城市公共交通系统通勤出行特征提取分析方法*

翁剑成[1] 涂 强[2] 袁荣亮[2] 王月玥[3]

（1. 北京工业大学交通工程北京市重点实验室，100124，北京；2. 北京市城市规划设计研究院，100044，北京；
3. 北京市轨道交通指挥中心，100101，北京//第一作者，副教授）

摘 要 公共交通出行者的出行特征是地铁及公交线网规划与运营优化的重要依据。基于多模式刷卡数据，提出城市公共交通系统出行链提取方法，利用存在换乘的出行链调查数据进行验证，提取成功率达96.1%。基于出行者历史刷卡数据构建了多种机器学习分类器以识别通勤人群，经过精度比较，发现随机森林分类器效果最优，准确度达99.96%。利用分类器和出行链提取方法，对北京市公共交通系统出行链结构、换乘特征等进行初步分析。该方法可以有效提取分析通勤人群出行特征，为公共交通系统方案的优化提供数据支持。

关键词 公共交通；地铁；出行特征；出行链；机器学习

中图分类号 U491.1

DOI：10.16037/j.1007-869x.2019.06.015

Characteristic Extra ction and Analysis of Commute Trip by Public Transit System

WENG Jiancheng，TU Qiang，YUAN Rongliang，WANG Yueyue

Abstract The trip characteristic by public transit is an important basis of public transit network planning and operation optimization. Relying on the multiple-mode smart card data，an extraction method of public transit trip chain（PTTC）is proposed，and the extraction success rate is 96.1%. Classifiers of machine learning are formed to distinguish the commuters and non-commuters based on the history card data of passengers. Through accuracy comparison，the Random Forest model classifier presents the optimal effect with 99.96% accuracy. The structure of PTTC and the transfer characteristics in Beijing are analyzed by using the proposed classifier and trip chain extraction method，which can effectively extract and analyze the characteristics of commuters and provide support for the optimization of public transit system.

Key words public transit；metro；trip characteristic；PTTC；machine learning

First-author′s address Beijing Key Laboratory of Traffic Engineering，Beijing University of Technology，100124，Beijing，China

随着智能公共交通系统的不断发展及其技术突破，城市公共交通运行、服务等方面的动态数据持续积累，智能卡刷卡交易及车辆GPS（全球定位系统）位置等数据已形成了海量的规模。

基于良好的数据基础，很多学者利用智能卡数据在公共交通用户出行行为分析方面做了大量研究，主要包括出行者的出行起点/终点（OD）、出发时间、行程时间和换乘特征等方面。文献[1]利用伦敦市的公交智能卡数据，研究了地铁与其他公交系统换乘之间出行阶段的连接时间阈值。文献[2]基于韩国智能卡数据记录信息，对乘客公共交通出行时间及换乘特征进行了分析。文献[3]基于智能卡数据，提出了用于预测公共交通出行者的活动目的、出行地点、出行时间、持续时间的方法。文献[4]利用刷卡数据揭示了深圳市通勤人群在出发时间、出行耗时、换乘特征等方面的规律。文献[5]提出了基于智能卡信息采集技术的公交客流及出行信息的分析方法。文献[6]提出基于多源数据的公共交通通勤出行特征提取方法，但通勤人群识别仅依据一周的出行频次，可靠度较低。文献[7]利用智能卡和问卷调查数据，建立了基于决策树模型的通勤人群分类器，精度较高。

这些研究都是基于智能卡数据，以单次刷卡记录为研究对象对公共交通出行者出行行为进行详细分析。然而，在城市公共交通系统网络化、出行模式多样化的背景下，应重点解决换乘行为的判别问题，注重从“完整出行”的角度研究乘客的出行行为，以期更客观、准确地描述出行者的出行特征与需求时空分布。此外，与非通勤人群相比，通勤人群在换乘特征、出行频率等方面有明显的差异性。现有研究在出行行为分析时缺乏对不同出行者的科学分类，无法确切表达通勤人群的出行特点与资源时空需求。

