文章编号：0253-374X(2023)03-0405-11

DOI：10.11908/j.issn.0253-374x.22011

基于集成学习的信号控制交叉口排队长度估计

吴　浩[1]，刘　磊[2]，唐克双[1]

(1. 同济大学 道路与交通工程教育部重点实验室，上海 201804；2. 中共兴平市委组织部，陕西 兴平 713100)

摘要：基于电子警察(LPR)数据和网联车辆轨迹数据，提出了一种基于集成学习的信号控制交叉口排队长度估计方法。通过分析不同数据条件下估计方法的适用条件和精度水平，运用随机森林方法设计集成学习器，并构建电子警察和网联车辆轨迹感知信息及不同方法估计结果和真实排队长度之间的非线性映射关系。仿真结果表明：本方法的平均绝对误差为1.3 m•周期$^{-1}$•车道$^{-1}$，平均绝对百分比误差为1.4%。

关键词：信号控制交叉口；排队长度；电子警察(LPR)数据；网联车辆轨迹数据；集成学习；随机森林

中图分类号：U491.4　　**文献标志码**：A

Queue Length Estimation at Signalized Intersection Based on Ensemble Learning

WU Hao[1], *LIU Lei*[2], *TANG Keshuang*[1]

(1. Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University, Shanghai 201804, China; 2. Organization Department of CPC Xingping Municipal Committee, Xingping 713100, China)

Abstract: Based on license plate recognition (LPR) data and connected vehicle trajectory data, an ensemble learning method was deployed to estimate the intersection queue length. By analyzing the applicability and accuracy of different queue length estimation methods, the random forest method was applied to design a base ensemble learner and formulize the nonlinear mapping relationship among the LPR data, connected vehicle trajectory data, estimation results of the existing queue length methods and real queue length values. Simulation results show that the proposed method overperforms the existing queue length methods, with a mean absolute error of 1.3 m•cycle^{-1}•lane^{-1} and a mean absolute percent error of 1.4%.

Key words: signalized intersections; queue length; license plate recognition (LPR) data; connected vehicle trajectory data; ensemble learning; random forest

作为城市道路网络中的重要节点，信号控制交叉口是交通拥堵的常发地点。排队长度能够形象、直观地反映信号控制交叉口的拥挤程度，是交叉口运行效率评价的重要指标之一，也是交叉口信号控制优化的重要参数[1]。精确且有效地估计交叉口排队长度，一方面可以评价当前交叉口各流向排队状态，评估信号控制方案的合理性；另一方面，可以根据当前排队长度参数，制定相应的信号控制策略，优化交叉口信号配时参数。

根据检测器数据的不同，现有的信号控制交叉口排队长度估计方法可分为基于定点检测器(线圈、地磁等)数据的方法和基于移动检测器(浮动车等)数据的方法。基于定点检测器数据的方法又可分为输入输出模型[2-6]与交通波模型2类[7-12]。前者通过分析车辆到达率与消散率之间的关系，得到累计到达车辆数与累计驶离车辆数的差值，从而对排队长度进行估计；后者则认为在车辆排队形成与消散过程中，会形成向上游传递的交通波(集结波与消散波)，2个交通波相交位置即为最大排队长度所在位置。基于移动检测器数据的方法可分为基于交通波理论的确定性模型[13-18]与基于概率论理论的非确定性模型[19-23]2类。前者基于轨迹数据分析车辆停止与启动关键点，并结合线性回归等模型对交通波进行拟合，从而对排队长度进行估计；后者则认为车辆到达是一个随机过程，然后利用概率统计方法推导车辆排队长度分布，进而计算排队长度期望值。

近年，由于平安城市建设和交通执法管理的需

收稿日期：2022-01-07
基金项目：国家自然科学基金(61673302)
第一作者：吴　浩(1996—)，男，博士生，主要研究方向为数据驱动的信号控制评估与优化。
E-mail：mas@tongji.edu.cn
通信作者：唐克双(1980—)，男，教授，博士生导师，工学博士，主要研究方向为交通信号控制、智能交通。
E-mail：tang@tongji.edu.cn

论文拓展介绍

要,电子警察设备在我国城市道路广泛布设,电子警察数据可提供所有车辆通过交叉口停车线的时刻、所在车道及车辆ID等断面全样观测信息,相比于线圈等传统定点检测器,其数据质量更高、覆盖范围更广。另一方面,随着智能网联汽车、移动导航等新技术的推广和应用,海量网联车辆轨迹数据的实时获取成为可能,网联车辆轨迹数据可提供抽样车辆的个体连续观测信息。上述2类数据在交通流观测上形成了时间和空间的互补,也为交叉口信号控制评价和优化提供了新机遇。因此,国内外学者基于电子警察和网联车辆轨迹数据提出了很多信号控制交叉口排队长度估计方法,包括基于电子警察数据的方法[5-6,24]、基于网联车辆轨迹数据的方法[14,22-23]和基于电子警察和网联车辆轨迹数据融合的方法[25]。然而,上述方法只适用于一定的交通状态(饱和度等)或数据条件(网联车渗透率),适用范围和可移植性存在局限。例如,基于电子警察数据的方法大多不适用于过饱和条件,而基于网联车辆轨迹数据的方法通常要求较高的渗透率。

提出了一种基于集成学习的信号控制交叉口排队长度估计方法,有效集成基于电子警察和网联车辆轨迹数据的排队长度估计方法的优点,可适应多样的交通状态和数据条件,从而实现更加可靠且准确的排队长度估计。

1 研究综述

根据数据源的不同,现有信号控制交叉口排队长度估计方法可分为基于固定式检测设备数据[2-12]和基于移动检测器数据[13-23]的方法。前者可结合车辆离散等假设对排队长度进行较为合理的估计,或基于交通波模型直观地反映车辆的排队以及消散过程,但该类模型多基于上下游双截面的定点检测场景,而且精度受限于传统定点检测器质量。随着电子警察数据的可获取,部分学者尝试基于电子警察数据估计排队长度[24,26-27],受限于电子警察设备检测机理,现有估计方法仅适用于未饱和场景,并且适用场景受到限制。后者可结合轨迹数据与线性回归等模型对交通波进行拟合,或认为车辆到达是一个随机过程,然后利用概率统计方法推导车辆排队长度分布,进而计算排队长度期望值。然而,该类模型多需要假定到达类型、浮动车轨迹采样率等部分参数已知,在轨迹数据采样率较低的条件下,模型估计精度较低,鲁棒性较差。

