MUSTAFA ÖZUYSAL, Ph.D.
E-mail: mustafa.ozuysal@deu.edu.tr
Dokuz Eylül University
Department of Civil Engineering
Izmir 35160, Turkey
GÖKMEN TAYFUR, Ph.D.
E-mail: gokmentayfur@iyte.edu.tr
Izmir Institute of Technology
Izmir 35430, Turkey
SERHAN TANYEL, Ph.D.
E-mail: serhan.tanyel@deu.edu.tr
Dokuz Eylül University
Department of Civil Engineering
Izmir 35160, Turkey

Science in Traffic and Transport
Original Scientific Paper
Accepted: Apr. 30, 2011
Approved: Dec. 20, 2011

PASSENGER FLOWS ESTIMATION OF LIGHT RAIL TRANSIT (LRT) SYSTEM IN IZMIR, TURKEY USING MULTIPLE REGRESSION AND ANN METHODS

ABSTRACT

Passenger flow estimation of transit systems is essential for new decisions about additional facilities and feeder lines. For increasing the efficiency of an existing transit line, stations which are insufficient for trip production and attraction should be examined first. Such investigation supports decisions for feeder line projects which may seem necessary or futile according to the findings. In this study, passenger flow of a light rail transit (LRT) system in Izmir, Turkey is estimated by using multiple regression and feed-forward back-propagation type of artificial neural networks (ANN). The number of alighting passengers at each station is estimated as a function of boarding passengers from other stations. It is found that ANN approach produced significantly better estimations specifically for the low passenger attractive stations. In addition, ANN is found to be more capable for the determination of trip-attractive parts of LRT lines.

KEYWORDS

light rail transit, multiple regression, artificial neural networks, public transportation

1. INTRODUCTION

Passenger flow modelling of Light Rail Transit (LRT) Systems is a rarely studied area of public transportation. On the other hand, such modelling is essential especially for the developing countries like Turkey where new rail transit projects are still under construction and are constructed gradually necessitating a long period for completion. The completed and opened-to-service part of these projects may guide to the next steps for determining the location of new stations and mode integration with existing public transit systems. It is also important to ensure that any infrastructure investment will have beneficial effects on the overall transport system and those affected by it [1]. Hence, in the decision-making process, the modelling approach will be very beneficial.

Izmir is the third biggest metropolitan city of Turkey with over 3 million of population. The city has a new transportation master plan proposing many supplementary transit lines. Therefore, the locations of new transit stations and transfer points have to be examined by using the statistics of existing systems. Izmir LRT is one of the newly constructed modern rail transit systems located at the south-east of the city centre with an approximately linear track between the west and the east. The current LRT system is a small range transit application having 11.6km of total line length, 10 stations and a feasibility capacity of 11,000 passengers per hour per direction. Although the daily demand of Izmir LRT is about 100,000 passengers, it is expected to show a considerable increase when the other supplementary transit lines are opened to service in a few years. General map of the current system in service is given in *Figure 1*. The dashed lines in the figure show the intersecting rail transit project which is being constructed for the North-South connection for the public transportation of Izmir. The LRT line also has extension projects connected by Bornova, Üçyol, and Halkapinar stations.

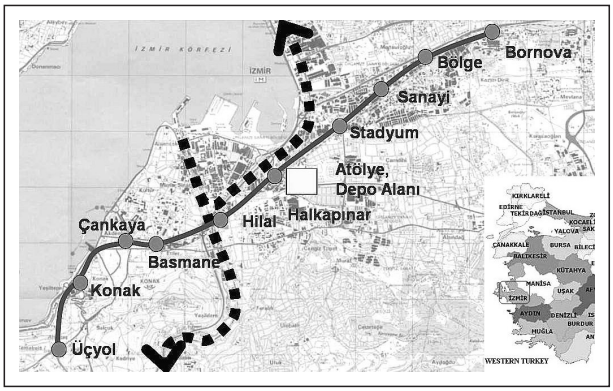

Figure 1 - İzmir LRT system in service

The operational plan of Izmir LRT system is based on some statistical data of passenger flows. The statistical data are obtained from the prepaid ticket machines at the stations including time, usage and station locations. By using these data, passenger numbers getting into every track according to time and the day of week are obtained for a minimum time period of 5 minutes. The assigned time intervals between tracks are controlled whether total boarding passengers of each station exceed the physical capacity of the tracks or not. However, there is an important detail which cannot be neglected in this operation planning. There is no information available on how many of the passengers come to the stations and to which direction on the line they go. This is because all of the passengers who enter use the same prepaid ticket machines regardless of trip directions. The number of passengers getting off at any station is also neglected in this application. Therefore, the used statistics may not reflect the real situation and therefore the operational plans are captive of the trial and error approximation. This is another necessity for the passenger flow modelling of Izmir LRT system which can make the operational plan more reasonable by predicting the amount of alighting passengers at each station.

There are some studies in literature about passenger flows in the public transportation facilities. However, rather than the flow prediction, these studies generally focused on passenger flow management of busy stations or station stop time and departure time optimization [2, 3, 4]. Lee et al. studied the modelling of the flow weight distribution and found a power-law behaviour for Seoul subway system [5]. There are some other studies involving the application of ANNs for predicting daily trend of total public transit flows that provide practical benefits for the operational planning and decision support [6, 7].

ANN is one of the recently explored technologies, which show promise in the area of transportation engineering. Neural networks have the ability to learn from their environment and to adapt to it in an interactive manner similar to their biological counterparts. This is an exciting prospect because of the vast possibilities that exist for performing certain functions with ANN [8]. Therefore, the use of ANNs in passenger flow prediction may reduce the dependency on probabilistic approaches used for the flow weight distributions and increase the significance of the past flow statistics on planning practice.

In this study, the Izmir LRT trip flow predictions by the regression and ANN models are explored. The estimation performance of trip-productive stations which has great importance in decision-making for feeder line projects has been also investigated.

2. DATA

In the study the daily total numbers of boarding and alighting passengers have been used for the models of each station. The total number of alighting passengers at each station is estimated by using the total number of boarding passengers of other stations. The data belong to nine consecutive months from October 2007 to June 2008 and a total of 240 days are included in which 20 items of record (10 boarding and 10 alighting sum) are arranged. The data set includes also weekend days in which travel demand changes dramatically. On the other hand, the data of some specific days like national holidays in the given interval have been eliminated for preventing the inclusion of extreme observations which may decrease the estimation performance. Thus, a reasonable heterogeneity of data is obtained which can make the distinction clear between the estimation capabilities of regression and ANN approaches.

Although peak and off-peak hour distinction for trip flow estimation is necessary for a more rational analysis, it is not possible in practice because the alighting passenger numbers are recorded from mechanical counters at the end of the day for each exit of all stations. Since the exit gates cannot be monitored, a detailed record can be easily obtained from entry gates through the electronic ticket machines. This can lead to reduced prediction capability since boarding and alighting activity observation necessitate short time lags between the two activities. The dynamic passenger flow analysis cannot be realized by this kind of data. On the other hand, this study which aims at predicting the source stations of general outputs of each station for a general trip flow analysis can also show the prediction capability of the developed models with limited conditions.

