



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

学 校

宁波大学

参赛队号

21116460002

1.陈康鑫

队员姓名

2.张强

3.谢敏

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：

本文研究了空气质量的二次预报问题，主要创新点在于：(1)(2) 针对任务一，使用了基于半参数的差值网络模型来完成数据缺失值的处理，并结合箱型图提出异常值。结合给定的 AQI 计算方法求出了 2020 年 8 月 25 日到 8 月 28 日每天实测的 AQI 和首要污染物。

针对任务二，充分考虑了时间滞后性问题，通过计算各气象条件峰值与污染物浓度峰值的时间差，降低了时间滞后性带来的 AQI 计算不准确性。在计算了 AQI 之后，通过 AQI 的等级，将污染物分成了 3 类，最后采用了谱聚类的方法，计算在不同污染物浓度类别下的气象条件分类情况，最终将气象条件分成了三类。

针对任务三，首先根据问题描述的 A、B、C 之间的影响忽略不计而舍弃掉对地理位置有影响的近地风向这一解释变量，接着通过分析各数据之间的共性关系完成对影响因子的筛选，为后期的预测模型提供有关自变量。在预分析完成之后，进行并行处理，第一步分别将 6 项污染物作为响应变量，每次选取剩余的影响因素中的一项作为解释变量，进行 GAM 分析，并行步则是分别将 6 项污染物作为响应变量，与所有的预分析后的影响因素进行 GAM 模型分析，综合并行处理的结果，选取出对 6 项污染物浓度变化解释率等较高的气象条件作为预测模型的变量，将最终选定的多项解释变量送入 GAM 模型，完成最终污染物浓度的预测模型的搭建。

针对任务四，为了实现位宽自动优化并综合考虑性能和资源两种因素，考虑到资源与芯片实现面积和功耗有关，首先建立资源使用量化评价函数。然后以 RSNR 代价作为性能评定指标，以不同位宽向量下的 RSNR 代价与资源使用评价函数分别赋予权重因子，组成综合评价目标函数。最后提出基于禁忌搜索算法的位宽自动优化方法，以得到性能和代价综合评价目标函数的最优 Pareto 解集，即最优位宽设计方案。

最后本文对 CR 算法的相位噪声提取效果以及整个信号传输系统的抗噪声性能进行了检验，检验结果为：该算法能较好地提取出信道中的相位噪声；该算法提高了整个系统的抗噪声性能。

关键词： 仿真框架；CR 算法；噪声耦合；去耦合；ASIC 算法设计；量化噪声

目录

1. 问题重述	3
1.1 问题背景	3
1.2 问题提出	3
2. 模型假设	3
3. 符号说明	4
4. 问题一的模型建立与求解	4
4.1 问题一的描述分析	4
4.2 数据预处理	4
4.3 模型求解和分析	7
5. 问题二的模型建立与求解	7
5.1 问题二的描述分析	7
5.2 时间滞后性的考虑	8
5.3 模型求解和相关性分析	8
5.4 数学模型的建立	8
6. 问题三的模型建立与求解	10
6.1 问题三的描述分析	10
6.2 数学模型的建立	10
6.3 模型求解和分析	12
7. 问题四的模型建立与求解	15
7.1 问题四的描述分析	15
7.2 数学模型的建立	15
7.3 模型求解和分析	15
8. 模型评价	17
8.1 模型的优点	17
8.2 模型的缺点	17
参考文献	17
附录 A 主程序源代码	18

1. 问题重述

1.1 问题背景

大气污染系指由于人类活动或自然过程引起某些物质进入大气中，呈现足够的浓度，达到了足够的时间，并因此危害了人体的舒适、健康和福利或危害了生态环境 [1]。污染防治实践表明，建立空气质量预报模型，提前获知可能发生的大气污染过程并采取相应控制措施，是减少大气污染对人体健康和环境等造成的危害，提高环境空气质量的有效方法之一。

目前常用 WRF-CMAQ 模拟体系（以下简称 WRF-CMAQ 模型）对空气质量进行预报。WRF-CMAQ 模型主要包括 WRF 和 CMAQ 两部分：WRF 是一种中尺度数值天气预报系统，用于为 CMAQ 提供所需的气象场数据；CMAQ 是一种三维欧拉大气化学与传输模拟系统，其根据来自 WRF 的气象信息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的变化过程，继而得到具体时间点或时间段的预报结果。

但受制于模拟的气象场以及排放清单的不确定性，以及对包括臭氧在内的污染物生成机理的不完全明晰，WRF-CMAQ 预报模型的结果并不理想。因此题目提出了二次建模概念：即指在 WRF-CMAQ 等一次预报模型模拟结果的基础上，结合更多的数据源进行再建模，以提高预报的准确性。

针对一次预测不准确的问题，本文二次建模的主要目的有：一是结合已有数据与一次预测结果对二次污染物的生成进行分析；二是提升预测精度。

1.2 问题提出

通过数学建模，解决了以下问题：

任务一：使用附件 1 中的数据，按照附录中的方法计算监测点 A 从 2020 年 8 月 25 日到 8 月 28 日每天实测的空气质量指数 (Air Quality Index, AQI) 和首要污染物，将结果按照附录“AQI 计算结果表”的格式放在正文中。

任务二：在污染物排放情况不变的条件下，某一地区的气象条件有利于污染物扩散或沉降时，该地区的 AQI 会下降，反之会上升。使用附件 1 中的数据，根据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。

任务三：使用附件 1、2 中的数据，建立一个同时适用于 A、B、C 三个监测点（监测点两两间直线距离 $>100\text{km}$ ，忽略相互影响）的二次预报数学模型，用来预测未来三天 6 种常规污染物单日浓度值，要求二次预报模型预测结果中 AQI 预报值的最大相对误差应尽量小，且首要污染物预测准确度尽量高。并使用该模型预测监测点 A、B、C 在 2021 年 7 月 13 日至 7 月 15 日 6 种常规污染物的单日浓度值，计算相应的 AQI 和首要污染物，将结果依照附录“污染物浓度及 AQI 预测结果表”的格式放在论文中。

任务四：相邻区域的污染物浓度往往具有一定的相关性，区域协同预报可能会提升空气质量预报的准确度。如图 4，监测点 A 的临近区域内存在监测点 A1、A2、A3，使用附件 1、3 中的数据，建立包含 A、A1、A2、A3 四个监测点的协同预报模型，要求二次模型预测结果中 AQI 预报值的最大相对误差应尽量小，且首要污染物预测准确度尽量高。使用该模型预测监测点 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，计算相应的 AQI 和首要污染物，将结果依照附录“污染物浓度及 AQI 预测结果表”的格式放在论文中。并讨论：与问题 3 的模型相比，协同预报模型能否提升针对监测点 A 的污染物浓度预报准确度？说明原因。

2. 模型假设

1. 假设近地面臭氧污染形成机制中的化学反应实现了完全转化，且臭氧前产物如 NO_2 的含量会在自然界得到补充；
2. 假设臭氧浓度数值在二次污染中，只受给定的气象数据以及其余五项污染物的浓度影响，汽车尾气、工业生产、全球变化等题目中未出现的影响因素不在本模型的考虑范围内；
3. 本文以 $O_3 - 8h$ 的滑动平均值为一次臭氧的空气质量影响，不进行 $O_3 - 1h$ 的空气质量评价；
4. 假设在二次污染期间，高层大气中的氧气不会经过系列转化形成新的臭氧；
5. 假设在二次污染期间，没有雷雨天气产生臭氧；
6. 假设空气质量只受地表臭氧浓度影响，高空如平流层的臭氧浓度不影响空气质量的判定；
7. 假设四季变化对空气质量没有影响
8. 假设本文的二次污染物只有臭氧，没有有机颗粒物等其他污染物质；
9. 假设二次污染时，臭氧的形成只需考虑 NO_2 的含量影响，不需要考虑 VOCs 以及 NO_x 的含量影响；
10. 假设本文模型不受地形特点影响；

3. 符号说明

符号	说明
φ_d	低通平滑插值函数
δ_d	高通平滑插值函数
τ_d	强度函数
r_k	时间点
L_{dn}	时间序列中总的观测数
t_d	第 d 维时间序列
p_d	第 d 维观测值列表
$\gamma_{dd'}$	可学习相关性
σ_d	瞬态分量
λ_d	平滑交叉维度插值
T_2	近地 2 米温度
T	地表温度
R_Hu	比湿
Hu	湿度
Win_S	近地 10 米风速
$Win_$	近地 10 米风向
$Rain$	雨量
$Cloud$	云量
$High_B$	边界层高度
$Atmos$	大气压
Sen_H	感热通量
Lat_H	潜热通量
$Long_W$	长波辐射
$Short_W$	短波辐射
$Solar_W$	地面太阳能辐射
$k(\mu_i)$	连接函数
h_i	平滑函数
$X_i\theta$	全参数模型成分
ε_i	残差

4. 问题一的模型建立与求解

4.1 问题一的描述分析

问题一要求利用题目所提供的数据和公式，计算空气质量指数。题目提供的数据来自监测点长期空气质量预报基础数据，包括污染物浓度一次预报数据、气象一次预报数据、气象实测数据和污染物浓度实测数据。本文首先对给定数据进行了可视化分析，发现存在缺失值以及异常值，针对异常值采用了箱型图剔除法；针对缺失值，采用了基于半参数的插值网络，该网络充分考虑了时序性问题。在完成数值的预处理后利用给定的 AQI 计算方式完成问题一的求解。

