文章编号:1672-5328(2023)04-0087-12　中图分类号:U491.1+4　文献标识码:A　DOI:10.13813/j.cn11-5141/u.2023.0033

基于频繁序列模式挖掘的卡口短时交通量预测

刘冉[1]，李岩[1]，毛海虓[1]，钱剑培[1]，王继峰[1]，马悦[2]
(1.中国城市规划设计研究院，北京 100037；2.住房和城乡建设部标准定额研究所，北京 100835)

摘要：基于数据的城市交通管理和控制方法是广大学者和交通管理部门的关注重点。以频繁序列模式挖掘算法为基础，对卡口车辆轨迹序列进行时空特征分析。选用7种典型的机器学习算法进行预测，并分析了卡口空间区位、交通量以及连接道路等级对预测精度的影响。研究结果表明，集成学习算法特别是RF的预测性能最好，误差较小且训练速度快；SVR和神经网络算法(MLP、LSTM)在预测误差表现上相近，但是基于神经网络算法的预测模型耗时较长。此外，不同模型的预测误差在空间上的分布具有相似性，在卡口密布的区域预测精度更高，在外围边缘区域误差较大；卡口交通量越大、连接的道路等级越高，预测精度越高。随着城市交通电子卡口设备在路网中的完善，该预测方法的准确性可以进一步提高。

关键词：短时交通流量预测；频繁序列模式挖掘；机器学习

Short-Term Traffic Flow Forecasting at Intersections Based on Frequent Sequence Pattern Mining

LIU Ran[1], LI Yan[1], MAO Haixiao[1], QIAN Jianpei[1], WANG Jifeng[1], MA Yue[2]
(1. China Academy of Urban Planning & Design, Beijing 100037, China; 2. Research Institute of Standards and Norms Ministry of Housing and Urban-Rural Development, Beijing 100835, China)

Abstract: Data-based urban traffic management and control strategies are major focus of scholars and traffic management departments. This paper conducts a spatiotemporal features analysis on the trajectory sequences based on frequent sequence pattern mining algorithms. Seven typical machine learning algorithms are employed for short-term traffic flow prediction, and the impact of spatial location, traffic volume, and grade of intersecting roads on prediction accuracy are analyzed. The results reveal that ensemble learning algorithms, particularly the RF model, demonstrate superior predictive performance with smaller errors and faster training speeds. SVR and neural network algorithms (MLP, LSTM) show comparable predictive error performance, but neural network-based models are more time-consuming. Besides, the prediction errors of the model were similar in space. The prediction accuracy is higher in the area where the checkpoints are densely distributed, and lower in the periphery area. The prediction error is lower where traffic volume is larger and the connected roads with higher grades. With the improvements of electronic devices at checkpoint in urban road network, the forecasting accuracy can be further enhanced.

Keywords: short-term traffic flow forecasting; frequent sequence pattern mining; machine learning

收稿日期：2023-01-10
基金项目：国家重点研发计划资助项目“基于城市高强度出行的道路空间组织关键技术”(2020YFB1600500)
作者简介：刘冉(1996—)，女，河南信阳人，硕士，助理工程师，主要研究方向：城市交通规划。E-mail: liuran563491@163.com
通信作者：王继峰(1981—)，男，辽宁朝阳人，博士，教授级高级工程师，注册城乡规划师，城市交通研究分院综合交通所所长，主要研究方向：城市综合交通体系规划、区域交通体系规划、交通与城镇化等。E-mail: wangjifeng@gmail.com

0 引言

随着大数据快速发展以及智能交通电子设施的不断完善，交通调查手段越来越多样化，基于数据的城市交通管理和控制方法逐渐成为广大学者和交通管理部门的关注重点。其中，城市道路高清摄像卡口监控系统通过图像识别技术识别车牌号码，对卡口地

点、时间、车牌号码、进口方向等信息进行记录并上传至终端，形成过车记录数据。车牌号码具有唯一性，因此，相比于传统断面检测设备，卡口监控系统具有记录车牌号码、识别车辆行驶路径，以及在此基础上对交通量进行分析和预测的优势[1]。根据相关研究，卡口设备对车辆号牌自动识别的精度达到90%左右，建成期3年以内的卡口识别率为95%以上[2]，在有效性方面基本满足交通量分析的数据要求。本文通过挖掘历史及实时交通数据中隐藏的潜在规律，对实时交通运行状态进行分析评价，并对未来短时间内交通量和交通状态进行预测，此项研究结果有助于道路交通管理部门对路网进行优化管理，进而提高交通运行效率。

1 研究综述

常见的短时交通流预测方法主要包括参数方法和非参数方法。参数方法基于传统统计学理论且以线性理论为基础，主要包括回归分析[3]、时间序列[4-5]等方法。然而，如今城市交通流的非线性特征越来越突出，导致这类方法的预测精度降低，难以满足需要。非参数方法是基于无数学模型的智能算法，从历史数据中挖掘出输入变量和输出变量之间的映射关系，但映射的具体形式则类似于一个“黑箱”。这种算法可移植性强且具有较高的精度，对于具有不确定性和非线性特性的问题有良好的适应性，逐渐成为短时交通流预测的主流算法，例如K近邻(K-Nearest Neighbor, KNN)、随机森林、支持向量机、神经网络等算法。李翠 等[6]利用K近邻和主成分分析(Principal Components Analysis, PCA)相结合的KNN-PCA法预测高速公路短时交通量，证明了该算法对交通流变化趋势具有良好的预测能力。李浩 等[7]提取了卡口流量数据及其日期、节假日、周期三个方面的特征，提出了基于周期性变化的卡口交通量随机森林(Random Forest, RF)预测模型，并验证了模型的有效性。Sun Z Q等[8]对特征变量进行降维，然后采用支持向量回归(Support Vector Regression, SVR)方法对交通量进行预测，发现模型预测性能良好且运行速度大大提高。此外，集成学习算法(例如极端梯度增强算法(eXtreme Gradient Boosting algorithm, XGBoost)、梯度提升回归树(Gradient Boosting Regression Tree, GBRT)算法)预测交通流的合理性以及对不同交通流状态的良好适应性也得到了检验[9-11]。针对检测器采集数据出现的长度不定、采样不规则以及数据缺失等问题，Tian Y等[12]利用多尺度时间平滑方法对数据进行补齐，并利用长短期记忆神经网络方法(Long Short-Term Memory, LSTM)获得了精度较高的交通流预测结果。

基于非参数方法的短时交通流预测模型日趋成熟，精度高并且运行效率较高。在特征变量的选择方面，早期的研究仅考虑了交通量的时间相关性，通过分析交通流的周期性、波动性规律来筛选对预测时段有较大影响的历史时段交通流作为输入。但是交通流是一个时空维度上的概念，对于卡口交通量而言，车辆在空间中流动，按照时间顺序行经不同的卡口，上下游的卡口流量都会对当前卡口的流量产生影响[13]。因此，很多研究者开始关注联合交通量的“时间—空间”特征并进行预测。李巧茹 等[14]提出一种时空支持向量机模型，以本路段及相邻路段的流量序列作为输入，利用支持向量机进行训练和预测。尹韬霖[15]通过分析与目标路段直接相连的三个相邻路段的相关系数来确定空间相关性，并利用随机森林等集成学习模型进行预测。

