

DOI:10.19853/j.zgjsps.1000-4602.2025.23.012

XGBoost结合建筑特征信息预测生活小区污水量

陈 成¹, 王福鹏¹, 闫 钰², 刘学勇², 董艳红², 董文艺^{1,3,4},
王宏杰^{1,3,4}

(1. 哈尔滨工业大学<深圳> 生态环境学院, 广东 深圳 518055; 2. 中国市政工程东北设计研究总院有限公司, 吉林 长春 130021; 3. 城乡水资源与水环境全国重点实验室, 黑龙江 哈尔滨 150090; 4. 深圳市水资源利用与环境污染控制重点实验室, 广东 深圳 518055)

摘 要: 生活小区污水排放量数据是采用模型法进行污水管网检漏的重要基础条件,但现有的直接监测法、定额法和给水量法存在监测量大、精度低和数据难获取等问题。考虑到生活小区的建筑特征信息(如建筑面积、住户数量等)与排水行为密切相关,通过收集上述易获取的建筑特征信息,结合极端梯度提升(XGBoost)机器学习算法的优异非线性拟合能力,具有实现生活小区污水量高精度预测的潜力。为此,对多个典型生活小区的污水排放量进行监测,基于小区建筑特征信息并结合XGBoost算法,建立生活小区污水量预测模型。结果表明,预测模型的主要参数为住户数、建筑面积、占地面积、竣工时间和排水时间,经预处理和超参数优化后,模型在一天中08:00—24:00的污水量预测误差较小,测试集的平均相对误差为11%,优于定额法和给水量法的42%和17%。

关键词: 污水管网; 外水诊断; 生活小区污水量预测; 建筑特征信息; XGBoost

中图分类号: TU992 **文献标识码:** A **文章编号:** 1000-4602(2025)23-0077-07

Predicting Sewage Volume in Residential Communities Based on Building Characteristic Information Using XGBoost Algorithm

CHEN Cheng¹, WANG Fu-peng¹, YAN Yu², LIU Xue-yong², DONG Yan-hong²,
DONG Wen-yi^{1,3,4}, WANG Hong-jie^{1,3,4}

(1. College of Ecology and Environment, Harbin Institute of Technology <Shenzhen>, Shenzhen 518055, China; 2. China Northeast Municipal Engineering Design and Research Institute Co. Ltd., Changchun 130021, China; 3. State Key Laboratory of Urban Water Resource and Environment, Harbin 150090, China; 4. Shenzhen Key Laboratory of Water Resource Utilization and Environmental Pollution Control, Shenzhen 518055, China)

Abstract: The volume of domestic sewage discharge in residential communities serves as a fundamental input for leakage detection in sewer networks using model-based approaches. However, conventional methods such as direct monitoring, quota-based estimation, and water consumption-based estimation suffer from limitations including extensive monitoring requirements, low accuracy, and poor data accessibility. Given the close correlation between sewage discharge behavior and building

基金项目: 深圳市科技计划项目(KCXFZ20240903093500001); 中国市政工程东北设计研究总院有限公司科技研发项目
通信作者: 王宏杰 E-mail: whj1533@qq.com

characteristics (such as total floor area and number of households), this study explored the potential of using easily accessible building characteristic information, combined with the strong nonlinear fitting capabilities of the XGBoost machine learning algorithm, to achieve high-accuracy predictions of sewage discharge volumes. To this end, sewage discharge data were collected from multiple representative residential communities. A predictive model for community-scale sewage discharge was developed by integrating building characteristic variables with the XGBoost algorithm. The results indicated that key predictive characteristics included the number of households, total floor area, land area, completion year, and typical drainage hours. After data preprocessing and hyperparameter optimization, the model achieved high prediction accuracy during 08:00 and 24:00, with an average relative error of 11% on the test set, which outperformed the quota-based (42%) and water consumption-based (17%) methods.

