Advances in Applied Mathematics 应用数学进展, 2023, 12(12), 5057-5071
Published Online December 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.1212497

Hans汉斯

基于LSTM多步预测模型的空气质量预测与预警

王子航，王子睿，柳卓行，杨　娟*

江汉大学人工智能学院，湖北　武汉

收稿日期：2023年11月11日；录用日期：2023年12月4日；发布日期：2023年12月15日

摘　要

本文首先通过斯皮尔曼相关性分析及灰色关联分析，以两方式相对比的方式筛选出与PM2.5浓度变化有关的因素，通过随机森林回归算法得出因素对PM2.5浓度的影响程度。然后将LSTM神经网络模型与多步预测模型相结合，构建LSTM多步预测模型，并设置步长用于预测PM2.5的值，根据均方根误差检验对模型效果进行评估。最后，将数据集带入LSTM多步预测模型并设置步长用以预测AQI的值。

关键词

灰色关联分析，随机森林算法，LSTM多步预测模型，均方根误差

Air Quality Prediction and Early Warning Based on LSTM Multi-Step Prediction Model

Zihang Wang, Zirui Wang, Zhuoxing Liu, Juan Yang*

School of Artificial Intelligence, Jianghan University, Wuhan Hubei

Received: Nov. 11th, 2023; accepted: Dec. 4th, 2023; published: Dec. 15th, 2023

Abstract

In this paper, firstly, the factors related to the changes of PM2.5 concentration are screened out by Spearman correlation analysis and grey relation analysis through two relative comparisons, and the degree of influence of the factors on the concentration of PM2.5 is derived by random forest regression. Then the LSTM neural network model was combined with the multi-step prediction model to construct an LSTM multi-step prediction model and the step size was set for predicting the value of PM2.5, and the model effect was evaluated according to the root mean square error

*通讯作者。

文章引用: 王子航, 王子睿, 柳卓行, 杨娟. 基于 LSTM 多步预测模型的空气质量预测与预警[J]. 应用数学进展, 2023, 12(12): 5057-5071. DOI: 10.12677/aam.2023.1212497

test. Finally, the dataset was brought into the LSTM multi-step prediction model and the step size was set to predict the value of AQI.

Keywords

Grey Relation Analysis, Random Forest Algorithm, LSTM Multi-Step Prediction Model, Root Mean Square Error

1. 引言

现代社会中，空气污染问题已经成为一个让全人类不得不重视的议题。空气中的污染物会对人类身体健康、生态环境和社会经济发展造成巨大的不利影响。为了减轻空气污染对人类生产生活造成的影响，环境空气质量指数(AQI)应运而生。通过 AQI 指数，可以掌握当日当地较为可靠的空气污染情况。与此同时，PM2.5 是当下对人体危害最严重的空气污染物。AQI 和空气中 PM2.5 浓度含量能对生产、环境保护等方面产生重要影响。因此，能否精准预测 PM2.5 浓度和 AQI 指数以达到保护环境的目的，成为很多研究者重点关注的课题。

2. 因素筛选及对 PM2.5 影响程度分析

在这部分，我们将筛选出与 PM2.5 浓度变化有关的因素，并量化分析筛选出的因素对 PM2.5 浓度影响的程度。分析过程由以下步骤构成：

步骤一：数据预处理，并通过正态性校验判断是否适用于斯皮尔曼相关性分析。

步骤二：对比斯皮尔曼相关性分析及灰色关联分析结果，筛选出相关因素。

步骤三：通过随机森林回归对筛选出的因素进行排名，确定影响程度。

这部分的处理流程见图 1。

2.1. 数据预处理

2.1.1. 数据问题描述

本文采集了 2015 年 1 月 1 日至 2023 年 4 月 29 日的 8 类污染物浓度数据：AQI、质量等级、PM10、O_3、SO_2、PM2.5、NO_2、CO，以及 5 类气象数据：降水量、平均气压、平均 2 分钟风速、平均气温、平均相对湿度。在原始数据中，大部分数据正常，但有些数据异常，有极少数数据为空值。

2.1.2. 预处理过程

本文首先对数据进行清洗[1]。去除异常值，对于缺失值，以三倍标准差的方法对数据进行填充。接着通过归一化处理将所有变量的值限于[0, 1]之间，此种方法可以将数值的绝对值转换为相对值，消除指标间的量纲影响，使各指标处于同一数量级，方便进行各指标间的综合对比评价。归一化结果见表 1。

接着，对数据进行正态性检验，判断其是否适用于皮尔森分析法以及斯皮尔曼分析法。皮尔森相关性分析需满足变量存在正态分布。但经检验，所有大气参数及污染物结果全部呈现显著性($\rho > 0.05$)，并不满足原假说，不服从正态分布。故本文采用斯皮尔曼相关系数分析法。

