基于随机森林的水质监测指标预测

李旭杰[1,2],史 灵[2,3],花思洋[3],孙 颖[2,4],黄凤辰[2]
(1. 河海大学 海洋与近海工程研究院,江苏 南通 226300; 2. 河海大学 计算机与信息学院,江苏 南京 210098;
3. 钛能科技股份有限公司,江苏 南京 211806; 4. 江苏开放大学 信息工程学院,江苏 南京 210017)

摘要:通过采集2020年6月至2021年6月南京市秦淮新河代表站的DO、WT、pH、COD、NH_3-N、TUR 6类水质监测指标数据,利用Pearson相关系数对监测指标间的相关程度进行分析,从而得到各监测指标间的相关系数,进一步通过多元线性回归算法得到高度相关的参数指标间的统计关系,利用回归方程的形式表示监测变量间的因果关系,最后通过随机森林算法利用水质监测中的自变量指标实现对因变量指标的预测,达到减少监测项目从而降低监测成本的目的。研究结果表明因变量水质监测指标的预测值和实际值几乎重合,有效说明随机森林模型能够实现因变量水质监测指标的准确预测。

关键词: Pearson相关系数; 多元线性回归算法; 随机森林模型; 秦淮新河
中图分类号:X522 **文献标识码:**B **文章编号:**1007-7839(2022)05-0006-0005

Water quality monitoring indicators prediction based on random forests

LI Xujie[1,2,3], SHI Ling[2,4], HUA Siyang[4], SUN Ying[2,5], HUANG Fengchen[2]
(1. *Institute of Ocean and Offshore Engineering, Hohai University, Nantong* 226300, *China*;
2. *College of Computer and Information, Hohai University, Nanjing* 210098, *China*;
3. *Talent Science and Technology Co., Ltd., Nanjing* 211806, *China*;
4. *School of Information Engineering, Jiangsu Open University, Nanjing* 210017, *China*)

Abstract: Through collecting the data of six kinds of water quality monitoring indicators of Qinhuai New River representative station from June 2020 to June 2021, including dissolved oxygen, water temperature, PH value, chemical oxygen demand, ammonia nitrogen and turbidity. Pearson correlation coefficient was used to analyze the correlation degree among monitoring indicators, so as to obtain the correlation coefficient among monitoring indicators. Further through multiple linear regression algorithm was highly related to the statistical relationship among the parameters, using the regression equation in the form of said monitoring the causal relationship among variables, finally by random forest algorithm using water quality monitoring in the independent variable of the dependent variable indicators forecast, achieve the goal of reducing monitoring project so as to reduce the monitoring cost. The results show that the predicted value and the actual value of the dependent variable water quality monitoring index almost coincide, which effectively indicates that the random forest model can achieve the accurate prediction of the dependent variable water quality monitoring index.

收稿日期:2021-12-28
基金项目:江苏省水利科技项目(2020028);南通市社会民生科技项目(MS22021042);广东省水利科技创新项目(2020-04);江苏省教育厅未来网络科研基金资助(FNSRFP-2021-YB-7);中国科学院无线传感网与通信重点实验室开放课题(20190914)
作者简介:李旭杰(1979—),男,副教授,博士,主要从事水利信息化技术研究。E-mail:lixujie@hhu.edu.cn

Key words: Pearson correlation coefficient; multiple linear regression algorithm; random forest model; Qinhuai New River

1　概　述

本文以江苏省南京市秦淮新河为代表站进行研究。秦淮新河属秦淮河水系,起于河定桥经西善桥至金胜村入江口,总长16.8 km,是下游入江分洪道的一条重要通道[1]。选取2020年6月至2021年6月的水质监测各类监测指标数据,利用Pearson相关系数对指标间进行相关性分析,得到各监测指标间的相关系数,对变量关系间的强弱进行有效度量,对影响水质的主导因素进行识别,然后采用多元线性回归算法进一步分析水质指标间的统计关系[2],确定变量之间的因果关系,并对多元线性回归算法的可信程度进行检验。根据符合评价标准的多元回归方程,通过随机森林模型用自变量指标对因变量指标做出进一步预测,若之后监测到的水质指标值与预测值相比有较大差异,则可以说明该河段水质有较大变化,可起到预警作用[3],能够对水质可能出现的问题进行有效预防,构建一个高效的水质监测预测模型,能够为秦淮新河的水环境保护提供科学指导依据。

