

基于高频在线水质数据异常的突发污染预警

史 斌¹,姜继平^{1,2*},王 鹏^{1,3} (1.哈尔滨工业大学环境学院,黑龙江 哈尔滨 150090; 2.南方科技大学环境科学与工程学院,广东 深圳 518055; 3.哈尔滨工业大学城市水资源与水环境国家重点实验室,黑龙江 哈尔滨 150090)

摘要: 在高频水质自动监测背景下,建立了基于软测量和水质时间序列异常检测的水体突发污染预警预报技术.假定突发污染事故会引起典型自动监测水质参数变化,采用回归分析建立水质参数和在线高频监测水质参数间的线性关系进行软测量,采用人工神经网络预测短程水质变化,建立基于预测残差的异常判断最小阈值,最终通过有序监督聚类进行水质突变检测从而对突发污染事故进行预警.采用美国弗吉尼亚州的Potomac River 流域在线监测数据进行算法验证和案例分析.分析受试者工作曲线(ROC)表明:该方法对 2 倍异常和 3 倍异常水平的检测准确率分别为 62.7%和 92.5%,且随着异常水平的增加准确率增加,通常突发污染事故中特定污染物浓度水平一般明显高于 3 倍,该方法具有较高的准确率.较其他突发污染水质预警技术,该技术有效缩短了平均检测时间,为流域污染预警预报和快速应急响应提供了新途径.

关键词: 突发水污染事故; 高频水质自动监测; 异常检测; 软测量; 人工神经网络

中图分类号: X522 文献标识码: A 文章编号: 1000-6923(2017)11-4394-07

Early warning of water pollution incidents based on abnormal change of water quality data from high frequency online monitoring. SHI Bin¹, JIANG Ji-ping^{1,2*}, WANG Peng^{1,3} (1.School of Environmental, Harbin Institute of Technology, Harbin 150090, China; 2.School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China; 3.State Key Laboratory of Water Resources and Water Environment, Harbin Institute of Technology, Harbin 150090, China). *China Environmental Science*, 2017,37(11): 4394~4400

Abstract: With the high frequency automatic monitoring of surface water quality, a technique for early warning of water pollution incidents was developed using the water quality soft measurement and abnormal detection of time series. This technique takes the assumption that water pollution incidents would cause the change of typical automatic monitoring water quality parameters, and then establishes the linear relationship between the water quality parameters and online high frequency monitoring water quality parameters. Using the artificial neural network, the change of water quality parameters in a short duration was predicted; using the time series of residual error, the threshold of abnormal change was determined. Finally, early warning of pollution incidents could be achieved through detecting abnormal change based on sequential leader clustering algorithm. To verify the technique, this study takes the online monitoring data obtained from the Potomac River in Virginia, USA as a case study. The analysis of the receiver operating characteristic curve (ROC) shows that the detection accuracies of double and triple abnormal levels can reach 62.7% and 92.5%, respectively. Because the concentration level of a water pollution incident is usually significantly higher than 3times, this technique can provide a relative high accurate early warning. Compared with traditional abnormal detection methods, this technique can shorten the detection time. Along with increasing improvement of automatic monitoring facilities, this study provided a new avenue for early warning of, and prompt response to, pollution incidents.

Key words: water pollution incident; high frequency automatic water quality monitoring; abnormal change; soft measurement; artificial neural network

当前,在世界范围内蓄意或非故意化学品泄露造成的水体突发污染仍不断发生^[1-2].尤其在中国,几十年来的经济高速发展,将我国带入了突发环境污染事件的高发期^[3-5].频发的水污染事件及其伴随的严重危害引起了各级政府关注,

收稿日期: 2017-04-25

基金项目: 中国博士后科学基金资助项目(2014M551249);国家自然科学基金资助项目(51779066);中央高校基本科研业务费专项基金资助项目(HIT. NSRIF.2017060)

* 责任作者, 助理教授, jiangjp@sustc.edu.cn

也促进了管理部门和环境专家们致力于开发污染预警响应技术,服务于水质安全保障。

随着国家对河流水质管理的重视,水体监测基础设施日益健全,数据积累日渐丰富,实现大数据支撑下的水环境管理已经日益成熟。在发达国家,如美国已经实现了对河流水温、DO、浊度、pH 值、电导率和硝酸亚硝酸盐等水质指标的每 5~15min 监测一次并每小时发布^[6]。我国 2008 年以来也布设了国家水质自动监测系统,但是在采样频率、监测指标、覆盖程度上还远远不够。

