Research Article

Melek Nar, Seher Arslankaya*

Passenger demand forecasting for railway systems

https://doi.org/10.1515/chem-2022-0124
received December 2, 2021; accepted December 23, 2021

Abstract: The rapid increase of the population and the number of motor vehicles brought about the transportation problem today. It has brought the efforts of the operators to determine the headway of the vehicles during the day in order to minimize the waiting times of the passengers at the stops and increase the satisfaction of the passengers, taking into account the passenger demand. Nowadays, especially during the current pandemic period (COVID-19), passenger demand forecasting becomes much more significant, so that measures can be taken and headway planning can be made to adjust the social distance by identifying the number of passengers in advance. In this study, the significance of demand forecasting in the railway sector is considered, and the study tackles the issue in two stages: on line and station basis that make the study different from others. In the first stage of the study, passenger demand forecasting is made on line basis with statistical techniques such as regression analysis and simple average, the mean absolute percentage error values are calculated and compared. Regression analysis is conducted with SPSS Statistics 21.0 programme. In the second stage of the study, passenger demand forecasting is made with artificial neural network and machine learning (ML) algorithms technique on station basis and the error values (mean absolute error, BIAS, mean squared error, mean absolute percentage error, and root mean squared error) are compared. As a result of the study, while the best demand forecasting method is simple average on line basis, it is seen that the most successful and reliable results for demand forecasting on station basis are obtained through decision tree, which is one of the ML algorithms.

Keywords: railway systems, artificial neural networks, demand forecasting, machine learning, regression analysis, COVID-19

1 Introduction

Urbanization, which started with industrialization in the world, has brought many problems with it. Transportation is one of these problems. Public transportation has come to the fore in developed countries in order to solve the problem of urban planning and traffic congestion and to realize efficient passenger transportation [1–4]. Public transportation types are supported and encouraged to prevent transportation problems. At this point, reliable, fast, and convenient urban rail systems come to the fore for public transportation [5].

Effective management of rail systems, which provide an effective solution to the transportation problem in big cities which get crowded day by day, is of great importance in terms of operational efficiency and passenger service satisfaction [6].

Regression models are powerful tools for characterizing the relationship between demands and other important factors, but they require complicated modeling techniques and enormous data to produce acceptable results. Expert system models are built up of rules for demand forecasting based on the knowledge of a human expert. It is extremely difficult to transform the knowledge of an expert to mathematical rules [7,8].

Demand forecasting is important for the correct planning of the supply of the service to be caused by the passenger demand. Passenger demand forecasting plays an important role in decision making and planning. In recent years, artificial neural networks (ANNs) are often preferred because they can be used in any linear or non-linear function. The biggest advantage of ANNs over other prediction models is learning and working with incomplete and improper data. ANNs are used in many areas such as banking, economy, energy demand, tourism forecast modeling, supply chain, and transportation...
[9–11]. Short-term forecasting is the key to the success of transportation operations planning such as train timetabling and operator allocation [8].

Machine learning (ML) is a subset of artificial intelligence, where the ML algorithm acts or performs the task without being explicitly programmed. The machine can learn automatically from the past raw data to generate predictive models based on predesigned algorithms. In general, there are two types of learning algorithms: supervised and unsupervised learning. Supervised ML algorithms learn from labeled data: input and output. The algorithm is responsible for finding the relationship between the input and the output and stops learning when it achieves an acceptable performance level [12].

As a result of the literature reviews, it has been observed that the studies carried out in the past mainly deal with the different stages of the rail transportation systems. In this study, the passenger demand of the future period is estimated by using the previous period data of the Yenikapı–Kirazlı metro line. Unlike the studies in the literature, this study includes two stages for demand forecasting. In the first stage of the study, line-based demand forecast is made for weekdays and weekends, while in the second stage of the study, the estimation is made on a station basis. Making the study both line based and station based contributes to the literature. Considering the critical importance of passenger demand forecasting, demand estimation has been made using ANNs and ML algorithms for station basis, and statistical techniques, such as regression analysis technique and simple average method, for line basis. Also, ML algorithms, such as decision tree, linear regression, random forest (RF), polynomial regression, and support vector regression (SVR), as well as an ensemble model of these, are used in building the prediction models. After obtaining the results of demand forecast, finally, their performances are evaluated using the mean absolute error (MAE), BIAS, mean squared error (MSE), mean absolute percentage error (MAPE), and the root mean squared error (RMSE).

To the best knowledge of the authors, this study is one of the first attempts to make use of both line-based and station-based forecasting simultaneously.

The article is organized as follows: Section 2 reviews the literature of passenger demand forecasting. Section 3 explains the materials and methods used in the study. Section 4 describes the structure and mathematical formulation of the proposed system. Section 5 compares the predictive performance between the proposed approaches. Finally, Section 6 gives the conclusion of the study.

2 Literature review

Travel demand model, also known as travel demand forecasting model, aims to establish a spatial distribution of trips and estimate the travel behavior and travel demand for a specific time frame based on certain assumptions. Travel demand forecasting is an attempt to predict the future travel pattern and to quantify it [13]. Knowing the number of passengers who use the public transportation vehicle is very important in terms of planning. When the literature is reviewed, it is seen that algorithms such as ANNs, deep learning, ML, and statistical methods are used to make this prediction.

