基于meanshift的OD计算与公交下车站点判定方法

**摘 要** 为了解城市公共交通出行的时空特征，改善城市公交的运营服务，提升居民公共交通出行效率。

研究并实现了计算公交乘客的下车站点及下车时间的方法。方法利用出行链[1]的思想，对连续出行的乘客采用阈值内最近邻算法判定公交下车站点信息，对一些出行链头部或尾部断裂的情况，通过分析乘客中长期的历史乘车记录，利用meanshift聚类算法得出每个乘客的居住地中心点O和工作地中心点D，采用阈值内最近邻算法计算离工作地中心或居住地中心距离最近的方法对下车站点进行进一步补充。实验数据表明，此方法能补充大部分公交交易数据的下车站点，且准确率较高。

**关键词** 出行链 公交下车站点 OD meanshift

Abstract …

# 0 引言

近年来，越来越多的人涌向北上广深一线城市，这些城市的公交地铁成为了绝大部分人的出行方式，因此，城市公交线网也面临着越来越大的压力，研究这些乘客的出行时间、空间以及公交系统的客流量，对交通服务升级、交通运行效率提升、公交线网优化、交通工具的科学布局组织等均有重要的意义。广州市公交车普遍安装了进出站监测系统和车载GPS导航定位设备，按时序记录了公交车的进出站信息及粗略的gps行车轨迹数据，所有连续运行的公交车辆和众多的乘客刷卡交易形成了城市公共交通运行和乘客出行大数据。目前，公交收费大都是一票制的，只有上车刷卡，没有下车刷卡，为了研究乘客出行方式，首先要确定公交下车站点及下车时间，以公交IC卡为主的数据分析方法，数据丰富有效，成本较低，是现在公交客流OD研究的主流方法[2−3]。国外对结合大数据的公交客流OD研究较早，也相对成熟。国外Barry等[4]、Jinhua Zhao等[5]基于出行链思想，利用AFC，车辆自动定位系统，依托地铁公交联合数据，利用大数据分析实现了公交客流的推导。Erik等[6]利用公交出行链，分析出了出行时间的变化特征，揭示了城市交通出行过程的连续性。国内学者研究主要集中于IC卡下车站点的推导，对于公交客流OD的推导研究较晚。胡郁葱等[7]通过IC卡数据挖掘技术获取了公交OD矩阵。陈峥嵘[8]将智能公交数据处理方法应用于公交客流OD研究。胡继华等[9]提出结合出行链的IC卡公交客流研究方法。

本文以广州市羊城通卡交易数据为实例进行分析研究，以出行链的思想，对每天的交易情况中每个乘客的连续交易行为进行分析，以公交IC卡交易数据、公交进出站数据和线路站点数据为基础，针对完整的公交出行链，提出了一种判定公交下车站点的最近邻下车站点算法，对一些出行链不完整的情况，根据乘客长期出行的时空规律性特点，通过分析用户中长期的历史交易情况，提出了采用meanshift算法计算乘客的居住地(O)和工作地(D)，再利用OD预测一些不完整出行链里面交易记录的下车站点的方法。通过这两种方法，对公交交易数据的补充能达到85%，通过抽取部分羊城通用户的出行下车站点进行验证，有非常高的准确率。

# 1 交易数据与公交数据简介

本文数据来源于广州羊城通卡的公交交易数据，地铁交易数据，公交公司提供的每天的进出站数据以及线路站点数据。在“一票制”的收费模式下，公交交易数据中不包含下车站点和下车时间信息，地铁交易数据包含地铁上车站点、上车时间以及地铁下车站点、下车时间。公交交易数据和地铁交易数据如表1、表2所示。

表1 公交交易数据

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 逻辑卡号 | 车辆大牌号 | 班次 | 线路编码 | 上车站点编码 | 上车站点名称 | 上车站点经度 | 上车站点纬度 | 上车时间 | 上下行 |
| 5100008512149944 | 粤A09576D | 2 | 00080 | 00080114 | 石涌口站 | 113.28334 | 23.109627 | 20191212070653 | 1 |
| 5100000280809967 | 粤A20742D | 4 | 00080 | 00080107 | 鹭江站2 | 113.308378 | 23.095455 | 20191212065635 | 1 |
| 5100000720473639 | 粤A24698D | 22 | 00080 | 00080119 | 惠福路（省医门诊部）站 | 113.259324 | 23.120017 | 20191212164133 | 1 |
| 5100001007976043 | 粤A09576D | 2 | 00080 | 00080107 | 鹭江站2 | 113.308378 | 23.095455 | 20191212065202 | 1 |
| 5100005967023913 | 粤A04085D | 25 | 00080 | 00080116 | 草芳围站 | 113.273331 | 23.112619 | 20191212183507 | 1 |
| … | … | … | … | … | … | … | … | … | … |

