

基于季节性 ARIMA 模型的上海空气污染指数时序分析

林元莘

2024 年 5 月 12 日

摘要

本次报告主要针对上海市 2014 年-2022 年的空气质量指数数据进行了深入分析，包括探索性数据分析、时间序列建模以及模型诊断。通过探索性数据分析，我们发现空气质量指数在不同年份内和不同季节内的变化规律，为后续时间序列建模提供了重要参考。通过时间序列建模，我们为每个空气污染指数建立了合适的 ARIMA 模型，并且对 2023 年以后的数据进行了预测。通过模型诊断，我们对模型的残差进行了正态性检验和自相关性检验，发现大部分模型的残差都通过了检验。在建模过程中，我们发现 SO₂ 的一阶差分并不平稳，通过对数变换后，建立了新的 ARIMA 模型，该模型的残差通过了正态性检验和自相关性检验。通过本次报告，我们对时间序列分析的方法有了更深入的了解，同时也对空气质量指数的变化规律有了更清晰的认识。

关键词：时间序列；ARIMA；假设检验

目录

| | |
|---------------------------------------|-----------|
| 1 数据集简介 | 1 |
| 1.1 数据集背景 | 1 |
| 1.2 数据集特征介绍 | 1 |
| 2 探索性数据特征分析 | 2 |
| 2.1 空气污染指数年内分布 | 2 |
| 2.2 空气污染指数年际分布 | 2 |
| 3 时间序列建模 | 4 |
| 3.1 时间序列图以及平稳性检验 | 4 |
| 3.2 ACF 与 PACF | 7 |
| 3.3 模型识别 | 9 |
| 3.4 模型预测 | 11 |
| 3.5 模型诊断 | 12 |
| 3.5.1 正态性检验 | 13 |
| 3.5.2 自相关性检验 | 14 |
| 3.6 讨论与反思 | 15 |
| 3.6.1 SO ₂ 的重新建模 | 15 |
| 4 总结与反思 | 19 |
| A 附录 | 19 |

1 数据集简介

1.1 数据集背景

本次报告所使用数据集来自 the World Air Quality Index project(世界空气质量指数项目) 网站上的实时数据 Shanghai Air Pollution: Real-time Air Quality Index。”The World Air Quality Index Project” 是一个全球性的空气质量监测项目，提供实时的空气污染数据。这个项目由一个独立的环保组织发起，目的是增加公众对空气污染问题的认识。项目利用全球范围内数千个监测站的数据，测量包括 PM2.5、PM10、二氧化硫、二氧化氮等多种空气污染物的浓度。数据通过项目的网站和移动应用程序提供给公众，支持多种语言，使得全球用户都能轻松获取和理解空气质量信息。

此外，这个项目还提供了一个开放数据平台，允许研究人员和开发者访问历史和实时数据，促进环境科学的研究和应用的发展。数据的准确性和时效性使得它成为研究环境政策、公共健康以及城市规划等领域的重要资源。

该数据集包括了从 2014 年 1 月 1 日到 2024 年 5 月 5 日上海市每日空气质量各项指数数据，包括 PM2.5、PM10、SO2、NO2、O3 等。

1.2 数据集特征介绍

本次报告主要针对中从 2014 年 1 月 1 日到 2024 年 5 月 5 日上海市每日空气质量各项指数数据进行数据分析。

该数据集含有 3694 行 *7 列数据。这个数据集主要包含了每日的空气质量指数数据，包括日期、PM2.5、PM10、SO2、NO2、O3、CO 等。数据集中的变量主要包括：

表 1: 数据集特征介绍

| 特征名称 | 特征意义 | 数据类型 |
|------|----------|------|
| date | - | 日期 |
| pm25 | PM2.5 指数 | 浮点数 |
| pm10 | PM10 指数 | 浮点数 |
| o3 | 臭氧指数 | 浮点数 |
| no2 | 二氧化氮指数 | 浮点数 |
| so2 | 二氧化硫指数 | 浮点数 |

| | | |
|----|--------|-----|
| co | 一氧化碳指数 | 浮点数 |
|----|--------|-----|

2 探索性数据特征分析

在深入到复杂的统计分析之前，我们首先尽量对数据集进行了详尽的探索性数据分析。这一步骤对于揭示数据的内在结构、识别关键变量以及检测任何异常或异常模式至关重要。EDA 不仅帮助我们更好地理解数据，还为之后特征工程以及选择合适的预测模型指明了方向。

2.1 空气污染指数年内分布

根据常识，空气污染指数随季节变化比较明显。由此，我们考虑查看每个空气污染指数在不同年内的变化情况。如图1所示

由图可知，每项空气污染指数在不同年份内的变化趋势大致相同，均呈现出明显的季节性变化。

- PM2.5、PM10、SO2、NO2、CO 等空气污染指数在冬季和春季较高，夏季和秋季较低。
- 臭氧（O3）指数在夏季较高，冬季较低。

2.2 空气污染指数年际分布

考虑查看每个空气污染指数年际变化情况。由于 2024 年只有到五月的数据，所以删去 2024 年的数据进行分析。如图2所示

由图可知，每项空气污染指数年际变化大部分较为明显、

- PM2.5、PM10、SO2、NO2、CO 等空气污染指数在 2022 年前呈现明显下降趋势。
- 臭氧（O3）指数总体变化不呈现趋势。
- 2022 年后部分空气污染指数呈现上升趋势。

