

侯兴汉, 彭悌珊, 于波, 等. 基于SVR插补高缺失率的地下水水质监测数据插补研究[J]. 安全与环境学报, 2026, 26(6): 2465–2474.

HOU X H, PENG T S, YU B, et al. Research on the application of Support Vector Regression (SVR) for data interpolation in groundwater quality monitoring series with high missing rates[J]. Journal of Safety and Environment, 2026, 26(6): 2465–2474.

基于SVR插补高缺失率的地下水水质监测数据插补研究*

侯兴汉¹, 彭悌珊², 于波¹, 李新峰¹, 刘国东²

(1 北京飞燕石化环保科技发展有限公司, 北京 102500; 2 四川大学水利水电学院, 成都 610065)

摘要:地下水水质监测数据常由多种原因导致缺失,严重影响数据分析与水质评价的准确性。研究以华北某石化区多年季度监测数据为例,针对高锰酸盐指数、pH值、总硬度和锰质量浓度4个关键水质指标具有污染指示意义但存在高缺失率的问题,提出一种基于滑动窗口的支持向量回归(Support Vector Regression, SVR)插补方法。通过分阶段网格搜索与5折交叉验证优化模型超参数,并将插补结果与常规线性插值方法进行对比。结果显示,SVR模型在高锰酸盐指数、pH值和总硬度的插补精度上均优于线性插值,表明其适用于样本量小、缺失率高、连续缺失多的水质时间序列数据;锰质量浓度受检出限影响其插补效果略低于线性插值,未来需单独分析该类缺失问题。研究为地下水监测数据缺失值处理提供了可靠的新方法,具有较强的实际应用价值。

关键词:环境工程学;石化区地下水;水质监测;支持向量回归;数据插补

中图分类号:X832 **文献标志码:**A **文章编号:**1009-6094(2026)06-2465-10

DOI: 10.13637/j.issn.1009-6094.2025.1337

0 引言

地下水作为水资源的重要组成部分,是人类及其生活环境必不可少的基本要素。在地下水环境监测过程中,水质数据的收集需要依赖实验室分析或高成本传感器,采样频率相对较低,监测数据量少^[1]。此外,地下水水质测量难度大,数据采集、传输等情况的发生导致数据缺失较多。

水质数据缺失一般包括完全随机缺失、随机缺失和非随机缺失,主流的插补方法包括删除法和填充法^[2],其中填充法包括统计学方法和机器学习方法^[3]。近年来,国内外诸多学者利用多种机器学习方法插补水质数据,并针对多种插补方法效果对比^[4]、水质数据高比例缺失^[5-7]、模型结果可解释性^[8]进行了研究,并取得了良好的效果。但不难看出,对于数据量较少、缺失率很高(10%~25%)且存在长时间连续缺失的序列数据,现有插补方法通常失效,且随机森林(Random Forest, RF)、长短期记忆网络(Long Short Term Memory, LSTM)等模型极易出现过拟合,插补方法的鲁棒性差。

1995年,Cortes等^[9]首次提出一种基于统计学习理论的机器学习工具,即支持向量机(Support

Vector Machine, SVM)。1997年,该团队基于支持向量首次提出支持向量回归(Support Vector Regression, SVR)用于回归分析^[10]。相较于RF、LSTM等机器学习方法,SVR可以在样本数据量有限的情况下,在模型复杂性和学习能力之间取得良好平衡,从而获得较好的泛化能力^[11]。因此,SVR更适用于水质监测数据稀缺的场景。薛同来等^[12]融合遗传算法构建了参数自适应寻优的支持向量回归水质预测模型,减少了模型选择参数的时间,并获得了良好的预测效果。张秀菊等^[13]探讨了支持向量回归理论及算法步骤,并将其运用于实际的水质预测中。石翠翠等^[14]利用高斯函数分布特性改进粒子群算法,建立了非线性动态惯性权重的粒子群算法(Particle Swarm Optimization with Nonlinear Decreasing Weight Based on Gaussian Function, GNIPSO),并优化了支持向量回归模型,提高了模型的精确度和全局搜索能力。陈霖等^[15]利用皮尔逊相关系数(Pearson Correlation Coefficient, PCC)、斯皮尔曼相关系数(Spearman Correlation Coefficients, SCC)、平均影响值算法(Mean Influence Value Algorithm, MIV)对输入参数进行相关性分析并实现降维,构建了基于支持向量回归的气浮单元水质预测模型。张鹏等^[16]基

* 收稿日期: 2025-07-28

作者简介: 侯兴汉,高级工程师,从事环境监测和环境工程研究, hxx2781@163.com; 刘国东(通信作者),教授,博士,从事地下水环境保护研究, 415145321@qq.com。

于最大信息系数 (Maximal Information Coefficient, MIC) 构建的支持向量回归模型在溶解氧预测中展现出最优拟合效果。然而,对于 SVR 在地下水监测数据插补中的应用探索仍显不足,尤其是较高缺失率条件下的模型适用性、插补精度与稳定性尚未形成系统化的验证体系。

本文聚焦于华北某石化区地下水污染监控需求,构建基于滑动窗口的支持向量回归的插补框架,采用多阶段网格搜索结合 k 折交叉验证的嵌套优化策略以选择合适的核函数和超参数,并对比线性插值插补结果,旨在验证高缺失率条件下支持向量回归模型修复水质数据的有效性。

1 模型理论研究

1.1 滑动窗口技术

滑动窗口技术 (Sliding Window, SW) 是一种将序列数据转换为监督学习问题的重要方法,滑动窗口将时间序列数据分割为按照一定方向滑动的连续窗口,窗口尺寸 (即单次输入数据个数) 可以直接固定,也可以不断扩容或缩小直至满足一定条件。结合支持向量回归 (SVR),采用固定窗口尺寸和固定步长分割的策略,可降低问题复杂度,提高计算效率,有序捕捉水质数据中的时间依赖性^[17-18]。

