International Journal of Latest Engineering Research and Applications (IJLERA) ISSN: 2455-7137

Volume – 07, Issue – 04, April 2022, PP – 01-07

Traffic data analysis by data mining algorithms

Quang Hoc Tran, Khanh Giang Le*

Faculty of Civil Engineering, University of Transport and Communications, Hanoi, Vietnam

**Corresponding author: gianglk@utc.edu.vn*

Abstract: Traffic congestion is one of the most difficult problems of most major cities in the world. To minimize traffic congestion, it is important to analyse traffic data to show real-time congestion locations or where congestion is frequent or short-term traffic volume forecasting. However, the traffic data is very large and complex. This requires a reasonable method of analysing such complex datasets to obtain the desired results. At present, there is not many articles mentioning the integration of data mining techniques to solve the problem of traffic congestion. Therefore, in this study, data mining techniques were integrated to investigate the traffic data in some main arterial roads in Hai Phong District, Vietnam as a case study. The results confirm the potential for integrating data mining techniques in analysing complex traffic datasets so that more reasonable recommendations can be made to solve the traffic problems. This study helps traffic managers have more tools to manage and control the traffic system effectively and reasonably. This is also the first study about this issue in Vietnam, so the contribution of the article will help the traffic authorities easily solve this problem not only in Hai Phong City but also in other cities.

Keywords: Cluster analysis, data mining,forecast,traffic classification.

I. INTRODUCTION

The development of information communication technology (ICT) and global positioning system (GPS) bringing many efficiencies in real-time traffic management. In addition, short-term traffic prediction is a necessary function of the intelligent transportation system (ITS) [1]. Short-term traffic prediction enables us to estimate traffic situations in the future and providing continuous feedback on traffic conditions [2]. Different from long-term traffic forecasting for traffic planning purposes, short-term forecasting focuses on forecasting in the short period from a few seconds to a few hours. Automated detecting devices collect the majority of the traffic data utilized in short-term forecasts [3].

Aside from rich real-time traffic data, fast evolving research in this sector, which employs standard statistical models and machine learning techniques, has accelerated the use of short-term traffic volume forecasts. For nearly 40 years, short-term traffic forecasting has been a traditional ITS study direction. Following Ahmed and Cook's use of Box-Jenkins' method [4,] a plethora of traditional statistical approaches such as historical average algorithms [5, 6], smoothing [5, 6], Kalman filtering [7], and ARIMA family models [8, 9] have become widely employed in this field. These well-founded mathematical approaches are usually parametric models that work well in model specification; however, when traffic patterns are complex and model parameters are difficult to modify responsively, they become insufficient [10, 11].

Data-driven empirical algorithms have grown successful in short-term traffic forecasting in the last 20 years, as autonomous traffic detecting equipment have become more extensively utilized and machine learning theories have advanced rapidly. These techniques have the benefit of not requiring any assumptions about the underlying model formulations or the uncertainty involved in predicting model parameters. K closest neighbor (KNN) [13, 14], Support Vector Machine (SVM) [15], Random Forest (RF) regression [16], and Artificial Neural Network (ANN) [17] are examples of these techniques.

The KNN, ANN, and ARIMA models are compared with each other, and it is concluded that the KNN is superior [18]. Furthermore, kernel neighborhoods revealed that the strategy gave predictions with similar accuracy to the seasonal version of an ARIMA model [19]. According to previous research, the KNN approaches employed are mostly the basic versions of KNN [20].

There are many methods to solve this problem. However, each method has its pros and cons. Currently, with the strong development of information technology, the application of data mining algorithms has been widely applied to solve problems related to traffic congestion. However, at present, there is no article mentioning the integration of data mining techniques to solve the problem of traffic congestion. Therefore, in this study, data mining techniques were integrated to investigate the traffic data in some main arterials in Hai Phong City, Vietnam as a case study.

.