4.2 数据预处理

在计算实测污染物的 AQI 之前，需要对数据进行统一的预处理，即数据清洗。其中，缺失值是本文主要处理的对象。从缺失数据的缺失影响因素来看，分为完全随机缺失 (Missing Completely At Random, MCAR)、随机缺失 (Missing At Random, MAR)、非随机缺失 (Missing Not At Random, MNAR) 角度考虑^[2]。一般来说，如果数据中缺失观测值的比例相对于观测值总数很小，最简单的方式是删除带有缺失项的样本，即完全数据分析。当缺失项较多时，由于一部分的数据信息缺失，完全数据分析方法的偏差很大。如本课题中，附件 2 中监测点 C 逐小时污染物浓度与气象实测数据中湿度信息的缺失就有 6147 条，缺失比率占整个湿度信息的 31.6%，直接删除包含缺失值的行可能会导致放弃有用的信息。从图1题目给定信息的数据缺失量化图可以看出，信息数据的丢失比较多，不能简单的将缺失值进行丢弃。除开缺失值外，还有异常数据，图2为六个监测点 A、B、C、A1、A2 的六个污染物浓度的箱型图，其余信息的箱型图请见附件，针对异常值的处理，本文采用箱型图进行剔除。

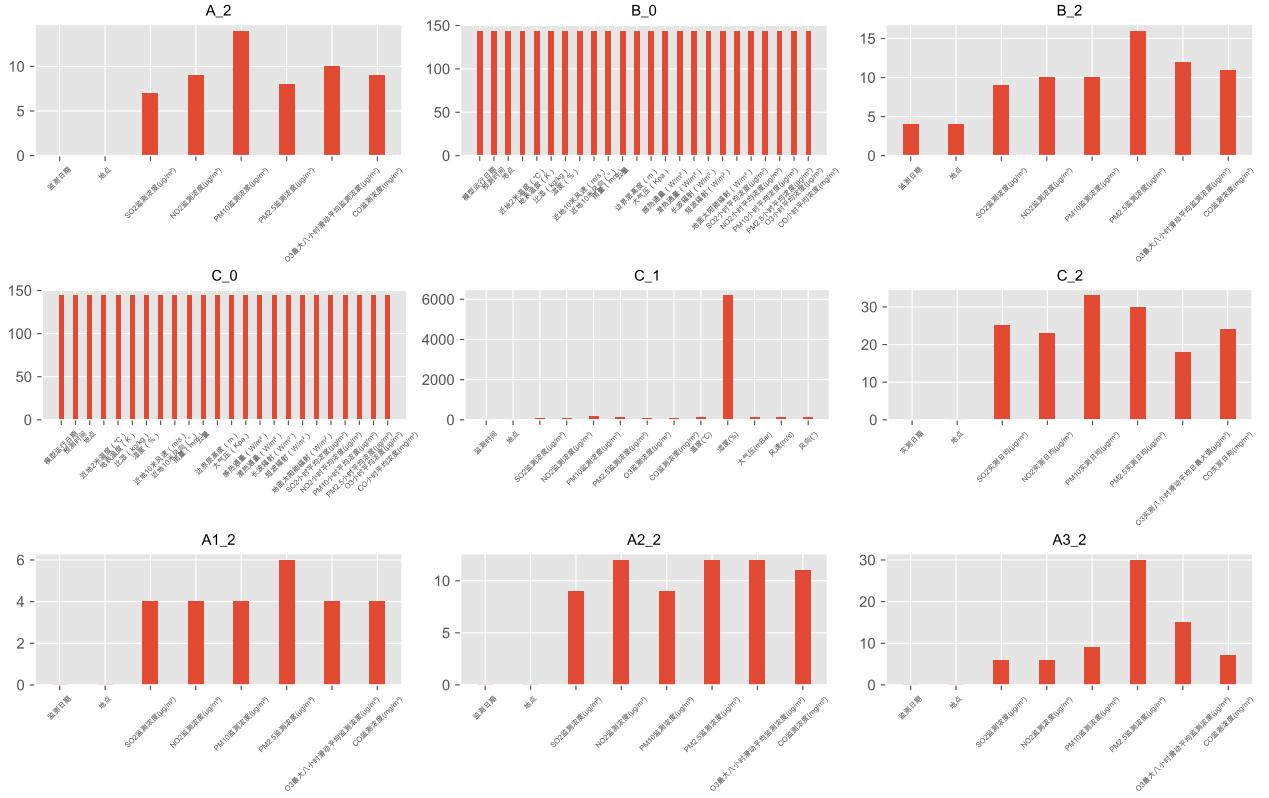


图 1: 数据缺失量化图

纠正由缺失数据导致的结论偏倚，缺失数据处理方法相继被提出。通过距离测量来识别相邻点，并通过相邻点观测值的完整值来估计缺失值的方法是一种常见的插补方法，如 K 近邻 (K-Nearest Neighbor, KNN) 算法^[3]。这种插值方法虽然有效快速，但针对时序问题时，利用距离信息进行插补的方法则显得捉襟见肘。本文是一个预测空气质量的问题，与时间息息相关，因此本文采用了基于半参数插值网络方法^[4]，这种插值方法允许在插值阶段跨多元时间序列的多个纬度共享信息。

半参数插值网络模型如图3所示，网络第一层是对 D 个时间序列中的每一个序列分别执行三个半参数单变量变换，每个变换都是基于径向基函数 (Radial Basis Function, RBF)，用以适应连续的时间观测。这三个变换分别是一个低通 (平滑) 插值函数 φ_d ，一个高通 (非平滑) 插值函数 δ_d 以及一个强度函数 τ_d 。三个转换在时间点 r_k 处的计算公式如下所示：

$$M(c, \mathbf{t}, \varphi) = \sum_{t \in \mathbf{t}} g(c, t, \varphi), \quad g(c, t, \varphi) = \exp(-\varphi(c - t)^2) \quad (4.1)$$

$$\tau_{ad} = f_\theta^\tau(c_a, \mathbf{t}_d, \mathbf{p}_d) = M(c_a, \mathbf{t}_d, \varphi_d) \quad (4.2)$$

$$\epsilon_{ad} = f_\theta^\epsilon(c_a, \mathbf{t}_d, \mathbf{p}_d) = \frac{1}{M(c_a, \mathbf{t}_d, \varphi_d)} \sum_{j=1}^{L_{dn}} g(c_a, t_{jd}, \varphi_d) p_{jd} \quad (4.3)$$

$$\delta_{ad} = f_\theta^\delta(r_a, \mathbf{t}_d, \mathbf{p}_d) = \frac{1}{M(r_a, \mathbf{t}_d, k\varphi_d)} \sum_{j=1}^{L_{dn}} g(c_a, t_{jd}, a\varphi_d) p_{jd} \quad (4.4)$$

其中， L_{dn} 为时间序列中总的观测数， t_d 表示第 d 维的时间序列， φ_d 采用的是参数为 φ_d 平方指数核， δ_d 采用的是参数为 $a\varphi_d$ 的平方指数核，其中 $a > 1$ 。 $\mathbf{s}_d = \{t_d, p_d\}$ 是一个第 n 个数据案例的时间序列 d 的远足， $\mathbf{t}_d = t_{1d}, \dots, t_{L_{dn}}$ 是定义观测值的时间点列表， $\mathbf{p}_d = p_{1d}, \dots, p_{L_{dn}}$ 是相应的观测值列表。

第二层通过考虑所有时间序列的可学习相关性 $\gamma_{dd'}$ ，合并每个参考时间点的所有 D 个时间序列的信息，即对每个 d 维的输入序列插入一个交叉维度插值 λ_d 。此外，本文为每个输入维度 d 定义了一个瞬态分量 σ_d ，

