Queue Length Estimation Based Defence Against Data Poisoning Attack for Traffic Signal Control

Xu Gao, Jiqiang Liu, Yike Li, Xiaojin Wang, Yingxiao Xiang, Endong Tong, Wenjia Niu, Zhen Han

Queue Length Estimation based Defence Against Data Poisoning Attack for Traffic Signal Control

Xu Gao[1], Jiqiang Liu[1], Yike Li[1], Xiaojin Wang[1], YingXiao Xiang[1], Endong Tong[1,*], Wenjia Niu[1,*], and Zhen Han[1]

Beijing Key Laboratory of Security and Privacy in Intelligent Transportation, Beijing Jiaotong University, 3 Shangyuan Village, Haidian District, Beijing 100044, China.{niuwj,edtong}@bjtu.edu.cn

Abstract. With the development of intelligent transportation systems, especially in the context of the comprehensive development and popularization of big data and 5G networks, intelligent transportation signal systems have been experimented and promoted in various countries around the world. As with other big data-based systems, specific attacks pose a threat to the security of big data-based intelligent transportation system systems. Targeting system vulnerabilities, certain simple forms of attack will have a huge impact on signal planning, making Actual traffic is congested. In this article, we first show a specific attack and then add more attack points, analyze the system's vulnerabilities, and model based on traffic waves and Bayesian predictions, so that the attack points can help the impact is weakened and the traffic can function normally. For experiments, we performed traffic simulation on the VISSIM platform to prove the impact of our attack and further verify the accuracy and effectiveness of the model.

Keywords: I-SIG · Connected Vehicle · Bayesian Prediction · Traffic Wave.

1 Introduction

Vehicle network technology is gradually changing the current transportation network, and not only in China but also in various countries around the world have launched a pilot program for connected vehicles. At present, the Internet of Vehicles technology is gradually applied to online maps and some electric vehicles. Chinese transportation departments and Internet companies are also conducting research and experiments on intelligent traffic signal systems in recent years.

In September 2016, the U.S. Department of Transportation launched a pilot program for intelligent traffic signal systems. In this program, the vehicle's infrastructure is connected via wireless communications, using technologies to optimize traffic planning and prevent traffic failures and congestion. In 2018, the intelligent traffic signal system has been tested in three cities including New York. In order to promote deployment in the United States, USDOT [5] has proposed to authorize all new light vehicles to be equipped with connected vehicle (CV) technology [8]. No matter what technology is, as long as it is applied

in the real-life field, its safety is very important to us. In order to ensure the safety of vehicle and transportation equipment infrastructure and the safety of drivers and pedestrians in the environment, understand the security loopholes in the systems we deploy are very important, which also guarantees the stability of subsequent deployments.

Intelligent traffic signal systems (I-SIG) have been widely used in various countries. Intelligent traffic signal systems carry vehicle data and traffic signal data. The Internet of Vehicles technology is the core module of the system. Its technology has gradually matured and related products have been applied to practice. The US Department of Transportation estimates that by 2020, the cost of assembling on-board units will be about $ 350, which will make the cost and benefit of car-to-vehicle deployment more beneficial to society. In this paper, we analyze the security of transportation systems based on the Internet of Vehicles technology and study the design-level security issues and challenges in the case of multi-point attacks. Finally, we use the Bayesian probability model to find the attack point for defense.

2 Related work

2.1 Congestion Attack Based I-SIG

The first safety analysis of the emerging CV-based I-SIG was performed in [1]. Aiming at a highly realistic threat model, that is, dispersing data from an attack tool, the author conducted a vulnerability analysis and found that the current signal control algorithm design and configuration choices are extremely vulnerable to congestion attacks. The evaluation results in the real environment verify the effectiveness of the attack and show that the attack can even produce a blocking effect that prevents the entire method. Then use these insights to discuss defense direction. This work is the first step in understanding new safety issues and challenges in the next generation of CV-based transportation systems.analysis of the internal structure and algorithms of the signal system, and the defense measures and methods we need.

