分类号 C8/394　　密级 公开
U D C 0005618　　编号 10741

兰州财经大学
LANZHOU UNIVERSITY OF FINANCE AND ECONOMICS

硕 士 学 位 论 文

论文题目 基于多视角数据融合的 $PM_{2.5}$ 浓度预测研究——以兰州市为例

研 究 生 姓 名：廖若雯

指导教师姓名、职称：黄恒君 教授

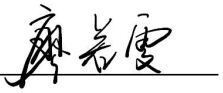
学 科、专 业 名 称：统计学 应用统计硕士

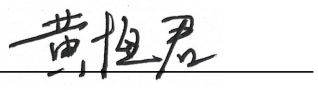
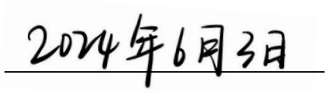
研 究 方 向：大数据分析及应用

提 交 日 期：2024 年 6 月 5 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：  签字日期： 

导师签名：  签字日期： 

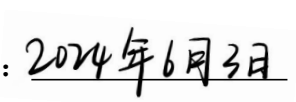
导师(校外)签名：_____ 签字日期：_____

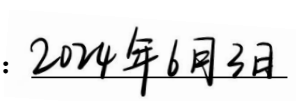
关于论文使用授权的说明

本人完全了解学校关于保留、使用学位论文的各项规定，  （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：  签字日期： 

导师签名：  签字日期： 

导师(校外)签名：_____ 签字日期：_____

Prediction of $PM_{2.5}$ concentration based on multi-view data fusion——A case study of Lanzhou City

Candidate: **Ruowen Liao**

Supervisor: **Hengjun Huang**

摘 要

在我国提出的三大攻坚战中，污染防治攻坚战是关乎人民群众身体健康和社会经济可持续发展的重大战役。空气质量问题是污染防治攻坚战的重点之一，对城市发展和居民健康造成了严重影响。近年来，由于工业排放、汽车尾气和人口密集等因素，我国许多地区的空气质量问题越来越严重。持续恶化的空气质量严重影响了城市的发展和居民的健康，空气质量问题成为当今时代的社会关注的热点问题，直径小于或等于 2.5μm 的细微颗粒物（$PM_{2.5}$）在空气质量监测中作为一个主要的污染源，是制约空气质量改善的主要因素。因此，本文将以 $PM_{2.5}$ 浓度值为预测对象，围绕着空气质量预测问题展开研究。根据目前对 $PM_{2.5}$ 浓度预测的研究现状和发展方向，对各种 $PM_{2.5}$ 浓度预测的方法进行了总结，从利用传统的统计方法到机器学习方法，以及随着深度学习的发展与应用，目前也有很多学者利用深度学习方法来预测 $PM_{2.5}$ 的浓度值等方面，分析了现有文献的优点与不足之处，发现现有的预测模型的准确性方面还有很大的提升空间，因此，本文的主要研究内容是基于空气质量数据、气象数据、兴趣点（POI）以及路网数据构建多视角图卷积门控递归单元（MGCN-GRU）模型框架和多视角图注意力长短期记忆（MGATs-LSTM）模型框架对 $PM_{2.5}$ 浓度值进行时空预测。

本文在已有的文献的基础上，挖掘了 $PM_{2.5}$ 浓度和时间特征、其他污染物特征、气象特征、POI 特征、路网结构特征之间的相关性。从时间和空间两个不同的维度，猜测具有时间序列的特征会对 $PM_{2.5}$ 浓度值的预测有时间依赖性，空间特征也会对 $PM_{2.5}$ 浓度产生空间相关性的影响，本文旨在挖掘这些特征对 $PM_{2.5}$ 浓度值预测的影响，这会有助于提高预测性能，并通过实验进行论证。

以甘肃省兰州市的 $PM_{2.5}$ 浓度数据为例，基于空气质量数据、气象数据、POI 和路网数据构建 MGCN-GRU 模型框架和 MGATs-LSTM 模型框架对 $PM_{2.5}$ 浓度值进行时空预测。其中对于 POI、路网这样的非时序数据，目前的特征提取方法存在忽视不同类别 POI 和路网之间的层次关系的问题。为解决这一问题，我们提出利用图结构学习非时序数据的特征表示，并将其作为辅助信息应用于

$PM_{2.5}$浓度预测中。最后利用平均绝对误差（*MAE*）、均方误差（*MSE*）、均方根误差（*RMSE*）、平均绝对百分比误差（*MAPE*）、决定系数（R^2）这 5 个指标对预测模型进行评价。结果证明在基于多视角数据融合的基础上加入了空间特征的时空预测模型比其他模型能够更加准确的进行$PM_{2.5}$浓度预测。再为了证明多视角数据融合对预测性能的重要影响，进而对城市$PM_{2.5}$浓度预测进行消融实验，证明数据融合在城市$PM_{2.5}$浓度预测中具有重要的优势，可以帮助我们更好地理解和预测空气质量，以支持决策和改善环境质量。

关键词：$PM_{2.5}$ 数据融合 图卷积网络 图注意力网络 预测

Abstract

Among the three major battles proposed by our country, the battle of pollution prevention and control is a major battle related to the health of the people and the sustainable development of society and economy. Air quality is one of the key issues in the battle against pollution, which has a serious impact on urban development and residents' health. In recent years, air quality problems in many parts of China have become more and more serious due to factors such as industrial emissions, vehicle exhaust and dense population. The deteriorating air quality has seriously affected the development of cities and the health of residents, and the issue of air quality has become a hot issue of social concern in today's era. Fine particulate matter ($PM_{2.5}$) with a diameter less than or equal to 2.5 μm is a major source of pollution in air quality monitoring, and a major factor restricting the improvement of air quality. Therefore, this paper will take $PM_{2.5}$ concentration value as the forecast object and carry out research on air quality prediction. According to the current research status and development direction of $PM_{2.5}$ concentration prediction, various $PM_{2.5}$ concentration prediction methods are summarized, from the use of traditional statistical methods to machine learning methods, as well as with the development and application of deep learning, there are many scholars using deep learning methods to predict $PM_{2.5}$ concentration. After analyzing the advantages and disadvantages of existing literature, it is found that there is still much room for improvement in the accuracy of existing prediction models. Therefore, Based on air quality data, meteorological data, point of interest (POI) and road network data, the main research content of this paper is to construct multi-view graph Convolutional gated recursive unit (MGCN-GRU) model framework and

multi-view graph Long Short-Term Attention memory (MGATs-LSTM) model framework to predict $PM_{2.5}$ concentration in time and space.

On the basis of existing literature, this paper explores the correlation between $PM_{2.5}$ concentration and time characteristics, other pollutant characteristics, meteorological characteristics, POI characteristics and road network structure characteristics. From the two different dimensions of time and space, it is speculated that features with time series will have a time dependence on the prediction of $PM_{2.5}$ concentration, and spatial features will also have a spatial correlation impact on $PM_{2.5}$ concentration. This paper aims to explore the impact of these features on the prediction of $PM_{2.5}$ concentration, which will help improve the prediction performance, and demonstrate through experiments.

Taking $PM_{2.5}$ concentration data in Lanzhou, Gansu Province as an example, MGCN-GRU model framework and MGATs-LSTM model framework were constructed based on air quality data, meteorological data, POI and road network data to predict $PM_{2.5}$ concentration in time and space. For non-time series data such as POI and road network, the current feature extraction methods ignore the hierarchical relationship between different types of POI and road network. To solve this problem, we propose to use the graph structure to learn the feature representation of non-time series data and apply it as auxiliary information to the prediction of $PM_{2.5}$ concentration. At last, the prediction model was evaluated using five indexes: mean absolute error (*MAE*), mean square error (*MSE*), root mean square error (*RMSE*), mean absolute percentage error (*MAPE*) and determination coefficient (R^2). The results show that the spatio-temporal prediction model with the addition of spatial features on the basis of multi-view data fusion can predict $PM_{2.5}$ concentration more accurately than other models. In order to prove the important impact

of multi-perspective data fusion on the prediction performance, ablation experiments were conducted for urban $PM_{2.5}$ concentration prediction to prove that data fusion has important advantages in urban $PM_{2.5}$ concentration prediction, which can help us better understand and predict air quality to support decision-making and improve environmental quality.

Keywords: $PM_{2.5}$; Data fusion; Graph convolutional network; Graph attention network; Prediction

目 录

1 绪论

随着工业化和城市化的加速发展，环境污染日益加剧，空气质量成为当今社会关注的焦点之一。空气中的颗粒物和有害气体对人类健康和环境造成严重影响，大气污染已成为全球性的问题，为此国家和个人都在为保护环境做出贡献。

1.1 研究背景

近年来，随着社会经济的飞速发展，人类在享受物质文明的同时，也面临着诸如大气污染、雾霾天气频发等环境问题[51]。由于烟雾天气引起的空气污染频繁发生，空气污染将对人类健康构成巨大威胁，严重影响了人们的健康和生活质量，进而影响生态环境以及经济发展[11]，城市大气污染已经成为人类面临的紧迫挑战，其中，直径小于或等于 2.5μm 的细微颗粒物（$PM_{2.5}$）对人体健康和环境质量的影响更大，它们对人体的伤害[39]都是无法挽回的。这些空气污染物造成了很严重的空气质量问题，限缓了社会的发展速度，给经济和环境带来了不良影响。为了改善人们的生活以及出行条件，需要定期监测和预测空气质量， 因此，准确预测 $PM_{2.5}$ 浓度的变化、实时提供未来空气质量信息，有利于人们及时采取防护措施，在一定程度上降低大气污染对人体的危害，提前规划和预防人们的行为，同时也可以在环境治理方面给政府提供环境治理建议。

在空气质量预测的研究中，交通尾气、工业排放、生活燃烧等源头排放是城市中空气污染的主要来源之一，且气象因素如风向、风速、温度等也对空气质量产生重要影响，因此，通过综合考虑多种数据和因素进行空气质量预测成为重要的研究方向。为了制定更有效的保护措施，需要对影响空气质量的各种因素做出全面分析并进行预测，以便提供更为可靠的环境保护依据，从而防止空气质量对人们出行和生活造成不良影响。随着人工智能、大数据和机器学习等技术的发展，空气质量预测研究也得到了更多的关注。通过运用先进的技术手段和分析方法，可以更准确地预测空气质量变化情况，并为环境管理和公众健康提供更好的支持。时空预测模型是一种能够结合时间和空间信息进行预测的模型，可以更准确地捕捉 $PM_{2.5}$ 浓度的时空变化规律。传统的预测模型往往忽视了时空维度的交互影响，而时空预测模型能够充分利用不同时间点和不同

空间位置的数据，从而提高预测的精度和准确性。在这个背景下，开展空气质量预测研究不仅有利于加深对环境问题的认识，还可以推动环保措施的制定和实施，为构建清洁、美丽的生态环境贡献力量。

1.2 研究意义

本研究旨在开展科学的$PM_{2.5}$污染预测，以便为实施有效的污染治理和控制提供科学依据和参考。由于空气质量受多种因素交织影响，包括地理环境、气象条件和污染物来源等，预测不够精准和准确，将妨碍污染治理和控制的实现，以及现有的监测方法和处理技术已经无法很好地满足$PM_{2.5}$污染治理的需要。因此，为提高$PM_{2.5}$浓度预测的准确性和稳定性，需利用多视角数据，探索开发新的方法来监测和处理$PM_{2.5}$污染，设计$PM_{2.5}$浓度预测的时空预测模型，来预测和监测$PM_{2.5}$浓度，以提高预测的准确性和稳定性。不仅如此，在$PM_{2.5}$污染治理和控制过程中，公众参与至关重要。通过预测研究，可以向公众提供有关$PM_{2.5}$浓度的信息，包括监测数据、预测结果和反馈意见等，从而促进信息共享和公开透明，让公众更好地了解$PM_{2.5}$污染的问题和治理进展，参与治理进程，促进可持续发展。

在理论层面上，通过对现有空气质量预测研究中存在的不足进行分析，基于文献资料和研究成果的交叉应用，提出了一种新颖的时空预测模型组合。通过整合距离、兴趣点和路网数据，以构建其他站点对目标站点的空间影响，从而提高空气质量预测的精准度。从时间和空间两个角度的完整的解释了预测问题，为构建时空预测模型提供了新的思路。

$PM_{2.5}$浓度预测研究对环境治理方面具有非常重要的意义。$PM_{2.5}$是一种直径小于2.5微米的可吸入颗粒物，进入人体后易长时间滞留，对人体健康有很大的危害。预测研究可以为各行业提供准确的监测结果，为治理提供有力、具体的方案，改善、维护环境空气质量，促进人民健康。可应用深度学习等新技术进行数据处理、模型预测，不仅能够应用到环境监测，在处理环境事务等方面也能起到作用。更好地发挥先进技术在环境治理中的作用，提升环境治理科技水平。

$PM_{2.5}$预测研究也是推动智慧城市建设中非常重要的一环。通过数据分析和模型预测，可以更精确地掌握城市$PM_{2.5}$浓度的变化，并及时采取调整和优化

措施，从而实现智慧化城市垃圾管控、交通监管，更好地调解人口迁移、城市规划等方面的矛盾。$PM_{2.5}$预测研究可以帮助我们了解$PM_{2.5}$的来源和转移规律，掌握环境背景，有助于预测和规划环保产业以及相关产业的发展布局，促进环保产业的发展，推动可持续发展。

总之，$PM_{2.5}$浓度预测研究的目的和意义是巨大的，不仅关系到人民的身体健康和生活质量，也涉及到环境治理、智慧城市建设和产业发展等领域，因此，在深度学习和数据融合思路的基础上，进一步加入反映城市特征的多视角信息（如监测站点的经纬度、兴趣点（POI）特征、路网结构等），针对单一城市空气质量预测，提出了多视角数据融合方法，建立多视角图卷积（MGCN-GRU）模型框架和多视角图注意力（MGATs-LSTM）模型框架，进行$PM_{2.5}$浓度时空预测实例研究具有非常重要的实用价值和理论指导意义。

1.3 国内外研究现状

随着人们对空气质量预测问题研究的不断深入，已出现多种预测的方法和模型。可以分为经典统计学习方法，机器学习方法和深度学习方法。

（1）传统预测方法在空气质量预测研究中的应用

许多基本的统计模型都可以应用到空气质量预测问题研究中，包括回归和分类模型，它们都是对基于气象数据与空气质量数据之间的相关性这一假设来进行预测的，其中有线性回归，时间序列等分析方法被更多的学者所采纳来运用于$PM_{2.5}$浓度预测当中。例如，Sun 等[21]（2013）将关键的气象因素纳入到模型中，并在先验假设中考虑到了$PM_{2.5}$浓度的非高斯分布，提出了一种服从对数正态分布的隐马尔科夫模型，利用这种方法来预测有效地降低了$PM_{2.5}$浓度超标的预警次数。Gupta 等[19]（2009）利用美国东南部 85 个$PM_{2.5}$监测站点 3 年同期监测的$PM_{2.5}$质量浓度，将监测点作为季节函数变量建立多元回归方程预测$PM_{2.5}$浓度值，该方法对气象信息的相关系数提高了 3 倍以上。徐东等[50]（2021）建立的多元线性回归模型，研究结果证明温度对$PM_{2.5}$的影响不显著，可吸入颗粒物（PM_{10}）和一氧化碳（CO）对$PM_{2.5}$有重要影响。

虽然传统的时间序列预测模型较为擅长提取数据中的线性趋势，但无法捕捉数据中的非线性特征。尤其是在处理涉及到高频复杂时间序列的情况下，如空气质量数据时，传统时间序列预测模型的预测效果非常有限。

（2）机器学习方法在空气质量预测研究中的应用

为了提高模型从数据中提取非线性特征的能力，解决线性统计模型存在的问题，机器学习技术被运用到有关预测问题的研究中，发现这种预测方法更适合描述关系复杂的问题。Dong 等[4]（2009）提出将时间结构添加到 HMM 中预测 $PM_{2.5}$ 浓度值，结果表明该模型可以提高对于高浓度 $PM_{2.5}$ 未来 24 小时预测的准确性。戴孝杰等[37]（2017）提出了一个未来 24 小时 $PM_{2.5}$ 的粒子群优化的支持向量机分析算法预测，其中在 6-12 小时是预测结果最好。李龙等[41]（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[26]（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 $PM_{2.5}$ 浓度数据输入模型，与其他模型的预测能力进行比较，发现该模型存在预测优势。Li 等[12]（2018）分析了上海市 18 个环境空气监测站的 8 种空气污染物的空间变化特征，采用 K 最近邻法将 18 个站点划分为 3 个级别组，结果表明，不同污染物的空间分布差异较大，空气质量与气态污染物浓度密切相关，预测气态污染物浓度对调节工厂和汽车排放起着决定性作用。Liu 等[13]（2017）着眼于美国周边地区 $PM_{2.5}$ 的长期分布，采用了基于随机森林的地质统计（回归克里格法）方法，以改进最常用的卫星衍生网格化 $PM_{2.5}$ 数据集，提高空间分辨率和精度，通过与现有 $PM_{2.5}$ 数据集的比较，验证了该方法的准确性和优势。

机器学习运用的本质是对序列进行有监督的学习，因此，时间窗特征对机器学习方法的效果至关重要。普通机器学习方法在处理时间序列问题时表现欠佳，尤其是在预测空气污染方面。机器学习方法在预测空气污染方面已经取得了不错的效果，但机器学习方法很难解决预测过程中的高维非线性长时间序列问题[18]。部分学者在对深度学习进行运用中发现，利用长短期记忆神经网络进行空气质量预测，可以有效克服机器学习的局限性，更好地处理时间维上的非线性空气质量数据。

（3）深度学习算法在空气质量预测研究中的应用

近年来，在采用经典统计模型[14]和机器学习方法[22]开展预测的基础上，人们倾向于采用深度学习方法，用以提高城市空气质量预测的准确性。提出利用

深度学习能力捕捉非线性关系，以此来解决非线性时间序列的问题。深度学习是机器学习的一个分支。深度学习的实质是通过构建具有多个隐层学习模型以及海量训练数据来学习数据中相关的有用特征，从而提高预测或分类的准确性[27]。例如，石峰等[47]（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[54]（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。Zhao 等[29]（2019）比较了人工神经网络（ANN）、长短期记忆网络（LSTM）和 LSTM-Fully Connected（LSTM-FC）模型在预测 $PM_{2.5}$ 浓度上的表现，他们发现 LSTM-FC 产生了更好的预测性能。蒋洪迅等[43]（2021）构建了双向长短期记忆网络的神经网络预测模型（DLENN），基于沈阳地区 11 个监测站 2016 至 2017 年空气质量和气象条件数据，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存在优势。于书玉[53]（2023）针对 $PM_{2.5}$ 预测中数据来源单一的问题，提出了长短期记忆网络（LSTM）融合神经网络预测模型，以北京市的空气质量监测站点 2010 至 2014 年的污染数据和天气数据进行实验，并将 LSTM 模型与其他预测模型进行对比，研究表明提出的 LSTM 融合模型具有更优的预测能力。以上研究表明，深度学习模型在数据特征提取和预测精度方面，普遍优于传统机器学习方法和经典统计模型。

（4）时空预测模型在空气质量预测研究中的应用

在之前的文献中可以看到，主要利用的数据集基本上仅仅只有历史气象数据与空气质量数据，这些数据源和模型结构相对单一。而空气质量预测中，所研究的角度不能只关注时间序列，还需要考虑空间属性的度量，将空间因素纳入模型。因此，从时间和空间两个角度进行分析，考虑 $PM_{2.5}$ 在空气中的传播情况，以及其他地区空气质量对目标区域的影响。在这样的背景下，大量学者提出了时空预测这一概念来更准确的解决空气质量的预测问题。

时空序列数据是具有空间相关性的时间序列的集合，是基于一般时间序列

延伸而来的，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，却没有考虑到可能存在的空间关联。例如，Zhao 等[29]（2019）提出了一种采用基于空间组合的全连接神经网络捕捉目标区域与相邻五个站点的相关性。Ge 等[7]（2019）首先利用张量分解法填补了缺失的历史空气质量数据，然后利用非时间序列数据（例如 POI 和道路网络）来评估不同区域之间的相似性，从而捕捉时间和空间的依赖性，改善空气质量的预测效果。Li 等[15]（2020）利用卷积神经网络和长短期记忆网络串联组合来对 $PM_{2.5}$ 浓度进行预测，CNN-LSTM 可以提高 $PM_{2.5}$ 预测的准确性。Tao 等[24]（2019）将一维的卷积神经网络与双向门控循环单元的组合模型来预测 $PM_{2.5}$ 浓度，并评估了模型的性能，发现组合模型预测的 $PM_{2.5}$ 浓度的效果更好。Chen 等[2]（2021）提出了一种基于卷积神经网络和循环神经网络（CNN-RNN）的 $PM_{2.5}$ 值预测框架，该框架由人工智能云计算驱动，对多模式数据进行解读最终发现该预测框架效果最优。马俊文等[44]（2022）联合长短期记忆网络（LSTM）提取的时间特征和图卷积神经网络（GCN）提取的空间特征，提出预测 $PM_{2.5}$ 浓度的 LSTM-GCN 组合模型。以北京市 35 个空气质量监测站 2018—2020 年监测数据进行仿真实验，并将 LSTM-GCN 模型与 LSTM 模型、GCN 模型以及时空地理加权回归模型（GTWR）进行对比，结果表明所提出 LSTM-GCN 模型在准确率上有所提升。叶如珊等[52]（2022）提出一种基于卷积神经网络（CNN）与双向长短期记忆网络（BiLSTM）的混合预测模型.该模型不仅考虑数据双向的时序特征,还关注不同特征之间的空间关联性，模型通过提取并融合数据的时空特征来实现 $PM_{2.5}$ 浓度预测。在北京市 2010 年-2014 年的天气和污染水平数据集上将该模型与 LSTM、BiLSTM、CNN-LSTM 模型进行 $PM_{2.5}$ 浓度预测实验对比，结果表明 CNN-BiLSTM 模型的预测误差明显小于其他模型,该预测模型具有更好的预测性能。曹旺等[35]（2022）采用门控循环单元（GRU）和图神经网络相结合的混合模型预测的方式，并进一步采用改进的门控循环单元提升网络效果，实验结果表明，提出的改进方法相比于现有的方法具有更好的性能以及泛化效果，在中国生态环境部提出的京津冀地区真实数据集上验证了方法的有效性，与现有网络相比预测准确率更高。傅颖颖等[38]（2021）提出了融合图卷积神经网络（GCN）和注意力机制（AttentionSeq2Seq）

的$PM_{2.5}$小时浓度多步预测模型。以2015—2016年北京市22个空气质量监测站点的空气质量数据为样本进行实例验证，与Seq2Seq模型和使用了图卷积神经网络、未使用注意力机制的GCNSeq2Seq模型进行了对照，结果表明组合的多步预测模型（GCNAttentionSeq2Seq）可有效应用于多种长度的$PM_{2.5}$浓度预测窗口。宋飞扬等[48]（2020）提出基于时空特征的KNN-LSTM的$PM_{2.5}$浓度预测模型，以哈尔滨市10个空气质量监测站的污染物数据进行仿真实验，并将KNN-LSTM模型与其他预测模型进行对比，结果表明所提KNN-LSTM模型能有效提高LSTM模型的预测精度。

（5）多视角数据融合在空气质量预测研究中的应用

提高城市空气质量预测准确性的另一个做法是多视角数据融合。其中，一类是时间序列数据的融合应用。例如，Huang等[8]（2018）使用历史空气质量污染物数据和气象数据，从双视角数据出发利用卷积神经网络和长短期记忆网络对$PM_{2.5}$浓度进行预测，结果表明利用融合后的数据进行预测效果更佳。Zheng等[30]（2015）提出基于多视图的混合模型，该模型考虑了来自周围几个监测站点的气象和空气质量数据，并使用线性回归和神经网络开展预测，预测未来48小时的空气质量监测站数据的效果较好。白盛楠等[34]（2019）通过对气象、空气质量污染物特征进行灰色关联度分析，得到与$PM_{2.5}$之间的关联强度，采用多视角长短期记忆模型预测$PM_{2.5}$浓度的日值变化趋势，发现多视角下能够较好地预测$PM_{2.5}$的日值变化趋势。另一类是开展时空数据的融合研究。例如，Qi等[20]（2019）利用图卷积网络提取不同站点之间的空间相关性，并使用长短期记忆网络捕获时间相关性，利用时空数据进行预测，其模型预测效果较好。Zhang等[28]（2018）又提出了Deep-Air模型，通过监测站历史浓度数据、兴趣点、路网、风向等一系列可对污染物造成影响的特征结合起来，对某个城市的监测站进行48小时的污染物趋势分析，通过预测得知此模型效果优秀，并已经投入使用，为政府提供污染物预测值。陈逸彬等[36]（2022）提出一个基于混合CNN-LSTM结构的$PM_{2.5}$预测模型,将CNN结构与LSTM模型结合，对单元和多元数据集展开实验，通过添加CNN结构可以提取更多有效信息并提升预测的精确性。张旭[55]（2022）提出了基于图注意力和长短期记忆网络的时空预测模型（GAT-LSTM），融合了三个视角下空间信息特征输入到长短期记忆力网