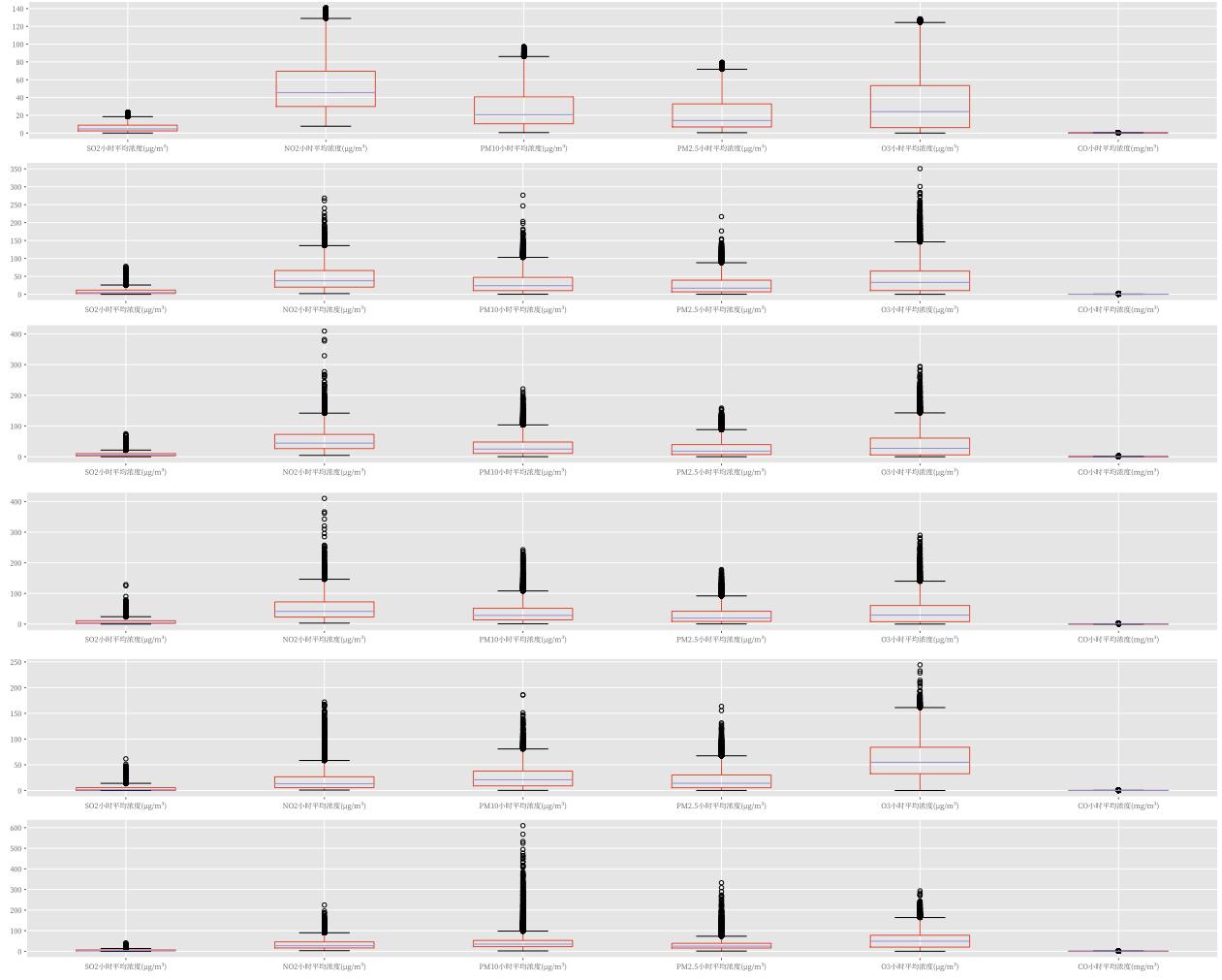


图 2: 数据异常值量化图

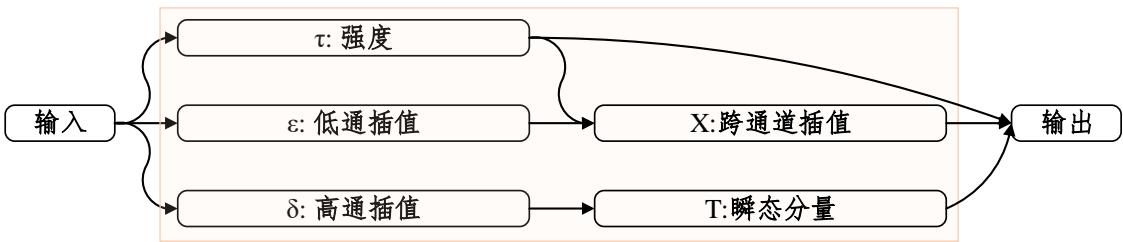


图 3: 插值网络

该值为第一层高通插值 δ_d 与平滑交叉维度插值 λ_d 之间的差值，计算公式如下：

$$\lambda_{ad} = f_\theta^\lambda(c_a, \mathbf{s}) = \frac{\sum_{d'} \gamma_{dd'} \beta_{ad'} \epsilon_{ad'}}{\sum_{d'} \beta_{ad'}}, \quad \sigma_{ad} = f_\theta^\sigma(c_a, \mathbf{s}) = \delta_{ad} - \lambda_{ad} \quad (4.5)$$

我们使用平滑交叉维度插值 λ_d 来获取平滑的趋势，瞬态分量 σ_d 来捕获瞬态，强度函数 τ_d 来捕获关于观测及时发生的信息。

4.3 模型求解和分析

对数据清洗完毕后，按照附录中的 AQI 计算方法可以测出监测点 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。在计算 AQI 之前需要计算各项污染物的空气质量分指数 (IAQI)，其计算公式如下：

$$IAQI_P = \frac{IAQI_{Hi} - IAQI_{LO}}{BP_{Hi} - BP_{LO}} \cdot (C_P - BP_{LO}) + IAQI_{LO} \quad (4.6)$$

其中， $IAQI_P$ 表示污染物 P 的空气质量分指数， C_P 表示污染物 P 的质量浓度值， BP_{Hi} 和 BP_{LO} 表示与 C_P 相近的污染物浓度限值的高位值与低位值， $IAQI_{Hi}$ 和 $IAQI_{LO}$ 表示与 BP_{Hi} 、 BP_{LO} 对应的空气质量分数。各项污染物项目浓度限值及对应的空气质量分指数级别如表1所示。

表 1: IAQI 及对应的污染物浓度限值

指数或污染物项目	IAQI 及对应的污染物浓度限值							单位
IAQI	0	50	100	150	200	300	400	500
CO_{24} 小时平均	0	2	4	14	24	36	48	60
SO_2 24 小时平均	0	50	150	475	800	1600	2100	2620
NO_2 24 小时平均	0	40	80	180	280	565	750	940
O_3 -Max8h	0	100	160	215	265	800	-	-
粒径小于等于 10 PM_{10} 24 小时平均	0	50	150	250	350	420	500	$\mu g/m^3$
粒径小于等于 2.5 $PM_{2.5}$ 24 小时平均	0	35	75	115	150	250	350	500

¹ 臭氧 (O_3) 最大 8 小时滑动平均浓度值高于 $800 \mu g/m^3$ 的，不再进行其空气质量分指数计算。

² 其余污染物浓度高于 $IAQI=500$ 对应限值时，不再进行其空气质量分指数计算。

在计算完单个 IAQI 之后，空气质量指数 (AQI) 取各分指数中的最大值，即空气质量等级范围根据 AQI 数值划分，等级对应的 AQI 范围如表2所示。

$$AQI = \max IAQI_{SO_2}, IAQI_{NO_2}, IAQI_{PM_{10}}, IAQI_{PM_{2.5}}, IAQI_{O_3}, IAQI_{CO} \quad (4.7)$$

其中 $IAQI_{SO_2}, IAQI_{NO_2}, IAQI_{PM_{10}}, IAQI_{PM_{2.5}}, IAQI_{O_3}, IAQI_{CO}$ 为各污染物的分指数，

表 2: 空气质量等级及对应空气质量质数 (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,+)

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；IAQI 大于 100 的污染物为超标污染物。

在完成数据预处理之后，经过 IAQI 的计算，得到了监测点 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，结果如表3所示。从表中数据可以看出，监测的四天中只有 2020 年 8 月 26 日无首要污染物，其余三天的首要污染物均为 O_3 。因为在计算当前结果时，没有考虑气象条件对污染物的影响 (如风速对监测点 A 的污染物浓度的影响)，也没有考虑二次污染的情况，所以从表中数据的 AQI 计算结果缺乏一定的真实性，需要在全面综合的考虑气象条件以及各污染物之间的相关性后重新计算，完成对空气质量的可靠估计。

5. 问题二的模型建立与求解

5.1 问题二的描述分析

问题二是在污染物排放情况不变的条件下，考虑气象条件。利用指定的数据，根据污染物对气象条件的影响程度，对气象条件进行合理分类，并对分类后的气象条件特性进行分析。针对这个问题，本文首先考虑到的是时间滞后性问题，通过分析各气象条件与 6 项污染物之间的峰值，计算出了各影响因子之间的时间滞后性，为使得 AQI 的结果更加贴近实际的预测，我们将计算出的滞后时间进行了调整，使得各影响因子的峰值实现了跨时组合。

表 3: AQI 计算结果

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
2020/8/27	监测点 A	108	O_3
2020/8/28	监测点 A	137	O_3

5.2 时间滞后性的考虑

空气质量指数受到多种气象条件的影响，而气象条件由于其本身特性，会导致 AQI 的变化产生一定的滞后性。

5.3 模型求解和相关性分析

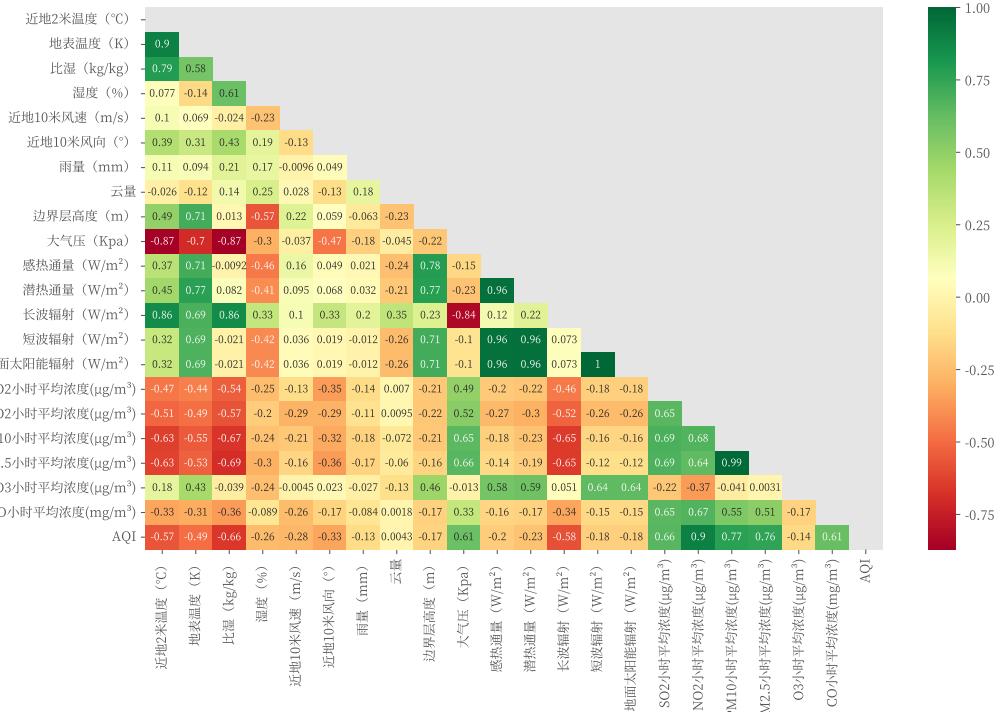


图 4: 相关性分析

根据如图所示的相关系数热力图(其他热力图请见附件)，可以直观地观察到，温度，湿度，风速，近地 2 米温度，地表温度，比湿，湿度，边界层高度，感热通量，潜热通量，短波辐射，地面太阳辐射这 12 个指标与污染物浓度相关性较高。