由于数据检测过程中的误差及干扰,因此基于单数据源的排队长度估计方法的可靠性、稳定性难以保证。随着信息技术的革新,数据融合技术与数据挖掘技术的不断发展,部分学者尝试融合多源检测数据。Badillo等[28]基于浮动车数据与采样间隔为20 s的上游路段定点检测数据,通过分析浮动车的启停关键点,结合交通波理论构建交叉口排队长度估计模型。Cai等[29]基于采样间隔为1 s的上游定点检测器数据与浮动车数据,根据浮动车启停位置、固定点检测器横断面、初始排队长度和最大排队长度位置关系的不同,分4种情况进行讨论,并基于交通波理论确定排队形成和消散过程中的关键点数据,从而建立4种情况下初始排队长度和最大排队长度指标的估计模型。Bhaskar等[30]基于线圈检测器数据与蓝牙数据,利用车辆累计到达-驶离曲线计算交叉口排队长度。吴翱翔[31]通过融合浮动车数据、射频识别数据以及视频数据建立交叉口排队长度估计模型,该方法首先将射频识别数据和视频检测数据所提供的行程时间信息进行特征级融合,然后对浮动车数据进行时间和空间匹配,并与前2种数据源进行决策级融合,最后基于行程时间和交叉口交通状态求得交叉口排队长度。总而言之,数据融合能发挥不同数据源的自身优势,得到更好的排队长度估计结果,但现有模型多基于传统定点检测数据,估计精度仍受限于传统定点检测数据质量,而基于射频识别、蓝牙等新型数据的估计模型虽获得了较高的估计精度,但其设备在我国大多数中小城市中布点有限,很难在工程实践中得到大范围应用。

随着电子警察设备在我国广泛布设,部分学者尝试融合电子警察数据与其他类型数据。陶晶晶[32]基于路段定点检测数据与电子警察数据,利用反向传播(BP)神经网络模型建立上下游交叉口流量的映射关系,并基于交通波理论分流向重构车辆轨迹,进一步求取交叉口排队长度。Qom等[33]基于上下游检测器数据,以5 min为时间粒度对数据信息进行集计,同时结合上下游电子警察设备计算路段平均行程速度,基于内插法构建排队长度估计模型。在此基础上,李爱杰[34]融合路段低频定点检测数据(60 s间隔)与电子警察数据,基于交通波理论和概率论方法构建交叉口排队长度估计模型。Tan等[25]基于车辆轨迹数据与电子警察数据对周期级排队长度进行估计,在该方法中,基于贝叶斯理论分别计算未饱和、过饱和条件下车道排队长度的极大似然估计值,进一步估计车道级排队长度。此类方法考虑了单截

面电子警察数据源条件下可能存在的无法计算剩余排队长度等缺点，通过与其他数据融合，可发挥不同数据源的优势，但模型估计的精度仍受限于传统定点检测设备的质量与移动检测数据的采样频率。

综上所述，基于电子警察数据的排队长度估计已有初步研究，但受限于其检测机理，单截面数据驱动的方法主要适用于未饱和场景，并且交通波的完整重构仍需多截面数据；而基于网联车辆轨迹数据的方法多需已知到达类型分布、浮动车采样率等参数，并且主要适用于轨迹数据渗透率高的场景。2种数据源在交通状态和数据条件上的适应性具有明显的互补性，虽已有融合了电子警察数据与网联车辆轨迹数据的排队长度估计方法，但仍主要适用于轨迹数据采样率较高的路口。集成学习可通过构建并结合多个机器学习器做最后的决策，发挥各类方法的优势互补性，被广泛应用于交通状态评估领域[35-36]，在处理本问题上具有明显的优势。因此，基于集成学习方法构建信号控制交叉口排队长度估计方法具有重要的意义。

2　3个基本模型

2.1　基于电子警察数据的排队长度估计模型

李爱杰[34]首次提出基于变点分析(CPA)的排队长度估计模型，研究场景如图1所示。在该方法中，基于相邻同属性车道排队长度趋于一致的假设，将电子警察数据所得到的车辆驶离车头时距序列进行切分。考虑到排队车队、非排队车队在绿灯期间消散流率的不同，选取平均值、方差指标构建特征函数，用来描述切分所得2个车头时距子序列的数值特征差异性，并通过判断特征函数最大值所处位置得到周期内车道级排队长度值。

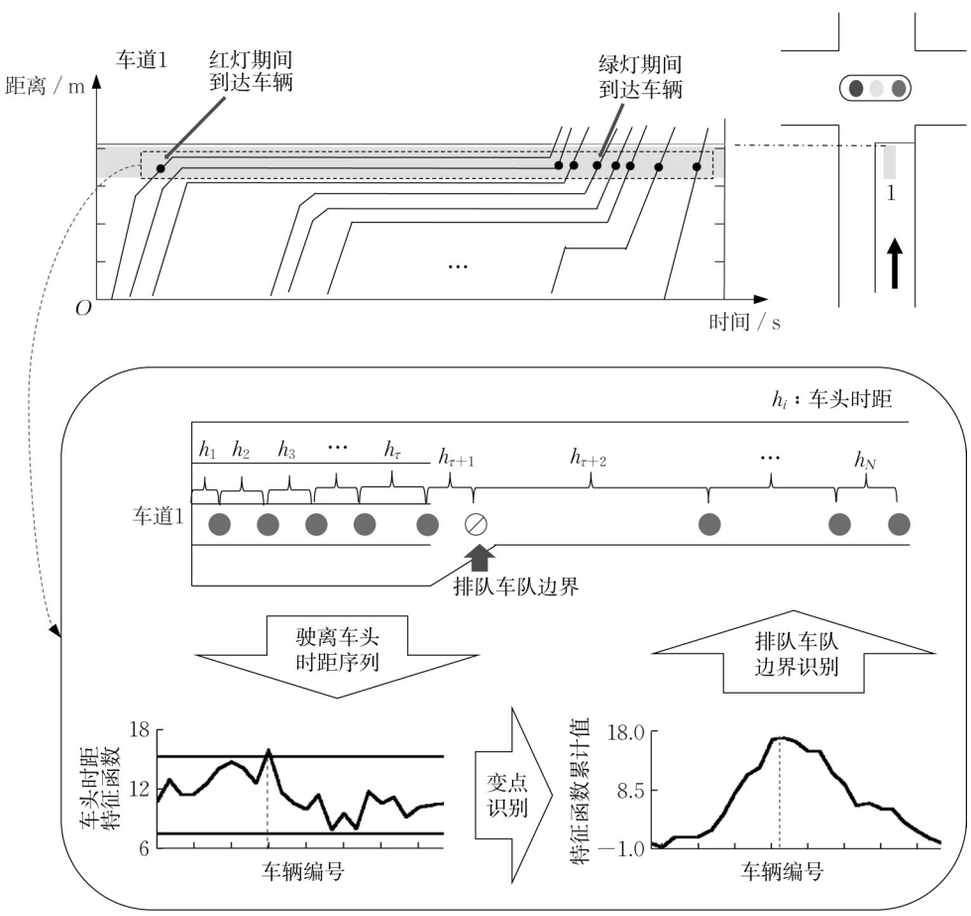


图1　变点分析模型研究场景

Fig.1　Research scenario of CPA model

基于连云港市朝阳路-通灌北路交叉口所采集的真实电子警察数据对该方法进行实证验证。结果表明，该方法的估计精度为80.4%。然而，该方法基于如下假设：排队车辆与非排队车辆的驶离车头时距均值与方差存在较大波动；最后一辆排队车辆与第一辆非排队车辆的车头时距明显增大，当周期内所捕获的电子

