基于多源城市交通出行数据的定制公交需求辨识方法研究

陈汐[1]，王印海[2]，代壮[3]，马晓磊[1,4]
1. 北京航空航天大学交通科学与工程学院，北京 100191；
2. 美国华盛顿大学土木和环境工程系，美国 西雅图 98195；
3. 西南交通大学交通运输与物流学院，四川 成都 610031；
4. 北京航空航天大学大数据科学与脑机智能高精尖创新中心，北京 100191

摘要
定制公交作为一种新的公交服务模式，对其需求的辨识具有重要实践意义，也是后续线路设计流程的基础。在大数据背景下，通过对海量出行数据中时空信息的挖掘分析，提出一个基于多源数据的定制公交需求分析框架，包括从公交和互联网用户中辨识通勤用户、出行需求的融合及站点选址方法。最后将该方法应用到成都市的出行数据中以验证其有效性，其需求分析结果可为定制公交的线路设计提供依据。

关键词
定制公交；多源数据；数据挖掘；需求辨识

中图分类号：U491.1+7 文献标识码：A doi: 10.11959/j.issn.2096-0271.2020057

Research on demand identification for customized bus based on multi-source mobility data

CHEN Xi[1], WANG Yinhai[2], DAI Zhuang[3], MA Xiaolei[1,4]
1. School of Transportation Science and Engineering, Beihang University, Beijing 100191, China
2. Department of Civil and Environmental Engineering, University of Washington, Seattle 98195, United States
3. School of Transportation and Logistics, Southwest Jiaotong University, Chengdu 610031, China
4. Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing 100191, China

Abstract
As a new type of public transit services, the demand identification of customized bus (CB) has great practical significance, as well as the basis of route design of CB. Under the context of big data, a methodology framework of CB demand identification based on multi-source data was proposed by mining spatial-temporal characteristics from large scale mobility data. The proposed framework includes several phases, which are the identification of commuters from transit and Internet users, data fusion of travel demands and stop deployment method. This study takes Chengdu city as an example to verify the effectiveness of the proposed methods. The results can provide a theoretical support of CB route design.

Key words
customized bus, multi-source data, data mining, demand identification

1 引言

优先发展公共交通是我国城市发展和交通发展的重大战略方针，且“公交都市”战略中也提出要构建多模式公交系统，以实现新时期城市交通的转型发展。目前公众对城市交通出行个性化、精细化和品质化的要求逐渐提高，因此发展多元化的公交出行服务模式已成为必要趋势。在此背景下，具有需求响应、运行灵活等特点的定制公交开始运营。定制公交通过整合个体的出行需求，为出行起终点（origin-destination，OD）、出行时段、服务水平相似的人群（如通勤用户）提供个性化的公共交通服务[1-2]。常规公交与定制公交的特点对比见**表1**。这种新的公交服务模式被认为能够有效吸引私家车出行用户转向乘坐公共交通出行。此外，全国各主要城市在新型冠状病毒肺炎疫情防控期间也推出了针对通勤用户的“复工定制公交”线路，保证乘客“一人一座”，以提升市民出行的便捷性和安全性。可见，定制公交可以作为个性化和精细化出行需求市场中一种很好的补充形式[4]。

近几年，“互联网+交通”的发展趋势有效地促进了定制公交这种新的出行模式在国内的推广、普及。运营企业通过搭建线上平台采集出行需求，用户通过手机App等渠道提出个性化的出行需求。基于大数据挖掘、人工智能算法、物联网等技术手段完成对出行需求的整合、线路的规划、运营车辆的调配以及服务信息发布等环节[1-4]。在线路的实际运营中，公交企业接收乘客反馈的建议，不断对现有线路进行调整、优化，逐步提升服务质量，使定制公交的运营模式形成完整的闭环。针对以上定制公交服务设计流程，已有文献对其中涉及的相关理论方法进行了研究。但大部分文献对出行需求的分析基于小范围出行需求进行调查，对于城市级的线路设计及大规模出行分析，存在研究群体相对较少、数据周期短、分析的准确度有偏差等问题。随着移动通信及互联网技术的发展，针对城市居民的出行需求，实现了由单一数据源到多源数据的采集，如从单一的调查数据的收集，到公交IC卡、手机导航应用的普及，这些技术手段可以采集到大量的出行信息。这些海量、多源的出行数据可以很好地解决单一调查数据难以挖掘城市居民出行规律的问题。因此，如何在多种出行模式下分析乘客的出行规律、融合出行需求，特别是辨识乘客的通勤行为、挖掘用户职住地，辅助定制公交的线路设计以提高其上座率和服务率，是值得深入探讨的问题。

多源交通出行数据符合大数据的“4V”特征，即规模性（volume）、多样性（variety）、价值性（value）和高速性

表1 常规公交与定制公交特点对比

对比项	常规公交	定制公交
服务范围与运行长度	无特殊要求	通勤线路长度长，微循环线路长度较短
站点布设	常规站点布设方法	从需求出发（优先利用现有站点）
站点数量	无特殊要求	较少
道路条件	无特殊要求	灵活选择主干通道
速度	15~30 km/h	一般高于常规公交
起终点	需布设公交场站	可在居住区与工作区设置
服务时间段	全天候	高峰期居多（通勤）
舒适度	无固定座位	一人一座

（velocity）[5]。第一，公共交通数据和新型互联网数据的体量巨大。据相关统计数据，来自政府及互联网企业的数据量正从TB量级增长到PB（EB）量级。第二，本文涉及的相关数据类型繁多，即异构多源，包括IC卡、车载GPS、出行导航和规划数据等。第三，多源出行数据具有实时性特点，即出行需求数据可以被实时地采集，并反馈给相关企业的调度和决策人员，研究者需要关注数据的计算效率问题。第四，多源交通出行数据的价值密度低，但是商业价值高。例如，车载GPS会实时地传回大量数据，但是对于特定的公交运营或调度目的，只需要对其中某些数据或字段进行分析。由于多源交通出行数据具有以上特征，在分析挖掘时会存在一些难点。第一，在海量数据中挖掘所需的有价值的信息是本文的难点之一；第二，本文重点关注异构多源这一特征。异构通常指不同形式或类型的数据[6]，多源指来源不同的数据，如公交数据包括静态数据、IC卡数据等，新型互联网数据包括导航数据和规划数据等。对单一数据源的分析与挖掘已有大量文献进行了相关讨论，如公交乘客上下车站点的推断、乘客出行行为分析等相关研究。相比之下，对多源数据的分析和应用依旧值得进一步探讨。本文讨论的多源出行数据具有多源和异构两种性质。因此，将多源数据进行融合，最大化地挖掘、提炼每种数据的价值以辅助定制公交的设计，是本文重点关注的问题，也是难点之一。