本文拟利用海量的智能卡交易数据，研究城市

* 国家自然科学基金（51578028）；交通运输部科技计划（2015318221020）

公共交通系统出行链的连接方法，并引入机器学习方法进行通勤人群判别，为实现公共交通出行行为的精细化分析提供技术支持，为城市公共交通规划与管理提供更为准确的指导。

1 数据基础与预处理方法

公共交通刷卡数据是进行出行链提取、出行者类型识别及出行行为特征分析的基础，包含道路公交智能卡数据和城市轨道交通自动售检票(AFC)系统刷卡数据两种来源。

1.1 城市轨道交通及道路公交刷卡数据特点

目前，城市轨道交通 AFC 系统主要用以记录用户卡号、进出站点编号及时间等信息。由于乘客在轨道交通网络内部换乘时不需要再次刷卡，因此 AFC 数据无法直接记录乘客在轨道交通系统内部的换乘行为，但根据其出行轨迹可获取不同出行 OD 所对应的换乘次数。

本文以北京市 2014 年 9 月的公共交通刷卡数据作为研究基础。2014 年北京市道路公共交通系统的计费方式同时包括一票制和分段计价制两种，可覆盖大多数城市的情况，具有普适性。其中：一票制只能准确记录乘客的上车站点信息，下车时间和站点位置缺失；分段计价制虽然上下车均需刷卡，但上车站点信息缺失的现象明显，且准确度较低，往往只有下车站点信息相对准确。以往基于智能卡数据推算道路公交上下车站点信息的研究较多，在此不作为重点研究对象。

1.2 数据预处理与整合步骤

为了完整分析公共交通出行者的出行过程，按照以下步骤剔除与出行特征分析无关的数据字段，并对轨道交通和道路公交的异源数据进行整合。

(1) 关键字段提取：从道路公交和轨道交通刷卡数据库中提取与出行特征相关的字段，包括用户卡号、进出线路号、进出站车站编号、进出站时间等 7 个有效字段；

(2) 数据整合：以卡号为关联条件，将同一用户的刷卡记录按照刷卡时间排序，为一票制、分段计价制道路公交和轨道交通线路等 3 类数据增加出行阶段类型的数据标记，分别记为 B1、B2 和 R。公共交通刷卡数据整合表如表 1 所示。

表 1 公共交通刷卡数据整合表

记录	一卡通卡号	上车时间	下车时间	上车线路	上车站号	下车线路	下车站号	车辆编号	标记类型
1	00001098	2014-04-15 15:39:08	2014-04-15 16:11:50	10	29	6	31		R
2	00001098	2014-04-15 8:49:31	2014-04-15 8:49:31	408	4	408		00083051	B1
3	00001098	2014-04-15 9:56:12	2014-04-15 9:56:12	103	18	103	6	00083054	B2

注：第 2 条记录上下车时间相同，是由于一票制(B1)只能记录上车时间，下车时间信息缺失；第 3 条记录上下车时间相同，是由于分段计价制(B2)只能准确记录下车时间，上车时间缺失

2 城市公共交通出行链结构提取方法

将城市公共交通出行链定义为从出行的起始站点到目的站点，由一个或多个地铁及道路公交的出行阶段按照时间顺序组成的一次完整的出行过程。其中，一个出行阶段指从道路公交出发站点刷卡上车或轨道交通进站起，经过在途出行(可包含轨道交通内部换乘)后刷卡下车或出站的过程。因此每一条刷卡数据记录都可表示一个出行阶段。出行阶段与出行链示意图如图 1 所示。

2.1 出行链结构提取方法

基于经过整合的公共交通刷卡数据，将所有刷卡记录按照时间顺序进行排序，利用一卡通卡号字段锁定同一用户，根据相邻出行记录时间差进行换乘关系识别，划分或者连接该用户的所有出行阶段。