由于路网中交通量的时空变化幅度大，且受设备分布不均影响，流量采集无法面面俱到，仅采用相邻路段进行预测不具备适用性。鉴于机动车在空间上行经不同的卡口构成车辆行驶轨迹序列，对机动车出行轨迹序列频繁模式进行挖掘，可以估算出卡口断面在空间上的关联程度，进而探究卡口交通量的空间相关性。陈玲娟[16]提出一种Apriori-LSTM预测模型，基于Apriori算法对轨迹序列进行频繁模式挖掘，将支持度高的路段作为关联路段输入LSTM预测模型，Apriori-LSTM预测结果优于传统的KNN，RNN，LSTM模型。杨冰健[17]基于卡口轨迹数据，利用频繁序列模式挖掘算法计算路段的重要度，挖掘关键路段之间的关联规则，为后续交通流预测奠定基础。贾若[1]基于Apriori算法提取卡口路径特征，构造了基于频繁序列关联规则的卡口路径特征向量，利用KNN、RF、GBRT等算法进行预测，提高了预测的准确性。J Ganapathy 等[18]基于频繁序

列模式挖掘 Prefixspan 算法对钦奈市高速公路上的 3 个收费站流量时间序列进行频繁模式挖掘和预测，发现预测误差明显低于多层神经网络和支持向量算法，但是预测仅仅考虑了目标站点本身流量序列的频繁模式特征，并未考虑站点在空间上的关联特征。之前的研究表明基于频繁序列模式挖掘的算法能够关注交通流的时空流向，充分挖掘空间关联关系，为交通流短时预测模型提供了更加准确、有效的输入，在交通流短时预测方面有着巨大潜力。因此，本研究采用频繁序列模式挖掘算法，基于卡口数据构造车辆行驶轨迹和具有时空特征的输入变量，结合典型的机器学习算法进行交通流短时预测。

2 研究方法

2.1 研究区域及数据

本文选取青岛市市南、市北区由高速公路、快速路、主干路、次干路组成的较高等级路网作为研究对象。市南区和市北区是青岛市典型的老城区，开发强度高，人口和就业岗位密度大，汽车使用较为集中，具有典型的交通拥堵特征，是青岛市交通管理部门重点关注的拥堵防治区域。

使用的卡口过车数据在空间上包括青岛市市南、市北两区(行政区划代码 AREAID 为 370202、370203)的 163 个交叉口、442 台卡口监控设备的过车记录，路网结构和卡口分布情况如图 1 所示；在时间上覆盖 2022 年 1 月 12—26 日，原始数据共计 71 708 634 条卡口过车记录，具备交通流数据的海量特征。卡口过车记录包括车牌号码、采集时间、卡口断面、行政区划等相关信息，本研究中使用的主要是车牌号码、采集时间、采集地点名称、方向编号以及行政区划代码，数据结构如表 1 所示。

原始数据中包括一些重复记录数据，即同一辆车短时间内在同一个卡口检测器被多次记录，因此需对数据进行清洗，对 5 s 内有重复记录的数据仅保留第一条记录。清洗后的过车记录共计 69 191 661 条。

2.2 频繁序列模式挖掘算法

频繁序列模式挖掘是指从指定的序列集合中挖掘出现频率(支持度)较高的序列模式，在考虑序列中各项关联关系的基础上，还关注了序列的先后顺序。常见的频繁序列模式挖掘算法主要包括两类：Apriori 算法和 Prefixspan 算法[19]。Apriori 算法的基本思想是：基于先验知识，通过迭代算法，从长度为 1 的频繁项集开始，生成长度为 2 的候选项集，再根据支持度阈值筛选长度为 2 的频繁项集，以此类推，利用 k 频繁项集生成长度为 $k+1$ 的候选项集，直到不存在更大的频繁项集。Prefixspan 算法的基本思想则是前缀投影，即通过递归策略，从长度为 1 的前缀开始，以频繁项的前缀为基础划分搜索空间，获得与该前缀相应的投影数据库(与前缀对应的后缀的集合)，并在投影数据库中对各项进行支持度统计，得到频繁的项，利用得到的频繁的项和原前缀合并，得到长度为 2 的前缀，接着用新的高一阶的前缀进行递归，不断地对增长的前缀进行投影，直到产生不了新前缀。

为了对比 Apriori 和 Prefixspan 算法的运算效率，本研究从基于不同的序列数据库大

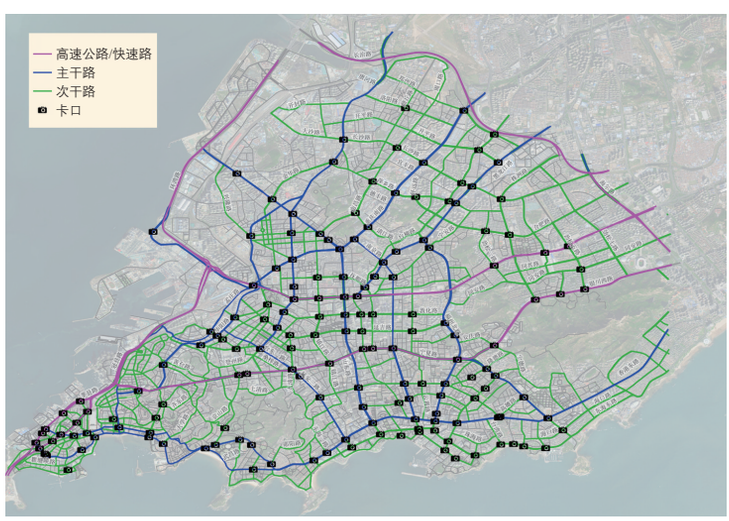


图 1 研究范围内路网及卡口分布情况

Fig.1 Road network and intersections distribution within the study area

表 1 卡口过车记录数据结构

Tab.1 Data structure of traffic records at checkpoints

字段名	字段类型	字段描述	说明
CCARNUMBER	VARCHAR2(16)	车牌号码	
DCOLLECTIONDATE	DATE	采集时间	“YYYY/MM/DD hh:mm:ss”
CCOLLECTIONADDRESS	VARCHAR2(200)	采集地点名称	
NDERICTRION	NUMBER(2)	方向编号	包括东、西、南、北四个进口方向
AREAID	NUMBER(8)	行政区划代码	国家公布的六位县级以上行政区划代码

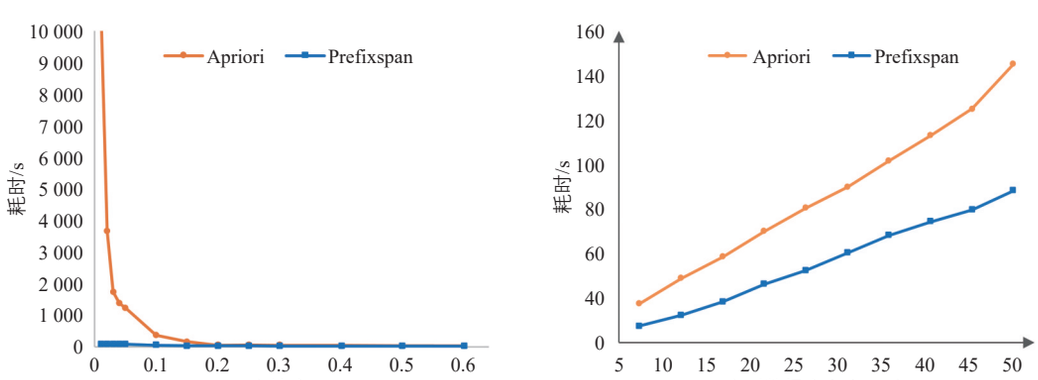
小 N 和不同的支持度阈值 min_support 两个方面统计算法的运行时间，其中序列数据库为从卡口数据中提取的 N 条机动车出行轨迹序列。实验在同一台电脑上完成，硬件平台为 3.60 GHz CPU、16 Gb 内存、Windows 10 操作系统、Python 3.9 编程环境。两种算法运行效率对比结果如图 2 所示。

