Hindawi
Journal of Advanced Transportation
Volume 2022, Article ID 8811139, 13 pages
https://doi.org/10.1155/2022/8811139

Research Article

A Simulation Approach to Detect Arterial Traffic Congestion Using Cellular Data

Shen Li,[1] Jian Zhang ,[2,3] Gang Zhong ,[4] and Bin Ran[3]

[1] *Department of Civil Engineering, Tsinghua University, Beijing 100084, China*
[2] *Engineering College, Tibet University, Lhasa 850000, China*
[3] *School of Transportation, Southeast University, #2 Southeast University Road, Nanjing 210096, China*
[4] *College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, #29 Jiangjun Avenue, Nanjing 211106, China*

Correspondence should be addressed to Jian Zhang; jianzhang@seu.edu.cn

Received 2 March 2020; Accepted 17 November 2021; Published 21 February 2022

Academic Editor: Zhixiang Fang

Cellular data provide a promising way for congestion detection with low cost and high coverage, and the simulation study is a feasible solution to verify the detection method. This paper presents a simulation approach that uses cellular data to detect traffic congestion on urban arterials based on the relationship between cellular data and traffic status. The virtual testbed, which includes three main modules, is developed to perform the cellular activities generation, collection, and aggregation process between cell phones and cell stations. Then, the correlation between cellular data and traffic status data is studied. Finally, three scenarios using the data from testbed are demonstrated to measure the performance of the proposed method under different conditions. The results indicate that the proposed approach is a feasible and efficient way to simulate cellular data generation, collection, and aggregation process. Also, it can be the base for further analysis to detect traffic congestion on arterials using cellular data.

1. Introduction

Recently, congestion has become a critical road traffic problem around the world. Detecting traffic status effectively is a key point to reduce congestion. There are a variety of mature techniques available for gathering raw data for congestion detection, such as loop detector and probe vehicle. [1–6]. During the last decade, cellular data generated between cell phones and cellular stations shows the potential that it is a promising data source to provide features indicating traffic congestion with better coverage and less cost [7–13].

Using real cellular data to build the congestion detection model directly without validation may not find the appropriate correlation between cellular data and traffic congestion because the complex data collection environment and the huge data sample size, especially for urban arterials [14–16]. Traffic simulation is an indispensable instrument in the field of the transportation research and engineering [17–20]. These techniques are normally used in freeways with different data resources. Bauza and Gozalvez proposed a large-scale congestion detection model based on a unique open-source simulation platform [21]. Cárdenas et al. published a paper to manage the congestion on freeways between two cities using VISSIM simulator [22]. The previous studies proved that simulation approach can be the base of congestion detection investigation. Meanwhile, some researchers also tried to apply the simulation approach to the cellular data to validate the congestion detection models. Zhang et al. published a paper to detect traffic congestion on freeways using the cellular data from freeway segments under simulation environment [23]. Yang et al. presented a method to analyze the traffic congestion using a hybrid traffic-and-wireless simulation network based on handoff data [24]. However, analyzing the cellular data to detect traffic congestion on arterials under a microscopic-level testbed has not been studied. Moreover, the data beside the handoff in the cellular data, and how to simulate these cellular data is the critical problem that needs to be discussed.

This paper presents a simulation approach that uses cellular data to detect traffic congestion on urban arterials based on the relationship between cellular data and traffic status. In this study, we first analyze the procedures of generating, collecting, and aggregating cellular data between cell phones and cell stations. A virtual testbed is established with the VISSIM traffic simulation tool. The traffic data from the city of Taicang (China) are used as inputs to the virtual test bed, including traffic signal phase and road geometry. The virtual testbed, which includes three main modules, that is, event generator, collector, and aggregator, is developed using the VISSIM COM APIs. The event generator module is accountable for generating all cellular communication activities between the cell phones on vehicles and the cellular stations in the simulated traffic road network. The event collector module running in the background will keep collecting all the cellular activities generated by the event generator. An independent event aggregator module aggregates and persists the desired statistics. The testbed performs the cellular activities generation, collection, aggregation process between cell phones and cell stations. Two kinds of major activities (i.e., location update and handoff) between cell phones and cell stations are generated. The handoff data are generated when a cell phone is passing the boundary of two adjacent cell stations, and the location update data are used to describe cell phone transmit data with the stations which includes location update, on/off cell phone, and texting/phone call. Then, based on the data output from the proposed testbed, the correlation between cellular data and traffic status data to detect traffic congestion is studied using support vector machine (SVM) algorithms with joint mutual information (JMI) feature-selection method. Furthermore, the performances in three scenarios including recurring traffic congestion, nonrecurring traffic congestion, and a small penetration rate of cell phones are measured. To validate these different scenarios, the data from morning peak hours and evening peak hours are selected for the recurring traffic congestion condition, one or two lanes will be closed on a designated arterial road during the certain period for the nonreccuring traffic congestion condition, and the percentage of cell phones that can transmit data with cell stations will be reduced for the small penetration rate condition. The results indicate that the proposed approach is a feasible and efficient way to simulate cellular data generation, collection, and aggregation process. Also, it can be the base for further analysis to detect traffic congestion on arterials using cellular data.

2. Virtual Testbed Design

Traffic simulation is very important in the field of the transportation research and engineering. In this study, a VISSIM-based simulation platform was developed to simulate the traffic network and the cellular communications among the cellular network. The platform, which includes three main modules, that is, event generator, collector, and aggregator, is developed on top of the VISSIM COM APIs.

2.1. Environment Setup. The proposed research site is a subdivision of the urban arterial networks in the city of Taicang, 30 miles northwest of Shanghai, see Figure 1. More specifically, research will focus on the two major arterial roads in Taicang, Xian and Shanghai road, along with the upstream, downstream, and minor roads. See Table 1 for the roads in the simulated arterial network.

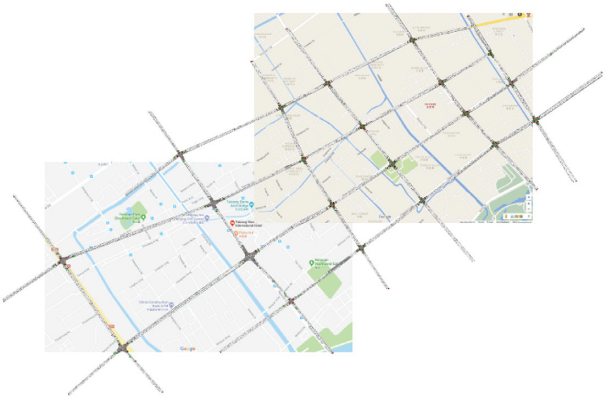


Figure 1: Simulated arterial network in the VISSIM.

In total, 12 intersections are chosen as the traffic volume collecting intersection in the simulation network. Those interactions are located in the downtown Taicang, which is the busiest area of the city. To collect traffic volumes, we put a data collection point on each lane of those intersections. See Table 2 for detailed information of the data collection points.

In the examined area, 98 cell stations are allocated according to the location of real-word cellular stations. See Figure 2 for the distribution of these cell stations. We map the cellular stations to a set of road network links. Each link is covered by one and exactly only one cell station.

