Scientia Iranica A (2022) 29(3), 1095–1106

Sharif University of Technology
Scientia Iranica
Transactions A: Civil Engineering
http://scientiairanica.sharif.edu

Short-term prediction of traffic state: Statistical approach versus machine learning approach

A. Rasaizadi, E. Sherafat, and S.E. Seyedabrishami*

School of Civil and Environmental Engineering, Tarbiat Modares University, Tehran, P.O. Box 14115-111, Iran.

Received 14 March 2021; received in revised form 12 May 2021; accepted 5 July 2021

KEYWORDS
Short-term prediction;
Traffic state;
Multinomial logit;
Support vector machine;
Deep neural network.

Abstract. Short-term traffic prediction helps intelligent transportation systems to manage future travel demands. The objective of this paper is to predict the state of traffic in the case of Karaj to Chaloos, a suburban road in Iran. To this end, two approaches, i.e., statistical and machine learning, are employed. In addition, the performance of the multinomial logit model is evaluated using Support Vector Machine (SVM) and Deep Neural Network (DNN) as two top machine learning techniques. The Principal Component Analysis (PCA) is considered to reduce the dimension of the data and make it possible to use the Multinomial Logit (MNL) model. SVM and DNN can predict the traffic state using both primary and reduced datasets (ALL and PCA). Moreover, MNL can be used to not only compare the accuracy of models but also estimate their explanatory power. SVM employing primary datasets outperforms other models with the accuracy rate of 79%. Next, the prediction accuracy rates for SVM-PCA, MNL, DNN-PCA, and DNN-ALL are equal to 78%, 73%, 68%, and 67%, respectively. SVM-ALL exhibits better performance in predicting light, heavy, and blockage states, while MNL can predict the semi-heavy state more accurately. Use of the PCA dataset increases the accuracy of DNN and decreases SVM accuracy by 1%. Greater precision is achieved for the first three months of testing than that in the second three months.

1. Introduction

As a result of an increase in suburban travel demands, especially during holidays in tourist destinations, traffic congestion on these roads can cause several problems. In addition, traffic congestion affects social and environmental issues. In such cases, alongside traffic supply management, travel demand management is essential [1]. According to previous studies, deployment of Advanced Travelers' Information Systems (ATIS) and Advanced Traffic Management Systems (ATMS) can guarantee a balance between travel demand and supply in near future [2]. One of the effective components of these systems is short-term prediction of traffic parameters [3]. Advanced passenger information systems inform system operators and users about the predicted parameters in near future [4]. System operators will be better prepared to handle the critical situation. In addition, road users can have better plans for their future travels and choose less congested traffic hours or a parallel path with low traffic and not to travel, if unnecessary [5].

One of the most important traffic variables that

*. *Corresponding author. Tel.: +98 21 82884914; Fax: +98 21 82884964*
E-mail addresses: arash_rasaizadi@modares.ac.ir (A. Rasaizadi); elahe.sherafat@gmail.com (E. Sherafat); seyedabrishami@modares.ac.ir (S.E. Seyedabrishami)

doi: 10.24200/sci.2021.57906.5469

can be predicted ranging from an upcoming hour to a few months is the traffic state. It includes light, semi-heavy, heavy, and blockage and exhibits road performance under different conditions. Compared to traffic volume and speed, this qualitative traffic parameter contains more significant information that is easily understandable for passengers who do not know other road specifications such as capacity and free-flow speed [6].

This study examines two different methods, including statistical approach and Machine Learning (ML) approach, to predict the traffic state. Each approach has its advantages and disadvantages [7]. The statistical approach is characterized by a well-established theoretical background; however, the ML approach aims to achieve the highest possible accuracy [8]. The statistical approach needs more prior assumptions. Yet, a majority of them are unable to depict the non-linear relationships; on the contrary, the ML approach is more flexible [9]. Compared to the statistical approach, ML approach can easily address outliers as well as missing and noisy data [10]. The statistical approach can be interpreted by estimating coefficients and elasticities, while the ML approach is labeled as a black-box model [11]. Among well-known statistical models are ARIMA [12] for continuous parameter and Multinomial Logit (MNL) [13] for nominal parameters. On the other hand, Neural Networks (NNs) [14], Support Vector Machine (SVM) [15], decision tree [16], and K-Nearest Neighborhood (KNN) [17] are widely used to predict both continuous and nominal traffic parameters.

These two different approaches can be complementary; therefore, this paper employed both of them to fulfill deficiencies. In this regard, the MNL as a statistical model and Deep Neural Network (DNN) and SVM as two ML models were used to predict the traffic state in Karaj-Chaloos suburban road in Iran. After feature extracting, 92 important features were defined. Since some of them are nominal, they must change to dummy features to be applicable in the statistical model [18]. As a result, the number of features increased to 280. Calibrating the MNL model with four utility functions and 280 variables seems to be very difficult; therefore, Principal Component Analysis (PCA) is used for dimension reduction [19].

The present study made four main contributions. First, this paper defines and predicts the traffic state which, in spite of being more informative than traffic volume and speed, has been insignificantly studied. Second, to the best of the authors' knowledge, it is the first time that both statistical (MNL) and ML (DNN and SVM) models are simultaneously used for predicting the traffic state. Third, two different datasets are considered: the first one with 92 extracted features and the second one with reduced features by PCA. The prediction results are compared. Finally, the suburban traffic data consisting of non-routine trips in Iran as a developing country are used.

2. Previous studies

A number of studies have employed ML models such as SVM [20], decision tree [21], and NNs [22] as well as statistical models such as count regressions [23], MNL [24], Nested-logit [25], and other logit families [26] for mode choice prediction.

Karlaftis and Vlahogianni [27] referred to the differences and similarities of the statistical models versus the NN model in transportation studies. They argued that although the NN model has been included in many of transportation studies, its application still remains vague due to poor explanation and inability to generate a unit answer. Moreover, making a comparison between the NN and statistical models is not fair because unlike linear statistical models, the complexity of the NN model makes it more compatible to analyzing the nonlinear relationship.

Golshani et al. [7] compared the performance of the NN, the most popular ML algorithm, with those of discrete choice models, continuous models, and continuous-discrete models to model travel mode and timing decisions. The obtained results pointed to the better performance of the NN model in predicting the travel mode and timing decisions as well as its simplicity and speed. To overcome the problem of poor explanation, they conducted sensitivity analysis to confirm the significant role of the independent variables in prediction accuracy.