Few studies exist in the literature for rail transportation systems, and, the current studies use SVR, fuzzy, and ANNs algorithms. Zhao and Mi [14] proposed a novel hybrid model, specially, the singular spectrum analysis–wavelet packet decomposition–SVR model, for short-term high-speed railway passenger demand forecasting that explicitly considers the relevance of neighbor time data. Li and Sheng [15] proposed empirical models for forecasting the market share of air and high-speed rail integration service using multinomial logit-based discrete choice models. A stated preference survey was conducted to estimate the parameters in the proposed models. Dou et al. [16] applied fuzzy set theory, portfolio optimization, and train operation adjustment theory. First, fuzzy passenger demand forecasting model is established to predict passengers during holidays. The results show that the fuzzy passenger forecasts’ predict is more accurate than autoregressive integrated moving average (ARIMA) model.

Then, authors study train deployment theory in sudden large passenger flow considering the total operation cost of the train dispatching, unserved passenger volume, the required number of trains, station capacity, section capacity, and train set configuration. Lastly, train dispatching optimization model is established, the validity of this model is illustrated with a case study. Çelebi et al. [8] adopted neural networks to develop short-term passenger demand forecasting models to be used in operational management of light rail services. A multi-layer perceptron model is preferred due to not only its simple architecture but also proven success of solving approximation problems. For eliminating the significant seasonality in time slots, each time slot is handled independent of the others, and an ANN based on daily data is developed for each.

Jin et al. [17] developed an approach for short term air passenger forecasting on the basis of the variational mode decomposition, autoregressive moving average model
(ARMA), and kernel extreme learning machine. Gong [18] formulated the “intercity passenger demand forecast problem.” To solve the problem, an algorithm termed as ARMA-GRNN was proposed. Cyril et al. [13] reported forecasts of public transport demand from Trivandrum district to five other districts in Kerala using the ARIMA model. Kim and Shin [19] developed a model for forecasting short-term air passenger demand using big data from search engine queries. To ensure that it had some predictive ability, time shifts ranging from 0 to 11 months, at 1-month intervals, were used to develop the forecasting model.

Ke et al. [20] explored the short-term passenger demand forecasting under the on-demand ride service platform via a novel spatio–temporal DL approach. Accurate real-time passenger demand forecasting can provide suggestions for the platform to rebalance the spatial distribution of cruising cars to meet passenger demand in each region, which will improve the car utilization rate and passengers’ degree of satisfaction. Li et al. [21] aimed to predict the passenger demand under hybrid ride-sharing service modes. Fuloria [22] compared exponential smoothing, multiple regression, and LSTM for forecasting accuracy in training as well as validation datasets. Bai et al. [23] proposed a novel deep learning framework for multi-step citywide passenger demand forecasting, and formulated the citywide passenger demand on a graph and employed the hierarchical graph convolution architecture to extract spatial and temporal correlations simultaneously.

Picano et al. [24] addressed the passengers demand forecasting problem acting with real data from Didi Chuxing, the most famous Transportation Network Companies (TNC) in China. In order to forecast the future behavior of passenger demands, authors proposed a Chaos Theory (CT) approach to deal with the corresponding non-linear scalar time series. Picano et al. [25] addressed the problem of the prediction of the service requests for the TNCs. In particular, different algorithms for different real datasets have been presented. The predictive methods designed for three analyzed datasets are based on the CT principles and the corresponding phase space has been reconstructed, the chaotic behavior studied, through the analysis of the largest Lyapunov exponent. Furthermore, a different CT based algorithm has been proposed for the different datasets studied. Alekseev and Seixas [26] developed models based on ANNs for the air transport passenger demand forecasting. It is found that neural processing can outperform the traditional econometric approach used in this field and can accurately generalize the learnt time series behavior, even in practical conditions, where a small number of data points are available.

Table 1 summarizes the methods in the literature. It can be obviously seen that the studies are line based. Unlike other studies, this study is both station-based and line-based that neural processing can outperform the traditional econometric approach used in this field and can accurately generalize the learnt time series behavior, even in practical conditions, where a small number of data points are available.

### 3 Materials and methods

This section briefly introduces the demand forecasting methods used: statistical analysis techniques, such as

| Author            | Methods  | Transportation systems | Demand forecasting |
|-------------------|----------|------------------------|--------------------|
| Jin et al. [17]   | Regression X | Railway X | Line/Region based |
| Li et al. [21]    | X        | Railway X X | Station based |
| Fuloria [22]      | X        | Railway X | Station based |
| Bai et al. [23]   | X        | Railway X | Station based |
| Picano et al. [24]| X        | Railway X | Station based |
| Picano et al. [25]| X        | Railway X | Station based |
| Zhao and Mi [14]  | X        | Railway X | Station based |
| Cyril et al. [13] | X        | Railway X | Station based |
| Ke et al. [20]    | X        | Railway X | Station based |
| Kim and Shin [19] | X        | Railway X | Station based |
| Li and Sheng [15] | X        | Railway X | Station based |
| Dou et al. [16]   | X        | Railway X | Station based |
| Gong [18]         | X        | Railway X | Station based |
| Çelebi et al. [8] | X        | Railway X | Station based |
| Alekseev and Seixas [26]| X | Railway X | Station based |

*In this study

*: Refers to the methods used by the authors of this article.

X: Refers to the methods used by other article authors in the literature.
regression analysis technique and simple average methods, and also ANNs, support vector machine (SVMs), RF, linear regression, polynomial regression, and decision tree which are machine learning algorithms.

### 3.1 Regression analysis technique

Unlike the time series models that create demand forecasts for future periods using the demand data of the past, regression analysis is a statistical forecasting method that uses the relationship between variables [27]. It is accepted that there are two variables in the method and the relationship between them is linear [28]. The equation of a line representing the linear relationship between dependent and independent variables is formulated with bivariate regression analysis [29]. The formula used in the method is given in equation (1) [30].

\[ Y_i = a + bX_i \]  

where \( Y_i \) = dependent variable, \( a \) = initial value of the regression line, \( b \) = regression line slope, and \( X_i \) = independent variable [31].

