学校代码：10270 分类号：F713.8;TP181 学号：192502693



硕士专业学位论文

基于LSTM模型的长三角地区空气质量指数的预测应用

学 院： 数理学院

专业学位类别： 应用统计硕士

专 业 领 域： 应用统计

研究生姓名： 徐浩

指 导 教 师： 赵红军

完 成 日 期： 年 月

论文独创性声明

本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。论文中除了特别加以标注和致谢的地方外，不包含其他人或机构已经发表或撰写过的研究成果。其他同志对本研究的启发和所做的贡献均已在论文中做了明确的声明并表示了谢意。

作者签名： 日期：

论文使用授权声明

本人完全了解上海师范大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其它手段保存论文。保密的论文在解密后遵守此规定。

作者签名： 导师签名： 日期：

## 摘 要

近年来，大气污染问题逐渐成为影响人们生活不可忽略的因素。因此，对空气质量指数进行预测和监控给人们出行生活以及相关政府部门制定环保政策时提供了强有力的参考意见。

## Abstract

## 第1章：绪论

### 1.1研究背景和意义

#### 1.1.1研究背景

工业革命以来，社会生产力得到大幅提高，人民生活水平也随之大幅改善。与此同时人类活动也对环境造成了深刻的影响，全球变暖，雾霾，酸雨，冰川融化导致的海平面上升以及近来频发的各种恶劣天气等等诸多问题。

我国目前的空气质量研究处于比较晚的阶段，伴随着长期以来形成的以煤炭为主的能源结构，导致我国目前的大气污染尤为严重。加之机动车保有量的不断攀升，更加剧了这一现象。同时，空气质量与人民的生活息息相关，时刻影响着每个人的健康。1996 年，我国首次采用标准化数值形式即空气污染指数 API（Air Pollution Index）来表示空气质量的情况，参与评价的大气污染物有可吸入颗粒物 (PM10)、 二氧化硫(SO2)、二氧化氮(NO2) [2]。2012 年 2 月，环保部发布空气质量指 数 AQI（Air Quality Index)技术规定（试行），并新增三个指标可入肺颗粒 物(PM2.5)、一氧化碳（CO)、臭氧(O3)，直至 2014 年全国城市空气质量评价指标已均为 AQI 了。2010年05月11日国务院办公厅转发环境保护部等部门关于推进大气污染联防联控工作改善区域空气质量指导意见的通知大气污染防治行动计划 ，2013年09月10日 国务院关于印发大气污染防治行动计划的通知.

我国正处在产业升级和产业转型的重要节点，但是环境问题却一直阻碍着我们的发展进程。通过对于空气质量的研究预测，有利于我们提前作出反应，规避在发展过程中可能出现的环境问题，更好的贯彻落实科学发展观，走可持续发展道路，推动我国产业升级。

随着“人类命运共同体”的提出，大气污染环境问题不仅仅是我们国家，它关系着全人类的未来。通过对空气质量的研究预测，不仅有助于我国解决目前现有的困境，还能为目前正在面临相同问题的其他国家提供指导意见，更好地打造“人类命运共同体”。

#### 1.1.2研究意义

长江三角洲地区是我国三大经济圈之一。自1978年国家确定改革开放的发展战略以来，经济的飞速发展和三大产业的高度集中对长江三角洲地区的环境产生了深的影响，由此而来的空气污染问题也日趋严峻。因此，对空气质量指数建立循环神经网络模型进行预测研究以及结合灰色关联分析对其进行主要污染源分析，给相关的环保或者政府部门制定相应的决策时提供了强有力的参考意见。

### 1.2研究内容、研究方法与技术路线

#### 1.2.1研究内容

循环神经网络(RNN)是一类以[序列](https://baike.baidu.com/item/%E5%BA%8F%E5%88%97/1302588)数据为输入，在序列的演进方向进行[递归](https://baike.baidu.com/item/%E9%80%92%E5%BD%92/1740695)且所有节点（循环单元）按链式连接的[递归神经网络](https://baike.baidu.com/item/%E9%80%92%E5%BD%92%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C/16020230)。循环神经网络具有记性性，参数共享，在序列问题处理具有一定优势。因此，在自然语言处理领域有着广泛应用，同时也广泛用于各中类型的时间序列预报。

长短期记忆模型(LSTM)是对RNN的改进，通过引入遗忘门解决了循环神经网路中参数矩阵的信息过载的问题，同时也避免了循环神经网络中存在的梯度消失，很难处理长序列的问题。因此，LSTM在工业界也得到了广泛应用。

