# 基于机器学习的地下水水质预测研究*

肖　燚[1)]　郭亚会[1)]　李明蔚[1)]　付永硕[1)†]　孙　峰[2)]
(1)北京师范大学水科学研究院, 100875, 北京;　2)水利部信息中心(水利部水文水资源监测预报中心), 100875, 北京)

**摘要**　基于实测的地下水水质数据(pH、总硬度、溶解性总固体、硫酸盐、氯化物、Fe、Mn 7 种)和气象数据(平均气温、最低气温、最高气温、平均最低气温、平均最高气温、20: 00—20: 00 降水量、日降水量≥0.1 mm 的时间、最大日降水量 8 种), 分别使用 BP 神经网络、随机森林(RF)和支持向量机(SVM)构建了地下水水质参数的机器学习预测模型. 对于每一种水质参数, 分别使用不同的机器学习算法基于不同滞后期的数据进行模拟, 将结果与实测水质进行对比, 选择精度最高的机器学习模型及其对应的滞后期作为该水质参数的最优模型和最佳滞后期. 结果表明, 不同机器学习方法和滞后期的选择对预测精度影响很大, BP 神经网络对 pH($R^2$=0.225, RMSE 为 2.411)、总硬度($R^2$=0.503, RMSE 为 47.973 mg·$L^{-1}$)、氯化物($R^2$=0.994, RMSE 为 0.544 mg·$L^{-1}$)和 Fe($R^2$=0.302, RMSE 为 7.772 mg·$L^{-1}$)的预测精度最高, RF 对硫酸盐($R^2$=0.908, RMSE 为 3.788 mg·$L^{-1}$)和 Mn($R^2$=0.522, RMSE 为 0.429 mg·$L^{-1}$)的预测精度最高, BP 神经网络、RF 和 SVM 对溶解性总固体的预测性能均较好($R^2$=0.994~0.996, RMSE 为 674.660~950.470 mg·$L^{-1}$). 此外, 硫酸盐和 Mn 预测模型对应的最佳滞后期为 0 个月, 溶解性总固体和氯化物预测模型对应的最佳滞后期为 1 个月, pH、总硬度和 Fe 预测模型对应的最佳滞后期为 2 个月.

**关键词**　地下水水质; BP 神经网络; 随机森林; 支持向量机
**中图分类号**　X832　　**DOI**: 10.12202/j.0476-0301.2021196

## 0　引言

地下水是水资源的重要组成部分, 是维持人们生产和生活的重要资源之一, 尤其在地表水资源相对匮乏的地区, 工业、农业及居民的生活用水更加依赖地下水[1−2]. 在过去几十年, 人口持续增长、城市化进程加快、气候变化、地下水的不合理开发利用等一系列因素导致地下水水量急剧减少、水质严重恶化, 水资源安全遭受了巨大威胁和挑战[3−5]. 水安全是人类健康、公共卫生、粮食安全的重要保障, 目前全球有超过 8.44 亿人面临饮用水危机, 地下水作为重要的饮用水源, 对其进行及时监测和科学管理, 对于确保地下水水质和水量都具有十分重要的意义[6]. 水质从本质上决定了水的可用性, 传统水质监测主要采用现场采样结合实验室测定的方法, 可测参数多, 准确度高, 但操作烦琐, 成本高昂[7−8]. 而在地下水的监测中, 往往需要借助打井, 然后取样分析, 这就使得地下水的水质监测成本和难度都比地表水高很多[9−10]. 借助新的方法通过少量的地下采样点, 选取恰当的预测因子和水质指标进行建模, 对过去水质进行重建和对未来水质变化进行预测, 是一种高效的方法. 基于这种方法进行地下水水质预测, 为保护和管理地下水资源提供了一条新的途径.

随着城市化进程的不断加快, 以全球变暖、降水变化、极端气候为代表的气候变化对水资源造成了严重影响, 加剧了水体污染[11−12]. 气候变化增加了未来水质状况的不确定性, 理解气候变化与水质的关系、制定适应气候变化趋势的水质保护和管理政策以应对这种不确定性, 成为未来研究的重点[13]. Valiallahi 等[14] 利用因子分析、聚类分析等多元统计分析方法评估了气候变化和人为因素对河流水质的影响, 结果表明季节和气象因素会对水质产生很大影响, 降雨导致地表径流增大, 从而加剧了水体污染; 安国英等[15] 对大理地区 1989—2019 年的 12 个站点数据进行分析, 结果表明洱海的综合营养状态指数和总氮(TN)年质量浓度与平均温度显著正相关, 总磷(TP)质量浓度与冬季气温、高锰酸盐指数年质量浓度与夏季或冬季气温显著正相关; 车蕊等[16] 通过研究连续极端降雨对东江流域水质影响, 发现降雨量与浊度、TP、氰化物、Pb、Fe 和 Mn 的质量浓度值均呈显著正相关, 与 pH、电导率和 Zn 的质量浓度值呈显著负相关. 综

* 国家重点研发计划课题资助项目(2018YFC0407702)
† 通信作者: 付永硕(1979—), 教授, 博士生导师. 研究方向: 生态遥感、生态水文过程. E-mail: yfu@bnu.edu.cn
收稿日期: 2021-08-20

上可见,气象因素与水质有较强的相关性,选择气象因素作为预报因子具有一定的合理性. 目前针对水质监测往往都是分析气象因素和地表水(湖泊、河流等)水质的相关性关系,对于气象因素和地下水之间关系的研究还主要集中在地下水水位[17]和补给量[18−19]等方面,针对气象因素和地下水水质的研究还很少. 因此需要深入探究气象因子与地下水水质指标之间的相关性关系,这对理解气象因子对地下水水质的影响,控制地下水污染,积极应对潜在的气候变化带来的水质变化的负面影响具有十分重要的意义.