Key words: sewer network; external water diagnosis; sewage volume prediction in residential communities; building characteristic information; XGBoost

我国城市污水管网普遍面临外水入侵的问题,因此亟须开展外水诊断工作^[1-2]。模型法通过分析管道预测流量与实测流量之间的差异^[3-5],识别是否存在外水入侵,具备较高的计算精度与效率,因而在实际应用中受到越来越多的关注。然而,模型法通常依赖于区域内排水户(生活小区、商场、工业园等)的实际排水数据,而居民生活小区是城市排水户的主要类型。由于小区数量众多,直接监测存在监测量大、监测设备多等问题,目前仍面临较大困难。传统的定额法通过人口和折减系数估算污水排放量,往往只体现日均水平,预测精度有限;另一种基于给水量并乘以折减系数的估算方式,未能充分考虑污水排放相对于给水行为存在的时间延迟及衰减特征,因而同样存在一定偏差。因此,亟须开发一种准确高效的生活小区污水量预测方法,从而为提升整个污水管网外水诊断的精度奠定基础。

建筑特征信息如户数、排水时间等变量与居民数量和用水排水习惯密切相关,因此用于污水量预测具备一定的理论依据。Jia等^[6]利用建筑物体积和人口密度估算建筑使用人数,将实测污水量按照建筑人数比例进行分配,从而进行污水管网建模并取得了较好效果,流量模拟的拟合优度达到0.91。何韶渺等^[7]利用地理信息数据识别建筑物类型,结合相关规范估算其污水量,纳什效率系数达到0.9,预测精度较高。但上述方法均依赖于管网实测流量和液位数据对小区预测流量进行分配调整,无法实现直接基于建筑物特征预测其污水量,因此难以用于外水诊断。若收集多个典型生活小区污水量

数据,建立小区污水量数据集,结合预测模型可实现基于建筑物特征直接预测污水量。由于建筑特征信息变量多,变量之间关系复杂,且存在交互作用和非线性关系,因此简单的线性回归方法难以实现高精度预测,而机器学习中的极端梯度提升(XGBoost)算法能生成多个预测函数^[8],并对预测函数进行不断优化迭代,因此其可充分挖掘数据规律,具有优异的非线性拟合能力和预测精度^[9]。由于其对数据准确性、数据与因变量相关性等要求较高,需要筛选出数据中重要的变量,并对数据进行降噪和特征提取等预处理,才能实现高精度预测。

基于上述背景,首先对典型生活小区开展污水量监测工作,提取和筛选影响排水量的关键因素;随后,构建融合建筑特征信息与XGBoost算法的生活小区污水量预测模型;最后,分析生活小区污水量预测模型的预测效果,优选预测误差较小的时间段,为城市污水管网健康管理提供技术支持。

1 材料与方法

1.1 研究数据获取

1.1.1 污水量数据获取

生活小区污水排放总口水量小、液位低,常低于多普勒流量计监测最低水位,因此需要开发一种低水位下的流量监测设备。三角薄壁堰通过堰板形成壅水建立流量和液位的特定关系,将监测堰上水头代入堰流公式即可得待测流量^[10],受低液位影响较小。因此,开发了一种管道三角堰流量计(见图1),其将堰板设计为圆形与管道形状匹配,通过充气膨胀的环形密封圈实现管道密封,液位数据通过

堰板背部的液位计探头读取。经率定,该流量计流量监测范围为 $0.51\sim 45\text{ m}^3/\text{h}$,平均相对误差为 4.6% 。

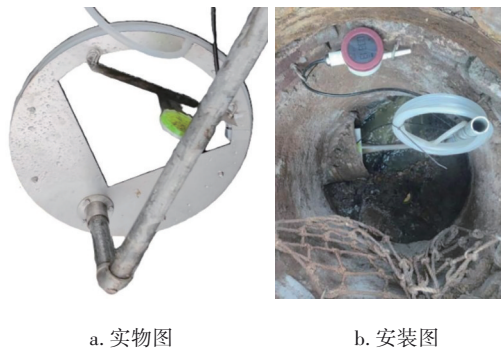


图1 管道三角堰流量计
Fig.1 Pipe triangular weir flowmeter

1.1.2 建筑特征信息数据获取

通过小区物业、售房平台、地图、排水管网GIS平台等渠道收集不同建筑特征信息数据。其中,户数和竣工时间:售房平台+小区物业,房价:售房平台,楼栋数和占地面积:地图量取,建筑面积:小区不同楼栋平面面积乘以楼栋层数之和,排口数:排水管网GIS平台。

