entropy

MDPI

Article

Recognition of Vehicles Entering Expressway Service Areas and Estimation of Dwell Time Using ETC Data

Qiqin Cai [1,2,3], Dingrong Yi [1], Fumin Zou [2,3,*], Zhaoyi Zhou [3], Nan Li [3] and Feng Guo [4]

[1] School of Mechanical Engineering and Automation, Huaqiao University, Xiamen 361021, China
[2] Fujian Key Laboratory for Automotive Electronics and Electric Drive, Fujian University of Technology, Fuzhou 350118, China
[3] Digital Fujian Traffic Big Data Research Institute, Fujian University of Technology, Fuzhou 350118, China
[4] College of Computer and Data Science, Fuzhou University, Fuzhou 350108, China
* Correspondence: fmzou@fjut.edu.cn; Tel.: +86-181-4405-6202

Abstract: To scientifically and effectively evaluate the service capacity of expressway service areas (ESAs) and improve the management level of ESAs, we propose a method for the recognition of vehicles entering ESAs (VeESAs) and estimation of vehicle dwell times using electronic toll collection (ETC) data. First, the ETC data and their advantages are described in detail, and then the cleaning rules are designed according to the characteristics of the ETC data. Second, we established feature engineering according to the characteristics of VeESA and proposed the XGBoost-based VeESA recognition (VR-XGBoost) model. Studied the driving rules in depth, we constructed a kinematics-based vehicle dwell time estimation (K-VDTE) model. The field validation in Part A/B of Yangli ESA using real ETC transaction data demonstrates that the effectiveness of our proposal outperforms the current state-of-the-art. Specifically, in Part A and Part B, the recognition accuracies of VR-XGBoost are 95.9% and 97.4%, respectively, the mean absolute errors (*MAE*s) of dwell time are 52 and 14 s, respectively, and the root mean square errors (*RMSE*s) are 69 and 22 s, respectively. In addition, the confidence level of controlling the *MAE* of dwell time within 2 min is more than 97%. This work can effectively recognize the VeESA and accurately estimate the dwell time, which can provide a reference idea and theoretical basis for the service capacity evaluation and layout optimization of the ESA.

Keywords: VR-XGBoost; K-VDTE; ETC data; ESAs; data mining

Citation: Cai, Q.; Yi, D.; Zou, F.; Zhou, Z.; Li, N.; Guo, F. Recognition of Vehicles Entering Expressway Service Areas and Estimation of Dwell Time Using ETC Data. *Entropy* **2022**, *24*, 1208. https://doi.org/10.3390/e24091208

Academic Editor: Ernestina Menasalvas

Received: 11 August 2022
Accepted: 26 August 2022
Published: 29 August 2022

1. Introduction

By the end of 2021, the total mileage of expressways in China was approximately 169,100 km, ranking first in the world [1]. As an essential and critical core node of expressways, ESA is of great significance in regulating road traffic flow and relieve traffic pressure. However, the infrastructure of most ESAs built in the early years has been unable to meet the demand of the rising traffic volume, resulting in frequent queues and congestion [2,3]. Therefore, a scientific and reasonable evaluation of the service capacity of ESA and further quantitative suggestions for the reconstruction and extension of ESA have become urgent issues at present [4,5]. The pause rate and dwell time of vehicles are essential parameters in the operation and management of ESA. It is not only an important metric for operation evaluation but also a premise for layout optimization. Therefore, it is of great practical significance and application value to accurately estimate the pause rate and dwell time for the quantification of the reconstruction and extension of ESA.

Currently, the main methods for pause rate estimation are the elastic coefficient method [6,7] and feature engineering method [8–18]. The elasticity coefficient method, proposed by the Japanese expressway design standards, mainly uses ESA pause rate survey data and national economic data to establish the pause rate trend model to estimate the future pause rate. Although this method is simple in principle and convenient in calculation, it has certain limitations and one-sidedness. On the other hand, the feature engineering method mainly considers multidimensional features such as major road traffic flow, average speed and human physiological demand and feeds the model for training and learning to estimate the pause rate. However, few studies have investigated the estimation of dwell time in ESA. The limited survey data obtained at a specific ESA are mainly used in existing studies to statistically analyze the vehicle dwell time according to categories such as vehicle type, diurnal differences, seasonal differences, etc. [19–24]. Therefore, the following challenges remain for pause rate and dwell time estimation. (1) Fewer and more difficult to obtain ESA data, resulting in model training effects that could be further improved. (2) Existing studies tend to consider only the overall estimation of the pause rate, ignoring the differences between individual vehicles. (3) The dwell time estimation fails to fully consider the ESA regionality, timeliness and kinematic principle in the vehicle travel law.

To address the aforementioned challenges, we propose a method for the recognition of vehicles entering ESAs (VeESAs) and the estimation of dwell times using ETC data. First, with the rapid development of Internet of Vehicles (IoV) technology in recent years [25,26], China built the world's largest IoV system—the ETC system—at the end of 2019, with a penetration rate of more than 80% of its users. Therefore, this study will utilize ETC data as experimental data to solve the problem of insufficient data. Then, ETC data preprocessing rules are designed by deeply mining the characteristics of ETC data. Second, we proposed an XGBoost-based VeESA recognition (VR-XGBoost) model based on a detailed analysis of the main factors affecting VeESA. On this basis, taking into full consideration the driving pattern of vehicles entering/exiting the ESA, we proposed a kinematics-based vehicle dwell time estimation (K-VDTE) method, which is expected to provide reference ideas for the scientific and reasonable evaluation of the service capacity of the ESA. This work can provide decision support for the layout optimization of ESA reconstruction and extension and improve the management level and high-quality development of ESA.

The main contributions of this study are as follows:

1. We proposed a VR-XGBoost model for recognizing vehicles entering expressway service areas based on ETC data, which not only achieves an effective estimation of the pause rate but also accurately recognizes individual vehicles driving into ESA.
2. Taking into full consideration the driving pattern of vehicles entering/exiting the ESA, we proposed a K-VDTE model for vehicle dwell time estimation.
3. The validity of the proposed method is verified by using real ETC data, which can provide a more scientific and reasonable reference basis for ESA reconstruction and extension.

The remainder of this work is organized as follows: Section 2 reviews related work regarding ESA pause rate and vehicle dwell time estimation. The proposed method, including the framework, data preprocessing, feature engineering, the VR-XGBoost model, and the K-VDTE model, is described in Section 3. Section 4 shows the experimental results and analysis. Finally, the conclusion is presented in Section 5.

2. Related Work

2.1. Pause Rate Estimation

In this section, an overview of pause rate estimation methods is presented. The elastic coefficient method (ECM) was proposed in early Japanese expressway design standards

for calculating the pause rate of various types of VeESAs [6]. Drawing on relevant experience in Japan, Sun et al. [7] concluded that ECM was also applicable to the development pattern of the ESA pause rate in Guangdong Province, China, and used the ECM to estimate the average growth rate of the pause rate to achieve prediction.

Considering the close relationship between the pause rate and ESA spacing, Cui et al. [8] proposed a new method for determining the pause rate based on the continuous vehicle travel time. Through an in-depth analysis of the relationship between the pause rate and traffic flow parameters [9,10], Chen et al. [11] proposed a pause rate estimation method based on a traditional linear regression model, which provided an important reference basis for the layout optimization and function design of ESA. In response to the low accuracy of pause rate prediction, a BP neural network-based ESA pause rate prediction model was constructed [12]. On the basis of previous work, Shen et al. [13] extracted multidimensional feature vectors from the data and constructed a tree-level BP neural network for pause rate prediction, which further improved the prediction accuracy. To further optimize the essential parameters of the wavelet neural network (WNN), some scholars introduced evolutionary algorithms, such as particle swarm optimization (PSO) [14] and genetic algorithm (GA) [15], to optimize the initial parameters of the WNN. The improved WNN-based pause rate prediction models were established, and the validity and reliability were verified on a real dataset. Under the premise of fully investigating the global optimal search capability of particles, Sun et al. [16] improved the topology of traditional PSO and fused it with the XGBoost algorithm to form a combined model for ESA traffic flow prediction. Experiments have demonstrated that the combined model has higher prediction accuracy and stronger generalization ability than a single model.

In the past few decades, deep learning methods [27,28], such as long short-term memory (LSTM) and convolutional neural networks (CNN), have achieved good performance in the field of transportation and are widely used in traffic flow prediction. Wang et al. [17] built a model based on LSTM for ESA instantaneous population analysis and prediction. The experimental results showed that it was able to accurately predict population mobility despite the relatively large population fluctuations. Zhao et al. [18] extracted spatiotemporal features using CNN, LSTM, and attention mechanism models and proposed a short-term traffic flow prediction model based on STL-OMS to achieve an accurate prediction of ESA traffic flow.

2.2. Vehicle Dwell Time Estimation

In this section, an overview of dwell time estimation methods is presented. King et al. [19] conducted an early field survey at nine locations in the United States. The results showed that the average vehicle dwell time in rest areas was 11.4 min, with a standard deviation of 12.87 min, a minimum dwell time of 1 min and a maximum dwell time of 3 h and 31 min. Recently, the Japanese Institute of Expressway General Technology noted through actual statistics that the average dwell time of small vehicles in most ESAs exceeded 25 min [20], while the dwell time for families with elderly and children was extended by an average of 10~20 min in ESAs [21]. Furthermore, analysis of dwell time by vehicle type showed that heavy vehicles had the longest average dwell time, significantly longer than other vehicle types [22]. Analysis of dwell time by seasonal differences showed that all categories of vehicles had longer dwell times in summer than in any other season [23]. Analysis of dwell time by diurnal differences showed that the average dwell time was significantly longer at night than during the day [24].

