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	西安邮电大学
参赛队号	21116640067
队员姓名	1.艾宇
	2.胥策
	3.杨玉蓉

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 基于 LSTM-FC 的大气污染物浓度预测模型

摘 要：

工业发展和城市化进程的加速使得大气污染问题日益突出，严重影响到人们的生活和健康，违背了经济可持续发展的理念，尤其是近年来许多大城市雾霾天气的频繁出现，引发了广泛的社会效应。随着我国步入“绿水青山就是金山银山”的“新时代”，大气污染治理已经势在必行。科学阐明大气污染的根源，探究大气污染物浓度变化的规律及气象条件对浓度的影响，实现污染物浓度精确预测，是国际大气化学研究领域最关注的科学问题，是制定区域大气污染协同控制方案的基础，对指导大气治理工作具有重要意义。

国内外学者展开了大量关于大气污染物浓度预测的研究，但现有大多大气污染物浓度预测模型忽视了空气质量监测数据的空间特性，不能很好地实现时间与空间相关性的耦合，造成预测精度不高等问题。针对上述问题，本文提出了一种基于长短期记忆网络-全连接神经网络 (Long Short-Term Memory - Full Connested, LSTM-FC) 的大气污染物混合预测模型。本文利用附件提供的多个空气质量监测点的实时监测数据、一次预测数据、以及气象条件等，通过大量的数据分析及处理，建立气象条件变化与大气污染物浓度变化的数学关系，并预测大气污染物浓度变化的趋势。结合 Pearson 相关性分析、k-means 聚类、神经网络、权重预测等相关数学方法和现代信息处理技术，通过 Python、MATLB、SPSS、MySQL 等工具辅助解决相关问题。

针对问题一，对附件一提供的大气污染物实测数据进行数据清洗，包括数据缺失检测及近邻填补、箱型法异常值检测、数据可视化等。其次，根据所给的各项污染物空气质量分指数（IAQI）计算模型计算当日空气质量指数（AQI）及首要污染物，并绘制 2020 年全年 AQI 波动曲线并统计首要污染物占比情况。根据计算结果，一年中臭氧为首要污染物的天数占比为 59.6%，说明臭氧对大气环境的危害程度较大。最后，展示了监测点 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。

针对问题二，对附件一中逐小时预报及实测数据进行预处理，并对污染物浓度及气象因子做相关性分析，为消除不同污染物浓度量纲对分析结果的影响，对数据进行标准化处理。为分析气象因子对污染物浓度的影响，建立不同气象因子与污染物浓度的多元线性回归模型，但经过大量数据训练和分析，最终选择相关性系数进行分析。其次，建立基于 k-means 聚类的气象分类模型，以手肘法作为聚类数量的判定条件，将气象因子划分为五类，分析每类气象条件的特征及其对空气质量的影响，最终建立完整的气象条件的分类模型。

针对问题三，为了建立适用于多个监测点的污染物浓度预报模型，首先利用 MySQL 对附件一和二中 A、B、C 三个监测点的数据进行整合，在数据预处理时选择适当的插值方法保证数据时间上的连续性，降低缺失值对模型的负面影响。将处理后的数据拆分为训练接、验证集、和测试集，污染物浓度的的实测值作为标签建立对应关系。其次，考虑到气象条件与氮氧化物对臭氧浓度的影响，分别建立了基于 LSTM-FC 网络的一次大气污染物浓度预测模型和基于 LSTM-FC 网络的臭氧浓度预测模型，并进行模型评价。根据模型训练结果，该模型对 NO_2，PM_{10}，CO 等污染物的预测准确率较高，对 SO_2 的预测效果略差。相比于 WRF-CMAQ 预测模型，一次污染物的预测模型 *MSE* 指标下降了平均 5.4%，R^2 指标提高了 4.6%；对于臭氧浓度预测模型，*MSE* 指标下降了 12.4%，R^2 指标提高了 9.3%，说明提出的模型有效改进了浓度预测精度。最终，展示了监测点 A、B、C 自 2021 年 7 月 13 日至 2021 年 7 月 15 日的各污染物的浓度、AQI 及首要污染物种类。

针对问题四，对附件一和三中的数据进行预处理和整合。其次，在问题三 LSTM-FC 网络的基础上考虑不同区域间数据的相关性，建立了基于时空加权融合的大气污染物预测模型 (STW-LSTM-FC)。通过加入图卷积神经网络 GCN 提取多个监测点的空气质量特征，将自身数据和邻近区域气象数据的隐藏特征聚合成新的特征结点，通过区域协同预报提高了空气质量预报的准确度。最终，展示了监测点 A、A1、A2、A3 自 2021 年 7 月 13 日至 2021 年 7 月 15 日的各污染物的浓度、AQI 及首要污染物种类。

关键词：空气质量指数 AQI，大气污染物浓度预测模型，相关性分析，k-means 聚类，LSTM-FC 网络模型，STW-LSTM-FC 网络模型，区域协同预报模型，评价指标

目录

一、 问题重述

在大气污染问题日益突出的背景下，空气质量监测和预报成为工作污染防治工作的的重要环节。据研究，我国是大气污染较为严重的国家之一，许多城市长期处于大气污染物浓度超标的状态，对我国经济发展和人口健康造成了巨大的影响。建立空气质量预测模型，提前预测到可能发生的污染就可以采取一定的措施来预防，这是减少环境污染对人体的危害以及保护生态环境的行之有效的方法。近几年，我国政府也采取了许多相关措施来进行环境监测与防治，在许多城市都建立了空气质量监测站，达到了较好的效果，但污染防治仍然还有很长的路要走。要想实现人与自然的可持续发展，国际和国内社会都应加大力度采取更多有效措施来对空气污染问题进行综合的治理与管控，有针对性地进行治理，尽可能地改善空气质量，这是人们的迫切愿望，也是当地政府的科学发展目标。

为了准确客观的对环境进行把控，常用 WRF-CMAQ 空气预报模型对空气质量进行预报，但人们对空气质量预测的要求不断增加。在实际自然环境中气象场和排放清单具有不确定性，WRF-CMAQ 的预测结果存在一定的不准确性，急需一种更完善的预判模型提高预报的准确性，从而更好地进行大气治理。在大气中的污染物按照其来源主要包括两个方面，一方面是由人类活动或者自然界直接排放的污染物，即一次污染物（SO2、NO2、PM10、PM2.5、CO）；另一方面是由大气中已经存在的其他污染物在特定气象条件下经过化学反应生成的新污染物，即二次污染物，本问题只考虑臭氧 O3 这一种二次污染物。对于上述六中污染物，尤其是臭氧的浓度预测是进行大气治理的关键问题。

基于上述的研究背景，本文给出了部分监测点的预报基础数据，以及时间、污染物浓度监测及预报数据、气象因子及变化情况等相关数据，本文需要解决的问题如下：

1、对于问题一，依据给出的监测点 A 的实测数据，按照问题给出的 AQI 计算方法，计算每天的 AQI 及首要污染物。

2、对于问题二，气象条件对各种污染物的扩散和沉降有很大的影响，依据给出的监测点 A 处的污染物实测数据与气象条件记录，对气象条件进行分类，并通过数学方法总结归纳各类气象条件的特征。

3、对于问题三，建立适用于 A、B、C 三个监测点的二次预报模型，提高预测的精确度。由于臭氧属于二次污染物，故其浓度预测较为复杂，需要考虑气象因子的影响等因素，故对一次污染物和二次污染物分别建立预测模型，并对模型进行评价，使得 AQI 预报值的相对误差尽量小，首要污染物预测准确度尽量高。

4、对于问题四，由于空气质量数据间存在时空相关性，在预测过程中若仅考虑时间相关性忽略对空间特征的建模分析，会降低模型预测性能。结合临近区域 A、A1、A2、A3 的污染物观测数据，建立区域协同预测模型，进一步提高大气污染物预测的准确度。模型预测单一污染物的单日浓度值变化趋势。

二、 模型假设及符号说明

2.1 模型假设

为了使问题易于理解，我们做出如下假设：

（1）假设所有附件数据来源于实际应用中，可以反映普遍的空气质量检测情况。

（2）假设数据监测地点具有代表性，不存在特殊性或极端地域条件，且附件中数据真实可靠。

（3）假设人类活动、城市效应等对大气污染物浓度的影响可以忽略不计。

2.2 符号说明

为了增加论文的可读性，如表 2.1 所示给出了本文模型建立过程中用到的符号及其说明。

表 2. 1 符号说明

变量名称	变量说明
$IAQI_P$	污染物的空气质量分指数
AQI	空气质量指数
ρ	总体相关系数
r	皮尔森相关系数
max	气象因子样本数据的最大值
min	气象因子样本数据的最小值
β	偏回归系数
T	温度
H	湿度
A	气压
W	风速
e	回归残差
k	聚类簇的数量
Y	神经网络的输出
X	神经网络输入矩阵
b	神经网络偏置
MSE	均方误差
R^2	拟合优度

三、 问题一模型的建立与求解

3.1 问题分析及建模思路

问题一要求根据附件一数据计算 AQI 及首要污染物。首先需要分析附件一数据，主要包含 2019/4/16 至 2021/7/13 监测点 A 的逐小时污染物浓度与气象一次预报数据，监测点 A 逐小时与逐点污染物浓度与气象实测数据，分别给出了 SO_2，NO_2，CO，O_3，PM_{10}，$PM_{2.5}$ 的监测浓度及温湿度、气压、风速、风向等实时情况。

其次，对数据进行预处理，包括脏数据检测、缺失值删除、填补、及异常值检测，增强后续模型的稳定性和可信度。通过提取各污染物的质量浓度值、与污染物的质量浓度值相近的污染物浓度限值的高位值与低位值等数据信息建立污染物的空气质量分指数（IAQI）模型，并根据各污染物项目的分指数计算最终的 AQI 及首要污染物。如图 3.1 所示为 AQI 模型建立流程图。

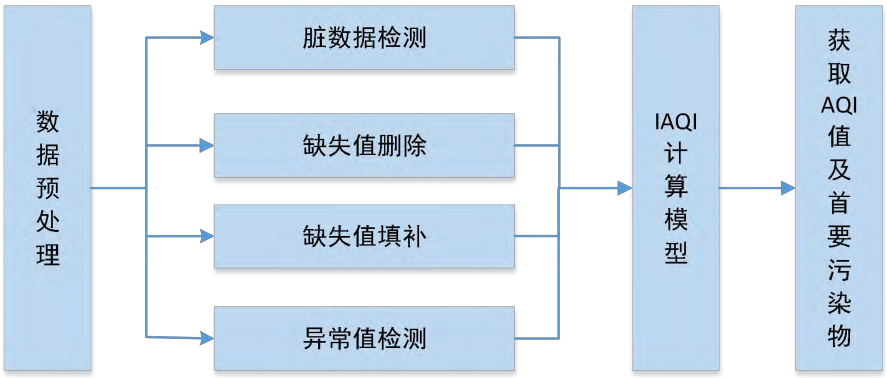


图 3. 1 AQI 计算模型流程图

3.2 据分析及预处理

本问题主要使用附件越高一“监测点 A 空气质量预报基础数据”计算 AQI 及首要污染物，故先对附件一中的数据进行整体预处理，方便后续使用。首先，对数据进行初步清洗，去除缺失数据及无效值。本文采用删除缺失数据行、填补缺失数据两种方式来处理缺失数据。对于完全缺失的数据直接进行删除，对于有部分缺失的数据，通过填补补全。

（1）脏数据检测

AQI 的计算主要是通过各项污染物的检测浓度来计算，即附件一的 sheet3 中污染物浓度实测数据。原始数据数据共有 823 项，共检测出 23 个脏数据。其中，表 3.1 所示为检测出的 3 个完全缺失数据，表 3.2 所示为部分缺失数据。

表 3. 1 完全缺失值检测结果

监测日期	地点	SO2 监测浓度(μg/m³)	NO2 监测浓度(μg/m³)	PM10 监测浓度(μg/m³)	PM2.5 监测浓度(μg/m³)	O3 最大八小时滑动平均浓度(μg/m³)	CO 监测浓度(mg/m³)
2019/12/8	监测点 A	NA	NA	NA	NA	NA	NA

2020/8/2	监测点 A	NA	NA	NA	NA	NA	NA
2020/11/21	监测点 A	NA	NA	NA	NA	NA	NA

表 3. 2 部分缺失检测结果

监测日期	地点	SO2 监测浓度(μg/m³)	NO2 监测浓度(μg/m³)	PM10 监测浓度(μg/m³)	PM2.5 监测浓度(μg/m³)	O3 最大八小时滑动平均监测浓度(μg/m³)	CO 监测浓度(mg/m³)
2019/4/24	监测点 A	5	20	NA	16	85	0.6
2019/5/24	监测点 A	3	32	NA	NA	88	0.7
2019/6/18	监测点 A	5	29	19	7	61	NA
2019/7/17	监测点 A	9	30	NA	36	236	0.8
2019/7/30	监测点 A	4	25	NA	7	77	0.8
2019/8/27	监测点 A	4	26	NA	9	56	0.6
2019/10/10	监测点 A	7	26	NA	32	170	0.9
2019/10/26	监测点 A	8	37	NA	44	102	0.9
2020/1/5	监测点 A	10	47	52	36	80	NA
2020/1/15	监测点 A	10	42	76	30	93	NA

（2）填补并整合数据

对于数据的缺失填充，经大量文献参考，主要包括均值填充、中值填充、众数填充、及近邻填充等方法[1]，通过对比分析，并结合本问题的数据特点，最终选择插值法来填补数据。该方法是通过搜索缺失数据周围的样本点，搜索最近的样本点并将其作为缺失数据的填补值，该方法计算量小、处理起来高效灵活、思路简单。经过处理以后的数据几乎不影响数据间的相关性，而且较好的保证了实验数据的完整性。补足并整合后的部分数据如下表 3.3 所示显示了前十个数据的最终补全结果。

表 3. 3 插值补全结果

监测日期	地点	SO2 监测浓度(μg/m³)	NO2 监测浓度(μg/m³)	PM10 监测浓度(μg/m³)	PM2.5 监测浓度(μg/m³)	O3 最大八小时滑动平均监测浓度(μg/m³)	CO 监测浓度(mg/m³)
2019/4/24	监测点 A	5	20	36	16	85	0.6
2019/5/24	监测点 A	3	32	33	22	88	0.7
2019/6/18	监测点 A	5	29	19	7	61	0.7

2019/7/17	监测点 A	9	30	47	36	236	0.8
2019/7/30	监测点 A	4	25	22	7	77	0.8
2019/8/27	监测点 A	4	26	31	9	56	0.6
2019/10/10	监测点 A	7	26	52	32	170	0.9
2019/10/26	监测点 A	8	37	70	44	102	0.9
2020/1/5	监测点 A	10	47	52	36	80	0.8
2020/1/15	监测点 A	10	42	76	30	93	0.9

（3）剔除异常个案数据

偏离数据总体范围的异常数据会极大地干扰我们的分析过程，影响模型的稳定性，导致最终得到的分析结果存在很大的不真实性。通常异常值的检测方法包括业务法、3σ原则、和箱型图，及曲线波动检测等[2]，经过对四种方法的对比分析，选择箱型图、和曲线波动检测来标识异常值。

箱型图可以并列显示多列数据，并显示各列数据的中位数、尾长、主要的分布区间，以及异常值。如图 3.2 为六项污染物自 2019/4/16 至 2021/7/13 期间的箱型图。图 3.3 为动态变化曲线图。根据曲线图，SO_2，NO_2，CO，O_3，PM_{10} 的监测范围均在合理范围内上下波动，但根据 $PM_{2.5}$ 的波动曲线，存在突变值，如表 3.4 所示为剔除的个例数据。