II. MATERIAL AND METHODS

2.1. Material

2.1.1. Study zone

The study was performed in Hai Phong city, Vietnam. In 2019, Hai Phong covers 1,562 km^2 and has a population of 2,028,514 people. The primary transportation means in Hai Phong are motorbikes, buses, taxis, and an increasing quantity of private cars in recent years. Hai Phong has the high growth rate of personal vehicles in Vietnam with more than 70,000 cars are in circulation and more thousands new registrations each month. The transportation infrastructure system in Hai Phong does not keep up with the current urbanization speed, leading to a very risky of traffic accident and traffic congestion. Meanwhile, the main causes of traffic congestion have not been considered. Therefore, the authors chose Hai Phong City as a case study to pilot the proposed method (Figure 1).

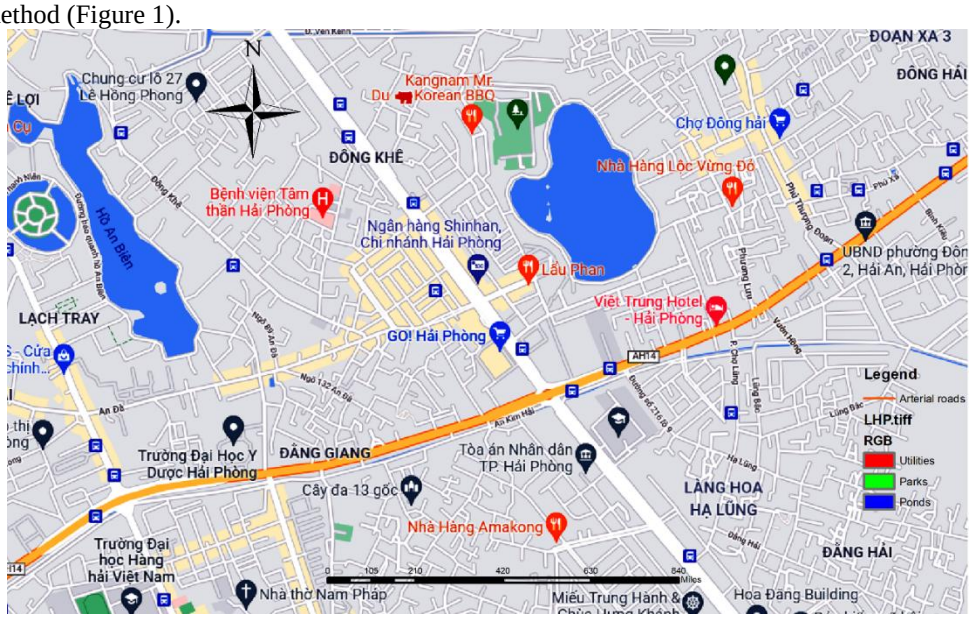

Figure 1. Study area

2.1.2. Data preparation

Traffic data is the key to the problem of reducing traffic congestion. Thus, this study also mentions a reasonable method in collecting traffic data. Two main roads selected in the study area are Le Hong Phong and Nguyen BinhKhiem roads (Figure 1).Removing noise, tackling missing values, and removing irrelevant attributes are very important in data mining techniques. This step makes the dataset available for later analysis [21,22]. The dataset used in this paper is from one of the traffic monitoring stations in Hai Phongcity in Vietnam. The experiment dataset was collected from February 1 to February 29, 2020. Creating time series data from raw travel data using GIS spatial query tools. Journey data is captured in seconds from a variety of various devices. As a result, the volume of data collected is very huge. These data are typically saved in JSON (JavaScript Object Notation) or BSON (Binary JSON) format and processed by MongoDB.

Table 1. An example of the raw data format

ID	Vehicle ID	Provider ID	Date time	Speed	x	y
1	30H02221	2025	1680518220	55	105.32196	11.81535835

Table 1 shows that the raw data only provide us information of each vehicle including coordinates, speed, and date time. We have no way of knowing which road this vehicle is on. Hence, at first, we need to matchthe vehicles with routes by applying Geospatial Queries in MongoDB.