固定窗口尺寸和固定步长分割的滑动窗口的基本原理为:假定某一地下水水质指标有 n 个时间序列数据,滑动窗口尺寸为 m ,滑动步长为 t ^[19],记 $\mathbf{x}_i$ 为第 i 个窗口的时间序列向量 ($i \leq n - m + 1$), a_i 表示在 i 时间的水质监测值,则整个时间序列的向量 $\mathbf{x} = (a_1, a_2, \dots, a_n)$,第 i 个窗口的数据子向量为 $\mathbf{x}_i = (a_i, a_{i+1}, \dots, a_{i+m-1})$,第 $i+1$ 个窗口的数据子向量为 $\mathbf{x}_{i+1} = (a_{i+t}, a_{i+t+1}, \dots, a_{i+t+m-1})$,见图 1。

1.2 支持向量回归模型

支持向量回归运用了支持向量机的思想,其核心目标是通过非线性变换将数据映射到高维特征空

间,构造一个能容忍指定误差范围 (ε) 的最优超平面^[20]。SVR 在预测函数 $f(x)$ 的两侧构造宽度为 2ε 的间隔带用于拟合数据,为了增强对噪声的鲁棒性,当真实值 y 在间隔带之外 ($|y - f(x)| > \varepsilon$) 时,使用松弛变量 ξ 量化预测误差。此外,SVR 基于结构风险最小化原则,在保证预测误差最小的同时降低模型复杂度,优化参数 ω 和 b ,并构造最终模型函数 $f(x)$,从而提高泛化能力,见图 2。

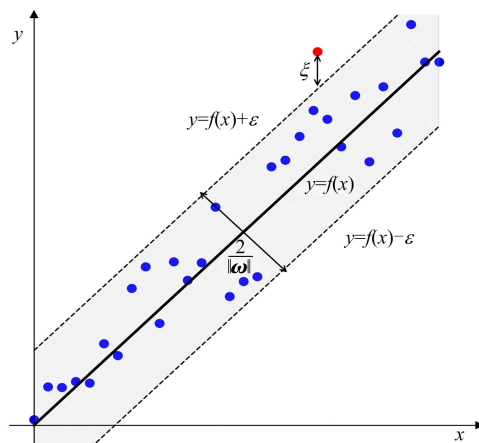


图2 最优超平面示意图

Fig.2 Schematic diagram of the optimal hyperplane

SVR 优化问题的线性回归数学形式及其约束条件为

$$\begin{cases} \min_{\omega, b, \xi, \xi^*} \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) \\ \text{s. t.} \begin{cases} y_i - (\omega^T \mathbf{x}_i + b) \leq \varepsilon + \xi_i \\ (\omega^T \mathbf{x}_i + b) - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \end{cases} \quad (1)$$

式中 $\mathbf{x}_i$ 为第 i 个样本的特征向量, y_i 为第 i 个样本目标值; ω 为权重向量; $\|\omega\|^2$ 表示支持向量之间的距离; C 为正则化参数; ξ_i, ξ_i^* 为上下边界松弛变量; b 为偏置项; ε 为不敏感损失函数。

对于非线性样本数据,SVR 将原始低维空间通过一个映射函数 ϕ 转换到高维空间以使样本数据线性可分,并在新空间中使用 SVR 线性回归思想进行求解。然而,直接进行这种映射会导致计算复杂度急剧增长,即“维数灾难”^[21]。为此,SVR 引入了核函数 $K(\mathbf{x}_i, \mathbf{x}_j)$,使得在低维空间的计算结果等于数据在高维特征空间中的内积,使非线性回归问题变成一个近似线性回归问题。目前,常用的核函数包括线性核函数、多项式核函数、高斯核函数、拉普拉斯核函数、Sigmoid 核函数等^[21]。经拉格朗日乘子

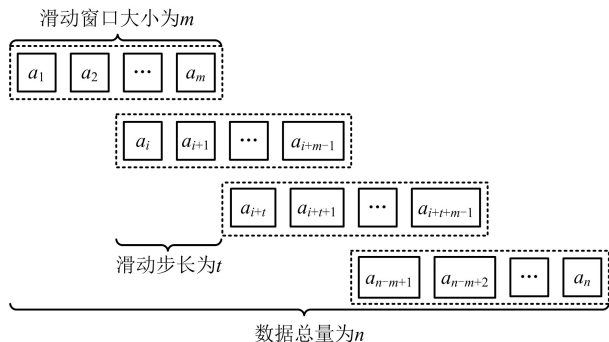


图1 滑动窗口原理图

Fig.1 Sliding window principal diagram

法变换、对偶原理的推导及核函数的引入,SVR 的最终形式为

$$f(x) = \sum_{i=1}^n (\alpha_i + \alpha_i^*) K(\mathbf{x}_i, \mathbf{x}_j) + b \quad (2)$$

式中 $\alpha_i, \alpha_i^* > 0$, 为拉格朗日乘子; $K(\mathbf{x}_i, \mathbf{x}_j)$ 为核函数。

支持向量回归模型在处理高维数据和小样本问题时表现优异,很好地契合了地下水水质数据样本总量较少、数据参数间存在复杂的非线性关系的情况。

1.3 参数优化

SVR 的性能高度依赖于核函数与超参数的选择。超参数是在模型训练前需预先设定的参数,会控制模型的结构与学习过程,其最优值无法从训练数据中直接获得,须通过优化算法确定。研究对核函数类型与正则化参数 C 、核函数系数 γ 、不敏感损失参数 ε 3 个核心超参数,采用分阶段网格搜索与 5 折交叉验证相结合的策略进行优化。