络中提取空气质量信息的时间相关性，最终提高了模型的预测精度。孙小新[49]（2022）提出了图注意力和长短期记忆网络的时空预测模型（GAT-LSTM）和图注意力和门控递归单元构成的时空预测模型（GAT-GRU）利用空气质量污染物数据和气象数据双视角下对$PM_{2.5}$浓度进行预测，在融合后的双视角下能够较好的进行预测$PM_{2.5}$的变化趋势。以上研究均表明，多视角数据融合方法在时序数据以及时空数据的预测精度方面，普遍优于其他非融合方法。

通过对国内外研究现状的讨论，空气质量预测方法在当前面临一系列挑战。其中，现有方法在考虑因素时存在着一定的不足之处，主要表现在对于时间序列（污染物数据、气象数据）和非时间序列数据（地理距离、POI、路网）处理的不全面性以及在区域间相互影响方面的缺失。但是影响城市空气质量的因素却是多样的，具体而言，这些方法往往忽略了气象因素的影响，未对非时间序列数据进行细致的特征提取，且对不同区域间相互影响的程度差异缺乏考虑。这些局限性导致了预测模型存在一定的局限性和偏差。但是实际上不同区域对于待预测目标区域空气质量的影响程度是不同的，所以现有的一些方法不能很好地捕获空气质量在空间上的相关性。

综上所述，现在存在的模型仍然存在以下不足：

（1）在对空气质量进行预测是，未充分考虑空间相关性，导致存在预测偏差的问题。因为污染物扩散会受附近区域空气流动的影响，忽略空间因素会使预测结果产生偏差。仅通过分析历史数据变化趋势，没有考虑空间因素，无法有效提高模型的预测能力。

（2）现有预测模型虽然具备提取时空关联特征的能力，但大多适用于欧式空间数据，而真实世界的时空数据常以非欧式形式存在。将非欧式数据强行转换为图片数据以提取时空特征可能导致欧式空间模型处理数据时产生偏差，尤其在涉及POI、路网等数据时，这种转换可能限制预测模型的精度。导致预测结果产生偏差的问题。

（3）随着数据特征提取技术的不断进步，数据变得更加复杂，高维度化可能导致数据处理和计算效率下降，增加了预测的困难度。这些不足之处使得空气质量预测方法的改进和优化面临着更为严峻的挑战。

1.4 研究内容和创新点

1.4.1 研究内容

空气质量问题是当今时代的社会关注的热点问题，由于$PM_{2.5}$在空气质量监测中作为一个主要的污染源，是制约空气质量改善的主要因素。因此，本文将以$PM_{2.5}$浓度值为预测对象，围绕着空气质量预测问题展开研究。根据目前对$PM_{2.5}$浓度预测的研究现状和发展方向，对各种$PM_{2.5}$浓度预测的方法进行了总结，从利用传统的统计方法到机器学习方法，以及随着深度学习的发展与应用，目前也有很多学者利用深度学习方法来预测$PM_{2.5}$的浓度值，分析了现有文献的优点与不足之处。因此，在现有的预测模型的准确性方面还有很大的提升空间。本文的主要研究内容是基于空气质量数据、气象数据、POI 数据、路网数据对$PM_{2.5}$浓度值构建 MGCN-GRU 模型和 MGATs-LSTM 模型进行时空预测。本研究的具体内容及主要目标如下：

（1）数据处理部分

由于获取的数据存在缺失、以及收集产生的数据误差的情况，需要对不同数据源的数据进行预处理工作，目标是形成可融合的数据类型，以支持时空预测模型的构建，包括：

本研究在对$PM_{2.5}$浓度做出有关分析之前，采用合适的方法对数据集进行了缺失值填补、对数据进行归一化以及对空间特征进行了范围筛选，将数据融合为整体数据集。目的是综合利用空气质量污染物数据、气象数据、空气质量监测站点的经纬度信息、兴趣点（POI）数据和路网等多种数据源，融合不同数据视角的信息，以提高预测结果的准确性和可靠性。

（2）$PM_{2.5}$浓度多维度相关性分析

本研究在已有的文献的基础上，对$PM_{2.5}$浓度预测进行了研究。挖掘了$PM_{2.5}$浓度和时间特征、其他污染物特征、气象特征、POI 特征、路网结构特征之间的相关性。从时间和空间两个不同的维度，猜测具有时间序列的特征会对$PM_{2.5}$浓度值的预测有时间依赖性，空间特征也会对$PM_{2.5}$浓度产生空间相关性的影响。目的旨在挖掘这些特征对$PM_{2.5}$浓度值预测的影响，这会有助于提高预测性能，并通过实验进行论证。

（3）基于多视角图卷积网络（MGCN-GRU）的$PM_{2.5}$浓度预测研究

以甘肃省兰州市的数据为例，分别利用 AMRM、LSTM、GRU、GCN-LSTM、MGCN-GRU 模型进行 $PM_{2.5}$ 浓度值预测，最后利用平均绝对误差（*MAE*）、均方误差（*MSE*）、均方根误差（*RMSE*）、平均绝对百分比误差（*MAPE*）、决定系数（R^2）这 5 个指标对预测模型进行评价。结果证明 MGCN-GRU 模型比其他预测模型能够更加准确的进行 $PM_{2.5}$ 浓度预测。为了证明模型的实用性，本文对未来 3、6、9、12、15、18 小时的 $PM_{2.5}$ 浓度值都进行了预测，最终实验结果都比其他模型的效果要好。再为了证明多视角数据融合对预测性能的重要影响，进而对城市 $PM_{2.5}$ 浓度预测进行消融实验，证明数据融合在城市 $PM_{2.5}$ 浓度预测中具有重要的优势。

（4）基于多视角图注意力网络（MGATs-LSTM）的 $PM_{2.5}$ 浓度预测研究

以甘肃省兰州市的数据为例，分别利用 MGCN-GRU、GAT-LSTM、GAT-GRU、MGATs-GRU、MGATs-LSTM 模型进行 $PM_{2.5}$ 浓度值预测，最后利用平均绝对误差（*MAE*）、均方误差（*MSE*）、均方根误差（*RMSE*）、平均绝对百分比误差（*MAPE*）、决定系数（R^2）这 5 个指标对预测模型进行评价。结果证明 MGATs-LSTM 模型比其他时空预测模型能够更加准确的进行 $PM_{2.5}$ 浓度预测。为了证明模型的实用性，本文对未来 3、6、9、12、15、18 小时的 $PM_{2.5}$ 浓度值都进行了预测，最终实验结果都比其他模型的效果要好。再为了证明多视角数据融合对预测性能的重要影响，进而对城市 $PM_{2.5}$ 浓度预测进行消融实验，证明数据融合在城市 $PM_{2.5}$ 浓度预测中具有重要的优势。

总之，基于多视角数据融合的 $PM_{2.5}$ 浓度预测研究的主要目标是在数据融合、预测模型、结果展示和模型效果验证等方面进行优化和提升，最终实现更加准确、稳定和可靠的 $PM_{2.5}$ 浓度预测结果，为环境保护工作提供科学依据。

1.4.2 创新点

本研究的创新点如下：

（1）在数据层面，综合利用多视角数据信息。本研究不仅考虑了空气质量污染物数据、气象数据、空气质量监测站点的经纬度信息、兴趣点（POI）数据和路网等多种数据源，将这些数据信息综合起来，提高预测精度和可信度。

（2）在实现层面，采用深度学习的方法，强调时空相关性。本研究引入了深度学习技术，构建了多视角数据融合的时空预测模型框架，将时间和空间因

素考虑进去，以充分利用与$PM_{2.5}$浓度相关的地理和气象信息。模型框架的构建从时间和空间角度共同出发，本研究的 MGCN-GRU 模型框架是基于数据融合方法利用 GRU 模型融入了 GCN 模型同时处理历史$PM_{2.5}$浓度的时间信息与空间信息，在捕获时序数据的时空信息的同时，MGCN-GRU 模型分别基于空气质量监测站点的地理距离、POI 特征以及路网结构的信息，来构建站点之间是否存在连边的图结构，并通过归一化处理将三个图结构的信息进行融合，实现了基于多图图卷积的时空数据融合。本研究的 MGATs-LSTM 模型框架是利用 LSTM 模型融入了改进的双层 GAT 模型同时处理历史$PM_{2.5}$浓度的时间信息与空间信息，在捕获时序数据的时空信息的同时，MGATs-LSTM 模型框架分别基于空气质量监测站点的地理距离、POI 特征以及路网结构的信息，赋予不同的注意力权重来构建站点之间的图结构，并通过归一化处理将三个图结构的信息进行融合，实现了基于多图多层图注意力机制的时空数据融合。这样的方法可提高预测精度，并且有助于提升科学防空气污染能力。

1.5 论文结构安排

1.5.1 论文结构

论文一共包含六个章节，其中每个章节具体内容如下所示：

第一章：绪论，本章在总体上先介绍了本文研究的$PM_{2.5}$浓度对空气质量的影响的现实背景，以及对其做出合理的预测能够产生的理论意义和现实意义。其次是介绍了国内外相应学者以及研究机构当前对$PM_{2.5}$浓度预测从不同方法上研究的现状以及较为优秀的研究成果，最后是对论文的研究内容做了整体介绍以及论文的组织架构。

第二章：相关技术介绍，本章介绍了本文在处理数据方面、相关性分析方面以及模型预测方面需要用到的一些理论方法，主要介绍了$PM_{2.5}$浓度的时空特征分析的方法，主要包括 KNN 算法、归一化方法以及 Spearman 相关系数法，之后对图的构建利用了地理距离计算方法、余弦距离相似性度量方法做出介绍，最后对图卷积网络（GCN）、图注意力网络（GAT）、门控递归单元（GRU）、长短期记忆网络（LSTM）的模型与算法流程公式进行了概述。

第三章：$PM_{2.5}$各维度相关分析，本章对$PM_{2.5}$浓度影响因素做出了多维度的相关分析。分析之前对其问题进行了一个简单描述，对数据的来源进行介绍，

并对数据做出了有关的处理，之后就是分别对时间特征、其他污染物特征、气象特征、POI 特征、路网特征做出了有关分析。

第四章：基于多视角图卷积网络的 $PM_{2.5}$ 浓度预测研究，以甘肃省兰州市为例，本章在研究前对研究问题进行简单描述，对本文提出的 MGCN-GRU 模型框架做出了总体概述，对实验设置做出了模型的训练集和测试集的合理划分，以及对所使用的模型性能、实验环境做出了简单介绍，对所得到的实验结果进行了相应分析，为了验证本文模型的有效性，分别运用 ARMA、LSTM、GRU、GCN-LSTM、MGCN-GRU 模型对 $PM_{2.5}$ 浓度值在 3、6、9、12、15、18 小时进行预测。再为了证明多视角数据融合对预测性能的重要影响，进而对城市 $PM_{2.5}$ 浓度预测进行消融实验，证明数据融合在城市 $PM_{2.5}$ 浓度预测中具有重要的优势，可以帮助我们更好地理解和预测空气质量，以支持决策和改善环境质量。

第五章：基于多视角图注意力网络的 $PM_{2.5}$ 浓度预测研究，以甘肃省兰州市为例，研究之前对研究问题进行了简单描述，对本文提出的 MGATs-LSTM 模型框架做出了总体概述，以及模型中对时序特征的提取和非时序特征的提取进行详细说明，并且对实验设置做出了模型的训练集以及测试集的划分、所使用的模型性能评价指标以及实验环境做出了介绍，对所得到实验结果进行了相应的分析。为了验证本文模型的有效性，分别运用上一章节的 MGCN-GRU 模型和 GAT-LSTM、GAT-GRU、MGATs-GRU、MGATs-LSTM 模型对 $PM_{2.5}$ 浓度值在 3、6、9、12、15、18 小时进行预测。再为了证明多视角数据融合对预测性能的重要影响，进而对城市 $PM_{2.5}$ 浓度预测进行消融实验，证明数据融合在城市 $PM_{2.5}$ 浓度预测中具有重要的优势，可以帮助我们更好地理解和预测空气质量，以支持决策和改善环境质量。

第六章：总结与展望。对本文做出一个工作内容总结，以及可以开展后续研究内容的展望，并给出相应的结论，进一步阐述了本文对 $PM_{2.5}$ 浓度预测的研究在时空预测研究中的意义。其次，对本文提出的模型框架所存在的一些不足之处以及为进一步提高模型的训练及预测效果提供研究方向，并对未来可以改进的地方提出了展望。

1.5.2 技术路线

本研究从空气质量污染问题出发，以多视角数据为基础，以数据融合为处

理手段，构建时空预测模型，进而能够为预测区域直接提供一套 $PM_{2.5}$ 浓度预测的研究方法，达到为更有效地控制和治理 $PM_{2.5}$ 污染提供有力的支持和参考。研究框架如下图 1.1 所示。

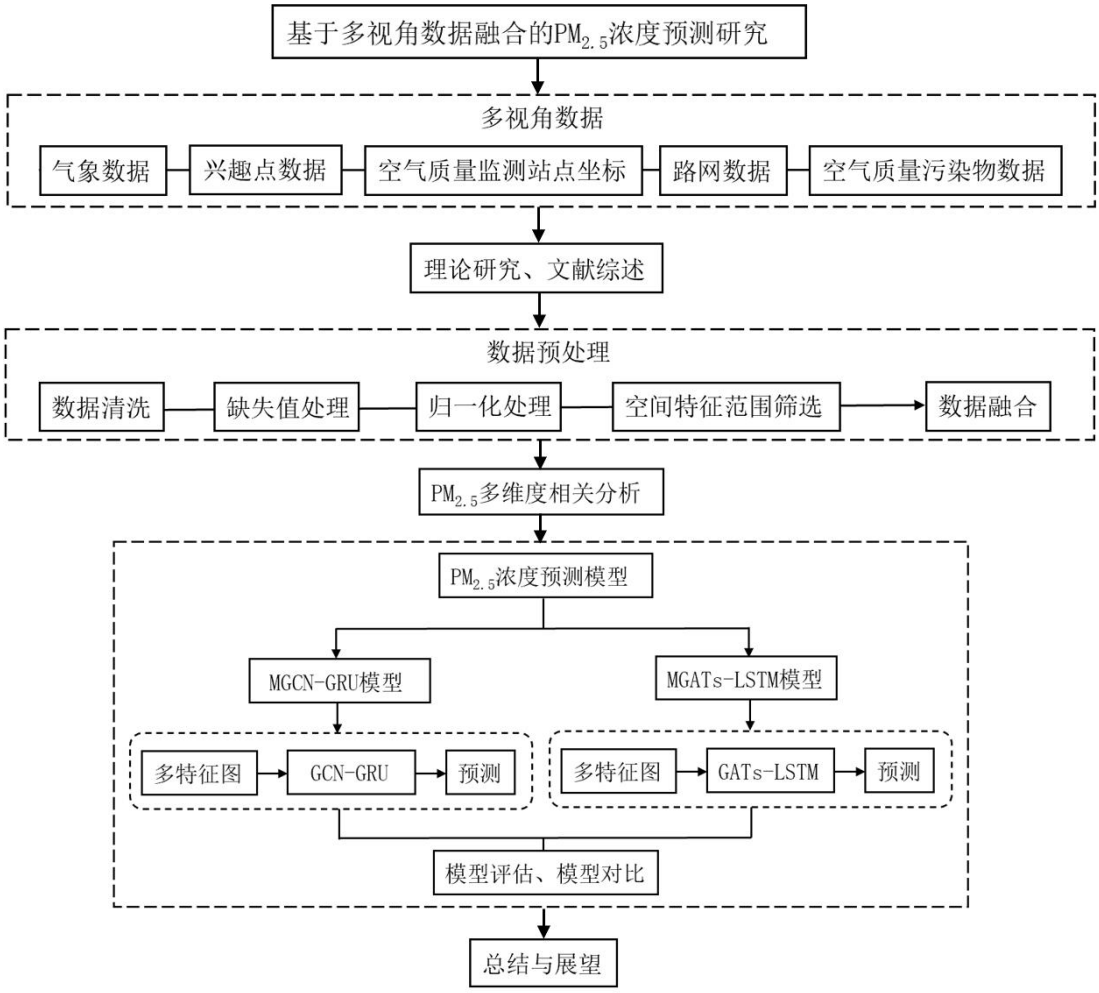


图 1.1 研究技术路线

2 相关技术介绍

2.1 数据预处理

2.1.1 K 近邻算法

K 近邻（K-Nearest Neighbor，KNN）算法填补缺失值的基本思想就是通过对距离的测量来识别数据集中 K 个样本的空间相似性，利用这 K 个样本来估计当前缺失位置的值，对于每个缺失值都使用其数据集附近相似的 K 个邻域的平均值进行插补。这样就能够将数据中缺失的数值进行完全的填补。对于存在缺失值时的距离计算，KNN 算法是基于欧氏距离的最短距离点被认为是最近邻点。在存在缺失坐标的情况下，KNN 算法是通过忽略缺失值并放大非缺失坐标的权重来计算欧几里德距离。其计算公式如（2.1）所示。

$$D(x,y)=\sqrt{\sum_{i=1}^{n}|x_i-y_i|^2} \tag{2.1}$$

其中，$\boldsymbol{D}$表示两个向量$\boldsymbol{x}$和$\boldsymbol{y}$之间的欧氏距离，n是向量$\boldsymbol{x}$和$\boldsymbol{y}$的维度，$\boldsymbol{x}_i$表示向量$\boldsymbol{x}$的第i个元素，$\boldsymbol{y}_i$表示向量$\boldsymbol{y}$的第i个元素。

2.1.2 归一化

空气质量数据的各项指标值都具有高波动性以及不平稳性，其对预测结果的精准度有着较为重要的影响。归一化旨在处理数据中具有的波动性、不平稳性等特征，确保所有预测因素都在类似的范围内，这是模型预测中的一个重要步骤，因为数值较大的输入会不成比例地掩盖数值较小的输入的影响。本文选取的归一化公式如（2.2）所示。

$$\tilde{x}_i^n=\frac{x_i-x_{min}}{x_{max}-x_{min}} \tag{2.2}$$

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

2.1.3 Spearman 相关系数

相关性分析是探究变量之间相关程度的一种方法。是通过度量 $PM_{2.5}$ 浓度序列与其他污染物和气象因子的相似度，来排除相关性较弱的特征序列以提高模型预测精度。在特征选择中，计算相关系数可保留相关性较高的特征、去除不相关特征，简化模型并降低干扰。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。

2.2 图结构

图被认为是包含丰富潜在价值的复杂结构[31]，是一种十分常用的数据结构，它一般是所有顶点的集合所组成。顶点之间的连边对应着不同研究对象之间的相互关系。图[9]通常可以表示为$\boldsymbol{G(V,E)}$。其中，$\boldsymbol{V=\{v_1,v_2,...,v_N\}}$表示的是图$\boldsymbol{G}$中所涉及所有节点的集合，$\boldsymbol{v_i}$表示的是所涉及所有节点集合中的第$i$个节点，$N$表示的是图中的节点数量。图$\boldsymbol{G}$中的节点之间连边集合可以表示为$\boldsymbol{E=\{e_1,e_2,...,e_m\}}$，其中，$m$表示的是图中节点之间的边数。

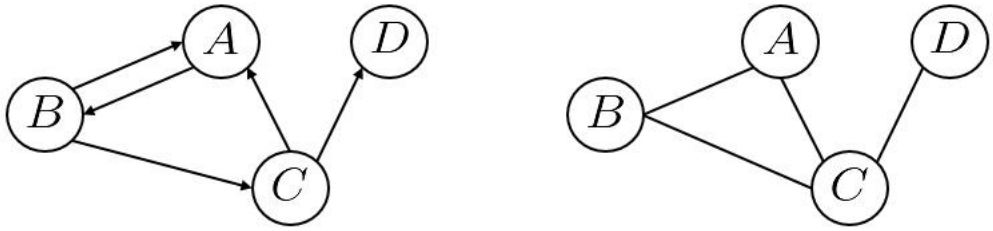


图 2.1 有向图（左）和无向图（右）

针对图的构建，如图 2.1 所示，图可以分为有向图（左）和无向图（右）。本文研究所涉及的图均为无向图，故有$\boldsymbol{e_{ij}=e_{ji}}$。以下研究的图默认均指无向图。关系图的构建主要考虑不同节点之间的邻近关系，邻近关系可以靠相似度来进行测度。

相似度是指待测点与样本点之间在某一地理环境配置上的近似程度[32]。相似度计算是 $PM_{2.5}$ 浓度预测模型中非常重要的一个过程。本节将介绍两种相似度的度量方法，分别是地理空间距离相似性度量和余弦距离相似性度量方法。

2.2.1 地理空间距离相似性度量方法

对于邻近关系，根据地理学第一定律，表明变量在相同的区域内同一类型的数据之间存在一定程度的联系，其相关性与距离有一定的关系，距离越近空间相关性越强，距离越远，相关性相对就较弱[1][6]。因此，本文研究了事物间的

相邻的空间关系，即通过相邻关系来确定节点之间是否有连边。在地理分析中，地理环境的相似性概念被广泛运用，依靠一组地理变量来评估其他区域内地理变量的相似程度。

本文中的模型利用各空气质量监测站点之间的相关数据提取空间特征，例如计算空气质量监测站点之间的距离等特征的相关信息，以确定图注意力网络模型中的邻接矩阵。因此，本小节旨在介绍地理空间距离相似性度量的计算方法。在地理学研究中，球面模型距离计算方法是计算地理空间距离方法之一，其采用标准球体模型，将地球形状简化为球体，用以计算球面上两点之间的最短距离，简单且具有一定的精度。本文所应用的地理空间距离计算方法为球面模型计算[42]。

本文应用 Haversine 公式进行地球上两个监测点的距离计算[46]，Haversine 公式在球面余弦公式基础上进行了一些变换，公式中对球面余弦公式的$cos(x_j - x_i)$这一项公式进行了替换，由于监测站之间距离相对于地球的半径是很小的，所以 Haversine 公式在距离很近的情况下，也可以保证数字距离的精准度。Haversine 公式中的半正矢函数如公式（2.4）所示。

$$haver(\phi) = \frac{1-cos(\phi)}{2} \tag{2.4}$$

利用 Haversine 公式和两个空气质量监测站点的经纬度信息还有地球的半径r，可以计算出两个站点之间的地理平面距离，计算过程如公式（2.5）所示。

$$haver(\frac{d_{ij}}{r}) = haver(x_j - x_i) + cos(x_j)cos(x_j)haver(y_j - y_i) \tag{2.5}$$

其中，d_{ij}为所求的监测站i和监测站j之间的距离，x_i，x_j为两个监测站的维度值，r是地球的半径。

2.2.2 余弦距离相似性度量方法

在向量空间中，余弦距离是利用两个向量之间的夹角余弦来度量它们的不相似性。当余弦值接近 1 时，表示夹角接近 0 度，即两个向量相似度高，称为“余弦相似性”。相较于欧氏距离，余弦距离更注重向量的方向差异。具体计算公式如（2.6）所示。

$$cos(\boldsymbol{\theta}) = \frac{\boldsymbol{a}\cdot\boldsymbol{b}}{||\boldsymbol{a}||\times||\boldsymbol{b}||} = \frac{\sum_{i=1}^{n}(X_i \times Y_i)}{\sqrt{\sum_{i=1}^{n}(X_i)^2}\times\sqrt{\sum_{i=1}^{n}(Y_i)^2}} \tag{2.6}$$

该公式就是向量$\boldsymbol{a}$和向量$\boldsymbol{b}$的余弦计算公式。其中，$\boldsymbol{a}$和$\boldsymbol{b}$是两个n维的向量。

X和Y为不同监测站之间的特征数量。

2.3 图神经网络

2.3.1 图卷积网络（GCN）

本文采用了可以由 CNN 推广到图卷积神经网络（GCN），GCN 是一种深度学习方法，能够处理不规则图结构数据。该方法引入邻接矩阵$\boldsymbol{A}$，并在傅里叶域构造一个滤波器，用于作用于图的节点并获取节点的信息，而且还通过利用其一阶邻域来捕捉各个节点之间的空间特征，结合多个卷积层，构建出一个 GCN 模型，该模型可以用公式（2.7）表示为：

$$\boldsymbol{H}^{(l+1)} = \boldsymbol{\sigma}(\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\widehat{\boldsymbol{A}}\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{H}^{(l)}\boldsymbol{\theta}^{(l)}) \tag{2.7}$$

$\boldsymbol{A}$为邻接矩阵，$\boldsymbol{D}$为度矩阵，$\boldsymbol{H}^{(l)}$是l层的输出，$\boldsymbol{\theta}^{(l)}$是包含该层的参数，$\boldsymbol{\sigma}(\cdot)$表示非线性模型的$sigmoid$函数。