此外，不同类型的出行数据来自多个部门和企业，存在数据单位或者数据存储格式不一致的问题。因此在融合过程中还需要考虑数据一致性问题。

基于上述原因，本文在“互联网+交通”的大背景下，探究了如何利用多源出行数据挖掘城市出行需求，以辅助定制公交的服务设计，并探索了一套从数据处理到出行需求分析的流程。特别是针对新型互联网数据，提出了互联网用户的通勤需求识别方法、互联网数据与传统公交数据的融合算法。最后，将该分析流程应用到成都市的出行数据中，以验证该方法在处理城市级规模问题中的有效性。

2 基于多源数据的挖掘分析框架

Liu T和Ceder A在参考文献[1]中针对国内定制公交的运营与规划进行了系统的梳理，其线路设计如**图1**所示。其中，出行需求的提取是后续设计流程的基础，也是本文重点讨论的问题，不同于以往研究偏重于利用调查数据或单一数据源，本文讨论利用多源出行数据对需求进行挖掘分析，具体步骤如下。

步骤1：出行需求分析。出行需求的获取可以为后续线路规划模型的设计提供准

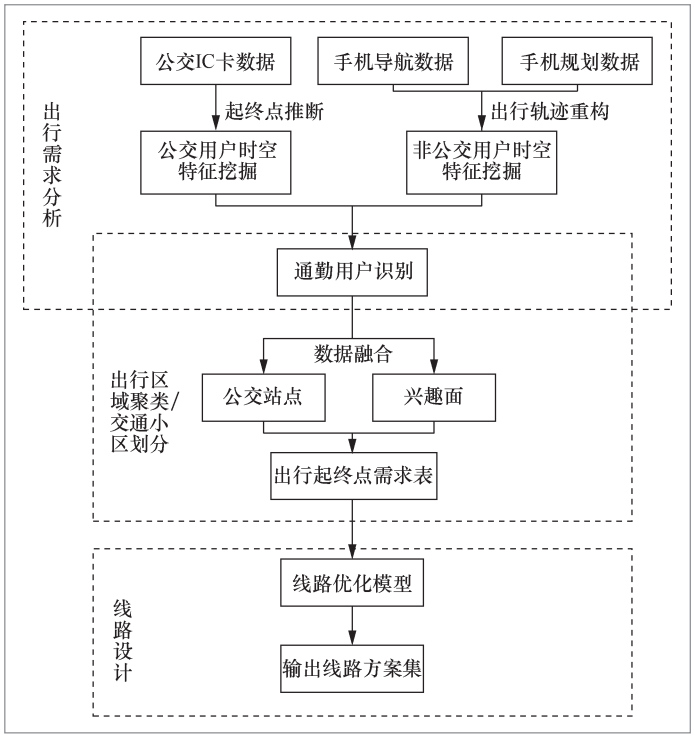

图1 多源数据挖掘分析及线路设计框架

确的数据基础。对于定制公交，该步骤的主要目的是获得乘客准确的出发地与目的地等资料。不同于出行调查直接获取相关信息，本文通过多源数据挖掘乘客的出行信息。因此，如何在海量数据中识别乘客的出行起终点信息，进一步辨识乘客的出行行为，特别是通勤行为，是需要解决的难点之一。

步骤2：出行区域聚类/交通小区划分。在设计线路时，每个出行区域或交通小区都是线路的构成元素。若出行区域范围划分得过大，则线路中的走行区域减少，有利于降低线路成本，但是会导致乘客到站点的走行距离变长，降低定制公交的吸引力，且定制公交的目的是为出行需求相似的用户提供个性化、便捷的公交服务。因此，目前多数研究通过聚类方法将出行起终点相似的需求进行聚类，并考虑每个区域的覆盖半径，以控制乘客到站点的走行距离，达到平衡乘客和运营者双方利益的目的。在出行区域划分完毕之后，运营者可在区域内灵活选择上下车站点位置。

步骤3：线路设计。在确定好每个出行区域，并且选择好站点位置之后，可将提取的需求投影到各出行区域（站点）之中，进而得到一张起终点需求表。以此需求为基础，可以构建线路规划模型。模型的优化可以从乘客、运营商、环境和社会效益几个方面来考虑，再通过算法进行求解，最终得到线路运营方案。

值得注意的是，在实际运营中，该框架是反复迭代的过程，乘客的出行需求是动态变化的，应根据运营状况，结合出行数据和乘客反馈，定期更新需求及线路，以提升服务质量。本文重点讨论步骤1和步骤2，对于步骤3中的线路设计问题，很多文献将其归结为车辆路径问题（vehicle routing problem，VRP）或者装卸货（接乘）问题（pickup and delivery problem，PDP），并进行了很多讨论。本文仅在实例分析中展示部分线路设计的结果，不将其作为本文重点研究的问题。

3 多源交通出行数据

本文涉及的数据来源主要包括两个部分：一类是传统公交数据，包括IC卡、车辆GPS数据、公交静态数据；另一类是新型互联网数据，即用户使用手机导航App产生的出行记录，包括手机规划数据和手机导航数据。

3.1 传统公交数据

（1）公交IC卡数据

不同运营商对应的IC卡数据结构可能不相同，如单一票制收费系统中只需上车刷卡，无下车刷卡记录。而北京等城市采用分段计费模式，刷卡系统中可以记录乘客的上下车站点信息。因此，目前我国的城市公交还没有形成统一的刷卡数据采集和存储规范[7]。但IC卡中存储的数据依旧可以反映城市居民的出行情况。**表2**给出了目前单一票制IC卡系统中存储的主要字段及其说明。

