分类号_____　　密级 _____
U D C _____　　编号 10741

硕 士 学 位 论 文

论文题目 基于“分解-聚类-集成”的 $PM_{2.5}$ 时空预测研究及其应用

研 究 生 姓 名：周尧民

指导教师姓名、职称：黄恒君 教授

学 科、专 业 名 称：应用经济学 统计学

研 究 方 向：经济与社会统计

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：周龙民 签字日期：2022年5月30日

导师签名：黄柏素 签字日期：2022年5月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：周龙民 签字日期：2022年5月30日

导师签名：黄柏素 签字日期：2022年5月30

Spatio-temporal prediction of $PM_{2.5}$ based on decomposition-clustering-integration and its application

Candidate: Yaomiin Zhou

Supervisor: Hengjun Huang

摘 要

随着城市的发展、人口的集聚，城市汽车保有量持续增加，周边工厂排放的空气污染物等，造成城市环境恶化，居民出行以及健康状况受到严重影响。由此，利用城市空气质量数据、气象数据、空间 POI 等等城市大数据，构建精准的空气质量模型，从而更好的帮助居民制定出行计划，辅助政府制定环保决策。

在构建空气质量预测模型中从时间维度和空间维度共同出发，不仅丰富了研究角度，并且将数据融合的构想运用在研究中，将时间序列与空间信息进行了融合。本文以$PM_{2.5}$污染物浓度为例，探究$PM_{2.5}$序列在时间维度与空间维度的特征提取方式，将其纳入预测模型，并将时间与空间维度的预测结果动态地结合，提升预测效果。主要工作如下：

第一，探究时空模型理论算法，包括数据缺失值、离群点的处理，以及对运用相关性理论进行特征选取。在此基础上，本文深入研究了$PM_{2.5}$预测的各类前沿算法：模态分解、时间序列聚类、深度神经网络等，构建$PM_{2.5}$时空预测模型理论架构。

第二，构建时空预测模型，在分析$PM_{2.5}$预测在时间维度和空间维度特性的基础上，分别从两个维度出发构建时间与空间预测器。在时间维度上，利用模态分解提取$PM_{2.5}$数据波动特征，运用时间序列聚类算法对分量进行重构，并基于 ELSTM 模型构建时间预测器；在空间维度运用拉普拉斯算子从图模型角度提取站点的空间关系，以此构建空间预测器；最后，运用 XGBoost 将两部分结果进行动态聚合，完成 LX-M-CEEMDAN-VMD-LSTM 模型的构建。

第三，利用兰州市空气污染物浓度数据、气象数据以及地理信息对$PM_{2.5}$浓度序列进行预测。在时间预测模块中，运用 CEEMDAN 与 VMD 分解构建二层分解方法提取时序信息，并进行聚类重构，不仅提高了时序序列预测的精度，而且运用聚类进行数据重构进一步简化了模型。在空间预测模块中，拉普拉斯矩阵有效的提取数据的空间特征，提升了空间预测精度。基于 XGBoost 提取各类特征重要性，则将时间与空间特征动态结合，弥补了各自维度的不足。并以均方根误差(RMSE)、绝对值误差(MAE)和和平均绝对误差百分比（MAPE）三个评价指标以及 DM 检验对比此模型的优越性、有效性。结果表明，本文所构建模型的预测精度显著提高，在各项指标中均优于对照模型。

关键词：时空预测 模态分解 时间序列聚类 拉普拉斯算子 LSTM 神经网络

Abstract

With the development of the city and the agglomeration of the population, the number of cars in the city continues to increase, and the air pollutants emitted by the surrounding factories have caused the deterioration of the urban environment, and the travel and health of the residents have been seriously affected. Therefore, using urban big data such as urban air quality data, meteorological data, spatial POI, etc., to build an accurate air quality model, so as to better help residents make travel plans and assist the government to make environmental protection decisions.

Starting from the time dimension and the space dimension in the construction of the air quality prediction model not only enriches the research angle, but also applies the concept of data fusion in the research, and integrates the time series with the spatial information. Taking the $PM_{2.5}$ pollutant concentration as an example, this paper explores the feature extraction method of $PM_{2.5}$ sequence in the time dimension and space dimension, incorporates it into the prediction model, and dynamically combines the prediction results of the time and space dimensions to improve the prediction effect. The main work is as follows:

Firstly, explore the theoretical algorithms of spatiotemporal models, including the processing of missing data and outliers, and feature selection using correlation theory. On this basis, this paper deeply studies

various cutting-edge algorithms for $PM_{2.5}$ prediction: modal decomposition, time series clustering, deep neural network, etc., and constructs the theoretical framework of $PM_{2.5}$ spatiotemporal prediction model.

Secondly, build a spatiotemporal prediction model, and build a temporal and spatial predictor on the basis of analyzing the characteristics of $PM_{2.5}$ prediction in the time dimension and space dimension. In the time dimension, the modal decomposition is used to extract the fluctuation characteristics of $PM_{2.5}$ data, the time series clustering algorithm is used to reconstruct the components, and the time predictor is constructed based on the ELSTM model; In the spatial dimension, the Laplacian operator is used to extract the spatial relationship of the site from the perspective of the graphical model, so as to construct the spatial predictor; finally, XGBoost is used to dynamically aggregate the two parts of the results to complete the LX-M-CEEMDAN-VMD-LSTM model 's build.

Thirdly, the $PM_{2.5}$ concentration sequence is predicted by using lanzhou air pollutant concentration data, meteorological data and geographic information. In the time prediction module, CEEMDAN and VMD were used to construct a two-level decomposition method to extract time series information, and then cluster reconstruction was carried out, which not only improved the accuracy of time series prediction, but also

further simplified the model by clustering data reconstruction. In the space prediction module, the Laplace matrix can effectively extract the spatial features of data and improve the accuracy of space prediction. Based on XGBoost, the importance of various features is extracted, and the temporal and spatial features are dynamically combined to make up for the deficiency of their respective dimensions. Root mean square error (RMSE), absolute error (MAE) and mean absolute error percentage (MAPE) were used to compare the advantages and effectiveness of the model. The empirical results show that the prediction accuracy of the proposed model is significantly improved, and it is superior to the control model in all indicators.

Keywords： Spatiotemporal prediction; Modal decomposition; Time series clustering; Laplacian; LSTM neural network

目 录

1 绪论.....1
1.1 研究背景.....1
1.2 研究意义.....2
1.3 国内外研究综述.....3
1.4 研究内容及创新点.....7
1.5 研究结构安排.....8
2 理论基础与方法.....10
2.1 数据预处理.....10
2.2 模态分解.....11
2.2.1 集成经验模态分解（EEMD）.....11
2.2.2 自适应完备集成经验模态分解（CEEMDAN）.....12
2.2.3 变分模态分解（VMD）.....13
2.3 基于形状的时间序列聚类算法.....14
2.3.1 时间序列形状相似度.....14
2.3.2 时间序列形状提取.....14
2.3.3 基于形状的时间序列聚类.....16
2.4 ELSTM 神经网络.....17
2.5 粒子群算法.....18
2.6 本章小结.....20
3 时空预测模型构建.....21
3.1 模型架构.....21
3.2 时间预测模块.....22
3.3 空间预测模块.....23
3.3.1 拉普拉斯空间特征嵌入法.....24
3.3.2 模型构建.....25
3.4 预测结果聚合.....25

3.5 本章小结.....26
4 时空预测模型实证分析.....27
4.1 数据描述.....27
4.2 评判标准.....31
4.3 时空预测模型实证.....32
4.3.1 时间预测流程及结果.....33
4.3.2 空间预测流程及结果.....39
4.3.3 时空模型集成结果.....40
4.4 模型比较及鲁棒性.....41
4.4.1 模型比较.....42
4.4.2 鲁棒性分析.....44
4.5 本章小结.....45
5 总结与展望.....46
5.1 研究工作总结.....46
5.2 预测系统构想.....47
5.3 未来工作展望.....47
参考文献.....49
附 录.....53
后 记.....56

1 绪论

当前社会对环境的重视与日俱增，大气污染的治理，空气质量的提升不仅被专家学者一次次的强调，大众也已经广泛的认同并成为社会关注的问题。而空气监测站能够实时监测当地区域的空气质量，但未能对未来空气质量的情况进行准确的预报，结果的不准确会致使出现严重污染而无法及时进行预防和治理。目前已有的预测方法在空气质量预测中，对其影响因素并没有充分考虑，尤其是在地理区位影响以及经济发展状况等并没有纳入模型计量，往往仅依据数据自身序列特征进行预测。因此充分考虑多种影响因素，做好空气质量的准确预测，亟待深入研究。

1.1 研究背景

在经济与人口快速增长的带动下，工业化水平不断提升，城市地域不断扩大，化石燃料排放增多的同时产生了各种有害废气。这也进一步使得空气中的有害物质急剧增多，出现雾霾天气，极大的危及了人们的健康以及自然环境。环保部门也已开始高度关注此类问题，并不断加大相关投入力度，以加强对环境污染的监测，并对大气污染严重超标的区域进行整治，着力控制温室气体以及其他污染废气的排放。此举虽然在一定层面上改善了空气质量水平，但各地为促进经济发展，保证城市发展水平，对于空气污染的治理水平仍然十分落后，甚至部分地区仍以破坏环境为代价推动经济的增长，这样的短视行为必定会造成大气污染，空气质量严重超标等不可逆问题出现。所以我们不仅要从治理段着手，而且需要从检测端做好预防，而反观监测环节，一方面，监测站的建设所需资金数目庞大，并且由于地理条件等因素不宜大量安装与部署，同时部分地区监测站由于地理环境等影响，设备频繁损坏，导致无法持久稳定的实现空气细粒度的质量监测。另一方面，监测站并未能够做好数据的相关预测，一般只提供实时数据，使得当地只能进行事后整治，而其治理效果往往并不理想。

$PM_{2.5}$作为城市空气污染的首要污染物，对空气质量和能见度等影响重大。尤其在于，$PM_{2.5}$直径小，并含有大量的有害物质，可轻易被吸入体内附着在呼吸道及肺叶上，对人体中的一系列系统产生危害，其中的有害物质、重金属等可溶解在血液中，对人体健康产生更大的伤害。如果长期处于高水平的$PM_{2.5}$浓度下，公共卫生健康则会受到严重影响、社会经济也会造成巨额损失。早在 2015

年，全球疾病负担研究中心（GBD）就曾发布相关研究报告，$PM_{2.5}$造成的空气质量污染，致使全球超过 400 万人过早死亡，中国则超过 100 万人。$PM_{2.5}$指标过高不仅对大众的健康造成了损害，同时也增加了国家在环境治理方面的支出。在京津冀及周边地区，仅治理好 $PM_{2.5}$这一项指标，就能够每年实现 600 亿元以上的健康效益，因此，$PM_{2.5}$污染及相关研究已经受到广泛关注。

1.2 研究意义

做好空气质量情况的预测，提前获得较为精准的预测结果，能够对大气污染严重、以及过去时常出现极端天气的区域提供数据支撑，也有利于该区域做好空气质量的预防和治理。此外，如果有较为精准的预测结果，政府等相关部门便可利用预测值来辅助决策，决定当前城市中学校是否需要停课，通知市民穿戴好防护口罩等工具或减少外出，以保护身体健康。因此，完善好空气质量预测系统，提供精准的预测结果，能够辅助政府在大气污染治理方面提出更有针对性的措施，也能够为人们制定合理的出行计划提供参考，并在自然环境的保护和社会的发展方面具有极大的益处与前景。

理论意义上，对已有的文献资料以及研究成果进行梳理，分析已有研究在空气质量预测问题中存在的不足，将不同学科内容进行交叉并运用于空气质量问题研究，提出一种新颖的组合模型，并将模型中的参数进行优化，以期在空气质量预测及相关问题研究中有所创新，提高空气质量预测的精准度。同时，从空间角度利用距离以及兴趣点数据（Point of Information，POI）构造其他站点对目标站点的空间影响，从而将空气质量预测问题在时间与空间上进行完整阐述，为构建时空类模型提供新思路。

现实意义上，利用各类可获得的城市大数据，诸如除历史空气质量数据之外的气象数据、POI地理信息数据，以此充分考虑各种影响因素，设计完备的时空预测模型及系统，为城市居民提供精准的空气质量预报，为城市决策者制定合理的环保决策提供数据支撑。此外，城市空气质量监测站点数量有限，并且城市面积愈发辽阔，空气质量监测站点覆盖的区域愈发不足。监测站点无法对未来空气质量进行预测，只能够提供实时监测值等问题突出。因此，设计一套空气质量时空预测模型，将时间与空间因素同时纳入模型进行预测，以期对$PM_{2.5}$浓度数据进行更精准的预测。

1.3 国内外研究综述

国内外研究学者针对空气质量污染物浓度预测，提出了许多预测算法与模型，根据其研究发展的先后顺序，主要有统计分析模型、支持向量机模型、神经网络模型等。统计分析模型将空气污染物浓度序列进行分解，将多个信号分量运用线性模型预测。支持向量机模型通过合适的参数与核函数来构建预测模型，并通过训练集学习数据特征，并构建参数优化的支持向量机，进而完成预测。然而统计分析模型在运用的前提中，要求数据要有较高的平稳性，支持向量机计算复杂，则在运用于大量数据集时，模型的运算效率较低，预测完成时间较长。随着神经网络在非线性数据的拟合与预测方面的优势被人们，因此相关学者将神经网络及深度学习领域的方法应用在空气质量预测中，来改进模型的特征提取能力以提高预测精度。

通过对城市空气污染的文献进行梳理与研究，相关学者将$PM_{2.5}$浓度预测归结对时间序列的研究中，而众多学者都对时间序列的预测问题展开了深入研究。现有文献中所构建的模型主要包括经典统计模型、支持向量机模型、神经网络模型等。经典统计模型中大多集中在线性回归，Sun 等（2013）为将某些关键气象因素纳入模型，并考虑$PM_{2.5}$浓度在先验中表现出的非高斯分布，提出了服从对数正态分布的隐马尔科夫模型，有效地降低了针对$PM_{2.5}$浓度超标的预警次数。龚明等（2016）首先对残差进行了拟合，并在此基础上将灰色模型和马尔科夫模型进行融合，提高了预测精度。但该类方法难以应对空气污染物的复杂变化，尤其是难以确切反映数据中的非线性与高波动性。因此，受其因素限制下预测精度难以提升。且经典统计模型并没有很好的处理数据中的非平稳性、含噪声等特点，模型预测精度也因此而下降。

因此，时间序列方面的分析方法被更多的学者所采纳，运用于$PM_{2.5}$浓度预测当中。早期 Le（2012）等采用自回归移动平均模型提取数据序列的线性特征，并得到了气压、风速等气象因素与细微颗粒物浓度之间的显著相关性。但传统的时间序列预测模型虽能较好的提取数据当中的线性趋势，但并不能捕捉数据中的非线性特征，尤其在空气质量这类高频复杂的时间序列中，传统的时间序列预测模型预测效果十分有限。为提高模型对数据中的非线性特征的提取能力，相关研究学者探寻机器学习方法的运用。其中 Sun 等（2017）利用主成分分析法和最小

二乘支持向量机的混合模型对$PM_{2.5}$浓度进行短期预测，李龙等（2017）使用最小二乘支持向量机模型结合气象因素和污染物浓度特征以预测$PM_{2.5}$浓度，提高了支持向量机模型的泛化能力及预测精度。Zhou 等（2019）将多任务算法和多输入支持向量机结合，通过任务算法寻找多输入支持向量机最优模型参数，将台北市各站点$PM_{2.5}$浓度数据送入模型中，比较该模型与其他模型的预测能力。但空气质量序列，往往变化十分频繁，样本量巨大，而 SVM 则由于计算复杂，应对较大数据集的处理时，效率低下。于是有学者提出将神经网络与其他算法进行配合，并将其运用于较大规模数据的计算，效果显著，因此众多学者将其应用于空气污染预测之中。石峰等（2017）通过灰狼优化算法改进神经网络预测模型的参数，并考虑了与空气污染物密切相关的气象数据，以上海市$PM_{2.5}$数据为实证案例进行拟合，其结果显著优于反向传播神经网络等模型。周杉杉等（2018）提出自组织递归模糊神经网络方法学习$PM_{2.5}$浓度序列的非线性特征，并利用主成分分析科学的筛选出与$PM_{2.5}$指标具有较强相关性的特征变量，作为神经网络的输入数据，采用偏最小二乘算法调整网络结构，使得模型即简洁又保持了较高的预测精度。

