Time Series Prediction of GNSS Elevation Coordinates Based on Anomaly Detection

Li Wei[1]，Lu Tieding[1]，Cao Xiaoming[2]，Han Zhao[3]

1. School of Surveying and Mapping Engineering , East China University Of Technology ,Nanchang330013 ,China

2. Xi'an Academy of Surveying and Mapping, Xi'an710054,China

3. Overseas Branch of CCCC Road and Bridge Construction Co. Ltd., Beijing100000,China

1. liwei.times@qq.com, 2. tdlu@whu.edu.cn

Abstract: A combined LOF-Prophet-RF prediction model based on local outlier factor (LOF) is proposed for GNSS elevation coordinate time series with many outliers and large and unsteady data distribution intervals. Firstly, the GNSS elevation coordinate data are detected by LOF algorithm, and then the missing data are interpolated by Prophet model to obtain the new elevation coordinate time series after rejecting the detected outliers, and finally the prediction is carried out by Random Forest (RF) model. The experimental results show that the combined model can effectively reject the coarse outliers and has good sensitivity to local outliers; it can better represent the change trend of elevation coordinate time series and obtain more accurate prediction data than the traditional prediction methods.

Keywords: GNSS; Coordinate time series; Outlier detection; Interpolation; Prediction

基于异常值探测的GNSS高程坐标时间序列预测

李威[1]　鲁铁定[1]　曹小明[2]　韩钊[3]

1．东华理工大学测绘工程学院，南昌，中国，330013

2．西安市勘察测绘院，西安，中国，710054

3. 中交路桥建设有限公司海外分公司，北京，中国，100000

1. liwei.times@qq.com, 2. tdlu@whu.edu.cn

【摘要】针对GNSS高程坐标时间序列异常值多、数据分布区间大且不平稳等特点，提出了一种基于局部异常值因子（local outlier factor，LOF）的LOF-Prophet-RF组合预测模型。首先对GNSS高程坐标数据使用LOF算法进行异常值探测，剔除探测到的异常值后对缺失数据采用Prophet模型进行插值，得到新的高程坐标时间序列，最后使用随机森林模型（Random Forest，RF）进行预测。实验结果表明，该组合模型能够有效的剔除粗差，对局部异常值有良好的敏感性；较传统预测方法能更好的表现高程坐标时间序列的变化趋势，并得到更高精度的预测数据。

【关键词】GNSS；坐标时间序列；异常值探测；插值；预测

1 引言

随着GNSS观测技术与精度的提高，20多年来不断累积的GNSS坐标观测数据为进一步的地球物理研究与基准站的数学规律和特征分析提供了坚实的数据支撑[1]。但测站在测量时难以避免的受到外界环境和机器内部的影响从而导致异常值的产生，如多路径效应、电离层延迟、接收机信号故障、周跳等[错误!未找到引用源。]等。异常值对观测结果的可靠性与数据分析的准确性产生较大影响，因此准确的异常值探测是GNSS坐标时间序列重要的数据预处理步骤。而由于高程方向坐标时间序列表现出的非平稳性与非线性导致异常值探测更加困难，以及异常值对高程方向预测精度的影响需进一步分析。

国内外学者对 GNSS 坐标序列进行了大量研究，且在预测领域有着深入的研究与应用。坐标时间序列的准确预测，对研究基准站的长期变化规律、为灾害预防提供方法支撑等科学研究有着重要意义。现已有诸多方法应用于坐标时间序列的预测领域中，包括统计分析、指数平滑、周期项模型、最小二乘拟合、ARMA、神经网络、广义加法模型等方法[3]-[5]。

资助信息：国家重点研发计划（2016YFB0501405，2016YFB0502601-04）；国家自然科学基金（42061077；42064001）；江西省科技落地计划项目（KJLD12077）；江西省自然科学基金（20202BAB214029，2017BAB203032，20202BABL214055，20202BABL211007）；江西省教育厅科学技术研究项目（GJJ204015）。