In addition, Hirai et al. [29] estimated the total dwell time in the service area for the whole trip by mining the ETC trip data using the average travel speed method. The correlation analysis of the dwell time distribution characteristics and rest behavior [30] was expected to construct the next rest behavior model. At the same time, the driver's rest behavior was used to characterize the distribution of vehicle travel time [31] to further construct driving behavior characteristics [32]. A method for calculating the number of stranded vehicles across time was proposed through statistical analysis of vehicle dwell

time and rest behavior characteristics [33], and then a mathematical model for ESA scale design was proposed [34], which was used to optimize the ESA layout [35,36].

3. Methodology

3.1. Framework

In this section, we present the framework of this study, as shown in Figure 1. First, we perform data preprocessing, including extraction of required data, ETC trajectory construction, data cleaning, data fusion and forming of structured data. Second, we consider features such as speed features, spatiotemporal features and external features to construct feature engineering, thus building an XGBoost-based VeESA recognition model. On this basis, a kinematics-based vehicle dwell time estimation model is proposed. This study not only enables the effective recognition of VeESA but also further estimates their dwell time in the service area.

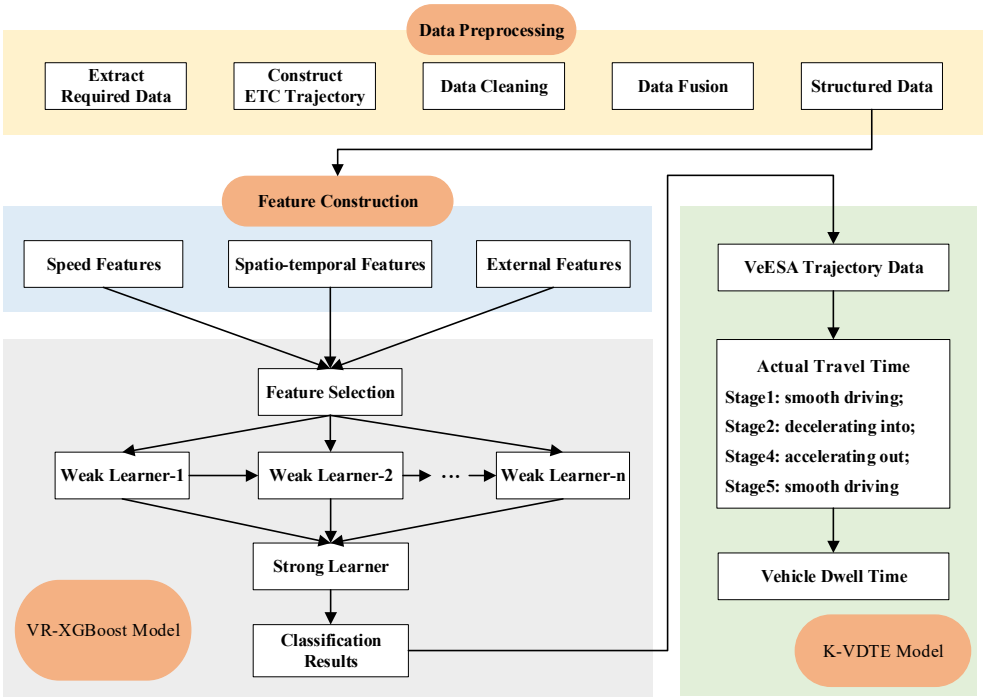

Figure 1. Overall framework.

3.2. Data Overview and Preprocessing

3.2.1. Data Overview

The experimental datasets in this work contain the ETC dataset and ESA dataset. The ETC data were collected by more than 1000 ETC gantries deployed in the whole road network of the Fujian Provincial Expressway. Specifically, as the world's largest IoV system, the ETC system uses radio frequency identification (RFID) technology to enable mobile vehicles equipped with an onboard unit (OBU) to communicate with roadside units (RSU) for data collection [37]. The collection period was from September 3 to 10, 2020. We obtained a total of 42,964,489 ETC data, including vehicle ID (after desensitization), transaction time, gantry ID, vehicle type, etc., as shown in Table 1. According to the classification of vehicle types and tolls of China's expressway, vehicles can be divided into 4 categories of buses, 6 categories of trucks and 6 categories of special operating vehicles. The total number of vehicles is approximately 1.72 million in the dataset. Specifically, each transaction data contains all field information.

Table 1. Description of partial fields in ETC data.

	Field Name	Description	Example
1	VehID	vehicle ID	A000001
2	VehClass	vehicle type	1
3	EnWeight	entrance gross axle weight	1500
4	EnStation	entrance ID	1002
5	EnTime	entrance time	2020/9/5 00:00:00
6	GantryID	gantry ID	G000335001000120020
7	TradeTime	transaction time	2020/9/5 01:00:00
8	Workday	workday	0

The ESA data were collected by the cameras at the entrance and exit of Yangli ESA Part A/B. Specifically, the camera uses the technology of license plate recognition to obtain information about the vehicles entering the service area [38]. The collection period is consistent with the ETC data. We obtained more than 30,000 data points, including vehicle ID (after desensitization), capture time, service area ID and entrance/exit information, as shown in Table 2. The total number of vehicles is approximately 18,000 in the dataset. It is worth noting that this dataset is only used for experimental validation to evaluate the recognition effect and the estimation accuracy of vehicle dwell time.

Table 2. Description of fields in ESA data.

	Field Name	Description	Example
1	SAID	service area ID	Yangli Part A
2	EnEx	entrance/exit	0/1
3	VehID	vehicle ID	A000001
4	CapTime	capture time	2020/9/5 00:00:00

In this work, only the data of Yangli ESA and two ETC gantries before and after it are used, whose deployment locations are shown in Figure 2. To facilitate the following explanation, we have made relevant definitions as follows.

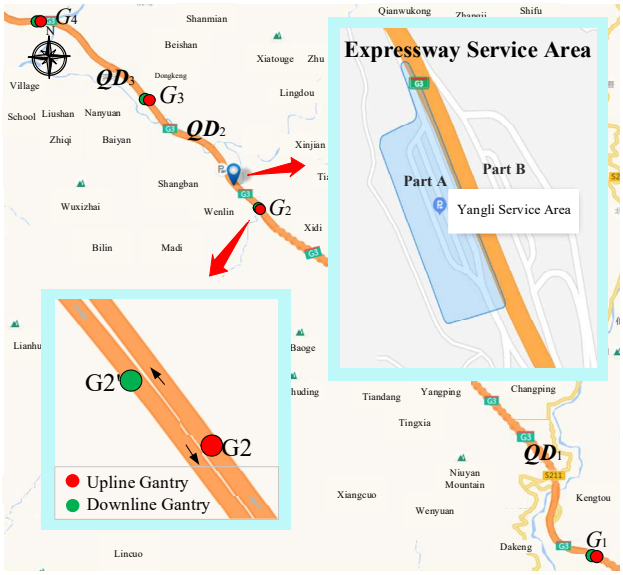

Figure 2. Visualization of ETC gantries and ESA locations.

Expressway Section QD [39]: Each ETC gantry and the entrance/exit of an expressway toll station is collectively called a node $\boldsymbol{G}$, and two adjacent nodes constitute an expressway section, referred to as $\boldsymbol{QD}$:

$$\boldsymbol{QD} = < G_1, G_2 > \quad (1)$$

where G_1 and G_2 are the start and end points of $\boldsymbol{QD}$.

Taking road upline as an example, it can be seen from Figure 2 that G_1 and G_2 constitute **Section** 1 ($\boldsymbol{QD}_1$), G_2 and G_3 constitute **Section** 2 ($\boldsymbol{QD}_2$), where the ESA is located, and G_3 and G_4 constitute **Section** 3 ($\boldsymbol{QD}_3$). It can be found from the partial enlarged detail that the gantries always appear in pairs, which are distributed along the upline and downline of the road, such as G_2 and G_2'. Therefore, the discrete ETC data need to be processed into vehicle trajectories and fused with the ESA data, as detailed in Section 3.2.2.

3.2.2. Data Preprocessing

The prerequisite for effective data mining is to ensure data quality. However, there is a large amount of "dirty" data in ETC data, which is caused by various objective factors such as equipment failure, wireless signal crosstalk and bad weather in the process of ETC data collection, transmission and storage, which seriously affects the potential value of ETC data mining. There are 3 main problems in the "dirty" data as follows:

(1) Data Redundancy

Generally, it is generated by repeated uploading of data in the transmission process or repeated copying in the storage process. This tends to cause an increase in the data scale and serious interference with data mining. In addition, the continuous communication between vehicle OBUs and ETC antennas due to traffic congestion and anchor failure within the antenna coverage area is also a cause of data redundancy. In general, it is sufficient to keep only one of the instances of data and delete the rest directly.

(2) Data Missing

Due to equipment failure, bad weather and other reasons, the vehicle OBU does not communicate or communicates unsuccessfully with the ETC antenna, which results in missing data. At the same time, there is also the possibility of missing data due to network packet loss during data transmission.

(3) Data Abnormality

With the influence of wireless signal crosstalk and other factors, the vehicle OBU of the vehicle traveling on the road upline communicates successfully with the ETC antenna deployed on the road downline, and the dataset generates records that do not comply with expressway driving rules.