The polygon is then decomposed into a point collection. A GIS programming tool can execute all of these actions. The Arc object programming library is used in this research. After filtering the polygon's points. The following step is to calculate average velocity over various time intervals. We divided a day into 96 equal periods (15 minutes per segment) for this investigation. Each trip point will be included in each corresponding

time interval based on the time of collection (Table 1). The average velocity of the segment within that duration will be determined using all locations in the same timeframe. Table 2 illustrates the dataset after processing.

Table 2. Anexample of the dataset after processing

No	Timeline	Speed	Total Signal	Density
1	2020-02-01T00:00:00.000Z	29.3	58	10
2	2020-02-01T00:15:00.000Z	31.5	55	13
3	2020-02-01T00:30:00.000Z	38.6	92	24
4	2020-02-01T00:45:00.000Z	36.2	62	14
5	2020-02-01T01:00:00.000Z	38.8	45	21
6	2020-02-01T01:15:00.000Z	36.5	83	18
7	2020-02-01T01:30:00.000Z	34.6	68	21
8	2020-02-01T01:45:00.000Z	29.1	45	15
9	2020-02-01T02:00:00.000Z	35.6	33	13
10	2020-02-01T02:15:00.000Z	41.4	33	19
...	...	...	...	...

The data includes only vehicles with an Itinerary monitoring equipment (Only business vehicles, including passenger and goods vehicles, are required to install GPS devices to store and transmit travel information to transportation business units, according to Decree No.86 of the Vietnam Government on transportation business's conditions). According to Table 2, each signal is 15 minutes apart, giving a total of 96 signals every day. Speed: the average speed of all vehicles on the road is considered in these 15 minutes. Total Signal: total number of signals received from vehicles in this 15-minute period. Density: number of vehicles traveling on this stretch of road during this 15-minute period.

Missing value is unavoidable when detecting device is not completely reliable. The numbers of records every day in experiment dataset are shown in Table 2. There is no record on February 28th. For other dates that contain missing value more or less,the interpolation method is used to solve this issue. Traffic volume pattern between holiday and normal dates can be easily identified as Figure 2. Figure 2 shows that the density on weekends is much lower than on weekdays.

