

DOI:10.19853/j.zgjsps.1000-4602.2024.03.008

# 供水管网流量监测数据异常值检测方法对比分析

胡诗苑<sup>1</sup>, 高金良<sup>1</sup>, 钟丹<sup>1</sup>, 武睿<sup>2</sup>, 刘路明<sup>3</sup>

(1. 哈尔滨工业大学 环境学院, 黑龙江 哈尔滨 150090; 2. 广东粤海水务投资有限公司, 广东 深圳 518021; 3. 哈尔滨工业大学水资源国家工程研究中心有限公司, 黑龙江 哈尔滨 150090)

**摘要:** 随着信息化技术的发展,水务企业迎来了智慧化转型升级。数据采集与预处理作为水务企业实现智慧管理的重要前序步骤,为后续数据挖掘、运营管理、调度决策提供了基础。由于环境的影响、管网中的随机扰动、管网事故等原因,监测数据的质量问题广泛存在,因此寻求有效的供水管网流量监测数据的异常值检测方法至关重要。基于此,首先根据供水管网流量监测数据的基本特征和时间维度的相关性,将常见异常归纳为3个类型;其次,以东南沿海某市的真实小区流量监测数据为例,分别探究基于统计、密度和预测的Boxplot、LOF与Prophet异常值检测模型在不同类型异常数据检测中的性能。结果表明,Boxplot与LOF模型能够较准确地识别出异常数据,但Boxplot对异常的判断标准较宽泛,容易将部分非异常数据识别为异常点,Prophet对于不稳定性较高的流量数据识别效果有限。

**关键词:** 流量监测数据; 异常值检测; Boxplot; LOF; Prophet

**中图分类号:** TU991 **文献标识码:** A **文章编号:** 1000-4602(2024)03-0053-07

## Comparison of Methods for Flow Monitoring Data Outlier Detection in Water Distribution Network

HU Shi-yuan<sup>1</sup>, GAO Jin-liang<sup>1</sup>, ZHONG Dan<sup>1</sup>, WU Rui<sup>2</sup>, LIU Lu-ming<sup>3</sup>

(1. School of Environment, Harbin Institute of Technology, Harbin 150090, China; 2. Guangdong Yuehai Water Investment Co. Ltd., Shenzhen 518021, China; 3. National Engineering Research Center of Urban Water Resources Co. Ltd., Harbin Institute of Technology, Harbin 150090, China)

**Abstract:** With the development of information technology, water enterprises are undergoing intelligent transformation and upgrading. Data collection and preprocessing is an important pre-step for water enterprises to realize intelligent management, and provides a foundation for subsequent data mining, operation management and scheduling decision. Due to the reasons such as environmental factors, random disturbance in the pipe network and pipe network accident, monitoring data quality issues exist widely, making it is very important to find an effective method for flow monitoring data outlier detection in water distribution network. The common anomalies were firstly classified into three categories according to the basic characteristics and temporal correlation of flow monitoring data in water distribution network. Then,

基金项目: 国家重点研发计划项目(2022YFC3203800); 国家自然科学基金资助项目(51978203); 黑龙江省重点研发计划项目(2022ZX01A06); 揭榜制科研项目(CE602022000203)

通信作者: 高金良 E-mail: gjl@hit.edu.cn

the performance of Boxplot, LOF and Prophet outlier detection models based on statistics, density and prediction in the detection of different types of real flow monitoring data outliers was explored in a southeast coastal city of China. Boxplot and LOF models identified outliers more accurately. However, Boxplot had broad criteria for outlier identification, and it was easy to identify some non-abnormal data as outliers. Prophet had limited effectiveness in identifying unstable flow data.

