基于时序分解特征的水质溶解氧预测

李 慧

(常州大学外国语学院，江苏 常州 213164)

摘要： 为提高水体生态系统的稳定性，提出一种基于变分模态分解(VMD)的水质溶解氧混合预测模型，利用VMD将溶解氧进行分解得到多个本征模态函数(IMFs)，根据不同IMFs的特点，分别建立不同的模型。非线性序列部分使用主成分(PCA)和粒子群(PSO)优化最小二乘向量机(LSSVM)模型；线性序列部分使用差分整合移动平均自回归模型(ARIMA)。将各部分预测结果结合起来，得到溶解氧的预测值。将本文模型运用于京杭大运河常州段某监测点进行验证，结果显示平均绝对误差为0.168、均方根误差为0.211、平均绝对百分率误差为4.576，混合预测模型具有较高的预测精度，能够满足现代化水质管理的高要求。

关键词： 水质溶解氧预测；时序分解；变分模态分解；最小二乘支持向量机；差分整合移动平均自回归

中图分类号： X832；TP391 **文献标志码：** A **文章编号：** 1671－1807(2024)24－0273－07

水质中溶解氧(DO)的含量直接影响着水中生物生存及水质质量，利用现有的数据预测水体中溶解氧含量对水质环境改善具有重大的指导意义。溶解氧时间序列具有随机性、非线性、波动性和非平稳等特点，将溶解氧分解成多个分量进行单独研究，才能更好地描述这些特性。变分模态分解(variational mode decomposition，VMD)算法是一种变分问题自适应、准正交、完全非递归的分解方法，能够很好地降低时间序列不稳定性[1]。姜建国等[2]通过VMD算法将光伏序列进行分解，减少了随机分量和噪声数据对预测结果的影响，使得模型预测效果最佳。罗潇远等[3]使用VMD对风电功率数据进行分解，为超短期风电功率预测模型提供了较为稳定的数据序列。

目前，对于数据的预测方法主要包括时间序列模型[4-6]、人工神经网络模型[7-10]和支持向量机模型[11-12]等。支持向量机模型计算开销小，具有较强的非线性计算能力，最小二乘法支持向量机模型(least squares vector machine model，LSSVM模型)是对支持向量机模型的优化，改善了模型性能。隗海林等[13]使用LSSVM对怠速时间进行预测。Zhang等[14]将LSSVM用于风电功率预测。差分整合移动平均自回归模型(autoregressive integrated moving average model，ARIMA模型)是时间序列预测分析方法之一，能较好地对差分平稳时间序列拟合。吴静等[15]通过将ARIMA模型用于溶解氧含量预测，预测精度显著提高。刘畅和李蕴峰[16]使用ARIMA模型来预测海南省旅游数据和新冠肺炎之间的影响。

本文为研究水质中溶解氧变化，提出了“分解-预测-重构”[17-18]组合预测模型。使用VMD将溶解氧序列分解成多个分量，分解得到的本征模态函数(intrinsic mode function，IMFs)结合各自特征使用不同的预测方法，其中线性序列使用ARIMA进行预测，非线性序列使用粒子群(particle swarm optimization，PSO)优化的LSSVM进行预测，并在非线性序列预测前采用主成分分析方法(principal component analysis，PCA)[19]对输入数据进行筛选，消除变量间的冗余信息，最后将各分量的预测结果相加得到最终溶解氧的预测值。模型运用于京杭大运河常州段某监测点，实验结果表明预测精度有明显提高，方法有效可行，研究结果可为水质管理提供依据。

1 理论和方法

1.1 变分模态分解

使用合适的分解算法分解溶解氧序列，才能更好地体现溶解氧特征。VMD算法具有自适应、非递归等优点，其实质上就是构造和求解变分的过程。当VMD求解约束变分问题时，模型可以写成：

收稿日期： 2024-07-24

作者简介： 李慧(1996—)，女，江苏南通人，硕士研究生，研究方向为信息智能化。

$$\min\left(\sum_{k=1}^{K}\left\|\partial_t\left\{\left[\delta(t)+\frac{\mathrm{j}}{\pi t}\right]u_k\right\}\mathrm{e}^{-\mathrm{j}\omega_k t}\right\|_2^2\right)$$
$$\text{s.t.}\sum_k u_k = f \tag{1}$$

