

DOI: 10.11992/tis.201706071

网络出版地址: <http://kns.cnki.net/kcms/detail/23.1538.TP.20171109.1255.026.html>

# 出租车数据的城市道路网路段通行时间估计方法

黄顺伦, 杜春, 宋宝泉, 李军, 陈浩

(国防科技大学 电子科学与工程学院, 湖南 长沙 410073)

**摘要:** 城市路段通行时间估计能够更好地运营和管理城市交通。针对包含起点-终点位置, 行程时间和距离信息的GPS行程数据, 提出了一种城市道路网短时通行时间的估计模型。首先将城市道路网按照交叉路口分解为多个路段, 并基于k-最短路径搜索方法分析司机行进路线。然后针对每一个路段, 提出了双车道通行时间多项式关联关系模型, 既能提升道路网通行时间精细度, 又能避免因训练数据不足导致的路网通行时间过拟合问题。最后以最小化行程期望时间和实际行程时间之间的均方误差为优化目标, 拟合道路网通行时间。在纽约出租车数据集上的实验结果表明, 所提模型及方法相对于传统单车道估计方法能够更准确地估计城市道路网路段的通行时间。

**关键词:** 通行时间估计; GPS-出租车; 城市道路网; 双车道模型

**中图分类号:** TP311 **文献标志码:** A **文章编号:** 1673-4785(2017)06-0790-09

中文引用格式: 黄顺伦, 杜春, 宋宝泉, 等. 出租车数据的城市道路网路段通行时间估计方法[J]. 智能系统学报, 2017, 12(6): 790-798.

英文引用格式: HUANG Shunlun, DU Chun, SONG Baoquan, et al. Urban link travel time estimation using taxi data[J]. CAAI transactions on intelligent systems, 2017, 12(6): 790-798.

## Urban link travel time estimation using taxi data

HUANG Shunlun, DU Chun, SONG Baoquan, LI Jun, CHEN Hao

(School of Electronic Science and Engineering, National University of Defense Technology, Changsha 410073, China)

**Abstract:** The accurate estimation of urban link travel time plays a significant role in urban traffic monitoring and supervision. Using taxicab GPS trip data, which contains origin and destination locations, travel time, and distances, this paper establishes a model to estimate average short-term urban link travel times. Firstly, the urban road network is divided into many segments based on crossings, and the running route of the driver was analyzed using the k-shortest path search algorithm. Then, for each road segment, a polynomial incidence relation model of the travel time in double lanes is proposed; this increases precision and avoids the overfitting of the travel time of the road network caused by insufficient training data. Finally, by minimizing the mean square error between the expected path travel time and the observed path travel time as the optimization objective, the travel time of the road network is fitted. The results of experiments conducted on New York taxi datasets show that, relative to the traditional single-lane estimation method, the proposed model and method more efficiently estimate the travel time of the road segments in urban road networks.

**Keywords:** travel time estimation; GPS-enabled taxicab; urban road networks; two-lane model

城市道路通行时间的准确估计和预测对于改善城市交通状况是至关重要的, 其目标在于计算准确的道路网通行时间信息, 以便选择道路网中更好的路线使通行时间最小。若欲准确评估路段通行时间, 最核心的就是从道路传感器中获取良好的车辆实时信息。然而, 在大多数情况下, 只能获得离散

的车辆速度和位置信息, 具体的时空轨迹信息难以实时获取。因此, 必须开发适当的方法来估计道路网路段通行时间。

目前, 针对城市道路通行时间估计和预测的研究主要包括两类方法, 即基于传感器数据的预测以及基于城市全球定位系统(global positioning system, GPS)数据的预测方法。

第一类方法的研究主要依赖于各种类型传感器采集的数据进行预测, 主要包括: 环形线圈检测

收稿日期: 2017-06-22. 网络出版日期: 2017-11-09.

基金项目: 国家“863”计划项目(2015AA123901).

通信作者: 陈浩. E-mail: [hchen@nudt.edu.cn](mailto:hchen@nudt.edu.cn).

器<sup>[1-3]</sup>、自动车辆识别<sup>[4-5]</sup>、摄像机、远程通信微波传感器<sup>[6]</sup>和自动化牌照识别<sup>[7]</sup>等。通常,这些数据需要相应精度级别的传感器来获取。然而由于传感器的安装和维护费用高昂,导致基于传感器获取数据的道路通行时间预测方法的应用难以普及。

第二类方法是基于GPS数据估计城市地区交通动态需求和道路网动态变化,因其具有极大的应用价值而引起国内外学者的广泛关注。通过车辆或移动手机中的GPS设备获取的数据可以成为监测城市交通量的可行来源<sup>[8]</sup>。随着从车辆和手机中获得的GPS数据越来越多,基于这些大规模分散数据估计路段通行时间已变为现实。因为手机数据涉及个人隐私等问题,大量研究主要以车辆GPS数据为主。Zhan等<sup>[9]</sup>基于轨迹数据估计城市交通流量;Zheng等<sup>[10]</sup>基于稀疏车辆数据提出了估计城市路段通行时间的ANN模型;Hunter等<sup>[11]</sup>利用GPS车辆数据统计路段通行时间;Herring等<sup>[12]</sup>认为出租车相比普通车辆在城市中具有更高渗透率,利用出租车GPS数据可以更好表现道路网情况,因此他们基于500辆出租车的GPS数据,估计和预测了旧金山城市范围内离散的交通状况。然而,所有上述方法仅适用于GPS轨迹数据。但是,现实世界中大量出租车GPS数据仅含起点终点(origin-destination, OD)信息,如纽约公布的出租车行驶数据集<sup>[13]</sup>等。由于全球定位系统起点终点(GPS-OD)数据中,仅包含出租车一次运营过程的起点和终点,而不包含本次运营的路线信息,于是基于GPS-OD数据进行路网通行时间估计,不仅需要拟合路段时间,还需要分析出租车运行路线,给城市道路网路段通行时间估计问题带来了新的挑战。Zhan<sup>[14-15]</sup>等利用纽约出租车GPS-OD数据,估计道路网路段通行时间。但是,他只考虑了道路单车道对车辆行驶的影响,当路段较宽时,车道数可能更多,单车道不能很好地刻画道路网精细化程度。

