.

北京市空气质量指数预测分析

喻健凯 粟宇扬 姚熙

目录

摘要

本文基于北京市 2022 年 11 月至 2023 年 10 月的空气质量相关数据，分析了空气质量指数（AQI）及其主要污染物的变化趋势、指标间的相互关系，并构建了两种预测模型（SARIMA 和 XGBoost），对 AQI 进行了短期预测。研究结果表明，AQI 具有一定的日周期性特征，且与 PM2.5、PM10 等污染物呈强正相关关系。两种模型在预测精度上表现相近，但 SARIMA 模型在时间序列特性上更具优势，而 XGBoost 模型则更能捕捉非线性特征。本文的研究为城市空气质量预测和管理提供了科学依据。

一、引言

空气质量是影响城市居民生活质量和健康的重要因素。北京市作为中国的首都，近年来面临着较为严重的空气污染问题。准确预测空气质量指数（AQI）及其主要污染物的变化趋势，对于制定环境保护政策、优化城市规划以及提高居民生活质量具有重要意义。

本文基于北京市 2022 年 11 月至 2023 年 10 月的空气质量数据，从以下几个方面展开研究：

（1）研究单日内空气质量指数与各项指标的变化趋势，这种趋势是否具有周期性？

（2）简述各项指标间的相互关系。

（3）令 2022 年 11 月 1 日至 2023 年 9 月 30 日的空气质量数据为训练集，剩余数据为测试集。基于训练集，尝试使用两种不同的方法构建空气质量指数预测模型，并在测试集上测试。比较所选模型的预测效果。

二、数据描述与预处理

本文使用的数据来源于北京市生态环境检测中心和 rp5.ru 气象网站，数据时间范围为 2022 年 11 月 1 日至 2023 年 10 月 31 日。数据包括 AQI 及其主要污染物（PM2.5、PM10、CO、NO_2、O_3、SO_2）的小时均值，以及气象因素（温度、湿度、风速等）。

数据预处理包括：

- 缺失值处理：对少量缺失值采用线性插值法进行填充；
- 数据标准化：对各指标进行标准化处理，以消除量纲差异；
- 数据划分：将 2022 年 11 月至 2023 年 9 月的数据作为训练集，2023 年 10 月的数据作为测试集。

