CETRA2024 15-17 May 2024, Cavtat, Croatia
8th International Conferenceon Road and Rail Infrastructure

DAILY TRAVEL DEMAND PREDICTION IN RAIL SYSTEMS BY USING DEEP LEARNING TECHNIQUES

Yalçın Alver, Halil Uğur Ercan
Ege University, Civil Engineering Department, Türkiye

Abstract

Future travel demands should be predicted accurately in order to plan, make operational decisions, and manage urban public transportation systems. The success of the developed prediction model will directly affect the success of the transportation plan. Many factors, such as day of the week, weather, whether there is a large organization in the city, whether schools are open, affect the demand for urban public transportation. Organizations such as celebrations, festivals, sports competitions, or changes in the weather may cause a different travel demand than expected. The unexpected increase in travel demand makes it difficult to manage the transportation system. Daily travel demand predictions should be considered when making many operational decisions, such as arranging the frequency of services and determining the number of personnel to serve. Trip data of public transportation systems can be obtained easily by utilizing smart transportation cards. This large-scale dataset allows modelling the relationship between the above-mentioned factors and public transport usage demand using deep learning techniques. This paper presents a comprehensive study on the development of a daily passenger demand prediction model for rail systems using deep learning techniques. The study incorporates a wide range of external factors, including weather conditions, day of the week, public holidays, and the occurrence of specific events such as football matches, to create a prediction model. Various deep learning models with different variable sets were developed using the daily travel data of the 2019 Istanbul M2 Yenikapı – Hacıosman metro line. The impact of various external factors on travel demands were systematically examined by assessing the prediction performances of five different deep learning models created with different set of variables.

Keywords: public transit, demand prediction, deep learning, railway systems

1 Introduction

In order to plan transportation well, it is necessary to create a good prediction model with the data obtained from transportation systems. The success of the created model in predicting future demand will directly affect the success of the planning. Many factors, such as the day of the week, air temperature, precipitation, whether there is an organization such as celebrations, festivals, sports competitions, etc., whether the schools and universities are in course period, affect the demand for public transport. Prediction of daily travel demand in public transport systems should be considered while making many operational decisions, such as setting trip frequency, determining the number of personnel to work. Also, the unexpected increase in travel demand makes it difficult to manage emergency situations. Accurate prediction of travel demand is of great importance for daily operational safety management and emergency prevention studies [1].

DOI: https://doi.org/10.5592/CO/CETRA.2024.1512

With technological developments, it has become much easier to collect data, and the concept of big data has emerged with these large-scale data. The data provided by cameras, sensors, smart cards that are used in intelligent transportation systems and GPS data of mobile phones constitute the big data that contributes to the field of transportation. Since big data is a collection of large and complex data sets, it is very difficult to process this amount of data using traditional data processing techniques [2]. In this study, it is desired to process large-scale transportation data, which is difficult to process with traditional methods, with an innovative method by using of deep learning techniques. The use of deep learning techniques in the field of transportation has recently become very popular.
In the literature, there are studies that aim to use deep learning techniques to find a solution in the transportation field. Zhu et al. [3] proposed an artificial neural network model that predicts the number of daily entrances and exits at the station by using the data of air temperature, precipitation, working days and traffic conditions of the road near the station, with the station entry-exit data obtained with the smart card. Similarly, Liu et al. [4] developed artificial neural network models that predict the station-based passenger demand of the rail system based on variables such as current month, day of the month, day of the week, whether that day is a holiday or not. The point where the models showed differences is that not all variables are included in all models. It was stated that the model in which all variables were included was the model with the highest prediction success. Xiong et al. [1] proposed LSTM and CNN models that predict daily and daily short-time travel demand (for 10 minutes later) with the passenger demand data of the metro system consisting of 15 lines and 47 stations in Beijing, China. In addition, the success of the models that are developed by three different traditional methods and the LSTM and CNN models were compared. It has been stated that LSTM and CNN models give much better results than traditional methods. In this study, only the spatial and temporal variables were used. A similar study was carried out in Chongqing, China, and a model that predicts the number of passengers leaving the subway station for 10 minutes in the future according to the spatial-temporal variables in the subway system was proposed and the model was named ST-LSTM (Spatio-Temporal LSTM) [5]. This developed model was compared with the models created with SARIMA, PSO-SVR, LSTM techniques and it was stated that the most successful results were obtained with the proposed model.
This study aims to develop daily passenger demand prediction models with different group of variables and investigate the effect of the variables on the prediction results.

