# 基于差分隐私的轨迹隐私保护方法

袁水莲,皮德常,胥 萌
(南京航空航天大学计算机科学与技术学院, 江苏南京 211106)

**摘 要:** 针对现有的轨迹隐私保护模型大多难以抵御复杂背景知识攻击的问题,本文提出了一种基于差分隐私的轨迹隐私保护方法.首先结合地理不可区分机制对原始轨迹数据添加半径受限的拉普拉斯噪音;其次构造数据映射模型将原始数据和噪音数据映射到新的发布位置,使攻击者无法获取真实轨迹数据;接着应用最优数据映射函数发布最优的轨迹位置以提高发布数据的可用性;最后利用差分隐私抵御非敏感信息推理攻击,进一步保护用户隐私.实验结果表明,本文算法既能有效保护轨迹数据中用户的隐私,也能保证数据的可用性.

**关键词:** 轨迹数据; 隐私保护; 差分隐私; 地理不可区分; 背景知识攻击; 推理攻击
**中图分类号:** TP309.2 **文献标识码:** A **文章编号:** 0372-2112 (2021) 07-1266-08
**电子学报URL:** http://www.ejournal.org.cn **DOI:** 10.12263/DZXB.20200827

# Trajectory Privacy Protection Method Based on Differential Privacy

YUAN Shui-lian, PI De-chang, XU Meng
(*College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu* 211106, *China*)

**Abstract:** Aiming at the problem that most of the existing trajectory privacy protection models are difficult to withstand complex background knowledge attacks, this paper proposes a trajectory privacy protection method based on differential privacy. Firstly, the Laplacian noise with limited radius is added to the original trajectory data by combining the mechanism of geographic indistinguishability. Secondly, a data mapping model is constructed to map the original data and noise data to the new publishing location, so that the attacker cannot obtain the real trajectory data. Then the optimal data mapping function is applied to publish the optimal trajectory position to improve the availability of published data. Finally, differential privacy is used to defend against non-sensitive information inference attack to further protect user privacy. The experimental results show that the algorithm in this paper can not only effectively protect the privacy of users in the trajectory data, but also ensure the availability of the data.

**Key words:** trajectory data; privacy protection; differential privacy; geographical indistinguishability; background knowledge attack; inference attack

## 1 引言

随着具有定位功能的移动设备的普及,轨迹数据的收集分析越来越便捷,随之而来的隐私泄露问题日益突出.差分隐私模型具有坚实的数学基础,可以动态调整、定量分析隐私保护水平,即使攻击者拥有丰富的背景知识也能有效保护用户隐私,因此成为隐私保护领域的研究热点,但将其合理应用到轨迹隐私保护中极具挑战.轨迹数据保护必须充分考虑轨迹的时空特性和实际应用场景,对于非位置敏感信息也应给出相应的保护策略.

针对以上问题,本文提出一种基于差分隐私的轨迹隐私保护方法.该方法通过对轨迹数据添加噪音抵御攻击,通过数据映射模型和最优数据映射确定发布位置,通过差分隐私对非位置敏感信息添加噪音提高保护效果.

## 2 相关工作

近年来,隐私保护备受关注,取得了丰富的研究成果.杨等[1]提出使用不变后随机响应方法对数据集进行扰动,使之满足局部差分隐私.叶等[2]引入位置服务信息的邻近共享机制,降低用户对服务器的依赖.陈

收稿日期:2020-08-03;修回日期:2020-11-27;责任编辑:覃怀银

等[3]通过抑制位置点或添加假轨迹,保证处理轨迹数据集时达到最大收益.

随着研究的深入,差分隐私逐渐被应用于轨迹隐私保护. Wang 等[4]提出 P-STM 隐私保护方法,首先处理差分隐私下轨迹的异质性,然后重写轨迹提高相似性评估效用. Deldar 等[5]提出 PDP-SAG 保护机制,通过对敏感属性噪音值进行一致性约束,使噪音频率保持一致. Feldman 等[6]提出轨迹保护算法 PKM,允许从 GPS 数据库进行聚类,并允许用户控制引入的噪声.

## 3　问题定义

**定义 1(位置点)**　轨迹位置点由 $l=(x, y, t)$ 表示,其中,$x$ 和 $y$ 为位置点的经度和纬度,$t$ 为经过位置点的时间.

**定义 2(轨迹)**　轨迹 $t$ 是由一系列位置点按照时间顺序形成的序列,可表示为:

$$t = l_1 \rightarrow l_2 \rightarrow \cdots \rightarrow l_{|n|} \tag{1}$$

其中,$|n|$ 表示位置点数目.

**定义 3(轨迹数据集)**　轨迹数据集 $T$ 是由一系列轨迹序列组成的集合,可表示为:

$$T = \{t_1, t_2, \cdots, t_{|T|}\} \tag{2}$$

其中,$|T|$ 表示轨迹数目.

