基于城市交通小区的共享单车投放量预测

杜书

【摘要】共享单车是一种新型分时租赁模式的共享经济。为了确定单车资源的合理配置方式，共享单车企业需要提前预测从一个区域到另一个区域的骑行需求数量，这个问题被表述为原点—目的地矩阵预测（ODMP）问题。为了有效地解决这一问题，本文提出对单车订单数据的普遍性处理思路，针对2019年8月上海市哈啰单车订单数据进行处理分析，结合上海市行政区划数据、土地利用及社会经济属性等对原始特征进行衍生，揭示共享单车使用的时空特征。应用先进的机器学习技术XGBoost模型，分析OD需求影响因素，预测未来OD需求量。预测结果结合交通设施承载力、用户出行成本、车辆与调度成本，可作为共享单车投放决策的数据依据，为相关企业带来经济效益的同时减轻城市交通负担，平衡供需关系，促进城市交通的可持续发展。

【关键词】共享单车；OD需求；交通规划；短时交通流量预测

1 引言

当前，我国城市交通体系正面临诸多严峻挑战，包括空气污染、道路拥堵、噪音污染以及频繁的交通事故等问题，这些问题与机动车保有量的持续增长紧密相关[1-2]。在这种背景下，步行和自行车等慢行交通方式变得尤为重要。公共自行车系统以其便捷性、环保性、高效率以及解决“最后一公里”问题的能力而受到广泛关注[3]。公共自行车系统经历了从政府运营的有桩模式到企业主导的无桩共享单车模式的转变，并逐渐演化为基于互联网的共享单车时代，这一转变显著推动了共享单车的普及。然而，共享单车的发展也面临着车辆随意停放、供需不平衡及城市道路交通管理等问题，这些问题突显了进行科学规划和管理共享单车投放的迫切需求[4]。

有效的单车投放规划不仅需要预测未来的共享单车需求量，还需考虑如何优化车辆分布、提高使用效率并减少资源浪费。这包括设置合理的停车点、平衡车辆在高需求区域的供应以及制定应对高峰时段和特殊事件的策略[5]。机器学习方法，如神经网络、支持向量机、聚类分析和回归树分析等，为这些预测和规划任务提供了强大的工具[6]。Parsa[7]利用XGBoost和Catboost多个模型分别分析预测未来某区域内交通事故发生的数量，通过对预测结果的准确率分析，表明Boost方法尤其是XGBoost在预测时间序列数据有着较强的优势。Kidando等人[8]通过对美国佛罗里达州的主要公路的数据进行分析，根据地理信息、道路占有率等重要交通特征，利用改进的Bayesian理论在线性模型方法，完成了交通流预测的任务。陈喜群[9]等针对路网本身的几何拓扑关系，结合交通流数据的时间相关性以及空间相关性，将图卷积网络理论运用到道路网，以交叉口为节点同时以路段为边，建立了城市道路交

叉口短时流量预测模型，得到了较为准确的预测结果。杜圣东等[10]在原始 LSTM 方法的基础上进行改进，加入了编码层以及解码层，并且结合注意力机制（Attention mechanism）来实现模型端到端的功能，注意力机制的引入可以显著提升模型对序列中重要部分的建模能力，从而提升了模型的性能和对长序列的处理能力。模型可以自主学习到与城市交通流量相关的特征，再通过特征分析和衍生操作，得到了非常优秀的预测效果。

本研究将对相关理论进行改进拓展，结合实际出行数据，通过对交通小区的共享单车流量进行精确预测，更加真实的了解共享单车用户的出行需求，为相关企业的决策支持和政府的交通管理部门提供更为科学的数据依据，从而优化共享单车的整体运营效率，实现更可持续的城市交通解决方案[11]。

2 基于共享单车使用特性的出行数据分析与预处理

2.1 原始数据结构

共享单车原始出行数据具有巨量订单和信息的特点，要从这些海量数据中提取所需的信息，首先需要了解共享单车出行数据的结构。本研究使用的共享单车出行数据包含上海市 2019 年 8 月的哈啰出行数据，一条完整的哈啰单车出行记录包括 10 个部分：订单号、用户 ID、车辆 ID、骑行起始、结束日期时间、骑行起始、结束的经纬度和骑行轨迹。如表 1 所示是共享单车原始出行数据清单。