The histograms and closeness to normal distribution of the data set can be seen in *Figure 2*. The abbreviation “B” means boarding data and “A” corresponds to alighting data. As it can be seen, the data of Basmane, Halkapinar and Stadyum stations are the farthest from normal distribution and the values seem

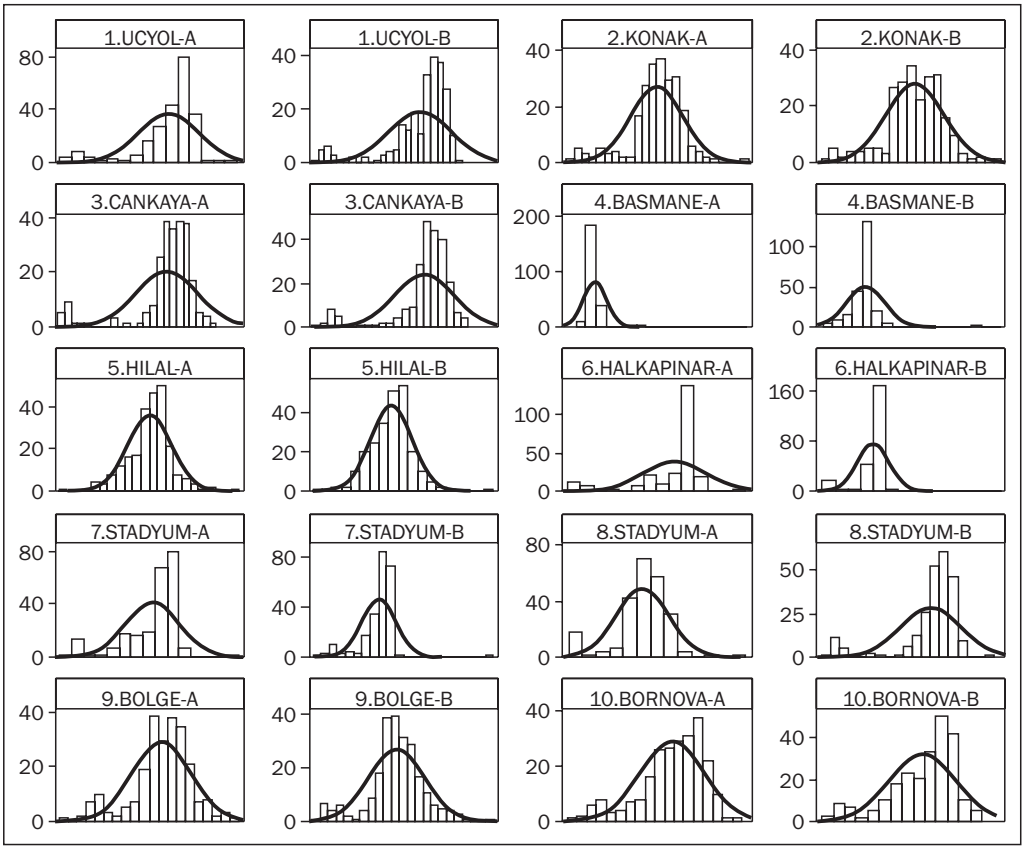

Figure 2 - Histograms of İzmir LRT passenger data

grouped together around a constant value. This may be unfavourable for prediction phase of trip flows of these stations. The data sets for other stations seem somewhat skewed to the left specifically for Ucyol, Konak and Cankaya stations. This skewness results probably from the data of weekend days in which the trip flow is considerably lower compared to working week days.

3. REGRESSION MODELS

For a dynamic passenger flow model, one can easily think that the passengers boarding from a specific station split to groups going to other stations and there is a simple linear relationship between the total boarding number and the divided alighting number. Therefore, the regression analysis may seem the only sound method for demonstrating this linear relationship. However, for the common case in which the dynamic boarding and alighting data are not available, the splitting percentages of each boarding station may not be obtained easily.

The regression analysis utilized for the passenger flow is based on the linear estimate of alighting passenger numbers of a specific station depending on the number of boarding passengers of other stations:

$$Y_{i,a} = \beta_0 + \beta_1 \cdot X_{1,b} + \beta_2 \cdot X_{2,b} + \ldots + \beta_{i-1} \cdot X_{i-1,b} + \\ + \beta_{i+1} \cdot X_{i+1,b} + \ldots + \beta_n \cdot X_{n,b} + \varepsilon_{i,a} \quad (1)$$

where "n" is the number of stations, "$Y_{i,a}$" the number of alighting passengers at the station, "$X_{i,b}$" the number of boarding passengers at other stations, "β_0" the constant term, "β_i" the coefficients of explanatory variables and "$\varepsilon_{i,a}$" the regression residual.

The ordinary least square method is applied by varying data type, the number of explanatory boarding stations and the consideration of constant term. Eight different regression types are obtained. Since the ranges of the passenger numbers using each station are considerably different, data standardization can increase the efficiency of the regression. Therefore, beside the regressions with raw data, the standardized data are also applied for the purpose of comparison. Eq.2 is utilized for the standardization which is also being used for ANN applications.

$$x_{s,i} = (1.8(x_{n,i} - x_{n,\min(i)})/(x_{n,\max(i)} - x_{n,\min(i)})) - 0.9 \quad (2)$$

where $x_{\max(i)}$ and $x_{\min(i)}$ are the maximum and minimum values of "i"th station. "s" and "n" indices indicate the standardized and natural cases respectively. The used standardization method compresses the data into [-0.9, 0.9] range which intends to prevent upper and lower limit saturation problem in ANN analysis. The same standardization method with ANN approach is preferred to get homogeneity in performance comparison.

The regression models are also utilized for both cases of eliminated and non-eliminated boarding sta-

tions for explanatory variables. It is expected that the elimination causes a decrease in the estimation performance. However, it can present the effective stations if the decrease is negligible. The boarding stations are eliminated from the explanatory groups by using "stepwise" approach in which the variables having F-test values smaller than 0.05 are included and the ones having the values over 0.10 are eliminated. The regressions are also diversified by the inclusion and exclusion of the constant term which may be significant in the case of ineffective boarding stations.

The estimation performances of the regression models are compared by using Root Mean Square Errors (RMSE) (Eq.3) and Efficiency Factor (EF) (Eq.4). RMSE is a frequently used measure of the differences between the predicted values and the actual (observed) values and serves to aggregate the residuals into a single measure of predictive power.

$$\mathrm{RMSE} = \left(\left(\sum_{i=1}^{N} (Y_i^{obs} - Y_i^{pre})^2 \right) / N \right)^{1/2} \tag{3}$$

$$\mathrm{EF} = 1 - \left(\sum_{i=1}^{N} (Y_i^{obs} - Y_i^{pre})^2 / \sum_{i=1}^{N} (Y_i^{obs} - \overline{Y}_i)^2 \right) \tag{4}$$

where "Y_i^{obs}" and "Y_i^{pre}" are the observed and predicted values of "i^{th}" alighting passenger observation and "$\overline{Y}_i$" is the mean of alighting passengers for each model. EF accounts for model errors in estimating the mean of the observed data set which ranges from minus infinity to 1.0. "EF = 1" corresponds to a perfect match of modelled alighting passenger numbers to the observed data. "EF = 0" indicates that the model predictions are as accurate as the mean of the observed data and an efficiency less than zero ($-\infty < \mathrm{EF} < 0$) shows worse prediction than the mean.

For a more accurate comparison, RMSE is given as the ratio of the observed mean for obtaining an impartial comparison (*Table 1*). Coefficients of determination values (R) of the regressions are not provided because they may be elusory for the regressions without constant term.

When the table is investigated in a station base, the six stations (Ucyol, Konak, Cankaya, Halkapinar, Stadyum and Bornova) seem to indicate successful estimations with EF values close to "1". The performances of different regression types are also similar for these stations. However, this cannot be said for the regressions of other four stations (Basmane, Hilal, Sanayi and Bolge). The identical property of these stations is their considerably low number of passengers for both boarding and alighting flows. Besides, Basmane station is close to the biggest social and commercial fair area of Izmir and this causes high fluctuation in trip demand depending on the time and size of the activity at the fair.

In general, there is no remarkable difference between the given performance statistics of the regression types. However, it can be said that the exclusion of the constant term decreases the estimation capability especially for the mentioned four stations. The elimination of the boarding stations also has a minor decreasing effect on the performance. It means that the flow effective stations can be easily distinguished. The statistics of the post regressions between observed and predicted values of the regression models are presented in *Table 2*.

