分类号_____　　密级 _____
UDC _____　　编号 10741

兰州财经大学
LANZHOU UNIVERSITY OF FINANCE AND ECONOMICS

硕 士 学 位 论 文

（专业学位）

论文题目 基于 EEMD-LSTM-ARIMA 的兰州市空气质量预测研究

研 究 生 姓 名：李 娜

指导教师姓名、职称：杨盛菁 教授

学 科、专 业 名 称：统计学 应用统计硕士

研 究 方 向：大数据分析

提 交 日 期：2022 年 5 月 30 日

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果.尽我所知,除了文中特别加以标注和致谢的地方外,论文中不包含其他人已经发表或撰写过的研究成果.与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意.

学位论文作者签名：李娟 签字日期：2022.5.30

导师签名：[signature] 签字日期：2022.5.30

关于论文使用授权的说明

本人完全了解学校关于保留、使用学位论文的各项规定,同意（选择“同意”/“不同意”）以下事项：

1.学校有权保留本论文的复印件和磁盘,允许论文被查阅和借阅,可以采用影印、缩印或扫描等复制手段保存、汇编学位论文；

2.学校有权将本人的学位论文提交至清华大学“中国学术期刊（光盘版）电子杂志社”用于出版和编入 CNKI《中国知识资源总库》或其他同类数据库,传播本学位论文的全部或部分内容.

学位论文作者签名：李娟 签字日期：2022.5.30

导师签名：[signature] 签字日期：2022.5.30

Lanzhou City based on EEMD-LSTM-ARIMA Air Quality Prediction Study

Candidate : Li Na

Supervisor:Yang Shengjing

摘要

随着经济的快速持续发展，大量的能源需求和过度开发，造成植被破坏，废气排放量增多等不良影响。这直接影响空气质量的优良程度，从而影响人民的身体健康。而兰州市作为群山环绕的工业城市，大气污染物消散困难，造成空气质量下降。本文利用兰州市 2014 年至 2020 年空气污染物浓度数据，探究兰州市空气质量的变化规律，分析兰州市的空气质量的特性为后续的空气质量指数预测提供条件。在对几种模型效果对比分析的基础上，选择了组合模型对 AQI 空气质量指数进行预测，该模型有效的提高了预测精度。主要结果如下：首先，将兰州市 2014 年到 2020 年的空气质量数据进行年度、季度、月度的划分，分别探究兰州市空气质量的变化规律。发现兰州市空气质量呈周期性波动，季度特征较为明显，夏季空气污染程度低，冬季污染物程度高。其次，选取各个单一模型进行对比，将对比后精度较高的单一模型投入组合模型的建立中。最后运用 LSTM、EEMD-LSTM、EEMD-LSTM-ARIMA 模型对 AQI 指数进行预测，对比各个模型的预测结果，最终得出 EEMD-LSTM-ARIMA 模型对 AQI 指数的预测最为精准，同时，选取不同的数据集验证模型的普适度，最终证明组合模型能够为空气质量预测提供相应的依据。

关键词：空气质量指数 空气质量预测 EEMD-LSTM-ARIMA；

Abstract

With the rapid and continuous economic development, the large amount of energy demand and over-exploitation has caused adverse effects such as the destruction of vegetation and increased emissions of exhaust gases. This directly affects the excellent degree of air quality and thus the health of the people. And as an industrial city surrounded by mountains, Lanzhou City has difficulties in dissipating atmospheric pollutants, which causes a decline in air quality. In this paper, we use the data of air pollutant concentration in Lanzhou City from 2014 to 2020 to explore the change pattern of air quality in Lanzhou City and analyze the characteristics of air quality in Lanzhou City to provide conditions for the subsequent air quality index prediction. Based on the comparative analysis of the effects of several models, a combined model is selected for AQI air quality index prediction, which effectively improves the prediction accuracy. The main results are as follows: firstly, the air quality data of Lanzhou City from 2014 to 2020 were divided into annual, quarterly and monthly, and the change pattern of air quality in Lanzhou City was explored separately. It was found that the air quality of Lanzhou City showed cyclical fluctuations with more obvious quarterly characteristics, low air pollution level in summer and high pollutant level in winter. Secondly, each single model was selected for comparison, and the single model with higher accuracy after comparison was put into the establishment of the combined model.

Finally, the LSTM, EEMD-LSTM, and EEMD-LSTM-ARIMA models are used to predict the AQI index, and the prediction results of each model are compared, and finally the EEMD-LSTM-ARIMA model has the most accurate prediction of the AQI index. The combined model was finally proved to provide the corresponding basis for air quality prediction.

Keywords: Air quality index; Air quality prediction; EEMD-LSTM-ARIMA;

目录

1 绪论

1.1 研究背景及意义

在干燥和洁净的大气中，微气体的组成可以忽略不计。然而，在大气层的某一区域，以前不存在的微生物出现了，其数量和存续时间会影响和损害人、动物、植物和材料。当微生物含量过高，大气结构复杂程度加大，从而构成大气污染物浓度提高，自然资源与环境、人们的生活场所和生产生活方式都受到的较大程度的影响，将此类破坏环境，危害人们身体健康的现象称作空气污染，也可以称为大气污染。空气质量主要受自然因素和人的生产生活方式的影响。自然因素包括气温的变化，空气湿度以及气压或辐射度的关系造成的空气质量波动，与自然因素相比，人为因素给空气质量带来的影响更为显著，如工厂排放废气、煤炭燃烧、汽车尾气等。随着城市的现代化建设，人们的生产生活方式发生转变，需求增多消耗增多，为了更多更好的满足人们的精神物质需求，会加大对于资源的利用从而排放出大量的废气，燃烧更多的燃料，使得大气环境每况愈下，尤其是以重工业为主的城市以及人流量较多的地区。空气污染等问题已经引起了各国的重视，因为空气质量直接影响到该区域内人们的身体健康。如果空气遭受了污染，那么通过呼吸这种空气，人们的身体健康将会受到严重威胁。例如由于人们取暖所燃烧的燃料生成的气体长期存在于人们的生活区域，导致空气中的硫氧化物浓度增加，在空气中发生一定的化学反应生成硫酸凝结在空气中，这些污染物一旦进入人体，会对人们的呼吸道产生危害，进而危害到人体的生命安全。当然，空气污染不仅仅只会使人患有慢性病，还有其他的多种危害，主要表现为呼吸系统疾病和生理功能紊乱，以及刺激眼、鼻粘膜组织等疾病。因此，为了将空气质量量化，找到其变化规律并有效防治，推出了空气质量指数（AQI），作为衡量生活质量的一个指标。

中国作为一个发展中国家，发展中国家的一个通病就是会在一段时间内快速发展，将产业发展到一个水平后意识到快速发展带来的一些弊端，破坏生态资源，造成了大气环境负担，在经济增长的背后，看不清为了经济增长所付出的环境和资源的代价。改善空气质量举措深度契合了“绿水青山就是金山银”理念，同时为

我国绿色发展提供了坚实基础，如果单纯的将 GDP 增长量作为衡量政绩的标准，可能确实会推动社会的发展，但很有可能会造成不顾生态环境，以透支未来为代价造成经济增长的假象。调查显示，为了弥补快速经济增长所带来的危害，每年要花费大量的资金在医疗保健和支持可持续发展等领域。中国天气状况的持续恶化已成为对公众健康和可持续经济发展的严重威胁。

而兰州市一直以来都是以重工业为主的生产发展方式，且兰州市地势狭长，四面环山，不利于空气污染物的扩散，导致长期面临较差的空气质量，人们出行不便，也会对人们的呼吸道等方面产生健康威胁。且冬末春初浮沉，沙尘更是困扰兰州市空气质量的巨大难题。曾一度被冠以“黑帽子”，体现出过去兰州市空气质量确实较差，但在 2013 年兰州市经过一系列的措施，产业升级、增大绿色覆盖面积、能源多元化、大力发展新能源，有效的改善了空气质量。但这并不意味着空气质量一直保持在高质量阶段，目前仍然存在冬季春季空气质量较差的问题，本文旨在针对兰州市空气质量指数做出预测，从而找出预测精度较高的模型，为恶劣天气预警提供科学依据，保障人们的健康出行，也为后续政府针对空气质量出台更有针对性的扶持政策与应急响应机制提供坚实基础。为了改善中国逐渐恶化的天气状况，政府相继出台了许多相关政策并采取有效措施，确实对改善中国的空气质量起到了积极作用。然而，随着中国人口和车辆数量的增加，供暖和建筑工地的粉尘等污染源在短期内无法完全消除，中国的空气污染仍然是一个问题。

空气污染质量既复杂又具有挑战性，不仅需要政府的大力支持，社会各界要积极响应相关政策，从自身做起，积极的参与到绿色发展的事业中来，采取各种手段各种措施保护环境。而对空气质量的科学预测，应积极开展对大气状况的预警等活动，让政府及时采取预防措施，避免损害公众健康。随着经济的不断发展，脱贫攻坚的完美收官，使得我国居民的生活水平和生活质量逐渐提高，对于人居环境，出行便利，身体健康更加关注。环保意识已深入根治到人们心中，空气质量的好坏也成为了他们日渐关注的话题，应向他们提供有价值的信息，供他们决策，空气质量预报将为此做出巨大贡献。

目前，中国已经建立了比较科学的空气质量检测体系，但目前的空气质量监测体系还有需要完善的地方。针对六种主要的大气污染物收集地面观测数据 ，再将收集到的数据处理后计算出监测数据的日平均浓度，空气质量指标计算出当

天的具体空气质量指数，用于评估当天的空气质量状况。这表明，现有监测系统的实时监测具有一定的滞后性，无法提前预测空气污染状况，因此无法向政府提供可靠的预警信息。传统的空气质量预测方法大多是统计回归模型、数值预测模型和浅层机器学习算法，这些方法使用方便，但仍存在一些误差，不能合理利用海量的数据。深度学习方法可以弥补这些缺陷，并利用其自身强大的学习能力来预测空气质量，效果更好。本文介绍了空气质量预测中的深度学习算法，寻找高效准确的预测模型，同时将传统模型优化，将更适合的预测工具投入实际应用，为相关部门提供科学依据，确保人们的日常出行得到保障。

1.2 国内外研究现状

1.2.1 空气质量预测的国内外现状

空气质量预测可精准的为群众提供空气质量状况的参考，方便了人们出行，也让政府可以采取相关措施及时应对恶劣天气。国外发达国家早在 20 世纪 70 年代就建立起适合自己国家的空气质量检测系统。我国空气质量预测起步较晚，直至 2016 年才建立起初步的检测系统。空气质量预测方法主要分为：潜势预测，数值预测和统计预测。

（1）空气质量潜势预测法：

潜势预测在对空气质量进行预测时只考虑天气情况和气象参数，不考虑污染物浓度的影响。预测的准确性与天气预报的准确性有关，当气象条件超过气象质量的标准线时，就可以判定为空气污染天气。潜势预报最早发源于美国，随着技术的发展与进步，潜势预报也逐渐被数字和统计预报所取代。中国对于潜势预报的研究较西方发达国家晚，但随着我国的科技的发展以及综合性人才的培养，我们自己的科学家也在对空气质量的预测中有所成就，快速比肩外国预测手段直至实现弯道超车。尤莉等人（2003）选择气象参数，如天气状况、风和反转温度，构建了一个潜势预测模型，对内蒙古省会呼和浩特的空气质量预测[1]。赵慧芳（2009）等将气象因素引入空气质量预测模型中，建立了多影响因素的预测模型，玩车个了对晋江市空气质量的预测，并将分类法与趋势推断法相结合，实现了较好的预测效果[2]。

潜势预测法只能根据天气情况进行预测，参照条件过于单一，导致在预测精度上不高，常常错误的将无污染地区判定为高污染，预测结果产生较大误差。由于潜势预测只能将空气质量定性，无法给出定量的具体数值，参考价值有限，随着技术不断地更迭，逐渐的被取代。

（2）空气质量数值预测法

数值预测法所运用的原理是大气动力学，是在探究大气规律的过程中运用的物理与化学相融合的方式所确定空气质量的一种方法。通过对大气污染物的传输分布进行建模，用计算机数值方法求方程解，通过最终求解的数值来确定空气污染物在时间上的运动规律。

缪启龙等（2006）针对银川市空气质量建立动力预测模型，该系统结果显示，可以对大气污染物浓度进行 24 小时的预测[3]。Jose 等人（2008）运用 MM5-CMAQ-EMIMO 模型，打造欧洲地区空气质量预测系统[4]。Lee 等（2011）为提高模型的预测精度，将 CMAQ 模型进行改进，在原始数据中加入了化学条件，得到了良好的预测结果[5]。沈进等（2011）对珠江三角洲的臭氧浓度进行预测，采用 Models-3/CMAQ 和 CAMx 模型相结合的方式预测了较为精准的臭氧浓度[6] 。谢敏等(2012)将前 24 小时的监测数据作为验证集，利用改进的 CMAQ 方法预测原始浓度值，考虑了 CMAQ 模型预测的浓度变化趋势，最终得出了修正后的预测结果[7]。

虽然数值预测方法取得了一些进展，但数值预报方法仍然存在许多问题。如设备花费大、成本高，需要获取专业的气象知识，需要的预先假设较多，且常以主观经验判断模型参数，存在不可避免的误差等。为了获得更方便快捷，且人力成本低的预测方法，统计学的各界学者们发挥他们的专业，弥补了这一方面的空白，以统计学为基础的预测方法应运而生。

（3）空气质量统计预测法：

基于统计学的预测模型统称为统计预测法，它以特定地区和天气状况的历史数据积累为基础。统计预测方法不以空气污染物的物理和化学变化为基础，只以历史信息为基础的空气质量数据的线性和非线性变化为基础，使用起来便捷且降低成本。而且它不需要专业的气象知识，具有普遍适用性。且不需要特别专业的气象知识，具有普遍适用性 。

Liu（2015）对 ARIMA、BP 和指数平滑模型进行组合，根据熵权重法确定各个模型的权重，将组合成的模型运用到实际预测中，与单一模型比较，结果发现组合模型的精度好于单一模型[8]。Yang 和 Jian（2017）利用信号分解即差分进化改进 Elman 神经网络模型，经过原始信号的分解，消除数据噪声，大大提升了模型的精准度[9]。倪志伟等（2016）在传统 SVM 模型的基础上，对原有模型采取人工鱼群方法优化，运用三个城市的数据集验证，结果显示较传统的 SVM 模型预测效果较好，具有较高的稳定性和可信性[10]。彭艺（2020）将 SVM 模型的预测结果与其他单一模型进行对比，发现经典模型的预测误差不理想，提出了 ARIMA-SVM 组合预测方法，最终验证此方法的效果更好[11]。几位专家学者对统计预测方法的研究增加了统计预测方法的可能性，提高了空气质量预测的精度，为后续的空气质量提供了坚实的基础。

总之，经济发展总会带来一定的代价，大气结构复杂，任何的经济活动都有可能造成空气质量的波动，随着人们对生活水平的高要求，更加注重生活环境水平，因此愈加严重的空气质量问题逐渐得到重视。各界学者也在自己的领域针对空气质量问题展开研究，但通过梳理各位学者的研究发现目前的空气质量预测方法仍存在一定程度的缺点，所预测出的数据与原始数据仍存在一定差距。在上述文献中发现统计预测方法是更方便预测效果也较其他方法更优，在统计预测方法中属组合模型效果最好，故本文选取统计预测方法中的组合模型探究兰州市空气质量的预测。

1.2.2 模态分解的国内外研究现状

同一现象不同时间发生的事物活动数据的变化，将数据记录下来就成为时间序列数据。而气象数据大多都为时序数据，空气质量数据也是如此，时间序列数据反映了空气质量的一系列变化规律。时间序列数据在收集时由于事物活动并不统一，所以可能产生随机波动，这类数据为平稳数据；也有可能是有一定趋势的数据，这类数据则是非平稳数据。非平稳数据中又包含趋势、季节性、周期性或随机噪声等数据成分。所以对时间序列数据进行分级，有利于探索数据序列中的各种成分，从而清晰的判别数据中所蕴涵的规律与特性。在预测过程中，将数据分解后的各个分信号，进行数据重构，利用不同的预测方法预测适合的数据，将

有效提高预测精度。

（1）小波分解

小波分解已经被运用到各种各领域。Osowski 等（2007）在对空气污染物浓度预测中，采取小波分解，利用其多分辨率分析法，建立小波分解与支持向量回归模型的组合，结果显示，组合模型的预测误差小于支持向量回归的预测误差[12]。刘向丽等（2015）通过对时间序列数据进行小波分析高低频数据的分解，分层逼近建立 ARIMA 模型从而以股权期货数据验证模型的可靠性[13]。代军（2021）提出集合经验模态分解与小波分解相结合的方式对单一降噪方式进行改进，将改进模型实际应用得出较好的成果[14]。王振华等（2021）利用小波分解改善了 GLAS 获得的全波数据噪声复杂的问题，切实提高了垂直结构数据的精准度[15]。Singla（2021）利用小波分解将太阳辐照度原始数据进行分解，将不同的信号投入到 BiLSTM 模型中，对太阳辐照度前 24 小时数据进行预测[16]。马宁等（2022）由于传统风电功率模型波动较大，预测精度不稳定，提出了基于经验小波变换预测模型，将分解后各模态信号分别预测再汇总，最终得到较高精度的风电功率预测值[17]。

上述学者的研究领域中包括风电功率、股票和太阳辐射。可见，小波分解在去除噪声，多尺度分解上虽有比较广泛的应用，但小波基函数有多种选择，并不具备唯一性,因此小波分析在实际应用过程中对于最佳小波基函数的选取具有一定困难，很有可能因为选取不当反而失去了其特性，因此在实际操作过程中还需要多加选取小波基函数。

（2）EMD

经验模态分解最早在 1998 年，由于其能够分解出数据中非平稳信号而被逐步推广。董小刚等（2016）预测$PM_{2.5}$发展趋势，采取 EMD 算法，对其未来发展趋势进行一定分析[18]。Jian feng Zhang 等（2018）在西部旱区灌区 14a 数据预测中，采用 EMD-LSTM 模型，对比五个子区数据，反映出，该模型具有较强的学习能力，能够有效提高预测精度[19]。刘铭等（2020）在对天津市空气质量预测时，选取$PM_{2.5}$浓度指标，采用 EMD-LSTM 组合模型提高预测精度，得到较为精准的结果[20]。涂锦等（2020）针对时间序列的非线性数据预测模型进行改进，由于非线性数据中存在大量的噪声各高度的波动性，因此，首先将数据进行经验模态分