式中：t 为时间变量；u_k 为带宽有限的第 k 个 IMF；k 为相应的中心频率；f 为原始输入信号；ω_k 为各模态分量的中心频率；j 为虚数单位，用于构造解析信号；δ 为狄拉克函数，用于构造约束变分问题。

为得到最优解，引入二次惩罚项 α 和拉氏乘子 $\lambda(t)$ 通过交替对 u_k、ω_k 和 λ 更新，最终求解非应变变分问题。VMD 通过原始信号的特性划分频带，对每个 IMF 及其对应的中心频率进行连续更新，最终根据约束条件实现了信号的分解。

1.2 非线性序列预测算法

1.2.1 主成分分析

水质中各因素之间存在着相关性，为消除冗余，需要对数据进行降维。PCA 是一种重要的统计方法，其通过将可能存在关联的众多变量转化为综合性强的变量，转化后的变量称为主成分。设 $\boldsymbol{X}_1$，$\boldsymbol{X}_2$，…，$\boldsymbol{X}_p$ 为随机变量，记为 $\boldsymbol{X}_i=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_p\}^{\mathrm{T}}$，$\boldsymbol{X}$ 中 n 个样品的简单随机样本 $\boldsymbol{x}_i=\{x_{i1},x_{i2},\cdots,x_{ip}\}^{\mathrm{T}}(i=1,2,\cdots,n)$，构造样本矩阵并对其进行标准化变换，得标准化矩阵 Z，并计算出标准化矩阵 Z 的相关系数矩阵并其特征方程，得到 p 个特征根，计算其贡献比 τ_i 和累计贡献比 η_i 如下：

$$\tau_i=\frac{\lambda_i}{\sum_{j=1}^{p}\lambda_j},\quad i=1,2,\cdots,p \tag{2}$$

$$\eta_i=\frac{\sum_{j=1}^{i}\lambda_j}{\sum_{j=1}^{p}\lambda_j},\quad i=1,2,\cdots,p \tag{3}$$

其中，对于特征值 λ_j，$j=1,2,\cdots,n$，都有对应的单位特征向量 $\boldsymbol{b}_j$。

1.2.2 最小二乘支持向量机

非线性序列反映了溶解氧的非线性、非平稳性，对其需要使用合适的预测方法。LSSVM 是用于解决模式分类问题和函数估计问题的新型支持向量机方法，在非线性序列预测方面有一定的优势。对于 LSSVM，输入为 $\{\boldsymbol{x}_i,\boldsymbol{y}_i\}_{i=1}^{N}$ 的原问题可以写成：

$$f(w,e)=\frac{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+\gamma\sum_{i=1}^{N}\boldsymbol{e}_i^2}{2}$$
$$\text{s.t. }\boldsymbol{y}_i=\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x}_i)+\boldsymbol{b}+\boldsymbol{e}_i \tag{4}$$

式中：$\boldsymbol{w}$ 为权重；γ 为正则化参数；$\boldsymbol{b}$ 为偏置；$\boldsymbol{e}_i$ 为回归误差。引入拉氏乘子 α_i，将带有约束的问题转变为无约束的问题，并根据最优化理论转化为线性方程组，求解最终得到 LSSVM 的函数：

$$K(\boldsymbol{x},\boldsymbol{x}_i)=\exp[-\|\boldsymbol{x}-\boldsymbol{x}_i\|/(2\sigma^2)] \tag{5}$$

$$\hat{\boldsymbol{y}}(\boldsymbol{x})=\sum_{i=1}^{N}\alpha_i K(\boldsymbol{x},\boldsymbol{x}_i)+b \tag{6}$$

式中：σ 为 RBF 的宽度。

1.2.3 粒子群改进最小二乘支持向量机

预测算法的参数选取对预测结果有很大影响，因此选择合适的参数优化算法对预测算法至关重要。LSSVM 中正则化参数 γ 和 σ 的选取对整个模型的预测精度有较大影响。本模型采用 PSO 对 LSSVM 参数进行优化，具体步骤如下。