2.2 Prediction Algorithm of Traffic Wave

The queue prediction is an important content in the field of transportation. The original intention of the prediction is to be able to accurately estimate the real situation of traffic when the traffic situation is uncertain. To this problem, [17] and [18] proposed two different estimation methods. Both draw on the content of physics, and from their work, we need to clean and segment the acquired data to estimate the queues for actual traffic conditions. We must connect the related parts and conduct the experiment.

Traffic wave theory [13] is contained in the field of transportation, and its original intention is to combine the changes in traffic flow with the relevant content of fluid mechanics in physics. Traffic waves are divided into evacuation

waves and aggregate waves. Evacuation waves refer to the movement of the interface where traffic waves change from a high-density state to a low-density state. Decreasing the density of traffic flow will generate evacuation waves, and the opposite is the aggregate wave. This model based on physics can be easily combined with the actual traffic situation [11], and simulated and calculated by mathematical models. In this article, we have borrowed relevant knowledge in the field of transportation.

2.3 Bayesian Hieratical Model

The [19] study seeks to investigate the variations associated with lane lateral locations and days of the week in the stochastic and dynamic transition of traffic regimes(DTTR). In the proposed analysis, hierarchical regression models fitted using Bayesian frameworks were used to calibrate the transition probabilities that describe the DTTR. Datasets of two sites on a freeway facility located in Jacksonville, Florida, were selected for the analysis. The traffic speed thresholds to define traffic regimes were estimated using the Gaussian mixture model(GMM). These findings can be used in developing effective congestion countermeasures, particularly in the application of intelligent transportation systems, such as dynamic lane-management strategies.

In the field of transportation, there is also a related work to predict the length of lanes by using Bayesian and Internet of Vehicles technology [2]. Its work is based on Bayesian probability models, supplemented by vehicle data. By modeling and analyzing the data, and determining confidence degree. This Bayesian model is based on the Bayesian principle and uses the knowledge of probability statistics to classify a sample data set. Because of its solid mathematical foundation, the false positive rate of the Bayesian classification algorithm is very low [16]. The method is characterized by combining the prior probability and the posterior probability, which avoids the subjective bias of using only the prior probability and also avoids the over-fitting phenomenon using the sample information alone [12]. The Bayesian classification algorithm shows higher accuracy in the case of large data sets.

3 Defense Model

3.1 Single Point Attack

As shown in Fig. 1, the intelligent traffic signal system involves various units of actual traffic, including a road condition monitoring unit at an intersection and a vehicle-mounted unit that sends vehicle data. It was found in previous work that the security of traditional traffic infrastructure is weak, and an attacker can easily control it completely.

Therefore, in the work of this article, we focus on the safety of the on-vehicle unit, specifically by attacking the on-vehicle unit to send the wrong vehicle data to the system to affect the planning of the transportation system, and there is

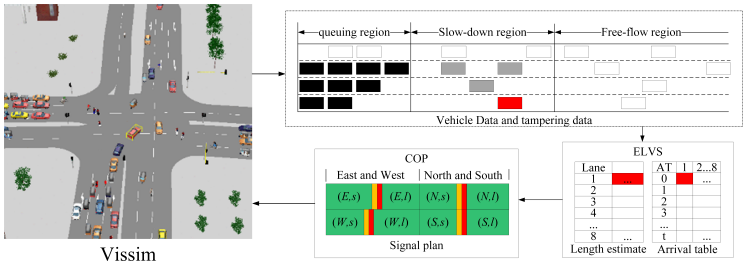

Fig. 1. The flow chart of single point attack experiment.

already a working proof There are simple attack methods that can affect normal traffic.

