symmetry

MDPI

Article

Trajectory Privacy-Protection Mechanism Based on Multidimensional Spatial–Temporal Prediction

Ji Xi [1,*], Meiyu Shi [1], Weiqi Zhang [1], Zhe Xu [1] and Yanting Liu [2]

[1] School of Computer Information Engineering, Changzhou Institute of Technology, No. 666, Liaohe Road, Changzhou 213022, China; 21030523@czust.edu.cn (M.S.); zhangwq@czust.edu.cn (W.Z.); xuz@czust.edu.cn (Z.X.)

[2] School of Software and Big Data, Changzhou College of Information Technology, Changzhou 213032, China

* Correspondence: xiji@czust.edu.cn

Abstract: The popularity of global GPS location services and location-enabled personal terminal applications has contributed to the rapid growth of location-based social networks. Users can access social networks at anytime and anywhere to obtain services in the relevant location. While accessing services is convenient, there is a potential risk of leaking users' private information. In data processing, the discovery of issues and the generation of optimal solutions constitute a symmetrical process. Therefore, this paper proposes a symmetry–trajectory differential privacy-protection mechanism based on multi-dimensional prediction (TPPM-MP). Firstly, the temporal attention mechanism is designed to extract spatiotemporal features of trajectories from different spatiotemporal dimensions and perform trajectory-sensitive prediction. Secondly, class-prevalence-based weights are assigned to sensitive regions. Finally, the privacy budget is assigned based on the sensitive weights, and noise conforming to localized differential privacy is added. Validated on real datasets, the proposed method in this paper enhanced usability by 22% and 37% on the same dataset compared with other methods mentioned, while providing equivalent privacy protection.

Keywords: local differential privacy; trajectory prediction; trajectory data publishing; location-based services

check for updates

Citation: Xi, J.; Shi, M.; Zhang, W.; Xu, Z.; Liu, Y. Trajectory Privacy-Protection Mechanism Based on Multidimensional Spatial–Temporal Prediction. *Symmetry* **2024**, *16*, 1248. https://doi.org/10.3390/sym16091248

Academic Editor: Lorentz Jäntschi

Received: 29 July 2024
Revised: 13 September 2024
Accepted: 15 September 2024
Published: 23 September 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The Internet industry, mobile communications, cloud computing, the Internet of Things, and other emerging technologies are rapidly developing, and smart devices with positioning functions are rapidly becoming popular. Human beings have opened a new era with intelligent interconnection and information sharing as the main symbols. These new intelligent experiences are mainly based on location-based services (LBSs), which provide specific and precise location-related services to provide users with convenient and favorable experiences. These location and trajectory data contain sensitive and complex information with essential commercial and immeasurable academic value in urban planning, disaster warning, and other public security [1,2]. Location information is a publicly available resource, but malicious attackers with ulterior motives to connect it with the relevant users can lead to serious privacy leakage problems and even crises for users' personal and property safety.

Trajectories' privacy protection methods are divided into the following types: trajectory generalization, trajectory suppression, trajectory encryption, dynamic pseudonyms, and trajectory protection methods based on differential privacy techniques. Trajectory generalization is a classical privacy protection method for location data, and most such protection methods are based on the k-anonymity technique [3]. The basic idea is to require each data release to make each package of released data indistinguishable from the other $k - 1$ entries. Trajectory suppression methods [4], on the other hand, are accomplished under the assumption that third-party anonymization servers are entirely reliable, and the most basic idea is to remove from a trajectory specific locations with sensitive identifiers that are frequently accessed by users, before the trajectory data is published. Track encryption encrypts the user's LBNS query information. The private information retrieval (PIR) technique [5,6] is a method of location information protection based on cryptographic techniques and theories that allows the user to retrieve needed information from a database without revealing the information to be retrieved. Dynamic pseudo-anonymization replaces the user's accurate ID information with a pseudonym when the user sends a request. The privacy-preserving model based on the differential privacy (DP) technique [7] provides strict data definition in terms of privacy preservation. It does not need to consider the background knowledge possessed by the attacker and is not affected by the change of a particular piece of data. DP was initially applied to querying databases to protect the individuals in the databases when releasing statistical information. Noise conforming to the Laplace distribution is added to the trajectory data to ensure that the query results satisfy the definition of differential privacy protection [8,9]. Since then, applying DP techniques in spatial geography has opened a new chapter in trajectory information privacy protection. The problem that all the current differential privacy-based methods for protecting user trajectory information must face is the relationship between privacy protection and data availability. How can errors caused by noisy data be reduced on the basis of protecting the user's privacy information to improve the data's usability? How can the degree of privacy protection of trajectory data be maximized while ensuring that the privacy protection scheme satisfies differential privacy?

Compared with previous approaches, the time-attentive sensitive area prediction mechanism we propose is highly innovative. Previous methods predicted in only the spatial dimension, whereas our work extends this prediction to the temporal dimension on the basis of spatial considerations. This not only enhances the accuracy of the prediction results but also ensures that the generated pseudo-trajectories align with the characteristics of user mobility.

The rest of this paper is organized as follows. Section 2 presents our related work on trajectory privacy-preserving methods. In Section 3, we discuss the relevant notions of trajectory privacy from the literature. In Section 4, we then describe several components and definitions of our TPPM-MP mechanism and introduce our temporal–spatial constraints areas-of-interest detection algorithm and trajectory-publishing method based on LDP. The experiment and evaluation are presented in Section 5.

2. Related Work

Differential privacy-preserving models are becoming a mainstream technique in the privacy-preserving field because of their excellent level of privacy preservation and portability. The basic idea is to randomly add noise conforming to Laplace distribution to the query result of the original data so that adding or deleting a particular record in the dataset does not affect the query result. Therefore, to realize privacy protection, no matter how much background knowledge the attacker has, it is challenging to infer through the query result whether the target data are in the queried dataset. Chen et al. [10] applied a differential privacy protection mechanism to the privacy protection of location data, adding the noise conforming to the Laplace distribution to the original trajectory dataset to make the published trajectory data satisfy the definition of differential privacy, to protect the user's trajectory information. Previous research [11] proposed a data mining algorithm for differential privacy, using the quadtree spatial decomposition technique to preprocess the location points to realize differential privacy. Xiao et al. [12] transformed the geographic coordinate system into two coordinate systems. They assigned a privacy budget to each location, and the report mentioned that a Markov model represents the user's location relationship in the associated moments. A spatiotemporal location protection scheme based on differential privacy was proposed in the literature. Lu et al. [13] proposed a method called a Lagrange multiplier-based differentially private algorithm to optimize the budget of the privacy mechanism to prevent the budget from being too large or too small, which would result in adding too little or too much noise. The protected trajectory privacy data can effectively defend against the problem of attackers inferring social relationships through the trajectory data. The above literature realizes trajectory protection from the perspective of noise generation methods or considering sensitive locations in the user's trajectory location and adding noise to the whole trajectory. These approaches make overly strong assumptions about trajectory protection while ignoring the differences in the degree of privacy protection required at different locations in the trajectory. Not all location points leak the user's private information. Not all locations need to be protected, leading to too low data availability and even less guarantee of the service quality of the user's LBNSs. Therefore, as described in this paper, we predict sensitive areas through multidimensional spatial–temporal attention based on user movement patterns and then protect the detected sensitive areas.

