Journal of Advanced Transportation

Advanced Data Intelligence Theory and Practice in Transport

Lead Guest Editor: Inhi Kim

Guest Editors: Dong-Kyu Kim, Young-Jae Lee, and Zhibin Li

Advanced Data Intelligence Theory and Practice in Transport

Journal of Advanced Transportation

Advanced Data Intelligence Theory and Practice in Transport

Lead Guest Editor: Inhi Kim
Guest Editors: Dong-Kyu Kim, Young-Jae Lee, and Zhibin Li

Associate Editors

Academic Editors

Contents

Hindawi
Journal of Advanced Transportation
Volume 2022, Article ID 1107048, 11 pages
https://doi.org/10.1155/2022/1107048

Research Article

Applying Clustered KNN Algorithm for Short-Term Travel Speed Prediction and Reduced Speed Detection on Urban Arterial Road Work Zones

Hyun Su Park,[1] **Yong Woo Park**,[2] **Oh Hoon Kwon**,[2] **and Shin Hyoung Park**[3]

[1]*The Institute of Urban Science, University of Seoul, Seoul 02504, Republic of Korea*
[2]*Department of Transportation Engineering, Keimyung University, Daegu 42601, Republic of Korea*
[3]*Department of Transportation Engineering, University of Seoul, Seoul 02504, Republic of Korea*

Correspondence should be addressed to Shin Hyoung Park; shinhpark@uos.ac.kr

Received 18 June 2021; Revised 20 December 2021; Accepted 5 January 2022; Published 4 February 2022

Academic Editor: Inhi Kim

This study developed and verified a travel speed prediction model based on the travel speed and work zone statistics collected from the advanced traffic management system (ATMS) real-time data in Daegu, South Korea. A clustered K-nearest neighbors (CKNN) algorithm was used to predict travel speed, resulting in a 6.9% average mean absolute percentage error (MAPE) using the data from 1,815 work zones. Furthermore, road network impact due to road work was calculated by comparing the travel speed prediction results obtained from the historical speed data. The predicted travel speed data in a work zone generated from this study is expected to allow drivers to select optimized paths and use them for traffic management strategies to operate in a work zone efficiently.

1. Introduction

A downside involved in road works is the reduction in the capacity of a road, leading to traffic congestion and inconveniencing drivers. In Daegu Metropolitan City, South Korea, an average of 20 road works are carried out per day with limited information provided in advance, such as construction schedules. Other essential details, including expected traffic congestion and estimated travel time based on road works, are not disclosed. Therefore, drivers navigating or near the work zone may experience significantly longer travel times than expected due to the restricted notices. Predicting the network impact of road constructions can provide drivers with opportunities to choose detours [1, 2] and allow road managers to use it as data for establishing traffic operation strategies in case of traffic congestion. It is necessary to have a system that predicts the impact of work on traffic flow to reduce congestion caused by frequent road works while providing information to drivers or road managers ahead of time. In addition, an algorithm for predicting the speed of neighboring road links after road work and a method for understanding the effect of road construction on the network should be developed.

Daegu Metropolitan City operates the advanced traffic management system (ATMS) to provide traffic condition information on urban roads. However, the data generated by the system does not reflect real-time traffic information, so inconsistencies in the actual road conditions arise. It is necessary to provide the system's users with estimated travel speed or time to change their travel plans appropriately based on the provided traffic information. Meanwhile, numerous studies have been conducted to predict traffic conditions in urban networks [3], but studies on predicting traffic conditions, particularly in urban construction sections, are relatively insufficient [4]. This study aims to develop an algorithm to predict traffic speed after road work in an urban area and present a method for determining whether the road network is affected. The travel speed prediction model for the work area was developed through the clustered K-nearest neighbors (CKNN) algorithm using traffic statistics collected from Daegu ATMS and work data gathered from the urban traffic information system (UTIS). Furthermore, road network impact due to road construction was calculated by comparing the driving speed prediction results obtained from the historical speed data.

This study is composed of five sections. Section 2 reviews previous studies related to this study, while Section 3 describes data used for this study and its preprocessing. Moreover, Section 4 describes how to design and verify a CKNN model, and Section 5 discusses the research result and the follow-up study.

1.1. Literature Review. Predicting travel speed or travel time has been an active research topic for decades, and as a result, various predictive models have been developed [5]. In early studies, parametric methods such as autoregressive moving average (ARMA), Kalman filter [6], autoregressive integrated moving average (ARIMA) [7, 8], and seasonal autoregressive integrated moving average (SARIMA) [9] were utilized [10, 11]. However, parametric methods were difficult to implement in real-time traffic systems due to some problems such as model calibration, validation, and computational challenges [5]. In addition, they have been proven to encounter poor performance compared to nonparametric methods in unstable traffic conditions and complex road settings [12, 13]. Neural network (NN), K-nearest neighbors (KNN) [14], Bayesian network (BN) [15], and support vector machine (SVM) [10, 16] are the representatives of nonparametric algorithms [17]. Such approaches were advantageous as they are free of assumptions regarding the underlying model formulation and the uncertainty in estimating the model parameters [18]. Recently, studies using deep learning techniques have been conducted to improve the prediction accuracy of traffic conditions [17, 19]. These include long short-term memory (LSTM) [20, 21], deep belief network (DBN) [22], stacked autoencoder (SAE) [23, 24], and convolutional neural network (CNN) [25], which were widely used and had achieved good results in predicting traffic conditions [26]. Nevertheless, a large amount of traffic data was required to utilize nonparametric methods and deep learning strategies, which increased algorithm execution times, making it difficult to present prediction results in real-time [27, 28].

This study considered two factors when selecting the travel speed prediction algorithm. First, the travel speed prediction algorithm must be implemented in the traffic information and management systems. Second, the algorithm should be easily understandable through a traffic manager's level of knowledge and experience. The parametric process was unsuitable for this study because of its complications in implementing it in the system based on these two points. Conversely, a nonparametric method was easier to apply and provided superior prediction performance than a parametric method in unstable traffic conditions. In particular, due to their excellent prediction results, neural networks and KNN algorithms have been used in related studies for a long time. However, it was difficult for analysts and managers to understand neural network models since they use numerous neurons, complex structures, and nonlinear functions [29–31]. Therefore, a KNN model implemented in a real-time traffic system that traffic managers can easily understand would be more appropriate.

KNN algorithms could perform relatively accurate predictions as the data increases, but computation time becomes longer. Liu et al. [32] recognized this problem and used a clustering method to improve it. Clustering is a procedure for grouping data with similar characteristics, and when used in the KNN algorithm, it shortens prediction times and maintains good performance [33]. Hence, a clustering method was applied to the KNN algorithm to compress prediction times and improve the accuracy.

Most previous studies on travel speed prediction in work zones were conducted on highways [34–38]. Prior studies focused on work zones of urban arterial roads, but these were limited to specific links or routes [4, 39]. Based on the cited studies emphasizing the efficacy and potential of the KNN algorithm, this study aims to develop an algorithm that predicts traffic speeds based on changing traffic conditions caused by work zones on urban roads. It also aims to present a method to understand its effect on the road network. This study is not limited to a specific link or route but conducts a travel speed prediction targeting all arterial roads in Daegu Metropolitan City. In addition, this study presents a difference in that few studies suggest a method for judging the effect on networks due to construction on urban arterial roads.

2. Data Description and Preparation

2.1. Standard Node Link Data. Standard node link is Korea's standard transportation network database with a unified identification (ID) system. Among them, link data includes various road information (link ID, number of lanes, road name, speed limit, etc.), as shown in Table 1. In Daegu, the two major systems that collect, process, and provide traffic information are UTIS and ATMS, which efficiently use standard node link-based link IDs to match data between systems. The calculation time of the travel speed prediction algorithm in a work zone was shortened, and the accuracy of the prediction result was improved by clustering 1,672 links according to their attributes' similarities.

2.2. ATMS Data. Daegu Metropolitan City provides real-time traffic information to road users by building ATMS, a part of intelligent transportation systems (ITS). ATMS collects individual vehicle travel information such as vehicle IDs and detection time through dedicated short-range communication (DSRC) when a vehicle equipped with an onboard unit (OBU) passes through roadside equipment (RSE) installation points. The collected traffic speed data was generated using the vehicle detection times. Meanwhile, the distance between the roadside devices was then processed into traffic speed data in units of five minutes for each road link. As shown in Table 2, ATMS data fields include standard node link-based ID (STD_LINK_ID), aggregated time

Table 1: Field items of standard link data.

Field name	Description	Field name	Description
LINK_ID	Link ID	ROAD_NAME	Road name
F_NODE	Start node ID	ROAD_TYPE	Road type
T_NODE	End node ID	MAX_SPD	Max. speed limit
LANES	Number of lanes	REST_VEH	Restricted vehicle
ROAD_RANK	Road grade	REST_W	Restricted width
ROAD_NO	Road number	REST_H	Restricted height

Table 2: Field items of advanced traffic management system (ATMS) data.

Field name	Description
STD_LINK_ID	Link ID
GENERATEDDATE	Collected time
SPEED	Speed

(GENERATEDDATE), and speed (SPEED) calculated by aggregating data collected over five minutes. In this study, the traffic speed of the work zone was predicted using ATMS data collected for a total of eight months, from November 2018 to June 2019.

2.3. UTIS Data. UTIS contains event information such as traffic accidents, road construction, events, and weather conditions that happened on the road. Table 3 shows the UTIS data, including various information such as event ID, link ID based on standard node link, event start and end date, event information, and location of occurrence. After collecting the UTIS data from November 2018 to June 2019, the results were used to extract (1,815 cases) ATMS data at the time of road work through link ID matching.

2.4. Data Preprocessing. This study applied travel speed data on arterial roads (1,672 links) collected through ATMS and road works statistics (1,815 cases) for eight months (November 2018 to June 2019). For the same links, ATMS speed data was classified into days with or without road works, and some links showed that road work was performed twice or more within eight months. Since the day of the week was one of the many factors that affect the travel speed, details were constructed by classifying the days from Sunday to Saturday so that the characteristics of the day can be reflected in the travel speed prediction model for the work zones. The network statistics with road works in progress were extracted and used as training data, and the statistics on the networks under the normal condition without road works were utilized to analyze the network impact caused by these construction or maintenance activities.

Because the traffic data generates random noise of measured values from its stochastic characteristics, it is required to remove the noise in historical speed data through smoothing [40]. Thus, moving average, which is considered a smoothing method, was implemented using five-minute time intervals (travel speed for ten minutes before and after including the speed of the i-th time) as in equation (1). The historical speed data was transformed into a smoother form of the time series data with outliers removed, as shown in Figure 1.

$$V_t = \frac{1}{5}\sum_{i=-2}^{i=2} V_{t+i}^{h}, \text{for all samples.} \tag{1}$$

In equation (1), V_t is the speed in time t at which the smoothing operation was performed, and V^h is the historical speed data.

3. Methodology

3.1. Cluster Analysis. Cluster analysis refers to grouping data having a similar pattern [33]. In this study, cluster analysis was performed to improve the accuracy of prediction results and the computation time required for prediction. Travel speed, which is affected by various factors such as road environment (e.g., speed limit, number of lanes, etc.), can be predicted more accurately because the noise from inconsistent data can be removed when clustered by links with similar road environments [33]. Moreover, it is possible to improve the prediction speed of the KNN algorithm by grouping data, which deteriorates as the number of samples increases [32, 41].

As a partitional clustering method, the k-means clustering algorithm was applied because the concept is relatively simple, making it easier for traffic managers to understand. The calculation time is short, making it effortless to use in a real-time information system [42]. The number of lanes and speed limit were used as input variables for k-means clustering. As seen in Table 1, the link contains numerous pieces of information, but the input variables used for k-means clustering analysis are limited. For example, since the road grade or road type means the hierarchy of roads (expressway/general road, highway/urban road/rural road, etc.) rather than link information, it is challenging to use them to cluster similar links. Conversely, the number of lanes and speed limit affect the capacity of the construction section network [43, 44] since network capacity is related to the traffic speed [45]. Therefore, similar links were classified as k-means clustering input variables using the number of lanes and speed limit.

Table 3: Field items of urban traffic information system (UTIS) data.

Field name	Description	Field name	Description
INCIDENTID	Event ID	TROUBLEGRADE	Event grade
LINKID	Link ID	INCIDENTTITLE	Event title
STARTDATE	Start date	INCIDENTINFO	Event information
ENDDATE	End date	INCIDENTCODE	Event type
COORDX	Longitude	INCIDENTSUBCODE	Event subtype
COORDY	Latitude	LOCATION	Address

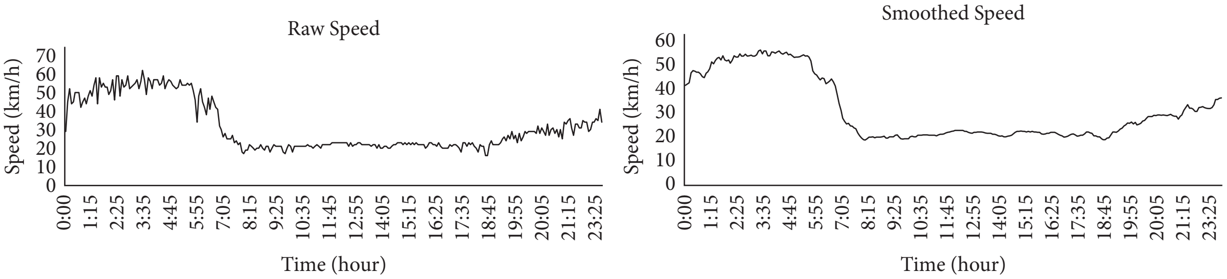


Figure 1: Smoothed historical speed data. (a) Raw data. (b) Smoothed data.

Next, the optimal number of clusters (k) was determined. Various methods for determining k include elbow method, gap statistic, silhouette coefficient, and canopy [42]. The elbow method is utilized in this study, which is the most frequently used k determination method. It is used to select k as the point at which cluster variability (within-cluster sum of squares) becomes smooth with an increase in the number of clusters [46]. For that reason, it was appropriate when the value of k is 3, which is the inflection point of the graph, as illustrated in Figure 2. The three clusters classified through this can be characterized as follows: Cluster 1 is a road with a speed limit of 50 km/h or slower and three lanes or less. Meanwhile, Cluster 2 is a road with a speed limit of 60 km/h or higher and four lanes or more, and Cluster 3 is a road with a speed limit of 60 km/h or higher and three lanes or less. The cluster analysis result was used to find the travel speed pattern most similar to the past when predicting the travel speed of the work zone through the KNN algorithm.

3.2. CKNN Algorithm. The KNN algorithm is a nonparametric method used for classification or regression. It predicts situations by referring to the K training data, which is most similar to the input data [33]. The measure of similarity usually uses the Euclidean distance, which is preferred in predicting a short-term traffic condition because its basic model and calculation time are short, with data matching based on simple similarity. In particular, its prediction is excellent for complex nonlinear problems and can reflect traffic conditions with incidents or traffic jams.

Moreover, the KNN algorithm used in this study predicted the speed up to the forecast duration by referring to the training data of K numbers. This was most similar to the travel speed pattern data during the lag duration before the road work's starting time (t). The detailed analysis procedure of the KNN algorithm is presented in Figure 3. When designing the KNN algorithm, 1,453 out of 1,815 units of data were used as training data for predictive model design, and 362 units of data were utilized as test data for finding the K and algorithm verification.

$$d = \sqrt{\left(V_t^c - V_t^h\right)^2 + \left(V_{t-1}^c - V_{t-1}^h\right)^2 + \left(V_{t-2}^c - V_{t-2}^h\right)^2 + \cdots + \left(V_{t-l}^c - V_{t-l}^h\right)^2}. \tag{2}$$

Data with the same day of the week and link cluster number as the input data is filtered from historical travel speed statistics to find the most similar travel speed pattern to the past. This study referred to the CKNN algorithm because the cluster results were used to find similar travel speed patterns [33]. Then, equation (2) was used to calculate the Euclidean distance between travel speeds of lag duration, and K training data was selected in order of the smallest Euclidean distance.

Here, d is Euclidean distance, V_t is speed data at the road work start time t, V^c is real-time speed data, V^h is historical speed data, and l is lag duration.

Finally, the travel speed for each period (in five-minute intervals) since the start of road work of the most similar K training data was reflected in equation (3) to predict travel speed after the current road work commenced. The method for pattern matching using Euclidean distance is shown in Figure 4.

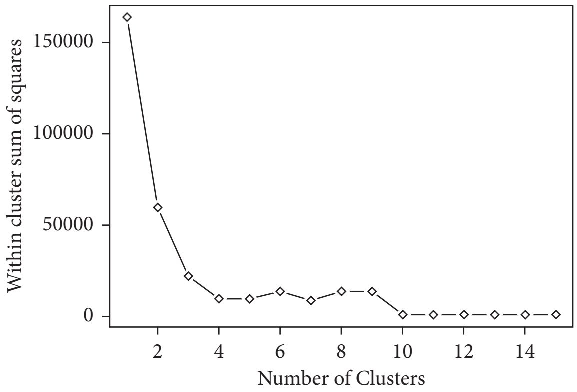


Figure 2: Determination of k using the elbow method.

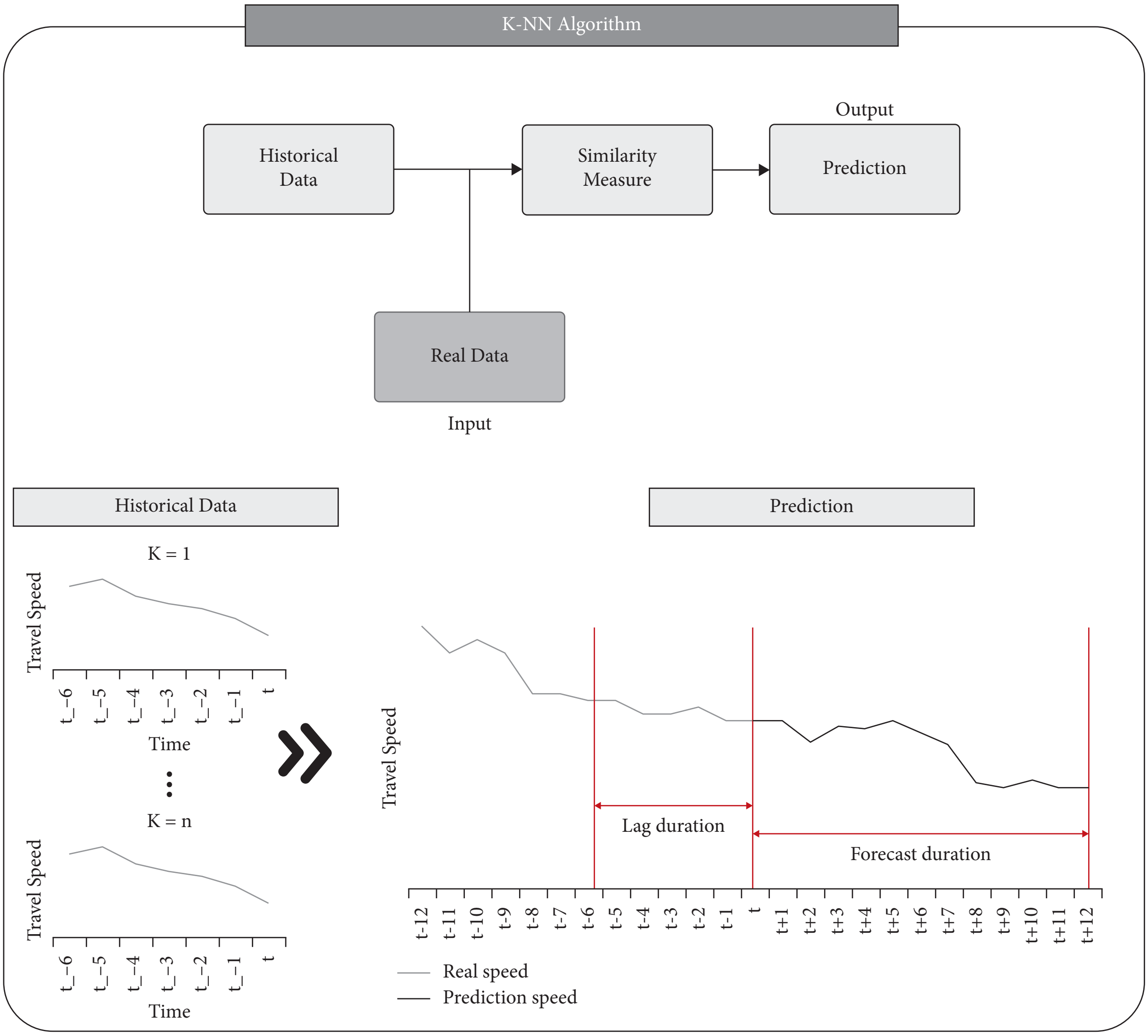
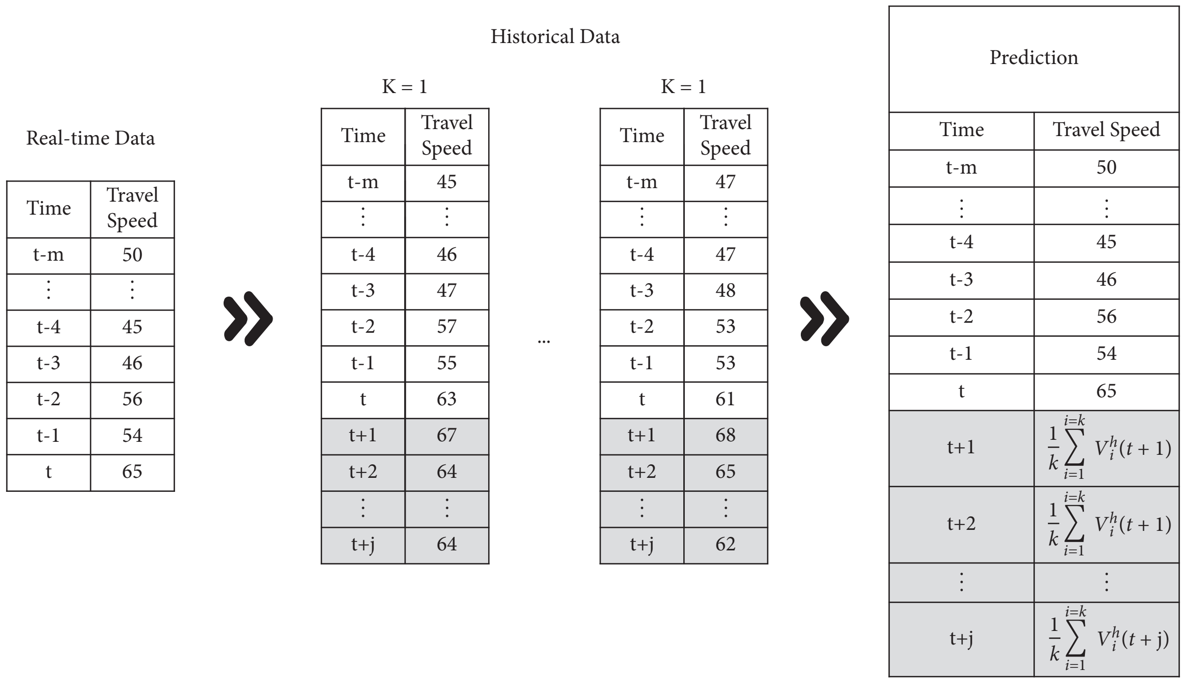


Figure 3: Framework of clustered K-nearest neighbors (CKNN) algorithm.



Historical Data

Real-time Data

Time	Travel Speed
t-m	50
⋮	⋮
t-4	45
t-3	46
t-2	56
t-1	54
t	65

K = 1

Time	Travel Speed
t-m	45
⋮	⋮
t-4	46
t-3	47
t-2	57
t-1	55
t	63
t+1	67
t+2	64
⋮	⋮
t+j	64

...

K = 1

Time	Travel Speed
t-m	47
⋮	⋮
t-4	47
t-3	48
t-2	53
t-1	53
t	61
t+1	68
t+2	65
⋮	⋮
t+j	62

Prediction

Time	Travel Speed
t-m	50
⋮	⋮
t-4	45
t-3	46
t-2	56
t-1	54
t	65
t+1	$\frac{1}{k}\sum_{i=1}^{i=k} V_i^h(t+1)$
t+2	$\frac{1}{k}\sum_{i=1}^{i=k} V_i^h(t+1)$
⋮	⋮
t+j	$\frac{1}{k}\sum_{i=1}^{i=k} V_i^h(t+\mathrm{j})$

Figure 4: Speed prediction method based on Euclidean distance.

$$V^p_{t+j} = \frac{1}{K}\sum_{i=1}^{i=\mathrm{K}} V^h_{t+j}(i). \qquad (3)$$

In equation (3), V^p is the predicted travel speed, V^h is the historical speed data, $V(i)$ is i-th nearest neighbor speed, K is the number of nearest neighbors, and j is the prediction horizon (every five minutes).

3.3. Selection of Optimal K and Appropriate Lag Duration (in CKNN Algorithm). When designing the CKNN algorithm, it is imperative to determine the lag duration required for Euclidean distance calculation and the optimal K and forecast duration. The lag duration and K can be selected using mean absolute percentage error (MAPE), mean absolute error (MAE), and root mean square error (RMSE), which are methods for evaluating the predictive power of a model. In this study, predictions were performed by varying the lag duration and K value using test data, and the results were compared to select the most suitable lag duration and K for the model. The lag duration and the optimal K were chosen based on the prediction accuracy of the CKNN algorithm using equations (4)–(6) or the three error criteria.

$$\mathrm{MAPE}(\%) = \frac{1}{n}\frac{1}{m}\sum_{i=1}^{i=n}\left(\sum_{j=1}^{j=m}\left|\frac{V^a_{t+j}(i) - V^p_{t+j}(i)}{V^a_{t+j}(i)}\right|\right) \times 100, \qquad (4)$$

$$\mathrm{MAE} = \frac{1}{n}\frac{1}{m}\sum_{i=1}^{i=n}\left(\sum_{j=1}^{j=m}\left|V^a_{t+j}(i) - V^p_{t+j}(i)\right|\right), \qquad (5)$$

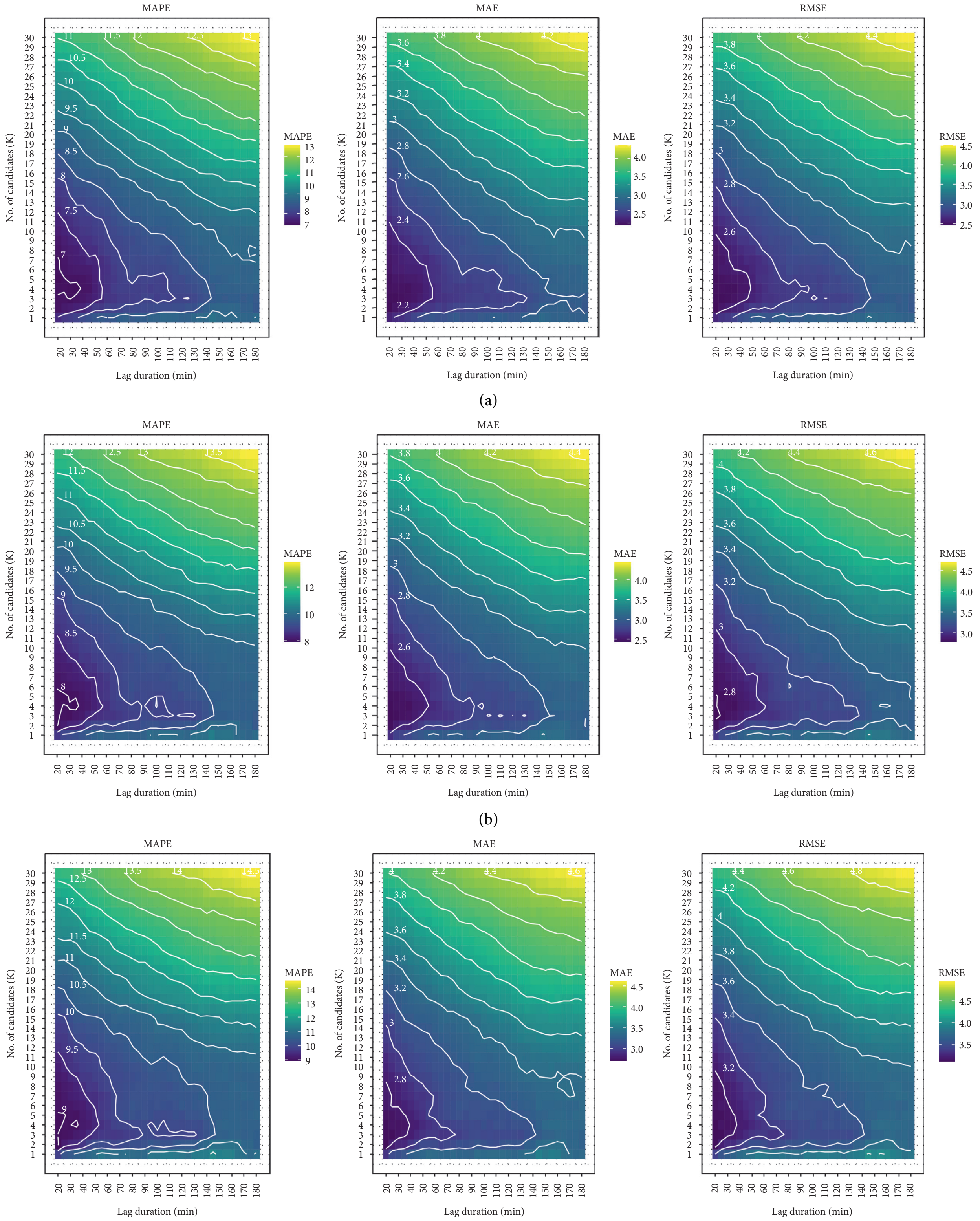
$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\frac{1}{m}\sum_{i=1}^{i=n}\left(\sum_{j=1}^{j=m}\left(V^a_{t+j}(i) - V^p_{t+j}(i)^2\right)\right)}. \qquad (6)$$

Here, V^a is actual speed data, V^p is the predicted travel speed, $V(i)$ is i-th work zone link speed, n is the number of data on the road works, and m is the number of time intervals for forecast duration.

3.4. Determination of the Impact of Road Work on the Network and Travel Speed Degradation. Road work in an urban area may or may not reduce the network traffic speed depending on the scale or type of work. Thus, a probability distribution model was applied to determine whether the actual road work causes the decrease in the travel speed under road work. The impact of road work on the network was determined by comparing the speed under normal network conditions with speed predicted through the CKNN algorithm by checking if the confidence level at 95% is met. Assuming that speeds under normal conditions were the standard normal distribution when the predicted value satisfied the 95% confidence level of the average speed under normal conditions, the road work had no impact on the network.

$$z = \left|\frac{V_i^p - \overline{v}_i^h}{\sigma}\right|. \qquad (7)$$

In equation (7), V_i^p is the predicted travel speed in time i, $\overline{v}_i^h$ is the average speed in time i, and σ is the standard deviation in normal speed in time i.



(a)

(b)

(c)

Figure 5: Impact of lag duration and number of candidates on prediction error. (a) Forecast duration: 1 hour. (b) Forecast duration: 2 hours. (c) Forecast duration: 3 hours.

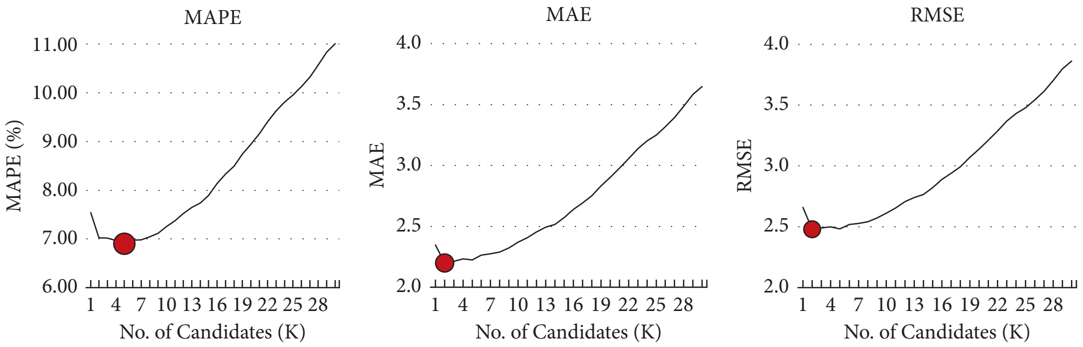


Figure 6: Optimal K given the lag duration is 20 min.

If the z-score calculated through equation (7) is higher than $z_{0.025}$, or the critical value of the 95% confidence interval of the average travel speed at normal conditions, the road work affected the network. When calculating the predicted network degradation caused by road work, equation (8) must be applied to determine the speed degradation against the average speed.

$$\text{speed degradation}\,(\%) = \frac{V_i^p - \overline{v}_i^h}{\overline{v}_i^h} \times 100. \qquad (8)$$

3.5. Case Study. In this section, the forecast accuracy of the prediction algorithm was evaluated by selecting the optimal K and lag duration when predicting the travel speed of the work zone. The predicted speed during road work and the normal speed without road work were determined, including the work zone's impact on network and travel speed degradation.

First, the CKNN algorithm and 362 test data were used to select the optimal K and lag duration. Based on the three error criteria presented above (equations (4)–(6)), the accuracy of the CKNN algorithm was analyzed according to the increase in K and lag duration for each prediction time (one hour, two hours, three hours), as illustrated in Figure 5.

Second, the prediction accuracy based on the three error criteria was most appropriate when the lag duration was 20 minutes. Thus, the speed pattern for the previous 20 minutes should be used when designing the CKNN algorithm. Forecast duration was less accurate when forecasting for a long time, so one hour was identified as the most suitable. Lastly, to find the optimal K, a predictive power evaluation was performed according to the change of the K value when the lag duration was 20 minutes and the forecast duration was 1 hour (Figure 6).

Table 4 shows the values of MAPE, MAE, and RMSE when K has values from 1 to 10. Figure 6 shows that when the K value is ten or more, the value of each criterion continues to increase upward; hence it is unable to find the minimum value. As a result, MAPE obtained the minimum value when K was 5, MAE reached the minimum value when K was 2, and RMSE acquired the minimum value when K was 2 and 5. Since the difference between the MAE values when K was 2 and 5 was detected as small and with scale-dependent errors, the K value was then identified as 5, which minimized MAPE.

Table 4: Mean absolute percentage error (MAPE), mean absolute error (MAE), and root mean square error (RMSE) according to K values when lag duration is 20 minutes.

K	MAPE (%)	MAE	RMSE
1	7.54	2.35	2.66
2	7.02	2.20	2.48
3	7.02	2.21	2.49
4	6.97	2.23	2.50
5	6.90	2.22	2.48
6	6.97	2.26	2.52
7	6.98	2.28	2.53
8	7.05	2.29	2.54
9	7.12	2.32	2.57
10	7.26	2.37	2.61

The test details were used as the input data to verify the model for predicting travel speed in a work zone and the travel speed for an hour after the road work was predicted. The result is presented in Figure 7.

As a result of predicting the test set using the CKNN algorithm, the average MAPE, 6.9%, exhibited excellent predictive power, as indicated in Table 5. In some cases, the MAPE for the predicted value exceeded 15%, but most of them were predictable within 10%. Thus, the model accuracy was considered high [47].

Table 6 specifies the results of analyzing the network impact due to the road work carried out at 10:25 am on Friday, February 15, 2019. The network was classified as Cluster 1. Using the CKNN algorithm, the travel speed one hour after road work was predicted and compared with the network under normal conditions.

Consequently, there was no difference between the speed predicted by CKNN and the normal speed at the beginning of the road work. It was found that the network effect occurred from about 25 minutes after the start of the work, and the speed decreased by about 11%–17% compared to normal conditions, suggesting that road or traffic managers need to establish a strategy to reduce congestion about 30 minutes after starting. Predicting the speed and judging the network impact can also forecast congestion intensity by time, enabling more active and preemptive traffic and congestion management.

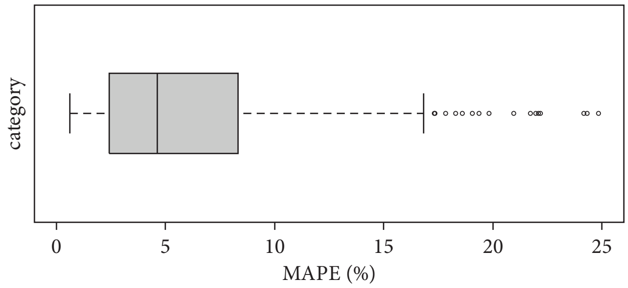


Figure 7: Mean absolute percentage error (MAPE) results at 1-hour prediction ($k = 5$, lag duration = 20 min).

Table 5: Prediction reliability based on mean absolute percentage error (MAPE).

MAPE	Prediction reliability
0% ≤ MAPE < 10%	Very accurate prediction
10% ≤ MAPE < 20%	Accurate prediction
20% ≤ MAPE < 50%	Reasonable prediction
50% > MAPE	Not accurate prediction

Table 6: Impact analysis result of road work based on actual data.

Time		Predicted travel speed (km/h)	Average travel speed at normal conditions (km/h)	Standard deviation of travel speed at normal conditions (km/h)	z-score	Whether the construction affects the speed	Speed degradation (%)
$t+1$	10:30	19.36	21.29	1.51	1.28	No	—
$t+2$	10:35	19.48	21.65	1.50	1.45	No	—
$t+3$	10:40	19.52	21.65	1.45	1.47	No	—
$t+4$	10:45	19.48	21.77	0.92	2.49	No	—
$t+5$	10:50	19.44	21.97	1.35	1.87	Yes	−11
$t+6$	10:55	19.28	21.58	1.12	2.06	No	—
$t+7$	11:00	18.84	21.84	1.21	2.47	Yes	−11
$t+8$	11:05	18.64	22.35	1.60	2.32	Yes	−14
$t+9$	11:10	18.60	22.39	1.60	2.36	Yes	−17
$t+10$	11:15	18.56	22.45	1.43	2.71	Yes	−17
$t+11$	11:20	18.72	21.94	0.89	3.60	Yes	−15
$t+12$	11:25	19.04	22.10	0.79	3.87	Yes	−14

4. Conclusion

It is crucial to prepare an appropriate traffic management strategy for the expected congestion level by predicting the travel speed after road work to prevent congestion caused by road works. This study developed a model that predicts the travel speed of the work zone using the CKNN algorithm. Furthermore, a method to grasp how much the traffic speed decreases due to road work was compared with the normal speed pattern.

Most proposed methodologies for short-term speed prediction presented by several existing studies were methods for predicting speed in normal road conditions. Since roads in the work zone were entirely or partially blocked, a speed pattern differed from normal road conditions. Applying the proposed methodology through a case study can accurately predict the speed from the start of road construction up to an hour later. Furthermore, it was likewise feasible to provide useful information for preemptive traffic congestion management by detecting the timing of link speed degradation caused by capacity reduction due to road work.

However, this study had limitations that need improvement through future studies. First, the established model for predicting travel speeds in a work zone filtered the data using the day of the week and link clusters classified according to road characteristics. Still, it is necessary to use work type as a filter. For example, road works that block or occupy roads largely affect traffic conditions. However, work conducted on drains or sidewalks will only slightly influence traffic conditions. Therefore, better results can be achieved if the travel speed of the work zone can be predicted by considering the work type.

Second, a prediction model was developed using eight-month data for major arterial roads installed with traffic information collection devices. Although the amount of data was not small, it was still insufficient for securing details similar to the input data.

Third and last, this study used only the CKNN algorithm for speed prediction. However, evaluating the appropriateness of the methodology proposed in this study compared to results predicted by other clustering methods such as support vector machines, random forests, and neural networks is required.

Data Availability

The data used to support the findings of this study are not publicly made available according to the data security policy of Daegu Metropolitan City.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the 2021 Research Fund of the University of Seoul.

References

[1] N. G. Polson and V. O. Sokolov, “Deep learning for short-term traffic flow prediction,” *Transportation Research Part C: Emerging Technologies*, vol. 79, pp. 1–17, 2017.

[2] E. J. Kim, H. C. Park, S. Y. Kho, and D. K. Kim, “A hybrid approach based on variational mode decomposition for analyzing and predicting urban travel speed,” *Journal of Advanced Transportation*, vol. 2019, Article ID 3958127, 12 pages, 2019.

[3] H. C. Park, S. Kang, S. Y. Kho, and D. K. Kim, “Investigation of effects of inherent variation and spatiotemporal dependency on urban travel-speed prediction,” *Journal of Transportation Engineering, Part A: Systems*, vol. 146, no. 5, 2020.

[4] Y. Hou, P. Edara, and C. Sun, “Traffic flow forecasting for urban work zones,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 1761–1770, 2014.

[5] Z. Zheng and D. Su, “Short-term traffic volume forecasting: a K-Nearest neighbor approach enhanced by constrained linearly sewing principle component algorithm,” *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 143–157, 2014.

[6] J. Guo, W. Huang, and B. M. Williams, “Adaptive Kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification,” *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 50–64, 2014.

[7] S. Lee and D. B. Fambro, “Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting,” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1678, no. 1, pp. 179–188, 1999.

[8] D. Billings and J. S. Yang, “Application of the ARIMA models to urban roadway travel time prediction-a case study,” *In 2006 IEEE International Conference on Systems, Man and Cybernetics*, vol. 3, pp. 2529–2534, 2006.

[9] A. M. Khoei, A. Bhaskar, and E. Chung, “Travel time prediction on signalised urban arterials by applying SARIMA modelling on Bluetooth data,” in *Proceedings of the 36th Australasian Transport Research Forum (ATRF)*, Brisbane, Australia, 2013.

[10] M. Castro-Neto, Y.-S. Jeong, M.-K. Jeong, and L. D. Han, “Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions,” *Expert Systems with Applications*, vol. 36, no. 3, pp. 6164–6173, 2009.

[11] L. Kang, G. Hu, H. Huang, W. Lu, and L. Liu, “Urban traffic travel time short-term prediction model based on spatio-temporal feature extraction,” *Journal of Advanced Transportation*, vol. 2020, Article ID 3247847, 16 pages, 2020.

[12] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, “Short-term traffic forecasting: where we are and where we’re going,” *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 3–19, 2014.

[13] L. Li, S. He, J. Zhang, and B. Ran, “Short-term highway traffic flow prediction based on a hybrid strategy considering temporal-spatial information,” *Journal of Advanced Transportation*, vol. 50, no. 8, pp. 2029–2040, 2016.

[14] B. Sun, W. Cheng, P. Goswami, and G. Bai, “Short-term traffic forecasting using self-adjusting k-nearest neighbours,” *IET Intelligent Transport Systems*, vol. 12, no. 1, pp. 41–48, 2017.

[15] H.-C. Park, D.-K. Kim, and S.-Y. Kho, “Bayesian network for freeway traffic state prediction,” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2672, no. 45, pp. 124–135, 2018.

[16] Y. Cong, J. Wang, and X. Li, “Traffic flow forecasting by a least squares support vector machine with a fruit fly optimization algorithm,” *Procedia Engineering*, vol. 137, pp. 59–68, 2016.

[17] P. Wu, Z. Huang, Y. Pian, L. Xu, J. Li, and K. Chen, “A combined deep learning method with attention-based LSTM model for short-term traffic speed forecasting,” *Journal of Advanced Transportation*, vol. 2020, Article ID 8863724, 15 pages, 2020.

[18] F. G. Habtemichael and M. Cetin, “Short-term traffic flow rate forecasting based on identifying similar traffic patterns,” *Transportation Research Part C: Emerging Technologies*, vol. 66, pp. 61–78, 2016.

[19] H. F. Yang, T. S. Dillon, and Y. P. P. Chen, “Optimized structure of the traffic flow forecasting model with a deep learning approach,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2371–2381, 2016.

[20] Y. J. Lee and O. Min, “Long short-term memory recurrent neural network for urban traffic prediction: a case study of Seoul,” in *Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1279–1284, Maui, HA, USA, November, 2018.

[21] Z. Zhao, W. Chen, X. Wu, P. C. Y. Chen, and J. Liu, “LSTM network: a deep learning approach for short-term traffic forecast,” *IET Intelligent Transport Systems*, vol. 11, no. 2, pp. 68–75, 2017.

[22] W. Huang, G. Song, H. Hong, and K. Xie, “Deep architecture for traffic flow prediction: deep belief networks with multitask learning,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 5, pp. 2191–2201, 2014.

[23] Y. Jia, J. Wu, and Y. Du, “Traffic speed prediction using deep learning method,” in *Proceedings of the 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1217–1222, Rio de Janeiro, Brazil, December, 2016.

[24] Y. Lv, Y. Duan, and W. Kang, “Traffic flow prediction with big data: a deep learning approach,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 865–873, 2014.

[25] C. Song, H. Lee, C. Kang, and W. Lee, “Traffic speed prediction under weekday using convolutional neural networks concepts,” in *Proceedings of the 2017 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1293–1298, Los Angeles, CA, USA, June, 2017.

[26] C. Zhang, J. J. Q. Yu, and Y. Liu, “Spatial-temporal graph attention networks: a deep learning approach for traffic forecasting,” *IEEE Access*, vol. 7, pp. 166246–166256, 2019.

[27] X. Luo, D. Li, Y. Yang, and S. Zhang, "Spatiotemporal traffic flow prediction with KNN and LSTM," *Journal of Advanced Transportation*, vol. 2019, Article ID 4145353, 10 pages, 2019.

[28] X. Yin, G. Wu, J. Wei, Y. Shen, H. Qi, and B. Yin, "Deep learning on traffic prediction: methods, analysis and future directions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 4, 2021.

[29] M. Lippi, M. Bertini, and P. Frasconi, "Short-term traffic flow forecasting: an experimental comparison of time-series analysis and supervised learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 2, pp. 871–882, 2013.

[30] Y. Zhang, Y. Zhang, and A. Haghani, "A hybrid short-term traffic flow forecasting method based on spectral analysis and statistical volatility model," *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 65–78, 2014.

[31] Y. Wu, H. Tan, L. Qin, B. Ran, and Z. Jiang, "A hybrid deep learning based traffic flow prediction method and its understanding," *Transportation Research Part C: Emerging Technologies*, vol. 90, pp. 166–180, 2018.

[32] T. Liu, J. Ma, W. Guan, Y. Song, and H. Niu, "Bus arrival time prediction based on the K-Nearest neighbor method," in *Proceedings of the 2012 Fifth International Joint Conference on Computational Sciences and Optimization*, pp. 480–483, Heilongjiang, China, June, 2012.

[33] M. Akbari, P. J. v. Overloop, and A. Afshar, "Clustered K nearest neighbor algorithm for daily inflow forecasting," *Water Resources Management*, vol. 25, no. 5, pp. 1341–1357, 2011.

[34] M. Kamyab, S. Remias, E. Najmi, S. Rabinia, and J. M. Waddell, "Machine learning approach to forecast work zone mobility using probe vehicle data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2674, no. 9, pp. 157–167, 2020.

[35] H. Wang, L. Liu, S. Dong, Z. Qian, and H. Wei, "A novel work zone short-term vehicle-type specific traffic speed prediction model through the hybrid EMD-ARIMA framework," *Transportation Business: Transport Dynamics*, vol. 4, no. 3, pp. 159–186, 2016.

[36] D. R. Taylor, S. Muthiah, B. T. Kulakowski, K. M. Mahoney, and R. J. Porter, "Artificial neural network speed profile model for construction work zones on high-speed highways," *Journal of Transportation Engineering*, vol. 133, no. 3, pp. 198–204, 2007.

[37] X. Jiang and H. Adeli, "Dynamic wavelet neural network model for traffic flow forecasting," *Journal of Transportation Engineering*, vol. 131, no. 10, pp. 771–779, 2005.

[38] H. T. Zwahlen and A. Russ, "Evaluation of the accuracy of a real-time travel time prediction system in a freeway construction work zone," *Transportation Research Record*, vol. 1803, no. 1, pp. 87–93, 2002.

[39] Q. Meng and J. Weng, "Cellular automata model for work zone traffic," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2188, no. 1, pp. 131–139, 2010.

[40] P. Cai, Y. Wang, G. Lu, P. Chen, C. Ding, and J. Sun, "A spatiotemporal correlative K-Nearest neighbor model for short-term traffic multistep forecasting," *Transportation Research Part C: Emerging Technologies*, vol. 62, pp. 21–34, 2016.

[41] H. Xiaoyu, W. Yisheng, and H. Siyu, "Short-term traffic flow forecasting based on two-tier K-Nearest neighbor algorithm," *Procedia - Social and Behavioral Sciences*, vol. 96, pp. 2529–2536, 2013.

[42] C. Yuan and H. Yang, "Research on K-value selection method of K-means clustering algorithm," *D-J Series*, vol. 2, no. 2, pp. 226–235, 2019.

[43] Transportation Research Board, *Highway Capacity Manual 2010*, TRB National Research Council, Washington, DC, USA, 2010.

[44] A. H. Mashhadi, M. Farhadmanesh, A. Rashidi, and N. Marković, "Review of Methods for Estimating Construction Work Zone Capacity," *Transportation Research Record*, vol. 2675, 2021.

[45] A. Dhamaniya and S. Chandra, "Influence of operating speed on capacity of urban arterial midblock sections," *International Journal of Civil Engineering*, vol. 15, no. 7, pp. 1053–1062, 2017.

[46] T. M. Kodinariya and P. R. Makwana, "Review on determining number of cluster in K-means clustering," *International Journal*, vol. 1, no. 6, pp. 90–95, 2013.

[47] C. D. Lewis, *Industrial and Business Forecasting Methods: A Practical Guide to Exponential Smoothing and Curve Fitting*, Butterworth Scientific, London, UK, 1982.

Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 4738177, 10 pages
https://doi.org/10.1155/2021/4738177

Research Article

Expression and Validation of Online Bus Headways considering Passenger Crowding

Shengyu Yan,[1] Jibiao Zhou ,[2,3] and Zhuanzhuan Zhao[4]

[1]*School of Automobile, Chang'an University, Middle Section of Nan'er Huan Rd., Xi'an 710064, Shaanxi, China*
[2]*Department of Transportation Engineering, Tongji University, Caoan Rd. #4800, Shanghai 201804, China*
[3]*School of Civil and Transportation Engineering, Ningbo University of Technology, Fenghua Rd. #201, Jiangbei District, Ningbo 315211, Zhejiang, China*
[4]*School of Automobile Engineering, Shaanxi College of Communication Technology, Wenjing Rd. #19, Weiyang District, Xi'an 710019, Shaanxi, China*

Correspondence should be addressed to Jibiao Zhou; zhoujibiao@tongji.edu.cn

Received 26 April 2021; Revised 30 July 2021; Accepted 18 August 2021; Published 7 September 2021

Academic Editor: Zhibin Li

Passenger crowding in a city bus is uneven and the most crowded area always appears in the wheelbase of the carriage. The present study aimed to provide a sensitive indicator of the most crowded area to schedule bus headways online using a binocular camera sensor. The algorithm of standee density in the wheelbase area (SDWA) was given by a nonlinear regression model considering standees' preferences for the standing area, and its goodness of fit and continuity were tested. Considering the characteristics of city bus operation, the proportion of the number of interstops determined from the SDWA was used as a judgment index for passenger crowding. Based on the SDWA algorithm and the judgment index, an online headway model of city buses was proposed, and the feasibility of such a model was verified through a case study in Xi'an city. The proposed model might be beneficial to bus scheduling, seating provision, and bus design.

1. Introduction

Standee density refers to the number of standing passengers in a unit effective area. It is an important indicator that reflects whether the bus selected matches the line and whether the headway is rational. The limit load of a city bus in Europe and the United States is 5–6 pax/m^2 [1, 2], while the number of standees approved in China is 8 pax/m^2 [3]. According to surveys, the standee density in a bus in Xi'an city during peak hours reaches 9–10 pax/m^2 and often exceeds the threshold of 8 pax/m^2; however, in actual operation, this reduces ride comfort and overloads buses.

The standee density at various positions within the carriage is actually uneven, and sparsely populated areas at the front and in a rear aisle might affect the true standee density characteristics and impair sensitivity to changes in the standee flow. Therefore, the number of standees cannot truly reflect the maximum passenger crowdedness. The position preference of passengers on each bus line can be determined. According to the passengers' preference for selecting a standing position, this study proposed some important areas to synthesize a standee density algorithm that could sensitively express the most crowded areas in buses to schedule the headway of the bus in real time [4].

Currently, to meet passengers' travel requirements, a drivers' workload on a bus line is fixed for a period of time by public transport enterprises in China, notwithstanding the operational cost. Likewise, the total scheduling frequency is also assigned based on the fixed workload on the line, but the headway varies; the fixed workload refers to that the departure frequency of each bus is constant to ensure the demands of operation. Public transport enterprises in China are totally state-owned enterprises. Regardless, if they are in debt, local governments pay it by the end of the year, as long as they guarantee the necessary services. Consequently, in this case, different headways per day have a slight impact on the operational cost. In the bus scheduling station, the dispatchers are unable to obtain information on the real-time passenger flow. As a consequence, many problems arise, such as personal judgment subjectivity in attendance and the headway not being rational to the bus line [5, 6]. Therefore, overcrowding of passengers often occurs. The main problem, therefore, is the allocation of online headways in the trough hour, off-peak hour, and peak hour.

Similar studies considered the operational cost and passenger waiting time as a balance index for determining the offline bus headway. The main objective of the present study was to determine the online bus headway without the operational cost; correspondingly, the passenger waiting time would surely be a unique index. As the arrival time of each passenger at the bus stop is random and the passenger waiting time is hard to determine online, this study proposed introducing passenger crowding instead of the waiting time. For online scheduling, first, the online passenger flow data obtained by the collector equipped on the front and rear doors of the bus were used. Second, the number of standees was obtained in real time, which represented passenger discomfort during the ride. Therefore, the aforementioned problem led to another problem: determining the online headway according to the standee density.

Due to standee density at various positions being actually uneven, standee density in city buses cannot truly display the most crowded area. In the present study, the number of standees on the bus floor was allocated for each specified area to evaluate the most crowded area. A method for determining the bus headway was established based on the areas of higher standee density on the bus. This model was a pragmatic approach to improve the efficiency of bus transportation and increase the bus travel sharing rate of bus lines with large passenger groups in every city of China.

2. Literature Review

In recent years, considerable attention has been given to witnessing an increased interest in the model of scheduling frequency on public transport. Many bus scheduling models have been established based on offline passenger flow data, resulting in positive effects on the public transport quality of service. In terms of multidimensional analysis of passenger crowdedness, Tirachini optimized the scheduling frequency of subway vehicles considering passenger demand and the supply and operation of public transport [7]. From the standpoint of cost, passenger travel and operational costs were integrated into the newsboy model by Herbon and Hadas [8], who proposed the simulation results of the scheduling frequency of subway vehicles. The standee density is a multipurpose indicator used in pricing strategy, seat capacity, and scheduling arrangement [9]. A route planning and scheduling model has also been proposed based on passenger density and travel distance [10]. In particular, Jara-Díaz proposed an extension of Jansson's model for a single period based on the effect of vehicle size on operational costs and that of crowdedness on the value of time [11].

Assuredly, the impact of standee density on the bus design and travel cost was evaluated from different perspectives. Tirachini developed a social welfare maximization model with externalities of crowdedness, exposing the interplay between congestion and crowdedness in the design of bus systems [12]. The concept of passenger crowdedness involved sitting passengers and standees. It was a coordinating algorithm for the number of passengers in the carriage. In addition, the crowdedness cost had an internal relationship with passenger crowdedness by estimating the willingness of passengers to choose a moderately relaxed trip at different standee densities [13]. However, with a larger scale passenger flow, the standee density was related to the serviceability of the subway. Therefore, a model for calculating the standee density was entrenched, and conclusive recommendations for its standard were proposed [14].

Furthermore, studies discussed the formation mechanism of standee density and key influencing factors in terms of bus door position and passenger preference in choosing a standing area. In addition, a crowd behavior control model was established, simulation studies were conducted at various crowd densities, and the results were used in the decision support tool of crowd control systems [15]. A follow-up survey proposed that door crowdedness was affected by multiple bus design parameters, including door placement, aisle length, presence of a front seating area, and service type [16].

However, the number of standees during morning and evening peak hours is significantly greater than that of sitting passengers in China, and standees have little chance of getting a seat on the buses. Passengers can get on and off a subway from the same door, although they are allowed only to get on from the front door and off from the rear door of almost all buses in China. Hence, the passing flow on board is difficult to determine, which is the root cause of unevenness in standee density [17, 18]. A train mock-up was especially constructed to examine the impact of door width, seat type, platform edge doors, and horizontal gap on the time taken by passengers to board and alight [19].

Batarce explained that the public transport selection preference showed the application requirements of crowdedness cost, and a random discrete selection probability model was established [20]. Moreover, a baseline-category logit model for selecting standing areas was created considering the travel distance of passengers and the standee density in subways; it was also closely related to the door position [21].

In summary, many studies have reported the characteristics of standee density and offline headway. However, few studies have been conducted on the unevenness of standee density to define the most crowded area, aiming to establish an online bus headway model. Most of the aforementioned studies proposed the calculation method of standee density, determined its threshold, and analyzed the travel mode selection and cost-benefit issue based on passenger crowdedness [22]. Thus, these studies considered the operational cost and passenger waiting time to modify the offline bus headway. The present study proposed a model to overcome these challenges to determine the online headway in trough, off-peak, and peak hours according to the standee density.

3. Data Collection

3.1. Online Data Collection. For better efficiency of getting off a bus, usually, the operation mode is paying the bus fare in cash or by a prepaid card without limit. Consequently, the data for the number of passengers getting off are lost. To overcome this problem, the passenger flow data collector was introduced. It automatically collected the number of standees getting on and off the bus at every bus stop to determine the online scheduling arrangement.

The data collector consisted of an analyzer and two binocular camera sensors (Figures 1(a) and 1(b)). It used the human head calibration algorithm. The cameras installed on the front and rear doors of the bus collected video images (Figures 1(c) and 1(d)). The number of passengers getting on and off the bus was processed and transmitted to the monitoring host through the CAN system and then to the information processing platform via 3G/4G wireless communication [23, 24].

The data collection system must be verified manually for the accuracy of passenger flow. After the system started to run, the data collector was arranged for 12 surveys on the bus, although 28 bus stops existed on the surveyed line (Table 1).

The accuracy of data collection slightly reduces when the passenger flow is dense, but the accuracy can still reach 98.8%. A bus line in Xi'an has 20 buses equipped with data collectors, and an increasing number of bus enterprises have adopted binocular camera sensors to monitor passenger traffic in China.

3.2. Manual Survey Data Collection. A manual survey was adopted due to the unevenness of standee density in each area and the inability of passenger flow data collectors to collect the number of standees in each area of a bus carriage. This featured high precision but involved a high labor cost [25]. According to the stipulation in Xi'an, each bus is operated by 2 drivers for 6 round trips per day, and each bus line is equipped with at least 20 buses. When the bus reaches the highest speed between two bus stops, the standee density is relatively stable. Every two investigators were appointed to investigate the number of standees in the designated areas of one bus [26]. After a survey, 79 round-trip passenger flow data points were obtained in this study. The standees moved to the rear door in basically four areas (Table 2).

Since Areas 2 and 3 were close to rear doors, the changes in the standee density in both areas were more sensitive than those in Areas 1 and 4. Although the cost of a manual survey was high, the number of standees in each area of the bus could be flexibly mastered [27]. Each area of the bus floor was measured during the manual survey process. Basically, the standing area within the wheelbase, Areas 2 and 3, was spacious.

4. Methodology

4.1. Headways Based on the Standee Density. The headway is closely related to the time of the first and last buses, the number of buses available, the scheduling task, and the trough, off-peak, and peak hours [28, 29]. Generally, the public transport company's operation workload is fixed to ensure the operational needs of the bus line and the annual review of the vehicle production task [30]. For example, for public transport companies in Xi'an city, it is stipulated that the bus runs six round trips per day and three round trips per driver, and it is recorded as C. If a bus line has m buses and the ratio of the number of buses being repaired and rested per day to the total number of buses is d, then the scheduling frequency available to the dispatcher per day is $Cm(1-d)$ times [31].

Suppose that the time difference between the first and last buses of the line is T_d (min) and that the parking time of the first and last buses is T_c. The first and last buses are scheduled outside of $T_d - T_c$ according to dispatchers, in which the scheduling frequency of the number of buses available for the dispatcher is $Cm(1-d)-2$. Therefore, the headway of the bus line during off-peak hours η_0 is

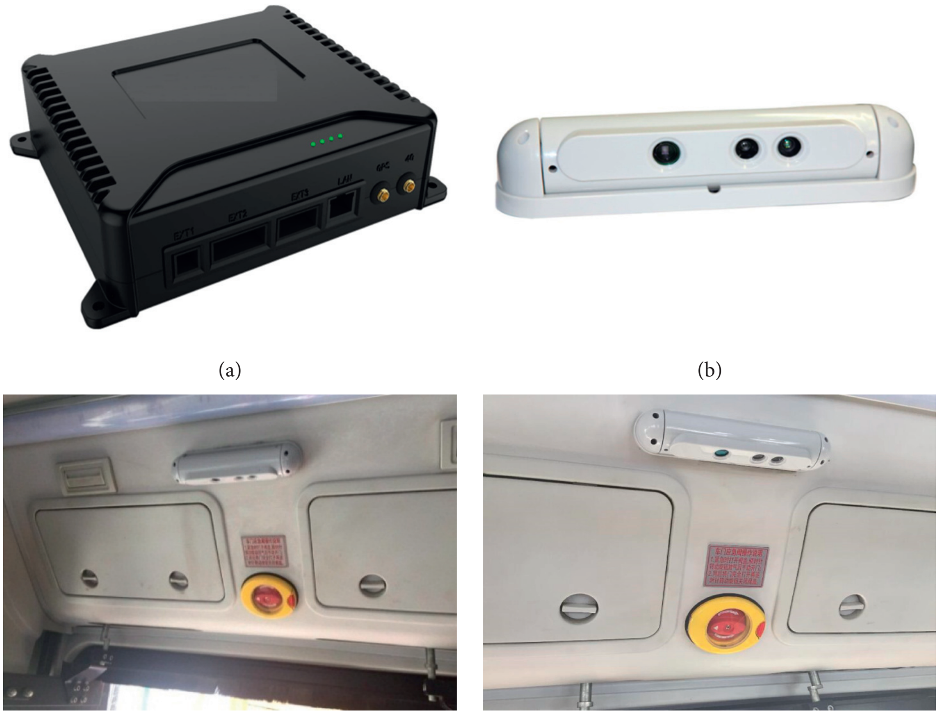
$$\eta_0 = \frac{(T_d - T_c)}{[Cm(1-d)-2]}. \tag{1}$$

The result obtained in equation (1) is actually the average headway of a day, which is the headway of off-peak hours. However, the actual online scheduling arrangement cannot be implemented only by this value, and it also needs to be processed based on the standee density, which means that η_0 will vary with the standee density. If the standee density is high during peak hours, the appropriate reduction should be made to η_0, shortening the headways. Conversely, η_0 should be appropriately increased to reduce the frequency of scheduling in trough hours.

The problem is that the standee density of each area shows unevenness, and a key indicator is needed to determine online headway according to the density of the most crowded areas. Actually, the areas on the bus floor designated for passenger seating and standing are limited. Standing in a spacious area is an instinct response of passengers. The wheelbase area is spacious enough and often more crowded than the other areas. The standing area is divided into four areas with different densities of standees [32]. According to the manual survey, Figure 2 shows the division of the standing area.

Defining the interstop as the bus line between every two stops, $i\,(i = 1, 2, \ldots, n)$ refers to the interstops, and $j\,(j = 1, 2, 3, 4)$ refers to the designated areas. The premise of the standee density algorithm is the number of standees present in the interstop, which is described as follows:

$$\begin{aligned}
Q_i &= \sum_{i=1}^{n} (Q_{ui} - Q_{di}) - (Q_a - 1),\\
Q_i &= \sum_{j=1}^{4} Q_{ij},\\
S &= \sum_{j=1}^{4} S_j,\\
\rho_{ij} &= \begin{cases} \dfrac{Q_{ij}}{S_j}, & \text{if } Q_{ij} > S_j,\\[2ex] 0, & \text{if } Q_i \le S_j, \end{cases}
\end{aligned} \tag{2}$$



(a)

(b)

(c)

(d)

Figure 1: Passenger flow data collection system. (a) Passenger flow analyzer. (b) Binocular camera sensor. (c) Front door installation position. (d) Rear door installation position.

Table 1: Statistical results of each survey period.

Survey period	Data collected by the instrument (pax)		Data collected by manual validation (pax)		Relative accuracy (%)
	Number of get-on passengers	Number of get-off passengers	Number of get-on passengers	Number of get-off passengers	
07:30–09:00	653	649	647	647	99.1
10:00–11:30	474	474	474	474	100.0
13:00–14:30	501	498	501	501	99.4
18:0–19:30	729	726	720	720	98.8

Table 2: Description of each standing area.

Area ID	Description of each area	Area reversibility
1	From the front door to the end of the horizontal seat	Irreversible
2	From the end of the horizontal seat to the rear door	Irreversible
3	From the beginning of the rear door to the end of it	Irreversible
4	From the end of the rear door to the end of the rear aisle	Reversible

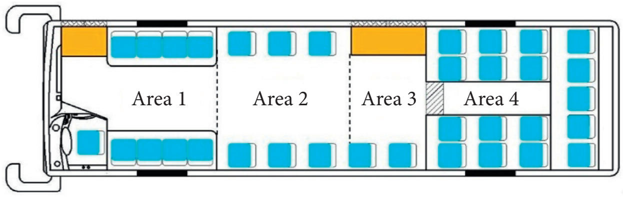


Figure 2: Position of each standing area.

where Q_i is the number of standees; Q_{ij} is the number of standees in area j; $\rho_{ij}$is the standee density in area j; Q_{ui} and Q_{di} are the number of passengers getting on and getting off the bus, respectively; Q_a is the seating capacity of the bus, including one driver's seat; S is the total standing area supplied; and S_j is the j area supplied, which comes from field measurement.

4.2. A Suitable Standee Density for Scheduling. Standing areas are suitable as the key indicator for scheduling buses needs to be defined. According to the manual survey, changes in ρ_{ij} were asynchronous. Figure 3 reveals the propensity of passengers to choose each standing area with a gradual increase in Q_i.

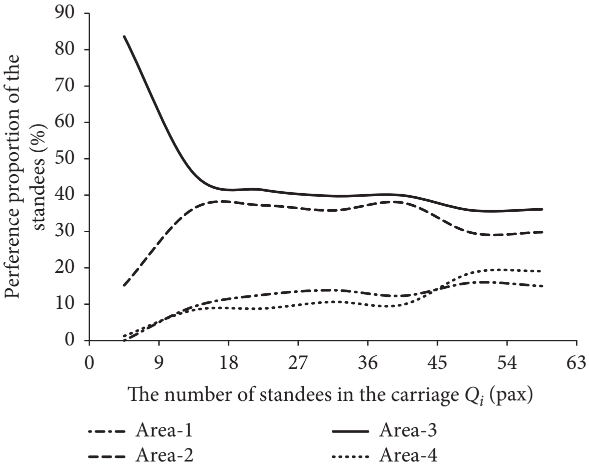


Figure 3: Changes in the proportion of passengers choosing each area.

Q_i did not exceed 45; the proportion of passengers choosing to stand in Areas 1 and 4 was basically maintained within 20%. ρ_{i1} and ρ_{i4} both had a tendency to slow down with the increase in Q_i. This result indicated that if less standing space was available, passengers would seek another area to stand. When Q_i did not exceed 18, Area 2 diverted the passenger flow of Area 3. When it was more than 45, the flow tended to divert to Areas 3 and 4. Therefore, Areas 2 and 3 were critical positions and both were in the wheelbase area. Therefore,Q_i could not give the true expression of the crowdedness degree of the truly crowded area of the carriage, as Areas 1 and 4 were disruptive factors. Therefore, factors unrelated to the standee density in the wheelbase area (SDWA) were excluded.

According to the manual survey, the discrete values of surveys ρ_{i2} and ρ_{i3} were positively correlated with Q_i, but the growth trends varied in different $Q_i$ranges. After nonlinear regression, the change rule of the discrete values of surveys ρ_{i2} and ρ_{i3} can be described as follows:

$$\rho_{i2} = \begin{cases} 0, & 0 < Q_i \le 1.8, \\ 0.13Q_i - 0.20, & 1.8 < Q_i \le 9, \\ 1.86 \ln Q_i - 3.11, & 9 < Q_i \le 45, \\ 2.03e^{0.016Q_i} - 0.03, & Q_i > 45, \end{cases}$$

$$\rho_{i3} = \begin{cases} 0.20Q_i + 0.38, & 0 < Q_i \le 9, \\ 1.36e^{0.034Q_i} + 0.34, & 9 < Q_i \le 45, \\ 7.42 \ln Q_i - 21.63, & Q_i > 45. \end{cases} \tag{3}$$

The *Pearson* correlation coefficient was introduced to test the goodness of fit of ρ_{i2} and $\rho_{i3}$with the discrete value ρ_{sj} of the corresponding area. The number of samples was N, and the correlation between ρ_{ij} and ρ_{sj} was expressed by the product difference correlation coefficient R_j as follows:

$$R_j = \frac{N \sum \rho_{ij}\rho_{sj} - \sum \rho_{ij} \sum \rho_{sj}}{\sqrt{N \sum \rho_{ij}^2 - \left(\sum \rho_{ij}\right)^2} \cdot \sqrt{N \sum \rho_{sj}^2 - \left(\sum \rho_{sj}\right)^2}}. \tag{4}$$

In the grade correlation coefficient R_j level, the closer the absolute value of R_j to 1.0, the greater the correlation [33]. When $Q_i \in (0,9]$, $Q_i \in (9,45]$, and $Q_i \in (45,+\infty)$, the goodness-of-fit values tested were 0.997, 0.905, and 0.951 in Area 2 and 0.996, 0.996, and 0.970 in Area 3, indicating that the correlation between ρ_{i2} and $\rho_{i3}$with the discrete values of the survey was extremely good, and the goodness-of-fit test results were significant.

ρ_{i2} and ρ_{i3} were not reciprocal independent indicators, and neither could fully reflect the true level of the SDWA. Although a weighted algorithm could be introduced to synthesize ρ_{i2} and ρ_{i3}, both of them were segmented functions of different distribution types. The difference between the SDWA values calculated with ρ_{i2} and ρ_{i3}conformed to the trend of the logarithmic curve when $Q_i \in (9,45]$, and the SDWA algorithm was continuous at the turning points. As a result, based on the two indicators, the SDWA indicator was established as follows:

$$\rho_{i0} = \frac{(\alpha\rho_{i2} + \beta\rho_{i3})}{(\alpha + \beta)} + \gamma \ln\left(\frac{Q_i}{9}\right), \tag{5}$$

where α and β are the weight coefficients of ρ_{i2} andρ_{i3} accordingly, and $\alpha + \beta = 1.0$; γ is the allocation factor of Q_i. If $\gamma = 0$, according to the definition of the continuity of the segmented function, considering $Q_i = 9$ and $Q_i = 45$ in the first derivative of equation (5), two weight coefficients were obtained, with $\alpha_1 = 0.68, \beta_1 = 0.32, \alpha_2 = 0.72$, and$\beta_2 = 0.28$. However, Q_i calculated with the two weight coefficients was not continuous when $Q_i = 45$. As the difference between ρ_{i2} and ρ_{i3} conformed to the logarithmic curve, when $Q_i \in (9,+\infty)$, $\gamma \ne 0$. Moreover, when the first derivative of ρ_{i0} was continuous at $Q_i = 45$, $\gamma = 0.08$ could be obtained. In summary, ρ_{i0} could be expressed as follows:

$$\rho_{i0} = \begin{cases} 0.16Q_i - 0.02, & 0 < Q_i \le 9, \\ 0.43e^{0.034Q_i} + 1.26 \ln Q_i - 1.94, & 9 < Q_i \le 45, \\ 1.46e^{0.016Q_i} + 2.14 \ln Q_i - 6.34, & Q_i > 45. \end{cases} \tag{6}$$

Figure 4 shows the plotted curve of ρ_{i0} to test and compare the numerical stability of ρ_{i0} and the sensitivity to changes in passenger flow.

When $Q_i \in (0, 45]$, let $\rho_{i0} = Q_i/9$, and Q_i could be 39. Moreover, $\rho_{i0} > Q_i/9$ in the range of $Q_i \in (0, 39)$. In the range of $Q_i \in (0, 45]$, due to the influence of Areas 1 and 4 on passenger flow, the larger the value of ρ_{i0} is, the more crowded the areas. In the range of $Q_i \in (0, 15]$, the first derivative greater than 0 indicated that the index had great volatility. When $Q_i \le 15$, passengers were basically free to select positions and had a higher propensity for Areas 2 and 3. In summary, the judgment performance of ρ_{i0} as an indicator was better than that of Q_i.

4.3. Judgment Logic of the Status on the Bus Line. As the transit capacity and quality of service manual mentions, it is suitable for a new public transport system to define the peak hours by a passenger density of 2 pax/m^2 in America. However, the code for the design of metros in China recommends that the passenger crowding density should be within 5 pax/m^2, and the proportion of interstops (referring to the section between every two bus stops) with a passenger crowding density exceeding 5 pax/m^2 should be controlled within 20% of the total based on ergonomics [3]. As a result, a statistical indicator was introduced.

The proportions of the number of interstops λ_k falling into $\rho_{i0} \le 1$, $1 < \rho_{i0} \le 5$, and $\rho_{i0} > 5$ to the total number of interstops were taken as the statistical indicator to define the peak hour, off-peak hours, and tough hours online and avoid personal judgment subjectivity:

$$\lambda_k = \frac{\sum_{i=1}^{n} a_{ik}}{n-1}, \qquad a_{ik} = \begin{cases} 1, & \rho_{i0} \ge 5, \\ 0, & \rho_{i0} < 5 \end{cases}, \quad k = 1, 2, 3, \tag{7}$$

where $k = 1, 2, 3$ refers to$\rho_{i0} \le 1$, $1 < \rho_{i0} \le 5$, and $\rho_{i0} > 5$ accordingly; a_{ik} is the number of interstops based on k.

Based on the 79 round-trip passenger flow data points of the manual survey with large passenger traffic during off-peak hours in Xi'an, which were surveyed, the proportion of the number of stops classified was based on ρ_{i0} (Table 3).

When $\rho_{i0} \in (1, 5]$, the corresponding proportion of interstops λ_2 fluctuated approximately 50%, indicating the index properties of the SDWA in the off-peak hours. $\lambda_2 =$ 50% was set as the state judgment threshold. The real-time judgment logic of the passenger flow data collection system was proposed (Table 4).

Importantly, the threshold $\lambda_2 = 50\%$ is the reference value; it is necessary to determine the specific conditions of the bus lines. The aforementioned judgment result only applies to the bus executing its task during the operation period but does not indicate an increase or decrease in the extent of η_0. Hence, it is also necessary to determine the online headways for the trough and peak hours.

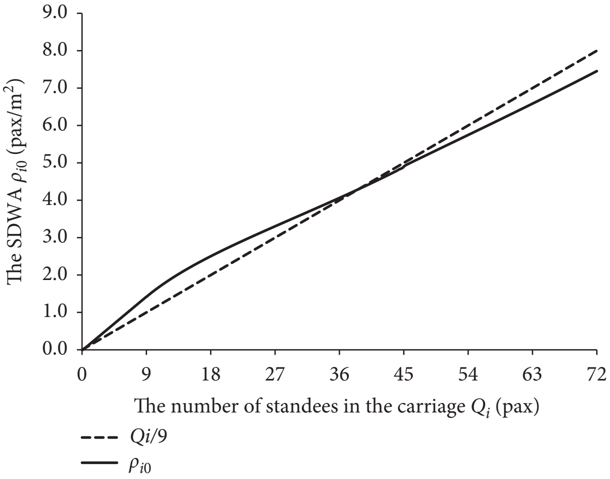


Figure 4: Variation trend of ρ_{i0} with Q_i.

4.4. Online Headways Based on the Standee Density. According to the judgment logic of λ_k and η_0, the division of peak, off-peak, and trough hours was performed based on the key threshold of the proportion of interstops.

As $\lambda_1 + \lambda_2 + \lambda_3 = 100\%$, interstops of proportion in trough hours $\lambda_1 = 30\%$ could be derived. However, the calculated η_0, η_{peak}, and η_{trough} were not integers. Hence, using the rounding function INT(η) to integrate the original noninteger headway, the headways during the trough and peak hours were obtained. To achieve better passenger flow dissipation effects during peak hours and higher transportation efficiency during off-peak and trough hours, the rounding function was appropriately decreased for η_{peak} and increased for η_0 and η_{trough}.

During the peak hours, the headway η_{peak} was shortened, resulting in $\eta_{\text{peak}} < \eta_0$, but η_{peak} could not be shortened without limit. Equation (1) shows that when all buses were put into operation $(d = 0)$, that is, no repaired or rested buses were present, the minimum value was taken. Hence, the online headway during peak hours η_{peak} was as follows:

$$\mathrm{INT}\left(\eta_{\text{peak}}\right) = \begin{cases} \dfrac{0.2\eta_0}{\lambda_3}, & \eta_{\text{peak}} > \eta_{\min}, \\ \eta_{\min} = \dfrac{T_d - T_c}{Cm - 2}, & \eta_{\text{peak}} \le \eta_{\min}, \end{cases} \tag{8}$$

$$\mathrm{INT}\left(\eta_0\right) = \eta_0 + X_0, \tag{9}$$

$$\mathrm{INT}\left(\eta_{\text{peak}}\right) = \eta_{\text{peak}} - X_{\text{peak}}, \tag{10}$$

where $\eta_{\min}$ is the minimum headway and X_0 and X_{peak} are the rounded decimal places of η_0 and η_{peak}, respectively, with values greater than zero.

The headway was prolonged during trough hours, resulting in $\eta_{\min} > \eta_0$, but to meet the passengers' travel requirements and bus operation tasks, η_{trough} could not be extended without limit. Equation (1) shows that when only 70% of the buses were put into operation $(d = 30\%)$, that is, repaired and rested buses accounted for 30% of the number

Table 3: Proportion of interstops classified during off-peak hours.

Off-peak hours	Proportion of stops λ_1 when $\rho_{i0} \leq 1$ (%)	Proportion of stops λ_2 when $1 < \rho_{i0} \leq 5$ (%)	Proportion of stops λ_3 when $\rho_{i0} > 5$ (%)
9:00–11:30 a.m.	51.87	43.90	4.23
11:00 a.m.–1:00 p.m.	52.67	47.33	0.00
1:00–3:00 p.m.	42.34	57.66	0.00
8:00–10:00 p.m.	39.89	52.24	7.87

Table 4: The real-time judgment logic.

Operational period	Judgment logic	Headway control
Peak hours	$\lambda_3 \geq 20\%$	Shorten η_0
Off-hours	$\lambda_3 < 20\%$ and $\lambda_2 \geq 50\%$	Continue η_0
Trough hours	$\lambda_3 < 20\%$ and $\lambda_2 < 50\%$	Increase η_0

of buses in the line, the maximum value was taken. Hence, the online headway during the trough hours η_{trough} was as follows:

$$\text{INT}\left(\eta_{\text{trough}}\right) = \begin{cases} \dfrac{\lambda_1 \eta_0}{0.3}, & \eta_{\text{trough}} < \eta_{\max}, \\ \\ \eta_{\max} = \dfrac{T_d - T_c}{Cm(1-d) - 2}, & \eta_{\text{trough}} \geq \eta_{\max}, \end{cases} \quad (11)$$

$$\text{INT}\left(\eta_0\right) = \eta_0 - X_0, \quad (12)$$

$$\text{INT}\left(\eta_{\text{trough}}\right) = \eta_{\text{trough}} + X_{\text{trough}}, \quad (13)$$

where $\eta_{\max}$ is the maximum headway and X_{trough} is the rounded decimal place of η_{trough}, with a value greater than zero.

5. Numerical Example and Sensitivity Analysis

5.1. Field Validation. A bus line with the largest passenger flow in Xi'an city was considered as an example. It was used to verify the feasibility of the headway model and measure the parameter range of the headway model [34]. The time of the first and last buses on the line was 6:00 a.m. and 12:00 p.m.; the number of buses available on the line was 20; and the number of bus interstops was 24 in total. The bus had 37 seats, and S was 8.96 m^2 [35, 36]. Moreover, 85% of the buses operated during off-peak hours. The passenger flow data were collected during the peak hours of a working day (Table 5). The calculation result of the SDWA was obtained from equation (6).

When Q_i increased from 0 to 36, the maximum deviation rate of ρ_{i0} relative to Q_i/S was 40.52%, which was due to the passengers diverting from Areas 1 and 4 to Areas 2 and 3. However, at this time, Q_i increased slowly and was not sensitive enough to the standee flow. When Q_i exceeded 36, the maximum deviation rate was –6.14% because Areas 1 and 4 diverted the passenger flow from Areas 2 and 3, alleviating the crowdedness of the SDWA. ρ_{i0} embodied the SDWA after the passenger flows of Areas 2 and 3 were diverted so that ρ_{i0} was slightly lower than Q_i/S. Therefore, it was more desirable to reflect the crowdedness of standees by using ρ_{i0} rather than Q_i.

The scheduling time length was 1080 min. As the number of buses during off-peak hours was 85% of the total number of buses, η_0 calculated using equation (1) was 10.8 min. Therefore, the headway in the off-peak hours was 11 min. According to the values (Table 5), λ_1, λ_2, and λ_3 were calculated according to equation (5) to be 12.5%, 50.0%, and 37.5%, respectively. As λ_3 was the preferred judgment index and exceeded 20.0%, it was determined to be the peak hours of passenger flow, and the judgment conclusion was consistent with the judgment logic. At this time, the bus line dispatched all the buses into operation, and the minimum headway of the evening peak hours was 9 min.

5.2. Value Analysis. Xi'an public transport enterprises have clear regulations on the daily running tasks of buses and the number of buses available for scheduling. Each bus runs six round trips per day, which is completed by two drivers. The number of buses available for scheduling should be maintained at more than 70% of the total buses. According to the aforementioned provisions, the minimum and maximum headways of buses can be calculated.

To avoid the phenomenon of dispatch overload on the bus line, that is, when $d = 0$, according to the data provided by the bus line, the lower limit $\eta_{\min}$ of η_{peak} was calculated to be 9 min. Let $\text{INT}\left(\eta_{\text{peak}}\right) = \text{INT}\left(\eta_{\min}\right)$. As the number of buses during off-peak hours was 85% of the total number of buses, it was obtained by equation (8).

When λ_3 approached 25% from 20%, that is, approached the scheduling load in the peak hours, the headway was scheduled by $\text{INT}\left(0.2\eta_0/\lambda_3\right)$.

When λ_3 exceeded 25%, the headway was still scheduled at 9 min to reach the bus scheduling load, and all buses ran on the line.

To meet the basic passenger's travel requirements and bus operation tasks, that is, when $d = 30\%$, according to the data provided by the bus line, the upper limit $\eta_{\max}$ of η_{trough} was calculated to be 14 min. Let $\text{INT}\left(\eta_{\text{trough}}\right) = \text{INT}\left(\eta_{\max}\right)$; as the number of buses in the off-peak period was 85% of the total number of buses, it was obtained by equation (11).