为了能将不同模型的优势进行集成，部分学者将不同模型进行组合，以探究组合模型预测效果。Wang 等（2015）提出通过泰勒展开对模型误差项进行修正，提出混合人工神经网络的概念，提高模型预测精度。但机器学习运用的本质是对序列进行有监督的学习，因此对数据中时间窗特征的学习效果，是机器学习类方法效果好坏的根本，而普通机器学习方法结果表现都不能尽如人意。而部分学者在对深度学习进行运用中发现，将长短期记忆神经网络运用于空气质量预测中，其很好的克服了机器学习在空气质量预测中的不足，能够较好的提取空气质量数据在时间维上的非线性。Huang 等（2018）更深一步的将卷积神经网模型长短期记忆神经网络（LSTM）模型相结合，运用 卷积提取$PM_{2.5}$浓度序列、风速小时数据等的时序特征后，将其提取的特征序列作为 LSTM 预测网络的输入数据进行预测，其预测结果的精度相较于 SVR、随机森林等机器学习模型或 LSTM 网络模型都有所提升。白盛楠等（2019）通过对气象、大气污染物因素进行灰色关联度分析，得到与$PM_{2.5}$之间的关联强度，采用多变量 LSTM 神经网络，较好地预测$PM_{2.5}$的日值变化趋势。蒋洪迅等（2021）构建了双向长短期记忆网络的神

经网络预测模型 DLENN，以双向 LSTM 分别提取了$PM_{2.5}$序列变化的趋势性和周期性，再以线性回归与神经网络进行结合，学习数据中的随机特性。将其实验结果与其他集成模型的预测结果进行对比，并表明 DLENN 模型在预测精度方面的优势。

而依托传统机器学习方法的组合模型，虽然预测精度有所改善，但对非平稳数据集的学习效果仍然不佳，并且此类模型过拟合问题突出。由此，部分学者提出了“分解-集成”的研究框架，以$PM_{2.5}$浓度序列作为原始信号进行分解，在完成分信号的预测后对预测结果进行集成。Xiong 等（2019）通过信号分解提取序列的非线性特征并降低数据的非平稳性，以此降低非线性与非平稳性对模型预测结果的影响。黄恒君等（2020）运用变分模态分解（VMD），将多模态分解处理后的数据作为深度学习的输入，并运用多视角学习进一步提取数据序列的非线性与非平稳性特征，以此提高了模型的预测精度。蒋峰等（2021）在对$PM_{2.5}$浓度序列进行分解后，采用样本熵提取各序列之间的相关性并进行重构，最后将重构的分序列作为极限学习机的输入，完成预测并集成预测结果，由此显著提高了模型的精度和稳健性。为进一步提升预测效果，部分学者提出将分信号进行二次分解，以提取复杂分信号的数据特征。Yin 等（2017）将经验模态分解产生的信号运用小波分解，进一步将分信号分解为更具平稳性的子序列，其模型预测结果得到稳步提升。由此，本文在数据分解中采用二次分解的结构，进一步提取$PM_{2.5}$浓度序列的特征。

而空气质量预测中，所研究的角度不应仅从对时间序列出发，还应包含空间属性的度量，将空间因素纳入模型。由此，考虑$PM_{2.5}$这类空气污染物在空气中的传播，度量其他地区空气质量对目标区域的影响，从时间和空间两个角度进行分析，从而更精准的解决预测问题。在这样的背景下，大量学者对空气质量数据进行分析与预测中提出了时空预测这一概念，而所运用的数据也因此在维度上则成倍增加，使得传统数据挖掘处理方法力不从心，亟需探索新的数据挖掘算法以匹配数据规模的迅速增长。时空序列数据即为指具有空间相关关系的时间序列的集合，其由一般时间序列延伸而出，已有的预测算法诸如循环神经网络、长短期记忆网络以及门控循环单元网络等，在提取时间维度的依赖性上表现优良，但没能够考虑其中可能出现的空间关联。Liu 等（2017）提出端到端的深度学习架构，

将卷积和 LSTM 网络结合形成 Conv-LSTM 模型，该模型将卷积神经网络和长短期记忆神经网络进行叠加，用卷积的方式提取空间特征，Liu 将该模型运用在交通流时空信息的提取当中，效果相较于只考虑时间的算法有较大的改进。

上述空间特征都是以卷积操作为主，卷积提取过程中，在池化层往往会遗失大量空间信息。为进一步提取时空数据的空间特征，He 等（2016）提出将数据转换为图片序列，利用卷积神经网络、残差神经网络等将数据转换为三维图片来提取空间特征。Zhang 等（2016）在 DeepST 模型中首次将此类想法应用于交通流量预测，将时点的交通流量数据转为图片数据，然后进行降采样，将其结果利用卷积进行特征提取，并对得到的特征进行融合以此完成预测。Zhang 等（2017）发现随着卷积模块层数的增加，预测精度会有所降低，于是对 DeepST 模型做出改进，在卷积的基础上增加了对残差信息的处理，降低卷积处理的信息损失水平，以此构建了 ST-ResNet 模型。上述模型虽然都能够处理数据中的空间特征，但对时间维度特征的提取较为粗略，仍有很大的提升空间。Li 等（2019）提出了 ST-DCCNAL 模型。将处于不同位置上的时空数据转换为图片数据，并利用 LSTM 网络以提取时序特征。在空间特征的提取上，则利用全连接网络来代替循环神经网络，弥补了循环神经网络存在信息损失的缺点。此外，ST-DCCNAL 模型将注意力机制加入 LSTM 中，从而自动更新不同时刻对应的权重，考虑到时间间隔越近，影响越大的现实影响。

综上所述，现有空气质量预测模型有各自优点，但仍存在以下几点问题：

（1）统计分析模型运用当中，需要数据具有良好的平稳性，而成因复杂或者不理想的周期信号数据，不稳定性明显，尤其是以空气质量数据为代表。如果不加以考虑就使得模型预测能力大幅下降。仅通过分析历史数据变化趋势，模型预测能力不会有效得到提升。

（2）随着训练样本逐步增大甚至上升为海量数据时，由于支持向量的求解过程是借助二次规划等复杂计算，数据量过大就会耗费大量的机器内存和运行时间，因此模型的总体复杂度也会急剧上升，运行效率降低。

（3）数据特征提取技术进步的同时，必然产生更为复杂、维度更高的数据，使得后续对数据处理及运算越发困难，预测效率显著下降。

空间分析模型中，主要存在数据获取成本与分析结果精度权衡问题，主要体

现在如下几个方面：

（1）尽管现如今的时空序列预测算法都能够做到提取数据中的时空关联特征，但其模型大多仅适用欧式空间的时空数据。但在实际中，大部分时空数据往往以非欧数据形式存在，例如 POI 等，强行转换为图片数据以此提取时空特征，并没有较强的逻辑关系，其预测精度也往往有限。

（2）上述时空预测模型中，都是将空间特征矩阵与原数据进行拼接，并送入神经网络进行学习，这严重加大了神经网络的结构复杂程度，模型学习效率不高。

1.4 研究内容及创新点

本文依据当前研究的现状和发展方向，对各类空气质量预测模型的优缺点进行了总结：从着力挖掘历史数据变化趋势，仅从时间序列相关性上考虑气象等相关因素，并进行预测的统计分析模型，发展到将多维矢量数据通过非线性算子映射到高维特征空间，以此提取数据非线性性的支持向量机模型。而随着深度学习的发展与运用，相关学者开始广泛使用神经网络对空气质量数据进行建模预测，但多数研究成果都仅采用短期数据进行建模预测，从而忽略了空气质量及其影响数据的长期趋势。此外，传统神经网络不仅存在过拟合的问题，也会因为训练数据量的增长出现梯度消失以及误差积累等问题。故此，现有的时空预测模型在准确性以及模型结构优化方面还有很大的提升空间。

本文在现有文献的基础上进一步深入挖掘影响 $PM_{2.5}$ 浓度序列变化的影响因素并提高空气质量模型的预测精度，从时间与空间角度出发，研究 $PM_{2.5}$ 浓度序列在时间上的时序依赖性，以及周边区域对目标区域的影响。旨在通过充分挖掘影响 $PM_{2.5}$ 浓度序列的空间因素和时间因素，并结合空气质量数据在短期内的波动以及长期内的趋势特点，进行模型的构建完成预测，以期提高空气质量预测精度。

创新点在于模型构建从时间和空间角度共同出发，时间维度中，基于“分解-聚类-集成”研究框架，提出 CEEMDAN 与 VMD 分解相组合的二层分解方法，深入提取 $PM_{2.5}$ 浓度序列的时序特征，并提出基于时间序列聚类方法下的 K-shape 算法将 $PM_{2.5}$ 浓度序列的分解结果进行聚类，降低模型计算复杂度。在空间维度中，基于图模型的原理，计算空间特征的拉普拉斯矩阵，并将所得结果与 $PM_{2.5}$

站点数据进行点乘的方式，将空间特征嵌入 $PM_{2.5}$ 数据中，完成空间特征的提取。此外，为将时间预测结果与空间预测结果进行动态结合，利用 XGboost 模型提取时间与空间特征的重要性权重，并与预测结果进行线性加权，完成时空预测模型构建。

1.5 研究结构安排

本研究总共分为五章进行阐述，论文的主要内容结构如图 1.1 所示：

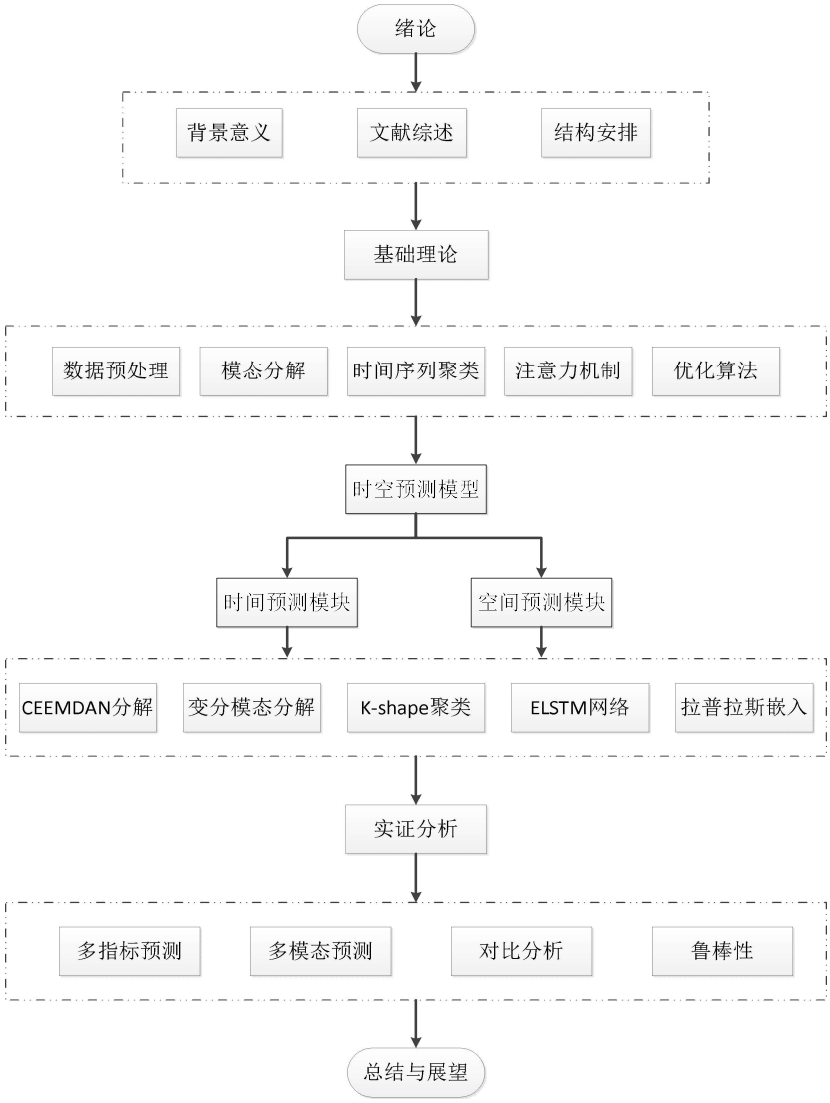

图 1.1 论文框架示意图

第 1 章：研究内容背景及意义，本章在阐述研究背景及意义的基础上，分析过往国内外空气质量预测方面的研究内容，阐释现有空气质量预测模型中所存在的一些问题与不足，并提出本文构想。

第 2 章：理论基础及方法介绍，本章主要论证数据缺失值、离群点的处理方法，并对模态分解算法、时间序列聚类、深度神经网络（ELSTM）、粒子群优化算法等模型建立过程中涉及的理论基础进行详细阐述。

第 3 章：时空预测模型构建，本章主要介绍空气质量时空预测模型(LX-M-CEEMDAN-VMD-LSTM)，详细介绍了数据在时间预测模块中的具体步骤及流程，并提出新颖的简易空间特征提取方法，以构造空间预测模块。最后，探究将时间预测模块与空间预测模块动态集成的方法，以完成时空模型的搭建。

第 4 章：本章主要以兰州$PM_{2.5}$为研究对象，将其他空气污染数据、气象数据、以及城市地理信息数据作为辅助，运用第三章提出的时空预测模型进行$PM_{2.5}$浓度预测，详细阐述了从数据采集、数据处理、模型预测等各个环节所做的研究，并运用误差指标对模型进行评价以及鲁棒性分析。

第 5 章：总结与展望。在完成本文主体研究内容后，对论文的工作进行了思考和总结，进一步阐述了空气质量预测当中时空研究的意义，并对其进一步可以充实的工作重点和方向进行了深入构思。

2 理论基础与方法

本章首先介绍了数据预处理和特征选择相关内容，并阐述了建模过程中的模态分解、时间序列聚类以及集成预测的有关基础理论和方法，为下章提出的时空预测模型做准备。

2.1 数据预处理

（1）缺失值填充

由于数据均为小时数据，考虑空气质量数据在相近的数小时内变化不大，对则对连续缺失值不超过三个数据位置，根据前后数据求均值进行补全。其公式如（2.1）所示：

$$X_t = \frac{1}{2}(X_{t-1} + X_{t+1}) \tag{2.1}$$

其中，X_t是t时刻的缺失值，X_{t-1}与X_{t+1}是前后时刻的值，即根据前后的值估计出缺失值，以此捕捉时间上的连续性。

若连续缺失值大于 3 个，则使用 KNN 算法，以此利用其它站点未缺失的数据对缺失值进行填充。通过回归建模的方式，寻找 K 个近邻值中相似度最高的数据进行填充。

（2）平稳化处理

进行平稳化处理是为了应对数据中具有的波动性、不平稳性等特征特征，尤其在空气质量数据中，各项指标值都具有高波动性以及不平稳性，并且对预测结果的精准度具有十分重要的影响。常用的平稳化处理有差分处理、平方根变换、对数变换、归一化变换等方法。

本文在模型训练之前采用了归一化的方式处理空气质量数据和气象数据。将每个数据点都映射到[0,1]之间，归一化如公式（2.2）所示：

$$x_i^* = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{2.2}$$

其中，x_i^*表示变化之后的数据值，$x_{\max}$为x_i的最大值，$x_{\min}$为最小值。

（3）特征选择

为后续时空预测模型构建提供充分的数据基础，本节需探究 $PM_{2.5}$ 浓度序列与其他空气污染物浓度序列以及气象因子之间的具体关系。而相关性分析则是探究两组或两组以上变量之间的相关程度的基本方法，通过度量 $PM_{2.5}$ 浓度序列与

其他空气污染物浓度序列以及气象因子的相似度，将相关性较弱的特征序列排除在模型输入之外，防止其序列对模型的预测精度造成干扰。在统计研究中，常用的相关性分析方法包括相关系数、回归分析以及方差-协方差矩阵。而在特征选择中，通过计算相关系数就可以通过度量变量之间的相关性，以此保留相关性程度较高的特征，去除相关性较低的特征，精简模型的同时降低不相关特征对模型的干扰。

本节通过计算 $PM_{2.5}$ 浓度序列与其他污染物、气象因子之间的相关系数矩阵进行相关分析。相关系数的计算方法主要有三种：Pearson 相关系数、Kendall 相关系数和 Spearman 相关系数。其中，Pearson 相关系数要求繁杂，其适用条件为变量间为线性关系、变量总体为正态单峰分布、且观察值是成对独立。Kendall 相关系数和 Spearman 相关系数，则并无对样本量大小以及总体分布形态等要求。故本文采用 Spearman 相关系数度量 $PM_{2.5}$ 浓度序列与其他污染物、气象因子之间的相关程度，计算公式如（2.3）所示：

$$\rho = \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{2.3}$$

这里，d_i为两组待计算数据 X 和 Y 分别排序后对应的x_i和y_i的差值，n为样本数。（2.4）式为 Pearson 相关系数计算公式：

$$\rho = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2(y_i-\bar{y})^2}} \tag{2.4}$$

2.2 模态分解

2.2.1 集成经验模态分解（EEMD）

集成经验模态分解（EEMD），是将待分解的信号中加入白噪声，再利用经验模态分解（EMD）的方法进行分解，补充了经验模态分解中丢失的信息。模态分解是一类将原始信号分解为多个分信号的时频分析方法，能够捕捉原始信号中非线性和非平稳特征。但分解过程中，出现振动幅度较大的情况时，会使 EMD 的分解的结果中出现模态混叠现象。即相近的特征时间尺度在不同的 IMF 都有出现，或一个 IMF 分量中出现差别较大的特征时间尺度，使得相邻两个 IMF 波形混叠难以区分，致使分解结果不能满足拆分非线性以及波动性的目的。而 EEMD 则能够较为有效的改善 EMD 产生的模态混叠现象，其步骤可以简述为：