2　国内外研究现状

传统的水质监测一般是进行人工操作,这种监测方法不能及时、准确地获得水质不断变化的动态数据。而通过各类监测水质指标的传感器实现对水体中的COD、NH_3-N、pH值等进行在线精确监测,能有效提升水质监测效率,避免手工测定的耗时费力甚至不精确的一系列缺陷,结合计算机以及通信等技术手段,可以对所采集到的数据进行分析处理,为进一步产生和研究数据奠定基础。

2017年,郑德论[4]通过监测汕头湖沟中上游河段水体的水质状况,采用单因子评价方法确定该河段水体的主要污染物。2018年,汤云[5]针对闽江流域的多项水质指标监测数据,利用小波分解、遗传算法改进的BP神经网络方法,分析闽江流域内水质时空分布特征并解析污染源,构建基于小波分解和遗传算法改进的BP神经网络的水质预测模型。2019年,杨娜等[6]以雄安新区白洋淀水质为研究对象,用灰色聚类分析法并结合变异系数法赋权,对其水质进行分级与评价,为白洋淀的综合治理提供更加客观科学依据。2020年,秦无双等[7]对蓬溪县主要地表水体进行水质现状分析,采用主成分分析法确定了主要污染因子。2021年,国内杨志民[8]针对契爷石水库进行水质监测,采用综合水质评价法和模糊综合评价法对水库水质进行综合评价。

水质自动监测在我国出现的时间较晚,尚处于起步阶段。就现阶段而言,我国水质监测较国外尚存在一定的差距。目前对水质监测数据的自动采集研究比较多,集中在对水质的在线自动监测上,而对于水质监测所采集到的数据进行分析处理的研究还不多,此方面研究有待拓展。本文对水质监测数据进行分析,构建水质监测指标预测模型,提供一定水质监测的科学依据。

3　水质监测指标预测模型构建

构建水质监测指标预测模型体系,采用统计分析方法,包括相关性分析以及多元回归分析,结合随机森林模型,对水质监测指标进行主导性因素指标的相关预测,构建预测模型见图1。对获取水质监测项目监测过程中所用的各类监测指标数据进行清洗,剔除无效、异常数据。无效、异常数据主要指超过各类水质监测指标传感器的测量范围内的异常数据。

3.1　水质监测指标相关性分析

利用相关系数对各类水质监测指标进行相关性分析,常见的相关系数主要有Pearson相关系数、Spearman相关系数以及Kendall秩相关系数,其中Pearson相关系数适用于衡量线性相关关系,针对其适用性,本文采用Pearson相关系数来对6类水质监测指标进行相关程度强弱的度量,定义其公式为

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{1}$$

式中:n为样本量;x_i和y_i分别为2个监测指标的变量值;$\bar{x}$和$\bar{y}$分别为x_i和y_i样本的平均值。

图2是6类水质监测指标间Pearson相关系数矩阵热力图,根据热力图颜色的深浅分别可以得到各监测指标间的相关程度强弱。其中,COD_{cr}表示化学需氧量,NH_3-N为氨氮,DO为溶解氧,WT为水温,