When λ_1 approached 30% from 42%, that is, approached the dispatching load in the peak hours, the headway was scheduled by $\text{INT}\left(\lambda_1\eta_0/0.3\right)$.

When λ_1 exceeded 42%, the headway was scheduled to be 14 min according to $\text{INT}\left(\eta_{\max}\right)$.

Table 5: Passenger flow data collected during peak hours.

Bus stop ID	Number of get-on passengers (pax)	Number of get-off passengers (pax)	Number of passengers on board (pax)	Number of standees (pax)	SDWA, ρ_{i0} (pax/m^2)
1	34	—	34	0	0.00
2	16	0	50	13	2.03
3	18	0	68	31	3.74
4	19	0	87	50	5.56
5	9	5	91	54	5.95
6	12	10	93	56	6.14
7	0	3	90	53	5.85
8	2	1	91	54	5.95
9	13	5	99	62	6.73
10	13	14	98	61	6.63
11	3	19	82	45	5.08
12	2	4	80	43	5.08
13	1	2	79	42	4.76
14	0	2	77	40	4.56
15	1	3	75	38	4.37
16	0	0	75	38	4.37
17	0	3	72	35	4.10
18	4	8	68	31	3.74
19	2	13	57	20	2.77
20	3	4	56	19	2.67
21	5	14	47	10	1.63
22	2	4	45	8	1.31
23	0	4	41	4	0.65
24	—	41	0	0	0.00

The aforementioned calculations assumed the number of buses available on the line to be 20. If the number of buses allocated to the line could be increased on this basis, the value taking a range of λ_k was eased.

By examining the real-time passenger flow data obtained from the collector, Q_i of each interstop was given in real time. Taking Q_i as a dependent variable, a more sensitive indicator, ρ_{i0}, was obtained. By the end of each bus in operation, the proportion of the number of interstops was calculated using the model. Then, the proportion of interstops was used to determine the headway of the bus being set out. The headway model was simple, and it was easy to realize the automatic arrangement of the headway. The model was based on the standee density algorithm, which was more suitable for bus lines with variations in passenger flow.

6. Conclusions

The present study proposed the SDWA for determining the online headway; additionally, the feasibility of the method was verified by numerical examples.

First, after discussing the unevenness of the standee density on the bus floor, the SDWA was capable of sensitively reflecting the variation in standee flow. Passengers were more likely to choose the wheelbase area for standing if no seat was available. If the number of standees did not exceed 5S, the proportion of passengers who chose to stand in the wheelbase area surely exceeded 80%.

Second, the interstop proportion based on the SDWA exceeding 5 pax/m^2 should be given priority. Taking the proportions of interstops as the evaluation criterion for determining the headway of the bus being set out was surely objective and feasible.

Finally, as the arrival time of each passenger at the bus stop was hard to determine, using the proportion of interstops of the former bus to determine the headway of the buses to schedule online enabled the elimination of accidental factors. This method might be of great benefit to the bus headway, passenger evacuation, seat layouts, and emergency security.

Further studies should concentrate on evaluating the matching degree of seat layout and standee density to determine the criteria for guiding bus selection for public transport enterprises.

Data Availability

The data used to support this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work and do not have any commercial or associative interest that represents conflicts of interest in connection with the work submitted.

Acknowledgments

This study was supported by the Natural Science Basic Research Program of Shaanxi (2021JQ-292), Natural Science Foundation of Zhejiang Province (LQ19E080003), the

National Natural Science Foundation of China (Grant 52002282), and the MOE (Ministry of Education in China) Project of Humanities and Social Sciences(21YJCZ790137). And the authors also thank the Transportation Science Institute of Chang'an University and the Key Laboratory for Automotive Transportation Safety Enhancement Technology of the Ministry of Communication PRC.

References

[1] Transportation Research Board, *Capacity and Quality of Service Manual*, Transportation Research Board, Washington DC, USA, 2nd edition, 2003.

[2] American Public Transportation Association, *Standard Bus Procurement Guidelines RFP*, American Public Transportation Association Standards, Washington, DC, USA, 2013.

[3] Ministry of Housing and Urban-Rural Development of the People's Republic of China, *Code for Design of Metro*, China Architecture Publishing & Media Co., Ltd., Beijing, China, 2013.

[4] S. Yan and R. Xiao, "Development of driving cycle of Xi'an bus and CNG consumption verification," *Journal of Chang'an University (Natural Science Edition)*, vol. 35, no. 3, pp. 136–141, 2015.

[5] S. M. Amiripour, A. Ceder, and A. S. Mohaymany, "Hybrid method for bus network design with high seasonal demand variation," *Journal of Transportation Engineering*, vol. 140, no. 6, pp. 1–11, 2014.

[6] B. Alonso, J. L. Moura, A. Ibeas, and F. J. Ruisánchez, "Public transport line assignment model to dual-berth bus stops," *Journal of Transportation Engineering*, vol. 137, no. 12, pp. 953–961, 2012.

[7] A. Tirachini, D. A. Hensher, and J. M. Rose, "Crowding in public transport systems: effects on users, operation and implications for the estimation of demand," *Transportation Research Part A: Policy and Practice*, vol. 53, no. 7, pp. 36–52, 2013.

[8] A. Herbon and Y. Hadas, "Determining optimal frequency and vehicle capacity for public transit routes: a generalized newsvendor model," *Transportation Research Part B: Methodological*, vol. 71, no. 1, pp. 85–99, 2015.

[9] A. de Palma, M. Kilani, and S. Proost, "Discomfort in mass transit and its implication for scheduling and pricing," *Transportation Research Part B: Methodological*, vol. 71, no. 1, pp. 1–18, 2015.

[10] E. Nasibov, A. C. Diker, and E. Nasibov, "A multi-criteria route planning model based on fuzzy preference degrees of stops," *Applied Soft Computing*, vol. 49, no. 12, pp. 13–26, 2016.

[11] S. Jara-Díaz and A. Gschwender, "Towards a general microeconomic model for the operation of public transport," *Transport Reviews*, vol. 23, no. 4, pp. 453–469, 2003.

[12] A. Tirachini, D. A. Hensher, and J. M. Rose, "Multimodal pricing and optimal design of urban public transport: the interplay between traffic congestion and bus crowding," *Transportation Research Part B: Methodological*, vol. 61, no. 3, pp. 33–54, 2014.

[13] L. Haywood and M. Koning, "The distribution of crowding costs in public transport: new evidence from Paris," *Transportation Research Part A: Policy and Practice*, vol. 77, no. 7, pp. 182–201, 2015.

[14] Q. Wu, F. Chen, and Y. Gao, "Computation model of standing-passenger density in urban rail transit carriage," *Journal of Traffic and Transportation Engineering*, vol. 15, no. 4, pp. 101–109, 2015.

[15] M. Kapałka, "Simulation of human behavior in different densities as a part of crowd control systems," *Lecture Notes in Computer Science*, Springer, vol. 9012, no. 3, pp. 202–211, Berlin, Germany, 2015.

[16] D. Katz and L. Garrow, "The impact of bus door crowding on operations and safety," *Journal of Public Transportation*, vol. 15, no. 2, pp. 71–93, 2012.

[17] G. Björklund and J. E. Swärdh, "Estimatiing policy values for in-vehicle comfort and crowding reduction in local public transport," *Transportation Research Part A: Policy and Practice*, vol. 106, no. 12, pp. 453–472, 2017.

[18] S. Yan, J. Cao, and Z. Zhao, "Seating provision and configuration of a 12m city bus considering passenger crowding," *International Journal of Automotive Technology*, vol. 21, no. 5, pp. 1223–1231, 2021.

[19] R. Thoreau, C. Holloway, G. Bansal, K. Gharatya, K. Roan, and N. Tyler, "Train design features affecting boarding and alighting of passengers," *Journal of Advanced Transportation*, vol. 50, no. 8, pp. 2077–2088, 2016.

[20] M. Batarce, J. C. Muñoz, and J. D. D. Ortúzar, "Valuing crowding in public transport: implications for cost-benefit analysis," *Transportation Research Part A: Policy and Practice*, vol. 91, no. 9, pp. 358–378, 2016.

[21] R. Shi, B. Mao, Y. Ding, and L. Liu, "Pedestrain choice behavior analysis of standing position in subway carriage," *Journal of Transportation Systems Engineering and Information Technology*, vol. 17, no. 2, pp. 142–148, 2017.

[22] M. M. Rahman, S. C. Wirasinghe, and L. Kattan, "Users' views on current and future real-timebusinformation systems," *Journal of Advanced Transportation*, vol. 47, no. 3, pp. 336–354, 2013.

[23] M. S. Ghanim and G. Abu-Lebdeh, "Real-time dynamic transit signal priority optimization for coordinated traffic networks using genetic algorithms and artificial neural networks," *Journal of Intelligent Transportation Systems*, vol. 19, no. 4, pp. 327–338, 2015.

[24] V. Nasri, "Design and construction of the second avenue subway project in New York," *Geomechanics and Tunnelling*, vol. 6, no. 5, pp. 528–541, 2013.

[25] Y. Ji, L. Gao, D. Chen, X. Ma, and R. Zhang, "How does a static measure influence passengers' boarding behaviors and bus dwell time?," *Transportation Research Part A: Policy and Practice*, vol. 110, no. 4, pp. 13–25, 2018.

[26] D. Corsar, P. Edwards, J. Nelson, C. Baillie, K. Papangelis, and N. Velaga, "Linking open data and the crowd for real-time passenger information," *Journal of Web Semantics*, vol. 43, no. 3, pp. 18–24, 2017.

[27] B. Marco, M. J. Carlos, and O. J. Dios, "Valuing crowding in public transport: implications for cost-benefit analysis," *Transportation Research Part A.*vol. 91, no. 9, pp. 358–378, 2016.

[28] E. Hans, N. Chiabaut, and L. Leclercq, "Investigating the irregularity of bus routes: highlighting how underlying assumptions of bus models impact the regularity results," *Journal of Advanced Transportation*, vol. 49, no. 3, pp. 358–370, 2015.

[29] A. Tirachini, R. Hurtubia, T. Dekker, and R. A. Daziano, "Estimation of crowding discomfort in public transport: results from Santiago de Chile," *Transportation Research Part A: Policy and Practice*, vol. 103, no. 9, pp. 311–326, 2017.

[30] T. Stasko, B. Levine, and A. Reddy, "Time-expanded network model of train-level subway ridership flows using actual train

movement data," *Transportation Research Record*, vol. 2540, no. 11, pp. 92–101, 2016.

[31] W. Wu, *Highway Transport Operations Research*, China Communications Press Co., Ltd., Beijing, China, 2017.

[32] D. Hörcher, D. J. Graham, and R. J. Anderson, "The economics of seat provision in public transport," *Transportation Research Part E: Logistics and Transportation Review*, vol. 109, no. 1, pp. 277–292, 2018.

[33] J. C. Rayner, W. O. Thas, and D. J. Best, *Smooth Tests of Goodness of Fit: Using R*, Wiley, Hoboken, NJ, USA, 2009.

[34] A. Yamamoto, M. Fukuda, and H. Utsumi, "Vehicle management and travel data analysis of E-bus adopted in JR kesennuma line," *World Electric Vehicle Journal*, vol. 8, no. 1, pp. 122–130, 2005.

[35] A. A. Carlos and C. Sharon, "An observational comparison of the older and younger bus passenger experience in a developing world city," *Ergonomics*, vol. 59, no. 6, pp. 840–850, 2015.

[36] International Union of Public Transport, "A manufacturer's view," *Public Transport International*, vol. 54, no. 3, p. 52, 2005.

Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 5283283, 11 pages
https://doi.org/10.1155/2021/5283283

Research Article

Detecting Invalid Associations between Fare Machines and Metro Stations Using Smart Card Data

Pengfei Zhang,[1] Zhenliang Ma ,[2] and Xiaoxiong Weng[1]

[1] *School of Civil Engineering and Transportation, South China University of Technology, Guangzhou 510000, China*
[2] *Institute of Transport Studies, Department of Civil Engineering, Monash University, Clayton, VIC 3168, Australia*

Correspondence should be addressed to Zhenliang Ma; mike.ma@monash.edu

Received 24 April 2021; Accepted 31 May 2021; Published 11 June 2021

Academic Editor: Inhi Kim

Data quality is essential for its authentic usage in analysis and applications. The large volume of automated collection data inevidently suffers from data quality issues including data missing and invalidity. This paper deals with an invalid data problem in the automated fare collection (AFC) database caused by the erroneous association between the fare machines and metro stations, e.g., a fare machine located at Station A is wrongly associated with Station B in the AFC database. It could lead to inappropriate fare charges in a distance-based fare system and cause analysis bias for planning/operation practice. We propose a tensor decomposition and isolation forest-based approach to detect and correct the invalid associated fare machines in the system. The tensor decomposition extracts features of passenger flows and travel times passing through fare machines. The isolation forest coupled with a neural network (NN) takes these features as inputs to detect the wrongly associated fare machines and infer the correct association stations. Case studies using data from a metro system show that the proposed detection approach achieves over 90% accuracy in detecting the invalid associations for up to 35% invalid associations. The inferred association has a 90% accuracy even when the invalid association ratio reaches 40%. The proposed data-driven invalid data detection method is useful for large-scale data management in terms of data quality check and fix.

1. Introduction

Smart card data collected from the automatic fare collection (AFC) system (i.e., AFC data) enable many beneficial applications in the public transportation system such as collective and individual mobility analysis, system state monitoring, and operation planning and control [1]. The usefulness of these analysis applications is highly dependent on the data quality. The AFC data are collected online and in a large scale that may inevitably encounter data quality issues such as data missing and invalidity.

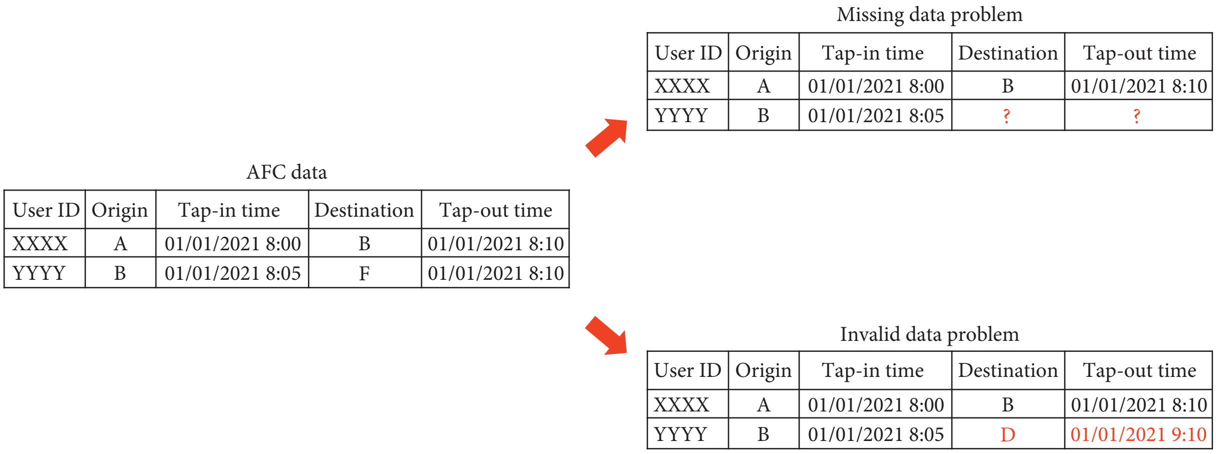
Data problems are prone to happen due to the following reasons:

(i) Human factors: in the AFC system, the transaction records may be missing if passengers fail to tap in/out properly.

(ii) Infrastructure failure: for example, AFC records are triggered when a passenger taps in/taps out through an entrance/exit fare machine. The malfunctioning of fare machines may lead to issues of missing data (machine fails to record or upload transactions) and invalid data (erroneous transactions).

(iii) Inadequate data management. Daily data management for transportation systems is a complex practice. Missing and invalid data may happen in the process of database merging, maintenance, or system update.

Among those data problems, missing and invalid data problems are the most critical and common ones. Figure 1 illustrates the characteristics of these two problems and also their difference. The missing data are cognizable and clearly identifiable via the data structure. For example, some AFC



Missing data problem

User ID	Origin	Tap-in time	Destination	Tap-out time
XXXX	A	01/01/2021 8:00	B	01/01/2021 8:10
YYYY	B	01/01/2021 8:05	?	?

AFC data

User ID	Origin	Tap-in time	Destination	Tap-out time
XXXX	A	01/01/2021 8:00	B	01/01/2021 8:10
YYYY	B	01/01/2021 8:05	F	01/01/2021 8:10

Invalid data problem

User ID	Origin	Tap-in time	Destination	Tap-out time
XXXX	A	01/01/2021 8:00	B	01/01/2021 8:10
YYYY	B	01/01/2021 8:05	D	01/01/2021 9:10

Figure 1: Missing and invalid data problems in the AFC dataset.

transactions may have missing data on tap-out records (empty cells). However, the invalid data are impossible to be directly recognised since the data structure is exactly the same as the valid data. Generally, the invalid data problem can be divided into two categories: data record and association errors. The data record error originates from the facility malfunctioning in the AFC system as mentioned above. The data association error occurs in the process of merging different sources of data (e.g., AFC fare machine records and station dictionaries). The data association error may come from the incomplete information inference and invalid information matching.

The paper deals with the invalid data problem to detect the hidden association errors of the complete and seemingly valid data. Specifically, it aims to detect the invalid association between fare machines and stations in the AFC data. For example, fare machine 001# is located in Metro Station A, but wrongly associated to Station B in the AFC database (Figure 2). The problem is prone to happen as the fare machines are frequently added, replaced, etc. in the metro systems, but the fare machine-station dictionary may fail to update timely. The consequences of invalid associations could be significant, e.g., under/over charging for a large amount of passengers. In addition, it is costly to fix this problem by manpower. One should manually check all the machines in metro stations to rebuild the correct association between fare machines and stations. Especially, it is impossible to manually detect such problem in the historical dataset since the fare machine distribution may not consistent with the current system.

We develop a data-driven approach, based on tensor decomposition and machine learning techniques, to automatically detect such invalid associations using AFC data, and also infer the correct association stations that a fare machine belongs to. The approach works in two steps: the tensor decomposition is utilized to extract the flow volume and travel time patterns of each fare machine. Then, the isolation tree technique and NN models are designed to detect the incorrect linked fare machines and infer their correct association stations based on the extracted features from tensor decomposition.

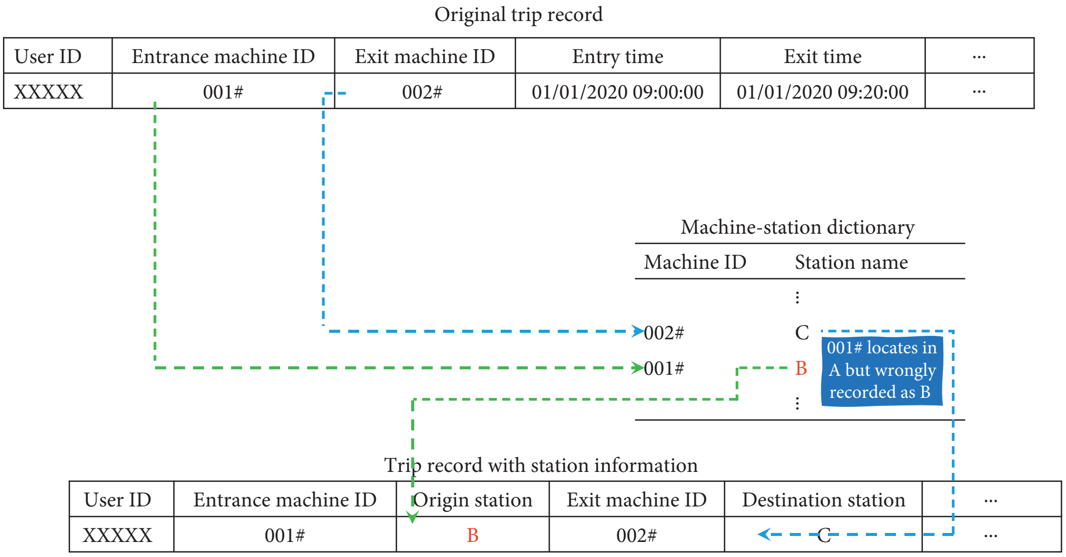
The remaining is organised as follows: Section 2 reviews the relevant studies on data quality issues, including overview of data quality problems, feature extraction techniques, and anomaly detection; the problem formulation and methodology are presented in Section 3; Section 4 reports the case study using the AFC data from a large metro system; the final section concludes the paper and discusses potential further studies and applications.

2. Related Work

2.1. Data Quality Problems. Data quality is one of the most important issues in big data area. Low or bad data quality is costly. For example, it is reported that bad data or poor data quality costs US businesses 600 billion dollars annually [2]. For metro systems, AFC systems collect massive transaction data of metro passengers. The literature has reported plenty of data quality problems related to AFC data. Robinson et al. [3] reported that the reasons of AFC data quality problems can be grouped into 4 categories: (1) software; (2) data; (3) hardware; (4) user. A recurrent information missing problem of the boarding station in Beijing Metro has been reported by Ma et al. [4]. Liu et al. [5] reported a time synchronisation problem of the AFC and AVL system, which causes the recorded boarding time information to be invalid in a large scale. Network, scheduling, fare table, etc. are important data stored in the AFC database. Errors in these data will lead to significant consequences. For example, the London Oyster smart card system crashed on Saturday 12th July 2008 due to erroneous data resulting in over 40,000 Oyster cards having to be replaced [6].

Although many studies deal with missing data in transportation, to the best of our knowledge, there is no study on detecting or fixing the association errors in transportation or other related areas, particularly the fare machine-station invalid association problem.

2.2. Feature Extraction Techniques. The key idea for a data-driven detection approach is to extract the passenger flow or/and travel time patterns between fare machines and stations.



Original trip record

User ID	Entrance machine ID	Exit machine ID	Entry time	Exit time	···
XXXXX	001#	002#	01/01/2020 09:00:00	01/01/2020 09:20:00	···

Machine-station dictionary

Machine ID	Station name
	⋮
002#	C
001#	B
	⋮

Trip record with station information

User ID	Entrance machine ID	Origin station	Exit machine ID	Destination station	···
XXXXX	001#	B	002#	C	···

Figure 2: Invalid association problem between fare machine and metro station.

Feature extraction is one of the most important issues in the machine learning field. Feature extraction reduces the resources required to characterize a large set of data or/and a huge dimensions of input information. Plenty of methods are proposed in the machine learning community dealing with the feature extraction. These methods can be roughly divided into two parts: conventional statistical learning methods and deep learning-based method. Conventional statistical learning methods such as principle component analysis (PCA) [7], Isomap [8], and partial least squares (PLS) regression [9] mainly based on the statistical learning-based algorithms. The advantages of these methods are they are robust to small dataset, i.e., do not need large amount of samples to maintain the performance. However, the disadvantages are also critical. For example, they are not robust to noisy samples, and the feature extraction quality is highly dependent on specific tricks in different tasks, thus which are less generalized. Deep learning-based feature extraction methods become more and more popular recently. Variety forms of neural networks, e.g., convolutional neural network (CNN) [10] and long short-term memory (LSTM) [11] neural network. can be treated as feature extraction models. Different from the statistical learning-based algorithms, they extract the features in a latent, end-to-end manner. The advantage is that the extracted features are more representative and comprehensive. However, these models always require a large dataset in the training procedure; thus, they are not suitable in the few-shot scenario. In conclusion, there is no a generalized feature extraction method for all the tasks. Feature extraction methods should be designed based on the characteristics of the focused problem.

In our problem, passenger flow and travel time patterns are related to multiple modes, e.g., time and location. Tensor is a nature choice to represent and capture these patterns. Tensor is a multidimensional extension of matrix [12]. Tensor has been widely used in transportation area to deal with multidimension data. Tan et al. [13] utilized a tensor decomposition approach to capture the multimode correlations in traffic data and recover missing traffic data by reconstructing the traffic flow tensor. The results show that the proposed algorithm performs well even when the missing ratio is high. Chen et al. [14] proposed singular value decomposition (SVD)-combined tensor decomposition framework to complete the traffic data using traffic speed information. Sun and Axhausen [15] utilize a probabilistic tensor decomposition method to mine the urban mobility patterns. Mobility patterns of different passenger groups (e.g., students, adults, and elders) are explored. In our study, we also use tensor decomposition to extract the flow pattern related to each fare machine.

2.3. Anomaly Detection. The invalid associations (between fare machines and stations) are treated as anomalies. Anomaly detection is an important topic in data mining. The anomaly detection could be roughly divided into three categories, statistical, machine learning, and deep learning models.

(1) Statistical method: statistical methods are the early explorations of the anomaly detection. The methods in this category first make assumptions of the distribution of the studied dataset. The samples with low probabilities are treated as anomalies. Rousseeuw and Driessen [16] proposed an anomaly detection method based on the Gaussian assumption of the data. The performance of statistical anomaly detection methods highly depends on the fitting between the assumption and the reality, thus exhibiting limited performance.

(2) Machine learning-based methods: the most widely used anomaly detection methods are the machine learning-based methods, which generally have two categories: supervised and unsupervised methods. Supervised methods [17, 18] refer to the models applying to the dataset that the training data are

labeled with "nominal" or "anomaly." The models are trained with the labeled data and use to identify new instances. Unsupervised methods deal with the dataset without labels. These methods automatically detect the anomalies based on certain criteria. Popular unsupervised methods include LOF [19], DBSCAN [20], k-means [21], and the isolation forest [22] method.

(3) Deep learning-based methods: the emerging deep learning models bring new opportunities to better solve the anomaly detection problem. Hundman et al. [23] propose an LSTM network-based framework for anomaly detection; [24] utilized a generative adversarial network (GAN) to detect the anomalies in time series data. Nguyen et al. [25] detect the anomalies by constructing the model snapshot and outputting the ensembles of the NN models. Deep learning-based methods tend to have more a promising performance compared to other techniques. However, these methods require a large amount of training data to produce reasonable results. Its performance is low in scenarios with a small set of training data, e.g., the fare machine-station association problem studied in this paper.

3. Methodology

3.1. Problem Formulation. Let m be a fare machine, and $S_m, \widehat{S}_m \in \Delta$ its actual station and current association station in the AFC dataset, respectively, where $\Delta = \{S^1, S^2, \ldots, S^s\}$ contains all the stations in the metro system. Note that different fare machines could share the same station, i.e., located in the same station. If $S = \widehat{S}$, fare machine m is defined as valid association fare machine; if $S \neq \widehat{S}$, fare machine m is defined as invalid association fare machine. The fare machine-station association detection problem is defined as follows.

Given an AFC dataset $\mathbf{D}$ and a set of fare machines Φ recorded in $\mathbf{D}$, detect invalid association fare machines and infer their associated stations for fare machines m in Φ.

Mathematically, the problem is defined as follows:

(i) *Invalid Association Detection.* Find $\phi \subset \Phi$, s.t. $\phi = \left\{m_{S\longrightarrow\widehat{S}} | S \neq \widehat{S}\right\}$, and $\Phi - \phi = \left\{m_{S\longrightarrow\widehat{S}} | S = \widehat{S}\right\}$

(ii) *Station Inference.* For each fare machine m in ϕ, find $\widehat{S}$ s.t. $S = \widehat{S}$

3.2. Fare Machine Features. For convenience, we define the concept of fare machine-related passenger flow (MRF). For an entrance fare machine, MRF refers to the passenger flow tapping in an entrance fare machine of the origin station and tapping out at a destination station (using any machine) during a certain time slot. For an exit fare machine, MRF represents the passenger flow tapping in at an origin station (using any machine) and tapping out at an exit fare machine during a certain time slot. MRF can be characterized using different features, such as flow volume and travel time. Indicators extracted from the MRF features can be used to characterize fare machines. The hypothesis is that MRF features share more similar patterns if the fare machines are located at the same station than at different stations.

The flow volume and travel time are selected to characterize the MRFs of fare machines. These two features reflect system dynamics from both the demand (mobility patterns) and supply (network and operations) points of view as well as their interactions. They provide complementary knowledge and therefore give a more comprehensive view of the MRF patterns. They are defined for entrance and exit fare machines separately:

(i) For entrance fare machines, MRF flow volume measures the number of passengers passing through each fare machine at an origin station and going to a destination station. For exit fare machines, it represents the number of passengers entering the metro system at an origin station and tapping out through an exit fare machine. MRF flow volume reflects the mobility behavior of passengers.

(ii) MRF travel time indicates the average travel time from a fare machine to a destination station for entrance fare machines and from an origin station to a fare machine for exit fare machines. It reflects the supply characteristics of the metro system, e.g., geographical relationship between stations and scheduling, but also demand characteristics of certain stations as it includes time waiting to board a train under capacity constraints.

Figure 3 shows the overview of the proposed framework. It consists of three modules: MRF feature extraction, invalid association detection, and associated station inference:

(i) MRF feature extraction module: it constructs the MRF flow volume and travel time tensors to characterize fare machines and extracts latent MRF flow and travel time features using the tensor decomposition technique.

(ii) Invalid association detection module: it detects the invalid associations (between fare machines and stations) in two steps. The valid and invalid associations are initially detected using the isolation forest method. Then, the invalid associations are reinspected (the feedback arrow) using neural networks (trained with the valid association data).

(iii) Association station inference: it infers the station that a fare machine (detected as invalid association) belongs to using the trained neural networks.

3.3. MRF Tensor Construction. For data representation, tensors are used to characterize the MRF flow volume and travel time. A tensor is a high-order generalization of a matrix. The multiway property of a tensor fits the nature of MRF features. For example, MRF flow volume can be characterized by "machine mode" (M), "time mode" (T), "day mode" (D), and "station mode" (S). For entrance fare machines, "machine mode" denotes the related fare machine ID, "time mode" represents the time interval of a day (e.g., 6:

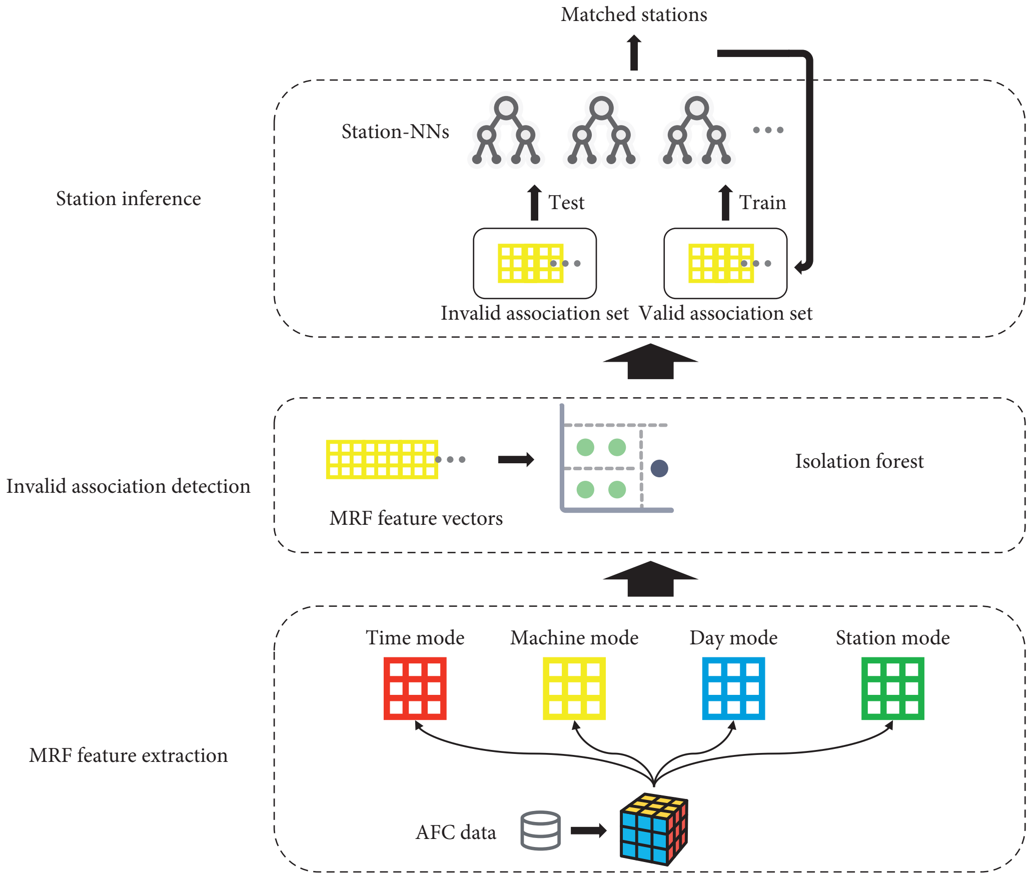


Figure 3: Overview of the proposed framework.

00 to 7:00 AM), "day mode" denotes the date, and "station mode" denotes a destination station ID. For exit fare machines, the definitions of tensor modes are the same with entrance fare machines, except for the "station mode." The "station mode" of an exit fare machine is the origin station ID. In this way, two 4-way tensors are used to represent the MRF flow volume of entrance and exit fare machines, respectively. For example, an entry: 50 at (A, 8:00 to 9:00 AM, January 1, B) of entrance machine tensor represents "the passenger flow volume passing through entrance machine A in the interval 8:00 to 9:00 AM on January 1 and exiting at Station B is 50 passengers." The methodology for fare machine-station association is the same for entrance and exit fare machines. Entrance fare machines are used to illustrate the proposed framework. Unless stated, the "fare machines" and "MRF tensors" refer to entrance fare machines and entrance MRF tensors, respectively.

To construct the MRF flow volume tensor, the mode variables above are transformed into numerical indices:

(i) Machine mode: the fare machines are labeled from 1 to M. Then, the machine IDs belong to a set $M = \{1, 2, \ldots, M\}$, where M represents the total number of fare machines.

(ii) Time mode: the hourly interval is used to represent the tap-in time $T = \{1, 2, \ldots, T\}$. Note that only the operating hours of the metro system are considered, where the i^{th} element in T denotes the i^{th} operating hour of the day.

(iii) Day mode: day mode represents the date, thus $D = \{1, 2, \ldots, D\}$ where 1 and D represent the first and the last day of the studied time span, respectively.

(iv) Station mode: the stations are labeled from 1 to S $S = \{1, 2, \ldots, S\}$, where S denotes the set of stations in the metro system.

The MRF flow volume is represented by a size $M \times T \times D \times S$ tensor $\mathcal{V}$. Figure 4 shows the structure of the MRF flow volume tensor. Each entry of $\mathcal{V}$, denoted as $\mathcal{V}_{mt\ ds}$, represents the MRF flow volume entering through fare machine m and exiting at destination station s during the t^{th} time interval of day d. For the exit fare machines, the tensor construction procedure is the same as the entrance machines. Accordingly, the entry $\mathcal{V}'_{mtds}$ denotes the MRF volume entering though station s and tapping out though fare machine m during the t^{th} time interval of day d.

Similarly, the MRF travel time tensor is denoted as $\mathcal{T} \in \mathbb{R}^{M \times T \times D \times S}$. An entry $\mathcal{T}_{mtds}$ in $\mathcal{T}$ represents the average travel time of all passengers entering through fare machine m and traveling to destination station s during the t^{th} time interval of day d.

The properties of MRF flow volume and travel time tensors are different, though they share the same structure. The difference stems from the tensor cells that have no AFC

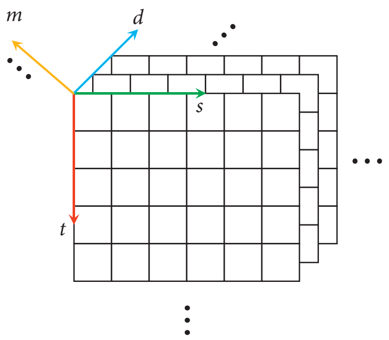


Figure 4: Structure of MRF flow volume tensor. MRF flow volume tensor consists of 4 modes, i.e., time (t) mode, day (d) mode, station (s) mode, and machine (m) mode.

observation. For the MRF flow volume tensor, the value of such cells is 0 since the MRF flow volume for the corresponding $[m, t, d, s]$ is 0 (no passenger flow). However, for the travel time tensor, cells having no observation cannot be directly filled with a zero. No observation in the MRF travel time tensor only means that no passengers traveled for the specified $[m, t, d, s]$ case. However, the corresponding travel time cannot be 0. An initial idea is filling these cells using the average travel time of such OD pairs in the historical data. Unfortunately, nonobservation cells always account for a large ratio of the MRF travel time tensor (e.g., 63.5% in the studied AFC dataset). Therefore, it is hard to estimate a reasonable average travel time for each cell based on limited information. Instead, "NaN" values are used to fill those cells to represent the unknown travel times.

3.4. Tensor Decomposition. Tensor decomposition is used to extract fare machine features from the MRF flow volume and travel time tensors. Given the different properties of these two tensors, different tensor decomposition methods are developed to extract the MRF flow volume and travel time features, respectively.

3.4.1. Tensor Decomposition of MRF Flow Volume. For MRF flow volume tensor $\mathscr{V}$, the CANDECOMP/PARAFAC (CP) decomposition [12] is used to extract the fare machine features. CP decomposition factorizes a tensor into a summation of a series of rank-1 tensors. A rank-1 tensor $\mathscr{V} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_n}$ (I_i is the dimension of mode i) is an outer product of N vectors: $\mathscr{X} = \mathbf{a}^{(1)} \circ \mathbf{a}^{(2)} \circ \cdots \circ \mathbf{a}^{(n)}$, where $\mathscr{X}_{i_1 i_2 \ldots i_n} = a^{(1)}_{i_1} a^{(2)}_{i_2}, \ldots, a^{(n)}_{i_n}$, $\mathbf{a}^{(i)}$ denotes a vector, $a^{(i)}_k$ denotes the k^{th} element of $\mathbf{a}^{(i)}$, and the symbol $\circ$ denotes the outer product of vectors.

The CP decomposition of $\mathscr{V} \in \mathbb{R}^{M \times T \times D \times S}$ can be formulated as follows:

$$\widehat{\mathscr{V}} = \sum_{r=1}^{R} \mathbf{m}^r \circ \mathbf{t}^r \circ \mathbf{d}^r \circ \mathbf{s}^r, \tag{1}$$

where R represents the total number of components, $\mathbf{m}^r \in \mathbb{R}^M$, $\mathbf{t}^r \in \mathbb{R}^T$, $\mathbf{d}^r \in \mathbb{R}^D$, and $\mathbf{s}^r \in \mathbb{R}^S$ represent the component vector of the machine, time, day, and station modes, respectively. Figure 5 illustrates the process of CP decomposition of $\mathscr{V}$.

Computing the CP decomposition of $\mathscr{V}$ can be treated as an optimization problem. The goal is to find a CP decomposition $\widehat{\mathscr{V}} = \sum_{r=1}^{R} \mathbf{m}^r \circ \mathbf{t}^r \circ \mathbf{d}^r \circ \mathbf{s}^r$ with R components that could best approximate $\mathscr{V}$. The decomposition $\mathscr{V}$ is the solution of the following optimization problem, i.e., find

$$\widehat{\mathscr{V}}^* = \underset{\widehat{\mathscr{V}}}{\arg\min} \|\mathscr{V} - \widehat{\mathscr{V}}\|_F, \tag{2}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. This optimization problem can be solved using the alternating least squares (ALS) method [26]. Details of the solution procedure can be found in [12].

The feature matrix $\mathbf{M}_{\mathscr{V}} = [\mathbf{m}^1, \mathbf{m}^2, \ldots, \mathbf{m}^R]$ is constructed utilizing all the component vectors $\mathbf{m}^r$ in $\widehat{\mathscr{V}}^*$. Since each entry in $\widehat{\mathscr{V}}^*$ is calculated as the outer product of all the 4 component vectors, $\mathbf{M}_{\mathscr{V}}$ could be treated as an indicator of the hidden information of all the other 3 modes. The entries in $\widehat{\mathscr{V}}^*$ that are related to the i^{th} fare machine are calculated only using the elements in i^{th} row of $\mathbf{M}_{\mathscr{V}}$. Therefore, each row of $\mathbf{M}_{\mathscr{V}}$ can be used as a latent feature vector to represent each fare machine's MRF flow volume pattern.

3.4.2. Tensor Decomposition of MRF Travel Time. CP decomposition cannot be applied directly to extract travel time features. This is because the travel time tensor has nonnumerical (i.e., NaN) entries, which makes the operation $\mathscr{T} - \widehat{\mathscr{T}}$ infeasible. A variation of CP decomposition, CP Weighted OPTimization (CP-WOPT) [27], is used to deal with the MRF travel time tensor decomposition. CP-WOPT is widely used to recover tensors with missing entries. CP-WOPT utilizes a weight tensor to indicate the location of NaN entries. The formulation is as follows:

$$\widehat{\mathscr{T}}^* = \underset{\widehat{\mathscr{T}}}{\arg\min} \|\mathscr{W} * (\mathscr{T} - \widehat{\mathscr{T}})\|_F. \tag{3}$$

The weight tensor $\mathscr{W} \in \mathbb{R}^{M \times T \times D \times S}$ has the same shape as $\mathscr{T}$ and is defined as

$$\mathscr{W}_{mtds} = \begin{cases} 0, & \text{if } \mathscr{T}_{mtds} \text{ is NaN}, \\ 1, & \text{otherwise}. \end{cases} \tag{4}$$

In the initialization phase, NaN cells are filled with random values. As these values are multiplied by 0 during the optimization, they do not influence the results of the optimization objective (optimal solution). After optimization, $\widehat{\mathscr{T}}^*$ can represent features of the observed travel time data well. As there exists strong relationship between the cells in $\mathscr{T}$, the features of the entries without observations can also be represented in the reconstructed tensor $\widehat{\mathscr{T}}^*$. A feature matrix $\mathbf{M}_{\mathscr{T}}$ is constructed using the machine mode component vectors in $\widehat{\mathscr{T}}^*$ to represent the multimode travel time features of fare machines. Details about the CP-WOPT method can be found in [27].

The MRF flow volume and travel time feature vectors of each fare machine are concatenated into one single vector to characterize the corresponding fare machine.

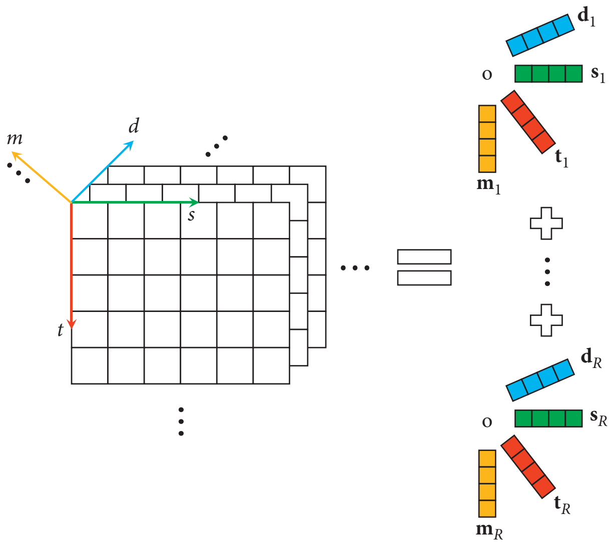


Figure 5: CP decomposition of the MRF flow volume tensor.

3.5. Fare Machine-Station Association. As fare machines at the same station share similar surrounding Point of Interests (POIs), the MRF features of these fare machines tend to be similar. Therefore, we should first extract the MRF feature of each station. Then, the MRF feature of each machine is compared to the station MRF feature. If a fare machine has a similar MRF feature with a station, then this station is likely to be the association station of the fare machine. We divide the inference process into two successive problems P1 and P2.

3.5.1. P1: Invalid Association Fare Machine Detection. To solve P1, we first give two assumptions: (1) the MRF features of the invalid associations are anomalies to their recorded stations. More formally, let $C(\cdot)$ be the count function, anomaly means $C(m_{S_1 \longrightarrow S_2}) \ll C(m_{S_2 \longrightarrow S_2})$ for $\forall S_1 \in \Phi, S_1 \neq S_2$. This indicates that the number of fare machines with association station S_1 but recorded station S_2 should be far less than the number of valid association fare machines in S_2. Note that this assumption does not mean the total number of invalid association fare machines of S_2 is less than the valid fare machines. We only restrict that fare machines recorded as S_2 but actually associated with S_1 should be minority to S_2. This assumption is reasonable since the error leads to fare machine-station invalid association tends to be random; for example, it is unlikely to have many fare machines located in the same station wrongly recorded as another station simultaneously. (2) The invalid associations happen randomly. This assumption indicates that for a fare machine m in station S, it experiences equal probability being wrongly associated to all the other stations in the system. This assumption is reasonable since the invalid associations mainly because of the inadequate data management in the process of database merging, maintenance, or system update.

Based on this assumption, the isolation forest method is adopted to solve P1. The isolation forest model is an unsupervised model for anomaly detection, which could be directly used for the contaminated dataset. The only requirement of this method is that the outlier should be few and different with the normal instances. This exactly fits the aforementioned assumption. The isolation forest detects the outliers using a special measurement: partitions. The isolation forest "isolates" observations by randomly selecting a dimension of the MRF feature vector and then randomly splitting the space between the maximum and minimum values of the selected dimension. Since recursive partitioning can be represented by a tree structure, the number of splittings required to isolate an MRF feature is equivalent to the path length from the root node to the terminating node. This path length, averaged over a forest of such random trees, is a measure of normality. Random partitioning produces noticeably shorter paths for anomalies. Hence, when a forest of random trees collectively produce shorter path lengths for particular fare machines, they are highly likely to be anomalies [22].

Based on the results from the isolation forest, we can divide the fare machine MRF feature vectors into two parts:

$\mathbf{F}_\phi$ contains all the MRF feature vectors that are inferred as invalid (i.e., abnormal) by the isolation forest

$\widehat{\mathbf{F}}_\phi$ contains all the MRF feature vectors that inferred as valid (i.e., normal) by the isolation forest

The fare machines with their MRF features in $\widehat{\mathbf{F}}_\phi$ are detected as valid, while the fare machines in $\mathbf{F}_\phi$ are reinspected in the process of solving P2.

3.5.2. P2: Association Station Inference. In P2, a reinspection of the fare machines in $\mathbf{F}_\phi$ is conducted to refine the detection results from P1. The reinspection detects which associations are wrongly detected as invalid in $\mathbf{F}_\phi$. In practical applications, the inference provides a certain sense about the data quality in their AFC database. The model outputs the potential association stations of the detected invalid association fare machines, which facilities effective field investigation and reduces manpower.

Neural network (NN) is used to model the station MRF feature using the MRF features in $\widehat{\mathbf{F}}_\phi$ (detected as valid). As the number of samples (i.e., fare machines) are limited (e.g., 2000 fare machines in the studied network), the NN training may face underfitting issues. We built one shallow neural network for each station, which denotes as the station-NN. For a certain station-NN $\mathcal{N}_i$ of station S_i, we label the fare machines with the recorded station S_i in $\widehat{\mathbf{F}}_\phi$ as 1 and label other fare machines in $\widehat{\mathbf{F}}_\phi$ as 0. It is inadequate to directly train the station-NN with the labeled features. Since a metro system has many stations (e.g., 90 stations in the studied metro system), for one certain station, the number of positive samples (i.e., MRF features labeled as 1) is much less than the negative samples (MRF features labeled as 0), which will lead to the learning bias. We utilize the adaptive synthetic sampling (ADASYN) [28] approach to oversample the positive samples, ensuring that the number of the oversampled positive samples is similar with the number of negative ones. $\mathcal{N}_i$ is then trained with the oversampled MRF features and their corresponding labels. After $\mathcal{N}_i$ is well-trained, the output of the network will be the probability that the input fare machine MRF feature belongs to this station.

For an MRF feature $\mathbf{v}$ in $\mathbf{F}_\phi$, we input it into all the well-trained station-NNs. Let $\mathbf{P} = [P_1, P_2, \ldots, P_S]$ denote the output probability from each station-NN, and $\mathbf{P}_\pi = [P_{\pi(1)}, P_{\pi(2)}, \ldots, P_{\pi(S)}]$ is the descend order permutation of $\mathbf{P}$, where $P_{\pi(i)} > P_{\pi(j)}$, given $i < j$. The top-k stations $\mathcal{K} = \{\pi(1), \pi(2), \ldots, \pi(k)\}$ would be the most possible association station of the corresponding fare machine of $\mathbf{v}$.

Using $\mathcal{K}$, the reinspection for P1 is conducted for the fare machine in $\mathbf{F}_\phi$ with the following rule: given a fare machine $m_{S\longrightarrow s} \in \mathbf{F}_\phi$, if $s \notin \mathcal{K}$, m is inferred as invalid, otherwise as valid. For the fare machines inferred as invalid after the reinspection, the top-k station $\mathcal{K}$ is treated as the potential association stations set. In the implementations, one can first check the stations in this set to find if this fare machine is there.

4. Case Study

We utilize AFC data from an urban metro system to evaluate the proposed detection and inference approach. The data cover 7 days from January 15 to 21 in 2018. The fare machine-station association information is carefully checked to ensure its validity for benchmarks. Figure 6 illustrates the statistic of the number of machines in the metro system during the studied time span.

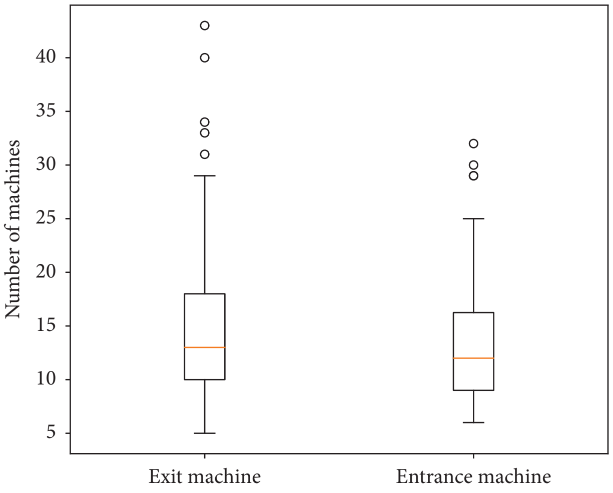


FIGURE 6: Number of entrance and exit fare machines in the studied metro system.

4.1. Experimental Setup. We randomly select 1000 entrance fare machines and 1000 exit fare machines and collect the corresponding AFC transaction records to construct the experimental dataset. We randomly choose a set of fare machines and modify their associated stations (invalid associations). The proposed approach is validated with the ratio of invalid associated fare machines ranging from 5% to 40%. The approach runs 20 times per scenario to avoid random errors. Table 1 summarizes the model parameters used in the experiments.

4.2. Performance Evaluation. Table 2 shows the tabularised relations between truth/falseness of the detection and valid/invalid association.

A set of performance metrics is used to comprehensively evaluate the model performance, including accuracy (Accu), true positive rate (TPR), and false positive rate (FPR):

$$\text{Accu} = \frac{N_T}{N_{\text{PN}}}, \tag{5}$$

where N_{PN} is the total number of associations (or fare machines) and N_T the number of correctly detected associations (between fare machines and stations). The correctly detected fare machines include cases that are truly positive and negative:

$$\text{TPR} = \frac{N_{\text{TP}}}{N_P}, \tag{6}$$

where N_{TP} is the number of truthfully detected invalid association (correctly inferred an invalid association as invalid), and N_P is the number of invalid associations. TPR measures the model's sensitivity towards invalid associations:

$$\text{FPR} = \frac{N_{\text{FP}}}{N_N}, \tag{7}$$

where N_{FP} is the number of falsely detected valid associations (falsely inferred a valid association as invalid) and N_N is the total number of valid associations. FPR measures the misjudgement rate of the valid associations.

4.2.1. Evaluation of Invalid Association Detection (P1). Figure 7 shows the detection results of associations with the invalid association ratio ranging from 5% to 40%. The results indicate that the isolation forest model is robust to the invalid associations when the invalid association ratio is less than 20% (the detection accuracy is over 96%). It can still achieve a detection accuracy of 87%, and even 40% of the fare machines are wrongly associated with stations in the data. The TPR is an essential characteristic of the detection of invalid associations in P1, since there is no reinspection of the invalid associations in $\widehat{\mathbf{F}}_\phi$ in the following procedures of the approach. That is, the wrongly associated fare machines in $\widehat{\mathbf{F}}_\phi$ will remain undetected which may eventually impact practical applications in reality. Also, it is favorable to detect more invalid associations to ensure a clean MRF feature set for each station, which benefits the correction of invalid associations in P2. The TPR is over 90% when the invalid association is less than 20%, which indicates the promising performance of the proposed approach in detecting the invalid associations. The falsely detected valid associations (FPR) is very low (less than 5%), and it decreases with the increase of the invalid association ratio as expected.

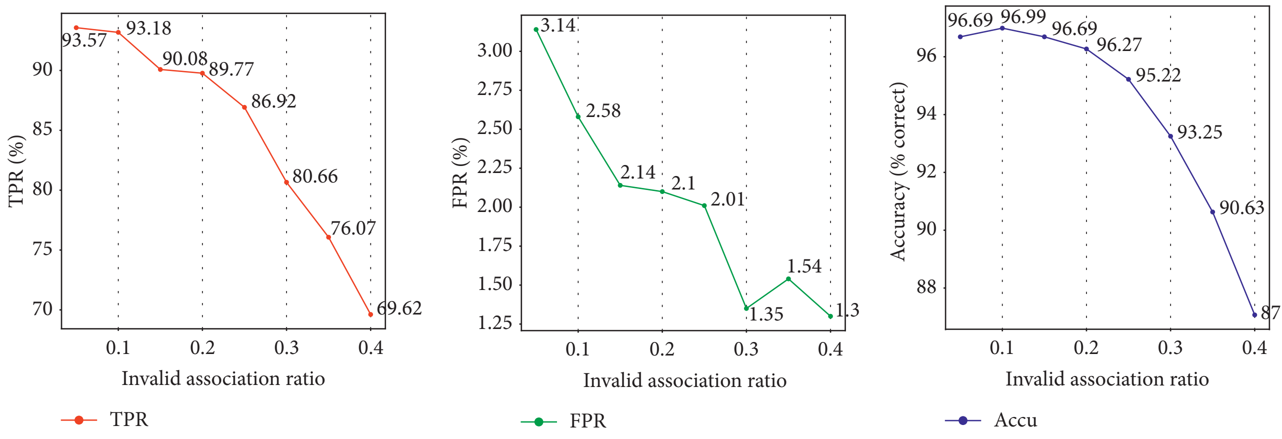
4.2.2. Evaluation of Association Inference (P2). For the P2 evaluation (rematching wrongly associated fare machines to stations), we quantify the model's capability to effectively

Table 1: Model parameters.

Tensor decomposition	Optimal value (potential values)
Number of components (R)	8 (1–15)
Optimization algorithm	ALS (ALS refers to the alternating least squares algorithm)
Error tolerance	$1e-6$ ($1e-3$–$1e-8$)
Maximum number of iterations	100 (10, 100, 500, 1000)
Isolation forest	Value
Threshold score (the threshold score is calculated with the *decision_function* in *sklearn.IsolationForest* package under *Python 3.7*)	0
Number of estimators	1000 (200, 500, 1000, 1500)
Station-NN	Value
The number of top stations in P1 reinspection	5
Number of hidden layers	2 (1–5)
Optimizer	Adam (Adam refers to the optimization algorithm proposed in [29])
Number of neurons	(16, 5)

Table 2: Confusion matrix of the valid association detection.

	Invalid association (positive)	Valid association (negative)
True detection (true)	True positive (TP)	True negative (TN)
False detection (false)	False positive (FP)	False negative (FN)



(a)

(b)

(c)

Figure 7: Model detection performance (a) TPR, (b) FPR, and (c) accuracy with invalid association ratio ranging from 0.05 to 0.6.

allocate large probabilities to the correctly matched stations. We use the top-k accuracy measure. Depending on different k values, it measures the probability that the inferred set of the top-k stations (ordered by probabilities) includes the actual associated station in reality:

$$\text{top} - k\,\text{Accuracy} = \frac{N_c}{N_{\mathbf{F}_\phi}}, \tag{8}$$

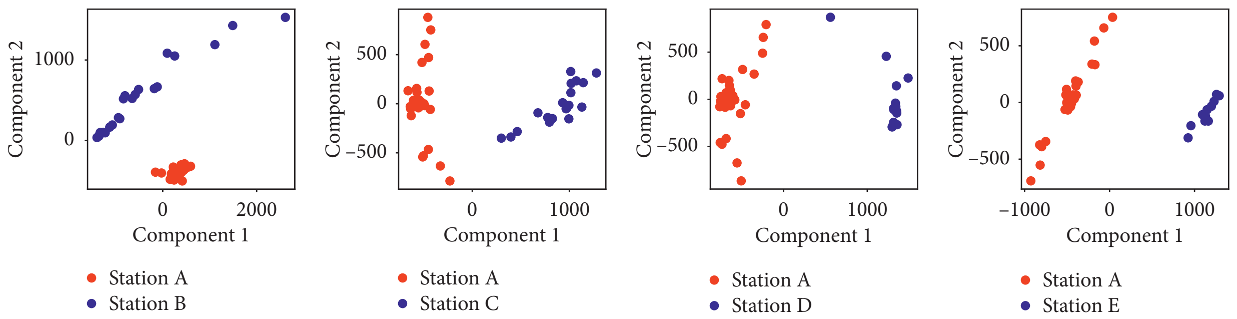
where N_c is the number of fare machines in $\mathbf{F}_\phi^*$ with their matched station contained in $[\pi(1), \pi(2), \ldots, \pi(k)]$ and $N_{\mathbf{F}_\phi^*}$ is the number of fare machines in $\mathbf{F}_\phi^*$.

Table 3 summarizes the model performance of P2 with varied levels of invalid association ratios in the dataset.

Table 3: Model performance in rematching invalid associated fare machines.

	Invalid association ratio (%)							
	5	10	15	20	25	30	35	40
Top-1	76.4	77.8	77.1	78.9	75.7	74.1	70.9	69.1
Top-2	86.8	87.1	87.2	88.5	87.4	85.1	82.1	81.4
Top-3	90.8	91.1	91.7	91.8	91.7	89.9	88.3	87.7
Top-4	93.1	93.4	94.4	93.6	93.9	92.5	90.9	91.0
Top-5	94.1	95.0	95.9	95.3	95.1	94.1	92.9	93.3

The results show that the top-k accuracy exceed 90% when k is greater than 3, regardless the invalid association ratio. It indicates that the top 3 inferred stations from the



(a) (b) (c) (d)

Figure 8: MRF feature vectors between Station A and (a) Station B, (b) Station C, (c) Station D, and (d) Station E.

model are highly likely to include the correctly associated station of the studied fare machine. This provides an important implication for further field investigations to these probable stations in practice, i.e., checking the most likely stations that the invalid associated fare machines may belong to.

4.3. Latent Feature Analysis. The foundation of the detection or inference model being effective is the quality of the MRF features. That is, the fare machines at different stations are preferable to have significantly different MRF features. To explore the feature quality, we utilize the principle component analysis (PCA) [7] to reduce the dimension of the MRF feature vector to two. We randomly choose 5 stations in the studied metro system, select one station as the reference station, and compare its MRF feature vector to that of the other 4 stations, respectively.

Figure 8 shows the MRF feature visualization results. The results show that the MRF features between stations exhibit significant differences, which indicates a high quality. This benefits the model to formulate relatively distinct MRF feature for each station, thus which is effective to detect the invalid associations and infer the associated station of the fare machines. For different stations, the MRF feature of fare machines appears different patterns. For example, the MRF features of machines in Station E (Figure 8(d)) are very similar to each other, while the MRF features of Station B (Figure 8(a)) appear a distributed manner. The reason partly lays in the different layout of the stations. For some large stations (e.g., transfer stations in the commercial center), there are many gates entering/exiting the stations, which may lead to variances in travel time between the same OD pairs. It would be the main reason for the miss and wrongly detection of the proposed model.

5. Conclusion

Ensuring data quality is essential for its effective use in practice. The paper proposes a model to detect the invalid data in the AFC dataset, caused by the erroneous association between fare machines and stations (e.g., due to delayed updating dictionaries or incorrect data merging). It combines tensor decomposition, isolation forest, and NN methods to detect the invalid associations in the recorded dataset and infer the correct association station that a fare machine belongs to.

The model is validated using the AFC data in a busy metro system. The experiment results show that the invalid association can be detected with more than 90% accuracy when the invalid association ratio is low. Also, the model is robust to invalid associations and it can still achieve 69.62% accuracy in the extreme case when the invalid association ratio is 55%. The association station inference results indicate that the top 3 inferred stations from the model are highly likely to include the correctly associated station of the studied fare machine (around 90%). This provides an important implication for further field investigations to these probable stations in practice.

The proposed model provides useful knowledge for the AFC data management in terms of data quality check and fixing invalid data. Though the study focuses on the invalid data detection problem, the model is general and can be generalized to inference applications, e.g., inferring the alighting stations for the bus system having only the boarding records. As the extracted MRF features are meaningful, further studies could focus on the analysis based on the MRF features, for example, analysing the different utilization of fare machines in different gates of the same station to improve the infrastructure efficiency.

Data Availability

The AFC data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] H. N. Koutsopoulos, Z. Ma, P. Noursalehi, and Y. Zhu, "Transit data analytics for planning, monitoring, control, and information," in *Mobility Patterns, Big Data and Transport Analytics*, pp. 229–261, Elsevier, Amsterdam, The Netherlands, 2019.

[2] X. L. Dong, L. Berti-Equille, and D. Srivastava, "Truth discovery and copying detection in a dynamic world,"

Proceedings of the VLDB Endowment, vol. 2, no. 1, pp. 562–573, 2009.

[3] S. Robinson, B. Narayanan, N. Toh, and F. Pereira, "Methods for pre-processing smartcard data to improve data quality," *Transportation Research Part C: Emerging Technologies*, vol. 49, pp. 43–58, 2014.

[4] X. Ma, Y.-J. Wu, Y. Wang, F. Chen, and J. Liu, "Mining smart card data for transit riders' travel patterns," *Transportation Research Part C: Emerging Technologies*, vol. 36, pp. 1–12, 2013.

[5] Y. Liu, X. Weng, J. Wan, X. Yue, H. Song, and A. V. Vasilakos, "Exploring data validity in transportation systems for smart cities," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 26–33, 2017.

[6] D. Everett: 60000 Oyster Cards Corrupted, [EB/OL], 2008, https://www.yumpu.com/en/document/read/27117143/60000-oyster-cards-corrupted-smart-card-news.

[7] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[8] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[9] S. Wold, M. Sjöström, and L. Eriksson, "Pls-regression: a basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109–130, 2001.

[10] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*, Vol. 1, MIT Press, Cambridge, UK, 2016.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[12] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, 2009.

[13] H. Tan, G. Feng, J. Feng, W. Wang, Y.-J. Zhang, and F. Li, "A tensor-based method for missing traffic data completion," *Transportation Research Part C: Emerging Technologies*, vol. 28, pp. 15–27, 2013.

[14] X. Chen, Z. He, and J. Wang, "Spatial-temporal traffic speed patterns discovery and incomplete data recovery via SVD-combined tensor decomposition," *Transportation Research Part C: Emerging Technologies*, vol. 86, pp. 59–77, 2018.

[15] L. Sun and K. W. Axhausen, "Understanding urban mobility patterns with a probabilistic tensor factorization framework," *Transportation Research Part B: Methodological*, vol. 91, pp. 511–524, 2016.

[16] P. J. Rousseeuw and K. V. Driessen, "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, vol. 41, no. 3, pp. 212–223, 1999.

[17] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pp. 427–438, Dallas, TX, USA, May 2000.

[18] F. Angiulli and C. Pizzuti, "Fast outlier detection in high dimensional spaces," in *Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery*, pp. 15–27, Springer, Helsinki, Finland, August 2002.

[19] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pp. 93–104, Dallas, TX, USA, May 2000.

[20] M. Çelik, F. Dadaşer-Çelik, and A. Ş. Dokuz, "Anomaly detection in temperature data using DBSCAN algorithm," in *Proceedings of the 2011 International Symposium on Innovations in Intelligent Systems and Applications*, pp. 91–95, IEEE, Istanbul, Turkey, June 2011.

[21] G. Münz, S. Li, and G. Carle, "Traffic anomaly detection using k-means clustering," in *Proceedings of the GI/ITG Workshop MMBnet*, pp. 13-14, Hamburg, Germany, September 2007.

[22] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422, IEEE, Pisa, Italy, December 2008.

[23] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, and T. Soderstrom, "Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 387–395, London, UK, August 2018.

[24] D. Li, D. Chen, B. Jin, L. Shi, J. Goh, and S.-K. Ng, "Mad-gan: multivariate anomaly detection for time series data with generative adversarial networks," in *Proceedings of the International Conference on Artificial Neural Networks and Machine Learning-ICANN 2019: Text and Time Series*, pp. 703–716, Springer, Munich, Germany, September 2019.

[25] D. T. Nguyen, C. K. Mummadi, T. P. N. Ngo, T. H. P. Nguyen, L. Beggel, and T. Brox, "Self: learning to filter noisy labels with self-ensembling," 2019, https://arxiv.org/abs/1910.01842.

[26] J. D. Carroll and J.-J. Chang, "Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, 1970.

[27] E. Acar, D. M. Dunlavy, T. G. Kolda, and M. Mørup, "Scalable tensor factorizations for incomplete data," *Chemometrics and Intelligent Laboratory Systems*, vol. 106, no. 1, pp. 41–56, 2011.

[28] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: adaptive synthetic sampling approach for imbalanced learning," in *Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pp. 1322–1328, IEEE, Hong Kong, China, June 2008.

[29] D. P. Kingma and B. J. Adam, "A method for stochastic optimization," 2014, https://arxiv.org/abs/1412.6980.

Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 5573650, 23 pages
https://doi.org/10.1155/2021/5573650

Research Article

A Hybrid LSTM-Based Ensemble Learning Approach for China Coastal Bulk Coal Freight Index Prediction

Wei Xiao,[1,2,3] **Chuan Xu**,[1,2,3] **Hongling Liu**,[1] **and Xiaobo Liu**[1,2,3]

[1] *School of Transportation and Logistics, Southwest Jiaotong University, No. 111 Erhuanlu Beiyiduan, Chengdu 610031, China*
[2] *National United Engineering Laboratory of Integrated and Intelligent Transportation, Southwest Jiaotong University, Chengdu, Sichuan, China*
[3] *National Engineering Laboratory of Integrated Transportation Big Data Application Technology, Chengdu 611756, Sichuan, China*

Correspondence should be addressed to Chuan Xu; xuchuan@swjtu.edu.cn

Received 19 February 2021; Revised 13 April 2021; Accepted 13 May 2021; Published 26 May 2021

Academic Editor: Dong-Kyu Kim

China Coastal Bulk Coal Freight Index (CBCFI) reflects how the coastal coal transporting market's freight rates in China are fluctuated, significantly impacting the enterprise's strategic decisions and risk-avoiding. Though trend analysis on freight rate has been extensively conducted, the property of the shipping market, i.e., it varies over time and is not stable, causes CBCFI to be hard to be accurately predicted. A novel hybrid approach is developed in the paper, integrating Long Short-Term Memory (LSTM) and ensemble learning techniques to forecast CBCFI. The hybrid LSTM-based ensemble learning (LSTM-EL) approach predicts the CBCFI by extracting the time-dependent information in the original data and incorporating CBCFI-related data, e.g., domestic and overseas thermal coal spot prices, coal inventory, the prices of fuel oil, and crude oil. To demonstrate the applicability and generality of the proposed approach, different time-scale datasets (e.g., daily, weekly, and monthly) in a rolling forecasting experiment are conducted. Empirical results show that domestic and overseas thermal coal spot prices and crude oil prices have great influences on daily, weekly, and monthly CBCFI values. And in daily, weekly, and monthly forecasting cases, the LSMT-EL approaches have higher prediction accuracy and a greater trend complying ratio than the relevant single ensemble learning algorithm. The hybrid method outperforms others when it works with information involving a dramatic market recession, elucidating CBCFI's predictable ability. The present work is of high significance to general commerce, commerce-related, and hedging strategic procedures within the coastal shipping market.

1. Introduction

Nowadays, 90% of global trades are completed via sea transportation [1]. Sea transport is critical to the business system on the globe and also in the domestic trade system. China Coastal Bulk Coal Freight Index (CBCFI) [2] is built for timely reflecting how coastal coal transporting market's freight rates in China are fluctuated, by complying with the present mechanism of China Coastal Bulk Freight Index (CBFI) [3]. This system can publicize the complex index and spot ratios in terms of various routes/kinds pertaining to vessels of the coastal coal service market on a day-to-day basis. China (Coastal) Bulk Freight Index Panelist offers CBCFI freight data per weekday (Shanghai Shipping Exchange publicizes CBCFI at the official website and as well as http://www.chineseshipping.com.cn at 15:00 (Beijing Time) on each index publication day).

CBCFI represents the voyage-charter freight rate conditions pertaining to the market of bulk coal shipping in coastal areas. This indicates the unstable property pertaining to the coal bulk shipping market, as well as reflecting the developing states pertaining to the China economic condition and domestic business tendency. Thus, it refers to the "weatherglass" pertaining to the market of bulk coal shipping in coastal areas. Due to the mentioned feature, numerous internal personnel and specialists attempt at estimating the subsequent tendency pertaining to the market of bulk coal shipping in coastal areas by accurately predicting the bulk coal freight index for guiding company strategy decisions. Additionally, forecasting CBCFI value enables operating personnel and decision-making personnel for managing market trends and avoiding risks inside the coastal coal shipping market. Furthermore, it helps industries and manufactories with the domestic shipping system.

Previous studies indicate that the shipping freight index usually presents complicated instability features to be uncertain, cyclical, and nonlinear properties [4–6]. To achieve the satisfactory freight index prediction, a variety of freight indices predicting methods developed previously, where the econometric time series predicting approaches exhibit drawbacks in nonstationary and nonlinear properties, while the single neural network (NN) methods show disadvantages in overfitting, local minimum point, and parameter selection problems. Because both single econometric and NN approaches are limited in freight index prediction, recently, empirical studies consistently demonstrate the adaptability of hybrid techniques. Hybrid approaches are capable of combining individual approaches and make the respective advantage remedy others' deficiencies.

Recently, ensemble learning algorithms are extensively employed for analyzing the multivariable prediction. Ensemble learning algorithms (e.g., random forest (RF) and gradient boosting regression tree (GBRT)) are effective in determining the essential variables of a time series and investigate the inner relations among variables [7]. Over the past few years, ensemble learning algorithms have been extensively employed for studying the mentioned time series to be stock prices, Baltic Dry Index, and traffic flow, leading to the production of essential outcomes [8–10].

Ensemble learning algorithms' fast advancement presents one novel idea for the way of utilizing multidata and improves the readability from the original data. We here combine AI and ensemble learning algorithms for formulating one emerging hybrid approach, an ensemble learning (LSTM-EL) approach (e.g., LSTM-GBRT and LSTM-RF) by exploiting Long Short-Term Memory (LSTM) for CBCFI forecasting. Inside the approach, the LSTM layer obtains details dependent on time within the data, and the GBRT/RF layer shows great robustness of ensemble learning in terms of the training of approach.

The present work has the following organization. Section 2 presents the review of the analyses of the market of shipping freight. Section 3 presents the designed approaches. Section 4 introduces the data collection. In Section 5, the approach performance receives the comparison and analysis. Lastly, Section 6 gives the concluding remarks.

2. Literature Review

Since freight rates are uncertain and unstable, the approaches to quantitatively analyze the rates arouse long concern from the shipping industry. As a result, increasing literature proposes approaches to predict freight rates, many of which use the Baltic Dry Index (BDI) [11]. Both BDI and CBCFI are shipping indices that effectively assess the current situation or shipping market, and they have some similar features. First, CBCFI and BDI are both daily number-issued leading indicators that measure the costs of shipping raw materials. Specifically, BDI mainly measures shipping costs for dry bulk commodities, including coal, grain, iron ore, and metals, while CBCFI mainly measures shipping costs for coal; second, in terms of the composition, BDI takes 23 shipping routes measured on a time charter and voyage basis while CBCFI takes 14 shipping routes measured on a voyage basis; third, as for the type of vessel, BDI looks at the ships that can carry 15,000 deadweight tons (DWT)–80,000 DWT of cargo (90 percent of the global fleet), while CBCFI focuses on the ships that can carry 15,000 DWT–60,000 DWT [3, 12, 13]. As the BDI and CBCFI share these similar features, we refer to the abundant research of the BDI to obtain a clear view of current prediction approaches. The analytic methodology of BDI among others is split into three types.

Traditional econometric approaches are initially covered, including vector error correction (VEC), generalized autoregressive conditional heteroskedasticity (GARCH), vector autoregression (VAR), and the autoregressive integrated moving average (ARIMA) approaches. Cullinane et al. [14] first developed a Baltic Dry Bulk Index (BDI) study approach method through the ARIMA approach. Kavussanos and Alizadeh-M [15] developed one season-related ARIMA approach for one independent variant as well as one VAR approach for investigating seasonality in the dry bulk shipping market. Batchelor et al. [16] discussed the performances of VAR and ARIMA as well as vector equilibrium correction (VECM) approaches for the prediction of spot and forward freight rates. To improve the accuracy of BDI forecasts, Tsioumas et al. [17] developed a multivariate vector autoregressive approach with exogenous variables (VARX) approach and the results demonstrate that the VARX approach outperforms the ARIMA approach. Chen et al. [18] applied the ARIMA and the VAR approaches to predict the spot rates of several dry bulk routes and found out that the VAR approach performed better than the ARIMA approach in test-sample forecasts. Adland et al. [19] showed a cointegration-based method for analyzing regional ocean freight rates' dynamic properties.

According to Stopford [20], sea prediction is difficult for statistical and traditional econometrical approaches for capturing the nonlinearity of dry bulk freight rates [21]. Accordingly, a second type of analytic approach currently employs several nonlinearity and artificial intelligence (AI) approaches, comprising artificial neural network (ANN), machine learning algorithm, and nonlinear methods. Li and Parsons [22] applied neural systems to predict monthly tanker freight rates from the short to the long term and found the neural systems outperforming the ARIMA time sequence approach for prediction in the long run. Yang et al. [23] investigated and predicted the freight rate instability alarming pertaining to China Coastal Bulk Freight Index (CCBFI), China Containerized Freight Index (CCFI), and Baltic Freight Index (BFI) by exploiting support vector machine (SVM). Thalassinos et al. [24] adopted a chaos approach for predicting the BDI using the invariable coefficients pertaining to the strange attractor under reconstruction, governing the system's evolving process. Guan et al. [25] developed one SVM-based multistep prediction approach to predict weekly Baltic Supermax Index (BSI) data. Şahin et al. [26] develop three ANN approaches based on BDI data and results show that their performances are close, whereas the highest consistency pertains to the ANN by exploiting the past two weekly observations of the BDI data. Among the deep-learning approaches, the long short-term memory (LSTM) neural network is regarded as a practical technique for handling time series problems [27–29]. For example, Nelson et al. [27] employed the LSMT network to predict the future trends of stock prices based on the historical price, alongside technical analysis indicators; Duan et al. [30] implemented the LSTM model for multistep ahead travel time prediction. The above studies demonstrated that NN methodologies generate more effective forecasting performance than conventional time series and econometric approaches.

Though NNs are capable of handling nonlinearity and show better robustness, it is hard to determine the configuration of the NN algorithm; in addition, it tends to fall into over or lacked training, easily resulting in local minimum trapping [11]. This prompts one shift into the third category hybrid methodologies, which usually integrates one noise-reducing method by adopting one NN-based algorithm. Leonov and Nikolov [31] proposed a hybrid approach of wavelets and neural networks to investigate the fluctuation in the freight rates of the Baltic Panamax routes 2A and 3A. Bulut et al. [32] established one vector autoregressive fuzzy combined logical predicting approach in terms of time charter rates. Zeng et al. [33] propose a hybrid approach of empirical mode decomposing process (EMD) as well as ANN. Uyar et al. [34] presented one trained recurrent fuzzy neural system with the use of a genetic algorithm for improving long-term dry cargo freight rates prediction to be more accurate. As an efficient strategy to improve the forecasting ability of a single model, ensemble learning has been widely used to improve the model performance [35–38]. For example, Kamal et al. [35] developed a deep ensemble recurrent network of recurrent neural network (RNN), long-short-term memory (LSTM), and gated rectified unit neural network (GRU) to improve the BDI predictive performance, and results showed that the ensemble method outperforms the single deep-learning approach. Tan et al. [37] proposed an LSTM-based deep ensemble learning model that combined bagging, random subspace, and boosting to forecast ultra-short-term industrial power demand, and they found out that the proposed model obtains higher accuracy and robustness than LSTM, eXtreme Gradient Boosting (XGBoost), and other time series methods. Liu et al. [36] proposed a deep air pollution model for forecasting PM2.5 concentrations based on a wind-sensitive attention mechanism, LSTM, and XGBoost, and experiments illustrate that the proposed approach is superior to a single multilayer perceptron (MLP), SVM, LSTM, and XGBoost. The above studies indicate that models combining machine learning and neural networks usually obtain better predictive results than any single model.

Although much research has been conducted to solve the shipping indices forecasting issues, most do not consider different spot rates or other related factors information. Accordingly, these methods cannot effectively reflect the critical factors that contribute to strengthening the predictive performance. To address this issue, we propose a hybrid model of LSTM and ensemble learning algorithm to handle the CBCFI forecasting problems. The LSTM approach has the ability to acquire the data determined by time and noticeably impacts the predicting process of time series, whereas this approach fails to appropriately mine the implicit relations between exogenous variables in predicting inflection point data. Thus, to better utilize feature information, it is necessary to incorporate an applicable ensemble learning algorithm to optimize the feature combination in order to construct the feature set that reflects the short/long-term trend of the CBCFI. For the ensemble learning part, boosting and bagging are two of the important ideas; they both combine a set of weak learners to create a strong learner that obtains better performance. In all kinds of machine learning methods, Gradient Boosting Regression Tree (GBRT) [39] and Random Forest (RF) [40] have received much attention and are often used as representatives of ensemble learning algorithms. GBRT and RF are decision-tree-based ensemble learning algorithms that use a boosting and bagging framework, respectively. Because of the implementation of a gradient boosting algorithm or a bootstrap sampling method, GBRT or RF can handle variables fast, making it suitable for complicated tasks. Accordingly, this study adopts GBRT and RF as ensemble learning algorithms for CBCFI forecasting.

3. Data Description

This section firstly presents the data source. As this dataset includes both historical freight index and other impacted variables, variable correlations are introduced and conducted in this section as well.

3.1. Data Source. All datasets in the present study are supported by the Wind Economic Database [41]. Wind pairs more than 1.3 million macroeconomy-related and industry time series based on effective graphics and data study equipment for elucidating China's economy to finance-related professionals.

The China (Coastal) Bulk Coal Freight Index (CBCFI) is a compound weekday index (null on holiday and weekend) that considers 14 shipping routes [2]. It takes 1st September 2011 as the base period, and the base index is 1,000 points. Now, CBCFI is extensively employed by practitioners of industry and considered a vital economic indicator in coastal coal transportation. Figure 1 displays the composite routes of CBCFI and its corresponding weights. It can be seen that CBCFI mainly describes the shipping market of transporting coal from the north to the south. According to the histogram at the bottom of Figure 1, we notice that six routes whose weights are over 10% hold dominant positions, in which most of them depart from Qinhuangdao Port. For easily reading, in Figure 1, the dominant routes are plotted in red, and for those weight values less than 10% but nonzero,

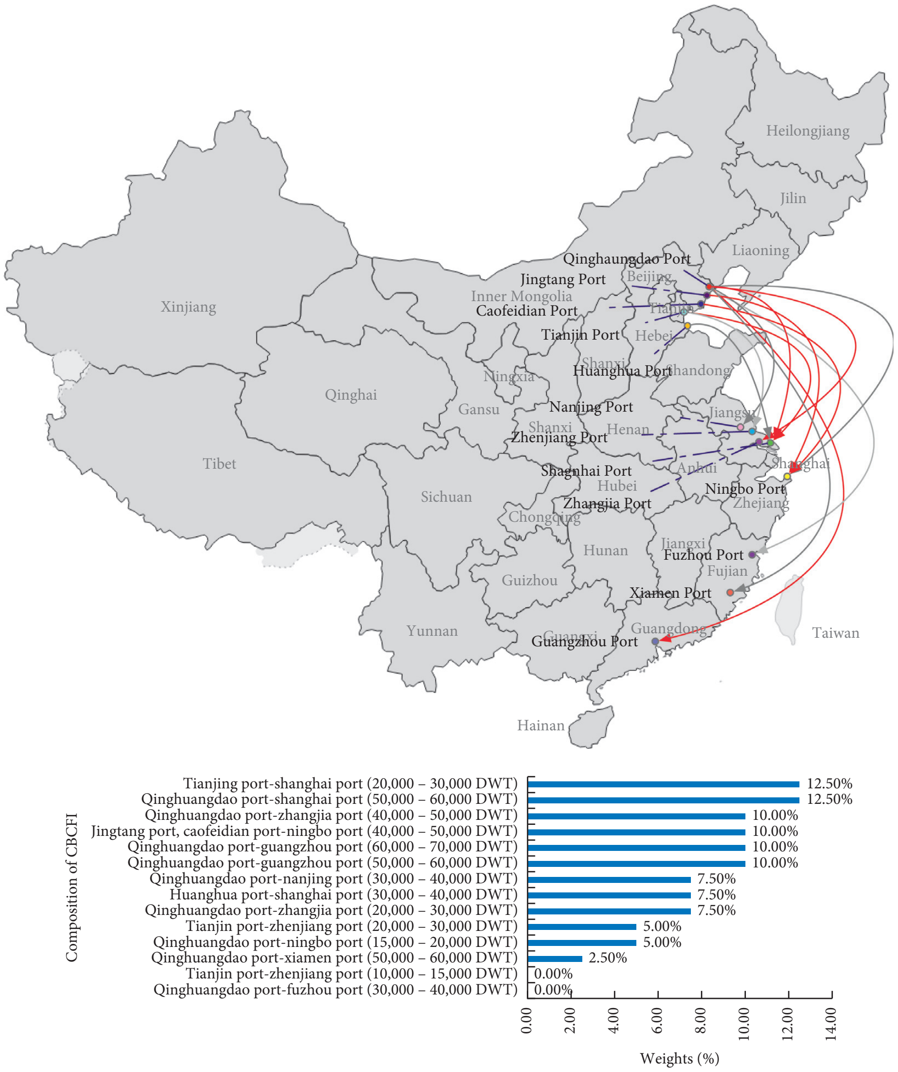


FIGURE 1: Illustration of the CBCFI composition and the corresponding shipping routes (DWT in bracket represents the "deadweight tonnage (DWT)" of the ship).

routes are marked in dark grey, while those with zero weights are in light grey.

Previous studies of traditional shipping index (e.g., BDI) generally discuss the daily, weekly, and monthly forecasting cases to evaluate the performance of approaches in short-, mid-, and long-term predictions [11, 16, 22]. Thus, in this paper, daily, weekly, and monthly forecasting experiments are conducted and the full available CBCFI data from January 2012 to October 2020 are employed as the experimental datasets, with 2129 daily observations, 422 weekly observations, and 10 yearly observations.