（1）$s(t)$为原始信号序列，$v^i(t)$代表第 i 次实验中添加的白噪声序列，其分布

为标准正态分布。第 *i* 次的信号序列表示为$S^i(t)=s(t)+v^i(t)$，其中$i=1$，…，i表示实验的次数。

（2）将分信号序列 *s*^{i}(*t*)运用 EMD 分解，得到IMF_k^i，其中$k=1$，…，*K*，表示分解的模态个数。

（3）则有 *s*(*t*)的 *k* 个模态分量为 *IMF*$_k$，并对IMF_k^i进行平均得到$\overline{IMF_k}$，即：$\overline{IMF_k}=1/K\ \sum_{i=1}^{K} IMF_k^i$。

2.2.2 自适应完备集成经验模态分解（CEEMDAN）

经验模态分解中添加白噪声序列，会产生破坏原始信号特征的问题。针对以上问题，Torres 等提出了自适应加噪声的完备经验模态分解（CEEMDAN），引入自适应高斯白噪声，即将有限次的自适应白噪声添加至每个分解阶段中，以此达到在较少的计算下，其重构误差接近 0。因此，CEEMDAN 可以有减少模态分解过程中混叠问题的发生，也有效改善了经验模态分解方法中时常存在的分解不完整的缺点。CEEMDAN 算法可描述如下：

（1）利用 EEMD 算法分解得到第一个模态分量，如公式（2.5）所示：

$$IMF_1=\frac{1}{I}\sum_{i=1}^{I} IMF_i \tag{2.5}$$

（2）在第一阶段（$k=1$）计算第一个余量，如公式（2.6）所示：

$$R_1[n]=X[n]-IMF_1[n] \tag{2.6}$$

（3）分解$R_1[n]+\varepsilon_1 E_1(\omega_i[n])$，（$i=1$，…，$I$）到第一个模态分量，则第二个模态分量表示为，如公式（2.7）所示：

$$IMF_2[n]=\frac{1}{I}\sum_{i=1}^{I} E_1(R_1[n]+\varepsilon_1 E_1(\omega[n])) \tag{2.7}$$

（4）对于$k=2$，…，*K*，计算第 *k* 个余量，如公式（2.8）所示：

$$R_k[n]=R_{(k-1)}[n]-IMF_k[n] \tag{2.8}$$

（5）分解$R_k[n]+\varepsilon_k E_k(\omega_i[n])$，（$i=1$，…，$I$）分解到第 *k* 个模态分量上，第 *k*+1 个模态分量可表示为，如公式（2.9）所示：

$$IMF_{(k+1)}[n]=\frac{1}{I}\sum_{i=1}^{I} E_k(R_k[n]+\varepsilon_k E_k(\omega_i[n])) \tag{2.9}$$

（6）重复第四步至第五步直到当残差分量不适应被分解时，停止分解。最

终的余量满足：，$R[n]=X[n]-\sum_{k=1}^{K} IMF_k$ 其中 K 表示分解得到的固有模态函数的数量，参数 $X[n]$表示为$X[n]=\sum_{k=1}^{K} IMF_k+R[n]$。

2.2.3 变分模态分解（VMD）

2014 年由 Dragomiretskiy 和 Zosso 提出变分模态分解（VMD），即一种新数据处理技术。VMD 分解是将信号分解为 K 个本征模态函数（IMF），通过寻找一系列模态及各模态的中心频率，重构原始数据。VMD 分解目的是将 K 个本征模态函数的带宽之和达到最小，进而运用 L_2 范数的平方最小即得到上述要求。故此，分信号的瞬时频谱具有一定的现实物理意义。具体步骤如下：

（1）构造变分问题，假设原始信号 f 被分解为 k 个分量，为保证分解序列为具有中心频率的有限带宽的模态分量，同时确保各模态的估计带宽之和最小，其约束条件为所有模态之和与原始信号 f 相等，具体变分约束表达式如（2.10）所示：

$$\min_{\{u_k\}\{\omega_k\}}\left\{\sum_k\left\|\partial\left[\left(\delta(t)+\frac{j}{\pi t}\right)*u_k\right](t)e^{j\omega_k t}\right\|_2^2\right\} \quad s.t.\sum_{k=1}^{K}u_k=f \tag{2.10}$$

式中：$\{u_k\}=\{u_1,u_2,\cdots,u_k\}$是模态，$\{\omega_k\}=\{\omega_1,\omega_2,\cdots,\omega_k\}$是模态的中心频率。另外，$\delta(t)$为狄利克雷函数，*为卷积运算。

（2）求解变分问题，引入拉格朗日乘子λ，转变为无约束变分问题，得到增广拉格朗日表达式如（2.11）所示：

$$L(\{u_k\},\{\omega_k\},\lambda)=\alpha\sum_k\left\|\partial_t\left[\left(\delta(t)+\frac{j}{\pi t}\right)*u_k(t)\right]e^{-j\omega_k t}\right\|_2^2+\left\|f(t)-\sum_k u_k(t)\right\|_2^2+\left\langle\lambda(t),f(t)-\sum_k u_k(t)\right\rangle \tag{2.11}$$

式中：α为平衡参数，其作用是降低高斯噪声的干扰。在本文中，式（2.11）的优化问题采用迭代方向乘子法（ADMM），既通过式（2.12）-式（2.14），迭代更新 u、ω、λ。

$$\hat{u}_k^{n+1}(\omega)=\frac{\hat{f}(\omega)-\sum_{i\neq k}\hat{y}_i(\omega)+\frac{\hat{\lambda}(\omega)}{2}}{1+2\alpha(\omega-\omega_k)^2} \tag{2.12}$$

$$\omega_k^{n+1}=\frac{\int_0^\infty\omega\left|u_k^{n+1}(\omega)\right|^2d\omega}{\int_0^\infty\left|u_k^{n+1}(\omega)\right|^2d\omega} \tag{2.13}$$

$$\hat{\lambda}^{n+1}(\omega)=\hat{\lambda}^{n}(\omega)+\gamma\left(f(\omega)-\sum_{k}\hat{u}_{k}^{n+1}(\omega)\right) \tag{2.14}$$

其中：γ为噪声容忍度，$u_k^{n+1}(\omega)$、$u_i(\omega)$、$f(\omega)$和$\hat{\lambda}(\omega)$分别对应$u_k^{n+1}(t)$、$u_i(t)$、$f(t)$和$\hat{\lambda}(t)$的傅里叶变换。

2.3 基于形状的时间序列聚类算法

由于时序数据的特殊性，其聚类方法应与截面数据聚类有所区别。Gravano 基于一种可度量的迭代优化过程提出 K-shape 聚类算法。首先，基于动态规划原理，对时间序列进行扭曲，进行必要的错位。然后，计算出最合适的距离，依据时序数据的形状相似性，将形状相似的序列聚为一类。

2.3.1 时间序列形状相似度

以往的时间序列聚类都是通过振幅和相位畸变的距离进行测度，Gravano 则通过互相关测度确定时间序列的相似性。计算互相关测度时保持其中一个序列 y 不变，将 x 在 y 上滑动，计算 x 的每一个位移 s 的内积。将所有内积结果表示为，$CC_w(x,y)=(c_1,c_2,\cdots,c_w)$，并且得到的互相关序列长度为$2m-1$，则有下定义：

$$CC_w(x,y)=R_{w-m}(x,y),w\in 1,2,\cdots,2m-1 \tag{2.15}$$

$R_{w-m}(\vec{x},\vec{y})$则由下式（2.16）可得：

$$R_x(\vec{x},\vec{y})=\begin{cases}\sum_{i=1}^{m-k}x_{i+k}\cdot y_i, & k\geq 0\\ R_{-k}(\vec{y},x), & k<0\end{cases} \tag{2.16}$$

则计算出使得$CC_w(x,y)$最大的ω，进而得到x相对于y的最佳移动：$s=w-m$。进而得到距离测度如公式（2.17）所示：

$$SBD(x,y)=1-\max_w\left(\frac{CC_w(x,y)}{\sqrt{R_0(x,x)*R_0(y,y)}}\right) \tag{2.17}$$

取值范围为[0,2]，结果越小表示两个序列越相似。

2.3.2 时间序列形状提取

时间序列分析中的许多任务依赖于通过一个序列有效地总结一组时间序列的方法。这个序列此处称为平均序列，在聚类中则被称为质心。其目标是找到与类内所有其他时间序列之间距离平方和的最小值。因此就变为一个优化问题：

$$\mu_k=\arg\max_{\mu_k}\sum_{x_i\in P_k}NCC_c(x_i,\mu_k)^2$$

$$= \arg\max_{\mu k} \sum_{P_k} \left(\frac{\max_{\omega} CC_{\omega}(x_i, \mu_k)}{\sqrt{R_0(x_i, x_i), R_0(\mu_k, \mu_k)}} \right)^2 \tag{2.18}$$

式（2.18）中需要对类内所有的时间序列计算一个最佳的偏移。因为在这里提到的方法是用在迭代聚类当中，所以把前一次计算得到的聚类中心作为参考并把所有的序列与这个参考的序列对齐，并根据式（2.15）与（2.16）化简得到公式（2.19）：

$$\mu_k = \arg\max_{\mu_k} \sum_{x_i \in P_k} \left(\sum_{l \in [1,m]} x_{il} \cdot \mu_{kl} \right) \tag{2.19}$$

为了简单起见，我们用向量表示此方程，并假设序列已经进行了归一化处理。得到公式（2.20）：

$$\begin{aligned} \mu_k &= \arg\max_{\mu_k} \sum_{x_i \in P_k} \left(x_i^T \cdot \mu_k \right)^2 \\ &= \arg\max_{\mu_k} \mu_k^T \cdot \sum_{x_i \in P_k} \left(x_i \cdot x_i^T \right) \cdot \mu_k \end{aligned} \tag{2.20}$$

引入$M = Q^T \cdot S \cdot Q$，令$\mu_k = \mu_k Q$，其中$Q = I - \frac{1}{m} O$，I是单位矩阵，O是全幺矩阵，用S代替$\sum_{x_i \in P_k} \left(x_i, x_i^T \right)$，得到公式（2.21）：

$$\mu_k = \arg\max_{\mu_k} \frac{\mu_k^T \cdot Q^T \cdot S \cdot Q \cdot \mu_k}{\mu_k^T \cdot \mu_k} = \arg\max_{\mu_k} \frac{\mu_k^T \cdot M \cdot \mu_k}{\mu_k^T \cdot \mu_k} \tag{2.21}$$

最大值μ_k即为求瑞利商（Rayleigh quotient）最大化问题，同时最大值为矩阵M对应最大特征值的特征向量。具体算法过程如下：

算法 1：提取平均序列

输入：数据矩阵$X(n * m)$，具有标准化后的时间序列；

C是一个m维列向量，其参考序列与X的时间序列对齐。

输出：C'是一个具有质心的m维列向量。

$X' \leftarrow \varnothing$

For $i \leftarrow 1$ *to* n *do*

$[dist, x'] \leftarrow SBD\left(C, X\left(i\right)\right)$

$X' \leftarrow [X'; x']$

$S \leftarrow X'^T \cdot X'$

$Q \leftarrow I - \frac{1}{m} \cdot O$

$M \leftarrow Q^T \cdot S \cdot Q$

$C' \leftarrow Eig(M, 1)$

2.3.3 基于形状的时间序列聚类

K-shape 算法的聚类时采用迭代的方式进行，迭代分为两步：（1）在分配步骤中，将每个时间序列与平均序列比较，并将其分配给最接近平均序列的一类以此更新聚类中的成员关系；（2）在细化步骤中，重新计算质心，比较每个序列与新质心的距离，并进行重新分配。算法重复这两个步骤，循环迭代到组内序列没有变化，或达到允许的最大迭代次数为止。具体算法流程如下：

算法 2：K-shape 聚类算法

输入：数据矩阵X（$n*m$），最大迭代次数为 100

输出：IDX是一个n维列向量，包含n个时间序列的k个聚类信息。

C是一个$k*m$的矩阵，包含k个聚类中心。

While $IDX \neq IDX'$ *and* $iter < 100$ *do*

$IDX \leftarrow IDX'$

for $j \leftarrow 1$ *to* k *do*

$X' \leftarrow \varnothing$

for $i \leftarrow 1$ *to* n *do*

If $IDX(i) = j$ *then*

$X' \leftarrow [X'; X(i)]$

根据式（24）计算聚类中心

for $i \leftarrow 1$ *to* n *do*

$\min dist \leftarrow \infty$

for $j \leftarrow 1$ *to* n *do*

$[dist, x'] \leftarrow SBD(C(j), X(i))$

if $dist < \min dist$ *then*

$\min dist \leftarrow dist$

$IDX(i) \leftarrow j$

2.4 ELSTM 神经网络

长短时记忆神经网络最早是由 Hochreiter 和 Schmidhuber 提出，是循环神经网络（RNN）的一个改进与发展。LSTM 由于其独特的单元结构，在处理长期相关关系方面具有较明显的优势，其结构如图 2.1 所示：

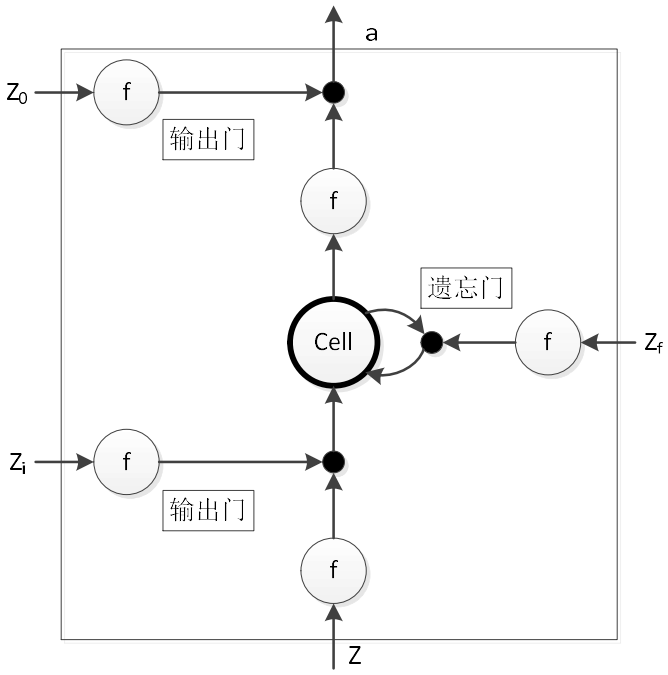

图 2.1 LSTM 神经网络的神经元结构图

LSTM 包含一个或者多个储存器和三个自适应乘法门，分别为输入门、输出门和遗忘门。其中，输入门是为了控制是否允许写入，而遗忘门是为了控制记忆单元的值是否需要更新，输出门是控制是否允许输出，通过这三个门实现信息的保存与控制。在时刻t，设 x_t 代表 $PM_{2.5}$ 的时间序列，y_t 代表 LSTM 的预测结果，c_t 和 h_t 分别为神经元状态值和隐藏层状态值，则 LSTM 各单元更新情况如（2.22）-（2.27）所示：

$$i_t = \sigma\left(W_{ih}h_{t-1} + W_{ix}x_t + W_{ic}c_{t-1} + b_i\right) \tag{2.22}$$

$$f_t = \sigma\left(W_{fh}h_{t-1} + W_{fx}x_t + W_{fc}c_{t-1} + b_f\right) \tag{2.24}$$

$$c_t = f_t * c_{t-1} + i_t * g\left(W_{ch}h_{t-1} + W_{cx}x_t + b_c\right) \tag{2.24}$$

$$o_t = \sigma\left(W_{oh}h_{t-1} + W_{ox}x_t + W_{oc}c_t + b_o\right) \tag{2.25}$$

$$h_t = o_t * \tanh\left(c_t\right) \tag{2.26}$$

$$y_t = \phi\left(W_{yh}h_t + b_y\right) \tag{2.27}$$

其中，W_{ih}、W_{fh}、W_{ch}、W_{oh} 分别是隐藏层状态值 h_t 的权重矩阵；W_{ix}、W_{fx}、W_{cx}、W_{ox} 分别是时间序列 x_t 的权重矩阵；W_{ic}、W_{fc}、W_{oc} 分别是神经元状态值 c_t 与三个门函数的对角矩阵；b_i、b_f、b_c、b_o 分别是偏置向量；W_{yh} 和 b_y 是 LSTM 网络的输出权重与偏置向量；σ(*)是 *sigmoid* 激活函数；g(*)和 h(*)都是 *tanh* 激活函数；Φ为 *softmax* 激活函数。上述模型中所需要学习训练的参数有：各个节点间的有偏连接权重、神经元内部的输入连接权重和神经元递归连接权重。并对其权重学习设置了不同系数的正则化项。进而使用弹网惩罚项将其添加到目标函数中，防止模型学习过程中的过拟合，即公式（2.28）所示：

$$\min_{\omega}\left\{\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(y_{ti}-\overline{y}_{ti}\right)^2\lambda_1\|\omega\|_1+\lambda_2\|\omega\|_2^2\right\} \tag{2.28}$$

