计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(12):251−256 [doi: 10.15888/j.cnki.csa.006685]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于大规模 GPS 轨迹数据的出租车换道行为研究①

康　军[1,2], 温兴超[1], 段宗涛[1,2], 唐　蕾[1]

[1](长安大学 信息工程学院, 西安 710064)
[2](陕西省道路交通智能检测与装备工程技术研究中心, 西安 710064)
通讯作者: 温兴超, E-mail: wen.xingchao@foxmail.com

摘　要: 出租车换道行为的统计特性对研究经济、心理等人类动力学有重要的意义. 结合大数据分析技术, 基于西安市出租车 GPS 轨迹数据对出租车司机的换道行为进行了定量研究. 设计了一种基于出租车 GPS 轨迹数据的出租车司机换道行为识别模型, 利用大数据平台对出租车司机换道次数按不同时段进行了定量统计, 对出租车司机换道次数、出租车平均行驶速度和出租车司机的收入之间进行了相关性分析. 分析结果表明, 出租车频繁换道行为对司机收益呈现负相关影响, 进一步说明出租车司机驾驶习惯和和心理对整个出租车运营有显著影响.

关键词: 智能交通; 大数据; 出租车司机驾驶行为; GPS 轨迹数据; Spark

引用格式: 康军,温兴超,段宗涛,唐蕾.基于大规模 GPS 轨迹数据的出租车换道行为研究.计算机系统应用,2018,27(12):251–256. http://www.c-s-a.org.cn/1003-3254/6685.html

Study on Lane-Changing Behavior Based on Large Scale GPS Trajectory Data

KANG Jun[1,2], WEN Xing-Chao[1], DUAN Zong-Tao[1,2], TANG Lei[1]

[1](School of Information Engineering, Chang'an University, Xi' an 710064, China)
[2](Shaanxi Road Traffic Detection and Equipment Engineering Technology Research Center, Xi'an 710064, China)

Abstract: The statistical characteristics of taxi behavior have important significance to study the economic and psychological of human dynamics. Based on the taxi GPS track data in Xi'an, the lane-changing behavior was quantitatively studied by big data analysis technology. The model of a lane-changeing behavior recognition was designed, combined with the big data analysis technology, the number of taxi drivers' lane-changing was quantified in terms of different time periods, and the correlation analysis between taxi drivers' lane-changing times, taxi average driving speed, and taxi drivers' income was carried out. The results show that there is a significant negative correlation between the income of taxi drivers and the average driving speed of taxis, which further indicates that taxi drivers' habits and psychology have a significant impact on the whole taxi operation.

Key words: intelligent transportation; big data; taxi driver driving behavior; GPS trajectory data; Spark

近年来, 基于轨迹数据挖掘人类移动规律和兴趣爱好已成为研究人员的热点研究领域之一. 在轨迹数据处理方面, 高强等人[1]对基于轨迹大数据的处理技术进行了综述, 介绍了多种轨迹处理分析方法; 在出租车 GPS 轨迹分析方面, 郑运鹏等人[2]通过聚类方法识别乘客热点区域并对出租车司机进行热点推荐; 在基于出

① 基金项目: 陕西省工业科技攻关项目 (2015GY002); 国家自然科学基金青年科学基金 (61303041); 陕西省重点科技创新团队项目 (2017KCT-29); 陕西省国际科技合作计划项目 (2017KW-015); 陕西省重点研发计划项目 (2017GY-072, 2018GY-136)
Foundation item: Industrial Science and Technology Plan of Shaanxi Province (2015GY002); Young Scientists Fund of National Natural Science Foundation of China (61303041); Major Science and Technology Innovative Team Plan of Shaanxi Province (2017KCT-29); International Science and Technology Cooperation Program of Shaanxi Province (2017KW-015); Key Research and Development Program of Shaanxi Province (2017GY-072, 2018GY-136)
收稿时间: 2018-05-15; 修改时间: 2018-06-08; 采用时间: 2018-06-19; csa 在线出版时间: 2018-12-03

