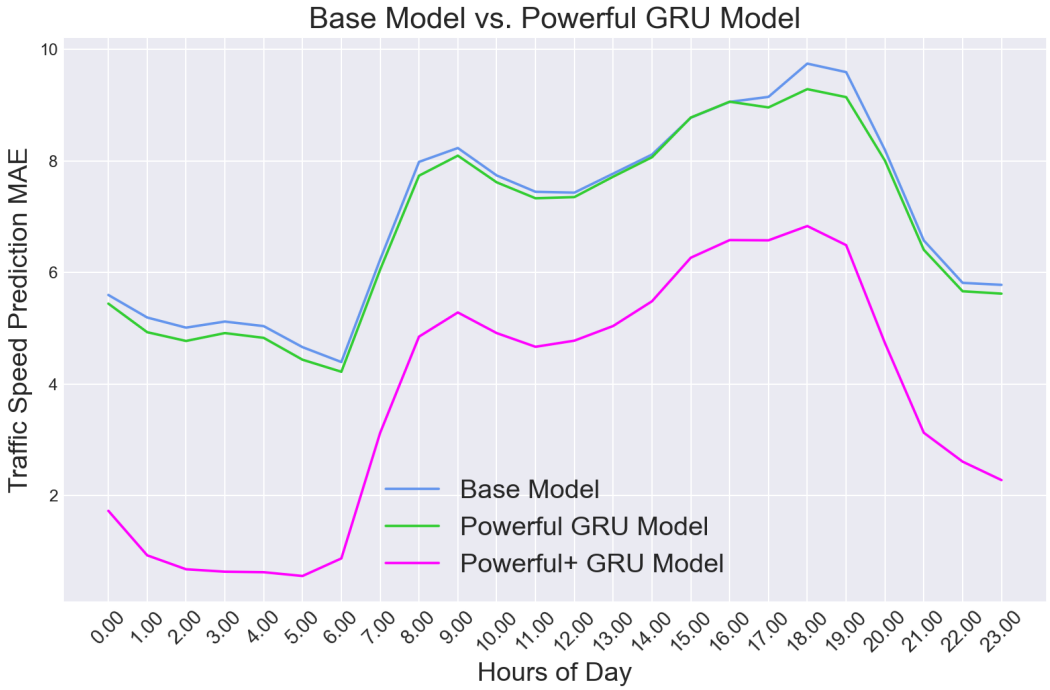


Figure 7.48 Powerful+ Model HOD Improvements (MAE)

