doi:10.11887/j.cn.202406016 http://journal.nudt.edu.cn

考虑噪声影响的 MEMD-XGBoost 方法在 GNSS 高程时间序列建模和预测中的应用

鲁铁定[1,2],李 祯[1*],贺小星[3]

(1. 东华理工大学 测绘与空间信息工程学院,江西 南昌 330013;
2. 自然资源部 环鄱阳湖区域矿山环境监测与治理重点实验室,江西 南昌 330013;
3. 江西理工大学 土木与测绘工程学院,江西 赣州 341000)

摘 要:全球导航卫星系统(global navigation satellite system,GNSS)高程时间序列研究有助于监测和分析地壳板块运动,可以为研究人员判断区域运动趋势提供依据。基于经验模态分解和极端梯度提升算法构建了 MEMD-XGBoost 模型来预测分析 GNSS 高程时间序列。为了验证模型的预测性能,实验选取 8 个 GNSS 站高程时间序列数据进行预测实验,特征构造结果显示,多次经验模态分解可以准确地提取原始时间序列信息,提供有效特征。建模结果表明,MEMD-XGBoost 模型可以有效改善数据质量。预测结果表明,MEMD-XGBoost 模型预测结果具有较高的精度和准确率,误差离散程度较小,模型具有较强的稳定性和鲁棒性,可以较好地预测出 GNSS 站高程方向的运动趋势和季节性变化。因此,该模型可以应用于 GNSS 高程时间序列建模和预测研究。

关键词:GNSS 时间序列;经验模态分解;极端梯度提升;建模;预测

中图分类号:P228 **文献标志码**:A **文章编号**:1001-2486(2024)06-149-10

论文拓展

Noise-aware MEMD-XGBoost method for GNSS vertical time series modeling and prediction

LU Tieding[1,2], *LI Zhen*[1*], *HE Xiaoxing*[3]

(1. School of Surveying and Geoinformation Engineering, East China University of Technology, Nanchang 330013, China;
2. Key Laboratory of Mine Environmental Monitoring and Improving around Poyang Lake, Ministry of Natural Resources, Nanchang 330013, China;
3. School of Civil Engineering and Surveying & Mapping Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, China)

Abstract: The study of GNSS(global navigation satellite system) vertical time series is helpful for monitoring and analyzing the movement of crustal plates, and can provide an important basis for judging the movement trend. A MEMD-XGBoost model was constructed based on empirical mode decomposition and extreme gradient boosting algorithm for GNSS vertical time series prediction and analysis. In order to verify the prediction performance of the model, the vertical time series data of 8 GNSS stations were selected for prediction experiments. The feature construction results show that multiple empirical mode decomposition can accurately extract the original time series information and provide effective features. The modeling results show that the MEMD-XGBoost model can effectively improve the data quality. The prediction results show that the prediction results of the MEMD-XGBoost model have high precision and accuracy, and the degree of error dispersion is small, the model has strong stability and robustness, and can better predict the movement trend and seasonal changes in the U direction of the GNSS station. Therefore, the model can be applied to GNSS vertical time series modeling and prediction research.

Keywords: GNSS time series; empirical mode decomposition; extreme gradient boosting; modeling; prediction

收稿日期:2022-06-30

基金项目:国家自然科学基金资助项目(42374040, 42061077, 42064001, 42104023);江西省自然科学基金资助项目(20202BABL213033;20202BAB212010);江西理工大学高层次人才科研启动资助项目(205200100564)

第一作者:鲁铁定(1974—),男,陕西富平人,教授,博士,博士生导师,E-mail:tdlu@whu.edu.cn

***通信作者**:李祯(1998—),男,河南新乡人,硕士研究生,E-mail:lizhenhd@163.com

引用格式:鲁铁定, 李祯, 贺小星. 考虑噪声影响的 MEMD-XGBoost 方法在 GNSS 高程时间序列建模和预测中的应用[J]. 国防科技大学学报, 2024, 46(6): 149-158.

Citation: LU T D, LI Z, HE X X. Noise-aware MEMD-XGBoost method for GNSS vertical time series modeling and prediction[J]. Joural of National University of Defense Technology, 2024, 46(6): 149-158.

过去30年来,来自全球20 000多个全球导航卫星系统(global navigation satellite system, GNSS)站的观测资料不断积累,为地球科学领域各主题的研究提供了庞大的数据库[1-4]。这些数据可以有效地描述地球物理效应引起的长期趋势和非线性变化[5]。分析GNSS坐标时间序列有助于开展监测地壳板块运动[6-7]、水坝或桥梁变形[8-10]、全球或区域参考框架[11-13]等研究。通过分析GNSS坐标时间序列,可以预测连续时间点的坐标,为确定运动趋势提供重要依据[14]。

在现有的时间序列预测研究中,一种结合信号分解和预测算法的预测模式在诸多研究中得到了良好的应用[15-17],该预测模式首先通过信号分解算法对GNSS时间序列进行分解,然后对各个分量进行逐一预测,最后将各个分量的预测值等权相加得到预测时间序列。该预测模式虽然可以良好应用于时间序列预测研究,但GNSS高程时间序列属于非平稳时间序列,且噪声模型丰富[18],使用上述预测模式进行预测容易随着分量数量级降低,预测精度和准确率下降。因此,本文提出了一种基于极端梯度提升算法(extreme gradient boosting,XGBoost),通过多次经验模态分解(multi-empirical mode decomposition,MEMD)改进的GNSS高程时间序列预测模型——MEMD-XGBoost模型。

在GNSS相关研究中,经验模态分解(empirical mode decomposition,EMD)算法被研究人员广泛应用于GNSS站速度估计[19]、GNSS时间序列异常值探测[20]、高频GNSS信号去噪[21]等研究中。XGBoost算法在GNSS领域的研究主要集中在电离层闪烁[22-23]、电离层电子含量[24]、土壤水分反演[25]等研究。本文通过EMD和XGBoost算法构建了MEMD-XGBoost模型,该模型通过EMD算法分解得到的重构时间序列,依次得到若干个重构时间序列,并将其作为特征取代使用XGBoost模型预测时提取时间特征作为特征的步骤,辅助原始时间序列的预测工作,从而达到优化先分解再预测的预测模式、削弱噪声影响和弥补XGBoost算法提取非平稳时间序列特征能力欠佳的目的。