租车 GPS 轨迹数据研究城市交通状况方面, 王晗等人[3]提出使用大规模出租车 GPS 轨迹数据对城市交通状况进行建模和预测, 构建基于大规模出租车 GPS 轨迹的宏观交通密度动态模型, 预测未来的交通密度,何雯等人[4]提出了基于 GPS 轨迹的规律路径挖掘算法; 在基于 GPS 轨迹数据研究人类活动方面, 段宗涛等人[5]以西安市 GPS 轨迹为例基于 Hadoop 分布式系统设计了出租车服务策略分析模型. Moreira Matias L 等人[6]通过出租车 GPS 轨迹实时数据和出租车站牌的乘客需求预测在短时间内 (几分钟) 的出租车乘客空间分布从而进行智能派遣.

本文基于西安市大规模的出租车 GPS 数据, 利用大数据分析平台, 首先对不同时段的司机收入进行量化; 再根据各时段司机收入分布情况进行轨迹数据筛选, 然后利用过滤数据对出租车的换道行为进行识别和次数统计, 并计算出租车的平均速度; 最后与司机收入结合分析出租车司机换道行为的习惯和心理对运营的影响.

1 换道行为轨迹分析

1.1 数据预处理

本文基于西安市出租车调度系统采集的 GPS 数据; 数据格式依次为: ‘序号’, ‘车辆牌照’, ‘时间’, ‘经度’, ‘纬度’, ‘水平速度’, ‘方向’, ‘状态位’(0 无状态位 1 防劫 2 签到 3 签退 4 空车 5 重车 6 点火 7 熄火). 将 GPS 数据转化为文本格式, 上传到 HDFS 分布式存储平台. 原始的 GPS 轨迹数据可能存在数据缺失, 漂移, 时间错乱, 不完整等问题. 因此本文通过数据预处理过程从 12 000 多辆出租车 GPS 数据中, 选取了 3189 辆数据完备和准确的 GPS 数据进行分析, 根据统计学原理, 为了保障精度, 假设在整个样本区间 95% 的置信度下, 0.05 的误差限, 此时认为数据选取足够可靠.

在数据预处理过程中, 首先对 GPS 轨迹数据进行二次排序, 获得车辆的完整时空轨迹, 然后对其进行以下几个方面的处理:

(1) 错误数据. 一天出租车状态位为‘4’即空车的数据, 对于研究是无用数据, 对于超出西安市范围 (东经 108 度~109 度, 北纬 33.65 度~34.65 度) 的数据均认为无用的错误数据, 对错误数据直接删除.

(2) 重复数据. GPS 数据时间重复现象, 如果重复时间数据没有断点现象, 直接删除一条重复数据, 如果某时刻的重复数据造成断点现象, 删除重复数据后, 还需利用上下两条数据进行插值处理.

(3) 缺失数据. 原始的 GPS 轨迹存在部分字段缺失 (速度或航向角) 现象, 如果连续时间点中间某一字段缺失, 利用上下两条数据插值补足, 如果存在大量连续时间点字段缺失, 直接删除.

(4) 异常数据. 在原始 GPS 数据中, 存在 speed<0 或者 speed >120(单位为 km/h) 的异常速度值, 还存在相邻时刻的经纬度之差大于 0.012 阈值的 GPS 轨迹误差, 这些认为是异常数据, 这样的数据比较少直接剔除比较简单并对后续研究没有影响.

1.2 换道运动轨迹关系

汽车的行驶过程符合物理运动规律, 如图 1 所示建立汽车行驶运动的平面图. 图 1 中i, j, k代表连续时间间隔车辆所在的经纬度点, V_i, V_j, V_k为车辆的瞬时速度, δ_i, δ_j, δ_k表示车辆行驶的航向角 (即车辆瞬时速度与正北方向沿顺时针形成的角度), θ_{ij}, θ_{jk}表示轨迹方向 (即相邻上一时刻 GPS 点与当前 GPS 点的连线与正北方向的顺时针角度).

