

## 基于出租车轨迹数据挖掘的乘客候车时间预测\*

齐观德, 潘遥, 李石坚, 潘纲

(浙江大学 计算机科学与技术学院, 浙江 杭州 310027)

通讯作者: 李石坚, E-mail: shijianli@zju.edu.cn

**摘要:** 随着城市人口的增长与交通问题的加剧, 出行也变得越来越困难. 出租车是重要的公共交通工具; 对于乘客来说, 了解等候出租车需要花费的时间有助于他们安排行程与选择最优的等候地点. 针对这个问题, 使用出租车轨迹的历史数据, 预测乘客在某时某地等候出租车需要的时间. 算法使用参数与非参数模型, 通过对空车到达过程的建模, 预测单个乘客等候出租车需要的时间. 使用大规模的杭州出租车数据, 结合仿真来验证预测算法. 实验评估结果显示, 参数模型对等候时间预测的平均误差约为 4.5 分钟, 并且参数模型预测误差小于 5 分钟的概率约为 83%.

**关键词:** 乘客候车时间预测; 空车到达时间间隔; 出租车轨迹

中文引用格式: 齐观德, 潘遥, 李石坚, 潘纲. 基于出租车轨迹数据挖掘的乘客候车时间预测. 软件学报, 2013, 24(Suppl. (2)): 14-23. <http://www.jos.org.cn/1000-9825/13019.htm>

英文引用格式: Qi GD, Pan Y, Li SJ, Pan G. Predicting passengers' waiting time by mining taxi traces. Ruan Jian Xue Bao/Journal of Software, 2013, 24(Suppl. (2)): 14-23 (in Chinese). <http://www.jos.org.cn/1000-9825/13019.htm>

### Predicting Passengers' Waiting Time by Mining Taxi Traces

QI Guan-De, PAN Yao, LI Shi-Jian, PAN Gang

(Department of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China)

Corresponding author: LI Shi-Jian, E-mail: shijianli@zju.edu.cn

**Abstract:** Travelling in metropolis is becoming increasingly difficult for people due to the growing population and busier traffic. Since taxi is important public traffic tools in cities, knowing how long it will take to find a taxi can be helpful for customers to plan their schedule and choose the best place to wait. This paper presents a method to predict the waiting time for a customer at a given time and location from historical taxi trajectories. With a parametric and a non-parametric model for arriving flow of vacant taxis, it predicts the time a single passenger needs to wait for a taxi service. Emulation experiment with large-scale taxi dataset in Hangzhou was conducted to verify the new method. The evaluation results validate that the mean prediction errors for both the parametric and non-parametric model are about 4.5 minutes. Moreover, the probability of that the prediction error is below 5 minutes for parametric model is about 83 percent.

**Key words:** prediction of passenger waiting time; arriving interval of vacant taxis; taxi traces

随着移动设备的普及与定位技术的成熟, 人或物的移动轨迹被记录并产生大量的轨迹数据. 这些数据取自日常生活中的一些片段, 反映了个人行为、人际关系与城市动态. 常见的移动数据来自移动设备, 如手机<sup>[1]</sup>、iPad, 或交通工具, 如公共自行车<sup>[2]</sup>、公共交通卡<sup>[3]</sup>、出租车<sup>[4]</sup>.

出租车是城市的重要交通工具, 出租车 GPS 数据记录了城市交通、人群移动的动态变化. 据统计, 杭州市平均每天有 60 万次的出租车载客轨迹, 占公共交通的 1/7. 深圳市每天有一百万乘客使用出租车出行, 占交通客流总量的 1/9. 另一方面, 出租车 GPS 轨迹数据, 由于覆盖广、精度高、隐私好的优点<sup>[5]</sup>, 已经成为文献研究中使用的一种重要的移动轨迹数据.

\* 基金项目: 国家重点基础研究发展计划(973)(2013CB329504); 国家科技支撑计划(2012BAH94F01); 浙江省钱江人才计划(2011R10078)

收稿时间: 2012-06-15; 定稿时间: 2013-07-22

出租车服务的形式有两种,一种是乘客通过服务中心,提交打车需求,服务中心发送消息给出租车并安排车辆;另一种是乘客在路边等候,直到有空车路过。在国外,出租车服务的主要形式是预约;而在国内,由于巨大的城市客流量,出租车主要以第2种形式服务。目前,有一些应用试图通过增加第1种服务的比例来方便人们的出行。例如,“快的打车”应用试图通过额外付费的方式鼓励出租车司机响应附近乘客的请求。但是,这些应用中乘客在发出叫车请求后仍然需要等待出租车的响应;而且额外付费的方式也阻碍了叫车服务成为惯常的服务形式。

