基于 ARIMA 改进的实时动态滑坡预测模型

冯文杰[1,2] 王习东[1,2] 叶 永[3]
FENG Wenjie WANG Xidong YE Yong

摘 要 针对应急救援环境下，滑坡实时动态位移测量存在数据波动范围广、噪声大、模态混叠等问题，提出了一种基于 CEEMDAN-Kmeans-ARIMA 的组合预测模型。首先基于自适应噪声完备集合的经验模态分解算法，将添加 PPP 定位偏差噪声的斋藤模型信号分解为多个本征模态函数，并且根据 K-means 聚类算法将物理意义相近的本征模态函数进行聚类重构；然后针对多个聚类重构分量，构建最优的差分自回归移动平均预测模型；最后将聚类重构分量的预测值进行叠加，得到组合模型的预测值。实验结果表明，组合模型的 MAPE 指标相对于 ARIMA 单模型提高了 2.16%，解决了 ARIMA 预测模型存在的突变量不敏感、剩余滑坡预测时间不准等问题。在救援环境下，采用所提出的模型对实时动态滑坡进行预警预测具有一定的工程应用价值。

关键词 滑坡；集合自适应噪声经验模态分解；聚类；时序数据预测；模态混叠；PPP

doi：10.3969/j.issn.1672-9528.2024.02.002

0 引言

滑坡灾害是全球范围内发生频率最高、分布范围最广、造成损失最重的自然灾害之一[1]。科学监测是预警预报的前提，合理体系是救援抢险的保障。我国在地质灾害防治方面，建立了较为完善的监测预警机制，每年成功预报和避让的地质灾害近千起，但我国每年有数十万次的救援人员参与各种滑坡、坍塌山体的救援任务，亟需加强救援现场的监测预警水平，完善救援体系，科学施救，保障救援人员的生命安全[2]。

滑坡预测受滑坡体地质条件和众多环境因素的影响，具有复杂、非线性等特点。赵淑敏[3]利用集合经验模态分解（EEMD）将数据分解为主趋势项和误差项，然后分别采用优化径向基神经网络和马尔科夫链对主趋势项和误差项进行预测。张明岳等人[4]利用变分模态分解（VMD）将位移数据分解为趋势项、周期项和随机项分量，通过双向长短时记忆（Bi-LSTM）神经网络对累计位移进行预测。Nguyen 等人[5]通过研究越南安沛省穆庚寨县的滑坡数据，提出了多重增强朴素贝叶斯数模型，该模型在滑坡空间预测上具有出色表现。Bezak 等人[6]采用多元线性回归模型和随机森林模型分别进行了降水与地下水位关系建模和地下水位与滑坡关系建模，验证了两次建模预测比直接预测的效果更佳。

现阶段有大量的文献讨论了滑坡易发点的问题，但是对于实际的短期预测研究较少[7]。基于上述认知，本文提出了一种基于 CEEMDAN-Kmeans-ARIMA 组合的预测模型。该模型为提高预测实时性，采用实时厘米级的 PPP 定位数据作为滑坡监测数据[8]。首先利用集合自适应噪声经验模态方法（CEEMDAN）分解信号，并利用 K-means 聚类算法将具有类似特征的本征模态函数（IMF）进行聚类。然后对聚类后的数据进行重构，采用差分自回归移动平均（ARIMA）模型对重构后的数据进行预测。最后将预测后的数据再进行重构，得到最终的预测信号。

1 理论基础

1.1 CEEMDAN 分解算法

经验模态分解（EMD）是时频分析中常用的信号处理方式，在分析非线性非平稳信号时具有显著的优势。CEEMDAN 是在 EMD 的基础上发展而来的，CEEMDAN 相比于 EEMD 具有更低的运算成本，相比于互补集合经验模态分解（CEEMD）具有更好的重构效果。CEEMDAN 是结合了 EMD 的分解理念、EEMD 的统计特性改进而来的，具体算法流程如下。

（1）在原始时序信号 $x(t)$ 中添加一系列服从正态分布的自适应白噪声 $\varepsilon_0\omega^i(t)$，得到 $x_i(t)$ 信号：

$$x_i(t)=x(t)+\varepsilon_0\omega^i(t),i=1,2,3,\cdots,n \quad (1)$$

1. 水电工程智能视觉监测湖北省重点实验室 湖北宜昌 443002
2. 三峡大学计算机与信息学院 湖北宜昌 443002
3. 三峡大学水利与环境学院 湖北宜昌 443002
[基金项目] 国家重点研发计划资助项目（No. 2021YFC3001903）

