中国研究生创新实践系列大赛

“华为杯”第十八届中国研究生

数学建模竞赛

题 目 空气质量预报二次建模

摘 要：

大气污染物是指由于人类活动或者自然过程排入大气中，有害于环境和人的物质，主要包括硫氧化物、碳氮化物、氮氧化物以及微粒。污染物在进入大气之后，就会受到气象条件的影响。目前，污染物的预测主要依靠 WRF-CMAQ 是气象与化学双向耦合的模型，增加了气溶胶对于短波、长波的直接影响、臭氧对于长波的直接影响和气溶胶对于云雨生成的间接影响，从而使得空气质量模拟更接近于实际情况。但由于 WRF 模拟的气象场和排放清单的不确定性，以及 CMAQ 中污染物的生成机理不完全明晰等原因，WRF-CMAQ 模型的预报结果也存在不足之处，题目提出了二次建模的思想。

问题一已知各个污染物的质量浓度值，根据附录中表 1 可以得出污染物浓度限值的高位值和低位值以及分别对应的空气质量分数。并计算得出各项污染物的空气质量分指数，取六种污染物空气质量分指数的最大值作为空气质量指数（AQI），并判定这种污染物为首要污染物

问题二本文计算出监测点 A 逐日污染物浓度实测数据中的 AQI 值，再对六种污染物以及 AQI 值进行相关性分析得出对 AQI 值影响最大的污染物最为代表污染物，最后利用多元线性回归模型对监测点 A 逐小时气象实测数据与代表污染物进行回归分析，从而对气象条件进行分类。

问题三计算 A、B、C 三个监测点一次模型的预测值和真实值的误差，得到三地的平均误差值，以此来建立时间序列模型预测新的误差值，进行二次模型的预测。

问题四本文使用 CNN 模型对历史实测数据中六种污染物浓度分别进行分析得出污染物的空间特征，而后利用 LSTM 模型对空间特征数据进行时序提取并结合预测的各类气象数据通过全连接层，得出污染物的时空特征并进行预测。然后使用标准误差来评价模型来评价模

型的性能。

目录

一、问题重述.....3
1.1 背景.....3
1.2 本文需要解决的问题.....4
二、模型假设与符号说明.....5
三、问题一模型的建立与求解.....5
3.1 问题一分析.....6
3.2 问题一求解.....6
四、问题二模型的建立与求解.....7
4.1 问题二分析.....7
4.2 模型建立.....7
4.3 模型求解.....8
五、问题三模型的建立与求解.....9
5.1 问题三分析.....9
5.2 模型建立.....9
5.3 模型求解.....10
六、问题四模型的建立与求解.....17
6.1 问题四分析.....17
6.2 模型建立.....17
6.3 模型求解.....17

参考文献：.....17
附录：.....18

一、问题重述

1.1 背景

生态环境与人类的生存与发展息息相关，但随着人类社会的不断发展，人类对自然环境的破坏也越来越严重。特别是从工业革命以来，人类为了发展经济已经让我们赖以生存的地球家园受到了无法挽回的破坏。城市化和工业化的发展也带来了严重的大气污染问题，严重危害人体健康，也给自然生态环境和气候变化带来极为不利的影响。目前，解决大气污染问题刻不容缓，成为全人类最棘手的问题之一，因为大气污染带来的危害无国界，时刻威胁着我们的生存条件。就我国而言，大气污染也是不容乐观，虽然近年来在国家领导人的带领以及全国人民的积极配合下，蓝天保卫战期间大气环境有所改善，但受不利气象条件影响，冬季持续性、区域性的雾霾和重污染天气过程和夏季区域性臭氧污染天气过程仍时有发生，因此大气污染防治措施仍不能松懈。

大气污染是指大气中污染物质的浓度达到有害程度，从而导致生态系统和人类正常生存和发展的条件被破坏，进而危害人和物的现象。而大气污染物是指由于人类活动或者自然过程排入大气中，有害于环境和人的物质，主要包括硫氧化物、碳氮化物、氮氧化物以及微粒，它们主要来源于工业废气、生活燃煤、汽车尾气等人类活动以及火山喷发、森林火灾、自然尘、森林植物释放、海浪飞沫颗粒物等天然源。污染物在进入大气之后，就会受到气象条件的影响。例如，风向风速影响污染物的迁移扩散方向和速度；空气湿度过大时，空气中的水蒸气和污染物会形成稳定的气溶胶，导致污染物不容易扩散等。

根据《环境空气质量标准》（GB3095-2012），用于衡量空气质量的常规大气污染物共有六种，分别为二氧化硫（SO_2）、二氧化氮（NO_2）、粒径小于 10 μm 的颗粒物（PM_{10}）、粒径小于 2.5 μm 的颗粒物（$PM_{2.5}$）、臭氧（O_3）、一氧化碳（CO）。其中粒径小于 10 μm 的颗粒物（PM_{10}）、粒径小于 2.5 μm 的颗粒物（$PM_{2.5}$）属于气溶胶状态污染物，二氧化硫（SO_2）、二氧化氮（NO_2）、一氧化碳（CO）属于气体状态污染物，而臭氧（O_3）是由于氧化型大气污染物在阳光照射下引起光化学反应生成的二次污染物。这些污染物不仅给大气环境造成严重影响，还会对人体产生极大的危害。比如SO_2、NO_2会引起心肺疾病，CO 能引起严重缺氧症状等，因此大气污染防治是全人类都不可避免的问题。

由于大气污染过程与气象条件密切相关，因此我们通过建立空气质量预报模型，先发

制人，提前采取防治措施应对可能发生的大气污染，从而减少大气污染带来的种种危害，提高环境空气质量。

针对空气质量预测问题，国内外学者也展开了充分的研究。例如，徐爱兰等[1]以南通市为例，基于 K-means 聚类方法和 CNN-LSTM 混合深度学习模型对多监测站点的数据进行训练和预测，给出了 $PM_{2.5}$ 的浓度预报；Gilik Aysenur 等[2]针对巴塞罗那、科卡利、伊斯坦布尔等城市为例，结合卷积神经网络和长短时记忆深度神经网络模型，通过时空关系对城市多位置的空气质量进行预报；Lei Zhang 等[3]针对空气数据的时空特征利用时空正交立方体模型(STOR-cube)和时空动态对流模型(ST-DA)对空气质量进行长期预报；Seng Dewen 等[4]提出了一种基于长短时记忆(LSTM)的多输出多指标有监督学习(MMSL)的综合预测模型，对 SO_2 、 NO_2 、 $PM_{2.5}$ 、 O_3 、 CO 浓度进行预测；Mao Wenjing 等[5]提出了具有时间滑动长短时记忆扩展模型(TS- LSTME)的神经网络这一深度学习框架对京津冀地区的 $PM_{2.5}$ 24 小时平均浓度进行预测；我国城市环境空气质量预报现有的模型主要包括多元线性回归、人工神经网络、NAQPMS、CMAQ、CAMx、WRF-Chem 及多模式集合预报体系等，但这些模型还未考虑到空气数据的时空相关性。[6]

目前，WRF-CMAQ 模拟体系是常用的空气质量预报模型，其中 WRF 是集数值天气预报、大气模拟、数据同化于一体的模型系统，主要用于大气环境模拟、天气研究、气象预报等，并为空气质量模型提供气象场；而 CMAQ 模型是第三代空气质量模型系统，能够根据 WRF 模型提供的气象场信息以及场域内的污染排放清单，基于物理和化学反应原理模拟污染物等变化过程，从而得出具体时间点或时间段的预报结果。

WRF-CMAQ 是气象与化学双向耦合的模型，增加了气溶胶对于短波、长波的直接影响、臭氧对于长波的直接影响和气溶胶对于云雨生成的间接影响，从而使得空气质量模拟更接近于实际情况。但由于 WRF 模拟的气象场和排放清单的不确定性，以及 CMAQ 中污染物的生成机理不完全明晰等原因，WRF-CMAQ 模型的预报结果也存在不足之处，题目提出了二次建模的思想。实际上，气象条件是影响空气质量的重要因素，且污染物浓度的实测数据的变化情况对预报结果具有一定的参考价值，所以我们考虑在一次预报结果的基础上，结合这两部分的数据源进行再次建模，从提高预报结果的准确性，尽可能减少空气污染给人类生产生活带来的影响。

1.2 本文需要解决的问题

问题一：使用附件 1 中的数据，利用附录中环境空气质量指数计算公式计算监测点 A 从 2020 年 8 月 25 日到 8 月 28 日每天实测的 AQI 和首要污染物。（按照附录“AQI 计算结果表”的格式放在正文中。）