1 MEMD-XGBoost模型原理

1.1 EMD算法

EMD是由Huang等[26]在1998年提出的一种适用于非线性和非平稳过程的自适应信号处理技术,EMD算法被研究人员广泛应用于生物医学[27]、语音识别[28]、系统建模[29]、钢铁工业[30]等领域研究。基于时间尺度、EMD局部特征(局部最大值、局部最小值和过零),EMD将信号分解为多个本征模态函数(intrinsic mode function, IMF)和一个残差,这些IMF相互正交,且模态分解数量由信号本身决定。EMD算法的分解步骤如下:

1)寻找原始时间序列$X(t)$的极值点,然后通过曲线插值法拟合极值点,得到原始时间序列的上包络线$X_{\max}(t)$和下包络线$X_{\min}(t)$。

2)求$X_{\max}(t)$和$X_{\min}(t)$的平均值$m_1(t)$,即:

$$m_1(t)=\frac{X_{\max}(t)+X_{\min}(t)}{2} \tag{1}$$

3)通过$X(t)$和$m_1(t)$相减得到余下信号$d_1(t)$。此时,由于GNSS高程坐标时间序列属于非平稳时间序列,得到的第一阶模态函数并不准确,即$d_1(t)$并不满足IMF的两个条件(在整个数据范围内,局部极值点和过零点的数目必须相等,或者相差数目最多为1,并且在任意时刻,上包络线和下包络线的平均值必须为零),所以需要继续筛选。

4)对$d_1(t)$重复步骤1~3,直至筛分门限值S_D小于门限值,从而得到第一阶模态分量$c_1(t)$,即IMF1,S_D计算公式可表示为:

$$S_D=\sum_{t=0}^{T}\frac{|d_{k-1}(t)-d_k(t)|^2}{d_{k-1}^2(t)} \tag{2}$$

5)对$X(t)$和$c_1(t)$求差得到第一阶残差量$r_1(t)$,然后对$r_1(t)$重复步骤1~5,重复n次后得到第n个IMF分量$c_n(t)$和残差量$r_n(t)$。原始时间序列$X(t)$经EMD算法分解为:

$$X(t)=\sum_{1}^{n}c_n(t)+r_n(t) \tag{3}$$

1.2 XGBoost算法

Chen等基于梯度提升决策树(gradient boosting decision tree, GBDT)模型改进并提出了XGBoost模型[31]。传统的GBDT模型只使用了一阶泰勒展开,比较复杂,容易出现过拟合,即模型在训练集和测试集的预测精度差异较大。XGBoost模型采用二阶泰勒展开,并增加了正则化项,使模型简化,减少了过拟合的发生。XGBoost模型预测原理[32]如下:

假设有一个数据集$D=\{(x_i,y_i)\}(|D|=n, x_i\in\mathbf{R}^m, y_i\in\mathbf{R})$,数据集$D$中含有$n$个观测值,每个观测值有$m$个特征。将$x_i$在第$t$轮的预测值定义为$\hat{y}_i^{(t)}$,$x_i$的最终预测值可以表示为:

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{4}$$

式中,$\hat{y}_i^{(t-1)}$ 表示第 t 轮前的预测值,$f_t(x_i)$ 表示在第 t 轮新加入的函数。为了防止节点过多导致过拟合,XGBoost 算法引入惩罚项降低过拟合的风险,惩罚函数 $\Omega(f_t)$ 可以表示为:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{5}$$

式中,γT 表示惩罚,$\frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2$ 表示惩罚项,λ 为系数,T 为叶节点数,j 为样本数量,ω_j 为权重。由损失函数 L 和正则化惩罚项 Ω 组成的目标函数 $O^{(t)}$ 可以表示为:

$$O^{(t)} = \sum_{i=1}^{n} L(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \Omega(f_t) + c \tag{6}$$

式中,c 为一个常数项。

XGBoost 算法采用二阶泰勒展开对目标函数进行优化,展开公式可以表示为:

$$\begin{cases} O^{(t)} \simeq \sum_{i=1}^{n} \left[L(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \\ \qquad \Omega(f_t) + c \\ g_i = \partial_{\hat{y}^{(t-1)}} L(y_i, \hat{y}_i^{(t-1)}) \\ h_i = \partial^2_{\hat{y}^{(t-1)}} L(y_i, \hat{y}_i^{(t-1)}) \end{cases} \tag{7}$$

然后去掉常数项(真实值与前一轮预测值之差),目标函数仅依赖于误差函数对每个数据点的一阶导数和二阶导数。目标函数最终被简化为:

$$OBJ^{(t)} \simeq \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{8}$$

1.3　构建 MEMD-XGBoost 模型

通过 EMD 算法分解时间序列后容易出现模态混叠现象,即不同时间尺度特征成分被分解到一个特征模态函数分量,或者同一时间尺度成分出现在不同的特征模态函数中。为了削弱这一现象对于预测的影响以及提高重要特征的权重,实验通过将每次得到的 IMF 分量叠加得到重构时间序列并作为特征,多次使用 EMD 算法分解上一步得到的有用信号,重构得到的时间序列会逐渐丢失次要特征,在剔除噪声的同时提高时间序列主要特征的权重。需要注意的是,为了保证预测模型的严谨性,应当仅使用训练集数据构造特征。

GNSS 高程时间序列数据通常为一维时间序列数据,其具有统一的时间间隔[5]。将 GNSS 高程数据按照时间顺序一维排列为:

$$X_1, X_2, X_3, \cdots, X_{n-1}, X_n \tag{9}$$

将 GNSS 高程时间序列数据定义为 D,其多次分解过程可以被表示为:

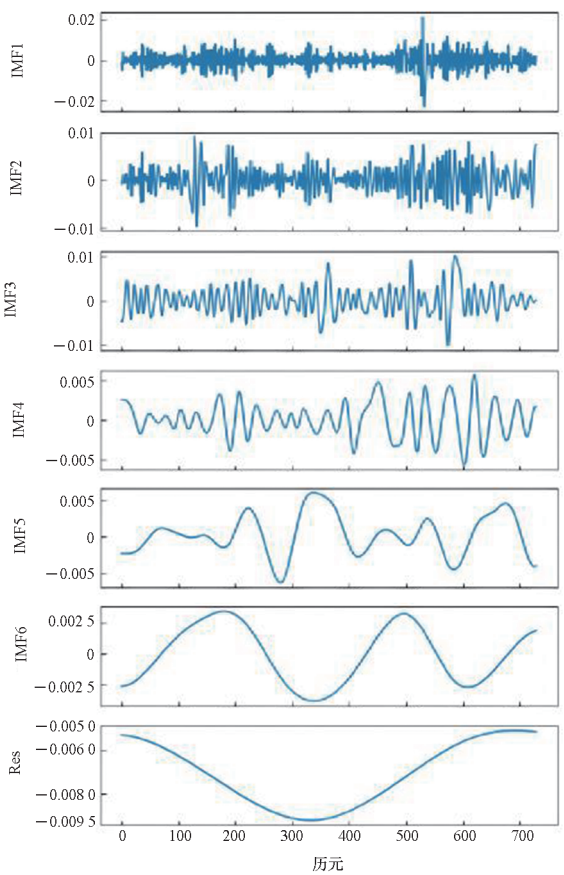
$$\begin{cases} D \rightarrow D_1 + r_1 \\ D_1 \rightarrow D_2 + r_2 \\ \vdots \\ D_{n-1} \rightarrow D_n + r_n \end{cases} \tag{10}$$

