# PM2.5 的预测与 LSTM 模型构建

刘东升

北京市第一六一中学，北京 100053

## 摘要

预测空气质量 $PM_{2.5}$ 势在必行。影响 $PM_{2.5}$ 指数的因素很多，很多影响因素具有不确定性和非线性特征，因而对 $PM_{2.5}$ 预测的效率等都存在一定的问题。为了监测和估算 $PM_{2.5}$ 浓度，采用的长短时记忆（LSTM）预测模型能有效地预测空气质量 $PM_{2.5}$ 指数。为了评估预测 $PM_{2.5}$ 的 LSTM 模型的性能，采用了均方误差（MSE）、平均绝对误差（MAE）、均方根误差（RMSE）等测量指标对预测的数据和原始数据进行误差对比分析。通过实验证明，LSTM 预测模型能有效地预测空气质量 $PM_{2.5}$ 指数。本文的主要贡献是根据所获得的污染物浓度因子数据以及地表气象因子数据的不用，在 LSTM 预测模型基础上，建立了具体的不同数据来源的 $PM_{2.5}$ 预测模型，从而使 $PM_{2.5}$ 的预测更具有实用性、可行性和灵活性。

关键词：北京 $PM_{2.5}$；预测；LSTM 模型

# Forecasting Model of $PM_{2.5}$ and LSTM Model Construction

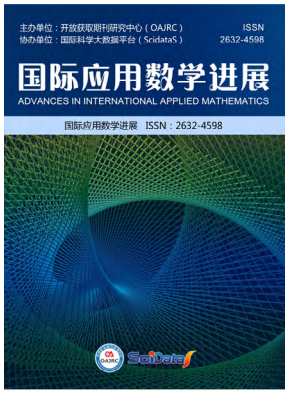
Dongsheng Liu

Beijing No. 161 Middle School， Beijing 100053

## SUMMARY

Forecasting air quality $PM_{2.5}$ is imperative. There are many factors that affect the $PM_{2.5}$ concentration. Many influencing factors have uncertainty and non-linear characteristics, so there are certain problems with the efficiency and accuracy of $PM_{2.5}$ prediction. In order to monitor and estimate $PM_{2.5}$ concentration, the LSTM forecasting model can effectively predict the $PM_{2.5}$ concentration of air quality. In order to evaluate the performance of the LSTM model of $PM_{2.5}$, the data and the original data are compared and analyzed by measuring indexes such as mean square error (MSE), average absolute error (MAE) and mean square root error (RMSE). Experiments show that the LSTM forecasting model can accurately predict the $PM_{2.5}$ concentration of air quality. The main contribution of this paper is based on the data of pollutant concentration factor and the use of surface meteorological factor data. Based on the LSTM forecasting model, $PM_{2.5}$ forecasting model with specific data sources is established. This makes the $PM_{2.5}$ forecast more practical, feasible and flexible.

**Key words:** Beijing $PM_{2.5}$; Projections; LSTM model




http://aam.oajrc.org

OPEN ACCESS

DOI: 10.12208/j.aam.20200001

收稿日期：2019-11-18

出刊日期：2020-01-10

刘东升，北京市第一六一中学，北京 100053;

邮箱：yangchunwo@163.com

电话：13439127918

基金项目：本项目为中国科协和教育部组织的“英才计划”研究项目，指导老师为北京大学数学学院卢朓副教授。

## 1. 前言

### 1.1 选题的背景

北京的 $PM_{2.5}$ 对空气的污染已经成为人们关注的焦点。特别是在冬季供暖时期，北京空气质量更是令人担忧，北京采暖季基本上采用烧天然气（部分农村烧散煤），排放 NOx 是雾霾的主要成因，环境问题十分严峻。因而预测 $PM_{2.5}$ 迫在眉睫。基于此，本文提出了一种基于 LSTM 模型来预测北京的 $PM_{2.5}$ 指数。

目前，机器学习取得非常显著的成就。机器学习能够对强大的、高维度的数据进行分析、特征提取、数据拟合、数据训练、数据验证和数据评估等，这些成果奠定了本文研究的基础。[1] 深度学习中 LSTM（Long Short-Term Memory，长短时记忆）[2] 循环神经网络的理论的丰富和发展，使本课题对于 $PM_{2.5}$ 的预测成为可能。本文建立在机器学习的基础上，基于通过对影响 $PM_{2.5}$ 指数的相关数据的特征的分析和提取，构建预测 $PM_{2.5}$ 的 LSTM 模型。实验数据表明本文提出 LSTM 预测模型，可以有效的预测 $PM_{2.5}$ 浓度值。

以北京地区的天气特征和空气污染指数为基础，循环神经网络 LSTM 为主要网络结构，构造出一套能够有效预测 $PM_{2.5}$ 浓度值的具有时序变化规律的预测模型，同时将地表天气、地表污染物浓度的数据作为模型输入参数，提高预报的精度，降低预报误差。

### 1.2 关于 $PM_{2.5}$ 的预测的研究前期成果

$PM_{2.5}$ 的预测是属于时间序列分析问题的，有学者用自回归（AR）模型、滑动平均（MA）模型、自回归平均滑动（ARMA）模型、自回归积分移动平均（ARIMA）模型进行预测，如陈菊芬等采用多模态支持向量回归的来预测 $PM_{2.5}$，MSVR 模型具有更好的预测效果，与单一 SVR 模型相比，MSVR 模型具有更好的预测效果。[3] 有学者用非线性方法处理的，如任才溶等采用随机森林和气象参数的方法来预测 $PM_{2.5}$ 浓度以及等级预测。[4] 或者采用深度学习进行预测，陈志文等利用 OpenStack 云计算平台，采用 BP 神经网络来预测 $PM_{2.5}$。[5] 传统的学习其特征提取步骤是需要人工和专业专业知识参与，而深度学习通过自动学习，通过训练数据，提取复杂的特征，建立模型，更科学、更合理、更有效。

许多学者在 $PM_{2.5}$ 预测算法和模型上取得了不少突破性进展，但是都存在着模型复杂、难度大等问题。本文提出的基于 LSTM 循环神经网络 $PM_{2.5}$ 预测算法模型，在易用性等方面存在优势，特别是利用时间序列的预测的研究成果和理论来预测，具有前沿性。

## 2. $PM_{2.5}$ 预报预测原理、方法与模型

Hochreiter [6] 等提出的 LSTM 模型是递归神经网络（RNN）的一个改进方案。如图所示：

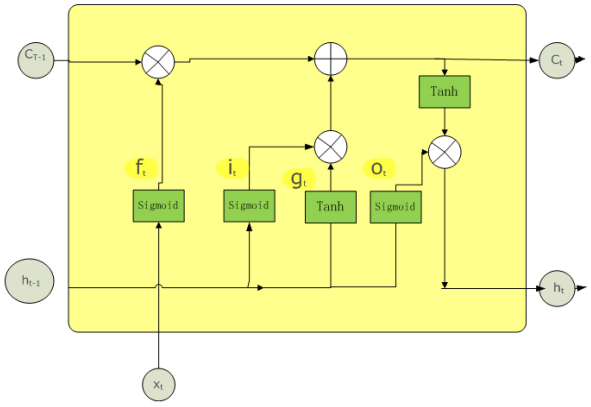



图 1 LSTM 神经网络

### 2.1 神经网络 LSTM

LSTM 是时间循环神经网络，可以用来预测具有时间序列特征的事件。其基本结构为神经 cell 单元，一个神经 cell 单元中有三个门：输入门 $i_t$、遗忘门 $f_t$ 和输出门 $o_t$，分别控制着输入信息、遗忘信息和输出信息；存在两种状态：$h_t$ 为短期状态，$C_t$ 为长期状态；$x_t$ 为当前输入值，$h_{t-1}$ 为先前短期状态，$C_{t-1}$ 为先前长期状态，$\widetilde{C}_t$ 为随时间步更新的状态。一个信息进入 LSTM 中，根据规则决定去留，符合要求的留下，不符合的删除。[7]