### 3.2 Simple average method

The simple average method is to collect the past period data one by one and divide them by the number of periods [32]. The advantage of the method is that it provides a flattened forecasting by using all periods and is easy to apply. The mathematical equation of the simple average method is shown in equation (2) [33].

\[ F_{t+1} = \frac{1}{t} \sum_{i=1}^{t} Y_i \]  

where \( t \) = period, \( F_{t+1} \) = predicted value of the next period, and \( Y_i \) is the actual demand value in the period \( i \).

When a new observation, \( Y_{t+1} \), is available, this new value is added to equation (2) while creating a forecast for time \( t + 2 \), and equation (3) is obtained [33].

\[ F_{t+2} = \frac{1}{t+1} \sum_{i=1}^{t+1} Y_i \]  

#### 3.3 ANNs

ANNs are computer systems developed to automatically perform the functions of the human brain, such as learning, understanding, and gaining experience such as revealing new information. Artificial neural networks consist of input layer, output layer and hidden layer. These layers generate models for the data in the computer network. Using these models, they can make a decision by looking at the examples of the events, making generalizations about the relevant event and collecting information, using the information learned about the examples for situations they will encounter later. ANNs consist of layers connected in parallel. These layers are a structure simulated according to the nervous system in the human brain. ANNs [34,35] consist of three layers, including the input layer, one or more hidden layers, and output layers. The connections between these layers form the function of the network. By adjusting the weight values of the layers that are connected with each other, the network is trained for the realization of a certain function. Thus, an output is produced for an input in the

![Figure 1: Steps of the study.](image-url)
network. A neuron input is the output of another neuron. Outputs are transmitted by synapse links. Synaptic connection weights are expressed in numerical values [9].

3.4 ML techniques

ML techniques are able to learn patterns and solve complex problems just by processing (very often) large size databases. Probably, the most classical machine learning approach is constituted by the ANN paradigm [36].

(a) Decision tree: A decision tree is a tree whose internal nodes can be taken as tests (on input data patterns) and whose leaf nodes can be taken as categories (of these patterns). These tests are filtered down through the tree to get the right output to the input pattern. Decision tree algorithms can be applied and used in various different fields [37].

(b) RF regression: RF is an ensemble learning classification and regression method suitable for handling problems involving grouping of data into classes. The algorithm was developed by Breiman and Cutler [38,39].

(c) Linear regression: Linear regression is the most common predictive model to identify the relationship among the variables. Apart from univariate or multivariate data types the concept is linear [40–42].

(d) Polynomial regression: Polynomial regression is a regression algorithm that models the relationship between a dependent variable (y) and independent variable (x) as nth degree polynomial [43].

(e) SVR: SVM is one of the supervised learning models for classification and regression [9,10]. SVM for regression is specifically said to be SVR. SVM can be linear or non-linear using respective kernel functions [40,44,45].

3.5 Evaluation of different methods

We use five different measures of forecast errors for evaluating the model performance and the accuracy of the methods; they are MAE, MSE, BIAS, and MAPE [10,46–48] and RMSE.

Assume $X_1, X_2, \ldots, X_n$ are actual data and $F_1, F_2, \ldots, F_n$ are forecasted data, and then the $n$ values of forecast errors, $e_1, e_2, \ldots, e_n$, are given by $e_1 = F_1 - X_1$, $e_2 = F_2 - X_2, \ldots, e_n = F_n - X_n$.

Table 2: Model summary for weekdays (Friday, Monday, Tuesday, Wednesday, Thursday – 2019)

| Model   | $R$ | $R$ square | Adjusted $R$ square |
|---------|-----|------------|---------------------|
| 1       | 0.977 | 0.954    | 0.949               |

Table 3: Model summary for weekends (Saturday, Sunday 2019)

| Model   | $R$ | $R$ square | Adjusted $R$ square |
|---------|-----|------------|---------------------|
| 1       | 0.927 | 0.860    | 0.854               |

Table 4: Coefficients for weekdays

| Model          | Unstandardized coefficients | Standardized coefficients | $t$ | Sig. |
|----------------|-----------------------------|---------------------------|-----|------|
|                | $B$                         | Std. error                | Beta|      |
| 1 (Constant)   | -202980.131                 | 106543.892                | -1.905 | 0.063 |
| Monday_2019    | 1.678                       | 0.292                     | 0.252 | 5.749 | 0.000 |
| Tuesday_2019   | 1.727                       | 0.325                     | 0.258 | 5.314 | 0.000 |
| Wednesday_2019 | 0.457                       | 0.547                     | 0.065 | 0.834 | 0.409 |
| Thursday_2019  | 2.011                       | 0.609                     | 0.308 | 3.299 | 0.002 |
| Friday_2019    | 1.353                       | 0.345                     | 0.248 | 3.916 | 0.000 |

Table 5: Coefficients for weekends

| Model           | Unstandardized coefficients | Standardized coefficients | $t$ | Sig. |
|-----------------|-----------------------------|---------------------------|-----|------|
|                | $B$                         | Std. error                | Beta|      |
| 1 (Constant)   | 1077192.668                 | 105946.505                | 10.167 | 0.000 |
| Saturday_2019  | 3.613                       | 0.473                     | 0.784 | 7.643 | 0.000 |
| Sunday_2019    | 0.932                       | 0.588                     | 0.163 | 1.586 | 0.119 |
MAE: It measures the average significance of the forecast errors, where all individual errors have equal weights:

\[ \text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |e_i|. \]  

(4)

