北京林业大学学报
JOURNAL OF BEIJING FORESTRY UNIVERSITY

基于3种时间序列模型的北京市每日花粉浓度预测

张鑫 杨华 董玲玲 张宏远

Prediction of daily pollen concentration in Beijing based on three time series models

Zhang Xin, Yang Hua, Dong Lingling, Zhang Hongyuan

引用本文:
张鑫, 杨华, 董玲玲, 张宏远. 基于3种时间序列模型的北京市每日花粉浓度预测[J]. 北京林业大学学报, 2025, 47(6):90–100. doi: 10.12171/j.1000–1522.20230232
Zhang Xin, Yang Hua, Dong Lingling, Zhang Hongyuan. Prediction of daily pollen concentration in Beijing based on three time series models[J]. Journal of Beijing Forestry University, 2025, 47(6):90–100. doi: 10.12171/j.1000–1522.20230232

在线阅读 View online: https://doi.org/10.12171/j.1000–1522.20230232

您可能感兴趣的其他文章
Articles you may be interested in

西双版纳热带季节雨林辐射长期变化特征
Long–term variation characteristics of radiation in the tropical seasonal rainforest in Xishuangbanna, southwestern China
北京林业大学学报, 2021, 43(4): 56–67. https://doi.org/10.12171/j.1000–1522.20200270

长期施氮和降水减少对长白山阔叶红松林凋落物量的影响
Effects of long–term nitrogen addition and precipitation decreasing on the litterfall production of broadleaved Korean pine forest in Changbai Mountains of northeastern China.
北京林业大学学报, 2017, 39(4): 29–37. https://doi.org/10.13332/j.1000–1522.20160171

造林与间伐对东北温带弃耕地土壤温室气体排放的长期影响
Long–term effects of afforestation and thinning on greenhouse gas emissions from temperate abandoned–land soil in the Northeast of China
北京林业大学学报, 2017, 39(10): 13–23. https://doi.org/10.13332/j.1000–1522.20170017

北京地区8种树种枯死可燃物含水率预测模型及变化规律
Prediction models and changing rules of dead fuel moisture content of 8 forest species in Beijing area
北京林业大学学报, 2020, 42(6): 80–90. https://doi.org/10.12171/j.1000–1522.20190370

基于机器学习的落叶松毛虫发生面积预测模型
Studies on prediction models of *Dendrolimus superans* occurrence area based on machine learning
北京林业大学学报, 2017, 39(1): 85–93. https://doi.org/10.13332/j.1000–1522.20160205

基于积温的文冠果开花物候期预测模型的构建
Establishment of the flowering phenological model of *Xanthoceras sorbifolium* based on accumulated temperature
北京林业大学学报, 2019, 41(6): 62–74. https://doi.org/10.13332/j.1000–1522.20180128

DOI:10.12171/j.1000−1522.20230232

基于 3 种时间序列模型的北京市每日花粉浓度预测

张 鑫[1] 杨 华[1] 董玲玲[1] 张宏远[2]
（1. 北京林业大学林木资源高效生产全国重点实验室，北京 100083；2. 北京市十三陵林场管理处，北京 102200）

摘要:【目的】分析花粉高峰期持续时间和浓度峰值，构建北京市每日花粉浓度的最优预测模型，为科学预测未来每日花粉浓度提供数据支持。**【方法】**采用多重插补法处理 2015—2020 年北京市每日花粉浓度时间序列中的缺失数据，2015—2019 年数据用于建立 SARIMA、LSTM 和 Prophet 3 种时间序列模型，预测未来一年（2020 年，共计 182 d）的花粉浓度变化。**【结果】**（1）随机森林法、贝叶斯线性回归法、观测值中随机取样法和加权预测均值匹配法 4 种多重插补法中，随机森林法的第 3 个插补数据集 P 值最小（$P = 0.002$），为最优插补数据集。（2）2015—2020 年每日平均花粉浓度数据显示，春季高峰期集中在 3—6 月，4 月初达到峰值（792 粒/(10^3 mm^2)）；秋季高峰期集中在 8 月至 9 月末，在 9 月初达到峰值（449 粒/(10^3 mm^2)）。2015—2019 年花粉浓度总体呈逐年下降趋势，2020 年呈现阶跃式上升；其中，2015 年高峰期持续时间最长（春季 107 d，秋季 65 d），2018 年最短（春季 60 d，秋季 46 d）；2020 年花粉浓度峰值达到最高水平，而 2019 年花粉浓度峰值最低。（3）3 种时间序列模型中，LSTM 模型对北京市每日花粉浓度时间序列的描述和预测效果最佳。当 LSTM 模型的时间步长（look_back）为 60 时，模型预测效果最佳，RMSE、MAE 均为最小，$R^2 = 0.78$。相比之下，Prophet 模型效果较差，无法灵敏捕捉浓度峰值，预测值存在负数情况，预测效果不佳。SARIMA 模型拟合效果尚可，但预测效果不理想，预测值存在为负的情况。**【结论】**与 SARIMA 和 Prophet 模型相比，LSTM 模型更适用于北京市每日花粉浓度时间序列模型的建立与长期预测。未来研究应完善花粉浓度数据，优化模型性能，以更准确地预测花粉高峰期的起止时间、持续时间及高峰浓度，为过敏性疾病的防控提供更可靠的依据。

关键词：多重插补法；花粉浓度；长短期记忆神经网络；长期预测
中图分类号：S685　　**文献标志码：**A　　**文章编号：**1000−1522(2025)06−0090−11
引文格式：张鑫，杨华，董玲玲，等. 基于 3 种时间序列模型的北京市每日花粉浓度预测［J］. 北京林业大学学报，2025，47(6)：90−100. Zhang Xin, Yang Hua, Dong Lingling, et al. Prediction of daily pollen concentration in Beijing based on three time series models［J］. Journal of Beijing Forestry University, 2025, 47(6): 90−100.

Prediction of daily pollen concentration in Beijing based on three time series models

Zhang Xin[1] Yang Hua[1] Dong Lingling[1] Zhang Hongyuan[2]
(1. State Key Laboratory of Efficient Production of Forest Resources, Beijing Forestry University, Beijing 100083, China;
2. Beijing Shisanling Forest Farm Management Office, Beijing 102200, China)

Abstract:［**Objective**］The duration and concentration peak of pollen were analyzed, and the optimal prediction models of daily pollen concentration in Beijing were built to provide data support for scientific prediction of future daily pollen concentration.［**Method**］Based on the time series data of daily pollen concentration in Beijing from 2015 to 2020, multiple interpolation methods were used to handle missing data, and three time series models (SARIMA, LSTM and Prophet) were used to predict pollen concentration