问题二：依据各类气象体条件对污染物浓度的影响程度，对气象条件进行合理分类，并阐述各类气象条件的特征。

问题三：忽略检测点之间的相互影响，利用附件 1、2 中的数据，建立一个适用于 A、B、C 三个监测点，同时可以最小化 AQI 预报值的最大相对误差，并提高污染物预测准确度的二次预报数学模型。利用所建立的模型来预测监测点 A、B、C 在 2021 年 7 月 13 日至 7 月 15 日 6 种常规污染物单日浓度值，并且计算相应的 AQI 和首要污染物。（按照附录“AQI 计算结果表”的格式放在正文中。）

问题四：使用附件 1、3 中的数据，建立包含 A、A1、A2、A3 四个监测点，同时可以最小化 AQI 预报值的最大相对误差，并提高污染物预测准确度的协同预报模型。利用所建立的模型预测监测点 A、A1、A2、A3 在 2021 年 7 月 13 日至 7 月 15 日 6 种常规污染物的单日浓度值，并且计算相应的 AQI 和首要污染物。（按照附录“AQI 计算结果表”的格式放在正文中。）说明协同预报模型与二次预报数学模型相比，能否提升针对监测点 A 的污染物浓度预报准确度的理由。

二、模型假设与符号说明

符号	意义
C_{O_3}	臭氧（O_3）最大 8 小时滑动平均
C_t	臭氧在某日 $t-1$ 时至 t 时的平均污染物浓度
$IAQI_P$	污染物 P 的空气质量分指数，结果取整数
C_P	污染物 P 的质量浓度值
BP_{Hi}, BP_{Lo}	与 C_P 相近的污染物浓度限值的高位值和低位值
$IAQI_{Hi}, IAQI_{Lo}$	与 BP_{Hi}, BP_{Lo} 对应的空气质量分指数
AQI	空气质量指数

三、问题一模型的建立与求解

3.1 问题一分析

问题一要求我们利用附件 1 中的数据计算监测点 A 从 2020 年 8 月 25 日到 8 月 28 日每天实测的 AQI 和首要污染物。已知各个污染物的质量浓度值，根据附录中表 1 （空气质量分指数（IAQI）及对应的污染物项目浓度限值）可以得出污染物浓度限值的高位值和低位值以及分别对应的空气质量分数。并计算得出各项污染物的空气质量分指数，取六种污染物空气质量分指数的最大值作为空气质量指数（AQI），并判定这种污染物为首要污染物。

3.2 问题一求解

首先，通过附件 1 中的表 3 可知 2020 年 8 月 25 日到 8 月 28 日臭氧(O_3)最大 8 小时滑动平均浓度值分别为 $112ug/m^3$、$92ug/m^3$、$169ug/m^3$、$201ug/m^3$，均低于 $800ug/m^3$，且其余五种污染物的浓度均低于 $IAQI=500$ 对应限值，因此各污染物的空气质量分指数都需要进行求解。

其次，通过空气质量分指数的公式：

$$IAQI_p=\frac{IAQI_{Hi}-IAQI_{Lo}}{BP_{Hi}-BP_{Lo}}\cdot\left(C_p-BP_{Lo}\right)+IAQI_{Lo} \quad (1)$$

我们求出了 2020 年 8 月 25 日到 8 月 28 日各污染物的空气质量分指数，如下表所示：

表 1　2020 年 8 月 25 日-8 月 28 日各污染物的空气质量分指数

检测时间	$IAQI_{CO}$	$IAQI_{SO_2}$	$IAQI_{NO_2}$	$IAQI_{O_3}$	$IAQI_{PM_{10}}$	$IAQI_{PM_{2.5}}$
2020/8/25	12	8	15	60	27	27
2020/8/26	12	7	20	46	24	24
2020/8/27	15	7	39	108	37	37
2020/8/28	18	8	38	137	47	47

由 AQI 的计算公式:

$$AQI=\max\left\{IAQI_{SO_2},IAQI_{NO_2},IAQI_{PM_{10}},IAQI_{PM_{2.5}},IAQI_{O_3},IAQI_{CO}\right\} \quad (2)$$

可知，每天空气质量指数是取当天的各空气质量分指数的最大值，于是我们得到了 2020 年 8 月 25 日到 8 月 28 日每天的空气质量指数以及每日首要污染物，如下表所示：

表 2　2020 年 8 月 25 日-8 月 28 日每天实测的 AQI 和首要污染物

监测日期	地点	AQI 计算	
		AQI	首要污染物

2020/8/25	监测点 A	60	O_3
2020/8/26	监测点 A	46	无首要污染物
2020/8/27	监测点 A	108	O_3
2020/8/28	监测点 A	137	O_3

四、问题二模型的建立与求解

4.1 问题二分析

问题二要求我们依据各类气象体条件对污染物浓度的影响程度，对气象条件进行合理分类。已知气象条件会影响污染物的扩散或沉降，从而影响 AQI 值。我们考虑把气象条件分为两类，一类气象条件容易使污染物扩散使 AQI 上升，另一类气象条件容易导致污染物沉降降低 AQI。首先，我们计算出监测点 A 逐日污染物浓度实测数据中的 AQI 值，再对六种污染物以及 AQI 值进行相关性分析得出对 AQI 值影响最大的污染物最为代表污染物，最后利用多元线性回归模型对监测点 A 逐小时气象实测数据与代表污染物进行回归分析，从而对气象条件进行分类。

4.2 模型建立

相关系数用来研究变量之间的线性相关程度。我们采用了 Spearman 相关系数来度量监测点 A 逐日污染物浓度实测数据六种污染物以及 AQI 值进行相关性。对于内容变量为 n 的样本，原始数据进行转换变为等级数据，相关系数为：

$$\rho=\frac{\sum_i(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_i(x_i-\bar{x})^2\sum_i(y_i-\bar{y})^2}}$$

多元线性回归模型的一般表达式为

$$Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\cdots+\beta_kX_{ki}+\mu_i,i=1,2,\cdots n$$

其中，k 为解释变量的数目，$\beta_j(j=1,2,\cdots k)$ 称为回归系数。上式也被称为总体回归函数的随机表达式，它的非随机表达式为

$$E\left(Y\middle|X_{1i},X_{2i},\cdots,X_{ki}\right)=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\cdots+\beta_kX_{ki}$$