汤鸿霄^[7]指出,下一步环境科学和技术的研究要充分运用卫星遥感、数值信息、模拟模式、多媒体图像等学科技术,再加上互联网,大数据,人工智能等新增创新手段,试图以数字化来描述大范围环境体系变化,达到准确阐释和预报环境动态和灾害。

基于高频水质监测,结合数据流异常检测技术,可以开发新型的突发水污染预警技术,用于流域的智能化管理。目前,基于高频时间序列的异常检测问题广泛应用于卫星遥测、金融诈骗、网络入侵、安全防御、电信数据管理、传感器异常分析等诸多领域^[8-9],随着数据流异常检测技术在各个领域的广泛应用^[10]和目前水环境数据采集技术的不断发展和丰富^[11],基于水质监测数据的异常检测研究开始引起了环境专家们的兴趣。目前大多研究是针对城市供水管网在线水质监测系统的异常检测^[12-13]。而河流水质由于受到水文、气象、污染物等多种因素的综合影响,其变化规律难以捕捉,目前为止国内外基于河流在线水质监测数据的突发污染预警研究较少。

本研究针对目前突发污染事件中存在迟知、未知的问题,基于在线水质自动监测站高频水质监测数据,研究建立在线突发污染预警技术,基于水质参数多元回归模型构建关键非在线监测水质参数的软测量方法,并结合神经网络水质预测残差和聚类分析检测下一时刻的水质异常变化,预报可能出现的突发污染。通过对建立的预警方法采用美国弗吉尼亚州的实际观测数据构建典型情景进行验证,对所建立的方法在应用过程可能存在的问题也进行了讨论。

1 研究方法

突发污染不同于常规污染情景,污染物种类存在极大的不确定性,目前整个自然界有超过 10 万种的毒性污染物,即使世界上最先进的检测仪器也无法全部检测出单个物质。但由于水质常规高频自动监测指标对大部分污染物都会表现出不同程度的敏感性^[14-15],开展基于水质常规高频监测指标的污染物浓度软测量是可行的^[16]。

回顾我国突发污染案例,其主要污染物可分为有机物、重金属或酸碱性无机物等^[17],结合自动监测指标的类型,可以选用 pH 值、浊度、电导率和 UV254 等 4 个指标作为监测数据集,pH 值代表水体的酸碱度,浊度代表水体受悬浮有机物污染的程度,电导率代表水中离子浓度,而 UV254 在一定程度上可以代表水体受有机污染程度。

基于上述在线水质监测指标,通过分析水质参数之间的相关关系,找出水质变化规律;并基于变化规律采用数据驱动模型预测当前水质情况;与实际监测数据进行对比,计算预测残差;以异常判断最小阈值为基准,采用有序监督聚类方法判断水质参数预测残差与最小阈值的大小,来实现异常检测判断。

1.1 水质参数相关性分析

对于同一河段,水体中任何一种污染物组分的变化都可能会引起上述几种自动监测指标的变化^[18]。Jeffery 等^[19]研究了如何利用高频水质参数如电导率和浊度表示不能进行高频监测的水质参数,总结出如下回归方程:

$$y_i = mx_i + b + e_i \quad i = 1, 2, \dots, n \quad (1)$$

式中: y_i 为第 i 个目标污染物; m 和 b 为参数; e_i 为相对误差; x_i 为水质自动监测参数; n 为监测次数。

回归方程可以近似表达突发污染物进入水体后引起常规水质指标的变化,然后结合指标权重进行叠加分析,通过比较该指标预测时间序列和实际监测时间序列的残差,实现突发污染异常检测。

