Traffic Speed Prediction for Short and Long Holidays: Special Approaches

Fatih Ecevit, Yusuf Mert Çelikarslan
Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi, 34220 Istanbul, Türkiye
{fatih.ecevit, mert.celikarslan}@yildiz.edu.tr

***Özetçe* —projede İstanbul sınırları içerisinde hız tahmini yapılması planlanıp, İstanbul Büyükşehir Belediyesi tarafından tutulan veriler kullanılarak sayısal verilerin görselleştirilmesi, analiz edilmesi ve yorumlanması adımları sonrasında uzun ve kısa tatil günleri için trafik tahmini yapan bir araç geliştirilmiştir. Analiz adımında İstanbul'da birçok lokasyondan alınan verilerde harita üzerindeki konumu, zaman-hız verileri esas alınarak grafiksel olarak görselleştirme işlemi gerçekleştirilmiştir. Görselleştirme sonrası yorumlama kısmı ile belirli günlerin diğer günler ile benzerlikleri tespit edilip Random Forest, GradientBoosting, CatBoost, LGBM, XGBRegressor modelleri kullanılarak ortalama %12.63 hata oranı ile kısa ve uzun tatil günleri hız tahminini gerçekleştirebilmektedir.**

Anahtar Kelimeler—traffik hız tahmini, uzun süreli, tatil, xgboost, denetimli öğrenme, zaman serisi

***Abstract*—In this project, a tool that forecasts traffic for long and short holidays was developed after the steps of visualizing, analyzing and interpreting the digital data using the data kept by Istanbul Metropolitan Municipality. In the analysis step, data from many locations in Istanbul are mapped. The graphical visualization process was carried out on the basis of time-speed data. With the interpretation part after visualization, the similarities of certain days with other days are determined and Random Forest, GradientBoosting, CatBoost, LGBM, XGBRegressor models are used to predict the speed of short and long holidays with an average error rate of 12.63%.**

Keywords—Traffic speed prediction, long term, holiday, XGBoost, supervised learning, time-series

I. Introduction

A review of the existing literature shows that many models have been developed for traffic forecasting since the 1960s. Initially, the statistical models ARIMA and its improved models SARIMA and SARIMAX were used in traffic forecasting problems. Later, the success of machine learning models on regression data led to the use of these models in time series. Machine learning models are more successful than statistical models in capturing instantaneous changes in the data, but they perform less well in predicting seasonal changes. Nowadays, with the development of deep learning models, various applications have been tried in time series problems. The advantage of neural networks over machine learning models is the ability to store historical data for long-term predictions. As with any time series, traffic data is divided into two types: univariate and multivariate. The accuracy of the model can be increased by supplementing traffic data with different features such as weather, temperature, road conditions, etc. Traffic data is also differentiated within itself. Forecasting models in the literature use traffic flow, traffic density, number of vehicles on the road and traffic speed. In this project, traffic speed data for 2018 and 2019 sampled at 5-minute intervals in Istanbul Metropolitan Municipality is used. In the literature review, it was observed that [1]DFT-SVR, [2]XGBoost and deep learning models such as [3]LSTM, CNN, ANN are more successful in long-term traffic prediction. Various tests on SVR model and linear regression models have been conducted and the expected results were not obtained. Although acceptable results were found in short-term forecasts, it was concluded that the SVR model was not sufficient for this research topic, considering the variable characteristics of both long-term and holiday days. Therefore, XGBoost and Deep learning algorithms were used. As an error measurement metric, MAPE is used to calculate the model's error of inference and MAE is used to determine the parameters accordingly by finding the worst and best predictions.

II. Methodology

Traffic flow data face continuous and uncertain changes, which complicates the forecasting process. In this paper, we propose a hybrid model based on ENSEMBLE (Random Forest, XGB Regressor, Catboost) and SVR to optimize traffic forecasts during holiday periods. Traffic flow data are often characterized by certain trends and high frequency errors. By transforming the traffic time series into the frequency domain of time via DFT, trend analysis is performed over a certain threshold. Taking into account the natural changes in traffic flow during holiday periods, the basic characteristics of the trend are usually constant. In this case, the trend component is estimated by analyzing historical data. For the remaining data, unusual events and sudden changes are first identified. These abrupt changes are usually of a constant nature, while unusual events involve uncertainty. For unusual events, a special preprocessing process is applied and these data are forecast using the ENSEMBLE method. The final forecast values are obtained by combining trend and unusual data.