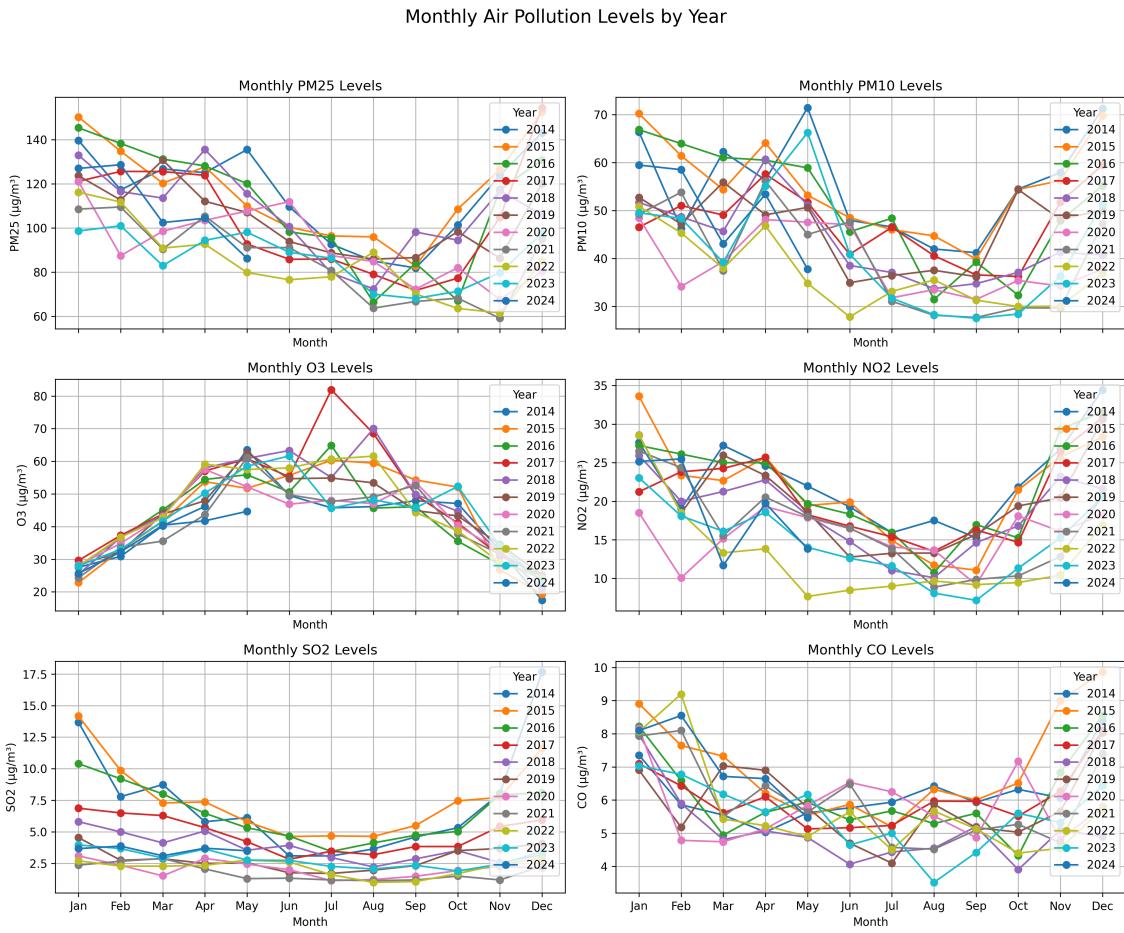


图 1: 空气污染指数随年份变化

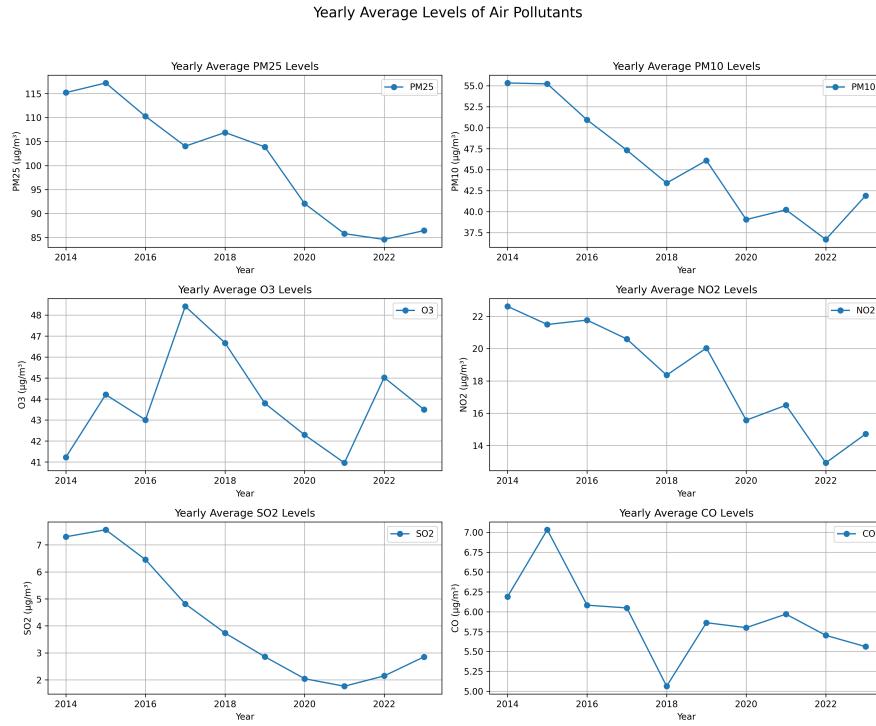


图 2: 每年平均空气污染指数

由上，时间序列建模需要更具有季节性，并且有一定的趋势效果。

3 时间序列建模

3.1 时间序列图以及平稳性检验

作出每个空气污染指数的时间序列图，如图3所示同时，对每个时间序列图进行 ADF 单位根检验

ADF 单位根检验一种用于检查时间序列是否具有单位根，从而判断序列是否非平稳的统计方法。由以下思路构成

- **模型设定：** ADF 检验函数会尝试以下常用时间序列模型，常数趋势以及线性趋势

$$\nabla Z_t = \alpha + (\beta t +) \gamma Z_{t-1} + \sum_{i=1}^p \phi_i \nabla Z_{t-i} + a_t$$

- **检验统计量：** ADF 测试关注的是 Z_{t-1} 的系数 γ 。如果这个系数显著不为 0，则推断序列没有单位根，是平稳的。即

$$H_0 : \gamma = 0 \text{(存在单位根, 序列非平稳)}, H_1 : \gamma < 0 \text{(不存在单位根, 序列平稳)}$$

- **滞后阶数选择:** 实际上, 模型会根据 AIC(赤池信息准则) 来选择合适的滞后阶数 p 。

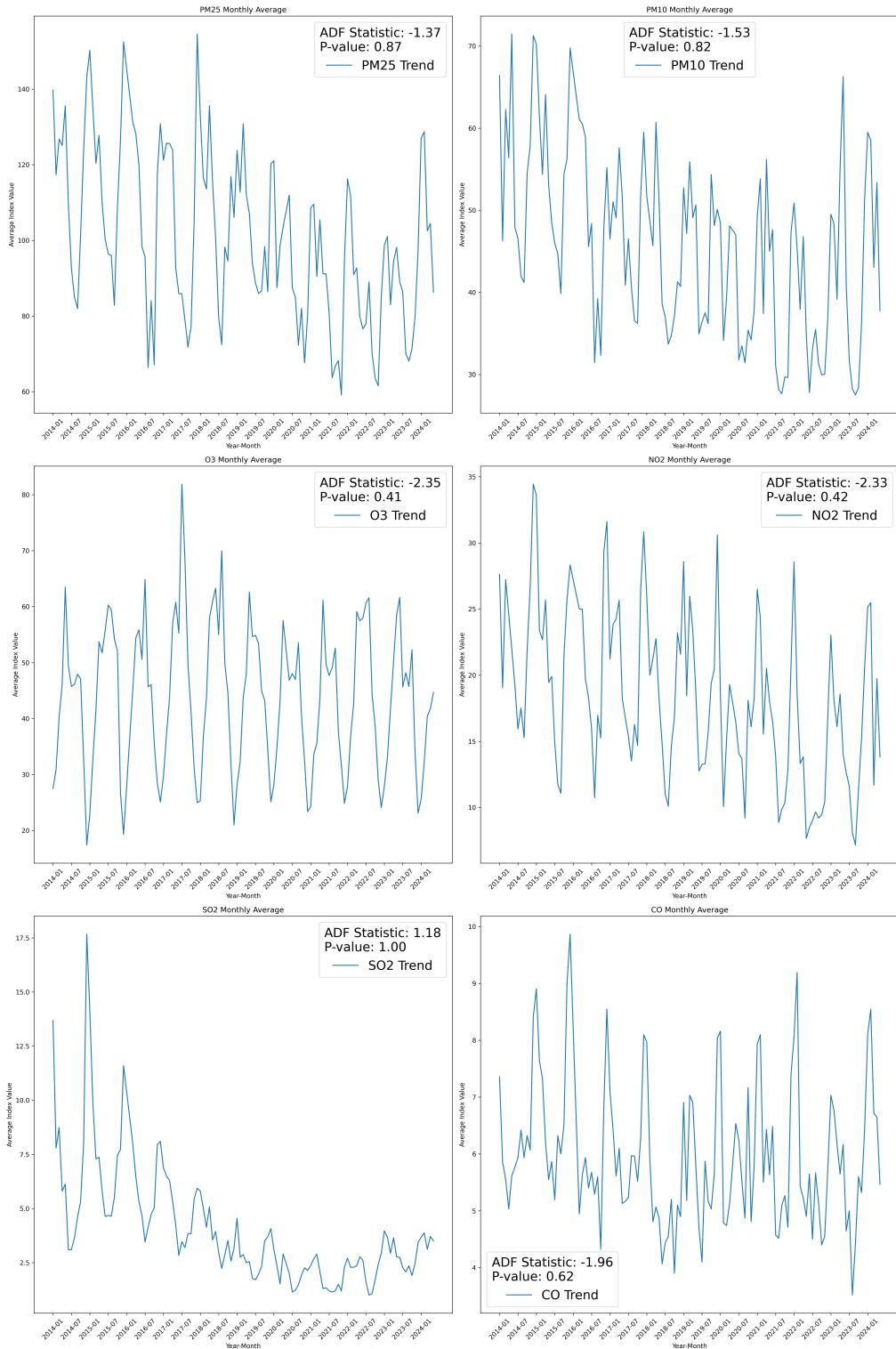


图 3: 空气污染指数时间序列图以及单位根检验

表 2: ADF 单位根检验

| 特征名称 | ADF 统计量 | P-值 |
|------|---------|------|
| pm25 | -1.37 | 0.87 |
| pm10 | -1.53 | 0.82 |
| o3 | -2.35 | 0.41 |
| no2 | -2.33 | 0.42 |
| so2 | 1.18 | 1 |
| co | -1.96 | 0.62 |

所有列出的变量的 P-值都远高于常用的显著性水平（如 0.05, 0.01）。这意味着，对于所有列出的污染物（PM2.5, PM10, O3, NO2, SO2, CO），我们没有足够的证据拒绝原假设 H_0 。因此，基于这些 ADF 检验结果，我们不能认为这些时间序列是平稳的。

于是考虑一阶差分，即令 $Y_t = \nabla Z_t$ ，考虑 $\{Y_t\}$ 时间序列的平稳性。

同样的，我们做出一阶差分后的时间序列图，以及 ADF 单位根检验的结果。

表 3: 一阶差分后 ADF 单位根检验

| 一阶差分的特征名称 | ADF 统计量 | P-值 |
|-----------|---------|----------|
| pm25 | -9.78 | 6.80e-17 |
| pm10 | -9.07 | 4.25e-15 |
| o3 | -9.96 | 2.33e-17 |
| no2 | -11.46 | 5.71e-21 |
| so2 | -2.97 | 0.03 |
| co | -8.57 | 8.02e-14 |