**Key words:** flow monitoring data; outlier detection; Boxplot; LOF; Prophet

随着通信、新技术和算力等新基建信息基础设施的发展,水务行业迎来了自动化到信息化乃至智慧化的转型升级。通过物联网、互联网、大数据等技术手段对供水管网进行实时监控,并对海量的流量、水压、水质、水基础设施信息数据进行采集、预处理、挖掘与分析,为供水管网实现智慧化运营管理、调度决策提供了依据<sup>[1]</sup>。数据采集与预处理作为智慧水务全过程管理的前序步骤,决定了后续数据挖掘分析和水务企业智慧化管理的质量<sup>[2]</sup>。

目前以SCADA为代表的供水管网实时监测和远程控制系统的建设已经取得了一定的进展,但由于环境的影响、管网中的随机扰动和监测仪器发生机械故障等原因,导致监测数据的质量问题广泛存在<sup>[3]</sup>。具有异常值的数据限制了后续需水量预测模型、调度模型、事故预警模型等智慧化管理模型的准确性,由此产生错误的需水量预测结果、不合理的调度建议、误报或漏报管网中的爆管事故等<sup>[4]</sup>。因此对搜集到的数据采用异常值检测等预处理手段来提高数据质量,是水务企业实现智慧化发展的重要基础保障工作<sup>[5]</sup>。刘书明等<sup>[6]</sup>根据时段和季节将供水管网流量监测数据进行切分,构建了自回归滑动平均模型,并对人工模拟含异常序列中的异常值进行识别。黄国东等<sup>[2]</sup>使用基于支持向量机的供水管网数据清洗方法,能够较好地修复数据错误。现有研究进行异常检测时对训练集要求较高,然而水务公司获取的供水管网流量监测数据通常已包含一定的异常值,难以满足训练集的高质量要求。为了水务公司在实际操作中的便利性和可行性,笔者探究了不同异常值检测方法在包含噪声和异常值的原始数据中的性能表现。具体来说,首先分析了供水管网流量监测数据产生异常的原因,并按数据分布特性与时间维度的相关性将供水管网流量监测数据分为3种类型,随后分别采用基于统计、密度和预测的异常数据检测方法对我国东南沿海某

市的真实小区流量监测数据进行异常值识别,探究不同异常值识别技术在三种不同异常类型检测中的优缺点,旨在为水司在实际工程中进行数据清洗提供依据。

### 1 供水管网异常监测数据产生原因及类型

供水管网异常监测数据与数据采集及传输设备异常、用水行为的随机扰动和供水管网事故相关。数据采集设备与传输设备可能会受到周围环境的干扰,如路面振动、供电不稳定及电磁波干扰等产生并传输异常数据,不仅如此,设备本身的机械故障也可能产生异常数据<sup>[7]</sup>。而用户用水具有复杂性和随机性,会对供水管网的流量等参数造成影响,尤其是靠近用水端的管段,当随机扰动超出正常范围时,则产生异常。除以上两类管网状态处于正常状态时产生的异常数据,供水管网中的异常事故也会产生异常监测数据。此类数据虽为供水管网的真实监测值,但其与正常运营状态数据具有一定差距,预示着供水管网中爆管等事故的发生,该类数据在异常事件预警时为有效数据<sup>[8]</sup>,但在需水量预测或供水管网日常运营调度模型训练中会产生模型误差。

根据Chandola等<sup>[9]</sup>关于异常检测相关研究的概述,本研究结合数据分布的基本特征与供水管网流量监测数据时间维度的相关性,将流量监测异常数据分为监测数据正常波动范围外的异常和不符合时间序列变化规律的点与子序列异常,即点异常、上下文异常与集合异常,如图1所示。点异常:点异常是与正常数据相差较大、在监测数据正常波动范围之外的异常点,是最易于发现的异常类型,如图1(a)所示。上下文异常:上下文异常点通常处于监测数据的正常波动范围内,仅观测该点与正常数据波动范围很难发现其异常,但将该点与上下文数据联系起来,可发现其存在破坏了监测数据在时间上的连贯性,不符合监测数据的上下文波动规律,但由

于其易淹没在正常数据中,较难被发现,如图1(b)所示。集合异常:集合异常是指流量监测数据时间序列的某段子序列不符合流量数据随时间变化的规律,该子序列中的某些点可能是正常数据,但将其作为子序列整体出现时,则为异常集合,如图1(c)所示。

