DOI:10.20040/j.cnki.1000-7709.2023.20222548

基于随机森林回归模型和高频数据的鄱阳湖子湖电导率预测

刘丽贞[1],黄 琪[2],迟殿委[3],方朝阳[2],楚明航[1]

(1. 江西省科学院微生物研究所,江西 南昌 330096; 2. 江西师范大学鄱阳湖湿地与流域研究教育部重点实验室,江西 南昌 330022; 3. 烟台理工学院人工智能学院,山东 烟台 264005)

摘要:电导率是衡量水质的重要参数,高频监测获取水体中电导率对水质管理具有重要作用。由于野外条件的变化复杂性引起设备故障导致数据缺失时有发生,为优化野外监测体系和插补缺失数据,基于高频监测获取的气象和水体物理指标,结合机器学习模型,预测水体中电导率值。结果表明,随机森林回归模型预测效果最优,其决定系数 R^2 可达 0.996,均方根误差 R_{RMSE} 为 1.31 μS/cm,平均相对误差 M_{MRE} 为 0.38%;pH 值贡献率最大,是影响电导率的主导因素。研究结果利于优化野外高频监测系统平台,健全高频监测数据,为水质管理提供科学依据。

关键词:电导率;随机森林回归模型;高频监测数据;鄱阳湖

中图分类号:X83;[TV11] **文献标志码**:A **文章编号**:1000-7709(2023)10-0050-04

1 引言

电导率是反映水体离子浓度的指标,通常用于反映水质的污染状况,表征人类活动对水质的影响,并对藻华暴发区位具有指示作用。电导率也可以作为反映淡水水生生物群落结构差异的水质变量,如浮游植物功能群、浮游动物群落结构、敏感类群落的分布、栖息藻类生长和底栖动物群落组成等。基于野外数据建立大型底栖动物电导率水质基准[1],并基于水体电导率指示太湖长时间污染过程[2],因此水体电导率的持续监测和应用在水质监管和污染防控中具有十分重要的指示意义。连续高频监测日益成为水质管理的关键手段,目前获取电导率高频数据可结合浮标于野外原位获取[3]。然而随着水质监测传感器搭载的增多,后期维护成本和不定因素也会增加。野外实际环境的复杂性可导致高频监测相关设备(如发生故障)传回的数据缺失,若设备未得到及时维护,会影响后续数据挖掘和分析。因此,健全高频监测数据、完善野外高频监测系统平台是关键。鉴于水质和气象因素存在联系,且多个水质因素间存在关联,在野外布设浮标搭建,选择监测指标时,可针对监测目标利用机器学习算法用于估算其他水质指标,健全高频监测数据。已有利用机器学习如使用极限学习机预测土壤电导率的研究,如董凡[4]利用小波-神经网络混合模型预测沿海地区地下水电导率;高鹏等[5]基于气象因素利用广义回归神经网络和长短期记忆神经网络预测了土壤电导率。基于机器学习结合气象因素来预测电导率是可行的,但目前鲜有研究应用气象指标结合水体理化指标来预测自然水体电导率变化趋势。为此,本文利用气象参数和水体简易物化参数数据集预测水体电导率,以期发展更优的高频水质监测体系,为水质监控提供科学依据。

2 研究区域和数据

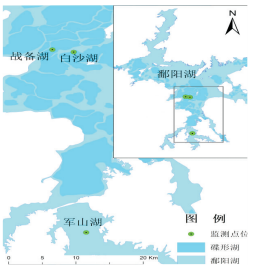
在鄱阳湖周边水体白沙湖、东湖和军山湖湖体内各布设 1 个野外高频监测浮标,具体布置点位见图 1(a)。浮标监测实景图见图 1(b)。监测

收稿日期:2022-12-07,**修回日期**:2023-01-06

基金项目:江西省重点研发计划(20212BBG73014, 20192ACB70014);江西省主要学科学术和技术带头人培养计划——青年人才项目(20212BCJ23034);江西省青年重点基金项目(20192ACBL21022);鄱阳湖湿地与流域研究教育部重点实验室开放基金项目(PK2019006);江西省科学院杰出青年人才培育计划(2021YSBG50004)