- **P-值极低**除了 SO2 外，所有变量的 P-值非常低，远低于通常的显著性水平(0.05, 0.01)。这意味着对于 PM25, PM10, O3, NO2, 和 CO, 我们有充足的证据在统计上拒绝原假设，认为这些序列在一阶差分后是平稳的。
- **SO2 的情况** SO2 的 P-值为 0.03，也低于 0.05 的显著性水平，表明即使这个变量的统计量不如其他变量那么极端，我们仍然有足够的理由认为其在一阶差分后是平稳的。

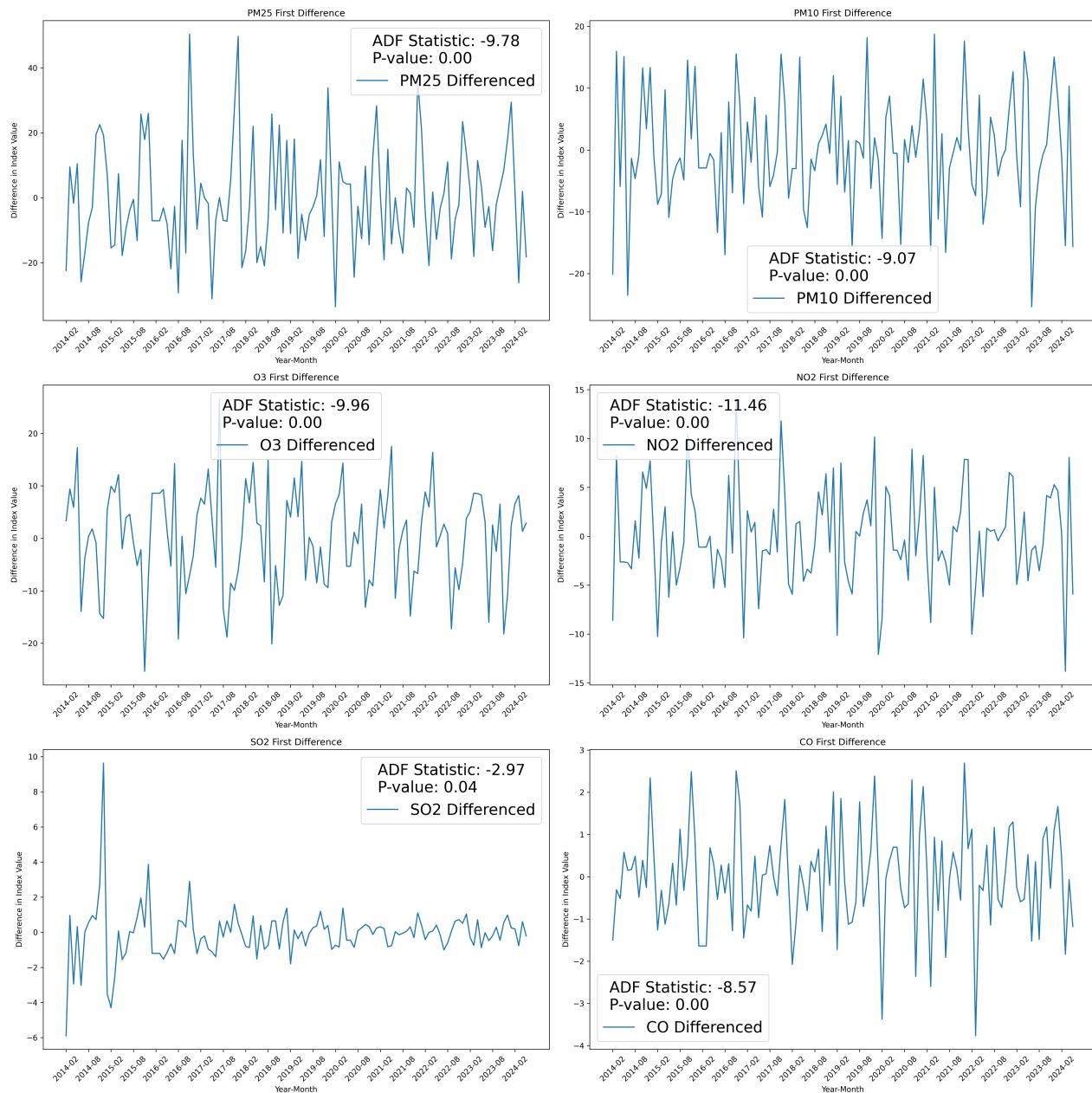


图 4: 一阶差分后空气污染指数时间序列图以及单位根检验

3.2 ACF 与 PACF

我们绘制每个变量及其一阶差分后的自相关函数和偏自相关函数图5，并且有季节性差分后的图6。从图中可以发现

- 除了 O3 以外，每个污染指数在一阶差分后呈现快速衰减性质。并且出现季节性。
- O3 在一阶差分后 ACF 仍然缓慢衰减，考虑继续进行差分。

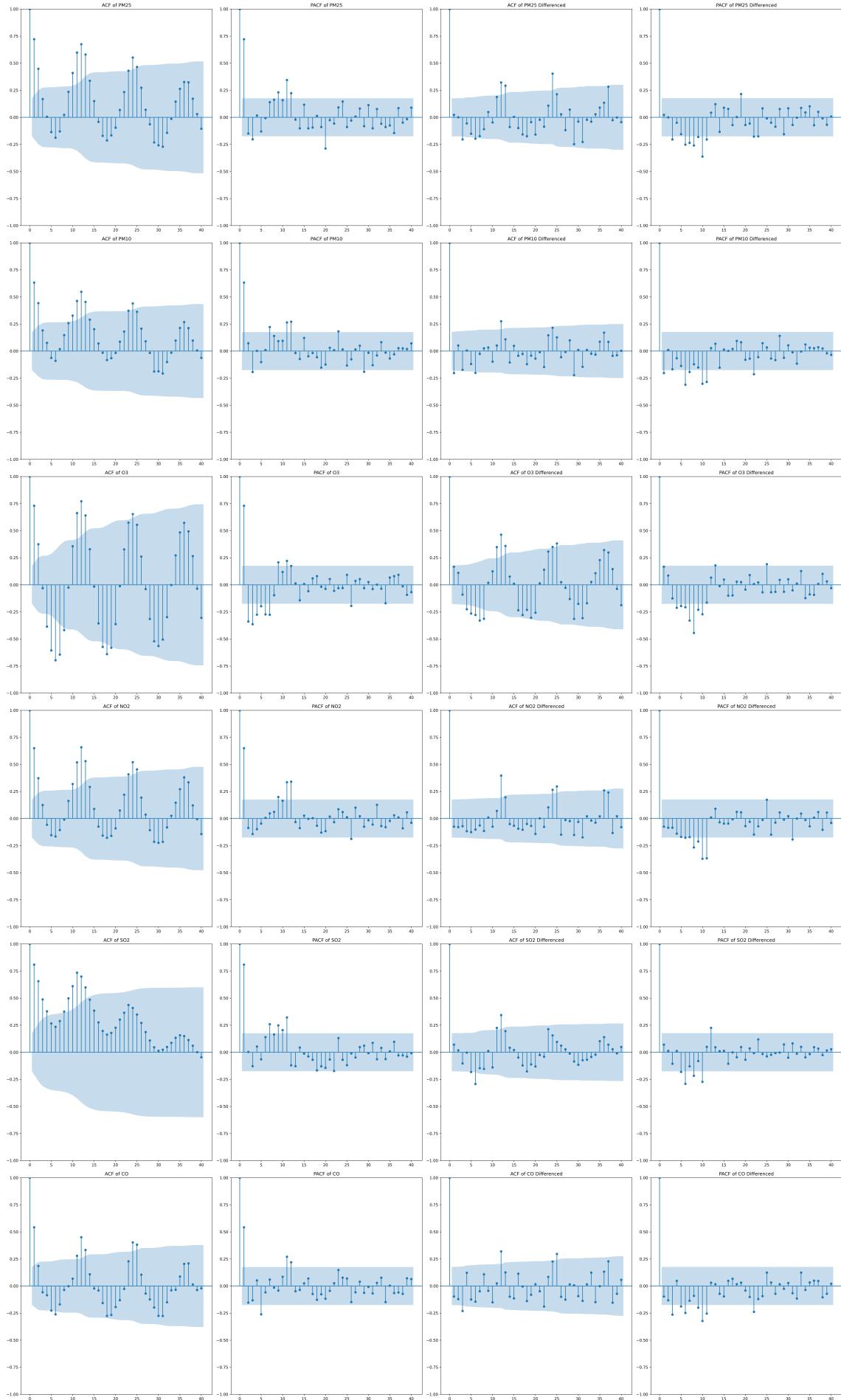


图 5: ACF 与 PACF 变化图

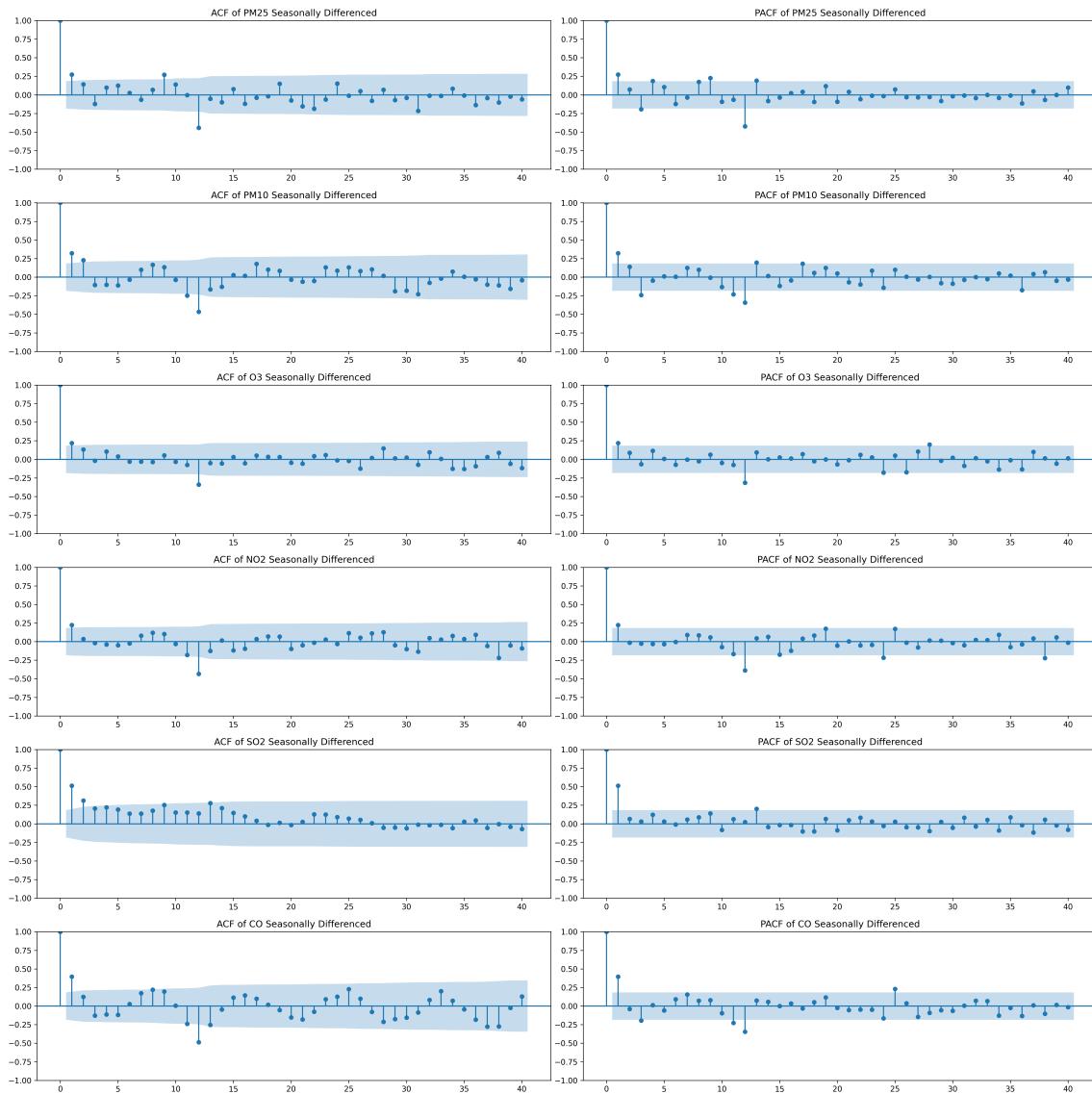


图 6: 季节性差分 ACF 与 PACF 变化图

3.3 模型识别

根据以上各种 ACF 与 PACF 图，我们可以分别为每个空气污染指标提供推荐模型

1. PM2.5

- 样本 ACF 下降缓慢，PACF 在初期截尾，显示季节性峰值。
- 一阶差分在滞后为 3 后非季节性处迅速衰减，仍有季节效应。
- 季节差分 ACF 与 PACF 都在滞后为 1 以后截尾且季节处突起。
- 推荐： $ARIMA(3, 1, 1) \times (1, 1, 1)_{12}$