In the era of big data, statistical models and machine learning methods cannot handle large-scale multivariate time series data with high dimensionality and nonlinearity. With the rapid rise of deep learning, the field of time series forecasting has been further developed. A significant advantage of deep learning models is that they can extract features from shallow information for analysis, and these features can further generate deep features [14]. As a result, deep learning models are more effective for solving complex problems than traditional models. Various improved neural networks such as RNNs, convolutional neural networks (CNNs), and graph neural networks (GNNs) have been proposed to mine temporal and spatial dependencies within time series. These are the most popular, efficient, and widely used deep learning techniques. Liu et al. [15] proposed a dual-stage two-phase model (DSTP)-based approach for extracting spatial correlations simultaneously, spatiotemporal relationships at different times, and temporal relationships between different sequences. GNNs are more widely used with multiple time-series data, e.g., traffic flow data based on road networks and air quality monitoring data in multiple areas in a city. Wang et al. [16] designed a graph convolutional network (GCN) to learn the topology of a sensor network to capture spatial correlations for traffic safety prediction. Song et al. [17] proposed a spatiotemporal synchronization mechanism to capture local spatiotemporal correlations for traffic flow prediction. However, these dynamic spatial correlations were localized due to the limitation of the neighborhood range. To solve these problems, Wang et al. [18] introduced geospatial convolution to obtain complex spatial relationships between regions for traffic accident risk prediction. Although the above GCN-based models have achieved significant performance results, some limitations remain. For multivariate time series data with actual geographic locations, not only do the data from different locations interact with each other, but they are also affected by exogenous factors relating to the current location. However, the above methods learn only one graph structure, which makes it challenging to capture the spatial–temporal correlations at different scales.

3. Related Definitions

The concept of differential privacy was first proposed as a definition by Dwork [19] in response to the problem of privacy leakage in statistical databases. This approach aims to make database query results insensitive to changes in individual records in the data set. First, the model is based on a rigorous mathematical theory, which provides a strict definition of privacy protection and scientific and rigorous proof of the level of privacy protection. Second, DP rigorously defines a privacy-preserving model entirely independent of background knowledge and theoretically assumed to be resistant to any attack originating from background knowledge. The key definitions of differential privacy are presented below.

3.1. Definition 1 (ε-Differential Privacy)

Suppose there are two neighboring datasets D and D', that differ by only one record, and there exists an algorithm M; $Range(M)$ is the set of all possible output values of the algorithm M. If any output $S \subseteq Range(M)$ is satisfied for any pair of neighboring datasets D and D', the following applies:

$$\Pr(M(D) = S) \leq \Pr\big(M(D') = S\big) \times \mathrm{e}^{\varepsilon} \tag{1}$$

Then, the algorithm M is said to satisfy ε-differential privacy [20], with $\varepsilon \in (0,1)$ denoting the degree of privacy protection. In general, the smaller ε is, the more noise needs to be added and the higher the degree of privacy protection.

3.2. Definition 2 (Sensitivity)

There exists a function $f : D \to R^d$; the input is a dataset D, and the output is a d-dimensional vector of real numbers. Then, for any neighboring datasets D and D', the sensitivity of the function query $f(D)$ is as follows:

$$S_f = \max_{D,D'} \|f(D) - f(D')\|_1 \tag{2}$$

where $\|f(D) - f(D')\|_1$ denotes the first-order paradigm distance of the query function to the query result on the neighboring dataset [21], and S_f denotes the sensitivity of the function f.

The primary implementation of differential privacy is to add noise to the query data results, and the typical noise mechanisms are categorized into the Laplace mechanism for numerical data and the exponential mechanism for non-numerical data. The exponential mechanism [22] for non-numerical data requires introducing a scoring function to obtain a score for each possible output, which is normalized to the probability value returned by the query. This paper focuses on using the Laplace mechanism to generate numeric data. The idea of the mechanism is to generate noise that meets the Laplace distribution and add the noise data to the original data; the formula is shown below:

$$f'(D) = f(D) + N \tag{3}$$

The noise N obeys the Laplace distribution $Lap(\mu, b)$, where $\mu = 0, b = \frac{S_f}{\varepsilon}$, and satisfies Equation (1) after noise addition, which results in the probability density function of the noisy data conforming to the Laplace distribution:

$$p\left(x|0, \frac{S_f}{\varepsilon}\right) = \frac{\varepsilon}{2\Delta f} \times \mathrm{e}^{\frac{-\varepsilon|x|}{\Delta f}} \tag{4}$$

In order to visualize the noisy data more, we draw its probability density image based on $Lap(\mu, b)$, as shown in Figure 1.

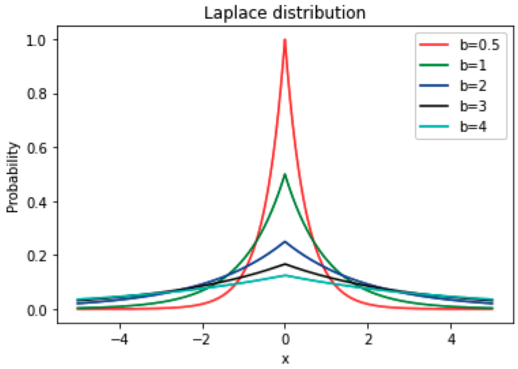


Figure 1. Laplace noise probability density.

From Figure 1, we can observe that the image exhibits central symmetry. Among these parameters, μ and b determine the shape of the Laplace probability density function; the lower the value of b, the more noise we need to add.

The main difference between local differential privacy (LDP) and centralized differential privacy (CDP) is the different processing methods. Differential privacy scrambles the collected data and thus requires a trusted central server to aggregate and process the data. In contrast, local differential privacy pre-processes the data locally and saves it for uploading without the intervention of a central server to process the data, thus better protecting privacy.

3.3. Definition 3 (Local-Differential Privacy)

Any localized differential privacy function $f(l)$ with domain of definition $Dom(f)$ and domain of values $Ran(f)$ has for any inputs l and $l' \in Dom(f)$ and output $l^* \in Ran(f)$:

$$-\varepsilon \leq \ln\left(\frac{\Pr(f(l) = l^*)}{\Pr(f(l') = l^*)}\right) \leq \varepsilon \tag{5}$$

According to the above formula, local differential privacy [21] ensures that the function $f(l)$ satisfies ε-local differential privacy by controlling the similarity of the output results of any two records; the smaller ε is, the higher the similarity of the output results of the two records, and vice versa. The mathematical definition of local differential privacy and differential privacy is the same, but the realization of the mechanism is very different. The mainstream data perturbation technique of local differential privacy is a randomized response. Local differential privacy can be applied without considering the background knowledge of the attacker and without relying too much on a trusted third-party centralized service provider.

