文章编号:1005-0523(2021)03-0061-06

多因素轨道交通客流量预测模型研究

黄海超,陈景雅,王 爽,王方伟

(河海大学土木与交通学院,江苏 南京 210098)

摘要:针对传统预测模型只关注时间因素的不足,提出一种引入天气因素同时考虑日期属性的预测模型。首先通过显著性检验确定天气因素与客流量的相关程度,再采用灰色关联度分析(GRA)计算各天气因素与客流量的非线性关联度,逐步筛选关联度低的天气因素。每次筛选后利用双向长短期记忆(BiLSTM)神经网络进行预测,提出GRA-BiLSTM预测模型。结果表明:将GRA值低于0.6的天气因素作为变量会降低预测精度,逐步剔除关联度低的天气因素获得的GRA-BiLSTM相较于传统LSTM,无论工作日还是非工作日,预测误差均显著降低,同时收敛速度与鲁棒性也优于传统机器学习。

关键词:城市交通;客流预测;多因素;灰色关联度分析;双向长短期记忆神经网络

中图分类号:U293.5 **文献标志码**:A

本文引用格式:黄海超,陈景雅,王爽,等. 多因素轨道交通客流量预测模型研究[J]. 华东交通大学学报,2021,38(3):61-66.

DOI:10.16749/j.cnki.jecjtu.20210706.018

Multi-factor Rail Transit Passenger Flow Prediction Model

Huang Haichao, Chen Jingya, Wang Shuang, Wang Fangwei

(College of Civil and Transportation Engineering, Hohai University, Nanjing 210098, China)

Abstract: Traditional prediction model only focuses on time factors. Aiming at that deficiency, a prediction model which introduces weather factor and considers date attribute was proposed. Firstly, the degree of correlation between weather factors and passenger flow was determined by significance test. Then, to gradually screen the low-relevant weather factors, grey relation analysis (GRA) was adopted to calculate the non-linear correlation between various weather factors and passenger flow. After each screening, the bi-directional LSTM neural network was used to forecast and the GRA-BiLSTM prediction model was proposed. The results show that taking weather factors with GRA value less than 0.6 as input will reduce the prediction accuracy. Compared with the traditional LSTM, the prediction error of GRA-BiLSTM, which is obtained by gradually eliminating the weather factors with low correlation, is significantly reduced on both work days, and non_work days and the convergence speed and robustness are better than the traditional machine learning as well.

Key words: urban traffic; passenger flow prediction; multi-factor; grey relation analysis; Bi-directional long-short-term memory network

Citation format: HUANG H C, CHEN J Y, WANG S, et al. Multi-factor rail transit passenger flow prediction model[J]. Journal of East China Jiaotong University, 2021, 38(3): 61-66.

收稿日期:2021-01-16

基金项目:国家自然科学基金项目(52078190);教育部人文社会科学研究规划基金(18YJAZH119)

作者简介:黄海超(1996—),男,硕士研究生,研究方向为深度学习在智能交通领域的应用。E-mail:hhc123@hhu.edu.cn。

通信作者:陈景雅(1967—),女,教授,研究方向为互通立交安全评估。E-mail:13805151397@139.com。

轨道交通客流量预测是城市智能交通系统的重要组成部分,客流量不仅受历史客流量、日期属性等时间因素影响,也受空间因素如:温度、风速、气压等影响。时间因素与空间因素对客流量的影响存在复杂非线性关系,是多种因素共同作用的结果[1]。