式中：$x_i(t)$ 是第 i 次添加白噪声后的时间序列，ε_0 为第一次添加白噪声的自适应因数，$\omega^i(t)$ 为第 i 次添加的白噪声，n 为添加白噪声的次数。

（2）采用 *EMD* 算法对上述生成的时间序列 $x_i(t)$ 进行分解，得到第一个本征模态函数 $IFM_{1i}(t)$，然后将 $IFM_{1i}(t)$ 进行平均，得到原始时序信号 $x(t)$ 的 $IFM_1(t)$：

$$IMF_{1i}(t)=EMD_1\left(x_i(t)\right) \tag{2}$$

$$IMF_1(t)=\frac{1}{n}\sum_{i=1}^{n}IMF_{1i}(t) \tag{3}$$

式中：$EMD_1(\cdot)$ 运算表示求 $IFM_1(t)$ 运算。

（3）求得 $x(t)$ 的 $IFM_1(t)$ 后，就可以得到余量 $R_1(t)$：

$$R_1(t)=x(t)-IMF_1(t) \tag{4}$$

（4）然后再对 $R_1(t)$ 添加一系列服从正态分布的自适应白噪声分量 $\varepsilon_1 EMD_1(\omega^i(t))$，得到 $R_{1i}(t)$ 信号：

$$R_{1i}(t)=R_1(t)+\varepsilon_1 EMD_1\left(\omega^i(t)\right) \tag{5}$$

（5）将 $R_{1i}(t)$ 替换步骤 2 中的 $x_i(t)$，求得 $IMF_2(t)$，然后计算出余量 $R_2(t)$，再添加噪声分量得到 $R_{2i}(t)$，重复上述步骤，进行迭代计算。$R(t)$ 不能继续分解时，结束迭代，得到 k 个 $IMF(t)$ 分量，原始信号最终被分解为：

$$x(t)=R(t)+\sum_{j=1}^{k}IMF_j(t) \tag{6}$$

1.2 K-means 聚类算法

K-means 聚类算法属于分区聚类结构，通过给定的聚类数目和聚类初始点进行迭代，使各个类别的质点到其余点的距离和最小，完成迭代。最终将数据划分为指定数量、互不相交、非层次的多种类别 [9]。K-means 算法通过欧式距离来度量数据之间的差异，本文采用 K-means 对 CEEMDAN 分解得到的 *IMF* 进行聚类，因此设数据的样本数为 m，样本维度为 n，那么样本点 $\mathbf{x_i}$ 为：

$$\mathbf{x_i}=\left(x_{i1},x_{i2},x_{i3},\cdots,x_{in}\right),i=1,2,3,\cdots,m \tag{7}$$

那么有任意两个样本点 $\mathbf{x_i},\mathbf{x_j}\left(i,j=1,2,3,\cdots,m\right)$ 的欧式距离 $d_{ij}=d\left(\mathbf{x_i},\mathbf{x_j}\right)$ 为：

$$d_{ij}=\sqrt{\sum_{k=1}^{n}\left(x_{ik}-x_{jk}\right)^2} \tag{8}$$

如果要将 m 个样本数分成 k 类，那么就得确定 k 个初始聚类中心，首先通过公式（8）计算所有点之间的距离，选择样本中距离最远的两个点$\mathbf{x}_\alpha,\mathbf{x}_\beta$作为初始点，那么有：

$$d_{\mathbf{x}_\alpha\mathbf{x}_\beta}=d\left(\mathbf{x}_\alpha,\mathbf{x}_\beta\right)=max\left\{d_{ij}\right\} \tag{9}$$

再确定下一个初始聚类点$\mathbf{x}_\gamma$，使得$\mathbf{x}_\gamma$到$\mathbf{x}_\alpha,\mathbf{x}_\beta$的欧式距离最小值等于样本中其余点到$\mathbf{x}_\alpha,\mathbf{x}_\beta$的欧式距离较小值的最大值，也即是：

$$\left\{\begin{array}{c} min\left\{d_{\mathbf{x}_\alpha\mathbf{x}_\gamma},d_{\mathbf{x}_\beta\mathbf{x}_\gamma}\right\}=max\left\{min\left\{d_{\mathbf{x}_\alpha\mathbf{x}_i},d_{\mathbf{x}_\beta\mathbf{x}_i}\right\}\right\} \\ s.t.\ \mathrm{i}\neq\alpha,\beta \end{array}\right\} \tag{10}$$

