基于 LSTM 的空气质量历史关联模型研究

文丽霞* 郑乃瑞 王浩宇 邓小超
西南交通大学地球科学与环境工程学院 四川成都 610036

摘 要：空气质量关乎个人健康和社会的发展，科学有效的空气质量预测具有十分重要的意义。空气质量指数（AQI）具有时序性，以往的研究大多忽略了其在时间维度上的特性，也没有对污染物排放量与空气质量的关系进行深入的探讨。为此，本研究充分考虑 AQI 数据的时序特征，结合大气污染物排放量，采用长短期记忆网络（LSTM），建立基于时间序列的空气质量历史关联模型。提出 AQI 间接预测模式，即先预测空气质量分指数，再根据 AQI 计算公式得出当日 AQI 值。设计四组实验方案，将 AQI、空气质量分指数以及大气污染物排放量之间进行变量组合寻优，对比直接预测模式与间接预测模式，以及加入大气污染物排放量前后的预测效果。研究结果表明，间接预测模式可以达到相对更好的预测效果。

关键词：AQI 间接预测；时间序列预测；大气污染物排放量；LSTM

Study on Historical Correlation Model of Air Quality Based on LSTM

Lixia Wen, Nairui Zheng, Haoyu Wang, Xiaochao Deng
Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu,Sichuan,610036, China

Abstract: Air quality is crucial for personal health and social development, and scientifically accurate air quality prediction is of great significance. The Air Quality Index (AQI) exhibits temporal characteristics, but previous studies have mostly overlooked its temporal dimension and lacked in-depth exploration of the relationship between pollutant emissions and air quality. Therefore, this study fully considers the temporal characteristics of AQI data and combines them with atmospheric pollutant emissions to establish a time series-based historical correlation model for air quality using Long Short-Term Memory (LSTM) networks. An indirect prediction mode of AQI is proposed, where the sub-indices of air quality are first predicted, and the AQI value for the day is then calculated using the AQI formula. Four sets of experimental designs are implemented to optimize the variable combinations among AQI, sub-indices, and atmospheric pollutant emissions. The prediction effects of the direct prediction mode, indirect prediction mode, and the addition of atmospheric pollutant emissions are compared. The research results demonstrate that the indirect prediction mode achieves relatively better prediction performance.

Keywords: AQI indirect prediction; Time series prediction; The amount of air pollutants discharged; LSTM

引言

AQI 是非线性的时间序列[1]，其时间序列的特性使得过去的空气质量会对现在及将来的空气质量变化产生影响。因此可以根据历史 AQI 序列进行统计分析，捕捉其发展规律，从而预测出 AQI 在未来一段时间内的变化情况。目前常用的时间序列预测模型有差分整合移动平均自回归模型(ARIMA)、循环神经网络（RNN）和卷积神经网络（CNN）等。Chhikara 等[2]建立了基于 CNN-LSTM 的印度德里空气质量预测模型，结果显示 RMSE 为 221.682。Sethi 等[3]使用 ARIMA 模型预测印度古鲁格拉姆（Gurugram）的空气质量，在考虑污染物浓度以及气象参数的情况下，其预测精度 RMSE 为 66.8。以上方法在预测空气质量时，考虑到了 AQI 时序性。然而，这些方法都存在一定的局限性，如 ARIMA 通常不能捕捉非线性关系，预测精度在很大程度上受到其线性映射能力的限制；RNN 在长时间序列的应用中存在梯度消失和爆炸的问题。而 LSTM 作为一种深度学习模型，具有长时间记忆功能，改善了 RNN 中存在的长期依赖问题，常用于处理复杂的非线性时间序列。区域空气质量受气象、污染物排放和历史空气质量等因素综合影响。仅考虑单一要素进行预测既不能获得较高的准确性，也不足以帮助我们应对空气污染，多要素的预测方法有望发挥更大的价值。因此有必要根据 AQI 的时间序列特征以及大气污染物排放量综合预测当前空气质量。

一、数据来源与研究方法

1.数据来源

本研究淄博市张店区为研究对象，数据集包含 2017 年 1 月 1 日至 2020 年 12 月 31 日的日均 AQI，6 项污染物（PM_{10}、$PM_{2.5}$、SO_2、NO_2、CO、O_3）的日均浓度，以及张店区内记录在册的大型工业企业的大气污染物排放量（数据来源：淄博市生态环境局）。因 2018 年 11 月 27 日至 11 月 30 日，以及 2018 年 12 月 3 日（共 5 日）有异常值，删除处理后，数据量变为 1456×11 组。用 Python 中的 describe 函数得到数