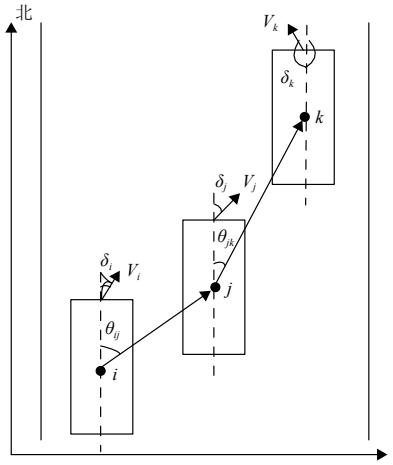

图 1 汽车行驶运动轨迹

出租车在行驶过程中会发生航向角和轨迹方向的波动, 而直线换道行为的轨迹波动会呈现规律变化, 如图 2 所示, $\Delta\delta_1$, $\Delta\delta_2$, $\Delta\delta_3$表示航向角偏差 (相邻两个航向角之差), $\Delta\theta_1$, $\Delta\theta_2$表示轨迹偏差 (相邻两个轨迹方向角之差), 连续的四个 GPS 轨迹中, $\Delta\delta$先变大, 再逐渐变小, 之后又变大满足关系$\Delta\delta_1 \approx \Delta\delta_3 > \Delta\delta_2$; 对于轨迹方向, 换道行为发生在同一道路方向, 变化不大, 即$\Delta\theta_1 \approx \Delta\theta_2$. 上述规律换道行为发生在同一道路方向, 图 2 中示例的是右拐弯, 左拐弯也符合这一规律, 如果是十字路口拐弯则满足$\Delta\delta_1 \approx \Delta\delta_3 < \Delta\delta_2$且$\Delta\theta_1 < \Delta\theta_2$, 环岛行

为满足$\Delta\delta_3 > \Delta\delta_1 > \Delta\delta_2$且$\Delta\theta_1 < \Delta\theta_2$, 因此根据航向角偏差和轨迹偏差的关系与变化规律可以区别不同的换道行为, 本文研究的是在同一道路方向的换道行为即直道换道行为.

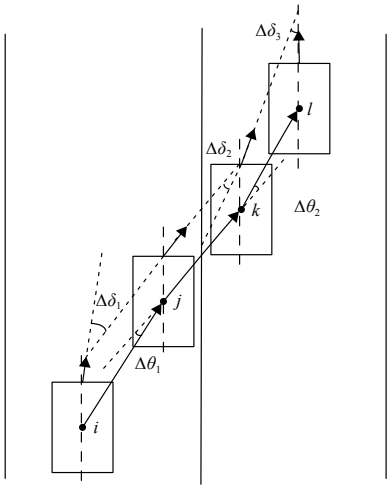

图 2 换道运动轨迹关系

2 换道行为识别模型

2.1 换道行为模型特征向量的选取

以换道行为的航向角偏差和轨迹方向偏差为特征, 建立特征向量样本集合$T=\{(x_1,y_1),\cdots,(x_n,y_n)\}$, 其中$x_i=(\Delta\delta_{i1},\Delta\delta_{i2},\Delta\delta_{i3},\Delta\theta_{i1},\Delta\theta_{i2}),1\le i\le n$, 其中$\Delta\delta_{i1}$, $\Delta\delta_{i2}$, $\Delta\delta_{i3}$表示连续四个 GPS 轨迹点之间的航向角偏差, 同理$\Delta\theta_{i1}$, $\Delta\theta_{i2}$表示连续四个 GPS 轨迹点之间的连续两个轨迹方向偏差; 以D_1表示 1 次出租车换道行为, D_2表示其它情况; y_i表示样本x_i属于D_1类或D_2类的数学值, 则:

$$y_i=\begin{cases}1, & x_i\in D_1\\ -1, & x_i\in D_2\end{cases} \tag{1}$$

2.2 换道行为识别支持向量机模型的实现

支持向量机 (Support Vector Machine, SVM) 是机器学习领域可以用于分类的经典模型, 它在解决有限样本, 数据非线性适应, 模式识别中表现出许多优势, 本文利用 Spark 的 SVMWithSGD 进行换道行为识别.