步骤 1：对 PSO 中的各个参数进行初始化。

步骤 2：对每个粒子的 LSSVM 模型预测，其适应度为预测误差，并与粒子本身的最优适应度比较，若前者更优，则粒子当前的位置为其最优位置。

步骤 3：将粒子的最优位置适应度与群体最优位置适应度进行比较，若前者更优，说明粒子最优位置为群体的最优位置。

步骤 4：更新粒子的速度和位置，计算惯性权值。

1.3 线性序列预测算法

线性序列反映了溶解氧时间序列形成的线性过程，ARIMA 是线性序列预测方法之一，用于对差分平稳时间序列进行拟合。ARIMA 模型有三个参数：数据滞后数(p)、差分化阶数(d)、预测误差滞后数(q)，其数学模型的一般形式如下

$$u_t=a+\varphi_1 u_{t-1}+\ldots+\varphi_p u_{t-p}+\varepsilon_t+\theta_1\varepsilon_{t-1}+\ldots+\theta_q\varepsilon_{t-q} \tag{7}$$

式中：u_t 为差分后的平稳序列；φ 为自回归的系数；θ 为移动平均的系数；a 为常数；ε_t 为零均值白噪声序列。

ARIMA 的具体步骤如下：

步骤 1：对输入的时间序列进行平稳性检测，将非平稳性序列差分转化为平稳性序列。

步骤 2：对平稳序列计算和分析自相关系数和偏自相关系数，得到 p 和 q。

步骤 3：使用 p 和 q 建立模型，并进行模型检验。

步骤 4：将最佳参数代入模型，使用模型对输入数据进行预测。

2 基于时序分解特征的水质溶解氧混合预测模型

2.1 原始数据处理

京杭大运河江南段又称江南运河，常州段是江

南运河的典型段落。江南运河河道水位落差较小,若水质管理不及时,会导致水体水质加速恶化,严重影响沿岸的工农业发展和群众的身体健康。

延陵大桥监测点是京杭大运河常州河段的重要监测点,选取延陵大桥监测点2018年10月30日到2019年10月29日共1 968条监测数据,其中前1 868条数据为训练集,后100条数据为测试集,采样时间间隔为4 h。监测变量包括pH、氨氮、溶解氧、化学需氧量、总磷、电导率、水温、浊度等。图1展示了延陵大桥监测点溶解氧序列分布。

2.1.1　主成分选取

水质监测站获取的数据类型较多,各数据之间存在一定的相关性,采用PCA对变量进行分析确定其主成分,即pH(X_1)、氨氮(X_2)、化学需氧量(X_3)、总磷(X_4)、电导率(X_5)、水温(X_6)、浊度(X_7)。

本文利用SPSS实现数据的降维,对各变量进行方差和主成分贡献率计算,结果如表1所示,由表1可知,前2个主成分的特征值均大于1,但其累计贡献率仅为70.857%,故考虑第三主成分。第三主成分的特征值接近1,其累积贡献率达到82.258%,因此选取前三个主成分。

对前三个主成分的特征值进行读取,组成成分矩阵如表2所示,将成分矩阵与原始数据进行计算可得到降维后的数值,降低了数据维度和计算的复杂度。

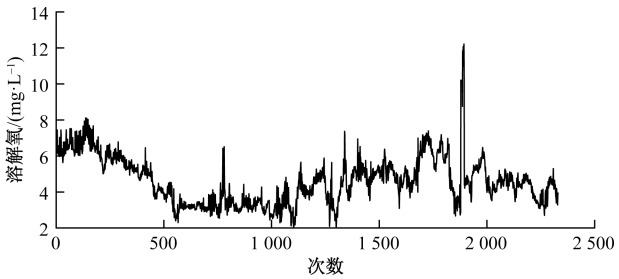


图1　溶解氧序列分布

表1　方差和成分贡献率

成分	初始特征值		
	总计	方差百分比/%	累积/%
1	3.106	44.368	44.368
2	1.854	26.489	70.857
3	0.798	11.401	82.258
4	0.588	8.396	90.654
5	0.298	4.259	94.913
6	0.194	2.771	97.684
7	0.162	2.316	100.000