警察数据满足任一条件时才能进行有效估计。因此，当饱和度从0.4增加至0.8时，排队车辆与非排队车辆特征差异变得不明显，该方法的精度显著降低，平均绝对误差增加了1.0辆•周期$^{-1}$[24]。

2.2 基于网联车辆轨迹数据的排队长度估计模型

Li等[16]提出一种基于网联车辆轨迹数据的周期级排队长度估计模型，研究场景如图2所示。首先，将车辆状态分为运动状态、停车状态以及爬行状态3类，并基于带有约束条件的最小二乘拟合方法判断车辆的关键启停点，即车辆加入排队、离开排队的时间及位置；然后，基于交通波理论与分段线性拟合方法求取排队车辆集结波与信号配时参数，进一步得到周期最大排队长度及剩余排队长度。

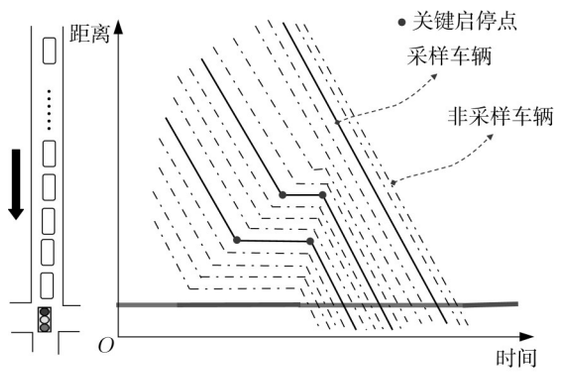


图2 交通波模型研究场景

Fig.2 Research scenario of shockwave-based model

基于微观仿真软件Vissim对该方法进行了仿真验证及敏感性分析。结果表明，在轨迹渗透率为20%的情况下，上传时间间隔为5、15、25 s时平均绝对误差分别为1.8、2.6、3.2辆。由于该方法需基于车辆轨迹点识别车辆运行状态及关键启停点，其估计精度取决于轨迹数据的渗透率，在实际情况下轨迹数据的渗透率多低于10%[22]，相应的估计误差在3辆车以上，因此当轨迹数据渗透率较低时(0%～20%)，该方法的平均绝对误差为3.2～3.6辆，适用性显著降低。

2.3 基于电子警察和网联车辆轨迹数据融合的排队长度估计模型

Tan等[25]提出了一种基于电子警察数据与网联车辆轨迹数据融合的交叉口排队长度估计模型。该模型在数据级层面融合了电子警察数据与轨迹数据，研究场景如图3所示。首先，分析了历史电子警察数据并拟合了排队车辆与非排队车辆消散车头时距的概率分布；然后，基于贝叶斯方法提出了一种最大后验概率的排队长度估计方法，通过求取排队车辆与非排队车辆的最大概率分割点得到了最大排队长度；最后，通过采样轨迹车辆的排队位置信息对所计算的排队长度进行修正，提高了排队长度估计精度。

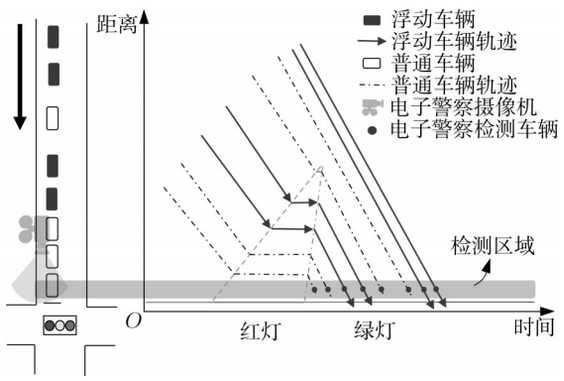


图3 贝叶斯模型研究场景

Fig.3 Research scenario of Bayesian-based model

基于深圳市福中路-皇岗路交叉口真实数据对该方法进行了实证验证与仿真验证。结果表明，该模型的平均绝对误差为3.1辆•周期$^{-1}$，平均绝对百分比误差为14.0%。由于该方法需要通过采样轨迹车辆的排队位置信息对所计算的排队长度进行修正，估计精度仍取决于轨迹数据渗透率；此外，仿真验证结果表明，在饱和度接近1、渗透率从50%降至5%时，估计误差从1.0辆•周期$^{-1}$增加至2.5辆•周期$^{-1}$。在实际情况下轨迹数据的渗透率多低于10%[22]，因此该方法仍主要适用于轨迹数据采样率较高的路口。

3 基于集成学习的交叉口排队长度估计

3.1 建模思路

由前期相关研究成果可知，基于电子警察数据的变点识别方法适用于排队车队与非排队车队消散过程具有明显差异的场景[34]，而基于网联车辆轨迹数据的交通波方法适用于轨迹数据渗透率高的场景[16]。现有的在数据级层面融合了电子警察数据与网联车辆轨迹数据的贝叶斯模型，相较于单数据源可提升排队长度估计的精度，但仍主要适用于轨迹数据采样率较高的路口[25]。真实数据条件下，交通量存在短时波动性，排队车队与非排队车队的消散特征取决于上游车辆到达模式，同时轨迹数据采样率等参数存在随机波动，因此上述3个模型的估计精度会受到影响。虽然3种方法对于不同的交通状

态和数据条件覆盖度有着良好的互补性，但是各种方法的误差与交通状态密切相关，而且此类关系往往是复杂非线性的。集成学习方法广泛应用于分类问题集成、回归问题集成、特征选取集成等问题，可通过构建并结合多个机器学习器做最后的决策，具有模型结构简单、参数较少、能处理特征变量多重共线问题、泛化能力较强的特点，近年来被广泛应用于交通状态评估领域[35-36]，在处理本研究问题上具有很大优势。因此，本研究拟基于集成学习方法，以电子警察和网联车辆轨迹感知信息以及3种方法估计结果为输入特征，构建与真实排队长度之间的非线性映射关系，使得各种方法在时间、空间维度上进行互补，实现排队长度的精准估计。

根据基学习器间关系的不同，集成学习可分为串行集成方法与并行集成方法。前者假设各基学习器间存在较强的依赖性，各基学习器的学习模型仅能按顺序生成，训练耗时长，常见方法为AdaBoost算法、梯度提升树[37]等；后者则认为各基学习器间不存在强依赖关系，可通过并行训练节省训练所需时间，常见方法为Boostrap抽样、随机森林[37]。其中，随机森林是一种随机选取训练集解释变量的子集进行训练，从而获得一系列决策树的集合的方法，因其泛化能力强、对数据集适应能力强、训练高度并行化、可处理高维数据的特点，被广泛应用于交通领域的分类（回归）集成问题[38-39]。考虑到排队长度指标实时计算所需速度及交通采集数据的高维性，本研究拟选取并行集成方法中的随机森林构建集成学习模型。

