引用格式：赵鹏军,曹毓书.基于多源地理大数据与机器学习的地铁乘客出行目的识别方法[J].地球信息科学学报,2020,22(9):1753-1765. [Zhao P J, Cao Y S. Identifying metro trip purpose using multi-source geographic big data and machine learning approach[J]. Journal of Geo-information Science, 2020,22(9):1753-1765.] **DOI:**10.12082/dqxxkx.2020.200134

基于多源地理大数据与机器学习的地铁乘客出行目的识别方法

赵鹏军,曹毓书

北京大学城市与环境学院城市规划与交通研究中心,北京 100871

Identifying Metro Trip Purpose using Multi-source Geographic Big Data and Machine Learning Approach

ZHAO Pengjun*, CAO Yushu

The Centre for Urban Planning and Transport Studies, College of Urban and Environmental Sciences, Peking University, Beijing 100871, China

Abstract: Identifying metro trip purpose using Smart Card Data (SCD) is important to expand the application of SCD in transport research and transport planning. This paper integrates different types of big data and combines the theories on the interaction between transport and land use. By taking Beijing as a case, we firstly analyze the metro trip purposes of individual passengers using travel survey data from 5565 respondents. Secondly, we investigate the land use features of trip origin and destination using Point of Interest(POI) data . Thirdly, a metro trip dataset is developed which includes the information of trip purpose, trip duration, and spatial distribution of trip origin and destination. Fourthly, a Random Forest (RF) algorithm is used to establish a RF classifier using the metro trip dataset as training data. Finally, this trained classifier is used to classify each metro trip recorded by the SCD to identify the metro trip purpose and the spatial distribution of metro trips for different purposes. The results of analysis show that the random forest classifier trained in this study can effectively identify metro trip purposes from SCD. For trips with "go to work" and "go home" purposes, the accuracy of identification can reach over 90%. One reason for the high identification accuracy is that land use information is included in the RF classifier. Our results confirm the theory of spatial-temporal interactions between transport and land use. There is an increasing availability of multi-source geographic big data and traffic survey data of residents in large cities, which means that the method developed in this study would have a high value in metro trip predicting and monitoring, transport planning, and land use policy-making around the metro stations. Also, our results enhance our knowledge of metro travel behavior in megacities.

Key words: Metro trips; trip purpose; travel survey data; smart card data; point of interest data; Random Forest algorithm; land use; spatial-temporal interactions; Beijing

***Corresponding author:** ZHAO Pengjun, E-mail: pengjun.zhao@pku.edu.cn

---

收稿日期：2019-03-22；**修回日期**：2020-07-04.

基金项目：国家自然科学基金项目(41925003)；英国研究理事会全球挑战基金项目(R48843)。 [**Foundation items:** National Natural Science Foundation of China, No.41925003; Research Councils of United Kingdom Global Challenges Research, No.R48843.]

作者简介：赵鹏军(1975—)，男，陕西延安人，教授，博士生导师，主要从交通与空间规划研究。E-mail: pengjun.zhao@pku.edu.cn

摘要：探索地铁乘客出行目的识别方法，有助于突破智能卡数据（Smart Card Data，SCD）在具体应用场景中的局限性，提升SCD在交通出行研究、交通发展规划等领域的应用价值。本文融合多源地理大数据，基于城市交通与土地利用时空间互动理论，以北京市居民地铁出行为例，在交通出行调查数据中提取5565个地铁出行样本及其对应的出行目的和出行特征相关变量。基于兴趣点（Point of Interest，POI）数据得到各样本起止站点的土地利用特征相关变量，形成包含每次地铁出行的出行目的、出行特征、土地利用特征的地铁出行数据集。使用基于随机森林（Random Forest，RF）算法对地铁出行数据集进行训练完成的分类器对SCD记录的每一次地铁出行进行分类，获得该次出行的出行目的及其不同目的地铁出行时空间分布规律。研究结果表明，本识别方法可有效预测地铁乘客的出行目的，其中，“上班”、“回家”2类出行目的的预测准确率均超过90%；纳入土地利用特征相关变量可显著提升RF分类器预测准确率，印证了城市交通与土地利用的时空间互动理论。鉴于当前SCD的可获取性逐渐提高，该项技术在居民地铁出行监测与预测、地铁线网布局和地铁周边土地利用规划等实践方面，具有很强的推广性，有助于更全面地认知大城市居民的地铁出行行为。

关键词：地铁出行；出行目的识别；交通调查数据；智能卡数据；兴趣点数据；随机森林；土地利用；时空间互动；北京

1 引言

交通出行是居民为特定目的或活动转换到对应位置或场所而产生的在空间上的位移需求[1]。城市交通出行的空间分布，与人口、设施的空间分布密切相关[2]。分析城市不同类型出行需求的空间分布规律，并探讨其与城市空间结构[3]、城市土地利用[4]的关系，是城市交通与土地利用一体化发展的基础[5]。交通大数据具有样本量大、时效性强的特征[6]，能够弥补传统交通调查数据样本量相对不足的局限性，已逐渐成为智能交通规划与管理的关键技术之一。随着各国“公交都市”等绿色交通理念的普及和政策实施，基于智能卡数据（Smart Card Data, SCD）的居民出行监测与管控技术体系正在受到普遍关注[7-12]，被广泛应用于城市交通流时空分布[13-15]、居民出行规律[16-17]等相关研究中。然而，SCD缺乏出行目的、持卡人社会经济属性等详细信息，使得SCD在信息挖掘方面受到限制。相关研究针对SCD技术的这一缺陷，探索出2种SCD出行目的的识别方法。① 依据对每次出行终点的土地利用特征、在终点驻留时间等设定的假定条件判断每个持卡人每一次出行的出行目的地是属于工作地、居住地或其他地点。接下来通过计数或聚类方法，筛选出每个持卡人的居住地和就业地[14-15,17]。在此基础上，可根据每个持卡人每次出行的目的地识别该次出行的出行目的。该方法简单有效，但没有真实值对假定条件进行检验，存在着基于主观经验、缺乏统一标准等问题。② 首先对某一地铁乘客进行出行调查，并基于调查数据，训练用于地铁出行目的识别的分类器。随后使用分类器对该地铁站SCD记录的每一次地铁出行进行分类，预测该次出行的出行目的[16]。相对于方法①，该方法可基于真实调查数据进行检验，更具说服力。然而仅采用一个地铁站的SCD与出行调查数据，会忽略城市内部来自不同区域乘客在社会经济属性、出行行为习惯等方面的差异。从规划管理与政策制定的角度，对整个市域或区域地铁出行目的识别方法的研究更具实践价值。

