文章编号:2096-1472(2026)10-0019-06
DOI:10.19644/j.cnki.issn2096-1472.2026.010.004

基于智能邻接矩阵 GCN-LSTM 的空气质量预测

王依卓,李 林

(上海理工大学管理学院,上海 200093)
✉ 398743495@qq.com; ll_ft@163.com

摘 要:为突破传统单站点空气质量预测模型捕捉污染物空间关联的局限,提出一种融合智能邻接矩阵的图卷积-长短期记忆网络(GCN-LSTM)模型。该模型通过构建综合地理距离、$PM_{2.5}$浓度时序相关性与站点数据质量的智能邻接矩阵,精准刻画多站点间的动态空间依赖关系。结合图卷积网络(GCN)与注意力机制提取空间特征,并与长短期记忆(LSTM)协同建模时序动态以实现目标站点 $PM_{2.5}$浓度高精度预测。实验结果表明,相较于基准 LSTM 模型,GCN-LSTM 模型预测精度显著提升,决定系数(R^2)达 0.899 6,均方根误差(RMSE)和平均绝对误差(MAE)分别降至 4.93 和 3.32,验证了智能空间特征建模的有效性,为城市空气质量精准预测提供新思路。

关键词:空气质量预测;图卷积网络;长短期记忆网络;时空预测
中图分类号:TP183 **文献标识码**:A

Air Quality Prediction Based on GCN-LSTM with Intelligent Adjacency Matrix

WANG Yizhuo, LI Lin

(*Business School, University of Shanghai for Science & Technology, Shanghai* 200093, *China*)
✉ 398743495@qq.com; ll_ft@163.com

Abstract: To overcome the limitation of traditional single-site air quality prediction models in capturing the spatial correlation of pollutants, a graph convolutional network-long short-term memory (GCN-LSTM) model incorporating an intelligent adjacency matrix is proposed. By constructing an intelligent adjacency matrix that integrates geographical distance, temporal correlation of $PM_{2.5}$ concentrations, and site data quality, the model accurately characterizes the dynamic spatial dependencies among multiple sites. It combines graph convolutional networks and attention mechanisms to extract spatial features and synergistically model temporal dynamics with LSTM, thereby achieving high-precision prediction of $PM_{2.5}$ concentration at the target site. Experiments show that compared with the baseline LSTM model, the prediction accuracy of the proposed GCN-LSTM model is significantly improved, with the coefficient of determination (R^2) reaching 0.899 6 and the root mean square error (RMSE) and mean absolute error (MAE) reduced to 4.93 and 3.32 respectively. This validates the effectiveness of intelligent spatial feature modeling and provides a new approach for accurate urban air quality prediction.

Key words: air quality prediction; Graph Convolutional Network; Long Short-Term Memory network; spatiotemporal prediction

0 引言(Introduction)

近年来,随着我国生态文明建设的深入推进和“双碳”战略的全面实施[1],推动经济社会发展全面绿色转型、持续改善生态环境质量已成为高质量发展的核心内涵。在此战略指引下,大气污染防治工作正迈向精准溯源、科学施策的攻坚阶段。因此,发展高精度、高时效的空气质量预测技术,为实现公共健康防护与污染的精准溯源和靶向治理提供了关键技术支撑,具有至关重要的现实意义。

当前空气污染预测方法主要可分为数值预报方法、统计方法与人工智能方法三类[2]。数值预报方法基于大气物理与化

收稿日期:2025-11-26

www.rjgczz.com

学机理，通过构建数学模型模拟污染物在大气中的迁移、扩散、化学转化及干湿沉降等关键过程[3]，具备良好的物理可解释性与空间覆盖能力。但它对排放清单准确性依赖较高、计算资源消耗大，且在复杂下垫面与边界层条件下存在机理不确定性，导致它在实时快速预测与局部精度方面存在一定的局限[4]。

统计方法主要依托历史观测数据构建预测模型，常见方法包括时间序列分析与回归分析等。该类方法能够有效捕捉数据中的时序趋势与统计关联，提高预测的准确性和可靠性。李颖若等[5]采用季节性自回归积分滑动平均模型对臭氧浓度进行预测，结果表明该模型能有效刻画污染物季节性变化特征；王凯文等[6]构建了基于多元线性回归的空气质量指数预测模型，通过引入气象因子提升了预测的实用性。

人工智能方法进一步可分为机器学习与深度学习两类。机器学习方法如随机森林，在处理高维非线性数据方面表现出较强适应能力；深度学习方法如长短期记忆(Long Short-Term Memory，LSTM)网络，则通过深层结构自动提取时序依赖特征。杨思琪等[7]基于随机森林构建多因子空气质量预测模型，结果显示该模型的预测性能优于 Boosting 等对比模型；白盛楠等[8]则结合灰色关联分析与 LSTM 网络，构建多变量 $PM_{2.5}$ 预测模型，实现了对日浓度变化的有效捕捉。