根据式（10）进行迭代计算，就可以求得 k 个初始聚类点集合 $S^{(0)}$，然后通过欧式距离最小原则完成其余样本点的聚类，得到 k 个样本聚合集合 $C^{(0)}$：

$$S^{(0)}=\left\{x_1^{(0)},x_2^{(0)},\cdots,x_k^{(0)}\right\} \tag{11}$$

$$C^{(0)}=\left\{C_1^{(0)},C_2^{(0)},\cdots,C_k^{(0)}\right\} \tag{12}$$

求得聚合集合 $C^{(0)}$ 后，计算每个聚类集合的质心 $\mathbf{x_i}^{(1)}$，得到新的聚类初始点集合 $S^{(1)}$：

$$S^{(1)}=\left\{x_1^{(1)},x_2^{(1)},\cdots,x_k^{(1)}\right\} \tag{13}$$

迭代求解 p+1 次聚合集合得到 $C^{(p)}$，然后求得下一次初始聚类点 $S^{(p+1)}$。当 $S^{(p+1)}=S^{(p)}$、$C^{(p+1)}=C^{(p)}$，也即是两次迭代结果一致时，迭代结束，此时的 $C^{(p)}$ 集合即为 K-means 算法的聚类结果。

1.3 ARIMA 预测模型

ARIMA 是一种非线性、非平稳时间序列预测模型，通过对非平稳时间序列数据进行差分平稳化处理，然后采用拟合和参数估计的方法，建立数学模型 [10]。ARIMA 模型由差分模型、自回归模型、移动平均模型 3 部分组成，其构成如下：

$$ARIMA(p,d,q)=AR(p)+Diff(d)+MA(q) \tag{14}$$

式中：$AR(p)$ 是自回归模型，$Diff(d)$ 是差分模型，$MA(q)$ 是移动平均模型，p 是自回归模型阶数，d 是差分模型阶数，q 是移动平均模型阶数。

1.3.1 差分模型

差分模型是将非线性、非平稳的数据进行差分，得到平稳数据的过程。设非线性、非平稳时间序列为 $x(t)$，$diff(\cdot)$ 为差分运算符，一阶差分表达式 $\Delta_1x(t)$ 为：

$$\Delta_1x(t)=diff_1\left(x(t)\right)=x(t)-x(t-1) \tag{15}$$

检验一阶差分序列 $\Delta_1x(t)$ 是否为平稳信号，若不为平稳信号，则继续求解二阶差分：

$$\Delta_2x(t)=diff_2\left(x(t)\right)=diff_1\left(\Delta_1x(t)\right) \tag{16}$$

通过式（15）进行迭代求解，直至 $x(t)$ 转化为平稳时间序列 $\Delta_dx(t)$，完成 d 阶差分过程。

1.3.2 自回归模型

自回归模型是通过过去 p 个时刻的值对未来时刻值进行线性拟合，然后加上常数项和下一时刻的随机噪声。设经过 d 阶差分之后的平稳时间序列为 $y(t)$，下一时刻的预测值为 $y(t+1)$，下一时刻的随机噪声为ξ_{t+1}，常数项为 c，自相关系数为 β_i，那么 $AR(p)$ 模型预测为：

$$y(t+1)=c+\sum_{i=1}^{p}\beta_i y(t-i+1)+\xi_{t+1} \tag{17}$$

1.3.3 移动平均模型

移动平均模型是通过过去 q 个时刻的随机噪声对未来

时刻值进行线性拟合，然后加上常数项和下一时刻的随机噪声。设经过 d 阶差分之后的平稳时间序列 $y(t)$，$y(t)$ 的白噪声序列为 $e(t)$，下一时刻的预测值为 $y(t+1)$，下一时刻的随机噪声为 $e(t+1)$，常数项为 μ，白噪声系数为 θ_i，那么 $MA(q)$ 模型预测为：

$$y(t+1)=\mu+\sum_{i=1}^{q}\theta_i e(t-i+1)+e_{t+1} \tag{18}$$

2 模型建立

PPP 定位监测滑坡动态位移本身就具有不平稳的波动特性，又由于 PPP 定位时受到 GNSS 信号质量、星历数据、多路径效应、天线相位中心偏移、大气延迟等因素的影响，使得 PPP 定位具有一定程度的非线性和非平稳特性。若直接采用 ARIMA 模型进行预测，会导致预测精度不高。基于上述问题和采用算法的基本原理，同时考虑到信号模态混叠特征及预测效率，本文提出了一种基于 CEEMDAN-Kmeans-ARIMA 的组合滑坡预测模型。具体流程框图如图 1 所示。

