ISSN 2616-5775 Vol. 7, Issue 1: 1-8, DOI: 10.25236/AJCIS.2024.070101

Personalized trajectory differential privacy protection mechanism based on spatiotemporal correlation prediction

Yanmei Shen[1], Rui Hua[2], Hui Wang[1,*], Zihao Shen[2], Peiqian Liu[1]

[1]School of Software, Henan Polytechnic University, Jiaozuo, China
[2]School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, China
[*]Corresponding author: wanghui_jsj@hpu.edu.cn

Abstract: *Aiming at the problems of user trajectory privacy budget and personalized demand of existing trajectory differential privacy protection technology, a personalized differential privacy protection mechanism based on sliding window and prediction perturbation is proposed. First, based on the road network topology, the sensitive road sections are classified into levels, and the allocation of personalized privacy budget is achieved by customizing the sensitivity of semantic locations. Then, recurrent neural networks and exponential perturbation methods are utilized to predict the perturbed locations that satisfy the differential privacy and temporal correlation requirements, and service similarity is introduced to detect location availability; If successful, the predicted location is directly used instead of the location of the differential perturbation, which reduces the privacy overhead from successive queries and further improves the utilization of the privacy budget. Finally, a w sliding window based trajectory budget allocation mechanism is designed to dynamically adjust the degree of privacy protection for each position in the trajectory according to the privacy needs of the user of the position. Experiments on real datasets show that the method can better achieve the balance between privacy and utility of trajectory data and improve the usability of published data while strictly protecting privacy.*

Keywords: *trajectory privacy; differential privacy; privacy budget allocation; temporal correlation*

1. Introduction

In recent years, with the rapid development of intelligent terminals with positioning function and mobile communication technology, a variety of location-based services have become increasingly popular, and now cover all aspects of national economy and social life, due to the fact that mobile users need to provide their location information to location-based services when enjoying convenient services [1], so that a large number of users' location information has been obtained by untrustworthy third parties, which may cause the users to suffer from serious location privacy leakage and jeopardize the user's privacy and security. Researchers have conducted studies on location privacy protection techniques and achieved rich research results [2].

Geographic indistinguishability is modeled by adding controlled noise to the user's location through a Laplace perturbation mechanism in a polar coordinate system, making it nearly impossible for an attacker to distinguish the difference between the approximate location and the true location, thus protecting the user's true location within a circular region of radius r [3].The definition of differential location privacy is based on a rigorous mathematical statistical model and can be used to control the level of privacy protection by adjusting the privacy parameters, thus attracting much attention. However, existing research on location differential privacy protection still suffers from the following problems; existing location differential protection mechanisms are only effective for single or sporadic queries, but in the case of multiple queries, the user's true location may still be exposed [4].This is due to the spatio-temporal correlation between successive queried locations, and thus the query not only leaks the privacy cost of the current location, but also increases the privacy risk of other locations. Therefore, even if a single location satisfies the ε-differential privacy requirement, it does not ensure that the entire trajectory satisfies ε-differential privacy [5].

Location prediction is used to estimate the user's current location by looking at publicly available information (e.g. historical trajectories). Since the prediction is independent of the user's real location and does not provide more useful information to the attacker, the privacy loss is minimal or even negligible [6]. Therefore, this paper attempts to reduce the privacy cost and minimize the privacy risk by adopting a prediction mechanism instead of differential perturbation. In this context, this paper proposes a personalized trajectory differential privacy protection mechanism based on spatio-temporal correlation prediction, and its main contributions are as follows.

(1) We propose a differential query strategy that integrates predictive perturbation to optimize the privacy cost in continuous queries. Leveraging recurrent neural networks and exponential perturbation, we predict query locations that meet differential privacy and spatio-temporal criteria. A service similarity map assesses the availability of these locations. Queries are executed at predicted locations if available, or at differentially perturbed locations otherwise, enhancing query efficiency and reducing privacy costs.

(2) To maintain data availability, we introduce a sliding window mechanism. It dynamically adjusts privacy protection levels for each location in a trajectory, considering location predictability and significance.

(3) By analyzing the topological relationships in road networks, we categorize privacy levels for road segments near sensitive areas. This system allows users to customize location sensitivity, enabling personalized privacy budget allocation and optimizing its utilization.