Randomized response techniques are the dominant perturbation mechanism to achieve local differential privacy and reduce errors caused by respondents' wrong answers in sensitive question situations. For example, depending on the optional answer and sensitive questions, there are two cases of yes or no; in this paper, it is required to give an answer based on whether the user gives an answer based on the heads or tails of a uniform coin in n location blocks, assuming that the probability that the coin lands heads-up is p and the probability that it lands tails- up is $1-p$. The user responds to either the proper answer or an answer contrary to the truth, based on the result of the coin toss. If the actual situation is that the proportion of users in some of the n blocks is α, the position of the result answering yes against the block is k and answering otherwise is $n-k$. Then, the proportion of points answering yes or no according to the above is given:

$$P(ans = \text{'}Yes\text{'}) = \alpha p + (1-\alpha)(1-p) \tag{6}$$

$$P(ans = \text{'}No\text{'}) = (1-\alpha)p + \alpha(1-p) \tag{7}$$

The excellent likelihood estimate of the proper proportion is as follows:

$$\hat{\alpha} = \frac{p-1}{2p-1} + \frac{k}{(2p-1)n} \tag{8}$$

The mathematical expectation of $\hat{\alpha}$ is as follows:

$$\begin{aligned} E(\hat{\alpha}) &= \frac{1}{2(p-1)}\left[p - 1 + \frac{1}{n}\sum ans\right] \\ &= \frac{1}{2(p-1)}[p - 1 + \alpha p + (1-\alpha)(1-p)] \\ &= \alpha \end{aligned} \tag{9}$$

The result guarantees that $\hat{\alpha}$ is an unbiased estimator of the proper proportion α and is correctable when the estimate N of some of the n blocks is as follows:

$$N = \hat{\alpha} \times n = \frac{p-1}{2p-1}n + \frac{k}{2p-1} \tag{10}$$

Therefore, the estimation results satisfy the definition of local differential privacy, and the privacy preserving budget is set as follows:

$$\varepsilon = \ln \frac{p}{1-p} \tag{11}$$

However, such a randomized response technique does not satisfy our need for accurate and pseudo-location outputs. In other words, we perturb the mechanism to control the output of any location in the location candidate set to satisfy localized differential privacy. Therefore, we use a randomized response technique that can directly randomize the response to a secure location set L_k containing $k(k \geq 2)$ candidate values. Suppose RM is our randomized perturbation mechanism for any output $loc_i, loc_i^* \in L_m$.

For any output $l^* \in L$, the output of its response $\hat{l^*} \in L$ is generated with the following equation:

$$\mathrm{P}\left(\hat{l}^* \middle| l^*\right) = \frac{1}{k-1+\mathrm{e}^{\varepsilon}} \times \begin{cases} \mathrm{e}^{\varepsilon} & \hat{l}^* = l^* \\ 1 & \hat{l}^* \neq l^* \end{cases} \tag{12}$$

That is, responding to any of the remaining k -1 answers with probability $\frac{1}{k-1+\mathrm{e}^{\varepsilon}}$ and responding to the actual answer with probability $\frac{\mathrm{e}^{\varepsilon}}{k-1+\mathrm{e}^{\varepsilon}}$ results in satisfying the ε-LDP.

A multivariate time series is composed of multiple exogenous and target sequences. Given n exogenous sequences, $X = \left(x^1, x^2, \ldots, x^n\right) = (x_1, x_2, \ldots, x_{\mathrm{T}}) \in \mathbb{R}^{n \times T}$, where T denotes the time search window size, the exogenous sequence of a time window is constructed as a tree graph structure, where T is the number of nodes and the exogenous sequence $x_t = \left(x_t^1, x_t^2, \ldots, x_t^n\right) \in \mathbb{R}^n$ of timestamp t is used as a feature of node t.

$X = (x_{1,}x_2, \ldots, x_{t-1})$ denotes the node characteristics (i.e., timestamp history information) of the neighboring timestamps of node t, given the history values of the target sequence $(y_1, y_2, \ldots, y_{\mathrm{T}})$ where $y_t \in \mathbb{R}$, and the history values of the n exogenous sequences $(x_1, x_2, \ldots, x_{\mathrm{T}}) \in \mathbb{R}^{n \times \mathrm{T}}$. The purpose of this predictive model is to learn from the graph structure to discover hidden features and predict future values $\widetilde{y}_{T \to h}$:

$$\widetilde{y}_{T \to h} = F(y_1, y_2, \cdots y_{\mathrm{T}}, \mathrm{X}_1, \mathrm{X}_2, \cdots \mathrm{X}_{\mathrm{T}}) \tag{13}$$

where $\widetilde{y}_{T \to h}$ denotes the predicted value across h timestamps, and when h = 1, the model is used to make the next prediction for subsequent location prediction based on historical data. $F(\cdot)$ is a nonlinear mapping function.

4. Trajectory Privacy Protection and Prediction Mechanisms

The proposed trajectory privacy protection system TPPM-MP consists of two parts, the user side, and the server side, as shown in Figure 2. The user side collects the user's trajectory data, uses the trajectory processing mechanism based on localized differential privacy, and then publishes the processed trajectory data to the server side, which achieves the basic usability of the data published by the user without disclosing the user's privacy information. The significance of using the localized privacy protection mechanism is that the trustworthiness of the third-party service provider can be disregarded.

The user side part is divided into trajectory prediction and privacy protection mechanisms. The trajectory prediction part mainly introduces the prediction function and the objective function. Considering the multi-dimensional spatial–temporal correlation, the one-dimensional spatial–temporal feature D_t and the two $-$ dimensional spatio $-$ temporal feature H_t are simply aggregated as follows:

$$C_{t'} = [D_{t'} : H_{t'}] \tag{14}$$

The final prediction of the trajectory prediction mechanism is made via a nonlinear mapping of the aggregated features:

$$\begin{aligned} \left(\hat{y}_{t',T+1}, \hat{y}_{t',T+2}, \cdots, \hat{y}_{t',T+h}\right) &= \mathcal{F}(\hat{y}_{t'}, X_{t'}) \\ &= v_y^T\left(C_{t'}W_y + b_\omega\right) + b_v \end{aligned} \tag{15}$$

where the parameter matrices $W_y \in \mathbb{R}^{p+q}$ and $b_w \in \mathbb{R}^T$ map the new features to dimension T. The final prediction is generated using the linear transformations $v_y \in \mathbb{R}^{h\times T}$ and $b_v \in \mathbb{R}^h$. The final prediction results use a modified hybrid strategy with the same model, which accepts multiple inputs and predicts multiple outputs. This strategy balances the drawbacks of direct, iterative, and MIMO strategies, avoids conditional independence assumptions, and allows for time dependence between the output data, while the simultaneous outputs of multiple models allow some flexibility. Figure 3 illustrates the architecture of the improved hybrid strategy proposed in this chapter, where F denotes the trajectory prediction model proposed.