网格搜索(Grid Search, GS)是一种超参数优化方法,可通过给定参数范围建立超参数组合的网格,并遍历所有网格数据选取参数值^[22]。为了减少计算量,网格搜索可采用分阶段参数搜索策略。第一阶段粗略搜索确定核函数类型并定位参数的大致范围,第二阶段固定核函数并在粗搜结果的范围内寻找参数最优解。 k 折交叉验证法是一种评估模型泛化能力的无偏估计统计方法,即将训练集均匀划分为 k 个互斥子集,依次取其中一个子集作为验证集,其余 $k-1$ 个子集为训练集用于模型训练。重复该过程 k 次后,将 k 次泛化误差的平均值作为最终交叉验证误差。采用优化后的 GS 法得到不同的参数组合后,对每一个组合训练模型通过交叉验证评估性能,选择表现最好的最佳参数组合进行研究。

2 应用

2.1 研究区概况

华北某石化企业是我国大型石化企业之一。该石化企业所在区域地形以山区为主,地势总体上西高东低、北高南低。西北一带为海拔 500 ~ 1 000 m 的低山区,东南为海拔 150 ~ 200 m 的丘陵区,并向东南呈簸箕状与山前倾斜平原相连,海拔在 30 ~ 80 m,地表起伏不平,冲沟较发育。石化区地下水类型为松散岩类孔隙水和基岩裂隙水。其中,松散岩类孔隙水赋存于第四纪砂土卵石层、砂土层和粉土层,富水性差;基岩裂隙水主要赋存于风化裂隙,构造裂

隙发育较弱,富水性差。第四系孔隙水与基岩裂隙水构成复合含水层组,水位埋深及水量变化幅度大,且受地形控制由北向南流动。研究区属于温带季风气候区,降水集中在 6—9 月,年内分配不均的降水特征对地下水动态平衡产生显著影响。

然而,该石化企业在创造巨大经济效益的同时,不可避免地对地下水环境产生了一定影响。根据企业的地下水例行监测数据分析,石化区内地下水环境质量总体较好,但高锰酸盐指数、pH 值、总硬度及锰质量浓度等部分水质指标存在超标现象。

2.2 水质监测数据

为了掌握石化区地下水环境质量,了解石化生产对地下水环境的影响,防止由于生产活动对地下水环境造成大范围污染,该企业在石化区内陆续布置了 30 口地下水环境质量观测井,并选取《地下水质量标准》中的 24 项常规指标和 20 项特征指标每季度取样检测一次。受天气条件、检测设备、水文地质条件(部分井在某些年份的枯水期干涸)以及早期对地下水环境保护重视程度低等多种因素影响,导致观测数据序列不连续,缺失较多。高锰酸盐指数作为有机污染物的综合指标,是衡量石化区地下水有机污染的核心参数;pH 值能反映水体酸碱平衡,可直接指示石化生产过程中潜在的酸碱污染;总硬度可表征水体中钙镁离子的含量,并用于评估锅炉排水、循环冷却水排污等石化生产环节对水体的影响;锰作为石化工艺的特征污染物,其质量浓度变化对污染溯源具有特殊指示意义。为了解历年各季度地下水的环境状况及其与生产活动的关系,高锰酸盐指数、pH 值、总硬度及锰质量浓度指标监测数据需为完整序列,为此对其缺失数据进行插补。

在水文研究中,数据缺失率在 10% 以内、连续性良好的水文系列,采用线性插值、KNN 插值等常规插补法即可满足数据的修复需求。本文选取 2013 年第二季度至 2024 年第二季度共计 45 个季度观测周期为研究时段,以高锰酸盐指数、pH 值、总硬度及锰质量浓度 4 个地下水水质指标作为分析对象进行研究。

相比于一般的水文系列资料,本文研究的地下水水质监测频次低、数据量较小。同时,各观测井中的 4 个水质指标数据总体缺失比例高,缺失率集中在 10% ~ 25%,显著高于常规的水文序列。此外,数据缺失结构复杂,既存在完全随机缺失,也存在由枯水期井干涸、设备故障等导致的连续多个季度的连续缺失片段。这种高比例与非随机性并存的缺失

特性,对插补方法的鲁棒性提出了更高要求。

2.3 地下水水质数据插补流程

本文基于 Python 3.9.13 编程环境,使用 pandas、numpy、sklearn、openpyxl 4 个库,构建了一套完整的地下水水质数据插补算法。该算法通过滑动窗口技术将时间序列数据转换为适合 SVR 模型训练的输入输出对,并使用分阶段网格搜索和 5 折交叉验证方法为 SVR 模型选择合适的超参数。此外,为了保证模型的完整性,算法添加了备用填充方法以解决有效样本数不足的情况,最终生成完整的地下水水质数据集,具体步骤如下。

(1)数据预处理。将每个水质指标作为一个独立的工作表,按列整理各个观测井时间序列的水质监测数据。对水质数据进行缺失值标记及异常值处理,将原始数据转化为适合模型训练的规范化时序特征。对于锰质量浓度,部分数据低于方法检出限,其实际浓度无法准确定量。为确保数据的真实性,此类数据按缺失值处理,未纳入统计。本文统计原始数据缺失率及连续缺失长度,以为后续插补提供数据基础。

(2)滑动窗口数据集创建。为了构建滑动窗口,根据数据缺失的情况,确保训练数据量,须生成尽可能多的训练样本以缓解小样本问题。本文将滑动窗口尺寸设置为 4,滑动步长设置为 1,采用逆向窗口预测机制进行研究。此外,以后续 3 个观测值($x_{i+1}, x_{i+2}, x_{i+3}$)作为输入特征,当前观测数据(x_i)为预测目标进行探索。

(3)SVR 模型训练与预测。按照数量比为 4:1 的比例将数据集划分为训练集和验证集,构建 SVR 模型,在模型训练前利用 standardscaler 对数据进行 Z-score 标准化处理。标准化处理为

$$Z = \frac{x - \mu}{\sigma} \quad (3)$$

式中 x 为某一水质数据监测值; μ 为该水质指标单个观测井的样本均值; σ 为样本标准差。

采用分阶段网格搜索策略对 SVR 模型的超参数进行调优。第一阶段进行粗粒度搜索,确定 SVR 模型的核函数 kernel、正则化参数 C 、核函数系数 γ 和不敏感损失参数 ε 的范围,网格搜索策略的参数