$\beta_j(j=1,2,\cdots k)$ 也被称为偏回归系数。

4.3 模型求解

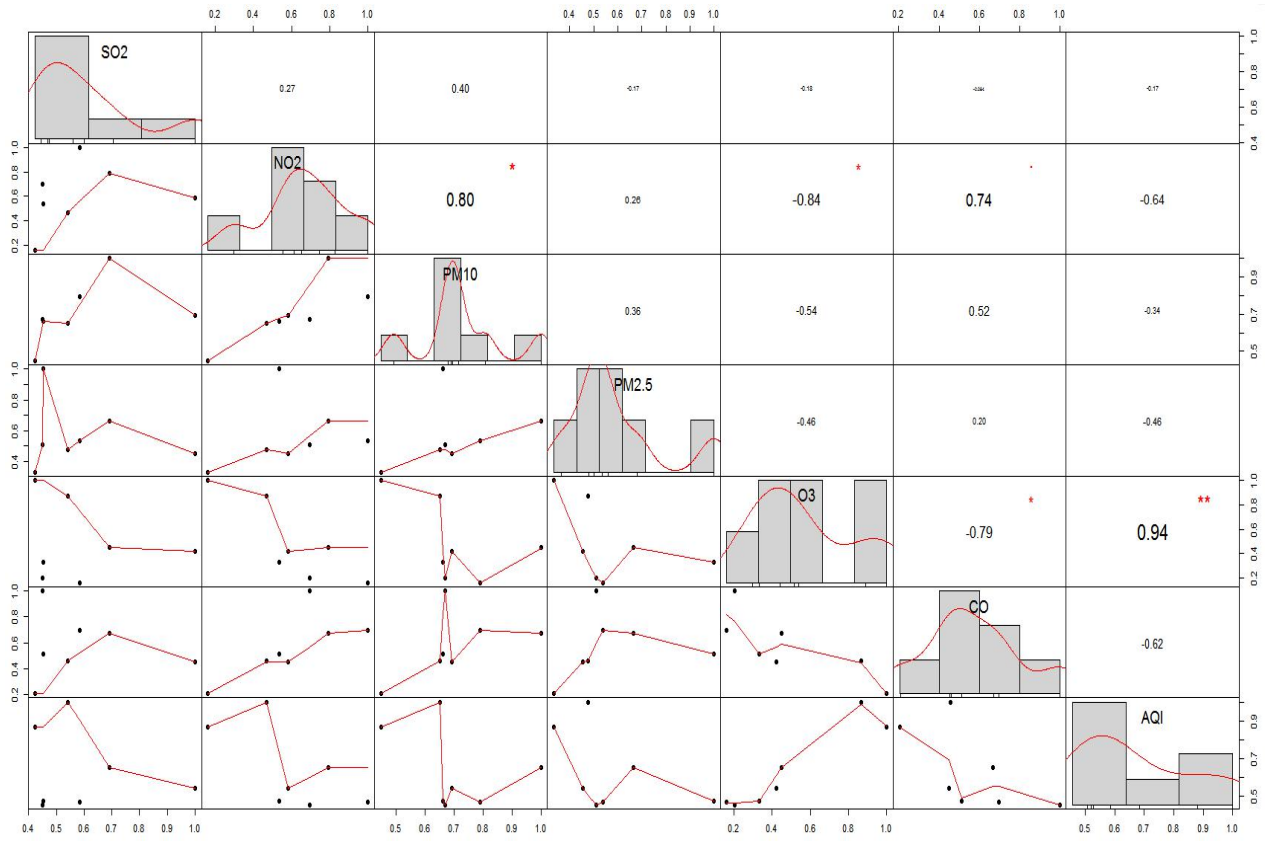


图 1 监测点 A 逐日污染物浓度实测数据六种污染物以及 AQI 值进行相关性分析图

从图 1 中可以看出臭氧和 AQI 值的相关性最高，相关系数为 0.94，所以我们选取臭氧为代表污染物，观察空气中臭氧浓度和各类气象条件的关系，以此来对气象条件进行分类。

表 3 监测点 A 逐小时气象实测数据与代表污染物进行回归结果

气象条件	污染物					
	SO2	NO2	PM10	PM2.5	O3	CO
温度	0.0396**	-1.4103**	0.3664**	-0.0321**	4.5978**	-0.0128**
湿度	-0.1025**	-0.2230**	-0.7022**	-0.3464	-1.7578**	-0.0014**
气压	0.0897**	-0.2023**	1.0907**	0.7194**	1.1042**	0.0014**
风向	-0.0001	-0.0127**	0.0022	0.0032**	-0.0310**	0.0001**
风速	-1.2000**	-17.075**	-14.0410**	-10.6719**	1.0064*	-0.1141**

（注：*表示通过置信度为 0.01 的显著性检验，**表示通过置信度为 0.001 的显著性检验）

从表 3 中，我们可以看出温度、气压、风速三个气象条件对空气中臭氧浓度有正向影响，而湿度和风向与臭氧浓度呈负相关关系。所以我们将温度、气压、风速分为第一类气象条件，这一类气象条件数值的上升会导致空气中的 AQI 值增高，造成空气的污染。将湿

度和风向分为第二类气象条件，这里以气象条件数值的升高会引起 AQI 值的下降，空气污染程度更低。

五、问题三模型的建立与求解

5.1 问题三分析

问题三要求我们建立一个适用于 A、B、C 三个监测点，同时可以最小化 AQI 预报值的最大相对误差，并提高污染物预测准确度的二次预报数学模型。我们通过计算 A、B、C 三个监测点一次模型的预测值和真实值的误差，得到三地的平均误差值，以此来建立时间序列模型预测新的误差值，进行二次模型的预测。并得出监测点 A、B、C 在 2021 年 7 月 13 日至 7 月 15 日 6 种常规污染物单日浓度值，计算得到相应的 AQI 和首要污染物。

5.2 模型建立

利用 2020.7.24-2021.7.12 数据建立二次预报的数学模型，以期得到更高的预测精度。

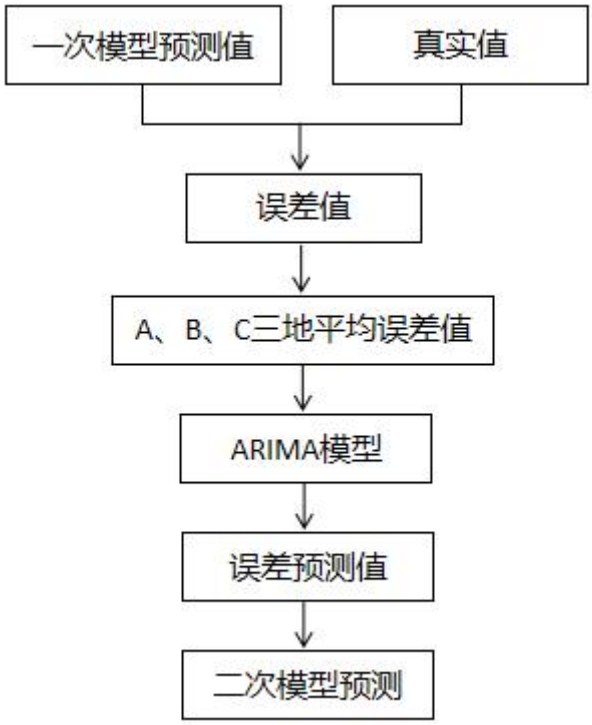
首先，利用附件 1 中附表“监测点 A 逐小时污染物浓度与气象一次预报数据”中该时间段数据，提取主要污染物的每日平均值（其中 CO、SO2，NO2，PM10，PM2.5 为 24 小时平均值，O3 为最大 8 小时滑动平均），即一次模型每日污染物预测值。

其次，利用附件 1 中附表“监测点 A 逐日污染物浓度实测数据”的对应时间，作为每个污染物真实数据。

接着，将真实值减去一次模型的预测值即误差值。

对 A、B、C 三个地点的误差值取平均，计算平均误差值。将平均误差值利用时间序列模型进行拟合预测，得到未来三天 2021.7.13-2021.7.15 的污染物预测误差值。

将预测误差值加上一次模型的预测值即为本文建立的二次预测模型。



5.3 模型求解

一次预测模型基于 WRF-CMAQ 模拟系统对空气进行预报，将第 i 地的第 j 种污染物预报值记录为 y_{ij}（$i=$A地，B地，C地；$j=CO$，CO_2，NO_2，PM_{10}，$PM_{2.5}$，O_3），将真实污染物实测值记录为 y_{ij}（$i=$A地，B地，C地；$j=CO$，CO_2，NO_2，PM_{10}，$PM_{2.5}$，O_3），计算其误差值记录为 $e_{ij}=y_{ij}-y_{ij}$，将 A、B、C 三个地点的各个污染物误差平均值记录为 $e_j=e_{Aj}+e_{Bj}+e_{Cj}$。

如图 2 为 3 地 6 种污染物误差值，由图可以看出，对于每种污染物，其三个地点的一次预测模型的预测误差都具有类似的变化趋势，并且误差波动情况是基于 0 上下波动，随着时间变化，其误差值上下波动，具备有平稳的趋势，为进一步探究其变化规律，本文提出利用时间序列模型 ARMA 进行误差预测，对 A、B、C 三个地点的各个污染物误差平均值 e_j 进行模拟预测，将预测情况记录为 e_{ij}。