表 1 共享单车原始出行数据清单

序号	字段	描述
1	orderid	订单号
2	userid	用户 ID
3	bikeid	车辆 ID
4	start_time	骑行起始日期时间
5	end_time	骑行结束日期时间
6	start_location_x	骑行起始经度
7	start_location_y	骑行起始纬度
8	end_location_x	骑行结束经度
9	end_location_y	骑行结束纬度
10	track	骑行轨迹

2.2 出行数据预处理

数据预处理指的是对原始出行数据进行清洗，去除重复数据、处理缺失值和异常值。再根据需求和领域知识，进行特征提取和构造以及特征衍生。常见的数据预处理步骤包括数据清理、数据集成、数据变换。其中数据清理是对共享单车原始数据集中的缺失值、异常值、离群点、噪声进行处理；数据集成是将多个数据源中的数据结合成一致的数据存储；数据变换主要在于利用 Python 语言将单车数据使用的国测局坐标系(GCJ02)转换为 GPS 原始坐标系(WGS84)。

由于原始数据的特征较少，在进行建模之前，需要对原始数据进行特征衍生，可以将原始特征转换成更加有用或具有独特解释力的特征。可以通过组合两个或多个特征来创建交互项，使用特征工程技巧，如多项式特征、离散化、分箱等来生成新的特征，有助于提高模型的预测能力，帮助发现数据中潜在的关联关系。表 2 所示为特征工程处理后的特征。

表 2 特征工程处理后的特征

序号	特征	特征描述
1	distance	订单距离
2	trackpoints	订单定位轨迹点数量
3	duration	订单持续时间
4	year	年份
5	month	月份
6	day	该月的第几天
7	weekdays	是否为工作日
8	dayofweek	该周的第几天
9	quarter	季度
10	period_5	该天第几个 5 分钟段内
11	period_15	该天第几个 15 分钟段内
12	day_part	凌晨、早晨、上午、中午、下午、傍晚、晚上、深夜
13	count_user	该订单用户本月使用共享单车次数
14	count_bike	该订单共享单车本月被使用次数
15	cross_region	订单是否跨行政区

3 基于多源数据处理的共享单车特征提取与衍生

3.1 每日单车使用次数分析

将共享单车数据进行特征衍生，得到如图 1 所示的每日单车次数。从总体趋势上看，每日单车使用次数在一个月内呈现出上升的趋势。开始时使用次数较少，但随着时间的推移，使用次数逐渐增加。

具体地，通过统计一个月中每一天的订单单数，可以得到这批数据的一些变化趋势。在月初（1 号至 7 号），单车的使用次数相对较低，在 2000 次以下；在月中（8 号至 21 号），使用次数逐渐增加，在 2000 至 4000 之间波动；从 22 号开始到月底，单车的使用次数明显增加，超过 4000 次，并在 31 号达到近 5000 次。