为了克服上述的问题,本文基于出租车数据提出了一种城市道路网路段通行时间估计方法,主要贡献在于:

1)建立了基于出租车GPS-OD数据集的双车道道路网通行时间估计模型。假设道路网每条路段为双车道,能够更准确地描述道路网通行情况,为了避免训练数据量不足而导致的过拟合问题,建立了双车道间通行时间多项式关联关系模型。

2)采用优化非线性最小二乘方法估计路段每小时平均通行时间,从而实现路网通行时间拟合。

3)设计多组实验,分析双车道通行时间之间不同多项式关系对道路网路段通行时间估计结果的影响,确定效果最优的多项式关系。通过多组估计不

同时段路段通行时间的实验,验证了本文所提双车道预测方法相比于单车道方法能够更准确地估计道路网路段的通行时间。

## 1 道路网路段通行时间估计模型

本节将介绍估计道路网路段通行时间模型,该模型的总体框架如图1所示,主要包括以下4个步骤:1)地图匹配。GPS数据中的起点和终点映射到道路网中,以减小GPS误差带来的影响,将原始数据转化为可用数据。2)路径选择。采用k-最短路径算法<sup>[16]</sup>,构建每段行程的路径集合,并根据MNL(multinomial logit model)<sup>[17]</sup>模型计算司机选择不同路径的可能性,最后筛选合理的路径集合作为估计路段时间过程的基础。3)双车道通行时间模型构建。为精细刻画道路网通行时间程度,提出双车道通行时间多项式关联关系模型。4)路段通行时间估计。将步骤2)中筛选出的多条路径作为出租车某次行程的可能发生事件以计算每次行程的期望时间,最后将路段通行时间估计问题转换为行程观测时间与期望时间均方误差最小问题。

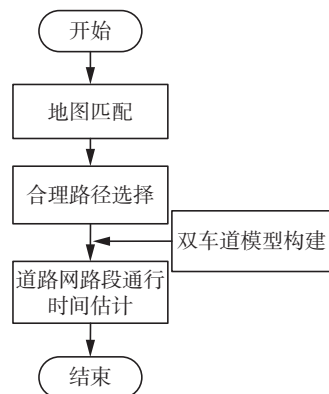


图1 模型总体框架

Fig. 1 General framework for model

下面将详细介绍地图匹配、路径选择模型、路段双车道通行时间多项式关联关系模型,道路网路段通行时间估计方法。

### 1.1 地图匹配

GPS数据因接收设备老化,信号传播延迟等原因存在一定定位误差,需要预先对原始GPS数据进行地图匹配,其具体作用将起点和终点映射到道路网中,将原始数据转换为可用数据,便于道路网分析。

图2说明了数据地图匹配过程,其中端点 $(A_1, A_2, B_1, B_2)$ 为路径相交节点,首先将原始起点和终点 $(A, B)$ 匹配到最近路段的垂足上 $(A', B')$ ,匹配后的点的位置用路段两个端点 $(A_1$ 和 $A_2, B_1$ 和 $B_2)$ 和 $\alpha_1, \alpha_2 (\alpha_1 = d'_k/d_k, \alpha_2 = d'_l/d_l)$ 表示。对于位于单向街道的起点和终点,两个端点在给定路段的方向信息情

况下很容易被识别。对于位于双向街道上的点,这个路段的两个端点都可用,不同端点组合的 $\alpha_1$ 、 $\alpha_2$ 也可能是同样的记录。

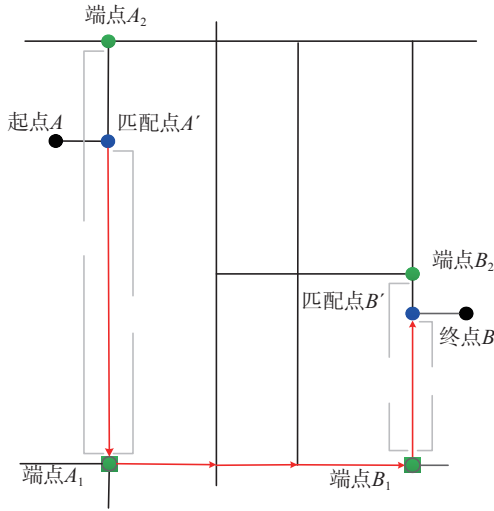


图2 地图匹配示例

Fig. 2 Illustration of data mapping

## 1.2 路径选择模型

将 GPS-OD 数据进行地图匹配后,得到了道路网中每次行程的起点终点数据。由于路径选择信息的缺失,在估计城市道路网时间时需要推断实际的路径。但在城市的道路网中,对于某一确定的出租车行程,所有路径的集合是非常大的。考虑到交通网中观测数目过于庞大,对整个空间进行路径搜索非常耗时,减少路径数目是很有必要的。这里,采用 k-最短路径搜索算法生成最初的路径集合,然后利用数据中记录的行程距离来排除不合理的路径。

计算每段行程的可选路径集合后,由于缺少司机社会和行为特征,不能用传统的计量经济学模型来估计司机所选择的路径。因为司机做出决策之前不可能知道实际的路径时间。但是,他们可以通过经验推测道路网通行时间,因此本文基于 MNL 模型利用可选路径集合中的路径时间和距离表示路径成本  $C_m$  求不同路径的选择概率。为降低复杂性,定义路径选择模型为

$$P_m(t, d, \theta) = \frac{\exp[-\theta C_m(t, d_m)]}{\sum_{j \in R_i} \exp[-\theta C_j(t, d_j)]} \quad (1)$$