范围见表 1。第二阶段在该范围基础上进行精细化搜索;kernel 固定为第一阶段的最佳 kernel 值; C 以第一阶段最佳值 C_1 为中心,取 $[0.5C_1, 2C_1]$ 内的 5 个等距点进行优化; ε 在 $[0.1, 1]$ 内取 5 个等距点进行细调; γ 需根据第一阶段结果,若最佳取值为具体数值,则在 $[0.1, 1]$ 内取 5 个等距点继续调优,若值为字符串 'scale' 或 'auto' 则在此阶段保持不变。

以平均绝对误差 (Mean Absolute Error, E_{MAE}) 和均方根误差 (Root Mean Square Error, E_{RMSE}) 为评估指标,通过 5 折交叉验证评估模型性能,最终确定最优参数组合。 E_{MAE} 和 E_{RMSE} 计算公式分别为

$$E_{MAE} = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (4)$$

$$E_{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (5)$$

式中 n 为样本数目; $\hat{y}_i$ 和 y_i 分别表示第 i 个水质数据的插补预测值和实际监测值。为了确保插补的有效性,系统过滤了包含多个缺失值的窗口段,最小有效样本阈值设置为 10。

(4)备用填充方法。对于有效样本数不足的情况,若缺失点前或后存在有效数据,且水质指标具有时间稳定性,优先使用向前填充和向后填充填补缺失值;若仍有缺失值,在观测井历史数据大于最小有效样本阈值前提下使用该水质指标对应观测井的全局均值进行填充。

(5)插补效果对比验证。为了评估插补结果的可靠性,本文进一步采用线性插值法对原始数据进行插补,并将插补结果与 SVR 模型的预测结果进行对比分析。线性插值公式为

$$x = x_1 \frac{x_2 - x_1}{t_2 - t_1} \times (t - t_1) \quad (6)$$

式中 x 为 t 时刻的缺失值; x_1 和 x_2 分别为 t_1 和 t_2 时刻的地下水水质监测值。

3 结果与讨论

根据华北某石化区多年地下水水质监测数据集,分别采用基于滑动窗口的支持向量回归模型 (SVR) 对高锰酸盐指数、pH 值、总硬度及锰质量浓度 4 个水质指标进行插补,并通过 5 折交叉验证方

表 1 第一阶段网格搜索参数范围(粗粒度)

Table 1 Parameter ranges for grid search in the first stage (coarse-grained)

参数	kernel	C	γ	ε
范围	RBF, linear	0.1, 1, 10	'scale', 'auto', 0.1, 1	0.1, 0.5, 1

法对插补质量进行评估。结果显示,pH 值插补效果最优,73.3% 观测井的 E_{MAE} 均小于 0.5, E_{RMSE} 均小于 0.6;54.5% 观测井高锰酸盐指数的 E_{MAE} 小于 1 mg/L, E_{RMSE} 小于 1.1 mg/L;45.5% 观测井总硬度的 E_{MAE} 小于 50 mg/L, E_{RMSE} 小于 55 mg/L;40.9% 观测井锰质量浓度的 E_{MAE} 小于 0.1 mg/L, E_{RMSE} 小于 0.1 mg/L。采用 Kolmogorov-Smirnov 双样本检验,发现高锰酸盐指数、pH 值、总硬度、锰质量浓度通过检验 ($p > 0.05$) 的观测井比例分别为 31.8%、50%、54.5%、31.8%。本文选取缺失率最大观测井 K2#、缺失率最小观测井 7#、缺失率适中观测井 1#作为代表性观测井进行研究,4 个水质指标的最优 SVR 模型超参数组合情况见表 2,采用 SVR 方法插补得到的结果见图 3~6。

为了定量评估插补效果,计算各个观测井中水质指标插补数据均值与原始数据均值的绝对偏差,并以数据缺失率为横坐标、插补偏差为纵坐标,绘制两种方法插补的散点分布图。为了分析插补偏差随缺失率变化的非线性关系及局部变化趋势,研究采用局部加权回归散点平滑法 (Locally Estimated Scatterplot Smoothing, LOESS),拟合高锰酸盐指数、pH 值、总硬度及锰质量浓度 4 个关键水质指标的变化趋势曲线,见图 7。对比 SVR 法和线性插值法结果可以发现,除了锰质量浓度,其余水质指标 SVR 插补的散点大多位于线性插值的下方,即 SVR 插补误差趋势线总体均位于线性插补趋势线下方,表明使用 SVR 法插补结果的插补偏差更小,插补效果

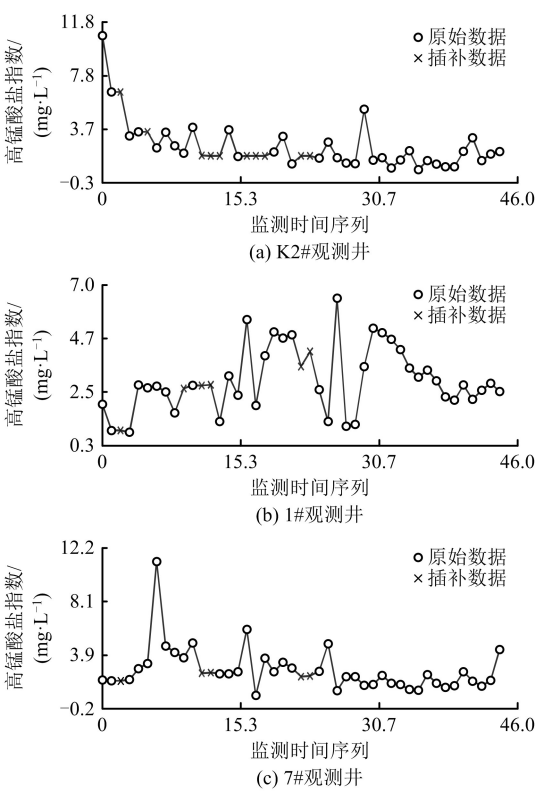


图 3 代表性观测井高锰酸盐指数插补结果
Fig. 3 Interpolation results for permanganate index in representative observation wells