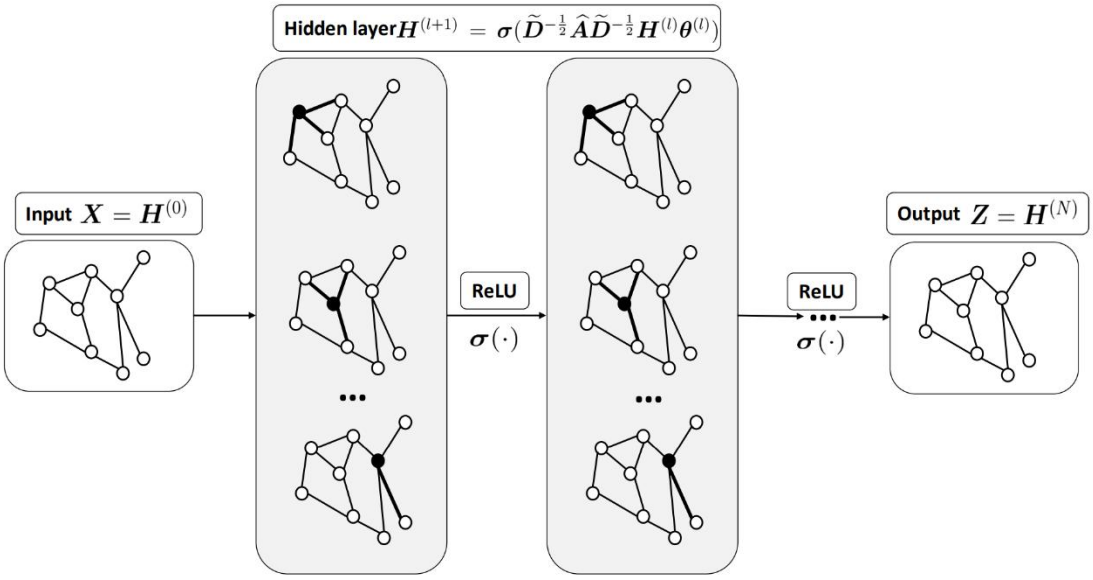


图 2.2 GCN 模型结构图

GCN 模型结构如图 2.2 所示，以两层 GCN 模型为例来获得空间依赖性，其可以用公式（2.8）表示为：

$$f(\boldsymbol{X}, \boldsymbol{A}) = \boldsymbol{\sigma}(\widehat{\boldsymbol{A}}\, ReLU(\widehat{\boldsymbol{A}}\boldsymbol{X}\boldsymbol{\theta_0})\boldsymbol{\theta_1}) \tag{2.8}$$

其中$\widehat{A}$表示的是加了自循环的邻接矩阵，$\boldsymbol{\theta_0} \in \boldsymbol{R}^{P\times H}$表示的是第一层网络的参数矩阵，P 是特征矩阵的长度，H 是隐含层的单元数。$\boldsymbol{\theta_1} \in \boldsymbol{R}^{H\times T}$表示的是第二次网络的参数矩阵。$f(\boldsymbol{X}, \boldsymbol{A}) \in \boldsymbol{R}^{N\times T}$表示当预测步长为$\boldsymbol{T}$时网络的输

出。$ReLU$表示的是激活函数，其公式为$max(0, \boldsymbol{X})$，$ReLU$激活函数会将输入的$\boldsymbol{X}$为负的值全部变为 0，$ReLU$函数能够缓解模型过拟合问题的发生，而且还能尽可能的防止梯度消失。因此，本文使用$ReLU$函数作为该 GCN 网络中的激活函数。

2.3.2 图注意力网络（GAT）

图注意力网络（Graph Attention Network，GAT）[25]更关注节点的信息聚合过程。自从注意力机制[23]诞生以来，在信息融合方面取得了极大的成功。GAT 建立在注意力机制之上，对当前节点的邻域信息进行加权融合，动态确定各邻接节点的权重。空间依赖关系的建模就需要使用到 GAT 模型，就是将注意力机制引入到图卷积网络后所产生的的模型，其目的是为了克服图卷积的缺陷：在消息传递和聚合时简单地将所有邻接信息取平均值，不区分不同邻接的重要性。GAT 网络结构如图 2.3 所示。

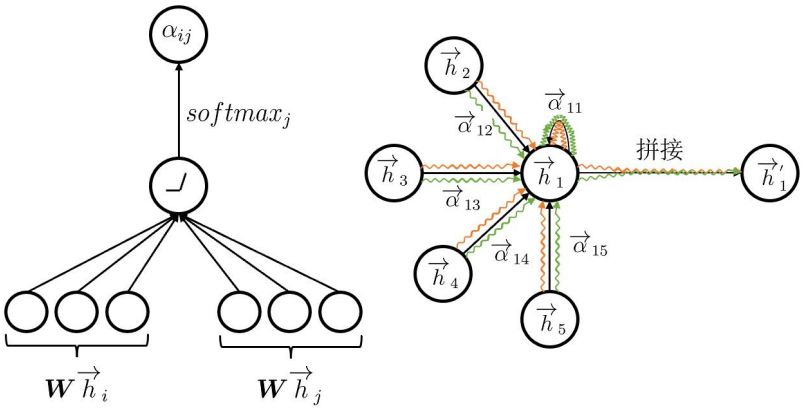


图 2.3 GAT 网络结构图

左图表示动态权重的计算过程，右图表示邻接信息的融合过程。记当前节点为i，其邻接节点为$j \in \mathcal{N}_i$，节点的向量表示为$\vec{h}$，则j对i的归一化注意力权重如公式（2.9）所示。

$$\alpha_{ij} = \frac{exp(LeakyReLU(\vec{a}^T[\boldsymbol{W}\vec{h}_i \| \boldsymbol{W}\vec{h}_j]))}{\sum_{K \in \mathcal{N}_i} exp(LeakyReLU(\vec{a}^T[\boldsymbol{W}\vec{h}_i \| \boldsymbol{W}\vec{h}_k]))} \tag{2.9}$$

其中exp表示指数函数，$LeakyRuLU$为激活函数，$\vec{\alpha}$和$\boldsymbol{W}$为可训练的转换向量和矩阵，$\|$表示矩阵的拼接操作。在信息融合的过程中，右图是 GAT 使用了多头注意力机制，对最终的信息进行多方面聚合，得到更新后的当前节点向

量表示$\overrightarrow{h'_i}$，相关公式如（2.10）所示

$$\overrightarrow{h'_i} = \|_{k=1}^{K}\sigma(\sum_{j\in\mathcal{N}_i}\alpha_{ij}^{k}\boldsymbol{W}^{k}\overrightarrow{h_j}) \tag{2.10}$$

GAT 模型在聚合过程中使用注意力，整合多个模型的输出，并生成面向重要目标的随机游走。GAT 是一种基于空间的图卷积网络，注意力机制用于确定节点邻域的权重，以聚合特征信息。GAT 的提出极大地增强了节点之间的信息聚合能力，将注意力机制以一种共享的方式应用在图的所有边上，提升了模型的鲁棒性。与此同时，该类做法适用于任何空域图神经网络，具有极好的普适性。

与图卷积网络（GCN）类似的地方在于，都是将相邻节点的特征信息聚合到中心顶点上，但不同的是 GCN 使用的是拉普拉斯矩阵，GAT 使用的是注意力系数。在 GAT 图中，每个节点根据相邻节点的特征可以获得不同的权重分配；引入注意力机制后，GAT 只与相邻节点相关，无需获取整张图的信息。因此，GAT 在一定程度上来说其预测效果要比 GCN 更好，因为它可以更好的整合顶点特征之间相关信息，并能够更好的传递到模型中去。

2.4 门控递归单元（GRU）

门控递归单元（GRU）是递归神经网络的一种门控机制，它是一种循环神经网络，具有可以有效地在长序列之间捕捉到语义关联的能力，能够减轻梯度消失或者梯度爆炸现象。门控递归单元与 RNN 和 LSTM 具有相同的输入和输出结构。具体而言，它们都包括了具有当前时刻的输入$\boldsymbol{x}_t$和之前时刻的节点传递过来的隐藏特征$\boldsymbol{h}_{t-1}$，这些隐藏特征$\boldsymbol{h}_{t-1}$的矩阵中都携带了关于之前节点的历史信息。结合$\boldsymbol{x}_t$和$\boldsymbol{h}_{t-1}$，GRU 模型会获得当前隐藏节点的输出，并将隐藏特征$\boldsymbol{h}_t$传递给下一个节点。GRU 模型的具体流程示意图[①]如图 2.4 所示。

[①] 图来源于 T-GCN_ A Temporal Graph Convolutional Network for Traffic Prediction.

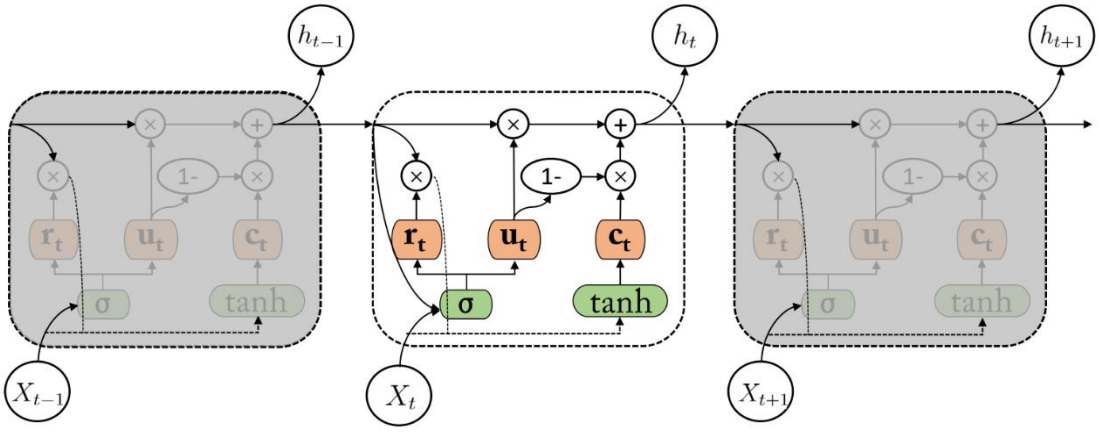


图 2.4 GRU 模型流程示意图

在 GRU 模型中，更新门决定了上一层状态中有多少信息量传递到当前状态，更新门向量中的值越大，则表示引入前一状态的信息越少，当前的信息写入的就越多。具体过程则是将前一个时刻神经单元的输入$\boldsymbol{h}_{t-1}$和当前时刻的输入$\boldsymbol{x}_t$输入到更新门$\boldsymbol{u}_t$中。而重置门$\boldsymbol{r}_t$用于控制当前时刻的候选值中有多少信息是来自前一个时刻，$\boldsymbol{r}_t$中的值越大，表示之前状态信息量被写入到候选值控制单元中的就越多。模型中用来存储记忆内容，并生成候选值$\boldsymbol{c}_t$的部分，其输入包括前一个时刻神经单元的输入$\boldsymbol{h}_{t-1}$、当前时刻的输入$\boldsymbol{x}_t$以及计算出来的$\boldsymbol{r}_t$；在计算完成$\boldsymbol{r}_t$、$\boldsymbol{u}_t$、$\boldsymbol{c}_t$之后，就可以生成该单元的输出值并传递到输出层。具体计算过程如公式（2.11）-（2.14）所示。

$$\boldsymbol{u}_t = \boldsymbol{\sigma}(\ \boldsymbol{W}_u[\ \boldsymbol{X}_t, \boldsymbol{h}_{t-1}] + b_u) \tag{2.11}$$

$$\boldsymbol{r}_t = \boldsymbol{\sigma}(\ \boldsymbol{W}_r[\ \boldsymbol{X}_t, \boldsymbol{h}_{t-1}] + b_r) \tag{2.12}$$

$$\boldsymbol{c}_t = tanh(\ \boldsymbol{W}_c[\ \boldsymbol{X}_t, (\ \boldsymbol{r}_t \odot \boldsymbol{h}_{t-1})] + b_c) \tag{2.13}$$

$$\boldsymbol{h}_t = \boldsymbol{u}_t \odot \boldsymbol{c}_t + (\ 1 - \boldsymbol{u}_t) \odot \boldsymbol{h}_{t-1} \tag{2.14}$$

获取时间依赖性是预测 $PM_{2.5}$ 浓度的另一个关键问题。要获取时间依赖性，这就需要利用好门控机制作用，尽可能多地来记忆长期信息。前文也有所提及，在门控机制中，GRU 模型的内部结构对于 LSTM 模型的结构来说要更简单一些，其所需要训练的参数也较少，不仅如此，该模型训练所需的时间也相应的减少

了一些。本文涉及到的门控递归单元模型以$t-1$时刻保存的历史信息和当前的序列信息作为输入，获得时刻t的样本变化情况。

2.5 长短期记忆网络（LSTM）

长短期记忆网络（Long Short Term Memory, LSTM）是一种循环神经网络（RNN）的变体，可以增强记忆力，可用来解决稳定性和梯度消失问题[10]。LSTM 作为 RNN 的一个改进版，通过添加一个单元状态来解决长期存储问题，影响输出的有当前和上一时刻的输出。网络中较为重要的组成是细胞状态和门控单元。其中，细胞状态用来存储网络中的信息，而门控单元是用来控制细胞状态的网络结构，由一个网络层和一个逐点相乘运算组成。LSTM 的每一个模块内有 4 层结果结构，包括 3 个$Sigmoid$层和一个$tanh$层。其中的$Sigmoid$激活函数会将经过该门控单元的值映射到 0 到 1 之间。映射的输出值表示相应元素的重要性，0 表示不允许传递任何信息，1 表示所有信息都可以传递。门控单元包括 3 个门控单元：遗忘门、记忆门和输出门。LSTM 网络结构如图 2.5 所示。

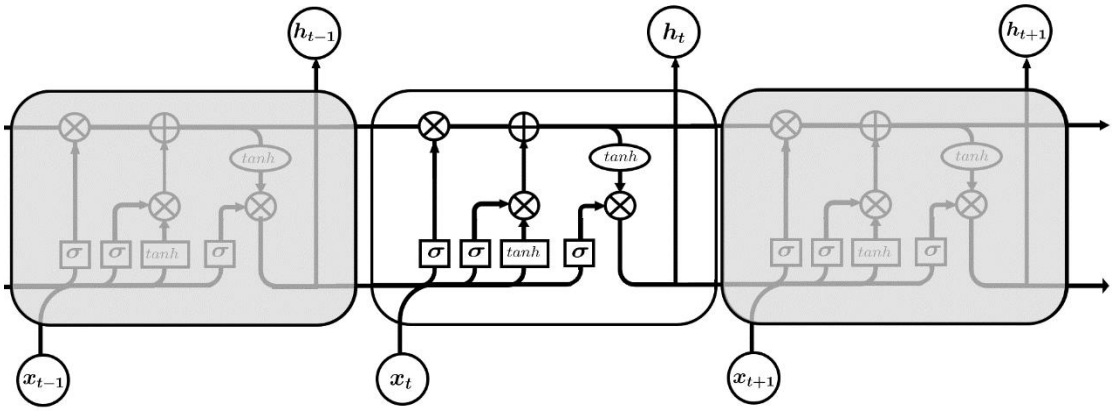


图 2.5 LSTM 网络结构图

其中，遗忘门的作用就是门控单元决定需要遗忘和保留的信息，记忆门就是当遗忘门丢弃了一部分信息后，记忆门会计算筛选出新的信息进行记忆补充，也就是上一步的状态$\boldsymbol{C}_{t-1}$已经被忘记了一部分，接下来就是把有效信息添加进去，再让$\boldsymbol{C}_{t-1}$经过遗忘门丢弃部分信息，加上更新部分信息，生成新的状态$\boldsymbol{C}_t$，输出门就是将当前时刻的输入$\boldsymbol{x}_t$和上一时刻的隐藏状态$\boldsymbol{h}_{t-1}$通过神经元进行非线性映射得到的输出$\boldsymbol{o}_t$，然后将$\boldsymbol{o}_t$和投入到$tanh$函数中的$\boldsymbol{C}_t$相乘得到隐藏信息$\boldsymbol{h}_t$，最后得到输出值的候选项最终运用到预测中去。具体的计算公式如（2.15）-

（2.20）所示。

$$\boldsymbol{f_t} = \boldsymbol{\sigma}(\boldsymbol{W}_f \cdot [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + b_f) \tag{2.15}$$

$$\boldsymbol{i}_t = \boldsymbol{\sigma}(\boldsymbol{W}_i \cdot [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + b_i) \tag{2.16}$$

$$\tilde{\boldsymbol{C}}_t = tanh(\boldsymbol{W}_C \cdot [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + b_C) \tag{2.17}$$

$$\boldsymbol{C}_t = \boldsymbol{f}_t \cdot \boldsymbol{C}_{t-1} + \boldsymbol{i}_t \cdot \tilde{\boldsymbol{C}}_t \tag{2.18}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o \cdot [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + b_o) \tag{2.19}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \cdot tanh(\boldsymbol{C}_t) \tag{2.20}$$

其中，$\boldsymbol{x}_t$表示的是当前时刻的输入，$\boldsymbol{h}_{t-1}$表示上一时刻的隐藏状态，σ表示$Sigmoid$激活函数，$\boldsymbol{W}_f$和b_f分别是权重和偏置项。将前一步信息$\boldsymbol{h}_{t-1}$与本次输入信息$\boldsymbol{x}_t$拼接起来，得到的矩阵线性运算后传给$Sigmoid$函数，映射输出$\boldsymbol{f}_t$，该$\boldsymbol{f}_t$就是 0 到 1 之间的值，该值直接决定了状态信息保留多少，也就是代表信息被遗忘的概率大小是多少。记忆门主要有两层结构：一个$tanh$层用来产生更新值的候选项$\boldsymbol{C}_t$，也就是$tanh$激活函数映射后得到的输出值$\boldsymbol{C}_t$表示学习到的新知识，$tanh$的输出在$[-1, 1]$上，说明细胞状态在某些维度上需要增强，在某些维度上需要减弱，还有一个$Sigmoid$层，其输出值要乘到$tanh$层的输出上，起到一个缩放的作用，极端情况下$Sigmoid$输出 0 说明相应维度上的细胞状态不需要更新。

2.6 本章小结

本章节主要旨在介绍了本文在处理数据、相关性分析以及模型预测方面所需要涉及到的理论方法，主要介绍了 $PM_{2.5}$ 浓度的时空特征分析方法，主要包括 KNN 算法、归一化方法以及 Spearman 相关系数法，之后对图的构建利用了地理距离计算方法、余弦距离相似性度量方法做出介绍，最后对图卷积网络（GCN）、图注意力网络（GAT）、门控递归单元（GRU）、长短期记忆网络（LSTM）模型与算法流程公式进行了概述。

3 $PM_{2.5}$与各维度相关分析

3.1 问题描述

对$PM_{2.5}$浓度序列进行预测仅考虑它与其他五种污染物之间的影响情况，并不能得到很好的效果。事实上，影响$PM_{2.5}$浓度的因素还有很多，如温度、湿度、风速、风向等气象数据[16]，除此之外，$PM_{2.5}$浓度还可能受到POI、路网等非时序数据的影响[33]。不同的空气质量监测站点监测到的$PM_{2.5}$浓度值对待预测的空气质量监测站点监测的$PM_{2.5}$浓度值的相关程度是不同的，例如，离目标站点较近的空气质量监测站点中的影响因素对$PM_{2.5}$浓度预测具有较高的参考价值，反之，离目标站点较远的监测站点中的影响因素对其预测的参考价值较小。除此之外，还需考虑不同因素对目标空气质量监测站点的影响也会随时间的推移而有所变化，比如，在某一时刻目标空气质量监测站点下过一场雨，那么该空气质量监测站点中的影响因素对该站点监测到的$PM_{2.5}$浓度会比过去没下雨时影响更大。

3.2 数据来源

在本次实验中，研究区域以甘肃省兰州市为例，使用的数据集包括空气质量数据、气象数据、空气质量监测站点经纬度信息、气象站点经纬度信息、兴趣点数据以及路网数据。

（1）空气质量数据：取自中国环境监测总站的全国城市空气质量实时发布平台[①]，本研究收集了甘肃省兰州市8个空气质量监测站点自2021年1月1日至2023年4月1日每隔3小时的空气质量监测数据，空气质量记录数据总数超过31.5万条，数据中包含6个污染物属性，分别是细颗粒物（$PM_{2.5}$）、可吸入颗粒物（PM_{10}）、二氧化硫（SO_2）、二氧化氮（NO_2）、臭氧（O_3）、一氧化碳（CO）。如表3.1所示。

[①] 中国环境监测总站网址：http://www.cnemc.cn/sssj/

表 3.1 空气质量污染物属性表

编号	数据名称	单位
1	$PM_{2.5}$	$\mu g/m^3$
2	PM_{10}	$\mu g/m^3$
3	SO_2	$\mu g/m^3$
4	NO_2	$\mu g/m^3$
5	O_3	$\mu g/m^3$
6	CO	mg/m^3

（2）气象数据：取自 NCDC（美国国家气候数据中心[①]，National Climatic Data Center），本研究收集了甘肃省 11 个气象站点自 2021 年 1 月 1 日至 2023 年 4 月 1 日每隔 3 小时的气象数据，气象数据记录总数超过 5.1 万条，数据中包括 6 个属性，分别是气温（Temperature）、露点（Dew point）、气压（Pressure）、风向（Wind direction）、风速（Wind speed）、降雨量（Precipitation）。如表 3.2 所示。

表 3.2 气象数据属性表

编号	数据名称	单位
1	气温	℃
2	露点	℃
3	气压	hpa
4	风向	——
5	风速	m/s
6	降雨量	mm

（3）站点经纬度数据：空气质量监测站点和气象监测站点的经纬度坐标，利用各个站点的经纬度得到站点之间的欧式距离。针对不同站点监测的空气质量数据和气象数据，本文利用泰森多边形法 计算气象监测站点周围的 $PM_{2.5}$ 浓度，将气象监测站点监测到的 $PM_{2.5}$ 浓度与其相似的空气质量站点监测到的数据进行合并。利用泰森多边形计算出每个气象监测站点所属的泰森多边形的母

[①] 美国国家气候数据中心网址：https://www.ncei.noaa.gov/

站点索引，通过计算每个气象站点到所有空气质量监测站点的距离，最终得到气象监测站点和空气质量监测站点的冯洛诺伊（Voronoi）图，结果如图 2 所示。

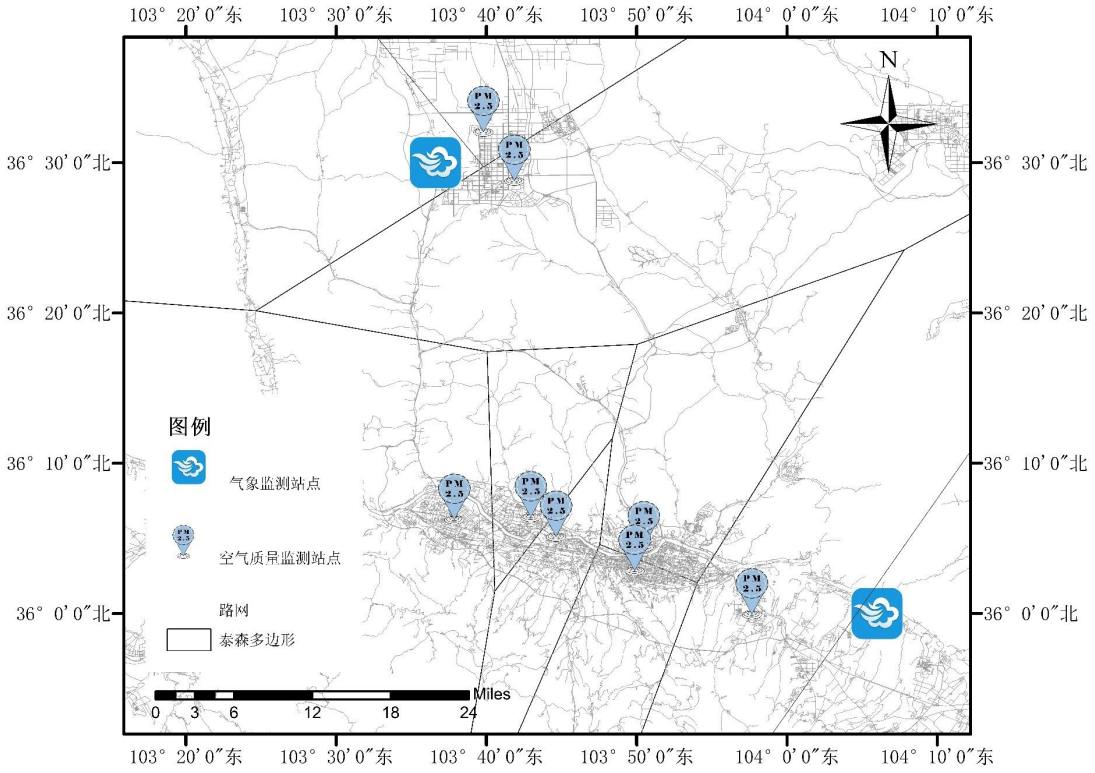


图 3.1 气象监测站点和空气质量监测站点的 Voronoi 图

图 3.1 为气象监测站点和空气质量监测站点的 Voronoi 图，利用泰森多边形计算气象站点周围的 $PM_{2.5}$ 浓度情况后，由图 3.1 可以看出，左上方三个空气质量监测站点的 $PM_{2.5}$ 的平均浓度可以作为左上气象站点监测到的 $PM_{2.5}$ 浓度。其余空气质量监测站点监测到的 $PM_{2.5}$ 浓度平均值作为右下气象监测站点监测到的 $PM_{2.5}$ 浓度。在合并空气质量和气象数据的过程中，该部分将空气质量数据和气象数据中的监测站点和时间列合并为监测数据的关键值。采用监测站点列和时间列值作为数据融合的条件，当两个关键值相同时，进行相应的空气质量数据和气象数据的融合。表 3.3 是获取到的甘肃省兰州市 8 个空气质量监测站点的经纬度数据列表。