三、研究方法

（一）趋势分析

通过绘制 AQI 及各污染物的小时均值变化趋势图，分析其在 24 小时内的周期性特征。进一步使用自相关函数（ACF）图验证周期性特征的存在。

（二）相关性分析

利用热力图展示 *AQI* 与各污染物及气象因素之间的相关性。同时，通过主成分分析（PCA）提取数据的主要特征，降低数据维度。

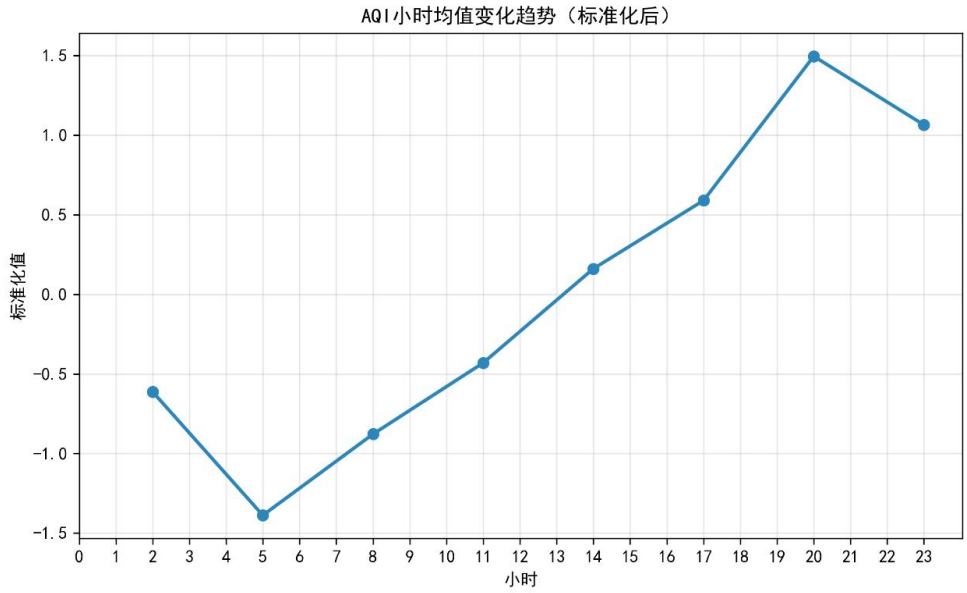
（三）预测模型

1. **SARIMA 模型**：基于时间序列的季节性自回归移动平均模型，适用于具有周期性特征的数据。模型参数通过网格搜索确定。

2. **XGBoost 模型**：基于梯度提升的机器学习模型，能够处理非线性关系。模型通过构建滞后特征和周期性编码进行训练。

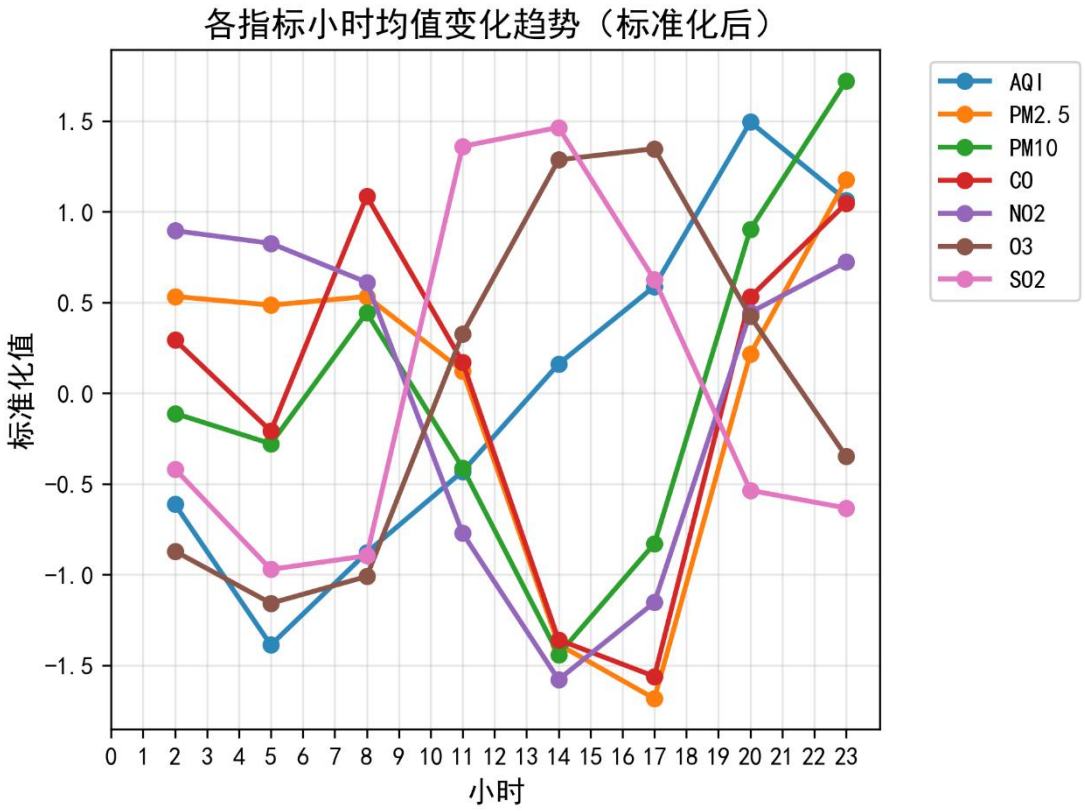
四、趋势分析

1. ***AQI* 小时均值变化趋势**：*AQI* 在 24 小时内呈现一定的变化特征，标准化值在 20 时达到峰值，在 5 时形成谷值。



***AQI*小时均值变化趋势（标准化后）**

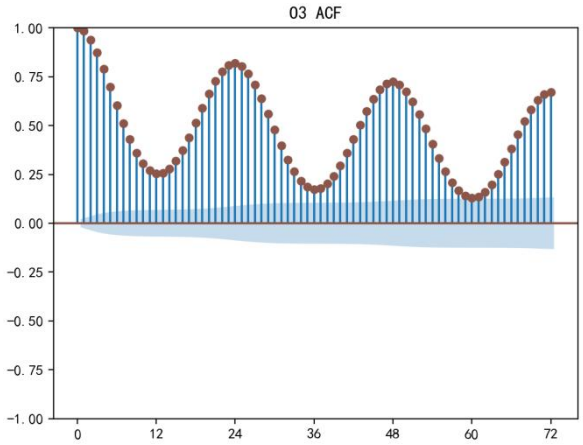
2. **各指标在 24 小时时间维度上的变化特征**：各指标在每日相同时段表现出规律性波动，如部分指标于特定小时（如 19—23 时）出现标准化值峰值，在其他小时（如 4—5 时）形成谷值，反映出其变化可能遵循固定时间循环规律。