A. Decomposition of Traffic Flow Time Series with DFT

Traffic flow time series are characterized by their complex and constantly changing structure. In order to better understand and predict these variable structures, transformational analysis of time series is important. In this context, Discrete Fourier Transform (DFT) provides

an effective method for transforming traffic flow data from time domain to frequency domain. With the DFT, the fundamental frequency components of the traffic flow and their changes over time can be distinguished and analyzed. This decomposition process is a fundamental step for identifying the main trends and anomalies of traffic mobility. This method provides valuable information for the development of traffic management and forecasting models.

B. Estimation of Common Trend

The trend component refers to the trends that traffic flow data consistently shows. As travelers become more environmentally conscious, changes in their travel habits and their environmental impacts become more pronounced. During holiday periods, the trend component of traffic flow changes in a general trend, with a marked increase between years. While the extent of this increase in Western countries varies, it follows a fundamentally similar trend. In this study, a forecast is made by overestimating the trend component based on historical data.

C. Different Characteristics of Holidays

The impact of [4]public holidays on traffic is more limited than long-term holidays and usually falls on a weekend or weekday. For example, on weekday holidays such as April 23, people's holiday plans may be short and focused on destinations close to the city, which causes changes in traffic. Single-day holidays are celebrated on specific days, such as May 1 or August 30, and how these holidays affect traffic is important in the modeling process. During three-day holidays, people may have longer travel plans, creating a different challenge for traffic patterns. People's travel habits may differ during holidays that fall on the weekend, and this informs how holidays should be handled in the modeling process. Special days such as New Year's Eve may create a unique density in traffic, and people's behavior and traffic flow may differ on such days. In summary, the impact of holidays on traffic patterns varies depending on the type of holiday, its duration and people's behavior, and being able to accurately model these factors increases the accuracy of traffic forecasts.

III. Case Study

Traffic congestion is a major problem for people living in metropolitan areas. Public transportation and new road systems cannot solve this problem. Therefore, traffic forecasting systems help both the public and the government to fully understand this problem. When the term traffic forecasting is mentioned, it is divided into two parts: speed and flow. When the average speed data is low, it indicates congestion and when it is high, it indicates no congestion. Flow data can be interpreted in a similar way to speed, but flow data tries to make this interpretation using the number of vehicles on the road at that moment. In this project, we used speed data for Istanbul for the years 2018 and 2019. Istanbul is the largest metropolis in western Turkey, with around 15 million people living in the city. There are [5]7554 different road segments in Istanbul and these roads have different characteristics. There is more population on the European side than on the Anatolian side. This causes traffic to flow in the opposite direction from the European side to the Anatolian side in the morning and evening. In addition, most of the people in Istanbul have migrated to this metropolitan city from Anatolia to work in the past. Therefore, these people return to their hometowns during long vacations. Considering all these parameters, predicting traffic congestion during long holidays becomes a challenging problem. This problem can be divided into short holidays (April 23rd, May 1st, etc.) and long holidays (Eid al-Fitr, Eid al-Adha). For short holidays, if the holiday does not cover a weekend, we forecast a 3-day holiday including a day before, a day after and the holiday itself. If the holiday covers the weekend, we make a 5-day forecast including a before, an after and the 3-day holiday itself. For long holidays, there is no need to separate holidays, so we estimate 13 days, which includes 2 before, 2 after and the 9-day holiday itself.

For the Discrete Fourier Transform (DFT):

$$x_k = \sum_{n=0}^{N-1} x_n \exp\left(-\frac{2\pi i k n}{N}\right), \quad \text{where} \quad k = 0, 1, \ldots, N-1. \tag{1}$$

For the Inverse Discrete Fourier Transform (IDFT)

$$x_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k \exp\left(\frac{2\pi i k n}{N}\right) \quad \text{where} \quad k = 0, 1, \ldots, N-1. \tag{2}$$

The FFT is an efficient algorithm for computing the DFT and its inverse. It reduces the number of operations from $O(N^2)$ to $O(N \log N)$, making it much faster for large sequences.

