

基于随机森林算法的美国芝加哥空气质量预测研究

山 昊¹, 范晓沁¹, 杨玉兰¹, 房志明^{1,2}, 黄中意^{1,2*}

¹上海理工大学管理学院, 上海

²上海理工大学智慧应急管理学院, 上海

收稿日期: 2025年9月3日; 录用日期: 2025年9月13日; 发布日期: 2025年10月13日

摘 要

随着城市化的加速, 空气污染问题日益严重, 因此准确预测影响空气质量的污染物含量对于改善城市空气质量具有重要意义。本文基于随机森林算法建立了空气质量预测模型。该模型仅利用与空气质量指数(AQI)相关性较强的污染物历史数据, 即可对未来的空气质量进行短期预测, 模型操作简单, 预测结果精确度也较高。首先, 选取了美国芝加哥市2019年1月至2021年1月的空气质量综合指数, 对该城市的AQI数据进行相关性分析, 识别出与AQI相关的主要污染物指标。然后, 采用随机森林算法建立空气质量预测模型。实验结果表明, 本文建立的空气质量预测模型对于芝加哥城市O₃浓度的训练集上的R²达到了0.92, 验证集上的拟合优度为0.87, 对于PM_{2.5}、NO₂、SO₂等其他主要污染物也有较好的拟合效果。本研究基于AQI提取的相关性污染物指标的随机森林预测模型能显著提升空气质量预测的准确性, 能够为空气质量提供有效的决策支持。

关键词

随机森林, AQI相关性, 空气质量预测

A Study on Air Quality Prediction in Chicago, USA Based on Random Forest Algorithm

Hao Shan¹, Xiaoqin Fan¹, Yulan Yang¹, Zhiming Fang^{1,2}, Zhongyi Huang^{1,2*}

¹Business School, University of Shanghai for Science and Technology, Shanghai

²School of Intelligent Emergency Management, University of Shanghai for Science and Technology, Shanghai

Received: September 3, 2025; accepted: September 13, 2025; published: October 13, 2025

*通讯作者。

文章引用: 山昊, 范晓沁, 杨玉兰, 房志明, 黄中意. 基于随机森林算法的美国芝加哥空气质量预测研究[J]. 建模与仿真, 2025, 14(10): 97-107. DOI: 10.12677/mos.2025.1410609

Abstract

With the acceleration of urbanization, air pollution has become increasingly severe. Therefore, accurately predicting pollutant concentrations that affect air quality is crucial for improving urban air quality. This paper establishes an air quality prediction model based on the Random Forest algorithm. This model can make short-term predictions of future air quality using only historical data of pollutants strongly correlated with the Air Quality Index (AQI). The model is simple to operate and delivers highly accurate prediction results. First, the comprehensive air quality index data for Chicago, USA, from January 2019 to January 2021 were selected. Correlation analysis of the city's AQI data identified key pollutant indicators correlated with AQI. The Random Forest algorithm was then employed to build the air quality prediction model. Experimental results demonstrate that the proposed air quality prediction model achieves a coefficient of determination of 0.87 for O_3 concentrations in Chicago, with comparable fitting performance for other primary pollutants, including $PM_{2.5}$, NO_2 , and SO_2 . This random forest model, trained on correlation-based pollutant indicators extracted from AQI data, significantly enhances the accuracy of air quality forecasting and provides effective decision support for air quality management.

Keywords

Random Forest, AQI Correlation, Air Quality Prediction

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

近年来,随着工业生产的发展、能源消耗的提高以及城市人口的迅速增长,大气污染问题日趋严重,这给人们的生命健康和生产生活带来了严重威胁[1],也对地球的生态环境造成了严重危害。因此在大力发展生产力的同时,我们迫切需要保护和改善大气环境。在这样的情况下,要认真考虑经济发展与环境质量之间的协调关系。如何在环境固定的承载力范围之内,使得经济得到较好的可持续发展,这已经成为全社会面临的重要问题。而开展环境空气质量预测工作是及时应对重污染天气的重要技术保障,对区域大气污染的联合减排等具有指导意义。

空气质量指数(Air Quality Index, AQI)是空气污染物总效应的无量纲指数,它反映了空气的污染水平以及对人体健康的影响程度。在公共卫生领域及城市地区,主要有二氧化硫(SO_2)、二氧化氮(NO_2)、颗粒物($PM_{2.5}$ 、 PM_{10})、臭氧(O_3)和一氧化碳(CO)空气污染成分。许多学者基于 AQI 开展了广泛的研究。例如,王璐基于 SPSS 分析了甘肃省兰州地区空气质量指数与各气象要素之间的关系[2]。郑舒天等利用 2015~2019 年文山市气象监测站常规地面观测资料和文山市环境监测站空气质量指数(AQI)逐日平均值、各类污染物逐日平均浓度值等观测资料,通过统计学方法对不同季节气象要素和 AQI 做相关性分析,选取相关性好的因子做回归分析并分季节建立预报方程[3]。刘诗韵等利用 2015~2019 年乌兰察布市环境监测中心污染物浓度数据及 AQI,结合同期气象观测数据,分析乌兰察布市空气质量状况,研究空气质量与气象要素间的相关性,利用逐步回归分析筛选出关键影响因子,建立不同时段 AQI 预报模型并比较预测效果[4]。

此外,随着人工智能的发展,机器学习凭借其强大的数据学习和规律挖掘能力,在处理复杂数据结构、大数据方面展现出巨大的优势,已有大量学者基于神经网络、决策树、随机森林、支持向量机等算