1.2 研究区域概况

本研究需要对多个小区进行污水量监测,选择小区的标准主要有:污水排放总口数量为 $1\sim 2$ 个,总口监测条件良好,无淤积堵塞,供水水表正常工作,小区类型具有代表性,户数、面积、竣工时间等参数总体分布均匀。最终选择9个小区,其竣工时间分别为1995年、2017年、2004年、2012年、2000年、2020年、2004年、2002年、2012年,除了小区C和G有2个排口外,其他小区均为1个,其他信息见表1。

表1 样本小区建筑特征信息
Tab.1 Building characteristic information of sample communities

小区	户数/户	建筑面积/ m^2	占地面积/ m^2	房价/元	楼栋数/栋
A	280	49 870	8357	55 200	4
B	433	58 904	10 600	43 100	3
C	272	32 370	7 750	56 400	8
D	363	95 000	3 509	33 650	2
E	240	23 757	11 400	64 200	3
F	672	103 000	22 300	55 800	6
G	548	79 675	31 600	58 300	9
H	360	46 958	13 600	51 000	3
I	1 106	122 008	24 700	54 000	5

应用管道三角堰流量计对上述9个生活小区进行流量监测,监测结果如图2所示。小区A~I的流量范围分别为 $0.68\sim 11.50$ 、 $1.87\sim 25.53$ 、 $1.11\sim 17.47$ 、 $1.02\sim 16.69$ 、 $1.44\sim 9.78$ 、 $5.42\sim 35.43$ 、 $1.82\sim 29.93$ 、 $2.36\sim 18.25$ 、 $5.82\sim 44.94\text{ m}^3/\text{h}$,平均流量分别为 3.98 、 7.82 、 6.26 、 6.12 、 4.17 、 15.61 、 11.58 、 8.05 、 $18.44\text{ m}^3/\text{h}$ 。

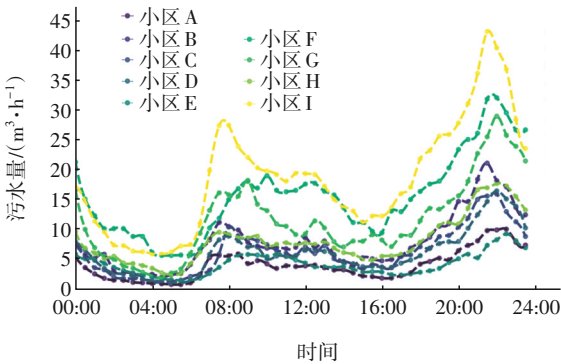


图2 样本小区污水量随时间变化规律
Fig.2 Changes in sewage volume of sample communities with time

1.3 研究方法

生活小区污水量预测方法构建涉及数据集构建、预测算法确定、变量筛选、数据降噪、数据特征提取、超参数优化等环节。

1.3.1 数据集构建

按照7:2的比例将污水量数据划分为训练集和测试集,样本数分别为672和192组。

1.3.2 预测算法

采用XGBoost算法对污水量进行预测。建筑特征信息变量多,而且变量之间存在交互作用和非线性关系,而机器学习中的XGBoost算法具有较强的非线性拟合能力,是一种基于梯度提升框架的机器学习算法,广泛应用于回归问题分析中^[11]。其核心思想是通过逐步构建并加权集成多个弱分类器(通常为决策树)^[8],以最小化目标损失函数,从而提升模型的整体预测性能。因此相比于传统预测方法,XGBoost算法与本研究的多变量非线性拟合需求相符,但其也存在对数据准确性、数据与因变量相关性要求较高的问题,因此需要对数据进行预处理,常用的预处理方法有变量筛选、数据降噪和数据特征提取。

1.3.3 变量筛选方法

科学合理地选择与污水量高度相关的变量进

行建模,是确保预测准确性的基础。本研究采用皮尔森相关系数对建筑特征信息和污水量之间的相关性进行量化。皮尔森相关系数是衡量两个连续变量之间线性相关程度的统计量,其数值范围为 $[-1, 1]$,其中值为1表示完全正线性相关,值为-1表示完全负线性相关,值为0表示无线性相关关系。