式中:  $P_m$  表示某一行程中可选路径  $m$  的选择概率与道路网路段通行时间  $t$ 、行程中各个可选路径的距离  $d$ 、参数  $\theta$  有关。  $C_m$  表示路径  $m$  的成本与道路网路段通行时间  $t$  和路径距离  $d_m$  有关,参数  $\theta$  用于表示司机感知不同时间段道路网通行时间不同时的路径成本变化,  $\theta$  大表示感知错误小,司机倾向于选择成本小的路径,而  $\theta$  小意味着感知错误较大,成本越大的路径越有可能被选择。在此模型中,  $\theta$  和道路网通行时间都是待估参数。

假设每个司机在同一起点终点的行程下,更偏好选择行程时间和距离更短的路径,那他们就能够行驶更多行程数量,获得更多收益。在建立合理的路径集合时,设置阈值用于排除违反上述选择行为假设的路径。路径距离在行程观测距离一定比例内的才会被使用。因为数据中记录的行程距离不精确(只到 160 m),设置工作日行程距离阈值为 15%~25%,周末为 20%~25%,消除那些偏离记录中行程距离太多的不合理路径。阈值设定取决于一个小时内可用的行程数据量。

根据城市出租车计价规则:开始行程收取基本费用,超过基本乘车距离和时间,按相应比例收取叠加费用。考虑到实际情况票价计算的复杂,采用行程时间和距离的线性模型表示行程成本,如式(2)所示。

$$\text{fare} = \beta_0 + \beta_1 \cdot \text{time} + \beta_2 \cdot \text{distance} \quad (2)$$

式中: fare 表示行程成本,  $\beta_0$  为常数,  $\beta_1$ 、 $\beta_2$  是行程时间和距离的成本系数。根据文献[14],  $\beta_1$ 、 $\beta_2$  的估计值为 0.275/min 和 2.516/km。

$$C_m(t, d_m) = \beta_1 \cdot g_m(t) + \beta_2 \cdot d_m \quad (3)$$

式中:  $d_m$  是路径  $m$  的距离,路径  $m$  的通行时间  $g_m(t)$  定义为

$$g_m(t) = a_1 t_0 + a_2 t_D + \sum_{l \in L} t_l \quad (4)$$

式中:  $t_0$  是起点所在路段的通行时间,  $t_D$  是终点所在路段的通行时间,  $L$  是道路网的路段集合,  $t_l$  是路段  $l$  的通行时间,  $\delta_{ml}$  是路径与路段的关系值,取值为 0、1, 1 表示路径  $m$  经过路段  $l$ , 0 则相反,  $\alpha_1$ 、 $\alpha_2$  是距离比例。

## 1.3 路段双车道通行时间之间多项式关联关系模型

在城市道路网中许多道路分为左侧车道、直行车道和右侧车道。左侧和直行车道在行驶过程中会出现等待红绿灯的情况,右侧车道则可以直接通行。若只考虑单车道情况,将会忽略左侧和直行车道上等待红绿灯的时间。若加入多车道,但不考虑车道间关系,可能导致待估变量数目太多,拟合效果较差,或样本数不足的情况。假定同一路段上左侧车道和直行车道上的车辆通行时间服从相似分布,并根据路段车道间车流量会相互影响的实际情况,我们认为路段上为双车道,且存在一定的多项式关系,如式(5)所示。

$$y = \sum_{i=2}^k \gamma_i x^i + \gamma_1 x + \gamma_0, k = 2, 3, \dots \quad (5)$$

式中:  $x$  是路段上一条车道的通行时间,  $y$  表示与  $x$  相关的另一条车道的通行时间,多项式  $\gamma$  为待估参数。

## 1.4 道路网路段通行时间估计

道路网路段通行时间估计是最小化行程观测时间与期望时间之间的均方差,将出租车实际路径选择作为隐含变量,路段通行时间  $t$ 、双车道之间的多



项式关系参数  $\gamma$  和比例参数  $\theta$  作为待估参数, 观测  $i$  的期望时间  $E(Y_i|R_i)$  可写成

$$E(Y_i|R_i) = \sum_{m \in R_i} g_m(t, \gamma) P_m(t, \gamma, d, \theta) \quad (6)$$

式中:  $Y_i$  是观测  $i$  的时间变量,  $R_i$  是根据观测  $i$  的 OD 行程记录建立的可能路径集,  $t$  是道路网路段通行时间向量,  $\gamma$  是双车道间多项式关系参数,  $d$  是  $R_i$  的所有路径距离,  $g_m(t, \gamma)$  是路径  $m$  的行程时间,  $P_m(*)$  是选择路径  $m$  的可能性,  $\theta$  是比例参数。

对于一条路径, 其距离是确定的, 道路网路段通行时间向量  $t$ , 双车道间多项式关系参数  $\gamma$  和比例参数  $\theta$  是待估参数, 那么  $E(Y_i|R_i)$  可以表示为一个与  $R_i, t, \gamma, \theta$  有关的函数:

$$E(Y_i|R_i) = f(R_i, t, \gamma, \theta) \quad (7)$$

进一步, 行程观测时间  $y_i$  与行程期望时间  $E(Y_i|R_i)$  之间的误差可以定义为

$$r_i = y_i - f(R_i, t, \gamma, \theta) \quad (8)$$

则误差平方定义为

$$S(t, \gamma, \theta) = \sum_{i \in D} r_i^2 = \sum_{i \in D} (y_i - f(R_i, t, \gamma, \theta))^2 \quad (9)$$