作者简介:刘丽贞(1987-),女,副研究员,研究方向为水环境生态,E-mail:woliulizhen2007@126.com

通讯作者:黄琪(1985-),男,博士、助理研究员、硕导,研究方向为自然地理学,E-mail:huangq@jxnu.edu.cn



(a) 鄱阳湖野外高频监测浮标点位

(b) 浮标监测实景图

图 1　鄱阳湖野外高频监测浮标点位和浮标监测实景图

Fig. 1　High frequency monitoring buoy location in Poyang Lake and real picture of buoy monitoring

指标包括风向(WD)、风速(WS)、相对湿度(RH)、pH 值(pH)、水温(WT)、电导率(EC),各数据集的详细信息见表 1。各指标同步收集的数据共达 110 516 个。

表 1　数据集信息

Tab. 1　Data set information

数据监测点	监测指标	监测时间段	监测频率	数据个数/个
白沙湖	WD、WS、RH、pH、WT、EC	2018-03-04 14:19～2018-03-06 14:52、2018-03-12 10:23～18:17、2018-03-15 09:42～16:17、2018-03-15 19:21～2018-03-16 19:45、2018-10-30 17:50～2018-11-03 2:15、2018-12-03 9:53～2018-12-03 11:19、2019-01-04 14:26～2019-01-04 18:04、2019-01-07 14:11～2019-01-10 10:01、2019-05-20 16:36～2019-05-29 3:54、2019-06-26 16:35～2019-06-27 17:03、2019-06-30 9:55～2019-07-11 9:00	每 3 min 获取 1 次数据	108 477
东湖		2019-01-09 10:38～2019-01-10 10:01、2019-05-20 16:36～2019-05-20 16:51	每 1 min 获取 1 次数据	613
军山湖		2019-01-09 10:38～2019-01-10 09:48、2019-05-20 16:36～2019-05-20 16:51	每 10 s 获取 1 次数据	1 426

3　数据分析和模型精度评价方法

数据集采用 CSV 格式保存。统计分析采用 SPSS 软件进行分析,模型(包括随机森林模型(RF)、广义加性模型(GAM)、支持向量机回归模型(SVR)及多元线性回归模型(MLR))采用 Python 3 编程软件进行水体电导率预测分析,对比分析这四种模型对鄱阳湖子湖水体电导率的估算效果。分析过程使用 Pandas、Numpy、Math、Sklearn 库来实现。将指标 WD、WS、RH、pH 值、WT 作为输入自变量,输出变量为 EC。选择数据集的 75%作为训练样本用于模型构建,剩余 25%作为验证样本用于模型验证[6]。

采用决定系数(R^2)、均方根误差(R_{RMSE})、平均相对误差(M_{MRE})作为检验模拟值与实测值是否一致的判定方法。其中 R^2 值与模型精度成正比,越接近于 1,代表模型拟合精度越高;M_{MRE}、R_{RMSE} 均反映实测值与模型预测值之间差异程度,M_{MRE}、R_{RMSE} 越接近 0,代表模型实测值与预测值偏差越小,预测能力越强。

4　结果与讨论

4.1　数据描述和相关分析

鄱阳湖子湖风向均为西南风,从偏北风 2.37°变化至西北风 359.28°;风速均值为轻风 3.48 m/s,变化范围为 0.35～20.31 m/s;相对湿度均值为 84.23%,变化范围为 30%～100%;pH 值最小值为 2.68。pH 值均值为 6.57,由强酸雨降雨引起[7]。水温由冬季的 0.42 ℃变为夏季的 39.9 ℃。总体而言,鄱阳湖子湖水体电导率较低,均值为 118.91 μS/cm,低于太湖 1996 年以前水体电导率均值(239.43 μS/cm)[8],由此反映出鄱阳湖水体受污染程度较低。电导率最大值为 477.95 μS/cm,为有沉积物存在时的电导率[8],但明显小于太湖北部入湖河流最大值 1 521 μS/cm[2],说明鄱阳湖水体存在局部污染状况,需引起关注。解释变量(WD、WS、RH、pH、WT)与 EC 的线性 Spearman 相关系数分别为 0.01、0.022、−0.009、0.039、−0.035,均很小,这说明仅利用线性统计模型,并不能很好地估算电导率数值。