1.3.4 数据降噪方法

采用了高斯平滑对数据进行降噪,平滑程度参数 σ 在后文进行优化。污水管网环境恶劣,导致流量监测数据存在异常数据噪声,从而影响预测效果。数据去噪算法能够识别并优化数据中存在的随机波动,使得数据更加平滑,减小误差^[12]。高斯平滑是一种常用的去噪和平滑技术,广泛应用于图像处理、信号处理、时间序列分析等领域^[13]。它基于高斯分布对数据进行加权平均,从而平滑局部波动、去除噪声,同时尽量保留整体趋势。

1.3.5 数据特征提取方法

利用径向基函数变换将时间变量转换为多个径向基函数,从而凸显出时间数据的循环特征,提高模型预测精度。径向基函数变换是一种常用于机器学习特征工程中的数据变换技术,其能够将原始数据转化为能更代表预测目标特征的数据,从而提高模型表现。其核心思想是通过生成多个数据中心点,利用重复基函数捕捉数据与中心点的接近程度,反映数据变化规律和循环特点^[14]。径向基函数变换的参数中, input_range 为数据变化范围,本研究设置预测时间间隔为 2 h,预测总周期为 24 h,因此设置 input_range 为 (0, 24); n_periods 参数为要创建的基函数数量,不同取值对预测效果有较大影响,因此将该值作为超参数在后文进行优化。

1.3.6 超参数优化方法

超参数优化是指在训练机器学习模型时,搜索和选择最优超参数组合的过程。超参数与模型的训练过程直接相关,合理的超参数选择可以显著提高模型的性能、精确度和泛化能力,因此需要对超参数进行优化。由于 XGBoost 模型可优化的超参数较多,因此采用逐个参数单独优化以降低复杂度。调参目标函数设置为均方误差 (MSE),对于每组超参数,利用 K 折交叉验证对训练集进行 K 次划分,划分结果为 $K-1$ 份新的训练集和 1 份验证集,计算 K 次划分验证集均方误差的平均值,作为该组超参数的目标函数并对其进行最小化,本研究中将折数 K

设置为 5。

2 结果与讨论

2.1 生活小区污水量变化的主要影响因素

实现生活小区污水量的有效预测,首要步骤是明确影响生活小区污水量变化的主要因素。收集多个与生活小区污水量变化存在联系的建筑特征信息变量,解析它们与污水量的联系。其中,户数和建筑面积反映了小区的居住人数,从而影响小区排水量;占地面积反映小区的规模,从而间接反映小区的人均用水量;楼栋数越多,使用人数可能越多,产生的污水量可能越大;竣工时间反映了建筑的设备老化程度,从而反映小区污水的漏损率;房价越高,居住人群的消费水平越高,用水量及污水量可能越大;气温越高,小区用水量及污水量可能越大;最后,人类排水活动具有明显的时间规律性。可见,这些相关信息均在物理意义上与污水量存在联系。计算它们与生活小区污水量之间的相关性,并绘制相关系数热图,结果如图 3 所示。

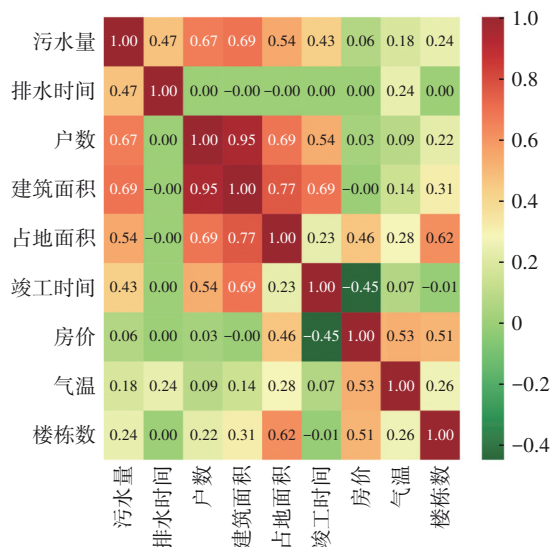


图3 污水量影响因素相关系数热图

Fig.3 Heat map of correlation coefficients of factors affecting sewage volume

从图 3 可以明显看出,污水量与户数和建筑面积有较强相关性 (相关系数 > 0.60),表明这两类物理属性在一定程度上能够反映污水量排放水平。此外,污水量与排水时间、占地面积及竣工时间的相关性处于中等水平 (相关系数 > 0.40),但与房价、气温、楼栋数等变量的相关性较弱 (相关系数 < 0.25),表明其影响较小。因此保留相关性中等及