本文基于城市交通与土地利用交互作用理论，在方法② 基础上纳入对土地利用特征的考虑，进一步探索城市全域范围内地铁乘客出行目的的识别方法。根据城市交通与土地利用交互作用理论[5]，地铁站周边的土地利用与地铁出行量息息相关[18-21]，同时也会影响居民出行方式[22-24]以及出行目的地[25-27]的选择。本文以北京市作为研究区域，融合多源地理大数据进行随机森林（Random Forest，RF）分类器训练，在分类器训练过程中融入出行特征的同时纳入土地利用特征；随后使用RF分类器，对SCD记录的每一次地铁出行进行分类，识别对应地铁出行的出行目的，并对不同目的地铁出行时空间分布特征进行可视化。最后对比分类器①仅包括出行特征RF分类器，分类器②同时包括出行特征、土地利用特征的RF分类器效果，检验纳入土地利用特征对RF分类器效果是否有提升，以映证城市交通与土地利用交互作用理论。

2 研究理论基础、数据与方法

2.1 理论基础

研究基于城市交通与土地利用的时空互动理论。该理论表明，交通出行需求源于城市经济活动的空间分布，因而土地利用可以通过影响社会经济活动的空间分布来影响交通流的时空分布特征；交通需求时空分布与交通服务设施同时也会从城市

土地价格方面影响城市土地利用，由此形成交通与土地利用的时空互动。

交通与土地利用时空互动同时体现在集计与非集计2个层面[5]：① 在非集计（个体）层面，依据Alonso单中心城市模型，居民会在交通出行成本与住房成本中进行权衡、以实现自身效用最大化，由此形成微观个体层面交通出行与土地利用之间的互动[28]。② 在集计层面，依据Hansen“土地利用与交通系统之间存在动态循环”假设，不同土地利用类型的空间分离使得工作、居住、购物、休闲等活动在不同的城市区位进行。居民在日常活动区位切换过程中产生交通出行，交通出行汇总后形成不同区域之间的交通流，引发交通服务设施供给的变化，进一步影响区域可达性，并影响居民个体的出行决策，导致土地利用变化。因此，城市交通与土地利用之间存在时空间互动作用[29]。

交通出行与土地利用之间的时空互动作用显著体现在地铁站周边区域。居民个体出行行为、整体出行量与地铁周边土地利用情况呈现出显著的相关关系。地铁出行量与地铁站周边土地利用、人口密度显著相关[20]。居民也会依据地铁站点周边的土地利用特征，进行出行方式选择[19,22,30]、出行目的地决策[25]、是否在地铁站点周边进行居住、购物、休闲等日常活动[27]。

2.2 数据来源

本文基于多源地理大数据进行地铁乘客的出行目的识别，包括北京市居民出行调查数据，北京市兴趣点（Point of Interest，POI）数据，北京市智能卡数据等，具体数据类型、来源与对应时间如表1所示。

2.2.1 交通调查数据

交通出行调查数据为北京市2015年交通出行调查数据。通过调查问卷对北京市各个地区出行人群进行抽样、调查并获取受访人的一日活动轨迹。数据记录了受访对象一天内一次或多次出行的出发时刻、出发地所在交通小区，目的地所在交通小区以及到目的地的活动（即出行目的）。到目的地的活动（出行目的）主要包括上班、回家、上学、购物、就医、休闲娱乐、外出就餐、探访亲友等。数据同时记录了受访者每一次出行的换乘信息，包括换乘地点、换乘时刻、换乘后交通出行方式。该调查数据用来筛选出居民地铁出行调查记录，并提取出每条地铁出行记录的起点站、终点站等位置信息，出发时刻、到达时刻、出行时长等出行特征，以及被用于识别的出行目的信息。

使用调查问卷数据筛选受访人每次地铁出行的相关信息，包括出发地所在交通小区、出发时刻、目的地所在交通小区和出行目的。由于问卷中没有记录地铁出行对应的起止地铁站名称和地铁出行结束时刻，仅包含上车站点与下车站点所在地铁线路。为获得地铁出行具体出发站点、到达站点名称及时刻，采用以下3个步骤。① 根据每次地铁出行出发地与目的地对应的交通小区，寻找对应地铁线路上距离交通小区中心点最近地铁站，分别作为该次地铁出行的起始站点。② 基于起始站点与到达站点名称，调用百度placeapi路径规划功能（http://lbsyun.baidu.com/index.php?title=webapi/guide/webservice-placeapi）以获得该次地铁出行的时长。③ 将每次地铁出行前的出发时刻和出行时长进行加总，即可获得到达时刻。最终获得共5565条地铁出行记录，每条记录包含出发站点、出发时刻、到达站点、到达时刻、出行时长与出行目的等出行特征。由于以上学、购物、外出就餐、休闲娱乐、就医等为目的的地铁出行占比较少，因此将其汇总成为“其他”类型，最终汇总为3类，如表2所示。

2.2.2 POI数据

POI数据用于表征地铁站空间位置以及地铁站点周边土地利用特征。POI数据主要包括：① 2015年百度POI数据，用于反映各类设施空间分布情

表1　所用数据来源与简要信息

Tab. 1　Data sources and brief description

数据类型	数据描述	数据年份	数据来源
居民出行调查数据	居民一日出行链	2015年（对应2014年北京市居民出行情况）	北京市交通委员会（http://jtw.beijing.gov.cn/）
SCD智能卡数据	共计约1434万条地铁出行数据	2018年（7月1日至7月7日）	北京市交通委员会（http://jtw.beijing.gov.cn/）
百度POI数据	用于反映城市服务设施的空间分布情况	2015年	百度地图开放平台（http://lbsyun.baidu.com/）
地铁站点数据	北京市地铁站点空间分布情况	2014年、2018年	北京地铁（https://www.bjsubway.com/）
住房交易价格数据	单位面积成交价格	2015年	北京链家网（https://bj.lianjia.com/）

表2 交通调查数据中不同出行目的地铁出行数量及占比

Tab. 2 Number and proportion of metro trips by purpose intraffic survey data

出行目的	样本数量/条	占比/%
回家	3270	58.76
其他	498	8.95
上班	1797	32.29
总计	5565	100.00

况。② 地铁站点数据，来源于北京地铁官网，用于反映地铁站空间位置，其中地铁站点坐标通过地理编码方式获取。地铁站点数据包含2014年、2018年的数据，分别用于对应交通调查数据及SCD数据年份。③ 2015年住房交易价格数据，数据来源为北京链家网（https://bj.lianjia.com/），用于反映地铁站点周边的住房价格。

本文使用POI数据，从就业机会、住房机会与公共服务设施3个方面表征地铁站点周边土地利用特征。其中表征就业机会数据来源于百度POI数据，包括高收入工作相关兴趣点，如银行、写字楼、政府机构、教育培训等，以及低收入工作相关兴趣点，包括园区、农林园艺、厂矿等；表征住房机会POI数据来源于百度POI数据和住房交易价格数据，住宅区、宿舍等百度POI数据用于表征住房密度，住房交易价格数据用于表征地铁站周边房价；公共服务设施主要是指生活设施，来源于百度POI数据，包括娱乐场所、医院、学校、商店、餐馆、购物中心、开敞公共空间（公园广场）等类型POI。