In this article, we first realize the analysis of traffic congestion by sending different vehicle data while controlling the vehicle-mounted unit. Further, we add single attack points to analyze the impact of the attack points on the traffic and the interaction between the attack points. With effect. The model we build is presented by the simulation software VISSIM [4] in the transportation field. It provides the Component Object Model (COM) required to build the model. This interface can input signal control into the VISSIM simulation environment, and can also obtain the corresponding intersection from vehicle data, we use code to integrate control input and vehicle output to complete the overall experimental process operation in this way. First of all, the attack process is at the stage of obtaining vehicle data. We tamper with the data obtained by the attacked vehicle to affect the subsequent traffic signal planning. The traffic planning process draws on the latest ISIG system provided by the US Department of Transportation.

3.2 Estimate Queue Length

Traffic congestion has a great impact on us. It not only increases the transit time of vehicles but also increases the probability of traffic accidents. In order to resist the impact of the attack on the traffic, we use the traffic wave model researched in the field of transportation to estimate the queue length of the current lane.

We analyzed the cause of the congestion, and we found that vehicles that were motionless in the system would have a huge impact on the I-SIG algorithm for predicting the lane queue length. In consideration of this, we observed the traffic flow at the intersection of the cycle before the attack. To predict the lane queuing length at intersections in the current cycle, we predict the current queuing length based on the traffic wave model based on the original data. The calculation formula is as follows:

$$L_p = |w_q|t_w + |w_s|t_w + |w_f|t_w \tag{1}$$

Among them, w_q and w_s are the stopping waves of the waiting section and the decelerating section, w_f is the starting wave of the free passage section, and

w is the duration of a signal cycle, which is calculated by the formula:

$$w_i = \frac{k_i v_i \ln(k_i/k_j)}{k_i - k_j} \tag{2}$$

k_i and k_j are the density of the divided i section and the i previous section, the previous section of the slow-down section is the waiting section, the density of the previous section of the waiting section is equal to itself, and v_i is the average speed of the vehicles in i section.

3.3 Bayesian Defense Model

We propose a method for estimating the maximum queue length of a vehicle at a signalized intersection using high-frequency trajectory data of the vehicle. The estimated queue length is estimated from the distribution of multiple adjacent periods by the maximum posterior method. The data of these adjacent cycles are obtained through the simulation environment. We estimate the queue length at the next moment from the traffic data at the previous moment and compare it with the real value. The method of predicting the length uses the traffic wave model proposed in the previous section.

We can use the Bayesian model to predict that the current queue length meets the conditions with a certain confidence α by L_p. The specific steps are as follows:

• **Step1**: Use the traffic wave model to predict the queue length L_w^t of the current stage t based on the data of stage $t-1$.

• **Step2**: Calculate the absolute valued of the predicted value L_w^t and the true value L_r^t in the current stage.

• **Step3**: Calculate d_m.

$$d_m = \max \left\{ d | \min_D \frac{\sum_i^{\{D\}} d_i}{\sum_j^N d_j} < \alpha \right\} \tag{3}$$

• **Step4**: In the case where the confidence is, the probability P that the predicted queue length L_p is smaller than the confidence interval using the Bayesian model is

$$P(L_p < d_m | lane = l) = \sum_{d<d_m} \frac{P(L_p = d) \prod_k P(lane = k | L_p = d_m)}{\prod_k P(lane = k)} \tag{4}$$

• **Step5**: If P=0, there is an attacked vehicle within the predicted length, otherwise it does not.

The following is our algorithm flow for finding attack vehicles. D represents a data set constructed by collecting the difference between the predicted length of the previous cycle and the actual sample queue length in a sample of multiple traffic flows. We use the model trained by Bayesian method to compare the

predicted length at the test moment. If the predicted length is within the confidence interval, we believe that there is no attack point at that moment, and it is possible that the attack did not cause system planning impact. If it is beyond the range of the confidence interval, we will find the ID of the attacked vehicle and exclude it from the calculation. This effectively prevents traffic congestion at the attack point. α is the confidence of our principle.

