基于多源大数据的城市出行特征研究
——以青岛市为例

王振 张志敏 禚保玲 陈如梦

【摘要】道路网络中运行的公交车、小汽车及出租车能够实时反映城市的运行状态，是一个城市的基本脉动。选取公交刷卡数据、出租车 GPS 数据及车牌识别数据，运用大数据挖掘方法获取不同方式的出行矩阵，借助地图路线导航功能，对不同出行 OD 进行线路规划。通过对路线导航结果进行居民出行特征分析，并从出行距离和时间两个方面得到以下结论：1）9km 为居民出行选择公共交通和私家车的临界优势点，长距离出行选择私家车的频次更高；2）公共交通的服务水平有待提高，公交与小汽车出行时间之比高于深圳市提出实施的“公交提速 1.5 战略”。

【关键词】多源数据；出行矩阵；路线导航；出行特征分析

传统的居民出行特征研究往往基于调查问卷的方式[1-5]，对居民出行耗时、出行距离、出行方式等指标进行统计。伴随着多源数据的发展和分析手段日趋多元化，基于公交刷卡数据[6]、出租车 GPS 数据[7]、车牌识别数据[8]、RFID 数据[9]的居民空间职住和活动特征研究多以单一数据源为研究对象，缺少不同数据之间的相互校核，或者缺少对居民活动时间、距离的分析，或者计算距离采用欧式距离等。基于以往的分析方法，在获取居民出行矩阵后，我们往往通过 Transcad 或者 EMME 建立交通分析模型[10-12]，运用四阶段的方法进行交通流分配，从而得到出行个体的路线；但是此方法存在路网更新慢、交通阻抗参数需反复调整校验等问题，具有一定的局限性。目前，高德地图、百度地图等通过与政府机构及企业合作掌握着城市路网的基础资料和城市运行的诸如速度、时间等数据，且能够根据交通运行情况实时动态的更新和发布交通指数，解决了模型中交通阻抗难以准确获取的问题。目前，高德和百度地图均能够能够提供基于 GPS 数据的路线导航服务，并对外开放 API 接口，用户可以根据出行起终点数据，在指定出行方式和出行策略后，申请获取对应路线的出行时间、距离及路径等信息，在一定程度上解决了以往获取出行矩阵后无法获取实际路径的问题。

本文拟基于公交车、出租车及车牌卡口某一天的数据，利用数据挖掘方法分析方法对城市运行中的多源数据进行处理，获取出行时空矩阵，借助地图路线导航功能最终得到实际出行路线。

1 数据处理技术路线

公共车和小汽车作为城市交通机动化运行的主要方式，具有出行量大、覆盖范围广的特点。如下图所示，基于公交刷卡（IC）和到离站（AVL）数据、小汽车卡口过车数据以及出租车 GPS 数据等多源海量数据，通过 PYTHON、SQL 等分析工具提取不同出行方式的出行 OD 矩阵，运用地理编码工具获取起终点对应的经纬度，在高德地图路线导航服务下进行路线规划，进而对不同方式的出行特征进行分析。