According to Figure 2, the CBCFI fluctuations exhibit irregular properties and high frequencies as well as large amplitudes. In the long run, from 2012 to the first half of 2016, CBCFI had a few inflection points. And in the fourth quarter of 2013, CBCFI reached the highest value of 1450.39 since its release and then returned to the low level at the end of the year. From the second half of 2016 to now, CBCFI has experienced a large fluctuation range, especially from 2017 to 2020. Compared with the previous years, the rise and fall of freight rates have been much larger, and the frequency of rise and fall also increases obviously, which reflects the changeable market situation of coastal coal transport. Moreover, the global Coronavirus Disease 2019 (COVID-19) outbreak has generated a public health crisis that started in December 2019, which has certain consequences for the economics of the shipping industry. For seasonal fluctuations, from the trend of each year, except for 2014, CBCFI shows the phenomenon that its index rises in summer and winter in other years, but the increase rate is different each year. The main reason is that demand for coal tends to rise in summer and winter, traditionally the peak seasons for electricity use, leading to a seasonal rise in coal transport prices. Although not every summer or winter in the past 7 years has seen an increase in rates, historical statistics indicate a high probability of rates rising in both summer and winter each year.

To justify the uncertain and nonlinear characteristics of the CBCFI series, descriptive statistics of the daily CBCFI and its rate of change are provided, and the results are listed

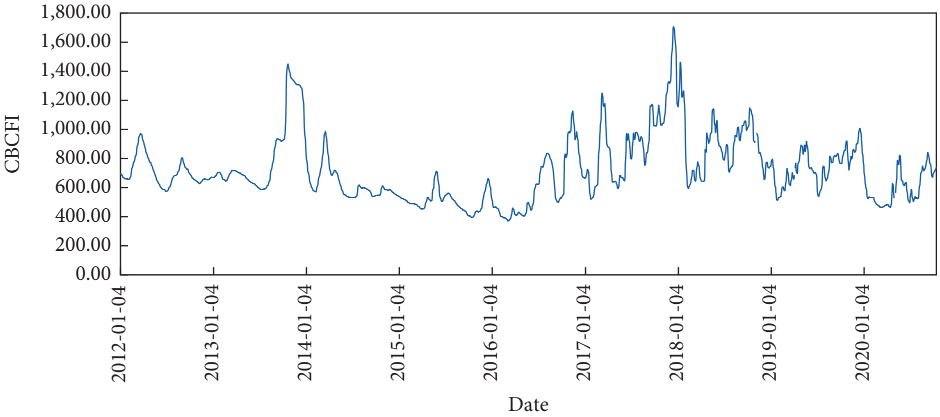


FIGURE 2: Tendency of daily CBCFI dataset adopted for predicting process.

in Table 1. The rate of change for the CBCFI is $R_t = ((y_t - y_{t-1})/y_{t-1})$.

From the above statistics, it can be observed that, in terms of the skewness and kurtosis, the skewness of CBCFI series and its rate of change series are greater than zero, and the kurtosis of each is either less than or greater than three. Therefore, the daily CBCFI and its rate of change series are left-skewed and right-skewed, respectively, which indicates that they have the characteristics of a sharp peak and a thick tail. In addition, the J-B test results indicate that both time series do not follow a normal distribution because they reject the hypothesis of the Jarque-Bera (J-B) statistic at 5% significant level. Moreover, both daily CBCFI and its rate of change series are indicated to be nonstationary, based on the results of the augmented Dickey–Fuller (ADF) unit root test; specifically, the t-statistics values summarized in Table 1 are less than the value of the critics (−2.86)). Furthermore, the Ljung-Box statistics to the squared residuals (also known as Q-Statistics of Square Residuals) is applied to test the nonlinearity of the CBCFI and its rate of change series. For both series at lag 6, we reject the hypothesis of no autocorrelation at the 5% significance level. Autocorrelated squared residuals are indications of nonlinearity [2]. The above statistics indicate that the CBCFI has the uncertain, nonstationary, and nonlinear properties, and those complicated features justify the need for the proposed ensemble approach.

TABLE 1: Descriptive statistics of daily CBCFI and rate of change of CBCFI.

Statistic	Daily CBCFI	Rate of change for CBCFI
Mean	722.6500	−0.0002
Median	675.4800	−0.0012
Standard deviation	223.8494	0.0228
Maximum	1,706.2000	0.1186
Min	370.9900	−0.1337
Kurtosis	1.7801	5.6084
Skewness	1.2186	0.2550
J-B	658.9484	626.4104
p value	0.000	0.000
ADF [lags]	−4.5298 [12]	−11.3960 [12]
p value	0.0130	0.0164
Q [lags]	26.571 [6]	15.95 [6]
p value	0.000	0.000
Observation	2129	2128

Note. J-B represents the Jarque-Bera (J-B) test in which a goodness-of-fit test is performed for examining whether the time series follows a normal distribution [42]. ADF represents the augmented Dickey-Fuller (ADF) test with a null hypothesis of the existence of autocorrelation in the sample series [43]. Q represents the Ljung-Box statistics to the squared residuals (also known as Q-Statistics of Square Residuals) in which a nonlinearity test [44] under the null hypothesis of linearity and residuals of a properly specified linear model should be independent [45].

3.2. CBCFI Impact Factor Data. Within other markets, freight rates inside the shipping industry receive the formation based on the interacting processes pertaining to various factors, such as the prices of the transported cargo, the demand and supply, and the cost. Thus, if these elements influence the CBCFI, then they should be investigated. The demand for coastal transport arises from the need of exporters and importers to transport the coal to specifically domestic destinations. The "derived" demand is mainly affected by domestic economy and trade, such as the Import and Export Trade (IET), the industry production, domestic coal inventory, the contract rates, and spot rates of domestic thermal coal. As the domestic economy is improved, domestic trade will be promoted, and shipping transport will be more demanded. Moreover, random shocks based on emergency events (e.g., 2015 Tianjin explosions and 2019 COVID-19 outbreak) and cyclical and seasonal market movements of the coal transported by sea further substantiate that the demand for shipping transport depends on macroeconomic factors. Besides, transportation consumption and costs constitute other CBCFI determinates. On the promise of collecting data, the relevant data of the CBCFI are considered maximally, which could roughly be classified into three categories:

(i) *Domestic and Overseas Thermal Coal Spot Prices.* It is worth mentioning that Qinhuangdao Port is the greatest coal export port on the globe, and the routes of CBCFI are the dominant carriers for thermal coal transportation from north to south. Thus, the spot price of thermal coal is considered as one of the critical factors affecting the fluctuation of CBCFI. Furthermore, with the enlarged opening of the market, the influence of the international market on the domestic market will be increasingly intensified, and the correlation between the domestic and foreign markets will be further enhanced. Even at this stage, the price of thermal coal in the international market will affect the export of coal, which will be reflected through the price of thermal coal.

(ii) *Coal Inventory*. Coal inventory refers to a critical element that impacts the coal price based on the economic new normal [46]. Coal inventory is the result of coal production, transportation, consumption, and other factors. It is basically the same as the factors of price formation and has a leading effect on the price change of thermal coal. In recent years, Qinhuangdao coal inventory has become the weathervane of thermal coal price change [47].

(iii) *Fuel Oil and Crude Oil Prices*. On the one hand, fuel oil prices have a significant impact on coal prices as fuel oil prices influence the cost of shipping. When the price of fuel oil increases, the shipping cost will increase; when the price of fuel oil declines, the shipping cost will decrease. On the other hand, coal and crude oil are the most basic energy sources, and the sharp rise in oil prices has also contributed to the rise in thermal coal prices [48]. The influence of crude oil price on thermal coal price is reflected in the following aspects. First, crude oil, that is, one type of vital fossil energy source, refers to a dominant substitute to coal. Thus, crude oil price instability can impact coal demand and price in a corresponding manner [49]. When crude oil rises in price, the demand for coal would rise as the substitute. The fluctuation of coal price will influence the bulk coal shipping cost somehow and thus influence the CBCFI values. Note that the "prices" discussed in the present study contain the spot prices and futures prices (a spot price is an offer to complete a commodity transaction immediately, while a future contract locks in a price for future delivery). Table 2 summarizes the abovementioned factors in the basic feature set and will be optimized later. Variables are numbered sequentially for the feature description.

3.3. Variables Correlation Analysis. To give a clear and simple view of the correlations between CBCFI and relevant impact factors, the Pearson correlation coefficient [50] is calculated and the results are presented in Table 2. Domestic and overseas thermal coal spot rates (u_1~u_8) show obvious correlation with CBCFI, particularly, "Qinhuangdao Port-Q5000 Index-FOB," "Jintang Port-Q5000 Index-FOB," and "Guangzhou Port-Q5500 Index (Indian Coal)-EXT" with the largest Pearson correlation values; on the other hand, coal inventories ($u_9 \sim u_{12}$) show negative correlations with the CBCFI series. Based on the results, we remove two variables with the smallest Pearson coefficient value (u_9 and u_{11}), and the rest of the 22 variables are selected as the input variables.

In addition, to demonstrate the necessary property for selecting the ensemble learning algorithm to deal with the CBCFI prediction problem, the relationships between the 24 variables and CBCFI are checked. Then, one color-coded Pearson correlation matrix is generated. The numerical value one with the expression of dark blue indicates one overall positive linear correlation of two characteristics, whereas chartreuse indicates zero, demonstrating no linear correlation. As is shown in Figure 3, there is an interrelation of different degree between the 24 variables. For example, fuel oil spot and futures prices ($u_{14} \sim u_{19}$) are highly correlated with crude oil spot and futures prices ($u_{20} \sim u_{24}$). Therefore, the ensemble learning algorithms dealing with multidata relations are considered to solve the CBCFI prediction problem.

4. Methodology

In this section, we first give a problem statement to present an overview of the prediction problem researched in this work. Then, the core concept and the flowchart with algorithm pseudocode of the proposed hybrid model structure are presented. At last, the prediction accuracy measurements are described.

4.1. Problem Statement. The goal of this work is to predict the CBCFI values of the next day given historical data. We define the historical observations of the target CBCFI as $Y = (y_1, \ldots, y_t, \ldots, y_T)^T \in \mathbb{R}^T$, where T represents time window size and y_t is the CBCFI value at time t. Similarly, auxiliary factors are represented by $X = (X_1, \ldots, X_t, \ldots, X_T)^T \in \mathbb{R}^{T\times D}$, where D specifies the number of related factors. $X_t \in \mathbb{R}^D$ is the values of all D related factors at time t, and $X^d \in \mathbb{R}^T$ is the value of the d^{th} factor in time window T. Thus, the prediction target y_{T+1} could be defined as follows:

$$y_{T+1} = F(y_1, \ldots, y_t, \ldots, y_T; X_1, \ldots, X_t, \ldots, X_T), \quad (1)$$

where $F(\cdot)$ is the mapping function we are aiming to learn.

4.2. Preparing Data. Given a CBCFI time series $\tilde{s}$ with length N, $\tilde{s} = \{s(t_i)\}_{i=1}^N$, the time series firstly should be preprocessed. Data cleaning and data normalization are the key data preparation tasks for further forecasting tasks. As the given time series has no missing value and to preserve the characteristics of real-world data, no noise reduction or data smoothing was performed on data. In this study, we only normalized the data with min–max normalization algorithm. The time series $\tilde{s}(t_i)$ is normalized and the resulting normal data are expressed as $s = \{s(t_i)\}_{i=1}^N$:

$$s(t_i) = \frac{\tilde{s}(t_i) - \min(\tilde{s})}{\max(\tilde{s}) - \min(\tilde{s})}, \quad (2)$$

Table 2: Impact factors description and the corresponding Pearson coefficient values.

Variable category	Variable description	Feature number	Pearson coefficient values
Domestic and overseas thermal coal spot rates	Qinhuangdao Port-Q5500 Index-FOB (a spot price is an offer to complete a commodity transaction immediately, while a futures contract locks in a price for future delivery. It leaves its point of origin)	u_1	0.550
	Qinhuangdao Port-Q5000 Index-FOB	u_2	0.575
	Qinhuangdao Port-Q5500K Index-FOB	u_3	0.512
	Qinhuangdao Port-Q5000K Index-FOB	u_4	0.530
	Jintang Port-Q5500 Index-FOB	u_5	0.554
	Jintang Port-Q5000 Index-FOB	u_6	0.571
	Guangzhou Port-Q5500 Index (Australian Coal)-EXT (EX-tank (EXT) refers to the price of coal shipped from the warehouse, including the price before the coal is put into the warehouse and the warehouse usage fee)	u_7	0.545
	Guangzhou Port-Q5500 Index (Indian Coal)-EXT	u_8	0.571
Coal inventories	Qinhuangdao Port	u_9	0.007
	Caofeidian Port	u_{10}	−0.157
	Guangzhou Port	u_{11}	−0.021
	Tianjin Port	u_{12}	−0.295
	Jintang Port	u_{13}	−0.197
Fuel oil prices	Fuel Oil 180 Singapore-FOB	u_{14}	0.202
	Fuel Oil 380 Singapore-FOB	u_{15}	0.205
	China fuel oil futures closing price (a closing price is the final price at which it trades during regular market hours on any given day) (continuous contract (a continuous contract is a reinsurance contract that does not have a fixed contract end date, which will continue to be renewed and be in effect until one of the parties in the contract terminates it. Continuous contracts are different from standard reinsurance contracts in that they do not provide coverage for only a fixed period of time))	u_{16}	0.265
	China fuel oil futures closing price (active contract (an active contract means that this future contract can be traded for a specific amount of time))	u_{17}	0.272
	China fuel oil futures settlement price (the settlement price is the average price at which a contract trades, calculated at both the open and close of each trading day) (continuous contract)	u_{18}	0.266
	China fuel oil futures settlement price (active contract)	u_{19}	0.273
Crude oil prices	Brent crude oil spot price	u_{20}	0.152
	West Texas Intermediate (WTI) crude oil spot price	u_{21}	0.157
	Dubai crude oil spot price	u_{22}	0.152
	WTI crude oil futures settlement price	u_{23}	0.181
	Brent crude oil futures settlement price	u_{24}	0.187

where $\min(\tilde{s})$ is the minimum value of $\tilde{s}$ and $\max(\tilde{s})$ is the maximum value of $\tilde{s}$.

4.3. LSTM-Ensemble Learning (LSTM-EL) Approaches for CBCFI Forecasting.

Traditional freight indices prediction methods usually use the historical time series data of the target with ignorance of other impact factors. Generally speaking, the trend of CBCFI is reflected in two ways: historical CBCFI information and impact factor information. The historical time series information sometimes is sparse and thus not enough to produce accurate prediction, while some close impact factor information could reflect the movement of CBCFI from different aspects to a certain extent and with the support of the powerful database of Wind Economic Database (Wind Economic Database, refer to https://www.wind.com.cn/en/edb.html).

Our proposed LSTM-EL model is composed of two layers: in the first layer, a cluster of LSTMs is constructed to generate the embedding features, and in the second layer, an ensemble learning method for final CBCFI prediction.

Figure 4 illustrates the overall framework of the bilevel hybrid LSTM-EL configuration; it consists of an LSTM method and two parallel ensemble learning methods. Note that the LSTM-EL model includes two different hybrid models, LSTM-GBRT and LSTM-RF. GBRT behaves similarly to RF in the manner of fitting multiple trees, but it instead fits them in a sequential manner, and hence another expectation of this paper aims to explore which ensemble algorithm fits the CBCFI forecasting problem better. For detailed descriptions of single RF, GBRT, and LSTM approaches, see Appendix A.

In the first layer, the dataset is first split into the in-sample and the out-of-sample. The preliminary embedded LSTM focuses on extracting the time-dependency information from variables of the in-sample and generates embedding features from the last LSTM layer, GBRT/RF is taken as an ensemble learning method to make the final

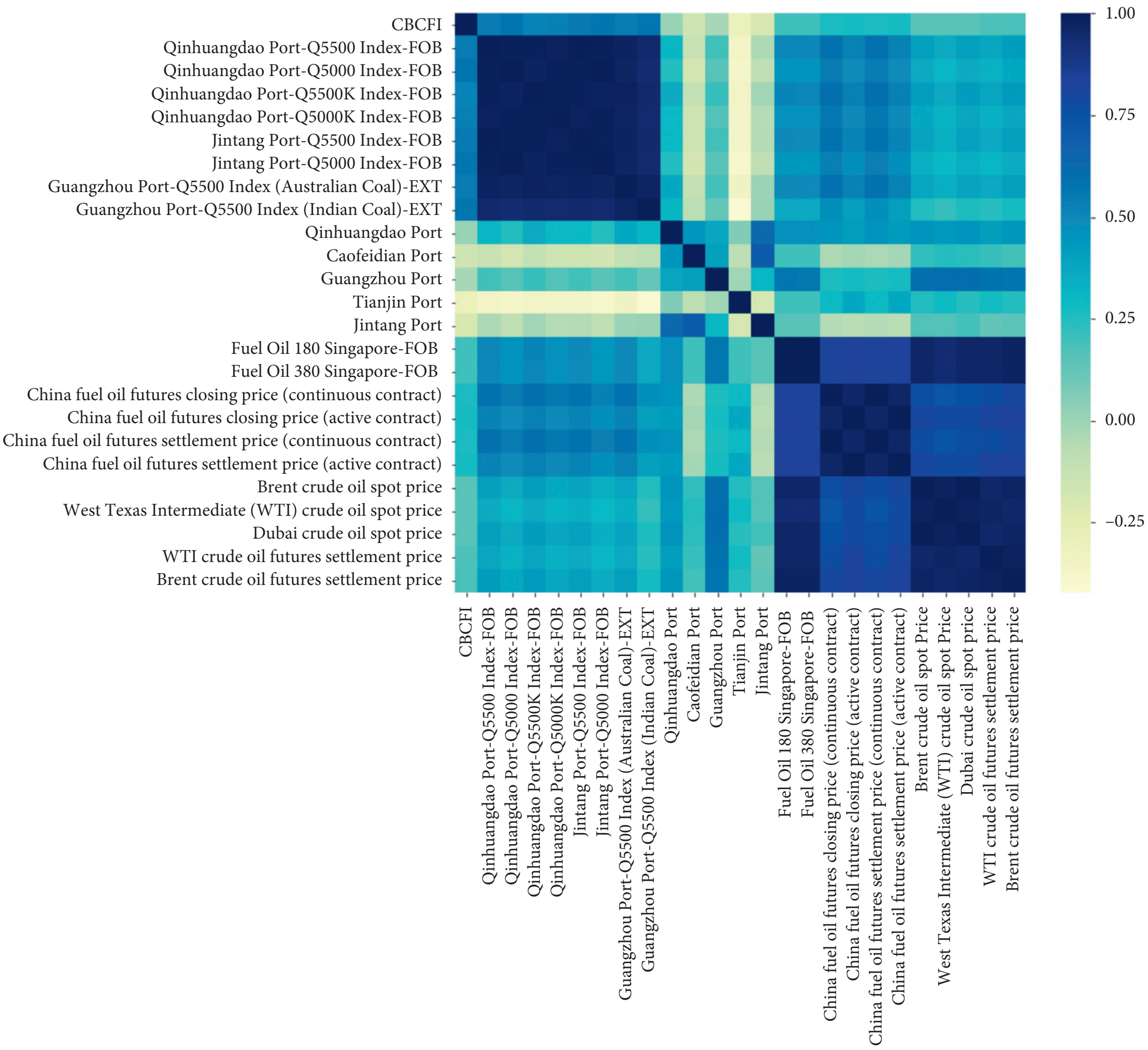


Figure 3: The Pearson correlation matrix of all features.

predictions by combining preliminary embedding features from LSTM, and the prediction values of each individual tree are summed up to get the final value. The details of the proposed LSTM-EL model are illustrated in Algorithm 1.

The idea behind GBRT is that each iterator is used to reduce the previous residual. To reduce these residuals, a new tree in the direction of the gradient descent of the loss function is created. After LSTM forms the training samples, the recursive form regression tree is as the equation to calculate $F_{g,m}(s_{t+1}')$ in Algorithm 1.

Before building the bilevel LSTM-EL prediction architecture, several hyperparameters should be determined. For the upstream model LSTM, the LSTM network with optimization of multiple hyperparameters has achieved acceptable performance when applied on sequence data [51]. For the time series problem, the key hyperparameters include the number of LSTM layers, the number of nodes in each LSTM layer, the number of fully connected layers, and the time-lags, and for ensemble learning algorithm, the number of trees and the maximum depth of a tree are the most essential parameters. In our work, the time-lag and embedding size are the most important hyperparameters.

(1) Time-lag: the time-lag parameter has a significant impact on the performance of time series forecasting [52], as it determines the length of the historical sequence that should be included for the training.

(2) Embedding size: that is, the number of neurons for the last layer in the LSTM network represents the input-data dimension of the downstream ensemble learning models and further determines the complexity of GBRT and RF. If the embedding size is very high, then the LSTM will be overfitting on training instances and increase the training difficulties of the downstream models, and if its size is too small, then it will be unable to memorize the time-dependency information collected from the time-lag sequences.

However, to the best of our knowledge, there are no general rules to choose the time-lag and the hidden layers' size. Therefore, we investigated the effect of key parameters

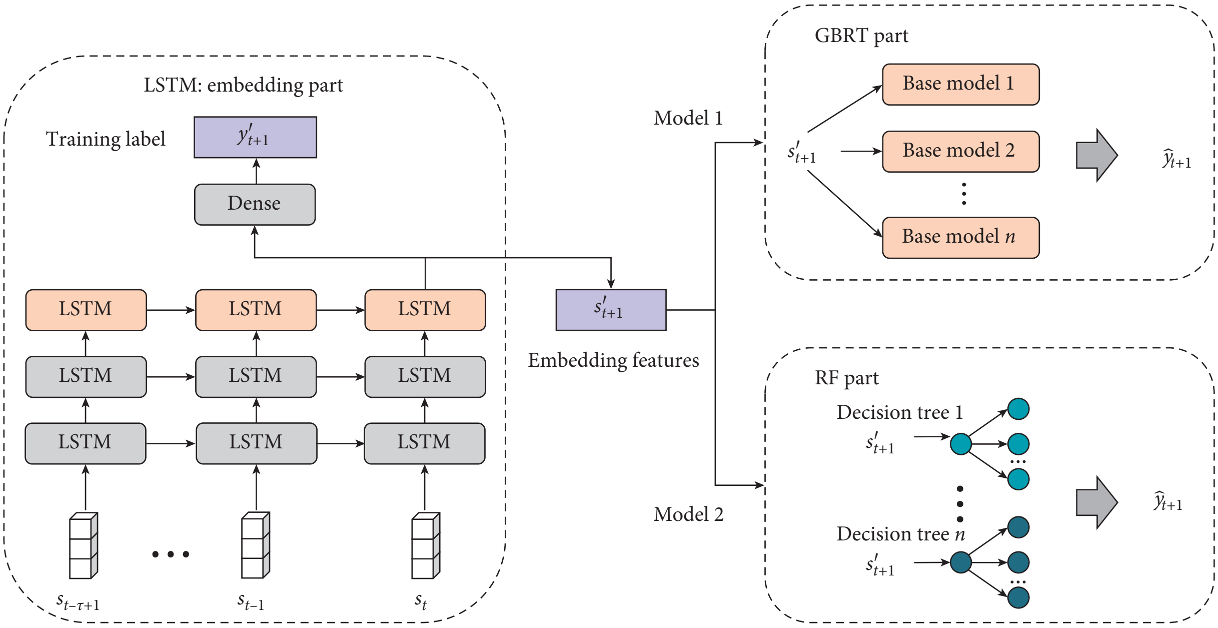


Figure 4: Structure diagram of LSTM-EL approach.

```
Input: historical observations: $\{X_1, \ldots, X_t, \ldots, X_T; Y_1, \ldots, Y_t, \ldots, Y_T\}$, $X_t = (x_t, x_{t-1}, \ldots, x_{t-\tau+1})$, $Y_t = (y_t, y_{t-1}, \ldots, y_{t-\tau+1})$, $X_t \in \mathbb{R}^{\tau\times D}$, $Y_t \in \mathbb{R}^{\tau\times 1}$, $x_t \in \mathbb{R}^D$, $y_t \in \mathbb{R}$; Length of input sequence (time-lag): $\tau$; Feature size: $D$ (i.e., 22 in this paper).
Output: learned LSTM-EL model.
//construct training instance
for all available time interval $t\,(1 \le t \le T)$ do:
    $S_t = (X_t; Y_t)$, $S_t \in \mathbb{R}^{\tau\times(D+1)}$//Embedding part- LSTM model to extract the features with time-dependent information.
Given a training instance $S_t$:
Step 1. Embedding features:
    $s'_{t+1} = \mathrm{LSTM}(S_t)$//embedding features produced by LSTM, which fused the temporal and internal correlations of $S_t$ into a new vector $s'_{t+1}$ with lower dimension, $s'_{t+1} \in \mathbb{R}^{d\prime}$ ($d\prime$ equals to the number of units in the last LSTM layer and the last dense layer only has one unit). $s'_{t+1}$ is then being used to predict $y_{t+1}$ in the second layer.
Step 2. LSTM prediction:
    $y'_{t+1} = \mathrm{Dense}(\mathrm{LSTM}(S_t))$//$y'_{t+1} \in \mathbb{R}$ is the prediction value given by LSTM part. Note that we just use $y'_{t+1}$ to optimize the parameters of the embedding part and the final prediction of our method is obtained by the downstream methods, GBRT and RF.
Step 3. Optimization:
    $\min_\theta \| y_{t+1} - y'_{t+1} \|_2^2$//The embedding part is trained by minimizing the objective function shown above and its parameters $\theta$ are updated via backpropagation. //GBRT or RF is taken as downstream method to make the final predictions.
//Model 1: GBRT part
Given the embedding features $s'_{t+1}$ generated from $S_t$:
Input: training set $\{s'_{t+1}\}_{t=1}^{T}$, differentiable loss function $L(y_{t+1}, F_g(s'_{t+1}))$, and the maximum number of trees $M$. //$F_g$ is the decision tree mapping function, and the optimal $F_g$ can be obtained through minimizing the loss function.
For $m = 1$ to $M$, do
    For $t = 1$ to $N$, do
        $F_{g,m}(s'_{t+1}) = F_{g,m-1}(s'_{t+1}) + \mathrm{lr} * \rho_m g_m(s'_{t+1})$. //$\rho_m$ is the step-size, $g_m$ is the base learner, and lr is the learning rate. For each step, a new decision tree is aimed at correcting the error made by its previous base learner.
Output model $\hat{y}_{t+1} = \mathrm{GBRT}(s'_{t+1})$//
//Model 2: RF part
Step 4: given the embedding features $s'_{t+1}$ generated from $S_t$
Input: training set $\{s'_{t+1}\}_{t=1}^{T}$, the number of trees in forest $M$
For $m = 1$ to $M$, do
    $D_k = \mathrm{Boostrap}(D)$//$D$ is the training set
    $h_k =$ Decision Tree with random feature selection $(D_k, f)$//$f$ denotes the number of attributes to use at each node, picked uniformly at random new features for every split
    Prune tree to minimize out-of-bag error
Average all $M$ Trees
Output model $\hat{y}_{t+1} = \mathrm{RF}(s'_{t+1})$.
```

Algorithm 1: LSTM-EL.

while keeping the other parameters fixed, and grid search [53] was applied to find the optimum hyperparameters. To do so, our model requires some basic settings. In our work, we built our model using two LSTM layers and one fully connected layer. The number of the neurons for the first layer is equal to that of the second layer. The search space of the above four critical parameters is illustrated in Table 3.

Table 3: Value specified for key hyperparameters.

Hyperparameters	Search range
Time-lag τ	[2, 4, 6, 8, 10]
Embedding size d'	[16, 32, 64, 128]
Number of trees	[1, 1000]
Depth of trees	[1, 50]

For each combination of the time-lag and embedding size, the LSTM-EL model is designed and trained, and the corresponding optimal combination of the number of trees and depth of trees is selected using the grid search. Here, for the sake of brevity, different combinations of the time-lag τ and the embedding size $d\prime$ are evaluated by root mean square error (RMSE) and mean absolute percentage error (MAPE) and the results are shown in Figures 5 and 6. According to previous studies, Li et al. [54] showed that a small time-lag cannot guarantee enough long-term memory inputs for our LSTM-EL model; thus, the model cannot fully exploit the LSTM for long-term memory modeling. Large time-lags permit an increased number of unrelated inputs, which increases the model's complexity and the difficulty of learning useful features. It can be observed that the effect of the number of nodes in each neuron layer shows that, with the increase in the number of neuron nodes, the prediction performance improves slightly. Thus, we set different time-lag τ and embedding size in the successive experiment to optimize both accuracy and time efficiency for both LSTM-GBRT and LSTM-RF, as indicated by the RMSE and MAPE. Based on the results of the experiments, Table 4 summarizes the best parameters of the obtained model.

4.4. Accuracy Measurement. For measuring the prediction precision pertaining to the developed approach, several evaluation criteria are implemented, such as the RMSE as well as the MAPE. Generally, with the decline of MAPE and RMSE, the approach will be more precise. However, it is well known that, for a given prediction, actual outcomes above and below the prediction are treated asymmetrically when using MAPE and RMSE [55]. For the mentioned reason, to measure the data fluctuations (e.g., upward, stable, or downward), direction matching rate (Dsta) is employed. In addition, we utilize mean absolute scaled error (MASE) to assess if the developed prediction approach outclasses the naïve prediction method [56]. For detailed description of MAPE, RMSE, Dsta, and MASE criteria, see Appendix B.

5. Empirical Results

5.1. Daily, Weekly, and Monthly CBCFI Forecasting. After determining the best network architecture for the prediction task, the training set was utilized to train our LSTM-EL model until convergence. Evaluations were conducted using the test set. To analyze the generality of the hybrid LSTM-EL structure, we use a dataset with day-to-day, week-to-week, and month-to-month bases. Specifically, the weekly and monthly data are calculated as the average of daily CBCFI. In addition, to avoid overfitting problem, early stopping and validation sets are utilized in the present study, and the percentages for training, testing, and validating sets are 60%, 20%, and 20%, respectively.

In our prediction of each forecasting approach, LSTM-RF, LSTM-GBRT, GBRT, and RF models are applied, all of which are evaluated by calculating MAPE, RMSE, Dsta, and MASE. A rolling approach is implemented to conduct a next-day/weekly/monthly CBCFI forecast. The approach uses the actual value of the predictor variable in the previous period for making a prediction in the testing set. Note that the time-lag is fixed, and new data are added for further $t+1$ prediction. Figure 7 displays how the rolling approach works.

Note that, in weekly forecasting, each point represents the weekly CBCFI value and a new weekly CBCFI value is calculated by every new 5 daily CBCFI values (only workdays data). Likewise, in monthly forecasting, each point represents the monthly CBCFI value and a new monthly CBCFI value is calculated by the working days in each new month, automatically excluding weekends (Saturday and Sunday).

We next conduct the daily, weekly, and monthly CBCFI forecasting experiments, respectively. The CBCFI data from January 2012 to October 2020 are sample data. To evaluate the predictive performance of LSTM-EL models, we split the data into training data, validating data, and testing data. The ratio for each dataset is 6:2:2. Figure 8 shows the learning curves of mean square error (MSE) for the validation data and training data for 100 epochs in daily, weekly, and monthly forecasting cases. The learning curves show a good fit because of the decrease in training and validation loss to one stable data with a minimal gap between the two final loss values.

Table 5 compares the predictive performance of two hybrid LSTM-EL approaches with the corresponding single EL approaches in the rolling forecasting approaches. For daily CBCFI forecasting cases, we found that the predictive performances for orientation matching and errors are enhanced through the introduction of the proposed hybrid structure. And all the values of MASE less than 1 indicate that four approaches outperform the average one-step naïve forecast. Note that the predictive performance in MAPE, RMSE, and Dsta is improved by considering hybrid structure forecasting. Compared to the predictive performances of corresponding single approaches, the improvement percentages (the improvement percentage is calculated by the difference between the evaluation index value of the hybrid model and the conventional model over that of the conventional model) in MAPE of LSTM-GBRT and LSTM-RF are 22.47% and 24.59%, respectively, of RMSE are 41.54% and 24.73%, respectively, and of Dsta are 72.84% and 69.36%, respectively. The above results indicate that LSTM-GBRT exhibits the most significant improvement level.

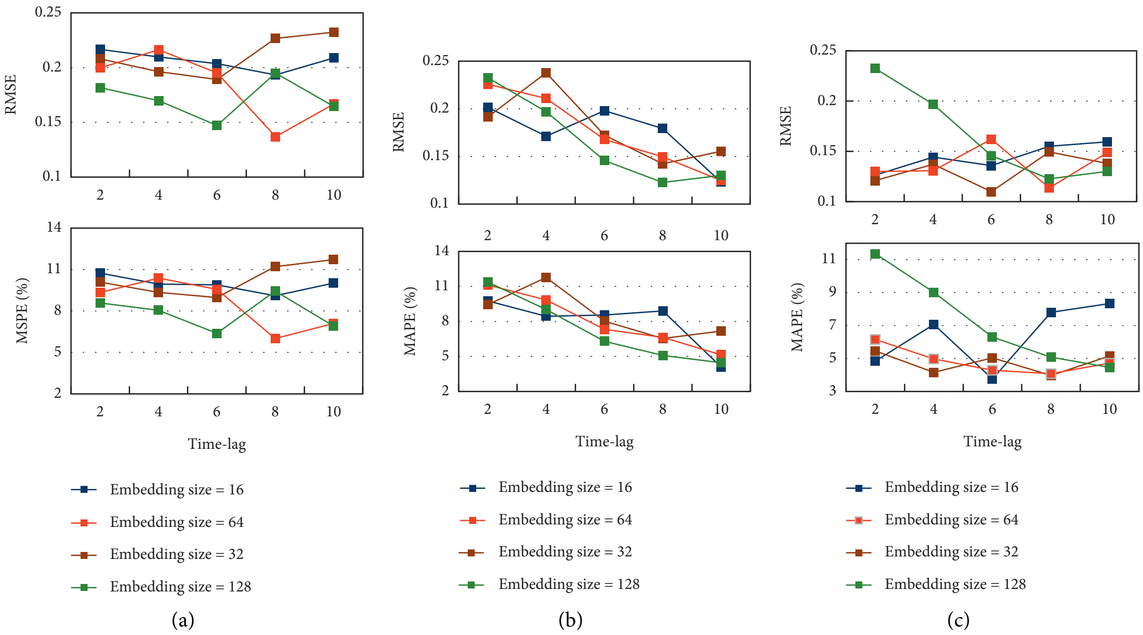


Figure 5: Key parameter sensitivity analysis for LSTM-GBRT under three time scales: (a) daily, (b) weekly, and (c) monthly CBCFI forecasting.

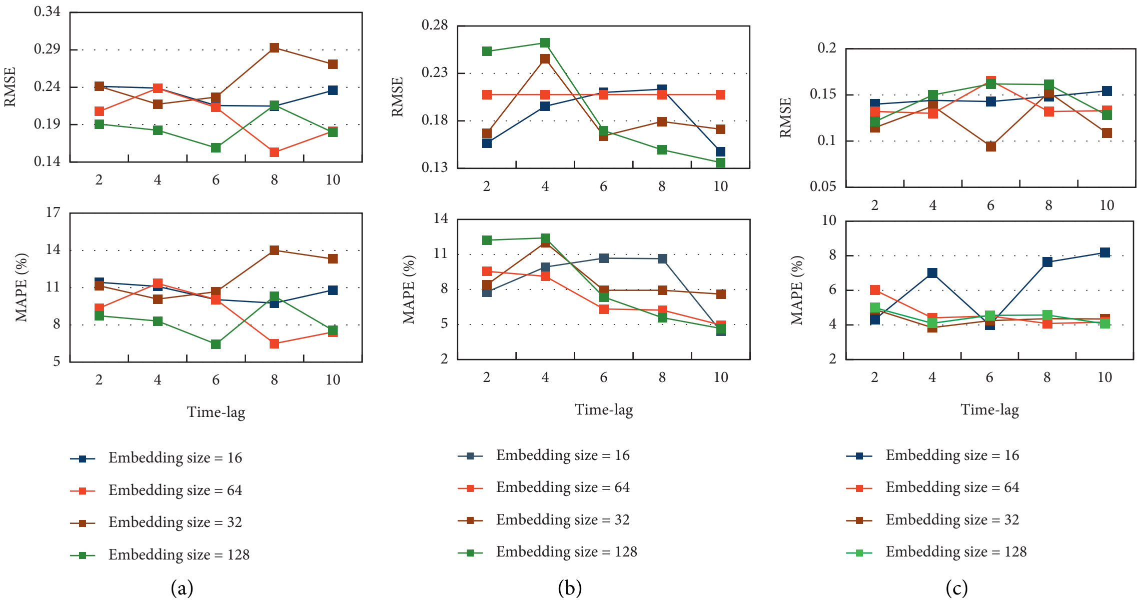


Figure 6: Key parameter sensitivity analysis for LSTM-RF under three time scales: (a) daily, (b) weekly, and (c) monthly CBCFI forecasting.

For weekly CBCFI forecasting cases, consistent with daily CBCFI predicting processes, performance enhancement rates are positive for both errors and Dsta. The improvement percentages in MAPE of LSTM-GBRT and LSTM-RF are 54.62%, 39.48%, respectively, of RMSE are 32.12% and 11.77%, respectively, and of Dsta are 50.91% and 67.17%, respectively. Moreover, the improvement in MAPE for all two LSTM-EL approaches shows higher significance for the weekly CBCFI predicting processes than the daily prediction. Thus, using a hybrid structure to extract the time-dependent characteristics between features in LSTM-based prediction enhances accuracy. In comparison with daily information, the hybrid approach shows significant improvements for the weekly CBCFI predicting process.

Since predicting long-term CBCFI with low-frequency time-scale data raises several challenges, this study examines how the hybrid approach promotes monthly data forecasting to be accurate. Consistent with daily and weekly forecasting, hybrid approaches outperform the single ensemble learning approaches. Notably, the hybrid structure promotes the

Table 4: The best configurations of the proposed model and benchmark models.

Algorithm	Best configurations	Forecasting cases		
		Daily	Weekly	Monthly
LSTM-GBRT	LSTM layer			
	Time-lag	8	10	6
	Number of hidden layers	2	2	2
	Number of units in the hidden layers	64	16	32
	GBRT layer			
	Number of trees	101	31	41
	Depth of trees	1	1	1
LSTM-RF	LSTM layer			
	Time-lag	8	10	6
	Number of hidden layers	2	2	2
	Number of units in the hidden layers	64	16	32
	RF layer			
	Number of trees	41	11	21
	Depth of trees	11	11	1
GBRT	Number of trees	191	11	11
	Depth of trees	21	1	1
RF	Number of trees	11	31	11
	Depth of trees	1	1	1
LSTM	Time-lag	8	10	8
	Number of hidden layers	2	2	2
	Number of units in the hidden layers	64	32	64

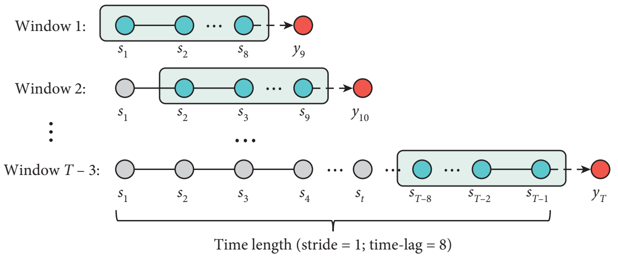


Figure 7: Illustration of rolling forecasting.

precision noticeably, MAPE improvement for GBRT and RF are 53.72% and 61.40%, respectively. RMSE improvements for LSTM-GBRT and LSTM-RF are 50.32% and 57.83%, respectively. Moreover, the MASE of the monthly data of hybrid approaches is less than 1, indicating that hybrid approaches in this approach outperform the average one-step naïve forecast.

Table 6 compares the general predictive performances of hybrid LSTM-EL approaches over three time scales. The $\overline{\text{MAPE}}$ and $\overline{\text{RMSE}}$ values indicate that hybrid LSTM-EL approaches perform best in weekly CBCFI forecasting and achieve the most obvious improvement of accuracy in monthly forecasting. And the performance indicators of different approaches (Table 5) show that the LSTM-GBRT approach outperforms LSTM-RF in daily forecasting while LSTM-RF achieves a higher accuracy in weekly and monthly forecasting, which indicates that LSMT-GBRT is better capable of dealing with high-frequency data while LSTM-RF is more suitable for mid- or low-frequency data.

Figure 9 shows the comparisons between the real CBCFI values and the predicted values produced by LSTM-EL approaches and their corresponding single EL approaches. The inset square of Figure 9(a) shows how the hybrid and single EL approaches perform at the uptrend, bottom, and downtrend. It is found that four approaches are capable of predicting the daily, weekly, and monthly tendency of CBCFI, but the outputs from LSTM-EL approaches derive less than from the actual CBCFI values. In the three situations, single EL predictions show obvious fluctuations that actual CBCFI values do not have. For example, in situation I beginning in January 2019, the GBRT and RF predictions display heavily up and down while they do not happen in the actual CBCFI trend, LSTM-GBRT and LSTM-RF still keep close to the real CBCFI values, and similar phenomena are observed in situations II and III.

Figures 9(b) and 9(c) show that single GBRT and RF models produce large errors in weekly and monthly forecasting, especially at the bottom or top of the trendline. In contrast, the hybrid LSTM-EL forecasting approach reproduces CBCFI trends and generates relatively small errors. For example, in weekly forecasting beginning in December 2019, the CBCFI dropped from approximately 1,000 points to a historic low of roughly 450 points, since the outbreak of COVID-19 hits the economy across the globe. And the CBCFI was kept at a low point until May 2020, despite the epidemic was gradually under control. In this case, only LSTM-GBRT and LSTM-RF predictions reproduce the CBCFI trend, while other approaches display large fluctuations not existing in the actual CBCFI itself. Likewise, in the monthly forecasting, the single EL approaches GBRT and RF predictions deviate, and the LSTM-GBRT and LSTM-RF predictions are very close to the actual CBCFI.

5.2. Diebold–Mariano (DM) Test.

To evaluate whether there is any statistically significant difference between the hybrid

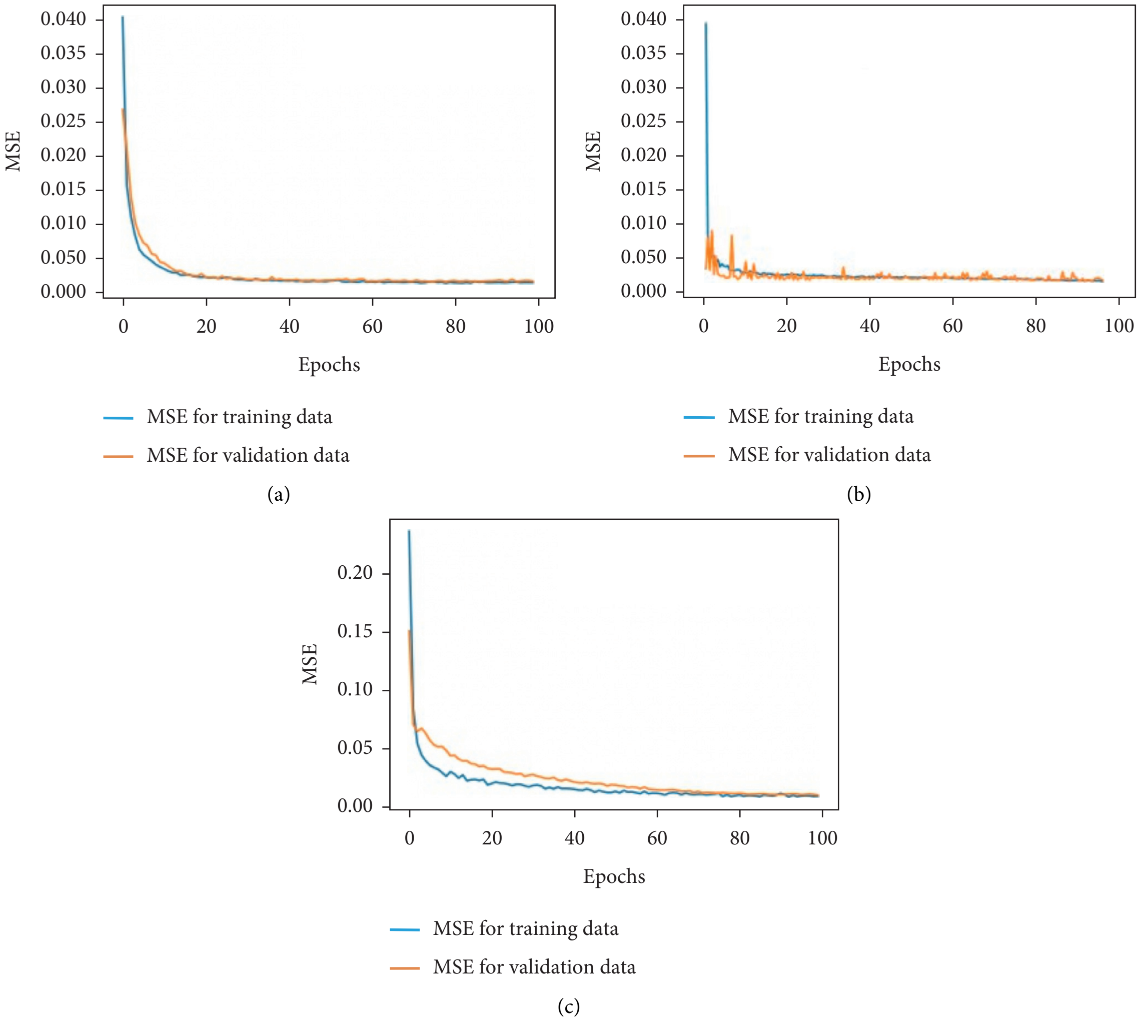


Figure 8: MSE for validation and training data for (a) daily, (b) weekly, (c) monthly forecasting cases.

and conventional models, the Diebold–Mariano (DM) test [57] is implemented to compare the testing-sample prediction results. The DM test is widely used in determining whether the differences of time series predicting accuracy by different models are substantially crucial from a statistical perspective [58]. Table 7 summarizes the results of the DW test for the daily, weekly, and monthly CBCFI datasets, respectively. p value less than 0.05 indicates the rejection of the null hypothesis that there is no difference between the two compared forecasting models. It can be seen that LSTM-EL models have significantly different accuracies when compared to other benchmarks. For the daily dataset, LSTM-RF and LSTM-GBRT present statistical differences in predictive performance when compared with other models but there is no statistical difference in accuracies between LSTM-RF and LSTM-GBRT. Similarly, in weekly and monthly forecasting cases, the predictive improvement offered by incorporating ensemble learning algorithms is statistically significant while the predictive improvements between different conventional ensemble learning algorithms or LSTM-based hybrid models do not display statistically predictive performances.

5.3. Feature Importance Analysis. Moreover, Figure 10 presents the features ranked by complying with the clarified variance the respective feature facilitates the LSTM-GBRT approach. In this case, the features are plotted against their relative importance, that is, the percent significance regarding the critical feature. For brevity, Figure 10 only presents the top 6 features with the sum of feature importance over 98%. It can be clearly seen that coal inventory at Tianjin Port influences the monthly CBCFI values but less important for daily and weekly data, which indicates that coal inventory is more likely to impact the long-term forecasting but not short- or midterm forecasting. In addition, domestic and overseas thermal coal spot rates and crude oil prices have obvious impacts on daily, weekly, and monthly CBCFI values, while coal inventory and fuel oil price are less important for daily and weekly CBCFI. Specifically, Guangzhou Port-Q5500 Index (Australian Coal)-EXT shows a great impact on daily, weekly, and monthly CBCFI values. Moreover, WTI crude oil spot price has an obvious impact on daily CBCFI values, while weekly CBCFI is more sensitive to Dubai crude oil spot price. This may result from these two crude oil indices that mainly serve

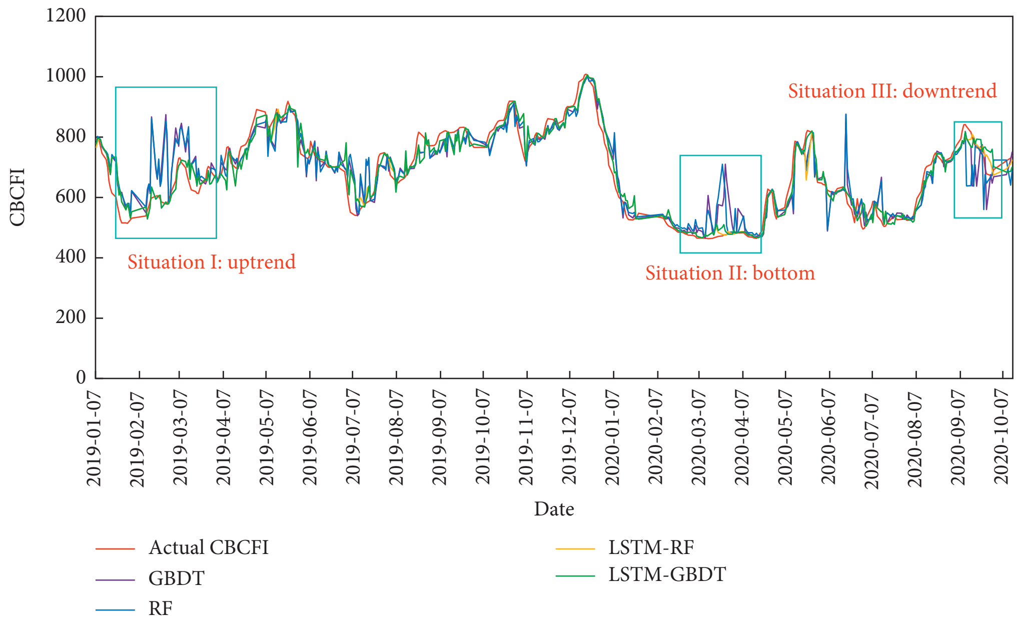
Table 5: Predictive performance of hybrid LSTM-EL and benchmark models for CBCFI prediction.

Methods	MAPE (%)	RMSE	MASE	Dsta
Daily CBCFI forecasting *Training period: 2012/1/4~2017/4/10* *Testing period: 2019/1/7~2020/10/13*				
LSTM	9.15	0.2419	1.1532	0.5703
GBRT	7.84	0.2342	0.8915	0.5145
LSTM-GBRT	6.00	0.1369	0.9132	0.8893
RF	8.62	0.2034	0.8370	0.5042
LSTM-RF	6.50	0.1531	0.9560	0.8539
Weekly CBCFI forecasting *Training period: 2012/1/9~2017/3/24* *Testing period: 2019/1/21~2020/9/28*				
LSTM	9.47	0.1931	1.6532	0.4303
GBRT	8.99	0.1815	0.6173	0.5324
LSTM-GBRT	4.08	0.1232	0.9835	0.8034
RF	7.32	0.1673	0.5525	0.4853
LSTM-RF	4.43	0.1476	0.8941	0.8113
Monthly CBCFI forecasting *Training period: 2012/1/4~2017/3/4* *Testing period: 2019/4/1~2020/10/13*				
LSTM	11.53	0.2571	1.0126	0.5863
GBRT	10.89	0.2212	0.7940	0.5932
LSTM-GBRT	5.04	0.1099	0.9203	0.9024
RF	11.01	0.2236	0.7864	0.6072
LSTM-RF	4.25	0.0943	0.8825	0.9043

Table 6: Predictive performance of hybrid LSTM-EL over three time scales.

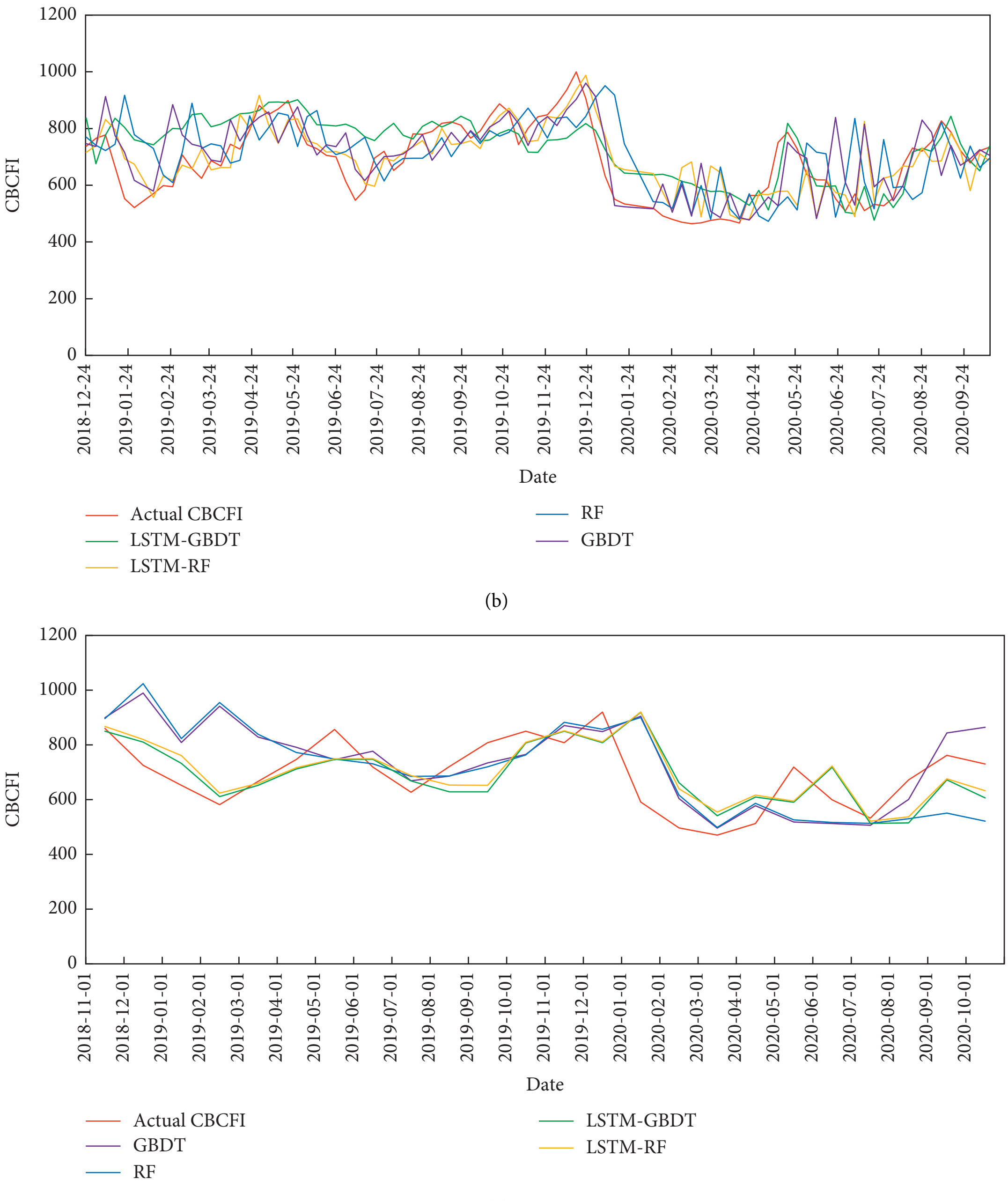
	Forecasting accuracy			Improvement percentage		
	Daily forecasting	Weekly forecasting	Monthly forecasting	Daily forecasting	Weekly forecasting	Monthly forecasting
$\overline{\text{MAPE}}$	6.25%	4.26%	4.65%	23.53%	47.05%	57.56%
$\overline{\text{RMSE}}$	0.1450	0.1354	0.1021	33.14%	21.95%	54.08%

Note. $\overline{\text{MAPE}}$ and $\overline{\text{RMSE}}$ represent the average values of MAPE and RMSE. Each average MAPE/RMSE is calculated by the average MAPE/RMSEs of LSTM-GBRT and LSTM-RF from Table 5.



(a)

Figure 9: Continued.



(b)

(c)

Figure 9: Actual CBCFI and forecasting results through LSTM-EL approaches and corresponding EL approaches for (a) daily, (b) weekly, and (c) monthly data.

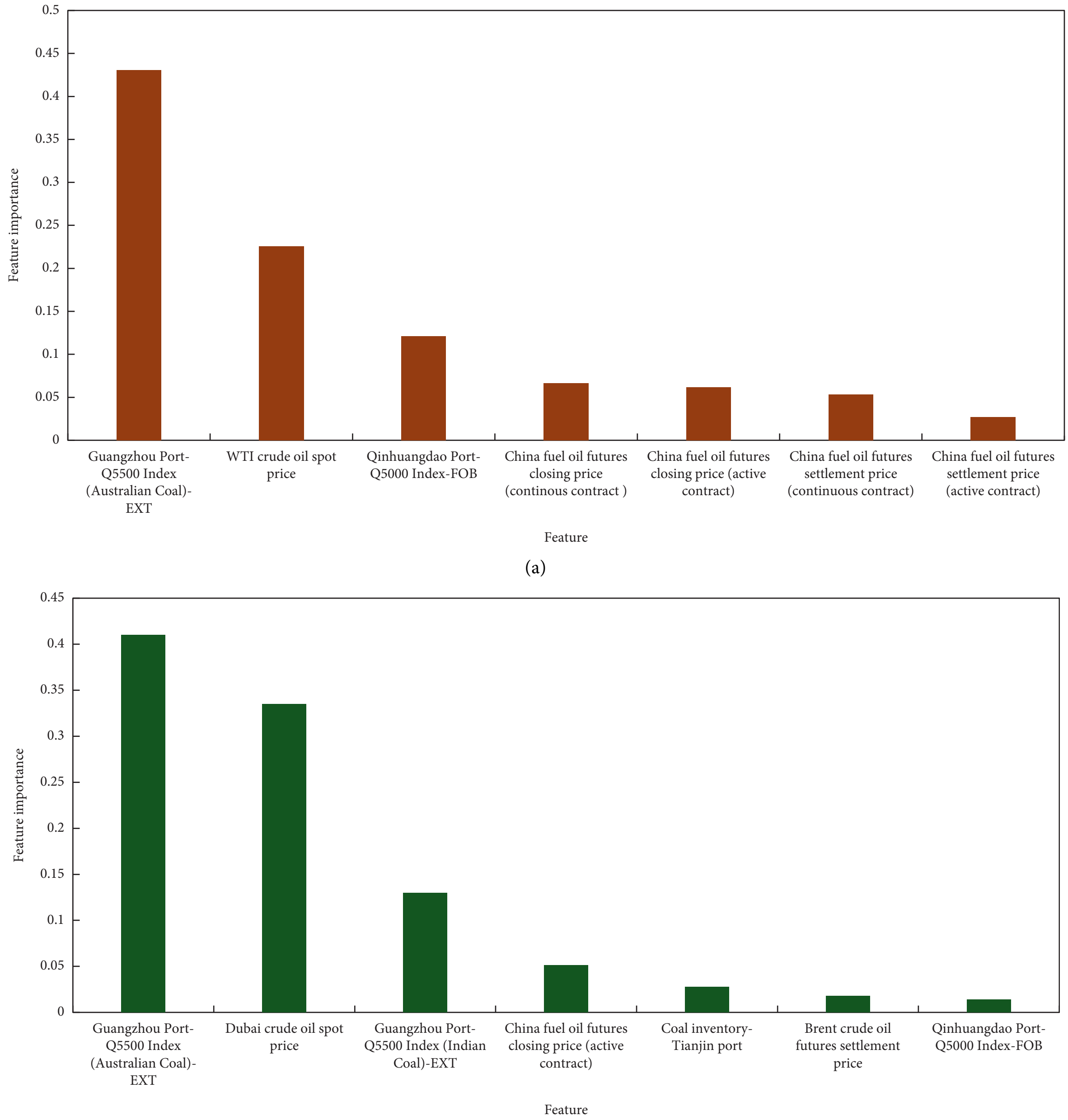
Table 7: DM test results for hybrid models and the benchmarks.

Data type	Tested model	Reference model				
		LSTM	GBRT	RF	LSTM-RF	LSTM-GBRT
Daily CBCFI forecasting	LSTM	—				
	GBRT	2.4328**	—			
	RF	2.4103**	−1.6784	—		
	LSTM-RF	2.5123**	2.2341**	2.1231**	—	
	LSTM-GBRT	2.2763**	2.2910**	2.8723**	−1.2432	—

Table 7: Continued.

Data type	Tested model	Reference model				
		LSTM	GBRT	RF	LSTM-RF	LSTM-GBRT
Weekly CBCFI forecasting	LSTM	—				
	GBRT	−2.7084**	—			
	RF	2.2034**	−1.6110	—		
	LSTM-RF	2.2361**	2.3414**	2.6535**	—	
	LSTM-GBRT	2.2276**	2.3012**	2.5541**	2.6287	—
Monthly CBCFI forecasting	LSTM	—				
	GBRT	−4.5123**	—			
	RF	−5.2341**	−1.9883	—		
	LSTM-RF	−5.2011**	−6.1094**	−5.3312**	—	
	LSTM-GBRT	−5.1998**	−6.2014**	−6.4234**	−2.6536	—

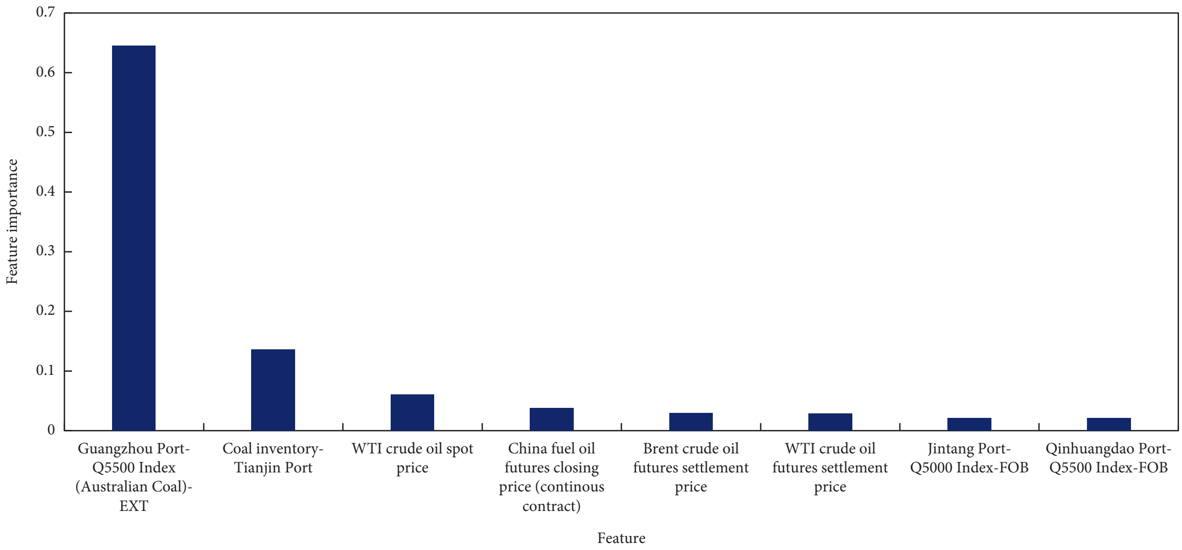
Note. **The value is significant at 5%.



(a)

(b)

Figure 10: Continued.



(c)

Figure 10: Feature importance for (a) daily, (b) weekly, and (c) monthly CBCFI forecasting cases.

different regions. China, the world's biggest oil importer, has ramped up purchases of American crude for years [59]. WTI crude oil refers to oil extracted from wells in the US and sent via pipeline to Cushing, Oklahoma, which is the main benchmark for oil consumed in the United States [60]. Crude oil as a substitute for coal and thus Chinese domestic crude oil price would be more sensitive to the daily changes of the US crude oil prices. In addition, because crude oil is a substitute for coal, the fluctuation of coal price will influence the bulk coal shipping cost somehow, and thus the CBCFI values will be sensitive to the WTI crude oil prices in the short time scale. On the other hand, as Dubai crude oil price index is the main reference for Persian Gulf oil delivered to the Asian market and roughly half (44.8%) of Chinese imported crude oil originates from nine Middle Eastern nations [61], the CBCFI index will be affected by the fluctuation of Dubai crude oil price as well. Moreover, Nanovsky [62] introduced an oil price-distance interaction variable to explain how global trade behaves as a result of oil price changes. He found that when oil prices increase, international trade becomes more localized in that countries begin trading relatively more with their neighbors. In contrast, when they decrease, trade becomes more dispersed in that the distance between countries becomes less relevant. Thus, the price of crude oil usually presents volatility in one week, and thus, the weekly Chinese coastal shipping cost may look at the price of the Asian crude oil market more.

5.4. The Impact of the Supply and Demand on CBCFI. As CBCFI is an index for shipping price, in addition to the factors discussed above, supply and demand are essential factors. For CBCFI, the supply should be available bulk fleet and the demand should be the amount of coal that needs to be shipped.

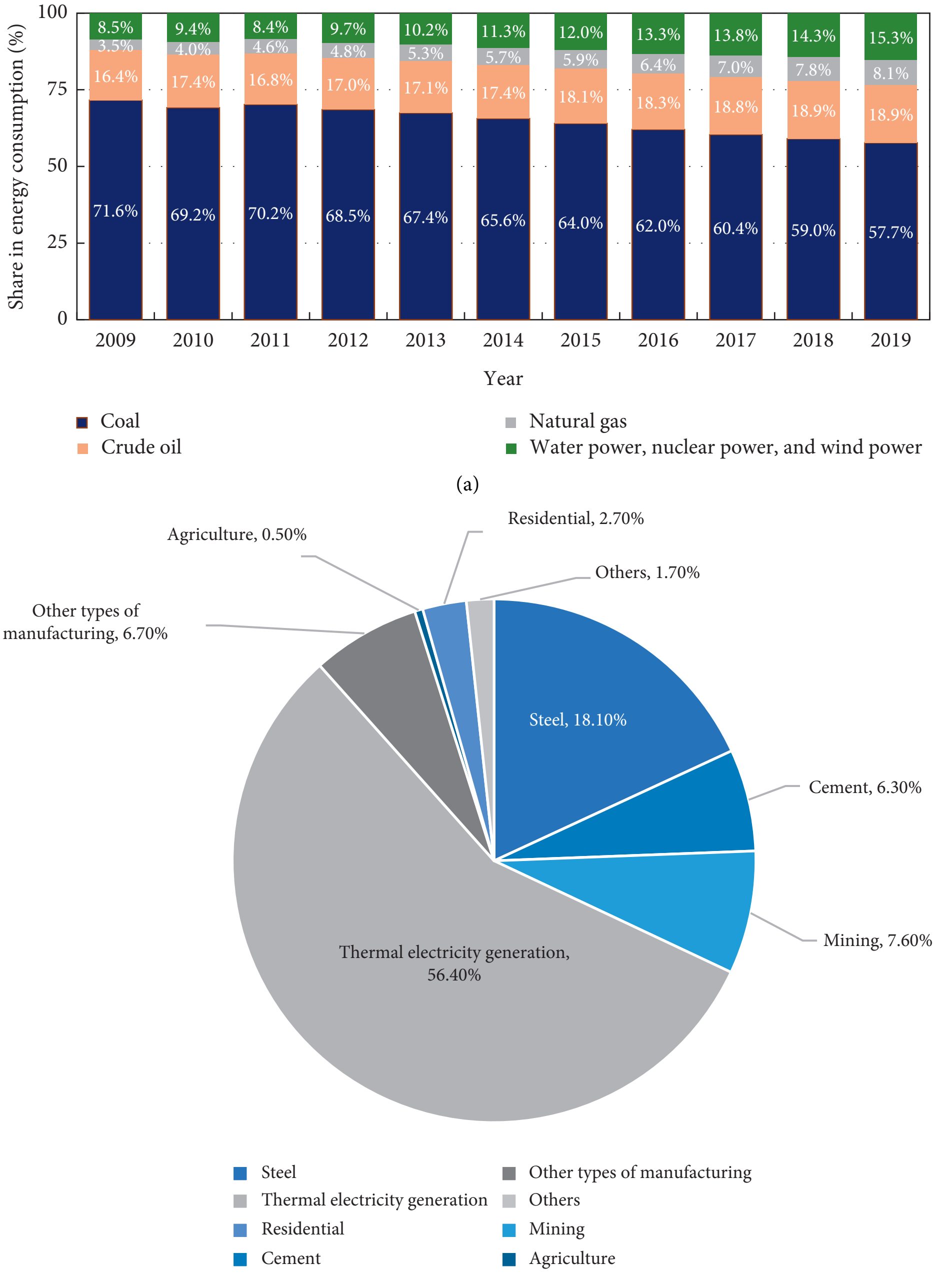
On the supply side, specifically, according to the routes and ports of CBCFI, the supply should be the available bulk fleet at Tianjin Port, Jintang Port, Qinhuangdao Port, Caofeidain Port, and Huanghua Port. However, we did not find any available open-source fleet data.

On the demand side, specifically, the CBCFI represents the coal transportation need mainly from northern China to southern China. The national statistics (Figure 11) indicate that most of the coal was used for thermal power generation (56.40%) and steel-making (18.1%). Therefore, the southern thermal power generation and steel production are considered as indexes of the CBCFI-related demand. However, we only found the above data on a nationwide basis and cannot get the data only for southern China. To estimate the demand, we use the sum of utility electricity consumption of the southern coastal provinces with at least one port city included in the CBCFI routes, including Shanghai, Jiangsu, Zhejiang, Fujian, and Guangdong, as a substitution to the southern thermal power generation. For the steel production, we use the total domestic steel production as a substitution index of southern provinces' steel production.

Besides, the coal sources in China consist of 90% self-produced mainly from northern China and 10% coal imports [47]. The coal production and coal imports also may have impacts on the CBCFI-related demand. Specifically, both the coal production increase and coal imports decrease may increase the CBCFI-related demand. Therefore, domestic coal production and coal imports are also considered as the demand-related factors.

Note that the above demand and demand-related factors are only available on a monthly basis. Accordingly, southern utility electricity consumption, domestic steel production, domestic coal production, and coal imports are added to the monthly forecasting model.

Table 8 displays the improvement percentage of monthly predictive performance with adding demand and demand-related factors. It is found that the predictive performances of different models have positive improvements with adding



(a)

(b)

Figure 11: Energy consumption in China. (a) China's energy consumption by source (data source: Samantha. W. Energy consumption in China from 2009 to 2019, by source. Available at https://www.statista.com/statistics/278669/energy-consumption-in-china-by-source/); (b) China's coal consumption by end market in 2020 (Data source: National Bureau of Statistics. Available at http://www.stats.gov.cn/tjsj/).

demand and demand-related factors, particularly an obvious improvement in MAPE (12.67%) and RMSE (10.01%) of LSTM. And for hybrid models, the improvement percentages in MAPE of LSTM-GBRT and LSTM-RF are 5.32% and 5.16%, respectively, and of RMSE are 5.18% and 5.01%, respectively. Figure 12 presents the top 7 features with the sum of feature importance over 98%. Southern utility electricity consumption ranks the top, followed by Guangzhou Port-Q5500 Index (Australian Coal)-EXT, domestic steel consumption, and coal inventory at Tianjin Port. As not all self-produced coal needs to be transported, domestic coal production shows little importance for CBCFI forecasting. The above results illustrate that demand factors could lead to a higher monthly predictive accuracy.

Table 8: Improvement percentage of predictive performance with adding demand and demand-related factors.

Evaluation criteria	Models	Improvement percentage of monthly forecasting (%)
MAPE	LSTM	12.67
	GBRT	9.97
	LSTM-GBRT	5.32
	RF	9.81
	LSTM-RF	5.16
RMSE	LSTM	10.01
	GBRT	8.90
	LSTM-GBRT	5.18
	RF	7.87
	LSTM-RF	5.01

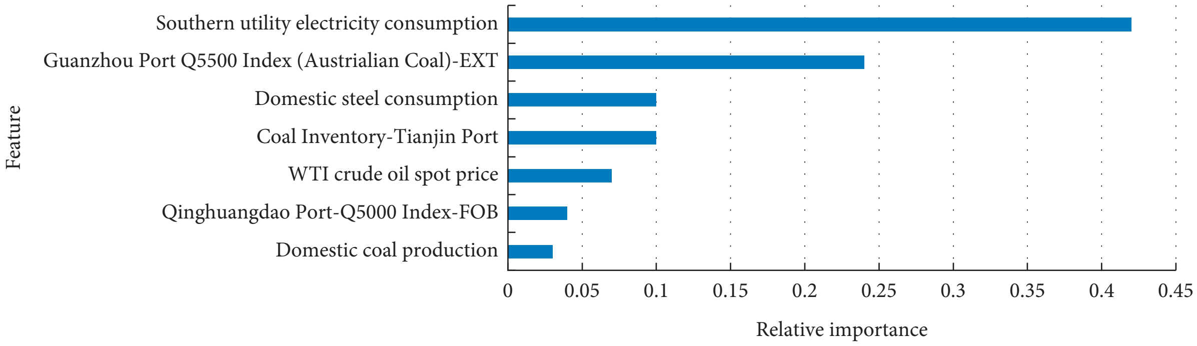


Figure 12: Feature importance for CBCFI monthly forecasting with adding demand and demand-related factors.

6. Conclusion

The present study attempted at enhancing the forecasting accuracy of the CBCFI by formulating a novel hybrid LSTM-EL approach, which is capable of extracting the useful time-dependent information in the data by combining the LSTM technique and ensemble learning algorithms. A rolling forecasting approach is developed for assessing LSTM-EL's forecasting accuracy in comparison with its corresponding single ensemble algorithms. Furthermore, critical factors that influence CBCFI values are discussed and experiments under daily, weekly, and monthly time scales in the rolling forecasting approach are conducted in order to test the performance generality of LSTM-EL approaches.

The major intellectual advantages here consist of the emerging method by exploiting artificial neural network and ensemble learning methods to be the useful approach to obtain the shipping freight market's nonlinear and non-stationary features. According to the empirically achieved outcomes, domestic and overseas thermal coal spot rates and crude oil prices have obvious impacts on daily, weekly, and monthly CBCFI values, while coal inventory and fuel oil price are less important for daily and weekly CBCFI. In terms of forecasting accuracy, LSTM-EL approaches outperform the single EL models in three time-scale forecasting cases and generate better results than the naïve forecasts. Moreover, the accuracy improvement by LSTM-EL approaches for different CBCFI time-scale datasets is validated. Results indicate that hybrid LSTM-EL approaches perform best in daily CBCFI forecasting but achieve the most obvious improvement of accuracy in weekly forecasting. In addition, a DM test is implemented to evaluate whether there is any statistically significant difference between the hybrid and conventional models, and the results illustrate that LSTM-based hybrid models present statistical difference in predictive performance when compared with other models but there is no statistical difference in accuracies between LSTM-RF and LSTM-GBRT, so do the weekly and monthly forecasting cases. Overall, the LSTM-EL method has a high prospect to predict the CBCFI index in an accurate manner.

The mentioned emerging method is capable of acting to be one effective tool to make the decisions regarding chartering and shipping based on uncertain properties and further being incorporated into management toolkit by shipping industry practitioners. The developed method and outcomes widen freight rates forecasting study and indicate probable subsequent study in relevant aspects fields.

Appendix

A. Basic Model Theory

The deep-learning approach exhibits one prominent performance as opposed to the conventional statistics-related approach since it is capable of mapping the initial information for a nonlinear approach, which generates more effective influence. And long short-term memory (LSTM) based on the concept of recurrent neural network (RNN) presents an outstanding ability in time series predictions. On the other hand, ensemble learning methods refer to machine learning technique that combines several bases approaches in order to minimize the causes of error in learning approaches, such as noise, bias, and variance, for improving the overall predictive performance of the approach. In this paper, two prevailing approaches are focused on, (i) Random Forest (RF) and (ii) gradient boosting regression tree (GBRT).

(1) Long short-term memory (LSTM)

LSTM, developed by Hochreiter and Schmidhuber, refers to on e special type of recurrent neural networks (RNN) [63]. LSTM consists of a unique set of memory cells that trains the data through a time back-propagation algorithm, capable of solving the long-term dependence problems of RNN, and thus is suitable for time series problems. The schematic diagram of the LSTM approach is displayed in Figure 13.

LSTM has options for adding or deleting the data of its cell condition, as achieved with the use of cell gates. The standard LSTM can be expressed as follows. The respective step t and its corresponding input sequence are denoted as $X = \{x_1, x_2, \ldots, x_t\}$, and the three types of gates are input gate i_t, output gate o_t, and forget gate f_t. The passed information can be determined whether to be remembered or forgotten by the output of the hidden layer h_{t-1} and input x_t of the current layer. The activation function for limiting the outputted data inside the range of [0, 1] is as follows:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right). \tag{A.1}$$

The input gate aims at determining the appropriate input information (h_{t-1}, x_t) to the cell. Its activated function is to set the forgetting gate. $\widetilde{C}_t$ denotes a "candidate" hidden state determined from the previous hidden state and the current input. C_t expresses the internal memory of the unit. For achieving the purpose of retaining the corresponding information, C_t integrates the previous memory, under the multiplication from the gate that is forgotten, and the newly hidden state below under the multiplication from the input gate.

$$\begin{aligned} i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \\ \widetilde{C}_t &= \tan h(W_c \cdot [h_{t-1}, x_t] + b_c), \\ C_t &= f_t \cdot C_{t-1} + i_t \cdot \widetilde{C}_t. \end{aligned} \tag{A.2}$$

The output gate controls the data able to be outputted. Likewise, the activation function aims at setting the gate this is forgotten. After the memory cell state gets updated by the $\tan h$ activation function, the multiplication of dot determines the to-be-output information.

$$\begin{aligned} o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), \\ h_t &= o_t \circ \tan h(C_t), \end{aligned} \tag{A.3}$$

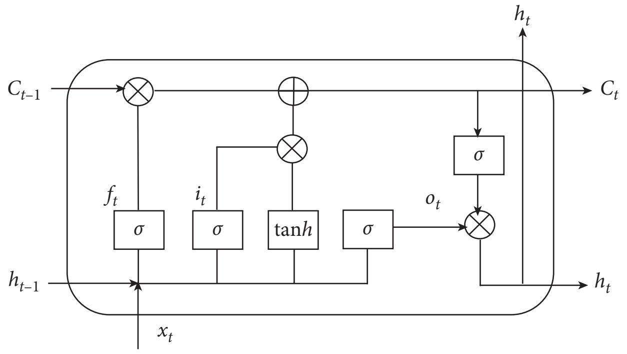


Figure 13: LSTM structure diagram.

where W_i, W_f, and W_o denote the matrices of weight and b_i, b_f, b_c, and b_o express biased vectors.

(2) Random Forest (RF)

RF expresses one combined learning algorithm containing decision trees, introduced by Breiman [40]. Thus, each tree of RF receives the training in a separate manner on an independent training set under the selection in a random manner. As opposed to the conventional decision tree, RF superiority has the reflection within two aspects. On the one hand, the trees built show inconsistency as generated in various training sets of a bootstrap subsampling and various random subdivided sets pertaining to characteristics for the split in terms of the respective tree node. On the other hand, the subsets of features are selected randomly. In this case, RF can achieve both low bias and low variance output and is not easy to fall into overfitting [64]. For regression, the final prediction is the average of the predictions from the set of decision trees. In RF, out-of-bag (OOB) error rate generally serves to be one way for measuring evaluation indices' significance [64].

(3) Gradient boosting regression tree (GBRT)

In terms of an established set D with n examples and m features $D = \{(x_i, y_i)\}$ $(|D| = n,\ x_i \in \mathfrak{R}^m,\ y_i \in \mathfrak{R})$, a tree ensemble approach applies K additive functions for predicting the output.

$$\widehat{y}_i = \phi(x_i) = \sum_{K=1}^{K} f_k(x_i), \quad f_k \in \Gamma, \tag{A.4}$$

where $\Gamma = \left\{f(x) = \omega_{q(x)}\right\} (q\colon \mathfrak{R}^m \longrightarrow T,\ \omega \in \mathfrak{R}^T)$ denotes the regression tree space. Specifically, q denotes the configuration of the respective tree mapping one instance to the relevant leaf index, T expresses the number of leaves in the tree, and the respective f_k represents one single tree structure q and leaf weight ω. Inconsistent with decision trees, the respective regression tree covers one continuous score on the respective leaf, and ω_i is used for representing the score on the ith leaf.

B. Evaluation Criteria

Model performance evaluation criteria MAPE and RMSE have the following formulation:

$$\begin{aligned} \text{MAPE} &= \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y_r - Y_p}{Y_r}\right|, \\ \text{RMSE} &= \sqrt{\frac{\sum_{t=1}^{n}\left(Y_r - Y_p\right)^2}{n}}, \end{aligned} \tag{B.1}$$

where n denotes the size pertaining to the dataset under the testing process and Y_r and Y_p represent the actual and forecasted data at time t, separately. Generally, the lower the RMSE and MAPE data are, the more accurate the approach will be. Nevertheless, specific to a set predicting process, practical results over and under the forecast receive the asymmetrical treat when using MAPE and RMSE [55]. Thus, direction matching rate (Dsta) is used to measure the data fluctuations (e.g., upward, stable, or downward), which is defined by

$$\begin{aligned} \text{Dsta} &= \frac{1}{n}\sum_{t=1}^{n}u(t), \\ u(t) &= \begin{cases} 1, & \left(Y_r(t+1) - Y_r(t)\right) * \left(Y_p(t+1) - Y_r(t)\right) \geq 0, \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{B.2}$$

The range of Dsta value is [0, 1]. The more approaching the Dsta data is to 1, the greater the precision pertaining to the direction-related predicting process concerned with the approach will be, and vice versa.

In addition, we utilize mean absolute scaled error (MASE) for assessing whether the proposed predicting approach outperforms naïve forecasting method [56].

$$\text{MASE} = \text{mean}\left|\frac{\left|e_t\right|}{(1/(N-1))\sum_{i=2}^{N}\left|Y_p - Y_{p-1}\right|}\right|, \tag{B.3}$$

where e_t denotes the prediction error determined to be $(Y_r - Y_p)$ and $(Y_p - Y_{p-1})$ refers to the naïve forecast's error. That is, MASE less than 1 indicates that the approach generates a more effective prediction than the calculated naïve predictions.

Data Availability

The data applied here originate from the public Wind Economic Database, IEA, National Bureau of Statistics of China, and Statista, and the data could receive the assessment from https://www.wind.com.cn/en/edb.html, https://www.iea.org/data-and-statistics, http://www.stats.gov.cn/tjsj/, and https://www.statista.com/statistics, respectively.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was financially supported by the Key Research and Development Plan of the Ministry of Science and Technology, China (no. 2018YFB1601402), the Energy Foundation (project: Preliminary Investigation and Research on Freight Industry, no. VQ99QT1136Y17005), and the Open Research Fund for National Engineering Laboratory of Integrated Transportation Big Data Application Technology (no. CTBDAT201912).

References

[1] M. Stopford, "How shipping has changed the world and the social impact of shipping," *Global Maritime Environmental Congress*, vol. 7, pp. 1–10, 2010.

[2] SSE, "About CBCFI," 2020, https://en.sse.net.cn/indices/introduction_cbcfi_new.jsp.

[3] SSE, *Popularization of CBFI*, SSE, Perth, UK, 2020.

[4] X. Ding, S. Dai, F. Chen et al., "Long memory and scaling behavior study of bulk freight rate volatility with structural breaks," *Transportation Letters*, vol. 10, no. 6, pp. 343–353, 2018.

[5] Y. Kou, M. Luo, and Y. Zhao, "Examining the theoretical-empirical inconsistency on stationarity of shipping freight rate," *Maritime Policy & Management*, vol. 45, no. 2, pp. 145–158, 2018.