表2 地铁交易数据

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 逻辑卡号 | 进站站点名称 | 进站站点经度 | 进站站点纬度 | 进站时间 | 出站站点名称 | 出站站点经度 | 出站站点纬度 | 出站时间 |
| 5100000463880707 | 燕岗 | 113.271988 | 23.075142 | 20191212180552 | 天河客运站 | 113.3436 | 23.171 | 20191212184628 |
| 5100000661412876 | 白云大道北 | 113.298178 | 23.222382 | 20191212215134 | 人和 | 113.295995 | 23.335224 | 20191212220917 |
| 5100003001572874 | 竹料 | 113.368332 | 23.348496 | 20191212102338 | 纪念堂 | 113.263431 | 23.132179 | 20191212112113 |
| 5100008504680653 | 万胜围 | 113.384796 | 23.097712 | 20191212114836 | 大学城北 | 113.385639 | 23.058091 | 20191212115813 |
| 5100000487561440 | 磨碟沙 | 113.341285 | 23.098689 | 20191212204457 | 江泰路 | 113.28057 | 23.082231 | 20191212210541 |
| … | … | … | … | … | … | … | … | … |

公交车的进出站监测系统记录了每天所在公交车辆的进站出站数据，公交交易数据和公交进出站数据之间存在着对应关系。每辆车都有一个智能调度编号，根据这个编号，对进出站数据预先划分好了把班次，如表3所示

表3 公交车辆班次信息

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 车辆大牌号 | 班次 | 站点编码 | 编号 | 进站时间 | 经度 | 纬度 | 线路编码 | 线路名称 |
| 粤A37951D | 1 | 00910002 | 2 | 20191201060057 | 23.063393 | 113.13251 | 00910 | 91线 |
| 粤A37951D | 1 | 00910003 | 3 | 20191201060342 | 113.13629 | 23.060812 | 00910 | 91线 |
| 粤A37951D | 1 | 00910004 | 4 | 20191201060536 | 113.13501 | 23.059775 | 00910 | 91线 |
| … | … | … |  | … | … | … | … | … |

线路站点数据也是由公交公司提供的，它描绘了最近一段时间广州市公交的上行线路或下行线路所包含的站点信息，经纬度经过验证都是准确的，属于火星坐标。如表4所示

表4 公交线路站点信息

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 线路编码 | 线路名称 | 站点编码 | 站点名称 | 上下行 | 站点序号 | 经度 | 纬度 |
| 00012 | 大学城1线 | 00012001 | 广外总站 | 0 | 1 | 113.28815 | 23.200052 |
| 00012 | 大学城1线 | 00012002 | 广外（白云山西门）站2 | 0 | 2 | 113.28177 | 23.20192 |
| 00012 | 大学城1线 | 00012003 | 白云国际会议中心站2 | 0 | 3 | 113.27467 | 23.187626 |
| …… …… …… …… …… …… …… …… | | | | | | | |
| 00077 | K3 | 00077128 | 天丰路口站 | 1 | 28 | 113.4583733 | 23.1624833 |
| 00077 | K3 | 00077129 | 广东国防教育基地总站 | 1 | 29 | 113.424504 | 23.155417 |
| 00077 | K3 | 00077130 | 玉树大街口站 | 1 | 30 | 113.43099 | 23.155539 |

在下车站点判定过程中，本文提出了采用地铁数据和公交数据联合分析的方法，采用基于出行链的乘客下车站点概率模型进行下车站点的判定。

# 2 最近邻推算下车站点算法

基于广州市的公交交易和地铁交易的融合数据，按照每个用户的刷卡进站或上车时间对所有交易排序，对于每天有N(N>1)次乘车记录的乘客，先拟定这些乘客的出行链是闭合的，得出的用户出行方式可以归纳为以下2类情况。如图1



图1 闭合出行链

第一类为对称出行，这类乘客最典型的就是上班族或学生，他们的出行归纳为H->W->W->H,去的时候中间没有换乘或有一两次换乘，回去时也没有换乘或有一两次换乘。本文将广州市常驻人口每天的出行划分为居住地和工作地，居住地和工作地范围内都会聚集一些公交站点或地铁站点。大多数上班族用户每天上午从居住地选取一个站点出发，下午从工作地选取一个站点乘车回家。第二类为非对称出行，这些乘客的出行不具有时空规律性，可能是临时去某些地方，比如去旅游，购物，聚会等等，他们的出行归纳为H->T1->T2->T3->H(Ti指临时地)。中间可能经过一次或N次换乘，但最终都回到了居住地。