当$\lambda_1 \neq 0$，$\lambda_2 \neq 0$上式即为弹网惩罚，构成 ELSTM 模型，以提高模型的泛化能力。

2.5 粒子群算法

粒子群算法是一种进化计算技术，能够对模型中的参数组合进行自动寻优，其原理是通过个体的协作与信息共享来找寻最优方案，其构思来源于鸟群捕食行为的模仿。

算法设计了一种无质量的粒子来模拟鸟群中的个体，其具有移动的方向及距离。速度代表移动的快慢，位置代表移动的方向。二者即为每个粒子的变量。所有粒子根据自身历史记录中的最佳位置，以及群体中共享的当前全局最优解来调整自身的速度和位置。其位置更新过程如图 2.2 所示：

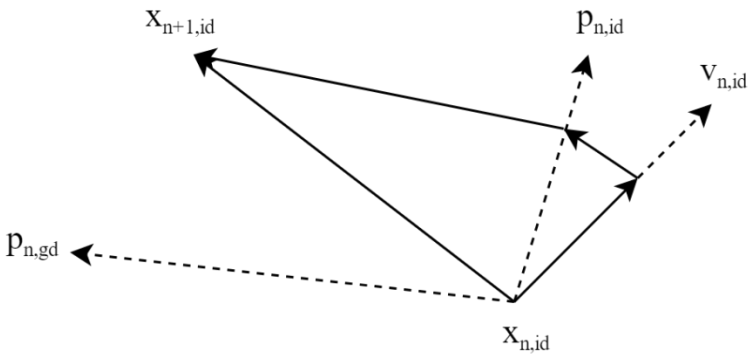

图 2.2 粒子位置更新示意图

假设在d维空间中有m个粒子，粒子的速度与位置更新公式为（2.29）与（2.29）：

$$v_{id}^{n+1} = \omega v_{id}^{n} + c_1 r_1 \left(p_{id}^{n} - x_{id}^{n}\right) + c_2 r_2 \left(p_{gd}^{n} - x_{id}^{n}\right) \tag{2.29}$$

$$x_{id}^{n+1} = x_{id}^{n} + v_{id}^{n+1} \tag{2.30}$$

式中，v_{id}^{n}是第n次迭代中第i个粒子的第d维速度分量，x_{id}^{n}是第n次迭代中第i个粒子第d维位置分量，p_{id}^{n}是第n次迭代中第i个粒子的个体极值的第d维分量，p_{gd}^{n}是第n次迭代中第i个粒子第的全局极值的第d维分量，c_1和c_2是学习率，r_1和r_2为 0 到 1 之间的随机数，ω是权重。其基本过程如图 2.3 所示：

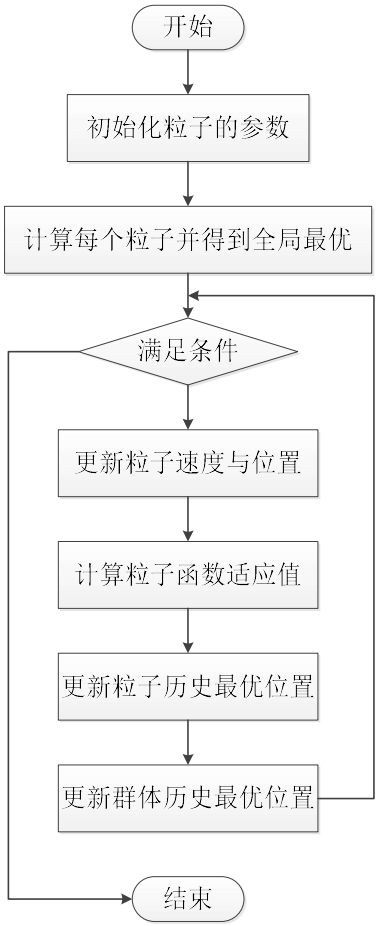

图 2.3 粒子群算法流程图

粒子群算法参数少且结构简单，所以寻优效率高，尤其在多目标参数优化中，能咦较高的效率得到较优的参数组合。因此，本文运用粒子群算法寻找 LSTM 神经网络模型的最优超参数组合。

2.6 本章小结

本章首先介绍数据预处理中的异常值的筛查以及缺失值填充，并探究模型中所运用的模态分解（EEMD、CEEMDAN、VMD）与基于时间序列的聚类算法（K-shape）以及神经网络模型（LSTM 模型）与粒子群参数优化的原理，然后对本文提出的 $PM_{2.5}$ 预测模型相关的技术进行介绍。

3 时空预测模型构建

针对现有预测模型对空气质量数据中的非平稳性与非线性特征提取不足，以及传统神经网络携带时间序列的长期趋势特征等问题，本章提出一种基于时空数据的 $PM_{2.5}$ 预测模型（LX-M-CEEMDAN-VMD-ELSTM）。

3.1 模型架构

首先，做好数据处理的基础性工作。对空气污染数据以及气象数据中的缺失值进行填充，对数据异常值进行预处理。统计空气质量监测站点周围一公里 POI 的类型及数量反映空气质量监测站点的空间相似度。通过本文上一章节介绍的模态分解，时间序列聚类，ELSTM 网络以及在构建网络时起辅助作用的粒子群优化算法，构建时间预测模型，即基于“分解-聚类-集成”的深度神经网络模型，获取 $PM_{2.5}$ 时序数据中的短期波动性与长期趋势性。并运用多变量 ELSTM 模型获取 $PM_{2.5}$ 与其他污染物数据、气象数据的序列相关特征，以此反映其他时序数据对 $PM_{2.5}$ 的影响。并将预测结果进行集成构成时间预测模型（LX-M-CEMMDAN-VMD-ELSTM）。在空间上，充分考虑影响空气质量预测的空间因素，一方面考虑由经纬度数据获取的站点之间的欧式距离，另一方面考虑 POI 数据，由于 POI 数据是空间上非地理意义的点，如商店，酒吧，加油站，医院，车站等，由此很好的反映了城市空间布局。运用上述空间数据计算拉普拉斯矩阵提取空气污染检测站点的空间特征并与各站点 $PM_{2.5}$ 数据进行点乘，以此将空间特征嵌入数据，构造其他站点对目标站点的空间影响，并作为多变量 ELSTM 模型的输入，完成空间预测模型的构建。图 3.1 是本文提出的模型与研究内容逻辑关系图。

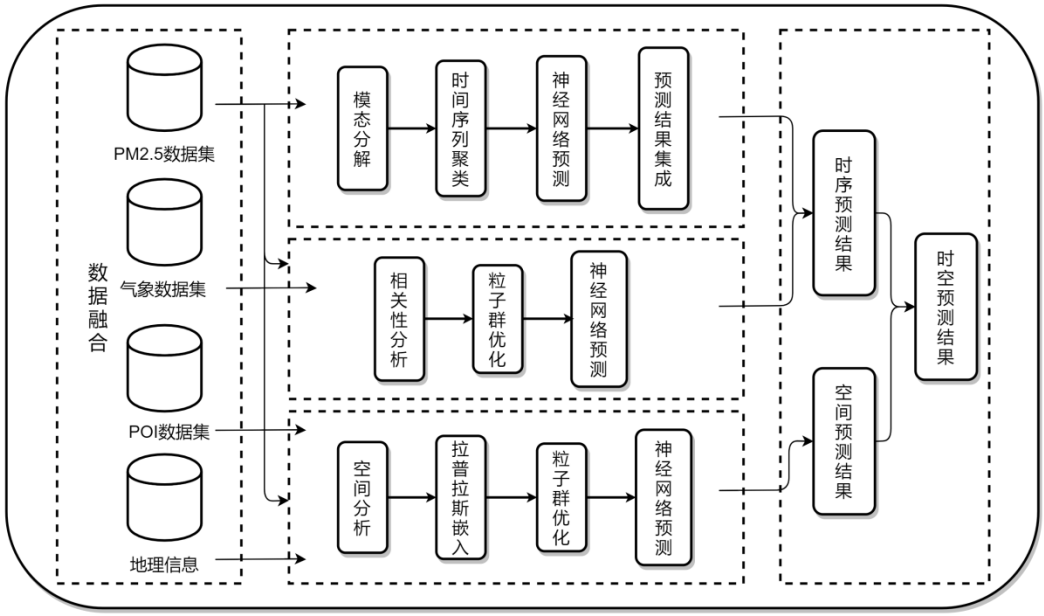

图 3.1 模型构建图

3.2 时间预测模块

在时间预测模块中，首先将 $PM_{2.5}$ 原始序列利用模态分解的方法分解为多个子序列，利用子序列提取 $PM_{2.5}$ 序列自身的高波动性与长期趋势。再将子序列作为神经网络的输入，经过神经网络学习各子序列特征后，将子序列的预测结果进行线性集成。此外，利用气象数据，例如降水、风速、气压等，其他污染物数据，例如 PM_{10}，SO_2，NO_2 等。先构建各个特征与 $PM_{2.5}$ 的相似水平，提取相关程度较高的特征，运用多变量 LSTM 模型构建变量之间的关系形成预测结果。最后将两部分预测结果结合构建时间预测模块。

前文中，已经有部分学者将模态分解后的结果进行二次分解，更深一步的提取到了时间序列中的复杂特征。但运用二次分解后会显著提高模型的计算时间，导致模型运行效率降低。本文则对分解后的结果进行聚类处理，重构 ELSTM 神经网络的输入，以完成合理减少子序列数量的目标，从而提高模型运行效率。在模态分解集成中主要进行如下研究：

（1）对数据进行预先处理，运用 KNN 算法填补缺失值并进行相关性分析，再将序列转化为有监督序列以配合 LSTM 模型的输入；（2）运用 CEEMDAN 方法将原序列进行分解，由此避免模态混叠现象并改进 EEMD 对信号分解的不

完整性。进一步将高频信号运用 VMD 分解，提取复杂分信号的潜在特征；（3）利用基于形状相似度的时间序列聚类算法（K-shape）对分信号进行聚类。用来区分所有成分之间的差异，并且依据它们的特性对这些数据划分成 K 类。具体而言，各分信号和残差成分的预测值可以被它划分成不同的类别，每个类别中的序列具有彼此相似的特征；（4）将 LSTM 神经网络加入正则化项，在规避递归神经网络的梯度爆炸问题的同时，提高模型的稳定性及泛化能力；（5）采用“分解-聚类-集成”框架作为组合模型的运行机制，更好的适应于时间序列中非平稳、高波动、含噪声的特点；具体而言，为了改善 $PM_{2.5}$ 浓度序列的预测精度，本文在“分解-聚类-集成”的研究范式下，从信号分解，信号预测以及对结果进行集成方面对已有模型进行优化和改进，具体算法流程如下：

算法 3：分解-聚类-集成预测算法

输入：数据$X_{[n,1]}$，n为样本个数，h为预测个数

过程：令$L()$为 LSTM 预测网络，$f()$为 CEEMDAN 分解，$g()$为 VMD 分解

1. 分解：

（1）：求解$X^{*}_{[n,m]}$：$X^{*}_{[n,m]}{}' \leftarrow f(X)$

$$令X^{*}_{[n,m]} := \left[X^{*}_{[n,1]}, X^{*}_{[n,2]}, X^{*}_{[n,m-2]}\right]$$

（2）：求解$X^{**}_{[n,2m]}$：$X^{**}_{[n,2m]} \leftarrow g\left(X^{*}_{[n,1]}\right), g\left(X^{*}_{[n,2]}\right)$

2. 聚类：利用 K-shape 聚类算法将$X^{**}_{[n,2m]}$聚为 K 类，记为$C_1, C_2, \cdots, C_K$

3. 预测：令预测结果$\left[R_1, R_2 \cdots, R_K, R\left(X^{*}_{[n,m-2]}\right)\right]$，求解：

$$\left[R_1, R_2 \cdots, R_K, R\left(X^{*}_{[n,m-2]}\right)\right] \leftarrow G\left(\left[C_1, C_2 \cdots, C_K, X^{*}_{[n,m-2]}\right]\right)$$

4. 集成：$Y_{[h,1]} \leftarrow R_1 + R_2 + \cdots + R\left(X^{*}_{[n,m-2]}\right)$

输出：预测值$Y_{[h,1]}$

3.3 空间预测模块

如今在城市中均部署有空气质量监测站点，且数量在逐年上升，空间布局也日趋合理。监测目标包括固体颗粒污染物（$PM_{2.5}$ 和 PM_{10}）和气体污染物（SO_2、NO_2、CO 和 O_3）。而决定 $PM_{2.5}$ 浓度水平的因素十分复杂，例如空气流动的带动下各类污染物均具扩散性，因此，区域内的 $PM_{2.5}$ 浓度水平则会受到周围区域的其他污染物浓度、风速、雨雪天气等多方面的影响。在对 $PM_{2.5}$ 浓度序列预测时，本文在考虑监序列时序特征的同时，加入空间特征，即收集周围监测站点的 $PM_{2.5}$ 浓度数据，并运用经纬度坐标以及 POI 等地理信息数据构建与目标站点之间的空间关系，以期构建更为精准的预测模型，提升空气质量的预测精度。

由此，本文提出基于图模型中拉普拉斯矩阵，利用各个站点的空间属性，即经纬度坐标以及站点周围 POI 统计信息构造拉普拉斯矩阵，与站点 $PM_{2.5}$ 浓度数据进行点乘，将空间属性嵌入 $PM_{2.5}$ 浓度数据。并作为 ELSTM 神经网络的输入，以此构建空间预测模型。

3.3.1 拉普拉斯空间特征嵌入法

（一）拉普拉斯矩阵原理

拉普拉斯矩阵是图论中用到的一种重要矩阵，给定一个有 n 个顶点的图 $G=(V,E)$，其中，$V=\{v_1,v_2,\cdots v_n\}$，$E=\{(v_i,v_j)\,|i,j=[1,\cdots,n]\}$，$(v_i,v_j)$表示定点$v_i$与$v_j$之间的边。定义$\omega_{ij}$为边$(v_i,v_j)$的权重，对于有边连接的顶点$v_i$与$v_j$，$\omega_{ij}>0$；对于没有边连接的顶点$v_i$与$v_j$，$\omega_{ij}=0$。

将G使用邻接矩阵表示，得到W，定义如（3.1）式所示：

$$W=\begin{pmatrix}\omega_{11} & \cdots & \omega_{1n}\\ \vdots & \ddots & \vdots\\ \omega_{n1} & \cdots & \omega_{nn}\end{pmatrix} \tag{3.1}$$

对于G中任意一个顶点v_i，它的度定义为和它相连的所有边的权重之和，即：$d_i=\sum_{i=1}^{n}\omega_{ij}$，进而得到度矩阵$D$，定义如（3.2）式所示：

$$D=\begin{pmatrix}d_1 & \cdots & \cdots\\ \vdots & \ddots & \vdots\\ \cdots & \cdots & d_n\end{pmatrix} \tag{3.2}$$

其拉普拉斯矩阵被定义$L=D-A$，D其中为图的度矩阵，A为图的邻接矩阵。

（二）特征嵌入流程

利用各个站点的经纬度数据、POI 的种类与数量分别构造含有空间信息的特征向量；利用 KNN 算法构建最近邻集合，得到拉普拉斯矩阵 L；将构造的拉普拉斯矩阵 L 与站点数据进行矩阵 X 相乘的操作，将不同的空间特征信息嵌入各个站点的 $PM_{2.5}$ 浓度序列；将构造的空间数据，转为监督序列，并输入进 LSTM 多变量预测模型，得到空间预测结果。

3.3.2 模型构建

空间预测模块从空间角度考虑周围空气污染对目标区域的影响，进而将其纳入模型，进而以目标站点的 $PM_{2.5}$ 浓度序列数据为对象进行预测。本文运用拉普拉斯矩阵提取周围空气污染数据的空间特征，并将结果作为深度神经网络的输入特征。并与时序模型形成呼应，构建多维 LSTM 神经网络。经建模试验后发现，当隐藏层数由一层增加至 2 层时，$PM_{2.5}$ 预测的精度显著提升，隐藏层层数继续增加时，模型的预测结果增加效果并不明显，同时运行时间大幅提升，因此，从模型整体运行效率出发，本文空间预测器中的神经网络的隐藏层数设置为 2 层。同时模型训练时所需的参数与时间预测模块的训练参数保持一致，并将结果与时间预测模块的结果进行结合，完成预测任务，构建时空预测模型。

3.4 预测结果聚合

空间与时间两个维度的预测能够将不同特征带入模型，为模型预测提供更多修正结果的可能性。而在此基础上，还需要将时间和空间维度的预测结果在不同情形下有所侧重，当目标站点风速较大时，$PM_{2.5}$ 以及其他空气污染物的扩散水平就会加快，周围的空气质量水平就会极大的影响目标区域，则考虑空间预测模块的权重应相对更大，又如在发生极端天气时，未来数小时的结果都应受其影响，则考虑时间预测模块的权重更大。因此，本文以机器学习方法，通过自动提取各类指标对预测目标的重要性，设置时空聚合器，使时间预测模块与空间预测模块的预测结果能够动态的进行集成。

本文所提出的预测聚合器使用 XGBoost（eXtreme Gradient Boosting）来聚合时间预测和空间预测的结果，利用 XGboost 模型来动态获取各自维度上特征的重要性。预测聚合器的基本思想是利用时间预测模块和空间预测模块的特征数据，作为模型的输入，构建树模型；并通过样本训练修剪分支，选择误差最小的优化子树，得到各自特征的权重，并分别在时间和空间维度上进行加总获得各自