法研究空气质量的预报。牛艳飞以鹤壁市 2021 年 6 月 1 日~12 月 31 日的空气质量指数(AQI)日报数据为样本, 基于时间序列分析建立预测模型[5]。Li 等提出了一种新颖的长短期记忆神经网络扩展(LSTME)模型, 该模型本身就考虑了时空相关性, 用于空气污染物浓度预测[6]。史雨梅等将图卷积神经网络(Graph Convolutional Networks, GCN)应用于空气质量指数(Air Quality Index, AQI)的预测, 基于 GCN 特征聚合的特点, 充分考虑邻近区域之间的空间相关性。文中以安徽省 16 个地级市为例, 构建了城市之间的依赖关系图, 预测各个城市的 AQI [7]。文昌俊等针对 BP 神经网络预测精度不佳、预测结果不稳定的问题, 引入 Tent 混沌映射克服初始种群分布不均的缺点, 针对黏菌算法位置更新的随机性和后期容易陷入局部最优等问题引入领导者策略和莱维飞行策略, 利用自适应反向学习策略扩大搜索空间并用 23 组基准函数加以测试。随后利用 ISMA 算法优化 BP 网络模型的初始权值和阈值, 构建 ISMA-BP 空气质量指数预测模型[8]。路凯丽等提出基于卷积神经网络和门控循环单元的集成深度学习模型(CNN-GRU)对 AQI 进行预测。其中, 卷积神经网络提取污染气体浓度和 AQI 的时空特征并完成特征映射, 门控循环单元(GRU)建模时序关系并高效完成计算。选取 2014~2022 年北京市和广州市的 6 种主要污染气体($PM_{2.5}$ 、 SO_2 、 CO 、 NO_2 、 O_3)日平均质量浓度和 AQI 进行实例研究, 使用 CNN-GRU 模型对 AQI 进行预测, 与多元宇宙优化的广义回归神经网络模型(MVO-GRNN)、遗传算法优化的 BP 神经网络模型对 AQI 的预测进行对比分析[9]。

本文针对现有研究提出一种基于随机森林算法的空气质量预测模型, 并将与 AQI 相关性较高的污染物指标作为预测模型的输入, 显著提升空气质量预测的准确性。

2. 随机森林算法

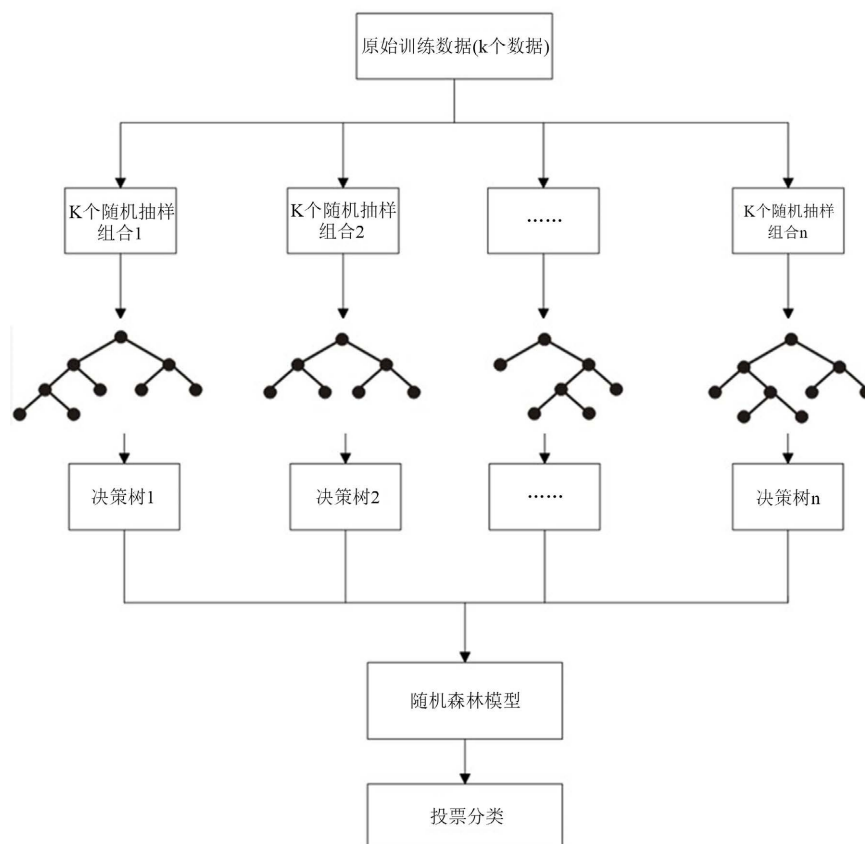


Figure 1. Schematic diagram of the random forest algorithm
图 1. 随机森林算法的原理图

随机森林算法是一种被广泛应用的机器学习算法，其工作原理主要表现在从输入数据集中抽取一个子集作为样本，然后在该子集上构建多棵决策树，如图 1 所示。这些决策树相互独立，每棵树仅使用随机选择的特征，并在分裂节点时使用随机选择的特征值。不同于使用所有可用特征和特征值，随机森林模型采用特征选择的方法，仅选择那些随机抽取的特征，以此减少过拟合的风险，并提高模型的泛化能力。同时，在分裂节点时使用随机选择的特征值，使得模型能够更好地捕捉数据的分布和变化，进一步提高模型的预测精度。最后，这些独立构建的决策树结果会被综合起来，以得出最终的预测结果。

基于随机森林算法的预测模型构建包含采样、构建决策树、预测结果三步。首先，数据采样采用 Bootstrap 重抽样方法，即从原始训练集中进行有放回的随机抽样，生成多个样本子集。每个子集用于训练一棵独立的决策树。由于采用有放回抽样策略，可有效增加子集的数量，提高模型的多样性。同时，每轮抽样中未被抽取的样本(称为“袋外样本”，Out-of-Bag (OOB))可用于模型性能的评估。