[6] J. Liu and F. Chen, "Asymmetric volatility varies in different dry bulk freight rate markets under structure breaks," *Physica A: Statistical Mechanics and Its Applications*, vol. 505, pp. 316–327, 2018.

[7] J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, "Recent advances and applications of machine learning in solid-state materials science," *NPJ Computational Materials*, vol. 5, no. 1, p. 83, 2019.

[8] D. Wolff and U. Neugebauer, "Tree-based machine learning approaches for equity market predictions," *Journal of Asset Management*, vol. 20, no. 4, pp. 273–288, 2019.

[9] J. Yang, C. Zhao, H. Yu, and H. Chen, "Use GBDT to predict the stock market," *Procedia Computer Science*, vol. 174, pp. 161–171, 2020.

[10] S. Yang, J. Wu, Y. Du, Y. He, and X. Chen, "Ensemble learning for short-term traffic prediction based on gradient boosting machine," *Journal of Sensors*, vol. 2017, Article ID 7074143, 15 pages, 2017.

[11] X. Zhang, M. Y. Chen, M. G. Wang, Y. E. Ge, and H. E. Stanley, "A novel hybrid approach to Baltic dry index forecasting based on a combined dynamic fluctuation network and artificial intelligence method," *Applied Mathematics and Computation*, vol. 361, pp. 499–516, 2019.

[12] BalticExchange, "BDI overview," 2021, https://www.balticexchange.com/en/index.html.

[13] W. Hansen, "Understanding the Baltic dry index. Learning markets," 2021, https://www.learningmarkets.com/understanding-the-baltic-dry-index/.

[14] K. P. B. Cullinane, K. J. Mason, and M. Cape, "A comparison of models for forecasting the baltic freight index: box-Jenkins revisited," *International Journal of Maritime Economics*, vol. 1, no. 2, pp. 15–39, 1999.

[15] M. G. Kavussanos and A. H. Alizadeh-M, "Seasonality patterns in dry bulk shipping spot and time charter freight rates," *Transportation Research Part E: Logistics and Transportation Review*, vol. 37, no. 6, pp. 443–467, 2001.

[16] R. Batchelor, A. Alizadeh, and I. Visvikis, "Forecasting spot and forward prices in the international freight market," *International Journal of Forecasting*, vol. 23, no. 1, pp. 101–114, 2007.

[17] V. Tsioumas, S. Papadimitriou, Y. Smirlis, and S. Z. Zahran, "A novel approach to forecasting the bulk freight market," *The Asian Journal of Shipping and Logistics*, vol. 33, no. 1, pp. 33–41, 2017.

[18] S. Chen, H. Meersman, and E. Voorde, "Forecasting spot rates at main routes in the dry bulk market," *Maritime Economics & Logistics*, vol. 14, 2012.

[19] R. Adland, F. E. Benth, and S. Koekebakker, "Multivariate modeling and analysis of regional ocean freight rates," *Transportation Research Part E: Logistics and Transportation Review*, vol. 113, pp. 194–221, 2018.

[20] M. Stopford, *Maritime Economics*, Routledge, Evanston, IL, USA, 2013.

[21] S. Dai, F. Chen, Y. Zeng, and X. Zeng, "Scaling behavior of bulk freight rate volatility before and after noise reduction," *Journal of Shanghai Jiaotong University (Science)*, vol. 21, no. 6, pp. 655–661, 2016.

[22] J. Li and M. G. Parsons, "Forecasting tanker freight rate using neural networks," *Maritime Policy & Management*, vol. 24, no. 1, pp. 9–30, 1997.

[23] H. Yang, F. Dong, and M. Ogandaga, "Forewarning of freight rate in shipping market based on support vector machine," in *Proceedings of the 6th International Conference of Traffic and Transportation Studies Congress (ICTTS)*, Nanning, China, August 2008.

[24] E. Thalassinos, M. Hanias, P. Curtis, and J. Thalassinos, *Marine Navigation and Safety of Sea Transportation: STCW, Maritime Education and Training (MET), Human Resources and Crew Manning, Maritime Policy, Logistics and Economic Matters*, CRC Press, Boca Raton, FL, USA, 2013.

[25] F. Guan, Z. Peng, K. Wang, X. Song, and J. Gao, "Multi-step hybrid prediction model of Baltic supermax index based on support vector machine," *Neural Network World*, vol. 26, no. 3, pp. 219–232, 2016.

[26] B. Şahin, S. Gürgen, B. Ünver, and İ. Altin, "Forecasting the Baltic dry index by using an artificial neural network approach," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 26, pp. 1673–1684, 2018.

[27] D. M. Q. Nelson, A. C. M. Pereira, and R. A. d. Oliveira, "Stock market's price movement prediction with LSTM neural networks," in *Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 1419–1426, Anchorage, AK, USA, May 2017.

[28] C. Wu, C. Lu, Y. Ma, and R. Lu, "A new forecasting framework for bitcoin price with LSTM," in *Proceedings of the 2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 168–175, Singapore, November 2018.

[29] Z. Zhao, W. Chen, X. Wu, P. Chen, and J. Liu, "LSTM network: a deep learning approach for short-term traffic forecast," *IET Intelligent Transport Systems*, vol. 11, pp. 68–75, 2017.

[30] Y. Duan, L. V. Yisheng, and F.-Y. Wang, "Travel time prediction with LSTM neural network," in *Proceedings of the 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, Rio de Janeiro, Brazil, November 2016.

[31] Y. Leonov and V. Nikolov, "A wavelet and neural network model for the prediction of dry bulk shipping indices," *Maritime Economics & Logistics*, vol. 14, no. 3, pp. 319–333, 2012.

[32] E. Bulut, O. Duru, and S. Yoshida, "A fuzzy integrated logical forecasting (FILF) model of time charter rates in dry bulk shipping: a vector autoregressive design of fuzzy time series with fuzzy c-means clustering," *Maritime Economics & Logistics*, vol. 14, no. 3, pp. 300–318, 2012.

[33] Q. Zeng, C. Qu, A. K. Y. Ng, and X. Zhao, "A new approach for Baltic dry index forecasting based on empirical mode decomposition and neural networks," *Maritime Economics & Logistics*, vol. 18, no. 2, pp. 192–210, 2016.

[34] K. Uyar, Ü. ilhan, and A. İlhan, "Long term dry cargo freight rates forecasting by using recurrent fuzzy neural networks," *Procedia Computer Science*, vol. 102, pp. 642–647, 2016.

[35] I. M. Kamal, H. Bae, S. Sunghyun, and H. Yun, "DERN: deep ensemble learning model for short- and long-term prediction of baltic dry index," *Applied Sciences*, vol. 10, no. 4, p. 1504, 2020.

[36] D. R. Liu, S. J. Lee, Y. Huang, and C. J. Chiu, "Air pollution forecasting based on attention-based LSTM neural network and ensemble learning," *Expert Systems*, vol. 37, Article ID e12511, 2020.

[37] M. Tan, S. Yuan, S. Li, Y. Su, H. Li, and F. H. He, "Ultra-short-term industrial power demand forecasting using LSTM based hybrid ensemble learning," *IEEE Transactions on Power Systems*, vol. 35, no. 4, pp. 2937–2948, 2020.

[38] L. Duan and M. Binbasioglu, "An ensemble framework for community detection," *Journal of Industrial Information Integration*, vol. 5, pp. 1–5, 2017.

[39] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[40] L. Breiman, "Random forests," *Machine LearningMachine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[41] Wind https://www.wind.com.cn/en/default.html 2021.

[42] C. M. Jarque and A. K. Bera, "Efficient tests for normality, homoscedasticity and serial independence of regression residuals," *Economics Letters*, vol. 7, no. 4, pp. 313–318, 1981.

[43] D. A. Dickey and W. A. Fuller, "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the American Statistical Association*, vol. 74, no. 366a, pp. 427–431, 1979.

[44] N. Davies and J. D. Petruccelli, "Detecting non-linearity in time series," *The Statistician*, vol. 35, no. 2, pp. 271–280, 1986.

[45] R. S. Tsay, "Chapter 4.2 nonlinear tests," in *Analysis of Financial Time Series*, Wiley, Hoboken, NJ, USA, 3rd edition, 2010.

[46] L. Yang, C. Haozhou, R. Yuqi, and Y. Kunpneg, *Analysis of Factors Influencing Coal Prices under the New Normal of China's Economy*, Reseach Square, Durham, NC, USA, 2020.

[47] S. Cornot-Gandolphe, *China's Coal Market: Can Beijing Tame "King Coal"?*, Oxford Institue for Energy Studies, Oxford, UK, 2014, https://www.oxfordenergy.org/wpcms/wp-content/uploads/2014/12/CL-11.pdf.

[48] N. Zamani, "The relationship between crude oil and coal markets: a new approach," *International Journal of Energy Economics and Policy*, vol. 6, no. 4, pp. 801–805, 2016.

[49] X. Guo, J. Shi, and D. Ren, "Coal price forecasting and structural analysis in China," *Discrete Dynamics in Nature and Society*, vol. 2016, Article ID 1256168, 7 pages, 2016.

[50] P. Schober, C. Boer, and L. A. Schwarte, "Correlation coefficients: appropriate use and interpretation," *Anesthesia & Analgesia*, vol. 126, no. 5, 2018.

[51] N. Reimers and I. Gurevych, "Optimal hyperparameters for deep LSTM-networks for sequence labeling tasks," 2017, https://arxiv.org/abs/1707.06799.

[52] G. H. T. Ribeiro, P. S. G. d. M. Neto, G. D. C. Cavalcanti, and I. R. Tsang, "Lag selection for time series forecasting using particle swarm optimization," in *Proceddings of the 2011 International Joint Conference on Neural Networks*, pp. 2437–2444, San Jose, CA, USA, 2011.

[53] Scikit-learn.org, "Parameter estimation using grid search with cross-validation—Scikit-learn 0.19.1 documentation," 2021, http://scikit-learn.org/stable/auto_examples/model_-selection/plotgrid_search_digits.html.

[54] X. Li, L. Peng, X. Yao et al., "Long short-term memory neural network for air pollutant concentration predictions: method development and evaluation," *Environmental Pollution*, vol. 231, pp. 997–1004, 2017.

[55] J. McKenzie, "Mean absolute percentage error and bias in economic forecasting," *Economics Letters*, vol. 113, no. 3, pp. 259–262, 2011.

[56] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *International Journal of Forecasting*, vol. 22, no. 4, pp. 679–688, 2006.

[57] X. Diebold Francis and R. S. Mariano, "Comparing predictive accuracy," *Journal of Business & Economic Statistics*, vol. 20, no. 1, pp. 134–144, 1995.

[58] G. Li, H. Song, and S. F. Witt, "Recent developments in econometric modeling and forecasting," *Journal of Travel Research*, vol. 44, no. 1, pp. 83–99, 2005.

[59] WorldOil, "China buys millions of barrels of U.S. oil to comply with trade deal," 2021, https://www.worldoil.com/news/2020/8/18/china-buys-millions-of-barrels-of-us-oil-to-comply-with-trade-deal.

[60] Energy Information Administration, "Benchmarks play an important role in pricing crude oil," 2020, https://www.eia.gov/todayinenergy/detail.php?id=18571.

[61] D. Workman, "Top 15 crude oil suppliers to China," 2021, http://www.worldstopexports.com/top-15-crude-oil-suppliers-to-china/.

[62] S. Nanovsky, "The impact of oil prices on trade," *Review of International Economics*, vol. 27, no. 1, Article ID e0001, 2019.

[63] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[64] R. Díaz-Uriarte and S. Alvarez de Andrés, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol. 7, no. 1, p. 3, 2006.

Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 5575557, 8 pages
https://doi.org/10.1155/2021/5575557

Research Article

Enhancing Railway Maintenance Safety Using Open-Source Computer Vision

Donghee Shin,[1] Jangwon Jin,[2] and Jooyoung Kim[2]

[1] *Railroad Operation Company, NEO TRANS Co. Ltd., Seongnam-Si, Gyeonggi-do 13524, Republic of Korea*
[2] *Graduate School of Transportation, Korea National University of Transportation, Uiwang-si, Gyeonggi-do 16106, Republic of Korea*

Correspondence should be addressed to Jooyoung Kim; jykim@ut.ac.kr

Received 9 February 2021; Revised 16 April 2021; Accepted 20 April 2021; Published 28 April 2021

Academic Editor: Inhi Kim

As high-speed railways continue to be constructed, more maintenance work is needed to ensure smooth operation. However, this leads to frequent accidents involving maintenance workers at the tracks. Although the number of such accidents is decreasing, there is an increase in the number of casualties. When a maintenance worker is hit by a train, it invariably results in a fatality; this is a serious social issue. To address this problem, this study utilized the tunnel monitoring system installed on trains to prevent railway accidents. This was achieved by using a system that uses image data from the tunnel monitoring system to recognize railway signs and railway tracks and detect maintenance workers on the tracks. Images of railway signs, tracks, and maintenance workers on the tracks were recorded through image data. The Computer Vision OpenCV library was utilized to extract the image data. A recognition and detection algorithm for railway signs, tracks, and maintenance workers was constructed to improve the accuracy of the developed prevention system.

1. Introduction

With increasing construction of new high-speed railways, more railway maintenance work is necessary to ensure seamless operation. This, however, is accompanied by the frequent occurrence of accidents involving maintenance workers on the tracks.

According to a press release [1] from the Ministry of Land, Infrastructure and Transport, from 2007 to 2017, the number of railway accidents decreased by 14.6%, although the overall length of railway tracks had increased. However, the number of casualties continues to increase every year. The victims of such casualties are predominantly maintenance workers who are hit by a running train.

According to the Korea Transport Safety Authority, the majority of domestic railway accidents are either railway casualty accidents or railway safety accidents. The casualty accidents involve railway track workers accounting for about 60% of all accidents in a year. This suggests that workers are not adequately protected by existing safety management systems. In addition, a report from the Korea Transport Safety Authority shows that the fatality rate for domestic railway workers is approximately three times that for the leading European countries. Accordingly, appropriate railway safety measures are urgently required. As new high-speed railways are constructed and existing railways are improved, more maintenance work is required; thus, maintenance workers are more frequently exposed to unsafe situations that often result in accidents. When a maintenance worker is hit by a train, it almost always results in a casualty, which is a serious concern. The main safety measures for maintenance work are installing signs—which are expected to be recognized by train drivers—to mark a work zone and allocate a safety guard who notifies train drivers of the work zone. In other words, the existing safety measures rely heavily on human capacity. Unfortunately, it is entirely possible for a train driver to miss a work zone sign, and safety guards at the track side can also be exposed to accidents, along with maintenance workers. Thus, these measures are essentially incapable of preventing railway accidents. With the objective of preventing railway accidents, this study attempted to develop an improved safety system for maintenance workers on the tracks by utilizing the tunnel monitoring system installed on trains and railway vehicles. Image data of the tunnel monitoring system are obtained from PTZ (Pan-Tilt-Zoom) cameras installed throughout a train. The tunnel monitoring system monitors detect electric car lines and railway tracks in real time, as the train is running. This study involved constructing an algorithm for extracting images of railway signs, tracks, and track workers from the image data of a tunnel monitoring system by means of Computer Vision OpenCV library and recognizing those images. As the proposed method can detect track workers, tracks, and work zone signs as objects, it is expected to provide train or railway vehicle drivers with track information in advance, thus alerting them to the possibility of an accident.

This study was conducted on trains running on the Shinbundang Line. The total length of this line is 31.1 km, extending from Gangnam station in Seoul to Gwanggyo station in Gyeonggi-do. This study utilized images stored by the tunnel monitoring systems in trains running on the Shinbundang Line. Image data obtained between April 15 and 21, 2019, were used in this study. Images were stored in the tunnel monitoring system for one week.

1.1. Literature Review. The existing management methods for railway maintenance work can be classified as follows: (i) GPS-based notification systems that inform maintenance workers of a train or railway vehicle approaching a work zone near the track. However, these systems are inefficient for underground or tunnel sections. There are also problems related to communication fees and security licenses. (ii) Wireless communication system that alerts workers to approaching trains or railway vehicles by means of infrared sensors. These systems require a separate detector for approaching trains or railway vehicles to be installed. Moreover, as the number of work zones increases, the cost of installing detectors also increases. (iii) Frequency-based methods to inform railway maintenance workers of the approach of a train or railway vehicle. As these methods require a separate frequency transmitter/receiver to be installed, they are expensive to construct and maintain. (iv) A safety fence is installed for works that are carried out near tracks on which trains and railway vehicles are running. The safety fence should be installed only in a work zone, and a safety manager and other staff should be available to warn maintenance workers of an approaching train or railway vehicle. This method becomes impractical when there are several work zones.

The four aforementioned railway safety measures and the studies that have been conducted on them have common limitations regarding human error. These existing systems are dependent on the working conditions of maintenance workers and, thus, have drawbacks related to human error and efficiency. As workers need to attach an alert device both on the safety helmet and on their body while they are working with their tools, the efficiency of maintenance work is severely affected. Among existing studies on this issue, in order to prevent casualty accidents caused by large construction machines, Nieto developed an alert system in which GPS receivers are attached to large construction machines, and an alarm is sent to all workers when such a machine is approaching [2]. As each large construction machine is equipped with a camera system and a GPS receiver, if there is a worker in a predetermined work zone, an alarm is sent to the driver so that an appropriate action can be taken. Teizer proposed a system (that uses radio frequency (RF)) that sends an alarm to the driver of a construction machine and to a construction worker when the machine is approaching the worker, thereby preventing an accident [3]. RTSA (2012) developed automatic track warning systems to enhance the safety of track workers. Saito's system utilizes GPS to send an alarm to track workers through a mobile radio when a train approaches [4]. Hjort proposed an electronic data-transmission software program (ETW) using GPS in order to improve the safety of track workers [5]. D'Arco demonstrated that GNSS (Global Navigation Satellite System) produced fewer time and distance errors than GPS in railway maintenance sites. The GNSS improved the error of GPS in the measurement of distance. Besides, the GNSS achieved a higher estimation accuracy than GPS by combining relative distance estimates, and each track worker had to wear a receiver. Eirini Konstantinou (2019) eliminated noise with Kalman Filter, an algorithm for computer vision, and suggested using support vector machines (SVMs) to track, control, and monitor workers' locations. Mingyuan Zhang (2020) proposed a method to assess the safety level of construction workers based on computer vision and fuzzy reference, noting that construction workers have accidents in the environment of the construction site.

Existing studies have focused mainly on alerting workers to approaching trains. However, the working conditions (such as noise) near tracks and unauthorized works continue to cause accidents during maintenance works. Maintenance workers at the track side and a train driver cannot be relaxed until the train completely passed through the work zone. All of them are under heavy pressure. This study aimed to improve the environment of maintenance work done either on or near railway tracks on which trains would be running. To achieve this, an algorithm was developed to detect objects on or near the tracks, thereby avoiding accidents and facilitating the safe passage of a train through the work zone.

This study presents a method that is different from those used in earlier studies. The target train of this study was equipped with PTZ (Pan-Tilt-Zoom) cameras both at the front and the back. The PTZ camera-based object image detection data were utilized to show objects such as maintenance workers, tracks, and work zone signs to the train driver. This study developed an evacuation system that gives such railway track information to the train driver and helps the workers and the train avoid any contact. This would help the driver take appropriate action to avoid an accident.

This study is distinguishable from earlier studies in the following three ways.

First, text information of maintenance (work) signs was detected from PTZ camera images and provided to the train driver in advance. This study focused on the fact that the maintenance work signs should be installed 500 m before the work zone.

Second, the railway track was recognized and maintenance workers on or near the tracks were detected as objects by using PTZ camera images. Information thus obtained was provided to the train driver.

Third, the region of interest (ROI) for the recognition of railway tracks was set in order to detect other objects, in addition to maintenance workers. Information was provided to the train driver in real time so that the train could pass safely through the work zone.

Intel's OpenCV library and OpenCV_Python 3.6 were used for image processing and object detection.

2. Methods

This study constructed three algorithms by utilizing the OpenCV library. To construct these algorithms, a license-plate extraction algorithm, a lane-detection algorithm, and a vehicle and object-recognition algorithm were utilized. These algorithms are being actively studied in the field of intelligent transport system. The details of the method of this study can be summarized as follows. First, existing studies on text and license-plate extraction algorithms were reviewed. Considering that maintenance (work) signs are installed 200 and 500 m before the work zone, an algorithm was constructed, which extracts the text "500 m from work zone" and provides it to the driver of a train or railway vehicle before the train or vehicle enters the work zone. Second, the lane-detection algorithm, which had been extensively studied in the field of intelligent transport service, was utilized to construct a railway-track-detection algorithm. Third, the vehicle and object recognition algorithm of the intelligent transport service was utilized to construct an object-detection algorithm to detect maintenance workers near tracks.

2.1. Algorithm to Detect and Recognize Railway Signs. For sign recognition, there are two methods: shape recognition and color recognition. As the shape-recognition method is highly likely to capture similar shapes in the background, the recognition efficiency is poor and the recognition itself takes a long time. However, this method has a faster processing rate than the RGB color-recognition method and is less affected by the surroundings. Accordingly, a template-matching algorithm was adopted as the railway-sign recognition and detection algorithm, as it is less affected by the surroundings of railway tracks and can quickly detect and recognize specific signs (work sign).

Template matching is a technique for finding a specific image from an original image. In this study, the original image and a target image were processed by using gray scale. Subsequently, the target image was detected and recognized by using a specific red-color box. The template-matching function was used to recognize a work sign in a railway sign image. The process of template matching is presented in Figure 1.

The template-matching function was effective in detecting a work sign image using the original image. However, although this matching function could solve the problem of translation, recognition of rotated and scaled objects proved to be difficult, even with template rotation and scaling. Accordingly, this study utilized the RANSAC (Random Sample Consensus) algorithm, which accurately filters the matching result of key points between two images. The RANSAC algorithm assumes the existence of homography transformation between two images. This algorithm filters out incomplete matching results and retains only the results satisfying the motion model between two images [6]. The principle of the RANSAC algorithm is to create a model by random sampling from the data and determine how different the model is supported by, i.e., how many data have a distance from the model less than a constant value (T). The RANSAC algorithm is applied in the following order:

Two points are selected randomly.

$F(x)$ of the straight line passing between the two points is obtained.

The number of datasets C' is calculated, in which the distance of the above $F(x)$ $r_i = |u_i - f(\varepsilon)|$ is T and less.

In case C is larger than the saved C', a new C is saved.

After the above process is iterated N times, the optimal C is returned to $F(x)$.

A result is derived by applying the least-squares method to datasets satisfying $F(x)$.

Here, N is to be selected so that at least one dataset among all available datasets can satisfy the probability Q consisting of a model and inlier (appropriate point).

If the probability of all data being inlier is u, the probability of all data being outlier is $v = 1 - u$. From this, N can be calculated as follows:

$$N = \frac{\log(1-Q)}{\log\ (1-v)^{n}}. \tag{1}$$

The above process is iterated N times to determine the ultimate model [7]. When images were rotated and then matched by means of the RANSAC algorithm, target images that had not been shown by template-matching could be detected. After rotating the image picture using the RANSAC algorithm, the image was not detected in the template matching, but the RANSAC algorithm was detected, as shown in Figure 2.

2.2. Algorithm to Detect and Recognize Railway Tracks. The majority of the existing studies on lane detection have used conventional cameras to acquire images and survey the road in front and set it as ROI, before applying the lane-detection algorithm [8]. The lanes are then detected using

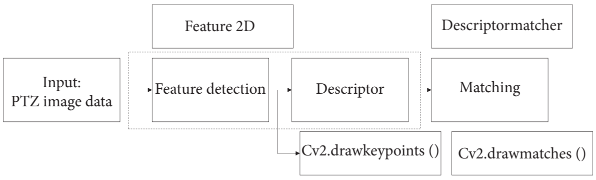


Figure 1: Template matching.

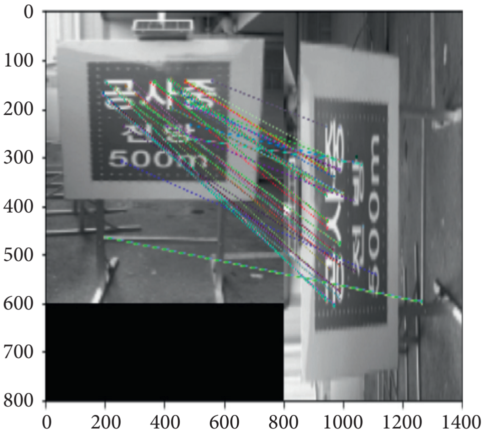


Figure 2: Result of application of RANSAC algorithm.

edge detection [9], Hough transform [10], template matching [11], and so on. Template matching is the most widely used technique. This method detects lanes by constructing a top-hat filter, which utilizes the brightness difference between lane and road and applies the corresponding template to the ROI. However, this method is not as effective in detecting curved lanes as it is in detecting straight lanes. A few attempts have been made to solve this problem by dividing an image into multiple sections and applying the Hough transform to each section [12]. Accordingly, this study constructed a Hough transform algorithm to detect railway tracks. Unlike road lanes, railway tracks cannot be distinguished by colors. For this reason, various methods, including the Gaussian filter, were used to construct an algorithm to detect and recognize the features of railway tracks.

There are various color models, such as RGB, YUV, and HSV. As road lanes are clearly distinguished by yellow and white, they are suitable for a color model. On the other hand, railway tracks are difficult to distinguish by colors. Accordingly, various color modes were applied, and the HSV model was selected to detect railway tracks. HSV is a color space that expresses images by hue (H), saturation (S), and value (V). The darkness and lightness of a color are expressed by the saturation channel, and the brightness is determined by the value. The HSV color space does not indicate a combination of colors; it indicates the color itself. It thereby achieves good intuitiveness. In case an object needs to be detected from an image by using colors, the HSV space seems to be more appropriate than the RGB space [13]. Railway track images were converted to the HSV color space using the OpenCV library, and the railway tracks were extracted from HSV. All objects other than the railway tracks were colored black; Figure 3 shows the result. There was a loss of railway track. To obtain a clearer image, an ROI (Region Of Interest) was set. By setting the ROI, the irrelevant objects were expressed in black, and, thus, the railway track images were detected and extracted. Straight railway tracks were clearly recognized in images. However, all the tracks were not straight. Thus, HSV alone was not sufficient. To solve this problem, a different HSV color model and the edge detection algorithm were used in conjunction.

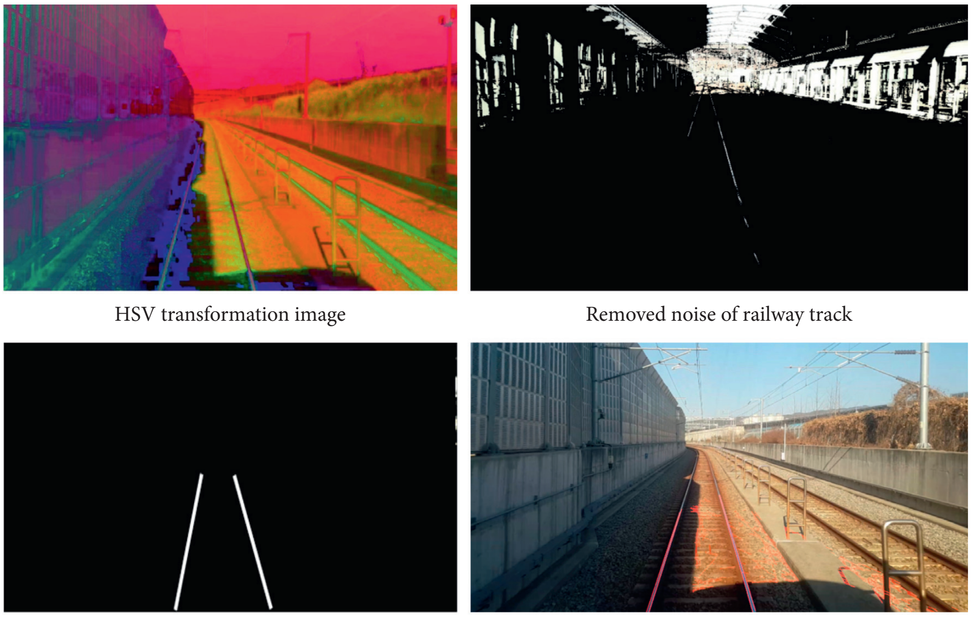
Among various edge detection methods, the Canny edge operator is most widely used as it is the most clearly defined. This method is recognized as the best optimized edge detection method from the following aspects, which are the conventional criteria for evaluating the performance of edge detection operators: efficiency of edge detection (Good Detection), locality of edge detection (Good Localization), and single response to an edge [14]. This study utilized Canny edge detection to address the insufficiency of HSV-based detection. However, it was necessary to recognize curved tracks in a different way from straight tracks. Accordingly, a Hough transform algorithm was implemented to detect representative lines, which were then applied to curved tracks.

2.3. Algorithm to Recognize Maintenance Workers. In the field of intelligent traffic service, several algorithms have been developed to detect pedestrians, and several other methods are still being studied. This study used HOG [1] and SVM [2] algorithms, which are most widely used and verified in the field of intelligent traffic service, to recognize maintenance workers on tracks. The details of the process were as follows: first, an image was inputted for recognition, and a feature vector was extracted by using the HOG feature. After that, a pretrained SVM was used to distinguish maintenance workers on the tracks. In the next step, images for training were inputted. After a HOG feature vector was extracted from the inputted training images, the SVM was trained, and training data were extracted and then utilized to recognize workers. Generally, maintenance workers always wear a uniform during work. If this feature is extracted, classified, and detected, the processing rate may be accelerated according to image background, and the workers can be recognized more quickly.

The flow chart for recognizing maintenance workers at the tracks using HOG descriptor and SVM algorithms is presented in Figure 4.

When only the HOG descriptor algorithm was implemented, only one of two workers at the trackside was detected. To address this and accurately detect workers at the tracks, the features of workers were classified using the SVM classifier.

2.4. Experimental Application and Evaluation of Algorithms. After algorithms were constructed based on images of railway signs, tracks, and maintenance workers, they were applied to real images for verification. For images of actual railways, the image data stored by the railway image



HSV transformation image

Removed noise of railway track

Setting of ROI zone

Recognizing and detection image in curve tracks

Figure 3: Process of algorithm to detect and recognize railway tracks.

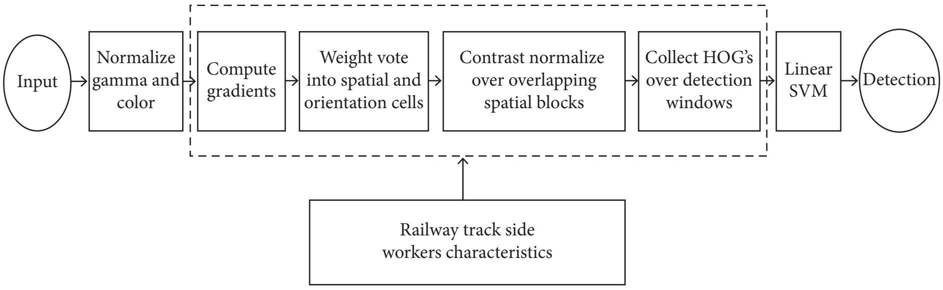

Input
Normalize gamma and color
Compute gradients
Weight vote into spatial and orientation cells
Contrast normalize over overlapping spatial blocks
Collect HOG's over detection windows
Linear SVM
Detection
Railway track side workers characteristics

Figure 4: HOG descriptor and SVM algorithms.

recording device of the Shinbundang Line were used. The algorithms were applied to the image data and were verified. This study attempted to recognize and detect railway signs, tracks, and maintenance workers while focusing on the safety of the workers. Template matching, RANSAC, and OCR were implemented for railway sign recognition. Color transform HSV, Canny edge, Gaussian blue, ROI, and Hough transform algorithms were applied to railway track images. An HOG descriptor and SVM model were implemented for detection of maintenance workers at the tracks by using the OpenCV library. Color images having a resolution of 1920 × 1080 pixels were converted to 800 × 600-pixel images to evaluate the constructed algorithms. The test images were trackside images captured either during daytime (in the open) or in a tunnel.

2.5. Application of Algorithm to Image Data. The recognition and detection of railway signs were evaluated using images stored by the image-recording device of a train running through tunnels on the Shinbundang Line. Figure 5 represents the application of algorithm to image data. To detect maintenance workers at the tracks, a morphological operation was used to remove noise, detect outlines, and apply the Gaussian blur. Moving objects in binary images were detected and recognized and were marked with red squares. Initially, there were several errors in detection. However, with continued use, the maintenance workers, who were classified by using the SVM classifier, could be recognized and detected.

3. Results of Algorithm Verification

To numerically evaluate the algorithms, this study adopted the concepts of precision and recall, which have been used for performance verification and evaluation of object-recognition and detection algorithms in several studies.

Precision is a measure of accuracy. Precision is the ratio of true detections to all detection results. It can be expressed by the following equation:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{\text{TP}}{\text{all detections}}. \qquad (2)$$

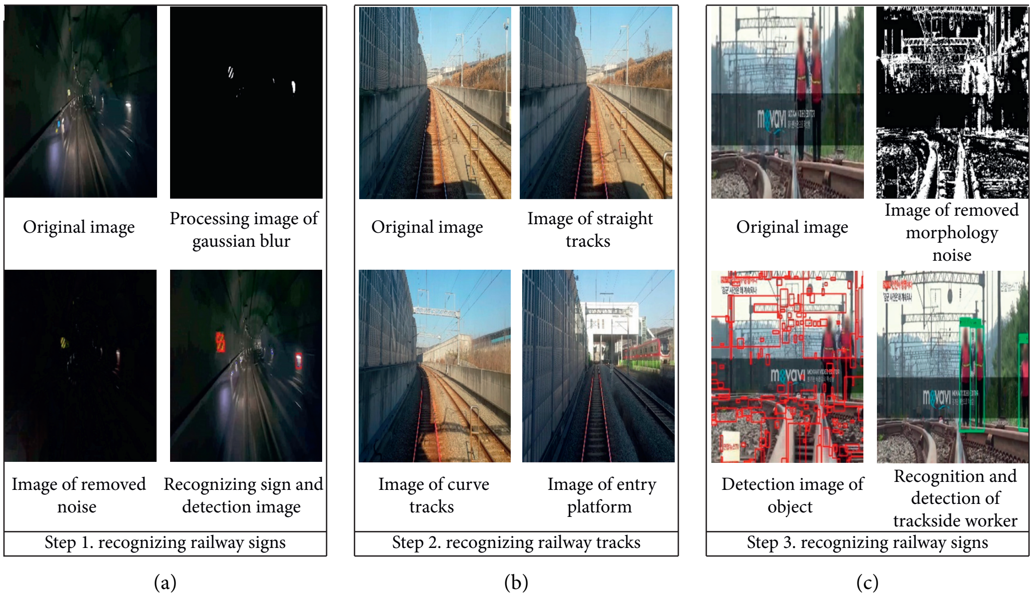

Original image
Processing image of gaussian blur
Image of removed noise
Recognizing sign and detection image
Step 1. recognizing railway signs
(a)
Original image
Image of straight tracks
Image of curve tracks
Image of entry platform
Step 2. recognizing railway tracks
(b)
Original image
Image of removed morphology noise
Detection image of object
Recognition and detection of trackside worker
Step 3. recognizing railway signs
(c)

Figure 5: Application of algorithm to image data.

Here, TP stands for true positive, which indicates accurate detection, while FP stands for false positive, which indicates incorrect detection. Hence, precision is the percentage of the accurate detections among all the detections made by an algorithm. If an object detection algorithm detects five objects, of which four are TPs, the precision is 4/5 = 0.8.

Recall denotes a detection rate or a recall rate. In other words, it is the ratio of true detections to all the targets. Recall can be expressed by

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\text{TP}}{\text{all ground truth}}. \tag{3}$$

Here, FN stands for false negative, which indicates objects that are to be detected but have not been detected yet. TP (true positive) refers to a case where a target object is accurately detected. FP (false positive) refers to a case where a nontarget object is wrongly recognized or detected. FN (false negative) refers to a case where a target object has not been recognized and detected. TN (true negative) refers to a case where a wrong thing that is not a target is accurately sensed and has not been recognized and detected. The above classification can be summarized in Table 1.

It is not sufficient to use only precision or only recall for performance evaluation of an object-detection algorithm. Let us assume that there are 10 objects and four out of 5 objects are correctly detected. Then, precision = 4/5 = 0.8 and recall = 4/10 = 0.4. Precision indicates good performance but recall does not. It is noteworthy that the values of precision and recall are always between 0 and 1, and, when the precision is high, the recall tends to be low and vice versa. Accordingly, it is almost the same to evaluate the performance of an algorithm using either of these parameters. It is necessary to apply both of them for evaluating an algorithm accurately. In this regard, the precision-recall curve and AP are needed. Furthermore, there is a criterion to judge whether an object has been detected correctly: this criterion is the intersection over union (IoU). This study utilized the performance verification index for object-detection algorithms, which was proposed by Everingham [15]. This method is IoU (Intersection over Union). As shown in Figure 6, let us assume that there is an image labelled with a ground truth boundary box. The ground truth boundary box wraps the object that is to be recognized and detected. When the ground truth box of the image was not given, the boundary box was detected by an object-recognition and detection algorithm, as follows.

Table 1: Definition of precision and recall.

Ground truth	Predict result	
	Positive	Negative
Positive	TP (true positive)	FN (false negative)
Negative	FP (false positive)	TN (true negative)

IoU measures the area of overlap between a recognized and detected boundary box and the ground truth boundary box and then divides the overlapped area by the union area. The equation is presented below. If the IoU value is 0.5 and above, the result is judged to be true. Otherwise, if the IoU value is less than 0.5, the result is judged to be false.

$$\text{IoU} = \frac{(R \cap G)}{(R \cup G)}. \tag{4}$$

Here, R is the boundary box detected by algorithm and G is the ground truth.

The performance of the algorithm for recognizing and detecting railway signs was verified by analyzing image data in real time. The data were acquired by the railway track and tunnel-monitoring system. The focus of the performance verification was whether the algorithm could recognize and

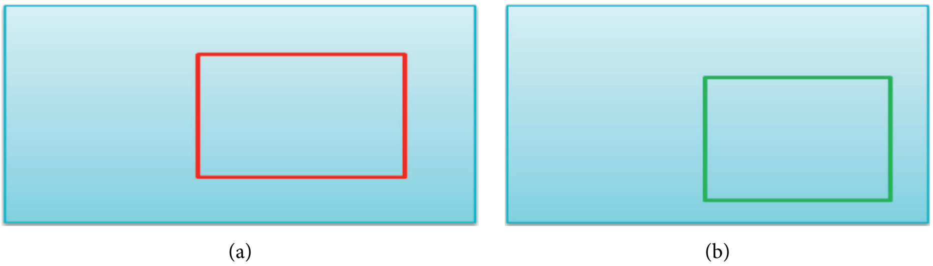


Figure 6: Ground truth boundary box: (a) ground truth and (b) the boundary box detected by algorithm.

detect railway signs while the train was running at 50 km/h, 70 km/h, and 90 km/h in a tunnel.

The number of false negatives is that of railway signs that were not recognized and detected. The number of false positives is that of fluorescent lights or trains on the opposite side, which were recognized and detected. In other words, objects other than railway signs were also recognized and detected. The detection rate is the number of detected railway signs among all the railway signs. When a train was running on Shinbundang Line, the algorithm was evaluated using the image data of the real-time tunnel monitoring system.

When the recognition and detection algorithm for railway signs was applied for 0–50 km/h image data, there were no false negatives. However, there was one false positive where a fluorescent light in the tunnel was detected instead of a railway sign. When the train was moving at 51–70 km/h, there was one false negative, and there were three false positives. The traffic control system at the trackside and fluorescent lights in the tunnel were wrongly detected. When the train was moving at 71–90 km/h, there were two false negatives and four false positives. Table 2 presents the results of false negatives and false positives.

The algorithm for recognizing and detecting maintenance workers cannot be verified according to the velocity of train. For this reason, this object-detection algorithm was verified by classifying cases as follows: first, the workers were scattered or grouped (gathered). Second, they were gathered in the longitudinal direction. And third, they were gathered within a facility on the ground as shown in Table 3.

Although a worker was recognized and detected, when the object detection IoU was <0.5, the result was false. When the workers were scattered (i.e., they were a certain distance away from each other), the algorithm showed neither any false negatives nor any false positives. However, when the workers were arranged in the longitudinal direction, false negatives were obtained. In addition, when the workers at the trackside were grouped together, their faces, arms, legs, and bodies were concealed such that the number of the workers could not be accurately detected in some cases. Although false positives and false negatives were obtained, the algorithm showed performance indices of 0.5 and above. Thus, it can be concluded that the algorithm for recognizing and detecting workers at the trackside performs sufficiently well.

4. Conclusions and Further Scope

Appropriate signs should be installed 200 and 500 m before a work zone, in order to alert train drivers in advance. Drivers are expected to pay careful attention while operating their trains or railway vehicles; a guard should also be available to send a signal to all drivers. However, not only maintenance workers at the trackside but also the guards are often hit by trains and become victims of casualty accidents. Accordingly, to prevent such accidents, this study utilized the tunnel monitoring system installed on trains to recognize and detect maintenance workers at the tracks. This study attempted to develop an algorithm to recognizing a work zone and warn or even stop an approaching train, in addition to the existing alert system that lets maintenance workers know of any approaching trains.

However, this study cannot perfectly ensure the safety of maintenance workers at the tracks. It has limitations that need to be addressed in the future. An image-processing library was used to recognize and detect workers at the trackside. However, not all trains are equipped with the tunnel-monitoring system. In addition, urban railways undergo maintenance work only at night, which imposes a time limitation. The functions, models, and library, which were used to recognize and detect railway signs, tracks, and workers, are not final solutions. As there are several ongoing research and development projects, various methods need to be considered. Nevertheless, this study is significant as it developed a new approach. Existing systems only alert maintenance workers to the approach of a train or a railway vehicle. The proposed system enables train drivers to recognize and detect a work zone in advance and to be prepared for an emergency. Therefore, this study contributes to enhancing the safety of workers at railway tracks.

This study has the following limitations that need to be addressed. First, the proposed algorithm of this study is difficult to generalize. In order to address this, various methods need to be applied and analyzed. Second, this system recognized and detected objects at the trackside (railway signs, railway tracks, and maintenance workers) by using only the image data of the tunnel-monitoring system. However, it is necessary to collect diverse data and extend the spatial scope of research. Third, this study collected and utilized only the data of Shinbundang Line as the image data of the tunnel-monitoring system. As urban railways are equipped with various tunnel-monitoring systems, the scope

Table 2: Verification of recognizing and detecting algorithm railway signs.

Classification	Railway sign	True positive	False negative	False positive	Precision (TP/TP + FP)	Recall (TP/TP + FN)
0~90 km/h	95	84	3	8	0.96	0.91
0~50 km/h	40	39	—	1	0.98	1.0
51~70 km/h	35	31	1	3	0.91	0.97
71~90 km/h	20	14	2	4	0.78	0.86

Table 3: Verification of recognizing and detecting algorithm trackside workers.

Classification	# of trackside workers	True positive	False negative	False positive	Precision (TP/TP + FP)	Recall (TP/TP + FN)
# of trackside workers	1	1	—	—	1.0	1.0
	2	2	—	—	1.0	1.0
	5	4	1	—	1.0	0.8
	6~7	4	3	—	1.0	0.57
Trackside workers (longitudinal)	4	2	2	—	1.0	0.5
Trackside workers + ground facility	7	5	1	1	0.83	0.83

of research needs to be extended by using data of the tunnel-monitoring systems of other railway lines. In spite of these limitations, this study will contribute to preventing accidents caused by the mistakes of train or railway vehicle drivers. Moreover, this study provides a basis for future studies aimed at preventing maintenance workers from being hit by trains or railway vehicles.

Data Availability

Image data of the tunnel monitoring system are obtained from PTZ (Pan-Tilt-Zoom) cameras installed throughout a train.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] Ministry of Land, “Infrastructure and transport,”Ministry of Land, Report 3.25, 2018, http://www.molit.go.kr/USR/NEWS/m_71/dtl.jsp?id=95080549.

[2] A. Nieto, S. Miller, and R. Miller, “GPS proximity warning system for at-rest large mobile equipment,” *International Journal of Surface Mining, Reclamation and Environment*, vol. 19, no. 1, pp. 75–84, 2005.

[3] J. Teizer, B. S. Allread, C. E. Fullerton, and J. Hinze, “Autonomous pro-active real-time construction worker and equipment operator proximity safety alert system,” *Automation in Construction*, vol. 19, no. 5, pp. 630–640, 2010.

[4] S. Teruaki, O. Masaru, and S. Atsushi, “Development of a train approach alarm for lines without track circuits,” *JR EAST Technical Review*, vol. 30, 2014.

[5] H. Graham and H. Ben, “Digital track occupancy authorities for staff working on track,” *AusRAIL PLUS*, vol. 21-23, 2017.

[6] A. Spizhevoy, *OpenCV 3 Computer Vision with Python Cookbook*, Acorn Publishing Co., Cleveland , OH, USA, 2019.

[7] A. F. Martin and R. C. Bolles, “Random Sample Consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” *ACM*, vol. 24, pp. 381–395, 1981.

[8] C. S. Bae, J.-H. Lee, and S.-B. Cho, “Lane detection algorithm using morphology and color information,” *System and Control*, vol. 15–24, 2011.

[9] B. H. Kim, Y. Han, and H. Hahn, “Lane detection using moore-neighbor edge trace algorithm,” *Proceedings of IEIE*, vol. 31, pp. 857–85, 2010.

[10] X. Fangfang et al., “Real-time lane detection for intelligent vehicles based on monocular vision,” in *Proceedings of the Control Conference (CCC)*, Guangzhou, China, 2012.

[11] C.-H. Kum, D.-C. Cho, and W.-Y. Kim, “Development of lane detection system using surrounding view image of vehicle,” *Journal of Broadcasting Engineering*, vol. 36, 2013.

[12] W. Li, “Human-like driving for autonomous vehicles using vision-based road curvature modeling,” *International Journal of Hybrid Information Technology*, vol. 6, no. 5, pp. 103–116, 2013.

[13] H. Lee, *A Technique of Handwritten Resident Registration Number Detection and Extraction Using Character Recognition Based on CNN*, Soongsil University, Seoul, South Korea, 2016.

[14] Y. Choi, *Study on a Face Recognition and a Performance Comparison by Threshold Change of Canny Edge Operator*, Graduate School of Electronics and Communications Engineering, Kwangwoon University, Seoul, South Korea, 2006.

[15] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (VOC) challenge,” *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 6676092, 15 pages
https://doi.org/10.1155/2021/6676092

Research Article

Driver Lane-Changing Behavior Prediction Based on Deep Learning

Cheng Wei,[1] **Fei Hui**,[1] **and Asad J. Khattak**[1,2]

[1]*School of Information and Engineering, Chang'an University, Xi'an, ShaanXi 710064, China*
[2]*Civil and Environmental Engineering Department, University of Tennessee, Knoxville, TN 37996, USA*

Correspondence should be addressed to Fei Hui; feihui@chd.edu.cn

Received 9 November 2020; Revised 12 March 2021; Accepted 9 April 2021; Published 28 April 2021

Academic Editor: Young-Jae Lee

A correct lane-changing plays a crucial role in traffic safety. Predicting the lane-changing behavior of a driver can improve the driving safety significantly. In this paper, a hybrid neural network prediction model based on recurrent neural network (RNN) and fully connected neural network (FC) is proposed to predict lane-changing behavior accurately and improve the prospective time of prediction. The dynamic time window is proposed to extract the lane-changing features which include driver physiological data, vehicle kinematics data, and driver kinematics data. The effectiveness of the proposed model is validated through the experiments in real traffic scenarios. Besides, the proposed model is compared with five prediction models, and the results show that the proposed prediction model can effectively predict the lane-changing behavior more accurate and earlier than the other models. The proposed model achieves the prediction accuracy of 93.5% and improves the prospective time of prediction by about 2.1 s on average.

1. Introduction

Driver lane-changing behavior is a key factor in driving safety. An improper lane-changing behavior may cause a vehicle collision [1, 2] or even a traffic accident [3–5]. In [6], it was indicated that nearly 18% of the total number of traffic accidents were caused by improper lane changing. Using a prediction model in the Advanced Driver Assistance Systems (ADASs) [7–9] could reduce the risk of accidents. Therefore, a model for accurate prediction of a driver lane-changing behavior using multiple data fusion is needed. Substantial research regarding the lane-changing prediction has been conducted. At present, there are two mainstream groups of methods for prediction models of lane-changing, namely, mathematical methods and artificial intelligence approaches. One of the lane-changing prediction models based on a mathematical method was introduced in [10, 11]. Also, in [12],the logistic regression method was used in a lane-changing prediction model, wherein the distances to the front and adjacent rear vehicles, forward time-to-collision (TTC), and turn signal were taken into account, and the results showed that this model performed well in certain circumstances. Baumann et al. [13] improved a cognitive model to characterize driver behavior in an automotive environment and proved the correlation between drivers' cognitive processes and their driving movements. Salvucci [14] introduced an adaptive control of rational cognitive structures to monitor the lane-changing process in a multilane highway environment, and their model demonstrated how cognitive architectures could facilitate understanding of driver behavior. However, some of the mentioned studies were conducted in specific traffic scenarios such as highway entrance and ramp; besides, a part of the traffic scenarios was simulated instead of a real one. Therefore, those studies may not consider the lane-changing behavior in real traffic scenarios. Moreover, the above studies achieved the prediction accuracy ranges from 80% to 85%, so it has considerable space for improvement.

With the aim to build a more intelligent lane-changing prediction model, researchers have adopted machine learning. A lane-changing behavior prediction model based on the support vector machine (SVM) classifier and Bayesian filtering (BF) was proposed in [15], and it was shown that this model could predict driver lane-changing behavior 1.3 s in advance. In addition, a lane-changing behavior and trajectory prediction model based on the Hidden Markov Model (HMM) was presented in [16], and the results showed that this model predicted the lane-changing trajectories very well,which makes it suitable for prediction of human-like lane-changing maneuvers. Hou et al. [17] proposed a lane-changing prediction model based on fuzzy logic, which was developed for the case of a forced lane-changing under the lane-descending conditions. The prediction accuracy of this model for nonmerging behavior and merging behavior was 86.3% and 87.5%, respectively. Furthermore, a novel lane-changing intention recognition algorithm which combines the HMM and BF models was proposed in [18], where the model input consisted of three signals from the CAN bus (steering angle, lateral acceleration, and yaw rate), and the output was behavior classification. The results revealed that the HMM-BF could achieve an average recognition accuracy of 91.9%. Zheng et al. [19] proposed a machine learning-based segmentation and classification algorithm consisting of three stages. The first stage includes data preprocessing and prefiltering, and its function is to reduce noise and remove clear left and right turning events. The second stage employs a spectral time frequency analysis segmentation approach to generalize all potential time-variant lane-changing and lane-keeping candidates. The third stage includes two possible classifications: lane-changing and lane-keeping. The results showed that the average accuracy of this three-stage algorithm exceeded 83.22%. Furthermore, in [20], a dynamic Bayesian network (DBN) was used to predict the lane-changing maneuvers,and the test using the real data showed that the lane changing was detected 1 s in advance. However, the machine learning methods are not suitable for multisource data fusion, and even a single-input data may result in lower accuracy.

In recent years, the deep learning algorithms have been widely used in lane-changing prediction because of its powerful high-dimensional data processing and autonomous learning capabilities, which is in sharp contrast with conventional mathematical methods. Xie et al. [21] proposed a deep learning-based method to predict the future trajectory of vehicles and achieved good results. It is demonstrated that the deep learning model can mine potential features from high-dimensional data and also indicated the feasibility of deep learning in lane-changing research. In [22], a backpropagation (BP) neural network was used as a controller of an automatic vehicle system, and a camera image was used as a neural network input to construct a lane-changing model. The results showed that one hidden layer was enough to provide good performance of a time-varying nonlinear dynamic system. Tomar et al. [23] proposed a method based on a multilayer perceptron to predict a lane-changing trajectory accurately. The prediction results showed that this model was able to predict the future path accurately only for discrete patches of a trajectory, but not for the complete trajectory. Ding et al. [24] developed a BP neural network-based model to predict the lane-changing trajectories, and they compared prediction results of the BP neural network with that of the Elman network. It was found that the BP neural network-based model achieved better prediction performance under different sectionsand generated more reliable simulation results than the Elman network-based model. In [25], a fully connected neural network was applied to predict the lane-changing behavior of drivers; especially, the network model input consisted of multivehicle data, and the prediction accuracy of more than 90% was achieved. Moreover, a multifeature fusion neural network [26] that takes into account the physiological factors such as driver's head rotation was proposed to predict driver lane change behavior and the prediction accuracy exceeded 85%, while the prospective time was 1.5 s. Dou et al. [27] introduced a prediction model based on the SVM and BP neural network, which combined the results of SVM and BP neural network to improve the prediction accuracy, and the results showed that the average combined accuracy exceeded 92%. Furthermore, an MTSDeepNet using a convolution kernel to process the multivariate time series data and a fully connected neural network to classify the lane-changing behavior were designed to predict the lane changing in [28], and the accuracy of this model exceeded 91.0%. Considering the driver's driving style, Li et al. [29] proposed a lane-changing intention estimation model based on Bayesian network and Gaussian mixture model, which achieved a good prospective time, but its accuracy is low. However, the lane-changing process is determined by both the driver and traffic environment, but the aforementioned research studies did not consider all the factors affecting the driver lane-changing behavior. Also, using a fully connected network for lane-changing prediction may cause data loss which further reduces model performance.

In summary, the existing studies have the limitations of low prediction accuracy and short prospective time. Two reasons for these limitations are "data problem" and "prediction model structure problem." Aiming at these two problems, a hybrid neural network driven by multiple types of data is proposed. The first level of the hybrid network is composed of Seq2Seq, a variant of RNN [30, 31], which is mainly used for time series data processing to reduce invisible data loss. The second level consists of a fully connected neural network for data fusion and lane-changing classification. There are three contributions of this study. (1) Three different types of data including vehicle kinematics data, driver kinematics data, and driver physiological data are collected and used. (2) A hybrid network with a two-level training model is proposed to deepen the number of network layers while avoiding the problem of gradient dispersion. (3) A dynamic time window algorithm is proposed to ensure the consistency and homogeneity of the model input data and extend the prediction prospective time.

The remainder of the paper is organized as follows. Section 2 describes the data source and introduces the data processing method. Section 3 elaborates the working principle and structure of the proposed Seq2Seq-FC neural network, as well as its mathematical relationships. Section 4 validates the model generalization ability. Lastly, Section 5 concludes the paper.

2. Data Collection and Processing

Since the data used in this study was derived from real traffic scenarios and different sensing equipment, the original data was first filtered and segmented, and then, the timestamp alignment was conducted, and the data with the same label was extracted by a time window.

2.1. Data Collection. In the data collection process, the speed and acceleration of a vehicle were collected by the Cohda-wireless dedicated short-range communication (DSRC) installed on the vehicle. The steering angle and angular velocity of the steering wheel were obtained by a corner tester mounted at the steering wheel. The electroencephalogram (EEG) and driver's head movement data were obtained by the brain wave analyzer. Besides, the electrocardiogram (ECG) stemmed from a heart rate tester. The number of driver's head rotations in the horizontal direction was determined using the driving video obtained by a driving recorder. The equipment used for data acquisition is shown in Figure 1. Also, a different data was used for model training and validating. The data was collected in the same way but using different traffic routes. The route used for the collection of training data is presented in Figure 2, and the route started at the Chang'An University going through the main roads in Xi'an and ended at Xi'an Cheng'nan Passenger Transport Center. During the process of data collection, the procedure was repeated for several times. Since all data collection equipment are connected to the same PC, a timestamp is added to the header of each data based on Beijing time to synchronize the data collected by different equipment. In addition, the purchased brainwave analyzer and DSRC equipment have supporting data receiving software, which can denoise and filter the data and other self-designed equipment receive data using the serial communication protocol, and the data are denoised using the period-average padding method and PauTa criterion built into hardware. Therefore, the data output by a different equipment has been denoised and can be used directly.

It should be noted that, during the research, we found the impact of other factors on the subject vehicle (such as surrounding environmental factors,traffic environmental factors, and driving purpose) will ultimately be reflected in the driver's control of the vehicle. For example, if the number of vehicles around the subject vehicle increases, the speed of the subject vehicle will decrease. If the subject vehicle is driving on a slippery road, its overall speed will be significantly lower than the speed on a normal road. Therefore, only the subject vehicle's own data and the driver's own data were collected.

2.2. Features Extraction. The time window is the basis for data processing, and the data in time window should contain the data referring to the entire lane-changing process,not only a part of it. As shown in Figure 3, during the lane-changing process, the β bands of brain wave signal change greatly at the beginning of the lane-changing, and vehicle's steering angle tends to become stable at the end of the lane-changing process. Therefore, the point where the β bands change drastically is recorded as the starting point of the time window, and the stable point of the steering angle after the lane changing is completed and is recorded as the end point of the time window. Because the time taken for each lane changings is different, each lane-changing behavior has its corresponding time window. Considering the prospective time of the prediction model, the method of time window shrinking is adopted in this work. As shown in Figure 3, without considering the prediction accuracy, the endpoint of time window 2 has longer prospective time than that of the endpoint 1. Therefore, in the process of model training, the model is verified by shrinking the time window while comparing the accuracy. During data processing,the data

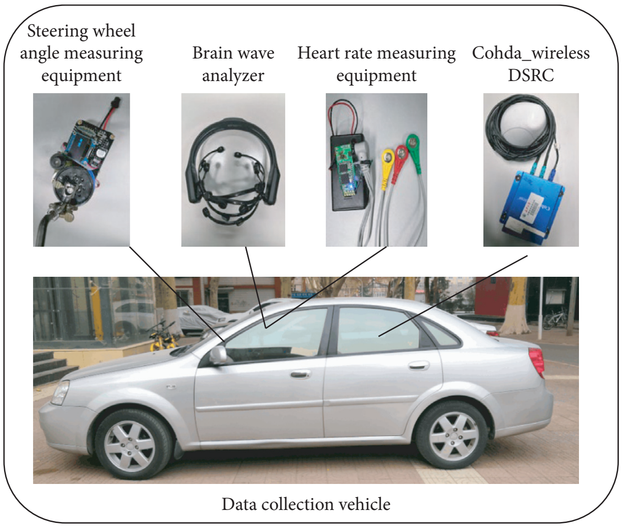

Steering wheel angle measuring equipment
Brain wave analyzer
Heart rate measuring equipment
Cohda_wireless DSRC
Data collection vehicle

Figure 1: Equipment used for data collection.

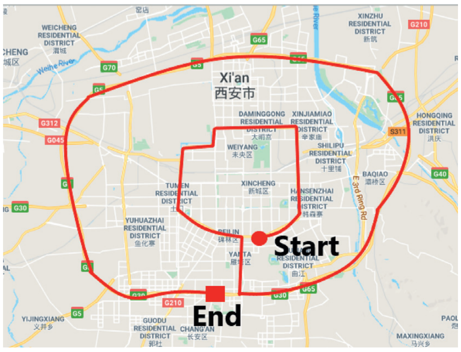

Xi'an
西安市
Start
End
E 3rd Ring Rd
G65
G70
G5
G312
G045
G30
G210
S311
G40
WEICHENG RESIDENTIAL DISTRICT
XINZHU RESIDENTIAL DISTRICT
DAMINGGONG RESIDENTIAL DISTRICT
XINJIAMIAO RESIDENTIAL DISTRICT
HONGQING RESIDENTIAL DISTRICT
WEIYANG
SHILIPU RESIDENTIAL DISTRICT
BAQIAO
XINCHENG
TUMEN RESIDENTIAL DISTRICT
HANSENZHAI RESIDENTIAL DISTRICT
YUHUAZHAI RESIDENTIAL DISTRICT
BEILIN
YANTA
YIJINGXIANG
GUODU RESIDENTIAL DISTRICT
CHANG'AN
MAXINGXIANG

Figure 2: Data collection route.

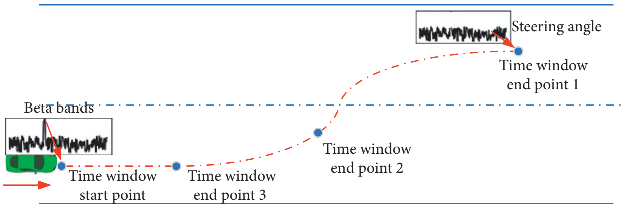


Figure 3: Lane-changing process.

length of a current time window is important because the RNN-Seq2Seq used in this work requires input data of a fixed dimension. However,since a dynamic time window is used, a data extraction method is needed. The data extraction method not only ensures the consistency of input data but also minimizes data loss. Besides, the maximum speed of a driver's head in a three-dimensional space and the number of driver's head rotations in the horizontal direction within a time window are directly fed to the input of the fully connected neural network without being processed by the RNN-Seq2Seq.

2.3. Data Processing. Vehicle kinematics data processing. The length of the time window was dynamically adjusted, to ensure that the dimension of data extracted from each time window was equal, during the vehicle kinematics data processing, for each time series data in a time window, the maximum, minimum, average, and variance values were used. Table 1 shows the characteristic of data within a time window which were used for further data processing. When the time window was gradually shrunk, the label data of the Seq2Seq was a caudal data of the original time window, but the processing method stayed the same. As can be seen in Table 1, 15 data features were extracted and used. There are two reasons for selecting these 15 features. The first is that such feature extraction method can ensure the same dimension and consistency of model input. The second is that the selected 15 features can reflect the vehicle movement situation in a time window no matter how long the time window is, which is also the meaning of those 15 features.

Driver's physiological and kinematics data processing. The EEG data were filtered by a bandpass filter which was a Chebyshev type II filter with a lower cut-off frequency of 4 Hz and an upper cut-off frequency of 30 Hz. The δ bands of 1–3 Hz in the EEG with an amplitude of 20–200 μV were removed because the δ wave appears in the human infant stage or immature mental development period or when an adult is under extreme fatigue and lethargy or anesthesia. Therefore, these bands do not change significantly when the driver considers the lane-changing. In addition, the θ bands with a frequency of 4–7 Hz and an amplitude of 5–20 μV, the α bands with a frequency of 8–13 Hz and an amplitude of 20-100 μV, and β bands with a frequency of 14–30 Hz and an amplitude of 100–150 μV were needed. Namely, the waveforms corresponding to these three bands types changed significantly when people are in the depression state, normal state, and excited state, so they should be used as input to the prediction model because they are changeable and important pointers to the people emotions and states. Similarly, the average, maximum, and minimum values and the variance of θ, α, and β bands of the EEG data within a time window were included. Six frequency domain features are extracted from each of the three EEG channels: $\theta/(\theta+\alpha)$, $\alpha/(\theta+\alpha)$, $(\theta+\alpha)/\beta$, α/β, $(\theta+\alpha)/(\theta+\beta)$, and θ/β. The reason for selecting these six features is that Martensson's research [32] shows that these six features can reflect the driver's brain activity at a certain time and can also reflect the driver's thinking while driving. Like the method of vehicle kinematics data extraction, the selected 18 brainwave data related features can reflect the driver's brain thinking when performing lane changing.

Table 1: Vehicle kinematics data processing method.

Feature	Process method
Speed	Maximum, minimum, average, and variance
Acceleration	Maximum, minimum, and average
Steering angle	Maximum, minimum, average, and variance
Steering angular velocity	Maximum, minimum, average, and variance

The heart rate signal was processed similarly as the other data, and the maximum, minimum, and average values and variance of the heart rate in a time window were taken as the input data regarding the motion data of a driver's head in a time window, and we used the speed maximum value in a three-dimensional space and the number of rotations of a driver's head in the horizontal direction. These four features were not processed by the RNN, instead they were fed directly to the input of the fully connected network because, to a certain extent, the maximum speed of head movement can reflect the urgency of the driver to change lanes. For example, if the driver frequently turns his head during driving and the speed is high, the driver has a higher probability of performing lane changing. In summary, 26 data features were extracted and used as input parameters for data processing. The extracted features and their labels are shown in Table 2. And, the 26 data features in Table 2 can reflect the driving situation of the vehicle in two aspects (vehicle and driver) when input as a model. It also ensures that the dimensions of the input data are not affected by the length of the time window.

After the original data was processed as described above, the dimension of the input data will be the same in every time window regardless of its length. In this way, data consistency during model training and testing was ensured. In Figures 4 and 5, the vehicle kinematic data and the brain wave during the lane-changing process are presented, respectively. The reason why the steering angle shown in Figure 4 does not have a negative value is, when designing the steering wheel angle measuring equipment, we set its equilibrium state (middle value) to 0 and design it as a positive value regardless of turning to the right or turning to the left.

3. Proposed Model

As already mentioned, the proposed model is based on an RNN-Seq2Seq network and a fully connected neural network. The RNN-Seq2Seq network is used to extract the features of the original data and predict the value of the

Table 2: The extracted features and labels.

Parameters	Definition
$\theta_{\max}\ \theta_{\min}\ \theta_{ave}\ \theta_{\text{var}}$	Maximum, minimum, average and variance of theta bands
$a_{\max}\ a_{\min}\ a_{ave}\ a_{\text{var}}$	Maximum, minimum, average, and variance of alpha bands
$\beta_{\max}\ \beta_{\min}\ \beta_{ave}\ \beta_{\text{var}}$	Maximum, minimum, average, and variance of beta bands
$p_{\max}\ p_{\min}\ p_{ave}\ p_{\text{var}}$	Maximum, minimum, average, and variance of heart rate
$(\theta/(\theta+\alpha))\ (\alpha/(\theta+\alpha))$ $(((\theta+\alpha)/\beta\alpha)/\beta)$ $(\theta+\alpha)/((\theta+\beta)\ \theta/\beta)$	Six absolute power of theta, alpha, and beta bands
$\omega_{\max}\ \omega_{\min}\ \omega_{ave}\ \omega_{\text{var}}$	Maximum, minimum, average, and variance of steering angle
$\omega_{a\max}\ \omega_{a\min}\ \omega_{aave}\ \omega_{a\text{var}}$	Maximum, minimum, average, and variance of angular velocity
$v_{\max} v_{\min} v_{ave} v_{\text{var}}$	Maximum, minimum, average, and variance of vehicle speed
$a_{\max} a_{\min} a_{ave}$	Maximum, minimum, average, and variance of vehicle acc
$H_{x\max} H_{y\max} H_{z\max}$	Maximum speed of the head in three dimensions
H_c	Number of rotations of the head within the time window

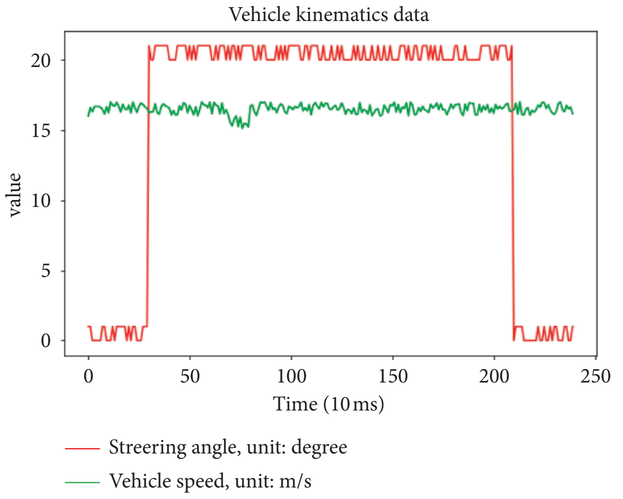


Figure 4: Vehicle kinematics data during the lane-changing process.

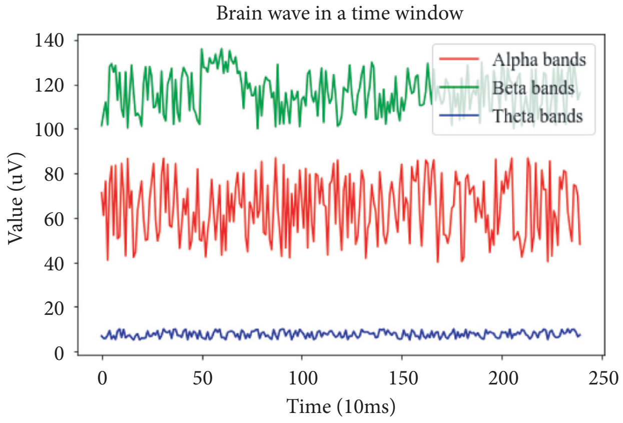


Figure 5: Brain wave during the lane-changing process.

next point which will be used as an input for the following fully connected neural network. In this work, 41 features are extracted and used for data processing. Except the data related to the driver's head rotation, the others features are subjected to the RNN-Seq2Seq processing. Therefore, the Seq2Seq has 37 inputs and 37 outputs. In addition, the fully connected neural network has 41 inputs and 2 outputs. The data processing procedure is shown in Figure 6. It should be noted that when the entire prediction model is trained, the accuracy and prospective time of the model are compared by shrinking the time window.

3.1. Seq2Seq Layer. RNN is a kind of neural network that can model data sequences. The main advantage of this neural network type is that it can process time series data well. Since the lane-changing behavior often lasts for a long period of time, if the time series data is directly processed by a fully connected neural network, there will be inevitable data loss, which will decrease model accuracy. Additionally, when processing the time series data, an RNN considers the correlation between data in the data sequence, so the use of a data in a time window can be maximized.

Since our goal is to develop the prediction model whose input is consisted of time sequence data, an RNN structural variant, and the Seq2Seq structure, containing the encoder and decoder, is chosen, which represents an enhanced version of a normal RNN and consists of an encoder and a decoder. The computational kernel of both encoder and decoder is an LSTM (long short-term memory unit) or a GRU (gated recurrent unit).

Typical Seq2Seq structure denotes the encoder-decoder framework shown in Figure 7. The working mechanism of a Seq2Seq structure uses the encoder to map the input data to the semantic space to get a decoding vector c which represents the semantics and then use the decoder to obtain the required output.

As shown in Figure 7, the encoder-decoder framework has two inputs: one is $x = \{x_1, x_2, \ldots x_n\}$ which represents the encoder input and the other is $y = \{y_1, y_2, \ldots, y_n\}$ which represents the decoder input. The inputs x and y are sequentially passed to the network in the respective order.

Assuming that the input sequence of the encoder is $\{x_1, x_2, \ldots x_n\}$ and according to the RNN characteristics, the hidden state at time t in the input process is a function of the

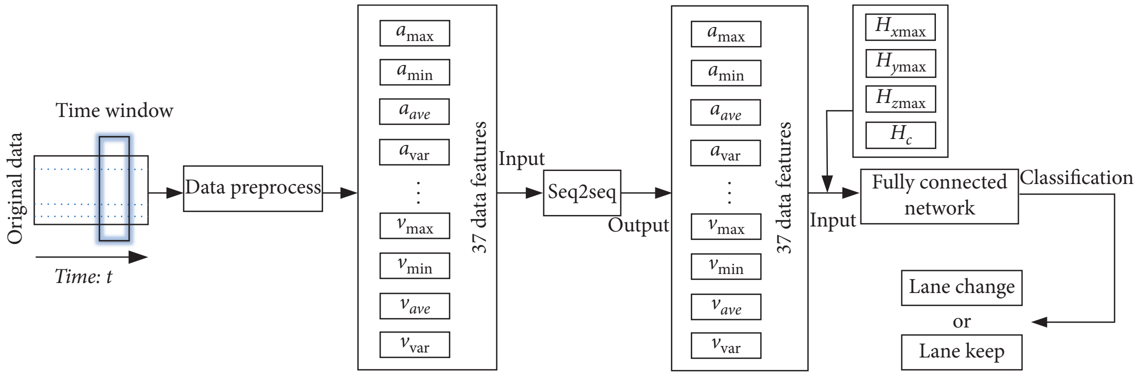


Figure 6: Data processing procedure.

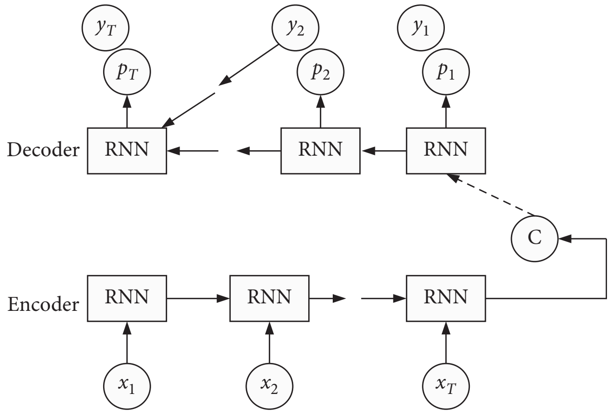


Figure 7: The Seq2Seq decoder-encoder based structure.

current input signal and the hidden layer output in the previous given as

$$h_t = f\left(h_{t-1}, x_t\right), \tag{1}$$

where f is a nonlinear activation function and the decoding vector c is the last state of h_t. Finally, a fixed-length decoding vector c is obtained. The decoder can be regarded as another RNN. When the decoding vector c is sequentially fed to the decoder, the decoder output at the time t can be expressed as

$$h_t = f\left(h_{t-1}, y_{t-1,}c\right). \tag{2}$$

Then, the conditional probability of y at time t is given by

$$p\left(y_t | y_{t-1}, y_{t-2}, \ldots, y_1, c\right) = g\left(h_t, y_{t-1}, c\right). \tag{3}$$

where f and g are the given activation functions, and it must produce a valid probability. The two components of the encoder-decoder structure are jointly trained to maximize the conditional log-likelihood which is given by

$$\max_{\theta} \frac{1}{N} \sum_{n=1}^{N} \log\, p_\theta\left(y_n | x_n\right), \tag{4}$$

where θ is the set of model parameters and $\{x_n\ y_n\}$ denotes the data pair in the training dataset.

Thus, in the Seq2Seq training, y participates in loss calculation and node operation, unlike general RNN, which is used for loss supervision. Assuming the encoder input is $\{x_n\ y_n\}$, the calculation process is as follows:

$$\begin{cases} h_t = z h_{t-1} + (1-z) h_t', \\ h_t' = \left(\tanh\left[W x_t\right] + \left[U\left(r h_{t-1}\right)\right]\right), \\ z = \sigma\left(\left[W_z x_t\right] + \left[U_z h_{t-1}\right]\right), \\ r = \sigma\left(\left[W_r x_t\right] + \left[U_r h_{t-1}\right]\right). \end{cases} \tag{5}$$

In the above formulas, h_t', z, and r represent the intermediate variables, W and U are the training parameters, and σ is the activation function. When the hidden state of the encoding process is completed, the decoding vector is given by

$$c = \tan h\left(V h^n\right), \tag{6}$$

where h^n is the final value of the encoder output after n epochs and V is the training parameter. After the decoding vector c is obtained, the decoder starts the decoding process initializing the initial hidden state $h_0^{''}$ which is given by

$$h_0'' = \tanh\left(V' c\right). \tag{7}$$

The hidden state of the decoder at time t is given by

$$\begin{cases} h_t'' = z' h_{t-1}'' + (1-z')\widetilde{h}_t, \\ \widetilde{h}_t = \tan h\left(\left[W' y_{t-1}\right] + r'\left[U' h_{t-1}'' + Cc\right]\right), \\ z' = \sigma\left(\left[W_z' y_{t-1}\right] + \left[U_z' h_{t-1}''\right] + C_z c\right), \\ r' = \sigma\left(\left[W_r' y_{t-1}\right] + \left[U_r' h_{t-1}''\right] + C_r c\right), \end{cases} \tag{8}$$

where $W, U,$ and C and its deformations are variable training parameters. After obtaining the last hidden state, the condition probability is calculated by

$$p\left(y_t | y_{t-1}, \ldots, y_1, X\right) = \frac{\exp\left(g s_t\right)}{\sum_1^k \exp\left(g s_t\right)}, \tag{9}$$

where k is the output dimension. Also, it holds that

$$\begin{aligned} s_t &= \max\{s_1 s_2, \ldots, s_k\}, \\ s_i &= O_h h_t'' + O_y y_{t-1} + O_c c, \quad (i = 1, 2, \ldots, k), \end{aligned} \tag{10}$$

where $O_h, O_y,$ and O_c are variable training parameters.

Once the Seq2Seq finishes processing a given data sequence, its output is directly fed to the input of the fully connected neural network to perform lane-changing classification.

3.2. Fully Connected Layer. A fully connected network is the most basic and simple neural network, yet such a network performs well in the multiparameter fusion, so it is generally used in the complex nonlinear classification tasks. Since the lane-changing classification represents a nonlinear task including a large-amount input data, a fully connected network is used as a classification network.

In Figure 8, $x(1), x(2), \ldots, x(n-1), x(n)$ are the fully connected layer inputs that are from the Seq2Seq layer output and w_{ji}^{l} represents the weight of the ith synapse of the jth neuron in the lth layer. The induced local domain in the lth layer is

$$v_j^l(n) = \sum_i w_{ji}^l(n) y_i^{(l-1)}(n), \tag{11}$$

where $y_i^{(l-1)}(n)$ is the output of the ith neuron in the previous layer (i.e., the $(l-1)$th layer) after n iterations. For $i = 0$, it holds that $y_0^{(l-1)}(n) = 1$and $W_{j0}^{(l)}(n) = b_j^{(l)}(n)$, where $b_j^{(l)}$denotes the bias of the jth neuron in the lth layer. Using the SoftMax function as an activation function, the output of the jth neuron in the lth layer is given by

$$y_j^{(l)} = \varphi_j\left(v_j(n)\right). \tag{12}$$

If the neuron j is in the first hidden layer, then it holds that

$$y_j^{(0)} = x_j(n), \tag{13}$$

where $x_j(n)$is the jth element of the input vector $x(n)$.

Besides, if neuron j is in the output layer ($l = L$, L is the depth of the network), then it holds that

$$y_j^{(L)} = o_j(n). \tag{14}$$

Therefore, the error is given by

$$e_j(n) = d_j(n) - o_j(n), \tag{15}$$

where $d_j(n)$ is the jth element of the expected response vector$d(n)$.

After the forward propagation is completed, the backpropagation is performed to complete the weight optimization. The backpropagation is given by

$$\delta_j^{(l)}(n) = \begin{cases} e_j^{(L)}(n)\varphi_j'\left(v_j^{(L)}(n)\right), & \text{Output layer } L \text{ neurons } j, \\ \varphi_j'\left(v_j^{(L)}(n)\right)\sum_k \delta_k^{(l+1)}(n) w_{kj}^{(l+1)}(n), & \text{Hidden layer } L \text{ neurons } j. \end{cases} \tag{16}$$

where $\delta_j(n)$is the local gradient and φ_j' is the differentiation of an independent variable. After the local gradient is obtained, the weight updating process is performed. The weight updating process in the nth iteration is given by

$$w_{ji}^{(l)}(n+1) = w_{ji}^{(l)}(n) + \alpha\left[w_{ji}^{(l)}(n-1)\right] + \eta\delta_j^{(l)}(n) y_i^{(l-1)}(n), \tag{17}$$

where α is the momentum constant and η is the learning rate.

3.3. Model Training. According to the survey statistic [33, 34], most drivers are more inclined to the car-following behavior than to the lane changing, so there are not a lot of lane-changing data. Therefore, we had to collect the data needed for the prediction model training and testing. During the data collection process, to avoid the situation where drivers deliberately change the lanes, drivers were not told the true purpose of their trip before the driving started. In addition, uncertain factors such as the driver's driving skills, driving style, and travel purpose will affect the predictive performance of the model. Therefore, in the process of data collection, the research team will provide the driver with some brief driving information (such as recreational driving or emergency driving) when inviting the driver and to make the data more generalized. Different drivers were invited to participate in the data collection process in order to collect as many diverse data as possible under different conditions. After data collection, 7000 features of the lane-changing behavior were extracted from the collected dataset. Since the amount of data was not large, the model was trained and validated by the 10-fold cross-validation method whose pseudocode is shown in Algorithm 1.

We used Google's open source deep learning library TensorFlow to build the hybrid neural network described above. The computational kernel of Seq2Seq was GRU; during the backpropagation training, the Adam optimizer in TensorFlow was used, which automatically adjusts the learning rate parameters during training to avoid overfitting. The initial weights of the entire network were filled with several sets of data obeying the positive distribution, and the initial learning rate is 0.005.

The model training process and its discussion items are shown in Figure 9. As the time window shrinks, the model training process advances. The first step in training is to use the original time window; since the time window is not shrunk, there is no label data for the Seq2Seq. The fully connected neural network is added after the Seq2Seq, and since the lane-changing signs are the label of the entire network, only the accuracy is shown in this step. The second

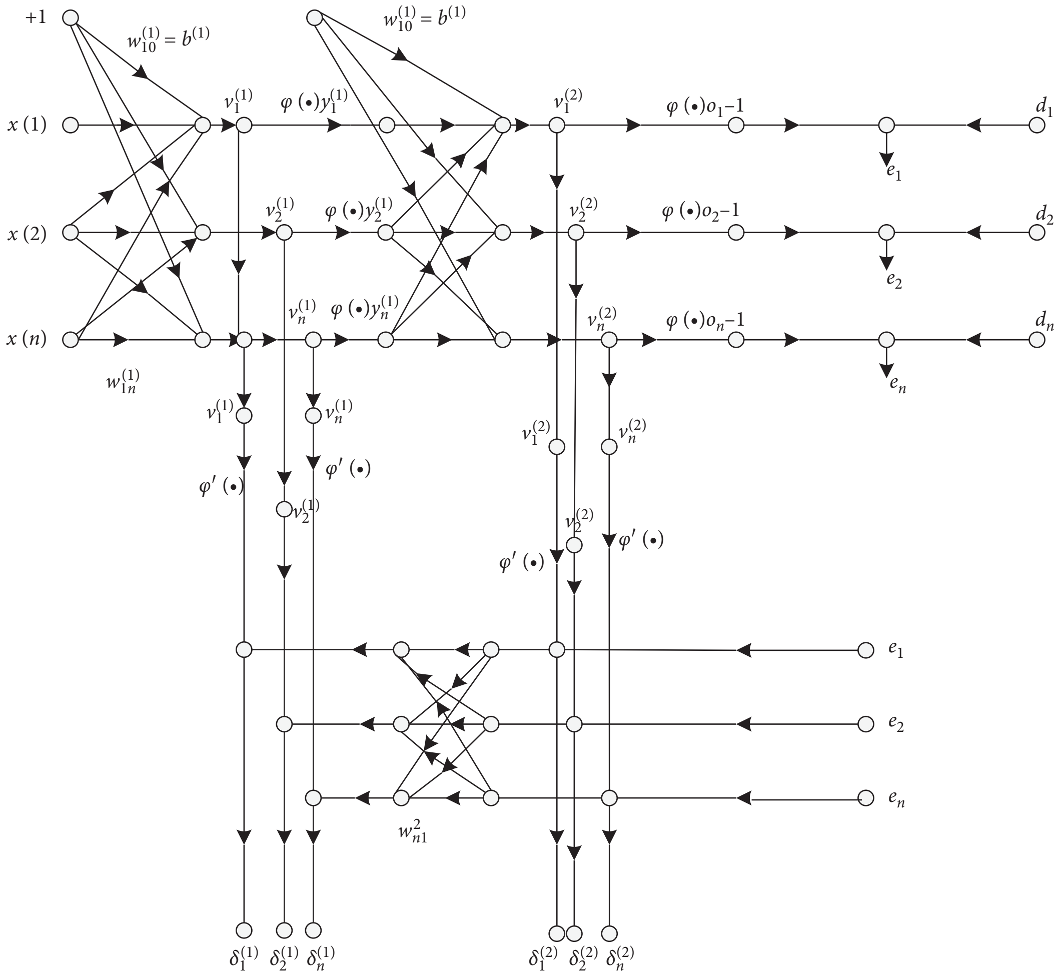


Figure 8: Fully connected neural network data flow diagram.

```
    Input: training set and validation set
    Output: classification results
(1) The entire dataset is randomly divided into 10 parts
(2) Use one of 10 parts as a test set and the other 9 as a training set, and loop beginning at this step
(3) Train the model with 9 training sets and record the relevant data
(4) Test the model with the test set and record the relevant data
(5) Return to 2, select new test set and training set, the sets must differ from the previous ones, and repeat this process until all 10 parts
    are used as a test set, and the loop ends
(6) Compare the real value and prediction value and calculate the prediction accuracy of each training and validation step
(7) Model training/verification complete
```

Algorithm 1: 10-fold cross-validation method.

step is shrinking the time window to 2/5 of the original one, and then, the model's prospective time and accuracy are determined. Similarly, the third step denotes the time window shrinking to 1/5 of the original size and then determines the new model's prospective time and accuracy. After shrinking the time window, the new prediction performance of the Seq2Seq is also displayed. This time window shrink mechanism enables the model to predict lane-changing behavior using only a small part of the header data of the entire lane-changing data, instead of using all lane-changing data for lane-changing recognition. At the same time, when processing the data, we found that, within 1/5 of the data of the original window, lane-changing behavior did not occur from a macroperspective. Therefore, it is reasonable to use this part of data to predict whether the vehicle will change lanes in the future.