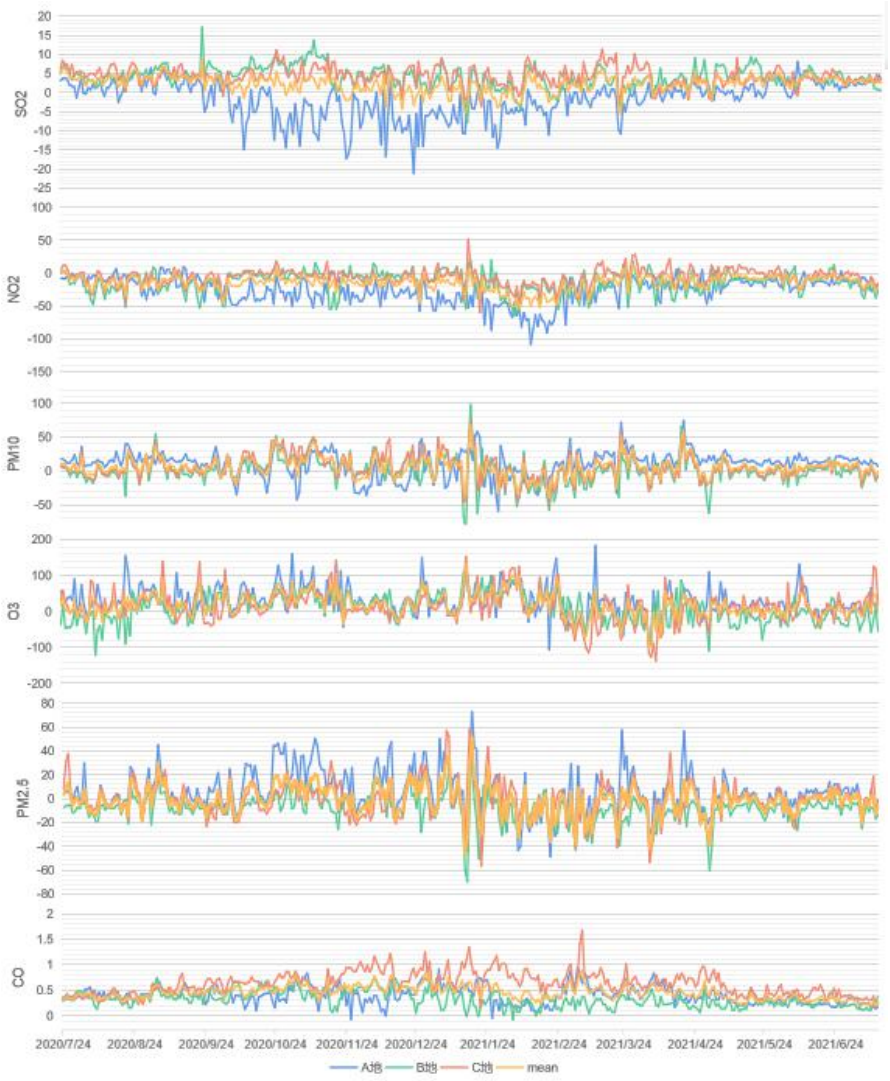


图 2 3 地 6 种污染物一次预测模型的误差情况

导入 2020 年 7 月 24 日到 2021 年 7 月 13 日监测点 A、B、C 三地的二氧化硫浓度预测误差均值序列之后，我们首先绘制了如图 3（a）所示的时序图，经过 ADF 检验，结果显示 P 值等于大于显著性水平（$\alpha = 0.05$），并且经过延迟 6 阶和 12 阶的纯随机性检验，结果显示 P 值小于显著性水平（$\alpha = 0.05$），因此可以判断该序列是非平稳非白噪声序列。

于是我们对该序列进行 1 阶差分，差分后时序图，见图 6（a）显示，差分后的序列实现了平稳。接着我们考察 1 阶差分后序列的自相关图与偏自相关图，见图 7（a）、图 8（a），经过模型识别以及综合分析，我们对三地的二氧化硫浓度预测误差均值序列拟合 $ARIMA(2,1,2)$ 模型。基于该拟合模型，对序列进行为期 3 期的预测，如图 9（a）显示。

导入 2020 年 7 月 24 日到 2021 年 7 月 13 日监测点 A、B、C 三地的二氧化氮浓度预测误差均值序列之后，我们首先绘制了如图 3（b）所示的时序图，经过 ADF 检验，结果显示 P 值等于大于显著性水平（$\alpha = 0.05$），并且经过延迟 6 阶和 12 阶的纯随机性检验，结果显示 P 值小于显著性水平（$\alpha = 0.05$），因此可以判断该序列是非平稳非白噪声序列。

于是我们对该序列进行 1 阶差分，差分后时序图，见图 6（b）显示，差分后的序列实现了平稳。接着我们考察 1 阶差分后序列的自相关图与偏自相关图见图 7（b）和图 8（b），经过模型识别以及综合分析，我们对三地的二氧化硫浓度预测误差均值序列拟合 $ARIMA(3,1,1)$ 模型。基于该拟合模型，对序列进行为期 3 期的预测，如图 9（b）显示。

按照上述的方法，我们分别对检测点 A、B、C 三地的 PM_{10}、$PM_{2.5}$、O_3、CO 浓度预测误差均值序列拟合了 $ARIMA(1,0,2)$、$ARIMA(0,1,3)$、$ARIMA(2,1,2)$、$ARIMA(1,1,1)$ 模型，并基于相应的拟合模型对序列进行了 3 期预测。