李洁等[2]在传统预测模型的基础上引入出行日期特征及节假日因素,建立季节性差分自回归滑动平均模型(SARIMA),有效提高预测精度。Sui等[3]建立了周末客流量与天气因素之间关系的多元回归模型,分析两者的内在联系。Xue等[4]着重研究降雨因素对客流的影响。许熳灵等[5]进一步引入不同的天气因素,并验证其对地铁客流量时空分布的影响。Ni等[6]考虑地铁附近的媒体事件与客流变化规律,提高客流量的预测精度。Tselentis等[7]证明了单一模型具有局限性,加之如今庞大的数据规模,基于深度学习的混合预测模型[8]逐步替代了传统机器学习。滕靖等[9]在考虑日期属性和天气因素的同时,采用粒子群算法(PSO)优化长短期记忆神经网络(LSTM)并证明混合模型优于传统的LSTM。谢宇等[10]引入突发事件因素建立混合模型,提高模型对突发事件的预测精度。此外,不引入外部因素,考虑邻近站点的空间关联的预测模型,如时空长短期记忆(ST-LSTM)[11],多模式深度融合(MPDF)[12]等也一定程度上提高预测精度。但是引入越多因素必然导致模型的复杂度提高,需要在保证模型精度的同时提高模型效率。李泽文等[13]提出主成分分析(PCA)对引入的多因素进行降维,减少预测时间。徐先峰等[14]通过K近邻算法提取特征,结合双向长短期记忆(BiLSTM)兼顾预测精度与速度。李梅等[15]提出Pearson相关分析法提取显著影响因子,确定引入因素的维数,具有良好的适用性。

由于空间、时间因素与客流量之间存在复杂的非线性关系,目前针对多因素与客流量关系的研究较少。本文提出一种多因素建模方式,首先采用显著性检验确定采集的天气因素是否与客流量序列相互独立。再通过GRA提取天气因素与客流量潜在的非线性关系并量化,根据GRA值逐步筛选关联度低的天气因素,降低模型复杂度。最后采用深度学习的BiLSTM进行预测,构建GRA-BiLSTM混合模型。同时与传统预测模型进行对比,验证模型可靠性。

1 数据预处理

1.1 数据预处理及相关性分析

轨道交通客流量数据来自明尼苏达州明尼阿波利斯至圣保罗2017-01-01~2018-09-30每小时轨道交通客流量,明尼阿波利斯与圣保罗分别是明尼苏达州第一、二大城市,旅客往来频繁,客流规模大;天气数据来自美国国家海洋和大气管理局,天气数据共7类,包括:大气压、温度、露点、风速、云量、可见度、天气状况。对完全相同的、明显异常及缺失的数据进行清洗,获得8种数据,每种数据15 246条。考虑到工作日与非工作日,人群出行规律,受天气因素影响不同,从而使得客流量数据在工作日与非工作日呈现不同的特征,将数据划分为工作日(约8.7万条)与非工作日(约3.4万条),分别训练。

采用Person相关系数检验天气变量与客流量的相关程度,利用t双边检验确定天气因素是否与客流量服从不同分布。结果如表1所示,Person系数均不超过0.4,表明天气因素与客流量并非简单的线性关系,所有天气因素在0.05水平上存在显著差异,可作为独立影响因素。

表1 天气因素相关性分析

Tab.1 Correlation analysis of weather factors

天气因素	Person系数	P值
风速	0.042 0*●	0
可见度	0.016 5*●	0
露点	0.023 8*●	0
大气压	0.017 1*●	0.000 1
温度	0.128 4*●	0
云量	0.137 5*	0.029 8
天气状况	-0.007 6*●	0

注:* 表示在0.05水平上存在显著差异,*●表示在0.01水平上存在显著差异。

1.2 灰色关联度分析

GRA是灰色系统理论中的一种多因素统计分析方法,对一个系统发展变化态势进行定量描述和

比较，适用于探究非线性相关性。其基本思想是通过确定参考序列和若干个比较序列的几何形状相似程度来衡量因素间关系的强弱。天气因素对客流量的影响不是简单的线性关系，采用 GRA 对其潜在的非线性关系进行进一步分析，步骤如下：

1）确定系统中的参考序列（客流量）和比较序列（天气因素）。

2）对客流量序列与天气因素序列进行无量纲化处理，消除不同量纲的影响，考虑到之后要将其作为神经网络的输入变量，将其归一到[−1，1]。

3）计算客流量序列与天气序列的关联系数

$$\xi_i(k)=\frac{\min\limits_i\min\limits_k|x_0(k)-x_i(k)|+\rho\min\limits_i\min\limits_k|x_0(k)-x_i(k)|}{|x_0(k)-x_i(k)|+\rho\min\limits_i\min\limits_k|x_0(k)-x_i(k)|} \tag{1}$$