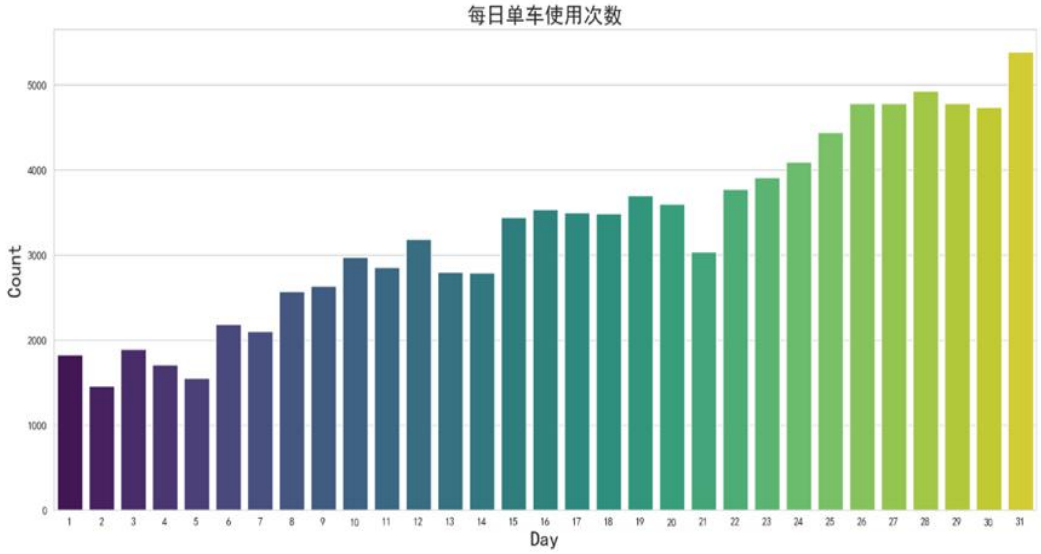

图 1 每日单车使用次数图

我们可以得出以下结论：

(1)单车使用次数随着时间的推移呈现明显的增长趋势；

(2)月初和月中的使用次数相对较低，可能与工作日和周末的差异有关；

(3)月底的使用次数明显增加，可能与人们的出行需求增加有关；

(4)可能存在周期性的变化，例如每周末使用次数增加，工作日使用次数减少。

3.2 共享单车不同时段使用量分析

通过对订单在一天中开始的时刻进行处理，统计各个时段的订单单数，可以得到单车的使用规律。如图 2 折线显示了一周中每天不同时间使用共享单车的计数。所有工作日的线条在早晨 6 点左右都有一个显著的峰值，同样晚上 18 点左右也有另一个峰值；反映出早高峰和晚高峰的时候使用共享单车的人最多；在两个高峰期之前的时间段内，共享单车的使用量都呈现上升趋势，在结束两个高峰期之后共享单车的使用量呈现下降趋势；在工作日和周末

的使用模式可能不同：工作日可能更加集中在上下班时间，而周末可能更加分散。周末的单车使用量大多数情况下都要低于周一到周五，除了 10 点到 16 点的时间段，这个时间段可能的原因是由于周末路况较为复杂，因此很多人选择了单车出行。

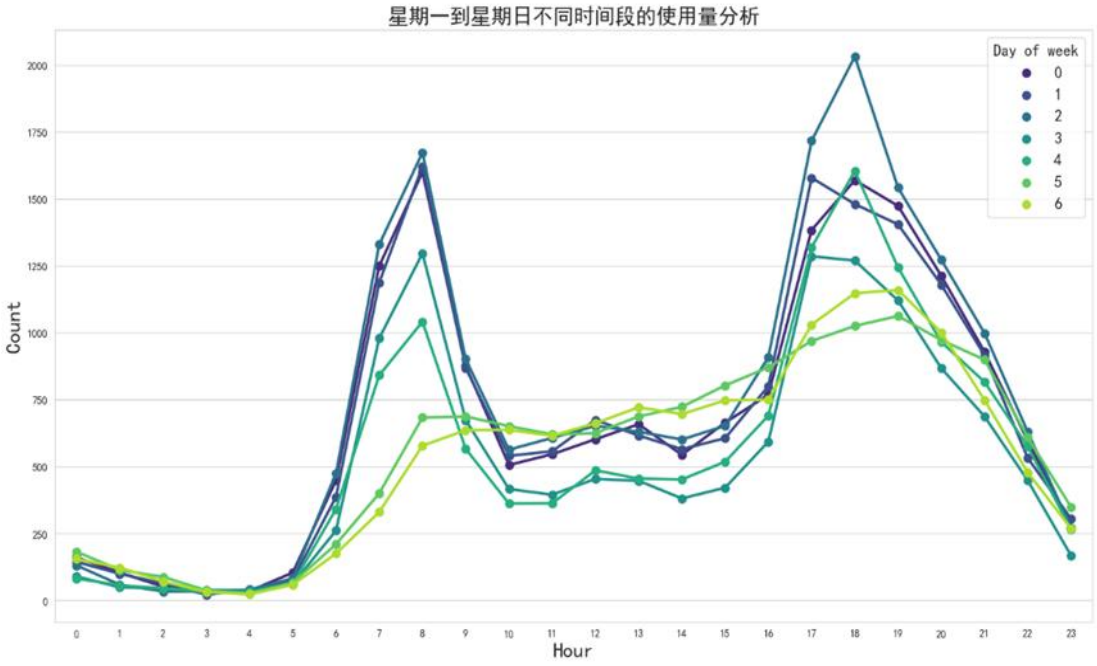

图 2 共享单车不同时段使用量变化

3.3 骑行距离比例分布与用户使用频次分析

原始数据中提供了订单的经纬度坐标，通过经纬度可以计算每个订单的距离，我们又做了骑行距离比例分布与用户使用频次分析的工作

从图 3 的骑行距离比例分布饼形图，我们可以得出：在骑行不同距离范围的比例分布上，最大的部分是 0-1 km，其次是 1-2 km，然后是 2-3 km，3-4 km 和超过 4 km 的部分较小；反映出绝大多数共享单车的使用者的骑行距离范围为 0-2 km。