4.2　电导率预测模型精度评价

分别构建水体电导率的 RF 模型、GAM 模型和 SVR 模型三种机器学习算法估算模型,并引入传统 MLR 模型进行比对,模型评价指标结果见表 2。

表 2　测试集中 RF 模型及其他常用模型预测精度评价及比较

Tab. 2　Evaluation and comparison of prediction accuracy of random forest model and other commonly used models in test set

预测模型	R^2	R_{RMSE} /(μS·cm^{-1})	M_{MRE} /%	运行速度
RF 模型	0.988 0	2.12	0.46	较快,适合非线性复杂关系
GAM 模型	0.590 0	12.75	4.21	较快,精度不高
SVR 模型	0.250 0	17.47	1.66	慢,仅适合小样本数据
MLR 模型	0.079 8	18.97	4.97	较快,仅适合于线性回归

由表 2 可知,在四种建模方法中,MLR 模型验证集 $R^2=0.079\ 8$、$R_{RMSE}=18.97$ μS/cm、$M_{MRE}=4.97\%$,相比于三种机器学习模型估算效果最差,估算能力最弱。这是由生态系统中各因素之间复杂的非线性关系所致。

对比分析三种机器学习模型的验证集的 R^2、R_{RMSE}、M_{MRE} 可知,RF 模型预测精度最高,验证集中的决定系数 R^2 可达 0.988,R_{RMSE} 为 2.12 μS/cm,M_{MRE} 为 0.46%。GAM 模型预测精度次之。通过综合评估四种模型的评价参数可知,对水体电导率估算的建模效果依次为 RF>GAM>

SVR＞MLR，说明 RF 模型可精准地定量估算水体电导率。图 2 为 RF 模型数据集的预测效果。图 2 中，n 为训练数据集个数。由图 2 可知，R^2 可达 0.996，R_{RMSE} 为 1.31 μS/cm，M_{MRE} 为 0.38%。

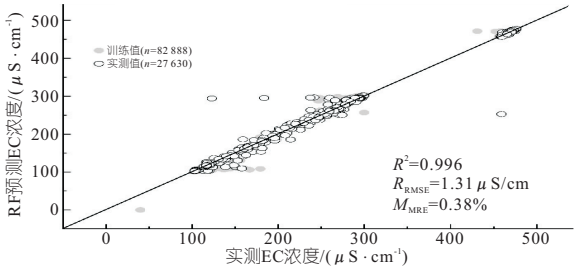


图 2 RF 模型预测值和实测值

Fig. 2 Predicted values and measured values calculated by forest regression model

此外，针对同样大的数据集，使用同一个运行软件，四种预测模型的运行速度也有所差异。其中 SVR 模型运行最慢，且预测精度较低，由此可见，SVR 算法并不适用于电导率预测。这是由于支持向量机是以小样本统计理论为基础，在解决小样本、非线性回归问题中具有显著优势[9-10]。相比之下，随机森林算法不仅具有很高的预测准确性，且对异常值和噪声均具备很好的容忍度，适合计算大规模数据，运行速度很快，在环境研究尤其是针对高频监测数据领域具有很大的应用潜力[8]。对于本文监测数据集，RF 算法只需运行不到 1 min 就完成。由此可见，RF 模型在高频监测数据估算预测中值得推广。鉴于本数据集收集了四季数据，且湖泊水体电导率变幅相对较大(101.2～477.95 μS/cm)，因此基于 RF 模型来估算水体电导率可进一步拓展到其他气候相近的同类型湖泊。可结合数据预处理如噪声数据光滑处理、优化 RF 算法进行优化等过程，以提升预测精度。扩大来自不同类型水体和季节的数据量来训练优化模型，提升泛化能力。进而优化野外监测系统平台，实时有效地监控水质，助力水生态文明建设。