2.1.1 具体算法公式为：

$$f_t = \text{Sigmoid}（w_f \cdot [h_{t-1}，x_t] + b_f）\quad（1）$$

$$i_t = \text{Sigmoid}（w_i \cdot [h_{t-1}，x_t] + b_i）\quad（2）$$

$$\widetilde{C}_t = \text{Tanh}（w_c \cdot [h_{t-1}，x_t] + b_c）\quad（3）$$

$$C_t = f_t^* C_{t-1} + i_t^* \widetilde{C}_t \quad（4）$$

$$o_t = \text{Sigmoid}（w_o \cdot [h_{t-1}，x_t] + b_o）\quad（5）$$

$$h_t = o_t^* \text{Tanh}（C_t）\quad（6）$$

其中，w 为每一层输入信息的权重矩阵，b 为偏差项。$w_f$、$w_i$、$w_c$、$w_o$ 分别为对应的权重系数矩阵；$b_f$、$b_i$、$b_C$、$b_o$ 分别为对应的偏置项。首先，利用上一时刻的隐藏层输出和当前层输出，通过（1）、（2）、（3）式计算输入门、输出门、遗忘门的系数。然后，通过（4）式得到当前神经元的候选状态 $C_t$。再通过遗忘门和输入门确定 $C_t$ 和 $C_{t-1}$ 在当前细胞状态中的比例关系，使用（5）式对其进行更新。最后，使用（6）式计算当前时刻的隐藏层输出量。

2.1.2 门

门可以有选择地决定是否让信息通过。由 Sigmoid 神经网络层和点的信息作乘法运算，通过 Sigmoid 神经网络层输出 0 和 1 之间的数值，0 表示不通过任何信息，1 表示全部通过。

“忘记门”由 Sigmoid 层来实现。当前输入值 $x_t$ 和先前短期状态 $h_{t-1}$ 经过时，“忘记门”为单元格先前长期状态 $C_{t-1}$ 中的每个数字输出 0 和 1 之间的数字。1 代表完全保留，而 0 代表彻底删除。

2.1.3 激活函数

在 LSTM 网络中，LSTM 具有删除或添加的能力，这个能力是由被称为门（Gate）的结构所赋予的。而在门中有两个关键性的激活函数在起作用，激活函数 Sigmoid 函数和 tanh 双曲函数，给神经元引入了非线性因素，使得神经网络可以逼近非线性函数。

sigmoid 函数

公式为：$f（x）= \dfrac{1}{1+e^{-x}}$

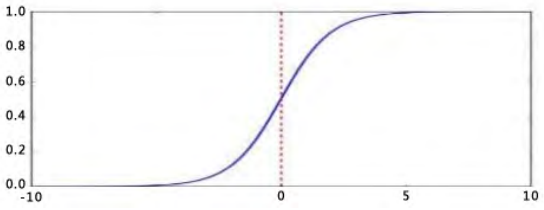



图 2 sigmoid 函数

tanh 双曲函数

公式为：$f（x）= \dfrac{e^x - e^{-x}}{e^x + e^{-x}}$

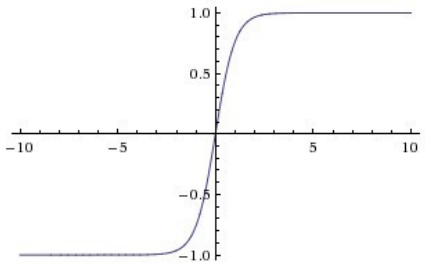



图 3 tanh 双曲函数

sigmoid 函数其输出是在（0，1）开区间内，tanh 双曲函数其输出是在（−1，1）开区间内。

## 2.2 LSTM 的优势

LSTM 单元在输入门 $i_t$、遗忘门 $f_t$ 和输出门 $o_t$ 三个门的控制下，可以学会通过“输入门”提取重要特征信息，“遗忘门”删除不重要的信息和保留重要的信息，“输出门”输出重要的信息，这是 LSTM 的优势所在。LSTM 在长短双重记忆中，可以很好地预测预测时间序列的时间，因而可以预测 $PM_{2.5}$ 的值。

# 3. 数据的采集和分析

本文的主要研究区域为北京地区，研究对象为北京的主要大气污染物 $PM_{2.5}$ 的预测。

使用的数据包括三类：$PM_{2.5}$ 浓度、空气污

染物数据和地表气象数据。本文利用的 $PM_{2.5}$ 浓度、空气污染物数据主要来源于“$PM_{2.5}$ 历史数据”网站[①]，主要选取的数据是北京市 2015 年的空气质量数据，总计下载 300 多条。气象预报数据来自美国气象网站[②]，通过这些数据，分析 $PM_{2.5}$ 与北京市 2015 年地表气温、地表湿度、风向风级之间的关系。在本实验中，利用这些因子在过去的相关 $PM_{2.5}$ 浓度、空气污染物数据和地表气象数据来预测下一个时间段的 $PM_{2.5}$ 浓度。这三种影响因素的数据将被整合到机器学习模型中，进行监督学习和分析，实现准确的预测。中国环境监测总站数据，所有城市监测点的数据都有，而且提供的是原始数据。美国大使馆监测数据，仅支持 $PM_{2.5}$ 的监测，不提供其他污染物的监测。这些数据都可以用来预测 $PM_{2.5}$。

### 3.1 北京 $PM_{2.5}$ 数据的采集时间序列分布特征

北京 $PM_{2.5}$ 时间序列分布折线图如下所示：

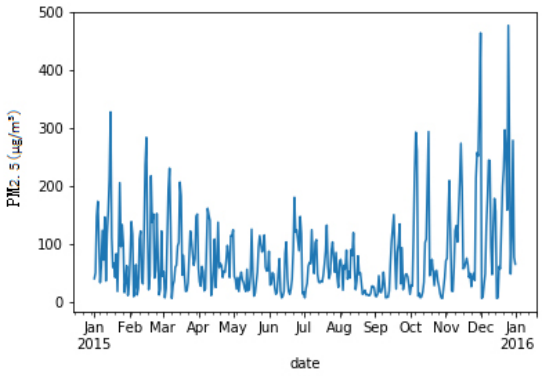



图 4 2015 年北京 $PM_{2.5}$ 月变化图

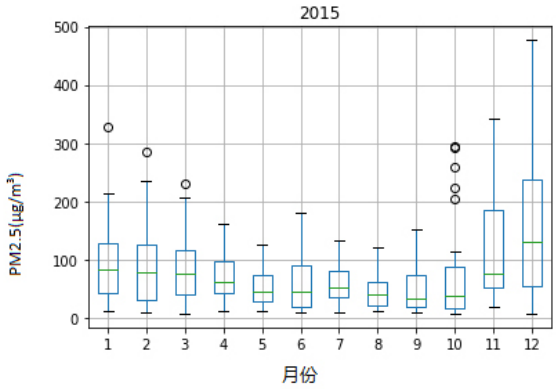



图 5 2015 年 $PM_{2.5}$ 月贡献情况图

通过图 4 可以看出：北京 $PM_{2.5}$ 指数的分布呈现出一定的季节性波动特征，四个季节中冬季 $PM_{2.5}$ 指数最高，空气污染严重，空气质量差。主要是由于冬季北京采暖造成的，北京采暖依靠烧煤，增加了颗粒物的污染，并且北京冬季天气干燥，降雨量少且持续时间较短，因此北京冬季 $PM_{2.5}$ 指数高，空气质量相对较差。通过图 5 对数据的挖掘，可以发现，时间对 $PM_{2.5}$ 的影响相当重要，因此，在需要把 Month、Week、Day 等不同的时间尺度作为特征加入到数据去训练模型。