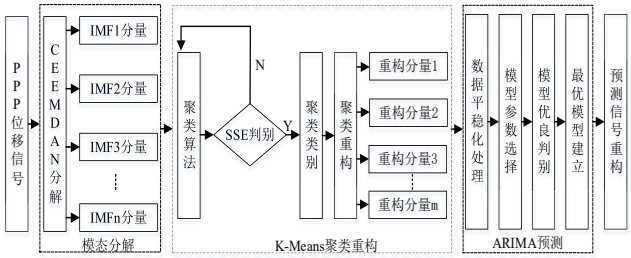

图 1 模型整体架构

具体步骤如下。

（1）采用 CEEMDAN 算法将滑坡位移信号分解成一系列的 IMF 分量。

（2）通过 K-means 聚类算法对 IMF 分量进行聚类，根据平方误差和（SSE）确定聚类类别数，将类别相同的信号进行重构。

（3）采用 AIC、BIC 信息准则，选择拟合优度和可解释性最佳的 ARIMA 模型。

（4）对聚类重构后的数据进行预测，然后将预测数据进行重构，得到最终的预测数据。

2.1 实验数据

为验证 CEEMDAN-Kmeans-ARIMA 组合模型对 PPP 定位监测数据的预测效果，首先通过 PPP 定位接收机每 6 s 获取一次静态定位信息，采集 7.5 h 数据，选取 PPP 定位收敛后的 6.05 h 的数据作为实验数据。然后考虑模型建立时间及现场救援人员紧急撤离时间，再对数据进行抽取，使得定位数据为每分钟一次，共 363 组数据。图 2 为抽取之后 PPP 定位信号的定位偏差。

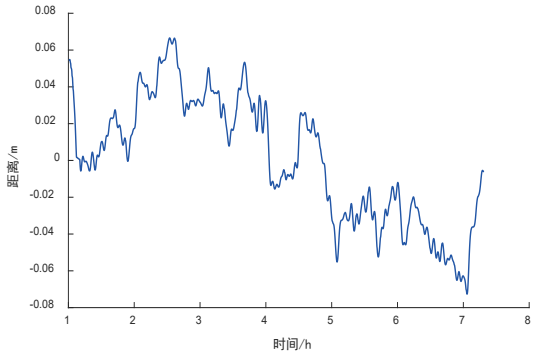

图 2 PPP 定位偏差信号

通过实验得知，该 PPP 定位接收机纬度方向的静态定位的方差为 3.41 cm。为进一步验证本文提出的组合模型的预测可行性，同时满足实际应急救援现场中需要快速对滑坡发生时间进行初判，引进斋藤时间预报模型进行仿真预测。斋藤时间预报模型仅通过位移时间序列，在不借助第三方软件的条件下即可计算预报结果，具体方法为：

$$T=\frac{0.5(t_2-t_1)^2}{(t_2-t_1)-0.5(t_3-t_1)} \tag{19}$$

式中：T 为滑坡发生剩余时间；t_1、t_2、t_3 分别是位移时间序列中的三个点，但是其时间间隔内的变形量相等。然后参考龙井村滑坡 1# 位移计获取的精细化变形过程 [11]，通过相应函数构造三段式的斋藤模型，并加入对应的 PPP 定位噪声，具体实验数据如图 3 所示。