机器学习是人工智能的重要分支,它涵盖概率论知识、统计学知识、近似理论知识和复杂算法知识[20]. 机器学习可以通过学习输入数据的结构和内在模式,挖掘数据的潜在信息,并应用于解决分类、回归、聚类等问题[21]. 由于机器学习具有适用性强、精度高等显著优势,目前已有许多学者将其应用于水资源领域的研究: Mohapatra 等[22]比较了自适应神经模糊系统(ANFIS)、深度神经网络(DNN)和支持向量机(support vector machines, SVM)对农业生态区地下水水位季节性变化的预测效果,结果表明 DNN 算法的预测效果最优,能有效预测大多数农业生态区地下水位的季节性变化; Tran 等[23]评估了随机森林回归(RFR)、极端梯度提升树回归(XGBR)、CatBoost 算法和轻梯度提升树回归(LGBR)4 种机器学习方法对越南湄公河地区地下水盐度的预测效果,其中 CatBoost 算法的精度最高(均方根误差 RMSE 为 205.96,决定系数为 $R^2$=0.84). 在水质的建模中,水质数据和环境变量间的关系往往是非线性的,而机器学习恰巧可以很好地解决复杂非线性问题[24]. 有研究表明,机器学习算法用于估计和预测水质指数具有显著的准确性,能节省大量的时间和精力[8]. Wang 等[25]设计了一种 CA-NARX 算法,采用改进的动态聚类算法对水体富营养化程度进行分类,综合考虑气温、水温、水面蒸发量、降雨量 4 个气象因子,采用正向动态回归神经网络对水体中 TN 和 TP 进行预测,取得了较高的精度. 上述研究探讨了利用机器学习进行水质预测的可行性,但大多是利用机器学习对气象因子和地表水水质进行建模,很少有研究涉及机器学习在气象因子与地下水水质建模中的应用,借助机器学习构建气象因子和地下水水质间的关系还有待于进一步的探究.

本文以佳木斯市 5 个站点 2018—2020 年地下水水质监测数据为基础,结合佳木斯市气象数据,利用 BP(back-propagation)神经网络、随机森林(random forest, RF)、SVM 分别构建地下水水质和气象因子之间的关系,获得了针对特定站点的地下水水质监测模型,对模型进行了检验,最后基于构建的模型,结合 2015—2019 年月尺度的气象数据对同时段的各水质参数进行重建. 在此过程中还研究分析了气象因子对地下水水质影响的滞后性,滞后期分别设置为 1、2、3 个月,选择精度最高的作为地下水水质监测/预测的最佳模型.

## 1 研究区与数据

**1.1 研究区概况** 采样点福隆、永兴、长胜、望江、福胜位于佳木斯市,该地区位于三江平原西南部,黑龙江省东北部,坐标范围 129°29′~135°5′E, 45°56′~48°28′N,属中温带大陆性季风气候,年均气温 3℃,年均降水量 510 mm,年有效平均积温 2521 ℃[26]. 采样站点地理坐标及采样日期见表 1.

**表 1 研究区采样站点信息**

| 站点 | 经纬度 | 采样日期 |
|---|---|---|
| 福隆 | 46°49′N, 130°25′E | 2018-09-14、2019-07-25、2020-09-28 |
| 永兴 | 46°46′N, 130°32′E | 2018-09-23、2019-07-18、2020-09-15 |
| 长胜 | 46°51′N, 130°16′E | 2018-09-23、2019-07-18、2020-09-15 |
| 望江 | 46°51′N, 130°10′E | 2018-09-23、2019-07-18、2020-09-15 |
| 福胜 | 46°54′N, 130°27′E | 2018-09-23、2019-07-20、2020-10-08 |

**1.2 气象数据** 收集了 2015—2019 年间佳木斯市以气温和降水为代表的逐月气象数据,包括平均气温、最低气温、最高气温、平均最低气温、平均最高气温、20: 00—20: 00 降水量、日降水量≥0.1 mm 的时间和最大日降水量 8 个气象因子,数据统计结果如图 1 所示.

佳木斯地区气温和降水呈现明显的季节差异,夏季湿润炎热,平均气温在 20 ℃ 左右,个别月份降水量>100 mm,冬季寒冷干燥,平均气温<−10 ℃,降水量<30 mm. 总体上,降水量的逐年波动趋势相对高于气温逐年波动趋势.

**1.3 水质数据** 采集了 2018—2020 年间佳木斯市 5 个站点年尺度的实测水质数据,具体包括色度、pH、总硬度、溶解性总固体、硫酸盐、氯化物、Fe、Mn、Cu、Zn、Al、挥发性酚、阴离子表面活性剂、高锰酸盐指数、氨氮、浊度、硫化物和 Na 等共 18 种水质指标. 本文对实测水质参数的有效性进行甄别,最终筛选出 7 种可行的水质指标来构建长时间序列的水质数据,部分水质实测结果如表 2 所示.