随着支持度阈值的降低，两种算法耗时均增加，但是 Apriori 算法耗时的增加幅度高于 Prefixspan 算法；在支持度为 0.01 时，Apriori 算法运行耗时超过 3 h，而 Prefixspan 算法耗时在 100 s 以内，说明在支持度较低时 Apriori 算法性能较低，这也验证了 Jian P 等[20]的研究。随着数据集的增加，特别是在数据量较大的情况下，Apriori 算法的耗时明显高于 Prefixspan 算法。Apriori 算法通过逐层迭代，多次对数据库进行扫描并产生候选集，生成的候选集数量庞大且在每次验证候选集时都要对数据库进行扫描，复杂度较大，运行效率较低，因此数据量较大时不适用。基于模式增长策略的 Prefixspan 算法采用分治思想，不断产生序列数据库的多个更小的投影数据库，然后对每个投影数据库单独进行挖掘。由于该算法不产生候选序列，减少了对数据库的扫描次数。因此，数据量较大时 Prefixspan 算法在频繁序列模式挖掘中具有较高的效率，在支持度较低时更为明显。由于本研究卡口过车记录数据量较大且区域卡口较多，因此采取 Prefixspan 算法进行卡口车辆轨迹序列的频繁路径挖掘。

3 卡口交通流时空分布关联特征

3.1 时间分布关联特征

对数据清洗后的每个卡口断面每日的过车记录以 15 min 为间隔进行统计，得到对应的 15 min 交通量。以山东路—延吉路交叉口北进口流量数据为例，图 3 为该卡口在 2022 年 1 月 12—26 日共 15 天内每 15 min 交通量时序变化图。卡口交通量呈现明显的周期性和波动性，在 1 d 和 7 d 的尺度上呈现出周期性，并在每日不同时段有波动变化规律。1)每



a N =200 000, min_support 变化　　b min_support=0.25, N 变化

图 2 两种算法运行时耗对比

Fig.2 Efficiency comparison of the Apriori and Prefixspan algorithms

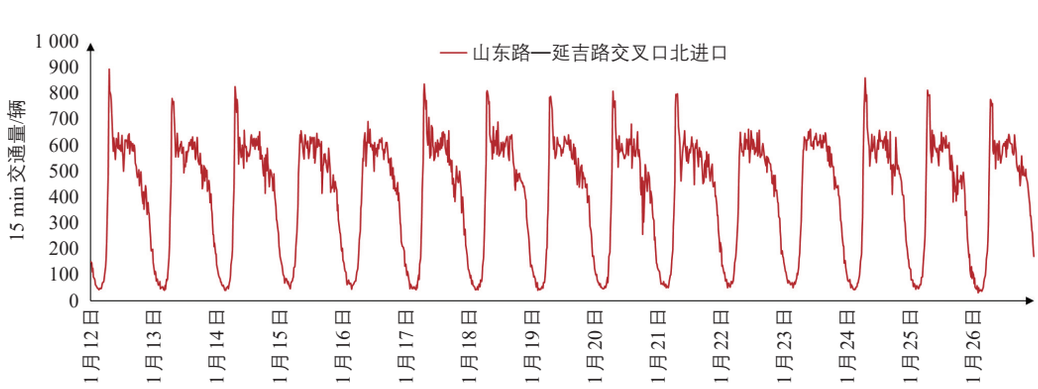


图 3 山东路—延吉路交叉口北进口 15 min 交通量时序变化

Fig.3 Time series of 15-minute traffic flow at the north entrance of the intersection of Shandong Road and Yanji Road

日各时段流量的变化趋势具有相似性，早高峰7:00—9:00达到峰值，日间平稳波动，晚高峰17:00—19:00过后逐渐下降，次日凌晨5:00以后逐渐回升；2)非工作日(15日、16日、22日、23日)的流量峰值明显低于工作日，平峰时段的流量与工作日大致相似。

其余卡口交通流亦呈现类似的规律。对青岛市市南、市北两区其余441个卡口断面在2022年1月12—26日的15 min交通量进行分析，发现研究范围内卡口断面检测到的交通量在时间上均呈现1天和7天的周期性规律。于是可以认为，卡口 $t+1$ 时刻通过的交通量和1天前 t 时刻以及一周前 t 时刻的交通量具有相关性。因此，除了 t 时刻交通量以外，1天前 t 时刻以及一周前 t 时刻的交通量也可作为 $t+1$ 时刻卡口交通量预测的特征变量输入预测模型。

3.2 空间分布关联特征

交通流的流动性决定了流量在空间分布

表2 部分卡口断面的频繁上下游卡口分析结果

Tab.2 Partial results of frequent upstream and downstream traffic at some checkpoint sections

序号	卡口断面	频繁上游卡口		频繁下游卡口	
		卡口断面	支持度	卡口断面	支持度
1	山东路与延吉路交叉口东进口	延吉路与徐州路交叉口东进口	0.73	山东路与敦化路交叉口南进口	0.35
2	山东路与延吉路交叉口西进口	延吉路与镇江北路交叉口西进口	0.45	延吉路与徐州路交叉口 西进口	0.39
3	山东路与延吉路交叉口南进口	山东路与江西路交叉口南进口	0.37	山东路与敦化路交叉口南进口	0.82
4	山东路与延吉路交叉口北进口	山东路与敦化路交叉口北进口	0.71	山东路与江西路交叉口北进口	0.45
5	东海西路与南京路交叉口东进口	东海西路与福州南路交叉口东进口	0.62	东海西路与山东路交叉口东进口	0.63
6	东海西路与南京路交叉口西进口	东海西路与山东路交叉口西进口	0.70	东海西路与福州南路交叉口西进口	0.78
7	东海西路与南京路交叉口南进口	澳门路与普宁路交叉西进口	0.41	东海西路与山东路交叉口东进口	0.40
8	东海西路与南京路交叉口北进口	香港中路与南京路交叉口北进口	0.45	东海西路与福州南路交叉口西进口	0.33
9	香港西路与延安三路交叉口东进口	香港中路与山东路交叉口东进口	0.50	香港西路与太平角六路交叉口东进口	0.29
10	香港西路与延安三路交叉口西进口	香港西路与太平角六路交叉口西进口	0.56	香港中路与山东路交叉口西进口	0.73
11	香港西路与延安三路交叉口南进口	东海西路与延安三路交叉口东进口	0.68	江西路与延安三路交叉口南进口	0.37
12	香港西路与延安三路交叉口北进口	江西路与延安三路交叉口北进口	0.32	香港中路与山东路交叉口西进口	0.52
13	重庆南路与人民路交叉口东进口	重庆南路与南京路交叉口北进口	0.46	鞍山路与人民路交叉口北进口	0.27
14	重庆南路与人民路交叉口西进口	温州路与宁化路交叉口西进口	0.28	人民路与嘉定路交叉口南进口	0.28
15	重庆南路与人民路交叉口南进口	鞍山路与人民路交叉口南进口	0.36	人民路与嘉定路交叉口南进口	0.45
16	重庆南路与人民路路口北进口	人民路与嘉定路路口北进口	0.47	鞍山路与人民路交叉口北进口	0.45