图2 相邻两次交易的下车站点判定

假设乘客在一次乘车记录中上车站点为Si，SiLA,所属线路是LA{S1,S2,S3,…},则此次的下车站点Sj满足如下条件SjLi,j>i。

最近邻下车站点判定可以描述为

1. 按天输入某个用户的按进站时间排序的交易信息列表lno:{on\_station\_time1,on\_station\_time2…,on\_station\_timei,… on\_station\_timelast}
2. 如果只有一条，先不作处理。如果有多于2条记录，按顺序读取第i条和i+1条，如果i是最后一条记录，则读取第i条和0条。分别拿到相邻两次刷卡站点是Si和S(i+1)%last
3. 判断遍历到的相邻两条记录是否属于B->B,B->R这类换乘方式，如果属于这两种情况，计算Si站点所在线路下游站点LA{Sj},j>i。计算下游每个站点Sj与S(i+1)%last的欧式距离D(Sj,S(i+1)%last)，如果D(Sk,Sj)<*d*(*d*为指定的一个距离阈值)，则放入备选集合*R*{S1,S2,S3…}。如果相邻这两条交易记录是R->R,R->B则不需作处理。
4. 结果集*R*中拿取距离D(Sj,S(i+1)%last)最小的一个站点作为第i条公交交易记录的下车站点。如果*R*为空，说明出行链断裂，先不作处理

如果*R*为空，说明下游每个站点Sj与S(i+1)%last的距离都不小于*d*，认为出行链断裂，对于第i(1<i<last)条公交交易记录，属于中途断裂的，无法确定其下车站点。对于R集中，如果最小的站点距离为0.0，则属于同站换乘。

# 3 OD进一步补充下车站点

## 3.1 meanshift算法计算OD

每个乘客的居住地或工作地附近会有很多不同的公交站点或地铁站点你，乘客也会选择不同的站点乘车上下班，分析乘客中长期的这些交易记录，采用求密集点中心点的方法得出居住地或工作地的中心点。meanshift算法其实通过名字就可以看到该算法的核心，mean（均值），shift（偏移），简单的说，也就是有一个点x,它的周围有很多个点xi，计算点x移动到每个点xi所需要的偏移量之和，求平均，得到平均偏移量，（该偏移量的方向是周围点分布密集的方向）该偏移量是包含大小和方向的。然后点x往平均偏移量方向移动，再以此为新的起点不断迭代直到满足一定条件结束。

给定d维空间Rd的n个样本点 ,i=1,…,n,在空间中任选一点x，那么Mean Shift向量的基本形式定义为:



Sk是一个半径为h的圆形区域，满足以下关系的y点的集合,



k表示在这n个样本点xi中,有k个点落入Sh区域中.在二维坐标中，任选一个点，然后以这个点为圆心，h为半径做一个圆。落在这个圆内的所有点和圆心都会产生一个向量，向量以圆心为起点，以落在圆内的点为终点，然后把这些向量相加求平均，结果就是meanshift向量。

如图所以。其中蓝色箭头就是Mh（meanshift向量）。再以meanshift向量的终点为圆心，再作一个二维的圆。如图3所示，重复以上步骤，可再得到一个meanshift向量，如此重复下去，meanshift向量可以收敛到概率密度最大得地方，也就是最稠密的地方。



图3 求密集点的中心点过程

结合到现有的业务场景，广州市常驻人口的出行分析，始终围绕着工作地和居住地展开，根据乘客通勤出行的时空规律性,分析地面和轨道交通中长期的海量交易数据，拟定每个用户工作日上午最早的一次交易记录或非工作日第一次交易记录的上车站点坐标<home\_lon,home\_lat>属于居住地范围，当然少数情况下，乘客前一天可能在其它地方居住，当天的最早乘车不是从居住地出发。工作日下午最晚的一次交易记录的上车站点坐标<work\_lon,work\_lat>属于工作地范围，少数情况下可能先以非公交方式去某些地方，再坐公交回去。这些少数情况下非工作地或居住地的坐标较为离散，拿到多天的居住地拟坐标{<home\_lon1,home\_lat1>,<home\_lon2,home\_lat2>,<home\_lon3,home\_lat3>…}，采用meanshift算法求密集点的中心点,得到用户居住地的中心点作为O，类似的方法，得到用户工作地的中心点D。