表 3.3 空气质量监测站点列表

监测站点 ID	监测站点名称	所在城市	经度	纬度
1476A	兰炼宾馆	兰州	103.631	36.1031
1478A	生物制品所	兰州	103.8412	36.0726
1479A	铁路设计院	兰州	103.831	36.0464
3186A	教育港	兰州	103.7158	36.1058
3241A	百合公园	兰州	103.7439	36.0842
3242A	和平	兰州	103.9611	35.998
3245A	新区管委会	兰州新区	103.6633	36.533
3246A	舟曲中学	兰州新区	103.6975	36.4789

（4）POI 数据：取自 POI 规划云[①]。根据每个 POI 的类别标签，我们将这些数据划分到 9 个大的类别，具体类别以及数量如表 3.4 所示。

表 3.4 POI 类别表

编号	POI 类别
1	餐厅
2	工厂
3	公园
4	购物中心
5	酒店
6	商务写字楼
7	学校
8	医院
9	住宅小区

（5）路网数据：取自 OpenStreetMaps 网站[②]的数据，道路可以根据其性质分为五类: 高速公路、一级公路、二级公路、三级公路和四级公路。如表 3.5 所示。

[①] POI 规划云网址：http://guihuayun.com/poi/

[②] OpenStreetMaps 网站：https://www.openstreetmap.org

表 3.5 道路等级

编号	道路等级
1	高速公路
2	一级公路
3	二级公路
4	三级公路
5	四级公路

3.3 数据预处理

本次实验所使用到的初始数据集不能直接使用，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，所以需要先进行处理得到我们想要的数据才能进行下一步的实验，所以下面介绍对 $PM_{2.5}$ 浓度预测实验所使用到的数据集的处理方式。

3.3.1 污染物、气象数据处理

（1）缺失值填充

由于数据为小时数据，考虑到空气质量数据在相近的数小时内变化不大，利用 K 最近邻算法对缺失值进行填充。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。本文将使用 Python 中的 sklearn 的 impute 模块中的 KNNImputer 函数对空气质量数据和气象数据进行缺失值的填充。

（2）数据归一化

获取到的污染物数据集和气象数据集之间存在着量纲、单位和数据类型的不同，所以，本文在模型训练之前采用了归一化的方式将每个数据点都映射到[0,1]之间，后续方便输入到预测模型中。

3.3.2 POI 特征、路网数据处理

兴趣点（POI）是指城市中一些关键性的地标点，POI 的不同分类会对附近的空气质量产生不同的影响，一个区域内 POI 的数量分布情况能够反映出该区域的类型[40]，因此，根据 POI 的不同属性，事先已经将 POI 划分成 9 种类别。

在实验中我们需要统计出兰州市 8 个空气质量监测站点周围的 9 种 POI 的

数量，我们通过 Arcgis 软件，将 POI 数据映射到地图当中，甘肃省兰州市的 8 个空气质量监测站点周围 POI 分布情况如图 3.2 所示。

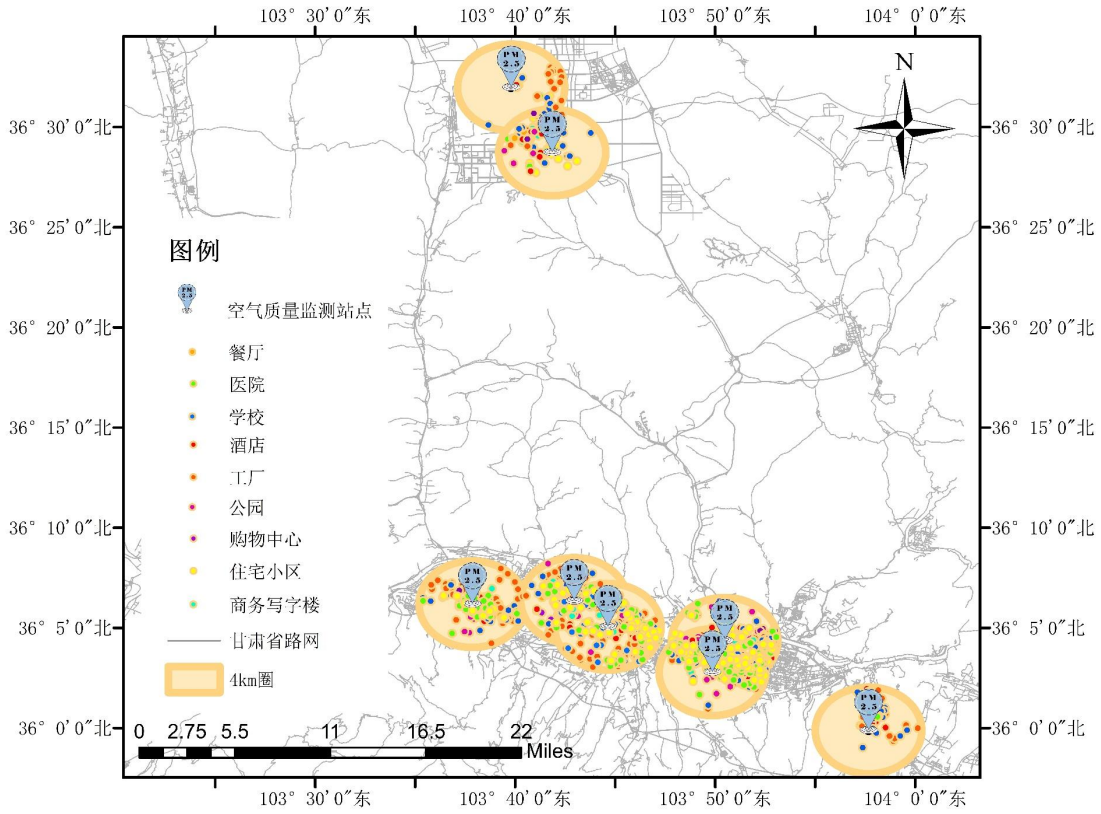


图 3.2 空气质量监测站点周围 4 千米 POI 类别数量分布图

图 3.2 为兰州市空气质量监测站点与其周围 POI 类别数量分布图，蓝色标点表示兰州市 8 个空气质量监测站点，蓝色标点周围的黄色区域表示以空气质量监测站点为圆心的 4 公里区域[56]。通过软件统计出每个空气质量监测站点周围 4 千米内每个类别 POI 的数量，为每个空气质量站点构建一个 9 维向量，其中每个维度表示一种 POI 的数量。图中的黄色区域内为 9 种 POI 类别个数，其中包括餐厅、医院、学校、酒店、工厂、公园、购物中心、住宅小区和商务写字楼 9 类公共服务点。

3.3.3 路网数据处理

众所周知，交通状况会明显的影响周围的空气质量。汽车尾气排放被认为是城市污染的一大主要来源，同时，路网结构与实际交通状况密切相关，因此，本实验在模型中融入不同类型的路网特征是非常有必要的。

路网数据的处理过程和 POI 数据类似，根据路网结构的不同属性，事先已经根据路网的道路等级标签划分成 5 个等级类别。将路网数据投射到地图中，

甘肃省兰州市的 8 个空气质量监测站点周围路网的分布情况，如图 3.3 所示。

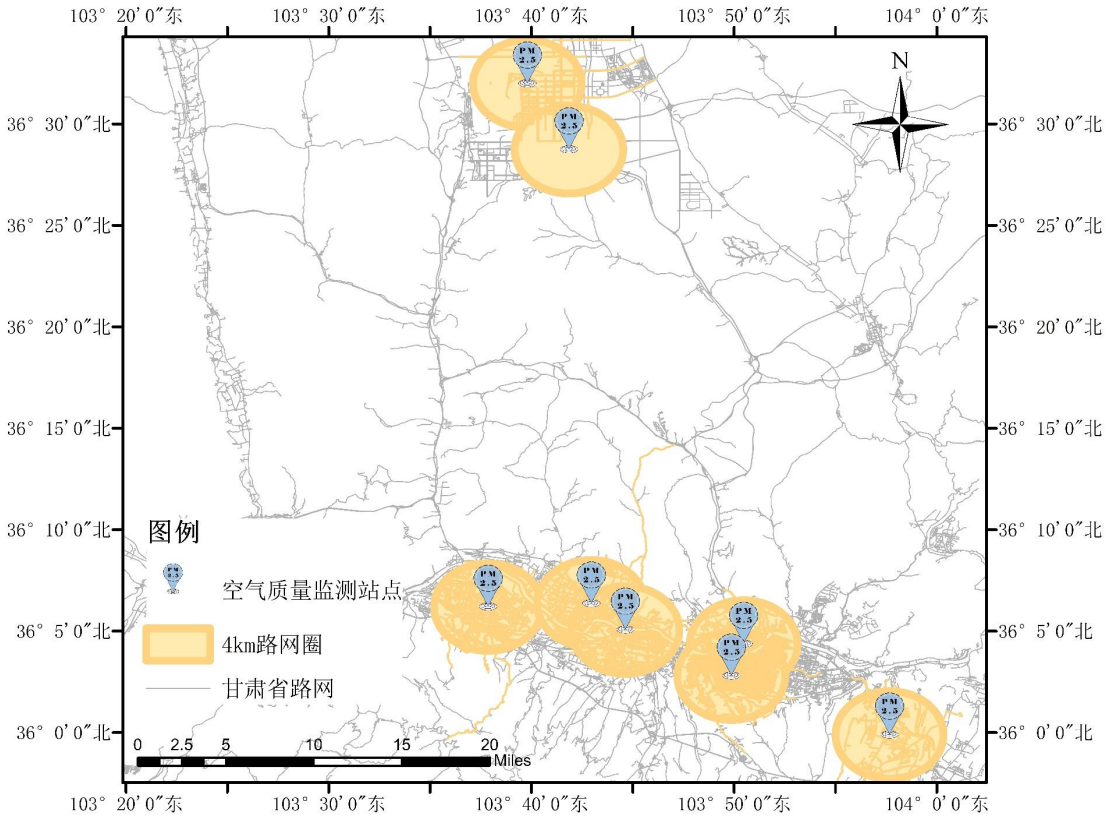


图 3.3 空气质量监测站点周围 4 千米路网分布图

图 3.3 为兰州市空气质量监测站点与其周围 4 千米路网分布图，蓝色标点表示兰州市 8 个空气质量监测站点，蓝色标点周围的黄色区域表示以空气质量监测站点为圆心的 4 公里区域[56]。通过 Arcgis 软件，将每个空气质量监测站点周围 4 千米内的路网分别切割独立，然后统计每个空气质量监测站点周围不同等级路网的路网长度，可以构造出一个 5 维的路网向量，其中每个维度表示一种路网的长度数据。图中的黄色区域内为 4 公里区域的路网结构，其中包括了高速公路、一级公路、二级公路、三级公路和四级公路 5 个类别。

3.4 $PM_{2.5}$ 等级划分

为研究 $PM_{2.5}$ 浓度对空气质量影响的情况，根据中国环境科学研究院发布的《环境空气质量标准》文件可知[57]，空气质量等级 24 小时 $PM_{2.5}$ 平均值按其数值在 0-35 微克评级为优，35-75 微克评级为良，75-115 微克评级为轻度污染，115-150 微克评级为中度污染，150-250 微克评级为重度污染，250 微克以上评级为严重污染。如表 3.6 所示。

表 3.6 空气质量等级 24 小时 $PM_{2.5}$ 平均值分布表

空气质量等级	24 小时 $PM_{2.5}$ 浓度平均值
优	0~35$\mu g/m^3$
良	35~75$\mu g/m^3$
轻度污染	75~115$\mu g/m^3$
中度污染	115~150$\mu g/m^3$
重度污染	150~250$\mu g/m^3$
严重污染	大于 250$\mu g/m^3$

根据上表将获取到的空气质量数据中的 $PM_{2.5}$ 浓度值按照其平均值在 75 微克以上的认为该天空气质量被污染，本章统计出甘肃省兰州市 8 个空气质量监测站点近两年被污染的天数。如图 3.4 所示。

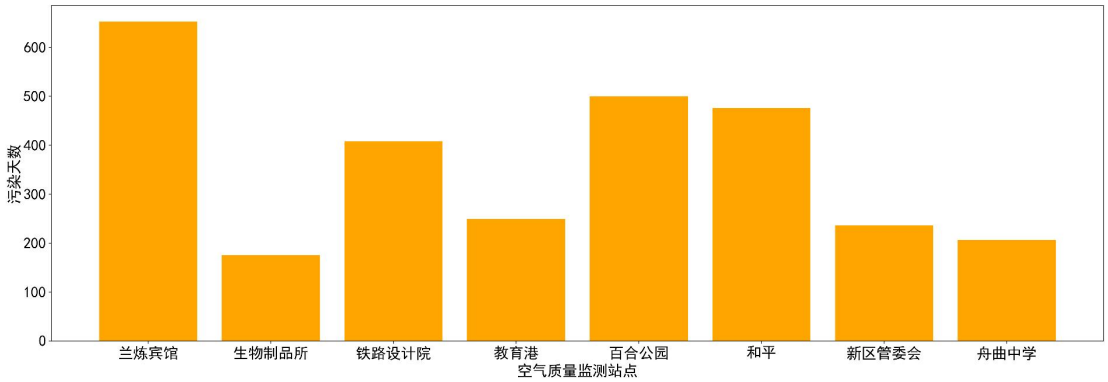


图 3.4 甘肃省兰州市 8 个空气质量监测站点近 2 年的污染天数

图 3.4 中可以看出不同的空气质量监测站点监测到被污染的天数都有所不同，影响空气质量监测站点监测到的 $PM_{2.5}$ 浓度值变化的因素有很多，猜测可能包括时间因素、其他污染物因素、气象因素、周围 POI 数量以及路网结构等。

3.5 $PM_{2.5}$ 与时间的关系

为研究 $PM_{2.5}$ 和一天中时间的关系，根据前文《环境空气质量标准》将 $PM_{2.5}$ 浓度按空气质量等级划分，对某个空气质量监测站点进行统计当天的一段时间内 $PM_{2.5}$ 浓度值的分布情况，如图 3.5 所示。

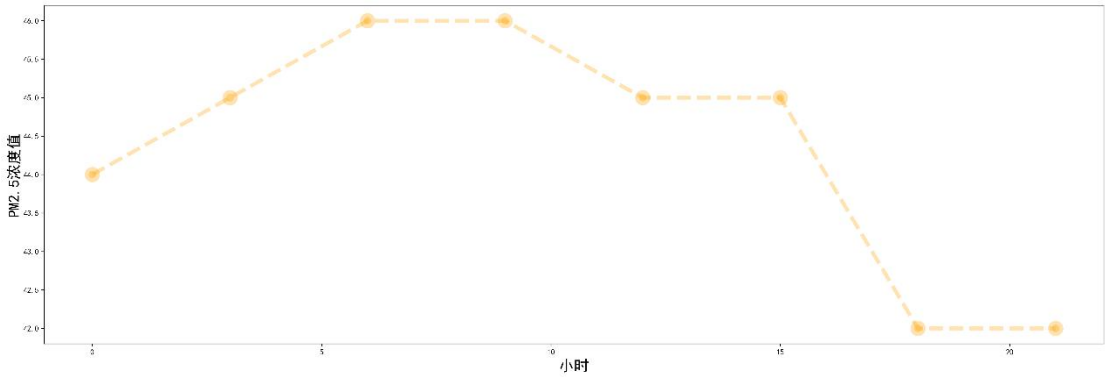


图 3.5 一天中 $PM_{2.5}$ 浓度值变化情况

可以发现在晚上和凌晨的时候，$PM_{2.5}$ 浓度数值是偏低，可能是由于凌晨人们的出行活动相较于白天来说减少很多，交通、工厂等人为造成的污染源减少了，而在早上到下午这段时间，$PM_{2.5}$ 的浓度值是偏高的，可能是因为这个时间段人们活动的较为频繁，导致 $PM_{2.5}$ 浓度值升高。

为了研究 $PM_{2.5}$ 的浓度在一周内的分布情况，对某个空气质量监测站点进行统计，统计其在一段时间内 $PM_{2.5}$ 浓度值的分布情况，如图 3.6 所示。

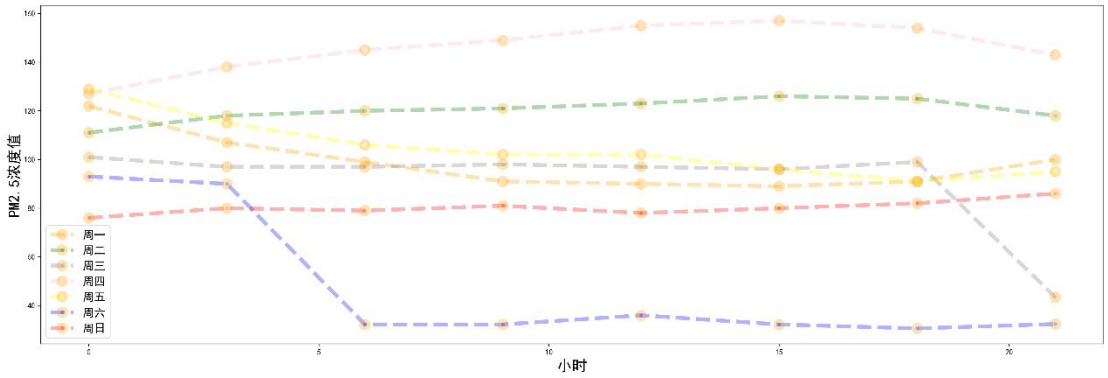


图 3.6 一周 $PM_{2.5}$ 浓度变化情况

可以发现，在周六和周日的 $PM_{2.5}$ 浓度值是比工作日的 $PM_{2.5}$ 浓度值要低得，可能是因为周末人们倾向于在家休息，使得该空气质量监测站点周边的地区活动、人流量减少，引起 $PM_{2.5}$ 的浓度值降低。对于其他空气质量监测站点监测的 $PM_{2.5}$ 浓度变化趋势可能不一致，但总体上可以看出，$PM_{2.5}$ 和时间上是存在着一定的联系的。

3.6 $PM_{2.5}$ 与污染物、气象特征之间的关系

在考虑到气象一系列特征对 $PM_{2.5}$ 浓度的影响因素时，有很多其他因素并没有被充分考虑到，污染物之间会产生化学反应导致污染物浓度的变化，致使

污染物之间的浓度也会互相影响，为研究 $PM_{2.5}$ 与污染物、气象因素之间的关系，获取了的 6 个空气质量污染物特征，包括 $PM_{2.5}$、PM_{10}、SO_2、NO_2、O_3、CO，与气象因素的 6 个特征，包括气温、露点、气压、风向、风速、降雨量。

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素之间的具体关系，为后续时空预测模型构建提供充分的数据基础。将这 12 个特征进行相关分析，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，构建出相关关系矩阵，并绘制出相关关系热力图，如图 3.7 所示。

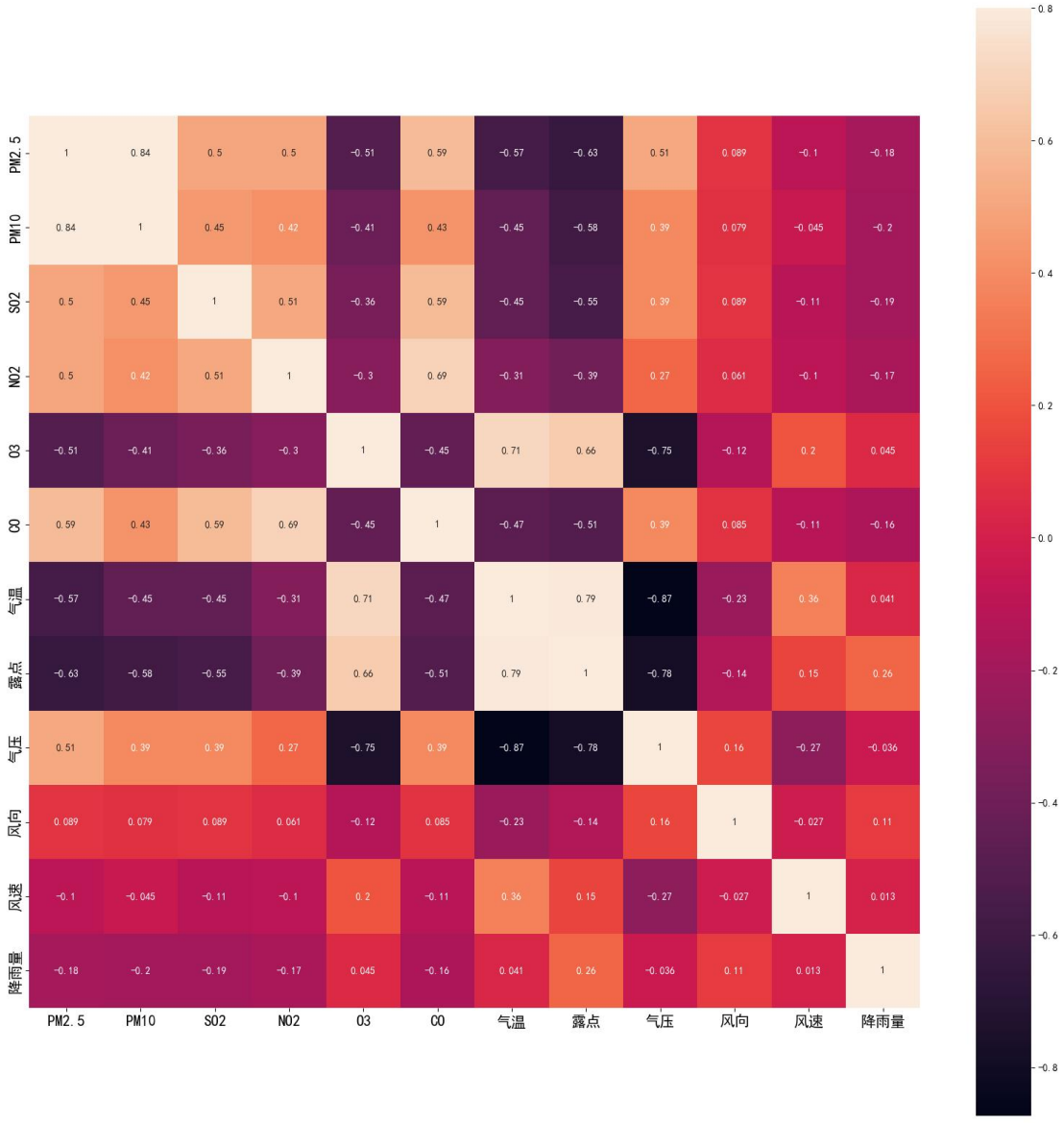


图 3.7 气象特征与污染物特征的相关关系热力图

从图 3.7 中观察可以得出甘肃省兰州市的 $PM_{2.5}$ 浓度序列与其他空气污染物

浓度序列及气象因素之间的相关性，大体可以看出 $PM_{2.5}$ 浓度与 PM_{10}、SO_2、NO_2、O_3、CO、气温、露点、气压、风速和降雨量呈高度相关，与风向相关性较弱。本节计算了 $PM_{2.5}$ 与其他 5 污染物浓度、6 个气象属性的 Spearman 等级相关性，旨在能够更加仔细地度量 $PM_{2.5}$ 浓度跟其他影响因子之间的相关程度，其相关系数值如表 3.7 所示。

表 3.7 $PM_{2.5}$ 与其他污染物、气象因子之间的相关系数值

影响因子	Spearman 相关系数值	影响因子	Spearman 相关系数值
$PM_{2.5}$	1.000	气温	-0.570
PM_{10}	0.836	露点	-0.626
SO_2	0.503	气压	0.506
NO_2	0.501	风向	0.0891
O_3	-0.506	风速	-0.102
CO	0.593	降雨量	-0.181