5.4 数学模型的建立

常用聚类方法有 K-Means (K 均值) 聚类，用高斯混合模型 (GMM) 的最大期望 (EM) 聚类，谱聚类等等。

本题采用谱聚类方式。因为谱聚类相对于传统的 K-Means 算法，谱聚类对数据分布的适应性更强，聚类效果也很优秀，同时聚类的计算量也小很多。

聚类结果如下图所示：

簇中心气象指标：

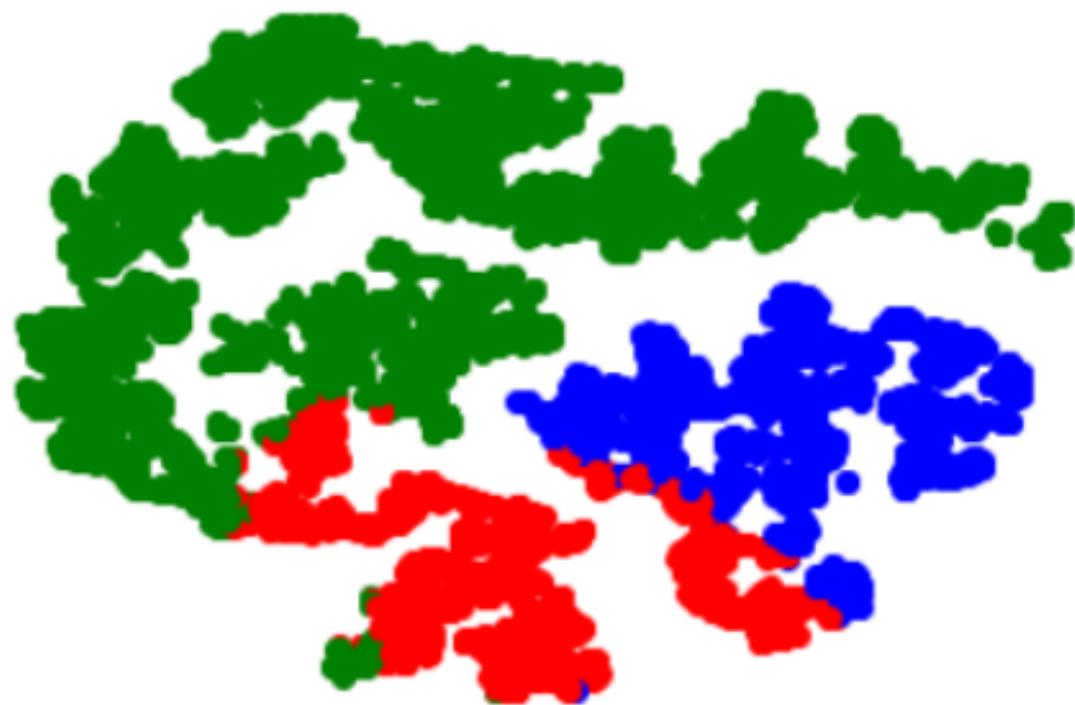


图 5: 聚类结果

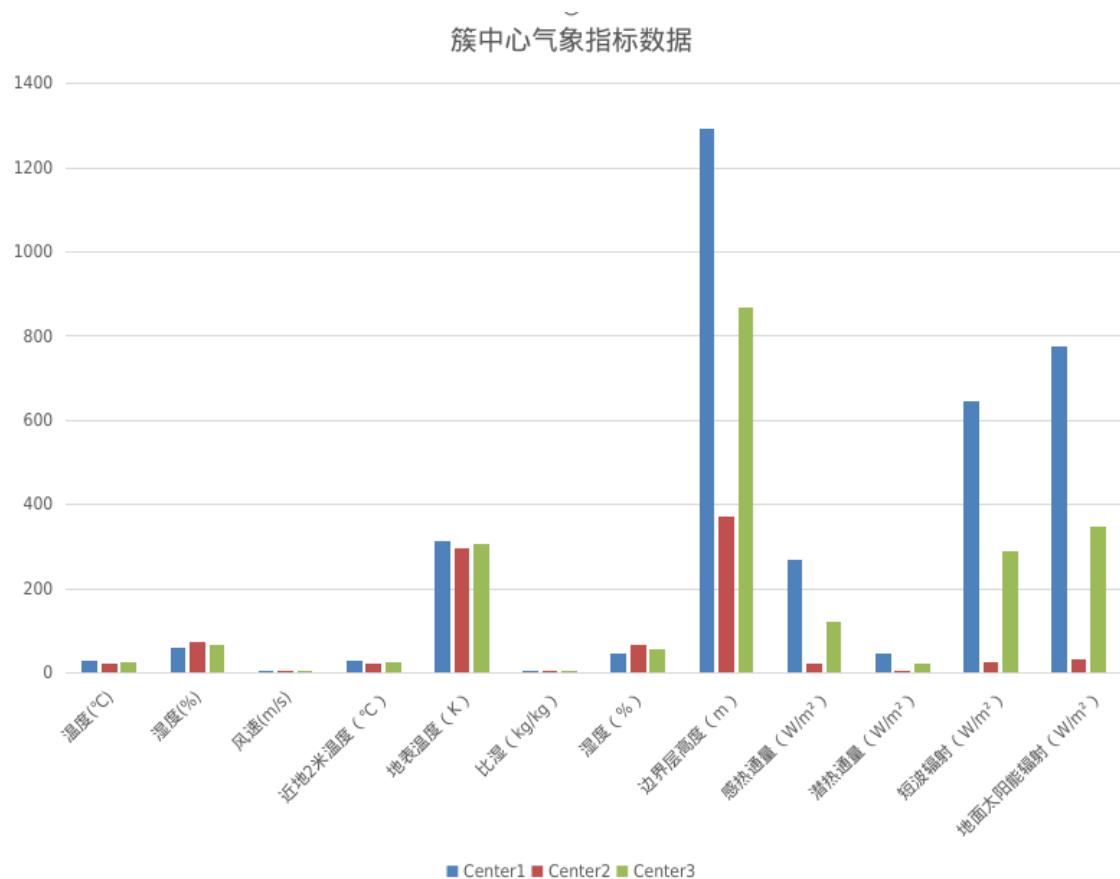


图 6: 簇中心气象指标

由聚类结果图和中心气象指标图示可知，分成三类：一类气象条件，湿度低但温度风速大二类气象条件正好相反，温度低，湿度高，风速低三类气象条件处于一类和二类之间，温度适中

6. 问题三的模型建立与求解

问题三要求利用附件 1 和 2 中的数据，建立一个同时适用于 A、B、C 三个监测点的二次预报数学模型，三个监测点之间的直线距离大于 100km 且互不影响，用该模型预测 A、B、C 三个监测点在 2021 年 7 月 13 日至 7 月 15 日这三天的 6 种常规污染物单日浓度值，要求二次预报模型预测结果中 AQI 预报值的最大相对误差应尽量小，且首要污染物预测准确度尽量高。

已知信息

问题三提供的数据有三个：

- (1) 三个监测点 A、B、C 的逐小时污染物浓度与气象一次预报数据表；
- (2) 三个监测点 A、B、C 的逐小时污染物浓度与气象实测数据表；
- (3) 三个监测点 A、B、C 的逐日污染物浓度实测数据表。

6.1 问题三的描述分析

因为题目中说 A、B、C 三个站点之间的影响可以忽略，但气象条件下包含了风向等受地理位置影响的解释变量，所以不能直接将此类的数据直接用于预测模型的构建，需要将此类的解释变量进行剔除，再进行其他分析。本文首先对解释变量进行了预分析，剔除掉对地理位置有影响的变量以及存在共线性的变量，降低多重共线性对模型的影响。预分析之后进行一个并行处理，第一步分别将 6 项污染物作为响应变量，每次选取剩余的影响因素中的一项作为解释变量，进行广义相加模型 (Generalized Additive Model, GAM) 分析^[5]，并行步则是分别将 6 项污染物作为响应变量，与所有的预分析后的影响因素进行 GAM 模型分析，综合并行处理的结果，选取对 6 项污染物浓度变化解释率等较高的气象条件作为预测模型的变量，将最终选定的多项解释变量送入 GAM 模型，完成最终污染物浓度的预测模型的搭建。

多重共线性是指线性回归模型中的解释变量之间存在高度相关关系而使模型估计失真或难以准确估计^[6]。本文采用皮尔斯相关系数来判别解释变量之间是否存在的共线性，如果解释变量间相关系数较大，则两个解释变量之间通常存在严重共曲线性关系，需要剔除一个解释变量，用留下的解释变量指标来代表丢弃变量指标。

GAM 是一种非参数回归模型，模型中部分或全部的自变量采用平滑函数，以此来降低线性拟合带来的风险，属于组合模型。该模型不需要预假设模型中自变量是否线性相关于因变量，对于模型的假定比较宽松，可以解决 Logistic 回归多解释变量时引发的维度灾难问题。

6.2 数学模型的建立

本文首先将检测点 B 和 C 中的明显缺失值采用了问题一的方法进行填充，[加图]为检测点 C 湿度条件的缺失值补充结果图，从图中结果可以看出，本文采用的缺失值填补方法能有效的维护数据原始的分布特征特性。缺失值处理完毕后，本文对数据进行了预分析，得到了解释变量自身的频率分布直方图与密度分布曲线，如图 8 所示，从图中可以看出，21 个解释变量基本符合正态分布类型，其中，近地 2 米温度和地表温度的数据分布走向基本类似，比湿与湿度的主要数据分布基本类似，考虑到题目中所说的 A、B、C 三个监测点之间的影响忽略不计，需要剔除近地风向这一影响因素。