式中,D_n 和 r_n 分别表示第 n 次分解得到的有用信号和残差项。以 BJYQ 站为例,图 1 为 BJYQ 站训练集数据 3 次 EMD 结果。

从图 1 可以看出,经历过 3 次分解后得到的残差项数量级已经很小,继续分解意义不大,因此,实验将 EMD 次数设置为 3。然后,实验通过前文假设和 3 次 EMD 分解可以将特征 1(F1)、特征 2(F2)和特征 3(F3)分别定义为 D_1、D_2 和 D_3。

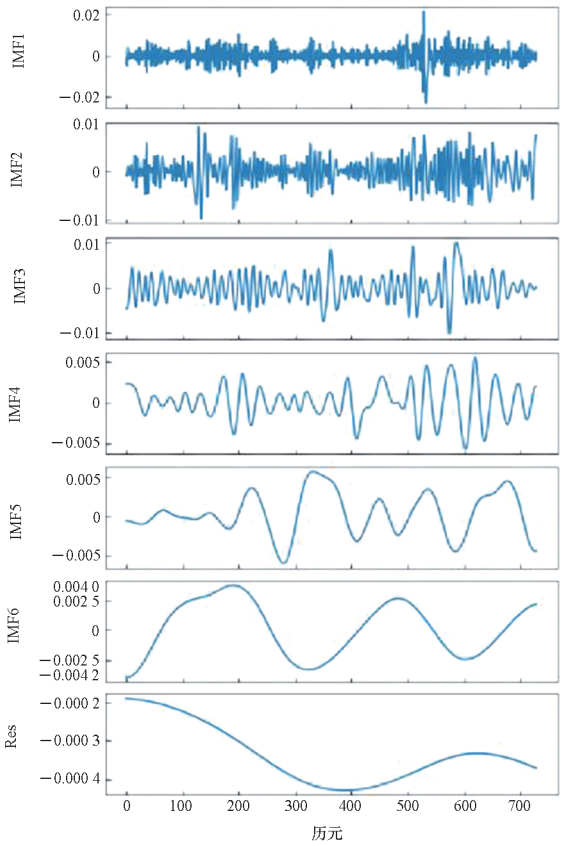
通过上述步骤后,MEMD 算法可以为 XGBoost 模型提供 3 个特征,并构成特征集,特征集可整合为:

$$\begin{bmatrix} X_{1F1} & X_{1F2} & X_{1F3} \\ X_{2F1} & X_{2F2} & X_{2F3} \\ \vdots & \vdots & \vdots \\ X_{nF1} & X_{nF2} & X_{nF3} \end{bmatrix} \tag{11}$$



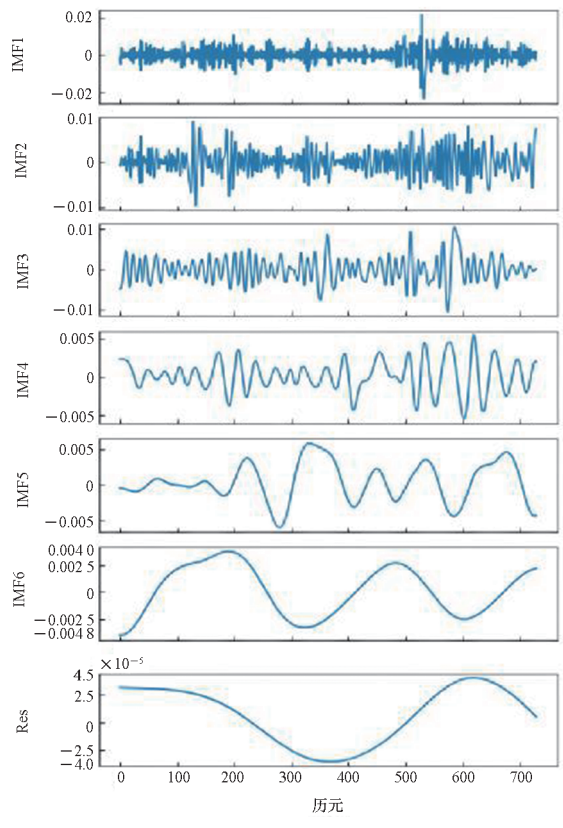
(a) 第 1 次分解结果

(a) The 1st decomposition result



（b）第 2 次分解结果

（b）The 2nd decomposition result



（c）第 3 次分解结果

（c）The 3rd decomposition result

图 1　BJYQ 站 3 次 EMD 结果

Fig. 1　Three EMD results of BJYQ station

式中，X_{nF1} 表示第 n 个观测值对应的第 1 个特征。将整合的 3 维时间序列加入原始时间序列生成一个 4 维时间序列：

$$\begin{bmatrix} X_{1F1} & X_{1F2} & X_{1F3} & X_1 \\ X_{2F1} & X_{2F2} & X_{2F3} & X_2 \\ \vdots & \vdots & \vdots & \vdots \\ X_{nF1} & X_{nF2} & X_{nF3} & X_n \end{bmatrix} \tag{12}$$

在 XGBoost 模型中，将生成的 4 维时间序列中的前 3 列时间序列数据作为特征取代 XGBoost 模型提取时间特征的步骤；将原始时间序列数据作为目标序列进行预测，从而得到预测结果，图 2 为 MEMD-XGBoost 模型预测流程图。