收稿日期: 2023−09−11　**修回日期:** 2024−04−01
基金项目: 国家林草局应急揭榜挂帅项目（202303）。
第一作者: 张鑫。主要研究方向：森林资源监测技术。Email: 18810961930@163.com　地址：100083 北京市海淀区清华东路 35 号北京林业大学林学院。
责任作者: 杨华，教授。主要研究方向：森林资源监测技术。Email: huayang8747@bjfu.edu.cn　地址：同上。董玲玲，博士，高级实验师。主要研究方向：森林生态。Email: donglingling1120@163.com　地址：同上。
本刊网址: http://j.bjfu.edu.cn; http://journal.bjfu.edu.cn

changes in the next year (year 2020, totaling 182 d). [**Result**] (1) Among the four multiple interpolation methods, including random forest imputation, Bayesian linear regression imputation, simple random sampling imputation, and predictive mean matching with distance-aided donor selection, the third imputed dataset from the random forest method had the smallest value (P = 0.002), which was the best imputation dataset. (2) Average daily pollen concentration data from 2015 to 2020 showed that the spring peak occurred between March and June, reaching peak in early April (792 grain/(10^3 mm^2)); the autumn peak occurred between August and the end of September, reaching peak in early September (449 grain/(10^3 mm^2)). From 2015 to 2019, the overall pollen concentration showed a decreasing trend, while there was a step upward trend in 2020. The peak duration was the longest in 2015 (107 d in spring, 65 d in autumn), while it was the shortest in 2018 (60 d in spring, 46 d in autumn). The peak value of pollen concentration was the highest in 2020, while the peak value was the lowest in 2019. (3) Among the three models, the LSTM model was the best in describing and predicting time series of daily pollen concentration in Beijing. When the time step (look_back) of the LSTM model was 60, the model provided the best prediction, with the smallest RMSE and MAE, and R^2 of 0.78. In contrast, the Prophet model performed poorly, unable to capture the peak concentration sensitively, with predicted values sometimes negative. The SARIMA model had a reasonable fit but did not perform well in prediction, with some negative predicted values. [**Conclusion**] Compared with SARIMA and Prophet models, the LSTM model is more suitable for establishing and long-term predicting time series model of daily pollen concentration in Beijing. Future research should focus on improving the pollen concentration data and optimizing model performance to more accurately predict the starting and ending times, duration, and peak concentration of pollen periods.

Key words: multiple interpolation method; pollen concentration; long and short-term memory neural network; long-term forecast

随着城市绿地面积的增加，空气中的花粉颗粒浓度升高，花粉浓度的监测工作变得愈发重要。目前，北京市[1−2]、承德市[3]和呼和浩特[4]等多个城市，已开展花粉监测工作。花粉监测数据主要用于分析花粉浓度的季节性变化规律[3]、花粉种类变化与分布现状[1−2,5]。随着监测数据的积累与技术的进步，花粉浓度模型逐步发展，为花粉浓度的预测预报提供了理论基础。前人研究[6−8]大部分以气候因子作为自变量，部分研究中引入叶面积指数（LAI）[6]、增强型植被指数（EVI）[9]等变量，基于花粉浓度建立了广义加性模型（GAMs）[9]、机器学习模型[6−7]和时空模型[8]，主要用于短期预测。然而，Seo 等[10]基于深度神经网络（DNN），同样添加气候因子后进行了未来两年花粉浓度的长期预测。

每日花粉浓度数据是长期、具有季节性的时间序列数据。时间序列模型在数据分析与预测中已被广泛应用。常见模型包括季节自回归差分移动平均模型（seasonal autoregressive integrated moving average，SARIMA 模型）、长短期记忆神经网络模型（long short-term memory，LSTM 模型）以及 Prophet 模型。SARIMA 模型适用于周期性时间序列的预测，可通过对模型参数的量化体现数据的变化趋势，反映每年同一时间的季节性变化规律；Milenković 等[11]构建的铁路月客流时间序列的 SARIMA 模型表现出了良好的预测性能。LSTM 模型是一种特殊的 RNN 模型（递归神经网络），在 RNN 每个重复模块链的基础上增加了遗忘门、输入门和输出门，以控制信息的输入、输出和保留，克服了 RNN 权重影响过大、容易产生梯度消失和爆炸的缺点[12]；与 RNN 由于梯度消失只能进行短期记忆相比，LSTM 可以结合短期记忆与长期记忆，在长期预测方面的效果更优。LSTM 模型的论文首次发表于 1997 年[13]，收敛快，训练性能出色，预测精度较高，常用于对长时间序列数据进行学习预测；Zhao 等[14]构建的 6 种短期交通量预测模型中，LSTM 模型在预测时间较长时效果最优；Chimmula 等[15]和任树顺等[12]构建的 LSTM 模型对数据短期以及长期的预测均取得了较优的效果。Prophet 模型是 Facebook 于 2017 年发布开源的时间序列预测框架，擅长处理具有季节性特征的大规模时间序列数据，可处理异常值情况，预测速度快，泛用性高，尚未被广泛应用[16]；张骏等[17]运用 Prophet 模型可较精准地对 5 种污染物浓度进行预测。Toharudin 等[18]气温长时间序列模型的建立结果表明，Prophet 和 LSTM 模型 RMSE 的值差异不显著，表现接近且均较好。综上，可尝试将 SARIMA、LSTM 和 Prophet 3 种时间序列模型应用于每日花粉

浓度数据的长期预测。

本研究以北京市每日花粉浓度为研究对象，主要针对完备化花粉浓度数据集，花粉浓度高峰期和峰值特征，SARIMA、LSTM、Prophet 3 种时间序列模型对花粉浓度长期预测的可行性的研究问题开展研究。不同于前人通常基于温度等气候因子构建的预测短期花粉浓度的多元模型，本研究不添加气候因子，从花粉浓度的季节性入手，构建能够科学预测长期(2020 年，共计 182 d)花粉浓度的时间序列模型。通过该试验寻找最优花粉浓度时间序列模型，预测未来一年花粉浓度的变化趋势，有利于城市根据预测结果提前在宏观层面发布和制定控制空气中花粉浓度的政策与措施。

1 研究区概况

北京市地处华北平原西北端(115°20′～117°30′E，39°28′～41°05′N)，总面积为 16 410 km^2，其中平原区面积为 6 200 km^2，约占总面积的 38%。属于暖温带半湿润半干旱季风气候，春秋短暂，冬季寒冷干燥，夏季高温多雨，其中 80% 的降雨集中在 6—8 月，大雨通常发生于 7、8 月份。植被类型以暖温带落叶阔叶林为主，并间有温性针叶林分布。近年来，北京市实施了两轮百万亩造林绿化、生态涵养区建设等一系列生态工程，森林覆盖率已达 44.8%，城市绿化覆盖率达到 49.3%，野生植物种类多达 2 088 种。

北京市的花粉季节通常从 2 月份下旬开始，持续至 10 月份上旬。李挚等[19]基于 2010—2013 年花粉监测数据研究发现，北京市常见的致敏花粉来源包括蒿属(*Artemisia*)、葎草属(*Humulus*)、悬铃木属(*Platanus*)、藜科(Chenopodiaceae)和柏科(Cupressaceae)等。目前，北京市春季常见致敏花粉为榆科(Ulmaceae)、柏科、松科(Pinaceae)以及杨柳科(Salicaceae)花粉，而夏秋季则以藜科、蒿属和桑科(Moraceae)花粉为主。

2 材料与方法

2.1 数据来源与数据预处理

研究数据集为北京市城区 2015—2020 年 4 月 1 日—9 月 29 日空气中每日花粉浓度数据，来源于“气象北京”微博。花粉采集采用重力沉降法，曝片时间为 24 h。采集样本经染色处理后，由专业人员读取花粉数量并分类，据此计算花粉浓度。为处理数据缺失，采用多重插补法，使用随机森林法(random forests)、贝叶斯线性回归法(Bayesian linear regression)、观测值中随机取样法 simple random sampling)和加权预测均值匹配法(Imputation by predictive mean matching with distance aided donor selection)分别填补数据集缺失值。每种方法均设置 5 个插补数据集，并基于广义线性模型拟合[20]，比较统计学检验结果的 P 值，取 P 值最小，结合插补数据集密度图以及插补数据与监测数据的分布图，确定最优插补数据集。依据李俊等[21]提出的空气花粉浓度对人群变应性鼻炎致敏基准剂量的 95% 可信下限值(21.38 粒/ 10^3 mm^2)，本研究取 21.38 粒/ 10^3 mm^2 为门限值，作为识别花粉浓度高峰期开始与结束时间的判断依据。

2.2 3 种时间序列模型的建立

将数据集分为训练集和测试集，其中，2015—2019 年数据作为训练集，用于模型建立与拟合效果分析；2020 年数据作为测试集，利用训练集建立的模型对 2020 年的花粉浓度进行预测后生成预测值，进行预测效果分析。数据共 1 092 个，其中训练集 910 个，测试集 182 个。