这些方法的应用与创新提升了坐标时间序列的预测精度，但普遍存在以下问题：传统模型往往存在建模复杂、参数不易确定、稳定性差等缺陷，以及预测精度易受到异常值影响，建立的模型不具备长期预测能力[6]。

针对以上问题，本文提出一种 LOF-Prophet-RF 组合预测模型，该组合模型能够较准确的探测局部异常值，并得到符合序列变化的插值数据，以及随机森林模型对 GNSS 坐标时间序列有良好的适用性，具有稳定的预测精度。本文首先对坐标实测数据进行异常值探测与插值，并将处理前后的序列分别建立 RF 预测模型，以分析异常值对高程坐标时间序列预测精度的影响。

2 模型原理与方法

2.1 LOF异常值探测

局部异常因子（LOF）是由 Breunig 等提出的无监督的异常值探测算法[7]，通过计算点p与其邻域点的密度来判断该点是否为异常点，当点p的密度越低时，越可能是异常点。LOF 算法会对数据集中每一个数据点都计算一个离群异常因子 LOF，并根据 LOF 得分是来判断是否为异常值点。LOF 算法的主要定义与方法如下：

定义 1（k-邻近距离）

在数据集D中，距离数据点p最近的几个数据点中，第k个最近的数据点与点p之间的距离称为点p的k-邻近距离$(k-distance)$，记为$d_k(p)$；将两个数据点p和o的距离记为$d(p,o)$。任意两个数据点间的距离可以采用欧氏（Eucildean）距离、马氏（Mahalanobis）距离、闵可夫斯基（Minkowski）距离等方法计算。

定义 2（可达距离）

在给定参数k时，数据点p到点o的可达距离可记为：

$$reach-dist_k(p,o)=max\{k-distance(o),\ d(p,o)\} \tag{1}$$

定义 3（局部可达密度）

与点p的距离小于或等于k-邻近距离的数据点集合称作点p的k-距离领域，记作$N_k(p)$；点p与邻近数据点的平均可达距离的倒数为点p的局部可达密度，即：

$$lrd_k(p)=1/\left(\frac{\sum_{o\in N_k(p)}reach-dist_k(p,o)}{|N_K(P)|}\right) \tag{2}$$

定义 4（局部离群因子）

数据点p的局部离群因子表示为点p邻域$N_k(p)$的局部可达密度与数据点p的局部可达密度的平均比值，即：

$$LOF_k(P)=\frac{\sum_{o\in N_k(p)}\frac{lrd_k(o)}{lrd_k(p)}}{|N_K(P)|} \tag{3}$$

通过以上原理计算出局部数据点的可达距离，再根据可达距离计算局部可达密度，最后得到局部离群因子。当数据点的 LOF 值接近 1 时，表明该点的局部可达密度与邻域点的局部可达密度相似，越可能为正常点；而当 LOF 值大于 1 时，该点的密度则小于领域点的密度，可能是异常值[8]。

2.2 Prophet模型基本原理

由 Facebook 的数学家 Taylor 等在 2017 年构建的 Prophet 模型[9]，是一种分析时间序列的新模型，包括处理时间序列数据中异常值、缺失值，以及预测时间序列的变化规律。Prophet 模型通过广义加法模型对时间序列进行贝叶斯曲线拟合[6]，并在拟合预测过程中自动填补缺失值，从而进行插值。该模型的表达为：

$$y(t)=g(t)+s(t)+\hbar(t)+\varepsilon(t) \tag{4}$$

式(4)中，Prophet 模型由趋势项、季节性、假日项以及残差项四部分构成。其中，趋势项$g(t)$是 Prophet 模型的核心部分，表示时间序列中的非周期部分。$g(t)$趋势变化可用逻辑回归函数表示：

$$g(t)=C/(1+\mathrm{e}^{-k(t-m)}) \tag{5}$$

式(5)中，C表示模型的承载能力，k表示增长率，d表示偏移量。

$s(t)$表示周期项或季节项，常以周、月、季度或年为单位，该项的拟合主要依靠傅里叶级数构造灵活的周期模型，$s(t)$表示为：

$$s(t)=\sum_{n=1}^{N}\left(a_n cos\left(\frac{2\pi nt}{T}\right)+b_n sin\left(\frac{2\pi nt}{T}\right)\right) \tag{6}$$