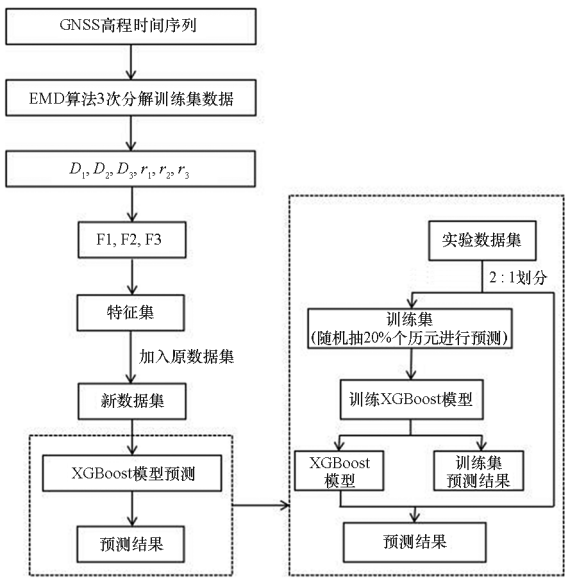


图 2　MEMD-XGBoost 模型预测流程图

Fig. 2　Prediction flow chart of MEMD-XGBoost model

基于多次 EMD 算法分解改进的 GNSS 高程时间序列预测模型——MEMD-XGBoost 模型具体预测步骤如下：

1）数据准备。GNSS 高程时间序列是通过实际观测或求解得到的，它应在周、天、小时、秒等维度具有一致性。本文选取 8 个 GNSS 观测站单日解高程时间序列数据作为实验数据。

2）依次使用 3 次 EMD 算法重构训练集时间序列。首先通过 EMD 模型进行时间序列分解，然后将分解得到的各 IMF 分量叠加得到重构时间序列，最后将得到的重构时间序列再次进行 EMD，共进行 3 次。实验通过该步骤重构得到 3 个时间序列，并将其作为特征辅助原始时间序列的建模和预测。

3）构造数据集。将子时间序列和原始时间序列放入同一数据集，将子时间序列作为原始时间序列的特征，并通过 2∶1 的比例划分训练集和

测试集。

4)XGBoost 模型预测。首先将实验数据集中的训练集输入 XGBoost 模型,在训练集中随机抽取 20% 个历元进行预测并通过五折交叉验证得到最终预测模型、输入特征的特征评价结果以及训练集预测结果。然后将实验数据集输入 XGBoost 模型进行测试集目标时间序列的预测。

5)统计 MEMD-XGBoost 模型预测结果。通过设置不同的预测精度指标评判模型预测的有效性、稳定性、误差主要方向和离散程度。

2　数据分析

本文数据均来自中国地震局 GNSS 数据产品服务平台,实验选取中国大陆构造环境监测网络的 8 个 GNSS 站 2013—2015 年的数据验证 MEMD-XGBoost 模型的有效性。图 3 为 8 个 GNSS 站的位置分布图。

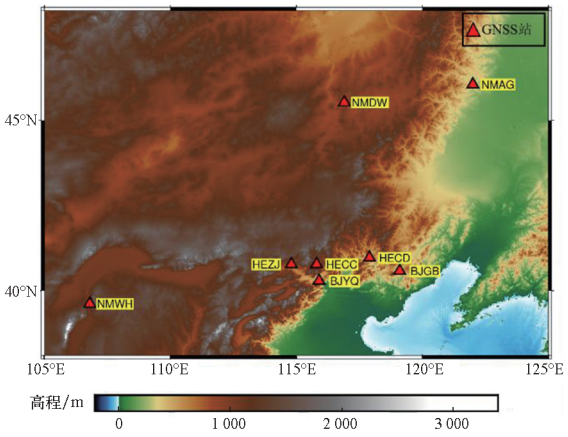


图 3　GNSS 站位置分布

Fig. 3　Location distribution of GNSS stations

实验在进行观测站筛选时,选取的 GNSS 站点数据完整率较高,因此实验选取的 8 个观测站数据虽均含有缺失数据,但缺失数据较少。基于上述情况,实验通过缺失值临近历元的观测值进行插值处理,插值方法可表示为:

$$X_m = \frac{X_{mp1} + X_{mn1}}{2} \tag{13}$$

式中,X_m 表示第 m 个历元的缺失值,X_{mp1} 表示第 m 个历元前含有真值的最临近历元的观测值,X_{mn1} 表示第 m 个历元后含有真值的最临近历元的观测值。该插值方法适用于缺失值较少的情况,其可以有效保留原始时间序列的超短期变化趋势。

经过插值处理后,每个 GNSS 站含有 1 095 个历元的数据,实验按照 2∶1 的比例划分训练集和测试集,即训练集包含各 GNSS 站 730 个历元数据,测试集包含各 GNSS 站 365 个历元数据。

为了分析模型的抗异常值干扰能力,实验通过四分位数法对各 GNSS 站数据进行异常值探测[33],图 4 为各 GNSS 站异常值探测结果,图中 U 表示 GNSS 高程时间序列的值。

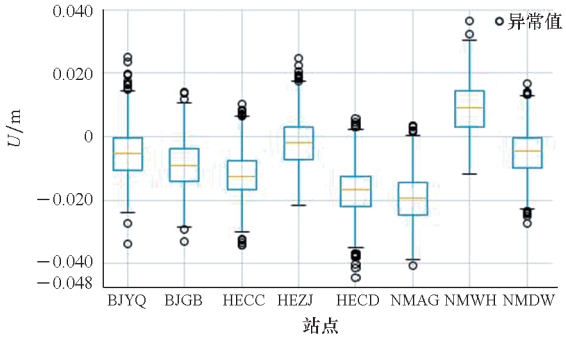


图 4　GNSS 站异常值探测结果

Fig. 4　Outlier detection results of GNSS stations

图 4 中,黑色圆圈表示异常值,箱状图表示数据的分布情况。从图 4 可以看出,各 GNSS 站均含有异常值。

3　实验结果与分析

3.1　精度评价指标

本文使用平均绝对误差(mean absolute error, MAE)、均方根误差(root mean square error, RMSE)和对称平均绝对百分比误差(symmetric mean absolute percentage error, SMAPE)作为模型预测精度的评价指标[34-35]。MAE、RMSE 和 SMAPE 可表达为:

$$E_{\mathrm{MAE}} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{14}$$

$$E_{\mathrm{RMSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{15}$$

$$E_{\mathrm{SMAPE}} = \frac{100\%}{n}\sum_{i=1}^{n} \frac{2|\hat{y}_i - y_i|}{|\hat{y}_i| + |y_i|} \tag{16}$$

其中,y_i 为原始值,$\hat{y}_i$ 为预测值。MAE、RMSE 和 SMAPE 的值越小代表模型的预测精度越高,更适用于该时间序列;反之,则代表模型的预测精度越低,在该时间序列中适用性较差。SMAPE 指标与平均绝对百分比误差(mean absolute percentage error, MAPE)作用相同,可以衡量模型预测的准确率,但 SMAPE 可以有效避免 MAPE 值因真实值小而计算结果太大的问题。