## 2 研究方法

本文选取 BP 神经网络、RF 和 SVM 3 种常用的机器学习方法进行建模,将 2018—2019 年的实测地

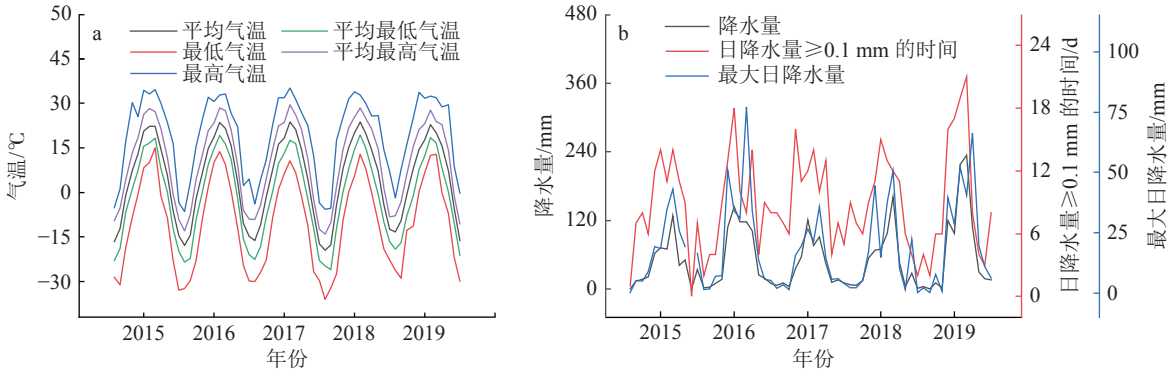


**图 1　佳木斯市多年气象因子统计**

**表 2　水质实测数据**

| 站点 | 采样时间 | pH | 总硬度/(mg · L$^{-1}$) | 溶解性总固体/(mg · L$^{-1}$) | 硫酸盐/(mg · L$^{-1}$) | 氯化物/(mg · L$^{-1}$) | 铁/(mg · L$^{-1}$) | 锰/(mg · L$^{-1}$) |
|---|---|---|---|---|---|---|---|---|
| 福隆 | 2018-09-14 | 6.400 | 49.500 | 78.000 | 3.260 | 1.350 | 10.400 | 0.330 |
| | 2019-07-25 | 6.500 | 42.100 | 190.000 | 4.310 | 1.900 | 9.900 | 0.450 |
| | 2020-09-28 | 7.570 | 56.000 | 272.000 | 3.320 | 2.100 | 1.190 | 0.560 |
| 永兴 | 2018-09-23 | 6.500 | 75.700 | 98.000 | 0.160 | 0.410 | 3.210 | 0.790 |
| | 2019-07-18 | 6.800 | 84.900 | 266.000 | 0.798 | 0.665 | 5.360 | 0.180 |
| | 2020-09-15 | 7.900 | 104.000 | 384.000 | 3.160 | 0.700 | 17.000 | 1.370 |
| 长胜 | 2018-09-23 | 6.500 | 57.100 | 73.000 | 1.010 | 0.470 | 2.060 | 0.320 |
| | 2019-07-18 | 6.800 | 40.200 | 192.000 | 1.510 | 0.814 | 18.500 | 0.340 |
| | 2020-09-15 | 7.300 | 96.000 | 155.000 | 3.830 | 0.700 | 13.600 | 0.660 |
| 望江 | 2018-09-23 | 6.600 | 57.500 | 81.000 | 3.290 | 1.570 | 2.250 | 0.270 |
| | 2019-07-20 | 6.800 | 54.700 | 2488.000 | 3.220 | 1.650 | 14.900 | 0.140 |
| | 2020-09-15 | 7.300 | 110.000 | 2974.000 | 3.400 | 1.800 | 11.800 | 0.570 |
| 福胜 | 2018-09-23 | 6.500 | 53.500 | 97.000 | 16.200 | 8.160 | 4.060 | 0.680 |
| | 2019-07-20 | 6.600 | 42.200 | 112.000 | 10.300 | 5.300 | 23.400 | 0.760 |
| | 2020-10-08 | 6.900 | 76.000 | 340.000 | 6.700 | 5.700 | 20.800 | 0.910 |

下水水质数据和对应时段的气象因子用于模型的训练，在模型构建的过程中充分考虑滞后期的影响，分别将单个站点当月、1 个月前、2 个月前、3 个月前的全部气象因子与单一的水质参数进行建模，并对 5 个站点的水质数据分别进行重建. 最后利用不同滞后期下重建的 2015—2019 年各水质参数的平均值和 2020 年的实测水质数据进行对比，完成对模型精度的检验，并根据 $R^2$ 和 RMSE 筛选出符合条件的模型，技术路线如图 2 所示.

**2.1　BP 神经网络**　人工神经网络是一种常用的非线性统计性数据建模工具，主要通过自学习寻找目标值与输入变量之间的映射关系[27]，与传统方法相比，人工神经网络模型在需要更少的先验假设的情况下，能获得更高的精度，适合于求解非线性和不确定性问题[28].