**定义 4(非敏感信息推理攻击)**　推理攻击[7]指攻击者从非敏感信息中推断出敏感信息的方法. 假设 $\gamma = \{\inf_1, \inf_2, \cdots, \inf_m\}$ 为轨迹中的非敏感信息,且 $\inf_i$ 的取值为 0 或 1,对前 $k$ 行属性求和得 $\mathrm{sum}_k = \{\mathrm{sum}_{k1}, \mathrm{sum}_{k2}, \cdots, \mathrm{sum}_{km}\}$,通过求解 $\mathrm{sum}_k$ 和 $\mathrm{sum}_{k-1}$ 的差值可得第 $k$ 条轨迹对应的非敏感信息,然后查询轨迹数据库即可推理出用户轨迹.

**定义 5($\varepsilon$-差分隐私[8])**　假设数据集 $T$ 经随机算法 $M$ 处理后的输出结果集合为 $Y$,$Y$ 的任意子集为 $D$,对于任意邻近数据集 $T$ 和 $T'$,若算法 $M$ 满足不等式:

$$\frac{\Pr(M(T)=D)}{\Pr(M(T')=D)} \leqslant \mathrm{e}^{\varepsilon} \tag{3}$$

则称算法 $M$ 提供 $\varepsilon$-差分隐私保护.

**定义 6(全局敏感度[9])**　对于任意邻近数据集 $T$ 和 $T'$,给定查询函数 $f: T \rightarrow R$,定义敏感度为:

$$\Delta f = \max_{T,T'} \|f(T) - f(T')\|_1 \tag{4}$$

其中,$\|\cdot\|_1$ 为曼哈顿距离.

**定义 7(拉普拉斯机制[8])**　若随机算法 $M$ 满足 $\varepsilon$-差分隐私,则其输出结果为:

$$M(T) = f(T) + Y \tag{5}$$

其中,随机噪音 $Y$~Lap($\Delta f/\varepsilon$).

本文引入地理不可区分机制,只考虑距离不超过 $d$ 的有限区域内的位置点隐私.

**定义 9(地理不可区分性[10])**　随机算法 $M$ 满足地理不可区分性当且仅当 $M$ 满足:直径为 $d$ 的区域圆 $C$ 内的任意两点 $X_1$, $X_2$ 和任意输出结果集合 $Y$,集合 $Y$ 的任意子集为 $D$,有:

$$\frac{\Pr(M(X_1)=D)}{\Pr(M(X_2)=D)} \leqslant \mathrm{e}^{\varepsilon d} \tag{6}$$

## 4　基于差分隐私的轨迹隐私保护方法

### 4.1　截取有限区域

为了构造合理的后验概率分布,需要将噪音位置限制在有限区域,以减少因位置数据集过大造成的误差,本文限制区域圆的最大半径 $r_{\max}$ 为:

$$r_{\max} \leqslant \left|(x,y)_{\max} - (x,y)_{\min}\right| \leqslant 2r \tag{7}$$

本文通过聚类方式创建区域圆,并重新定义全局敏感度. 假设相邻区域圆的圆心为 $(C_x, C_y)$ 和 $(C_{x'}, C_{y'})$,轨迹位置点为 $(x_i, y_i)$ 和 $(x_i', y_i')$,位置点个数为 $n$ 和 $n'$. 相邻区域圆的位置点个数仅相差 1,即 $|n - n'| = 1$,取 $n' = n - 1$. 那么二维空间的全局敏感度如下:

$$\begin{aligned}\Delta f &= |C - C'| = \left|(C_x, C_y) - (C_{x'}, C_{y'})\right| \\ &= \left|\left(\frac{\sum x_i}{n} - \frac{\sum x_i'}{n'}, \frac{\sum y_i}{n} - \frac{\sum y_i'}{n'}\right)\right| \\ &= \left|\left(\frac{\left(\sum x_i - \sum x_i'\right) - C_{x'}}{n}, \frac{\left(\sum y_i - \sum y_i'\right) - C_{y'}}{n}\right)\right|\end{aligned} \tag{8}$$

记 $\left(\sum x_i - \sum x_i', \sum y_i - \sum y_i'\right)$ 取最大值的点为:

$$\left(X_{\max}, Y_{\max}\right) = \begin{pmatrix}\max\left(\sum x_i - \sum x_i'\right), \\ \max\left(\sum y_i - \sum y_i'\right)\end{pmatrix} \tag{9}$$

由全局敏感度定义并结合式(8)和(9)可得:

$$\begin{aligned}\Delta f &= \left(\frac{X_{\max} - C_{x'}}{n}, \frac{Y_{\max} - C_{y'}}{n}\right) \\ &= \frac{\left|X_{\max} - C_{x'}\right| + \left|Y_{\max} - C_{y'}\right|}{n}\end{aligned} \tag{10}$$

结合差分隐私理论和地理不可区分性下的差分隐私结果,得出二维平面的概率密度函数为:

$$p_{\varepsilon}(l)(l') = \frac{(\varepsilon/\Delta f)^2}{2\pi} \cdot \mathrm{e}^{-\frac{\varepsilon}{\Delta f} d(l,l')} \tag{11}$$

其中,$l$ 和 $l'$ 分别表示真实位置点和噪声位置点,$d(l,l')$ 表示位置点 $l$ 和 $l'$ 的欧氏距离.