2. PM10

- 样本 ACF 下降缓慢，PACF 在初期截尾，显示季节性峰值。
- 一阶差分 ACF 在滞后为 1 后截尾，MA 在滞后为 1 后截尾。
- 季节差分 ACF 与 PACF 都在滞后为 1 以后截尾。季节处突起。
- 推荐： $ARIMA(0, 1, 1) \times (1, 1, 0)_{12}$

3. O3

- 样本 ACF 下降缓慢，PACF 在滞后为 6 后截尾，显示季节性峰值。
- 一阶差分 ACF 在滞后为 1 时截尾。
- 季节差分 ACF 与 PACF 都在滞后为 1 以后截尾季节处突起。
- 推荐： $ARIMA(1, 1, 0) \times (1, 1, 0)_{12}$

4. NO2

- 样本 ACF 下降缓慢，PACF 在滞后为 1 后截尾，显示季节性峰值。
- 一阶差分 ACF 与 PACF 在初期截尾，展现周期性
- 季节性差分滞后为 1 截尾并且季节处突起
- 推荐： $ARIMA(0, 1, 1) \times (1, 1, 1)_{12}$

5. SO2

- 样本 ACF 下降缓慢，PACF 在滞后为 1 后截尾，显示季节性峰值。
- 一阶差分 ACF 与 PACF 在初期截尾，展现周期性
- 季节性差分滞后为 1 截尾,PACF 周期处突出
- 推荐： $ARIMA(0, 1, 0) \times (0, 1, 1)_{12}$

6. CO

- 样本 ACF 下降缓慢，PACF 在滞后为 1 后截尾，显示季节性峰值。

- 一阶差分 ACF 与 PACF 在初期截尾，展现周期性
- 季节性差分滞后为 1 截尾并且季节处突起
- 推荐： $ARIMA(0, 1, 0) \times (1, 1, 1)_{12}$