2. Trajectory privacy protection method.

Based on the observation that the predicted locations are usually randomly distributed around the user's real location, the attacker often cannot infer the user's real location through the predicted locations, so the prediction mechanism can be used as a kind of random perturbation[7]. Since the privacy overhead of the prediction mechanism is very small, the prediction mechanism can be utilized to replace differential privacy thus effectively reducing the privacy overhead caused by continuous queries. At the same time, because of the differentiation of user privacy, according to the user's demand for different location points, the differential privacy budget can be rationalized and allocated, which can effectively improve the utilization rate of the privacy budget and save the budget expenditure. Trajectory location has temporal correlation, if only the location release on a single moment is considered and the correlation between the trajectory data is ignored, although a single location satisfies ε-differential privacy, it does not ensure that the trajectory satisfies ε-differential privacy. Therefore, based on the above observation, this paper proposes a personalized trajectory differential privacy protection mechanism based on spatio-temporal correlation prediction, which mainly contains the following strategies. The framework of the proposed algorithm is illustrated in Figure 1:

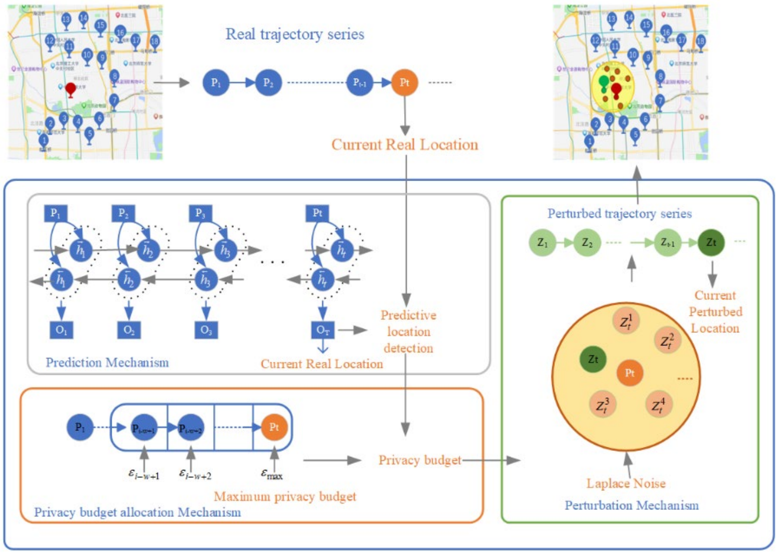


Figure 1: Personalized trajectory differential privacy protection mechanism based on spatiotemporal correlation prediction

2.1. Sensitivity Processing.

The method assigns different privacy level values based on the user's preset sensitive locations and travel modes, and sets different differential privacy budgets to provide different degrees of privacy protection. However, if the geographic topological relationship is ignored and only the set of predefined sensitive locations is anonymized, other location points in the trajectory may also lead to the leakage of privacy information.To address the above issues, this paper considers the overall connectivity between location points, combines the user's travel mode, and according to the results of the region division and the initial location set of sensitive points, the whole map can be divided into three parts, SL for the initial sensitive location set, NA for the logically unavailable regions or regions that have a great deviation from the user's frequent stops, and NSL for the initial location set of non-sensitive locations transforms the map according to the semantics into a The map is transformed into an undirected graph according to the semantics, and the distances between the location points are represented by Euclidean distances. Considering the overall connectivity of the locations, a certain sensitivity is assigned to the semantic locations that have high correlation with the sensitive locations. The set of highly correlated location points near the sensitive location points is defined as , and the spatio-temporal dimensional features of the trajectory data can be extracted using a graph attention network, using the correlation coefficients and Euclidean distances of the historical data between individual nodes to define the magnitude of the weights, specifically, given the datasets of any two nodes A and B on a continuous time series in the past as $H_A = \{a_1, a_2, \cdots a_n\}$ and $H_B = \{b_1, b_2, \cdots b_n\}$ The correlation coefficient of A and B is set as :

$$r = \frac{\mathrm{Cov}(H_A, H_B)}{\sqrt{DH_A}\sqrt{DH_B}} = \frac{\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{n}(a_i - \bar{a})^2}\sqrt{\sum_{i=1}^{n}(b_i - \bar{b})^2}} \tag{1}$$

