Journal of Highway and Transportation Research and Development
2024,18(2):82-88
ISSN 2095-6215
DOI: 10. 26599/HTRD. 2024. 9480017

Vehicle Trajectory Generation Based on Generation Adversarial Network

Zhonghe He*, Renchi Shao, Sijia Xiang**

(School of Electrical and Control Engineering, North China University of Technology, Beijing 100144, China)

Abstract: With the development of networked vehicles, location information-based transportation systems have proven to provide significant benefits. However, the exposure of vehicle location information also raises important privacy issues. Current typical methods for protecting vehicle location privacy protection methods such as anonymity and pseudonymity, still carry the risk of the vehicle being tracked, leading to data security issues. This paper proposes a kind of vehicle trajectory generation algorithm based on Generative Adversarial Networks (GAN). The algorithm utilizes vehicle movement trajectory data to train both the discriminator and generator models to generate virtual trajectory data that matches the distribution of real trajectory data. Therefore, virtual trajectory data can obscure vehicle information, addressing the privacy concerns associated with moving trajectory data and enhancing the security of applications. In this paper, the vehicle travel time of sample trajectory data and virtual trajectory data is used as indicators for statistical analysis. The experiment demonstrated that the cumulative probability distribution of travel time for the sample data and virtual data passed the Kolmogorov-Smirnov (K-S) test at permeabilities ranging from 10% to 100% and at significance levels of 0. 01 and 0. 05. Both datasets accepted the hypothesis that they originate from the same distribution. The reliability of the proposed method for generating virtual trajectories has been demonstrated.

Key words: traffic engineering; vehicle location privacy protection; traffic data Security; generative adversarial network; virtual trajectory

1 Introduction

With the development of the automotive industry, the importance of automotive transportation is becoming increasingly prominent. Particularly, the widespread application of various mobile traffic sensors, such as automatic toll collection (ETC) tags, global positioning system (GPS) devices, Bluetooth Mac address matching, has gaimed attention.[1] The large amount of trajectory data collected by vehicles can be used for the construction of urban intelligent transportation to better prevent the waste of traffic resources and traffic congestion problem[2-3]. But the exposure of vehicle location information can also raise important privacy concerns. If this location information falls into the hands of a "person of interest", vehicle users will face serious threats[4-5].

In the original in-vehicle self-assembly network, privacy was divided into three categories: the identity of the vehicle, the location of the vehicle, and the data generated while the vehicle was in motion. Since data can be protected by cryptographic processing, the primary issue is to protect the vehicle identity and location. To cope with such problems, anonymity and pseudonym change strategies have emerged, and typical anonymous authentication mechanisms include pseudonyms, random assignment, group signatures, ring signatures, blind signatures, smart cards, etc.[6] Simple anonymization does not protect privacy well, and an effective K-anonymization mechanism is proposed in the literature[7], but optimizing this mechanism requires high computational efficiency, which cannot be maximized by the limited resources of mobile vehicles. The use of pre-generated pseudonyms when vehicles broadcast their privacy information and the use of pseudonym change strategies to protect vehicles from being tracked have been proposed in the literature[8-9]. There are three types of pseudonym strategies: time-based hybrid strategies[10], where vehicles change their pseudonyms periodically, but this type of methods can be easily broken, even in some cases, tracking accuracy reaches 100%[11]; With area-based mixing strategies, vehicles enter predefined road sections and change their names with the coordination of roadside units. These road sections are also called hybrid zones[12]. However, the experimental results of Buttyan

Received 30 May, 2023; Revised 6 August 2023; Accepted 1 September 2023

* *E-mail address*: zhonghehe@ ncut. edu. cn

** Corresponding author, xiangsijia0823@ 163. com

© The Author(s) 2024. Published by Tsinghua Uhiversity Press. This is an open access article under the CC BY license (http://creativecommons. org/licenses/by/4. 00/).

et al.[13] show that the success rate of tracking vehicles is up to 70%. The context-based hybrid pseudonym strategy, which improves the drawbacks of the above two strategies by adding speed and roadway information to the context[14-15], although it effectively enhances the security of vehicle information. However, changing names for a long time will increase the probability of data leakage.