为了更进一步量化解释变量之间的相关性，本文采用了双变量的 Pearson 相关系数分析，结果如表 4 所示，从表中数据可以看到，近地 2 米温度与地表温度、长波辐射，长波辐射与比湿，潜热通量与感热通量、短波辐射、地面太阳能辐射，短波辐射与感热通量、潜热通量、地面太阳能辐射，感热通量与地面太阳能辐射之间的相关系数均在 85% 以上，在 $P < 0.05$ 和 $P < 0.01$ 上显著相关，而所有的数值中，短波辐射和地面太阳能辐射之间相关系数最高，从概率分布直方图和密度分布曲线也可以看出，两个影响因素具有极高的共线性，那是因为太阳辐射波长较地面和大气辐射波长小得多，所以通常又称太阳辐射为短波辐射，因此在计算相关系数时，呈现了高度的互相关性，考虑到太阳能辐射与其他几个也呈现了较高的相关性，于是本文舍弃掉了太阳能辐射这一变量，采用其他的标量指标来代替太阳能辐射，以此减少构建多变量曲线模型能产生的解释变量共曲线性问题。

因为 6 项污染物中只有 O_3 涉及二次污染，其他的 5 项的数据可以直接从给定材料获取，二次污染的数据则需要考虑 O_3 的产生方式来判断新的数据量。地表 O_3 含量一部分来自平流层的输入，但大部分来自自然界排放的碳水化合物和人类活动排放的 NO_x 、 $NMHC$ 和 CO 等臭氧前体物，这些前体物经过光化学反应可以促进低层大气产生臭氧和过氧乙酰硝酸酯等二次污染物 O_3 。臭氧的形成需要一定的气象条件，由于反应过程需要紫外线的参与，因此在光照好、湿度高、温度低的夏季白天往往会出现臭氧浓度的升高，在无风

表 4: 影响因素间 Person 相关系数

	T_2	T	R_Hu	Hu	Win_S	Win_D	Rain	Cloud	High_B	Atmos	Sen_H	Lat_H	Long_W	Short_W	Solar_R	SO2	NO2	PM10	PM2.5	O3	CO
T	1	.901***	.790**	.077***	.102***	.389***	.105***	-.026***	.493***	-.867***	.366***	.449***	.320***	.858***	.469***	-.515***	-.632***	.184***	-.330***		
T		.901***	1	.581**	-.142**	.069***	.307***	.094***	-.116***	.711***	.700***	.769***	.693***	.686***	.441***	-.490***	-.553***	-.529***	.431***	-.309***	
R_Hu	.790**		.581**	1	.607***	-.024***	.024***	.428***	.208***	.136***	.013*	-.871***	-.0009	.082***	.859***	-.021***	.537***	-.574***	-.671***	-.039***	-.364***
Hu	.077***		.581**	1	.607***	-.142**	.069***	.307***	.094***	-.116***	.711***	.700***	.769***	.693***	.686***	.441***	-.490***	-.553***	-.529***	.431***	-.309***
Win_S	.102***		.607***	1	.607***	-.142**	.069***	.307***	.094***	-.116***	.711***	.700***	.769***	.693***	.686***	.441***	-.490***	-.553***	-.529***	.431***	-.309***
Win_D	.389***		.069***	1	.024***	-.232***	.193***	.193***	.173***	.173***	.136***	.136***	.136***	.136***	.136***	.136***	.136***	.136***	.136***	.136***	.136***
Rain	.307***		.024***	1	.049***	-.134***	.1	.049***	-.134***	.1	.049***	-.134***	.1	.049***	-.134***	.1	.049***	-.134***	.1	.049***	-.134***
Cloud	.105***		.094***	1	.173***	-.134***	.193***	.193***	.173***	.173***	.136***	.136***	.136***	.136***	.136***	.136***	.136***	.136***	.136***	.136***	.136***
High_B	.493***		.136***	1	.116***	-.116***	.028***	.028***	.028***	.028***	.028***	.028***	.028***	.028***	.028***	.028***	.028***	.028***	.028***	.028***	.028***
Atmos	.867***		.711**	1	.013*	-.568***	.250***	.059***	-.231***	.1	.063***	-.231***	.1	.218***	.781***	.713***	.230***	.208***	.217***	.171***	.460***
Sen_H	.366***		.867***	1	.700***	-.295***	.037***	-.037***	.045***	1	.182***	-.045***	1	.182***	-.149***	.1	.103***	.103***	.103***	.103***	.333***
Lat_H	.449***		.710**	1	.009	-.460***	.157***	.049***	.049***	1	.021***	-.245***	1	.965***	.957***	.121***	.957***	.201***	.143***	.583***	.156***
Long_W	.858***		.693***	1	.082***	-.407***	.055***	.055***	.055***	1	.032***	-.207***	1	.965***	.965***	.218***	.965***	.223***	.296***	.592***	.170***
Short_W	.636***		.636***	1	.320***	-.021***	.021***	.021***	.021***	1	.327***	-.231***	1	.348***	.230***	.121***	.073***	.073***	.073***	.051***	.337***
Solar_R	.320***		.686***	1	.021***	-.417***	.036***	.036***	.036***	1	.019***	-.417***	1	.019***	-.258***	.01012	.007	.046***	.523***	.651***	.002
SO2			.686***	1	.021***	-.417***	.036***	.036***	.036***	1	.019***	-.417***	1	.019***	-.258***	.01012	.007	.046***	.523***	.651***	.002
NO2			.636***	1	.021***	-.441***	.049***	.049***	.049***	1	.019***	-.352***	1	.019***	-.258***	.01012	.007	.046***	.523***	.651***	.002
PM10			.636***	1	.021***	-.537***	.055***	.055***	.055***	1	.019***	-.258***	1	.019***	-.258***	.01012	.007	.046***	.523***	.651***	.002
PM2.5			.636***	1	.021***	-.671***	.055***	.055***	.055***	1	.019***	-.258***	1	.019***	-.258***	.01012	.007	.046***	.523***	.651***	.002
O3			.184***	1	.021***	-.364***	.039***	.039***	.039***	1	.019***	-.245***	1	.019***	-.245***	.01012	.007	.046***	.523***	.651***	.002
CO			.330***	1	.021***	-.309***	.039***	.039***	.039***	1	.019***	-.259***	1	.019***	-.259***	.01012	.007	.046***	.523***	.651***	.002

1 ** 表示在 0.01 水平(双侧)上显著相关, * 表示在 0.05 水平(双侧)上显著相关

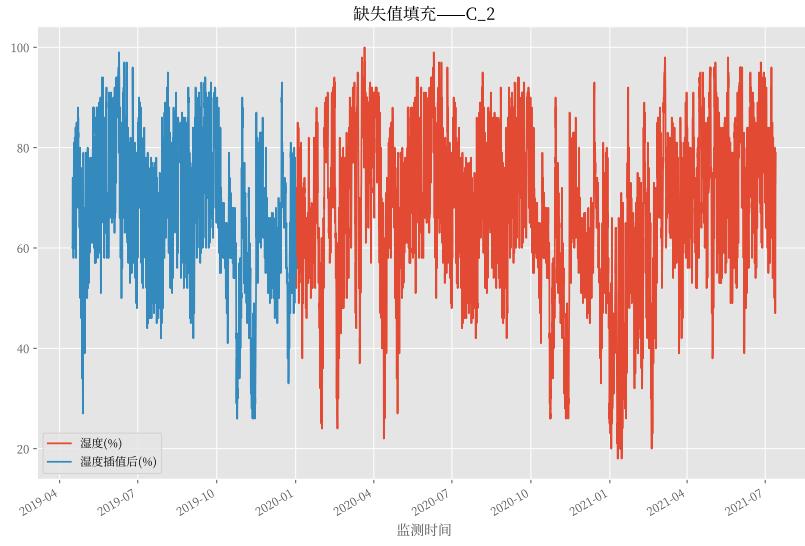


图 7: 监测点 C 湿度条件缺失值填充结果图

的情况下臭氧还可以不断累积，最终出现臭氧污染。臭氧与氮氧化合物反应过程如图所示^[7]，主要化学方程式如下：



本模型将臭氧污染形成的反应过程简化为：



从上式可以看出， NO_2 与 O_3 的比例是 3 : 1，因此在二次污染产生臭氧时，可根据 NO_2 的浓度来判定新产生的臭氧浓度值。通过与一次污染生成的臭氧浓度值线性相加，得到二次污染后新的臭氧浓度值，提升关于臭氧浓度值预测模型的鲁棒性。

6.3 模型求解和分析

在构建 GAM 模型中，构建的模型方程是：

$$k(\mu_i) = X_i\theta + h_1(x_{1i}) + h_2(x_{2i}) + \dots + h_j(x_{ji}) + \varepsilon, i = 1 \dots n \quad (6.3)$$

其中， i 为第 i 天， n 为观测的天数， j 为气象条件的个数， μ_i 为相应变量的期望值， $k(\mu_i)$ 是连接函数， h_j 的气象条件 x_{ij} 的平滑函数，代表了各污染物浓度与气象因子间的复杂关系， $X_i\theta$ 代表全参数模型成分， ε 表示残差。本文选取的平滑是惩罚 3 次回归样条，来实现各污染物浓度与气象条件之间的非线性相应。

图10-12为检测点 A、B、C 的浓度结果走势图，从图中结果可以看出，本文对于 CO 、 NO_2 、 PM_{10} 、 $PM_{2.5}$ 、 SO_2 的拟合效果比较好，曲线的走势基本符合规律，针对 O_3 的预测，在曲线走势上基本和实际数据一致，但具体的数值却小于实际的数值，是因为本文假设二次污染时只有简单的 NO_2 化学反应生成新的 O_3 ，而在实际的场景中，还可能有其他的复杂条件，使得生成了更多的二次污染物 O_3 。表5为污染物浓度及 AQI 预测结果表，从表中可以看出，对于检测点 A 来说，这三天的主要污染物为 CO ，而对于检测点 B 和 C 来说，首要污染物为 NO_2 。