在构建每棵决策树时，算法从全部特征中随机选取 m 个特征，并在这些特征中进一步选择 k 个特征用于节点分裂。分裂点的选择依据均方误差(Mean Square Error, MSE)最小化原则。对于当前节点样本集 D ，其均方误差计算公式如下：

$$\text{MSE}(D) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

其中， $\text{MSE}(D)$ 表示样本集 D 的均方误差， n 为样本数量， y_i 为第 i 个样本的真实值， $\hat{y}_i$ 为其预测值。在决策树分裂过程中，分裂后左右子节点的加权均方误差按下式计算：

$$\text{MSE}_{\text{split}} = \frac{n_{\text{left}}}{n} \text{MSE}(D_{\text{left}}) + \frac{n_{\text{right}}}{n} \text{MSE}(D_{\text{right}})$$

其中， n_{left} 和 n_{right} 分别代表分裂后左右子节点的样本数。

在预测阶段，对于回归任务，每棵决策树输出一个预测值，最终结果为所有树预测值的平均值；对于分类任务，则采用投票机制，选择多数树所支持的类别作为最终预测结果。

3. 美国芝加哥空气质量预测研究

3.1. 数据来源与预处理

本研究中使用美国芝加哥市空气质量数据，包括 $\text{PM}_{2.5}$ 、 PM_{10} 、 SO_2 、 NO_2 、 O_3 等污染物数据，主要来源于美国环境保护署(EPA)的 AirNow 网站。这些数据涵盖了 2019 年至 2021 年期间的六种空气质量参数。

由于受到监测数据权限、有关监测设备功能和获取数据渠道等原因的限制，致使最终没能得到部分气象指标的实测数据，即出现了数据缺失的现象。KNN (K-Nearest Neighbor) 算法即 K 最邻近法，是一种非参的、惰性的算法模型。KNN 算法模型不会对数据进行假设，也就是说 KNN 建立的模型结构是由数据直接决定的，这与现实情况的符合度较高。相比其他算法，K 最邻近法简单易用，是一种相对简洁明了的算法，并且模型的预测效果好。由于原始数据里有部分的缺失值(异常值)，KNN 算法对于异常值不敏感。因此本文选择 K 最邻近法进行缺失的填补。KNN 算法的原理是当预测一个新的值的时候，根据它距离最近的 K 个点是什么类别来判断这个新值属于哪个类别。

对于本文的缺失值补充的 KNN 算法步骤是：1) 首先进行缺失值和已知值之间的距离计算。距离计算方法有欧氏距离、曼哈顿距离等。2) 选择参数 K，也是最重要的一步。3) 根据多数表决规则，确定缺失值进行填补。部分结果如表 1 所示。

Table 1. Example of results after missing value imputation**表 1.** 缺失值补充后结果示例

wind speed	1	1	0.8	1	缺失值	1.2	1.6
wind speed	1	1	0.8	1	1.4	1.2	1.6

3.2. 空气质量指数相关性分析

根据《环境空气质量指数(AQI)技术规定(试行)》(HJ633-2012), 空气质量指数(AQI)可用于判别空气质量等级。首先需得到各项污染物的空气质量分指数(IAQI), 其计算公式如下:

$$IAQI_p = \frac{IAQI_{Hi} - IAQI_{Lo}}{BP_{Hi} - BP_{Lo}} \cdot (C_p - BP_{Lo}) + IAQI_{Lo}$$

其中, $IAQI_p$ 为污染物 P 的空气质量分指数, C_p 为污染物 P 的质量浓度值, BP_{Hi} 、 BP_{Lo} 指的是与 C_p 相近的污染物浓度限值的高位值与低位值, $IAQI_{Hi}$ 、 $IAQI_{Lo}$ 指的是与 BP_{Hi} 、 BP_{Lo} 对应的空气质量分指数。

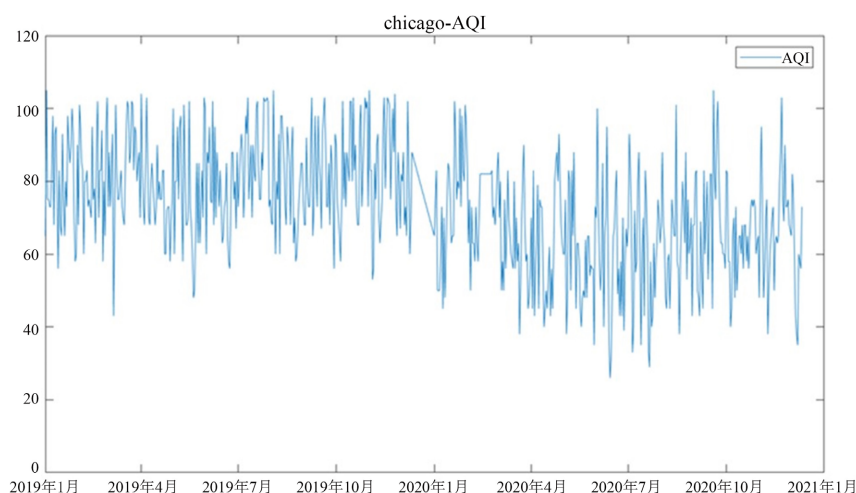
空气质量指数(AQI)取各分指数中的最大值, 即

$$AQI = \max \{IAQI_1, IAQI_2, IAQI_3, \dots, IAQI_n\}$$

其中, $IAQI_1, IAQI_2, IAQI_3, \dots, IAQI_n$ 为各污染物项目的分指数。最后, 对于 AQI 的计算仅涉及附表 1 中的六种污染物, 因此计算公式如下:

