·创新与探索·

污染源自动监测异常数据识别规则及处理方法探索

魏艳[1,2],赖静娴[2],周启龙[2],彭雨林[3],姜继平[4]

(1.平安数字信息科技(深圳)有限公司,广东　深圳　518000;2.深圳衡伟环境技术有限公司,广东　深圳　518000;3.深圳市生态环境局坪山管理局,广东　深圳　518118;4.南方科技大学环境科学与工程学院,广东　深圳　518055)

摘　要:基于大量污染源自动监测数据的特征分析与异常原因解析,探索建立针对自动监测异常数据的识别规则与标志处理方法,并通过模型训练实现了异常数据的自动标志。经实例验证,该方法可识别异常偏高、异常偏低、异常为 0、迟滞不变、逻辑错误等 5 种类型的异常数据,按照数据有效性及异常原因进行标志处理,可以为后续数据分析及各类模型训练提供数据基础和保障。

关键词:污染源;自动监测;异常数据;识别规则;自动标志;模型训练

中图分类号:X830.3　　**文献标志码**:B　　**文章编号**:1006-2009(2022)02-0056-04

Exploration on Identification Rules and Processing Methods of Abnormal Data in Automatic Monitoring of Pollution Sources

WEI Yan[1,2], LAI Jing-xian[2], ZHOU Qi-long[2], PENG Yu-lin[3], JIANG Ji-ping[4]

(1. *Ping An Digital Information Technology (Shenzhen) Co., Ltd., Shenzhen, Guangdong* 518000, *China*; 2. *Shenzhen Hopeway Environmental Technology Co., Ltd., Shenzhen, Guangdong* 518000, *China*; 3. *Pingshan Administration Bureau of Shenzhen Ecological Environment Bureau, Shenzhen, Guangdong* 518118, *China*; 4. *School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen, Guangdong* 518055, *China*)

Abstract: Based on characteristic analysis and abnormal cause analysis of a large number of automatic monitoring data of pollution sources, the identification rules and identification processing methods for automatic monitoring abnormal data were explored and established, and the automatic identification of abnormal data was realized through model training. Examples showed that this method could identify five types of abnormal data, including abnormally high, abnormally low, abnormally zero, hysteresis invariance and logic error. According to data validity and abnormal reasons, this identification and processing method could provide data base and guarantee for subsequent data analysis and various model training.

Key words: Pollution source; Automatic monitoring; Abnormal data; Identification rule; Automatic identification; Model training

随着政府与公众对环境质量要求的提升,环境监管力度不断加强,污染源自动监测系统在我国快速发展[1]。以深圳市为例,截至 2020 年,全市安装污染源自动监测系统的企业已达 800 余家,实现了重点排污企业的在线监测全覆盖。然而,由于监测系统运维工作难以完全到位、企业有意无意逃避监管、在线监测设备质量参差不齐等原因,监测数据

收稿日期:2020-12-16;**修订日期**:2022-01-14

基金项目:国家自然科学面上基金资助项目(51979136);广东省普通高校特色创新(自然科学类)基金资助项目(2018KTSCX201);深圳市科创委技术攻关基金资助项目(JSGG20180508151852303)

作者简介:魏艳(1985—),女,重庆人,工程师,硕士,主要从事环境监测与环境信息化工作。

质量难以得到保障,也无法保证污染源监管的准确性[2-3]。国内关于自动监测数据质量和有效性判别方面的研究主要基于前端管控设备、质量控制方法等[4-12],《水污染源在线监测系统(COD_{Cr}、NH_3-N 等)数据有效性判别技术规范》(HJ 356—2019)(以下简称《规范》)也为如何进行有效数据的人工判别提供了流程和方法指引。目前针对污染源自动监测数据计算机自动审核方面的研究相对较少,如何对监测数据异常进行总结,发掘其变化模式,形成通用判别准则,是污染源监测监管的一个重要基础问题。今基于监测数据特征分析与异常原因解析,探索污染源自动监测异常数据的自动审核判别方法,并通过模型训练实现异常数据的自动标志,提高数据质量与处理效率,保障污染源监管的准确性和及时性。