SVMWithSGD 通过求得最优分类面$g(x)=w^{\mathrm{T}}x+b$, 将D_1和D_2分开. SVMWithSGD 引入损失函数和正则化函数求解最优分类面[7], 转化为目标函数

$$\min_{w,b}\sum_{i=1}^{n}\max(0,1-y_i(wx_i+b))+\lambda\|w\|^2 \tag{2}$$

最后采用随机梯度下降求解. 根据以上分析, 换道行为识别算法如表 1.

表 1 换道行为识别算法

算法步骤	算法说明
1	获取 GPS 轨迹数据, 计算航行角偏差和轨迹方向偏差
2	建立训练集特征向量(x_i,y_i)
3	SVMWithSGD 训练出 SVMModel 即最优分类面 $g(x)=w^{\mathrm{T}}x+b$
4	SVM 测试, 计算测试样本$\boldsymbol{x}_i$的分类结果

2.3 换道行为识别支持向量机模型的验证

经过出租车 (装配 GPS 设备) 在城市道路的实测, 如表 2 所示得到 200 个由航向角和轨迹方向偏差与标签组成的特征数据, 标签 1 表示换道行为, 0 表示其他, 将 200 个特征数据随机分为训练样本 (120 个样本, 大概有 50 多次换道行为) 和测试样本. 利用 spark 的 SVMWithSGD(stepSize, numIterations, gradient, updater, optimizer) 进行模型训练. 其中, stepSize: 迭代步长, 默认为 1; numIterations: 迭代次数, 设置为 50; gradient: 梯度下降的损失函数, 设置为 hinge loss; updater: 正则化, 设置为$L2$范数; optimizer: 随机梯度下降优化计算.

表 2 部分特征数据集

序号i	$\Delta\delta_{i1}$/度	$\Delta\delta_{i2}$/度	$\Delta\delta_{i3}$/度	$\Delta\theta_{i1}$/度	$\Delta\theta_{i1}$/度	标签
1	16.0	4.0	16.0	4.0	3.0	1
2	18.0	64.0	12.0	6.0	21.0	0
…	…	…	…	…	…	…
198	36.0	10.0	28.0	1.5	2.0	1
199	18.0	26.0	35.0	18.0	26.0	0
200	64.0	21.0	58.0	4.0	3.0	1

采用 80 个测试样本 (40 次换道行为) 进行测试, 测试结果如表 3 所示, 可以看出, Spark 的 SVM 算法可以准确的识别换道行为.

表 3 SVM 分类准确率

测试样本数	正确匹配数	准确率 (%)
10	10	100
30	29	97
50	48	96
70	67	96
80	77	96

3 出租车司机换道行为的大数据分析方法实现

司机的换道行为的分析主要是研究不同时段司机换道行为发生的次数与出租车平均行驶速度的关系, 首先将一天划分成 12 个时段 (即每两个小时为一个时

段), 并按时段量化司机收入, 大量的收入数据满足大数中心定律, 近似服从 T 分布, 对平均收入进行区间估计并过滤置信范围内的 GPS 轨迹数据, 数据过滤后, 利用 SVM 进行换道行为识别, 并统计各时段换道行为发生次数和出租车平均行驶速度. 换道行为的分析的关键是数据的定量统计和分析, 由于数据量规模比较大, 传统的的数据库和数据分析工具 SPSS 等分析方法比较耗时, 而且本文分析的统计量计算步骤繁复, 需进行大量的迭代操作. Spark 是一个分布式系统架构的计算引擎, 它是基于内存计算的框架, 计算速度快, 适合大规模数据迭代运算[8], 并且 spark 包含数据统计库函数, 可以直接调用进行复杂的数学统计, 为了分析计算方便, 因此本文采用大数据平台 spark 进行分析数据.