Lee et al. [28] compared four types of NN models with the traditional logit model in terms of mode choice. NN models include Back-Propagation Neural Network (BPNN), Radial Basis Function Network (RBFN), Probabilistic Neural Network (PNN), and Clustered Probabilistic Neural Network (CPNN). The prediction accuracy of the employed models was compared using the cross-validation method. In addition, the importance of independent variables was determined through sensitivity analysis. According to the findings, the PNN model with the prediction accuracy of 80% was superior to the MNL model with the prediction accuracy of 70%.

Cheng et al. [29] evaluated the impact of different parameters on the travel mode choice and predicted it. Stochastic Random Forest (RF) models as a powerful technique for implementing the decision tree and MNL model can predict travel mode. The results pointed to the greater accuracy and lower cost of the RF model than those of the MNL model. The proposed method also estimated the relative importance of the explanatory variables and how they might affect the travel mode.

Wang and Ross [30] compared the MNL model with Extreme Gradient Boosting (XGB) learning technique to predict the travel mode based on the decision tree algorithm. Their obtained results pointed to the higher accuracy and superiority of the ML technique (XGB) over the MNL model.

Hensher and Ton [31] employed nested logit models and NN to predict the travel mode in Melbourne, Sydney, and Pooled Cities (Melbourne-Sydney) and compared the generalizability of the models. The results confirmed the superiority of the nested logit model in terms of matching the overall market share; however, the NN model exhibited better performance in matching the market share of individuals.

Filho and Maia [32] predicted the traffic flow using the PCA to reduce the dimensions of data. They utilized the local linear k-mean model to predict the traffic flow. Model validation through real-time network data confirmed the proper model performance in real conditions.

Jin et al. [33] used PCA and Support Vector Regression (SVR) for simultaneous prediction of traffic flow, travel time, and traffic speed. They compared the performance of the SVR model with those of ARIMA and NN models. SVR prediction reached the highest accuracy. In another study, Jin et al. [34] applied Robust Principal Component Analysis (RPCA) to the Beijing traffic data to detect abnormal traffic flow pattern isolation and loop detector faults.

3. Methodology

3.1. Deep Neural Network (DNN)

NNs are regarded as a useful model for time-series prediction [35]. Typically, an NN consists of an input layer (receives inputs), hidden layer(s) (improves the learning ability), and output layers (represents the results). Nodes (neurons) in different layers are known as the Processing Elements (PEs). Each PE in the hidden layer receives the output of the connected PEs from previous layers, and the output of the current layer can be generated [7] through a transformation function. Traditional (shallow) NN only contains 2–3 hidden layers. Upon increasing the number of hidden layers and PEs, the NN is converted to DNN which can ensure better performance than that in the shallow models in terms of accuracy [36]. Deep-learning methods are representation-learning methods with multiple levels of representation (several hidden layers) [37].

NN training algorithms are diverse. Momentum, Levenberg-Marquardt (LM), and Conjugate Gradient (CG) algorithms are the most known algorithms. The present study employed the CG algorithm to train a DNN. The CG is an iterative algorithm searching for a numerical solution of the objective function. This search is done along with conjugate directions, which typically yield faster convergence than gradient descent directions [38].

Here, w denotes the weights in DNN, d the training direction vector, and g the opposite direction of d. For each iteration ($i = 0, 1, \cdots$), the training direction vector is obtained through Eq. (1) [38]:

$$d^{(i+1)} = g^{(i+1)} + d^{(i)} \cdot \gamma^{(i)}, \tag{1}$$

where γ is the conjugate parameter. Then, weights are improved using Eq. (2) [38]:

$$w^{(i+1)} = w^{(i)} + d^{(i)} \cdot \eta^{(i)}, \tag{2}$$

where η is the training rate calculated by minimization.

The DNN is designed to be trained by the traffic data from the beginning of the period available to time t and then, it predicts the traffic from time $t+1$ to the expected time. This prediction is considered a test for model performance. The scheme of DNN is shown in Figure 1.

It is essential to find the appropriate structure of the model, thus achieving more accurate predictions. The tests on different structures were repeated for validation dataset to obtain the best structure for the DNN model using one input layer, 22 hidden layers, and one output layer. This model can be implemented using R-studio.

3.2. Multinomial Logit (MNL)

As a model based on statistics and probabilities, the MNL model is widely used in modeling the nominal dependent variables [39,40]. The effect of independent variables on the dependent variable is determined through an estimation of the related coefficients, and the statistical significance of the coefficients is examined using the t-test. The MNL model is based on the choice theory. For traffic state prediction, the traffic state that has the greatest utility will have the highest probability of occurrence [41]. The traffic state with the highest occurrence probability is regarded as the model prediction. Utility functions are a linear function of independent variables, coefficients, and the error term. The MNL assumes that the error term is independent and identically distributed (iid) [41]. Based on this assumption, the occurrence probability of the traffic state i at time q is shown in Eq. (3):

Figure 1. Prediction steps by Deep Neural Network (DNN).

$$P_{qi} = \frac{\exp(\beta_i x_{qi})}{\sum_{j=1}^{J} \exp(\beta_j x_{qj})}, \tag{3}$$

where P_{qi} is the occurrence probability of the traffic state i at time q; x_{qi} is the vector of independent variables in the utility of traffic state i in time q; and β_is are the coefficients of the corresponding independent variables in the utility of the traffic state i, which must be estimated by the model. The exponential log-likelihood function is defined by Eq. (4) [26]:

$$\log L = \sum_{q=1}^{Q} \sum_{i=1}^{I} M_{qi} \log(P_{qi}), \tag{4}$$

where M_{qi} is the indicator variable whose value equals one if the traffic state i occurs at time t; otherwise, it equals zero. Also, Q is the total number of hourly observations and I the total number of traffic states.

This function should be maximized to estimate the model coefficients. To this end, R-studio is used. Numerical methods including SANN, Nelder Mead, and BFGS are used for optimization [42].

Finally, the output of the MNL model is the probability of occurrence of each traffic state. The maximum probabilistic method is employed to determine the final prediction of the model. In this method, the traffic state with the highest probability of occurrence is the model prediction. Random probability method and average share are other methods for achieving the final prediction [43–47].