表2　成分矩阵

变量	1	2	3
PH	0.783	−0.426	0.328
氨氮	−0.268	0.786	0.367
化学需氧量	0.284	0.849	0.114
总磷	0.778	0.359	−0.260
电导率	0.751	−0.220	0.565
水温	0.703	0.385	−0.167
浊度	0.822	−0.100	−0.360

2.1.2　分解溶解氧序列

利用VMD对溶解氧序列进行分解,通过观察VMD的分量瞬时频率可知溶解氧序列的最佳分解个数为两个IMFs。图2为溶解氧序列的分解结果。

由图2可以看出,IMF1具有明显的非线性特性,IMF2具有波动特性,但其波动范围较小,且散射点均匀分布在0附近,说明其非线性不明显。为检验IMFs的特征,对IMFs进行单位根检验。试验结果表明,IMF1的单位根检验$H=0$,拒绝原假设,数据不平稳;IMF2的单位根检验$H=1$,不拒绝原假设,数据平稳,即VMD分解后IMF1为非平稳序列,IMF2平稳序列。因此对IMF1建立基于PCA和PSO优化LSSVM的预测模型,对IMF2建立基于ARIMA的预测模型。

2.2　模型建立和对比

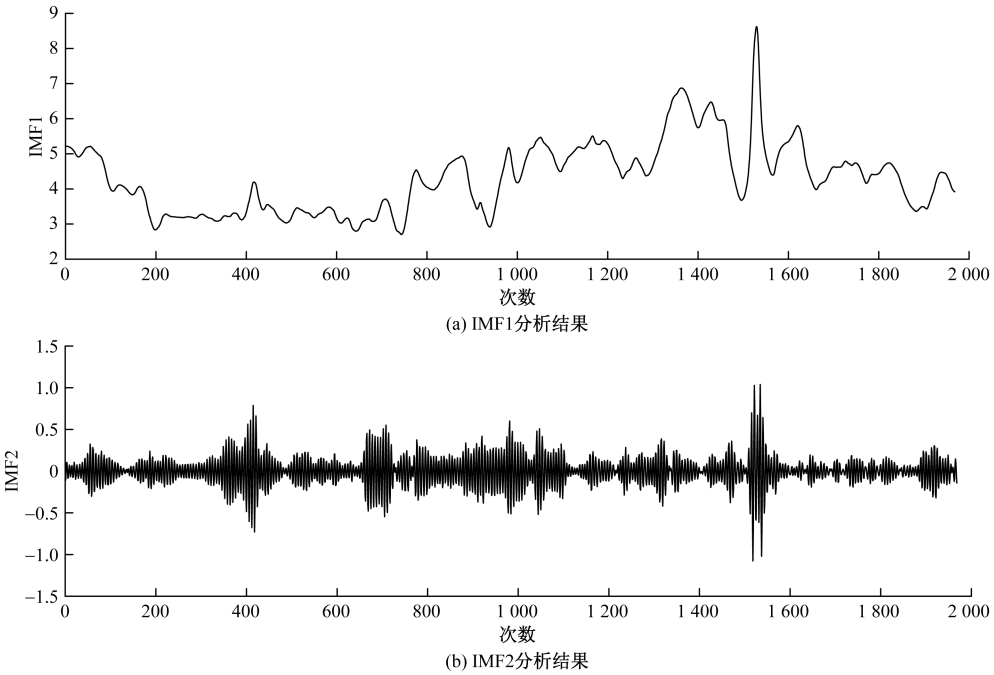
溶解氧预测模型的实验过程主要分为4个部分:原始数据处理、建立混合模型、建立对比模型、模型评价。模型具体步骤如图3所示。

(1)原始数据处理:采用VMD分解原始溶解氧序列,得到多个IMFs;使用PCA对其余水质变量提取主成分,降低数据维度。

(2)建立混合模型:判断IMFs特征,若IMF是非线性序列,则对该IMF建立基于PCA和PSO优化LSSVM的预测模型;若IMF是线性序列,则对其建立基于ARIMA的预测模型;对预测得到的IMFs求和,即为最终溶解氧预测值。

