

基于在线支付数据的公交客流特征分析方法

——以三明市清流县为例

张琪胜

[上海市政工程设计研究总院(集团)有限公司,上海市 200092]

摘要:为了给公交系统规划方案提供数据支撑,以公交在线支付数据为基础,通过数据收集、代刷数据剔除、换乘数据判断、下车站点判断等步骤,结合统计方法,得到了公交分担比例、时间分布、距离分布、线站分布、特定群体公交服务情况等公交运行特征,并以三明市清流县为例,验证了数据分析方法在公交客流特征分析中的适应性。

关键词:在线支付数据;下车站点估算;公交客流特征;公交规划

中图分类号: U941

文献标志码: A

文章编号: 1009-7716(2023)12-0206-05

0 引言

城市公交作为一种低碳、经济、公平的出行方式,在交通规划领域得到越来越多的关注。在公交系统规划过程中需要对公交客流特征进行详细分析。传统方法多基于问卷、跟车调查等方式所获取的数据来分析或主观判断公交客流特征,一般情况下效率较低,成本较高,并且覆盖面有限,客观性不足。随着公交在线支付数据的大规模产生与应用,及其作为结构化数据的全面性、准确性、便捷性、实时性、经济性等特点的逐步展现,公交客流特征的获取已逐步从传统方法转变为依靠在线支付数据分析方法。但是由于公交运营过程中多采用一票制,数据中一般仅包含上车站点信息,缺少下车站点信息,且乘客有可能存在代替他人支付等行为,为直接依靠在线支付数据进行特征分析带来不便。因此,在统计特征前需先进行剔除代刷数据、判断换乘行为、判断下车站点等一系列操作,以形成可用数据。

本文以公交在线支付数据为基础,经数据收集、代刷数据剔除、换乘数据判断、下车站点判断等步骤,结合统计方法,研究了公交运行特征,并以三明市清流县为例,验证数据分析方法在公交客流特征分析中的适应性。

1 数据分析流程及方法

1.1 分析流程

数据分析是指用适当的统计分析方法对收集来的大量数据进行分析,将它们加以汇总和理解并消化,以求最大化地开发数据的功能,发挥数据的作用^[1]。数据分析主要包括数据收集、数据清洗、数据分析这3个层面。数据收集是指根据数据分析的需求收集相关的数据。随着信息化、数字化的发展,公交客流特征分析所能收集到的数据越来越精确和广泛。公交在线支付数据大多为结构化数据,严格遵循数据格式与长度规范,并且公交运营公司逐渐采用数据化管理平台,数据格式规范统一,便于进行统计、筛选、分析。一般情况下,公交在线支付数据包括上车站点、上车时间、交易金额、卡号、线路名称、线路方向等信息,具体数据类型见表1。数据清洗是指对收集的数据进行修正、填补、剔除,如线路、站点名称的统一,时间格式的处理,代刷数据的判断,换乘行为的判断等。数据分析则是通过建立数据模型来提取有用信息,形成分析结论,如判断下车站点、分析客流时变特征、分析线路站点客流分布特征等。公交客流数据分析流程图见图1。

1.2 数据清洗

1.2.1 基础数据处理

公交在线支付数据大多以天为单位,为体现公交特征,保证数据处理量,在数据处理前需要聚合形成月度或年度数据。收集数据可能存在一定程度的重复、无效、异常数据。在进行数据分析之前需要将这类

收稿日期:2022-12-05

作者简介:张琪胜(1994—),男,硕士,工程师,从事交通规划工作。

表1 公交线上支付数据字段属性表

字段	类型	字段	类型
上车站点	Text	卡号	Text
上车时间	Datetime	线路名称	Text
交易渠道	Text	线路方向	Bit
交易金额	Decimal	车牌号	Text

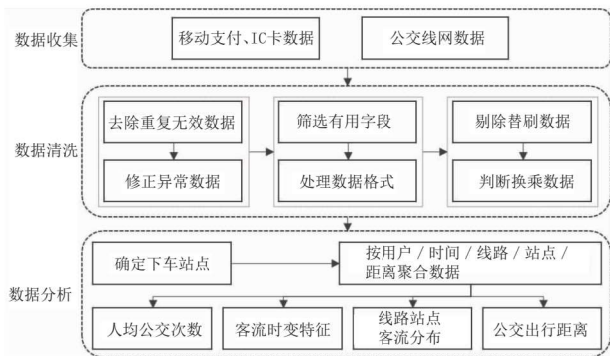


图1 公交客流数据分析流程图

数据进行剔除、修正。重复数据指收集数据的各字段值完全相同的数据；无效数据指关键字段值缺失且无法修补的数据；异常数据指某一些字段值在正常范围外。此外,还存在同一线路名称在各条数据中不一致、同一站点名称在各条数据中不一致、在线支付获取的线路站点名称数据与公交线网中不一致等情况。