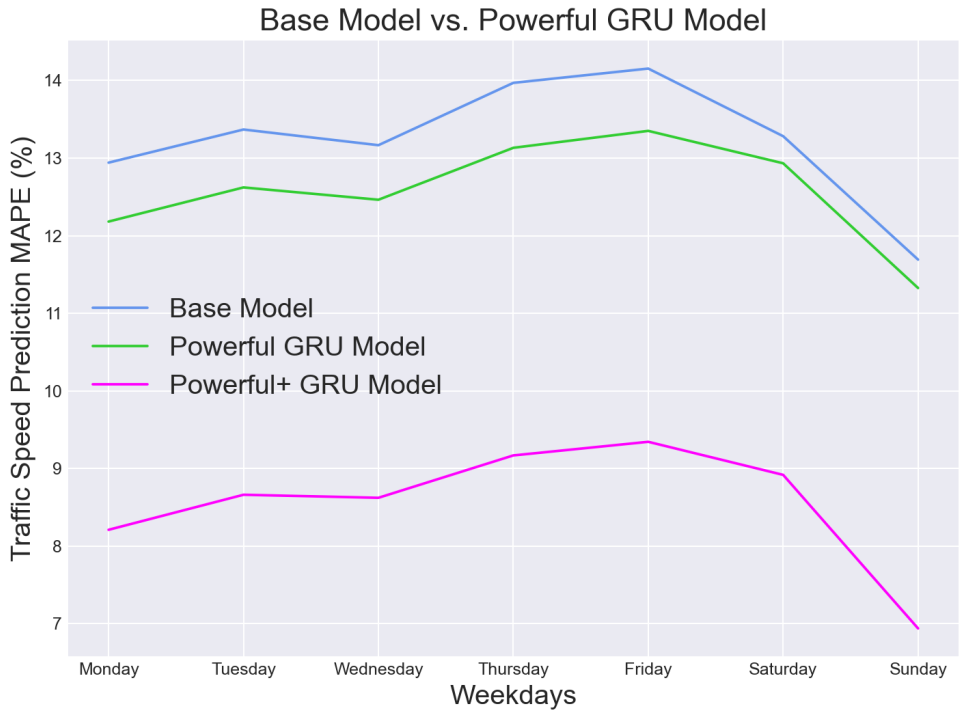


Figure 7.49 Powerful+ Model DOW Improvements (MAPE)

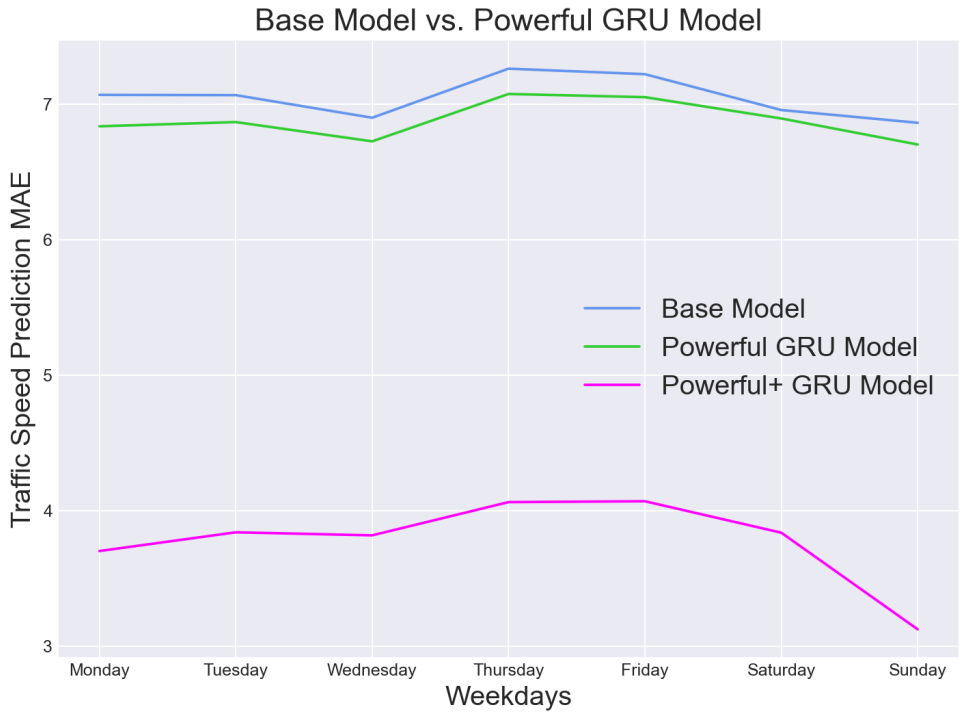


Figure 7.50 Powerful+ Model DOW Improvements (MAE)