Algorithm 1 Defense algorithm

Require: Veh_t, Veh_{t-1}, α
Ensure: ID
initialize: Set ID, d, $L_p = 0$, D$\leftarrow$ Veh_{t-1}
sort: D
for $i = 1, 2, ..., len(D)$ **do**
  if $\sum_{j=1}^{i} \text{D}_i/\text{sum(D)} > \alpha$ **then**
    $\text{d} = D_{i-1}$
    break:
  end if
end for
for j = 1,2,...,len(Veh_t) **do**
  if $L_j > d$ **then**
    ID = j
  end if
end for
return ID

4 Experiment

4.1 Experimental Setup

Traffic Intersection Setup. By collecting data on real traffic environment, we mainly analyze the structure of the intersection and the number of lanes at each intersection. In our simulation environment, the maximum speed is limited to 40km/h. This speed refers to the maximum speed of traffic restrictions in Chinese cities and the intersection range monitored by the system is set to be about 500m. This range is determined based on statistical traffic flow changes. The speed that VISSIM generates vehicles is eight vehicles per ten seconds. The generated speed refers to daily life near the school counted by our students. The average traffic volume. The simulation program is integrated into Visual Studio. Each attack simulation lasts about 30 minutes.

The traffic light has several phases. The time and process after a straight and left turn in both directions at an intersection are called phases. The phase is determined after calculation based on the traffic volume at each intersection. The phase of each intersection is different. Start by measuring the flow, measuring various traffic flows [11] in all directions in a period, including pedestrian, non-motorized and motor vehicle traffic, and peak traffic. Then calculate the time

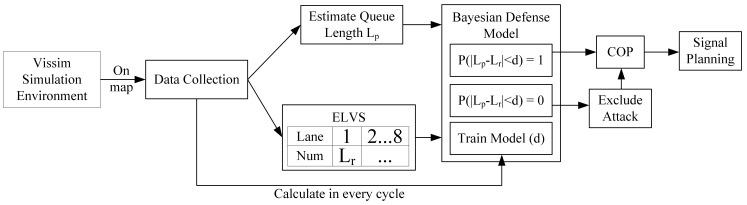

Fig. 2. The data flow of vehicle driving data in the experiment. Vehicle data comes from I-SIG based simulation environment.

required for them to run according to the law of traffic flow, and then determine the time of the red and green lights at each intersection, and then determine the phase. The signal lights are divided according to the diagram [9]. The odd numbers represent the left turn direction and the even numbers represent the straight direction.

Attack Data Generation. For the generation of attack data, we refer to the known work using the longest distance attack method, adding multiple attack points for the loophole length algorithm vulnerability in the system, the number of attack points is less than or equal to four, and the data structure of the attack points satisfies the basic requirements of the system. In the attack phase, we assume that the corresponding number of vehicles are controlled by us. We directly modify the corresponding number of vehicle data in the integration program and complete the modification before the signal planning. Each attack will attack the eight phases of the intersection 1000 times each time. The seeds are generated by random vehicles to ensure randomness. In the initial stage of the experiment, we used a large interval assembly rate setting for a set of experiments.

Data Collection. To prepare the data set for subsequent experiments, we integrate the system's estimation algorithm and planning algorithm with the VISSIM interface in VS. Each simulation will record the vehicle data at the current moment and the vehicle data for the next cycle. The evaluation file generated by VISSIM after the simulation is completed.

4.2 Simulation

The experimental work in this paper is demonstrated through the VISSIM simulation platform. The simulation experiment simulates the actual traffic situation at the intersection. Under the current development of the Internet of Vehicles, it cannot satisfy all vehicles with vehicle-mounted units. This assumption is in line with reality. Traffic conditions, so our experiment set the parameter of the penetration rate, the experiment with different penetration rates. Fig. 2 shows the reality used in real experiments. One of the currently known attacks is an attack on the maximum queue length. After analysis, it is determined that it is an algorithm that estimates the queue length of vehicles at the intersection. This attack in a specific traffic situation passing a single point can have a greater