式(6)中，T表示时间序列的周期，$2n$表示模型中周期的期望个数，a_n、b_n表示平滑参数。

$\hbar(t)$为不规律的假日项和特殊变动，无法通过周期性模型进行建模，因此 Prophet 模型为每个假日项构建独立的模型，但 GNSS 坐标时间序列通常不存在此类数据，因此采用 Prophet 模型默认模式即可。$\varepsilon(t)$为残差项，表示模型中未预测到的随机趋势[错误!未找到引用源.]。

2.3 随机森林基本原理

随机森林模型（Random Forest，RF）是一种基于 Bagging（可生产多个决策树分类器）和决策树

的集成学习算法[10]，由 Breiman 等在 2001 年提出，通常用来解决回归与分类的问题。随机森林是一种适用性强、建模高效的机器学习模型，在未调整超参数的情况下也能得到理想的结果。随机森林增强了模型的可解释性，且不易过拟合，这是因为随机森林引入了袋装法和特征子空间两种随机策略，对噪声和异常值具有良好的抵抗性。图 1 为随机森林的预测流程[12]。

随机森林的预测步骤：

1）采用 Boostrap 重采样方法从原始数据样本集N中随机生成n个训练样本集，以及各训练样本集对应的决策树。

2）决策树的特征维数为 M，从中随机选取m个特征作为子集。对于决策树每个节点，从 m 个特征中选择最优的属性进行分裂，且 m 的大小在随机森林的生长过程中保持不变。

3)使每颗决策树从上而下递归分裂直至生长的最大限度，且不进行任何减枝操作。

4）将各颗决策树预测得到的结果等权相加，并取其平均值作为最终预测值。

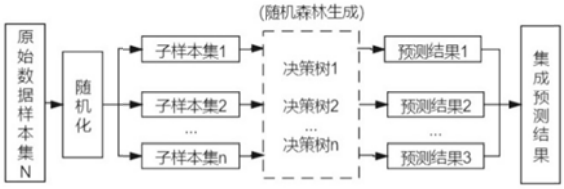

Figure 1. Random forest prediction process

图 1.随机森林预测流程

2.4 组合思想与方法

GNSS 高程坐标时间序列存在部分异常值，对于极端异常值也就是粗差，传统的异常值探方法能够进行准确的探测，但由于高程方向的序列往往表现出周期性与不平稳性，导致传统方法对于局部异常值的探测效果较差。而 LOF 算法通过计算每个数据点与它邻近数据点的局部可达密度来判断数据点的异常情况[13]，因此能够较好的对局部数据的异常值进行准确的探测。

对探测到的异常值进行剔除后，数据中会产生空值，因此在预测之前需要对缺失数据的序列进行插值。传统插值方法对小比例非连续的随机缺失数据能够有良好的插值效果，但对于连续缺失数据的插值效果较差，因此，本文采用基于预测性质的 Prophet 方法进行插值，该方法能够较好的拟合原始数据并对数据中的缺失数据进行填补。

最后对 LOF 算法与 Prophet 模型处理后的数据选用随机森林进行预测，随机森林在回归预测时通过平均决策树，从而降低过拟合的可能性。该算法具有很好的稳定性，只有半数以上的基分类器产生差错时才会得到错误的预测。本文方法的实验流程如图 2 所示。