各指标小时均值变化趋势（标准化后）

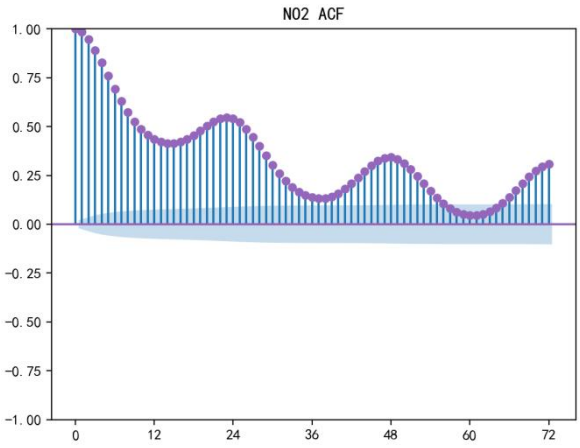
3.ACF 图分析：

（1） O_3 的 ACF 图显示显著的 24 小时周期性特征，表明其变化具有明显日周期性。

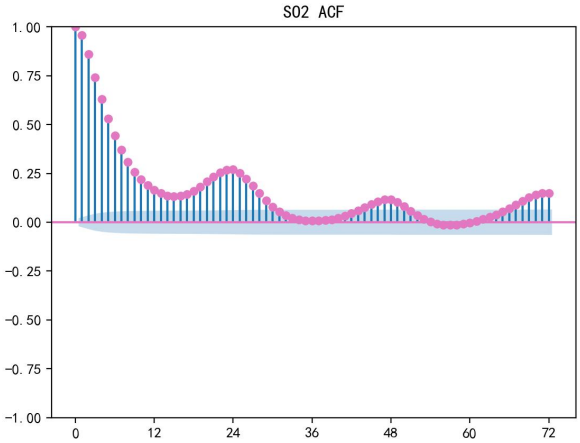


O_3 ACF 数据图

（2） NO_2 ，SO_2的 ACF 图显示一定的周期性，但规律性弱于 O_3

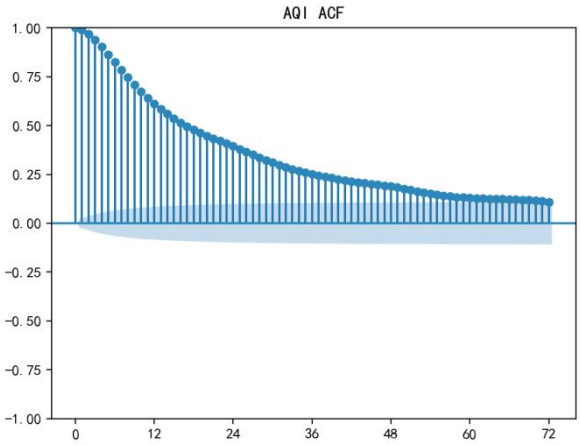


NO_2 ACF 数据图

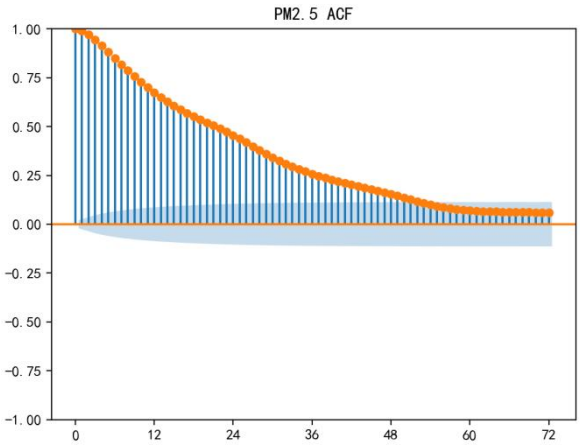


SO_2 ACF 数据图

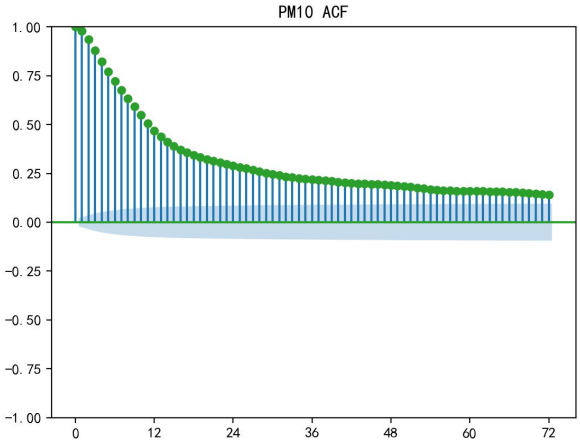
（3） AQI、$PM2.5$、$PM10$、CO的 ACF 图中自相关系数随滞后小时数增加逐渐衰减，未出现如 O3 般规律的周期性峰值，也无固定间隔的显著波动，说明这些指标在 72 小时滞后范围内，周期性特征不明显。