3.3. Support Vector Machine (SVM)

SVM is a well-known ML method based on the statistical learning theory, Vapnik-Chervonenkis dimension theory, and structural risk minimization principle [48]. SVM can effectively deal with classification problems. Support vectors include a set of points in the n-dimension space based on which the boundaries of the clusters are determined. In other words, SVM is a classifier that determines the best separation between each cluster. There are a large number of boundaries that can separate clusters. In case the data are linearly separable, a simple way to find vectors is to calculate the distance between the boundaries and support vectors and select vectors with the largest distance from each class (Figure 2). If the data are distributed non-linearly, it is necessary to use mathematical functions such as Kernel functions and map data into another space, where the data can be separable and the support vectors can be easily determined (Figure 3). Linear, polynomial, sigmoidal, and Radial Basis Function (RBF) are some common kernel functions. This study employed the RBF kernel function as a widely used function in the literature [48]. The formulation of RBF function is given in Eq. (5) [49]:

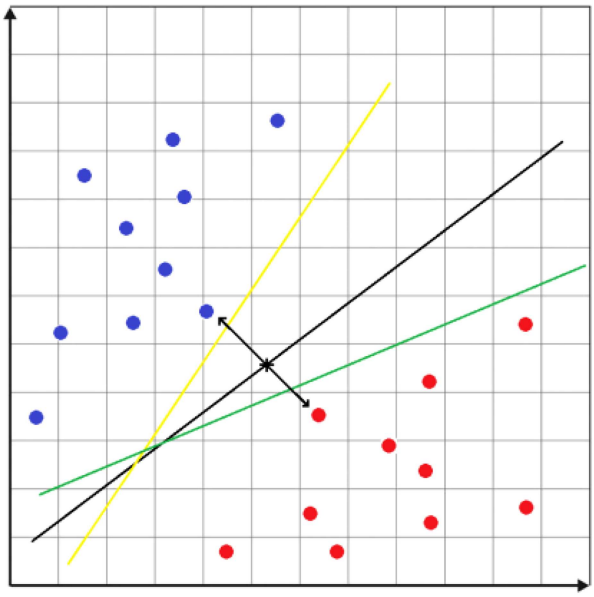

Figure 2. Optimal boundary for separating clusters.

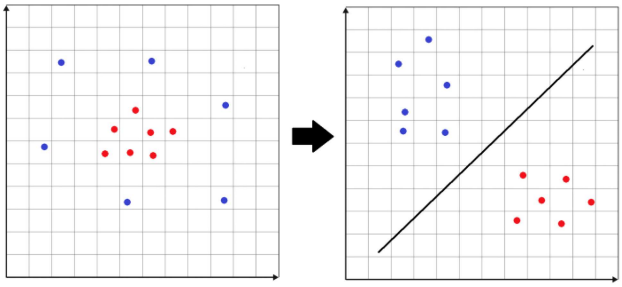

Figure 3. Mapping data into a separable space.

$$K(X_i, X_j) = \exp\left(\frac{\|X_i - X_j\|^2}{2\sigma^2}\right), \tag{5}$$

where σ is a free parameter to be calibrated. $\|X_i - X_j\|^2$ is the squared Euclidean distance between the two feature vectors X_i and X_j. In this study, the SVM model was implemented using R-studio.

3.4. Principal Component Analysis (PCA)

PCA is a multivariate technique for finding a reduced set of non-redundant features that explains substantial variance in the original data [50]. In some cases where the number of features in the data is large, PCA can improve the performance of the statistical model such as MNL [51]. The PCA identifies the most important components with a maximum share of the total variance explanation. Principal Components (PCs) are computed as a linear combination of initial features. The first PC is the coordinate axis with the most substantial variance of the features around it. The next PC is determined by the same criterion and perpendicular to the first PC and then, the other PCs will be determined. In this study, based on the unit

Table 1. Definition of traffic states.

	V/C			
S/Sf	**Under 0.5**	**0.5–0.7**	**0.75–1**	**Over 1**
Over 0.8	Light	Light	Semi-heavy	Semi-heavy
0.5–0.8	Light	Semi-heavy	Semi-heavy	Heavy
0.2–0.5	Heavy	Heavy	Heavy	Heavy
Under 0.2	Blockage	Blockage	Blockage	Blockage

vector method, the PCs are obtained. The unit vector of each PC is called the Eigenvector, and the sum of the least-squares of distances of the records for each PC from the origin is called the Eigenvalue (Eq. (6)) of that PC [52]:

$$\text{The eigenvalue for } PC1 = \sum_{i=1}^{m} d_i^2, \tag{6}$$

where m is the number of records and d_i the distance of the record of each PC from the origin.

4. Dataset

The data set used in this study was collected by a loop detector in Karaj-Chaloos road in Iran. Each record consists of macroscopic traffic parameters including traffic speed, volume, and state collected in one-hour periods. These records were collected from March 2018 to September 2019. The traffic state parameters consist of light, semi-heavy, heavy, and blockage. Traffic state is determined by the ratio of the hourly average speed to the road free-flow speed as well as the ratio of hourly traffic volume to the road capacity. Table 1 shows how a traffic state is defined. This type of traffic state definition is provided by Iran road maintenance and transportation organization (http://www.rmto.ir/en).

In Table 1, V/C is the ratio of the hourly traffic volume to the road capacity, and S/Sf the ratio of the hourly average speed to the road free-flow speed.

Followed by matching the solar and lunar calendars with the traffic parameters, a clear relation between them was observed. These traffic parameters depend on holidays; therefore, it is essential that the calendar-related features be considered to predict traffic parameters. Since such holidays in Iran are based on both solar and lunar calendars, both of them were taken into account. Other features pertaining to weather conditions and blockage of each road direction and parallel paths that directly affect traffic parameters were also considered in predictive models. The effective extracted features are presented in Table 2.

One-year records from March 2018 to March 2019 were used to train models, and the next month records from March 2019 to April 2019 were used as validation dataset to tune parameters of ML models. The next

Table 2. Description of the features used in predictive models.

Feature name	Description	Type
Season	Including spring, summer, fall, and winter	Nominal
Solar month	Including 12 solar months	Nominal
Lunar month	Including 12 lunar months	Nominal
Day of a solar month	Including 29–31 days of a solar month	Nominal
Day of a lunar month	Including 29–30 days of a lunar month	Nominal
Time of day	Including 24 hours a day	Nominal
Day or night	Including day and night	Dummy
Number of holidays	The number of sequential holidays	Continuous
Holidays	Includes 1 for holidays and 0 for other days	Dummy
Holiday type	Type of holidays	Nominal
Days before holidays	Equal to 1 if there is at least one holiday in the next 3 days	Dummy
Type of ahead holidays	Including the type of holiday in the next 3 days	Nominal
Days after holidays	Equal to 1 if there is at least one holiday in the past 3 days	Dummy
Type of previous holidays	Including the type of holiday in the previous 3 days	Nominal
Blockage	Blockage of the road by accidents or by police	Dummy
Blockage of the opposite direction	Blockage of the opposite direction by accidents or by police	Dummy
Blockage of parallel paths	Blockage of parallel paths by accidents or by police	Dummy
Weather	Including sunny, rainy, and snowy	Nominal