Considering the spatial correlation between nodes, the equivalent distance between nodes A and B is characterized using $eDis$. The inverse of $eDis$ is used to describe the magnitude of the correlation.

$$e\mathrm{Dis} = \begin{cases} 1, & g = 0 \\ \exp\left(-\left(\frac{Dis * g}{\delta}\right)^{(1-r)}\right), & g \neq 0 \end{cases} \tag{2}$$

where c is the shortest hop between nodes of the road network, $neighborSet = \{g \,|\, g.eDis < b\}$, where b is a threshold set by the user, and finally the privacy sensitivity assigned to any location in the connection set of α, as shown in Eq(10).

$$g.pl = \frac{\left[\frac{1}{g.eDis}\right]}{\sum_{g' \in neighborSet} \frac{1}{g'.eDis}} \times (a.pl) \tag{3}$$

In order to facilitate the calculation, this paper generates a sensitive map by calculating the sensitivity of each region from the gridded map using the above process. First, the sensitivity of the location is obtained based on the user's real location, and if the sensitivity of the location is less than the threshold set by the user, the user's real location is directly utilized for querying services without executing the privacy protection algorithm, which can improve the efficiency of trajectory privacy protection.

2.2. Predictive Mechanisms.

The development of recurrent neural networks has given rise to GRU (Gate recurrent unit), which can effectively deal with the above problems and capture long range dependencies[8]. In order to improve the accuracy of the prediction mechanism, BiGRU improves the GRU, BiGRU network realizes the full understanding of the historical data through the forward and reverse dimensions, which has a great performance improvement compared to the unidirectional GRU network, and the prediction results are more stable its structure[9]. Using BiGRU network as a prediction framework, its model is:

$$\begin{cases} \vec{h}_t = GRU \quad {}^{t}\mathbf{s}_i, \vec{h}_{t-1} \\ \overleftarrow{h}_t = GRU \quad {}^{t}\mathbf{s}_i, \overleftarrow{h}_{t-1} \end{cases} \tag{4}$$

Where: $\vec{h}_t$ denotes the forward hidden state of the forward GRU network output at moment t; $\overleftarrow{h}_t$ denotes the reverse hidden state of the reverse GRU network output at moment t. The output of BiGRU is further processed as follows. In addition, the output results of BiGRU are further processed:

$$\hat{h}_t = \omega_{t1}\vec{h}_t + \omega_{t2}\overleftarrow{h}_t + b_t \tag{5}$$

Where: $\hat{h}_t$ is the input for the output of ${}^{t}\mathbf{s}_i$ network; ω_{t1}, ω_{t2} are the forward and back propagation weight matrices, respectively; b_t is the bias matrix; BiGRU is utilized to train the real dataset according to the temporal correlation between the user's real position to complete the prediction of trajectory position, in the process of the model training, BiGRU refers to both sides of the prediction point at the same time, to get the corresponding trajectory characteristics of the moment of the attention weights. The location points with higher sensitivity will be given higher weights; finally the output vector is weighted and summed to derive the predicted location point set, due to the lower probability of some elements in the location point set derived through the computation, combined with the ξ-location set to filter out the elements with lower probability to get the candidate set ΔXt .For privacy preserving effect this paper selects the exponential mechanism for its selection, $E(\varepsilon_e)\Delta Xt \to l$, where ε_D is the privacy budget of the exponential selection mechanism; the scoring function is $f = f^{(t)} = \{f_1^{(t)}, \cdots, f_m^{(t)}\}$, and m is the number of elements in the ξ-location set. This mechanism adds noise to the output result, but the elements with high probability are still output with larger probability, thus making the output result more private and more reasonable. Through the privacy budget allocation mechanism, this paper obtains the privacy budget ε_D of the exponential mechanism respectively . $E(\varepsilon_D, \Delta Xt)$ is the privacy budget for ε_D , through the exponential perturbation mechanism in outputting the predicted position O_T . Algorithm1 gives the process of predicted location generation.