2.2. Virtual Testbed Work Flow. The cellular event generator is the engine of the simulation, and all business logics reside there. First, it loads the simulation setup files to initialize the road and cellular network. Later, it proceeds to start the VISSIM in discrete simulation mode and generates the vehicle movements and the cellular events accordingly.

Figure 3 roughly summarizes the simulation process as follows: it begins with the event generator to generate cellular communication events and randomly assign the events to vehicles while the vehicle is running on the road network. The event collector module is running in the background to keep collecting and processing the events. Finally, the cellular data aggregator calculates the statistics and persist them to DB, HDFS, or in-memory storage. For future research, it could also be used as online aggregation, machine learning, and visualization.

During the events generation, two types of the cellular events are randomly generated (i.e., LU or Handoff) and attached to the running vehicles. The time step and lifetime of the events are randomly distributed as well.

The simulator will simulate 24-hour traffic to cover the morning and evening peak hour and the night time traffic. See Table 3 for the hourly traffic volume in different periods of a day.

VISSIM is a microscopic, behavior-based multipurpose traffic simulation platform to analyze and optimize traffic

Table 1: Roads included in the simulated arterial network.

North-south roads	East-west roads
G204	Zhenghe Road
Renmin Road	Xianfu Street
Taiping Road	Shanghai Road
Dongcang Road	Chaoyang Road
Dongting Road	
Loujing Road	

Table 2: Data collection points distributed at the examined intersections in the simulated arterial network.

No.	Name	Data collection points
1	XianFu@204	1–12
2	Shanghai@204	13–24
3	Xianfu@Renmin	25–38
4	Shanghai@Renmin	39–48
5	Xianfu@Taiping	49–63
6	Shanghai@Taiping	64–79
7	XianFu@Dongcang	80–95
8	Shanghai@Dongcang	96–111
9	Xianfu@Dongting	112–127
10	Shanghai@Dongting	128–143
11	Xianfu@Loujiang	156–164
12	Shanghai@Loujiang	144–155

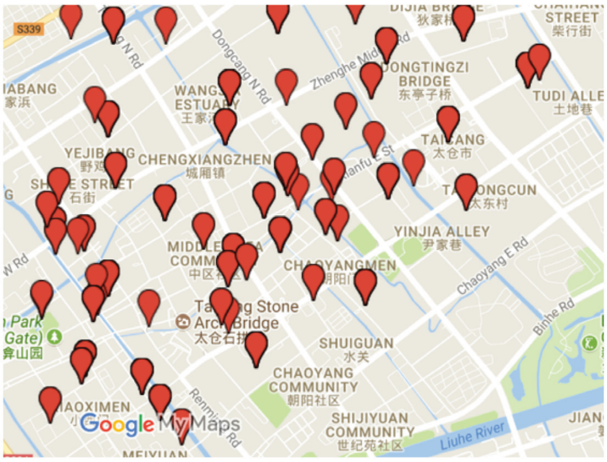


Figure 2: Cell station distribution of Taicang.

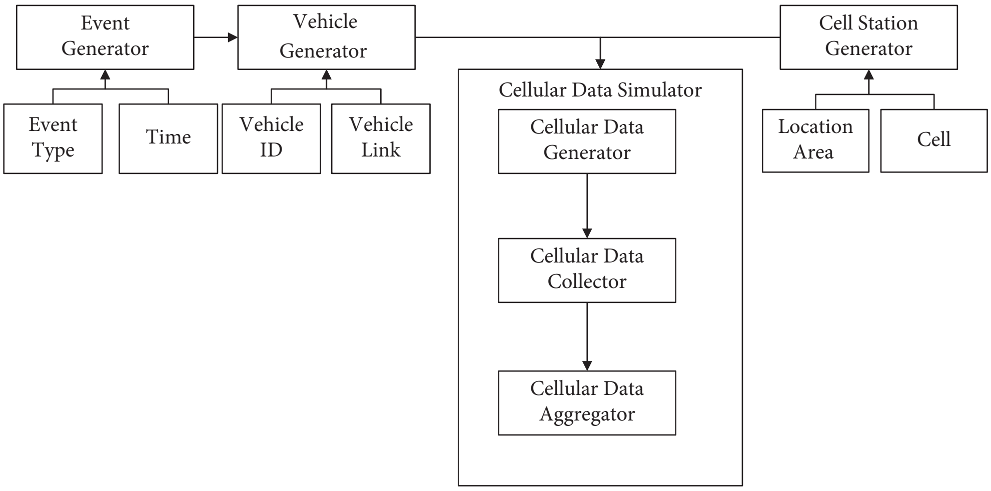


Figure 3: The framework of the simulation platform.

Table 3: Example vehicle inputs in the VISSIM.

Link	Volume (6:00–7:00)	Volume (7:00–8:00)	Volume (8:00–9:00)	Volume (9:00–10:00)	Volume (10:00–11:00)
1	1000	1000	1000	750	750
2	1000	1000	1000	500	500
3	1000	1000	1000	500	500
4	500	1000	1000	500	500

flows. This simulation platform adapts the lateral and longitudinal fundamental core models as the underlying vehicle dynamics mathematical models for the car-following and lane-change behaviours. Meanwhile, it adapts the dynamic traffic assignment (DTA) models for the underlying route choice behaviours.

3. Data Processing Framework

After the simulation, two major data sets (traffic status data and cellular data) need to be organized as appropriate data format. Traffic status data need to be labeled as the baseline to divide the whole data set into three levels, which will be used to train the simulated cellular data.

3.1. Simulation Data Processing. Two data sets are generated using VISSIM simulation platform—the cellular events that are collected by the event collector and the traffic volumes that are collected by the loop detectors. We use the cellular events as the sole data source for the detection of the traffic conditions and the traffic volumes as the ground truth to calibrate and train the statistical models.

3.2. Traffic Volume Data. Traffic volume data generated by VISSM is the key to help build the ground truth. The simulation platform collects the following four fields from raw traffic volume data, from which the traffic status can be determined (Table 4). Data collector id represents the unique id for a data collector. T (Entry) represents the time when the vehicle enters the collector, and T (Exit) represents the time when the vehicle leaves the collector. The last column represents the unique id for a vehicle.

3.3. Cellular Events Data. Cellular events data are the only data source for the congestion detection. The simulation platform generates the similar format as the real-world cellular events. See Table 5 as an example.

3.4. Traffic Density Labeling. The study uses the data collected from each loop detector every 30 minutes as a data sample; 264 samples were extracted from the original data for further analysis. To simplify this issue, the study uses the data from VISSIM simulator convert to the traffic density parameter. To get the ground truth labels of the traffic density for each data sample, the traffic data distribution of the collected data samples was analyzed to categorize each data sample into the low-, medium-, and high-density group. Four thresholds were used to categorize the data samples to determine which one is used to separate traffic status of the designed classifier, including 10%, 20%, 30%, and 40%. Using the 20% as an example, if the detected vehicle number were less than the 20-percentile number, the data sample would be labeled as low-density (Figure 4). Otherwise, if the detected vehicle numbers were more than the 80-percentile number, the data sample would be labeled as high-density, that is, traffic congestion.