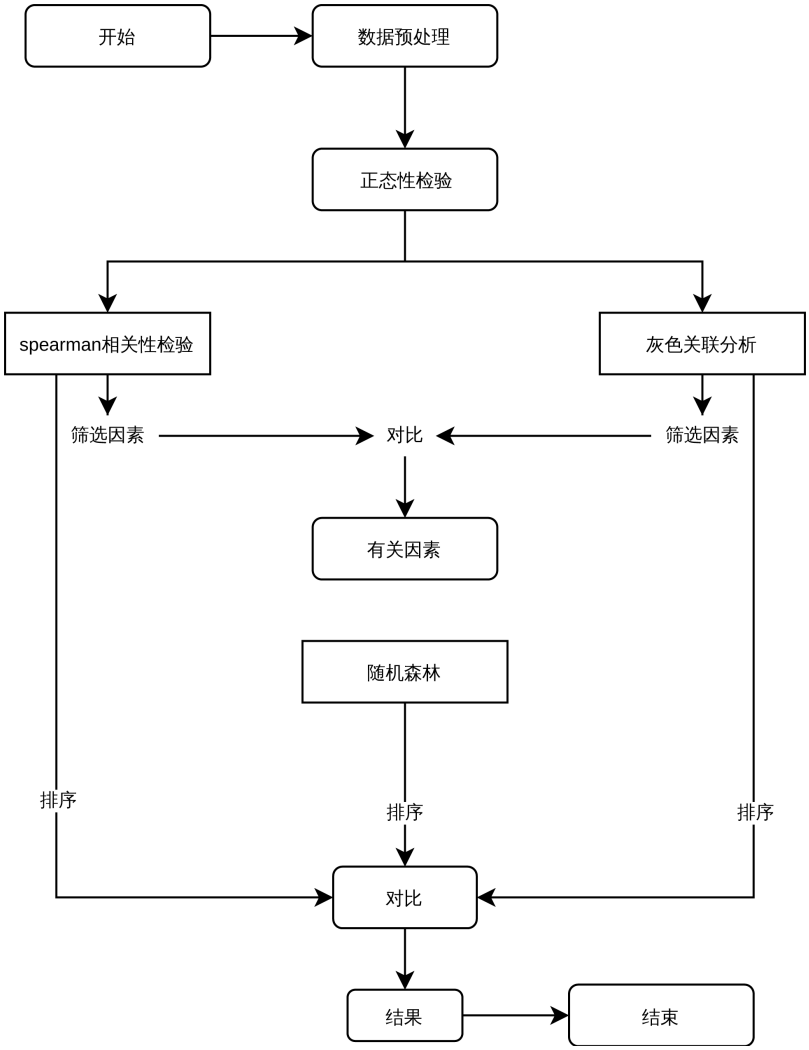

Figure 1. Flow chart of factor screening and its impact extent on PM2.5
图 1. 因素筛选及对 PM2.5 影响程度流程图

Table 1. Normalized results
表 1. 归一化结果

PM2.5	PM10	O_3	SO_2	CO	NO_2	降水	气压	风速	气温	湿度
0.16204	0.357357	0.286821	0.579710	0.304347	0.45	0	0.9597544	0.052	0.0424886	0.35483871
0.29231	0.432432	0.077519	0.797101	0.434782	0.625	0	0.9413369	0.032	0.068285	0.44086021
0.38737	0.546546	0.170542	0.6086	0.521739	0.70833	0	0.9178035	0.036	0.0644916	0.61290322
0.468	0.6486	0.1434	1.00	0.6521	0.9416	0	0.891200	0.032	0.086494	0.6272401

……

2.2. 模型建立

模型一. 斯皮尔曼相关系数分析

斯皮尔曼相关系数分析法被定义成等级变量之间的皮尔逊相关系数[2]。相关系数计算公式为：

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

但在实际应用中，变量与变量间的连结通常没有相关联系，基于此可通过更为简易的步骤计算相关系数。公式如下：

$$1 - \frac{6\sum di^2}{n\left(n^2 - 1\right)} \tag{2}$$

根据此模型，以大气参数及大气污染物为变量构建热力图判断其相关关系。

模型二. 灰色关联分析

灰色分析方法，是一种定量描述系统中因素对该系统发展态势的量化比较方法，其基本思想是通过序列几何曲线的相似程度来确定系统中因素对系统发展的联系程度。曲线相似程度越高，则系统中相应序列的关联度越大；相似程度越小，关联度也越小[3]。

经过分析可知，附件中各个变量序列的量纲不同，如果不对参数进行处理，直接分析，分析结果则会受到影响。因此在分析之前，需要进行去量纲处理。公式如下：

$$\eta_i(k) = \frac{\min_i \min_k \left|x_i(k) - x_0(k)\right| + \rho \max_i \max_k \left|x_i(k) + x_0(k)\right|}{\left|x_i(k) - x_0(k)\right| + \rho \max_i \max_k \left|x_i(k) - x_0(k)\right|} \tag{3}$$

在上式中，$\left|x_i(k) - x_0(k)\right|$ 为 $x_0(k)$ 和 $x_i(k)$ 第 k 个点的绝对误差；$\min_i \min_k \left|x_i(k) - x_0(k)\right|$ 为两级最小差；ρ 为分辨率，$0 < \rho < 1$，一般取 $\rho = 0.5$；ρ 越大，分辨率越小，ρ 越小，分辨率越大。式中的 $\min_i \min_k$，$\max_i \max_k$ 分别为大气参数及污染物样本中的最小值与最大值。

计算关联度的公式如下：

$$r_i = \frac{1}{n}\sum_{k=1}^{n} \eta_i(k) \tag{4}$$

其中，r_i 为 x_i 对 x_0 的关联度。

模型三. 随机森林回归

随机森林的算法步骤如下：

1) 从整合后的表格中抽取训练集。进行多轮抽取，从每次抽取的原始样本中使用 Bootstrapin 的方法抽取 X 个训练样本；

2) 使每个训练集都得到一个样本；

3) 针对回归问题，这里不作预测，而是通过算法得到特征重要性。

2.3. 模型求解

2.3.1. 因素筛选

方法一：采用斯皮尔曼相关系数分析法

以大气参数及大气污染物为变量构建热力图判断其间相关关系，斯皮尔曼相关性分析可以判断两变量间是否存在统计上的相关性，相关系数绝对值越大，说明两值其间的关系越紧密。基于此，本文通过热力图从宏观角度呈现大气参数、大气污染物与 PM2.5 间的关联程度。

其热力图结果如图 2 所示。

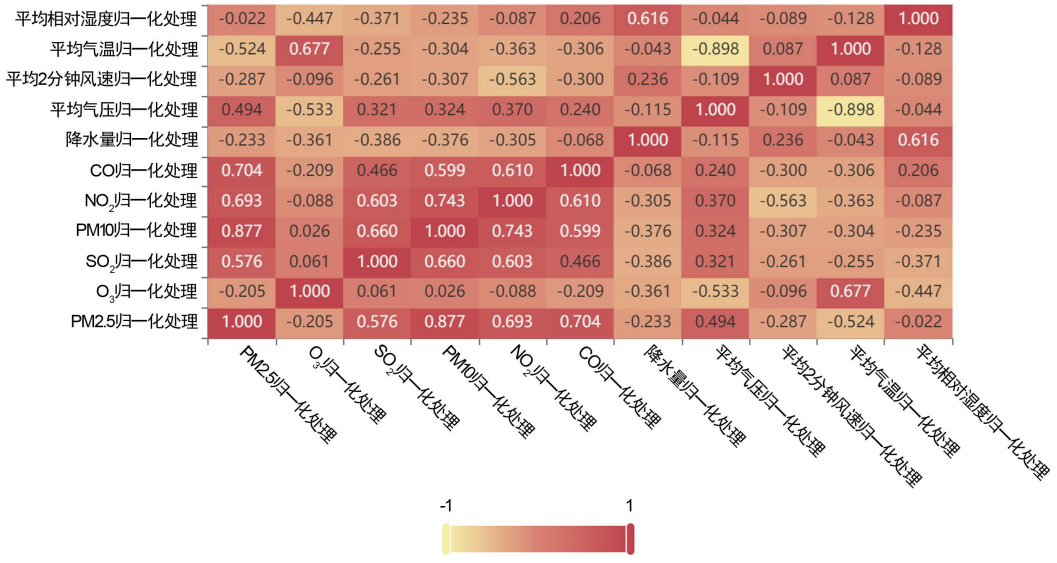

Figure 2. Heat map
图 2. 热力图

根据相关系数绝对值的大小进行排序，可得到以下结果，见表 2。

Table 2. Spearman correlation coefficient table
表 2. 斯皮尔曼相关系数表

PM10	CO	NO_2	SO_2	气温	气压	风速	降水	O_3	湿度
0.877	0.704	0.693	0.576	−0.524	0.494	−0.287	−0.233	−0.205	−0.022