3.1 出租车司机收入量化

司机收入量化即计算司机每个时隙的收入, 通过计算每个时隙出租车司机载客的距离, 根据出租车运营价格大致得到司机的收入. 利用司机载客状态的 GPS 数据相邻时间经纬度的变化, 计算出间隔的距离, 累积一个时隙得到司机的载客距离, 从而得到司机的大致收入. 利用 spark 对出租车司机收入量化, 主要用到三个算子: map, groupByKey 和 mapValue. map 算子将原始数据切分成 key_value 型数据 ((车牌号, 日期, 时隙), 原始数据), groupByKey 将 (车牌号, 日期, 时隙) 这种 key 相同的数据聚合到一个组. mapValue 对 group 成一组的数据进行操作, 对司机收入进行量化. 算法 1 中的 V 表示原始数据, 原始数据: (车牌号,日期,时段,经度,纬度,航向角,状态)=(carid,date,slot,lon,lat,A,state).

算法 1. 司机收入并行量化算法

输入: 按时间排序的 GPS 轨迹数据.
输出: 各个时隙司机的收入.

```
1. map:
2. 切分数据为 key_value 形式: ((carid,date,slot),V);
3. groupByKey:
4. 将 key 相同的数据聚合成一组,即
(carid,date,slot),iterable(V));
5. mapValue:
6. 对每组数据的 value 进行操作
7. for (line<- iterable(V)) {
8. if(state==5){
9. 累积计算载客状态的每两个相邻时间间隔 GPS 点距离, olon, olat 代表上一个经纬度;
10. distance=GetDistace(olon,olat,lon, lat);
11. 累积计算一个时隙载客总距离;
12. Dis=Dis+distance;
13. }
14. 累积计算一个时隙司机的收入;
15. Salary=Dis*运营单价;
16. }
17. 整合数据为 ((carid,date,slot),Salary).
```

3.2 按出租车收入的置信区间提取 GPS 轨迹数据

对于出租车司机的收入, 由于存在很多非营运因素造成出租车司机收入过高或过低, 影响出租车换道行为的分析结果. 因此, 首先假设出租车司机收入服从 T 分布, 然后对出租车司机收入进行区间估计, 在后续分析过程中, 仅保留收入在置信水平为 95% 的置信区间内的出租车的轨迹数据, 以消除收入极值的影响. 出租车司机收入的置信区间如下式

$$\left(\bar{x}-\frac{S}{\sqrt{n}}t_{\alpha/2}(n-1),\bar{x}+\frac{S}{\sqrt{n}}t_{\alpha/2}(n-1)\right) \tag{3}$$

其中, n为出租车数量, $\bar{x}$为出租车收入均值的估计值, S为出租车收入方差的估计值, $t_{\alpha/2}(n-1)$ 为 T 分布水平 α的分位数, 其中显著水平α为 0.05. 在 spark 平台上提取各时段平均收入置信区间内数据, 算法 2 首先对量化收入用 map 数据转化为 (时隙, (收入, 原始数据)), 然后进行 groupByKey 和 mapValue 计算, 得到各个时段的置信区间, 然后对收入数据进行 filter 操作, 过滤出置信数据.

算法 2. 司机收入置信区间范围内轨迹提取算法

输入: 算法 1 数据 ((carid,date,slot),Salary)
输出: 置信区间的轨迹数据

```
1. map:
2. 转换数据为 (slot,( Salary,V));
3. groupByKey:
4. 将 key 相同数据聚合为一组, 即
5. (slot, Iterator(Salary,V));
6. mapValue:
7. 计算各时隙平均收入x̄和方差 S
8. x̄=getAveSalary(Iterator(Salary,V));
9. S=getVar(Iterator(Salary,x̄));
10. map:
11. 得出每个时隙收入的置信区 condis;
12. filter:
13. 从 (slot, (Salary, V)) 过滤置信数 data;
14. data=filter(condis.contains(Salary) ).
```

3.3 出租车司机换道次数提取

上述置信区间的轨迹数据, 算法 3 在 Spark 上通过 groupByKey 和 mapValue 计算航向角和轨迹方向的偏差和平均行驶速度, 并建立特征向量, 通过 map 算子利用 Spark 已建立 SVM 模型进行换道行为判别, 最后通过 reduceByKey 统计换道次数.