$$AQI = \max \{IAQI_{SO_2}, IAQI_{NO_2}, IAQI_{PM_{10}}, IAQI_{PM_{2.5}}, IAQI_{O_3}, IAQI_{CO}\}$$

空气质量等级范围根据 AQI 数值划分, 等级对应的 AQI 范围见附表 2。图 2 展示了芝加哥在 2019 年 1 月至 2021 年 1 月期间的空气质量指数(AQI)变化情况。从图中可以看出, AQI 值在不同时间段内波动较大, 反映出空气质量的不稳定。根据空气质量等级及对应 AQI 范围的划分, 当 AQI 小于或等于 50 时, 空气质量评价为“优”, 此时无首要污染物; 而当 AQI 大于 50 时, IAQI 最大的污染物为首要污染物。图中显示, 芝加哥的 AQI 值在多数时间内处于“良”(51~100)和“轻度污染”(101~150)之间, 表明空气质量总体上可接受。然而, 也有部分时间 AQI 值超过了“轻度污染”的范围, 进入“中度污染”(151~200)甚至“重度污染”(201~300)区间, 这可能会对更广泛人群的健康产生影响。

**Figure 2.** AQI index for Chicago**图 2.** Chicago 的 AQI 指数

为了进一步探究 AQI 与各污染物指标之间的相关性,采用皮尔逊相关系数来度量不同指标变量之间线性相关程度,以下是计算两变量之间线性相关系数的公式:

$$p_{x,y} = \text{cov}(x,y) = \frac{\text{cov}(x,y)}{\delta_x \delta_y} = \frac{E[(x-\mu_x)(y-\mu_y)]}{\delta_x \delta_y}$$

图 3 中的热力图展示了芝加哥不同污染物指标之间的相关性,其中颜色的深浅表示相关性的强弱,正相关用蓝色表示,负相关用红色表示。根据热力图的分析,可以发现 $\text{PM}_{2.5}$ 与 PM_{10} 之间的相关性较强,相关系数为 0.6251,这表明 $\text{PM}_{2.5}$ 和 PM_{10} 的浓度变化趋势较为一致,这可能是因为 $\text{PM}_{2.5}$ 和 PM_{10} 都主要来源于交通排放、工业排放等共同的污染源。此外,二氧化氮(NO_2)与一氧化碳(CO)之间的相关系数为 0.6116,显示出较强的正相关性,这可能是因为交通排放是二氧化氮和一氧化碳的共同来源,且在燃烧过程中往往同时产生这两种污染物。此外,从图中可以明显看出空气质量指数(AQI)与 $\text{PM}_{2.5}$ 、 PM_{10} 、 NO_2 、 CO 等污染物均表现出较强的正相关性。



Figure 3. Heatmap of pollutant correlation in Chicago

图 3. Chicago 的污染物指标相关性热力图

3.3. 预测模型建立

3.3.1. 模型构建与参数优化

针对模型参数优化,采用网格搜索与交叉验证相结合的方法,对随机森林的关键超参数进行系统性调优。重点调整了决策树数量($n_estimators$)、树的最大深度(max_depth)和节点最小样本数($min_samples$)这三个关键参数。具体而言,树的数量通过网格搜索在 100 到 1000 范围内以 100 为步长进行遍历,结合 5 折交叉验证评估不同树数量对模型泛化能力的影响,最大深度(max_depth)则设置从 5 到 20 的候选值。特征排序如图 4 所示,其中温度、气压和湿度位列前三。

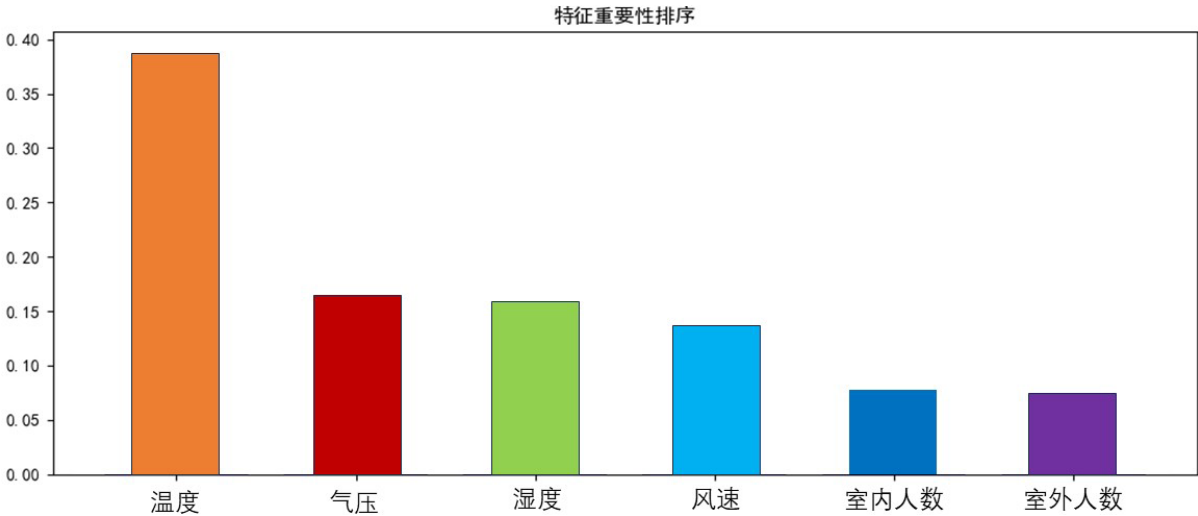


Figure 4. Feature importance ranking chart
图 4. 特征重要性排序图

通过五折交叉验证，确定了预测模型的参数组合。其中，决策树数量为 500，最大深度为 15，节点分裂所需的最小样本数为 2，训练结果如图 5 所示，预测值和实际值展现出高度的一致性。此外，模型在训练集上的均方根误差 MSE 为 1.93，相关系数 R^2 为 0.92。