在这些出租车服务中,乘客经常需要等候出租车的到来。候车问题本质在于交通供需在时空上的随机性与不均衡。首先,出租车数量是有限的,不可能随时随地提供出租车服务;而乘客的出行需求是随时随地都可能发生的,因此乘客需要等候;其次,出租车不同于公交车,其服务是不定点的,具有一定的偶然性,故而候车的时间难以预测。

未知的候车时间会给乘客带来诸多不便:(1) 等候,尤其是无法预期的等候,是一个无聊的、令人烦躁的过程;(2) 未知的等候时间不利于乘客计算出行时间,安排行程;(3) 时间即金钱,缺乏对候车时间的认识不利于乘客选择交通工具。预测候车时间有利于:(1) 帮助乘客建立心理预期,减少焦虑;(2) 统筹安排行程;(3) 选择合适的地点进行等待。

为了解决候车时间预测的问题,我们根据出租车历史轨迹数据建立模型,预测乘客等候出租车时间。候车时间预测问题是指给定热点位置  $L$ ,当前时间段  $H$ ,计算乘客等到空车的候车时间。通过提取出租车乘客上车位置数据,对乘客上车位置进行聚类,获取乘客上车的热点区域。然后使用每个区域每个时段的空车到达间隔来计算乘客在该区域该时段的候车时间。本文是 HHME 2012 论文“基于出租车轨迹数据挖掘的乘客候车时间预测”的扩充版本,初步结果已在 HHME 会议论文集中发表<sup>[6]</sup>。

## 1 相关工作

车辆轨迹数据的研究可以为与交通相关的所有主体,包括市民、驾驶者与政府改善交通服务。(1) 预测交通费用:交通费用是乘客关注的重要问题;但由于出租车计价考虑行驶时间,使用出租车出行的交通费用并不固定而难以准确预测。新加坡管理大学的 Balan 等人分析出租车服务的历史数据,告诉乘客两地的历史通行时间与理论花费<sup>[7]</sup>。(2) 最优路径导航:如何寻找最优路径是一个重要问题。由于道路状况不同(如拥堵),最优路径并不是简单的最短路径。微软亚洲研究院的谢幸和郑宇等人认为出租车司机熟悉城市道路状况,因而往往选择最好的路径。他们从出租车轨迹数据中挖掘出租车司机中的专家,从专家的轨迹中学习最优的路线,提供导航服务<sup>[8-10]</sup>。(3) 预测道路状况:城市交通状况的主要信息是道路的拥塞状况,出租车可以作为传感器感知城市道路的拥塞状况。深圳先进技术研究院使用出租车辆行驶历史数据中的速度信息,预测城市道路路况<sup>[11]</sup>。(4) 监管交通服务:对于政府而言,出租车司机的服务有好又坏,如何评价和监管出租车司机的服务是重点。张大庆教授等发现,有一些出租车司机的行驶轨迹不同于大部分出租车司机,他认为这些轨迹是异常轨迹,异常轨迹的监测有助于监督出租车的驾驶行为<sup>[12]</sup>。(5) 评价交通规划:交通规划是城市规划的重要组成,交通规划直接影响人们出行的路线选择。微软亚洲研究院的郑宇和谢幸等通过分析轨迹数据中出租车绕路、低速行驶等现象,认为这些现象反映了目前交通规划中的不足<sup>[13]</sup>。Pan 等人<sup>[14]</sup>综述了出租车轨迹数据挖掘的工作。出租车乘客等候时间预测是交通领域的典型问题。Deng 等人<sup>[15]</sup>使用排队论模型研究出租车调度方案,考虑以一定概率满足乘客需求时应该需要的车辆数目,以及在此情形下的乘客等候时间。Yang 等人<sup>[16]</sup>使用联立方程模型对城市整体的出租车数量、出租车计价、乘客数量、乘客等候时间等进行建模。Yang 等人<sup>[17]</sup>还使用网络模型对大面积区域之间乘客流动建立模型,预测出租车运营的供求关系以及在供求计算平衡时的乘客等候时间与出租车服务质量。这些研究仍然没有充分解决出租车乘客等候时间预测问题,排队论模型缺乏出租车与乘客到达速率的实际数据;而联立方程与网络模型往往只能简单地对城市大面积区域进行建模,忽略了乘客等候时间在不同小区域之间的差异性。本文使用真实出租车轨迹数据,细粒度地分析城市小区域中乘客等候时间问题,通过实际数据中的空车到达间隔数据,建模并预测乘客等候时间。

## 2 前期工作

### 2.1 定义

**定义 1(出租车轨迹).** 出租车轨迹是指一个 GPS 数据点的时间序列,  $Tr: p_1 \rightarrow p_2 \rightarrow \dots \rightarrow p_i \rightarrow \dots \rightarrow p_n$ , 其中每个数据点  $p_i$  是一个包含经纬度、时刻与载客状态的四元组:  $p_i \triangleq (lat, long, t, s)$ .