由此, 所估计的道路网路段通行时间为

$$t = \arg \min_t S(t, \gamma, \theta) \quad (10)$$

## 2 道路网路段通行时间求解

利用 Levenberg-Marquardt(LM)<sup>[18]</sup>方法解决非线性最小二乘问题。该方法是一种广泛用于求解最小二乘拟合和非线性规划问题的优化算法。在各种问题上, 它优于一般的梯度下降方法和著名的高斯-牛顿 (GN) 方法<sup>[18]</sup>。传统的高斯-牛顿法是计算代价高的线性搜索法。更新的高斯-牛顿法类似于牛顿法, 当近似的 Hessian 矩阵近似奇异时变成了数学问题。如果利用不恰当的初始值, 则容易不能收敛到最优。另一方面, Levenberg-Marquardt 方法利用信任域策略而不是线性搜索方法, 在更新步骤前确定步长。在 LM 中利用不同的 Hessian 近似方法也有助于确保每次迭代时矩阵的正定性, 具有更好的鲁棒性, 这意味着在许多情况下, 即使初始值远离最终优值, Levenberg-Marquardt 法也能找到一个近似解。在 Bonnans 和 Gilbert<sup>[19]</sup>中表明 Levenberg-Marquardt 具有快速局部收敛性能。

本文中, 利用 LM 算法求解的目标函数为行程期望时间:

$$E(Y_i|R_i) = f(R_i, t, \gamma, \theta) = \sum_{m \in R_i} g_m(t, \gamma) \frac{P_m(t, \gamma, \theta)}{S_{R_i}(t, \gamma, \theta)} \quad (11)$$

简单起见, 定义

$$P_m(t, \gamma, \theta) = \exp(\theta(-\beta_1 \cdot g_m(t, \gamma) - \beta_2 \cdot d_m)) \quad (12)$$

式 (1) 的分母写成

$$S_{R_i}(t, \gamma, \theta) = \sum_{j \in R_i} \exp(\theta(\beta_1 \cdot g_m(t, \gamma) - \beta_2 \cdot d_m)) \quad (13)$$

求解 Jacobian 矩阵为

$$J_{il} = \frac{\partial f(R_i, t, \gamma, \theta)}{\partial t_l}, l = 1, 2, \dots, N \quad (14)$$

$$J_{ip} = \frac{\partial f(R_i, t, \gamma, \theta)}{\partial \gamma_p}, p = N+1, \dots, N+k+1 \quad (15)$$

$$J_{iq} = \frac{\partial f(R_i, t, \gamma, \theta)}{\partial \theta}, q = N+k+2 \quad (16)$$

Jacobian 矩阵是一个  $N_D * (N+k+2)$  的矩阵, 其中  $N_D$  是数据集中行程观测数目,  $N$  是道路网中路段数目,  $k$  是双车道间多项式关系的阶次。

路段通行时间  $t$ , 车道间多项式关系参数  $\gamma$ , 比例参数  $\theta$  在第  $v$  次迭代更新为

$$\begin{aligned} t &\approx t^{v+1} = t^v + u_t^{(v)} \\ \gamma &\approx \gamma^{v+1} = \gamma^v + u_\gamma^{(v)} \\ \theta &\approx \theta^{v+1} = \theta^v + u_\theta^{(v)} \end{aligned} \quad (17)$$

得到待估参数在第  $v$  次迭代求解增量正规方程时的更新方向为  $u^{(v)} = (u_t^{(v)}, u_\gamma^{(v)}, u_\theta^{(v)})$ , 增量正规方程为

$$(J^{(v)T} J^{(v)} + \lambda I) u^{(v)} = J^{(v)T} r_i \quad (18)$$

当迭代结束后, 式 (17) 所得  $t$  值即为所估计的道路网路段通行时间,  $\gamma$  为双车道间多项式关系参数,  $\theta$  为表示司机对道路网感知程度的比例参数。

通过分析可以发现, 上述函数非凸, 可能有多多个局部最优点。在具体求解时, 考虑将初始值默认为整个道路网当前时段下的平均速度, 能较快较好地收敛到合适的最优的值。

## 3 实验结果与分析

我们采用纽约出租车行程数据集, 数据由城市出租车豪华轿车委员会 (New York City Taxi and Limousine Commission, NYTLC) 收集。其特点是每个出租车都安装了 GPS 设备采集数据。纽约有北美最大的出租车市场, 12 779(2006 年) 辆黄色纪念章出租车每年大约服务 2.4 亿人次。在曼哈顿, 乘坐出租车人数是所有出行人数的 25%<sup>[20]</sup>。数据集包含 2010—2015 年出租车行程数据, 其中包括行程开始和结束 (OD 数据) 的地理位置、行程距离、时间和票价等信息, 而缺少出租车的确切轨迹。但是, 大量的数据 (一天 450 000 ~ 550 000 的记录数量) 可以推断出租车可能路线, 并进一步估计道路网的路段通行时间。

基于 Python 语言编程实现前面部分讨论的模型。硬件配置为 i5 处理器, 3.2 GHz CPU, 4 GB 内存。在实验中利用均方根误差 (root mean square error, RMSE) 和平均绝对百分比误差 (mean absolute percentage error, MAPE) 来评估估计结果:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (T_i^{Pr} - T_i^{Ob})^2} \quad (19)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{T_i^{Pr} - T_i^{Ob}}{T_i^{Ob}} \right| \times 100\% \quad (20)$$

式中:  $T_i^{\text{Pr}}$  是模型估计的通行时间,  $T_i^{\text{Ob}}$  是实际观测的通行时间,  $n$  是通行观测数目。