BP 神经网络模型，即前馈神经网络模型，是目前应用最广泛的一种人工神经网络模型，由输入层、隐含层和输出层组成. 输入层负责接收输入信号，输出层负责输出计算结果，隐含层负责描述问题的层次关

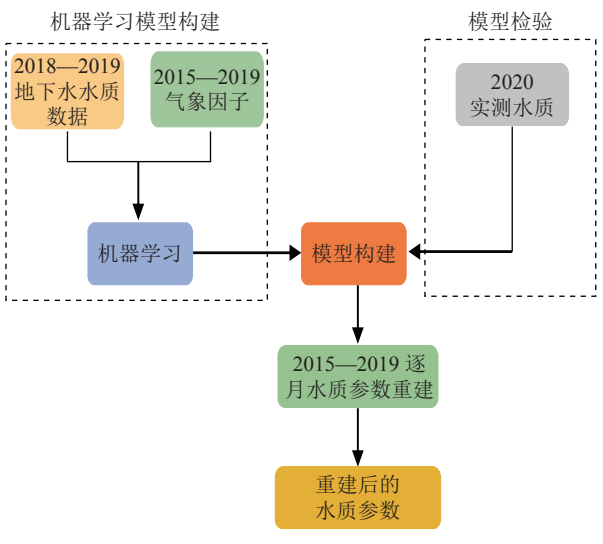


**图 2　技术路线**

系. 本文选择 BP 神经网络进行建模, 隐含层设定为一层, 输入层、隐含层、输出层中神经元的个数分别为 8、10、7.

隐含层第 $j$ 个神经元输出值为 $O_j$, 计算式为

$$O_j = \Phi(n_j) = \Phi\left(\sum_{i=1}^{N} \omega_{ij} x_i + \theta_j\right), \tag{1}$$

输出层第 $k$ 个神经元输出值为 $O_k$, 计算式为

$$O_k = \Psi(n_k) = \Psi\left(\sum_{j=1}^{M} \epsilon_{jk} y_j + \mu_k\right), \tag{2}$$

式中: $n_j$ 和 $n_k$ 分别为隐含层第 $j$ 个神经元和输出层第 $k$ 个神经元的输入信号; $\omega_{ij}$为输入层第 $i$ 个神经元到隐含层第 $j$ 个神经元的权值, $\theta_j$为对应的阈值; $\epsilon_{jk}$为隐含层第 $j$ 个神经元到输出层第 $k$ 个神经元的权值, $\mu_k$为对应的阈值; $N$ 和 $M$ 分别为输入层和隐含层的神经元个数; $\Phi$为隐含层传递函数, $\Psi$为输出层传递函数.

本文构建的人工神经网络模型的层次为 8-11-7, 神经网络结构如图 3 所示.

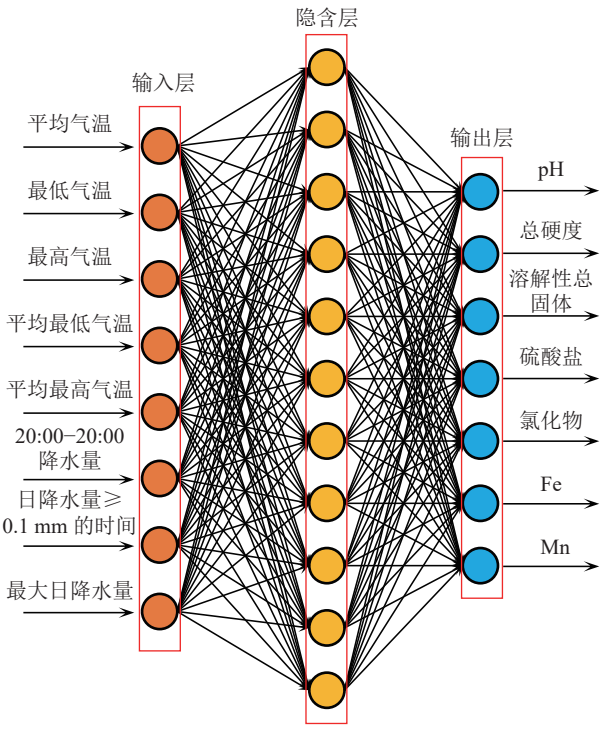


图 3 人工神经网络结构

**2.2 随机森林** RF 是一种常用的机器学习方法, 被广泛地应用于解决分类和回归问题. 利用自助法(Bootsrap)重抽样方法从原始样本中抽取多个 Bootsrap 样本进行决策树建模, 然后组合多棵决策树的预测方法, 并最终通过投票或取均值的方法得到预测结果[29], 如图 4 所示.

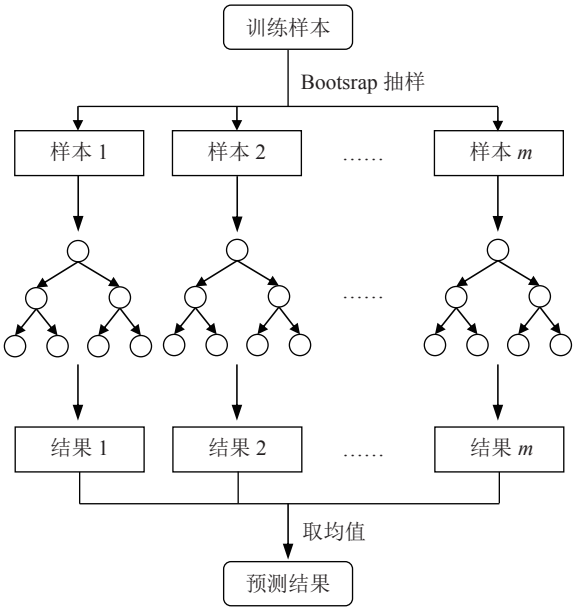


图 4 随机森林结构

RF 能很好地处理变量间的非线性关系, 具有预测准确率高、对噪声和异常值容忍度高、运算较快、不易出现过拟合等显著优势[30]. 此外, RF 能在观测变量较少的情况下出色地完成多变量预测[31], 适合于受实地测量多种条件限制, 采集数据量较少的地下水水质建模与分析. 本文建立的 RF 模型设定决策树数量为 300, 每次树模型重建时节点分裂的次数为 6.