式中：$\xi_i(k)$为客流量序列与第 i 个天气序列的关联系数，k 为序列元素的索引；$x_0(k)$为归一化的客流量序列；$x_i(k)$为归一化的天气因素序列；ρ 为分辨系数，ρ 越小，关联系数间差异越大，区分能力越强，通常取 0.5。

4）计算灰色关联度

$$r_i=\frac{1}{m}\sum_{k=1}^{m}\xi_i(k) \tag{2}$$

式中：m 为天气因素的数目。

灰色关联度如表 2 所示，天气因素与客流量相关性强弱为：大气压>温度>露点>风速>云量>天气状况>可见度。

表 2　天气因素灰色关联度

Tab.2　Grey relational degree of weather factors

天气因素	大气压	温度	露点	风速	云量	天气状况	可见度
灰色关联度	0.681 3	0.674 0	0.657 4	0.643 0	0.620 5	0.598 1	0.583 3

2　BiLSTM 神经网络

LSTM 是深度循环神经网络的代表，通过门控单元实现时空记忆功能，同时有效缓解循环神经网络的梯度消失和爆炸问题。通过引入输入门、遗忘门和输出门等门控单元系统，LSTM 可以控制何时忘记历史信息或使用新信息更新单元状态，在解决非线性时间序列问题时效果极佳，其过程可表达为

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_{xi}\boldsymbol{X}_t+\boldsymbol{W}_{hi}\boldsymbol{H}_{t-1}+\boldsymbol{W}_{ci}\boldsymbol{C}_{t-1}+\boldsymbol{b}_{\mathrm{i}}) \tag{3}$$

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_{xf}\boldsymbol{X}_t+\boldsymbol{W}_{hf}\boldsymbol{H}_{t-1}+\boldsymbol{W}_{cf}\boldsymbol{C}_{t-1}+\boldsymbol{b}_{\mathrm{f}}) \tag{4}$$

$$\boldsymbol{c}_t=\boldsymbol{f}_t\boldsymbol{c}_{t-1}+i_t\tanh(\boldsymbol{W}_{xc}\boldsymbol{X}_t+\boldsymbol{W}_{hc}\boldsymbol{H}_{t-1}+\boldsymbol{b}_{\mathrm{c}}) \tag{5}$$

$$\boldsymbol{o}_t=\sigma(\boldsymbol{W}_{xo}\boldsymbol{X}_t+\boldsymbol{W}_{ho}\boldsymbol{H}_{t-1}+\boldsymbol{W}_{co}\boldsymbol{C}_{t-1}+\boldsymbol{b}_{\mathrm{o}}) \tag{6}$$

$$H_t=\boldsymbol{o}_t\tanh(\boldsymbol{C}_t) \tag{7}$$

式中：$\boldsymbol{i}_t$，$\boldsymbol{f}_t$，$\boldsymbol{o}_t$，$\boldsymbol{c}_t$ 分别为输入门 i，遗忘门 f，输出门 o，细胞状态 c 在 t 时刻的输出；$\boldsymbol{W}_{xi}$，$\boldsymbol{W}_{xf}$，$\boldsymbol{W}_{xo}$，$\boldsymbol{W}_{xc}$ 分别为输入 $\boldsymbol{X}_t$ 与输入门 i，遗忘门 f，输出门 o，细胞状态 c 的权重；$\boldsymbol{W}_{hi}$，$\boldsymbol{W}_{hf}$，$\boldsymbol{W}_{ho}$，$\boldsymbol{W}_{hc}$ 为隐藏层输出 $\boldsymbol{H}_t$ 与相应门的权重；$\boldsymbol{W}_{ci}$，$\boldsymbol{W}_{cf}$，$\boldsymbol{W}_{co}$ 为细胞状态输出 $\boldsymbol{C}_t$ 与相应门的权重；$\boldsymbol{b}_{\mathrm{i}}$，$\boldsymbol{b}_{\mathrm{f}}$，$\boldsymbol{b}_{\mathrm{o}}$，$\boldsymbol{b}_{\mathrm{c}}$ 为偏置向量；σ 为激活函数。