Algorithm 1 Predictive Mechanisms

```
Input: raw trajectory dataset D, ξ-position collection
Output: predicted location:O_T;
for T∈D do;
  for p∈T do;
    x_BiGRU=BiGRU(D,x);
    y_BiGRU=BiGRU(D,y);
    end for;
end for;
Calculate the ξ-position collection Set $\Delta Xt$,based in ξ;
O_T ← S( $\varepsilon_D$ ,$\Delta Xt$);
return O_T.
```

In order to evaluate the service quality of the predicted location, this paper proposes a service similarity based detection function. Where the detection function based on service similarity $S(\varepsilon_\theta, \alpha, O_T)$:

$$S\left(\varepsilon_\theta, \alpha, O_T\right)(x) = \begin{cases} 0, sim(p, O_T) & \geq \alpha + Lap(\varepsilon_\theta) \\ 1, & other \end{cases} \tag{6}$$

WhereO_T is the predicted location; $sim(p, O_T)$ is the service similarity between location p and O_T , and the output "0" indicates successful prediction and "1" indicates failed prediction; $Lap(\varepsilon_\theta)$ denotes the perturbation value of $Lap(\varepsilon_\theta)$ when the privacy budget is ε_θ. The detection function in this paper also introduces a perturbation mechanism in order to ensure the security. Since the detection function will inevitably leak part of the user's location privacy, in order to ensure security, this paper also introduces the perturbation mechanism of $Laplace$ in the detection function. Although the detection function uses up part of the privacy budget, the value of the perturbation privacy budget that satisfies the ε-geographic indistinguishability is smaller, thus saving the privacy budget and improving the service quality at the same time.

2.3. Privacy budget allocation mechanism for w-sequences satisfying ε-differential privacy

From the query mechanism, it can be seen that each location needs to be allocated a privacy budget of ε_D , ε_θ and ε_N in the exponential perturbation phase, the detection function detection phase and the perturbation phase of geographic indistinguishability, respectively. If a timestamped fake location is predicted successfully, the predicted location is utilized as the query location. The process of generating the query location goes through two phases, the exponential perturbation mechanism, and the detection function, and thus costs a privacy budget of $\varepsilon_D + \varepsilon_\theta$. If the detection fails, the query point is generated using the perturbation of geographic indistinguishability, which undergoes three stages and thus costs the sum of ε_D , ε_θ , and ε_N privacy budget. If the prediction fails, instead, more privacy budget is spent. To ensure that a trajectory satisfies w-trajectory sequence differential privacy, we use the *w*-sliding window mechanism to assign a corresponding privacy budget to each position in the trajectory. A *w*-sliding window is a sequence of locations under w consecutive timestamps, i.e., consecutive trajectory segments of length *w*, of the form shown in Figure 2.

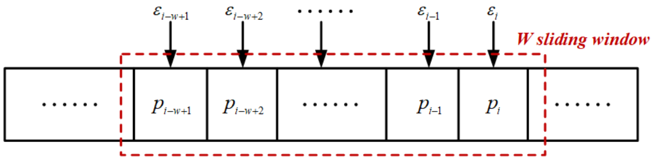


Figure 2: w sliding window indication

In order to make the trajectory satisfy ε-differential privacy, the parameters *S*, *D* and *N* are introduced to regulate the privacy budget, and $S + D + N = 1$. For the exponential perturbation mechanism, each position is assigned a privacy budget ε_D of $\frac{D\varepsilon}{w}$. In the exponential scoring mechanism, because of the sequential combinatorial property of the scoring function, each position in the ξ-position set is assigned a privacy budget $\frac{\varepsilon_D}{m}$. In the detection phase, each location is assigned a privacy budget ε_θ of $\frac{S\varepsilon}{w}$ .The privacy budget allocation method is shown in Algorithm2.