以上的变量,利用户数、建筑面积、排水时间、占地面积、竣工时间对污水量进行预测。

2.2 生活小区污水量预测模型参数优化

2.2.1 降噪强度参数优化

采用不同平滑强度的高斯平滑对所有污水量监测数据进行处理,随着平滑强度参数 σ 值的提高,污水量数据曲线逐渐变平滑,说明高斯平滑有效减少了随机误差带来的数据波动,当 σ 达到 1.5 时,平滑后的曲线基本平滑且保留了原有的变化趋势,无剧烈抖动,继续提高 σ 值对曲线噪声去除无明显效果,因此最终选择 $\sigma=1.5$ 时的降噪强度对污水量数据进行降噪。

2.2.2 模型超参数优化

对径向基函数变换中的基函数数量和 XGBoost 模型中的 5 项超参数进行优化,以 MSE 作为目标函数,尝试不同超参数组合使得 MSE 最小,优化过程见图 4。通过逐个超参数的调整, MSE 降至 10.78,说明模型已经达到了较高的精度,优化结果为基函数数量=24,超参数 $n_estimators=101$, $max_depth=2$, $learning_rate=0.08$, $colsample_bytree=0.22$, $subsample=0.14$ 。

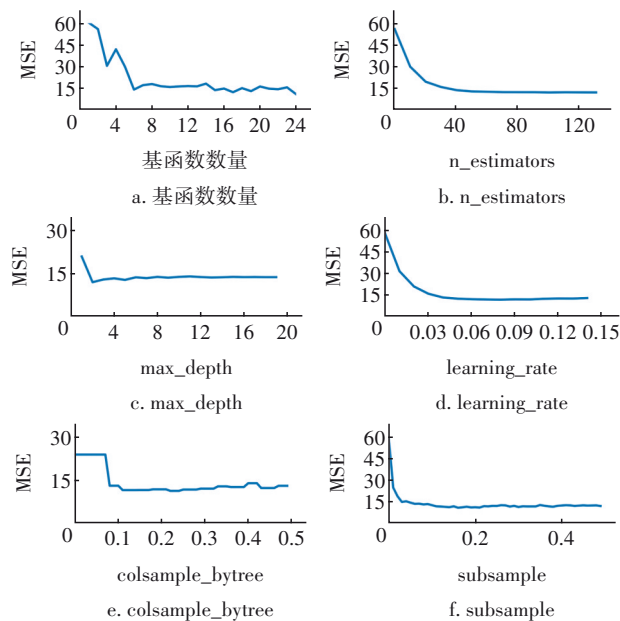


图 4 模型超参数优化过程

Fig.4 Model hyperparameter optimization process

2.3 生活小区污水量预测模型效果分析

计算 XGBoost 模型预测结果,将训练集和测试集的所有样本的特征值代入建立的预测模型中,得到预测值并与真实值进行比较,结果见图 5。可见,

XGBoost 模型预测效果较好,置信区间和预测区间窄,预测精度高,线性拟合线与正比例函数基本重合,预测准确度好,测试集决定系数 R^2 和平均相对误差 (MRE) 分别为 0.92、14%。