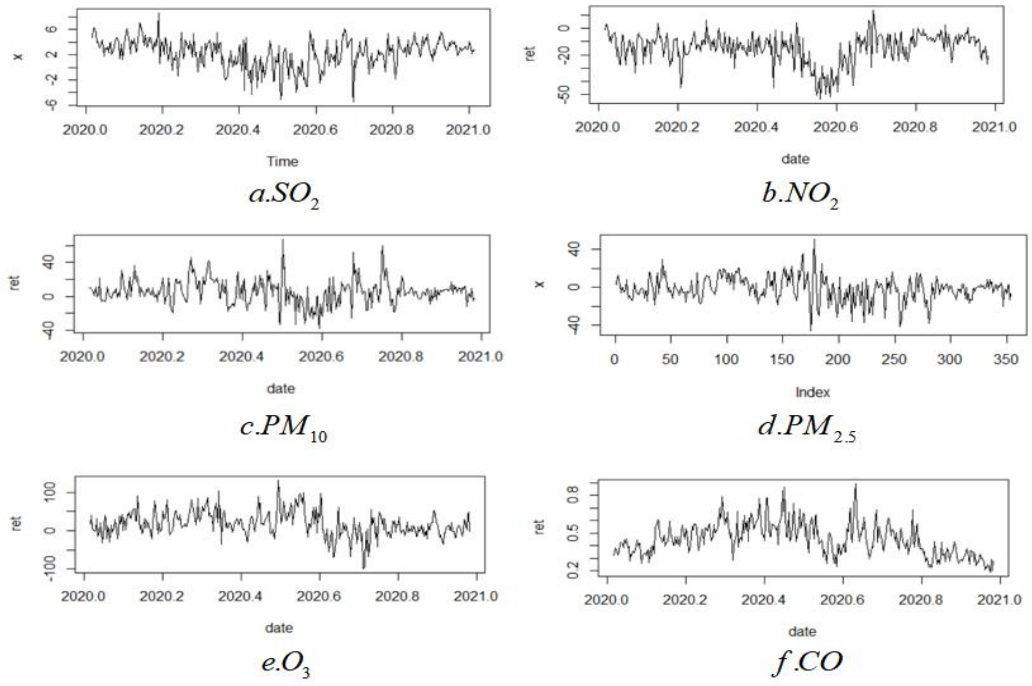


图 3　三地的六种污染物浓度预测误差均值序列时序图

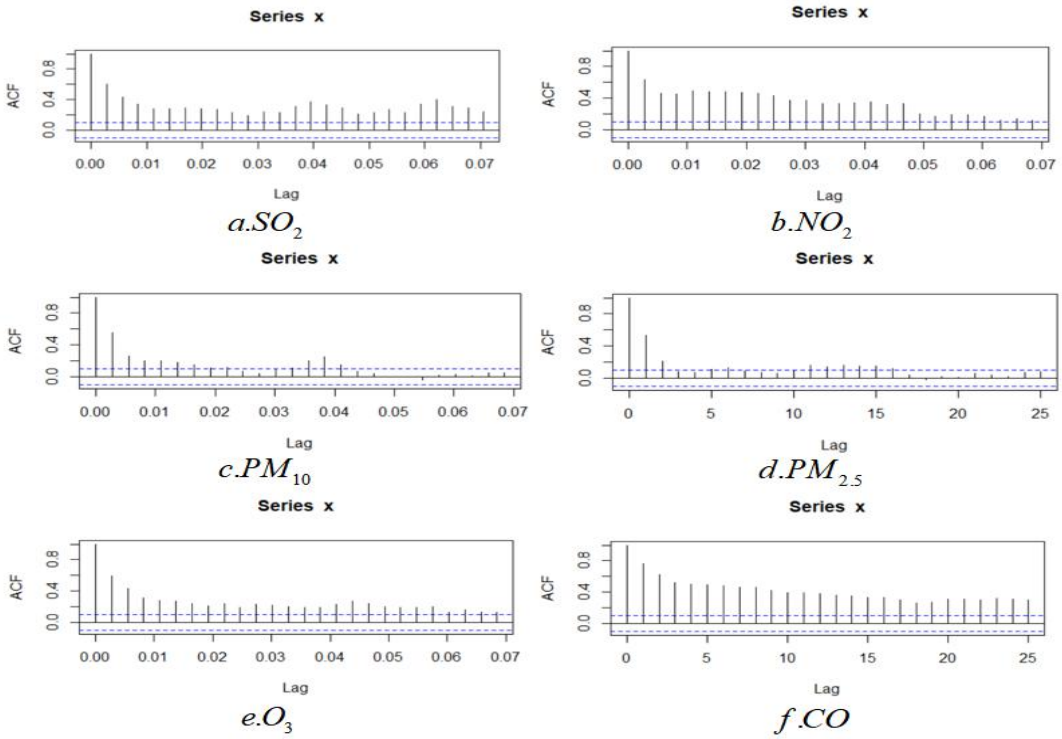


图 4　三地的六种污染物浓度预测误差均值序列自相关图

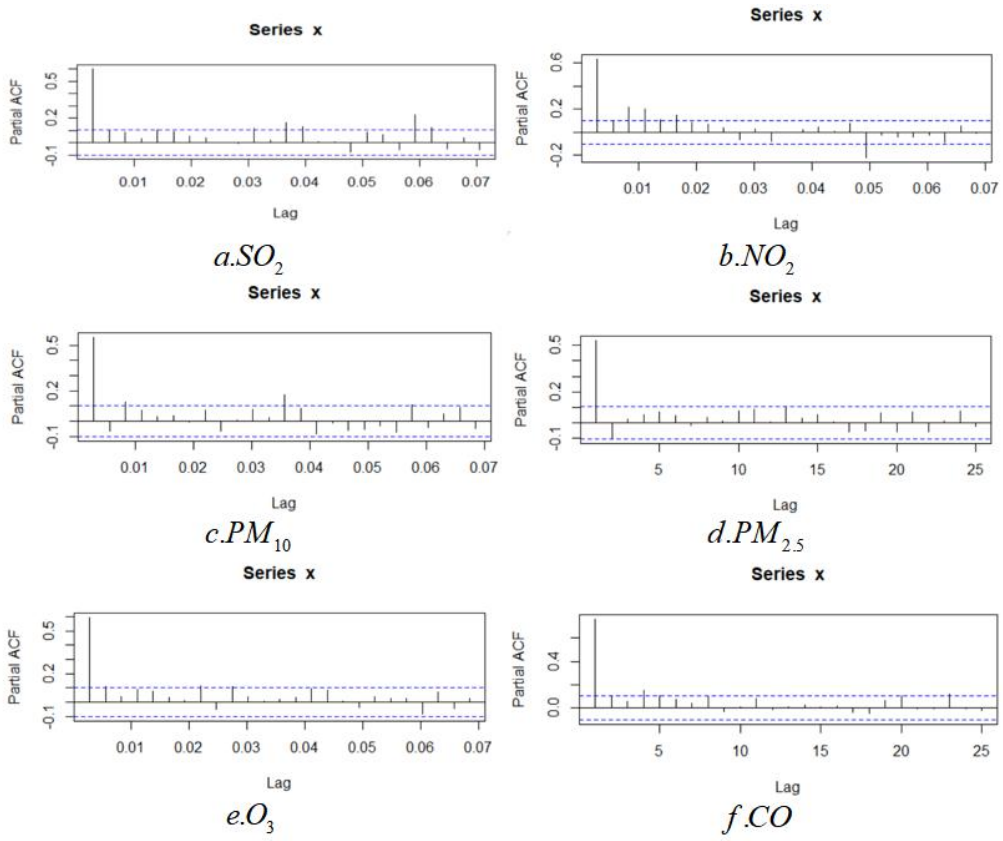
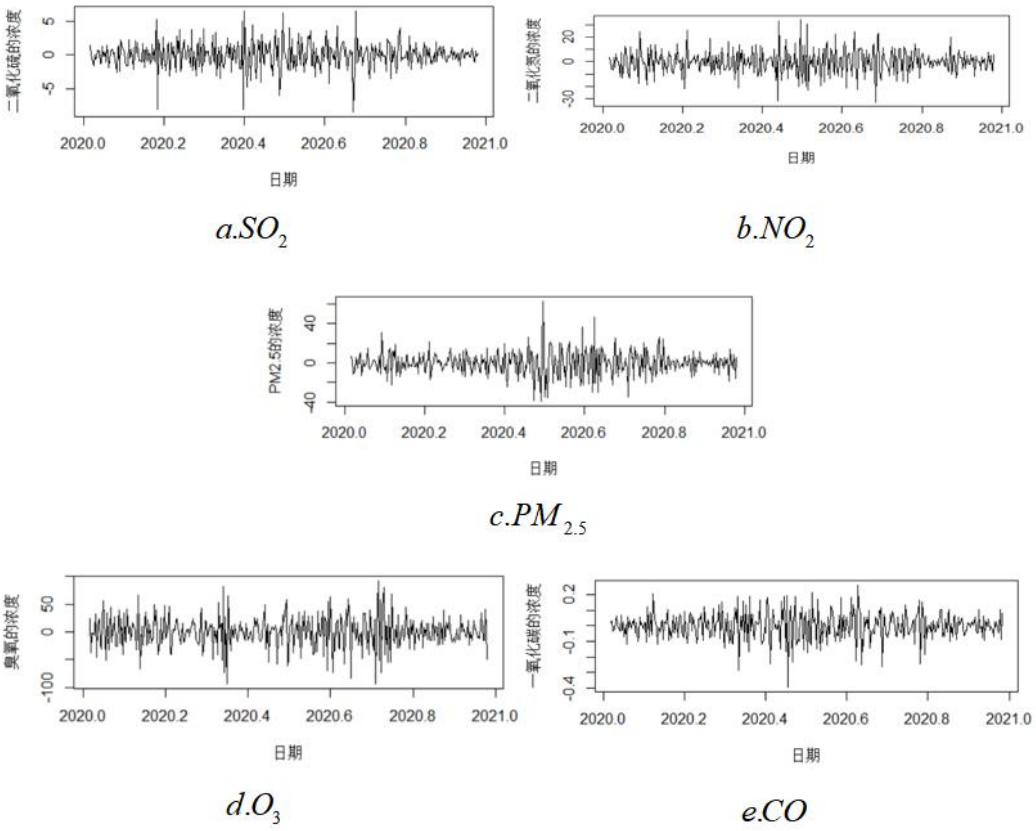