（2）公交车辆GPS数据

公交车辆GPS通过经纬度定位记录了该车辆实时的运行状态，通常每10 s左右产生一条位置记录。此外，在GPS的数据结构中还记录了车辆的营运线路以及到站、离站等信息。通过分析、融合车辆的状态信息可以还原其运行轨迹。**表3**给出了GPS数据包含的主要字段及其说明。

（3）公交静态数据

公交静态数据描述的是公交系统的整体情况，主要包括车辆、人员、线路及站点等信息，其中与本文讨论相关的是站点和线路信息。公交站点的地理信息数据记录了站点对应的编号、类型、经纬度等信息。该信息可以与GPS数据进行匹配，从而推

断车辆到离站的具体信息[7]。此外，由于一条公交线路包含多个站点，而一个公交站点可能会出现在多条公交线路中，因此公交站点与线路信息表主要用于记录线路与站点的对应关系，包含的字段内容主要有线路编号、站点的序号及编号。每条线路的基本信息一般被单独存放在一张表中，包括线路的编号、线路运营长度、走行方向、起终站点及线路类型等信息[7]。

3.2 新型互联网数据

（1）手机导航数据

手机导航数据记录了用户在使用App时产生的实时位置信息，一般每隔5~30 s记录一次，原始的导航数据字段内容包括用户ID、时间戳、速度及经纬度信息。为获得用户每次行程的起终点，需要对原始的轨迹数据进行预处理。根据参考文献[8]，可以设定一个时间阈值5 min，若某用户两条连续的出行记录大于此阈值，那么将这两条记录的位置信息分别记作上一次行程的终点和下一次行程的起点。对每个用户进行上述操作，可以得到具体的导航起终点信息，见**表4**。本文涉及的导航数据也可被视为私家车（驾车）出行需求。

（2）手机规划数据

用户在出行之前会查询导航软件，以提前对自己的行程进行规划，这时系统会记录查询的相关信息，该信息即手机规划数据，具体内容见**表5**。

表 2 公交 IC 卡数据结构说明

字段	名称	说明
ID	刷卡记录编号	每次使用记录的唯一标识
CARD_ID	卡编号/用户编号	每张卡的唯一标识
TIME	刷卡时刻	例如，2016-11-07 7:09:09
LINE_ID	线路编号	刷卡所对应的线路编号
CAR_ID	车辆编号	刷卡所对应的车辆编号
CONSUM	消费金额	每次刷卡的消费金额
BALANCE	剩余金额	刷卡后的余额

表 3 车载 GPS 数据结构说明

字段	名称	说明
ID	GPS记录编号	每条GPS记录的唯一标识
TIME	记录时间	例如，2016-11-07 7:10:06
LINE_ID	线路编号	GPS数据对应的线路编号
CAR_ID	车辆编号	GPS数据对应的车辆编号
LON	经度	地理信息经度
LAT	纬度	地理信息纬度
SPEED	速度	定位速度
DATATYPE	数据类型	GPS状态，如运行、到离站

（3）兴趣面数据

本文中的兴趣面（area of interest，AOI）数据是指互联网地图（如高德地图）中的兴趣面。不同于兴趣点（point of interest，POI），AOI用来描述地图中区域状的实体，如居住小区、办公楼、大型商圈等。每个AOI有自己的编号和用地属性。本文涉及的AOI数据包含3种用地类型，即居住类型、办公类型和商业类型。

3.3 数据预处理

在进行多源大数据挖掘分析之前，通常要对数据进行清洗或预处理工作[7]。其主

表 4 手机导航数据示例

用户ID	起点经度	起点纬度	终点经度	终点纬度	导航开始时间	导航结束时间
8e65**c9	104.0704° E	30.75361° N	104.0684° E	30.71848° N	2016-11-12 14:16:33	2016-11-12 14:25:08

表 5 手机规划数据示例

用户ID	起点经度	起点纬度	终点经度	终点纬度	查询时间	出行方式
5370**d5	103.920741° E	30.572845° N	103.920722° E	30.573240° N	2016-11-17 13:58:33	驾车

要目的是检测原始数据中的错误，剔除对分析结果有影响的数据，从而提升数据的质量。本节介绍了相关的数据预处理步骤。

（1）剔除无用字段

在第3.1节和第3.2节中详细介绍了每种数据的主要字段内容。实际上，一些字段与公交的出行需求分析无关，因此可以删除这些字段，以提升数据的分析效率。例如，在公交IC卡数据中，只需保留用户的卡号、刷卡时间、车辆号、线路号等信息。在车载GPS数据中，公交车的行驶速度等字段信息对需求分析无影响，可以在预处理阶段进行删除。

（2）剔除冗余数据和错误数据

冗余数据主要指信息重复的数据。冗余数据的存在会对后续的分析结果产生影响。例如，在分析公交IC卡时，重复的数据信息会使出行需求的总量偏高。因此，可根据具体的分析目标对冗余数据进行处理，如进行删除操作，以提高分析的准确性。

受数据采集终端或通信网络故障影响，采集到的数据记录会出现信息错误的情况。例如，对于时间字段，会出现“24:15:00”的情形。此外，数据中还会出现部分字段缺失的情况。为确保后续分析结果的准确性，应对数据的有效性进行检查，剔除错误或缺失的数据。

4 分析方法

本节根据多源数据挖掘分析框架的几个主要步骤，从乘客的出行需求分析及站点选址（交通小区划分）两个方面对相关的理论方法进行介绍。

4.1 用户出行轨迹重构

大数据等信息技术的应用使得交通信息的采集实现了从单一数据源到多源数据源的发展。在本文研究涉及的领域，传统的公交IC卡数据、公交车辆GPS数据、公交静态数据、新型互联网数据为多元化的出行服务提供了丰富的数据基础。在定制公交设计与优化理论中，首要的任务是准确捕获乘客的出行需求。因此，在海量数据资源下，如何有效提取用户的出行轨迹、挖掘完整的出行信息是需要解决的问题之一。本节主要讨论几种数据类型的出行信息挖掘方法。

