基于随机森林和模糊综合评价的地表水水质评价

王鑫民[1]　李伟英[1,2]　周　宇[1]　景昕宇[1]　陈　胜[1]

(1 同济大学环境科学与工程学院，上海　200092；2 同济大学长江水环境教育部重点实验室，上海　200092)

摘要：基于 M 河自上游到下游 4 处水质监测点 10 年间(2011－2020 年)的每月水质监测数据，采用模糊综合评价和随机森林的方法，对 M 河水质进行系统性综合评价。结果表明采集的水样超过 50%低于Ⅲ类水水质标准，下游污染明显高于上游；基于随机森林的水质评价模型，以各水质指标作为模型特征参数，对水质评价分类的准确率为 99%，揭示各水质指标在水质评价中的重要性排序，其中总氮、粪大肠菌群和溶解氧为最重要的水质指标。本方法及结论为 M 河水污染溯源、水质评价、水污染防治及水资源管理提供了理论基础与技术支撑。

关键词：水质；模糊综合评价；随机森林；关键指标

中图分类号：TU992；X799　　**文献标识码**：A　　**文章编号**：1002－8471(2022)02－0128－05

DOI：10. 13789/j.cnki.wwe1964. 2021. 09. 10.0005

引用本文：王鑫民，李伟英，周宇，等. 基于随机森林和模糊综合评价的地表水水质评价[J]. 给水排水，2022，48(2)：128-132. WANG X M，LI W Y，ZHOU Y，et al. Surface water quality evaluation based on random forests and fuzzy comprehensive evaluation[J]. Water & Wastewater Engineering，2022，48(2)：128-132.

Surface water quality evaluation based on random forests and fuzzy comprehensive evaluation

WANG Xinmin[1]，LI Weiying[1,2]，ZHOU Yu[1]，JING Xinyu[1]，CHEN Sheng

(*1 College of Environmental Science and Engineering，Tongji University，Shanghai 200092，China*；
2 Key Laboratory of Yangtze River Water Environment <Ministry of Education>，Tongji University，Shanghai 200092，China)

Abstract：M River is the drinking water source of a southern city of China，we collected monthly water quality data for a 10-year period (2011－2020) at four water quality monitoring sites from upstream to downstream of the M River. Fuzzy comprehensive evaluation and random forests method were used to comprehensively evaluate the water quality of M River. The results showed that more than half of the water samples of M River exceeded the Class Ⅲ standards，and there was a clear trend of deterioration from upstream to downstream. Random forests models effectively calculate the feature importance of water quality parameters. Total nitrogen，fecal coliform and dissolved oxygen were the most important water quality parameters. These findings have comprehensive analyze the water quality of M River，and have provided strong support for the water quality management of M River.

基金项目：国家自然科学基金 (51979194)；长江上游水质生物安全风险跨区域联防联控机制与战略研究(04002380114)；供水管材内外壁腐蚀特征与安全性能研究 (kh0040020210434)。

Keywords: Water quality; Fuzzy comprehensive evaluation; Random forests; Key index

0　引言

目前，地表水水质评价方法主要有单因子指数评价法、内梅罗污染指数法、综合污染指数评价法、模糊评价法、随机森林评价法等[1]。模糊数学评价法能将影响环境质量的主要污染因子进行归一化处理，同时还能综合评价不同年份、不同区域环境质量及其变化趋势。张洪伟[2]把模糊综合评判法应用于甘肃省正宁县区域内，针对不同河流和水库地表水的水质指标检测数据做出了定量的综合评价；申震[3]将水质模糊数学评价法和常规水质评价法进行了对比，发现模糊评价法更多考虑了污染物权重，评价结果更为客观。闫佰忠[4]提出了基于随机森林模型的评价方法评价了安阳市 2017 年 8 个地下水监测点水质情况，发现随机森林模型的水质评价模型能够准确评价水质的同时，拥有更高训练效率与稳定性。

已见报道的研究方法虽考虑了水的单一属性与综合属性，实现综合评估，但在予以权重分配的时候容易受到数据噪声或者专家赋权造成的主管臆断，影响了评估的准确性[5,6]。M 河作为中国南方某城市的水源，为沿途 1 400 余万人提供饮用水水源，在生活、生产建设等各方面发挥重要作用。本文依据《地表水环境质量标准》(GB 3838－2002)采用模糊综合评价和随机森林评价的方法，在获得 M 河水质评价结果的基础上，针对水质指标重要性进行了深入分析，以期为 M 河水污染溯源、水质评价、水污染防治及水资源管理提供了理论基础与技术支撑。