据的均值、标准差、最大值以及最小值等基本统计特征（表1）。

表 1 观测数据基本特征

统计指标	数据总数	均值	标准差	最小值	最大值
AQI	1456	99.21	47.72	13	326
$PM_{2.5}$ ($\mu g/m^3$)	1456	58.02	37.95	4	276
PM_{10} ($\mu g/m^3$)	1456	106.94	56.2	4	396
SO_2 ($\mu g/m^3$)	1456	24.93	16.68	4	154
NO_2 ($\mu g/m^3$)	1456	43.64	17.84	8	116
CO (mg/m^3)	1456	1.24	0.58	0.2	6.7
O_3 ($\mu g/m^3$)	1456	115.6	58.39	6	289
SO_2 排放量 (t)	1456	1.87	0.6	0.52	5.34
NO_X 排放量 (t)	1456	6.95	2.12	2.6	16.09
颗粒物排放量 (t)	1456	0.48	0.31	0.13	4.45
总废气排放量 (亿 m^3)	1456	1.45	0.36	0.63	2.02

2.基于 LSTM 神经网络模型的构建

长短期记忆网络（Long Short-Term Memory，LSTM）是一种特殊的递归神经网络。与前馈神经网络相比，LSTM 可以利用时间序列对输入进行分析，通过在隐藏层中添加记忆单元来控制时间序列数据的记忆信息。LSTM 具有较长期的记忆功能，常作为时间序列预测模型被应用至空气质量预测、自然语言处理、文本识别和计算机视觉等多个领域[4]。

LSTM 通过在隐藏层中添加记忆单元和遗忘单元来控制时间序列数据的记忆信息，这种记忆单元被称之为门控系统。门控系统可以让序列中的信息选择性通过，有助于 LSTM 解决时间长期依赖问题。LSTM 由遗忘门、输入门和输出门组成。遗忘门用来加强记忆某些历史信息，同时对输入的信息进行过滤，即输入门的作用是决定让多少新的信息加入到细胞单元中。输入门来加强记忆某些历史信息，同时对输入的信息进行过滤，即输入门的作用是决定让多少新的信息加入到细胞单元中。输出门决定待输出的状态信息，生成输出信号至下一个序列单元。

LSTM 网络结构如图 1 所示，图中σ为激活函数，X_{t-1}（Cell）为 t-1 时刻的细胞状态，即 LSTM 的记忆存储单元（短期记忆），h_{t-1}为上一神经元的输出，即 t-1 时刻的隐藏层状态。将 t-1 时刻的细胞状态和隐藏层状态传入当前时刻，使用门控单元处理 t 时刻的细胞状态，选择性的记忆和遗忘信息，并更新细胞状态和隐藏层状态，达到更新记忆的目的。

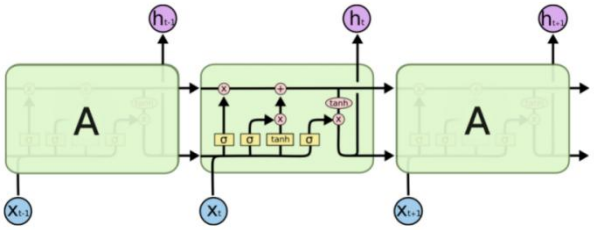

图 1 LSTM 网络结构图[5]

二、空气质量历史关联模型预测方法及结果分析

设计 4 组不同输入变量的实验方案，各实验方案及其对应的输入和输出变量具体见实验 1 至实验 4。选取 2017.1.1-2020.9.30 的数据作为模型的训练集和验证集，占比分别为 70%和 30%，以 2020.10.1-2020.12.31 的数据作为模型的测试集来评估模型的性能。实验结果采用平均绝对百分比误差（MAPE）、均方根误差（RMSE）以及平均绝对误差（MAE）来衡量。

1.实验 1 单变量直接预测

实验 1 采用直接预测模式，模型的输入变量仅 AQI 时间序列，输出变量为 AQI。同时以 ARIMA 模型和高斯隐马尔可夫模型（GHMM）作为对比模型。三个模型的预测结果见表 2。由表 2 可知，LSTM 神经网络模型的 AQI 预测精度明显优于 ARIMA 模型和 GHMM 模型。将 AQI 真实值与三个模型预测值进行对比（图 2），由图 2 可知，尽管 ARIMA 模型的预测精度数值小于 GHMM，但 ARIMA 模型缺乏对非线性序列的映射能力，导致模型的后期预测值呈直线变化。