4.1.1 公交IC卡数据上下车站点推断方法

在使用手机App乘坐公共交通普及之前，国内大部分城市的公交刷卡类型属于单一票制[9]，即乘客只需上车刷卡，且刷卡时不会记录站点位置信息。因此，为获得乘客的起终点信息，需要进行上下车站点的推断。对于单一票制的推断，国内外已有大量研究，且方法成熟[7,9-12]。本节根据国内公交IC卡数据的特征，简要介绍推断方法的主要流程。

（1）乘客上车站点识别方法

由于公交IC卡数据不含站台等位置信息，因此需要结合GPS数据进行上车站点的推断。上车站点识别算法步骤如下。

步骤1：公交GPS在运营过程中受到外部环境影响，会存在数据记录精度有限的情况，因此首先要进行GPS数据修正[7,12]，并与公交静态数据表进行匹配。该步骤的目的是得到车辆准确的到站时刻表，该表包含线路编号、车辆编号、站点编号及到站时间等信息[7]，见**表6**。

步骤2：利用SQL数据库软件，将公交IC卡数据表中的线路编号、车辆编号与**表6**中的线路编号、车辆编号进行“等号”关联。

步骤3：将公交IC卡数据表中的刷卡时间与**表6**中的“到站时间”进行“大于（>）”关联，与**表6**中的“下一站到站时间”进行“小于（<）”关联。

GPS数据的记录事件产生由开启和关

闭公交车门的行为引起，因此在上述算法中，假设乘客上车刷卡时间要晚于**表6**中车辆到达上车站点时间（或车辆开门时间），且早于下一站到站时间。基于以上假设，根据线路的到站时间表，若某用户刷卡时间晚于车辆到达第i个站点的时间，且早于第i+1个站点的到站时间，则推断出用户的上车站点为第i个站点。

（2）乘客下车站点识别方法

在单一票制中，由于乘客在下车时无须刷卡，因此在IC卡数据字段中无乘客下车相关信息。但在出行需求分析中，目的地是十分关键的信息，因此如何提取乘客的出行目的地也是需要解决的问题。目前研究中对此类问题的有效解决方法是基于乘客的出行链进行下车站点推断[7,12]。

现有文献认为公共交通的出行链可以被描述为一名乘客在一天的出行中至少乘坐了两次及以上的公共交通[7]。在参考文献[13]中，基于出行链的推断有两个假设条件：第一，在出行链中乘客本次出行的起点为上一次出行的终点；第二，乘客在一天之中最后一段出行过程的终点与当天第一段出行的起点是相同的。此外，本文讨论的公交出行不包含地铁。

基于以上假设，并考虑到公交出行通勤占比很大，具有很强的早晚出行规律性，例如，早高峰从居住地到达工作地，晚高峰从工作地回到居住地。因此对下车站点的识别可采取基于连续性的推断方法[12,14]。

例如，对于某乘客当日的末次出行，假设乘客会选择距当日首次出行的上车站点最近的站点下车。因此若末次出行乘坐的线路为R_i，首次出行的线路为R_j，则将线路R_i中距离R_j中上车站点最近的站点判定为R_i的下车站点。对于某乘客当日的非末次出行，乘客先乘坐线路R_m，再乘坐R_n。因此，选取R_m中距离R_n上车站点最近的站点作为乘客乘坐线路R_m的下车站点。

对于不满足公交出行链描述的公交出行，已有文献中的处理方法是通过乘客出行距离分布并结合公交站台的吸引特征[9]，构建基于概率的下车站点推断模型，具体方法可参考文献[7,9,12]。

4.1.2 用户通勤行为判别

前文中提到，通勤用户是定制公交主要的需求来源。因此，挖掘通勤乘客也是定制公交线路设计的重要任务之一。本节讨论了几种用户类型的通勤行为判定方法。

（1）公交用户通勤判别

公交出行通勤用户在整个出行总量中占比很大，具有明显的潮汐规律性，因此可分析乘客在一定周期（如一个月）内的早晚高峰出行规律或刷卡规律，通过刷卡频次并结合刷卡站台周边的用地信息（如POI信息）进行判定[15]。具体步骤如下。

步骤1：查阅已有文献，根据频次设定判定标准。例如，一周工作日中早高峰两次及以上在同一个站台刷卡时，可将该站台视为居住地。类似地，晚高峰两次及以上在同一个站台刷卡时，可将该站台视为工作地。

步骤2：对于某用户，根据步骤1中的通勤标准，判定其居住地和工作地。对于居住地，在推断其上车站点的基础上，统计在一个周期内的早高峰时段在不同站台的刷卡频次，若频次最高所对应的站台达到通勤判定标准，将该站台标记为居住地，否则判定失败，该用户居住地为空。类似地，统计在一个周期内晚高峰时段在不同站台的

表 6 公交到站时刻表示例

线路编号	上下行	车辆编号	站点编号	到站时间	下一站到站时间
97	1	21119006	42357	2016-11-16 10:17:33	2016-11-16 10:29:42

刷卡频次，并进行工作地判定。

步骤3：对数据库中每个公交用户的IC卡记录进行上述步骤的统计和判定，输出乘客的工作地和居住地。

（2）规划数据通勤判别

不同于公交IC卡数据的起终点对应站台，规划数据的起终点信息对应的是经纬度，这可能导致用户每次规划路径时所对应的位置信息（经纬度）不一致的现象出现。因此，本文提出了一个基于AOI数据的通勤用户判定方法。该方法的主要思路为利用AOI数据将经纬度信息映射到更大的区域，以方便统计每次规划路径的位置信息，从而得到用户出行起终点的对应区域。具体判定步骤如下。

算法1 规划数据用户通勤判别算法

输入：规划数据用户一个周期内（如一个月）的出行轨迹信息

输出：用户居住地与工作地

步骤1：将规划数据中所有用户的经纬度信息映射到AOI上，得到每个经纬度信息对应的AOI编号，以及对应的AOI类型。这使得用户每次的规划位置从经纬度投影到AOI。

步骤2：查阅已有文献，根据频次设定判定标准。例如，一周工作日中，规划数据用户两次及以上在同一个AOI并且其类型为“居住类型”规划路径时，可将该AOI视为居住地。类似地，用户两次及以上在同一个AOI并且其类型为“办公类型”规划路径时，可将该AOI视为工作地。