3.2　模型构建

3.2.1　训练数据准备

在基于随机森林的集成学习模型中，通过不同的饱和度、轨迹渗透率组合生成训练数据，并且一个周期即是一个样本。此外，原始输入数据主要分为数据层变量与决策层变量2类，前者可直接从2类数据源中获取信息，即电子警察数据所包含的周期车流量、轨迹数据所包含的周期内排队轨迹数与周期内非排队轨迹数，而后者包含上述3种方法的排队长度估计方法的结果。训练数据集格式定义如下所示：

$$D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_N,y_N)\} \tag{1}$$

式中：$\boldsymbol{x}_i=(v_i,m_i,n_i,r_{1i},r_{2i},r_{3i})^{\mathrm{T}}$为输入特征向量；$y_i$为标签变量，即第$i$个周期的真实排队长度（$i=1,2,\cdots,N$）。在输入特征向量中，$v_i$表示当前第$i$个周期内电子警察设备所采集到的交通流量，$m_i$、$n_i$分别表示当前第$i$个周期内采样轨迹数据中排队车辆轨迹与非排队车辆轨迹数量，r_{i1}、r_{i2}、r_{i3}分别表示基于当前第i个周期所采集的基于电子警察数据的变点识别方法、基于网联车辆轨迹数据的交通波方法以及基于数据融合的贝叶斯方法的估计结果。

3.2.2　基学习器训练

基学习器通常基于现有学习算法从训练数据中产生，如决策树算法、BP神经网络算法、支持向量机等。其中，随机森林是一类主要以决策树模型为基学习器的模型。因此，以CART（classification and regression tree）决策树算法[40]为例，对基学习器的训练过程进行说明。主要流程如下所示：

（1）对任意的划分特征$\boldsymbol{x}_i$，对应的任意划分点s，在基于原始数据集D得到训练集R后，可进一步把训练集R划分成2个子集R_1与R_2。通过遍历特征向量中的变量j，求取使得2个子集R_1与R_2以及2个子集之和的均方差最小的对应划分点s和最优切分变量j，计算式如下所示：

$$\min_{j,s}\left(\min_{\hat{c}_1}\sum_{\boldsymbol{x}_i\in R_1(j,s)}(y_i-\hat{c}_1)^2+\min_{\hat{c}_2}\sum_{\boldsymbol{x}_i\in R_2(j,s)}(y_i-\hat{c}_2)^2\right) \tag{2}$$

式中：$\hat{c}_m$（$m=1,2$）表示第m个子区域内真实排队长度y的均值。

（2）用选定的数值对(j,s)划分区域并决定相应的输出值，计算式如下所示：

$$R_1(j,s)=\{\boldsymbol{x}|\boldsymbol{x}_j\leqslant s\},R_2(j,s)=\{\boldsymbol{x}|\boldsymbol{x}_j>s\} \tag{3}$$

$$\hat{c}_m=\frac{1}{N_m}\sum_{\boldsymbol{x}_i\in R_m(j,s)}y_i,\boldsymbol{x}_i\in R_m,m=1,2 \tag{4}$$

式中：N_m表示第m个区域中训练集的记录数。

（3）继续对2个区域R_1与R_2分别调用第（1）步与第（2）步进行训练集划分，直至满足停止条件，即达到决策树的深度限制或叶子节点的个数限制，此时共可得到M个区域。

（4）将输入空间划分为M个区域$(R_1,R_2,\cdots,R_M)$，可生成决策树，如下所示：

$$f(\boldsymbol{x})=\sum_{m=1}^{M}\hat{c}_mI(\boldsymbol{x}\in R_m) \tag{5}$$

式中：$f(\boldsymbol{x})$表示基学习器对应的输出结果；M表示最终生成的区域个数；$I(\boldsymbol{x}\in R_m)$表示判定$\boldsymbol{x}$所属区域的函数，当$\boldsymbol{x}\in R_m$时$I(\boldsymbol{x}\in R_m)=1$，反之$I(\boldsymbol{x}\in R_m)=0$。

以上便是基于CART决策树算法的基学习器训练过程。基于上述步骤，给定一组特征数据$\boldsymbol{x}_i=$

$(v_i, m_i, n_i, r_{1i}, r_{2i}, r_{3i})^{\mathrm{T}}$，可基于该基学习器得到对应的输出结果$f(\boldsymbol{x})$。

3.2.3 基学习器集成

基于上述基学习器训练方法可获取各基学习器的输出结果，在此基础上需进一步确定结合策略以求得系统集成的输出。根据实现方法的不同，集成学习的结合策略可分为投票法、平均法与学习法。平均法相较于其余方法，适用于大规模集成，并且无需与其他学习器结合[37]。因此，本研究采用加权平均集成策略，对各基学习器赋予加权系数，并根据加权系数和基学习器结果进行结合以获得集成系统的输出。

假设该集成学习器共包含T个基学习器$\{f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \cdots, f_T(\boldsymbol{x})\}$，并且每个基学习器在样本空间$\boldsymbol{x}$上的排队长度估计结果为$f_i(\boldsymbol{x})$，最终的排队长度$H(\boldsymbol{x})$计算式如下所示：

$$H(\boldsymbol{x}) = w_i \sum_{i=1}^{T} f_i(\boldsymbol{x}) \tag{6}$$

式中：w_i为基学习器$f_i(\boldsymbol{x})$所对应的权重系数，通常要求$w_i \geqslant 0$且$\sum_{i=1}^{T} w_i = 1$。

根据各基学习器的计算结果，即CART决策树结果进行基学习器集成，可在给定输入数据$\boldsymbol{x}$条件下，得到排队长度输出结果$H(\boldsymbol{x})$。在此基础上，训练集数据$\boldsymbol{x}_i$与排队长度输出结果$H(\boldsymbol{x})$可作为集成学习的模型输入，进一步估计排队长度。

3.2.4 排队长度估计

在将得到的训练集数据$\boldsymbol{x}_i$与排队长度输出结果$H(\boldsymbol{x})$作为模型输入的基础上，本研究基于随机森林方法进行集成学习，主要分为数据准备、模型训练以及模型评估3个阶段，建模流程如图4所示，主要步骤解释如下：