维度的总权重；最后，对预测结果进行赋权聚合，输出预测的 $PM_{2.5}$ 预测值。

3.5 本章小结

本章主要介绍了时空模型框架及流程分，在时间维度上，利用相关性提取 $PM_{2.5}$ 与气象数据和其他污染物数据的特征，利用模态分解提取 $PM_{2.5}$ 数据自身的高波动特征、短期与长期特征。利用基于时间序列的聚类算法对序列进行重构，提高预测部分的运行效率，并基于 LSTM 模型构建预测模块；在空间维度基于拉普拉斯算子，并利用经纬度、POI 数据提取监测站点之间的空间相关性，从图模型角度构建了空间预测模块；最后，运用 XGBoost 将两部分结果进行动态聚合，得到 $PM_{2.5}$ 预测值。

4 时空预测模型实证分析

本章对兰州市的空气质量预测任务展开实证。从时间角度与空间角度构建模型，时间角度先利用兰州市气象(平均气温、气压、降水、风速、风向)和空气质量(PM_{10}、SO_2、NO_2、O_3、CO)的原始数据进行 LSTM 神经网络多指标建模；针对 $PM_{2.5}$ 目标预测序列，则利用提出的模态分解集成预测模型从数据驱动角度提取历史 $PM_{2.5}$ 数据中的高波动性与长期趋势，进行空气质量指数的预测，完成时间预测模块实证。空间角度，利用拉普拉斯算子提取经纬度与 POI 数据中目标站点与其他空气质量监测的空间关系，并将空间特征嵌入各站点 $PM_{2.5}$ 数据中，并将上述结果作为 LSTM 神经网络的输入进行建模，完成空间预测实证。将两部分预测结果通过 XGboost 进行动态聚合，完成时空预测模型的组合架构，并将预测结果与照组模型进行对比，以此证明本文模型在预测结果上的优越性。

实证部分具体内容包括：首先，对此次研究中的数据内容进行简介，包括数据来源、数据类目和数据总量，并对数据进行描述性统计分析与预处理操作；其次，对本研究结果的评价方法进行简要介绍，包括评价指标、验证方式等；再次，阐述时空数据建模的各个环节及评价结果；最后，就 $PM_{2.5}$ 浓度序列进行时空预测实证分析，利用各项模型评价指标以比较与其他模型在预测精度方面的优劣，并运用对照组数据验证模型的稳定性。

4.1 数据描述

本文收集兰州市 7 个空气质量监测站点数据，范围为 2020/10/1 到 2021/9/30 的小时空气质量数据($PM_{2.5}$、PM_{10}、SO_2、NO_2、O_3、CO)和 4 个气象站点的小时气象数据(平均气温、气压、降水、风速、风向 5 个指标)，以及各个站点的经纬度，兰州市的 POI 数据。百合公园与和平空气质量监测站为 2021 年新设置站点，故 2020 年相关数据则用相邻站点数据求平均得到。其中，训练集取 2020 年 10 月至 2021 年 8 月的数据，测试集取 2021 年 9 月至 10 月的数据。数据集中包含部分缺失值、异常值，且各数据集之间量纲、单位和数据类型均有所不同，其中风向含有特殊代码。

（1）兰州市空气质量数据，取自中国空气质量在线监测平台。获取了近一年 6 项主要空气指标的小时数据，具体包括 $PM_{2.5}$、PM_{10}、SO_2、NO_2、O_3、CO 污染物浓度数据。

（2）气象数据，通过中国天气网获取 5 项主要气象因子的历史小时数据，具体包括气温、气压、降水、风速和风向。

（3）经纬度：空气质量监测站的经纬度坐标，利用各个站点的经纬度得到站点之间的欧式距离。

（4）兴趣点数据（POI）：由于数据获取的限制，当前收集到的数据集自高德地图 2019 年的 POI 数据，利用 Arcgis 软件找到空气质量监测站点周围 1 千米距离范围内的兴趣点，并统计种类数目与数量，具体 POI 数据信息如下表 4.1 所示，处理结果如图 4.1 所示。

表 4.1 兰州市 POI 类别以及数据量

名称	数量	名称	数量
餐饮服务	281	地名地址信息	660
风景名胜	21	公共设施	73
公司企业	497	购物服务	493
交通设施服务	401	金融保险服务	415
科教文化服务	500	汽车服务	40
商务住宅	496	生活服务	302
体育休闲服务	67	医疗保健服务	86
政府机构及社会团体	363	住宿服务	305

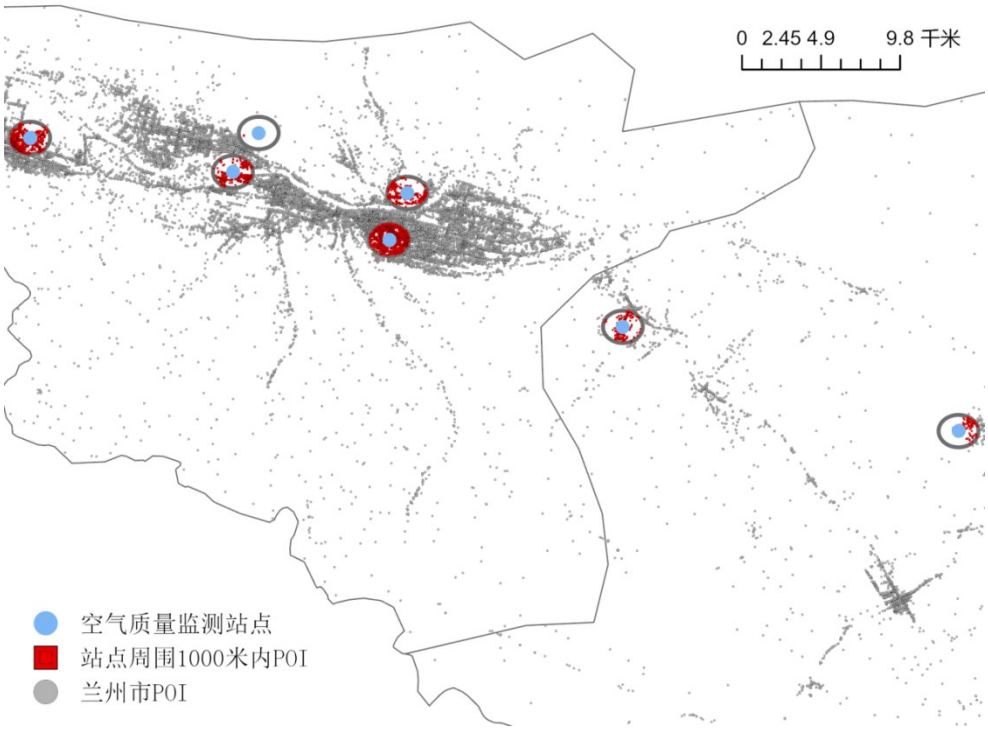

图 4.1 兰州各站点周围 1 公里 POI 信息提取结果

上图显示了兰州市 7 个空气质量监测站点，灰色点表示 POI 信息，将监测站点周围一公里内的 POI 信息运用 Arcgis 软件进行提取（用红色方块表示），进行各个站点数据信息统计，得到下表 4.2 所示结果：

表 4.2 兰州各站点周围 1 公里 POI 信息

站点名称	POI 数量	POI 种类数
兰炼宾馆	943	15
榆中兰大校区	135	11
生物制品厂	624	15
铁路设计院	2401	16
教育港	3	3
百合公园	534	16
和平	342	13

在此，以兰州市榆中县的空气质量监测站和气象监测站的站点数据，并以

$PM_{2.5}$ 浓度和气温数据为例说明空气质量数据、气象数据的基本特点。如表 4.3 所示，表中显示了榆中县监测站点 $PM_{2.5}$ 浓度序列和气温时间序列的描述性统计信息。从表中可以看出空气质量数据通常包含高波动性，其中 $PM_{2.5}$ 序列较高的偏度值说明了空气质量数据是非对称分布的，峰度值为正表明数据较标准正态分布更为陡峭；气温的峰度显示为负、则有理由拒绝分布的正态性假定，偏度也为负值，说明数据呈左偏分布。图 4.2 和 4.3 所示为兰州的 $PM_{2.5}$ 浓度值和气温情况，能明显地看出其不平稳、非正态、含极端值等特点。

表 4.3 兰州 $PM_{2.5}$ 浓度和气温的数据特征

指标	均值	中位数	标准差	最大值	最小值	偏度	峰度
$PM_{2.5}$	28.78	23	37.76	607	0	10.36	132.49
气温	8.28	8.9	10.92	33.9	-20.6	-0.15	-0.69

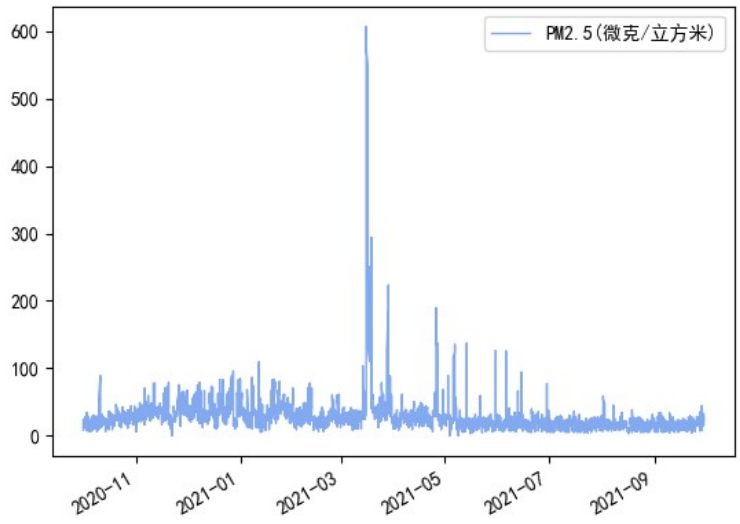

图 4.2 兰州市 $PM_{2.5}$ 浓度数据

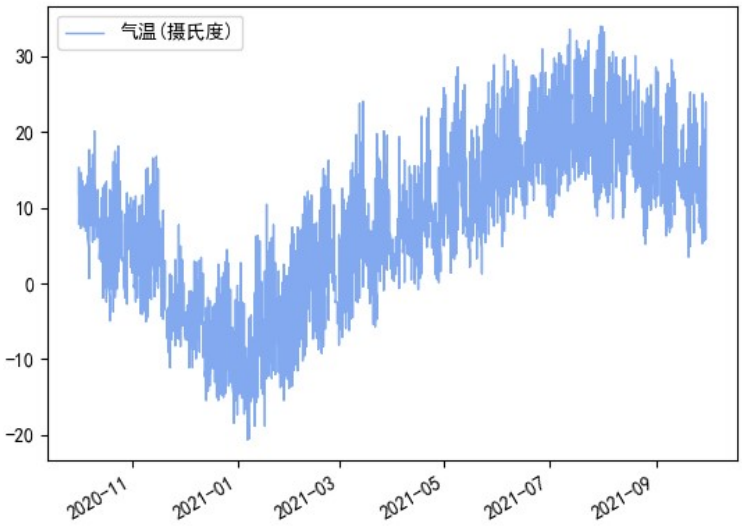

图 4.3 兰州市气温数据

4.2 评判标准

本文运用轮廓系数作为聚类结果的评价指标，轮廓系数计算组内的内聚度和组外的分离度，常用于评价算法对聚类结果所产生的影响。样本 i 的轮廓系数如公式（4.1）所示：

$$s(i)=\frac{b(i)-a(i)}{\max(a(i),b(i))} \tag{4.1}$$

其中 a 为某个样本与其所在簇内其他样本的平均距离，b 为某个样本与其他簇样本的平均距离。计算所有样本对应的轮廓系数并取均值作为该聚类结果的评价指标，其取值范围为[-1,1]。越接近 1 则聚类效果越好。

为了检验所提出模型的有效性，本文选取了下述三种被广泛采用的误差分析方法，即平均绝对误差（MAE）、均方根误差（RMSE）和平均绝对误差百分比（MAPE），这三种误差分析方法的计算公式（4.2）-（4.4）所示：

$$MAE=\frac{1}{N}\sum_{t=1}^{N}|\hat{y}(t)-y(t)| \tag{4.2}$$

$$RMSE=\sqrt{\frac{1}{N}\sum_{t=1}^{N}(\hat{y}(t)-y(t))^2} \tag{4.3}$$

$$MAPE=\frac{1}{N}\sum_{t=1}^{N}\left|\frac{\hat{y}(t)-y(t)}{y(t)}\right| \tag{4.4}$$

其中，y_t和$\hat{y}_t$表示t时刻真实值和预测值，N表示测试集的大小。

为保证所提出模型与对照模型有显著的差异性，则通过 Diebold-Mariano 统

计量进行验证，以保证本文模型的优异性。DM 检验通过假设检验的方式进行证明，原假设为本文所提出模型和对照模型的预测精度没有显著差异，备择假设为本文所提出模型的预测结果显著优于对照模型。DM 统计量定义如公式（4.5）所示：

$$S_{DM} = \frac{\bar{g}}{\sqrt{\hat{V}/T}} \tag{4.5}$$

其中，$\bar{g} = \sum_{t=1}^{T} g_t/T$ （$g_t = (x_t - \hat{x}_{A,t})^2 - (x_t - \hat{x}_{B,t})^2$），$\hat{V}_{\bar{g}} = \gamma_0 + 2\sum_{l=1}^{\infty}\gamma_1(\gamma_l)$，$\gamma_l = \mathrm{cov}(g_t, g_{t-l})$。$\hat{x}_{A,t}$和$\hat{x}_{B,t}$代表 t 时刻模型 A、B 的预测结果。当 S_{DM} 值与 p 值小于显著性水平α（0.05 或 0.01）则拒绝原假设。

为适应时间序列数据前后位置固定的要求，本文采用窗口滑动的方式对数据进行预测。即随模型对数据进行不断预测时，不断将前一小时的 $PM_{2.5}$ 真实数据送入到历史数据集中，将下一小时的数据纳入窗宽。此方法更符合预测时间序列过程中的实际，同时能够保证考虑最近历史数据影响的同时平衡较远历史数据的影响。

4.3 时空预测模型实证

模型基于 TensorFlow 框架实现，实现的详细架构则从时间与空间两部分展开，时间预测模块主要分为多变量预测与模态分解预测，多变量预测基于其他污染物浓度以及气象因素对 $PM_{2.5}$ 浓度序列的影响，在特征选择后，构建多变量预测模块。对于 $PM_{2.5}$ 自身的高波动性以及短期特征与长期趋势的提取，则运用“分解-聚类-集成”的研究范式，将 $PM_{2.5}$ 浓度序列经由模态分解的方式分解为多个子序列，子序列中则提取了原序列中的高频，低频信息以此表征 $PM_{2.5}$ 浓度序列的短期特征与长期趋势特征。基于 VMD 方法，将一层模态分解结果中预测效果相对较差的序列，进行二次分解，充分提取序列中的时序特征。采用时间序列聚类算法（K-shape），将具有相似特征的分信号进行聚类，对分信号进行重构，以此提高模型的运行效率。最后运用深度学习的方法即 ELSTM 神经网络进行预测并将预测结果线性集成。空间预测模块将由经纬度构建的各个站点之间的欧式距离，以及由 POI 数据提取的 POI 数量及种类，完成空间特征的构造，由此计算拉普拉斯算子并与 $PM_{2.5}$ 站点数据矩阵进行点乘，嵌入空间特征。最后采用 LSTM 预测模块进行预测。为将时间预测模块与空间预测模块动态集成，运用

XGboost 回归对预测结果动态赋权，完成时空预测结果。

4.3.1 时间预测流程及结果

（1）数据预处理

在空气质量数据和气象数据实时采集过程中，都必须依赖于监测站有效运转，但设备老化、损坏、以及设备维护中就会造成数据缺失等问题。本文就此问题展开研究，重点分析数据的异常处理与缺失数据的填充。在此次研究中，对数据进行收集时，其中存在异常内容，气象数据中部分风向数据值显示为 999907 远大于正常值 360，在查阅相关资料后发现 999907 代表无风状态，故修改为 0；气压等气象数据也存在数值过大的异常值，则进行删除，并采用第二章介绍的处理缺失数据的方法进行填充。

（2）特征选择

将采集到的空气污染物数据以及气象因子数据进行相关分析，首先计算 Spearman 相关系数矩阵，并利用矩阵散点图进行分析，如图 4.3 所示，为兰州市 $PM_{2.5}$、PM_{10} 等 6 个污染物指标和平均温度、气压等 5 个气象指标的散点分布图：