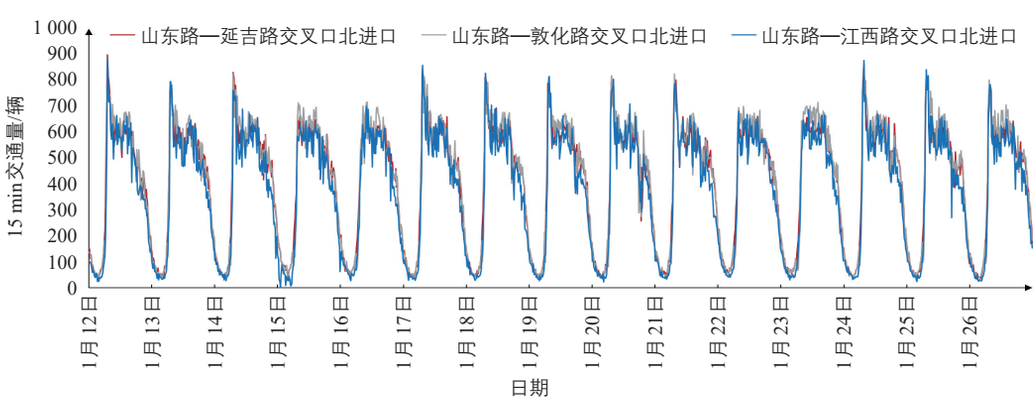


图4 山东路—延吉路交叉口北进口卡口和频繁上下游卡口15 min交通量时序变化

Fig.4 Time series of 15-minute traffic flow at the north entrance of the intersection of Shandong Road and Yanji Road, along with its frequent upstream and downstream traffic at checkpoints

上的相关性。机动车随着时间的变化在空间上发生位移，构成其轨迹序列，通过判断轨迹序列数据库中频繁出现的卡口对，即可识别和目标卡口关联度最高的卡口断面，进而将其流量作为目标卡口断面流量的影响因素。本研究利用Prefixspan算法提取每个卡口断面的频繁上下游卡口，由此建立起频繁上下游卡口对预测时段 $t+1$ 时刻目标卡口交通量来源的支撑关系。最终以频繁上下游卡口的实时流量作为特征变量输入预测模型，对目标卡口的短时交通量进行预测。

表3 卡口 $t+1$ 时刻交通量影响因素相关性分析

Tab.3 Correlation analysis of influencing factors on traffic flow at $t+1$ period

相关性	t 时刻该卡口过车流量	前一天 t 时刻该卡口交通量	一周前 t 时刻该卡口交通量	t 时刻频繁上游卡口交通量	t 时刻频繁下游卡口交通量
Pearson相关性	0.974	0.953	0.977	0.959	0.965
Sig.	<0.001	<0.001	<0.001	<0.001	<0.001

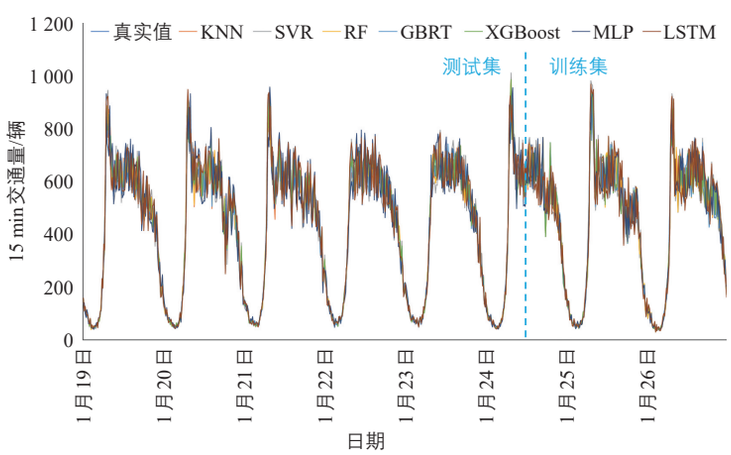


图5 不同模型预测值与真实值的对比

Fig.5 Comparison of predictions and observations by different models

表4 不同模型的预测性能

Tab.4 Prediction performance of different models

方法	训练集			测试集			耗时/s
	R^2	MAPE/%	RMSE	R^2	MAPE/%	RMSE	
KNN	0.976	7.51	35.51	0.966	8.55	40.50	3.62
SVR	0.965	10.92	42.83	0.953	12.53	47.26	5.43
RF	0.985	6.07	27.84	0.960	9.50	38.04	0.70
GBRT	0.989	6.20	28.61	0.964	9.71	41.91	0.78
XGBoost	0.989	6.19	29.53	0.967	9.47	45.70	2.92
MLP	0.962	11.77	44.05	0.964	9.22	42.10	15.70
LSTM	0.964	9.38	43.42	0.954	11.81	46.41	29.32

每个卡口断面频繁上下游卡口的提取过程如下：

1）将经过数据清洗后的全日过车记录数据集根据时间(DCOLLECTIONDATE)按照一定顺序(上游频繁卡口采用逆序，下游频繁卡口采用顺序)排列，将卡口名称(CCOLLECTIONADDRESS)以及进口方向(NDERICTRION)进行拼接，组成新的字段卡口断面(ADD&DIR)，并根据车牌号进行分组，将同一分组内的过车记录合并成一行作为该车牌号码的全日出行轨迹序列 $TR=\{tr_1, tr_2, \cdots, tr_n\}$，其中，$tr_i$ 为出行轨迹中依次经过的卡口断面(包含卡口名称及进口方向)；

2）对于每个卡口而言，序列数据库S是所有经过该卡口的机动车全日出行轨迹的集合，即S=<ID, TR>，支持度阈值min_support设为0.25，即当子序列 $tr=\{tr_a, tr_b, \cdots, tr_k\}$ 在S中的支持度大于0.25时，认为该子序列为频繁序列；

3）利用Prefixspan算法递归搜索序列数据库S的频繁序列；

4）在按照时间逆序排列的轨迹序列数据库S的所有频繁二项序列集中，若 $\{tr_n, tr_u\}$ 在所有 $\{tr_n, \_\}$[①]中支持度最大，则认为卡口 tr_u 是与 tr_n 的频繁上游卡口；在按照时间顺序排列的轨迹序列数据库S的所有频繁二项序列集中若 $\{tr_n, tr_l\}$ 在所有 $\{tr_n, \_\}$ 中支持度最大，则认为卡口 tr_l 是 tr_n 的频繁下游卡口。

基于Prefixspan算法挖掘卡口断面的频繁上下游卡口的部分结果如表2所示。以山东路—延吉路交叉口北进口卡口交通量数据为例，山东路—敦化路交叉口北进口和山东路—江西路交叉口北进口分别是其频繁上下游卡口，因此可以认为山东路—延吉路交叉口北进口交通量受上游山东路—敦化路交叉口北进口以及下游山东路—江西路交叉口北进口的影响显著。进一步分析卡口断面在研究周期内每15 min交通量数据的相关性(见图4)，可以看出这三个卡口在同一天内的流量虽然有所差异，但是总体趋势基本保持一致，峰值点也大致重合。

3.3 时空特征变量和目标变量Pearson相关性分析

Pearson相关性可以用来检验两连续变量之间的相关性，衡量这两个变量之间关联

的密切程度；相关系数在±1范围之内，其中，正数代表正相关，负数代表负相关；Sig.值可以用来判断相关性是否具有统计学意义。对于案例卡口，通过计算 $t+1$ 时刻卡口交通量和5个时空影响因素之间的Pearson相关系数来分析交通量与不同影响因素之间的相关性，结果如表3所示。

时间维度上，$t+1$ 时刻该卡口交通量和该卡口 t 时刻、前一天 t 时刻以及一周前 t 时刻交通量具有显著相关性(Sig.<0.001)。空间维度上，$t+1$ 时刻该卡口交通量和 t 时刻频繁上下游卡口交通量均具有显著的相关性(Sig.<0.001)。其他卡口也有类似的规律。因此，将表3中的5个变量作为卡口交通量的影响因素并用于预测模型的输入是可行的。