3.5. JMI-Based SVM Classifier. After the collection of cellular data, a preliminary data processing is executed to extract features from the raw cellular data. The study will use an optimal feature-selection model to find the most relevant features among handoff events, location update, text message, and phone call. To achieve this goal, a candidate feature set will be extracted from the original cellular traffic volume data, and an algorithm based on joint mutual information (JMI) will be applied to select a subset from the candidate features. The method that proposed by Brown et al. [25]. The basic idea of the method is to maximize a conditional likelihood considering the mutual information between features.

Then, we use SVM classification in two passes to first classify the data into low-level traffic condition and other conditions, and then the second pass classifies the other cellular data to medium and high condition.

Using the SVM as the classifier in the condition detection on arterials, it yields relatively good accuracy and low computation cost, in a near real-time manner. Meanwhile, overfitting is one of the key issues for the statistical machine-learning models; the next section will address this issue.

4. Scenario Design

Different scenarios of the proposed approach need to be validated. First, the JMI-based feature extraction and classification algorithm need to be proved that it can be applied to traffic condition detection on arterials using cellular data as the sole data source. Second, the proposed method needs to be tested to see its performance under different conditions.

The study validates the model in three conditions: recurring traffic congestion, nonrecurring traffic congestion, and a small penetration rate of cell phones (Figure 5). To validate recurring congestion condition, the simulator will generate data from morning peak hour and evening peak hour. For the nonrecurring condition, we will simulate it by closing one or two lanes on a designated arterial road during the certain period. Finally, with a small penetration rate of cell phones (a.k.a less cell phone events), the proposed model will be applied to see how it performs.

Table 4: Example of raw traffic status data from VISSIM.

Data collector ID	T (entry)	T (exit)	Vehicle ID
13	60.37	60.69	1
13	68.58	68.84	43
23	114.56	114.83	16

Table 5: Example of cellular data from VISSIM.

MSID	LA_TO	CELL_TO	LA_FROM	CELL_FROM	TYPE	TIME
1645	20,969	5415	20,969	25,425	LU	10
1761	20,822	54,672	20,822	25,487	LU	10
1983	20,696	54,271	20,969	54,273	HO	10

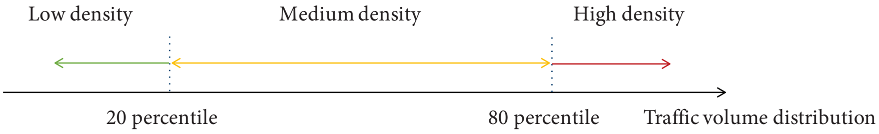


Figure 4: An illustration of traffic density labeling.

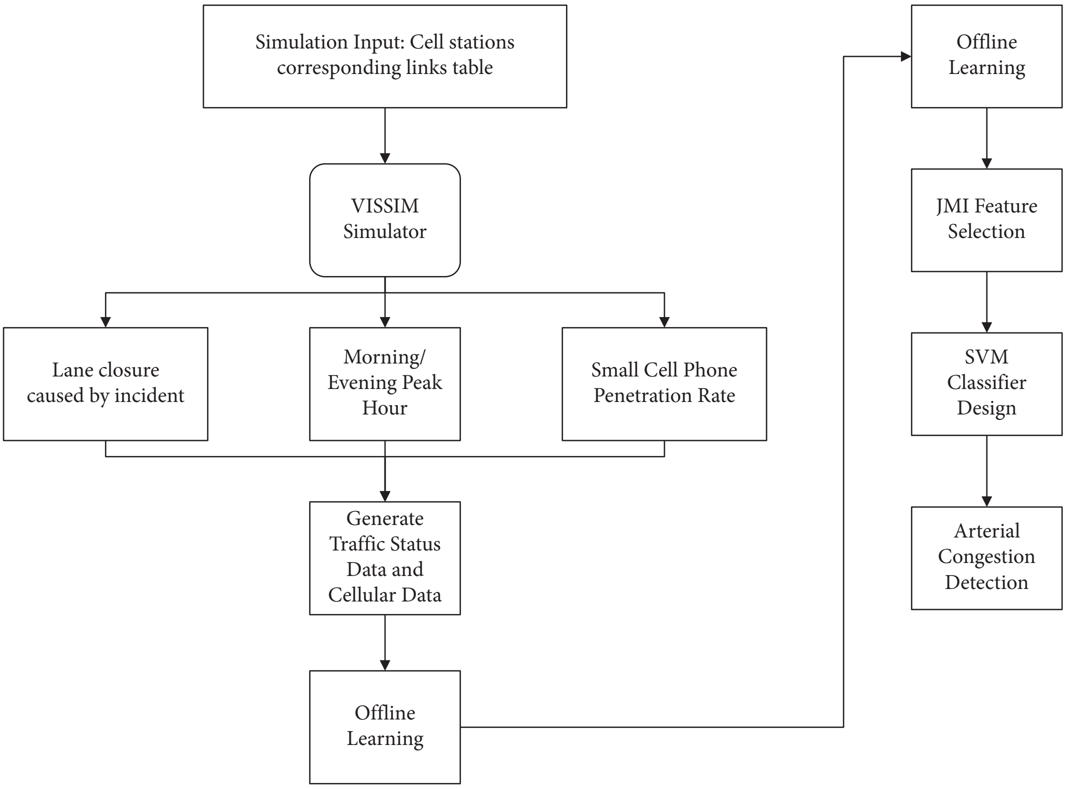

Simulation Input: Cell stations corresponding links table
VISSIM Simulator
Lane closure caused by incident
Morning/ Evening Peak Hour
Small Cell Phone Penetration Rate
Generate Traffic Status Data and Cellular Data
Offline Learning
Offline Learning
JMI Feature Selection
SVM Classifier Design
Arterial Congestion Detection

Figure 5: Flow chart of model validation process.

4.1. Recurring Traffic Congestion. Recurring congestion is generally the consequence of factors that act regularly or periodically on the traffic system, such as morning peak hour and weekend out-of-town and back-in-town trips. The recurring congestions happening in the morning and evening peak hours typically are caused by the traffic demand increasing beyond the supply. The main characteristics of recurring congestion are that the location of the congestion has strong upstream and downstream relations. Meanwhile, multiple congestion locations can occur simultaneously in the network.

The main purpose of this validation is to prove the correctness and the adaptability of the proposed method. The basic setting of the simulation environment is as follows:

(i) 80% vehicle owns a cell phone and all cell phones are under the power on situation

(ii) 40% vehicles driving on roads make at least one phone call

(iii) The traffic simulation contains both morning peak hour and evening peak hour

4.2. Nonrecurring Traffic Congestion. Another condition needs to be discussed is the congestion caused by accidents in nonpeak hours. The main reason is that this type of congestions is different from the peak hours on account of the underlying mechanisms and network dynamics. Unexpected, unplanned, or large events cause nonrecurrent congestion (e.g., work zone and crashes) that will impact the parts of traffic system randomly and, as such, cannot be easily predicted or modeled. There are many reasons that are responsible to generate nonrecurring congestion. Incidents, of course, have a major role to play in it. Crashes, vehicle breakdown, bad weather, special events, and work zones are all examples that can give rise to extreme congestions [26].

Analyzing nonrecurrent congestion is important for several reasons. First, nonrecurring congestions were defined as unexpected or unusual congestion caused by an unexpected incident and typically the impact period is transient, which means the model that is used to handle recurring congestion may have a problem to deal with the nonrecurring ones. Second, the share of nonrecurring congestion is relativity high among all congestions, FHWA estimates the share as high as 55% [7]. Third, even under low traffic volume condition, none of the events may result in congestion but along with the increase of the traffic demand, congestion can easily result from events. Reducing the impacts of these unexpected events during moderate to high-volume conditions is one of the major goals of traffic management systems [27].

