DOI:10.20040/j.cnki.1000-7709.2023.20220927

基于组合模型的月降水量预测研究

程桂芳,王雪敏

(郑州大学数学与统计学院,河南 郑州 450001)

摘要:近年来,由降水量过多或过少引起的灾害日益增加,因此准确地预测降水量对人类的生活和社会的发展具有重大意义和实际应用价值。基于郑州市1990~2019年的月降水量数据,分别利用SARIMA、Prophet、LSTM单一模型对郑州市2020~2021年的月降水量进行预测。为了提高月降水量的预测精度,提出SARIMA-EMD-LSTM、Prophet-EMD-LSTM两种组合模型,实证表明这两种组合模型预测精度更高,均方根误差显著减少,其中Prophet-EMD-LSTM组合模型的预测效果相对较优。最后利用该模型对郑州市2022年4~12月的月降水量进行了预测,精度较高。

关键词:降水量预测;SARIMA;Prophet;EMD;LSTM;组合模型

中图分类号:TV124 **文献标志码**:A **文章编号**:1000-7709(2023)04-0013-04

1 引言

近年来,随着人类社会的发展以及生存空间的不断扩张,人类对大自然的过度开采和利用导致自然环境逐渐恶劣,水资源环境也受到很大影响。水资源包括江河、井、湖泊、降水量、地下水等,其中降水量是水文研究过程中关注的重要因素之一,降水量异常给人们的日常生活、社会生产活动造成很大影响。降水量过多或过少都会对社会生产活动造成负面影响。因此,合理准确地预测某一地区月降水量非常重要。目前用于预测时间序列数据的模型主要有统计学模型[1,2]、机器学习和深度学习[3-6]模型。其中统计学模型仅能较好地提取数据中的线性特征,不能有效地处理降水量数据中的非线性变化;深度学习模型在处理非线性信息方面具有很大的优势。因此,本文采用统计学模型与深度学习模型相结合的方式分析月降水量数据,以提高模型的预测精度。实证表明该方法有效可行。

2 研究数据与方法

2.1 数据来源与处理

基于郑州市1990年1月至2022年3月的月降水量数据(数据来源于国家气象信息中心)进行分析,将每日的降水量数据处理为月降水量数据,保留精度为0.1 mm。由于2021年郑州市"7·20"特大暴雨事件为极端异常降水情况,导致7月的月降水量高达902.4 mm。"7·20"事件是郑州市有气象记录以来发生的一次极端情况,如果使用该数据进行分析及预测,则会由于降水量数据波动极其异常,导致分析陷入歧途。因此,在预测2022年降水量时,将其视为缺失值处理,并用后推法填补数据。

此外,将数据集划分为两部分:①1990年1月至2019年12月的月降水量数据作为训练集;②2020年1月至2021年12月的月降水量数据作为测试集。

2.2 研究方法

2.2.1 SARIMA模型

季节性差分自回归移动平均模型(SARIMA)是ARIMA乘法模型中的一种[1],通常记为SARIMA$(p,d,q)(P,D,Q)S$,其中p为季节性自回归的阶数;q为季节性移动平均项的阶数;P为季节性自回归的最大阶数;Q为季节性移动平均项的最大阶数,具体的结构形式为:

$$\phi(B)\phi(B^S)(1-B)^d(1-B^S)^D=\theta(B)\theta(B^S)\varepsilon_t \quad (1)$$

其中 $$\phi(B)=1-\phi_1B-\phi_pB^p$$

$$\theta(B)=1-\theta_1B-\theta_qB^q$$

收稿日期:2022-04-05,**修回日期**:2022-05-27

基金项目:国家自然科学基金项目(11971444);2021年度河南省高等教育教学改革研究与实践重点项目(2021SJGLX060);2022年河南省医学科技攻关联合共建项目(LHGJ20220518);2022年度河南省社科联调研课题项目(SKL-2022-405)