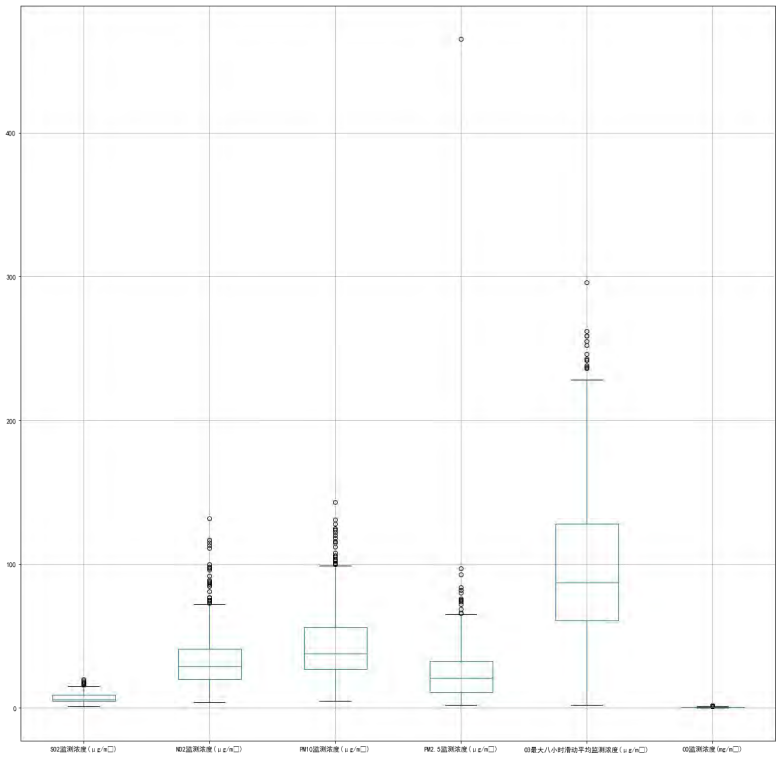
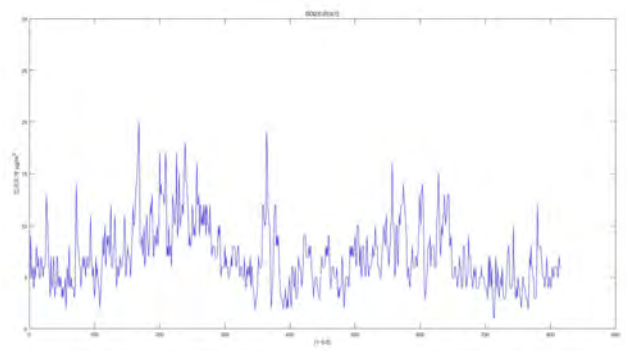
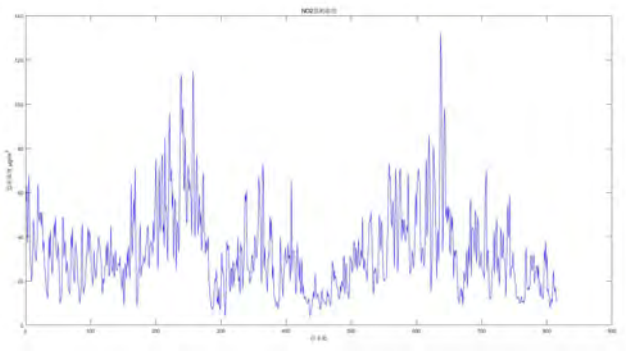


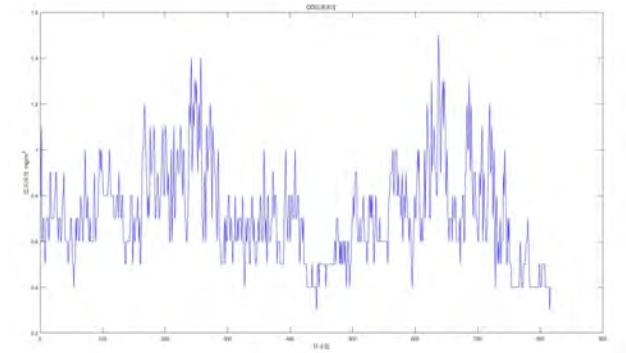
图 3.2 六项污染物数据箱型图可视化



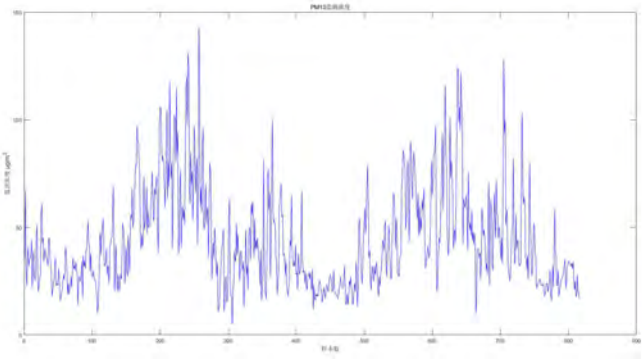
（a）SO_2 逐日动态变化曲线



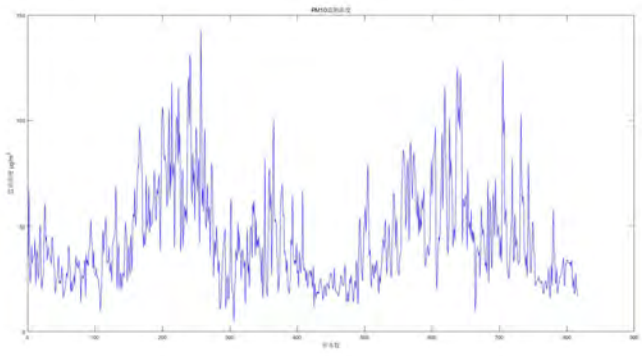
（b）NO_2 逐日动态变化曲线



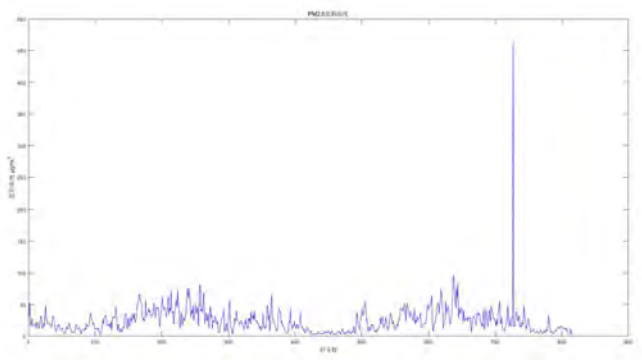
（c）CO 逐日动态变化曲线



（d）PM_{10} 逐日动态变化曲线



（e）PM_{10}逐日动态变化曲线



（f）$PM_{2.5}$逐日动态变化曲线

图 3. 3 六项污染物浓度逐日变化曲线

表 3. 4 异常数据剔除

监测日期	地点	SO2 监测浓度(μg/m³)	NO2 监测浓度(μg/m³)	PM10 监测浓度(μg/m³)	PM2.5 监测浓度(μg/m³)	O3 最大八小时滑动平均监测浓度(μg/m³)	CO 监测浓度(mg/m³)
2021/4/14	监测点 A	4	18	34	465	83	0.7

3.3 模型建立

空气质量指数（AQI），是通过定量来描述空气质量状况的一种无量纲指数，其取值范围一般在在 0-500 之间，其数值越大，说明污染程度就越高。通常，空气质量的优劣，主要是通过对空气中不同的污染物的浓度进行实时检测获得的，用于空气质量状况评价的污染物指标主要有：一氧化碳（CO）、二氧化硫（SO_2）、二氧化氮（NO_2）、臭氧最大 8 小时滑动平均（O3_8h）、粒径小于等于 10 μm 颗粒物（PM_{10}）、粒径小于等于 2.5 μm 颗粒物（$PM_{2.5}$）等[3]。根据最新的《环境空气质量指数（AQI）技术规定（试行）》（HJ633-2012），建立 AQI 评价模型。首先，建立各项污染物的空气质量分指数计算模型：

$$IAQI_P = \frac{IAQI_{Hi} - IAQI_{Lo}}{BP_{Hi} - BP_{Lo}} \cdot (C_P - BP_{Lo}) + IAQI_{Lo} \qquad (3\text{-}1)$$

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} 表示与 C_P 值相近的污染物浓度限值的高位值；BP_{Lo} 表示与 C_P 值相近的污染物浓度限值的低位值；$IAQI_{Hi}$ 表示在指数表中与 BP_{Hi} 相对应的空气质量分指数；$IAQI_{Lo}$ 表示在指数表中与 BP_{Lo} 相对应的空气质量分指数。

根据附录表 1 中所给的空气质量分指数及对应的污染物项目浓度限值，以及附件一中的污染物监测数据，可以得出各个污染项的空气质量分指数 $IAQI$ 的相应数值。再从各类污染物的 IAQI 值中选取最大的一个数值，即：

$$AQI = max\{IAQI_1, IAQI_2, IAQI_3, IAQI_4, IAQI_5, IAQI_6\} \quad (3\text{-}2)$$

作为 AQI 的最终值。其中，贡献了最大数值 IAQI 的污染物，被确定为当天的首要污染物。

3.4 模型求解结果

根据 3.3 节中的 AQI 计算模型，计算出附件一 sheet3 中逐日的 AQI 值及对应的每日首要污染物，具体结果查看支撑材料中 “问题一 AQI 计算及首要污染物计算结果”。其中，2020/1/1 至 2020/12/31 一整年的 AQI 变化曲线如图 3.4 所示。根据曲线图可知，在 2020 年的四五月份及八九月份，AQI 的值较高。如图 3.5 所示是一年内首要污染物的占比情况[4]，其中臭氧是首要污染物占比为 59.6%，占一年中一大半的时间，是大气污染的罪魁祸首，此外，NO_2 占比为 28.2%，PM_{10} 占比 8.2%，$PM_{2.5}$ 占比 3.2%，而 CO 仅占 0.5%，SO_2 占比为 0。

如表 3.6 所示，是从最终计算结果中提取出的监测点 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。其中，25、27、28 日的主要污染物均为 O_3，由于 26 日 AQI 值小于 50，故没有首要污染物。根据如表 3.7 所示的空气质量等级及对应 AQI 范围，这四天的空气质量等级如表 3.8 所示。

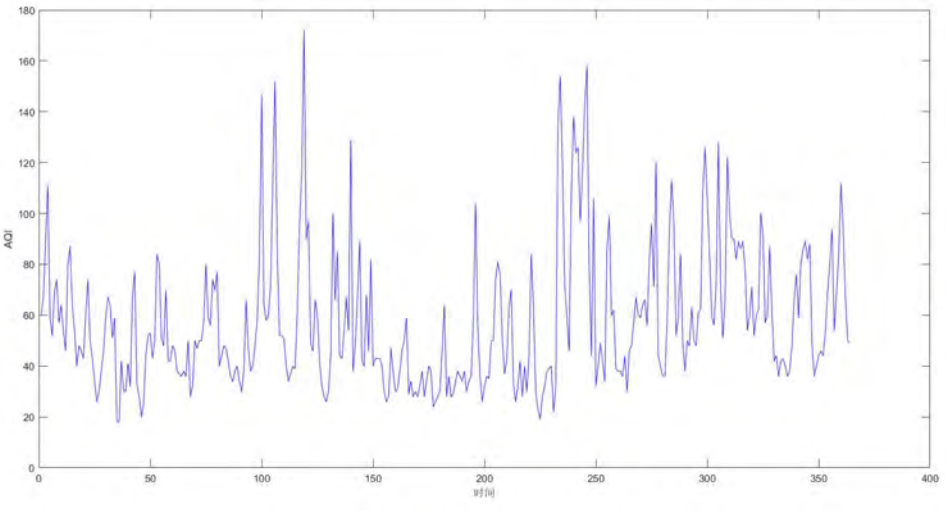


图 3. 4 2020 年 AQI 变化曲线

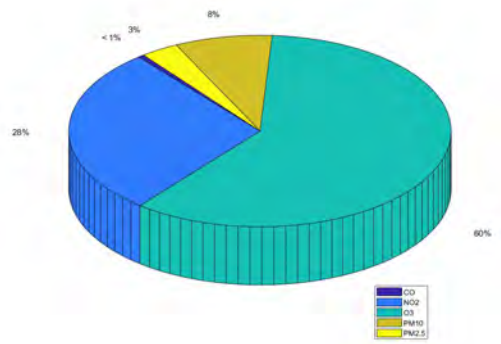


图 3. 5 一年内首要污染物的占比情况

表 3. 5 最终 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	109	O_3
2020/8/28	监测点 A	138	O_3

表 3. 6 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,+∞)

表 3. 7 8 月 25 日到 8 月 28 日空气质量等级

监测日期	地点	空气质量等级
2020/8/25	监测点 A	良
2020/8/26	监测点 A	优
2020/8/27	监测点 A	轻度污染
2020/8/28	监测点 A	轻度污染

根据最终的 AQI 结果及首要污染物，发现臭氧污染频繁成为首要污染物，是环境污染的最大杀手。原因在于，在日照较较强时，尤其在春夏阶段，空气中大量的的氮氧化物和挥发性有机化合物（VOCs）经紫外线照射会发生光化学反应，生成臭氧。根据图 3.3 的波动曲线，也容易知道，在光照充足的季节，臭氧浓度较高，所以 AQI 值较大。故在 8 月 25 日到 8 月 28 日光照较为充足的夏季，会产生大量的臭氧。

四、 问题二模型的建立与求解

4.1 问题分析及建模思路

大气污染物的浓度与气象条件有着密切的关系，在检测大气污染的同时还需测定温湿度、气压、风速、风向等气象参数的影响。本节需要根据附件一数据对气象条件进行分类，并分析气象条件对污染物的影响程度。附件一中的“监测点 A 逐小时污染物浓度与气象实测数据”包含不同污染物的检测浓度及多个气象情况，故使用该数据

进行气象分类。为了便于分析，本节提取具有代表性的 2020/1/1 至 2020/12/31 一整年的数据进行分析。

首先，对数据进行清洗，去除缺失值或进行缺失数据填充。其次，对不同气象因子与污染物浓度之间的相关性进行初步分析，可以知道每个气象因子对不同污染物浓度的整体的抑制和促进作用。其次，不同气象因子的量纲完全不同，如湿度单位为百分比，而风向的单位是度，且数据间的数值差异较大，故应对数据进行标准化处理，通过去量纲减小数据差异的影响。

不同气象因子对不同污染物的作用及作用程度不同，某种污染物浓度可能受一种气象因子或多种气象因子的综合影响，故针对每一个污染物浓度，建立气象因子与污染物浓度的多元线性回归模型并进行回归检测。

为了对气象条件分类[5]，对每种气象条件进行初步的等级划分，根据划分等级进行 k-means 聚类[6]，最终，根据聚类中心的值即可对气象条件进行分类。

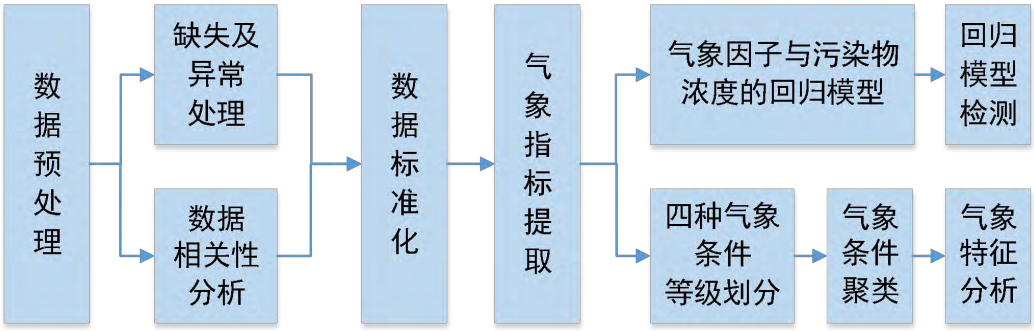


图 4. 1 建立气象条件分类模型流程图

4.2 数据分析及预处理

4. 2. 1 数据预处理

（1）缺失检测及处理

根据 3.2 节数据预处理过程处理附件一中 sheet2 的数据。首先进行数据缺失检测，存在大量无效数据，对于完全缺失的数据直接删除处理，对于有一项缺失的数据进行最近邻插值。原始数据由 19432 条记录组成，共检测出 407 条缺失记录，其中有 44 条完全缺失的数据。