BiLSTM 基本原理与普通 LTSM 相同，结合正向、反向 LSTM 同时对输入时间序列 $\boldsymbol{X}_t$ 进行前向和后向两次训练，获得预测交通流序列 $\boldsymbol{Y}_t$。BiLSTM 考虑数据间的关联性，进一步提高 LSTM 特征提取的全局性和完整性。BiLSTM 结构如图 1 所示。

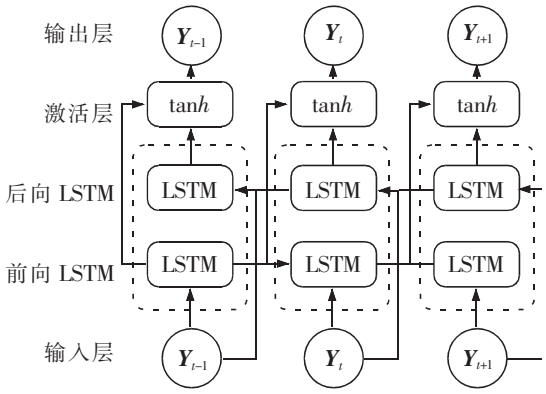


图 1　BiLSTM 神经网络结构

Fig.1　BiLSTM neural network structure

3　应用实例

3.1　模型构建与天气因素分析

将天气数据与客流量数据归一化处理，并以前 70 周数据作为训练集，最后 21 周作为测试集验证模型性能。将前 3 h 天气因素、客流量（24 维）及当前天气因素（7 维）共 31 维数据作为输入，当前客流量作为输出。迭代次数取 200，初始学习率设为 0.005，同时为防止学习率过大，模型来回震荡，采用衰减法动态调整学习率，每迭代 50 次学习率衰减 50%。损失函数采用均方根误差，通过 AdamOptimizer 优化器对模型进行优化。为优化隐含层神经元个数，随机抽取 20%作为样本进行试

验,结果如图 2 所示。当隐含层神经元超过 64 个后训练时间大幅上升，训练误差减小效果不明显,确定隐含层神经元个数为 64。

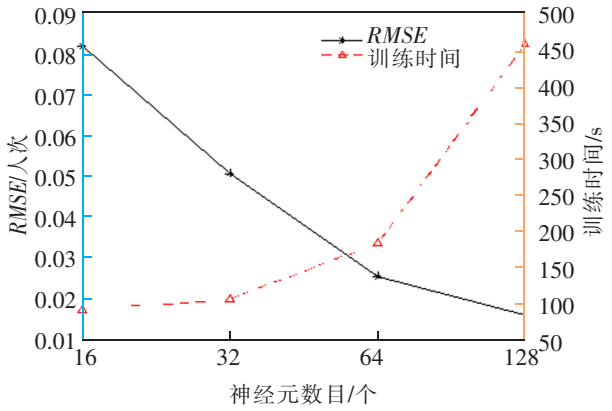
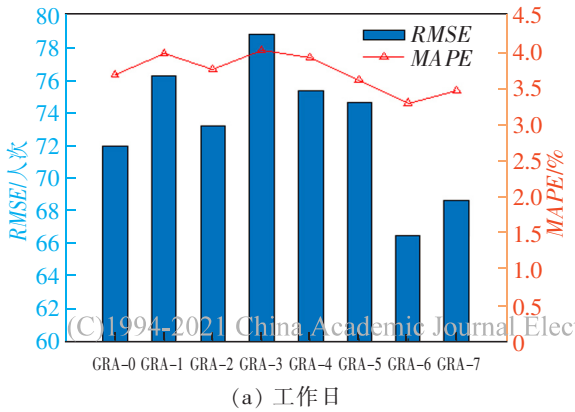