为了评价模型预测的稳定性,实验通过增量误差(delta error, DE)衡量预测误差的离散程度。

$$\begin{cases} D_{\mathrm{E}} = \dfrac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) \pm \sigma_\delta \\ \delta = \hat{Y} - Y \end{cases} \tag{17}$$

其中，D_{E} 表示增量误差，σ_δ 表示预测误差的标准差，δ 表示预测误差时间序列，$\hat{Y}$ 表示模型预测得到的时间序列，Y 表示原始时间序列。

3.2　特征构造结果

实验以 EMD 算法为基础对原始时间序列进行重构，并将重构时间序列作为特征辅助原始时间序列的建模。图 5 为 BJYQ 站特征构造结果。

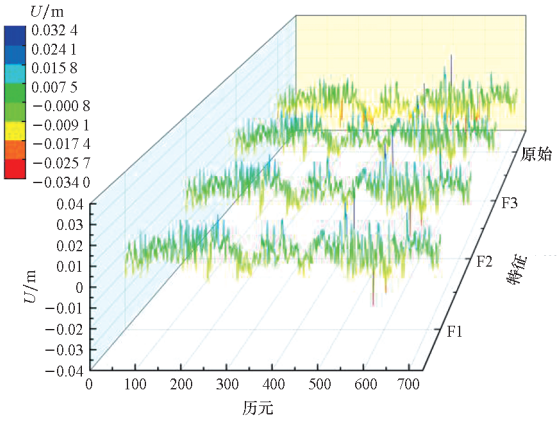


图 5　BJYQ 站特征构造结果

Fig. 5　Characteristic construction results of BJYQ station

图 5 所示均为高程值映射曲线，因此曲线由不同的颜色构成。从图 5 可以看出，通过 EMD 算法重构得到的特征时间序列的运动趋势和原始时间序列较为一致，但在初始历元阶段，特征时间序列的高程映射颜色与原始时间序列呈现了差异，这是由 EMD 算法的端点效应造成的。

3.3　MEMD-XGBoost 模型预测结果

实验使用 MEMD-XGBoost 模型对 8 个 GNSS 站高程时间序列数据进行建模，以 HECD 站为例，以 XGBoost1（以时间为特征）、XGBoost2（以邻近观测站数据为特征）[14]、长短期记忆（long short-term memory，LSTM）网络、CNN-LSTM[36-38] 模型作为参照模型验证 MEMD-XGBoost 模型的有效性，表 1 为 HECD 站模型建模精度。

表 1　HECD 站模型建模精度

Tab. 1　Model accuracy for HECD station

模型	MAE/mm	RMSE/mm
XGBoost1	6.56	8.40
XGBoost2	3.38	4.31
LSTM	5.50	7.18
CNN-LSTM	4.89	6.58
MEMD-XGBoost	1.48	2.05

由表 1 可以看出，MEMD-XGBoost 模型的 MAE 和 RMSE 值较小，验证了 MEMD-XGBoost 模型的有效性。图 6 为 HECD 站测试集建模结果。

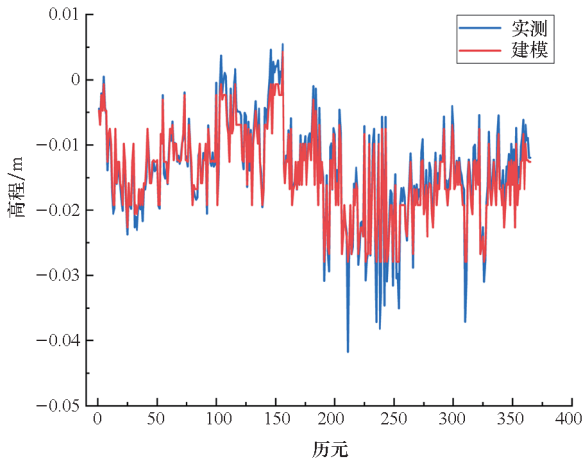


图 6　HECD 站测试集建模结果

Fig. 6　Modeling result of HECD station test set

图 6 中，蓝色曲线为 HECD 站原始时间序列数据，红曲线为 MEMD-XGBoost 模型建模结果。从图 6 可以看出，MEMD-XGBoost 模型可以较好地保留 HECD 站高程方向的运动趋势和季节性变化，但在第 200 ~ 250 历元间建模精度出现了下滑。通过图 4 中 HECD 站的异常值探测情况可以得到建模精度出现下滑的原因，HECD 站的异常值主要集中在极小值点，与 HECD 站的预测误差主要来源相同，从而印证了 MEMD-XGBoost 模型具有一定的抗异常值干扰能力，通过 MEMD-XGBoost 模型进行建模，也可以有效削弱异常值对于原始时间序列的影响。图 7 为 HECD 站测试集原始时间序列和建模时间序列的异常值探测结果。

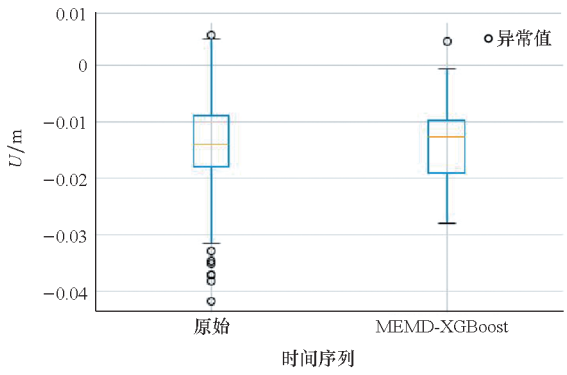


图 7　HECD 站测试集异常值探测结果

Fig. 7　Outlier detection results on the test set of HECD station

通过图 7 可以看出，建模前后高程值的主要分布范围较为一致，异常值个数明显减少，因此

HECD 站的原始时间序列质量可以通过 MEMD-XGBoost 模型建模得到进一步的提升。

为了更好地验证 MEMD-XGBoost 模型的稳定性,实验引入多个精度评价指标相互印证 MEMD-XGBoost 模型在 8 个 GNSS 站的建模精度,表 2 为各 GNSS 站建模精度。

表 2　GNSS 站建模精度

Tab. 2　Modeling accuracy of GNSS stations

站点	MAE/mm	RMSE/mm	D_E/mm	SMAPE/%
BJYQ	0.84	1.09	−0.47 ±0.99	21.13
BJGB	0.82	1.08	0.12 ±1.07	20.56
HECC	0.73	0.99	−0.35 ±0.93	19.19
HECD	1.48	2.05	−0.53 ±1.98	20.43
HEZJ	0.86	1.26	0.30 ±1.22	28.76
NMAG	1.30	1.74	−0.10 ±1.74	17.57
NMDW	1.09	1.53	−0.20 ±1.52	28.89
NMWH	0.26	1.79	−0.29 ±1.77	13.89