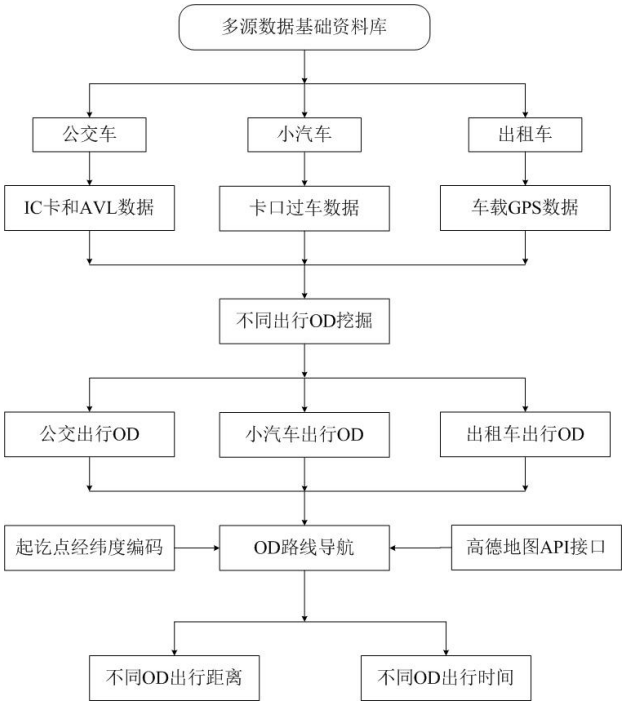

图 1 多源数据分析技术路线图

2 多源数据 OD 矩阵挖掘

2.1 出租车 GPS 数据

出租车 GPS 数据又称浮动车交通信息采集系统（简称浮动车系统，FCD），采集时间为 2015 年某工作日，主要字段属性如表 1 所示，包括车牌号码（CPH）、上传时间（SJ）、速度（SD）、经度（JD）、纬度（SD）、运行状态（ZT）等，时间发送间隔为 30s。

表 1 车辆 GPS 数据结构表

CPH	WD	JD	SD/km	ZT	SJ
鲁 BT02**	36.07506	120.403	4.9	0	2015/3/2 13:23
鲁 BT02**	36.0742	120.3984	0	0	2015/3/2 13:26
鲁 BT02**	36.06782	120.396	28.7	1	2015/3/2 13:27

提取出租车 OD 矩阵的步骤如下所示：

Step1 基于 Python 的 Pandas 模块将一天的数据读入一个 Dataframe，以车牌号（CPH）、WD（纬度）、经度（JD）、速度（SD）、载客状态（ZT）、采集时间（SJ）为 key 值。

Step2 将所有数据分别按照车牌号和采集时间升序排列，得到每辆出租车一天运行轨迹的时间序列。

Step3 对同一车牌号对应的全天记录，基于载客状态字段推断乘客上下客点，其中 0 表示空车，1 代表载客：

1）状态由 0 变为 1 对应记录所在的位置即为乘客的上客点（O）；

2）状态由 1 变为 0 对应记录所在的位置即为乘客的下客点（D）。

Step4 对上一步获取的上、下客点根据时间顺序匹配出 OD，并进行有效性分析：若下客点与上客点对应的时间之差小于 10 分钟，将作为无效数据进行剔除。

Step5 将不同车牌号的出行 OD 进行汇总，即得到市域所有的出行 OD 数据。

2.2 公交数据

公交刷卡（IC 卡）数据采集时间为 2015 年某工作日，字段属性如表 2 所示，包括卡号、交易日期时间，线路名称、线路号、POS 机编号等信息。

表 2 公交车刷卡（IC 卡）数据结构表

线路编号	线路名称	POS 机编号	卡号	交易日期时间
000878	温馨巴士西海岸站	370020022382	2660000004691365	6:04:35
000878	温馨巴士西海岸站	370020022382	2660000000492582	6:05:08
000878	温馨巴士西海岸站	370020022382	2660000004025960	6:06:28

公交到离站（AVL）数据采集时间为 2015 年某工作日，字段属性如表 3 所示，包括线路、车牌号、报站时间、到离站编号、站点编号、站点名称等信息。

表 3 公交车到离站（AVL）数据结构表

车牌号	线路	站点	经度	纬度	报站时间
鲁 BJ6011	773	蔚蓝群岛北	120/20/5232E	36/15/5690N	22:47:31
鲁 BJ6215	773	蔚蓝群岛北	120/20/5203E	36/15/5699N	22:27:25
鲁 BJ6206	773	蔚蓝群岛北	120/20/5227E	36/15/5687N	21:28:32