可得出与 PM2.5 浓度变化有关的因素有：PM10、CO、NO_2、SO_2 以及气温。其中 PM2.5 与 PM10、CO、NO_2、SO_2 间具有强正相关，气温与 PM2.5 间存在强负相关。五个因素对 PM2.5 的影响排名如下：

PM10 > CO > NO_2 > SO_2 > 气温

方法二：采用灰色关联分析

通过对问题一的分析，可知需要建立一个 PM2.5 含量与气象参数和空气污染物浓度的相关模型。灰色关联分析是一种定量描述系统中因素对该系统发展态势的量化比较方法。它可以从几何角度出发，根据固定量列和对比数据列的相似程度来判断参考数据和比较数据间是否存在紧密联系[2]。

因此，选择灰色分析方法确定气象参数和空气污染物浓度对空气中 PM2.5 含量的影响程度。将 $\rho = 0.5$ 代入公式，用 MATLAB 编程可计算得到 PM2.5 与 PM10、CO、NO_2、SO_2、气温、风速、降水、O_3、湿度间的关联系数。部分结果见表 3。

Table 3. Grey relational coefficient table
表 3. 关联系数表

	PM10	SO_2	CO	NO_2	O_3	风速	气温	降水	湿度	气压
$\eta(1)$	0.800302	0.544872	0.634567	0.778465	0.755249	0.385302	0.819626	0.807044	0.72173	0.719116
$\eta(2)$	0.699513	0.497628	0.600479	0.778265	0.631073	0.435161	0.657634	0.690592	0.770972	0.781123
$\eta(3)$	0.697524	0.693189	0.609058	0.788209	0.563469	0.485246	0.587294	0.607629	0.689169	0.758546
$\eta(4)$	0.606112	0.484673	0.513721	0.731203	0.516349	0.541814	0.533995	0.566993	0.758874	0.734993

Continued

$\eta(5)$	0.509884	0.591528	0.694882	0.877329	0.482308	0.587041	0.51295	0.541016	0.791808	0.941167
$\eta(6)$	0.398784	0.724863	0.636296	0.922382	0.383845	0.834949	0.419654	0.39901	0.79549	0.515409
……	……	……	……	……	……	……	……	……	……	
$\eta(3039)$	0.681256	0.953647	0.731681	0.858704	0.845658	0.385978	0.96783	0.865718	0.504648	0.909916
$\eta(3040)$	0.686514	0.965865	0.702018	0.771482	0.81773	0.391695	0.888504	0.882195	0.487701	0.885877
$\eta(3041)$	0.651197	0.874803	0.968119	0.759287	0.759765	0.401272	0.828482	0.990714	0.786943	0.553698

接着计算 PM2.5 与 PM10、CO、NO_2、SO_2、气温、风速、降水、O_3、湿度间的关联度，结果见表 4。

Table 4. Grey relation degree table
表 4. 关联度表

PM10	SO_2	Co	NO_2	O_3	风速	气温	降水	湿度	气压
0.88858	0.888244	0.855305	0.764862	0.688445	0.82769	0.813349	0.776017	0.507679	0.411858

显然可得与 PM2.5 浓度变化有关的因素有：PM10、CO、NO_2、SO_2、气温、风速。各个因素对 PM2.5 的影响排名如下：

PM10 > SO_2 > CO > 风速 > 气温 > NO_2

将方法一和方法二的结果综合对比可得：与 PM2.5 浓度变化有关的因素有：PM10、CO、NO_2、SO_2 以及气温。

2.3.2. 对 PM2.5 影响程度分析

采用随机森林回归对各数据进行分析，此处运用随机森林模型并不是为了预测，而是通过算法找寻各因素对 PM2.5 的重要性程度。将结果与前两种方法进行对比，从而得到最为精准的结果。相较于从统计学出发斯皮尔曼相关性分析及从几何角度出发的灰色关联分析，通过随机森林回归可以从自身元素属性出发，避免决策单一化所造成的影响，可以有效提高结果准确度。结果见图 3。

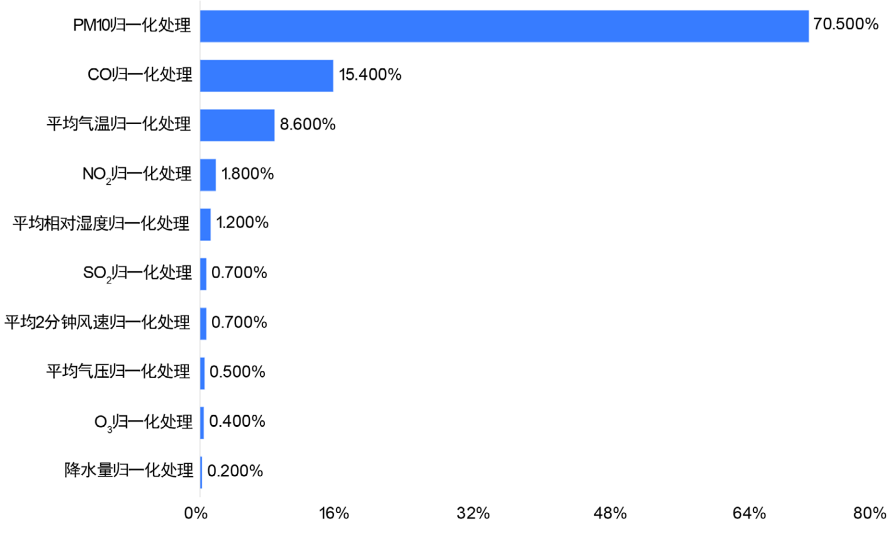

Figure 3. Ranking of importance of each factor on PM2.5
图 3. 各因素对 PM2.5 特征重要性排行

由图 3 可得：PM10、CO、气温对 PM2.5 浓度有强影响。NO_2、SO_2 对 PM2.5 浓度有较强影响。其他量影响较弱。将斯皮尔曼相关性分析、灰色关联性分析以及随机森林回归所得结果相对比，可得出以下结论：与 PM2.5 浓度变化有关的因素有：PM10、CO、NO_2、SO_2 以及气温。其中 PM10、CO、气温对 PM2.5 浓度有强影响。NO_2、SO_2 对 PM2.5 浓度有较强影响。