尽管上述方法在不同层面推动了空气质量预测技术的发展，现有研究在空间依赖性建模方面仍存在不足，尤其在图结构构建过程中对站点间异质关联的刻画尚不充分。为此，本文提出一种融合智能邻接矩阵的 GCN-LSTM 预测模型，主要内容包括：(1)设计了一种综合考虑地理距离、$PM_{2.5}$ 浓度时序相关性及各站点数据质量的多阈值智能邻接矩阵构建方法，以更精准地量化站点间的动态空间依赖强度；(2)构建了一个端到端的 GCN-LSTM 模型框架，通过图卷积网络(Graph Convolutional Network，GCN)模块提取多站点的空间特征，并结合注意力机制与 LSTM 模块协同建模时空动态，以实现 $PM_{2.5}$ 浓度的高精度单步预测。

1 基于 GCN-LSTM 的预测模型(Prediction model based on GCN-LSTM)

1.1 GCN 模型原理

图卷积网络(GCN)是一种基于频谱图理论的深度学习架构，它的核心机制是通过聚合节点的局部邻域信息以生成有效节点嵌入，从而实现对图结构数据中复杂依赖关系的建模[9]。本文将上海市的空气质量监测网络视为一个图结构，其中每个监测站点是图中的一个节点，站点之间的空间关联关系构成边。GCN 的核心思想是通过聚合邻居节点的信息来更新当前节点的特征表示，从而捕捉复杂的空间依赖关系。

1.1.1 图卷积网络基本原理

传统卷积神经网络(Convolutional Neural Network，CNN)所依赖的离散卷积操作受限于规整的欧几里得空间数据(如图像、网格)[10]。然而，在空气质量预测任务中，监测站点在地理空间上呈不规则分布，它的拓扑结构本质上是非欧几里得空间的，无法直接应用传统 CNN。GCN 通过频谱图理论将卷积操作泛化至图结构数据，从而实现对此类非规则空间依赖关系的建模[11]。GCN 的结构示意图如图 1 所示。

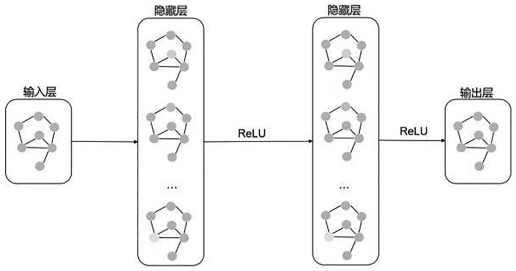

图 1 GCN 结构示意图

Fig. 1 Schematic diagram of GCN structure

GCN 的核心的层间传播规则如式(1)所示：

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

其中：$\widetilde{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I}_N$ 是添加自动连接的邻接矩阵($\boldsymbol{I}_N$ 为单位矩阵)；$\widetilde{\boldsymbol{D}}$ 是 $\widetilde{\boldsymbol{A}}$ 的度矩阵，它的对角线元素 $\widetilde{D}_{ii}=\sum_j \widetilde{A}_{ij}$；$\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}$ 是对邻接矩阵进行对称归一化，用于稳定训练过程并防止梯度爆炸；$\boldsymbol{H}^{(l)}$ 是第 l 层的节点特征矩阵；$\boldsymbol{W}^{(l)}$ 是第 l 层可训练的节点特征矩阵；$\sigma(\cdot)$ 是 ReLU 激活函数。

1.1.2 空间注意力机制

标准 GCN 模型固有的局限性在于它对各邻居节点施加了等权重的聚合策略，忽略了节点间相互作用的异质性。为克服这一局限，使模型能够依据实际数据动态评估空间关联的强度(如上风向站点相较于下风向站点通常具有更强的影响力)，本文引入了空间注意力机制[12]，以实现对非对称空间依赖关系的自适应建模。

空间注意力机制能够自适应地学习每个邻居节点对目标节点的重要性权重。它的计算过程如式(2)所示：

$$\alpha_i=\mathrm{Softmax}(\boldsymbol{W}_2\cdot\mathrm{ReLU}(\boldsymbol{W}_1\cdot h_i)) \tag{2}$$

其中：$h_i\in\mathbb{R}^H$ 是站点 i 的 GCN 输出特征；$\boldsymbol{W}_1\in\mathbb{R}^{H\times H/2}$、$\boldsymbol{W}_2\in\mathbb{R}^{H/2\times 1}$ 是可学习参数；Softmax 确保所有权重之和为 1：$\sum_{i=1}^{N}\alpha_i=1$。该注意力机制能够根据实时气象条件、污染物分布等动态因素，自适应调整各邻居站点的贡献度，从而更准确地反映实际的大气污染扩散过程。

在获得空间注意力权重后，对目标站点特征和邻居站点特征进行加权融合：

$$h_{\mathrm{fused}}=\lambda\cdot h_{1147\mathrm{A}}+(1-\lambda)\cdot(\sum_{j=1}^{N}\alpha_j\cdot h_j) \tag{3}$$

其中：$\lambda=0.7$ 是目标站点特征的权重系数，该权重系数是基于大量实验验证的最优配置，既保证了目标站点自身特征的主导地位，又充分利用了空间邻居的辅助信息；$\lambda_{1147\mathrm{A}}$ 是目标站点特征。