Another privacy-preserving approach is to implement location blurring, which protects the actual vehicle location by mapping it to an imprecise or inaccurate location. SAXENA et al.[16] proposed a probability-based metric to quantify the privacy of fuzzy processing mechanisms and illustrated how to express some existing positional fuzzy processing using a formal model. The literature[17] also proposes a user-centric location blurring method and experiments with real data to verify its effectiveness in hiding user location information. Implementing location blurring is an effective way to protect privacy, but the possibility of data theft increases as the number of users around increases.

In recent years, researchers have focused more on the balance between application modeling and privacy protection. LIU, et al.[18] proposed a deep learning model based on traffic flow prediction using a large number of real data sets and introduced a privacy-preserving machine learning approach to improve the prediction accuracy of traffic flow data while achieving good privacy protection. However, the large amount of real data leads to a large limitation of the experiment. The generation of virtual trajectories[19] further improves the privacy preservation. The literature[20] proposes a virtual trajectory generation algorithm based on the rotation algorithm. The corresponding virtual points are generated at each location according to the user's needs, and virtual trajectories are formed by connecting the virtual location points. Although the more virtual trajectories, the lower the probability of trajectory leakage, there is still a probability of trajectory leakage.

Based on this, this paper applies the generative adversarial networks (GANs) to vehicle virtual trajectory data generation, colling real trajectory data as samples, training the network structure parameters in GAN, and generating large amounts of high-quality virtual trajectory data that conform to the real data distribution for various traffic modeling application analysis. Since the data released to the application is completely virtual, erasing any real vehicle information, it can solve the problem of location privacy protection of vehicle track data and improve the security of traffic data application.

2 Gan-based trajectory generation

2.1 Introduction to virtual track generation

Vehicle virtual trajectory generation firstly collects the vehicle trajectory data of road network area; after that, the processing data corresponds to each road section to generate dataset to provide available training samples for the next step of network training; through generative adversarial network training dataset, finally outputs a large amount of high quality virtual trajectory data that meet the real trajectory data distribution. Figure 1 shows the flow chart of the structure of algorithm.

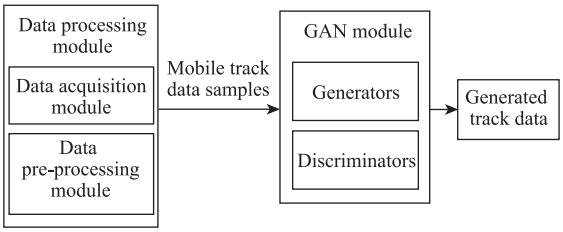

Fig. 1 Structured flowchart

2.2 Data processing algorithms

Firstly, the mobile track data of the vehicle is collected, the original data of the mobile track contains SIM card number, longitude, latitude, time, speed, speed direction and other information, the longitude and latitude data of each road section in the trunk line can be obtained through map matching and other methods, and the collected original data is matched to the corresponding road section; The mobile trajectory data of each road section of the trunk line obtained from the data acquisition module is processed, the longitude and latitude offset data, time abnormal data, and incomplete data are identified as faulty data, and the available data are cleaned and screened to complete the data pre-processing process, and the available and complete mobile trajectory data are obtained and input to the generative adversarial network module. Figure 2 shows the data pre-processing process.