As it was known, the squared R and the slope of the post regression (β_1) should be close to "1" and the constant term (β_0) should be close to "0" for a sound model. For these criteria the regression models of Halkapinar station give the highest reliable estimations. Halkapinar is located at the middle of the LRT line and it is the centre of inter-modal public transit. Therefore, this considerable success at Halkapinar station is very important to make inferences about the efficiency of transfer points. For the models constructed with standardized data, the exclusion of the constant term indicates smaller decrease in the performance compared with the models of the raw data. The four stations un-

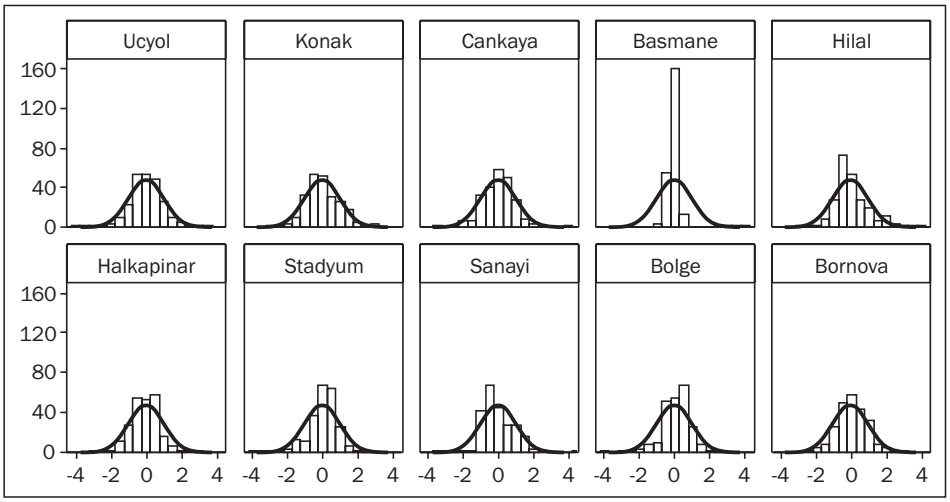

Figure 3 – The histograms of standardized residuals for RE model

Table 1 - Performance of the regression models

Alighting Stations		Statistics	Raw Data				Standardized Data			
			All boarding stations		Eliminated boarding stations		All boarding stations		Eliminated boarding stations	
			With constant term	Without constant term	With constant term	Without constant term	With constant term	Without constant term	With constant term	Without constant term
			RAC	RA	REC	RE	SAC	SA	SEC	SE
1	Ucyol	RMS/mean:	0.039	0.043	0.039	0.044	0.039	0.040	0.039	0.040
		EF:	0.909	0.902	0.908	0.903	0.909	0.907	0.908	0.905
2	Konak	RMS/mean:	0.048	0.049	0.048	0.049	0.048	0.049	0.048	0.049
		EF:	0.833	0.838	0.832	0.833	0.833	0.827	0.832	0.827
3	Cankaya	RMS/mean:	0.047	0.049	0.048	0.049	0.047	0.048	0.048	0.048
		EF:	0.920	0.908	0.919	0.907	0.920	0.919	0.919	0.919
4	Basmane	RMS/mean:	0.457	0.457	0.468	0.468	0.457	0.604	0.468	0.605
		EF:	-7.045	-7.490	-12.679	-10.313	-7.045	-1.247	-12.679	-1.200
5	Hilal	RMS/mean:	0.089	0.138	0.089	0.139	0.089	0.089	0.089	0.089
		EF:	-0.644	-0.567	-0.672	-0.547	-0.644	-0.645	-0.672	-0.675
6	Halkapinar	RMS/mean:	0.061	0.061	0.062	0.062	0.061	0.064	0.062	0.064
		EF:	0.920	0.919	0.917	0.916	0.920	0.912	0.917	0.911
7	Stadyum	RMS/mean:	0.079	0.079	0.079	0.079	0.079	0.083	0.079	0.083
		EF:	0.819	0.821	0.817	0.824	0.819	0.792	0.817	0.791
8	Sanayi	RMS/mean:	0.118	0.119	0.120	0.125	0.118	0.119	0.120	0.119
		EF:	0.563	0.521	0.540	0.347	0.563	0.561	0.540	0.555
9	Bolge	RMS/mean:	0.076	0.077	0.077	0.080	0.076	0.076	0.077	0.077
		EF:	0.651	0.671	0.634	0.640	0.651	0.652	0.634	0.635
10	Bornova	RMS/mean:	0.069	0.072	0.070	0.073	0.069	0.070	0.070	0.070
		EF:	0.784	0.708	0.776	0.699	0.784	0.773	0.776	0.768

R: raw data, S: standardized data, A: including all stations, E: eliminated stations, C: including constant term

der question exhibit more sensitivity for the elimination of boarding stations. When the models with elimination are compared, RE model can be stated as the most successful in general. In order to recognize a regression estimator as a model, the residuals should provide some important criteria like fitness of normal distribution. The standardized residual histograms of RE model compared with normal distributions are given in *Figure 3*. It is clear that RE model fairly satisfies the criterion for most of the stations, specifically Ucyol, Konak, Cankaya, Halkapinar and Bornova. Some stations like Hilal, Stadyum, Sanayi and Bolge have a bit of a bias with normal distribution. However, the estimation residuals of Basmane station indicate a distribution considerably far from the normal distribution.

Figure 4 represents the flow scheme of Izmir LRT obtained by the stepwise elimination of the stations of RE model. Ucyol, Konak and Cankaya stations seem as the most effective stations for trip attraction and production. The stations at the middle section of the LRT line, like Basmane, Hilal, Halkapinar and Stadyum indicate lower dependence on other stations. Konak, Cankaya and Bornova stations show attractiveness for long trips rather than the short ones.

Consequently, it can be said that the regression models give high estimation capability for the sections of LRT line where the trip demand demonstrates stable and relatively higher trend compared to its average. However, as presented above, the regression models perform poorly for the stations where there are fluctuations in trip demand and therefore a more reliable modelling approach may be required.

4. ANN MODELS

Neuro-computing is concerned with processing information which first involves a learning process within an artificial neural network architecture that adaptively responds to inputs according to a learning rule. After the neural network has learned what it needs to know, the trained network can be used to perform certain tasks depending on the particular applications [8].

ANN can have one or more layers consisting of many neural cells which are connected by the con-