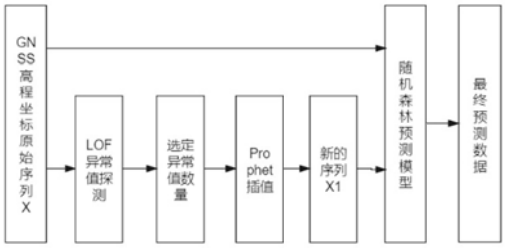

Figure 2. Experimental method flow

图 2.实验方法流程

2.5 模型评价指标

本文选用均方根误差（RMSE）与平均绝对误差（MAE）作为预测数据精度的评判标准，定义如下：

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(g_i - h_i)^2} \tag{7}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|g_i - h_i| \tag{8}$$

式中，g_i表示不同日期的预测值，h_i表示不同日期的原始值。当 RMSE 和 MAE 的值越小时，表示模型的精度越高，预测数据与原始数据的偏差就越小。

3 LOF-Prophet-RF 组合模型实验

由于 GNSS 坐标时间序列的水平方向通常呈现出长期的线性趋势，而不包含季节项，传统方法即有良好的效果，因此本文主要研究具有非线性的高程方向时间序列。本文选用 ZHNZ 站 2008 年 1 月 1 日~2016 年 1 月 30 日的高程坐标实测数据作为实验数据，数据来源于中国地震局 GNSS 数据产品服务平台（http://www.cgps.ac.cn/)。其中，2008 年 1 月 1 日~2015 年 12 月 31 日为训练样本集（图 3），2016 年 1 月 1 日~2016 年 1 月 30 日为测试样本集。由于本文专注异常值对预测数据的精度影响分析，以及各预测模型的精度对比，因此不对训练集的拟合精度进行分析。

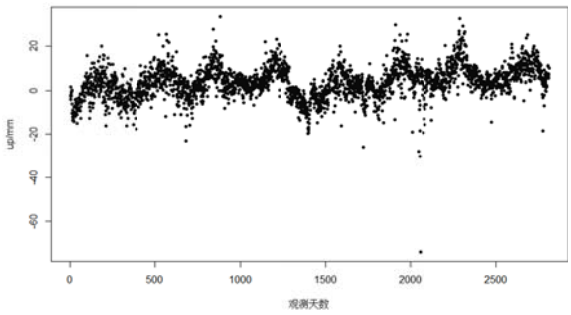

Figure 3. ZNHZ station elevation coordinates measured data
图 3.ZNHZ 站高程坐标实测数据

3.1 LOF异常值探测

由于 ZHNZ 站高程坐标时间序列中存在粗差与局部异常值，因此先使用 LOF 算法对 ZHNZ 站训练样本集进行异常值探测。由 LOF 的原理可知，当数据点的 LOF 得分大于 1 时，该点可能为异常值，但由图 4 的密度曲线可知，坐标数据中存在大量得分大于 1 的数据点，因此将大于 1 的数据点都当作异常值是不合理的。本文实验最终选取k=7，LOF 阈值为 1.5，训练样本集中共检测到得分大于 1.5 的异常值共 71 个，在图 5 中用红色点表示，从图中可知，LOF 能够准确探测极端异常值，并能探测到大部分局部异常值，但仍会有误判数据。对于误判数据，仍选择进行剔除，这是因为通过高精度的插值模型能够削弱甚至抵消误判带来的影响。