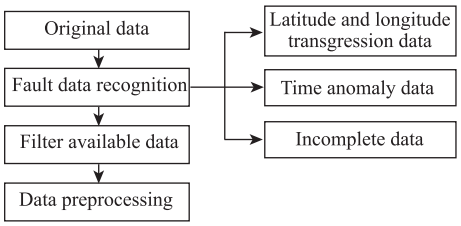

Fig. 2 Data processing flow

2.3 Network input

Assuming the real sample data of the real vehicle trajectory dataset for a certain road segment or a path or a region over a period of time, the label set $k_i \in \{1,\cdots,K\}$, this paper uses an unsupervised (i.e., all are unlabeled data) learning setup, i.e., the real set is labeled as $k=1$ and the set of pseudo data is $k=0$. For the sample dataset of vehicle trajectory $x^{(i)} \sim p_{\text{data}}[x^{(i)}]$, the distribution $p_{\text{data}}[x^{(i)}]$ of the dataset is estimated, where the trajectory dataset of each vehicle is denoted as $x^{(i)}=(x_1^1, y_1^1),(x_1^2,y_1^2),\cdots,(x_1^n,y_1^n)$ and (x_1^1,y_1^1) is the latitude and longitude coordinates of the first vehicle collected for the first time, and the trajectory dataset is reshape as $(x_1^1,x_1^2,\cdots,x_1^n),(y_1^1,y_1^2,\cdots,y_1^n)$ according to the input rules of neurons, after which the normalization process is performed and input to the nodes of the neural network.

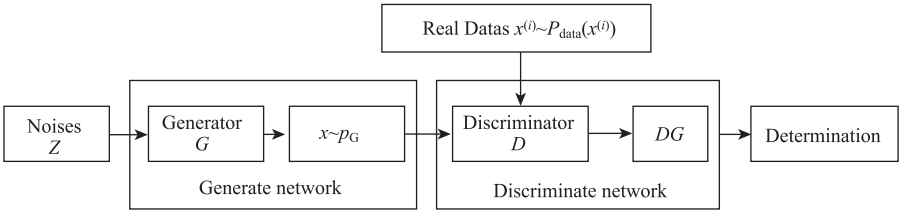

Fig. 3 Generative adversarial network architecture

2.4 Training process

We first read in for the collection of matching vehicle movement trajectory data set, and use the normalized trajectory data with Gaussian noise to generate the label set as the input of the generator, which generates the trajectory data conforming to the real distribution and input to the discriminator for determination, and the discriminator makes the determination according to the error result and feeds back to the generator, which updates the weights with the discriminator network. Specifically, after the batch read-in vehicle movement trajectory, a vehicle trajectory is taken as an example for illustration, after inputting a vehicle trajectory, the trajectory points of the movement trajectory data are first normalized so that they are spaced at $[0,1]$, and then used as the input of the generator together with Gaussian noise, and the vehicle trajectory points are generated through the operation of the generator network, and the generator hands the vehicle trajectory points to the discriminator and waits for the decision. The discriminator network makes a determination based on the error between the normalized sample trajectory data and the generated virtual trajectory data. If the determination is true, the label of this generated trajectory is recorded as 1; otherwise, it is 0. The target loss function for training is Equation (1):

$$V(G,D)=E_{x\sim P_{\text{data}}}\cdot \ln D(x) + E_{x\sim P_G}\cdot \ln[1-D(x)], \tag{1}$$

where $D(x)$ is the output value of the discriminator to determine whether x is the real trajectory.

Gradient descent is one of the most frequently used methods in solving the model parameters of machine learning algorithms, i.e., unconstrained optimization problems. When solving for the minimum of the loss function, the gradient descent method can be used to solve it step by step iteratively to obtain the minimized loss function and model parameter values. Conversely, if we need to solve for the maximum value of the loss function, then we need to iterate with the gradient ascent method. Update its network parameters by incrementing the stochastic gradient of the discriminator:

$$\square_{\theta_d}\frac{1}{m}\sum_{i=1}^{m}\{\ln D(x^{(i)}) + \ln[1-D(G(z^{(i)}))]\}. \tag{2}$$

Update network parameters by decreasing the random gradient of the generator:

$$\square_{\theta_g}\frac{1}{m}\sum_{i=1}^{m}\ln[1-D(G(z^{(i)}))] \tag{3}$$

According to the principle of the text generative ad-

versarial network, the two games, the reference[22] proved the convergence of the loss function, fixed G, the global optimum of the discriminator when $D_G^*(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x)+p_g(x)}$, at this time $D_G(x) = \frac{1}{2}$, the minimum value of the generator loss function is $C(G) = -\ln(4)$. At this time, the distribution of the generated trajectory data is the same as the distribution of the sample trajectory, and the discriminator discriminates between true and false trajectories with equal probability.