图 2 隐含层神经元数目寻优

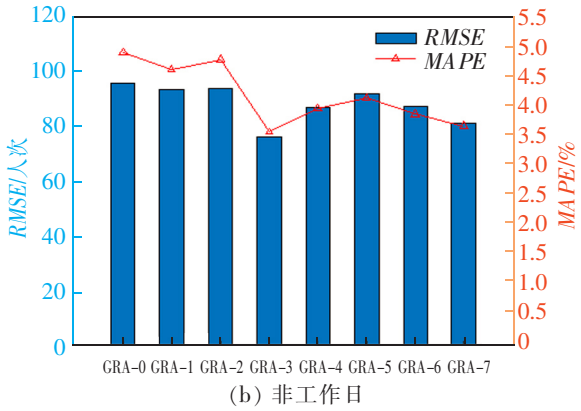
Fig.2 Optimization of neurons number in the hidden layer

根据以上参数,建立 BiLSTM 神经网络,将包含所有天气因素的神经网络命名为 GRA-0，根据 GRA 值,逐步筛选关联度低的天气因素,如 GRA-1 为剔除“可见度”因素,输入变量为 27 维的模型,GRA-2 为剔除“可见度”、“天气状况”后输入变量为 23 维的模型,最终 GRA-7 即为剔除 7 种天气因素,输入变量仅为历史客流量的模型,也就是传统的神经网络模型。

平均绝对百分误差(*MAPE*)体现了真实值与预测值的相对偏差，可直接衡量预测结果的好坏,常被用于评价预测模型的优劣。均方根误差(*RMSE*)可直接体现真实值与预测值的绝对差值,且特大或特小误差对指标影响明显,可有效弥补 *MAPE* 的不足。采用这两种指标评价模型性能。考虑到日期属性对客流量的影响，工作日与非工作日分别训练,每种模型训练 3 次取平均值,结果如图 3 所示。



(a) 工作日



(b) 非工作日

图 3 GRA-BiLSTM 工作日与非工作日预测结果

Fig.3 GRA-BiLSTM prediction results on weekdays and weekend

从图 3 可以看出,无论是否处于工作日,不考虑天气因素的神经网络(GRA-7),其预测误差均小于考虑所有因素的神经网络 (GRA-0)。这说明将所有天气因素作为输入并不能提高预测精度，反而因为关联度低的天气因素含有大量与客流量无关的信息，相当于训练过程拟合了样本噪声,预测精度反而降低。通过 GRA 逐步筛选关联度低的天气因素,预测误差有所降低,这是因为通过剔除低关联度的因素相当于样本降噪。

天气因素对工作日与非工作日客流量的影响存在明显差异。在工作日期间,仅考虑“大气压”的神经网络(GRA-6)误差最小,这说明工作日客流量受天气因素影响较小,大部分天气因素相当于冗余的噪声，降低模型的学习效率；在非工作日期间,同时考虑“大气压”、“温度”、“露点”、“风速”的神经网络(GRA-3)误差最小,这说明非工作日的客流量受天气影响比较明显,这些天气因素都能给神经网络带来正向影响。由表 2 可知,最佳神经网络模型考虑的天气因素,其 GRA 值均大于 0.6。当天气因素与客流量的 GRA 值高于 0.6 时,才可认为该因素与客流量存在潜在关联,能提高神经网络的预测精度。

模型的收敛速度能反映所需的训练时间,收敛越快,训练时间相应越短。为验证根据 GRA 筛选天气因素对神经网络收敛速度的影响。分别取工作日与非工作日中预测效果最佳的神经网络与 GRA-0，GRA-7 训练过程的 LOSS 曲线进行对比。结果如图 4 所示。