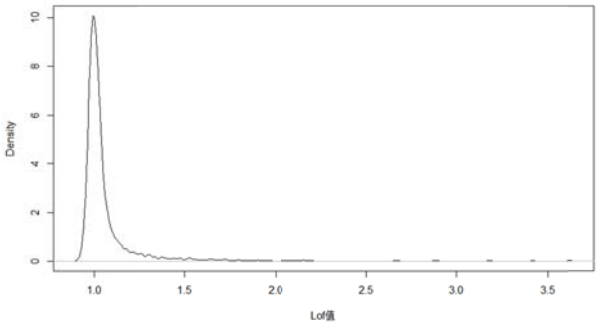

Figure 4. LOF density curve of ZHNZ station
图 4.ZHNZ 站 LOF 密度曲线

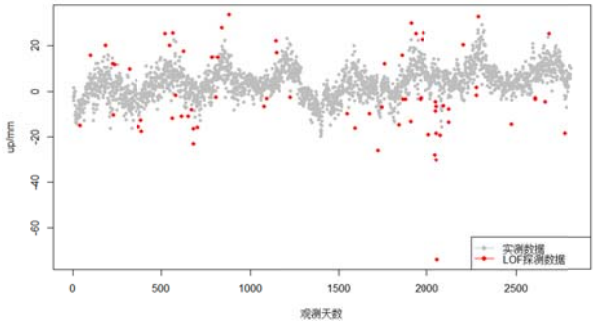

Figure 5. LOF anomaly detection at ZHNZ station
图 5.ZHNZ 站 LOF 异常值探测

3.2 Prophet插值

对 2.1 中探测到的异常值进行剔除后，数据中就产生部分缺失值，因此在预测前需要对其进行插值。Prophet 模型插值的本质是通过对缺失值进行预测从而达到插值的效果，即模型在拟合序列时自动填补空值从而补全数据。图 6 是 Prophet 模型对 ZHNZ 站缺失数据的插值效果图，蓝色点表示 Prophet 插值数据，从图中可以看出，Prophet 模型在处理缺失值时的鲁棒性较强，并且得到的插值数据能够较好的契合时间序列的变化规律。

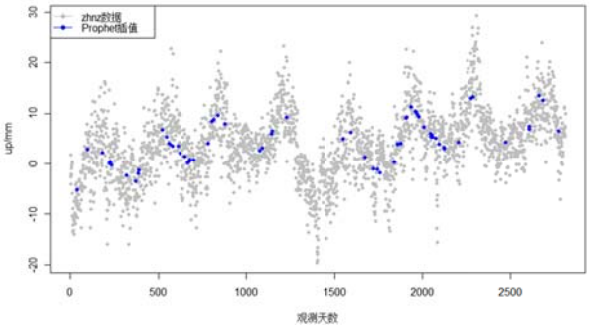

Figure 6. Prophet interpolation at ZHNZ station
图 6.ZHNZ 站 Prophet 插值

3.3 随机森林预测

对 ZHNZ 站原始时间序列 X 进行 LOF 异常值探测与 Prophet 模型插值处理后，得到新的时间序列 X1。分别对时间序列 X 与 X1 建立随机森林预测模型，得到未来 30 天的预测数据，并与测试样本集数据进行对比，从而分析异常值对预测精度的影响。实验中选用 1000 颗决策树，由图 7 可知，随机森林在对 GNSS 高程坐标时间序列的建模过程中，具有良好的抗过拟合能力，并能较好的扑捉到序列中的周期性，得到的拟合数据更加规律和平稳。

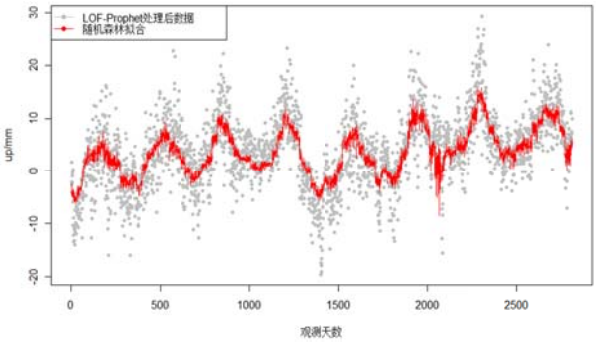

Figure 7. Random forest fitted sequence X1
图 7.随机森林拟合序列 X1

表 1 统计了随机森林模型对时间序列 x 与 X1 的预测精度，由于随机森林模型每次训练的精度有一定波动，为了保证实验结果的可靠性与准确性，

分别对序列 X 与 X1 进行十次预测实验，并计算平均值作为最终的预测精度。由表 1 可知，随机森林的预测精度十分稳定，在一定范围内进行变化；序列 X1 较 X 的预测精度有稳定的提升，MAE 与 RMSE 分别提高 3.9%与 2.5%，这表明异常值对预测精度有一定的负面影响，在预测前进行异常值探测与修正是必要的。