For the Absolute Percentage Mathematical Error (MAPE):

$$\text{MAPE} = \frac{100}{N} \sum_{i-1}^{n} \frac{|y_i - \hat{y}_i|}{y_i} \tag{3}$$

For the Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (A_t - F_t)^2} \tag{4}$$

A. Data Description

In this paper, the data used to evaluate the performance of the proposed model are collected from the toll stations in Istanbul province from 2018 to 2019. Istanbul is the largest metropolis in western Turkey, with around 15 million people living in the city. There are 7554 different road segments in Istanbul and these roads have different characteristics. We have 2 years of speed data collected at 5 minute intervals. The values of the data on holidays are meaningful for us and only 106 segments of data reflecting the general traffic characteristics of Istanbul were used. We used 2018 data as the training set and in addition to this, we used days that may be similar starting from 1 week before the data used as the test set in 2019. National and religious holidays in 2019 were used as the test set. When the data set was analyzed, it was seen that some segments (2102) had missing data

for nearly 15 days. For this reason, the process of filling the missing data in these segments was carried out based on the same days of the previous months. Missing data in the test set were discarded from the test set and not included in the model performance.

B. Decomposition of Traffic Flow Time Series with FFT

Traffic speed data increases from year to year, but in general there is not much increase in traffic speed data over a two-year period. The difference in speeds between years is usually influenced by roadworks or road regulations specific to that segment. Therefore, there is no clear trend curve between the two years. This curve can be seen in Figure 2. However, seasonal weekly and daily curves are observed, which can be seen in Figure 2. In order to make this seasonal curve meaningful in the data, it is necessary to remove the noise in the data.

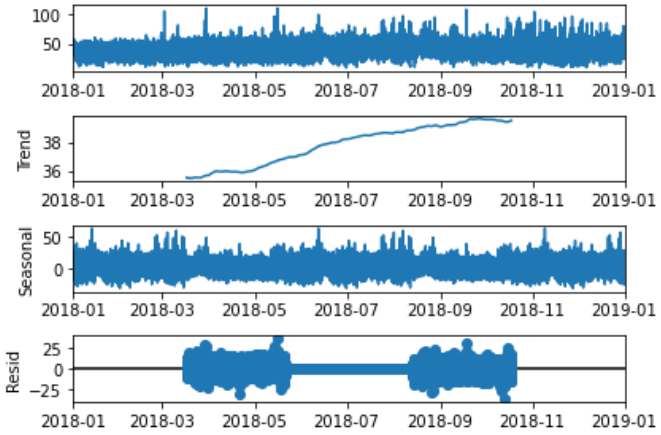

Figure 1 Seasonal Decomposition of Data from 2018 to 2019

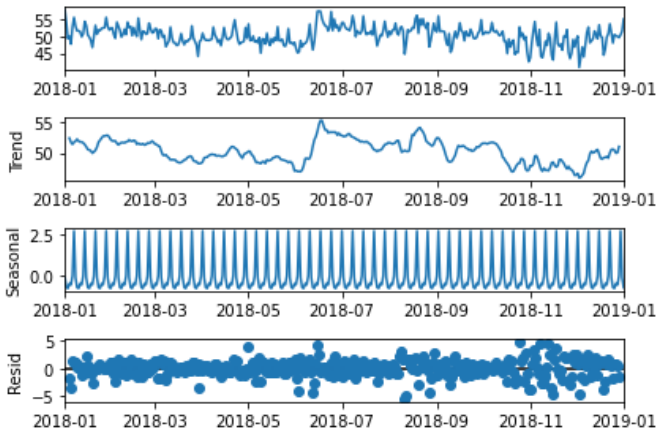

Figure 2 Seasonal Decomposition of Daily Data

This can be done either with readily available decomposition libraries or with the decomposition methods used in signal filtering applications. We have used the FFT algorithm, which is a faster version of the DFT algorithm also available in this paper. This algorithm allows us to filter out noisy data with a low frequency according to a given cutoff frequency. This greatly improved our model performance. Approximately 1 or 2 percent depending on the day forecasted decreased our MAPE error rate. Filtered data and real data can be seen in Figure 3. This filtering process is depended on the cutoff frequency as we mentioned earlier. When cutoff frequency was chosen too low, noisy data can affect performance of proposed model. On the contrary, if it was chosen too high we can lose meaningful data that model needs to forecast speed data of test set.