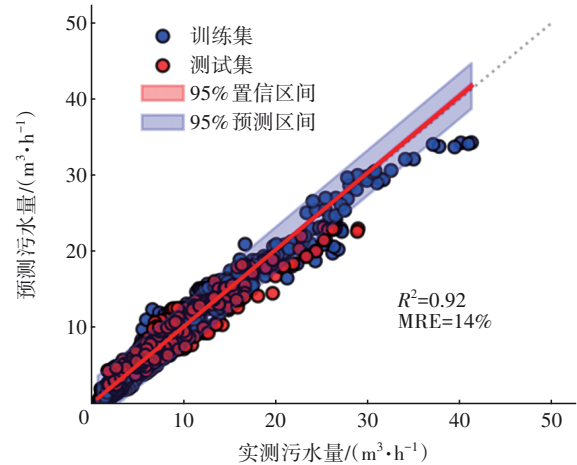
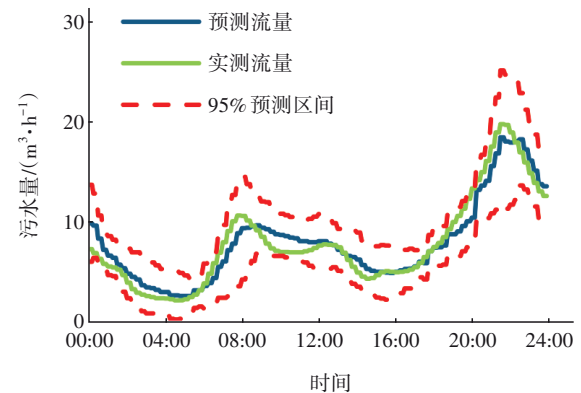


图 5 XGBoost 模型预测值和实测污水量对比

Fig.5 Comparison of measured and predicted sewage volume using XGBoost model

2.4 生活小区污水量预测模型优化

构建生活小区污水量预测区间能够在污水量预测中考虑预测误差,从而对预测结果的不确定性进行衡量。设某小区的瞬时污水量 y 服从均值为 μ 和标准差为 σ 的正态分布,则其 95% 预测区间即为 $[\mu \pm 2\sigma]$ 。由于总体均值 μ 和标准差 σ 无法获取,因此利用模型预测污水量 $\hat{y}$ 来近似表示均值 μ ,利用预测误差(即模型残差)的标准差 s_e 来替代标准差 σ ,构建 95% 预测区间 $[\hat{y} \pm 2s_e]$ 。选择测试集中的两个小区分别计算其 95% 预测区间,绘制实测流量、预测流量和预测区间,结果见图 6。可以看出,实测流量基本位于预测区间内,说明建立的污水量预测区间有效。