更好。
为了消除量纲影响并综合评估不同方法的插补稳定性,将所有插补偏差进行归一化处理,见图 8。为了更精确地量化其统计特征,表 3 给出了相应偏

表 2 代表性观测井各水质指标最优 SVR 模型超参数组合

Table 2 Optimal SVR model hyperparameter combinations for key water quality indicators in representative observation wells					
水质指标	观测井	核函数 kernel	正则化参数 C	核系数 γ	不敏感损失 ε
高锰酸盐指数	K2#	RBF	0.05	0.1	0.55
	1#	RBF	0.875	1	0.1
	7#	RBF	5.0	0.8	0.2
pH 值	K2#	RBF	5	1	0.1
	1#	RBF	0.05	scale	0.775
	7#	RBF	5	0.775	0.1
总硬度	K2#	RBF	20	1	1
	1#	RBF	20	1	1
	7#	RBF	20	scale	0.1
锰质量浓度	K2#	RBF	20	0.325	0.325
	1#	RBF	0.05	scale	1
	7#	RBF	0.05	1	0.1

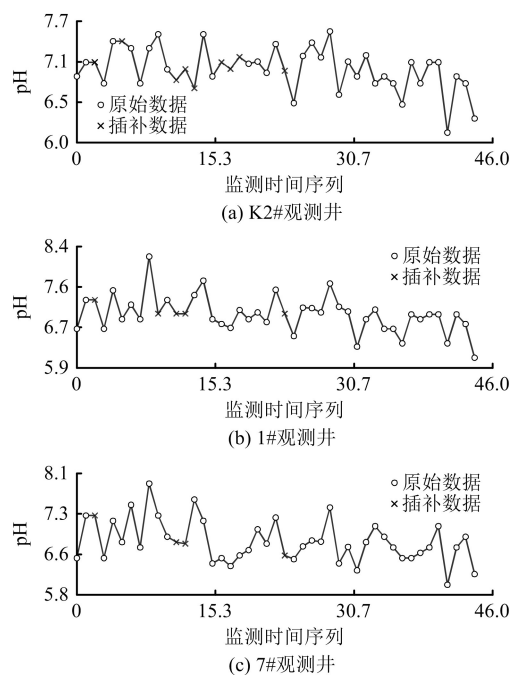


图4 代表性观测井 pH 值插补结果

Fig. 4 Interpolation results for total hardness in representative observation wells

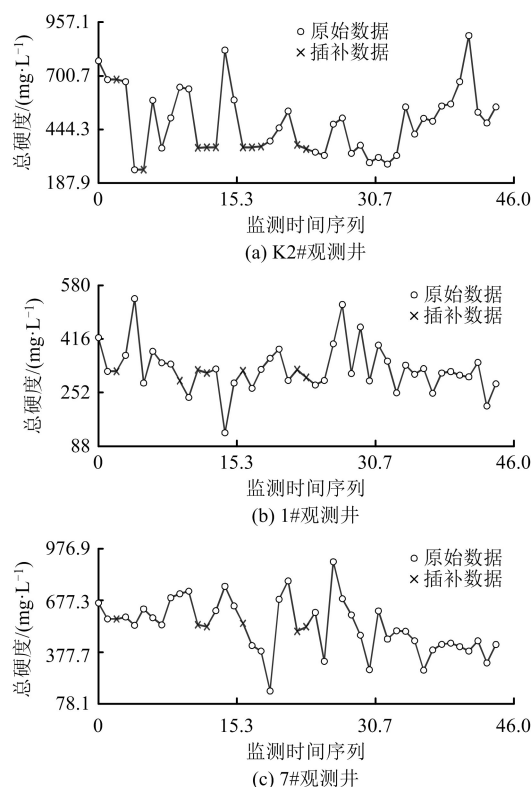


图5 代表性观测井总硬度插补结果

Fig. 5 Interpolation results for total hardness in representative observation wells

差的中位数、平均值、标准差和四分位距。本文将不同观测井数据缺失率第 25 ~ 75 百分位数 (P25 - 2470

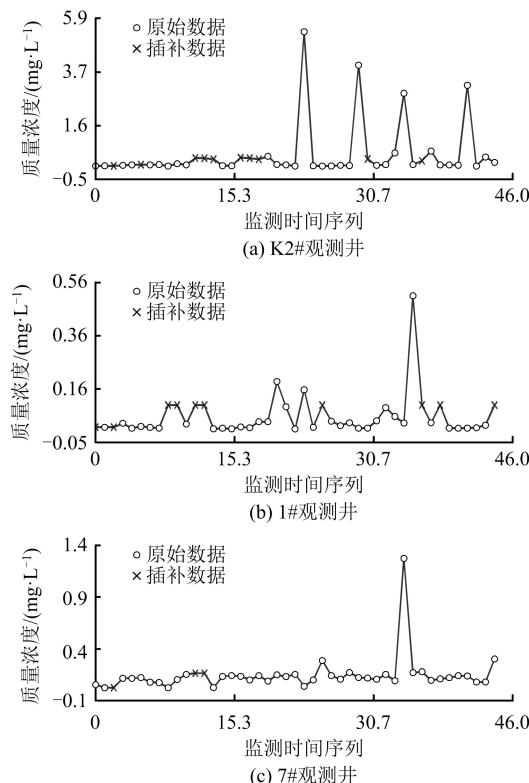


图6 代表性观测井锰质量浓度插补结果

Fig. 6 Interpolation results for manganese mass concentration in representative observation wells