### 3.2 北京 $PM_{2.5}$ 与空气污染物的相关性特征

为了考量各特征之间的相关程度，需要计算皮尔森相关系数，使用最小二乘法对数据进行直线拟合。相关系数越接近 1 或 −1，说明不同特征之间的相关度越强，相关系数越接近 0，说明不同特征之间的相关度越弱，计算公式如下：

$$\rho_{x,y} = \frac{\mathrm{cov}(X,Y)}{\sigma_x \sigma_y} = \frac{E[(X-\mu_x)(Y-\mu_y)]}{\sigma_x \sigma_y} \quad (7)$$

北京 $PM_{2.5}$ 指数与其他 5 种空气污染物 $PM_{10}$、$SO_2$、CO、$NO_2$、$O_{3\_8h}$ 之间的相关关系散点图北京 PM2.5 指数与其他 5 种空气污染物相关性如图所示：

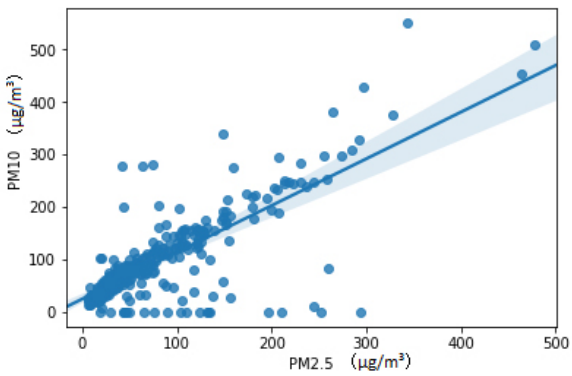



图 6 $PM_{10}$ 与 $PM_{2.5}$ 相关性散点图

[①]数据来源网站为“$PM_{2.5}$ 历史数据”，网址为 https://www.aqistudy.cn/historydata/。

[②]数据来源网站网址为 https://www.wunderground.com/history/daily/cn/beijing/ZBNY/date/2018-7-18

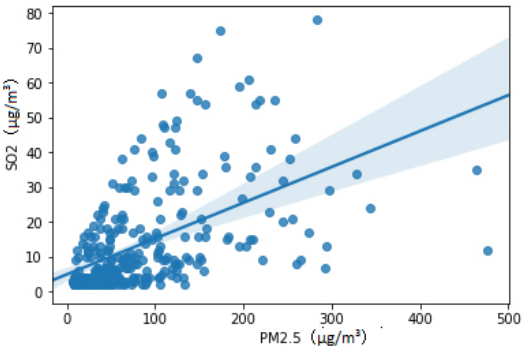



图 7 $SO_2$ 与 $PM_{2.5}$ 相关性散点图

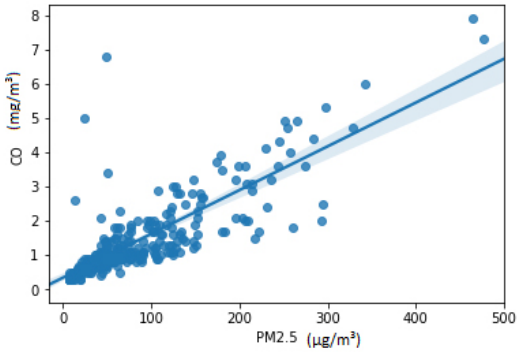



图 8 CO 与 $PM_{2.5}$ 相关性散点图

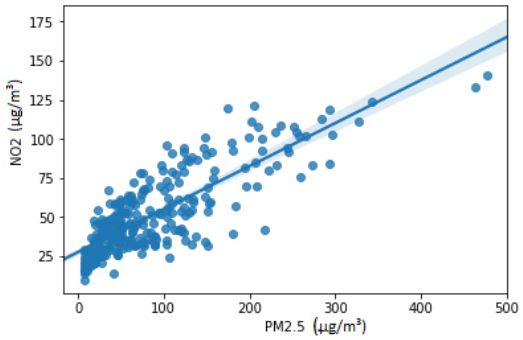



图 9 $NO_2$ 与 $PM_{2.5}$ 相关性散点图

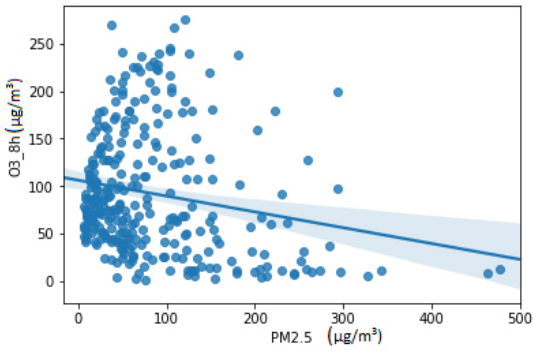



图 10 $O_{3\_8h}$ 与 $PM_{2.5}$ 相关性散点图

相关性散点图展示了 $PM_{2.5}$ 与其它空气污染物 $PM_{10}$、$SO_2$、CO、$NO_2$、$O_{3\_h}$ 的相关性。由图可以发现 $PM_{2.5}$ 与 $PM_{10}$ 、CO 存在着很强的相关性，且与 $PM_{2.5}$ 与 $PM_{10}$ 的相关性最高。

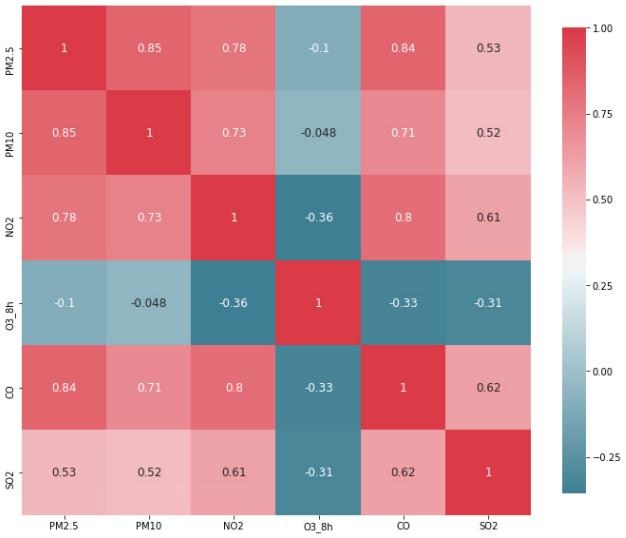



图 11 $PM_{2.5}$ 与各个各种污染物之间的相关性以及相关系数

各种污染物与 $PM_{2.5}$ 指数的相关性如下：$PM_{10}$ 和 $PM_{2.5}$ 指数呈正相关，两者之间相关性最大，为 0.85；CO 与 $PM_{2.5}$ 指数呈正相关，两者之间相关性较大，为 0.84；$NO_2$ 与 $PM_{2.5}$ 指数呈正相关，两者之间相关性居中，为 0.78；$SO_2$ 与 $PM_{2.5}$ 指数呈正相关，两者之间相关性较小，为 0.53；$O_{3\text{-}8h}$ 与 $PM_{2.5}$ 指数呈负相关，两者之间相关性小，为 -0.1。但是，$PM_{2.5}$ 与 $PM_{10}$ 的相关系数为 0.85、$PM_{2.5}$ 与 CO 的相关系数 0.84，超过了 0.8，$PM_{2.5}$ 与 $NO_2$ 的相关系数为 0.78；同时 $NO_2$ 与 $PM_{10}$ 的相关系数为 0.73，CO 与 $PM_{10}$ 的相关系数为 0.71，即各因素间存在多重共线相关性，不满足相互独立的条件。