**定义 2(上下客次数).** 出租车上(下)客对应于数据中载客状态值从空车(有人)转变为有人(空车). 一个区域的上下客次数可以用区域内发生的载客状态转变次数来表示.

**定义 3(热点).** 热点是指上车乘客较为密集的区域. 本文从应用角度出发, 主要关注热点区域的候车时间预测问题. 此外, 我们还抛弃了诸如汽车站与火车站这种有乘客等候队列的区域, 并假设大部分热点区域乘客可以在空车到达时立即搭载.

**定义 4(空车时间间隔).** 空车时间间隔是同一个区域相邻两辆车经过的时间间隔. 空车时间间隔的计算需要判断空车是否经过热点. 我们通过将 GPS 点匹配到地图生成的轨迹与热点是否相交来判断.

**定义 5(候车时间预测).** 给定热点位置  $L$ , 当前时间段  $H$ , 我们要预测平均等候时间  $WT_{L,H}$ .

### 2.2 数据集描述

使用出租车历史轨迹数据<sup>[18]</sup>来计算乘客等候时间, 典型的出租车轨迹数据每条记录包含的主要信息有:

- 出租车 ID: 出租车在数据集中的唯一匿名标识;
- 时间: 数据采样时刻, 包括年月日, 时分秒信息;
- 经度: 采样时刻出租车所处经度;
- 纬度: 采样时刻出租车所处纬度;
- 载客状态: 采样时刻出租车是否载客.

### 2.3 数据预处理

**区域定义:** 本文使用热点, 即区域上下客频繁的区域, 作为研究对象. 乘客的上下客位置可以看作是二维地图上的点, 我们使用改进的 DBSCAN 方法对这些点聚类, 获得城市范围内的热点区域, 每个热点表示为地图上的一个多边形区域. 在本文中, 我们使用一周的出租车数据, 生成乘客下车的位置, 通过聚类, 生成热点数量为 952 个<sup>[4]</sup>(如图 1 所示).



图 1 热点位置可视化, 不同颜色对应热点区域的功能

道路地图表示:为了计算出租车是否经过区域,我们需要道路地图来匹配出租车 GPS 数据.道路地图可以表示为一个图,  $G = \langle V, E \rangle$ .  $V$  表示道路段端点的集合,每个端点  $v_i$  有地图坐标  $(x_i, y_i)$ ;  $e_{i,j} = 1$  表示在  $v_i$  和  $v_j$  两个点之间存在直线段道路连接.

GPS 轨迹匹配:如何判断车辆在任意两个相邻采样点  $p_i : (x_i, y_i)$  和  $p_{i+1} : (x_{i+1}, y_{i+1})$  之间的轨迹?对于任意的 GPS 坐标  $(x, y)$ , 将其匹配到距离最近的路段  $e_{i,j}$  上的点  $(x', y')$ . 不妨设  $(x_i, y_i)$  匹配至路段  $e_{i',j'}$  的  $(x'_i, y'_i)$ ;  $(x_{i+1}, y_{i+1})$  匹配至路段  $e_{i'+1, j'+1}$  的  $(x'_{i+1}, y'_{i+1})$ . 在图  $G$  中临时添加顶点  $v_s : (x'_i, y'_i)$ ,  $v_d : (x'_{i+1}, y'_{i+1})$ , 边  $e_{s,i'}$ ,  $e_{s,j'}$ ,  $e_{d,i'+1}$ ,  $e_{d,j'+1}$ . 计算  $v_s$  和  $v_d$  之间的最短路径作为车辆轨迹.

空车到达间隔计算:空车到达表现在数据上是车辆轨迹与区域相交.对每个区域,我们获得其每天的空车到达时间序列,该序列的一阶差分为空车到达间隔时序数据.任意两个相邻采样点确定一段轨迹;假设车辆在相邻两个采样点之间匀速行驶,我们可以得知车辆行经这段轨迹上任意点的时刻.遍历所有区域与所有空车轨迹,将每段空车轨迹与每个区域的交叉点对应时刻作为空车到达时间.每个区域一天的空车到达时间数据的有序排列即为空车到达时间序列数据.