4 卡口交通量预测

4.1 模型构建及指标选取

对于第 i 个卡口断面 E_i 每天的过车记录数据，以15 min为间隔进行计数，则一天被划分为96个时段，得到其 n 天的15 min交通量时间序列 $Y^i=\{q_1^i,q_2^i,\cdots q_{96}^i,\cdots q_t^i,\cdots q_{96\times n}^i\}$。对于 $t+1$ 时刻的交通量，构造时空特征变量 $X^i=\{q_t^i,q_{one\_day\_t}^i,q_{one\_week\_t}^i,q_t^{i\_u},q_t^{i\_l}\}$，其中，$q_t^i$ 是卡口断面 E_i 的 t 时刻实时交通量，$q_{one\_day\_t}^i$ 是 E_i 在一天前 t 时刻的交通量，$q_{one\_week\_t}^i$ 是 E_i 在一周前 t 时刻的交通量，$q_t^{i\_u}$ 是 E_i 的频繁上游卡口 E_{i_u} 在 t 时刻的实时交通量，$q_t^{i\_l}$ 是 E_i 的频繁下游卡口 E_{i_l} 在 t 时刻的实时交通量。为了评估模型的预测性能，将数据集按照7:3划分为训练集和测试集。将上述特征变量 X^i 输入训练模型，选用典型的机器学习算法(支持向量回归SVR、K近邻KNN)、集成学习算法(随机森林RF、梯度提升回归树GBRT、极端梯度增强算法XGBoost)以及神经网络算法(多层感知机MLP和长短时记忆神经网络LSTM)构建7种模型进行预测。

为了评估不同预测模型的性能，选取平均绝对误差百分比(Mean Absolute Percentage Error, MAPE)、均方根误差(Root Mean Squared Error, RMSE)以及决定系数 R^2 作为衡量预测精度的指标，同时记录训练模型的训练时间作为衡量模型训练效率的指标。

$$MAPE=\frac{1}{m}\sum_{k=1}^{m}\left|\frac{f_k-y_k}{y_k}\right|\times 100\%,$$

$$RMSE=\sqrt{\frac{1}{m}\sum_{k=1}^{m}(f_k-y_k)^2},$$

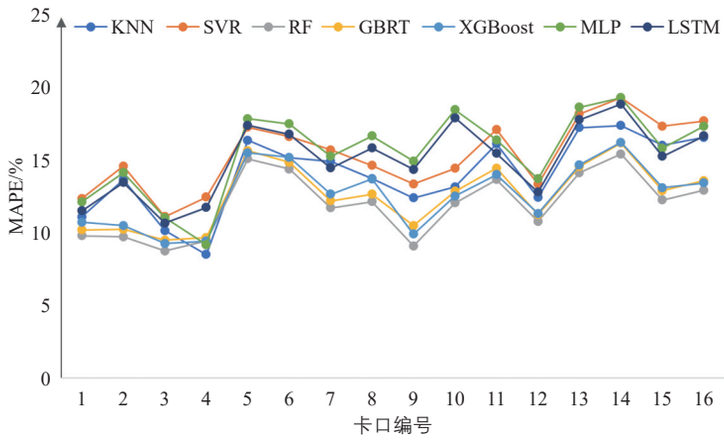


图6 不同卡口断面预测性能对比

Fig.6 Prediction performance comparison of different checkpoint sections

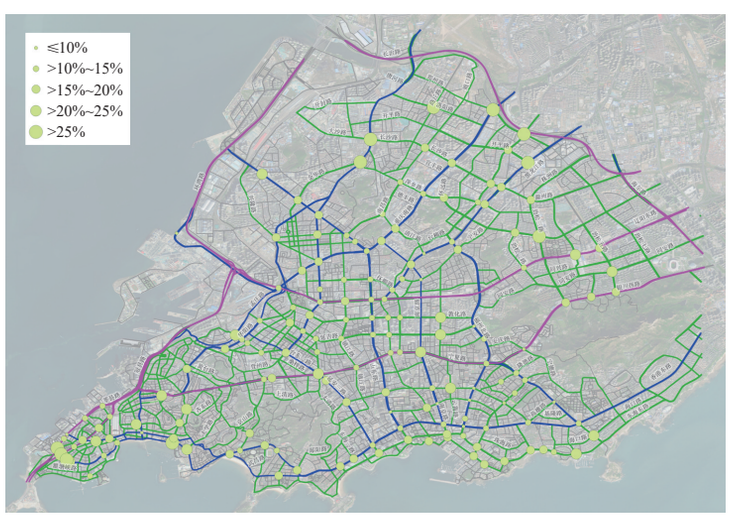


图7 KNN预测模型的MAPE

Fig.7 MAPE of KNN prediction model

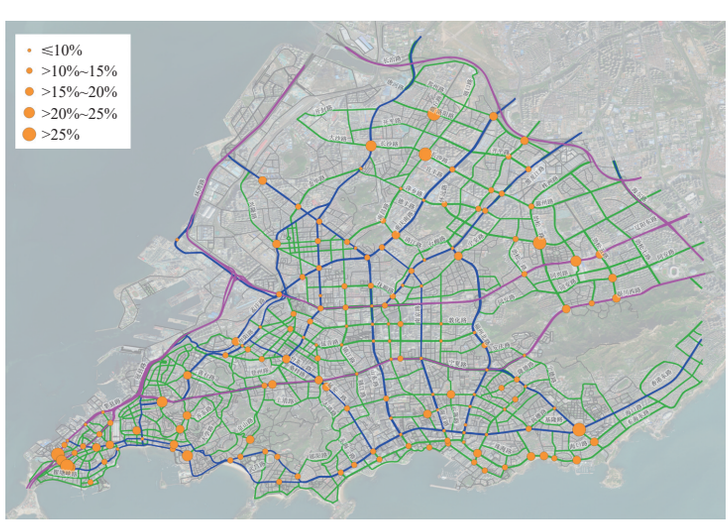


图8 RF预测模型的MAPE

Fig.8 MAPE of RF prediction model

$$R^2=1-\frac{\sum_{k=1}^{m}(f_k-y_k)^2}{\sum_{k=1}^{m}(\bar{y}-y_k)^2},$$

式中：f_k 为预测值；y_k 为真实值；$\bar{y}$ 表示真实值的均值；m 为样本数量/个。

4.2 卡口交通量预测结果

1）不同模型对单一卡口的交通量预测结果。

山东路—延吉路交叉口北进口卡口交通量预测结果如图5和表4所示，所有模型的测试集和训练集的决定系数 R^2 都在0.95以上，可见7种模型均具有良好的预测精度。其中，基于RF算法的预测模型性能最优，训练集和测试集的MAPE分别为6.07%和9.50%，训练集和测试集的RMSE最低，分别为27.84和38.04。其次是GBRT和XGBoost，稍逊于RF。XGBoost调参过程较为烦琐，耗时较多，比RF和GBRT多出2 s。KNN算法耗时较集成学习算法更长，虽然在训练集的预测精度不如GBRT和XGBoost，但是在测试集的预测精度下降较小，较为稳定。SVR和神经网络(MLP、LSTM)模型在预测误差上表现相近，但MLP和LSTM模型的耗时较长，训练迭代50次耗时分别达到15.7 s和29.32 s，交通量预测的实时性表现较差。本案例中RF和GBRT模型对卡口交通量预测的训练速度最快，分别是0.70 s和0.78 s，能够较好地满足短时交通量预测的实时性要求。综上所述，基于集成学习算法特别是随机森林RF算法的卡口交通量预测方法在准确性和实时性上表现最好。