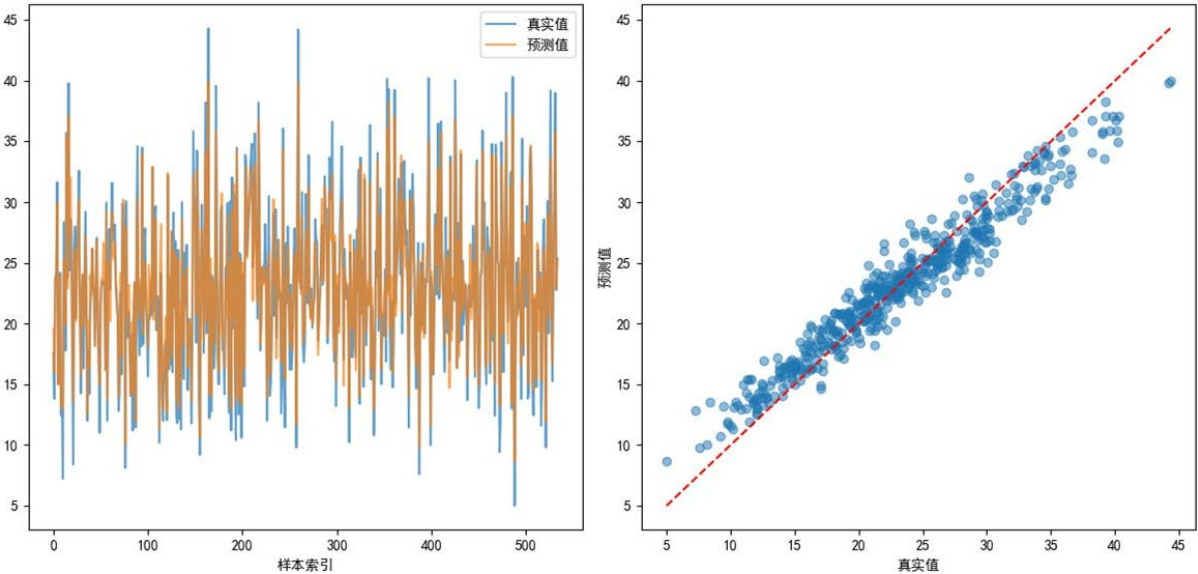


Figure 5. Comparison of actual values and predicted values for the model
图 5. 模型真实值与预测值的对比图

3.3.2. 模型验证与预测

将芝加哥城市的 O_3 数据分为训练集和验证集，图 6 展示了芝加哥城市 O_3 浓度的实际值与预测值的拟合图，随机森林预测模型的拟合优度为 0.87。

从图 7 中可以看出，显示出较高的拟合程度，这表明该模型能够较为准确地描述和预测所研究的现象。尽管模型在某些时间点上存在一定的偏差，但总体上能够较好地捕捉 O_3 浓度的变化趋势。