2.5 Algorithmic process

The specific algorithm steps are as follows:

Step 1: Sampling a small batch of m noise samples $\{z^{(1)},\cdots,z^{(m)}\}$ from the noise prior distribution $p_g(z)$.

Step 2: Sampling a small batch of m vehicle trajectory dataset $\{x^{(1)},\cdots,x^{(m)}\}$ from data generation distribution $p_{\text{data}}(x)$.

Step3: Update the discriminator by the iterative process in Equation (2) of stochastic gradient ascent.

Step 4: Loop steps 1-3 and sample a small batch of m noise samples $\{z^{(1)},\cdots,z^{(m)}\}$ from the noise prior distribution $p_g(z)$.

Step5: Update the generator by the iterative process in Equation (3) of stochastic gradient descent.

3 Simulation

To verify the reliability of the trajectory generation method based on generative adversarial networks in this paper, we use travel time as an indicator in our experiments to verify the effectiveness of using the generated dataset in traffic modeling analysis. For this purpose, the model and algorithm proposed in this paper are evaluated in the microscopic traffic simulation model VISSIM, where the traffic states of two adjacent road sections of the South Ring Road in Changping District, Beijing, during the peak period (from 18:00 to 19:00) are selected to build simulation models in the VISSIM simulation platform, using real traffic flow data and signal timing data. In the experiments of this paper, the latitude and longitude data of the numbered movement trajectories of all vehicles are collected as samples of the training data set, and the accuracy and influencing factors of the trajectory generation model of this paper are analyzed under different penetration rates. Figure 4 shows the simulation model of two adjacent sections of South Ring Road in Changping District built by the simulation software.

Since the mobile trajectory data we obtain are applied in traffic modeling such as travel time estimation, queue length estimation and other traffic state discriminations, vehicle latitude data can be ignored for set road sections, which not only eliminates the influence factor of latitude data in the training process, but also reduces the data size and improves the operation efficiency.

In the experiments of this paper, the sampling interval is 5 s. We use a constant learning rate, set to 0.001, and we take the time interval of one hour for each analysis, i.e., we use trajectory data within 3 600 s for training.

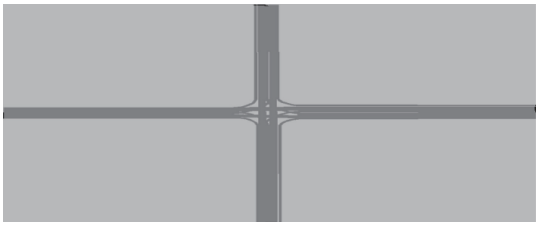

Fig. 4 Simulation model of adjacent sections of south ring road

In the simulation experiments, we compared the travel times of the generated data with the actual travel times. Testing hypothesis H_0: Kolmogorov-Smirnov test is a standard nonparametric test to indicate whether the sample comes from a hypothetical distribution (unlike other tests such as the t-test, which tests for unique means, or the chi-square test, which tests for data normality). The test was based on the K-S statistic, which was calculated as the maximum difference between the experimental and hypothetical cumulative distributions. The test provides a p-value indicating the strength of the fit. A low p-value indicates that the data do not accept the distribution of the hypothesis, and we reject the hypothesis H_0 with a p-value below the significance level a. The significance level a corresponds to the percentage of Type I errors allowed by the test (rejecting the original hypothesis when it is true). The parameter a is generally set to 0.05 or 0.01.

Figure 5 shows all the collected trajectory datasets of all vehicles, showing the variation of distance from upstream detection line to downstream detection line with

travel time for all vehicles, and Figure 6 shows the fitted graph of virtual trajectory dataset generated by all vehicles based on the trajectory generation algorithm in this paper. Figure 7 shows the cumulative probability distribution of travel time for both, and the hypothesis of $H_0 = 0$ was accepted by K-S test at significance levels $a = 0.05$ and 0.01 when the complete vehicle data was a sample.