(3)建立对比模型:基于水质变量,创建LSSVM模型;基于其他水质变量和IMFs,建立了VMD-PSO-LSSVM模型;基于数据处理中的主成分和溶解氧,建立了PCA-PSO-LSSVM模型;根据IMFs特征构造不同的ARIMA模型,建立VMD-ARIMA模型。

(4)模型评价:选择平均绝对误差(MAE)、平均绝对百分率误差(MAPE)和均方根误差(RMSE)作为误差评价指标,预测误差范围、平均值作为稳定性评价指标。MAE、RMSE、MAPE越小,预测精度越高。



(a) IMF1分析结果

(b) IMF2分析结果

图 2　溶解氧序列 VMD 分解结果

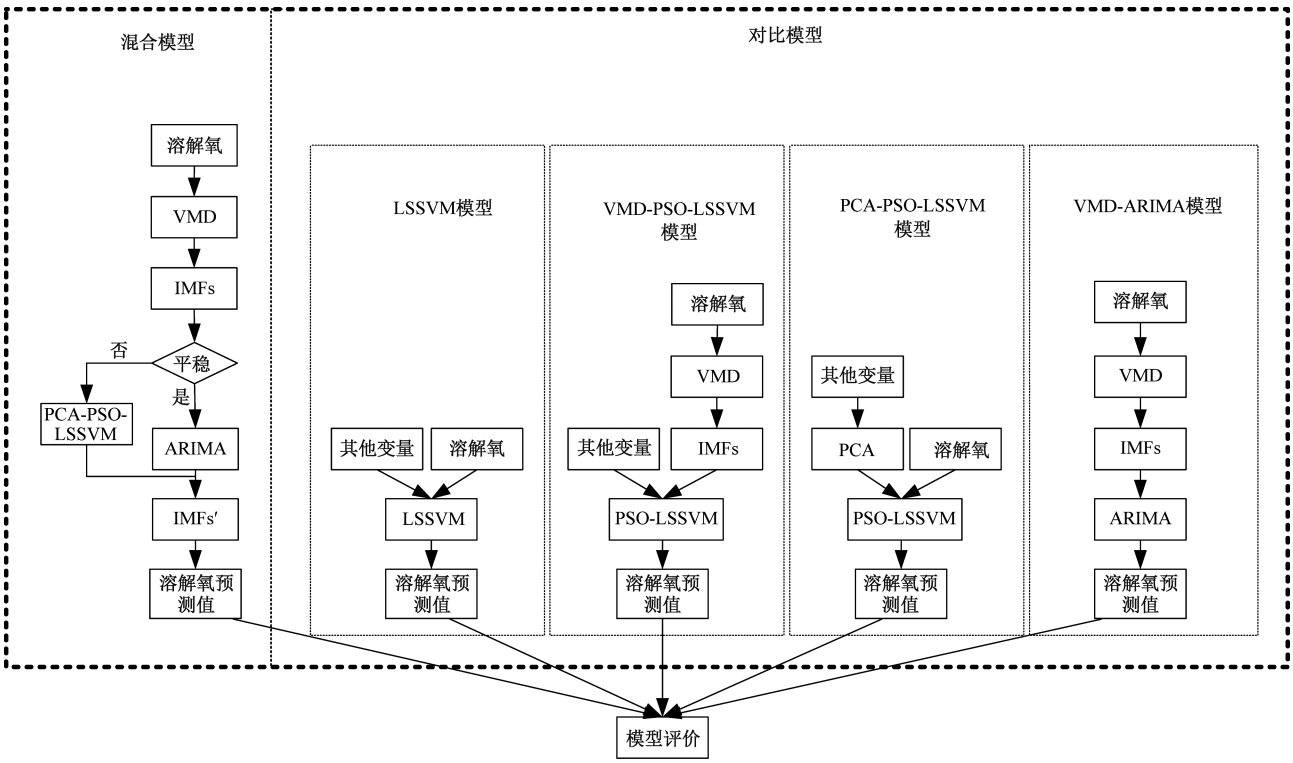


图 3　混合模型的建立和对比

2.3　结果分析

将原始数据处理后，对其进行混合预测及误差分析，并将混合模型预测结果与其他 4 种模型对比，结果如图 4 所示。图 4(a)为未对数据进行分解的模型预测曲线，图 4(b)为对数据进行分解的模型预测曲线。比较图 4 中的预测模型发现，VMD-PSO-LSSVM、PCA-PSO-LSSVM、VMD-ARIMA 及本文建立的混合模型预测曲线拟合度高于单一的 LSSVM 预测曲线。对比 VMD-PSO-LSSVM 算法与 PCA-PSO-LSSVM 预测算法，VMD-PSO-LSSVM