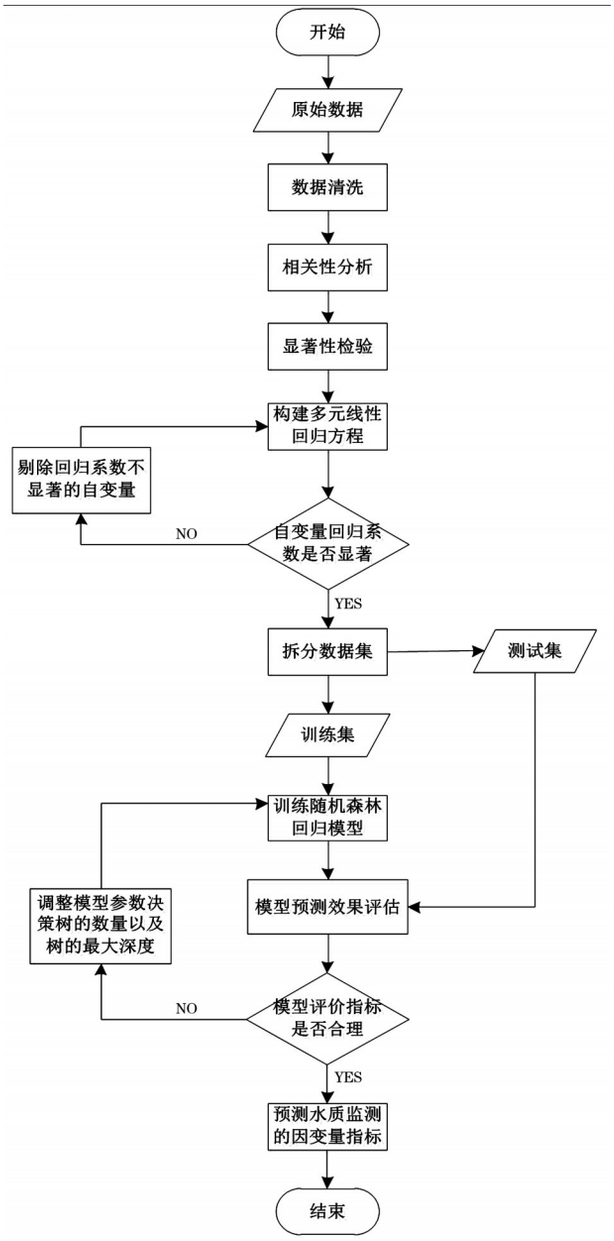

图1 水质监测指标预测模型

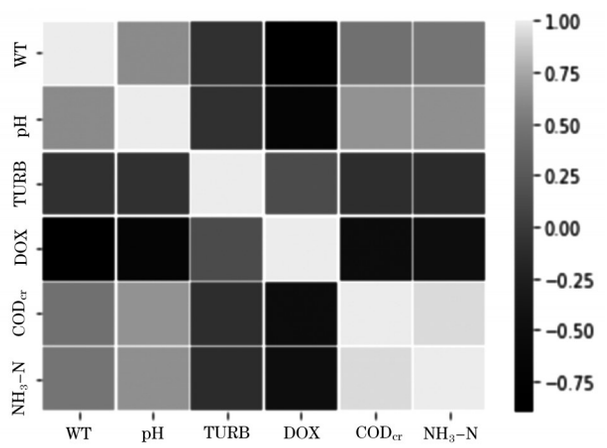

图2 Pearson相关系数矩阵热力

TURB为浊度。

Pearson相关系数矩阵如表1所示。其中，Pearson相关系数的绝对值结果越接近于1表示变量之间的相关性越强，结果越接近于0表示变量之间的相关性越弱。其绝对值结果在0~0.3之间，呈现弱相关性；在0.3~0.5之间，呈现低相关性；在0.5~0.8之间，呈现显著相关性；在0.8~1之间，呈现高度相关性[9]。根据表1的分析结果可以得出高度相关的变量有两对，一对是WT和DO，2个监测指标间的Pearson相关系数为-0.8965，明显呈现高度负相关性；另一对是COD_{cr}指标和NH_3-N指标，它们的相关系数为0.9478，明显呈现高度正相关性。

然后，采用显著性检验的方法对Pearson相关系数进行检验，验证各监测指标间的相关性非偶然因素引起，所得结果能够代表总体指标数据上的相关程度。在本文中，显著性检验的P值均小于选定的显著性水平0.05，故变量之间的相关性都通过显著性检验，显著性检验结果P值列于表2。

3.2 多元线性回归算法

相关性分析是回归分析的基础和前提，而回归分析则是认识变量间相关程度的具体形式。采用构建多元线性回归方程的方法可进一步得到监测指标间相关程度的具体形式。本文通过回归方程的形式，进一步分析水质指标间的统计关系。

利用多元回归算法，设因变量为y，k个自变量分别为$x_1, x_2, \cdots, x_k$，描述因变量y如何依赖自变量$x_1, x_2, \cdots, x_k$和误差项ε的方程。多元线性回归方程可表示如下：

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k+\varepsilon \quad (2)$$

式中：β_0，β_1，β_2，$\cdots$，β_k是模型的参数，ε为误差项，式(2)表明y是$x_1, x_2, \cdots, x_k$的线性函数加上误差项ε。误差项反映了除$x_1, x_2, \cdots, x_k$与y的线性关系之外的随机因素对y的影响，是不能由$x_1, x_2, \cdots, x_k$与y之间的线性关系所解释的变异性[10]。

对得到的多元线性回归方程，剔除回归系数异常以及不显著的自变量，此时的多元线性回归方程可得以成立。回归系数反映的是回归方程中表示自变量x对因变量y影响大小的参数，异常回归系数为回归系数的正负号与Pearson相关系数相反的数值，不显著的回归系数是指不符合回归系数检验的数值。根据高度相关的变量对可得到2个多元线性回归方程为

$$\rho(COD_{cr})=2.5255\times pH\text{值}+23.0224\times\rho(NH_3\text{-}N) \quad (3)$$