The congestion caused by accidents is an extension process that usually spreads from the downstream to the upstream, from the point to the surface. Good adaptability is achieved, if the proposed model can capture both the recurring and nonrecurring congestions.

4.3. Traffic Congestion in Low Cell Phone Penetration Rate. The last condition the study needs to validate is that whether the model can still perform well in the scenario of small penetration rate of cell phone usage. The approach of the study is the data-driven one, so the quantity and quality of data heavily affect the model.

Small sample size caused by small penetration rate of a cell phone or law enforcement that prohibits the usage of the cell phones on the road may make an impact to the performance of the model. Small sample size may create the sampling error, a critical problem to the proposed method. Estimating the impact of random sampling needs to be considered as well. This problem is not only associated with simulated data, it may also occur in real data.

To verify these questions, the study needs to validate model under small penetration of the cell phone. The basic parameters of the condition are as follows:

(i) 50% vehicles own a cell phone, and all cell phones are under the power on situation

(ii) 10% vehicles driving on roads make at least one phone call

4.4. Performance Measures. The result of the congestion detection based on the SVM algorithm is measured by the classic confusion matrix (Table 6).

True positive (TP) represents a positive sample that is correctly classified as a positive one. False negative (FN) demonstrates a positive data sample that is classified as a negative one. False positive (FP) means that a negative data sample is classified as a positive one, and the true negative (TN) indicates the number of negative cases correctly classified as negative ones.

The number of samples that are correctly classified can be described by a classification rate which is an important measure to indicate the performance of the classifier.

$$\text{classification rate} = \frac{\text{\# of TP cases} + \text{\# of TN cases}}{\text{\# of all cases}}. \tag{1}$$

The study also introduces ROC curve to illustrate the relationship between TPR and FPR at different thresholds.

$$\begin{aligned} \text{true positive rate (TPR)} &= \frac{\text{\# of TP cases}}{\text{\# of TP cases} + \text{\# of FN cases}}, \\ \text{false positive rate (FPR)} &= \frac{\text{\# of FP cases}}{\text{\# of FP cases} + \text{\#of TN cases}}. \end{aligned} \tag{2}$$

The more a curve in the ROC space bends to the up-left corner, the better the classification performance of the classifier is (Figure 6). To quantitatively describe the ROC curve, the machine-learning community usually employs the area under the curve (AUC) statistic for model comparison. A higher AUC value indicates a closer bending to the up-left corner of the ROC curve, which proves better performance of the designed classifier.

The description on the five-fold cross-validation method is illustrated in Figure 7. In the five-fold cross-validation, the samples are divided into five subsets of equal size. Sequentially, one subset is tested using the classifier trained on the remaining four subsets. Thus, each subset will be predicted once, so the cross-validation accuracy is the percentage of data which were classified correctly. The cross-validation procedure could prevent the overfitting problem.

The performance of an SVM classifier with Gaussian kernel relies on the selection of regularization parameter C and kernel spread gamma (γ). The best combination of C and γ was selected using a grid search method. As reported by Chang and Lin, exponentially growing sequences of C and γ is a practical method to find a narrower range where the best pair of (C, γ) values may exist (for example, $C = 2^{-5}, 2^{-3}, \ldots, 2^{5}, \gamma = 2^{-5}, 2^{-3}, \ldots, 2^{5}$) [15]. This is called

Table 6: The confusion matrix.

		Predicted label	
		Positive	Negative
True label	Positive	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

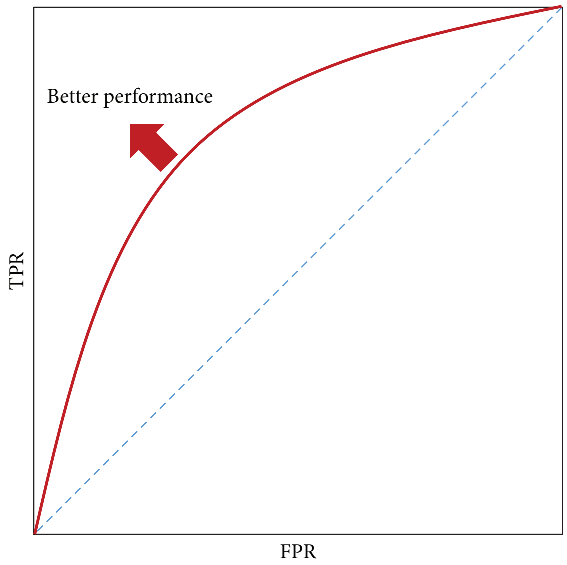
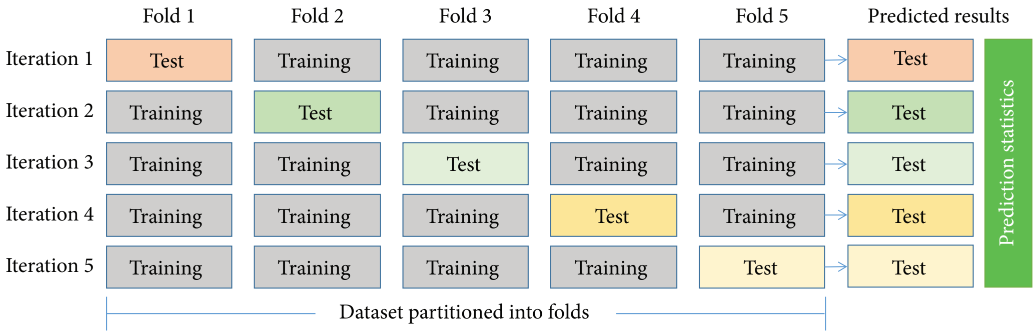


Figure 6: An illustration of an ROC curve.



	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Predicted results
Iteration 1	Test	Training	Training	Training	Training	Test
Iteration 2	Training	Test	Training	Training	Training	Test
Iteration 3	Training	Training	Test	Training	Training	Test
Iteration 4	Training	Training	Training	Test	Training	Test
Iteration 5	Training	Training	Training	Training	Test	Test

Dataset partitioned into folds

Prediction statistics

Figure 7: Description on the five-fold cross-validation.

the loose grid search. When the narrower range is detected, the grid search method would be applied repeatedly to find an even narrower range till the best pair of (C, γ) values is found, which is called the fine grid search.

After the best pair of (C, γ) values is selected, the study will process the data using the algorithm to divide the whole cellular data into three traffic statuses.

5. Results and Discussions

5.1. Recurring Traffic Congestion. To select an optimized subset from the 294 candidate features, a conditional likelihood maximization method based on mutual information was employed. See Figure 8 for the relationship between the number of selected features and the classification accuracy when selecting different labeling thresholds. Results show that the highest classification accuracies are 88.9%, 94.8%, 85.0%, and 83.7% for the labeling thresholds of 10%, 20%, 30%, and 40%, respectively. Using 20% labeling threshold, the number of features are five where the classification accuracy achieved is 94.8%.