（2）异常值检测及处理

对数据进行缺失检测后，进行异常检测，发现部分污染物浓度的检测结果为负值，可能是由于设备调试或偶然因素造成，为了减小异常值对数据的影响，直接删除含有异常数据的行记录。共检测出 268 个异常值。如表 4.1 为检测出的部分异常数据。

表 4. 1 检测出的部分异常数据

监测时间	SO2 监测浓度(μg/m³)	NO2 监测浓度(μg/m³)	PM10 监测浓度(μg/m³)	PM2.5 监测浓度(μg/m³)	O3 监测浓度(μg/m³)	CO 监测浓度(mg/m³)
2019/4/20 2:00	5	78	52	38	-1	0.8
2019/4/21 2:00	4	67	32	24	-1	0.7
2019/4/21 3:00	4	74	25	21	-1	0.7
2019/4/21 22:00	3	52	31	18	-1	0.7
2019/4/21 23:00	2	45	33	22	-1	0.5

2019/4/22 0:00	1	44	17	22	-1	0.6
2019/4/22 1:00	2	41	29	18	-1	0.6
2019/4/22 2:00	3	40	32	22	-1	0.5

4.2.2 污染物浓度与气象因子的相关性

认识和掌握不同气象因子的变化对污染物的影响，可以在大气污染防治过程中通过气象条件减少大气污染造成的经济损失及社会危害。对不同气象因子与污染物浓度之间的相关性进行初步分析，可以知道每个气象因子对不同污染物浓度的整体的抑制和促进作用。本节采用皮尔森相关系数 (Pearson correlation coefficient) 计算气象因子与污染物浓度之间的相关性。

皮尔森相关系数是最常用的相关系数之一，是一种线性相关系数。将系数记为 r，用来反映两个变量 X 和 Y 的线性相关程度。如表 4.2 所示，r 值介于-1 到 1 之间，绝对值越大表明相关性越强，r 约接近于 0，相关性越弱。若系数为负值，则呈负相关，系数为正，呈正相关。将两个变量 X 和 Y 之间的协方差和标准差乘积的比值记为总体相关系数ρ，其计算方式如下：

$$\rho_{X,Y}=\frac{cov(X,Y)}{\sigma_X\sigma_Y}=\frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y} \tag{4-1}$$

估计每个污染物浓度及气象因子之间的协方差和标准差，即可得到样本的皮尔森相关系数 r：

$$r=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \tag{4-2}$$

表 4.2 所示为计算的最终相关性系数，图 4.2 为系数热力图。根据相关系数，温度及风速对 SO_2，NO_2，CO，PM_{10}，$PM_{2.5}$ 的浓度有抑制作用，即温度越高风速越大，这些污染物的浓度越低；温度及风速对 O_3 的浓度则有促进作用，即温度越高风速越大，臭氧的浓度越高，说明高温环境有利于臭氧的生成。同理，湿度对所有的污染物浓度都有抑制作用，湿度越高，污染物浓度越低，但湿度与二氧化氮的相关性系数为−0.054，表明湿度对二氧化氮的浓度影响较小。而气压对污染物浓度的影响与温度刚好相反，气压越大，O_3 的浓度会略低，但是系数为−0.049，影响较为微弱。而风向与各项污染物的相关系数都较小，说明对污染物浓度的影响十分微弱。此外，根据各气象因子与污染物浓度的相关系数，发现 AQI 主要受湿度和气压条件的影响。

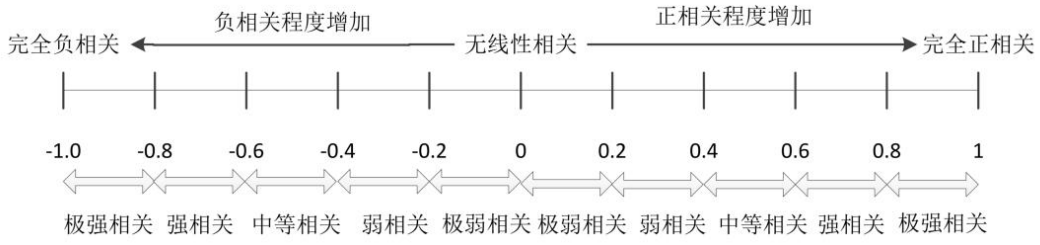


图 4.2 皮尔森相关系数

表 4.2 气象因子与污染物浓度及 AQI 的相关性系数

	SO_2	NO_2	PM_{10}	$PM_{2.5}$	O_3	CO	AQI
温度	-0.137	-0.358	-0.218	-0.286	0.325	-0.282	0.04
湿度	-0.447	-0.054	-0.42	-0.307	-0.555	-0.024	-0.54
气压	0.287	0.308	0.372	0.391	-0.049	0.232	0.15
风速	-0.137	-0.437	-0.233	-0.306	0.184	-0.199	-0.03
风向	-0.099	-0.079	-0.019	-0.014	-0.050	-0.065	-0.03

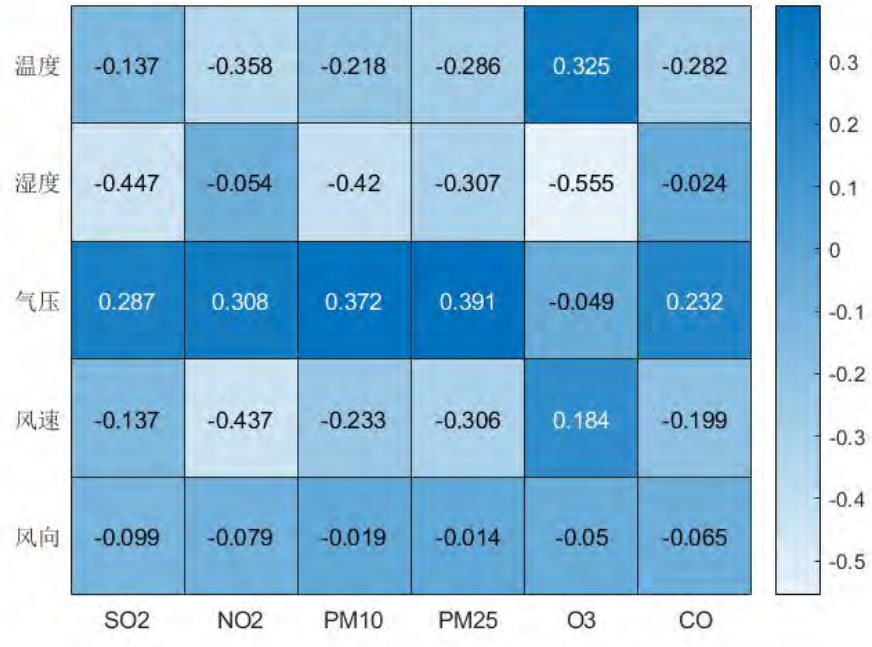


图 4.3 污染物浓度与气象因子的相关性系数热力图

4.2.3 数据标准化处理

在数据分析中，数据来源会导致数据的量纲、数据的量级产生巨大的差异，为了让这些数据具备可用性和可比性，需要对数据进行标准化来消除这些差异。数据的标准化 (Normalization) 就是将原始各指标数据按比例缩放，去除不同属性数据的单位限制，将其转化为无量纲的纯数值，这样，不同单位或量级的指标就能够进行加权或者比较。

数据标准化方法多种多样，但不同的标准化方法，对数据的建模结果会产生不同的影响。常见的标准化的方法有：min-max 标准化（min-max normalization）、log 函数转换、z-score 标准化（zero-mena normalization）、atan 函数转换、模糊量化法等。经过对比分析，采用 min-max 标准化方法进行数据标准化。

min-max 标准化方法也称为离差标准化，是对数据的线性压缩变换，使结果落入到 [0,1] 区间内。记 *max* 为各气象因子样本数据的最大值，*min* 为气象因子样本数据的最小值，需要进行标准化的数据序列为 *x1,x2,x3...xn*，那么，依次数据进行标准化的变换公式为：

$$y_i = \frac{x_i - \min\limits_{1\le j\le n}\{x_j\}}{\max\limits_{1\le j\le n}\{x_j\} - \min\limits_{1\le j\le n}\{x_j\}} \tag{4-3}$$

数据的最终标准化结果请查看提交的支撑材料“问题二数据标准化结果”。

4.3 建立气象因子分类模型

4.3.1 多元线性回归模型

某种污染物浓度可能受一种气象因子或多种气象因子的综合影响，故针对每一个污染物浓度，建立气象因子与污染物浓度的多元线性回归模型，通过回归函数可以清晰地得到气象条件对污染物浓度的影响，从而进一步对气象进行分类，并分析气象条件对 AQI 的影响。

多元线性回归的一般形式可以表示为：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_n X_n + e \tag{4-4}$$

Y 近似地表示为自变量 $X_1, X_2, \ldots X_n$ 的线性函数，β_0 为常数项，β_1，β_2，$\ldots \beta_m$ 为不同自变量对应的偏回归系数，e 是将 n 个自变量对 y 的影响去除后的随机误差，即回归模型的残差值。

本节将气象因子作为模型的自变量，污染物浓度作为因变量。根据第二节相关性分析，风向对各污染物浓度的影响较为微弱，且根据资料查找及基本常识判断，风的方向对周边区域的监测点的污染物浓度有一定影响，但对本区域的污染物浓度影响可以忽略，故仅取温度（T）、湿度（H）、气压（A）、风速（W）作为自变量来建立回归模型，进行回归分析。最终建立的回归模型的形式如（4-5）所示：

$$Y_\Omega = \beta_0 + \beta_1 T + \beta_2 H + \beta_3 A + \beta_4 W \tag{4-5}$$

其中，Ω 表示任意一种污染物。

对于线性回归模型，需要对其进行检测验证模型的准确性和稳定性，从而衡量回归效果，一般包含偏差平方和 q，平均标准差 s，复相关系数 r 等。其计算方式如下：

$$q = \sum_{i=1}^{n}\left(y - \beta_0 - \sum_{j=1}^{k}\beta_j x_j\right)^2 \tag{4-6}$$

$$s = \sqrt{\frac{q}{n}} \tag{4-7}$$

$$r = \sqrt{1 - \frac{q}{d_{yy}}} \tag{4-8}$$

式中 $d_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$，$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$，当 r 接近于 1 时，说明相对误差 $\frac{q}{d_{yy}}$ 接近于零，线性回归效果好。

4.3.2 基于 k-means 聚类的气象分类模型

为了对气象条件进行分类，首先对温度、湿度、气压、风速进行等级划分，划分依据的根据每个气象因子的最大最小区间进行等区间划分，划分为低、中、高三个等级，分别用数值 0、1、2 来表示。如表 4.3 所示为各气象因子等级划分情况。

表 4.3 各气象因子等级划分

等级	0	1	2
温度(℃)	(0,16.5]	(16.5,27.2]	(27.2,45]
湿度(%)	(0,42]	(42,70]	(70,100]
气压(MBar)	(900,1005]	(1005,1017]	(1005,1035]
风速(m/s)	(0,1.86]	(1.86,3.71]	(3.71,8]

根据以上划分方式对标准化后的气象情况进行划分，划分结果可查看支撑材料“问题二气象因子等级划分结果”。对划分结果进行 k-means 聚类。

k-means 算法是一种经典是无监督学习聚类方法。给定一个训练集*{x(1)，…，x(m)}*(其中 *x(i) ∈Rn*)，将数据分组成几个内聚的“簇”。通过不断迭代逐次更新各聚类中心的值，直至得到最好的聚类结果。k-means 算法的基本思想是初始随机给定 *k* 个簇中心，按照最邻近选取的原则把待分类的样本点分到各个簇中。然后，按平均法重新计算各个簇的质心，从而确定新的簇心。一直迭代，直到簇心的移动距离小于某个给定的值。k-means 聚类算法主要分为 4 个步骤，如下：

Step1：随机选取 k 个点作为初步聚类的中心；

Step2：计算其他数据到每个聚类中心的距离，将数据归入到与其距离最近的聚类中心的簇中；

Step3：对这 k 个聚类的数据计算均值，作为新的聚类中心；

Step4：继续以上的三个步骤，直到新的聚类中心与上次的聚类中心的值相等时结束算法。

该方法存在 k 值选取、初始聚类中心选择、及终止条件选取的问题，故要选取合适的方法进行预设。

（1）*k* 值选取

对于 *k* 的计算方法包括层次聚合、稳定性方法、手肘法等，本文采用手肘法进行 *k* 的选取。手肘法（elbow 法）的核心指标是误差平方和（*SSE*），即所有样本的聚类误差，随着聚类数 *k* 的增加样本划分会更加精细，则 *SSE* 就会逐渐减小。*k* 小于真实聚类数时，随着 *k* 的增加 *SSE* 会大幅下降，而达到真实聚类数后，*SSE* 的小降幅度会减小并趋于平缓，也就是 *SSE* 和 *k* 的关系图是手肘的形状，而肘部对应的 *k* 值就是数据的真实聚类。

（2）初始质心的选取：

选择适当的初始质心是基本 k-means 算法的关键步骤。常见的方法是随机选取，但是这样簇的质量常常很差。选取初始质心问题的一种常用方法有多种：

第一种是通过多次运行，每次使用一组不同的随机初始质心，然后选取具有最小 *SSE*（误差的平方和）的簇集。这种策略简单，但是效果可能不好，取决于数据集和寻找的簇个数。第二种方法是取一个样本，用层次聚类技术聚类。从层次聚类中提取 *k* 个簇，并用这些簇的质心作为初始质心。该方法通常很有效。第三种是随机地选择第

一个点，或将所有点的质心作为第一个点，对于每个后继初始质心，选择离已经选取过的初始质心最远的点。这种方法，确保了选择的初始质心不仅是随机的，还是散开的，但这样可能会选中离群点。

（3）算法停止条件：

有两种方法来终止迭代，一种方法是设定迭代次数 T，到达第 T 次迭代，终止迭代。此时，所得的类簇即为最终的聚类的结果；另一种是采用误差平方和准则函数，函数模型如下：

$$J=\sum_{k=1}^{K}\sum_{x_i\in C_k} dist(x_i, Center_k) \tag{4-9}$$

（4）类簇中心的重新计算：

k-means 算法每次迭代对应的类簇中心要重新计算进行更新。定义第 k 个类簇的类簇中心为 $Center_k$，则中心更新方式如下：

$$Center_k=\frac{1}{|C_k|}\sum_{x_i\in C_k} x_i \tag{4-10}$$

C_k 表示第 k 个类簇中数据对象的个数，求和是指类簇 C_k 中所有元素在每个属性上的和。

4.4 模型求解结果

4. 4. 1 回归模型结果及检验

根据第三节中的多元线性回归模型求解出的关于气象因子的回归模型分别如下：

$$Y_{SO_2}=1.010T-8.661H+2.889A-7.045W+12.167 \tag{4-11}$$

$$Y_{CO}=-0.447T-0.123H-0.006A-0.678W+1.212 \tag{4-12}$$

$$Y_{NO_2}=-45.660T-18.289H-6.205A-96.942W+96.319 \tag{4-13}$$

$$Y_{O_3}=151.995T-145.762H+45.842A+3.396W+35.570 \tag{4-14}$$

$$Y_{PM_{10}}=10.167T-59.165H+37.188A-81.559W+76.640 \tag{4-15}$$

$$Y_{PM_{2.5}}=-1.571T-29.201H+24.096A-61.292W+45.717 \tag{4-16}$$

根据权重系数可以判断不同气象因素对污染物浓度的扩散和抑制的影响，及影响程度的大小。为了证明回归模型的回归性能，对每个污染物浓度回归模型进行估计与检测。以二氧化硫的回归模型为例，展示其检测结果，其余检测结果保存在支撑材料“问题二回归模型检测结果”中。