Table 2 – Post-regression statistics of the regression models

Alighting Stations		Statistics	Raw Data				Standardized Data			
			All boarding stations		Eliminated boarding stations		All boarding stations		Eliminated boarding stations	
			With constant term	Without constant term	With constant term	Without constant term	With constant term	Without constant term	With constant term	Without constant term
			RAC	RA	REC	RE	SAC	SA	SEC	SE
1	Ucyol	R^2	0.917	0.903	0.916	0.903	0.917	0.914	0.916	0.912
		β_0	1,556.8	540.9	1,572.4	365.5	1,556.8	1,499.4	1,572.4	1,535.2
		β_1	0.917	0.970	0.916	0.979	0.917	0.919	0.916	0.917
2	Konak	R^2	0.857	0.855	0.856	0.853	0.857	0.853	0.856	0.853
		β_0	1,967.5	1,663.4	1,977.2	1,804.7	1,967.5	2,029.9	1,977.2	2,030.6
		β_1	0.857	0.878	0.856	0.868	0.857	0.853	0.856	0.853
3	Cankaya	R^2	0.926	0.921	0.925	0.920	0.926	0.925	0.925	0.925
		β_0	1,002.9	1,444.7	1,012.2	1,425.2	1,002.9	957.6	1,012.2	942.7
		β_1	0.926	0.894	0.925	0.895	0.926	0.928	0.925	0.929
4	Basmane	R^2	0.111	0.110	0.068	0.068	0.111	0.007	0.068	0.008
		β_0	4,300.4	4,318.4	4,505.6	4,464.2	4,300.4	4,757.4	4,505.6	4,740.5
		β_1	0.111	0.107	0.068	0.075	0.111	0.069	0.068	0.074
5	Hilal	R^2	0.378	0.055	0.374	0.054	0.378	0.378	0.374	0.374
		β_0	734.1	899.9	738.8	897.3	734.1	734.4	738.8	739.6
		β_1	0.378	0.228	0.374	0.230	0.378	0.378	0.374	0.374
6	Halkapi-nar	R^2	0.926	0.926	0.923	0.923	0.926	0.920	0.923	0.919
		β_0	202.2	212.9	209.2	219.0	202.2	243.1	209.2	247.1
		β_1	0.926	0.922	0.923	0.920	0.926	0.914	0.923	0.913
7	Stadyum	R^2	0.847	0.846	0.846	0.844	0.847	0.829	0.846	0.828
		β_0	866.8	827.4	873.8	754.2	866.8	1,008.6	873.8	1,014.7
		β_1	0.847	0.853	0.846	0.866	0.847	0.825	0.846	0.825
8	Sanayi	R^2	0.696	0.691	0.685	0.670	0.696	0.693	0.685	0.691
		β_0	678.0	741.2	702.3	920.7	678.0	680.3	702.3	687.1
		β_1	0.696	0.669	0.685	0.590	0.696	0.696	0.685	0.693
9	Bolge	R^2	0.742	0.737	0.732	0.714	0.742	0.741	0.732	0.732
		β_0	1,414.9	1,252.5	1,466.6	1,327.2	1,414.9	1,409.4	1,466.6	1,457.4
		β_1	0.742	0.770	0.732	0.757	0.742	0.742	0.732	0.733
10	Bornova	R^2	0.822	0.807	0.817	0.806	0.822	0.817	0.817	0.816
		β_0	3,607.3	5,298.5	3,718.1	5,443.5	3,607.3	3,850.3	3,718.1	4,015.5
		β_1	0.822	0.741	0.817	0.734	0.822	0.812	0.817	0.805

R: raw data, S: standardized data, A: including all stations, E: eliminated stations, C: including constant term

nection links having a certain direction determined by the network architecture. Each connection link has an associated weight that represents its connection strength and each neuron typically applies a nonlinear transformation, called an activation function, to its net input to determine its output signal. The network is trained by using an expected output in a manner that the weights of connection links are updated according to the selected learning method in a typical iteration step called epoch [9].

The neural networks as global approximation tool have been widely used due to the ability to process and map external data and information based on past experience to generate successful forecasts. One of the developing application areas of ANN is transportation engineering. Murat and Ceylan investigated the applicability of ANN models in forecasting of transport energy demand and found consistent results [10]. Zhang et al. used ANN for the reconstruction of vehicle crash accidents and they claim that the pre-impact velocity of

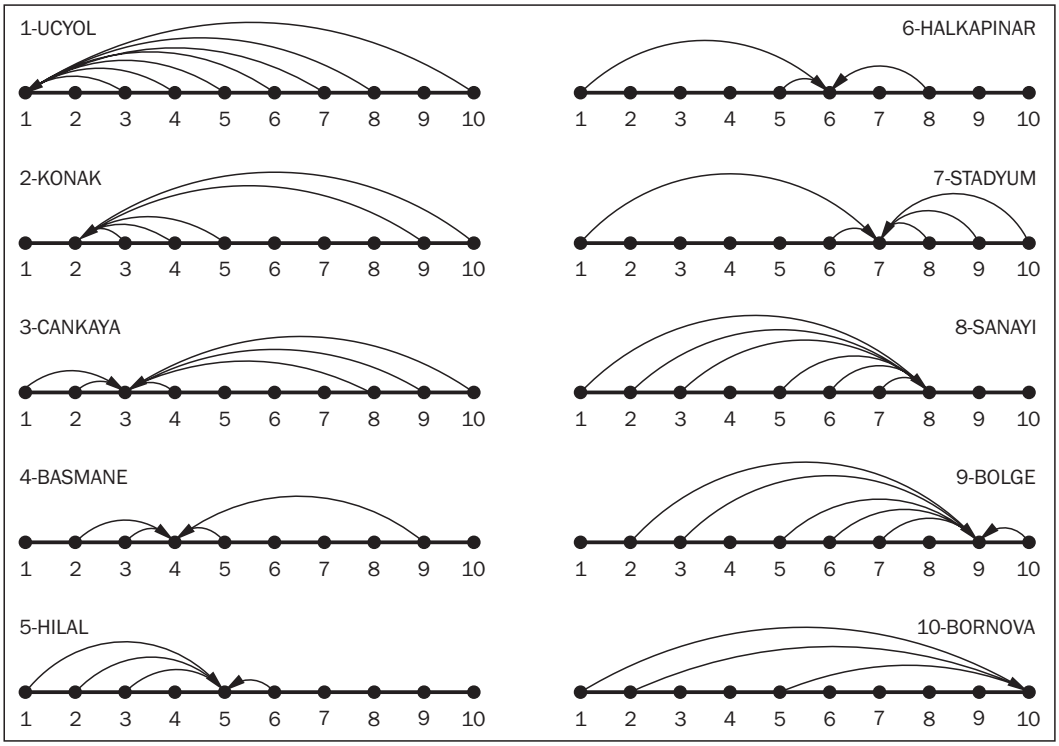

Figure 4 - Trip flow scheme for RE model

vehicles without tyre marks could be predicted by ANN model [11]. Murat used ANN to estimate vehicle delay for non-uniform and over-saturated conditions [12].

In this study "Feed Forward" perceptron with "Back Propagation" training algorithm (FFBP) type of ANN, which is the most widely used type and a remarkable alternative for the regression approach, is chosen for the application. In Feed Forward ANN, the nodes are arranged in layers and they are connected to those in the next layer; however, not to those in the same layer. The information flows only in the forward direction, from the input layer to the hidden and output layers.

The Back Propagation is a supervised training algorithm in which an input-output training set is used and it consists of mainly two activities: forward pass and backward pass. In forward pass, the training pairs from the input data sets are selected and fed into the input neurons and the activity is propagated from input layer to hidden and then output layers. In backward pass, the propagation occurs in a reverse direction and the errors are computed for each output unit. Layer by layer, the error for each hidden unit is computed by the propagating errors. The weights are updated by the generalized delta rule which is based on the steepest gradient descent with the direction vector being set to negative of the gradient vector. Consequently, the solution often follows a zigzag path while trying to reach a minimum error position. Therefore, it is sometimes possible to get trapped by a local minimum. "Gradient descent with momentum" technique is a successive way to avoid this problem in which the weights of the next epoch are determined by including the effect of the weight difference between the past two epochs [13]:

$$\Delta\omega_{ij(n)} = -\eta(\partial E/\partial\omega_{ij}) + \alpha \cdot (\Delta\omega_{ij(n-1)}) \quad (5)$$

where, "$\Delta\omega_{ij(n)}$" are the present iteration differences of the weights, "$\Delta\omega_{ij(n-1)}$" the past iteration differences of the weights, "η" the learning rate, "E" the error function depending on the weights and "α" the momentum factor.

It is known that the extrapolation capability of ANN is relatively weak if compared with interpolation [14]. Therefore, an attempt is made to distribute the minimum and maximum values comparable for the training and testing parts of the data set. For this purpose, a rank number is attained for each day of 20 data columns in such a manner that the maximum value of the column has the biggest rank. Then, the numbers of ranks are summed up for each row (day of record) and data is sorted according to the summation. The sorted data are distributed to training and testing set one by one for each row and consequently 120 training and 120 testing pairs are obtained.