The five selected features include: on-call mobile phone number (N_Handoff) at the nearest station, power-on mobile phone number (N_LU) at the nearest station, total mobile phone number (N_Total) at the second-nearest station, on-call mobile phone number (N_Handoff) at the second-nearest station, and power-on mobile phone number (N_LU) at the third-nearest station. Among the five selected features, two of them are from the mobile phone traffic volume features about the nearest station, another two from the second-nearest station, and the other one from the third-nearest station.

We will focus on discussing the simulation results with the labeling threshold at 20% in this paper.

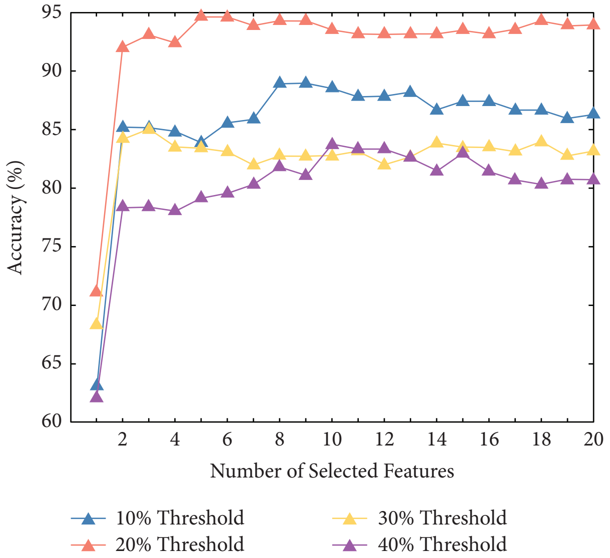


Figure 8: Relationship between the number of selected features and the classification accuracy.

Based on the selected features, the grid search method is employed again to find the best pair of (C, γ) values. Results show that when $C = 4.5$ and $\gamma = 0.5$ using the 20% labeling threshold, the classifier achieves its best performance. See Figure 9 for the fine grid search results on (C, γ) values.

See Table 7 for the confusion matrix of the traffic density classification results based on the selected feature, labeling with 20% threshold. The overall classification rate is 94.8%. The correct classification rates for the low-, medium-, and high-density groups are 80.7%, 98.7%, and 97.3%, respectively. About 13.7% of the low-density samples are classified into the medium group. The number of medium-density samples classified into the low group is four, accounting 1.0% of all the medium samples. As for the high-density samples, 97.3% of the samples are recognized correctly.

Taking the low-, medium-, or high-density samples as positive samples separately, the corresponding TPR and FPR values could be computed. The ROC curves and AUC values are illustrated in Figure 10. The AUC for the medium-density group is 0.98. All numbers are 0.99 for the other traffic density groups. The high AUC values show that the classifier performance based on the optimized subfeature set selected from the 294 features is satisfactory.

5.2. Nonrecurring Traffic Congestion. Nonrecurring is different from the peak hours on account of the underlying mechanisms and network dynamics. Thus, the model that is used to handle recurring congestion may have a problem to deal with the nonrecurring ones. This section discusses the proposed model performance under nonrecurring traffic congestion condition.

A conditional likelihood maximization method based on mutual information is employed. See Figure 11 for the relationship between the number of selected features and the classification accuracy. Results show that the highest classification accuracy is 81.8% at the 20% labeling thresholds, and the number of features is six when the highest classification accuracy is achieved.

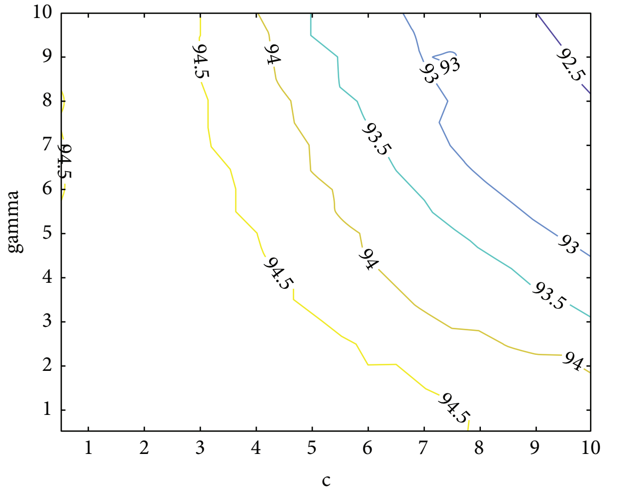


Figure 9: Using the grid search method to find the best pair of (C, γ) values based on the selected features.

The six selected features include: total mobile phone number (N_Total) at the nearest station, power-on mobile phone number (N_LU) at the nearest station, total mobile phone number (N_Total) at the third-nearest station, total mobile phone number (N_Total) at the second-nearest station, power-on mobile phone number (N_LU) at the fourth-nearest station, and power-on mobile phone number (N_LU) at the third-nearest station. Among the six selected features, three of them are from the nearest two stations. The other three are from the stations farther.

Based on the selected traffic volume features, the grid search method is employed again to find the best pair of (C, γ) values. Results show that when $C = 1$ and $\gamma = 1.5$ using the 20% labeling threshold, the classifier achieves its best performance. See Figure 12 for the fine grid search results on (C, γ) values.

See Table 8 for the confusion matrix of the traffic density classification results based on the selected traffic volume features of mobile phone use when labeling with the 20% threshold. The overall classification rate is 81.8%. The correct classification rates for the low-, medium-, and high-density groups are 69.2%, 82.5%, and 92.3%, respectively. About 15.4% of the low-density samples are classified into the medium group. The number of medium-density samples classified into the low group is five, accounting 12.5% of all the medium samples. As for the high-density samples, 92.3% of the samples are recognized correctly.

The result shows that the accuracy is relatively low under low-density condition. It means higher cellular data may not represent higher traffic density. It may cause by the increasing of nonvehicle cell phone users. As the penetration rate of cell phone users in the vehicle is small, it may not represent the traffic condition status very well. On the contrary, the proposed model performs well under medium- and high-density condition. It means higher traffic density will cause higher cellular data, and the false alarm of the proposed model is low.

Table 7: Confusion matrix of the traffic density classification results based on the selected traffic volume features of mobile phone use.

		Predicted density class		
		Low	Medium	High
True density class	Low	80.7% (171/212)	13.7% (29/212)	5.6% (12/212)
	Medium	1.0% (6/626)	98.7% (618/626)	0.3% (2/626)
	High	1.8% (4/218)	0.9% (2/218)	97.3% (212/218)