表1　Pearson相关系数矩阵

	WT	pH	TURB	DO	COD_{cr}	NH_3-N
WT	1.0000	0.6090	−0.0811	−0.8965	0.4541	0.4777
pH	0.6090	1.0000	−0.0750	−0.6593	0.6553	0.6382
TURB	−0.0811	−0.0750	1.0000	0.1463	−0.0999	−0.1146
DO	−0.8965	−0.6593	0.1463	1.0000	−0.5045	−0.4837
COD_{cr}	0.4541	0.6553	−0.0999	−0.5045	1.0000	0.9478
NH_3-N	0.4777	0.6382	−0.1146	−0.4837	0.9478	1.0000

表2　Pearson相关系数显著性检验结果P值

	WT	pH	TURB	DO	COD_{cr}	NH_3-N
WT		2.19×10^{-83}	0.02	6.34×10^{-288}	1.83×10^{-42}	2.06×10^{-47}
pH	2.19×10^{-83}		0.03	3.62×10^{-102}	1.56×10^{-100}	7.71×10^{-94}
TURB	2.10×10^{-2}	3.29×10^{-2}		2.92×10^{-5}	4.43×10^{-3}	1.08×10^{-3}
DO	6.35×10^{-288}	3.62×10^{-102}			1.65×10^{-53}	1.01×10^{-48}
COD_{cr}	1.83×10^{-42}	1.56×10^{-100}		1.65×10^{-53}		
NH_3-N	2.06×10^{-47}	7.71×10^{-94}		1.01×10^{-48}		

ρ(DO)=−0.1628×WT值−0.7034×pH值+0.3417×TURB值　　(4)

3.3　随机森林模型

构建多元线性回归方程后,对整个数据集进行拆分,得到训练集和测试集:训练集用于训练随机森林回归模型,测试集用于模型预测效果的评估。在本文中,取测试集样本数为数据集总样本数的25%。建立随机森林回归模型,采用5折交叉验证方法利用训练集对模型进行训练。相关参数设置随机值random_state=0,通过5折交叉验证寻找到模型的最佳参数,不重复抽样将原始数据随机分成5份;每次挑选其中1份作为测试集,剩余4份作为训练集用作模型训练;重复该步骤5次,使得每个子集拥有一次作为测试集的机会,其余机会作为训练集;计算5组测试结果的平均值作为模型的准确率[11-12]。

基于上述方法,针对COD_{cr}指标的预测模型可得到决策树的数量n_estimators=50,树的最大深度max_depth=7;针对DO指标的预测模型得到决策树的数量n_estimators=150,树的最大深度max_depth=6。然后,利用测试集对随机森林回归模型进行预测效果评价,可求得该随机森林回归模型的3种评价指标的值分别为:均方根误差RMSE=0.2883、平均绝对误差MAE= 0.1813、确定系数R^2=0.9831。当随机森林回归模型的确定系数R^2已经达到最接近于1,则停止对随机森林回归模型的参数继续调整。根据训练得到的最佳随机森林回归模型,利用水质监测指标中自变量监测指标对因变量指标进行预测。

4　实验结果分析

根据随机森林模型,得到根据水质监测指标中的自变量指标对因变量指标的预测,由水质监测指标的相关性分析以及通过构建多元线性回归方程,得到高度相关的变量对及其因果关系。由训练结果得到相应的随机森林回归模型预测值与水质监测指标中的因变量指标实际值对比图。其中,化学需氧量监测指标实际值与预测值的对比图,见图3,溶解氧监测指标实际值与预测值的对比图,见图4。图3~4中因变量水质监测指标的实际值为蓝色线,因变量水质监测指标的预测值为黄色线。从图3~4中明显可见因变量水质监测指标的预测值和实际值几乎重合,有效说明本文提出的基于随机森林回归模型的水质监测指标预测方法能够实现因变量水质监测指标的准确预测。