Table 1. Effect of outliers on the prediction accuracy of RF
表 1.异常值对随机森林预测精度的影响

实验次数	X		X1	
	MAE	RMSE	MAE	RMSE
1	2.310	3.208	2.228	3.127
2	2.352	3.234	2.213	3.142
3	2.307	3.205	2.225	3.121
4	2.337	3.236	2.212	3.110
5	2.333	3.231	2.232	3.140
6	2.319	3.217	2.226	3.163
7	2.337	3.211	2.243	3.144
8	2.286	3.183	2.234	3.147
9	2.331	3.231	2.230	3.158
10	2.327	3.219	2.231	3.131
平均值/mm	2.32	3.22	2.23	3.14

为了验证本文方法的有效性以及随机森林在 GNSS 高程坐标时间序列中的适用性，选用 ARMA 模型、线性回归（LM）、弹性网络（Elastic Net）作为对比方法进行实验。由表 2 可知，本文方法的预测精度最高，较传统方法有较大提升；且单一的 RF 模型也能保持较高的预测精度，表明 RF 方法对 GNSS 坐标时间序列的适用性强。图 8 是 LOF-Prophet-RF 预测数据与原始数据的对比，可以看出预测数据较原始数据的稳定性更好，没有产生较大的异常波动，且与原始数据一样呈现微弱的下降趋势，这表明本文方法能够准确的对坐标数据进行预测。

Table 2. Comparison of the accuracy of different models
表 2.不同模型的精度对比

模型	精度指标/mm	
	MAE	RMSE
LOF-Prophet-RF	2.23	3.14
RF	2.32	3.22
ARIMA(0,1,2)	2.38	3.42
LM	2.77	3.53
Elastic Net	2.86	3.61

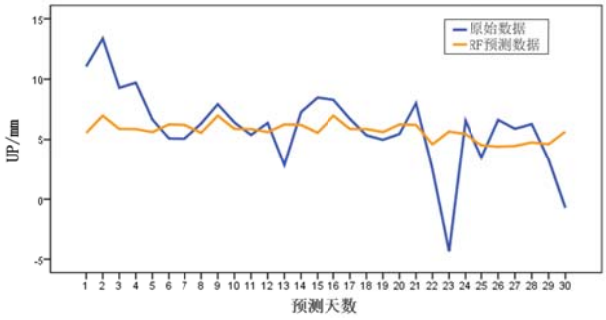

Figure 8. Random forest prediction
图 8.随机森林预测

4 结论

本文对 GNSS 高程坐标时间序列进行的预测实验，并顾及异常值对预测精度的影响，因此借助 LOF 算法对数据进行异常值探测，同时使用 Prophet 模型进行插值，对处理后的数据建立随机森林模型进行预测。本文以陆态网 ZHNZ 站实测高程数据为例进行实验，得到以下结论：

（1）LOF 算法能够准确的探测粗差及部分异常值，但仍可能出现误判的情况，以及阈值的选择仍具有较大主观性。因此对 LOF 算法在 GNSS 坐标时间序列中的适用性有待进一步研究，包括对异常值数量的合理选择，以及k值的选取对探测结果的影响分析。

（2）异常值对数据分析会产生一定的影响，因此对坐标数据中的异常值进行探测与修正是十分必要的，合适的插值方法也至关重要。由本文实验可知，进行异常值处理后的数据能够得到更高精度的预测数据。

（3）随机森林模型对 GNSS 坐标数据有良好的适用性，拟合数据能够体现出原始序列的周期与趋势，且预测精度较传统方法也有一定提升。在预测前对数据进行数据预处理可以在一定程度上提升预测精度。

References (参考文献)

[1] Qiu Xiaomeng，Wang Fengwei，Zhou Shijian，et al. Application of local mean decomposition and singular value decomposition in noise reduction of GNSS station coordinate time series signal[J]. Bulletin of Surveying and Mapping，2020(05)，85-89. 邱小梦,王奉伟,周世健,等.局部均值分解和奇异值分解在GNSS站坐标时间序列信号降噪中的应用[J].测绘通报,2020(05):85-89.