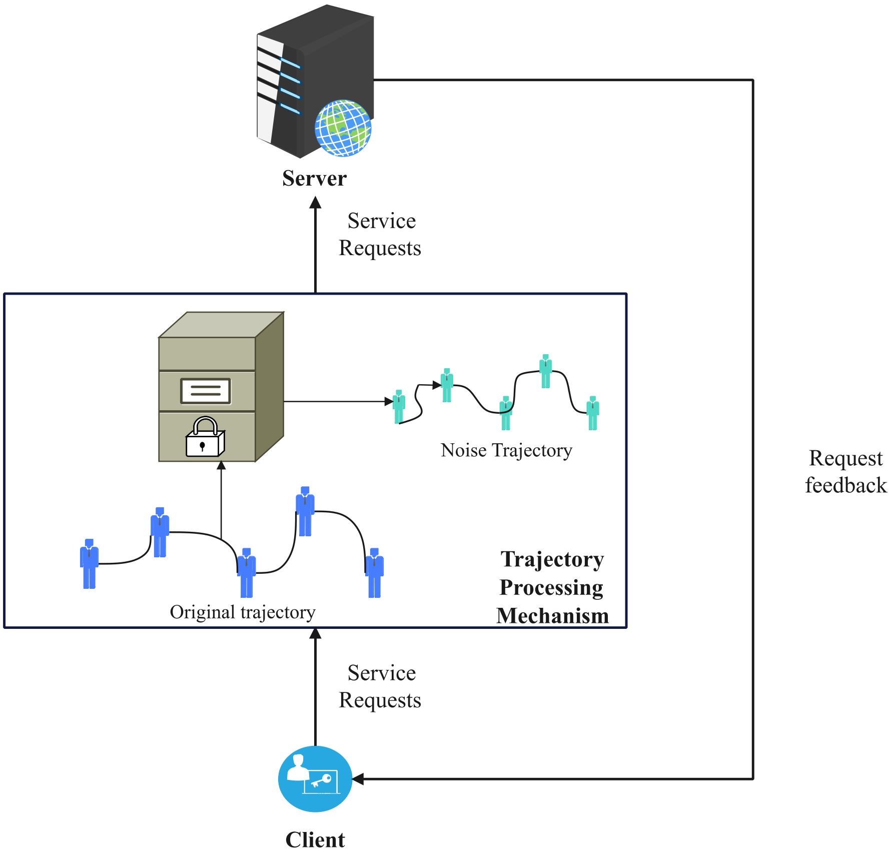


Figure 2. TPPM-MP system architecture diagram.

All parameters of the method proposed in this paper are learned using MSE as an objective function, which can be formulated as follows:

$$\mathcal{L}(\theta) = \frac{1}{N\times h}\sum_{i=1}^{N}\sum_{t=1}^{h}\left(y_t^i - \hat{y}_t^i\right)^2 \tag{16}$$

where θ denotes the learnable parameter, N denotes the number of training samples, h denotes the length of the predicted time step, and y_t^i and $\hat{y}_t^i$ denote the true and predicted values of time step t, respectively. Finally, Adam is used to optimize the objective function.

Subsequently, the corresponding privacy budget is allocated based on the predicted values of the historical data. Assuming that each mobile user has p locations to be protected, the importance of location $c_{u_i}^h (h \in [1, p])$ is $\omega_h(c_{u_i}^h)$, and the sensitivity of the location point is $\Delta\omega_h$, from which is derived the probability that the location will be selected:

$$\Pr(c_{u_i}^h) = \frac{exp\left(\frac{\varepsilon}{2\Delta\omega_h} * \omega_h(c_{u_i}^h)\right)}{\sum_{h=1}^{p} exp\left(\frac{\varepsilon}{2\Delta\omega_h} * \omega_h(c_{u_i}^h)\right)} \tag{17}$$

Given the privacy budget, our mechanism can allocate the budget according to the probability of each hotspot being selected by the utility function, $\Pr(c_{u_i}^h)$, and the budget ε_h allocation for each hotspot can be calculated by the following formula:

$$\varepsilon_h = \varepsilon * \left(1 - \frac{\Pr(c_{u_i}^h)}{\sum_{h=1}^{p} \Pr(c_{u_i}^h)}\right) \tag{18}$$

The pseudo-location candidate set of predicted location points can be generated by assigning the corresponding privacy budget value, which is denoted as $Tr_{sp} = \{(lon_j, lat_j, t_j, loc_j), (lon_{j+1}, lat_{j+1}, t_{j+1}, loc_{j+1}), \cdots, (lon_k, lat_k, t_k, loc_k)\}$, where $1 \leq j < k \leq n$, and the generalized location residency is then generated based on the set of candidate locations in the location region. The generalized location residency given by the following equation:

$$lon^* = \frac{lon_j + lon_{j+1} + \cdots + lon_k}{k - j + 1} + lonNoise \tag{19}$$

$$lat^* = \frac{lat_j + lat_{j+1} + \cdots + lat_k}{k - j + 1} + latNoise \tag{20}$$

where lon^* and lat^* represent the precision and latitude coordinates of the noise location data available for publication. The *lonNoise* and *latNoise* are noise that satisfies the Laplace distribution.

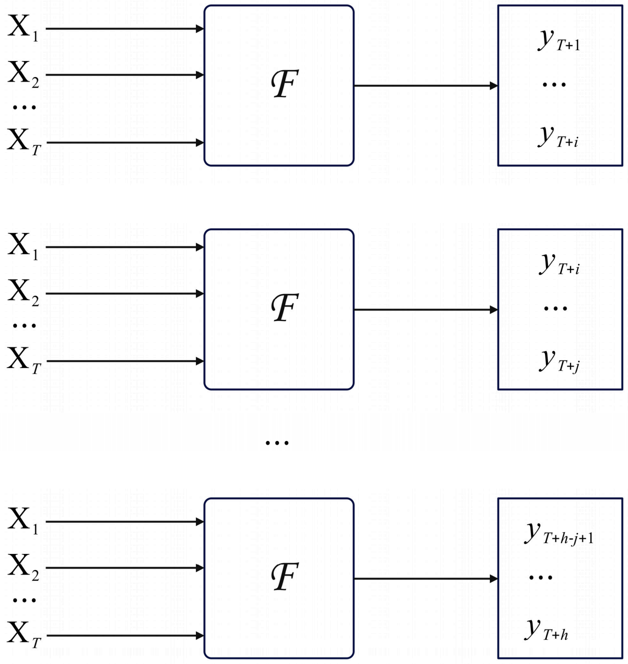


Figure 3. Structure of trajectory prediction mechanism.

5. Experimental Evaluation

In this section, we report the evaluative testing of our experimental methodology; all algorithmic experiments were implemented on Python. Our datasets were derived from typical datasets published on the web, Geolife [23–26] and Gowalla [27]. The Geolife dataset is derived from the GPS track data of 182 users, collected by Microsoft Research Asia. The data points are in chronological order, each containing longitude, latitude, and altitude information. There are 17,621 trajectories with a total distance of more than 1.2 million km and a total duration of more than 50,000 h. The data records the location of users' homes and workplaces and tracks a wide range of outdoor activities such as shopping, traveling, touring, biking, etc. The Gowalla dataset, collected by Stanford University, is a location-based social networking site that allows users to share information about their location by checking in. The dataset includes 6,442,890 check-in locations and 19,651 users' check-in location information.