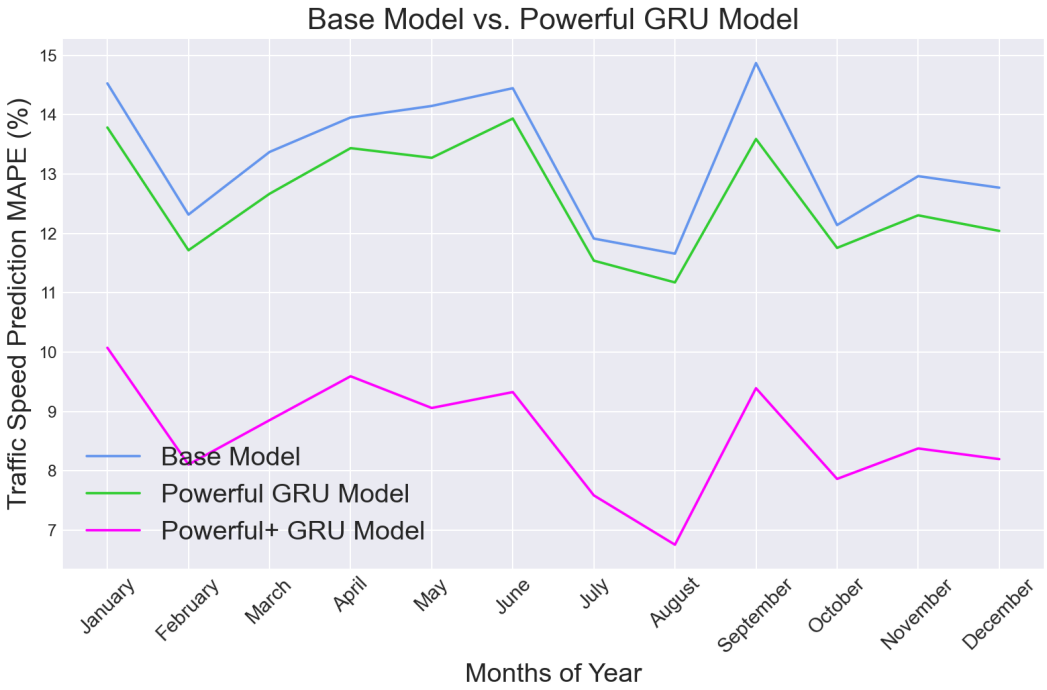


Figure 7.51 Powerful+ Model MOY Improvements (MAPE)

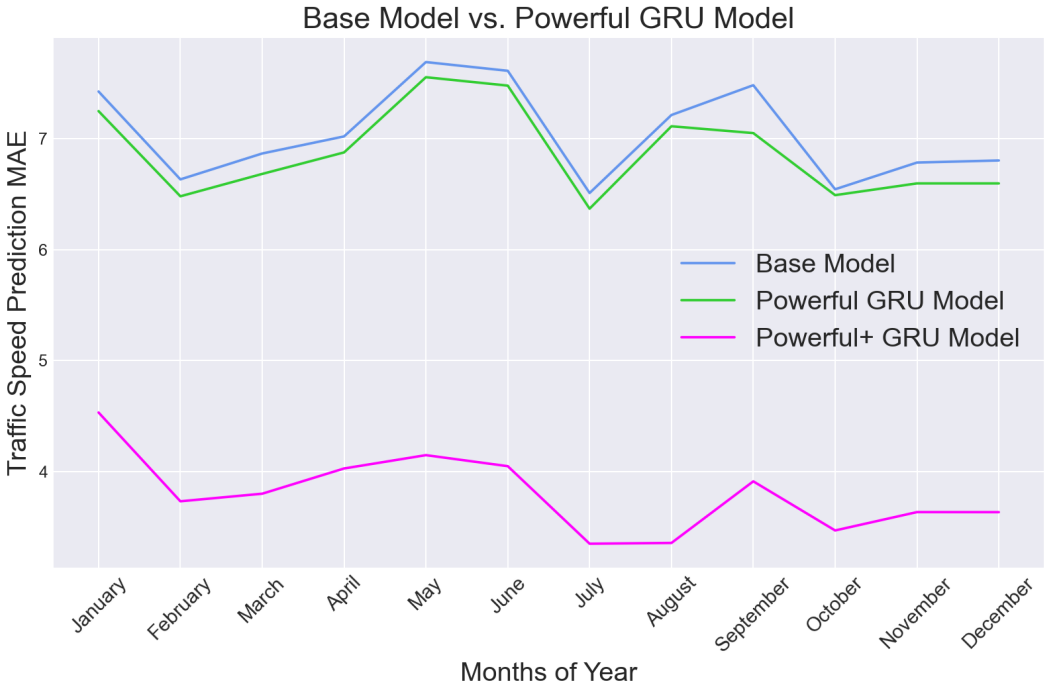


Figure 7.52 Powerful+ Model MOY Improvements (MAE)

7.6 Time Cost of Traffic Speed Estimation

In this section, the time performances of the proposed system architecture, time costs according to algorithms, and hardware acceleration details are given.