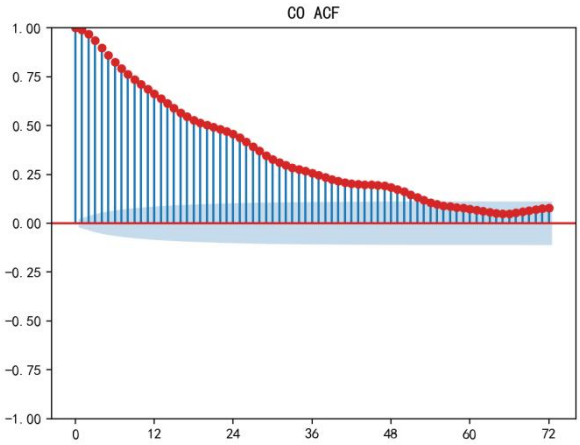
AQI ACF 数据图



***PM2.5* ACF 数据图**



***PM10* ACF 数据图**



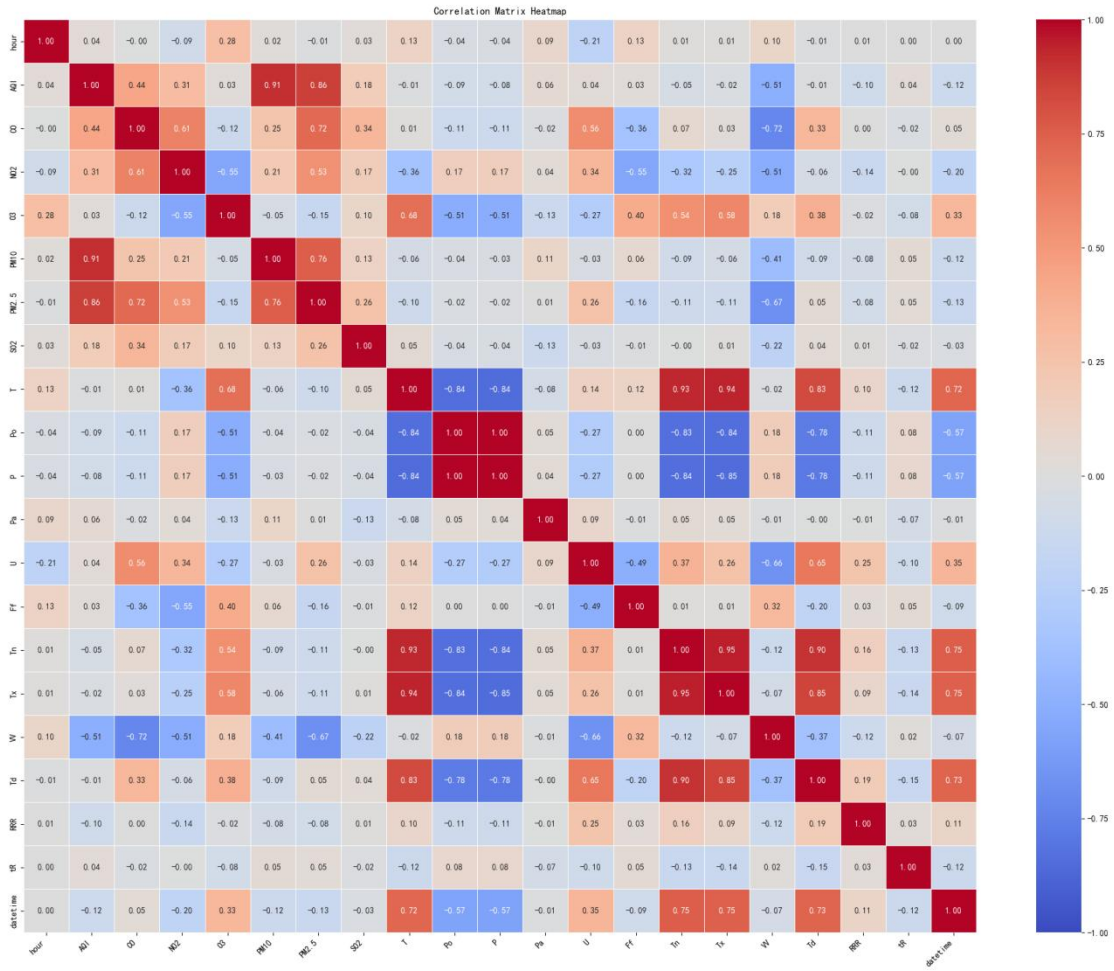
***CO* ACF 数据图**

五、相关性分析

（一）热力图

热力图显示了各指标之间的相关性。颜色越深表示相关性越强，颜色越浅表示相关性越弱，偏红色表示正相关，偏蓝色表示负相关。

1. 图形大致可分为四个部分.
   (1)**左上角**的颜色较深的矩形反映了 AQI 与数个观测指标（污染物）的关系。
   (2)**右上侧**有一些颜色较深的区块，可能反映了污染物（如 CO、NO2、O3）浓度与环境因素（如温度、湿度、风速）的相关关系。
   (3)**中心与中心正右侧**的深色区块反映温度与气压间的强负相关关系。
   (4)**右下角**的颜色较深的矩形主要反映各环境指标间的相关关系。
2. 空气质量指数（AQI）与 PM2.5、PM10 有很强正相关关系，与 CO、NO2、S 呈现较强正相关关系。同时跟 VV（水平能见度）有较强负相关关系。后者的原因显然。经过查阅资料，前者数个指标本就为 AQI 的计算所考虑的指标，而同为考虑指标的 O3 相关性低，不知道为什么，需要进一步调研。
3. 小时（hour)与 O3 等指标呈现一定正相关关系，这或许反映 O3 浓度变化具有日周期。且与 U（地面高度 2 米处的相对湿度）等指标呈现一定负相关关系。