MSE: It also measures the significance of the forecast errors, and larger errors get penalized more due to squaring:

\[ \text{MSE} = \frac{1}{n} \sum_{i=1}^{n} e_i^2. \]  

(5)

BIAS: This is an indication of whether the forecast is overestimating or underestimating the actual supply over the forecast horizon:

\[ \text{BIAS} = \sum_{i=1}^{n} e_i. \]  

(6)

MAPE: It measures the relative significance of forecasting errors in percentage terms:

\[ \text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{e_i}{X_i} \right| \times 100. \]  

(7)

RMSE: It measures how much error there is between two data sets:

\[ \text{RMSE} = \sqrt{\text{MSE}}. \]  

(8)

4 Proposed system

In this section, passenger demand forecast for 2020 has been made for Yenikapi M1–Kirazli M1 line by using 2019 data. Demand forecasting is made by regression analysis technique, simple average method, ANNs, and ML algorithms. Demand forecasting in this study consists of two stages. First, line-based demand forecasting is

![Weekdays Demand Forecast](image1)

Figure 2: Comparison of regression analysis results and test values for weekdays.

![Weekdays Demand](image2)

![Weekends Demand Forecast](image3)

![Weekends Demand](image4)

Figure 3: Comparison of regression analysis results and test values for weekends.
made using the statistical techniques such as regression analysis technique and simple average method. In the second stage of the study, station-based daily forecasting is made for all stations on Yenikapi M1–Kirazlı M1 line using ANN and ML algorithms. The steps of the study are shown in Figure 1.

5 Results and discussions

In this section we perform a thorough presentation of results, with clear discussion on the model structure.

5.1 Regression analysis technique

In this section, passenger demand forecasts for 2020 are made for Yenikapi M1–Kirazlı M1 line by using 2019 data. Demand forecasting is made by using the least squares method in SPSS Statistics 21.0 programme. With this method, weekday and weekend passenger demand forecasts were made on line basis. First, the test data are divided into two groups as weekdays and weekends. Then, the data are transferred to SPSS Statistics 21.0 programme. According to the results of the regression analysis, the model is successful approximately 95% according to the weekday result of 2019 (as shown in Table 2) and it is 86% successful according to the weekend result (as shown in Table 3). The values in the coefficients table give the regression coefficients and their significance levels to be used in the regression equation (Tables 4 and 5).

When the coefficient values in the table are transformed into a regression model, the equation (9) is obtained: The work done for Monday was applied for 7 days in a week.

$$b_1 = \text{Monday},$$

$$y_i = a + b_1 X,$$

$$y_i = -202980.131 + 1.678X. \tag{9}$$

When the X value is substituted in the equation (9), the forecasting values are obtained.

When the test values are tested with this equation, MAPE (average absolute percentage error) values of 0.01 for the weekdays and 0.04 for the weekends are obtained. Figures 2 and 3 show the actual value and regression output. While Figure 2 shows the comparison of the results of regression analysis and test values for weekdays, Figure 3 shows the comparison of the results of regression analysis and test values for weekends.
5.2 Simple average method

The simple average method is calculated by taking the average of all passengers in the previous period. The previous period data are divided into groups as weekdays and weekends, then simple average method is applied as determined in equations (2) and (3). Previous period data and forecasting data obtained by simple average method are shown in Table 6.

With this method, when test values are tested, MAPE (average absolute percentage error) values of 0.001 for weekdays and 0.002 for weekends are obtained.

Figures 4 and 5 show the actual value and simple average method output for Weekdays and Weekends, respectively.

The MAPE is calculated for the forecasts found by regression analysis and simple average method as determined in equation (7). MAPE values for both methods are given in Table 7. As it can be seen from the forecasting results obtained by regression analysis and simple

![Figure 4](image_url)

**Figure 4:** Values obtained as a result of simple average method with test values for weekdays.

![Figure 5](image_url)

**Figure 5:** Values obtained as a result of simple average method with test values for weekends.

| Method              | Regression analysis technique | Simple average method |
|---------------------|-------------------------------|-----------------------|
| Weekdays            | 0.01                          | 0.001                 |
| Weekends            | 0.04                          | 0.002                 |

**Table 7:** MAPE values for regression analysis technique and simple average method

![Figure 6](image_url)

**Figure 6:** The artificial neural network architecture with the most successful results.
average method, the simple average method has been more successful in estimating passenger demand on line basis.

5.3 ANN

In this stage of the study, test data are used for station based forecasting. Training and test data for ANNs method are transferred to Matlab software environment. In order to determine the hidden layer numbers in the ANN structure, the number of neurons in these layers, the activation function of the layers, $\alpha$ learning rate, and momentum coefficients, various combinations have been tried by using the trial and error method and the parameters with the lowest error have been obtained. The most successful experiment is the ANN with an input layer of 5 neurons, a hidden layer of 20 neurons, and an output layer of 1 neuron. Levenberg-Marguardt was used as an educational function. The ANN architecture with the lowest error is given in Figure 6.

The most successful combination compared to the actual data is shown in Figure 7. It is seen that the ANN generally gives very close results to the station-based test data of the total number of passengers. When comparison is made on a daily basis, it is seen that the ANN gives very close results to the daily test data of the total number of passengers. Comparison of daily test data and demand forecast with ANN of Yenikapi station is given in Figure 8. The work done for Yenikapi station is applied for other 12 stations in Yenikapi M1–Kirazli M1

---

**Figure 7:** Comparison of actual data and demand forecast with ANNs on station basis.

**Figure 8:** Comparison of daily test data and demand forecast with ANN of Yenikapi station.
line. Station basis comparison is made via Matlab programme. The actual data for 2019 and ANN outputs are given in Table 8.