7.6.1 Time Cost of the Suggested System Architecture

Table 7.4 shows one segment's training/testing/storing time costs for a year according to the techniques used with GRU. The table also includes the average MAPE values obtained from the experiments. The values in the table show the results of the experiments running on the Nvidia RTX A5000 graphics card using Tensorflow and Keras. Experiments on an MSI laptop (with Tensorflow and Keras) run on average, 43.3% slower. Producing one prediction for a segment with the base GRU model was measured in approximately 8 ms. Since hundreds of forecasts can be taken simultaneously on a prediction model, this time is noticeably shorter in bulk predictions. When predictions are made in 5-minute periods in one month, approximately 8900 predictions are produced. Such an estimate is completed in about a second and a half. Based on this value, a 1-month forecast can be produced for 106 segments in approximately 150 seconds in a sequentially working estimator.

Table 7.4 Prediction Models Time Costs for 1 Road Segment's 12 Months with GRU

Model Definition	Paramset	Avg. MAPE	Consumed
Base Model	A	13.45%	29.09 mins
Hybrid Weather	A	13.29%	38.33 mins
Dynamic Distance Spatial	A	13.19%	43.91 mins
Speed Group	B	13.04%	37.58 mins
Weather & Speed Group	B	13.19%	39.35 mins
Weather & Spatial	A	13.16%	80.64 mins
Weather & Spatial	B	13.19%	55.83 mins
Speed Group & Spatial	B	12.99%	49.54 mins
Weather & Speed Group & Spatial	B	12.96%	62.03 mins
Weather & Speed Group & Spatial	A	12.99%	88.98 mins

In addition to these times, the data preparation process is also used for estimation. The data preparation process varies according to the needs of the model used. For example, one-month forecast data preparation for a segment for the base model takes about 2.6 seconds. This value can be up to 14 seconds when weather speed groups and spatial information are used. When the data preparation time is calculated by considering the distribution ratios of the data participating in the estimation in the powerful model used, it can be said that there is a preparation period of approximately 6 seconds for the one-month estimation of each road

segment. Figure 7.53 gives an example of the time cost for a road segment. If one-month forecasts are required for 106 segments in a sequential working CPU, a data preparation period of up to 10 minutes and 36 seconds is expected. In this scenario, it can be said that the estimation time, including the data preparation time, can be up to 13 minutes. In real-world usage scenarios, this time drops to under 10 seconds with the multi-threading method. Thus, as the most up-to-date weather data or new information about the future comes in, getting new traffic speed forecasts through forecast models is almost instantaneous.

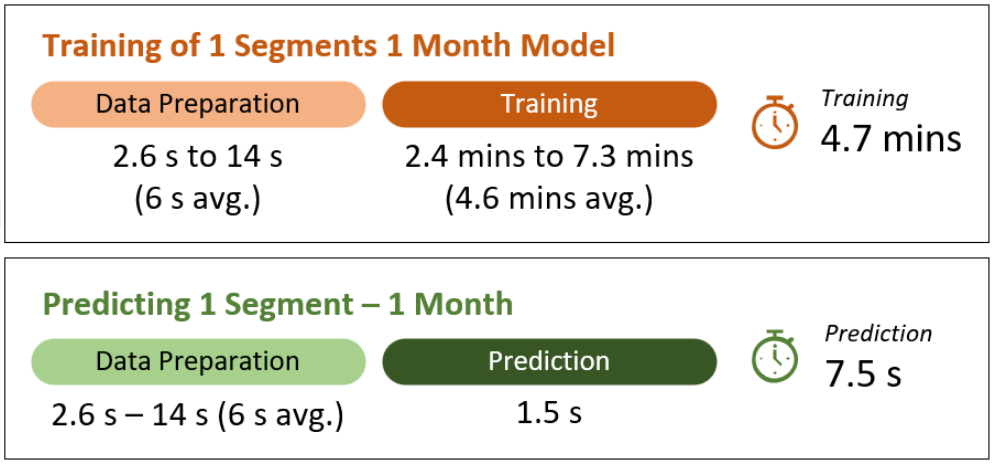


Figure 7.53 Time Cost of 1 Segment's 1 Month Model

7.6.2 Time Cost of Other Algorithms

Table 7.5 Time Costs of Base Prediction Models Time Costs for 1 Road Segment

Model Definition	Avg. MAPE	Consumed
Polynomial Regression	16.51%	1.89 mins
CNN Model	13.81%	25.51 mins
LSTM Model	13.45%	29.34 mins
GRU Model	13.23%	29.09 mins