3.4 模型预测

我们采用 2014 年-2022 年的数据，来预测 2023 年以后的数据变化

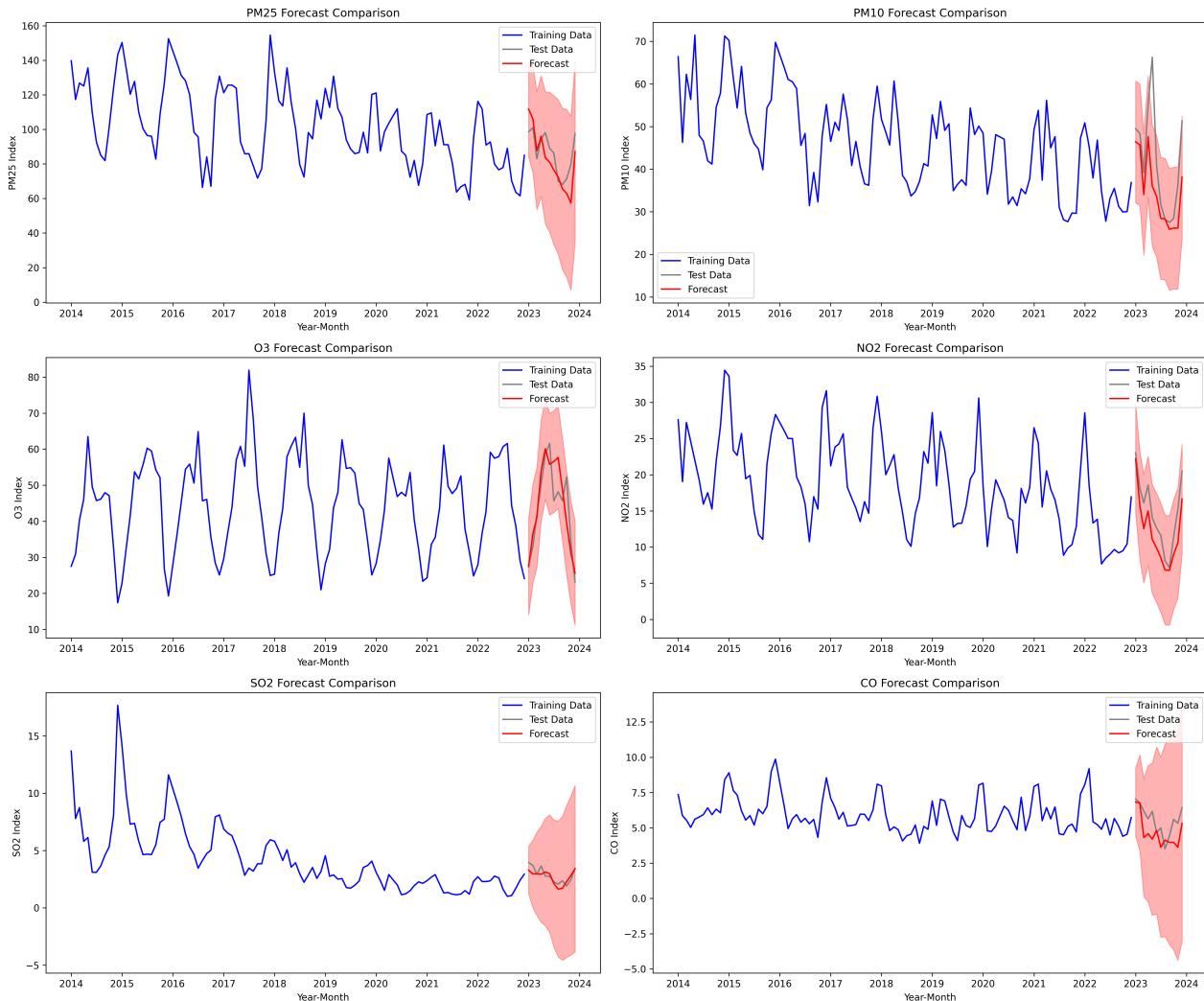
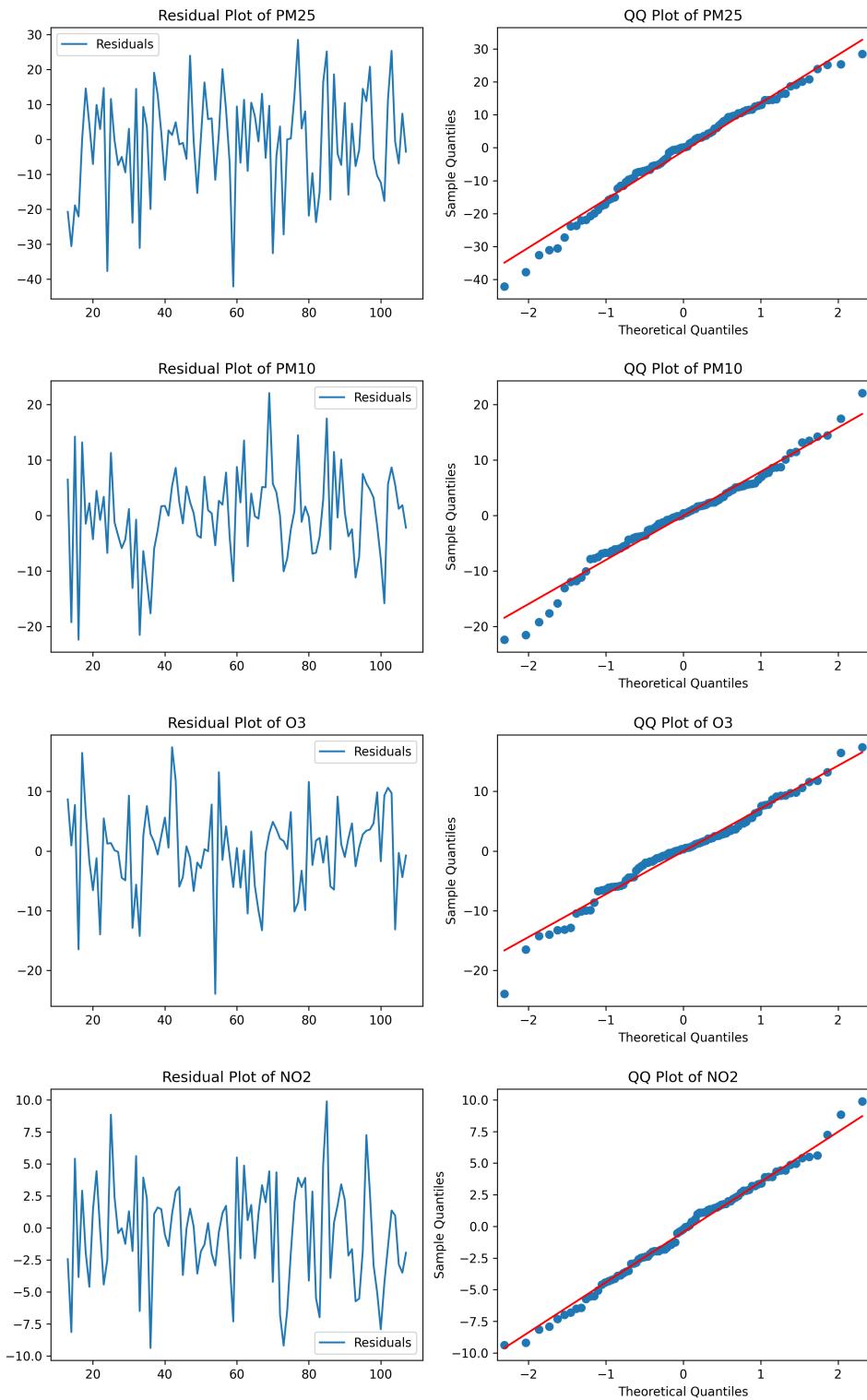


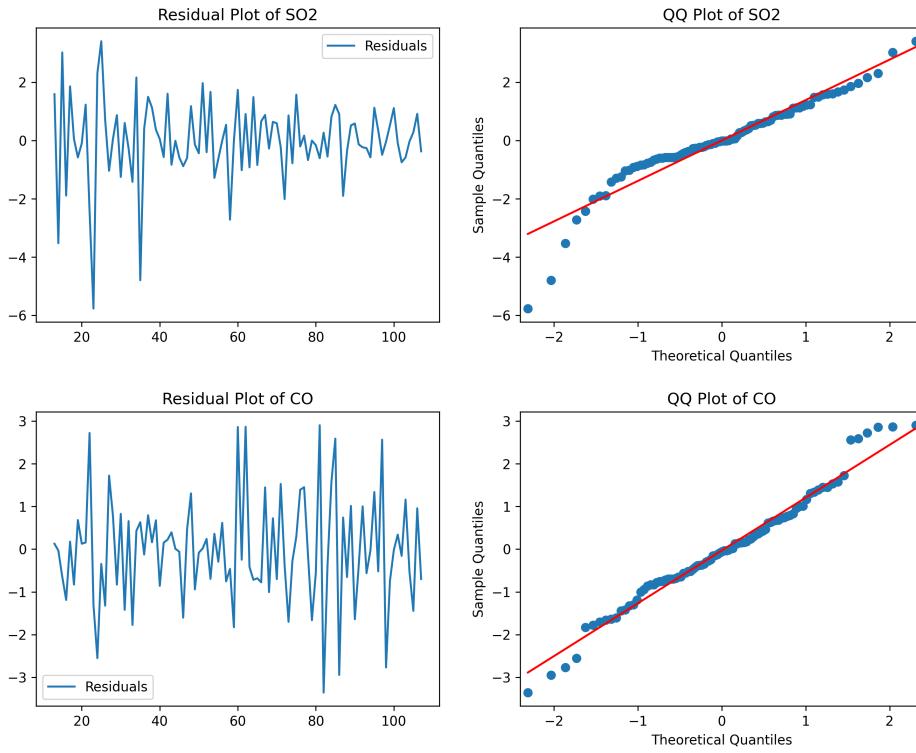
图 7：空气污染指数预测及其置信区间图

可以看到，大部分模型都能较为准确地刻画出 2023 年空气污染指数的变化趋势。

3.5 模型诊断

我们对残差进行分析，画出残差序列的时间图以及 Q-Q 图??，以检验模型的合理性。





3.5.1 正态性检验

我们对残差进行正态性的 Shapiro-Wilk 检验⁴。该假设为 H_0 ：该数据为正态分布。满足统计值：靠近 1 表示数据更接近正态分布。P 值：高于某个显著性水平（通常为 0.05 或 0.01）时，我们不能拒绝原假设，认为数据是正态分布的。