### 3 候车时间预测模型

我们使用参数和非参数的方法对候车时间进行预测.参数方法建立空车到达间隔的指数分布模型,然后对乘客的候车时间进行预测.由于指数分布的特性,参数方法不需要假设乘客到达的分布.非参数方法假设乘客随机均匀地到达,计算历史的空车到达时间,计算历史的等候时间,用历史平均值来预测候车时间.

#### 3.1 基于空车到达间隔分布的候车时间预测

##### 3.1.1 空车到达间隔建模

一定时间内,到达区域的车辆数量一般认为服从泊松分布,即车辆到达的间隔服从指数分布.出租车实际运营随时间变化的差异性决定了一般指数分布是不合适的.因为当空车到达数量服从泊松分布时,空车到达数量随时间而平稳地增长.这种平稳的增长速率与实际的出租车运营是相悖的,例如,晚上去景区的空出租车显然比白天更少.

现有的研究表明,一定时间内,到达区域的车辆数量服从非齐时泊松分布<sup>[19,20]</sup>,即不同时刻车辆到达的速率不同.车辆到达的速率是一个随时间变化的函数  $\lambda = f(t)$ , 本文假定这个函数是一个分段常值函数,即到达速率在一个时间段内是恒定的  $f(t) = \lambda_i, t_i \leq t < t_{i+1}$ . 进而,在第  $i$  个时间段,空车到达间隔  $\Delta t$  服从分时的指数分布  $p_i(\Delta t) = \lambda_i e^{-\lambda_i \Delta t}$ .

##### 3.1.2 乘客候车时间建模

如图 2 所示,单个乘客在每个区域的候车时间期望可以基于空车到达间隔的分布计算.以一天的第  $j$  个时间段为例,假设其中空车间隔  $\Delta t$  服从分布  $p_j(\Delta t) = \lambda_j e^{-\lambda_j \Delta t}$ . 不妨设前一辆车在  $t$  时刻经过,乘客以  $f(t_0)$  的概率密度在  $t + t_0$  时刻出现,则等候时间  $t_w$  分布为

$$p(t_w) = \frac{\int f(t_0) \lambda_j e^{-\lambda_j (t_w + t_0)} dt_0}{\int \int f(t_0) \lambda_j e^{-\lambda_j (t_w + t_0)} dt_w dt_0} = \frac{\lambda_j e^{-\lambda_j t_w} \int f(t_0) \lambda_j e^{-\lambda_j t_0} dt_0}{\int f(t_0) \lambda_j e^{-\lambda_j t_0} dt_0 \int \lambda_j e^{-\lambda_j t_w} dt_w} = \lambda_j e^{-\lambda_j t_w} \quad (1)$$

故而,无论单个乘客出现的概率分布是如何的,乘客等候时间的分布都是仍然服从指数分布.等候时间的期望为

$$E(t_w) = \int p(t_w) t_w dt_w = \int \lambda_j e^{-\lambda_j t_w} t_w dt_w = \frac{1}{\lambda_j} \quad (2)$$

如何估计  $\lambda_j$ ?  $\lambda_j$  表示第  $j$  个时间段空车到达的速率,可以通过历史数据进行参数估计.我们用最大似然法估算  $\lambda$  的取值:

$$\hat{\lambda} = \arg \max \hat{l}(\lambda | \Delta t_1, \dots, \Delta t_n) = \arg \max_{\lambda} \lambda^n e^{-\lambda \sum \Delta t_i} = \sum t_i / n = \frac{1}{\mu} \quad (3)$$

其中,  $\Delta t_1, \dots, \Delta t_n$  为样本中的空车到达时间间隔,  $\mu$  为到达时间间隔的样本均值.



图2 空车时间间隔与等候时间的关系

### 3.2 候车时间非参数估计

候车时间的非参数估计直接使用历史的空车到达时间;通过随机产生(重复一万次)乘客到达时间,计算乘客的候车时间,作为历史值;进而将历史值的平均用作预测候车时间.历史计算需要假定乘客到达时间的概率分布,给定一个长为  $l$  的时间段,我们假设乘客以  $1/l$  的概率密度均匀出现于该时刻.此外,不妨假设在第  $j$  个时间段中,已知的空车到达时间分别为  $t_1, t_2, \dots, t_n, 0 = t_0 \leq t_1 \leq t_2 \leq \dots \leq t_n \leq l \leq t_{n+1}, t_{n+1}$  是该时间段后首次空车到达的时间.