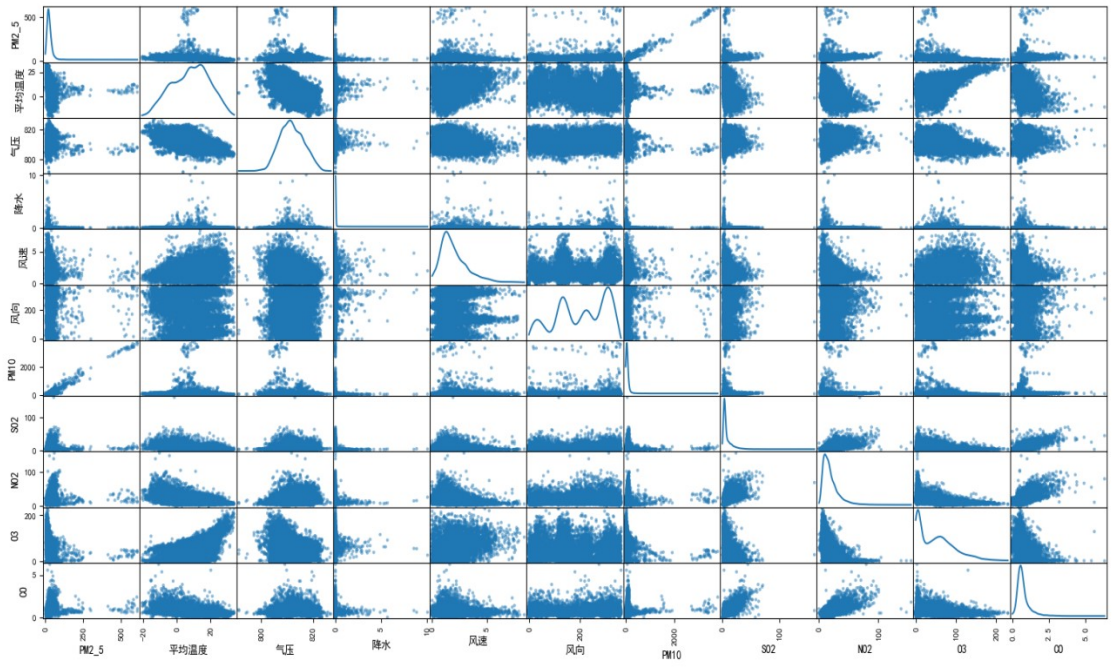

图 4.4 矩阵散点图

从图 4.4 中可以得出兰州市 $PM_{2.5}$ 浓度序列与其他空气污染物浓度序列、气

象条件的相关性，$PM_{2.5}$ 浓度与 PM_{10}、NO_2、SO_2、气压、温度和降风速相关性较强，与降水的相关性较弱。其中，既有正的相关关系，也有负的相关关系。

为了更加仔细地度量 $PM_{2.5}$ 浓度跟其他影响因子之间的亲密程度，本节计算了 $PM_{2.5}$ 与其他污染物浓度、5 个气象属性的 Spearman 等级相关系数。相关系数如表 4.4 所示：

表 4.4 $PM_{2.5}$ 与其他污染物和气象因素 Spearman 系数

其他空气污染物因素		气象因素	
PM_{10}	0.763	平均气温	-0.588
SO_2	0.385	气压	0.242
NO_2	0.471	降水	-0.062
O_3	-0.409	风速	-0.152
CO	0.380	风向	0.147

从上表反映了 $PM_{2.5}$ 与 PM_{10}、SO_2 等空气污染物以及气温、气压等气象因素之间的相关联系水平。将相关系数取绝对值，并且采用学界一致的度量水平，认为 0-0.1 为没有相关性，0.1-0.3 为弱相关，0.3-0.5 为中等相关，0.5-1.0 为强相关[37]。从上表中，能够简便的得到空气污染物、气象因素之间的因果或关联关系，为下文建模提供支持。本文主要探究特征因素对 $PM_{2.5}$ 的影响程度，由 Spearman 等级相关系数得到，在其他空气污染物中 PM_{10} 与 $PM_{2.5}$ 有强的正相关性，其他空气污染物因素与 $PM_{2.5}$ 有中等程度的相关性。气象因素中气温跟与 $PM_{2.5}$ 有较强的负相关性，气压、风速、风向与 $PM_{2.5}$ 有较弱的相关性，降水则与 $PM_{2.5}$ 相关性较差。本研究则选取相关系数绝对值在 0.1 以上为特征选择依据。

（二）预测流程

运用较为流行的多变量预测方式，即多变量 LSTM 预测模型，提取其他空气质量污染物数据以及气象数据对 $PM_{2.5}$ 浓度序列的影响。训练集与测试集与其他模块保持一致，分别取数据的 80%作为训练集，20%作为测试集，模型超参数包括时间窗口大小、卷积核数量、批处理大小、训练次数和隐藏层单元个数，取值如下：时间窗口大小设定为 12，卷积核数量设定为 64、批处理大小为 72，LSTM

层隐藏层单元个数为 128。经过反复试验，发现当模型迭代到 120 次左右误差损失变化趋于稳定，损失函数呈收敛状。因此设定模型的训练次数为 120 次。预测结果如图 4.5 所示：

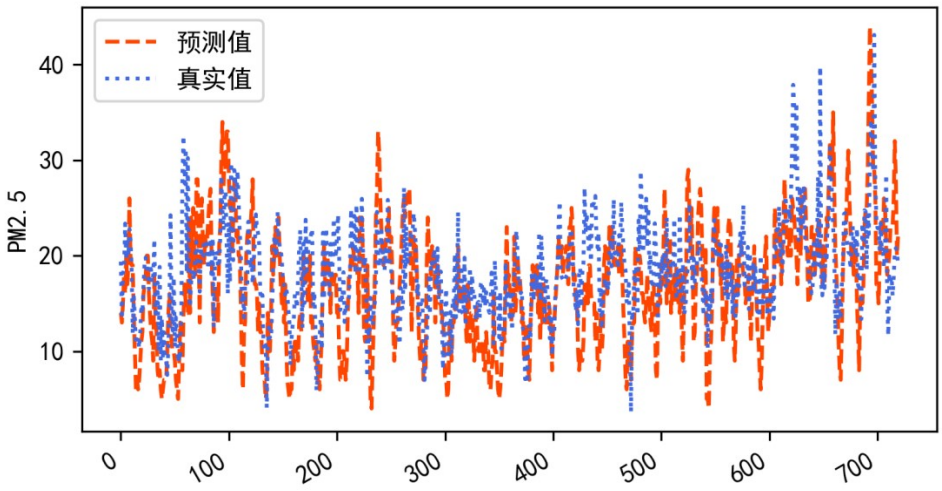

图 4.5 多变量预测结果

多变量预测效果并不理想，为提高预测精度，需提取 $PM_{2.5}$ 自身序列特征，本文则运用模态分解的方式，对 $PM_{2.5}$ 浓度序列自身具有的高波动性，并基于模态分解的优势，提取序列短期特征与长期趋势。将原始 $PM_{2.5}$ 浓度时间序列运用 CEEMDAN 方法分解为多个子序列，结果如图 4.6 所示：

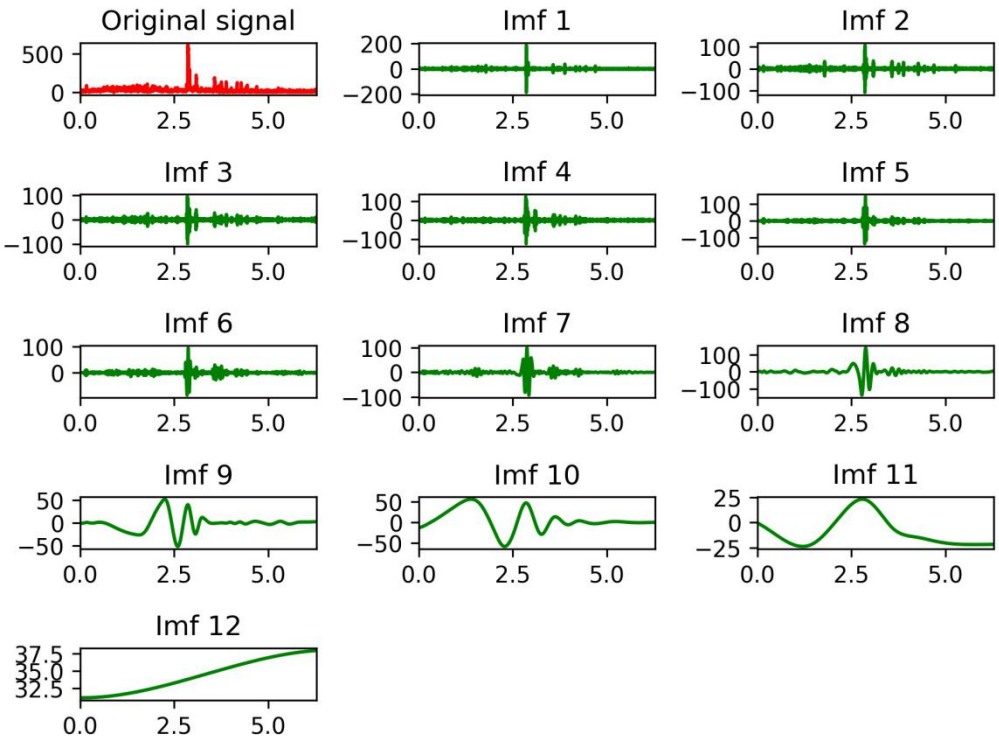

图 4.6 $PM_{2.5}$ 原始序列 CEEMDAN 分解图

在图 4.5 所示中，$PM_{2.5}$ 浓度序列被分解为 11 条分量以及 1 条趋势序列，且按照分量波动频率从高到低进行依次排列，不同频率的分量提取了不同的波动。高频分量主要内涵原始 $PM_{2.5}$ 浓度序列的震荡特征，低频序列以及趋势序列则内涵原序列的趋势特征。在模型运行前，对分解后的序列进行一阶差分，以缓解数据的不平稳性，然后运用 ELSTM 网络模型学习每一条分量以及趋势序列的特征。由于采用小时数据，认为 $PM_{2.5}$ 小时浓度序列中午夜与白天的 $PM_{2.5}$ 浓度有较明显的不同，故在对测试集进行预测时，采用原序列中连续 12 个数据点预测第 13 个数据点并向后连续滑动。算法的迭代次数为 100 次，每个小批量中包含的样本数为 50，并将以上超参数运用于本文的对比模型中，确保在模型比较环节中的公平性和有效性。基于上述设定，使用 ELSTM 网络对 CEEMDAN 分解后的各个分量进行预测，并在此处展示高频分量预测结果，如图 4.7 所示：

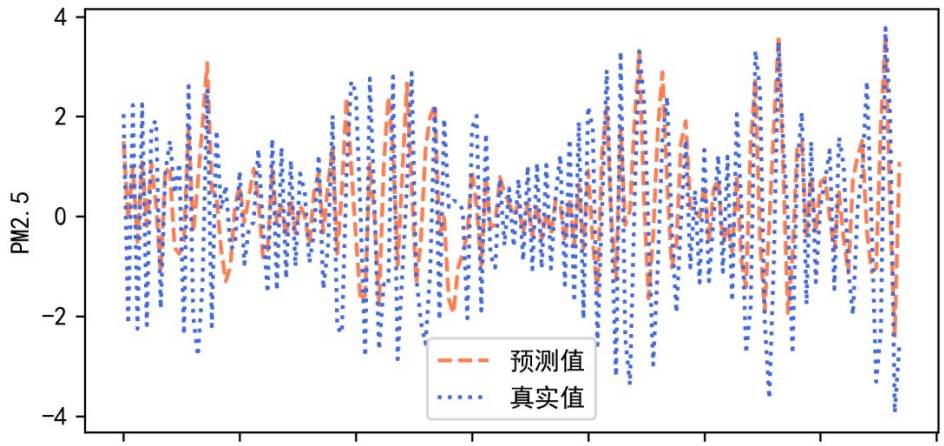

图 4.7 高频信号序列的 ELSTM 预测图

在图 4.7 所示的预测结果中，预测值与真实值较为接近，但频率波动较大，包含的非线性特征明显，预测效果也较差。由此，充分提取高频信号中依然存在较为复杂的非线性性和非平稳性，从而提高模型精度，本文对高频分量采取 VMD 分解，以获得分信号的数据特征。将分解模态数 K 设定为超参数，本文通过各模态中心频率接近程度确定分解模态数，分量中出现中心频率较为接近的情况，即 VMD 出现过分解。高频信号选定不同 K 值分解后，各分量中心频率如表 4.5 所示：

表 4.5 不同 *K* 值对应中心频率

K 值	IMF5	IMF6	IMF7	IMF8	IMF9	IMF10	IMF11	IMF12	IMF13
5	23940								
6	15261	23954							
7	14736	19023	24383						
8	14235	17124	21315	25699					
9	11337	14345	17181	21355	25715				
10	11290	14168	16737	19508	22937	26124			
11	10491	12356	14606	16973	19659	23071	26161		
12	8051	11275	14065	16498	18990	21437	24060	26437	
13	7448	11453	12278	14514	16790	19137	21526	24094	26449

由表 4.5 可知，当模态数为 13 时分量 6 与分量 7 的中心频率相较为接近，即出现模态混叠现象，故确定分信号数量为 12。图 4.7 即为高频信号经 VMD 分解后的时域图，如图 4.8 所示：

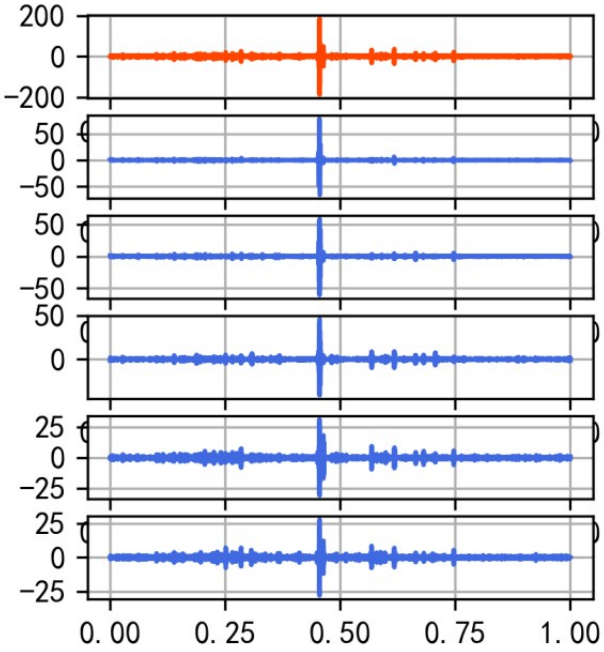

图 4.8 高频信号 VMD 分解时域图

将高频信号做二次分解以获取高频信号中的非线性特征，但同时会使子序列的数量成倍增加，加之 ELSTM 神经网络的训练过程本就复杂，使得预测模块的训练时间过长。为此运用时间序列聚类算法（K-shape），将具有相似特征的分信号进行聚类，聚类结果如表 4.6 所示：

表 4.6 高频分量聚类结果

高领分量	聚类个数	轮廓系数
IMF1	5	0.115
IMF2	6	0.257
IMF3	7	0.273

由表 4.6 所示，将高频分量的 VMD 分解结果进行聚类，并取轮廓系数最大的聚类个数作为聚类结果。减少分信号的个数并完成数据重构。

利用上述运算得到的数据，以及预训练确定的模型参数作为预测的初始参数，对每个分信号数据集划分训练集与测试集、并将数据转化为监督问题的可训练形式，与 ELSTM 神经网络的数据输入要求进行匹配。模型运行结束，需要将结果进行逆差分、逆标准化，以还原为正常的 $PM_{2.5}$ 预测数值。最后将各个分信号的预测结果进行叠加，作为最终的预测值。

将模态分解预测结果与多变量预测结果进行集成，一方面能够提取 $PM_{2.5}$ 浓度数据的历史特征，另一方面有效的结合了其他污染物数据、气象数据对 $PM_{2.5}$ 浓度序列的影响，最终预测结果如图 4.9 所示：

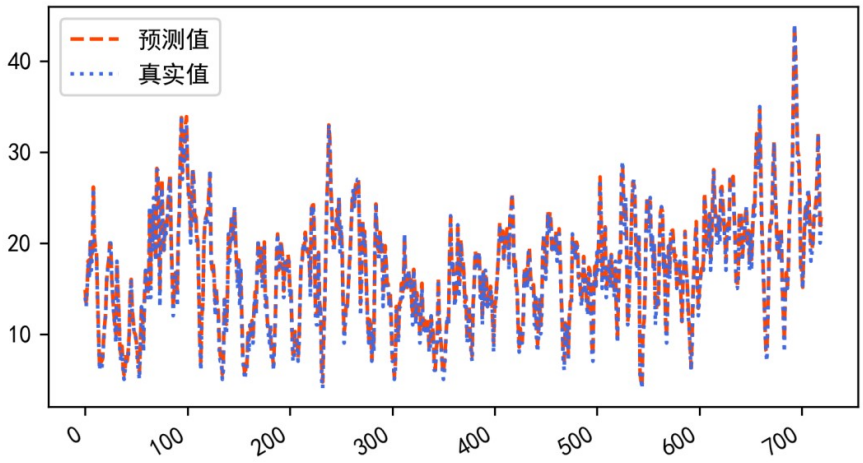

图 4.9 时间预测效果图

4.3.2 空间预测流程及结果

本文获取兰州市 7 个空气质量监测站与 4 个气象监测站近一年的数据，并运用本文提出的拉普拉斯嵌入法，将空间距离特征与 POI 地理信息特征嵌入 $PM_{2.5}$ 浓度序列，并将 2020 年 10 月至 2021 年 8 月的数据作为训练数据，使模型学习数据中的特征，并生成神经网络结构。超参数包括：每个小批量中要包括的样本数（64），隐藏层神经元个数（128）。模型运行结果如图 4.10 所示，预测模型损失函数 MSE 在经过一次次的训练学习后，显著缩小并在模型迭代次数为 200 次左右达到稳定，证明在迭代训练后 LSTM 网络中各个参数的组合达到最优。由于兰州市榆中县的 $PM_{2.5}$ 浓度序列数值在范围为 0-600 中，相对集中分布在 40-50 左右，但也存在极端天气的影响，因此模型初始训练时损失函数值较大，