解，再将分解后数据利用 ANN 模型进行预测，有效提高了非线性数据的预测效果[21]。金秀章等（2021）针对煤电厂脱硫系统出口SO_2浓度不稳定的问题，提出基于 EMD-LSTM 组合预测模型，经 EMD 分解后作为输入数据，经 LSTM 模型训练，最终得出稳定的预测值，降低传统模型预测误差[22]。

综上所述，经验模态分解在原始数据上以自适应的方式分解，免去了小波分解在选取小波基函数的操作，在实际运用中更方便快捷，但其也不是最佳的数据分解方法，EMD 存在模态混叠现象，在经验模分解的过程中，由于局部极值在短期间隔内发生多次突变，从而使信号之间区分边界不清，容易出现混乱的现象，而集成经验模态分解解决了这一棘手问题。

（3）EEMD

EEMD 基于 EMD 方法发展而来的，针对 EMD 方法的不足，提出加入随机噪声，使其均匀的分布在整个时频空间内，使得最终分解的信号持久稳固。秦喜文等（2016）针对北京市空气质量进行深入探究，选取$PM_{2.5}$污染物浓度指标，利用 EEMD-SVR 组合模型，预测其未来的波动趋势[23]。Yang 等（2017）为构建空气质量监控预警平台，采取 EEMD-Elman 组合方法，凸显出混合模型稳定优质效果[24]。许德合（2021）在对新疆维吾尔自治区干旱趋势预测中，采取多尺度降水数据，利用 EEMD-ARIMA 组合模型，将数据分解并投入到时间序列模型中分别预测，最终将预测结果汇总，对比其预测效果发现，组合模型高于单一模型预测准确度[25]。史学良等（2021）针对传统空气质量预测模型进行调整，由于在预测空气质量指数时，大气中含有多种污染物，使得 AQI 预测更为复杂，为解决时间序列数据中存在的波动问题，提出利用 EEMD 对数据进行信号分解，再利用改进后的 LSTM 模型进行预测，比传统模型泛化能力强，预测精度高[26]。

总之，对于时间序列的分解模型中，集成经验模态分解模型拥有较好的可操作性和较好的分解效果，通过合理的信号分解消除数据中存在的噪声和波动性，从而为后续的预测提供基础保障，而且在对比各位学者的研究中发现，集成经验模态分解方法经常与深度学习方法组合使用效果最好。因此，本文选取集成经验模态分解方法作为组合模型中的一种较为合适。

1.3 研究方法与思路

本文根据兰州市自 2014 年至 2020 年空气污染物数据对兰州市空气质量的演变规律进行分析探究。同时针对兰州市主要空气污染物及空气质量指数进行预测，对比于其他模型，找出预测精度最高的模型，为今后的有关预测提供高质量的预测工具。

为研究兰州市空气质量变化规律，将数据按年度、季度、月度进行划分，画出一系列统计图，以直观的显现出兰州市空气污染物浓度变化规律。同时，通过构建热力图从统计意义上探究兰州市各项空气污染物与 AQI 空气质量指数的相关性和强度，从而得出它们之间的影响规律。运用随机森林的特征选择，选取对 AQI 空气质量影响程度较大的因素，将重要程度大于 0.1 的因素作为 AQI 空气质量指数回归方程的自变量。找到影响空气质量指数程度较高的指标为下一步验证模型的普适性做准备。

本文主要运用长短期记忆网络（LSTM）模型，对空气质量指数进行预测，同时由于空气质量数据为非平稳数据，存在噪声和波动，需要进行数据预处理首先对数据进行 EEMD 分解，分别将高频数据及低频数据输入到特点不同的模型中训练并预测，将预测结果线性加总成为最终预测结果。为后续预测提供更适合的数据。在传统预测方法的基础上，本文增设滑动窗口，对空气质量进行预测，提高预测的精准度。并与单一模型进行对比，验证模型的可靠度。

1.4 研究内容

第一章是绪论，介绍了本文研究的背景及意义，国内外文献综述包括对时间序列预测模型和数据分解模型的了解，研究问题和思路，以及研究内容。

第二章是相关概念和理论方法，主要介绍了本文所涉及的空气质量指数的相关原理，AQI 计算方法，其分级标准和范围以及模型的原理及模型建立步骤，介绍长短期记忆网络（LSTM）、随机森林（RF）、支持向量回归（SVR）、集成经验模态分解（EEMD）、时间序列模型（ARIMA）以及相关的概念和理论。

第三章探究概括了兰州市空气质量的时空特征及演变规律，了解目前兰州市的空气质量情况，将数据的来源和预处理简单说明。探究兰州市各污染物浓度与

AQI 空气质量指数直接的相关性及变化趋势，为后续的预测提供基础保障。

第四章是模型的选择，选取随机森林、长短期记忆网络、支持向量回归模型对 AQI 指数预测，对比预测效果，为组合模型提供依据。

第五章基于上述模型的概念和原理，建立 LSTM 模型，并针对数据进行集成经验模态分解，分解出高频项、低频项和趋势项，更好的进入到训练阶段，预测出更精准的结果。同时，为了验证模型的普适性，增加对$PM_{2.5}$和PM_{10}数据集的验证，对训练好的模型投入实际运行，预测未来 7 天的空气质量数据。

1.5 可能的创新点

本文主要探索空气质量的预测模型，对比三种经典深度学习模型的预测效果，选取其中精度较高的模型作为最终的预测工具，选定 LSTM 模型作为最终的组合模型，对 LSTM 模型进行优化，建立滑动窗口，并不断调整参数，令数据在训练过程中有效的记忆，使得数据拟合更好，预测结果更精确。而后将集成经验模态分解后的数据，根据数据特点选取不同类型的模型进行重构集成，将各部分数据输入到适合其变化的模型中训练，得到更贴合真实的数据，减小预测数据的误差，为今后的空气质量预测提供可靠依据。

2 相关概念和理论方法

2.1 空气质量指数概念

衡量空气质量的指标为空气质量指数，具体数值取决于污染物浓度。空气污染最主要的来源是人为和自然的气象条件的活动变化，使得空气质量波动，包括汽车、船舶和飞机排放、工业污染、住宅和供暖排放以及垃圾焚烧。除此之外城市人口密度、地势和天气变化也成为了空气质量变化的主要因素。因此，空气质量指数通过使用函数关系将每种污染物转换为单一的数字形式，通过比较各个污染物所占的空气质量分指数来确定。所有污染物的最高 IAQI 值是 AQI。因此，空气质量指数的确定准确反映的空气质量的范围和被污染程度。如果 AQI 值大于 50，则该项污染物则被人文是对空气质量影响较大，即为主要污染物。

2.2 空气质量指数的计算方法

计算过程可分为以下几步：

第一步：是将分级浓度进行对比，找出相对应的浓度限值，以 API 的浓度限值为标准，以细颗粒物、硫氧化物、氮氧化物、臭氧、一氧化碳等各空气污染物的含量计算各各指标的 IAQI 指数：

$$IAQI_p = \frac{IAQI_{Hi} - IAQI_{Lo}}{BP_{Hi} - BP_{Lo}}\left(C_p - BP_{Lo}\right) + IAQI_{Lo} \tag{2.1}$$

式中：

$IAQI_p$——污染物 P 的空气质量分指数；

C_p——污染物 P 的质量浓度值；

BP_{Hi}——污染物 P 的浓度上线

BP_{Lo}——污染物 P 的浓度下线

第二步：通过第一步的计算，计算出每个污染物浓度的 IAQI 最大值为 AQI，当 AQI 大于 50 时将对应的污染物确定为首要污染物：

$$AQI = max\{IAQI_1, IAQI_2, IAQI_3, \ldots, IAQI_n\} \tag{2.2}$$

式中：

IAQI——空气质量分指数；

N——污染物种类。

第三步：将最终的 AQI 指数与分级标准对照。通过对照确定当下的空气质量级别和污染程度，为精准施策提供依据。

2.3 空气质量指数等级划分

空气质量指数（AQI）常被用来衡量空气质量的好坏，是评价空气质量的定量标准。随着科技的发展，AQI 空气质量指数的发布也基本可以实现实时监测，过去一天才能发布一次，现在可以通过手机软件随时观看，方便了人们对空气质量的评判也方便了人们的出行。在我国，过去所用的空气质量指数与 AQI 原理相差无几。AQI 的空气质量划分标准是世界通用的，都以 50 为一级，依次叠加。我国使用不同颜色来区分空气质量等级，依次为：优、良、轻度污染、中度污染、重度污染和高度污染。事实上，其他国家也用大致相同的颜色来表示这六个级别。

表 2.1 AQI 等级表

AQI	级别	颜色标识
0-50	优	绿色
51-100	良	黄色
101-150	轻度污染	橙色
151-200	中度污染	红色
201-300	重度污染	紫红色
>300	严重污染	褐红色

2.4 时间序列分解算法

2.4.1 经验模态分解

EMD 最早是由 NE. Huang 等人提出的，EMD 是将信号分为不同模态的方法。EMD 相比与小波分解的优点是它不用选择定基函数，它是根据被分析的信号自适应地生成自己的模式函数。EMD 分解基于几个假设：1.信号至少有两个离群值，一个极大值和一个最小值；2.时间尺度属性由两个离群值之间的时间尺度决定。EMD 分解的目的是将信号 f(t)分解为 N 个内部模式 EMD 分解的目的是将原始信号分解为 N 个内分信号（IMF）和一个趋势项。每个分量信号（IMF）

应同时满足如下条件：1）在原始数据内，局部极值和跨零点的数量必须等于或最多相差一个；2）局部最大值的包络（上层包络）和局部最小值的包络（下层包络）的平均值必须在任何特定时间为零。

EMD 分解步骤如下：

在第一阶段，确定信号的所有局部最大和最小点，然后使用曲线拟合技术构建上下包络，将各个最大值结合起来，使原始信号包络在上下包络中。

在第二步中，可以从上下包络中构建它们的平均 m(t)曲线，并将平均曲线从原始信号 f(t)中提出，这样得到的就是分量信号（IMF）。

在第三步，信号经过前两个步骤后可能存在不满足条件的情况，因此需要重复上述步骤，直到某一时刻的 SD 值小于阈值，可将该信号输出，这样的到的第一个信号就是 IMF1，SD 的计算公式如下：

$$SD = \sum_{t=0}^{r} |H_{k-1}(t) - H_k(t)|^2 \tag{2.3}$$

第四步：残差$r(t) = f(t) - H(t)$，重复第一、二、三步，直到 r(t)满足预先设定的条件。

2.4.2 集成经验模态分解

EEMD 的原理相对简单：信号的偏振点会影响 IMF，如果分布不均匀，就会出现模态混叠。为了解决这一现象，将白噪声逐步加到原始信号的分析中，让白噪声在数据刻度上发挥自身作用。使得白噪声得以均匀分布。在经过几次平均计算后，噪声就会相互抵消，所以可以直接看到集成平均的结果是最终结果。随着集成平均数的增加，集成平均数和原始信号之间的差异也会减少。

EEMD 分解步骤如下:

第一步：设定总体平均次数 M;

第二部：将一个具有标准正态分布的白噪声$n_i(t)$加到原始信号 x(t)上，以产生一个新的信号：

$$x_i(t) = x(t) + n_i(t) \tag{2.4}$$

式中：$n_i(t)$表示第 i 次加白噪声序列，$x_i(t)$表示第 i 次试验的附加白噪声信号，$i = 1,2,3 \dots M$

第三步：对所得含噪声的信号$x_i(t)$分别进行 EMD 分解，得到各自 IMF 和的

形式：

$$x_i(t) = \sum_{j=1}^{J} c_{i,j}(t) + r_{i,j}(t) \tag{2.5}$$

式中：$c_{i,j}(t)$为加入白噪声后分解得到的 IMF，$r_{i,j}(t)$是残余函数，代表信号的趋势项，J 是 IMF 的数量；

第四步：重复上述步骤，将每次分解后的信号加入白噪声得到 IMF 的集合：

$$c_{1,j}(t), c_{2,j}(t), \ldots c_{M,j}(t), j = 1,2, \ldots J \tag{2.6}$$

第五步：将得到的 IMF 的集合计算平均值，得到 EEMD 分解后最终的 IMF，即：

$$c_j(t) = \frac{1}{M}\sum_{i=1}^{M} c_{i,j}(t) \tag{2.7}$$

式中: $c_j(t)$是 EEMD 分解的第 j 个 IMF，i=1,2,…,M，j=1,2,…,J。

2.5 LSTM 长短期记忆网络方法

2.5.1 RNN

RNN 即循环神经网络，RNN 能够作用于连续的时间序列数据,且数据的长度和尺度可任意变换。人工神经网络只能建立层与层之间的连接，但细化的层面无法进行处理，但 RNN 可连接层与层之间的神经元，是人工神经网络的进阶，能够更好的捕捉数据间的波动。

RNN 在输入阶段，除了输入层的X_t之外，还有一个边缘循环提供上一时刻的隐层状态S_t。在任意时刻，每一个神经元读取了输入信息和上一时刻的隐藏层之后会产生新的隐藏状态S_t,从而输出O_t。RNN 当前的状态是由上一时刻的状态S_{t-1}和当前的输入X_t共同决定的。对于一个序列数据，将数据输入 RNN 网络模型后，RNN 网络中输入层和循环层的结构开始运转，可处理当前的输入信息，也可以是对下一时刻数据的预测。RNN 网络结构是每一时刻都有数据输入，但不一定会将数据输出，有可能作为上一时刻的信息状态传入下一层中。

前向传播：

$$S_t = f(U * X_t + W * S_{t-1}) \tag{2.8}$$

$$o_t = gV_{st} \tag{2.9}$$

其中 t 代表时间步长，t 时刻隐藏层的值S_t取决于本时刻输入值X_t和上一时刻输出值S_{t-1}的共同作用。其中 f 和 g 均为激活函数，f 激活函数有 tanh，Relu 和 sigmoid 函数，g 激活函数通常为 softmax。

反相传播：

由于 RNN 中每一步的输出不仅与当前的输入数据有关，还受到前一步隐藏层状态的影响，因此，这种将输出的数据结果反向计算误差进行传递的算法被称为反向传播算法。再根据权重的变化输入。每一次的输出值o_t都会产生一个误差值e_t则总的误差可以表示为:

$$\mathrm{E} = \sum_t e_t \tag{2.10}$$

$$\nabla \mathrm{U} = \frac{\partial E}{\partial U} = \sum_t \frac{\partial e_t}{\partial U} \tag{2.11}$$

$$\nabla \mathrm{V} = \frac{\partial E}{\partial V} = \sum_t \frac{\partial e_t}{\partial V} \tag{2.12}$$

$$\nabla \mathrm{W} = \frac{\partial E}{\partial W} = \sum_t \frac{\partial e_t}{\partial W} \tag{2.13}$$

综上所述，循环神经网络 RNN 操作的主要步骤如下：

（1）将每个网络节点的输出值正向传播。

（2）再将数据计算误差后的结果反向传播。

（3）计算梯度值，利用随机梯度下降算法对网络权重值进行更新。

2.5.2 长短期记忆网络 LSTM

LSTM 是一种时间上的递归神经网络，能够准确抓取相对较长时间序列数据中的波动规律，从而处理和预测由于数据波动形成的重点事件。LSTM 和 RNN 的主要区别在于，它在算法中加入了一个 "处理器"，通过这个处理器决定数据信息的去留，相当于神经网络中的神经元。长短期记忆网络结构与传统的神经网络相比增加了三个门，称为输入门、遗忘门和输出门。当细胞进入到 LSTM 框架中，如果被规则认为是有用信息，则会传输到下一步骤，若被认定为无用则会被遗忘门舍弃。

（1）遗忘门

在 LSTM 中的遗忘门是整个预测过程中的第一步，遗忘门在信息输入的过程中从前一个状态h_{t-1}和当前输入的信息x_t中选择可留下的部分，最终输出一个在 0 到 1 的数值给每个在细胞状态C_{t-1}中的数字。1 表示“完全保留”，0 表示“完全舍弃”。

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2.14}$$

其中h_{t-1}表示的是上一个 cell 的输出，x_t表示的是当前细胞的输入。σ表示 sigmod 函数。

（2）输入门

通过遗忘门处理后的信息传送到输入门，输入门可决定所有留下的细胞中有哪些信息可以继续作为新信息传入到新的细胞状态中。在输入门的操作过程中需要首先经过 “input gate layer ”的 sigmoid 层决定哪些信息需要更新；在将信息处理后，经过一个 tanh 层生成一个向量，也就是激活层备选的用来更新的内容，$\widetilde{C_t}$在下一步，把这两部分联合起来，对 cell 的状态进行一个更新。

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2.15}$$

$$\widetilde{C_t} = \tanh\,(W_c \cdot [h_{t-1}, x_t] + b_c \tag{2.16}$$

在细胞更新中，我们把C_{t-1}与f_t相乘，将需要遗忘的无用信息丢弃。接着加上$i_t * \widetilde{C_t}$这就是新的候选值，重复操作，不断更新细胞中的信息。。

（3）输出门

最终，输出门将决定最终模型会输出的结果。根据前两部的操作选择细胞胞，将细胞中的信息有选择性输出。首先，通过 sigmoid 层来确定细胞中哪些信息可以作为输出信息。其次，通过 tanh 进行处理并将它和 sigmoid 门的输出相乘，最终仅仅会输出确定输出的那部分。

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{2.17}$$

$$h_t = o_t * \tanh\,(C_t) \tag{2.18}$$

2.5.3 激活函数

激活函数在神经网络模型中发挥重要作用，它可以将非线性因素引入网络中。通过神经元中的加权求和，应用于函数。激活函数是为了在神经层中加入非线性

因素，如果没有激活函数，即使有很多层，每一层也等于一个矩阵的乘法。一般的激活函数如下：

（1）Sigmoid 函数

Sigmoid 函数是生物学中经常遇到的 S 形函数，也被称为 S 形增长曲线。在计算机科学中，由于其独特的增长特性以及独特的反增长特性，Sigmoid 函数经常被用作神经网络的阈值函数，在 0,1 之间映射变量。其公式如下:

$$f(x) = \frac{1}{1+e^{-x}} \tag{2.19}$$

（2）Tanh 函数

Tanh 是双曲函数中的一个，Tanh 双曲正切。在数学中，双曲正切"Tanh"是由基本双曲函数双曲正弦和双曲余弦推导而来。公式如下：

$$f(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{2.20}$$

（3）Relu 函数

Relu 激活函数(The Rectified Linear Unit)，用于隐层神经元输出。公式如下：

$$f(x) = max\ (0, x) \tag{2.21}$$

2.6 随机森林

随机森林是 bagging 算法的演变，并做了一些小的修改。我们知道，bagging 是一种随机抽样方法，它从原始数据集中抽取 m 个子样本，用这些 m 个子样本来训练 m 个基础学习者，从而减少模型的方差。随机森林方法通过 bagging 或整合的思想来避免过度拟合，这实际上相当于同时对样本和特征进行采样。算法步骤如下：