当乘客在  $t_i$  和  $t_{i+1}$  之间出现的时候,平均等候时间为  $wt_i$ ;处于此区间的概率为  $p_i$  与区间长度  $l_i$  成正比,即  $l_i/l$ ;总平均等候时间为  $\sum_i p_i wt_i$ .其中,

$$wt_i = \begin{cases} t_1/2, & i=0; \\ (t_{i+1}-t_i)/2, & 1 \leq i < n; \\ t_{n+1}-(l+t_n)/2, & i=n; \end{cases} \quad (4)$$

$$p_i = \begin{cases} t_1/l, & i=0; \\ (t_{i+1}-t_i)/l, & 1 \leq i < n; \\ (l-t_n)/l, & i=n. \end{cases} \quad (5)$$

## 4 实验与模型评估

本文使用假设检验评估模型的适用性,并从候车时间的绝对误差和误差分布来评估模型预测性能.由于实际乘客到达时间数据的缺乏,实验采用半模拟的方式评估预测性能.本文的实验评估考虑了时间段长度、训练数据量、预测时间以及不同模型对预测性能的影响.

### 4.1 实验方法

本文使用卡方检验来评估空车到达间隔是否服从指数分布.我们聚集每个区域,每个时间窗内相同时间段的空车到达间隔数据,进行卡方检验.

本文使用两种不同的模型来预测候车时间并评估其性能.主要实验方法为:(1) 确定实验参数:训练数据长度(时间窗长)、预测有效性、时间段长度;(2) 使用一定时间窗内的数据作为训练数据,预测未来某天每个时间段的候车时间;预测方法如前所述;(3) 对应每个预测值,我们使用半模拟的方法计算仿真的候车时间作为真实值(即假设乘客到达是随机均匀的,重复地随机产生乘客并计算候车时间);(4) 计算预测值与真实值的绝对误差以及误差的分布来评估模型.

### 4.2 卡方检验

我们使用卡方检验判断空车时间间隔是否服从指数分布.我们将每个区域按时间窗和时段将数据划分为一个个单独的样本;对每个样本的空车时间间隔数据进行卡方检验,统计通过检验(置信度  $> 0.05$ )的比例.卡方检验的实验使用的时间窗长度是 10 天,时段长度是 15 分钟.实验结果表明,79.02%的样本可以通过卡方检验.

### 4.3 参数优化

本文预测每个地点每个时间段的候车时间,时间段长度影响模型的精细程度以及数据量,进而影响模型的预测误差.时间段长度越短,模型越精细,模型预测误差减小;但这同时使得每个时间段内的数据量越少,模型预测误差增大.

本文采用一定天数的出租车历史轨迹数据训练模型.在*i.i.d* (独立同分布)假设下,不同天同一时间段的数据服从相同的分布,从而训练数据量越大,模型训练越充分,预测误差越小.但是,如果不同天的训练数据服从不同的分布,训练数据长度的增加可能会导致模型的偏离,产生更大的预测误差.在训练数据分布不遵循*i.i.d* 假设的情况,采用较小的数据量反而是更有效的.

本文使用基于空车间隔的参数模型与非参数模型预测等候时间.我们对不同时间段长度和训练数据长度取值,考虑时间段长度从0.5小时~12小时,训练数据长度从10天~210天;观察参数和非参数模型预测准确率的变化情况,以获得最佳的模型参数.图3表明,当时间段长度为6小时,训练数据长度为30天时,参数模型的预测误差最小.图4表明,当时间段长度为0.5小时,训练数据长度为210天时,非参数模型的预测误差最小.