由于概率密度函数只与 $d(l,l')$ 有关,为简便计算,可将其转化为极坐标下的概率密度函数:

$$p_\varepsilon(r,\theta)=\frac{(\varepsilon/\Delta f)^2}{2\pi}re^{-\frac{\varepsilon}{\Delta f}r} \tag{12}$$

由式(12)求得半径$r$和角度$\theta$的边缘概率密度函数为:

$$p(r)=\int_0^{2\pi}p_\varepsilon(r,\theta)\mathrm{d}\theta=\left(\frac{\varepsilon}{\Delta f}\right)^2re^{-\frac{\varepsilon}{\Delta f}r} \tag{13}$$

$$p(\theta)=\int_0^{\infty}p_\varepsilon(r,\theta)\mathrm{d}r=\frac{1}{2\pi} \tag{14}$$

由于$p_\varepsilon(r,\theta)=p(r)p(\theta)$,所以$r$和$\theta$相互独立,因此可以从$p(r)$和$p(\theta)$中求解出$r$和$\theta$的值. 由式(14)知,$p(\theta)$为常量,所以$\theta$是$[0,2\pi)$服从均匀分布的随机数.

求解$r$的值,首先计算$r$的累积分布函数:

$$F_\varepsilon(r)=\int_0^r p(r)\mathrm{d}r=1-\left(1+\frac{\varepsilon}{\Delta f}r\right)e^{-\frac{\varepsilon}{\Delta f}r} \tag{15}$$

将式(10)的全局敏感度带入式(15)得:

$$F_\varepsilon(r)=1-\left(1+\frac{n\varepsilon}{\left|X_{\max}-C_{x'}\right|+\left|Y_{\max}-C_{y'}\right|}r\right)\cdot e^{-\frac{n\varepsilon}{\left|X_{\max}-C_{x'}\right|+\left|Y_{\max}-C_{y'}\right|}r} \tag{16}$$

通过求解式(16)的逆函数计算出$r$:

$$r=-\frac{\left|X_{\max}-C_{x'}\right|+\left|Y_{\max}-C_{y'}\right|}{n\varepsilon}\left(W_{-1}\left(\frac{\beta-1}{e}\right)+1\right) \tag{17}$$

其中,$\beta$是$[0,1)$服从均匀分布的随机数,$W_{-1}$是Lembert W函数的$(-\infty,-1)$分支.

由$r$和$\theta$求解直角坐标系下的噪音位置点为:

$$l'=l+(r\cdot\cos\theta,r\cdot\sin\theta) \tag{18}$$

噪音位置点需根据坐标精度划分单元网格,并在网格$G$上将其转换为区域$R=A\cap G$中近似的点. 转换后的点仍满足地理不可区分性,证明过程见文献[10]. 算法1描述了噪音数据集的生成过程.

**算法1 噪音轨迹数据集生成算法**

**输入:**轨迹数据集$T$,隐私预算$\varepsilon$,坐标精度$\delta$
**输出:**噪音数据集$T'$

```
1.  计算区域圆A的最大半径 r_max
2.  计算全局敏感度 Δf
3.  生成随机数β和θ,并计算半径r
4.  WHILE i=1 to len(T') DO
5.      FOR j=1 to len(t_i) DO
6.          计算 l'_ij = l_ij + (r·cosθ, r·sinθ)
7.          根据δ划分网格G
8.          将l'_ij映射为l^1_ij
9.          WHILE r ≤ r_max DO
10.             IF l^1_ij ∈ A THEN
11.                 l'_ij ← l^1_ij
12.                 j++
13.             ELSE IF l^1_ij ∉ A THEN
14.                 转换为R中近似点l^2_ij
15.                 l'_ij ← l^2_ij
16.                 j++
17.             END IF
18.             i++
19.         END WHILE
20.     END FOR
21. END WHILE
22. RETURN T'
```

算法1的第1~3行计算了区域圆的相应参数,第4~21行是生成噪音数据集的具体过程,其中第9~19行描述了如何转换极坐标下的噪音位置点.

### 4.2 构造后验概率

假设$A$是区域圆集合,每个区域圆$a$赋予不同的权重$w(a)$,且权重基于位置$l^*$的总体不确定性:

$$u(l^*)=\sum\zeta(l')p(l^*|l')d(l',l^*) \tag{19}$$

其中,$\zeta(l')$表示位置$l'$的先验概率.

由于较低的权重应分配给较高不确定性的区域,设置权重$w(a)$为:

$$w(a)=1-w_0\cdot\frac{u(l^*)-u_{\min}}{u_{\max}-u_{\min}} \tag{20}$$

其中,$u_{\max},u_{\min}$为最大和最小的总体不确定性,$w_0\in[0,1)$为最高不确定性区域的基本权重.

此外,所有区域圆的权重还应满足:

$$\sum_{a\in A}w(a)=1 \tag{21}$$

本文将权重$w(a)$作为区域圆的全局先验概率. 单个区域圆的$n$个位置点大致均匀分布,设置其先验概率$\zeta(l)=w(a)/n$.