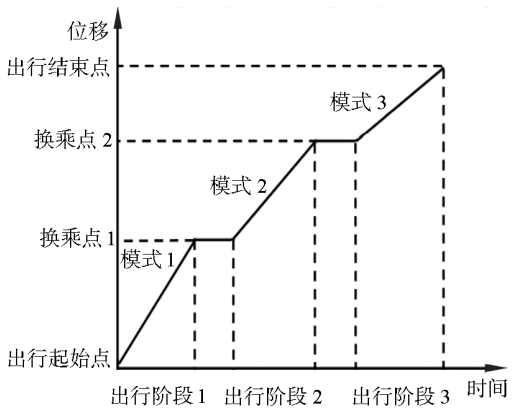

图 1 城市公共交通出行链二维结构图

由此方法确定的公共交通出行链可由一个或多个出行阶段组成，设第 i 个出行阶段的上、下车(或进、出站)刷卡时间分别为 T_{i_ON} 和 T_{i_OFF}，则相邻出行阶段之间的换乘时间可由 $T_{i+1\_ON}-T_{i\_OFF}$ 表示(见图 2)。

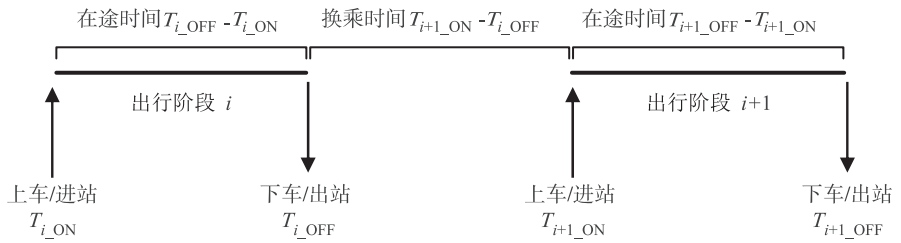

图 2　前后两个出行阶段时间分布示意图

2.1.1　换乘关系判别阈值

在刷卡数据中，单次（一票制）刷卡道路公交只记录上车时间 T_{B1_ON}、双次（分段计价制）刷卡道路公交只记录下车时间 T_{B2_OFF}，轨道交通同时记录进站时间 T_{R_ON} 和出站时间 T_{R_OFF}。因此，三种出行模式间的换乘交易时间差阈值包含了不同的时间组成，部分换乘结构的交易时间差阈值中包含公交在途时间（见表 2）。

根据道路公交站点服务水平及轨道交通站点吸引范围的相关研究[8-9]，确定道路公交与道路公交、道路公交与轨道交通间在理论上可接受的最大换乘时间（不含在途时间）。根据所有道路公交线路的运营里程和高峰时段的运行速度，确定 B1 或 B2 理论上的最大可接受在途时间。根据不同的公共交通换乘模式，共划分出 8 种换乘类型。选取一个月的多模式刷卡数据（约 1 500 万条/日），连接同一卡号用户的相邻出行阶段，分别计算这 8 种换乘类型的交易时间差。基于累计频率在 95%位的刷卡实际交易时间差，确定各换乘类型的交易时间差阈值（部分包含在途时间），如表 2 所示。

表 2　8 种出行阶段连接类型换乘关系判别实际交易时间差阈值

序号	换乘类型	交易时间差	时间组成	实际交易时间差阈值 /min
1	R−B1	$T_{B1\_ON} - T_{R\_OFF}$	轨道交通换乘道路公交时间　$T_{i+1\_ON} - T_{i\_OFF}$	<23
2	R−B2	$T_{B2\_OFF} - T_{R\_OFF}$	轨道交通换乘道路公交时间+B2 在途时间　$T_{i+1\_OFF} - T_{i\_OFF}$	<104
3	B2−R	$T_{R\_ON} - T_{B2\_OFF}$	道路公交换乘轨道交通时间　$T_{i+1\_ON} - T_{i\_OFF}$	<20
4	B1−R	$T_{R\_ON} - T_{B1\_ON}$	B1 在途时间+道路公交换乘轨道交通时间　$T_{i+1\_ON} - T_{i\_ON}$	<56
5	B2−B1	$T_{B1\_ON} - T_{B2\_OFF}$	道路公交换乘道路公交时间　$T_{i+1\_ON} - T_{i\_OFF}$	<25
6	B1−B1	$T_{B1\_ON} - T'_{B1\_ON}$	B1 在途时间+道路公交换乘道路公交时间　$T_{i+1\_ON} - T_{i\_ON}$	<68
7	B2−B2	$T_{B2\_OFF} - T'_{B2\_OFF}$	道路公交换乘道路公交时间+B2 在途时间　$T_{i+1\_OFF} - T_{i\_OFF}$	<112
8	B1−B2	$T_{B2\_OFF} - T_{B1\_ON}$	B1 在途时间 + 道路公交换乘道路公交时间 + B2 在途时间　$T_{i+1\_OFF} - T_{i\_ON}$	<137