The results of different algorithms according to their resource consumption are given in Table 7.5. The Polynomial Regression method, which can produce results quickly, is remarkable. With this simple method, which has a deficient speed of adaptation to changes, it is possible to get results fast. The prediction performance produced according to the promised time is also good. The running times of the GRU and LSTM models are pretty similar to each other. It has been shown that the GRU model works with 5% better performance. However, another time-consuming part of the models is the preprocessing stages in the data feeding stage to the model. When these times are taken into account, the performance difference is negligible. However, the GRU model gave slightly better performance. In this way, it has been

possible to obtain more successful results with a MAPE value of 0.1% compared to the LSTM model with the GRU model in the same amount of time.

7.6.3 CUDA Hardware Acceleration and Alternatives

Forecasting models need to be retrained when a change in the forecasting method is required. The one-year training period for 106 segments in the Powerful model takes 98 hours with a computer with an Nvidia RTX A5000 graphics card. This value is approximately 55 minutes for a segment. This time can be reduced to less than a day using four graphics cards with the power of an Nvidia RTX A5000 graphics card. Similarly, when we want to do a one-month training for 106 road segments, the total training time is 9 hours; The training time for a segment is approximately 5 minutes. This means that 1-month model training for 106 segments can be completed in about 2.5 hours in a similar hardware configuration.

The times given are the accelerated values thanks to the CUDA cores in the Nvidia graphics cards. It is also possible to train and predict using only the processor cores without parallel processing power. There are Intel's unique instruction sets and OneDNN infrastructure offered for this. In the experiments with the OneDNN library and the Intel Xeon processor, the batch size value was kept at 128 and below, and the results were approximately half-speed results obtained with the Nvidia RTX A5000 graphics card. It has been observed that this performance difference increases exponentially as the size of the data matrix fed to the prediction model increases.

8
CONCLUSION

In this section, the results and evaluations of the thesis are given. We have analyzed the factors that affect traffic and provided different techniques to utilize these factors for predicting traffic speed. After that, we proposed a system architecture to simplify feature integration to the traffic speed prediction network.

Although there are numerous models in the literature that address short-term traffic speed estimation, only a few studies have been conducted on long-term traffic speed estimation. There are several methods for making predictions, including deep learning and machine learning techniques like LSTM, ARIMA, CNN, Regression, and Bayesian modeling. This study compared the performance of Polynomial Regression, CNN, LSTM, and GRU models. The results showed that the LSTM and GRU models outperform the other algorithms in a great manner. In particular, the GRU model gave slightly better outcomes than LSTM. After analyzing the estimation results, it was found that the Polynomial Regression approach resulted in a 16.51% MAPE and 8.67 MAE, while the CNN method produced a 13.81% MAPE and 7.27 MAE. The LSTM model performed with a 13.45% MAPE and 7.09 MAE, followed by the GRU model with a 13.23% MAPE and 6.89 MAE. These results were obtained via the trained models with traffic speed data only. After analyzing the results of the base models, further analyses were conducted using LSTM and GRU models.

In order to produce valuable results, it is necessary to work with a comprehensive and consistent dataset. These results were obtained with data provided by the Istanbul Metropolitan Municipality. The dataset includes 106 road segments and was collected in 2018. Up to one-week forecasts with a resolution of 5 minutes data were produced using this dataset. Furthermore, the Segment Connection Tool (SCT) was developed in this thesis to build the connection map using the location data of 106 road segments. We utilized the weather data, including temperature, wind speed, cloudiness, precipitation, and visibility information obtained from Openweather. The accidents and events dataset from the Istanbul

Metropolitan Municipality was also utilized for post-improvement operations. A study was conducted to examine the impact of data sources (multi-feature spaces) on traffic. New approaches for utilizing information gathered from different resources to enhance the forecasting model have been proposed based on the obtained inferences.