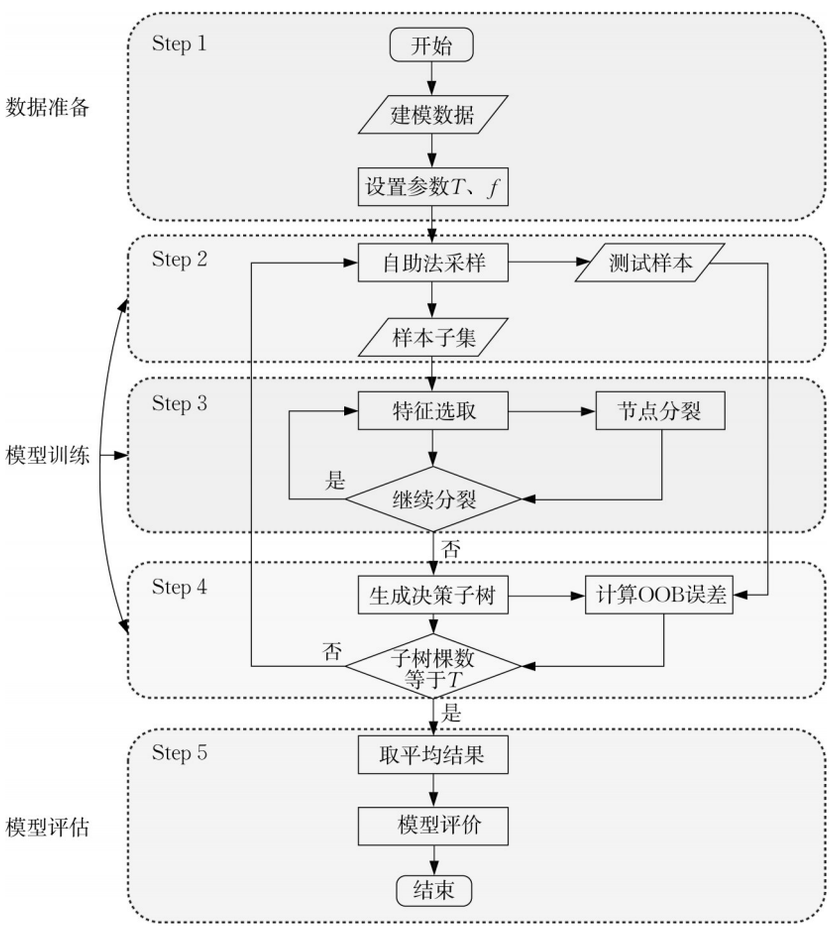


图4 基于随机森林的交叉口排队长度估计建模流程

Fig.4 Framework of queue length estimation based on random forest method

Step 1 设置随机森林算法的主要参数，即决策子树棵数T。

Step 2 假设$D=\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_N, y_N)\}$代表原始数据样本，基于抽样方法生成训练样本子集$B=\{B_1, B_2, \cdots, B_t\}$。在此过程中，考虑到轨迹数据渗透率等参数的随机波动性，其样本容量可

能在一定范围内波动，而自助法重采样技术适用于任一样本容量场景，可基于有限的样本资料多次重复抽样，并重新建立起足以代表母体样本分布的新样本，在解决本研究问题时具有更大优势[41]。因此，采用自助法重采样技术建立决策子树 $A=\{f(\boldsymbol{x},\theta_t),t=1,2,\cdots,T\}$。

Step 3　假设 K 代表总特征空间，基于随机子空间思想，从 K 中随机抽取 k 个特征（$k<K$），并基于CART决策树算法进行节点分裂。给定一组特征数据 $\boldsymbol{x}_i=(v_i,m_i,n_i,r_{1i},r_{2i},r_{3i})^{\mathrm{T}}$，可基于该基学习器得到对应的输出结果 $f(\boldsymbol{x})$。

Step 4　重复Step 2与Step 3，构建 T 棵决策子树。在此过程中，对于每棵决策子树不进行剪枝，任其自由生长，形成随机森林。

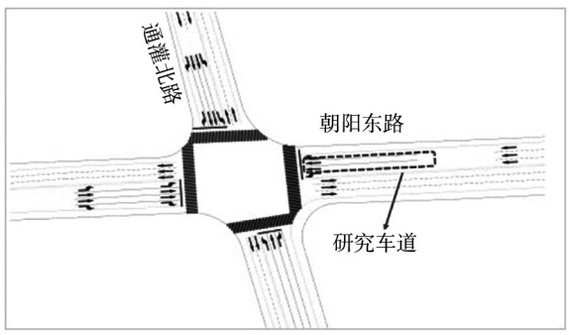
Step 5　基于所构建的 T 棵决策子树，对未知样本做出决策，即计算未知样本对应下的排队长度估计结果，并取平均值作为随机森林输出结果。

在基于自助法重采样技术生成训练样本的过程中，每次约有1/3的样本不会出现在所采集的样本集合中，此类数据称为袋外（OOB）数据，即OOB样本。OOB样本未参与决策树的建立，可用于所构建的随机森林模型的效果评估。首先计算每个样本作为OOB样本时决策树的分类情况，然后通过基学习器集成步骤求取该样本的分类结果，并用分类错误个数占样本总数的比率作为随机森林的OOB误分率，也称袋外错误率。袋外错误率越低，说明测试集上表现好，模型的泛化能力更强。

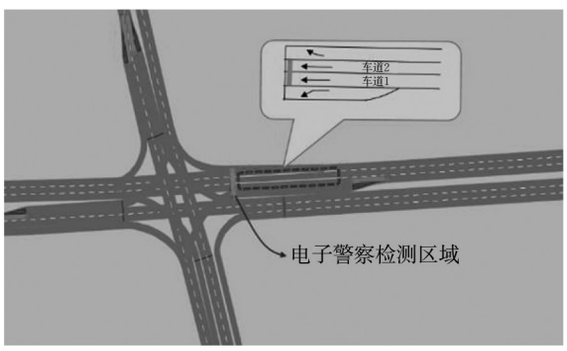
4　模型评价

4.1　验证场景

以连云港市朝阳东路-通灌北路交叉口为研究对象，基于微观仿真软件Vissim建立仿真模型，并将2条东进口直行车道作为目标车道，如图5a所示。该交叉口使用定时信号控制方案，周期长度为130 s，东直行流向绿灯时长为60 s。通过在交叉口停车线上游2 m位置设置数据采集点，可模拟真实电子警察数据，精确记录通过交叉口车辆的车辆类型、通过时刻以及车道等信息，如图5b所示。Vissim仿真软件可基于用户设置的时间间隔，精确记录车辆的全样运行轨迹数据，本研究基于3 s的采样间隔采集车辆轨迹数据，以还原车辆在路段上的实际运动状态；通过随机抽样的方法模拟网联车辆渗透率，从而得到所需验证场景。



a 交叉口渠化图



b Vissim仿真模型

图5　仿真场景

Fig.5　Simulation scenario

仿真模型的时长设置为9 000 s，仿真运行的前600 s为预热时间，不作为验证数据，因此可获取65个有效周期的数据。考虑车辆到达随机性对实验结果的影响，仿真中设置不同的随机种子多次运行，本实验中设置10个随机种子。此外，为验证饱和度、轨迹渗透率的影响，共设置11个饱和度场景（0.4～0.9，间隔0.05）、15个采样率范围（5%～20%，间隔2%；20%～50%，间隔5%）。每个场景包括65个周期，共计10 725（=65×11×15）组仿真数据。

4.2　特征选择与参数优化

由仿真模型所采集到的模拟电子警察数据与车辆轨迹数据得到原始特征向量，共包含了3个数据层变量（周期内车流量、周期内排队轨迹数、周期内非排队轨迹数）与3个决策层变量（基于变点识别方法的排队长度估计结果、基于交通波的排队长度估计结果、基于贝叶斯方法的排队长度估计结果）。为评估不同特征的重要性程度，首先基于随机森林模型对不同的特征进行选择测试，特征重要性结果如图6所示。由图6可知，整体而言，决策层的重要性程度明显大于数据层。对于决策层变量，贝叶斯估计结果的重要性最大，其次为交通波估计结果与变点识别估计结果。此外，对于数据层变量，电子警察流量重要性大于非排队轨迹数与排队轨迹数。