关键字段的数据格式也可能存在不一致或不利于下一步处理的情况,因此需要对数据格式进行统一。如数据库中时间数据类型可能为 Datetime、Timestamp 等形式,在数据处理之前需要进行统一。

1.2.2 代刷数据处理

公交支付管理中,公交公司可能存在允许代替他人支付的情况,在数据分析之前,需要剔除这部分代刷数据。判定思路为:同一用户前后两次刷卡的时间差在代刷时间阈值 T_0 内,则前次刷卡记录判定为代刷数据。具体流程为:逐个筛选用户,对某一用户的所有刷卡记录按照上车刷卡时间排序,新增代刷字段 delete,当判定为代刷时赋值为 1,否则赋值为 0。

乘客代刷判定流程见图 2。

1.2.3 换乘数据处理

代刷数据剔除后需要进行乘客换乘行为的判定。判定思路为:设定一次乘车和换乘时间阈值为 T_1 ,若同一乘客同一日相邻两次刷卡时间 T 在 $T_0 < T < T_1$,则后次刷卡记录为前次刷卡记录的换乘乘次^[2]。具体流程为:逐个筛选用户,对某一用户的所有刷卡记录按照上车刷卡时间排序,新增换乘字段

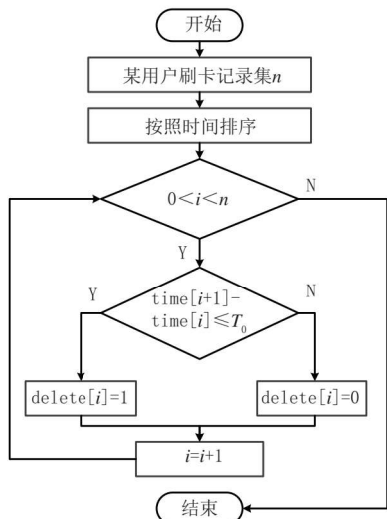


图2 乘客代刷判定流程图

transfer,当判定为换乘时赋值为 1,否则赋值为 0;若 transfer=1,则后次下车站点 end_station[i]等于前次上车站点 start_station[i+1]。

乘客换乘判定流程图见图 3。

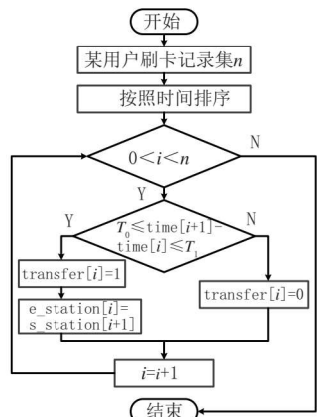


图3 乘客换乘判定流程图

1.3 下车站点估算

一般情况下,公交线上支付数据仅有乘客上车信息,缺少下车信息,这为后续研究公交出行特征带来极大的不便。针对这一问题,提出基于上车站点概率分布的下车站点估算方法。该方法在基础数据处理时需要固定数据格式,剔除替刷数据,判断换乘数据,在模型估算时基于乘客经常上车地点测算概率分布,考虑上下行方向和上下站点间距离约束来估计乘客下车站点。为合理简化研究,提出如下假设:

- (1)根据线上支付数据得到的公交出行特征与总体特征一致。
- (2)代刷数据所表现的公交出行特征与总体特征一致。
- (3)根据下车站点估算的基础集合不包括乘客历史上未曾上车的公交站点。

(4) 公交乘客总是基于公交线网的最短路径出行。

(5) 乘客下车站点符合其在同一线路相反方向上车站点的概率分布。

同时,为方便模型表达,引入如下符号以描述各参数:

n 为乘客乘坐线路的公交站点数量; s_c 为乘客上车站点在该线路相反运行方向上的对应站点; s_k 为乘客乘坐线路相反运行方向上的公交站点, $k=1,2,\dots,n$; p_k 为乘客在 s_k 对应公交站点下车概率, $k=1,2,\dots,n$; freqs_k 为某一固定时间段内(月度或年度),乘客在 s_k 站点的上车次数; dis 为乘客上下站点之间的公交出行距离; dis_0 为乘客乘车的最小公交出行距离; d 为乘客上车线路的站间距集合; e_{station} 为估算的下车站点。

下车站点估算模型为:

$$e_{\text{station}} = \text{randomchoice}(s_k, F\{X=s_k\})$$

s.t.