2.2.1 SARIMA 模型

SARIMA 模型为季节性时间序列模型，适用于具有季节性、趋势性和周期性的平稳数据序列，表达式为$(p, d, q)(P, D, Q)_s$。其中 s 为时间序列的变化周期，取值一般为 12[22]；p 为自回归阶数；d 为一般差分阶数；q 为移动平均阶数；P 为季节性自回归阶数；D 为季节性差分阶数；Q 为季节性移动平均阶数。建立 SARIMA 模型首先检验数据平稳性，人工定阶用 R 语言的 ndiffs( )函数确定一般差分阶数 d[23]，再进行季节性差分，通过相关图与偏相关图确定季节性差分阶数 D 以及 p、q、P、Q 的定阶[16]。

2.2.2 LSTM 模型

LSTM 网络在神经网络层级之间存在记忆模块，可对时间序列中的前后部分产生记忆；模块中包括由取舍信息的遗忘门和输入门以及决定输出值的输出门构成的激活单元，各个门的权重通过训练获得；该模块使 LSTM 网络能够克服梯度消失的问题，并用于处理复杂的时间序列问题。

LSTM 网络对数据的“标度”较为敏感，需先将数据缩放至 0~1 之间再进行模型建立，应先对数据进行标准化处理以提高模型精度。通过控制 keras 包中的 lstm 层和 dense 层构建 LSTM 模型，lstm 层具有长短期记忆功能，可防止较早期的信号在处理过程中逐渐消失，lstm 层的 activation(激活)设置为“tanh”(双曲正切)，recurrent_activation(用于循环时间步的激活函数)为“sigmoid”；dense 层为全连接层，将 lstm 层的输出转换为目标输出。lstm 层和 dense 层均需设置 units 参数，该参数代表单元内隐藏层尺寸。在确定 units、epochs(迭代次数)、batch_size

（每个训练批次样本数）时采用试错法。本研究在进行 LSTM 模型的预测时采取了滑动窗口的模式，要求输入一个表示要用多少个过去值预测下一个值的额外参数数值——look_back，look_back 取值根据数据特点与实际情景确定，根据植物开花规律设置 14、30、60、90 进行比较。

2.2.3　Prophet 模型

Prophet 模型是基于加法模型的时间分解模型，适用于具有强烈季节性影响的时间序列，且在处理具有缺失值、异常值以及趋势变化的数据时具有鲁棒性[17]。Prophet 模型构建方式如公式（1）所示。

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \tag{1}$$

式中：t表示时间序列中当前的时间点，用于计算函数的具体值；$g(t)$为趋势函数，分别为分段性函数和非线性逻辑回归函数，拟合非周期变化；$s(t)$为周期性变化，包括每周、每年等，使用傅里叶级数建模每年的季节分量；$h(t)$为假期变化，将来不同的节假日为 0，属于节假日为 1；ϵ_t为表示随机无法预测的波动的噪声项。

Prophet 模型在 R 软件中的构建需控制参数以期获得更好的预测效果，changepoint.range 表示对数据的前多少比例推断变化点，以便有足够的时间片段来预测未来趋势并避免过拟合；changepoint.prior.scale 可调整稀疏先验的强度，值越大趋势灵敏度越高；holidays 表示建模的假期或其他特殊事件，由于本研究不涉及气候因子等影响因素，未设置参数 holidays；季节性参数中的傅里叶级数默认值通常适用于大多数情况，不再进行修改；对于自定义季节性由于时间序列的长度超过 2 个周期，Prophet 将默认拟合每周和每年的季节性，而本数据有 5 个周期，因此也不再单独设置。

2.3　模型评估指标

采用均方根误差（RMSE）、平均绝对误差（MAE）以及决定系数（R^2）对模型进行评估。RMSE 以及 MAE 的值越小，模型预测效果越好；R^2 一般在 0～1 之间，越接近于 1 表明模型效果越好。评估指标计算见公式（2～4）。

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

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - y')^2} \tag{4}$$

式中：n为样本数量，y_i为第 i 个花粉浓度实际值，$\hat{y}_i$为第 i 个花粉浓度预测值，y'为花粉浓度实际值均值。

2.4　统计分析软件

R 软件中的 mice（）函数用于数值插补，选择最优的插补数据集。加载 forecast 包，ndiffs（）函数确定一般差分阶数，人工建立 SARIMA 模型，并使用 auto.arima（）函数自动建立 SARIMA 模型；根据 AIC 信息准则比较人工定阶和自动定阶的 SARIMA 模型；Box.test（）函数检验模型残差是否为白噪声序列；分别使用 keras 深度学习包以及 prophet（）函数，建立 LSTM 模型和 Prophet 模型。R 软件 ggplot 包与 Origin 软件绘制拟合与预测图。

3　结果与分析

3.1　建立插补数据集

检查花粉浓度数据集中缺失值，2015—2020 年缺失值比例分别为 0.28、0.28、0.15、0.12、0.24、0.08，均小于 0.50，符合缺失值插补的要求。4 种数据插补法中，随机森林法的第 3 个插补数据集 P 值最小（P = 0.002，表 1），表明其模型拟合效果最佳；结合插补数据集密度图（图 1），随机森林法插补数据集的分布与原始监测数据最为吻合，因此被选为最优插补数据集。2015—2020 年 4 月 1 日—9 月 29 日插补数据与监测数据的分布（图 2）显示，插补数据的变化趋势与监测数据基本一致，符合每日花粉浓度的实际变化情况，可用于后续变化趋势分析与时间序列模型建立。

3.2　每日花粉浓度变化

2015—2020 年北京市每日花粉浓度变化（插补数据集）显示，花粉浓度每年均呈现春、秋两个高峰期，以及 6—8 月低且平稳的低谷期。各年到达高峰期的时间不同，持续时间与浓度峰值也存在差异。2015 年高峰期持续时间最长（春季 107 d，秋季 65 d），2018 年最短（春季 60 d，秋季 46 d）；2020 年花粉浓度峰值最高，春季为 2 142 粒/（10^3 mm^2），秋季为 1 490 粒/（10^3 mm^2）；2019 年花粉浓度峰值最低，春季 397 粒/（10^3 mm^2），秋季 240 粒/（10^3 mm^2）。根据图 3，4 月 1 日的花粉浓度已达到 400 粒/（10^3 mm^2）左右，因此推断春季高峰期从 3 月份开始，集中在 3—6 月，在 4 月初达到峰值（792 粒/（10^3 mm^2））；秋季高峰期集中在 8 月至 9 月末，在 9 月初达到峰值（449 粒/（10^3 mm^2））。2015—2019 年花粉浓度总体呈逐年下降趋势，但在 2020 年呈现阶跃式上升，各年花粉平均浓度再次接近 2015 年的水平（图 4）。