Our analysis showed that the traffic speed could be decreased by 12% in harsh weather conditions. This result led us to utilize weather data. One of the standard methods used in the literature to use weather data in traffic speed estimation is the normalization method. The literature indicates that providing environmental data to the network in a similar format only yields a limited increase in success. Several studies have shown that providing different data improves performance by up to 8% in short-term traffic speed estimation. This thesis explored the impact of weather conditions on traffic speed, and a new method called Pre-tuning has been suggested, which involves preprocessing traffic speed data based on the quantitative change in weather conditions. This method uses Effect Rate (ER) and Polynomial Regression (PR) approaches to apply both linear and nonlinear changes in the weather data. This approach aims to enhance the accuracy of traffic speed prediction, particularly during harsh weather. It has resulted in a 26% improvement on rainy days with an amount of 4mm or more, a 20% boost on windy days with a speed of 12 m/s or higher, and a 6% of MAPE increase was observed in cool days where temperatures are below 15°C. By utilizing a weather hybrid model, the MAPE value was reduced from 13.23% to 13.06% in total. Although the forecast model is tailored towards extreme weather conditions and a substantial improvement is possible for such conditions, it has positively impacted the whole year's performance.

Another method we suggest to improve prediction performance is using models suitable for speed groups for road segments. Analyzing past traffic speed averages makes it possible to predict traffic speed with decent accuracy. Some studies have demonstrated the effectiveness of this approach, which relies on simple averaging methods rather than complex techniques like machine learning. Taking this approach one step further, training a model for the relevant speed groups and selecting the forecast model by looking at the region's past average speed values is a new method we introduce in this thesis. This proposed method improved the MAPE value by over 2% in difficult-to-predict areas. Especially in the months that are different from the characteristics of the previous months, such as September, the performance increase has been observed at the maximum level. This method makes it possible to predict time periods that differ significantly from the past more accurately. The MAPE value, which was 13.23% in the base model, was improved to 12.91% using the traffic speed grouping technique.

A road segment is affected by the traffic of neighboring road segments. The analysis has revealed that the diameter of the affected area differs for each road segment. For this reason, experiments were carried out on 106 road segments up to the 1st, 2nd, and 3rd neighbor distances, and the most successful distance of each road segment was found. It has been observed that this distance increases in regions with high vehicle density and shortens in cases where vehicle density decreases and traffic speed increases. This value obtained for each segment area is evaluated as a characteristic feature and matched with that road area. Test results show that there are improvements of up to 50% in terms of MAPE on each road segment. In addition, using spatial data, the MAPE value of 13.23% has been improved up to 12.99% MAPE for all segments.

The methods proposed in the thesis are combined under the hybrid system architecture. Since each road segment's characteristics are unique, segment-specific models are suggested. In this way, we could predict 106 road segments with a MAPE of 12.64%, which was predicted as 13.23% MAPE with the base GRU model. With the post-prediction improvement structure we proposed, we had the chance to customize and improve the traffic forecast. Thanks to post-prediction improvements, the MAPE value has been improved by up to 8.63% MAPE in the cumulative result. Upon analyzing post-prediction improvements, there were significant gains observed. Specifically, the MAPE value at the time of accidents and events improved from 55.45% to 42.62%. Further revisions were made by adjusting the estimates based on legal speed limits, resulting in an improved MAPE value of 9.45% from 12.64%. Increasing the estimation resolution to 15 minutes from 5 minutes for long-term predictions also improved the MAPE from 12.64% to 11.28%. After implementing the suggested solutions, an average improvement rate of 35% has been achieved for long-term traffic speed estimation.

Predicting the long-term traffic speed is challenging due to technical obstacles and lengthy model training sessions. In this thesis, we carried over 6,000 hours of model training for over 100,000 traffic speed forecasting models. These estimation models produced traffic speed estimations for more than 29,300,000 moments of 5 minutes. In order to achieve this, multiple hardware with parallel processing power had to be coordinated by following distributed system architecture principles. Also, the study utilized a significantly more extensive dataset than previous literature examples and focused on long-term traffic speed estimation instead of short-term. The analysis also examined the influence of environmental data on traffic speed and proposed system architecture for integrating features. The model outputs were made customizable through post-estimation improvements. These unique contributions to the literature have advanced the field of traffic speed estimation.