P75) 的范围定义为“主分布区间”, 则高锰酸盐指数、pH 值及总硬度数据的缺失率主分布区间分别为 11.1% ~ 15.6%、8.9% ~ 13.3% 和 13.3% ~ 17.2%, 表现出较高的数据集中度。然而锰质量浓度数据的缺失率主分布区间为 6.7% ~ 28.9%, 缺失率范围显著扩大, 这可能与锰质量浓度的检测限较高及分析方法灵敏度差异有关。分析表 3 与图 8 可知, 相较于线性插值, 高锰酸盐指数的偏差中位数降低 12.9%, 平均值降低 30.0%, 标准差降低 37.3%, 四分位距 (Interquartile Range, IQR) 缩减 46.2%; 总硬度的偏差中位数降低 38.3%, 平均值降低 28.5%, IQR 缩减 65.2%, 表明使用 SVR 插补的精度和稳定性均优于线性插值。对比 pH 值插补结果则可以看出, 两种方法的各项偏差统计量均较为接近, 但 SVR 在 pH 值指标中仅产生 1 个异常值, 显著少于线性插值的 3 个。锰质量浓度用线性插值的偏差中位数低于 SVR, 但两种方法的偏差平均值和标准差接近, 且 SVR 的四分位距相较于线性插值更低, 表明其插补结果中较大偏差相对更少。综合分析表明, SVR 方法在多数水质指标上具有更好的插补精度和稳定性, 尤其适用于高锰酸盐指数和总

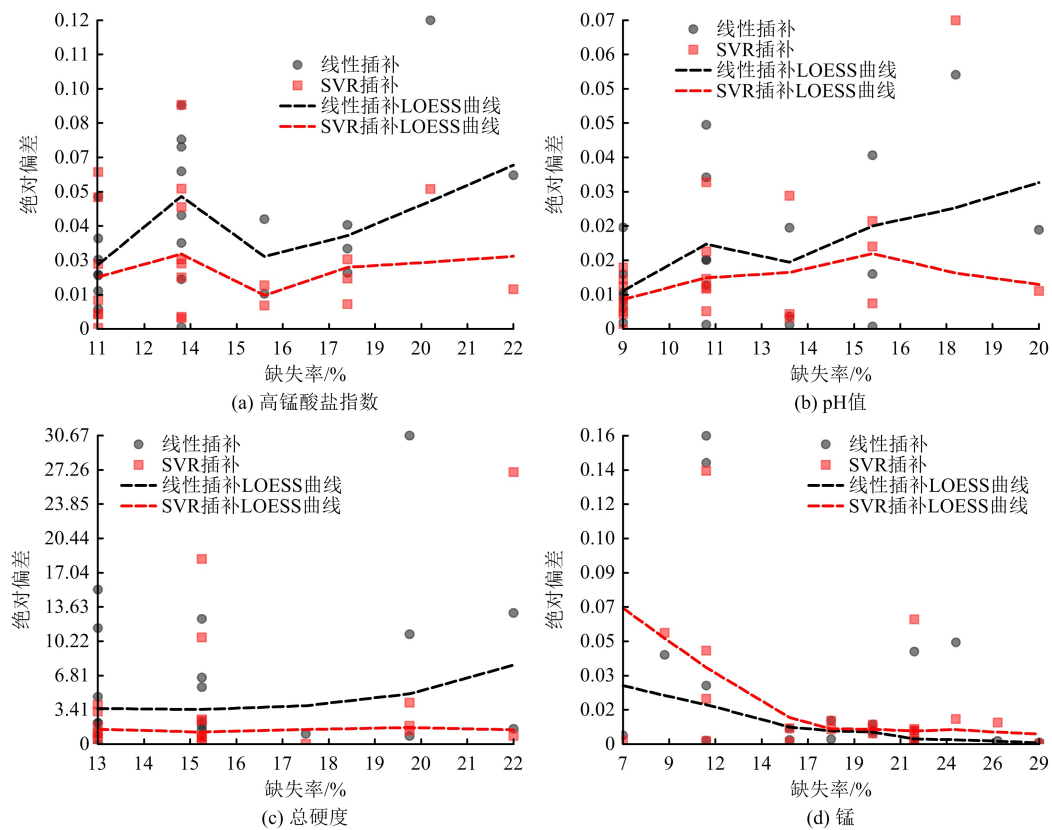


图 7 不同水质指标在不同缺失率下插补效果对比

Fig. 7 Comparison of interpolation effects for different water quality indicators under different missing rates

表 3 不同水质指标插补效果统计特征对比

Table 3 Comparison of statistical characteristics of interpolation effects for different water quality indicators

水质指标	缺失率主 分布区间/%	插补方法	偏差中位数	偏差平均值	偏差标准差	第一四分位数	第三四分位数
高锰酸盐 指数	11.1 ~ 15.6	线性插值	0.014 0	0.020 7	0.018 5	0.006 9	0.032 2
		SVR 插值	0.012 2	0.014 5	0.011 6	0.006 9	0.020 5
pH 值	8.9 ~ 13.3	线性插值	0.013 2	0.025 0	0.031 1	0.006 3	0.032 3
		SVR 插值	0.016 9	0.025 7	0.031 5	0.005 9	0.032 5
总硬度	13.3 ~ 17.2	线性插值	0.011 8	0.022 1	0.024 4	0.006 1	0.031 1
		SVR 插值	0.007 3	0.015 8	0.025 3	0.003 9	0.012 6
锰	6.7 ~ 28.9	线性插值	0.001 2	0.006 1	0.009 9	0.000 3	0.005 0
		SVR 插值	0.002 2	0.006 0	0.010 0	0.000 6	0.003 5

硬度等指标的缺失值插补。

综上所述,采用 SVR 插值方法对高锰酸盐指数、pH 值、总硬度的插补效果总体上更接近于实际监测值,具有更高的稳定性,优于线性插值的插补效果。然而,在锰质量浓度的插补中,线性插值法稍显优势,经统计,各观测井将低于检出限的锰质量浓度

作为缺失值处理的数量在 0~9 个,平均每个观测井的处理数量约为 4 个;各观测井由于检出限而按缺失值处理的锰质量浓度指标数占其总缺失量的 0~73%,其中不小于 50% 的观测井数量为 10 口,此时用时间序列数据直接进行插补将会使结果存在偏差,导致 SVR 插补方法在锰质量浓度指标中的表现