注：序号 6 中 T'_{B1_ON} 表示 B1−B1 换乘类型时的第一个出行阶段的上车时间，以示与第二个出行阶段的上车时间 T_{B1_ON} 的区别。序号 7 类同

2.1.2　出行链结构提取

基于城市公共交通系统内各换乘关系的时间判别阈值，可实现出行链结构的提取。在表 1 的基础上增加以下标记字段："CHAIN"代表该刷卡记录处于该公共交通卡用户的第 i 条出行链；"JS"代表该刷卡记录处于所属出行链的第 k 个阶段。基于公共交通卡卡号及上车时间字段，对表 1 中的刷卡记录进行排序，采用图 3 所示流程即可实现出行链结构的提取与标记。

2.2　出行链结构提取方法验证

选取 396 名志愿者，记录他们连续 10 个工作日的公共交通出行过程，包括通勤出行和非通勤出行。志愿者按照出行次序，完整记录每一次出行过程包含的所有乘车信息，包括乘坐的交通方式、线路号、上车和下车站点及刷卡时间等。

根据志愿者卡号，匹配刷卡交易记录中的数据，进行出行链提取，对比模型提取结果与实际出行过

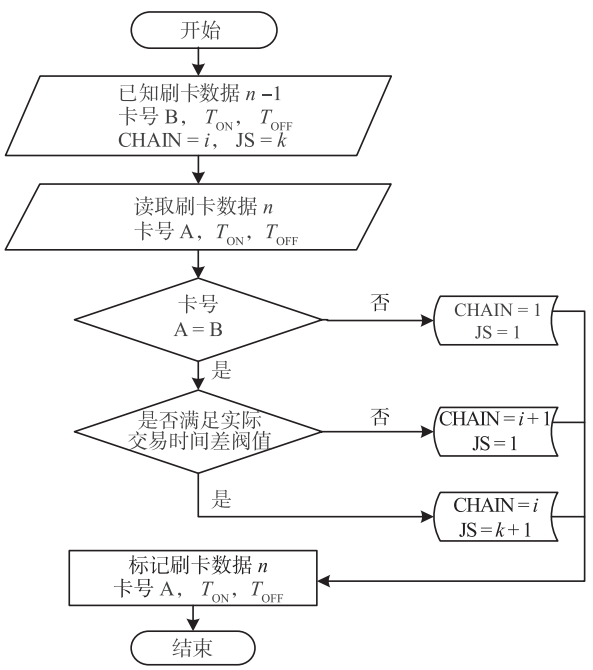

图 3　公共交通出行链结构信息标记流程

程的吻合程度。共记录了 284 条包含换乘的出行链，包含 577 个出行阶段，共有 15 种出行链结构。验证结果显示，模型的出行链结构提取成功率为 96.1%（见表 3）。道路公交出行记录信息的不完备造成部分换乘交易时间差阈值中包含了道路公交在途时间，对出行链提取的准确度影响较大。但目前的提取成功率可以基本满足换乘特征分析的精度要求。

表 3　公共交通出行链结构提取成功率

序号	出行链结构	实际调查数量	成功提取数量	提取成功率/%
1	R−B1	36	36	100
2	R−B2	33	31	94
3	B2−R	35	35	100
4	B1−R	22	21	96
5	B2−B1	42	42	100
6	B1−B1	41	37	90
7	B2−B2	32	30	94
8	B1−B2	34	32	94
9	B1−B1−B2	2	2	100
10	B1−B1−B1	1	1	100
11	B1−B2−B1	1	1	100
12	B1−R−B2	1	1	100
13	B1−R−B1	2	2	100
14	B2−R−B1	1	1	100
15	B2−R−B2	1	1	100
合计		284	273	96