经过多次迭代学习，预测模型的损失函数值显著减小，达到理想水平，表明预测性显著提高且趋于稳定状态，如图 4.10 所示：

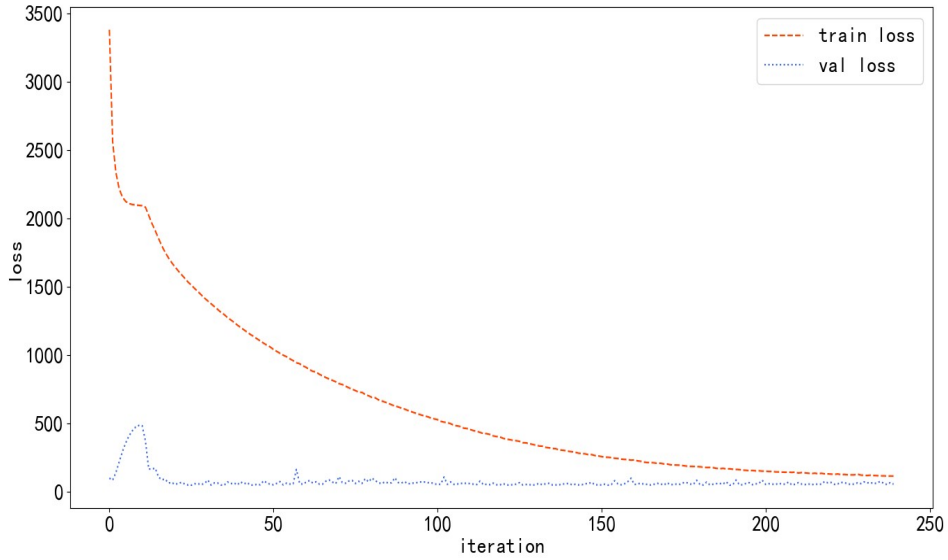

图 4.10　模型损失函数收敛图

训练空间预测器时，以榆中站点为预测目标，依次加入嵌入经纬度特征的分站点 $PM_{2.5}$ 数据和嵌入 POI 特征的分站点 $PM_{2.5}$ 数据，有效地提高了预测精度。其预测结果如图 4.11 所示：

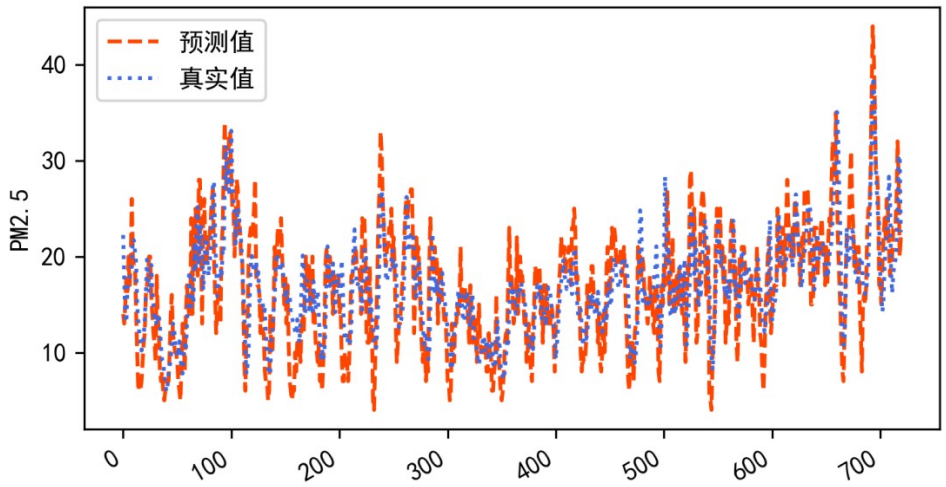

图 4.11　空间预测结果

4.3.3 时空模型集成结果

为了将时间预测模块和空间预测模块的预测结果动态集成，通过 XGboost 算法得到各维度特征的重要性，以此构造预测聚合器将两部分预测结果结合，以充分考虑不同时间段空气污染物数据、气象条件等对 $PM_{2.5}$ 预测造成不同程度的影响。特征重要性如图 4.12 所示：

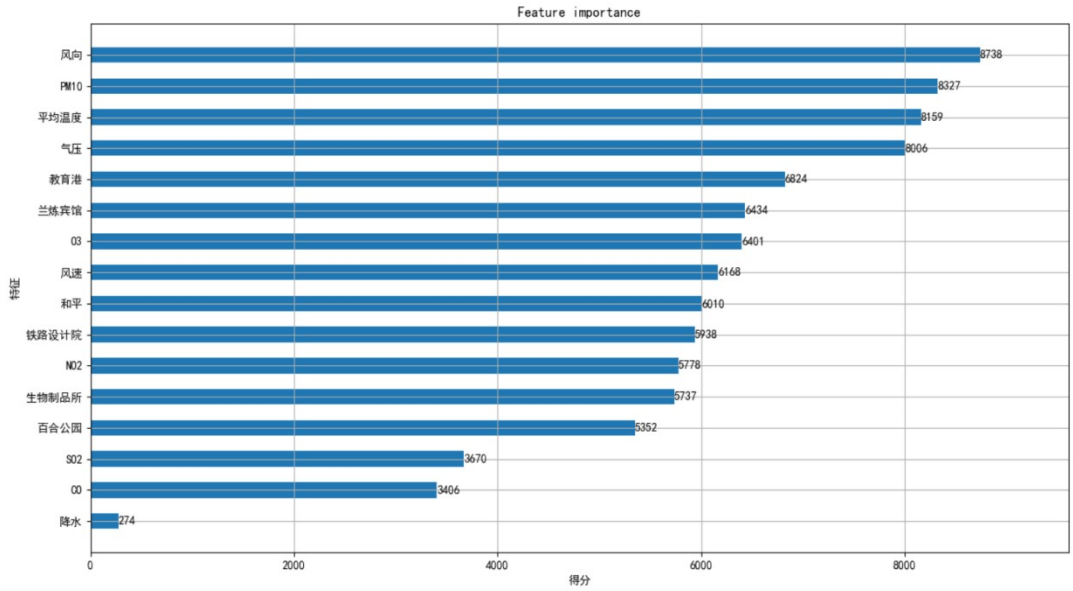

图 4.12 XGboost 特征重要性得分图

由于预测聚合器能够从数据样本中动态的学习各个特征对 $PM_{2.5}$ 数据的重要性，使得预测聚合器能够进一步提高模型的预测效果。如表 4.7 结果所示，时空聚合结果在 RMSE 中取得了最优的结果，在 MAE 与 MAPE 指标中提升了空间预测器的预测结果并与时间预测器的结果接近。

表 4.7 时空预测聚合结果

预测器	RMSE	MAE	MAPE
时间预测器	2.974	2.216	0.414
空间预测器	3.905	3.108	0.498
时空预测聚合器	2.675	2.293	0.427

4.4 模型比较及鲁棒性

4.4.1 模型比较

为了说明本文所提出的时空组合模型的有效性，时间预测模块将所提出模型与单一模型、一次分解集成模型进行对比，探究随着模型复杂度的提升，以及分解方法的组合对预测结果的影响，并利用 DM 检验分析模型之间的预测精度是否存在显著性差异。下表为各个模型在兰州 $PM_{2.5}$ 浓度预测中的误差值，分别从 RMSE、MAE 和 MAPE 对所有预测模型的性能进行评价。

表 4.8 时间预测模块结果对比

	RMSE	MAE	MAPE
单变量 LSTM	9.528	8.317	0.914
多变量 LSTM	5.872	4.857	0.623
结合一层模态分解	3.253	2.496	0.503
结合双层模态分解	2.974	2.216	0.454

由表 4.8 预测结果所示，结合其他空气污染物与气象特征时，预测效果较单序列有显著提升更好。基于“分解-聚类-集成”的组合模型在预测性能上均优于多变量 LSTM 模型，表明“分解-聚类-集成”研究范式可以有效克服因 $PM_{2.5}$ 浓度数据的高波动性、非线性性对模型预测精度造成的影响，显著提高模型的预测能力。二次分解模型在 RMSE 与 MAE 两类评价标准中表现效果最好，表明对预测效果提升起到了一定的作用。

表 4.9 空间预测模块结果对比

特征	RMSE	MAE	MAPE
单序列预测	4.708	3.623	0.723
附加经纬度预测	4.174	3.292	0.548
附加经纬度与 POI 预测	3.905	3.108	0.498

空间预测模块预测结果如表 4.9 所示，加入经纬度特征后，三个指标的预测结果均有较大提高，故说明空间信息的提取对预测精度的提升有较大的改善作

用，在加入反映经济社会因素的 POI 数据特征后，预测结果仍有所提升，由此也说明，空间信息的重要性，以及探讨 $PM_{2.5}$ 预测中的空间因素的必要性。

将时间预测模块中的模型与空间预测模块中的模型相结合，即得到时空空预测模型。模型充分考虑 $PM_{2.5}$ 预测时时间角度与空间角度的影响因素，并从评价指标中探寻模型优劣。模型预测效果如表 4.10 所示：

表 4.10 时空模型预测结果对比

	RMSE	MAE	MAPE
LX-M-LSTM	4.217	3.368	0.607
LX-M-CEEMDAN-LSTM	3.034	2.443	0.531
LX-M-CEEMDAN-VMD-LSTM	2.675	2.293	0.427

本文提出的模型 LX-M-CEEMDAN-VMD-LSTM，时间角度运用多变量与模态分解的方式挖掘 $PM_{2.5}$ 序列与其他污染物浓度序列、气象数据的关系，以及 $PM_{2.5}$ 自身的短期波动性和长期趋势，空间上运用拉普拉斯算子提取此站点与其他站点的空间信息。从模型指标上可以看出，模型的预测效果显著优于空间预测结果，故说明时空预测模型的有效性。

表 4.11 DM 检验结果

	LX-LSTM	LX-M-LSTM	LX-M-CEEMDAN-LSTM	LX-M-CEEMDAN-VMD-LSTM
LX-LSTM	—	-1.9597 （0.0483）	10.0727 （0.0000）	11.1307 （0.0000）
LX-M-LSTM	1.9597 （0.0483）	—	16.6922 （0.0000）	17.9841 （0.0000）
LX-M-CEEMD AN-LSTM	-10.0727 （0.0000）	-16.6922 （0.0000）	—	7.5061 （0.0000）
LX-M-CEEMD AN-VMD-LSTM	-11.1307 （0.0000）	-17.9841 （0.0000）	-7.5061 （0.0000）	—

为判断所提出模型的预测结果是否在统计学上显著优于基准模型，运用 DM 统计量对其进行检验。DM 检验即根据 DM 统计量的值判断模型之间的预测精度是否在统计意义上具有显著差异。结果如表 4.11 所示。以本文提出的 LX-M-CEEMDAN-VMD-LSTM 作为测试模型时，其预测精度在 0.01 的显著性水平下，认为显著优于其他的基准模型。

4.4.2 鲁棒性分析

为了进一步验证模型的有效性，检验数据变化对模型的鲁棒性影响，本文选取同样时间段生物制品所空气污染物检测站点的 $PM_{2.5}$ 浓度序列，城关区气象监测站的气象数据（与生物制品所监测站距离最接近的气象检测站点）验证模型是否仍能保持较好的预测性能以及稳定性。模型参数设置与前文保持一致。下图为 LX-M-CEEMDAN-VMD-LSTM 模型与基准模型在不同数据集的预测比较结果。如图 4.12 所示：

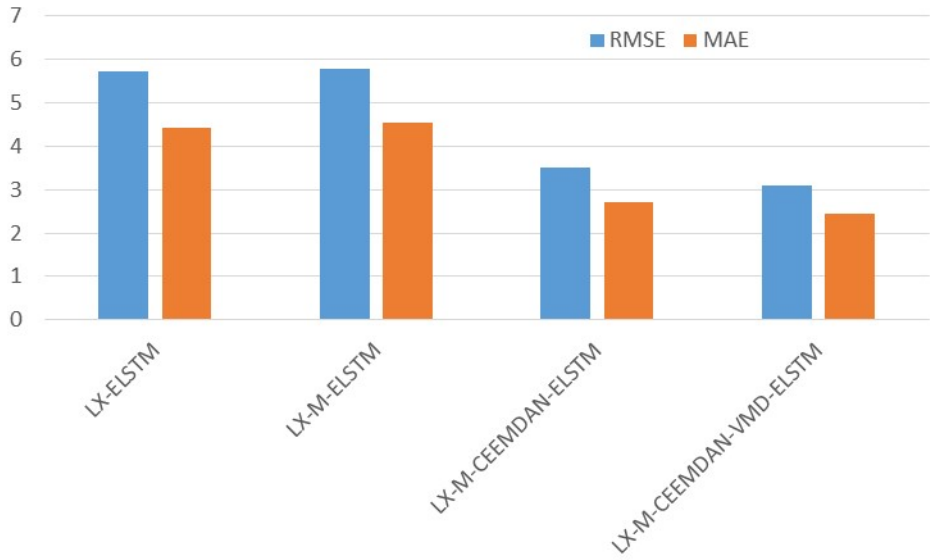

图 4.13 不同数据集上预测模型误差对比

由于生物制品所监测站点的原始 $PM_{2.5}$ 浓度序列存在数值 0，故 MAPE 指标不能用于此次评价，但在 RMSE 与 MAE 评价指标中，本文所提出的 LX-M-CEEMDAN-VMD-LSTM 组合模型在此数据集上的预测精度仍取得了最高水平，即模型能够对不同的 $PM_{2.5}$ 浓度时间序列做出较为准确的预测，具有良好的鲁棒性。

4.5 本章小结

本章主要介绍了本文的数据准备与实验环境，详细介绍了本文的时空预测方法的过程及其结果，并对模型的鲁棒性进行了分析。在时间预测模块中说明了基于 LSTM 神经网络预测模块的多指标、多模态建模的一般方法、建模流程，利用兰州市空气污染物浓度数据以及气象数据对 $PM_{2.5}$ 浓度序列进行预测，提出了“分解-聚类-集成”的基本框架、构造方法。在空间预测模块中，利用站点经纬度以及附近 1 公里内的 POI 数据构造站点之间的相关性，运用拉普拉斯算子将站点空间相关性嵌入 $PM_{2.5}$ 浓度序列中，完成空间特征的提取，对 $PM_{2.5}$ 浓度序列进行预测。为进一步说明 LX-M-CEEMDAN-VMD-LSTM 模型预测效果，本章设置了对照组实验以及鲁棒性实验。以 RMSE、MAE、MAPE 为模型评价指标佐证文章说提出时空组合模型的有效性，以不同站点数据重新完成模型预测流程，以此验证模型具有较好的鲁棒性。

5 总结与展望

5.1 研究工作总结

空气质量预测数据是为了得到某一地区在接下来的一段时间内的空气污染状况，通过精准的预测结果，为公众制定出行安排提供提示，为政府制定大气污染预防及治理政策提供数据支撑，同时也能为环境保护工作提供辅助作用，因此对于空气质量预测模型的研究具有实际意义。

本文通过研究大量空气质量及其时空数据预测方面的文献，对于影响 $PM_{2.5}$ 预测的时间与空间两个角度进行充分研究。在时间相关性特征提取中，运用多变量 LSTM 神经网络提取 $PM_{2.5}$ 浓度序列与其他污染物序列、气象数据在时间角度上的相关性，并提出基于“分解-聚类-集成”研究范式的多模态集成预测模型；在空间特征提取中，提出了一种拉普拉斯嵌入法，将经纬度反映的距离因素以及兴趣点数据反映的人员聚集、社会发展等空间因素对 $PM_{2.5}$ 浓度序列的影响，解决空气质量预测时对于空间影响因素考虑不充分问题，旨在提高 $PM_{2.5}$ 浓度序列的预测精度。并将时间预测模块与空间预测模块利用科学的方法进行动态聚合，形成 LX-M-CEEMDAN-VMD-ELSTM 的组合模型。本文的主要工作展示如下：

（1）首先，对 $PM_{2.5}$ 预测的研究背景和研究意义进行了介绍，对 $PM_{2.5}$ 预测模型目前的发展方向和现状进行深度研究，在时序预测问题中，从仅依靠历史数据变化，采用传统统计分析模型拟合数据变化趋势进行预测；到利用支持向量机模型，通过机器学习算法提高对数据非线性的提取；再到目前的神经网络模型，深度学习模型等，总结了前人的研究，对现有模型的优点和存在的问题进行阐述。并对时序预测问题中对其他时序特征以及历史特征提取进行了深入研究，并且分析了各种方法的优点与不足。