对POI数据的具体处理方法为：使用ArcGIS软件，分别以地铁起始和到达站点为圆心，以800 m为半径划定缓冲区，计算在缓冲区范围内特定类型的POI核密度值之和。选取缓冲区内特定类型兴趣点的核密度值之和、而非数量的原因在于：核密度值代表被计算要素在其周围邻域中的密度，因此相对于直接计算缓冲区范围内不同类型兴趣点数量，不仅能够反映缓冲区范围内不同类型的兴趣点密度、同时也能够反映缓冲区周边一定空间范围内不同类型兴趣点的密度信息。各类地理数据核密度的空间分布如图1所示。另外，利用地铁站到天安门的欧式距离，代表每次出行起始和到达地铁站点的到北京市中心的远近。

2.2.3 智能卡数据

SCD主要用于带入训练完成的分类器，识别每一次地铁出行对应的出行目的，并展示不同目的的

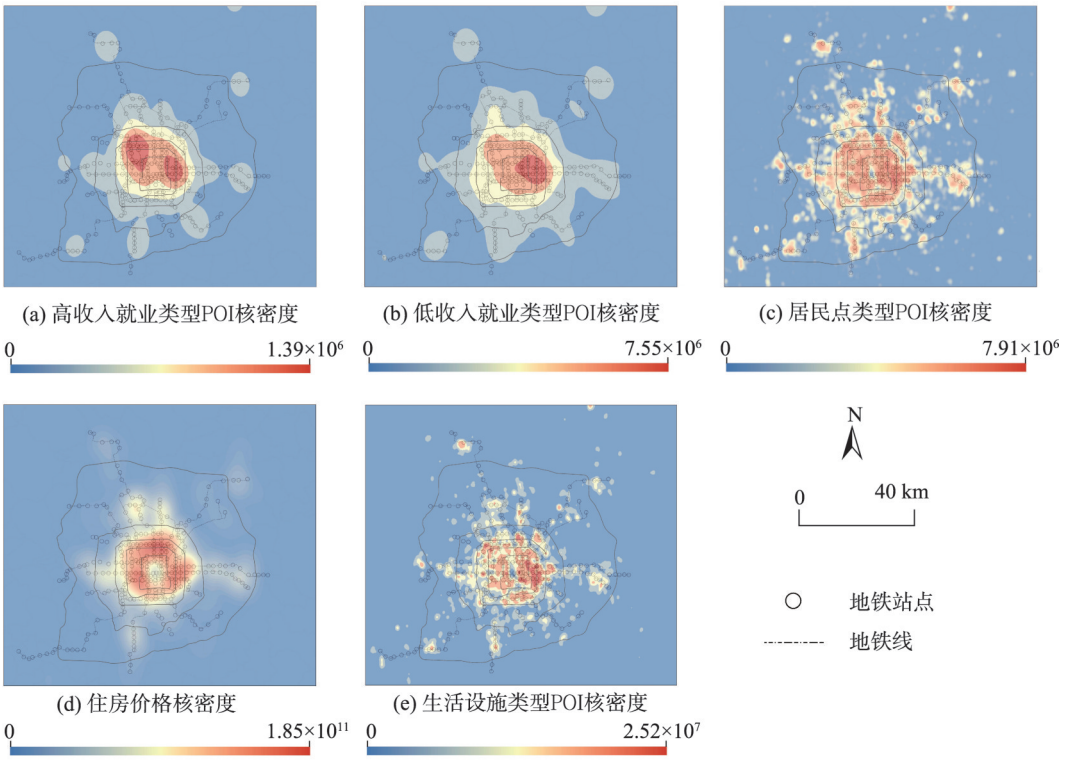


图1 各类型POI空间分布特征

Fig. 1 Spatial distribution of land-use related variables

地铁出行时空间特征。研究使用的SCD覆盖北京市2018年7月1日至2018年7月7日一周内共14 336 909次地铁出行，可有效代表北京市地铁出行的时空间分布规律。SCD包含每条刷卡记录对应ID，起止站点、出发时刻和到达时刻等信息。与以往研究中使用的公交刷卡数据不同[14-15,17]，本文采用的SCD不包含持卡人ID信息，而仅仅记录某一次地铁出行，无法获取同一持卡人一周内的活动轨迹。本文通过使用Microsoft SQL Server对一周内居民地铁出行记录进行统计，得到地铁出行时空间分布特征，如图2所示。

2.3 研究方法

2.3.1 地铁出行目的识别路线

实现地铁出行目的识别，共需5个步骤。① 基于居民出行调查数据筛选居民地铁出行记录，并提取出每条地铁出行记录的起点站、终点站等位置信息，以及出发时刻、到达时刻、出行时长、出行目的等出行特征信息。② 根据地铁站点位置信息，使用POI数据，分别对居民出行调查数据与SCD中每次出行起止地铁站周边土地利用情况进行表征。③ 采用简单随机抽样法，将地铁出行样本划分为训练集与测试集。基于训练集数据，以地铁出行出发时刻、到达时刻、出行时长等出行特征，出行起始站点和到达地铁站点周边的土地利用特征，以及出行目的为预测变量，对RF分类器进行训练。④ 使用样本测试集对RF分类器效果进行检验。⑤ 将2018年北京市SCD中每条记录对应出行变量与始末站点周边土地利用特征、空间位置变量带入RF分类器中，获得每条记录的出行目的识别结果。其中，步骤①与步骤②具体计算过程已分别在交通调查数据与POI数据部分介绍。步骤③至步骤⑤如图3(a)—图3(c)所示。与此同时，表3对RF分类器中所选取特征进行具体描述。

2.3.2 技术方法

本文主要采用RF算法，该算法由Breiman[31]在2001年提出，对非平衡和缺失数据具有较强的容忍度，能够有效分析处理高维、存在共线性或相互作用的数据，具有很高的预测准确率。在有监督分类算法中，RF是一种相对精确且时间复杂度较低的算法[32-33]。