REFERENCES

[1] S. Du, T. Li, X. Gong, Z. Yu, S. Horng, "A hybrid method for traffic flow forecasting using multimodal deep learning," *CoRR*, vol. abs/1803.02099, 2018. arXiv: 1803.02099. [Online]. Available: http://arxiv.org/abs/1803.02099.

[2] H. Chang, Y. Lee, B. Yoon, S. Baek, "Dynamic near-term traffic flow prediction: System-oriented approach based on past experiences," *Intelligent Transport Systems, IET*, vol. 6, pp. 292–305, Sep. 2012. doi: 10.1049/iet-its.2011.0123.

[3] J. Qu, X. Gu, L. Zhang, "Improved ugrnn for short-term traffic flow prediction with multi-feature sequence inputs," in *2018 International Conference on Information Networking (ICOIN)*, 2018, pp. 13–17. doi: 10.1109/ICOIN.2018.8343075.

[4] D. Ma, B. Sheng, S. Jin, X. Ma, P. Gao, "Short-term traffic flow forecasting by selecting appropriate predictions based on pattern matching," *IEEE Access*, vol. 6, pp. 75 629–75 638, 2018. doi: 10.1109/ACCESS.2018.2879055.

[5] Z. Duan, Y. Yang, K. Zhang, Y. Ni, S. Bajgain, "Improved deep hybrid networks for urban traffic flow prediction using trajectory data," *IEEE Access*, vol. 6, pp. 31 820–31 827, 2018. doi: 10.1109/ACCESS.2018.2845863.

[6] H. Bi, Z. Ye, H. Zhu, "Data-driven analysis of weather impacts on urban traffic conditions at the city level," *Urban Climate*, vol. 41, p. 101 065, 2022, issn: 2212-0955.

[7] M.-b. Pang, S.-s. Zheng, Z.-h. Cai, "Simulation of three lanes one-way freeway in low visibility weather by possible traffic accidents," *Physica A: Statistical Mechanics and its Applications*, vol. 433, pp. 161–170, 2015, issn: 0378-4371.

[8] J. Fior, L. Cagliero, "Correlating extreme weather conditions with road traffic safety: A unified latent space model," *IEEE Access*, vol. 10, pp. 73 005–73 018, 2022. doi: 10.1109/ACCESS.2022.3190399.

[9] H.-x. Hu, Z.-z. Lin, Q. Hu, Y. Zhang, W. Wei, W. Wang, "Multi-source information fusion based dlaas for traffic flow prediction," *IEEE Transactions on Computers*, pp. 1–11, 2023. doi: 10.1109/TC.2023.3236902.

[10] Y.-S. Jeong, Y.-J. Byon, M. M. Castro-Neto, S. M. Easa, "Supervised weighting-online learning algorithm for short-term traffic flow prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 4, pp. 1700–1707, 2013. doi: 10.1109/TITS.2013.2267735.

[11] J. van Lint, S. Hoogendoorn, H. van Zuylen, "Accurate freeway travel time prediction with state-space neural networks under missing data," *Transportation Research Part C: Emerging Technologies*, vol. 13, no. 5, pp. 347–369, 2005, issn: 0968-090X. doi: https://doi.org/10.1016/j.trc.2005.03.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X0500029X.

[12] K. Kumar, M. Parida, V. Katiyar, "Short term traffic flow prediction for a non urban highway using artificial neural network," *Procedia - Social and Behavioral Sciences*, vol. 104, pp. 755–764, 2013, 2nd Conference of Transportation Research Group of India (2nd CTRG), issn: 1877-0428. doi: https://doi.org/10.1016/j.sbspro.2013.11.170. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877042813045618.

[13] W. Huang, G. Song, H. Hong, K. Xie, "Deep architecture for traffic flow prediction: Deep belief networks with multitask learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 5, pp. 2191–2201, 2014. doi: 10.1109/TITS.2014.2311123.