1　数据及研究方法

1.1　研究区域

本文研究 M 河所在区域为城市区域，年平均气温 16～20 ℃，年降水量 1 500～2 000 mm。共设有四个监测站(WWP、FWP、SWP 和 CWP)，见图 1 所示。

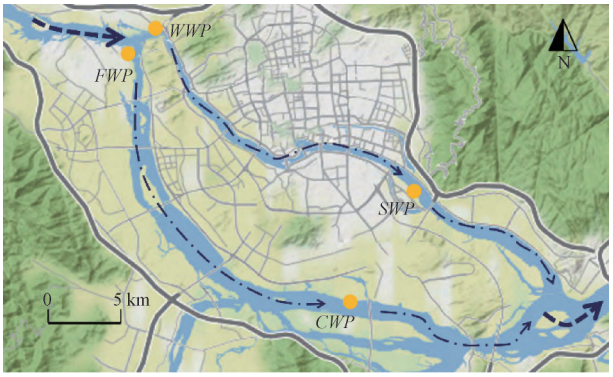

图 1　研究区位置及监测断面分布

Fig.1　Location map of study area and distribution monitoring sites

1.2　水质数据采集

根据《地表水环境质量标准》(GB 3838－2002)作为流域监测原则和方法，于 2010 年 10 月至 2020 年 8 月对 4 个监测断面每月进行 1 次水样采集。

1.3　水质数据分析

本研究采用模糊综合评价和随机森林的方法进行水质评价分析。

1.3.1　模糊综合评价法

(1)构建评价集合。采用 GB 3838－2002 为评价标准，基于水质监测中主要的污染指标作为评价因子并划分的水质级别，建立因子集和评价集。

(2)建立隶属度函数及模糊矩阵。通过隶属度函数[3]确定评价指标的实测值与相对应的评价标准值构建模糊矩阵 $\boldsymbol{R}$，隶属度函数公式见式(1)～式(3)：

$$j=1\text{ 时},r_{ij}=\begin{cases}0, & X_i \geqslant S_{i(j+1)}\\ \dfrac{S_{i(j+1)}-X_i}{S_{i(j+1)}-S_{ij}}, & S_{ij}<X_i<S_{i(j+1)}\\ 1, & X_i \leqslant S_{ij}\end{cases} \tag{1}$$

$$1<j<n\text{ 时},r_{ij}=\begin{cases}0, & X_i \leqslant S_{i(j-1)}\text{ 或 }X_i \geqslant S_{i(j+1)}\\ \dfrac{X_i-S_{i(j-1)}}{S_{ij}-S_{i(j-1)}}, & S_{i(j-1)}<X_i<S_{ij}\\ \dfrac{S_{i(j+1)}-X_i}{S_{i(j+1)}-S_{ij}}, & S_{ij} \leqslant X_i<S_{i(j+1)}\end{cases} \tag{2}$$

$$j=n\text{ 时},r_{ij}=\begin{cases}0, & X_i \leqslant S_{i(j-1)}\\ \dfrac{X_i-S_{i(j-1)}}{S_{ij}-S_{i(j-1)}}, & S_{i(j-1)}<X_i<S_{ij}\\ 1, & X_i \geqslant S_{ij}\end{cases} \tag{3}$$

其中，X_i 表示 i 因素的实测值；S_{ij} 表示 i 因素对应 j 级的标准值，r_{ij} 表示 i 因素对 j 级的隶属度。

由式(1)～(3)隶属度函数构建模糊矩阵，见式(4)：

$$\boldsymbol{R}=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mn} \end{bmatrix} \tag{4}$$