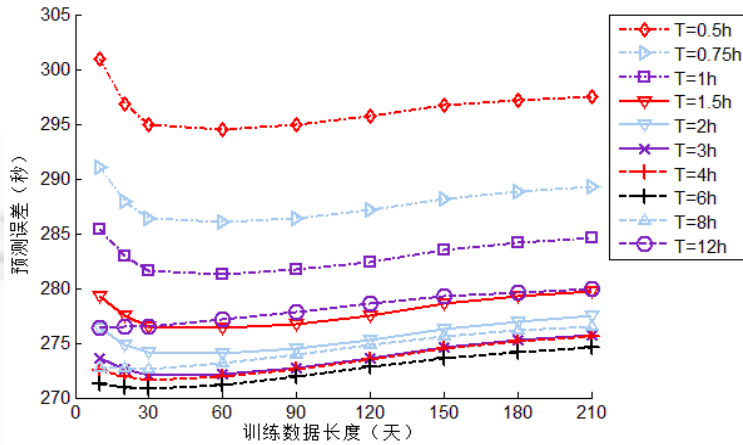


图3 不同时间段长度( $T$ )和训练数据长度下,参数模型的预测误差

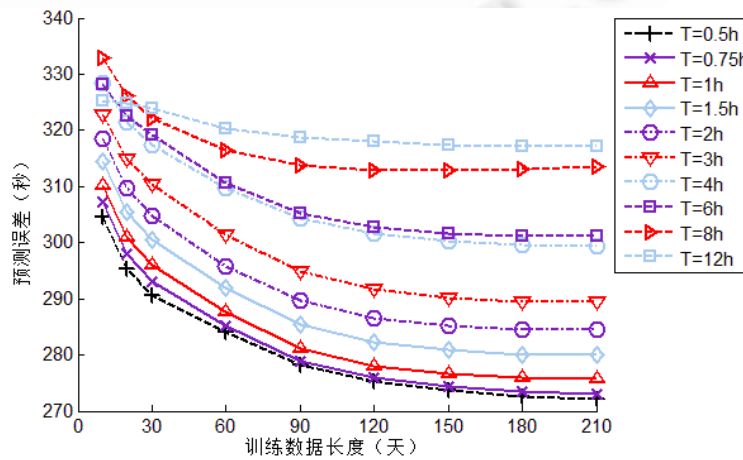


图4 不同时间段长度( $T$ )和训练数据长度下,非参数模型的预测误差

### 4.4 模型选择

本文使用基于空车间隔的参数模型与非参数模型预测等候时间.我们使用不同的时间段长度和训练数据量进行实验,比较参数与非参数模型的预测误差.

首先,我们分别选定两个模型最佳的训练数据量(参数模型 30 天,非参数模型 210 天),考虑时间段长度从 0.5 小时~12 小时对预测误差的影响.实验结果(如图 5 所示)表明,参数模型在绝大多数情形优于非参数模型.并且非参数模型随着时间段长度增加,预测误差增大.这种变化产生的原因模型泛化性能有限,随时间段增长而变得更加粗糙.

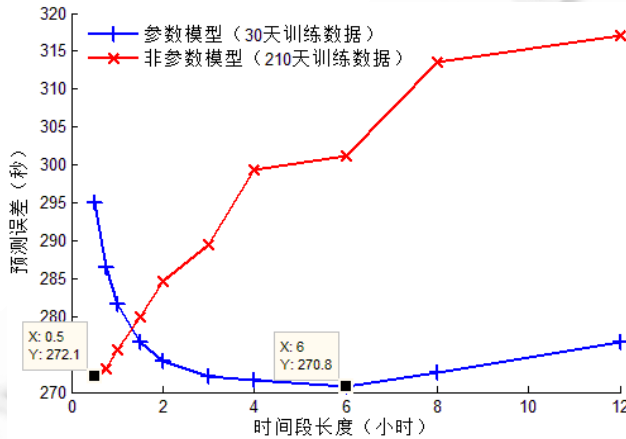


图 5 时间段长度对参数与非参数模型的预测误差影响

其次,我们选定两个模型最佳的时间段长度(参数模型 6 小时,非参数模型 0.5 小时),训练数据量在 10 天~210 天间变化,计算预测误差.实验结果(如图 6 所示)中参数模型在多数情形下优于非参数模型.实验表明非参数模型需要更多的训练数据,才能有更好的预测性能,这不仅意味着更长的训练时间,而且说明建立非参数模型需要更大的训练代价.相反,参数模型对于训练数据量则远没有那么敏感.

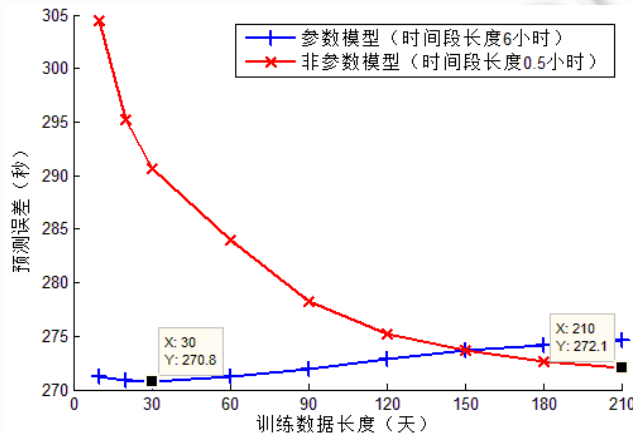


图 6 训练数据量对参数与非参数模型的预测误差影响

### 4.5 预测结果的可用性

本文已经选择了较好的预测模型与参数组合,并计算了模型的预测误差.但是实际应用中,我们不仅要考虑预测误差均值的大小,还要考虑预测误差的分布情况;不仅要考虑对相邻时间点的预测性能,还要考虑模型在长时间预测中的可用性.