When there is no middle label in the model, only the accuracy is discussed. The proposed model structure is presented in Figure 10, and the loss and the accuracy after the iteration are displayed in Figure 11.

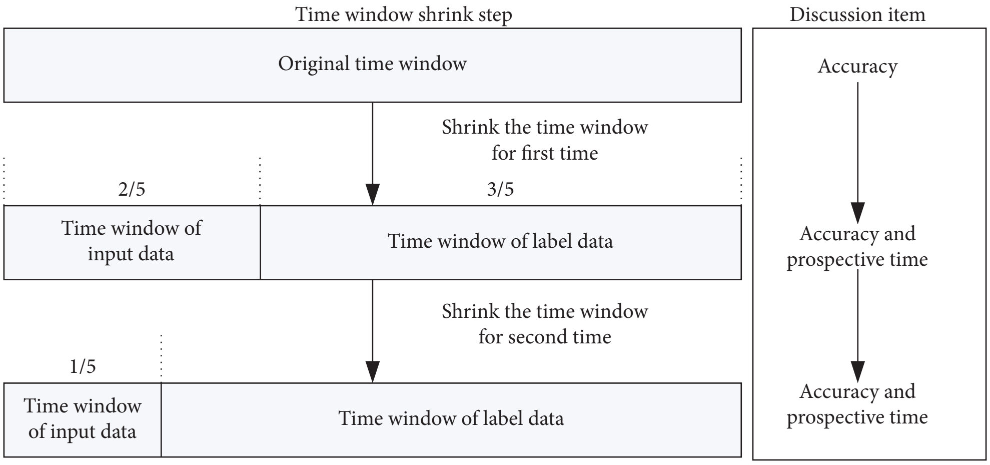

Time window shrink step
Discussion item
Original time window
Accuracy
Shrink the time window for first time
2/5
3/5
Time window of input data
Time window of label data
Accuracy and prospective time
Shrink the time window for second time
1/5
Time window of input data
Time window of label data
Accuracy and prospective time

Figure 9: Model training procedure.

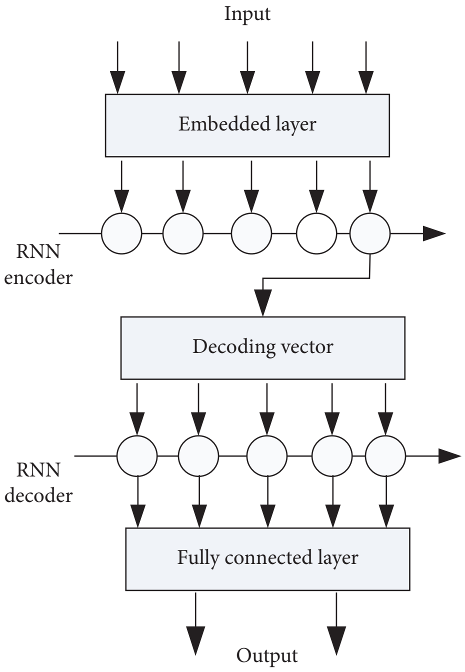
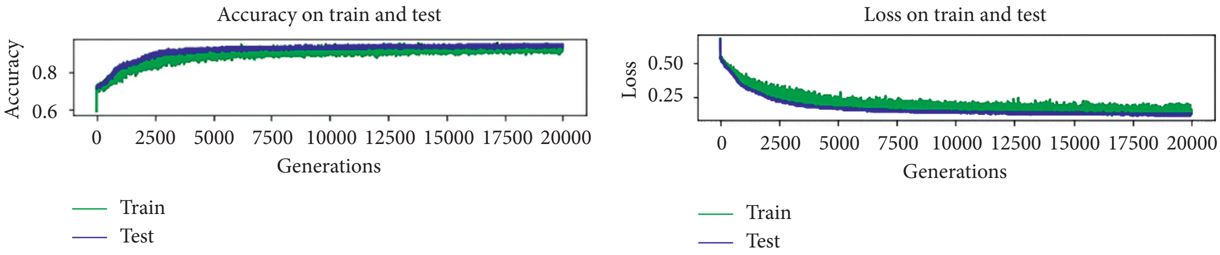


Figure 10: The proposed model structure.



(a)

(b)

Figure 11: The accuracy (a) and the loss (b) after 20,000 iteration.

As presented in Figure 11, after iterating for 20,000 times, the model achieves the convergence state with the accuracy of 0.9858 and the loss of 0.11, which shows that the proposed model can recognize the driver's lane-changing behavior more accurately.

The training process of the entire model after shrinking the time window is given in Figure 9. Seq2Seq uses front data of a time window to predict the caudal data. Because there are series of different input and output data as well the Seq2Seq layer which predicts the driver and vehicle status at the next time point, so the Euclidean distance is used to estimate the level of regression of the prediction result to the real result. The calculation formula is

$$\text{Average deviation} = \frac{\sqrt{\left(\text{actual value}_1 - \text{predictive value}_1\right)^2 + \cdots + \left(\text{actual value}_n - \text{predictive value}_n\right)^2}}{n}. \tag{18}$$

In Figure 12, the average Euclidean distance of the model training for once and twice shrunk time window is presented. It can be noticed that, after 200 iterations, the average deviation began to converge, and the deviation fluctuated within a small range, which shows that the Seq2Seq achieved a good prediction performance. Next, as already explained, the Seq2Seq prediction results were fed to the input of the fully connected network and subjected under the 10-fold cross-validation method. The obtained prediction accuracy and the loss for once shrunk time window are presented in Figure 13. When the number of iterations reached 20,000, the model converged with the accuracy of 0.935 and the loss of 0.12. The model accuracy was reduced due to the data loss caused by time window shrinking, but the prospective time was extended. The accuracy and the loss for twice shrunk time window are presented in Figure 14, where the accuracy increased to 0.938 and the loss was reduced to 0.11.

In Figures 15 and 16, the prospective time for once and twice shrunk time window is, respectively, presented. t_1 and T_1 denote the time needed that a vehicle changes a lane, t_2 and T_2 are the data extraction window for the Seq2Seq training and testing, $\mathrm{t}_3 + \mathrm{t}_4$ and $T_3 + T_4$ are the label extraction window for the Seq2Seq training and testing, and t_3 and T_3 are the time for predicting the lane changing, which is also the prospective time. When $T_3 > t_3$, the prediction method presented in Figure 16 had a longer prospective time than that presented in Figure 15.

In the model verification, the method presented in Figure 15 achieved the prediction accuracy of 0.935, and the model presented in Figure 16 achieved the prediction accuracy of 0.938. Thus, the latter method had better prediction performance. Besides, after statistical analysis of the data used in the paper, we came up with such a mathematical relationship:

$$\begin{aligned} T_2 &= T_4 = \frac{1}{5}T_1, \\ T_3 &= \frac{3}{5}T_1. \end{aligned} \tag{19}$$

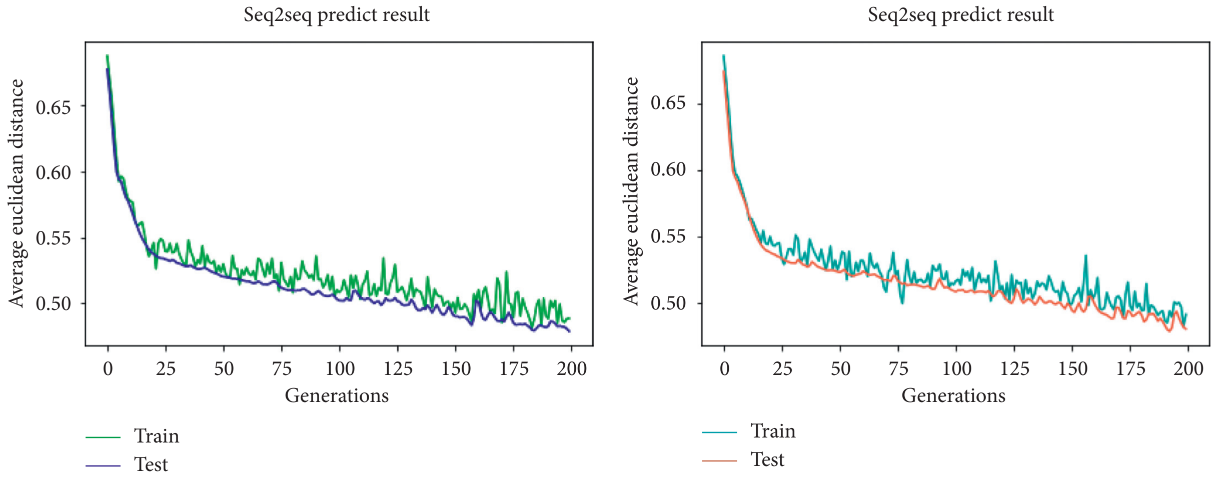
The time cost and the corresponding number of lane changings are presented in Table 3, where the lane changing conducted in the interval 3-4s accounted for 47% of the total number of lane changings, which means the prospective time was 1.8-2.4s. Therefore, the average perspective time of the model is 2.1s. This result indicates that the model could predict the lane changing well and achieve a high prediction accuracy. The comparison between the time cost of lane changing and the prediction time cost after sampling 50 times in all lane changing is presented in Figure 17, and the presented result shows that the prediction time was much shorter than the lane-changing time.

4. Model Validation

After the model was trained, all the weight parameters in the prediction model were optimized. In order to test the generalization ability of the developed model, a different data was used for model training and testing. Also, different drivers were invited to drive vehicles to collect data on another route to obtain the test data. The routes used for validation data acquisition is presented in Figure 18. This route also started at the Chang'an University but ended at the Xi'an Chengbei Passenger Station, and the route direction was different compared with the one used for collection of training data presented in Figure 2. In addition, according to the data scale, the detailed structure and parameters of the cascade model, as shown in Table 4, were determined.

The model was validated using the same data processing method as that used for model training, but the validation included only one input data, without any label, the data was used as the network input, and the focus was on the comparison between the lane changing predicted by the network and the real data. The cross-validation was performed using the data of 3000 lane changings. The accuracy of each validating batch and the average accuracy are presented in Figure 19, where the model accuracy decreased in the first few validating batches, but the overall performance was good, and the average accuracy of the test exceeded 93.5%. Moreover, the prospective time was the same as that of the training. The prospective time for randomly extracted 50 adjacent lane-changing data was 1/5 of the entire lane-changing time, as shown in Figure 20.

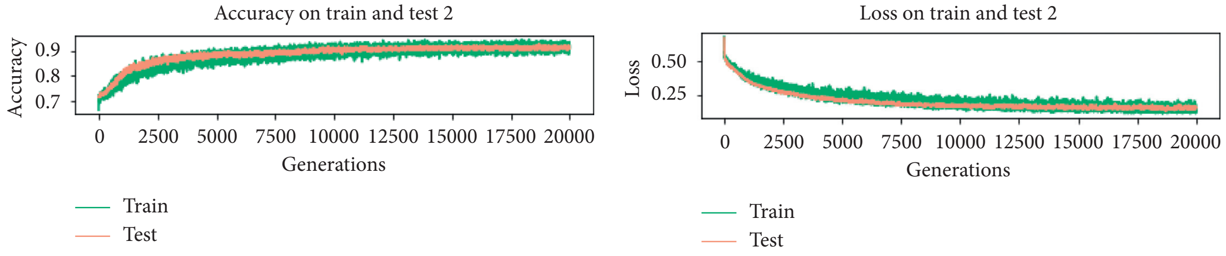
The proposed model was compared with the five most commonly used prediction models. The comparison results are presented in Figure 21 and Table 5. The performance parameters of MST-Net all come from the literature [27], so Table 5 does not include the calculation time of



(a)

(b)

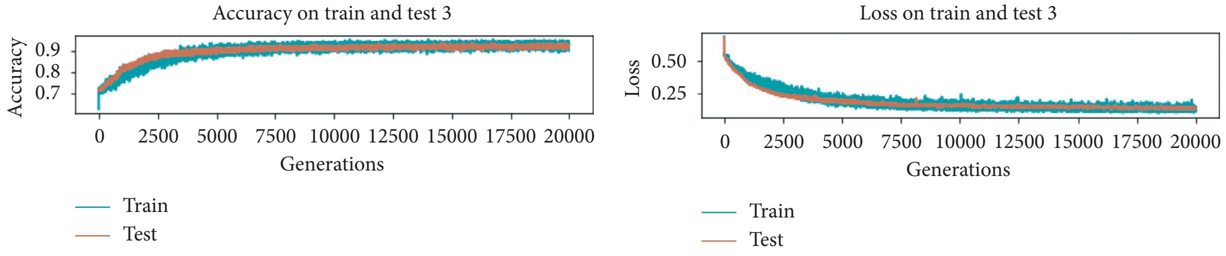
Figure 12: The average deviation for once (a) and twice (b) shrunk time window.



(a)

(b)

Figure 13: The accuracy (a) and the loss (b) for once shrunk time window.



(a)

(b)

Figure 14: The accuracy (a) and the loss (b) for twice shrunk time window.

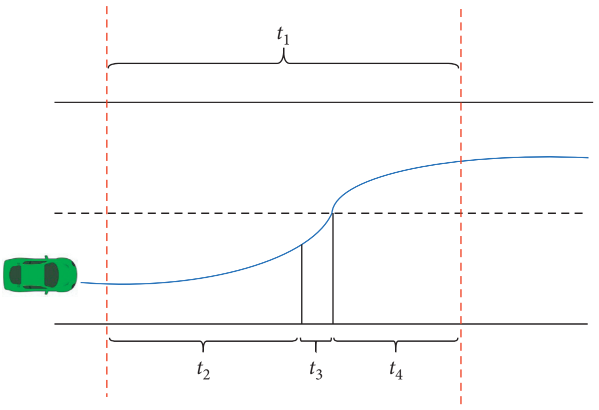


Figure 15: Prospective time for once shrunk time window.

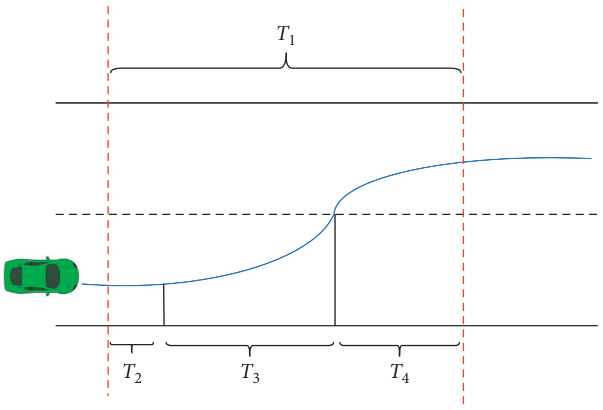


Figure 16: Prospective time for twice shrunk time window.

Table 3: Number of lane changes and time cost.

TC	0-1 s	1-2 s	2-3 s	3-4 s	4-5 s
NoLC	70	1372	1848	3290	420

TC : time cost; NoLC : number of lane changing.

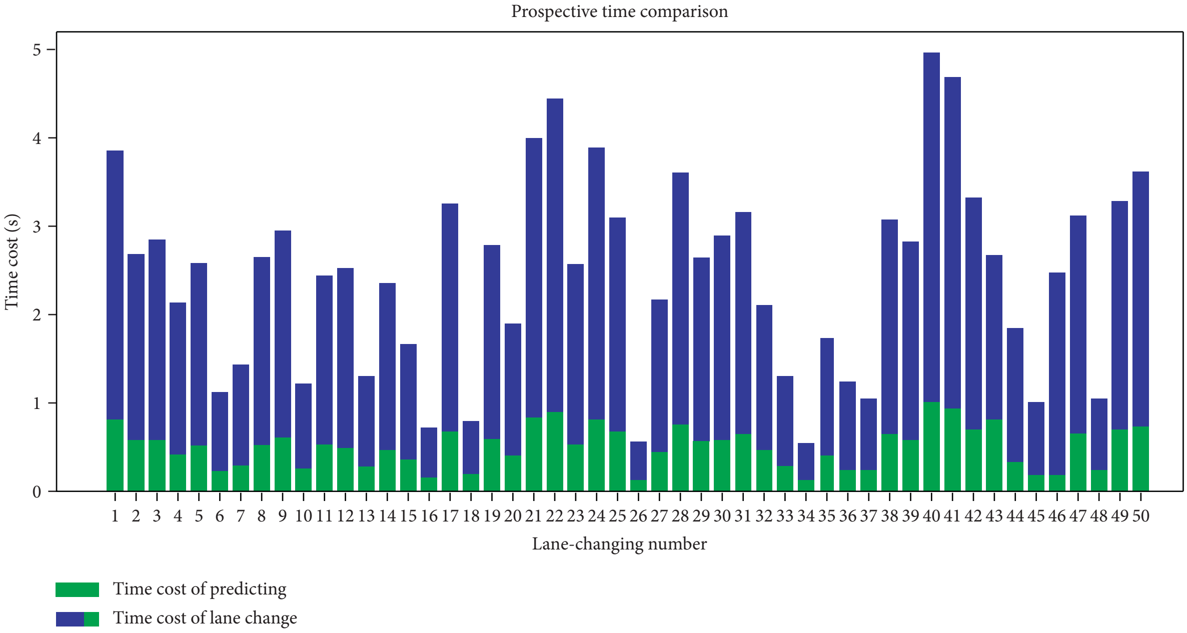


Figure 17: The time cost for different numbers of lane changes in testing.

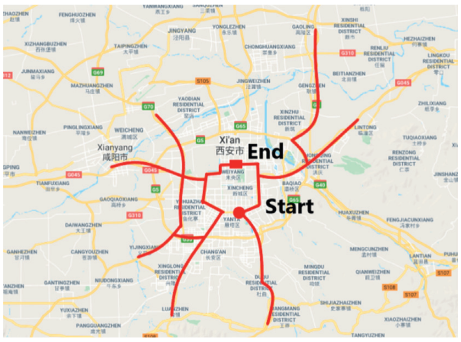


Figure 18: Validation data collection route.

Table 4: Cascade model structure during validation.

Level	Layer name	Number of nodes
1	Encoder-repeat-decoder	37-20-37
2	Input Hidden1 Hidden2 output	41-25-8-2

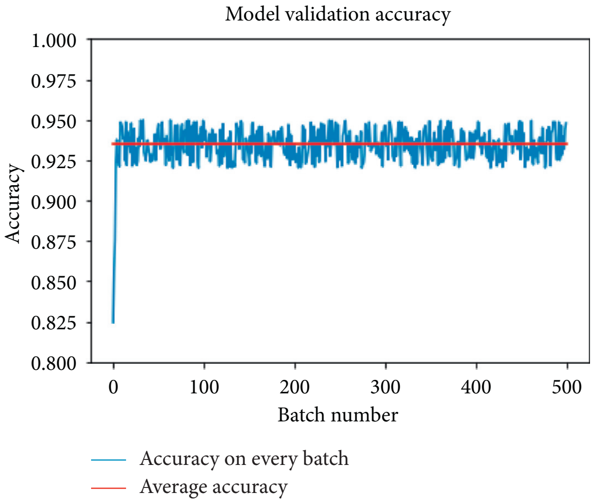


Figure 19: Validation accuracy.

MTSDeepNet. The accuracy of the Dynamic Bayesian Network, the Decision Tree, and the SVM prediction was 75.0%, 84.0%, and 91.1%, respectively, and the accuracy of the BP neural network and the MTSDeepNet was 91.6% and 92.0%, respectively. The accuracy of the proposed Seq2Seq-FC structure was 93.5%, which was better than that of the other five algorithms. Furthermore, although the computation time of the proposed model is higher than several other methods, the microsecond-level increase does not affect applications such as vehicle anticollision which are built on it.

To better illustrate the superiority of the model proposed in the paper, it is necessary to illustrate the data and algorithms used in the six comparison models. As shown in

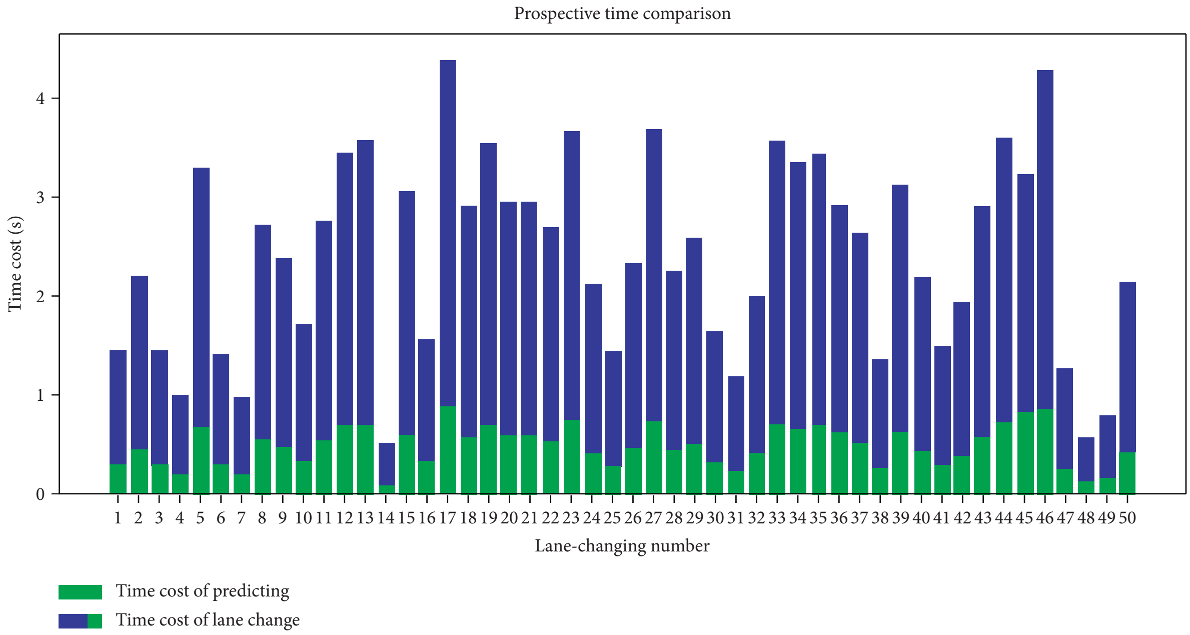


Figure 20: Time cost for different numbers of lane changes in validating.

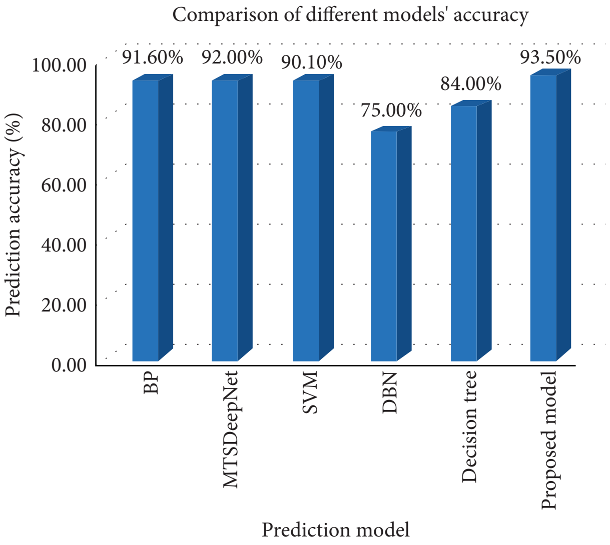


Figure 21: Comparison of the accuracy of different prediction models.

Table 5: Calculation time of different prediction models.

Prediction model	Signal data calculation time	Batch time (ms)
SVM	75us	22.8
DBN	64us	19.7
Decision tree	81us	24.7
Proposed model	160us	47.6

Table 6, it is the detailed information of the six models in the experimental process.

The word "other" means that the algorithms and data used in the comparison model are from other literatures; on the contrary, "this paper" indicates that the algorithms and data used are all derived from this paper. The accuracy of MSTDeepNet comes from [28], and the BP, SVM, DBN, and decision tree use the data which is extracted in this paper and the algorithms used in other literatures. In the experiment, we input the 41 values which are from second shrinking of the time window into different classification algorithms to obtain the above accuracy. The comparison results show that this paper has certain advantages in terms of data or prediction model.

Table 6: Details of the six comparison models.

Prediction model	Data used in the model	Algorithm used in model
MSTDeepNet	Other [28]	Other [28]
BP	This paper	Other [22]
SVM	This paper	Other [15]
DBN	This paper	Other [20]
Decision tree	This paper	Other [35]
Proposed model	This paper	This paper

5. Conclusion

In this paper, a Seq2Seq-FC neural network for prediction of driver lane-changing behavior is introduced. The proposed model has two levels, where the first level denotes the Seq2Seq network whose function is to process the time series data and the second level denotes a fully connected neural network which works as a nonlinear classifier. In the proposed prediction model, the vehicle kinematics data (VKA), the drivers' kinematics data (DKA), and the drivers' physiology data (DPA) are used as input data for the fully connected network. In addition, a dynamic time window is proposed to extract the features of the lane-changing process, and the method of time window shrinking is successively used to train the prediction model and improve the prospective time. Moreover, model testing was performed by data different from that used for model training to evaluate the model generalization ability. The test results showed that the proposed prediction model achieved a good performance.

In the data collection process, 35 drivers took part, and different routes were used for training and test data. The collected data consisted of 10,000 lane-changing samples, of which 7000 samples were used for model training, and the rest was used for model validating. The validating results proved the effectiveness and stability of the proposed model. Moreover, the proposed Seq2Seq-FC model was compared with five common prediction models: BP neural network, SVM, dynamic Bayesian network, decision tree, and MTSDeepNet. The comparison results showed that the Seq2Seq-FC network achieved higher prediction accuracy and longer prospective time than other models. The results presented in this study can be helpful to improve the practical effect of the ADAS and enhance lane-changing safety.

In our future research, we will improve the proposed model from several aspects. Firstly, many researchers have demonstrated that driver's decision to change a lane is also affected by vehicle type and driver's driving skills [36, 37]; for instance, a car has different lane-changing factor compared with a bus. However, in this study, we did not consider vehicle type because we used the vehicles of the same type. Therefore, in our future work, we will take different types of vehicles into account. Secondly, although many different road conditions were included in the traffic route, some of the road types such as rugged mountain road and country road were still not included, but they will be considered in our future research. Thirdly, since the neural network considers the fuzzy relationship between input and output, input data redundancy can be caused, and model calculation speed can be reduced. However, the sensitivity analysis can be used to eliminate some variables that have little effect on model classification, which may make the model more optimized and concise.

Data Availability

The data used to support the findings of the study are available from the first author ChengWei (chengwei@chd.edu.cn) upon reasonable request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by National Key R&D Program of China (Grant no. 2018YFB1600604).

References

[1] S. Gong and L. Du, "Optimal location of advance warning for mandatory lane change near a two-lane highway off-ramp," *Transportation Research Part B: Methodological*, vol. 84, pp. 1–30, 2016.

[2] H. Yu, H. E. Tseng, and R. Langari, "A human-like game theory-based controller for automatic lane changing," *Transportation Research Part C: Emerging Technologies*, vol. 88, pp. 140–158, 2018.

[3] M. Keyvan-Ekbatani, V. L. Knoop, and W. Daamen, "Categorization of the lane change decision process on freeways," *Transportation Research Part C: Emerging Technologies*, vol. 69, pp. 515–526, 2016.

[4] A. Yasir, Z. Zuduo, and H. Mazharul, "Connectivity's impact on mandatory lanechanging behavior: evidences from a driving simulator study," *Transportation Research Part B: Methodological*, vol. 93, pp. 292–309, 2018.

[5] J.-B. Sheu and S. G. Ritchie, "Stochastic modeling and real-time prediction of vehicular lane-changing behavior," *Transportation Research Part B: Methodological*, vol. 35, no. 7, pp. 695–716, 2001.

[6] Y. Dou, F. Yan, and D. Feng, "Lane changing prediction at highway lane drops using support vector machine and artificial neural network classifiers," in *Proceedings of the IEEE International Conference on Advanced Intelligent Mechatronics IEEE*, Delft, The Netherlands, 2016.

[7] H. Liu, H. Wei, T. Zuo, Z. Li, and Y. J. Yang, "Fine-tuning ADAS algorithm parameters for optimizing traffic safety and mobility in connected vehicle environment," *Transportation Research Part C: Emerging Technologies*, vol. 76, pp. 132–149, 2015.

[8] A. Simic, O. Kocic, M. Z. Bjelica, and M. Milosevic, "Driver monitoring algorithm for advanced driver assistance systems," *TELFOR*, vol. 7, 2016.

[9] Y. Saito, M. Itoh, and T. Inagaki, "Driver assistance system with a dual control scheme: effectiveness of identifying driver drowsiness and preventing lane departure accidents," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 5, pp. 660–671, 2016.

[10] M. Errampalli, M. Okushima, and T. Akiyama, "Fuzzy logic based lane change model for microscopic traffic flow simulation," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 12, no. 7, pp. 172–181, 2008.

[11] W. Shi and Y. P. Zhang, "Decision analysis of lane change based on fuzzy logic," *Applied Mechanics and Materials*, vol. 419, pp. 790–794, 2013.

[12] E. C. B. Olsen, *Modeling Slow Lead Vehicle Lane Changing*, Virginia Polytechnic Institute and State University, Blacksburg,VA, USA, 2013.

[13] M. Baumann and J. F. Krems, "Situation awareness and driving: a cognitive model," *Modelling Driver Behaviour in Automotive Environments*, vol. 12, pp. 253–265, 2007.

[14] D. D. Salvucci, "Modeling driver behavior in a cognitive architecture," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 48, no. 2, pp. 362–380, 2006.

[15] P. Kumar, M. Perrollaz, S. Lefevre, and C. Laugier, "Learning-based approach for online lane change intention prediction," *Intelligent Vehicles*, vol. 4, pp. 797–802, 2013.

[16] W. Yao, H. Zhao, P. Bonnifait, and H. Zha, "Lane change trajectory prediction by using recorded human driving data," *Intelligent Vehicles Symposium*, vol. 7, pp. 430–436, 2013.

[17] Y. Hou, P. Edara, and C. Sun, "A genetic fuzzy system for modeling mandatory lane changing," *Intelligent Transportation Systems*, vol. 12, pp. 1044–1048, 2012.

[18] K. Li, X. Wang, Y. Xu, and J. Wang, "Lane changing intention recognition based on speech recognition models," *Transportation Research Part C: Emerging Technologies*, vol. 69, pp. 497–514, 2016.

[19] Y. Zheng and J. H. L. Hansen, "Lane-change detection from steering signal using spectral segmentation and learning-based classification," *IEEE Transactions on Intelligent Vehicles*, vol. 2, no. 1, pp. 14–24, 2017.

[20] D. Kasper, G. Weidl, T. Dang et al., "Object-oriented bayesian networks for detection of lane change maneuvers," *IEEE Intelligent Transportation Systems Magazine*, vol. 4, no. 3, pp. 19–31, 2012.

[21] D.-F. Xie, "A data-driven lane-changing model based on deep learning," *Transportation Research Part C Emerging Technologies*, vol. 106, pp. 41–60, 2019.

[22] R. M. H. Cheng, J. W. Xiao, and S. Lequoc, "Neuromorphic controller for AGV steering," in *Proceedings of the 2010 IEEE International Conference on Robotics & Automation*, IEEE, Anchorage, AK, USA, 1992.

[23] R. S. Tomar, S. Verma, and G. S. Tomar, "Prediction of lane change trajectories through neural network," in *Proceedings of the 2010 International Conference on Computational Intelligence and Communication Networks*, pp. 125–146, Bhopal, India, 2010.

[24] C. Ding, W. Wang, X. Wang, and M. Baumann, "A neural network model for driver's lane-changing trajectory prediction in urban traffic flow," *Mathematical Problems in Engineering*, vol. 2013, Article ID 967358, 2013.

[25] J. Zheng, K. Suzuki, and M. Fujita, "Predicting driver's lane-changing decisions using a neural network model," *Simulation Modelling Practice and Theory*, vol. 42, pp. 73–83, 2014.

[26] J. Peng, Y. Guo, R. Fu, W. Yuan, and C. Wang, "Multi-parameter prediction of drivers' lane-changing behaviour with neural network model," *Applied Ergonomics*, vol. 50, pp. 207–217, 2016.

[27] Y. Dou, F. Yan, and D. Feng, "Lane changing prediction at highway lane drops using support vector machine and artificial neural network classifiers," in *Proceedings of the IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, pp. 901–906, Delft, The Netherlands, 2016.

[28] X. Wang, Y. L. Murphey, and D. S. Kochhar, "MTS-DeepNet for lane change prediction," in *Proceedings of the International Joint Conference on Neural Networks*, pp. 4571–4578, 2016.

[29] X. Li, W. Wang, and M. Roetting, "Estimating driver's lane-change intent considering driving style and contextual traffic," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 9, pp. 3258–3271, 2019.

[30] Y. Qiao, K. Hashimoto, A. Eriguchi, H. Wang, D. Wang, and Y. Tsuruoka, "Parallelizing and optimizing neural encoder-decoder models without padding on multi-core architecture," *Future Generation Computer Systems*, vol. 108, pp. 1206–1213, 2020.

[31] L. Zheng zhong and S. G. D. O. David, "The impact of encoding–decoding schemes and weight normalization in spiking neural networks," *Neural Networks*, vol. 108, pp. 365–378, 2019.

[32] H. Mårtensson, O. Keelan, and C. Ahlström, "Driver sleepiness classification based on physiological data and driving performance from real road driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 2, pp. 421–430, 2019.

[33] C. Marina Martinez, M. Heucke, F. Y. Wang, B. Gao, and D. Cao, "Driving style recognition for intelligent vehicle control and advanced driver assistance: a survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 3, pp. 666–676, 2018.

[34] Z. Zheng, "Recent developments and research needs in modeling lane changing," *Transportation Research Part B: Methodological*, vol. 60, no. 1, pp. 16–32, 2014.

[35] Y. Li, M. Dong, and R. Kothari, "Classifiability-based omnivariate decision trees," *IEEE Transactions on Neural Networks*, vol. 16, no. 6, pp. 1547–1560, 2019.

[36] E. Suzdaleva and I. Nagy, "An online estimation of driving style using data-dependent pointer model," *Research Part C: Emerging Technologies*, vol. 86, pp. 16–23, 2019.

[37] P. Cai, Y. Wang, G. Lu, P. Chen, C. Ding, and J. Sun, "A spatiotemporal correlative k-nearest neighbor model for short-term traffic multistep forecasting," *Transportation Research Part C: Emerging Technologies*, vol. 62, pp. 21–34, 2019.

Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 5568777, 16 pages
https://doi.org/10.1155/2021/5568777

Research Article

Bottom-Up Approach Ship Emission Inventory in Port of Incheon Based on VTS Data

Hyangsook Lee,[1] Hoang T. Pham,[1] Maowei Chen,[1] and Sangho Choo[2]

[1] *Graduate School of Logistics, Incheon National University, Incheon 22012, Republic of Korea*
[2] *Department of Urban Design & Planning, Hongik University, Seoul 04066, Republic of Korea*

Correspondence should be addressed to Sangho Choo; shchoo@hongik.ac.kr

Received 23 January 2021; Revised 24 February 2021; Accepted 19 March 2021; Published 10 April 2021

Academic Editor: Dong-Kyu Kim

As a result of the rapid growth of international trade, atmospheric pollution from transportation has been more topical than ever, especially in dense hub port-cities. The shipping industry should pay more attention corresponding to its contribution to local atmospheric pollution. This paper supports the application of data collected from the vessel tracking service system with a bottom-up approach to generate a comprehensive 2019 local ship emission inventory at Port of Incheon. The calculated emission inventory presented the dominance of CO_2 emission and the considerable contribution of NO_x and SO_x emissions, the significant contribution of auxiliary engines during the hotelling at berth during the year of 2019. Then, based on calculated emission inventory, this study suggested and simulated applicable green policies in the practice: (1) local emission control area realization, (2) vessel speed reduction program, (3) application of cold ironing, and (4) establishment of a national integrated emission platform. The combination of the three first policies could help reduce the significant volume of emitted CO (29%), NO_x (30%), SO_x (93%), PM10 and PM2.5 (64%), VOC (28%), NH_3 (30%), and CO_2 (30%).

1. Introduction

Transportation has been considered a remarkable contributor to atmospheric pollution [1]. The recent decades have detected a growing consideration of a leading global anthropogenic emissions contribution from in-port traffic [2]. With 90% of the global transport volume being carried by ship, maritime air pollution has exacerbated local society's health because of the high volume of emissions exhausted as well as the geographical condition. Approximately 70% of ship emissions are discharged within a radius of 400 km from the coastline [2]. As the result of the hub-port trend, ships often concentrate on well-located ports, where they often are situated close to busy industrial zones and densely populated cities. In-port ship emissions could supply 55–77% of total local emissions [3, 4]. Therefore, these port-cities and their hinterlands are grappling with a great burden from transport emissions and infrastructure [5]. It is estimated that approximately 230 million people, living in the top 100 world ports, are directly affected by exhaust shipping emissions [6]. Thus, the port industry has kept the momentum in green reform by reducing ship emissions and performs more of their social responsibility under higher pressure from public authorities [7, 8].

More much than it seems, it is stated that there are 450 different air pollutants which are emitted by ship engine combustion [9]. Greenhouse gases (GHGs), carbon monoxide (CO), nitrogen oxides (NO_x), sulphur oxides (SO_x), and particulate matter (PM) are stated as key ship-source air pollutants [10, 11]. Shipping SO_x and NO_x contribute markedly roles compared to other air pollutants in total national emissions [12]. Besides a negative affection to global climate, numerous research articles have proved a considerable negative tie between closed-to-land ship emissions and the local community's health (lead to asthma, respiratory and cardiovascular diseases, lung cancer, and premature death) [13]. Exhausted PM emissions from shipping activities are considered the main reason for the annual 60,000 cardiopulmonary and lung cancer deaths in Europe, East Asia, and South Asia littoral regions [14].

To reduce air pollution from ships, emission control legislations were contemplated by international regulation-making agencies. Likewise, the International Maritime Organization (IMO) is reducing ship emission through its International Convention on the Prevention of Pollution from Ships (MARPOL) Annex VI. The sulphur content in the fuel used onboard is reduced from 3.5% m/m to 0.5% m/m from 1 January 2020 [15]. The sulphur directive in the European Union (EU), the fourth federal standard, also called Tier 4, in the United States (U.S.), and emission control areas (ECAs) in China are promulgated to follow this regulation [16–18]. It is reported that the implementation of these regulations helped reduce 77% SO_x, 23% NO_x, and 27% PM10 emissions in the EU by 2017, compared with 2000 [19]. Despite the expected expansion in shipping, stricter emissions regulations could reduce the volume of SO_x, PM2.5, and NO_x emissions by 87%, 92%, and 56%, respectively, between 2015 and 2050 [20].

In Korea, the public has been enhancing national air quality through its 10-year comprehensive plans since the 1990s. Air quality management basic plan during 2015–2024 is established to manage an integrated air management system. Besides, in 2017, an other comprehensive plan of particular matter management was announced with an epicenter of 30% PM emission reduction and reinforcement control of the volatile organic compounds (VOCs) until 2022. [21] From December 2019, a vessel speed reduction (VSR) program has been conducted to reduce the PM level during winter in five key ports of Korea, including Port of Incheon (POI). A VSR area in each port spans 20 nm in radius [22]. However, almost all national policies and green measures are decided with the support of annual national emission inventories provided in the clean air policy support system (CAPSS), using national fuel supply statistics. However, these inventories revealed inconsistency and uncertainty with other local academic research articles, because of no actual traffic data [23]. Quantifying emitted emissions from shipping activities is a stride to explore the cause and scale of marine pollution [24]. It also provides valuable statistics to anticipate the future trend of pollution and then establishes counter-measures and policies [25]. Therefore, generating a more reliable in-port ship emission inventory is a pressing issue for producing appropriate regional policies and measures related to air pollutant management at ports.

In this study, a comprehensive annual ship emission inventory of target pollutants of CAPSS at POI in 2019 was estimated, according to geographical areas, operational phases, and ship types. This study applied another ship tracking system named vessel tracking service (VTS), operated by port authorities or coast guards to keep track of ship traffic and ensure navigational safety in the port area. This emission inventory applied a bottom-up approach with both EO and FC methods to see the difference in the estimating results of the two methods and then achieved a comprehensive multidimensional understanding of air pollutants from in-port ship operation. Then, applicable green policies for the POI situation in the practice will be suggested and simulated to evaluate their effects based on the calculated emission inventory. The remainder of this study is grouped as follows: Section 2 summarizes the previous studies related to ship-related emissions and then points out the research demand of this paper. Materials preparation is introduced in Section 3. Section 4 describes two applied methodologies for ship emission prediction. Section 5 presents and discusses calculation results. Several policies to reduce the volume of emissions emitted from ships operated at POI are suggested and simulated in Section 6 and Section 7 is the conclusion of this study.

2. Literature Review

Numerous previous studies have suggested estimation methods to generate ship emissions. The mainstream methods to prepare ship emission inventory can be broadly categorized into two main approaches: top-down approach (fuel-based) and bottom-up approach (activity-based). In the top-down approach, exhausted ship emissions are estimated from analyzed statistically marine fuel sales and fuel-related emission factors [26]. The marine fuel sale reports are mainly published by the Energy Information Administration, the International Energy Agency, and the United Nations Framework Convention on Climate Change [27]. This approach is recommended for a low level of traffic data availability situation. Due to the positive correlation between fuel consumption and emissions discharged from engine combustion, this approach would be the most accurate method if researchers could be confident about the marine fuel sales data collected. However, it proved that there is a significant discrepancy between banker fuel sales statistics and the actual fuel used by the global fleets [28]. The main reasons behind this difference could be the practice of offshore tinkering or the inaccurate fuel statistics in several countries [27].

Due to difficulties in collecting reliable inputs for the top-down approach, the bottom-up approach is recommended as another choice if accurate sailing statistics (e.g., actual travel distance with speed and port calling records with real-time operations) are available. Also, this method requires a higher level of input parameters such as detailed ship technical data (e.g., ship types, engine characteristics, and design information). An amount of emitted emissions is calculated for each specific ship activity and then scaled up over activities and trips to figure out the total volume of emissions [29]. The generic emission estimation equation of bottom-up can be expressed in the following equation:

$$E = \text{Energy}.\text{EF}\,(.\text{CF}), \tag{1}$$

where E is the emission amount of certain pollutants from the ship's engines; Energy is the energy demand; EF is the emission factor; and CF is a controlling factor used when the ship is equipped with reduction technologies or fuel correction factors. In general, the estimation of energy demand can be dichotomized into the following:

(a) The total energy output (EO) of operated engines during the operating time, by applying the following equation:

$$\text{Energy} = P.\text{LF}.T. \tag{2}$$

(b) The total engines' fuel consumption (FC) during the operating time, using the following equation:

$$\text{Energy} = \text{SFOC}\,(.\text{LF}).T, \tag{3}$$

where P denotes engine power; LF denotes load factor; T denotes activity time; and SFOC denotes specific fuel consumption.

Bottom-up approach ship-related emission studies since 2009 are summarized in Table 1. As shown in Table 1, the EO method dominated the bottom-up approach for estimating ship-related emissions, especially from 2014. The most considered pollutants in previous studies are CO_2, NO_x, CO, SO_x, and PM.

It is widely agreed that the bottom-up approach is generally more accurate than the former [52]. Therefore, over the last two decades, the latter has been adopted frequently by researchers [26]. Inaccuracy estimation from the top-down approach at the national level could lead to a completely different understanding and measures at the regional air quality management level. Therefore, at the regional level, the bottom-up approach is recommended to be a reliable source to establish effective green policies for better air condition [40, 53, 71, 72]. However, for larger scales, applications of the bottom-up approach have been finite because of data gaps and anomalies [52, 54]. Also, the diversity and details in input information could help improve the reliability of results; however, its corresponding complexity also increases the level of uncertainty, which is associated with the different applied models and assumptions, and how to collect the set of inputs [27, 29]. Especially, on a global scale, the use of average input parameters such as different load factors, operational activity time, and emission factors for a different size, age, and type of ships leads to considerable uncertainty in estimation [54, 73].

Recently, a vessel tracking system named automatic identification system (AIS), proposed in the IMO International Convention of Safety of Life at Sea (SOLAS) for better identification of ships, has been widely applied in the stream to enhance estimations [44, 46, 49, 51, 52, 56, 74] and considered as "the best method to report the activities and movements of ships" [26]. This system is equipped compulsively on passenger ships of all sizes and commercial ships with 300 gross tonnage (GT) or more operated at sea. Therefore, high-resolution ship movement AIS data could be a fount of reliable relative ship operational profile as vessel travel time and average speed between waypoints on the sea at short-time intervals and identify ship routes. Also, AIS collects and transmits ship characteristics as IMO identification number, size, weight, etc. [52]. Hence, ship's activities are better geographically characterized and analyzed, and consequently, it improves the reliability of emission inventory investigation. Despite the abovementioned benefits, the AIS system does not cover ships with less than 300 GT. Besides, with ships greater than or equals to 300 GT, in several cases, AIS data could not be fully accessible. Time gaps also occur in several cases if temporary signal disruption happens, resulting in erroneous position reports.

During the last decades, several investigations about Korean port-related ship emissions have been conducted domestically. At the national level, under CAPSS, national air pollutants emission inventories, including in-port ship emissions, had been updated annually, and the latest update is for the year 2016 [75]. However, CAPSS only considers hotelling at anchorage and berth and maneuvering process in port emission. The equations with values of parameters, provided by NIER [76], follow a top-down approach based on fuel consumption and are shown in the following equations:

$$E = \sum_{\text{ph}} \text{FC}_{\text{ph}} \times \text{EF}, \tag{4}$$

$$\text{FC}_{\text{hotelling}} = N \times \text{DF} \times 0.79 \times 0.2, \tag{5}$$

$$\text{FC}_{\text{manoeuvring}} = \frac{\sum (2 \times N \times D)}{\text{FE}}, \tag{6}$$

where FC represents fuel consumption; N represents annual total ship call; DF is average daily fuel consumption by ship type (ton/day); D represents average travel distance in port (km); FE represents average fuel economy by ship tonnage (km/kL); 0.79 is the average hotel time of a ship call per day; and 0.2 means the fuel consumption of hotelling at anchorage and berth will be assumed as 20% of the fuel consumption of full operation.

At the regional level, Shin and Cheong [39] generated the GHGs ship emissions at Port of Busan (POB) with a top-down approach. Chang et al. [44] firstly used the bottom-up approach to assess the GHGs emissions from ship operations at the Port of Incheon (POI) from January to October 2012 based on the FC method. Later, Chang et al. [49] continued to apply the FC method to update NO_x, SO_x, and PM emissions released from ships at POI during the same period. With the same method, Khan et al. [61] collected ship traffic data at POI from the AIS system to figure out ship GHG emissions, however, during only October 2014. Kwon et al. [1] developed a system for emission estimation and then considered the amount of CO, NO_x, SO_x, and PM10 emitted from ships around POI only in December 2017. Lee et al. [70] estimated a non-CO_2 emission inventory in POI during 2017 using VTS data and the EO method.

This paper supports the idea of the application of VTS data in the bottom-up approach studies as a promising confident alternative suggested by Lee et al. [70]. However, the contribution of this paper is that

(1) On one face, this paper considers an FC method, which was used to estimate emissions in POI several times before by Chang et al. [44], Chang et al. [49], and Khan et al. [61]. However, all of these three papers did not provide an up-to-date and full-year estimation. On the other hand, this paper also considers the EO method applied by Lee et al. [70]; however, this paper investigates additionally CO_2 emission, which is the most important pollutant but

Table 1: Summary of bottom-up approach studies.

Author	Method	Operation modes	Considered pollutants	General estimating equation
Kwon et al. [1]	EO	Cru, Man, Hot	NO$_x$, CO, SO$_x$, PM10	E = P. LF. T. EF
Song [2]	EO	Cru, Man, Hot	NO$_x$, CO, SO$_x$, CO$_2$, CH$_4$, N$_2$O, HC, PM10, PM2.5	E = P. LF. T. EF. CF
Sanabra et al. [5]	EO	Cru, Man, Hot	SO$_2$, PM2.5, NO$_x$, VOCs	E = P. LF. T. EF
Merk [6]	EO	Cru, Man, Hot	CH$_4$, CO, CO$_2$, NO$_x$, PM10, PM2.5, SO$_x$	E = P. LF. T. EF
EEA [12]	EO	Cru, Man, Hot	NO$_x$, CO, SO$_x$, NMVOC, TSP, PM10, PM2.5,. . .	E = P. LF. T. EF
Song and Shon [23]	EO	Cru, Man, Hot	SO$_2$, NO$_x$, CO$_2$, VOC, PM	E = P. LF. T. EF
ICF International [30]	EO	Cru, Dec, Man, Hot	NO$_x$, CO, SO$_x$, CO$_2$, HC, CH$_4$, PM10, PM2.5, N$_2$O	E = P. LF. T. EF
Corbett et al. [31]	FC	N/a	CO$_2$	E = SFOC. LF. T. EF
Joseph et al. [32]	EO	Cru, Man, Hot	NO$_x$, SO$_2$, TSP, PM10	E = P. LF. T. EF
Deniz and Kilic [33]	FC	Cru, Man, Hot	NO$_x$, SO$_2$, CO, CO$_2$, VOC, PM	E = SFOC. T. EF
Deniz et al. [34]	EO	Cru, Man, Hot	NO$_x$, SO$_2$, CO$_2$, HC, PM	E = P. LF. T. EF
Howitt et al. [35]	FC	N/a	CO$_2$	E = P. LF. SFOC. T. EF
Kiliç and Deniz [36]	EO	Cru, Man, Hot	NO$_x$, SO$_2$, CO$_2$, HC, PM	E = P. T. EF
Lonati et al. [37]	FC	Man, Hot	NO$_x$, SO$_x$, CO, VOC, PM10	E = SFOC. T. EF
Tzannatos [38]	EO	Man, Hot	NO$_x$, SO$_2$, PM	E = P. LF. T. EF
Shin and Cheong [39]	FC	Man, Hot	CO$_2$, N$_2$O, CH$_4$	Man: E = D. FE$^{-1.}$ EF Hot: E = (FC. 0.2). 0.79. EF
Villalba and Gemechu [40]	EO	Man, Hot	CO$_2$	E = P. LF. T. EF
Chang and Wang [41]	FC	Dec	NO$_x$, SO$_2$, CO$_2$, HC, PM	E = SFOC. LF. T. EF
Berechman and Tseng [42]	EO	Hot	NO$_x$, CO, SO$_2$, CO$_2$, HC, VOC, PM10, PM2.5	E = P. LF. T. EF
Yau et al. [43]	EO	Cru, Dec, Man, Hot	NO$_x$, SO$_2$, PM10	E = P. LF. T. EF
Chang et al. [44]	FC	Cru, Man, Hot	CO$_2$	E = SFC. LF. T. EF
McArthur and Osland [45]	EO	Cru, Dec, Man, Hot	NO$_x$, SO$_2$, PM10	E = P. LF. T. EF
Ng et al. [46]	EO	Cru, Dec, Man, Hot	SO$_2$, NO$_x$, CO, VOC, PM10	E = P. LF. T. EF
Saraçoğlu et al. [47]	EO	Cru, Man, Hot	NO$_x$, SO$_2$, CO$_2$, HC, PM	E = P. LF. T. EF
Tai and Lin [48]	FC	Cru, Man, Hot	NO$_x$, SO$_2$, CO$_2$, HC, PM	E = T. FE. EF
Chang et al. [49]	FC	Cru, Man, Hot	NO$_x$, SO$_2$, PM	E = SFC. LF. T. EF
Liu et al. [50]	FC, EO	Man, Hot	SO$_2$	E = FC. EF
E = P. LF. T. EF. CF				
Goldsworthy and Goldsworthy [51]	EO	Cru, Man, Hot	NO$_x$, CO, SO$_2$, CO$_{2e}$, PAH, VOC, PM10, PM2.5	E = P. LF. T. EF
Coello et al. [52]	EO	N/a	CO$_2$, NO$_x$, CO, NMVOC, SO$_x$, PM	E = P. LF. T. EF
Tichavska and Tovar [53]	EO	Cru, Man, Hot	CO$_2$, NO$_x$, CO, SO$_x$, PM2.5	E = P. LF. T. EF
Maragkogianni et al. [54]	EO	Man, Hot	NO$_x$, SO$_x$, PM	E = P. LF. T. EF
Cullinane et al. [55]	EO	Man, Hot	NO$_x$, CO, SO$_2$, CO$_2$, HC, PM10, PM2.5	E = P. LF. T. EF
Fan et al. [56]	EO	N/a	NO$_x$, CO, SO$_2$, NMVOC, PM10, PM2.5, OC, EC, V, Ni	E = P. LF. T. EF. CF
Papaefthimiou et al. [57]	EO	Man, Hot	NO$_x$, SO$_2$, PM2.5	E = P. LF. T. EF
Chen et al. [58]	EO	Cru, Man, Hot	NO$_x$, CO, SO$_2$, HC, PM10, PM2.5	E = P. LF. T. EF. CF
Styhre et al. [59]	EO	Cru, Dec, Man, Hot	CO$_{2e}$	E = P. LF. T. EF
Alver et al. [7]	EO	Cru, Man, Hot	NO$_x$, SO$_2$, HC, PM10	E = P. LF. T. EF
Knežević et al. [60]	EO	Man, Hot	NO$_x$, SO$_x$, VOC, PM	E = P. LF. T. EF
Khan et al. [61]	FC	Cru, Man, Hot	CO$_2$	E = SFC. LF. T. EF
Sun et al. [62]	EO	Cru, Dec, Man, Hot	NO$_x$, CO, HC, CO$_2$	E = P. LF. T. EF
Cao et al. [63]	EO	Cru, Man, Hot	NO$_x$, CO, SO$_2$, CO$_2$, HC, PM10, PM2.5	E = P. LF. T. EF. CF
Zhang et al. [64]	EO	Cru, Dec + Man, Hot	NO$_x$, CO, SO$_x$, CO$_2$, HC, PM10, PM2.5	E = P. LF. T. EF. CF
Ivce et al. [65]	EO	Cru, Man, Hot	CO$_2$	E = P. LF. T. EF
Wan et al. [66]	EO	Cru, Dec, Man, Hot	SO$_x$, NO$_x$, PM10	E = P. LF. T. EF. CF
Stazić et al. [67]	EO	Cru, Man, Hot	SO$_x$, CO$_2$, VOC, PM	E = P. LF. T. EF
Wan et al. [68]	EO	Cru, Dec, Man, Hot	NO$_x$, CO, SO$_x$, CO$_2$, HC, CH$_4$, NMVOC, PM10, PM2.5	E = P. LF. T. EF. CF
Ekmekçioğlu et al. [69]	EO	Cru, Man, Hot	CO, CO$_2$, NO$_x$, SO$_2$, PM, VOC	E = P. LF. T. EF
Lee et al. [70]	EO	Cru, Man, Hot	CO, NO$_x$, SO$_x$, PM, VOC, NH$_3$	E = P. LF. T. EF

was not considered by Lee et al. [70]. Then, the paper compares the results from the two methods to figure out the gap between them. Hence, it provides a comprehensive annual ship emission inventory by different calculating methods, which covers all ships operated at the port during all months in one calendar year as well as focuses on all key pollutants of CAPSS [77]: CO_2, CO, NO_x, SO_x, PM (including PM10 and PM2.5), VOCs, and ammonia (NH_3);

(2) This paper suggests several available policy implications to reduce the local ship-related emissions and then simulated the effects of them if they will be conducted in the future.

3. Materials Preparation

3.1. Geographical Scope. Incheon metropolitan city is South Korea's third most populous city after Seoul and Busan and also is the gateway of the northwestern area and Seoul capital area, the world's fourth-largest metropolitan area by population, with the Yellow Sea. After becoming an international port in 1883, POI has been the country's logistics hub that handles and distributes both general cargoes and containers. It is ranked as the global 27th busiest port in terms of cargo volume and the 50th biggest container port [78]. Currently, POI has expanded into 125 berths, with a total of 26,031 meters of berth length, distributed to five main component ports as North Port, Inner Port, Coastal Port, South Port, and New Port. The North Port is an industrial port specialized in managing raw materials as scrap iron, feed-by products, hardwood, etc. With a lock-gate, a calm water level in the Inner Port is available for handling semiconductor equipment, automobiles, and precision machine parts. It also is a multipurpose port that serves other cargoes as grain, fruit, and general cargo. The South Port is available for handling small and medium containers and general cargoes. The New Port is under-constructed and specialized for handling medium and large containers. The Coastal Port, with international passenger and ferry terminals, focuses on serving passenger cruise ships and car ferries. Also, around these component ports, POI operates three smaller specialized ports named Geocheom-do Port, Song-do Port, and Yeongheung-do Port. The Song-do Port is specialized in fossil energy products, while the Geocheom-do Port is used for handling sand. The Yeongheung-do Port is constructed to support Yeongheung Thermal Power Plants. The capacity of POI is described in Table 2.

The POI geographical segments are presented in Figure 1. The study not only covered ship operation inside the port boundary but also the "affected zone" (within 5 km from the port boundary) to consider the effect of close-to-port emitted ship emissions, following EPA [80].

3.2. Activity Phase. A typical ship call often contains consecutive activity phases, which consist of a series of continuous activities that have similar features, categorized as "cruising, "maneuvering," "hotelling at anchorage," and "hotelling at berth." Due to the VSR program was applied for a few last days in December 2019, the "decelerating phase" is assumed as not considered in this study. As defined in ICF [30] and by Song [2], at the "cruising" phase, the ship moves inside the port boundary, and all engines keep running. "Maneuvering" is the time that the ship transits between the breakwater (intersection of open sea and inland waterway) and berths at a slower speed. Even with tug assist, the propulsion engines are still in operation. "Hotelling at anchorage" is a period of waiting for berth call while "hotelling at berth" is the time that ships are docking at berth. In both "hotelling at anchorage" and "hotelling at berth" phases, only auxiliary engines (including boilers) still work with a peak load for providing on-board power. In the "hotelling at berth" phase, the ship can use shore power instead of turning on auxiliary engines, called cold ironing; however, cold ironing has been still not implemented in POI yet. The average speed and travel distance in each phase were provided by a local pilot company. Table 3 summarizes detailed activity phases information in POI.

3.3. Ship Classification. Ship characteristics (e.g., speed, engine size, and usage) vary considerably depending on the ship type. Previous studies classified ships in different ways; however, the common standard applied is a type of cargoes carried on the ship. This study followed ICF's classification with 11 types of ships [30]. However, in POI, there are eight types available, involving bulk carrier, container ship, cruise ship, general cargo, reefer, roll-on/roll-off (RORO), tanker, and miscellaneous.

3.4. Data Collection and Analysis

3.4.1. Data Source. Due to the limitations of the AIS system mentioned above, in this study, in-port ship traffic data were collected from the VTS-based Korean Port Management Information System (Port-MIS) to consider ships with GT less than 300. A total of 16,677 ship calls were recorded from Port-MIS, as shown in Table 4. Tanker fleet dominated ship calls at POI in 2019 with 42.6% of the total ship calls, following by the general cargo fleet and container ship fleet with 20.2% and 17.3%, respectively. Reefer fleet only contributed negligibly to total ship calls with 0.02% of total ship calls.

3.4.2. Data Cleaning and Make-Up. A total of 47,626 statuses describing the entire marine traffic at POI during 2019 were collected. These statuses were noted in order of occurrence inside POI, which is different from the ship call's timeline format. Two consecutive events in the same ship call usually are several hours apart; then, a ship call can last from a few hours to a few days. Also, a high level of traffic exaggerated the complexity level of the original data and made it an arduous task to handle nagging details. Therefore, a code, combined from call sign and time points, was applied to resort original data corresponding to the timeline of each ship call.

Waypoints are used to present ship activities in data. A typical ship call contains waypoints as (1) entrance, clearance from the port area; (2) anchoring, collecting anchor; (3) berthing, unberthing; and (4) shift within the port area between berths (if need). A combination of two consecutive waypoints identifies a corresponding activity phase of the ship call. The time information of each waypoint is provided in the YYYY/MM/DD HH: MM format. The time of an activity phase is identified as the subtractions of 2 corresponding waypoints' time. The average moving time and distance of the cruising and maneuvering process for each target port were calculated. In several cases, when only arrival and departure times were reported in the data, the total time of ship spent on the port is the difference between arrival time and departure time. Later, the corresponding average moving time and distance for the target port are applied, and then the hotelling time at the berth of that ship is the subtraction of average cruising and maneuvering time from the total time.

3.4.3. Ship Basic Data. Basic information about the ship, involving name, ship type, engine type, propulsion engine power, weight tonnage, and design speed, was obtained from the Korea Ship Safety Technology Authority and Korean Register. For missing propulsion engine power data, a simple linear approach to modeling the relationship between ship weight tonnage and propulsion engine power is applied by ship types. Linear models are shown in Figure 2. The coefficients of determination (R^2) are relatively high (>0.85), showing the high reliability of prediction.

Besides, installed auxiliary engine power also was not collected completely because there is a lack of information from manufacturers. However, it is impossible to apply a linear approach to estimate auxiliary engine power because there is no connection between installed auxiliary engine power and speed [51]. Also, it is difficult to recognize the using level of auxiliary engines in each activity phase during a ship call in the practice. Therefore, in previous studies, installed auxiliary engine power is commonly estimated by applying the default ratios by ship types with the total propulsion engine power of the ship [51]. This study applied the ratios revised from ICF [30], which is shown in Table 5.

Lastly, the real values of revolutions per minute (RPM) of engines also were obtained from the Korea Ship Safety Technology Authority and Korean Register. RPM value of the engine helps to classify those engine speed designations: high-speed diesel (HSD), medium-speed diesel (MSD), and slow-speed diesel (SSD). Then, based on engine speed designations, proper EF values are matched in the calculation. In the case of missing values, they were made up of average values, as shown in Table 6.

3.4.4. Activity Time. The ship transit time was calculated with the actual speed of the ship and the representative travel distance of each port (from border to berth). The average transit time by the port is shown in Table 7. Among ports in POI, Inner Port shows the largest value because of a lock-gate.

The actual times in hotelling at anchorage and berth were obtained from the difference in time between anchoring, collecting anchor, and berthing, unberthing. To avoid anomalies, the ceilings are applied as 7 days (168 hours) for hotelling at anchorage and 14 days (336 hours) for hotelling at berth. The average docking time by ship type after applying ceilings is shown in Table 8. Almost average values by ship type (except container ship's average time) are considerably greater than the national average time applied in CAPSS (0.79 days~18.96 hours). These differences may lead to considerable variances between estimated results and national estimations, especially tankers, general cargo ships, and container ships dominated the total number of ship calls at POI with the north of 80%.

4. Methodology

To provide a multidimensional overview of in-port ship pollution, this study applied both EO and FC methods to figure out port emissions emitted in POI during 2019 with the activity time data collected from the VTS system. The methodology of the bottom-up approach was developed by improving the general estimation equations from the previous works, which are reviewed in the literature review. The total volume of emissions discharged from a ship call is the aggregation of emissions from the combustion of all engines operated in all activity phases. Then, emission results from calls were scaled up to reach the entire POI emission in 2019. Also, CF is used for adjusting emission factors, and in this study, CF will be assumed to be 1, meaning there are no emission reduction technologies installed on ships.

4.1. EO Method. EO method adheres to the generic equation mentioned in the literature review part. The energy demand here considers the energy output of the engine over the operating time. The equation for the EO method applied in this study considers all activity phases, air pollutants, and engines in a ship call as follows:

$$E_{s,i} = \sum_{\mathrm{ph},j}\left[T_{s,ph}\sum_{j}\left(P_{s,j}\times \mathrm{LF}_{s,j}\times \mathrm{EF}_{i,j,\mathrm{ph}}\right)\right], \quad (7)$$

where E represents the total volume of emission emitted over a complete ship call (g); T represents activity time (hour); P represents engine power (kW); LF represents load factor (%); EF represents emission factor (g/kWh); s represents ship call; i represents pollutant; ph represents the activity phase of a ship call (hotelling at anchorage, cruising, maneuvering, and hotelling at berth); and j represents engine type (propulsion engine and auxiliary engine).

The propeller law is applied to calculate the propulsion engine load factor (LF_m), as shown in the following equation:

$$\mathrm{LF}_m = \left(\frac{\mathrm{AS}}{\mathrm{MS}}\right)^3, \quad (8)$$

where AS = ship actual speed (knots) and MS = ship maximum speed (knots), which is defined by the manufacturer. In the case of the auxiliary engine load factor (LF_a), because

Table 2: Summary of POI's capacity. Source: [79].

No.	Port	Max. Ship DWT	Berths	Handling Capacity Berth length (m)	Bulk (1000RT)	Container (1000TEU)	Main products
1	Inner Port	50,000	46	9,838	38,161		General cargo, iron, grains
2	South Port	100,000	25	3,642	17,600	762	Chemical, cement, sand
3	Coastal Port	50,000	9	1,429			Passenger, oil, LPG
4	North Port	100,000	26	6,421	13,900		Oil, general cargo, wood product
5	New Port	3,000	6	1,600		2,100	Container
6	Song-do	75,000	4	1,300			LPG, oil
7	Yeongheung-do	200,000	5	1,126	14,690		Bituminous coal, limestone
8	Geocheom-do	5,000	4	675	8,320		Sand
Summary		125	26,031	92,671	2,862	28,735.5	

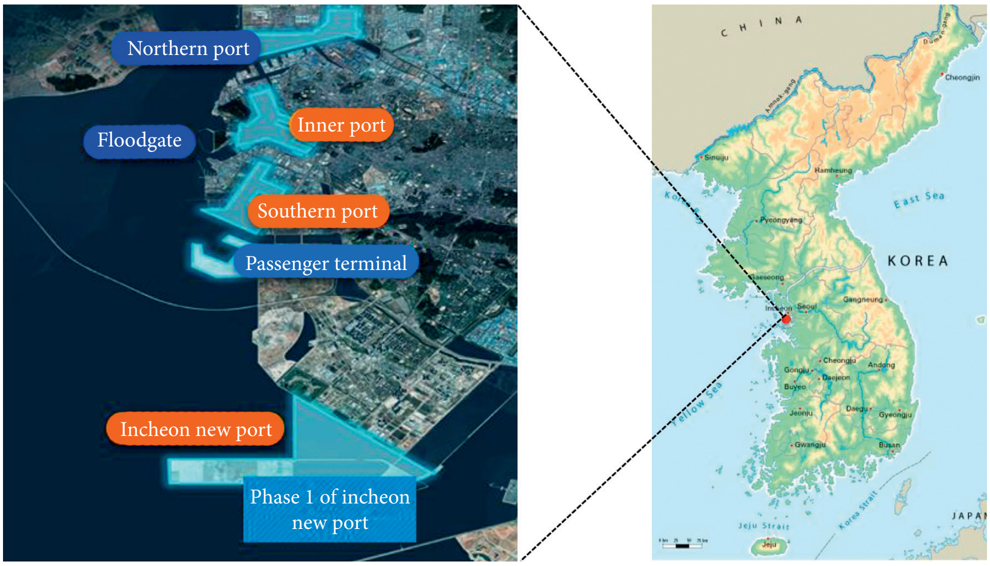


Figure 1: Geographical segments of Port of Incheon. Note: images were downloaded from the Internet and Incheon Port Authority (IPA) website.

Table 3: Ship activity segments and features in POI.

	Phase category	Propulsion engine	Auxiliary engine	Boiler	Avg. speed (knots)	Travel distance (nm)
1	Hotelling at anchorage	Off	On	On	0	0
2	Cruising	On	On	Off	12	Varies by berth
3	Maneuvering	On	On	On	Around 3.5	1
4	Hotelling at berth	Off	On	On	0	0

it varies by ship type and activity phases, and the data sources are limited, therefore, ICF [30] also suggested default LF_a assumptions in his study. This study referred to these values for calculation.

The non-CO_2 EFs used here were referred from Lee et al. [70]. For CO_2 emission, the carbon fraction in the fuel used is defined at 86.8% by weight, and a factor of molecular weight of C and CO_2 equals 44/12. Using SFCs referred from ICF [30], the EFs for CO_2 emission are calculated as

$$CO_2\ \mathrm{EF} = 0.868.\frac{44}{12}.\mathrm{SFC}. \qquad (9)$$

As an international hub port, most ships visited POI are defined as international ocean-going ships; therefore, international EFs estimation methods suggested in them are deemed adaptable for this study. Applied EFs in the EO method are summarized in Table 9.

4.2. FC Method. Similar to the EO method, the FC method's equation adheres to the generic equation and considers fuel consumption of the engine over the operating time to figure out energy demand. The detailed equation for fuel consumption applied in this study is referred from Corbett et al.

Table 4: Ship call statistics (unit: ship call). Source: Port-MIS.

Ship types	North Port	Inner Port	Coastal Port	South Port	New Port	Others	Total
Bulk carrier	253	304	—	104	—	—	661
Container ship	27	13	7	968	1,871	—	2,886
Cruise ship	6	3	6	7	—	1	23
General cargo	1,056	696	149	1,406	47	20	3,374
RoRo	2	777	738	330	—	—	1,847
Reefer	2	—	1	1	—	—	4
Tanker	4,144	473	639	836	165	847	7,104
Miscellaneous	113	61	77	204	29	294	778
Total	5,603	2,327	1,617	3,856	2,112	1,162	16,677

[31]; then, the final equation for the FC method is formed and shown in the following equation:

$$\begin{aligned} E_{s,i,ph} &= \sum_{\mathrm{ph},j} \mathrm{FC}_{\mathrm{ph},j} \times \mathrm{EF}_{j,i,\mathrm{ph}} \\ &= \sum_{\mathrm{ph}} \left\{ T_{\mathrm{ph}} \times \left[\mathrm{MF}_{\mathrm{ph}} \times \left(\frac{\mathrm{AS}}{\mathrm{MS}}\right)^3 \times \mathrm{EF}_{m,i,\mathrm{ph}} + \mathrm{AF}_{\mathrm{ph}} \times \mathrm{EF}_{a,i,\mathrm{ph}} \right] \right\}. \end{aligned} \tag{10}$$

5. Results and Discussion

This section reports a multidimensional bottom approach in-port ship emission inventory in POI during 2019, including two inventories estimated under the two methods mentioned above. These inventories figured out in-port ship emissions in seven geographical segments (at sea, five main ports, and other specialized ports), by eight types of air pollutants and by four ship activity phases. There was considerable variance between results from the two methods. EO method reported 201,612.26 tonnes of CO_2, 323.85 tonnes of CO, 4,097.79 tonnes of NO_x, 1,237.92 tonnes of SO_x, 145.06 tonnes of PM10 (including PM2.5), 133.32 tonnes of PM2.5, 136.21 tonnes of VOCs, and 0.44 tonne of NH_3, while FC method showed 193,981.71 tonnes, 452.88 tonnes, 3,573.27 tonnes, 1,224.01 tonnes, 88.21 tonnes, 82.33 tonnes, 116.09 tonnes, and 0.43 tonne, respectively. The total fuel consumption for ships at POI during 2019 (reported in FC method) was nearly 61,193 tonnes.

In general, the EO method reported higher values than the FC method's value with almost all pollutants, except for CO emissions. Also, there are big gaps (in percentage) between the calculated volumes of NO_x, CO, and PM emissions of the two inventories. However, both inventories mutualized in the rank of emitted pollutants in POI's emission inventory. CO_2 was the dominating air pollutant in both inventories at POI during 2019, accounting for about 97% of the total amount of emitted emissions. Among the rest, NO_x was the most serious air pollutant in both inventories at POI during 2019, which covers over 64% of the total amount of non-CO_2 emissions, followed by SO_x emissions with over 20%. The contribution of CO ranged from 5–8%. NH_3 just accounted for nearly 0.01% of the total amount of non-CO_2 emissions. Other pollutants' shares in both inventories were also insignificant with less than 2.5% of the total non-CO_2 volume. The comparison of EO with FC methods about non-CO_2 emissions is shown in Figure 3.

Figure 4 presents the inventories of ship emission by geographical areas during 2019, considering the sea area (when ships moved on the area of the sea from affected zone to berths) and port areas (when ships docked at berths, involving five main component ports and three others). Both inventories agreed that almost all ship emissions were exhausted on sea area. Also, the big difference between the two inventories happened in sea area (7%), while they showed similar volumes at berths. The emission share on sea area is 53–55% of the total amount of emissions. Among port areas, Inner Port was the most polluted port, accounting for 14% of the ship emissions, followed by North Port and South Port with a ratio of 9% and 7%, respectively. Coastal Port shared the smallest proportion, compared to five main component ports, with almost 4% of the total amount of emission.

The great contribution of Inner Port, North Port, and South Port to the air pollution at POI has been anticipated because they are the most important and busiest ports at POI. With a lock-gate, the Inner Port is the ideal port for handling car ferries; therefore, the RORO fleet (mainly car ferries) supplies mainly for the first rank of Inner Port. The tanker fleet was distributed in all ports of POI; however, the concentration in a large number of tankers in North Port pushed it to become the second polluted port at POI. In the case of South Port, general cargo and container ship fleets are the main emission contributors there.

Next, Figure 5 illustrates the emission contribution by ship types at POI during 2019. Tanker fleet was agreed as the most polluted fleet at POI, with around 29% of total in-port emission volume, followed by general cargo, RORO, and container ship. However, two inventories show a dissimilar view in the ranks of these three groups. While EO inventory reports that second rank belongs to container ship fleet, which contributed up to 20.1% of emission volume, followed closely by RORO fleet and general cargo fleet with the ratio of 19.8% and 19.6%, respectively; in case of FC inventory, the RORO fleet raised to occupy the second position with 20.3%, pushed general cargo fleet down to the third rank with 19.8%, and container ship fleet only contributed 19.7% of the total amount of emissions, ranked fourth. The top-4 polluted fleets supplied 88.7% of the total emission volume. In both inventories, the bulk carrier fleet also was a considerable polluted source, which contributed over 8.8% of the total

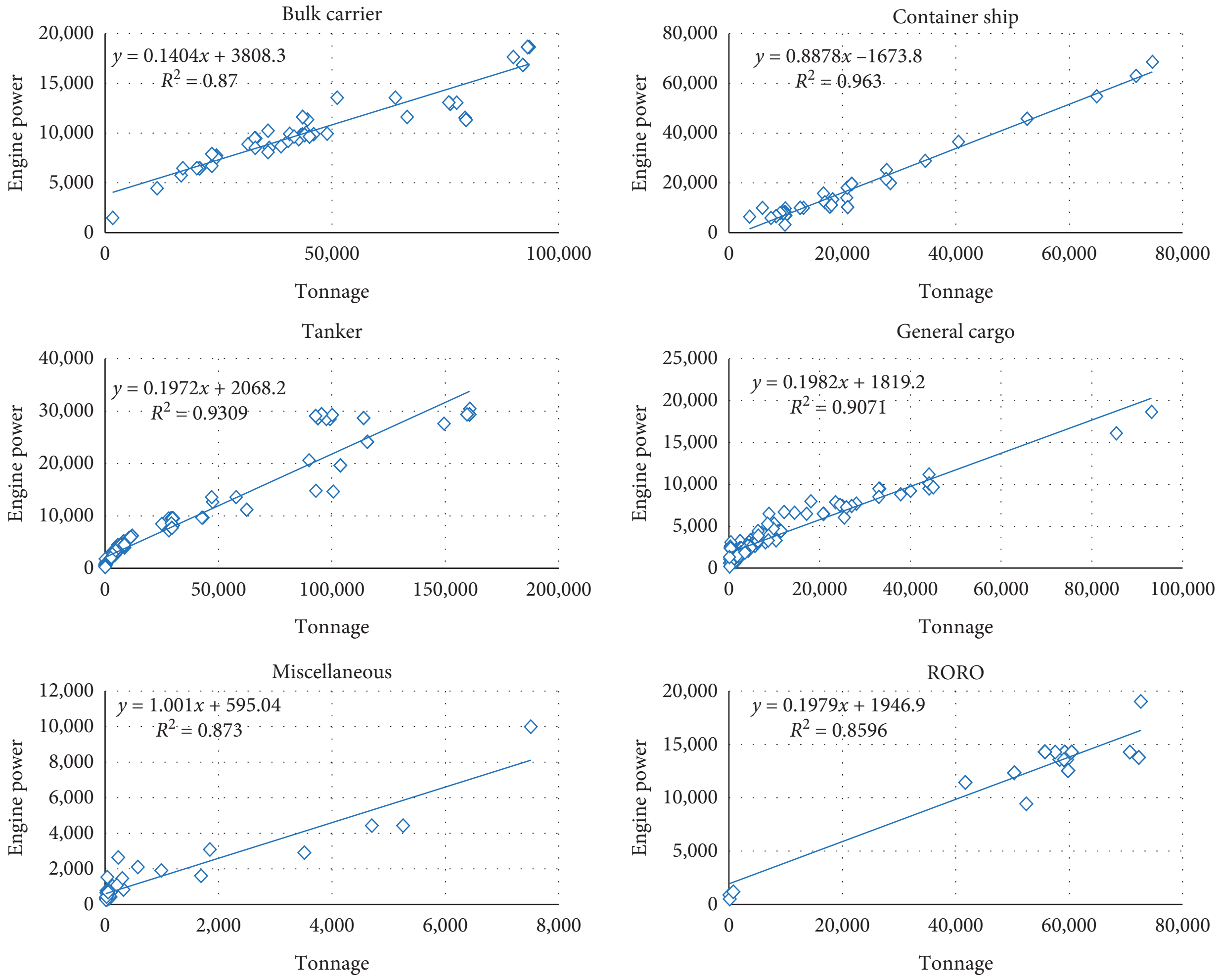


Figure 2: Linear regression of missing main engine data by ship type.

Table 5: The ratio for estimating auxiliary engine power. Source: ICF [30].

Ship type	Auxiliary to propulsion ratio
Bulk carrier	0.222
Container ship	0.220
Cruise ship	0.278
General cargo	0.191
RORO	0.259
Reefer	0.406
Tanker	0.211
Others	0.100

amount of emission, ranked fifth. Other fleets present minor contributions, fewer than 2%. Therefore, IPA should focus on emission reduction from tanker, general cargo, RORO, and container ship.

Then, Figure 6 shows ship emission inventories by activity phases. Although the rankings and contribution (in percentage) of cruising and hotelling at berth are interchanged in two inventories, however, it is undeniable that they overshadowed others in total emitted emissions. During hotelling at berth, the propulsion engine is in respite, and then auxiliary engines monopolize in emitting all emissions. Therefore, besides the implementation of ECA, to reduce the volume of in-port ship emissions, efforts to cut down emissions during cruising and hotelling at berth phases are indispensable.

6. Ship-Related Emissions Reduction Policies at POI

As discussed above, several policies are designed to reduce the volume of emissions emitted from ships operated at POI as (1) local ECA realization, (2) VSR program motivation, (3) application of cold ironing, and (4) establishment of a national integrated emission platform. In this study, the above-calculated EO emission inventory is adopted as the baseline emission inventory and solutions (1)–(3) are simulated and evaluated based on EO method if assuming that these suggested policies are applied at POI. The potential benefit of solution (4) also is discussed.

6.1. Assessing the Environmental Benefit of Designating the Local ECA. The great contribution of NO_x and SO_x emissions among non-CO_2 emissions to air pollution at POI suggests the necessity of designating a local NO_x and SO_x

Table 6: Average RPM values.

Ship type	Average RPM values of propulsion engines	Average RPM values of auxiliary engines
Bulk carrier	103	843
Container ship	131	847
Cruise ship	484	895
General cargo	271	1,225
RoRo	105	787
Reefer	311	1,043
Tanker	286	1,147
Miscellaneous	490	1,037

Table 7: Average transit time in POI 2019 (unit: hour).

Port	Average transit time
North Port	2.2
Inner Port	3.0
Coastal Port	1.6
South Port	1.7
New Port	1.4
Song-do Port	1.4
Geocheom-do Port	2.5
Yeongheung-do Port	0.75

emission control area (ECA). In ECA, ships will be required to follow strictly Tier III in Regulation 13 of Annex VI of MARPOL 73/78 about reducing NO_x emissions and the sulphur contents in the fuel used will be reduced to 0.1%. This idea was suggested firstly by Chang et al. [49]; however, it has not been implemented yet in Incheon or any other Korean ports. As of 2019, the required fuel used inside the POI area was MGO with the maximum sulphur content at 1.0% by mass. Then, with SO_x ECA, ships have to switch from high sulphur content fuel (1.0%) to a very lower one (0.1%). Assuming that energy efficiency is unchanged, since the volumes of SO_x and PM emissions released are dependent critically on the sulphur content in the fuel used according to estimating equations applied in Lee et al. [70], the SO_x ECA could help decrease 10 times the amount of SO_x emission and nearly 2.5 times the amount of PM emission. The cutdown volume of emissions will be 1114.12 tonnes of SO_x and 71.07 tonnes of PM10 (including 65.25 tonnes of PM2.5) if the SO_x ECA is established.

Different from SO_x emission, in the case of NO_x emission, the abatement is implemented mostly by technology improvement in the engine's combustion and selective catalytic reduction (SCR) system equipped onboard. Due to the limitation of the estimating method, it is difficult to simulate and estimate the expected reduction in NO_x emission; therefore, the reduction of NO_x is not considered in this study.