The data is standardized by using Eq.1 which is compatible with the chosen tangent hyperbolic activation function. This method compresses the data set to the range of -0.9 and 0.9 instead of -1 and 1 which prevents the upper and lower limit saturation. The saturation problem may cause insufficient learning because the activation functions give the values cumulated around "0" and "1" especially for the data having repeated patterns at minimum and maximum limits [15].

The independent variables which are boarding passengers of nine stations are standardized by using the maximum and minimum values of the whole data set. However, for back transposition of the dependent variable which is the number of alighting passengers from the model station, the maximum and minimum values of training data set are used. In this way, the output of test data is treated as unobserved.

In this study, two-hidden-layered network architecture is employed. The number of neurons in the first hidden layer was obtained by the trail and error procedure while 5 neurons were fixed in the second hidden layer. Consecutively 5, 10, 15, 20, 25 and 30 neurons are tried for the first hidden layer of ANN model. Thus, six different trainings and tests are applied for each station. The performance measures of post regression, RMSE and EF are calculated for both of the simulation results of test and whole data. The numbers of neurons that give the best performance for each station are obtained as shown in *Table 3* for testing data which are more critical than the training results. As it can be seen from the table, the testing data results give "15" as the optimum number of neurons. The resulting network structure is shown in *Figure 5*.

For the first stage, all of the boarding stations are included in the input layer of the network (ANN-AS). The network was successfully trained with 2,500 epochs, 0.05 learning rate and 0.9 momentum factor. The results of the first stage analysis obtained by using test data are given in *Table 4*. Beside the mentioned performance statistics, the percentages of discrepancy ratio (DR) are also presented in the Table. DR values of the estimations are calculated by Eq. 6 for each observed and predicted pair:

$$DR = \log_{10}(Y_i^{pre}/Y_i^{obs}) \quad (6)$$

Generally, it is accepted as good estimation if DR value is between -0.1 and 0.1 corresponding to 25% deviation from the observations. In the table, the percentages of the estimation having DR below -0.10 is indicated as "low estimation ratio" (LER), and the percentages over 0.10 DR as "high estimation ratio" (HER). The estimations between the DR of -0.10 and 0.10 are indicated as "proper estimation ratio" (PER).

The PER percentages of ANN models are satisfying in general. A tendency for low prediction is dominant

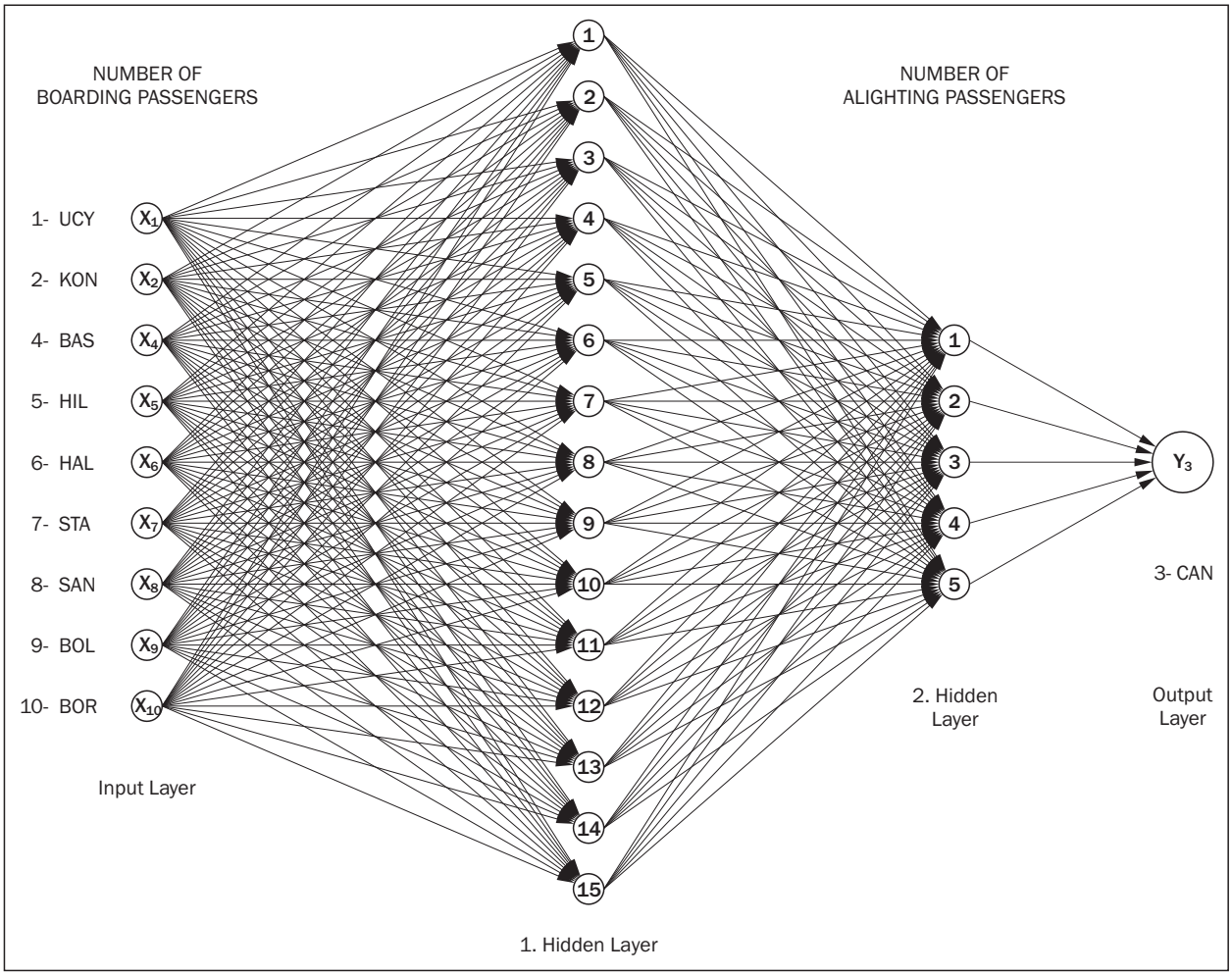

Figure 5 - The network architecture (example for Cankaya station)

Table 3 - The number of neurons giving the best performance for each station

Station	Convergence performance	β_1	β_0	R	RMSE	EF	General mean	General mode
1- Ucyol	15	15	15	5	5	5	10.6	15
2- Konak	15	10	10	20	20	20	17.5	20
3- Cankaya	10	20	20	30	30	30	23.8	30
4- Basmane	20	25	30	25	10	10	18.1	10
5- Hilal	20	20	20	25	15	15	21.9	20
6- Halkapinar	20	15	15	5	5	5	13.1	15
7 - Stadyum	15	15	15	15	15	15	15.0	15
8- Sanayi	20	15	15	20	20	20	19.4	20
9- Bolge	15	15	15	20	20	20	16.3	20
10- Bornova	15	20	5	15	15	15	13.8	15
Mean:	16.5	17.0	16.0	18.0	15.5	15.5	17.3	15
Mode:	15	15	15	20	20	20	16.7	15
Mean/Mode Diff.:	9.1%	11.8%	6.3%	11.1%	29.0%	29.0%	3.8%	0.0%

Table 4 - The performance of network simulation by using the testing data

Stations		Post Regression Statistics			General Performance		Discrepancy Ratio		
		R	β_1	β_0	RMSE	EF	LER	PER	HER
1	Ucyol	0.834	1.037	-388.989	1,766.5	0.515	1.25	98.75	0.00
2	Konak	0.898	0.963	544.860	824.8	0.776	0.42	99.58	0.00
3	Cankaya	0.965	0.974	342.394	619.5	0.930	0.42	99.17	0.42
4	Basmane	0.281	0.472	2,799.810	3,979.8	-1.888	9.58	86.67	3.75
5	Hilal	0.772	1.035	-48.659	113.8	0.271	0.83	97.50	1.67
6	Halkapinar	0.937	0.950	134.196	219.3	0.873	0.83	97.92	1.25
7	Stadyum	0.759	0.938	380.602	917.7	0.346	2.50	95.42	2.08
8	Sanayi	0.626	0.922	118.122	554.0	-0.341	5.00	86.67	8.33
9	Bolge	0.757	0.914	490.698	649.4	0.371	2.50	96.25	1.25
10	Bornova	0.905	0.906	2,069.241	1,448.4	0.808	1.25	98.33	0.42

LER: low estimation ratio, PER: proper estimation ratio, HER: high estimation ratio

for the most of the stations. Minimum PER percentages are obtained for Basmane and Sanayi which are the stations having the lowest and most inconsistent passenger activity.