### 3.1 测试数据和道路网

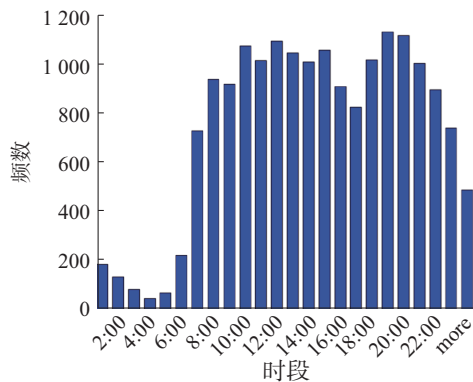
实验中利用纽约出租车两周 (3/2/2015—3/15/2015) 的 OD 行程数据测试所提出的方法。实验区域位于曼哈顿中央公园东南部一块  $1\,508\text{ m}^2$  的范围, 相关道路网如图 3 所示, 包含 208 个节点和 386 条边。道路网中有 348 条道路是单向街道, 38 条是双向街道。图 4、5 分别展示了在该范围内工作日 (3/2/2015 和 3/9/2015, 周一) 和周末 (3/7/2015 和 3/14/2015, 周六) 的行程频数。通过统计和观察图 4、5 可以发现, 该区域内工作日 (周一) 一小时内将近 1 200 条行程数, 周六大约 1 000 条的行程数。且每周同一天行程观测数近似服从同一分布。



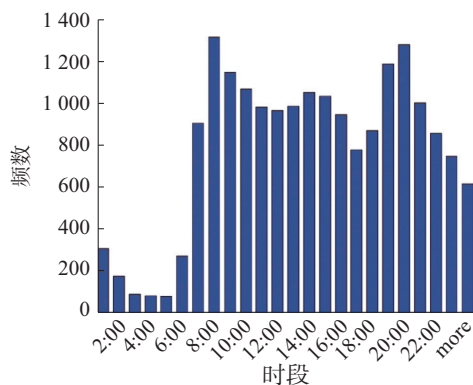
(a) 曼哈顿市中心地图 (b) 曼哈顿道路网图

图 3 研究区域测试道路网:曼哈顿市中心

Fig. 3 Test network of study region: midtown Manhattan



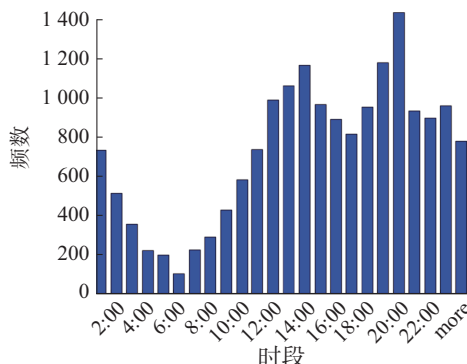
(a) 研究区域日期为 2015.3.2



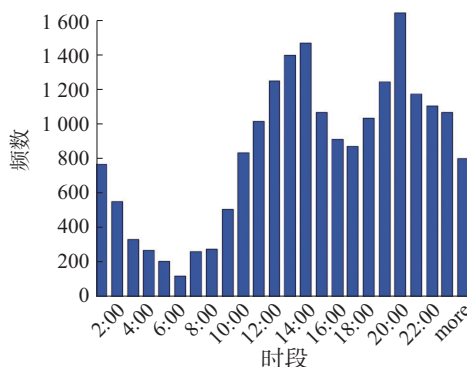
(b) 研究区域日期为 2015.3.9

图 4 研究区域内周一每小时观测数目直方图

Fig. 4 Histogram for number of hourly observations in the study region on Monday



(a) 研究区域日期为 2015.3.7



(b) 研究区域日期为 2015.3.14

图 5 研究区域内周六每小时观测数目直方图

Fig. 5 Histogram for number of hourly observations in the study region on Saturday

若实验数据以分钟为单位采样, 行程数和信息量太少不能保证良好的统计意义。若以天为单位采样, 不具有良好的代表性和研究意义。因此实验以小时为单位采样, 从相应的数据中估计道路网通行时间。

### 3.2 结果与分析

为了验证提出的算法性能, 在实验中引入了 Zhan<sup>[14]</sup>提出的单车道道路通行时间估计模型进行比较。设计了两组实验, 第一组实验分析单车道与双车道间不同多项式关系对估计道路网路段通行时间结果的影响, 并确定效果最优的多项式关系, 第二组为不同时段估计路段通行时间的实验。

#### 3.2.1 双车道间通行时间多项式关联关系模型下的估计误差实验

为了验证 2.3 节所提双车道模型的有效性, 以 3/2/2015—3/15/2015 中 9:00—10:00 为研究时段分别计算不同车道关系下的模型估计误差, 实验结果如表 1 所示。通过观察可以发现, 双车道模型整体上 RMSE 和 MAPE 相比 Zhan 等<sup>[14]</sup>提出的单车道方法要低: 当转换模型为二阶、三阶、四阶多项式时, 一周中有半数以上的时间段双车道模型误差低于单车道模型; 当转换模型为五阶、六阶多项式时, 一周中

的 RMSE 和 MAPE 全都低于单车道模型结果。当双车道通行时间之间多现实转换模型取为六阶多项式时,周一到周六实验所得的 RMSE 和 MAPE 分别比单车道低 3.45, 5.33, 0.21, 0.13, 0.06, 0.41 和

39.8%, 11.7%, 1.75%, 2.6%, 1.5%, 5.4%。上述分析证明,高阶多项式的双车道模型能够更好地刻画道路网的精细化程度,相比单车道模型能够更准确地估计道路网通行时间。

表 1 不同车道关系下的模型估计误差

Table 1 Model estimation error in different lane conditions

时间	误差	两车道之间的关系					
		单车道 <sup>[14]</sup>	二阶多项式	三阶多项式	四阶多项式	五阶多项式	六阶多项式
周一	RMSE	6.82	4.45	3.69	4.18	4.12	3.37
	MAPE/%	40.00	36.50	31.40	35.20	34.70	30.20
周二	RMSE	9.15	4.29	4.3	4.27	3.86	3.82
	MAPE/%	44.00	34.80	37.40	34.50	32.70	32.30
周三	RMSE	3.90	3.72	3.91	3.81	3.74	3.69
	MAPE/%	32.65	31.60	32.60	31.60	31.40	30.90
周四	RMSE	3.65	3.65	3.63	3.63	3.64	3.52
	MAPE/%	33.90	33.10	33.00	33.00	33.00	31.30
周五	RMSE	3.42	3.33	3.29	3.09	2.94	3.36
	MAPE/%	32.10	30.90	30.60	106.0	105.0	30.60
周六	RMSE	2.71	3.09	2.31	3.09	2.38	2.30
	MAPE/%	37.70	37.70	31.90	37.90	33.40	32.30