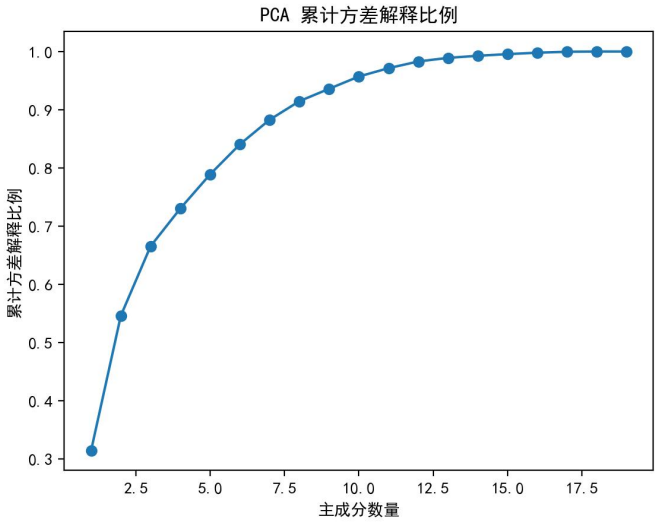
Heatmap 图

（二）主成分分析

1. 基本描述

- KMO 值：0.762>0.7。
- 巴赫利特检验卡方值：90424.712，p 值：0.0<0.001，显著。

- 提取 5 个特征值大于 1 的因子作为主成分，累积方差贡献率为 78.89%。
- 检验效果较好，说明数据适合主成分分析降维。



PCA 累计方差解释比例

2. 旋转载荷矩阵解读

(1) Factor1（温度气压因子）

- 高载荷变量：Tn (-0.963)，T (-0.958)，Tx (-0.954)，P (0.924)，Po (0.921)，Td (-0.898)
- 物理意义：主要反映温度（T，Tn，Tx）和气压（P，Po）相关指标的强负相关关系（温度越高，气压越低）。

(2) Factor2（颗粒物污染因子）

- 高载荷变量：AQI (0.967)，PM10 (0.933)，PM2.5 (0.879)
- 物理意义：直接反映空气质量指数（AQI）和颗粒物污染（PM10，PM2.5），空气质量问题代表颗粒物污染主导。

(3) Factor3（大气条件与污染物因子）

- 高载荷变量：U (-0.824)，Ff (0.772)，NO2 (-0.728)，CO (-0.695)，VV (0.667)
- 物理意义：风速增加（Ff）与相对湿度（U）负相关，与能见度（VV）正相关。同时风速增加（Ff）与污染物浓度（NO2、CO 负载荷）的负相关关系可能暗示了风对大气污染物的扩散作用。