3.3　时间序列模型预测结果

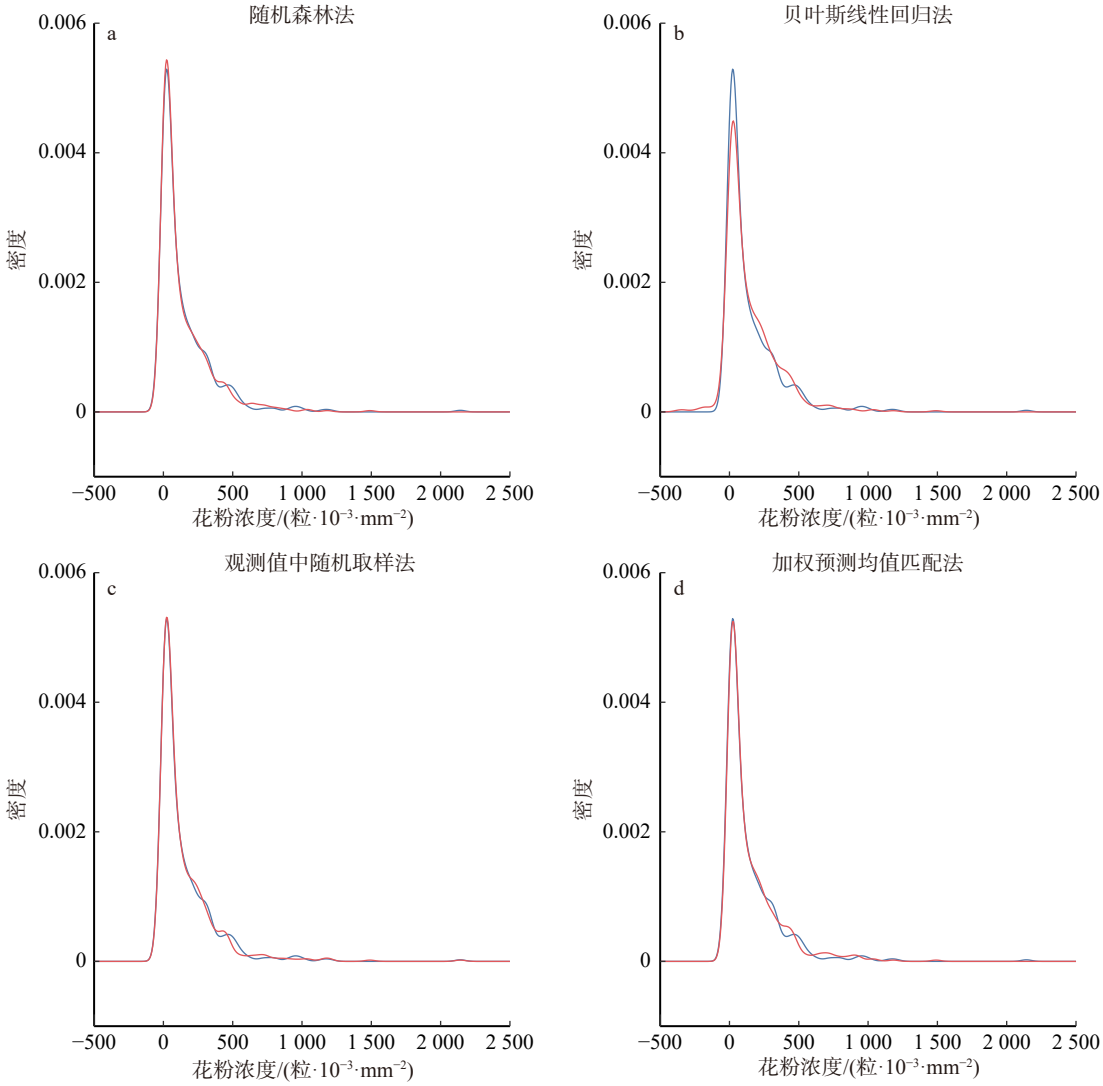
3.3.1　SARIMA 模型

对数据进行平稳性检验，P 值远小于 0.05，接受备择假设，表明序列平稳，可进行时间序列建模。首

表 1 不同插补数据集 *P* 值

Tab. 1 *P* values of different imputation datasets

数据插补法	插补数据集编号	*P*值	数据插补法	插补数据集编号	*P*值
随机森林法	1	0.014	观测值中随机取样法	1	0.008
	2	0.003		2	0.009
	3	0.002		3	0.010
	4	0.004		4	0.003
	5	0.005		5	0.003
贝叶斯线性回归法	1	0.199	加权预测均值匹配法	1	0.074
	2	0.094		2	0.093
	3	0.221		3	0.071
	4	0.022		4	0.033
	5	0.176		5	0.087

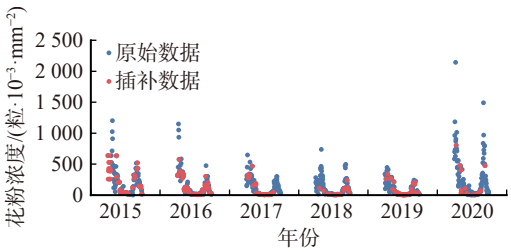


蓝色曲线表示原始数据，红色曲线表示插补后数据。

图 1 插补数据集密度图

Fig. 1 Density plots of interpolation dataset

先进行人工定阶，使用 ndiffs( )函数确定一般差分阶数为 $d=1$。从图 5、6 可以看出，一阶季节性差分后有效消除了数据的季节性影响，因此确定季节性差分阶数 $D=1$。拖尾表示(偏)自相关始终有非零取值，不会在阶数大于某个常数后就恒等于零或在 0 附近随机波动，截尾表示(偏)自相关在阶数大于



插补与监测数据为 2015—2020 年 4 月 1 日—9 月 29 日的北京市每日花粉浓度。

图 2　插补与监测数据分布散点图

Fig. 2　Scatter plot of interpolation and monitoring data distribution

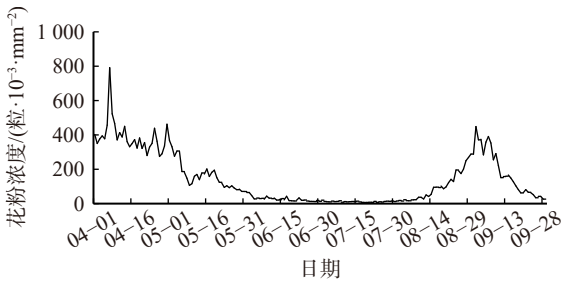


图 3　2015—2020 年每日平均花粉浓度

Fig. 3　Daily average pollen concentrations from 2015 to 2020

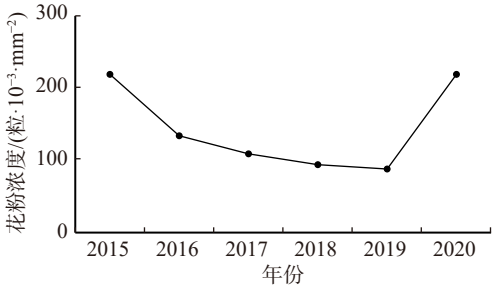


图 4　各年花粉平均浓度

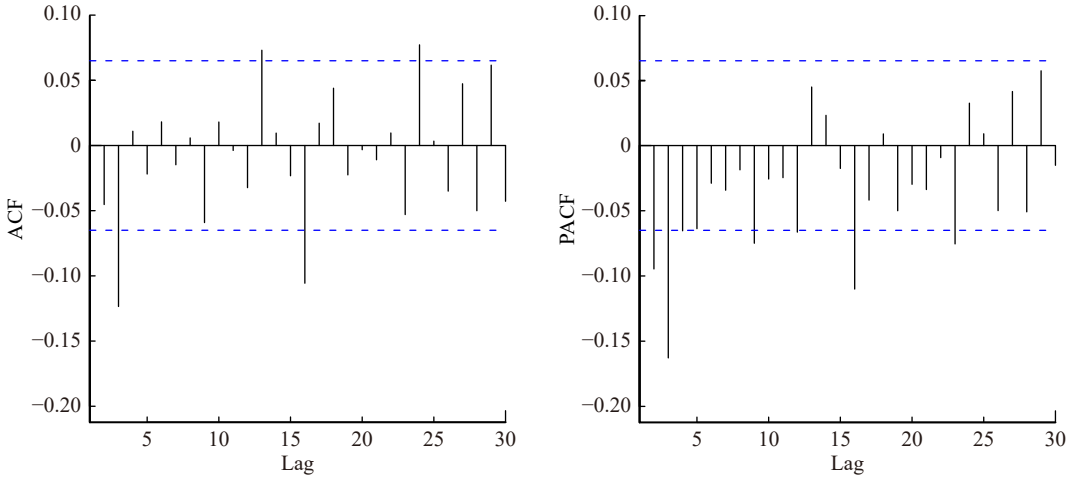
Fig. 4　Annual average pollen concentration