输入：训练数据集$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$，样本子集个数 T

输出：最终决策树的均值

（1）对 t=1,2,…,T 从样本中随机抽取 m 个点作为样本点，将样本点组合为训练集 D（t）

（2）从训练集中重复抽取数据用来训练决策树，训练决策树需要从这些特征中选择一些特征，将这些特征中选择最佳的截断点，再做左右子树的分割。

（3）随机森林的结果有两大类，一类是分类，另一类为预测。如果用于分类，则在最终的预测树中，哪一类的投票数最多，则为最终的类别。如果用于预测，则将所有决策数的均值作为最终的预测结果。

2.7 SVR 相关理论

(一)支持向量回归原理

SVM 算法非常高效：支持线性和非线性的分类和回归。其主要思想是将目标倒置：在分类问题中，是将最大的范围区分两个类别，限制违反范围；在 SVM 回归中；在 SVM 回归中，试图在限制违反区间的情况下拟合尽可能多的数据。由超参数 ϵ 控制。

在一般的回归问题中，给定训练样本D $= \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}, y_i \in R$,我们想学习尽可能接近 y 的 f(x)，其中 w 和 b 是待定的参数。在这个模型中，如果 f(x)和 y 相等，则损失为零，在支持向量回归中，我们在计算损失之前最多允许 f(x)和 y 之间有 ϵ，这相当于形成一个以 f(x)为中心的宽度为 2ϵ 的区间带，如果训练样本位于这个区间带中，则认为它们被正确预测了。因此，SVR 问题可以转化为：

$$\min_{w,b,\xi_i,\widehat{\xi_\iota}} \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\left(\xi_i + \widehat{\xi_\iota}\right) \qquad (2.22)$$

$$\text{s.t.} \quad f(x_i) - y_i \le \epsilon + \xi_i$$

$$y_i - f(x_i) \le \epsilon + \widehat{\xi_\iota}$$

$$\xi_i \ge 0, \widehat{\xi_\iota}, \ge 0, \mathrm{i} = 1,2,, \ldots, \mathrm{m}$$

引入拉格朗日乘子，可得拉格朗日函数：

$$\mathrm{L}\left(\mathrm{w}, \mathrm{b}, \alpha, \hat{\alpha}, , \xi, \hat{\xi}, \mu, \hat{\mu}\right)$$

$$= \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\left(\xi_i + \widehat{\xi_\iota}\right) - \sum_{i=1}^{m}\mu_i\,\xi_i - \sum_{i=1}^{m}\widehat{\mu_\iota}\widehat{\xi_\iota} \qquad (2.23)$$

$$+ \sum_{i=1}^{m}\alpha_i(f(x_i) - y_i - \epsilon - \widehat{\xi_\iota},$$

对四个遍历求偏导，令偏导数为零，可得：

$$w_i = \sum_{i=1}^{m}(\widehat{\alpha_\iota} - \alpha_i)x_i,$$

$$0 = \sum_{i=1}^{m}(\hat{\alpha}_i - \alpha_i) \tag{2.24}$$

$$C = \alpha_i + \mu_i$$

$$C = \hat{\alpha}_i + \hat{\mu}_i$$

将上式带入，即可得 SVR 对偶问题：

$$\max_{\alpha,\ \hat{\alpha}} \sum_{i=1}^{m} y_i\,(\hat{\alpha}_i - \alpha_i) - \epsilon(\hat{\alpha}_i + \alpha_i) - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\hat{\alpha}_i - \alpha_i)\,(\hat{\alpha}_j - \alpha_j)x_i^T x_j$$

$$\text{s.t. } \sum_{i=1}^{m}(\hat{\alpha}_i - \alpha_i) = 0 \tag{2.25}$$

$$0 \le \alpha_i, \hat{\alpha}_i \le C$$

上述过程需要满足 KKT 条件，即

$$\begin{cases} \alpha_i(f(x_i) - y_i - \epsilon - \xi_i) = 0 \\ \hat{\alpha}_i(y_i - f(x_i) - \epsilon - \xi_i) = 0 \\ \alpha_i\hat{\alpha}_i = 0, \xi_i\hat{\xi}_i = 0 \\ (C - \alpha_i)\xi_i = 0, (C - \hat{\alpha}_i)\hat{\xi}_i = 0 \end{cases} \tag{2.26}$$

最后可得 SVR 的解为：

$$f(x) = \sum_{i=1}^{m}(\hat{\alpha}_i - \alpha_i)\,x_i^T x + b \tag{2.27}$$

其中 b 为：

$$b = y_i + \epsilon - \sum_{i=1}^{m}(\hat{\alpha}_i - \alpha_i)\,x_i^T x \tag{2.28}$$

（二）核函数

在上述定义中我们假定训练样本是线性可分的，但是在常见的问题中，所遇到的数据大多为线性不可分，面对线性不可分问题可以将二维空间转化为更高维空间，想要样本在这个特征空间内线性可分，这就需要用到核函数：

$$k(x_i, x_j) = \phi(x_i)^T\phi(x_j) \tag{2.29}$$

其中$\phi(x)$表示将 x 映射后的特征向量，$k(x_i, x_j)$就是核函数。

表 2.2 常用核函数

名称	表达式	参数
线性核	$k(x_i, x_j) = x_i^T x_j$	
多项式核	$k(x_i, x_j) = (x_i^T x_j)^d$	$d \ge 1$为多项式次数
高斯核	$k(x_i, x_j) = \exp\left(-\frac{\|\|x_i - x_j\|\|^2}{2\sigma^2}\right)$	$\sigma > 0$为高斯核的带宽
拉普拉斯核	$k(x_i, x_j) = \exp\left(-\frac{\|\|x_i - x_j\|\|}{\sigma}\right)$	$\sigma > 0$

续表 2.2 常用核函数

Sigmoid 核	$k(x_i, x_j) = \tanh(\beta x_i^T x_j + \theta)$	tanh 为双曲正切函数，$\beta > 0$，$\theta < 0$

2.8 ARIMA 模型

2.8.1 模型简介

ARIMA 经常被用于时间序列数据的预测。它所处理的时间序列数据是动态的而非静止的，尤其是对随机过程的动态处理更为合适。但是前提是数据必须为平稳数列。然而，如果将非平稳数量应用到该模型中，就必须首先进行差分，经过差分会得到一个平稳数列。ARIMA 模型的大致原理是从自身的历史数据中学习，一旦学会，将应用于今后的预测当中。ARIMA（p，d，q），d 是差分的步长（差分的阶数指的是进行多少次差分。比如步长为 n 的一阶差分$diff(x) = f(x) - f(x-n)$,而二阶步长为 n 的差分：$diff(x) = f(x) - f(x-n)$, $diff(x-n) = f(x-n) - f(x-n-n)$, diff 二阶差分$(x-n) = diff(x) - diff(x-n)$，经过差分将会得到平稳数列。p 为自回归项，q 是移动平均项数。

（1）自回归模型 AR

自回归模型通过历史数据的趋势来对后续的数据进行预测，在对历史数据的趋势和规律进行分析时建立动态的预测模型用于预测。自回归模型的输入数据必须时平稳数据。自回归模型的运用中首先需要确定一个序列中的多少阶 p 作为历史数据来预测下一期的值。p 阶的自回归模型可以表达如下：

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t \tag{2.30}$$

y_t是预测值，μ是常数项，p 是阶数，γ是相关系数，ϵ_t是误差项。

（2）移动平均模型 MA

移动平均模型对模型回归中的误差项进行累加，q 阶自回归过程的公式定义如下：

$$y_t = \mu + \epsilon_t + \sum_{i=1}^{p} \theta_i \epsilon_{t-i} \tag{2.31}$$

同时移动平均模型能有效地消除预测中的随机波动。

（3）自回归移动平均模型 ARMA

将自回归和移动平均模型相结合，我们就得到了自回归移动平均模型 ARMA(p,q)，计算公式如下：

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t + \sum_{i=1}^{p} \theta_i \epsilon_{t-i} \quad (2.32)$$

2.8.2 参数确定

（1）p,q 的确定

表 2.3 ARIMA 模型参数确定

模型	AR(p)	MA(q)	ARMA(p,q)
自相关函数	拖尾	q 阶后截尾	拖尾
偏自相关函数	P 阶后截尾	拖尾	拖尾

（2）信息准则法

模型结束的确定对预测结果的误差起决定性作用，在相同的预测误差情况下，根据奥斯卡姆剃刀准则，模型越小越好。我们可以根据信息准则函数法，来确定模型的阶数。预测误差通常用平方误差即残差平方和来表示。常用的信息准则法：

$$AIC = 2K - \ln(L) \quad (2.33)$$

$$BIC = Kln(n) - 2\ln(L) \quad (2.24)$$

其中 k 是模型参数个数，n 为样本数量，L 是似然函数。

2.8.3 模型的检验

在建模之后进行假设检验，诊断残差序列是否为白噪声。主要的检验值：

（1）检验参数估计的显著性（t 检验）

（2）检验残差序列的随机性，即残差之间是独立的，残差序列的随机性可以通过自相关函数法来检验，即做残差的自相关函数图。

3 兰州市空气质量现状分析

3.1 兰州市空间地理环境分析

兰州隶属于我国西北几省之一的甘肃省，市中心位于北纬 36 度和东经 103 度。兰州市位于我国西北部，西北部许多城市的特点便是地貌复杂多样，拥有我国大部分的地貌特点，但大多以沙漠和戈壁为主。地势由西南向东北倾斜，地形狭长，四面环山，形成相对封闭的环境气候。空气相对静止，风力较小，不利于空气污染物的扩散。兰州位于西北内陆。难以借助海洋的温度和湿度，常年降雨较少，气候干燥。属于大陆性温带季风气候。兰州的季节特点是冬季寒冷干燥，春秋气温反复无常，夏季降雨少，温度高，秋季多雨导致气温骤降。各个地区间气温差距较大，且早晚温差也颇为明显。全省各地年降水量 36.6~734.9mm，且降水分布不均，东南部分地区降水多，而西北地区相对较少。受气候影响， 6～8 月是降水的密集时间段，占全年降水量的 50%～70%。降水区域分布不均，有的地区降水量堪比其他城市的总和，但有的地区甚至全年的降水量只有 300 毫米。可能是地势的原因，导致空气中的相对湿度较小，不利于空气中悬浮颗粒物或浮尘的沉淀。气候变化大，生态环境复杂多样。由此可见，兰州的自然环境和复杂地形对空气质量有很大影响。

3.2 兰州市空气质量时间变化分析

3.2.1 数据来源及预处理

“兰州市沿黄河而建,四面群山环绕,地形狭长。自古以来，由于兰州市沟壑地势特点，地理位置险要，空气相对静止，不易扩散，导致空气中的污染长期存续，导致空气质量较差。故本文选取兰州市空气质量指数数据作实证研究，希望能够提供兰州市空气质量指数较为精准的预测结果，为有关部门针对空气质量情况预警提供有力依据，告知群众空气质量情况，以提前做好准备，防患于未然。

本文空气质量的数据来源于《中国环境检测总站》。采用 2014 年到 2020 年的日空气质量数据，共 2556 条有效数据。其中空气质量数据包括 AQI，$PM_{2.5}$、

PM_{10}、SO_2、NO_2、CO、O_3。选取兰州市空气污染物与空气治理指数数据，分析兰州市空气质量变动情况，分析 AQI 空气质量指数受空气污染物的影响，运用 Python、SPSS、Excel 软件对数据进行整理、模型的建立，对空气治理指数预测，探索更精准的预测模型为空气治理的应急响应机制提供坚实基础。下面对数据进行预处理：

（1）数据归一化

Min-Max 标准化（Min-Max Normalization）

由于数据的不同种类以及不同的规范单位，如果直接进行数据的分析很有可能造成数据间差异巨大，导致分析结果产生巨大误差，不利于后续研究的继续，不能提供科学的数据基础。数据标准化的意义就在于可以将不同量纲的数据经过处理，消除量纲进而可将数据间的真实关系更准确的显示，让数据具有可比性。原始数据经过数据标准化处理后，各组数据处于同一量纲，适合进行综合对比评价。公式如下：

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3.1}$$

Z-score 标准化方法

零-均值规范化也称标准差标准化，零-均值规范化主要是通过将所有数据通过缩减或放大，将数据的范围控制在（0，1）区间内，经过处理的数据的均值为 0，标准差为 1。转化公式为：

$$x_{new} = \frac{x - \mu}{\sigma} \tag{3.2}$$

经过 Z-score 标准化使得各数据值统一，数据集也具有可比性。除此之外，数据标准化还有利于初始化的进行，从而避免给梯度数值的更新带来数值问题，还有利于学习率数值的调整，加快寻找最优解速度。

（2）异常值处理

对空气质量指数进行未来预测时，历史空气质量数据的真实有效性具有非常重要的作用。但是常用的空气质量指数数据大多是仪器采集的，由于设备本身问题、数据传导问题或者网络链接问题会导致数据采集的异常或缺失。在传送过来的数据集中就会存在异常数据，因为仪器会将手机过程中的缺失值以某一数字进行填充。如果异常数据不处理常常会导致数据集不完整或数据误差范围较大，导致预测数据的真实性准确性收到一定程度的影响。所以我们需要对所获数据进行

异常值处理，保持输入数据的完整和准确。主要的数据处理方式有：均值填补、多重插补和删除法。

首先将数据的缺测值进行填补，一般的填补方法为均值填补或多重插补方法，对于却测数据不多的可以采取删除的方式。而数据异常值的检验可以根据简单的统计分析，做一个描述统计，人为的观察判断数据是否在正常的区间范围，从而确定是否异常。还可以运用 3σ原则，在正常情况下，数据应该在 3σ范围内，若超过则判定为异常值。也可以运用箱线图进行分析，箱线图的选取值较为客观，在判断异常值和离散点方面具有其特有的优势。在本文的输入数据中，有 7 条缺测数据，采取平均值填补的方式，将数据集补充完整，为后续的预测方法提供合理依据。

3.2.2 空气质量年度周期性特征

由于过去我国过于追求 GDP 增长指标，而忽略了绿色 GDP 的价值理念，没有将可持续发展作为首要目标，造成人居环境的破坏，空气质量得不到保证，尤其是以发达地区的问题更加明显，空气质量下降，导致人们群众的身体健康受到影响。近年来国家对于空气污染问题的重视程度日益上升，在经过实施“蓝天工程”后，空气质量得到一定程度改善。但仍存在由于人为活动或自然气象的因素造成大气污染的情况，且兰州市依河而建，四面环山，由于城市的主要建设是在黄河两岸，地势狭长。若无风力的加持，空气中的污染物长期存在，无法扩散。更重要的是，兰州拥有的几家工业基地，产生的废气或工业污染，加重了兰州的大气污染程度，一度被冠以“黑帽子”的别称，特别是浮沉、扬沙等恶劣天气不便于人们出行。因此，本文基于兰州市 2014 年到 2020 年的空气污染浓度及空气质量指数数据，对兰州市空气质量的波动幅度和主要节点分析，找出兰州市空气质量的变化趋势。影响空气质量的主要污染物为$PM_{2.5}$、PM_{10}、SO_2、NO_2、CO、O_3。本文大气污染物浓度数据均来自于真气网中数据。在时间区分上选取季节年，即将 3 月-5 月认定为春季，6 月-8 月为夏季，9 月-11 月为秋季，12 月-次年 2 月为冬季，季节年包括了一整个冬季，可以完整的体现出季节特征。

表 3.1 部分空气质量数据

日期	AQI	$PM_{2.5}$ $(\mu g/m^3)$	PM_{10} $(\mu g/m^3)$	SO_2 $(\mu g/m^3)$	NO_0 $(\mu g/m^3)$	O_3 (mg/m^3)	CO $(\mu g/m^3)$
2014/1/1	158	121	216	53	38	29	1.8
2014/1/2	99	74	147	46	29	33	1.6
2014/1/3	96	69	142	53	33	29	1.5
2014/1/4	121	92	164	51	37	33	1.6
…	…	…	…	…	…	…	…
2020/12/30	55	28	60	18	41	58	1.3
2020/12/31	78	46	99	28	62	30	2.1

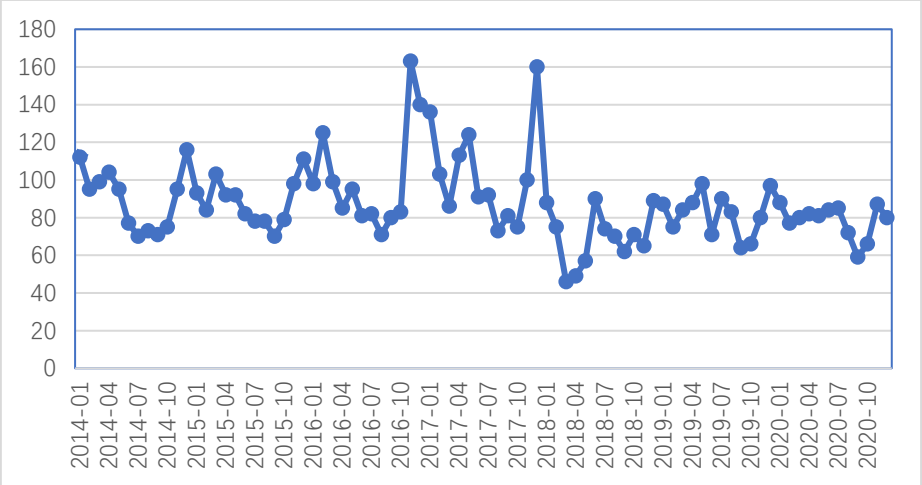

图 3.1 2014-2020 年 AQI 指数变化

由图 3.1 可以看出，兰州市空气质量指数成周期性波动，除 2016 年和 2018 年某个月份出现大幅波动外，其他年份的波动趋势大致相同。一般在 12 月份达到最高，由于冬天的取暖刚需，对于煤炭的需求量较大，从而导致大气中污染物浓度较高，空气质量下降。8 月份相对最低，由于夏季气候适宜，在降雨后污染物浓度下降，且随着温度升高，污染物扩散加快，使得空气质量较冬季有所缓解。可以看出在 2018 年夏季之后，兰州市空气质量有所改善。