In order to verify the effectiveness of the proposed location prediction model in the multi-step prediction task, a statistical model and a deep learning model with excellent performance were chosen as the comparison methods in this experiment. The comparison methods are described as follows:

ARIMA: (Autoregressive Integrated Moving Average Model) [28] is a typical univariate time series forecasting statistical model. It involves a combination of difference operation and ARMA (Auto-Regressive and Moving Average Model), firstly converting the non-smooth time series into smooth data by difference operation and then using ARMA to fit the differenced series.

LSTM: (long short-term memory) [29] is a widely used RNN (recurrent neural network) variant designed to mine hidden long-term temporal dependencies in time series.

DA-RNN: Data Associated Recurrent Neural Network model uses a two-stage attention mechanism with an input attention mechanism and a temporal attention mechanism. First, the input attention mechanism adaptively selects relevant exogenous sequences. In the second stage, the temporal attention mechanism automatically selects the relevant encoder hidden states for all time steps. The DA-RNN [30] can be utilized to predict the value of the next moment efficiently.

Three different evaluation metrics were used to assess the performance of the forecasting models [31]. Two evaluation metrics, mean absolute error (MAE) and root mean squared error (RMSE), are widely used in time series forecasting to measure the error between predicted and observed values. Smaller values of MAE [32] and RMSE [33] indicate the model's lower prediction error and more accurate prediction. In addition, the coefficient of determination (R squared, R^2) was also utilized to determine the fitting effect of the model. The range of values of R^2 [34] was determined as [0,1], and a value of R^2 closer to 1 indicates that the model is fitted better. Assuming that y_t is the true value of time step t, $\hat{y}_t$ is the predicted value of time step t, $\overline{y}$ is the average of the true value, and N is the number of samples, the evaluation index is defined as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum\nolimits_{i=1}^{N}\left(y_t^i - \hat{y}_t^i\right)^2} \tag{21}$$

$$\mathrm{MAE} = \frac{1}{N}\sum\nolimits_{i=1}^{N}\left|y_t^i - \hat{y}_t^i\right| \tag{22}$$

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{N}\left|y_t^i - \hat{y}_t^i\right|}{\sum_{i=1}^{N}\left|y_t^i - \overline{y}_t^i\right|} \tag{23}$$

The multi-step time series prediction results of the TPPM-MP and the comparison methods using two real datasets are given below. For a fair comparison, only the best evaluation results of each method with different parameter settings are shown. Table 1 shows the evaluation results for single-step time series prediction. To ensure clarity of results and facilitate observation, we used $1 - \mathrm{R}^2$ as the metric. From the table, it can be observed that the TPPM-MP model proposed in this paper achieved optimal performance on all datasets. The results of single-step prediction and multi-step prediction are analyzed below.

Table 1. Single-step time series prediction results on Geolife dataset and Gowalla dataset.

Methodologies	Geolife			Gowalla		
	MAE	**RMSE**	$\mathbf{1-R^2}$	**MAE**	**RMSE**	$\mathbf{1-R^2}$
ARIMA	0.3890	0.7355	0.0812	0.3412	0.2348	0.0190
LSTM	0.3760	0.7346	0.0812	0.2780	0.2336	0.0174
DA-RNN	0.3550	0.7155	0.0735	0.1302	0.0967	0.0059
TPPM-MP	**0.2806**	**0.6251**	**0.0526**	**0.0616**	**0.0454**	**0.0019**

For single-step prediction, each method involves predicting the value of the next time step (h = 1). It can be observed from Table 1 that the MAE and RMSE values of the ARIMA model were both higher than the other compared methods or TPPM-MP. On the Geolife and Gowalla datasets, the MAE ratio of ARIMA was higher than that of TPPM-MP in both cases. This indicates that ignoring exogenous factors reduced the model's performance. Although LSTM performed better than ARIMA, TPPM-MP had lower MAE values than LSTM on each dataset. This was because the LSTM network focused on extracting the long-term dependencies of all time series rather than selecting relevant features. The above experimental results suggest that TPPM-MP's use of the attention mechanism to capture spatial–temporal correlations helps it achieve better prediction performance.

The following section describes the privacy performance and usability analysis of the method proposed in this paper for publishing trajectory data. We also consider different prediction scenarios and privacy-preserving budgets, and compare the algorithm in this paper with other privacy-preserving methods, DP-Srat [35] and N-gram [36]. We evaluate our privacy-preserving mechanism using four metrics, i.e., relative error, accuracy P, recall R, and $F_{\text{–value}}$, as follows:

$$P = \frac{\left|Q\left(Traj'_{u_i}\right) \cap Q\left(Traj_{u_i}\right)\right|}{\left|Q\left(Traj'_{u_i}\right)\right|} \tag{24}$$

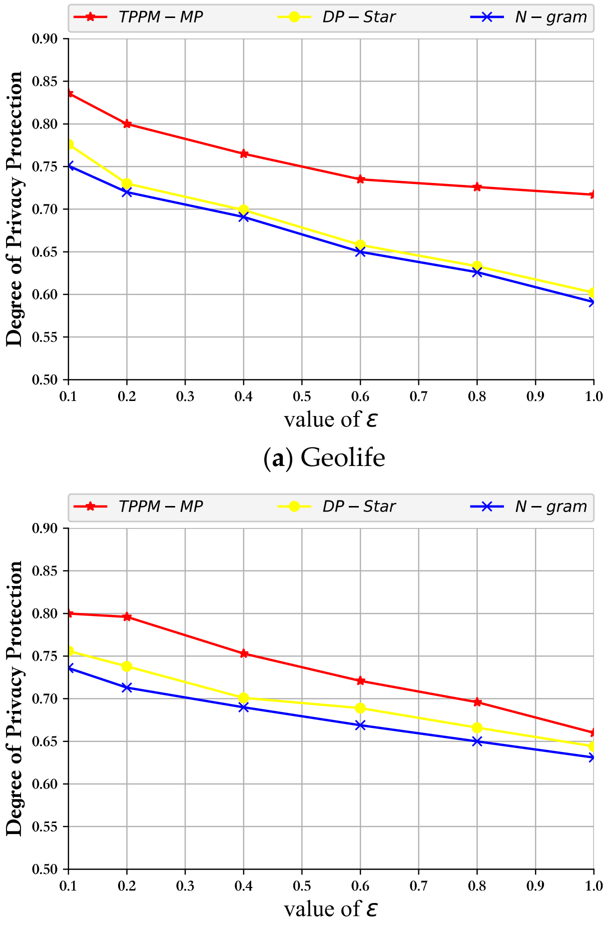
$$R = \frac{\left|Q\left(Traj'_{u_i}\right) \cap Q\left(Traj_{u_i}\right)\right|}{\left|Q\left(Traj_{u_i}\right)\right|} \tag{25}$$

$$F_{-value} = \frac{(\mu+1)P*R}{\mu*P+R} \tag{26}$$

where μ is the tuning parameter. In this paper, we set P and R to be equally important, so μ = 1.