### 5.4 ML algorithms

In this section, passenger demand forecasts for 2020 are made for Yenikapı M1–Kirazlı M1 line by using 2019 data, and the data are used for station based forecasting.

Decision tree, linear regression, polynomial regression, SVM, and RF are applied via Python programme in order to forecast the demand. While applying the mentioned ML algorithms “pandas” library of the Python programme is used [49,50]. First, we transferred the actual data format into csv. After transferring, the csv file is read for each station separately. Machine is trained for decision tree algorithm and forecasting is made. The same process is also applied for linear regression algorithm, and the forecasting data are obtained. For RF algorithm, the machine is trained. Random_state makes the output of the model non-reproducible, so when the value random_state is specified, the same parameters will produce the same results if the same training data are given. n_estimators indicates the number of decision trees to be created. We ensure that the machine we train using the RF algorithm makes a prediction according to the data we provide. While carrying out the process for polynomial regression, we transform the values in the training column with polynomial features, before training the machine. The degree parameter here is the degree of the polynomial, the more the degree is increased, the healthier the result. The forecasting is made after training the machine for polynomial regression algorithm. Finally, the actual data are

| Stations          | Actual values | Forecasting values of ANNs |
|-------------------|---------------|----------------------------|
| Yenikapı M1       | 21,580,918    | 22,359,554                 |
| Aksaray           | 10,512,904    | 15,114,789                 |
| Eminiyet          | 6,786,245     | 8,131,640                  |
| Ulubatlı          | 4,976,793     | 4,717,769                  |
| Bayrampaşa        | 4,068,294     | 4,793,001                  |
| Sağmalcılar M1    | 5,940,868     | 5,993,238                  |
| Kocatepe          | 11,540,092    | 13,806,921                 |
| Otogar            | 6,335,930     | 9,274,460                  |
| Esenler           | 4,968,780     | 4,015,545                  |
| Menderes          | 5,295,528     | 5,722,780                  |
| Üçüzü打好        | 4,733,464     | 4,481,730                  |
| Bağcılar Meydan   | 5,718,710     | 5,319,202                  |
| Kirazlı M1        | 14,012,546    | 14,844,106                 |

| Stations           | Decision tree | Linear regression | Polynomial regression | SVR |
|--------------------|---------------|-------------------|-----------------------|-----|
| Yenikapı M1        | 21,604,501    | 20,450,369        | 21,029,558            | 22,033,163 |
| Aksaray            | 10,559,756    | 10,336,055        | 10,386,937            | 10,554,932 |
| Eminiyet           | 6,789,393     | 6,077,873         | 6,077,873             | 6,789,393 |
| Ulubatlı           | 4,975,577     | 4,660,477         | 4,660,477             | 4,975,577 |
| Bayrampaşa         | 4,082,877     | 3,884,522         | 3,884,522             | 4,082,877 |
| Sağmalcılar M1     | 5,942,894     | 5,942,894         | 5,942,894             | 5,942,894 |
| Kocatepe           | 11,565,433    | 11,866,587        | 11,866,587            | 11,565,433 |
| Otogar             | 6,274,932     | 6,546,065         | 6,546,065             | 6,274,932 |
| Esenler            | 6,372,133     | 6,398,997         | 6,398,997             | 6,372,133 |
| Menderes           | 3,903,469     | 4,148,825         | 4,148,825             | 3,903,469 |
| Üçüzü打好          | 4,103,849     | 4,186,663         | 4,186,663             | 4,103,849 |
| Bağcılar Meydan   | 6,096,633     | 5,989,601         | 5,989,601             | 6,096,633 |
| Kirazlı M1         | 11,382,133    | 12,058,786        | 12,058,786            | 11,382,133 |

| Stations           | Decision tree | Linear regression | Polynomial regression | SVR |
|--------------------|---------------|-------------------|-----------------------|-----|
| Yenikapı M1        | 22,009,946    | 20,450,369        | 21,029,558            | 22,033,163 |
| Aksaray            | 10,559,756    | 10,336,055        | 10,386,937            | 10,554,932 |
| Eminiyet           | 6,789,393     | 6,077,873         | 6,077,873             | 6,789,393 |
| Ulubatlı           | 4,975,577     | 4,660,477         | 4,660,477             | 4,975,577 |
| Bayrampaşa         | 4,082,877     | 3,884,522         | 3,884,522             | 4,082,877 |
| Sağmalcılar M1     | 5,942,894     | 5,942,894         | 5,942,894             | 5,942,894 |
| Kocatepe           | 11,565,433    | 11,866,587        | 11,866,587            | 11,565,433 |
| Otogar             | 6,274,932     | 6,546,065         | 6,546,065             | 6,274,932 |
| Esenler            | 6,372,133     | 6,398,997         | 6,398,997             | 6,372,133 |
| Menderes           | 3,903,469     | 4,148,825         | 4,148,825             | 3,903,469 |
| Üçüzü打好          | 4,103,849     | 4,186,663         | 4,186,663             | 4,103,849 |
| Bağcılar Meydan   | 6,096,633     | 5,989,601         | 5,989,601             | 6,096,633 |
| Kirazlı M1         | 11,382,133    | 12,058,786        | 12,058,786            | 11,382,133 |
scaled and then forecasting is made according to the scaled data for SVR after machine is trained.

Forecasting data obtained by using “pandas” library of the Python programme are given in Table 9. Figure 9 displays the comparison of the forecast results of ML algorithms for the stations in the line. The comparison of the total number of passengers based on stations (actual data, ANN, and decision tree) is given in Figure 10.