[14] Y. Lv, Y. Duan, W. Kang, Z. Li, F.-Y. Wang, "Traffic flow prediction with big data: A deep learning approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 865–873, 2015. doi: 10.1109/TITS.2014.2345663.

[15] K. Wang, C. Ma, Y. Qiao, X. Lu, W. Hao, S. Dong, "A hybrid deep learning model with 1dcnn-lstm-attention networks for short-term traffic flow prediction," *Physica A: Statistical Mechanics and its Applications*, vol. 583, p. 126 293, 2021, issn: 0378-4371. doi: https://doi.org/10.1016/j.physa.2021.126293. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378437121005665.

[16] Y. Zhang, S. Wang, B. Chen, J. Cao, Z. Huang, "Trafficgan: Network-scale deep traffic prediction with generative adversarial nets," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 1, pp. 219–230, 2021. doi: 10.1109/TITS.2019.2955794.

[17] Y. Jia, J. Wu, M. Xu, "Traffic flow prediction with rainfall impact using a deep learning method," *Journal of Advanced Transportation*, vol. 2017, pp. 1–10, Aug. 2017. doi: 10.1155/2017/6575947.

[18] S. Dunne, B. Ghosh, "Weather adaptive traffic prediction using neurowavelet models," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 1, pp. 370–379, 2013. doi: 10.1109/TITS.2012.2225049.

[19] K. C. Dey, A. Mishra, M. Chowdhury, "Potential of intelligent transportation systems in mitigating adverse weather impacts on road mobility: A review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 3, pp. 1107–1119, 2015. doi: 10.1109/TITS.2014.2371455.

[20] A. Nasser, V. Simon, "A novel method for analyzing weather effect on smart city traffic," in *2021 IEEE 22nd International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, 2021, pp. 335–340. doi: 10.1109/WoWMoM51794.2021.00061.

[21] A. Nigam, S. Srivastava, "Hybrid deep learning models for traffic stream variables prediction during rainfall," *Multimodal Transportation*, vol. 2, no. 1, p. 100 052, 2023, issn: 2772-5863. doi: https://doi.org/10.1016/j.multra.2022.100052. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772586322000521.

[22] L. Zhongjian, J. Xu, K. Zheng, H. Yin, P. Zhao, X. Zhou, "Lc-rnn: A deep learning model for traffic speed prediction," Jul. 2018, pp. 3470–3476. doi: 10.24963/ijcai.2018/482.

[23] W. Yuankai, H. Tan, B. Ran, Z. Jiang, "A hybrid deep learning based traffic flow prediction method and its understanding," *Transportation Research Part C: Emerging Technologies*, vol. 90, May 2018. doi: 10.1016/j.trc.2018.03.001.

[24] W. Chen *et al.*, "A novel fuzzy deep-learning approach to traffic flow prediction with uncertain spatial–temporal data features," *Future Generation Computer Systems*, vol. 89, Jun. 2018. doi: 10.1016/j.future.2018.06.021.

[25] Y. Li, S. Chai, Z. Ma, G. Wang, "A hybrid deep learning framework for long-term traffic flow prediction," *IEEE Access*, vol. 9, pp. 11 264–11 271, 2021. doi: 10.1109/ACCESS.2021.3050836.

[26] T. AYAR, F. Atlinar, M. A. GÜVENSAN, H. İ. TÜRKMEN, "Long-term traffic flow estimation: A hybrid approach using location-basedtraffic characteristic," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 30, no. 3, pp. 562–578, 2022.

PUBLICATIONS FROM THE THESIS

Conference Papers

1. E. Bilgin, H. I. Turkmen and M. A. Guvensan, "A Weather Oriented Pre-Tuning Methodology For Long-term Traffic Speed Estimation," 2023 IEEE 24th International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), Boston, MA, USA, 2023, pp. 451-456, doi: 10.1109/WoWMoM57956.2023.00079.

Projects

1. Tübitak 1001 - Uzun Vadeli Trafik Akış Hızı Tahmini için Çoklu Özellik Uzayı Tabanlı Bir Modelin Oluşturulması ve Gerçeklenmesi (TUBITAK1001-120E357)