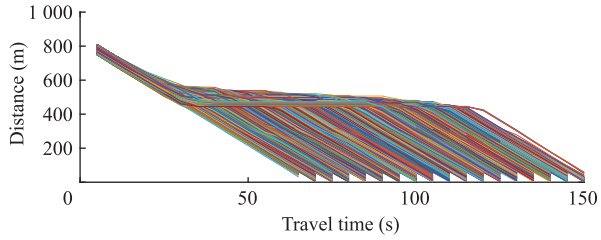

Fig. 5 All collected vehicle trajectory datasets

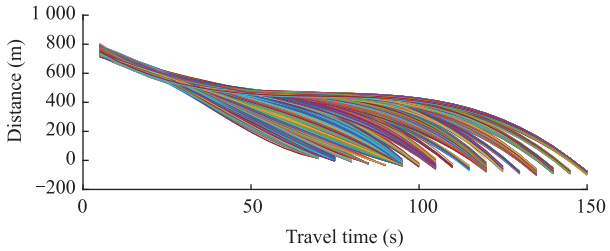

Fig. 6 Virtual trajectory dataset

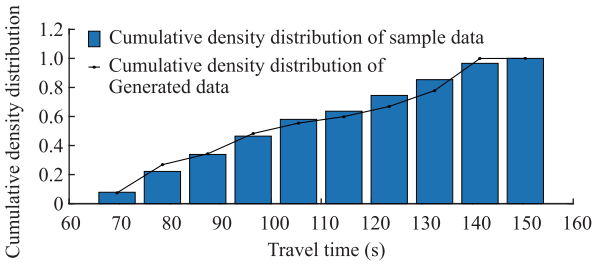

Fig. 7 Travel time cumulative probability distribution

To further analyze the effect of permeability on the accuracy of generated trajectories, a comparative analysis of trajectory travel times at different permeabilities was performed, with permeability gradually increasing from 10% to 100%, and for each permeability, a random sampling procedure was executed to extract the corresponding portion of trajectories from the complete set of trajectories for training. Taking the penetration rate of 50% as an example, the hypothesis that $H_0 = 0$ is accepted when the significance level is 0.01 and 0.05, i.e., the sample and generated data are from the same distribution, and the maximum distance k value between the two distributions is 0.1 and the p-value is 0.99. Figure 8 shows the vehicle trajectory dataset collected with 50% penetration, Figure 9 shows the virtual trajectory dataset generated with 50% penetration, and Figure 10 shows the cumulative probability distribution of travel time. Table 1 shows the results of the travel time K-S test at different penetration rates (form 10% to 100%), and the results show that the travel time of the virtual trajectory and the cumulative density function of the travel time of the trajectory samples accept the hypothesis of accepting $H_0 = 0$ at all penetration rates, the maximum distance k values of both distributions are less than 0.3, the probability from the same distribution, i.e., the P value, is greater than 0.6, and even at the lower penetration conditions, the method still has good results.