However, due to the highly discrete characteristic of ETC data, it is difficult to achieve effective judgment of data abnormalities due to isolated data points. Therefore, it is necessary to rely on the trajectory semantic context formed by the topology of the expressway ETC gantry network to accurately detect and repair the above situation. For this purpose, we further define it as follows: **ETC Trajectory** $\boldsymbol{eTr}$: The sequence of ETC gantry nodes formed by a vehicle passing through a continuous expressway **Section** $< \boldsymbol{QD_1}, \boldsymbol{QD_2}, \ldots, \boldsymbol{QD_{n-1}} >$ is called an ETC trajectory $\boldsymbol{eTr}$:

$$\boldsymbol{eTr} = < \boldsymbol{tr_1}, \boldsymbol{tr_2}, \ldots, \boldsymbol{tr_n} > \quad (2)$$

where $\boldsymbol{tr_1}$ and $\boldsymbol{tr_n}$ are the start and end points of the trajectory, respectively. $\boldsymbol{tr_i}$ is the transaction data when the vehicle travels through the ETC gantry, which contains information such as gantry ID $\boldsymbol{tr_{i.N}}$, transaction timestamp $\boldsymbol{tr_{i.T}}$, vehicle ID $\boldsymbol{tr_{i.P}}$, vehicle type $\boldsymbol{tr_{i.C}}$, entrance gross axle weight $\boldsymbol{tr_{i.EW}}$, entrance ID $\boldsymbol{tr_{i.EID}}$, entrance timestamp $\boldsymbol{tr_{i.ET}}$ and workday $\boldsymbol{tr_{i.H}}$ (consistent with Table 1). n indicates the total number of nodes that the vehicle passes through.

The ETC data cleaning algorithm (Algorithm 1) includes the construction of the vehicle trajectory, data cleaning and data repair. First, the ETC data are grouped by vehicle ID $\boldsymbol{tr_{i.P}}$, entrance ID $\boldsymbol{tr_{i.EID}}$, and entrance timestamp $\boldsymbol{tr_{i.ET}}$. Second, we eliminate duplicate data after sorting by transaction timestamp $\boldsymbol{tr_{i.T}}$ for each set of data. Third, we obtain two adjacent data in each set of data and judge the correctness by its topological information, which mainly includes the removal of redundant data generated by the opposite gantries and the repair of missing data. It is worth noting that the topology dataset includes two subsets: $\boldsymbol{TP}$ and $\boldsymbol{TP'}$, which is a collection of topologies (e.g., $< G_1, G_2 >$). Specifically, $\boldsymbol{TP} = \{< G_1, G_2 >, < G_2, G_3 >, < G_3, G_4 >, \ldots\}$ denotes normal topology data and $\boldsymbol{TP'} = \{< G_1, G_2' >, < G_2, G_3' >, < G_3, G_4' >, \ldots\}$ denotes opposite topology data. The topologies in $\boldsymbol{TP}$ and $\boldsymbol{TP'}$ always appear in pairs, such as $< G_1, G_2 >$ and $< G_1, G_2' >$. Finally, the vehicle trajectories that meet the requirements are added to the trajectory dataset $\boldsymbol{eTRAJ}$. The specific algorithm is shown as follows:

Algorithm 1: ETC data cleaning algorithm

Input: ETC data **eData**, Topology data $\boldsymbol{TP}$, Opposite topology data $\boldsymbol{TP'}$
Output: ETC trajectory dataset $\boldsymbol{eTRAJ}$

1: G_eData = **eData**.Groupby([$\boldsymbol{tr_P}$, $\boldsymbol{tr_{EID}}$, $\boldsymbol{tr_{ET}}$]) ; # Grouping
2: **For** $\boldsymbol{eTr^j} \in$ G_eData **do**: # Traversal operation for each set of data
3: eTr$^j \leftarrow$ eTr^j.$sorted(by = \boldsymbol{tr_T})$ # Sorted by transaction time
4: $\boldsymbol{eTr^j} \leftarrow \boldsymbol{eTr^j}.drop\_duplicates()$ # Data deduplication
5: While (i=1, i < len($\boldsymbol{eTr^j}$)):
6: $\boldsymbol{tp} \leftarrow <\boldsymbol{tr_{i.N}}, \boldsymbol{tr_{i+1.N}}>$
7: **IF** $\boldsymbol{tp} \in \boldsymbol{TP}$ **:**
8: i+= 1;
9: continue;
10: **Else IF** $\boldsymbol{tp} \in \boldsymbol{TP'}$:
11: $\boldsymbol{tp'} \leftarrow <\boldsymbol{tr_{i.N}}, \boldsymbol{tr_{i+2.N}}>$
12: ***IF*** $\boldsymbol{tp'} \in \boldsymbol{TP}$:
13: delete $\boldsymbol{tr_{i+1}}$ # Delete opposite gantry transaction data
14: i+= 2;
15: *Else*:
16: $\boldsymbol{tp''} \leftarrow <\boldsymbol{tr_{i.N}}, \boldsymbol{tr_{i+1.N'}}>$, $\boldsymbol{tp'''} \leftarrow <\boldsymbol{tr_{i+1.N'}}, \boldsymbol{tr_{i+2.N}}>$
17: **IF** $\boldsymbol{tp''} \in \boldsymbol{TP}$ && $\boldsymbol{tp'''} \in \boldsymbol{TP}$:
18: $\boldsymbol{tr_{i+1.N}} \leftarrow \boldsymbol{tr_{i+1.N'}}$ # Replacement of opposite gantry ID
19: i+= 2;
20: Else:
21: break;
22: End IF
23: End IF
24: Else:
25: break;
26: End IF
27: IF i = len($\boldsymbol{eTr^j}$)-1:
28: $\boldsymbol{eTRAJ}.append(\boldsymbol{eTr^j})$;
29: End IF
30: End While
31: End For

The ETC driving trajectory through data cleaning also needs to be fused and matched with the service area traffic data as the label data for subsequent experiments. Therefore, we designed algorithm for fusion of ETC trajectory and ESA data (Algorithm 2). As seen from Section 3.2.1, the ESA is located in $\boldsymbol{QD_2}$. Therefore, only the ETC driving trajectory data and service area data vehicle data must be obtained, and at the same time, the service area entrance and exit capture time in the 2nd and 3rd gantry transaction time periods can match the VeESA to the corresponding ETC driving trajectory. The remaining unmatched driving trajectories are not driven into the service area trajectories.

Notably, the gantry system and the service area entrance/exit camera system appear to be clocked out of sync. Therefore, the time difference delta is set. We make the transaction time of G_2 Δt hours ahead and the transaction time of G_3 Δt hours behind, i.e., $tr_{2.T}^j - \Delta t$ and $tr_{3.T}^j + \Delta t$. By expanding the time range, we ensure that VeESA is fully matched. After the experiments, the time difference in this work is set to 1 h, i.e., $\Delta t = 1h$. The specific algorithm is shown as follows:

Algorithm 2: Fusion of ETC trajectory and ESA data

Input: ETC trajectory dataset ***eTRAJ***, ESA dataset ***sData***, time difference Δt

Output: final trajectory data ***eTr***

1: VIDSet = unique(***sData***.VehID)
2: **For** $\boldsymbol{eTr^j} \in \boldsymbol{eTRAJ}$ **do**:
3: $\boldsymbol{eTr^j} =< tr_P^j, tr_{1,T}^j, tr_{2,T}^j, tr_{3,T}^j, tr_{4,T}^j, tr_C^j, tr_{EID}^j, tr_{ET}^j, tr_W^j, tr_H^j >$;
4: $tr_l^j \leftarrow 0$; $tr_{PCT}^j \leftarrow null$; $tr_{NCT}^j \leftarrow null$;
5: **If** tr_P^j **in** $VIDSet$:
6: sdTmp = ***sData***[***sData***.VehID == tr_P^j]
7: **For** row **in** sdTmp.iterrows():
8: **IF** $tr_{2.T}^j - \Delta t <$ row.CapTime $< tr_{3.T}^j + \Delta t$:
9: $tr_l^j \leftarrow 1$;
10: **IF** row.ExEn = 0:
11: $tr_{PCT}^j \leftarrow$ row.CapTime;
12: **Else**:
13: $tr_{NCT}^j \leftarrow$ row.CapTime;
14: **End IF**
15: **Else**:
16: continue;
17: **End IF**
18: **End For**
19: **Else:**
20: continue;
21: **End IF**
22: $\boldsymbol{eTr^j}.append(< tr_{PCT}^j, tr_{NCT}^j, tr_l^j >)$
23: End For

Through data cleaning and data fusion, a total of approximately 44,000 and 39,000 trajectories were obtained in Yangli Part A and Part B, respectively. The final data samples are shown in Table 3. In these trajectories, the total ETC trajectories of entering Part A and Part B are approximately 7800 and 6700, respectively, and the pause rates of both Parts A and B are approximately 17%. It is worth noting that due to equipment failure and other reasons, there is a missing situation of service area entrance/exit capture data in the experimental dataset. However, this problem does not affect the experiments on the recognition of VeESA in this work. In other words, only one valid capture of data needs to exist in the ESA entrance/exit data to complete the tagging work. Subsequent vehicle dwell time estimation experiments will be conducted by selecting the trajectories where both entrance and exit capture data exist.