某个常数后快速趋于 0。根据图 6，自相关图与偏自相关图在阶数 12 以内均拖尾，因此自回归阶数 p = 4（取值 PACF 首次趋于 0 时对应的阶数），移动平均阶数 q = 2（取值 ACF 首次趋于 0 时对应的阶数）。当 PACF 12 阶后拖尾则 P 取值 1，若截尾则 P 取值 0；当 ACF 12 阶后拖尾则 Q 取值 1，若截尾则 Q 取值 0。图 6 中阶数大于 12 后均拖尾，因此季节性自回归阶数 P = 1，季节性移动平均阶数 Q = 1。最终，人工定阶后的 SARIMA 模型为$(4,1,2)(1,1,1)_{12}$。对模型进行白噪声检验，结果显示统计学检验值为 0.983 5，远大于 0.05，表明残差为白噪声序列，建模成功。再对模型自动定阶，结果为(4,1,1)，进行白噪声检验，统计学检验值为 0.958 5，远大于 0.05，证明残差是白噪声序列，建模成功。然而，自动定阶的模型未识别到时间序列的季节性特征，但是数据本身具有明显季节性，因此 SARIMA 自动定阶模型不能说明数据的分布情况。

人工定阶的 AIC 值为 10 434.74，明显低于自动定阶的 AIC 值（10 546.39），表明人工定阶的模型效果更好，SARIMA 最优模型表达式为$(4,1,2)(1,1,1)_{12}$。基于该模型，对未来 182 天的每日花粉浓度进行了预测，SARIMA 模型的拟合效果尚可，对高峰期和部分高峰值的拟合情况较好，但是对 6—8 月较平稳趋近于 0 的数据拟合情况有偏差（图 7a、8a）；模型预测结果欠佳，数据小幅度波动，预测值存在负数，无法预测到数据的高峰期与高峰值，预测线斜率将近 0，与实际不符（图 7b、8a）。

3.3.2　LSTM 模型建立结果

经试错法设置模型参数初始值 dense 层 units 为 60，lstm 层 units 为 182；运行训练时训练步数 epochs 为 100，每次迭代的批次大小 batch_size 为 30。研究中设置 look_back 取值参数分别为 14、30、60、90，则对于 2020 年花粉数据预测得到的预测值



Lag 表示滞后阶数。当 Lag = k 时，表示将时间序列向后移动 k 个时间点，考察当前值与过去 k 期值的关系。下同。

图 5　一阶一般差分自相关（ACF）与偏自相关（PACF）

Fig. 5　First order general difference autocorrelation (ACF) and partial autocorrelation (PACF)

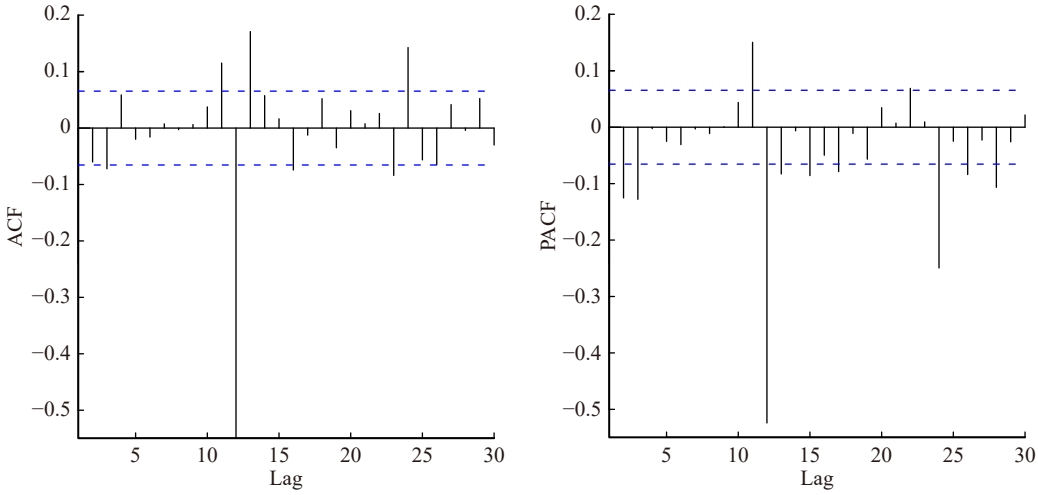


图 6 一阶季节性差分自相关(ACF)与偏自相关(PACF)

Fig. 6 Post-seasonal difference autocorrelation (ACF) and partial autocorrelation (PACF)

分别从 2020 年 4 月份第 15、31、61、91 天开始。比较选择 4 个模型中拟合与预测效果最好的模型，4 个 LSTM 模型统计学分析见表 2。可以看出，look_back = 90 时模型拟合效果最佳，look_back = 60 时模型预测效果最佳；但是 look_back = 60 时，拟合与预测的评估指标相差更少，效果更稳定。因此，以寻找花粉浓度预测效果最好的时间序列模型为目标，look_back = 60 时为最优 LSTM 模型。LSTM 模型(look_back = 60)对于高峰期、高峰值、6—8 月数据平稳期的拟合与预测效果好，拟合和预测线与验证线也相差不远，与实际值相近(图 7c、7d、8b)。

3.3.3 Prophet 模型

经尝试 Prophet 模型参数值设置为默认值时模型的拟合与预测效果更好，因此 Prophet 模型参数设置为：changepoint.range = 0.8，changepoint.prior.scale = 0.05。Prophet 模型拟合、预测线与验证线相差较多，且预测线的斜率为负，表示预测值与实际值相差过多(图 7e、7f)；Prophet 模型的拟合效果以及预测效果都不理想，对于训练集的高峰值无法灵敏捕捉，对训练集的高峰期不能准确拟合，尤其是预测值的分布为相同多次大幅度波动的曲线，且有负值，不能预测到数据的高峰期与高峰值，与实际值相差甚大(图 8c)。