1　数据异常特征与原因分析

1.1　监测数据异常的基本特征

参考《规范》要求,将流量为 0(排污口未排水)时段的监测数据归为停产时段数据,不纳入分析识别范围。《规范》中描述的"水质自动分析仪、数据采集传输仪及监控中心平台接收的数据误差大于 1%"等情况,仅能通过现场核查发现异常数据,无法通过计算机程序审核实现自动判别,可通过人工补充标志的方法将数据分类。

基于常年监测运维数据进行特征分析,将自动监测数据的异常表现总结归纳为 6 种类型,具备以下 6 类特征的自动监测数据均未在正常变化范围内波动,可初步判别为异常数据。

(1)异常偏高。某污染因子监测数值高于该企业规定的排放浓度限值。

(2)异常偏低。某污染因子监测数值低于监测设备检出限。如某企业排放废水中总镍质量浓度为 0.01 mg/L ~ 0.02 mg/L,低于广东省地方标准《镍水质自动在线监测仪技术要求》(DB44/T 1718—2015)规定的检出下限 0.05 mg/L,判别此段总镍监测数据异常偏低。

(3)异常为 0。监测数据出现 0 值。

(4)迟滞不变。一定时间(超出测量频率)范围内监测数据始终维持在一个数值。如某企业排放废水中总磷监测值持续 7 h 为 0.98 mg/L,而废水测量频率一般设置为 2 h/次,判别此段数据为迟滞不变数据。

(5)逻辑错误。监测数据不符合正确的逻辑关系,即违反学科认识的基本规律。如按照监测因子之间的关系,正确的逻辑关系应为氨氮 < 总氮、六价铬 < 总铬,违反此逻辑出现氨氮 > 总氮、六价铬 > 总铬即为逻辑错误数据。

(6)其他错误格式数据。其他因仪器噪声、传输错误所产生的个别错误格式数据。此类数据基本为孤立异常值,在处理过程中直接清洗,后续不再分析。

1.2　数据异常原因分析

根据异常数据核查记录与监测系统运行原理,分析各类异常数据产生的原因。

(1)异常偏高。主要原因一是企业排放污水浓度超标,实际数据偏高;二是在仪器校准校验过程中产生;三是仪器故障造成测量错误。

(2)异常偏低。主要原因一是废水中所含此类污染物浓度很低,已经低于设备检出限;二是仪器故障造成测量错误。

(3)异常为 0。主要原因一是废水中所含此类污染物浓度极低,已经低于传输标准要求的小数位数,如总铜按照《污染物在线监控(监测)系统数据传输标准》(HJ 212—2017)应保留小数点后 3 位,当监测数值低于 0.001 mg/L 时传输到平台的数据为 0,此种情形下的 0 值数据真实准确;二是仪器故障造成测量错误;三是监测设备正在进行空白样(零标)校准或比对测试。

(4)迟滞不变。主要原因一是监测设备故障,无法正常测量,按照规则此时会产生设备报错,并且传输上一个测量数据,而监测仪器状态信息无法传回数据平台,导致平台持续收到同一个完全不变的数值;二是设备正在进行校准等质控操作,此时设备不再测量水样,而是持续上传上一个测量值。

(5)逻辑错误。主要原因是监测设备异常引起的测量不准确,可能是分量参数或总量参数测量不准确造成,也可能是两项参数测量均不准确造成。因此,在逻辑错误的异常数据中一定存在无效数据。

2　异常数据识别规则与处理方法

2.1　异常数据识别规则

排除因停工停产而产生的无效数据,针对其他监测数据,基于上述 6 类污染源异常数据特征表

现,整理出相应的识别规则(见表 1)。表 1 中各项数据识别规则仅根据数据的具体表现建立,并不能说明该数据的真实性、准确性,还需要进一步进行数据的标志处理。

表 1　污染源异常数据识别规则

Table 1　The methods of identification of abnormal data

异常类型 ID	数据类型	识别规则①	解释
类型 1:Abn-High	异常偏高	$C_{obs} > C_s$	监测数据 > 排放因子限值
类型 2:Abn-Low	异常偏低	$C_{obs} < C_{lim}$	监测数据 < 监测设备检出限
类型 3:Abn-Zero	异常为 0	$C = 0 \mid T > T0$	监测数据为 0
类型 4:Abn-Stil	迟滞不变	$C > C0 \mid T > T0$	超过测量频率(如 2 h)监测数据维持在同一个数值
类型 5:Abn-Log	逻辑错误	$C_{WQ1} > C_{WQ2} \mid WQ1 \subset WQ2$	某化学元素的分量 > 总量
类型 6:Abn-Other	其他错误格式数据	——	格式检验