1.1.3 智能邻接矩阵构建

邻接矩阵的构建质量直接决定了 GCN 的空间特征提取效果。传统图构建方法通常基于地理距离度量，采用 Delaunay 三角剖分或 k 近邻方法建立对称的邻接关系[13-14]。然而，空气污染物的传播受到风向、风速、地形障碍物等多种因素的影响，呈现出显著的非对称特性[15]。简单的对称距离模型难以准确刻画这种复杂的空间依赖关系。为此，本研究提出了一种智能邻接矩阵构建方法，综合考虑地理距离、时序相关性和数据质量三重因素。具体而言，邻接矩阵 $\boldsymbol{A}\in\mathbb{R}^{N\times N}$ 的构建采用多阈值筛选策略：设定距离阈值 $d_{ij}\leqslant 30$ km 以捕捉空间邻近性，相关性阈值 $p_{ij}>0.6$ 以确保时序模式的相似性，并引入质量阈值(缺失率 $m_j<0.2$、零值 $z_j<0.3$)以筛选数据可靠的站点。

www.rjgcz.com

对于任意 2 个不同站点 i 和 j，连接权重 A_{ij} 的计算公式如式(4)所示：

$$A_{ij}=\text{SpatialWeight}(i,j)\times\text{CorrelationWeight}(i,j)\times\text{QualityWeight}(j) \tag{4}$$

各分类计算如下：

①空间临近权重：

$$\text{SpatialWeight}(i,j)=\exp\left(-\frac{d_{ij}^2}{2\sigma^2}\right),\quad \sigma=15 \tag{5}$$

基于站点间的经纬度，使用 Haversine 公式计算大圆距离 d_{ij}。距离越近，空间关联性通常越强，并通过高斯核函数计算其权重贡献。

②时序相关性权重：

$$\text{CorrelationWeight}(i,j)=|p_{ij}| \tag{6}$$

其中：p_{ij} 为站点与目标站点 i 历史 $PM_{2.5}$ 浓度的皮尔逊相关系数。相关性越高，表明两站点污染模式越相似，连接应越强。

皮尔逊相关系数 p_{ij} 的计算公式如式(7)所示：

$$p_{ij}=\frac{\sum_{t=1}^{T}(x_t^i-\bar{x}^i)(x_t^j-\bar{x}^j)}{\sqrt{\sum_{t=1}^{T}(x_t^i-\bar{x}^i)^2}\cdot\sqrt{\sum_{t=1}^{T}(x_t^j-\bar{x}^j)^2}} \tag{7}$$

其中：x_t^i 和 x_t^j 分别表示站点 i 和站点 j 在时间 t 的 $PM_{2.5}$ 浓度值；$\bar{x}^i$ 和 $\bar{x}^j$ 分别表示站点 i 和站点 j 在整个时间序列上的 $PM_{2.5}$ 浓度平均值；T 是时间序列的总长度(时间点的总数)。

③数据质量权重：

$$\text{QualityWeight}(j)=(1-m_j)\times(1-z_j) \tag{8}$$

其中：m_j 为站点 j 的数据缺失率；z_j 为 0 值率。数据质量越差，其权重贡献越低。

以目标站点 1147A(静安监测站)为例，它的有效邻居站点共 14 个，按影响权重降序排列，见表 1。

表1　目标站点的邻居站点信息

Tab.1 Analysis of the influence of neighboring sites on the target site

邻居站点	距离/km	相关系数	影响权重	空间关系
1141A	2.7	0.964	0.948	极近邻，强影响
3273A	4.0	0.965	0.931	近邻，强影响
1142A	5.6	0.947	0.883	近邻，强影响
1144A	6.9	0.954	0.859	中近邻，强影响
1143A	9.2	0.937	0.776	中距离，较强影响
1149A	10.3	0.940	0.743	中距离，较强影响
1145A	11.4	0.930	0.693	中距离，中等影响
3265A	11.8	0.924	0.677	中距离，中等影响
3268A	12.2	0.901	0.648	中距离，中等影响
1150A	14.6	0.926	0.577	中远距离，较弱影响
3271A	14.6	0.920	0.574	中远距离，较弱影响
3270A	18.4	0.851	0.403	远距离，弱影响
1148A	26.7	0.896	0.183	边缘距离，极弱影响
3272A	28.7	0.871	0.140	边缘距离，极弱影响

1.2 LSTM 模型原理

长短期记忆(LSTM)网络是一种特殊的循环神经网络(Recurrent Neural Network，RNN)，它通过引入精密的门控机制和细胞状态，有效解决了传统 RNN 在训练长序列数据时难以避免的梯度消失(Vanishing Gradient，VG)和梯度爆炸(Exploding Gradient，EG)问题[16-17]。LSTM 单元的结构示意图如图 2 所示。它的核心在于 3 个门控：遗忘门 $\boldsymbol{f}_t$ 决定从历史细胞状态 $\boldsymbol{C}_{t-1}$ 中丢弃何种信息；输入门 $\boldsymbol{i}_t$ 控制将多少新信息写入当前细胞状态 $\boldsymbol{C}_t$；输出门 $\boldsymbol{o}_t$ 基于更新后的细胞状态，计算当前时刻的输出 $\boldsymbol{h}_t$。