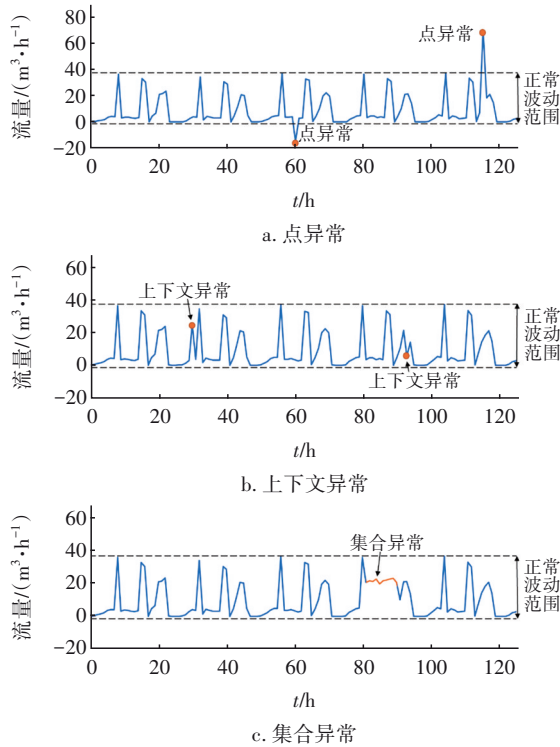


图1 流量监测数据异常类型

Fig.1 Outlier types of flow monitoring data

## 2 供水管网异常数据检测方法

### 2.1 基于统计的异常数据检测方法

基于统计的异常数据检测方法根据数据统计特征推断数据的合理分布范围,以此判别异常数据。基于统计的异常数据检测方法主要包括3sigma准则、Z分数、Boxplot等。3sigma准则与Z分数等方法以数据服从正态分布为前提,而实际流量监测数据通常无法严格服从正态分布。Boxplot根据数据实际分布情况确定异常值范围,无需假定数据分布形式,且对异常值具有耐抗性,其上下四分位数以外的数据(包含异常数据)不会对其异常值判别标准产生扰动,呈现较客观的异常值判别结果<sup>[10]</sup>。

故本研究以Boxplot为代表,分析基于统计的异常数据检测方法的性能。箱型图通常包含5个基本

数据特征,即最小值、下四分位数( $Q_1$ )、中位数(Median)、上四分位数( $Q_3$ )、最大值。Boxplot异常值检测方法以 $Q_3$ 和 $Q_1$ 为基础,计算正常数据的上限( $Q_{upper}$ )与下限( $Q_{lower}$ ),见式(1)~(3)。超出范围( $Q_{lower}, Q_{upper}$ )的离群值则被认为是异常流量监测数据,如图2所示。

$$Q_{upper} = Q_3 + 1.5 \times IQR \quad (1)$$

$$Q_{lower} = Q_1 - 1.5 \times IQR \quad (2)$$

$$IQR = Q_3 - Q_1 \quad (3)$$

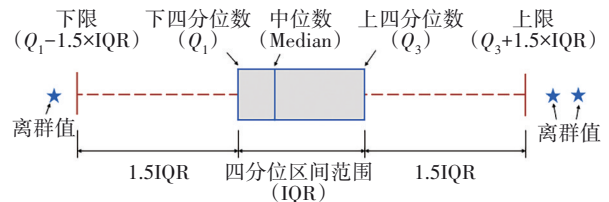


图2 Boxplot异常值检测原理

Fig.2 Principle of Boxplot outlier detection

### 2.2 基于密度的异常数据检测方法

由于异常值通常具有远离正常数据的趋势,且占比较低。根据此特征,基于密度的异常数据检测方法通过将数据划分为不同的簇,并计算簇的密度来判断异常值。局部异常因子(LOF)为典型的基于密度的异常值检测方法,其基本原理为<sup>[11]</sup>:①根据所有监测数据的第 $k$ 距离邻域 $N_k(P)$ ,计算其第 $k$ 距离 $d_k(P)$ ,其中 $N_k(P)$ 为点 $P$ 第 $k$ 距离内所有的点, $d_k(P)$ 为点 $P$ 到第 $k$ 邻近点的距离;②计算各点的局部可达密度 $\rho_k(P)$ ,见式(4);③计算各点的局部离群因子,见式(5)。

$$\rho_k(P) = \left[ \frac{\sum_{O \in N_k(P)} d_k(P, O)}{|N_k(P)|} \right]^{-1} \quad (4)$$

$$LOF_k(P) = \frac{\sum_{O \in N_k(P)} \rho_k(O)}{|N_k(P)| \rho_k(P)} \quad (5)$$