a. 小区 B

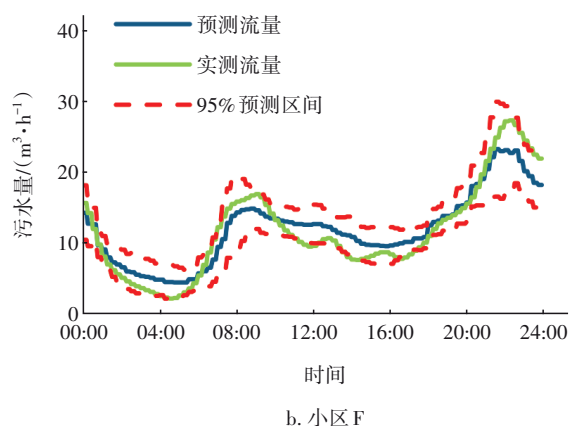


图6 生活小区污水量95%预测区间

Fig.6 95% prediction interval of sewage volume in residential communities

分析所有生活小区在不同时间的污水量预测相对误差,结果表明08:00—24:00的污水量预测相对误差较小,平均误差为12.6%,变化范围为8%~20%(见图7)。

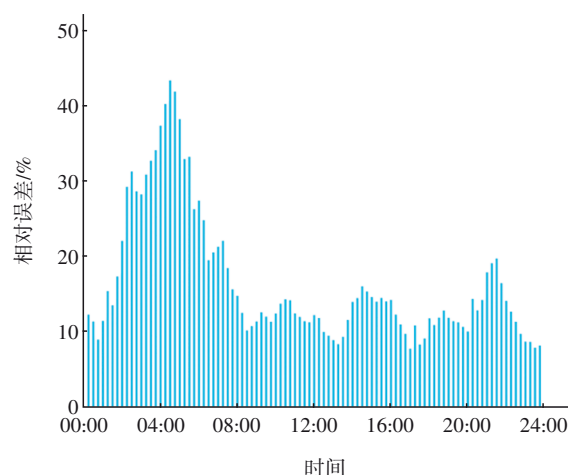


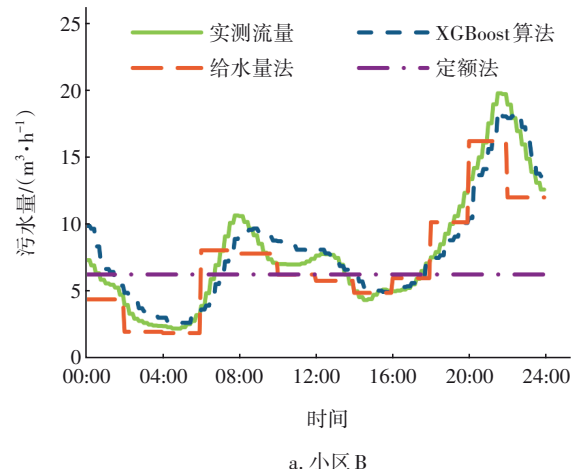
图7 生活小区污水量预测相对误差

Fig.7 Relative error of sewage volume prediction in residential communities

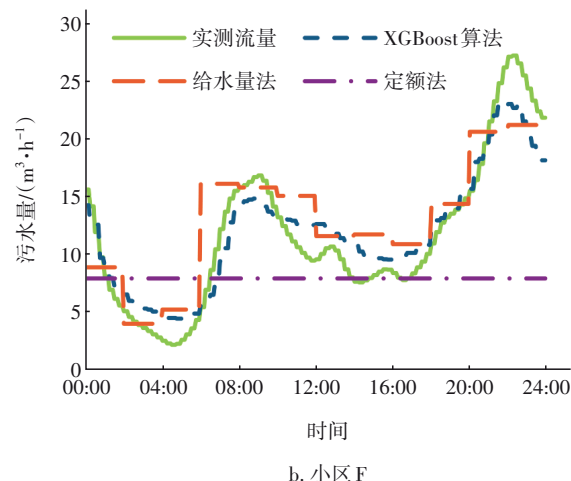
2.5 不同生活小区污水量预测方法对比

分别采用XGBoost算法、定额法和给水量法预测测试集中B和F两个生活小区的污水量,计算其在08:00—24:00的平均相对误差,结果见图8。其中XGBoost算法的预测效果最好,预测流量变化平滑,与实测流量拟合程度较高,小区B和F的平均相对误差分别为9%和13%;其次是给水量法,其基本反映了实测流量变化规律,但精度不够高,平均相对误差分别为15%和19%;定额法预测效果最差,预测结果为固定值,无法反映真实污水量逐时的变

化规律,平均相对误差分别达到40%和44%。



a. 小区 B



b. 小区 F

图8 不同方法预测生活小区B、F污水量结果

Fig.8 Results of different methods for predicting sewage volume in residential communities

3 结论

① 构建了一种基于XGBoost算法的建筑特征信息预测生活小区污水量方法,生活小区污水量的主要影响因素为户数、建筑面积、占地面积、竣工时间和排水时间,因此优选上述因素为污水量预测模型的输入参数。

② 构建的污水量预测模型在08:00—24:00的预测误差较小,其测试集平均相对误差为11%,优于传统定额法和给水量法的42%和17%。

参考文献:

- [1] 邢玉坤,曹秀芹,柳婷,等.我国城市排水系统现状、问题与发展建议[J].中国给水排水,2020,36(10):19-23.
XING Yukun, CAO Xiuqin, LIU Ting, et al. Current