RF是一种基于决策树的集成算法，属于Bag-

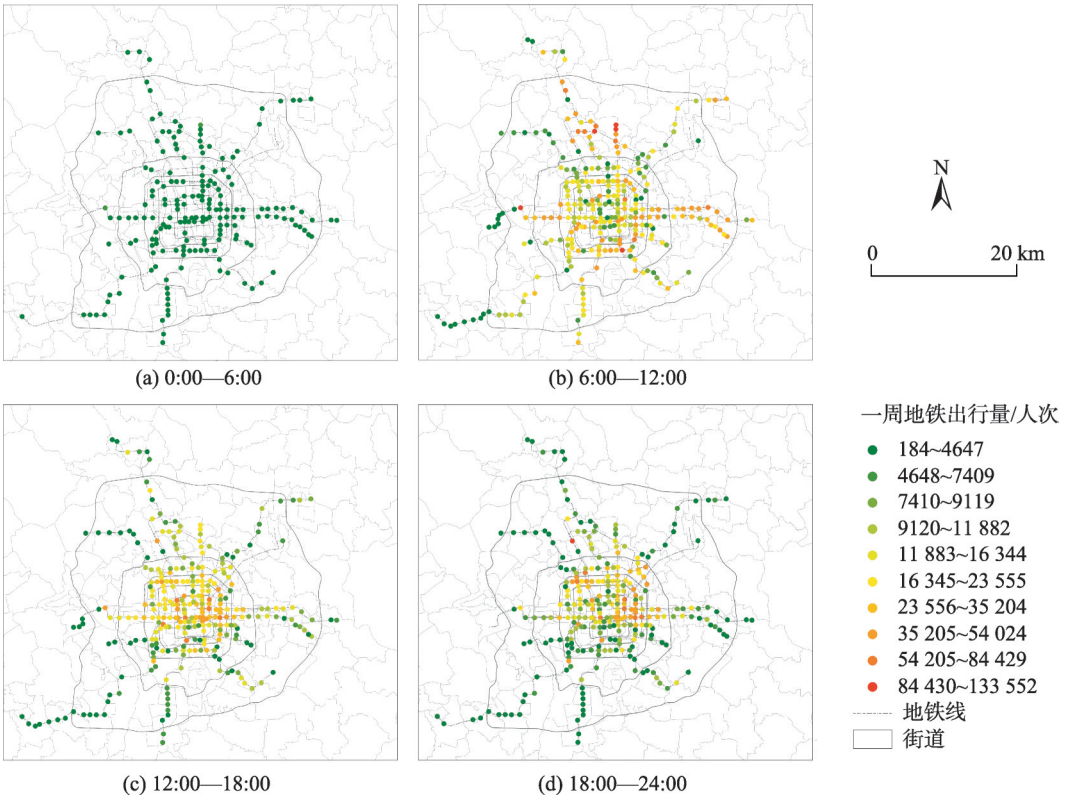
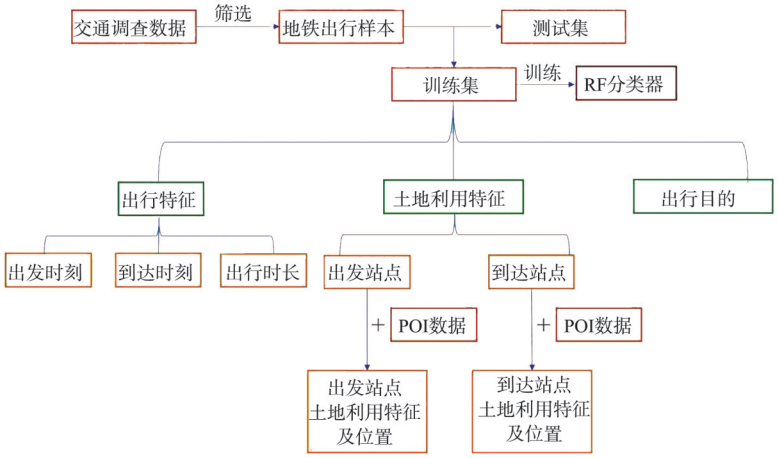
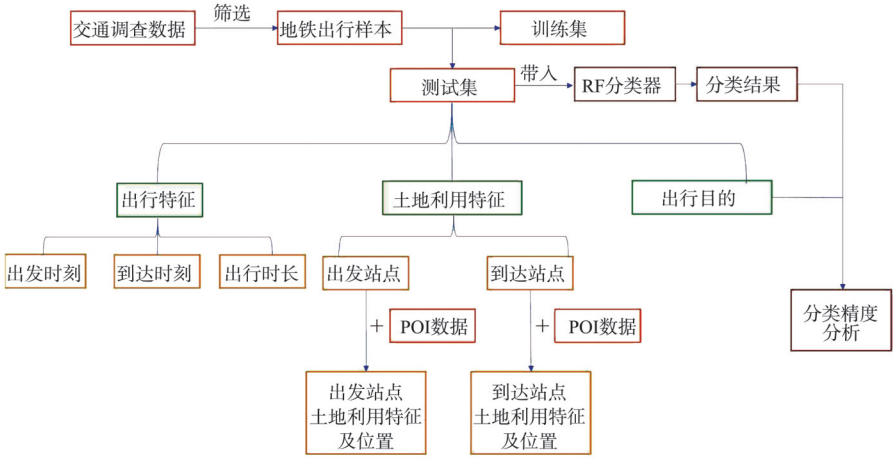


图2　一周内地铁平均出行记录时空分布

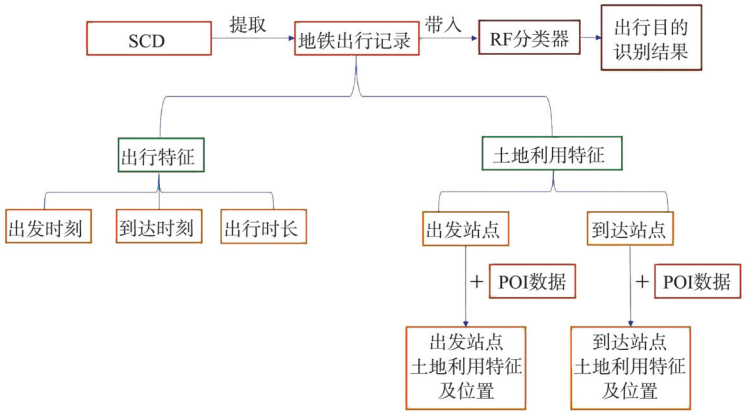
Fig. 2　Spatial distribution of metro trip records in a week



(a) RF分类器训练过程



(b) RF分类器精度评估



(c) 地铁出行目的识别

图3 地铁刷卡数据出行目的识别研究路线

Fig. 3 Schematic diagram of estimating trip purpose of the smart card transactions

ging(Bootstrap Aggregating,引导聚集算法)类型。它首先在原始数据集上通过有放回抽样重新选出k个新数据集;随后根据选出的k个数据集训练k个决策树并组成RF分类器,并以决策树投票的方式决定分类结果,将得票最高的类别为最终标签。基于以上过程,RF算法可以提升抗过拟合能力以及识别精度。

表3 RF分类器包含特征

Tab. 3 Variables included in the random forest classifier

特征名称	特征描述
出行目的	被识别变量(上班、回家、其他)
出行特征	出发时刻、到达时刻、出行时长
土地利用特征	起止点周边高收入、低收入工作场所类型POI核密度值 起止点周边居民点类型兴趣点与住房价格核密度值 起止点周边公共服务与生活服务设施类型POI核密度值 起止点到市中心欧式距离

在RF对原始数据集进行有放回抽样过程中，原始数据集中每个样本未被抽取的概率p为$(1-1/n)^n$，当n足够大时，p约等于0.368，此时大约有37%的样本未被抽取，这些未被抽取的样本统称为袋外样本(Out of Bag，OOB)。利用袋外样本可以对RF中每棵决策树进行精度估计，另外，对RF分类其中所有决策树OOB精度估计取平均，也可得到RF自身性能的泛化精度估计[34]。