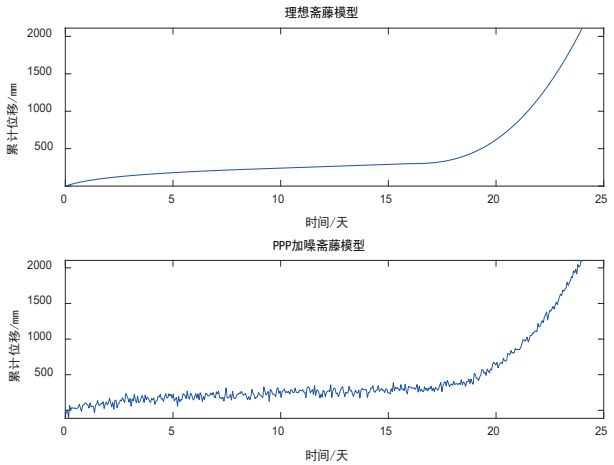

图 3 斋藤模型与加噪斋藤模型

2.2 数据 CEEMDAN 分解

采用 CEEMDAN 方法对 PPP 加噪斋藤模型信号进行分解，得到 6 个序列，分解后的子序列如图 4 所示。IMF 分量随着分解次数的增加，信号频率不断降低，也即是信号的不同模态被分离开来。IMF1 ～ IMF3 为高频分量，IMF4 和 IMF5 为低频分量，Res 为趋势项。从趋势项可以看出，滑坡存在初始变形、匀速变形和加速变形三个阶段，与模拟的实

际情形相吻合。

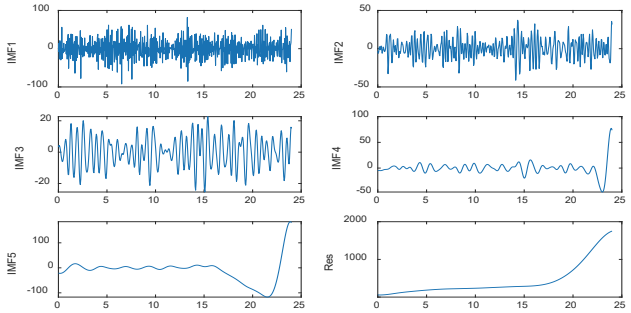

图 4 PPP 定位信号 IMF 分量

2.3 数据 K-means 聚合及重构

将 CEEMDAN 分解后的 IMF 分量进行聚类，首先通过误差平方和（SSE）与聚类类别的关系可知，在聚类类别为 4 时，SSE 数值趋近于平缓，因此将 K-means 的聚类数规定为 4。然后将 IMF 分量进行聚类，聚类结果如图 5 所示。

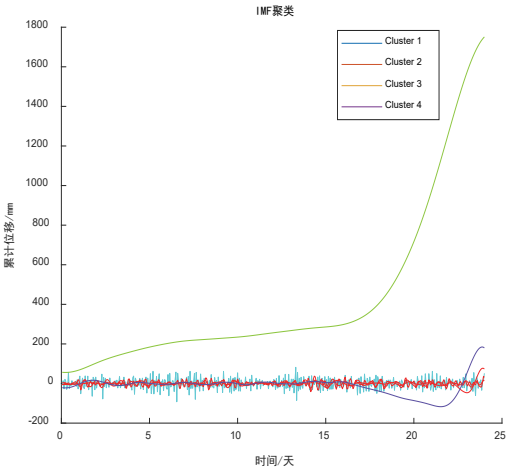

图 5 IMF 分量 K-means 聚类

K-means 算法将 IMF 分量进行分类之后，将同一类的 IMF 分量进行重构，重构的结果如图 6 所示。重构之后减少 IMF 分量，从而减小了预测时的计算量，提高了预测效率。

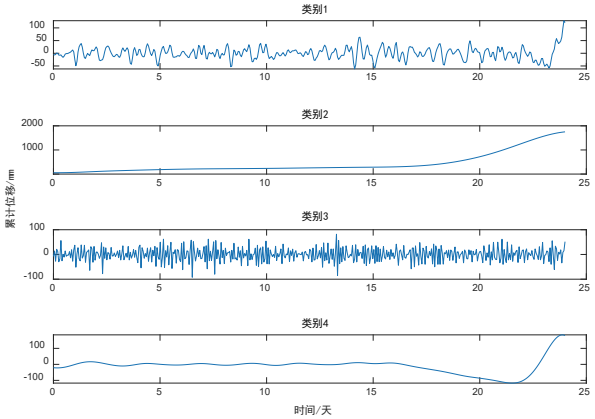

图 6 聚类数据重构

2.4 组合模型数据预测

采用 ARIMA 模型对聚类重构后的数据进行预测，判断聚类重构后的各数据是否具有平稳特性，将各数据平稳化处理后，观察数据的自相关函数和偏相关函数，确定合适的 p、q 范围后，引进信息准则 AIC、BIC，通过轮询计算和信息准则最小原则确定最优的模型参数，完成最优模型建立。为了更好地检验模型的外推能力，本文采用滚动预测的方式对模型进行训练和预测，也即是将最后 20% 的数据用于预测，前面 80% 的数据全部用于训练。然后每预测完一个数据，就将该时刻的数据归为训练集，再次构建预测模型，预测下一时刻的值，直至将后 20% 的数据预测完成，具体预测结果如图 7 所示。