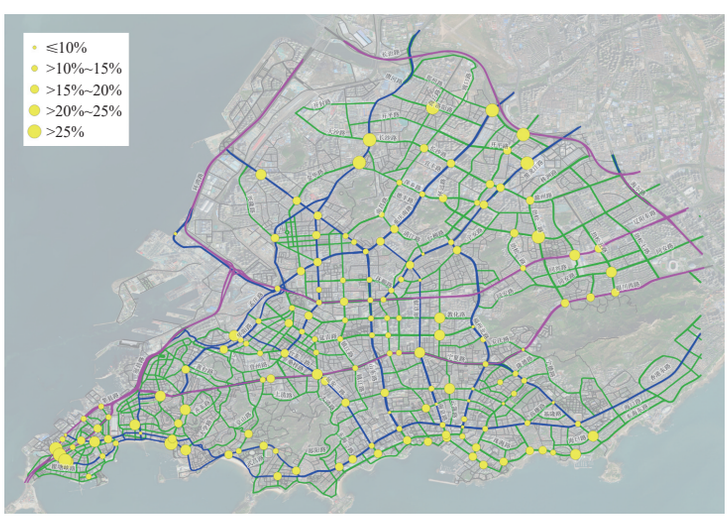


图9 GBRT预测模型的MAPE

Fig.9 MAPE of GBRT prediction model

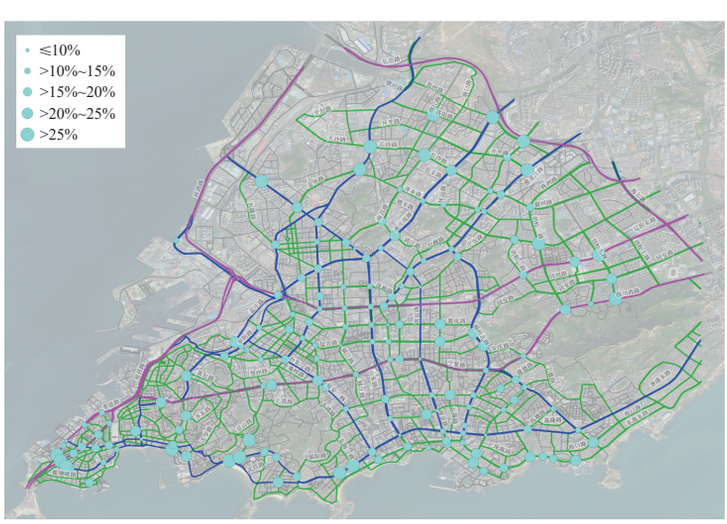


图10 LSTM预测模型的MAPE

Fig.10 MAPE of LSTM prediction model

2）不同模型对不同卡口的交通量预测结果。

由于不同卡口之间的实际交通量有所差异，为了使卡口之间的预测误差具有可比性，采用平均误差百分比MAPE进行对比。16个卡口断面预测性能对比如图6所示，在不同卡口断面，模型的预测精度具有较大差异，且7种模型的预测精度在不同卡口断面的变化具有相似性，特别是同种类型算法之间，这可能是受卡口断面实际交通量波动的影响。

图7~图10展示了不同模型对所有卡口交通量的预测误差在空间上的分布(为了便于可视化，同一交叉口各进口道卡口断面交通量MAPE值取均值，且由于模型预测误差规律相似，仅展示部分模型的MAPE空间分布结果)。可以发现，MAPE在所研究的路网中呈现“内高外低、密高疏低”的趋势。造成“内高外低”的原因可能是因为研究范围不是青岛市完整路网，外围边缘卡口(例如贵州路—台西三路交叉口、台柳路—郑州路交叉口、辽阳西路—劲松五路交叉口等)的频繁上下游卡口并不一定位于研究范围之内，对频繁路径的挖掘并不精准。“密高疏低”指卡口分布比较密集区域(例如山东路—延吉路交叉口、镇江北路—敦化路交叉口、福州南路—香港中路交叉口等)的预测精度较高，卡口分布较为稀疏的西南和东北区域(例如大学路—太平路交叉口、江苏路—莱芜一路交叉口、合肥路—劲松三路交叉口、四流南路—长沙路交叉口等)的预测精度稍低。同时，受电子交通设施分布不均的

影响，高清卡口无法将所有城市交叉口一一覆盖，因此卡口稀疏区域的频繁上下游卡口挖掘误差较大，导致该区域的卡口断面交通量预测精度降低。但是总体而言，随着城市交通设施的不断完善，基于本研究方法的预测精度将会进一步提升。

7种模型预测的MAPE的频数分布情况如图11所示，包含研究范围内163个交叉口。整体来看，RF效果最好，MAPE均值为13.47%；GBRT和XGBoost的MAPE均值为14%左右；SVR和LSTM的MAPE均值近似，分别为17.32%和17.77%，但LSTM耗时较长；MLP模型误差最大，MAPE均值为18.42%。虽然各种方法的误差均值都在20%以内，但比较而言，集成学习算法如RF，GBRT，XGBoost的预测精度较高且速度较快，神经网络算法(MLP和LSTM)的预测误差稍大且训练耗时较长。