从表 2 可以得到,MEMD-XGBoost 模型建模结果的 MAE 和 RMSE 值较为稳定,说明 MEMD-XGBoost 模型可以有效地进行 GNSS 高程时间序列建模研究。D_E 指标可以有效地反映模型建模结果的主要误差方向和误差值的离散程度,从表 2可以看出,模型在 BJGB 和 HEZJ 站的主要误差方向为正,即建模值偏大,在其余 6 个主要误差方向为负,即建模值偏小;D_E 指标中 ± 号后的数字主要反映误差的离散程度,从表 2 可以看出,8 个 GNSS 站的值均小于 2,反映了模型具有较好的稳定性,在大部分建模历元可以得到较高精度的建模结果。SMAPE 指标可以衡量模型建模的准确率,从表 2 可以看出,在 8 个 GNSS 站,MEMD-XGBoost 模型建模结果的 SMAPE 值可以控制在 13.89% ~ 28.89%,说明 MEMD-XGBoost 模型建模结果具有较高的准确率。图 8 为 8 个 GNSS 站绝对误差值分布情况。

从图 8 可以看出,大部分历元的误差可以在 2 mm 的精度范围内,验证了 MEMD-XGBoost 模型具有良好的稳定性。

对于时间序列,建模属于一种随机行为,会影响时间序列本身的稳定性。因此,实验通过分析原始时间序列和建模时间序列的功率谱密度,分析其频域上的特性。图 9 所示为 HECD 站的功率谱密度。

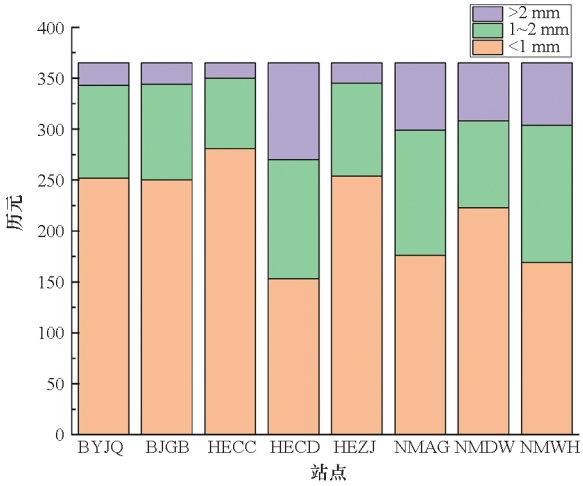
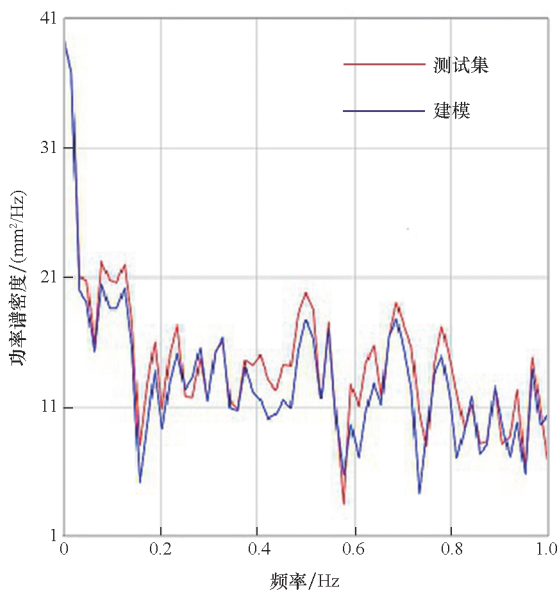


图 8　GNSS 站绝对误差值分布

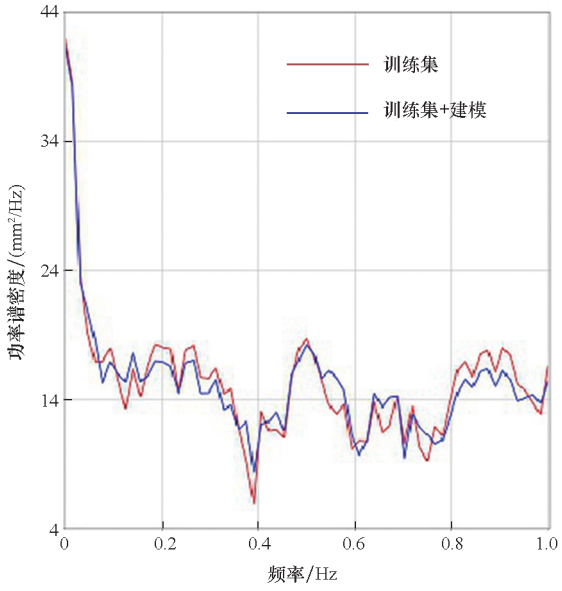
Fig. 8　Absolute error value distribution of GNSS stations

图 9 中,通过设置 3 组不同的功率谱密度对比,分析 MEMD-XGBoost 模型的建模性能。通过对比测试集中原始时间序列和建模时间序列的功率谱密度可以得到:在低频上两时间序列表现出较好的一致性,说明模型可以较好地保留原始时间序列的主要信息;在高频噪声上,模型建模的时间序列功率明显低于原始时间序列,说明 MEMD-XGBoost 模型具有抗异常值干扰能力,具有较好的鲁棒性。通过对于是否含有建模部分的时间序列功率谱密度可以得到,模型建模虽然对训练集时间序列的稳定性造成了影响,但影响较小,说明 MEMD-XGBoost 模型可以较为完善地学习到训练集中的信息。通过对比建模前后时间序列的功率谱密度可以得到,MEMD-XGBoost 模型建模结果