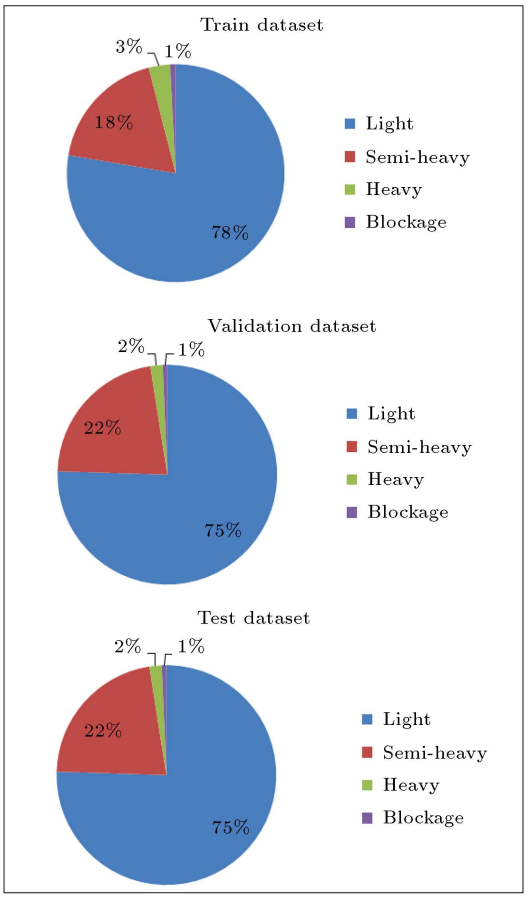

Figure 4. The number of records and share of each traffic state in each dataset.

five-month records from April 2019 to August 2019 were also employed to test the accuracy of predictions. Figure 4 shows the number of records and share of each traffic state in each dataset.

By defining dummy features, 98 primary features were converted to 280 features. To reduce the data dimensions and use them for the MNL model, PCA was used and 30 PCs were defined. These 30 new features could explain 37.7 % of the total variance of 280 features. The first PC (PC1) explains 3.4% of the total variance and this value reaches 0.7% for the 30th PC. Figure 5 shows the distribution of records in terms of the two first PCs, and Figure 6 presents the share of the total variance explanation of each PC. Table 3 shows three features with the highest weight for each PC.

In the next section, the predictive models were trained based on these two datasets: (a) ALL dataset with all 92 primary features and (b) PCA dataset with 30 PCs.

5. Results

To evaluate the prediction accuracy in this study, two criteria of accuracy and F-measure are used. Let

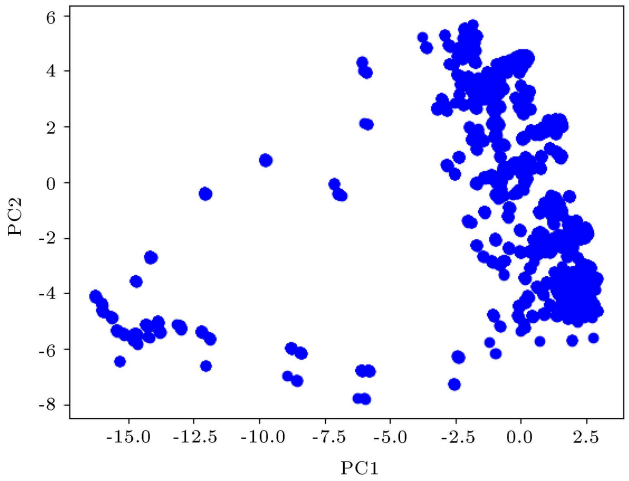

Figure 5. The distribution of records in terms of two first Principal Components (PCs).

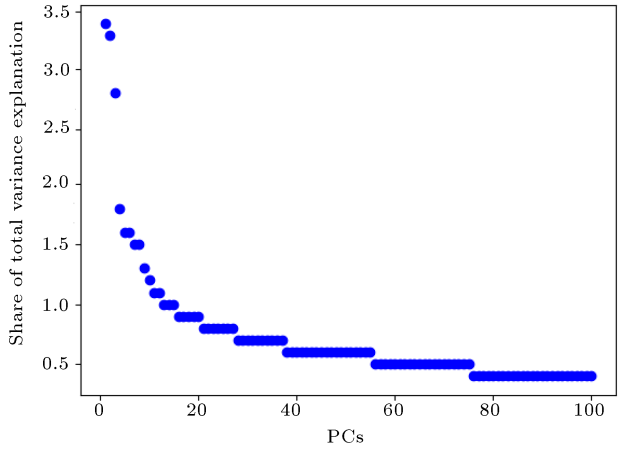

Figure 6. The share of total variance explanation of each Principal Components (PC).

CM be an $N \times N$ confusion matrix where N is the total number of traffic states. $CM(i,j)$ stands for the number of the state i assigned to state j by the predictive model. Then, the accuracy (Acc) and F-measure (F_1) formulas are:

$$Acc = \frac{\sum_{m=1}^{N} CM(m,m)}{\sum_{m=1}^{N}\sum_{n=1}^{N} CM(m,n)}, \tag{7}$$

$$F_1(i) = \frac{2Re(i).Pr(i)}{Re(i) + Pr(i)}, \tag{8}$$

where $Re(i)$ and $Pr(i)$ are state i recall and precision:

$$Re(i) = \frac{CM(i,i)}{\sum_{m=1}^{N} CM(i,m)}, \tag{9}$$

$$Pr(i) = \frac{CM(i,i)}{\sum_{m=1}^{N} CM(m,i)}. \tag{10}$$

A machine with a four-core 2.80 GHz processor and

Table 3. Three features with the highest weight for each Principal Component (PC).