$$X \sim F\{X=s_k\} = p_k = \frac{\text{freqs}_k}{\sum_{c=1}^m \text{freqs}_k}, k=c+1, \dots, n \quad (1)$$

$$\text{dis}(s_c, s_k) \geq \text{dis}_0 \quad (2)$$

$$\text{dis}(s_c, s_j) = \begin{cases} \sum_{i=1}^{j-1} d^i, & (i, j=1, 2, \dots, n, i < j) \\ \sum_{j=1}^{i-1} d^j, & (i, j=1, 2, \dots, n, j < i) \end{cases} \quad (3)$$

模型的目标为估算乘客下车站点,即在明确乘客所乘公交线路及上车站点的前提下,在上车站点(不包含)至终点(包含)这一样本空间内,按照该乘客在该线路相反运行方向各上车站点概率,取值下车站点。式(1)表示根据乘客在当前线路相反运行方向的上车次数,计算该乘客下车站点在上车站点至终点区间的概率分布;式(2)表示该乘客乘坐公交的出行距离需大于固定值;式(3)表示根据站间距计算的上下车站点间的公交出行距离。

1.4 特征提取方法

基于公交在线支付数据所能得到的信息主要包括公交分担比例、客流时间分布、出行距离分布、线路及站点客流分布和特定群体公交服务情况等特征。其中客流时间分布、出行距离分布、线路及站点客流分布只需根据属性按精度分段统计即可得到。公交分担比例则需结合线下统计数据得到公交出行总量,通过手机信令等手段得到统计范围内总出行量,进而计算二者比例。特定群体公交服务情况主要

根据支付金额分类为免票群体(主要为老年群体)、学生群体、其他群体,进而分析日均客流、人均乘坐次数等指标。

2 案例分析

三明市清流县中心城区范围约 86 km^2 , 2020 年常住人口 5.06 万, 现有常规公交线路 4 条, 总长度 32 km, 建成区线网密度 3.13 km/km^2 。

通过三明市公交管理平台获取月度乘客在线支付数据, 结合线下统计数据计算线上支付率约为 70%, 足以反映总体情况。在剔除代刷数据、判断下车站点后, 进行公交客流特征分析。

2.1 公交出行总量

公交营收数据可以间接反映出行总量特征。从图 4 可以看出, 受新冠疫情影响, 城区公交客流下降明显, 2020 年较前一年缩量 28%; 全年中寒暑假期间为城区公交客流量的阶段低点, 通学和非通学月份公交出行客流总量比例约为 1.4 : 1, 数据表明学生群体在公交乘客群体中占有相当比例。