作者简介:程桂芳(1979-),女,博士、副教授、硕导,研究方向为统计与大数据分析,E-mail:gfcheng@zzu.edu.cn

$$\phi(B^S)=1-\phi_1 B^S-\phi_p B^{pS}$$

$$\theta(B^S)=1-\theta_1 B^S-\theta_q B^{qS}$$

式中，B 为延迟算子；$\phi(B)$、$\theta(B)$ 分别为非季节的自回归项系数、滑动平均项系数的多项式；$\phi(B^S)$、$\theta(B^S)$ 分别为季节性自回归项系数、滑动平均项系数的多项式；S 为季节性周期的长度；d 为逐期差分的步数；D 为季节性差分的步数。

2.2.2 Prophet 模型

Prophet 模型本质上是基于时间序列可分解思想的自加性模型，根据时间序列$\{Y_t\}$中可能存在的影响因子，将其分解为 4 个核心因子。模型的结构形式[2]为：

$$y(t)=g(t)+s(t)+h(t)+\varepsilon(t) \quad (2)$$

式中，$g(t)$ 为时间序列去除周期效应后的大概趋势，记为趋势项；$s(t)$ 为时间序列周期性变动或季节性变动的规律，记为周期项；$h(t)$ 为时间序列中包含的某些周期不固定的节假日引起的波动，记为节假日项；$\varepsilon(t)$ 为时间序列中不能被模型解释的随机波动，记为误差项或剩余项。

2.2.3 经验模态分解

经验模态分解(EMD)可将一个高频的复杂时间序列分解成若干个信号分量的线性和，包括本征模函数(IMF)和趋势项(R_{es})，其分解的表达式[5,6]为：

$$Y_t=\sum_{i=1}^{n}\mathrm{IMF}_i(t)+R_n(t) \quad (3)$$

式中，Y_t 为原始的时间序列数据；$\mathrm{IMF}_i(t)$ 为分解出的第 i 个信号分量；$R_n(t)$ 为趋势项；t 为时间；n 为分解的信号分量的个数。

2.2.4 长短时记忆神经网络模型

长短时记忆(LSTM)[4]神经网络(图 1)由 1 个输入层、1 个输出层及 1 个或多个隐藏层构成，通过设计“门”结构在细胞状态中遗忘、添加或筛选信息。

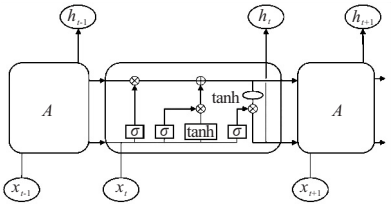

图 1 长短时记忆神经网络单元结构

Fig. 1 Short-term memory neural network unit structure

3 基于组合模型的月降水量预测研究

3.1 组合模型试验步骤

由于月降水量数据具有明显的季节效应，统计学模型能较好地提取季节效应信息及原始数据中的线性特征，并具有参数较少、模型简单的优势，而深度学习模型在函数逼近方面有很大优势，能很好地挖掘原始数据中的非线性信息。因此，本文考虑综合统计学模型和深度学习模型，即分别拟合 SARIMA-EMD-LSTM、Prophet-EMD-LSTM 组合模型对郑州市的月降水量序列进行分析与预测。步骤如下。

步骤 1 基于训练集的数据分别拟合 SARIMA、Prophet 模型并进行预测，得到统计学模型的拟合值 $\hat{y}$ 及预测值 $\hat{y}_1$，然后用原始数据 y 减去拟合值得到模型残差 ε，即：

$$\varepsilon=y-\hat{y} \quad (4)$$

步骤 2 利用 EMD 对模型残差进行分解，得到信号分量 IMF_1，IMF_2，…，IMF_k 和剩余项 R_{es}。

步骤 3 对各分量进行重构。采用零均值假设检验针对分量的统计性质进行重构，即基于零均值假设对各分量进行单样本检验，通过检验统计量或 P 值的大小，判断是否拒绝原假设。若不能拒绝零假设，则认为本征模函数分量属于高频项；若拒绝零假设，则认为本征模函数分量属于低频项。然后根据本征模函数分量的正交性及独立性性质，直接通过线性相加的方法组合同一类的本征模函数分量，最终将分解出的信号分量重构为高频项 H_{IMF}、低频项 L_{IMF} 和趋势项 R_{es}。