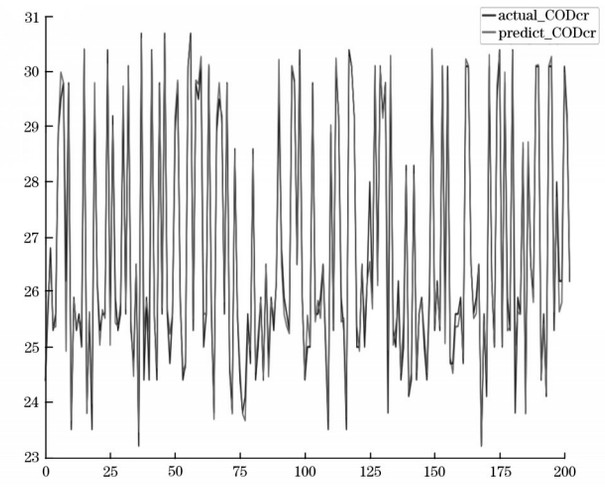

图3 COD监测指标对比

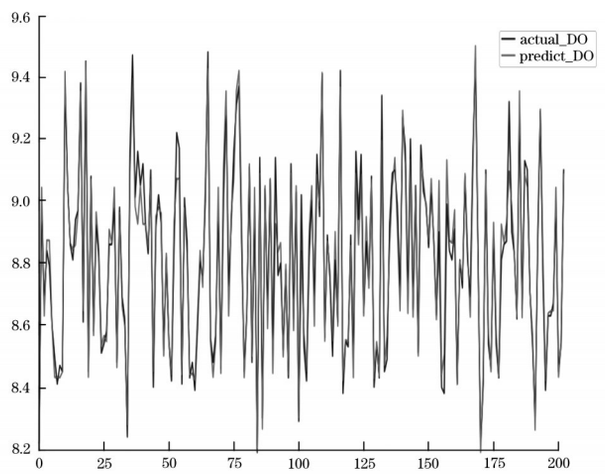

图4 DO监测指标对比

5 结 语

近年来我国水质监测发展迅速，对水质状况进行监测时往往监测的指标种类众多，监测成本较高且信息量巨大，难以从中提取有效信息，对数据进行有效分析势在必行。为有效降低监测成本，对监测指标项目进行合理降维，利用Pearson相关系数对指标间进行相关性分析，对变量关系间的强弱进行有效度量，但由于并未对变量之间的关系进行固化形成模型，无法利用这种关系对数据进行预测，需要进一步进行回归分析，在实际应用中一个参数指标往往受到多个参数指标的影响，多元线性回归算法易于实施，具有较大的应用前景。

在水质监测对各类指标进行监测的实践中，利用多元线性回归得到高度相关的监测指标间统计方程，可以准确得知自变量指标和因变量指标之间的关系，对因变量指标进行有效预测可减低监测成本。利用随机森林模型中的回归模型对监测指标进行预测，相对于其他模型而言，随机森林回归模型具有预测准确度高、泛化能力强的优势。实验结果也能够有效表明随机森林回归模型可利用因变量水质监测指标实现对自变量水质监测指标的准确预测，随机森林回归模型在水质监测指标分析预测的应用研究具有重要意义。

参考文献：

[1] 邵园园，戴庆云，蒋涛，等. 秦淮河流域水生态（环境）调度研究[J]. 江苏水利，2020(12)：44-47，69.

[2] 马振，周密. 聚类分析在秦淮河水质指标相关性研究中的应用[J].水文，2018，38(1)：77-80.

[3] WEIQI H E. Water quality monitoring in a slightly-polluted inland water body through remote sensing-case study of the Guanting Reservoir in Beijing, China [J]. Frontiers of Environmental Science & Engineering in China，2008，2(2)：163-171.

[4] 郑德论. 汕头龙湖沟河段水质状况监测与评价对策[J]. 广州化工，2017，45(20)：117-119.

[5] 汤云. 闽江流域水质时空分布特征分析及水质预测[D]. 福州：福州大学，2018.

[6] 杨娜，陈国鹰. 基于灰色聚类分析的白洋淀水质评价[J]. 科技创新与应用，2019(34)：43-45.

[7] 秦无双，胡艳丽，肖娟，等. 蓬溪县主要地表水体水质现状及水质改善对策[J]. 绿色科技，2020(12)：61-63.

[8] 杨志民. 契爷石水库水质监测评价及水质预测研究[J]. 中国水能及电气化，2021(11)：64-68.

[9] 刘玲，孙玉豪. 基于Pearson & BIM云的工程造价信息数据共享研究[J]. 工程经济，2021，31(8)：20-24.

[10] 贾俊平，何晓群，金勇进. 统计学[M].6版.北京：中国人民大学出版社，2014.

[11] 彭一晋. 基于智能手机软件数据对用户性别的预测[D]. 大连：大连理工大学，2021.

[12] 刘海军，刘韵锋，陈侨，等. 基于随机森林和贝叶斯优化的TiO_2光催化污染物降解速率预测模型研究[J]. 信息记录材料，2021，22(8)：25-27.