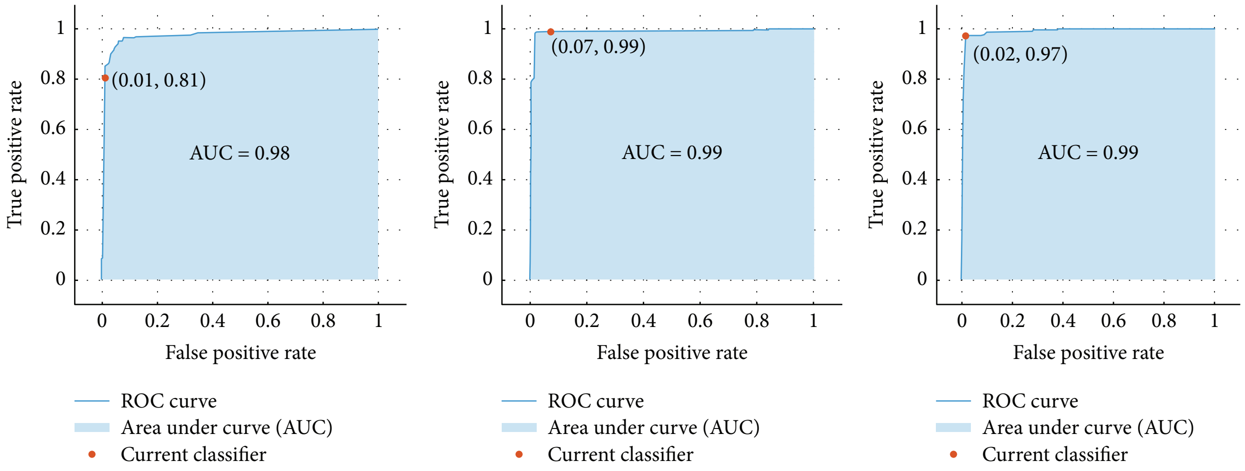


Figure 10: ROC curve AUC to describe the performance of the binary classifier for the low-, medium-, and high-density group based on the selected features.

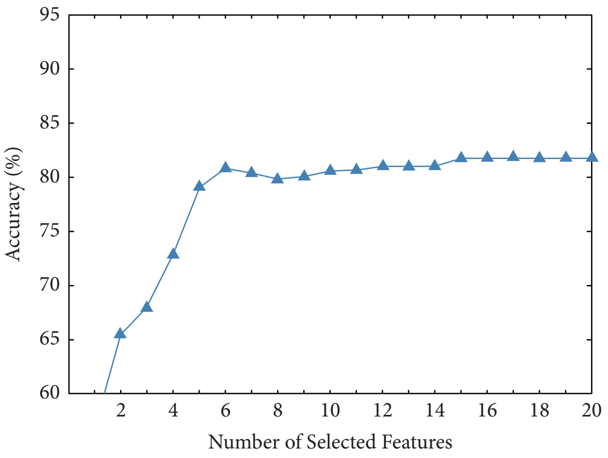


Figure 11: Relationship between the number of selected features and the classification accuracy.

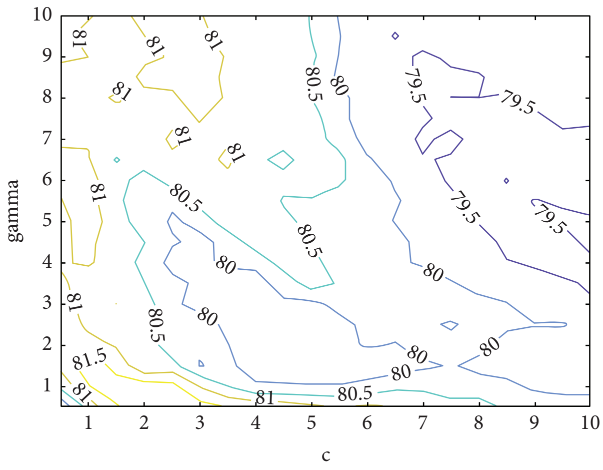


Figure 12: Using the grid search method to find the best pair of (C, γ) values based on the selected features.

Compared with results from recurring delay, the overall accuracy is lower, especially for the low traffic density condition. The results prove that the proposed model can handle medium- and high-density condition well but play poorly in low-density condition under nonrecurring congestion condition.

Taking the low-, medium-, or high-density samples as positive samples separately, the corresponding TPR and FPR values are computed. The ROC curves and AUC values are illustrated in Figure 13. The AUC for the low-, medium-, and high-density groups are 0.83, 0.94, and 0.96.

5.3. Traffic Congestion in Low Cell Phone Penetration Rate. The small sample size may make the sampling error become a critical problem to the proposed method. Estimating the impact of random sampling needs to be considered.

A conditional likelihood maximization method based on mutual information is employed. See Figure 14 for the relationship between the number of selected features and the classification accuracy when selecting different labeling thresholds. Results show that the highest classification accuracies are 79.2%, 83.3%, 76.5%, and 72.3% at the labeling thresholds of 10%, 20%, 30%, and 40%, respectively. Using

Table 8: Confusion matrix of the traffic density classification results based on the selected traffic volume features of mobile phone use.

		Predicted density class		
		Low	Medium	High
True density class	Low	69.2% (9/13)	15.4% (2/13)	15.4% (2/13)
	Medium	12.5% (5/40)	82.5% (33/40)	5% (2/40)
	High	0.0% (0/13)	7.7% (1/13)	92.3% (12/13)

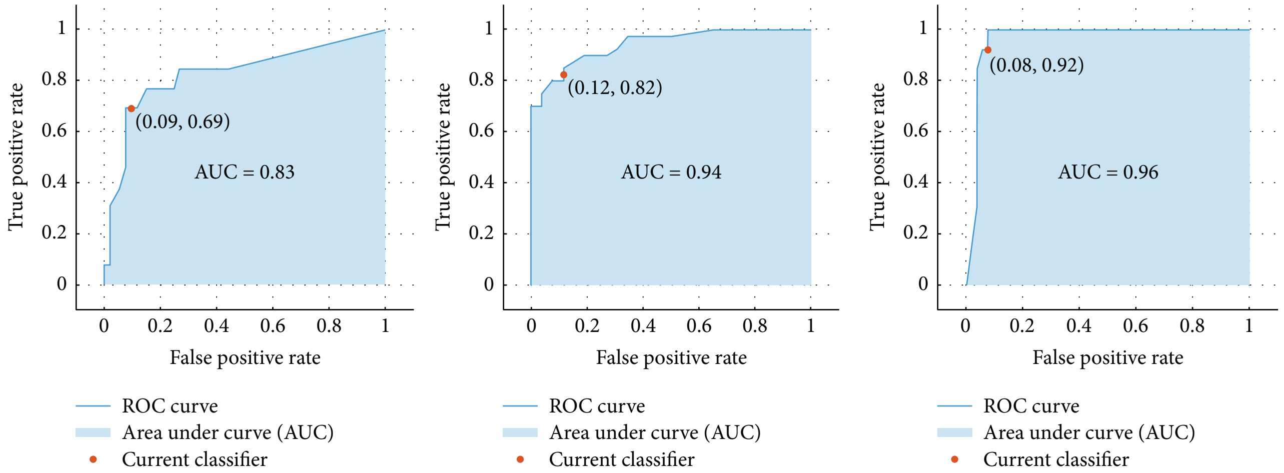


Figure 13: ROC curve AUC to describe the performance of the binary classifier for the low-, medium-, and high-density group based on the selected features.