3　基于机器学习的通勤人群鉴别

在数据挖掘技术中，“分类识别”十分重要且具有广泛的应用价值。目前，机器学习分类器的核心算法种类多样[10]，各类算法的分类原理、适用范围和精度特点各有差异。

机器学习分类器的建立过程可分训练和测试两部分，构建过程与步骤如图 4 所示。

3.1　样本数据采集与预处理

采用网络问卷、现场调查等方式，并通过对调查样本一周刷卡数据特征进行校验，最终确定了 978 位公共交通出行者为样本人群，其中包括 490 位通勤出行者和 488 位非通勤出行者。

为了使分类器能够了解各类出行人群的出行特征，从而增强分类器的泛化性和推广性，在基础数据选择时需要注重样本的多样性。因此，在选择样本数据时，考虑了出行人群在性别比例、年龄结构和出行结构等方面的均衡性。此外，在样本数据选择时还考虑了样本数据的出发时间、出行距离和出行时间等要素。

通勤出行具有以下特点：出行的往返性、出发时间的规律性、出行方式选择的固定性、线路选择的多样性。而非通勤出行的上述特征并不明显。因此，可选取上车和下车刷卡时间、上车和下车线路编号、上车和下车站点编号作为特征值描述每个公共交通出行者的出行特征。

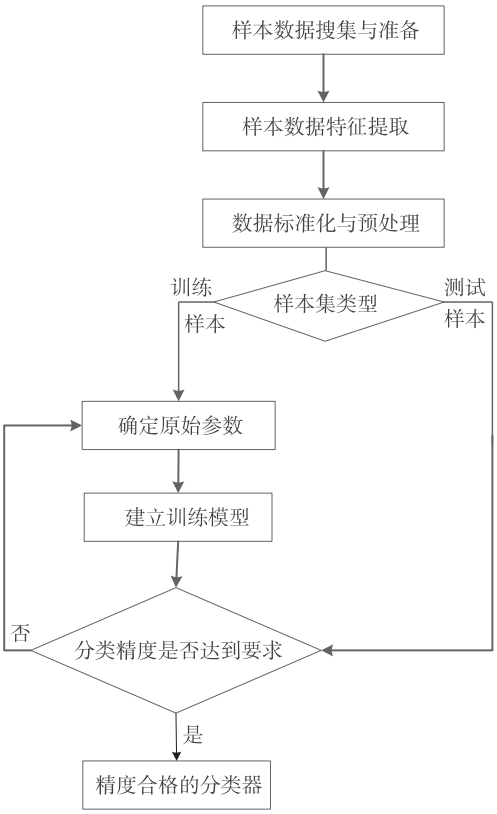

图 4　机器学习分类模型的建立过程

3.2　模型训练

（1）训练与测试集准备：将全部样本数据按照 7:3 的比例随机划分为训练集与测试集。

（2）参数标准化：根据刷卡样本数据的特点，选取 Z-score 标准化方法，以消除分类数据特征参数的量纲和自身变异对数据大小的影响。

（3）模型训练：选取多种机器学习算法进行模型训练，包括决策树（Decision Tree）、逐步增强法（AdaBoost）、感应器（Perception）、支持向量机（SVM）、随机森林（Random Forest）、梯度提升树（Gradient Boosting Tree）等，基本涵盖了常用的机器学习算法。

（4）模型评价：采用分类准确度 A、召回率 R 和精准度 P 来评估模型的分类效果。三个参数的计算公式如下：

$$A = \frac{T_P + T_N}{P_S + N_S} \quad (1)$$

$$R = \frac{T_P}{P_S} \tag{2}$$

$$P = \frac{T_P}{T_P + F_P} \tag{3}$$

式中：

P_S——通勤人群的样本数量；

N_S——非通勤人群的样本数量；

T_P——可正确识别的通勤人群的数量；

T_N——可正确识别的非通勤人群的数量；

F_P——把非通勤人群识别为通勤人群的数量。

基于测试集的293个样本采用不同的算法进行模型评价，计算结果如图5所示。结果显示，随机森林算法的分类准确度最高，达99.96%，且召回率和精准度也明显高于其他算法。与已有的基于决策树的通勤人群鉴别方法[7]（准确度98.1%，召回率81.0%）相比，模型精度有明显提升。因此，随机森林算法在出行人群分类中具有最好的适用性，可实现高精度的通勤人群鉴别。