PCs	Feature 1	Feature 2	Feature 3
PC1	Number of sequential holidays	Tomorrow is a holiday	Three days ago is a holiday
PC2	Number of sequential holidays	Tomorrow is a holiday	Three years later is a holiday
PC3	Tomorrow is a holiday	Three days later is a holiday	Three days later is a holiday
PC4	It is a holiday	Tomorrow is a holiday	Yesterday is a holiday
PC5	Weekdays	Year	Year
PC6	Year	Year	Season
PC7	Season	Season	Solar months
PC8	Weekdays	Three days later is a holiday	Weekdays
PC9	Season	Season	Lunar months
PC10	Lunar months	Solar months	It is a holiday
PC11	Two days ago is a holiday	Three days ago is a holiday	Three days later is a holiday
PC12	Lunar months	Solar months	Three days later is a holiday
PC13	Hour	Day or night	Hour
PC14	Hour	Two days ago is a holiday	Day or night
PC15	Lunar months	Tomorrow is a holiday	Two days ago is a holiday
PC16	Three days ago is a holiday	Yesterday is a holiday	Two days ago is a holiday
PC17	Tomorrow is a holiday	Three days later is a holiday	Two days ago is a holiday
PC18	Two days ago is a holiday	It is a holiday	Yesterday is a holiday
PC19	Number of sequential holidays	Tomorrow is a holiday	Three days ago is a holiday
PC20	Two days ago is a holiday	Yesterday is a holiday	Lunar months
PC21	Solar months	Lunar months	Lunar months
PC22	Solar months	Lunar months	Yesterday is a holiday
PC23	Three days ago is a holiday	Yesterday is a holiday	It is a holiday
PC24	Three days ago is a holiday	Tomorrow is a holiday	Yesterday is a holiday
PC25	Three days ago is a holiday	Three days later is a holiday	Two days ago is a holiday
PC26	Number of sequential holidays	Tomorrow is a holiday	Three years later is a holiday
PC27	Three days ago is a holiday	Two days ago is a holiday	Two days ago is a holiday
PC28	Two days ago is a holiday	Three days later is a holiday	Three days ago is a holiday
PC29	Day of a lunar month	Day of a solar month	Tomorrow is a holiday
PC30	Tomorrow is a holiday	Three days ago is a holiday	Day of a solar month

Table 4. Computation time taken to train each model.

Model	MNL	DNN-ALL	DNN-PCA	SVM-ALL	SVM-PCA
Computation time consumed (sec)	196	334	81	490	126

Table 5. Traffic state prediction accuracy.

Dataset	MNL	DNN-ALL	DNN-PCA	SVM-ALL	SVM-PCA
Train	83	99	92	100	92
Test	75	67	68	79	78

32 GB memory, running Windows 10, was used to train models. Table 4 shows the computation time taken to train each model.

According to Table 4, in the case of using 30 PCs for training models, the DNN has the least computation time consumed. In addition, training ML models with 92 primary features increased the computation time consumed dramatically. Table 5 shows the accuracy of each model.

According to Table 5, the accuracy of the models varies from 83% to 100% for the training dataset and from 67% to 79% for the testing dataset. The prediction accuracy results show the superiority of the SVM. Moreover, MNL prediction is more accurate than DNN prediction. Application of 30 PCs improves the accuracy of the DNN prediction and decreases that of SVM prediction. The differences in the accuracy values for testing and training datasets in DNN and

Table 6. Model accuracy for each predicted traffic state in the testing dataset.

Traffic state	MNL	NN-ALL	NN-PCA	SVM-ALL	SVM-PCA	Number of records
Light	77	78	79	100	99	6236
Semi-heavy	69	26	28	34	11	1463
Heavy	3	18	17	42	21	244
Blockage	0	0	0	9	0	68

Table 7. Model F-measures for each predicted traffic state in testing and training datasets.

Models	MNL		SVM-ALL		SVM-PCA		DNN-ALL		DNN-PCA	
Dataset	Test	Train	Test	Train	Test	96	78	Train	Test	Train
Light	84	91	88	100	87	78	26	100	80	97
Semi-heavy	55	50	21	100	19	59	12	99	29	83
Heavy	5	6	59	99	32	0	0	99	14	62
Blockage	0	0	0	0	0	96	78	0	0	0

SVM models resulting from ALL 92 primary features are 36% and 21%, respectively, while these differences are reduced to 29% and 14% using 30 PCs.

Tables 6 and 7 demonstrate the accuracy and F-measures of models for each predicted traffic state.

Since 78% of the train dataset records are labeled as light traffic state, as a result of this imbalance in test dataset, the highest accuracy and F-measure of models are achieved and used for predicting the light traffic state and the least accuracy and F-measure are related to the blockage traffic state with less than 1% of total records. Only SVM-ALL can predict some correct blockage traffic state. Of note, 100% accuracy of the SVM-ALL was obtained in predicting light traffic state. This model can predict the most iterative state in the best possible way. It is the MNL model that can predict semi-heavy more accurately than other models. Further, SVM-ALL yields a more accurate prediction for the heavy state.

Table 8 shows the accuracy of models for each solar month in the test dataset.

According to Table 8, the accuracy of the traffic state prediction in the first two months is greater than that in the second three months for all five models. Due to the short-term nature of these predictions, a decrease in prediction accuracy over time is expected. From 21 April to 20 May and from 22 June to 22 July, the MNL is the most accurate model. For other periods, the SVM-ALL model can predict the traffic state more accurately than other models. Interestingly, the accuracy of SVM-ALL and SVM-PCA models is equal in all periods. The difference between the accuracy rates of DNN-ALL and DNN-PCA models is negligible during these five months.

In MNL model as a statistical model, the relationship between the features and traffic state and statistical significance of features can be determined by estimating features coefficient and t-state. This paper used MNL for not only traffic prediction but also fulfilling the lack of interpretability of ML models. Table 8 shows the results of the MNL model. In Table 9, positive coefficients in traffic state utilities indicate that PCs increase the occurrence probability of that state compared to blockage state which is set as the base state. In addition, according to Table 3, the influencing features can be determined. For instance, the coefficient of PC1 in light utility is estimated significantly. The negative sign indicates that PC1 decreases the occurrence probability of light compared to blockage. The first three high weighted features of PC1 are the numbers of sequential holidays; tomorrow is a holiday and three days ago was a holiday. In other words, these three holiday-related features increase the probability of blockage occurrence compared to the

Table 8. Model accuracy for each solar month in the testing dataset.

Models	21 April–21 May	22 May–21 June	22 June–22 July	23 July–22 August	23 August–22 September
MNL	91	84	81	57	57
DNN-ALL	78	81	65	56	44
DNN-PCA	81	83	65	60	44
SVM-ALL	89	88	77	62	60
SVM-PCA	89	88	77	62	60

Table 9. Result of the Multinominal Logit (MNL) model.