式中: $d_k(P, O)$ 为点 $P$ 到点 $O$ 的可达距离,是 $d_k(O)$ 与点 $P$ 到点 $O$ 实际距离中的最大值。

局部离群因子 $LOF_k(P)$ 代表了点 $P$ 附近的第 $k$ 距离邻域 $N_k(P)$ 内所有点的局部可达密度与点 $P$ 自身局部可达密度的比值。 $LOF_k(P)$ 越大,则说明点 $P$ 自身局部可达密度小于邻域点的密度,点 $P$ 越有可能为异常值;而 $LOF_k(P)$ 越接近1,则说明点 $P$ 自身



局部可达密度与邻域点的密度相当,为同类簇,点 $P$ 为异常值的可能性较低。

### 2.3 基于预测的异常数据检测方法

基于预测的异常值检测方法包括深度学习生成模型(生成对抗网络、变分自编码器等)和基于回归的模型(ARIMA、Prophet等)。深度学习生成模型为异常值检测领域的前沿方法,在工业、医疗等领域已有相关研究<sup>[12]</sup>。深度学习生成模型需对正常数据进行训练后,根据预测值与监测数据的差异判别异常数据,否则可能会产生较大误差。但供水管网中获取的流量监测数据通常已包含一定异常数据,需要先对数据进行人工处理获得正常值后再进行训练,因此深度学习生成模型更适用于供水管网新出现的异常状态预警,而不适用于预处理环节。本研究通过Prophet对供水管网流量时间序列进行回归,将流量监测数据视为时间的函数,回归提取数据主要信息后,通过回归值与实际值的差值来判别异常数据,其原理见式(6)<sup>[13]</sup>。

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t \quad (6)$$

式中: $g(t)$ 为趋势项,包括线型模型和logistic回归模型; $s(t)$ 为季节项,通过傅里叶展开来逼近; $h(t)$

为节假日等外部项因素; $\varepsilon_t$ 为误差项。

## 3 实际案例应用及讨论

### 3.1 案例流量数据简介

为探究不同方法在实际供水管网小区流量监测数据异常值识别中的性能,以我国东南沿海某市的小区入口流量监测点数据为案例数据,取3组监测点,分别编号为flowno1、flowno2、flowno3。选取的3组监测点所对应的小区具有不同日需求模式,flowno1与flowno2的日需求模式具有3个峰且均为尖峰;flowno3的日需求模式具有2个峰,流量变化相对较平缓。不同监测点流量数据分布的密集程度也具有明显的差别,分布从稀疏到密集排序依次为flowno1、flowno2、flowno3。不仅如此,3组案例监测点数据均包含了不同类型的真实异常数据。选取的案例监测点能反映不同异常值检测方法在不同分布的监测数据上针对不同异常类型的检测性能,具有一定的普适性。flowno1实验数据集为2016年4月23日—6月24日的流量数据,flowno2实验数据集为2016年9月1日—11月2日的流量数据,flowno3实验数据集为2016年5月14日—7月15日的流量数据,数据基本特征见表1。

表1 案例监测点流量数据基本特征

Tab.1 Basic characteristics of flow data of case monitoring points

项目	平均流量/ ( $\text{m}^3 \cdot \text{h}^{-1}$ )	最小流量/ ( $\text{m}^3 \cdot \text{h}^{-1}$ )	四分之一分位数/ ( $\text{m}^3 \cdot \text{h}^{-1}$ )	四分之三分位数/ ( $\text{m}^3 \cdot \text{h}^{-1}$ )	最大流量/ ( $\text{m}^3 \cdot \text{h}^{-1}$ )	方差/ ( $\text{m}^3 \cdot \text{h}^{-1}$ ) <sup>2</sup>
flowno1	13.22	0.00	4.26	20.64	123.84	14.28
flowno2	7.12	0.00	1.32	6.87	136.69	10.32
flowno3	9.08	0.00	5.87	12.14	60.82	4.74