(3)构建模糊权重矩阵。采用超标倍数赋权法归一化计算各个评价对象的权重值 a_i，见式(5)，并构建模糊权重矩阵 A，见式(6)。

$$a_i=\frac{X_i/S_i}{\sum_{i=1}^{n}\frac{X_i}{S_i}} \tag{5}$$

其中，S_i 表示 i 因素评价标准的均值。

$$A=(a_1,a_2,,\cdots,a_n) \tag{6}$$

(4)构建模糊综合评价结果矩阵。将模糊权重矩阵 A 和模糊关系矩阵 R 相乘，求得分级模糊综合评价结果矩阵 B，见式(7)。

$$\boldsymbol{B}=\boldsymbol{A}\times R=(b_1,b_2,\cdots,b_n) \tag{7}$$

1.3.2 随机森林

随机森林[5,7]是基于分类回归树原理构建的一种有监督学习的集成模型，即利用 bootstrap 抽样方法进行随机抽取等概率的结构变量，通过对各组合样本进行决策树构建，从而寻找得分最高的分类结果，以作为分类的优选。随机森林算法中 OOB 分值愈高，表明该参数重要性愈大。因此，OOB 可作为水质评价模型中各项指标因子重要性分值的评价参数，依据其数值大小，可以判定水质各指标特征重要性并对其进行排序。随机森林法不仅可有效分析样本间的差别，而且在部分水质参数缺失的条件下，仍可保持准确度较高的水质评价结果[7]。

2 结果与分析

2.1 单一水质指标分析

基于 M 河自上游到下游连续 10 年间(2011—2020 年)的每月水质监测大数据分析结果表明，BOD_5、铜、锌、硒、砷、汞、石油类、挥发酚等均低于检出限，对 M 河水质评价无指示作用。因此，本研究选取 pH、水温(WT)、DO、总氮(TN)、氨氮(NH_3-N)、硝酸盐(NO_3^--N)、总磷(TP)、高锰酸盐指数(COD_{Mn})、氯化物(Cl^-)、硫酸盐(SO_4^{2-})、粪大肠菌群(*F. coli*)、铁(Fe)、锰(Mn)和氟化物(F^-)等水质参数开展 M 河水质评价研究分析。

M 河 pH 值分布范围 6.47～7.60，64%的样品 pH<7，处于相对偏弱酸性的状态，上游到下游水质 pH 无明显变化。M 河总氮长期处于严重超标状态(平均浓度 1.52 mg/L)，且从上游到下游亦无明显变化，说明研究区域段 M 河无总氮类污染物输入，由此表明 M 河总氮污染超标可能是研究区域上游段的农村乡镇地区污染物排入导致。SO_4^{2-}、NO_3^-、F^-、COD_{Mn}、Cl^-、NH_3-N 总体符合地表水Ⅲ类水水质标准。DO、COD_{Mn}、TP、*F. coli*、NH_3-N、Fe 和 Mn 在不同程度上超过了Ⅲ类标准，其中 Fe 和 Mn 的浓度分布范围分别为 1.26～3.2 mg/L 和 0.16～1.52 mg/L，且在研究流域段发生了明显的数值升高，且 Fe、Mn 之间高度相关(相关系数为 0.8，见图 2)，这表明 Fe 和 Mn 可能由同源污染导致。水温和 DO 存在高度负相关(相关系数为－0.63，见图 2)，这表明温度是水中藻类生长及其生化反应的重要影响因素，从而影响 M 河中 DO 的含量，即水温越高 DO 越低。

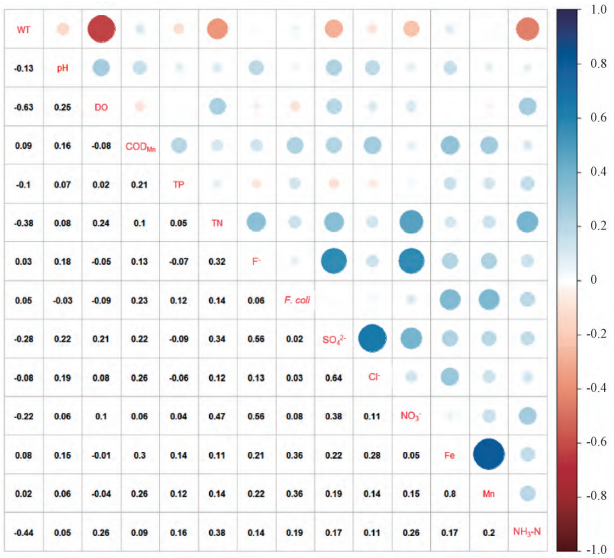

图 2 M 河水质指标相关性

Fig.2 Correlation of water quality parameters of M river

综上所述，M 河存在部分水质指标存在不同程度的超标问题，这有可能是因为 M 河上游流域现代农业发展迅速、化肥农药的滥用和污水未做处理排入江河所致。与此同时，随着 M 河流域工业化、城镇化的不断推进，采矿、城市建设、交通发展等因素导致工业废水非法排放点增多，导致 Fe 和 Mn 污染