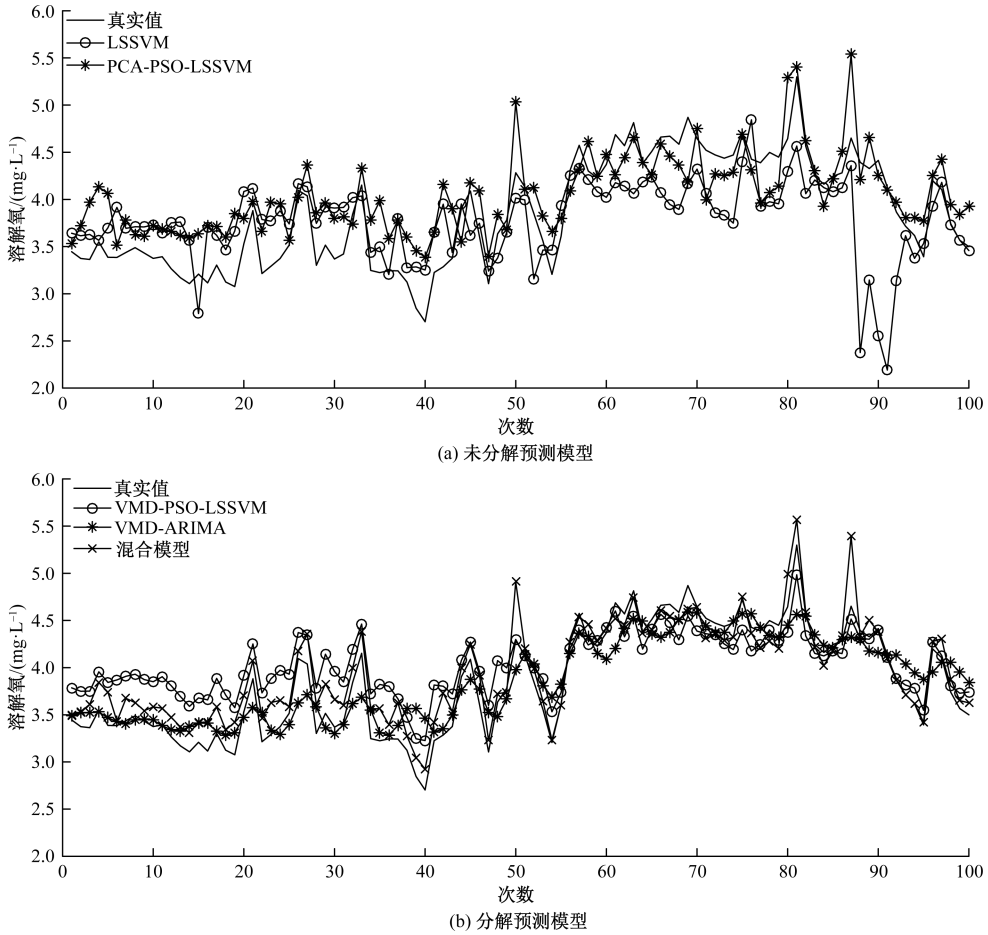


图 4　溶解氧预测结果

的预测曲线拟合程度更好,说明 VMD 处理过的算法预测精度提高。利用 VMD 分解后根据各 IMF 的特征建立的 ARIMA 模型比其余 3 种对比模型性能更优,说明预测模型中根据数据自身特征建立模型的重要性。

表 3 为预测模型的评价,通过对 5 种模型的 RMSE、MAE、MAPE 比较,发现 LSSVM、PSO-PSO-LSSVM、VMD-PSO-LSSVM、VMD-ARIMA、混合模型误差指标依次呈降趋势,即预测性能呈上升趋势。与其他 4 种模型的 RMSE、MAE、MAPE 相比,本文建立的混合模型 3 个指标均显著降低,预测精度较高。