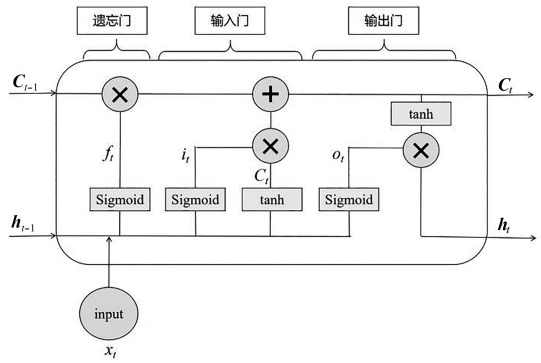

图 2　LSTM 结构示意图

Fig. 2 Schematic diagram of LSTM structure

①遗忘门(Forget Gate)

遗忘门决定从先前细胞状态中丢弃哪些信息。它通过 Sigmoid 激活函数查看当前输入 x_t 和前一时刻隐藏状态 $\boldsymbol{h}_{t-1}$，为细胞状态 C_{t-1} 中的每个元素输出一个 0～1 之间的数值，其中 0 表示“完全丢弃”，1 表示“完全保留”。

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_f\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_f) \tag{9}$$

其中：$\boldsymbol{f}_t$ 是遗忘门输出向量；σ 是 Sigmoid 激活函数，$\sigma(\boldsymbol{x})=\frac{1}{1+e^{-x}}$；$\boldsymbol{W}_f$ 是遗忘门的权重矩阵；$[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]$ 表示将两个向量拼接；$\boldsymbol{b}_f$ 是遗忘门的偏置向量。

②输入门(Input Gate)与候选细胞状态

输入门决定哪些新信息将存入细胞状态。它包含两个部分：Sigmoid 层决定要更新的值；tanh 层生成候选值向量 $\widetilde{\boldsymbol{C}}_t$。

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_i\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_i) \tag{10}$$

$$\widetilde{\boldsymbol{C}}_t=\tanh(\boldsymbol{W}_C\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_C) \tag{11}$$

其中：$\boldsymbol{i}_t$ 是输入门的输出向量；$\widetilde{\boldsymbol{C}}_t$ 是候选细胞状态向量；tanh 是双曲正切函数：$\tanh(\boldsymbol{x})=\frac{e^x-e^{-x}}{e^x+e^{-x}}$。

③细胞状态更新

基于遗忘门和输入门的输出，更新当前时刻的细胞状态 C_t：

$$\boldsymbol{C}_t=\boldsymbol{f}_t\odot\boldsymbol{C}_{t-1}+\boldsymbol{i}_t\odot\widetilde{\boldsymbol{C}}_t \tag{12}$$

其中：$\odot$ 表示逐元素乘法(Hadamard 积)；$\boldsymbol{f}_t\odot\boldsymbol{C}_{t-1}$ 控制历史信息的保留程度；$i_t\odot\widetilde{C}_t$ 控制新信息的添加程度。

④输出门(Output Gate)与隐藏状态

输出门决定下一个隐藏状态的内容。隐藏状态 h_t 包含了先前序列的信息，用于预测输出。

$$\boldsymbol{o}_t=\sigma(W_o\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_o) \tag{13}$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t\odot\tanh(\boldsymbol{C}_t) \tag{14}$$

其中：$\boldsymbol{o}_t$ 是输出门的输出向量；$\boldsymbol{h}_t$ 是当前时刻的隐藏状态，也是该 LSTM 单元的输出。

1.3 基于 GCN-LSTM 的空气质量预测

前述 GCN 与 LSTM 分别擅长处理空间与时间维度的特征，本研究将二者结合，构建了一个端到端的 GCN-LSTM 混合预测模型，以实现对 PM2.5 浓度的精准时空预测。模型的整体架构遵循“多站点输入 → GCN 时空特征提取 → 时空特征融合 → LSTM 时序建模 → 单站点输出”的流程，旨在协同挖掘数据中的时空模式。GCN-LSTM 预测模型的示意如图 3 所示。