基于上述分析,将贝叶斯定理与式(11)的概率密度函数以及式(20)的权重相结合,得出后验概率分布为:

$$\xi(l)=\frac{\zeta(l)\cdot e^{-\frac{\varepsilon}{\Delta f}d(l,l^*)}}{\sum_{a\in A}w(a)\cdot e^{-\frac{\varepsilon}{\Delta f}d(a,l^*)}} \tag{22}$$

### 4.3 数据映射模型

数据映射模型的目标是通过最小化贝叶斯条件隐私期望函数来选择映射位置$l^*$,若贝叶斯条件隐私期望函数为:

$$\sum\xi(l)d(l,l^*) \tag{23}$$

则数据映射函数定义为最小化贝叶斯条件隐私期望函数:

$$l^*=M(l)=\min_{l^*}\sum_l\xi(l)d(l,l^*) \tag{24}$$

当有多个$l^*$满足式(24)时,每个$l^*$被攻击者随机选中的概率均为$h(l^*|l')$,式(24)应更新为:

$$l^* = M(l) = \min_{l^*} \sum_{l} \xi(l) d(l, l^*) h(l^*|l') \tag{25}$$

当且仅当$h(l^*|l')$用于计算$l^*$时$h(l^*|l')>0$,其他情况$h(l^*|l')=0$. 若仅有一个$l^*$满足式(25),随机选择该$l^*$的概率为1,即$h(l^*|l')=1$,此时式(25)退化为式(24).

### 4.4 最优数据映射

由于可能存在多个满足数据映射模型的位置$l^*$. 为了发布最优位置,本文提出最优数据映射方法.

对于任意$l^*$,$\hat{l}$,其数据质量损失与加权距离和,若满足:

$$\begin{aligned} &\mathrm{Qloss}(\hat{l}) \leqslant \mathrm{Qloss}(l^*) \\ &\mathrm{WD}(\hat{l}) \leqslant \mathrm{WD}(l^*) \end{aligned} \tag{26}$$

则称$\hat{l}$为最优发布位置,式(26)为最优数据映射.

本文将数据质量损失定义为真实位置和发布位置之间的贝叶斯无条件隐私期望:

$$\mathrm{Qloss} = \sum \zeta(l) p(l^*|l) d(l, l^*) \tag{27}$$

通过制定线性规划使数据质量损失最小化:

$$\begin{aligned} &\arg\min \sum \zeta(l) p(l^*|l) d(l, l^*) \\ &\text{s.t.}\quad p(l^*|l) \leqslant \mathrm{e}^{\varepsilon} \cdot p(l^*|l') \\ &\sum \zeta(l) p(l^*|l) = \frac{1}{n} \\ &p(l^*|l) \geqslant 0 \\ &\sum p(l^*|l) = 1 \end{aligned} \tag{28}$$

本文采用针对连续选址模型提出的Weiszfeld算法来选择最优位置,此时最优位置选择问题转换为位置理论中的韦伯问题,即:寻找一个位置点$\hat{l}=(x,y)$使得该点到给定位置点$l^*=(a_i,b_i)$的加权距离和最小. 其数学模型为:

$$\mathrm{WD} = \min C = \min \sum w_i d_i(\hat{l}, l^*) \tag{29}$$

其中,$d_i(\hat{l}, l^*) = \sqrt{(x-a_i)^2 + (y-b_i)^2}$.

通过构造迭代法求解加权距离和的最小值:

$$(x_{k+1}, y_{k+1}) = \left( \frac{\sum w_i a_i / d_i(\hat{l}_k, l_k^*)}{\sum w_i / d_i(\hat{l}_k, l_k^*)}, \frac{\sum w_i b_i / d_i(\hat{l}_k, l_k^*)}{\sum w_i / d_i(\hat{l}_k, l_k^*)} \right) \tag{30}$$

Weiszfeld($l^*$)表示迭代结果,当前后两次迭代点之间的距离下降到$\varepsilon$时停止迭代,即:

$$\frac{\| (x_{k+1}, y_{k+1}) - (x_k, y_k) \|_2}{\| (x_k, y_k) \|_2} < \varepsilon \tag{31}$$

其中,$\varepsilon$根据实际情况设置,$\|\cdot\|_2$表示Euclid范数.

算法2描述了本文提出的基于差分隐私的轨迹隐私保护方法(OptDMM).

**算法2　基于差分隐私的轨迹隐私保护方法(OptDMM)**

**输入**:轨迹数据集$T$,噪音数据集$T'$,隐私预算$\varepsilon$

**输出**:发布数据集$\hat{T}$

```
1.  通过聚类算法构造区域圆A
2.  WHILE |A|>0 DO
3.      FOR i=1 to len(T') DO
4.          FOR j=1 to len(t_i) DO
5.              构造先验概率ζ和后验概率ξ
6.              由式(24)计算数据映射函数 l*_ij
7.              IF num(l*_ij) = 1THEN
8.                  l̂_ij ← l*_ij
9.                  j++
10.             ELSE IF num(l*_ij) > 1 THEN
11.                 由约束条件式(27)和式(28)计算 l*_ij
12.                 l̂_ij ← Weiszfeld(l*_ij)
13.                 j++
14.             END IF
15.             i++
16.         END FOR
17.     END FOR
18. END WHILE
19. RETURN T̂
```

算法2首先通过聚类方式构造区域圆(第1行). 接着构造先验概率和后验概率,并计算数据映射函数. 然后判断映射后的新位置数目,若有一个则直接发布;若有多个则通过最优数据映射发布最优位置(第2~18行).