4.3 解释变量重要性分析

由各解释参数的重要性(图 3)可知，RF 算法中的特征重要性参数可给出各自变量影响力的大小，其中 pH 值占的比重最大，比值为 62.1%，其次为相对湿度(20.9%)。水温和风速的影响相对较弱，分别为 4.1%、1.8%。水体电导率是溶液传导电流的能力，反映水中电解质的含量，其大小主要由溶解在水体的离子种类、浓度和水温等决定[8]。pH 值可直接影响水体离子的存在形态和含量，进而对电导率的影响较大，如尼洋河流域水体 pH 值越低可导致溶解物质变化，从而影响电导率[11]。周熠[12]发现南昌市新城区大气降水中 pH 值越小，电导率越大，降水的污染越严重。江西省全省均有酸雨污染[13]，且严重酸雨区主要受矿山废石堆的硫化矿物影响[7]。大量低 pH 值降雨冲刷流域，溶解并携带大量离子汇入鄱阳湖，提高了鄱阳湖电导率增加的风险，应引起一定的重视。此结论与太湖流域输入是引起水体电导率变化的主要外源性因素相一致[8]。进一步分析表明，水温贡献相对较小，尽管温度是影响溶液电导率变化最大的外在因素，但野外电导率传感器本身进行了温度校正[14]，缩小了设备上电导率参数随温度变化导致的不稳定差异，增加了数值的可比性。其次，鄱阳湖水温在日变化时间尺度呈“S”型变化[15]，而电导率日变化趋势不显著[16]，这也是引起水温对电导率贡献较小的原因。对于气象因子来说，相对湿度是影响电导率的重要气象因素，这是由于相对湿度往往与降雨、水面蒸发有关，相对湿度与降雨量呈正相关，与水面蒸发呈负相关，且降雨中 pH 值较低，显著影响水体电导率。

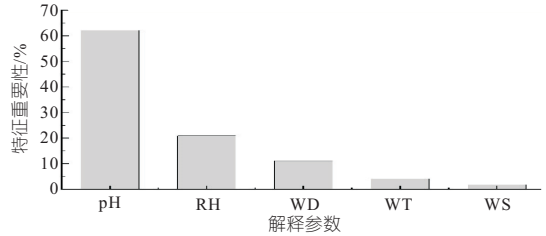


图 3 各解释参数的重要性

Fig. 3 Importance of each interpretation parameter

5 结论

a. 鄱阳湖子湖总体电导率较低，均值为 118.91 μS/cm，但最大值为 477.95 μS/cm，存在局部污染现象，应引起一定的重视。

b. 建立了鄱阳湖水体电导率估算随机森林模型，其确定系数可达 0.996，进一步分析得出，pH 值对电导率影响最大。

c. 可结合数据预处理，如噪声数据光滑处理、优化随机森林算法、扩大数据量进行优化等过程，来提高模型预测精度。

参考文献：

[1] 张远，丁森，赵茜，等. 基于野外数据建立大型底栖动物电导率水质基准的可行性探讨[J]. 生态毒理学报，2015，10(1)：204-214.

[2] WU T H，ZHU G W，ZHU M Y，et al. Use of conductivity to indicate long-term changes in pollution processes in Lake Taihu，a large shallow lake[J].

Environmental science and pollution research,2020,27(17):21376-21385.

[3] 杨伟启.浮标式水质在线监测系统在我国水质环境监测管理体系中的应用与思考[J].中国科技博览,2014(37):312.

[4] 董凡.莱州湾南岸典型剖面地下水位和电导率动态变化特征及预测[D].长春:吉林大学,2022.

[5] 高鹏,谢家兴,孙道宗,等.基于物联网和 LSTM 的柑橘园土壤含水量和电导率预测模型[J].华南农业大学学报,2020,41(6):134-144.

[6] 宋利明,任士雨,张敏,等.基于集成学习的大西洋热带海域黄鳍金枪鱼渔情预报[J].中国水产科学,2021,28(8):1069-1078.

[7] 何纪力,陈宏文,胡小华,等.江西省严重酸雨地带形成的影响因素[J].中国环境科学,2000,20(5):477-480.

[8] 王瑞,代丹,张弛,等.太湖不同介质电导率时空变化特征[J].环境科学,2019,40(10):4469-4477.

[9] 孔宪喻,苏荣国.基于支持向量机的黄东海富营养化快速评价技术[J].中国环境科学,2016,36(1):143-148.