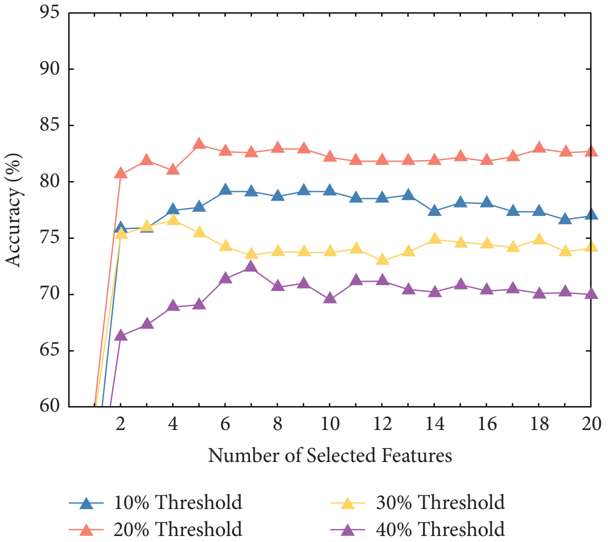


Figure 14: Relationship between the number of selected features and the classification accuracy when labeling with different thresholds.

the 20% labeling threshold, the number of features are also five when the highest classification accuracy is achieved.

The five selected features include: total mobile phone number (N_Total) at the nearest station, total mobile phone number (N_Total) at the second-nearest station, power-on mobile phone number (N_LU) at the fifth-nearest station, on-call mobile phone number (N_Handoff) at the sixth-nearest station, and power-on mobile phone number (N_LU) at the eighth-nearest station. Among the five selected features, two of them are from the mobile phone traffic volume of the nearest two stations. The other three are from the stations farther.

We will focus on discussing the simulation results with the labeling threshold at 20% in all the subsections within Section 5.3. Simulation results with other labeling thresholds will be discussed later.

Based on the selected traffic volume features, the grid search method is employed again to find the best pair of (C, γ) values. Results show that when $C = 4$ and $\gamma = 9.5$ using the 20% labeling threshold, the classifier achieves its best performance. See Figure 15 for the fine grid search results on (C, γ) values.

See Table 9 for the confusion matrix of the traffic density classification results based on the selected traffic volume features of mobile phone use when labeling with the 20% threshold. The overall classification rate is 83.3%. The correct classification rates for the low-, medium-, and high-density groups are 86.8%, 88.6%, and 64.2%, respectively. About 13.2% of the low-density samples are classified into the medium group. The number of medium-density samples classified into the low group is 3, accounting for 7.0% of all the medium samples. As for the high-density samples, 64.2% of the samples are recognized correctly.

The result shows that the accuracy is relatively low under high-density condition. It represents that the false alarm of the proposed model is relatively high, which means higher traffic density may not cause higher cellular data. It may cause by the small penetration of cellular data from on-road users. The changing traffic density may not reflect well from cellular data because the difference between medium density

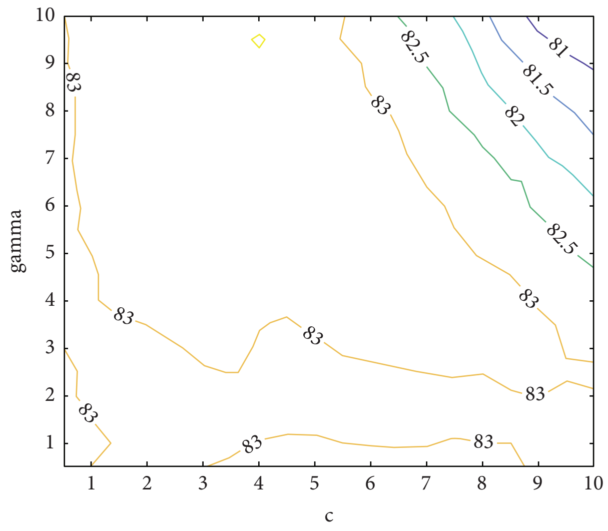


Figure 15: Using the grid search method to find the best pair of (C, γ) values based on the selected features.

Table 9: Confusion matrix of the traffic density classification results based on the selected traffic volume features of mobile phone use.

		Predicted density class		
		Low	Medium	High
True density class	Low	86.8% (46/53)	13.2% (7/53)	0.0% (0/53)
	Medium	4.4% (7/158)	88.6% (140/158)	7.0% (11/158)
	High	0.0% (0/53)	35.8% (19/53)	64.2% (34/53)

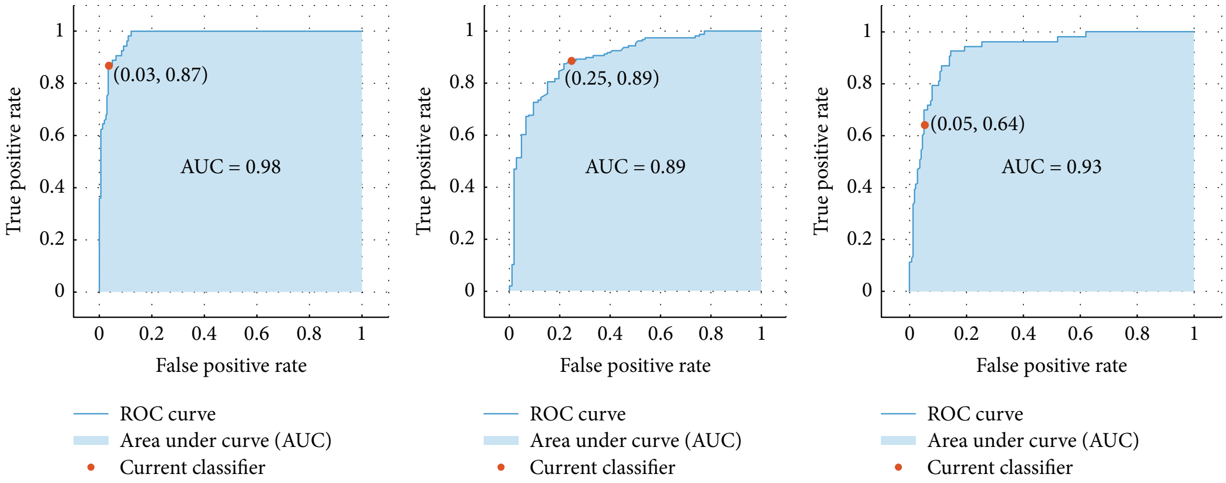


Figure 16: ROC curve AUC to describe the performance of the binary classifier for the low-density group based on the selected features.

and high density may not have enough features to show. On the contrary, the proposed model performs well under low- and medium-density condition.

Compared with results from recurring delay, the overall accuracy is low. Taking the low-, medium-, or high-density samples as positive samples separately, the corresponding TPR and FPR values are computed. The ROC curves and AUC values are illustrated in Figure 16. The AUC for the low-, medium-, and high-density groups are 0.98, 0.89, and 0.93, respectively.

6. Conclusion

This paper presented a simulation approach to detect traffic congestion on urban arterials using cellular data. Based on determining the procedures of cellular data generation, collection, and aggregation, a microscopic-level virtual testbed was established. The correlation between cellular data and traffic status data generated from testbed was studied to detect traffic congestion using support vector machine (SVM) algorithm with joint mutual information

(JMI) feature selection method. Some typical scenarios were selected to validate the designed testbed. The performances from different conditions showed that the proposed simulation approach can detect traffic congestion effectively. Some conclusions are obtained as follows:

(1) A virtual testbed on the VISSIM COM APIs to simulate cellular data and traffic data of the arterial networks is developed. The testbed is useful to output cellular data and corresponding traffic status data under microscopic-level condition by considering the complicated arterial and cellular networks.

(2) Three scenarios are designed to describe different conditions of traffic congestion. The validation process proved the proposed testbed can detect traffic congestion appropriately on urban arterial networks.