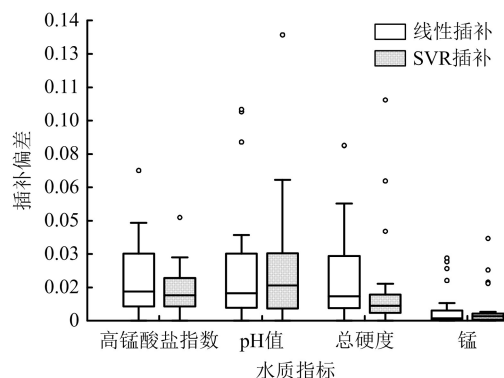


图8 不同水质指标归一化插补偏差分布对比

Fig.8 Comparison of normalized interpolation deviation distributions for different water quality indicators

相对受限。因此,在后续研究中,须单独分析由检出限导致的数据插补问题。

4 结论

本文针对石化区地下水多年水质监测数据高缺失率的问题,提出了基于滑动窗口技术与支持向量回归(SVR)相结合的数据插补模型,并以华北某石化区多个水质指标的实际监测数据进行了验证,得到以下结论。

(1)支持向量回归模型可用于数据规模较小、缺失率较高、连续缺失数据较多的时间序列插补,在多个关键指标插补精度上显著优于传统方法。SVR对总硬度的插补结果偏差中位数较线性插值降低38.3%,平均值降低28.5%,四分位距缩减65.2%;高锰酸盐指数与pH值同样在多个误差指标上优于线性插值。

(2)引入滑动窗口技术能提高局部结构的捕捉能力,增强模型的鲁棒性,稳定插补结果。

(3)应用多阶段网格搜索与5折交叉验证策略,在有限样本条件下合理选择核函数与超参数,提高了建模精度与计算成本,并获得了较好的计算效率。

(4)在对锰质量浓度缺失值进行插补时,研究对锰质量浓度由检出限问题导致的数据插补问题进行缺失值标记,使得该指标插补效果相对受限,SVR插补效果略低于线性插值,未来可考虑结合研究区相关资料引入贝叶斯统计对该部分进行单独分析;此外,在样本量进一步扩大时,可对比本文方法与RF、LSTM等更复杂的机器学习模型在不同缺失场景下的性能差异,以更全面地评估各类方法的性能

边界及其适用范围。

参考文献 (References):

- [1] 程一蓓. 浅谈地下水环境监测存在的主要问题及技术性建议[J]. 皮革制作与环保科技, 2024, 5(23): 69-71.
CHENG Y B. Brief discussion on the main problems in groundwater environmental monitoring and technical suggestions[J]. Leather Manufacture and Environmental Technology, 2024, 5(23): 69-71.
- [2] 熊中敏, 郭怀宇, 吴月欣. 缺失数据处理方法研究综述[J]. 计算机工程与应用, 2021, 57(14): 27-38.
XIONG Z M, GUO H Y, WU Y X. Research review on missing data processing methods [J]. Computer Engineering and Applications, 2021, 57(14): 27-38.
- [3] BERTSIMAS D, PAWLOWSKI C, ZHUO Y D. From predictive methods to missing data imputation: an optimization approach[J]. Journal of Machine Learning Research, 2018, 18: 196.
- [4] 陈湛峰, 李晓芳. 基于时间序列分析的水质自动监测数据插补方法[J]. 环境与发展, 2023, 35(5): 48-52.
CHEN Z F, LI X F. Interpolation method for automatic water quality monitoring data based on time series analysis [J]. Environment and Development, 2023, 35(5): 48-52.
- [5] LARSON D M, BUNGULA W, LEE A, et al. Reconstructing missing data by comparing interpolation techniques: applications for long-term water quality data [J]. Limnology and Oceanography: Methods, 2023, 21(7): 435-449.
- [6] CHOI J, LIM K J, JI B. Robust imputation method with context-aware voting ensemble model for management of water-quality data [J]. Water Research, 2023, 243: 120369.
- [7] 高海燕, 刘畅, 马文娟. 地表水水质监测时序数据插补方法比较及应用[J]. 水文, 2024, 44(5): 61-67.
GAO H Y, LIU C, MA W J. Comparison and application of interpolation methods for time-series data in surface water quality monitoring[J]. Journal of China Hydrology, 2024, 44(5): 61-67.
- [8] 杨国华, 李婉露, 孟博. 基于机器学习方法的地下水氨氮时空分布规律[J]. 吉林大学学报(地球科学版), 2022, 52(6): 1982-1995.
YANG G H, LI W L, MENG B. Spatiotemporal distribution of ammonia nitrogen in groundwater based on