[10] 张玉超,钱新,钱瑜,等.支持向量机在太湖叶绿素 a 非线性反演中的应用[J].中国环境科学,2009,29(1):78-83.

[11] 张涛,李颖智,蔡五田,等.尼洋河流域不同水体电导率变化特征及影响因素[J].干旱区资源与环境,2018,32(4):126-131.

[12] 周煜.南昌市新城区大气降水的化学特征与环境指示意义[D].南昌:南昌大学,2014.

[13] 陈梦成,罗晶,许开成.江西省酸雨特征分析及预测模型[J].环境科学与技术,2014,37(10):167-170.

[14] 姬忠勇.水质电导率的测量原理分析及方法介绍[J].科技信息,2013(10):347.

[15] 徐火生,汪泽培.鄱阳湖水温时空变化规律分析[J].水文,1989(6):28-32.

[16] 王传海,李宽意,文明章,等.苦草对水中环境因子影响的日变化特征[J].农业环境科学学报,2007,26(2):798-800.

Prediction of Electrical Conductivity in Sub-lake of Poyang Lake Based on Random Forest Regression Model and High-frequency Data

LIU Li-zhen[1], HUANG Qi[2], CHI Dian-wei[3], FANG Chao-yang[2], CHU Ming-hang[1]

(1. Institute of Microbiology of Jiangxi Academy of Sciences, Jiangxi Academy of Sciences, Nanchang 330096, China; 2. Key Laboratory of Poyang Lake Wetland and Watershed Research, Ministry of Education, Jiangxi Normal University, Nanchang 330022, China; 3. College of Artificial Intelligence, Yantai Institute of Technology, Yantai 264005, China)

Abstract: Conductivity is an important parameter to measure water quality. High-frequency monitoring of water conductivity plays an important role in water quality management. Due to the complexity of field conditions, equipment failure often leads to data loss. In order to improve the high-frequency monitoring data, machine learning model was used to predict the conductivity content in water body based on the meteorological and physical indexes obtained from high-frequency monitoring. The results show that the random forest regression model has the best prediction effect, with its determination coefficient R^2 reaching 0.996, root mean square error (R_{RMSE}) 1.31 μS/cm, and mean relative error (M_{MRE}) 0.38%. The pH value contributed the most and was the dominant factor affecting the conductivity. The results are conducive to optimizing the field high-frequency monitoring system platform, improving the high-frequency monitoring data, which provides scientific basis for water quality management.

Key words: conductivity; random forest regression model; high-frequency monitoring data; Poyang Lake

(上接第 45 页)

Experimental Study on Influence of Unsteady Inflow Process on Hydrodynamics of Compound Open Channel

FU Jun[1a], ZHANG Tao-tao[2,1b], LIU Jia-ming[1a], LIU Jie-qing[1a]

(1a. College of Water Conservancy and Hydropower Engineering; 1b. Institute of Water Science and Technology, Hohai University, Nanjing 210098, China; 2. School of Environmental Science and Engineering, Suzhou University of Science and Technology, Suzhou 215009, China)

Abstract: The unsteady inflow process changes the hydrodynamic characteristics of the compound open channel and then determines the processes of material transfer and energy transfer in the floodplain and main channel. In order to study the influence of different unsteady inflow processes on the hydrodynamic characteristics of the compound open channel, indoor flume experiments were carried out to analyze the influence of different unsteady inflow processes on the water level, flow instability, and discharge per unit width process of the compound open channel, taking into account three factors: discharge amplitude, unsteady period, and rising duration. The results show that with the increase of discharge amplitude, the slope of wave peak growth of water level is greater than that of wave valley, and the wave height increases linearly. The discharge process of the unit width of the channel is deformed over time, and there is an obvious slow change process in the descending process of the main channel and the rising process of the floodplain. With the increase of the unsteady period, the wave peak of the water level increases, the wave valley decreases, and the wave height also increases linearly. The phenomenon of deformation of the unit width discharge of the channel over time disappears. The water level peaks and valleys change synchronously with the hydrologic skewness. When the hydrologic skewness (ratio of fluctuation water duration) is greater than 1, the rising process of the unit width discharge of the floodplain and main channel will also have an obvious slow change stage.

Key words: compound open channel; unsteady; inflow process; main channel; floodplain