①C_{obs}为排污口自动监控数据;C_s为排污许可证中规定的该排放因子浓度限值;C_{lim}为该排放因子监测设备的检出下限;T 为监测时间;$T0$ 为监测频率;C_{WQ1}和 C_{WQ2}分别为某化学元素的分量和总量。

2.2　异常数据标志处理

建立异常数据标志处理方法,实际上是将数据分类,将各类异常数据可能产生的原因汇总(见表 2)。表 2 中真实有效的数据可用于监管执法,设备故障(包含设备维修运维时段)和校准校验都无法反映企业排放废水的真实情况,应视作无效数据。基于以上分析,异常数据根据其真实性可分为有效、无效两种。在有效数据中,异常偏高反映企业废水污染物浓度超过排放标准,不符合环保管理规定,而偏低或异常为 0 均为正常的浓度偏低,符合环保管理规定,因而有效数据可分为有效-正常、有效-超标两类。针对无效数据,可根据其产生原因分为无效-设备故障、无效-校准校验两类。综上,污染源异常监测数据的类型可标志为 2 种 4 类,即有效-正常、有效-超标、无效-设备故障、无效-校准校验。结合监测系统现场核查,可准确标志各类异常数据,在分类数据积累充分后,可利用机器学习训练开发自动标志模型。

表 2　数据异常原因

Table 2　The potential drivers lead to data abnormal

数据类型	数据真实有效	设备故障	校准校验
异常偏高	√	√	√
异常偏低	√	√	
异常为 0	√	√	√
迟滞不变		√	√
逻辑错误		√	

3　自动标志模型训练

利用深度学习方法,训练针对异常偏高、异常偏低、异常为 0、迟滞不变、逻辑错误等 5 个不同异常类型数据的分类模型,实现异常数据的自动标志。模型训练过程主要包括数据预处理、特征工程、模型训练及应用。以异常偏高数据的二分类自动标志模型训练为例说明(真实超标数据量较少,以校准校验、设备故障训练二分类模型为例),其余各类型数据的模型训练参照该方法。

3.1　数据预处理

整理已正确标志的异常偏高历史数据,每个样本使用 1 d 的数据量,以分钟为采样频率,每个样本包含 1 440 个数据。核查数据的真实性、完整性,并对样本进行去重、补缺、去除异常样本等处理。使用有监督学习方式,数据格式重组为(样本数 $n \times 1\ 440$ 行,1 列),各样本的量纲大小不一致,对数据进行 0 -均值标准化。将各类对应原因的标志以数字替代,0 表示校准校验,1 表示设备故障,完成训练数据的预处理。

3.2　特征工程

分析校准校验和设备故障样本,发现可以从以下 3 个方向进行区分:①偏高部分前后数据形状差异。校准校验数据表现为前后数据较平稳,无剧烈波动,而设备故障数据则存在较大波动。②偏高部分形状差异。校准校验数据一般呈现先低后高、先高后低、单高 3 种规律形状,设备故障数据无明显固定形状。③偏高发生的时间差异。校准校验偏高数据发生时间大多为中午—下午,设备故障偏高数据多发生于上午和夜晚。

3.3　模型训练及应用

使用 Python 的 Keras 深度学习框架,将 80% 的数据划分为训练数据、20% 的数据划分为测试数据进行模型训练并调参。经过反复试验,当 LSTM 使

用单层神经元个数为 64、CNN 网络使用两层神经元个数依次为 32 和 64、卷积核大小为 5 时效果最佳，训练准确率为 97. 08%，测试准确率为 94. 74%，神经网络部分参数与范围见表 3。训练后在真实生产环境中取异常偏高 31 个样本预测，准确率达到 87%。