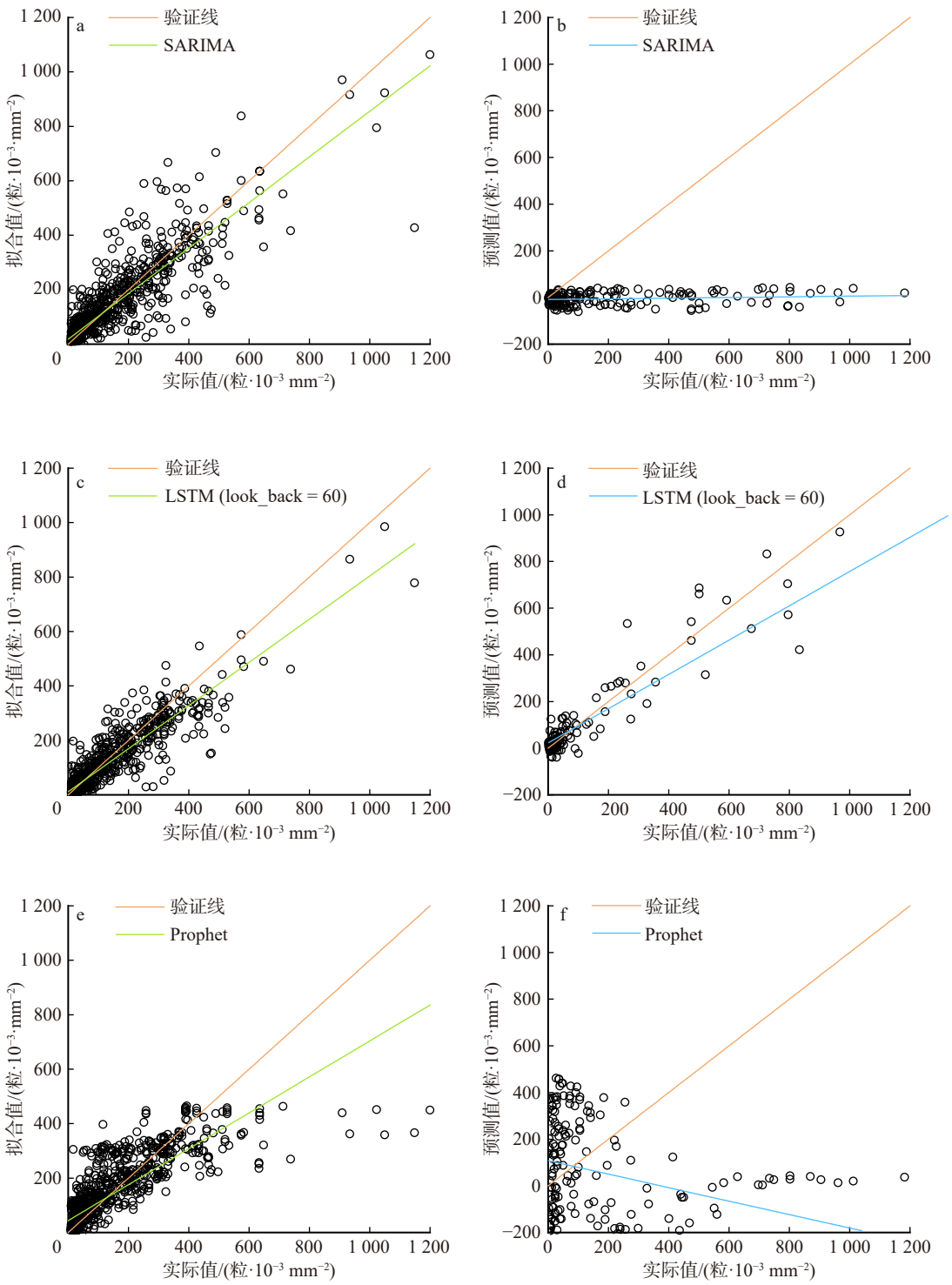
3.4 模型评估分析

由表 3 可看出，对于 RMSE 值和 MAE 值，各模型排序为 LSTM 模型 < SARIMA 模型 < Prophet 模型；对于 R^2，各模型排序为 LSTM 模型 > SARIMA 模型 > Prophet 模型。因此，LSTM 模型(look_back = 60)拟合与预测的 RMSE 与 MAE 最小，误差最小，R^2 为 0.78，优于其他模型。Prophet 模型的拟合效果最差，且 R^2 为 0.66，小于 0.7；SARIMA 模型以及 Prophet 模型预测值 R^2 均小于 0，表明模型预测效果差，均不适用于每日花粉浓度时间序列的预测。

4 讨 论

4.1 缺失值插补

依据 Shi 等[24]的研究结果，缺失值比例小于 50% 时产生的插补数据集与完整数据结果类似，在进行多重插补时应选择缺失值比例小于 0.5 的数据。共获取了北京市 2015—2022 年共 8 年的每日花粉浓度，经检查 2021 年数据缺失值比例超过了 0.5，2015—2020 年和 2022 年缺失值比例均小于 0.5。当 2015—2020 年数据作为训练集且 2022 年数据作为预测集时，预测时间间隔一年，有可能导致结果不理想；因此选取 2015—2019 年数据作为训练集，2020 年数据作为预测集。采用 4 种多重插值法对缺失值进行插补，其中随机森林法的第 3 个插补数据集 P 值最小(P = 0.002)，最适用于每日花粉浓度时间序列的缺失值插补。Shi 等[24]与 Avtar 等[25]的研究结果均表明多重插补法在处理数据缺失值时比其他方法得到的结果更好。多重插补法可对每一种插补方式生成多个插补数据集并整合，将空缺值视为随机样本，考虑到了由于数据插补而产生的不确定性，从而提高插补数据集的精度，为花粉变化规律分析和时间序列模型建立做好准备。多重插补法对花粉浓度缺失数据插补有效可行，与 Everett 等[26]多重插值法可有效处理缺失数据集的研究结论一致。缺失机制分为完全随机缺失、随机缺失、非随机缺失 3 种[27]，本研究数据为花粉数据监测集，缺失机制不属于非随机缺失，可采用多重插补法标准填补过程进行缺失值插补[27]，未来还可做缺失机制的研究。多重插补法属于统计学插补方法，通过设置 method 参数选



绿色实线代表拟合值与实际值线性拟合的函数图像，蓝色实线代表预测值与实际值线性拟合的函数图像，橙色实线为 1∶1 验证线。

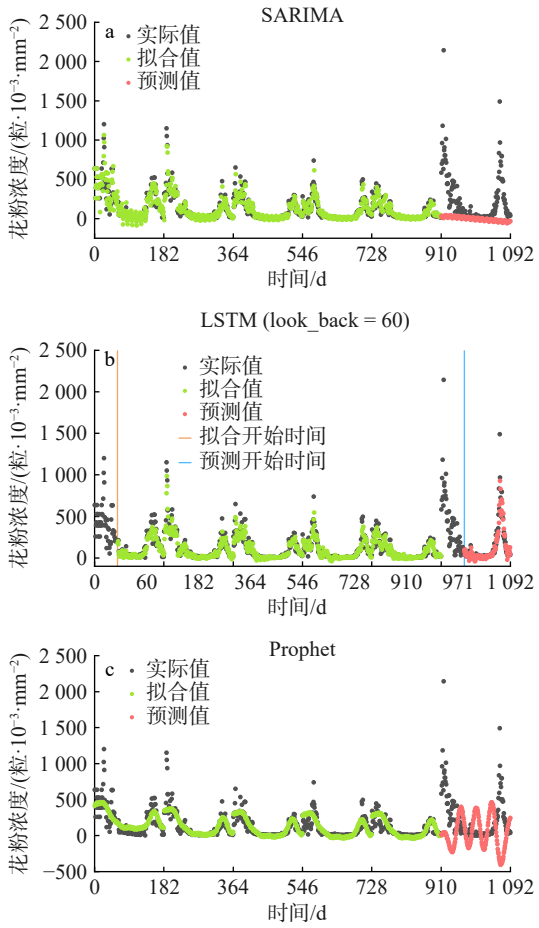
图 7　花粉浓度拟合值、预测值与实际值对比

Fig. 7　Comparison of fitted value, predicted value and actual value of pollen concentration

择缺失值填补所使用的方法，包括随机森林在内共 28 种；其他机器学习填补方法也可获得较高的插补数据集准确性，例如传统的 K 最近邻算法、支持向量机填补模型等，以及在传统基础上改进的插补方式（基于主成分分析和长短期记忆网络的缺失值插补等）[27−28]。将来可继续探寻机器学习插补法处理花粉浓度时间序列数据集缺失的问题，以期获得精确性相对较高的完备化数据集。

4.2　北京市每年花粉浓度变化特征

2015—2020 年北京市各年每日花粉浓度有两个高峰期，春季高峰期集中在 3—6 月，在 4 月初达到峰值（792 粒/ 10^3 mm^2）；秋季高峰期集中在 8 月—9 月末，在 9 月初达到（449 粒/ 10^3 mm^2）；双峰起止时间以及高峰值时间与孟龄等[29]对北京城区气传花粉季节分布特征的研究结果一致。花粉浓度在 2015—2019 年总体呈逐年下降趋势，但 2020 年呈