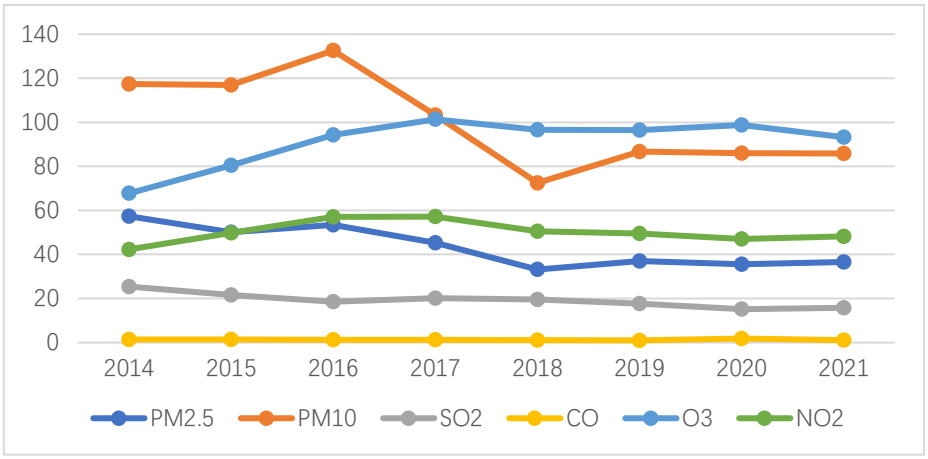

图 3.2 大气污染物浓度变化

由图 3.2 所示，$PM_{2.5}$、PM_{10}、SO_2、NO_2、CO以及O_3污染物浓度呈先上升后下降的趋势，自 2014 年以来至 2017 年，兰州市各项污染物浓度逐年上升，这可能是由于兰州的地理位置狭长，群山环绕，导致污染物不易扩散，从而使得空气质量不佳。但在 2017 年之后呈下降趋势，说明在 2017 年兰州市空气质量得到了有效改善。

3.2.3 空气质量明显的季度特征

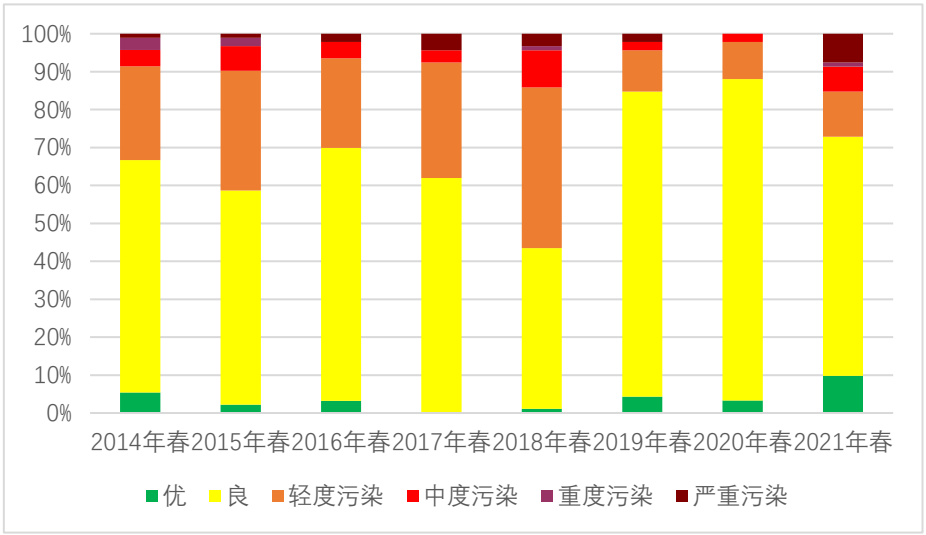

图 3.3 春季污染特征

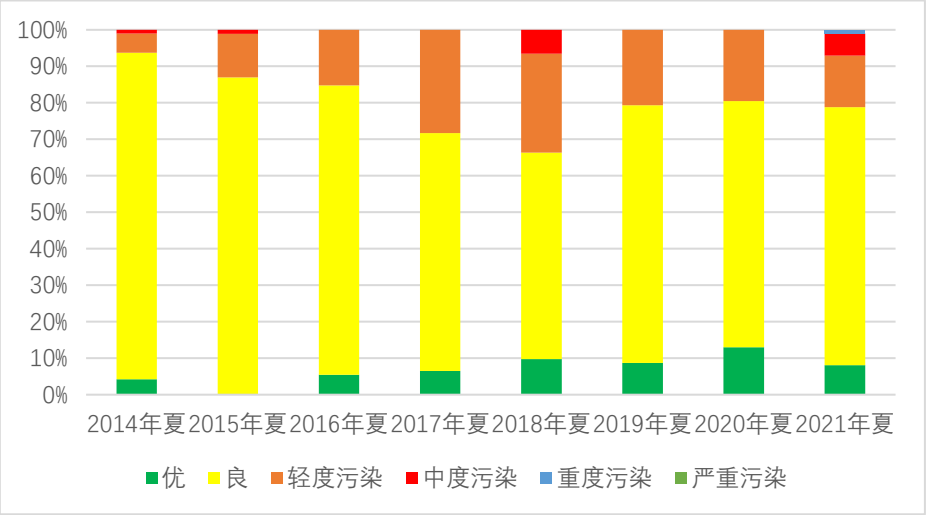

图 3.4 夏季污染特征

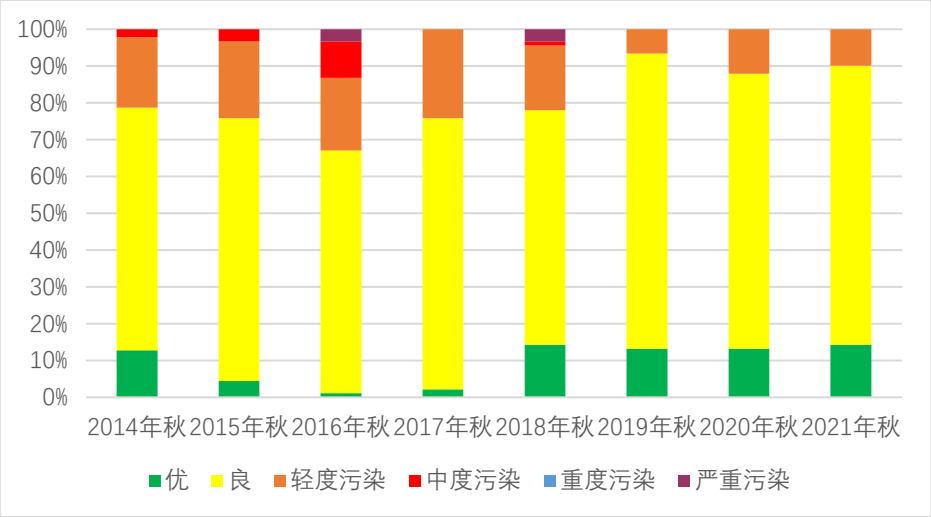

图 3.5 秋季污染特征

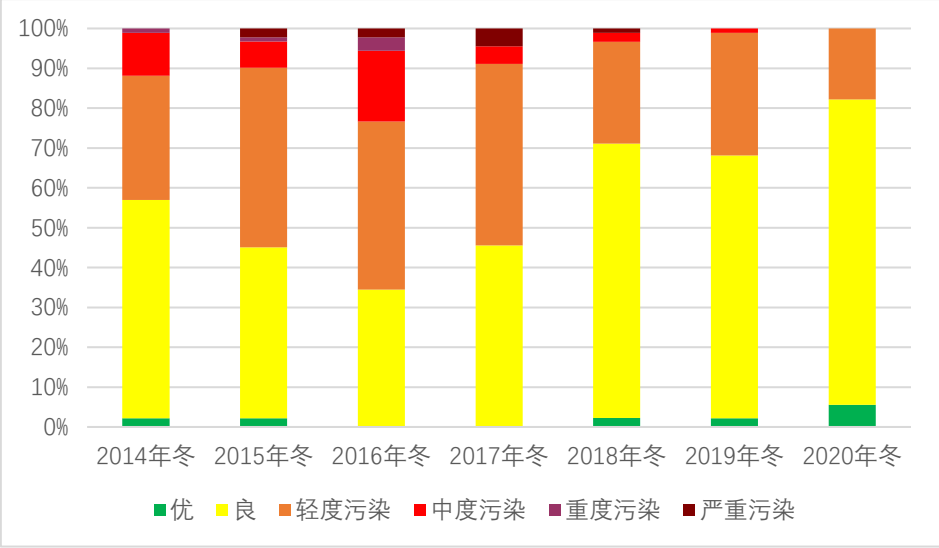

图 3.6 冬季污染特征

由图 3.3 到 3.6 可知，兰州市空气污染物浓度的季节性特征较为明显，可以看出冬季污染程度普遍较重，夏季程度最轻，而秋季和春季污染程度在夏冬之间，其中春季污染程度较秋季稍显严重。这是由于在冬季要燃烧燃料进行取暖以及秸秆焚烧处理。燃料的燃烧会产生CO和SO_2，若不经处理向大气中排放废气，大气中污染物浓度增加，导致冬季的空气质量恶化。而在夏季中，由于夏天空气温度高，降雨多空气中相对湿度较大，可以减少空气中的污染物浓度，降雨可以将空气中的悬浮物沉淀到地面，改善空气质量。说明在一年当中夏天的空气质量是最好的，且气候条件也较为良好，这时人们日常出行以及进行休闲娱乐生活也比较适宜。而冬天由于天气寒冷，雨水较少，对于空气质量的影响较大，空气质量差，不利于人们的出行活动。

3.2.4 空气质量月度变化呈“U”型特征

由图3.7可知，兰州市空气污染物浓度整体随时间变化呈“U”型，其中$PM_{2.5}$、PM_{10}、SO_2、CO、NO_2污染物在 1 月份之后呈下降趋势，在 7、8 月份达到最低，自 9 月份开始呈上升趋势。是由于在夏天风力和垂直气温的变化，使得垂直气温上下波动幅度大，当风速较高时，风吹过地势的起伏，加强了风的流速，在这种物理现象下，空气流动，空气中的污染物会随着风力的作用稀释和扩散。大多数污染物浓度在夏季呈现下降趋势，但臭氧在夏季不降反升，这是由于夏季多雨加上空气的流动，大气污染物随雨水沉淀及风力推动浓度迅速降低，是一年当中空气质量较好的季节，人们在夏天经常可以看到蓝天白云。但是，臭氧的形成是由空气中的氮氧化物和有机物在太阳的照射下温度升高，从而发生了化学反应，生成的物质将臭氧氧化。夏季温度高，太阳照射时间长，在这样的条件下，臭氧浓度升高。因此，在夏季，臭氧会随着气温上升而增多。

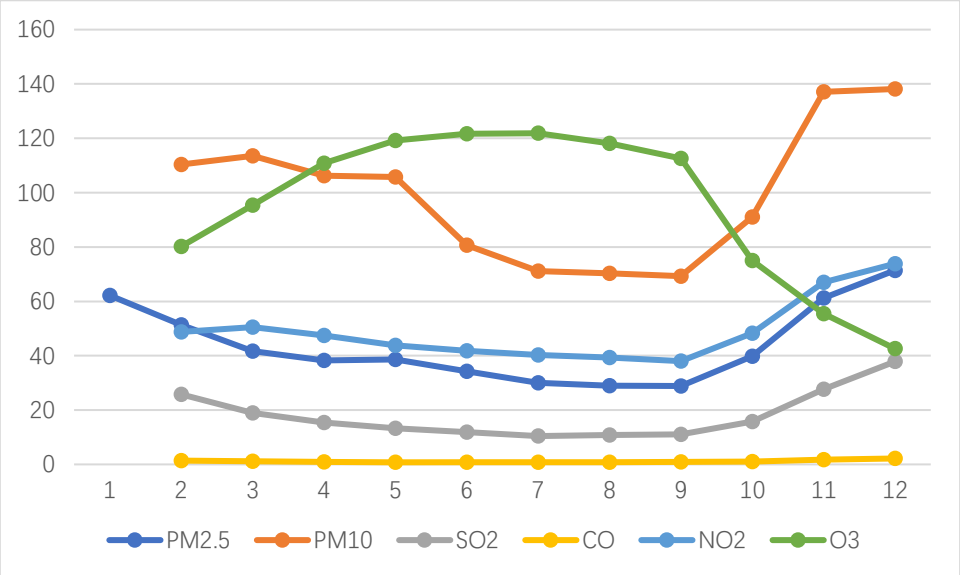

图 3.7 污染物浓度月度变化特征

3.3 主要污染物分析

目前空气中存在的氮氧化物、硫氧化物以及可吸入颗粒是影响空气质量的主要原因。其中氮氧化物是指空气中含氮的污染物，主要为NO和NO_2。氮氧化物是较容易产生的一种空气污染物，在很多的生活场景中都有可能产生。比如汽车尾气、工厂废气等，一氧化氮并不可怕，但是氧化氮可被臭氧氧化，一旦排放到空气中，被空气中的臭氧氧化为NO_2，NO_2浓度过高对人们的身体会产生不利影响。硫氧化物中对空气质量产生较大形象的为SO_2，SO_2在是由燃烧和汽车尾气等排放出污染物，由于目前我国的能源结构单一，虽然新能源汽车的极力推广，一定程度上主力了产业升级，但是由燃烧的其他行业也亟待整改和优化。二氧化硫对人体健康有重要影响，其化学性质不稳定容易和其他的物质结合，产生酸性的凝结物滞留在空气中。人们通过呼吸吸入肺中，将会对人体的呼吸系统产生重大影响。可吸入颗粒：主要指分散悬浮在空气中的液态或固态物质，其粒度在微米级，包括气溶胶、烟、尘、雾和炭烟等多种形态。可吸入颗粒一旦进入人体，将会给呼吸系统造成巨大负担。

3.3.1 AQI 与空气污染物间相关性分析

从上述分析中可以看出，空气质量程度呈现出明显的季节性特征，在规律背

后是大气中各种污染物浓度的变化造成的。因此，对空气质量指数影响较大的污染物进行探究，需要对大气污染物浓度与 AQI 指数进行相关分析，找出影响 AQI 空气质量指数程度最大的大气污染物，从而提供科学依据，能够使政府根据不同的污染物采取更有针对性的治理措施。

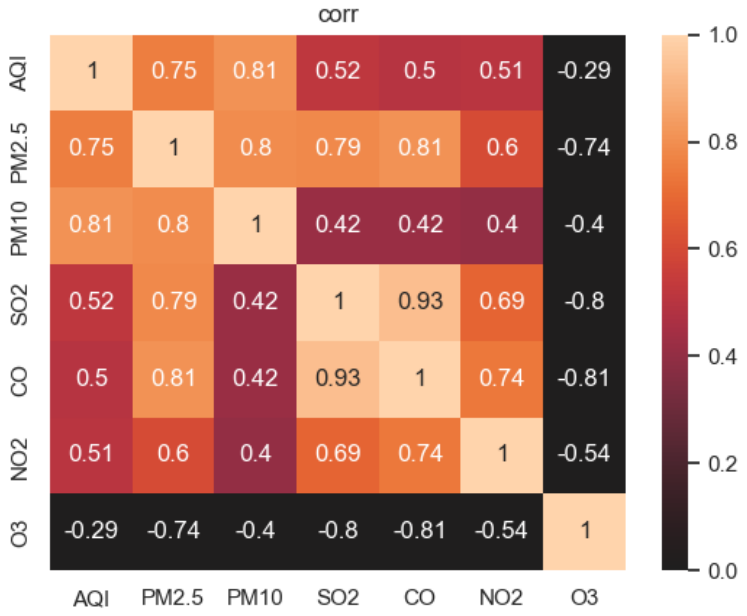

图 3.8 污染物浓度相关热力图

这里选用 2014 年 1 月至 2020 年 12 月期间共 2556 条有效数据对各主要污染物之间的相关关系进行了分析，如图 4.1 所示，对比图中右侧的刻度和颜色可以看出不同指标之间的相关关系。各污染物浓度之间具有较强的相关关系，其中，O_3与$PM_{2.5}$和CO的之间具有较强的负相关关系，与其他污染物浓度呈弱负相关性，其中与 PM_{10}的相关性最弱。 AQI 与$PM_{2.5}$和PM_{10}的相关关系较强，相关系数分别为 0.75 和 0.81，说明$PM_{2.5}$和PM_{10}的浓度变与 AQI 数值同向变化。而 AQI 与O_3存在较弱的负相关性。综上所述，影响 AQI 的主要因素是$PM_{2.5}$和PM_{10}。相关部门要针对$PM_{2.5}$和PM_{10}采取更有针对性的措施进行空气治理。

3.3.2 空气污染物对 AQI 的特征重要性选择

利用随机森林进行指标的选取，可以直观的看出哪些因素对于 AQI 空气质量指数的变化影响最大，对于分析空气质量的影响因素至关重要。具体步骤如下：

1）首先针对决策树计算袋外数据误差，即通过重复抽样得到的数据计算其

与训练集数据的误差，记为err_{ooB1}。通过同样的操作训练决策树。由于是重复抽样，所以一定有没有被抽中的数据，这部分数据无法参与决策树的训练就作为评估数据，用来验证决策树的性能。

2）其次通过对没有参与决策树建立的数据加入随机噪声干扰，来改变不同数据对样本特征的决策，再次计算袋外数据误差，记为err_{ooB2}。

3）假设森林中有 N 棵树，$X = \sum \frac{err_{ooB2} - err_{ooB1}}{N}$。所计算出的 X 就能够代表特征的重要性。原理是再加入随机噪声后，若袋外数据前后误差较大，说明此特征在影响预测结果上发挥着重要作用，从而体现该特征的重要性程度。

本文利用$PM_{2.5}$、PM_{10}等空气污染物进行相关分析，可以看出各污染物浓度的变化对 AQI 空气质量指数的重要程度。得到特征重要性排序如表 3.2 所示：

表 3.2 特征重要性

指标名称	重要性	排序
$PM_{2.5}$	0.680861	1
PM_{10}	0.271771	2
SO_2	0.151462	3
NO_2	0.013103	4
O_3	0.010105	5
CO	0.009014	6