- machine learning methods[J]. Journal of Jilin University (Earth Science Edition), 2022, 52(6): 1982–1995.
- [9] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273–297.
- [10] DRUCKER H, BURGESS C J C, KAUFMAN L, et al. Support vector regression machines[C]//MOZER M C, JORDAN M I, PETSCHKE T. Advances in Neural Information Processing Systems 9. Cambridge: MIT Press, 1997: 155–161.
- [11] 刘杰, 李佟, 李军. 基于改进支持向量回归机的污水处理厂出水总氮预测模型[J]. 环境工程学报, 2018, 12(1): 119–126.
- LIU J, LI T, LI J. Prediction model of total nitrogen in effluent from wastewater treatment plant based on improved support vector regression[J]. Chinese Journal of Environmental Engineering, 2018, 12(1): 119–126.
- [12] 薛同来, 赵冬晖, 韩菲. 基于 GA 优化的 SVR 水质预测模型研究[J]. 环境工程, 2020, 38(3): 123–127.
- XUE T L, ZHAO D H, HAN F. Research on water quality prediction model based on GA optimized SVR[J]. Environmental Engineering, 2020, 38(3): 123–127.
- [13] 张秀菊, 安焕, 赵文荣, 等. 基于支持向量机的水质预测应用实例[J]. 中国农村水利水电, 2015(1): 85–89.
- ZHANG X J, AN H, ZHAO W R, et al. Application example of water quality prediction based on support vector machine[J]. China Rural Water and Hydropower, 2015(1): 85–89.
- [14] 石翠翠, 刘媛华, 陈昕. 基于粒子群算法优化支持向量回归的水质预测模型[J]. 信息与控制, 2022, 51(3): 307–317.
- SHI C C, LIU Y H, CHEN X. Water quality prediction model based on particle swarm optimization and support vector regression[J]. Information and Control, 2022, 51(3): 307–317.
- [15] 陈霖, 晏欣, 唐智和, 等. 基于优化支持向量回归机的气浮单元水质预测模型[J]. 工业水处理, 2025, 45(5): 157–165.
- CHEN L, YAN X, TANG Z H, et al. Water quality prediction model for air floatation units based on optimized support vector regression machines[J]. Industrial Water Treatment, 2025, 45(5): 157–165.
- [16] 张鹏, 梅书浩, 石成春, 等. 基于机器学习的水库溶解氧预测模型比较研究[J]. 华北水利水电大学学报(自然科学版), 2025, 46(1): 87–95.
- ZHANG P, MEI S H, SHI C C, et al. Comparative study on dissolved oxygen prediction models in reservoirs based on machine learning[J]. Journal of North China University of Water Resources and Electric Power (Natural Science Edition), 2025, 46(1): 87–95.
- [17] 谭宏强, 牛强. 基于滑动窗口及局部特征的时间序列符号化方法[J]. 计算机应用研究, 2013, 30(3): 796–798.
- TAN H Q, NIU Q. Symbolization method for time series based on sliding window and local features[J]. Application Research of Computers, 2013, 30(3): 796–798.
- [18] 石秀峰, 王进, 揣新, 等. 基于改进 PSO–ELM 的坑湖水质预测与评价[J]. 合肥工业大学学报(自然科学版), 2025, 48(2): 145–150.
- SHI X F, WANG J, CHUAI X, et al. Water quality prediction and evaluation of pit lakes based on improved PSO–ELM[J]. Journal of Hefei University of Technology (Natural Science Edition), 2025, 48(2): 145–150.
- [19] 郭利进, 李博伦. 基于 GNFA–SVR 污水出水总氮软测量研究[J]. 工业水处理, 2022, 42(10): 111–117.
- GUO L J, LI B L. Soft measurement of total nitrogen in wastewater effluent based on GNFA–SVR[J]. Industrial Water Treatment, 2022, 42(10): 111–117.
- [20] 杜尊峰, 樊涛, 赵亚琦, 等. 基于多核 SVR 集成模型与主动学习的结构可靠性分析方法[J/OL]. 工程力学: 1–13[2025–09–25]. <https://doi.org/10.6052/j.issn.1000-4750.2024.01.0060>.
- DU Z F, FAN T, ZHAO Y Q, et al. Structural reliability analysis method based on multi-kernel SVR ensemble model and active learning[J/OL]. Engineering Mechanics: 1–13[2025–09–25]. <https://doi.org/10.6052/j.issn.1000-4750.2024.01.0060>.
- [21] 孟滔. 支持向量回归水质预测模型的研究进展[J]. 绿色科技, 2021, 23(8): 77–79.
- MENG T. Research progress of water quality prediction models based on support vector regression[J]. Journal of Green Science and Technology, 2021, 23(8): 77–79.
- [22] SHAMS M Y, ELSHEWEY A M, EL-KENAWY E S M, et al. Water quality prediction using machine learning models based on grid search method[J]. Multimedia Tools & Applications, 2024, 83(12): 35307–35334.

Research on the application of Support Vector Regression (SVR) for data interpolation in groundwater quality monitoring series with high missing rates

HOU Xinghan¹, PENG Tishan², YU Bo¹, LI Xinfeng¹, LIU Guodong²

(1 Beijing Feiyan Petrochemical Environmental Protection Science and Technology Development Co., Ltd., Beijing 102500, China;

2 College of Water Resource and Hydropower, Sichuan University, Chengdu 610065, China)

Abstract: A petrochemical area in North China encountered significant challenges due to extensive missing groundwater quality monitoring data and a limited sample size. This study developed a data imputation method that integrates sliding window processing with Support Vector Regression (SVR). This approach was utilized to reconstruct missing time-series data for four key water quality indicators: permanganate index, pH, total hardness, and manganese. Based on the characteristics of the missing data, a sliding window size of 4 was employed, with a sliding step size of 1. The model used backward window prediction, leveraging the subsequent three observations to predict the current value. The study employed a phased grid search for tuning the hyperparameters of SVR. In the initial phase, coarse-grained range searches were conducted for the kernel function, regularization parameter C , kernel weight gamma, and insensitivity loss parameter epsilon (ε). Following this phase, a refined parameter search was performed within narrowed ranges. Using five-fold cross-validation, the evaluation process identified the optimal parameter combination based on the minimum mean absolute error and root mean square error. Additionally, the study utilized the absolute deviation between the mean values of the interpolated water quality indicators and the mean of the original data from each observation well as an evaluation metric. The study also conducted a comparative analysis between the SVR interpolation method and conventional linear interpolation. The results indicate that the SVR model significantly outperforms linear interpolation for the interpolation of permanganate index, pH, and total hardness, with the best performance observed specifically in total hardness interpolation. However, for the manganese indicator, the temporal continuity was disrupted by a substantial number of values falling below the detection limit. This posed challenges for model training, resulting in the linear interpolation method performing slightly better than SVR in this particular case. This study offers a reliable method for imputing missing values in time-series-correlated environmental monitoring data. It is recommended that future research adopt Bayesian methods to address issues related to detection limits. Additionally, as sample sizes increase, the performance of more complex models, such as Random Forest and Long Short-Term Memory, should be compared across various missing data scenarios.

Key words: environmental engineering; petrochemical site groundwater; water quality monitoring; support vector regression; data interpolation