图 5　三地的六种污染物浓度预测误差均值序列偏自相关图

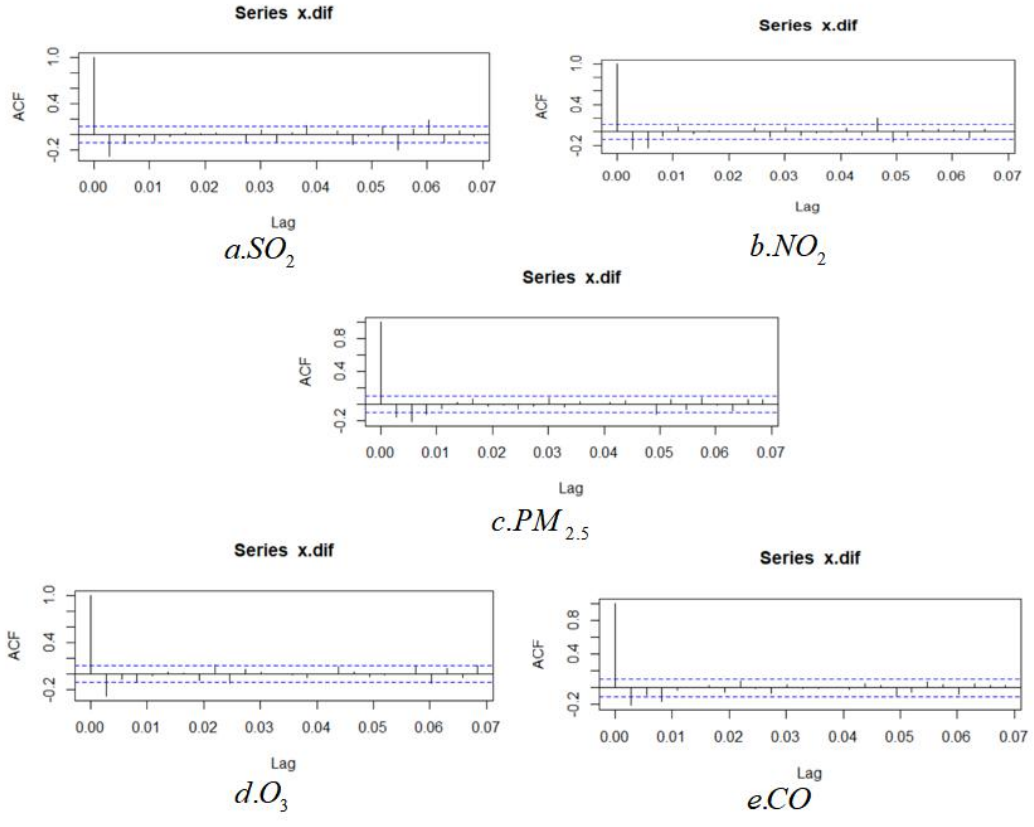


$a.SO_2$ $b.NO_2$

$c.PM_{2.5}$

$d.O_3$ $e.CO$

图 6 三地的五种污染物浓度预测误差均值 1 阶差分序列时序图

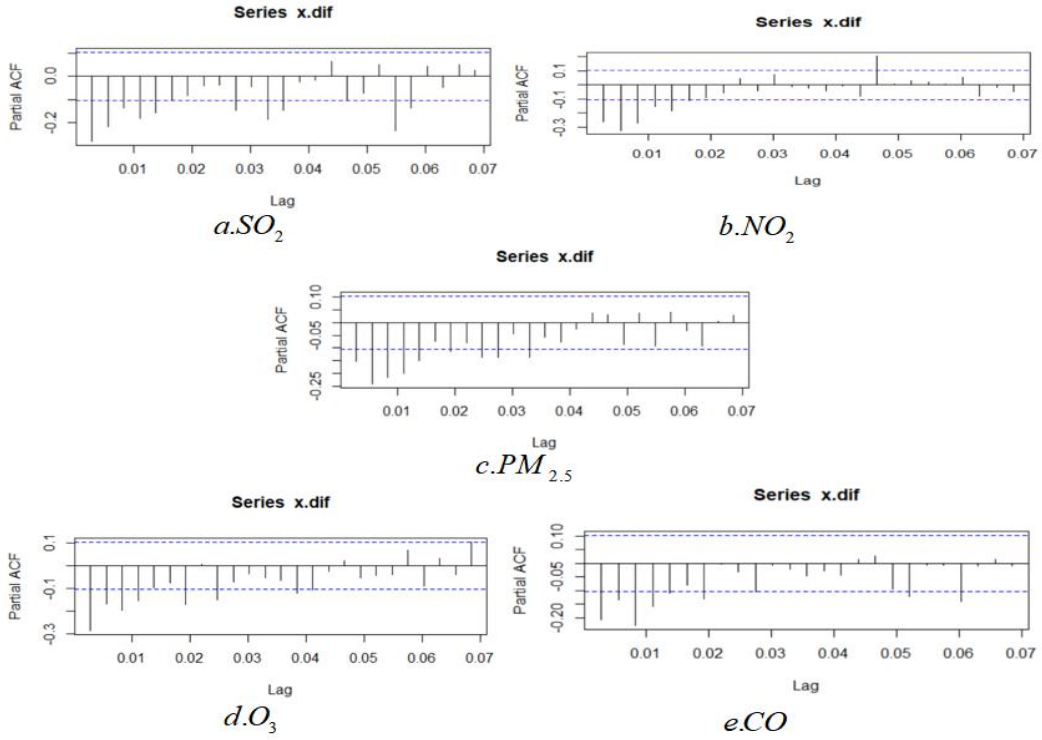


$a.SO_2$ $b.NO_2$

$c.PM_{2.5}$

$d.O_3$ $e.CO$

图 7 三地的五种污染物浓度预测误差均值 1 阶差分序列自相关图

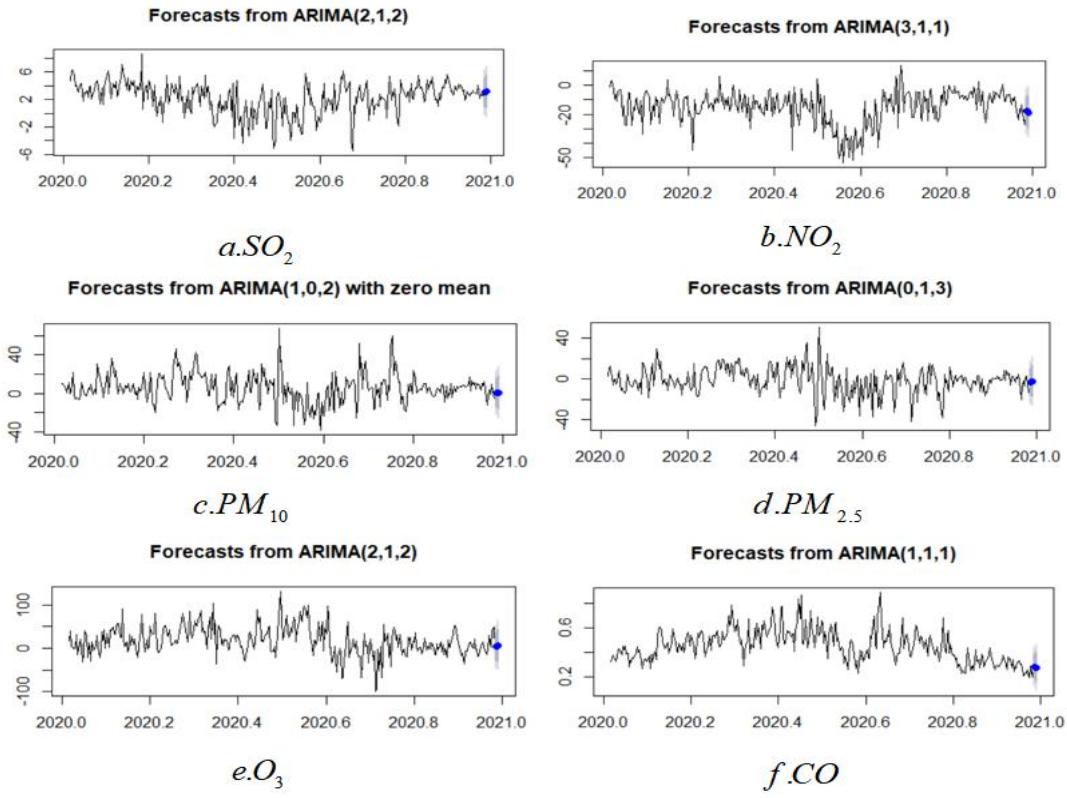


$a.SO_2$ $b.NO_2$

$c.PM_{2.5}$

$d.O_3$ $e.CO$

图 8　三地的五种污染物浓度预测误差均值 1 阶差分序列偏自相关图



$a.SO_2$ $b.NO_2$

$c.PM_{10}$ $d.PM_{2.5}$

$e.O_3$ $f.CO$

图 9　三地的六种污染物浓度预测误差均值序列预测图

由此，建立二次预测模型 $\hat{Y}_{ij} = \hat{y}_{ij} + e_{ij}$，计算 A、B、C 三地二次预测模型的 AQI 值，并根据预测 AQI 值、预测首要污染物与真实情况进行对比，衡量模型准确率，A 地基于二次预测模型的 AQI 值与真实 AQI 值对比如图 10 所示表 4 所示，为三地 AQI 预测误差、首要污染物预测准确率。

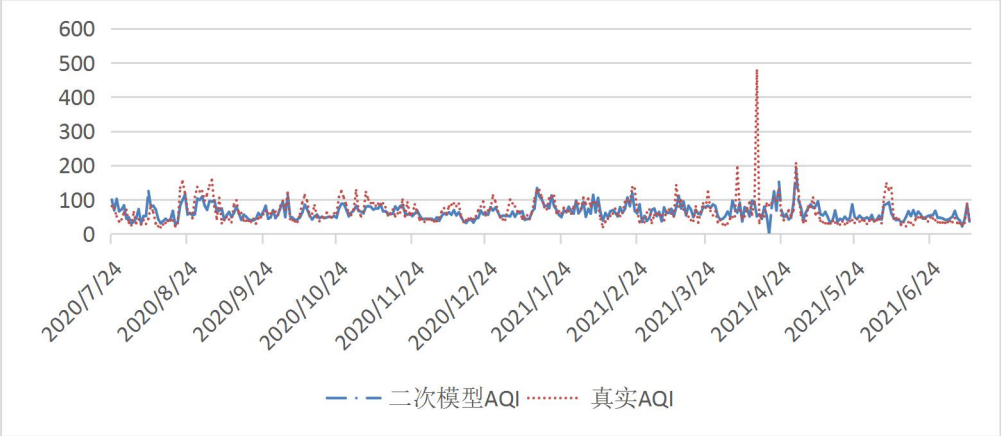


图 10　A 地基于二次预测模型的 AQI 值与真实 AQI 值对比

表 4 三地 AQI 预测误差、首要污染物预测准确率

地点	A 地	B 地	C 地
AQI 预测误差	25.943	27.138	25.771
首要污染物预测准确率	0.84	0.72	0.8

A 地污染物浓度及 AQI 预测结果

预报日期	地点	二次模型日值预测							
		SO_2 ($\mu g/m^3$)	NO_2 ($\mu g/m^3$)	PM_{10} ($\mu g/m^3$)	$PM_{2.5}$ ($\mu g/m^3$)	O_3 最大八小时滑动平均 ($\mu g/m^3$)	CO (mg/m^3)	AQI	首要污染物
2021/7/13	监测点 A	7.10	10.35	9.44	2.74	73.28	0.50	37.00	无
2021/7/14	监测点 A	7.78	17.85	9.68	4.12	90.33	0.52	45.00	无
2021/7/15	监测点 A	6.82	19.22	10.22	4.28	84.94	0.54	42.00	无
2021-7-11	监测点 B	3.17	0.00	8.62	0.78	41.28	0.36	21.00	无
2021-7-12	监测点 B	3.25	0.00	6.26	1.12	34.93	0.38	17.00	无
2021-7-13	监测点 B	3.27	0.00	6.45	0.52	49.96	0.35	25.00	无
2021-7-11	监测点 C	7.08	1.54	18.39	9.35	122.82	0.48	69.00	03
2021-7-12	监测点 C	6.65	1.41	19.78	11.56	123.61	0.49	70.00	03
2021-7-13	监测点 C	8.17	18.14	27.41	17.14	127.27	0.52	73.00	03