表 4: Shapiro-Wilk 检验

| 特征名称 | Shapiro-Wilk 统计量 | P-值 |
|------|------------------|----------|
| pm25 | 0.98 | 0.089 |
| pm10 | 0.98 | 0.174 |
| o3 | 0.98 | 0.193 |
| no2 | 0.99 | 0.823 |
| so2 | 0.93 | 4.506e-5 |
| co | 0.98 | 0.123 |

由图??和表4知，除 SO2 以外，所有空气污染指数的模型残差，都通过了正态性检验。

3.5.2 自相关性检验

对残差的 ADF 进行分析，作出残差的 ADF 图

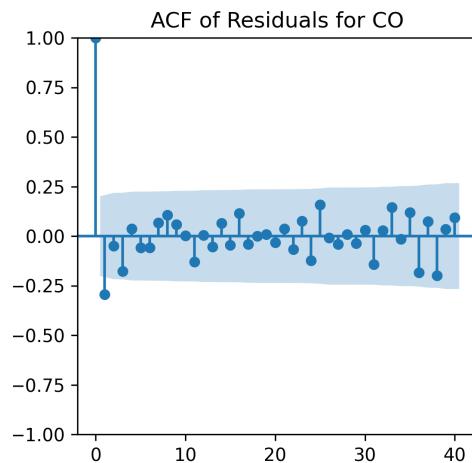
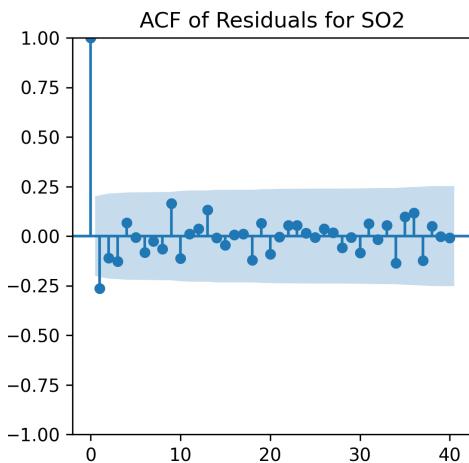
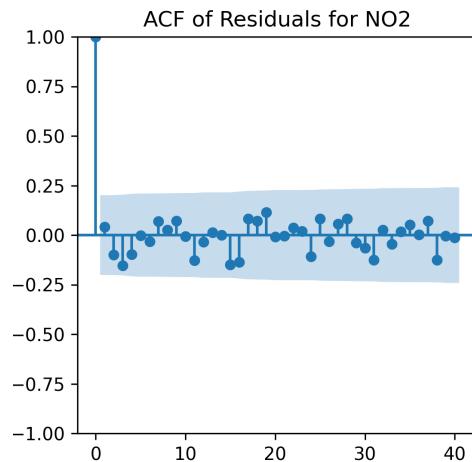
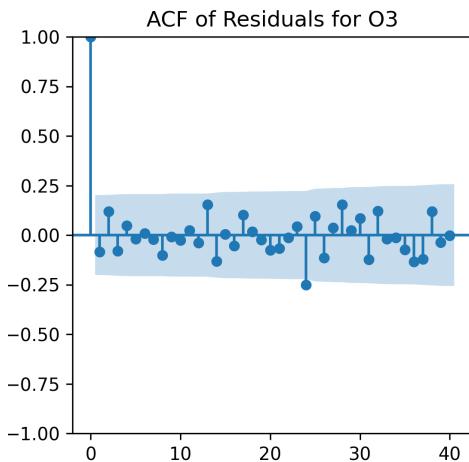
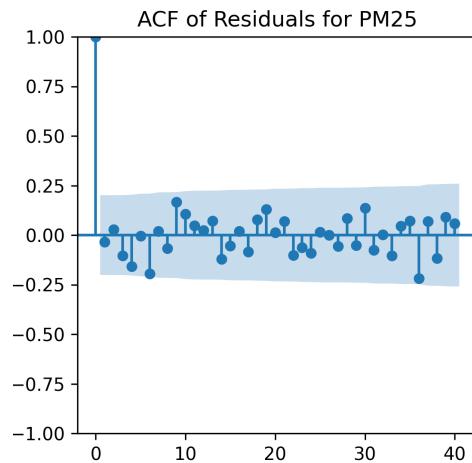
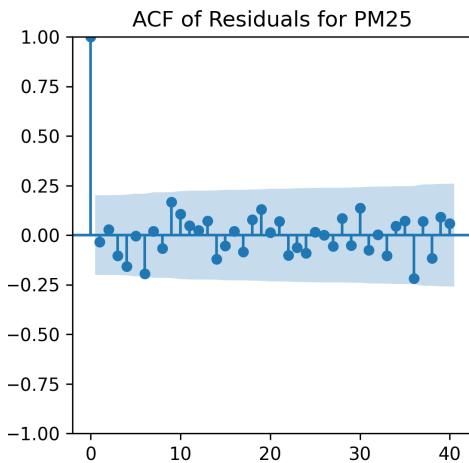


表 5: Ljung-Box 检验

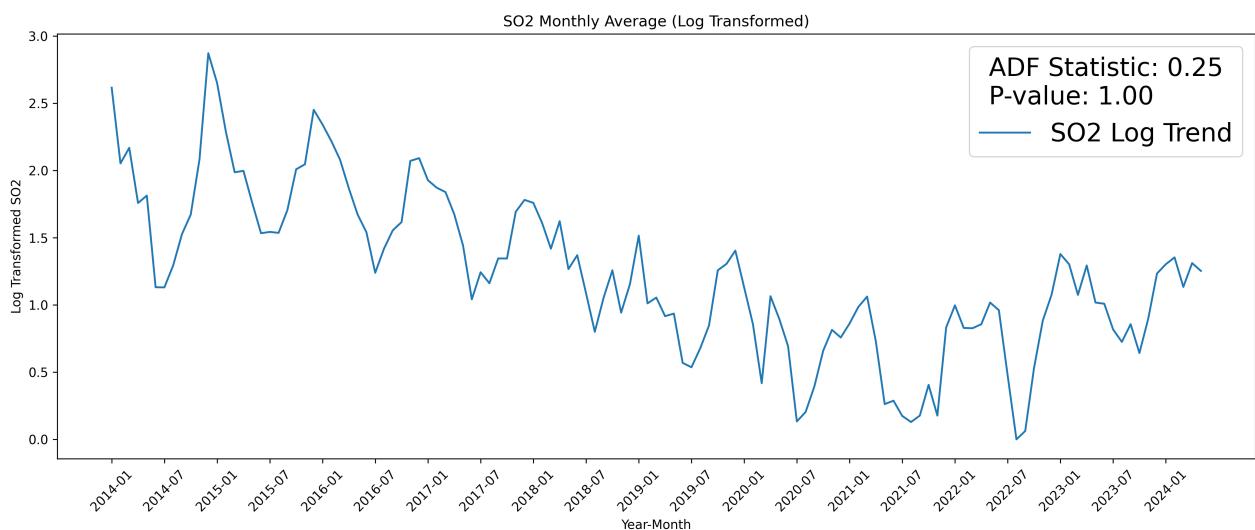
| 特征名称 | Ljung-Box 统计量 | P-值 |
|------|---------------|-------|
| pm25 | 21.84 | 0.589 |
| pm10 | 30.14 | 0.180 |
| o3 | 20.477 | 0.669 |
| no2 | 17.07 | 0.846 |
| so2 | 22.03 | 0.577 |
| co | 22.88 | 0.527 |