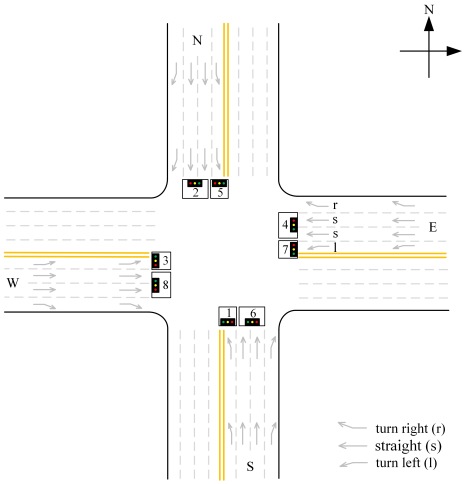

Fig. 3. This is the experimental scenario for the I-SIG system.Two phases in the four directions of the intersection are digitally marked, and each cycle calculates the time allocated for different phases in each direction.

impact on the traffic situation of the entire intersection. This attack has the characteristics of being far from the intersection and still. We conduct follow-up research experiments through this key point. During the experiment, we first set an attack point in each direction and applied the characteristics obtained from the analysis to our attack experiments. We controlled each attack point. The speed is 0, and control its position in the intersection.

Adding attack points and feeding back the planned results to the VISSIM simulation environment is achieved through the VISSIM interface. Assuming we control the attacked vehicle through the vulnerability of the vehicle unit, we mark the ID of the attacked vehicle in the experiment. The experimental code changes its original vehicle data through using COM, delivers false data to the intelligent traffic signal system, and the system gets the plan after the attack and enters it into the VISSIM simulation environment.

5 Analysis

In Fig. 3, we describe the calculation and prediction of the average queue length of the lane. The diagonal line in the left half represents the traffic flow in different stages. For example, the green diagonal line in the lower left L_1 represents the traffic flow in the free passage phase of L_1. The angle between it and the vehicle's running direction is the average vehicle speed at this stage, and the vertical lines represent the corresponding distances at different stages. During the experiment, we set the distance between the intersection and the intersection to 1 km, which is also a reference. Real environment at the intersection near the school. In the

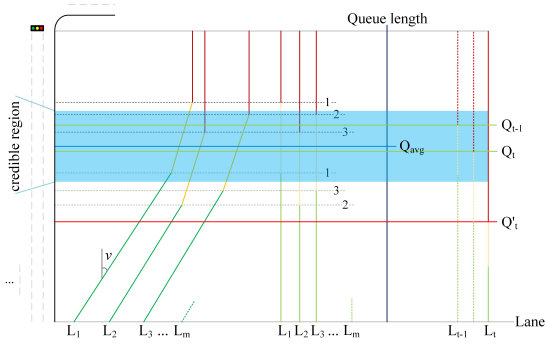

Fig. 4. Prediction of queue length based on traffic waves and Bayesian Model.The right vertical axis is the projection of the queue length, and the horizontal axis marks the different lanes.

experiment, we collected the difference between the predicted value and the real value before and after the two cycles to build the Bayesian model dataset. We used the intel i5-9400f CPU to run the program in order to collect the simulation data. The simulation results can be collected about every 30s We have compiled a total of 100,000 data sets to build a probability model. In Fig. 4, i in Li represents the number of lanes, the point where the horizontal axis line and the blue solid line intersect represents the queue length of the lane, and Q represents the queue length. The blue area indicates that we have obtained 95% confidence in the dataset. Q_{t-1} represents the queue length of the vehicle in the previous cycle. Q_t represents the vehicle data of the previous cycle without attacking the vehicle. The queue length, $Q_t^{'}$ represents the predicted queue length in the case of attacking vehicles. If the predicted queue length is outside the confidence zone, it means that there is an abnormality in the current traffic situation, that is, some vehicles maliciously send false data.