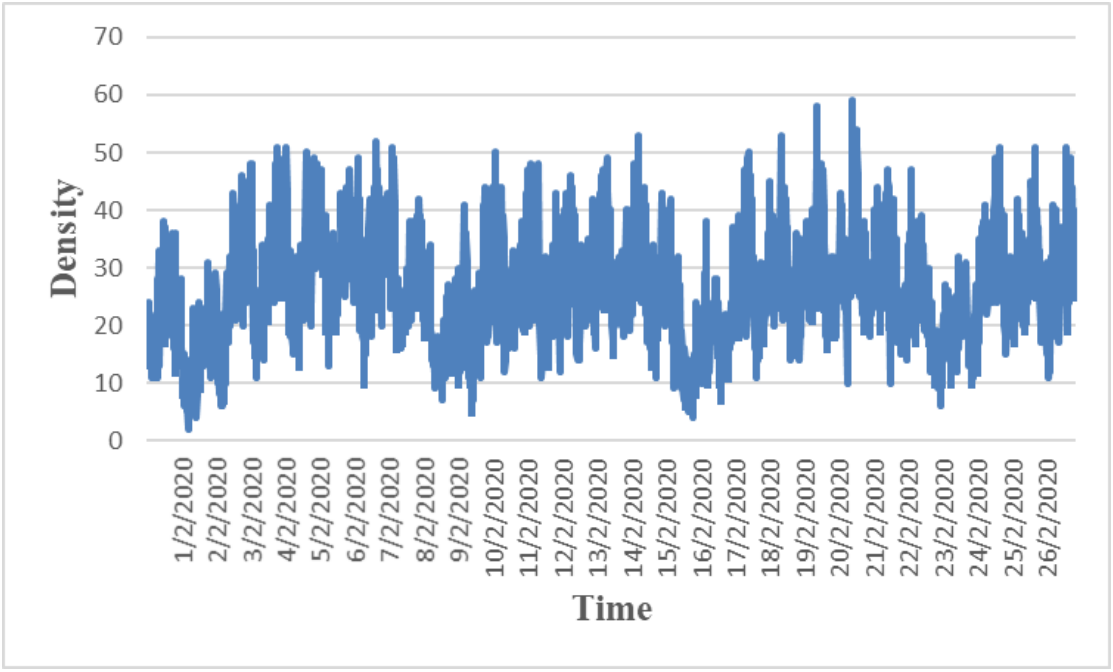

Figure 2. Traffic volume pattern

2.2. Methods

2.2.1. Two-step clustering algorithm

The two-step cluster technique was used in this investigation. Based on Euclidian and log-likelihood distance criteria, the dataset is scanned to see if the current data can be sorted into one of the previously formed clusters or if it should start a new group. Second, values that are inappropriate for any situation will be classified as outliers and removed[23].

To identify the number of clusters, the indicator BIC (Schwarz's Bayesian Information Criterion) or AIC (Akaike's Information Criterion) is computed for each number of clusters from a defined range.

$$\mathrm{BIC(J)} = -2\sum_{j=1}^{J} \xi_j + m_J \log(N) \quad (1)$$

$$\mathrm{AIC(J)} = -2\sum_{j=1}^{J} \xi_j + 2m_J \quad (2)$$

$$\text{where} m_J = J\left(2K^A + \sum_{k=1}^{K^B}(L_k - 1)\right) \quad (3)$$

2.2.2. AutoClass

We use an implementation of a probabilistic model-based clustering technique called AutoClass to compare with the outputs from two-step clustering algorithm. This algorithm allows for the automatic selection of the number of clusters and the soft clustering of the data.

Clustering process is made by AutoClassand two-step clustering methods. Concordance of clustering algorithms were evaluated with Kappa statistics. Thestatistical significance level was 0,05 and WEKA and SPSS were utilized for the analysis.

2.2.3. Time-series Forecasting Method

Autoregressive integrated moving average forecasting methods is a univariate process. These techniques, often called the Box-Jenkins forecasting methodology, have the following steps:Model identification and selection; Estimation of autoregressive (AR), integration or differencing (I), and moving average (MA) parameters; Model checking. Current values of a data series are correlated with past values in the same series to produce the AR component. Current values of a random error term are correlated with past values to produce the MA component. An I component is added to correct for a lack of stationarity through differencing.

III. RESULTS AND DISCUSSION

All numerical variables descriptive statistics were given as mean, standard deviation, minimum and maximum in Table 3.

Table 3.Descriptive statistics for numerical variables

Variables	Mean	Std. Deviation	Minimum	Maximum
Speed	30.77	11.53	0.00	51.27
Total Signal	97.92	58.86	0.00	340
Density	23.50	11.79	0.00	59

3.1. Cluster Analysis

3.1.1. Two-step cluster algorithm

All variables mentioned in Table 2 were used in this process. The number of clusters was chosen through the indicators of BIC or AIC. The outputs obtained from both indicators BIC and AIC are the same. The largest rate of distance measures is 2.166 for four clusters. Therefore, the number of clusters was chosen automatically is four was showed in Table4. However, in order to evaluate the number of clusters, we carried out rerun the model with the number of clusters is 3, 5, 6, and 7, respectively. Finally, four clusters are an optimal result.

Table 4 Size of clusters by Two-step cluster algorithm.

Cluster	1	2	3	4
Size (%)	38.5	27.6	23.3	10.5
Density	24.98	36.95	15.67	0.13
Speed	36.02	35.84	29.86	0.24

3.1.2. AutoClassalgorithm

Table 5.Size of clusters by AutoClass algorithm.

Cluster	1	2	3	4	5
Size (%)	35	10	19	17	19
Density	28.18	0	18.73	16.56	38.85
Std. dev.	3.10	0	4.17	5.11	4.87
Speed	35.60	0	36.98	27.18	35.66
Std. dev.	3.71	0	3.89	4.50	3.24

Table 5 shows the relative size for the five-cluster solution. This output is not quite good because the sizes of the clusters are not quite similar. However, an advantage of this method is that the centroids for each numerical variable and the frequencies for each categorical variable were presented separately.