In this example, the degree of privacy protection is regulated and controlled according to the privacy budget ε of DP, and the distance between the actual location and the pseudo location is obtained from the probability as e^{ε}. To facilitate the measurement of the degree of privacy protection, the result is restricted to [0,1], so the formula for the privacy protection degree (***PPD***) is $PPD = e^{-\varepsilon}$. The results obtained using the comparative experimental real-world datasets Geolife and Gowalla are shown in Figure 4.

As shown in Figures 5–7, the experimental results of TPPM-MP, N-gram, and DP-Star under different privacy budgets were assessed. Firstly, analyzing the curve shapes, that obtained via our mechanism was similar to those of the previously proposed mechanisms. Throughout the experimental results, the values of the three metric mechanisms, precision, recall, and F-value, all increased with epsilon. This is mainly because our privacy-preserving mechanism is less tolerant of noise; therefore, when epsilon increased, it added a small amount of noise. This was also the reason why the data availability could be better realized. In addition, our approach had more outstanding performance metrics than the other two schemes, mainly because our privacy budget allocation mechanism and noise control output mechanism played a good role in ensuring that the published pseudo-trajectory data was highly similar to the actual trajectory data.

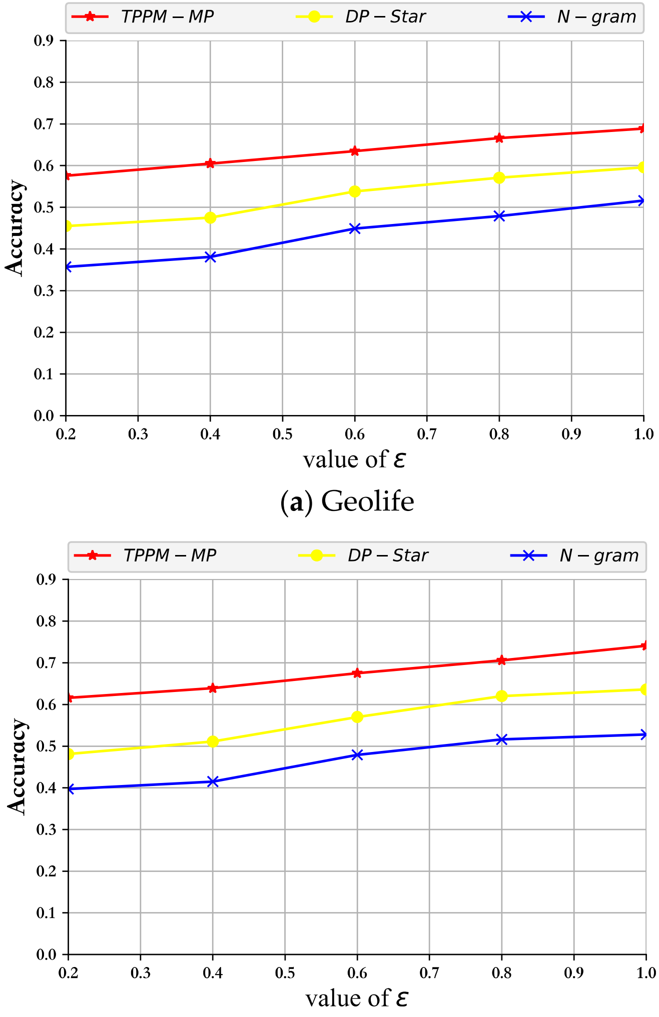


(**a**) Geolife

(**b**) Gowalla

Figure 4. The effect of ε vs. PPD.

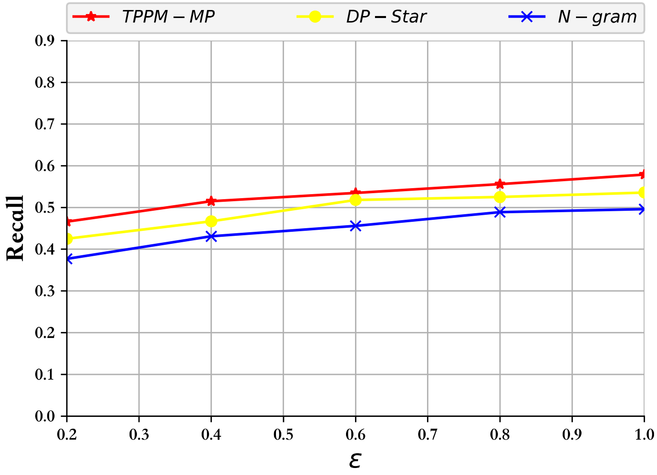
The N-gram method first extracts the sensitive information in the trajectory data using a variable-length n-gram model. Then, it adds noise adaptively for each data in the trajectory sequence. This scheme adds too much noise data compared to our DPTP-LICD scheme, resulting in a significant difference between the generalized output trajectory data and the original trajectory. DP-Star also provides excellent control over the privacy budget compared with our mechanism but applies the minimal description principle to generalize the original trajectory sequence into a series of points that are representative of the trajectory. While this saves storage space and reduces budget allocation, it also guarantees the accuracy of some queries. However, this approach also makes the trajectory data more different from the original trajectory. In order to reflect the fairness and impartiality of the performance comparison experiment, we did not use the best p-value for the comparison experiment but used the average relative error to make the comparison, reconciling good p-values and bad p-values, so that the comparison experiment was closer to the fact and more objective. Especially as seen in Figure 8, we were able to achieve a relatively low average relative error with increasing privacy budget via DPTP-LICD compared with N-gram and DP-Star.



(**a**) Geolife

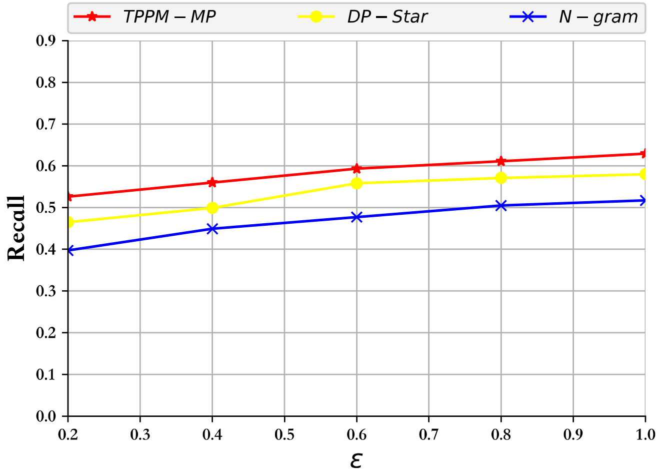
(**b**) Gowalla

Figure 5. Variation of accuracy with privacy budget.