## 3.2 OD补充断裂的出行链

对于一些乘客的出行记录都是在上午的情形，lno:{on\_station\_time1,on\_station\_time2…,on\_station\_timei,…on\_station\_timelast}, on\_station\_timei<130000，通过前面最近邻站点算法计算，会判断当天最后一次交易记录和首次交易记录的结果集R是否为空，如果不为空，这些乘客在上午完成一次H->T->H(T指临时地)；如果为空，说明乘客出行链断裂，不是回家，对于这些出行链尾部断裂的乘客，拟定这些乘客是前往工作地。如图3



图3 尾部出行链断裂

对于当天上午最后一次公交交易记录，计算下游站点与工作地中心点D的距离D(Sj,D)，设定另外一个阈值*d2*,从小于*d2*的结果集*R*中选取距离最小的站点作为最后一次公交的下车站点。

对于一些乘客的出行记录都在下午的情形，lno:{on\_station\_time1,on\_station\_time2…,on\_station\_timei,…on\_station\_timelast}, on\_station\_timei>130000，拟定这些乘客最终都是要回居住地的，无论是否是上班。可能是从工作地回居住地，也可能从临时地回居住地。对于这些出行链头部断裂的乘客，计算与居住地中心点O的距离小于*d2*的结果集*R*，从*R*中选取距离最小的站点作为最后一次公交的下车站点。如图2



图4 出行链头部断裂

上述两部作处理后，对于*R*集合为空的情况，这类乘客的出行链划分为不可预测的断裂出行链。这类乘客可能中途因为其它原因选择非公交的其它出行方式，或者下班后去其它地方，或者上午不去上班，无法作预测。如图5



图5 无法预测下车站点的情形

## 3.3 下车时间补充

进出站信息表里记录了每辆公交在每个班次进每个站点的时间enter\_time。通过上述两阶段补充能确定下车站点名称off\_station\_name和下车站点编号off\_station\_no，根据乘客当天所乘公交车的车辆大牌号，所属班次信息，联合当天进出站信息里对应车辆的大牌号，班次信息，根据下车站点编号能找到对应的下车站点信息，拿到这个enter\_time时间作为这条交易记录里面下车站点的下车时间。

# 4 实验结果与分析

本实验以广州市2019年12月12日的公交交易和地铁交易数据进行研究。剔除掉站点坐标有误差的交易记录，提取样本总的交易记录1413202条，根据站点服务便利水平[10]，设置连乘距离阈值*d*=1km,第一阶段补充后，补充下车站点的数据为769370，占到总数据的55%，如表5

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 逻辑卡号 | 车辆大牌号 | 班次 | 上下行 | 上车站点编号 | 上车站点名称 | 上车站点经度 | 上车站点纬度 | 下车站点编号 | 下车站点名称 | 下车站点经度 | 下车站点纬度 |
| 5100000100689285 | 粤ACZ533 | 5 | 0 | 01060001 | 锦城花园（东风东）总站 | 113.308924 | 23.131261 | 01060006 | 三角市站 | 113.273828 | 23.125047 |
| 5100000101438503 | 粤A09136D | 2 | 1 | 00250113 | 同福中站 | 113.259323 | 23.10421 | 00250118 | 华海大厦站1 | 113.276039 | 23.092644 |
| 5100000101438503 | 粤A30951D | 11 | 0 | 07680004 | 海印桥南站2 | 113.283340 | 23.100723 | 07680007 | 江南大道中站2 | 113.269028 | 23.107765 |
| … | … | … | … | … | … | … | … | … | … | … | … |

表5 第一阶段补充下车站点

采用meanshift计算乘客的居住地和工作地的中心坐标，分析近2个月的公交地铁交易数据，得出用户的OD(居住地的中心和工作地的中心)，如表6

|  |  |  |
| --- | --- | --- |
| 逻辑卡号 | 居住地中心点坐标 | 工作地中心点坐标 |
| 5100001001990234 | 113.315391,23.092266 | 113.311124,23.102409 |
| 5100000946167060 | 113.395548,23.124465 | 113.357939,23.110331 |
| 5100008530906080 | 113.488704,23.177012 | 113.463858,23.123432 |
| … | … | … |