为确定模型预测是否稳定,将混合模型与其他 4 种对比模型预测误差进行对比,表 4 为 5 种预测模型的误差。

由表 4 可知,混合模型的误差范围为 0.001～0.742,误差均值为 0.112,与 LSSVM 模型、VMD-PSO-LSSVM 模型、PCA-PSO-LSSVM 模型、VMD-ARIMA 模型相比,混合模型波动范围小,误差均值小,表明混合模型具有较小的误差分散度,模型较稳定。

表 3　五种模型预测评价

变量	LSSVM	VMD-PSO-LSSVM	PCA-PSO-LSSVM	VMD-ARIMA	混合模型
RMSE	0.528	0.346	0.387	0.249	0.211
MAE	0.395	0.295	0.322	0.191	0.168
MAPE	10.227	8.338	8.927	5.145	4.576

表 4　5 种模型预测误差

误差	LSSVM	VMD-PSO-LSSVM	PCA-PSO-LSSVM	VMD-ARIMA	混合模型
最小误差	0.001	−0.012	0.011	0.004	−0.001
最大误差	−2.016	0.627	0.886	0.759	0.742
平均误差	−0.176	0.185	0.196	0.118	0.112

表 5 为 5 种模型在不同时间段中预测误差评价,其中 1D 表示预测时间为 1 d,5D 表示预测时间为 5 d,10D 表示预测时间为 10 d,15D 表示预测时间为 15 d。从表中可以看出,VMD-ARIMA 模型和混合模型性能明显优于其他 3 种模型,其中预测时长为 1 d 和预测时长为 5 d 时,VMD-ARIMA 模

表5　五种模型不同时长预测误差

变量	1D(1 d)		5D(5 d)		10D(10 d)		15D(15 d)	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
LSSVM	0.301	0.260	0.396	0.362	0.362	0.314	0.510	0.400
VMD-PSO-LSSVM	0.394	0.392	0.481	0.473	0.422	0.388	0.362	0.316
PCA-PSO-LSSVM	0.508	0.459	0.436	0.389	0.431	0.375	0.399	0.334
VMD-ARIMA	0.105	0.090	0.185	0.144	0.248	0.187	0.242	0.182
混合模型	0.244	0.220	0.237	0.220	0.230	0.197	0.220	0.179

型略优于本文提出的混合模型，但是VMD-ARIMA模型的预测性能波动较大，随着时间的延长，预测性能越来越差，反观本文提出的混合模型，总体预测性能波动不大，其预测较为准确。表明混合模型随着预测时长的增加，预测性能波动不大，模型预测稳定。

3　总结

本文在“分解-预测-重构”模型基础上提出的基于VMD分解特征的混合溶解氧预测方法，利用VMD来分解溶解氧时序，降低溶解氧时序的不稳定性，与未进行分解的预测模型相比，提出的模型具有较小的误差分散度，模型较稳定。针对VMD分解得到的IMFs结合各自特征使用不同预测算法进行预测，其中评价指标RMSE为0.211，MAE为0.168，MAPE为4.576，性能优于未结合数据特征进行预测的混合模型，进一步验证了模型的可靠性和稳定性。因此本文所建立的混合模型在水质溶解氧预测方面能有效跟踪水体溶解氧的变化规律，对水质环境的改善具有重要的现实意义。

参考文献

[1] DRAGOMIRETSKIY K, ZOSSO D. Variational mode decomposition[J]. IEEE Trans Signal Process, 2014, 62(3): 531-544.

[2] 姜建国，杨效岩，毕洪波. 基于VMD-FE-CNN-BiLSTM的短期光伏发电功率预测[J]. 太阳能学报，2024，45(7)：462-473.

[3] 罗潇远，刘杰，杨斌，等. 基于改进鱼鹰优化算法与VMD-LSTM的超短期风电功率预测[J/OL]. 太阳能学报：1-9[2024-08-06]. https://doi.org/10.19912/j.0254-0096.tynxb.2023-1942.

[4] 聂淑媛. 基于四类时间序列模型的GDP预测效果比较[J]. 统计与决策，2024，40(2)：63-67.