### 4.5.1 预测误差分布

预测误差分布对实际应用非常重要,因为即使预测误差均值较小,预测误差离散度越大时,预测出错的几率会越大.我们选取参数模型,设定其时间段长度为 6 小时,训练数据量为 30 天,计算预测误差的概率分布.如图 7 所示,本文绘制了预测误差的概率分布直方图和累积分布曲线;从图中可知,平均预测误差约为 4.5 分钟,预测误差小于 5 分钟的概率是 83.21%.

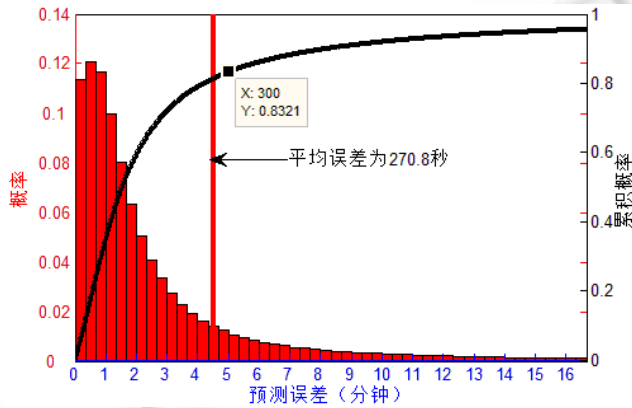


图 7 参数模型在时间段长度为 6 小时,训练数据为 30 天时的误差分布

### 4.5.2 长时间预测中的误差

在长时间预测实验中,我们使用参数模型,30 天数据训练,时间段长度为 6 小时,考虑预测长度对预测误差的影响.预测长度是指训练数据与预测数据之间的距离.例如,当训练长度数据量为 30 天,预测长度是 10,我们使用第 1 个月的数据训练,预测第 41 天的等候时间.如果预测误差随着预测长度改变很小,说明模型可以有效地进行长时间预测.实验结果(如图 8 所示)表明,随着预测长度增加,预测误差呈现增大的趋势,但是增大的误差并不太多.

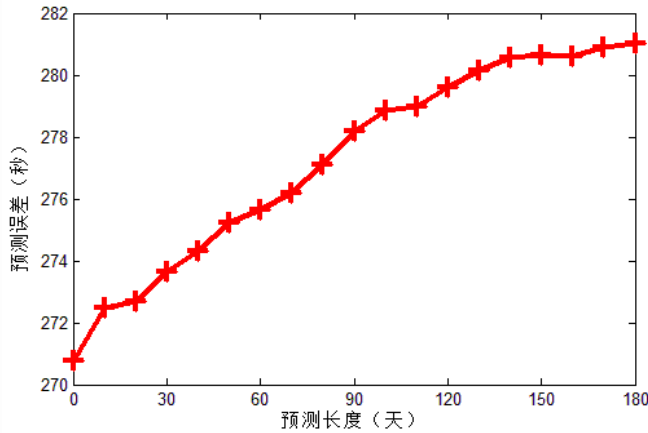


图 8 预测长度对预测误差的影响

## 5 总结与展望

本文建立模型对一个区域在一个时间段内的候车时间进行了建模与预测.首先我们根据出租车的历史轨迹数据提取区域空车到达事件.我们使用非齐时泊松过程对空车到达建模,并使用历史数据来评估模型参数:空车到达间隔的期望.然后,本文提出了参数与非参数模型基于空车到达间隔预测乘客等候时间.

本文对参数和非参数模型进行了参数优化与模型评估,并验证了在一个较短的时间段中,空车到达间隔确



实服从指数分布,也验证了非齐时泊松过程建模空车到达过程是可行的.我们使用真实数据,结合仿真的方法对预测模型进行了评估,发现等候时间的预测平均误差为 4.5 分钟,并且误差小于 5 分钟的概率约为 83%.

在现有工作中,我们假设乘客在空车到达时可以立即搭载出租车,并未考虑存在乘客队列时的情况,未来拟改进该部分工作.