Table 3. Examples of experimental data.

	tr_p	$tr_{1.T}$	$tr_{2.T}$	$tr_{3.T}$	$tr_{4.T}$	tr_C	tr_{EID}	tr_{ET}	tr_W	tr_H	tr_{PCT}	tr_{NCT}	tr_l
Part A	A0000001	2020-09-05 08:06:03	2020-09-05 08:08:20	2020-09-05 08:14:12	2020-09-05 08:23:02	23	6101	2020-09-05 06:29:55	18.8	1	2020-09-05 08:01:59	2020-09-05 08:04:08	1
	A0000002	2020-09-03 06:28:34	2020-09-03 06:30:46	2020-09-03 06:43:52	2020-09-03 06:52:07	22	6103	2020-09-03 04:24:28	11.4	0	2020-09-03 06:24:42	2020-09-03 06:33:32	1
	A0000003	2020-09-10 23:38:27	2020-09-10 23:40:24	2020-09-10 23:43:03	2020-09-10 23:50:57	1	2202	2020-09-10 23:19:23	0	0			0
	A0000004	2020-09-07 03:51:13	2020-09-07 03:54:27	2020-09-07 03:59:52	2020-09-07 04:11:46	11	6101	2020-09-06 22:46:11	14.3	0			0
	A0000005	2020-09-03 21:14:13	2020-09-03 21:17:24	2020-09-04 04:56:05	2020-09-04 05:06:41	16	6307	2020-09-03 19:33:43	45.1	0	2020-09-03 21:12:24		1
Part B	A0000006	2020-09-04 17:17:52	2020-09-04 17:32:36	2020-09-04 17:48:00	2020-09-04 17:50:17	16	6707	2020-09-04 16:48:35	50.1	0		2020-09-04 17:36:41	1
	A0000007	2020-09-08 13:42:53	2020-09-08 13:54:00	2020-09-08 14:05:48	2020-09-08 14:07:57	2	6707	2020-09-08 13:23:20	0	0			0
	A0000008	2020-09-06 10:47:19	2020-09-06 10:55:17	2020-09-06 11:19:16	2020-09-06 11:21:21	3	2903	2020-09-06 09:52:21	0	1	2020-09-06 10:47:21	2020-09-06 11:08:32	1
	A0000009	2020-09-06 16:58:22	2020-09-06 17:07:12	2020-09-06 17:09:52	2020-09-06 17:12:13	12	6707	2020-09-06 16:37:20	7.6	1			0
	A0000010	2020-09-10 21:51:28	2020-09-10 22:01:59	2020-09-10 22:21:59	2020-09-10 22:24:20	14	6707	2020-09-10 21:25:11	17.9	0	2020-09-10 21:53:26	2020-09-10 22:09:52	1

3.3. XGBoost-Based VeESA Recognition

3.3.1. Feature Vector Modeling

There are numerous factors that affect the pause rate and dwell time of ESA, which have highly nonlinear characteristics. Therefore, we summarize the previous research results [15,18] and construct feature vectors from 3 dimensions, such as speed features, spatiotemporal features, and external factors. The details are as follows:

(1) Speed Features

The speed features are the key features for the recognition of VeESA. When a vehicle enters the ESA, the average speed of ESA section ($\boldsymbol{QD}_2$) will be significantly lower than $\boldsymbol{QD}_1$ and $\boldsymbol{QD}_2$. Meanwhile, it will also be lower than the overall average speed of other vehicles of the same type in this section. Therefore, we construct the speed feature vector as follows:

$$\boldsymbol{v} = (v_1, v_2, v_3, v_4)^T \tag{3}$$

where $v_1 \sim v_3$ represent the driving state of the individual vehicle during the whole trip. Among them, $v_1 = d_1/(tr_{2.T} - tr_{1.T})$ indicates the average speed of the vehicle in $\boldsymbol{QD}_1$, $v_2 = d_2/(tr_{3.T} - tr_{2.T})$ represents the average speed of the vehicle in $\boldsymbol{QD}_2$, $v_3 = d_3/(tr_{4.T} - tr_{3.T})$ represents the average speed of the vehicle in $\boldsymbol{QD}_3$, $d_1 \sim d_3$ represent the total mileage of $\boldsymbol{QD}_1$, $\boldsymbol{QD}_2$ and $\boldsymbol{QD}_3$, respectively, and $v_4 = \frac{1}{n}\sum_{j=1}^{n} v_2^j$ represents the overall average speed of vehicles of the same type, except for the vehicle in $\boldsymbol{QD}_3$. v_4 mainly avoids the disturbance caused by the reduction in v_2 in certain time periods due to special conditions or traffic congestion.

(2) Spatiotemporal Features

In general, the longer a vehicle spends on the expressway, the more demands on the ESA for drivers and passengers. Therefore, we construct the actual cumulative travel time of the vehicle from the entrance of the toll station to the ESA as one of the spatiotemporal features. At the same time, people's needs for ESA are also different during different times

of the day and on non-workdays. For example, the pause rate of ESA is generally higher at meal times, after midnight and on non-workdays. Therefore, we construct the spatio-temporal features vector as shown below.

$$\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3)^T \tag{4}$$

where γ_1 represents the actual cumulative travel time of the vehicle from the entrance of the toll station to the ESA, γ_2 represents the time period feature, which divides the whole day into 24 time periods by hour, whose value range is 0~23, and γ_3 is a variable for the workday, and its value is 0 (workday) or 1 (non-workday).

(3) External Features

Vehicle type is also an important feature in road traffic. Different types of vehicles have different demands on the ESA. At the same time, the difference in passenger/freight volume will also have some influence on the pause rate of ESA. For example, the more passengers a bus carries, the more stops it needs for rest, dining, etc. Fully loaded large trucks often require services such as breaks and water refills. Therefore, the feature vector is constructed as follows:

$$\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^T \tag{5}$$

where θ_1 represents vehicle type. From the data source of Section 3.2.1, the vehicle types are divided into 16 categories, θ_2 represents passenger/freight volume, and θ_3 represents the traffic flow of the same time slice.

Feature vector modeling is completed by constructing all feature values into vector form.

3.3.2. Modeling of Recognition of VeESA

XGBoost is an integrated learning method based on the boosting algorithm, whose learner usually chooses the decision tree model [40], as shown in Figure 3. The model learns the residuals of the true values and the predicted values of the decision tree by iteratively generating new decision trees. Eventually, the results of all trees are accumulated as the final result to obtain better classification accuracy, i.e., the weak classifiers are combined into a stronger classifier. Therefore, we introduce XGBoost to build a VeESA recognition model.

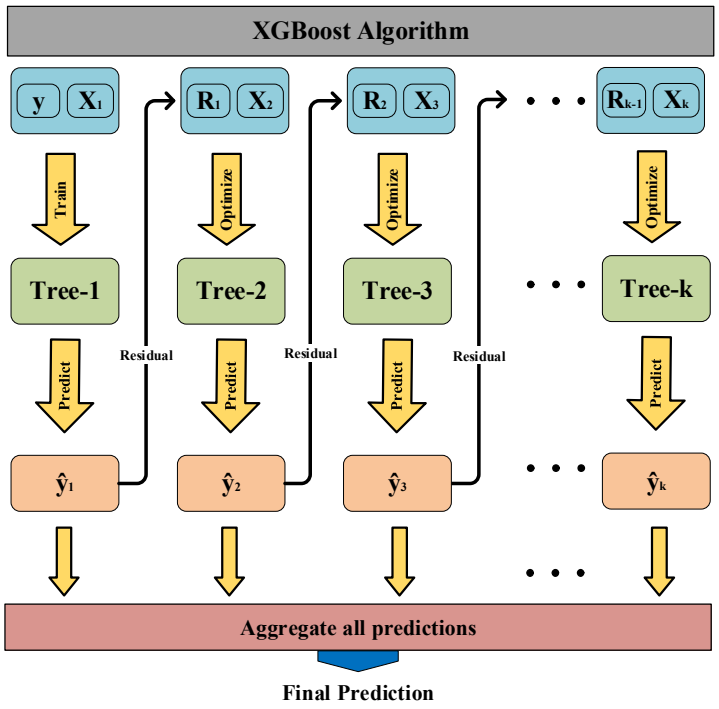

Figure 3. XGBoost Schematic.

We abstracted a 10-dimensional feature vector from the raw ETC data with known label information to form the sample dataset. We set the dataset as $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where $x_i = (v_1, v_2, v_3, v_4, \gamma_1, \gamma_2, \gamma_3, \theta_1, \theta_2, \theta_3)^T (i = 1,2, \ldots, N)$ represents the feature vector of the i-th sample. $y_i = 0/1 (i = 1,2, \ldots, N)$ represents the classification label value corresponding to x_i. We assume that VR-XGBoost integrates K decision trees, and the prediction result is shown in Equation (6):

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \ , f_k \in F \tag{6}$$

where K represents the number of trees, $f_k(x_i)$ represents the predicted value of the k-th decision tree on sample x_i, and F represents the integrated classifier composed of all decision trees.

The objective function of XGBoost consists of the loss function and the regularization item, as shown in Equation (7):

$$Obj = \sum_{i=1}^{n} loss(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{7}$$

where $loss$ represents the logistic regression loss function used for classification.

$$loss(y_i, \hat{y}_i) = \ y_i \ ln(1 + e^{-\hat{y}_i}) + (1 - y_i) \ ln(1 + e^{\hat{y}_i}) \tag{8}$$

$\Omega(f_k)$ represents the L1 regularizer, which is used to prevent the model from overfitting. The formula for the regularizer is Equation (9):

$$\Omega(f_k) = \alpha T_k + \frac{1}{2} \alpha \parallel w_k \parallel_1 \tag{9}$$

where α represents the regularization penalty coefficient, which takes values in the range of [0, 1]. T_k presents the number of leaves of the k-th tree and w_k represents the leaf weight of the k-th tree.

The XGBoost algorithm adopts an additive stepwise integration strategy in the training process. Tree-1 is optimized first, followed by Tree-2 until Tree-K has been optimized.

$$\hat{y}_i^{(0)} = 0 \tag{10}$$

$$\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \tag{11}$$

$$\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \tag{12}$$

…

$$\hat{y}_i^{(k)} = \hat{y}_i^{(k-1)} + f_k(x_i) \tag{13}$$

We improve the prediction accuracy by adding an incremental function f_k to optimize the objective function during the iterative process, which is calculated as in Equation (14):

$$Obj^{(k)} = \sum_{i=1}^{n} loss(y_i, \hat{y}_i^{(k-1)} + f_k(x_i)) + \Omega(f_k) + c \tag{14}$$

where c represents the constant term and $\hat{y}_i^{(k-1)}$ denotes the predicted value in the $k-1$st iteration on sample x_i.