物增加。

2.2　水质模糊综合评价

M 河四个监测站(WWP、FWP、SWP 和 CWP)的水质评价集合,选择在水质监测结果中主要的污染指标作为评价因子(表 1),其因子集 U={ DO, COD_{Mn}, TP, TN, *F. coli*, NH_3-N },评价集 V={ Ⅰ、Ⅱ、Ⅲ、Ⅳ、Ⅴ }。

表 1　地表水环境质量标准限值

Tab.1　Environmental quality standard and ranking for surface water

水质参数	Ⅰ类	Ⅱ类	Ⅲ类	Ⅳ类	Ⅴ类	Ⅰ～Ⅴ类标准平均值
DO/(mg·d^{-1})	7.5	6	5	3	2	4.7
COD_{Mn}/(mg·d^{-1})	2	4	6	10	15	7.4
TP/(mg·d^{-1})	0.02	0.1	0.2	0.3	0.4	0.204
TN/(mg·d^{-1})	0.2	0.5	1	1.5	2	1.04
F. coli/个	200	2 000	10 000	20 000	40 000	14 440
NH_3-N/(mg·d^{-1})	0.15	0.5	1	1.5	2	1.03

采用最大隶属度原则确定河流断面的水质类别。基于各监测点连续 10 年水质监测数据,M 河各监测断面的模糊综合评价分析结果表明,上游水质明显优于下游,其中监测点 FWP、WWP、SWP、CWP 中Ⅰ类水占比分别为 26%、16%、9%和 8%;监测点 CWP 中Ⅳ类和Ⅴ类水质占比 72%,说明 CWP 断面受污染严重(见图 3)。

图 3　M 河模糊综合评价

Fig.3　Fuzzy comprehensive evaluation of the M river

通过超标倍数赋权法归一化可获得 M 河六项水质指标的模糊权重矩阵,从各水质指标的权重分配情况比较出各监测点的主要污染水质指标(见图 4)。结果表明 TN 是污染权重最高的水质指标,其主要污染源可能来自农药化肥施用、农村生活污染等,DO 的模糊权重从上游监测点 FWP 到下游监测点 CWP 降低(由 26%下降至 19%),说明 M 河上游到下游的水质特性发生了变化;下游地区 *F. coli* 的权重增大(由 10%增加至 27%),说明下游存在的微生物污染导致水质出现一定程度的恶化。模糊综合评价水质指标权重的变化验证了 M 河研究区域可能存在城镇生活污染源或工业污染源等相关污染源影响水质变化。

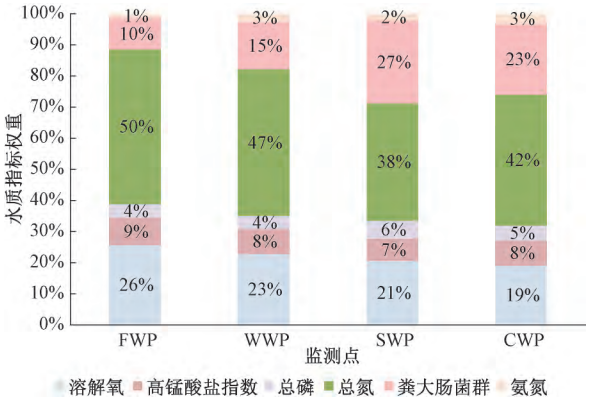

图 4　M 河模糊综合评价水质指标权重

Fig.4　Water quality parameter weights of the M river

2.3　随机森林水质评价

随机森林评价法能够客观地反映水质的实际状况,有效分析得出水体污染程度、水质类别和水质评价中的关键水质指标,在水质评价方面应用广泛。

基于 10 年(2011－2020 年)的 M 河水质监测数据,以各水质指标作为特征参数,依据 GB 3838－2002 得到的地表水水质分类作为分类结果,将水样划分训练集和测试集(4∶1),训练得到随机森林水质评价模型,模型准确率(所有预测正确的样本占总样本的比重)为 99%。结果表明总氮特征重要性最大,说明其是 M 河水质评价的关键影响因子;粪大肠菌群的特征重要性次之,特别是在 SWP 河 CWP 监测点,粪大肠菌群特征重要性明显增大,与模糊综合评价的结果吻合,验证了研究区域存在粪大肠菌群污染的问题;溶解氧和高锰酸盐指数的特征重要性占比为 10%～20%,对水质评价具有一定的影响;而 TP 和 NH_3-N 的特征重要性较小,说明其在水质评价中影响较低。随机森林水质评价模型证明模糊综合评价模型评价结果的正确性,说明研究区域存在相关的污染源问题,表明需要重点排查研究区域上游城镇生活污水处理厂、农业面源污染等总氮污染源以及研究区域中下游微生物污染源等异常