From the comparison results between the two methods mentioned above, we can see that the two-step clustering method gives us a number of clusters of fourwhile autoclass algorithm gives us a number of clusters of five. However, the size of clusters in autoclass algorithm is quite different from that of method two-step clustering method.

3.2. Time Series Modeller

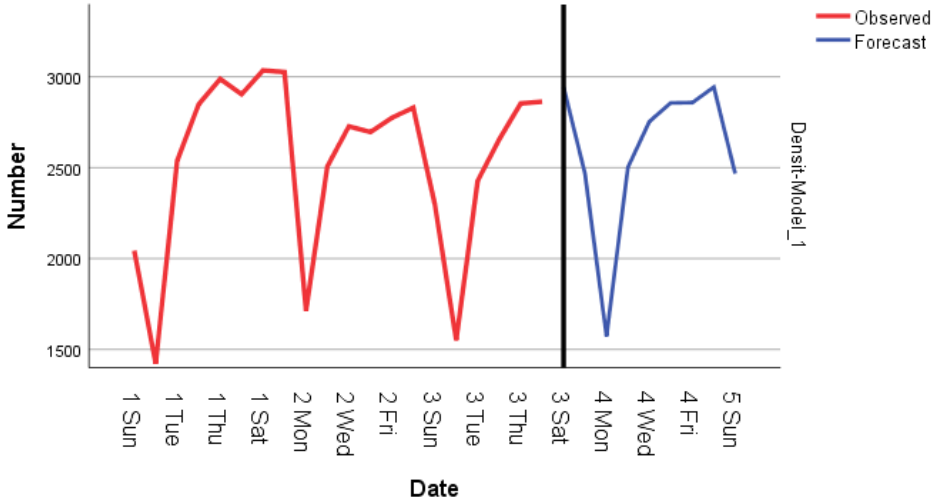

Figure 3.Density Forecasting Modeller

Table 6.Model Fit

Fit Statistic	Mean	SE	Minimum	Maximum	Percentile						
					5	10	25	50	75	90	95
Stationary R-squared	.743	.	.743	.743	.743	.743	.743	.743	.743	.743	.743
R-squared	.869	.	.869	.869	.869	.869	.869	.869	.869	.869	.869
RMSE	170.915	.	170.915	170.915	170.915	170.915	170.915	170.915	170.915	170.915	170.915
MAPE	5.176	.	5.176	5.176	5.176	5.176	5.176	5.176	5.176	5.176	5.176
MaxAPE	23.238	.	23.238	23.238	23.238	23.238	23.238	23.238	23.238	23.238	23.238
MAE	119.615	.	119.615	119.615	119.615	119.615	119.615	119.615	119.615	119.615	119.615
MaxAE	462.139	.	462.139	462.139	462.139	462.139	462.139	462.139	462.139	462.139	462.139
Normalized BIC	10.533	.	10.533	10.533	10.533	10.533	10.533	10.533	10.533	10.533	10.533

Analysis time series model helps forecast traffic flow in a short time with high accuracy. The model helps managers to forecast the traffic to have reasonable adjustment options. Experimental results are shown in

Figure 3, Figure 4, and Table 6. The time series prediction of the ARIMA for all of the average speed on segment 1 is depicted in Figure 3. Figure 3 shows that the blue line is a density prediction model based on observational data. The forecast model is quite accurate with the vehicle density decreasing at the end of the week and high in the middle of the week. As it can be seen, the prediction accuracy of ARIMA with respect to the actual vehicle speed time series is quite acceptable (R-squared = 0.896). Here, we show the results with both the train and test datasets.