Next, we expand the second-order Taylor formula and discard the constant term to speed up the solution and reduce the running time, which is calculated as Equation (15):

$$\begin{aligned} Obj^{(k)} &= \sum_{i=1}^{n} [l\left(y_i, \hat{y}_i^{(k-1)}\right) + g_i f_k(x_i) + \tfrac{1}{2} h_i f_k^2(x_i)] + \Omega(f_k) = \\ &= \sum_{j=1}^{K} [(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \alpha {w_j}^2)] + \alpha T \end{aligned} \tag{15}$$

where $I_j = \{i | q(x_i) = j\}$ denotes the sample set of leaf j, and g_i and h_i are the first derivative and the second derivative of the loss function, respectively.

The objective function is transformed into a quadratic $Obj^{(k)}$ minimization problem on w_j. Then, we obtain the optimal prediction of each leaf node and the minimum value of the objective function, that is, the optimal value:

$${w_j}^* = -\frac{G_j}{H_j + \alpha} \tag{16}$$

$$(Obj^{(k)})^* = -\frac{1}{2} \sum_{j=1}^{T} \frac{{G_j}^2}{H_j + \alpha} + \alpha T \tag{17}$$

where $G_j = \sum_{i \in I_j} g_i$, $H_j = \sum_{i \in I_j} h_{i.}$

3.4. Kinematics-Based Dwell Time Estimation

The vehicle dwell time is also an essential parameter in the operation and management of ESA. Therefore, after the recognition of VeESA, we need to further estimate the dwell time in the ESA. From the location of the service area between Gantry 2 and Gantry 3, we know that the total travel time of the section consists of the actual travel time of the vehicle in the section and the dwell time. Therefore, the dwell time Δt_s can be obtained as follows:

$$\Delta t_s = \Delta t_{QD2} - \Delta t_r \tag{18}$$

where $\Delta t_{QD} = tr_{3.T} - tr_{2.T}$, Δt_r represents the actual travel time, which is an unknown parameter.

Therefore, the vehicle dwell time estimation is converted into the actual vehicle travel time estimation. Since the traffic conditions of the expressway are relatively smooth, the expressway can approximate the free-flow state in noncongested and nonemergency conditions. Vehicles usually travel smoothly on the highway, so the average speed of $\boldsymbol{QD}_1$ and $\boldsymbol{QD}_3$ can be used as the speed of $\boldsymbol{QD}_2$, and thus the actual travel time of $\boldsymbol{QD}_2$ can be estimated:

$$\Delta t_r = \frac{d_2}{\frac{v_1 + v_2}{2}} = \frac{2d_2}{v_1 + v_2} \tag{19}$$

By substituting Equation (19) into Equation (20), we can obtain the following:

$$\Delta t_s = tr_{3.T} - tr_{2.T} - \frac{2d_2}{v_1 + v_2} \tag{20}$$

Although the average speed method is simple and straightforward, it does not take into account the kinematics of the VeESA during the entrance/exit ramp. In general, VeESA goes through a total of five kinematic stages, including smooth driving upstream, decelerating into the ESA, dwelling in the service area, accelerating out of the ESA and smooth driving downstream, as shown in Figure 4.

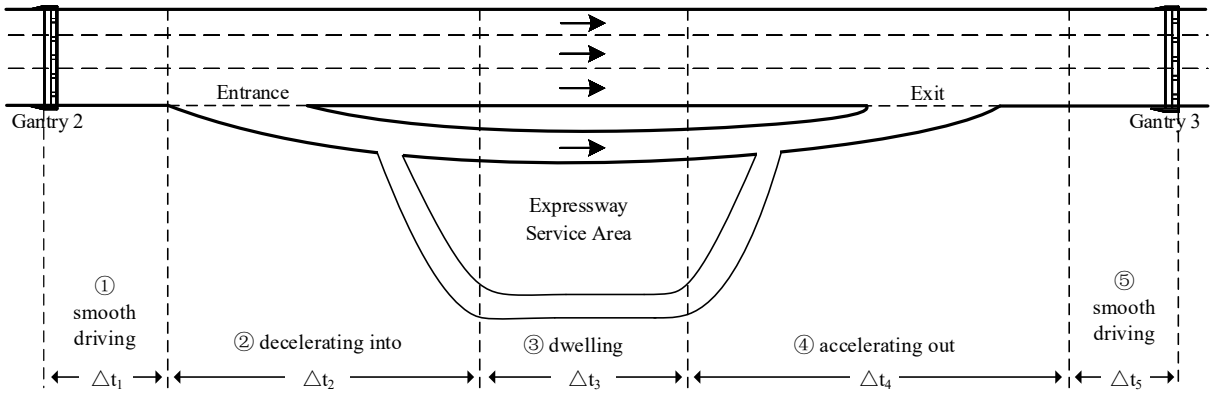

Figure 4. Kinematic process of vehicles driving in and out of the service area.

Therefore, we construct a kinematics-based model for estimating the dwell time, where the actual travel time Δt_r is redefined as follows:

$$\Delta t_r = \Delta t_1 + \Delta t_2 + \Delta t_4 + \Delta t_5 \tag{21}$$

where $\Delta t_1 \sim \Delta t_5$ correspond to the time spent in each of the above five stages.

Stage 1: smooth driving upstream

According to the principle of inertia, the driving state of this stage can be considered as the continuation of the previous section ($\boldsymbol{QD}_1$). Therefore, we approximate Stage 1 as uniform motion. We take the average travel speed of $\boldsymbol{QD}_1$ as the travel speed of Stage 1, and we can obtain the time spent in Stage 1:

$$\Delta t_1 = \frac{\Delta s_1}{v_1} \tag{22}$$

where Δs_1 is the distance from Gantry 2 to the diversion point of the entrance ramp and v_1 is the average travel speed of $\boldsymbol{QD}_1$.

Stage 2: decelerating into the ESA

To a certain extent, the ramps in ESA are similar to the ramps at the entrance and exit of the expressway toll station. However, the ramps at the entrances and exits of toll sta-

tions are usually designed with large curvature, while the ramps at service areas are generally of small curvature or even similar to straight lines. This makes the vehicle smoother when driving in/out of the service area. Therefore, we approximate Stage 2, i.e., the deceleration driving process into the service area entrance ramp, as uniform deceleration linear motion. From stage 1, the initial velocity of uniformly decelerating linear motion is v_1. Let the velocity at the moment Δt_2 be $v_{\Delta t_2}$, the displacement be Δs_2 and the acceleration be a^-, which gives the following.

$$v_{\Delta t_2} = v_1 + a^- \Delta t_2 \tag{23}$$

$$v_{\Delta t_2}^2 - v_1^2 = 2a^- \Delta s_2 \tag{24}$$

We combine Equations (23) and (24) to obtain the following.

$$\Delta t_2 = \frac{2\Delta s_2}{v_1 + v_{\Delta t_2}} \tag{25}$$

where Δs_2 is the distance from the diversion point of the entrance ramp to the service area.

Stage 4: accelerating out of the ESA

Similarly, we approximate Stage 4, i.e., the service area exit ramp acceleration process, as uniformly accelerated rectilinear motion until the driving speed reaches a steady state. From stage 4, it can be seen that the vehicle reaches a smooth state after moment Δt_4, whose driving speed is v_3. Meanwhile, we assume the initial velocity v_{30} and acceleration a^+ of uniformly accelerated linear motion. From Equation (23), we can obtain the time spent in Stage 4:

$$\Delta t_4 = \frac{v_3 - v_{30}}{a^+} \tag{26}$$

In general [41], $a^+ = 0.8 \sim 1.2 m \cdot s^{-2}$.

Stage 5: smooth driving downstream

This stage is similar to the smooth driving upstream. Therefore, we approximate Stage 5 as uniform motion. The driving state of the next section ($\boldsymbol{QD}_3$) can be considered a continuation of Stage 5. We take the average travel speed of $\boldsymbol{QD}_3$ as the travel speed of Stage 5, and we can obtain the time spent in Stage 5 as follows:

$$\Delta t_5 = \frac{\Delta s_5}{v_3} \tag{27}$$

where v_3 is the average travel speed in the back section of the service area, and Δs_5 is the distance from the smooth point in stage 4 to Gantry 3, which is expressed as follows:

$$\Delta s_5 = v_{30} \Delta t_4 + \frac{1}{2} a^+ {\Delta t_4}^2 \tag{28}$$

We substitute Equations (22), (25), (26) and (27) into Equation (21) to obtain the following.

$$\Delta t_r = \frac{\Delta s_1}{v_1} + \frac{2\Delta s_2}{v_1 + v_{\Delta t_2}} + \frac{v_3 - v_{30}}{a^+} + \frac{\Delta s_5}{v_3} \tag{29}$$

After finishing, we obtain the mathematical model for the estimation of vehicle dwell time based on kinematics.

$$\Delta t_s = \Delta t_{QD2} - \left(\frac{\Delta s_1}{v_1} + \frac{2\Delta s_2}{v_1 + v_{\Delta t_2}} + \frac{v_3 - v_{30}}{a^+} + \frac{\Delta s_5}{v_3}\right) \tag{30}$$

It can be generally considered that the velocity $v_{\Delta t_2}$ in uniformly decelerating linear motion and the initial velocity v_{30} in uniformly accelerating linear motion are both zero, which can be simplified as follows:

$$\Delta t_s = \Delta t_{QD2} - \frac{\Delta s_1 + 2\Delta s_2}{v_1} - \frac{v_3(v_3 - 2)}{2a^+} \quad (31)$$

4. Experiments

The experimental platform is a Centos Linux release 7.9.2009 (Core) operating system based on an Intel(R) Core (TM) i9-10900K CPU @ 3.70 GHz and 64 GB RAM, and all experiments were implemented on the open-source web application Jupyter Notebook using Python version 3.8.8.