Algorithm 2 Privacy budget allocation

Input: p-current location, O_T-predicted mechanism, α, SenMap, θ, *w*-window size, z, SimMap, ε-total privacy budget, *S*, *D*, *N*;

Output: allocated privacy budget ε_i;

```
p.pl← lookup(SenMap,p);
   if p.pl≤θ then;
      z←p;
   else;
      $\varepsilon_D \leftarrow D\frac{\varepsilon}{w}$;
$\varepsilon_\theta \leftarrow S\frac{\varepsilon}{w}$ , $\varepsilon_N \leftarrow N\frac{\varepsilon}{w}$;
End if
If test(p,α, SimMap , $\varepsilon_\theta$)=0 then;
z← $O_T$   $\varepsilon_i = \varepsilon_D + \varepsilon_\theta$;
else;
$\varepsilon_N = \frac{N\varepsilon\theta}{w(g.pl)}$;
$\varepsilon_i = \varepsilon_D + \varepsilon_\theta$;
End if;
```

Due to the variability of privacy, equal perturbation at each point does not fulfill the requirement. Therefore, this paper introduces a personalized privacy protection strategy to improve the utilization of privacy budget. For locations with higher sensitivity, a smaller privacy budget is allocated, i.e., the larger perturbation noise is added, thus obtaining a higher degree of privacy protection. First, the sensitivity radius R , and the initial sensitivity θ corresponding to each sensitive location point are calculated based on the user's pre-set sensitive location points, non-reachable points, and user-acceptable error distance value Δ etc. Assume that $\mu = (x_0, y_0)$ is the user's current true location, and

$q=(x_0+rr\cos\theta, y_0+rr\sin\theta)$ is the generated false location. The distortion distance between the user's true position and the false position can be expressed as follows:

$$rr=\left(p_0,q\right)=\sqrt{\left(x-x_0\right)^2+\left(y-y_0\right)^2} \tag{7}$$

The distortion distance between the user's true location and the false location needs to satisfy the following relationship based on geographic indistinguishability:

$$rr=\frac{-1}{\varepsilon_i}\left(W_{-1}\left(\frac{\tau-1}{e}\right)+1\right), \tau=rand(0,1) \tag{8}$$

If you want to generate a false position that satisfies the user's needs, the distortion distance between the user's true position and the false position must be less than the user's acceptable error distance value, based on the above formula, there is the following formula:

$$R=\frac{-Sum}{\varepsilon\Delta}\left(W_{-1}(\frac{\tau-1}{e})+1\right)\le d_{i,j} \tag{9}$$

Considering the location sensitivity, the higher the sensitivity, the lower the privacy budget allocated for the noise perturbation phase and the smaller the protection radius. Thus, for any sensitivity $g.pl$, the choice of its radius is $R=\frac{R\theta}{g.pl}$,such that for perturbations that satisfy geographic indistinguishability, each location is assigned a privacy budget ε_N of :

$$\varepsilon_N=\frac{N\varepsilon\theta}{w\left(g.pl\right)} \tag{10}$$

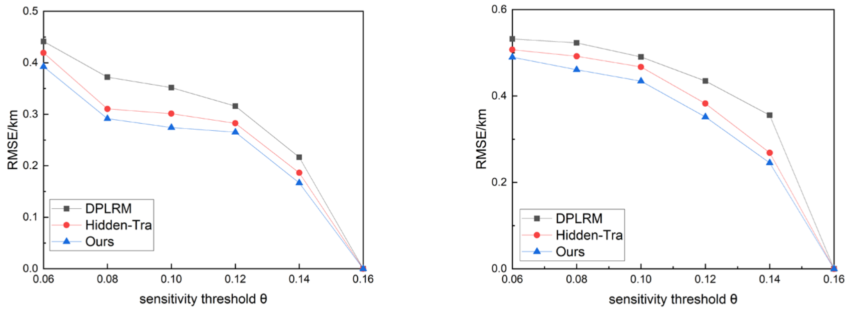
where S + D ≤ N and $g.pl\in[\theta,1]$,. In order to save the privacy budget, the budget value in the prediction phase is smaller than the budget value in the perturbation phase, so S + D ≤ N.