**2.3 支持向量机** SVM 是建立在统计学习理论基础上的一种数据挖掘方法, 在解决小样本、非线性和高维模式识别问题中具有独特的优势[32]. SVM 的基本理论是寻找满足分类要求的最优分类平面, 即超平面[33]. 这一平面不但能将 2 类正确分开, 而且能够使分类间隔最大, 距离超平面最近的向量称为支持向量, 超平面的表示方法为

$$\Omega^{\mathrm{T}} x + b = 0, \tag{3}$$

式中: $\Omega$为法向量, 决定超平面的方向; $b$ 为位移项, 决定超平面与原点之间的距离.

SVM 最早主要应用于解决模式识别问题, 后来又逐步扩展到解决回归问题, 基于 SVM 方法的回归估计能够以可控制的精度逼近任一非线性函数, 与传统算法相比优势显著[34]. 假设输入样本集可以表示为 $x_1, x_2, \cdots, x_m$, SVM 结构如图 5 所示, SVM 回归函数的表达式为

$$f(x) = \sum_{i=1}^{m} (\alpha_i^* - \alpha_i) K(x_i, x) + b, \tag{4}$$

式中: $K(x_i, x)$为核函数; $\alpha_i^* - \alpha_i \neq 0$ 对应的样本为支持向量; $b$ 为偏置项[31].

本文建立的 SVM 模型核函数采用径向基核函数

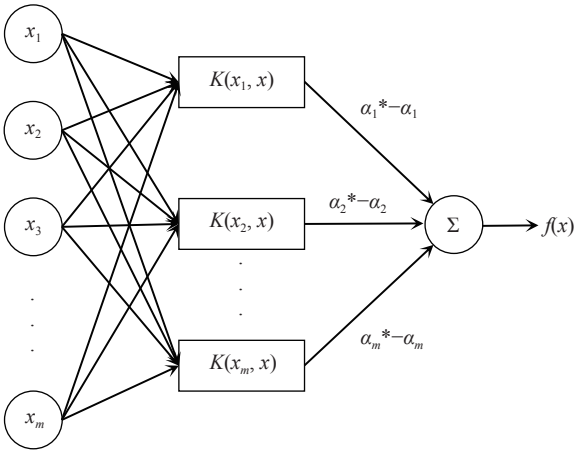


**图 5　SVM 结构**

(radial basis function, RBF), 设定惩罚参数 Cost 为 1, 核参数 Gamma 为 0.1.

## 3　结果与分析

将利用 BP 神经网络、RF、SVM 重建的各不同滞后期下 2015—2019 年逐月水质数据的平均值和 2020 年实测水质数据进行对比, 得到各水质参数基于不同机器学习模型在不同滞后期下的精度验证结果(表 3).

结果表明, 采用不同机器学习方法建模对地下水水质参数的预测精度有较大影响. 不同地下水水质参数的最优建模方法不同, pH、总硬度、氯化物和 Fe 最适合采用 BP 神经网络进行建模(pH: $R^2$=0.225,

**表 3　水质参数模型精度检验结果**

| 水质参数 | 滞后期/月 | BP神经网络 | | RF | | SVM | |
|---|---|---|---|---|---|---|---|
| | | $R^2$ | RMSE | $R^2$ | RMSE | $R^2$ | RMSE |
| pH | 0 | 0.002 | 2.005 | 0.001 | 0.931 | 0.012 | 0.905 |
| | 1 | 0.050 | 3.086 | 0.005 | 0.881 | 0.018 | 0.897 |
| | 2 | 0.225 | 2.411 | 0.013 | 0.903 | 0.017 | 0.898 |
| | 3 | 0.136 | 1.556 | 0.024 | 0.888 | 0.016 | 0.899 |
| ρ(总硬度) | 0 | 0.294 | 44.660 | 0.380 | 35.336 | 0.362 | 36.089 |
| | 1 | 0.499 | 49.019 | 0.330 | 36.474 | 0.357 | 36.354 |
| | 2 | 0.503 | 47.973 | 0.361 | 36.161 | 0.358 | 36.309 |
| | 3 | 0.299 | 40.900 | 0.351 | 36.615 | 0.358 | 36.292 |
| ρ(溶解性总固体) | 0 | 0.995 | 891.799 | 0.996 | 950.470 | 0.995 | 815.484 |
| | 1 | 0.996 | 885.855 | 0.994 | 674.660 | 0.995 | 769.421 |
| | 2 | 0.995 | 880.187 | 0.995 | 802.907 | 0.995 | 777.200 |
| | 3 | 0.994 | 842.341 | 0.994 | 724.839 | 0.995 | 780.221 |
| ρ(硫酸盐) | 0 | 0.898 | 3.040 | 0.908 | 3.788 | 0.899 | 3.477 |
| | 1 | 0.886 | 2.485 | 0.906 | 3.753 | 0.896 | 3.371 |
| | 2 | 0.889 | 1.737 | 0.898 | 3.448 | 0.896 | 3.389 |
| | 3 | 0.831 | 3.111 | 0.892 | 3.270 | 0.897 | 3.396 |
| ρ(氯化物) | 0 | 0.886 | 1.154 | 0.981 | 0.734 | 0.985 | 0.573 |
| | 1 | 0.994 | 0.544 | 0.983 | 0.709 | 0.986 | 0.519 |
| | 2 | 0.975 | 1.173 | 0.985 | 0.558 | 0.986 | 0.528 |
| | 3 | 0.964 | 0.427 | 0.987 | 0.466 | 0.986 | 0.531 |
| ρ(Fe) | 0 | 0.000 | 8.978 | 0.073 | 9.215 | 0.000 | 8.208 |
| | 1 | 0.082 | 8.744 | 0.050 | 8.370 | 0.004 | 7.908 |
| | 2 | 0.302 | 7.772 | 0.000 | 8.124 | 0.003 | 7.957 |
| | 3 | 0.005 | 8.642 | 0.015 | 7.644 | 0.002 | 7.976 |
| ρ(Mn) | 0 | 0.262 | 0.945 | 0.522 | 0.429 | 0.336 | 0.457 |
| | 1 | 0.054 | 1.073 | 0.173 | 0.490 | 0.276 | 0.467 |
| | 2 | 0.114 | 1.112 | 0.319 | 0.460 | 0.286 | 0.466 |
| | 3 | 0.000 | 0.676 | 0.221 | 0.477 | 0.289 | 0.465 |