### 3.2 模型构建

由于上下文异常通常位于正常监测数据波动范围内,Boxplot与LOF根据数据分布进行异常值识别,若使用流量监测数据总集进行建模,可能导致上下文异常识别困难,故将流量监测数据按小时分为24类后分别进行处理。Boxplot无需对参数进行调节,LOF中flowno1、flowno2、flowno3的 $n\_neighbors$ 分别选择50、10、50,contamination在(0, 0.05)范围内进行微调。Prophet直接对整个时间序列进行拟合,daily\_seasonality与weekly\_seasonality均选择True,3组监测数据的置信区间分别为0.99、0.95、0.99。

### 3.3 结果与讨论

不同监测点流量监测数据分别使用Boxplot、

LOF与Prophet方法进行异常值识别以后,按一天中不同小时绘制散点图,结果如图3所示。可以看出,Boxplot、LOF、Prophet模型对监测数据正常波动范围之外的点异常均能够进行有效识别。Boxplot与LOF能够识别在监测数据正常波动范围之内的上下文异常,但是Boxplot的识别效果依赖于数据分布情况,对于流量数据分布较广的监测点,例如flowno1与flowno2,虽然能够识别上下文异常,但是也将相当一部分正常数据判别为异常值;而流量数据分布较密集的监测点,例如flowno3,Boxplot能够获得很好的识别效果。LOF可以对异常值的比例进行调节,能够获得较好的异常值识别效果,而且尽可能少地将正常值判断为异常值,以保留更多的监测数据信息。