Features	Light		Semi-heavy		Heavy		Blockage
	Coefficients (t-state)	Level of significance	Coefficients (t-state)	Level of significance	Coefficients (t-state)	Level of significance	Coefficients
Constant	–1.72 (–34.2)	(* * *)	–4.88 (–22.6)	(* * *)	–8.27 (–10.9)	(* * *)	—
PC1	–0.07 (–5.53)	(* * *)	–0.12 (–5.85)	(* * *)	–0.18 (–3.24)	(* * *)	—
PC2	–0.37 (–25.6)	(* * *)	–0.66 (–13.9)	(* * *)	–0.73 (–7.04)	(* * *)	—
PC3	–0.01 (–0.75)	—	0.2 (4.67)	(* * *)	0.07 (0.64)	—	—
PC4	–0.02 (–1.03)	—	0.01 (0.42)	—	0.05 (1.32)	—	—
PC5	0.29 (13.81)	(* * *)	0.32 (4.68)	(* * *)	0.58 (2.8)	(* * *)	—
PC6	–0.24 (–11.2)	(* * *)	–0.02 (–0.32)	—	0.18 (0.95)	—	—
PC7	–0.33 (–17.2)	(* * *)	–0.36 (–6.34)	(* * *)	0.06 (0.29)	—	—
PC8	0.32 (15.29)	(* * *)	0.25 (4.4)	(* * *)	0.71 (4.09)	(* * *)	—
PC9	–0.01 (–0.27)	—	0.11 (1.96)	(* *)	–0.45 (–2.54)	(* * *)	—
PC10	–0.15 (–6.71)	(* * *)	–0.1 (–1.79)	(*)	0.19 (1.1)	—	—
PC11	0.01 (0.34)	—	–0.28 (–4.12)	(* * *)	–0.83 (–3.93)	(* * *)	—
PC12	–0.06 (–1.95)	(*)	0.12 (1.45)	—	–0.16 (–0.56)	—	—
PC13	–0.28 (–10.8)	(* * *)	–0.39 (–6.83)	(* * *)	–0.24 (–1.61)	—	—
PC14	0.44 (15.44)	(* * *)	0.48 (6.26)	(* * *)	1.07 (4.33)	(* * *)	—
PC15	0.07 (2.15)	(* *)	0.3 (3.9)	(* * *)	1.29 (3.65)	(* * *)	—
PC16	0.27 (7.01)	(* * *)	0.41 (4.31)	(* * *)	–1.55 (–3.29)	(* * *)	—
PC17	0.04 (1.92)	(*)	–0.31 (–4.3)	(* * *)	–0.61 (–2.28)	(* *)	—
PC18	0.13 (2.74)	(* * *)	0.24 (1.74)	(*)	0.99 (4.44)	(* * *)	—
PC19	0.03 (1)	—	–0.09 (–0.73)	—	–0.66 (–3.91)	(* * *)	—
PC20	0.18 (5.63)	(* * *)	0.52 (5.72)	(* * *)	0.78 (2.79)	(* * *)	—
PC21	0.03 (1.16)		0.24 (2.78)	(* * *)	0.34 (1.57)	—	—
PC22	0.23 (7.9)	(* * *)	0.31 (4.62)	(* * *)	0.25 (1.66)	(*)	—
PC23	0.01 (0.44)	—	0.19 (2.03)	(* *)	0.54 (1.77)	(*)	—
PC24	–0.29 (–10.8)	(* * *)	–0.24 (–3.59)	(* * *)	–0.51 (–2.43)	(* * *)	—
PC25	–0.02 (–0.81)	—	0.22 (3.35)	(* * *)	0.38 (1.83)	(*)	—
PC26	0.17 (5.82)	(* * *)	0.45 (6.53)	(* * *)	0.23 (1.58)	—	—
PC27	–0.17 (–5)	(* * *)	0.11 (1.55)	—	0.15 (0.77)	—	—
PC28	–0.26 (–8)	(* * *)	–0.28 (–4.03)	(* * *)	–0.46 (–2.32)	(* *)	—
PC29	–0.16 (–5.7)	(* * *)	–0.34 (–5.69)	(* * *)	–0.43 (–2.28)	(* *)	—
PC30	–0.01 (–0.2)	—	0.08 (1.41)	—	–0.21 (–1.34)	—	—

light state, which seems logical. Other coefficients can be interpreted in the same manner. Moreover, the stars point to the significance of these features. Three, two, and one stars are indicative of significance levels of 99%, 95%, and 90%, respectively. Coefficients without any stars are statistically insignificant.

6. Conclusion

This paper aims to predict traffic state through two different approaches, i.e., statistical and machine learning approaches. To this end, Multinomial Logit (MNL), Deep Neural Network (DNN), and Support Vector Machine (SVM) models were employed, and the prediction capability of these methods was assessed by comparing prediction accuracy. Since there was a limitation concerning the number of features in MNL, Principal Component Analysis (PCA) was used to reduce the data dimension. After converting 92 primary features to 280 features by defining dummy features, PCA reduced the number of features to 30, explaining 37.7% of the total variance. SVM and DNN were trained using two different databases of ALL (92 primary features) and PCA (30 PCs), and the results were compared. One-year records were taken into account to train models and the next six-month records were employed to test the accuracy of predictions.

Overall, it can be concluded that the SVM-ALL model reached the highest total accuracy equal to 79%. After SVM-ALL, the prediction accuracy values of SVM-PCA, MNL, DNN-PCA, and DNN-ALL were 78%, 75%, 67%, and 66%, respectively.

Light, heavy, and blockage traffic states were predicted more accurately using SVM-ALL, compared to other models. In the case of predicting the semi-heavy traffic state, MNL outperformed SVM-all. In general, the prediction accuracy for the first three months of the test dataset was greater than that in the second three months. DNN and SVM failed to assess the explanatory role of features. To address this issue, MNL model parameters including coefficient and t-state elaborated the relationship between features and traffic state and statistical significance of features.

Finally, it is important to determine what parameter is more important for transportation engineers and policymakers: the accuracy of predictions or discovering the effect of independent variables on traffic state. To achieve a greater prediction accuracy, machine learning models like the SVM were proposed; in addition, to have interpretable findings about the relationship between dependent and independent variables, statistical models such as the MNL were suggested. Also, these models could complement each other by employing the MNL first and, then, detecting independent variables that affect traffic state and next train machine learning regarding the results of MNL.

Using predicted traffic state and providing them to travelers and transportation agencies through intelligent transportation systems can ensure a balance between travel demand and supply in the near future, which is the main aim of this study.

As a limitation, this paper only uses the maximum probabilistic method to determine the prediction of MNL. For further research, it is suggested that the random probability method be used for both machine learning and statistical approaches.