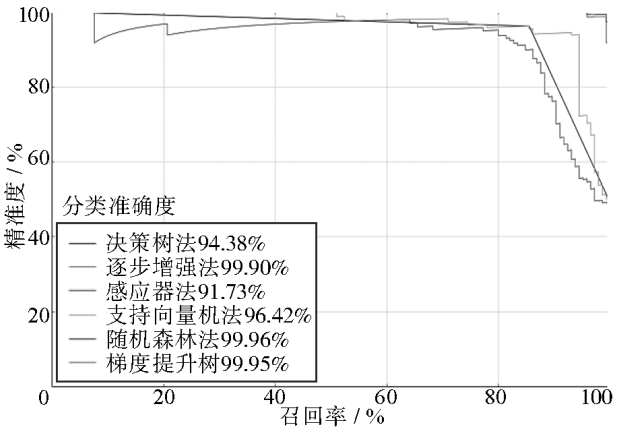

图5 机器学习分类效果评价截图

4 案例分析

利用提出的出行链提取方法和基于机器学习的出行人群分类模型，选取了北京市2014年9月一周的公共交通刷卡数据（当时尚未实施公交票改，数据普适性较好），对公共交通的出行人群结构、出行链与换乘特征进行了初步分析。

4.1 出行人群结构分析

通过分析可知，北京市每天采用公共交通通勤出行的人数在270万左右，出行量较为稳定，占公共交通日均出行总人数的52.5%。

在公共交通出行资源使用方面，通勤出行的公共交通使用频次明显高于非通勤出行。通勤出行日均刷卡次数为750万次，占刷卡总量的58.6%。同时，一周的不同工作日，通勤人群的出行特征和构成比例也相对稳定。

4.2 出行链结构分析

通勤人群出行链结构特征如表4所示。由表4可知：无换乘出行链（不包含轨道交通线网内的换乘）的通勤人群占比约为66.4%。此外，在含有轨道交通模式的通勤出行链中，约有28%的通勤者乘坐轨道交通前后需要采用道路公交接驳的方式完成出行，这反映北京市轨道交通线网在可达性方面有待提高。变异系数表示各结构类型出行链数量在统计期内的稳定程度。结果表明，不同结构类型的出行链在每天的数据量和占比中均较稳定，变异系数均在3%以内。

表4 通勤人群出行链结构特征分析表

结构类型	一周日均数量/万条	变异系数/%	通勤人群占比/%
R	182.6	1.42	34.93
B	164.8	2.16	31.52
B−B	64.3	1.84	12.30
R−B	43.5	1.55	8.32
B−R	40.3	1.99	7.70
B−B−B	9.3	1.86	1.77
B−R−B	8.4	2.84	1.60
其他	9.7	2.56	1.86

4.3 换乘特征分析

出行者平均换乘系数是衡量出行直达程度、反映乘车方便程度的指标。换乘系数越低，表明乘客出行直达程度越高，计算方法如下：

$$\text{乘客平均换乘系数} = \frac{\text{出行链总数+换乘总次数}}{\text{出行链总数}} = \frac{\text{出行阶段总数}}{\text{出行链总数}} \tag{4}$$

本案例的乘客平均换乘系数计算结果如表5所示。

表5 乘客平均换乘系数

出行类型	换乘系数
总体出行	1.37
通勤出行	1.43
非通勤出行	1.30

表5的计算结果表明，通勤出行者的平均换乘系数明显高于非通勤出行者。这说明受到出行时耗和工作地点的限制，通勤人群出行过程中存在更多换乘。

5 结语

研究利用公共交通刷卡数据，建立了城市公共

交通系统出行链连接方法和基于机器学习的出行人群分类模型，通过实际出行调查和测试样本集验证了出行链连接方法和出行人群分类模型的精度，并初步分析了北京市居民出行特征。结果表明，该特征提取分析方法可以有效识别通勤人群的城市公共交通系统出行链结构及换乘特性。