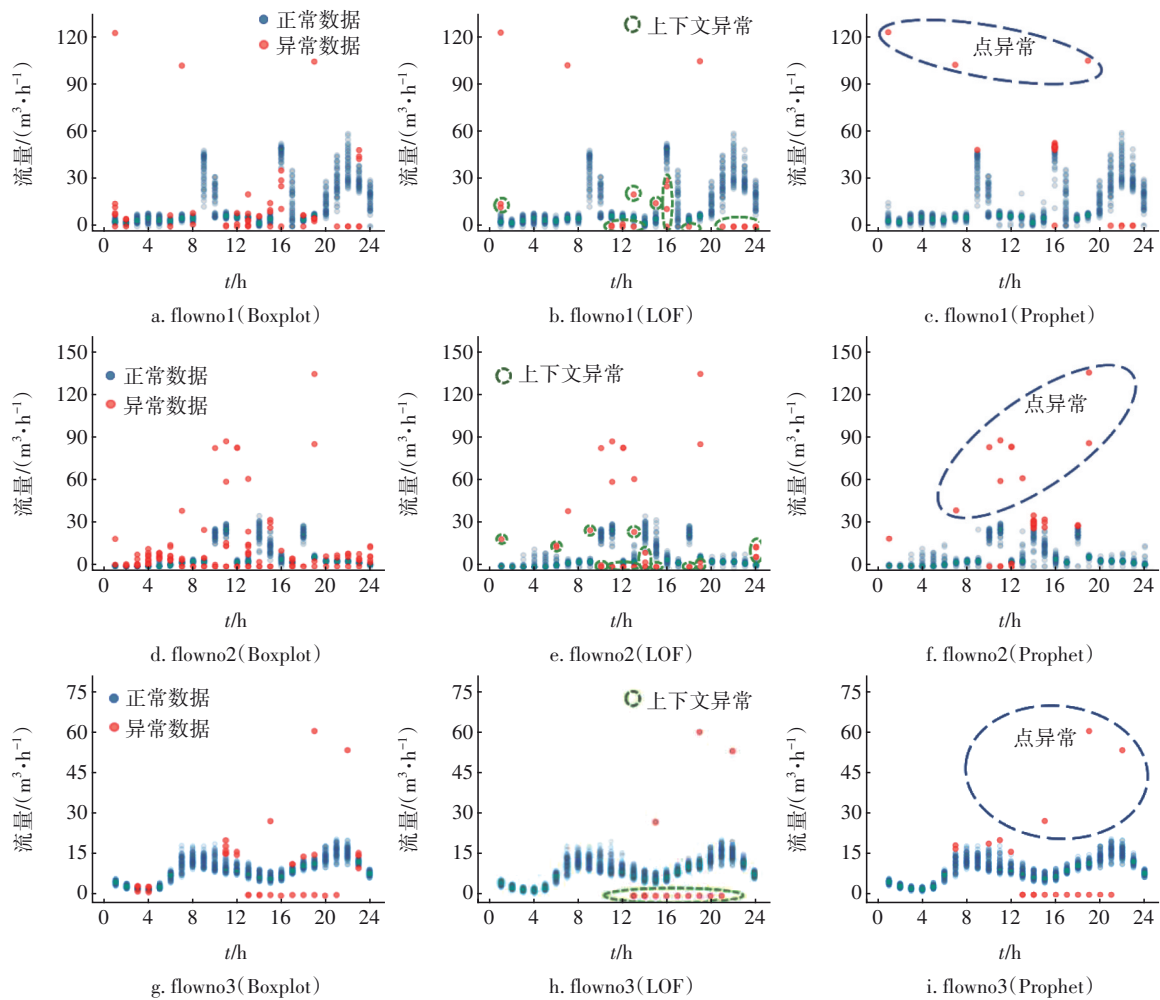
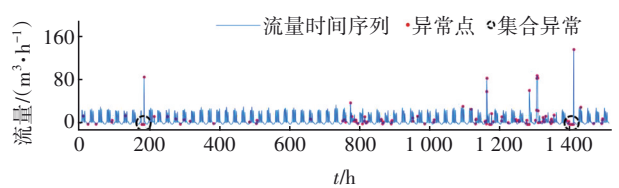
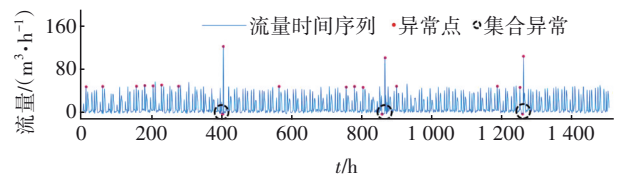
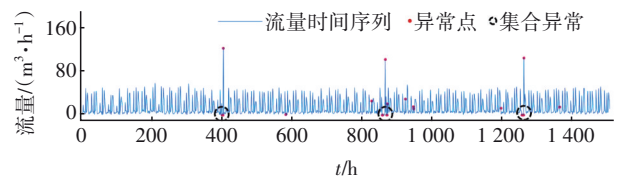
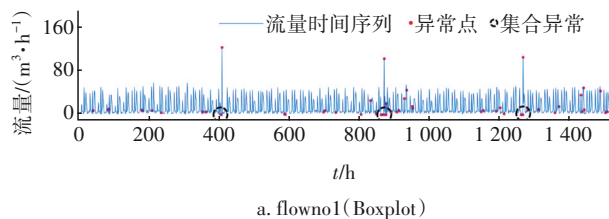


图3 异常值识别结果

Fig.3 Results of outlier detection

Prophet对淹没在正常数据波动范围内的上下文异常识别效果较差,这是由于流量数据具有较强不确定性,影响了预测模型的性能。对于flowno3,即稳定性较强的监测点,其能获得良好的性能;而对于不确定性较大的flowno1与flowno2,Prophet易将高峰时段的一些高流量值误识别为异常值,影响数据质量。