将 2015—2020 年 4 月 1 日—9 月 29 日每日花粉浓度数据按时间前后整合为从 1 到 1 092 的连续数字序列并作为的横坐标。

图 8 模型拟合预测

Fig. 8 Model fitting and prediction

现阶跃式上升。2015 年高峰期持续时间最长(春季 107 d，秋季 65 d)，2018 年高峰期持续时间最短(春季 60 d，秋季 46 d)；2020 年花粉浓度高峰值最高，2019 年花粉浓度高峰值最低。因此，北京市每年的花粉浓度虽有相同的双峰特征，但是各年也有数据波动。受数据限制，本文花粉浓度数据从 4 月开始，缺少 3 月份数据，因此不能确定春季高峰期的开始与持续时间，对未来花粉浓度数据更全面的监测将有助于分析花粉季节的变化特征。

4.3 3 种时间序列预测模型的比较

通过对各模型拟合和预测结果的误差分析与图表对比选择每日花粉浓度最优模型。SARIMA 模型的拟合结果尚可但预测结果不理想，预测 R^2 为负数，但拟合 R^2 为 0.77，存在模型过拟合的问题；该模型预测值存在为负的情况，不适用于北京市每日花粉浓度时间序列的长期预测。Zhao 等[14]的研究结果表示 ARIMA 模型可能不适用于海量数据的预测，SARIMA 模型属于 ARIMA 模型的特殊情况，可能对于少量数据的预测会更加准确；本研究中 SARIMA 模型预测开始即第一个值就与实际值相差甚多，因此 SARIMA 模型的预测效果不良可能与捕捉数据季节性和趋势性的灵敏度的关系更大，与实际可预测的步长区间的关系较小。预测结果出现负数可能是由于 2015—2019 年花粉浓度数据呈现递减趋势，但通过了平稳性检验，导致后续预测出现负值。本研究结果与前人 SARIMA 具有较好的预测结果不同[11,20]，这可能是用于建模的数据不同造成的，前人均采用月度数据进行模型构建与预测，预测步长也较短，而本研究数据量大，预测步长也相对较长，SARIMA 模型的建立与预测都出现了偏差。LSTM 模型拟合以及长期的预测结果较好，RMSE 和 MAE 最小，R^2 最大，对高峰期、峰值和 6—8 月份的数据

表 2 look_back 取值对比

Tab. 2 look_back value comparison

look_back取值	拟合值			预测值		
	RMSE/(粒·10⁻³·mm⁻²)	MAE/(粒·10⁻³·mm⁻²)	R^2	RMSE/(粒·10⁻³·mm⁻²)	MAE/(粒·10⁻³·mm⁻²)	R^2
14	72.74	37.75	0.79	115.42	59.66	0.73
30	63.89	35.94	0.80	114.27	60.44	0.73
60	60.72	36.15	0.80	109.10	49.24	0.78
90	57.43	33.81	0.83	128.39	60.50	0.76

表 3 3 个模型误差分析

Tab. 3 Error analysis of 3 models

模型	拟合值			预测值		
	RMSE/(粒·10⁻³·mm⁻²)	MAE/(粒·10⁻³·mm⁻²)	R^2	RMSE/(粒·10⁻³·mm⁻²)	MAE/(粒·10⁻³·mm⁻²)	R^2
SARIMA	77.99	45.02	0.77	380.30	224.64	−0.49
LSTM	60.72	36.15	0.80	109.10	49.24	0.78
Prophet	93.83	56.38	0.66	502.85	393.91	−1.61

平稳期均能取得较好的拟合与预测结果; 时间步长 look_back 的取值不同, 模型的拟合与预测精度也有所不同, 模型建立时需选择最适值。在未来推广适用时, 对于不同的数据需设置不同的 LSTM 模型参数值。本研究与前人对于 LSTM 研究不同的一点是通过向 LSTM 模型中添加滑动窗口的方式进行花粉浓度预测[14−15], 效果比较理想, 日后也可进一步推广滑动窗口的应用, 提高模型的预测精度。Prophet 模型对于北京市每日花粉浓度时间序列的拟合以及预测结果最差, 对高峰值无法捕捉, 对高峰期、峰值以及 6—8 月份的数据平稳期预测结果均与实际值相差较大, 不适用于描述北京市每日花粉浓度时间序列的分布并对其进行长期预测,与 Prophet 模型在空气污染物预测方面取得了精确的预测结果不同[17]。这可能是由于 Prophet 模型的结构与其他两个模型相比更简单且不能像 LSTM 模型一样进行滚动预测, 导致进行模型训练时无法准确学习数据的复杂模式, 预测精度较低。

对于北京市每日花粉浓度, 建立 LSTM 模型(长短期记忆神经网络模型)明显能够更好地对数据进行拟合以及长期预测。该结果同 Zhao 等[14]以及 Chimmula 等[15]的研究结论一致。由于训练的数据缺少 3 月份数据, 北京市花粉季节在 3 月初已经开始, 此时花粉浓度已经达到较高水平, 一定程度上影响了模型精度。

5　结　　论

采用 2015—2020 年 4 月 1 日—9 月 29 日北京市每日花粉浓度数据, 使用多重插值法插补缺失数据, 方法可行。北京市每年花粉浓度存在春季、秋季两个高峰期, 每年达到花粉高峰值的时间不同。每年花粉平均浓度在 2015—2019 年呈现连续下降趋势, 2019 年花粉浓度峰值最低; 在 2020 年呈阶跃式上升趋势, 2020 年花粉浓度峰值最高; 2015 年高峰期持续时间最长, 2018 年高峰期持续时间最短。应用 LSTM 模型拟合和预测北京市每日花粉浓度效果最佳; Prophet 模型效果较差; SARIMA 模型的拟合结果尚可但预测结果差, 后两种模型不适用于北京市每日花粉浓度的长期预测。

本研究对每日花粉浓度时间序列进行模型建立并进行了 182 d 的长期预测, 可预测未来长期的花粉浓度趋势, 提前预知花粉季节以及花粉高峰期、高峰值的到来, 并为未来长期花粉浓度预测提高科学依据。未来监测数据成为大数据库, 数据量会不断增长, 可尝试采用 Python 软件自动调整 LSTM 模型的 units、epochs、batch_size 参数, 从手工到编程自动完成调参, 将 AI 技术应用其中; 探讨在 LSTM 模型中构建更多 lstm 层, 或通过改变滑动窗口预测策略提高模型精度, 进一步完善花粉浓度的监测方法并提高花粉浓度监测数据的完整性, 提高日花粉浓度预测预报的准确度。

参　考　文　献

[1] 孙爱芝, 张海红, 李雪银, 等. 北京市北部空气花粉类型及浓度变化特征研究 [J]. 地理科学, 2023, 43(4): 737−744.
Sun A Z, Zhang H H, Li X Y, et al. Characteristics of airborne pollen types and concentrations in the north of Beijing, China[J]. Scientia Geographica Sinica, 2023, 43(4): 737−744.

[2] 李翔宇, 吴娟, 徐柳溪. 北京市密云区 2015—2019 年空气花粉浓度特征分析 [J]. 南方农业, 2022, 16(9): 150−152.
Li X Y, Wu J, Xu L X. Analysis of air pollen concentration characteristics in Miyun District, Beijing from 2015 to 2019[J]. South China Agriculture, 2022, 16(9): 150−152.

[3] 刘克利, 孙红斌, 王旭东, 等. 呼和浩特市花粉浓度与气象因子关系的分析 [J]. 内蒙古气象, 2009(6): 22−24.
Liu K L, Sun H B, Wang X D, et al. Analysis of relationship between pollen concentration and meteorological factors in Hohhot[J]. Meteorology Journal of Inner Mongolia, 2009(6): 22−24.

[4] 郑家华, 李健, 李清华, 等. 承德市区气传花粉浓度监测及意义 [J]. 中国耳鼻咽喉头颈外科, 2021, 28(5): 301−304.
Zheng J H, Li J, Li Q H, et al. Monitoring and clinical significance of airborne pollen concentration in Chengde[J]. Chinese Archives of Otolaryngology-Head and Neck Surgery, 2021, 28(5): 301−304.