(3) The results show that both LU and HO features from virtual testbed were effective for traffic status detection, and the cellular events from multinearby stations could be promising indicators for traffic status detection.

Our future work will try to explore more validation scenarios to improve the performances of the simulation method. Meanwhile, the detection approach can be extended to the scenarios of other countries, whereas the real traffic situations of China are not the same as those in the United States and Europe. In this case, more refined simulation framework should be established to model the interaction behaviour between vehicles and other traffic participants [28].

Data Availability

The data used to support the findings of this 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 the Key R&D Program of Jiangsu Province in China (grant no. BE2020013), the Industrial technology basic public service platform construction project (grant no. CEIEC-2020-ZM02-0100), and the China Merchants Testing Vehicle Technology Research Institute Co., Ltd. The work of the corresponding author was supported in part by the Chongqing Engineering Research Center of Research and Testing for Automated Driving System and Intelligent Connected Vehicle and the R&D project of China Communications Construction Company (grant no. 2019-ZJKJ-ZDZX02-2).

References

[1] Transport Research Centre, *Managing Urban Traffic Congestion*, Organisation for Economic Co-operation and Development and European Conference of Ministers of Transport. OECD, MI, USA, 2007.

[2] P. J. Bickel, C. Chen, J. Kwon, J. Rice, E. V. Zwet, and P. Varaiya, "Measuring traffic," *Statistical Science*, vol. 22, pp. 581–597, 2008.

[3] D. Work, S. Blandin, O. P. Tossavainen, B. Piccoli, and A. Bayen, "A distributed highway velocity model for traffic state reconstruction," *Applied Research Mathematics eXpress (ARMX)*, vol. 1, pp. 1–35, 2010.

[4] R. C. Long, C. Xie, and D. H. Lee, "Probe vehicle population and sample size for arterial speed estimation," *Computer-Aided Civil and Infrastructure Engineering*, vol. 17, no. 1, pp. 53–60, 2002.

[5] S. Li, G. Li, C. Yang, and B. Ran, "Urban arterial traffic status detection using cellular data without cellphone GPS information," *Transportation Research Part C: Emerging Technologies*, vol. 114, 2020.

[6] Y. Liu, Z. Liu, and R. Jia, "DeepPF: a deep learning based architecture for metro passenger flow prediction," *Transportation Research Part C: Emerging Technologies*, vol. 101, pp. 18–34, 2019.

[7] F. Zheng and H. Van Zuylen, "Urban link travel time estimation based on sparse probe vehicle data," *Transportation Research Part C: Emerging Technologies*, vol. 31, pp. 145–157, 2013.

[8] M. G. Demissie, G. H. D. A. Correia, and C. Bento, "Intelligent road traffic status detection system through cellular networks handover information: an exploratory study," *Transportation Research Part C: Emerging Technologies*, vol. 32, pp. 76–88, 2013.

[9] F. Yang, Z. Yao, P. J. Jin, and D. Yang, "Performance evaluation of handoff-based cellular traffic monitoring systems using combined wireless and traffic simulation platform," *Journal of Intelligent Transportation Systems*, vol. 20, no. 2, pp. 113–124, 2016.

[10] E. Jenelius and H. N. Koutsopoulos, "Travel time estimation for urban road networks using low frequency probe vehicle data," *Transportation Research Part B: Methodological*, vol. 53, pp. 64–81, 2013.

[11] S. B. Kotsiantis, "Supervised machine learning: a review of classification techniques," *Informatica*, vol. 31, pp. 249–268, 2007.

[12] A. D. Patire, M. Wright, B. Prodhomme, and A. M. Bayen, "How much GPS data do we need?" *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 325–342, 2015.

[13] L. Li, J. Zhang, Y. Wang, and B. Ran, "Missing value imputation for traffic-related time series data based on a multi-view learning method," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 8, pp. 2933–2943, 2019.

[14] Y. B. Y. Yim and R. Cayford, *Investigation of Vehicles as Probes Using Global Positioning System and Cellular Phone Tracking: Field Operational Test*, California Partners for Advanced Transportation Technology, Berkeley, CA, USA, 2001.

[15] C.-C. Chang and C.-J. Lin, "Libsvm," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011.

[16] J. Zhang, Y. Cheng, S. He, and B. Ran, "Improving method of real-time offset tuning for arterial signal coordination using probe trajectory data," *Advances in Mechanical Engineering*, vol. 9, no. 1, pp. 1–7, 2017.

[17] M. Piórkowski, M. Raya, A. L. Lugo, P. Papadimitratos, M. Grossglauser, and J.-P. Hubaux, "TraNS," *ACM SIGMOBILE - Mobile Computing and Communications Review*, vol. 12, no. 1, pp. 31–33, 2008.

[18] L. Shen, L. Du, X. Yang, X. Du, J. Wang, and J. Hao, "Sustainable strategies for transportation development in emerging cities in China: a simulation approach," *Sustainability*, vol. 10, no. 3, p. 844, 2018.

[19] S. Baldi, I. Michailidis, V. Ntampasi, E. Kosmatopoulos, I. Papamichail, and M. Papageorgiou, “A simulation-based traffic signal control for congested urban traffic networks,” *Transportation Science*, vol. 53, no. 1, pp. 6–20, 2019.

[20] J. Zhang, X. Jiang, Z. Liu, L. Zheng, and B. Ran, “A study on autonomous intersection management: planning-based strategy improved by convolutional neural network,” *KSCE Journal of Civil Engineering*, vol. 25, no. 10, pp. 3995–4004, 2021.

[21] R. Bauza and J. Gozalvez, “Traffic congestion detection in large-scale scenarios using vehicle-to-vehicle communications,” *Journal of Network and Computer Applications*, vol. 36, no. 5, pp. 1295–1307, 2013.

[22] O. Cárdenas, A. Valencia, and C. Montt, “Congestion minimization through sustainable traffic management: a microsimulation approach,” *LogForum*, vol. 14, 2018.

[23] J. Zhang, S. He, W. Wang, and F. Zhan, “Accuracy analysis of freeway traffic speed estimation based on the integration of cellular probe system and loop detectors,” *Journal of Intelligent Transportation Systems*, vol. 19, no. 4, pp. 411–426, 2015.

[24] F. Yang, Z. Yao, P. J. Jin, and Y. Xiong, “Arterial link travel time estimation considering traffic signal delays using cellular handoff data,” *IET Intelligent Transport Systems*, vol. 13, no. 3, pp. 461–468, 2019.

[25] G. Brown, A. Pocock, M. J. Zhao, and M. Luján, “Conditional likelihood maximisation: a unifying framework for information theoretic feature selection,” *Journal of Machine Learning Research*, vol. 13, pp. 27–66, 2012.

[26] European conference of ministers of transport, *Managing urban traffic congestion*, OECD, Paris, France, 2007.

[27] J. M. Kopf, J. Nee, J. M. Ishimaru, M. E. Hallenbeck, and D. Bremmer, “Measurement of recurring and non-recurring congestion,” *Phase*, vol. 2, 2005.

[28] P. Liu, J. Wu, H. Zhou, and J. Bao, “Estimating queue length for contraflow left-turn lane design at signalized intersections,” *Journal of Transportation Engineering, Part A: Systems,* vol. 145, no. 6, 2019.