表 3.7 反映了 $PM_{2.5}$ 浓度与其他 5 种空气质量污染物以及 6 类气象因素之间的相关程度。PM_{10} 浓度、SO_2 浓度、NO_2 浓度、CO 浓度与 $PM_{2.5}$ 浓度呈现强正相关性，而 O_3 与 $PM_{2.5}$ 浓度呈现强负相关性。气温、露点、气压与 $PM_{2.5}$ 之间关联较强，风速、降雨量对 $PM_{2.5}$ 浓度的影响较弱，风向则与 $PM_{2.5}$ 浓度的关联性较差。基于学界一致的相关性度量水平， 0-0.1 表示无相关性，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[45]。本研究选取 Spearman 相关系数绝对值在 0.1 以上作为特征选择依据，可简便地揭示空气污染物与气象因素之间的因果或关联关系，为后续建模提供支持。

3.7 $PM_{2.5}$ 与 POI 的关系

兴趣点（POI）是一个与人们生活息息相关的地理位置点，一个区域各类 POI 的数量及其密集程度可以反映该区域的功能和土地利用情况，而这些因素可能会影响这个区域的 $PM_{2.5}$ 浓度，例如空气质量监测站点周围工厂较多的要比周围是公园景区的 $PM_{2.5}$ 的浓度要高一些。

前文已经事先将 POI 分成了 9 个类别，并且在前文的讨论中已知，空气质量监测站点能够监测到的范围大约在 4 公里范围内，所以，为研究 $PM_{2.5}$ 与 POI

之间的关系，也就获取了空气质量监测站点周围 4 公里的 POI 的数量。

本文统计了各个空气质量监测站点周边的 POI 总量，根据前文所给出的根据我国的《环境空气质量标准》将 $PM_{2.5}$ 的浓度值按空气质量等级划分，统计出 2021 年 1 月至 2023 年 3 月被污染天数，即轻度污染及以上的天数。分析 $PM_{2.5}$ 与 POI 数量之间的关系，如图 3.8 所示。

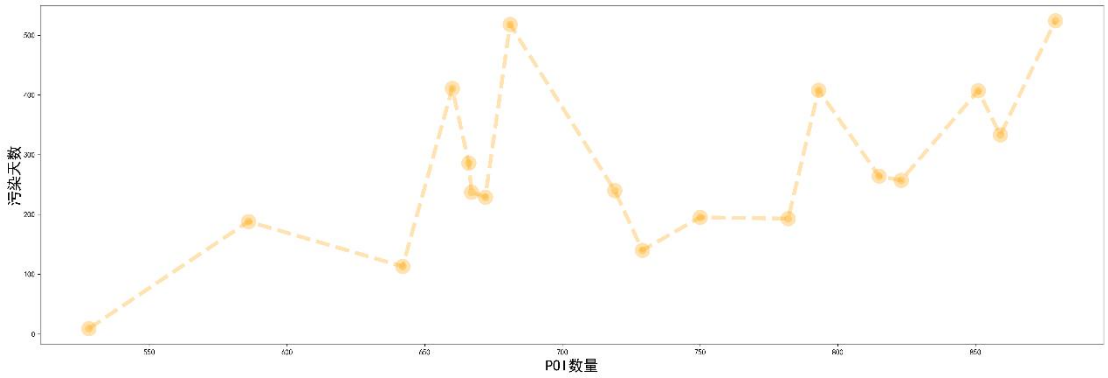


图 3.8 POI 数量与污染天数关系图

从 $PM_{2.5}$ 与 POI 数量之间的关系可以发现，整体上看，POI 的数量越多，污染天数就越多，可能是由于附近 POI 数量越多，在该区域的土地使用率较高，各种人为活动也随之增多，随之产生的 $PM_{2.5}$ 的浓度也会偏高。

3.8 $PM_{2.5}$ 与路网的关系

前文已经事先将道路等级划分成了 5 个等级，并且在前文的讨论中已知，空气质量监测站点能够监测到的范围大约在 4 公里范围内，所以，为研究 $PM_{2.5}$ 与不同等级道路之间的关系，也就获取了空气质量监测站点周围 4 公里的不同等级道路的总长度。本文统计了各个空气质量监测站点周边不同等级道路的总长度，根据前文所给出的环境空气质量标准，将 $PM_{2.5}$ 的浓度值按空气质量等级划分，统计出 2021 年 1 月至 2023 年 3 月被污染天数，即轻度污染及以上的天数。分析 $PM_{2.5}$ 与空气质量监测站点周围 4 千米的路网长度之间的关系，如图 3.9 所示。

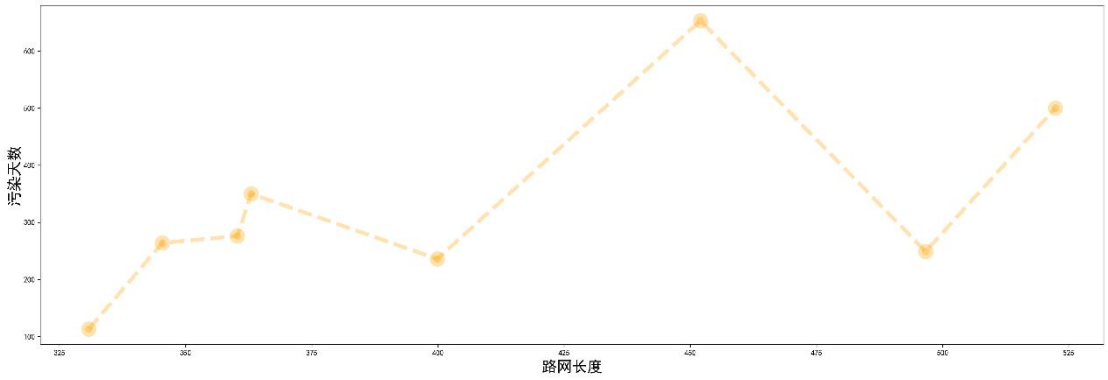


图 3.9 路网长度与污染天数的关系图

从 $PM_{2.5}$ 与空气质量监测站点周围 4 千米的路网长度之间的关系可以发现，整体情况下，周边的道路总长度越长，污染天数就越多，有可能是因为附近道路状况越复杂，说明在该区域的道路通达度较高，各种人为活动也随之增多，随之产生的 $PM_{2.5}$ 的浓度也会偏高。

3.9 本章小结

本章首先介绍了 $PM_{2.5}$ 在甘肃省兰州市内的一个分布情况，每一小节简单的分析了 $PM_{2.5}$ 与时间、其他空气污染物、气象因子、POI 特征以及路网结构之间的关系，认为上述特征均会对 $PM_{2.5}$ 的浓度变化产生影响，并由此提出如下假设，若要预测某个空气质量监测站点未来一段时间内的 $PM_{2.5}$ 的浓度值，除了考虑当前监测站点和周边监测站点的历史 $PM_{2.5}$ 的浓度之外，以下四种数据对该预测都有一定的辅助预测作用：

（1）当前状态的时间特征，包括当前时间属于一天当中的第几个小时，以及属于一周当中的第几天。

（2）当前空气质量监测站点以及周边空气质量监测站点的气象情况，包括气温、露点、气压、风向、风速和降雨量。

（3）POI 特征，包括不同类别的 POI 数量，以及 POI 总量。

（4）路网数据，包括不同等级的道路总长度。

本文将在下一章给出设计预测 $PM_{2.5}$ 浓度的预测模型，并用一系列实验验证上述猜想。

4 基于多视角图卷积网络的 $PM_{2.5}$ 浓度预测

4.1 问题描述

前文也提到，对 $PM_{2.5}$ 浓度序列进行预测不仅仅考虑其他污染物因素对其影响，还应该考虑到气象因素、POI 特征以及道路状况对其产生的影响，上一章节分析了 $PM_{2.5}$ 浓度在多维度上与众多特征的相关性，由于在这些特征中存在相关性小或者不相干的序列，若直接将所有特征的时序数据作为输入用来捕获不同空气质量监测站点之间的相关性，将会导致很高的计算成本并且降低预测性能[17]。为此，本章节只考虑了上一章节 $PM_{2.5}$ 浓度特征与其他特征相关系数绝对值大于 0.1 为特征选择条件，将符合的特征作为输入，且空气质量站点在空间上表现为一种拓扑图结构，以选择好的特征来捕捉空气质量监测站点之间的相关性的依据，本章节考虑借鉴图卷积网络的思想来提取数据中存在的空间信息。并利用 GRU 模型提取空气质量监测站点的时间信息。

4.2 拓扑图的构建

4.2.1 地理距离网络图

空气质量监测站点之间的地理距离也影响着其相关程度，即距离越近，相关性越强。其地理距离一般由两两站点的经纬度进行计算，根据第二章节给出的地理距离计算公式（2.5）可以计算出两两站点之间的地理距离，根据实际需求定义阈值 D 来界定节点之间是否有连通，计算公式如（4.1）所示。

$$e_{ij}^{D}=\begin{cases}1, & h_D(x_i,x_j)\le 50km\\ 0, & \text{其他}\end{cases} \tag{4.1}$$

$h_D(x_i,x_j)$是根据地理空间距离公式求出的两个空气质量监测站点之间的地理距离。将空气质量监测站点之间在根据实际需求定义的阈值下取的有效距离作为其边构建出空气质量监测站点的信息网络结构图。本文取 50 公里的阈值[56]，小于 50 公里则连接空气质量监测站点之间的边，即e_{ij}^{D}为 1，反之为 0。这样，就得到了一个基于有效地理距离的信息网络图。可用$\boldsymbol{G_1}(\boldsymbol{V},\boldsymbol{E})$来表示两两空气质量监测站点之间的距离关系。$\boldsymbol{V}$表示的是空气质量监测站点的集合，$\boldsymbol{E}$表示的是任意两个空气质量监测站点之间边$e_{ij}^{D}$的集合。各空气质量监测站点以有效

距离为边组成的信息网络结构如图 4.1 所示。

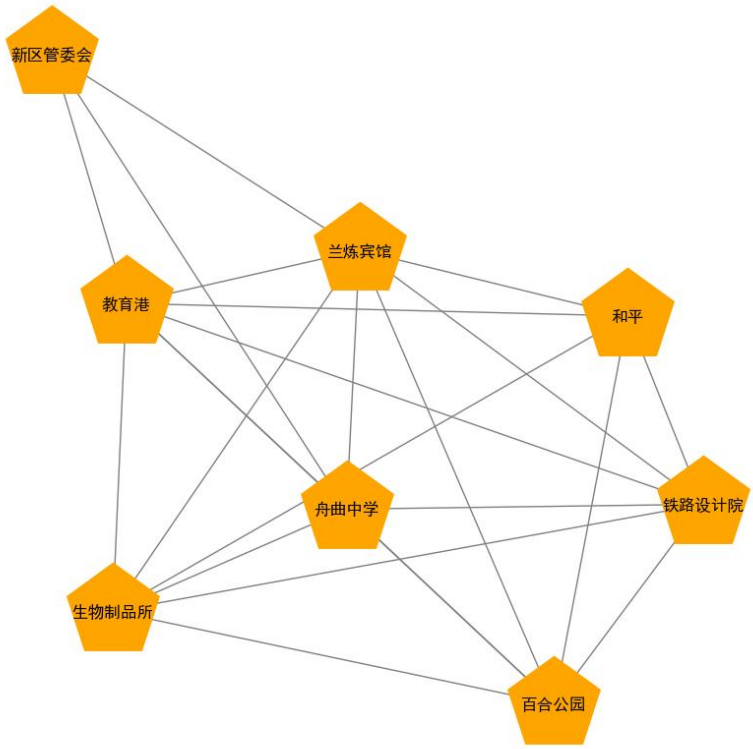


图 4.1 以有效地理距离为边组成的信息网络图

4.2.2 兴趣点网络图

兴趣点是指日常生活中常见的点状地理要素，对附近的空气质量监测站点监测 $PM_{2.5}$ 浓度也有一定的影响。上一章节给出了空气质量监测站点能够监测到的范围大约在 4 公里范围内，所以，为研究 $PM_{2.5}$ 与 POI 之间的关系，也就获取了空气质量监测站点周围 4 公里的 POI 的数量。将空气质量监测站点和 POI 类别之间的特征向量表示转换为矩阵形式。对于每个空气质量监测站点，使用 9 个 POI 类别数量来表示其周围的 POI，具体而言，这个矩阵的每一行包含一个空气质量监测站点周围 POI 类别的数量信息。这里使用相似度度量的方法来建立空气质量监测站点之间的联系。根据两者之间的余弦相似性来测量它们之间的相似度作为它们之间的边。对于每个空气质量监测站点，计算出其与其他空气质量监测站点之间的余弦相似度。计算公式如（4.2）所示。

$$e_{ij}^{p}=\begin{cases}1, & cos_P(x_i,x_j)\geq 0.8\\ 0, & \text{其他}\end{cases} \tag{4.2}$$

本文取 0.8 阈值[58]，大于 0.8 则连接空气质量监测站点之间的边，这样，就

得到了一个基于余弦相似度的 POI 类别数量的信息网络图。可用$\boldsymbol{G_2(V,E)}$来表示两两空气质量监测站点之间的 POI 类别数量关系。$\boldsymbol{V}$表示的是空气质量监测站点的集合，$\boldsymbol{E}$表示的是任意两个空气质量监测站点之间边e_{ij}^{P}的集合。各空气质量监测站点以有效 POI 数量信息为边组成的信息网络结构如图 4.2 所示。

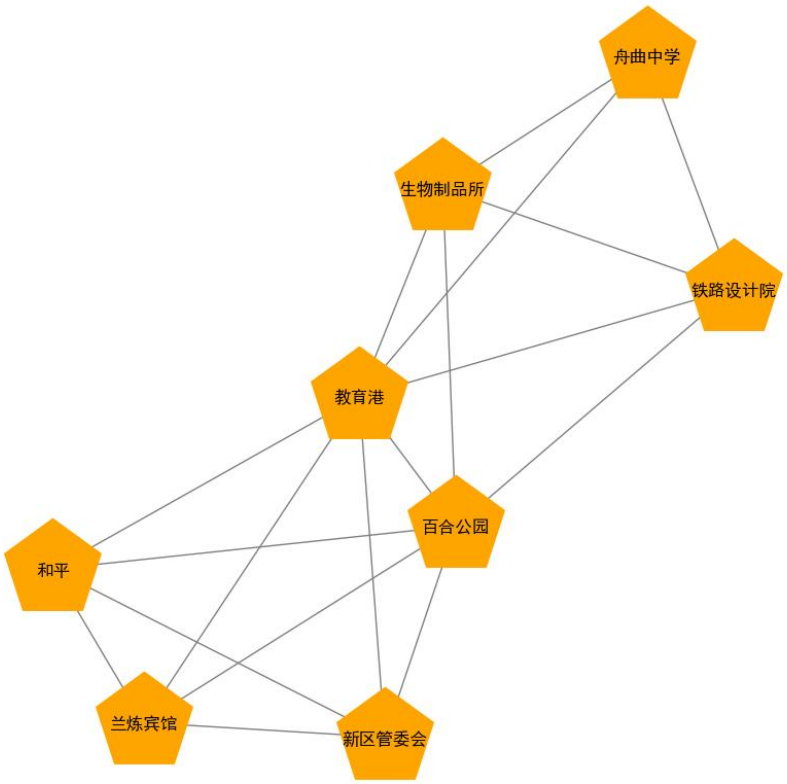


图 4.2 有效 POI 数量为边组成的信息网络图

4.2.3 可达性网络图

空气质量的好坏也会受到交通状况的强烈影响，路网结构的复杂度能反映该道路等级的交通状况，应提取不同的路网类型作为特征，输入到模型中对空气质量进行预测。将空气质量监测站点和不同等级道路之间的特征向量表示转换为矩阵形式。对于每个空气质量监测站点，使用 5 个道路等级来表示其周围的道路状况，具体而言，这个矩阵的每一行包含一个空气质量监测站点周围不同等级道路总长度的信息。这里使用相似度度量的方法来建立空气质量监测站点之间的联系。根据两者之间的余弦相似性来测量它们之间的相似度作为它们之间的边。对于每个空气质量监测站点，计算出其与其他空气质量监测站点之间的余弦相似度。如果两个站点之间的余弦相似度大于某个阈值，则连接它们

之间的边。计算公式如（4.3）所示。

$$e_{ij}^{R}=\begin{cases}1, & cos_{R}(x_i,x_j)\geq 0.8\\ 0, & 其他\end{cases} \tag{4.3}$$

本文取 0.8 阈值[58]，大于 0.8 则连接空气质量监测站点之间的边，这样，就得到了一个基于余弦相似度的不同等级道路的信息网络图。可用$\boldsymbol{G_3(V,E)}$来表示两两空气质量监测站点之间的不同等级的道路长度的关系。$\boldsymbol{V}$表示的是空气质量监测站点的集合，$\boldsymbol{E}$表示的是任意两个空气质量监测站点之间边e_{ij}^{R}的集合。各空气质量监测站点以有效道路为边组成的信息网络结构如图 4.3 所示。

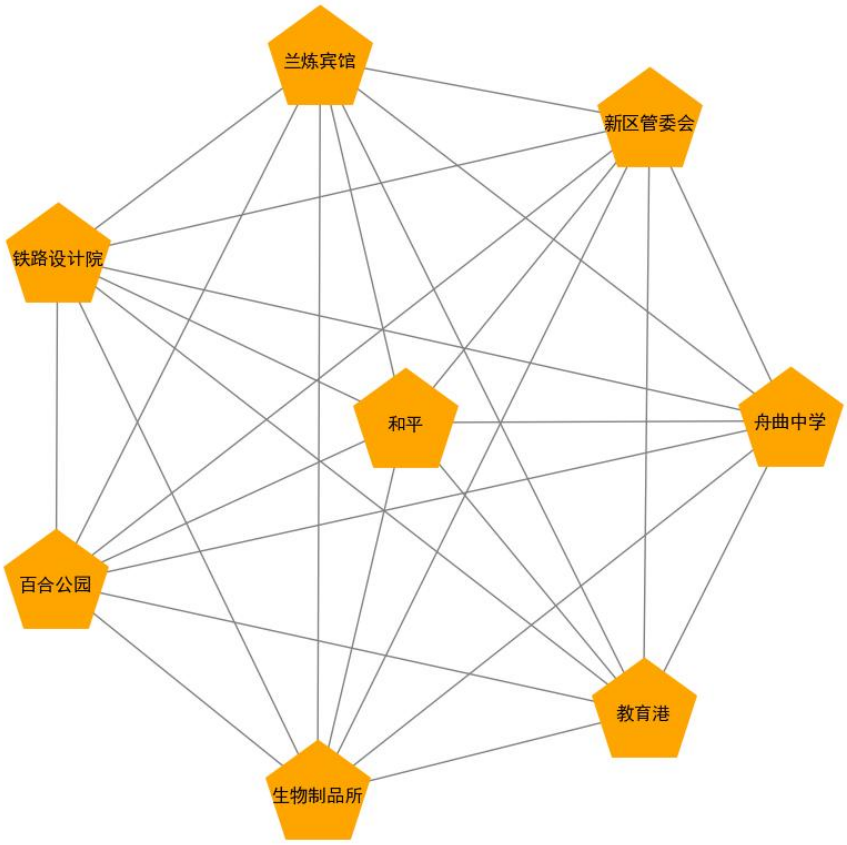


图 4.3 有效道路为边组成的信息网络图

4.2.4 三个网络图的组合

根据上面定义的三张图结构，将得到的以空气质量监测站点之间的有效地理距离、POI 类别数量、路网为边的三个图结构融合到一起，用$\boldsymbol{G}$来表示两两空气质量监测站点之间的关系。具体公式如（4.4）所示。

$$\boldsymbol{G}=\begin{cases}1, & \boldsymbol{G}=\sum_{i=1}^{N}\boldsymbol{G}_i\geq 2\\ 0, & \text{其他}\end{cases} \tag{4.4}$$

其中，$\boldsymbol{G}$的每个节点依旧是每个空气质量监测站点，每个边反映了两个节点之间的连接情况。各空气质量监测站点最终组成的信息网络结构如图 4.4 所示。

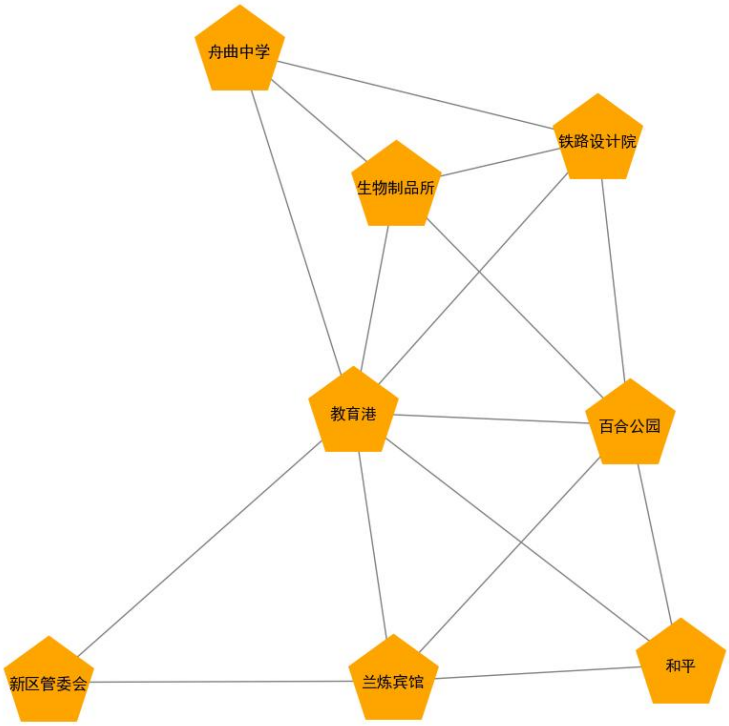


图 4.4 空气质量监测站点之间的信息网络图组合结果

为了提取更多与空气质量相关的非时序特征，将空气质量监测站点之间的信息网络图输入到 GCN 网络中，以提取非时序数据特征，并作为最终表示参与到空气质量的预测中。

4.3 MGCN-GRU 模型框架

为了提高 $PM_{2.5}$ 浓度预测的准确性，本章提出一个基于多视角数据融合的图卷积网络网络框架。通过将不同视角图结构中的信息进行融合，建立多视角图卷积（MGCN-GRU）时空预测模型，旨在更加有效地理解和预测空气质量的变化趋势。MGCN-GRU 模型网络结构如图 4.5 所示。

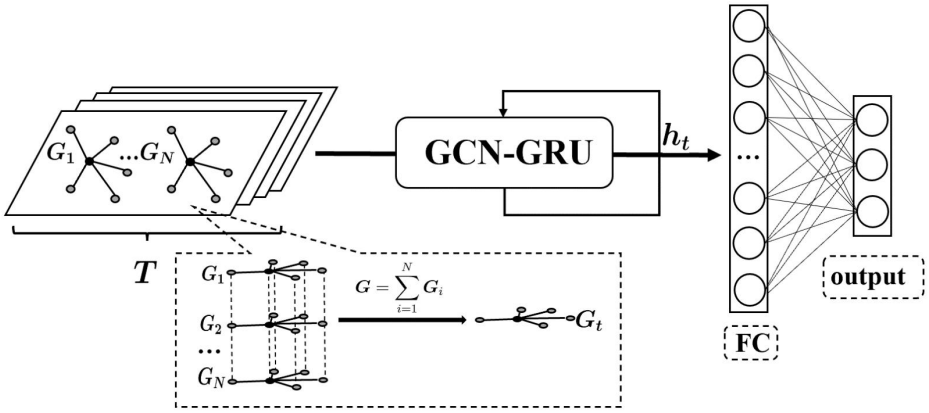


图 4.5 MGCN-GRU 模型网络结构图

图 4.5 为 MGCN-GRU 模型的网络结构，其中，$\boldsymbol{G}_1, \boldsymbol{G}_2, ..., \boldsymbol{G}_N$分别表示的是不同视角下的图结构，$N$是指第$N$个视角，$\boldsymbol{G}_t$为融合之后得到的$t$时刻的图结构。

具体而言，本文的 MGCN-GRU 模型是在基于多视角数据融合方法的基础上结合了图卷积（GCN）模型和门控递归单元（GRU）的优点，即可同时处理历史数据的时间信息与空间信息。该模型主要由以下两部分组成。

（1）基于 GCN 模型提取的非时序特征

对于捕捉空间信息部分而言，MGCN-GRU 模型利用了多视角数据融合方法，使用多图结构来表示不同视角的数据。每个图结构都代表了特定的视角，分别基于空气质量监测站点的地理距离、POI 特征以及路网结构信息作为图的连边权重构建站点之间的信息网络图。具体如式（4.5）所示。

$$\boldsymbol{G} = \sum_{i=1}^{N} \boldsymbol{G}_i \tag{4.5}$$

其中，$\boldsymbol{G}$的每个节点是每个空气质量监测站点，$\boldsymbol{G}_i$分别表示的是基于不同视角数据（即站点经纬度信息、POI 类别数量以及不同道路等级的路网长度）构造的信息网络图，N是指第N个视角。本章节取信息最大化作为空气质量监测站点之间的连边，捕获空气质量监测站点之间的空间依赖性。