青岛市公交车均为上车刷卡，因此上客站点 O 即为刷卡站点，但需要通过关联 IC 卡数据和 AVL 数据获取对应的站点信息，提取 OD 矩阵的步骤如下：

Step 1 通过 Python 的 Pandas 模块将 IC 卡数据读入 Dataframe，并以 IC 卡编号和交易时间进行升序排列，得到每个 IC 卡一天所有的刷卡记录。

Step 2 同样的方法导入车辆电子路单数据，并以车牌号字段进行升序排列，得到每辆车一天的 GPS 运行轨迹点位。

Step 3 对每一条 IC 卡数据，通过 POS 编号字段与到离站数据的车牌号字段的对应关系找到此刷卡记录对应车辆的运行轨迹点位；并找到时间最接近的站点为对应的刷卡站点。

乘客下客站点判断基于当日乘客前一次出行的终点是其进行下一次出行的起点，相关步骤如下：

Step 4 对上述 Step3 匹配后的数据，针对同一个 IC 卡号，若线路编号相同则认为第二条记录的上车站点为第一条线路的下客站点；

Step 5 若线路编号不相同，则计算第一条记录中线路各站点与第二条记录中上车站点的最短距离；

Step 6 若最短距离小于 1km，则认为第一条线路中的对应站点即为下客站点；

Step 7 继续进行下一条记录的判断，直至最后一条数据。

2.3 车牌识别数据

车牌识别数据采集时间为 2015 年某工作日，字段属性如表 4 所示，包括车牌号(CCARNUMBER)、数据采集时间(DCOLLECTIONDATE)、卡口编号(CADDRESSCODE)以及卡口名称(CCOLLECTIONADDRESS)等信息，其中通过卡口编号和卡口点位数据表(表 5)能够获取卡口对应的经纬度。

表 4 车牌识别数据结构表

CCARNUMBER	DCOLLECTIONDATE	CADDRESSCODE	CCOLLECTIONADDRESS
鲁 M4**2 挂	2017/6/6 4:45	611261001000	黄张路（S328)第九中学
鲁 B7**8E	2017/6/6 4:45	611301003000	漓江西路与青云山路路口
鲁 DW**9H	2017/6/6 4:45	102040237650	G204 烟上路-九赵路

表 5 卡口点位数据结构表

POINTCODE	POINTNAME	LONGITUDE	LATITUDE
10000205057	杭鞍高架路(温州路-人民路)	120.34921	36.09716
10000205063	杭鞍高架路(山东路-南京路)	120.37147	36.09524
30000205037	杭鞍高架路(人民路-鞍山一路)	120.35879	36.09464

车牌识别数据分析步骤如下：

Step 1 通过 Python 的 Pandas 模块将车牌识别数据读入 Dataframe，并以车牌号码和抓拍时间进行升序排列，得到每辆车一天所有的被抓拍记录。

Step 2 根据号牌类型字段刷选出小汽车的抓拍记录；并将车牌号码中前三位为“鲁 BT”和“鲁 UT”的出租车记录剔除。

Step 3 根据同一辆车被抓拍的记录大于 1 次，且分析时段（早高峰）内第一次与最后一

次被抓拍的时间差大于的 15 分钟两个标准识别出车辆的起终点。

3 OD 路线导航分析

3.1 地图 API 路线导航规划

将出租车 GPS 数据、公交车刷卡数据以及车牌识别数据提取的 OD 进行汇总，格式如表 6 所示，主要包含起终点的经纬度、时间及从起点至终点所采用的交通方式。

表 6 车辆 OD 数据结构表

OD_ID	O_WD	O_JD	O_SJ	D_WD	D_JD	D_SJ	Type
1	36.10413	120.4562	8:30:13	36.303266	120.399002	8:56:39	Taxi
2	36.29539	120.391	9:04:55	36.189841	120.402439	9:29:36	Bus
3	36.16272	120.4184	9:45:11	36.126918	120.420541	9:57:25	Car