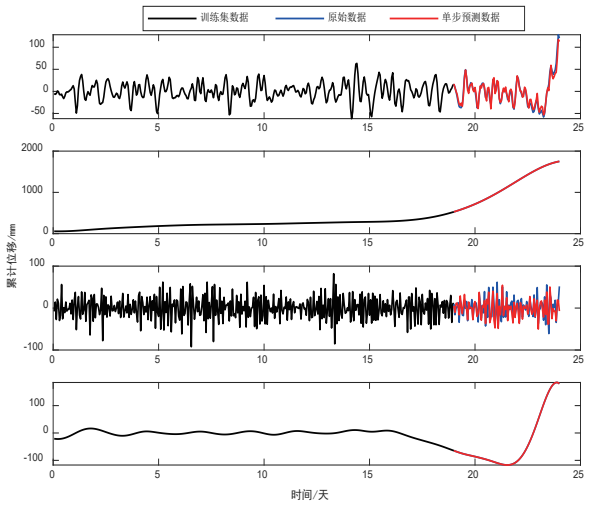

图 7 K-means 聚类重构数据预测

从预测结果可以看到，当数据的频率成分较为单一时，原始数据与单步预测的结果基本重合；当数据的频率成分复杂时，单步预测数据与原始数据变化趋势具有一致性，相关性较好。将预测分量进行重构，得到最终的预测信号，预测信号与实际信号对比结果如图 8 所示。

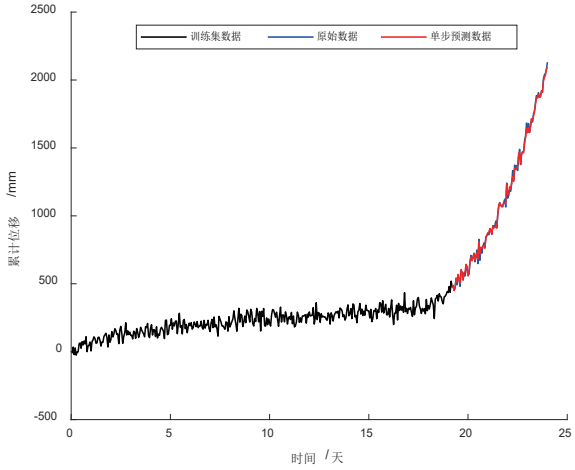

图 8 组合模型预测结果

从图 8 中可以观察到，预测信号与实际信号随时间变化的趋势一致，且预测信号能及时预测到信号的突变量，这说明预测模型在预测滑坡时具有较好的可信度。然后仅采用 ARIMA 单模型进行预测，预测结果如图 9 所示，单模型预测只能预测到滑坡的趋势量，不能较好地预测到突变量，与滑坡实时监测预警的理念不符。

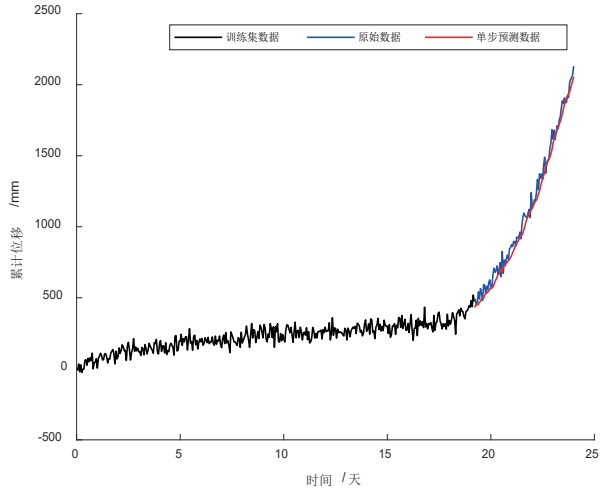

图 9 单模型预测结果

为更客观分析单模型和组合模型的预测效果，引进平均绝对百分比误差（MAPE）来衡量预测模型的准确性。单模型预测的 MAPE 值为 5.11%，组合模型的 MAPE 值为 2.95%，从 MAPE 评价指标上看，组合模型的预测效果在单模型预测效果基础上提高了 2.16%。然后通过斋藤模型剩余滑坡时间计算公式可知，组合模型预测结果与数值模拟的过程一致，随着滑坡速率的增加，剩余滑坡时间越来越短。单模型预测的结果在最后一段预测时，呈现出剩余滑坡时间增加的现象，与实际结果不符，也即是预测结果出现偏差导致预测剩余时间不准确。斋藤模型剩余时间预测的具体结果如表 1 所示。