3. 建立 PM2.5 浓度多步预测模型

3.1. 模型准备

在这部分，我们将构建 PM2.5 浓度多步预测模型，并通过用均方根误差(RMSE)对 3 步、5 步、7 步、12 步预测效果进行评估。由以下步骤构成：

步骤一：数据预处理，将原始数据表格中年月日合并，将质量等级通过数字化呈现。

步骤二：设置步长，在 Python 中导入筛选因子，通过 LSTM 网络进行编译。

步骤三：通过均方根误差对结果进行检验，将检验结果可视化。

处理流程图如图 4 所示。

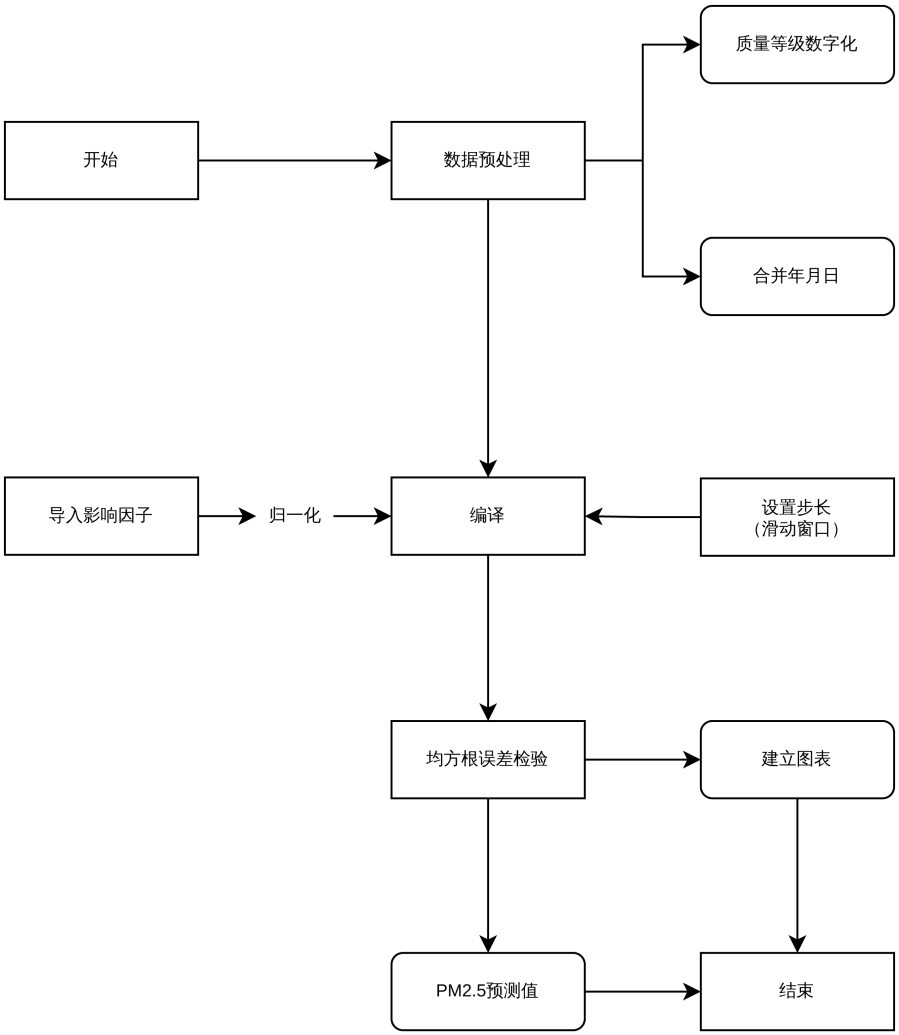

Figure 4. Flowchart of PM2.5 multi-step prediction model and its visualization
图 4. PM2.5 多步预测模型及可视化流程图

3.2. 模型建立

模型一. 多步预测模型

多步预测算法是一种基于时间序列数据的预测方法，可以预测多个未来时刻的数值。以下是一个简单的多步预测算法的伪代码。输入时间序列数据 x，以及预测的步长 n，将数据集分成训练集和测试集，使用训练集进行模型训练，得到预测模型对测试集进行预测，得到预测结果。

将测试结果中的最后 $n-1$ 个值作为输入数据，再次使用预测模型进行预测，得到下一个时刻的预测值 $y\_\{\text{test}+n\}$。将 $y\_\{\text{test}+n\}$ 加入预测结果 y_pred 中重复以上步骤，直到得到 n 个预测值再返回预测结果。在实现上述算法时，需要选择合适的模型和相关参数，如何选择模型和参数，需要根据具体的情况进行分析和实验。同时，还需要进行数据预处理、特征工程、模型训练等步骤才能实现一个高效的多步预测算法。简而言之，就是采用滑动窗口来进行预测。