3. Experimental analysis and evaluation

This paper analyzes the impact of model parameters on availability in two real datasets. Meanwhile, this paper's scheme is compared and analyzed with the DPLRM [10] mechanism and the Hidden-Tra [11] mechanism, so as to verify the effectiveness of this paper's scheme. In this paper, we use two real datasets, Geolife and TaxiService, attributes such as user number, timestamp, dimension, and longitude are selected for experiments. The RMSE is defined in the following equation:

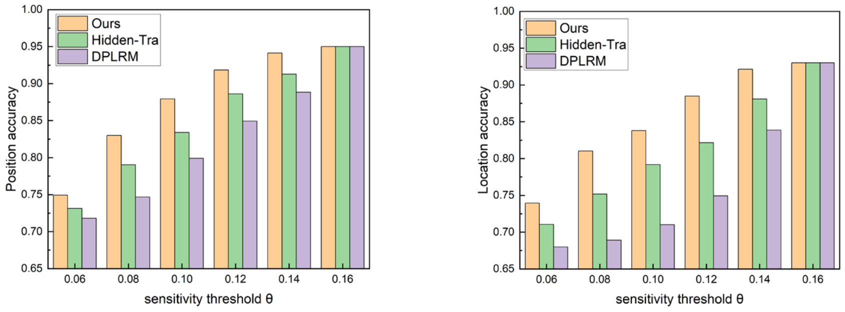
$$RMSE=\frac{1}{n}\sum_{i=1}^{n}d\left(p_i,z_i\right) \tag{11}$$

First, this paper analyzes the effect of sensitivity threshold θ on RMSE and location correctness in Geolife and TaxiService datasets, respectively, and the results are shown in Figure3. From Figure4, it can be seen that the usability of this paper's scheme increases with θ. The reason for this trend is twofold: first, as θ rises, the privacy budget allocated to each location in the perturbation phase improves, thus increasing availability; second, more locations are less sensitive than θ and thus do not need to be perturbed, further improving availability. When θ is set to 0.16, the RMSE drops to 0, as our sensitivity segmentation method deems all map locations non-sensitive, allowing direct publication and resulting in zero average RMSE for trajectories. Moreover, the Geolife dataset shows better availability compared to the TaxiService dataset, which uses larger grid divisions and has greater distances between locations, reducing availability. Our paper's approach also surpasses the Hidden-Tra and DPLRM schemes in usability. It achieves a higher prediction success rate than Hidden-Tra and considers personalized privacy settings, enabling more effective privacy budget use, reduced perturbation uncertainty, and enhanced usability. While the DPLRM scheme balances privacy and usability, considering the impact of published locations on current and past locations, it faces constraints. Our scheme introduces usability detection, further improving usability.



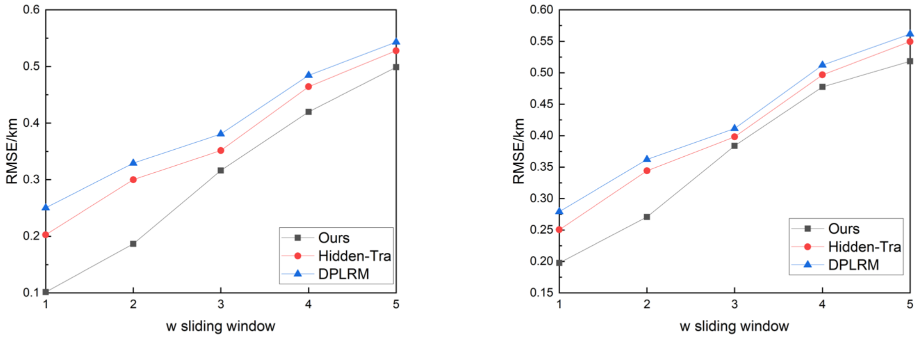
(a) Geolife Data Set *(b) TaxiService Data Set*

Figure 3: The impact of sensitivity threshold θ on RMSE



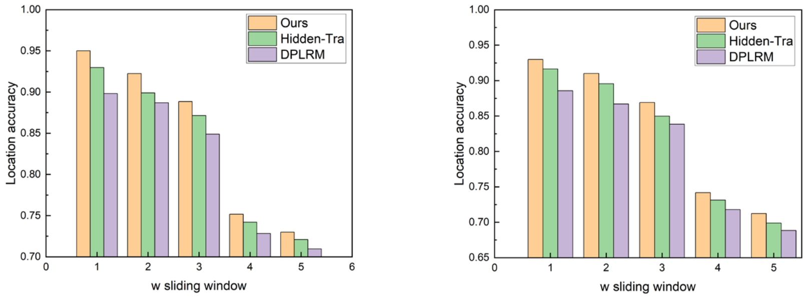
(a) Geolife Data Set *(b) TaxiService Data Set*