2 Method

2.1 Data

The COVID-19 pandemic has affected all parts of life, including transportation systems. Restrictions due to the pandemic had been continuously changed in Turkey after 2020. So that, the data from before the pandemic were used in this study to get more consistent results. The daily smart card data that were recorded at every station of M2 metro line of İstanbul in 2019, were provided by İstanbul Metropolitan Municipality for this study. Besides, the weather data, which include daily average temperature, amount of precipitation and average wind speed, were extracted for every single day of 2019. The national holidays, day of the week, month of the year were added to the dataset. Most students use public transport to get to school or university. So that, the academic schedules of the three universities (Istanbul University, Istanbul Technical University and Boğaziçi University) that are near to M2 metro line and the education schedule for middle and high schools were examined. The data on whether these educational institutions are in the course period was added to the dataset. Another important case that significantly affects the use of public transport is an event that takes place in the city with high participation. In this case, there is a football stadium that is

so close to one of the metro stations called “Seyrantepe”. Galatasaray football team plays its football matches in this stadium almost every two weeks. So that, it is expected that passenger demand at this station will increase on the days that there is a football match in this stadium. The data of whether there is Galatasaray’s football match was added to the dataset. The data was separated into two: a train dataset and a test dataset. The last day of the sequential six days was separated to the test dataset, and the rest of them were included to the train dataset. In this way, it is guaranteed that the number of each day of the week will be separated equally within datasets.

2.2 Deep learning model

Five deep neural network (DNN) models that have the same architecture were developed and trained for this study. Each of them has a different group of input variables. Input variables and groups are shown in Table 1. Different learning rates, loss functions, and optimizers were tried while developing the deep learning model. After this empirical process, mean square error was selected as a loss function, adam algorithm was selected as an optimizer, default learning rate 0.001 was used since these are the best performing parameters. The mean square error, mean absolute error and R-squared score were used to evaluate the model. The training processes were carried out on an open source TensorFlow Library.

Table 1 Variable groups that were used in the development of the models.

Variable Type	Variable description	Group 1	Group 2	Group 3	Group 4	Group 5
Numerical Variables	Temperature	X	X		X	
	Precipitation	X	X		X	
	Wind Speed	X	X		X	
Categorical Variables	Day of Week	X	X	X	X	X
	Month	X	X	X	X	X
	Holiday	X	X	X	X	X
	Middle and high schools are in course period or not	X	X	X		
	İstanbul University is in course period or not	X	X	X		
	İstanbul Technical University is in course period or not	X	X	X		
	Boğaziçi University is in course period or not	X	X	X		
	There is Galatasaray’s football match or not	X				

The model was also retrained with different groups of input variables to evaluate the effects of variables on the model’s success. So that, the effect of each variables on public transit usage were revealed in this study.

3 Results

The evaluation metrics of developed models are presented in this section. As specified in the method section, the models were trained with different groups of inputs. Table 2 describes the evaluation metrics for a model that includes all variables (Group 1). Also, the number of actual total and predicted total passengers for all test sets are given in this table. It is seen that the model metrics are varied among the stations. The model was fitted well to the real data with more than 0.80 R-squared score for 12 of 16 stations. At the stations named Haliç, Şişhane, Taksim, and Seyrantepe, the model was not fitted as well as other stations. It should be noted that RMSE and MAE values are higher at the stations that have more passenger demand than other stations, even if the model fits the real data well.

Table 2 Evaluation metrics of the test set of the model that was trained with group 1 variables.