（2）基于 GRU 的 $PM_{2.5}$ 浓度预测

为进一步增强模型的能力，以捕捉数据中的时间依赖性，为此，将上面获取到的非时序特征和时序特征的结果相连接，在由 GCN 组成的空间层之后，对输出张量进行整形，作为 GRU 模型的输入信息来预测未来t时刻的 $PM_{2.5}$ 浓度值，将捕捉到的空间信息馈送到 GRU 网络中，达到最终的预测目的。

4.4 实验结果分析

4.4.1 评价指标

本章采用 5 个指标来评估模型的预测性能。分别为平均绝对误差（*MAE*）、均方误差（*MSE*）、均方根误差（*RMSE*）、平均绝对百分比误差（*MAPE*）、决定系数（R^2），这 5 个评价指标计算公式如（4.6）-（4.10）所示：

（1）平均绝对误差（*MAE*）

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \widehat{y_i}| \tag{4.6}$$

（2）均方误差（*MSE*）

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y_i})^2 \tag{4.7}$$

（3）均方根误差（*RMSE*）

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y_i})^2} \tag{4.8}$$

（4）平均绝对百分比误差（*MAPE*）

$$MAPE = \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{y_i - \widehat{y_i}}{y_i}\right| \tag{4.9}$$

（5）决定系数（R^2）

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(\overline{y}_i - y_i)^2} \tag{4.10}$$

其中，$\hat{y}_i$表示预测值，y_i为实际值，$\overline{y}_i$表示实际值y_i的均值，N表示观测的样本数目。

4.4.2 参数设置

MGCN-GRU 模型超参数包括时间步长、学习率、批处理大小、隐藏层单元个数和训练次数，利用 K 折交叉验证的结果，根据每一组参数设置下的模型性能，考虑多个参数一起变化时对模型的综合影响，最终取值如下：时间步长设定为 56，学习率设定为 0.001，批处理大小为 64，GRU 层隐藏层单元个数为 64，经过多次试验后，认为模型迭代到 100 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 100 次。将 80%数据作为训练

集，剩余 20%的数据作为测试集，预测未来兰州市 8 个空气质量监测站点 $PM_{2.5}$ 在 3、6、9、12、15、18 小时浓度变化情况。

将预测结果对参数变化的敏感性在一定的范围内进行了测试。预测结果对参数变化的敏感性变化情况如图 4.6 所示。

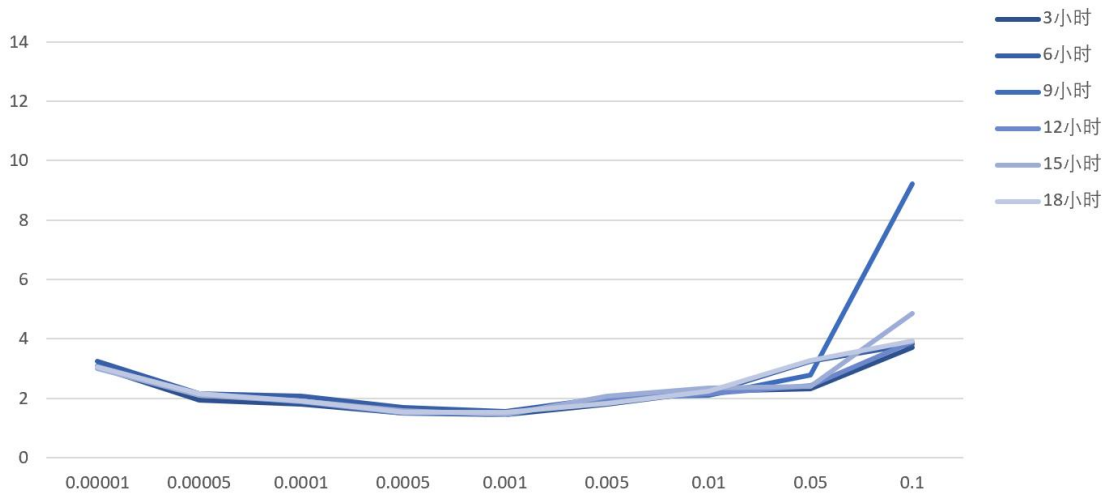


图 4.6 预测结果对参数变化的敏感性变化图

图 4.6 是作出了预测结果对参数变化的敏感性在不同时间段给出的一个指标变化图，由图 4.6 可以看出，模型的预测结果对一定范围内的参数变化并不是很敏感，这说明模型在一定范围内对参数的变化有稳健的预测能力，模型在面对输入的参数的微小变化时能够保持相对一致的预测结果。

4.4.3 实验结果

4.4.3.1. 多模型结果比较

对于现有文献提到的运用了传统的时间序列预测方法，循环神经网络以及图卷积网络分别与长短期记忆力网络、门控循环神经网络对 $PM_{2.5}$ 进行预测等研究，为了验证本文模型的有效性，分别运用 ARMA、LSTM、GRU、GCN-LSTM、GCN-GRU、MGCN-GRU 模型对 $PM_{2.5}$ 浓度值在 3、6、9、12、15、18 小时进行预测。以兰州市兰炼宾馆空气质量监测站点为预测目标，结合空气质量污染物特征和气象特征利用 LSTM 模型和 GRU 模型对该站点的 $PM_{2.5}$ 浓度做出预测，利用 GCN-LSTM 模型、GCN-GRU 模型和 MGCN-GRU 模型将空气质量污染物数据、气象数据、POI 特征和路网数据一起参与到 $PM_{2.5}$ 浓度预测中去，实现了更多数据视角的融合。4 种模型的预测结果如图 4.7 所示。

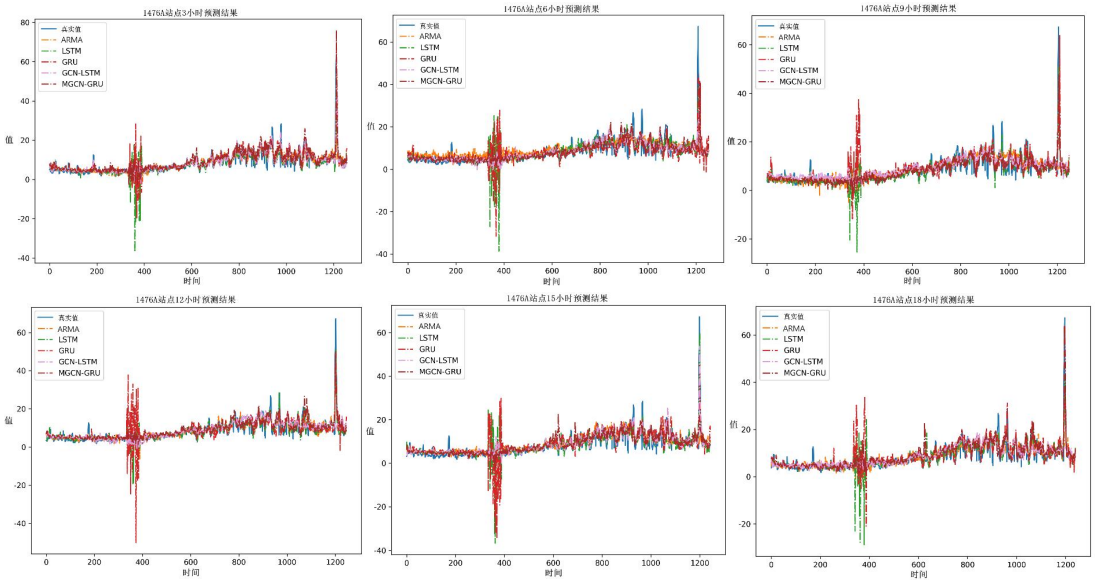


图 4.7 5 个模型在 6 个时间点对 $PM_{2.5}$ 浓度预测效果图

图 4.7 展示了 5 个模型（ARMA、LSTM、GRU、GCN-LSTM、MGCN-GRU）在 3、6、9、12、15、18 小时对 $PM_{2.5}$ 浓度的预测结果。由图 4.7 可以看出，MGCN-GRU 模型的预测效果总体上要优于其他预测模型，该模型的 $PM_{2.5}$ 浓度的预测值与实际值更加贴合。

为了说明本文模型的有效性，将不同的模型分别应用在对 $PM_{2.5}$ 浓度值每隔 3、6、9、12、15、18 小时进行预测。并利用模型性能评价指标 *MAE*、*MSE*、*RMSE*、*MAPE* 和 R^2 分别对这几个小时的 $PM_{2.5}$ 浓度值的预测进行评价。性能评价如表 4.1 所示。

表 4.1 5 个模型在 6 个时间点的预测性能评价

时间	评价指标	ARMA	LSTM	GRU	GCN-LSTM	MGCN-GRU
3 小时	*MAE*	2.0891	1.9616	2.0578	1.7055	**1.5093**
	MSE	12.6604	12.2206	12.5370	9.8559	**6.2160**
	RMSE	3.7098	3.4958	3.5408	3.1394	**2.4932**
	MAPE	36.2772	33.2749	36.2634	24.3940	**23.1463**
	R^2	0.4733	0.4899	0.4767	0.5886	**0.7405**
6 小时	*MAE*	2.4438	2.1888	2.4038	2.0077	**1.6731**
	MSE	19.3232	18.5566	15.2407	15.9654	**6.5995**
	RMSE	4.3372	4.3077	3.9039	3.9957	**2.5690**
	MAPE	39.1924	41.2722	38.5323	27.3748	**25.8994**
	R^2	0.3547	0.2265	0.3649	0.3345	**0.7249**
9 小时	*MAE*	2.2442	1.9362	2.1644	1.9835	**1.6373**
	MSE	15.7636	12.7650	17.4052	14.7378	**5.7646**
	RMSE	4.5527	3.5728	4.1720	3.8390	**2.4010**
	MAPE	38.5224	37.8584	38.5259	**26.7081**	28.4980
	R^2	0.4489	0.4685	0.2752	0.3863	**0.7600**
12 小时	*MAE*	1.9874	1.8214	2.0698	1.9068	**1.4955**
	MSE	15.0082	9.2293	16.0753	14.1855	**4.9059**
	RMSE	3.8911	3.0380	4.0094	3.7664	**2.2149**
	MAPE	36.1979	31.2189	35.7958	27.3687	**25.2047**
	R^2	0.3556	0.6160	0.3311	0.4097	**0.7959**
15 小时	*MAE*	2.1138	**1.6406**	2.2833	2.0937	1.8243
	MSE	9.7656	8.3577	20.8696	15.6907	**6.5636**
	RMSE	5.4433	2.8910	4.5683	3.9612	**2.5620**
	MAPE	42.8722	**28.7436**	41.3596	30.9596	31.1227
	R^2	0.5896	0.6524	0.6321	0.6475	**0.7270**
18 小时	*MAE*	1.6844	1.6604	1.9974	1.7022	**1.6086**
	MSE	11.9823	10.2436	11.7590	7.3134	**5.7640**
	RMSE	3.3341	3.2006	3.4291	2.7043	**2.4008**
	MAPE	33.7781	28.6987	31.5106	26.9365	**26.1975**
	R^2	0.5548	0.5743	0.5113	0.6961	**0.7604**

由表 4.1 的预测性能评价结果可以看出，传统的时间序列模型 ARMA 预测 $PM_{2.5}$ 浓度值时，其效果要稍微差一些，利用长短期记忆网络和门控递归单元对 $PM_{2.5}$ 浓度值进行预测时，其效果要比 ARMA 模型预测效果有些提升；同时考虑其他污染物和气象数据，站点的空间信息和时间信息，利用时空预测模型对 $PM_{2.5}$ 浓度进行预测的效果更佳，而结合了其他污染物、气象、POI、路网这些特征利用 MGCN-GRU 模型来预测 $PM_{2.5}$ 浓度值时，预测效果达到最优，这说明多视角数据融合时空预测模型的预测能力有所提高。 MGCN-GRU 模型的评价标准在 6 个预测时间尺度上基本都达到了最好，说明 MGGC-GRU 模型对 $PM_{2.5}$ 浓度值的预测能力还是不错的，能够对预测效果的提升起到了一定的作用。

4.4.3.2. 消融实验

为了证明多视角数据融合对预测性能的重要影响，进而对城市 $PM_{2.5}$ 浓度预测进行消融实验，通过进行消融实验，逐步去除其他可能的影响因素，仅保留空气质量污染物数据和气象数据，比较模型在完整数据和消融后的数据下的预测性能差异，可以从定量角度说明数据融合的必要性和优势，以及其他不同视角的因素对城市 $PM_{2.5}$ 浓度预测的贡献。

以兰州市兰炼宾馆空气质量监测站点（1476A）为预测目标进行消融实验，即逐步去除城市的 POI 特征、路网结构信息以及站点经纬度信息，分别对 3、6、9、12、15、18 小时的城市 $PM_{2.5}$ 浓度进行预测，预测结果如图 4.8 所示。

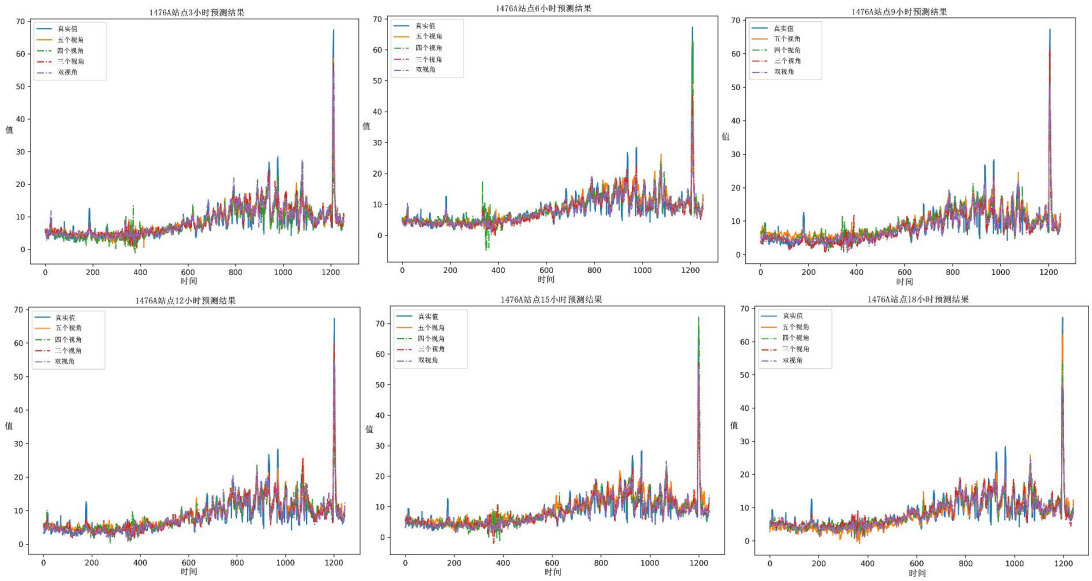


图 4.8 多视角下 $PM_{2.5}$ 浓度在 6 个时间点的预测效果图

图 4.8 展示了多视角下 $PM_{2.5}$ 浓度在 3、6、9、12、15、18 小时的预测效果。

其中，由五个视角，即包含 POI 特征、路网结构信息、空气质量监测站点经纬度信息、空气质量污染物数据和气象数据，以及逐步剔除了前三个视角之后，得到的包含路网结构信息、经纬度信息、空气质量污染物和气象数据的四个视角，包含经纬度信息、空气质量污染物和气象数据的三个视角，以及空气质量污染物和气象数据双视角的预测效果图。由图 4.8 可以看出不断增加多视角对 $PM_{2.5}$ 浓度进行预测，其模型的预测效果越好。

为了证明数据融合对城市 $PM_{2.5}$ 浓度预测起着重要作用，利用模型性能评价指标 *MAE*、*MSE*、*RMSE*、*MAPE* 和 R^2 分别对这几个小时的 $PM_{2.5}$ 浓度值的多视角预测进行评价。性能评价如表 4.2 所示。

表 4.2 多视角下 $PM_{2.5}$ 浓度在 6 个时间点的预测性能评价

时间	评价指标	五个视角	四个视角	三个视角	双视角
3 小时	*MAE*	**1.5093**	1.5744	1.6342	1.6777
	MSE	**6.2160**	7.6639	7.8832	7.8765
	RMSE	**2.4932**	2.7684	2.6452	2.7991
	MAPE	**23.1463**	23.8485	23.9484	24.5112
	R^2	**0.7405**	0.6801	0.6242	0.6211
6 小时	*MAE*	1.6731	**1.6567**	1.7121	1.7334
	MSE	**6.5995**	6.6509	6.7798	6.7833
	RMSE	**2.5690**	2.5789	2.5993	2.6221
	MAPE	**25.8994**	27.2646	26.9242	27.3557
	R^2	**0.7249**	0.7228	0.7029	0.6889
9 小时	*MAE*	1.6373	1.6537	**1.6244**	1.7136
	MSE	**5.7646**	5.8921	5.9566	6.1187
	RMSE	**2.4010**	2.4533	2.5182	2.5461
	MAPE	28.4980	**28.0867**	29.2764	29.1221
	R^2	**0.7600**	0.7219	0.7009	0.6893
12 小时	*MAE*	**1.4955**	1.5801	1.5972	1.6243
	MSE	**4.9059**	5.9352	5.8963	6.3434
	RMSE	**2.2149**	2.4362	2.4859	2.5213
	MAPE	25.2047	**24.1622**	25.9871	26.7734
	R^2	**0.7959**	0.7532	0.7366	0.6679
15 小时	*MAE*	**1.8243**	1.9314	2.7322	2.5047
	MSE	**6.5636**	6.6905	6.9253	7.7811
	RMSE	**2.5620**	2.8717	5.4424	3.3986
	MAPE	31.1227	32.8690	**28.7322**	30.6976
	R^2	**0.7270**	0.6928	0.7070	0.6454
18 小时	*MAE*	1.6086	**1.5584**	1.6236	1.6771
	MSE	**5.7640**	6.1099	6.5630	6.5544
	RMSE	**2.4008**	2.4718	2.5127	2.5245
	MAPE	26.1975	**24.0147**	26.2707	27.0121
	R^2	**0.7604**	0.7461	0.7220	0.7219

由表 4.2 的预测性能评价结果可以看出，通过对多个视角的数据进行消融实验，我们发现随着视角的增加，预测性能不断提升。当我们只使用空气质量数据和气象数据进行预测时，模型的性能较为有限。然而，随着 POI 特征、路网数据和站点经纬度信息的引入，我们观察到预测性能进一步提升。这说明多个视角的数据融合能够更全面、准确地捕捉到影响城市 $PM_{2.5}$ 浓度的各种因素。因此，数据融合在城市 $PM_{2.5}$ 浓度预测中具有重要的优势，可以帮助我们更好地理解和预测空气质量，以支持决策和改善环境质量。

4.4.4 实验设置

本次实验所采用的电脑硬件配置为：14Vcpu Intel® Xeon® Gold 6330 CPU、显卡为 RTX A5000(24GB)的独立显卡。

实验所用的软件配置：在 Windows11（64 位）操作系统下进行实验。数据预处理、模型的搭建及训练使用 Python（版本 3.8）。

本文编写的所有模型代码均基于深度学习的 Pytorch 框架，该框架具有易扩展、易实现的特性，能够快速进行实验编码。

4.5 本章小结

本章提出了考虑到时空特性的 MGCN-GRU 网络框架，并利用该框架对甘肃省兰州市的 $PM_{2.5}$ 浓度值做出预测，其中，对该网络框架结构进行了详细的介绍，对时空特性问题和网络框架结构创建的想法做出详细的说明，最后通过实验证明了本文提出的 MGCN-GRU 网络框架的有效性以及证明数据融合对城市 $PM_{2.5}$ 浓度预测起着重要作用。

5 基于多视角图注意力网络的$PM_{2.5}$浓度预测

5.1 问题描述

前文也提到，对$PM_{2.5}$浓度序列进行预测不仅仅考虑其他污染物因素对其影响，还应该考虑到气象因素、POI 特征以及道路状况对其产生的影响，上一章节利用了图卷积网络捕捉不同空气质量监测站点之间的空间相关性，但是站点之间的连边却没有考虑各种因素对站点附近监测到的$PM_{2.5}$浓度变化的影响程度，为此，引入空间注意力机制，来捕捉不同空气质量监测站点之间的动态空间相关性，从而把握不同特征对$PM_{2.5}$浓度序列的影响程度。

空气质量站点在空间上表现为一种拓扑图结构，而图注意力网络可以很好地处理这类图结构数据，并且在处理图数据过程中还能合理的考虑到节点之间的权重信息，因此本文考虑借鉴图注意力网络的思想来提取数据中存在的空间信息。并利用 LSTM 模型提取空气质量监测站点的时间信息。

5.2 注意力图的构建

在应用图注意力网络之前，需要构建两个数据结构，包含了各空气质量监测站点之间的关联关系，也就是需要构建节点原始特征矩阵和节点的邻接矩阵，以便于模型的训练。

为获取各空气质量监测站点之间的关联关系，根据第三章对$PM_{2.5}$浓度序列与 POI、路网结构的相关性分析可以看出，像 POI 和路网结构这两类非时序数据通常也会对空气质量造成不同程度的影响，因此，可以将非时序数据作为空气质量预测的辅助信息，在对 POI 和路网数据进行相关特征提取时，将每个区域内不同类别的 POI 数量和不同道路等级的路网长度作为非时序数据的相关特征，并且讨论 POI 和路网不同类别之间的层次信息[3]。

因此，为了更好地获取隐藏的相关特征，本文将所有要测量的空气质量监测站点的坐标、 POI 与道路网络、空气质量监测站点之间的距离、 POI 类别的数量以及不同道路等级的道路总长度等作为连接空气质量监测站点边的权重，以权重作为各个空气质量监测站点的邻接矩阵，将其和空气质量监测站点的特征矩阵，组成了不同权重的空气质量监测站点的信息网络结构图，将每个信息网络结构图拼接成一个具有三个权重的最终空气质量监测站点的信息网络图，

并输入到 GAT 模型中，进一步提取潜在的局部特征，最后将获得的非时间序列特征作为 $PM_{2.5}$ 浓度预测的辅助信息，从而提高 $PM_{2.5}$ 浓度的预测效果。

接下来将介绍由空气质量监测站点的坐标，POI 以及路网结构这类非时序数据获取到的不同空气质量监测站点之间连边权重组成的信息网络图。下文展示了甘肃省兰州市的 8 个空气质量监测站点的信息网络图结构。

5.2.1 地理距离网络图

空气质量监测站点之间的地理距离越近，相关性越强。可用$\boldsymbol{G_1(V,E)}$来表示两两空气质量监测站点之间的距离关系。$\boldsymbol{V}$表示的是空气质量监测站点的集合，$\boldsymbol{E}$表示的是任意两个空气质量监测站点之间边$\boldsymbol{e_{ij}}$的集合，用$\boldsymbol{W_{1ij}}$来表示边$\boldsymbol{e_{ij}}$的权重，这里也就是$\boldsymbol{e_i}$和$\boldsymbol{e_j}$两空气质量监测站点之间的距离。各空气质量监测站点以距离为边权重组成的信息网络结构如图 5.1 所示。

图 5.1 以距离为边权重组成的信息网络图

图 5.1 是利用了第二章提到的地理空间距离计算公式（2.5）计算出每个空气质量监测站点之间的地理距离，将空气质量监测站点之间的距离作为其边的权重构建出空气质量监测站点的信息网络结构图。

5.2.2 兴趣点网络图

可用$G_2(V,E)$来表示两两空气质量监测站点之间的 POI 类别数量关系。V表示的是空气质量监测站点的集合，E表示的是任意两个空气质量监测站点之间边e_{ij}的集合，用W_{2ij}来表示边e_{ij}的权重，这里也就是e_i和e_j两空气质量监测站点之间的 POI 不同类别的数量。各空气质量监测站点以不同类别的 POI 数量信息为边权重组成的信息网络结构如图 5.2 所示。