算法 3. 基于 SVM 的出租车换道次数提取算法

输入: 置信区间的轨迹数据
输出: 换道次数和平均速度

1. groupByKey:
2. 将 (slot,V) 按 key 聚合为一组;
3. mapValue:
4. 计算每个时隙的平均速度 meanv
5. meanv=getAverageSpeed(Iterator(V));
6. mapValue:
7. 计算连续四个 GPS 轨迹的航向角偏差和轨迹方向偏差 $\boldsymbol{x}_i=(\Delta\delta_{i1},\Delta\delta_{i2},\Delta\delta_{i3},\Delta\theta_{i1},\Delta\theta_{i2})$;
$\boldsymbol{x}_i$=getVector(Iterator(V));
8. map:
9. 判别换道行为为 1, 其他为 0;
10. state=SVMModel.predict(x_i);
11. reduceByKey:
12. 统计换道次数 S;
13. 整合数据为 (slot, (meanv, S)).

4 实例分析

通过司机收入量化, 置信数据提取, 和换道行为的分析, 得到了各个时隙的换道行为次数, 司机的收入, 和出租车平均速度. 最后, 对换道行为的次数和出租车平均速度, 司机收入进行相关性计算, 评估三者的关联关系. 实验环境: Spark 集群, 13 台 Workers, 104 个 Cores,64G Memory; GPS 数据集: 西安市 2015 年 9 月一个月的数据, 约 70 GB 数据, 包含 1 万多辆出租车.

本文相关性分析采用皮尔逊相关系数评估算法, 通过相关系数临界值表得知, 当自由度为 1000, 显著水平 $\alpha = 0.05$时, 显著相关的相关系数临界值为 0.041, 即对于 1000 组样本数据, 其相关系数绝对值大于等于 0.041, 则相关性显著. 以下各图分析样本的数量均大于等于 1000.

在图 3 中, 可以看出 0 点到 5 点, 出租车司机收入与置信范围的边界值低于其它时段, 说明乘客量比较少, 自 6 点后, 收入开始增加, 在 8 点到 9 点和 18 点到 19 点, 出租车司机收入相比前一时段增幅比较大, 说明这几个时段是客流量高峰期.

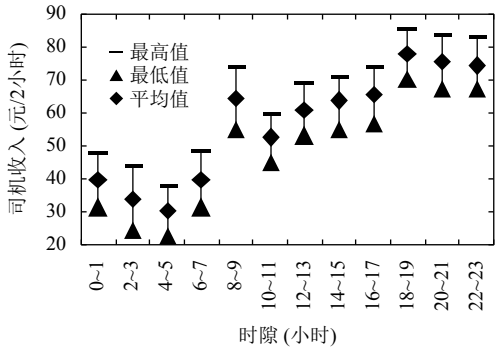

图 3 各个时隙出租车司机的平均收入与其置信范围的最低和最高值

在图 4 中, 0 点到 5 点速度高于其它时段, 速度与收入的相关系数低于临界值 0.041, 不显著相关, 可以理解在这期间司机的驾驶速度快, 与收入没有直接的关联关系, 自 6 点后的其它时段, 速度降低, 相关系数大于临界值, 呈现正相关, 表明在这期间, 司机的驾驶速度变快, 收入就会增加, 并且在速度较低的时段相关系数明显大于临界值, 说明在车流量多的时候速度与收入正相关性越显著.

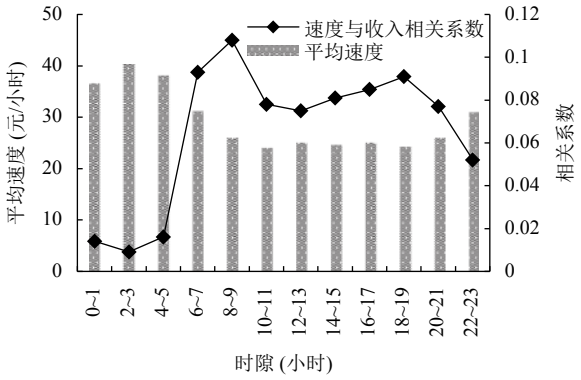

图 4 各个时隙出租车的平均速度及其与司机收入平均值的相关系数

图 5 中 0 点到 5 点平均换道次数少于其他时段, 速度与换道次数的相关系数为负值, 小于临界值-0.041, 可以看出换道次数与平均速度没有直接关系; 6 点以后换道次数增多, 在车流量高峰时明显高于其他时段, 如 8 点到 9 点, 18 点到 19 点, 换道次数与速度的的相关系数呈现显著的负相关, 说明速度越慢, 换道次数越多. 结合图 4 速度与收入的相关系数, 可以得出结论, 0 点到 5 点, 在这期间, 车流量少, 出租车司机都以较快的驾驶速度运营, 换道次数少, 速度与换道次数和