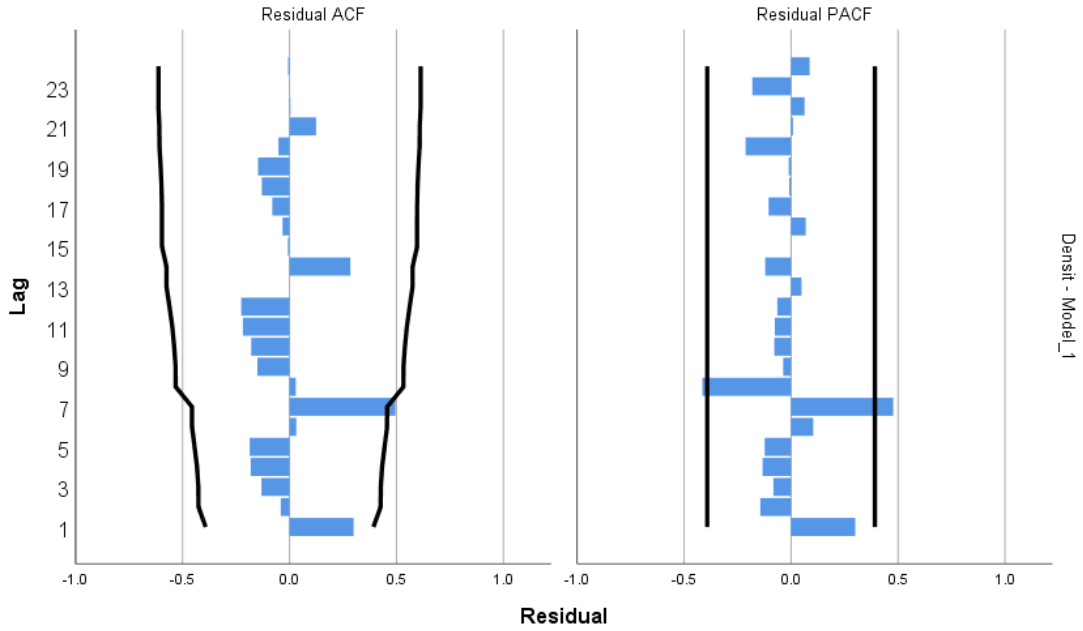

Figure 4.Residual of Time Series Modeller.

Figure 4 shows that the method produces forecasts that appear to account for all available information. The mean of the residuals is close to zero and there is no significant correlation in the residuals series. The time plot of the residuals shows that the variation of the residuals stays much the same across the historical data, apart from the one outlier, and therefore the residual variance can be treated as constant. This can also be seen on the histogram of the residuals. The histogram suggests that the residuals may not be normal — the right tail seems a little too long, even when we ignore the outlier. Consequently, forecasts from this method will probably be quite good, but prediction intervals that are computed assuming a normal distribution may be inaccurate.

IV. CONCLUSION

In this paper, the authors have presented a solution for collecting and processing big data collected from transport business vehicles fitted with itinerary monitoring equipment. The MongoDB database is used to read, process, and compute that big dataset.

Besides, the paper applies data mining techniques to explore traffic data sets. The results show that the combination of data mining techniques has yielded more valuable results than using the techniques alone because each technique has its advantages and disadvantages.

In addition, the article has shown that the number of clusters generated from methods two-step cluster and autoclassis different. Method two-step cluster creates four clusters and method autoclasscreates five clusters. Each cluster represents characteristic parameters including density and speed. When applying the clustering methods, the traffic manager will easily classify traffic, from which there will be a reasonable traffic regulation plan.

Analysis time series model helps forecast traffic flow in a short time with high accuracy. The model helps managers to forecast the traffic to have reasonable adjustment options.

Acknowledgements

This research is funded by the Ministry of Education and Training, Vietnam under grant number CT.2019.05.02.