从图 4 的用户使用频次饼形图，我们可以得出：图由三个部分组成，每部分用不同的颜色表示：蓝色代表 0-5 次，黄色代表 5-10 次，绿色代表大于 10 次；蓝色区域占据了饼形图的大部分面积，说明大多数用户使用频次在 0-5 次之间，黄色和绿色区域较小，表示使用频次在这两个范围内的用户较少。

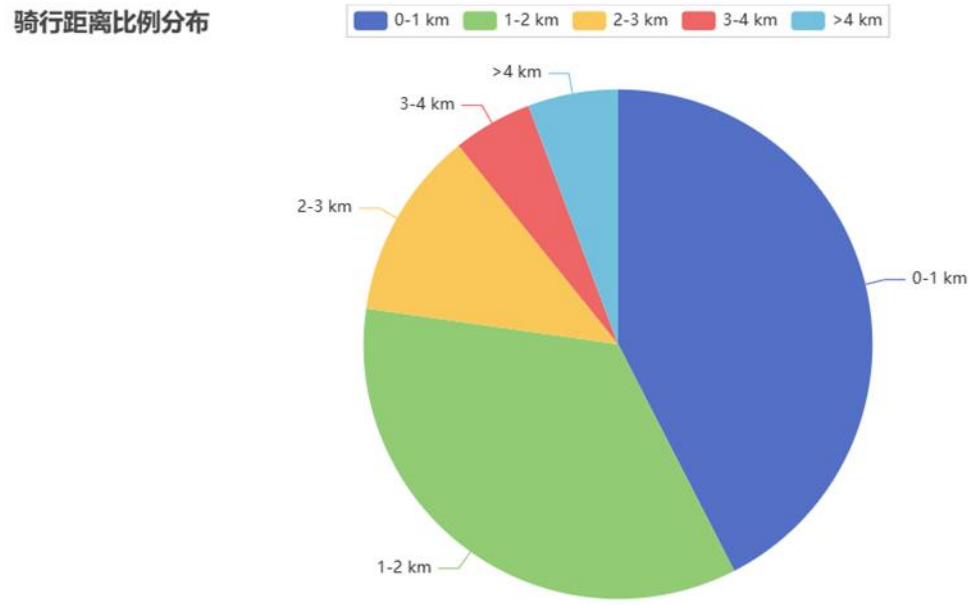

图 3 骑行距离比例分布

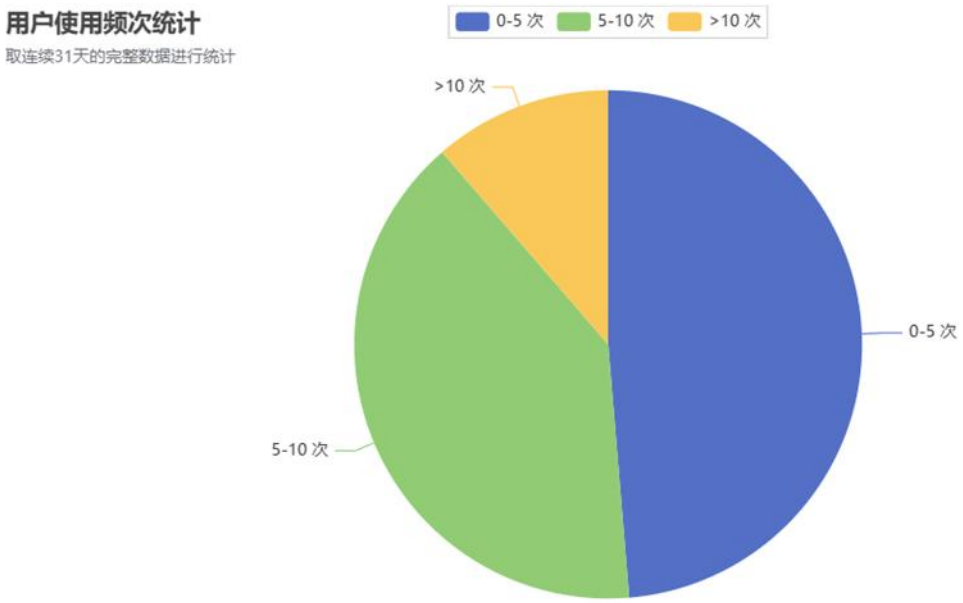

图 4 用户使用频次统计

3.4 高峰时段订单起点位置分析

根据共享单车数据对早晚高峰时间的订单起始点位做了 3D 图分析。早高峰时段为 8-9 点，单车起点位置分布如图 5；晚高峰时段为 18-19 点，单车起点位置分布如图 6。单车起点主要集中在一定区域内，个别行政区的订单量稀疏。Z 轴表示数量，颜色渐变表示不同数量的单车起点。大部分单车起点都集中在 Z 轴上 5 以下的区域，说明起点重合的概率是较小的。红色和黄色线条表示单车订单高频发生的点位。