根据预测模型，可以得出最终的预测结果，对于检测点 A、B、C 的污染物浓度预测结果如下表5所示。

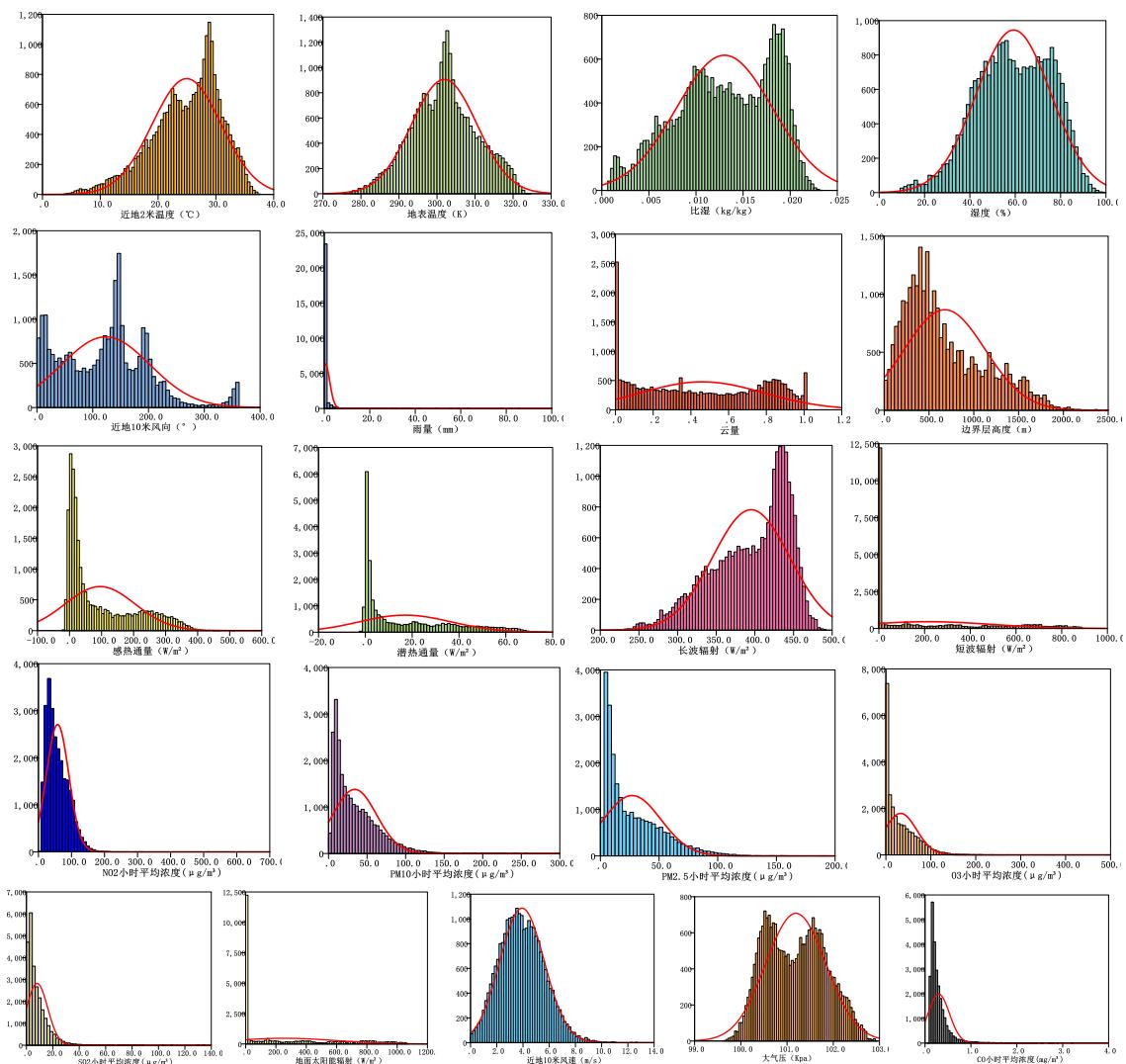


图 8: 影响因素的频率分布直方图与密度分布曲线

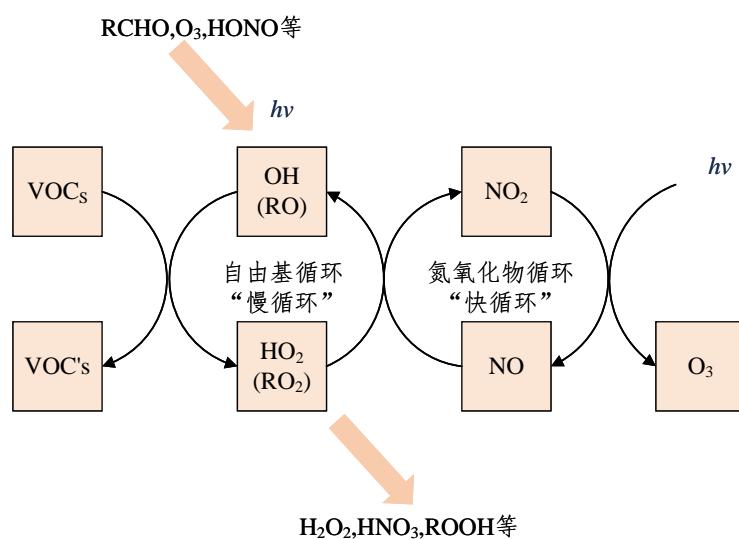


图 9: 臭氧与氮氧化物之间相互转化的反应过程

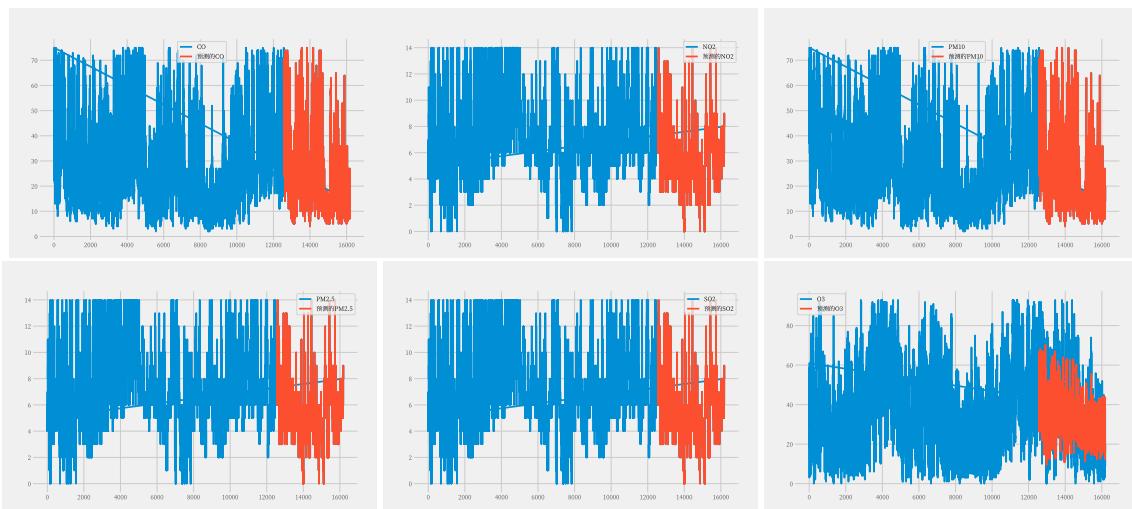


图 10: 预测 A 地各污染物浓度结果

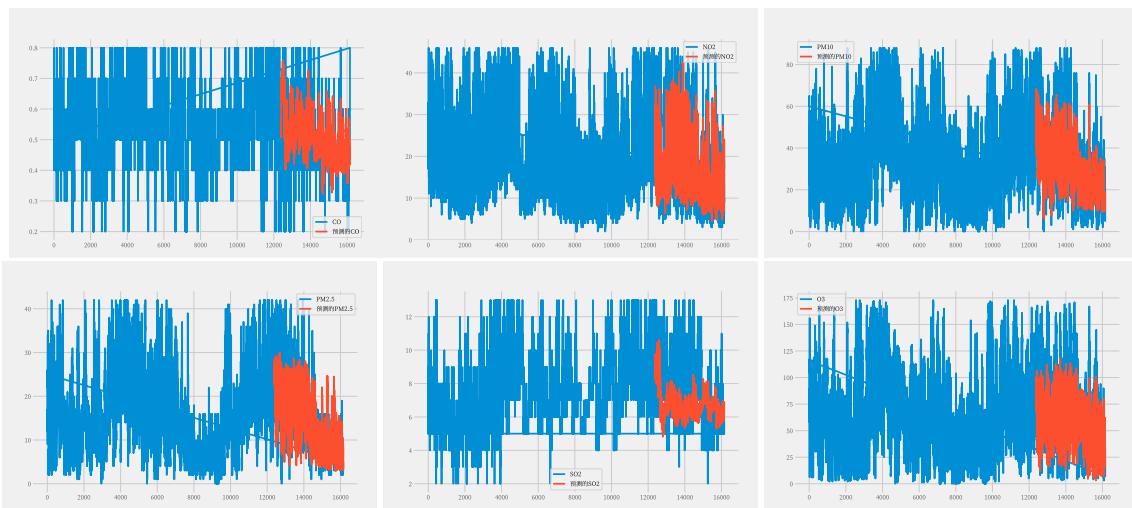


图 11: 预测 B 地各污染物浓度结果

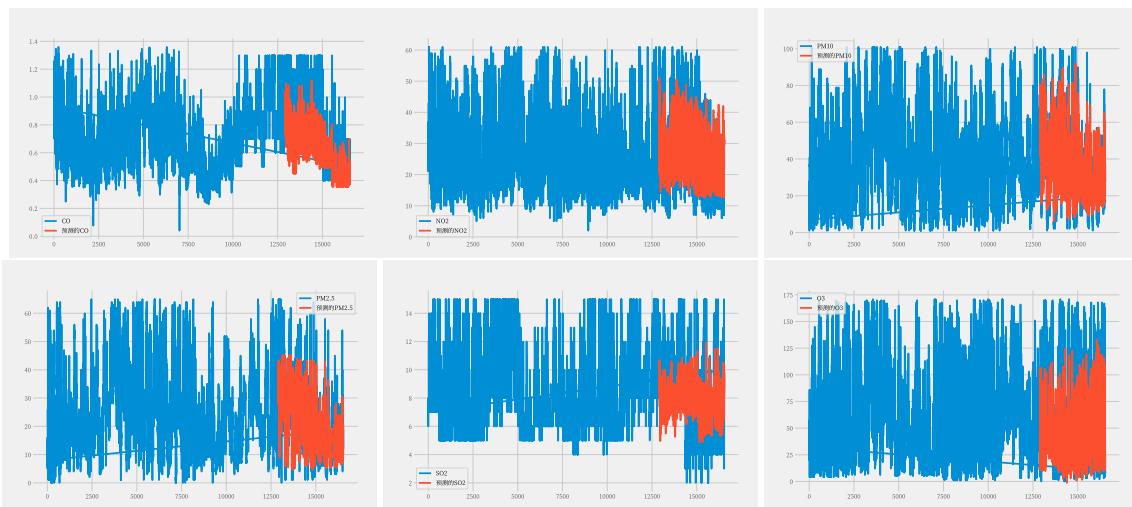


图 12: 预测 C 地各污染物浓度结果

表 5: 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							AQI	首要污染物
		SO_2 ($\mu\text{g}/\text{m}^3$)	NO_2 ($\mu\text{g}/\text{m}^3$)	PM_{10} ($\mu\text{g}/\text{m}^3$)	$PM_{2.5}$ ($\mu\text{g}/\text{m}^3$)	O_3 最大八小时滑动平均 ($\mu\text{g}/\text{m}^3$)	CO (mg/m^3)			
2021/7/13	监测点 A	5.999991894	5.999991894	29.00006866	5.999991894	25.69160652	29.00007	242	CO	
2021/7/14	监测点 A	5.999991894	5.999991894	29.00006866	5.999991894	25.69160652	29.00007	242	CO	
2021/7/15	监测点 A	5.999991894	5.999991894	29.00006866	5.999991894	25.69160652	29.00007	242	CO	
2021/7/13	监测点 B	6.735873699	15.42243481	15.10441017	7.833026409	31.08938217	0.401187	19	NO2	
2021/7/14	监测点 B	6.793331623	15.42243481	15.10441017	7.833026409	31.08938217	0.400969	19	NO2	