通过调用高德地图路线规划 API (Application Programming Interface)接口，输入表 6 中的每条 OD 记录对应的坐标、时间、方式等参数，在时间最短的策略下，进行出行路线规划，得到每条 OD 的出行距离和耗时等数据，进而可以获取每条线路的导航信息，如表 7 所示。

表 7 OD 路线导航结果

OD 编号	出行耗时/s	出行距离/m	出行方式
1	1231	6418	小汽车
2	1298	6434	出租车
3	2219	10390	公交车

3.2 出行特征分析

城市范围内的出行受城市尺度的影响，距离分布相对集中，统计发现公交车有 95%以上的出行在 20km 以下，而小汽车有 85%的出行在 20km 以下，因此从数据分布集中的角度考虑，本文仅对 20 公里以下的出行进行研究。

表 8 不同方式的出行距离指标统计（km）

index	bus	car	taxi
mean	6.6	8.1	7.2
std	4.1	4.8	3.8
25%	3.4	4.4	4.7
50%	5.7	7.5	6.7
75%	9.1	11.2	9.2
count	125549	19043	20012

通过对 125549 次公交出行 OD、19043 次小汽车出行 OD、20012 次出租车出行 OD 导航结果进行分析，公交车的平均出行距离最短，为 6.6km，高于青岛市第三次交通出行调

查（2016 年）[13]的 5.7km。出租车作为公共交通的组成部分，其服务的空间范围是有别于公交车的，在平均出行距离上要高于公交车，为 7.2km。如表 8 所示，公共交通出行中（含公交车和出租车），高达 75%的出行是小于 9km，在一定程度上反应了距离对于居民出行方式选择的敏感性；与此同时，从图 2 不同出行方式的距离分布图上可以看出，从 9km 开始小汽车出行的频率开始显著高于公交和小汽车出行，表明居民出行选择私家车和公共交通工具存在竞争优势的临界点，9km 以下居民更多选择公共交通工具，大于 9km 则选择私家车的比例占比更高。

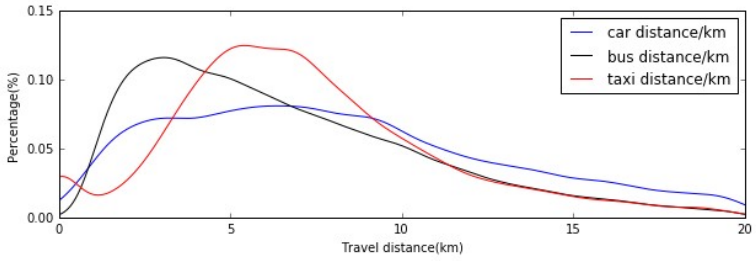

图 2 不同出行方式的出行距离分布图

同时从图 2 可以看出，不同出行方式分布规律具有一定的差异性，公共交通（含公交车和出租车）为单波峰形式，且公交车波峰对应的距离要低于出租车；小汽车的出行分布更为均衡，标准差高于公交车和出租车，这由小汽车灵活方便的适用性决定。

表 9 不同方式的出行耗时指标统计（min）

index	bus	car	taxi
mean	38.1	21.1	20.7
std	12.2	11.1	9.7
25%	29.0	12.7	14.4
50%	36.4	20.0	19.5
75%	45.5	28.5	26.5
count	125549	19043	20012

从出行时间上来看，如表 9 所示公交车的平均出行时间为 38 分钟，是小汽车平均出行时间的 1.9 倍，高于深圳市提出实施的“公交提速 1.5 战略”，即公交出行时间降低至小汽车 1.5 倍以内。同时我们从图 3 可以看出，受公共交通服务水平的影响，公交车出行在 15 分钟以内的占比几乎为 0，这与公交车候车时间的密切相关。