[2] Wei Shiyu，Li Chuan. The gross error detection for GNSS automatic monitoring data based on kalman filter[J]. The Chinese Journal of Geological Hazard and Control，2017，28(01)：146-150+155. 魏世玉,李川.基于卡尔曼滤波的GNSS 自动化监测数据粗差分析[J].中国地质灾害与防治学报,2017,28(01):146-150+155.

[3] Feng Shengtao，Liu Xuelong，Wang You. Least-Squares Fit Used for Primary Analysis of Position Time Series of GNSS[J]. Science of Surveying and Mapping，2015，40(10)：157-160. 冯胜涛,刘雪龙,王友.最小二乘拟合 GNSS 位置时间序列分析[J]. 测绘科学,2015,40(10):157-160.

[4] Zhang Mingmin，Liu Pan，Zhou Hailong，et al. Comparison and Analysis of the Accuracy of Two Elevation Coordinate Forecasting Models[J]. Engineering of Surveying and Mapping，2019，28(04)：13-18. 张明敏,刘盼,周海龙,等.两种高程坐标预测模型的精度对比分析[J].测绘工程,2019,28(04):13-18.

[5] Li Xia，Sun Maojun，Huang Yongsheng. Research on Application of LSTM Neural Network Model in GPS Deformation Monitoring[J]. Journal of Gansu Sciences，2019，31(3)：24-27. 李霞,孙茂军,黄永生.LSTM 神经网络模型在 GPS 变形监测中的应用研究[J].甘肃科学学报,2019,31(03):24-27

[6] Li Wei，Lu Tieding，He Xiaoxing，et al. Prediction and Analysis of Prophet-RF Model in GNSS Vertical Coordinate Time Series[J]. Journal of Geodesy and Geodynamics，2021,41（02）：116-121. 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（02）：116-121.

[7] Breunig M M , Kriegel H P , Ng R T , et al. LOF: Identifying Density-Based Local Outliers[C].Acm Sigmod International Conference on Management of Data. ACM, 2000.

[8] Dong Ze，Jia Hao，Outlier detection method for thermal process data based on EWT-LOF[J]. Chinese Journal of Scientific Instrument，2020，41(02)：126-134. 董泽,贾昊.基于 EWT-LOF 的热工过程数据异常值检测方法[J].仪器仪表学报,2020,41(02):126-134.

[9] Taylor S J，Letham B. Forecasting at Scale[J]. American Statistician，2017，72（1）：100-108

[10] Teng Jinling，Liu Pingzeng，Zhang Yan，et al. Research on ginger price forecast based on Prophet[J].Journal of Chinese Agricultural Mechanization，2020,41(08):211-216.滕金玲,柳平增,张艳,等.基于 Prophet 的生姜价格预测研究[J].中国农机化学报,2020,41(08):211-216.

[11] Fang Kuangnan，Wu Jianbin，Zhu Jianping，et al. A Review of Technologies on Random Forests[J]. Statistics &Information Forum，2011，26(03)：32-38.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(03):32-38

[12] Zhen Yiwei，Hao Min，Lu Baohong，et al. Research of Medium and Long Term Precipitation Forecasting Model Based on Random Forest[J]. Water Resources and Power，2015，33(06)：6-10. 甄亿位,郝敏,陆宝宏,等.基于随机森林的中长期降水量预测模型研究[J].水电能源科学,2015,33(06):6-10.

[13] Guo Zhankun，Jin Yongwei，Liang Xiaozhen，et al. Prediction Model of Port Container Throughput Based on Outlier Detection[J]. Mathematics in Practice and Theory，2019，49(17)：26-34. 郭战坤,金永威,梁小珍,等.基于异常值检测的港口集装箱吞吐量预测模型[J].数学的实践与认识,2019,49(17):26-34.