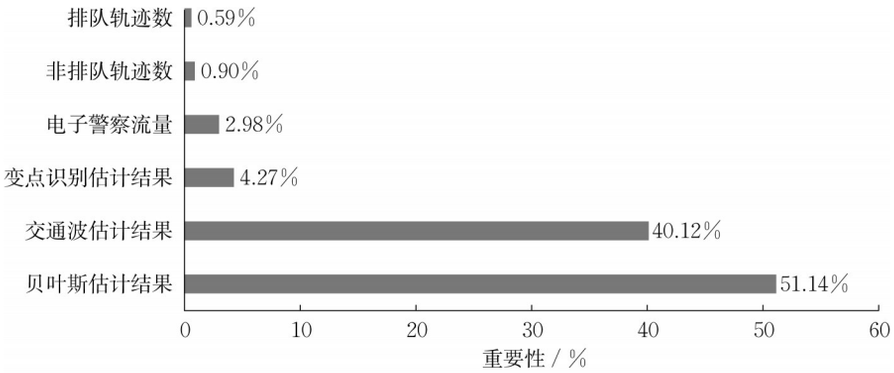


图6 特征重要性

Fig.6 Importance of features

特征个数是影响随机森林模型估计精度的重要因素之一。为评估特征个数对模型精度的影响，基于如图6所示的特征重要性分析结果，调整随机森林模型原始数据的特征向量维度，并使用OOB估计精度评估模型性能，从而选择最佳的特征个数。将特征个数选择范围设置为3～6，并按照如图6所示的重要性依次选择特征变量，对应的OOB估计精度结果如图7a所示。随着特征个数的增加，模型估计精度提升不显著。由此可见，随机森林模型对多维数据的处理能力较强，有着较强的特征选择能力，在本研究中不需考虑特征之间的多重共线问题，可选取原始特征向量所包含的全部6个特征。

相较于决策树算法，随机森林模型的优势主要在于多棵子树的随机性，因此决策子树棵数是影响随机森林估计精度的另一重要参数。图7b为决策子树棵数与OOB误差以及模型训练时间的关系。由图7b可知，随着决策子树棵数的增加，模型训练时间呈正比例增加。当决策子树棵数从1增加至5时，模型OOB误差从0.63快速下降至0.11左右；当决策子树棵数从5增加至15时，模型OOB误差的下降幅度逐渐放缓；当决策子树棵数超过15时，模型OOB误差不再发生显著变化。综上所述，当决策子树棵数设置为15时，模型OOB误差降至0，对应的模型训练时间仅为10.5 s，是较为合理的决策子树棵数。

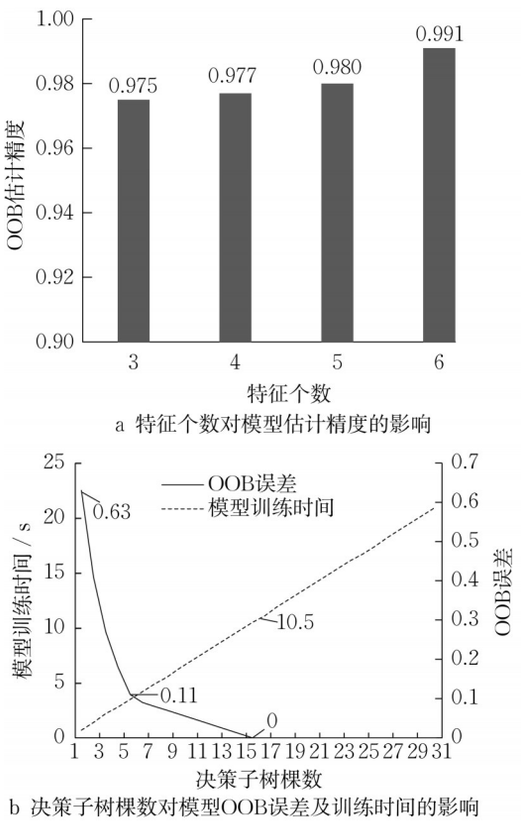


图7 标定参数对模型估计精度的影响

Fig.7 Effect of calibrated parameters on model estimation accuracy

4.3 仿真结果分析

基于4.2节的特征选择与参数优化结果，将随机森林模型的特征个数设置为6，最优决策子树棵数设置为15，在此基础上进行仿真验证。此外，还将基于电子警察数据的变点识别方法、基于网联车辆轨迹数据的交通波方法、基于电子警察与网联车辆轨迹数据融合的贝叶斯方法进行对比验证。在随机森林建模过程中，基于随机抽样方法选取75%的原始数据(8 042组)用于模型训练的数据集，选取剩余25%的原始数据(2 683组)作为验证数据集，并选用平均绝对误差(e_{MAE})与平均绝对百分比误差(e_{MAPE})对模型精度进行评价。e_{MAE}和e_{MAPE}计算式如下所示：

$$e_{MAE}=\frac{1}{m}\sum_{i=1}^{m}\left|\hat{Z}-Z\right| \tag{7}$$

$$e_{\mathrm{MAPE}}=\frac{1}{m}\sum_{i=1}^{m}\left|\frac{\hat{Z}-Z}{Z}\right|\times 100\% \tag{8}$$

式中：$\hat{Z}$表示基于集成学习的排队长度估计值；Z表示排队长度真实值；m表示当前研究时段内所含的周期个数。

在相同数据条件下，基于电子警察数据的变点识别方法、基于网联车辆轨迹数据的交通波方法、基于电子警察与网联车辆轨迹数据融合的贝叶斯方法，以及本方法的估计精度结果如图8所示。相较于单一数据源的估计方法（基于电子警察数据的变点识别方法、基于网联车辆轨迹数据的交通波方法），多源数据融合的估计方法（贝叶斯模型、随机森林）的估计精度明显更高。相较于贝叶斯模型，基于集成学习的排队长度估计精度明显提升，平均绝对误差为1.3 m，即小于1 辆•周期$^{-1}$，平均绝对百分比误差为1.4%，显著优于其他3类方法。本方法在轨迹数据渗透率随机变化、交通饱和度变化的复杂环境中适用性更强。

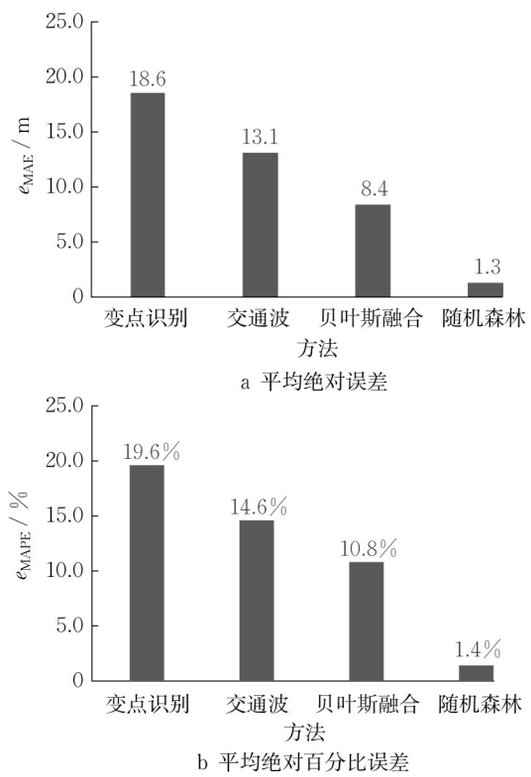


图8　不同排队长度估计模型精度对比

Fig.8　Comparison of estimation accuracy between different queue length estimation models