### 3.2.2 单车道模型与双车道六阶多项式关联关系模型估计道路网通行时间的实验

该实验分为训练和测试两个阶段,首先基于第3节训练求出道路网路段通行时间 $t$ ,双车道间六阶多项式参数 $\gamma$ 、 $\theta$  3个值,然后在测试阶段输入新的OD行程记录,并根据式(6)求出该行程记录的期望时间和误差。

基于两周的GPS数据(3/2/2010-3/15/2010),其

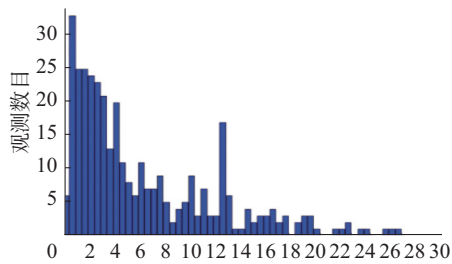
中每周同一天数据的80%作为训练样本,剩下20%第2周的数据作为测试样本,估计一天中4个时间段(9:00—10:00, 13:00—14:00, 19:00—20:00, 21:00—22:00)的道路网路段通行时间。分别采用Zhan<sup>[14]</sup>单车道模型和本文双车道为六阶多项式关系的模型对道路路段通行时间进行估计,通过观察表2可见,双车道模型在更多数据情况下结果都优于单车道模型。

表 2 模型估计误差

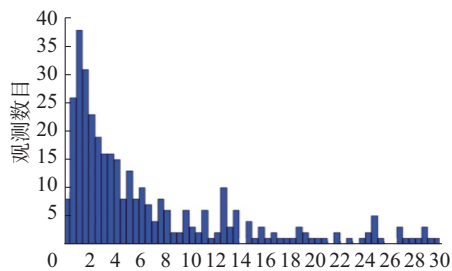
Table 2 Model estimation error

日期	误差	时间段							
		9:00—10:00		13:00—14:00		19:00—20:00		21:00—22:00	
		单车道	双车道六阶多项式	单车道	双车道六阶多项式	单车道	双车道六阶多项式	单车道	双车道六阶多项式
周一	RMSE/min	6.82	3.37	4.97	4.88	3.95	3.14	3.90	3.58
	MAPE/%	40.00	30.20	47.10	38.90	39.10	36.00	38.20	35.20
周二	RMSE/min	9.15	3.82	5.50	4.76	5.34	5.60	3.52	2.87
	MAPE/%	44.00	32.30	44.50	37.40	50.80	45.90	36.10	30.20
周三	RMSE/min	3.90	3.69	4.75	4.41	5.46	5.31	35.80	4.13
	MAPE/%	32.65	30.90	41.34	39.50	45.80	40.80	22.00	135.00
周四	RMSE/min	3.65	3.52	4.79	4.05	5.37	2.62	3.93	2.56
	MAPE/%	33.90	31.30	39.50	34.50	44.90	29.60	43.50	31.50
周五	RMSE/min	3.42	3.36	4.00	3.54	5.49	3.80	7.15	2.42
	MAPE/%	32.10	30.60	37.90	31.40	43.40	32.00	52.60	31.60
周六	RMSE/min	2.71	2.30	3.57	3.34	10.70	3.88	7.50	4.48
	MAPE/%	37.70	32.30	38.80	36.30	50.90	33.00	51.00	39.30
周日	RMSE/min	2.93	2.38	17.20	4.27	3.30	2.79	12.90	2.92
	MAPE/%	38.50	33.40	80.40	37.10	46.10	36.30	62.80	34.30

除了时间段(3月9日周三 21:00—22:00), 双车道模型路段通行时间估计结果的 MAPE 低于 40%, 可以观察到单车道模型和双车道模型在周三 21:00—22:00 误差值最大。可发现周三(3/11/2015)有纽约洋基对战巴尔的摩金莺的橄榄球比赛, 比赛结束后可能导致大量拥堵以及密集人群流动, 该事件可能与误差结果有较大关系。且双车道模型遇到异常情况时, 效果更加稳健, 其结果比单车道模型低 94%。

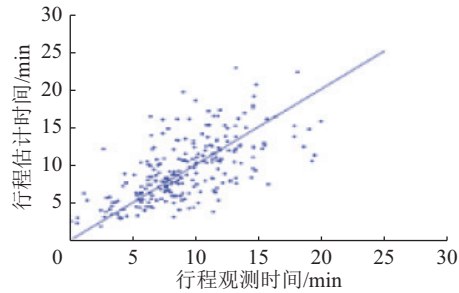


(a) 周一 9:00—10:00

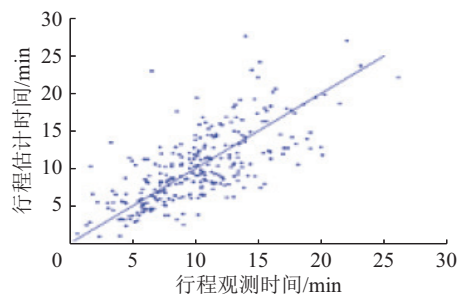


(c) 周二 9:00—10:00

我们用道路网路段通行速度而不是道路网路段通行时间直观表示估计结果, 图 6 表示周一、周二 9:00—10:00, 路段估计时间直方图和行程观测时间与估计时间之间的关系图, 其中  $X$  轴表示路段通行速度,  $Y$  轴表示该速度的路段数目。子图中  $X$  轴为行程观测时间,  $Y$  轴为模型估计时间。图 7 表示周三, 周六 13:00—14:00 的关系。其他时间段的关系与其相似, 不多作赘述。