所有模型残差都通过了自相关性检验。

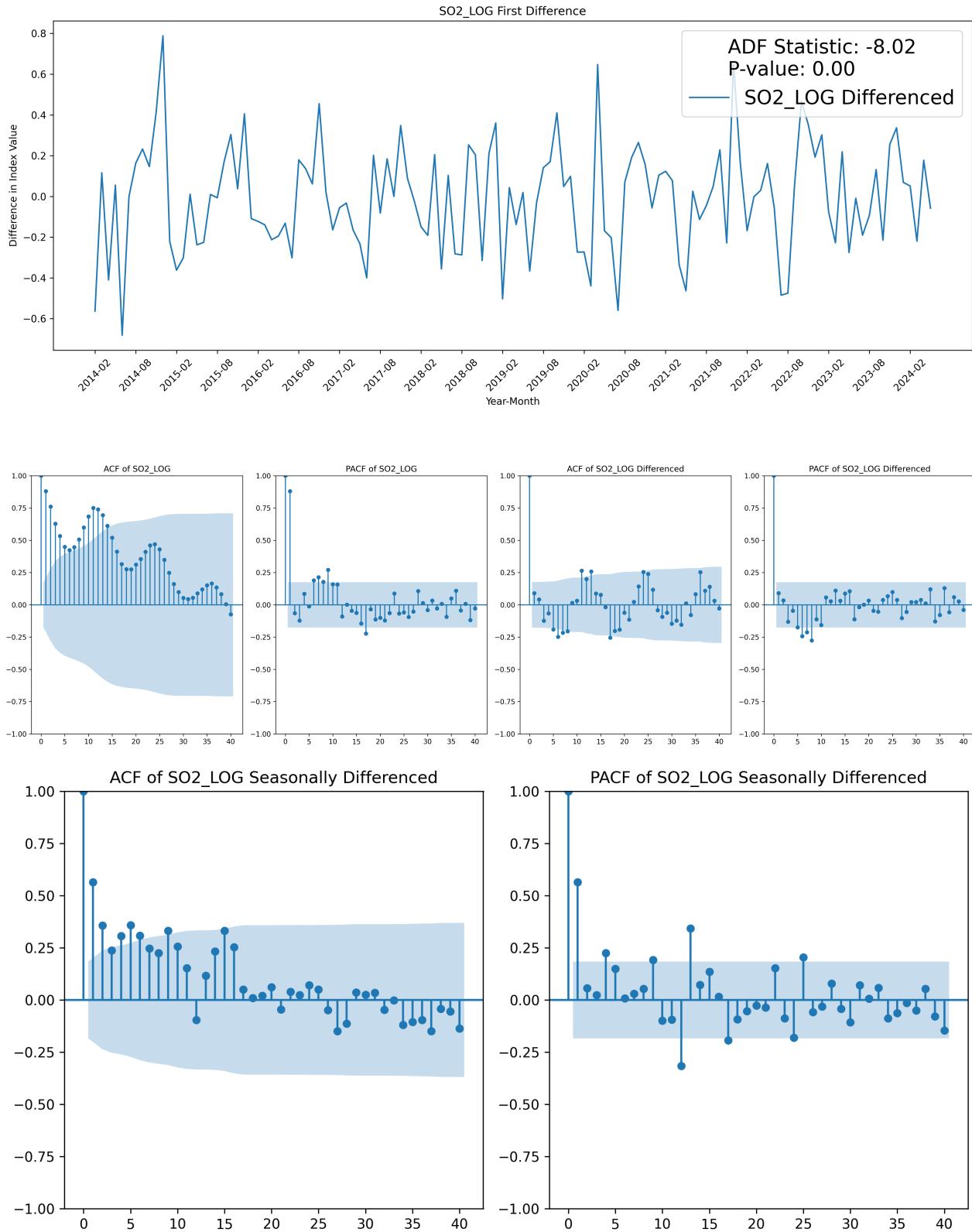
3.6 讨论与反思

3.6.1 SO2 的重新建模

由上可知, SO2 的一阶差分平稳性并不非常显著, 而后为其创建的模型也是无法通过正态性检验。由其数据特性, 我们考虑进行对数变换。



显然, 对数变换后一阶差分后的 SO2 指数出现了显著的平稳性。考虑其 ACF 与 PACF



由图可为其推荐建立季节性 ARIMA 模型 $ARIMA(0, 1, 1) \times (1, 1, 0)_{12}$ 。则有模型预测图像

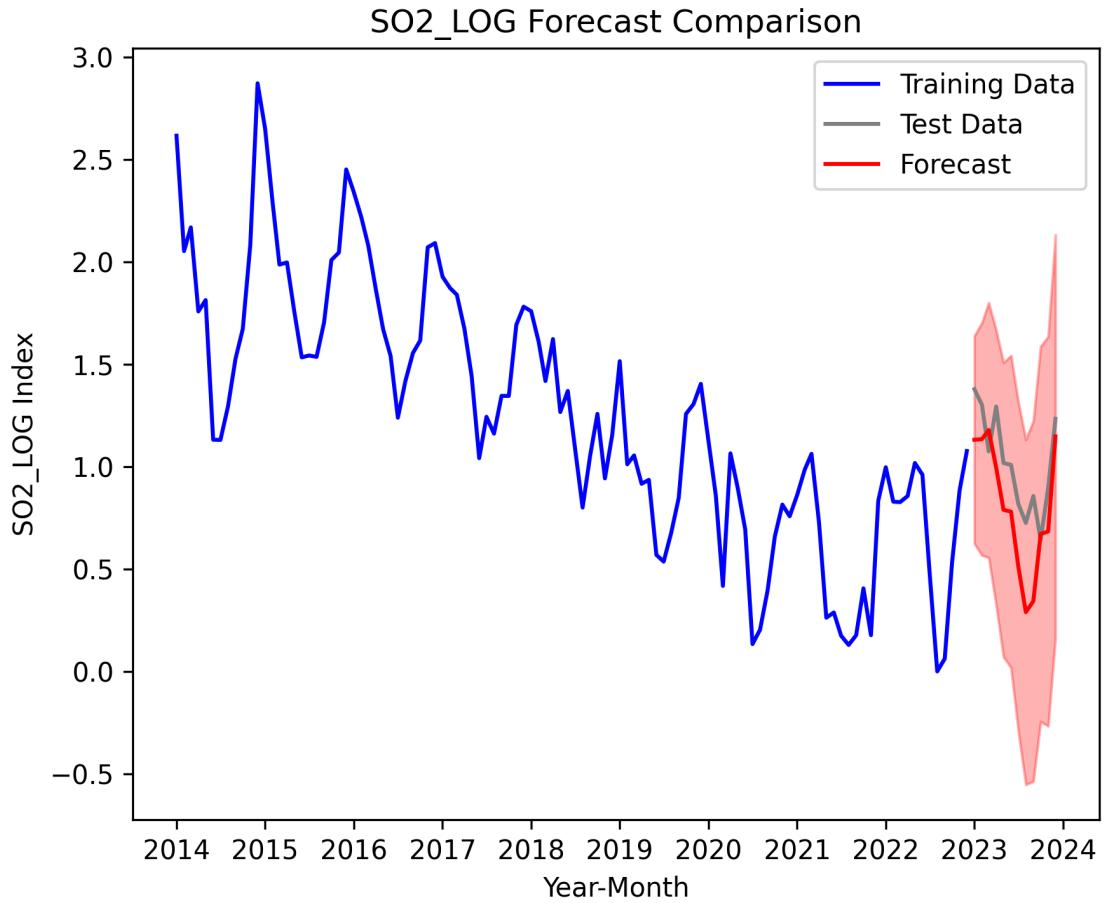


图 12: SO2_LOG 预测图

以及其残差的正态性与自相关性检验

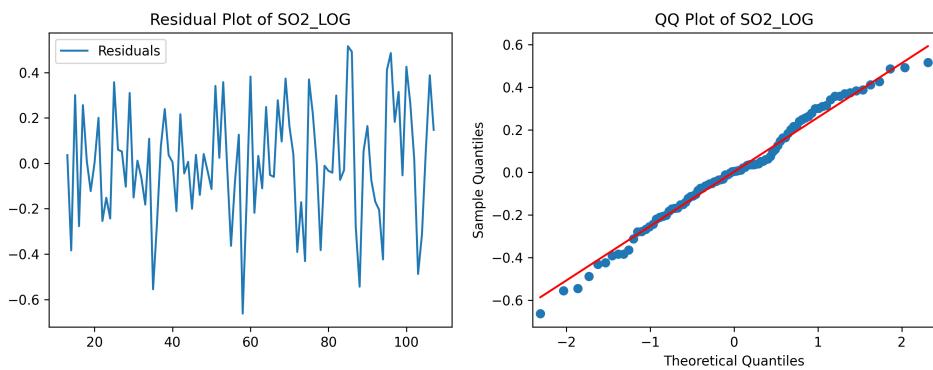


图 13: SO2_LOG 残差图

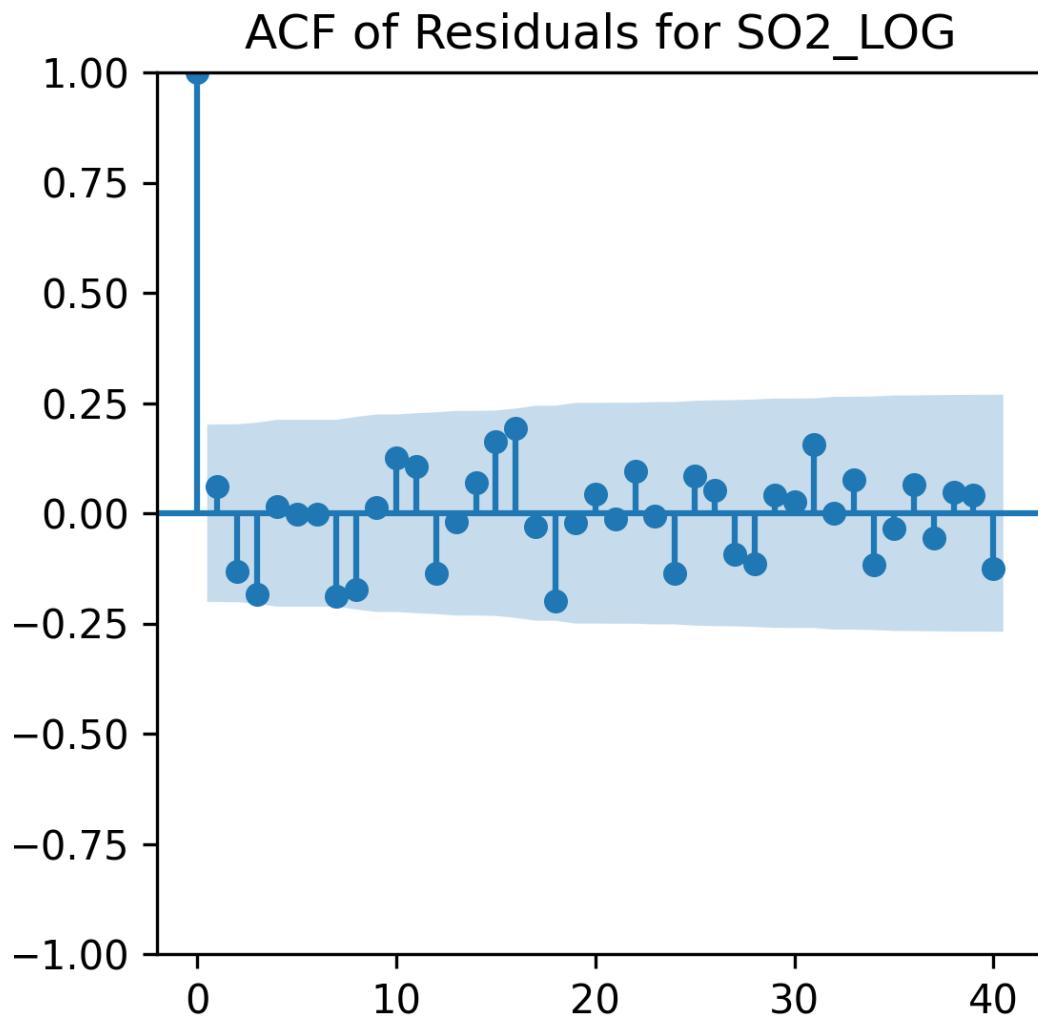


图 14: SO2_LOG 残差 ACF 图

表 6: Shapiro-Wilk 与 Ljung-Box 检验

| 特征名称 | Shapiro-Wilk 统计量 | P-值 |
|---------|------------------|--------|
| so2_log | 0.986 | 0.419 |
| 特征名称 | Ljung-Box 统计量 | P-值 |
| so2_log | 33.908 | 0.0863 |

显然, 对数变换后一阶差分后的 SO2 指数在该模型下有良好的残差正态性以及无自相关性。

4 总结与反思

- 本次报告对上海市 2014 年-2022 年的空气质量指数数据进行了深入分析，包括探索性数据分析、时间序列建模以及模型诊断。
- 通过探索性数据分析，我们发现空气质量指数在不同年份内和不同季节内的变化规律，为后续时间序列建模提供了重要参考。
- 通过时间序列建模，我们为每个空气污染指数建立了合适的 ARIMA 模型，并且对 2023 年以后的数据进行了预测。
- 通过模型诊断，我们对模型的残差进行了正态性检验和自相关性检验，发现大部分模型的残差都通过了检验。
- 在建模过程中，我们发现 SO₂ 的一阶差分并不平稳，通过对数变换后，建立了新的 ARIMA 模型，该模型的残差通过了正态性检验和自相关性检验。
- 通过本次报告，我们对时间序列分析的方法有了更深入的了解，同时也对空气质量指数的变化规律有了更清晰的认识。
- 当然，本次报告还有很多缺陷。例如未通过趋势 + 周期 + 残差的分解来对时序进行建模。
- 未对季节性 ARIMA 内的参数做进一步讨论，没有考虑并不显著的参数调整。
- 未构造足够多的推荐模型进行 AIC、BIC 检验得出最佳模型。

A 附录

表 7: SARIMAX 模型结果汇总表

| 参数 | 系数 | 标准误 | P 值 | 置信区间 |
|--|--------|-------|-------|----------------|
| PM2.5 模型: SARIMAX(3, 1, 1)x(1, 1, 1, 12) | | | | |
| ar.L1 | 0.2785 | 0.130 | 0.032 | [0.024, 0.533] |
| 接下一页 | | | | |

表 7 – 续前页

| 参数 | 系数 | 标准误 | P 值 | 置信区间 |
|--|----------|----------|-------|--------------------|
| ar.L2 | 0.0855 | 0.134 | 0.523 | [-0.177, 0.348] |
| ar.L3 | -0.1769 | 0.106 | 0.094 | [-0.384, 0.030] |