为验证集成学习方法在处理本研究问题的优势，选取8种常用模型进行对比分析，其中包括4种常用线性模型（岭回归、套索回归、弹性网络回归、贝叶斯回归）与4种常用的非线性模型（随机森林、梯度下降树（GBDT）、AdaBoost回归、多层感知机）。为确保不同模型评价指标的一致性，基于随机抽样方法，选取75% 的原始数据（8 042组）用于模型建立，选取剩余25% 的原始数据（2 683组）作为验证数据集，不同模型的平均绝对误差与平均绝对百分比误差如图9所示。

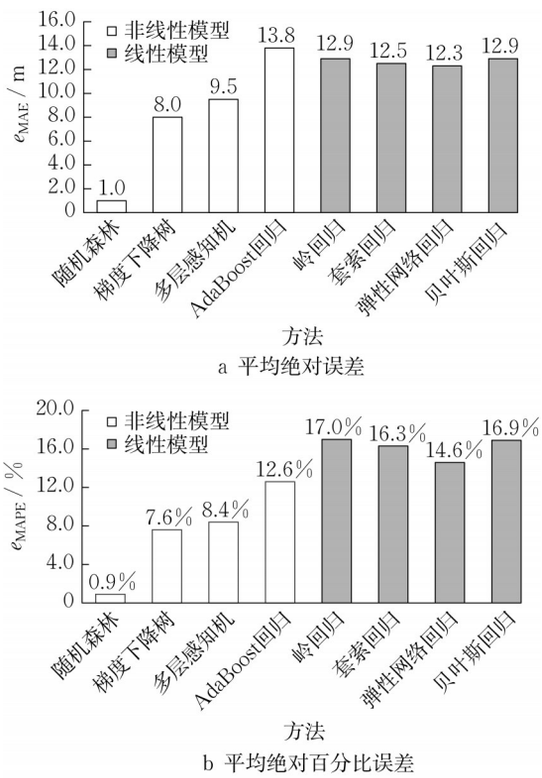


图9　不同学习模型精度对比

Fig.9　Comparison of estimation accuracy between different learning models

由图9可知，线性模型平均绝对误差均在12.3 m以上（约为1.8 辆•周期$^{-1}$），平均绝对百分比误差达到14.6% 以上。非线性模型在绝大多数情况下，其平均绝对误差小于9.5 m（约为1.4 辆•周期$^{-1}$），平均绝对百分比误差小于12.6%，非线性模型普遍优于线性模型。在集成学习算法中，随机森林具有明显优势，平均绝对误差为1.0 m，即小于1 辆•周期$^{-1}$，平均绝对百分比误差小于1%，显著低于其余7种模型，验证了集成学习在排队长度估计方面的优越性。

5　结语

通过融合我国城市道路交通检测中的2类新型

数据源——电子警察数据和网联车辆轨迹数据,提出了一种基于集成学习的信号控制交叉口排队长度估计方法。分析3种现有的排队长度估计方法的适用性,基于集成学习方法(随机森林)建立了2类数据源、3种方法估计结果与真实排队长度之间的非线性映射关系,并考虑饱和度、轨迹渗透率等因素生成仿真实验数据训练和优化集成学习器,实现对上一信号周期内车辆排队长度的后估计。相较于3种现有的排队长度估计方法,本研究的平均绝对误差低于1辆•周期$^{-1}$,精度得到明显提升。此外,随机森林方法的估计精度显著优于其他7类线性或非线性回归方法,具有良好的可移植性。

本研究仍存在一定的不足。例如,该方法需基于大量历史数据训练模型,对原始数据积累具有较高的要求。真实数据条件下,电子警察设备存在时钟漂移现象,轨迹数据也存在漂移点,2类数据源的时钟校准结果不匹配可能导致排队长度估计精度的降低。进一步的工作将解决上述问题,通过构建数据治理模块,对2类数据源进行校核与修复,提高排队长度的估计精度,并在此基础上开展基于真实场景的实证验证分析。

作者贡献声明:

吴 浩:模型构建,算法验证,论文撰写。

刘 磊:模型构建,算法验证。

唐克双:模型构建,论文撰写与修订。

参考文献:

[1] 杨晓光, 赵靖, 马万经, 等. 信号控制交叉口通行能力计算方法研究综述[J]. 中国公路学报, 2014, 27(5): 148.
YANG Xiaoguang, ZHAO Jing, MA Wanjing, *et al*. Review on calculation method for signalized intersection capacity [J]. China Journal of Highway and Transport, 2014, 27(5): 148.

[2] SHARMA A, BULLOCK D M, BONNESON J A. Input-output and hybrid techniques for real-time prediction of delay and maximum queue length at signalized intersections [J]. Transportation Research Record: Journal of the Transportation Research Board, 2007, 2035(1): 690.

[3] VIGOS G, PAPAGEORGIOU M. A simplified estimation scheme for the number of vehicles in signalized links[J]. IEEE Transactions on Intelligent Transportation Systems, 2010, 11(2): 312.

[4] LEE S, WONG S C, LI Y C. Real-time estimation of lane-based queue lengths at isolated signalized junctions [J]. Transportation Research, Part C: Emerging Technologies, 2015, 56: 1.

[5] ZHAN X, LI R, UKKUSURI S. Lane-based real-time queue length estimation using license plate recognition data [J]. Transportation Research, Part C: Emerging Technologies, 2015, 57: 85.

[6] ZHAN X, LI R, UKKUSURI S V. Link-based traffic state estimation and prediction for arterial networks using license-plate recognition data [J]. Transportation Research, Part C: Emerging Technologies, 2020, 117: 102660.

[7] SKABARDONIS A, GEROLIMINIS N. Real-time monitoring and control on signalized arterials [J]. Journal of Intelligent Transportation Systems, 2008, 12(2): 64.

[8] 姚荣涵, 王殿海. 拥挤交通流当量排队长度变化率模型[J]. 交通运输工程学报, 2009, 9(2): 93.
YAO Ronghan, WANG Dianhai. Change rate models of equivalent queue length for congested traffic flow[J]. Journal of Traffic and Transportation Engineering, 2009, 9(2): 93.

[9] LIU H X, WU X, MA W, *et al*. Real-time queue length estimation for congested signalized intersection [J]. Transportation Research, Part C: Emerging Technologies, 2009, 17(4): 412.

[10] 贾利民, 陈娜, 李海舰, 等. 基于单个地磁传感器的交叉口排队长度估计[J]. 吉林大学学报(工学版), 2016, 46(3): 8.
JIA Limin, CHEN Na, LI Haijian, *et al*. Intersection queue length estimation with single magnetic sensor [J]. Journal of Jilin University (Engineering and Technology Edition), 2016, 46(3): 8.