4.1. VR-XGBoost Evaluation

4.1.1. Construction of Feature Vector

We constructed the feature vector dataset for the training of VR-XGBoost by using 10 statistical features, as shown in Table 4. In the feature vector dataset, each vector contains 10 dimensions of attributes and its classification label l, where l = 0 represents non-VeESA and l = 1 represents VeESA. It is worth noting that the cumulative travel time γ_1 is not directly available in the ETC data. We replaced it with the cumulative travel time from the entrance of the toll station to the front gantry of ESA, i.e., γ_1 is the cumulative travel time from the entrance of the entrance to G_2.

Table 4. Sample of ESA feature vectors.

	v_1	v_2	v_3	v_4	γ_1	γ_2	γ_3	θ_1	θ_2	θ_3	l
Part A	114.6	21.4	109.2	85.7	0.88	14	0	2	0	4	1
	92.0	93.4	84.9	92.1	1.01	10	0	2	0	3	0
	68.0	7.2	66.6	54.1	12.18	21	1	13	13.54	10	1
	75.4	69.6	69.3	48.6	1.86	21	0	14	15.9	11	0
	77.4	60.5	64.8	44.9	18.55	22	1	15	30.28	8	0
	70.0	64.4	77.8	68.4	2.72	15	0	21	0	4	0
Part B	67.6	21.2	79.6	84.1	0.81	17	0	12	9.3	6	1
	80.4	88.3	81.3	83.8	0.73	18	1	12	7.5	8	0
	77.0	20.1	86.1	74.4	0.56	20	0	11	4.6	16	1
	67.1	76.2	72.2	66.3	0.81	21	0	11	0	22	0
	90.1	9.7	104.7	94.9	0.42	22	1	1	0	69	1
	91.6	102.5	99.3	96.7	2.35	23	0	1	0	20	0

Notes: v_1~v_4: km/h; γ_1:h; γ_2: o'clock; θ_2: t; θ_3: veh; γ_3, θ_1, l: dimensionless.

The correlation heatmap is further inscribed for correlation analysis of the feature vectors, as shown in Figure 5. In the figure, blue indicates a positive correlation between vectors, and red indicates a negative correlation between feature vectors. At the same time, when the color is more prominent, the correlation between vectors is stronger. The speed features were positively correlated with the traffic flow and negatively correlated with the cumulative travel time γ_1, vehicle type θ_1 and entrance gross axle weight θ_2. Specifically, there is a strong positive correlation among v_1, v_3 and v_4, which are both weakly positively correlated with v_2. The two features of γ_2 and γ_3 have a very low correlation with other features. Through heatmap analysis, we can clearly understand the correlation between feature vectors.

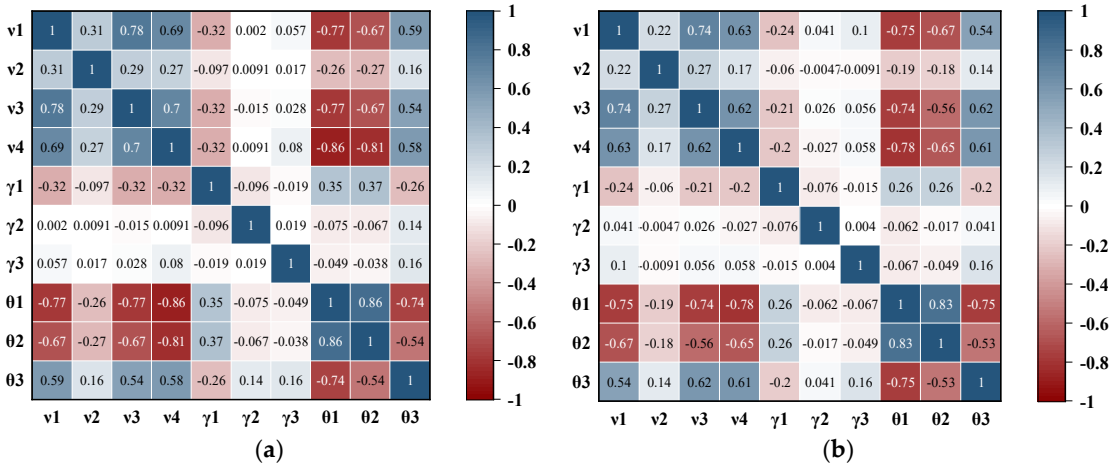

Figure 5. Correlation analysis of feature vectors. (**a**) Part A; (**b**) Part B.

4.1.2. Parameters Selection

The XGBoost classification algorithm has numerous parameters, including the following three aspects.

(1) General Parameters: booster, silent, nthread, etc.
(2) Booster Parameters: the number of decision trees (n_estimators), learning rate (learn_rate), maximum depth of the tree (max_depth), minimum weight in leaf nodes (min_child_weight), parameter that controls the number of leaves (gamma), proportion of sample sampling (subsample), scale of feature sampling (colsample_bytree), etc.
(3) Learning Task Parameters: objective and evaluative (eval_metric).

The general parameters and learning task parameters are set directly according to the model needs, while the booster parameters should be parameter-seeking by the tuning method. At present, the tuning method is mainly the grid search method, which is combined with the K-fold cross-validation method to achieve the optimal parameters [42]. In this work, we also used this method for tuning the parameters and set the cross-validation parameter K = 5. The search range, step size and optimal values of parameters for each parameter are shown in Table 5.

4.1.3. Comparative Analysis of Classification Models

To comprehensively evaluate the effectiveness of the VR-XGBoost model, this work introduced evaluation metrics such as accuracy, precision, recall, and F1-score, as shown below.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{32}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{33}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{34}$$

$$\text{F1} - \text{score} = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{35}$$

We compared and analyzed this experimental model with commonly used machine learning models (e.g., RF, GBDT, KNN), and the experimental results are shown in Table 6. The experimental results showed that VR-XGBoost, RF and GBDT all obtained good

recognition results with accuracy above 95%, while DT performed the worst due to its tendency to overfit. In particular, VR-XGBoost achieved the best results in evaluation metrics. Specifically, in Part B, the accuracy of VR-XGBoost was as high as 97.4%. This result showed the significant superiority of the VR-XGBoost model for the recognition of VeESA.

Table 5. Optimal combination of parameters.

	Parameter	Search Range	Step Size	Optimal Value
General Parameters	booster	gbtree/gblinear		gbtree
	silent	0/1		0
	nthread			4
Booster Parameters	n_estimators	[100, 1000]	100	300
	learn_rate	[0, 0.5]	0.01	0.1
	max_depth	[1, 10]	1	5
	min_child_weight	[1, 10]	1	1
	gamma	[0, 0.5]	0.1	0
	subsample	[0.6, 1]	0.05	0.8
	colsample_bytree	[0.6, 1]	0.05	0.8
Learning Task Parameters	objective	reg:linear/reg:logistic/ binary:logistic/…		binary:logistic
	eval_metric	error/auc/rmse/…		auc

Table 6. Performance comparison of classification models.

Model	Part A				Part B			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
GaussianNB	0.937	0.94	0.937	0.937	0.962	0.962	0.962	0.962
SVM	0.954	0.954	0.954	0.954	0.973	0.974	0.973	0.973
KNN	0.955	0.956	0.955	0.955	0.973	0.974	0.973	0.973
DT	0.913	0.914	0.914	0.914	0.947	0.947	0.947	0.947
AdaBoost	0.941	0.942	0.941	0.941	0.969	0.97	0.969	0.969
LR	0.947	0.947	0.947	0.947	0.966	0.966	0.966	0.966
RF	0.958	0.96	0.958	0.958	0.973	0.974	0.973	0.973
GBDT	0.958	0.959	0.958	0.958	0.973	0.974	0.973	0.973
VR-XGBoost	**0.959**	**0.96**	**0.959**	**0.959**	**0.974**	**0.974**	**0.974**	**0.974**

Next, the feature contributions are further analyzed, as shown in Figure 6. As a whole, the feature contribution ranking from largest to smallest is speed features, external features, and spatiotemporal features. In particular, the contribution of the speed feature in Part A and Part B, both of which exceed 65%, is much higher than that of the spatiotemporal feature and external features. Specifically, the feature contribution of v_2 in speed features is more than 50%, indicating that the feature is the most important. In contrast, the contribution rates of features, such as the actual cumulative travel time γ_1, the time period feature γ_2, the passenger/freight volume θ_2, and the traffic flow θ_3, etc., are all less than 5%. These features seem less important. Through quantitative analysis of contribution rate, we can clearly know the importance of each feature.