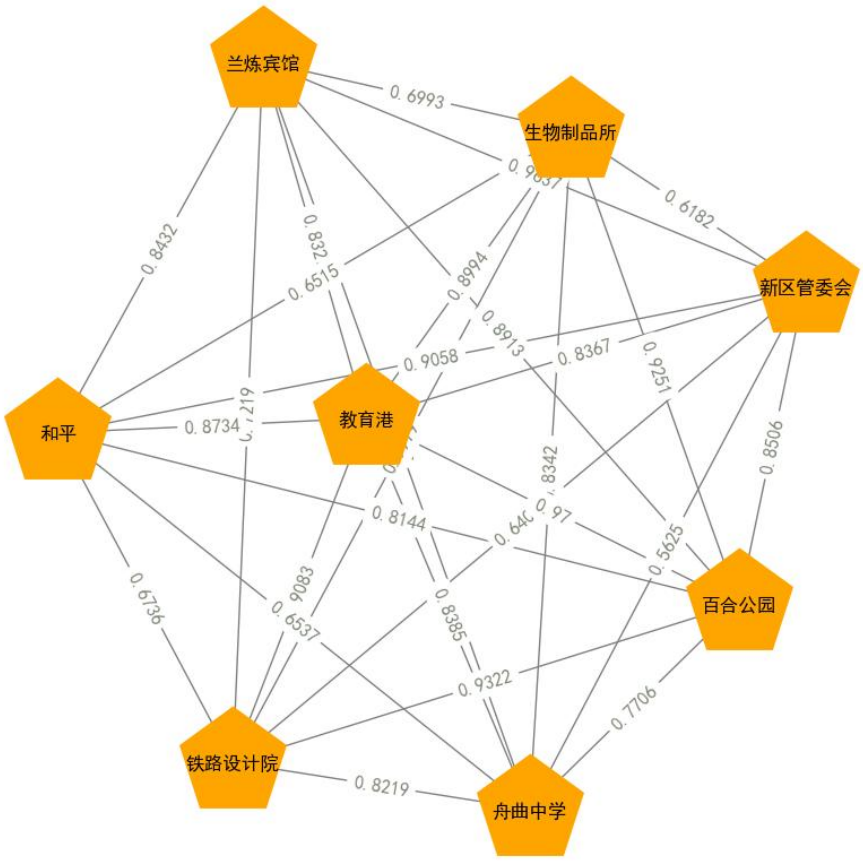


图 5.2 以不同类别的 POI 数量为边权重组成的信息网络图

将空气质量监测站点和 POI 类别之间的特征向量表示转换为矩阵形式。对于每个空气质量监测站点，使用 9 个 POI 类别数量来表示其周围的 POI，具体而言，这个矩阵的每一行包含一个空气质量监测站点周围 POI 类别的数量信息。这里使用相似度度量的方法来建立空气质量监测站点之间的联系。根据两者之间的余弦相似性来测量它们之间的相似度作为它们之间的边。对于每个空气质量监测站点，计算出其与其他空气质量监测站点之间的余弦相似度。这样，就

得到了一个基于余弦相似度的 POI 类别数量的信息网络图。

5.2.3 可达性网络图

可用$G_3(V,E)$来表示两两空气质量监测站点之间的不同等级的道路长度的关系。V表示的是空气质量监测站点的集合，E表示的是任意两个空气质量监测站点之间边e_{ij}的集合，用W_{3ij}来表示边e_{ij}的权重，这里也就是e_i和e_j两空气质量监测站点之间的不同等级道路的总长度。各空气质量监测站点以不同道路等级的总长度为边权重组成的信息网络结构如图 5.3 所示。

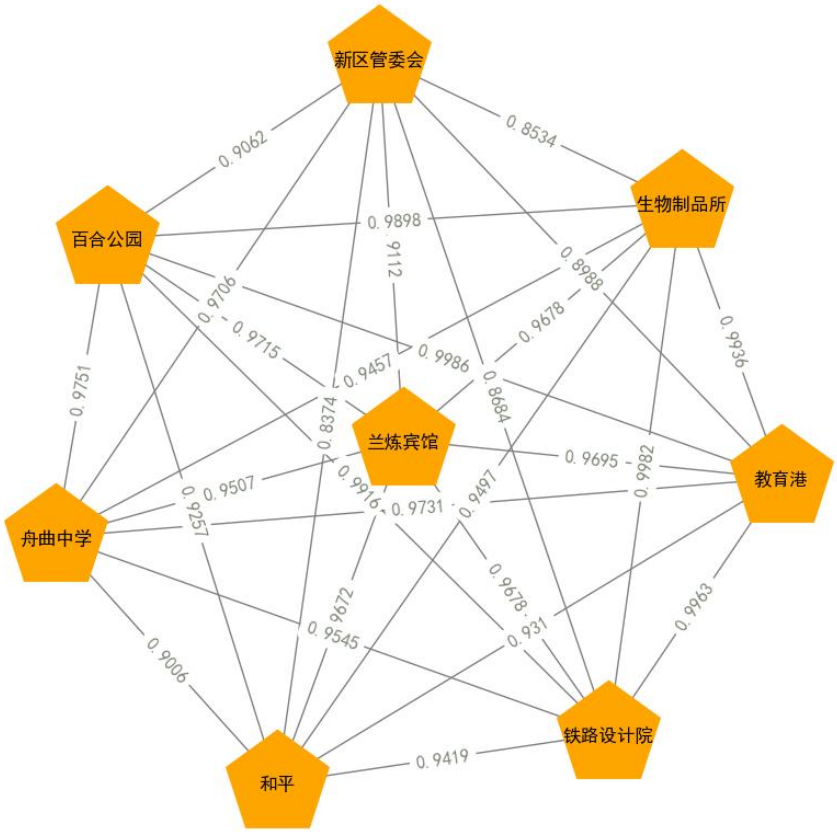


图 5.3 以不同等级道路总长度为边权重组成的信息网络图

将空气质量监测站点和不同等级道路之间的特征向量表示转换为矩阵形式。对于每个空气质量监测站点，使用 5 个道路等级来表示其周围的道路状况，具体而言，这个矩阵的每一行包含一个空气质量监测站点周围不同等级道路总长度的信息。这里使用相似度度量的方法来建立空气质量监测站点之间的联系。根据两者之间的余弦相似性来测量它们之间的相似度作为它们之间的边。对于每个空气质量监测站点，计算出其与其他空气质量监测站点之间的余弦相似度。

这样，就得到了一个基于余弦相似度的不同等级道路的信息网络图。

5.2.4 三个网络图的组合

根据上面定义的三张图结构，将得到的以空气质量监测站点之间的距离、POI 类别数量、不同等级道路总长度为边权重的三个图结构融合到一起，用$\boldsymbol{G}$来表示两两空气质量监测站点之间的关系。具体公式如（5.1）-（5.2）所示。

$$\boldsymbol{G} = \boldsymbol{G_1} + \boldsymbol{G_2} + \boldsymbol{G_3} \tag{5.1}$$

$$\boldsymbol{W_{ij}} = \boldsymbol{W_{1ij}} + \boldsymbol{W_{2ij}} + \boldsymbol{W_{3ij}} \tag{5.2}$$

其中，$\boldsymbol{G}$的每个节点依旧是每个空气质量监测站点，每个边反映了两个节点之间的连接情况。其边的权重则是$\boldsymbol{W_{1ij}}$、$\boldsymbol{W_{2ij}}$、$\boldsymbol{W_{3ij}}$的和。各空气质量监测站点最终组成的信息网络结构如图 5.4 所示。

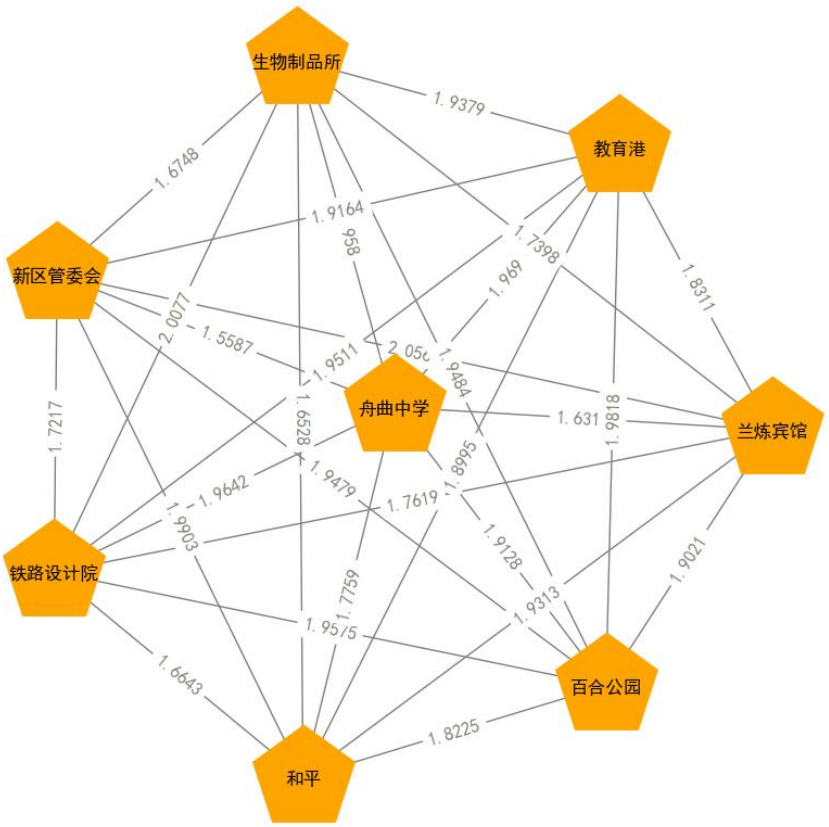


图 5.4 空气质量监测站点之间的信息网络图组合结果

由于三张图之间边的权重值存在量纲上的差异，在构造空气质量监测站点之间最后的信息网络图之前，将其权重值进行了归一化处理之后，再将三张图之间边的权重值进行加和，得到最终的两两空气质量监测站点之间边的权重。

为了提取更多与空气质量相关的非时序特征，将空气质量监测站点之间的信息网络图输入到 GAT 网络中，以提取非时序数据的潜在特征，并作为非时序数据的最终表示参与到空气质量的预测中。

5.3 MGATs-LSTM 模型框架

为了提高 $PM_{2.5}$ 浓度预测的准确性，本章提出一个基于多视角数据融合的多层图注意力机制网络框架。通过将不同视角图结构中的信息进行融合，建立多视角图注意力（MGATs-LSTM）时空预测模型，旨在更加有效地理解和预测空气质量的变化趋势。MGATs-LSTM 模型网络结构如图 5.5 所示。

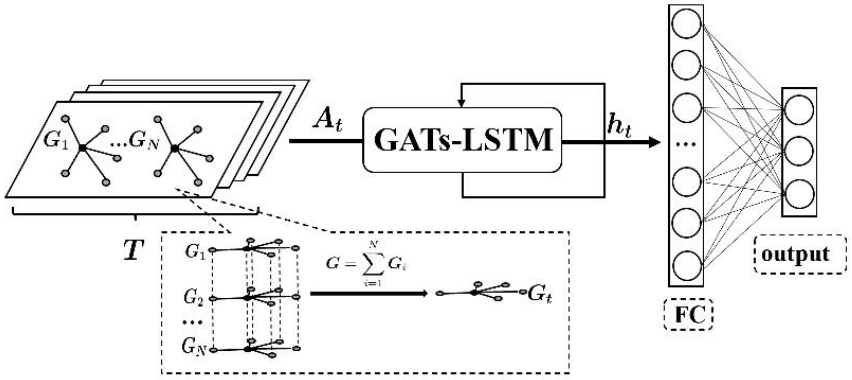


图 5.5 MGATs-LSTM 模型网络结构图

图 5.5 为 MGATs-LSTM 模型的网络结构，其中，$\boldsymbol{G}_1, \boldsymbol{G}_2, ..., \boldsymbol{G}_N$ 分别表示的是不同视角下的图结构，N 是指第 N 个视角，$\boldsymbol{G}_t$ 为融合之后得到的 t 时刻的图结构，$\boldsymbol{A}_t$ 为邻接矩阵。

具体而言，本文的 MGATs-LSTM 模型是在基于多视角数据融合方法的基础上结合了 Dogan Aykas 改进的双层图注意力（GATs）模型[5]和长短期记忆（LSTM）模型的优点，即可同时处理历史数据的时间信息与空间信息。该模型主要由以下两部分组成。

（1）基于 GAT 模型提取的非时序特征

对于捕捉空间信息部分而言，MGATs-LSTM 模型利用了多视角数据融合方法，使用多图结构来表示不同视角的数据。每个图结构都代表了特定的视角，分别基于空气质量监测站点的地理距离、POI 特征以及路网结构信息作为图的连边权重构建站点之间的信息网络图。具体如式（5.3）所示。

$$\boldsymbol{G} = \sum_{i=1}^{N} \boldsymbol{G}_i \tag{5.3}$$

其中，$\boldsymbol{G}$的每个节点是每个空气质量监测站点，$\boldsymbol{G}_i$分别表示的是基于不同视角数据（即站点经纬度信息、POI 类别数量以及不同道路等级的路网长度）构造的信息网络图，N是指第N个视角。模型中的两层 GAT 用于捕获空气质量监测站点之间的空间依赖性，并赋予不同的空间注意力权重给各个空气质量监测站点之间的边，表示空气质量监测站点对目标空气质量监测站点的影响程度。即$(a_t^1, a_t^2, ..., a_t^N)$,其中，$a_t^j$表示在$t$时刻空气质量监测站点$j$对于目标空气质量监测站点的影响程度。为了实现数据融合，利用图注意力（GAT）模型的注意力机制捕捉图节点的依赖关系，根据节点特征和关系来计算注意力权重，在消息传递过程中聚焦于相关信息，所得的注意力得分α_{ij}计算如式（5.4）所示。

$$\alpha_{ij} = \frac{exp(LeakyReLU(\overrightarrow{a}^T[\boldsymbol{W}\overrightarrow{h}_i \| \boldsymbol{W}\overrightarrow{h}_j]))}{\sum_{K\in\mathcal{N}_i} exp(LeakyReLU(\overrightarrow{a}^T[\boldsymbol{W}\overrightarrow{h}_i \| \boldsymbol{W}\overrightarrow{h}_k]))} \tag{5.4}$$

式（31）中$[\boldsymbol{W}\cdot h_i \| \boldsymbol{W}\cdot h_j] \in \mathbb{R}^{2d'}$表示第$i$个和第$j$个顶点对的特征被转换为$d'$维并进行拼接操作，其中，$h_i \in \mathbb{R}^d$和$h_j \in \mathbb{R}^d$分别指第$i$个和第$j$个顶点的特征向量。$\boldsymbol{W} \in \mathbb{R}^{d'\times d}$是对每个顶点的特征进行线性变换的共享权矩阵。

模型中采用了多层图注意力机制，在二维 GAT 层中，节点可以具有二维特征，即特征变量和时间步长。这里扩展了 GAT 模型，使其可以对二维特征进行操作，并计算每个特征变量的注意力得分。自动学习每个图结构中节点之间的重要关系，并分配适当的注意力权重，可以聚焦于融合后的结果为最重要的节点信息。在每一层的图注意力机制中，通过对节点和边进行消息传递和信息汇聚来实现信息融合。通过迭代这个过程，网络逐渐融合了多个图结构中的信息，并生成了综合的时空数据表示。具体如式（5.5）所示。

$$\alpha_{ij} = \sum_{n\in N} \alpha_{ij}^n \tag{5.5}$$

其中，α_{ij}是指不同图结构中的第i个和第j个站点之间的注意力得分，N是指第N个视角，α_{ij}^n是指第n个视角的第i个和第j个站点之间的注意力得分。并应用在多头注意力机制上，并将结果连接起来，获得节点的最终表示。具体如式（5.6）[5]所示。

$$\hat{h}_i = ||_{k=1}^{K}\sigma(\sum_{j\in\mathcal{N}} \boldsymbol{W}^k \overrightarrow{h}_j[diag(\alpha_{ij}^1,...,\alpha_{ij}^N)]^k) \tag{5.6}$$

其中，$\boldsymbol{W}$大小为$\mathbb{R}^{t'\times t}$，权重矩阵$\boldsymbol{W}$用于变换时间步长维度，h_i大小为$\mathbb{R}^{t\times d}$，将获得的$\hat{h}_i$和图中的邻接矩阵相结合，则第i个节点的最终表示具体如式（5.7）[5]所示。

$$\begin{cases} \hat{\boldsymbol{D}}_{ii} = \sum_j \hat{\boldsymbol{A}}_{ij}, \\ \widetilde{h}_i = \hat{h}_i(\hat{\boldsymbol{D}}^{-\frac{1}{2}}\hat{\boldsymbol{A}}\hat{\boldsymbol{D}}^{-\frac{1}{2}}) \end{cases} \tag{5.7}$$

其中，$\boldsymbol{D}$为度矩阵。$\widetilde{h}_i$为第i个节点的最终表示，即所提取到特征的空间信息。

（2）基于 LSTM 模型的 $PM_{2.5}$ 浓度预测

为进一步增强模型的能力，以捕捉数据中的时间依赖性，为此，将上面两部分分别获取到的时序特征和非时序特征的结果相连接，在由两个不同 GAT 层组成的空间层之后，对输出张量进行整形，作为 LSTM 模型的输入信息来预测未来t时刻的 $PM_{2.5}$ 浓度值，将捕捉到的空间信息馈送到由单个递归层组成的 LSTM 网络中，达到最终的预测目的。模型结构如图 5.6 所示。

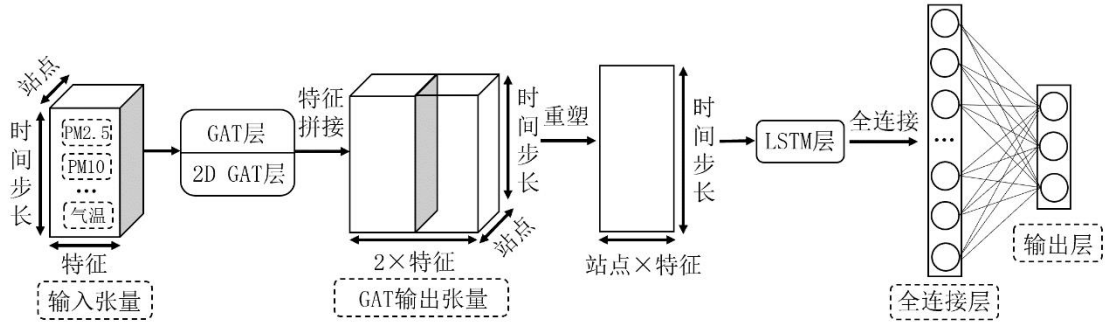


图 5.6 时序特征提取结构图

5.4 实验结果分析

5.4.1 评价指标

本章采用 5 个指标来评估模型的预测性能。分别为平均绝对误差（*MAE*）、均方误差（*MSE*）、均方根误差（*RMSE*）、平均绝对百分比误差（*MAPE*）、决定系数（R^2），这 5 个评价指标计算公式如（5.8）-（5.12）所示：

（1）平均绝对误差（*MAE*）

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \widehat{y_i}| \tag{5.8}$$

（2）均方误差（*MSE*）

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\ y_i - \widehat{y_i})^2 \tag{5.9}$$

（3）均方根误差（*RMSE*）

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\ y_i - \widehat{y_i})^2} \tag{5.10}$$

（4）平均绝对百分比误差（*MAPE*）

$$MAPE = \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{y_i - \widehat{y_i}}{y_i}\right| \tag{5.11}$$

（5）决定系数（R^2）

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(\overline{y}_i - y_i)^2} \tag{5.12}$$

其中，$\hat{y}_i$表示预测值，y_i为实际值，$\overline{y}_i$表示实际值y_i的均值，N表示观测的样本数目。

5.4.2 参数设置

MGATs-LSTM 模型中所涉及到的超参数包括时间步长、学习率、批处理大小、隐藏层单元个数和训练次数，利用 K 折交叉验证的结果，根据每一组参数设置下的模型性能，考虑多个参数一起变化时对模型的综合影响，最终取值如下：时间步长设定为 56，学习率设定为 0.001，批处理大小为 64，LSTM 层隐藏层单元个数为 64，经过反复试验后，发现模型迭代到 100 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 100 次。将 80%数据作为训练集，剩余 20%的数据作为测试集，预测未来兰州市 8 个空气质量监测站点 $PM_{2.5}$ 在 3、6、9、12、15、18 小时浓度变化情况。并将预测结果对参数变化的敏感性在一定的范围内进行了测试。预测结果对参数变化的敏感性变化情况如图 5.7 所示。

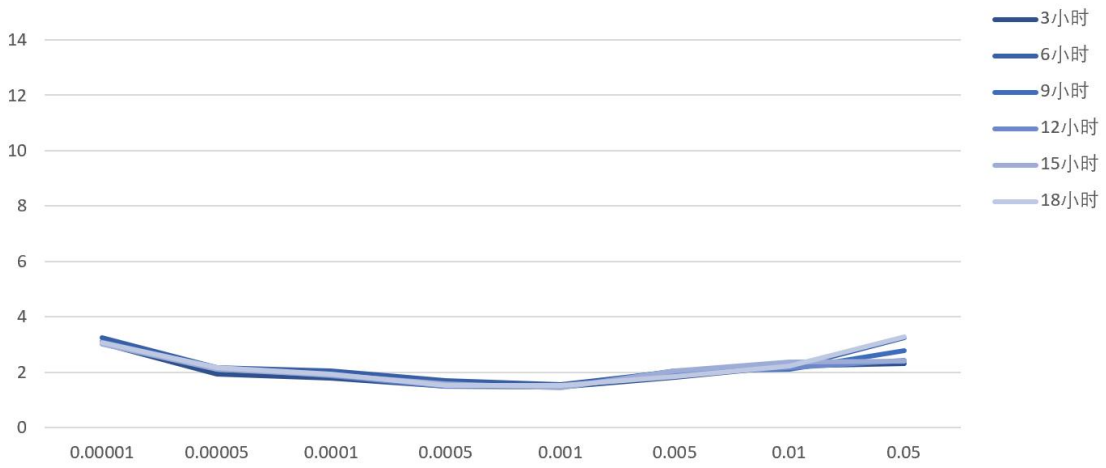


图 5.7 预测结果对参数变化的敏感性变化图

图 5.7 是作出了预测结果对参数变化的敏感性在不同时间段给出的一个指标变化图，由图 5.7 可以看出，模型的预测结果对参数的变化不敏感，这说明模型在一定范围内对参数的变化有稳健的预测能力，模型在面对输入的参数的微小变化时能够保持相对一致的预测结果。

5.4.3 实验结果

5.4.3.1. 多模型结果比较

对于现有文献提到的运用了图注意力网络，分别与长短期记忆力网络、门控循环神经网络对 $PM_{2.5}$ 进行预测问题，为了验证本文模型的有效性，分别运用上一章节的 MGCN-GRU 模型和 GAT-LSTM、GAT-GRU、MGATs-GRU、MGATs-LSTM 模型对 $PM_{2.5}$ 浓度值在 3、6、9、12、15、18 小时进行预测。以兰州市兰炼宾馆空气质量监测站点为预测目标，结合空气质量污染物特征和气象特征利用 GAT-GRU 模型和 GAT-LSTM 模型对该站点的 $PM_{2.5}$ 浓度做出预测，利用 MGATs-GRU 模型和 MGATs-LSTM 模型将空气质量污染物数据、气象数据、POI 特征和路网数据一起参与到 $PM_{2.5}$ 浓度预测中去，实现了更多数据视角的融合。5 种模型的预测结果如图 5.8 所示。

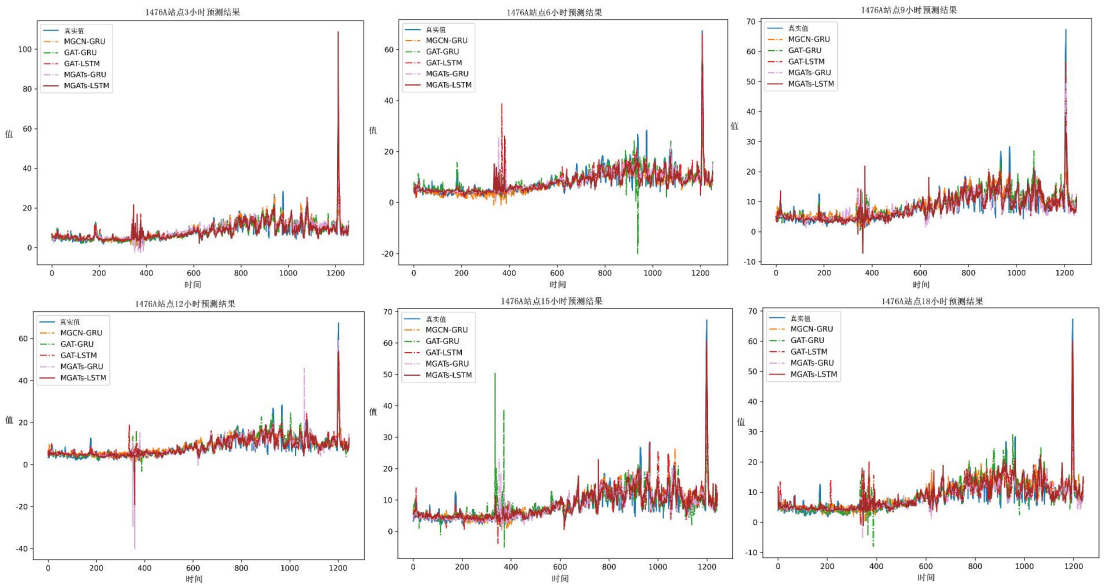


图 5.8 5 个模型在 6 个时间点对 $PM_{2.5}$ 浓度预测效果图

图 5.8 展示了 5 个模型（MGCN-GRU、GAT-LSTM、GAT-GRU、MGATs-GRU、MGATs-LSTM）在 3、6、9、12、15、18 小时对 $PM_{2.5}$ 浓度的预测结果。由图 5.8 可以看出，MGATs-LSTM 模型的预测效果要优于其他 4 个时空预测模型，该模型的 $PM_{2.5}$ 浓度的预测值与实际值更加贴合。