When the slopes of the post regressions (β_1) of test data simulations are compared, it can be said that Konak station gives the best result for trip flow prediction. Ucyol, Halkapinar, Sanayi and Cankaya stations follow consecutively. The efficiency factors (EF) and coefficients of determination (R) also indicate good prediction performance for the Ucyol and Halkapinar stations. However, Bornova station takes over instead of the Sanayi and Cankaya for EF and R values. Thus, the ANN model including whole boarding stations gives the highest performance for the critical three points of the LRT line (the edges and the main transfer points) and reasonable estimation capability for other stations.

In the second stage of ANN analysis, the capability of the estimation for flow effective stations is tried by eliminating some stations from nine boarding stations (ANN-ES). One by one, a station is eliminated from nine input stations and the corresponding performance is evaluated. This procedure is applied for all the stations; however, for the sake of brevity, we present only the results for Ucyol station in *Table 5*. As seen in *Table 5*, for example, the constant term of the post regression (β_0) is getting closer to "0" by the single elimination of the boarding data of the 4th, 6th and 10th stations (Basmane, Halkapinar and Bornova). According to these performance improvements indicated by different statistics in the table, the stations which have higher improvements and occur more frequently are selected for the combined elimination. The elimination is gradually continued while observing negligible decrease in the estimation performance. For example, 4-6, 4-6-5, 4-6-5-2-10 combinations are eliminated gradually for the Ucyol station.

Table 5 – Performances of ANN model after single station eliminations for Ucyol.

Eliminated boarding station		Achieved training performance	Post-regression statistics			General performance		Discrepancy ratio		
			R	β_1	β_0	RMSE	EF	LER	PER	HER
none		0.009	0.918	0.985	323.9	1,094.1	0.819	0.41	98.76	0.83
2	Konak	0.007	0.951	0.960	983.0	964.2	0.893	0.00	100.00	0.00
3	Cankaya	0.008	0.938	0.879	2,431.6	1,056.9	0.871	0.74	99.26	0.00
4	Basmane	0.007	0.957	1.004	15.7	897.8	0.907	0.00	99.26	0.74
5	Hilal	0.007	0.964	0.930	1,482.5	816.9	0.923	0.00	100.00	0.00
6	Halkapinar	0.007	0.957	1.006	14.5	904.8	0.906	0.00	99.26	0.74
7	Stadyum	0.007	0.940	0.851	3,038.4	1,085.3	0.864	0.74	99.26	0.00
8	Sanayi	0.007	0.955	0.893	2,252.9	947.5	0.896	0.00	99.26	0.74
9	Bolge	0.006	0.947	0.924	1,558.1	975.7	0.890	0.00	99.26	0.74
10	Bornova	0.009	0.945	1.023	-367.4	1,041.1	0.875	0.00	98.53	1.47
The boarding stations suitable for the elimination according to the increase in performance criterion:			4	4	4	5	5	4	2	2
			5	6	6			5	5	3
			6		10			6		5
										7

LER: low estimation ratio, PER: proper estimation ratio, HER: high estimation ratio

Table 6 - The performance of network simulation by using eliminated boarding stations

Stations		Post-regression statistics			General performance		Discrepancy ratio		
		R	β_1	β_0	RMSE	EF	LER	PER	HER
1	Ucyol	0.834	1.037	-389.0	1,766.5	0.515	1.25	98.75	0.00
2	Konak	0.898	0.963	544.9	824.8	0.776	0.42	99.58	0.00
3	Cankaya	0.965	0.974	342.4	619.5	0.930	0.42	99.17	0.42
4	Basmane	0.281	0.472	2,799.8	3,979.8	-1.888	9.58	86.67	3.75
5	Hilal	0.772	1.035	-48.7	113.8	0.271	0.83	97.50	1.67
6	Halkapinar	0.937	0.950	134.2	219.3	0.873	0.83	97.92	1.25
7	Stadyum	0.759	0.938	380.6	917.7	0.346	2.50	95.42	2.08
8	Sanayi	0.626	0.922	118.1	554.0	-0.341	5.00	86.67	8.33
9	Bolge	0.757	0.914	490.7	649.4	0.371	2.50	96.25	1.25
10	Bornova	0.905	0.906	2,069.2	1,448.4	0.808	1.25	98.33	0.42

LER: low estimation ratio, PER: proper estimation ratio, HER: high estimation ratio

The performance results of the combined elimination are summarized in *Table 6*. The results revealed that the combined elimination produces markedly different results than the single elimination. A reasonable decrease in the prediction performance can be seen for Ucyol, Stadyum and Sanayi stations (see *Table 4* and *Table 6*). On the other hand, Konak, Cankaya and Hilal stations indicate better performance after the elimination. Consequently, the western part of the LRT line, which is closer to the central business district (CBD) exhibits distinguishable performance with ANN model after the selection of trip-effective stations. Consequently, the CBD-based trips can be evaluated as having more predictable flow for LRT lines.

The trip flow scheme for ANN-ES model is shown in *Figure 6*. When it is compared with *Figure 4* given for RE regression model, it can be seen that ANN-ES model with elimination provides more numbers of stations as explanatory variables for Konak, Basmane, Hilal, Sanayi and Bornova stations which have relatively low estimation capability for RE model. For other stations, specifically Ucyol and Cankaya, few numbers of boarding stations can be sufficient to demonstrate the alighting stations for ANN-ES model without considerable decrease in prediction success.

5. COMPARISON OF REGRESSION AND ANN MODELS

Since the different variations of the multiple regression models give similar estimation performance, two