6.2. Assessing the Environmental Benefit of the VSR Program. In the last few days of December 2019, the Korean Ministry of Oceans and Fisheries started designating the "Vessel Speed Reduction Program" in five Korean main ports, including POI, to lessen PM levels during winter. Ships are motivated to transit slower than certain speeds (12 knots for container ships and car-carriers and 10 knots for others). To facilitate the analysis, it is assumed that there was no RSZ in POI during 2019 when calculating the baseline emission inventory and then simulating the effect of RSZ if it is designated. With lower cruising speed (10 knots), the ships will take more cruising time but lower load factor for propulsion engines (see equation (6)). Then, an RSZ can help reduce 7-8% amount of each pollutant: 22.77 tonnes of CO, 295.83 tonnes of NO_x, 87.03 tonnes of SO_x, 10.45 tonnes of PM10 (including 9.61 tonnes of PM2.5), 11.15 tonnes of VOC, 0.03 tonne of NH_3, and 14,167.93 tonnes of CO_2.

6.3. Assessing the Environmental Benefit of Applying Cold Ironing. In the case of hotelling at berth, emission could be cut down through reducing docking time or cut-downing emissions from docking time. In practice, reducing docking time is complicating and challenging because it is related to overhauling, optimizing, and scheduling the operation of a chain including other port facilities such as quay cranes, internal trucks, and yard cranes. In contrast, reducing emissions from docking time could be achieved easily by using an on-shore power supply. Assuming that 50% of hotelling time in each port is applied this technology, it will help reduce 22-23% amount of each pollutant: 72.26 tonnes of CO, 913.14 tonnes of NO_x, 278.54 tonnes of SO_x, 32.19 tonnes of PM10 (including 29.56 tonnes of PM2.5), 27.59 tonnes of VOC, 0.1 tonne of NH_3, and 45,375.03 tonnes of CO_2.

If three of them are conducted together, this will help cut down 95.03 tonnes of CO, 1208.97 tonnes of NO_x, 1150.68 tonnes of SO_x, 92.51 tonnes of PM10 (including 84.97 tonnes of PM2.5), 38.74 tonnes of VOC, 0.13 tonne of NH_3, and 59,542.96 tonnes of CO_2 corresponding to 29%, 30%, 93%, 64% (64%), 28%. 30% and 30% of the total emitted amount of each pollutant in the 2019 baseline inventory.

6.4. Establishment of a National Integrated Emission Platform. Besides counter-measures, a good assessment and management system is indispensable for managing green reform at ports. With the widespread Internet, the combination of the platform industry and assessments of environmental situations around seaport would be an interesting and promising research topic shortly. With the platform, users can access and interact with the system actively and easily. This study suggests an idea about a national integrated platform that can standardize and systematize procedures in

Table 8: Average time in hotelling at anchorage and berth in POI 2019 (unit: hour).

Types of ship	Average time in hotelling (hour)	Compare to the national avg. time in CAPSS (%)
Bulk carrier	83.3	439.3
Container ship	13.6	71.7
Cruise Ship	50.3	265.3
General cargo	39.9	210.4
RORO	25.9	136.6
Reefer	89.0	469.4
Tanker	22.0	116.0
Miscellaneous	51.9	273.7

Table 9: Applied EFs for EO method (unit: g/kWh). Source: Lee et al. [70].

Engine type	Phase	CO	CO_2	NO_x	SO_x	PM10	PM2.5	VOCs	NH_3	BSFC
Propulsion-HSD	Cruising	1.1	646.14	12.0	3.97	0.47	0.43	0.21	0.0014	203
Propulsion-MSD	Cruising	1.1	646.14	13.2	3.97	0.47	0.43	0.63	0.0014	203
Propulsion-SSD	Cruising	0.5	588.85	17.0	3.62	0.45	0.42	0.53	0.0013	185
Propulsion-HSD	Maneuvering	2.2	709.8	9.6	4.36	0.50	0.46	0.63	0.0016	223
Propulsion-MSD	Maneuvering	2.2	709.8	10.6	4.36	0.50	0.46	1.58	0.0016	223
Propulsion-SSD	Maneuvering	1.0	649.32	13.6	3.99	0.47	0.44	1.90	0.0014	204
Auxiliary	All	1.1	690.71	13.9	4.24	0.49	0.45	0.42	0.0015	217

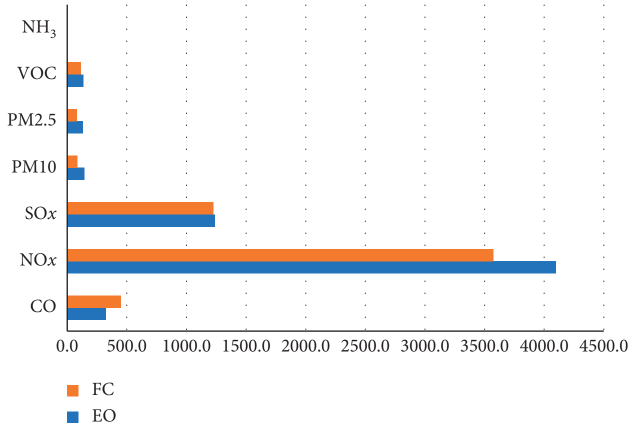


Figure 3: Non-CO_2 emissions comparison between FC and EO methods (unit: tonnes).

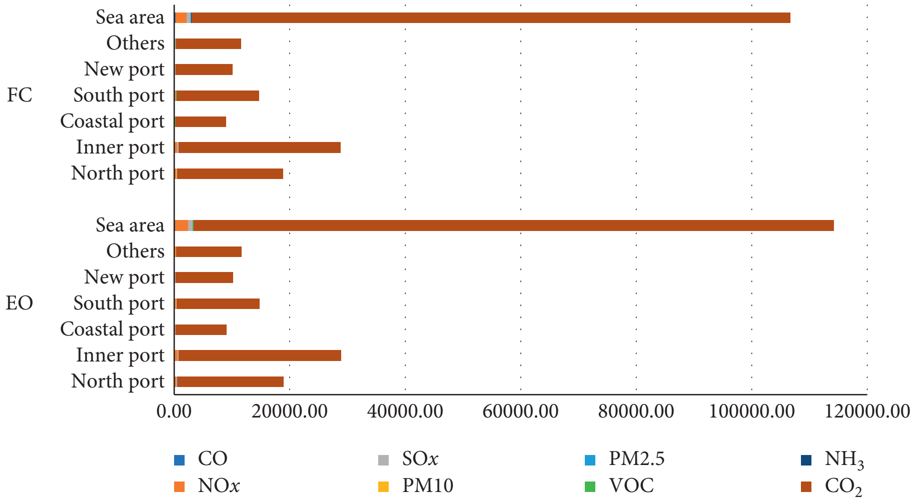


Figure 4: POI 2019 ship emission inventories by geographical areas (Unit: tonnes).

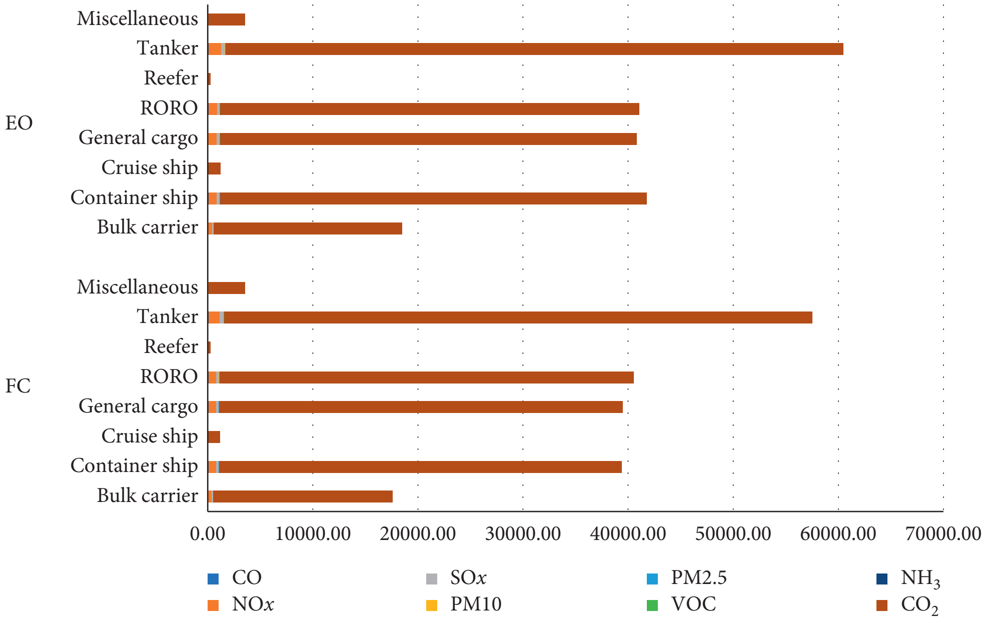


Figure 5: POI 2019 ship emission inventory by ship type (unit: tonnes).

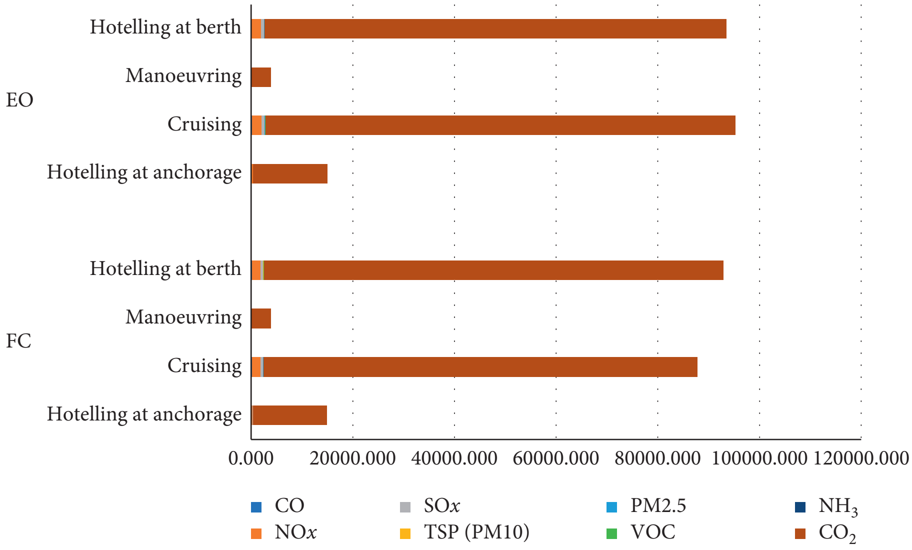


Figure 6: POI 2019 ship inventory by ship activity phase (unit: tonnes).

monitoring harmful ship-related emissions released from ports in Korea. This platform will provide three main functions: (1) data collection and analysis; (2) estimation, modification, and visualization; and (3) prediction. The abovementioned three functions are proposed to ensure and emphasize the systematicness to generate a better air pollutant assessment and management.

The necessary input data for emission estimation is linked with and automatic-synchronized simultaneously from the national Port-MIS system. Then, based on collected data, the system analyses, estimates, and visualizes the volume of emissions emitted or environment indexes promptly to warn the port operator if the negative impacts on the local community from in-port emissions exceed a certain "safe" level. Then, input data adjustment in the modification function allows port operators to adjust or reschedule the port operation, through suppositions and corresponding input adjustment, to reduce geographic emissions to the "safe" level. Also, by machine-learning algorithms and statistical analyses, the prediction function helps port operators predict the trend of air pollution at ports and then plan port operations appropriately. Also, the researchers use modification and prediction function in annual or seasonal emission inventory with their recommended air management policies or counter-measures to assess the performances of them.

7. Conclusion

As a result of the quick spreading of global trade, atmospheric pollution from transportation has been more topical than ever, especially in dense hub port-cities. Given environmental legislation that is enforced currently, the shipping industry has not been paid adequate attention to its relatively considerable volume of emitted emissions. Besides published national reports, the awareness of atmospheric pollution from growing port traffic has invoked reliable and up-to-date regional comprehensive assessments of egregious environmental impacts from in-port ship traffic with timely counter-measures. This study contributes to the stream of in-port ship emission research both theoretically and practically. Besides the sole application of AIS for inputs in the bottom-up approach studies, VTS could be considered as a promising confident alternative. In Korea, with support from the government, the VTS-based Port-MIS system would be a reliable data source for local studies.

This study generated a comprehensive annual ship emission inventory at POI in 2019, according to geographical areas, operational phases, and ship types by the bottom-up approach. Although there are small differences between the results of the two methods EO and FC, and it is impossible to demonstrate which one is better, however, two inventories agree about the dominance of CO_2 emission and the considerable volume of NO_x and SO_x emissions in the total volume of port emissions, the significant contribution of auxiliary engines during hotelling at berth phase and the 4-most polluted ship types at POI during the year of 2019. Also, this study suggests and simulates the effects of four applicable green policies for the POI situation in the practice. The combination of the three abovementioned policies could help slash the significant volume of emitted CO (29%), NO_x (30%), SO_x (93%), PM10 and PM2.5 (64%), VOC (28%), NH_3 (30%), and CO_2 (30%).

Given most ships visited POI could be considered as international waterborne navigation, using default inputs in the estimation process (e.g., EF, auxiliary characteristics data, and engine load factor), that referred from international studies, is deemed acceptable. Moreover, boilers also were not collected completely and considered because there is a lack of information from manufacturers and ship owners. Therefore, for better estimation, the authors also recommend the adoption of local values to reduce uncertainty. However, currently, Korean data are limited; therefore, it garners the attention of investigating local-specific inputs. Moreover, an idea of the automatic input collecting system be synchronized with our proposed integrated system could be another interesting topic for further research.

Although both inventories also reported an insignificant volume of PM emissions, in terms of weight contribution, however, if compared their size and weight to other pollutants, they would be a potential threat to the port environment as well as the local community. Therefore, besides figuring out the weight of the total volume of PM emitted, the evaluation of spatial seasonal PM dispersion in the port area would be extremely necessary. This evaluation will contribute considerably to ship emission inventory and the deep understanding of the affection of in-port ship PM emissions.

Finally, the port operation also contains other important activities with various types of cargo handling equipment, railcars, and drayage trucks, and they contribute significantly to port pollution. Therefore, a broader view of a comprehensive in-port emission inventory or a port-scale integrated system, which presents emissions not only from ships but also from other port-related land-based vehicles, should be considered soon to understand completely about port pollution. Of course, this complete inventory would be a great baseline for other in-port environment evaluating studies in Incheon as well as other similar regions.

Data Availability

Data are available from Port-MIS (https://new.portmis.go.kr/portmis) operated by Ministry of Oceans and Fisheries.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by an Incheon National University Research Grant in 2020.

References

[1] Y. Kwon, H. Lim, Y. Lim, and H. Lee, “Implication of activity-based vessel emission to improve regional air inventory in a port area,” *Atmospheric Environment*, vol. 203, pp. 262–270, 2019.

[2] S. Song, “Ship emissions inventory, social cost and eco-efficiency in Shanghai Yangshan port,” *Atmospheric Environment*, vol. 82, pp. 288–297, 2014.

[3] J. H. J. Hulskotte and H. A. C. Denier van der Gon, “Fuel consumption and associated emissions from seagoing ships at berth derived from an on-board survey,” *Atmospheric Environment*, vol. 44, no. 9, pp. 1229–1236, 2010.

[4] K. Cullinane and S. Cullinane, “Atmospheric emissions from shipping: the need for regulation and approaches to compliance,” *Transport Reviews*, vol. 33, no. 4, pp. 377–401, 2013.

[5] M. C. Sanabra, J. J. Usabiaga Santamaría, and F. X. Martínez De Osés, “Manoeuvring and hotelling external costs: enough for alternative energy sources?” *Maritime Policy & Management*, vol. 41, no. 1, pp. 42–60, 2014.

[6] O. Merk, *Shipping Emissions in Ports (Discussion Paper No. 2014-20)*, International Transport Forum, Paris, France, 2014.

[7] F. Alver, B. A. Saraç, and Ü. Alver Şahin, “Estimating of shipping emissions in the samsun port from 2010 to 2015,” *Atmospheric Pollution Research*, vol. 9, no. 5, pp. 822–828, 2018.

[8] J. S. L. Lam and T. Notteboom, “The greening of ports: a comparison of port management tools used by leading ports in Asia and Europe,” *Transport Reviews*, vol. 34, no. 2, pp. 169–189, 2014.

[9] L. Bilgili and U. B. Celebi, “Emission routing in maritime transportation,” in *Energy, Transportation and Global Warming*, pp. 837–849, Springer, Cham, Switzerland, 2016.

[10] J. S. Carlton, S. D. Danton, R. W. Gawen et al., *Marine Exhaust Emissions Research Programme*, Vol. 63, Lloyd's Register Engineering Services, London, UK, 1995.

[11] V. Eyring, H. W. Köhler, J. Van Aardenne, and A. Lauer, "Emissions from international shipping: 2. Impact of future technologies on scenarios until 2050," *Journal of Geophysical Research*, vol. 110, no. D17, 2005.

[12] European Environment Agency (EEA), *EMEP/EEA Air Pollutant Emission Inventory Guidebook 2019 (Report No. 13/2019)*, Publications Office of the European Union, Luxembourg, UK, 2019.

[13] Natural Resources Defense Council, *Harboring Pollution Strategies To CleanUp U.S. Ports*, Natural Resources Defense Council, NewYork, NY, USA, 2004.

[14] J. J. Corbett, J. J. Winebrake, E. H. Green, P. Kasibhatla, V. Eyring, and A. Lauer, "Mortality from ship emissions: a global assessment," *Environmental Science & Technology*, vol. 41, no. 24, pp. 8512–8518, 2007.

[15] International Maritime Organization, "Sulphur 2020–cutting Sulphur oxide emissions," 2020, http://www.imo.org/en/MediaCentre/HotTopics/Pages/Sulphur-2020.aspx.

[16] European Union, "Directive (EU) 2016/802 of the European parliament and of the council of 11 May 2016 relating to a reduction in the sulphur content of certain liquid fuels," 2020, https://eur-lex.europa.eu/eli/dir/2016/802/oj.

[17] Environmental Protection Agency, "Control of emissions of air pollution from nonroad diesel engines and fuel; final rule," *Federal Register*, vol. 69, no. 124, pp. 38958–39273, 2004.

[18] "North China: emission control areas (update)," 2020, https://www.nepia.com/industry-news/china-emission-control-areas-update/.

[19] European Environment Agency, *Air Quality in Europe—2019 Report (Report No 10/2019)*, Publications Office of the European Union, Luxembourg, UK, 2019.

[20] International Institute for Applied Systems Analysis, "The potential for cost-effective air emission reductions from international shipping through designation of further emission control areas in EU waters with focus on the Mediterranean Sea," 2020, http://www.iiasa.ac.at/web/home/research/researchPrograms/air/Shipping_emissions_reductions_main.pdf.

[21] Korea Ministry of Environment, *Comprehensive Plan on Fine Dust Management*, Korea Environmental Policy Bulletin, vol. 15, no. 2, Republic of Korea, 2018.

[22] Korean Ministry of Oceans and Fisheries, "Vessel speed reduction (VSR) program to start December this year," 2020, https://www.mof.go.kr/en/board.do?menuIdx=1491&bbsIdx=30629.

[23] S.-K. Song and Z.-H. Shon, "Current and future emission estimates of exhaust gases and particles from shipping at the largest port in Korea," *Environmental Science and Pollution Research*, vol. 21, no. 10, pp. 6612–6622, 2014.

[24] A. Glowacz, "Fault diagnosis of single-phase induction motor based on acoustic signals," *Mechanical Systems and Signal Processing*, vol. 117, pp. 65–80, 2019.

[25] L. Huang, Y. Wen, X. Geng, C. Zhou, C. Xiao, and F. Zhang, "Estimation and spatio-temporal analysis of ship exhaust emission in a port area," *Ocean Engineering*, vol. 140, pp. 401–411, 2017.

[26] R. A. O. Nunes, M. C. M. Alvim-Ferraz, F. G. Martins, and S. I. V. Sousa, "The activity-based methodology to assess ship emissions - a review," *Environmental Pollution*, vol. 231, pp. 87–103, 2017.

[27] H. N. Psaraftis and C. A. Kontovas, "CO_2 emission statistics for the world commercial fleet," *WMU Journal of Maritime Affairs*, vol. 8, no. 1, pp. 1–25, 2009.

[28] J. J. Corbett and H. W. Koehler, "Updated emissions from ocean shipping," *Journal of Geophysical Research: Atmospheres*, vol. 108, no. D20, 2003.

[29] A. Miola and B. Ciuffo, "Estimating air emissions from ships: meta-analysis of modelling approaches and available data sources," *Atmospheric Environment*, vol. 45, no. 13, pp. 2242–2251, 2011.

[30] ICF International, *Current Methodologies in Preparing Mobile Source Port-Related Emission Inventories: Final Report*, US Environmental Protection Agency (USEPA), Fairfax, VA, USA, 2009.

[31] J. J. Corbett, H. Wang, and J. J. Winebrake, "The effectiveness and costs of speed reductions on emissions from international shipping," *Transportation Research Part D: Transport and Environment*, vol. 14, no. 8, pp. 593–598, 2009.

[32] J. Joseph, R. S. Patil, and S. K. Gupta, "Estimation of air pollutant emission loads from construction and operational activities of a port and harbour in Mumbai, India," *Environmental Monitoring and Assessment*, vol. 159, no. 1–4, p. 85, 2009.

[33] C. Deniz and A. Kilic, "Estimation and assessment of shipping emissions in the region of Ambarlı Port, Turkey," *Environmental Progress & Sustainable Energy*, 2009.

[34] C. Deniz, A. Kilic, and G. Cıvkaroglu, "Estimation of shipping emissions in Candarli gulf, Turkey," *Environmental Monitoring and Assessment*, vol. 171, no. 1–4, pp. 219–228, 2010.

[35] O. J. A. Howitt, V. G. N. Revol, I. J. Smith, and C. J. Rodger, "Carbon emissions from international cruise ship passengers' travel to and from New Zealand," *Energy Policy*, vol. 38, no. 5, pp. 2552–2560, 2010.

[36] A. Kiliç and C. Deniz, "Inventory of shipping emissions in izmit gulf, Turkey," *Environmental Progress & Sustainable Energy*, vol. 29, no. 2, pp. 221–232, 2010.

[37] G. Lonati, S. Cernuschi, and S. Sidi, "Air quality impact assessment of at-berth ship emissions: case-study for the project of a new freight port," *Science of the Total Environment*, vol. 409, no. 1, pp. 192–200, 2010.

[38] E. Tzannatos, "Ship emissions and their externalities for the port of Piraeus-Greece," *Atmospheric Environment*, vol. 44, no. 3, pp. 400–407, 2010.

[39] K. Shin and J.-P. Cheong, "Estimating transportation-related greenhouse gas emissions in the Port of Busan, S. Korea," *Asian Journal of Atmospheric Environment*, vol. 5, no. 1, pp. 41–46, 2011.

[40] G. Villalba and E. D. Gemechu, "Estimating GHG emissions of marine ports-the case of Barcelona," *Energy Policy*, vol. 39, no. 3, pp. 1363–1368, 2011.

[41] C.-C. Chang and C.-M. Wang, "Evaluating the effects of green port policy: case study of Kaohsiung harbor in Taiwan," *Transportation Research Part D: Transport and Environment*, vol. 17, no. 3, pp. 185–189, 2012.

[42] J. Berechman and P.-H. Tseng, "Estimating the environmental costs of port related emissions: the case of Kaohsiung," *Transportation Research Part D: Transport and Environment*, vol. 17, no. 1, pp. 35–38, 2012.

[43] P. S. Yau, S. C. Lee, J. J. Corbett, C. Wang, Y. Cheng, and K. F. Ho, "Estimation of exhaust emission from ocean-going vessels in Hong Kong," *Science of The Total Environment*, vol. 431, pp. 299–306, 2012.

[44] Y.-T. Chang, Y. Song, and Y. Roh, "Assessing greenhouse gas emissions from port vessel operations at the Port of Incheon," *Transportation Research Part D: Transport and Environment*, vol. 25, pp. 1–4, 2013.

[45] D. P. McArthur and L. Osland, “Ships in a city harbour: an economic valuation of atmospheric emissions,” *Transportation Research Part D: Transport and Environment*, vol. 21, pp. 47–52, 2013.

[46] S. K. W. Ng, C. Loh, C. Lin et al., “Policy change driven by an AIS-assisted marine emission inventory in Hong Kong and the Pearl river delta,” *Atmospheric Environment*, vol. 76, pp. 102–112, 2013.

[47] H. Saraçoğlu, C. Deniz, and A. Kılıç, “An investigation on the effects of ship sourced emissions in Izmir Port, Turkey,” *The Scientific World Journal*, vol. 2013, Article ID 218324, 8 pages, 2013.

[48] H.-H. Tai and D.-Y. Lin, “Comparing the unit emissions of daily frequency and slow steaming strategies on trunk route deployment in international container shipping,” *Transportation Research Part D: Transport and Environment*, vol. 21, pp. 26–31, 2013.

[49] Y.-T. Chang, Y. Roh, and H. Park, “Assessing noxious gases of vessel operations in a potential emission control area,” *Transportation Research Part D: Transport and Environment*, vol. 28, pp. 91–97, 2014.

[50] T.-K. Liu, H.-Y. Sheu, and J.-Y. Tsai, “Sulfur dioxide emission estimates from merchant vessels in a port area and related control strategies,” *Aerosol and Air Quality Research*, vol. 14, no. 1, pp. 413–421, 2014.

[51] L. Goldsworthy and B. Goldsworthy, “Modelling of ship engine exhaust emissions in ports and extensive coastal waters based on terrestrial AIS data—an Australian case study,” *Environmental Modelling & Software*, vol. 63, pp. 45–60, 2015.

[52] J. Coello, I. Williams, D. A. Hudson, and S. Kemp, “An AIS-based approach to calculate atmospheric emissions from the UK fishing fleet,” *Atmospheric Environment*, vol. 114, pp. 1–7, 2015.

[53] M. Tichavska and B. Tovar, “Port-city exhaust emission model: an application to cruise and ferry operations in Las Palmas Port,” *Transportation Research Part A: Policy and Practice*, vol. 78, pp. 347–360, 2015.

[54] A. Maragkogianni, S. Papaefthimiou, and C. Zopounidis, *Mitigating Shipping Emissions in European Ports: Social and Environmental Benefits*, Springer International Publishing, Berlin, Germany, 2016.

[55] K. Cullinane, P.-H. Tseng, and G. Wilmsmeier, “Estimation of container ship emissions at berth in Taiwan,” *International Journal of Sustainable Transportation*, vol. 10, no. 5, pp. 466–474, 2016.

[56] Q. Fan, Y. Zhang, W. Ma et al., “Spatial and seasonal dynamics of ship emissions over the Yangtze river delta and East China Sea and their potential environmental influence,” *Environmental Science & Technology*, vol. 50, no. 3, pp. 1322–1329, 2016.

[57] S. Papaefthimiou, A. Maragkogianni, and K. Andriosopoulos, “Evaluation of cruise ships emissions in the Mediterranean basin: the case of Greek ports,” *International Journal of Sustainable Transportation*, vol. 10, no. 10, pp. 985–994, 2016.

[58] D. Chen, X. Wang, P. Nelson et al., “Ship emission inventory and its impact on the PM2.5 air pollution in Qingdao Port, North China,” *Atmospheric Environment*, vol. 166, pp. 351–361, 2017.

[59] L. Styhre, H. Winnes, J. Black, J. Lee, and H. Le-Griffin, “Greenhouse gas emissions from ships in ports—case studies in four continents,” *Transportation Research Part D: Transport and Environment*, vol. 54, pp. 212–224, 2017.

[60] V. Knežević, R. Radonja, and Č. Dundović, “Emission inventory of marine traffic for the port of Zadar,” *Pomorstvo*, vol. 32, no. 2, pp. 239–244, 2018.

[61] S. Khan, Y.-T. Chang, S. Lee, and K.-S. Choi, “Assessment of greenhouse gas emissions from ships operation at the Port of Incheon using AIS,” *Journal of Korea Port Economic Association*, vol. 34, no. 1, pp. 65–79, 2018.

[62] X. Sun, Z. Tian, R. Malekian, and Z. Li, “Estimation of vessel emissions inventory in Qingdao port based on big data analysis,” *Symmetry*, vol. 10, no. 10, p. 452, 2018.

[63] Y.-L. Cao, X. Wang, C.-Q. Yin et al., “Inland vessels emission inventory and the emission characteristics of the beijing-hangzhou grand canal in Jiangsu province,” *Process Safety and Environmental Protection*, vol. 113, pp. 498–506, 2018.

[64] Y. Zhang, J. C. H. Fung, J. W. M. Chan, and A. K. H. Lau, “The significance of incorporating unidentified vessels into AIS-based ship emission inventory,” *Atmospheric Environment*, vol. 203, pp. 102–113, 2019.

[65] R. Ivce, A. Zekic, R. Radonja, and B. Reljac, “Emission inventory of ships calling at the port of Bršica (bay of Raša),” in *Proceedings of the 2019 International Symposium ELMAR*, Zadar, Croatia, September 2019.

[66] Z. Wan, Q. Zhang, Z. Xu, J. Chen, and Q. Wang, “Impact of emission control areas on atmospheric pollutant emissions from major ocean-going ships entering the Shanghai Port, China,” *Marine Pollution Bulletin*, vol. 142, pp. 525–532, 2019.

[67] L. Stazić, R. Radonja, V. Pelić, and B. Lalić, “The port of Split international marine traffic emissions inventory,” *Pomorstvo*, vol. 34, no. 1, pp. 32–39, 2020.

[68] Z. Wan, S. Ji, Y. Liu, Q. Zhang, J. Chen, and Q. Wang, “Shipping emission inventories in China’s Bohai bay, Yangtze river delta, and pearl river delta in 2018,” *Marine Pollution Bulletin*, vol. 151, Article ID 110882, 2020.

[69] A. Ekmekçioğlu, S. L. Kuzu, K. Ünlügençoğlu, and U. B. Çelebi, “Assessment of shipping emission factors through monitoring and modelling studies,” *Science of the Total Environment*, vol. 743, Article ID 140742, 2020.

[70] H. Lee, D. Park, S. Choo, and H. T. Pham, “Estimation of the non-greenhouse gas emissions inventory from ships in the port of Incheon,” *Sustainability*, vol. 12, no. 19, p. 8231, 2020.

[71] D. McCollum and C. Yang, “Achieving deep reductions in US transport greenhouse gas emissions: scenario analysis and policy implications,” *Energy Policy*, vol. 37, no. 12, pp. 5580–5596, 2009.

[72] M. Tichavska and B. Tovar, “External costs from vessel emissions at port: a review of the methodological and empirical state of the art,” *Transport Reviews*, vol. 37, no. 3, pp. 383–402, 2017.

[73] V. Eyring, I. S. A. Isaksen, T. Berntsen et al., “Transport impacts on atmosphere and climate: shipping,” *Atmospheric Environment*, vol. 44, no. 37, pp. 4735–4771, 2010.

[74] L. Johansson, *Emission Estimation of Marine Traffic Using Vessel Characteristics and AIS-Data*, Aalto University, Espoo, Finland, 2011.

[75] National Institute of Environmental Research, *National Air Pollutants Emission Service (in Korea)*, National Institute of Environmental Research, Incheon, South Korea, 2016, https://airemiss.nier.go.kr/mbshome/mbs/airemiss/index.do.

[76] National Institute of Environmental Research, *National Air Pollutant Emissions Handbook*, National Institute of Environmental Research, Incheon, South Korea, 2020, https://airemiss.nier.go.kr/user/boardList.do?command=view&page=1&boardId=85&boardSeq=127&id=airemiss_040200000000in Korea.

[77] National Institute of Environmental Research, *Air Pollutants*, National Institute of Environmental Research, Incheon, South Korea, 2020, http://airemiss.nier.go.kr/mbshome/mbs/airemiss/subview.do?id=airemiss_020200000000In Korea.

[78] American Association of Port Authorities, *World Port Rankings 2016*, American Association of Port Authorities, New York, NY, USA, 2020, https://www.aapa-ports.org/unifying/content.aspx?ItemNumber=21048.

[79] Incheon Port Authority, "IPA introduction," https://www.icpa.or.kr/content/view.do?menuKey=114&contentKey=44.

[80] Environmental Protection Agency, *National Port Strategy Assessment: Reducing Air Pollution and Greenhouse Gases at U.S. Ports (EPA-420-R-16-011)*, Office of Transportation Air Quality, Washington, DC, USA, 2016.

Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 5530814, 14 pages
https://doi.org/10.1155/2021/5530814

Research Article

Exploring Route Choice Behaviours Accommodating Stochastic Choice Set Generations

Shin-Hyung Cho[1] and Seung-Young Kho[2]

[1]*School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA 30324, USA*
[2]*Department of Civil and Environmental Engineering and Institute of Construction and Environmental Engineering, Seoul National University, Seoul 08826, Republic of Korea*

Correspondence should be addressed to Seung-Young Kho; sykho@snu.ac.kr

Received 14 January 2021; Revised 27 February 2021; Accepted 5 March 2021; Published 9 April 2021

Academic Editor: Young-Jae Lee

Modelling route choice behaviours are essential in traffic operation and transportation planning. Many studies have focused on route choice behaviour using the stochastic model, and they have tried to construct the heterogeneous route choice model with various types of data. This study aims to develop the route choice model incorporating travellers' heterogeneity according to the stochastic route choice set. The model is evaluated from the empirical travel data based on a radio frequency identification device (RFID) called dedicated short-range communication (DSRC). The reliability level is defined to explore the travellers' heterogeneity in the choice set generation model. The heterogeneous K-reliable shortest path- (HKαRSP-) based route choice model is established to incorporate travellers' heterogeneity in route choice behaviour. The model parameters are estimated for the mixed path-size correction logit (MPSCL) model, considering the overlapping paths and the heterogeneous behaviour in the route choice model. The different behaviours concerning the chosen routes are analysed to interpret the route choice behaviour from revealed preference data by comparing the different coefficients' magnitude. There are model validation processes to confirm the prediction accuracy according to travel distance. This study discusses the policy implication to introduce the traveller specified route travel guidance system.

1. Introduction

Many studies have focused on the modelling behaviours of choosing routes using mathematical and empirical solutions. They have used various stochastic models to provide mathematical approaches for searching the available paths and choosing the most feasible routes [1–4]. Also, transportation researchers have attempted to formulate the route choice behaviours using empirical data. Recently, the development of intelligent transportation system (ITS), such as the vehicle detection system (VDS), automatic video system (AVS), closed-circuit television (CCTV), and variable message sign (VMS), has made it possible to collect and process the various data. These various types of travel information provide drivers' judgement about alternative routes [5].

Nevertheless, many travellers usually acquire limited information from experienced travel time [6]. The enormous amounts of data have allowed researchers to analyse travel behaviours and consider mathematical solutions. The process of generating a set of routes has been constructed using the travellers' cognitive process in choosing a route, and a reasonable number of routes have been derived from increasing the accuracy of the modelling process [7]. Furthermore, the researcher's interest in travel time reliability has increased during the last decades. The travel time reliability problem has required consideration of individuals' perceptions of the uncertainty of the travel time. The travel time perceived by an individual has been defined as a cumulative distribution function based on travellers' experiences [8, 9]. The concepts of perceived travel time and travel time reliability have been widely used to evaluate the traffic states in transportation operation.

Each traveller has a perceived travel time for a specific origin and destination (OD) pair to set up the travel time with travel time uncertainty when they start to travel. Even though there are several effective routes, travellers choose another route due to their travel experience. The process of generating a choice set involves modelling the cognitive process incorporating traveller's heterogeneity. The previous models based on rationality are somewhat limited in their ability to explain irrational choice behaviours. This problem results from the limited information about alternative routes. Many studies have recently explored these personal characteristics in the models, e.g., prospect theory, bounded rationality, and choice inertia. The route choice behaviour modelling should incorporate travellers' heterogeneity in the choice set generation and the route choice model. This study aims to construct the route choice model accommodating heterogeneous route choice behaviour for travel time reliability. The route choice model is developed through the heterogeneous K-α-reliable shortest path searching (HKαRSP) method using the reliability level derived by comparing the travel time budget (TTB) of the network of the individuals. Section 2 of this study reviews the choice set generation model and the route choice model comparing travel time reliability studies. In Section 3, the definitions of the terms are presented to establish the models. Section 4 introduces the methodology of determining the size of the choice set and the modelling of route choice behaviour which is developed using the concept of travel time reliability. An empirical analysis is conducted in Section 5 to estimate the route choice behaviour using processed travel data. Section 6 is the conclusion, which summarises the results of this study and discusses future research.

2. Literature Review

The traditional K shortest paths were modelled to determine the shortest paths concerning the travel time, assuming the determined link travel time. However, link travel times in the network have consistently been observed to have stochastic characteristics recognised by travellers. Several methodologies have been proposed to measure travel time reliability, e.g., the probability of on-time arrival, the TTB, α-reliable mean-excess travel time, and α-reliable travel time for generating the route choice set [10]. Among these methodologies, the on-time arrival probability was applied to the route choice model. The probability distribution function enabled calculating the probability of occurrence for each path, and it generated K paths according to their probabilities [11]. Mathematical modelling has been introduced to determine the optimal path from the sum of the distribution probabilities of the link travel time with the TTB [12, 13]. Researchers have employed the label-correcting algorithm to analyse the time-dependent problem and search for reliable paths [14]. The deviation-based path set generation models have defined the distribution of link travel time as normal distribution and constructed the stochastic travel time between the OD pairs to compare the network's reliable paths. These studies made it possible to remove nondominant paths and derive dominant paths through the various constraints [15, 16]. Other studies derived a user equilibrium model by dividing the travel time variation into predictable and unpredictable travel times in the route choice process using the α-reliable, mean-excess approach [17, 18]. A route choice model was constructed which reflected risk-aversion characteristics by generating the probability using the TTB [19, 20]. The other stochastic travel time-based models employed the TTB for travel time reliability to reflect individuals' heterogeneous risk-averse characteristics [21, 22]. The α-reliable travel time was used to determine the optimal path based on travellers' risk preference using the TTB. The models classified individual risk preference levels into risk-seeking, risk-neutral, and risk-averse travellers to derive the optimal paths for each scenario [16, 23]. There was other research dealing with the system optimum model reflecting the fuzzy network theory. Research has considered the perceived travel time and risk-taking properties in traffic assignment problem by incorporating fuzzy utility theorem. They discussed the differences between conventional and fuzzy network theory-based equilibrium model [24–29].

Numerous studies have been conducted to reflect the individual's choice behaviour in the model. McFadden (1973) developed the multinomial logit (MNL) model, which is a general form of the random utility-based choice model [30]. In transportation demand analysis, the MNL model has been applied at the mode choice stage before using the route choice model. The probit model and the MNL model also were used in stochastic or probabilistic assignment models [31, 32]. Many researchers have used explanatory variables to make the models more feasible, such as landmark dummy, percentage of the major road, percentage of uninterrupted flow, and delay percentage [1, 3, 33–35]. The MNL model had some drawbacks in the route choice model, i.e., (1) it does not consider the identification of an individual traveller's choice set, (2) it does not reflect the overlapping links in routes, and (3) it does not consider travellers' heterogeneity in choice behaviour. Several models have been proposed to improve the MNL model, including extended logit models to overcome overlapping links among routes. These models were composed of a deterministic term and a random error term that includes the additional overlapping term in the utility function. A modified MNL model, called the C-logit model, was proposed by subtracting the utility function's commonality [5]. Researchers have developed the implicit availability/perception (IAP) logit model by aggregating the path generation model and the route choice model using travellers' perceptions of routes [36]. The most considered route choice model to overcome the overlapping problem was the path-size logit model (PSL). The PSL model was introduced to modify the MNL model and considered the degree of overlapping of the routes in the MNL model [37–39]. The other researchers also proposed an improved PSL model, known as the path-size correction logit model (PSCL), by suggesting detailed and systematic derivations of the assumption for correcting the path-size factor [35]. There were other types of models based on the generalised extreme value (GEV) theory that considers the hierarchical structure for choices to capture the error component of overlapping links. Such models included the paired combinational logit (PCL) model [40, 41], the cross nested logit (CNL) model [34, 41, 42], the generalised nested logit (GNL) model [34, 43, 44], and the mixed multinomial logit (MMNL) model [45, 46] summarised in Table 1.

Since individual travellers have different travel experiences, they have different characteristics in determining choice sets and route choice behaviours. This study contributes a new approach for the generation of choice sets that incorporate individual travel behaviours and uses actual travel data to validate the choice behaviour. Significant differences compare the shortest path problem by perceived travel time from the previous research [16, 29]. The different methodologies of choice set generation are compared to improve the accuracy of route choice models, i.e., the k shortest paths-based choice model (KSP), the k reliable shortest paths-based choice model (KαRSP), and the heterogeneous k reliable shortest paths-based choice model (HKαRSP). The route choice models are used to compare the accuracy of choice probability according to the choice set generation methodologies from various choice models, i.e., MNL, PSL, PSCL, CNL, PCL, and MMNL models. The route choice model makes it possible to determine whether the choice set generation models are well-formulated.

3. Measuring of Individual Travel Time Reliability

3.1. Travel Time Budget (TTB). Travel time reliability is generated from the travel experience of individual travellers in specific OD pairs. It defines the distribution function of perceived travel time to obtain the probability of on-time arrival. The TTB is introduced to identify the risk preferences from the distribution as the value determined by the confidence level. The TTB has been defined the minimum total travel time threshold satisfied the reliability requirement with constraints, concerning the percentile of total travel time distribution specified by decision-makers using the confidence level, α. The meaning of this definition is interpreted as the value derived from the distribution of travel time by using the predetermined confidence level [9, 47]. Based on actual travel data and previous research, a lognormal distribution, a nonnegative and asymmetrical distribution, represents the stochastic travel time [16, 47]. Therefore, the travel time distributions for OD pairs are assumed to follow a lognormal distribution, lognormal (μ, σ), represented to the probability density function (PDF) and cumulative distribution function (CDF). In this study, TTB means that travellers plan for the travel time before departure to achieve their requirement of travel time reliability, which is expressed by the distribution of travel time experienced in the network for the confidence level, α, and the reliability level, α_l. There are three kinds of TTBs, i.e., TTB in the network, TTB of route k, and TTB for an individual. The TTB is required to achieve an α confidence level in the network from the origin, i, to the destination, j. The TTB in the network is $\mathrm{TTB}_{T^{ij}}(\alpha)$ in equation (1). TTB of the k$^{\mathrm{th}}$ path required achieving the α confidence level from the origin, i, to the destination, j. The TTB of route k in the network is $\mathrm{TTB}_{T_k^{ij}}(\alpha)$ in equation (2). The TTB required to achieve the α confidence level for individual l from the origin node i to the destination node j. The TTB for individuals is $\mathrm{TTB}_{T^{ijl}}(\alpha)$ in equation (3):

$$\mathrm{TTB}_{T^{ij}}(\alpha) = \exp\left(\mu^{ij} + \Phi_{T^{ij}}^{-1}(\alpha) \times \sigma^{ij}\right), \tag{1}$$

$$\mathrm{TTB}_{T_k^{ij}}(\alpha) = \exp\left(\mu_k^{ij} + \Phi_{T_k^{ij}}^{-1}(\alpha) \times \sigma_k^{ij}\right), \tag{2}$$

$$\mathrm{TTB}_{T^{ijl}}(\alpha) = \exp\left(\mu^{ijl} + \Phi_{T^{ijl}}^{-1}(\alpha) \times \sigma^{ijl}\right), \tag{3}$$

where i is the origin, j is the destination, k is the order of the α-reliable path or the predetermined number of the route choice set, l is the individual traveler, α is the confidence level (i.e., on-time arrival probability), μ^{ij} is the mean of the travel time distribution from the origin, i, to the destination, j, μ_k^{ij} is the mean of the travel time distribution of the k^{th} α-reliable path from the origin, i, to the destination, j, μ^{ijl} is the perceived mean of travel time distribution for individual, l, from the origin, i, to the destination, j, σ^{ij} is the standard deviation of travel time distribution from the origin, i to the destination, j, σ_k^{ij} is the standard deviation of the travel time distribution of the k^{th} α-reliable path from the origin, i, to the destination, j, and σ^{ijl} is the perceived standard deviation of the travel time distribution for individual, l, from the origin, i, to the destination, j.

3.2. Risk Preferences. The TTB has a structure combined with the predictable travel time in the travel time distribution. Travellers accept the predictable risk from their experiences to meet the predetermined travel time, which is defined as the TTB in the OD pair. Individual travellers set up a TTB for a specific OD pair using the perceived travel time based on their experience. The distribution of perceived travel time is expressed more clearly as individual travellers accumulated travel time for a specific OD pair. The travel time distribution in the network causes individual travellers to incur late arrivals because the distribution of the individual travel time is different from the distribution of the travel time determined by the network.

Reliability level (α_l) means that individuals determine the value of the probability of on-time arrival by cumulative distribution for a specific OD pair in the network based on repeated travels. When the TTB of an individual at the confidence level α has the same TTB of the network for a specific OD pair, the TTB represents the reliability level, α_l on the cumulative distribution of the network. The reliability level, α, is expressed as the on-time arrival probability for an individual's perceived TTB from the travel time distribution in this study. Risk preference is defined that the travellers have the characteristics of risk-taking for travel failure or delay due to travel time reliability. Since individual travellers have different risk preferences based on their travel experiences for each specific OD pair (i, j), the reliability level, α_l, are determined individual risk preference. The

Table 1: Summary of route choice models.

Contents		Model	Description	Study
MNL	MNL (multinomial logit)	$P(r\|Q_d) = (e^{V_r}/\sum_{r' \in Q_d} e^{V_{r'}})$	—	Dial [1], Fisk [3], Bekhor et al. [33], Prato and Bekhor [34]
	C-logit	$PP(r\|Q_d) = (e^{V_r - \alpha CF_r}/\sum_{r' \in Q_d} e^{V_{r'} - \alpha CF_{r'}})$	Subtracting commonality factor	Cascetta et al. [36], schussler and Axhausen [5], zhou et al. [18]
	PSL (path-size logit)	$P(r\|Q_d) = (e^{V_r + \beta_r l\text{n}(\text{PS}_r)}/\sum_{r' \in Q_d} e^{V_{r'} + \beta_r \ln(\text{PS}_{r'})})$	Adding ln(size) (path size)	Frejinger et al. [38], schussler and Axhausen [5], Li et al. [39]
	PSCL (path-size correction logit)	$P(r\|Q_d) = (e^{V_r + \beta_r l\text{n}(\text{PSC}_r)}/\sum_{r' \in Q_d} e^{(V_{r'} + \text{PSC}_{r'})})$	Adding PSC (path size correction)	Bovy et al.(2008)
GEV	PCL (paired-combinational logit)	$P(r\|Q_d) = e^{(\mu V_r/1 - \sigma_r)}/\sum_{r' \in Q_d} e^{(\mu V_{r'}/1 - \sigma_{r'})}$	Multiplications of unobserved probability(P_{ij})	Bliemer and Bovy [35]
	CNL (cross-nested logit)	$P(r\|Q_d) = (\kappa_{mr} e^{\mu V_r}/\sum_{r' \in Q_d} \kappa_{mr'} e^{\mu V_{r'}})$	Multiplications of Marginal(nested) probability	Prato and bekhor [34], bliemer and Bovy [35]
	GNL (generalized nested logit)	$P(r\|Q_d) = (\alpha_{mr} e^{\mu V_r}/\sum_{r' \in Q_d} \alpha_{mr'} e^{e^{\mu V_{r'}}})$	Including the allocation parameter(m, α_{nm})	Prato and bekhor [34], wen and Koppelman [44]
MNW	C-weibit	$P_h^{rs}(c^{rs}) = (e^{-\theta(c_h^{rs} + cf_h^{rs})}/\sum_{l \in c^{rs}} e^{-\theta(c_l^{rs} + cf_l^{rs})})$	Weibull distribution based model (open form)	Xu et al. [47]
	PSW (path-size weibit)	$P_{ik} = ((c_k^i - \xi_i^0)^{-\beta_i}/\sum_{s \in K_i} (c_k^i - \xi_i^0)^{-\beta_i})$		Castillo et al. [52], kitthamkesorn and Chen [53]
Mixed logit	MMNL (mixed multinomial logit)	$P_k = \Lambda(k\|\zeta) = (\exp(\mu(X_k\beta) + F_k T\zeta)/\sum_{r' \in Q_d} \exp(\mu(X_{r'}\beta) + F_{r'} T\zeta))$	Factor analytic specification	Ramming [51], Prato and Bekhor [34], Alizadeh et al. [54], Lee et al. [55]

characteristics of individual travellers were categorised, as shown below [16]:

$\alpha_l > 0.5$, risk-averse for on-time arrival

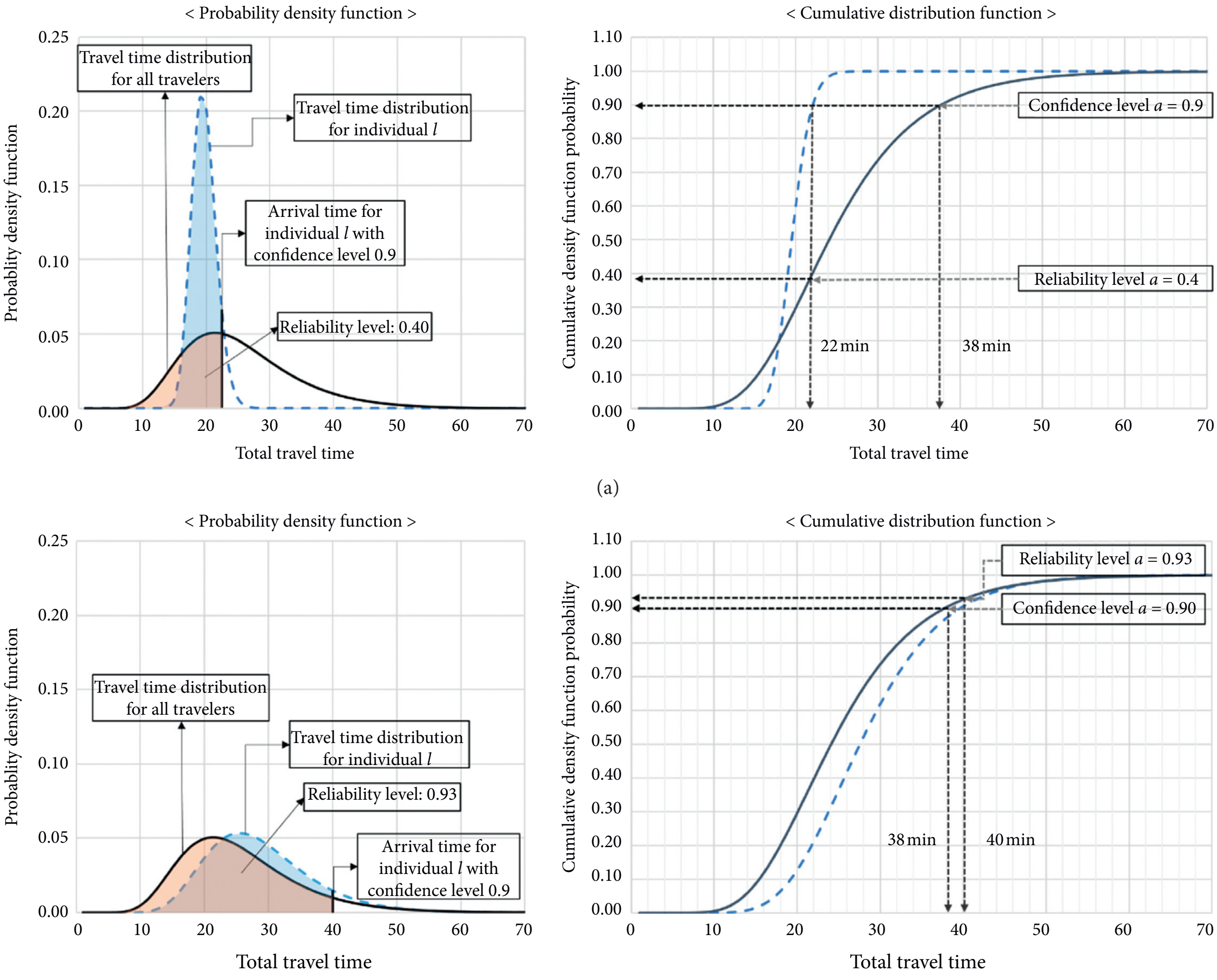
$\alpha_l = 0.5$, risk-neutral for on-time arrival

$\alpha_l < 0.5$, risk-seeking for on-time arrival

Risk preference is an essential factor in the choice set generation model. The process of choice set generation modelling is formulated using the reliability level, α_l, which is referred to as risk preference. The reliability level, α_l, is determined according to the difference between the individual's perceived travel time and the travel time provided by the network, so a difference occurred in generating the choice set. This analysis develops a route choice model that reflected travellers' behaviour according to whether they were risk-seeking or risk-averse in Figure 1. When the TTB for an individual is derived from the confidence level, α, according to the mean and standard deviation in the travel time distribution, it is possible to compare the TTB for the individual and the network confidence level, α. In other words, when the travel time experienced by an individual is less than the travel time in the network, the traveller would be concerned about late arrival based on the perceived travel time, in case of which it is defined as the risk-seeking characteristic. However, individual travellers' experiences indicate that they have more travel time than the network's travel time because they have experienced more travel time for the specific OD pair (i, j). Risk preference makes travellers calculate the TTB to arrive on time, which is a characteristic of risk-averse travellers.

4. Model Specifications

4.1. Route Choice Behaviour. The travel behaviour models have developed the following structure by dividing the choice set generation and route choice model. Researchers have tried to construct the modelling framework of route choice behaviour [7, 46]. The model is constructed to determine the size of the consideration set and individual choice set. Consideration choice set is derived by the number of experienced routes using the observed data from the universal set occurring in the network for a specific OD pair. A modelling process also includes a different choice set for individuals using TTB and risk preference in the individual choice set generation. The route choice model using the individual choice set is derived from the collective individual travel data. The individual choice set is a set of routes for incorporating traveller's heterogeneity. It is



(a)

(b)

Figure 1: Distributional characteristics for risk preferences: (a) risk-seeking travellers; (b) risk-averse travellers.

important to determine the choice set by the different travel behaviour for individuals. Likewise, travellers consist of their own considered choice set of routes from information and experience. They set for their own choice set to choose the proper route of the travel. In comparison between the cognitive and modelling process, the constructing set of choice is crucial for interpreting route choice behaviour. The cognitive process and modelling process for route choice behaviour is shown in Figure 2:

4.2. Choice Set Generation Model. Since there are millions of alternatives in the network, it is time-consuming to analyse using all alternatives in the choice set, and travellers do not consider the enormous size of the choice set for travel. Determining the choice set in the route choice model is essential because it affects the prediction accuracy in modelling results [41]. Since it is important to know the routes considered in the network, this study employs actual travel data to derive the size of the consideration set and the individual choice set. Travellers identify the choice set for their travel by travelling the known routes and determining the alternative routes. The travellers recognise the optimal path between specific OD pairs according to their individual experiences, and the individuals choose the observed route from their optimal choice sets. Thus, all observed choice sets could be the optimal paths experienced by individuals for the OD pairs. Travellers repeat creating and determining a route from the set of choices by considering their specific situations.

The route searching algorithm is developed to generate a choice set for individuals using TTB and risk preference. This algorithm generates a set of considered paths using the CDF of travel times. A set of individual paths is determined according to the number of paths specified in advance. It is necessary to generate the choice set with an appropriate size to estimate the choice probability. There are experienced paths that could be used to determine the proper size of the choice set, making it possible to know the exact path for each traveller. Also, the single alternative chosen by a traveller is one of the experienced paths. Fiorenzo-Catalano [48] mentioned the importance of determining the choice set by

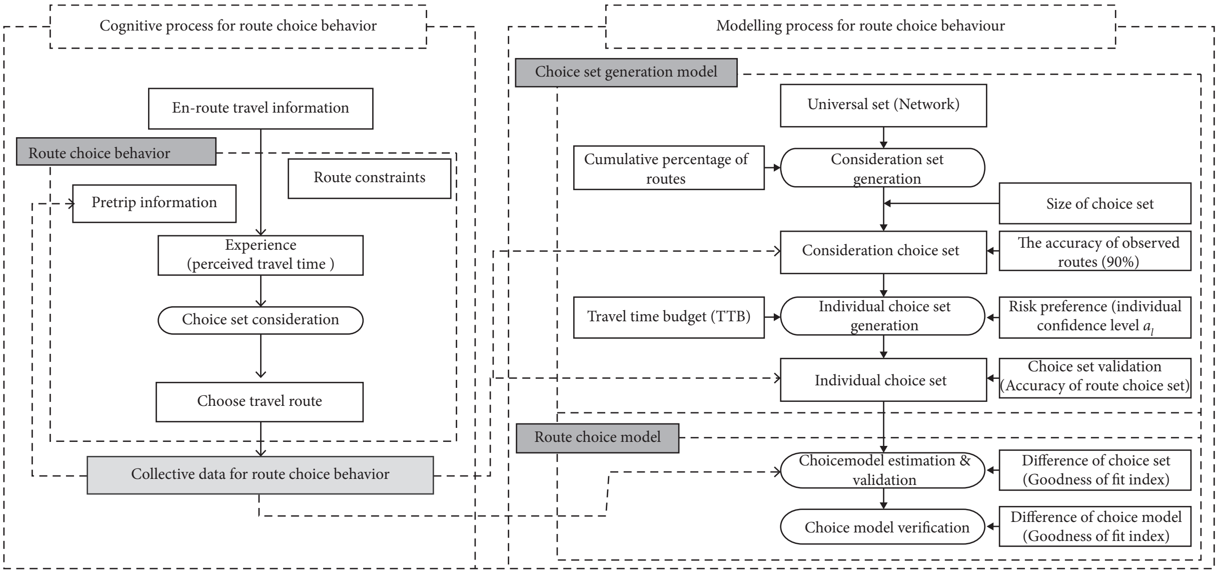

Cognitive process for route choice behavior
En-route travel information
Route choice behavior
Route constraints
Pretrip information
Experience
(perceived travel time)
Choice set consideration
Choose travel route
Collective data for route choice behavior
Modelling process for route choice behaviour
Choice set generation model
Universal set (Network)
Cumulative percentage of routes
Consideration set generation
Size of choice set
Consideration choice set
The accuracy of observed routes (90%)
Travel time budget (TTB)
Individual choice set generation
Risk preference (individual confidence level a_l
Individual choice set
Choice set validation
(Accuracy of route choice set)
Route choice model
Choicemodel estimation & validation
Difference of choice set
(Goodness of fit index)
Choice model verification
Difference of choice model
(Goodness of fit index)

Figure 2: Modelling process for route choice behaviours.

considering the researchers' perspectives because there are differences between travellers' perspectives and researchers' perspectives [48]. Since the researchers do not know individual travellers' choice sets, some assumptions are required in the choice set generation model. Two sets identify the appropriate choice set in the route choice model, i.e., the consideration set and the individual choice set. The consideration set includes the paths that most travellers are likely to choose. Besides, individual choice sets have the proper size for individual travellers to make their route choices.

The k-α-reliable shortest path searching algorithm for generating an individual choice set is consisted of eight steps, as shown below. First, the observed travel time is extracted from the database for a specific OD pair. Next, confidence level, α, is specified for the travel time reliability to achieve network performance with the value of 0.9 or more, as suggested in the previous study [16, 49]. Then, it is necessary to calculate the travel time distribution in the network and the TTB to derive the reliability level, α_l ($\mathrm{TTB}_{T^{ij}}(\alpha)$). Next, the travel time distribution is formed for each route TTB ($\mathrm{TTB}_{T_k^{ij}}(\alpha)$). From this process, the reliability level, α_l, is calculated according to the individual travel time distribution from the individual TTB ($\mathrm{TTB}_{T^{ijl}}(\alpha)$). Finally, the choice set for an individual is derived by calculating TTB according to reliability level, α_l. The algorithm for searching choice set includes the procedure for probabilistic reliable path searching algorithm for k-α-reliable shortest paths (PRPSA-KαRSP). (Algorithm 1)

Figure 3 illustrates an example to understand the differences in travel time budgets. There are five alternatives to choose the proper route for the traveller. Some travellers choose the dominant route A among the alternatives due to the fastest mean travel time; on the contrary, the other travellers are willing to choose route B for the reason of reliable travel route. In addition, the travellers varied with the formations of different choice sets considering travel experiences with reliability level, α_l.

The above algorithm is revised to generate the individual route choice set considering the observed route travels. Generating individual choice sets among various OD pairs should be repeated to model travellers' heterogeneity. Due to the different characteristics of individuals' observed choice set, it is possible to implement and derive the different perceived choice sets using the algorithm above. Even though the same travellers are on the other OD pairs, they would have different choice sets between the OD pairs due to their different experiences. The models are compared to the other models to evaluate each model's accuracy by developing the route choice model based on the choice set generation models.

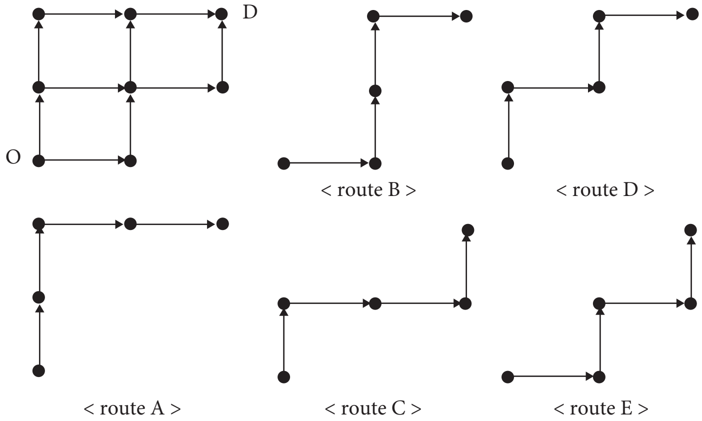
4.3. Route Choice Model.

There are various choice models to deal with the overlapping problems and cognitive process in the models. We explored which types of models are suitable for using data types and behavioural differences. There are overlapping problems in the route choice model, so it is necessary to propose an appropriate form. Also, a model that incorporated the heterogeneity of the travellers' route choice behaviour was suggested. There are various types of models, such as MNL [33], PSL and PSCL [35, 41], GNL [34], MMNL [45, 46], and MPSCL, based on the three kinds of choice set generation models. We compare those types of choice models considering data type and goodness-of-fit indexes.

Researchers tried to develop the improved model form in the overlapping problem. They developed the path-size logit model (PSL) for the improved MNL model, considering the degree of overlapping links. Bovy et al. proposed the improved path-size logit model [35]. Since there is no satisfactory derivation based on theoretical arguments, it is necessary to employ the correction terms. The model

```
Step 1. Choosing the OD pair to observe the path $(i, j)$
Step 2. Setting the confidence level, $\alpha$, for the satisfaction of the level of service, i.e.,$\alpha = 0.9$
Step 3. Building the distribution of travel time ($T^{ij}$) for travel from origin $i$ to destination $j$, and calculating the TTB ($\mathrm{TTB}_{T^{ij}}(\alpha)$), concerning the confidence level, $\alpha$
Step 4. Building the distribution of travel time for the $k^{\mathrm{th}}$ path ($T_k^{ij}$), $\forall k \in (1, \ldots, U)$, for each observed travel from origin $i$ to destination $j$
Step 5. Building the distribution of travel time for individual $l$ ($T^{ijl}$) for each traveller from origin $i$ to destination $j$
Step 6. Evaluating the reliability level, $\alpha_l$, for each traveller $l$, $\alpha_l = \Phi_{T^{ijl}}(\ln(\mathrm{TTB}_{T^{ij}}(\alpha) - \mu^{ijl})/\sigma^{ijl})$, normal distribution $\Phi(x) = (1/2)[1 + \mathrm{erf}(x/\sqrt{2})]$, where erf = error function
Step 7. Calculating the TTB ($\mathrm{TTB}_{T_k^{ij}}(\alpha_l)$), for each path concerning the reliability level, $\alpha_l$
Step 8. Choosing the $K$-$\alpha$-reliable shortest paths for individual $l$
```

Algorithm 1: PRPSA-KαRSP.



< TTB for alternative route between OD pair >

$TTB_{TA}^{OD}(0.9) = exp(2.6 + 1.282 \times 0.5) = 25.56$

$TTB_{TB}^{OD}(0.9) = exp(3.0 + 1.282 \times 0.1) = 22.83$

$TTB_{TC}^{OD}(0.9) = exp(2.8 + 1.282 \times 0.3) = 24.16$

$TTB_{TD}^{OD}(0.9) = exp(2.9 + 1.282 \times 0.3) = 26.70$

$TTB_{TE}^{OD}(0.9) = exp(3.1 + 1.282 \times 0.2) = 28.69$

Figure 3: Route choice behaviour from perceived travel time.

considers the impact of choice set in the route choice model in equations (4) and (5):

$$P(i|A_n) = \frac{e^{V_{in}+\beta_{\mathrm{PSC}} * \mathrm{PSC}_{in}}}{\sum_{j\in A_n} e^{V_{jn}+\beta_{\mathrm{PSC}} * \mathrm{PSC}_{jn}}}, \qquad (4)$$

$$\mathrm{PSC}_i = -\sum_{a\in\Gamma_i}\left(\left(\frac{l_a}{L_i}\right)\ln\sum_{j\in A_n}\delta_{aj}\right), \qquad (5)$$

where $\mathbf{V}$ is the A_n by K matrix of variables, β is the column vector of K unknown parameters for variables, PSC_{in} is the A_n by one vector of path-size correction term, L_i is the length of the travelled route of alternative i, l_a is the overlapped link a, and δ_{aj} is the binary variable if the link a exists in route L_i, 1, otherwise 0.

Moreover, researchers have proposed a mixed logit model to overcome the limitations of the logit model by adding error terms in the equation to account for the correlation among routes [39]. Since travellers' perceived routes are correlated, the error term is added to illustrate the relationship based on the topology of paths. The error term is divided into two parts in the model. One part represents correlation and heterogeneity, and the other part describes *i.i.d* (independently identically differentiated) extreme value. The equation of MPSCL is presented as

$$\mathbf{P_n(i) = \Lambda(i|\xi) = \frac{\exp(\mu(X_{in}\beta + F_{in}T\xi) + \ln(PSC_{in}))}{\sum_{J\in C_n}\exp(\mu(X_{jn}\beta + F_{jn}T\xi) + \ln(PSC_{jn}))}}, \qquad (6)$$

where $\mathbf{X}$ is C_n by K matrix of variables, β is the column vector of K unknown parameters for variables, $\mathbf{F_{in}T}\xi$ is C_n by one vector of error terms, $\mathbf{F}$ is the C_n by M factor loading matrix, $\mathbf{T}$ is M by M lower triangular matrix of unknown parameters, ζ is M by one vector of i.i.d standard normal variables as unobservable factors, ν is M by M lower triangular matrix of unknown parameters, and $\Gamma(k\,|\,\zeta)$ is the probability of chosen route k with given ζ.

5. Revealed Preference Routing Data

5.1. Data Descriptions. A case study was performed to apply the proposed methodology to solve the HKαRSP problem. The actual travelled data were used on the road network in the Daegu metropolitan area in South Korea. The actual path travel data were constructed by processing the information collected by the roadside equipment (RSE) installed on the intersections between arterial roads. Information of vehicular travel was collected using telecommunications between the RSE device installed on the road and the on-board unit (OBU) device installed in vehicles by a dedicated short-range communication (DSRC) device.

The DSRC was a useful technique for collecting traffic information, such as the number of vehicles passing by a specific location. The data were more accurate than the GPS data used in previous research. However, since it collected point data, a conversion process was required to track the OBU ID of an individual vehicle observed from the RSE to convert the data into individual route data. The model included a process of generating routes to track an individual's chosen routes. It used the route data with the high frequency for a specific OD pair to model an individual's perceived travel time. From a brief analysis of the data, basic statistics and study area are shown in Table 2.

5.2. Data Processing and Missing Correction. Since the DSRC data was a type of point-based data observed at an intersection of the arterial roads, it was necessary to convert them into route data. The process for tracking the travellers with the same vehicle ID (OBU ID) was conducted to identify each route. The model was constructed based on methods of classifying and generating route data by tracking individual vehicles.

It was necessary to identify the individual vehicles to change from point data to route data. The observed-time variable was used to construct this process. If the observed times for individuals on RSE were arranged in order, it was possible to produce the individuals' route data. The link travel time was calculated while moving from node to node, and it included checking whether the path was configured using the link travel time. To generate the route travel time for specific OD pairs, it is necessary to produce the route travel data from point observation data. When the link travel time was excessive from a certain marginal value (divided by 10 minutes), it was divided into different travels [50]. We also scattered the plot using the observed travel time to separate the route travel, including about 98% of travels in 10 minutes. The route data process has presented the step by step to generate each travellers' route travel, sequentially listing the data observed at the point (see Figure 4(a)). However, it was impossible to confirm whether the link between the two nodes is connected or not. There were the following three types of missing data. (1) Missing data between nodes on an arterial road by straightway, (2) missing data between nodes for the type of road with the uninterrupted flow, and (3) observation of one node on two different observations. It was necessary to define the links between the nodes to ensure whether they were related links. If the produced route data were the case of missing data, it was necessary to identify the target nodes or links. With the missing correction method, the route data were connected with the other node. This process was performed using all of the missing data. The developed algorithm performed the missing correction procedures (see Figure 4(b)).

6. Results

6.1. Structure of Route Choice Model. The specified model has required the actual data to generate individual choice sets based on the distribution of perceived travel times. The individual choice set was a set of paths that incorporated the travellers' heterogeneity. The choice set generation model determines the individual choice sets based on the different travel experiences. The route choice model that incorporated the heterogeneous choice set generation model was used to compare the travel behaviours.

There were more than 30 thousands of possible OD pairs among nodes. It was necessary to choose the feasible data for analysis of route choice behaviour. Since some OD pairs were too close or far away to analyse the travel behaviours, available OD pairs were selected, having more than twenty thousand observed trips and proper distances within 5 km to 25 km between OD pairs. The 76 OD pairs were chosen for the analysis to describe the heterogeneous travel behaviours. From the observations, 40 thousands travellers having frequent observed trips were selected for the final analysis. As mentioned before, it was important to determine the appropriate set to be considered from the thousands of alternatives. A methodology was established for choosing the choice set to be considered using actual travel data. The consideration choice sets should include all of the possible choice sets for most of the travellers. According to the assumptions presented above, we determined the possible number of consideration set and the individual's number of the choice set (K) using the observed data. Since the use of all observed paths was against the assumption, the size of the consideration set was determined to 16 observed paths considering 90% of the coverage probability as consideration choice set. It is necessary to determine how many travelled paths were chosen in the choice set for individuals from the observed data. To determine the alternative K for each individual, the 80% observed routes for each individual were calculated on average 3.12 routes except for observed at once, and the number of individual choices set was determined as four paths in the model.

The developed model used the actual travel data to analyse the route choice model. The NLOGIT 6.0 program, which is generally used for econometric analysis, was used to analyse the route choice model in this study. The MNL model was developed for estimating the parameters in the choice model with maximum likelihood estimation (MLE) methods. Generally, the more explanatory variables make a better goodness-of-fit index, but the correlated variables decrease the accuracy of parameter estimation. Even though there are many other kinds of variables from the raw data, it is necessary to analyse the correlation among variables to identify the effects of parameters appropriately. The explanatory variables were compared to whether the variables improve the goodness-of-fit or multicollinearity, and Pearson correlation analysis was employed to choose the appropriate variables. The model was developed to compare the relative size of variables between alternatives in the model without alternative specific constants (ASCs). It was necessary to retain the dummy variables to avoid biasedness [51]. Since travellers tended to consider more travel attributes than an immanent attribute of alternatives in route choice, the additional variables were needed instead of ASCs in the model. The variables were used to analyse the route

Table 2: Data description and empirical study area (Daegu metropolitan area).

Division	Value	Unit	
# of observed travellers	About 0.6 million per month	Travellers	
# of OD pair	31,152	Pair	
Mean of travelled route	About 30 million per month	Trips	
Mean of link distance	1.25	Km	
Mean of OD trips	6,015	Trips	

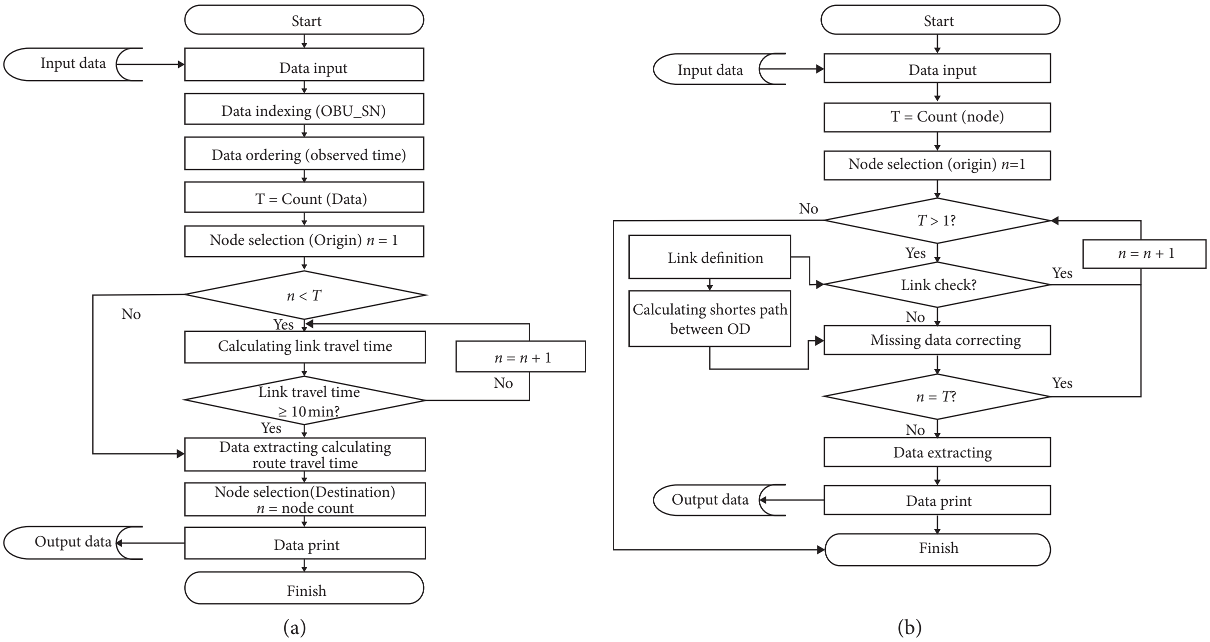


Figure 4: Data processing and correction: (a) algorithm for route data processing; (b) algorithm for the correction of missing data.

choice model using the DSRC data, i.e., travel time, buffer time, distance, ratio of uninterrupted flow road, tolls, and number of bridges. The final model was established with the several chosen variables of the following equation:

$$V_k = \beta_1 \mu_k^{ij} + \beta_2 \mathrm{BT}_k^{ij} + \beta_3 \mathrm{DIST}_k^{ij} + \beta_4 \mathrm{UNINT}_k^{ij} + \beta_5 \mathrm{TOLL}_k^{ij} + \beta_6 \mathrm{BRIDGE}_k^{ij}, \tag{7}$$

where V_k is the utility function for alternative k, β_i are the parameters, μ_k^{ij} is the mean travel time for alternative k from origin i to destination j, BT_k^{ij} is the buffer travel time for alternative k from origin i to destination j, DIST_k^{ij} is the distance travelled for alternative k from origin i to destination j, UNINT_k^{ij} is the ratio of uninterrupted flow for alternative k from origin i to destination j, TOLL_k^{ij} is the toll for alternative k from origin i to destination j, and BRIDGE_k^{ij} is the number of bridges for alternative k from origin i to destination j.

6.2. Heterogeneous Route Choice Models. The model was determined by evaluating the data, modelling structure, and goodness-of-fit index among the various other models, i.e., MNL, PSL, PSCL, MMNL, and MPSCL. This research employed the MPSCL model reflecting the overlapping links and considering the traveller's heterogeneity. The MPSCL model is necessary to analyse the route choice behaviour considered the route overlapping, which has a significant impact on the model's estimation, and the model had a much improved ρ^2 compared to the other models. The result showed the route choice model based on the different choice set generation model. The results of the model comparison are presented in Table 3.

The proposed model provided the most precise prediction of a route's choice probability using choice set generation with traveller heterogeneity. Due to the coincidence of consideration set generation and path-size correction term, the model had better model fitness indexes. Consideration of identified choice set for travellers was adopted in the MPSCL model. The model had a better accuracy of prediction for route choice probability in HKαRSP model than the KαRSP and KSP model.

The estimated model parameters had the appropriate value in the model and drew the significance at 1% level for most models. The parameters represented the variables' variations; in other words, the variables had a different effect on the choice for individual travellers, which is modelled by the random parameters. The mean and standard deviation parameters of travel time affected the model in the MPSCL

Table 3: Result of MPSCL model with truncated normal distribution.

Explanatory variables		HKαRSP	KαRSP	KSP
Level of service (LOS) attribute variable				
Mean travel time (μ_k^{ij})	Constant	−0.3121***	−0.0404***	−0.0343***
	Standard deviation	1.3852***	0.3073***	0.2865***
Buffer time (BT_k^{ij})		−0.3129***	−0.2300***	−0.4009***
Travel distance (DIST_k^{ij})		−0.5840***	−1.1117***	−0.9177***
Network attribute variable				
The ratio of uninterrupted flow road (UNINT_k^{ij})		4.5800***	2.4270***	—
Toll fare (TOLL_k^{ij}) (100won)		−0.0959***	—	−0.0454***
Number of bridge (BRIDGE_k^{ij})		−1.1636***	−0.9013***	−0.0508***
Path-size correction (PSC)		−4.0237***	−3.5107***	−1.7035***
Goodness of fit				
Observations		40,000	40,000	40,000
# of parameters		8	8	8
LL (0)		−55,451.8	−55,451.8	−55,451.8
LL (β)		−35,558.1	−36,551.9	−44,550.1
ρ^2		0.3588	0.3408	0.1966
$\bar{\rho}^2$		0.3586	0.3407	0.1964

* is 10% significance level, ** is 5% significance level, *** is 1% significance level.

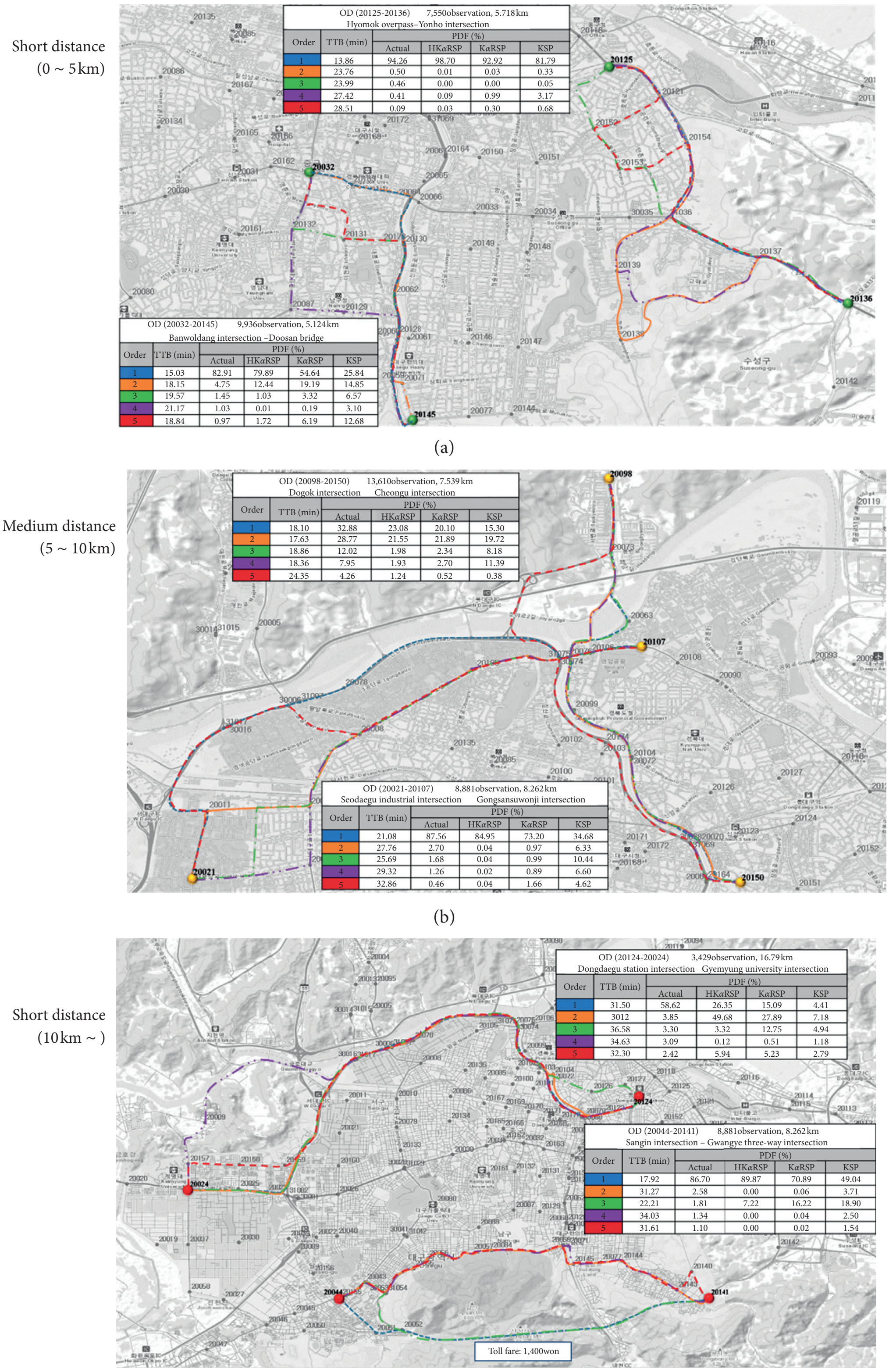
model, which means that travel time makes the differences in choice probability with traveller heterogeneity. The random parameters positively affected the log-likelihood estimation compared to basic models to improve parameters' accuracy. All parameters of the models had the appropriate values and were significant at 1%. From the HKαRSP model result, there was the effect of the level of service and attribute network variables in the route choice model. Travellers considered the variation of mean travel time (β_1 constant: −0.3121/standard deviation: 1.3852) by the heterogeneous choice behaviour. Travellers had the tendency not to increase the mean travel time, but they also were sensitive to the variation of travel time for the travel. The model is more sensitive to travel time reliability (BT, β_2: −0.3129) than average travel time (β_1: −0.3121). The ratio of uninterrupted flow road (β_4: 4.5800) had a significant positive effect on route choice behaviour due to the convenience of driving. They tended to choose the routes having many ratios of the uninterrupted travel route. Also, travellers did not want to choose the route with the bridges (β_6: −1.1636), and it seemed to make congestion on the bridges.

There was an impact on the level of service attribute variables in the route choice model. The use of buffer time derived the better goodness of fit than the standard deviation of travel time for travel time reliability. The model had more sensitive to the travel time reliability (buffer time; BT) than average travel time (μ_k^{ij}). Less distance made the better model fit than the mean of travel time in route choice. Furthermore, there was an impact of additional network attribute variables in the route choice model. The higher ratio of uninterrupted flow revealed a higher choice probability in the model. There was a tendency for less preference to use of toll road for travel in the urban area. Travellers tended to avoid crossing the bridge in the model due to traffic congestion. Reflecting the traveller heterogeneity in the mixed logit model made the accuracy of estimations. This was due to the consistency of the structure for a choice set generation model and route choice model. Also, we evaluated the best fit for the MPSCL model with HKαRSP choice set generation model. The model had better model fitness indexes which resulted from the coincidence of consideration set generation and path-size correction term.