表 1 斋藤模型剩余时间预测结果

累计位移 /mm		0	500	1000	1500	2000
单模型	实际时间 / 天	0	16.3	18.04	19.08	19.92
	滑坡时间 / 天			18.24	4.38	5.21
组合模型	实际时间 / 天	0	16.17	17.92	19.04	19.83
	滑坡时间 / 天			18.13	4.9	3.80

3 结语

本文构建了一个基于自适应噪声完备集合的经验模态分解、K 均值聚类和差分自回归移动平均的组合滑坡位移动态预测模型，首先获取 PPP 静态定位的定位偏差，然后构建斋藤模型并添加对应的 PPP 定位偏差得到实验数据，最后利用实验数据进行了单模型和组合模型验证，得出以下结论。

（1）采用 CEEMDAN 对数据进行分解，可以将 PPP 定位数据的不同模态分离开来，赋予分解后的 IMF 分量对应的物理意义，提高了预测准确性。

（2）采用 K-means 聚类算法将具有相似物理意义的 IMF 分量聚合，提高预测效率。

（3）本文提出的组合模型比 ARIMA 预测模型的 MAPE 指标提高了 2.16%，同时在剩余滑坡时间上的预测也更为准确，更具有实际参考意义。

本文仅通过 PPP 定位噪声和数值模拟实验对模型进行了评价，后期欲采用自主研发的监测设备获取实际滑坡监测数据，以进一步佐证其准确性。

参考文献：

[1] 许强，朱星，李为乐，等．“天 - 空 - 地”协同滑坡监测技术进展 [J]. 测绘学报，2022, 51(7): 1416-1436.

[2] 马海涛．重特大地质灾害应急救援现场监测预警保障体系研究 [J]. 中国安全生产科学技术，2022, 18(S1): 5-10.

[3] 赵淑敏．基于信息分解条件的滑坡变形预测 [J]. 水土保持通报，2021,41(3):181-186.

[4] 张明岳，李丽敏，温宗周．基于变分模态分解和双向长短时记忆神经网络模型的滑坡位移预测 [J]. 山地学报，2021, 39(6): 855-866.

[5] NGUYEN P T, TUYEN T T, SHIRZADI A, et al. Development of a novel hybrid intelligence approach for landslide spatial prediction[J]. Applied sciences-basel, 2019, 9(14):25.

[6] BEZAK N, JEMEC A M, MIKOŠ M. Application of hydrological modelling for temporal prediction of rainfall-induced shallow landslides[J]. Landslides, 2019, 16(7):1273-1283.

[7] COLLINI E, PALESI L A I, NESI P, et al. Predicting and understanding landslide events with explainable AI[J]. IEEE access, 2022,10: 31175-31189.

[8] 王利，张勤，黄观文，等．GPS PPP 技术用于滑坡监测的试验与结果分析 [J]. 岩土力学，2014(7): 2118-2124.

[9] 何选森，何帆，徐丽，等．K-means 算法最优聚类数量的确定 [J]. 电子科技大学学报，2022, 51(6): 904-912.

[10] 李颖若，韩婷婷，汪君霞，等．ARIMA 时间序列分析模型在臭氧浓度中长期预报中的应用 [J]. 环境科学，2021, 42(7): 3118-3126.

[11] 亓星，朱星，许强，等．基于斋藤模型的滑坡临滑时间预报方法改进及应用 [J]. 工程地质学报，2020,28(4): 832-839.

【作者简介】

冯文杰（1999—），男，湖北黄冈人，硕士，研究方向：智能化检测装备与技术。

王习东（1976—），男，湖北黄冈人，讲师，博士，研究方向：光电检测技术及系统、弱磁检测、FPGA 应用开发等。

（收稿日期：2023-11-08）