步骤3：在步骤2的判定标准基础上，识别规划用户的居住地和工作地，具体方法同公交IC卡用户的居住地和工作地的识别方法。

步骤4：对规划数据中每个用户的记录进行上述步骤的统计和判定，输出乘客的居住地和工作地。

（3）导航数据通勤判别

导航数据的起终点位置信息对应的也是经纬度，且由于导航的出行轨迹很多是间断的，其终点位置不一定是实际的终点位置（一些用户会在行程中关闭导航）。因此，对导航用户通勤的判别也需要结合AOI数据及导航起点位置信息进行判断。本文提出一个基于出行频次的通勤用户判别算法，具体步骤如下。

算法2 导航用户通勤判别算法

输入：导航用户一个周期内（如一个月）的出行轨迹信息

输出：用户居住地与工作地

步骤1：与规划数据处理方法相同，将导航数据中所有用户的每次轨迹行程的出发位置信息映射到AOI上，得到每个经纬度信息对应的AOI编号，以及对应的AOI类型。

步骤2：将一天24 h分割成24个单元，每个单元为1 h，初始化两个$N\times3$的空矩阵$\boldsymbol{M}_1$和$\boldsymbol{M}_2$，其中N为一个周期的天数（如一个月30天）。

步骤3：对于第i天，针对该用户所有轨迹的出发地位置，找到早高峰第一条轨迹对应的出发地AOI编号、AOI类型及导航行为发生时对应的时间区间，将其存储到$\boldsymbol{M}_1$的第i行。

步骤4：对于第i天，针对该用户所有轨迹的出发地位置，找到晚高峰第一条轨迹对应的出发地AOI编号、AOI类型及导航行为发生时对应的时间区间，将其存储到$\boldsymbol{M}_2$的第i行。

步骤5：按步骤3和步骤4重复操作一个周期内n天的轨迹信息，将相应的出行信息记录到矩阵$\boldsymbol{M}_1$和$\boldsymbol{M}_2$中。

步骤6：在$\boldsymbol{M}_1$中找到该用户AOI类型为“居住类型”的频次数最高（最频）的出行AOI编号，并记录其次数N_h；在$\boldsymbol{M}_2$中找到该用户AOI类型为“办公类型”的最频出行AOI编号，并记录其次数N_w。

步骤7：查阅已有文献，根据频次设定

判定标准[16]。例如，如果N_h大于8，则将该AOI区域标记为用户居住地，否则判定失败；类似地，如果N_w大于8，则将该AOI区域标记为用户工作地，否则判定失败。

步骤8：对导航数据中每个用户的记录按步骤1~步骤7进行统计和判定，输出乘客的居住地和工作地。

4.2 站点选址/出行区域聚类

站点的选址布局在定制公交的服务设计中是重要的一环。为吸引更多的出行者选择定制公交，乘客到站点的步行距离不应过大。但前文中提到，乘客的需求在一个区域内是不均匀的，即公交用户的需求点对应的是站台，而导航软件用户的需求点对应的是AOI。因此，为保证区域内用户与定制公交站点的走行距离在合理范围之内，通常通过聚类方法将出行起终点相似的需求聚类为一类，在聚类后的区域中选择站点位置。一些文献中也将该问题定义为定制公交的交通小区划分问题[3,17-18]。

不同于其他相关文献中处理的是单一的数据源（公交IC卡数据、出租车数据或调查数据）[3-4,18-19]，本文涉及多源数据，且需求点坐标对应的格式不一致。因此，在交通小区划分问题中如何融合多源数据需求是本文面临的难点之一。针对此问题，本文提出了一种多源数据融合方法。

首先，运用ArcGis软件计算每个AOI的中心坐标，将该坐标视为互联网用户的需求点。公交站台坐标可从数据库中直接抽取。然后，本文提出一个改进的具有噪声的基于密度的聚类（density-based spatial clustering of applications with noise，DBSCAN）算法，使用该算法对传统公交数据及互联网数据进行融合。由于所需处理的定制公交需求的数据规模通常较大，且存在很多孤立点，此外，在需求区域的划分中往往不能预先确定所需的聚类类别数，因此一般的K-means等聚类算法并不适合本文讨论的情况。本文采用DBSCAN算法，该算法不需要预先指定聚类类别的个数。但该算法易出现“聚类成片”的问题，即相邻的临时聚类簇容易聚成一类。此外，该算法也不能考虑每个聚类区域半径的大小。也就是说，在划分定制公交出行区域时，可能会存在某个区域划分过大，这样在后续布设站点时，该区域部分乘客到站点的走行距离会过大，这会降低定制公交的服务质量。因此，本文对该算法进行了改进，具体算法步骤如下。

算法3：改进的DBSCAN算法

输入：多源出行需求点，聚类半径r，每个类别中的最少需求点个数minPts

输出：每个需求点坐标所属类别

算法描述：

初始化类别标记$C=0$

Repeat

对于提取的需求点坐标P：

If坐标P的标签为空：

计算得到坐标P在聚类半径r内的所有相邻点P_n

If P_n的个数小于minPts：

将P标记为噪声点

Else

$C = C + 1$

将P标记为类别C

将P_n中所有标签为空的需求点标记为类别C

Else

继续循环下一个需求点

Until数据库中没有要处理的需求点

输出每个需求点所属类别

在该算法中，聚类半径r和每个类别中最少需求点数目minPts是两个十分关键的参数。对于聚类半径r，假设乘客步行的速度为80 m/min，可接受的步行时间为5 min，