RF可以对变量重要性进行度量，作为变量选择的依据。平均准确率的减少(Mean Decrease Accuracy，MDA)是RF特征选择方法之一，是一种基于OOB的误差估计方法。某一特征的MDA数值越大，说明RF估计精度下降越多，该特征越重要[35]。某一特征v对应的MDA具体计算方法为：① 训练随机森林模型，利用袋外样本数据测试模型中每棵决策树的OOB误差；② 随机打乱袋外样本数据中该特征的数值，并重新测试每棵决策树OOB误差；③ 计算2次测试的同一决策树OOB误差差值的平均值，即为单棵树对变量v重要性的度量值，计算公式(式(1))：

$$MDA(v)=\frac{1}{nTrees}\sum_{t=1}^{nTrees}(errOOB_t-errOOB_t^{'}) \quad (1)$$

式中：$MDA(v)$为RF分类器特征v的“平均准确率的减少”指标数值；$nTrees$为RF分类器中决策树的数量；$errOOB_t$为决策树t的OOB误差；$errOOB_t^{'}$为决策树t在随机打乱袋外样本数据后OOB误差重新测试结果。

RF分类器也可以使用准确率、混淆矩阵等方法，通过将测试集带入训练完成的RF分类器中，评估其分类精确程度。其中，准确率也称精度，是指分类准确的样本数占测试集样本总数的比例。准确率仅能用于评估分类器总体分类精度，对于每种识别类型对应结果，可以用混淆矩阵(Confusion Matrix)进行判断。混淆矩阵又被称为可能性表格或是错误矩阵。它是一种用来可视化算法性能的特定矩阵，多用于监督学习。其每一列代表预测值，每一行代表实际值，因而可以非常直接地显示多个类别预测值与实际值之间的混淆情况，便于计算分类器每个类别的预测准确率[36]。本文基于R语言，使用R studio编译平台，实现RF分类器训练和地铁出行目的识别过程。

3 结果及分析

本文使用5565个地铁出行样本对分类器进行训练。为能进一步验证RF分类器的准确性以及土地利用特征对于提升模型分类精度的有效性，本文需要有真实值对RF分类器进行检验。虽然RF分类器可以基于OOB误差衡量有效性，不需要划分训练集与测试集。但通过划分出的测试集，可以更加直观地对比仅包含出行特征RF分类器与同时包含土地利用特征及出行特征RF分类器的分类精度。因此研究采用简单随机抽样法，将5565个样本的75%作为训练集，其余25%作为测试集，分别对RF分类器进行训练和测试。首先基于训练集对RF分类器进行特征重要性评估、选择以及参数标定。

3.1 特征重要性评估

基于出行特征与土地利用特征的地铁出行识别过程中，较高维度的特征可以提高自动识别的精度，但过多的特征会提升RF算法时间复杂度与空间复杂度，不适合的特征也会降低RF分类器精度。为避免以上问题，基于MDA指标，对RF分类器中采用的特征进行重要性评估，避免将不合适的特征纳入其中，特征评估结果(MDA值)如图4所示。所有特征MDA平均值为41.57，中位数为40.36，标准差为12.95。在选择的特征中，出发时刻、到达时刻出行时长等出行特征变量的MDA值最大，分别为76.44、60.93、45.98；其余特征的MDA数值(即重要性)比较接近，其中住房价格核密度与高收入就业机会类型POI核密度等土地利用特征MDA值相对较小，起始站点住房价格核密度、起始站点高收入就业机会类型POI核密度、到达站点住房价格核密度、到达站点高收入就业机会类型POI核密度MDA值分别为27.57、26.05、32.04、31.79。另外，从MDA值也可以看出，土地利用特征对于

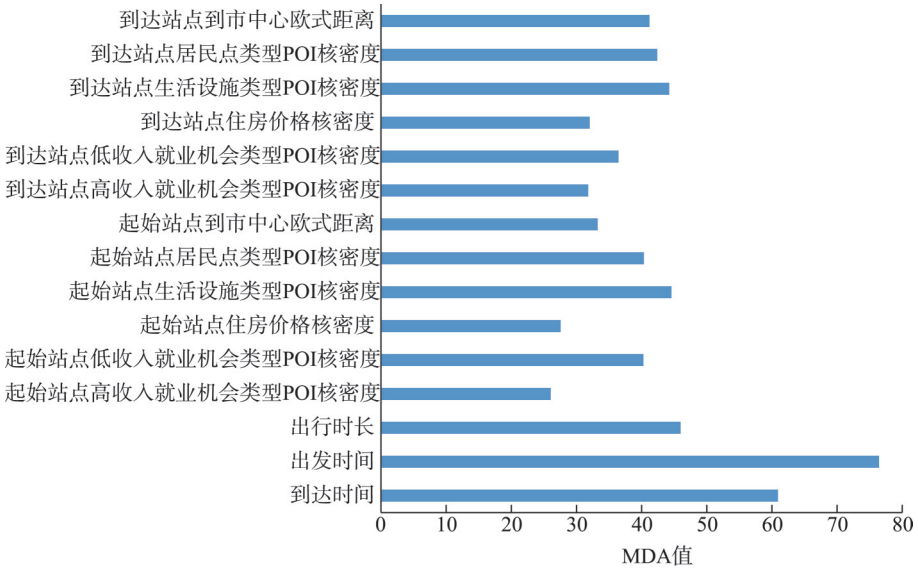


图4 RF分类器中不同特征MDA值

Fig. 4 MDA values of different feature importance in the RF classifier

RF分类器具有重要作用，在一定程度上印证了城市交通与土地利用的时空间互动关系。

3.2 RF分类器特征选择与参数标定

RF分类器作为集成分类器，需要同时对选取特征数量以及决策树数量进行参数标定，使得泛化OOB估计精度最大，从而让RF分类器获得整体上最佳分类效果。在标定特征数量N时，通过计算*nTrees*为100、200、400、800时不同的N值对应的OOB精度的均值，进行参数N的标定实验，结果如图5所示。其中变量的选取按照MDA值进行从大到小排序，随着N的增加，依次将剩余特征中具有最大MDA值的特征纳入RF分类中，并进行OOB精度计算。由图5可以看出，除特征数量N为3之外，特征数量N为其他数值时RF分类器泛化OOB精度变化不大，当N等于15时，RF分类器泛化OOB精度最高，约为0.898。另外，不同特征的MDA值绝对值差异不大，因此将所有出行特征与土地利用特征均纳入RF分类器中，N标定为15。

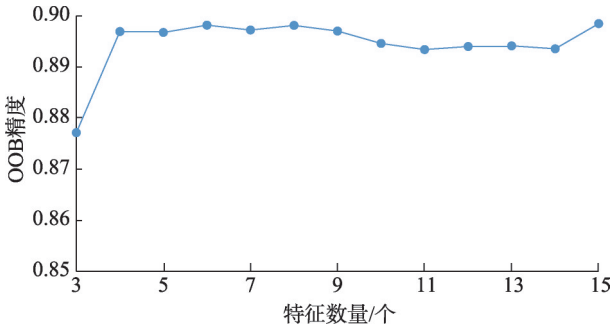


图5 RF分类器预测精度随特征数量N变化情况

Fig. 5 The OOB accuracy of the RF classifier changes with the number of features

随后对RF分类器中决策树数量进行标定，首先在RF分类器中输入选取的15个特征，改变决策树数量*nTrees*的值，计算不同决策树数量对应OOB精度，结果如图6所示。根据图6，RF分类器精度总体上先随着分类器中决策树的数量增加而增加，而后逐渐稳定。当决策树数量大于800时，OOB精度不再随着决策树增多或有较大的波动，最终RF分类器中决策树数量*nTrees*为800。