步骤 4 利用归一化方法对重构后的分量及残差进行标准化，然后基于 LSTM 对残差进行拟合与预测，得到 $\hat{\varepsilon}$ 与 $\hat{\varepsilon}_1$。

步骤 5 将残差预测值与统计学模型预测值相加，得到组合模型的最终预测值 $\tilde{y}$，即：

$$\tilde{y}=\hat{y}_1+\hat{\varepsilon}_1 \quad (5)$$

步骤 6 利用单一模型对郑州市的月降水量数据进行拟合与预测，并以均方根误差(R_{RMSE})、均方误差(M_{MSE})为标准，比较不同模型的预测精度，然后利用预测精度较好的模型预测 2022 年 4～12 月的郑州市月降水量。

实证分析流程见图 2。

3.2 组合模型预测结果分析

先分别利用 SARIMA 和 Prophet 模型基于 1990 年 1 月至 2019 年 12 月的月降水量数据拟合模型，预测 2020 年 1 月至 2021 年 12 月的月降水量。其中 SARIMA 模型的预测结果见图 3(a)。实证表明该模型能较好地拟合出郑州市月降水量季节性变化的趋势，但将拟合值与原始数据相比，发现模型针对具体值的拟合效果较差。

Prophet 模型的预测结果见图 3(b)。深色圆点表示郑州市真实的月降水量值，红色粗线条表

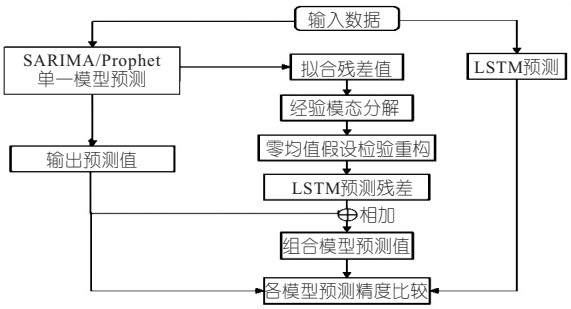

图 2　实证分析流程

Fig. 2　Empirical analysis of the process

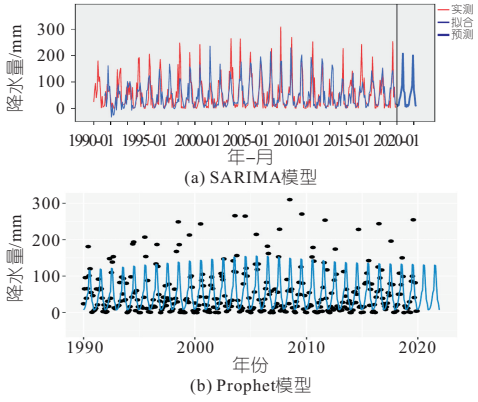

图 3　SARIMA 模型和 Prophet 模型拟合与预测图

Fig. 3　SARIMA model and Prophet model fitting and forecasting graph

示 Prophet 模型拟合的月降水量值,蓝色细线覆盖的范围为模型预测值的置信区间,最后两个周期为模型的预测值。Prophet 模型能准确地拟合出模型中的季节趋势,并可较为精确地拟合一些极小值点及变化幅度较小的中间值点,但模型对一些突变幅度很大的极大值点的拟合精度较差,拟合值通常低于真实值。

利用 EMD 分别将 SARIMA、Prophet 模型的残差分解成 6 个本征模函数(IMF)和 1 个趋势项(R_{es}),并基于零均值假设检验的结果,把重构后的高频项、低频项及剩余项设置为 LSTM 的输入项,与之对应的 SARIMA 模型残差设置为输出项,拟合出模型并进行分析。由于各分量和残差的水平、量级不同,需先使用归一化方法对数据进行预处理。