由表 3.2 可知，在各个空气污染物中$PM_{2.5}$对空气质量指数的影响程度最大，其重要性系数为 0.681，影响程度排名第二的为PM_{10}，其重要性系数为 0.272。可见，对兰州市空气质量影响较大的是$PM_{2.5}$、PM_{10}和SO_2，其中CO和O_3对兰州市空气质量影响较小，在空气中臭氧和一氧化碳的含量占比较小，且由于其化学性质不稳定，容易被氧化，所以对 AQI 空气治理指数的影响有限，而在燃料燃烧、汽车尾气的排放中SO_2和悬浮颗粒浓度偏高，将会对空气质量产生较大的影响。

3.3.3 空气污染物对 AQI 的回归分析

在对兰州市各个污染物指标进行重要性筛选后，得出相应的重要性系数，将重要系数大于 0.1 的指标作为影响因素，对 AQI 空气质量指数进行回归分析，探究各个影响因素与空气质量之间的关系。利用 SPSS 建立回归模型，因变量选取 AQI 指数，自变量选取$PM_{2.5}$、PM_{10}和SO_2共 3 种因素进行模型建立回归模型。

需要进行多重共线性检验：

若自变量的 VIF 大于 10，且容忍度小于 0.1，说明自变量间存在严重的多重共线性，由表 3.3 可以看出，本节中所选变量的容忍度均大于 0.1，且 VIF 小于 10，因此，为剔除不显著变量，本章采取逐步回归方法，对 AQI 空气质量的影响因素进行分析。

表 3.3 自变量的共线性检验

自变量	VIF	容忍度
$PM_{2.5}$	5.502	0.182
PM_{10}	3.938	0.254
SO_2	1.961	1.961

由表 3.4 可知，调整后 R 的平方为 0.660，F 值为 187.484，显著性水平为 0.000 小于 0.05。因此，AQI 与各空气污染物有显著的线性关系，证明可以用逐步回归模型对 AQI 与空气污染物进行分析。

表 3.4 回归模型分析表

项目	系数
R 的平方	0.813
调整后 R 的平方	0.660
标准误差	28.475
F	187.484
显著性	0.000

如表 3.5 所示，经过逐步分析后将二氧化硫剔除，得到$PM_{2.5}$和PM_{10}这两个因素对 AQI 空气质量指数的影响最大，且二者均与 AQI 呈正比。

表 3.5 回归系数表

自变量	回归系数	T 检验统计量	P 值
常数	43.547	41.977	0.000
$PM_{2.5}$	0.240	31.047	0.000
PM_{10}	0.449	14.520	0.000

最终回归方程为：

$$y = 43.547 + 0.240x_1 + 0.449x_2 \tag{3.1}$$

可吸入颗粒浓度越高，AQI 空气指数越大，相应的空气质量越差。主要是因为兰州经常出现沙尘，扬尘等天气，空气中颗粒物长期存在。最近几年兰州地铁的修建、道路施工扬起的工业粉尘也会转化为空气悬浮物，导致空气质量下降。而且燃烧燃料，垃圾秸秆等的燃烧也会造成空气中悬浮物过多，造成雾霾等恶劣天气。因此，在对空气治理治理时，应主要针对悬浮颗粒采取措施，如人工降水、

增加绿色覆盖面积和加大对垃圾分类的监管力度，不断优化垃圾处理手段，减少燃烧，改善空气质量。

4 基于深度学习模型的空气质量预测

4.1 评价指标

本节采用 2014 年到 2020 年的日空气质量数据，共 2556 条有效数据。其中空气质量指标为 AQI，$PM_{2.5}$、PM_{10}、SO_2、NO_2、CO、O_3。评价指标采用 RMSE、MAE 和R^2。

（1）拟合优度：R^2作为回归方程拟合整体程度的衡量指标，它表达了因变量与所有自变量之间的总体关系。R^2等于回归平方和比总平方和的比率，计算实际误差时，采取实际值与平均值最小的原则，回归误差与剩余误差是此消彼长的关系，即当回归误差增大则剩余误差就会减小，反之亦然。因而回归误差从正面测定线性模型的拟合优度，剩余误差则从反面来判定线性模型的拟合优度。计算公式如下：

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \tag{4.1}$$

其中y_i为第 i 个数据真实值，$\hat{y}_i$为插补值，$\bar{y}$为平均值。

（2）平均绝对误差：MAE（Mean Absolute Error）是绝对误差的平均值，能更好地反映预测值误差的实际情况。计算公式如下：

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|(\hat{y}_i - y_i)| \tag{4.2}$$

其中 n 为缺失值的样本数量。

（3）均方根误差：又叫标准误差（Root Mean Square Error）对一组测量中的特大或特小误差反映非常敏感，所以，标准误差能够很好地反映出预测数据的精密度。计算公式如下：

$$RMSE = \sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.3}$$

上述四种方法的拟合度都比较高，不能从曲线上分辨哪一种模型更好，因此我们用拟合优度，平均绝对误差和均方根误差进行对比：

4.2 预测模型的选择

对 AQI 空气质量指数单因素进行预测分析，目前常见的预测模型有 LSTM、RF、SVR 三种模型，且根据阅读文献后发现，三种模型对于不同的数据表现不同，但都较为精准，故本节选取选取此三种模型进行对比，检验各类模型对于空气质量预测的可行性，最终选取效果较好的模型作为最终预测工具。

4.2.1 LSTM 模型的建立

其中 LSTM 模型的建立是基于 python 语言中 Keras 框架，主要分为三步：数据预处理、确定模型参数、还原训练输出值。

（1）选择样本，构建训练集。在机器学习中，数据制备是较为复杂，且十分重要。模型能够正常运行的前提是要制备好训练集和测试集，训练样本过多将会导致训练难以拟合，若训练样本过少就会增大训练过程中的拟合误差。故本次 LSTM 模型的训练集与预测集采取 8：2 划分，即训练集为 2054 条数据，测试集为 502 条数据。

（2）数据预处理

本节的输入数据为兰州 2014-2020 年的空气质量数据，所能收集到的所有空气质量数据都是由地面观测站自动传送到终端，有可能出行网络中断或机器故障的情况，导致数据集不完整，为保证后续操作的精准性，需要将数据集补充完整。故先将缺失数据以前后平均值填充，构建一个完整的数据集之后再做后续的操作，为了让模型加快收敛，需要对数据进行归一化处理。

（3）确定模型参数

在模型中设置输入输出维度，本文经过参数调整得到最优参数如下所示，分析确定隐藏节点数为 32，训练批次为 32，迭代次数为 150。经过对优化器及损失函数的实验对比，确定 adame 和 mae 为本文的优化器和损失函数。本文试验基于 Keras 框架，在确定参数及配置后，由 Keras 传输至 TensorFlow 后端进行整个模型的训练与计算。

①滑动窗口大小：初始设定 units 为 16，epochs 为 100，batch_size 为 64，根据不同滑动窗口大小选择误差最小的值，作为最终滑动窗口参数。由表 4.1 所知，

当滑动窗口依次增大时，预测误差不断增大，在滑动窗口为 25 时，误差出现了拐点，再次增大滑动窗口大小，误差继续增大，可见滑动窗口大小为 25 时误差最小。因此，确定滑动窗口大小为 25。

表 4.1 滑动窗口验证误差

滑动窗口大小	RMSE	MAE	R^2
5	19.06	14.58	0.53
10	18.92	14.45	0.53
15	18.68	14.39	0.52
20	18.62	14.23	0.54
25	18.50	14.15	0.54
30	18.65	14.24	0.53
35	18.93	14.47	0.53

②epochs 确定：由表 4.2 可知，为能够尽快的试验出最佳迭代次数，在试验时，以 50 每量级增加，随着迭代次数的增加误差不断减小，在迭代次数为 150 时，预测误差最小，超过 150 预测误差在逐渐增大，因此，本文确定迭代次数为 150。

表 4.2 epochs 验证误差

epochs	RMSE	MAE	R^2
50	19.30	15.02	0.48
100	18.58	14.17	0.54
150	18.48	14.13	0.54
200	18.52	14.16	0.53
250	18.59	14.25	0.52

③batch_size 确定：由表 4.3 可知，根据阅读相关文献，选取大多数模型中出现频率最高的迭代批次进行试验，找出相对最优的迭代批次，可见 batch_size 为 32 时预测误差最小，因此，本文确定 batch_size 为 32。

表 4.3 batch_size 验证误差

batch_size	RMSE	MAE	R^2
16	18.47	14.10	0.53
32	18.42	14.04	0.55
64	18.50	14.15	0.54
128	18.59	14.20	0.54

④units 确定：units 是隐藏神经元，隐藏神经元的个数也学要仔细斟酌，他决定了信息的输入输出，过多或过少都可能影响输出结果，通过一次次试验找出最佳的隐藏神经元个数。由表 4.4 可知，在 units 取 32 时得到预测误差最小值，因此，本文 units 参数确定为 32.

表 4.4 units 验证误差

units	RMSE	MAE	R^2
16	18.51	14.18	0.54
32	18.48	14.15	0.54
64	18.51	14.18	0.53
128	18.52	14.15	0.53

4.2.2 随机森林模型的建立

其中随机森林的建模步骤为：

（1）数据预处理

RF 模型使用数据依然为 AQI 空气质量指数数据，选取兰州市 2014-2020 年的空气质量数据。为保证模型对比可信度，RF 的训练集与测试集数量要与 LSTM 模型数量一致。即训练集为 2054 条，测试集为 502 条。

（2）决策树的选择

RF 模型的建立是基于 python 语言中 sklearn 库中的 ensemble 包。其主要参数为决策树个数，决策树过多会导致训练时间过长，且不会提高模型效果，而决策树量越少会容易出现过拟合，因此，选择合适的决策树数量，对得到良好的预测结果有较大的帮助。具体结果如下：

由表 4.5 可知，随着决策树的不断增加，虽然误差的变化细微，但在决策树为 40 时标准误差最小为 20.183，随着决策树数量的增加，标准误差变大，故最终确定决策树个数为 40。

表 4.5 决策树评价

n_estimators	RMSE
10	20.390
20	20.299
30	20.260
40	20.183
50	20.278

除了决策树的数量之外，其他重要的 RF 模型具体参数如表 4.6 所示：

表 4.6 RF 算法参数设置

参数	参数取值
n_estimators	40
max_features	Auto
max_depth	None
bootstrap	True

4.2.3 支持向量回归模型的建立

SVR 模型的建立是基于 python 语言中 sklearn 库中的 SVR 包，主要分为两步：数据预处理、核函数选取。

（1）数据预处理

由于不同的数据间的量纲会影响网络学习，所以需要对数据进行归一化处理，提升模型的训练速度。SVR 模型使用数据依然为 AQI 空气质量指数数据，为保证模型对比可信度，SVR 训练集也为 2054 条，测试集为 502 条。

（2）核函数的选取

选择合适的核函数有助于得到较好的预测效果，选择核函数一般采取如下几种方法：一种为经验法，即利用专家的经验以及专业知识来选择核函数，能够更好的匹配模型；二是采用 Cross-Validation 方法，这种方法较于经验法具有更好的科学性，主要操作时将不同的核函数依次替换试验，通过对核函数的对比，选择误差最小的做为本次试验最终确定的核函数。在本节当中主要将常用的核函数 Linear、Poly 和 Rbf 进行对比，找出最适合的核函数运用到数据预测中，能够提供更好的预测效果。由表 4.7 可知，在常见核函数中，选用 RBF 核函数在相同条件下所预测的误差值最小，因此最终选取 RBF 核函数，具体情况见表 4.7：

表 4.7 核函数评价

核函数	RMSE
Linear	19.158
Poly	21.408
Rbf	19.110

SVR 模型的具体参数如表 4.8 所示：

表 4.8 SVR 算法参数设置

参数	参数含义	参数取值
Kernel	核函数	RBF
C	惩罚系数	1.0
ϵ	不敏感损失函数	0.0001

4.2.4 模型预测与结果分析

由图 4.1-4.3 可知，各模型最终预测值的曲线随整体趋势上大致相同，但仍有细微的差别，途中蓝色曲线为预测数据所构成的曲线，橙色曲线为测试集曲线，与 LSTM 模型相比，随机森林模型的预测曲线在中间偏后一段，预测值曲线与测试值曲线差距过大。同样，支持向量回归模型的整体预测曲线都与真实值曲线有一定差距，而长短期记忆网络模型的预测值和真实值曲线拟合度更高。

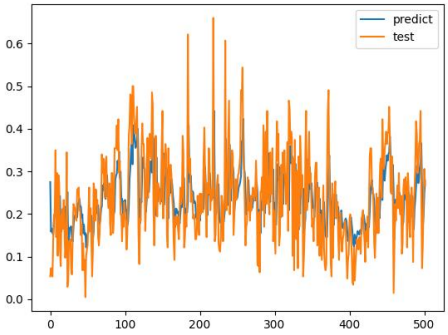

图 4.1 LSTM 模型预测结果

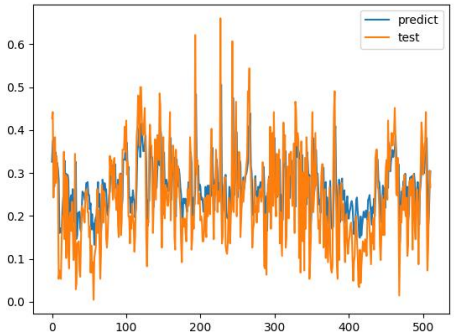

图 4.2 RF 模型预测结果

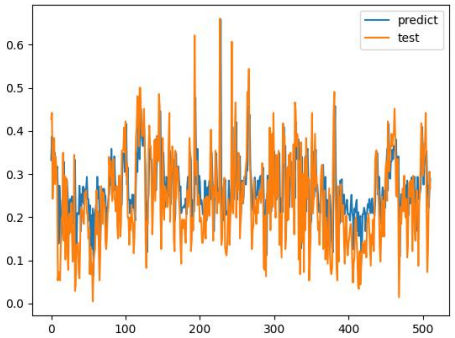

图 4.3 SVR 模型预测结果

由表 4.9 可知，在三种模型的预测结果看，单一因素 LSTM 模型效果由于 SVR 和 RF，其中 SVR 各项指标优于 RF。预测值与真实值之间 RMSE 为 18.94，MAE 为 14.34，拟合优度为 0.52。而 RF 和 SVR 模型的 RMSE、MAE 值都高于

LSTM 且拟合优度不足 0.5，说明其拟合效果不好。证明 LSTM 模型可以作为空气质量的预测工具。

因此，在经过与其他两种常见的模型预测精度进行对比后，可以看出，LSTM 模型在对空气质量指数预测中表现突出，可为下一步组合模型提供基础。

表 4.9 各模型效果评估

	RMSE	MAE	R^2
LSTM 模型	18.94	14.34	0.52
RF	20.18	15.58	0.41
SVR	19.11	14.68	0.44

5 基于组合模型的空气质量预测

5.1 EEMD-LSTM 模型的空气质量预测

本章数据选用与第四章相同数据，为了可以与 LSTM 单一模型和 EEMD-LSTM 模型进行对比。选用 2014 年到 2020 年 AQI 空气质量指数共 2556 条有效数据。评级指标分别为 RMSE、MAE 和R^2。

5.1.1 数据分解

运用 EEMD 方法将原始信号分解，分解之后的数据可以根据数据的波动频率进行区分，由图 5.1 可知，AQI 空气质量指数的原始数据可以分为 9 个 IMF 分量以及 1 个趋势项。从上到下的曲线波动频率依次递减，其中最上面 Singal 曲线为原始数据，IMF1-IMF9 为各 IMF 分量，RES 为其趋势项。

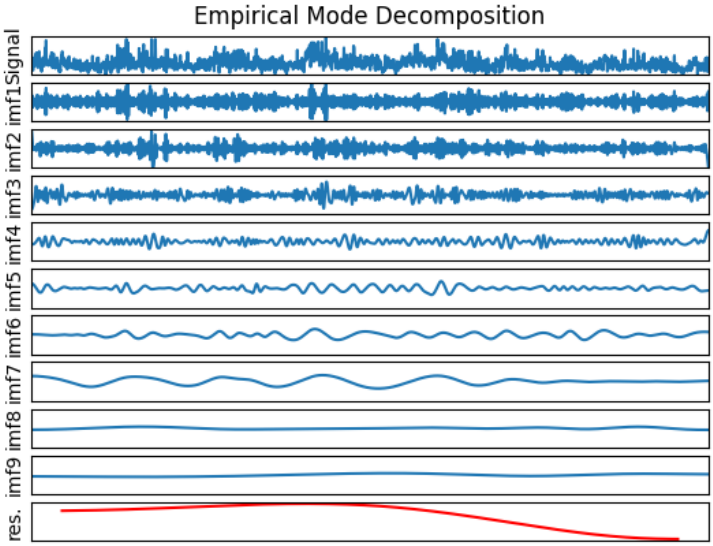

图 5.1 原始序列 EEMD 分解

5.1.2 EEMD-LSTM 建模步骤

本节将 EEMD 分解方法和 LSTM 网络模型进行结合。首先运用集成经验模

态分解法对原始数列进行分解，经过分解后得到不同的 IMF 分量信号及趋势，将每一个 IMF 分量信号作为输入数据输入到 LSTM 模型训练，通过对信号的波动和规律学习后，将预测结果输出。将每一个信号所得到的预测值进行加总得到最终的预测值。本节介绍了 EEMD-LSTM 组合预测模型“输入、预处理、分解、预测、集成”这一基本结构，按照步骤建立模型，选择神经元、滑动窗口等重要参数并对其进行调整。具体步骤如下所示：

（1）数据预处理

首先进行数据遍历，寻找缺失数据，用均值填充，其次由于数据的尺度不同需要将数据进行归一化处理。调用 python sklearn 库中相关函数 min-max，将数据都处理为（0，1）之间的标准数据，输出数据结果文件，作为模型的输入数据。

（2）EEMD 分解

该函数是由三层函数嵌套构成的，最里层函数计算上下包络线的均值和拟合程度，经最里层函数计算后输入到中间层判断数值是否满足 IMF 条件，若不满足则重复上述步骤，在满足 IMF 条件后进入最外层的计算，最外层是针对剩余序列加入白噪声后计算。在实际操作中，首先将原始数据输入，通过自主适应函数对数据进行分解，通过不同尺度的分解会形成多个分量信号，各分量信号按照波动频率排列，最上层为原始信号，最下层为趋势项，作为最终的结果输出。