（1）SO_2 的回归预测模型的方差分析

由表 4.4 的模型汇总结果可以初步分析出回归性能。由表中信息可知，模型的复相关系数 R=0.494，距离 1 较远，说明自变量和因变量之间的相关性一般。根据可决系数 R^2=0.1 和调整后的可决系数 R^2=0.244 来看，该模型可解释 24.4%的因变量差异，说

明模型的拟合程度一般，解释能力也较为一般。

表 4.4 模型汇总结果

R	R^2	调整后的 R^2	标准估算的误差
0.494[a]	0.244	0.244	3.2509502

a. 预测变量: (常量), 风速, 气压, 湿度, 温度。

b. 因变量: SO2 监测浓度 μg/m³。

由表 4.4 中的方差分析结果可知，回归方程分析的显著性检验值是 0.000，表明全部系数为 0 的原假设均被拒绝，该模型方程高度显著，说明能见度与特征变量之间的线性关系是成立的。此外，可以看出，T 检验显著性水平小于 10%的回归系数超过一半，说明回归方程中相应偏回归系数均显著。根据显著性水平检验分析得知数据拟合模型具有统计意义，所得线性方程成立且有效。

表 4.5 模型方差分析

模型	平方和	df	均方	F	Sig
回归	65327.478	4	16331.869	1545.309	0.000[b]
残差	202051.975	19118	10.569	-	-
总计	267379.453	19122		-	-

a. 因变量: SO2 监测浓度 μg/m³

b. 预测变量: (常量), 风速, 气压, 湿度, 温度。

（2）共线性检测

如表 4.6 所示的共线性诊断，可以看出，特征值具有三个几乎等于零的特征值，而条件 指标也具有三个大于 10 的值，证明存在多重共线性。

表 4.6 共线性的诊断结果

维数	特征值	条件索引	（常量）	温度	湿度	气压	风速
1	4.539	1.000	0.00	0.00	0.00	0.00	0.01
2	0.204	4.714	0.00	0.02	0.06	0.04	0.27
3	0.197	4.798	0.00	0.01	0.01	0.06	0.45
4	0.056	9.029	0.00	0.15	0.47	0.02	0.22
5	0.004	35.836	1.00	0.82	0.46	0.89	0.05

a. 因变量: SO2 监测浓度 μg/m³

（3）残差的正态性检验

如图 4.4 所示的残差的直方图和图 4.4 所示的累计概率图可以看出，残差的分布近似满足正态分布，通过分析，该二氧化硫浓度模型满足各种假设，具有较好的拟合效果和解释能力，具有一定的统计意义。因此，该模型可用于分析和解释气象影响因子对二氧化硫浓度的影响。

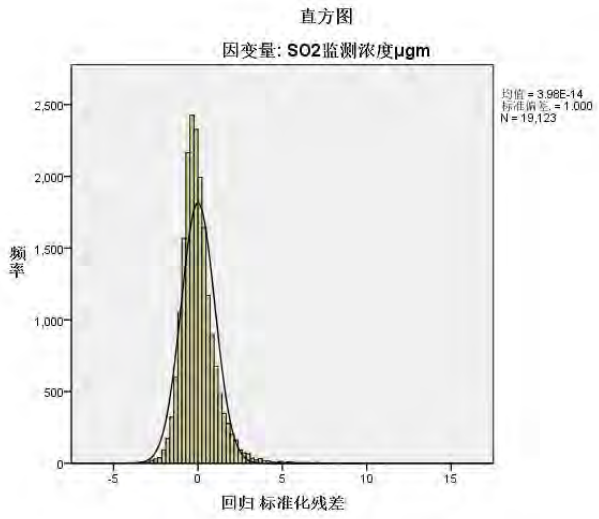


图 4.4 残差直方图

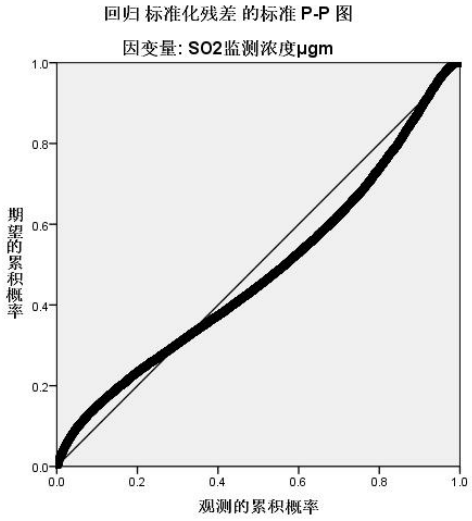


图 4.5 回归标准化残差图

4.4.2 气象分类结果及气象特征分析

采用手肘法进行聚类数量 k 的选取，如图 4.6 所示，为聚类中心方差随个数变化，当聚类个数为 5 时，误差平方和（SSE）有大幅度的降低。图 4.7 所示是不同聚类数下的模型打分，当 k=5 时，打分最高，与手肘法的判断结果一致。

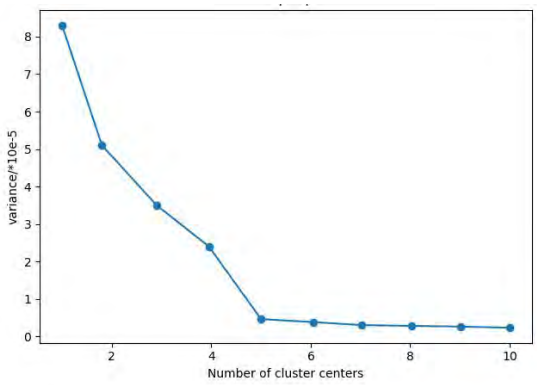


图 4.6 聚类中心方差随个数变化

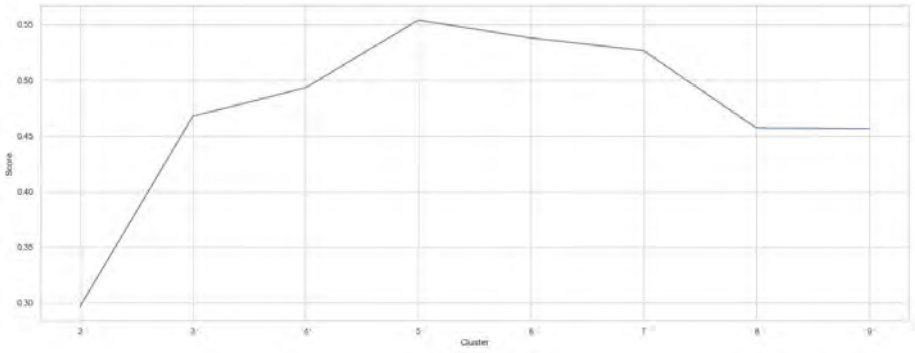


图 4.7 聚类打分

表 4.7 所示为最终的 5 个聚类簇的中心。聚类中心越大说明该气象因子的值越大。如温度和气压相对较高、湿度和气压相对较低时，属于第一类聚类簇。在实际应用中，根据各气象因子的等级计算其与每个簇中心的距离，距离最近的是该时间气象情况所

归属的类。如表 4.8 所示为最终分类结果中每个簇的个例的数量，及各类气象条件的特征分析。

表 4.7 k-means 聚类中心

气象分类等级	第一类	第二类	第三类	第四类	第五类
温度	1.86	0.35	1.28	1.61	0.24
湿度	0.33	0.51	1.84	1.43	0.84
气压	0.42	1.78	0.45	0.33	1.64
风速	1.39	1.23	0.22	1.62	0.67

根据各气象因子的等级划分与各气象分类等级的聚类结果以及气象因子与污染物的相关性，分析可知五类气象分类等级的气象特征如下：

第一类气象等级，温度区间为 16.5℃到 27.2℃之间，湿度区间为 0%到 42%，气压区间为 900 到 1005Mar，风速区间为 1.86 到 3.71m/s，此类气象由于湿度低、温度高等高影响因子不适宜污染物扩散，气压低、风速适中作为中、低影响因子的污染物扩散条件，虽有利于扩散，并不能对污染物扩散起到决定作用，所以此类气象条件不利于污染物的沉降，但同时存在适中污染物扩散有利条件。

第二类气象等级，温度区间为 0 度到 16.5℃，湿度区间为 0%到 42%，气压区间为 1005 到 1017MBar，风速区间为 1.86 到 3.71m/s;此类气象条件由于低温、低湿、压强较高等都是对污染物扩散不利的高影响因子，仅有低影响因子的风速处于适中状态，不会对污染物扩散产生大的影响，所以此类气象条件不利于污染物沉降，但存在扩散有利因素。

第三类气象等级，温度区间为 16.5℃到 27.2℃，湿度区间为 42%到 70%，气压区间为 900 到 1005MBar，风速区间为 0 到 1.86m/s；此类气象条件湿度适中、温度适中、压强低等高影响因子气象条件都有利于污染物扩散，仅低影响的风速较低不利于污染物的扩散，所以此类气象条件极有利于污染物沉降，但是污染物扩散因素欠缺。

第四类气象等级，温度区间为 16.5℃到 27.2℃之间，湿度区间为 42%到 70%，气压区间为 900 到 1005MBar，风速区间为 1.86 到 3.71m/s，此类气象条件温度适中、湿度适中、压强低、风速适中，所有因素都适合污染物扩散，所以此类气象条件最适宜污染物扩散与沉降。

第五类气象等级，温度区间为 0 度到 16.5℃，湿度区间为 0%到 42%，气压区间为 1005 到 1017MBar，风速区间为 0 到 1.86m/s，此类气象条件温度低、湿度低、压强高、风速低都不适宜污染物扩散，所以此类气象条件最不适宜污染物扩散与沉降。

表 4.8 k-means 聚类结果及气象分类特征

气象分类等级	个数	气象特征
第一类	3732	中温、低湿、低压、中风 不利于污染物沉降，存在扩散适中有利条件
第二类	3957	低温、低湿、中压、中风 极不利于污染物沉降，存在扩散低有利条件
第三类	4551	中温、中湿、低压、低风 极有利于污染物沉降，但扩散条件欠缺
第四类	3203	中温、中湿、低压、中风

第五类	3680	最适宜污染物扩散 低温、低湿、中压、低风 最不适宜污染物扩散

五、 问题三模型的建立与求解

5.1 问题分析及建模思路

首先，为了建立同时适用于 A、B、C 三个监测点的污染物浓度预报模型，应首先对每个监测点的一次预报数据及实测数据进行预处理。在对缺失数据进行处理的过程中需考虑数据时间上的连续性，故应先对缺失数据进行简单的波动检测，再决定处理方式。此外，由于要进行不同监测点的污染物浓度的预测准确度分析，故需要将附件一和附件二中的实测数据和预测数据进行整合。

其次，建立污染物浓度预测模型。通过大量的文献查找，总结出常用的大气污染物预测方法主要包括多元回归预测、灰色 GM(1,1)预测、支持向量机 SVM、RNN 预测、长短期记忆网络（LSTM，Long Short-Term Memory）预测[7]等模型，并对比了不同方法的优缺点，如表 5.1 所示。考虑到大气污染物浓度数据具有时序性和非线性的特点，以及不同方法的比较，本问题选择建立基于 LSTM-FC 的大气污染物浓度预测模型。

表 5. 1 不同大气污染物预测方法的比较

方法	优点	缺点
多元回归预测	原理简单、易实现。	未考虑大气污染物浓度的时序性和非线性特点。
灰色 GM(1,1)预测	原理简单、易实现。	未考虑大气污染物浓度的时序性和非线性特点。
SVM	考虑了时序性特点。	参数的选取依赖主观选择。
RNN 预测	擅长处理具有连续时间序列的数据。	容易出现梯度消失。
LSTM	擅长处理具有时序及非线性的数据，适用性强，可以有效防止梯度消失。	在并行处理上存在一定的劣势。

如图 5.1 所示是基于 LSTM-FC 的污染物浓度预测模型建立流程图。对于神经网络的训练，需要拆分出训练集、验证集以及测试集，其中训练集用于模型训练，验证集可以用来调整分类器的参数，确定网络结构，而测试集合可以用来检验模型的预测性能和泛化能力。本问题将数据按照 7:2:1 的比例进行训练集、验证集和测试集的拆分，将训练集与验证集输入 LSTM-FC 神经网络进行训练，利用训练集与验证集的效果，不断调整训练参数，通过反复迭代得到污染物浓度预测效果最好的模型。对测试集中的数据进行浓度预测，验证训练好的模型的准确率。

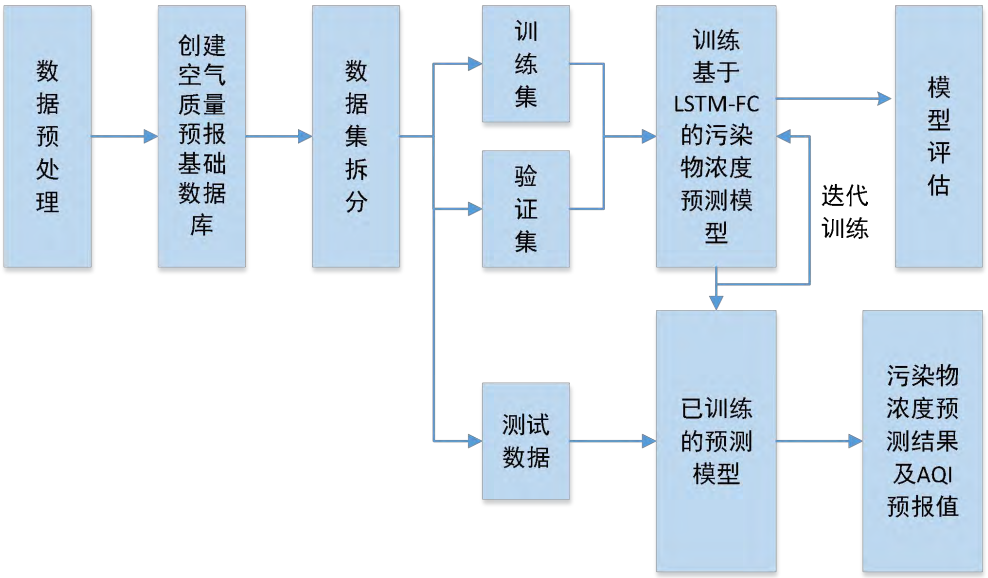


图 5.1 建立基于 LSTM 的污染物浓度预测模型流程图

此外，本问题中臭氧 O_3 属于二次污染物，气象条件对臭氧 O_3 的生成有很大影响。主要的气象因子如温度、相对湿度、太阳辐射、以及风速会影响臭氧生成有关的光化学反应，和大气运动对臭氧及其前体污染物的清除速率。故本问题建立两个污染物浓度预测模型，其中一个是关于一次污染物浓度的预测模型，一个是关于二次污染物臭氧的浓度预测模型，如图 5.2 所示。

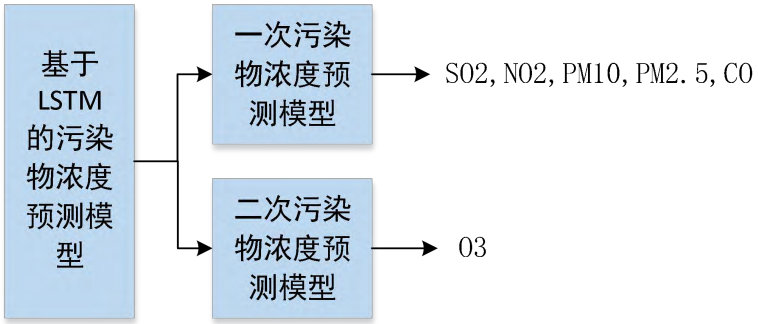


图 5.2 不同污染物预测模型

5.2 数据分析及预处理

附件中模型运行日期范围为 2020-07-23 至 2021-07-13，实测数据的范围为 2019-04-16 至 2021-07-13，为了将预测时间和实测数据相对应，只提取 2020-07-23 至 2021-07-13 的数据进行预处理。