Many studies have recently been conducted to provide a new concept of transportation services such as smart mobility, mobility-as-a-service (MaaS), and an autonomous vehicle. These studies focused on identifying individual preferences and providing more convenient service by combining various travel modes suitable for those preferences. From this perspective, analysing route choice behaviour based on individual travel experience would be an important process in introducing new transportation services. The results derived through this study were judged to establish a more efficient transportation operation strategy by providing information on the reliable route for an individuals' preference. The provision of transportation services should provide faster information from individuals' experiences, and such information makes the entire system operate efficiently.

6.3. Model Validation.

We validated the prediction accuracy for the route choice probability using the estimated parameters. There are differences according to the distance between OD pairs, and it is necessary to divide with the three categories based on the distance (short/medium/long distance). The prediction results were calculated by the observed travel attributes for each OD pair considering types of



OD (20125-20136) 7,550observation, 5.718 km
Hyomok overpass–Yonho intersection

Order	TTB (min)	PDF (%)			
		Actual	HKαRSP	KαRSP	KSP
🟦 1	13.86	94.26	98.70	92.92	81.79
🟧 2	23.76	0.50	0.01	0.03	0.33
🟩 3	23.99	0.46	0.00	0.00	0.05
🟪 4	27.42	0.41	0.09	0.99	3.17
🟥 5	28.51	0.09	0.03	0.30	0.68

OD (20032-20145) 9,936observation, 5.124 km
Banwoldang intersection –Doosan bridge

Order	TTB (min)	PDF (%)			
		Actual	HKαRSP	KαRSP	KSP
🟦 1	15.03	82.91	79.89	54.64	25.84
🟧 2	18.15	4.75	12.44	19.19	14.85
🟩 3	19.57	1.45	1.03	3.32	6.57
🟪 4	21.17	1.03	0.01	0.19	3.10
🟥 5	18.84	0.97	1.72	6.19	12.68

(a)

OD (20098-20150) 13,610observation, 7.539 km
Dogok intersection Cheongu intersection

Order	TTB (min)	PDF (%)			
		Actual	HKαRSP	KαRSP	KSP
🟦 1	18.10	32.88	23.08	20.10	15.30
🟧 2	17.63	28.77	21.55	21.89	19.72
🟩 3	18.86	12.02	1.98	2.34	8.18
🟪 4	18.36	7.95	1.93	2.70	11.39
🟥 5	24.35	4.26	1.24	0.52	0.38

OD (20021-20107) 8,881observation, 8.262 km
Seodaegu industrial intersection Gongsansuwonji intersection

Order	TTB (min)	PDF (%)			
		Actual	HKαRSP	KαRSP	KSP
🟦 1	21.08	87.56	84.95	73.20	34.68
🟧 2	27.76	2.70	0.04	0.97	6.33
🟩 3	25.69	1.68	0.04	0.99	10.44
🟪 4	29.32	1.26	0.02	0.89	6.60
🟥 5	32.86	0.46	0.04	1.66	4.62

(b)

OD (20124-20024) 3,429observation, 16.79 km
Dongdaegu station intersection Gyemyung university intersection

Order	TTB (min)	PDF (%)			
		Actual	HKαRSP	KαRSP	KSP
🟦 1	31.50	58.62	26.35	15.09	4.41
🟧 2	3012	3.85	49.68	27.89	7.18
🟩 3	36.58	3.30	3.32	12.75	4.94
🟪 4	34.63	3.09	0.12	0.51	1.18
🟥 5	32.30	2.42	5.94	5.23	2.79

OD (20044-20141) 8,881observation, 8.262 km
Sangin intersection – Gwangye three-way intersection

Order	TTB (min)	PDF (%)			
		Actual	HKαRSP	KαRSP	KSP
🟦 1	17.92	86.70	89.87	70.89	49.04
🟧 2	31.27	2.58	0.00	0.06	3.71
🟩 3	22.21	1.81	7.22	16.22	18.90
🟪 4	34.03	1.34	0.00	0.04	2.50
🟥 5	31.61	1.10	0.00	0.02	1.54

(c)

Figure 5: Model validation considering travel distance.

choice set generation methods. The validation results are shown in Figure 5.

7. Conclusion

In this study, the distributional characteristics were employed to model the uncertainty concerning the travel time for individual travellers. We used the concept of TTB and the probability that travellers would arrive at their destinations on time. The definition of risk preference was introduced according to the difference between the TTB considered by individual travellers and the TTB presented in the network. There was a process for generating an individual choice set based on the accumulated experience of individuals. The process of route choice was performed to consider a different choice set for each traveller. The reliability level, α_l, generated a path set by the cumulative travel time distribution for each path. We constructed a model for generating the choice set for individual travellers to incorporate the traveller's heterogeneity. The results obtained from actual path travel data showed that most travellers might consider the dominant path and select alternative paths similar to it if one dominant path exists. Also, travellers chose reliable paths to ensure on-time arrivals by the generated choice set.

The travellers were more sensitive to travel distance than travel time in the level of service attributes. The coefficients of travel time were in the range from −0.3121 to 0.0077, and the coefficients of travel distance were in the range from −1.1117 to −0.5840 in the level of service attributes. The travellers tended to have preferences for the use of uninterrupted flow and bridges, and they preferred not to use toll roads. The coefficients of the ratio of uninterrupted flow were in the range from −0.1099 to 4.7544, the estimation result of toll roads was in the range from −0.0959 to −0.0342, and the parameters of bridges were in the range from −1.1636 to −0.0508. The model had a better accuracy of prediction for route choice probability in the HKαRSP model than the KαRSP and KSP models. We derived better prediction according to the different travel distances. The results are applicable to transportation planning and traffic management by clarifying the choice set considered in the existing network. Moreover, it was possible to establish a strategy for providing route information using individuals' behavioural characteristics concerning transportation operation. Depending on the individual's risk preference, a different set of paths was considered, and a set of paths was established to provide information that is tailored to the individual reliability level, α_l. This study contributes to increasing the efficiency of traffic operation and planning according to individuals' route attributes.

There is further research from the additional improvements in modelling. The choice set generation model derives the appropriate number of sets as a necessary process for constructing the route choice model. It is necessary to compute travel time distribution following the time-dependent model to compare the differences in the choice sets. Also, the methodology for estimating the route travel time can be developed based on the difference between an individual's actual travel time on a given route and the estimated route travel time from the link travel time distribution. Finally, this research can extend the stochastic user equilibrium model according to travellers' risk preferences using the route choice model, such as the fuzzy traffic assignment model.

Data Availability

The data used in this research were provided by the Trlab research programme conducted at the Seoul National University, Seoul, Republic of Korea. The data used to support the findings of the study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (2020R1F1A1061802). The authors express their gratitude for the support.

References

[1] R. B. Dial, "A probabilistic multipath traffic assignment model which obviates path enumeration," *Transportation Research*, vol. 5, no. 2, pp. 83–111, 1971.

[2] J. Y. Yen, "Finding the *K* Shortest loopless paths in a network," *Management Science*, vol. 17, no. 11, pp. 712–716, 1981.

[3] C. Fisk, "Some developments in equilibrium traffic assignment," *Transportation Research Part B: Methodological*, vol. 14, no. 3, pp. 243–255, 1980.

[4] M. Chen and A. S. Alfa, "A network design algorithm using a stochastic incremental traffic assignment approach," *Transportation Science*, vol. 25, no. 3, pp. 215–224, 1991.

[5] N. Schuessler and K. Axhausen, "Processing raw data from global positioning systems without additional information," *Journal of the Transportation Research Board*, vol. 2105, pp. 28–36, 2009.

[6] X. Pan Zuo and K. Liu, "Parameter calibration in cumulative prospect theory for travellers' route choice behaviour," in *Proceedings of the COTA International Conference of Transportation Professionals*, pp. 2696–2708, Beijing, China, July 2015.

[7] P. H. L. Bovy, "On modelling route choice sets in transportation networks: a synthesis," *Transport Reviews*, vol. 29, no. 1, pp. 43–68, 2009.

[8] W. H. K. Lam, S. C. Wong, and H. K. Lo, *Transportation and Traffic Theory 2009: Golden Jubilee*, Springer, Berlin, Germany, 2009.

[9] C. Sun, L. Cheng, S. Zhu, F. Han, and Z. Chu, "Multi-criteria user equilibrium model considering travel time, travel time reliability and distance," *Transportation Research Part D: Transport and Environment*, vol. 66, pp. 3–12, 2019.

[10] K. K. Srinivasan, A. A. Prakash, and R. Seshadri, "Finding most reliable paths on networks with correlated and shifted log-normal travel times," *Transportation Research Part B: Methodological*, vol. 66, pp. 110–128, 2014.

[11] H. Frank, "Shortest paths in probabilistic graphs," *Operations Research*, vol. 17, no. 4, pp. 583–599, 1969.

[12] Y. Nie and Y. Fan, "Arriving-on-time problem: discrete algorithm that ensures convergence, transportation research record," *Journal of the Transportation Research Board*, vol. 1964, pp. 193–200, 2006.

[13] B. Y. Chen, W. H. K. Lam, A. Sumalee et al., "Finding reliable shortest paths in road networks under uncertainty," *Networks and Spatial Economics*, vol. 13, no. 2, pp. 123–148, 2013.

[14] Y. M. Nie and X. Wu, "Reliable a priori shortest path problem with limited spatial and temporal dependencies," in *Transportation and Traffic Theory 2009: Golden Jubilee*, pp. 169–195, Springer, Berlin, Germany, 2009.

[15] B. Y. Chen, W. H. K. Lam, A. Sumalee, and Z.-I. Li, "Reliable shortest path finding in stochastic networks with spatial correlated link travel times," *International Journal of Geographical Information Science*, vol. 26, no. 2, pp. 365–386, 2012.

[16] B. Y. Chen, Q. Li, and W. H. K. Lam, "Finding the *K* Reliable shortest paths under travel time uncertainty," *Transportation Research Part B: Methodological*, vol. 94, pp. 189–203, 2016.

[17] A. Chen and Z. Zhou, "The α-reliable mean-excess traffic equilibrium model with stochastic travel times," *Transportation Research Part B: Methodological*, vol. 44, no. 4, pp. 493–513, 2010.

[18] A. Chen, Z. Zhou, and W. H. K. Lam, "Modeling stochastic perception error in the mean-excess traffic equilibrium model," *Transportation Research Part B: Methodological*, vol. 45, no. 10, pp. 1619–1640, 2011.

[19] H. K. Lo and Y.-K. Tung, "Network with degradable links: capacity analysis and design," *Transportation Research Part B: Methodological*, vol. 37, no. 4, pp. 345–363, 2003.

[20] H. K. Lo, X. W. Luo, and B. W. Y. Siu, "Degradable transport network: travel time budget of travellers with heterogeneous risk aversion," *Transportation Research Part B: Methodological*, vol. 40, no. 9, pp. 792–806, 2016.

[21] X. Wu and Y. Nie, "Implementation issues for the reliable a priori shortest path problem, transportation research record," *Journal of the Transportation Research Board*, vol. 2091, pp. 51–60, 2009.

[22] X. Wu, "Study on mean-standard deviation shortest path problem in stochastic and time-dependent networks: a stochastic dominance based approach," *Transportation Research Part B: Methodological*, vol. 80, pp. 275–290, 2015.

[23] A. Chen and Z. Ji, "Path finding under uncertainty," *Journal of Advanced Transportation*, vol. 39, no. 1, pp. 19–37, 2005.

[24] H. Ramazani, Y. Shafahi, and S. E. Seyedabrishami, "A fuzzy traffic assignment algorithm based on driver perceived travel time of network links," *Scientia Iranica*, vol. 18, no. 2, pp. 190–197, 2011.

[25] G. E. Cantarella and V. Fedele, "Fuzzy utility theory for analysing discrete choice behaviour," in *Proceedings of the Fourth International Symposium on Uncertainty Modeling and Analysis, ISUMA 2003*, pp. 148–154, College Park, MD, USA, September 2003.

[26] M. Ridwan, "Fuzzy preference based traffic assignment problem," *Transportation Research Part C: Emerging Technologies*, vol. 12, no. 3-4, pp. 209–233, 2004.

[27] T. Lotan, "Effects of familiarity on route choice behavior in the presence of information," *Transportation Research Part C: Emerging Technologies*, vol. 5, no. 3-4, pp. 225–243, 1997.

[28] M. Miralinaghi, Y. Lou, Y. T. Hsu, R. Shabanpour, and Y. Shafahi, "Multiclass fuzzy user equilibrium with endogenous membership functions and risk taking behaviors," *Journal of Advanced Transportation*, vol. 50, no. 8, pp. 1716–1734, 2016.

[29] M. Miralinaghi, Y. Shafahi, and R. S. Anbarani, "A fuzzy network assignment model based on user equilibrium condition," *Scientia Iranica A*, vol. 22, no. 6, pp. 2012–2023, 2015.

[30] D. McFadden, *Conditional Logit Analysis of Qualitative Choice Behaviour,*University of California at Berkeley, Berkeley, CL, USA, 1973.

[31] Y. Sheffi and W. B. Powell, "An algorithm for the equilibrium assignment problem with random link times," *Networks*, vol. 12, no. 2, pp. 191–207, 1982.

[32] E. Cascetta, F. Russo, F. A. Viola, and A. Vitetta, "A model of route perception in urban road networks," *Transportation Research Part B: Methodological*, vol. 36, no. 7, pp. 577–592, 2002.

[33] S. Bekhor, M. E. Ben-Akiva, and M. S. Ramming, "Evaluation of choice set generation algorithms for route choice models," *Annals of Operations Research*, vol. 144, no. 1, pp. 235–247, 2006.

[34] C. Prato and S. Bekhor, "Applying branch-and-bound technique to route choice set generation, Transportation Research Record," *Journal of the Transportation Research Board*, vol. 1985, pp. 19–28, 2006.

[35] P. Bovy, S. Bekhor, and C. Prato, "The factor of revisited path size: alternative derivation," *Journal of the Transportation Research Board*, vol. 2076, pp. 132–140, 2008.

[36] E. Cascetta, A. Papola, F. Russo et al., "Implicit availability/perception logit models for route choice in transportation networks," *World Transport Research: Selected Proceedings of the 8th World Conference on Transport Research World Conference on Transport Research Society*, vol. 3, 1999.

[37] M. E. Ben-Akiva and J. L. Bowman, *Activity Based Travel Demand Model Systems: Equilibrium and Advanced Transportation Modelling*, pp. 27–46, Springer, Berlin, Germany, 1998.

[38] E. Frejinger, M. Bierlaire, and M. Ben-Akiva, "Sampling of alternatives for route choice modeling," *Transportation Research Part B: Methodological*, vol. 43, no. 10, pp. 984–994, 2009.

[39] D. Li, T. Miwa, T. Morikawa, and P. Liu, "Incorporating observed and unobserved heterogeneity in route choice analysis with sampled choice sets," *Transportation Research Part C: Emerging Technologies*, vol. 67, pp. 31–46, 2016.

[40] F. S. Koppelman and C.-H. Wen, "The paired combinatorial logit model: properties, estimation and application," *Transportation Research Part B: Methodological*, vol. 34, no. 2, pp. 75–89, 2000.

[41] M. Bliemer and P. Bovy, "Impact of route choice set on route choice probabilities," *Journal of the Transportation Research Board*, vol. 2076, pp. 10–19, 2008.

[42] P. Vovsha and S. Bekhor, "Link-nested logit model of route choice: overcoming route overlapping problem," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1645, no. 1, pp. 133–142, 1998.

[43] S. Bekhor and J. N. Prashker, "Stochastic user equilibrium formulation for generalized nested logit model," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1752, no. 1, pp. 84–90, 2001.

[44] C.-H. Wen and F. S. Koppelman, "The generalized nested logit model," *Transportation Research Part B: Methodological*, vol. 35, no. 7, pp. 627–641, 2001.

[45] M. Bierlaire and E. Frejinger, "Route choice models with subpath components," in *Proceedings of the Swiss*

Transportation Research Conference, Ascona, Switzerland, August 2005.

[46] S. Hess and J. M. Rose, "Allowing for intra-respondent variations in coefficients estimated on repeated choice data," *Transportation Research Part B: Methodological*, vol. 43, no. 6, pp. 708–719, 2009.

[47] X. Xu, A. Chen, L. Cheng et al., "Modeling distribution tail in network performance assessment: a mean-excess total travel time risk measure and analytical estimation method," *Transportation Research Part B: Methodological*, vol. 66, pp. 32–49, 2014.

[48] M. S. Fiorenzo-Catalano, *Choice Set Generation in Multi-Modal Transportation Networks*, Netherlands TRAIL Research School, Delft, Netherlands, 2007.

[49] X. Wu and Y. Nie, "Modeling heterogeneous risk-taking behavior in route choice: a stochastic dominance approach," *Transportation Research Part A: Policy and Practice*, vol. 45, no. 9, pp. 896–915, 2011.

[50] X. Zhan, S. Hasan, S. V. Ukkusuri, and C. Kamga, "Urban link travel time estimation using large-scale taxi data with partial information," *Transportation Research Part C: Emerging Technologies*, vol. 33, pp. 37–49, 2013.

[51] M. S. Ramming, *Network Knowledge and Route Choice*, Massachusetts Institute of Technology, Cambridge, MA, USA, 2001.

[52] E. Castillo, J. M. Menendez, P. Jimenez, and A. Rivas, "Closed form expressions for choice probabilities in the Weibull case," *Transportation Research Part B: Methodological*, vol. 42, no. 4, pp. 373–380, 2008.

[53] S. Kitthamkesorn and A. Chen, "A path-size weibit stochastic user equilibrium model," *Transportation Research Part B: Methodological*, vol. 57, pp. 378–397, 2013.

[54] H. Alizadeh, B. Farooq, C. Morency, and N. Saunier, "On the role of bridges as anchor points in route choice modeling," *Transportation*, vol. 45, no. 5, pp. 1181–1206, 2017.

[55] J.-H. Lee, S.-H. Cho, D.-K. Kim, and C. Lee, "Valuation of travel time reliability accommodating heterogeneity of route choice behaviors," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2526, pp. 86–93, 2016.

Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 5585542, 10 pages
https://doi.org/10.1155/2021/5585542

Research Article

Survey Data Analysis on Intention to Use Shared Mobility Services

Eunjeong Ko,[1] Hyungjoo Kim,[2] and Jinwoo Lee[1]

[1] *The Cho Chun Shik Graduate School of Green Transportation, Korea Advanced Institute of Science and Technology, Daejeon 34051, Republic of Korea*
[2] *Advanced Institute of Convergence Technology, Changeop-ro 42, Sujeong-gu, Seongnam, Gyeonggi-do 13449, Republic of Korea*

Correspondence should be addressed to Jinwoo Lee; lee.jinwoo@kaist.ac.kr

Received 5 February 2021; Revised 6 March 2021; Accepted 22 March 2021; Published 7 April 2021

Academic Editor: Zhibin Li

Shared mobility is a service that allows users to share various transportation modes and use them with reservations when necessary. It started with private automotive car-sharing and ride-sharing services. Currently, it operates on a wider range, including personal mobility devices such as electric bicycles and scooters. The purpose of this study is to derive a direction for providing future shared mobility services through analysis of factors affecting the usage intention of both current and prospective users. The survey targets 753 citizens living in Gyeonggi Province, Korea. The survey period is from February 12, 2020, to February 26, 2020. In this study, a logistic regression analysis is conducted to investigate the factors affecting the use intention of shared mobility. The analysis results show that gender, car ownership, and education, among variables reflecting socio-demographic characteristics, have significant effects on intention to use shared mobility. Moreover, we find that experience factors, including mainly used transportation modes, ownership of shared mobility device, past experience in similar services, satisfaction of existing shared mobility services, and distance from the home to the nearest bus stop, are also statistically influential. The analysis results are expected to lay the foundation for the introduction of shared mobility services and can be used as data for planning smart mobility services in the future.

1. Introduction

As negative impacts of autoownership have gradually increased, such as significant energy consumption, traffic congestion, inefficient land use, and excessive vehicle purchasing and operating costs, drivers have begun to look for alternatives [1]. Accordingly, shared mobility, a service that allows users to share various transportation modes and use them when necessary [2], has emerged as a major solution to existing transportation problems [1, 3–5]. This can be further divided into conventional automotive vehicle-sharing and personal mobility-sharing services in terms of types of vehicles to share [6]. The former involves car-sharing and ride-sharing services [7]: car sharing is when people share vehicles instead of using private vehicles (e.g., Drivy and Zipcar), often associated with a subscription fee, and ride sharing is when multiple passengers share their routes either partially or completely by a single vehicle (e.g., BlaBlaCar and Via).

Car sharing was introduced in Zurich, Germany, in 1948, as a car rental service, but it did not gain much popularity until in the late 1980s [1]. Besides this, personal mobility-sharing services were first introduced in Amsterdam, the Netherlands, in 1965, in a system that allowed people to share public bicycles; however, it was difficult to proceed with long-term planning as a result of bicycle damage or theft issues [8]. However, demand for alternative mobility has been continuously increasing according to its advantages (e.g., low cost, autonomy, flexibility, and rental in recent years) [9] and the disadvantages of owning personal cars (e.g., increased urban problems and uncertainty in future operation expenses) [1].

In recent years, with the successful introduction of shared mobility service in Europe thanks to advancements in telecommunication systems, many related studies have been carried out from different perspectives such as infrastructure planning for car-sharing and bicycle-sharing services [10, 11], estimation of benefits and impacts [12–16], social norms [17–19], infrastructure resiliency and influence of social networks [20–22], and user demand and characteristics analysis [23–25].

Especially, analysis of the user's intentions is attracting much attention in which it is closely related to future demand forecasting and economic evaluation. For example, Simsekoglu and Klöckner analyzed psychological attitudes and social determinants of electric bicycle use through structural equation modeling analysis [23]. Matyas and Kamargianni used a stated preference survey method to obtain initial insight into whether sharing modes can be used as a successful or feasible mobility tool [24]. Most recently, Ho et al. investigated user preferences for shared mobility. They identified the impact of current on-demand mobility services and obtained insight from customer demand analysis in terms of socio-economic conditions and travel needs [25]. However, such studies of user intention still have limitations in which they usually focused on a single specific shared mobility mode.

While the spread of shared mobility has increased, most user-related studies have analyzed factors influencing the use of shared transportation modes that are already widely used. In the analysis based on the shared mobility services already implemented, there is a possibility that established services did not fully reflect future user intention in the planning stages. Thus, such post-hoc analyses can be biased in terms of what potential users actually want and need, varying according to their individual situations. Thus, to predict the correct direction of the provision of a service based on the exact needs of users, it is necessary to analyze the intention to use a shared mobility system that has not been implemented yet.

To fill this research gap, we have conducted an online survey for citizens living in Gyeonggi Province, Korea, and carried out an ex-ante analysis of intention to use types of shared mobility services through a logistic regression model. The intended contributions of this study can be summarized threefold:

(1) We conduct an ex-ante analysis of potential user needs for shared mobility services to overcome the biases possibly inherent in the previous post-hoc analyses

(2) We identify factors that influence user needs based on their statistical significance and quantify the extent of influence using survey data collected from Gyeonggi Province, Korea

(3) We qualitatively suggest the direction of provision of shared mobility service in consideration of potential users' intention

The following section describes the questionnaire and sample characteristics used in this study. In addition, this section shows the statistics of the current usage status of various transportation modes. Section 2 presents the statistical method used to analyze the collected data. The results are documented Section 3. Section 4 summarizes the findings and suggests policy directions to enhance the effectiveness and efficiency of shared mobility plans for the future.

2. Data Description

2.1. Survey Overview and Questionnaire Design. An online survey has been conducted to analyze the characteristics and factors affecting the potential user's intention of shared mobility services. It targets people aged from 19 to 65 who live in Gyeonggi Province, Korea, and its survey period is from February 12, 2020, to February 26, 2020. This province is characterized by dense traffic zones that are satellite cities connected to Seoul City, the capital of Korea. Each of the province and the capital city is home to more than 10 million people, and there are a large number of people commuting to Seoul City from the surrounding satellite cities. While 781 responses were collected, we select a total of 753 samples due to the incompleteness of responses [26, 27]. In detail, samples are excluded in the following cases: (i) if there is no response to a question that needs to be answered and (ii) if multiple responses are made to a question that requires one answer.

The survey is comprised of three parts. The first part is a respondent socio-demographic characteristic (e.g., gender, age, job, household size, household income, and residential area) [28]. The second part is about the current and recent past use of conventional transportation modes and existing shared mobility services in the province. Specifically, the survey provides a detailed description of the shared mobility services (Table 1) and asks if the respondents are currently using or have used them and how satisfied they are to understand how their past experiences affect the future intention. The reason we consider only the past month's experiences is twofold. First, conventional transportation modes are mainly for commuting. Thus, the one-month survey can reflect the actual usage patterns. Second, the shared mobility services and their penetration rates are rapidly changing and still at a pilot phase, so the current perception of shared mobility services could not be highly influenced by old experiences obtained when the service environments were different from now. Since satisfaction is difficult to observe and measure explicitly, it is collected using the Likert scale [28, 29]. The third part is a future preference survey assuming that shared mobility services will be provided. The survey questionnaire is summarized in Table 2.

2.2. Sample Characteristics. An analysis of respondent characteristics confirms that the composition of the respondents reflects the entire socio-demographic characteristics of commuters in Korea. The sample numbers for different gender are almost even. Moreover, all age groups except for the 60s have almost the same sample number. As for the job, full-time workers occupy the majority of the respondents, so the collected sample reflects commuter traffic well [9]. Table 3 shows the statistical characteristics of the collected samples.

2.3. Usage Status of Transportation Modes and Shared Mobility Services. We survey the usage status of conventional transportation modes and shared mobility services, focusing on the experiences of respondents. We focus on the frequently used transportation modes in the past month of each

Table 1: Description of the shared mobility services.

Service	Concept	How to use
Car sharing	A short-term rental service for members	(1) Search for nearby parking lot using the smartphone app (2) Select and reserve a vehicle (3) Park in the parking lot after use
Ride sharing	A service that pays for and boards privately owned vehicles during rush hours	(1) Enter departure point, boarding time, and destination using the smartphone app (2) Take the vehicle and use it
Personal mobility sharing	A sharing service for single-person transportation modes powered by electric batteries	(1) Search for nearby electric bicycle or scooter using the smartphone app (2) Select and reserve an electric bicycle or scooter (3) Park freely after use

Table 2: Summary of survey questionnaire.

Part	Variables description
<Part 1> Sample socio-demographic factors	Gender, age, job, education, household size, household income, residence, etc.
<Part 2> Usage status and experiences of transportation and shared mobility service	Usage status of transportation in the past month Usage status of shared mobility service in the past month
<Part 3> Preference of shared mobility service	Intention to use the shared mobility

Table 3: Characteristics of the respondents.

Sample attributes		Number of the samples	%
Gender	Male	438	58.2
	Female	315	41.8
Age	20’s	169	22.4
	30’s	174	23.1
	40’s	176	23.4
	50’s	146	19.4
	60’s	88	11.7
Job	Professional/technical worker	156	20.7
	Administrative/office worker	381	50.6
	Service worker	56	7.4
	Production worker	38	5.0
	Self-employed worker	76	10.2
	Student	46	6.1
Education	High school	87	11.6
	University or higher	666	88.4
Household size	1	100	13.3
	2	119	15.8
	3	208	27.6
	4	265	35.2
	5 and above	61	8.1
Household income	Under 2,000,000 KRW*	53	7.0
	2,000,000~2,999,999 KRW	129	17.1
	3,000000~4,999,999 KRW	242	32.1
	5,000,000~6,999,999 KRW	176	23.4
	7,000,000~9,999,999 KRW	106	14.1
	10,000,000 KRW and above	47	6.2
Population	600,000 and above	290	38.5
	300,000~599,999	197	26.2
	Under 300,000	220	29.2
	County area	46	6.1

* denotes South Korean won.

respondent. Driving an owned vehicle is the most common mobility mode, followed by using conventional public transit such as subways, buses, and railroads. KTX presented in Table 4 is an abbreviation of the “Korea Train express” operated by Korail, a Korean railway company. The use of car-sharing and shared personal mobility services occupies around 1%, much lower than that of other conventional transportation modes. Thus, it means that this study focuses on people’s intentions who live in cities where shared mobility is not prevailing. In Table 4, the vehicle-sharing system, so-called macromobility, includes car sharing and ride sharing, and the personal mobility-sharing system, a component of micromobility, involves electric bicycle and scooter sharing.

We analyze the usage status and characteristics of respondents who have experienced shared mobility services more than once, which are documented in Tables 5 and 6, respectively. The results show car sharing has the highest number of once-experienced users, followed by shared electric bicycle, ride sharing, and shared electric scooter. For the purpose of usage, leisure/tourism has the highest proportion for each type of service, but the total usage for commuting is relatively low [30]. Furthermore, as can be intuitively expected, the travel distance and time per usage are longer in car sharing and ride sharing than personal mobility modes.

2.4. Analysis Method. Linear regression analysis reveals the correlation between one dependent variable and multiple independent variables if we assume that the dependent variable changes linearly by independent variables [31]. However, it is not suitable for dealing with binomial or discrete events. In contrast, logistic regression can analyze the nonlinear relationship between the dependent variable and independent variables that are binary or discrete. Therefore, it has the advantage of being able to understand the relationship between the binary dependent variable and several independent variables affecting various shared mobility choices [31]. For a shared mobility service, the willingness-to-use of a potential user, denoted by p, is mathematically expressed as equation (1). It is obviously a nonlinear function of n potentially influential factors indexed by i, $x_1, \ldots, x_n$, coupled with their coefficients, $\beta_0, \beta_1, \ldots \beta_n$. Even if the expression of p is nonlinear, the estimation of the coefficients can be efficiently done by transforming it to a linear form. The odds of p is $p/(1-p)$, and the logarithm of it is defined as the logit of p, $\text{Logit}(p)$. As shown in equation (2), $\text{Logit}(p)$ is a linear function of the inputs, so we can apply a linear regression to estimate the coefficients, $\beta_0, \beta_1, \ldots \beta_n$. In this paper, SPSS Statistics 25.0 software is used for the calculation:

$$p = \frac{e^{\left(\beta_0 + \sum \beta_i x_i\right)}}{1 + e^{\left(\beta_0 + \sum \beta_i x_i\right)}}, \tag{1}$$

$$\text{Logit}(p) = \log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i. \tag{2}$$

Table 4: Usage status of transportation as the primary mode.

Transportation modes	Number of the samples	%
Privately owned vehicle	333	44.2
Subway	193	25.6
City bus	134	17.8
Intercity bus	52	6.9
Shuttle bus	16	2.1
On foot	5	0.7
KTX	3	0.4
Vehicle-sharing system	9	1.2
Personal mobility-sharing system	8	1.1

In this study, the collected data about socio-demographic characteristics and past transportation usage statistics are considered as potential influential factors $(x_1, \ldots, x_n)$. The user intention of using a shared mobility is set as a dependent variable (p). Table 7 shows the input types of variables used in the model.

The detailed description of each variable is as follows: gender is divided into male and female; mainly, used transportation means whether you primarily use a vehicle or transportation when commuting; car ownership means you possess a vehicle that you can drive freely; ownership of shareable vehicle means respondents have at least one extra private vehicle that can be shared with other users; ownership of personal mobility device means respondents have electric bicycle or scooter; the education level is divided into twofold: high-school graduation and university or higher; previous experience refers to whether respondents have been used vehicle-sharing or personal mobility-sharing services; age is as natural number variables; distance from home to the nearest bus stop is as continuous variables in kilometers; satisfaction is measured on 5-point scale to show how satisfied the respondents were when they used the existing shared mobility services.

To identify variables associated with a significant correlation, we perform Pearson’s correlation analysis. Furthermore, a multicollinearity test has been conducted to check if the selected variables are free from multicollinearity with other variables. The analysis results through the correlation and the multicollinearity analysis are presented in Figure 1 and Table 8, respectively.

The titles of the rows and columns in Figure 1 means the following. Transportation is *mainly used transportation*; Car refers to *car ownership*; SV means *ownership of sharable vehicle*; PMD indicates *ownership of personal mobility device*; PEVS denotes *previous experience in vehicle sharing*; PEPMS represents *previous experience in mobility-sharing*; Distance stands for *distance from home to bus stop*.

Figure 1 shows the results of extracting significant variables related to user intention. In this figure, the overall result has a weak correlation, as the correlation is lower than or equal to 0.3 [32]. A few variables exceed 0.3, but we can consider them as moderate correlations since they are below 0.7 [32].

Table 5: Usage status of shared mobility services.

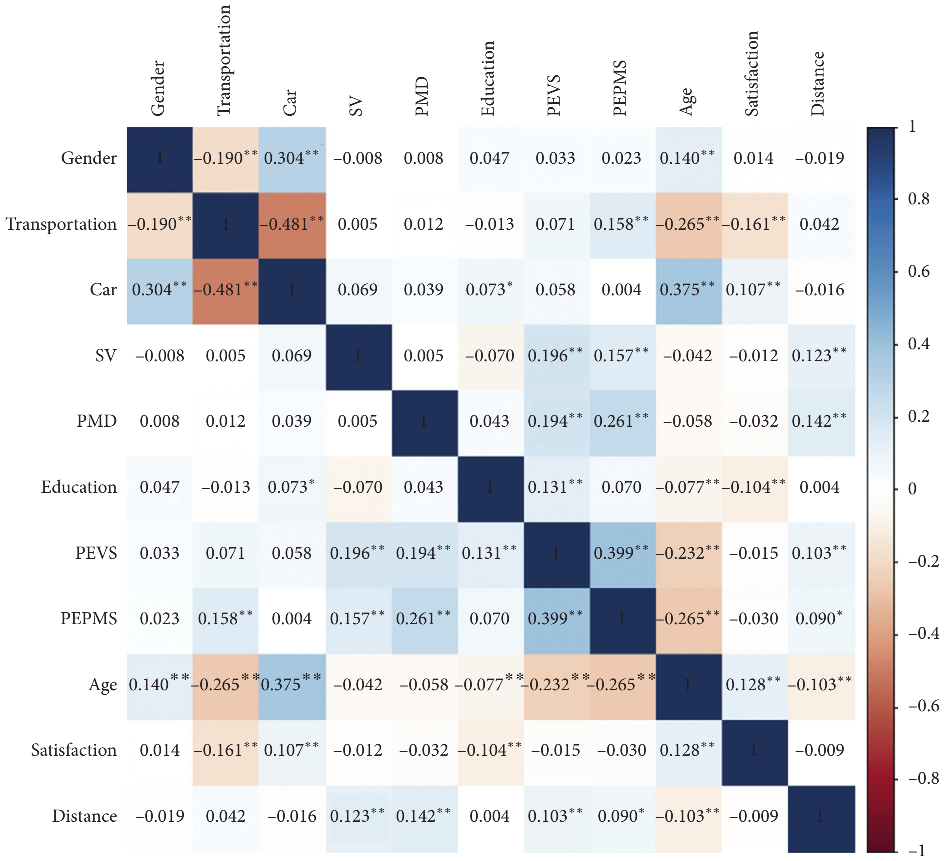
Sample attributes			Number of the samples	%
Usage	Car sharing	Yes	117	23.5
		No	576	76.5
	Ride sharing	Yes	72	9.6
		No	681	90.4
	Shared electric bicycle	Yes	139	18.5
		No	614	81.5
	Shared electric scooter	Yes	55	7.3
		No	698	92.7
Purpose of usage	Car sharing	Commuting	42	23.7
		Business	34	19.2
		Leisure/tourism	98	55.4
		Others	3	1.7
	Ride sharing	Commuting	20	27.8
		Business	17	23.6
		Leisure/tourism	35	48.6
	Shared electric bicycle	Commuting	37	26.6
		Business	17	12.2
		Leisure/tourism	84	60.4
		Others	1	0.7
	Shared electric scooter	Commuting	14	25.5
		Business	15	27.3
		Leisure/tourism	23	41.8
		Others	3	5.5

Table 6: Usage characteristics of shared mobility services.

Sample attributes	Average number of use	Distance per use (km)	Time per use (mins)
Car sharing	3.4	38.9	101.5
Ride sharing	3.2	16.3	34.4
Shared electric bicycle	3.2	10.6	38.4
Shared electric scooter	4.1	13.7	30.9

Table 7: Variable format of the estimation model.

Classification		Variables	Data format
Independent variable	Discrete variable	Gender	Female: 0 Male: 1
		Mainly used transportation	Vehicle: 0 Public transportation and others: 1
		Car ownership	No: 0 Yes: 1
		Ownership of shareable vehicle	No: 0 Yes: 1
		Ownership of personal mobility device	No: 0 Yes: 1
		Education	High school: 0 University or higher: 1
		Previous experience in vehicle sharing	No: 0 Yes: 1
		Previous experience in personal mobility sharing	No: 0 Yes: 1
	Continuous variable	Age	19~65
		Satisfaction	1(Unsatisfied)~5(Satisfied)
		Distance from home to the nearest bus stop	In kilometers
	Discrete variable	Intention to use the shared mobility	No: 0 Yes: 1



	Gender	Transportation	Car	SV	PMD	Education	PEVS	PEPMS	Age	Satisfaction	Distance
Gender	1	-0.190^{**}	0.304^{**}	−0.008	0.008	0.047	0.033	0.023	0.140^{**}	0.014	−0.019
Transportation	-0.190^{**}	1	-0.481^{**}	0.005	0.012	−0.013	0.071	0.158^{**}	-0.265^{**}	-0.161^{**}	0.042
Car	0.304^{**}	-0.481^{**}	1	0.069	0.039	0.073^{*}	0.058	0.004	0.375^{**}	0.107^{**}	−0.016
SV	−0.008	0.005	0.069	1	0.005	−0.070	0.196^{**}	0.157^{**}	−0.042	−0.012	0.123^{**}
PMD	0.008	0.012	0.039	0.005	1	0.043	0.194^{**}	0.261^{**}	−0.058	−0.032	0.142^{**}
Education	0.047	−0.013	0.073^{*}	−0.070	0.043	1	0.131^{**}	0.070	-0.077^{**}	-0.104^{**}	0.004
PEVS	0.033	0.071	0.058	0.196^{**}	0.194^{**}	0.131^{**}	1	0.399^{**}	-0.232^{**}	−0.015	0.103^{**}
PEPMS	0.023	0.158^{**}	0.004	0.157^{**}	0.261^{**}	0.070	0.399^{**}	1	-0.265^{**}	−0.030	0.090^{*}
Age	0.140^{**}	-0.265^{**}	0.375^{**}	−0.042	−0.058	-0.077^{**}	-0.232^{**}	-0.265^{**}	1	0.128^{**}	-0.103^{**}
Satisfaction	0.014	-0.161^{**}	0.107^{**}	−0.012	−0.032	-0.104^{**}	−0.015	−0.030	0.128^{**}	1	−0.009
Distance	−0.019	0.042	−0.016	0.123^{**}	0.142^{**}	0.004	0.103^{**}	0.090^{*}	-0.103^{**}	−0.009	1

$^{**}p < 0.01$, $^{**}p < 0.05$

Figure 1: Result of correlation.

For the variables with high significance shown in Figure 1, their VIF values are derived as shown in Table 8. VIF is a measure that calculates the association between a fixed independent variable and the other independent variables. If the VIF of a certain variable is 10 or more, we can understand that multicollinearity exists with the variable. Moreover, if this value exceeds 5, it is considered that attention is needed. In this case, it is inappropriate to put it into the model because the highly correlated variables can affect the model and result in undesirable biases [33]. As a result of the analysis, all estimated VIF values are lower than 5, so we find that multicollinearity does not exist.

A detailed description of each notation is in Table 1. *B* is a predicted value, meaning influence of the variable and Beta is the standardized value of *B*; S.E is the standard error, which estimates the variability; *t* is a value that is the difference between the predicted value divided by standard error, which compares the differences according to variability; Sig. judges whether it is valid within the significance level (95% confidence level in this study); allowance and variance inflation factor (VIF) are indicators of multicollinearity.

In general, the model is evaluated using R^2 of Cox and Snell and R^2 of Nagelkerke in regression analysis. However, for logistic regression, R^2 is generally low and depends on the dependent variable, so it is not appropriate to evaluate the adequacy of model [34, 35]. Thus, Hosmer and Lemeshow test, which is a goodness-of-fit method that performs a verification of the degree of agreement between a predicted value and an observed value using a chi-square distribution [34], is used to test the fit of the proposed model. If the results are greater than the set significance level, it can describe that the model is well estimated.

We set the final estimation model using the variables that are selected by using the backward elimination method, which has been known to be appropriate to prevent removing statistically meaningful variables related to the dependent variable compared to the alternative, the forward elimination method [36]. Table 9 shows the description and detailed verification results of the model through the Hosmer and Lemeshow test. The chi-square measure is estimated as 8.718, and the significance level is 0.367, which is greater than the standard, 0.05. Thus, the estimated model is statistically suitable to represent intention to use shared mobility services.

3. Discussion

Table 10 describes the model estimation results. The detailed description of notation is as follows: Wald means how important a variable is to describe a model; d.f is the degree

Table 8: Result of multicollinearity.

Variables	Nonstandard coefficient		Standard coefficient	t	Sig.	Collinearity value	
	B	S.E	Beta			Allowance	VIF
Constant	−0.121	0.111		1.082	0.280		
Gender	0.074	0.036	0.075	2.080	0.038	0.900	1.111
Mainly used transportation	0.097	0.039	0.099	2.462	0.014	0.727	1.376
Car ownership	0.103	0.048	0.091	2.129	0.034	0.631	1.584
Ownership of shareable vehicle	0.124	0.050	0.088	2.468	0.014	0.924	1.082
Ownership of personal mobility device	0.144	0.061	0.085	2.356	0.019	0.902	1.108
Education	0.131	0.054	0.085	2.431	0.015	0.950	1.053
Previous experience in vehicle sharing	0.107	0.042	0.097	2.510	0.012	0.774	1.293
Previous experience in personal mobility sharing	0.213	0.048	0.176	4.456	0.0001	0.750	1.334
Age	0.002	0.002	0.045	1.134	0.257	0.747	1.339
Satisfaction	0.016	0.017	0.032	0.927	0.354	0.953	1.050
Distance from home to the nearest bus stop	0.014	0.006	0.082	2.345	0.019	0.955	1.047

Table 9: Model verification.

Variables		Value
Model summary	Cox and Snell R^2	0.162
	Nagelkerke R^2	0.219
Hosmer and Lemeshow test	x^2	8.718
	Degree of freedom	8
	Significance	0.367

Table 10: Model estimation.

Variable	B	S.E	Wald	d.f	Sig.	Exp (B)	95% C.I for exp(B)	
							Lower	Upper
Constant	−0.643	0.430	2.235	1	0.135	0.526		
Gender	0.377	0.173	4.750	1	0.029	0.686	0.489	0.963
Mainly used transportation	0.514	0.196	6.883	1	0.009	1.671	1.139	2.453
Car ownership	0.589	0.224	6.925	1	0.008	1.802	1.162	2.794
Ownership of shareable vehicle	0.738	0.272	7.363	1	0.007	2.091	1.227	3.564
Ownership of personal mobility device	1.000	0.381	6.902	1	0.009	2.719	1.289	5.735
Education	0.608	0.258	5.541	1	0.019	1.837	1.107	3.048
Previous experience in vehicle sharing	0.468	0.212	4.884	1	0.027	1.598	1.054	2.421
Previous experience in personal mobility sharing	1.179	0.269	19.284	1	0.0003	3.252	1.921	5.504
Satisfaction			16.320	4	0.003			
Satisfaction (1)	−0.640	0.472	1.840	1	0.175	0.527	0.209	1.329
Satisfaction (2)	−0.323	0.364	0.786	1	0.375	0.724	0.354	1.479
Satisfaction (3)	−1.025	0.329	9.689	1	0.002	0.359	0.188	0.684
Satisfaction (4)	−0.521	0.331	2.467	1	0.116	0.594	0.310	1.138
Distance from home to the nearest bus stop	0.065	0.030	4.596	1	0.032	1.067	1.006	1.132

of freedom, which means the used amount of data information when estimating statistics; Exp(B) means how much influence a variable has; 95% C.I for exp(B) is a confidence interval in which a parameter is included.

As a result of model estimation, age, which is a respondent characteristic variable, is removed because it presents no significant effect on the intention to use shared mobility. In other words, people do not have heterogeneous preferences and use intentions to shared mobility services for different age groups. Gender ($B = 0.337$ and $p = 0.029$), mainly used transportation ($B = 0.514$ and $p = 0.009$), possession of a car ($B = 0.589$ and $p = 0.008$) or shared mobility ($B = 0.738$ and $p = 0.007$; $B = 1.000$ and $p = 0.009$, education ($B = 0.608$ and $p = 0.019$), thoughts on shared mobility ($B = 0.468$ and $p = 0.027$; $B = 1.179$ and $p = 0.0003$), satisfaction when using shared mobility ($p = 0.003$), and distance to the nearest bus stop from home ($B = 0.065$ and $p = 0.032$) are found to have significant effects. Especially, early adopters who have owned personal mobility devices ($B = 1.000$ and $p = 0.009$) or already experienced shared services ($B = 1.179$ and $p = 0.0003$) tend to think the future of shared mobility positively, as inferred from their highest B values.

In the case of gender, since the coefficient has a positive value ($B = 0.337$), it can be determined that men are more willing to use shared services than women. According to the sign of the coefficient, people have a higher intention to use shared mobility in the following cases: using public transportation as their primary mode, owning a private car or extra vehicle to share, with higher education level than high-school graduation, experienced shared mobility at least once, having high satisfaction with past usage of shared mobility services, and suffering from long distance from home to nearest bus stop. It is because people who mainly use public transportation will get more benefits by mitigating the accessibility-related inconvenience, and these benefits should be greater if the current accessibility is low. Last, those who own an extra car have a high intention because this vehicle can be used for vehicle sharing.

These findings can be summarized by the following. First, the experience of using or owning personal mobility devices, introduced relatively recently compared to other transportation modes, has the most positive effect on the user intention. Thus, to promote shared mobility services to public, it is considerable to offer free or cheap opportunities to use such personal mobility devices. Second, we find which groups of people are more willing to use shared mobility services, so this knowledge can be used in two ways: we need to (i) focus on providing services to the groups with higher intentions in order to increase usage rate or (ii) advertise the services to the groups with less intentions to increase the potential population of future users. Last, if there is an inconvenience in using existing public transportation modes, there is a high possibility of becoming a potential user of shared mobility. It means that we need to first introduce shared mobility services on the sites associated with low accessibility to public transportation systems.

4. Conclusions

We conduct an online survey on shared mobility and analyze factors affecting user intention. Correlation and multicollinearity analyses are performed to reveal hidden relations among measures of intention to use and potentially influential factors. This study aims to better understand the intent of using shared mobility for practitioners, which can be developed into various policy proposals for future mobility dissemination. Additionally, factors affecting future intention to use shared mobility services are quantitatively analyzed through logistic regression, which can estimate willingness to use according to individual characteristics and usage conditions.

For the factors reflecting socio-demographic characteristics, the analysis results show that gender, possession of a car, and education have significant effects on future intention, while age has no significant effect. Furthermore, for the indicators related to past transportation usage, mainly used transportation modes, possession of a shared mobility device, past related experiences, satisfaction when using shared mobility, and distance from home to the nearest bus stop are shown to have significant effects on intention to use.

Among them, previous experience is identified as one of the most important factors determining its intention to use. It appears that people who have a shareable vehicle or have experienced shared transportation have a high intention to use such systems. This also can affect the satisfaction of users with the use of shared mobility. Thus, citizens who have been exposed to shared mobility services in the past lower hurdles to use in the future. Past experiences and satisfaction can reduce uncertainty and anxiety about the introduction of new modes. It will raise expectations for the introduction of shared mobility services. Therefore, it is important to establish a social environment that can easily deliver the benefits of shared mobility to potential users when introducing new mobility.

They also have a high intention to use when they have been mainly using public transit services, and this intention becomes higher as the distance from their home to the nearest bus stop increases. This is another notable finding in this study. The expectations of people's future use of shared mobility are closely tied to the location of existing public transport facilities. Thus, if the installation of public transport is sparse or the distance between transfers of public transport is far, introducing shared mobility can increase the likelihood of choosing this. Therefore, policymakers may need to consider introducing shared mobility services when a number of citizens are in such situations.

It is possible to propose a more effective policy if we consider the analysis results of the user intention and the existing usage of transportation modes at the same time. We find that existing shared mobility is mainly used for leisure/tourism purposes rather than commuting. Therefore, in order to increase the utilization of shared mobility, introducing a service at a leisure or tourist complex may be considered. Moreover, it is found that vehicle-sharing mobility is more effective than personal mobility when trip distance is long. Likewise, personal mobility can be effectively introduced when the distance is relatively low, such as last-mile trips.

Coupled with the limitations of this study, we can consider some future research directions. This study focuses only on analyzing user intention in the future, assuming that shared mobility services are not currently settled and established. However, in many places in the world, services are in wide use, and their long-term usage and satisfaction statistics can be analyzed together to provide better services in the future. Moreover, to implement a shared mobility service in practice and suggest a policy, forecasting user intention and demand is not enough. A next step, a decision-making process, is needed to determine the scale and type of services. For this, both qualitative and quantitative methods can be adopted. As a qualitative method, it is possible to study cases of existing services based on the user demand, provided services, and user satisfaction. As a quantitative method, we can use an optimal decision-making framework to determine the most desirable scale and type of a newly introduced shared mobility service based on predicted demand and intentions. Last, in the current study, we identify which factors are actually influential, so it will be worth focusing on them for further detail. For example, education is shown to be significant in user intention. The group with a bachelor's degree or above, which accounts for 88.4% of respondents, can be subdivided into groups with Bachelor, Master, and Ph.D. degrees to see how much levels of higher education affect the awareness and use intention of various kinds of shared mobility services [37].

Data Availability

The data used to support the findings of this study is not publicly available according to the data security policy of Gyeonggi-do.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science, ICT and Future Planning (nos. 2020R1C1C1003296 and N01200545) and KAIST-KU joint Research Center, KAIST, South Korea.

Supplementary Materials

The supplementary description is as follows. Part 1: sample characteristics. Part 2: usage status of transportation and shared mobility service. Part 3: preference of shared mobility service. (*Supplementary Materials*)

References

[1] A. Millard-Ball, G. Murray, J. Ter Schure, C. Fox, and J. Burkhardt, "Car-sharing: where and how it succeeds," TCRP Report 104, Transit Cooperative Research Program (TCRP) Report, San Franciso, CA, USA, 2005.

[2] S. Shaheen, A. Cohen, and I. Zohdy, "Shared mobility: current practices and guiding principles," Report No. FHWA-HOP-16-022, University of California, Berkeley, CA, USA, 2016.

[3] S. A. Shaheen and A. P. Cohen, "Growth in worldwide carsharing," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1992, no. 1, pp. 81–89, 2007.

[4] B. Cohen and J. Kietzmann, "Ride on! mobility business models for the sharing economy," *Organization & Environment*, vol. 27, no. 3, pp. 279–296, 2015.

[5] P. DeMaio, "Bike-sharing: history, impacts, models of provision, and future," *Journal of Public Transportation*, vol. 12, no. 4, pp. 41–56, 2009.

[6] N. Davies, L. Blazejewski, and G. Sherriff, "The rise of micromobilities at tourism destinations," *Journal of Tourism Futures*, vol. 6, no. 3, p. 209, 2020.

[7] T. Kawaguchi, H. Murata, S. Fukushige, and H. Kobayashi, "Scenario analysis of car- and ride-sharing services based on life cycle simulation," *Procedia CIRP*, vol. 80, pp. 328–333, 2019.

[8] S. A. Shaheen, S. Guzman, and H. Zhang, "Bikesharing in Europe, the Americas, and Asia," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2143, no. 1, pp. 159–167, 2010.

[9] D. Efthymiou, C. Antoniou, and P. Waddell, "Factors affecting the adoption of vehicle sharing systems by young drivers," *Transport Policy*, vol. 29, pp. 64–73, 2013.

[10] G. Correia and J. M. Viegas, "Carpooling and carpool clubs: clarifying concepts and assessing value enhancement possibilities through a stated preference web survey in Lisbon, Portugal," *Transportation Research Part A: Policy and Practice*, vol. 45, no. 2, pp. 81–90, 2011.

[11] J. Firnkorn and M. Müller, "What will be the environmental effects of new free-floating car-sharing systems? the case of car2go in Ulm," *Ecological Economics*, vol. 70, no. 8, pp. 1519–1528, 2011.

[12] S. B. Kurth and T. C. Hood, "Car-pooling programs: solution to a problem?" *Transportation Research Record*, vol. 650, pp. 48–52, 1977.

[13] D. A. Prentice, D. T. Miller, and J. R. Lightdale, "Asymmetries in attachments to groups and to their members: distinguishing between common-identity and common-bond groups," *Personality and Social Psychology Bulletin*, vol. 20, no. 5, pp. 484–493, 1994.

[14] D. H. McKnight, L. L. Cummings, and N. L. Chervany, "Initial trust formation in new organizational relationships," *Academy of Management Review*, vol. 23, no. 3, pp. 473–490, 1998.

[15] S. Shaheen and C. Rodier, "Carsharing and carfree housing: predicted travel, emission, and economic benefits," in *Proceedings of the 83rd Annual Meeting of the Transportation Research Board*, Washington, DC, USA, January 2004.

[16] C. A. Rodier, "Review of the international modeling literature: transit, land use, and auto pricing strategies to reduce vehicle miles traveled and greenhouse gas emissions," in *Proceedings of the Transportation Research Board (TRB) Annual Meeting*, Washington, DC, USA, January 2009.

[17] G. J. Bruijn, S. P. Kremers, H. Schaalma, W. van Mechelen, and J. Brug, "Determinants of adolescent bicycle use for transportation and snacking behavior," *Preventive Medicine*, vol. 40, pp. 658–667, 2005.

[18] E. Heinen and S. Handy, "Similarities in attitudes and norms and the effect on bicycle commuting: evidence from the bicycle cities Davis and Delft," *International Journal of Sustainable Transportation*, vol. 6, no. 5, pp. 257–281, 2012.

[19] H. Sherwin, K. Chatterjee, and J. Jain, "An exploration of the importance of social influence in the decision to start bicycling in England," *Transportation Research Part A: Policy and Practice*, vol. 68, pp. 32–45, 2014.

[20] M. A. Ahmed, A. M. Sadri, and M. Hadi, "Modeling social network influence on hurricane evacuation decision consistency and sharing capacity," *Transportation Research Interdisciplinary Perspectives*, vol. 7, Article ID 100180, 2020.

[21] M. M. Mortula, M. A. Ahmed, A. M. Sadri, T. Ali, I. Ahmad, and A. Idris, "Improving resiliency of water supply system in arid regions: integrating centrality and hydraulic vulnerability," *Journal of Management in Engineering*, vol. 36, no. 5, Article ID 5020011, 2020.

[22] S. A. Morshed, M. Arafat, M. Ashraf Ahmed, and R. Saha, "Discovering the commuters' assessments on disaster resilience of transportation infrastructure," in *Proceedings of the International Conference on Transportation and Development 2020*, pp. 23–34, Seattle, WA, USA, August 2020.

[23] Ö. Simsekoglu and C. A. Klöckner, "The role of psychological and socio-demographical factors for electric bike use in

Norway," *International Journal of Sustainable Transportation*, vol. 13, no. 5, pp. 315–323, 2019.

[24] M. Matyas and M. Kamargianni, "The potential of mobility as a service bundles as a mobility management tool," *Transportation*, vol. 46, no. 5, pp. 1951–1968, 2018.

[25] C. Q. Ho, C. Mulley, and D. A. Hensher, "Public preferences for mobility as a service: insights from stated preference surveys," *Transportation Research Part A: Policy and Practice*, vol. 131, pp. 70–90, 2020.

[26] D. Dillman, J. D. Smyth, and M. Christian, *Internet, Mail and Mixed-Mode Surveys: The Tailored Design Method*, John Wiley & Sons, Inc., Hoboken, NJ, USA, 3rd edition, 2009.

[27] Y. Kwon, S. Son, and K. Jang, "Evaluation of incentive policies for electric vehicles: an experimental study on Jeju Island," *Transportation Research Part A: Policy and Practice*, vol. 116, pp. 404–412, 2018.

[28] Y. Kwon, S. Son, and K. Jang, "User satisfaction with battery electric vehicles in South Korea," *Transportation Research Part D: Transport and Environment*, vol. 82, Article ID 102306, 2020.

[29] R. A. Likert, "Technique for the measurement of attitudes," *Archiv für Psychologie (Frankf)*, vol. 55, p. 140, 1932.

[30] S. A. Shaheen, H. Zhang, E. Martin, and S. Guzman, "China's Hangzhou public bicycle," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2247, no. 1, p. 33, 2011.

[31] J. D. Cox and G. Pilcher, *Thermochemistry of Organic and Organometallic Compounds*, Academic Press, Cambridge, MA, USA, 1970.

[32] B. Ratner, "The correlation coefficient: its values range between +1/−1, or do they?" *Journal of Targeting, Measurement and Analysis for Marketing*, vol. 17, no. 2, pp. 139–142, 2009.

[33] I. Lee, *Easy Flow Regression Analysis*, HanNaRae, Seoul, South Korea, 2014.

[34] D. Hosmer and S. Lemshow, *Applied Logistic Regression*, A Wiley-Interscience Publication, Hoboken, NJ, USA, 2nd edition, 2000.

[35] J. Cohen, P. Cohen, S. West, and L. S. Aiken, *Applied Multiple Regression/Correlation Analysis for the Behavioural Sciences*, Lawrence Erlbaum Associates, Manhwah, NJ, USA, 3rd edition, 2003.

[36] R. Christensen, *Log-Linear Models and Logistic Regression*, Springer-Verlag, Berlin, Germany, 1997.

[37] S. Shaheen, A. Cohen, N. Chan, and A. Bansal, "Sharing strategies: carsharing, shared micromobility (bikesharing and scooter sharing)," in *Transportation Network Companies, Microtransit, and Other Innovative Mobility Modes*, E. Deakin, Ed., ResearchGate, Boston, MA, USA, 2020.

Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 6685004, 13 pages
https://doi.org/10.1155/2021/6685004

Research Article

Analysis of Travel Mode Choice in Seoul Using an Interpretable Machine Learning Approach

Eui-Jin Kim

Institute of Construction and Environmental Engineering, Seoul National University, Seoul 08826, Republic of Korea

Correspondence should be addressed to Eui-Jin Kim; kyjcwal@snu.ac.kr

Received 31 December 2020; Revised 19 February 2021; Accepted 22 February 2021; Published 2 March 2021

Academic Editor: Inhi Kim

Understanding choice behavior regarding travel mode is essential in forecasting travel demand. Machine learning (ML) approaches have been proposed to model mode choice behavior, and their usefulness for predicting performance has been reported. However, due to the black-box nature of ML, it is difficult to determine a suitable explanation for the relationship between the input and output variables. This paper proposes an interpretable ML approach to improve the interpretability (i.e., the degree of understanding the cause of decisions) of ML concerning travel mode choice modeling. This approach applied to national household travel survey data in Seoul. First, extreme gradient boosting (XGB) was applied to travel mode choice modeling, and the XGB outperformed the other ML models. Variable importance, variable interaction, and accumulated local effects (ALE) were measured to interpret the prediction of the best-performing XGB. The results of variable importance and interaction indicated that the correlated trip- and tour-related variables significantly influence predicting travel mode choice by the main and cross effects between them. Age and number of trips on tour were also shown to be an important variable in choosing travel mode. ALE measured the main effect of variables that have a nonlinear relation to choice probability, which cannot be observed in the conventional multinomial logit model. This information can provide interesting behavioral insights on urban mobility.

1. Introduction

The recent emergence of new travel modes such as ride-sourcing, ride-hailing, and autonomous vehicles and the evolution of new mobility services such as mobility as a service and mobility on demand (known as MaaS and MoD, respectively) is changing travel behavior significantly [1]. These emerging technologies present new sources of big data for understanding travel behavior and system performance [2]. New methods that leverage this big data are needed to analyze travel behavior changes and predict travel mode choices. The multinomial logit (MNL) model has dominated travel mode choice analysis due to its simplicity and readability. The simple MNL model and its variants have been applied to consider various effects in the context of travel mode choice based on the expert-designed model assumptions. Linear relationships in parameters of the simple MNL model can be intuitively interpreted as weights of the variables. Even nonlinear relationships in parameters such as willingness-to-pay for reduced travel time variability can be captured by combining the conventional utility functional form with a probability weighting function [3]. However, this approach requires prior assumptions for the functional form of the weighting function. The MNL can capture the interaction effects between correlated variables by adding appropriate interaction parameters that are based on empirical or experimental knowledge [4], but considering all of the interactions becomes impossible as the number of variables increases. Although the simple MNL model assumes the independence of irrelevant alternatives (IIA) causing misleading predictions, the correlations between travel modes have been addressed by the advanced structure of the MNL model such as the nested logit and mixed logit model [5]. However, it is very difficult to design an appropriate model structure of the MNL model that effectively captures a high degree of complexity in a dataset [6]. In summary, the existing MNL and its variants can take into account the various effects in the mode choice situations;

however, they rely on the model assumptions that should be determined by the subjective judgment of the researcher, and these assumptions affect the parameter estimates and the prediction performance.

Machine learning (ML) approaches are promising alternatives to the MNL-based model for modeling travel mode choice. It can represent complex relationships between mode choices and input variables in a data-driven manner rather than making strict assumptions about the data [7]. Many previous studies have reported the use of an ML approach to model travel mode choice [1, 6–11]. These authors have generally reported improvements in the prediction performance of ML approaches compared to MNL-based models. Recently, Wang et al. established an empirical benchmark by using 86 ML models to predict travel mode choice based on a 2017 U.S. national household travel survey dataset [12]. The authors found that ensemble models such as boosting, bagging, and random forest models exhibit performances superior to those of all other ML methods, including deep neural networks. However, due to the black-box nature of ML models, the authors could not explain the prediction results, making it difficult to find a suitable explanation for the relationship between the input variables and travel mode choices.

Several studies have performed additional analyses of the prediction results to complement the evaluation of performance. Wang and Ross proposed an extreme gradient boosting (XGB) model for predicting travel mode choice [1]. Using a relatively comprehensive dataset, the authors measured the relative importance of variables in the training process of the XGB and estimated the importance of correlated variables that cannot be explained using the MNL model. Hagenauer and Helbich measured the permutation-based importance of variables in predicting the choice of each travel mode, and their result showed that the critical variables varied with the predicted travel modes [7]. Lee et al. developed a choice model for alternatives related to autonomous vehicles using a gradient boosting machine (GBM) [10]. They measured the partial dependence (PD), which captures the marginal effects of attributes representing the relationship between the input variables and predicted output. Although the above researchers who conducted these three studies tried to explain the prediction results of their ML models with several meaningful interpretations, there is room for improvement by the application of various interpretation methods to reveal details of the characteristics of travel behavior.

In this study, model-agnostic interpretation methods were applied to explain the prediction results of ML models concerning mode choice behavior. XGB, random forest (RF), and artificial neural network (ANN) models were employed to predict travel mode choices from national household travel survey (NHTS) data in Seoul. Trip- and tour-related attributes were extracted from the NHTS data to construct the variable set. The tour refers to interconnected trips (i.e., trip chain) during a day. This dataset is enriched with traffic analysis zone (TAZ)-level spatial information. The performance of the models was evaluated regarding their prediction of each travel mode. Then, the best-performed XGB prediction results were analyzed to reveal choice behavior for urban travel modes. In doing so, two crucial issues were addressed, which are difficult to investigate using a conventional MNL model, i.e., (i) how each variable interacted with other variables and (ii) how the variable related to the probability of travel mode choice.

The remainder of this paper is organized as follows. In Section 2, the dataset and data-processing procedure applied in this study are described. Then, the ML models and model-agnostic interpretation methods are discussed in detail. In Section 3, performance evaluation of the ML models and interpretation of the XGB prediction results are presented. Finally, concluding remarks and future research directions are presented in Section 4.

2. Materials and Methods

2.1. Data Descriptions. The primary source of data for this study was a 2016 NHTS dataset in the Seoul, Korea [13]. These data included individual travel diaries that recorded every daily trip taken, with multiple trips on a given day expressed as a trip chain. The chained trips were divided by their trip purpose and established the major travel modes of the trip's purpose. For example, a person who uses the subway to go to work must first access the subway station on foot and then use the subway. In this case, the two chained trips, walking and subway, are combined into one subway trip as the primary travel mode. Walking is considered a primary travel mode only if it is used as the sole travel mode, but not as a means to access another travel mode. Seoul operates a public transit unified fare system for buses and subways, whereby charges are levied as if the person is using a single travel mode when transferring between these two forms of public transit. Therefore, this study makes no distinction between a bus and a subway, whereby the chained trips of a bus and subway with a transfer are considered to be one trip by public transit.

Table 1 describes the variables included in the travel mode choice model. Four categories of variables are used to train and test the mode choice model. Trip-related, tour-related, and individual attributes are extracted from the NHTS data, and built environment attributes are obtained from national spatial data [14] and population census [15] in Korea. The departure and arrival locations of NHTS data are recorded in the TAZ unit, which is within a radius of about 1 km; thereby, the NHTS data are merged with built environment attributes according to TAZ. The dependent variable is for primary travel modes: car, bike, transit, and walking. A single mode, which is assumed, is used for an entire tour because 89.9% of the respondents in the NHTS data used the one primary travel mode rather than a combination of modes. Trip-related attributes are extracted from single or sequential individual trips. The duration of an activity is calculated by the difference between the arrival time on the previous trip and the departure time on the next trip. The duration of activity on the last trip (i.e., the return trip home) is calculated by the difference between the arrival time of the last trip and the departure time of the first trip. Travel time includes in-vehicle and out-of-vehicle time, such

Table 1: Description of the independent and dependent variables.

Variable name	Explanation	Data types
Travel mode	Chosen travel mode for the trip (dependent variable): 1 = car, 2 = bike, 3 = transit, and 4 = walking	Categorical
Trip-related attributes		
Activity duration	Duration of the activity	Numeric
Travel time	Travel time of the trip	Numeric
Departure time	1 = the trip occurs in the morning or evening peak hours (8 A.M.–10 A.M. or 5 P.M.–7 P.M.); 0 = otherwise	Dummy
Trip type	Context of the trip: 1 = home-based work (HBW); 2 = home-based others (HBO); 3 = non-home-based others (NHBO); 4 = return home (RH)	Categorical
Tour-related attributes		
Sum of activity duration	Sum of activity duration during a day excluding the last trip	Numeric
Sum of travel time	Sum of travel time during a day	Numeric
Number of trips	Number of trips that occurred during a day	Categorical
Tour type	Context of the tour: 1 = home-work-home (HWH); 2 = home-other-home (HOH); 3 = home-work-other-home (HOWH)	Categorical
Individual attributes		
Age	Age of the traveller in years	Numeric
Gender	1 = the traveller is male; 2 = the traveller is female	Dummy
Car owner	1 = the household of traveller owns a car; 0 = otherwise	Dummy
Driver's license	1 = the traveller has a driver's license; 0 = otherwise	Dummy
Income	Monthly household income of the traveller (million KRW): low = income < 5; high = income ≥ 5	Dummy
Built environment attributes		
Land use in D: residential	The ratio of residential area to the total area at D in TAZ unit	Numeric
Land use in D: commercial	The ratio of commercial area to the total area at D in the TAZ unit.	Numeric
Population density at D	Density of the population (people/km^2) at the destination in the TAZ unit	Numeric
Number of workers at D	Number of workers at the destination in the TAZ unit	Numeric
Number of bus stops at D	Number of bus stops at the destination in the TAZ unit	Numeric
Number of subway stops at D	Number of subway stops at the destination in the TAZ unit	Numeric

Note. D = destination of a trip; 1,000 KRW = 0.84 USD.

as waiting time and access time. Departure time is divided into peak and nonpeak categories. Trip type is defined by the characteristics of the origin and destination, such as home, work, or other places. Tour-related attributes are extracted from all the trips of individuals during a single day. The sum of activity durations of trips is calculated, excluding the last trip, and the sum of travel time and the number of trips includes all the trips made during a day. Tour types are defined by the combination of trip types included in a tour. The Home-Other-Home (HOH) type includes the tour with more than three trips (e.g., H-O-O-H). Individual attributes include age, gender, car owner, driver's license, and income, and all of those attributes are directly collected in the NHTS data. Built environment attributes describe the spatial characteristics of a trip's destination (D). The variables for land use are defined as the ratio of a residential or commercial area to the total area. Population density, number of workers, number of bus stops, and number of subway stops are also used to characterize the destination in the TAZ unit. Although travel cost is an important variable in the travel mode choice, the NHTS data used in this study did not include the respondents' travel cost such as fuel cost, parking cost, and transit fares. Therefore, the effect of travel cost does not consider in the analysis like other studies using the NHTS data [1, 7, 8]. After a data-cleaning process, in which the trips were removed with very long activity duration and travel time, a total of 172,889 trips taken by 76,190 individuals were used. 75% of the NHTS data was used for training and 25% of those data for the test.

Table 2 shows the descriptive statistics of the variables. The distribution of the travel mode is imbalanced in that trips by walking, transit, car, and bike are 43.7%, 35.3%, 18.5%, and 2.5%, respectively. The mean activity duration is 490.2 minutes, which is slightly longer than the standard working time of eight hours, and the mean travel time of each trip is 21.7 minutes. The number of trips during a peak time is comparable to the number of trips at a nonpeak time. In terms of trip type, the percentage of HBW, HBO, NHBO, and RH are 31.8%, 16.7%, 4.8%, and 46.7%, respectively, indicating that more than 20% of noncommuting trips are included in the data. The sum of activity duration and the sum of travel time have a mean value of 509.2 minutes and 51.6 minutes, respectively. While 70.9% of travelers make two trips during a day, 29.1% make more than three trips. The people who made more than three trips may have tour types of HOH or HOWH, which are 27.0% and 21.4% of total tours, respectively. The percentages of females, car owners, driver's licenses, and those with a high income are

Table 2: Descriptive statistics of the variables.

Variable name	Category	%	Mean	Standard deviation
Travel mode	Car	18.5		
	Bike	2.5		
	Transit	35.3		
	Walking	43.7		
Trip-related attributes				
Activity duration (min)			490.2	251.1
Travel time (min)			21.7	15.9
Departure time	Peak	50.6		
	Nonpeak	49.4		
Trip type	HBW	31.8		
	HBO	16.7		
	NHBO	4.8		
	RH	46.7		
Tour-related attributes				
Sum of activity duration (min)			509.2	235.6
Sum of travel time (min)			51.6	33.4
Number of trips	2	70.9		
	3	10.5		
	4	16.4		
	5	1.4		
	6	0.8		
Tour type	HWH	51.6		
	HOH	27.0		
	HOWH	21.4		
Individual attributes				
Age			44.6	20.0
Gender	Female	51.7		
	Male	48.3		
Car owner	Yes	72.0		
	No	28.0		
Driver's license	Yes	54.7		
	No	45.3		
Income	High	33.0		
	Low	67.0		
Built environment attribute				
Land use in D: residential			0.49	0.20
Land use in D: commercial			0.29	0.20
Population density at D			42,862	11,771
Number of workers at D			32,787	75,271
Number of bus stops at D			125.2	85.5
Number of subway stops at D			1.0	1.2

Note. D = destination of a trip.

51.7%, 72.0%, 54.7%, and 33.0%, respectively. While the car owner indicates whether the household owns a private car, the driver's license indicates whether the individual owns a driver's license. The descriptive statistics of built environment attributes are also presented in Table 2.

2.2. Machine Learning Model for Predicting Travel Mode Choice. Three ML models, XGB, RF, and ANN, were applied to predict travel mode choices. Given a set of values of the input variable, the model predicts the probability that a specific travel mode will be chosen. To account for class imbalance, weight to the data instance is applied in inverse proportion to the frequency distribution of each class, and those class-specific weights are commonly used to train ML models. A hyperparameter is a parameter that controls the training process of the ML model. Since the hyperparameter affects the speed and quality of the training process, hyperparameter tuning is an essential task for evaluating an ML model's performance. The major hyperparameters of each ML model were tuned using a grid search technique based on 4-fold cross-validation. A comparable degree of a set of hyperparameter combinations is considered for each ML model.

2.2.1. Random Forest. The decision tree is a popular ML model due to its ability to capture complex structures in the data, although it suffers from an overfitting problem. To address this issue, ensemble models have been proposed. The RF [16] is a tree-based ensemble method related to the bagging approach, which averages noisy but approximately unbiased models to reduce the variance. An ensemble of independent trees on a random subset of a training dataset with randomly selected variables can achieve better generalized performance [9, 17]. The RF has also shown promising performance for predicting travel mode choice in previous studies [7, 8]. There are four significant hyperparameters used to tune the learning process of an RF model: the number of trees, the number of variables to split in each node, the maximum depth of each tree, which determines the model complexity of each tree, and the data-sampling rate used for training each tree. The RF model is implemented using the "ranger" package in R [18].

2.2.2. Extreme Gradient Boosting Model. The GBM is another tree-based ensemble method that has been successfully used to predict travel mode choice [1, 10]. Unlike the RF, the GBM builds a sequence of the low-depth decision tree, where each tree is trained to put more weight on the incorrect prediction of the previous trees [19]. The results of all the estimated trees collectively determine the result of the ensemble model. To implement GBM, an eXtreme Gradient Boost (XGB) proposed by Chen et al. [20] is employed. XGB is an efficient algorithm for constructing boosted trees using regularization terms and parallel processing. The five major hyper parameters of XGB are tuned, including the learning rate, maximum depth of each tree, number of variables considered in each tree, number of samples considered in each tree, and minimum value of the sum of instance weight of a node. The XGB model is implemented using the "xgboost" package in R [20].

2.2.3. Artificial Neural Network. The ANN is a widely used ML model for the training classification model. The promising performance of ANN rather than MNL for modeling travel mode choice has been reported in previous studies [6, 7]. A multilayer perceptron (MLP) is a conventional neural network including an input layer, one or more hidden layers, and an output layer. Nonlinear relationships in the data can be naturally captured by the MLP since it iteratively adjusts the weights and biases between neurons' interactions in multiple layers [21]. This study adopts an MLP with a single hidden layer, and a standard backpropagation algorithm with a decay term was used to train the MLP. The number of neurons in the hidden layer and a decay term are tuned. The ANN model is implemented using the "nnet" package in R [22].

2.3. Model-Agnostic Interpretation Methods. Interpretability is defined as the degree of understanding the cause of prediction [23]. Traditional interpretable models, such as logistic regression and decision tree, sacrifice prediction performance due to a simple model structure that improves interpretability. Recently, model-agnostic interpretation methods have been applied to make machine learning interpretable. Those interpretation methods commonly measure changes in prediction performance according to changes in the value of input variables. By doing so, the marginal effect of the variables is estimated to deduce the importance and interaction of variables. Also, the complex relationship between the input and outcome can be estimated. The target of the interpretation methods is divided into two perspectives: the entire model behavior (i.e., global interpretability) and a single prediction (i.e., local interpretability) [24]. This study focuses on the former by applying three model-agnostic interpretation methods.

2.3.1. Permutation-Based Variable Importance. When values of a variable are permutated so that their relationship with the predicted outcome is broken, the prediction error will increase. By calculating the increases in the model's prediction error, the importance of the variable is obtained. This study measures the importance based on the algorithm proposed by Fisher et al. [25]. The permutation-based variable importance can naturally consider all interactions with other variables (i.e., the sum of main and cross effects) by permutation. Therefore, highly correlated variables also can be directly interpreted. For the input variable matrix $\mathbf{X}$, the original error (e^{orig}) of the ML model ($\hat{f}$) is estimated by the defined loss function (L) between the predicted value ($\hat{f}(\mathbf{X})$) and the true value (y), as in equation (1). Then, the input matrix, including the permutated variable j ($\mathbf{X}^{\text{perm}_j}$) is used to compute the permutated error (e^{perm_j}), and the importance of variable j (VIMP_j) is calculated by ($e^{\text{perm}_j}/e^{\text{orig}}$), as shown in equation (2):

$$e^{\text{orig}} = L\left(y, \hat{f}(\mathbf{X})\right), \tag{1}$$

$$\text{VIMP}_j = \frac{L\left(y, \hat{f}\left(\mathbf{X}^{\text{perm}_j}\right)\right)}{e^{\text{orig}}} = \frac{e^{\text{perm}_j}}{e^{\text{orig}}}. \tag{2}$$

To measure the importance of the multiclass classification, the balanced accuracy of each travel mode (see equation (3)) is used as a L between the predicted value and the true value:

$$\begin{aligned} \text{Specificity} &= \frac{\text{TN}}{\text{TN} + \text{FP}}, \\ \text{Sensitivity} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{Balanced accuracy} &= \frac{\text{Specificity} + \text{Sensitivity}}{2}, \end{aligned} \tag{3}$$

where TN, FN, TP, and FP are the true negative, false negative, true positive, and false positive, respectively. Compared with the accuracy, the balanced accuracy can serve as a better judge of performance for the imbalanced classification problem where the difference in the number of negative and positive samples for each class is large [26]. The balanced accuracy in this study also measures the prediction performance of the ML model.

2.3.2. Variable Interaction. When variables are correlated, the effect of one variable depends on the value of other variables. The change in the prediction error also can be used to measure those correlations (i.e., variable interaction). Friedman's *H*-statistic is used to estimate the strength of the variable interaction quantitatively. This measurement indicates how much the variation in the prediction depends on the interaction of the variables [27]. The marginal effect of a variable on the model's prediction is represented by the partial dependence (PD) function, as in

$$\begin{aligned} \mathrm{PD}_j\left(x_j\right) &= \frac{1}{n}\sum_{i=1}^{n}\hat{f}\left(x_j, x_{-j}^{(i)}\right), \\ \mathrm{PD}_{jk}\left(x_j, x_k\right) &= \frac{1}{n}\sum_{i=1}^{n}\hat{f}\left(x_j, x_k, x_{-j-k}^{(i)}\right), \end{aligned} \tag{4}$$

where $\mathrm{PD}_j\,(x_j)$ is the PD function of a single variable j, $\mathrm{PD}_{jk}\,(x_j, x_k)$ is the 2-way PD function of two variables j and k, n is the total number of data points, i is a certain data point used to estimate the marginal effect, x_j and x_k are the variables used to calculate the marginal effects, and x_{-j} and x_{-j-k} are the other variables used in the ML model $(\hat{f})$. Mathematically, the interaction between variables j and k (i.e., two-way interaction) is estimated as in equation (5), and the interaction between variable j and any other variables (i.e., total interaction) is estimated as in equation (6) [28]:

$$H_{jk}^2 = \frac{\sum_{i=1}^{n}\left[\mathrm{PD}_{jk}\left(x_j^{(i)}, x_k^{(i)}\right) - \mathrm{PD}_j\left(x_j^{(i)}\right) - \mathrm{PD}_k\left(x_k^{(i)}\right)\right]^2}{\sum_{i=1}^{n}\mathrm{PD}_{jk}^2\left(x_j^{(i)}, x_k^{(i)}\right)}, \tag{5}$$

$$H_j^2 = \frac{\sum_{i=1}^{n}\left[\hat{f}\left(x^{(i)}\right) - \mathrm{PD}_j\left(x_j^{(i)}\right) - \mathrm{PD}_{-j}\left(x_{-j}^{(i)}\right)\right]^2}{\sum_{i=1}^{n}\hat{f}^2\left(x^{(i)}\right)}, \tag{6}$$

where $\mathrm{PD}_{-j}\,(x_{-j}^{(i)})$ is the PD function that depends on all variables except the jth variable. While the two-way interaction in equation (5) indicates the amount of the variance explained by the interaction between the two variables x_j and x_k among the variance of the output of the PD, the total interaction in equation (6) indicates the amount of the variance explained by the interaction between variables x_j and any other variable x_{-j} among the variance of the output of the entire function [28]. Therefore, if the *H*-statistic is zero, there is no interaction at all, and if all the effect of variables is applied as an interaction, the statistic would be one. When the *H*-statistic is larger than one, the interpretation would be difficult. In the case of two-way interaction, this can happen when the variance of two-way interaction is larger than the variance of the two-dimensional PD In the case of total interaction, this can happen when the variance of interaction between one variable and other variables is larger than the variance of the ML model.

2.3.3. Accumulated Local Effect. The promising performance of the ML model suggests that complex relationships exist between the input variables and predicted outcome in the real data, which may be nonlinear or polynomial. To represent these relationships, the ALE value was used, which shows the changes in the probability of a travel mode choice by the specific value (or category) of a variable. Generally, the marginal effect of the variables can be obtained using the PD function [10, 17]. However, the PD function assumes that the variables are not correlated with each other, which is unrealistic in real data. When the variables are highly correlated, the PD function includes unrealistic data when averaging the prediction results, which can substantially bias the estimated effect of the variable [28]. To address this issue, the accumulated local effect (ALE) is used, which is the unbiased alternative to PD [29]. The value of ALE can be interpreted as the main effect of the variable at a specific value compared to the average prediction value of the data. The ALE plots can depict any relationship, whether linear, monotonic, or more complex, between a variable and the predicted outcome. The ALE calculates the change in prediction results by replacing the target variable with grid values *z*. The average change in prediction is the effect for a specific interval, and its effect accumulates across all intervals as [29]

$$\hat{f}_{j,\mathrm{ALE}}\left(x_j\right) = \sum_{k=1}^{k_j^K(x)}\frac{1}{n_j(k)}\sum_{i:\,x_j^{(i)}\in N_j(k)}\left[f\left(z_{k,j}^K, x_{-j}^{(i)}\right) - f\left(z_{k-1,j}^K, x_{-j}^{(i)}\right)\right], \tag{7}$$

where $z_{k,j}^K$ is the partition of the minimum and maximum of x_j into K interval and $k_j^K(x) = k$ if $x \in (z_{k-1}^K, z_{k,j}^K]$, the average effects of all instances within an interval $(N_j\,(k))$ are calculated by dividing the sum of the difference of the prediction, i.e., $\sum_{i:\,x_j^{(i)}\in N_j(k)}[f\,(z_{k,j}^K, x_{-j}^{(i)}) - f\,(z_{k-1,j}^K, x_{-j}^{(i)})]$, by the number of instances in this interval $(n_j\,(k))$. The ALE is centered on having a zero mean, as shown in

$$\hat{f}_{j,\mathrm{ALE,cent}}\left(x_j\right) = \hat{f}_{j,\mathrm{ALE}}\left(x_j\right) - \frac{1}{n}\sum_{i=1}^{n}\hat{f}_{j,\mathrm{ALE}}\left(x_j^{(i)}\right). \tag{8}$$

While the intervals can be defined by the distribution of the numeric variables, the intervals for the categorical variables are determined by the similarity of categories since the categorical variables do not have a natural order. The similarity of the two categories is calculated by the sum of distances over the other variables. While the distance between the target category and other numeric variables is calculated by Kolmogorov–Smirnov distance, the distance between target category and other categorical variables is calculated by the relative frequency tables. More details are described in [28].