1.2 水质时间序列预测与残差辨识

由于水质异常通常由多种水质参数综合异常所致,而每种水质指标有不同的量纲,有必要在

水质指标叠加之前对原始数据进行标准化处理。我们选用 Z-score 标准化将原始时间序列转化为均值为 0, 方差为 1 正态分布的无量纲序列。其处理方程如公式(2)所示:

$$y_i = \frac{x_i - \bar{x}}{\sigma} \quad (2)$$

式中: $\bar{x}$ 为所有样本数据的均值, σ 为所有样本数据的标准差。

进一步采用人工神经网络等数据驱动模型, 对常态水质时间序列进行学习、计算、记忆、自适应以及智能处理。对水质发展趋势进行预测, 计算预测值和实际监测值残差, 从而将残差与指定阈值比较, 判断水质异常变化情况。

阈值选择的合理性对于异常检测算法起着极其重要的作用^[20], 为尽量提高算法异常检出率, 本研究采用最低实际阈值法来确定阈值:

$$\text{Thresh}_{\min} = \bar{\varepsilon} + 3\sigma \quad (3)$$

式中: $\bar{\varepsilon}$ 和 σ 分别为其残差分布的平均值及标准差。

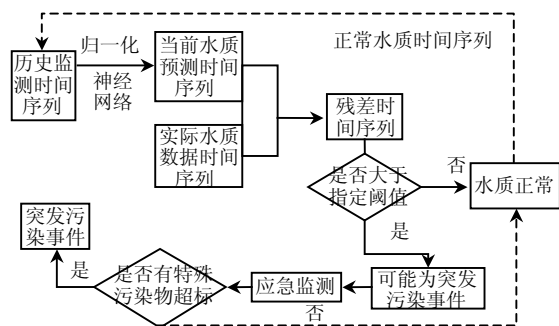


图1 基于高频水质时间序列的突发污染异常检测流程

Fig.1 The flow chart of abnormal detection for sudden water pollution event base on water quality time series of high-frequency monitoring

综上, 采用基于回归分析的神经网络预测方法进行异常检测的流程如图 1 所示, 主要步骤包括: (1) 分析变量的影响因素, 并确定主要因素及其影响程度; (2) 利用历史数据建立预测变量与主要影响因素之间的回归关系模型; (3) 把预测期主要影响因素的指标值, 代入网络进行预测; (4) 基于预测水质时间序列和实际监测水质时间序列的残差序列与指定阈值对比判断是否异常; (5) 启

动应急监测来监测是否有特殊污染物超标, 进一步确定是否为突发污染事件。

2 案例分析

2.1 Potomac River 水质监测数据收集

由于目前国内难以获取河流高频水质监测数据, 本研究所需的高频水质监测数据来源于美国地质调查局(<http://waterdata.usgs.gov/nwis>)。正常情况下水质监测频率为 15min, 监测数据现场存储, 并且每隔 1~4h 传输到美国地质调查局网站上。我们选择了位于弗吉尼亚州的 Potomac River 流域 01632900 号站点最近 3 个月的监测数据 (1~3 月, 共 55d), 其水质监测指标主要包括: 水温、电导率、溶解氧、pH、浊度和硝态亚硝态盐氮等 6 个。对于大量高频监测数据, 尽管每周都对自动监测站进行维护, 但是也难以避免存在间断性的数据缺失^[21]。对于单点缺失数据采用缺失点前后各两个点求均值的方法补充。对于连续缺失数据采用数据平移的方法补充, 即平移缺失点之前相同数量的数值填补缺失。另外, 对原始数据中的冗余数据和单位不一致数据, 分别采用人工剔除和数值单位转换的方法处理。

2.2 基于常态数据集的异常判断最小阈值构建

原始监测数据首先进行标准化处理, 图 2 为标准化后水质时间序列的变化情况。其中水温、pH 值、DO 和硝态亚硝酸盐氮 4 个指标呈现出明显的以天为单位的周期变化, 而浊度和电导率变化受环境变化比较敏感, 周期性变化不明显。

采用包括当前时刻的 96 个数据 (共 24h) 预测下一时刻 (第 97 个数据) 的水质监测数据, 当第 97 个监测数据获得后, 进行对比分析。

在本案例中采用 BP 神经网络, 设定两个隐含层, 分别选用 Log-sigmoid (输入值为任意值, 输出值在 0 和 1 之间) 和 Purelin 传递函数 (输入输出值均为任意值) 作为激活函数, 采用标准梯度下降算法训练神经网络。基于前 35d 共 3360 组监测数据, 通过测试调整神经网络结构, 依据“奥卡姆剃刀”原理, 最后确定隐含层分别为 5 个和 1 个节点神经元。以后 20d 共 1920 组监测数据作为检验集, 验证所训练的神经网络预测能力。