## 5 实验与分析

### 5.1 实验数据与环境

本文针对全局敏感度、隐私保护程度和数据可用性评估算法. Andrés等[10]提出的地理不可区分机制GIDP,Huang等[11]提出的TPA模型,Cunha等[12]提出的LPPM机制,Zhao等[13]提出的轨迹保护方法TLDP以及不使用最优数据映射函数DMM,均是本文(OptDMM)用于对比的算法.

本文使用爱丁堡信息学论坛步行数据库[14]中的轨迹数据,实验选取200条轨迹数据,其中包含6919个轨迹位置点. 实验环境为Windows 10 64位的操作系统,内存空间为12GB,处理器为2核Intel(R) Core(TM) i5-5200U CPU @2.20GHz.

### 5.2 全局敏感度度量

本节度量算法定义的全局敏感度. 全局敏感度越低,加入的噪声值越小. 下面对比相关工作中的P-STM[4]算法、PDP-SAG[5]算法以及PKM[6]算法.

实验选取3个区域圆度量全局敏感度. 区域圆的位置点数$n$分别为58、579、3304,单个移动对象访问的最大位置数$l_{\max}$分别为29、35、40,单个位置的最大访问次数$P$均为3. 度量结果见表1~3.

分析表格数据可知,本文定义的全局敏感度均低于其他算法,并且随着位置点数增加,本文的全局敏感度呈下降趋势,而其他算法呈上升趋势. 这表明当达到相同的隐私保护效果时,本文的全局敏感度可以有效降低添加的Laplace噪声.

**表1 全局敏感度度量$n$=58**

| 方法 | 计算公式 | 结果 |
|---|---|---|
| P-STM | $\max\{\log 2, \log P-\log(\log P)-1\}*l_{\max}$ | 8.730 |
| PDP-SAG | $l_{\max}$ | 29 |
| PKM | $8/3*\log(n)$ | 4.702 |
| Our | $(\lvert X_{\max}-C_{x'}\rvert+\lvert Y_{\max}-C_{y'}\rvert)/n$ | 3.966 |

**表2 全局敏感度度量$n$=579**

| 方法 | 计算公式 | 结果 |
|---|---|---|
| P-STM | $\max\{\log 2, \log P-\log(\log P)-1\}*l_{\max}$ | 10.536 |
| PDP-SAG | $l_{\max}$ | 35 |
| PKM | $8/3*\log(n)$ | 7.367 |
| Our | $(\lvert X_{\max}-C_{x'}\rvert+\lvert Y_{\max}-C_{y'}\rvert)/n$ | 0.394 |

**表3 全局敏感度度量$n$=3304**

| 方法 | 计算公式 | 结果 |
|---|---|---|
| P-STM | $\max\{\log 2, \log P-\log(\log P)-1\}*l_{\max}$ | 12.041 |
| PDP-SAG | $l_{\max}$ | 40 |
| PKM | $8/3*\log(n)$ | 9.274 |
| Our | $(\lvert X_{\max}-C_{x'}\rvert+\lvert Y_{\max}-C_{y'}\rvert)/n$ | 0.069 |

### 5.3 隐私保护程度分析

本节分析算法的隐私保护程度. 隐私预算$\varepsilon$越小,隐私保护程度越高. 下面证明算法满足$\varepsilon$-差分隐私. 若添加的噪声$Z$服从Laplace分布:

$$Z\sim\mathrm{Lap}\left(\frac{\left|X_{\max}-C_{x'}\right|+\left|Y_{\max}-C_{y'}\right|}{n\varepsilon}\right)$$

则本文机制KM满足$\varepsilon$-差分隐私. 证明如下:

假设$p_1$为原始数据集$T$的概率密度函数,$p_2$为噪声$Z$的概率密度函数,则

$$\begin{aligned}\frac{\Pr[\mathrm{KM}(T)\in D]}{\Pr[\mathrm{KM}(T')\in D]}&=\frac{\Pr[C[\mathrm{Clu}(T)]\in D]}{\Pr[C[\mathrm{Clu}(T')]\in D]}\\&=\frac{\Pr[c+z=d]}{\Pr[c'+z=d]}\\&=\frac{p_2[d-p_1(c)]}{p_2[d-p_1(c')]}\end{aligned}$$

其中$z\in Z, d\in D$. 由于$Z$服从Laplace分布,所以

$$\begin{aligned}\frac{p_2[d-p_1(c)]}{p_2[d-p_1(c')]}&\leqslant \mathrm{e}^{\frac{n\varepsilon}{\left|X_{\max}-C_{x'}\right|+\left|Y_{\max}-C_{y'}\right|}\cdot|c-c'|}\\&=\mathrm{e}^{\frac{n\varepsilon}{\left|X_{\max}-C_{x'}\right|+\left|Y_{\max}-C_{y'}\right|}\cdot\Delta f}\\&=\mathrm{e}^{\varepsilon}\end{aligned}$$

因此,本文机制KM满足$\varepsilon$-差分隐私.