为了探究不同异常值检测方法对于集合异常的识别效果,绘制不同监测点流量时间序列,如图4所示。



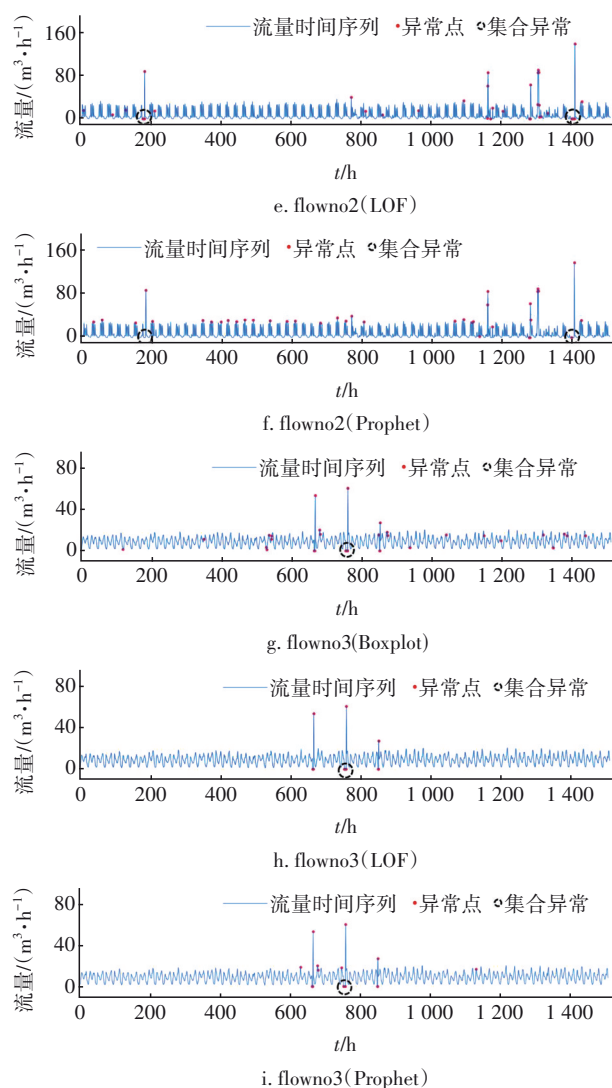


图4 flowno1、flowno2和flowno3集合异常识别结果

Fig.4 Results of collective outlier detection of flowno1, flowno2 and flowno3

从图4可以看出,Boxplot对不同流量监测点的集合异常均能很好地进行识别,但对流量数据分布较分散的flowno2,其产生的异常值比例过高,可能存在误识别,损失了真实流量数据。而LOF虽然能识别出所有的集合异常子序列,但由于集合异常子序列中个别点与其所在小时的正常数据分布相差不大,基于密度的LOF无法根据时间序列的规律将其识别为异常点,如图4(b)所示。Prophet对于具有较高不确定性、不稳定性的流量时间序列拟合较差,无法识别出所有的集合异常,如图4(f)所示,且易将流量数据正常波动范围边界上的点误识别为异常点;Prophet对于稳定性较强的flowno3识别效果较好。根据以上不同模型进行流量监测数据异

常检测的结果,汇总不同模型的优缺点。Boxplot模型无需调参,易于操作,根据数据分布可直接获得可能的异常点;但其易将部分非异常数据识别为异常点。LOF的效果较好,对于点异常、上下文异常均能进行较好地识别;但其需要根据数据分布情况调参,无法捕捉时间序列数据时间维度的相关性。Prophet考虑了时间序列数据的特点,但其效果一般,异常值检测效果受数据稳定性影响大。

#### 4 结论

供水管网流量监测数据受到环境等因素影响存在大量异常值,对后续需水量预测模型、调度模型等产生影响。本研究对异常数据产生的原因进行了分析,并将常见异常归纳为3种类型,随后采用基于统计、密度和预测的Boxplot、LOF与Prophet模型对我国东南沿海某市3个小区流量监测点实测数据进行异常值检测,探究不同模型对不同异常类型的检测性能。结果表明,Boxplot模型虽然易于操作,但其易将部分非异常数据识别为异常点;LOF效果最好,能有效识别点异常与上下文异常;Prophet模型受预测模型准确性影响较大,对于具有不稳定特性的流量监测数据,其识别效果受限。