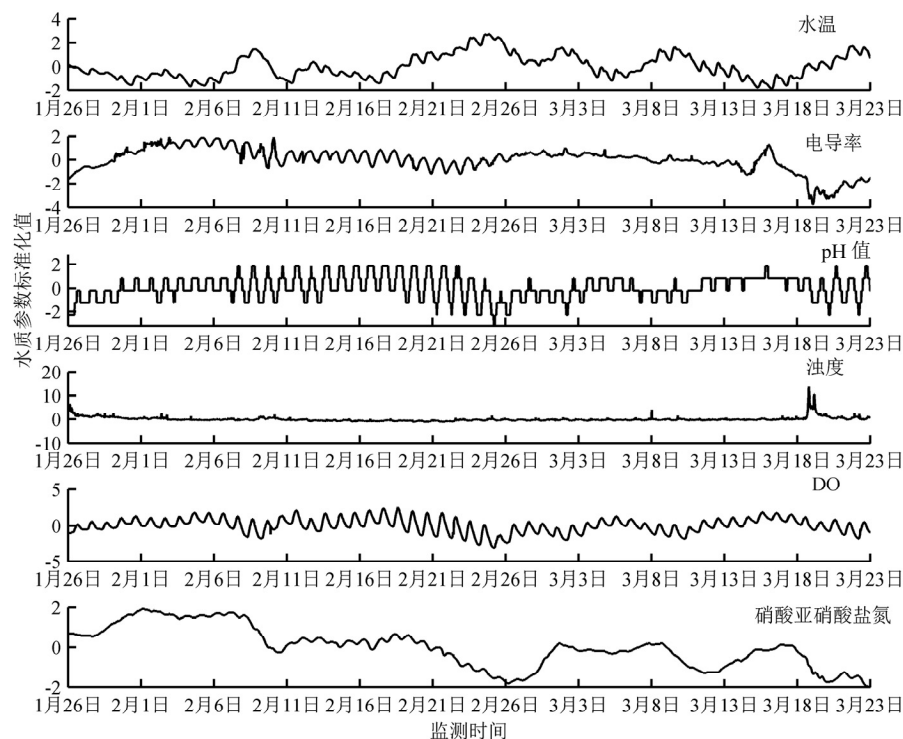


图 2 标准化水质自动监测数据时间序列

Fig.2 The time series of standardized automatic monitoring water quality data

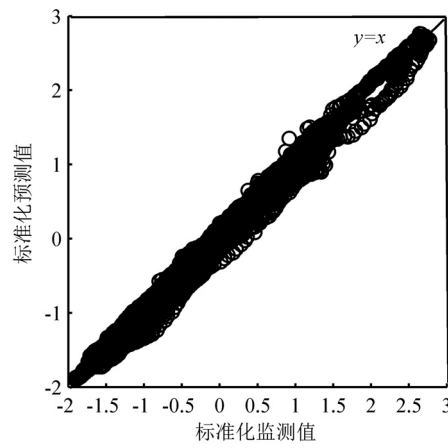


图 3 基于神经网络的预测值与实际监测值对比分布
Fig.3 The contrastive distribution of predicted values by ANN and monitored values

以浊度为例,采用公式(4)计算检验集预测残差,检验集的神经网络预测能力表现如图 3 所示,其平均预测残差为 0.0104,中位数残差为 0.0052,平均相对误差为 1.54%,有很高的预测精度.

RE=|预测值-监测值|

(4)

基于上述预测残差序列,结合异常判断阈值计算公式(3),计算得到异常判断最小阈值为 $\text{Thresh}_{\min} = \bar{\varepsilon} + 3\sigma = 0.18 + 3 \times 0.69 = 2.35$. 其中在阈值范围以内的数值为 97.7%,超出阈值的占 2.3%.