以步长为 3 为例，具体的预测流程图见图 5。

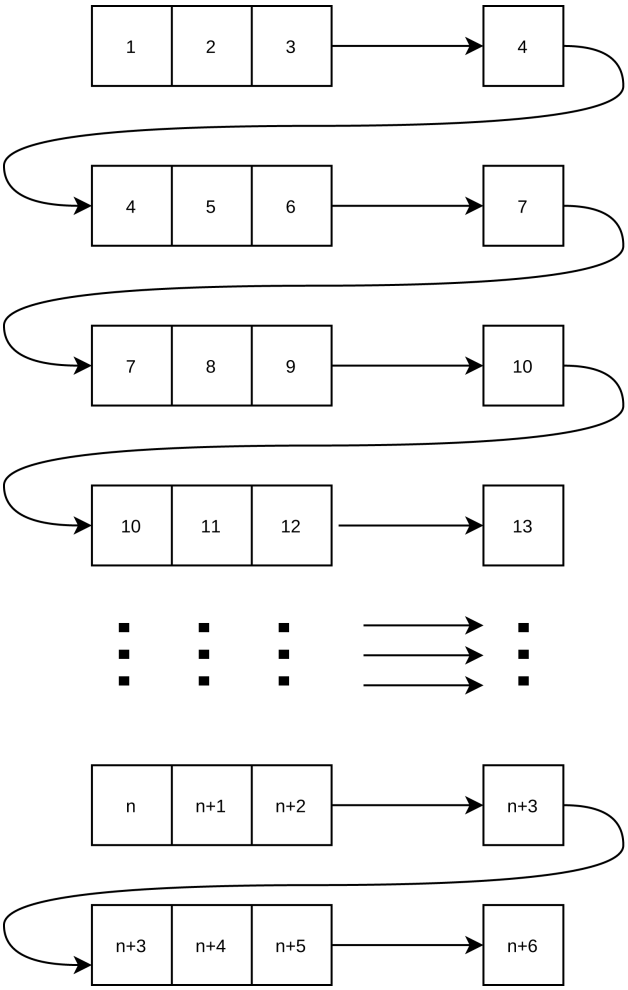

Figure 5. Step prediction flowchart (taking a step of three as an example)
图 5. 步长预测流程图(以步长为三为例)

模型二. LSTM 循环神经网络

传统的神经网络(BP 神经网络)分为以下几个层次：输入层、隐含层和输出层。每一层中都有若干个

神经元，每一层的神经元如同人脑中的突触般和下一层的神经元连接[2]。输入信号从输入层传入，经过隐含层的处理，进而从输出层输出。将输出层的数据和实际数据相比较，如果数据不相符，则进入反向传输，将误差分摊给隐含层的各个单元，从而建立一个稳定的模型。

但在前面的讨论中，PM2.5 的值不仅与其他参数有关，还和 PM2.5 的历史数据有关，但 BP 神经网络不具有“记忆性”，况且利用多步预测，预测误差会随步长的累计而增大，而 LSTM 网络(长短时记忆神经网络)能消除对数据的长期依赖。因此采用 LSTM 建立模型和预测数据。

LSTM 模型能够实现“记忆性”，又和普通的 RNN 模型不同，能够避免一般 RNN 模型对数据的长期依赖问题，而 LSTM 能够累计较远节点间的长期联系。因此选择 LSTM 模型对本问进行分析。LSTM 通过四个独特的结构实现以上功能，分别为记忆单元、输入门、输出门和遗忘门，而且各个结构之间采用特殊的方式进行连接。

输入门：新数据(第 2 节所得的影响因素数据)经由输入门传入 LSTM 中，输入门经过数据处理，决定保存新数据中的某些数据，此后被处理过的数据被存入记忆单元。

遗忘门：控制记忆单元中要遗忘哪些数据，从而更新数据，解决对数据的长期依赖问题。

输出门：选择性输出记忆单元中的数据(PM2.5 预测值)，决定数据是被直接输出还是传递至下一层 LSTM。回到循环起点，开启下一轮循环。

可以用简单表示 LSTM 对数据的处理过程，见图 6。

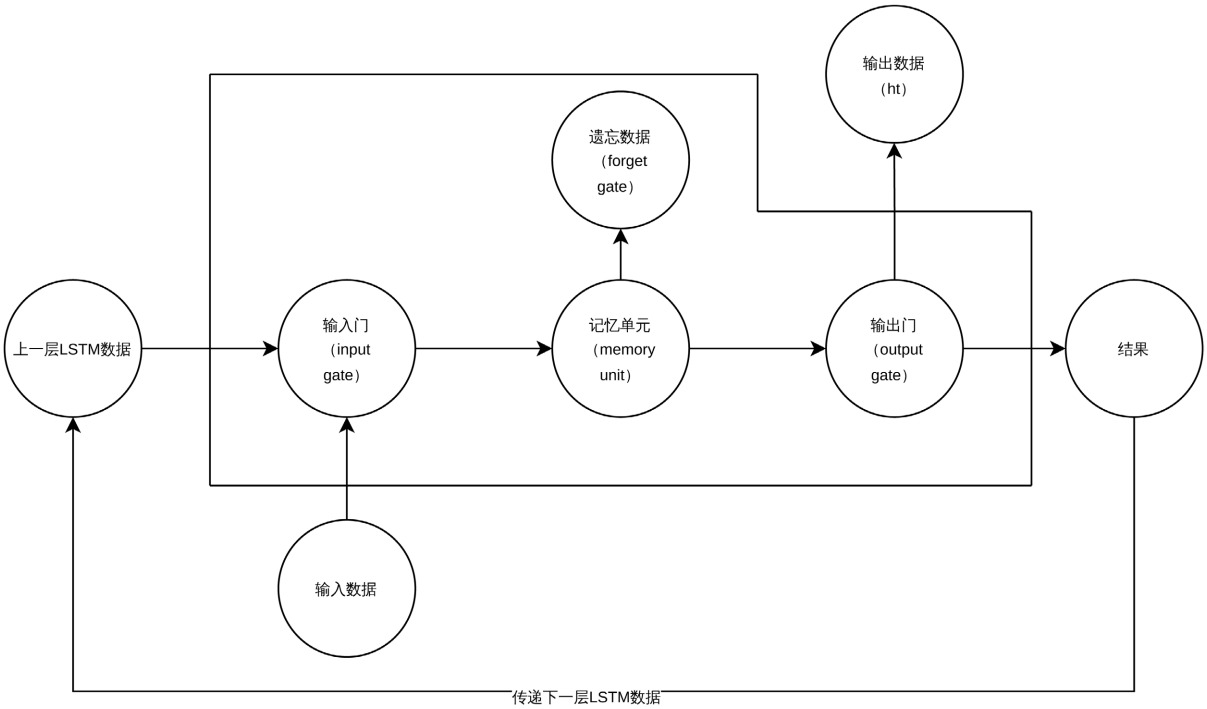

Figure 6. Flowchart of LSTM data processing
图 6. LSTM 数据处理流程图

检验方法：均方根误差检验

均方根误差与生活中常见的方差、标准差不同。其关系如图 7 所示。

均方误差的量纲通常与数据量纲不同，通常需要进行数据标准化，难以直观反映其间离散程度，故在均方误差上开二次方根，可得到均方根误差

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(\mathrm{Observed}_t-\mathrm{Predicted}_t\right)^2} \tag{5}$$