$PM_{10}$、$SO_2$、CO、$NO_2$ 各种污染物与 $PM_{2.5}$ 指数呈现正相关，说明这些因素对于 $PM_{2.5}$ 的形成起着至关重要的作用，在消解 $PM_{2.5}$ 的时候，要注意这些因素的作用。

### 3.3 北京 $PM_{2.5}$ 与气象因子之间的相关性特征

影响北京 $PM_{2.5}$ 气象数据包括地表温度（Temperature）、露点（Dew Point）、地表湿度（Humidity）、地表风速（Wind）、地表气压（Pressure）、降水（Precipitation）等，下面分析一下 $PM_{2.5}$ 与这些气象数据之间的关系如图 12。

进一步定量地计算 2015 年 $PM_{2.5}$ 与地表气象因子之间的 Pearson 相关系数矩阵，$PM_{2.5}$ 与平均温度、平均露点、平均湿度、平均风速、平均压力、平均降水等之间的相关系数如图 13。

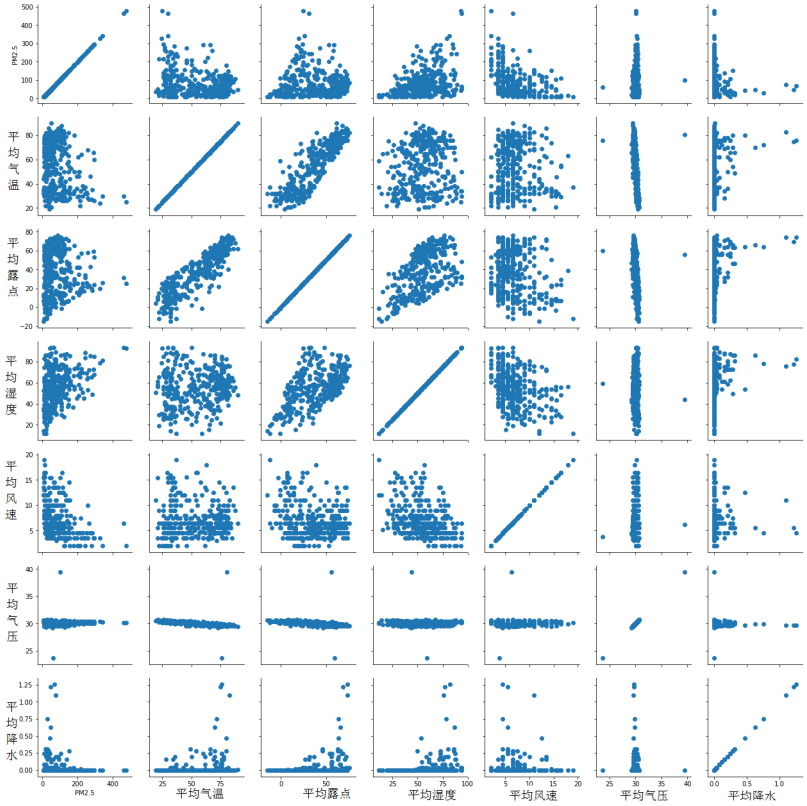



图 12 2015 年地表气象因子与 $PM_{2.5}$ 相关性散点图

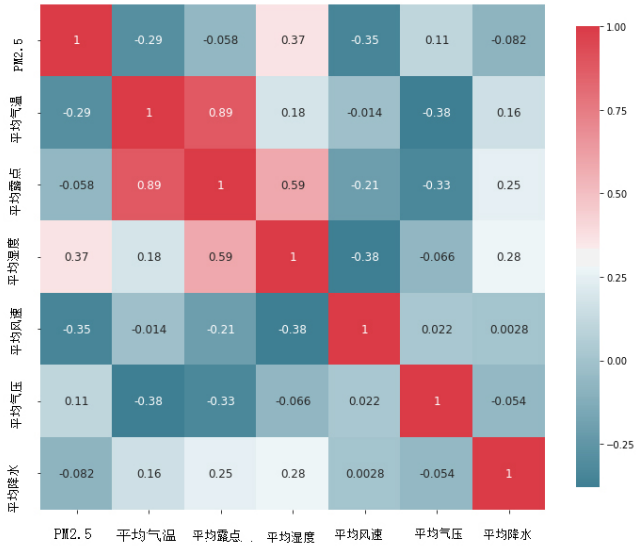



图 13 2015 年 $PM_{2.5}$ 与地表气象因子之间的相关性以及相关系数

$PM_{2.5}$ 与平均温度、平均露点、平均湿度、平均风速、平均压力、平均降水等之间的相关系数如下：平均风速与 $PM_{2.5}$ 指数呈负相关，且相关性较大，为 −0.35；平均温度和 $PM_{2.5}$ 指数呈负相关，且相关性较大，为 −0.29；平均降水与 $PM_{2.5}$ 指数呈负相关，且相关性小，为 −0.082；平均露点和 $PM_{2.5}$ 指数呈负相关，且相关性小，为 −0.058；平均湿度与 $PM_{2.5}$ 指数呈正相关，且相关性较大，为 0.37；平均压力与 $PM_{2.5}$ 指数呈正相关，且相关性小，为 0.11。

$PM_{2.5}$ 与气象因子之间都有一定的相关性。气温与气压之间、气温和气压也存在较强的相关关系，当气温升高时，气压会随之降低。从具体数据看，$PM_{2.5}$ 与气象因子之间的 Pearson 相关系数绝对值都没有超过 0.4。从相关系数上看，$PM_{2.5}$ 与平均压力、平均湿度之间的相关系数符号为正，说明随着降水量、气压的升高，$PM_{2.5}$ 质量浓度会有所上升；$PM_{2.5}$ 与平均风速、平均温度、平均降水、平均露点之间相关系数为负，说明随着这些气象因子数值的升高，PM2.5 浓度会有所下降。

当 $PM_{2.5}$ 与平均温度、平均露点、平均风速、平均降水等呈现负相关时，说明这些因素对于 $PM_{2.5}$ 的形成起着消解的作用，如果能够充分地利用这些因素，将缓解 $PM_{2.5}$ 的形成。

## 3.4 影响北京 $PM_{2.5}$ 因素在时间序列分布特征

污染物浓度因子（包括 $PM_{2.5}$、$PM_{10}$、$SO_2$、CO、$NO_2$、$O_{3\_8h}$）、地表气象因子（包括地表温度、地表湿度、地表风速风向、地表大气压强等）在时间上表现为明显的时间序列和时间周期特征，具体分布特征表现如下：