www.rjgcz.com

图 3 GCN-LSTM 结构示意图

Fig. 3 Schematic diagram of the GCN-LSTM structure

GCN-LSTM 模型以一个 4 维张量 $\boldsymbol{X}\in\mathbb{R}^{B\times N\times T\times F}$ 作为输入，其中 B 为批次大小，$N=19$ 为站点数量，$T=24$ 为历史时间步长(h)，$F=23$ 为特征维度(包括 6 维空气质量、6 维气象、8 维时间周期、3 维时间分类特征)。对于序列中的每一个时间步 $t=1,2,\cdots,T$，模型提取所有 N 个站点的当前特征 $\boldsymbol{X}_{:,:,t}\in\mathbb{R}^{B\times N\times F}$。将该特征张量与预先构建的智能邻接矩阵 $\boldsymbol{A}$ 一同输入 GCN 模块。GCN 模块通过信息传播与聚合，输出每个站点在当前时刻的空间编码特征，对目标站点及其邻居站点的 GCN 输出特征进行注意力加权融合，得到该时刻目标站点的最终空间特征表示 $h^{t}_{\text{fused}}\in\mathbb{R}^{B\times H}$($H$ 为 GCN 隐藏层维度)。将 T 连续时间步得到的空间特征 $h^{1}_{\text{fused}},h^{2}_{\text{fused}},\cdots,h^{T}_{\text{fused}}$ 按时间顺序堆叠，形成一个新的时空特征序列 $S\in\mathbb{R}^{B\times T\times H}$。将时空特征序列 S 输入 LSTM 网络中。LSTM 单元按其门控机制依次处理序列，学习其动态演化规律。取 LSTM 最后一个 $h_T\in\mathbb{R}^{B\times D}$($D$ 为 LSTM 隐藏层维度)作为对整个输入序列时空动态的概括性编码。将 h_T 通过一个全连接输出层，映射为最终标量输出 $\hat{y}\in\mathbb{R}^{B\times 1}$，即目标站点未来 1 h 的 $PM_{2.5}$ 浓度预测值。

2 空气质量预测应用(Application of air quality prediction)

2.1 数据简介与分析

本研究所使用的空气质量数据来源于中国环境监测总站的全国城市空气质量实时发布平台，时间跨度为 2022—2024 年，记录频率为逐小时，每个监测站点包含 26 093 条有效记录，共 19 个站点。原始数据集包含 $PM_{2.5}$、PM_{10}、SO_2、NO_2、CO、O_3 等 6 项主要污染物的实时浓度及其 24 h 滑动平均浓度等多个指标。各监测站点 $PM_{2.5}$ 年均浓度的空间分布如图 4 所示，呈现出明显的空间异质性。地理信息位置见表 2，基本覆盖了上海市全部行政区域，能有效代表全市不同功能区的空气质量状况。

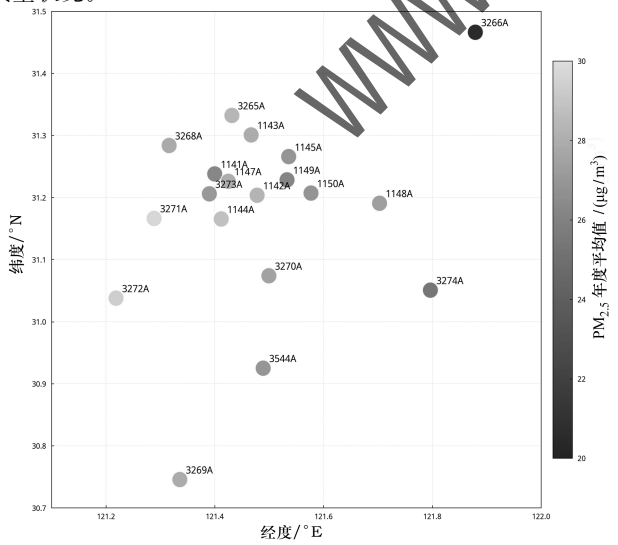

图 4 上海市 19 个监测站点 $PM_{2.5}$ 年均质量浓度空间分布

Fig. 4 Spatial distribution of PM2.5 annual average concentration at 19 monitoring stations in Shanghai

表2 空气质量监测站点信息

Tab. 2 Information of air quality monitoring stations

监测点编码	监测点名称	经度/(°E)	纬度/(°N)
1141A	普陀	121.400	31.238 0
1142A	十五厂	121.478	31.203 6
1143A	虹口	121.467	31.300 8
1144A	徐汇上师大	121.412	31.165 4
1145A	杨浦四漂	121.536	31.265 9
1147A	静安监测站	121.425	31.226 1
1148A	浦东川沙	121.703	31.190 7
1149A	浦东新区监测站	121.533	31.228 4
1150A	浦东张江	121.577	31.207 1
3265A	宝山庙行	121.431	31.332 2
3266A	崇明上实东滩	121.879	31.466 2
3268A	嘉定南翔	121.316	31.283 8
3269A	金山新城	121.335	30.745 6
3270A	闵行浦江	121.499	31.073 9
3271A	青浦徐泾	121.288	31.166 3
3272A	松江图书馆	121.218	31.037 8
3273A	长宁仙霞	121.390	31.205 9
3274A	浦东惠南	121.796	31.050 7
3544A	奉贤南桥新城	121.489	30.925 0

气象数据来源于美国国家气候数据中心(National Climatic Data Center，NCDC)，与空气质量数据同期，记录频率为逐 3 h，共 8299 条记录。数据包含温度(℃)、露点温度(℃)、海平面气压(hPa)、风向(°)、风速(m/s)、6 h 降水量(mm) 等关键气象要素。为与空气质量数据时间频率对齐，本文后续采用线性插值法将该数据集处理为逐小时数据。

2.2 数据预处理

2.2.1 空气质量数据清洗

针对不同字段的缺失率及其物理特性，本文采用差异化填充策略。对于低缺失率(<0.5%)的污染物(如 O_3、NO_2、SO_2)，采用前向填充结合后向填充的方法；对于中等缺失率(>0.6%)的污染物(如 $PM_{2.5}$、PM_{10}、CO 等)，采用小窗口线性插值并结合 9 点滚动均值进行平滑；对于高缺失率(38.3%)的 O_3_24 h 等衍生指标，则依据其定义重新计算(采用 24 h 滑动窗口)，以确保数据的物理合理性。