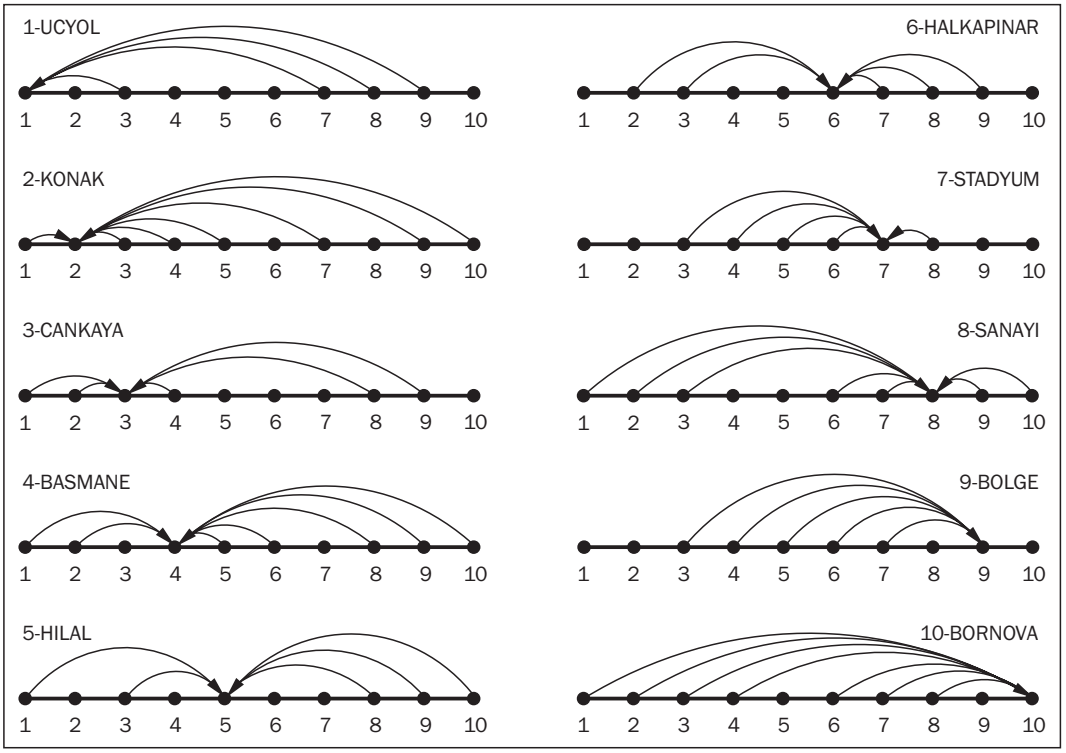

Figure 6 - Trip flow scheme for ANN-ES

Table 7 - Efficiency factors (EF) of regression and ANN models

Stations		Multiple Regression models		Artificial Neural Network models	
		RA	RE	ANN-AS	ANN-ES
1	Ucyol	0.902	0.903	0.756	0.689
2	Konak	0.838	0.833	0.862	0.806
3	Cankaya	0.908	0.907	0.952	0.931
4	Basmane	-7.490	-10.313	0.076	-0.026
5	Hilal	-0.567	-0.547	0.432	0.595
6	Halkapinar	0.919	0.916	0.960	0.876
7	Stadyum	0.821	0.824	0.721	0.573
8	Sanayi	0.521	0.347	0.545	0.399
9	Bolge	0.671	0.640	0.708	0.568
10	Bornova	0.708	0.699	0.826	0.808

of them (RA and RE) are selected for the comparison with ANN models (ANN-AS and ANN-ES). The efficiency factors (EF) of the four mentioned models are given in *Table 7*. As it is expected, the elimination of the boarding stations slightly decreases EF values which should be close to "1" for a proper estimation, for the both of regression and ANN approaches. Except for Ucyol and Stadyum stations, the ANN approach increases the prediction efficiency specifically for Basmane and Hilal stations which have poor estimations for the regression models. Accordingly, it can be said that the ANN approach produces considerably high capability of trip flow prediction for the cases in which the multiple regression is inadequate.

The difference between the two approaches is clearer when the DR percentages are compared. *Figure 7* shows the DR percentages of model predictions between -0.01 and 0.01 range which indicates the ratio of the predictions within the deviation of 2.3%. As can be seen from the figure, a considerably high success is obtained for the ANN models which produce reasonable predictions with 60% of the whole predictions. This is only 30% for the regression models. For the first five stations which have been constructed in the CBD, the ANN-ES model indicates higher performance than the ANN-AS model. Hence, rather than the regression models, the ANN models can allow the selection of trip-attractive stations for the LRT lines in CBD.

The statistics of the predictions having DR percentages out of -0.01~0.01 range is also important for evaluating the estimation performances. The percentages given in *Figure 8* are obtained by the difference in

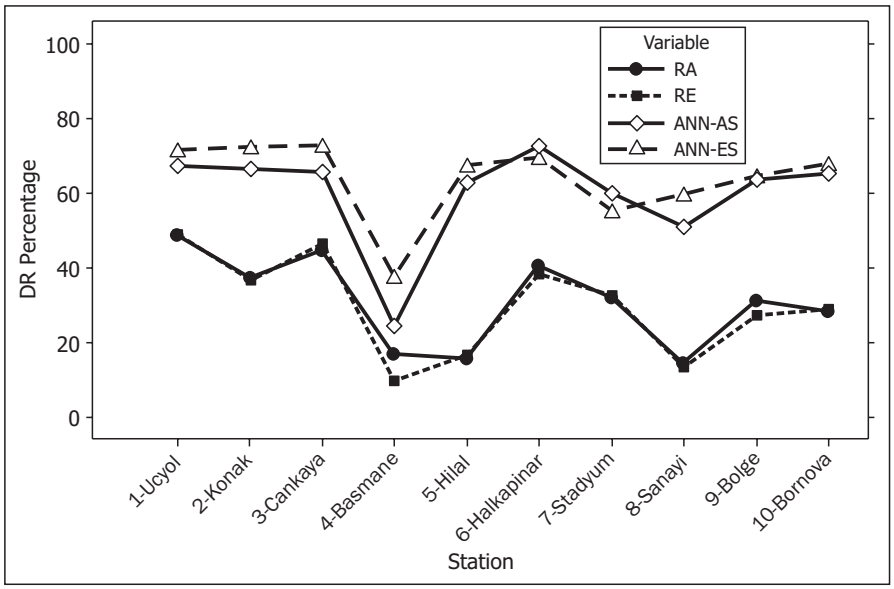

Figure 7 - DR percentage comparison for -0.01~0.01 range

percentages between high (over 0.01 DR) and low (under -0.01 DR) predictions. It is clear that the amount of high predictions is excessive for the regression models, specifically for Basmane and Sanayi. On the other hand, ANN models give a much closer trend around the "0" line. It can be said that ANN models produce more preferable estimations which are symmetrically distributed around the measured values. This proves the ability of determination of trip-attractive stations by ANN approach.

6. CONCLUSION

The most distinguishable difference between the two examined trip flow estimation approaches for Izmir LRT arises from the flow schemes represented in Figures 4 and 6. ANN model yields considerably different results in the selection stage of trip-effective stations. For the stations where the regression models produce poor estimations, ANN models show considerably high performance by the inclusion of some boarding stations. The ANN approach necessitates more explanatory variables (boarding stations), especially for the line section in CBD of Izmir. The station selections of ANN approach can be evaluated as more reliable for Izmir LRT because the discrepancies between the observed and predicted pairs produce findings in favour of this approach.

When the numbers of arrows are compared in the figures (*Figure 4* and *Figure 6*), Basmane, Hilal, Halkapinar and Stadyum stations seem less trip-attractive

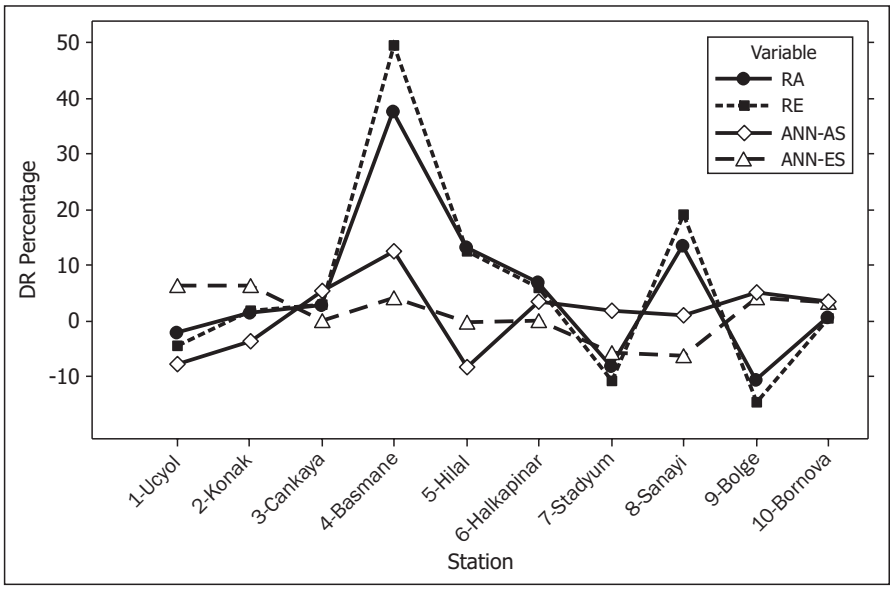

Figure 8 - DR percentage comparison for the differences out of -0.01~0.01 range

while Ucyol, Konak and Bornova stations show higher attraction. Accordingly, Izmir LRT system in its current form is found to be effective only for the trips between the two ends of the line. The inner trips having shorter distance are not reasonably supported by the system. Therefore, some feeder lines around the middle section of the system can provide higher travel demand and increase the efficiency of the LRT system in public transportation of Izmir.