根据输入层的变量个数,依次选取 12～48 作为模型的输入步长,将每 20 个样本数据组成一个批量进行训练,迭代次数选取 500 次,并根据网格搜索和交叉验证方法依次为不同步长的模型搜索最佳神经元个数。最终利用在训练集拟合的 LSTM 模型进行预测,得到各模型残差的预测值,并通过线性相加的方式得到月降水量的最终预测值,见图 4。由图 4 可看出,模型对 2020 年数据的预测结果较好,波峰处偏差较小,未出现过

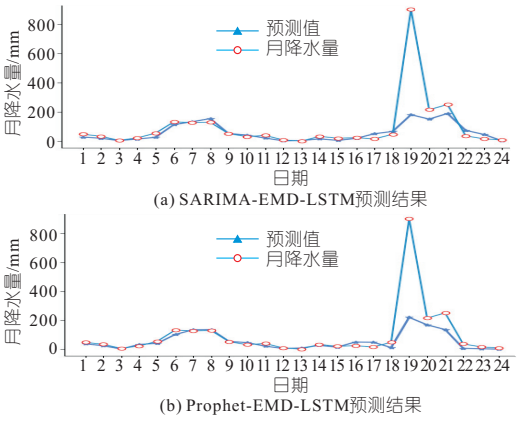

图 4　SARIMA-EMD-LSTM 和 Prophet-EMD-LSTM 预测结果拟合图

Fig. 4　SARIMA-EMD-LSTM and Prophet-EMD-LSTM forecast results matching graph

高或过低的情况,对极大值和极小值的预测较为准确,但随着时间的推移,模型预测的步数增加,对于月降水量的预测精度变差,对于极其异常的 2021 年 7 月份的预测,预测值偏差极大。

为了进一步比较单一模型与组合模型的预测精度差异,基于原始数据拟合 LSTM 模型并进行预测。分别选取 12 的倍数作为不同的滑动窗口长度 L 输入到网络中,设置模型的迭代次数为 1 000 次,加入 Dropout 项避免出现过拟合现象。预测结果见图 5。由图 5 可知,该模型对非极值点的月降水量预测较为准确,但对于极大值点 2020 年 6、8 月的预测有较大偏差,对 2021 年 7 月的预测值差别也极大。

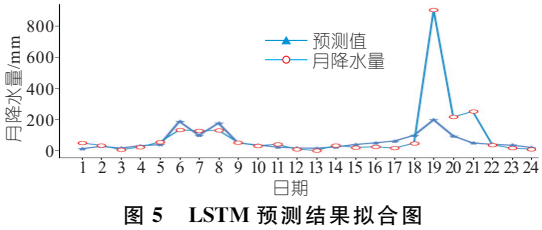

图 5　LSTM 预测结果拟合图

Fig. 5　LSTM forecast results matching graph

3.3　模型对比分析

根据各模型实例分析结果分别算出均方误差(M_{MSE})、均方根误差(R_{RMSE})(表 1),对比分析发现模型的预测精度从高到低依次为 Prophet-EMD-LSTM、SARIMA-EMD-LSTM、LSTM、Prophet、SARIMA。

表 1　模型预测效果对比

Tab. 1　Comparison of model prediction results

模型	M_{MSE}	R_{RMSE}
SARIMA	29 812.80	172.66
Prophet	26 851.53	163.86
LSTM	23 701.96	153.95
SARIMA-EMD-LSTM	22 331.19	149.44
Prophet-EMD-LSTM	20 311.03	142.52

与单一模型相比，组合模型结合深度学习和统计学模型，整体增强了模型的预测能力。SARIMA-EMD-LSTM 组合模型的均方根误差比 SARIMA 模型降低了 23.22 mm，比 LSTM 模型降低了 4.51 mm；Prophet-EMD-LSTM 组合模型比 Prophet 模型降低了 21.34 mm，比 LSTM 模型降低了 11.43 mm。因此，所提出的组合模型显著提高了月降水量的预测精度，其中 Prophet-EMD-LSTM 组合模型的预测精度最高。