此外, 该模型的可解释性也较强, 各个变量的系数和符号都符合预期, 并且展现出了卓越的拟合度和出色的泛化能力, 各主要污染物拟合效果如图 7~图 10 所示。

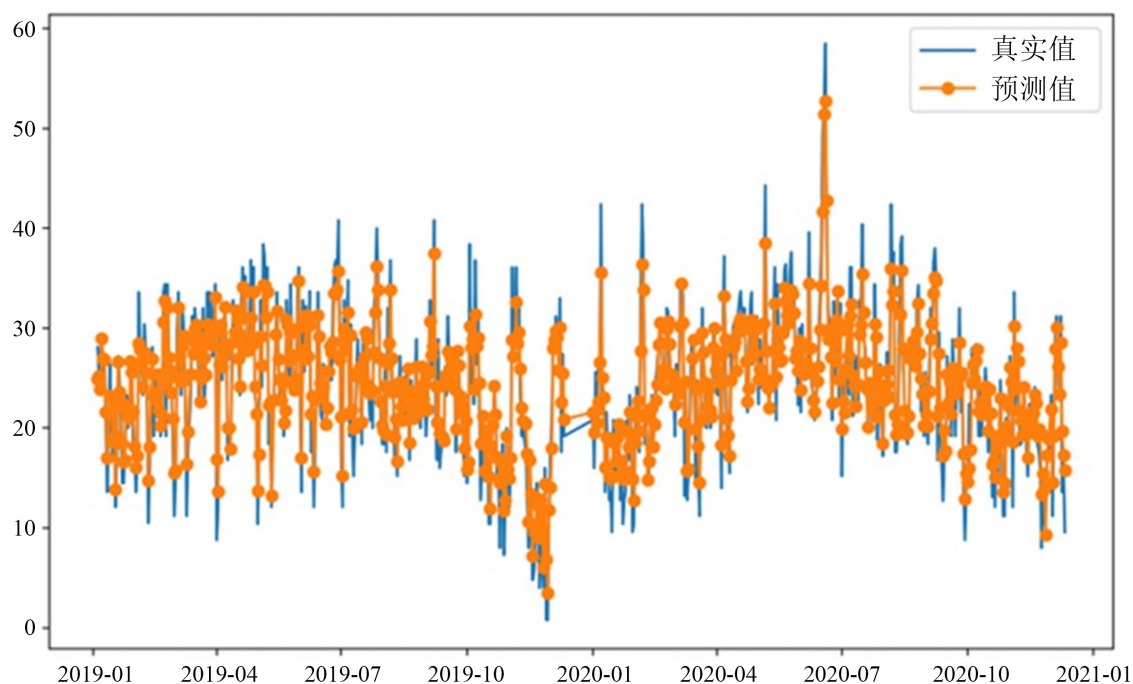


Figure 6. Fitting plot for O_3

图 6. O_3 的拟合图

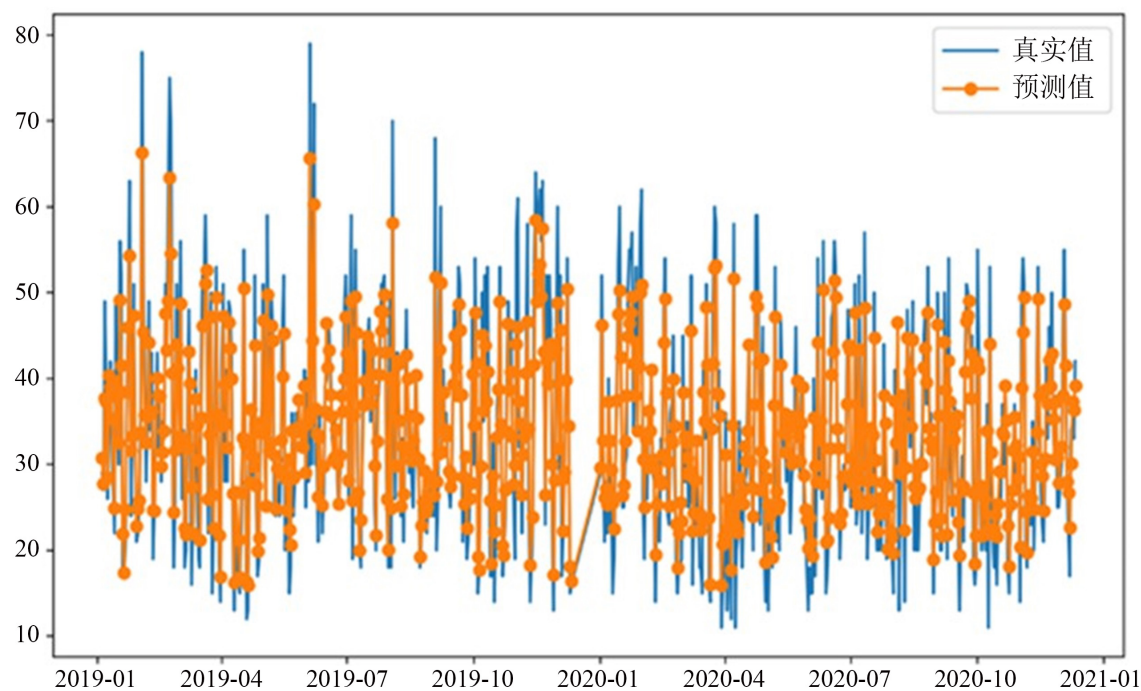


Figure 7. Fitting plot for $PM_{2.5}$

图 7. $PM_{2.5}$ 的拟合图

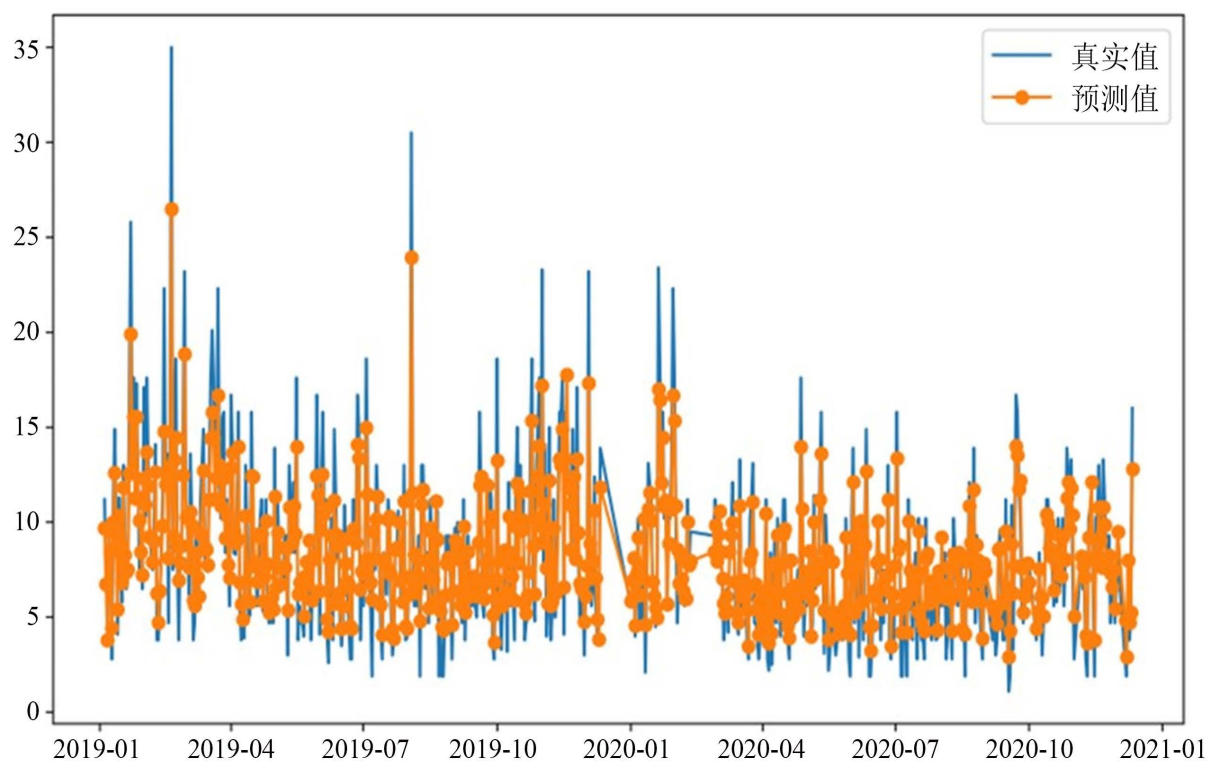


Figure 8. Fitting plot for NO₂

图 8. NO₂ 的拟合图

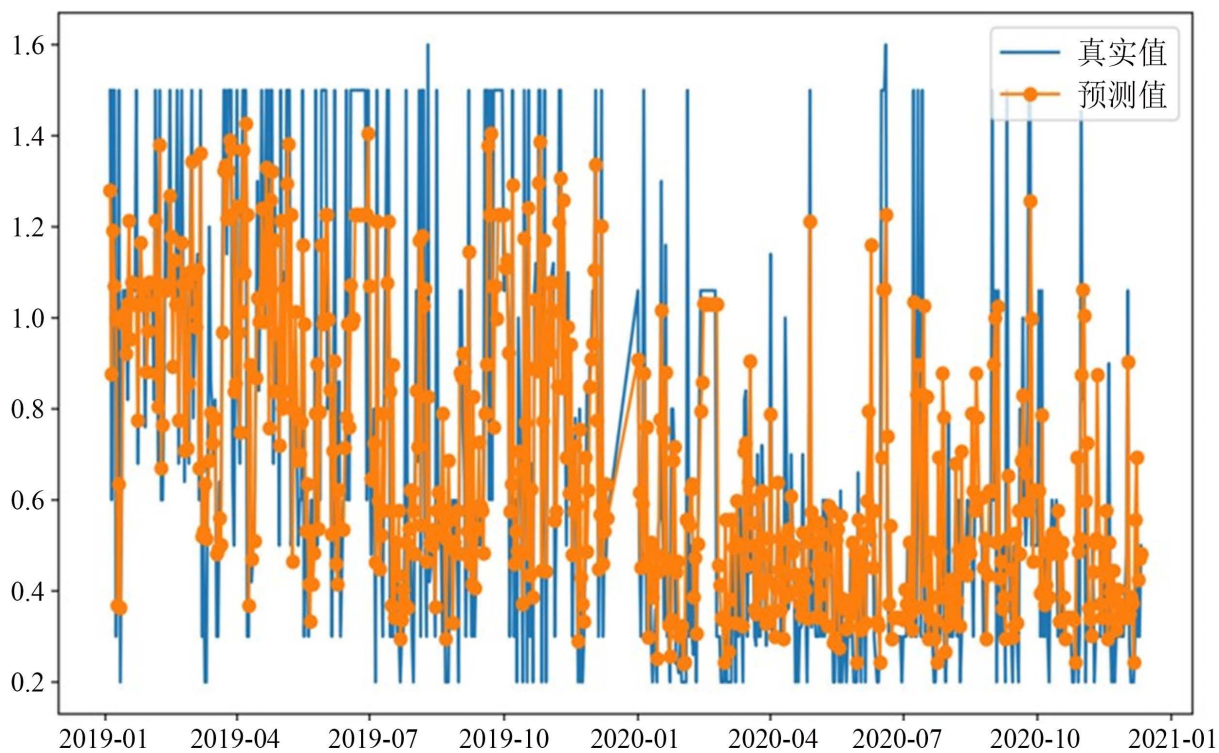


Figure 9. Fitting plot for SO₂

图 9. SO₂ 的拟合图

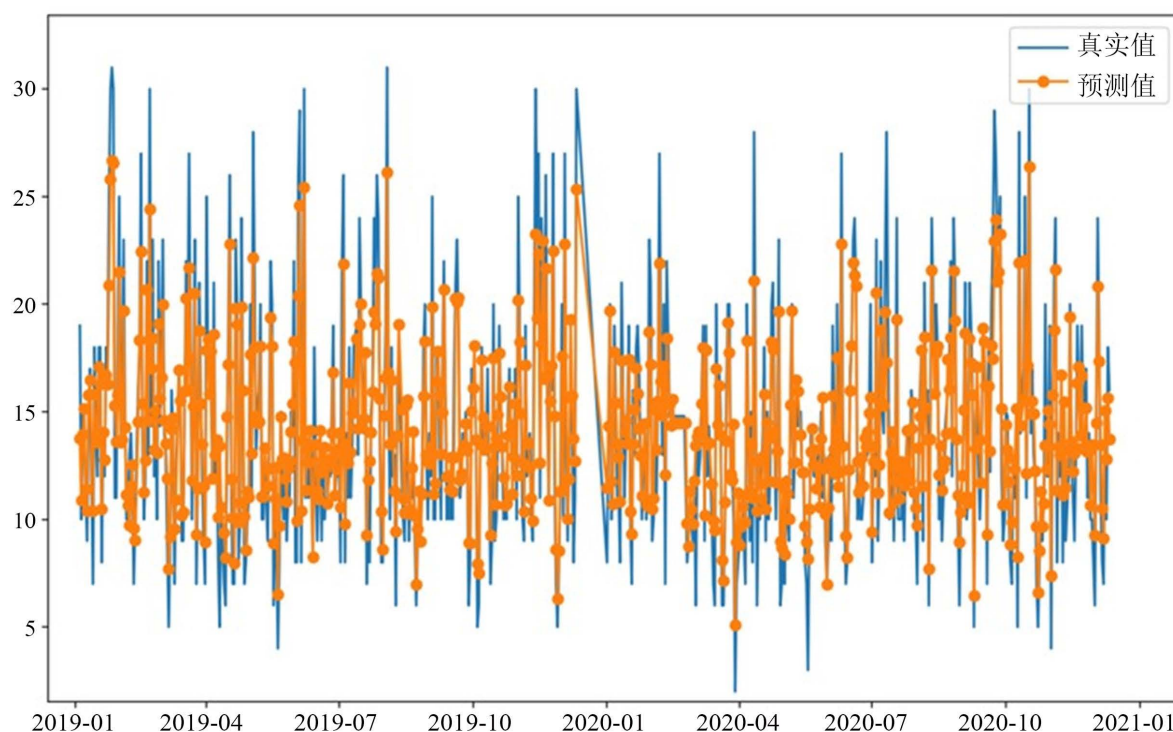


Figure 10. Fitting plot for PM₁₀

图 10. PM₁₀ 的拟合图

4. 结论

本文基于美国芝加哥市 2019 年 1 月至 2021 年 1 月的空气质量综合指数, 建立了随机森林空气质量预测模型。该模型仅利用与空气质量指数(AQI)相关性较强的污染物历史数据, 即可对未来的空气质量进行短期预测, 模型操作简单, 预测结果精确度也较高。

然而, 该模型虽然可以较为精确地预测美国芝加哥市空气质量的变化趋势, 但无法揭示导致该地区空气质量变化的影响因素和影响机制。关于这方面的不足, 还有待于进一步的深入研究和分析, 未来也将开展多因素影响下的空气质量预测研究。

参考文献

- [1] 陈雪芬. 新形势下环境工程中大气污染的危害和治理[J]. 石化技术, 2025, 32(6): 391-393.