注: 0代表用当月的气象和当月的水质进行建模, 1、2、3分别代表用提前1、2、3个月的气象和当月的水质进行建模; RMSE的单位除pH为1外, 其余参数的单位均为mg・$L^{-1}$.

RMSE 为 2.411; 总硬度: $R^2$=0.503, RMSE 为 47.973 mg·$L^{-1}$; 氯化物: $R^2$=0.994, RMSE 为 0.544 mg·$L^{-1}$; 铁: $R^2$=0.302, RMSE 为 7.772 mg·$L^{-1}$), 硫酸盐和锰最适合采用 RF 进行建模(硫酸盐: $R^2$=0.908, RMSE 为 3.788 mg·$L^{-1}$; Mn: $R^2$=0.522, RMSE 为 0.429 mg·$L^{-1}$). 此外, BP 神经网络、RF 和 SVM 对溶解性总固体的预测效果都较好($R^2$=0.994~0.996, RMSE 为 674.660~950.470 mg·$L^{-1}$). 各水质参数基于最优机器学习方法和滞后期的重建结果如图 6 所示, 图中 ML 表示选择的机器学习方法, $T$ 表示最佳滞后期.

## 4 结论

本文使用 BP 神经网络、RF 和 SVM 3 种机器学习方法, 将气象因子作为自变量和地下水水质进行建

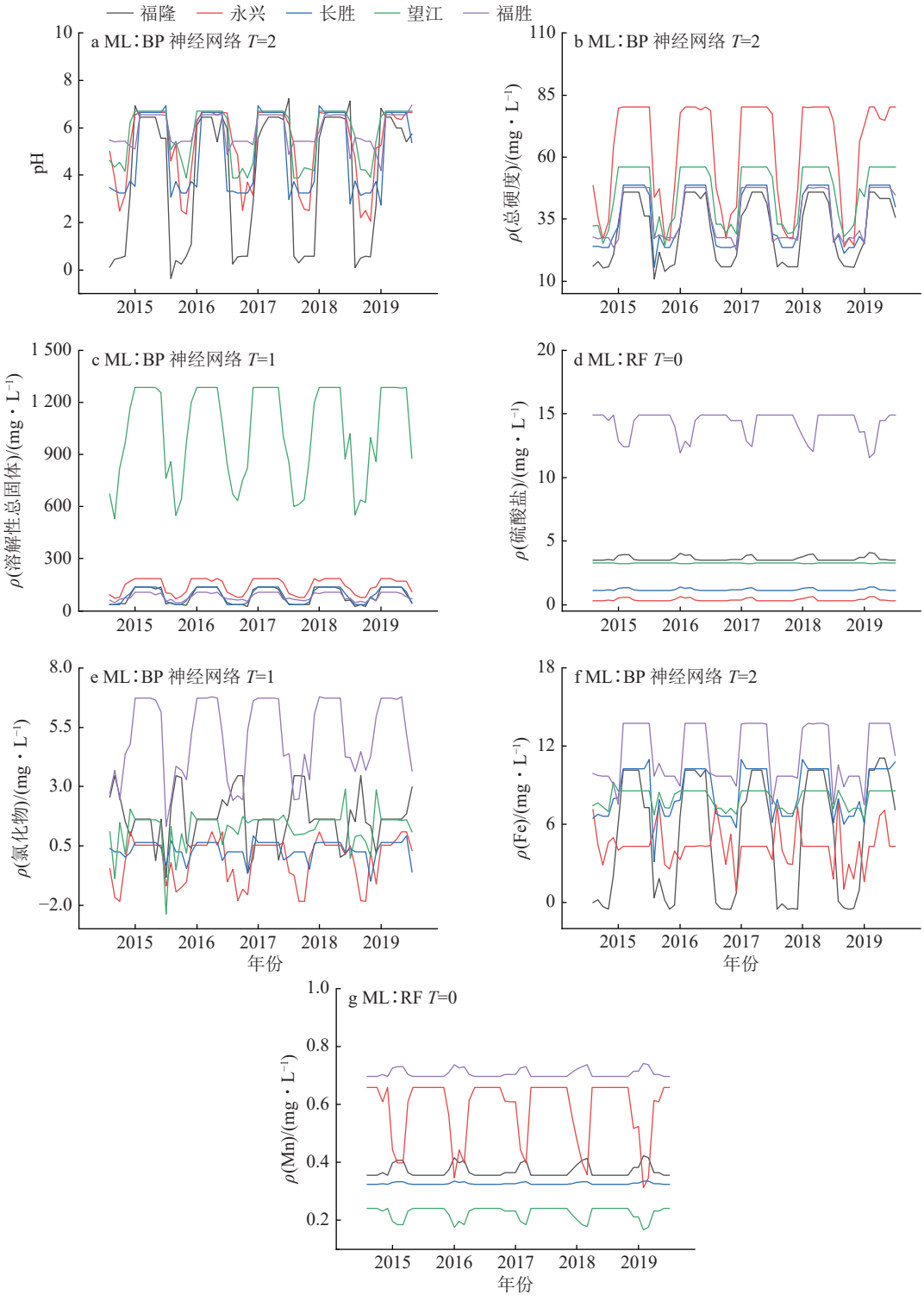


图 6 5 个站点典型水质模拟

模, 并将输出结果与实测水质进行了对比, pH、总硬度、溶解性总固体、硫酸盐、氯化物、Fe 和 Mn 的最优预测模型的 $R^2$ 分别为 0.225、0.503、0.996、0.908、0.994、0.302、0.522, RMSE 分别为 2.411、47.973、885.855、3.788、0.544、7.772、0.429 mg·$L^{-1}$. 结果表明预测取得了较高的精度. pH、总硬度、溶解性总固体、氯化物和 Fe 的最优预测模型是基于 BP 神经网络构建的, 最优滞后期分别为 2、2、1、1、2 个月, 硫酸盐和 Mn 的最优预测模型是基于 RF 构建的, 最优滞后期为 0 个月. 机器学习在模型构建中有着无可比拟的优势, 基于机器学习的水质监测/预测研究在未来的应用场景有很大潜力. 同时, 不同的机器学习模型在拟合时会存在很大的差异性, 同一个模型参数选取的不同也会导致差异性. 未来的地下水水质建模的研究, 需要考虑水质反演的机制性, 同时需要探究更先进的深度学习算法在水质预测中的有效性.

## 5　参考文献


[1] 喻朝庆. 水-氮耦合机制下的中国粮食与环境安全[J]. 中国科学:地球科学, 2019, 49(12): 2018

[2] 杨建青, 章树安, 陈喜, 等. 国内外地下水监测技术与管理比较研究[J]. 水文, 2013, 33(3): 18

[3] MARTINSEN G, 刘苏峡, 莫兴国, 等. 考虑地下水可持续开采约束条件的海河流域水资源优化配置[J]. Journal of Geographical Sciences, 2019, 29(6): 935

[4] SINGHA S, PASUPULETI S, SINGHA S S, et al. Prediction of groundwater quality using efficient machine learning technique[J]. Chemosphere, 2021, 276: 130265

[5] 戴长雷, 王思聪, 李治军, 等. 黑龙江流域水文地理研究综述[J]. 地理学报, 2015, 70(11): 1823

[6] RAKIB M A, SASAKI J, MATSUDA H, et al. Groundwater salinization and associated co-contamination risk increase severe drinking water vulnerabilities in the southwestern coast of Bangladesh[J]. Chemosphere, 2020, 246: 125646