2.3 假想数据集异常检测的准确性分析

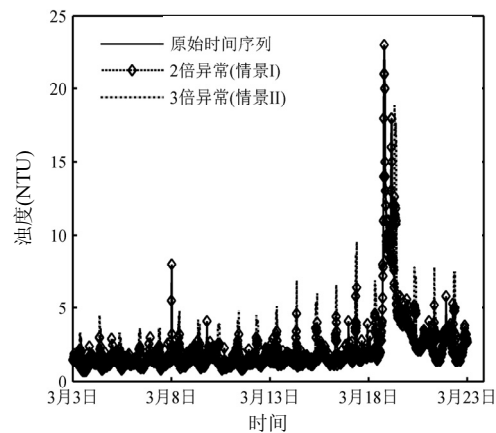


图 4 水质异常变化(浊度)的时间序列假设情景
Fig.4 The hypothetical scenario of water quality abnormal changes (Turbidity)

确定启动预警阈值后,结合人工构建的异常水质监测数据流,进行异常检测准确性分析.

以浊度监测数据为例,检验集每天 10:00~12:00 间的样本分别叠加 2 倍和 3 倍异常,即将观测数据人为放大 2 倍和 3 倍,构造出 2 组新的时间序列,记为情景 I 和情景 II,可代表发生突发污染的水质异常变化情形(例如每天上午 10:00~12:00 点有点源偷排),如图 4 所示.

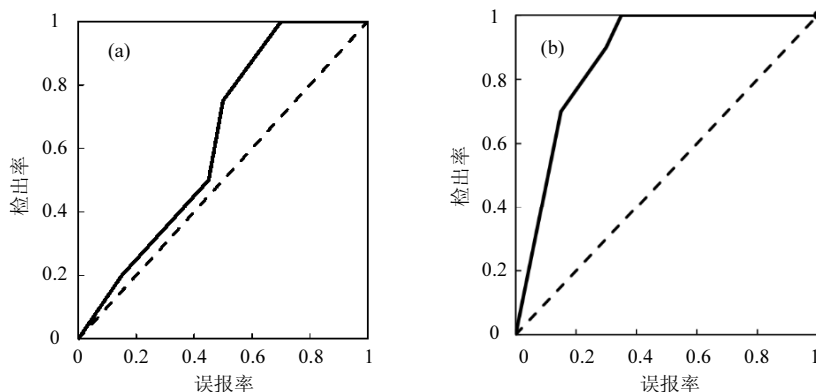


图 5 异常情景 I(a)和情景 II(b)时 ROC 曲线

Fig.5 The ROC curves of abnormal scenario I (a) and II (b)

2.4 基于软测量的非在线监测水质参数异常检测与水质预警

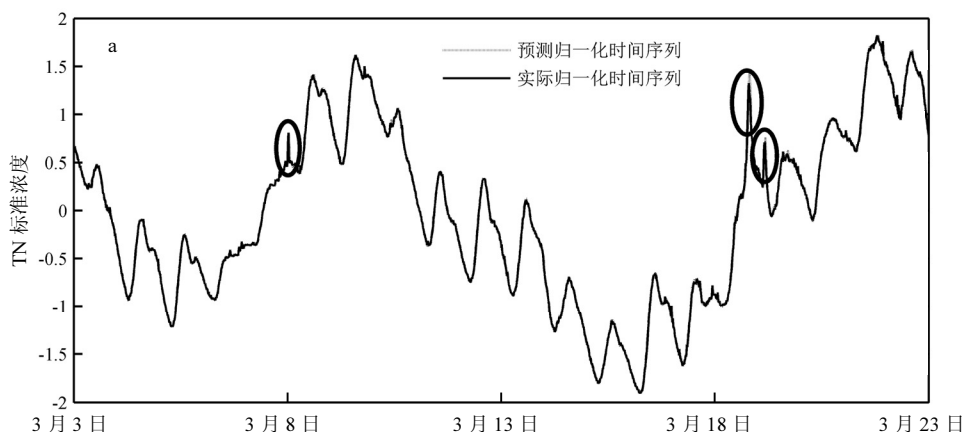
非在线监测水质参数,如总氮可采用类似公式(5)的回归公式^[13]进行软测量.结合软测量技术,本文提出的异常检测方法应用范围更为广泛.