从早晚高峰时间的订单起始点位分布可以得出：早晚高峰时段，人们更倾向于从特定区域开始骑行。这些热门区域可能是工作地点、学校、景点或其他活动场所。

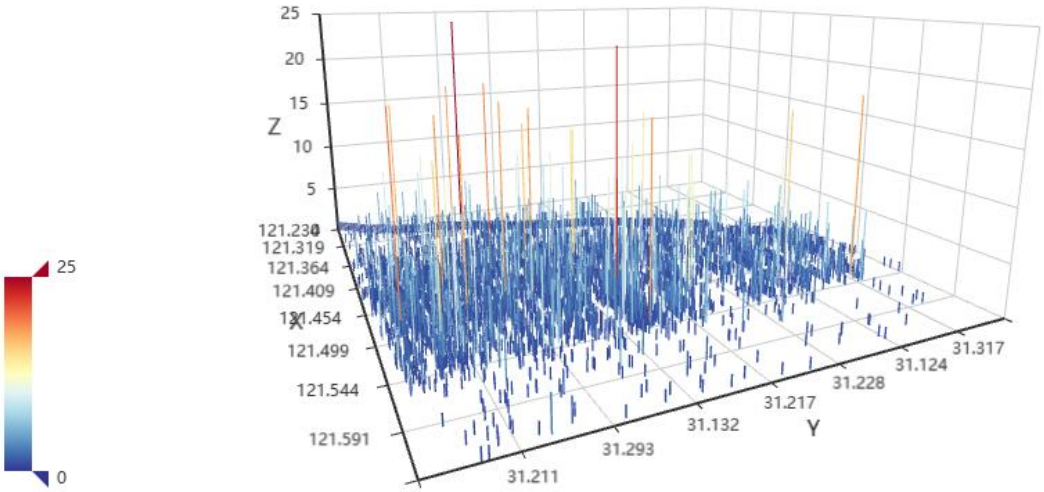

图 5 8-9 点早高峰单车起点位置分布

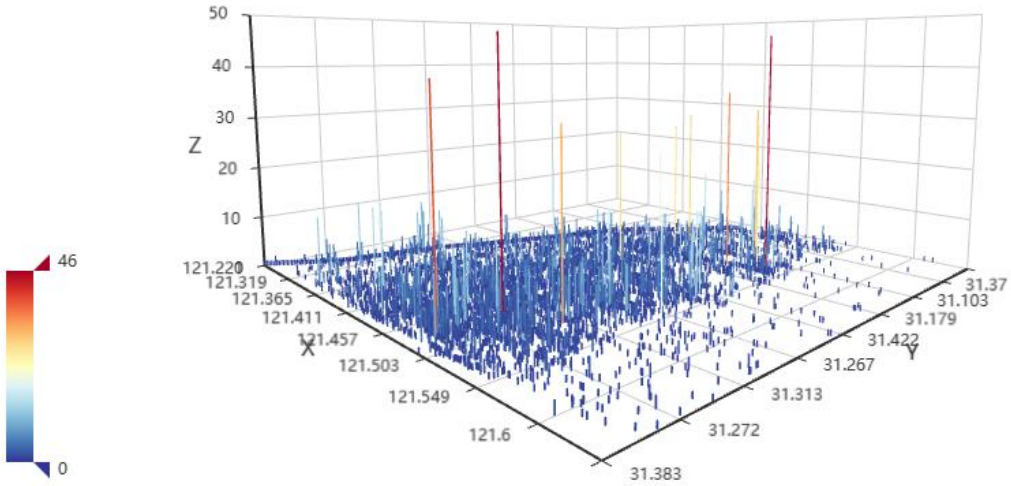

图 6 18-19 点晚高峰单车起点位置分布

3.5 订单持续时间分析

箱型图是由一组或多组连续型定量数据的来反映数据的分布情况的统计图。它的优势在于可以显示五个数据，包括最小值、25%位点值、中位值、75%位点值、最大值，这五个值一起组成了共享单车数据的分布形状、分布和集中趋势。它用于反映原始数据分布的特征，还可以进行多组数据分布特征的比较。如图 7 所示，为单车订单持续时间箱型图。