在采集大气污染物浓度过程中，时常存在数据质量参差不齐、异常或缺失值等数据质量问题，需要进行预处理等操作。在利用神经网络建立预测模型的过程中，如果不能选择合适的数据预处理方法，数据的分析结果可能会产生严重偏差，影响整个模型的预测准确率，因此，本问题需要对数据进行较为细致的预处理。预处理的具体流程如图 5.3 所示。

首先，使用 MySQL 数据库进行表合并，将预测数据和实测数据相对应，再进行其余的预处理操作。由于本节所用的数据是具有时序性的，故对数据进行缺失性检测后，不可以直接进行删除或补全，否则会影响数据的连续性及预测模型的准确性。故本文对缺失的数据进行波动性检测，分析缺失数据周围数据的发展规律，如果近邻数据呈

现一定的规律性，譬如单调递增递减、平稳或略微波动等，则根据周围的发展规律进行补齐；对于近邻数据不存在明显规律，起伏不定的，则直接删除。

对缺失数据进行处理后，使用箱型法对预测数据和实测数据进行异常检测，如图5.4所示是箱型图。根据检测结果，部分浓度数据存在负值，故删除这些存在负值的数据行。对三个监测点的预处理后的数据可查看支撑材料“问题三数据预处理结果”。

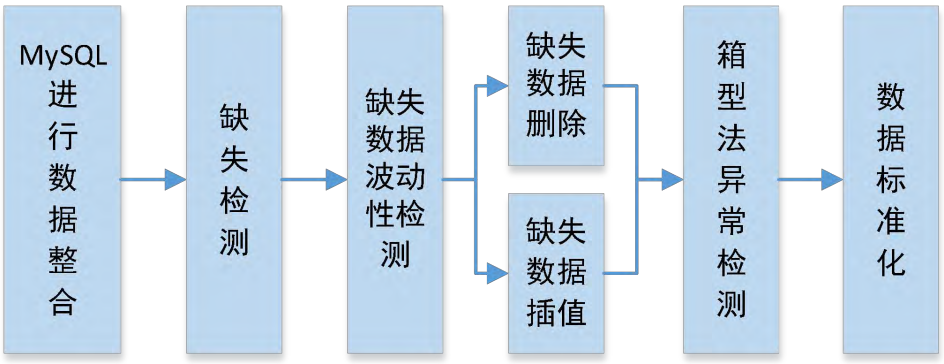


图 5.3 数据预处理的具体流程

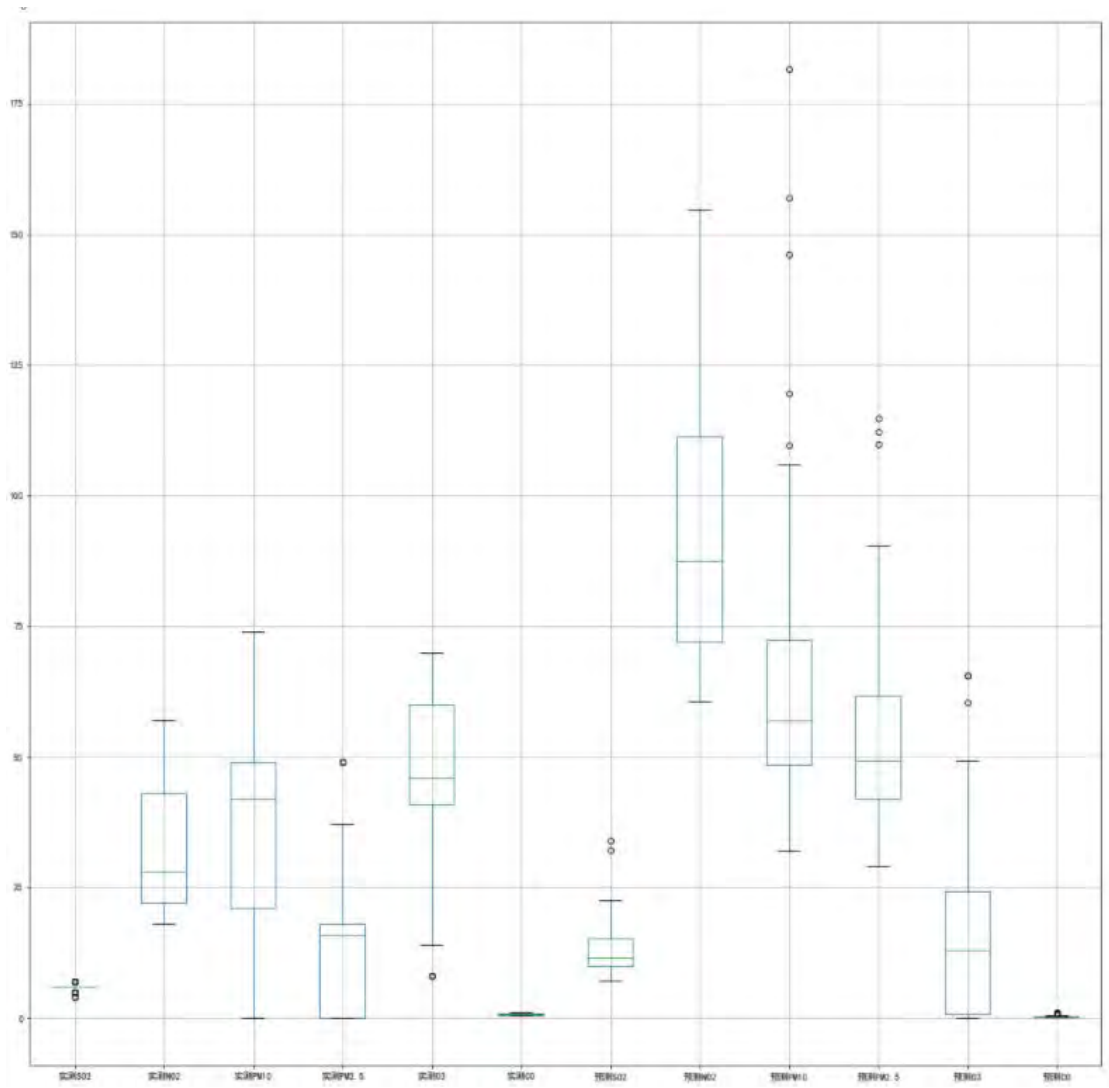


图 5.4 六种污染物浓度

5.3 建立基于 LSTM-FC 的大气污染物浓度预测模型

5.3.1 基于 LSTM-FC 神经网络模型

本问题构建出一种基于长短期记忆神经网络和全连接神经网络（Full Connested, FC）的混合型神经网络模型，即 LATM-FC 神经网络模型[8]。LSTM 能够解决气象与污染物序列的长时间依赖性，而全连接神经网络能够准确地捕捉气象与污染物序列的空间上的关联性，深入挖掘气象数据和污染物数据之间的时空特性。

（1）LSTM 网络模型建立

LSTM 神经网络[9][10]是递归神经网络（RNN，Recurrent Neural Network）的一种，能够对长期依赖的某种信息进行学习。RNN 网络的特点是可以将历史信息作为输入连接到对当前信息的判断上，适合处理序列数据，通常用来解决文本生成、机器翻译、和 DNA 序列的分析等问题。然而，传统的 RNN 存在梯度消失的问题，随着网络深度的逐渐加深，后续节点对前层节点的感知会渐渐变弱，所以，传统 RNN 只对短期的记忆信息较为敏感，不能有效地捕获长期记忆信息。问题三需要预测未来 3 天的 6 种常规污染物的单日浓度值，故选择使用在 RNN 基础上改良的 LSTM 网络模型，LSTM 增加了对信息的判断，即判断信息的有用性，有用则保留，无用则遗忘，更近似地模拟了人脑进行决策的过程，即长短期记忆网络。如图 5.5 所示为 LSTM 网络的单元结构。LSTM 增加了输入门、遗忘门和输出门，缓解了模型训练中梯度消失和梯度爆炸的问题，弥补了传统 RNN 模型的不足。

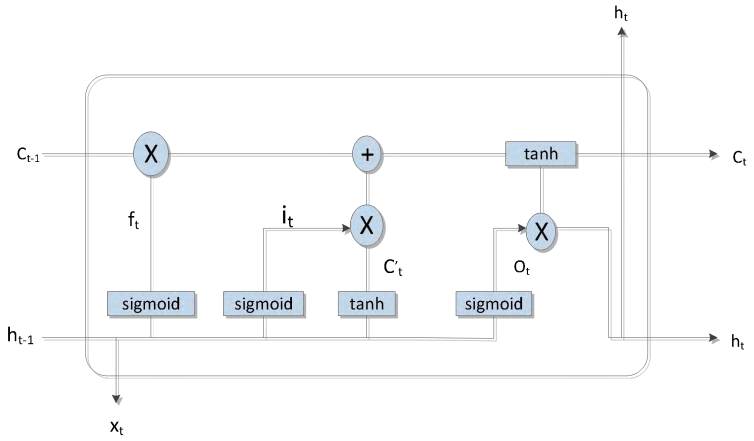


图 5. 5 LSTM 网络单元结构

如图 5.5，LSTM 网络的单元结构共有 4 层，h_t、h_{t-1} 分别为当前单元及上一个单元的输出；x_t 为当前单元的输入；*sigmoid*、*tanh* 为激活函数；图中的圆形结点均表示向量之间的某种算数规则；C_t 为神经元在 t 时刻的状态；f_t 为遗忘阈值，该阈值通过 *sigmoid* 激活函数控制细胞应该如何丢失信息；i_t 为输入阈值，该阈值决定了 *sigmoid* 函数需要更新的信息，然后通过 *tanh* 激活函数生成新的记忆 C_t，并最终控制应该向神经元添加的新信息；o_t 为输出阈值，该阈值决定了 *sigmoid* 函数输出神经元状态的哪些部分，并使用 *tanh* 激活函数处理神经元状态，得到最终的结果。计算公式分别如下：

$$f_t = sigmoid\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \tag{5-1}$$

$$i_t = sigmoid\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \tag{5-2}$$

$$C_t' = \tanh\left(W_c \cdot \left[h_{t-1}, x_t\right] + b_C\right) \tag{5-3}$$

$$o_t = sigmoid\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right) \tag{5-4}$$

其中，W_f、W_i、W_o、W_C 依次是遗忘门、输入门、输出门、以及神经元状态矩阵对应的权系数矩阵；b_f、b_i、b_o、b_C 依次分别表示对应的偏移常量值。根据以上的公式，可以进一步计算出神经元的状态 C_t 和输出 h_t。

$$C_t = f_t C_{t-1} + i_t C_t' \tag{5-5}$$

$$h_t = o_t \cdot \tanh\left(C_t\right) \tag{5-6}$$

根据以上步骤，通过 3 个控制门的机制，可以完成神经元的内部处理过程，保证 LSTM 网络模型能够有效地利用输入数据，探索数据的隐藏信息，对过去长期的数据形成记忆，从而学习长期以来的数据关系。

（2）全连接神经网络 FC

全连接神经网络是最常见的一种网络结构，能够探索数据的空间上的关联性。如图 5.6 所示，全连接神经网络通常是 3 层结构，分别为输入层、隐藏层、和输出层，其相邻层的每两个神经元之间都存在连接性。

全连接神经网络在解决复杂问题时表现出很好的性能，尤其是结构合理的全连接神经网络可以有效地解决其他结构不能解决的复杂问题。但 FC 也带来了参数过多、难以训练的问题。全连接神经网络一般包含线性和非线性部分，前者一般指在前向传播过程中将结点的输入与对应的权重做线性乘法，在这个过程中权重和偏置项起主要作用，这也是需要通过反向传播算法进行训练学习的参数，其数学公式为：

$$Y = W * x + b \tag{5-7}$$

其中，Y、W 分别为输出和权重向量，x 为输入向量，b 为偏置项。

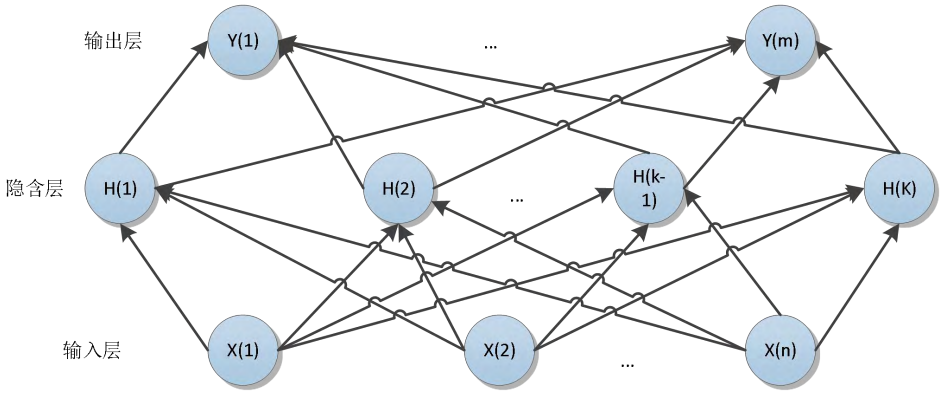


图 5. 6 全连接神经网络结构

全连接神经网络的非线性部分则是由激活函数来实现。现实中的大部分应用问题都是非线性的，如果只有线性的特性无法满足对数据进行拟合的要求，因此将激活函数引入到神经网络中去。激活函数将其非线性特性引入到全神经网络中，这样神经网络就可以逼近任意的非线性函数，从而扩展了神经网络的应用范围。

（3）激活函数

常用的激活函数有 *Sigmoid*、*Tanh*、*ReLU* 等等，本文的污染物浓度预测模型使用了 *ReLU* 和 *Elu* 两种激活函数，其数学形式如式（5-8）和式（5-9）所示。

$$\mathrm{Re}\,LU(x)=\begin{cases}0, x<0\\ x, x\geq 0\end{cases} \tag{5-8}$$

$$Elu(x)=\begin{cases}\alpha\left(e\hat{}x-1\right), x<0\\ x, x\geq 0\end{cases} \tag{5-9}$$

ReLU 激活函数的运算简单，能加快模型收敛速度并减少计算代价，由于其在大于 0 处导数恒为 1，避免了 *Sigmoid* 函数中存在的由于函数两端梯度接近于 0 而产生的梯度消失的问题，广泛用于神经网络模型中。

Elu 在 x>0 的区间取输入 *x*，减轻了梯度弥散的问题（x>0 区间处的导数为 1），而 *ReLU* 的输出值非负，则均值大于 0；当激活值的均值不为 0 时，会对下一层造成偏差 *bias*，如果激活值之间没有相互抵消作用（即均值非 0），会导致下一层的激活单元有偏差漂移 *Bias-shift*。通过叠加，单元越多 *Bias-shift* 就会越大。*Elu* 在输入取值较小时具有软饱和的特性，对噪声有很好的鲁棒性。

（3）Dropout

Dropout 是在本问题训练过程中采用的训练策略。在相邻层之间随机忽略一些神经元节点之间的连接，从而可以达到泛化模型的作用。在训练过程中，可以调整 Dropout 的比率来规定要忽略掉的神经元连接的数量，经验上设置为 0.5 或者 0.25。

（4）损失函数

损失函数可以用来衡量在训练过程中污染物浓度模型的预报值和真实监测值之间的差距，本问题采用的是较为常见的均方误差（Mean Squared Error，MSE），定义为：

$$MSE=\frac{\sum_{i=0}^{N}\left(\hat{y}_i-y_i\right)^2}{N} \tag{5-10}$$

其中，$\hat{y}_i$ 为模型的预测数值，y_i 为真实的数值，N 是样本的总数量。*MSE* 越趋近于 0 值，表明模型在对应数据集上的拟合效果就越好。那么，最小化损失函数就成为整个模型训练所寻求的目标。若在训练集和测试集上损失函数的数值都比较小，认为模型从数据中学到了某种模式，并具有一定的泛化能力。