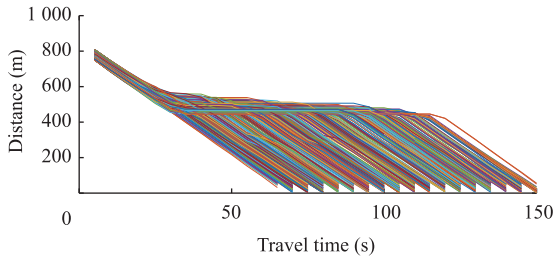

Fig. 8 Vehicle trajectory data set collected with penetration rate of 50%

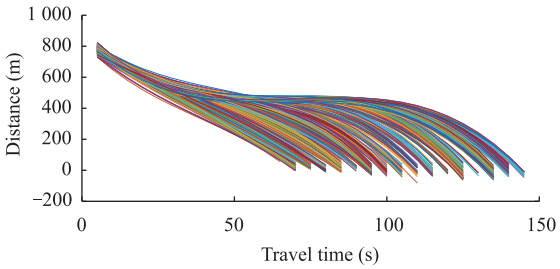

Fig. 9 A virtual trajectory dataset generated with a penetration rate of 50%

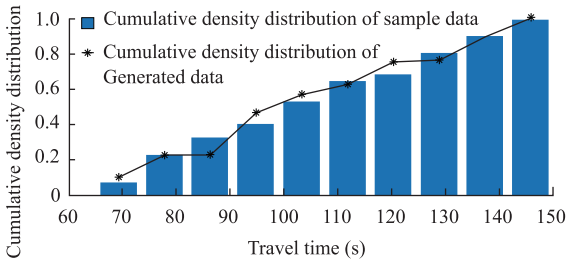

Fig. 10 Travel time cumulative probability distribution

Table 1 Travel time K-S test results under different penetration rates (form 10% to 100%)

Penetration rate (%)	100	90	80	70	60	50	40	30	20	10
H_0	0	0	0	0	0	10	0	0	0	0
P	0.99	0.97	0.97	0.99	0.67	0.99	0.99	0.67	0.97	0.67
K	0.1	0.2	0.2	0.1	0.3	0.1	0.1	0.3	0.2	0.3

4 Conclusions

In this paper, we propose a trajectory generation algorithm based on generative adversarial networks to protect the privacy of mobile trajectory data. We train the neural networks of generators and discriminators to produce a substantial amount of high-quality datasets that match the distribution of real data samples. This approach aims to ensure the privacy of mobile trajectory data while catering to diverse applications in traffic modeling analysis. In this paper, we develop a simulation model using real data from Changping District. We train the trajectory sample data with different penetration rates and then compare the reliability of the generated data with the travel time of the sample data. The experimental results show that the algorithm in this paper confirms the hypothesis of K-S test at various penetration rates with confidence levels of 0.01 and 0.05. This indicates that the two distributions satisfy the same distribution.

It can be seen that the algorithm in this paper has the advantages of protecting location privacy and being adaptable to sample sizes with different penetration rates. However, the algorithm's training capability for large data volumes in extensive areas is somewhat limited, and the network structure needs to be set based on experience. In future research, when dealing with larger datasets, we aim to enhance the generative adversarial network and choose a suitable network structure to broaden the training scope and improve the accuracy of the generated trajectory data.

Acknowledgement

This work is supported by National Key R&D Program Project (No. 2022YFB2601804) and Scientific Research Project of Beijing Educational Committee (No. KM201910009008).

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] ZHU N, LIU Y, MA S F, et al. Mobile Traffic Sensor Routing in Dynamic Transportation Systems [J]. IEEE Transactions. Intelligent Transportation Systems, 2014, 15 (5): 2273-2285.

[2] VACA M, TRAVER J E, MILANES V, et al. Automated Global Planner for Cybernetic Transportation Systems [C] //2016 14th International Conference on Control, Automation, Robotics and Vision (ICARCV). Piscataway: IEEE, 2016: 1-6.

[3] XIAO Z, CHEN Y X, ALAZAB M, et al. Trajectory Data Acquisition via Private Car Positioning Based on Tightly-coupled GPS/OBD Integration in Urban Environments [J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 23 (7): 9680-9691.

[4] HE W, HWANG K, LI D Y. Intelligent Carpool Routing for Urban Ridesharing by Mining GPS Trajectories [J]. IEEE Transactions. Intelligent Transportation Systems, 2014, 15 (5): 2286-2296.

[5] LI Z, PEI Q, MARKWOOD L, et al. Location Privacy Violation via GPS-Agnostic Smart Phone Car Tracking [J]. IEEE Transactions on Vehicular Technology, 2018, 67 (6): 5042-5053.

[6] ZHANG S, TAO J YUAN Y. Anonymous Authentication-Oriented Vehicular Privacy Protection Technology Research in VANET [C] //2011 International Conference on Electrical and Control Engineering. Piscataway: IEEE, 2011: 4365-4368.