模型	MAPE（%）	RMSE	MAE
ARIMA	33.67	49.95	34.31
GHMM	52.50	55.26	43.93
LSTM	20.75	26.48	19.06

表 2 AQI 预测精度对比

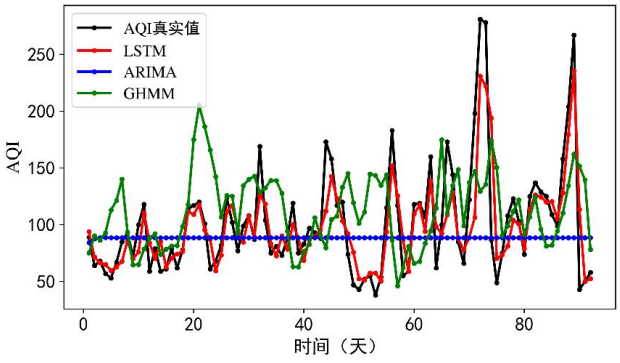

图 2 三个模型预测效果对比

2.实验 2 多变量间接预测

由于 AQI 是由空气质量分指数计算得出的，为了获得更加准确的 AQI，实验 2 通过间接预测模式预测 AQI。在预

测各分指数时，设计了不同的输入变量组合。预测出每日的各项 IAQI 后，计算出预测当日对应的 AQI。即实验 2 的模型输入变量为 IAQI 及其组合，输出变量为 IAQI。各项 IAQI 的最佳输入变量组合，对应的各分指数预测结果，以及计算出的 AQI 精度见表 3。从表 3 可知，六项 IAQI 中，$IAQI_{SO2}$ 的预测效果相对最好，其次为 $IAQI_{NO2}$，而 $IAQI_{O3}$ 的预测效果相对最差。比起实验 1，实验 2 采用间接预测模式预测的 AQI 精度更高。

表 3 分指数与 AQI 预测精度

预测对象	最佳变量组合方式	MAPE (%)	RMSE	MAE
$IAQI_{PM2.5}$	$IAQI_{PM2.5}$	25.23	26.92	18.50
$IAQI_{PM10}$	$IAQI_{PM10}$	22.67	22.20	16.86
$IAQI_{SO2}$	$IAQI_{SO2}$ + $IAQI_{PM10}$ + $IAQI_{PM2.5}$	27.00	6.27	4.80
$IAQI_{NO2}$	$IAQI_{NO2}$ + $IAQI_{SO2}$	19.49	15.85	12.31
$IAQI_{CO}$	$IAQI_{CO}$ + $IAQI_{NO2}$	31.01	10.34	7.46
$IAQI_{O3}$	$IAQI_{O3}$	40.42	14.93	10.93
AQI	/	18.00	25.39	16.85

3.实验 3 多变量直接预测

在实验 3 中，将大气污染物排放量与 AQI 进行组合，即模型输入变量为 AQI 与大气污染物排放量，输出为 AQI，得到的组合方式及相应的预测精度如表 4 所示。从表 3 可知，比起其他类型的输入组合，当输入数据为 AQI 与颗粒物排放量时，AQI 的预测精度相对最高。

表 4 实验 3 条件下的 AQI 预测精度

变量组合方式	MAPE (%)	RMSE	MAE
(a) AQI + SO_2 排放量	23.57	28.90	21.01
(b) AQI + NO_X 排放量	24.05	31.77	22.99
(c) AQI + 颗粒物排放量	20.37	27.50	19.96
(d) AQI + 总废气排放量	20.77	29.04	19.99
(e) AQI +（NOx + SO_2 + 颗粒物） 排放量	21.51	30.15	20.57

4. 实验 4 多变量间接预测

该实验中，考虑模型的输入变量组合为分指数与大气污染物排放量相结合，输出变量为分指数。结合实验 2，得到各分指数所对应的最佳输入变量组合及相应的预测精度（表 5）。同样地，根据分指数的预测值计算出 AQI，再将计算出的 AQI 与真实值进行对比计算，得出实验 4 条件下的 AQI 预测精度（表 5）。结合表 3 和表 5 可知，在预测 IAQI 时，采用 IAQI 与大气污染物排放量的变量组合预测后，除 $IAQI_{PM2.5}$、$IAQI_{CO}$、$IAQI_{O3}$ 以外，其余 IAQI 的预测精度都有所提升。