结合(5)式,构建接近真实情形的总氮高频时间序列,进行异常检测分析.依据 2.2 中所述步骤

训练 BP 神经网络,计算得到总氮的预测值,与实际监测值,对比得到对应的残差序列,如图 6 所示.基于残差序列,利用公式(3)计算得到异常检测最小阈值.

$$\text{TN} = 0.00317 \text{TURB} + 0.0234 \text{WT} - 0.0000655 \text{SC} + 0.469 \quad (5)$$

式中: TURB 为浊度标准化值, WT 为水体温度标准化值, SC 为电导率标准化值.



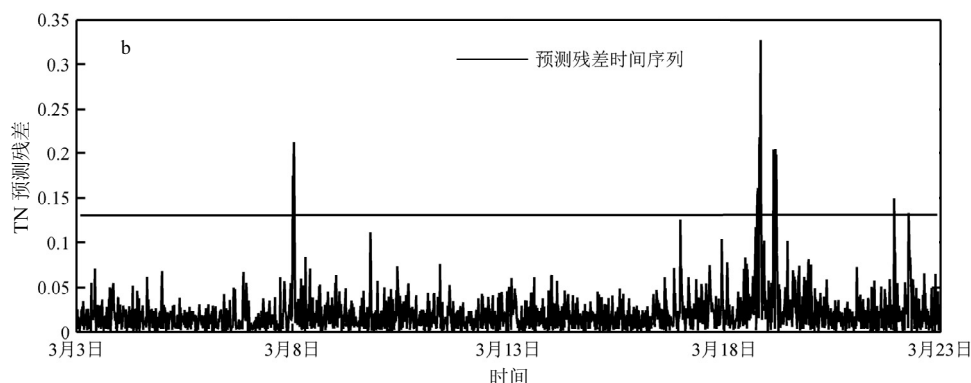


图6 TN 时间序列及预测残差

Fig.6 The time series of TN and predicted residual errors

基于神经网络训练后的异常判断最小阈值标准,针对需要检测水质参数时间序列,按照采样时间顺序,采用有序监督聚类方法^[19]进行聚类,对于小于最小阈值的判断为正常事件,大于最小阈值的判断为异常事件。从图6可以看出,总体上预测残差基本在最小阈值($\text{Thresh}_{\min}=0.1324$)范围内。在3月8日和3月19日出现了2次比较明显的水质突变,预测残差值也均明显高于指定阈值。通过查询对应时段的气象条件以及该站点的流量监测数据:3月8日平均流量为 $1.06\text{m}^3/\text{s}$,3月19日平均流量为 $1.38\text{m}^3/\text{s}$,均明显高于预测期平均流量 $0.83\text{m}^3/\text{s}$ 。分析其原因可能是由于降雨冲刷地表,将陆地上的含氮有机物冲刷进入水体造成,这种情况下应当启动应急监测,监测是否有特殊污染物进入水体以判断是否为突发污染事故。

2.5 讨论

在实际使用过程中,该方法还可能如下问题:(1)该异常检测方法的前提是常态水质时间序列训练得到的神经网络具有一定的预测精度和泛化能力。这就要求训练数据集要有代表性,例如针对不同时期的突发污染事件异常检测,取不同水文条件下的监测数据率定;(2)在标准正态分布的情形下, $\pm 3\sigma$ 所占的面积为99.74%,要比上面实际案例中的97.7%略大,考虑实际统计推断过程,该偏差是合理的;(3)若误差分布为非正态分布,需要构建其他统计准则进行判定;(4)上述异常判定过程中采用了残差作为基本度量,而不是相对残差,使得异常检测结果更为保守,不会略

掉可疑的水质异常波动。(5)可考虑结合小波技术等构建杂交预测模型,提高异常检测能力。

3 结论

3.1 本研究充分利用了水质自动监测站高频监测数据,与我国目前已有的水质在线监测方法和依据公众投诉、公共安全检查、常规抽样分析等传统异常检测方法相比,其对污染物浓度变化更敏感,检测效率更高。

3.2 本研究所提出的异常检测方法对较低浓度的水质异常检出准确率低,而且还存在一些误报现象,但是能够全部检测出实际异常时间序列,不会存在异常漏报现象。与后续的应急监测工作相结合即能够提高检出的准确率,并为真正的突发污染事件提供基础数据支持。