六、问题四模型的建立与求解

6.1 问题四分析

问题四要求我们使用附件 1、3 中的数据，建立包含 A、A1、A2、A3 四个监测点，同时可以最小化 AQI 预报值的最大相对误差，并提高污染物预测准确度的协同预报模型。我们考虑使用 CNN 模型对历史实测数据中六种污染物浓度分别进行分析得出污染物的空间特征，而后利用 LSTM 模型对空间特征数据进行时序提取并结合预测的各类气象数据通过全连接层，得出污染物的时空特征并进行预测。然后使用标准误差来评价模型来评价模型的性能。

6.2 模型建立

一个 CNN 的基本构成包含一个输入层、一个卷积层、一个输出层，其中输入层用于数据的输入；卷积层是使用给定核函数对输入数据进行特征提取，并依据核函数的数据产生若干个卷积特征结果；池化层对数据进行降维处理，从而减少数据特征；全连接层对已有数据特征进行重新提取并输出结果。

6.3 模型求解

在第三问中单个监测点仅选择单个监测点的污染物浓度历史数据，在多个监测点协同预报模型中需要输入 4 个监测点的污染物浓度数据，当某个监测点的污染物出现缺失值时，我们也要将其他站点的缺失值进行删除，用剩余的数据进行预测。输入 CNN-LSTM 模型得到预测结果。

参考文献：

[1] 徐爱兰,朱晏民,孙强,於香湘,彭小燕.基于 K-means 划分区域的深度学习空气质量预报[J].南通大学学报(自然科学版),2021,20(03):49-56.

[2] Gilik Aysenur and Ogrenci Arif Selcuk and Ozmen Atilla. Air quality prediction using CNN+LSTM-based hybrid deep learning architecture.[J]. Environmental science and pollution research international, 2021,

[3] Lei Zhang et al. Deep Spatio-temporal Learning Model for Air Quality Forecasting[J]. INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL, 2021, 16(2)

[4] Seng Dewen et al. Spatiotemporal prediction of air quality based on LSTM neural network[J]. Alexandria Engineering Journal, 2021, 60(2) : 2021-2032.

[5] Mao Wenjing et al. Modeling air quality prediction using a deep learning approach: Method optimization and evaluation[J]. Sustainable Cities and Society, 2020, : 102567-.

[6] 宋鹏程,张馨文,黄强,龙平,杜云松.我国城市环境空气质量预报主要模型及应用[J].四川环

境,2019,38(03):70-76.

附录：

数据预处理代码：

```
data = read.csv("q3.csv")#读取数据
library(mice)
library(ggplot2)
imputed = mice(data, method="rf", m=5,seed=100)#随机森林填充缺失值
imputed <- complete(imputed)
sapply(imputed, function(x) sum(is.na(x)))#检查
```

问题二代码：

```
##相关性分析##
df<-data.matrix(read.csv('timu2.csv',header=T,row.names = NULL))    #导入数据
df
cor(df,method = 'spearman') #相关系数矩阵

library(Hmisc)
res2<-rcorr(as.matrix(df))
res2
res2$r    #相关系数
res2$P    #p-value

#整合
flattenCorrMatrix <- function(cormat, pmat) {
  ut <- upper.tri(cormat)
  data.frame(
    row = rownames(cormat)[row(cormat)[ut]],
    column = rownames(cormat)[col(cormat)[ut]],
    cor  =(cormat)[ut],
    p = pmat[ut]
  )
}
flattenCorrMatrix(res2$r,res2$P)

df1<-flattenCorrMatrix(res2$r,res2$P)
abs(df1$cor)>0.25
df1[abs(df1$cor)>0.25,]

#可视化
library(PerformanceAnalytics)#加载包
chart.Correlation(res2$r, histogram=TRUE, pch=19)
```

```
write.csv(df1,'12.csv')#读出

##多元回归模型代码##
fund <- read.csv("aa.csv")#导入数据
summary(fund)#数据总览
#建立模型
mlr1 <- lm(SO2 监测浓度 ~ 温度 + 湿度 + 气压 + 风向 + 风速,data = fund)
mlr2 <- lm(NO2 监测浓度 ~ 温度 + 湿度 + 气压 + 风向 + 风速,data = fund)
mlr3<- lm(PM10 监测浓度 ~ 温度 + 湿度 + 气压 + 风向 + 风速,data = fund)
mlr4<- lm(PM2.5 监测浓度 ~ 温度 + 湿度 + 气压 + 风向 + 风速,data = fund)
mlr5<- lm(O3 监测浓度 ~ 温度 + 湿度 + 气压 + 风向 + 风速,data = fund)
mlr6<- lm(CO 监测浓度 ~ 温度 + 湿度 + 气压 + 风向 + 风速,data = fund)
#显著性检验
summary(mlr1)
summary(mlr2)
summary(mlr3)
summary(mlr4)
summary(mlr5)
summary(mlr6)
coef(mlr1)
coef(mlr2)
coef(mlr3)
coef(mlr4)
coef(mlr5)
coef(mlr6)
bb<-data.frame(coef(mlr1),coef(mlr2),coef(mlr3),coef(mlr4),coef(mlr5),coef(mlr6))#整合
write.csv(bb,'mm.csv')#读出
```