则聚类半径r为400 m，这保证了在该区域内设置站点时乘客的走行时间在可接受的范围内[19]。对于每个类别中需求点的个数，在参考文献[18]中提到，在每个区域中至少应存在一个上车站点和一个下车站点，以对应往返的出行需求。因此，在实际问题中，每个类别中的需求点个数至少为2。因此，基于上述算法，对于每个划分好的出行区域，可利用原有的公交站台作为定制公交站点，也可根据实际情况进行考察后确定站点位置。参考文献[4]讨论了站点的具体布设方法。

5 实例分析

本文利用成都市一个月（2016年11月1—30日）的出行数据对以上分析框架的可行性进行验证，采用的数据包括成都市公交数据、互联网导航及规划数据、AOI数据。本文使用的数据由成都公交集团和高德地图提供。由于数据敏感等原因，本文在结果部分只分析了数据的趋势。**图2**是公交IC卡数据、互联网导航和规划数据的数据量时间分布。从**图2**可以看出，公交IC卡数据从6:00开始刷卡量逐渐增加，到10:00逐渐平稳；从16:00开始刷卡量逐渐增加，到19:00高峰时段结束。因此，在通勤判别中将早晚高峰的时间段定义为6:00–10:00和16:00–19:00。而互联网用户的出行量在8:00–18:00时段内都保持在一个相对平稳的状态。因此，为了统一标准，在对数据进行分析、融合时，将互联网用户的通勤时间范围也定义在6:00–10:00和16:00–19:00这两个时段。

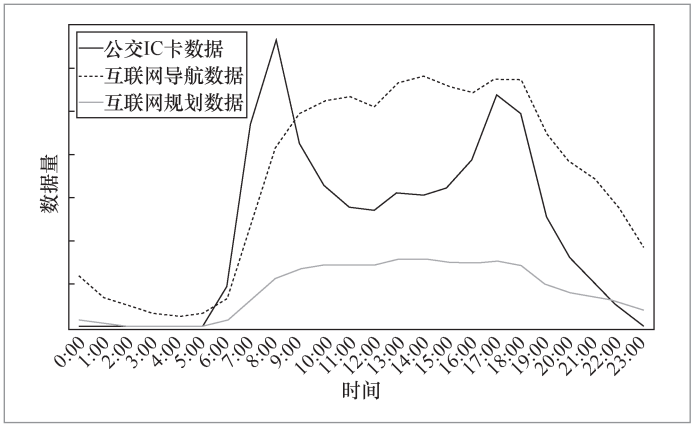

图 2 数据量时间分布

表 7 公交用户上车站点推断示例

卡号	刷卡时间	线路号	车辆号	上车站台
161075411	2016–11–07 7:34:33	98	21018422	30223

表 8 公交用户通勤判别示例

卡号	出发时段	刷卡站台	频次	返程时段	刷卡站台	频次
161075411	7:00–8:00	40675	3	19:00–20:00	30807	2

（1）首先将GPS数据和公交IC卡数据进行匹配以推断上车站点，最终约90%的用户上车站点可以被成功识别，从而得到乘客的上车站点匹配表，见**表7**。然后，结合下车站点推断算法，整理得到乘客在一个月内的上下车站台、刷卡时段及频次，见**表8**。最后，选取早高峰时段6:00–10:00、晚高峰时段16:00–19:00，设定频次阈值，得到成都市公交通勤用户识别结果。**图3**展示了不同频次阈值下通勤用户识别数量的趋势。

（2）对互联网导航和规划数据用户进行通勤识别，将导航数据和规划数据用户OD的经纬度坐标投影到AOI区域中。选取早高峰时段6:00–10:00、晚高峰时段16:00–19:00，设定频次阈值为8[16]，按第4.1节中的算法判定通勤用户，并得到对应的职住地信息，见**表9**。

（3）通过改进的DBSCAN算法融合公交与互联网数据需求，从而划分出行区域。首先计算每个AOI的中心坐标，再与公交站台坐标进行聚类融合，最终得到每个出行区域，聚类过程如**图4**所示，该方法通过聚类半径r限定了每个聚类区域的大小。最终将成都市划分为多个出行区域，其聚类中心空间分布如**图5**所示，再将出行需求

投影到每个出行区域中，得到成都市公交和互联网用户的潜在出行OD需求表，将其作为线路规划的数据输入。受运营成本等因素的影响，定制公交的服务不可能覆盖所有出行区域。因此，应优先考虑在出行的热点区域进行线路设计。在数据融合后，将出行OD需求表投影到地图中，得到数据融合后OD热点区域的空间分布，如图6所示。在图6中，圆圈区域是公交用户和互联网通勤用户共有的工作地热点区域。因此，在后续线路中可以优先在这两个区域提供定制公交服务，以保证定制公交上座率。

（4）得到出行OD需求后，可通过构建线路设计模型得到线路方案集。例如，以最大化定制公交服务率为目标构建模型。参考文献[20-21]对此方法进行了研究，本文不再讨论具体模型和算法。下面将本文挖掘分析的出行需求作为模型输入，展示了通勤定制公交的线路案例。图7是方案集中的某条定制公交线路，该线路的潜在出行需求约2 600人/天，线路长度为7 km，目前存在的常规公交线路是48路。表10给出了该定制公交线路在高峰时段每个站点间的潜在出行需求。表11对比了不同出行时段定制公交与常规公交线路的预计出行时间。