[11] 李爱杰, 唐克双, 董可然. 基于单截面低频检测数据的信号交叉口排队长度估计[J]. 交通信息与安全, 2018, 36(1): 57.
LI Aijie, TANG Keshuang, DONG Keran. Estimation of queuing length at signalized intersections using low-frequency point detector data [J]. Journal of Transport Information and Safety, 2018, 36(1): 57.

[12] YAO J, TANG K. Cycle-based queue length estimation considering spillover conditions based on low-resolution point detector data [J]. Transportation Research, Part C: Emerging Technologies, 2019, 109: 1.

[13] CHANG T H, LIN J T. Optimal signal timing for an oversaturated intersection [J]. Transportation Research, Part B: Methodological, 2000, 34(6): 471.

[14] BAN X J, HAO P, SUN Z. Real time queue length estimation for signalized intersections using travel times from mobile sensors [J]. Transportation Research, Part C: Emerging Technologies, 2011, 19(6): 1133.

[15] RAMEZANI M, GEROLIMINIS N. Queue profile estimation in congested urban networks with probe data [J]. Computer-Aided Civil and Infrastructure Engineering, 2015, 30(6): 414.

[16] LI F, TANG K, YAO J, *et al*. Real-time queue length estimation for signalized intersections using vehicle trajectory data [J]. Transportation Research Record: Journal of the Transportation Research Board, 2017, 2623(1): 49.

[17] YIN J, SUN J, TANG K. A Kalman filter-based queue length estimation method with low-penetration mobile sensor data at signalized intersections [J]. Transportation Research Record:

Journal of the Transportation Research Board, 2018, 2672(45): 253.

[18] ZHANG H, LIU H, CHEN P, *et al*. Cycle-by-cycle maximum queue length estimation at signalized intersections in connected vehicle environment[C]// 97th Annual Meeting of the Transportation Research Board. Washington DC: Transportation Research Board, 2018:1-9.

[19] COMERT G, CETIN M. Queue length estimation from probe vehicle location and the impacts of sample size[J]. European Journal of Operational Research, 2009, 197(1): 196.

[20] HAO P, BAN X, GUO D, *et al*. Cycle-by-cycle intersection queue length distribution estimation using sample travel times [J]. Transportation Research, Part B: Methodological, 2014, 68: 185.

[21] TIAPRASERT K, ZHANG Y, WANG X, *et al*. Queue length estimation using connected vehicle technology for adaptive signal control[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(4): 2129.

[22] TAN C, YAO J, TANG K, *et al*. Cycle-based queue length estimation for signalized intersections using sparse vehicle trajectory data [J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 22(1): 91.

[23] 谈超鹏，姚佳蓉，唐克双. 基于抽样车辆轨迹数据的信号控制交叉口排队长度分布估计[J]. 中国公路学报，2021，34(11)：282.
TAN Chaopeng, YAO Jiarong, TANG Keshuang. Queue length distribution estimation at signalized intersections based on sampled vehicle trajectory data [J]. China Journal of Highway and Transport, 2021, 34(11): 282.

[24] TANG K, WU H, YAO J, *et al*. Lane-based queue length estimation at signalized intersections using single-section license plate recognition data [J]. Transportmetrica B: Transport Dynamics, 2022, 10(1): 293.

[25] TAN C, LIU L, WU H, *et al*. Fuzing license plate recognition data and vehicle trajectory data for lane-based queue length estimation at signalized intersections[J]. Journal of Intelligent Transportation Systems, 2020, 24(5): 449.

[26] MA D, LUO X, JIN S, *et al*. Estimating maximum queue length for traffic lane groups using travel times from video-imaging data [J]. IEEE Intelligent Transportation Systems Magazine, 2018, 10(3): 123.

[27] LUO X, MA D, JIN S, *et al*. Queue length estimation for signalized intersections using license plate recognition data[J]. IEEE Intelligent Transportation Systems Magazine, 2019, 11(3): 209.

[28] BADILLO B, RAKHA H, RIOUX T, *et al*. Queue length estimation using conventional vehicle detector and probe vehicle data [C]//International IEEE Conference on Intelligent Transportation Systems. Anchorage: IEEE, 2012: 1674-1681.

[29] CAI Q, WANG Z, ZHENG L, *et al*. Shock wave approach for estimating queue length at signalized intersections by fusing data from point and mobile sensors [J]. Transportation Research Record: Journal of the Transportation Research Board, 2014, 2422(1): 79.

[30] BHASKAR A, QU M, CHUNG E. Bluetooth vehicle trajectory by fusing bluetooth and loops: motorway travel time statistics[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(1): 113.

[31] 吴翱翔. 基于多源数据的信号控制信号交叉口排队状态感知方法研究[D]. 上海：同济大学，2014.
WU Aoxiang. Research on the queue status sense of signalized intersections based on multi-source data[D]. Shanghai: Tongji University, 2014.

[32] 陶晶晶. 基于多源数据融合的单点信号控制交叉口排放估计与优化[D]. 上海：同济大学，2017.
TAO Jingjing. Emission estimation and optimization of signalized intersection based on multi-source data [D]. Shanghai: Tongji University, 2017.

[33] QOM S, HADI M, XIAO Y, *et al*. Queue length estimation for freeway facilities: based on combination of point traffic detector and automatic vehicle identification data [J]. Transportation Research Record: Journal of the Transportation Research Board, 2017, 2616(1): 19.

[34] 李爱杰. 基于路段定点检测器与电警数据融合的交叉口排队长度估计与预测[D]. 上海：同济大学，2018.
LI Aijie. Queue length estimation and prediction based on e-police and point detector data at signalized intersections[D]. Shanghai: Tongji University, 2018.

[35] XIAO J, XIAO Z, WANG D, *et al*. Short-term traffic volume prediction by ensemble learning in concept drifting environments [J]. Knowledge-Based Systems, 2019, 164: 213.

[36] CHEN X, CAI X, LIANG J, *et al*. Ensemble learning multiple LSSVR with improved harmony search algorithm for short-term traffic flow forecasting[J]. IEEE Access, 2018, 6: 9347.

[37] ZHANG C, MA Y. Ensemble machine learning: methods and applications[M]. Berlin: Springer Science & Business Media, 2012.

[38] LIU Y, WU H. Prediction of road traffic congestion based on random forest [C]//2017 10th International Symposium on Computational Intelligence and Design (ISCID). Hangzhou: IEEE, 2017, 2: 361-364.

[39] DOGRU N, SUBASI A. Traffic accident detection using random forest classifier[C]// 15th Learning and Technology Conference (L&T). Jeddah: IEEE, 2018: 40-45.

[40] LOH W Y. Classification and regression tree methods[M]// Encyclopedia of Statistics in Quality and Reliability. New York: Wiley, 2008.

[41] ABNEY S. Bootstrapping [C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Groningen: Association for Computational Linguistics, 2002: 360-367.