表6 用户OD坐标信息

第二阶段设置*d2*=1.5km，补充完下车站点后，补充的下车站点数据量为1201217，占到总数据量的85%。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 逻辑卡号 | 上车站点编号 | 上车站点名称 | 上车时间 | 下车站点编号 | 下车站点名称 | 下车时间 | 是否匹配 |
| 5100000782843700 | 00270124 | 东站汽车客运站 | 181902 | 00270125 | 瘦狗岭站 | 182306 | 是 |
| 5100000661121446 | 80050007 | 东华南路站 | 180429 | 80050013 | 石牌桥站B1 | 183449 | 是 |
| 5100000782843700 | 00270124 | 东站汽车客运站 | 211124 | 00270125 | 瘦狗岭站 | 211204 | 是 |
| 5100000722005355 | 80050007 | 东华南路站 | 190714 | 80050011 | 杨箕村站1 | 191951 | 是 |
| 5100000661121446 | 80050007 | 东华南路站 | 193017 | 80050013 | 石牌桥站B1 | 195040 | 是 |
| 5100000782843700 | 80110013 | 燕塘企业站1 | 215856 | 80110035 | 天寿路站1 | 232256 | 否 |
| 5100000661121446 | 01290021 | 晓港新村站 | 133356 | 01290027 | 东山口（东华北路）站 | 135325 | 是 |
| 5100000782843700 | 00540110 | 燕塘企业站1 | 181208 | 00540114 | 春晖苑站 | 182808 | 是 |
| 5100000782843700 | 00600108 | 广园天寿路口东站 | 080609 | 00600112 | 沙河站1 | 081606 | 是 |
| … | … | … | … | … | … | … |  |

抽取100名乘坐公交的乘客进行问卷调差，准确度为90%

# 5 结束语

本文通过出行链的方式，分析了广州市的公交交易数据和地铁交易数据，对于出行连续的乘客，采用本次乘车站点所在线路的下游站点与下次乘车上车站点的距离进行比较，取阈值内距离最小的站点作为下车站点。采用聚类算法进行工作地和居住地计算，使用工作地中心和居住地中心对公交交易数据作了进一步补充。实验证明次方法较为有效，准确率高。

# 参考文献

[1] KITAMURAR. Sequential, history dependent approach to trip chaining behavior[J]. Transportation Research Record,1983(944):13-22.

[2] 杨万波,王昊,叶晓飞,等.基于GPS和IC卡数据的公交出行OD推算方法[J].重庆交通大学学报(自然科 学版),2015, 34(4):117−121.

[3] 陆百川,邓捷,马庆禄,等.基于IC卡和RBF神经网络的短时公交客流量预测[J].重庆交通大学学报(自然 科学版),2015,34(6): 106−110.

[4] Barry J, Newhouser R, Rahbee A, et al. Origin and destination estimation in New York City with automated fare system data[J]. Transportation Research Record, 2002, 18(17): 183−187.

[5] ZHAO J H. The planning and analysis implications of automated data collection systems: rail transit OD matrix inference and path choice modeling examples[D]. Cambridge: Massachusetts Institute of Technology, 2004.

[6] JENELIUSE.Thevalueoftraveltimevariabilitywith trip chains, flexible scheduling and correlated travel times[J]. Transportation Research Part B Methodological, 2012,46(6):762-780.

[7] 胡郁葱,梁杰荣,梁枫明.基于IC卡数据挖掘获取公交OD矩阵的方法[J].交通信息与安全，2012, 30(4): 66−70. HU Yucong, LIANG Jierong, LIANG Fengming. A way to get bus regional OD matrix based on mining IC card information[J]. Journal of Transport Information and Safety, 2012, 30(4): 66−70.

[8] 陈峥嵘. 智能公共交通系统数据分析方法与应用研究 [D]. 南京: 东南大学, 2012. CHEN Zhengrong. Study on intelligent public transportation system data analysis methods and applications[D]. Nanjing: Southeast University, 2012.

[9] 胡继华, 邓俊, 黄泽. 结合出行链的公交IC卡乘客下车站点判断概率模型[J].交通运输系统工程与信息, 2014, 14(2): 62−67. HU Jihua, DENG Jun, HUANG Ze.Trip-chain based probability model for identifying alighting stations of smart card passengers[J]. Journal of Transportation Systems Engineering and Information Technology, 2014, 14(2): 62−67.

[10] 闫卫坡.大城市公共交通出行指数分析研究[D].北 京：北京工业大学，2012. [YANWP.Evaluationof public transportation index in metropolis[D]. Beijing: BeijingUniversityofTechnology，2012,.]