数据分布情况上，从箱型图中，可以看到不同时间段开始的订单的完成时间呈现出一定的变化。每个小时的箱子表示该小时内订单完成时间的四分位数范围。颜色从黄色渐变到紫色，随着时间的推移而变化。这意味着不同时间段的订单完成时间存在差异。

离散程度上，箱子和箱须的长度反映了订单完成时间的离散程度。如果箱子和箱须较长，说明该小时内订单完成时间的变化较大。

异常值上，黑色点表示离群值，即在某些小时内完成时间明显超过大多数订单的异常情况。这些离群值可能需要进一步调查，以确定其原因。

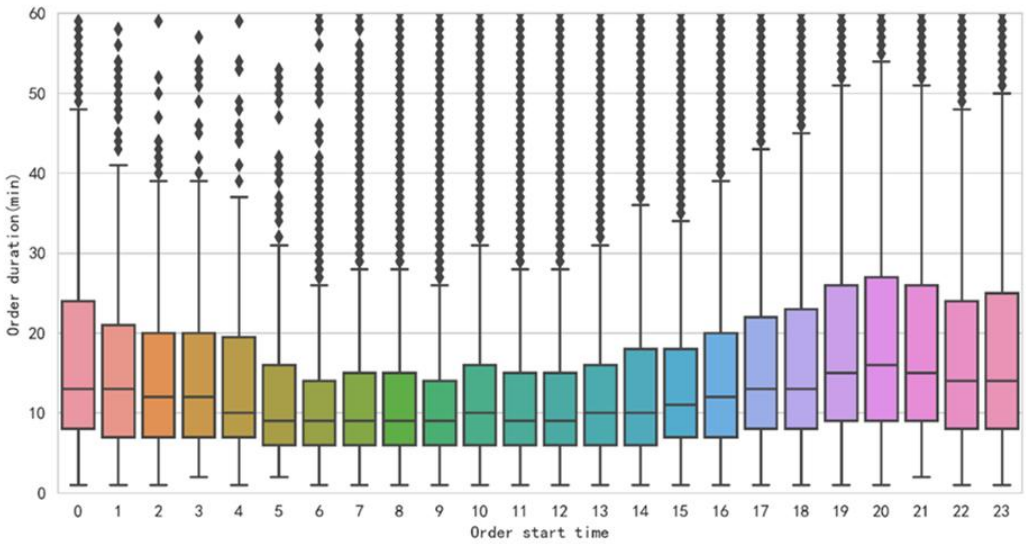

图 7 单车订单持续时间箱型图

另外，对共享单车不同时段使用量分析和订单时间进行箱型图分析，利于找到共享单车订单的显著特点，作为根据不同交通区和时间段投放共享单车方案的数据依据。

4 基于 XGBoost 的共享单车投放量需求预测

4.1 分区域的栅格化与 OD 流量统计

通过阿里云平台得到上海行政区划数据（geopandas 形式），对上海行政区划地图进行可视化处理，在地图上展示上海市的行政区边界。如图 8 所示，每个多边形代表一个行政区，它们是由爬取的数据中的‘geometry’列绘制而成，为后续对各区域进行栅格化和 OD 流量统计作基础。

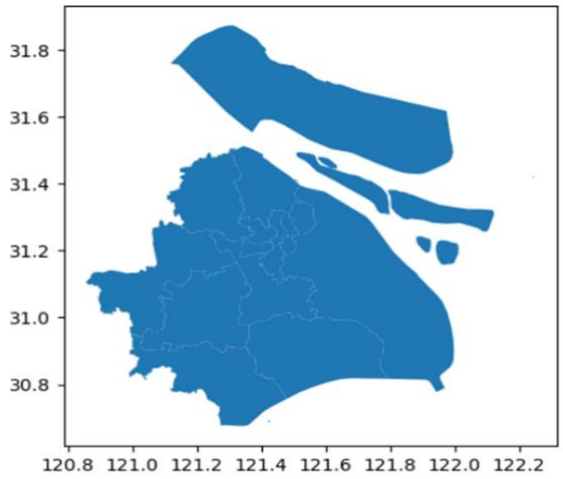

图 8 上海市行政区划可视化

数据栅格化处理是将地理信息数据转换为栅格数据的过程。这一步是结合原始订单的经纬度数据转化为交通小区 OD 流量的重要过程，在地理信息系统中，栅格化处理将连续的地理空间数据按照一定的分辨率和参考系统转换为离散的栅格数据集，结合其他来源数据后，每个栅格单元对应一个特定的空间区域，并包含该区域的属性信息。为了根据订单的经纬度坐标进行统计，得到各个区域内的 OD 流量。将地理区域划分为均匀的网格并进行编号，并且以小正方形的中心作为栅格的点位。如图 9 所示，得到栅格化处理后的上海行政区划地图可视化效果。