式中的 N 指的是样本总量，而 Predicted 及 Observed 分别指代 PM2.5 的预测值与真实值。

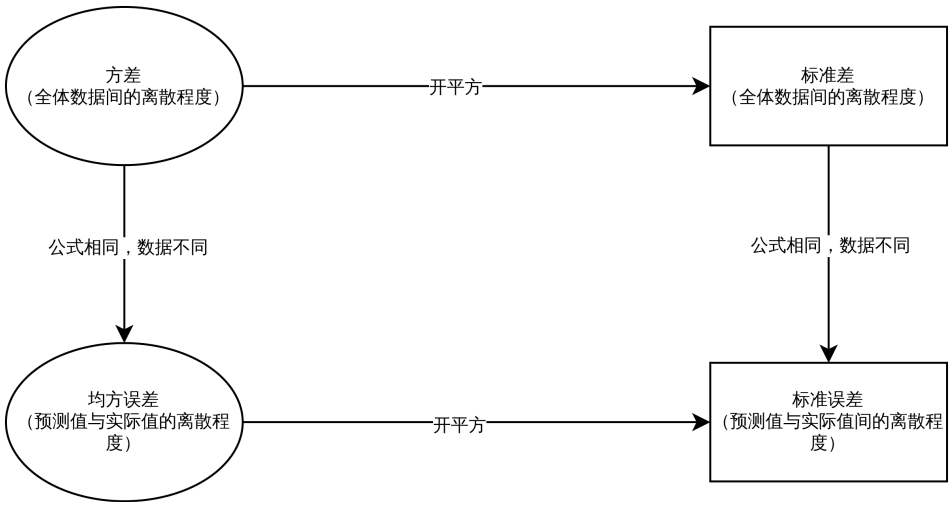

Figure 7. Relationship diagram of variance, standard deviation, mean square error, standard error
图 7. 方差、标准差、均方误差、标准误差关系图

均方误差通常用以衡量真实值与预测值间的误差。本文采用的是基于标准化数据的均方根误差，故均方根处在[0, 1]之间，越接近 0，说明模型的预测能力越好。

总结：由于多步预测模型预测误差会随步长的累计而增大，而 LSTM 神经网络能消除多步预测模型的误差。故本文选择将两模型相合并，构建 LSTM 多步预测模型。

3.3. 模型求解

首先进行数据预处理，将年月日合并为日期。将质量等级(优、良)以数据[0, 6]来体现。

本文使用 LSTM 循环神经网络进行多步预测，并通过均方根误差，对结果进行检验。多步预测存在随步数的累计，预测结果越不精准的问题。而相较于其他神经网络，LSTM 循环神经网络能够消除对累计数据的长期依赖，得到更为精准的预测结果[4]。

LSTM 模型能够实现“记忆性”，又和普通的 RNN 模型不同，能够避免一般 RNN 模型对数据的长期依赖问题，因此成为分析本问题的首选。经过 python 中的 Jupyter 进行编译，其结果见表 5~9。

Table 5. Table of RMSE for multi-step prediction (PM2.5)
表 5. 多步预测均方根误差表(PM2.5)

预测步长	3 步预测	5 步预测	7 步预测	12 步预测
RMSE	19.36	18.15	18.19	18.42

Table 6. Three-step prediction results (PM2.5)
表 6. 三步步长预测结果(PM2.5)

日期(年/月/日)	2023/4/30	2023/5/1	2023/5/2	2023/5/3	2023/5/4	2023/5/5
PM2.5	41.10	36.09	43.41	40.08	30.35	29.55

Continued

日期(年/月/日)	2023/5/6	2023/5/7	2023/5/8	2023/5/9	2023/5/10	2023/5/11
PM2.5	31.13	35.81	37.69	42.19	43.15	42.92

Table 7. Five-step prediction results (PM2.5)
表 7. 五步步长预测结果(PM2.5)

日期(年/月/日)	2023/4/30	2023/5/1	2023/5/2	2023/5/3	2023/5/4	2023/5/5
PM2.5	31.92	23.95	33.55	32.01	24.29	23.80
日期(年/月/日)	2023/5/6	2023/5/7	2023/5/8	2023/5/9	2023/5/10	2023/5/11
PM2.5	26.35	30.05	31.19	34.68	35.93	36.54

Table 8. Seven-step prediction results (PM2.5)
表 8. 七步步长预测结果(PM2.5)

日期(年/月/日)	2023/4/30	2023/5/1	2023/5/2	2023/5/3	2023/5/4	2023/5/5
PM2.5	37.43	29.64	34.49	32.04	24.58	24.37
日期(年/月/日)	2023/5/6	2023/5/7	2023/5/8	2023/5/9	2023/5/10	2023/5/11
PM2.5	24.87	27.30	27.83	30.77	32.52	34.54

Table 9. Twelve-step prediction results (PM2.5)
表 9. 十二步步长预测结果(PM2.5)

日期(年/月/日)	2023/4/30	2023/5/1	2023/5/2	2023/5/3	2023/5/4	2023/5/5
PM2.5	47.04	35.03	39.26	36.23	26.66	27.55
日期(年/月/日)	2023/5/6	2023/5/7	2023/5/8	2023/5/9	2023/5/10	2023/5/11
PM2.5	27.96	32.67141	36.21	39.66	39.57	40.05

接着，本文对预测数据进行了均方根检验，对模型的效果进行评估。经 Python 编译可视化结果如图 8 所示，由于长度问题，此处仅展示 3 步长的训练集与测试集的对比图，见图 8，红线代表真实值，蓝线代表预测值。经过多次迭代后，训练集和测试集呈高度拟合。可见预测效果较为准确。

4. 构建 AQI 多步预测模型

4.1. 模型准备

我们的工作：构建 AQI 多步预测模型，并通过用均方根误差(RMSE)对 3 步、5 步、7 步、12 步预测效果进行评估。并给出每天空气质量的预警等级颜色。由以下几个步骤构成：