The regression models provide better estimation capability for the sections of LRT line where the trip demand demonstrates stable and relatively higher trend compared to its average. However, the case of fluctuating demand and low trip attractions may cause a dramatic decrease in the estimation capability.

The ANN approach is more capable for the determination of trip-attractive stations because of unbiased DR values after the elimination of the stations. In addition, it is more reliable for the LRT sections constructed in CBD. Hence, it can be concluded that ANN is an effective tool for trip flow estimation.

The multiple regression models can be evaluated as more preferable from the simplicity and manageability points of view. Generally, in cases where the ANN model is ineffective, the regression models have better performance. The opposite of this case is also true according to the results.

In the light of these critiques, it can be concluded that the ANN approach should be considered as a "rescuer" technique, when the used data are unsuitable for the regression analysis. Otherwise, the regression analysis can be more practical and a user-friendlier method for trip flow prediction of LRT lines.

ACKNOWLEDGEMENT

The authors would like to thank the personnel of Izmir Metro Inc., Ilgaz Candemir, Emre Oral and Nurten Caliskan for providing the data of the study. Besides, Mustafa Özuysal appreciates TUBITAK, The Scientific and Technological Research Council of Turkey for doctorate scholarship.

Dr. **MUSTAFA ÖZUYSAL**
E-posta: mustafa.ozuysal@deu.edu.tr
Dokuz Eylül Üniversitesi, Mühendislik Fakültesi
İnşaat Mühendisliği Bölümü
35160 İzmir, Türkiye
Dr. **GÖKMEN TAYFUR**
E-posta: gokmentayfur@iyte.edu.tr
İzmir Yüksek Teknoloji Enstitüsü, Mühendislik Fakültesi
İnşaat Mühendisliği Bölümü
35430 İzmir, Türkiye
Dr. **SERHAN TANYEL**
e-posta: serhan.tanyel@deu.edu.tr
Dokuz Eylül Üniversitesi, Mühendislik Fakültesi
İnşaat Mühendisliği Bölümü
35160 İzmir, Türkiye

ÖZET

ÇOKLU REGRESYON VE YAPAY SİNİR AĞLARI (YSA) YÖNTEMLERİ KULLANILARAK İZMİR-TÜRKİYE'DEKİ HAFİF RAYLI SİSTEME (HRS) AİT YOLCU AKIMLARININ MODELLENMESİ

Toplu ulaşım sistemlerindeki yolcu akımlarının tahmin edilmesi, sistemin işletimi ile ilgili yeni kararlar ve mevcut sistemi destekleyici yeni hatların belirlenmesi açısından oldukça önemlidir. Mevcut bir toplu ulaşım hattına ait verimliliğin arttırılması için yolculuk üretim ve çekiminde düşük paya sahip istasyonların öncelikli olarak ortaya konması gerekmektedir. Bu tür bir inceleme aynı zamanda, sisteme yolcu taşıyacak yeni besleme hatlarının gerekli olup olmadığının belirlenmesini sağlamaktadır. Bu çalışmada, İzmir-Türkiye'deki bir hafif raylı sisteme (HRT) ait yolcu akımları, çoklu regresyon analizi ve ileri beslemeli - geri yayılımlı bir yapay sinir ağı (YSA) modeli kullanılarak tahmin edilmiştir. Her bir istasyonda inen yolcu sayısı, diğer istasyonlardan binen yolcu sayılarının bir fonksiyonu olarak modellenmiştir. YSA modeli ile özellikle düşük yolcu çekimine sahip istasyonlar için daha başarılı sonuçlar elde edilmiştir. Ayrıca YSA'nın yüksek yolcu çeken HRS kesimlerinin belirlenmesinde de daha yüksek başarım gösterdiği bulunmuştur.

ANAHTAR KELIMELER

hafif raylı sistem, çoklu regresyon, yapay sinir ağları, toplu ulaşım

LITERATURE

[1] **Gercek, H.**, **Karpak B.**, **Kilincaslan, T.A.**: *Multiple Criteria Approach for the Evaluation of the Rail Transit Networks in Istanbul*, Transportation, Vol. 31, No. 2, 2004, pp. 203-28

[2] **Li, J.P.**: *Train Station Passenger Flow Study*, Proceedings of the 2000 Winter Simulation Conference, eds.: J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, Orlando, Florida, U.S.A., December 2000, pp. 1173-1176

[3] **Harris, N.G.**, **Anderson, N.J.**: *An International Comparison of Urban Rail Boarding and Alighting Rates*, Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail & Rapid Transit, Vol. 221, No. 4, 2007, pp. 521-526

[4] **Takagi, R.**, **Goodman, C.**, **Roberts, C.**: *Optimization of Train Departure Times at an Interchange Considering Passenger Flows*, Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail & Rapid Transit, Vol. 220, No. 2, 2006, pp. 113-120

[5] **Lee, K.**, **Jung, W.S.**, **Park, J.S.**, **Choi, M.Y.**: *Statistical Analysis of the Metropolitan Seoul Subway System: Network Structure and Passenger Flow*, Physica A, Vol. 387, No. 24, 2008, pp. 6231-6234

[6] **Celikoglu, H.B.**, **Cigizoglu, H.K.**: *Public Transportation Trip Flow Modeling with Generalized Regression Neural Networks*, Advances in Engineering Software, Vol. 38, 2007, pp. 71-79

[7] **Celikoglu, H.B.**, **Cigizoglu, H.K.**: *Modelling Public Transport Trips by Radial Basis Function Neural Networks*,

Mathematical and Computer Modelling, Vol. 45, 2007, pp. 480-489.

[8] **Ham, F.M.**, **Kostanic, I.**: *Principles of Neurocomputing for Science and Engineering*, McGraw Hill, New York, 2001

[9] **Jang, J.R.**, **Sun, C.T.**, **Mizutani, E.**: *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*, Prentice Hall, Upper Saddle River NJ, 1997

[10] **Murat, Y.S.**, **Ceylan, H.**: *Use of Artificial Neural Networks for Transport Energy Demand Modeling*, Energy Policy, Vol. 34, No. 17, 2006, pp. 3165-3172

[11] **Zhang, X.**, **Jin, X.**, **Qi, W.**, **Guo, Y.**: *Vehicle Crash Accident Reconstruction Based on the Analysis 3D Deformation of the Auto-Body*, Advances in Engineering Software, Vol. 39, 2008, pp. 459-465

[12] **Murat, Y.S.**: *Comparison of Fuzzy Logic and Artificial Neural Networks Approaches in Vehicle Delay Modeling*. Transportation Research Part C, Vol. 14, No. 5, 2006, pp. 316-334

[13] **Tayfur, G.**, **Swiatek, D.**, **Wita, A.**, **Singh, V.P.**: *Case Study: Finite Element Method and Artificial Neural Network Models for Flow Through Jeziorsko Earthfill Dam in Poland*, ASCE Journal of Hydraulic Engineering, Vol. 131, No. 6, 2005, pp. 431-440

[14] **Tayfur, G.**, **Moramarco, T.**, **Singh, V.P.**: *Predicting and Forecasting Flow Discharge at Sites Receiving Significant Lateral Inflow*, Hydrological Processes, Vol. 21, No. 14, 2007, pp. 1848-1859

[15] **Tayfur, G.**: *Soft Computing Approaches in Hydrology. In: Hydrology and Hydraulics*, Ed.: V. P. Singh, Water Resources Publications, Colorado, 2008, pp. 113-144