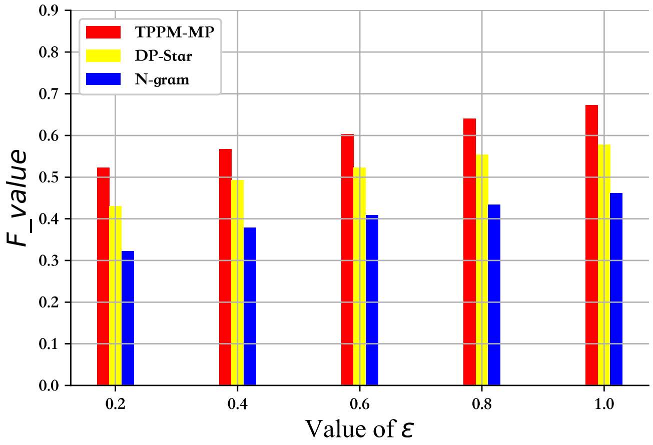
(**a**) Geolife

Figure 6. *Cont*.

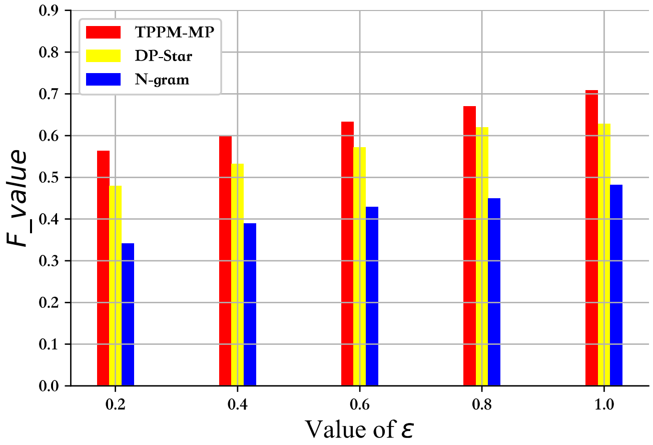


(**b**) Gowalla

Figure 6. Recall rate variation with privacy budget.

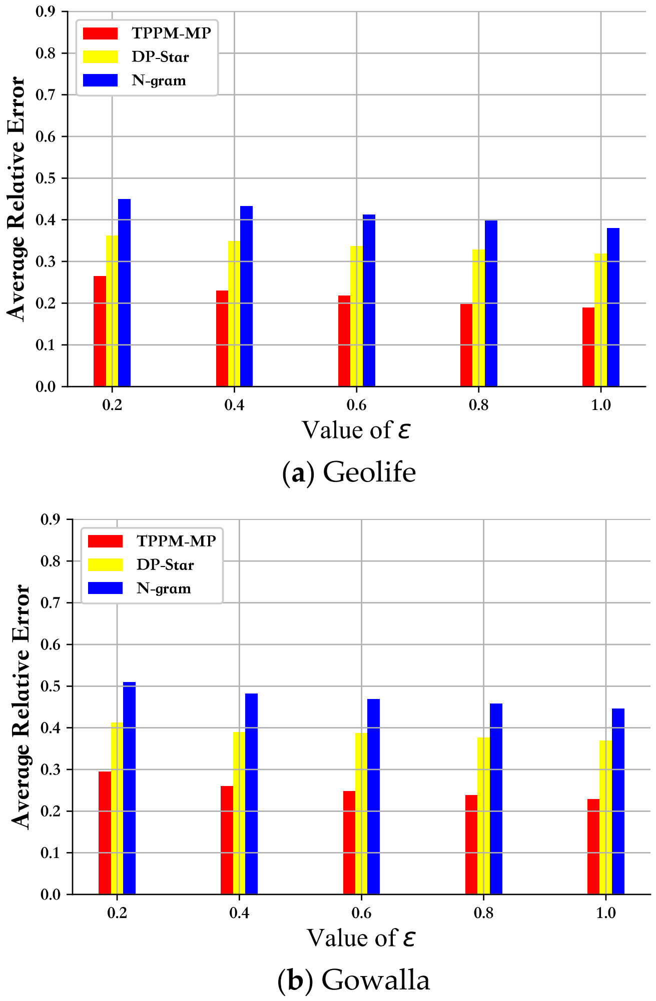


(**a**) Geolife



(**b**) Gowalla

Figure 7. F-value vs. privacy budget.



(**a**) Geolife

(**b**) Gowalla

Figure 8. Mean relative error vs. privacy budget.

6. Conclusions and Outlook

In this paper, we present an RNN network based on multidimensional spatial–temporal attention, which considers local spatiotemporal correlation and global spatial–temporal correlation from different spatiotemporal dimensions, fully exploiting the dependencies in the multivariate time series and performing trajectory data prediction. The model is a specialized network architecture that integrates a LSTM network based on correlation attention with an attention mechanism, where the former is designed to extract important node features and transform them into higher-level features, thereby endowing the node features with sufficient expressive power, while the latter is employed to compute the correlation strength between nodes. The privacy budget is allocated according to the importance of the predicted location in the trajectory, and noise based on the localized differential privacy technique is added to enhance the usability of the released data under the premise of guaranteeing the privacy of the user′s trajectory. The proposed method described in this paper enhanced usability by 22% and 37% on the same dataset compared with the other methods tested, while providing equivalent privacy protection. The prediction model proposed in this paper can be applied not only in the field of trajectory prediction but also in the field of cross-domain data prediction, to visualize development trends, which is of practical significance for optimizing the future actions of decision makers. In the future, we will work on discovering more feasible and effective privacy-preserving solutions for trajectories, and our future work will shift to practical applications, as we strive to create cutting-edge technologies that are more optimized and reusable.

Author Contributions: Conceptualization, J.X. and W.Z.; methodology, J.X. and M.S.; coding, Z.X. and M.S.; validation, Y.L.; investigation, M.S; writing-original draft preparation, J.X. and W.Z.; writing—review and editing, J.X. and W.Z.; visualization, Z.X.; supervision, J.X.; project administration, J.X.; funding acquisition, J.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (grant No. 62001215), Science and Technology Plan Project of Changzhou (CJ20210155, CJ20220151, CJ20220174), Natural Science Foundation of the Jiangsu Higher Education Institutions of China (23KJA520001, 21KJD520002).

Data Availability Statement: The original contributions presented in the study are included in the article; further inquiries can be directed to the corresponding author.

Acknowledgments: Thanks to the anonymous reviewers and editors from the journal.