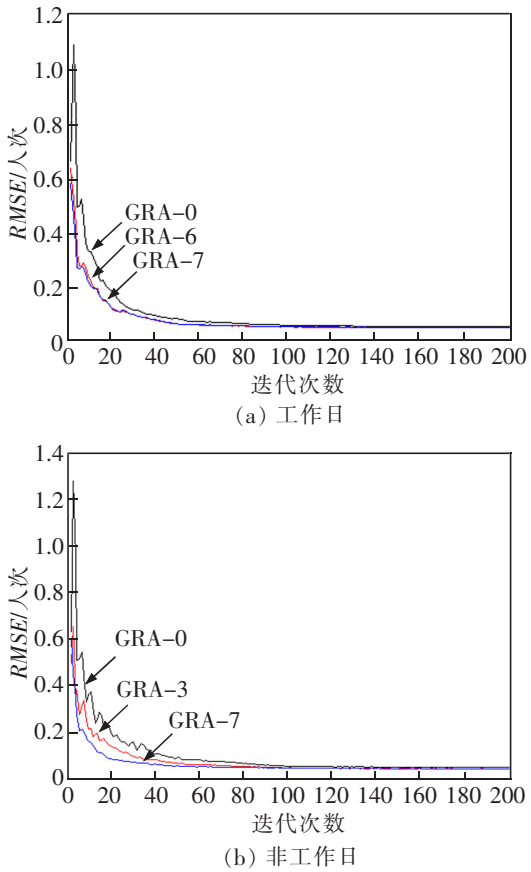


图 4　工作日与非工作日收敛速度对比

Fig.4　Comparison of convergence rates on weekdays and weekend

可以看出,输入维数最高的 GRA–0,早期波动较大,且收敛速度最慢,需要最长的训练时间。采用 GRA 降维后,神经网络收敛速度显著提高,训练效果更加稳定,且最终的训练效果优于传统模型 GRA–7。

3.2　模型对比

为验证模型可靠性,同时比较常用预测模型对多种天气因素的应答。分别采用 LSTM 神经网络、随机森林(RF)、最小二乘支持向量机(LSSVM)作为基准模型。LSTM 采用与原模型相同的参数设置,RF 预测模型根据经验公式确定子树数目,LSSVM 预测模型采用网格搜索确定最佳超参数 c,g,预测结果如图 5 所示。各模型经 GRA 处理后的预测误差 *RMSE* 汇总于表 3,最优模型的预测结果加横线表示。

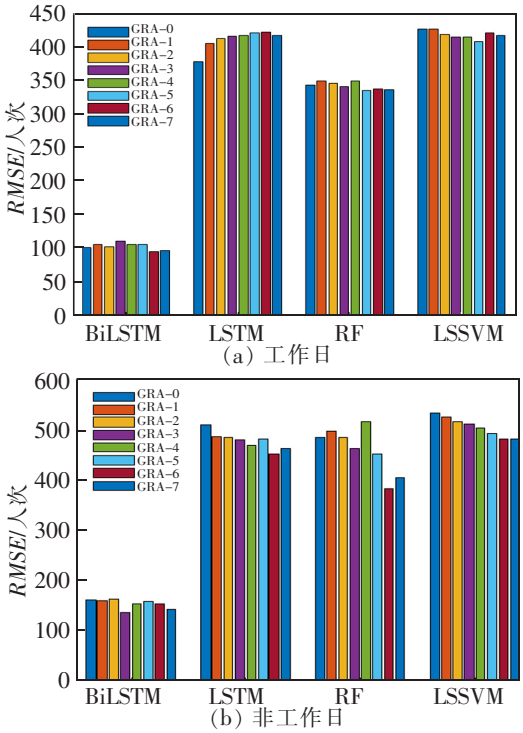


图 5　工作日与非工作日预测模型对比

Fig.5　Comparison of prediction model on weekdays and weekend

表 3　不同预测模型对天气因素的应答

Tab.3　Responses of different prediction models to weather factors

	模型	GRA–0	GRA–1	GRA–2	GRA–3	GRA–4	GRA–5	GRA–6	GRA–7
工作日	BiLSTM	101.820	106.594	103.404	111.683	107.197	107.143	96.074	97.801
	LSTM	379.748	406.818	414.948	417.665	419.368	423.477	424.399	419.875
	RF	345.448	350.999	347.065	343.107	351.159	337.072	339.018	338.503
	LSSVM	428.961	428.910	420.915	416.934	416.883	410.352	422.809	419.527
非工作日	BiLSTM	162.452	159.979	163.317	137.108	154.140	158.598	152.996	142.709
	LSTM	509.638	485.281	484.179	479.542	468.398	481.065	451.010	462.690
	RF	483.488	497.767	484.635	462.710	515.544	450.903	382.375	404.876
	LSSVM	533.444	525.766	515.329	511.010	503.062	492.631	481.774	481.143