2021/7/15	监测点 B	6.371338367	14.48451519	15.10441017	7.833026409	31.08938217	0.400969	18	NO2	
2021/7/13	监测点 C	7.335710526	19.0898056	21.65137482	11.45984268	33.03012085	0.428493	24	NO2	
2021/7/14	监测点 C	7.335710526	19.0898056	21.65137482	11.45984268	31.84996033	0.428493	24	NO2	
2021/7/15	监测点 C	7.18458271	19.0898056	21.65137482	11.45984268	31.44129181	0.428493	24	NO2	

7. 问题四的模型建立与求解

7.1 问题四的描述分析

问题四希望我们充分考虑区域协同性，因为相邻区域的污染物往往具有一定的相关性，区域协同预报可能会提升空气质量预报的准确度。在对站点 A 进行检测的同时，也需要考虑站点 A 周围的三个检测点 A1、A2、A3 的数据，建立一个包含这四个站点的协同预报模型，并希望尽量缩小 AQI 的预测值的最大相对误差，最后将该模型与问题 3 建立的模型进行性能比较，并讨论协同预报模型的现实有效性。这是一个是时间序列预测模型，传统的时间序列预测模型，如状态空间模型 (SSMS)^[10] 和自回归 (AR) 模型，被设计成独立地拟合每个时间序列，但需要从业者在手动选择趋势、季节性和其他组件方面的专业知识。为了解决上述挑战，深度神经网络^[11] 被提出作为一种替代解决方案，其中递归神经网络 (RNN)^[12] 被用于以自回归的方式建模时间序列。然而，由于梯度消失和爆炸的问题，RNNs 很难训练出了名的^[13]。尽管出现了各种变体，包括 LSTM^[14] 和 Gru^[15]，但这些问题仍未得到解决。Transformer^[16] 作为一种全新的架构被提出，利用注意机制来处理一系列数据。与基于 RNN 的方法不同，变压器允许模型访问历史的任何部分，这使得它可能更适合获取具有长期依赖性的循环模式。但标准的 Transformer 的空间复杂度随输入长度 L 的二次增长，导致了直接对细粒度长时间序列建模的内存瓶颈。本文参考了文献 [17] 的想法，把 Transformer 架构应用于时间序列预测，