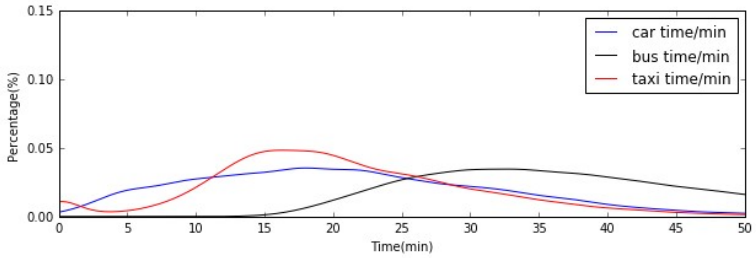

图 3 不同交通方式的出行时耗分布图

4 结论

基于海量多源传统交通运行数据，在传统数据挖掘方法和互联网分析技术的基础上，比较分析了公共交通和私家车不同出行方式的交通特征。通过研究发现，公共交通和小汽车在出行距离和出行时间均呈现不同的分布特点，公共车的平均出行距离为 6.6km，略高于青岛市第三次交通出行调查（2016 年）的 5.7km；公交车的平均出行时间为 38 分钟，是小汽车平均出行时间接近两倍。本文提供的方法在一定程度上解决了传统 OD 分析难以获取出行距离和时间的问题，但是同时也存在路线导航与实际路线可能会有偏差，这对宏观尺度上的城市规划研究是在可接受的误差范围内，同时也优于传统简单的距离和时间测算。

参考文献

[1] 千迪，王德，朱玮．上海市近郊大型社区居民的通勤特征——以宝山区顾村为例[J]．地理研究，2015，34(8)：1481-1491.

[2] 文婧，王星，连欣．北京市居民通勤特征研究——基于千余份问卷调查的分析[J]．人文地理，2012(5)：62-68.

[3] 刘定惠，朱超洪，杨永春．西部大城市居民通勤特征及其与城市空间结构的关系研究——以成都市为例[J]．人文地理，2014，29(2)：61-68.

[4] 王德起，许菲菲．基于问卷调查的北京市居民通勤状况分析[J]．城市发展研究，2010，17(12)：98-105.

[5] 夏晓敬，关宏志．北京市老年人出行调查与分析[J]．城市交通，2013(5)：44-52.

[6] 龙瀛，张宇，崔承印．利用公交刷卡数据分析北京职住关系和通勤出行[J]．地理学报，2012，67(10)：1339-1352.

[7] 相恒茂，姜伟，高浠舰，等．基于出租车上下客点的城市人口流动分析[J]．测绘与空间地理信息，2016(7)：31-33.

[8] 马金麟, 张宗博, 谢君平,等. 基于车牌识别数据的车辆 OD 矩阵获取研究[J]. 重庆理工大学学报:自然科学, 2017, 31(7):48-55.

[9] 赵晓晓, 杜威, 周旭,等. 基于 RFID 的城市路网 OD 矩阵获取方法及时空特性分析[J]. 交通信息与安全, 2016(1):30-36.

[10] 宋睿, 晏克非, 郑建. TransCAD 在四阶段交通需求预测中的应用[J]. 交通科技与经济, 2011, 13(1):79-81.

[11] 刘桢根, 卢士和. TransCAD 软件在交通分配中的应用[J]. 交通工程, 2006(7):40-42.

[12] 李国才, 秦文军, 梁成文,等. EMME/2 在沈阳市交通规划中的应用(英文)[J]. 上海理工大学学报, 1999(3):277-281.

[13] 青岛市城市规划设计研究院.青岛市第三次交通出行调查（2016 年）[R].青岛: 青岛市地铁工程建设指挥部, 2016.

作者简介

王振，男，硕士研究生，青岛市城市规划设计研究院，工程师。电子信箱：2227840807@qq.com

张志敏，女，硕士研究生，青岛市城市规划设计研究院，高级工程师。电子信箱：08010310126@163.com

禚保玲，女，硕士研究生，青岛市城市规划设计研究院，工程师。电子信箱：08010310126@163.com

陈如梦，女，本科，山东建筑大学。电子信箱：2319565731@qq.com