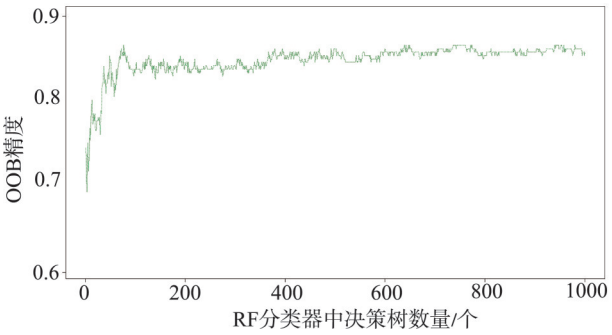


图6 随机森林分类器训练收敛情况及最佳树数量判断

Fig. 6 Convergence of training of random forest classifiers and judgment of the optimal number of trees

3.3 RF分类器训练与地铁出行目的识别结果

在标定RF分类器特征数量N与决策树数量*nTrees*后，本文选取准确率和混淆矩阵评估RF分类

器分类精度，以及地铁出行目的识别的可行性。结果如表4所示，RF分类器分类结果错误率为8.99%，即正确率为91.01%，预测准确率能够达到较高的标准。

表4　随机森林分类器混淆矩阵结果

Tab. 4　Random forest classifier confusion matrix results

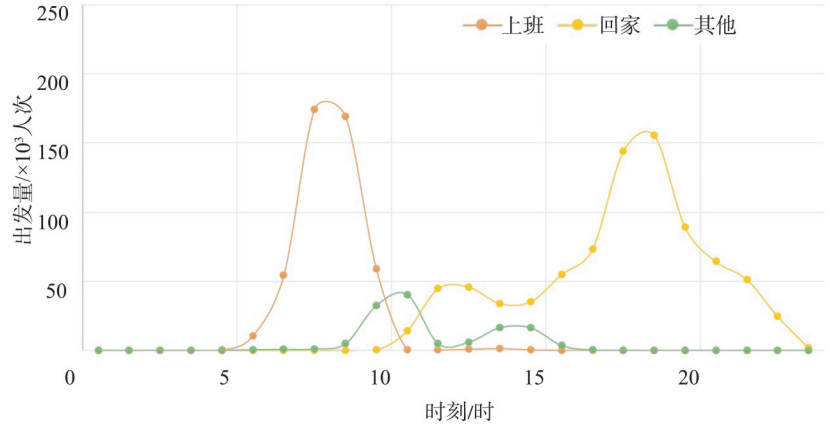
	样本数量/条			样本占比/%
	分类结果为“回家”	分类结果为“其他”	分类结果为“上班”	预测准确样本占比
真实值为“回家”	782	38	14	93.76
真实值为“其他”	8	45	23	59.21
真实值为“上班”	0	42	439	91.27

分类型结果方面，混淆矩阵如表4所示，对于每种类型的出行目的，通过混淆矩阵可观察出每个类别分类结果的准确率。可以发现，除“其他”这一类由于样本量较少、准确率相对较低（接近60%）外，“上班”、“回家”2类出行目的的预测准确率均超过90%，预测精度较高[36]。

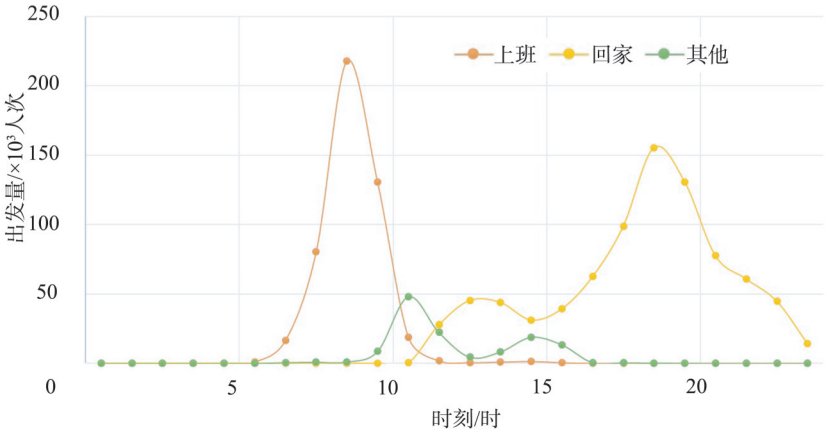
以上针对RF分类器训练结果相关指标的总结，可以说明RF分类器相对准确可靠，可以用来进行进一步分析。随后将地铁刷卡数据中每条出行记录出行时长、出发时刻、到达时刻始末站对应土地利用特征变量带入分类器中，识别每条刷卡记录的出行目的，其时空间分布特征如图7、图8所示。

3.4　交通与土地利用时空间互动理论映证

本文通过对比仅包含出行特征RF分类器与同时包括出行特征与土地利用特征的RF分类器的分类精度，说明土地利用特征对于提升RF分类器训练精度的有效性，以进一步验证交通与土地利用时



(a) 不同出行目的出发量时间分布



(b) 不同出行目的到达量时间分布

图7　不同出行目的地铁出行出发量与到达量时间分布特征

Fig. 7　Temporal distribution of metro trip departures and arrivals for different travel purposes