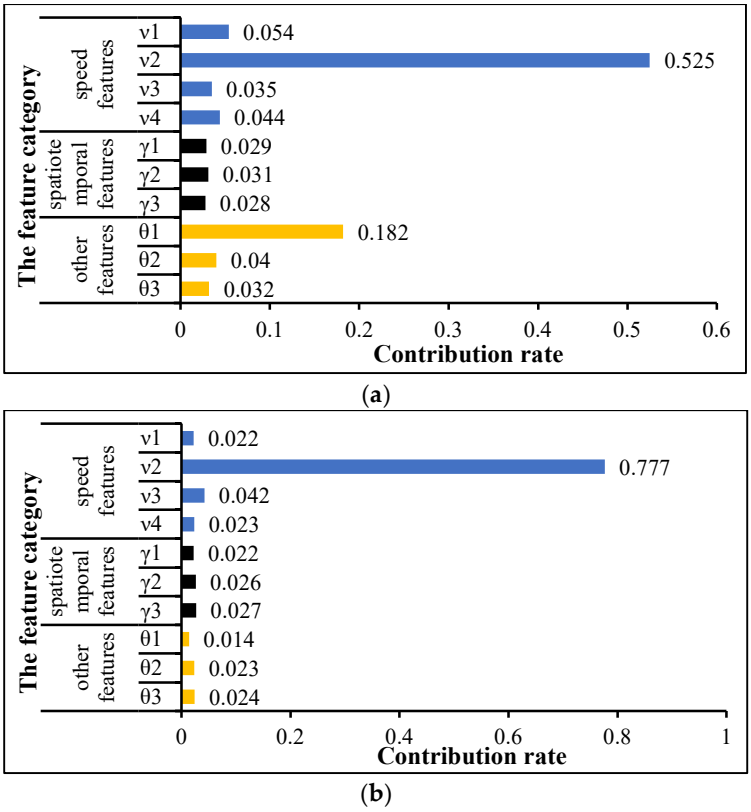

Figure 6. Contribution rate of features. (**a**) Part A; (**b**) Part B.

4.2. K-VDTE Evaluation

We sliced in 5-min increments to count the dwell time, and the distribution is shown in Figure 7. It can be seen that the dwell times in Part A and Part B both exhibit a long-tailed distribution, which indicates that most vehicles stay in the ESA only temporarily and briefly. Specifically, the number of vehicles with a dwell time of 5~10 min is the greatest, and more than 90% of the vehicles have a dwell time of less than 1 h in the ESA. Furthermore, the average dwell time in Part A and Part B was approximately 30 min, with a standard deviation of approximately 70 min, a minimum dwell time of less than 30 s and a maximum dwell time of more than 12 h. Through the statistical analysis, we can clearly understand the general situation of the vehicle dwell time in each ESA.

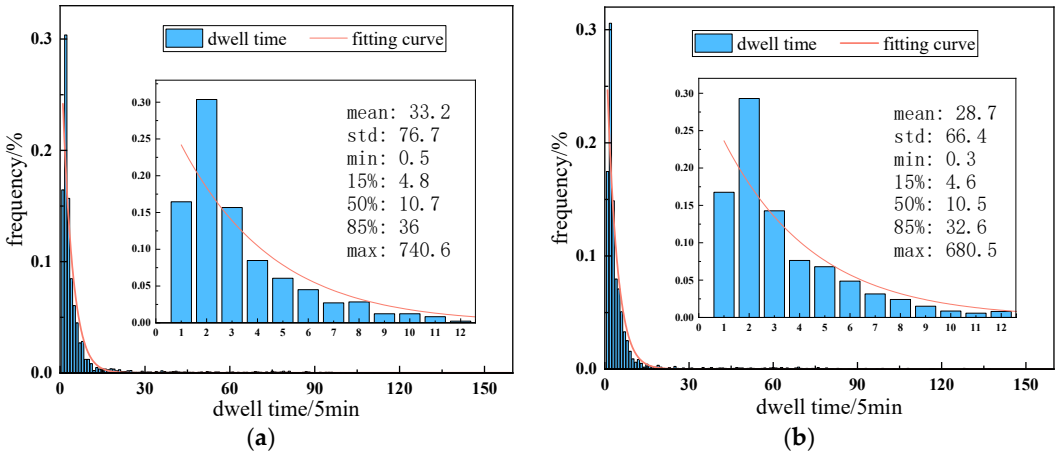

Figure 7. Dwell time distribution. (**a**) Part A; (**b**) Part B.

To evaluate the effectiveness of the K-VDTE model, the estimation errors are quantified using the evaluation metrics of root mean square error (*RMSE*), mean absolute error (*MAE*), and *R* coefficient, as shown below:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{36}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{37}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{38}$$

where $\hat{y}_i$ denotes the estimated dwell time obtained using the model, y_i denotes the true dwell time, $\bar{y}_i$ is the average dwell time, and n denotes the amount of data.

We compared the proposed K-VDTE model with the traditional averaging speed method and commonly used machine learning models, as shown in Table 7. The experimental results show that the proposed K-VDTE method performs the best, while the machine learning model performs the worst. Specifically, taking Part B as an example, the *MAE* of the K-VDTE model was only 14 s, which was not only more than one times better than the average speed method but also at least four times better than the machine learning models. This demonstrated the higher accuracy of our method. Moreover, comparing the *RMSE* of each model, the proposed K-VDTE model improved at least one order of magnitude over the machine learning models, which indicated that the proposed method was more robust.

Moreover, the integrated learning models, such as XGBoost, RF and GBDT, among machine learning models, perform better on evaluation metrics, while the single models, such as Lasso and KNN, obtain very poor results on all evaluation metrics. This result indicates that single models may not be suitable for vehicle dwell time estimation.

Table 7. Performance comparison of estimation models (unit: s).

Model	Part A			Part B		
	RMSE	***MAE***	$\boldsymbol{R^2}$	***RMSE***	***MAE***	$\boldsymbol{R^2}$
Lasso	4046	2095	0.275	3831	1823	0.2
KNN	3443	1148	0.475	3506	1073	0.33
AdaBoost	536	431	0.987	486	400	0.987
DT	318	90	0.995	263	65	0.996
ExtraTree	365	92	0.994	1248	146	0.915
RF	276	71	0.997	263	55	0.996
GBDT	272	72	0.997	315	61	0.994
XGBoost	242	70	0.997	263	62	0.996
AvgSpeed	85	71	1.000	36	30	1.000
K-VDTE	**69**	**52**	**1.000**	**22**	**14**	**1.000**

To investigate the estimated errors in depth, we performed a statistical analysis of the *MAE* of the dwell times and carved out the distribution of the cumulative probabilities, as shown in Figure 8. It can be seen that the distribution curves of the cumulative probabilities in Part A and Part B all exhibited a rapid increase with the increase in the dwell time estimation error until they stabilized after 2 min. Specifically, $P\{MAE \leq 120s\} > 95\%$ indicates that the probability of keeping the *MAE* within 2 min is more than 95%. Specifically, taking Part B as an example, the probability of controlling the *MAE* within 1 min and 2 min are $P_B\{MAE \leq 60s\} > 97\%$ and $P_B\{MAE \leq 120s\} > 99.8\%$, respectively. The results further validate that the K-VDTE model has strong robustness.

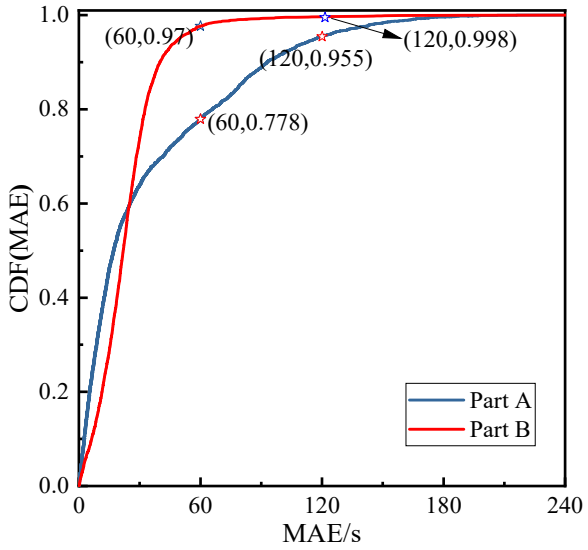

Figure 8. Cumulative probability distribution of *MAE*.

5. Conclusions

In this work, we proposed a method for the recognition of vehicles entering expressway service areas and the estimation of dwell time based on ETC data. This method provides reference ideas for scientific and reasonable evaluation of the service capacity of the ESA, which can also provide decision support for the optimization of the layout when reconstructing and extending the ESA. The specific conclusions are as follows:

(1) Experiments were conducted using real ETC data with a user penetration rate of over 80%. It not only solves the issue of insufficient data volume but also solves the geographical differences existing in different service areas in vehicle dwell time estimation. It can provide a more scientific and reasonable reference basis for the evaluation of the service capacity of ESA.

(2) Considering multidimensional information such as speed features, spatiotemporal features and external features, we constructed a VR-XGBoost model. This model can achieve not only the estimation of the overall pause rate of ESA but also the accurate recognition of vehicles entering the service area.

(3) After an in-depth study of the driving pattern of vehicles in the process of driving in/out of the ESA, we proposed a K-VDTE to realize vehicle dwell time estimation. The estimation accuracy of vehicle dwell time can be further improved by considering vehicle kinematics.

However, the present method also has certain limitations, whose expressway traffic state must approximate free-flow conditions. In the future, we will further explore the vehicle driving characteristics and laws under nonfree flow conditions to form a more scientific and reliable evaluation system.