异常值则采用基于统计规则与行业标准相结合的方法。参照国家环境空气质量标准识别明显超出合理范围的数值。对于无明确标准的字段，则根据其数据分布特性自动选择异常值识别方法：对近似正态分布的字段(如 O_3 浓度)采用 3σ 准则，对明显偏态分布的字段(如 $PM_{2.5}$、CO 浓度)采用 IQR 箱线法。所有识别出的异常值，统一使用相邻 3 点滚动均值进行替换。

2.2.2 气象数据清洗

缺失值的处理与空气质量数据的处理思路相同，根据字段缺失率对气象数据采取分级策略：对缺失率低于 5%的字段(如露点温度)，采用简单的前后向填充；对缺失率在 5%～50%之间的字段(如 6 h 降水量)，采用线性插值法；对于缺失率超过 50%的无效字段(如天空覆盖代码)，则直接剔除。

针对风速、6 h 降水量等易出现异常值的字段，使用 IQR 箱线法进行识别，并用相邻 3 点滚动均值进行替换，确保数据的合理性与平滑性。

最后将原始的逐 3 h 气象数据，通过线性插值降采样为逐小时数据，使其与空气质量数据的时序频率严格对齐。

www.rjgcz.com

2.2.3 特征工程

从原始空气质量数据中,选取 CO、NO_2、O_3、PM_{10}、$PM_{2.5}$、SO_2这 6 种污染物的实时浓度作为模型输入特征,避免使用其 24 h 滑动平均值以防止引入未来信息。并从时间戳中提取周期性特征,帮助模型学习污染物浓度的日、周、年变化规律,最后构建了时间分类特征,如是否周末、是否工作日以及季节特征。

2.3 GCN-LSTM 空气质量预测

为验证 GCN-LSTM 混合模型的有效性,实验在配备 NVIDIA RTX 4090 GPU 的服务器上进行,独立运行 10 次实验,每次使用不同的随机种子(42～51)。邻接矩阵构建中:距离阈值设定为 $d_{ij}\leqslant 30$ km,高斯核宽带 $\sigma=15$ km;相关性阈值 $p_{ij}>0.6$;数据质量阈值设定为缺失率 $m_j<0.2$,零值率 $z_j<0.3$。模型训练使用 Adam 优化器,初始学习率为 0.001,权重衰减为 1×10^{-5},采用均方误差作为损失函数,并设置了学习率调度与早停机制(最大训练轮数为 100、耐心值为 15)以防止过拟合。数据配置方面,使用过去 24 h 的序列数据进行预测,批大小设置为 32,数据集按照 70%、15%、15%的比例划分为训练集、验证集和测试集。基准模型采用与 GCN-LSTM 模型中 LSTM 部分结构相同的纯 LSTM 模型,以确保对比的公平性。GCN-LSTM 模型的参数设置见表 3。

表3　GCN-LSTM 模型参数设置

Tab. 3 Parameter settings of GCN-LSTM model

模块	参数名称	参数值	说明
GCN 编码器	输入维度	23	站点特征数量
	隐藏层维度	32	GCN 输出特征维度
	图卷积层数	2	两层 FixedGraphConvolution
LSTM 时序处理器	隐藏层维度	64	LSTM 网络隐藏状态维度
	LSTM 层数	2	堆叠的 LSTM 层数
	Dropout 率	0.2	防止过拟合的丢弃率
输出层	全连接层	[64, 32, 1]	两层全连接网络结构

GCN-LSTM 预测模型采用目标站点特征与邻居信息相结合的策略,其中目标站点特征权重为 0.7,邻居信息权重为 0.3,通过图卷积网络有效捕捉多站点间的空间依赖关系。

2.4 预测结果评价与统计显著性检验

为客观评估 GCN-LSTM 模型与基准模型的预测性能,本研究采用均方根误差(Root Mean Square Error,RMSE)、平均绝对误差(Mean Absolute Error,MAE)与决定系数(R^2)作为核心评价指标。所有评价均在独立的测试集上进行,各指标定义如式(15)～式(17)所示:

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{15}$$

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \tag{16}$$

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \tag{17}$$

为评估模型性能的稳定性,本研究独立运行了 10 次实验,每次使用不同的随机种子初始化网络权重。结果对比见表 4。GCN-LSTM 表现最优,其 RMSE 为(4.93±0.20)μg/m³。与次优模型相比,GCN-LSTM 的 RSME 相比 Transformer 模型(6.51±0.22)μg/m³ 降低了 24.3%,相比 LSTM 模型(6.10 ± 0.20)μg/m³ 降低了 19.2%。配对 t 检验结果表明,这些性能提升具有高度统计显著性(GCN-LSTM vs Transformer:$t=-14.748$,$p<0.001$。GCN-LSTM vs LSTM:$t=-13.246$,$p<0.001$)。

表4　各模型 10 次独立实验的平均预测性能对比

Tab. 4 Comparison of average prediction performance of different models over 10 independent experiments