（5）优化器

本文采用的优化算法是 Adam 算法。相比于传统的梯度下降等优化算法，Adam 优化算法能够自动调整权重的学习率，使得整个模型更快地收敛，并有效地改善陷入局部最优解的问题。

（6）LSTM-FC 模型

如图 5.7 所示，给出了构建的 LSTM-FC 模型的整体网络结构。LSTM-FC 模型网络结构由 3 个部分构成。模型的第一部分由两层的 LSTM 模型和 Dropout 构成，其输入需要转换为 *k*M*N* 的矩阵，其中 *k* 是数据的总条数，*M* 是输入的特征数，*N* 是时间步数。时间步长指的是采用先前多久的数据作为输入进行预测。两层的 LSTM 分别包含了 512 个节点和 64 个节点。

预测模型的第二部分包括两层全连接层和 Dropout。两层全连接层的节点数目分别为 128 个和 64 个，在两层之后设置 Dropout 是为了防止过拟合。本模型将 Dropout 率

设置为 0.25，并且在第二层的全连接层 FC 采用了激活函数 *ReLU*。

预测模型的第三部分是输出层。预测任务的最终输出为一个预测数值，故最后一层是只含一个节点的全连接层。该全连接层使用了 *ReLU* 作为激活函数。此外，本模型以 *MSE* 作为其损失函数，并使用 Adam 优化算法作为更新其权重的优化器。

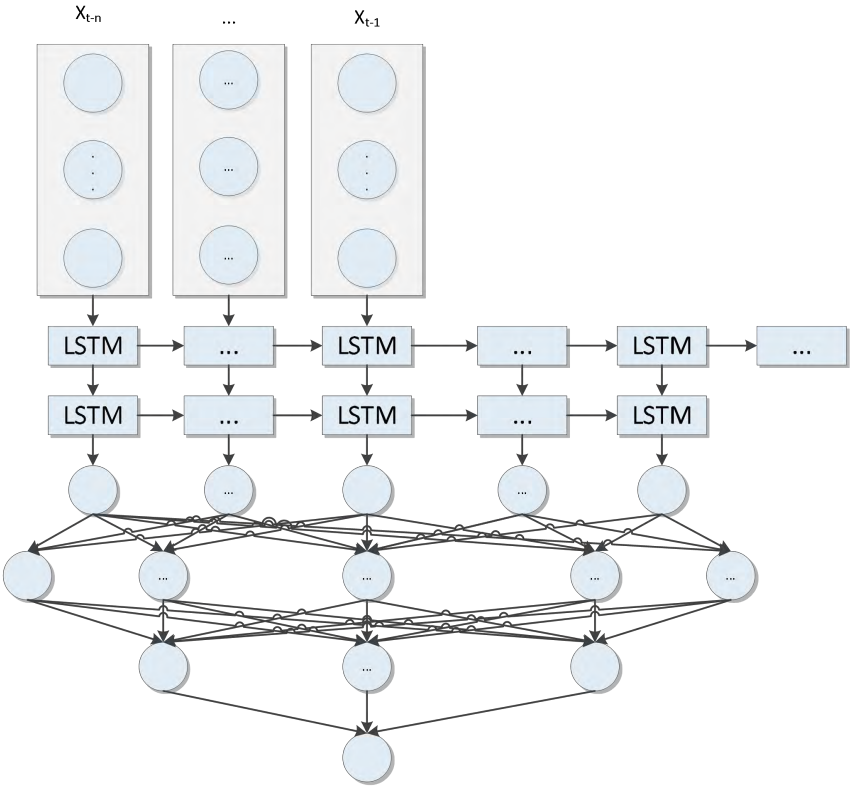


图 5.7 LSTM-FC 模型结构

5.3.2 基于 LSTM-FC 的一次污染物浓度的预测模型训练

如图 5.8 所示，五种一次污染物 SO2、NO2、CO、PM10、PM2.5 的浓度预测模型，输入数据主要是各该污染物的预报浓度，以实测的污染物浓度作为训练标签（label），经过模型训练与迭代，获得优化后的污染物浓度预测值。

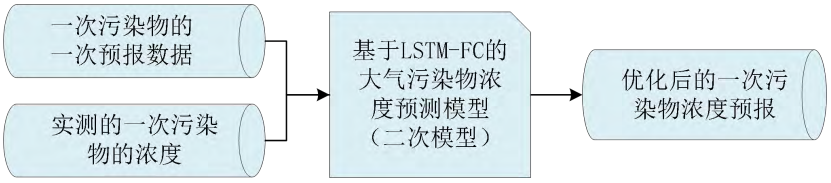


图 5.8 基于 LSTM-FC 的一次污染物浓度的预测模型

5.3.3 基于 LSTM-FC 的二次污染物（臭氧）浓度的预测模型训练

如图 5.9 所示，为臭氧 O_3 的浓度预测模型，由于臭氧属于二次污染物，要考虑各种气象条件以及氮氧化物的影响，故输入数据主要是臭氧的预报浓度、气象因子、二氧化氮的预报数据，以实测的臭氧浓度作为训练标签（label），经过模型训练与迭代，获得优化后的臭氧浓度预测值[11]。

对于气象因子选择，预报数据中给出了 15 种气象条件，但考虑到其对臭氧浓度的影响程度，需要对气象条件进行降维处理。可通过 PCA 降维法和相关性分析法进行气

象因子的筛选。考虑到运算量的大小，本文通过计算相关性选取气象因子。最终筛选出来的气象因子主要包括：温度、湿度、风速、太阳辐射、以及雨量。此外，为了保证选取的正确性，通过文献查找和标准查询，发现几种气象因子证是影响臭氧浓度的主要气象因素。

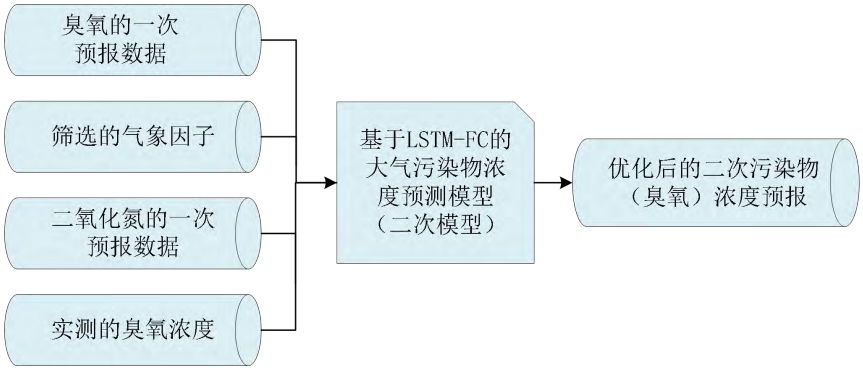


图 5.9 基于 LSTM-FC 的二次污染物（臭氧）浓度的预测模型

5.4 模型求解结果

5.4.1 模型评价指标

为了有效评价模型的性能本文使用了两种模型评价方法，分别是 *MSE* 和 R^2。

MSE 可以估计训练出的预测值与实际的浓度检测值之差的平方的期望值，*MSE* 的值越小，说明污染物浓度预测的精确度越高。

R^2 是指回归平方和与总离差平方和的一个比值，表示的是总离差平方和中可以由回归平方和来解释的比例。这一数值越大，模型就越精确，回归效果就越显著。其计算公式如下：

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y} - y_i)} \tag{5-11}$$

5.4.2 模型训练结果

如图 5.10 所示是模型训练过程中损失函数的动态变化结果，其中红色曲线是训练集损失值，黑色曲线表示验证集的损失值，蓝色是测试集的损失值。根据三条变化曲线可以看出，随着迭代次数的不断增加，三者的训练损失呈下降趋势，本文迭代次数为 50 次。

如图 5.11 至图 5.16 分别是二氧化硫 SO_2、二氧化氮 NO_2、PM_{10}、$PM_{2.5}$、臭氧 O_3，一氧化碳 CO 的实测数据与模型预测数据的曲线图，其中横坐标为测试的数量，纵坐标代表污染物浓度的实测值与预测值，红色是实测值，蓝色为预测值，两条曲线都较为接近，证明利用 LSTM-FC 网络进行大气污染物浓度预测的有效性。通过曲线图可以看出一氧化碳和二氧化氮的两个曲线的重合度较高，说明预测效果好，预测值与实际的检测值较为接近。而二氧化硫的预测效果较差。

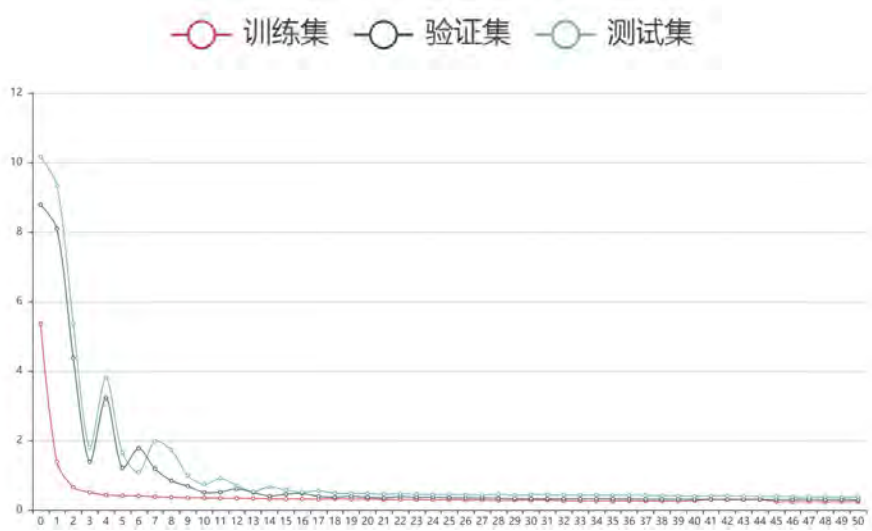


图 5. 10 模型训练结果

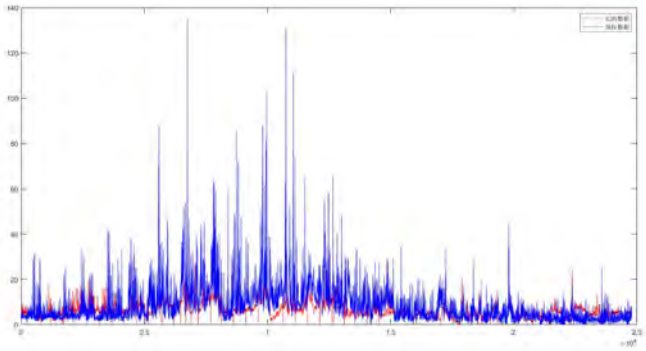


图 5. 11 SO_2 实测数据与模型预测数据对比曲线

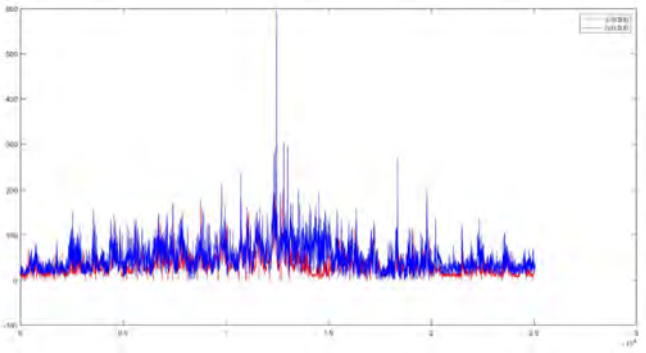


图 5. 12 NO_2 实测数据与模型预测数据对比曲线

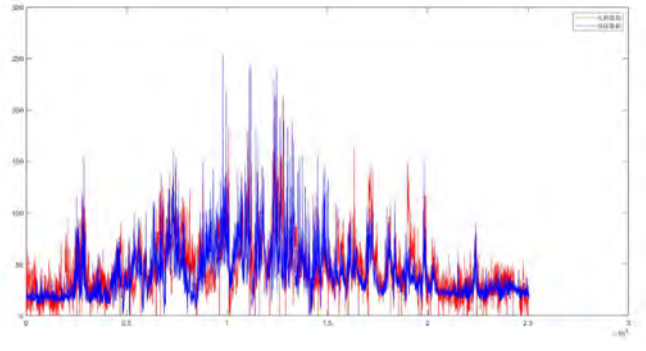


图 5. 13 PM_{10} 实测数据与模型预测数据对比曲线

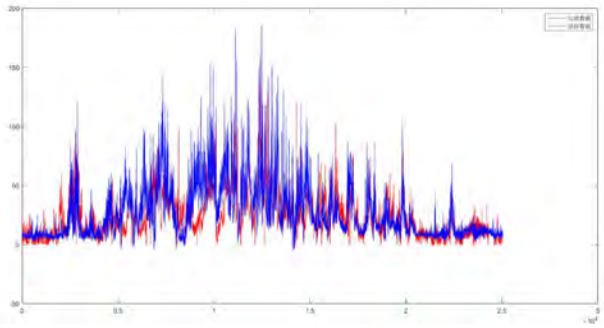


图 5. 14 $PM_{2.5}$ 实测数据与模型预测数据对比曲线

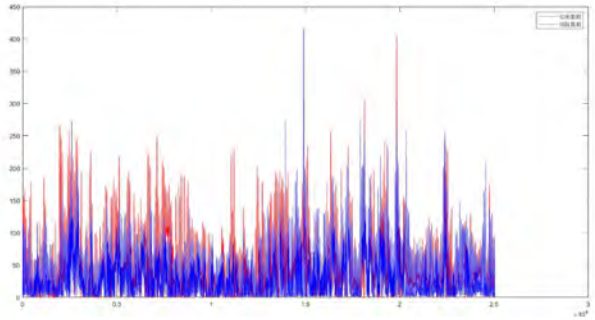


图 5. 15 O_3 实测数据与模型预测数据对比曲线

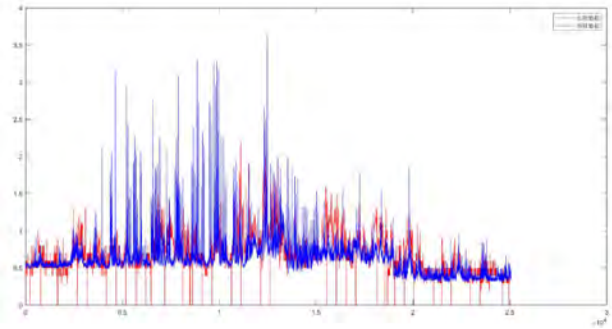


图 5. 16 CO 实测数据与模型预测数据对比曲线

如表 5.2 所示，是最终每个污染物浓度训练模型的模型评价，一些污染物浓度数值较大，主要是因为其原始的数值大，其相应的 *MSE* 误差就会大一些，但都在可接受的范围，根据评价结果，对二氧化氮和一氧化碳的预测效果较好，对二氧化硫的预测效果较差。总体而言，该模型能够达到较好的精度，具有较好的实用性。

表 5. 2 LSTM-FC 模型评估

污染物	*MSE*	R^2
SO_2 (μ g/m³)	9. 24	0. 70
NO_2 (μ g/m³)	0. 42	0. 95
PM_{10} (μ g/m³)	137. 42	0. 94
$PM_{2.5}$ (μ g/m³)	6. 75	0. 91
O_3 (μ g/m³)	2. 26	0. 92
CO (mg/m³)	0. 002	0. 97

如图 5.17 所示，分别是 WRF-CMAQ 模型臭氧 O_3 预测值和真实值差值曲线，以及本模型预测值与实测值差值曲线，根据曲线可以直观的看出，与原始的 WRF-CMAQ

预测模型相比，本模型预测效果较好，预测误差更小。对于一次污染物的预测模型，*MSE* 指标值平均下降了 5.4%，R^2 指标提高了 4.6%。对于二次污染物臭氧的预测模型，*MSE* 指标下降了 12.4%，R^2 指标提高了 9.3%。