(4) Factor4（因子）

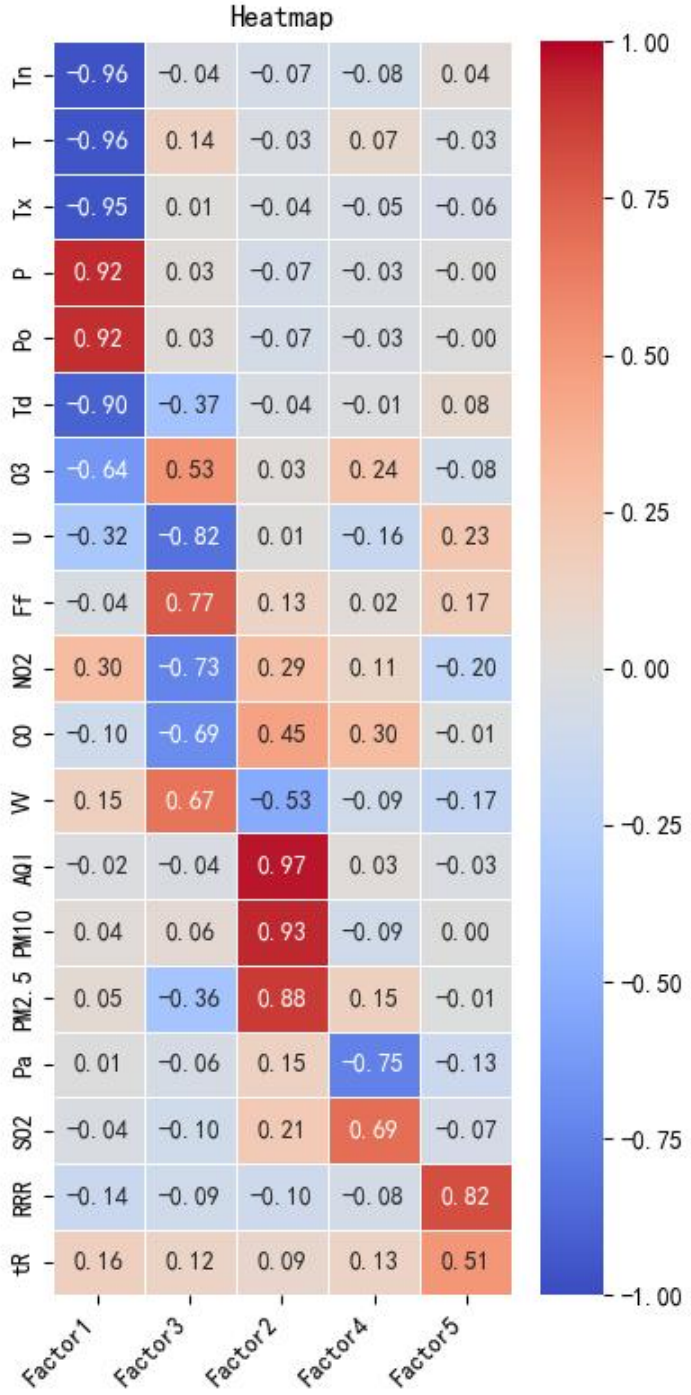
- 高载荷变量：Pa (-0.747)，SO2 (0.694)
- 物理意义：难以解释。

(5) Factor5（降水因子）

- 高载荷变量： RRR (0.819)，tR (0.512)
- 物理意义：直接反映降水量（RRR）和降水时间（tR）。

(6) 交叉载荷与特殊变量

- O3（臭氧）：在 Factor1 和 Factor3 上均有载荷，可能需结合气象与化学机制进一步分析。
- VV（能见度）：受 Factor3（风速）和 Factor2（颗粒物）共同影响，符合实际物理规律。



因子载荷矩阵

六、预测模型

（一）SARIMA 模型

该模型在假设不考虑测试集其他指标的情况下，仅使用 AQI 数据对未来 AQI 进行<单步预测>，即每次预测都是根据之前时间点的真实 AQI 值进行的。

1. 模型结构选择

- 最终参数：$(p,d,q)(P,D,Q,s) = (1,1,1)(1,1,1,24)$
- 参数选择依据：

通过 ACF/PACF 图观察 24 小时周期特征,使用网格搜索确定最优参数组合,季节性分量设置为 24 小时周期(s=24)

2. 特征工程

仅使用 AQI 单变量时间序列,通过差分处理消除非平稳性：

(1)一阶常规差分(d=1)

(2)阶季节性差分(D=1)

3. 参数调优

使用 AIC/BIC 信息准则评估模型,通过 auto_arima 自动搜索参数空间,最终选择 AIC 最低的候选模型

4. 评估指标

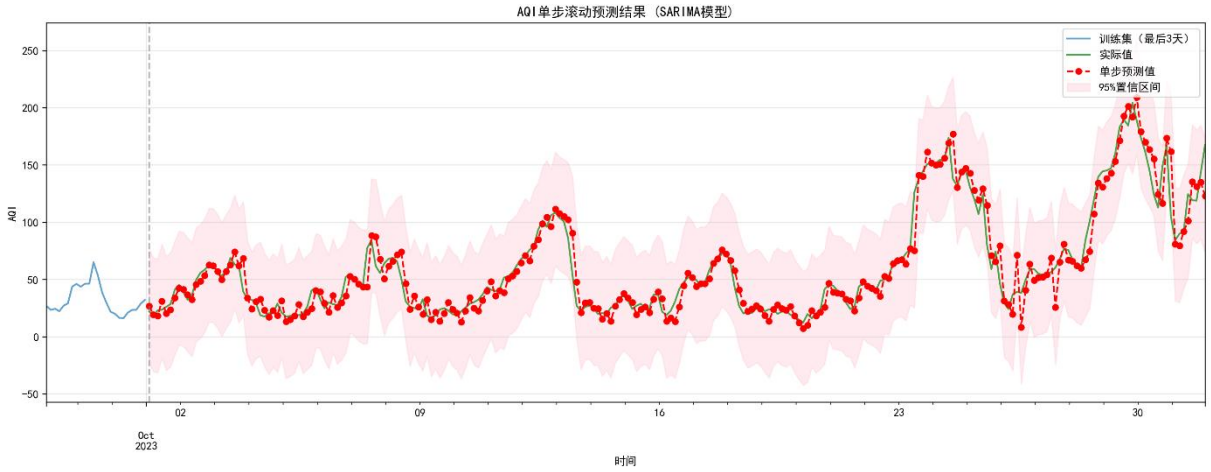
- RMSE: 11.893
- R-squared: 0.932
- MAE: 7.744