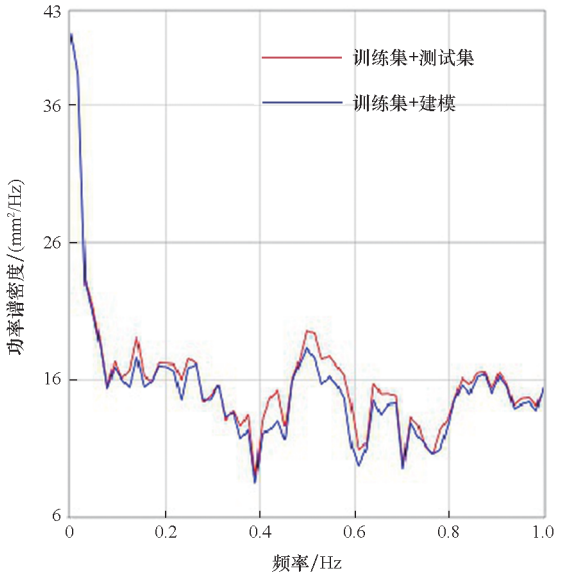
(a) 测试集建模前后功率谱密度

(a) Power spectral density of the test set before and after modeling



（b）有无建模结果功率谱密度

（b）Power spectral density with and without modeling results



（c）建模前后数据集功率谱密度

（c）Power spectral density of dataset before and after modeling

图9 HECD站功率谱密度

Fig. 9 Power spectral density of HECD station

可以有效保留原始时间序列中的主要信息，并通过模型的抗异常值干扰能力改善原始时间序列的数据质量，削弱异常值对于原始时间序列的影响。

上述实验说明了在建模过程中对目标时间序列通过降噪构造特征然后输入XGBoost模型可以在保留原始数据形态的情况下有效地改善数据质量。在进行预测实验时，通过邻近观测站[14]和部分地球物理观测值[39]构造特征可以达到实验目的。因此，实验选取目标观测站临近站点数据作为特征，通过MEMD-XGBoost模型对目标观测站进行预测。表3为8个观测站预测精度。

表3 8个观测站预测精度

Tab. 3 Prediction accuracy of 8 stations

目标站点	特征站点	MAE/mm	RMSE/mm
BJYQ	BJGB	3.64	4.67
BJGB	BIYQ	3.25	4.41
HECC	HEZJ	3.02	3.98
HECD	HECC	3.30	4.40
HEZJ	HECC	2.76	3.72
NMAG	NMDW	2.97	3.83
NMDW	NMAG	2.93	3.87
NMWH	HEZJ	3.20	4.12

从表3可以得到，当使用邻近观测站数据构造特征时，模型可以稳定有效地预测出目标观测站数据。该实验的预测精度与特征站点的选取有较强的关系，选取不同的特征观测站，实验结果会出现差异，具有高相关性的观测站更易获得高精度的预测结果。

4 结论

本文提出了一种通过多次经验模态分解改进的GNSS时间序列预测方法——MEMD-XGBoost模型，给出了模型的数据处理和预测策略，深入研究了模型的特点和预测结果的特性，并得出以下结论：

1）通过MEMD-XGBoost模型进行GNSS高程时间序列建模可以较好预测出GNSS站高程方向的运动趋势和季节性变化。

2）当使用目标时间序列构造特征时，实验通过4个精度评价指标衡量8个GNSS站的实验结果，结果表明，MEMD-XGBoost模型可以改善原始时间序列的数据质量，削弱异常值对于原始时间序列的影响。

3）当使用邻近观测站数据构造特征时，MEMD-XGBoost模型可以有效地预测出目标观测站高程方向的运动趋势，因此该模型也可用于区域网数据处理。此外，如何通过多源数据有效地构造特征需进一步研究。

参考文献(References)

[1] DENG L S, JIANG W P, LI Z, et al. Assessment of second- and third-order ionospheric effects on regional networks: case study in China with longer CMONOC GPS coordinate time

series[J]. Journal of Geodesy, 2017, 91: 207 – 227.

[2] HE X X, MONTILLET J P, FERNANDES R, et al. Review of current GPS methodologies for producing accurate time series and their error sources[J]. Journal of Geodynamics, 2017, 106: 12 – 29.

[3] 姜卫平, 王锴华, 李昭, 等. GNSS 坐标时间序列分析理论与方法及展望[J]. 武汉大学学报(信息科学版), 2018, 43(12): 2112 – 2123.
JIANG W P, WANG K H, LI Z, et al. Prospect and theory of GNSS coordinate time series analysis[J]. Geomatics and Information Science of Wuhan University, 2018, 43(12): 2112 – 2123. (in Chinese)

[4] 姚宜斌, 杨元喜, 孙和平, 等. 大地测量学科发展现状与趋势[J]. 测绘学报, 2020, 49(10): 1243 – 1251.
YAO Y B, YANG Y X, SUN H P, et al. Geodesy discipline: progress and perspective[J]. Acta Geodaetica et Cartographica Sinica, 2020, 49(10): 1243 – 1251. (in Chinese)

[5] WANG J, JIANG W P, LI Z, et al. A new multi-scale sliding window LSTM framework (MSSW-LSTM): a case study for GNSS time-series prediction[J]. Remote Sensing, 2021, 13(16): 3328.

[6] HOBBS B, ORD A. Nonlinear dynamical analysis of GNSS data: quantification, precursors and synchronisation[J]. Progress in Earth and Planetary Science, 2018, 5: 36.

[7] XU K K, HE R, LI K Z, et al. Secular crustal deformation characteristics prior to the 2011 Tohoku-Oki earthquake detected from GNSS array, 2003 – 2011[J]. Advances in Space Research, 2022, 69(2): 1116 – 1129.

[8] MONTILLET J P, SZELIGA W M, MELBOURNE T I, et al. Critical infrastructure monitoring with global navigation satellite systems[J]. Journal of Surveying Engineering, 2016, 142(4): 04016014.

[9] XI R J, JIANG W P, MENG X L, et al. Rapid initialization method in real-time deformation monitoring of bridges with triple-frequency BDS and GPS measurements[J]. Advances in Space Research, 2018, 62(5): 976 – 989.

[10] KONAKOGLU B, CAKIR L, YILMAZ V. Monitoring the deformation of a concrete dam: a case study on the Deriner Dam, Artvin, Turkey[J]. Geomatics, Natural Hazards and Risk, 2020, 11(1): 160 – 177.

[11] ALTAMIMI Z, REBISCHUNG P, MÉTIVIER L, et al. ITRF2014: a new release of the international terrestrial reference frame modeling nonlinear station motions[J]. Journal of Geophysical Research (Solid Earth), 2016, 121(8): 6109 – 6131.

[12] LAHTINEN S, JIVALL L, HÄKLI P, et al. Densification of the ITRF2014 position and velocity solution in the Nordic and Baltic countries[J]. GPS Solutions, 2019, 23(4): 95.

[13] LI Z, CHEN W, DAM T V, et al. Comparative analysis of different atmospheric surface pressure models and their impacts on daily ITRF2014 GNSS residual time series[J]. Journal of Geodesy, 2020, 94: 42.