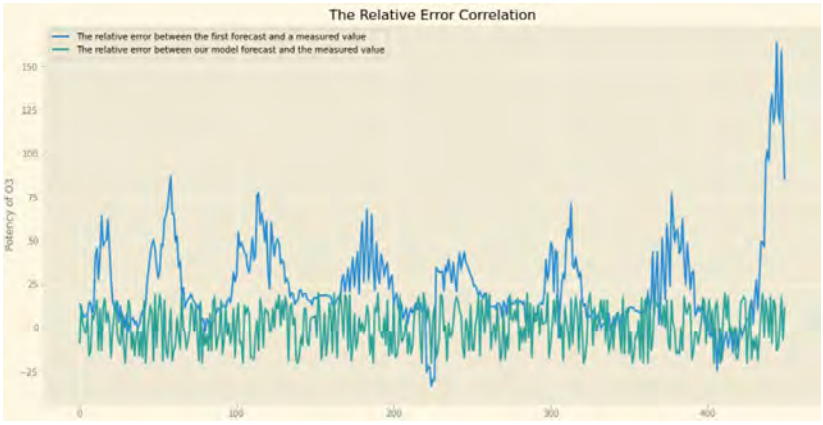


图 5. 17 WRF-CMAQ 模型、LSTM-FC 模型与实测数据的误差值曲线

表 5.3 至 5.5 所示为该模型预测的 2021/7/13 至 2021/7/15 日的监测点 A、B、C 的各污染物的浓度、AQI 数值及首要污染物种类。根据首要污染物，臭氧仍是出现概率最大的污染物。此外，在 2021/7/13 至 2021/7/15 日期间，监测点 B 的整体空气质量较好，根据空气质量划分等级，其三天空气质量均为“优”；监测点 C 的整体空气质量最差，空气质量为“良”或者“轻度污染”；监测点 A 的空气质量居于二者之间，三天空气质量均为“良”。

表 5. 3 监测点 A 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 ($\mu g/m^3$)	CO (mg/m^3)	AQI	首要污染物
2021/7/13	监测点 A	6. 505	15. 283	25. 910	7. 260	98. 678	0. 410	51	O_3
2021/7/14	监测点 A	6. 1285	19. 342	30. 986	10. 470	97. 306	0. 562	49	无
2021/7/15	监测点 A	5. 833	17. 820	33. 420	10. 580	130. 52	0. 583	76	O_3

表 5. 4 监测点 B 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 ($\mu g/m^3$)	CO (mg/m^3)	AQI	首要污染物
2021/7/13	监测点 B	5. 505	12. 323	19. 209	4. 233	56. 322	0. 590	29	无
2021/7/14	监测点 B	6. 130	12. 892	21. 521	5. 143	63. 788	0. 511	32	无
2021/7/15	监测点 B	6. 540	11. 720	20. 324	5. 311	73. 392	0. 469	37	无

表 5.5 监测点 C 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 (μg/m³)	NO_2 (μg/m³)	PM_{10} (μg/m³)	$PM_{2.5}$ (μg/m³)	O_3 (μg/m³)	CO (mg/m³)	AQI	首要污染物
2021/7/13	监测点 C	6.922	23.527	35.545	16.544	131.66	0.560	77	O_3
2021/7/14	监测点 C	8.900	23.470	34.834	16.983	134.70	0.519	79	O_3
2021/7/15	监测点 C	9.983	24.980	43.458	24.582	170.53	0.590	110	O_3

六、 问题四模型的建立与求解

6.1 问题分析及建模思路

大气气象及污染物浓度监测数据具有序列特性，也存在时间和空间上的特性，故空气质量会受到多种复杂因素的影响，而问题三只考虑了气象数据时间上的特性，但没有考虑不同区域之间的相互影响，如相邻区域的风速、风向等都会对当前区域的污染物浓度产生影响，忽略这一条件势必会降低浓度预测模型的性能。为了增强预报的准确性，本问题需要考虑相邻站点的气象基础数据[12][13][14]。

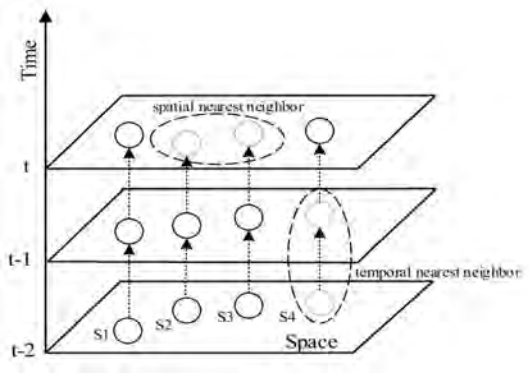


图 6.1 区域间的时空关联图

首先，应对附件三中的数据进行预处理，预处理方式与第五章的预处理方式相同。其次，对数据集进行训练集、验证集、以及测试集的拆分。考虑到相邻监测点与当前监测点的距离越大，对当前监测点的污染物浓度影响作用越小，本问题可根据相邻检测点与当前点之间的位置距离关系设置权重因子，本问题所给的 A1、A2、A3 三个监测点中，A1 与 A 的距离最远，故权重因子最小，A3 与 A 的距离最近，故权重因子最大。所以，为了确定具体的权重因子，需要选择合适的权重计算方式。

根据权重因子、相邻站点气象基础数据、本站点气象基础数据，在问题三 LSTM-FC 提出的中提出了基于时空加权融合的大气污染物预测模型，记为 STW-LSTM-FC（Spatial and Temporal Weighted - LSTM-FC），如图 6.2 所示为基本流程图。

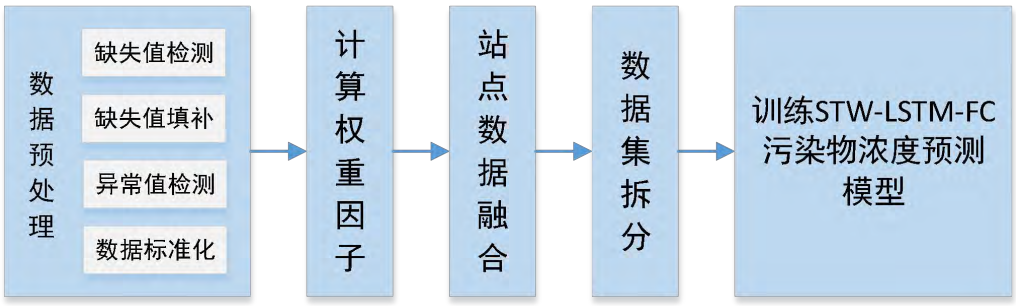


图 6.2 构建 STW-LSTM-FC 污染物浓度预测模型流程图

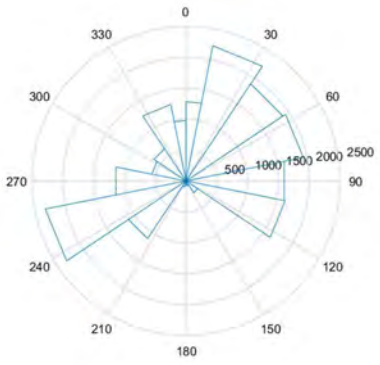
6.2 数据分析及预处理

6.2.1 数据预处理

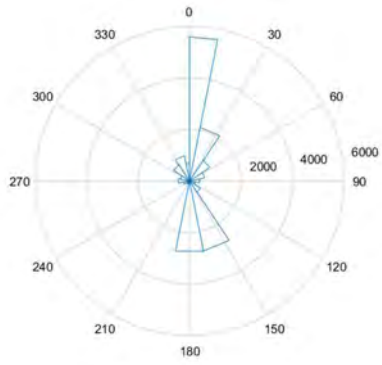
对于附件三中 A1、A2、A3 的气象及污染物浓度数据进行预处理，处理方式与问题二、三中的处理方式相同，主要进行缺失检测、异常检测、缺失值填补、及数据标准化。

6.2.2 数据分析

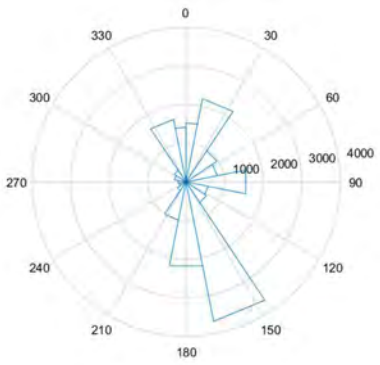
本问题需要考虑相邻区域间气象条件变化的互相影响情况，而相邻区域的影响因素有风向、风速等可能对当前区域有较大的的影响，故对风向数据进行了分析。如图 6.3 所示为检测点 A、A1、A2、A3 四个监测点的风向玫瑰风图，可以直观地显示不同风向的数据频次。



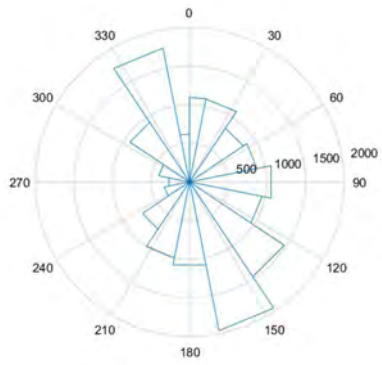
(a)监测点 A 风向图



(b)监测点 A1 风向图



(c)监测点 A2 风向图



(b)监测点 A3 风向图

图 6.3 各监测点风向图

6.3 建立 STW-LSTM-FC 模型

通过对污染物浓度预测问题的分析，空气质量数据间存在很强的时空相关性，如果在建模过程中只考虑时间上的相关性而忽略空间影响，势必降低预测的准确性。故本文同时考虑多个站点的气象数据，在问题三 LSTM-FC 模型的基础上，加入了图卷积神经网络 (Graph Convolutional Network, GCN) 模块，用于学习邻近区域的特征，如图 6.4 所示为 STW-LSTM-FC 模型的结构图。

本文中 A、A1、A2、A3 可以构成一个拓扑图，将多站点的气象基础数据整合成组成一组图数据，通过 GCN 对污染物图数据进行图卷积可以学习出监测点自身数据和邻近区域气象数据的隐藏特征，将特征聚合生成新的特征节点，有效利用了邻域数据信息。

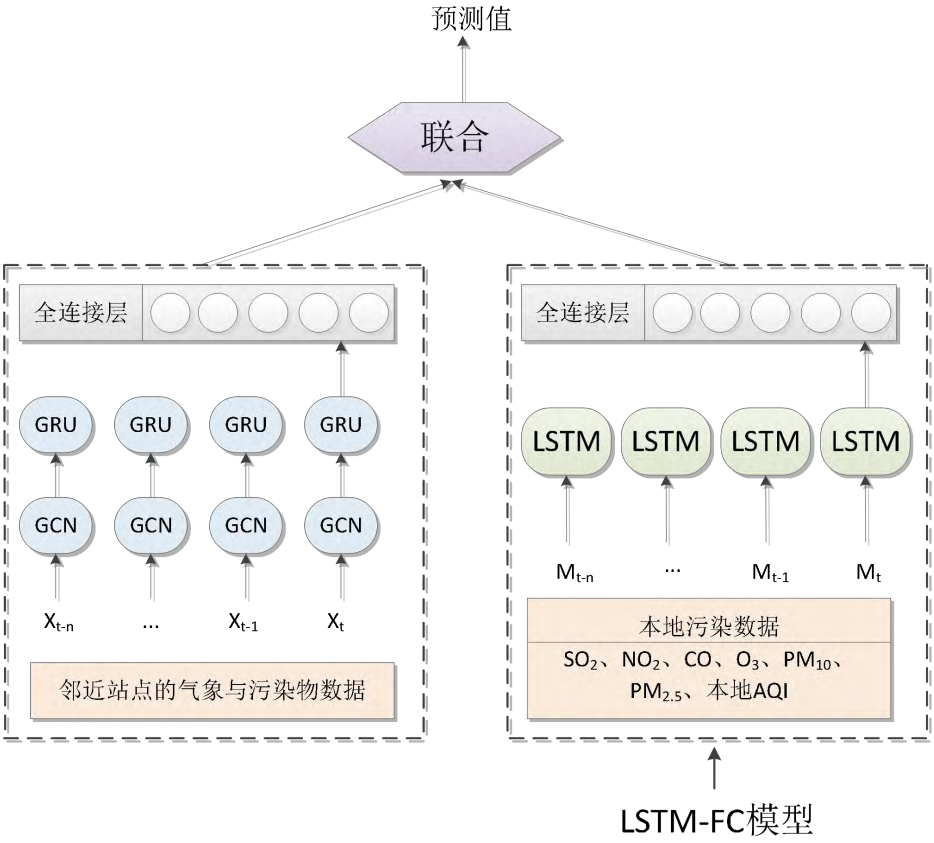


图 6. 4 STW-LSTM-FC 模型结构

如图 6.5 所示为用于学习多监测点数据特征的 GCN 模块，利用 CGN 提取多个监测点的空气质量特征的实现过程如下：

（1）根据 A、A1、A2、A3 监测点的位置信息可以构建监测点拓扑图 G，图的节点是每个监测点，边的权重与两结点之间的位置等关联信息有关，并将关联度信息表示为邻接矩阵 E。

（2）基于当前监测点 A 建立不同时刻的空气质量特征矩阵 X。

（3）对邻接矩阵 E 进行拉普拉斯变换得到生成矩阵 $\hat{E}$ 。本文生成矩阵的的大小为 4*4 的矩阵，如式（6-1）所示，其中 w_i^j 代表拉普拉斯矩阵中监测点 i 对监测点 j 的影响权重：

$$\text{生成矩阵} \rightarrow \begin{array}{c} \\ A \\ A1 \\ A2 \\ A3 \end{array} \begin{array}{c} \begin{array}{cccc} A & A1 & A2 & A3 \end{array} \\ \begin{bmatrix} W_A^A & W_A^{A_1} & W_A^{A_2} & W_A^{A_3} \\ W_{A_1}^A & W_{A_1}^{A_1} & W_{A_1}^{A_2} & W_{A_1}^{A_3} \\ W_{A_2}^A & W_{A_2}^{A_1} & W_{A_2}^{A_2} & W_{A_2}^{A_3} \\ W_{A_3}^A & W_{A_3}^{A_1} & W_{A_3}^{A_2} & W_{A_3}^{A_3} \end{bmatrix} \end{array} \tag{6-1}$$

（4）用下式转化相关联节点的特征来计算节点的新特征：

$$H^{(i+1)} = \sigma(\hat{E}H^{(i)}W^{(i)}) \tag{6-2}$$

其中，$\sigma(\bullet)$是非线性激活函数，$W^{(i)}$为第 i 层权值矩阵，$H^{(i)}$为第 i 层的激活值，且 $H^{(0)}=X$。将 CGN 提取的不同监测点数据的空间特征输入到门控循环单元网络 (Gated Recurrent Unit, GRU) 中，并将网络的输出输入进全连接神经网络做维度转换。