情况。

随机森林水质评价模型在本研究中水质分类和关键水质指标选择的问题上表现出优异的性能。同时，前人的研究也表明以随机森林为代表的集成学习算法相较于传统机器学习算法，表现出更高的模型准确率，可以优先考虑用于未来的水质监测和及时提供水质预警[8]。

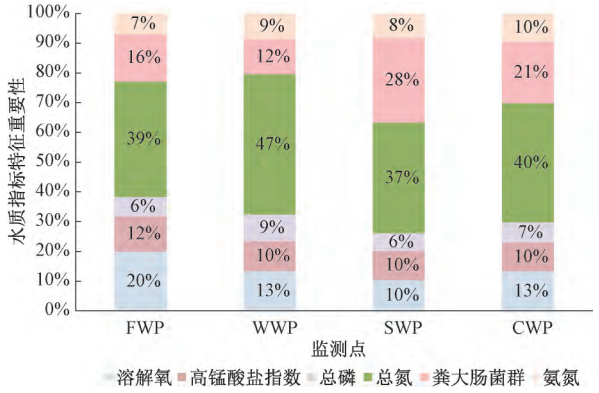

图 5 M 河水质随机森林评价模型指标特征重要性

Fig.5 The feature importance of water quality parameter of water quality evaluation model based on random forests

3 结论与展望

(1)水质污染。M 河超过一半的水样低于Ⅲ类水水质标准，主要的污染指标为总氮、粪大肠菌群、溶解氧、铁和锰。

(2)水质模糊综合评价。下游水体污染明显高于上游，下游水体Ⅳ类和Ⅴ类占比 72%；DO 的模糊权重自上游的 26%降低至下游的 19%；*F. coli* 的模糊权重自上游的 10%增大至下游的 27%，说明下游地区存在微生物污染，致使水质劣变。

(3)随机森林水质评价模型。建立的随机森林水质评价模型准确率为 99%，筛选出 M 河中重要的水质评价指标分别为总氮、粪大肠菌群和溶解氧。

本研究方法和成果已被当地政府采纳并通过对 M 河采取逐批、分段治理，M 河水质已经得到很大的改善。因此，本研究为地表水污染物溯源、水质评价方法、水污染防治及水资源管理提供了理论基础与技术施策借鉴。

参考文献

[1] 薛伟锋，褚莹倩，吕莹，等. 基于主成分分析和模糊综合评价的地下水水质评价——以大连市为例 [J]. 环境保护科学，2020，46(5)：87-92.

[2] 张洪伟，周添红，张国珍，等. 模糊综合评判法在地表水水质评价中的应用 [J]. 地下水，2017，39(1)：83-86.

[3] 申震. 模糊数学在水质评价中的应用 [J]. 市政技术，2017，35(6)：104-106.

[4] 闫佰忠，孙剑，安娜. 基于随机森林模型的地下水水质评价方法 [J]. 水电能源科学，2019，37(11)：66-69.

[5] 张颖，高倩倩. 基于随机森林分类算法的巢湖水质评价 [J]. 环境工程学报，2016，10(2)：992-998.

[6] 李婧，唐敏，梁亦欣. 2015—2018 年河南省辖海河流域水质改善效果评价 [J]. 环境工程，2020，38(5)：60-64.

[7] 董艳玲. 基于随机森林的北票市水质评价模型及应用 [J]. 水科学与工程技术，2018(3)：15-18.

[8] CHEN K，CHEN H，ZHOU C，et al. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data [J]. Water Research，2020.

§ 通信作者：李伟英，女，1963 年出生，安徽人，博士，教授级高级工程师。主要研究方向为饮用水安全保障理论与技术、膜法(金属膜)水处理理论与技术研究、供水系统水质生物安全评价与控制技术研究等。

通信处：200092 上海市杨浦区同济大学四平路校区明净楼 413A 室

E-mail：123lwyktz@tongji.edu.cn

收稿日期：2021－09－10