（3）划分数据集

将 EEMD 分解过后的数列，分别进行数据集的划分。按照一定的比例划分为训练集和测试集，训练集作为输入经过 LSTM 模型的训练，最终得出的预测值与测试集进行对比，以验证训练结果的好坏，根据前文单一模型的数据集划分，将个分量信号也按 8：2 的比例划分，即训练集为 2054，测试集为 502。

（4）LSTM 模型构建

调用 python 机器学习人工神经网络库 Keras，将 Keras 中各个核心层的列表信息传递到 Sequential 模型作为模型基本框架。在 Sequential 函数的框架中，通过 add 函数连接多层网络，将输入门、遗忘门、记忆细胞及输出门的结构以激活函数的形式依次加入到各层网络中。

（5）输出与效果评估

将每一个分量信号分别输入到 LSTM 模型中训练，经过训练后的各分量信

号预测结果求和，作为最终的预测结果。同时，需要将预测结果反归一，根据反归一后的数据评估指标验证模型的有效性。

5.1.3 模型预测与结果分析

通过上述步骤构建 EEMD-LSTM 模型，针对 AQI 数据进行预测，预测结果如图 5.2 所示：

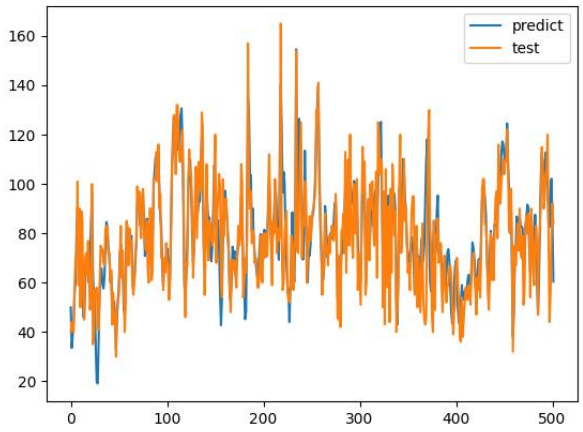

图 5.2 EEMD-LSTM 对 AQI 预测结果

由图 5.2 可知，EEMD-LSTM 组合模型曲线拟合度要高于单一长短期记忆网络模型的拟合度，只有个别的极端值预测有偏差，在其他数据上，两曲线基本重合，由此看出，将原始数据分解，通过分解后的数据去除原数据中的噪声和波动，对于预测 AQI 空气质量指数确实有一定帮助，预测结果更为精准。

从表 5.1 中可以看出 EEMD-LSTM 模型在对 AQI 预测中表现良好，虽有预测值偏离真实值的情况，但从曲线拟合度以及误差数据来看，整体预测程度在可接受范围内，其 RMSE 为 14.06，MAE 为 10.86，拟合优度为 0.53。

表 5.1 模型评价效果

	RMSE	MAE	R^2
EEMD-LSTM	14.06	10.86	0.53

5.2 EEMD-LSTM-ARIMA 模型的空气质量预测

5.2.1 数据分解

首先运用 EEMD 方法将原始信号分解，分解之后的数据可以分为 9 个 IMF 分量以及 1 个趋势项，分解结果如图 5.3 所示，可以看到所分解的数据波动频率逐渐降低，其中最上面曲线为原始数据，从上到下依次时 IM1-IMF9 的各分量信号，最下方的红色曲线则为趋势项。

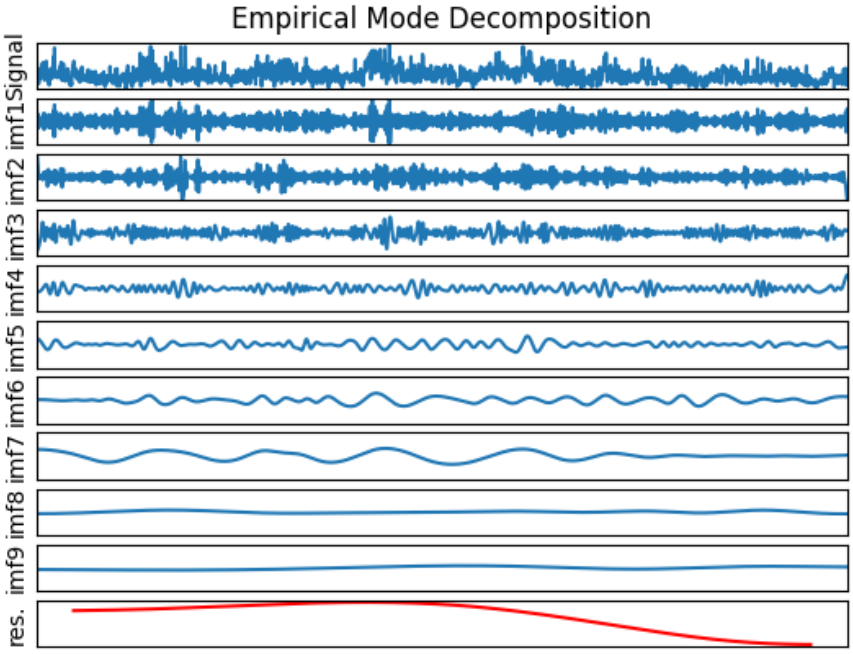

图 5.3 AQI 数据集 EEMD 分解结果

5.2.2 数据重构

由于空气质量指数 AQI 数据分解为 9 个信号和 1 个趋势向，分解后的 IMF 过多，不利于后续预测计算。因此，本章针对分解后的 IMF 分量进行 T 检验重构，若 p 值大于 0.05 则表示该分量具有良好的平稳性，划分为高频数据。若 p 值小于 0.05 则表示该分量具有非平稳下，应划分为低频数据。对于趋势项不进行划分，最后将各个分量划分为高频项、低频项和趋势项。各分量 T 检验结果如表 5.2 所示：

表 5.2 各分量 T 检验结果

IMF	T 统计值	P
IMF1	-1.629	0.103
IMF2	-0.984	0.325
IMF3	-1.319	0.187
IMF4	0.442	0.659
IMF5	-1.425	0.154
IMF6	-0.032	0.974
IMF7	2.386	0.017
IMF8	-9.171	0.001
IMF9	-14.022	0.001
RES	539.668	0.001

由表 5.2 可知，IMF1-IMF6 分量的 T 检验 p 值均大于 0.05，因此，将 IMF1-IMF6 定义为高频项，IMF7-IMF9 分量的 T 检验 p 值均小于 0.05，因此将 IMF7-IMF9 定义为低频项。剩余项作为趋势项。其中各分量的统计特征如表 5.3 所示：

表 5.3 各分量统计特征

IMF	平均周期（天）	Pearson	方差	方差贡献率（%）
IMF1	5.859	0.290	248.396	13.265
IMF2	7.391	0.297	180.003	11.452
IMF3	27.784	0.304	136.630	11.026
IMF4	34.152	0.210	79.106	10.580
IMF5	71.705	0.169	58.133	10.395
IMF6	121.678	0.214	63.786	10.258
IMF7	365.872	0.259	131.919	9.478
IMF8	639.992	0.008	19.894	8.581
IMF9	1278.667	0.048	72.970	8.327
RES	2556.375	0.162	288.350	6.637

由表 5.3 可知，较高频率分项周期短具有较强的随机性，而较低频率分项周期长具有较强的周期性。根据各分量与原 AQI 序列的 Pearson 相关性可知 IMF3 与原序列相关性最高为 0.304，其次是 IMF2 为 0.297。且 IMF1-IMF6 分量的方差贡献率高达 66.977%。因此，本文确定 IMF1-IMF6 分量重构为高频项，运用在对高频数据表现良好的 LSTM 模型进行高频项的预测，IMF7-IMF9 重构为低频项，运用在对低频数据表现良好的 ARIMA 模型进行低频项的预测。其中趋势项由于周期长，运用 ARIMA 模型进行趋势项的预测。

5.2.3 EEMD-LSTM-ARIMA 建模步骤

1.HIMF 建模过程

（1）数据预处理

本章输入数据为重构后的 HIMF 数据，对高频项数据进行归一化处理，划分训练集和测试集，按 8：2 划分数据，即 2054 条数据作为训练集，502 条数据作为测试集。

（2）确定模型参数

为了能与 LSTM 单一模型及 EEMD-LSTM 模型进行对比，本章依然选用隐藏节点数为 32，训练批次为 32，迭代次数为 150。定义优化器为 adam，损失函数为 mae 进行模型的训练。

（3）预测

将分解后重构的高频数据作为输入数据，经过 LSTM 模型的训练输出高频数据的预测值。

2.LIMF 建模过程

低频项数据运用 ARIMA 模型进行预测。

（1）对 LIMF 信号进行平稳性检验，ADF 检验若非平稳则 P 值大于 0.05，进而确定其差分的阶数：

（2）确定 ARMA 模型的阶数 p，q，并在初始估计中选择尽可能少的参数，在选择阶数时应根据 ADF 检验中差分的阶数进行确定；

AIC 准则：

假设X_t为 ARMA(p,q)模型，其中未知数的个数为 p+q+1 个，其中包括自回归系数$\varphi_1, \varphi_2, \dots, \varphi_p$，移动平均系数$\theta_1, \theta_2, \dots, \theta_q$和$\sigma_\epsilon^2$，那么 ARMA(p,q)的定阶准则为：

选取适当的 p 和 q 使得：

$$\mathrm{AIC} = \mathrm{n}\ln\widehat{\sigma_\epsilon^2} + 2(p+q+1) \qquad (6.1)$$

达到最小。

其中 n 为样本容量σ_ϵ^2与 p,q 有关。若当$\mathrm{p} = p', q = q'$时，AIC 值最小，则认为模型的阶数为p', q'，即为$\mathrm{ARMA}(p', q')$模型。

（3）选择模型，根据残差平方和，AIC 准则函数，DW 统计量等指标综合判断最终选定哪种模型。

（4）预测

首先判断 LIMF，RES 项是否为平稳序列，经 ADF 检验得出 LIMF 为平稳序列，因此，对 LIMF 低频项直接进行预测。其中根据 AIC 准则函数确定低频项的预测模型为 ARMA（5，3）。

3.RES 建模过程

趋势项数据运用 ARIMA 模型进行预测

（1）对 RES 信号进行平稳性检验，ADF 检验若非平稳则 P 值大于 0.05，需要进行差分，不同的数据选择不同的差分阶数：

（2）确定 ARMA 模型的阶数 p，q，根据拖尾结尾的标准选择合适的回归模型，并在初始估计中选择尽可能少的参数，在选择阶数时应根据 ADF 检验中差分的阶数进行确定；

（3）选择模型，根据残差平方和，AIC 准则函数，DW 统计量等指标综合判断最终选定哪种模型。

（4）预测

首先判断 RES 项是否为平稳序列，经 ADF 检验得出 RES 为平稳序列，因此，对 RES 趋势项不需要差分可直接进行预测。其中根据 AIC 准则函数确定趋势项预测模型为 ARMA（4，0）。

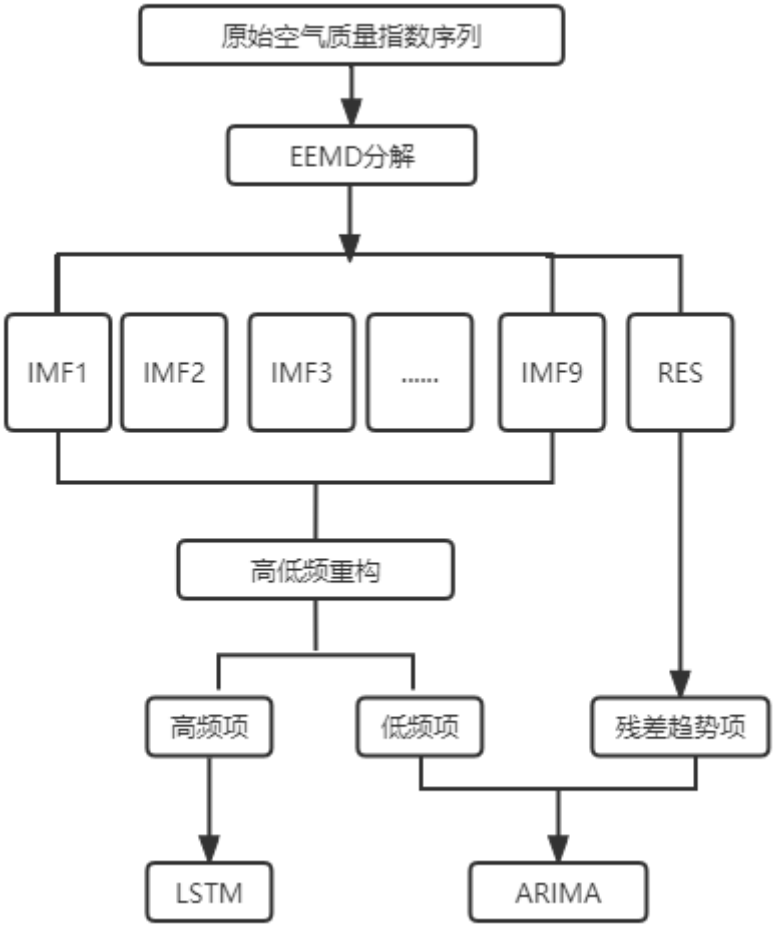

图 5.4 EEMD-LSTM-ARIMA 模型流程图

5.2.4 模型预测与结果分析

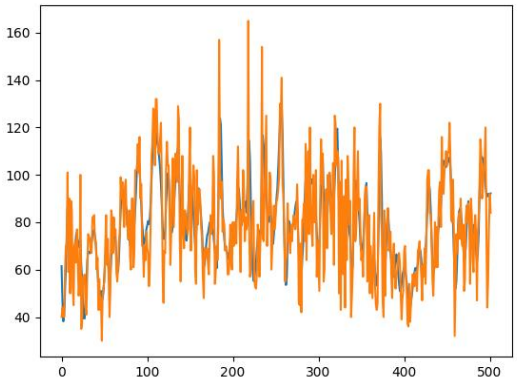

图 5.5 EEMD-LSTM-ARIMA 模型对 AQI 预测结果

由图 5.5 可看出 EEMD-LSTM-ARIMA 模型对 AQI 空气质量指数的预测效

果较好，虽然某些预测值有一定的误差，但整体趋势大致相同，且曲线在多个部分基本重合，说明组合模型的预测曲线和真实曲线整体拟合度较好。

由表 5.4 可以更加具体的看到，组合模型的误差有叫大幅度的递减，其中 RMSE 为 12.89，MAE 为 9.46，拟合优度为 0.65。可以看出 EEMD-LSTM-ARIMA 模型的预测效果良好。

表 5.4 EEMD-LSTM-ARIMA 模型的效果评估

	RMSE	MAE	R^2
EEMD-LSTM-ARIMA	12.89	9.46	0.65

5.3 各模型预测结果对比

根据上述几章建立的单一模型和组合模型，即单一变量的 LSTM、EEMD-LSTM 组合模型、EEMD-LSTM-ARIMA 组合模型，观察预测值与测试值的图像拟合优度以及评价参数，可以看出 EEMD-LSTM-ARIMA 组合模型的预测效果最好。

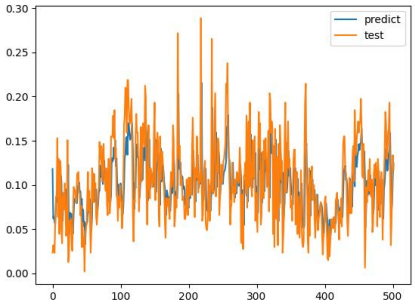

图 5.6 单一因素 LSTM 模型预测结果

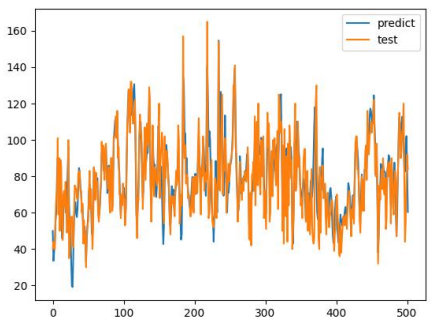

图 5.7 EEMD-LSTM 模型预测结果

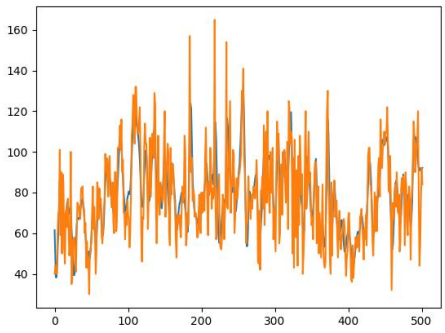

图 5.8 EEMD-LSTM-ARIMA 模型预测结果

图 5.6-5.8 中，蓝色线是预测值，橘色线是真实值，长短期记忆网络在预测

时由于数据波动较大且存在噪声的干扰，模型在学习时不能非常准确的抓取数据的波动趋势，导致预测结果会造成一定偏差，而 EEMD-LSTM 模型因为对数据分解，将原始数据中的噪声去除，通过每一个分量的预测值加总得到最终预测值，在这个过程中由于每一各分量的预测都会产生误差，虽然最终预测结果好于 LSTM 模型，但对于某些极端值的预测不能良好的反映。EEMD-LSTM-ARIMA 针对不同的数据特点建立预测模型，将每个模型的优势放大，在分解数据后运用在高频数据预测中表现良好的长短期记忆网络模型，低频项和趋势项则用时间序列模型进行模拟。经过不同模型的组合，对于某些数据的预测有了一定改善，最终的曲线拟合度也好于其他两个模型。

表 5.5 各模型预测值

真实值	LSTM	EEMD-LSTM	EEMD-LSTM-ARIMA
40	85.712	81.559	51.143
44	63.011	54.794	34.182
41	62.977	56.879	38.647
40	60.329	55.245	44.087
……	……	……	……
55	65.242	54.720	61.365
78	69.162	62.005	98.775
92	78.734	76.183	89.793
84	85.099	84.036	84.071

由表 5.5 可以看出，在各模型对空气质量指数进行预测的结果中，组合模型的效果高于单一模型，而组合模型中尤 EEMD-LSTM-ARIMA 表现最佳，虽然仍然有一定的差距，但由于大气中污染物含量较多，导致 AQI 数据的影响因素较多，预测较为复杂，在运用模型对其规律进行模拟时，会产生些许误差，但整体趋势大致相同。