最后，将 LSTM-FC 的训练结果和 CGN 的训练结果进行特征融合，通过适当的权重值将二者的时空特征联合起来，联合公式为：

$$y = \alpha o_v + (1-\alpha)o_c \tag{6-3}$$

其中，α 为组合权值，$0 \le \alpha \le 1$，o_v 是 CGN 模块的输出，o_c 是 LSTM-FC 网络的输出。

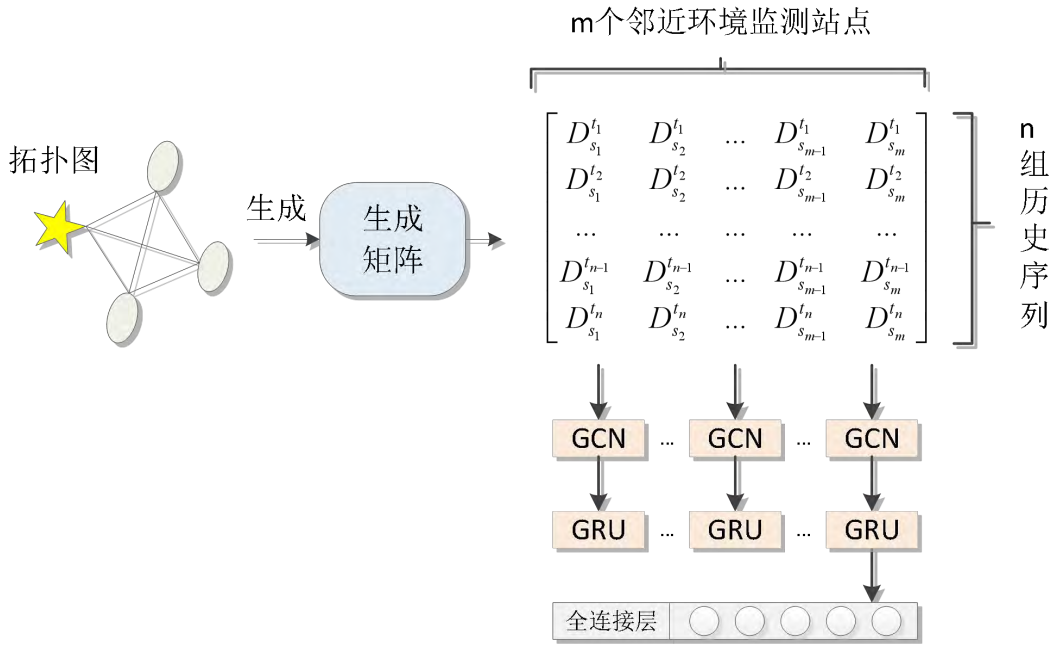


图 6. 5 用于多站点数据学习的 GCN 模块

6.4 模型求解结果

如图 6.1 所示是 STW-LSTM-FC 模型对监测点 A 污染物浓度进行预测的结果与真实值的误差曲线图，六种污染与浓度的平均预测误差为±0.2。

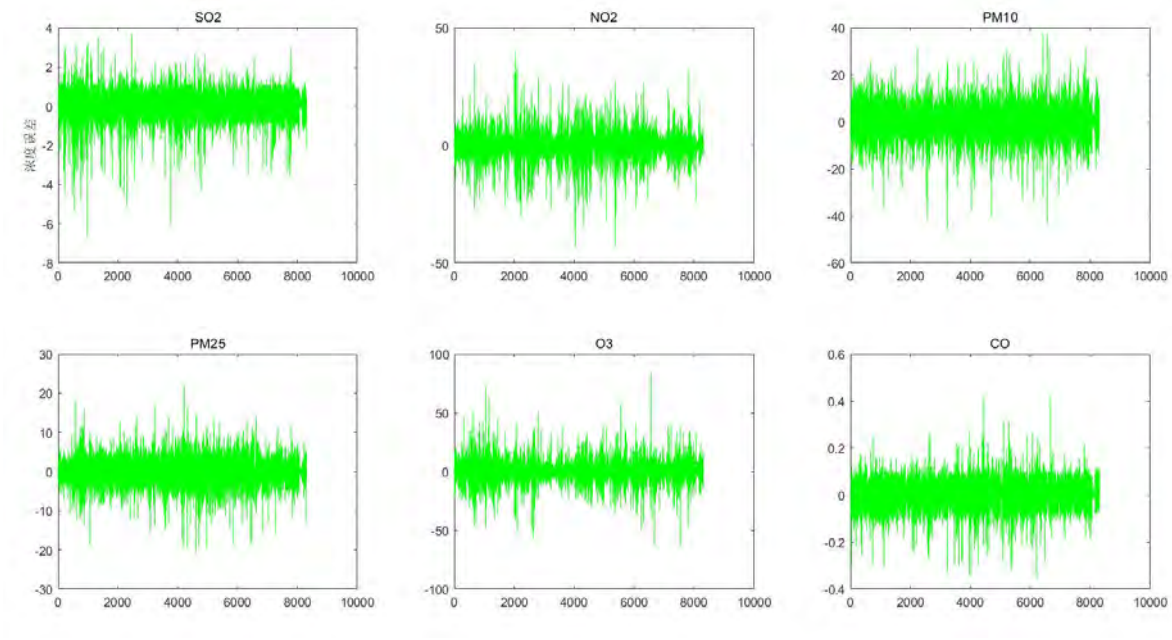


图 6.6 预测的结果与真实值的误差曲线图

如表 6.1 至 6.4 所示为四个监测点 2021/7/13 至 2021/7/15 的六项污染物预测模型及首要污染物。

表 6.1 监测点 A 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 (μg/m³)	NO_2 (μg/m³)	PM_{10} (μg/m³)	$PM_{2.5}$ (μg/m³)	O_3 (μg/m³)	CO (mg/m³)	AQI	首要污染物
2021/7/13	监测点 A	6.860	13.390	25.021	6.680	90.030	0.428	46	无
2021/7/14	监测点 A	5.765	10.038	26.169	10.131	86.790	0.483	44	无
2021/7/15	监测点 A	5.295	1.010	26.591	11.030	120.83	0.468	68	O_3

表 6.2 监测点 A1 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 (μg/m³)	NO_2 (μg/m³)	PM_{10} (μg/m³)	$PM_{2.5}$ (μg/m³)	O_3 (μg/m³)	CO (mg/m³)	AQI	首要污染物
2021/7/13	监测点 A1	8.510	12.290	29.269	11.480	92.32	0.351	47	无
2021/7/14	监测点 A1	12.740	23.441	48.318	22.340	121.38	0.511	68	O_3
2021/7/15	监测点 A1	12.789	25.060	50.571	21.781	140.09	0.579	84	O_3

表 6.3 监测点 A2 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 (μg/m³)	NO_2 (μg/m³)	PM_{10} (μg/m³)	$PM_{2.5}$ (μg/m³)	O_3 (μg/m³)	CO (mg/m³)	AQI	首要污染物

2021/7/13	监测点 A2	7.010	13.441	26.280	7.940	103.87	0.440	54	O_3
2021/7/14	监测点 A2	9.080	24.479	43.689	19.259	125.56	0.570	72	O_3
2021/7/15	监测点 A2	10.220	30.991	52.509	24.179	158.48	0.549	99	O_3

表 6.4 监测点 A3 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 (μg/m³)	NO_2 (μg/m³)	PM_{10} (μg/m³)	$PM_{2.5}$ (μg/m³)	O_3 (μg/m³)	CO (mg/m³)	AQI	首要污染物
2021/7/13	监测点 A3	4.881	8.745	15.033	5.930	90.044	0.340	46	无
2021/7/14	监测点 A3	4.681	11.324	20.100	10.101	75.634	0.449	38	无
2021/7/15	监测点 A3	4.889	10.533	24.434	13.665	109.08	0.471	58	O_3

七、 模型的总结与评价

本文基于空气质量监测相关的基础数据，对气象条件分类，并建立污染物浓度二次预测模型，建立的模型具有一定的优越性和局限性。

7.1 模型优缺点分析

7.1.1 模型的优点

本模型具有以下优点：

（1）本文利用多个专业软件进行数据的处理与问题求解，例如 MATLAB 仿真软件、MySQL 数据库、SPSS 数据分析软件、EXCEL、Python 等，经大量实验数据分析及拟合，得到的结果较为可靠；

（2）数据是模型建立的基础，对所给的附件数据均进行了细致的分析，并进行合理的预处理，包括剔除和填补缺失值、无意义值、异常值值，降低异常数据对模型的影响，使模型更稳定，可信度高；

（3）本文所使用数据来源于现实生活，在解决问题的过程中尽量结合现实情况，在建立模型的过程中，考虑到了多种气象因素对大气污染物浓度的影响，故建立的污染物预测模型很大程度上能反应实际的大气情况，对于环境空气质量监测及环境保护具有较大的参考价值，具有很好的实用性和可推广性；

（4）针对问题二提出的基于 k-means 聚类的气象分类模型原理简单、易于操作、执行效率高，得到了广泛的应用。

（5）针对问题三提出的 LMST-FC 混合模型擅长处理具有时序性和非线性的数据，对气象及污染物数据具有较强的适用性，可以防止梯度消失，能在多种污染物浓度的预测上达到较高的准确度，是一种可行的预测手段，在解决实际环境问题的过程中能够辅助决策。

（6）针对问题四提出的 STW-LSTM-FC 模型，相比问题三的 LSTM-FC 算法，考虑到气象时间与空间的耦合，能将高维度上复杂的映射关系相结合，充分挖掘相邻区

域对当前区域的污染物浓度影响特征，将具有时空属性的检测数据进行充分挖掘，通过区域协同预测有效提高了空气质量预报的准确度。

7.1.2 模型的缺点

本模型具有以下缺点：

（1）由于竞赛时间仓促，数据处理仍然不够完美，可能还有一些未考虑到的因素，具有一定误差。如果能对数据进行进一步优化预处理，得到的预测模型将会更准确、更可靠。

（2）本文建立的预测模型是在理想条件下构建的，没有考虑人口密度、人类活动、绿化程度、城市效应、以及地域差异等多种因素对污染物浓度的影响，此外，在构建模型过程中选择性的忽略了各种特征数据之间的相关性，所以，为了提高模型的实用性，此模型还需不断优化。

（3）在 LSTM-FC 及 STW-LSTM-FC 网络模型训练的过程中，会存在拟合精度不够的问题，尤其对某个污染物浓度的预测误差可能较大。此外，还可能存在样本筛选不合理的可能。

（4）O_3 作为大气污染的首要根源，对其准确预测尤为重要，本文在建立 O_3 浓度预测模型的过程中，仅考虑了一部分气象条件和氮氧化物的影响，下一步仍需考虑其它前体物如挥发性有机化合物 VOCs 和 CO 等的影响，不断完善规律，以提高浓度预报准确性。

7.2 模型的改进与推广

（1）污染物浓度的变化影响是多方面的，某一地区的空气质量状况受到多种复杂因素影响，在以后的研究中将会更深入地研究气象条件与大气污染物，尤其是臭氧的数据特性，分析气象条件与污染物、污染物与污染物之间的相互影响关系，构建更加完善的大气污染物预测模型。

（2）本文根据实际数据，提出了一种基于 LSTM-FC 网络的大气污染物预测模型，及考虑邻近区域影响的 STW-LSTM-FC 预测模型，能够高效精准预测大与污染物浓度，具有一定的学术研究价值，也在一定程度上满足工程应用的需求，具有较高的可推广性。

参考文献

[1] LOWE D G. Similarity metric learning for avariable-kernel classifier[J]. Neural Compution, 2014,7(1):72-85.

[2] 微秋凌白，数据分析之异常值分析、处理，https://www.jianshu.com/p/0deba6cdea37，2021.10.14。

[3] 若堃，城市空气质量分析预测，https://blog.csdn.net/sruokun/article/details/105489460，2021.10.15.

[4] yamgyutou，MATLAB_作图汇总(3D曲面图、热图、条形饼图折线等），https://blog.csdn.net/yamgyutou/article/details/119886234，2021.10.15.

[5] 张茹,张学杨,陆洪光,刘强. 基于层次分析和主成分分析的城市空气质量评价——以徐州市为例[J].安全与环境工程,2017,24(03):103-107.

[6] 张莉,谢亚楠,屈辰阳,汪鸣泉,常征,王茂华. 基于 K-Means 城市分类算法的夜光遥感电力消费估算[J].国土资源遥感,2020,32(04):182-189.

[7] 杜英魁,张乙芳,原忠虎,关屏,彭跃. 数据预处理对 LSTM 网络大气污染预测精度分析[J].计算机与数字工程,2021,49(07):1400-1404+1425.

[8]刘梦炀,武利娟,梁慧,段旭磊,刘尚卿,高一波. 一种高精度 LSTM-FC 大气污染物浓度预测模型[J].计算机科学,2021,48(S1):184-189.

[9] 赵彦明. 基于时空相关性的 LSTM 算法及 PM_(2.5)浓度预测应用[J].计算机应用与软件,2021,38(06):249-255+323.

[10]何哲祥,李雷. 一种基于小波变换和 LSTM 的大气污染物浓度预测模型[J].环境工程,2021,39(03):111-119.

[11]张灿,蒋昌潭,罗财红,刘姣姣,叶堤,谯捷,韩世刚. 气象因子对臭氧的影响及其在空气质量预报中的应用[J].中国环境监测,2017,33(04):221-228.

[12]王自发,王威. 区域大气污染预报预警和协同控制[J].科学与社会,2014,4(02):31-41.

[13]李丹阳. 基于时空融合的空气质量长短期预测模型研究[D].河北工程大学,2021.

[14]宋丹. 基于多源时空数据融合的空气质量预测算法研究[D].中国矿业大学,2019.

附录

程序 1：AQI 计算及首要污染物

```
'''
功能：AQI计算
'''

import math
def cal_linear(iaqi_lo, iaqi_hi, bp_lo, bp_hi, cp):...
def cal_co_iaqi(co_val):...
def cal_so2_iaqi(so2_val):...
def cal_no2_iaqi(no2_val):...
def cal_o3_iaqi(o3_val):...
def cal_pm10_iaqi(pm10_val):...
def cal_pm_iaqi(pm_val):...
def cal_aqi(param_list):...
import pandas as pd
def main():
    bys=pd.read_excel('bys_AQI.xlsx')
    list = []
    col_count = bys.shape[0]
    list_result = []
    list_first = []
    for x in range(col_count):
        param0 = bys.loc[x].values
        list = param0
        (aqi_val,iaqi_dic) = cal_aqi(list)
        list_result.append(aqi_val)
        for key in iaqi_dic:
            if iaqi_dic[key] == max(iaqi_dic.values()):
                list_first.append(key)
                break
    bys['AQI'] = list_result
    print(len(list_first))
    bys['首要污染'] = list_first
    print(bys.head())
    bys.to_excel('首要污染物_A3.xlsx')

if __name__ == '__main__':
    main()
```

程序 2：k-means 聚类

```python
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd


# 加载数据
def loadDataSet(fileName):
    data1 = pd.read_excel(fileName)
    dataset1 = np.array([(float(data1.iloc[row, 10]), float(data1.iloc[row, 4])) for row in range(data1.shape[0])])
    return dataset1


# 欧氏距离计算
def distEclud(x, y):
    return np.sqrt(np.sum((x - y) ** 2))  # 计算欧氏距离


# 为给定数据集构建一个包含K个随机质心的集合
def randCent(dataSet, k):
    m, n = dataSet.shape
    centroids = np.zeros((k, n))
    for i in range(k):
        index = int(np.random.uniform(0, m))  #
        centroids[i, :] = dataSet[index, :]
    return centroids


# k均值聚类
def KMeans(dataSet, k):
    m = np.shape(dataSet)[0]  # 行的数目
    # 第一列存样本属于哪一簇
    # 第二列存样本的到簇的中心点的误差
    clusterAssment = np.mat(np.zeros((m, 2)))
    clusterChange = True

    # 第1步 初始化centroids
    centroids = randCent(dataSet, k)
```

程序 3：LSTM-FC 模型

```python
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import r2_score

import tensorflow as tf
from tensorflow.keras import Sequential, layers, utils, losses
from tensorflow.keras.callbacks import ModelCheckpoint, TensorBoard

import warnings
warnings.filterwarnings('ignore')
dataset = pd.read_csv("o3.csv", parse_dates=['timestamp'], index_col=['timestamp'])
dataset.shape
dataset.head()
dataset.tail()
dataset.info()
dataset.describe()
plt.figure(figsize=(16,8))
sns.pointplot(x='t1', y='cnt', data=dataset)
plt.show()
plt.figure(figsize=(16,8))
sns.lineplot(x='t2', y='cnt', data=dataset)
plt.show()
plt.figure(figsize=(16,8))
sns.lineplot(x='hum', y='cnt', data=dataset)
plt.xticks([])
plt.show()
```