[7] WANG J B, CAI Z P, YU J G. Achieving Personalized k-Anonymity-Based Content Privacy for Autonomous Vehicles in CPS [J]. IEEE Transactions on Industrial Informatics, 2020, 16 (6): 4242-4251.

[8] ECKHOFF D, SOMMER C. Driving for Big Data? Privacy Concerns in Vehicular Networking [J]. IEEE Security & Privacy, 2014, 12 (1): 77-79.

[9] HAMIDA E B, NOURA H, ZNAIDI W. Security of Cooperative Intelligent Transport Systems: Standards, Threats Analysis and Cryptographic Countermeasures [J]. Electronics, 2015, 4 (3): 380-423.

[10] ZHANG Z X, FENG T Y, SIKDAR B, et al. A Flickering

Context-based Mix Strategy for Privacy Protection in VANETs [C] //ICC 2021-IEEE International Conference on Communications. Piscataway: IEEE, 2021: 1-6.

[11] SAMPIGETHAYA K, LI M Y, HUANG L P, et al. AMOEBA: Robust Location Privacy Scheme for VANET [J]. IEEE Journal on Selected Areas in Communications, 2007, 25 (8): 1569-1589.

[12] LU R X, LIN X D, LUAN T H, et al. Anonymity Analysis on Social Spot Based Pseudonym Changing for Location Privacy in VANETs [C] //2011 IEEE International Conference on Communications (ICC). Piscataway: IEEE, 2011: 1-5.

[13] MEMON I, ALI Q, ZUBEDI A, et al. DPMM: Dynamic Pseudonym-Based Multiple Mix-Zones Generation for Mobile Traveler [J]. Multimedia Tools & Applications, 2017, 76 (22): 24359-24388.

[14] LU Z J, QU G LIU Z L. A Survey on Recent Advances in Vehicular Network Security, Trust, and Privacy [J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 20 (2): 760-776.

[15] BOUALOUACHE A, SENOUCI S, MOUSSAOUI S. A Survey on Pseudonym Changing Strategies for Vehicular Ad-Hoc Networks [J]. IEEE Communications Surveys and Tutorials, 2018, 20 (1): 770-790.

[16] SAXENA A S, BERA D, GOYAL V. Modeling Location Obfuscation for Continuous Query [J]. Journal of Information Security and Applications, 2019, 44 (FEB): 130-143.

[17] SHAHID A R, PISSINOU N, IYENGAR S S, et al. KLAP for Real-World Protection of Location Privacy [C] //2018 IEEE World Congress on Services (SERVICES). Piscataway: IEEE, 2018: 17-18.

[18] LIU Y, YU J J Q, KANG J, et al. Privacy-Preserving Traffic Flow Prediction: A Federated Learning Approach [J]. IEEE Internet of Things Journal, 2020, 7 (8): 7751-7763.

[19] MA Z, ZHANG T, LIU X, et al. Real-Time Privacy-Preserving Data Release over Vehicle Trajectory [J]. IEEE Transactions on Vehicular Technology, 2019, 68 (8): 8091-8102.

[20] YU H L, LI G S, WU J H, et al. A Location-Based Path Privacy Protection Scheme in Internet of Vehicles [C] // IEEE INFOCOM 2020 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). Piscataway: IEEE, 2020: 665-670.

[21] ZHENG W B, WANG K F, WANG F Y. Background Subtraction Algorithm Based on Bayesian Generative Adversarial Network [J]. Acta Automatica Sinica, 2018, 44 (5): 878-890.

[22] MURPHY K P. Dynamic Bayesian Networks: Representation, Inference and Learning [D]. Berkeley: University of California, 2002.

[23] SUN Z B, ZAN B, BAN X G, et al. Privacy Protection Method for Fine-Grained Urban Traffic Modeling Using Mobile Sensors [J]. Transportation Research Part B: Methodological, 2013, 56: 50-69.