**致谢** 在此,我们向对本文研究提供过帮助的老师 and 同学表示感谢.

### References:

- [1] Calabrese F, Smoreda Z, Blondel VD, Ratti C. Interplay between telecommunications and face-to-face interactions: A study using mobile phone data. *PLoS ONE*, 2011,6(7):e20814.
- [2] Froehlich J, Neumann J, Oliver N. Sensing and predicting the pulse of the city through shared bicycling. In: Boutilier C, ed. *Proc. of the IJCAI'09*. AAAI Press, 2009. 1420–1426.
- [3] Lathia N, Capra L. How smart is your smartcard? Measuring travel behaviours, perceptions, and incentives. In: Landay J, eds. *Proc. of the Ubicomp'11*. New York: ACM, 2011. 291–300.
- [4] Pan G, Qi GD, Wu ZH, Zhang DQ, Li SJ. Land-Use classification using taxi GPS traces. *IEEE Trans. on Intelligent Transportation Systems*, 2013,14(1):113–123.
- [5] 潘纲,李石坚,齐观德,张王晟. 移动轨迹数据分析与智慧城市. *中国计算机学会通讯*, 2012,8(5):31–36.
- [6] 齐观德,李石坚,潘遥,潘纲. 基于出租车轨迹数据挖掘的乘客候车时间预测. 见:第 8 届和谐人机环境联合学术会议, 2012.
- [7] Balan RK, Khoa NX, Jiang LX. Real-Time trip information service for a large taxi fleet. In: Agrawala A, ed. *Proc. of the MobiSys 2011*. New York: ACM, 2011. 99–112.
- [8] Yuan J, Zheng Y, Xie X, Sun GZ. Driving with knowledge from the physical world. In: Apte C, ed. *Proc. of the KDD 2011*. New York: ACM, 2011. 316–324.
- [9] Yuan J, Zheng Y, Zhang CY, Xie WL, Xie X, Sun GZ, Huang Y. T-Drive: Driving directions based on taxi trajectories. In: Agrawala D, eds. *Proc. of the GIS 2010*. New York: ACM, 2010. 99–108.
- [10] Yuan J, Zheng Y, Xie X, Sun GZ. T-Drive: Enhancing driving directions with taxi drivers' intelligence. *IEEE Trans. on Knowledge and Data Engineering*, 2013,25(1):220–232.
- [11] Liu SY, Liu YH, Ni LM, Fan JP, Li ML. Towards mobility-based clustering. In: Rao B, eds. *Proc. of the KDD'11*. New York: ACM, 2010. 919–928.
- [12] Zhang DQ, Li N, Zhou ZH, Chen C, Sun L, Li SJ. iBAT: Detecting anomalous taxi trajectories from GPS traces. In: Landay J, ed. *Proc. of the Ubicomp 2011*. New York: ACM, 2011. 99–108.
- [13] Zheng Y, Liu YC, Yuan J, Xie X. Urban computing with taxicabs. In: Landay J, ed. *Proc. of the Ubicomp'11*. New York: ACM, 2011. 89–98.
- [14] Pan G, Qi GD, Zhang WS, Li SJ, Wu ZH. Trace analysis and mining for smart cities: Issues, methods, and applications. *IEEE Communications Magazine*, 2013,51(6):120–126.
- [15] Deng CC, Ong HL, Ang BW, Goh TN. A modelling study of a taxi service operation. *Int'l Journal of Operations and Production Management*, 1992,12(11):65–78.
- [16] Yang H, Lau YW, Wong SC, Lo HK. A macroscopic taxi model for passenger demand, taxi utilization and level of services. *Transportation*, 2000,27(3):317–340.
- [17] Yang H, Song SC, Wong KI. Demand-Supply equilibrium of taxi services in a network under competition and regulation. *Transportation Research Part B: Methodological*, 2002,36(9):799–819.
- [18] Qi GD, Li XL, Li SJ, Pan G, Wang ZH, Zhang DQ. Measuring social functions of city regions from large-scale taxi behaviors. In: Cook D, ed. *Proc. of the PERCOM Workshops*. New York: IEEE, 2011. 384–388.
- [19] Marguier PHJ, Ceder A. Passenger waiting strategies for overlapping bus routes. *Transportation Science*, 1984,18(3):207–230.
- [20] Siikonen ML. Elevator traffic simulation. *Simulation*, 1993,61(4):257–267.



齐观德(1987—),男,江西婺源人,博士生,主要研究领域为普适计算,数据挖掘与机器学习应用.

E-mail: qiguande@163.com



潘遥(1989—),男,博士生,主要研究领域为普适计算,分布式系统.

E-mail: yaopao@zju.edu.cn



李石坚(1979—),男,博士,副教授,主要研究领域为普适计算,移动互联网.

E-mail: shijianli@zju.edu.cn



潘纲(1976—),男,博士,教授,主要研究领域为普适计算,计算机视觉,智能系统.

E-mail: gpan@zju.edu.cn

www.jos.org.cn

www.jos.org.cn