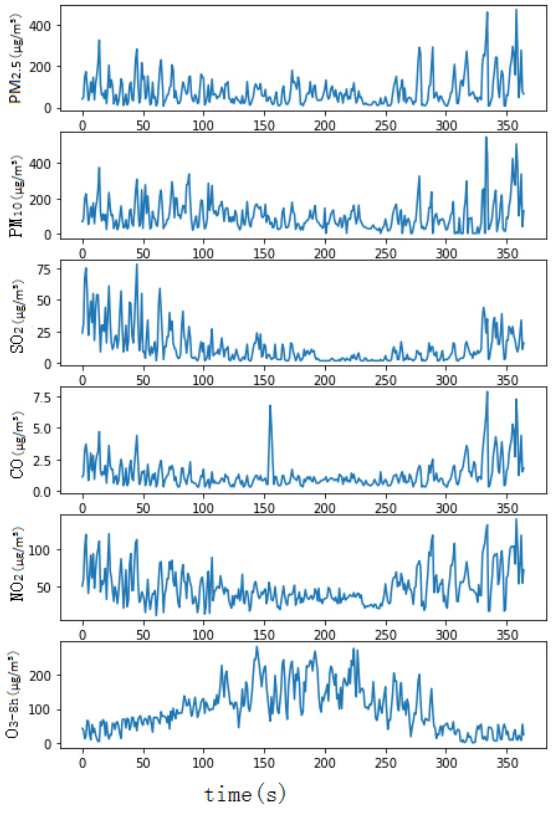



图 14 污染六项指标时间变化

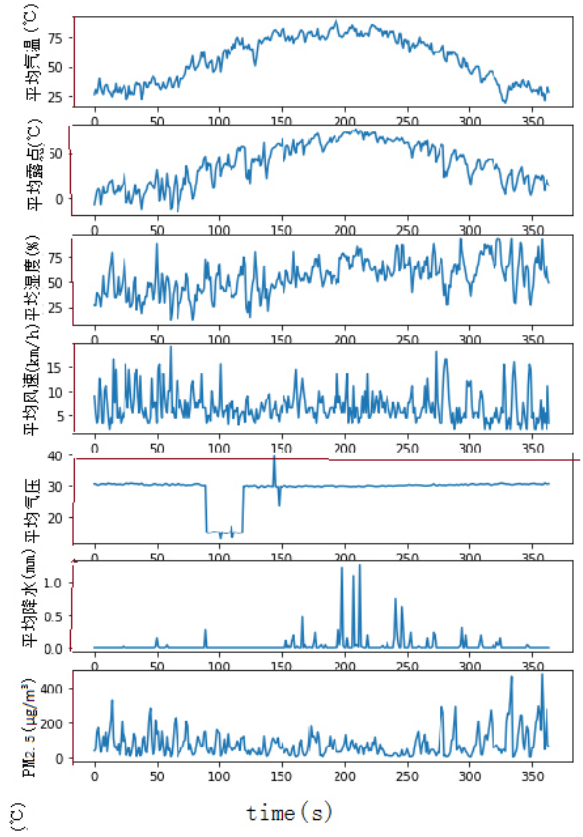



图 15 天气各项指标时间变化

说明空气污染物（$PM_{10}$、CO、$NO_2$、$SO_2$、$O_{3\_8h}$）和气象数据（包括地表温度、地表湿度、地表风速风向、地表大气压强等）在不同的时间段，对 $PM_{2.5}$ 有不同的影响。

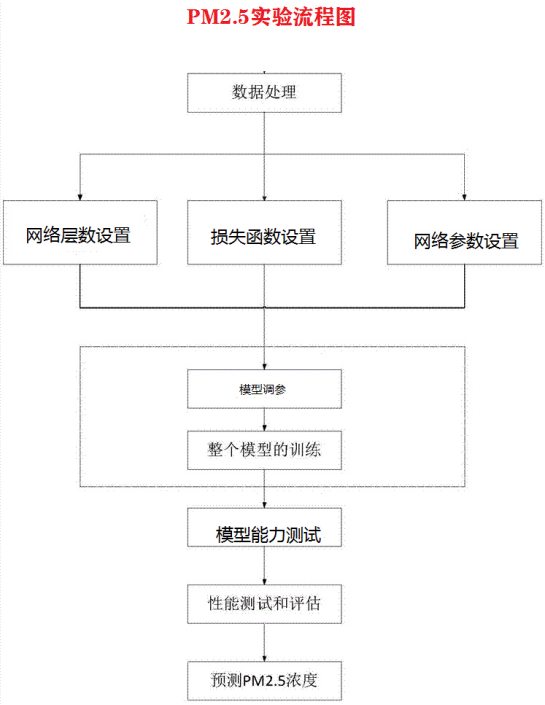
## 4. PM2.5 预测 LSTM 模型的构建

### 4.1 时间序列预测的重要性

在没有预测时间步长的情况下，求得的预测模型其实并没有很大的意义。因为在当前时刻，除了 $PM_{2.5}$ 没有值其余监测值都存在的情况下，$PM_{2.5}$ 的监测值也必然都可以监测给出，这个时候再去根据线性回归模型去预测 $PM_{2.5}$ 是没有必要的。所以，只有在一定的时间间隔的情况下，去预测 $PM_{2.5}$ 才是有意义的。

将时间序列问题转化为监督学习问题。具体来说，就是将数据组分为输入和输出模式，上一时间步长的观察值可作为输入 , 用于预测当前时间步长下的 $PM_{2.5}$ 观察值。

### 4.2 $PM_{2.5}$ 预测流程




### 4.3 时间序列预测框架

时间序列预测框架包括输入层、隐藏层、输出层、训练层和预测层。输入层模块对时间序列进行初步处理，输入数据通过标准化处理转变为 0 到 1 之间，以满足神经网络的输入的要求；隐藏层模块构建 LSTM 层，有相应的隐结点，决定输入元素是否存储；输出层模块获得预测结果，通过输入的 N 组变量对第 N+1 个值进行预测，并且通过反标准化处理，得到预测结果；训练层模块进行优化；预测层模块使用迭代的方法。

### 4.4 数据的处理

#### 4.4.1 数据预处理

1. 整合时间数据为时间格式

用 Strptime 函数（时间日期的格式控制函数）是为了将日期整合，将一个字符串格式化整合为日期格式格式。%Y 为带世纪部分的十制年份，%m 为十进制表示的月份，%d 为十进制表示的第几天，%H 为 24 小时制的小时。用 parse_dates 参数定义时间索引列。

2. 对输入数据和输出数据的处理

输入数据归一化（标准化处理）。数据归一化，就是为了消除不同变量量纲的差异性。使用 preprocessing.MinMaxScaler 函数，可以将不同变量的数据都缩放到最大值和最小值（通常为 0，1）之间。

$$\text{标准化数据（MinMaxScaler）} = \frac{\text{原始数据} - \text{最小值}}{\text{最大值} - \text{最小值}}$$

输出数据还原（反标准化处理）。为了保证神经网络输出的预测数据与输入数据是同一个层面上的数量，即数量级别是一致的，需要将输出的预测结果还原原始数据的数值，也就是反标准化处理。inverse_transform 函数，就是将标准化后的数据转换为原始数据。

### 4.5 具体参数设置

1）数据训练与测试的比例

本实验中总数为 4 年的数据的比例为，基本上是 1 年为训练数据，3 年为验证数据；总数为 1 年数据的比例为，基本上是 100 个数据为训练数据，200 个数据为验证数据；