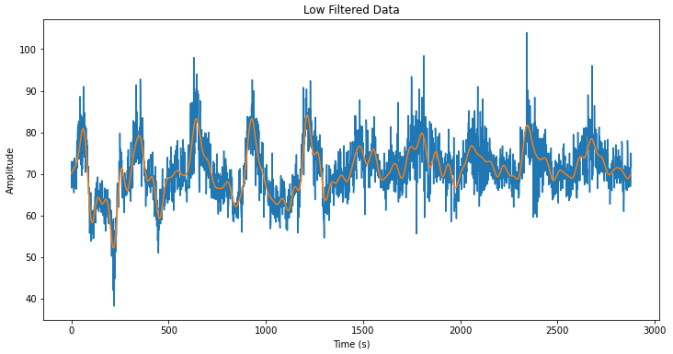

Figure 3 Low Frequency Filtered Data and Real Data

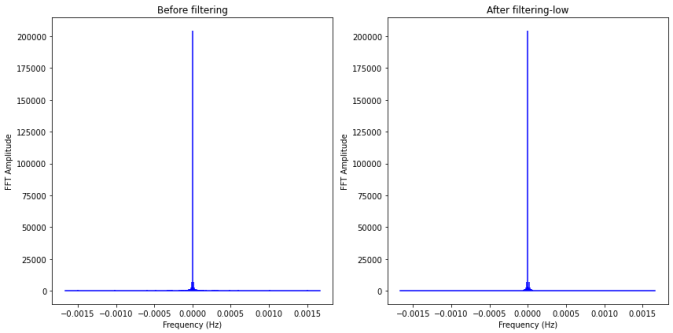

Figure 4 Low Bandwith FFT

C. Selected Days For Different Holidays

The training days selected for each holiday differ. These days can be seen in Table x. From a general point of view, for one-day holidays, it is a good approach to include the one-day holidays preceding that holiday in the training set, to use the same days of the previous weeks for the days before and after that holiday, and for holidays combined with a 3-day weekend, it is a good approach to include the previous weekends in the dataset along with the combined holidays. In addition, the days before the holiday and Fridays show similar characteristics. At the same time, the days after a holiday and Mondays are similar. This is completely different for long holidays. In 2018, Ramadan Eid, which falls on a weekend, was not successful in predicting Ramadan Eid, which is a 9-day holiday in 2019. Therefore, 2018 Eid al-Adha was included in the training

set to improve the model result. This is not the case for Eid al-Adha, but Eid al-Adha, which is a long holiday in both years, showed the best success. Based on this, it can be said that long holidays with similar characteristics provide sufficient results in prediction. There is no need to include Ramadan and Eid al-Adha together in the training set. As a result, long holidays give better results when there is sufficient data set.

D. Prediction Results and Performance Comparison

1) Forecasting Plots: Predicted speed for 2019 data across 2016 segments can be seen in Figure 13. If we look at the graph of short holidays, we can see the characteristic of holidays on feast days. On the preceding and following days there is a decline in the morning and evening working hours, which is not the case on feast days. For long holidays, this effect is spread over a period of 3 to 4 days in Eid al-fitr and a week in Eid al-adha.

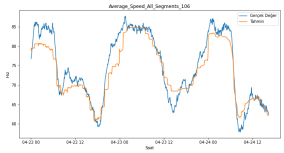

Figure 5 April 23th 2019

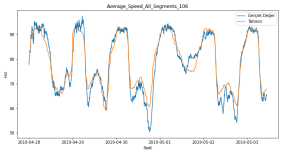

Figure 6 May 1st 2019

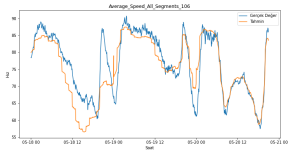

Figure 7 May 19th 2019

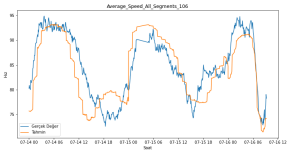

Figure 8 July 15th 2019

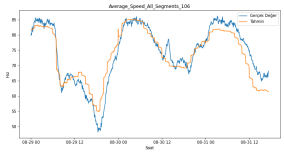

Figure 9 August 30th 2019

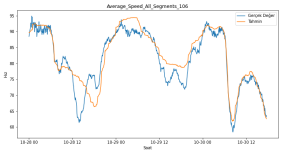

Figure 10 October 29th 2019

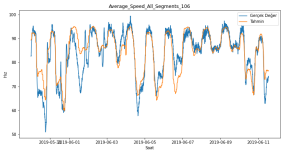

Figure 11 Eid al-fitr 2019

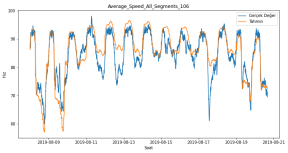

Figure 12 Eid al-adha 2019

Figure 13 'Actual and Predicted' Speed Value Graphs of 106 Segments for Holiday Days