[5] 鲁玉芬，方从严，开明. 灰色-时序组合模型在建筑物沉降预测中的应用[J]. 科技和产业，2022，22(8)：315-318.

[6] 王秋璐，许艳，黄海燕，等. 基于时空矩阵方法对福建省海湾水质变化特征分析[J]. 海洋学报，2019，41(2)：134-144.

[7] 谢小良，吴琳琳. 基于DA-CNN-BiLSTM的河流溶解氧浓度预测[J]. 人民黄河，2024，46(7)：92-97.

[8] 李晓晨，白音包力皋，李向东，等. 基于IPSO-BP神经网络的高含沙水体对鱼类影响预测方法[J]. 水利学报，2023，54(3)：291-301.

[9] MASOUD R, TAHER R, MUSTAFA E. Prediction of dissolved oxygen in River Calder by noise elimination time series using wavelet transform[J]. Journal of Experimental & Theoretical Artificial Intelligence, 2016, 28(4): 689-706.

[10] SALAR K, TAHER R. Modeling of dissolved oxygen concentration and its hysteresis behavior in rivers using wavelet transform: based hybrid models[J]. CLEAN-Soil, Air, Water, 2017, 45(2): 1-18.

[11] 刘丙杰，侯慕馨，冀海燕. 基于支持向量机集成的船舶舱室温湿度预测[J]. 海军工程大学学报，2024，36(3)：21-25.

[12] 周强，周杰，赵文洋，等. 基于支持向量机的砌体结构震害预测新方法研究[J]. 地震工程与工程振动，2023，43(5)：130-137.

[13] 隗海林，包翠竹，李洪雪，等. 基于最小二乘支持向量机的怠速时间预测[J]. 吉林大学学报(工学版)，2018，48(5)：1360-1365.

[14] ZHANG G, LIU H C, ZHANG J B, et al. Wind power prediction based on variational mode decomposition multi-frequency combinations[J]. Journal of Modern Power Systems and Clean Energy, 2019, 7(2): 281-288.

[15] 吴静，李振波，朱玲，等. 融合ARIMA模型和GAWNN的溶解氧含量预测方法[J]. 农业机械学报，2017，48(S1)：205-210.

[16] 刘畅，李蕴峰. 海南省旅游统计数据的ARIMA模型预测分析和新冠肺炎疫情影响分析[J]. 科技和产业，2023，23(8)：22-32.

[17] 薛东，段立强，高统彤，等. 考虑多重特征与不确定性度量的综合能源系统负荷预测研究[J]. 太阳能学报，2024，45(7)：379-388.

[18] 宦娟，曹伟建，秦益霖，等. 基于游程检测法重构集合经验模态的养殖水质溶解氧预测[J]. 农业工程学报，2018，34(8)：220-226.

[19] 陈英义，程倩倩，方晓敏，等. 主成分分析和长短时记忆神经网络预测水产养殖水体溶解氧[J]. 农业工程学报，2018，34(17)：183-191.

Prediction of Dissolved Oxygen in Water Quality Based on Time Series Decomposition Characteristics

LI Hui

(School of Foreign Studies, Changzhou University, Changzhou 213164, Jiangsu, China)

Abstract: In order to improve the stability of the water ecosystem, a mixed prediction model of dissolved oxygen based on variational mode decomposition (VMD) was proposed. Dissolved oxygen was decomposed by VMD to obtain multiple IMFs, and different models were established according to the characteristics of different IMFs. Principal component analysis (PCA) and particle swarm optimization (PSO) to optimize the least squares vector machine model (LSSVM) was uesed in the nonlinear sequence part autoregressive integrated moving average model (ARIMA) was used in the linear sequence part. The prediction results of dissolved oxygen were obtained by combining the prediction results of each part. The model in this paper was applied to a monitoring point in the Changzhou section of the Bei-jing-Hangzhou Grand Canal. The results show that the average absolute error is 0.168, the root mean square error is 0.211, and the average absolute percentage error is 4.576. The mixed model can predict dissolved oxygen well and has certain practical significance for the management of water quality environment.

Keywords: water quality dissolved oxygen prediction; time series decomposition; variational mode decomposition; least square support vector machine; autoregressive integrated moving average model