Figure 4: The impact of sensitivity threshold θ on position accuracy



(a) Geolife Data Set *(b) TaxiService Data Set*

Figure 5: The impact of w sliding window on RMSE



(a) Geolife Data Set *(b) TaxiService Data Set*

Figure 6: The impact of w sliding window on position accuracy

This paper examines the impact of sliding window length on usability, as depicted in Figure 5. The results show that the RMSE increases with the sliding window size. A larger w results in a smaller privacy budget for each location, reducing prediction success and necessitating more perturbations to achieve geographic indistinguishability, thereby lowering privacy availability. In the Geolife dataset, availability is higher compared to the TaxiService dataset. Our method outperforms both the DPLRM and Hidden-Tra schemes in terms of availability. The DPLRM scheme, with its stringent location publishing constraints, struggles with distant locations. Our approach enhances availability by incorporating availability detection. From Figure 6, it's evident that the location correctness rate declines with increasing w, yet our method maintains higher correctness rates than both the DPLRM and Hidden-Tra schemes, particularly in the Geolife dataset.

4. Conclusion

This paper introduces a personalized differential privacy mechanism for trajectory data, combining sliding window and predictive perturbation techniques. It emphasizes spatio-temporal correlation in its design and implementation, categorizing sensitive road sections to tailor semantic location sensitivity. This approach enhances privacy protection and data availability for personalized trajectory privacy schemes. The method employs recurrent neural networks and exponential perturbation to predict locations meeting differential privacy and temporal correlation criteria, with service similarity checks ensuring location availability and maintaining original trajectory spatio-temporal consistency. Additionally, a sliding window-based mechanism allocates trajectory budgets to boost data availability. Future work will focus on optimizing algorithm runtime and addressing privacy leaks in semantic trajectory protection.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (61300216), Doctoral Scientific Fund of Henan Polytechnic University (B2022-16) and Key scientific research funding projects for colleges and universities in Henan Province (23A520033).

References

[1] WANG F, LI G, WANG Y, et al. Privacy-aware traffic flow prediction based on multi-party sensor data with zero trust in smart city[J]. ACM Transactions on Internet Technology, 2023, 23(3): 1-19.

[2] SHEN S, ZHU T, WU D, et al. From distributed machine learning to federated learning: In the view of data privacy and security[J]. Concurrency and Computation: Practice and Experience, 2022, 34(16): e6002.

[3] YE A, ZHANG Q, DIAO Y, et al. A Semantic-Based Approach for Privacy-Preserving in Trajectory Publishing[J]. IEEE Access, 2020, 8: 184965-184975.

[4] ZHANG J, HUANG Q, HUANG Y, et al. DP-TrajGAN: A privacy-aware trajectory generation model with differential privacy[J]. Future Generation Computer Systems, 2023, 142: 25-40.

[5] XIONG X, LIU S, LI D, et al. Real-time and private spatio-temporal data aggregation with local differential privacy[J]. Journal of Information Security and Applications, 2020, 55: 102633.

[6] YANG H, VIJAYAKUMAR P, SHEN J, et al. A location-based privacy-preserving oblivious sharing scheme for indoor navigation[J]. Future Generation Computer Systems, 2022, 137: 42-52.

[7] QIU S, PI D, WANG Y, et al. Novel trajectory privacy protection method against prediction attacks[J]. Expert Systems with Applications, 2023, 213: 118870.

[8] CHOI S, KIM J, YEO H. TrajGAIL: Generating urban vehicle trajectories using generative adversarial imitation learning[J]. Transportation Research Part C: Emerging Technologies, 2021, 128: 103091.

[9] KIM J W, JANG B. Deep learning-based privacy-preserving framework for synthetic trajectory generation[J]. Journal of Network and Computer Applications, 2022, 206: 103459.

[10] WU Y, CHEN H, ZHAO S, et al. Differentially private trajectory protection based on spatial and temporal correlation[J]. Chinese journal of computers, 2018, 41(2): 309-322.

[11] JIA Jun-jie, QIN Hai-tao. Anonymity of dynamic trajectory based on genetic algorithm[J]. Computer Engineering & Science, 2021, 43(01): 142-150.