表 3　神经网络部分参数与范围

Table 3　The parameters and ranges of neural network

参数	范围
神经网络层数 *n*/层	1、2、3、4、5
神经元个数 *n*/个	8、16、32、64、128
Loss 函数	Binary_Crossentropy、Categorical_Crossentropy、Sparse_Categorical_Crossentropy
激活函数	Sigmoid、Tanh、ReLU、Softmax
梯度下降法	SGD、RMSprop、Adam、Nadam、Ftrl
Mini-batch	4、8、16、32、64、128、256

4　实际应用案例

利用上述数据处理方法，在深圳市某区的污染源监管系统进行试验。2020 年 7 月，该区共接收实时数据 1 300 余万个，利用识别规则共识别出异常数据 410 段（将一次异常发生的起止时间内的数据作为一段异常数据），占接收数据的 3. 4%。在 410 段异常数据中，异常偏高、异常偏低、异常为 0、迟滞不变、逻辑错误数据分别有 111 段、129 段、121 段、37 段、12 段，占比分别为 27%、31%、30%、9%、3%。

对 410 段异常数据中的 378 段进行标志处理，有效-正常、有效-超标、无效-设备故障、无效-校准校验数据分别有 206 段、7 段、51 段、114 段，占比分别为 54. 5%、2%、13. 5%、30%。在处理过程中，异常偏高数据采用模型自动标志，准确率达到 82%。由此可见，利用上述异常数据识别规则及处理方法，不仅能将无效数据从海量监测数据中筛选出来，还能为其他应用提供可靠的数据材料。

5　结语

研究通过分析污染源自动监测数据中异常数据的特征与产生原因，建立异常数据识别规则与标志处理方法，并应用深圳市某区的监测数据进行了实例分析。该方法不仅便于将海量自动监测数据中的异常数据自动识别出来，还可以将有效、无效数据自动识别分类，避免无效数据干扰监管执法、监督考核等环境管理应用，提高了监测数据的准确度、真实性。此外，基于大数据思维方式按产生原因及数据特征积累的大量数据，可用于挖掘异常数据出现的规律，优化改善运维和监管流程，利用异常数据出现的特征还可以进一步提出定量可考核的监测运维质量技术标准。

［参考文献］

［1］ 姚建松，蔡俊云，刘青阳，等. 环境保护工作中污染源自动监测系统的应用［J］. 节能与环保，2019（5）：96 - 97.

［2］ 刘燕. 污染源自动监测存在的问题及应对方法研究［J］. 黑龙江科技信息，2019（14）：170 - 171.

［3］ 张奇磊. 影响水质自动监测系统监测数据准确性的几个因素［J］. 干旱环境监测，2007，21（3）：184 - 186.

［4］ 董广霞，景立新，周冏，等. 监测数据法在工业污染核算中的若干问题探讨［J］. 环境监测管理与技术，2011，23（4）：1 - 4.

［5］ 徐薇薇，刘常永，王增国，等. 污染源自动监测设备动态管控系统技术及应用［J］. 环境监测管理与技术，2017，29（1）：69 - 71.

［6］ 王军霞，刘通浩，张守斌，等. 排污单位自行监测监督检查技术研究［J］. 中国环境监测，2019，35（2）：23 - 28.

［7］ 姜继平，王鹏，刘洁，等. 突发水污染预警应急响应研究与实践的方法学辨析［J］. 环境科学学报，2017，37（9）：3621 - 3628.

［8］ 彭刚华，梁富生，夏新. 环境监测质量管理现状及发展对策初探［J］. 中国环境监测，2006，22（2）：46 - 49.

［9］ 周良，尹卫萍. 浅谈环境监测质量管理工作［J］. 环境监测管理与技术，2012，24（5）：5 - 7.

［10］ 樊萍，孙健，李坤，等. 环境监测数据预审中部分指标的相关性分析［J］. 环境研究与监测，2006，19（4）：17 - 19.

［11］ 贾立明，滕曼，魏庆斌，等. 自动在线监测系统利用软件数据质控有效性研究［J］. 环境科学与管理，2013，38（5）：153 - 156.

［12］ 曲凯，李彦，崔志伟. 空气自动监测网络数据有效性的自动化判别［J］. 环境保护科学，2011，37（6）：1 - 3.