3.问题三代码：

```
filenames<-('F:/')
setwd(filenames)
Q1<-data.matrix(read.csv('Q1.csv'))
#多提取了 24 日 23：00 数据
Q1<-Q1[,c(1,2,7)]
cx<-matrix(NA,96,1)
for (i in 1:96){        #计算四天间隔两小时的 c 值   臭氧
    Q<-Q1[c(i,i+1),-c(1,2)]
    cx[i,]<-mean(Q)
}

CX<-matrix(NA,17,1)   #一天的 8 小时平均 c
CC<-matrix(NA,68,1)   #储存 4 天的平均 c
```

```
for (ii in c(1:4)) {     #计算 4 天的平滑平均 c 值（8h 平滑）
  p<-24*ii-23
  DAY<-cx[p:(p+23),]
  for (i in c(1:17)){   #8 小时的平均 c 值  只算了一天
    C_me<-mean(DAY[i:(i+7)])
    CX[i,]<- C_me
  }
  q<-17*ii-16
  CC[c(q:(q+16)),]<-CX
}
colnames(CC)<-c('O3 监测浓度')

MAXCO<-matrix(NA,4,1)
colnames(MAXCO)<-c('O3 监测浓度')
MINCO=MEANCO=MAXCO
for (i in 1:4){
  q<-17*i-16
  CO<-CC[q:(q+16),]   #一天的 CC 值
  MAXCO[i,]<-max(CO)
}
###################


pm10aqi<-function(x){
  if(x<=50){50*x/50}else
    if(x>50&x<=150){50*(x-50)/100+50}else
      if(x>150&x<=250){50*(x-150)/100+100}else
        if(x>250&x<=350){50*(x-250)/100+150}else
          if(x>350&x<=420){100*(x-350)/70+200}else
            if(x>420&x<=500){100*(x-420)/80+300}else
              if(x>500&x<=600){100*(x-500)/100+400}else
                if(x>600){0}

}
pm2.5aqi<-function(x){
  if(x<=35){50*x/35}else
    if(x>35&x<=75){50*(x-35)/40+50}else
      if(x>75&x<=115){50*(x-75)/40+100}else
        if(x>115&x<=150){50*(x-115)/35+150}else
          if(x>150&x<=250){100*(x-150)/100+200}else
            if(x>250&x<=350){100*(x-250)/100+300}else
              if(x>350&x<=500){100*(x-350)/150+400}else
                if(x<500){0}
```

```
}
so2aqi<-function(x){
   if(x<=50){50*x/50}else
      if(x>50&x<=150){50*(x-50)/100+50}else
         if(x>150&x<=475){50*(x-150)/325+100}else
            if(x>475&x<=800){50*(x-475)/325+150}else
               if(x>800&x<=1600){100*(x-800)/800+200}else
                  if(x>1600&x<=2100){100*(x-1600)/500+300}else
                     if(x>2100&x<=2620){100*(x-2100)/520+400}else
                        if(x>2620){0}
}
no2aqi<-function(x){
   if(x<=40){50*x/40}else
      if(x>40&x<=80){50*(x-40)/40+50}else
         if(x>80&x<=180){50*(x-80)/100+100}else
            if(x>180&x<=280){50*(x-180)/100+150}else
               if(x>280&x<=565){100*(x-280)/285+200}else
                  if(x>565&x<=750){100*(x-565)/185+300}else
                     if(x>750&x<=940){100*(x-750)/190+400}else
                        if(x>940){0}
}
coaqi<-function(x){
   if(x<=2){50*x/2}else
      if(x>2&x<=4){50*(x-2)/2+50}else
         if(x>4&x<=14){50*(x-4)/10+100}else
            if(x>14&x<=24){50*(x-14)/10+150}else
               if(x>24&x<=36){100*(x-24)/12+200}else
                  if(x>36&x<=48){100*(x-36)/12+300}else
                     if(x>48&x<=60){100*(x-48)/12+400}else
                        if(x>60){0}
}
o3aqi<-function(x){
   if(x<=100){50*x/100}else
      if(x>100&x<=160){50*(x-100)/60+50}else
         if(x>160&x<=215){50*(x-160)/55+100}else
            if(x>215&x<=265){50*(x-215)/50+150}else
               if(x>265&x<=800){100*(x-265)/535+200}else
                  if(x>800){0}
}
```

```
Q1.1<-read.csv('timu2.csv')
```

```
Q1.1
pm2.5<-Q1.1[,5]
pm10<-Q1.1[,4]
so2<-Q1.1[,2]
no2<-Q1.1[,3]
co<-Q1.1[,7]
o3<-Q1.1[,6]
len<-length(pm2.5)
IAQIpm2.5<-c()
IAQIpm10<-c()
IAQIso2<-c()
IAQIno2<-c()
IAQIo3<-c()
IAQIco<-c()
for (i in 1:len){ #pm2.5
  x<-pm2.5[i]
  aqi<-pm2.5aqi(x)
  aqi<-round(aqi)
  IAQIpm2.5<-c(IAQIpm2.5,aqi)
}
for (i in 1:len){ #pm10
  x<-pm10[i]
  aqi<-pm10aqi(x)
  aqi<-round(aqi)
  IAQIpm10<-c(IAQIpm10,aqi)
}
for (i in 1:len){ #pm10
  x<-so2[i]
  aqi<-so2aqi(x)
  aqi<-round(aqi)
  IAQIso2<-c(IAQIso2,aqi)
}
for (i in 1:len){ #pm10
  x<-no2[i]
  aqi<-no2aqi(x)
  aqi<-round(aqi)
  IAQIno2<-c(IAQIno2,aqi)
}

for (i in 1:len){ #pm10
  x<-o3[i]
  aqi<-o3aqi(x)
  aqi<-round(aqi)
  IAQIo3<-c(IAQIo3,aqi)
```

```
}
for (i in 1:len){ #pm10
   x<-co[i]
   aqi<-coaqi(x)
   aqi<-round(aqi)
   IAQIco<-c(IAQIco,aqi)
}

dimm<-dim(Q1.1)
MA<-matrix(NA,dimm[1],6)
MA[,1]<-IAQIco
MA[,2]<-IAQIso2
MA[,3]<-IAQIno2
MA[,4]<-IAQIo3
MA[,5]<-IAQIpm10
MA[,6]<-IAQIpm2.5
colnames(MA)<-c('co','so2','no2','o3','pm10','pm2.5')

AQII<-matrix(NA,dimm[1],3)
AQII[,1]<-apply(MA,1,max)
for (i in 1:dimm[1]){
   if(length(names(which(MA[i,]==AQII[i,1])))==1){
      AQII[i,2]<-names(which(MA[i,]==AQII[i,1]))
   }else if(length(names(which(MA[i,]==AQII[i,1])))==2){
      AQII[i,2:3]<-names(which(MA[i,]==AQII[i,1]))
   }
}

AQII[,1]<-as.numeric(AQII[,1])

write.csv(AQII,'Q2.2.csv')
for (i in 1:dimm[1]){
   id<-which(MA[i,]==AQII[i,])
   if (id==1){"

      }
}
setwd('D:/')
data<-seq(from = as.Date('2020/7/23'), to = as.Date('2021/7/12'), by = '1 day')
repp<-matrix(NA,8520,1)
repp[,1]<-as.character(rep(data,each=24))
write.csv(repp,'repp.csv')
```

```
####计算均值
A1<-read.csv('Q3.C 修正.csv')  #5:25 O3 在 24
r_nam<-as.character(seq(from = as.Date('2020/7/24'), to = as.Date('2021/7/12'), by = '1 day'))

A<-A1[,2:22]
nam<-colnames(A)
B<-matrix(NA,354,21)  #装均值的
#B<-matrix(NA,3,21)
i=1#算初始
 o3_0<-A[(24*i-23):(24*i),20][24]
#o3_0<-A[1,20]

for (i in 2:354){  #天数
  A2<-A[(24*i-23):(24*i),c(1:19,21)]  #除 o3 外的 24 小时均值
  B[i-1,c(1:19,21)]<-apply(A2,2,mean)
  c<-c()
  o3_1<-c(o3_0,A[(24*i-23):(24*i),20])
  o3_0<-A[(24*i-23):(24*i),20][24]
  for (ii in 1:25) {
    c<-c(c,(o3_1[ii]+o3_1[ii+1])/2)
  }
  c<-c[1:24]
  avg_c<-c()
    for (iii in 1:17) {
      avg_c<-c(mean(c[iii:(iii+7)]),avg_c)
    }
  C<-max(avg_c)
  B[i-1,20]<-C
}
colnames(B)<-nam
rownames(B)<-r_nam  #A 地平均
B
write.csv(B[,16:21],'Q3_C_predict_365day.csv')

dat<-read.csv('timu3C.csv')
#真实数据
colnames(dat)
colnames(B)
true<-dat[,which(colnames(dat)=="CO 实测日均.mg.m3."  )]
pre<-B[,which(colnames(B)== "CO 小时平均浓度.mg.m3.")]
e<-true-pre
#e_A<-e
#e_B<-e
#e_C<-e
```

```
e_A<-as.numeric(e_A)
e_B<-as.numeric(e_B)
e_C<-as.numeric(e_C)
AA<-data.frame(e_A,e_B,e_C)
write.csv(AA,'CO.csv')

plot(x,'l')
library('forecast')
ndiffs(x) #判断几阶段差分
x.dif<-diff(x,difference=2)
x.dif=ts(x.dif,start=1992)  #几阶差分就要原来的年份加几
plot(x.dif,xlab="年份",ylab="海洋捕捞产品产量（万吨）")

acf(x.dif)  #模型识别
pacf(x.dif)
auto.arima(x) #自动识别模型
x.fit<-arima(x,order=c(0,2,1),include.mean=F) #此处要用原序列
x.fit

accuracy(x.fit)  #MAAE>1 不好
#残差进行白噪声检验（p》0.05，是白噪声 符合期望）
for(i in 1:2)print(Box.test(x.fit$residual,type="Ljung-Box",lag=6*i))

#系数白噪声检验 df=序列长度-变量数 p<0.05 系数显著
t1=-0.6966/0.1575
#若参数估计值 t1 为负值，tail=T，否则为 F
pt(t1,df=28,lower.tail=T)

#预测
a.fore=forecast(x.fit,h=7)
a.fore
plot(a.fore)
#画图
L1<-a.fore$fitted-1.96*sqrt(x.fit$sigma2)
U1<-a.fore$fitted+1.96*sqrt(x.fit$sigma2)
L2<-ts(a.fore$lower[,2],start=2019)  #预测的开始年份
U2<-ts(a.fore$upper[,2],start=2019)
c1<-min(x,x.fit)
c2<-max(x,x.fit)
plot(x,type="o",pch=8,xlim=c(1963,1976),ylim=c(c1,c2))
lines(x.fit,col=4,lty=2)
```

```
write.csv(e,'e.csv')

length(e)
e1<-ts(e,frequency = 365,start = c(2020.07))
plot(e1,type = "l")
true_hour<-read.csv('aa.csv') #实测数据  每时
y<-e
XF<-B[,1:15]
```