| ma.L1 | -1.0000 | 51.150 | 0.984 | [-101.252, 99.252] |
| ar.S.L12 | -0.4226 | 0.183 | 0.021 | [-0.781, -0.064] |
| ma.S.L12 | -0.1698 | 0.218 | 0.436 | [-0.597, 0.258] |
| sigma2 | 166.7507 | 8528.603 | 0.984 | [-16504, 16837] |
| PM10 模型: SARIMAX(0, 1, 1)x(1, 1, [], 12) | | | | |
| ma.L1 | -0.9952 | 0.233 | 0.000 | [-1.452, -0.538] |
| ar.S.L12 | -0.4792 | 0.110 | 0.000 | [-0.694, -0.264] |
| sigma2 | 52.6734 | 11.658 | 0.000 | [29.823, 75.523] |
| O3 模型: SARIMAX(1, 1, 1)x(1, 1, [], 12) | | | | |
| ar.L1 | 0.1467 | 0.141 | 0.297 | [-0.129, 0.422] |
| ma.L1 | -0.9178 | 0.071 | 0.000 | [-1.056, -0.779] |
| ar.S.L12 | -0.4055 | 0.078 | 0.000 | [-0.559, -0.252] |
| sigma2 | 47.1473 | 6.941 | 0.000 | [33.542, 60.752] |
| NO2 模型: SARIMAX(0, 1, 1)x(1, 1, 1, 12) | | | | |
| ma.L1 | -0.9443 | 0.072 | 0.000 | [-1.085, -0.803] |
| ar.S.L12 | -0.1777 | 0.191 | 0.352 | [-0.552, 0.196] |
| ma.S.L12 | -0.4988 | 0.197 | 0.012 | [-0.886, -0.112] |
| sigma2 | 14.5765 | 2.669 | 0.000 | [9.346, 19.807] |
| SO2 模型: SARIMAX(0, 1, 0)x(0, 1, [1], 12) | | | | |
| ma.S.L12 | -0.1965 | 0.066 | 0.003 | [-0.326, -0.067] |
| sigma2 | 1.1423 | 0.166 | 0.000 | [0.817, 1.467] |
| CO 模型: SARIMAX(0, 1, 0)x(1, 1, [1], 12) | | | | |
| ar.S.L12 | -0.1470 | 0.203 | 0.469 | [-0.545, 0.251] |
| ma.S.L12 | -0.4161 | 0.212 | 0.050 | [-0.832, -0.000] |

接下一页

表 7 – 续前页

| 参数 | 系数 | 标准误 | P 值 | 置信区间 |
|---|---------|-------|-------|------------------|
| sigma2 | 1.5173 | 0.235 | 0.000 | [1.056, 1.979] |
| SO2_log 模型: SARIMAX(0, 1, 1)x(1, 1, [], 12) | | | | |
| ma.L1 | -0.4985 | 0.098 | 0.000 | [-0.690, -0.307] |
| ar.S.L12 | -0.5856 | 0.094 | 0.000 | [-0.769, -0.402] |
| sigma2 | 0.0670 | 0.011 | 0.000 | [0.045, 0.089] |