2）时间步数与特征数

输入层 = 时间步数 * 特征数。本研究中时间步数设置为 3，特征数则根据不同的预测模型，选取的特征数不同。

3）参数设置

本实验设置的参数为：损失函数：Mean Absolute Error（MAE）；优化算法；Adam；隐藏层 epochs 数目：50；Batch：72；学习率：0.01；迭代次数：500。epoch 的大小会对模型的预测精度和运行效率造成较大影响，较大的 epoch 导致运行效率低并可能伴随过拟合现象，较小的 epoch 会使模型训练不充分出现欠拟合现象。

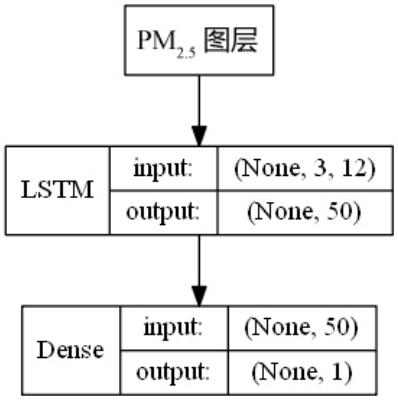



图 16 $PM_{2.5}$ 图层状况

4）预测方式

预测的依据主要是考虑前 3 天（或者小时）的 $PM_{2.5}$ 对于第 4 天（或者小时）的 $PM_{2.5}$ 数据影响程度。在 LSTM 神经网络的训练中，目标 $PM_{2.5}$ 研究数据被处理为每 3 条一组，将前 3 天（或者小时）的 $PM_{2.5}$ 数据读入神经网络模型中进行训练学习，预测出第 4 天（或者小时）的 $PM_{2.5}$ 并与第 4 天 $PM_{2.5}$ 的真实值进行比较。

## 4.6 评价指标

我们通常采用 MSE、RMSE、MAE 来评价预测算法。

均方误差：MSE（Mean Square Error）

$$\mathrm{MSE} = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

其中，$y_i - \hat{y}_i$ 为测试集上真实值 - 预测值。

均方根误差：RMSE（Root Mean Squared Error）

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

可以看出，RMSE=sqrt（MSE）。

平均绝对误差：MAE（Mean Absolute Error）

$$\mathrm{MAE} = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i|$$

## 5. $PM_{2.5}$ 实证预测分析

污染物浓度因子、地表气象因子、时间因子是影响区域 $PM_{2.5}$ 浓度预报精度的重要参量，本文除了引入污染物浓度因子来预测外，还引入了地表气象因子来预测，从而可以有效提高模型的预报精度而降低预报误差。地表气象因子对 $PM_{2.5}$ 浓度的变化是有影响的，在预测的过程中应当在模型中考虑。

基于 LSTM 的 $PM_{2.5}$ 浓度预报模型，目前是在很多的预测模型中表现很好。主要原因是因为 LSTM 系统，能够根据历史数据“预期”下一时间段的 $PM_{2.5}$ 数值，对于时间序列的预测非常有效。预测模型的拟合度相对高，均方误差、均方根误差及平均绝对误差相对低。

## 5.1 $PM_{2.5}$ 具体预测模型

（1）$PM_{2.5}$ 自身影响因子模型。只考虑 $PM_{2.5}$ 自身时间上的影响。

（2）污染物浓度因子模型。污染物浓度因子包括 $PM_{2.5}$、$PM_{10}$、$SO_2$、CO、$NO_2$、$O_{3\_8h}$。

（3）地表气象因子模型。地表气象因子包括地表温度平均温度、平均露点、平均湿度、平均风速、平均气压、平均降水等数据。

（4）污染物浓度因子 + 地表气象因子模型。其中污染物浓度因子包括 $PM_{2.5}$、$PM_{10}$、$SO_2$、CO、$NO_2$、$O_{3\_8h}$，地表气象因子包括平均温度、平均露点、平均湿度、平均风速、平均气压、平均降水。

（5）污染物浓度因子＋地表气象因子（包括最高、最低、平均数值）模型。其中污染物浓度因子包括 $PM_{2.5}$、$PM_{10}$、$SO_2$、CO、$NO_2$、$O_{3\_8h}$，地表气象因子温度（包括最高、最低、平均数值）、露点（包括最高、最低、平均数值）、湿度（包括最高、最低、平均数值）、风速（包括最高、最低、平均数值）、气压（包括最高、最低、平均数值）等。

## 5.2 实验结果

基于不同预报因子组合形成的 $PM_{2.5}$ 预测模型。污染物排放速率对区域 $PM_{2.5}$ 浓度的变化有显著影响、 应当在模型中得到考虑。

### 5.2.1 污染物浓度因子模型

实验以污染物浓度因子 $PM_{2.5}$、$PM_{10}$、$SO_2$、CO、$NO_2$、$O_{3\_8h}$ 来预测 $PM_{2.5}$。选取的数据主要来源于北京市 2015 年全年的 $PM_{2.5}$、$PM_{10}$、$SO_2$、CO、$NO_2$、$O_{3\_8h}$ 数据。

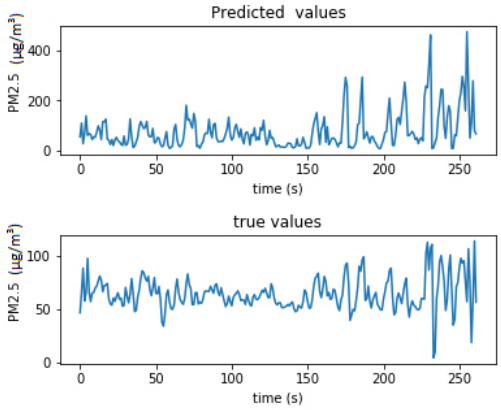



图 17 污染物浓度因子模型预测结果（2015 年）

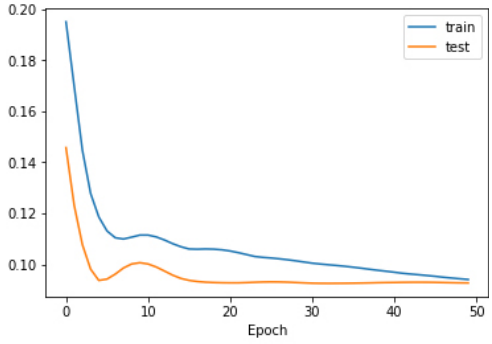



图 18 污染物浓度因子模型训练损失变化曲线（2015 年）

### 5.2.2 地表气象因子模型

实验主要针对 $PM_{2.5}$ 以及地表气象因子中平均温度、平均露点、平均湿度、平均风速、平均气压、平均降水等数据为 $PM_{2.5}$ 进行预测。选取的数据主要来源于北京市 2015 年全年的地表气象数据，来预测 $PM_{2.5}$。