- [2] 王璐. 兰州市 AQI 指数与气象要素相关性研究[J]. 化学工程与装备, 2023(2): 286-287, 301.
- [3] 郑舒天, 朱黎阳, 李自然, 等. 基于 AQI 的文山市空气质量统计预报研究[J]. 中国资源综合利用, 2022, 40(1): 77-80.
- [4] 刘诗韵, 云静波, 姜雨萌, 等. 乌兰察布市空气质量状况及 AQI 预报方法研究[J]. 内蒙古气象, 2023(2): 32-38.
- [5] 牛艳飞. 基于时间序列分析的鹤壁市 AQI 预测模型[J]. 中阿科技论坛(中英文), 2022(4): 84-88.
- [6] Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., et al. (2017) Long Short-Term Memory Neural Network for Air Pollutant Concentration Predictions: Method Development and Evaluation. *Environmental Pollution*, **231**, 997-1004. <https://doi.org/10.1016/j.envpol.2017.08.114>
- [7] 史雨梅, 余晓美. 图卷积神经网络在空气质量指数预测中的应用研究[J]. 三明学院学报, 2024, 41(3): 58-66.
- [8] 文昌俊, 陈洋洋, 何永豪, 等. 多策略 SMA-BP 神经网络的空气质量指数预测[J]. 电子测量技术, 2023, 46(22): 78-86.
- [9] 路凯丽, 杨露, 李涛. 基于集成深度学习模型的空气质量指数预测[J]. 南京信息工程大学学报, 2024, 16(1): 56-65.

附 录

Table S1. Air Quality Index (IAQI) and corresponding pollutant concentration thresholds