表 5.6 各模型预测结果评估

模型	RMSE	MAE	R^2
单一因素 LSTM	18.94	14.34	0.52
EEMD-LSTM	14.06	10.86	0.53
EEMD-LSTM-ARIMA	12.89	9.46	0.65

由表 5.6 可以看出，EEMD-LSTM-ARIMA 模型对 AQI 空气质量指数的预测最为精准，RMSE 为 12.89，比单一 LSTM 模型精度提高了 31.94%，拟合优度也提高了 25%，可见 EEMD-LSTM-ARIMA 模型可以为今后空气质量的预测提供有力依据。

5.4 其他数据集预测结果

为了验证模型，对于其他非平稳、非线性数据的空气质量数据也具有预测的有效性，因此，本节选取于 AQI 空气质量数据相关程度较大的$PM_{2.5}$和PM_{10}数据进行预测，因两种数据与 AQI 空气质量指数 数据拥有相似的数据结构，也具有较大的波动性和非线性，因此按照相同的步骤与流程，即利用集成经验模态分解将数据进行信号分解，接着通过数据重组，将分解后信号重构为高频项、低频项和趋势项，作为 LSTM 模型的输入数据，采用相同的训练集与测试集划分比例，将结果进行对比，来进一步验证模型的普适性和有效性。

5.4.1 $PM_{2.5}$数据集验证

由于$PM_{2.5}$属于常见的危害较大的空气污染物，长期存在会导致空气治理下降,影响人们的身体健康甚至会产生慢性疾病。AQI 空气质量指数波动影响受$PM_{2.5}$影响较大，且$PM_{2.5}$数据也具有 AQI 指数类似的数据特征，即具有非线性、非平稳的特征。因此选取$PM_{2.5}$数据可以有效证明该组合模型的普适性，为今后的空气质量预测提供科学依据。选取 2014 年到 2020 年的$PM_{2.5}$日度数据，有效数据为 2556 条。

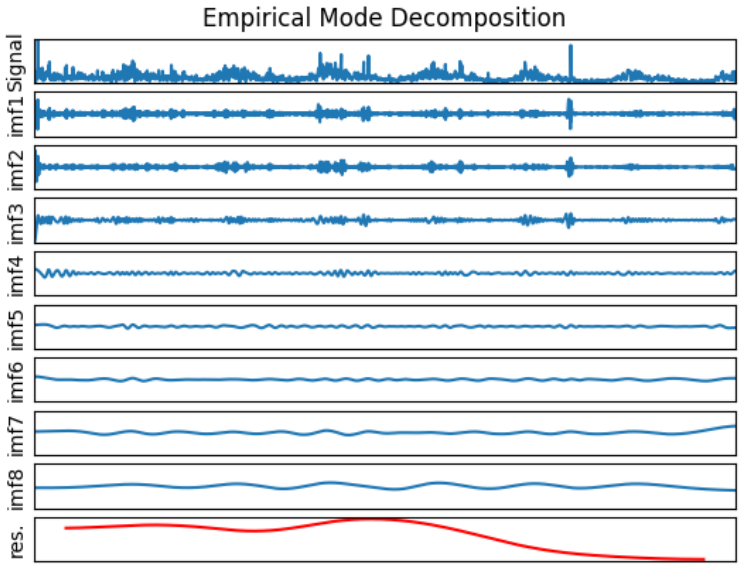

图 5.9 $PM_{2.5}$数据集 EEMD 分解结果

对$PM_{2.5}$数据集运用 EEMD 方法将原始信号分解，分解之后的数据可以分为，8 个 IMF 分量以及 1 个趋势项，分解结果如图 5.9 所示，可见差分后的各分量的波动频率越来越低，其中最上面曲线为原始数据，从上到下依次是各 IMF 分量，最后一条曲线则为趋势项。

$PM_{2.5}$数据分解为 8 个信号和 1 个趋势项，由于分解后的 IMF 过多，过多的分解信号会在后续的预测过程中增添不必要的麻烦，也可能会产生些许误差，从而给预测精度带来影响。因此，本节针对分解后的 IMF 分量进行 T 检验重构，若 p 值大于 0.05 则表示该分量具有良好的平稳性，将具有平稳性的数据划分为高频数据。若 p 值小于 0.05 则表示该分量具有非平稳性，应划分为低频数据。对于趋势项不进行划分，最后将各个分量划分为高频项、低频项和趋势项。各分量 T 检验结果如表 5.7 所示：

表 5.7 $PM_{2.5}$数据集各分量 T 检验结果

IMF	T 统计值	P
IMF1	-1.98	0.843
IMF2	-1.92	0.848
IMF3	-2.514	0.312
IMF4	0.797	0.426
IMF5	2.350	0.019
IMF6	5.481	0.001
IMF7	10.027	0.001
IMF8	-7.058	0.001
RES	303.466	0.001

由表 5.7 可知，IMF1-IMF4 分量的 T 检验 p 值均大于 0.05，证明 IMF1-IMF4 分量具有良好的平稳性。因此，将 IMF1-IMF4 定义为高频项，IMF5-IMF8 分量的 T 检验 p 值均小于 0.05，因此将 IMF5-IMF8 定义为低频项。剩余项作为趋势项。其中各分量的统计特征如表 5.8 所示：

表 5.8 各分量统计特征

IMF	平均周期（天）	Pearson	方差	方差贡献率（%）
IMF1	5.879	0.338	181.056	17.005
IMF2	10.463	0.342	166.775	15.672
IMF3	19.054	0.217	168.618	12.822
IMF4	34.779	0.233	54.186	11.406
IMF5	44.341	0.190	24.319	10.601
IMF6	121.712	0.191	30.421	9.365

续表 5.8 各分量统计特征

IMF7	182.472	0.171	119.965	8.921
IMF8	426.033	0.047	204.524	7.326
RES	1278.336	0.003	62.674	6.881

由表 5.8 可知，较高频率分项周期短具有较强的随机性，而较低频率分项周期长具有较强的周期性。根据各分量与原$PM_{2.5}$序列的 Pearson 相关性可知 IMF2 与原序列相关性最高为 0.342，其次是 IMF1 为 0.338。且 IMF1-IMF4 分量的方差贡献率高达 56.905%。因此，本文确定 IMF1-IMF4 分量重构为高频项，运用在对高频数据表现良好的 LSTM 模型进行高频项的预测，IMF5-IMF8 重构为低频项，运用在对低频数据表现良好的 ARIMA（4，3）模型进行低频项的预测。其中趋势项由于周期长，运用 ARIMA（2，6）模型进行趋势项的预测。

为对比模型的普适性和精准性，将单一 LSTM 模型、EEMD-LSTM 组合模型、EEMD-LSTM-ARIMA 模型进行对比，具体预测数值如表所示：

表 5.9 各模型$PM_{2.5}$预测值

真实值	LSTM	EEMD-LSTM	EEMD-LSTM-ARIMA
32	20.544	20.695	26.331
12	26.529	25.368	17.621
17	16.999	10.784	14.908
20	13.439	18.651	18.587
……	……	……	……
21	45.631	49.253	34.215
28	35.634	35.503	29.525
46	38.393	38.477	54.659
64	48.058	47.805	61.513

由表 5.9 可以看出在与各个模型的预测值对比后，发现 EEMD-LSTM-ARIMA 模型的预测值更为接近真实值，而且整体趋势也较为接近，与 LSTM、EEMD-LSTM 模型相比，后者的预测值虽然有个别值与真实值相差不多，但也有些值有所偏离，可能是由于$PM_{2.5}$空气质量指数变动的影响因素过多，数据较为复杂，导致模型在抓取数据规律时产生波动，造成预测有所偏差，而 EEMD-LSTM-ARIMA 模型的预测趋势较为稳定，所以，运用组合模型来预测$PM_{2.5}$也是可行的。

表 5.10 各模型预测结果评价

模型	RMSE	MAE	R^2
单一因素 LSTM	9.97	7.03	0.58

续表 5.10 各模型预测结果评价

EEMD-LSTM	9.88	6.96	0.59
EEMD-LSTM-ARIMA	9.22	6.59	0.66

由表 5.10 可以看出，EEMD-LSTM-ARIMA 模型对$PM_{2.5}$数据的预测最为精准，RMSE 为 9.22，比单一 LSTM 模型精度提高了 8.1%，拟合优度也提高了 13.8%，可见 EEMD-LSTM-ARIMA 模型也适用于对$PM_{2.5}$数据的预测。

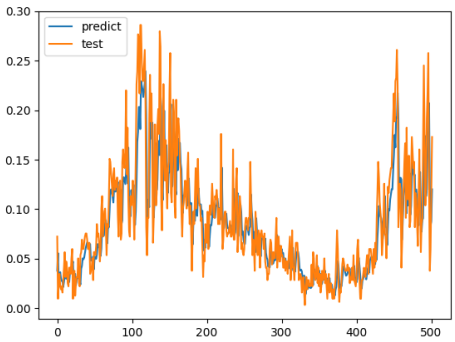

图 5.10 LSTM 模型预测结果

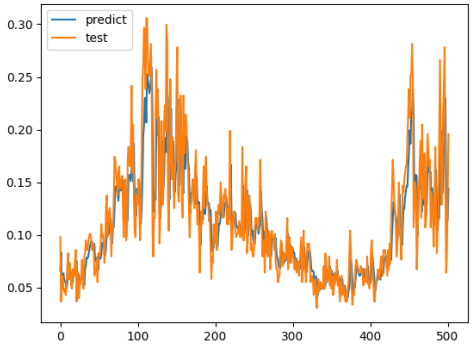

图 5.11 EEMD-LSTM 组合模型预测结果

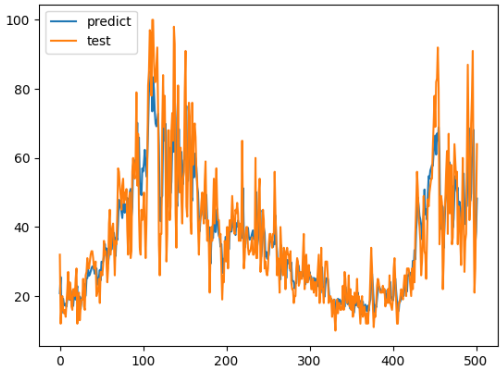

图 5.12 EEMD-LSTM-ARIMA 组合模型预测结果

由图 5.10-5.12 可见，EEMD-LSTM-ARIMA 模型在对$PM_{2.5}$数据集的预测上也有良好表现。虽然三个模型的预测值曲线都与真实值曲线拟合度较好，但可以看出组合模型的曲线重合度更好，橙色曲线表示真实值，蓝色曲线为预测值，组合模型的蓝色和橙色的重合度更高，而且通过各评价参数也可直观的看出，组合模型的预测误差较小。

5.4.2 PM_{10}数据集验证

PM_{10}对空气质量的影响也较大，AQI 空气质量指数波动影响受PM_{10}影响较大，且PM_{10}数据也具有 AQI 指数类似的数据特征，即具有非线性、非平稳的特

征。因此选取PM_{10}数据可以有效证明该组合模型的普适性，为今后的空气质量预测提供科学依据。选取 2014 年到 2020 年的PM_{10}日度数据，有效数据为 2556 条。

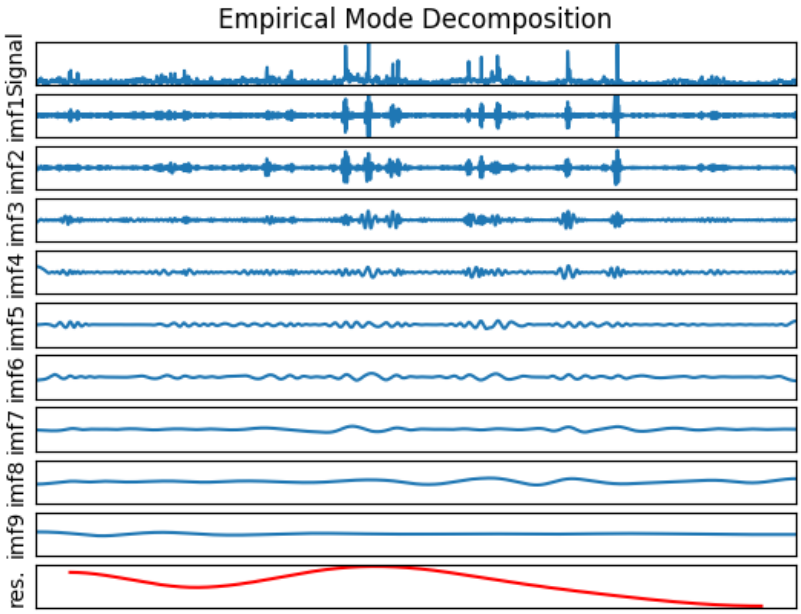

图 5.13 PM_{10}数据集 EEMD 分解结果

对PM_{10}数据集进行 EEMD 分解，得到 8 个 IMF 分量以及 1 个趋势项，分解结果如图 5.13 所示，可见差分后的各分量的波动频率越来越低，其中最上面曲线为原始数据，从上到下依次是各 IMF 分量，最后一条曲线则为趋势项。

PM_{10}数据分解为 9 个信号和 1 个趋势向，分解后的 IMF 过多，不利于后续预测计算。因此，本节针对分解后的 IMF 分量进行 T 检验重构，若 p 值大于 0.05 则表示该分量具有良好的平稳性，划分为高频数据。若 p 值小于 0.05 则表示该分量具有非平稳下，应划分为低频数据。对于趋势项不进行划分，最后将各个分量划分为高频项、低频项和趋势项。各分量 T 检验结果如表 5.11 所示：

表 5.11 PM_{10}数据集各分量 T 检验结果

IMF	T 统计值	P
IMF1	-0.060	0.952
IMF2	-2.331	0.207
IMF3	-0.683	0.495
IMF4	1.675	0.094
IMF5	-0.830	0.047
IMF6	-2.681	0.007

续表 5.11 PM_{10}数据集各分量 T 检验结果

IMF7	-5.563	0.001
IMF8	0.471	0.002
IMF9	-1.277	0.002
RES	292.311	0.001

由表 5.11 可知，IMF1-IMF4 分量的 T 检验 p 值均大于 0.05，因此，将 IMF1-IMF4 定义为高频项，IMF5-IMF9 分量的 T 检验 p 值均小于 0.05，因此将 IMF5-IMF9 定义为低频项。剩余项作为趋势项。其中各分量的统计特征如表 6.12 所示：

表 5.12 各分量统计特征

IMF	平均周期（天）	Pearson	方差	方差贡献率（%）
IMF1	3.233	0.318	2694.747	14.759
IMF2	9.671	0.204	2280.804	14.347
IMF3	21.495	0.290	1291.625	12.213
IMF4	23.765	0.057	954.762	9.679
IMF5	45.622	0.066	568.693	9.605
IMF6	106.293	0.111	659.053	9.080
IMF7	284.975	0.094	553.321	8.707
IMF8	319.309	0.026	1016.698	8.300
IMF9	511.323	0.018	242.065	7.334
RES	1278.654	0.023	427.865	5.977

由表 5.12 可知，较高频率分项周期短具有较强的随机性，而较低频率分项周期长具有较长的周期性。根据各分量与原PM_{10}序列的 Pearson 相关性可知 IMF1 与原序列相关性最高为 0.318，其次是 IMF3 为 0.290。且 IMF1-IMF4 分量的方差贡献率高达 50.998%。因此，本文确定 IMF1-IMF4 分量重构为高频项，运用在对高频数据表现良好的 LSTM 模型进行高频项的预测，IMF5-IMF9 重构为低频项，运用在对低频数据表现良好的 ARIMA（5，3）模型进行低频项的预测。其中趋势项由于周期长，运用 ARIMA（4，2）模型进行趋势项的预测。

为对比模型的普适性和精准性，将单一 LSTM 模型、EEMD-LSTM 组合模型、EEMD-LSTM-ARIMA 模型进行对比，具体预测数值如表所示：

表 5.13 各模型PM_{10}预测值

真实值	LSTM	EEMD-LSTM	EEMD-LSTM-ARIMA
88	56.050	63.102	79.491
25	49.652	43.214	33.184
39	40.601	23.805	36.536
41	49.508	28.067	43.926
……	……	……	……

续表 5.13 各模型PM_{10}预测值

44	57.405	15.514	40.123
60	68.205	74.847	67.637
99	78.984	84.575	101.451
120	104.925	106.722	110.363

由表 5.13 可知，在对PM_{10}的预测结果看，会比$PM_{2.5}$的预测结果相对差一些，但是在与各个模型进行对比中可以看到，组合模型的预测值与实际值的误差，仍然高于其他模型，说明组合模型仍然会较单一模型拥有较高的精准度也更为稳定，从而也适用于PM_{10}值的预测。

表 5.14 各模型预测结果评价

模型	RMSE	MAE	R^2
单一因素 LSTM	28.042	19.429	0.41
EEMD-LSTM	26.587	18.320	0.47
EEMD-LSTM-ARIMA	18.636	14.316	0.58

由表 5.14 可以看出，EEMD-LSTM-ARIMA 模型对PM_{10}数据的预测最为精准，从图 5.14-5.16 中也同样可以看出图像的拟合度上，组合模型预测值曲线的重合部分高于其他模型，曲线具有较高的拟合度。且组合模型的 RMSE 为 19.636，比单一 LSTM 模型精度提高了 42.8%，拟合优度也提高了 41.5%，可见 EEMD-LSTM-ARIMA 模型也适用于对$PM_{2.5}$数据的预测。