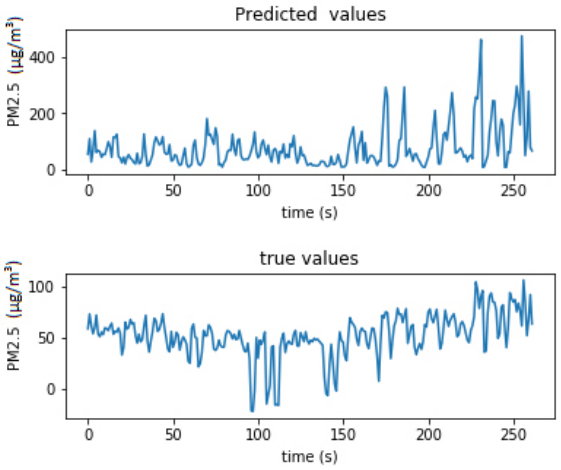



图 19 地表气象因子模型预测结果以及误差图（2015 年）

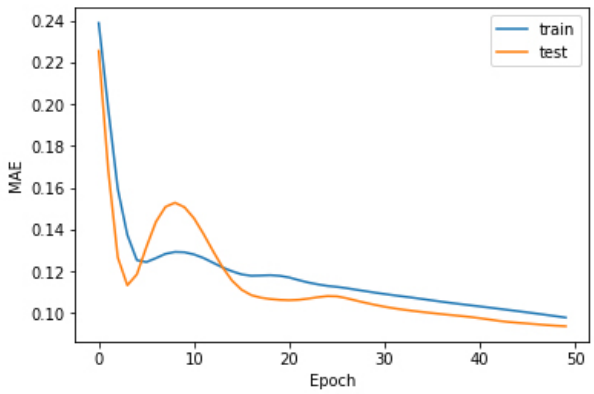



图 20 地表气象因子模型训练损失变化曲线（2015 年）

### 5.2.3 污染物浓度因子＋地表气象因子模型

实验主要针对污染物浓度因子 $PM_{2.5}$、$PM_{10}$、$SO_2$、CO、$NO_2$、$O_{3\_8h}$，以及地表气象数据中平均温度、平均露点、平均湿度、平均风速、平均气压、平均降水等数据为 $PM_{2.5}$ 进行预测，选取的数据主要来源于北京市 2015 年全年的数据，来预测下一天的 $PM_{2.5}$ 值。

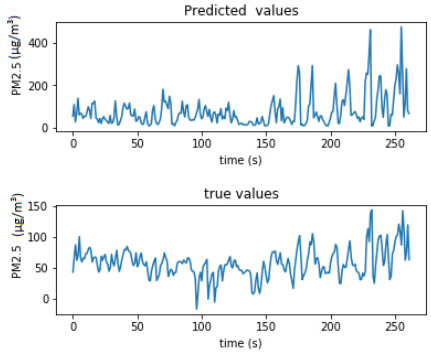



图 21 污染物浓度因子 + 地表气象因子模型预测结果（2015 年）

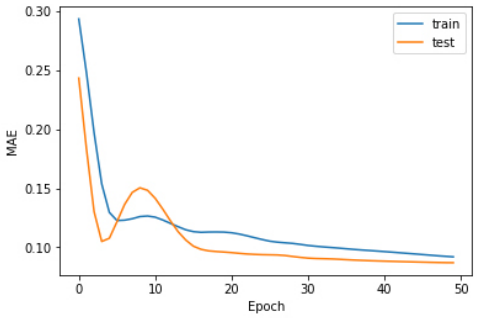



图 22 污染物浓度因子 + 地表气象因子模型训练损失变化曲线（2015 年）

5.2.4 污染物浓度因子 + 地表气象因子（包括最高、最低、平均数值）模型

实验主要针对 $PM_{2.5}$、$PM_{10}$、$SO_2$、CO、$NO_2$、$O_{3\_8h}$、温度最高、平均温度、温度最低、露点最高、平均露点、露点最低、湿度最高、湿度最低、风速最高、风速最低、气压最高、气压最低、平均降水等数据为 $PM_{2.5}$ 进行预测，选取的数据主要来源于北京市 2015 年全年的数据，来预测 $PM_{2.5}$。

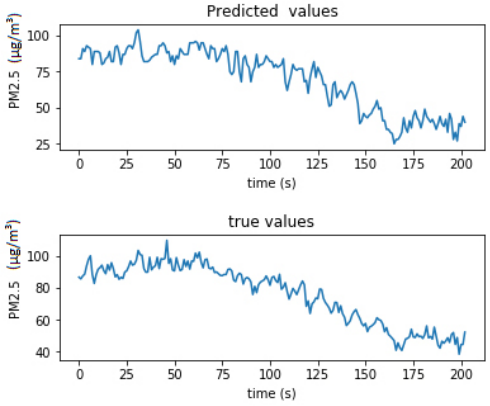



图 23 污染物浓度因子 + 地表气象因子（包括最高、最低、平均数值）模型预测结果（2015 年）

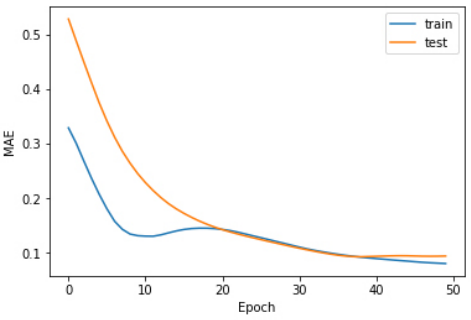



图 24 污染物浓度因子 + 地表气象因子（包括最高、最低、平均数值）模型训练损失变化曲线（2015 年）

5.2.5 按照数据的多少，又进行了两组实验

（a）污染物浓度因子模型（2013 年 12 月到 2018 年 12 月）

实验主要针对 6 项污染物 $PM_{2.5}$、$PM_{10}$、$SO_2$、CO、$NO_2$、$O_{3\_8h}$ 为 $PM_{2.5}$ 进行预测，利用的数据主要来源于“$PM_{2.5}$ 历史数据”网站[①]，分析选取的数据是北京市 2013 年 12 月到 2018 年 12 月的空气质量数据，总计下载 1850 条数据，来预测 $PM_{2.5}$。

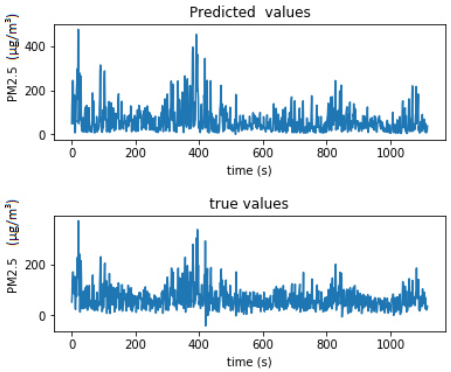



图 25 污染物浓度因子模型预测结果（2013 年 12 月到 2018 年 12 月）

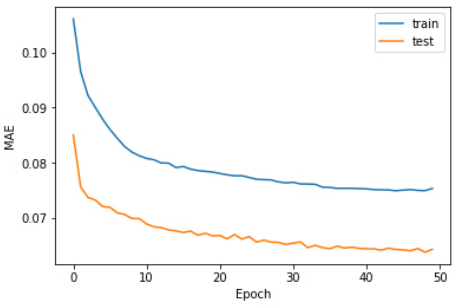



图 26 污染物浓度因子模型训练损失变化曲线（2013 年 12 月到 2018 年 12 月）

[①]数据来源网站为“$PM_{2.5}$ 历史数据”，网址为 https://www.aqistudy.cn/historydata/。

（b）地表气象因子模型（2010 年至 2014 年）

实验主要针对 2010 年至 2014 年美国驻华大使馆每小时报告的天气状况、污染程度来对 $PM_{2.5}$ 预测。该数据集包括 $PM_{2.5}$ 浓度、累积风速和累积降雨小时数。利用这些因子在过去 24 小时的信息来预测下一小时的 $PM_{2.5}$ 浓度。