3）卡口交通量和道路等级对预测结果的影响。

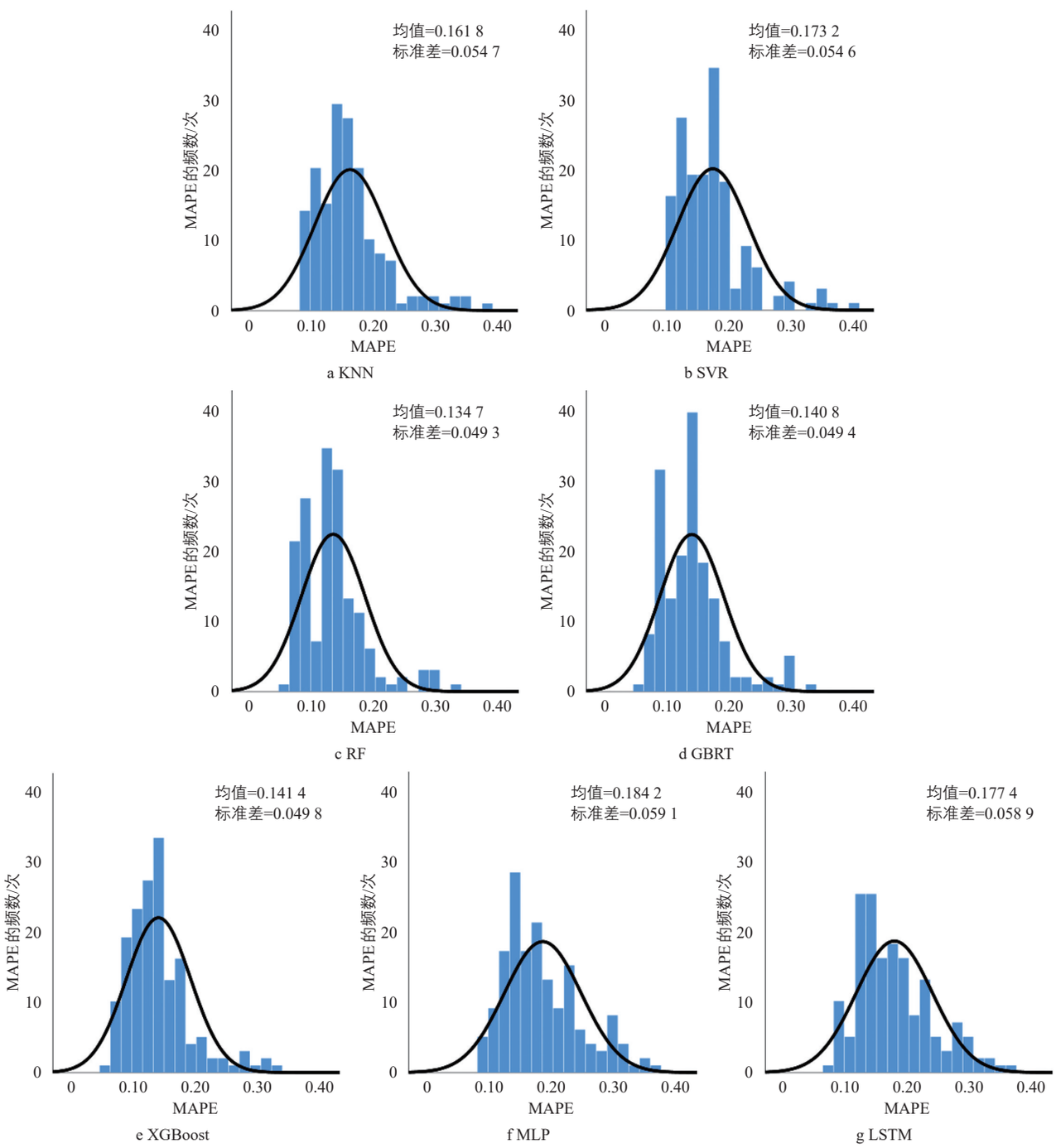


图11 7种模型预测的MAPE的频数分布情况

Fig.11 MAPE frequency distribution in the seven scenarios

除了将卡口的时空分布特征作为影响因素，本研究还分析了卡口15 min平均交通量以及连接道路类型对预测误差的影响。图12展示了各交叉口15 min平均交通量分布等级。本文利用方差分析比较不同交通量等级之间预测误差的差异性，结果如表5所示(以随机森林RF预测结果为例)，其中F值是F检验的统计量，Sig.值指显著性，Sig.<0.1认为组间差异性显著。可以发现，15 min交通量分级对于预测误差MAPE具有显著影响(F=5.217，Sig.=0.002<0.1)，随着交通量的增大，MAPE值降低，即高流量区域预测误差较小。同理，将卡口连接的道路类型分为主干路—主干路、主干路—次干路以及次干路—次干路三类(其余类型例如主干路-支路卡口样本量极少，忽略不计)。方差分析结果显示连接道路类型对MAPE有显著影响(F=2.388，Sig.=0.095<0.1)，连接较高等级道路的卡口交通量预测误差较小。而交通量较大的交叉口往往是拥堵治理关注的重点，较高的交通量预测精度有利于道路交通管理部门对未来的交通状态进行较为准确的预知和对路网进行优化管理，进而提高交通运行效率。

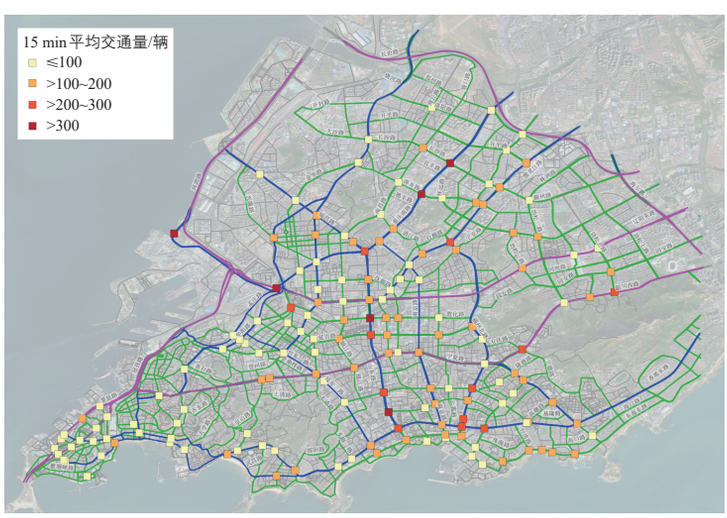


图12 交通量等级分布

Fig.12 Distribution of traffic flow grades

表5 不同因素对卡口交通量预测误差(MAPE)的影响

Tab.5 Influence of different factors on the traffic prediction error (MAPE) at checkpoints

因素	分类	均值	标准差	F	Sig.
15 min流量分级	0~<100	0.15	0.049	5.217	0.002
	100~<200	0.13	0.048		
	200~<300	0.10	0.033		
	≥300	0.09	0.028		
连接道路类型	主干路—主干路	0.10	0.045	2.388	0.095
	主干路—次干路	0.13	0.052		
	次干路—次干路	0.14	0.028		

5 结束语

频繁序列模式挖掘算法在卡口交通流特征挖掘及交通量预测中具有巨大潜能。本文首先基于卡口轨迹数据集对两类频繁序列模式挖掘算法——Apriori和Prefixspan的性能进行测试，验证了Prefixspan算法在对大规模轨迹数据集频繁路径挖掘中较Apriori算法更加高效，适合用于对卡口海量车辆轨迹序列的数据挖掘。为了预测卡口短时交通量，本文不仅考虑了交通量在时间上的相关性，发现卡口 $t+1$ 时刻的交通量与该卡口 t 时刻、一天前 t 时刻以及一周前 t 时刻的交通量高度相关，同时结合Prefixspan算法分析了卡口断面交通量在空间上的关联特征，挖掘出研究范围内各卡口断面的频繁上下游卡口，研究结果可以为城市道路交叉口疏堵提供思路。将与目标卡口交通量具有时空关联关系的5个时空影响因素作为特征变量输入预测模型，预测结果表明，集成学习算法特别是随机森林模型的整体表现更优，在保证预测误差最低的同时具有较高的运行效率，更适用于城市道路交通量的实时预测。除此之外，研究分析了卡口交通量预测精度的影响因素，发现在卡口分布较为密集、城市内部、交通量较大、连接道路等级较高等区域预测精度较高。这些结果表明了该预测方法的有效性，可以为交通管理部门提供决策管理支持以及出行诱导方面的数据支撑，并且随着未来电子卡口监控设备在城市空间分布的不断完善，预测精度将会进一步提高。

本研究尚存在一些不足，例如仅从卡口数据特征挖掘出发，没有考虑随机因素(天气状况、突发事件等)，未来有必要将这些因素纳入预测模型，以提高模型的普适性。此外，基于时空特征进行短时交通量预测的算法还有很多，例如时空注意力Transformer[21-22]、图卷积神经网络[23-24]等等，未来可以考虑和这些预测方法进行对比以及进行组合算法的研究。

注释：

Notes:

① $\{tr_n,\_\}$ 表示所有第一项为 tr_n 的频繁二项序列。

参考文献：

References:

[1] 贾若. 基于海量卡口过车数据的短时交通流预测[D]. 南京：东南大学，2018.
JIA R. Short- term traffic flow prediction based on massive bayonet data[D]. Nanjing: Southeastern University, 2018.

[2] 袁潜韬，邵晓波. 道路交通卡口车辆号牌识别准确率的分析与研究：以温州交警卡口现状为例[J]. 中国安防，2019(4)：86−90.

[3] 马飞虎，饶志强. 城市道路短时交通流预测方法研究[J]. 公路，2017，62(6)：192−196.
MA F H, RAO Z Q. Research on short term traffic flow forecast method of urban road[J]. Highway, 2017, 62(6): 192−196.