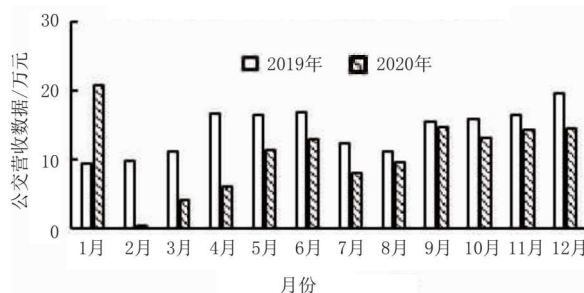


图4 2019~2020年三明市清流县公交营收情况

2.2 出行时间分布

按照数据日期属性聚合可以得到公交客运量的日变情况。2021年5月三明市清流县公交客流日变图见图5; 2021年7月三明市清流县公交客流日变图见图6。

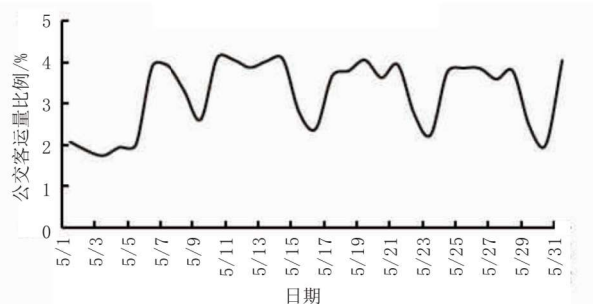


图5 2021年5月三明市清流县公交客流日变图

由图5、图6可知, 三明市清流县公交工作日客运量明显高于节假日和周末, 非假期月份二者比例约为 1.6 : 1, 假期月份二者比例约为 1.2 : 1。

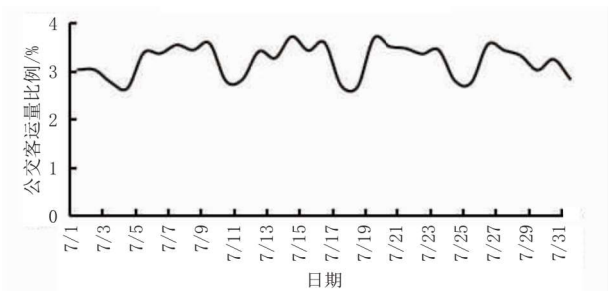


图 6 2021 年 7 月三明市清流县公交客流日变图

按照数据时间属性聚合可以得到客运量的时变情况。三明市清流县公交客运量时变图见图 7。

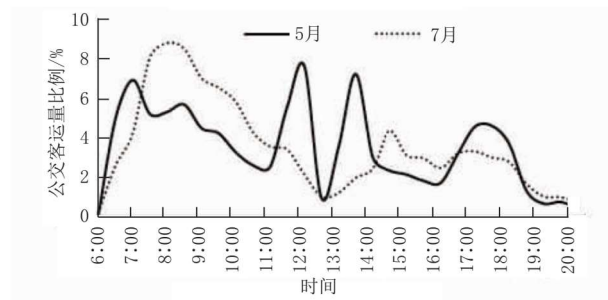


图 7 三明市清流县公交客运量时变图

由图 7 可知,通学月份早高峰时间为 7:00-9:00,午高峰时间为 11:30-12:30、13:30-14:00,晚高峰时间为 17:00-18:00;非通学月份早高峰时间为 7:30-9:00,午高峰时间为 14:00-15:00,晚高峰时间为 17:00-18:00,并且午晚高峰特征不明显。

从客流日变及时变图可以看出,上班及学生群体在公交乘客群体中占有相当比例,通学和非通学期间公交客流特征明显不同。公交部门可据此提供差异化的供给方案,如假期时运力高峰适当延后、午晚高峰适当削减运力。

2.3 出行距离分布

在判断下车站点后,可根据站间距计算公交出行距离。经计算,清流公交出行距离集中在 1~3 km 范围内,平均出行距离约为 2.2 km(见图 8)。一般情况下公交线路合理长度为 5 km~2.5 倍平均乘距^[3],因此公交线路长度可结合城区尺度,以 5~5.5 km 为基础进行调整。

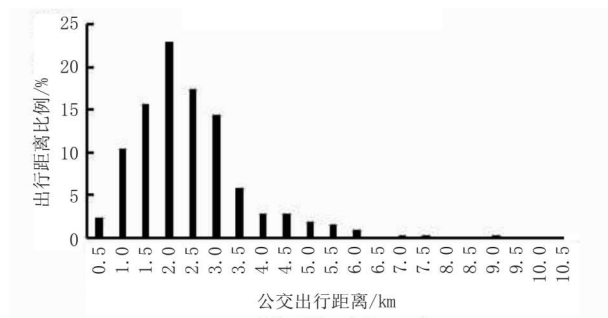


图 8 三明市清流县公交出行距离比例分布图

2.4 线站客流分析

按照数据线路和站点属性聚合可以得到公交线路客流特征。由图 9、图 10 可知,三明市清流县公交线路和站点集中度较高。线路层面,主要集中于 1 路和 2 路,5 路客流量较低,应对该线路进行优化调整。站点层面,联通站、清流国际大酒店车站、建行站工作日客流各占全日总客流量的 10%以上,并且前 5 位的站点客流约占总客流量的 47%(见表 2),应结合建设条件对客流集中站点进行港湾化改造。

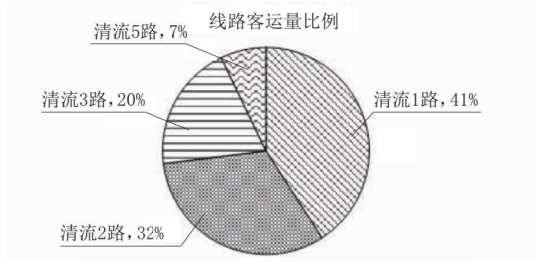


图 9 三明市清流县公交线路客运量分布图

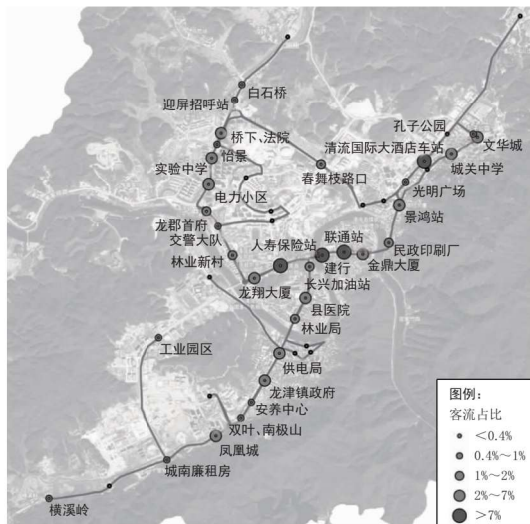


图 10 三明市清流县公交站点客运量分布图

表 2 前 5 位公交站点客流占比情况

排序	站点名称	流量占比 /%
1	联通站	12.1
2	清流国际大酒店车站	11.9
3	建行	10.3
4	人寿保险站	7.3
5	龙翔大厦	5.3
合计		46.9