可以看出,无论是工作日还是非工作日,最优GRA-BiLSTM预测误差显著低于各最优基准模型。由于关联度低的天气因素包含大量噪声,绝大部分GRA-0模型的预测误差高于GRA-7,且相差较大。BiLSTM在工作日与非工作日的误差波动均低于其他模型,体现良好的抗噪声干扰能力。

4 结论

本文同时考虑时间因素(历史客流、日期属性)及空间因素(天气),提出GRA-BiLSTM预测模型,有效降低预测误差,同时提高收敛速度及抗噪声干扰能力。

1)根据GRA逐步剔除低关联度变量,发现工作日与非工作日客流量对天气的应答不同,但GRA值0.6可作为天气因素对客流量是否具有正向影响的分界。

2)关联度低的天气因素可能与客流量相关的信息较少,并非完全无关。简单剔除仍会损失部分有效信息,天气因素对客流量如何影响还需进一步研究。

参考文献:

[1] WU J W,LIAO H. Weather,travel mode choice,and impacts on subway ridership in Beijing[J]. Transportation Research Part A:Policy and Practice,2020,135:264-279.

[2] 李洁,彭其渊,杨宇翔. 基于SARIMA模型的广珠城际铁路客流量预测[J]. 西南交通大学学报,2020,55(1):41-51.

[3] SUI T,JONATHAN C,FRANCISCO R,et al. To travel or not to travel:'Weather' is the question. Modelling the effect of local weather conditions on bus ridership[J]. Transportation Research Part C:Emerging Technologies,2018,86:147-167.

[4] XUE F,YAO E J,HUAN N,et al. Prediction of urban rail transit ridership under rainfall weather conditions[J]. Journal of Transportation Engineering,Part A:Systems,2020,146(7):04020061.

[5] 许嫚灵,付晓,汤君友,等.天气因素对城市地铁客流时空分布的影响——基于智能交通卡数据的实证研究[J]. 地理科学进展,2020,39(1):45-55.

[6] NI M,HE Q,GAO J. Forecasting the subway passenger flow under event occurrences with social media[J]. IEEE Transactions on Intelligent Transportation Systems,2017,18(6):1623-1632.

[7] TSELENTIS D I,VLAHOGIANNI E I,KARLAFTIS M G. Improving short-term traffic forecasts:To combine models or not to combine?[J]. IET Intelligent Transport Systems,2015,9(2):193-201.

[8] 李艳,彭春华,傅裕,等. 基于CNN-LSTM网络模型的风电功率短期预测研究[J]. 华东交通大学学报,2020,37(4):109-115.

[9] 滕靖,李金洋. 考虑日期属性和天气因素的铁路城际短期客流预测方法[J]. 中国铁道科学,2020,41(5):136-144.

[10] 谢宇,王丽清,徐永跃,等. 面向多因素影响的混合预测模型[J]. 计算机工程与设计,2020,41(10):2758-2764.

[11] TANG Q C,YANG M N,YANG Y,et al. ST-LSTM:A deep learning approach combined spatio temporal features for short term forecast in rail transit[J]. Journal of Advanced Transportation,2019:8392592.

[12] BAI Y,SUN Z Z,ZENG B,et al. A multi-pattern deep fusion model for short-term bus passenger flow forecasting[J]. Applied Soft Computing,2017,58:669-680.

[13] 李泽文,胡让,刘湘,等. 基于PCA-DBILSTM的多因素短期负荷预测模型[J]. 电力系统及其自动化学报,2020,32(12):32-39.

[14] 徐先峰,刘阿慧,陈雨露,等. 基于气象因素充分挖掘的BiLSTM光伏发电短期功率预测[J]. 计算机系统应用,2020,29(7):205-211.

[15] 李梅,李静,魏子健,等. 基于深度学习长短期记忆网络结构的地铁站短时客流量预测[J]. 城市轨道交通研究,2018,21(11):42-46.