[4] WILLIAMS B M, HOEL L A. Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: theoretical basis and empirical results[J]. Journal of transportation engineering, 2003(6):129.

[5] MOURAUD A. Innovative time series forecasting: auto regressive moving average vs deep networks[J]. Journal of entrepreneurship and sustainability issues, 2017, 4(3): 282−293.

[6] 李翠，黄侃，李霞. 基于K近邻与主成分分析的短时交通流预测[J]. 公路交通技术，2022，38(3)：138−144.
LI C, HUANG K, LI X. Short- term traffic flow forecasting based on K-nearest neighbors and principle component analysis[J]. Technology of highway and transport, 2022, 38(3): 138−144.

[7] 李浩，张杉，曹斌，等. 基于城市道路卡口数据的交通流量预测[J]. 重庆大学学报，2020，43(11)：29−40.
LI H, ZHANG S, CAO B, et al. Prediction traffic flow based on traffic data of urban road check points[J]. Journal of Chongqing University, 2020, 43(11): 29−40.

[8] SUN Z Q, GEOFFREY F. Traffic flow forecasting based on combination of multidimensional scaling and SVM[J]. International journal of intelligent transportation systems research, 2014, 12: 20−25

[9] 焦朋朋，安玉，白紫秀，等. 基于XGBoost的短时交通流预测研究[J]. 重庆交通大学学报(自然科学版)，2022，41(8)：17−23.
JIAO P P, AN Y, BAI Z X. Short-term traffic flow forecasting based on XGBoost[J]. Journal of Chongqing Jiaotong University (natural science), 2022, 41(8): 17−23.

[10] 郑乐军，文成林. 基于集成学习的交通流短时特性分析与神经网络预测方法[J]. 科学技术与工程，2021，21(4)：1615−1623.
ZHENG L J, WEN C L. Analysis of short-term characteristics of traffic flow based on ensemble learning and neural network prediction method[J]. Science technology and engineering, 2021, 21(4): 1615−1623.

[11] 张耀方，陈坚. 基于GBDT算法的高速公路分车型交通流短时预测模型[J]. 公路，2022，67(1)：221−227.
ZHANG Y F, CHEN J. Short term prediction model of expressway traffic flow by vehicle type based on GBDT algorithm[J]. Highway, 2022, 67(1): 221−227.

[12] TIAN Y, ZHANG K, LI J, et al. LSTM-based traffic flow prediction with missing data[J]. Neurocomputing, 2018, 318(27): 297−305.

[13] 胥鑫. 基于卡口过车数据的城市区域交通流量预测模型研究[D]. 南京：南京师范大学，2021.
XUE X. Research on traffic flow prediction model of urban area based on traffic bayonet data[D]. Nanjing: Nanjing Normal University, 2021.

[14] 李巧茹，赵蓉，陈亮. 基于SVM与自适应时空数据融合的短时交通流量预测模型[J]. 北京工业大学学报，2015，41(4)：597−602.
LI Q R, ZHAO R, CHEN L. Short-term traffic flow forecasting model based on SVM and adaptive spatio-temporal data fusion[J]. Journal of Beijing University of Technology, 2015, 41(4): 597−602.

[15] 尹韬霖. 基于集成学习的交通流量预测方法研究[D]. 北京：北方工业大学，2020.
YIN T L. Research on traffic flow forecasting method based on ensemble learning[D]. Beijing: North China University of Technology, 2020.

[16] 陈玲娟，杨任泉. 基于路段关联度的城市

交通短时流量预测[J]. 武汉理工大学学报(交通科学与工程版)，2023，47(3)：402-407.
CHEN L J, YANG R Q. Short term urban traffic flow prediction based on links correlation[J]. Journal of Wuhan University of Technology (transportation science & engineering), 2023, 47(3): 402-407.

[17] 杨冰健. 基于交通卡口数据的机动车轨迹提取与关键路段挖掘分析[D]. 北京：北京交通大学，2020.
YANG B J. Extraction of vehicle trajectory and mining of key sections based on traffic bayonet data[D]. Beijing: Beijing Jiaotong University, 2020.

[18] GANAPATHY J, PARAMASIVAM J. Prediction of traffic volume by mining traffic sequences using travel time based PrefixSpan [J]. Intelligent transport systems, 2019, 13 (7): 1199-1210.

[19] 刘扬. 基于序列模式挖掘的轨迹预测算法研究[D]. 沈阳：辽宁大学，2020.
LIU Y. Research on trajectory prediction algorithm based on sequential pattern mining [D]. Shenyang: Liaoning University, 2020.

[20] JIAN P, HAN J, MORTAZAVI-ASL B, et al. PrefixSpan: mining sequential patterns efficiently by prefix- projected pattern growth [C]//IEEE Computer Society. International Conference on Data Engineering, 2001.

[21] 周楚昊，林培群. 基于多通道Transformer的交通量预测方法[J]. 计算机应用研究，2023，40(2)：435-439.
ZHOU C H, LIN P Q. Traffic flow prediction method based on multi- channel transformer[J]. Application research of computers, 2023, 40(2): 435-439.

[22] 张力. 基于注意力机制的短时交通流预测方法研究[D]. 湖州：湖州师范学院，2022.
ZHANG L. Research on attention mechanism based short-term traffic flow prediction approach[D]. Huzhou: Huzhou University, 2022.

[23] 张壮壮，屈立成，李翔，等. 基于时空卷积神经网络的数据缺失交通流预测[J]. 计算机工程与应用，2022，58(7)：259-265.
ZHANG Z Z, QU L C, LI X, et al. Traffic flow prediction with missing data based on spatial- temporal convolutional neural networks[J]. Computer engineering and applications, 2022, 58(7): 259-265.

[24] 聂欣. 基于时空数据的交通流预测研究[D]. 哈尔滨：黑龙江大学，2021.

(上接第41页)

[9] 滕丽，钟楚捷，蔡砥. 广州市地铁TOD站域的空间类型分异[J]. 经济地理，2022(4)：103-111.
TENG L, ZHONG C J, CAI D. Study on the spatial type differentiation of Guangzhou Metro TOD Zones: based on “node- place- linkage” coupling model[J]. Economic geography, 2022(4): 103-111.

[10] 郭少锋，芦晓昀，刘义钰. 从TOD到TOR：存量语境下轨道交通引领城市更新策略研究[J]. 规划师，2022(3)：76-81.
GUO S F, LU X Y, LIU Y Y. From TOD to TOR: transit oriented renewal in built-up area redevelopment[J]. Planners, 2022(3): 76-81.

[11] 中国城市规划设计研究院，苏州规划设计研究院股份有限公司. 苏州市综合交通运输发展战略研究(2035)[R]. 苏州：苏州市交通运输局，2019.

[12] 苏州规划设计研究院股份有限公司. 苏州历史文化名城保护规划(2013—2030)[R]. 苏州：苏州市规划局，2013.

[13] 苏州规划设计研究院股份有限公司. 苏州历史文化名城保护规划(2021—2035)[R]. 苏州：苏州市自然资源和规划局，2021.

[14] 住房城乡建设部关于印发城市轨道沿线地区规划设计导则的通知(建规函[2015]276号)[A/OL]. (2015-12-10)[2022-09-10]. https://www.mohurd.gov.cn/gongkai/zhengce/zhengcefilelib/201512/20151210_225899.html.

[15] 苏州日报. 轨交建设与TOD“齐步走”[EB/OL]. (2021-08-06)[2022-09-10]. https://www.suzhou.gov.cn/szsrmzf/szyw/202108/d8ec1b446141484d97272d8ef0ed51ab.shtml.