3.3 随着河长制的逐步推行,该方法在明确环境责任,提高预警判定的准确性和时效性方面将会有广泛的应用。将来为了加强河段环境管理,在目标河段安装包含pH值、浊度、电导率和UV254等指标的高频传感器必将能够更能充分发挥该异常检测方法的优势。

参考文献:

- [1] 李尧远,曹蓉.我国应急管理研究十年(2004~2013):成绩、问题与未来取向[J].中国行政管理,2015,(1):83~87.
- [2] XUE P, ZENG W. Trends of environmental accidents and impact factors in China [J]. Frontiers of Environmental Science & Engineering in China, 2011,5(2):266~276.
- [3] 刘仁涛,郭亮,姜继平,等.环境污染应急处置技术的CBR-

- MADM 两步筛选法模型 [J]. 中国环境科学, 2015,35(3):943-52.
- [4] 赵艳民,秦延文,郑丙辉,等.突发性水污染事故应急健康风险评价 [J]. 中国环境科学, 2014,34(5):1328-1335.
- [5] Camp J V, Abkowitz M D, Leboeuf E J. Inland waterway resource and spill management needs in Southeastern USA [J]. Disaster Prevention And Management, 2010,19(4):483-497.
- [6] US EPA, 2017. National hydrography dataset high-resolution flowline data [EB/OL]. The National Map. <https://www.data.gov/>.
- [7] 汤鸿霄.环境科学技术的扩展融合趋势 [J]. 环境科学学报, 2017,37(2):405-406.
- [8] 李维铮,孟 桥.基于遥测数据动态特征的卫星异常检测方法 [J]. 空间科学学报, 2014,34(2):201-207.
- [9] 田嘉瑞.传感器网络数据流异常数据检测与修正 [D]. 哈尔滨:哈尔滨工程大学, 2008.
- [10] Zou S, Liang Y, Poor H V, et al. Nonparametric Detection of Anomalous Data Streams [J]. Computer Science, 2016,1:1-26.
- [11] 黄树成,曲亚辉.数据流分类技术研究综述 [J]. 计算机研究应用, 2009,26(10):3604-3609.
- [12] 陈 玥.城市供水水质异常检测方法研究 [D]. 杭州:浙江大学, 2013.
- [13] 魏 媛,冯天恒,黄平捷,等.管网水质多指标动态关联异常检测方法 [J]. 浙江大学学报(工学版), 2016,50(7):1402-1409.
- [14] Christensen V G. Characterization of surface-water quality based on real-time monitoring and regression analysis, quivira national wildlife refuge, south-central kansas, december 1998 through June 2001 [R]. Center for Integrated Data Analytics Wisconsin Science Center, 2001.
- [15] Horsburgh J S, Jones A S, Stevens D K, et al. A sensor network for high frequency estimation of water quality constituent fluxes using surrogates [J]. Environmental Modelling & Software, 2010,25(9):1031-1044.
- [16] 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(1):1-13.
- [17] Liu J, Guo L, Jiang J P, et al. Evaluation and selection of emergency treatment technology based on dynamic fuzzy GRA method for chemical contingency spills [J]. Journal of Hazardous Materials, 2015,299:306-315.
- [18] Helsel D R. Statistical methods in water resources / D.R. Helsel and R.M. Hirsch [M]. Amsterdam, New York: Elsevier, 1992.
- [19] Parmar K S, Bhardwaj R. Water quality management using statistical analysis and time-series prediction model [J]. Applied Water Science, 2014,4(4):425-434.
- [20] Mckenna S A, Hart D B, Murray R, et al. Testing and evaluation of water quality event detection algorithms [M]. Handbook of Water and Wastewater Systems Protection, 2011.
- [21] PE A-AYALA A. Educational data mining: A survey and a data mining-based analysis of recent works [J]. Expert Systems with Applications, 2014,41(4):1432-1462.

致谢: 感谢香港大学陈骥教授对论文英文部分的审阅和修订。

作者简介: 史 斌(1987-),男,山东郓城人,哈尔滨工业大学博士研究生,主要从事环境数学模型与决策支持系统方向研究.发表论文 10 余篇。