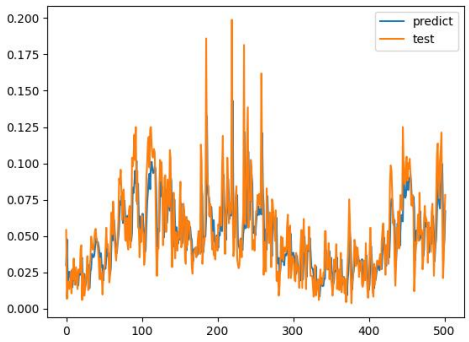

图 5.14 LSTM 模型预测结果

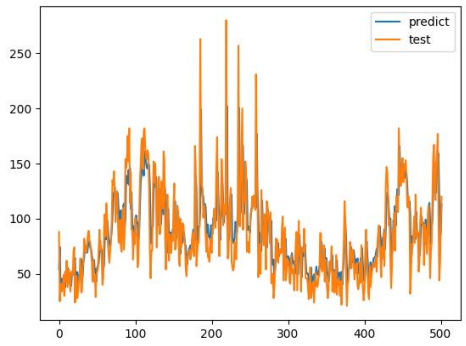

图 5.15 EEMD-LSTM 组合模型预测果

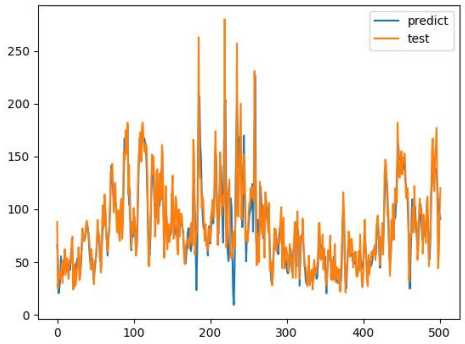

图 5.16 EEMD-LSTM-ARIMA 组合模型预测结果

由图 5.13-5.15 可知，在预测曲线的拟合度上，可直观的看出 EEMD-LSTM-ARIMA 模型的曲线重合度更高，虽然在个别极端值上出现偏差，但对于其他值的预测还是较为精准，在重合度上好于其他两种模型，说明组合模型对PM_{10}数据集的预测效果较好。综上所述，EEMD-LSTM-ARIMA 模型也适合对PM_{10}数据集的预测。

由上述两种不同数据集的验证，可以相对充实的证明本文所探究的组合模型，对空气质量数据的预测是有效的，相比长短期记忆网络模型和 EEMD-LSTM 组合模型，EEMD-LSTM-ARIMA 模型提供了更准确，误差更小的预测值，可以作为未来空气质量预测的有力工具。

5.5 模型实际应用

为检验模型的实用性，本节选取 2021 年 1 月 1 日至 2022 年 2 月 28 日数据，建立 EEMD-LSTM-ARIMA 模型，对 3 月 1 日至 3 月 7 日的空气质量指数进行预测，预测结果如下所示：

表 5.15 3 月 1 日-3 月 7 日 AQI 预测值

真实值	预测值	误差
104	106.724	2.619%
77	74.692	2.997%
73	73.757	1.037%
69	64.315	6.789%
81	86.089	6.282%
76	79.879	5.104%

从表 5.15 中可以看出，对未来一周的空气质量指数进行预测时，误差值较小，可以证明 EEMD-LSTM-ARIMA 模型可以用于预测空气质量数据，虽然有所

误差，但并不影响人们的判断，一般人们判断空气质量等级时依据空气质量指数的范围，每一等级的范围为 50，因此，即使空气质量数据有一定的偏差，但并不影响空气质量等级的判断，因此，该模型具有一定的参考价值。

6 结论与展望

6.1 结论

本文基于兰州市 2014 年到 2020 年的空气质量数据，探究兰州市空气质量的演变规律，并通过不同模型的对比，找出预测精度较高的模型对空气质量指数进行预测为后续的空气质量预测提供参考，且为了验证模型的普适性和实用性，增加了其他数据集的预测验证，之后将模型投入实际运用，具体的研究成果如下：

6.1.1 兰州市空气质量呈季节性和周期性变化

根据文中对空气质量年度特征的分析中可以看出，兰州市空气质量呈季节性和周期性变化。在 2014 年到 2017 年间，大气污染浓度及空气质量指数逐渐上升，在 2017 年时空气质量变化出现拐点，自 2017 年后产生下降趋势。且兰州市的空气质量呈现明显的周期性，冬天用来取暖的燃料燃烧加上汽车尾气的排放，使得空气中污染物浓度增多，导致冬天的空气质量多数处于轻度污染或重度污染。而夏季和秋季的重度污染天气较少，空气质量较好，这得益于兰州市采取了一些治理措施，像路面洒水，人工降雨等使空气中水蒸气含量增加，空气相对湿度较大，会加速可吸入颗粒等空气污染物的沉淀和扩散，从而减少空气中大气污染浓度。

6.1.2 组合模型对空气质量的预测效果较理想

为了在今后能够在一定程度上掌握空气质量的变化规律，本文针对 AQI 空气质量指数进行预测，在现有模型的基础上，选取三个效果较好，普遍使用的模型进行对比，优中选优，将单一模型中表现良好的模型用于组合。分别建立单一因素的 LSTM 模型、EEMD-LSTM 组合模型以及 EEMD-LSTM-ARIMA 组合模型对 AQI 空气质量指数进行预测，通过对比预测结果发现，EEMD-LSTM-ARIMA 模型的预测值更接近真实值，预测效果好于单一模型，经过各个模型的预测结果的对比发现，EEMD-LSTM-ARIMA 模型对空气质量的预测精度最为准确，且未

了验证模型的普适性，还将模型应用于$PM_{2.5}$和PM_{10}数据的预测中，通过对不同数据集的分析预测，验证了组合模型的实用性，且预测误差也在可接受范围内，虽仍有一定瑕疵，但相比于经典模型，预测精度有所提高，可以为今后的预测提供有力工具。

6.2 展望

本文主要针对空气质量的模型展开对比和改进，在最终的实验和应用结果中取得了一定的成效。虽然在一定程度上改善了预测效果，但在模型的数据选择，参数优化上仍然存在不足，需要对其进行有针对性的修改和完善。在接下来的研究分析中需要增加对以下几点的考虑的完善：

首先，数据方面。本章的实验使用的是兰州市某一个站点的空气质量数据，不能将各个区县的空气质量数据全部涵盖在内，没有将各区县的空气质量进行对和预测，缺少了一定的代表性。除此此外，污染物的扩散和变化是一个时空过程，如果在模型内加入空间因素或者其他因素，可能会对预测结果的精度有所提高。同时本次的训练数据集规模较小，仅有 2000 多个数据，可能无法全面的抓取数据的波动性和规律，在后续的研究中应该加大样本的数量。同时也要拓宽研究的广度和深度，增加多维的研究角度，如将数据划分为不同季节、不同监测点，以及不同的区域等，通过不同维度的考量，来提高训练模型的准确度。

其次，影响因素方面。由于 AQI 空气质量指数是根据空气中多种污染物的含量计算的，涉及变量过多，数据的波动程度大，所以在模型进行学习时，始终无法抓取最准确的波动规律。除此之外，气象因素，经济因素都有可能对空气质量产生影响。在今后的研究中，需要增加多种影响因素，提高模型对空气质量指数的预测效果。

最后，模型预测方面。本文虽然引入继承经验模态分解与时间序列模型进行改进，相对传统模型预测的准确度也有了提高；并且与其他经典深度学习模型相比，预测的误差较小，也更适合于空气质量数据的预测。但在模型的优化算法上可以试着从鲸鱼算法、灰狼算法、粒子群优化等算法上对模型参数优化等这些方面来对模型进行调整，更加科学严谨的训练模型，期望能够拥有误差更小的预测值，为空气质量预测提供更优的预测模型，从而改善空气质量。

参考文献

[1]尤莉,李彰俊,徐桂梅,等. 呼和浩特市空气污染潜势预报方法研究[J]. 环境保护, 2003, 15(3): 12-14.

[2]赵惠芳,陈雅莲,唐会荣,等. 晋江城市空气质量污染潜势统计预报方法初探[J]. 气象与环境学报, 2009, 25(5): 27-30.

[3]孙银川,缪启龙,李艳春,等.银川市空气质量动力预测系统及预测结果分析[J].干旱气象,2006(02):89-94.

[4] Jose R S, Juan L P, Jose L M, et al. European operational air quality forecasting system by using MM5-CMAQ-EMIMO tool[J]. Simulation Modelling Practice and Theory, 2008, 16(10): 1534–1540.

[5]Daegyun L,Daewon W B, Hyuncheol K, et al. Improved CMAQ predictions of particulate matter utilizing the satellite-derived aerosol optical depth[J]. Atmospheric Environment, 2011, 45(22): 3730-3741.

[6]沈劲,王雪松,李金凤,等. Models-3/CMAQ 和 CAMx 对珠江三角洲臭氧污染模拟的比较分析[J]. 中国科学, 2011, 41(11): 1750-1762.

[7]谢敏,钟流举,陈焕盛,等. CMAQ 模式及其修正预报在珠三角区域的应用检验[J]. 环境科学与技术, 2012, 35(2): 96-101.

[8]Liu D J, Li L. Application Study of Comprehensive Forecasting ModelBased on Entropy Weighting Method on Trend of PM2.5 Concentration inGuangzhou, China [J]. International journal of environmental research and publichealth, 2015, 12(6):7085-7099.

[9]Yang Z S, Wang J. A new air quality monitoring and early warningsystem: Air quality assessment and air pollutant concentration prediction[J].Environmental Research, 2017, 158:105-117.

[10]倪志伟,朱旭辉,程美英. 基于人工鱼群和分形维数融合的 SVM 的空气质量预测方法[J]. 模式识别与人工智能, 2016, 29(12): 1122-1131.

[11]杨涛锋,彭艺.基于改进 PSO 的 ARIMA-SVM 空气质量预测研究[J].云南大学学报(自然科学版),2020,42(05):854-862.

[12]Osowski, S, Garanty, K. Forecasting of the daily meteorological pollution using wavelets and support vector machine[J]. Engineering Applications of Artificial Intelligence.2007, 20(6):745-755.

[13]刘向丽，王旭朋.基于小波分析的股指期货高频预测研究.系统工程理论与实践.2015,(6)：1425-1432

[14]代军,叶幸玮.集合经验模式分解和小波变换方法的复合与应用[J].统计与决策,2021,37(13):155-158.

[15]王振华,刘晓丹,刘向锋.GLAS 全波形数据的小波与经验模态分解降噪[J].激光与光电子学进展,2021,58(23):364-371.

[16]Singla P,Duhan M,Saroha S. An ensemble method to forecast 24-h ahead solar irradiance using wavelet decomposition and BiLSTM deep learning network.[J]. Earth science informatics,2021,21(9):325-329.

[17]马宁,董泽,冯斌.基于信号分解和核极限学习机的风电功率预测[J].山东电力技术,2022,49(01):1-6.

[18]董小刚,周明眉,秦喜文,等．基于 EMD 的 PM_(2.5)趋势分析[J]．长春工业大学学报.2016,37(2):179-181.

[19]Jianfeng Z,Zhu Y,Xiaoping Z,et al.Developing a Long Short-Term Memory(LSTM) based model for predicting water table depth in agricultural areas[J].Journal of Hydrology,2018,561:918-929.

[20]刘铭,魏莱.EMD-LSTM 算法及其在 PM2.5 中的预测[J].长春工业大学学报,2020,41(04):322-327+417.

[21]涂锦,冷正兴,刘丁毅.基于 EMD 和神经网络的非线性时间序列预测方法[J].统计与决策,2020,36(08):41-44.

[22]金秀章,刘岳,于静,等.基于变量选择和 EMD-LSTM 网络的出口 SO_2 浓度预测[J].中国电机工程学报,2021,41(24):8475-8484.

[23]秦喜文,王强进,王新民,等.基于 VMD 和 LSTM 方法的北京市 PM_(2.5)短期预测[J].吉林大学学报(地球科学版),2022,52(01):214-221.

[24]Yang Z, Wang J. A new air quality monitoring and early warning system: air quality Assessment and air pollutant concentration prediction[J]. Environment,Research,2017,158:105-117.

[25]许德合,丁严,张棋,等.EEMD-ARIMA 在干旱预测中的应用——以新疆维吾尔自治区为例[J].中国农村水利水电,2021(07):1-11.

[26]史学良,李梁,赵清华.基于改进 LSTM 网络的空气质量指数预测[J].统计与决策,2021,37(16):57-60.

[27]Faruk D, Tolga E.Estimating national exhaust emissions from railway vehicles inTurkey [J].Science of The Total Environment, 2006, 374(1):127-134.

[28]Simone M .Influence of the public transportation systerm on the air quality of a major urban center.A case study : Milan ,Italy [J].Atmospheric Environment ,2008(42):7915-7923.

[29]李健,靳泽凡,苑清敏.京津冀空气质量环境库兹涅茨曲线及影响因素—基于2006—2017 年面板数据的分析[J].生态经济,2019,35(02):197-201+218.

[30]姜磊,周海峰,柏玲,等.中国城市空气质量指数(AQI)的动态变化特征[J].经济地理,2018,38(09):87-95.

[31]张如会.青岛市空气质量的影响因素分析及预测研究[D].青岛大学,2020.

[32]Aiiang T. Analysis on Factors Affecting the Air Quality in Beijing City Based on Grey Relation Theory[J].Advanced Materials Research,2014,3248(955-959):1583-1586.

[33] Zhou G Q, Xu J M, Xie Y, et al. Numerical air quality forecasting over eastern China: An operational application of WRF-Chem[J]. Atmospheric Environment, 2017, 153(9): 94-108.

[34]周秀杰,苏小红,袁美英. 基于 BP 网络的空气污染指数预报研究[J]. 哈尔滨工业大学学报, 2004, 36(5): 582-585.

[35]Gu K Y, Zhou Y, Sun H, et al. Prediction of air quality in Shenzhen based on neural network algorithm[J]. Neural Computing and Applications, 2019, 32(2):1-14.

[36]Chen L, Ding Y F, Lyu D D,et al. Deep Multi-Task Learning Based Urban Air Quality Index Modelling[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies,2019,3(1):1359-1367.

[37] Woody M C, Wong H W, West J J, et al. Multiscale predictions of aviation-attributable PM2.5 for U.S. airports modeled sing MAQ with plume-in-grid and an aircraft-specific 1-D emission model[J]. Atmospheric Environment, 2016,

147(13): 384-394.

[38]郑洋洋,白艳萍,侯宇超.基于 Keras 的 LSTM 模型在空气质量指数预测的应用[J].数学的实践与认识,2019,49(07):138-143.

[39]Vlachogianni A, Kassomenos P, Karppinen A. Evaluation of a Multi-ple Regression Model for the Forecasting of the Concentrations of NOx and PM10in Athens and Helsinki [J].Science of the Total Envi-ronment,2011,(8):119-127.

[40]Zhang X K, Zhang Q W, Zhang G, et al. A Novel Hybrid Data-Driven Model for Daily Land Surface Temperature Forecasting Using Long Short-Term Memory Neural Network Based on Ensemble Empirical Mode Decomposition[J]. International Journal of Environmental Research and Public Health, 2018, 15(5):1032-1032.

[41]Hao Z Z, Zhou Z, Russo A, et al. Impact of meteorological conditions at multiple scales on ozone concentration in the Yangtze River Delta.[J]. Environmental Science and Pollution Research, 2021, 28(44):62991-63007.

[42]刘永,郭怀成. 城市大气污染物浓度预测方法研究[J]. 安全与环境学报, 2004, 4(4): 60-63.

[43]方雪清,吴春胤,俞守华,等.基于 EEMD-LSTM 的农产品价格短期预测模型研究[J].中国管理科学,2021,29(11):68-77.

[44]Mrigank K, Srinidhi J, Jew D, et al. Air quality modelling using long short-term memory (LSTM) over NCT-Delhi, India[J]. Air Quality, Atmosphere & Health, 2019, 12(8):899-908.

[45]Pei Y, Lei Z L, Zeng Q H, et al. Load forecasting of refrigerated display cabinet based on CEEMD-IPSO-LSTM combined model[J]. Open Physics, 2021, 19(1):360-374.

[46]William H B, Casey D B, Viney P A, et al. Evaluating ammonia (NH3) predictions in the NOAA National Air Quality Forecast Capability (NAQFC) using in situ aircraft, ground-level, and satellite measurements from the DISCOVER-AQ Colorado campaign[J]. Atmospheric Environment, 2016, 140(15): 342-351.

[47]湛社霞,匡耀求,阮柱.基于灰色关联度的粤港澳大湾区空气质量影响因素分析

[J].清华大学学报(自然科学版),2018,58(08):761-767.

[48]高帅,胡红萍,李洋,等.基于改进的思维进化算法与 BP 神经网络的 AQI 预测[J].数学的实践与认识,2018,48(19):151-157.

[49]黄婕,张丰,杜震洪,等.基于 RNN-CNN 集成深度学习模型的 PM_(2.5)小时浓度预测[J].浙江大学学报(理学版),2019,46(03):370-379.

[50]郑洋洋,白艳萍,侯宇超.基于 Keras 的 LSTM 模型在空气质量指数预测的应用[J].数学的实践与认识,2019,49(07):138-143.

[51]黄厘博. 重庆市主城区空气质量分析及 PM_(2.5)浓度预测[D].西南大学,2020.

[52]李政毓.基于 EEMD-ARIMA-LSTM 组合模型对原油期货价格预测[D].山东大学,2021.

[53]袁燕,陈伯伦,朱国畅,等.基于社区划分的空气质量指数（AQI）预测算法[J].南京大学学报(自然科学),2020,56(01):142-150.

[54]李婷婷,田瑞琦,汪漂.基于经验模态分解的空气质量指数组合预测方法及应用[J].价值工程,2019,38(16):134-138.

[55]罗上学,张美玲,聂雅梅,等.基于 CEEMDAN-LSTM 模型的郑州市月降水量预测[J].水利规划与设计,2022(02):45-50.

后记

时光飞逝，回顾过往好似才刚刚踏入此时的校园，3 年的研究生生活是充实而又忙碌的，这段期间的学习让我提高了专业水平，锻炼了实践能力，如今即将毕业，百感交集，但最多的还是感激。

首先我要感谢我的导师，老师无论在生活中还是学习上对我们的帮助都是巨大的，老师和善细致，不仅在学术上严谨认真，在为人处事上更是使我获益匪浅。在论文题目与后续撰写中，帮我缕清了写作思路，在定稿之后，又悉心提出了针对性的建议，给予我很多帮助让我顺利完成论文。对此，我深表感激。

同时，我要感谢能在百忙之中抽出时间审阅论文的教授和专家们，感谢您们能够严谨对待我们的论文，精益求精，让论文更加完善。

其次，我要感谢在硕士期间的同门与同窗。感谢师兄师姐给予我学习生活经验，感谢师弟师妹们带给我的新鲜活力也感谢他们能够帮我分担压力，感谢室友们日常生活中给予我的照顾，同时也要感谢同窗们对我学习上的各种帮助，也感谢他们在这段学术生涯中的陪伴。

最后，我要感谢我的家人。感谢我的父母能够一直支持我，相信我。感谢他们一直以来都无私奉献。感谢我的家人们一直的鼓励与爱护。以前他们是我的避风港，以后我将继续努力，成为他们强有力的依靠。