#### 参考文献:

- [1] OBERASCHER M, RAUCH W, SITZENFREI R. Towards a smart water city: a comprehensive review of applications, data requirements, and communication technologies for integrated management[J]. Sustainable Cities and Society, 2022, 76: 103442.
- [2] 黄国东, 龙志宏, 朱子朋, 等. 基于支持向量机的供水管网监测数据清洗[J]. 给水排水, 2022, 48(9): 124-129.  
HUANG Guodong, LONG Zhihong, ZHU Zipeng, et al. Monitoring data cleaning for water distribution system based on support vector machine [J]. Water & Wastewater Engineering, 2022, 48(9): 124-129 (in Chinese).
- [3] 黄国东. 基于预测的供水管网异常监测数据清洗研究[D]. 杭州: 浙江大学, 2022.  
HUANG Guodong. Research on Abnormal Monitoring Data Cleaning of Water Supply Network Based on Prediction [D]. Hangzhou: Zhejiang University, 2022 (in Chinese).
- [4] YAN J R, TAO T. Unsupervised anomaly detection in

- hourly water demand data using an asymmetric encoder-decoder model [J]. *Journal of Hydrology*, 2022, 613: 128389.
- [5] 何黎,陈磊,纪莎莎,等. 基于K-shape聚类的连续液位监测数据异常检测方法[J]. *中国给水排水*, 2023,39(11):56-61.
- HE Li, CHEN Lei, JI Shasha, *et al.* Abnormal detection of continuous water level monitoring data based on K-shape clustering [J]. *China Water & Wastewater*, 2023,39(11):56-61(in Chinese).
- [6] 刘书明,吴以朋,车晗. 利用自识别的供水管网监测数据质量控制[J]. *清华大学学报(自然科学版)*, 2017,57(9):999-1003.
- LIU Shuming, WU Yipeng, CHE Han. Monitoring data quality control for a water distribution system using data self-recognition [J]. *Journal of Tsinghua University (Science and Technology)*, 2017,57(9):999-1003(in Chinese).
- [7] AYADI A, GHORBEL O, OBEID A M, *et al.* Outlier detection approaches for wireless sensor networks: a survey[J]. *Computer Networks*, 2017,129:319-333.
- [8] 吴以文,杜坤,吴汉清,等. 基于LSSVM交互预测的供水管网爆管检测[J]. *中国给水排水*, 2022,38(9): 58-63.
- WU Yiwen, DU Kun, WU Hanqing, *et al.* Water supply network burst detection based on least squares support vector machine interactive prediction [J]. *China Water & Wastewater*, 2022, 38 (9) : 58-63 (in Chinese).
- [9] CHANDOLA V, BANERJEE A, KUMAR V. Anomaly detection: a survey [J]. *ACM Computing Surveys*, 2009,41(3):15.
- [10] LI A H, FENG M Y, LI Y R Y, *et al.* Application of outlier mining in insider identification based on Boxplot method [J]. *Procedia Computer Science*, 2016, 91: 245-251.
- [11] BREUNIG M M, KRÖGER H P, NG R T, *et al.* LOF: identifying density-based local outliers [C]. DUNHAM M, NAUGHTON J F, CHEN W D, *et al.* Proceedings of the 2000 ACM Sigmod International Conference on Management of Data. New York: Association for Computing Machinery, 2000:93-104.
- [12] XIA X, PAN X Z, LI N, *et al.* GAN-based anomaly detection: a review [J]. *Neurocomputing*, 2022, 493: 497-535.
- [13] SAEED N, NGUYEN S, CULLINANE K, *et al.* Forecasting container freight rates using the Prophet forecasting method [J]. *Transport Policy*, 2023, 133: 86-107.
- 
- 作者简介:**胡诗苑(1994- ),女,湖南常德人,博士研究生,主要研究方向为供水管网智能优化调度。
- E-mail:**hsy\_hit@163.com
- 收稿日期:**2023-04-23
- 修回日期:**2023-07-14

(编辑:任莹莹)

完善水利基础设施网络  
增强水安全保障能力