为了验证OptDMM算法的隐私保护水平,本文与GIDP算法、TPA算法、LPPM算法以及TLDP算法对比噪音半径的累积分布函数.

本组实验测试不同隐私预算下的累积分布函数. 隐私预算的取值范围为0.1~0.5,取值间隔为0.05. 实验按照噪音半径的不同分为4组,每组的噪音半径分别为250,500,750,1000,每次实验执行10次,取平均值作为最终结果.

如图1所示,$x$轴表示隐私预算,$y$轴表示算法的累积分布函数变化曲线. 由曲线变化可知,隐私预算增大,累积分布函数值随之增大,并且本文算法的累计分布函数值均大于其余四种算法. 这表明当达到相同的隐私保护效果时,本文算法的隐私预算相对较小,隐私保护水平较高.

图2显示了不同噪音半径下的累积分布函数随隐私预算的变化情况. 噪音半径取100,150,200,300. 由图2可知,隐私预算相同时,噪音半径越大,累积分布函数越大,隐私保护水平越高;累积分布函数相同时,噪音半径越大,分配到的隐私预算越小,隐私保护程度越高.

### 5.4 数据可用性分析

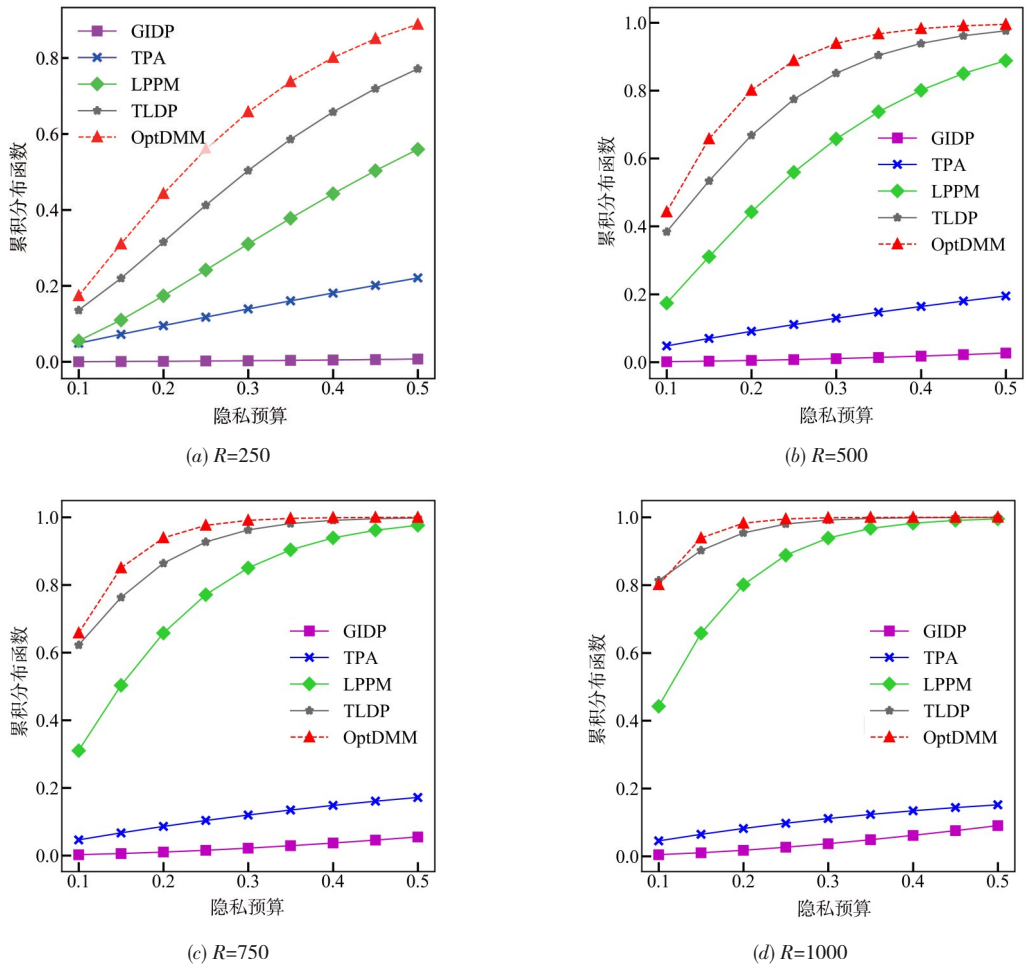
本节分析算法的数据可用性. 首先与其他算法对比平均误差.

$$\text{Average Error}=\frac{\sum_{i=1}^{n}|d(l_i,\hat{l}_i)|}{n}\tag{32}$$

其中,$d(\cdot)$为欧氏距离,$n$为轨迹位置点的数目.

本组实验测试不同隐私预算的平均误差. 隐私预算值取0.1,0.2,0.3,0.4,0.5,每次实验执行10次,取平均值作为最终结果.

如图3所示,$x$轴表示隐私预算,$y$轴表示算法的平均误差. 由图3可知,隐私预算增大,平均误差逐渐减小. 且本文算法的平均误差均低于其他算法,这表明本文算法的数据可用性相对较高. 观察图3(b)可知,隐私




(*a*) *R*=250　　(*b*) *R*=500

(*c*) *R*=750　　(*d*) *R*=1000

图1　不同算法的累积分布函数

预算的改变对本文算法的平均误差的影响不大,这表明本文算法相对稳定.