- status, problems and development suggestions of urban drainage system in China [J]. *China Water & Wastewater*, 2020, 36(10): 19–23 (in Chinese).
- [2] 罗坤,孙凌凯,马方凯,等. 基于水质水量监测的污水管网评估诊断与定量分析[J]. *中国给水排水*, 2024, 40(13): 109–114.
- LUO Kun, SUN Lingkai, MA Fangkai, *et al.* Evaluation, diagnosis and quantitative analysis of sewage pipe network based on water quality and quantity monitoring [J]. *China Water & Wastewater*, 2024, 40(13): 109–114 (in Chinese).
- [3] ZHAO Z, YIN H, XU Z, *et al.* Pin-pointing groundwater infiltration into urban sewers using chemical tracer in conjunction with physically based optimization model [J]. *Water Research*, 2020, 175: 115689.
- [4] 徐祖信,王思玉,刘淑雅,等. 基于地下水入渗反演解析的污水管网破损数值化定位[J]. *同济大学学报(自然科学版)*, 2022, 50(9): 1331–1338.
- XU Zuxin, WANG Siyu, LIU Shuya, *et al.* Numerical analysis and location of sewer network damage based on groundwater infiltration inversion [J]. *Journal of Tongji University (Natural Science)*, 2022, 50(9): 1331–1338 (in Chinese).
- [5] 胡鸿昊. 基于耦合动态模拟方法的污水管网入流渗入识别及定位[D]. 杭州:浙江大学, 2022: 13–15.
- HU Honghao. Identification and Pinpointing of Rain Derived Inflow and Infiltration Using Coupled Dynamic Simulation Method [D]. Hangzhou: Zhejiang University, 2022: 13–15 (in Chinese).
- [6] JIA Y, ZHENG F, ZHANG Q, *et al.* Foul sewer model development using geotagged information and smart water meter data [J]. *Water Research*, 2021, 204: 117594.
- [7] 何韶渺,武玉玲,马一祎,等. 基于POI流量估算与ES算法校核的污水管网模拟研究[J]. *给水排水*, 2024, 50(12): 139–145.
- HE Shaomiao, WU Yuling, MA Yiwei, *et al.* Simulation study of sewage pipe network based on POI flow estimation and ES algorithm calibration [J]. *Water & Wastewater Engineering*, 2024, 50(12): 139–145 (in Chinese).
- [8] 李媛媛,沈芳,陈嵩钰,等. 基于XGBoost的内陆河湖浊度反演与长时序分析[J]. *环境工程学报*, 2024, 18(2): 398–408.
- LI Yuanyuan, SHEN Fang, CHEN Songyu, *et al.* Inversion and long-term series analysis of turbidity in inland rivers and lakes based on XGBoost [J]. *Chinese Journal of Environmental Engineering*, 2024, 18(2): 398–408 (in Chinese).
- [9] 牛景辉. 基于GWO-XGBoost的工业污水水质关键数据预测算法[J]. *工业水处理*, 2024, 44(1): 184–190.
- NIU Jinghui. GWO-XGBoost prediction algorithm for industrial wastewater quality key data [J]. *Industrial Water Treatment*, 2024, 44(1): 184–190 (in Chinese).
- [10] 田欢,王文海,李芸,等. 雨水明渠流量测量常用设备及其选型[J]. *环境与可持续发展*, 2016, 41(2): 60–63.
- TIAN Huan, WANG Wenhai, LI Yun, *et al.* Research on usual equipments and the selection of rainwater flow measurement on open channel [J]. *Environment and Sustainable Development*, 2016, 41(2): 60–63 (in Chinese).
- [11] NIAZKAR M, MENAPACE A, BRENTAN B, *et al.* Applications of XGBoost in water resources engineering: a systematic literature review (Dec 2018–May 2023) [J]. *Environmental Modelling & Software*, 2024, 174: 105971.
- [12] 关思源,张巧珍. 基于滤波重构时间序列回归算法的需水量预测[J]. *中国给水排水*, 2025, 41(7): 63–68.
- GUAN Siyuan, ZHANG Qiaozhen. Water demand prediction utilizing filtering reconstruction time series regression algorithm [J]. *China Water & Wastewater*, 2025, 41(7): 63–68 (in Chinese).
- [13] WINK A M, ROERDINK J B. Denoising functional MR images: a comparison of wavelet denoising and Gaussian smoothing [J]. *IEEE Transactions on Medical Imaging*, 2004, 23(3): 374–387.
- [14] GULLU M. Coordinate transformation by radial basis function neural network [J]. *Scientific Research and Essays*, 2010, 5(20): 3141–3146.
-
- 作者简介:**陈成(2000–),男,重庆人,硕士研究生,主要研究方向为污水管网外水诊断。
E-mail: 1719559753@qq.com
收稿日期: 2025-06-08
修回日期: 2025-06-23

(编辑:李德强)