模型	RMSE/(μg·m⁻³)	MAE/(μg·m⁻³)	R^2
ARIMA	7.86 ± 0.17	8.59 ± 0.47	0.815 9 ± 0.011 1
SARIMA	7.40 ± 0.43	7.04 ± 0.38	0.803 8 ± 0.005 3
Transformer	6.10 ± 0.21	5.44 ± 0.31	0.854 3 ± 0.010 2
LSTM	6.51 ± 0.22	6.25 ± 0.41	0.856 7 ± 0.010 1
GCN-LSTM	**4.93 ± 0.19**	**3.32 ± 0.15**	**0.899 6 ± 0.007 9**

各模型 RMSE 的分布(误差棒)及显著性标记,直观反映了 GCN-LSTM 的优越性与稳定性,如图 5 所示。进一步展示了 GCN-LSTM 在测试集部分时段的预测曲线及其 95%置信区间,如图 6 所示,预测值与真实值吻合良好,且置信区间较窄,表明模型具有较高的预测稳定性。

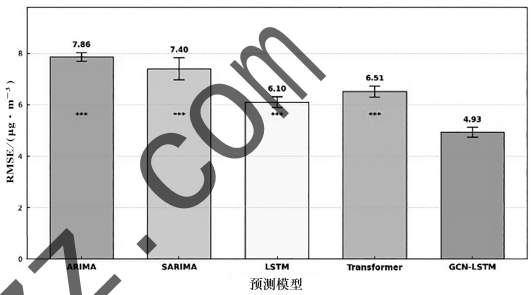

图 5　各模型误差棒图与显著性标记

Fig. 5 Error bar plots and significance markers of different models

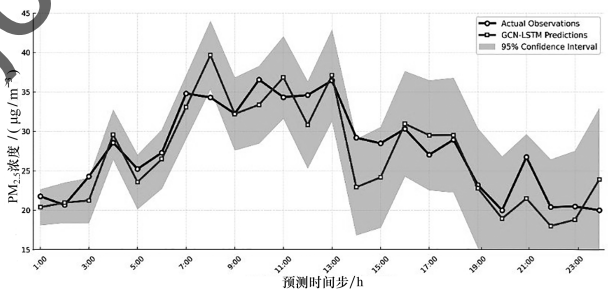

图 6　GCN-LSTM 预测曲线与 95%置信区间

Fig. 6 GCN-LSTM prediction curve and 95% confidence interval

2.5 GCN-LSTM 有效性分析

2.5.1 关键参数敏感性分析

对目标站点特征融合权重 λ 的敏感性分析见表 5。当 λ=0.7 时模型性能达到最优,平衡了目标站点自身信息与邻居站点空间信息的利用。

表5　目标站点特征融合权重 λ 的敏感性分析

Tab. 5 Sensitivity analysis of feature fusion weight λ for the target site

λ	RMSE/(μg·m⁻³)	MAE/(μg·m⁻³)	R^2
0.5	5.14	3.51	0.892 3
0.6	4.98	3.39	0.893 7
0.7	**4.93**	**3.32**	**0.899 6**
0.8	4.99	3.40	0.896 2
0.9	5.16	3.53	0.891 4

2.5.2 与基准图构建方法的对比

将本文 GCN-LSTM 中的邻接矩阵与 3 种常见基准方法对比:k 近邻(kNN)、Delaunay 三角剖分以及基于气象风场修正的

www.rjgczz.com

有向图。所有对比均在相同的 GCN-LSTM 框架下进行。见表6，本文提出的智能邻接矩阵取得了最低的 RMSE(4.93 $\mu g/m^3$)，显著优于所有基准方法。

2.5.3 智能邻接矩阵成分消融实验

为剖析智能邻接矩阵中“距离”“时序相关性”和“数据质量”3个因素的贡献，构建了仅包含单一或部分因素的变体模型进行对比。结果见表6。

表6 不同邻接矩阵构建方法的消融实验结果对比

Tab.6 Comparison of ablation experimental results under different adjacency matrix construction methods

模型	RMSE/(μg·m^{-3})	MAE/(μg·m^{-3})	R^2
S-GCN-LSTM	5.24	3.68	0.895 6
C-GCN-LSTM	5.19	3.68	0.905 6
Q-GCN-LSTM	6.5	3.79	0.864 8
GCN-LSTM	**4.93**	**3.32**	**0.899 6**
kNN-GCN-LSTM	5.14	5.58	0.887 6
Delaunay-GCN-LSTM	5.57	3.68	0.862 4
Wind-GCN-LSTM	5.94	4.47	0.875 6

三者融合(完整模型)性能最优，验证了多因素信息融合的协同增强效应。仅使用时序相关性(C-GCN-LSTM)优于仅使用地理距离(S-GCN-LSTM)，说明污染物浓度的时空模式相似性比单纯的地理邻近性更能刻画站点间的有效关联。仅使用数据质量筛选(Q-GCN-LSTM)表现相对较弱，表明质量因素需结合时空特征才能有效发挥过滤噪声的作用，而非独立决定连接关系。