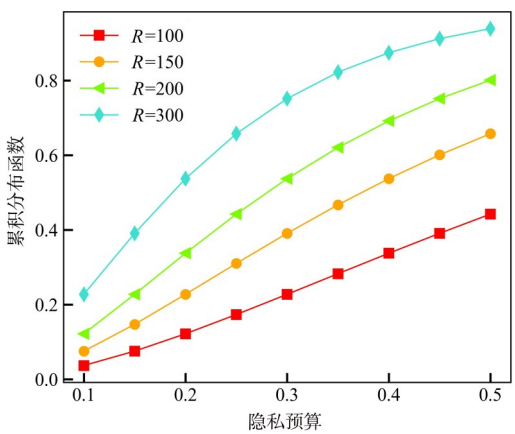



图2　不同噪音半径下的累积分布函数

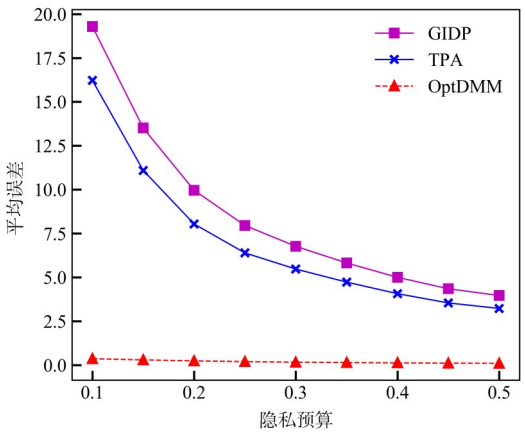
为使上述结论更具说服力,本文通过平均数据质量损失进一步衡量数据可用性.

$$\text{Average Qloss} = \frac{\sum_{i=1}^{n} |\text{Qloss}(\hat{l}_i)|}{n} \tag{33}$$

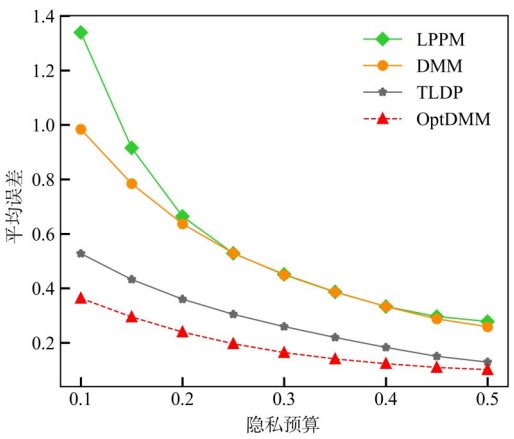
其中,$\text{Qloss}(\hat{l}_i)$表示发布轨迹点的数据质量损失.

本组实验测试不同隐私预算的平均数据质量损失. 隐私预算值取0.10,0.12,0.14,0.16,0.18,0.20,每次实验执行10次,取平均值为最终结果.

如图4所示,$x$轴表示隐私预算值,$y$轴表示算法的平均数据质量损失. 由图4可知,平均数据质量损失随隐私预算的增大而减小,且本文算法的平均数据质量损失总是低于其他算法,这进一步说明本文算法的数据可用性较高. 同时,随着隐私预算的增加,算法的平均数据质量损失变化较小,这也进一步表明本文算法的稳定性良好.

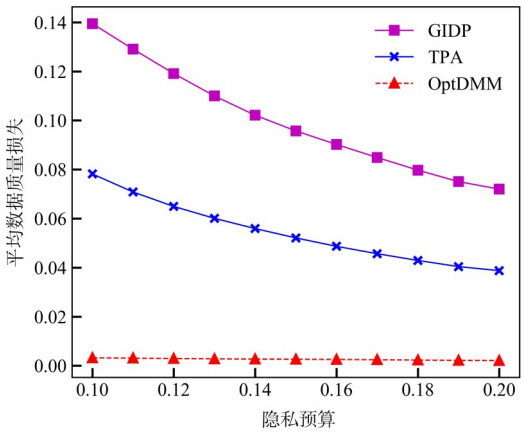



(*a*) 算法对比结果 1

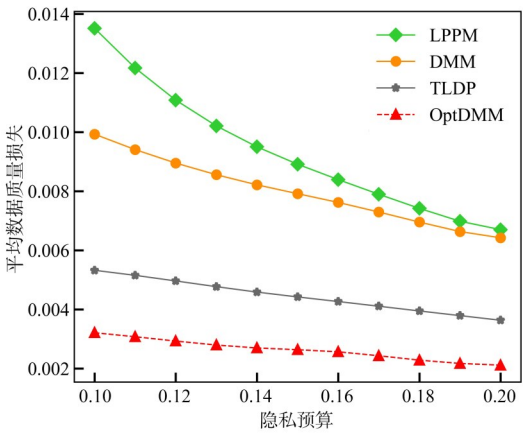



(*b*) 算法对比结果 2

图 3 不同算法的平均误差对比




(*a*) 算法对比结果 1




(*b*) 算法对比结果 2

图 4 不同算法的平均数据质量损失对比

## 6 结论

本文提出了一种用于轨迹数据的隐私保护方法. 该方法结合地理不可区分机制对原始轨迹数据添加拉普拉斯噪音,避免了噪音过大影响数据发布效果的问题;通过构造数据映射模型,混淆用户的真实轨迹位置;通过最优数据映射函数,提高发布数据的可用性. 实验结果表明,本文方法既能有效保护用户隐私,也能保证发布数据的有效性.