为了说明本文模型的有效性，将不同的模型分别应用在对 $PM_{2.5}$ 浓度值每隔 3、6、9、12、15、18 小时进行预测。并利用模型性能评价指标 *MAE*、*MSE*、*RMSE*、*MAPE* 和 R^2 分别对这几个小时的 $PM_{2.5}$ 浓度值的预测进行评价。性能评价如表 5.1 所示。

表 5.1 5 个模型在 6 个时间点的预测性能评价

时间	评价指标	MGCN-GRU	GAT-GRU	GAT-LSTM	MGATs-GRU	MGATs-LSTM
3 小时	*MAE*	1.5093	1.6409	1.5486	1.7834	**1.4528**
	MSE	6.2160	7.6127	5.7688	9.7010	**5.1512**
	RMSE	2.4932	2.7591	2.4018	3.1146	**2.2696**
	MAPE	23.1463	26.9932	24.1326	**20.3922**	22.7661
	R^2	0.7405	0.6822	0.7592	0.5950	**0.7850**
6 小时	*MAE*	1.6731	1.8364	1.6202	1.6739	**1.5624**
	MSE	6.5995	11.1908	**5.7792**	6.2133	6.0098
	RMSE	2.5690	3.3453	**2.4040**	2.4926	2.4515
	MAPE	25.8994	32.0257	25.7395	26.5853	**25.0356**
	R^2	0.7249	0.5335	**0.7591**	0.7410	0.7495
9 小时	*MAE*	1.6373	1.7689	1.5963	1.7030	**1.4583**
	MSE	5.7646	9.6213	6.0246	6.4866	**4.7396**
	RMSE	2.4010	3.1018	2.4545	2.5469	**2.1771**
	MAPE	28.4980	27.4700	25.2021	26.9598	**22.7397**
	R^2	0.7600	0.5994	0.7491	0.7299	**0.8026**
12 小时	*MAE*	1.4955	1.6027	1.5190	1.6376	**1.4964**
	MSE	4.9059	7.3620	5.6791	7.7793	**5.4456**
	RMSE	2.2149	2.7133	2.3831	2.7891	**2.3336**
	MAPE	25.2047	24.3303	**23.3003**	28.6409	23.5332
	R^2	0.7959	0.6937	0.7637	0.6763	**0.7734**
15 小时	*MAE*	1.8243	1.7425	1.4922	1.4859	**1.4488**
	MSE	6.5636	8.6912	5.3040	4.9111	**4.6149**
	RMSE	2.5620	2.9481	2.3030	2.2161	**2.1482**
	MAPE	31.1227	24.7438	23.8813	23.3547	**22.9243**
	R^2	0.7270	0.6386	0.7794	0.7958	**0.8080**
18 小时	*MAE*	1.6086	1.7696	1.6298	1.6779	**1.5229**
	MSE	5.7640	8.2209	7.0940	7.3862	**5.4799**
	RMSE	2.4008	2.8672	2.6635	2.7178	**2.3409**
	MAPE	26.1975	27.5867	25.2426	27.1031	**23.5802**
	R^2	0.7604	0.6583	0.7052	0.6930	**0.7722**

由表 5.1 可以看出，前两个组合模型是结合其他污染物与气象特征来预测 $PM_{2.5}$ 浓度值，其效果要稍微差一些，结合了其他污染物、气象、POI、路网这些特征利用 MGATs-GRU 模型和 MGATs-LSTM 模型来预测 $PM_{2.5}$ 浓度值，其效果要比前两个模型预测效果有显著提升。这说明多视角数据融合时空预测模型的预测能力有所提高，本文提出的模型在网络结构上采用在 LSTM 模型的基础上进行改进。MGATs-LSTM 模型的评价标准在 6 个预测时间尺度上基本都达到了最好，MGATs-LSTM 模型对 $PM_{2.5}$ 浓度值的预测能力还是不错的，能够对预测效果的提升起到了一定的作用。

5.4.3.2. 消融实验

为了证明多视角数据融合对预测性能的重要影响，进而对城市 $PM_{2.5}$ 浓度预测进行消融实验，通过进行消融实验，逐步去除其他可能的影响因素，仅保留空气质量污染物数据和气象数据，比较模型在完整数据和消融后的数据下的预测性能差异，可以从定量角度说明数据融合的必要性和优势，以及其他不同视角的因素对城市 $PM_{2.5}$ 浓度预测的贡献。

以兰州市兰炼宾馆空气质量监测站点为预测目标进行消融实验，即逐步去除城市的 POI 特征、路网结构信息以及站点经纬度信息，分别对 3、6、9、12、15、18 小时的城市 $PM_{2.5}$ 浓度进行预测，预测结果如图 5.9 所示。

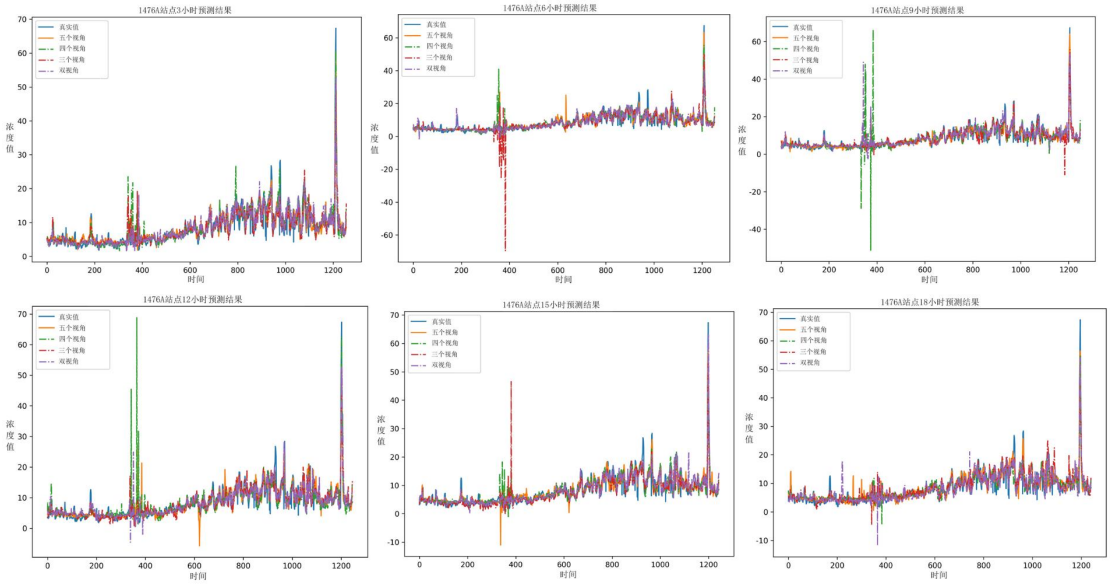


图 5.9 多视角下 $PM_{2.5}$ 浓度在 6 个时间点的预测效果图

图 5.9 展示了多视角下 $PM_{2.5}$ 浓度在 3、6、9、12、15、18 小时的预测效果。

其中，由五个视角，即包含 POI 特征、路网结构信息、空气质量监测站点经纬度信息、空气质量污染物数据和气象数据，以及逐步剔除了前三个视角之后，得到的包含路网结构信息、经纬度信息、空气质量污染物和气象数据的四个视角，包含经纬度信息、空气质量污染物和气象数据的三个视角，以及空气质量污染物和气象数据双视角的预测效果图。由图 5.9 可以看出不断增加多视角对$PM_{2.5}$浓度进行预测，其模型的预测效果越好。

为了证明数据融合对城市$PM_{2.5}$浓度预测起着重要作用，利用模型性能评价指标 *MAE*、*MSE*、*RMSE*、*MAPE* 和 R^2 分别对这几个小时的$PM_{2.5}$浓度值的多视角预测进行评价。性能评价如表 5.2 所示。

表 5.2 多视角下 $PM_{2.5}$ 浓度在 6 个时间点的预测性能评价

时间	评价指标	五个视角	四个视角	三个视角	双视角
3 小时	*MAE*	**1.4528**	1.5346	1.4849	1.5374
	MSE	**5.1512**	6.1503	5.4932	6.9905
	RMSE	**2.2696**	2.4800	2.3438	2.6440
	MAPE	**22.7661**	25.0075	24.0947	23.6360
	R^2	**0.7850**	0.7433	0.7707	0.7082
6 小时	*MAE*	**1.5624**	1.6846	1.7499	1.6136
	MSE	**6.0098**	7.4926	10.6884	7.1870
	RMSE	**2.4515**	2.7373	3.2693	2.6808
	MAPE	25.0356	26.9529	29.8520	**24.1837**
	R^2	**0.7495**	0.6877	0.5544	0.7004
9 小时	*MAE*	**1.4583**	1.7938	1.5266	1.5358
	MSE	**4.7396**	14.3181	6.1633	6.0798
	RMSE	**2.1771**	3.7839	2.4826	2.4657
	MAPE	**22.7397**	30.8858	24.1657	24.5206
	R^2	**0.8026**	0.4038	0.7434	0.7468
12 小时	*MAE*	1.4964	1.5708	**1.4613**	1.6180
	MSE	**5.4456**	6.9563	5.5884	7.0847
	RMSE	**2.3336**	2.6375	2.3640	2.6617
	MAPE	23.5332	25.3508	**22.6245**	25.6090
	R^2	**0.7734**	0.7105	0.7675	0.7052
15 小时	*MAE*	**1.4488**	1.4532	1.5065	1.4564
	MSE	**4.6149**	5.3936	6.1723	6.4300
	RMSE	**2.1482**	2.3224	4.4844	2.5357
	MAPE	22.9243	23.1412	23.2130	**21.9888**
	R^2	**0.8080**	0.7759	0.7433	0.7326
18 小时	*MAE*	**1.5229**	1.5686	1.5259	1.5672
	MSE	**5.4799**	6.8068	5.6223	6.0630
	RMSE	**2.3409**	2.6090	2.3711	2.4623
	MAPE	**23.5802**	23.7842	25.1680	25.1979
	R^2	**0.7722**	0.7171	0.7663	0.7480

由表 5.2 的预测性能评价结果可以看出，通过对多个视角的数据进行消融实验，我们发现随着视角的增加，预测性能不断提升。当我们只使用空气质量数据和气象数据进行预测时，模型的性能较为有限。然而，随着 POI 特征、路网数据和站点经纬度信息的引入，我们观察到预测性能进一步提升。这说明多个视角的数据融合能够更全面、准确地捕捉到影响城市 $PM_{2.5}$ 浓度的各种因素。因此，数据融合在城市 $PM_{2.5}$ 浓度预测中具有重要的优势，可以帮助我们更好地理解和预测空气质量，以支持决策和改善环境质量。

5.4.4 实验设置

本次实验所用的电脑硬件配置：14Vcpu Intel® Xeon® Gold 6330 CPU、显卡为 RTX A5000(24GB)的独立显卡。

实验所用的软件配置：在 Windows11（64 位）操作系统下执行。使用 Python（3.8 版本）进行数据预处理、模型的搭建以及训练。

本文编写的所有模型代码都是基于深度学习的 pytorch 框架，该框架能够快速的编码实验，其易扩展、易实现。

5.5 本章小结

本章提出了考虑到时空特性的 MGATs-LSTM 模型框架对甘肃省兰州市的 $PM_{2.5}$ 浓度值做出预测，并对模型框架进行了详细的介绍，首先对时空特性问题做出简单的阐述，对模型框架创建的想法及意义进行了描述，然后对模型的架构，计算过程，和各模块细节做出详细的说明，最后通过实验证明了本文提出的 MGATs-LSTM 模型的有效性以及证明数据融合对城市 $PM_{2.5}$ 浓度预测起着重要作用。

6 结论与展望

6.1 结论

从对兰州市的$PM_{2.5}$浓度进行预测得到的结果分析来看，主要得出如下结论：

（1）多视角数据融合是提高$PM_{2.5}$浓度预测可靠性和准确性的必要手段。在对$PM_{2.5}$浓度预测时，利用了时序数据（空气质量污染物数据、气象数据）和非时序特征（站点的地理距离、POI类别数量、路网数据）的多视角空间拓扑图构建使空间特征更加全面和准确，两者的结合提高了模型的预测性能，并增强了对数据的理解和解释能力。

（2）通过构建的MGCN-GRU和MGATs-LSTM时空预测模型框架，从空间和时间角度出发对兰州市的$PM_{2.5}$浓度进行了预测，模型结构更加科学合理。从预测结果可以看出，LSTM、GRU只考虑了时序特征，GAT-LSTM、GAT-GRU模型只考虑的时序特征的空间信息，模型对$PM_{2.5}$浓度的预测能力不足，效果欠佳。在基于多视角数据融合的基础上，MGCN-GRU和MGATs-LSTM模型提高了对$PM_{2.5}$浓度的预测能力，而进一步考虑了节点之间权重的MGATs-LSTM模型要较MGCN-GRU效果更好，该模型框架在原只考虑时序数据的改进两层图注意力网络上加入了非时序特征信息，使得模型可以同时利用时序数据和非时序特征用于空间信息的实时融合，采用对相邻节点的不同的空间信息进行聚合，促使模型结构更加科学合理。

（3）利用兰州市的多视角时空数据，包括空气质量数据、气象数据、站点经纬度信息、POI特征以及路网结构，验证了模型的预测性能和泛化能力。由于MGCN-GRU和MGATs-LSTM模型对地势复杂的兰州市8个空气质量监测站点的$PM_{2.5}$浓度都做出了预测，模型具有良好的泛化能力，可广泛应用于解决城市$PM_{2.5}$浓度预测问题。

总体来说，通过多视角数据融合的MGCN-GRU和MGATs-LSTM模型可以提高对$PM_{2.5}$浓度的准确预测，并为改善空气质量，促进城市环境建设做出贡献。模型的预测性能在兰州市的$PM_{2.5}$浓度预测中得到了验证，因此具有广泛的应用前景。

6.2 展望

本文致力于探讨空气质量预测问题，并通过分析及对比实验验证了所提出模型的可行性和有效性。虽然取得了一定成果，但仍存在一定的改进空间。在实验过程中，本文仅仅只是用到了甘肃省兰州市地区的真实数据，没有使用其他省份城市的数据进行验证来增强实验的说服力，因此，未来的研究方向可以包括利用其他省份城市的数据来验证模型的泛化能力，同时，可以开发空气质量可视化系统，以提升实验结果的说服力和可视化效果。

现行模型主要针对已建有监测站点的区域进行预测，却未考虑到监测站点稀少甚至不存在的地区情况。因此，未来研究可以针对不同地区的特点，对$PM_{2.5}$浓度进行预测，从而使模型具有更广泛的适用性和预测能力。

在短期预测方面，本文所提出的模型表现良好，但长期预测可能存在一定的误差。为此，未来的研究可以着重探讨长时间预测的方法和策略，以提高预测精准度和可靠性，从而更好地满足实际应用需求。通过持续深入的研究和改进措施，将能够使空气质量预测模型更加完善和可靠。

本文预测$PM_{2.5}$浓度只是一个具体的应用场景，希望未来可以将这种预测方法推广到其他相关领域的预测中。例如，空气质量预测模型可以推广到天气预测领域，通过监测空气中的污染物浓度来推断未来的天气状况。又或者应用到健康预测领域，通过监测$PM_{2.5}$浓度来预测人群的健康状况。这些领域的预测都可以受益于所构建的$PM_{2.5}$浓度预测模型框架，因此，希望未来可以将这种方法进行进一步的研究和推广，为更多相关领域提供有效的预测模型。

参考文献

[1] Brunsdon C, Fotheringham A S, Charlton M E. Geographically weighted regression: a method for exploring spatial nonstationarity[J]. Geographical analysis, 1996, 28(4): 281-298.

[2] Chen H.C, Putra K.T, Chun Wei, Lin J. A Novel Prediction Approach for Exploring $PM_{2.5}$ Spatiotemporal Propagation Based on Convolutional Recursive Neural Networks. 2021, 2101:6213.

[3] Chen L, Ding Y, Lyu 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), 1-17.

[4] Dong M, Yang D, Kuang Y, et al. $PM_{2.5}$ concentration prediction using hidden semi-Markov model-based times series data mining[J]. Expert Systems with Applications, 2009, 36(5): 9046-9055.

[5] D Aykas, S Mehrkanoon. Multistream Graph Attention Networks for Wind Speed Forecasting[J]. 2021 IEEE Symposium Series on Computational Intelligence (SSCI), 2021:1-8.

[6] Fotheringham A S, Charlton M E, Brunsdon C. Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis[J]. Environment and planning A, 1998, 30(11): 1905-1927.

[7] Ge L, Zhou A, Li H, et al. Spatially fine-grained air quality prediction based on DBU-LSTM[C]//Proceedings of the 16th ACM International Conference on Computing Frontiers. 2019: 202-205.

[8] 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): 2220-2242.

[9] H U Ttel F B, Peled I, Rodrigues F, et al. Deep Spatio-Temporal Forecasting of Electrical Vehicle Charging Demand[J]. arXiv preprint arXiv:2106.10940. 2021.

[10] Hochreiter S, Schmidhuber J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[11] Iram S Z, Javid M I, Sobia A, et al. Comparison of Ambient Air Quality among

Industrial and Residential Areas of a Typical South Asian City[J]. Atmosphere,2022,13(8): 1168-1180.

[12] Li C, Wang Z, Li B, Peng Z-R, Fu Q. Investigating the relationship between air pollution variation and urban form[J]. Build Environ, 2019,147:559-568.

[13] Liu Y, Cao G, Zhao N, Mulligan K, Ye X. Improve ground-level $PM_{2.5}$ concentration mapping using a random forests-based geostatistical approach[J]. Environ Pollut, 2018,235:272-282.

[14] 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.

[15] Li T, Hua M, Wu X. A Hybrid CNN-LSTM Model for Forecasting Particulate Matter ($PM_{2.5}$)[J]. IEEE Access, 2020, 8:26933-26940.

[16] Liu B, Yan S, Li J, et al. An attention-based air quality forecasting method[C]//2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2018: 728-733.

[17] Liang Y, Ke S, Zhang J, et al. Geoman: Multi-level attention networks for geo-sensory time series prediction[C]//International Joint Conference on Artificial Intelligence. 2018: 3428-3434.

[18] Ma J, Li Z, Cheng J C P, et al. Air quality prediction at new stations using spatially transferred bi-directional long short-term memory network[J]. Science of The Total Environment, 2020, 705: 135771.

[19] P. Gupta, S. A. Christopher. Particulate matter air quality assessment using integrated surface satellite and meteorological products: Multiple regression approach[J]. Atmos, 2009, 114:14205-14206.

[20] Qi Y, Li Q, Karimian H, et al. A hybrid model for spatiotemporal forecasting of $PM_{2.5}$ based on graph convolutional neural network and long short-term memory[J].Science of the Total Environment,2019,664:1-10.

[21] 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.

[22] Seyed Omid Nabavi, Leopold Haimberger, Esmail Abbasi, Assessing $PM_{2.5}$ concentrations in Tehran, Iran, from space using MAIAC, deep blue, and dark target AOD and machine learning algorithms[J]. Atmospheric Pollution Research, 2019, 10(3): 889-903.

[23] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Advancesin Neural Information Processing Systems 27: Annual Conference on Neural Information ProcessingSystems. Montreal, Quebec, Canada: Curran Associates, 2014: 3104-3112.

[24] Tao Q, Liu F, Li Y, Sidorov D. Air Pollution Forecasting Using a Deep Learning Model Based on 1D Convnets and Bidirectional GRU[J]. IEEE Access, 2019,7:76690-76698.

[25] Velickovic P, Cucurull G, Casanova A, et al. Graph attention networks[C]//6th International Conference on Learning Representations. Vancouver, BC, Canada: OpenReview net,2017:3-4.

[26] 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.

[27] Yann Lecun, Yoshua Bengio, Geoffrey Hinton.Deep learning[J]. Nature, 2015, 521(7553):436.

[28] Yi X , Zhang J , Wang Z , et al. Deep Distributed Fusion Network for Air Quality Prediction[C]// the 24th ACM SIGKDD International Conference. ACM, 2018.

[29] Zhao J, Deng F, Cai Y, Chen J. Long short-term memory - Fully connected (LSTMFC) neural network for $PM_{2.5}$ concentration prediction[J]. Chemosphere, 2019,220: 486-492.

[30] Zheng Y, Yi X, Li M, et al. Forecasting fine-grained air quality based on big data[C]//Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. 2015: 2267-2276.

[31] Zhou J, Cui G, Zhang Z, et al. Graph neural networks: A review of methods and applications[J].CoRR, 2018, abs/1812.08434.

[32] Zhu A, Liu J, Du F, et al. Predictive soil mapping with limited sample data [J].

European Journal of Soil Science, 2015, 66(3): 535-547.

[33] Zheng Y, Liu F, Hsieh H P. U-air: When urban air quality inference meets big data[C]//Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. 2013: 1436-1444.

[34] 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.

[35] 曹旺,王彤彤,张静怡.基于门控循环单元和图神经网络的 $PM_{2.5}$ 预测[J].现代计算机,2022,28(05):25-31.

[36] 陈逸彬,卢家璇,张书鸣等.基于混合 CNN-LSTM 结构的 $PM_{2.5}$ 浓度预测深度学习模型[J].信息与电脑(理论版),2022,34(04):53-55+65.

[37] 戴李杰, 张长江, 马雷鸣. 基于机器学习的 $PM_{2.5}$ 短期浓度动态预报模型[J]. 计算机应用, 2017, 37(11):3057-3063.

[38] 傅颖颖,张丰,杜震洪等.融合图卷积神经网络和注意力机制的 $PM_{2.5}$ 小时浓度多步预测[J].浙江大学学报(理学版),2021,48(01):74-83.

[39] 何强,井文涌,王翊亭.环境学导论[M].北京:清华大学出版社有限公司,2004.

[40] 黄亮东. 基于多源 POI 数据的天津市城市功能区识别与分析 [D]. 中国矿业大学, 2019.

[41] 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.

[42] 李雪佳, 封红旗, 梅宇,等. 基于三次多项式拟合三角函数的地理空间距离计算算法[J]. 计算机测量与控制, 2016(5):199-201.

[43] 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.

[44] 马俊文,严京海,孙瑞雯等.基于 LSTM-GCN 的 $PM_{2.5}$ 浓度预测模型[J].中国环境监测,2022,38(05):153-160.

[45] 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.

[46] 任硕. 基于膜计算的输电线路路径优化问题的研究与应用[D]. 山东师范大学, 2015:23-55.

[47] 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.

[48] 宋飞扬,铁治欣,黄泽华等.基于 KNN-LSTM 的 $PM_{2.5}$ 浓度预测模型[J].计算机系统应用,2020,29(07):193-198.

[49] 孙小新.基于图神经网络的 $PM_{2.5}$ 浓度预测算法研究[D].东北师范大学,2022:62-97.

[50] 徐东,杨晓芳.基于多元线性回归模型预测成都市 $PM_{2.5}$ 趋势[J].黑龙江科学,2021,12(06):36-37.

[51] 颜俨,姚柳杨,徐涛等.空气污染治理的公众偏好及政策评价——以西安市雾霾治理为例[J].干旱区资源与环境,2018,32(04):19-25.

[52] 叶如珊,王海波.基于 CNN-BiLSTM 模型的 $PM_{2.5}$ 浓度预测方法[J].数学的实践与认识,2022,52(07):181-188.

[53] 于书玉.基于 LSTM 融合神经网络预测模型研究[J].科学技术创新,2023,7:87-90.

[54] 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.

[55] 张旭.不完全数据下基于元学习的城市空气质量预测[D].哈尔滨工程大学,2022:19-22.

[56] 中国环境监测总站.环境空气质量监测点位布设技术规范(试行)[M].北京:中国环境科学出版社,2013:1-2.

[57] 中国环境科学研究院. 环境空气质量标准[M]. 中国环境科学出版社, 2012:8-9.

[58] 朱秀娟. 基于多站点的 PM2.5 预测模型研究[D].郑州大学,2022:19-21.

致 谢

值此论文完成之际，我要向你们致以最诚挚的谢意。

感谢老师在我学习期间的悉心教导和指导。在毕业论文写作过程中，老师给予了我许多宝贵的建议和意见。在我迷茫的时候，老师总是耐心地给予指引和鼓励。没有老师的支持和帮助，我无法完成这篇论文。

同时，我要感谢我的父母。他们对我的支持和鼓励是我最坚强的后盾。在我遇到困难的时候，他们总是在我身边给予我鼓励和支持。毕业论文虽然是我个人的成果，但是离不开父母的支持和鼓励。

此外，我要感谢我的同学和朋友。在我写论文的过程中，他们给予了我许多帮助和支持。我们相互鼓励和交流，一起成长。没有他们的陪伴和帮助，我研究生生活略显枯燥乏味。

最后，感谢学校给予我学习的机会和舞台。学校的教育为我打下了坚实的基础，让我能够在这个社会上游刃有余。学校的教育理念和师资力量为我提供了宝贵的学习资源，让我得以不断地成长。

在毕业的时刻，我要感谢所有给予我帮助和支持的人。有了你们的帮助，我才能够顺利完成学业并踏上新的征程。希望在以后的日子里，我能够继续努力，为社会做出更大的贡献。再次感谢大家！