Stations	Group 1			Actual total passenger	Predicted total passenger
	R-Squared	RMSE	MAE		
Yenikapı	0.84	5,206	3,580	3,881,761	3,925,133
Vezneciler	0.85	2,241	1,571	1,228,324	1,246,930
Haliç	0.59	1,825	1,139	776,836	769,874
Şişhane	0.69	2,867	1,428	1,350,810	1,389,763
Taksim	0.52	5,235	2,585	2,419,974	2,463,531
Osmanbey	0.80	6,356	3,291	2,730,735	2,807,518
Mecidiyeköy	0.84	4,000	2,647	3,301,242	3,353,153
Gayrettepe	0.85	4,806	2,628	2,155,575	2,195,408
Levent	0.85	4,074	2,325	2,413,554	2,435,266
4 Levent	0.87	2,581	1,621	1,817,777	1,843,376
Sanayi	0.86	1,435	831	664,829	673,481
İTÜ-Ayazağa	0.87	4,743	2,909	1,661,735	1,679,772
Atatürk Sanayi	0.85	1,253	700	448,565	454,585
Darüşşafaka	0.85	588	392	309,630	311,949
Hacıosman	0.88	1,334	906	1,052,945	1,069,338
Seyrantepe	0.52	3,240	1,656	474,010	467,424
Total	0.78	3,647	1,888	26,688,302	27,086,504

Evaluation metrics for the other models that trained with different groups of variables are shown in Table 2 and Table 3. Evaluation metrics vary among the models. The model that was trained with variable group 1 performed the best, and the model that was trained with variable group 5 performed the least, according to all evaluation metrics.

Table 3 Evaluation metrics of the test set of the models that were trained with group 2 and 3 variables.

Stations	Group 2			Group 3		
	R-Squared	RMSE	MAE	R-Squared	RMSE	MAE
Yenikapı	0.82	5,385	3,850	0.84	5,167	3,758
Vezneciler	0.85	2,254	1,562	0.84	2,356	1,662
Haliç	0.56	1,881	1,164	0.46	2,105	1,293
Şişhane	0.68	2,902	1,434	0.72	2,711	1,442
Taksim	0.46	5,568	2,844	0.50	5,376	2,811
Osmanbey	0.73	7,450	3,572	0.74	7,342	3,589
Mecidiyeköy	0.78	4,743	3,204	0.79	4,550	2,969
Gayrettepe	0.80	5,523	3,081	0.80	5,441	2,899
Levent	0.82	4,427	2,477	0.81	4,648	2,613
4 Levent	0.82	2,993	1,659	0.83	2,939	1,708
Sanayi	0.81	1,706	912	0.81	1,688	890
İTÜ-Ayazağa	0.86	4,970	2,914	0.85	5,167	3,012
Atatürk Sanayi	0.82	1,364	718	0.81	1,411	767
Darüşşafaka	0.82	647	403	0.80	680	457
Hacıosman	0.83	1,556	935	0.86	1,425	962
Seyrantepe	-0.09	4,889	2,564	-0.21	5,158	2,710
Total	0.71	4,111	2,081	0.70	4,099	2,096

Table 4 Evaluation metrics of the test set of the models that were trained with group 4 and 5 variables.

Stations	Group 4			Group 5		
	R-Squared	RMSE	MAE	R-Squared	RMSE	MAE
Yenikapı	0.80	5,718	3,850	0.84	6,519	4,512
Vezneciler	0.81	2,520	1,562	0.84	2,804	2,036
Haliç	0.55	1,910	1,164	0.46	2,128	1,421
Şişhane	0.68	2,919	1,434	0.72	3,286	1,619
Taksim	0.44	5,648	2,844	0.50	5,822	2,942
Osmanbey	0.74	7,355	3,572	0.74	7,899	3,721
Mecidiyeköy	0.78	4,733	3,204	0.79	4,712	3,093
Gayrettepe	0.80	5,523	3,081	0.80	5,831	3,228
Levent	0.81	4,604	2,477	0.81	4,837	2,929
4 Levent	0.81	3,100	1,659	0.83	3,347	2,147
Sanayi	0.81	1,706	912	0.81	1,741	989
İTÜ-Ayazağa	0.80	5,933	2,914	0.85	6,107	3,898
Atatürk Sanayi	0.81	1,418	718	0.81	1,375	775
Darüşşafaka	0.81	662	403	0.80	742	492
Hacıosman	0.82	1,584	935	0.86	1,775	1,128
Seyrantepe	-0.17	5,089	2,564	-0.21	5,187	2,739
Total	0.69	4,254	2,081	0.70	4,517	2,354

The trend of the passenger demand on the test dataset for Yenikapı station and the trend of the predicted passenger demand are shown in Figure 1.