步骤一：设置滑动步长，在 Python 中导入预测所需库，以 LSTM 神经循环网络进行编译。

步骤二：通过均方根误差对结果进行检验，将检验结果可视化。根据 AQI 值判断预警等级。

流程图如图 9 所示。

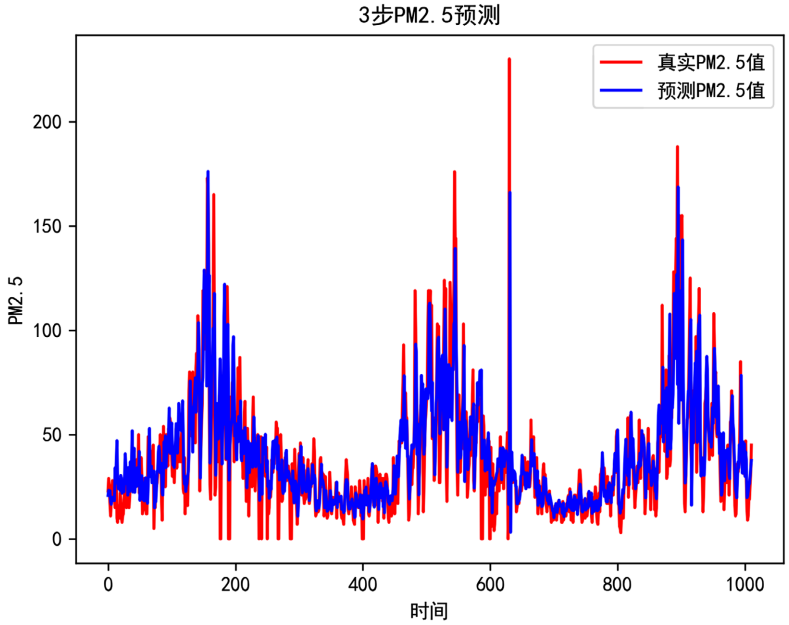

Figure 8. Visualization of test set and its predicted results (PM2.5)
图 8. 测试集及其预测结果的可视化(PM2.5)

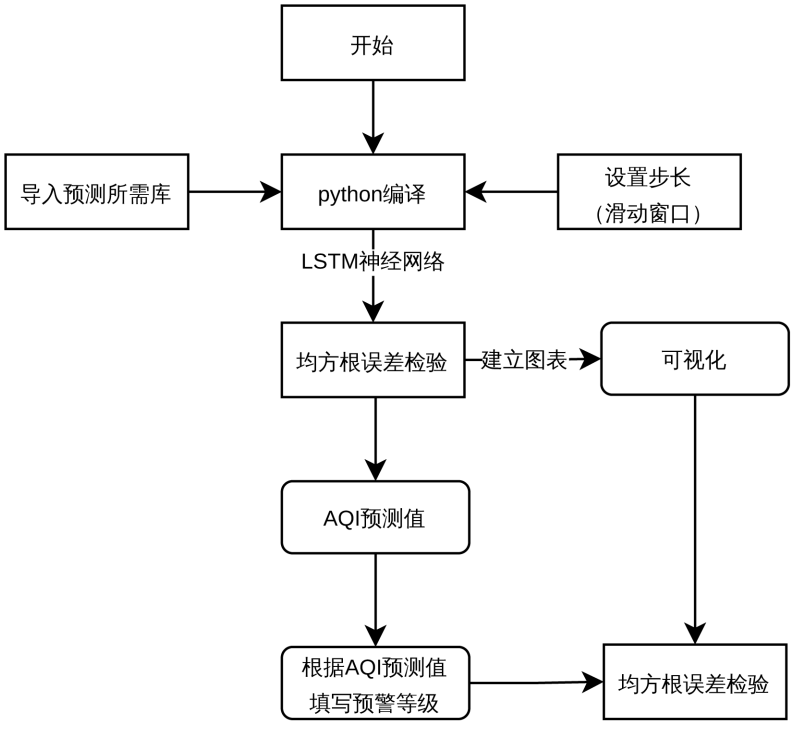

Figure 9. Flowchart of AQI multi-step prediction model and its visualization
图 9. AQI 多步预测模型及可视化流程图

4.2. 模型建立

均方根检验

在第 3 节中已有表述，此处不加赘述。

LSTM 循环神经网络多步预测模型

根据前文建立的模型，我们构建 LSTM 循环神经网络多步预测模型。多步预测算法是一种基于时间序列数据的预测方法，可以预测多个未来时刻的数值。简而言之，就是采用滑动窗口来进行预测。

以步长为 5 举例，具体流程图见图 10。

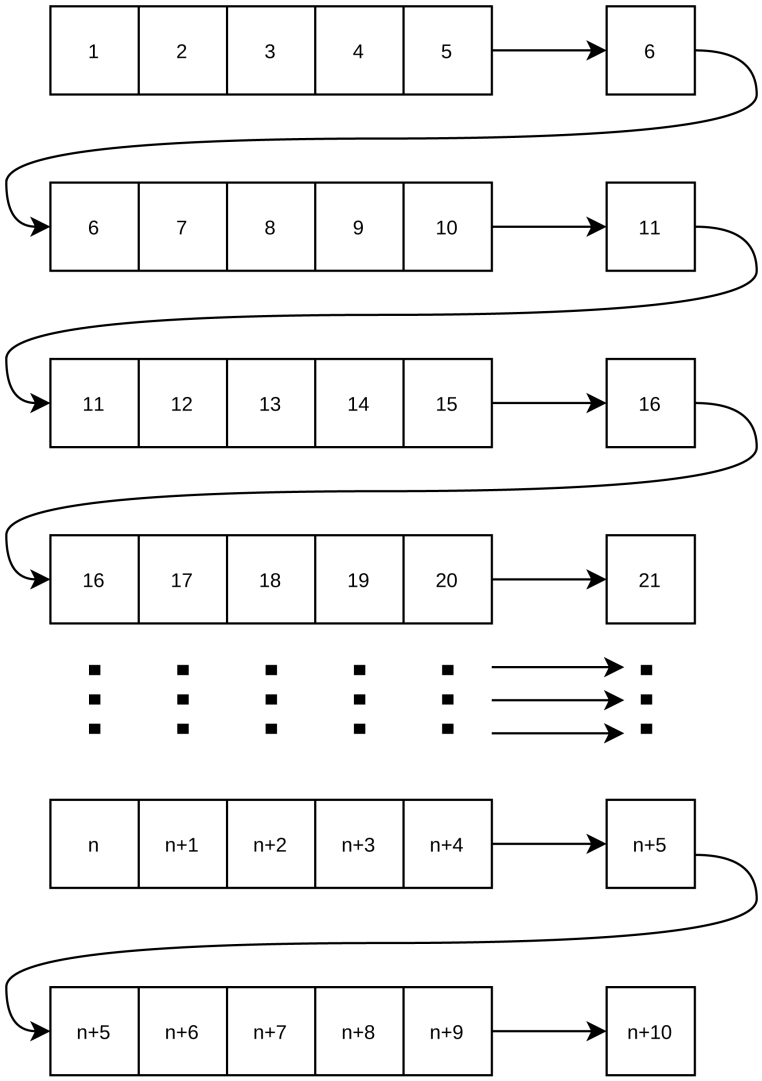

Figure 10. Step prediction flowchart (taking a step of five as an example)
图 10. 步长预测流程图(以步长为五为例)