The actual data for 2019 and the forecast values for ANNs and ML algorithms can be seen in Table 10. As it can be seen in Figure 11, the decision tree, which is a ML algorithm, made the best passenger demand forecasting.

After obtaining forecast values, evaluation with BIAS, MAE, MSE, MAPE, and RMSE are obtained. The errors are calculated separately for each algorithm. MAE, MSE, BIAS, MAPE and RMSE are calculated as determined in equations (4)–(8), respectively. Error measures obtained under the ANN and ML algorithms are given in Table 11.

6 Conclusion

In big cities like Istanbul, transportation is a huge problem. The transportation problem can be solved by public transportation. Rail transportation is almost a remedy for this problem. The success of strategic and detailed planning of public transportation highly depends on accurate
Table 10: The actual value and station based forecast values

| Stations       | Actual data | ANNs output | Decision tree output | Linear regression output | Random forest output | Polynomial regression output | SVR output |
|----------------|-------------|-------------|----------------------|--------------------------|----------------------|-------------------------------|------------|
| Yenikapı M1    | 21,580,918  | 22,359,554  | 21,604,501           | 20,450,369               | 22,009,946           | 21,029,558                    | 22,033,163 |
| Aksaray         | 10,512,904  | 15,114,789  | 10,336,055           | 10,554,032               | 10,372,547           | 10,422,918                    |            |
| Emniyet         | 6,786,245   | 8,131,640   | 6,789,353            | 7,080,301                | 6,232,495            | 6,958,298                     |            |
| Ulubatlı        | 4,976,793   | 4,717,769   | 4,975,577            | 5,103,849                | 4,759,887            | 5,084,261                     |            |
| Bayrampaşa      | 4,068,294   | 4,793,001   | 4,082,877            | 4,148,025                | 3,989,601            | 4,121,065                     |            |
| Sağmalcılar M1 | 5,940,868   | 5,993,238   | 5,942,904            | 6,056,633                | 5,749,993            | 6,003,844                     |            |
| Kocatepe        | 11,540,092  | 13,806,921  | 11,565,433           | 11,382,133               | 12,058,786           | 11,569,292                    |            |
| Otoğar          | 6,335,930   | 9,274,460   | 6,214,932            | 6,198,917                | 6,775,188            | 6,108,937                     |            |
| Esenler         | 4,968,780   | 4,015,545   | 4,976,127            | 4,992,412                | 4,998,213            | 5,015,730                     |            |
| Menderes        | 5,295,528   | 5,722,780   | 5,300,183            | 5,329,567                | 5,269,843            | 5,350,201                     |            |
| Üçüzüllü       | 4,733,464   | 4,481,730   | 4,736,340            | 4,789,098                | 4,681,280            | 4,787,869                     |            |
| Bağcılar        | 5,718,710   | 5,319,202   | 5,766,335            | 5,739,671                | 5,805,516            | 5,776,229                     |            |
| Meydan          |              |             |                      |                          |                      |                               |            |
| Kirazlı M1      | 14,012,546  | 14,844,106  | 14,030,227           | 13,324,303               | 14,302,624           | 13,614,183                    | 14,134,258 |

Figure 11: Comparison of the actual values and station based forecast values.

Table 11: Error measures obtained under the ANN and ML algorithms

| Error        | ANNs         | Decision tree | Linear regression | RF            | Polynomial regression | SVR           |
|--------------|--------------|---------------|-------------------|---------------|-----------------------|---------------|
| BIAS         | 778,636      | 23,583        | 1,130,549         | 429,028       | 551,360               | 452,245       |
| MAE          | 2,133        | 65            | 3,097             | 1,175         | 1,511                 | 1,239         |
| MSE          | 1,661,024,714| 1,523,720     | 3,501,756,278     | 504,287,739   | 832,870,821           | 560,343,945   |
| MAPE         | 0.0099       | 0.0003        | 0.0144            | 0.0054        | 0.0070                | 0.0057        |
| RMSE         | 46.1871      | 8.0381        | 55.6542           | 34.2844       | 38.8661               | 35.1998       |
demand information data. Passenger demand forecasting is very important in railway transportation systems in terms of accurate headway scheduling. If the passenger demand is estimated correctly, the frequency of headway is optimized, increasing passenger satisfaction and reducing operating costs.

In this study, although passenger demand estimation is made with eight different methods, the study consists of two stages. In the first stage of the study, demand forecasting was made for the Yenikapı–Kirazlı line. During this stage, simple average method and regression analysis are made for demand forecasting, also the regression is performed with SPSS Statistics 21.0 programme. As a result, the MAPE values of these two techniques were compared. It has been observed that the simple average method gives more accurate results.

In the second phase of the study, station-based demand forecasts were made for all stations on the Yenikapı–Kirazlı line. While making this forecasting, ANNs and decision tree, linear regression, RF, polynomial regression, and SVM methods, which are among ML algorithms, were used. BIAS, MAE, MSE, MAPE, and RMSE error rates were compared, and for this dataset, the lowest error rates were found as (0.03% MAPE) decision tree, (0.54%) RF, and (0.57%) SVR. As it can be seen from the results, the most accurate forecasting is obtained by ML algorithms.

This study is important as it will provide input for the headway scheduling studies in the future works of railway transportation systems.

And also, unlike the studies in the literature, this study includes two stages for demand forecasting. In the first stage of the study, line-based demand forecast is made for weekdays and weekends, while in the second stage of the study, the estimation is made on a station basis. Making the study both line based and station based contributes to the literature. So that unlike other studies in the literature, in this study, demand forecasting was made separately in a single study, taking into account the line density and station density. Thus, considering the line density, return stations can be determined according to the density of the stops, so the train may not have to complete the entire line. Cost profit can be achieved by not going to the last station every time, depending on the line density.