From Fig. 5 we can see that in the case of a high penetration rate, the impact on actual traffic is small. In the case of low penetration rate, because all vehicle information in the actual environment cannot be obtained, multiple attacked vehicles that send the wrong data has a greater impact on the signal planning caused by the system. Specific reasons Through careful analysis of the planning part of the code and its principle, the conclusion is that the data of each vehicle is normal for the system. The data of the attacked vehicle that is far away and stationary will cause the EVLS module to incorrectly wait for the vehicle queue.

Our attack uses the longest distance strategy, which is to modify the position data of multiple attacked vehicles to the farthest distance that the system can monitor and the vehicle speed is zero. It can be seen that our model can indeed be used at different preparation rates. Resist multi-point attacks against I-SIG. The reason is that when we find that the queue length is abnormal through the Bayesian model, the vehicle information outside the predicted area length is excluded and recorded to prevent repeated attacks in the future. It can be

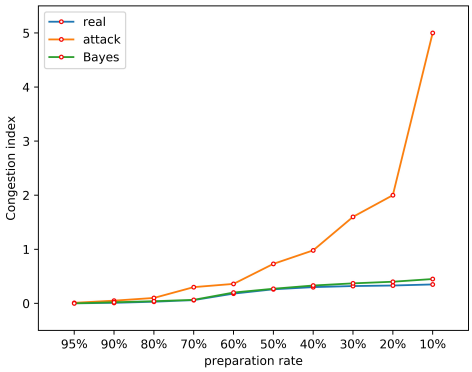

Fig. 5. Congestion index in different situations.

seen that when the preparation rate is smaller, the impact of our attack on the system is greater because the system cannot accurately obtain the queue length of the lane with less data because the calculation of the queue length needs to be more critical.

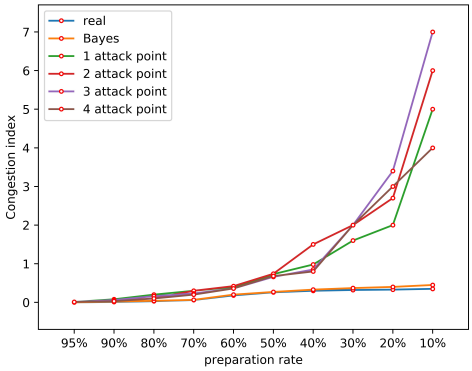

Fig. 6. The effect of adding different numbers of attack points.

Fig. 6 shows the impact of adding multiple attack points on the system. Different numbers have different effects on the system. This also proves that the traffic flow between the directions of the intersection is related. Multi-point attacks make the predicted length of lanes in different directions. At the same time, their changes also change the length of the traffic lights. We found that

increasing the number of attack points will affect normal traffic, but increasing the number of attack points does not necessarily enhance the effect of the attack.

6 Conclusion

The Internet of Vehicles technology is the most critical part of I-SIG. Our entire experiment is based on I-SIG. Through the simulation of the traffic simulation platform and the collection of experimental data, we gradually explore the internal implementation principle of the system and find its implementation. Problems and loopholes in the implementation process. These loopholes are not caused by inadequate consideration by designers, but because the Internet of Vehicles technology has not been widely popularized, causing the output under certain parameter settings to cause traffic congestion. In this paper, we first analyze the characteristics of single-point attacks systematically, and increase the number of attack points for experimental analysis. Finally, we propose a Bayesian model based on traffic waves to implement attacks against tampered vehicle data. The experimental results also show our model can defend well within a certain confidence range.

Although some of the multi-point attacks have no impact on the system and cannot be ruled out, this does not affect the effectiveness of the defense model. From the actual simulation results, the effect is still significant. During the experiment, we also tried to compare other models such as neural networks, decision trees, etc. From the current experimental results, these models do not have a good defense and generalization effect against multi-point attacks. Similarly, we also found that continuous addition attacks Vehicles, gradually affecting the system is a special attack method. The Bayesian model performs well, but it needs to be improved in future work.

7 Acknowledge

This research is supported by the Fundamental Research Funds for the Central Universities of China(No.2019RC008).