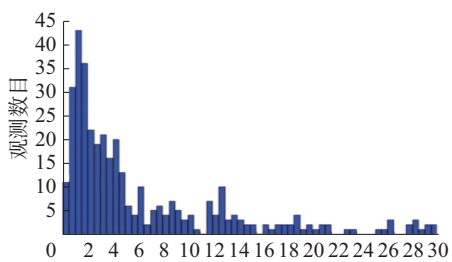
(b) 周一行程观测时间与估计时间关系



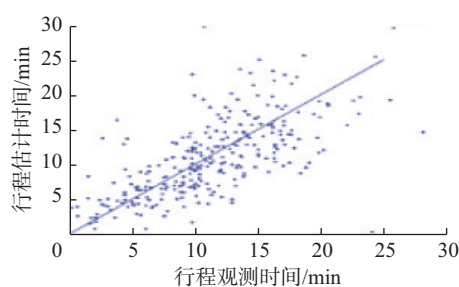
(d) 周二行程观测时间与估计时间关系

图 6 周一、周二路段估计时间直方图和行程观测时间与估计时间之间的关系

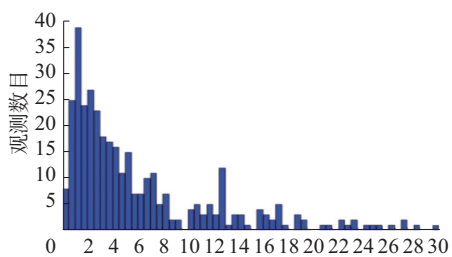
Fig. 6 Histogram of estimated link speed and correlation plot of observed and estimated path travel time for Monday, Tuesday



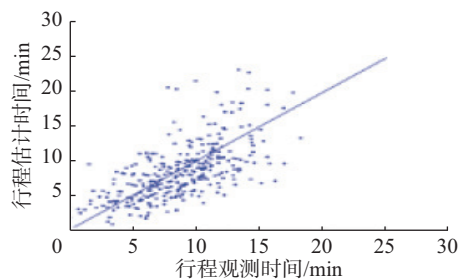
(a) 周三 13:00—14:00



(b) 周三行程观测时间与估计时间关系



(c) 周六 13:00—14:00



(d) 周六行程观测时间与估计时间关系

图 7 周三, 周六路段估计时间直方图和通行观测时间与估计时间之间的关系

Fig. 7 Histogram of estimated link speed and correlation plot of observed and estimated path travel time for Wednesday, Saturday



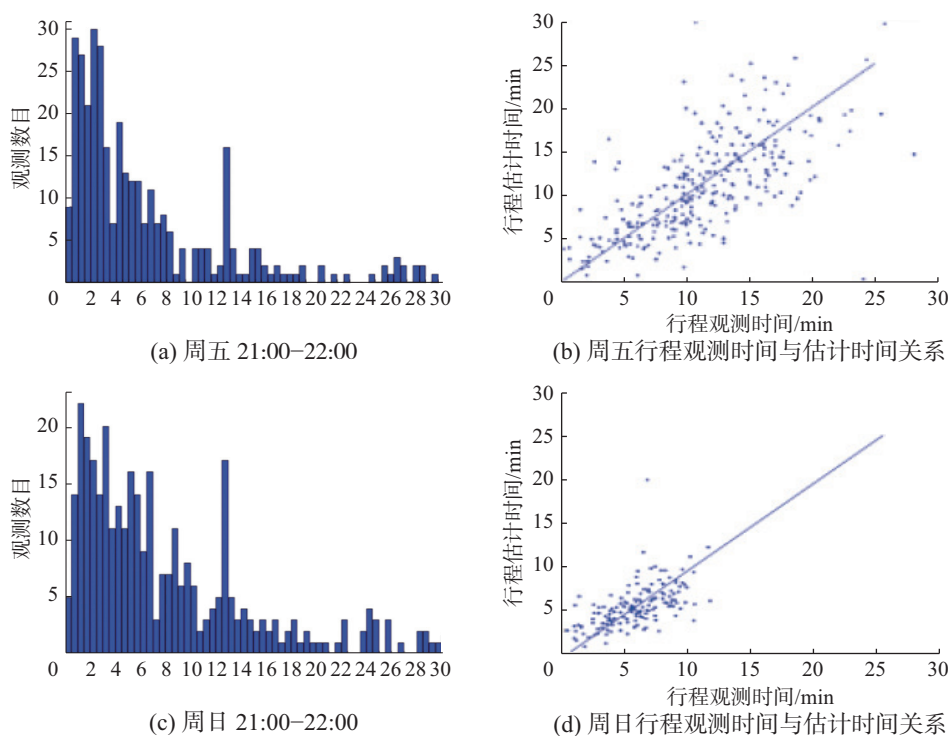


图8 周五、周日路段估计时间直方图和通行观测时间与估计时间之间的关系

Fig. 8 Histogram of estimated link speed and correlation plot of observed and estimated path travel time for Friday, Sunday

由于路段通行时间是以小时为单位估计的,所以一小时内道路的变化也会导致模型的误差 (Fosgerau 和 Fukuda<sup>[21]</sup>)。司机之间的选择偏好 (例如,一些司机驾驶速度快,偏好选择短路径,一些司机驾驶速度慢,偏好采取相对较长的路径等)也可能导致误差。观察到某些行程在测试的道路网中长达 20 min,这使得在路径选择中有很多不确定性,从而导致了一些误差。

## 4 结束语

本文提出了一种基于出租车 GPS-OD 数据来估计城市道路网通行时间的新模型。为了更精细地刻画道路网,该模型基于双车道估计行程期望时间,为了避免训练数据量不足而导致的过拟合问题,建立了双车道间通行时间多项式关联关系模型,并通过最小化行程期望时间和行程观测时间之间的误差来估计道路网通行时间。实验结果表明本文提出的方法能够有效地估计道路网每小时通行时间。为充分利用城市出租车数据估计道路网时间提供新的可能性。在下一步工作中,我们将利用 GPS 数据进一步研究城市交通流量的估计问题。

## 参考文献:

[1] COIFMAN B, CASSIDY M. Vehicle reidentification and travel time measurement on congested freeways[J]. Transportation research part a: policy and practice, 2002, 36(10):

899-917.