[5] 刘宜纲, 吕世华, 刘建忠, 等. 2012—2016 年海淀区气传花粉物候特征及其与气象要素的关系 [J]. 应用生态学报, 2019, 30(10): 3563−3571.
Liu Y G, Lü S H, Liu J Z, et al. Phenological characteristics of airborne pollen and its relationship with meteorological factors in Haidian District, Beijing, China during the period of 2012−2016[J]. Chinese Journal of Applied Ecology, 2019, 30(10): 3563−3571.

[6] 卞萌, 郭树毅, 王威, 等. 融合植被遥感数据的北京市次日花粉浓度预测 [J]. 地球信息科学学报, 2021, 23(9): 1705−1713.
Bian M, Guo S Y, Wang W, et al. Next-day prediction of pollen concentration in Beijing by integrating remote sensing derived leaf area index[J]. Journal of Geo-information Science, 2021, 23(9): 1705−1713.

[7] Iglesias-Otero M A, Fernández-González M, Rodríguez-Caride D, et al. A model to forecast the risk periods of Plantago pollen allergy by using the ANN methodology[J]. Aerobiologia, 2015, 31(2): 201−211.

[8] Nowosad J. Spatiotemporal models for predicting high pollen concentration level of *Corylus, Alnus*, and *Betula*[J]. International Journal of Biometeorology, 2016, 60: 843−855.

[9] Devadas R, Huete A R, Vicendese D, et al. Dynamic ecological observations from satellites inform aerobiology of allergenic grass

pollen[J]. Science of the Total Environment, 2018, 633: 441−451.

[10] Am Seo Y, Kim K R, Cho C, et al. Deep neural network-based concentration model for oak pollen allergy warning in South Korea[J]. Allergy, Asthma & Immunology Research, 2020, 12(1): 149−163.

[11] Milenković M, Švadlenka L, Melichar V, et al. SARIMA modelling approach for railway passenger flow forecasting[J]. Transport, 2015, 33(2): 1−8.

[12] 任树顺, 高萌, 王煦雯, 等. 基于 3 种时间序列模型的九龙江河流库区藻华预测 [J]. 环境科学学报, 2022, 42(11): 172−183.
Ren S S, Gao M, Wang X W, et al. Algal bloom prediction in the Jiulong River Reservoir based on three types of time series models[J]. Acta Scientiae Circumstantiae, 2022, 42(11): 172−183.

[13] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735−1780.

[14] Zhao Z, Chen W, Wu X, et al. LSTM network: a deep learning approach for short-term traffic forecast[J]. IET Intelligent Transport Systems, 2017, 11(2): 68−75.

[15] Chimmula V K R, Zhang L. Time series forecasting of COVID-19 transmission in Canada using LSTM networks[J]. Chaos, Solitons & Fractals, 2020, 135: 109864.

[16] 陆凯, 汤娟华, 袁焱. Prophet 模型和 SARIMA 模型在某市心脑血管疾病死亡人数预测中的应用比较 [J]. 中国卫生统计, 2023, 40(2): 285−287, 291.
Lu K, Tang J H, Yuan Y. Comparison of prophet model and SARIMA model in predicting the number of deaths from cardiovascular and cerebrovascular diseases in a certain city[J]. Chinese Journal of Health Statistics, 2023, 40(2): 285−287, 291.

[17] 张骏, 周磊, 王文梁, 等. 基于 Prophet 模型的空气污染物浓度预测方法 [J]. 中国环境监测, 2022, 38(6): 59−68.
Zhang J, Zhou L, Wang W L, et al. Air pollutant concentration prediction method based on prophet model[J]. Environmental Monitoring in China, 2022, 38(6): 59−68.

[18] Toharudin T, Pontoh R S, Caraka R E, et al. Employing long short-term memory and facebook prophet model in air temperature forecasting[J]. Communication in Statistics-Simulation and Computation, 2021, 52(2): 279−290.

[19] 李挚, 何海娟, 孙国强, 等. 北京市区与过敏相关的气传花粉 [J]. 基础医学与临床, 2015, 35(6): 734−738.
Li Z, He H J, Sun G Q, et al. Allergy-related airborne pollen in urban area of Beijing[J]. Basic & Clinical Medicine, 2015, 35(6): 734−738.

[20] Zhongheng Z. Multiple imputation with multivariate imputation by chained equation (MICE) package[J]. Annals of Translational Medicine, 2016, 4(2): 30.

[21] 李俊, 王洪源, 张志刚, 等. 空气花粉浓度与变应性鼻炎就医人次的暴露−反应关系初探 [J]. 环境与健康杂志, 2008, 25(9): 793−797.
Li J, Wang H Y, Zhang Z G, et al. Exposure-response relationship between air pollen concentration and allergic rhlnitls medical consultations[J]. Journal of Environment and Health, 2008, 25(9): 793−797.

[22] 刘天, 姚梦雷, 侯清波, 等. 7 种时间序列模型对全国肾综合征出血热发病率预测效果比较 [J]. 中国媒介生物学及控制杂志, 2022, 33(4): 548−554.
Liu T, Yao M L, Hou Q B, et al. Comparison of seven time series models in fitting and predicting the incidence of hemorrhagic fever with renal syndrome in China[J]. Chinese Journal of Vector Biology and Control, 2022, 33(4): 548−554.

[23] 袁依, 庞振陆, 崔庆霞, 等. 基于 SARIMA 模型或 SARIMA-GRNN 组合模型的急性出血性结膜炎发病预测 [J]. 郑州大学学报 (医学版), 2023, 58(1): 78−82.
Yuan Y, Pang Z L, Cui Q X, et al. Forecasting the incidence of acute hemorrhagic conjunctivitis in China based on SARIMA model or SARIMA-GRNN combination model[J]. Journal of Zhengzhou University (Medical Sciences), 2023, 58(1): 78−82.

[24] Shi D, Lee T, Fairchild A J, et al. Fitting ordinal factor analysis models with missing data: a comparison between pairwise deletion and multiple imputation[J]. Educational and Psychological Measurement, 2020, 80(1): 41−66.

[25] Avtar S S, Khuneswari G P, Abdullah A A, et al. Comparison between EM algorithm and multiple imputation on predicting children's weight at school entry[J]. Journal of Physics: Conference Series, 2019, 1366: 012124.

[26] Everett C J, Gebregziabher M. Estimating missing concentrations of p,p′-DDT in the 1999–2004 National Health and Nutrition Examination Survey using multiple imputation[J]. Environmental Science and Pollution Research, 2019, 26(6): 5885−5891.

[27] 熊中敏, 郭怀宇, 吴月欣. 缺失数据处理方法研究综述 [J]. 计算机工程与应用, 2021, 57(14): 27−38.
Xiong Z M, Guo H Y, Wu Y X. Review of missing data processing methods[J]. Computer Engineering and Applications, 2021, 57(14): 27−38.

[28] 张坤, 肖慧, 徐哈宁, 等. 基于主成分分析和长短期记忆网络的滑坡地表位移监测数据缺失插补算法 [J]. 科学技术与工程, 2023, 23(26): 11129−11135.
Zhang K, Xiao H, Xu H N, et al. Missing interpolation algorithm of landslide surface displacement monitoring data based on principal component analysis and long short term memory network[J]. Science Technology and Engineering, 2023, 23(26): 11129−11135.

[29] 孟龄, 王效科, 欧阳志云, 等. 北京城区气传花粉季节分布特征 [J]. 生态学报, 2013, 33(8): 2381−2387.
Meng L, Wang X K, Ouyang Z Y, et al. Seasonal dynamics of airborne pollen in Beijing urban area[J]. Acta Ecologica Sinica, 2013, 33(8): 2381−2387.

（责任编辑 范 娟
责任编委 徐程扬）