References

1. Qi Alfred Chen, Yucheng Yin, Yiheng Feng, Z. Morley Mao, and Henry X. Liu: Exposing Congestion Attack on Emerging Connected Vehicle based Traffic Signal Control, NDSS'18, San Diego CA.
2. Yu Mei, Weihua Gu, Edward C.S. Chung , Fuliang Li , Keshuang Tang: A Bayesian approach for estimating vehicle queue lengths at signalized intersections using probe vehicle data, Transportation Research Part C(2019)
3. MMITSS-AZ 1.0, https://www.itsforge.net/index. php/community/explore-applications/for-search-results#/30/63
4. PTV Vissim, http://vision-traffic.ptvgroup.com/en-us/products/ptv-vissim.

5. USDOT: Security Credential Management System (SCMS), https:// www.its.dot.gov/factsheets/pdf/CV SCMS.pdf.
6. H. A. E. AiJameel and M. A. H. AiJumaili, Analysis of traffic stream characteristics using loop detector data, Jordan J. Civil Eng., vol. 10, no. 4, pp. 403C416,2016.
7. Z. W. Li, J. Zhang, and H. Gu, Real-time traffic speed estimation with adaptive cruise control vehicles and manual vehicles in a mixed environment, in Proc. 16th COTA Int. Conf. Transp. Professionals, 2016, pp. 51C62.
8. M. Wang, W. Daamen, S. P. Hoogendoorn, and B. van Arem, Connected variable speed limits control and car following control with vehicle infra structure communication to resolve stop-and-go waves, J. Intell. Transp. Syst., vol. 20, no. 6, pp. 559-572, 2016.
9. Y. Feng, K. L. Head, S. Khoshmagham, and M. Zamanipour, A real-time adaptive signal control in a connected vehicle environment, Transportation Research Part C: Emerging Technologies, vol. 55, pp. 460-473, 2015.
10. G. Comert, Effect of stop line detection in queue length estimation at traffic signals from probe vehicles data, European Journal of Operational Research, vol. 226, no. 1, pp. 67C76, 2013.
11. Z. Sun and X. J. Ban, Vehicle trajectory reconstruction for signalized intersections using mobile traffic sensors, Transportation Research Part C: Emerging Technologies, vol. 36, pp. 268C283, 2013.
12. An, C., Wu, Y.J., Xia, J., Huang, W., 2018. Real-time queue length estimation using event-based advance detector data. Journal of Intelligent Transportation Systems 22 (4), 277-290.
13. Cai, Q., Wang, Z., Zheng, L., Wu, B., Wang, Y., 2014. Shock wave approach for estimating queue length at signalized intersections by fusing data from point and mobile sensors. Transportation Research Record: Journal of the Transportation Research Board 2422, 79-87.
14. Rompis, S. Y., Cetin, M., Habtemichael, F., 2018. Probe vehicle lane identification for queue length estimation at intersections. Journal of Intelligent Transportation Systems 22 (1), 10-25.
15. Newson, P., Krumm, J., Hidden markov map matching through noise and sparseness. In: Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems 2009. ACM, pp. 336-343.
16. Lai Zheng, Tarek Sayed:Bayesian hierarchical modeling traffic conflict extremes for crash estimation: A non-stationary peak over threshold approach, Analytic Methods in Accident Research 2019
17. Jereme Chien, Wen Shen: Stationary wave profiles for nonlocal particle models of traffic flow on rough roads, Nonlinear Differential Equations and Applications NoDEA, 2019, Vol.26 (6), pp.1-25
18. Changxi Ma, Ruichun He: Green wave traffic control system optimization based on adaptive genetic-artificial fish swarm algorithm, Neural Computing and Applications, 2019, Vol.31 (7), pp.2073-2083
19. Emmanuel Kidando, Ren Moses, Thobias Sando, Eren Erman Ozguven: An application of Bayesian multilevel model to evaluate variations in stochastic and dynamic transition of traffic conditions, Journal of Modern Transportation 2019