4.3. 模型求解

预测结果见表 10~14。

Table 10. Table of RMSE for multi-step prediction (AQI)
表 10. 多步预测均方根误差表(AQI)

预测步长	3 步预测	5 步预测	7 步预测	12 步预测
RMSE	19.36	18.15	18.19	18.42

Table 11. Three-step prediction results (AQI)
表 11. 三步步长预测结果(AQI)

日期(年/月/日)	2023/4/30	2023/5/1	2023/5/2	2023/5/3	2023/5/4	2023/5/5
AQI	99.83	88.61	87.46	85.57	83.36	81.88
预警等级颜色	无	无	无	无	无	无
日期(年/月/日)	2023/5/6	2023/5/7	2023/5/8	2023/5/9	2023/5/10	2023/5/11
AQI	80.81	79.97	79.32	78.82	78.44	78.15
预警等级颜色	无	无	无	无	无	无

Table 12. Five-step prediction results (AQI)
表 12. 五步步长预测结果(AQI)

日期(年/月/日)	2023/4/30	2023/5/1	2023/5/2	2023/5/3	2023/5/4	2023/5/5
AQI	100.93	91.52	88.86	88.58	89.03	88.34
预警等级颜色	无	无	无	无	无	无
日期(年/月/日)	2023/5/6	2023/5/7	2023/5/8	2023/5/9	2023/5/10	2023/5/11
AQI	87.48	86.86	86.42	86.08	85.76	85.49
预警等级颜色	无	无	无	无	无	无

Table 13. Seven-step prediction results (AQI)
表 13. 七步步长预测结果(AQI)

日期(年/月/日)	2023/4/30	2023/5/1	2023/5/2	2023/5/3	2023/5/4	2023/5/5
AQI	97.5	89.85	88.92	89.24	89.58	90.78
预警等级颜色	无	无	无	无	无	无
日期(年/月/日)	2023/5/6	2023/5/7	2023/5/8	2023/5/9	2023/5/10	2023/5/11
AQI	92.35	92.51	92.48	92.59	92.78	93
预警等级颜色	无	无	无	无	无	无

Table 14. Twelve-step prediction results (AQI)
表 14. 十二步步长预测结果(AQI)

日期(年/月/日)	2023/4/30	2023/5/1	2023/5/2	2023/5/3	2023/5/4	2023/5/5
AQI	102.12	94.54	91.8	89.11	87.39	86.8
预警等级颜色	无	无	无	无	无	无
日期(年/月/日)	2023/5/6	2023/5/7	2023/5/8	2023/5/9	2023/5/10	2023/5/11
AQI	87.14	88.13	89.4	90.43	91.86	93.36
预警等级颜色	无	无	无	无	无	无

预测等级可以通过 Excel 表格筛选或用 python 代码实现，本文采取了两种方法，取最优后得到结果，见表 15。

Table 15. Summary table of the number of warning level colors
表 15. 预警等级颜色次数汇总表

预警等级颜色	蓝色	黄色	橙色	红色	无	合计
天数(天)	311	0	40	5	2685	3041

接着，本文对预测数据进行了均方根检验，对模型的效果进行评估。经 Python 编译可视化结果如图 11。

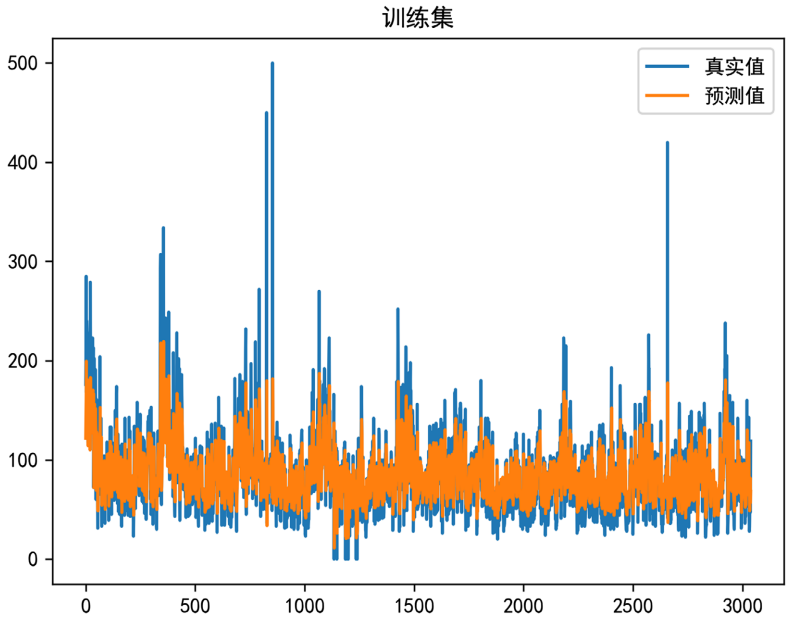

Figure 11. Visualization of test set and its predicted results (AQI)
图 11. 测试集及其预测结果的可视化(AQI)

根据图 11，蓝线代表测试集，橙线代表训练集。经过多次迭代后，训练集和测试集呈高度拟合。可见预测效果较为准确。

参考文献

[1] 史佳霖. 空气分离过程的数据驱动建模及预测方法研究[D]: [硕士学位论文]. 杭州: 杭州电子科技大学, 2021.
[2] 董亚伟. 基于时空注意力网络的 PM2.5 多步超前预测研究[D]: [硕士学位论文]. 兰州: 兰州大学, 2022.
[3] 刘迎军. 基于单步和多步模型的钱塘江南源流域水质检测[D]: [硕士学位论文]. 武汉: 武汉大学, 2021.
[4] 罗奥荣. 基于支持向量回归机的大气 PM2.5 浓度预测模型研究[D]: [硕士学位论文]. 北京: 北京工业大学, 2018.