2.5 特定群体服务分析

根据公交乘客的支付金额以及折扣比例可以间接判定乘客群体,如 8 折票为学生群体,免票一般为老年人群体。以学生群体为例,可以结合学校学生数量、周边站点设置及站点客流等情况分析公交对学

校的服务水平。经计算,城区小学学生 60%以上,中学学生 40%以上居住于城区 4 个社区(渔沧、凤翔、长兴、翠园),居住在城郊各村(城东、城南、桥下、南岐、供坊、大路口、俞坊)的学生极少;同时,居住于城区外部在城区内上学的中学学生比例较高,应设置节假日专线服务于此类群体。

城区各学校学生公交出行来源地比例分布表见表 3。

表 3 城区各学校学生公交出行来源地比例分布表 单位:%

来源地	城关小学	屏山小学	城关中学	实验中学	清流一中	高职	其他
渔沧	31	2	11	3	8	6	12
凤翔	9	16	6	30	27	32	43
长兴	24	10	11	10	16	10	22
翠园	11	30	21	3	10	6	11
城东	4	21	9	3	4	5	3
城南	7	1	4	2	4	6	0
桥下	1	6	0	6	4	3	1
南岐	0	0	0	1	1	1	0
供坊	0	0	1	1	1	1	0
大路口	0	2	2	1	1	0	0
俞坊	0	1	0	0	0	1	0
其他	13	11	35	40	24	29	8
合计	100	100	100	100	100	100	100

此外,城关小学、实验小学、职业中学、清流一中至公交站的步行距离大于 200 m,学生乘坐公交不便;实验小学、实验中学、清流一中、屏山小学的人均公交出行次数低于平均值,应对服务这些学校的公交线路站点进行优化调整,如服务于清流一中的一中站为线路特定绕行途径站点,只能下车不能上车,学生服务多有不便,应优化线路直接途径该站,提升公交对该校服务效率,同时避免线路绕行,具体数据见表 4。

表 4 三明市清流县公交学生群体服务情况表

学校	人数	站点	距学校 / m	人均公交出行次数
城关小学	2 200	联通站:1/2/5 路	250	0.23
		金鼎大厦:1/2/5 路	250	
实验小学 职业中学	2 800	人寿保险站:1/3/5 路	220	0.12
		北山 3:5 路	320	
屏山小学	1 500	春舞枝路口:3/5 路	300	0.06
		屏山小学:3 路	0	
城关中学	1 800	清流国际大酒店车站: 1/2/3/5 路	380	0.31
		城关中学:1 路	0	
实验中学	2 300	实验中学:1/3 路	0	0.08
清流一中	2 700	一中站:2 路	0	0.09
		龙翔大厦:1/3/5 路	450	
平均				0.15

3 结 语

基于在线支付数据的分析方法可以有效提取公交客流特征,并为公交系统优化提供方向。本文以三明市清流县为例,通过数据收集、代刷判断、换乘判断、下车估算等方法构造易于分析的数据结构,基于统计方法分析城市公交出行总量、时间分布、距离分布、线站分布、特定群体服务等情况,为公交系统规划方案提供数据支撑。需要说明的是,公交数据分析还可根据实际情况结合其他多源数据用于公交比例、公交出行目的、公交 OD 与出行 OD 匹配性、运送速度、断面客流等方面,本文限于篇幅不再赘述。

参考文献:

[1] 陶皖.云计算与大数据[M].西安:西安电子科技大学出版社,2017.
[2] 代位.基于刷卡数据和 GPS 数据的公交客流特征分析方法[J].交通与运输,2021,37(2):74-78.
[3] 陈小鸿.城市客运交通系统[M].上海:同济大学出版社,2008.

《城市道桥与防洪》杂志

是您合作的伙伴,为您提供平台,携手共同发展!

欢迎新老读者订阅期刊 欢迎新老客户刊登广告

投稿网站: <http://www.csdqyfh.com> 电话:021-55008850 联系邮箱: cdq@smedi.com