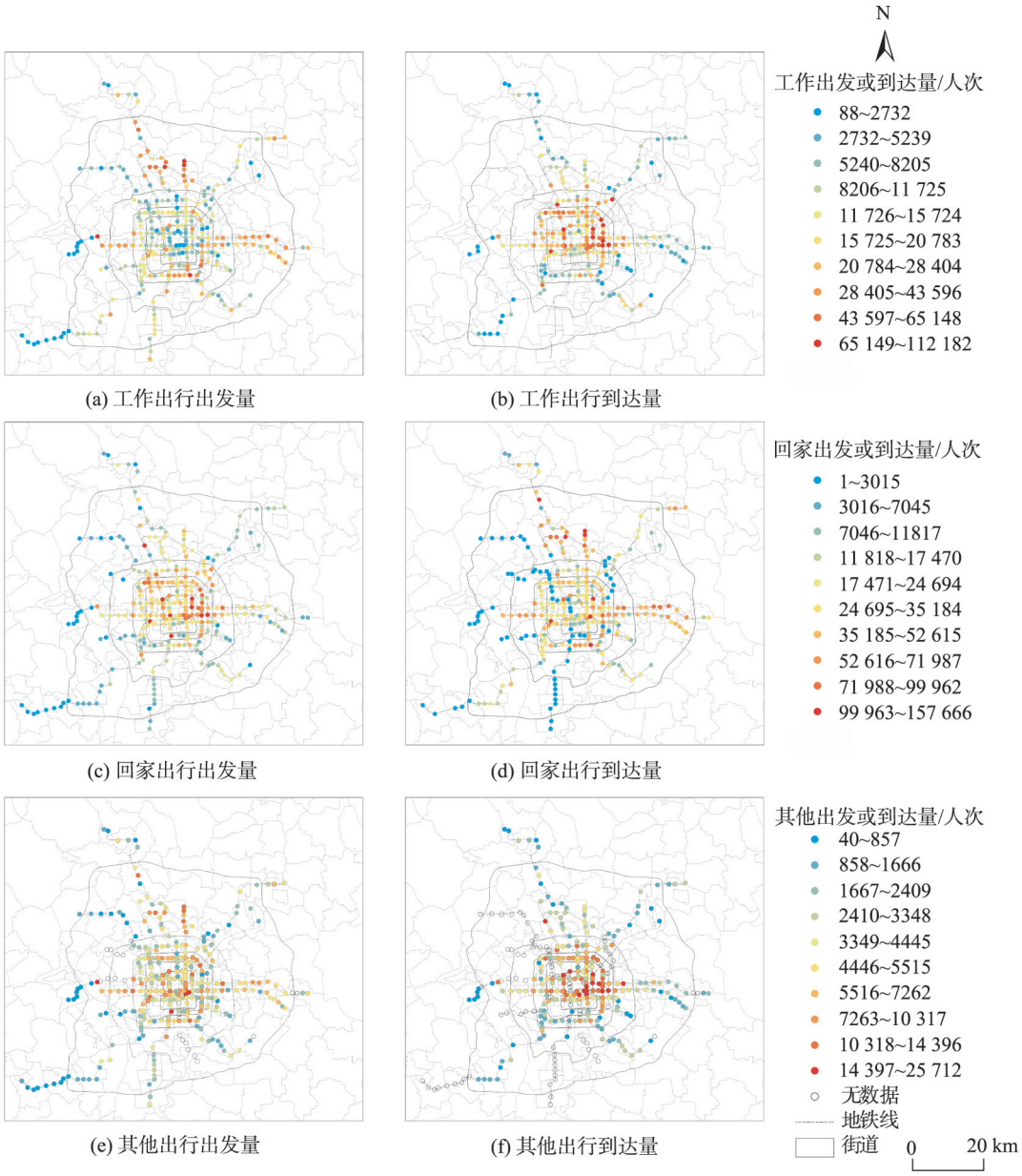


图8 不同出行目的地铁出行出发量与到达量空间分布特征

Fig. 8 Spatial distribution of metro trip departures and arrivals for different travel purposes

空间互动理论。其中仅包括出行相关特征RF分类器准确率为87.99%，低于同时包含出行特征与土地利用特征RF分类器模型91.01%的准确率；混淆矩阵如表5所示，仅包括出行特征RF分类器相较于同时包含出行特征与土地利用特征的RF分类器，在3种出行目的的分类准确率上均有下降。因此可以认为，土地利用特征可以提升RF分类器分类精度，从而进一步映证了交通与土地利用时空间互动理论。

4 结论与讨论

4.1 结论

研究基于交通与土地利用时空间互动理论，采用包括居民出行调查数据、POI数据、SCD等数据集的多源地理大数据对地铁乘客出行目的进行识别。首先通过交通调查数据与POI数据，获取地铁出行样本，并提取每个样本对应出行特征与土地利用特征，训练RF分类器；随后通过SCD数据与POI

表5 仅包括出行特征随机森林分类器混淆矩阵结果及对比

Tab. 5 Comparison of confusion matrix betweenrandom forest classifier with or without travel-related variables

	样本数量/条			样本占比/%	
	分类结果为“回家”	分类结果为“其他”	分类结果为“上班”	仅包含出行特征分类准确样本占比	初始RF分类器准确样本占比
真实值为“回家”	765	38	15	93.52	93.76
真实值为“其他”	22	31	33	36.05	59.21
真实值为“上班”	3	56	428	87.89	91.27

数据，提取SCD中出行记录对应的出行特征与土地利用特征，并带入至完成训练的RF分类器中，以识别出行记录对应的出行目的。为提升分类器对于非平衡与缺失数据的容忍度，提升识别准确率并降低时间、空间复杂度，采用RF算法，并基于OOB误差进行特征筛选与参数标定。与此同时通过准确率与混淆矩阵，对RF分类器效果进行评估。最后基于对SCD中出行记录的出行目的识别结果，对不同出行目的的地铁出行记录时空间特征进行可视化展示。文章主要结论包括：

（1）用于地铁出行目的识别的RF分类器效果较好，基于测试集的检验结果，RF分类器准确率可以达到91.01%；混淆矩阵结果表明“回家”、“上班”、“其他”3类出行目的的识别准确率分别为93.76%、91.27%、59.21%，除去“其他”这一项由于地铁出行样本较少而识别准确率较低外，“回家”、“上班”两类出行目的类别识别准确率较高。

（2）土地利用特征可以提升RF分类器的分类精度。仅包括出行特征RF分类器准确率为87.99%，低于同时包括出行特征和土地利用特征RF分类器91.01%的准确率。与此同时，仅包括出行特征RF分类器混淆矩阵结果中，“回家”、“上班”、“其他”3类出行目的识别准确率分别为93.52%、36.05%、87.89%，三者均低于同时包括出行特征及土地利用特征的RF分类器的结果。因此，土地利用特征可以提升RF分类器的分类精度，这一结论也同时映证了城市交通与土地利用的时空间互动理论。

4.2 讨论

研究基于RF算法探索地铁出行目的的识别方法，并获得了较好的分类效果。同时，也存在理论和应用层面的优势以及不足，具体包括：

（1）在理论层面，本文运用了交通与土地利用时空间互动理论，在RF分类器中纳入土地利用特征，具有更好的效果。相对于依赖经验判断的地铁出行目的识别方法，有助于进一步加强理论支撑，增强方法的可信度。

（2）在实际应用层面，在线地图平台的开放性、大城市交通出行调查的普及性以及智能卡的逐步推广，均决定了多源地理大数据的易获得性，因此本文地铁出行目的识别方法具有较强推广性。文中地铁出行目的识别方法可以被应用于居民地铁出行预测与监控、地铁线网布局和地铁周边土地利用规划等方面，具有实践意义。

（3）研究中“其他”目的的地铁出行样本占比较小，使得该类别识别准确率较低，仅为59.21%。这将造成SCD中地铁出行记录出行目的识别结果具有一定误差。另外，“其他”出行目的中包含了上学、购物、外出就餐、休闲娱乐、就医等出行目的类别，掌握这部分出行目的对应的地铁出行时空间分布规律，对于城市商业、休闲娱乐、教育医疗等土地利用规划及设施布局具有重要参考价值。但受限于样本数量，本文中使用的RF分类器不能精确识别出以上出行目的，无法进一步挖掘相关信息，在一些具体应用场景中仍存在局限。

参考文献(References):

[1] 王姣娥,杜方叶,靳海涛,等.基于交通出行链的就医活动识别理论框架与方法体系[J].地球信息科学学报,2020,22(4):805-815. [Wang J E, Du F Y, Jin H T, et al. Identifying hospital-seeking behavior based on trip chain data: Theoretical framework and methodological system[J]. Journal of Geo-information Science, 2020,22(4):805-815.]