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

References

1. Wang, S.; Gong, M.; Wu, Y.; Zhang, M. Multi-objective optimization for location-based and preferences-aware recommendation. *Inf. Sci.* **2020**, *513*, 614–626.
2. Jang, W.; Kim, S.; Chun, J.W.; Jung, A.R.; Kim, H. Role of recommendation sizes and travel involvement in evaluating travel destination recommendation services: Comparison between artificial intelligence and travel experts. *J. Hosp. Tour. Technol.* **2023**, *14*, 401–415.
3. Abul, O.; Bonchi, F.; Nanni, M. Never Walk Alone: Uncertainty for Anonymity in Moving Objects Databases. In Proceedings of the 2008 IEEE 24th International Conference on Data Engineering, Cancun, Mexico, 7–12 April 2008; pp. 376–385. [CrossRef]
4. Noman, M.; Benjamin, F.; Mourad, D. Walking in the crowd: Anonymizing trajectory data for pattern analysis. In Proceedings of the International Conference on Information and Knowledge Management, Proceedings, Hong Kong, China, 2–6 November 2009; pp. 1441–1444. [CrossRef]
5. Papadopoulos, S.; Bakias, S.; Papadias, D. Nearest Neighbor Search with Strong Location Privacy. *Proc. VLDB Endow.* **2010**, *3*, 619–629. [CrossRef]
6. Radomirović, J.; Milosavljević, M.; Kovačević, B.; Jovanović, M. Privacy Amplification Strategies in Sequential Secret Key Distillation Protocols Based on Machine Learning. *Symmetry* **2022**, *14*, 2028. [CrossRef]
7. Dwork, C. Differential privacy in new settings. In Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '10, Austin, TX, USA, 17–19 January 2010; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2010; pp. 174–183.
8. Thanh, C.P.; Hung, C.T. Consideration of Data Security and Privacy Using Machine Learning Techniques. *Int. J. Data Inform. Intell. Comput.* **2023**, *2*, 20–32.
9. Rammohan, S.R.; Jayanthiladevi, A. AI Enabled Crypto Mining for Electric Vehicle Systems. *Int. J. Data Inform. Intell. Comput.* **2023**, *2*, 33–39.
10. Chen, R.; Fung, B.C.; Mohammed, N.; Desai, B.C.; Wang, K. Privacy-preserving trajectory data publishing by local suppression. *Inf. Sci.* **2013**, *231*, 83–97. [CrossRef]
11. Ho, S.-S.; Ruan, S. Differential privacy for location pattern mining. In Proceedings of the 4th ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS, SPRINGL '11, Chicago, IL, USA, 1 November 2011; Association for Computing Machinery: New York, NY, USA, 2011; pp. 17–24.
12. Xiao, Y.; Xiong, L. Protecting locations with differential privacy under temporal correlations. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, CO, USA, 12–16 October 2015; pp. 1298–1309.
13. Ou, L.; Qin, Z.; Liao, S.; Hong, Y.; Jia, X. Releasing Correlated Trajectories: Towards High Utility and Optimal Differential Privacy. *IEEE Trans. Dependable Secur. Comput.* **2018**, *17*, 1109–1123. [CrossRef]
14. Cao, S.; Wu, L.; Wu, J.; Wu, D.; Li, Q. A spatio-temporal sequence-to-sequence network for traffic flow prediction. *Inf. Sci.* **2022**, *610*, 185–203.
15. Liu, Y. DSTP-RNN: A dual-stage two-phase attention-based recurrent neural networks for long-term and multivariate time series prediction. *Expert Syst. Appl.* **2020**, *143*, 113082.
16. Wang, J.; Chen, Q.; Gong, H. STMAG: A spatial-temporal mixed attention graph-based convolution model for multi-data flow safety prediction. *Inf. Sci.* **2020**, *525*, 16–36. [CrossRef]
17. Song, C.; Lin, Y.; Guo, S.; Wan, H. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 914–921.
18. Wang, B.; Lin, Y.; Guo, S.; Wan, H. GSNet: Learning Spatial-Temporal Correlations from Geographical and Semantic Aspects for Traffic Accident Risk Forecasting. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 4402–4409.
19. Beukers, F.; Vlasenko, M. Dwork Crystals II. *Int. Math. Res. Not.* **2021**, *6*, 4427–4444.
20. Zhao, Y.; Du, J.T.; Chen, J. Scenario-based Adaptations of Differential Privacy: A Technical Survey. *ACM Comput. Surv.* **2024**, *56*, 1–39.

21. Li, X.; Yan, H.; Cheng, Z.; Sun, W.; Li, H. Protecting Regression Models With Personalized Local Differential Privacy. *IEEE Trans. Dependable Secur. Comput.* **2023**, *20*, 960–974.
22. Jiang, H.; Pei, J.; Yu, D.; Yu, J.; Gong, B.; Cheng, X. Applications of Differential Privacy in Social Network Analysis: A Survey. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 108–127.
23. Zheng, Y.; Li, Q.; Chen, Y.; Xie, X.; Ma, W.-Y. Understanding mobility based on gps data. In Proceedings of the 10th International Conference on Ubiquitous Computing, UbiComp '08, Seoul, Republic of Korea, 21–24 September 2008; ACM: New York, NY, USA, 2008; pp. 312–321.
24. Zheng, Y.; Zhang, L.; Xie, X.; Ma, W.Y. Mining interesting locations and travel sequences from gps trajectories. In Proceedings of the 18th International Conference on World Wide Web, WWW '09, Madrid, Spain, 20–24 April 2009; ACM: New York, NY, USA, 2009; pp. 791–800.
25. Zheng, Y.; Xie, X.; Ma, W.-Y. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.* **2010**, *33*, 32–39.
26. Geolife Dataset. Available online: https://www.microsoft.com/en-us/download/details.aspx?id=52367 (accessed on 28 July 2024).
27. Gowalla Dataset. Available online: http://snap.stanford.edu/data/loc-gowalla.html (accessed on 28 July 2024).
28. Ray, S.; Lama, A.; Mishra, P.; Biswas, T.; Das, S.S.; Gurung, B. An ARIMA-LSTM model for predicting volatile agricultural price series with random forest technique. *Appl. Soft Comput.* **2023**, *149*, 110939.
29. Ordóñez, F.J.; Roggen, D. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors* **2016**, *16*, 115. [CrossRef]
30. Penny, S.G.; Smith, T.A.; Chen, T.; Platt, J.A.; Lin, H.; Goodliff, M.; Abarbanel, H.D.I. Integrating Recurrent Neural Networks With Data Assimilation for Scalable Data-Driven State Estimation. *J. Adv. Model. Earth Syst.* **2022**, *14*, e2021MS002843.
31. Bono, F.M.; Radicioni, L.; Cinquemani, S.; Conese, C.; Tarabini, M. Development of soft sensors based on neural networks for detection of anomaly working condition in automated machinery. In Proceedings of the Predictive Maintenance, and Communication and Energy Systems in a Globally Networked World, Long Beach, CA, USA, 6 March–11 April 2022; Volume 12049, pp. 56–70.
32. Chai, T.; Draxler, R.R. Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.* **2014**, *7*, 1247–1250.
33. Willmott, C.J.; Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim. Res.* **2005**, *30*, 79–82.
34. Kieu, T.; Luu, P.; Yoon, N. Multiple linear regression: Identify potential health care stocks for investments using out-of-sample predictions. *Teach. Stat.* **2020**, *42*, 98–107.
35. Holohan, N.; Leith, D.J.; Mason, O. Optimal Differentially Private Mechanisms for Randomised Response. *IEEE Trans. Inf. Forensics Secur.* **2017**, *12*, 2726–2735.
36. Rui, C.; Acs, G.; Castelluccia, C. Differentially Private Sequential Data Publication via Variable-Length N-Grams. In Proceedings of the ACM Conference on Computer & Communications Security, Raleigh, NC, USA, 16–18 October 2012; ACM: New York, NY, USA, 2012.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.