[2] ZHANG X, ZHANG B, LIU L, et al. Estimating foliar nitrogen concentration with hyperspectral remote sensing image[C]//Third International Asia-Pacific Environmental Remote Sensing Remote Sensing of the Atmosphere, Ocean, Environment, and Space. Beijing, 2003: 187-193.

[3] WU C C, LEE W M G. Control of vaporous naphthalene by scrubbing with surfactants[J]. Journal of environmental engineering, 2004, 130(3): 276-281.

[4] PARK D, RILETT L. Forecasting multiple-period freeway link travel times using modular neural networks[J]. Transportation research record: journal of the transportation research board, 1998(1617): 163-170.

[5] LI R, ROSE G. Incorporating uncertainty into short-term travel time predictions[J]. Transportation research part c: emerging technologies, 2011, 19(6): 1006-1018.

[6] YEON J, ELEFTERIADOU L, LAWPHONGPANICH S. Travel time estimation on a freeway using discrete time Markov chains[J]. Transportation research part B: methodological, 2008, 42(4): 325-338.

[7] HASAN S, CHOUDHURY C, BEN-AKIVA M, et al. Modeling of travel time variations on urban links in London[J]. Transportation research record: journal of the transportation research board, 2011(2260): 1-7.

[8] HERRERA J C, WORK D B, HERRING R, et al. Evaluation of traffic data obtained via GPS-enabled mobile phones: the mobile century field experiment[J]. Transporta-



- tion research part c: emerging technologies, 2010, 18(4): 568–583.
- [9] ZHAN X, ZHENG Y, Yi X, et al. Citywide traffic volume estimation using trajectory data[J]. IEEE transactions on knowledge and data engineering, 2017, 29(2): 272–285.
- [10] ZHENG F, VAN ZUYLEN H. Urban link travel time estimation based on sparse probe vehicle data[J]. Transportation research part c: emerging technologies, 2013, 31(0): 145–157.
- [11] HUNTER T, HERRING R, ABBEEL P, et al. Path and travel time inference from GPS probe vehicle data[J]. NIPS analyzing networks and learning with graphs, 2009, 12(1).
- [12] HERRING R, HOFLEITNER A, ABBEEL P, et al. Estimating arterial traffic conditions using sparse probe data [C]//13th International IEEE Conference on Intelligent Transportation Systems. Madeira Island, Portugal, 2010: 929–936.
- [13] Taxi data from new york taxi and limousine commission. [2016-05-14] <http://www.nyc.gov/html/tlc/html/home/home.shtml>.
- [14] ZHAN X, HASAN S, UKKUSURI S V, et al. Urban link travel time estimation using large-scale taxi data with partial information[J]. Transportation research part c: emerging technologies, 2013, 33(0): 37–49.
- [15] ZHAN X, UKKUSURI S V, YANG C. A Bayesian mixture model for short-term average link travel time estimation using large-scale limited information trip-based data[J]. Automation in construction, 2016, 72: 237–246.
- [16] YEN J Y. Finding the k shortest loopless paths in a network[J]. Management science, 1971, 17(11): 712–716.
- [17] DAGANZO C. Multinomial probit: the theory and its application to demand forecasting[M]. Elsevier, 2014.
- [18] CHEN Y, OLIVER D S. Levenberg–Marquardt forms of the iterative ensemble smoother for efficient history matching and uncertainty quantification[J]. Computational geosciences, 2013, 17(4): 689–703.
- [19] BONNANS J F, GILBERT J C, LEMARÉCHAL C, et al. Numerical optimization: theoretical and practical aspects [M]. Springer Science and Business Media, 2013.
- [20] KING D A, PETERS J R, DAUS M W. Taxicabs for improved urban mobility: are we missing an opportunity [C]//Transportation Research Board 91st Annual Meeting. Washington DC, USA, 2012 (12-2097).
- [21] FOSGERAU M, FUKUDA D. Valuing travel time variability: Characteristics of the travel time distribution on an urban road[J]. Transportation research part c: emerging technologies, 2012, 24(0): 83–101.
- [22] 孙锋, 黄玲, 叶盈, 等. 混行条件下直线式公交站点停靠车辆数优化[J]. 哈尔滨工程大学学报, 2015, 36(2): 152–155.
- SUN Feng, HUANG Ling, YE Ying, et al. Optimizing the number of buses stopping at on-line stops under mixed traffic conditions[J]. Journal of Harbin Engineering University, 2015, 36(2): 152–155.
- [23] 徐程, 曲昭伟, 陶鹏飞, 等. 动态交通数据异常值的实时筛选与恢复方法[J]. 哈尔滨工程大学学报, 2016, 37(2): 211–217.
- XU Cheng, QU Zhaowei, TAO Pengfei, et al. Methods of real-time screening and reconstruction for dynamic traffic abnormal data[J]. Journal of Harbin Engineering University, 2016, 37(2): 211–217.

#### 作者简介:



黄顺伦, 女, 1993 年生, 硕士研究生, 主要研究方向为交通数据分析、机器学习。



杜春, 男, 1983 年生, 讲师, 博士, 主要研究方向为机器学习、模式识别、图像处理。



宋宝泉, 男, 1980 年生, 讲师, 博士, 主要研究方向为大数据管理、智能信息处理。