在今后的研究中，可通过增加分类训练集的样本量，以提高分类器的准确性与泛化性；从出行链的出行时间、上下车位置和换乘过程等维度进一步进行分析与信息挖掘，为城市轨道交通及道路公交线路规划与站点布局优化等提供更准确的数据支撑。

参考文献

[1] SEABORN C,ATTANUCCI J,WILSON N H M. Analyzing multimodal public transport journeys in london with smart card fare payment data [J]. Transportation Research Record Journal of the Transportation Research Board,2009(2121):55.

[2] JANG W. Travel Time and transfer analysis using transit smart card data [J]. Transportation Research Record Journal of the Transportation Research Board,2010(2144):142.

[3] DEVILLAINE F, MUNIZAGA M, TREPANIER M. Detection of activities of public transport users by analyzing smart card data[J]. Transportation Research Record Journal of the Transportation Research Board,2012(2276):48.

[4] SHI X,HANGFEI L. The analysis of bus commuters′ travel characteristics using smart card data:the case of Shenzhen,China [C]//Transportation Research Board 93rd Annual Meeting. Washington:Transportation Research Board,2014:172.

[5] 陈学武,戴霄,陈茜.公交IC卡信息采集、分析与应用研究[J].土木工程学报,2004(2):105.

[6] 王月玥. 基于多源数据的公共交通通勤出行特征提取方法研究[D].北京:北京工业大学,2014.

[7] 孙世超,庄斌,黄伟.基于机器学习的公交卡数据中通勤人群辨识方法[J].交通工程,2017(1):58.

[8] 王淑伟,孙立山,荣建.北京市轨道站点吸引范围研究[J].交通运输系统工程与信息,2013,13(3):183.

[9] 郭淑霞,陈旭梅,于雷,等.轨道交通换乘常规公交平均候车时间模型[J].交通运输系统工程与信息,2010,10(2):143.

[10] WANG H,SHEN Y,WANG L,et al. Large-scale multimedia data mining using MapReduce framework[C]//4th IEEE International Conference on Cloud Computing Technology and Science Proceedings. Taipei:IEEE,2012:287.

（收稿日期：2017-07-13）

（上接第65页）

中型城市的空间特点，建设可实施性强。环线结构本身可内部独立运行，也可与干线共线运行，适应中小型城市的交通出行特征。这样的运行方式对运营管理的要求较高，在具体制式选择和配线设置上要进一步细化研究。

参考文献

[1] 中华人民共和国发展改革委员会. 国务院关于调整城市规模划分标准的通知[Z].北京：中华人民共和国发展改革委员会，2014.

[2] 叶霞飞，顾保南.城市轨道交通规划与设计[M].北京：中国铁道出版社.1999.

[3] 顾保南，曹仲明.城市轨道交通路网结构研究[J].铁道学报，2000,22(5):25.

[4] 曹仲明，顾保南.城市轨道交通网络结构的优化及其影响分析[J].城市轨道交通研究，1999(1):45.

[5] 王忠强，高世廉.城市轨道交通路网形态分析方法[J].城市轨道交通研究，1999(1):33.

[6] 中铁第四勘察设计院集团有限公司. 大理轨道交通线网规划报告[R].昆明：中铁第四勘察设计院集团有限公司西南设计院，2016.

[7] 中铁第四勘察设计院集团有限公司. 西双版纳旅游轨道交通线网规划报告[R].昆明：中铁第四勘察设计院集团有限公司西南设计院，2017.

[8] 中铁第一勘察设计院集团有限公司. 天水市城市轨道交通线网规划报告[R].武汉：中铁第四勘察设计院集团有限公司，2015.

（收稿日期：2017-10-24）

欢迎访问《城市轨道交通研究》网站
www. umt 1998. com