### 参考文献

[1] 杨高明,朱海明,方贤进,苏树智. 局部差分隐私约束的关联属性不变后随机响应扰动[J]. 电子学报, 2019, 47(5): 1079 – 1085.
YANG Gao-ming, ZHU Hai-ming, FANG Xian-jin, SU Shu-zhi. Invariant post-random response perturbation for correlated attributes under local differential privacy constraint [J]. Acta Electronica Sinica, 2019, 47(5): 1079 – 1085. (in Chinese)

[2] 叶阿勇,林少聪,马建峰,许力. 一种主动扩散式的位置隐私保护方法[J]. 电子学报, 2015, 43(7): 1362 – 1368.
YE A-yong, LIN Shao-cong, MA Jian-feng, XU Li. An active diffusion based location privacy protection method [J]. Acta Electronica Sinica, 2015, 43(7): 1362 – 1368. (in Chinese)

[3] 陈传明,林文诗,俞庆英,罗永龙. 一种基于单点收益的轨迹隐私保护方法[J]. 电子学报, 2020, 48(1): 143 – 151.
CHEN Chuan-ming, LIN Wen-shi, YU Qing-ying, LUO Yong-long. A trajectory privacy-preserving method based on single point gain [J]. Acta Electronica Sinica, 2020, 48(1): 143 – 151. (in Chinese)

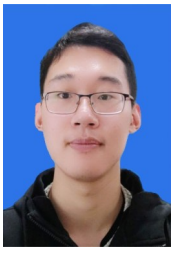
[4] Wang S, Nepal S, Sinnott R O, Rudolph C. P-STM: privacy-protected social tie mining of individual trajectories [A]. Proceedings of the 2019 IEEE International Con-

ference on Web Services [C]. Italy: IEEE, 2019. 1 – 10.
[5] Deldar F, Abadi M. PDP-SAG: Personalized privacy protection in moving objects databases by combining differential privacy and sensitive attribute generalization [J]. IEEE Access, 2019, 7: 85887 – 85902.
[6] Feldman D, Xiang C Y, Zhu R H, Rus D. Coresets for differentially private *k*-means clustering and applications to privacy in mobile sensor networks [A]. Proceedings of the 16th ACM/IEEE International Conference on Information Processing in Sensor Networks [C]. USA: ACM, 2017. 3 – 15.
[7] Dong Y L, Pi D C. Novel Privacy-preserving algorithm based on frequent path for trajectory data publishing [J]. Knowledge-Based Systems, 2018, 148: 55 – 65.
[8] Dwork C. Differential privacy [A]. Proceedings of the 33rd International Colloquium on Automata, Languages and Programming [C]. Italy: Springer, 2006. 1 – 12.
[9] McSherry F, Talwar K. Mechanism design via differential privacy [A]. Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science [C]. USA: IEEE, 2007. 94 – 103.
[10] Andrés M E, Bordenabe N E, Chatzikokolakis K, Palamidessi C. Geo-indistinguishability: differential privacy for location-based systems [A]. Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security [C]. Germany: ACM, 2013. 901 – 914.
[11] Huang H Y, Niu X, Chen C, Hu C Q. A differential private mechanism to protect trajectory privacy in mobile crowd-sensing [A]. Proceedings of the 2019 IEEE Wireless Communications and Networking Conference [C]. Morocco: IEEE, 2019. 1 – 6.
[12] Cunha M, Mendes R, Vilela J P. Clustering geo-indistinguishability for privacy of continuous location traces [A]. Proceedings of the 4th International Conference on Computing, Communications and Security [C]. Italy: IEEE, 2019. 1 – 8.
[13] Zhao X D, Pi D C, Chen J F. Novel trajectory privacy-preserving method based on clustering using differential privacy [J]. Expert Systems with Applications, 2020, 149: 113241.
[14] Majecka B. Statistical Models of Pedestrian Behaviour in the Forum [D]. UK: University of Edinburgh, 2009.

## 作者简介

**袁水莲**　女,1997年11月生于山东济宁.现为南京航空航天大学硕士研究生.主要研究方向为数据挖掘和隐私保护.
E-mail:shirley_ysl@nuaa.edu.cn

**皮德常(通信作者)**　男,1971年11月生于河南周口.现为南京航空航天大学教授、博士生导师.主要研究方向为数据挖掘和隐私保护.
E-mail:dc.pi@nuaa.edu.cn



**胥　萌**　男,1997年8月生于江苏盐城.现为南京航空航天大学硕士研究生.主要研究方向为数据挖掘.
E-mail: xu_meng@nuaa.edu.cn