Abbreviation

ARIMA	Autoregressive Integrated Moving Average
ATIS	Advanced Travelers Information Systems
ATMS	Advanced Traffic Management Systems
BFGS	Broyden-Fletcher-Goldfarb-Shanno
BPNN	Back-Propagation Neural Network
CG	Conjugate Gradient
CPNN	Clustered Probabilistic Neural Network
DNN	Deep Neural Network
DNN-PCAN	Deep Neural Network trained by principle components
DNN-PCA	Deep Neural Network trained by all features
KNN	K-Nearest Neighborhood
LM	Levenberg-Marquardt
ML	Machine Learning
MNL	Multinomial Logit
NN	Neural Networks
PC	Principal Component
PCA	Principal Component Analysis
PE	Processing Element
PNN	Probabilistic Neural Network
RBF	Radial Basis Function
RBFN	Radial Basis Function Network
RF	Random Forest
RPCA	Robust Principal Component Analysis
SANN	Simulated Annealing
SVM	Support Vector Machine
SVM-PCA	Support Vector Machine trained by principal components
SVM-ALL	Support Vector Machine trained by all features
SVR	Support Vector Regression
XGB	Extreme Gradient Boosting

References

1. Pan, T.L., Sumalee, A., Zhong, R.X., et al. "Short-term traffic state prediction based on temporal-spatial correlation", *IEEE trans. Intell. Transp. Syst.*, **14**(3), pp. 1242–1254 (2013).
2. Lee, W.H., Tseng, S.S., and Shieh, W.Y. "Collaborative real-time traffic information generation and sharing framework for the intelligent transportation system", *Inf. Sci.*, **180**(1), pp. 62–70 (2010).
3. Long, J., Gao, Z., Orenstein, P., et al. "Control strategies for dispersing incident-based traffic jams in two-way grid networks", *IEEE Trans. Intell. Transp. Syst.*, **13**(2), pp. 469–481 (2011).
4. Rasaizadi A., Ardestani, A., and Seyedabrishami, S.E. "Traffic management via traffic parameters prediction by using machine learning algorithms", *Int. J. Hum. Cap. Urban Manag.*, **6**(1) pp. 57–68 (2021).
5. Ishak, S. and Al-Deek, H. "Statistical evaluation of interstate 4 traffic prediction system", *Transp. Res. Rec.*, **1856**(1), pp. 16–24 (2003).
6. Liu, Q., Cai, Y., Jiang, H., et al. "Traffic state prediction using ISOMAP manifold learning", *Physica A*, **506**, pp. 532–541 (2018).
7. Golshani, N., Shabanpour, R., Mahmoudifard, S.M., et al. "Modeling travel mode and timing decisions: Comparison of artificial neural networks and copula-based joint model", *Travel. Behav. Soc.*, **10**, pp. 21–32 (2018).
8. Vlahogianni, E.I., Golias, J.C., and Karlaftis, M.G. "Short-term traffic forecasting: Overview of objectives and methods", *Transp. Rev.*, **24**(5), pp. 533–557 (2004).

9. Tang, J., Zheng, L., Han, C., et al. "Statistical and machine-learning methods for clearance time prediction of road incidents: A methodology review", *Anal. Methods Accid. Res.*, **27**, p. 100123 (2020).

10. Khanzode, K.C.A. and Sarode, R.D. "Advantages and disadvantages of artificial intelligence and machine learning: A literature review", *Int. J. Lib. Inf. Sci. (IJLIS)*, **9**(1), p. 3 (2020).

11. Loyola-Gonzalez, O. "Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view", *IEEE Access*, **7**, pp. 154096–154113 (2019).

12. Xu, D.W., Wang, Y.D., Jia, L.M., et al. "Real-time road traffic state prediction based on ARIMA and Kalman filter", *Front. Inf. Technol. Electron. Eng.*, **18**(2), pp. 287–302 (2017).

13. Seyedabrishami, S. and Izadi, A.R. "A copula-based joint model to capture the interaction between mode and departure time choices in urban trips", *Transp. Res. Rec.*, **41**, pp. 722–730 (2019).

14. Lopez-Martin, M., Carro, B., and Sanchez-Esguevillas, A. "Neural network architecture based on gradient boosting for IoT traffic prediction", *Future Gener. Comput. Syst.*, **100**, pp. 656–673 (2019).

15. Luo, C., Huang, C., Cao, J., et al. "Short-term traffic flow prediction based on least square support vector machine with hybrid optimization algorithm", *Neural Process. Lett.*, **50**(3), pp. 2305–2322 (2019).

16. Li, L., Dai, S., Cao, Z., et al. "Using improved gradient-boosted decision tree algorithm based on Kalman filter (GBDT-KF) in time series prediction", *J. Supercomput.*, **76**(9), pp. 1–14 (2020).

17. Yu, H., Ji, N., Ren, Y., et al. "A special event-based K-nearest neighbor model for short-term traffic state prediction", *IEEE Access*, **7**, pp. 81717–81729 (2019).

18. Hardy, M. and Reynolds, J., *Incorporating Categorical Information into Regression Models: The Utility of Dummy Variables*, Handbook of Data Analysis, pp. 229–255 (2004).

19. Juhos, I., Makra, L., and Tóth, B. "Forecasting of traffic origin NO and NO_2 concentrations by support vector machines and neural networks using principal component analysis", *Simul. Model. Pract. Theory.*, **16**(9), pp. 1488–1502 (2008).

20. Moons, E., Wets, G., and Aerts, M. "Nonlinear models for determining mode choice", *Portuguese Conference on Artificial Intelligence* (2007).

21. Xie, C., Lu, J., and Parkany, E. "Work travel mode choice modeling with data mining: decision trees and neural networks", *Transp. Res. Rec.*, **1854**(1), pp. 50–61 (2003).

22. Wang, S. and Zhao, J. "An empirical study of using deep neural network to analyze travel mode choice with interpretable economic information", *Transp. Res. Rec.* (2019). https://trid.trb.org/view/1573064

23. Nassiri, H. and Mohamadian Amiri, A. "Prediction of roadway accident frequencies: Count regressions versus machine learning models", *Sci. Iran.*, **21**(2), pp. 263–275 (2014).

24. Zhao, X., Yan, X., Yu, A., et al. "Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models", *Travel. Behav. Soc.*, **20**, pp. 22–35 (2020).

25. Abasahl, F., Kelarestaghi, K.B., and Ermagun, A. "Gender gap generators for bicycle mode choice in Baltimore college campuses", *Travel. Behav. Soc.*, **11**, pp. 78–85 (2018).

26. Rasaizadi, A. and Kermanshah, M. "Mode choice and number of non-work stops during the commute: Application of a copula-based joint model", *Sci. Iran.*, **25**(3), pp. 1039–1047 (2018).