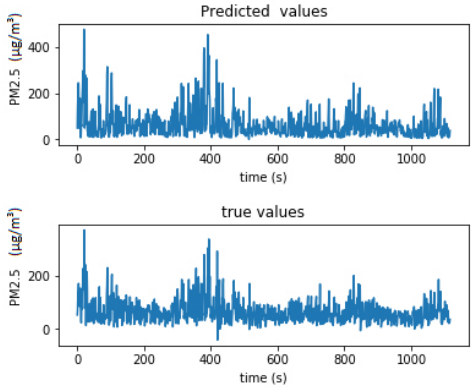



图 27 地表气象因子模型预测结果（2010 年至 2014 年）

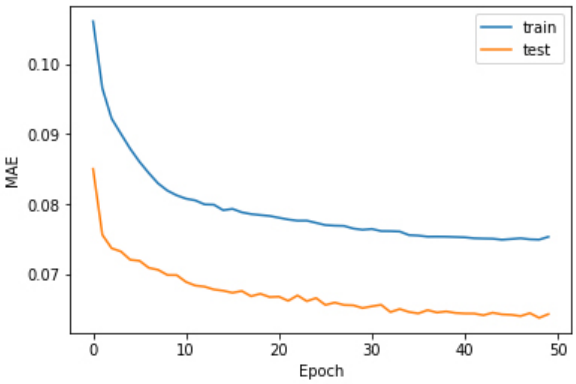



图 28 地表气象因子模型训练损失变化曲线 （2010 年至 2014 年）

5.2.6 LSTM 预测 $PM_{2.5}$ 模型的评判标准结果对比

对 LSTM 预测 $PM_{2.5}$ 模型的评判标准结果对比，结果如下：

表 1 2015 年 LSTM 预测 $PM_{2.5}$ 模型的评判标准（单位 %）

| | 污染物浓度因子模型 | 地表气象因子 | 污染物浓度 + 地表气象因子 | 全部因子 |
|---|---|---|---|---|
| 均方差 MSE： | 4598.95 | 5040.39 | 4240.56 | 90.42 |
| 方根误差 RMSE： | 67.81 | 70.99 | 65.11 | 9.50 |
| 平均绝对误差 MAE： | 42.93 | 44.07 | 40.90 | 7.74 |
| 解释度： | 0.203 | 0.18 | 0.28 | 0.87 |

## 6 实验结果分析

模型预测结果拟合成图形后如上图所示，横坐标为时间，纵坐标为 $PM_{2.5}$ 值，蓝色曲线为预测值，黄色曲线为真实值。观察以上几组拟合图的曲线，将预测结果与实际结果进行比较，可以发现：

（1）从 $PM_{2.5}$ 预测结果来看，预测的数据与真实数据在趋势上基本一致，也就是说预测值和实际值接近，预测的偏差较低。

（2）从时间上来看，预测值曲线与真实值曲线之间存在一定偏差，预测值存在着时间滞后性。对于时间序列的预测来说，时间滞后性是无法避免的，这有待于今后需要进一步调整算法或者模型。

（3）从数据大小来看，数据的大小和多少，直接影响到预测的效果，当输入数据量为 1 年的数据时，预测值曲线与真实值曲线的偏离程度明显，模型的预测结果不佳。而当输入数据由 1 年增加到 4 年时，模型预测结果的准确率便有了显著的提高，预测值曲线与真实值曲线偏离程度降低。

（4）从数据的组成来看，需要考虑多方影响因素。污染物浓度因子（$PM_{2.5}$ 浓度）、地表气象因子（地表温度、地表湿度、地表风速风向、地表大气压强等）、时间因子（预报时刻月份）是影响

区域 $PM_{2.5}$ 浓度预报精度的重要参量，在预测时，应该多方面考虑影响因素。本文污染物浓度因子的同时，成功地引入了地表气象因子作为预报模型的输入，可以有效提高模型的预报精度，降低预报误差。地表气象因子对 $PM_{2.5}$ 浓度的变化是是有深刻的影响，应当在模型中得到考虑。

## 7 关于 $PM_{2.5}$ 预测的未来展望

关于 $PM_{2.5}$ 预测的未来展望：

1. 考虑多时间空间气象数据。本课题收集的数据仅是时间维度的数据，还需要考虑空间维度上的数据，也就是说未来的研究还要考虑地形等地理特征等。如果能做到从时间和空间两个维度来构建模型和算法，将会使 $PM_{2.5}$ 的预更精准和有效。

2. 利用多源头和多维度的数据预测。本课题实验数据是空气污染物（包括 $PM_{2.5}$、$PM_{10}$、CO、$NO_2$、$SO_2$、$O_{3\_8h}$ 等），也包括地表气象数据（包括露点、温度、风向、风速、降雨量、气压等）数据，未来的研究还要利用更多的关于经济、交通等方面的数据，如交通数据、厂矿数据、城市等数据，从而使预测空气质量状况的视角更多源头、多维度、多全面化。

## 结论

影响空气质量 $PM_{2.5}$ 指数的因素很多且难以确定。有线性因素的影响，如 $PM_{10}$ 与 $PM_{2.5}$ 之间相关性很高，存在着线性关系；也存在着非线性之间的关系，如风向和 $PM_{2.5}$ 之间就存在着非线性之间的关系；还有很多难以确定的因素影响着 $PM_{2.5}$ 的指数，这些都为 $PM_{2.5}$ 的预测带来很多难度。用 LSTM（长短时记忆）模型来进行预测。影响 $PM_{2.5}$ 的因素，在时间上往往表现为很强的连续性和周期性等特点，而 LSTM 作为循环神经网络的一种，具有记忆性，可以将影响因素时间上的特征传递和继承到下一个时间段中，这就为监测和估算 $PM_{2.5}$ 浓度提供了很有效的预测方法，从而能有效地预测空气质量 $PM_{2.5}$ 指数。为了验证 LSTM（长短时记忆）模型预测的性能，将使用平均绝对误差（MAE）、均方根误差（RMSE）等测量指标，对预测的 $PM_{2.5}$ 的数据和原始数据进行误差分析，以验证 $PM_{2.5}$ 预测模型的性能。LSTM（长短时记忆）模型能较精准地预测空气质量 $PM_{2.5}$ 指数。根据获得的数据的不用，建立了具体的 $PM_{2.5}$ 预测模型。影响 $PM_{2.5}$ 污染指数的主要有污染物浓度因子（包括 $PM_{2.5}$、$PM_{10}$、$SO_2$、CO、$NO_2$、$O_{3\_8h}$）、地表气象因子（包括地表温度、地表湿度、地表风速风向、地表大气压强等）、时间因子（包括年、月、日、时等），根据能够获得的不同的影响因素数据，可以建立不同的数据预测模型，且这些数据模型具有很强的灵活性和组合性，在充分利用 LSTM（长短时记忆）模型优势的基础上，来预测 $PM_{2.5}$ 的，是具有实用性和可行性的。

## 参考文献

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

[2] Hochreiter S, Schmidhuber J.Long Short-Term Memory [J]. Neural Computation, 1997, 9(8)：1935-1780.

[3] 陈菊芬，李勇 . 基于多模态支持向量回归的 PM_(2.5) 浓度预测 [J]. 环境工程，2019，37(01):122-126+34.

[4] 任才溶，谢刚 . 基于随机森林和气象参数的 PM_(2.5) 浓度等级预测 [J]. 计算机工程与应用，2019，55(02):213-220.

[5] 陈志文，刘立 . 基于 BP 神经网络的 $PM_{2.5}$ 预测 [J]. 电子技术与软件工程，2019(05):143-144.

[6] S.Hochreiter and J.Schmidhuber, “Long Short-Term Memory,” Neural Computation[EB/OL], Vol.9, No.8, 1997, pp.1735-1780. http://dx.doi.org/10.1162/neco.1997.9.8.1735

[7] Understanding LSTM Networks[EB/OL]. http://colah.github.io/posts/2015-08-Understanding-LSTMs