[7] AHMED U, MUMTAZ R, ANWAR H, et al. Water quality monitoring: from conventional to emerging technologies[J]. Water Supply, 2020, 20(1): 28

[8] AGRAWAL P, SINHA A, KUMAR S, et al. Exploring artificial intelligence techniques for groundwater quality assessment[J]. Water, 2021, 13(9): 1172

[9] 范越, 卢文喜, 欧阳琦, 等. 基于Kriging替代模型的地下水污染监测井网优化设计[J]. 中国环境科学, 2017, 37(10): 3800

[10] 殷秀兰, 李圣品. 基于监测数据的全国地下水质动态变化特征[J]. 地质学报, 2021, 95(5): 1356

[11] AHMED T, ZOUNEMAT-KERMANI M, SCHOLZ M. Climate change, water quality and water-related challenges: a review with focus on Pakistan[J]. International Journal of Environmental Research and Public Health, 2020, 17(22): 8518

[12] KIM D, KIM J, JOO H, et al. Future water quality analysis of the Anseongcheon River basin, Korea under climate change[J]. Membrane Water Treatment, 2019, 10(1): 1

[13] 李洋, 李霞, 李国金. 气候变化背景下的水质研究: 研究进展及趋势[J/OL]. 水生态学杂志, [2022-03-02]. https://doi.org/10.15928/j.1674-3075.202001110010

[14] VALIALLAHI J, KHAFFAF ROUDY S. Application of multivariate statistical techniques for investigating climate change and anthropogenic effects on surface water quality assessment: case study of Zohreh River, Hendijan, Iran[J]. Applied Water Science, 2021, 11(6): 1

[15] 安国英, 郭兆成, 叶佩. 云南大理地区1989—2019年期间气候变化及其对洱海水质的影响[J/OL]. 现代地质, [2022-03-02]. https://doi.org/10.19657/j.geoscience.1000-8527.2021.102

[16] 车蕊, 林澍, 范中亚, 等. 连续极端降雨对东江流域水质影响分析[J]. 环境科学, 2019, 40(10): 4440

[17] JAVADINEJAD S, DARA R, JAFARY F. How groundwater level can predict under the effect of climate change by using artificial neural networks of NARX[J]. Resources Environment and Information Engineering, 2020, 2(1): 90