表 5 实验 4 条件下的分指数与 AQI 预测精度

预测对象	最优变量组合方式	MAPE (%)	RMSE	MAE
$IAQI_{PM2.5}$	$IAQI_{PM10}$ + 颗粒物排放量	24.37	27.56	18.88
$IAQI_{PM10}$	$IAQI_{PM2.5}$ + 颗粒物排放量	21.30	20.83	16.10
$IAQI_{SO2}$	$IAQI_{SO2}$ + $IAQI_{PM10}$ + $IAQI_{PM2.5}$ + SO_2 排放量	26.82	6.42	4.80
$IAQI_{NO2}$	$IAQI_{NO2}$ + $IAQI_{SO2}$ + NO_x 排放量	21.15	14.99	12.10
$IAQI_{CO}$	$IAQI_{CO}$ + $IAQI_{NO2}$ + NO_x 排放量	34.11	11.80	8.82
$IAQI_{O3}$	$IAQI_{O3}$ + SO_2 排放量	43.73	17.53	12.68
AQI	/	**19.37**	**26.56**	**18.58**

5.各实验结果对比分析

将实验 1 至实验 4 的结果进行整理，得到图 3，其中实验 3 的数据选择输入变量为颗粒物排放量和 AQI 组合的预测结果。输入变量仅 AQI 时，LSTM 神经网络模型的预测效果优于 ARIMA 模型和 GHMM 模型；直接预测 AQI 与间接模式预测 AQI 相比（实验 1 对比实验 2，实验 3 对比实验 4），间接预测模式预测的 AQI 精度更高；对比加入大气污染物排放量前后的实验（实验 1 对比实验 3，实验 2 对比实验 4），在加入大气污染物排放量后，AQI 的预测精度没有明显提升。

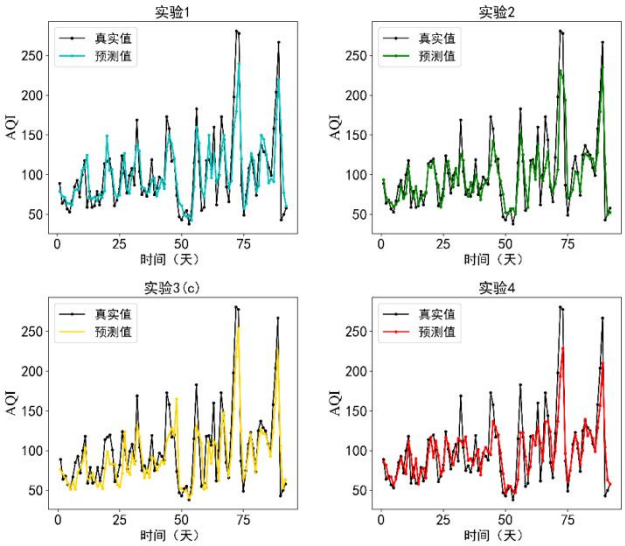

图 3 各实验条件下的 AQI 预测效果

三、总结与讨论

空气质量与大气污染物排放量以及历史空气质量等因素有关。本研究根据空气质量具有显著时间序列关系的特征，采用LSTM神经网络模型，建立了空气质量历史关联模型。在AQI预测方式上，提出了间接预测模式的新思路，更符合AQI计算原理。本研究以日均AQI、IAQI和大气污染物排放量数据，采用传统直接预测模式和间接预测新模式，进行了大量的对比实验。研究结果显示，在预测AQI时，基于空气

质量历史关联模型，采用间接预测模式的预测效果更好。

参考文献：

[1] Li H, Wang J, Li R, et al. Novel analysis‐forecast system based on multi-objective optimization for air quality index[J]. Journal of Cleaner Production. 2019, 208: 1365-1383.

[2] Chhikara P, Tekchandani R, Kumar N, et al. Federated Learning and Autonomous UAVs for Hazardous Zone Detection and AQI Prediction in IoT Environment[J]. IEEE Internet of Things Journal. 2021, 8(20): 15456-15467.

[3] Sethi J K, Mittal M. Analysis of Air Quality using Univariate and Multivariate Time Series Models[Z]. Noida, India: 2020823-827.

[4] Yong Yu, Xiaosheng Si, Changhua Hu, Jianxun Zhang; A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. NEURAL COMPUTATION. 2019, 31(7):1235-1270.

[5] Christopher O. Understanding LSTM Networks[GB/OL]. (2015-08-27),[2022-05-25]. http://colah.github.io/posts/2015-08-Understanding-LSTMs/.