（2）针对空间因素考虑不充分，缺乏简洁高效的特征提取方法获取空间特征，本文提出一种基于拉普拉斯嵌入法，通过经纬度数据、POI 数据分别构建拉普拉斯矩阵，从图模型的角度提取各个站点之间的空间关系，并与各个站点所形成的 $PM_{2.5}$ 数据矩阵进行矩阵相乘，以此将空间特征嵌入时序数据，解决空间影响因素没有充分考虑的问题。

（3）针对 $PM_{2.5}$ 浓度预测本质上是时序预测的问题，以此作为出发点，首先进行相关性分析，通过对实验结果的分析可以初步得到以下结论：$PM_{2.5}$ 浓度同

PM_{10} 及气温有较强的相关关系，与 SO_2、NO_2、O_3、CO 有中等程度的相关性，与气压、降水、风速、风向相关程度较弱。其次对 $PM_{2.5}$ 与其他空气污染物及其相关气象进行预测建模，运用当下时序预测中较为流行的深度神经网络进行多变量建模。同时，为更好的提取 $PM_{2.5}$ 浓度序列自身的短期高波动性以及长期趋势，本文提出基于“分解-聚类-集成”研究范式的二层分解多模态预测模型。最终将两个模型聚合实现 $PM_{2.5}$ 时序预测，提取 $PM_{2.5}$ 浓度序列和其他污染物浓度序列、气象数据在时间上的依赖性以及自身历史特征。最后空间预测模块与时间预测模块进行集成得到最终预测结果，完成 $PM_{2.5}$ 时空预测，并通过多组对比实验，验证本文提出的 LX-M-CEEMDAN-VMD-ELSTM 组合模型的预测精度和误差明显优于其他对比模型。

5.2 预测系统构想

在构建好空气质量时空预测模型，并得到 $PM_{2.5}$ 等污染物浓度序列的结果后，如何能够及时的发布给公众相关信息也是需要我们做好思考。在本文的最后，则对空气质量预测结果的发布系统提出一些想法，其目标是低成本，高效率的实现预测结果发布，做好从数据预测到惠及民生的最后一环。

平台的建设要做好数据的采集，数据的存储、数据的处理与融合、模型的运转与修正、以及预测信息的可视化。以兰州市为例，首先，在数据的采集方面，做好与政府的沟通，以便于后台程序能够实时的合法的提取空气质量数据与气象数据。在数据存储方面，可以根据数据集的大小搭建相对低廉的数据库进行数据储存。其次，模型的发布方面，在确保模型代码的稳健性后，将模型代码发布至服务器上，做好与数据库的连接，并做好管理员模块的搭建以方便项目工程人员能够便捷的调整模型参数。最后，与计算机编程人员合作完成系统前台页面的编写，包括管理员 $PM_{2.5}$ 界面、公众访问界面，确保公众的体验。

5.3 未来工作展望

本文以浓度序列为目标，对空气污染物浓度序列预测问题中时空因素进行了一系列研究，分别从时间与空间角度提出了新颖的方法并将二者进行集成，提出了时空预测模型。但对于目前的研究结果，仍存在一些不足，并有大量工作可以进一步进行拓展延伸，深入研究：

首先，本文仅选取兰州市的相关数据，尚未分析具有不同时空特征的城市数

据，应对其他城市的时空特征进行分析，研究不同时空特征下 $PM_{2.5}$ 时空预测问题，以此作为问题研究的补充；其次，$PM_{2.5}$ 浓度序列以及空气质量的影响因素不仅局限于空气本身和气象因素，当地的人口规模、产业结构和特殊的地理地貌等经济与地理因素也应需要充分发掘与空气质量的关系，并且这些因素如何进行转换，如何从多个视角进行度量与当地空气质量变化的影响，为空气质量预测做好支撑作用，十分具有现实意义；再次，关于时空预测模型中时间预测模型与空间预测模型的集成方面，本文仅通过线性组合的方式将不同模型进行了组合。而在实际问题中，应考虑不同时刻时间特征与空间特征对 $PM_{2.5}$ 浓度序列预测的影响可能不同，以此开展更进一步的研究。最后，有必要继续研究其相关预测工具或集成模式的性能，以便取长补短，继续提高预测性能。

参考文献

[1] Dragomiretskiy K, Zosso D. Variational Mode Decomposition[J]. IEEE Transactions on Signal Processing, 2014,62(3):531-544.

[2] Gravano, Luis, Paparrizos, et al. k-Shape: Efficient and Accurate Clustering of Time Series[J]. SIGMOD record: ACM SIGMOD (management of data), 2016,45(1):69-76.

[3] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[J]. IEEE, 2016:770-778.

[4] Huang C J, Kuo P H. A Deep CNN-LSTM Model for Particulate Matter ($PM_{2.5}$) Forecasting in Smart Cities[J]. Sensors, 2018,18(7):2220.

[5] Jiang F, Zhang C, Sun S, et al. A Novel Hybrid Framework for Hourly $PM_{2.5}$ Concentration Forecasting Using CEEMDAN and Deep Temporal Convolutional Neural Network[J]. arXiv preprint arXiv:2012.03781, 2020.

[6] Jitendra Kumar, Rimsha Goomer,Ashutosh Kumar Singh. Long Short Term Memory Recurrent Neural Network (LSTM-RNN) Based Workload Forecasting Model For Cloud Datacenters[J]. Procedia Computer Science, 2018,125:676-682.

[7] Le J, Yun Z, Zhu Y, et al. An application of ARIMA model to predict submicron particle concentrations from meteorological factors at a busy roadside in Hangzhou, China[J]. Science of The Total Environment, 2012,426:336-345.

[8] Li W, Tao W, Qiu J, et al. Densely Connected Convolutional Networks With Attention LSTM for Crowd Flows Prediction[J]. IEEE Access, 2019,7:140488-140498.

[9] Liu Y, Zheng H, Feng X, et al. Short-term traffic flow prediction with Conv-LSTM: 2017 9th International Conference on Wireless Communications and Signal Processing (WCSP), 2017[C]. IEEE.

[10] Niu M, Gan K, Sun S, et al. Application of Decomposition-Ensemble Learning Paradigm withPhase Space Reconstruction for Day-Ahead PM 2.5 Concentration Forecasting[J]. Journal ofEnvironmental Management. 2017,196: 110-118.

[11] Ping W, Yong L, Qin Z, et al. A novel hybrid forecasting model for PM_{10} and

SO_2 daily concentrations[J]. Science of the Total Environment, 2015,505:1202-1212.

[12] Sun W, Zhang H, Palazoglu A, et al. Prediction of 24-hour-average $PM_{2.5}$ concentrations using a hidden Markov model with different emission distributions in Northern California[J]. Science of the Total Environment, 2013,443:93-103.

[13] Torres M E, Colominas M A, Schlotthauer G, et al. A complete ensemble empirical mode decomposition with adaptive noise: 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2011[C]. IEEE.

[14] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need[J]. arXiv, 2017:6000-6010.

[15] Wang J, Wu C, Niu T.A Novel System for Wind Speed Forecasting Based on Multi-ObjectiveOptimization and Echo State Network[J]. Sustainability. 2019,11(2): 526.

[16] Wang Z, Hu L, Li J, et al. Magnitude, temporal trends and inequality in global burden of tracheal, bronchus and lung cancer: findings from the Global Burden of Disease Study 2017[J]. BMJ Global Health, 2020,5(10):e2788.

[17] Wei S, Jingyi S. Daily $PM_{2.5}$ concentration prediction based on principal component analysis and LSSVM optimized by cuckoo search algorithm[J]. Journal of Environmental Management, 2017,188:144.

[18] Xiong S, Wang C, Fang Z, et al. Multi-Step-Ahead Carbon Price Forecasting Based on Variational Mode Decomposition and Fast Multi-Output Relevance Vector Regression Optimized by the Multi-Objective Whale Optimization Algorithm[J]. Energies, 2019,12(1):147.

[19] Yanlai Z, Fi-John C, Li-Chiu C, et al. Multi-output support vector machine for regional multi-step-ahead $PM_{2.5}$ forecasting[J]. Elsevier, 2019,651:230-240.

[20] Yin H, Dong Z, Chen Y, et al. Peformance analysis of four decomposition-ensemble models for one-day-ahead agricultural commodity futures price forecasting[J]. Algorithms, 2017,10(3):108.

[21] Zhang J, Zheng Y, Qi D, et al. DNN-Based Prediction Model for Spatial-Temporal Data[J]. AMC, 2016:1-4.

[22] Zhang J, Zheng Y, Qi D. Deep spatio-temporal residual networks for citywide crowd flows prediction[J]. AAAI, 2017:1655-1661.

[23] 白盛楠, 申晓留. 基于 LSTM 循环神经网络的 $PM_{2.5}$ 预测[J]. 计算机应用与软件, 2019,36(01):67-70.

[24] 蔡韬. 基于城市大数据的细粒度空气质量预测与推测模型研究及应用[D]. 南京邮电大学, 2020.

[25] 龚明, 叶春明. 基于修正灰色马尔科夫链的上海市 $PM_{2.5}$ 浓度预测[J]. 自然灾害学报, 2016(25):97-104.

[26] 郝胜轩, 宋宏, 周晓锋. 基于近邻噪声处理的 KNN 缺失数据填补算法[J]. 计算机仿真, 2014,31(7):5.

[27] 侯俊雄，李琦，林绍福，等.门限重复单元的 $PM_{2.5}$ 浓度预报方法[J].测绘科学.2018,43(7);79-86.

[28] 黄恒君, 王伟科. 基于深度学习的 $PM_{2.5}$ 多模态集成预测应用[J]. 统计学报, 2020,1(2):39-47.

[29] 蒋锋, 乔雅倩. 基于样本熵和优化极限学习机的 $PM_{2.5}$ 浓度预测[J]. 统计与决策, 2021,37(3):166-171.

[30] 蒋洪迅, 石晓文, 孙彩虹, 等. 基于 DLENN 模型的沈阳地区 $PM_{2.5}$ 浓度预测[J]. 系统工程, 2021,39(2):13-21.

[31] 荆海航. 基于时空数据的空气质量预测模型研究[D]. 哈尔滨工程大学, 2020.

[32] 康晓明, 崔丽娟, 赵欣胜, 等. 北京市湿地削减大气细颗粒物 $PM_{2.5}$ 功能[J]. 生态学杂志, 2015,34(10):2807-2813.

[33] 黎维, 陶蔚, 周星宇, 等. 时空序列预测方法综述[J]. 计算机应用研究, 2020,37(10):2881-2888.

[34] 李波, 陈百菊. 基于变分模态分解和优化递归最小二乘的自适应波束成形算法[J]. 信息与控制, 2020,49(6):7.

[35] 李军, 李佳, 张世义, 等. 采用 EEMD 算法与互信息法的机械故障诊断方法[J]. 华侨大学学报（自然科学版）, 2018,39(1):7-13.

[36] 李龙, 马磊, 贺建峰, 等. 基于特征向量的最小二乘支持向量机 $PM_{2.5}$ 浓度

预测模型[J]. 计算机应用, 2014,34(8):2212-2216.

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

[38] 潘志松, 黎维. 基于深度学习的时空序列预测方法综述[J]. 数据采集与处理, 2021,36(3):436-448.

[39] 石峰, 楼文高, 张博. 基于灰狼群智能最优化的神经网络 $PM_{2.5}$ 浓度预测[J]. 计算机应用, 2017,37(10):2854-2860.

[40] 王伟科. 基于 LSTM 神经网络的空气质量多模态集成预测方法研究及其应用[D]. 兰州财经大学, 2020.

[41] 夏润，张晓龙.基于改进集成学习算法的在线空气质量预测[J] . 武汉科技大学学报. 2019,42(1):61-67.

[42] 于浩，王斌，肖刚，等．基于距离的不确定离群点检测[J].计算机研究与发展.2010,47(3);474-484.

[43] 于泉, 孙瑶. 基于粒子群优化小波神经网络的行程时间预测[J]. 交通运输研究, 2020,6(2):9.

[44] 张义, 王爱君. 空气污染健康损害、劳动力流动与经济增长[J]. 山西财经大学学报, 2020,42(03):17-30.

[45] 张袁元, 辛江慧, 周祥, 等. EMD 方法的改进研究及其在机械信号中的应用[J]. 机械设计与制造, 2016(2):98-102.

[46] 周杉杉, 李文静, 乔俊飞. 基于自组织递归模糊神经网络的 $PM_{2.5}$ 浓度预测[J]. 智能系统学报, 2018,13(4):509-516.

[47] 朱飞鸿.基于集成学习的空气质量预测模型分析研究[D].长安大学, 2018

[48] 夏茂森,江玲玲.变分模态分解模型中关键参数 K 的辨识研究——基于加权最大信息系数法[J].统计与信息论坛,2021,36(02):23-35.

附 录

CEEMDAN 分解主要代码：

```
E_imfNo = np.zeros(50, dtype=np.int)
max_imf = -1
N = len(X)
tMin, tMax = 0, 2 * np.pi
T = np.linspace(tMin, tMax, N)
ceemdan = CEEMDAN()
ceemdan.trials = 50
ceemdan.noise_seed(12345)
E_IMFs = ceemdan.ceemdan(X, T, max_imf)
imfNo = E_IMFs.shape[0]
c = np.floor(np.sqrt(imfNo + 1))
r = np.ceil((imfNo + 1) / c)
```

VMD 分解主要代码：

```
T = len(data)
fs = 1/T
t = np.linspace(1, T, num=T)/T
freqs = 2*np.pi*(t - 0.5 -1/T)/fs
f = data.iloc[:, 1]
f = [round(i,4) for i in f]
alpha = 2000
tau = 0
K = 12
DC = 0
init = 1
tol = 1e-7
u, u_hat, omega = vmd(f, alpha, tau, K, DC, init, tol)
```

拉普拉斯提取 MATLAB 主要代码：

```
clear;
fea = csvread('缺失值处理后站点数据.csv',1,0);
fea = fea';
options = [];
options.WeightMode = 'HeatKernel';
options.NeighborMode = 'KNN';
options.k = 5;
W = constructW(fea,options);
[nSmp, mFea]=size(fea);
DCol = full(sum(W,2));
D = spdiags(DCol,0,nSmp,nSmp);
D_mhalf = spdiags((DCol + ones(1,nSmp)') .^-.5,0,nSmp,nSmp) ;
Ag = D_mhalf * (W + eye(nSmp)) * D_mhalf;
LX = (Ag * fea)';
L2X = (Ag^2 * fea)';
```

ELSTM 预测模块主要代码：

```
dataset = pd.read_csv(filename,engine='python')
config = [12, 100, 50, 100, 1]
predictions = list()
    train, validation, test = train_test_split(dataset, n_validation,n_test)
    test = test.reset_index(drop=True)
    trian_validation = pd.concat([train,validation])
    history = [row for index,row in trian_validation.iterrows()]
    history = pd.DataFrame(history)
    model_list = individe_fit(train,cfg)
    for i in range(test.shape[0]):
        yhat = stacking_predict(train,history,cfg,model_list)
        predictions.append(yhat)
        history = history.append(test.iloc[i])
    error1 = measure_rmse(test['1'], predictions)
```

```
error2 = mean_absolute_error(test['1'], predictions)
error3 = mape(test['1'], predictions)
```

后 记

岁月不居，时节如流，光阴在指缝间溜走。在财大的这几年，虽没做出什么成绩，但过得很充实。在这 3 年的时间里，我感受过失败的挫折，也体会到成功的喜悦，在这段求学的历程中，我有很多需要感谢的人。

我能安心地继续在学校中充实自己，离不开家人的大力支持。本科毕业后的经历考研失败的我选择参加工作，但在工作中发现，想要实现自己的目标，必须更多的充实自己，于是选择冲击研究生入学考试，完成攻读研究生的目标。在父母深入交流后，父母也表示全力支持，一年的全身心备考，深刻的体会到父母对我无微不至的关照。在研究生阶段，父母也时刻提醒自己要坚定理想，做好学术研究，并且生活上也给予我更多的支持，让我的研究生生活即充实又美好。感谢我的父母，是你们的坚强后援，让我不必回头、一直向前。愿你们身体健康、平安喜乐。

在我的学习过程中，我要诚挚感谢我的导师黄恒君教授。对我们每一位学生都尽心尽力，尽职尽责。在科研上，无论是平时对于课程内容的消化理解，还是科研项目的难题处理，老师都倾力相助甚至亲自为我们排忧解难。在生活上，更是时常关心，尤其是在疫情期间，当我们需要在学校隔离时，老师则会定时举行会议，询问我们生活状态，为我们采买生活物资与医疗物资。在这里，衷心祝愿老师与师母身体健康、工作顺利、阖家幸福。

当然，我也要感谢这几年朝夕相处的同门，很怀念这三年互帮互助的日子，一起辛苦开展调研，一起努力参加比赛。正是有了你们，才会不断在学术上取得成果，在生活上充满欢乐。当然，我也要感谢我的舍友和同学们，一起早起学习，一起运动娱乐。衷心祝愿你们学业顺利，工作顺心。