附表 1. 空气质量分指数（IAQI）及对应的污染物项目浓度限值

指数或污染物项目	空气质量分指数及对应污染物浓度限值								单位
空气质量分指数(IAQI)	0	50	100	150	200	300	400	500	-
一氧化碳(CO) 24 小时平均	0	2	4	14	24	36	48	60	mg/m ³
二氧化硫(SO ₂) 24 小时平均	0	50	150	475	800	1600	2100	2620	
二氧化氮(NO ₂) 24 小时平均	0	40	80	180	280	565	750	940	
臭氧(O ₃)最大 8 小时滑动平均	0	100	160	215	265	800	-	-	
粒径小于等于 10 μm 颗粒物(PM ₁₀) 24 小时平均	0	50	150	250	350	420	500	600	μg/m ³
粒径小于等于 2.5 μm 颗粒物(PM _{2.5}) 24 小时平均	0	35	75	115	150	250	350	500	

注：(1) 臭氧(O₃)最大 8 小时滑动平均浓度值高于 800 μg/m³ 的，不再进行其空气质量分指数计算。(2) 其余污染物浓度高 IAQI = 500 对应限值时，不再进行其空气质量分指数计算。

Table S2. Air quality grades and corresponding Air Quality Index (AQI) ranges
附表 2. 空气质量等级及对应空气质量指数（AQI）范围

空气质量等级	优	良	轻度污染	中度污染	重度污染	严重污染
空气质量指数(AQI)范围	[0, 50]	[51, 100]	[101, 150]	[151, 200]	[201, 300]	[301, +∞)

注：当 AQI 小于或等于 50 (即空气质量评价为“优”)时，称当天无首要污染物；当 AQI 大于 50 时，IAQI 最大的污染物为首要污染物。若 IAQI 最大的污染物为两项或两项以上时，并列为首要污染物；当 IAQI 大于 100 的污染物为超标污染物。