[14] LI Z, LU T D. Prediction of multistation GNSS vertical coordinate time series based on XGBoost algorithm[C]// Proceedings of China Satellite Navigation Conference, 2022: 275 – 286.

[15] XU X H, REN W J. A hybrid model based on a two-layer decomposition approach and an optimized neural network for chaotic time series prediction[J]. Symmetry, 2019, 11(5): 610.

[16] 李福兴, 陈伏龙, 蔡文静, 等. 基于 EMD 组合模型的径流多尺度预测[J]. 地学前缘, 2021, 28(1): 428 – 437.
LI F X, CHEN F L, CAI W J, et al. Multiscale runoff prediction based on the EMD combined model[J]. Earth Science Frontiers, 2021, 28(1): 428 – 437. (in Chinese)

[17] HU H, ZHANG J F, LI T. A comparative study of VMD-based hybrid forecasting model for nonstationary daily streamflow time series[J]. Complexity, 2020, 2020: 4064851.

[18] HE X, BOS M S, MONTILLET J P, et al. Investigation of the noise properties at low frequencies in long GNSS time series[J]. Journal of Geodesy, 2019, 93: 1271 – 1282.

[19] 熊常亮, 贺小星, 鲁铁定, 等. 改进经验模态分解方法用于 GNSS 站速度估计[J]. 测绘科学, 2022, 47(3): 43 – 49.
XIONG C L, HE X X, LU T D, et al. Improved empirical mode decomposition method for velocity estimation of GNSS station[J]. Science of Surveying and Mapping, 2022, 47(3): 43 – 49. (in Chinese)

[20] 刘丹丹. 基于 EMD 的 GNSS 时间序列异常值探测算法[J]. 地球物理学进展, 2021, 36(5): 1865 – 1873.
LIU D D. New method of outlier detection for GNSS coordinate time series based on EMD approach[J]. Progress in Geophysics, 2021, 36(5): 1865 – 1873. (in Chinese)

[21] 李保金, 李艳艳. 高频 GNSS 信号去噪的小波和多方向主成分分析[J]. 全球定位系统, 2021, 46(4): 33 – 39.
LI B J, LI Y Y. Combined method of wavelet and MPCA for high-rate GNSS signal denoising[J]. GNSS World of China, 2021, 46(4): 33 – 39. (in Chinese)

[22] LIU H L, YANG L Q, LI L C. Analyzing the impact of climate factors on GNSS-derived displacements by combining the extended helmert transformation and XGBoost machine learning algorithm[J]. Journal of Sensors, 2021, 2021: 9926442.

[23] DEY A, RAHMAN M, RATNAM D V, et al. Automatic detection of GNSS ionospheric scintillation based on extreme gradient boosting technique[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 3091700.

[24] ZHUKOV A V, YASYUKEVICH Y V, BYKOV A E. GIMLi: Global ionospheric total electron content model based on machine learning[J]. GPS Solutions, 2020, 25: 19.

[25] JIA Y, JIN S G, SAVI P, et al. GNSS-R soil moisture retrieval based on a XGBoost machine learning aided method: performance and validation[J]. Remote Sensing, 2019, 11(14): 1655.

[26] HUANG N E, SHEN Z, LONG S R, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis[J]. Proceedings of the Royal Society of London Series A: Mathematical, Physical and Engineering Sciences, 1998, 454(1971): 903 – 995.

[27] BAJAJ V, PACHORI R B. Classification of seizure and nonseizure EEG signals using empirical mode decomposition[J]. IEEE Transactions on Information Technology in Biomedicine, 2012, 16(6): 1135 – 1142.

[28] PRASANNA KUMAR M K, KUMARASWAMY R. Single-channel speech separation using combined EMD and speech-specific information[J]. International Journal of Speech Technology, 2017, 20: 1037 – 1047.

[29] GUO Z H, ZHAO W G, LU H Y, et al. Multi-step

forecasting for wind speed using a modified EMD-based artificial neural network model [J]. Renewable Energy, 2012, 37: 241 – 249.

[30] LEI Z F, SU W B. Mold level predict of continuous casting using hybrid EMD-SVR-GA algorithm[J]. Processes, 2019, 7(3): 177.

[31] CHEN T Q, HE T, Benesty M, et al. XGBoost: extreme gradient boosting[EB/OL]. (2024 – 07 – 24)[2022 – 06 – 01]. https://cran. ms. unimelb. edu. au/web/packages/xgboost/vignettes/xgboost. pdf

[32] ZHANG X J, ZHANG Q R. Short-term traffic flow prediction based on LSTM-XGBoost combination model[J]. Computer Modeling in Engineering & Sciences, 2020, 125(1): 95 – 109.

[33] 杨可可, 刘立龙, 陈军. 基于滑动四分位距法的地震期间电离层 TEC 异常[J]. 桂林理工大学学报, 2019, 39(2): 427 – 432.
YANG K K, LIU L L, CHEN J. Abnormality of ionospheric TEC during earthquake based on sliding interquartile rang method[J]. Journal of Guilin University of Technology, 2019, 39(2): 427 – 432. (in Chinese)

[34] JIANG X C, LUO Y W, ZHANG B. Prediction of $PM_{2.5}$ concentration based on the LSTM-TSLightGBM variable weight combination model[J]. Atmosphere, 2021, 12(9): 1211.

[35] 李威, 鲁铁定, 贺小星, 等. 基于 Prophet-RF 模型的 GNSS 高程坐标时间序列预测分析[J]. 大地测量与地球动力学, 2021, 41(2): 116 – 121.
LI W, LU T D, HE X X, et al. Prediction and analysis of GNSS vertical coordinate time series based on prophet-RF model[J]. Journal of Geodesy and Geodynamics, 2021, 41(2): 116 – 121. (in Chinese)

[36] KIM T Y, CHO S B. Predicting residential energy consumption using CNN-LSTM neural networks[J]. Energy, 2019, 182: 72 – 81.

[37] HOU J W, WANG Y J, ZHOU J, et al. Prediction of hourly air temperature based on CNN-LSTM[J]. Geomatics, Natural Hazards and Risk, 2022, 13(1): 1962 – 1986.

[38] ZHANG X Y, LU X, LI W D, et al. Prediction of the remaining useful life of cutting tool using the Hurst exponent and CNN-LSTM[J]. The International Journal of Advanced Manufacturing Technology, 2021, 112: 2277 – 2299.

[39] GAO W Z, LI Z, CHEN Q S, et al. Modelling and prediction of GNSS time series using GBDT, LSTM and SVM machine learning approaches[J]. Journal of Geodesy, 2022, 96: 71.