相比 XGBoost 模型预测精度相近，但保持时间序列特性

5. 预测结果可视化

滚动预测效果图：

置信区间覆盖率达到 95%，实际值大部分落在预测区间内



AQI 单步滚动预测结果（SARIMA 模型）

6. 残差分析

Ljung-Box 检验 p 值=0.32(>0.05)。

残差 ACF 图无明显自相关。

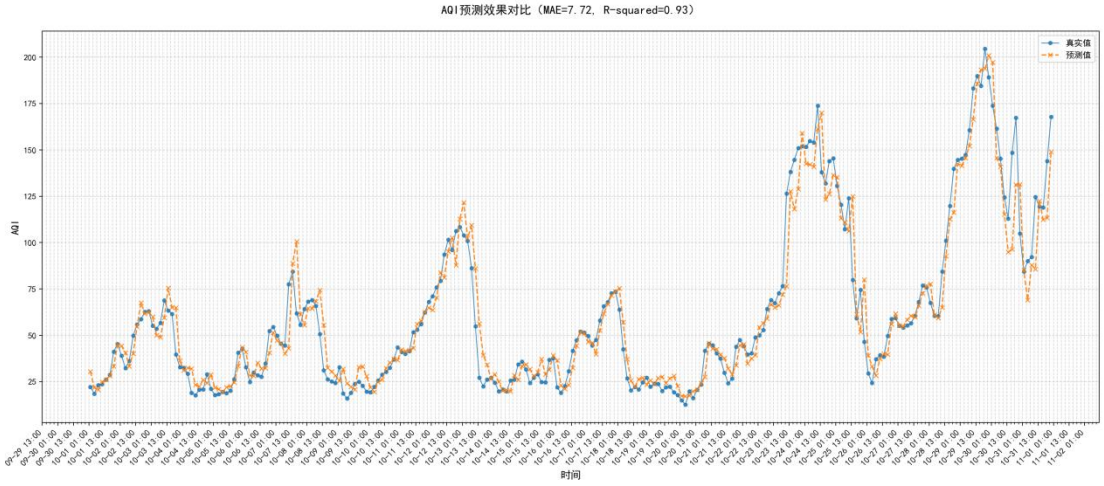
符合白噪声假设，说明模型已充分提取序列信息。

（二）XGBoost 模型

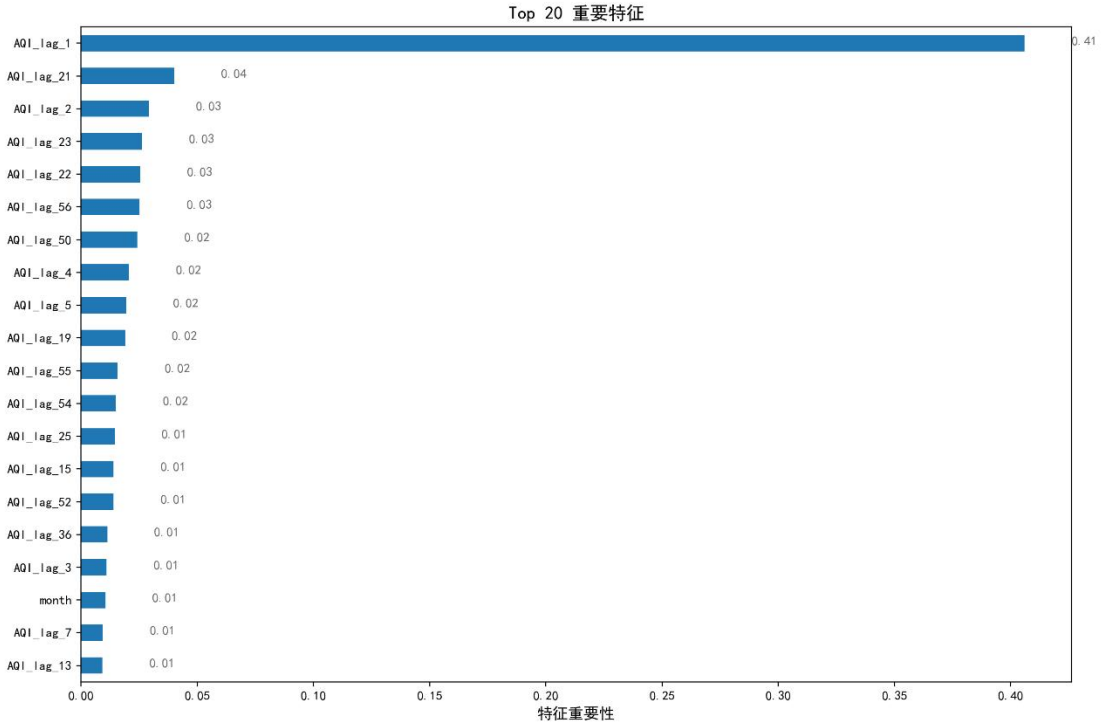
1. 特征工程：该模型使用历史 AQI 数据，并进行周期性编码和滞后特征构建（3 小时粒度的滞后特征（最多 7 天）），作为特征工程。