利用预测精度最高的 Prophet-EMD-LSTM 组合模型预测郑州市 2022 年 4～12 月的月降水量，并利用 2011～2020 年登封、嵩山、新密地区的月均降水量数据得到对应 3 个区域的月降水量预测值进行对比分析，结果见表 2。

表 2 2022 年 4～12 月月降水量预测结果

Tab. 2 Forecast of monthly precipitation from April 2022 to December 2022 mm

地区	月份								
	4	5	6	7	8	9	10	11	12
郑州	16.8	29.9	81.2	211.8	149.3	79.8	29.6	37.4	7.9
登封	14.8	28.4	73.3	248.3	116.8	74.3	27.8	14.8	28.4
嵩山	20.9	37.9	97.6	254.4	158.5	103.2	40.0	20.9	37.9
新密	17.3	27.4	74.9	249.8	161.6	76.2	30.0	17.3	27.4

将 Prophet-EMD-LSTM 组合模型预测的郑州市未来 9 个月的月降水量与 2022 年 1～3 月的月降水量真实值相加，得到预测的 2022 年的年降水量为 687.9 mm。基于历史数据计算得到 1990～2020 年的郑州市年平均降水量为 636.44 mm，由此可得 2022 年年降水量的距平百分率为 8.1%。

4 结论

a. 基于郑州市 1990～2019 年的月降水量数据分别拟合 SARIMA、Prophet、LSTM 单一模型，预测了 2020～2021 年的月降水量、并以均方根误差、均方误差为标准对比单一模型与组合模型的预测效果，结果表明所提组合模型的拟合误差、预测误差均低于单一模型，其中 Prophet-EMD-LSTM 组合模型的预测精度最高。

b. 组合模型运用模态经验分解能较好地提取残差中时间尺度上的特征，且零均值假设检验能够保证在尽可能不损失信息的前提下降低模型的复杂程度。

参考文献：

[1] SANI A S, AUWAL A M, ADENOMON M O. Application of sarima models in modelling and forecasting monthly rainfall in nigeria[J]. Asian journal of probability and statistics, 2021(6):30-43.

[2] 吴文培，宋亚林，魏上斐. 基于改进 Prophet 模型的用电量预测研究[J]. 计算机仿真，2021，38(11)：473-478.

[3] 贺玉琪，王栋，王远坤. BRR-SVR 月降水量预测优化模型[J]. 水利学报，2019，50(12)：1529-1537.

[4] 刘新，赵宁，郭金运，等. 基于 LSTM 神经网络的青藏高原月降水量预测[J]. 地球信息科学学报，2020，22(8)：1617-1629.

[5] 李栋，薛惠锋，张燕. 基于经验模态分解的降水量组合预测模型[J]. 计算机仿真，2019，36(3)：458-463.

[6] 罗那那，巴特尔·巴克，吴燕锋. 基于集合经验模态分解北疆降水多尺度变化特征[J]. 水土保持研究，2017，24(4)：362-367.

Prediction of Monthly Precipitation Based on Combined Model

CHENG Gui-fang, WANG Xue-min

(School of Mathematics and Statistics, Zhengzhou University, Zhengzhou 450001, China)

Abstract: In recent years, it causes disasters increasingly by too much or too little precipitation. Therefore, accurate prediction of precipitation is of great significance and practical application value to human life and social development. Based on the monthly precipitation data of Zhengzhou from 1990 to 2019, monthly precipitation was forecasted from 2020 to 2021 by utilizing SARIMA, Prophet and LSTM model, respectively. In order to improve the prediction accuracy of the model for monthly precipitation, two combined models of the SARIMA-EMD-LSTM and Prophet-EMD-LSTM were proposed. Empirical analysis shows that the proposed two combined models have higher prediction accuracy and decrease the root mean square error significantly. Furthermore, Prophet-EMD-LSTM model has comparatively better prediction effect. The monthly precipitations in Zhengzhou from April to December, 2022 were forecasted with higher precision.

Key words: precipitation forecasting; SARIMA; prophet; EMD; LSTM; combined model