本文进一步从出行时间和拥挤度方面对定制公交和常规公交线路进行对比。图8是常规公交48路63辆运营车辆在各时段的线路运营时间和乘客数量。从图8可以看

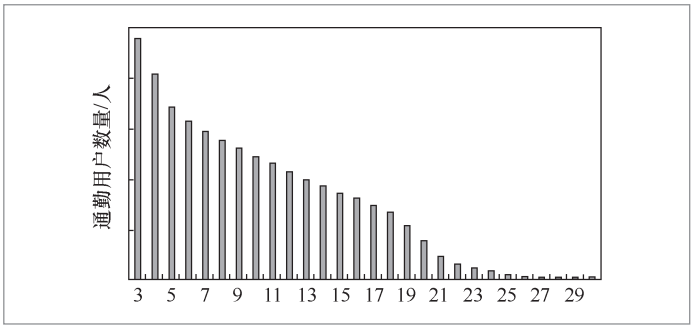

图 3 不同频次阈值下通勤用户数量趋势

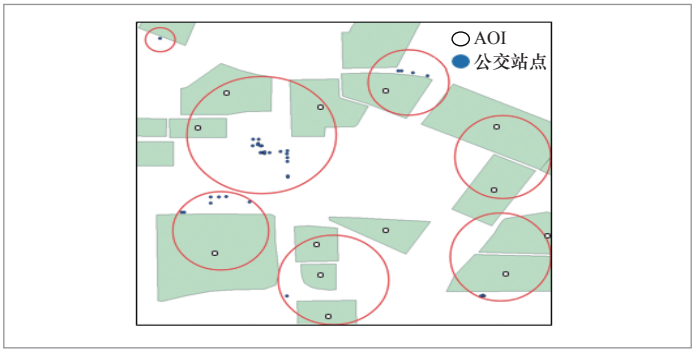

图 4 DBSCAN 算法聚类示意图

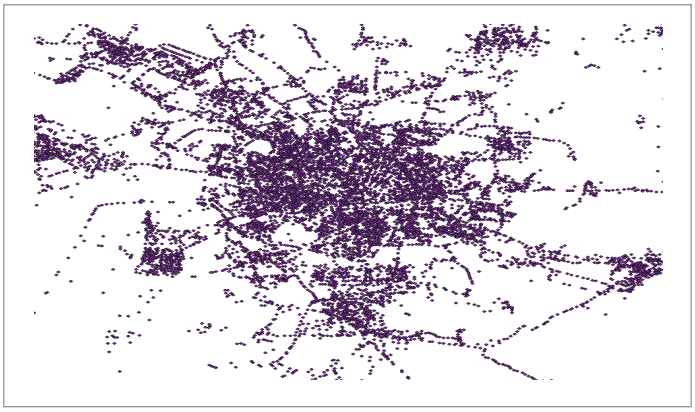

图 5 聚类中心空间分布

表 9 互联网导航和规划数据用户通勤判别示例

用户ID	早高峰出发时段	AOI编号	频次	晚高峰返程时段	AOI编号	频次
b4f***42a	7:00-8:00	10403	11	16:00-17:00	20577	15

表 10 定制公交线路的潜在出行需求（单位：人次）

站点	1	2	3	4	5	站点名称
1	–	300	648	704	226	星河路
2	–	–	–	321	161	沙湾路
3	–	–	–	195	4	马家花园
4	–	–	–	–	52	文殊院
5	–	–	–	–	–	春熙购物广场

表 11 定制公交与常规公交线路的预计出行时间对比

出行方式	预计出行时间/min		
	早高峰	平峰	晚高峰
定制公交	25	22	27
常规公交	35	33	38

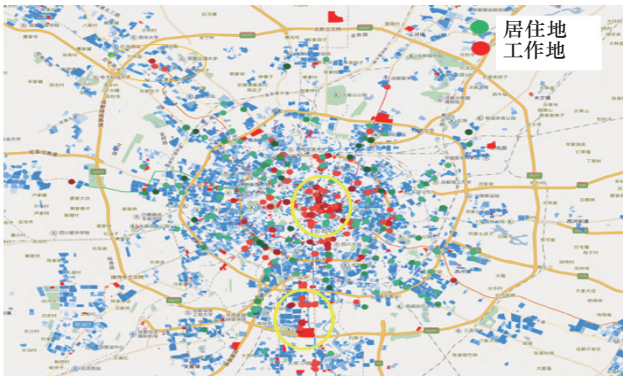

图 6 数据融合后 OD 热点区域空间分布

图 7 定制公交线路示例

出，由于该线路中间站点数量多，且通勤客流量大，造成停站时间过长，进而影响其运营时间和乘客出行体验。因此，如果引入定制公交线路，并提供“一人一座”等服务提高舒适度，可以很好地对常规公交线路进行补充。

6 结束语

本文提出了一个基于多源数据的定制公交需求辨识方法分析框架，并结合成都市的出行数据讨论了该框架的可行性。不同于传统基于调查问卷的分析方法，本文通过融合多源出行数据、挖掘居民的出行规律，获取城市居民的出行需求。本文介绍的方法适用于城市级的大规模数据，具有可操作性强、需求覆盖广、成本低、出行需求更新及时等优点，可作为定制公交服务设计的辅助手段。在未来的研究中，笔者将继续围绕定制公交的线路、时刻表、车辆调度及票价策略等运营规划方面开展工作。

参考文献:

[1] LIU T, CEDER A. Analysis of a new

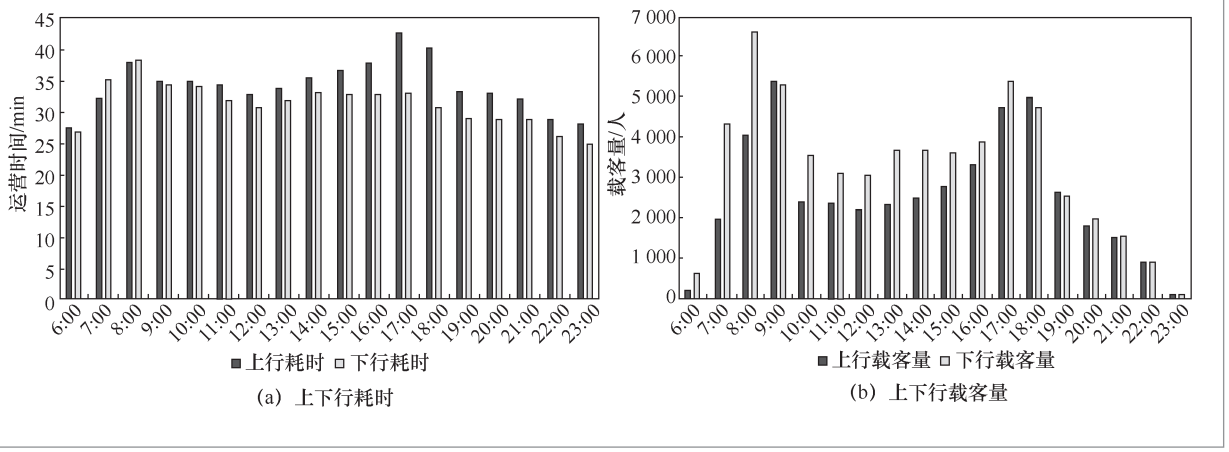

图 8 常规公交 48 路运营状态分析

public-transport-service concept: customized bus in China[J]. Transport Policy, 2015, 39: 63-76.