3. Results and Discussion

3.1. Prediction Performance. Since the travel modes are imbalanced, the prediction performance of the RF, XGB, and ANN models are evaluated using three metrics: specificity, sensitivity, and balanced accuracy, as shown in equation (3). Table 3 compares the prediction performances of the three models. Overall, the RF and XGB models exhibit better performance than the ANN model. Although class-specific weight was applied for training the ML models, all models show poor performance for the prediction of bike choice that is minority class (i.e., 2.5% of total). The performance of the XGB is comparable to that of RF and exhibited better performance for some travel modes and metrics. Compared with the RF, XGB shows slightly lower performance for predicting the choices of car and bike but shows better performance for predicting the choice of transit and walking. For all travel modes, the XGB shows the best performance for all metrics.

The number of FN explains the low sensitivity of the XGB for minor classes (i.e., car and bike). For example, in the case of car, the number of FN is 2,635, including 1,489 transit, followed by 1,111 walking and 35 bike. This result indicates that consideration of trip- and tour-related attributes cannot successfully identify the choice of car and public transit. This may be because the competitiveness of public transit (i.e., relative travel time for given OD) in Seoul is as high as that of cars [30]. The FN caused by walking indicates that car and walking share some travel characteristics. This result can be explained by travel patterns in Seoul where short-distance driving (i.e., trips of 5 km or less) represent 44% of all car driving [31]. The short-distance driving can indicate similar travel time to walking trip. In the case of bike, the number of FN is 754, including 401 walking, 219 transit, and 134 car. It also indicates that the travel characteristics of walking are similar to those of bike, such as travel time and trip type. To develop an understanding of mode choice behavior, the prediction results of the best-performing XGB model were analyzed using three model-agnostic interpretation methods in the following section.

3.2. Variable Importance. The permutation-based variable importance was measured based on the XGB model. Since decision makers have different objectives and application plans for each travel mode, the importance was measured for each travel mode. Figure 1 shows box plots of the importance of the top ten variables for each travel mode, which was calculated from 50 simulations to consider the randomness introduced by the permutation. Since this importance considers both the main and cross effects of a variable, it cannot be interpreted as the main effect of variables like the coefficient of MNL.

Although some variables are commonly important in predicting all mode choice, the ranking of other variables is somewhat different. Travel time and activity duration are important for all travel modes, and their influence is more significant on a tour level than on a trip level. The result can explain the recent success of the tour-based model in travel demand forecasting, compared with the trip-based model [32, 33]. While age, travel time, and activity duration commonly rank highly in importance among all travel modes, car owner, land use, and number of trips only influence a specific travel mode. This implies that policy-making needs to be carried out by focusing on different factors for each travel mode, based on the mode-specific analysis.

Regarding car, age is the most important variable in determining choice, which may indicate the varying preference for comfort and value of time by age [34]. Car ownership, of course, is the second important variable for the choice of a car. Two tour-related attributes, the sum of travel time and the sum of activity duration, are more critical than two corresponding trip-related attributes, travel time and activity duration.

Regarding bike, the small number of positive samples of bike results in a higher variance of importance than other travel modes. Low performance of the XGB may cause those variances, and the proposed box plot is useful in the case of those high variances. Similar to car, the age, sum of travel time, and sum of activity duration rank highly in terms of importance for bike, followed by gender. Unlike other travel modes, two land-use variables show considerable importance, indicating that land use affecting accessibility and mobility would influence the use of bikes [35].

Transit and walking present similar patterns of importance ranking. Both travel time for a trip and tour are important variables for the choice of transit and walking, followed by age and activity duration. As for walking, travel time is a dominant factor since only a short distance can be travelled, and, as for transit, travel time is a critical criterion for determining competitiveness over car and bike [36]. Both travel modes are significantly affected by the number of trips on tour and how the number of trips affects the choice of transit and walking is discussed in a later section using ALE.

3.3. Variable Interaction. Variable interaction was measured for each travel mode using the H-statistic. As shown in equations (5) and (6), the variable interaction can be divided into two cases, i.e., total interaction and two-way interaction. The left side of Figure 2 shows the total interaction of the top ten variables for the choice of each travel mode. Further investigation of total interaction is conducted by two-way interaction, as shown in the right side of Figure 2.

Regarding car, age, sum of activity duration, activity duration, sum of travel time, and travel time are found to have high interaction with other variables. The two-way interactions also indicate that their high interactions are caused mainly within them. This result reveals that their effects on prediction consist of main and significant cross effects, which cause the high variable importance of those variables (see Figure 1). For example, interaction strength between the sum of travel time and travel time is 0.37, which means 37% of the effect of those two variables on the prediction comes through the interaction. On the contrary, the car owner has a low interaction but high importance, indicating that the effect of the car owner appears mainly as

Table 3: Comparison of prediction performances of the ML models.

Travel modes	Specificity			Sensitivity			Balanced accuracy		
	ANN	RF	XGB	ANN	RF	XGB	ANN	RF	XGB
Car	0.806	0.881	**0.920**	0.752	**0.713**	0.670	0.779	**0.797**	0.795
Bike	0.985	0.990	**0.993**	0.116	**0.338**	0.291	0.550	**0.664**	0.642
Transit	0.879	**0.887**	0.883	0.515	0.639	**0.744**	0.697	0.763	**0.813**
Walking	0.819	0.834	**0.850**	0.739	0.826	**0.856**	0.779	0.830	**0.853**
All	0.882	0.909	**0.923**	0.647	0.727	**0.768**	0.764	0.818	**0.845**

Note. ANN = artificial neural network; RF = random forest; XGB = extreme gradient boosting.

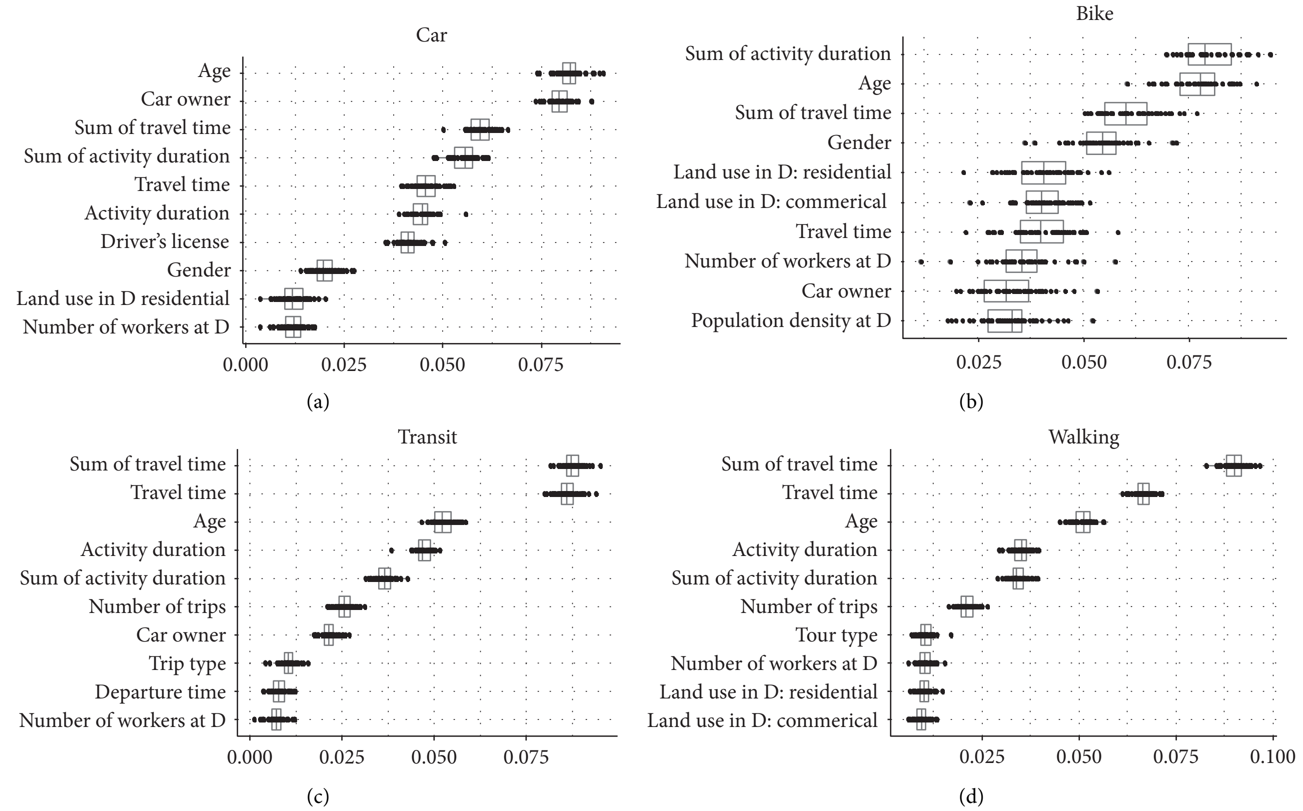
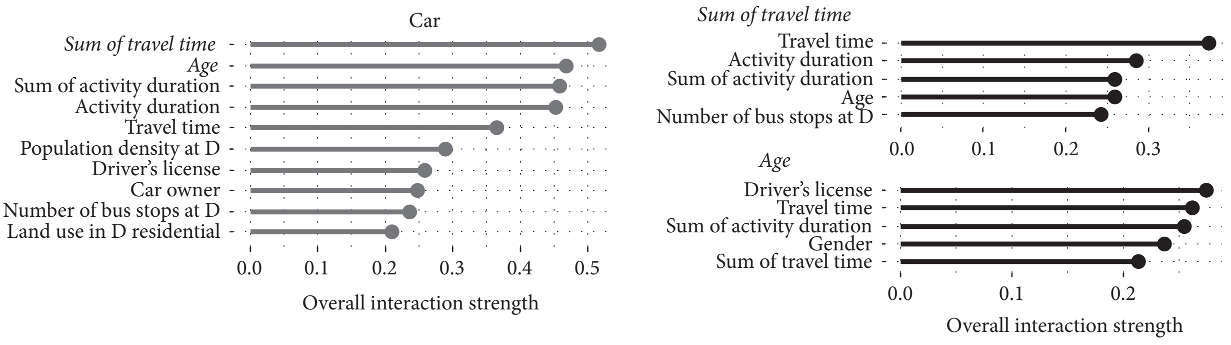
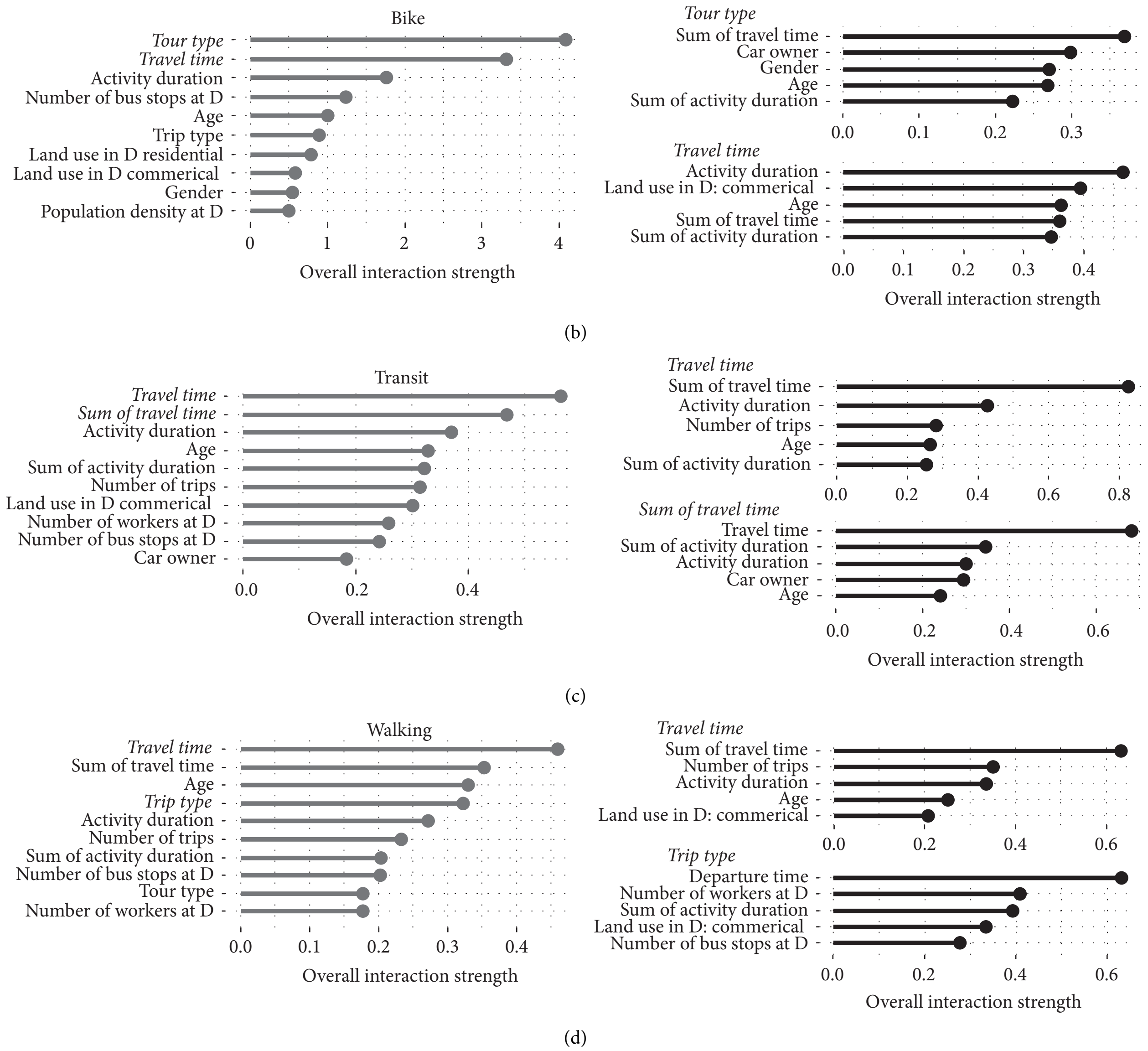


Figure 1: Permutation-based variable importance for each travel mode choice: (a) car, (b) bike, (c) transit, and (d) walking.



(a)

Figure 2: Continued.



(b)

(c)

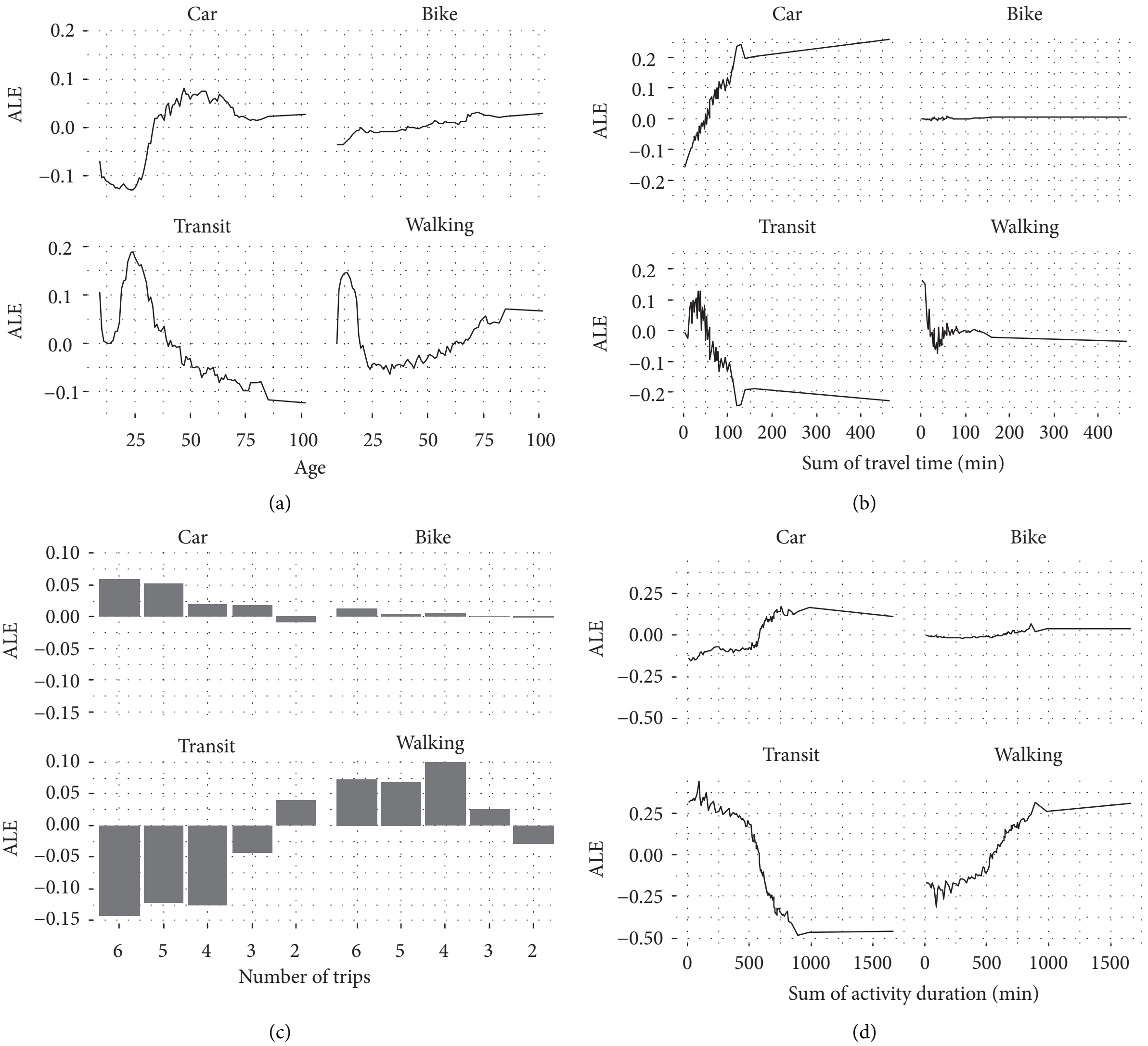
(d)

Figure 2: Variable interactions between one and other variables and between two variables, in predicting travel mode choice of (a) car, (b) bike, (c) transit, and (d) walking.

the main effect. Since the car owner indicates whether the household owns a private car, it can have low interaction with other individual attributes. Age shows the highest interaction with driver's license due to age restrictions on driver's license, although the car owner is more important for the choice of car than the driver's license. High interaction between age and gender indicates that gender, which is not top ten important variables, affects prediction mainly through interaction (i.e., cross effect).

Regarding bike, total interactions are higher than one, indicating that the variance of total interaction is larger than the variance of the ML model. This result can be caused by the low specificity (0.291) of the XGB model to bike choice, of which the changes in the value of a variable cannot thoroughly explain the changes in the class probability of bike. Therefore, it is difficult to extract significant meaning to the interpretation of the total interaction of the bike. Although the two-way interactions for bike have interaction strength smaller than one, significant interpretation is still challenging due to the result of total interaction.

Transit and walking show similar patterns of variable interaction, just like variable importance. Travel time, activity duration, and age have high total interactions for both travel modes. The two-way interaction of travel time for transit and walking choice indicate that, like car choice, the effects of travel time that are of high importance are derived from the significant cross effects among travel time, activity duration, and age. The number of trips is found to have high interaction with travel time for both transit and walking choice. This reveals that the use of transit and walking can be determined by a combination of travel time and the number of trips on tour, which measure the travel fatigue. Unlike



(a)

(b)

(c)

(d)

Figure 3: The ALE values of variables for predicting each travel mode: (a) car, (b) bike, (c) transit, and (d) walking.

other variables presenting a similar pattern for transit and walking, trip type has high total interaction for walking, while low total interaction for transit. Further investigation by two-way interactions shows that trip type is highly correlated with departure time, number of workers at D, activity duration, and land use, which are closely related to trip purpose [37]. The fact that walking includes both trip type and tour type in ten important variables also supports this result. This may be because the choice of walking is significantly linked to eating out and social/recreational trips or going school trip of the student [38].

3.4. Relationship between Variable and Travel Mode Choice. Although the variable importance and interaction tell us the magnitude of the importance and interactions, they do not present how they work. Based on variable importance and interaction, the significant variables are selected for further investigation by the ALE plots, as in Figure 3. While variable importance measures the total effect, including the cross and main effect, the value of ALE measures the main effect of a variable at a specific value (or specific category) on the prediction. Therefore, as shown in Figure 3, age that has a relatively high interaction and importance, and the number of trips that have relatively low interaction and importance can have a similar magnitude of ALE.

Age represents notable patterns of ALE for each travel mode. The choice probability of car gradually increases as age increases from the 20s to 60s, and decreases after the mid-60s, which may suggest a relationship between physical ability or social status and choice of car [38, 39]. The choice of bike gradually increases as age increases, but the difference is tiny. This result is caused by the lack of explanatory power of the XGB model in predicting bike choice. The choice of transit rises steadily until the mid-20s when people graduate from university and then decreases. Teenagers and older people in the study prefer walking as a travel mode more than those of other ages. The choice of walking, after reaching a high in the teenage years, declines toward the 30s and subsequently increases gradually. The peak ALE value of 0.15 among 14 year olds means the probability of walking being chosen is 15% higher for people who are 14 years old than the average age. The above nonlinear relationship between age and travel mode choice is valuable information that cannot be observed from conventional MNL assuming a linear relationship.

The ALE of the categorical variable is also calculated. As the number of trips increases, the choice probability of car and walking increases while the choice probability of transit decreases. This indicates that the number of trips would be a barrier to transit use as it is generally more burdensome to undertake multistop tours [40]. Meanwhile, a large number of trips would include trips of a relatively short distance, such as leisure and shipping trips, so the choice probability of walking would have increased. For bike, near-zero ALE appears, similar to age.

As the sum of travel time and the sum of activity duration increase, the tendency to choose car increases, while the tendency to choose transit decreases. Specifically, when the sum of travel time is more than 50 minutes, the choice probability of car and transit is symmetrical, and this pattern is also observed in the ALE of the sum of activity duration. This result intuitively indicates that car and transit are alternative to each other, depending on travel time and activity duration. When the sum of travel time and activity duration increases, the choice probability of car increases while those of transit decreases. The tendency to use walking as a travel mode decreases as the sum of travel time increases and is maintained after a slight rebound. This rebound may be related to the interaction between the number of trips and the travel time since a large number of trips would include more short-distance trips. People who perform activities for more than 500 minutes a day tend to use a car and walk more than transit. Considering that eight hours are regarded as the average number of working hours, the sum of activity duration is also an indicator for an additional trip activity after/before work, which would be short-distance trip. Therefore, the choice probability of walking continues to increase as the sum of activity duration increases.

4. Conclusions

This paper proposed interpretable ML approaches to predicting and analyzing travel mode choice. The XGB model performed best in the prediction of travel mode choice relative to the RF and ANN models. Understanding the decisions made by the XGB model is valuable both for improving prediction performance and providing insight to the practitioner. The three model-agnostic interpretation methods, i.e., permutation-based variable importance, H-statistic-based variable interaction, and ALE, were applied to investigate the influence of variables in predicting travel mode choices. These methods uncovered the correlated and nonlinear relationships between the behavioral attributes and travel mode choice.

Some interesting findings were highlighted by the results of three interpretation methods. The results of variable importance revealed that age, travel time, and activity duration have high importance for all travel modes. The interactions of those variables explained that such high importance is caused by large cross effects among those variables. These interrelated aspects of the significant variables revealed why the ML model considering the complex relationship of variables outperforms the traditional statistical models in predicting travel mode choice, as reported in the previous studies [1, 6–8]. Also, the tour-related attributes showed high interaction and importance for the choice of all travel modes, indicating that the tour-based analysis is necessary for mode choice, as reported in a modern travel demand forecasting model [41]. These findings regarding the complexity of mode choice emphasized the need to shift from the existing MNL model to a flexible ML model. The varying importance of some variables such as the car owner, tour type, land use, and number of trips according to travel mode indicated that mode-specific analysis should be conducted for targeting each travel mode. For example, to accurately predict the walking trips in the location, trip purpose-related attributes such as land use and activity duration should be collected. The ALE successfully represented the nonlinear relationship between the variables and the change in the choice probability of each travel mode, which is difficult to derive from a conventional MNL. The ALE intuitively showed the alternative patterns of travel mode through the symmetric patterns between travel modes. These results revealed the detailed modal shift patterns according to the behavior attributes such as age and the sum of travel time, which could be used to guide how to divide people into subgroups for predicting travel demand of each mode.

In future research, a proposed interpretation method is needed to extend a more in-depth and broader understanding of travel behavior. Bivariate ALE can be applied to represent the cross effect between variables that separated from the main effect, and it can enrich the explanation of variable interaction. Comparing the interpretation results of ML models with an advanced parametric model, such as a mixed logit model, would also be valuable to validate the model further. Deep learning models [11, 42] are reasonable alternatives for the XGB and RF and the proposed model-agnostic interpretation methods can still available for those models. Local interpretation methods such as local interpretable model-agnostic explanations (LIME) and Shapley additive explanations (SHAP) can contribute to better representation of the heterogeneity of individuals and groups [43, 44], which has also been a critical subject of behavior analysis. Although this study only considers a single primary mode due to the regional travel pattern, a tour-based mode choice model considering the exact combination of modes has been recently proposed to consider the dynamics among trips within the tour [45, 46]. Applying the proposed ML and interpretation methods to those complex modeling tasks would be meaningful future research in the regions with a high rate of multimodal trips.

Data Availability

All data used in this study are available from the author upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (no. 2020R1F1A1074395).

References

[1] F. Wang and C. L. Ross, "Machine learning travel mode choices: comparing the performance of an extreme gradient boosting model with a multinomial logit model," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2672, no. 47, pp. 35–45, 2018.

[2] E. J. Miller, "Modeling the demand for new transportation services and technologies," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2658, no. 1, pp. 1–7, 2017.

[3] D. A. Hensher, W. H. Greene, and Z. Li, "Embedding risk attitude and decision weights in non-linear logit to accommodate time variability in the value of expected travel time savings," *Transportation Research Part B: Methodological*, vol. 45, no. 7, pp. 954–972, 2011.

[4] C. R. Bhat, "Accommodating variations in responsiveness to level-of-service measures in travel mode choice modeling," *Transportation Research Part A: Policy and Practice*, vol. 32, no. 7, pp. 495–507, 1998.

[5] K. E. Train, *Discrete Choice Methods with Simulation*, Cambridge University Press, Cambridge, UK, 2009.

[6] D. Lee, S. Derrible, and F. C. Pereira, "Comparison of four types of artificial neural network and a multinomial logit model for travel mode choice modeling," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2672, no. 49, pp. 101–112, 2018.

[7] J. Hagenauer and M. Helbich, "A comparative study of machine learning classifiers for modeling travel mode choice," *Expert Systems with Applications*, vol. 78, pp. 273–282, 2017.

[8] S. Rasouli and H. J. P. Timmermans, "Using ensembles of decision trees to predict transport mode choice decisions: effects on predictive success and uncertainty estimates," in *Proceedings of the 17th International Conference of Hong Kong Society for Transportation Studies*, Hong Kong, China, 2012.

[9] A. Lhéritier, M. Bocamazo, T. Delahaye, and R. Acuna-Agost, "Airline itinerary choice modeling using machine learning," *Journal of Choice Modelling*, vol. 31, pp. 198–209, 2019.

[10] D. Lee, J. Mulrow, C. J. Haboucha, S. Derrible, and Y. Shiftan, "Attitudes on autonomous vehicle adoption using interpretable gradient boosting machine," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2673, no. 11, pp. 865–878, 2019.

[11] Z. Zhang, C. Ji, Y. Wang, and Y. Yang, "A customized deep neural network approach to investigate travel mode choice with interpretable utility information," *Journal of Advanced Transportation*, vol. 2020, Article ID 5364252, 11 pages, 2020.

[12] S. Wang, B. Mo, and J. Zhao, "Predicting travel mode choice with 86 machine learning classifiers: an empirical benchmark study," in *Proceedings of the 99th Annual Meeting of the Transportation Research Board*, Washington, DC, USA, 2020.

[13] KTDB (Korea Transport Database), *Household Travel Survey Data*, KTDB, Republic of Korea, 2016, https://www.ktdb.go.kr/eng/contents.do?key=263.

[14] MOLIT (2016), http://www.nsdi.go.kr/lxportal/?menuno=3085.

[15] KOSTAT (2016), http://kostat.go.kr/portal/eng/index.action.

[16] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[17] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Science & Business Media, Berlin, Germany, 2009.

[18] M. N. Wright and A. Ziegler, "Ranger: a fast implementation of random forests for high dimensional data in C++ and R," *Journal of Statistical Software*, vol. 77, pp. 1–17, 2017.

[19] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[20] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, and H. Cho, *"Xgboost: extreme gradient boosting," R Package Version 0.4-2*, pp. 1–4, 2015.

[21] E.-J. Kim, H.-C. Park, S.-Y. Kho, and D.-K. Kim, "A hybrid approach based on variational mode decomposition for analyzing and predicting urban travel speed," *Journal of Advanced Transportation*, vol. 2019, Article ID 3958127, 12 pages, 2019.

[22] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, Springer, New York, NY, USA, 4th edition, 2002.

[23] T. Miller, "Explanation in artificial intelligence: insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[24] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[25] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously," *Journal of Machine Learning Research*, vol. 20, pp. 1–81, 2019.

[26] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *Proceedings of the 2010 20th International Conference on Pattern Recognition*, Istanbul, Turkey, 2010.

[27] J. H. Friedman and B. E. Popescu, "Predictive learning via rule ensembles," *The Annals of Applied Statistics*, vol. 2, no. 3, pp. 916–954, 2008.

[28] C. Molnar, "Interpretable machine learning: a guide for making black box models explainable," 2019, https://christophm.github.io/interpretable-ml-book/.

[29] D. W. Apley and J. Zhu, “Visualizing the effects of predictor variables in black box supervised learning models,” 2016, https://arxiv.org/abs/1612.08468.

[30] H. Lee, H.-C. Park, S.-Y. Kho, and D.-K. Kim, “Assessing transit competitiveness in Seoul considering actual transit travel times based on smart card data,” *Journal of Transport Geography*, vol. 80, Article ID 102546, 2019.

[31] S. Lee, “A study on decreasing methods of short-distance auto trips in Seoul,” Report: 2004-R-08, Seoul Development Institute, Seoul, Republic of Korea, 2004.

[32] J. Freedman, J. Castiglione, and B. Charlton, “Analysis of new starts project by using tour-based model of San Francisco, California,” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1981, no. 1, pp. 24–33, 2006.

[33] M. Bradley, J. L. Bowman, and B. Griesenbeck, “SACSIM: an applied activity-based model system with fine-level spatial and temporal resolution,” *Journal of Choice Modelling*, vol. 3, no. 1, pp. 5–31, 2010.

[34] J. Ha, S. Lee, and J. Ko, “Unraveling the impact of travel time, cost, and transit burdens on commute mode choice for different income and age groups,” *Transportation Research Part A: Policy and Practice*, vol. 141, pp. 147–166, 2020.

[35] A. Faghih-Imani, N. Eluru, A. M. El-Geneidy, M. Rabbat, and U. Haq, “How land-use and urban form impact bicycle flows: evidence from the bicycle-sharing system (BIXI) in Montreal,” *Journal of Transport Geography*, vol. 41, pp. 306–314, 2014.

[36] T. A. Arentze and E. J. E. Molin, “Travelers’ preferences in multimodal networks: design and results of a comprehensive series of choice experiments,” *Transportation Research Part A: Policy and Practice*, vol. 58, pp. 15–28, 2013.

[37] E.-J. Kim, Y. Kim, and D.-K. Kim, “Interpretable machine learning models for estimating trip purpose in smart card data,” *Municipal Engineer*, Institution of Civil Engineers, London, UK, 2020.

[38] S. Kim and G. F. Ulfarsson, “Curbing automobile use for sustainable transportation: analysis of mode choice on short home-based trips,” *Transportation*, vol. 35, no. 6, pp. 723–737, 2008.

[39] J. Scheiner, “Social inequalities in travel behaviour: trip distances in the context of residential self-selection and lifestyles,” *Journal of Transport Geography*, vol. 18, no. 6, pp. 679–690, 2010.

[40] L. Yang, Q. Shen, and Z. Li, “Comparing travel mode and trip chain choices between holidays and weekdays,” *Transportation Research Part A: Policy and Practice*, vol. 91, pp. 273–285, 2016.

[41] J. Castiglione, M. Bradley, and J. Gliebe, *Activity-based Travel Demand Models: A Primer*, Transportation Research Board, Washington, DC, USA, 2015.

[42] E.-J. Kim, H.-C. Park, S.-W. Ham, S.-Y. Kho, and D.-K. Kim, “Extracting vehicle trajectories using unmanned aerial vehicles in congested traffic conditions,” *Journal of Advanced Transportation*, vol. 2019, Article ID 9060797, 16 pages, 2019.

[43] S. M. Lundberg and S. I. Lee, “A unified approach to interpreting model predictions,” *Advances in Neural Information Processing Systems*, pp. 4765–4774, MIT Press, Cambridge, MA, USA, 2017.

[44] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you? Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACMSIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, San Francisco, CA, USA, 2016.

[45] M. S. Hasnine and K. Nurul Habib, “Tour-based mode choice modelling as the core of an activity-based travel demand modelling framework: a review of state-of-the-art,” *Transport Reviews*, vol. 41, no. 1, pp. 5–26, 2021.

[46] P. Vovsha, J. E. Hicks, G. Vyas et al., “Combinatorial tour mode choice,” in *Proceedings of the 96th Annual Meeting of the Transportation Research Board*, Washington, DC, USA, 2017.

Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 6654096, 11 pages
https://doi.org/10.1155/2021/6654096

Research Article

Modeling of Merging Decision during Execution Period Based on Random Forest

Gen Li,[1] **Jianxiao Ma**,[1] **and Qiangru Shen**[2]

[1]*College of Automobile and Traffic Engineering, Nanjing Forestry University, Nanjing 210037, China*
[2]*School of Transportation and Civil Engineering, Nantong University, Nantong 226019, China*

Correspondence should be addressed to Gen Li; ligen@njfu.edu.cn

Received 9 November 2020; Revised 21 January 2021; Accepted 25 January 2021; Published 3 February 2021

Academic Editor: Zhibin Li

This study aims to investigate the key feature variables and build an accurate decision model for merging behavior during the execution period by using a data-driven method called random forest (RF). To comprehensively explore the feature variables during merging execution period, nineteen candidate variables including speeds, relative speeds, gaps, time-to-collisions (TTCs), and locations are extracted from a dataset including 375 noise-filtered vehicle trajectories. After the variable selection process, an RF model with 9 key feature variables is finally built. Results show that the gap between the merging vehicle and its putative following vehicle and the ration of this gap to the total accepted gap are the two most important feature variables. It is because merging vehicle drivers can easily observe the putative leading vehicles and control the relative speeds and positions to the putative leading vehicles and they tend to leave more space for their putative following vehicles. Relative speed between the merging vehicle and its following vehicle in the auxiliary lane is the only variable related to the vehicles in the auxiliary lane, which means merging vehicles mainly focus on the traffic condition in the adjacent main lane. Evaluation of the performance in comparison with the state-of-the-art method reveals that the proposed method can obtain much more accurate results in both training and testing datasets, which means RF is practical for predicting the merging decision behavior during execution period and has better transferability.

1. Introduction

As a basic driving task, lane changing has drawn great attention recently. Lane changing behavior was considered to be an important reason for traffic oscillations and accidents [1–4]. It was estimated that lane change crashes account for 4 to 10% of all crashes in the US [5]. Lane-changing behavior is complicated and risky because it is influenced by vehicles in both the current lane and the target lane. Several factors such as velocities and gaps should be taken into account during the lane changing process.

Luckily, with the rapid development of communication technology, driving assistance systems have been developed to help drivers to make safer decisions [6, 7]. Lane-changing decision assistance is one of the key functions of driving assistance systems. It can help drivers make safer decisions to start a lane change. Through the Vehicular Ad-hoc Network (VANET), vehicles can communicate with the surrounding vehicles and roadside unites [8–10]. The lane-changing decision assistance systems can well deal with the situation of discretionary lane-changing by using the data from surrounding vehicles and roadside unites. However, for merging areas on freeway, the judgment rules might be not applicable [11]. In merging areas, drivers need to change to the adjacent main lane within the limited distance, which may result in traffic congestions and even breakdowns [12–17].

As a sequential decision process, the whole merging process can be simplified as a sequential two-step model (gap searching and merging execution) or a three-step model (gap searching, merging position searching, and merging execution) [18–21]. However, most previous studies focused on the gap searching process but neglected the merging execution period. Several seconds are needed to execute the merging behavior and the traffic condition may change dynamically during the whole merging execution period. The ignorance of the merging execution process would lead to reduction of accuracy of traffic simulation and autonomous driving. Thus, there is a critical need to model the merging decision behavior during the execution period. During the merging execution period, the merging vehicles have interactions with putative leading (PL) and putative following (PF) vehicles in the adjacent main lane and the leading (L) and following (F) vehicles in the auxiliary lane. Various influencing factors might be considered for merging decision and should be analyzed in depth. However, previous studies [17] showed that there is multicollinearity between the variables. It was pointed by Balal et al. [22] that most of the lane changing related variables are highly correlated, implying that only a few representative or key variables might be sufficient to describe the interactions of vehicles. However, the selection of key variables is not an easy work. Therefore, the variable selection process should be conducted before building parametric models such as logit model. Improper selection of the key variables might make the performance of the model deteriorate too seriously to be applied to merging assistance systems.

Recently, data mining techniques have received a lot of attention in transportation fields due to their ability to deal with the large-scale data. Some of them can naturally overcome the multicollinearity problem and make full use of the training data. Thus, this study tried use a famous machine learning technique, random forest (RF), to model the merging decision behavior during execution period. It can not only produce more accurate prediction results but also excavate the hidden information among the data. More importantly, RF can effectively select the key variables. The main contribution can be summarized as follows: first, this study gives a comprehensive analysis of the influencing variables of merging decision. Second, the proposed RF method can accurately predict the merging decision during execution period, which can improve the safety and comfort level of driving assistance system if it could be incorporated into lane changing assistance system. Third, a key feature selection process is conducted to investigate the influencing factors. These contributions can not only help understand the diverse influences of different variables on the merging decision but also shed new insights for driver assistance systems and autonomous driving.

The remainder of the paper is organized as follows. Section 2 will provide a state-of-the-art review on the existing studies followed by section 3, which gives the methodology to build a RF model. Section 4 describes the NGSIM data used in this paper and comprehensively analyzes the influencing variables. Results and discussions are presented in section 5. Finally, the concluding remarks are presented in section 6.

2. Literature Review

Predicting merging decision has always been one of the focuses of transportation researches. A great number of models have been developed based on different theories. The first comprehensive lane changing framework was developed by Gipps [23] based on gap acceptance theory. Then, similar frameworks were adopted in other studies [24–27]. However, the gap acceptance theory has been criticized that it cannot reflect the real behavior of drivers. To overcome the deficiency, logistic and logit models were introduced by some researchers [15, 28, 29]. To account for the heterogeneity among drivers, mixed models were proposed by Weng et al. [30] and Li [31]. Game theory models were also developed to model the merging behavior [32, 33]. However, the prediction accuracy of the parametric models is barely satisfactory and the collinearity of influencing variables makes it difficult for researchers to choose appropriate variables to build accurate models [22].

Recently, data-driven methods, such as classification and regression tree (CART), Bayesian network, and fuzzy logic models, were used in building merging models or lane changing models and achieved promising results [16, 34–38]. CART was applied by Weng et al. [11] to model the merging decision in work zone area during execution period, in which time-to-collision (TTC) was considered as a risky factor. Considering the difference between cars and heavy vehicles, Moridpour et al. [39] presented the lane changing model based on fuzzy logic for heavy vehicles. A cooperative merging strategy was developed by Xu et al. [40] for vehicles with V2V and V2I networks, which is applicable to cooperative merging operations under saturated traffic conditions. However, the majority of previous studies separately considered speeds, relative speeds, and gaps as the influencing variables and ignored the interaction of variables. In addition, considering the complexity of merging behavior, a comprehensive analysis of all possible influencing factors should be conducted to better understand the merging decision during execution period.

Previous studies showed that the variables of lane changing behaviour were highly correlated with each other [17, 22, 31]. Thus, selecting some representative or key variables might better describe the interactions of vehicles. However, feature selection has never been an easy work. Feature selection methods can be classified into statistics based methods [41], information theory [42], manifold [43], and rough set [44]. Besides, data-driven methods are also widely used for feature selection [34, 45, 46]. In this study, a popular data-driven method called random forest was applied in this paper to model the merging decision during the execution period. Compared with other models in the literature, the RF has several unique features and advantages. First, it is able to handle multisource heterogeneous data without long-time data processing. Second, as an ensemble machine learning technique based on CART, RF inherits the advantage of CART that can automatically accommodate missing data of independent variables. Third, RF overcomes the deficiency of CART and can automatically resist outliers and is not easy to be affected by small perturbations in the training data. Finally, RF can select the key variables from high dimension data by the importance of all independent variables [45, 47]. RF has been successfully used in traffic prediction and produced promising results [48–51].

3. Methodology

Predicting merging decision can be simplified as a classification problem. Some classical machine learning techniques, such as CART, are very suitable for modeling merging decision. Though CART is efficient and easy-to-use, it is also easy to be affected by small perturbations in the training data [52]. To improve the robustness and generalization capacity of CART, an ensemble learning technique called random forest, which combines the bagging technique, CART, and random subspace method, was proposed by Breiman [45]. RF is an ensemble classifier composed of a group of decision tree classifiers and gets the prediction result by a simple majority vote. The RF model can improve the prediction accuracy of merging decision as well as help connected and autonomous vehicles (CAVs) make safer decisions during merging process. A brief description of random forest is given in this section and detailed fundamentals of mathematics can be referred to Breiman [45].

In RF, bootstrap aggregating (bagging) is the most basic theory. Suppose we have a training dataset (X, Y) with N training samples $\{(X_1, y_1), (X_2, y_2), \ldots, (X_N, y_N)\}$, where $X_i = \{x_i^1, x_i^2, \ldots, x_i^K\}$ and y_i represent the feature vector and the response variable of the sample i, respectively. Through bagging, RF generates B new training sets (X^b, Y^b) by sampling from (X, Y) uniformly and with replacement for N times. By sampling with replacement, some observations may be repeated in each data set (X^b, Y^b) and some may not appear. The probability that each sample in (X^b, Y^b) not selected is $(1 - (1/N))^N$.

Then, we can get

$$\lim_{N \longrightarrow \infty} \left(1 - \frac{1}{N}\right)^N = \frac{1}{e} \approx 0.368. \tag{1}$$

Equation (1) indicates that about 36.8% of the samples are not used in the training process, which is called OOB (Out of Bag) data. These data can be used for validation. Thus, cross-validation or separate test data are not necessary like other machine learning methods. In RF, the OOB error has been proved to be an unbiased estimation of generalization error.

The random subspace method is also used in RF. It can also be called attribute bagging or feature bagging, which means each tree is constructed based on a random subset of the feature variables. This method is designed to reduce the correlation between the trees and improve the generalization accuracy because the RF uses a simple majority vote of all the trees.

Combining the above two methods and CART, the basic steps of RF can be shown in Figure 1 and summarized as follows:

(I) Initiate the algorithm, set $b = 1$.

(II) Use the bootstrap sampling method to obtain a new data set (X^b, Y^b) by random sampling with replacement for N times, and the data that are not sampled will form a set called OOB set.

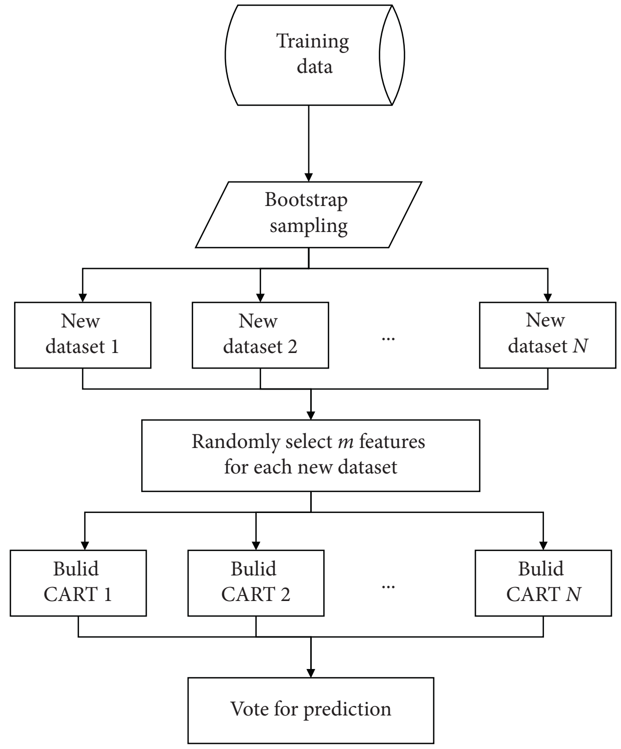


Figure 1: Flow chart of random forest.

(III) Randomly select m feature variables $(m < J)$ and use the selected variables for splitting to train a decision tree T_b based on the new sample set (X^b, Y^b). The decision tree will grow the deepest and is not pruned.

(IV) For $b = 2, \ldots, B$, repeats steps II-III.

The importance of the variables can be sorted by OOB data. RF can screen out important variables in the complex feature variable space, which is conducive to deepen the understanding of the research object. Assuming that the sample subset obtained by bootstrap method is $b = 1, 2, \ldots, B$, the process of using RF to calculate the importance of variable x_j is as follows:

(1) Suppose $b = 1$, and determine the OOB data $L_{b,j}^{\text{OOB}}$.

(2) Use T_b to predict OOB data $L_{b,j}^{\text{OOB}}$, and get the number of accurate predicted samples R_b^{OOB}.

(3) For the feature variable x_j, $j = 1, \ldots, J$, the following calculations are adopted:

  (a) Randomly change the variable values x_j in L_b^{OOB} to get a new data set $L_{b,j}^{\text{OOB}}$

  (b) Use T_b and L_b^{OOB} for prediction and get the number of correct classification R_b^{OOB}

  (c) Calculate the reduction value of classification accuracy, $R_b^{\text{OOB}} - R_{b,j}^{\text{OOB}}$

(4) For $b = 2, \ldots, B$, repeat steps (1–3), and calculate the average value of the reduced value of the classification accuracy to obtain the importance measurement of the variable x_j:

$$D_j = \frac{1}{B}\sum_{b}^{B}\left(R_b^{\mathrm{OOB}} - R_{b,j}^{\mathrm{OOB}}\right). \quad (2)$$

Previous studies have shown that the merging decision could be influenced by a number of highly correlated variables [22, 35]. Thus, the feature selection process must be conducted before building parametric merging decision models. By bagging and random space method, RF can naturally overcome the collinearity of influencing variables. Furthermore, the importance values can be utilized to rank the influencing variables and select the key feature variables through a forward stepwise or backward stepwise elimination process, which will be described in section 5.3.

4. Data Preparation

4.1. Data Description and Processing. In this section, vehicle trajectory data collected by the Federal Highway Administration (FHWA) in the NGSIM project are adopted to verify the proposed RF model. As an open-source dataset, the NGSIM dataset can provide rich and accurate vehicle trajectory data collected on both freeway and urban road [14]. It has been widely used in traffic studies such as traffic flow analysis and driving behavior modeling [18, 37, 53, 54].

Previous studies have shown that the US-101 dataset had the best accuracy and consistency [18, 55]. Thus, this dataset is chosen in this study. Figure 2 shows schematic diagram of data collecting site. One can find that the chosen 640 meters long segment is located between an on-ramp and an off-ramp with five main lanes and one auxiliary lane. Videos were captured from 7:50 a.m. to 8:35 a.m. on June 15, 2005, which was a sunny day. The dataset is updated at a resolution of 10 fps (frames per second) and contains three subsets containing 15 minutes trajectory data [56]. Table 1 shows the aggregate statics of speed and volume for every subset. The coordinates, speed, and acceleration of every vehicle at any instant can be easily obtained from the NGSIM dataset. Previous studies have shown that some random noises existed in the NGSIM data [55, 57]. Filtering and smoothing techniques should be adopted before using. In this study, a data smoothing technique called symmetric exponential moving average filter (sEMA) proposed by Thiemann et al. [57] is applied before further data analysis. In addition, the local coordinates of three subsets are unified to filter the inconsistency of the local coordinates. Detailed steps of data processing can be referred to Li and Sun [17], Li [31], and Li and Cheng [15]. After processing, trajectories of 375 merging vehicle trajectories are extracted from the dataset. All of the vehicles are passenger cars with lengths from 2.5 m to 7.8 m.

4.2. Data Extraction. After selecting the accepted gap, one merging vehicle needs several seconds to find the right time to merge into the adjacent lane and the driver may keep on adjusting the speed and relative position through acceleration deceleration during the execution period. At any time, a merging driver can either choose to continue merging or complete merging as shown in Figure 3. Let y_n^t define the n^{th} merging vehicle's decision at time t. Obviously, y_n^t is a binary variable, shown in the following equation:

$$y_n^t = \begin{cases} 1 \text{ merging vehicle } n \text{ selects to complete merging at time } t \\ 0 \text{ merging vehicle } n \text{ select to continue adjusting at time } t \end{cases}, \quad n = 1, \ldots, N, \; t = 1, \ldots, T_n. \quad (3)$$

Previous studies showed one second is suitable for a driver to make decisions [11, 28, 34, 37]. Thus, we also choose one second in this study. Then, T_n represents the total time to complete merging for vehicle n. Obviously, a merging vehicle can have several observations of $y_n^t = 0$, but only have one observation of $y_n^t = 1$. By extracting the trajectory data of 375 merging vehicles, 1583 observations are obtained in this paper, that is, 375 observations are selecting to merge ($y_n^t = 1$) and 1208 observations are not ($y_n^t = 0$). It means that it takes 3.23 seconds on average for a vehicle to complete merging after making the decision of gap selection.

During the process of merging execution, it has some certain influence on the additional lane and the main lane. At the same time, the merging behavior is also affected by the traffic flow state of the two lanes and the surrounding vehicles. Therefore, the main factors that affect the decision-making of merging vehicles are the speeds, relative speeds, and gaps in the adjacent main lane and the auxiliary lane. However, previous models considered the above variables separately and ignored the interaction between variables. Some studies showed that the gaps between the merging vehicle and PF vehicle in adjacent main line were linearly related to the total gap during the merging process [20]. Figure 4 shows the scatter plots of the PF gaps and the accepted gaps according to the dataset used in this study. A strong linear relationship can be found in Figure 4. One can also find that the range of the ratio of the PF gap to the accepted gap for $y_n^t = 1$ is rather smaller than that for $y_n^t = 0$, indicating that this ratio might be an important factor for merging decision. Therefore, the ratio of the PF gap to the accepted gap is also considered as the influence variable in this paper.

In addition, a surrogate safety measure combining vehicle speeds, space gap, and time-to-collision (TTC) was also considered, because merging driver needs to control vehicle to avoid rear end accidents with the surrounding vehicles. TTC is defined as

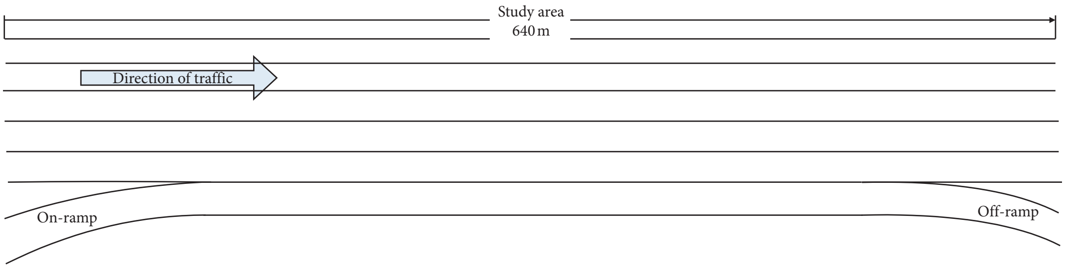


Figure 2: Schematic diagram of U.S. Highway 101.

Table 1: Aggregate statics of three subsets.

Time period	Main lane		Auxiliary lane	
	Volume (vph)	Time mean speed (km/h)	Volume (vph)	Time mean speed (km/h)
7:50 a.m.~8:05 a.m.	8148	44.00	464	63.99
8:05 a.m.~8:20 a.m.	7552	38.80	464	59.26
8:20 a.m.~8:35 a.m.	7108	33.61	496	55.44

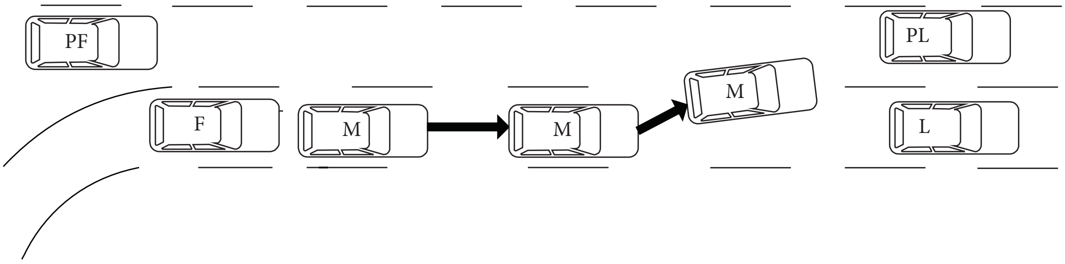


Figure 3: Merging decision during from the start to the end of the execution process.

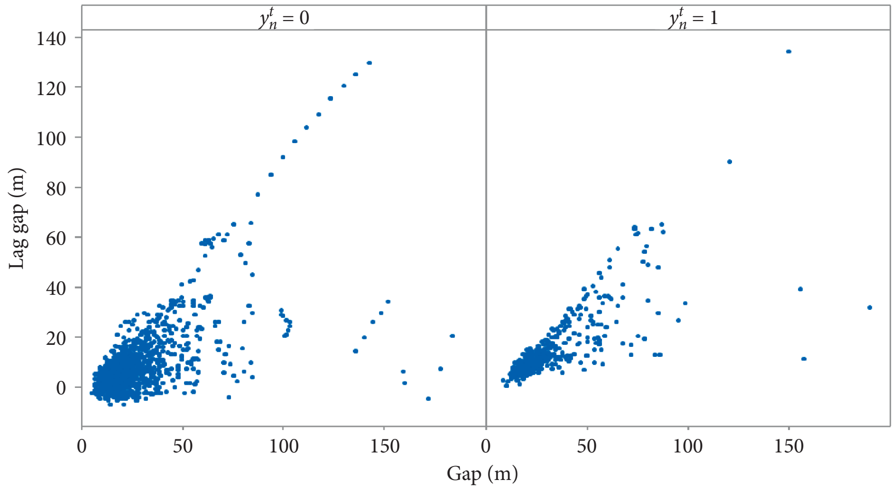


Figure 4: Relationship between the lag gap and the accepted gap in the main line.

$$\text{TTC} = \frac{x_L - x_F - L}{V_F - V_L}, \quad (4)$$

where x_L and x_F are the longitudinal position coordinates of the front bumper of the leading and following vehicle, respectively; V_L and V_F are the speeds of leading and following vehicle, respectively; and L is the length of leading vehicle.

Figure 5 shows the interactions between a merging vehicle and its surrounding vehicles. Table 2 shows the candidate variables and their explanations. It should be pointed out that TTC is negative when the following vehicle moves slower than the leading vehicle, which means that the collision would never occur. In addition, when the speed of the following vehicle is equal to or slightly larger than the

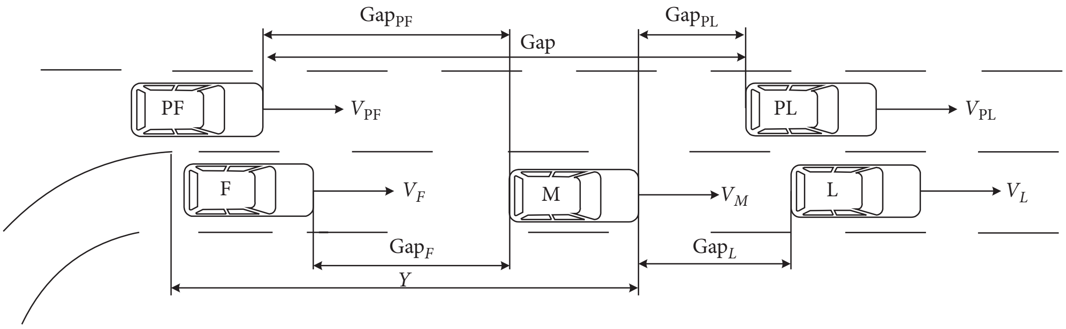


Figure 5: Schematic diagram of candidate variables.

Table 2: Candidate variables and explanations.

Candidate variables	Descriptions
V_M (m/s)	The speed of merging vehicle
V_{PL} (m/s)	The speed of putative leading vehicle
V_L (m/s)	The speed of putative following vehicle
ΔV_{PL} (m/s)	The relative speed between merging vehicle and its putative leading vehicle
ΔV_{PF} (m/s)	The relative speed between merging vehicle and its putative following vehicle
Gap_{PL} (m)	The gap size between merging vehicle and its putative leading vehicle
Gap_{PF} (m)	The gap size between merging vehicle and its putative following vehicle.
(Gap_{PF}/Gap)	The ratio of Gap_{PF} to the total gap
V_L (m/s)	The speed of leading vehicle in the auxiliary lane
V_F (m/s)	The speed of following vehicle in the auxiliary lane
ΔV_L (m/s)	The speed difference between merging vehicle and its leading vehicle in the auxiliary lane
ΔV_F (m/s)	The speed difference between merging vehicle and its following vehicle in the auxiliary lane
Gap_L (m)	The gap size between merging vehicle and its leading vehicle
Gap_F (m)	The gap size between merging vehicle and its following vehicle
TTC_{PL} (s)	The TTC between merging vehicle and putative leading vehicle
TTC_{PF} (s)	The TTC between merging vehicle and putative following vehicle
TTC_L (s)	The TTC between merging vehicle and leading vehicle in the auxiliary lane
TTC_F (s)	The TTC between merging vehicle and following vehicle in the auxiliary lane
Y (m)	The longitudinal position of merging vehicle

leading vehicle, TTC will be infinite or too large. In order to restrict these situations, we will set the TTC range to $(0, 100\text{ s}]$, that is, when TTC is negative or greater than 100 s, it is set to 100 s.

Table 3 shows the main statistical characteristics of the candidate variables for merging behavior. One can find that the merging vehicles move faster than both PF and PL vehicles and the PF vehicles have the lowest average speed. Both the leading and following vehicles in the auxiliary lane move faster than the merging vehicles. Additionally, the average speed of merging vehicles reduces from 12.477 m/s to 12.086 m/s during the merging process to accommodate for the mainline traffic speed, which can also be reflected by changes of average ΔV_{PL} and ΔV_{PF}. It is interesting to find that Gap_{PF} increases from 9.616 m to 16.081 m while Gap_{PL} does not change much. It means Gap_{PF} plays an important role and the PF vehicles tend to yield to the merging vehicles during the merging execution period. One can also find that the TTC_{PL} has the lowest average value during the merging process, indicating that the traffic conflicts between the merging vehicles and PL vehicles might be the most serious.

A Pearson's correlation analysis is conducted to correlation coefficients between dependent variable and independent variables, as shown in Table 4. Bold values are the insignificant correlation coefficients at 0.95 confidence level. One can find that the dependent variable y_n^t has significant correlations with several independent variables, such as V_{PL} and Gap_{PF}. It is interesting to find that there is no significant correlation between Gap_{PL} and y_n^t. (Gap_{PF}/Gap) has the strongest correlation with y_n^t.

5. Modelling Results

After extracting enough data, the RF model is trained and tested in this section to verify the effectiveness. A data mining software called Salford Predictive Modeler is used in this study [16]. The data is randomly divided into two parts: 80% of the lane change cases are randomly selected as the training data, and the remaining 20% is used as the test data for validation. Though RF can use the OOB data for validation, we still do this for comparison with the state-of-the-art methods.

5.1. Parameter Determination. The number of decision trees B is an important parameter of RF. When building decision trees, RF does not prune it. Thus, the modeling

Table 3: Statistics of candidate influencing variables.

Candidate variables	$y_n^t = 0$				$y_n^t = 1$			
	Average	Standard deviation	Maximum	Minimum	Average	Standard deviation	Maximum	Minimum
V_M (m/s)	12.477	3.610	23.389	1.539	12.086	3.269	23.265	2.481
V_{PL} (m/s)	10.997	3.324	19.839	1.578	11.454	3.129	19.967	1.794
ΔV_{PL} (m/s)	1.480	2.233	13.089	−5.481	1.092	1.928	10.895	−5.247
$\mathrm{Gap}_{\mathrm{PL}}$ (m)	13.699	16.781	172.256	0.491	13.821	15.507	152.789	0.631
V_{PF} (m/s)	10.320	3.157	18.681	0.501	10.912	2.925	18.868	1.923
ΔV_{PF} (m/s)	−2.157	2.247	4.344	−12.554	−1.175	1.845	4.113	−11.484
$\mathrm{Gap}_{\mathrm{PF}}$ (m)	9.616	13.660	129.836	0.202	16.081	14.100	134.491	0.410
$(\mathrm{Gap}_{\mathrm{PF}}/\mathrm{Gap})$	0.316	0.274	1.241	0.001	0.452	0.1864	0.900	0.040
V_L (m/s)	14.864	3.541	23.543	3.661	15.550	3.106	21.909	6.610
ΔV_L (m/s)	−2.103	3.135	5.656	−13.708	−3.173	3.396	−15.255	7.837
Gap_L (m)	54.27	38.87	186.94	1.030	56.08	39.54	189.46	2.260
V_F (m/s)	13.282	3.161	21.730	2.585	13.764	3.108	21.566	2.387
ΔV_F (m/s)	0.980	3.010	13.445	−11.474	1.082	3.070	10.741	−12.470
Gap_F (m)	46.84	46.12	105.83	1.610	43.52	43.64	0.66	192.96
$\mathrm{TTC}_{\mathrm{PL}}$ (s)	38.63	42.67	100	0.02	42.98	42.96	100	0.38
$\mathrm{TTC}_{\mathrm{PF}}$ (s)	80.54	35.92	100	0.01	71.21	41.17	100	0.82
TTC_L (s)	85.94	29.09	100	0.36	83.07	30.99	100	0.42
TTC_F (s)	75.61	37.92	100	0.06	68.94	41.33	100	0.06
Y (m)	82.96	68.45	350.86	0.05	94.50	74.49	361.15	0.97

Table 4: Correlation coefficients between dependent variables and independent variables.

	Correlation	Coefficient P value
V_M	**−0.047**	**0.062**
V_{PL}	0.059	0.019
V_{PF}	0.081	0.001
ΔV_{PL}	−0.164	0.0001
ΔV_{PF}	0.190	0.0001
$\mathrm{Gap}_{\mathrm{PL}}$	**0.003**	**0.901**
$\mathrm{Gap}_{\mathrm{PF}}$	0.196	0.0001
$(\mathrm{Gap}_{\mathrm{PF}}/\mathrm{Gap})$	0.224	0.0001
V_L	0.084	0.013
V_F	−0.065	0.021
ΔV_L	−0.140	0.0001
ΔV_F	**0.014**	**0.618**
Gap_L	**0.020**	**0.564**
Gap_F	**−0.031**	**0.270**
TTC_{PL}	0.043	0.086
$\mathrm{TTC}_{\mathrm{PF}}$	−0.106	0.0001
TTC_L	**−0.043**	**0.210**
TTC_F	−0.076	0.007
Y	0.072	0.004

accuracy of RF will increase rapidly with the increase of the number of decision trees at first. However, after reaching a certain number, generating more trees would not improve the model accuracy but increase the computational burden. Previous studies showed that the total number of trees should be set at 200–500 [45, 50]. To ensure the reliability of the modeling results, this paper sets the number of trees at 500.

In RF, a randomly selected subset of features is used to build each single tree. Reducing the number of sampled features m would bring down the correlation among decision tree, leading to less generalization error. However, a too small m would also make the single tree suffer from large prediction error. Different m has been used in different studies [49, 58]; thus, the number of sampled features m should be selected carefully. To select the best m, RF models are trained with an increasing number of m from 1 to 10. Table 5 shows the OOB errors with a different number of m. One can find that the OOB error has the lowest value when m is 3. Thus, the number of randomly sampled features m is set at 3 in this study.

5.2. Variable Importance. The variable importance can be easily obtained by RF according to equation (2). The rank and importance values of independent variables are shown in Table 6.

Table 5: OOB errors with different m.

m	1	2	3	4	5	6	7	8	9	10
OOB error	9.6%	9.4%	9.1%	9.5%	9.5%	9.4%	9.7%	10.1%	10.4%	10.8%

Table 6: Rank of variable importance.

Rank	Variables	Importance value (%)
1	Gap_{PF}	27.35
2	$(\text{Gap}_{\text{PF}}/\text{Gap})$	23.33
3	Gap_{PL}	9.82
4	Y	8.68
5	ΔV_{PF}	6.82
6	TTC_{PL}	5.77
7	TTC_{PF}	3.69
8	ΔV_{PL}	3.46
9	ΔV_F	2.58
10	Gap_F	1.36
11	V_{PF}	1.31
12	V_F	1.28
13	V_M	1.28
14	V_L	1.03
15	ΔV_L	1.02
16	V_L	0.98
17	TTC_F	0.95
18	Gap_L	0.68
19	TTC_L	0.18

According to Table 6, it can be seen that Gap_{PF} and $(\text{Gap}_{\text{PF}}/\text{Gap})$ are the most two important variables, whose importance values are much greater than other variables. The reason is probably that merging vehicle drivers can easily observe the PL vehicles and control the relative speeds and positions with them. Thus, they tend to leave more space for their PF vehicles. This finding is consistent with that of the previous studies [20].

5.3. Feature Variable Selection. From Table 6, one can find that the relative importance values of several variables are rather low, such as TTC_L (0.18%), indicating that there are some redundant or irrelevant variables in the RF model. Therefore, a feature variable selection process introduced by Genuer et al. [59] is applied in this study. The basic steps are shown as follows:

(1) Build a RF model with all candidate variables and rank the variables with the relative importance values in descending order

(2) Delete the variable with the lowest relative importance value and create a new variable set

(3) Build a new RF model with the new variable set and rank the variables with the relative importance values in descending order

(4) Repeat steps (2) and (3) until only one variable remains

(5) Rank all the RF models established in steps (1) to (4) according to the OOB error, and select the model and feature variable set with the lowest error

After feature variable selection, nine feature variables are remained and the OOB error is reduced from 9.1% to 8.9%, indicating that reducing the number of feature variables will not reduce the prediction performance. The values of variable importance in the model are shown in Table 7. It is easy to know from Table 7 that Gap_{PF} and $(\text{Gap}_{\text{PL}}/\text{Gap})$ are still the two most important factors. ΔV_F is the only variable related to the vehicles in the auxiliary lane, which means merging vehicle drivers mainly focus on the traffic condition in the main lane.

5.4. Accuracy of the Model. Table 8 shows the prediction accuracy for training data and testing data. For comparison, a binary logit model and a CART model are also built based on the same dataset. Significant variables are selected by stepwise selection method. The final binary logit model is shown as

$$P\left(y_n^t\right) = \frac{1}{1 + \exp\left(1.710 - 0.0829\Delta V_{\text{PL}} - 0.1481\Delta V_{\text{PF}} + 0.1321\Delta V_L - 0.01551\text{Gap}_{\text{PL}} - 2.076\left(\text{Gap}_{\text{PF}}/\text{Gap}\right) - 0.0405Y\right)}. \quad (5)$$

Table 7: Rank of variable importance after variable selection.

Rank	Selected variables	Importance values (%)
1	$\mathrm{Gap}_{\mathrm{PF}}$	30.59
2	$(\mathrm{Gap}_{\mathrm{PF}}/\mathrm{Gap})$	27.05
3	$\mathrm{Gap}_{\mathrm{PL}}$	12.99
4	Y	6.66
5	ΔV_{PF}	7.32
6	$\mathrm{TTC}_{\mathrm{PL}}$	7.14
7	$\mathrm{TTC}_{\mathrm{PF}}$	5.58
8	ΔV_{PL}	5.01
9	ΔV_F	4.31

Table 8: Prediction results and comparison of models.

Data set	Random forest (%)	Binary logit model (%)	CART model (%)
Training data	91.1	78.9	95.4
Testing data	88.3	72.5	76.29

The results show that the prediction accuracy of the RF model is much better than the binary logit model for both training data and test data. One can also find that CART has the highest prediction accuracy in training data. However, the performance of CART in testing data is much poorer than RF, indicating that RF has better ability to deal with problem of overfitting than CART. In addition, due to the influence of collinearity of variables, only six variables are included in the binary logit model. Some variables that may affect the merging decision behavior in a certain range are ignored by the binary logit model, such as $\mathrm{TTC}_{\mathrm{PL}}$ and ΔV_F. It is clear that RF can overcome the collinearity problem and deeply explore the complicated nonlinear relationships between merging decision and influencing variables. One can also find that the reduction of the accuracy in training and testing dataset is also much smaller than the logit model and CART model, showing that RF is practical for predicting the merging decision during execution period and has better transferability.

6. Conclusions

This study conducts a comprehensive analysis of the influencing variables of merging decision and employs the random forest (RF) to model the merging decision behavior during the execution period. The proposed RF method can accurately predict the merging decision during the execution period and investigate important influencing factors. The US-101 vehicle trajectory data are used to train and validate the RF model. To comprehensively explore the influencing factors during merging execution, 19 candidate variables are extracted including speeds, relative speeds, gaps, time-to-collisions (TTCs), and locations.

The modeling results show that $\mathrm{Gap}_{\mathrm{PF}}$ and $(\mathrm{Gap}_{\mathrm{PF}}/\mathrm{Gap})$ are the most two important variables, whose importance values are much greater than other variables. It is probably because that the merging vehicle drivers can easily observe the PL vehicles and control the relative speeds and positions with them and thus, they tend to leave more space for their PF vehicles. To select the effective variables, a feature variable selection process is adopted and 9 variables are selected in the RF model finally. $\mathrm{Gap}_{\mathrm{PF}}$ and $(\mathrm{Gap}_{\mathrm{PF}}/\mathrm{Gap})$ are still the two most important feature variables. ΔV_F is the only variable related to the vehicles in the auxiliary lane, which means merging vehicles mainly focus on the traffic condition in the adjacent main lane. Evaluation of the performances in comparison with the state-of-the-art method reveals that the proposed method can obtain much more accurate results in both training ant testing datasets. The reduction of the accuracy in training and testing dataset is also much smaller than that of logit model, showing that RF is practical for predicting the merging decision behavior during execution period and has better transferability.

Furthermore, it is obvious that merging drivers face more challenges and may make improper decisions under congested traffic conditions, which might cause long delays. In future, if vehicles can receive the real-time information about the traffic environment via VANETs, the proposed RF models can help the merging vehicles make safer decisions. Thus, the results of this study can also improve the safety and comfort of driving assistance systems and autonomous driving systems.

Data Availability

The NGISM data used to support the findings of this study have been deposited at the website: https://catalog.data.gov/dataset/next-generation-simulation-ngsim-vehicle-trajectories.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was funded by the Science and Technology Innovation Fund for Youth Scientists of Nanjing Forestry University, grant number CX2019021, and Scientific Research Start-up Funds of Nanjing Forestry University, grant number GXL2020012.

References

[1] L. Elefteriadou, R. P. Roess, and W. R. McShane, “Probabilistic nature of breakdown at freeway merge junctions,” *Transportation Research Record*, vol. 1995, no. 1484, pp. 80–89, 1995.

[2] M. Li, Z. Li, C. Xu, and T. Liu, “Short-term prediction of safety and operation impacts of lane changes in oscillations with empirical vehicle trajectories,” *Accident Analysis & Prevention*, vol. 135, Article ID 105345, 2020.

[3] X. Gu, M. Abdel-Aty, Q. Xiang, Q. Cai, and J. Yuan, “Utilizing UAV video data for in-depth analysis of drivers’ crash risk at interchange merging areas,” *Accident Analysis & Prevention*, vol. 123, pp. 159–169, 2019.

[4] C. Yin, J. Zhang, C. Shao, and Society, “Relationships of the multi-scale built environment with active commuting, body mass index, and life satisfaction in China: a GSEM-based

analysis," *Travel Behaviour and Society*, vol. 21, pp. 69–78, 2020.

[5] E. C. Olsen, S. E. Lee, W. W. Wierwille, and M. J. Goodman, "Analysis of distribution, frequency, and duration of naturalistic lane changes," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 46, no. 22, pp. 1789–1793, 2002.

[6] Z. Yao, L. Shen, R. Liu, Y. Jiang, and X. Yang, "A dynamic predictive traffic signal control framework in a cross-sectional vehicle infrastructure integration environment," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 4, pp. 1455–1466, 2019.

[7] Z. Liu, Y. Liu, Q. Meng, and Q. Cheng, "A tailored machine learning approach for urban transport network flow estimation," *Transportation Research Part C: Emerging Technologies*, vol. 108, pp. 130–150, 2019.

[8] Z. Yao, T. Xu, Y. Jiang, and R. Hu, "Linear stability analysis of heterogeneous traffic flow considering degradations of connected automated vehicles and reaction time," *Physica A: Statistical Mechanics and its Applications*, vol. 561, Article ID 125218, 2021.

[9] Z. Yao, R. Hu, Y. Jiang, and T. Xu, "Stability and safety evaluation of mixed traffic flow with connected automated vehicles on expressways," *Journal of Safety Research*, vol. 75, pp. 262–274, 2020.

[10] H. Wang, Y. Qin, W. Wang, and J. Chen, "Stability of CACC-manual heterogeneous vehicular flow with partial CACC performance degrading," in *Transportmetrica B: Transport Dynamics*, vol. 7, no. 1, pp. 788–813, 2019.

[11] J. Weng, S. Xue, and X. Yan, "Modeling vehicle merging behavior in work zone merging areas during the merging implementation period," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 4, pp. 917–925, 2016.

[12] M. Ardakani, J. Yang, and L. Sun, "Stimulus response driving behavior: an improved general motor vehicle-following model," *Advances in Transportation Studies*, vol. 39, 2016.

[13] M. K. Ardakani and J. Yang, "Generalized Gipps-type vehicle-following models," *Journal of Transportation Engineering*, vol. 143, no. 3, pp. 04016011.1–04016011.10, 2017.

[14] V. Alexiadis, J. Colyar, J. Halkias, R. Hranac, and G. McHale, "The next generation simulation program," *Institute of Transportation Engineers ITE Journal*, vol. 74, no. 8, pp. 22–26, 2004.

[15] G. Li and J. Cheng, "Exploring the effects of traffic density on merging behavior," *IEEE Access*, vol. 7, pp. 51608–51619, 2019.

[16] G. Li, S. Fang, J. Ma, and J. Cheng, "Modeling merging acceleration and deceleration behavior based on gradient-boosting decision tree," *Journal of Transportation Engineering*, vol. 146, no. 7, Article ID 05020005, 2020.

[17] G. Li and L. Sun, "Characterizing heterogeneity in drivers' merging maneuvers using two-step cluster analysis," *Journal of Advanced Transportation*, vol. 2018, Article ID 5604375, 2018.

[18] X. Wan, P. J. Jin, H. Gu, X. Chen, and B. Ran, "Modeling freeway merging in a weaving section as a sequential decision-making process," *Journal of Transportation Engineering, Part A: Systems*, vol. 143, no. 5, Article ID 05017002, 2017.

[19] K. I. Ahmed, *Modeling Drivers' Acceleration and Lane Changing Behavior*, Massachusetts Institute of Technology, Cambridge, MA, USA, 1999.

[20] C. F. Choudhury, V. Ramanujam, and M. E. Ben-Akiva, "Modeling acceleration decisions for freeway merges," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2124, no. 1, pp. 45–57, 2009.

[21] G. Li, Y. Pan, Z. Yang, and J. Ma, "Modeling vehicle merging position selection behaviors based on a finite mixture of linear regression models," *IEEE Access*, vol. 7, pp. 158445–158458, 2019.

[22] E. Balal, R. L. Cheu, T. Gyan-Sarkodie, and J. Miramontes, "Analysis of discretionary lane changing parameters on freeways," *International Journal of Transportation Science and Technology*, vol. 3, no. 3, pp. 277–296, 2014.

[23] P. G. Gipps, "A model for the structure of lane-changing decisions," *Transportation Research Part B: Methodological*, vol. 20, no. 5, pp. 403–414, 1986.

[24] Q. Yang and H. N. Koutsopoulos, "A microscopic traffic simulator for evaluation of dynamic traffic management systems," *Transportation Research Part C: Emerging Technologies*, vol. 4, no. 3, pp. 113–129, 1996.

[25] L. Wu, L. Yu, W. Wang, and J. Liu, "Prediction for the region disposition of Panama dry bulk fleet management," *IEEE Access*, vol. 7, pp. 136604–136615, 2019.

[26] P. Hidas, "Modelling vehicle interactions in microscopic simulation of merging and weaving," *Transportation Research Part C: Emerging Technologies*, vol. 13, no. 1, pp. 37–62, 2005.

[27] L. Bloomberg and J. Dale, "A comparison of the VISSIM and CORSIM traffic simulation models on a congested network," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1727, no. 1, pp. 52–60, 2000.

[28] J. Weng and Q. Meng, "Modeling speed-flow relationship and merging behavior in work zone merging areas," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 6, pp. 985–996, 2011.

[29] F. Marczak, W. Daamen, and C. Buisson, "Merging behaviour: empirical comparison between two sites and new theory development," *Transportation Research Part C: Emerging Technologies*, vol. 36, pp. 530–546, 2013.

[30] J. Weng, S. Xue, Y. Yang, X. Yan, and X. Qu, "In-depth analysis of drivers' merging behavior and rear-end crash risks in work zone merging areas," *Accident Analysis & Prevention*, vol. 77, pp. 51–61, 2015.

[31] G. Li, "Application of finite mixture of logistic regression for heterogeneous merging behavior analysis," *Journal of Advanced Transportation*, vol. 2018, Article ID 1436521, 2018.

[32] D. Arbis and V. V. Dixit, "Game theoretic model for lane changing: incorporating conflict risks," *Accident Analysis & Prevention*, vol. 125, pp. 158–164, 2019.

[33] K. Kang and H. A. Rakha, "Modeling driver merging behavior: a repeated game theoretical approach," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2672, no. 20, pp. 144–153, 2018.

[34] Q. Meng and J. Weng, "Classification and regression tree approach for predicting drivers' merging behavior in short-term work zone merging areas," *Journal of Transportation Engineering*, vol. 138, no. 8, pp. 1062–1070, 2012.

[35] E. Balal, R. L. Cheu, and T. Sarkodie-Gyan, "A binary decision model for discretionary lane changing move based on fuzzy inference system," *Transportation Research Part C: Emerging Technologies*, vol. 67, pp. 47–61, 2016.

[36] J. Tang, F. Liu, W. Zhang, R. Ke, and Y. Zou, "Lane-changes prediction based on adaptive fuzzy neural network," *Expert Systems with Applications*, vol. 91, pp. 452–463, 2018.

[37] Y. Hou, P. Edara, and C. Sun, "Modeling mandatory lane changing using Bayes classifier and decision trees," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 2, pp. 647–655, 2014.

[38] E. Wang and J. Sun, "Exploring freeway merging behavior using dynamic bayesian network models," in *Proceedings of*

the International Conference on Transportation and Development 2018: Traffic and Freight Operations and Rail and Public Transit, pp. 120–130, American Society of Civil Engineers, Reston, VA, USA, July 2018.

[39] S. Moridpour, M. Sarvi, G. Rose, and E. Mazloumi, "Lane-changing decision model for heavy vehicle drivers," *Journal of Intelligent Transportation Systems*, vol. 16, no. 1, pp. 24–35, 2012.

[40] L. Xu, J. Lu, B. Ran, F. Yang, and J. Zhang, "Cooperative merging strategy for connected vehicles at highway on-ramps," *Journal of Transportation Engineering, Part A: Systems*, vol. 145, no. 6, Article ID 04019022, 2019.

[41] M. Vasconcelos and N Vasconcelos, "Natural image statistics and low-complexity feature selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 228–244, 2008.

[42] H. Yang and J. Moody, "Data visualization and feature selection: new algorithms for nongaussian data," *Neural Information Processing Systems*, vol. 12, pp. 687–693, 1999.

[43] T. Ye, C. Zu, B. Jie, D. Shen, and D. Zhang, "Discriminative multi-task feature selection for multi-modality classification of Alzheimer's disease," *Brain Imaging and Behavior*, vol. 10, no. 3, pp. 739–749, 2016.

[44] W. Shu and H. Shen, "Incremental feature selection based on rough set in dynamic incomplete data," *Pattern Recognition*, vol. 47, no. 12, pp. 3890–3906, 2014.

[45] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[46] W. Tang and D. M. Levinson, "Deviation between actual and shortest travel time paths for commuters," *Journal of Transportation Engineering, Part A: Systems*, vol. 144, no. 8, Article ID 04018042, 2018.

[47] M. Belgiu, L. Drăguţ, and R. Sensing, "Random forest in remote sensing: a review of applications and future directions," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, 2016.

[48] B. Hamner, "Predicting travel times with context-dependent random forests by modeling local and aggregate traffic flow," in *Proceedings of the 2010 IEEE International Conference on Data Mining Workshops*, pp. 1357–1359, IEEE, Sydney, Australia, December 2010.

[49] L. Cheng, X. Chen, J. De Vos, X. Lai, and F. Witlox, "Applying a random forest method approach to model travel mode choice behavior," *Travel Behaviour and Society*, vol. 14, pp. 1–10, 2019.

[50] J. Cheng, G. Li, and X. Chen, "Developing a travel time estimation method of freeway based on floating car using random forests," *Journal of Advanced Transportation*, vol. 2019, Article ID 8582761, 2019.

[51] Y. Hou, P. Edara, and Y. Chang, "Road network state estimation using random forest ensemble learning," in *Proceedings of the 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1–6, IEEE, Yokohama, Japan, October 2017.

[52] Y. Zhang and A. Haghani, "A gradient boosting method to improve travel time prediction," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 308–324, 2015.

[53] Q. Wang, Z. Li, and L. Li, "Investigation of discretionary lane-change characteristics using next-generation simulation data sets," *Journal of Intelligent Transportation Systems*, vol. 18, no. 3, pp. 246–253, 2014.

[54] X. Wan, P. J. Jin, F. Yang, and B. Ran, "Merging preparation behavior of drivers: how they choose and approach their merge positions at a congested weaving area," *Journal of Transportation Engineering*, vol. 142, no. 9, Article ID 05016005, 2016.

[55] V. Punzo, M. T. Borzacchiello, and B. Ciuffo, "On the assessment of vehicle trajectory data accuracy and application to the next generation SIMulation (NGSIM) program data," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 6, pp. 1243–1262, 2011.

[56] Cambridge Systematics, "NGSIM US 101 data analysis: summary report," in *Prepared for Federal Highway Administration*Cambridge Systematics, Cambridge, UK, 2005.

[57] C. Thiemann, M. Treiber, and A. Kesting, "Estimating acceleration and lane-changing dynamics from next generation simulation trajectory data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2008, pp. 90–101, 2008.

[58] M. Ghasri, T. Hossein Rashidi, and S. T. Waller, "Developing a disaggregate travel demand system of models using data mining techniques," *Transportation Research Part A: Policy and Practice*, vol. 105, pp. 138–153, 2017.

[59] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2225–2236, 2010.