27. Karlaftis, M.G., and Vlahogianni, E.I. "Statistical methods versus neural networks in transportation research: Differences, similarities and some insights", *Transp. Res. Part C Emerg. Technol.*, **19**(3), pp. 387–399 (2011).

28. Lee, D., Derrible, S., and Pereira, F.C. "Comparison of four types of artificial neural network and a multinomial logit model for travel mode choice modeling", *Transp. Res. Rec.*, **2672**(49), pp. 101–112 (2018).

29. Cheng, L., Chen, X., De Vos, J., et al. "Applying a random forest method approach to model travel mode choice behavior", *Travel. Behav. Soc.*, **14**, pp. 1–10 (2019).

30. Wang, F. and Ross, C.L. "Machine learning travel mode choices: Comparing the performance of an extreme gradient boosting model with a multinomial logit model", *Transp. Res. Rec.*, **2672**(47), pp. 35–45 (2018).

31. Hensher, D.A. and Ton, T.T. "A comparison of the predictive potential of artificial neural networks and nested logit models for commuter mode choice", *Transport Res E-Log*, **36**(3), pp. 155–172 (2000).

32. Filho, R.H. and Maia, J.E.B. "Network traffic prediction using PCA and K-means", *2010 IEEE Network Operations and Management Symposium-NOMS 2010* (2010).

33. Jin, X., Zhang, Y., and Yao, D. "Simultaneously prediction of network traffic flow based on PCA-SVR", *International Symposium on Neural Networks* (2007).

34. Jin, X., Zhang, Y., Li, L., et al. "Robust PCA-based abnormal traffic flow pattern isolation and loop detector fault detection", *Tsinghua Sci. Technol.*, **13**(6), pp. 829–835 (2008).

35. Ferreira, Y.M., Lucas, R.J., Eduardo, P., et al. "Applying a multilayer perceptron for traffic flow prediction to empower a smart ecosystem", *International Conference on Computational Science and Its Applications* (2019).

36. Wu, Y., Tan, H., Qin, L., et al. "A hybrid deep learning based traffic flow prediction method and its understanding", *Transp. Res. Part C Emerg. Technol.*, **90**, pp. 166–180 (2018).

37. LeCun, Y., Bengio, Y., and Hinton, G. "Deep learning", *Nature*, **521**(7553), pp. 436–444 (2015).

38. Yuan, G., Li, T., and Hu, W. "A conjugate gradient algorithm for large-scale nonlinear equations and image restoration problems", *Appl. Numer. Math.*, **147**, pp. 129–141 (2020).

39. Seyedabrishami, S.E., Izadi, A.R., Rayaprolu, H.S., et al. "Car ownership: A joint model for number of cars and fuel types", *Transp. Res. Rec.*, **41**, pp. 572–576 (2019).

40. Jafari Shahdani, F., Rasaizadi, A., and Seyedabrishami, S. "The interaction between activity choice and duration: Application of copula-based and nested-logit models", *Scientia Iranica*, **28**(4), pp. 2037–2052 (2021).

41. Ben-Akiva, M.E., Lerman, S.R., and Lerman, S.R., *Discrete Choice Analysis: Theory and Application to Travel Demand*, **9**, MIT Press (1985).

42. Nash, J.C. "On best practice optimization methods in R", *J. Stat. Softw.*, **60**(2), pp. 1–14 (2014).

43. Iranitalab, A. and Khattak, A. "Comparison of four statistical and machine learning methods for crash severity prediction", *Accid. Anal. Prev.*, **108**, pp. 27–36 (2017).

44. Kabli, A., Bhowmik, T., and Eluru, N. "A multivariate approach for modeling driver injury severity by body region", *Anal. Methods Accid. Res.*, **28**, p. 100129 (2020).

45. Keya, N., Anowar, S., Bhowmik, T., et al. "A joint framework for modeling freight mode and destination choice: Application to the US commodity flow survey data", *Transport Res E-Log*, **146**, p. 102208 (2021).

46. Wang, K., Bhowmik, T., Yasmin, S., et al. "Multivariate copula temporal modeling of intersection crash consequence metrics: a joint estimation of injury severity, crash type, vehicle damage and driver error", *Accid. Anal. Prev.*, **125**, pp. 188–197 (2019).

47. Zhang, J., Li, Z., Pu, Z., et al. "Comparing prediction performance for crash injury severity among various machine learning and statistical methods", *IEEE Access*, **6**, pp. 60079–60087 (2018).

48. Wang, X., An, K., Tang, L., et al. "Short term prediction of freeway exiting volume based on SVM and KNN", *Int. J. Trans. Sci. Tech.*, **4**(3), pp. 337–352 (2015).

49. Han, S., Qubo, C., and Meng, H. "Parameter selection in SVM with RBF kernel function", *World Automation Congress 2012* (2012).

50. Skrandies, W. "Data reduction of multichannel fields: global field power and principal component analysis", *Brain Topogr.*, **2**(1–2), pp. 73–80 (1989).

51. Aguilera, A. and Escabias, M. "Solving multicollinearity in functional multinomial logit models for nominal and ordinal responses", *Func. Oper. Stat.*, pp. 7–13, Springer (2008).

52. Du, K. and Swamy, M. "Principal component analysis", *Neural Netw. Stat. Lear.*, pp. 373–425, Springer (2019).

Biographies

Arash Rasaizadi obtained his BSc degree in Civil Engineering from Shahid Bahonar University of Kerman, Kerman, Iran in 2013. He also received his MSc degree in Transportation Engineering from Sharif University of Technology (SUT), Tehran, Iran in 2015. He is currently a PhD candidate in the Transportation Engineering at Tarbiat Modares University (TMU), Tehran, Iran. His research interests include forecast and management of urban travel demand, short-term prediction of traffic parameters, and large data analysis.

Elahe Sherafat received her BSc degree in Civil Engineering at Yazd University, Yazd, Iran in 2015. She also received her MSc degree in Transportation Engineering from Tarbiat Modares University (TMU), Tehran, Iran in 2020. Her research interests include mainly the application of deep learning algorithms in transportation problems, especially traffic flow prediction and big data analysis.

Seyedehsan Seyedabrishami received his BSc degree in Civil Engineering and MSc degree and PhD in Transportation Engineering from Sharif University of Technology (SUT) in 2004, 2006, and 2011, respectively. He is now an Assistant Professor at the Civil and Environmental Engineering Department at Tarbiat Modares University (TMU) in Tehran. He has recently received a Georg Forster Fellowship as an experienced researcher from the Alexander von Humboldt Foundation in Germany. Since June 2017, he is a visiting professor at the Research Group on Modeling Spatial Mobility at the Technical University of Munich (TUM).