[2] 陈栩. 考虑多样化运营模式的定制公交线网规划研究[D]. 广州: 华南理工大学, 2017.
CHEN X. The study of customized bus network design in diversified operation modal[D]. Guangzhou: South China University of Technology, 2017.

[3] MA J, YANG Y, GUAN, W, et al. Large scale demand driven design of a customized bus network: a methodological framework and Beijing case study[J]. Journal of Advanced Transportation, 2017(3).

[4] LYU Y, CHOW C, LEE V, et al. CB-Planner: a bus line planning framework for customized bus systems[J]. Transportation Research Part C, 2019, 101: 233-253.

[5] 马晓磊, 丁川, 于海洋, 等. 公共交通大数据挖掘与分析[M]. 北京: 人民交通出版社, 2017.
MA X L, DING C, YU H Y, et al. Public transportation big data mining and analysis[M]. Beijing: China Communications Press, 2017.

[6] 项连城, 桑基韬, 徐常胜. 跨社交媒体网络大数据下的用户建模[J]. 大数据, 2016, 2(5): 32-42.
XIANG L C, SANG J T, XU C S. Cross-OSN user modeling in big data[J]. Big Data Research, 2016, 2(5): 32-42.

[7] 王周全. 基于IC卡数据与GPS数据的公交客流时空分布研究[D]. 成都: 西南交通大学, 2016.
WANG Z Q. Research on bus passenger flow space-time distribution based on IC card data and GPS data[D]. Chengdu: Southwest Jiaotong University, 2016.

[8] ZHANG S, LIU X, TANG J, et al. Urban spatial structure and travel patterns: analysis of workday and holiday travel using inhomogeneous Poisson point process models[J]. Computers Environment and Urban Systems, 2019, 73: 68-84.

[9] MA X, WANG Y, CHEN F, et al. Transit smart card data mining for passenger origin information extraction[J]. Journal of Zhejiang University Science C, 2012, 13(10): 750-760.

[10] MA X, WU Y, WANG Y, et al. Mining smart card data for transit riders' travel patterns[J]. Transportation Research Part C: Emerging Technologies, 2013, 36: 1-12.

[11] 马晓磊, 刘从从, 刘剑锋, 等. 基于公交IC卡数据的上车站点推算研究[J]. 交通运输系统工程与信息, 2015, 15(4): 78-84.
MA X L, LIU C C, LIU J F, et al. Boarding stop inference based on transit IC card data[J]. Journal of Transportation Systems Engineering and Information Technology, 2015, 15(4): 78-84.

[12] HUANG D, YU J, SHEN S, et al. A method for bus OD matrix estimation using multisource data[J]. Journal of Advanced Transportation, 2020(1): 1-13.

[13] 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, 1817(1): 183-187.

[14] 游婷, 范桂莲, 马兴慧. 基于公交IC卡信息的公交客流推算[J]. 交通工程, 2018, 18(6): 51-56.
YOU T, FAN G L, MA X H. Bus passenger flow calculation based on IC card information[J]. Journal of Transportation Engineering, 2018, 18(6): 51-56.

[15] MA X, LIU C, WANG Y, et al. Understanding commuting patterns using transit smart card data[J]. Journal of Transport Geography, 2017, 58: 135-145.

[16] SHOU Z, DI X. Similarity analysis of frequent sequential activity pattern mining[J]. Transportation Research Part C, 2018, 96: 122-143.

[17] 涂文苑. 定制公交的线网规划研究[D]. 北京: 北京交通大学, 2016.
TU W Y. The study of the customized bus network design[D]. Beijing: Beijing Jiaotong University, 2016.

[18] LI J, LYU Y, MA J, et al. Methodology for extracting potential customized bus routes based on bus smart card data[J].

Energies, 2018, 11: 2224.

[19] TONG L, ZHOU L, LIU J, et al. Customized bus service design for jointly optimizing passenger-to-vehicle assignment and vehicle routing[J]. Transportation Research Part C, 2017, 85: 451-475.

[20] 张敏捷, 冯偲, 吕晨曦, 等. 定制公交线路优化模型及求解算法[C]// 2014第九届中国智能交通年会大会论文集. 北京: 电子工业出版社, 2014: 347-354.
ZHANG M J, FENG S, LYU C X, et al. Custom bus routes optimization model and its algorithm[C]// The 9th China Intelligent Transportation Annual Meeting. Beijing: Publishing House of Electronics Industry, 2014: 347-354.

[21] CHEN X, WANG Y, MA X. Customized bus line design model based on multi-source data[C]// International Conference on Transportation and Development 2018: Traffic and Freight Operations and Rail and Public Transit. Virginia: American Society of Civil Engineers, 2018: 218-228.

作者简介

陈汐(1988-), 男, 北京航空航天大学交通科学与工程学院博士生, 主要研究方向为公共交通运营与规划。

王印海(1965-), 男, 博士, 华盛顿大学土木和环境工程系终身教授、博士生导师, 主要研究方向为交通检测、e交通学与大数据应用、交通控制、交通建模、智能交通系统、交通安全及交通仿真等。

代壮(1989-), 男, 博士, 西南交通大学交通运输与物流学院助理教授, 主要研究方向为交通数据分析和公共交通系统建模。

马晓磊(1985-), 男, 博士, 北京航空航天大学交通科学与工程学院交通运输工程系副教授、博士生导师, 主要研究方向为城市公交系统优化、交通数据挖掘与人工智能以及大规模交通网络建模与分析等。

收稿日期: 2020-05-18
通信作者: 马晓磊, xiaolei@buaa.edu.cn
基金项目: 国家自然科学基金资助项目(No.61773036)
Foundation Item: The National Natural Science Foundation of China (No.61773036)