收入三者没有相关关系; 6 点以后, 随着车流量增加, 收入与速度呈现正相关, 出租车司机通过提高速度来增加收入, 为了提高速度多次换道, 尤其在车流量高峰期, 而频繁的换道行为, 造成了交通不畅, 加剧了交通拥堵, 反而降低车的行进速度, 因此换道次数与速度呈现负相关.

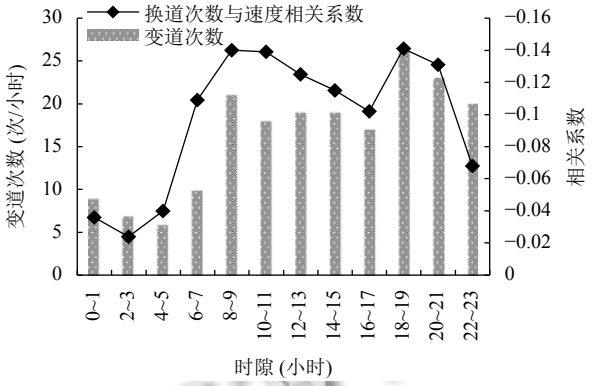

图 5 各个时隙出租车的换道次数及其平均速度的相关系数

5 总结

出租车司机的直道换道行为是一种司空见惯的驾驶行为, 一般的, 我们认为出租车司机的的换道行为会提高出租车的驾驶速度, 有利于出租车司机更好的服务乘客, 但这种行为会对城市交通状况产生什么影响, 进一步这一行为受到什么因素驱动, 本文通过大规模的 GPS 轨迹数据挖掘城市出租车司机的换道行为, 发现出租车 GPS 轨迹数据中隐藏的规律与特征.

通过结果分析, 出租车司机的直道换道行为受到收入的驱动, 当车流量较少时, 换道行为可以提高收入, 但当车流量较大时司机频繁换道行为致使车流整体速度降低, 造成交通不畅, 而出租车司机驾驶速度会影响它的收入, 速度越慢收入越低, 所以当车流量较大时, 司机的频繁直道换道行为会降低收入并且影响城市交通状况, 而且容易造成交通事故, 建议出租车司机在车流量较大时尽量减少不必要的换道行为, 养成良好的驾驶习惯, 促进城市交通和谐.

参考文献

1 高强, 张凤荔, 王瑞锦, 等. 轨迹大数据: 数据处理关键技术研究综述. 软件学报, 2017, 28(4): 959–992.
2 郑运鹏, 赵刚, 刘健. 基于出租车 GPS 数据的交通热区识别方法. 北京信息科技大学学报, 2016, 31(1): 43–47.
3 王晗. 基于出租车轨迹的路网交通流建模研究[硕士学位论文]. 北京: 北京交通大学, 2017.
4 何雯, 李德毅, 安利峰, 等. 基于 GPS 轨迹的规律路径挖掘算法. 吉林大学学报 (工学版), 2014, 44(6): 1764–1770.
5 段宗涛, 陈欣欣, 康军, 等. 基于 Hadoop 的出租车服务策略. 计算机系统应用, 2017, 26(1): 255–259.
6 Moreira-Matias L, Gama J, Ferreira M, *et al*. Predicting taxi–passenger demand using streaming data. IEEE Transactions on Intelligent Transportation Systems, 2013, 14(3): 1393–1402. [doi: 10.1109/TITS.2013.2262376]
7 闫辛. 半监督支持向量机模型与算法研究[博士学位论文]. 上海: 上海大学, 2016.
8 Apache 基金会 Spark 官方文档. http://spark.apache.org/, 2017-06-11/2017-06-28.