In addition, during the current pandemic period (COVID-19), passenger demand forecasting for public transportation has become more important, especially in metropolitan cities such as Istanbul, because the number of passenger demanding rail transportation is known in advance and social distance is ensured, it is important for the prevention of infectious diseases and the prevention of epidemics.

**Funding information:** There are no funding sources for this study.

**Author contributions:** Both authors have read and agreed to the published version of the manuscript.

**Conflict of interest:** There are no competing interests.

**Ethical approval:** The conducted research is not related to either human or animal use.

**Data availability statement:** All data generated or analyzed during this study are included in this published article.

**References**

[1] Ozcan C, Ceylan H, Haldenbilen S, Yaşar AB. Kentiçi Otobüs Taşımacılığında Talep Tahmini Ve Fiyat Analizleri: Denizli Örneği. Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Derg. 2010;12(1):47–61. Retrieved from https://dergipark.org.tr/tr/pub/deumffmd/issue/40834/492705.

[2] Gencer MA. Ankara metro M1 (Kızılay-Batikent) line scheduling of operating hours. (Master’s Thesis). Kırıkkale: Kırıkkale University Institute of Natural and Applied Sciences, Department of Industrial Engineering; 2016.

[3] Özcan T. Headway optimization in urban public transport systems. (Master’s Thesis). Denizli, Turkey: Pamukkale University Institute of Natural and Applied Sciences, Department of Civil Engineering; 2018.

[4] Uludağ N. Modeling bus lines with fuzzy optimization and linear target programming approaches. (PhD Thesis). Denizli, Turkey; Pamukkale University Institute of Natural and Applied Sciences, Department of Civil Engineering; 2010.

[5] Hamurcu M, Eren T. Decision making for rail system projects with AHP-GP and ANP-GP. Gazi J Eng Sci. 2017;3:1–13.

[6] Bayraktar D, Çelebi D, Ayaçoğlu D5O, Danış S, Dümbe F. Creating the most appropriate expedition schedules and machinist schedules in IMM rail transport systems. Project final report.

[7] Chow TW, Leung CT. Neural Network based short-term load forecasting using weather compensation. IEEE Trans Power Syst. 1996 Nov;11(4):1736–42.

[8] Çelebi D, Bolat B, Bayraktar D. Light rail passenger demand forecasting by artificial neural networks. 2009 International Conference on Computers & Industrial Engineering, Troyes, France: IEEE; 2009. p. 239–43.

[9] Efendigil T, Emirler Ö. The importance of demand estimation in the aviation sector: a model to estimate airline passenger demand. Tarım ve Gıda Değer Zirkülerinde Yönetilebilir Araştırmaları ve Endüstri Mühendisliği Özel Sayısı. 2017;12:14.
Jin F, Li Y, Sun S, Li H. Forecasting air passenger demand with a
Cyril A, Mulangi RH, George V. Modelling and forecasting bus
Kim S, Shin DH. Forecasting short
Zhao S, Mi X. A novel hybrid model for short
Dou F, Xu J, Wang L, Jia L. A train dispatching model based on
Shih H, Rajendran S. Comparison of time series methods and
Fuloria S. Passenger demand forecasting in the ridesharing
Picano B, Fantacci R, Han Z. Nonlinear dynamic chaos theory

Melek Nar and Seher Arslankaya

[10] Shih H, Rajendran S. Comparison of time series methods and
machine learning algorithms for forecasting Taiwan blood
services foundation’s blood supply. J Healthc Eng. 2019;Sep;2019:6123745.

[11] Kilimci ZH, Akyuz AO, Uysal M, Akyokus S, Uysal MO, Atak
Bulbul B, et al. An improved demand forecasting model using
deep learning approach and proposed decision integration
strategy for supply chain. Complexity. 2019;2019:2019.

[12] Aamer A, Eka Yani L, Alan, Priyatna I. Data analytics in the
supply chain management: review of machine learning applica-
tions in demand forecasting. Oper Supply Chain Manage: An
Int J. 2020;14(1):1–13.

[13] Cyril A, Mulangi RH, George V. Modelling and forecasting bus
passenger demand using time series method. 2018 7th
International Conference on Reliability, Infocom Technologies
and Optimization (Trends and Future Directions) (ICRITO).
Noida, India: IEEE; 2018. p. 460–6

[14] Zhao S, Mi X. A novel hybrid model for short-term high-speed railway
passenger demand forecasting. IEEE Access. 2019;7:175681–92.

[15] Li ZC, Sheng D. Forecasting passenger travel demand for air
and high-speed rail international service: A case study of
Beijing-Guangzhou corridor, China. Transp Res Part A Policy
Pract. 2016;94:397–410.

[16] Dou F, Xu J, Wang L, Jia L. A train dispatching model based on
fuzzy passenger demand forecasting during holidays. J Ind Eng
Manag. 2013;6(1):320–5.

[17] Jin F, Li Y, Sun S, Li H. Forecasting air passenger demand with a
new hybrid ensemble approach. J Air Transp Manage. 2020;83:101744.

[18] Gong W. ARMA-GRNN for passenger demand forecasting. 2010
Sixth International Conference on Natural Computation. Vol. 3.
Yantai, China: IEEE; 2010. p. 1577–81.

[19] Kim S, Shin DH. Forecasting short-term air passenger demand
using big data from search engine queries. Autom Constr.
2016;70:98–108.

[20] Ke J, Zheng H, Yang H, Chen XM. Short-term forecasting of
passenger demand under on-demand ride services: a spatio-
temporal deep learning approach. Transp Res, Part C Emerg
Technol. 2017;85:591–608.