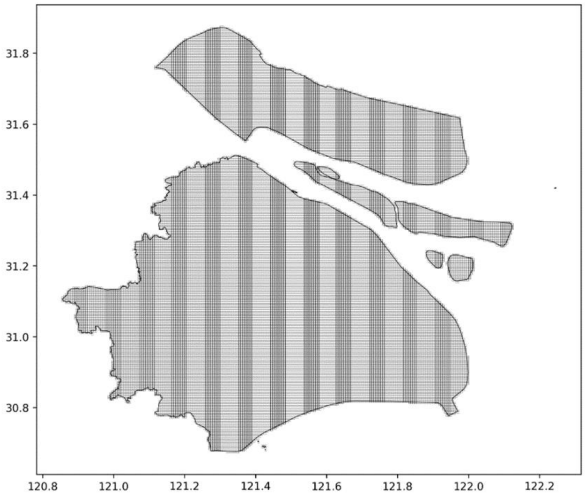

图 9 栅格化处理后的上海行政区划地图

栅格化提供了一种将多源数据集合并到统一框架中的方式。它使城市交通管理者能够整合共享单车出行数据、上海市的行政区划分数据、气候数据等。原始数据经过栅格化处理后才能转化为各个行政区的流量数据，是高效地应用机器学习模型来预测不同区域之间的骑行

流量前提，同时简化了复杂的地理形状，使得模型训练和预测过程更加迅速和精确，对评估骑行需求和计划共享单车投放策略至关重要。

4.2 XGBoost 机器学习模型预测

XGBoost（Extreme Gradient Boosting）是一种灵活、高效和准确的梯度提升决策树框架，在分类和回归预测等机器学习任务中被广泛运用。XGBoost 是基于原始梯度提升机（GBM）的构想，同时引入了许多创新的改进，是 Boosting 算法的一种实现方式。选取 80%数据进行训练,20%进行模型测试，使用了 XGBoost 包的分类器和 Scikit-learn 库，经过交叉验证后选择的最优超参数值，得到合理的预测结果，实现基于城市交通小区的共享单车投放量测算并且进行特征分析，对模型的参数设置如表 3 所示。

表 3 模型训练参数

模型训练参数	值
max_depth	3
min_child_weight	5
gamma	0
subsample	0.6
colsample_bytree	0.6
objective	squarederror
lambda	0
eta	0.05
seed	27

表中，Booster 表示指定学习算法的类型，这里设置为’gbtree’，表示使用基于树的模型。Max_depth 表示每颗树的最大深度，树的深度较大会使模型更复杂，可以学到更多复杂的关系。但是，深度过大容易导致过拟合。Min_child_weight 用来定义叶子节点上所有样本权重之和的最小值。用于控制叶子节点的分裂，较大的值能够防止过拟合。Gamma 表示控制叶子节点分裂的条件，只有当分裂后损失函数的减少大于 Gamma 时，才允许分裂，Gamma 值较大可以保守地进行分裂。Subsample 表示训练每颗树时使用的样本的比例。范围在(0,1]之间。较小的值使模型更加保守，防止过拟合。Colsample_bytree 表示训练每颗树时使用的特征的比例。范围在(0,1]之间。较小的值使模型更加保守，防止过拟合。Objective 表示指定学习任务的目标函数，这里设置为 Squarederror，表示回归问题，使用均方误差作为损失函数。Lambda 表示控制模型复杂度的 L2 正则化权重。较大的值使模型更加保守。Eta 表示学习率，即每一步权重调整的大小。较小的值可以使模型更加保守，缓解过拟合问题。Seed 表示随机种子，用于伪随机数生成。如果设置了相同的值，将会得到可复现的结果。

由于 OD 对数量太多（一共是 16×16=256 个），以及部分 OD 对之间流量较小，数据存在稀疏性，选择订单起终点为徐汇区的 OD_13 为例，得到下月 OD 流量预测结果和真实值基本吻合，说明该模型预测能力优秀，如图 10 所示。