[2] Ali Safwat K N, Magnanti T L. A combined trip generation, trip distribution, modal split and trip assignment model[J]. Transportation Science, 1988,22(1):14-30.

[3] Handy S. Methodologies for exploring the link between urban form and travel behavior[J]. Transportation Research Part D: Transport and Environment, 1996,1(2):151-165.

[4] Maria Kockelman K. Travel behavior as function of accessibility, land use mixing, and land use balance: Evidence

from San Francisco Bay Area[J]. Transportation research record, 1997,1607(1):116-125.

[5] 赵鹏军,万婕.城市交通与土地利用一体化模型的理论基础与发展趋势[J].地理科学,2020,40(1):12-21. [Zhao P J, Wan J. The key technologies of integrated urban transport-land use model: Theory base and development trends [J]. Scientia Geographica Sinica, 2020,40(1):12-21.]

[6] 陆化普,孙智源,屈闻聪.大数据及其在城市智能交通系统中的应用综述[J].交通运输系统工程与信息,2015,15(5):45-52. [Lu H P, Sun Z Y, Qu W C. Big data and its applications in urban intelligent transportation system[J]. Journal of Transportation Systems Engineering and Information Technology, 2015,15(5):45-52.]

[7] 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. [Long Y, Zhang Y, Cui C Y. Identifying commuting pattern of Beijing using bus smart card data[J]. Acta Geographica Sinica, 2012,67(10):1339-1352.]

[8] 韩昊英,于翔,龙瀛.基于北京公交刷卡数据和兴趣点的功能区识别[J].城市规划,2016,40(6):52-60. [Han H Y, Yu X, Long Y. Identifying urban functional zones using bus smart card data and points of interest in Beijing[J]. City Planning Review, 2016,40(6):52-60.]

[9] Zhang Y, Martens K, Long Y. Revealing group travel behavior patterns with public transit smart card data[J]. Travel Behaviour and Society, 2018,10:42-52.

[10] Liu K, Yin L, Ma Z, et al. Investigating physical encounters of individuals in urban metro systems with large-scale smart card data[J]. Physica A: Statistical Mechanics and its Applications, 2019:123398.

[11] Bagchi M, White P R. The potential of public transport smart card data[J]. Transport Policy, 2005,12(5):464-474.

[12] Pelletier M P, Trépanier M, Morency C. Smart card data use in public transit: A literature review[J]. Transportation Research Part C: Emerging Technologies, 2011,19(4):557-568.

[13] Morency C, Trepanier M, Agard B. Measuring transit use variability with smart-card data[J]. Transport Policy, 2007,14(3):193-203.

[14] Ma X, Wu Y J, Wang Y, et al. Mining smart card data for transit riders' travel patterns[J]. Transportation Research Part C: Emerging Technologies, 2013,36:1-12.

[15] Long Y, Thill J C. Combining smart card data and household travel survey to analyze jobs-housing relationships in Beijing[J]. Computers, Environment and Urban Systems, 2015,53:19-35.

[16] Kusakabe T, Asakura Y. Behavioural data mining of transit smart card data: A data fusion approach[J]. Transportation Research Part C: Emerging Technologies, 2014,46:179-191.

[17] Ma X, Liu C, Wen H, et al. Understanding commuting patterns using transit smart card data[J]. Journal of Transport Geography, 2017,58:135-145.

[18] 赵鹏军,李南慧,李圣晓. TOD建成环境特征对居民活动与出行影响——以北京为例[J].城市发展研究,2016(6):45-51. [Zhao P J, Li N H, Li S X. The impacts of the built environment on residents' acitivities and travel behavior in TOD areas: A case study of Beijing[J]. Urban Development Studies, 2016(6):45-51.]

[19] 文萍,赵鹏军,周素红. TOD对居民通勤模式的影响:以英国泰恩威尔都市区为例,国际城市规划,2019(8):1-15. [Wen P, Zhao P J, Zhou S H. Effects of TOD on residents' commuting patterns: Empirical evidence from tyne and wear, the UK[J]. Urban Planning International, 2019(8):1-15.]

[20] Cervero R. Transit-oriented development in the United States: Experiences, challenges, and prospects[C]. TCRP Report, 2004,102.

[21] Zhao P, Zhang Y. The effects of metro fare increase on transport equity: New evidence from Beijing[J]. Transport Policy, 2019,74:73-83.

[22] Cervero R, Day J. Suburbanization and transit-oriented development in China[J]. Transport Policy, 2008,15(5):315-323.

[23] Zhao P, Li S. Bicycle-metro integration in a growing city: The determinants of cycling as a transfer mode in metro station areas in Beijing[J]. Transportation Research Part A: Policy and Practice, 2017(99):46-60.

[24] Zhao P, Li P. Travel satisfaction inequality and the role of the urban metro system[J]. Transport Policy, 2019,79:66-81.

[25] Olaru D, Smith B, Taplin J H E. Residential location and transit-oriented development in a new rail corridor[J]. Transportation Research Part A: Policy and Practice, 2011,45(3):219-237.

[26] Li S, Zhao P. Exploring car ownership and car use in neighborhoods near metro stations in Beijing: Does the neighborhood built environment matter?[J]. Transportation research part D: Transport and Environment, 2017,56:1-17.

[27] Zhao P, Li S. Suburbanization, land use of TOD and lifestyle mobility in the suburbs[J]. Journal of Transport and Land Use, 2018,11(1):195-215.

[28] Alonso W. Location and land use: Toward a general theory of land rent[J]. Economic Geography, 1964,42(3):11-26.

[29] Hansen W G. How accessibility shapes land use[J]. Journal of the American Planning Association, 1959,25(2):73-76.

[30] 赵鹏军,孔璐. TOD对北京市居民通勤影响及其机制研究[J].人文地理,2017,32(5):125-131. [The impact of TOD on residents' commuting activities: A case of Beijing. Human Geography, 2017,32(5):125-131.]

[31] Breiman L. Random forests[J]. Machine learning, 2001,45(1):5-32.

[32] Liaw A, Wiener M. Classification and regression by randomForest[J]. R news, 2002,2(3):18-22.

[33] 曹泽涛,方子东,姚瑾,等.基于随机森林的黄土地貌分类研究[J].地球信息科学学报,2020,22(3):452-463. [Cao Z T, Fang Z D, Yao J, et al. Loess landform classification based on random forest[J]. Journal of Geo-information Science, 2020,22(3):452-463.]

[34] 方匡南,吴见彬,朱建平,等.随机森林方法研究综述[J].统计与信息论坛,2011,26(3):32-38. [Fang K N, Wu J B, Zhu J P, et al. A review of technologies on random forests [J]. Statistics & Information Forum, 2011,26(3):32-38.]

[35] Han H, Guo X L, Yu H. Variable selection using Mean Decrease Accuracy and Mean Decrease Gini based on Random Forest[C]. IEEE International Conference on Software Engineering & Service Science. IEEE, 2016: 219-224.

[36] Townsend J T. Theoretical analysis of an alphabetic confusion matrix[J]. Perception & Psychophysics, 1971,9(1): 40-50.