Author Contributions: Conceptualization, F.Z. and Q.C.; methodology, Q.C.; software, Q.C., Z.Z. and N.L.; validation, D.Y., Q.C. and F.Z.; formal analysis, Q.C. and F.Z.; investigation, F.G. and Q.C.; resources, F.Z.; data curation, Q.C. and D.Y.; writing—original draft preparation, Q.C., Z.Z. and N.L; writing—review and editing, F.Z., Q.C., D.Y., Z.Z. and N.L; visualization, Z.Z. and N.L.; supervision, Q.C.; project administration, F.G. and Q.C.; funding acquisition, F.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the National Natural Science Foundation of China (41971340), the Special Funds for the Central Government to Guide Local Scientific and Technological Development (2020L3014), the 2020 Fujian Province "the Belt and Road" Technology Innovation Platform (2020D002), and the Fujian Expressway Science and Technology Project (KJ202001).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Data were obtained from Fujian Expressway Information Technology Co., Ltd. and are available from the authors with the permission of Fujian Expressway Information Technology Co., Ltd.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ministry of Transport of the Peoples's Republic of China. *2021 Statistical Bulletin on the Development of the Transportation Industry*; Ministry of Transport of the People's Republic of China: Beijing, China, 2021.
2. Hernandes, S.; Poliak, M.; Poliaková, A. Draft of freight transport car parks facilities. *Arch. Motoryz.* **2019**, *85*, 41–47.
3. Seya, H.; Zhang, J.; Chikaraishi, M.; Ying, J. Decisions on truck parking place and time on expressways: An analysis using digital tachograph data. *Transportation* **2020**, *47*, 555–583.
4. Alkhatni, F.; Ishak, S.; Milad, A. Characteristics and Potential Impacts of Rest Areas Proximate to Roadways: A Review. *Open Transp. J.* **2021**, *15*, 260–271.
5. Hao, S.; Yang, P.; He, Z. *The Design and Implementation of Big Data Application Management Platform for Highway Service Areas. The International Conference on Cyber Security Intelligence and Analytics*; Springer: Cham, Switzerland, 2022; pp. 158–164.
6. Japan Highway Public Corporation. *Japan Highway Design Essentials;* Japan Highway Public Corporation: Tokyo, Japan, 1991.
7. Sun, X. Investigation and prediction of highway service area occupancy rate in Guangdong Province. *Guangdong Highw. Traffic* **2002,** 1, 50–52. https://doi.org/10.3969/j.issn.1671-7619.2002.01.0
8. Cui, H.; Liu, K. A New Determine Method on Pause Rate in Expressway Service Area Based on Vehicle Continuous Travel Time. *J. Hebei Univ. Technol.* **2008**, *37*, 100–104.
9. Wang, J.; Tang, Y. Transportation potential calculation model of pause rate in expressway service area. *China J. Highw. Transp.* **2008**, *21*, 109–114.
10. Ramli, I.; Hassan, S.; Hainin, M. Parking demand analysis of rest and service area along expressway in southern region, Johor Malaysia. *Malays. J. Civ. Eng.* **2017**, 29. https://doi.org/10.11113/mjce.v29.15687
11. Chen, Y.; Wang, J. A Study on Forecast Method of Pause Rate in Expressway Service Area. In Proceedings of the 22nd COTA International Conference of Transportation Professionals (CICTP 2022), to be held on July 8-11, 2022, in Changsha, China
12. Liu, J.; Shen, X.; Liu, H. Prediction Model for Percentage of Expressway Traffic Entering Rest Area Based on BP Neural Network. *Highway* **2012,** 6, 164–168. https://doi.org/10.3969/j.issn.0451-0712.2012.06.034.
13. Shen, X.; Wang, L.; Liu, H.; Yang, J. Estimation of the percentage of mainline traffic entering rest area based on BP neural network. *J. Appl. Sci.* **2013**, *13*, 2632–2638.
14. Lu, C. Research on Reasonable Allocation of Service Facilities in Expressway Service Area. Master's Dissertation, Beijing Jiaotong University, Beijing, China, 2017.
15. Shen, X.; Zhang, F.; Lv, H.; Liu, J.; Liu, H. Prediction of entering percentage into expressway service areas based on wavelet neural networks and genetic algorithms. *IEEE Access* **2019**, *7*, 54562–54574.
16. Sun, C.; Lyu, H.; Yang, R.; Wei, Z.; Hao, X.; Pei, L. Traffic volume prediction for highway service areas based on XGBoost model with improved particle swarm optimization. *J. Beijing Jiaotong Univ.* **2021**, *45*, 74–83.
17. Wang, J.; Zhang, K.; Zhi, P.; Xi, G.; Tang, X.; Zhou, Q. Analysis and Prediction Transient Population in Expressway Service Area based Long Short-Term Memory. In Proceedings of the 2021 IEEE 23rd Int Conf on High Performance Computing & Communications; 7th Int Conf on Data Science & Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys), Haikou, China, Dec. 20 2021 to Dec. 22 2021; IEEE: New York, NY, USA, 2021; pp. 1828–1832.
18. Zhao, J.; Yu, Z.; Yang, X.; Gao, Z.; Liu, W. Short term traffic flow prediction of expressway service area based on STL-OMS. *Phys. A Stat. Mech. Its Appl.* **2022**, *595*, 126937.
19. King, G. *NCHRP Report 324: Evaluation of Safety Roadside Rest Areas*; TRB, National Research Council: Washington, DC, USA, 1989.
20. Expressway Technical Research Institute. *Expressway Traffic Statistics Data Book.* Expressway Technical Research Institute: Machida, Japan, 2014, pp. 166–185.
21. Umayahara, A.; Akagi, T.; Suzuki, H. Study on The State of Rest Behavior at Expressway Rest Areas. *J. Archit. Plan. (Trans. AIJ)* **2017**, *82*, 1639–1647.
22. Perfater, M. Operation and motorist usage of interstate rest areas and welcome centers in Virginia. *Transp. Res. Rec.* **1989**, 1224, 46-53.
23. Twardzik, L.; Haskell, T. *Rest Area/Roadside Park Administration, Use and Operations*; Michigan State University: Lansing, Michigan, 1985.
24. Al-Kaisy, A.; Church, B.; Veneziano, D.; Dorrington, C. Investigation of Parking Dwell Time at Rest Areas on Rural Highways. *Transp. Res. Rec.* **2011**, *2255*, 156–164.

25. Li, Z.; Miao, Q.; Shehzad, A.; Chen, C. A provably secure and lightweight mutual authentication protocol in fog-enabled social Internet of vehicles. *Int. J. Distrib. Sens. Netw.* **2022**, *18*, 15501329221104332.
26. Wu, J.; Wei, M.; Srivastava, G.; Chen, C.; Lin, J. Mining large-scale high utility patterns in vehicular ad hoc network environments. *Trans. Emerg. Telecommun. Technol.* **2020**. https://doi.org/10.1002/ett.4168.
27. Liao, L.; Lin, J.; Zhu, Y.; Bi, S.; Lin, Y. A Bi-direction LSTM Attention Fusion Model for the Missing POI Identification. *J. Netw. Intell.* **2022**, *7*, 161–174.
28. He, Y.; Qin, Q.; Josef, V. A pedestrian detection method using SVM and CNN multistage classification. *J. Inf. Hiding Multimed. Signal Process.* **2018**, *9*, 51–60.
29. Hirai, S.; Xing, J.; Kobayashi, M.; Ryota, H.; Nobuhiro, U. Preliminary analysis on the resting behavior of expressway users with ETC data. In Proceedings of the 22nd ITS World Congress, Bordeaux, France, 5–9 October 2015.
30. Hirai, S.; Xing, J.; Kai, S.; Ryota, H.; Nobuhiro, U. Study of resting behavior on inter-urban expressways using ETC 2.0 probe data. In Proceedings of the 23rd ITS World Congress, Melbourne, Australia, 10–14 October *2016*; pp. 10–14.
31. Muramatsu, T.; Oguchi, T. Proposal and application of parking area performance measurement methodology. *Transp. Res. Procedia* **2016**, *15*, 628–639.
32. Jin, S. Study of the Forecast for the Volume of Freight Handled in Freeway Service Area. *Sci. Technol. Eng.* **2008**, *8*, 3233–3237.
33. Choe, Y.; Baek, S. A Study on Proper Size of Expressway Service Area. *J. Korean Soc. Transp.* **2009**, *27*, 7–18.
34. Kim, J.; Yang, T.; Yoon, T. Evaluation of Current Service Areas' Location to Improve Freight Car's Safety: A Case Study of Seohaean Expressway. *J. Korean Soc. Transp.* **2020**, *38*, 335–345.
35. Tint, T.; Maung, H.; Wyityi, W. Comparative Analysis on Highway Rest Centers Along Yangon-Mandalay Expressway. *Int. J. Emerg. Technol. Adv. Eng.* **2013**, 3, 41-48.
36. Wang, S.; Xu, Y. Calculation of reasonable scale of parking spaces in expressway service area based on queuing theory. *J. Highw. Transp. Res. Dev.* **2016**, 3, 116–119.
37. Kumar, V.; Kumar, R.; Jangirala, S.; Kumari, S.; Kumar, S.; Chen, C. An enhanced RFID-based authentication protocol using PUF for vehicular cloud computing. *Secur. Commun. Netw.* **2022**, *2022*, 8998339.
38. Liu, L.; He, D.; Ma, Y.; Zhang, X.; Huang, J.; Li, J.; Yao, J. A novel license plate location method based on deep learning. *J. Netw. Intell.* **2020**, *5*, 93–101.
39. Luo, S.; Zou, F.; Zhang, C.; Tian, J.; Guo, F.; Liao, L. Multi-View Travel Time Prediction Based on Electronic Toll Collection Data. *Entropy* **2022**, *24*, 1050.
40. Khanh, H.N. Classification of Concepts Using Decision Trees for Inconsistent Knowledge Systems Based on Bisimulation. *J. Inf. Hiding Multimed. Signal Process.* **2021**, *12*, 22–30.
41. Zhi, Y.; Zhang, J.; Shi, Z. Research on Design of Expressway Acceleration Lane Length and Merging Model of Vehicle. *China J. Highw. Transp.* **2009**, *22*, 93–97+115.
42. Zou, F.; Guo, F.; Tian, J.; Luo, S.; Yu, X.; Gu, Q.; Liao, L. The Method of Dynamic Identification of the Maximum Speed Limit of Expressway Based on Electronic Toll Collection Data. *Sci. Program.* **2021**, *2021*, 4702669.