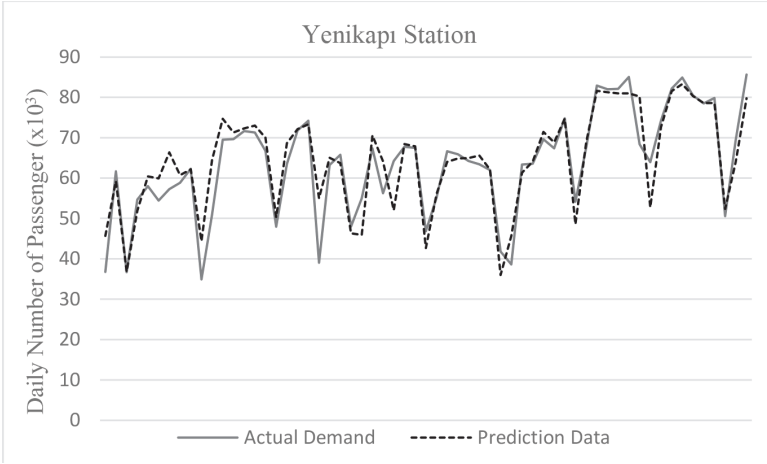

Figure 1 Trend of the test dataset and predicted passenger demand for Yenikapı station

4 Results

In this study, the daily passenger demand on the M2 Metro line of İstanbul was predicted by using deep neural networks with different independent variables. It is shown that passenger demand is affected by various variables. So that, the model success increases if different variables specific to the stations or metro line is used. The model that was developed with all input variables was performed the best in this study. It should be noted that the model accuracy is significantly increased for Seyrantepe station when the variable of whether there is Galatasaray's football match added to the model. After adding this variable R-squared score for this station increased from -0.09 to 0.52 (see Table 1 and Table 2). Also, there may be other organizations, such as national football matches, different tournaments, concerts, etc., in this stadium. The model accuracy may probably be increased if the variables about other organizations are added. The variable of whether there is Galatasaray's football match was the only variable that represented an event that took place in the city with high participation. The model's success can be increased if other organizational data for other stations are added to the prediction model.

The model success was low at the stations Haliç, Şişhane, and Taksim. This may be because of the region in which these stations are located. These locations are highly visited for touristic purposes. If the variables related to this situation are added to the prediction model, the prediction accuracy may probably be increased.

One of the purposes of this work was to forecast passenger demand. However, only one year of data could be obtained for now. One year of data is needed for training the forecasting model to make the model learn all temporal changes within a year, and then this model is needed to be tested with future data. Because of the lack of data, a prediction model for passenger demand was developed in this study. In the next study it is aimed to obtain the passenger demand data for 2018 and train the model with 2018 data, then test the model with 2019 data. So that, the forecasting objective can be satisfied. Besides, it is desired to extract more variables that can be used in a forecasting model. Using other deep learning techniques, such as RNN and LSTM is another objective for future work.

References

[1] Xiong, Z., Zheng, J., Song, D., Zhong, S., Huang, Q.: Passenger Flow Prediction of Urban Rail Transit Based on Deep Learning Methods, Smart Cities, 2 (2019) 3, pp. 371–387, DOI: 10.3390/smartcities2030023

[2] Neal, J.G.: Data, Data In the Cloud, Libraries Are Clearly Well-Endowed: Metadata/Linked Data and Issues of Opportunity, Responsibility, Authority and Unintended Consequences, Johns Hopkins University, OCLC Collective Insight Symposium, Baltimorei MD, 2013.

[3] Zhu, H., Yang, X., Wang, Y.: Prediction of Daily Entrance and Exit Passenger Flow of Rail Transit Stations by Deep Learning Method, Journal of Advanced Transportation, 2018, DOI: 10.1155/2018/6142724

[4] Liu, L., Chen, R.C.: A MRT daily passenger flow prediction model with different combinations of influential factors, 31st IEEE International Conference on Advanced Information Networking and Applications Workshops AINAW 2017, pp. 601–605, Taipei, China, 27-29.03.2017, DOI: 10.1109/WAINA.2017.19

[5] Tang, Q., Yang, M., Yang, Y.: ST-LSTM: A Deep Learning Approach Combined Spatio-Temporal Features for Short-Term Forecast in Rail Transit, Journal of Advanced Transportation, 2019, DOI: 10.1155/2019/8392592