2) Measuring Performance of the Proposed Prediction Method: There are different metrics in order to evaluate performance of models. Mean absolute percentage error, symmetric mean absolute percentage error and mean absolute error are some of them. In time series analysis, MAPE and SMAPE are the most used ones for performance metrics. The cause of that in time series forecasting test set can be varied length and speed data can differentiate between years and locations. The difference bitween MAPE and SMAPE, MAPE is biased towards high values. If the problem is included low values like probability of something SMAPE can be right choice, yet in our problem there is no need for SMAPE. Therefore, we used MAPE metric in this paper and in our model.

Table 1 Average MAPE Error Rate of Different Models Across 106 Segments

Models	April 23rd	May 1st	May 19th	July 15th	August 30th	October 29th	Eid al-fitr	Eid al-adha	Average
RandomForestRegressor	12.26%	14.67%	12.97%	11.28%	13.61%	12.10%	14.14%	11.20%	12.78%
XGBRegressor	11.50%	16.17%	13.15%	11.36%	13.88%	12.32%	14.07%	10.70%	12.89%
CatBoost	14.07%	14.86%	12.99%	11.26%	14.04%	12.17%	14.18%	10.89%	13.06%
Ensemble	12.25%	14.97%	12.97%	11.09%	13.59%	11.99%	13.96%	10.82%	12.71%

Table 2 MAPE Error Rate by Sample Segments

Segment Number	April 23rd	May 1st	May 19th	July 15th	August 30th	October 29th	Eid al-fitr	Eid al-adha	Average
614	5.84%	7.18%	5.66%	5.23%	12.80%	4.95%	7.08%	4.69%	6.68%
420	6.29%	8.55%	6.19%	4.44%	13.63%	9.41%	9.34%	8.28%	8.27%
312	13.68%	9.40%	18.10%	11.89%	15.77%	11.32%	14.51%	7.81%	12.81%
64	7.87%	11.46%	13.36%	7.92%	11.38%	10.04%	11.68%	7.55%	10.16%
321	9.83%	11.40%	8.15%	11.89%	16.60%	11.41%	11.19%	11.47%	11.49%

IV. Conclusions

In this project, a model that predicts the traffic speed in Istanbul is used. For holiday periods, they were examined according to time and speed, and national and religious holidays were grouped based on the similarity of traffic. It was determined that national holidays have different characteristics depending on the different days of the week, and religious holidays have different characteristics with their 9-day holiday period. Considering the similarities, five different models (Random Forest, GradientBoosting, CatBoost, LGBM, XGBRegressor) were used for prediction and the speed prediction of this model so far was made. The impact of weather was considered but not included as data. The average of missing data values was filled in, the estimated negative impact of the accidents was taken into account, and it was planned to develop the system with more comprehensive data sets in the future.

References

[1] X. Luo, D. Li, and S. Zhang, "Traffic flow prediction during the holidays based on dft and svr," *Journal of Sensors*, vol. 2019, 2019.

[2] B. Lartey, A. Homaifar, A. Girma, A. Karimoddini, and D. Opoku, "Xgboost: a tree-based approach for traffic volume prediction," in *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2021, pp. 1280–1286.

[3] N. C. Petersen, F. Rodrigues, and F. C. Pereira, "Multi-output bus travel time prediction with convolutional lstm neural network," *Expert Systems with Applications*, vol. 120, pp. 426–435, 2019.

[4] İ. TAKAK, H. GÖRMEZ, H. İ. TÜRKMEN, and M. A. GÜVENSAN, "Kısa, orta ve uzun vadeli trafik akış hızı tahmini ve görselleştirilme aracı," *International Journal of Advances in Engineering and Pure Sciences*, vol. 33, no. 4, pp. 568–580, 2021.

[5] Ş. Yaprak and A. Akbulut, "Trafik kaza ve denetim istatistikleri," *Polis Akademisi Yayınları*, vol. 75, pp. 8–13, 2019.