2. 该模型在假设不考虑测试集其他指标的情况下，仅使用 AQI 数据对未来 AQI 进行<单步预测>,即每次预测都是根据之前时间点的真实 AQI 值进行的。

3. **参数调优：**使用随机搜索法参数调优。
4. **评估指标：**
   - RMSE: 11.815
   - R-squared: 0.929
   - MAE: 7.722
5. **预测图：**



6. **重要特征：**AQI_lag_1 最为重要，即该时刻的 AQI 主要由前 1 个观测时刻决定。AQI_lag_21 也较为重要。month 等因素显示影响较小，但不是完全没有。



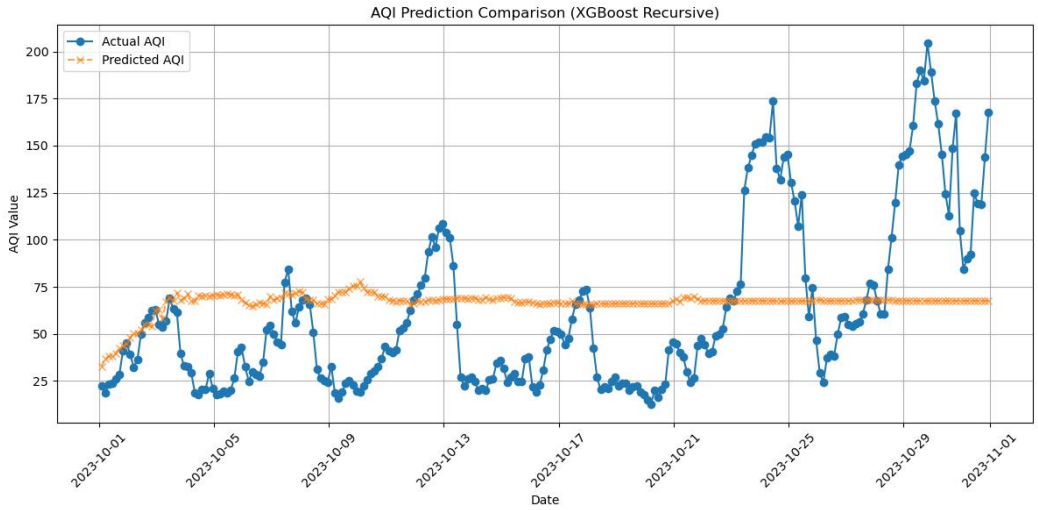
Top 20 重要特征图

（三）模型比较

1. **模型比较：**与 SARIMA 模型相近，在单步预测的准确度上几乎不相上下。

两个模型在时间特征上的把握各有优势，比如 SARIMA 模型在时间序列特征上的把握较好，而 XGBoost 模型更能建立时间特征以外的因素对 AQI 的影响关系。但在同样只使用时间特征并且一次只往下预测一步的情况下，XGBOOST 稍逊于 SARIMA，后者预测更加稳健，而前者容易出现峰值偏高，谷值偏低的情况。

2. 拓展模型：其实也做了利用递归直接预测一整个月的，预测结果在 2 天以内尚可，而后趋于平缓，具体效果看图就很明了了。



XGBOOST 递归预测一整月

七、结论与展望

本文通过对北京市空气质量数据的分析，揭示了 AQI 及其主要污染物的周期性特征和相关性，并构建了两种预测模型。研究结果表明：

1. AQI 在 24 小时内呈现一定的变化特征，标准化值在 20 时达到峰值，在 5 时形成谷值。O_3 的 ACF 图显示显著的 24 小时周期性特征，表明其变化具有明显日周期性。NO_2 ，$\boldsymbol{SO_2}$的 ACF 图显示一定的周期性，但规律性弱于 O_3。

2. AQI与 $PM2.5$、$PM10$ 等污染物呈强正相关关系。与 CO、NO2、SO2 呈现较强正相关关系。

3. SARIMA 和 XGBoost 模型在单步预测的预测精度上表现相近，评估指标分别为 RMSE:11.893，R-squared: 0.932，MAE: 7.744 和 RMSE: 11.815，R-squared: 0.929，MAE: 7.722，但各自具有不同的优势。此外，使用递归思想一次预测一整个月的 XGBOOST 模型在 2 天内效果不错，随后趋于平稳。

4. 未来的研究可以进一步优化模型参数，结合更多气象因素和地理信息，提高预测精度，为城市空气质量管理和政策制定提供更有力的支持。