[21] Li X, Zhang Y, Du M, Yang J. The forecasting of passenger
demand under hybrid ridesharing service modes: a combined
model based on WT-FCBF-LSTM. Sustain Cities Soc.
2020;62:102419.

[22] Fuloria S. Passenger demand forecasting in the ridesharing
context: a comparison of statistical and deep learning
approaches. IUP J Appl Econ. 2020;19:1.

[23] Bai L, Yao L, Kanhere S, Wang X, Sheng Q. Stg2seq: spatial-
temporal graph to sequence model for multi-step passenger
demand forecasting. ArXiv Prepr arXiv: 1905. 2019:10069.
doi: 10.24963/jcal.2019/274.

[24] Picano B, Chiti F, Fantacci R, Han Z. Passengers demand fore-
casting based on chaos theory. ICC 2019-2019 IEEE International
Conference on Communications (ICC); 2019. p. 1–6.

[25] Picano B, Fantacci R, Han Z. Nonlinear dynamic chaos theory
framework for passenger demand forecasting in smart city.
IEEE Trans Vehicular Technol. 2019;68(9):8533–45.

[26] Alekseev KP, Seixas JM. Forecasting the air transport demand
for passengers with neural modelling. VII Brazilian Symposium
on Neural Networks, 2002. SBNN 2002. Proceedings; 2002.
p. 86–91.

[27] Üreten S. Production operations management. 2nd edn.
Ankara: Nobel Publications; 1999.

[28] Demirbaş FP. Application of Demand forecasting methods in
Combi boiler production. (Master’s Thesis). Kocaeli, Turkey:
Kocaeli University Institute of Natural and Applied Sciences,
Department of Industrial Engineering; 2011.

[29] Olgun S. Demand forecasting methods in supply chain man-
agement and application of an artificial intelligence based
demand forecasting model. (Master’s Thesis).
Istanbul,Turkey: Istanbul University Institute of Natural and
Applied Sciences, Department of Industrial Engineering; 2009.

[30] Özsoy E. Creating a customer focused production plan based
on demand forecasting and an application. (Master’s Thesis).
Dokuz Eylül University Institute of Social Sciences, Production
Management and Industrial Management Program; 2006.

[31] Aksoy ZS. Demand forecasting methods and applications in
tenprise resource planning software. (Master’s Thesis).
Istanbul, Turkey: Istanbul University Institute of Natural and
Applied Sciences, Department of Industrial Engineering; 2008.

[32] Çağlar T. Methods used in demand forecasting and an enter-
prise application producing fence wire. (Master’s Thesis).
Kırıkkale, Turkey: Kırıkkale University Institute of Natural and
Applied Sciences, Department of Industrial Engineering; 2007.

[33] Meydan YA. Demand forecasting methods and its application
in a medium sized business. (Master’s Thesis). Istanbul,
Turkey: Istanbul Ticaret University Institute of Natural and
Applied Sciences; 2007.

[34] Arslankaya S. Estimation of hanging and removal times in
exolal with artificial neural networks. Emerg Mater Res.
2020;9(2):366–74. doi: 10.1680/jemmr.19.00191.

[35] Arslankaya S. Estimating the effects of heat treatment on
aluminum alloy with artificial neural networks. Emerg Mater
Res. 2020;9(2):540–9. doi: 10.1680/jemmr.20.00059.

[36] Ambrosio IK, Brentan BM, Herrera M, Luvisotto E, Ribeiro L,
Izquierdo I. Committee machines for hourly water demand
forecasting in water supply systems. Math Prob Eng.
2019;2019:2019.

[37] Navada A, Ansari AN, Patil S, Sonokamble BA. Overview of use
of decision tree algorithms in machine learning. 2011 IEEE
control and system graduate research colloquium. Shah Alam,
Malaysia: IEEE; 2011. p. 37–42.

[38] Breiman L, Cutler A. Random forests-classification description.
Department of Statistics Homepage; 2007. http://www.stat.
berkeley.edu/~breiman/RandomForests/cc_home.htm.

[39] Akinleye AA, Adewumi AO. Classification of phishing email
using random forest machine learning technique. J Appl Math.
2014;2014:2014.

[40] Kavitha S, Varuna S, Ramya R. A comparative analysis on
linear regression and support vector regression. 2016 Online
International Conference on Green Engineering and
Technologies (IC-GET). Coimbatore, India: IEEE; 2016. p. 1–5.
doi: 10.1109/GET.2016.7916627.
[41] Seber GAF, Lee AJ. Linear regression analysis. Wiley Series in Probability and Statistics; 2012.

[42] Montgomery DC, Peck EA, Vining GG. Introduction to linear regression analysis. Wiley Series in Probability and Statistics; 2015.

[43] https://www.javatpoint.com/machine-learning-polynomial-regression.

[44] Smola AJ, Schölkopf B. A tutorial on support vector regression. Stat Comput. 2004;14(3):199–222.

[45] Gunn SR. Support vector machines for classification and regression. ISIS Tech Rep. 1988;14(1):5–16.

[46] Nahmias S. Production and operations analysis. 6th edn. CA, USA: McGraw-Hill, Irwin; 2008.

[47] Ravindran A, Warsing DP. Supply chain engineering: models and applications. Boca Raton, FL, USA: CRC Press; 2013.

[48] Chopra S, Meindl P. Supply chain management: strategy, planning, and operation. 6th edn. Upper Saddle River, NJ, USA: Pearson-Prentice Hall; 2015.

[49] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.

[50] https://pandas.pydata.org/.