[18] POOL S, FRANCÉS F, GARCIA-PRATS A, et al. From flood to drip irrigation under climate change: impacts on evapotranspiration and groundwater recharge in the Mediterranean region of Valencia (Spain)[J]. Earth's Future, 2021, 9(5): e2020EF001859

[19] WANG S J, LEE C H, YEH C F, et al. Evaluation of climate change impact on groundwater recharge in groundwater regions in Taiwan[J]. Water, 2021, 13(9): 1153

[20] 徐艺. 机器学习算法及其应用研究 [D]. 长沙: 湖南大学, 2014

[21] 杨剑锋, 乔佩蕊, 李永梅, 等. 机器学习分类问题及算法研究综述[J]. 统计与决策, 2019, 35(6): 36

[22] MOHAPATRA J B, JHA P, JHA M K, et al. Efficacy of machine learning techniques in predicting groundwater fluctuations in agro-ecological zones of India[J]. Science of the Total Environment, 2021, 785: 147319

[23] TRAN D A, TSUJIMURA M, HA N T, et al. Evaluating the predictive power of different machine learning algorithms for groundwater salinity prediction of multi-layer coastal aquifers in the Mekong Delta, Vietnam[J]. Ecological Indicators, 2021, 127: 107790

[24] 徐萍. 基于多源遥感数据对松花江哈尔滨段水质反演研究[D]. 哈尔滨: 哈尔滨师范大学, 2020

[25] WANG J, GENG Y, ZHAO Q, et al. Water quality prediction of water sources based on meteorological factors using the CA-NARX approach[J]. Environmental Modeling & Assessment, 2021, 26(4): 529
[26] 朱长虹. 基于多维临界调控理论的佳木斯市水资源优化配置[D]. 哈尔滨: 东北农业大学, 2015
[27] 刘建华, 张启斌, YANG D, 等. 基于MCR-ANN-CA模型的包头市生态用地演变模拟[J]. 农业机械学报, 2019, 50(2): 187
[28] CHEN Y Y, SONG L H, LIU Y Q, et al. A review of the artificial neural network models for water quality prediction[J]. Applied Sciences, 2020, 10(17): 5776
[29] 方匡南, 吴见彬, 朱建平, 等. 随机森林方法研究综述[J]. 统计与信息论坛, 2011, 26(3): 32
[30] 任婷玉, 梁中耀, 陈会丽, 等. 基于模式识别方法的湖泊水质污染特征聚类研究[J]. 北京大学学报(自然科学版), 2019, 55(2): 335
[31] 豆荆辉, 夏瑞, 张凯, 等. 非参数模型在河湖富营养化研究领域应用进展[J]. 环境科学研究, 2021, 34(8): 1928
[32] 丁世飞, 齐丙娟, 谭红艳. 支持向量机理论与算法研究综述[J]. 电子科技大学学报, 2011, 40(1): 2
[33] CHEN Y B, XU P, CHU Y Y, et al. Short-term electrical load forecasting using the Support Vector Regression (SVR) model to calculate the demand response baseline for office buildings[J]. Applied Energy, 2017, 195: 659
[34] 杜树新, 吴铁军. 用于回归估计的支持向量机方法[J]. 系统仿真学报, 2003, 15(11): 1580

# Machine learning to predict groundwater quality

XIAO Yi[1)] GUO Yahui[1)] LI Mingwei[1)] FU Yongshuo[1)] SUN Feng[2)]

(1)College of Water Sciences, Beijing Normal University, 100875, Beijing, China; 2)Information Center(Hydrology Monitor and Forecast Center), Ministry of Water Resources, 100875, Beijing, China)

**Abstract** Groundwater quality data (pH, total hardness, total dissolved solids, sulfate, chloride, iron and manganese) and meteorological data (average temperature, minimum temperature, maximum temperature, average minimum temperature, average maximum temperature, daily (20:00-20:00) precipitation, daily precipitation ≥ 0.1 mm days, maximum daily precipitation) were subject to analysis by machine learning models, using BP neural network, random forest and support vector mechanism. For each groundwater quality parameter, different machine learning algorithms were used to simulate data in different lag phases, results were then compared with measured groundwater quality parameters. Machine learning model with highest accuracy and corresponding lag phase were selected as the optimal model. Different machine learning methods and choice of lag phase were found to have great influence on prediction accuracy. BP neural network showed the highest prediction accuracy for pH ($R^2$ = 0.225, RMSE is 2.411), total hardness ($R^2$ = 0.503, RMSE is 47.973 mg · $L^{-1}$), chloride ($R^2$ = 0.994, RMSE is 0.544 mg · $L^{-1}$) and iron ($R^2$ = 0.302, RMSE is 7.772 mg · $L^{-1}$). RF showed the highest prediction accuracy for sulfate ($R^2$ = 0.908, RMSE is 3.788 mg · $L^{-1}$) and Manganese ($R^2$ = 0.522, RMSE is 0.429 mg · $L^{-1}$). All methods used showed good predictive performance for total dissolved solids ($R^2$ = 0.994-0.996, RMSE is 674.660-950.470 mg · $L^{-1}$). The best lag phase of sulfate and Manganese monitoring model was 0 month, the best lag phase of chloride monitoring model was 1 month, the best lag phase of pH, dissolved total solids and total hardness monitoring model was 2 months.

**Keywords** groundwater quality; BP neural network; random forest; support vector machine

【责任编辑: 武 佳】