REFERENCES

[1]. Cheslow M., Hatcher G., Patel V., "An initial evaluation of alternative intelligent vehicle highway systems architectures," System Architecture, Article ID 92W0000063, 1992.
[2]. Vlahogianni E. I., Golias J. C., Karlafis M. G., "Short-term traffic forecasting: overview of objectives and methods," Transport Reviews, vol. 24, no. 5, pp. 533–557, 2004.
[3]. Wang Z., Ji S., Yu B., "Short-Term Traffic Volume Forecasting with Asymmetric Loss Based on Enhanced KNN Method," Mathematical Problems in Engineering, pp. 1–11, 2019, DOI:10.1155/2019/4589437
[4]. Ahmed M. S., Cook A. R., "Analysis of freeway traffic timeseries data by using box–jenkins techniques," Transportation Research Record, no. 722, pp. 1–9, 1979.
[5]. Smith B. L., Demetsky M. J., "Multiple-interval freeway traffic flow forecasting," Transportation Research Record, no. 1554, pp. 136–141, 1996.
[6]. Williams B. M., Durvasula P. K., Brown D. E., "Urban traffic flow prediction: application of seasonal autoregressive integrated moving average and exponential smoothing models," Transportation Research Record, no. 1644, pp. 132–141, 1998.
[7]. Okutani I., Stephanedes Y. J., "Dynamic prediction of traffic volume through Kalman filtering theory," Transportation Research Part B: Methodological, vol. 18, no. 1, pp. 1–11, 1984.
[8]. Levin M., Tsao Y. D., "On forecasting freeway occupancies and volumes," Transportation Research Record, vol. 773, pp. 47– 49, 1980.
[9]. Davis G. A., Niham N. L., Hamed M. M., Jacobson L. N., "Adaptive forecasting of freeway traffic congestion," Transportation Research Record, no. 1287, pp. 29-33, 1991.
[10]. Smith B. L., Williams B. M., Oswald R. K., "Comparison of parametric and non-parametric models for traffic flow forecasting," Transportation Research Part C: Emerging Technologies, vol. 10, no. 4, pp. 303–321, 2002.
[11]. Clark S., "Traffic prediction using multivariate nonparametric regression," Journal of Transportation Engineering, vol. 129, no. 2, pp. 161–168, 2003.
[12]. Vlahogianni E. I., Karlaftis M. G., Golias J. C., "Shortterm traffic forecasting: where we are and where we're going," Transportation Research Part C: Emerging Technologies, vol. 43, pp. 3–19, 2014.
[13]. Cheng S., Lu F., Peng P., Wu S., "Short-term traffic forecasting: anadaptive ST-KNN model that considers spatial heterogeneity," Computers Environment & Urban Systems, vol. 71, pp. 186–198, 2018.
[14]. Guo F., Polak J. W., Krishnan R., "Predictor fusion for short-term traffic forecasting," Transportation Research Part C: Emerging Technologies, vol. 92, pp. 90–100, 2018.
[15]. Sun Z., Fox G., "Traffic flow forecasting based on combination of multidimensional scaling and SVM," International Journal of Intelligent Transportation Systems Research, vol. 12, no. 1, pp. 20–25, 2014.
[16]. Hamner B., "Predicting travel times with context-dependent random forests by modeling local and aggregate traffic flow," The 10th IEEE International Conference on Data Mining Workshops, Sydney, Australia, December, 2010, pp. 1357–1359.
[17]. Van L. H., Van H. C., "Short-term traffic and travel time prediction models," Transportation Research Circular, vol. 22, no. 1, pp. 22–41, 2012.
[18]. Smith B. L., Demetsky M. J., "Traffic flow forecasting: comparison of modelling approaches," Journal of Transportation Engineering, vol. 123, no. 4, pp. 261–266, 1997.
[19]. Smith B. L., Williams B. M., Oswald K. R., "Parametric and nonparametric traffic volume forecasting," Transportation Research Board 79th Annual Meeting, 2000, p. 29.
[20]. Habtemichael F. G., Cetin M., "Short-term traffic flow rate forecasting based on identifying similar traffic patterns," Transportation Research Part C: Emerging Technologies, vol. 66, pp. 61–78, 2016.
[21]. Han J., Pei J., Kamber M., "Data Mining: Concepts and Techniques," in Morgan Kaufmann, 2011, pp. 1-200.
[22]. Tan P. N., Steinbach M., Kumar V., "Introduction to data mining," in Pearson Addison-Wesley, 2006, pp. 1-150.
[23]. Garson G. D., "Cluster analysis," Statistical Publishing Associates, 2014, pp. 1-100.