2.5.4 智能邻接矩阵稀疏度影响

智能邻接矩阵的稀疏度(由距离阈值 d_{ij} 控制)影响模型效率与信息聚合范围。如图7所示，模型性能(RMSE)随稀疏度变化呈现先改善后轻微上升的趋势，在稀疏度约 0.6(对应 $d_{ij}=30$ km)时达到较优平衡，过密或过疏的连接均会导致性能损失。图7中，$|E|/|V|^2$ 为智能邻接矩阵的稀疏度，其中|V|为站点总数，|E|为图中实际存在的边数。该指标随距离阈值 d_{ij} 的增大而增大，反映了站点间空间连接的稠密程度。

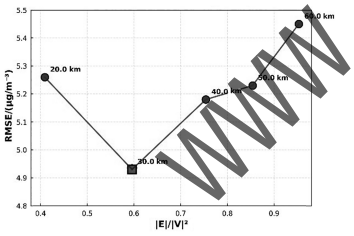

图7 稀疏度-误差曲线

Fig.7 Sparsity-error curve

3 结论(Conclusion)

本研究针对城市空气质量单站点 $PM_{2.5}$ 浓度预测中空间信息利用不足的问题，提出了一种结合 GCN 与 LSTM 的 GCN-LSTM 混合模型。该模型的核心在于通过一种智能邻接矩阵构建方法，将地理邻近性、时序相关性及站点数据质量融为一体，以精准量化多监测站点之间复杂的空间依赖关系。实验结果表明，在目标站点的预测任务中，GCN-LSTM 模型展现出全面优于纯 LSTM 基准模型的性能。

本研究仍存在一些不足，未来可在以下两方面进一步拓展：一是引入交通流量、工业排放、土地利用等多源动态外部数据，构建更完备的时空特征体系以增强模型对复杂城市大气系统的动态表征能力；二是将模型扩展至 $PM_{2.5}$、O_3、NO_2 等多污染物联合预测任务，通过图注意力机制建模污染物间的化学反应与传输耦合关系，揭示它们的相互作用机制，为区域复合污染协同防控提供更系统的科学依据。

参考文献(References)

[1] 中共中央、国务院. 关于完整准确全面贯彻新发展理念做好碳达峰碳中和工作的意见[N]. 人民日报，2021-10-25(6).

[2] 杨武霖，杨昊，周航，等.基于深度学习的空气污染多城市联合预测模型[J].软件导刊，2024，23(11)：107-115.

[3] Zhang Y，Bocquet M，Mallet V，et al. Real-time air quality forecasting，part I：history，techniques，and current status[J]. Atmospheric Environment，2012，60：632-655.

[4] Gao Z，Zhou X. A review of the CAMx，CMAQ，WRF-Chem and NAQPMS models：application，evaluation and uncertainty factors[J]. Environmental Pollution，2024，343：123183.

[5] 李颖若，韩婷婷，汪君霞，等.ARIMA 时间序列分析模型在臭氧浓度中长期预报中的应用[J].环境科学，2021，42(7)：3118-3126.

[6] 王凯文，李宏滨.基于多元线性回归的空气质量指数预测模型[J].信息记录材料，2024，25(2)：1-3，7.

[7] 杨思琪，赵丽华.随机森林算法在城市空气质量预测中的应用[J].统计与决策，2017，(20)：83-86.

[8] 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(1)：67-70，104.

[9] Kipf T N，Welling M. Semi-supervised classification with Graph Convolutional Networks[C/OL]// ICLR，2017，23：52.https://openreview.net/forum? id=SJU4ayYgl.

[10] LeCun Y，Bengio Y，Hinton G. Deep learning[J]. Nature，2015，521(7553)：436-444.

[11] Bronstein M M，Bruna J，LeCun Y，et al. Geometric deep learning：going beyond euclidean data[J]. IEEE Signal Processing Magazine，2017，34(4)：18-42.

[12] Veli č kovi ć P，Cucurull G，Casanova A，et al. Graph attention networks[C/OL]// ICLR，2018. https://openreview.net/forum? id=rJXMpikCZ.

[13] Qi Z，Wang T，Song G，et al. Deep air learning：interpolation，prediction，and feature analysis of fine-grained air quality[J]. IEEE Transactions on Knowledge and Data Engineering，2018，30(12)：2285-2297.

[14] Wang J，Tang J，Xu Z，et al. Spatiotemporal modeling and prediction in cellular networks：A big data enabled deep learning approach[C/OL]// IEEE INFOCOM，2017. https://doi.org/10.1109/INFOCOM.2017.8057090.

[15] Bai L，Wang J，Ma X，et al. Air pollution forecasts：an overview[J]. International Journal of Environmental Research and Public Health，2018，15(4)：780.

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

[17] Bengio Y，Simard P，Frasconi P. Learning long-term dependencies with gradient descent is difficult[J]. IEEE Transactions on Neural Networks，1994，5(2)：157-166.

作者简介：

王依卓(2000-)，女(汉族)，上海，硕士生。研究领域：空气质量预测与机器学习。本文通信作者。

李 林(1966-)，女(汉族)，营口，副教授，博士。研究领域：管理科学与工程，工业工程。

www.rjgcz.com