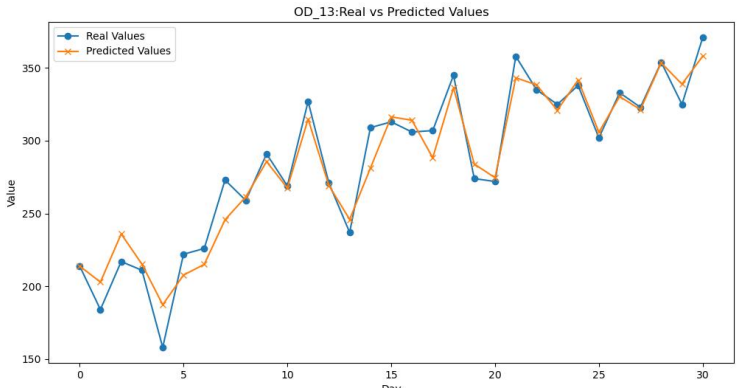

图 10 OD_13 预测结果和真实值对比图

5 结语

从数据上，本研究结合 2019 年 8 月上海市哈啰共享单车出行数据、交通行政分区划分数据、社会经济属性、目标区域气候数据等，对多源数据进行融合，揭示共享单车的使用特征。通过融合不同来源的数据，填补数据之间的空缺，减少数据的不一致性，提高数据的完整性和准确性。多源数据融合整合了不同角度和维度的数据，可以发现不同数据之间的关联和趋势，帮助企业获取更全面的信息和视角，更好地理解和分析数据，提高决策的准确性。

从方法上，本研究使用栅格化将原始订单数据转化为 OD 数据，可以根据企业的实际需求考虑对不同大小区域的投放量预测，从而实现投放方案的规划。

从预测模型的选用上，本研究对比了多种机器学习模型的结果，最终选用表现效果最好的 XGBoost 模型，预测精度高，可靠性强。

共享单车 OD 流量预测对城市交通小区的共享单车投放量测算有重要意义，而共享单车投放量研究又是共享单车系统发展的关键。进行共享单车出行影响机理的分析及预测，对共享单车交通道路网络和停车设施承载力进行测度，分析共享单车交通设施承载负荷；综合考虑交通小区内共享单车的借还需求和交通设施承载力的限制，考虑共享单车用户出行成本、投入车辆成本和调度工作量，最终可实现建立基于城市交通小区的共享单车最优投放方案。

参考文献

[1] 李晔,慢行交通系统规划探讨--以上海市为例[J].城市规划学刊,2008,1(3):78-81

[2] 云美萍,杨晓光,李盛.慢行交通系统规划简述[J].城市交通,2009,7(2):57-59

[3] 周杨,张冰琦,李强.公共自行车系统的研究进展与展望[J].城市发展研究,2014,21(9):118-123.

[4] 北京市规划和国土资源管理委员会.北京城市总体规划(2004 年-2020 年) [J].北京规划建设,2005,(2):5-51.

[5] 朱玮,庞宇琦,王德,等.上海市闵行区公共自行车出行特征研究[J].上海城市规划,2012(6):102-107.

[6] Zeng Q,Huang H L. A Stable and Optimized Neural Network Model for Crash Injury Severity Prediction [J]. Accident Analysis and Prevention 2014,73:35-81.

[7] Parsa A B,Movahedi A,Taghipour H, et al. Toward safer highways,application of XGBoost and SHAP for real-time accident detection and feature analysis [J]. Accident Analysis & Prevention,2020,136:105405.

[8] KIDANDO E,REN M,OZGUVEN E E,et al. Evaluating Traffic Congestion Using the Traffic Occupancy and Speed Distribution Relationship:An Application of Bayesian Dirichlet Process Mixtures of Generalized Linear Model[J]. 交通科技期刊(英文),2017(3):007.

[9] 陈喜群,周凌霄,曹震.基于图卷积网络的路网短时交通流预测研究[J].交通运输系统工程与信息,2020,20(4):7.

[10] 杜圣东,李天瑞,杨燕,等. 一种基于序列到序列时空注意力学习的交通流预测模型[J]. 计算机研究与发展,2020,57(8):14.

[11] 钱进. 城市公共自行车租赁点借还需求预测与分析[D]. 长安大学, 2015.

作者简介

杜书，男，硕士研究生，东南大学交通学院。电子邮箱：1210431058@qq.com