7.2 数学模型的建立

7.3 模型求解和分析

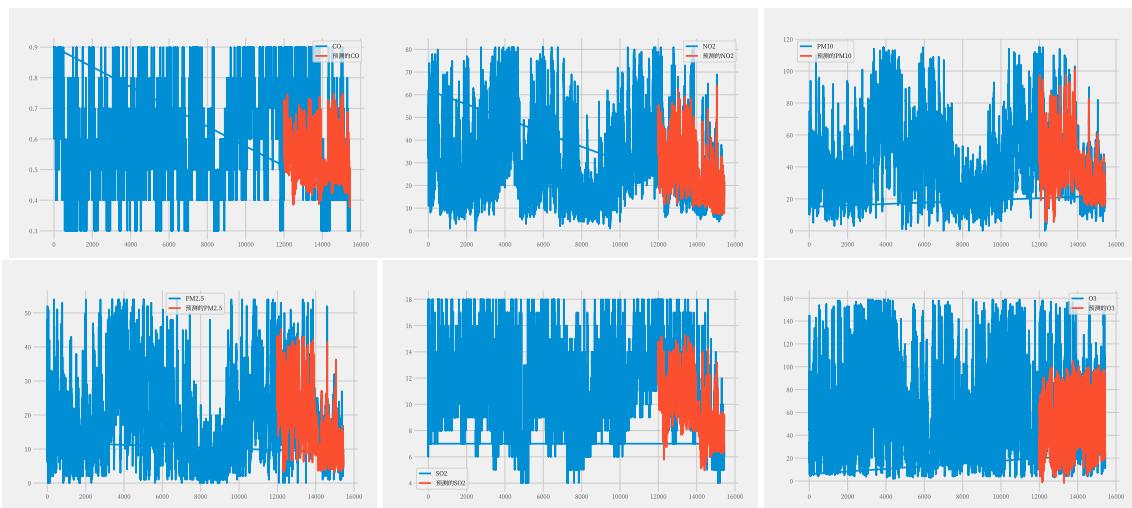


图 13: 预测 A1 地各污染物浓度结果

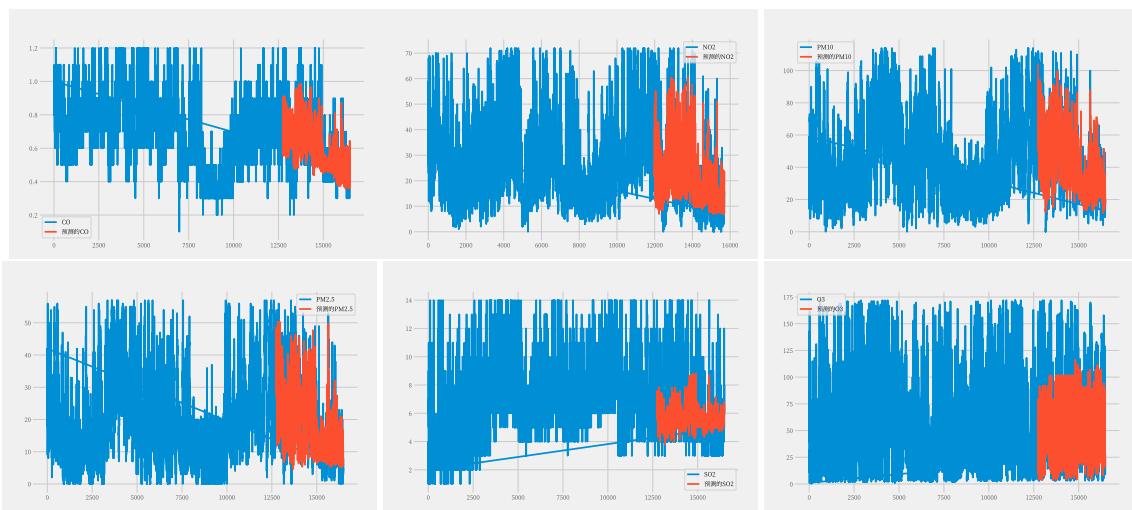


图 14: 预测 A2 地各污染物浓度结果

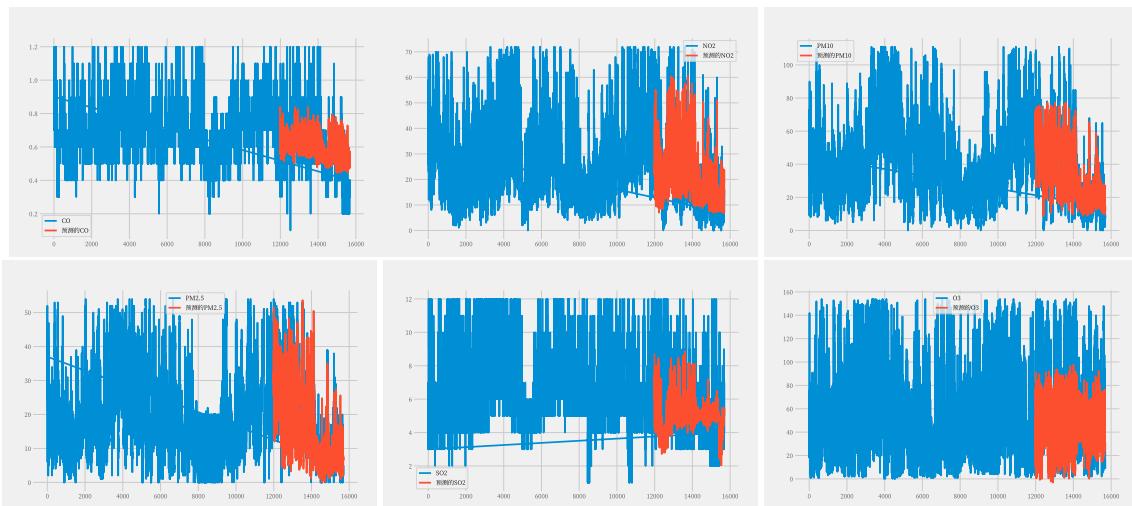


图 15: 预测 A3 地各污染物浓度结果

表 6: 监测点 A 日值预测

预报日期	地点	二次模型日值预测							首要污染物
		SO_2 ($\mu\text{g}/\text{m}^3$)	NO_2 ($\mu\text{g}/\text{m}^3$)	PM_{10} ($\mu\text{g}/\text{m}^3$)	$PM_{2.5}$ ($\mu\text{g}/\text{m}^3$)	O_3 最大八小时滑动平均 ($\mu\text{g}/\text{m}^3$)	CO (mg/m^3)	AQI	
2021/7/13	监测点 A	4.598626	19.22125	37.2189	19.35708	18.66456032	0.633464	237	O3
2021/7/14	监测点 A	4.598626	19.22125	37.2189	19.35708	18.66456032	0.651563	237	O3
2021/7/15	监测点 A	4.598626	17.5756	37.2189	19.35708	18.66456032	0.651563	237	O3
2021/7/13	监测点 A1	5.885881424	20.83291626	27.08069992	11.04520702	33.71400452	0.362359	27	PM10
2021/7/14	监测点 A1	5.885881424	23.1442337	27.08069992	11.04520702	33.71400452	0.362359	29	NO2
2021/7/15	监测点 A1	5.885881424	23.1442337	27.08069992	11.04520702	33.71400452	0.362359	29	NO2
2021/7/13	监测点 A2	7.19297123	22.57835579	31.01456833	9.943605423	29.4798317	0.41704	31	PM10
2021/7/14	监测点 A2	7.19297123	22.57835579	31.01456833	9.943605423	29.4798317	0.437403	31	PM10
2021/7/15	监测点 A2	7.19297123	22.57835579	31.01456833	9.943605423	29.4798317	0.437403	31	PM10
2021/7/13	监测点 A3	4.666842461	26.5647583	17.56330872	16.126091	32.95684814	0.509525	33	NO2
2021/7/14	监测点 A3	4.666842461	26.5647583	17.56330872	16.52332306	30.6361599	0.509525	33	NO2
2021/7/15	监测点 A3	4.666842461	26.5647583	17.56330872	16.59952545	30.6361599	0.509525	33	NO2

8. 模型评价

8.1 模型的优点

- (1) Transformer 网络能预测模型，可以解决传统神经网络模型无法处理时间序列的难题
- (2) 不需要二次测量污染物浓度数据，也能较好的实现二次天气质量预报

8.2 模型的缺点

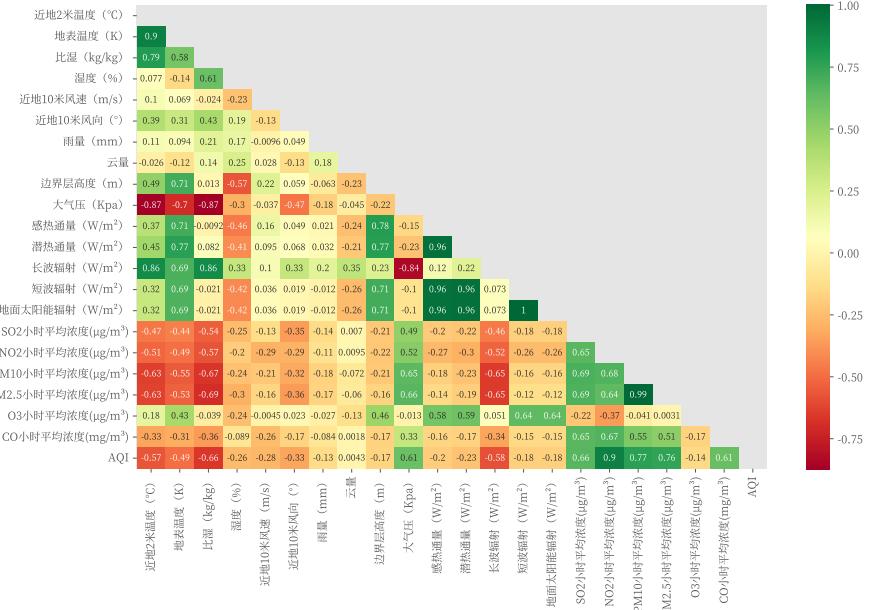
由于时间关系，问题 3 和 4 的可视化比较没有来得及进行试验。

参考文献

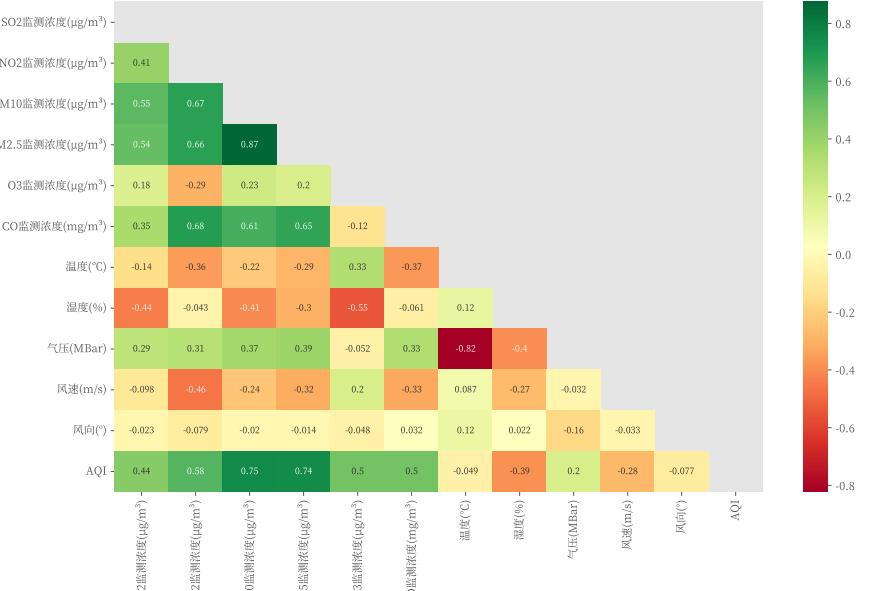
- [1] 郝吉明, 马广大, 王书肖. 大气污染控制工程 [M]. 北京: 高等教育出版社, 2010.
- [2] Rubin, Donald B. Inference and missing data. ETS Research Bulletin Series, 1975.
- [3] Mei, Gang, Nengxiong Xu, and Liangliang Xu. Improving GPU-accelerated adaptive IDW interpolation algorithm using fast kNN search. Springerplus 5.1 (2016): 1-22.
- [4] Shukla, Satya Narayan, and Benjamin M. Marlin. Interpolation-prediction networks for irregularly sampled time series. arXiv preprint arXiv:1909.07782 (2019).
- [5] Solanki, H. U., Dhyey Bhatpuria, and Prakash Chauhan. "Applications of generalized additive model (GAM) to satellite-derived variables and fishery data for prediction of fishery resources distributions in the Arabian Sea." Geocarto international 32.1 (2017): 30-43.
- [6] 贺祥, 林振山. 基于 GAM 模型分析影响因素交互作用对 $PM_{2.5}$ 浓度变化的影响 [J]. 环境科学, 2017, 38(01):22-32.
- [7] Li, Shiyang, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhua Chen, Yu-Xiang Wang, and Xifeng Yan. Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting. ArXiv:1907.00235 [Cs, Stat], January 3, 2020. <http://arxiv.org/abs/1907.00235>.
- [8] 刘昕, 辛存林. 陕甘宁地区城市空气质量特征及影响因素分析. 环境科学研究, 2019, 32(12): 2065-2074.
- [9] 陈敏东. 大气臭氧污染形成机制及研究进展 [J/OL] 2018, <https://max.book118.com/html/2018/0201/151478594.shtml>.
- [10] James Durbin and Siem Jan Koopman. Time series analysis by state space methods. Oxford university press, 2012.
- [11] Valentin Flunkert, David Salinas, and Jan Gasthaus. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. arXiv preprint arXiv:1704.04110, 2017.
- [12] Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long-and short-term temporal patterns with deep neural networks. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pages 95–104. ACM, 2018.
- [13] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In International conference on machine learning, pages 1310–1318, 2013.
- [14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.
- [15] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259, 2014.
- [16] Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. arXiv preprint arXiv:1606.01933, 2016.
- [17] Li, Shiyang, et al. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. Advances in Neural Information Processing Systems 32 (2019): 5243-5253.

附录 A 主程序源代码

```
# 问题1
print("HelloWorld!")
# 问题2
print("Open source is awesome!")
# 问题3
print("Open by NBU Professional Team")
# 问题4
```



(a) pic1.



(b) pic2.