LSTM神经网络的多种参数进行调整，才能获得较高的预测精度。目前对于LSTM神经网络预测模型参数选取多数是采取多网格搜索算法、控制变量精细调参，本质上都是暴力搜素寻找最优值的过程，计算资源消耗量大。

由于大范围参数组合寻优带来的计算资源消耗大，预测性能较差，无法寻找到LSTM神经网络最优参数组合的问题，所以本文引入MFO（飞蛾扑火算法）对LSTM参数进行优化。

本文拟通过MF0-LSTM对空气质量指数做短期预测,并与传统机器学习模型SVM中的回归预测模型进行对比，LSTM通过Keras深度学习框架进行模型训练和拟合。再通过灰色关联分析法(GRA)方法分析对空气质量经典6参数中（S02,03,PM2.5,PM10,PM10,NO2,N0）影响较大的因素，可以对大气污染治理有着指导性作用。

Keras是一个用Python编写的高级神经网络API,它能够TensorFlow，CNTK或者Theano作为后端运行。Keras的开发重点是支持快速的实验。能够以最小的时延把我们的想法转换为时延结果。

Keras具有如下优点：

1）模块化：模型被理解为由独立的、完全可配置的模块构成的序列或图。这些模块可以以尽可能少的限制组装在一起。特别是神经网络层、损失函数、优化器、初始化方法、激活函数、正则化方法，它们都是可以结合起来构建新模型的模块。

2）易扩展性： 新的模块是很容易添加的（作为新的类和函数），现有的模块已经提供了充足的示例。

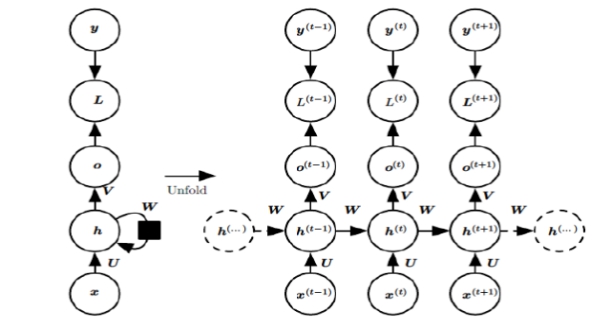
3）基于Python实现： Keras 没有特定格式的单独配置文件。模型定义在 Python 代码中，这些代码紧凑，易于调试，并且易于扩展。

#### 1.2.2研究方法与技术路线

##### 1.循环神经网络（RNN）

在深度学习网路（DNN）和卷积神经网络（CNN）中，训练样本的输入和输出是比较的确定的。但是有一类问题DNN和CNN不好解决，就是训练样本输入是连续的序列,且序列的长短不一，比如基于时间的序列：一段段连续的语音，一段段连续的手写文字。这些序列比较长，且长度不一，比较难直接的拆分成一个个独立的样本来通过DNN/CNN进行训练。

而对于这类问题，RNN则比较的擅长。那么RNN是怎么做到的呢？RNN假设我们的样本是基于序列的。比如是从序列索引1到序列索引的。对于这其中的任意序列索引号,它对应的输入是对应的样本序列中的。而模型在序列索引号位置的隐藏状态，则由和在位置的隐藏状态共同决定。在任意序列索引号，我们也有对应的模型预测输出。通过预测输出和训练序列真实输出,以及损失函数，我们就可以用DNN类似的方法来训练模型，接着用来预测测试序列中的一些位置的输出。



上图左边是RNN模型没有按时间展开的图，如果按时间序列展开，则是按上图中的右边部分。我们重点关注右边部分的图。

这幅图描述了在序列索引号附近RNN的模型。其中：

1）代表在序列索引号时训练样本的输入。同样的，和代表 在序列索引号和时训练样本的输入。

2）代表在序列索引号时模型的隐藏状态。由和共同决定。

3）代表在序列索引号时模型的输出。只由模型当前的隐藏状态决定。

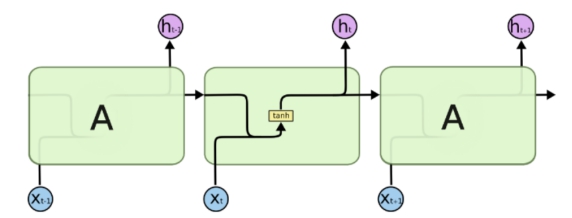
4）代表在序列索引号时模型的损失函数。

5）代表在序列索引号时训练样本序列的真实输出。

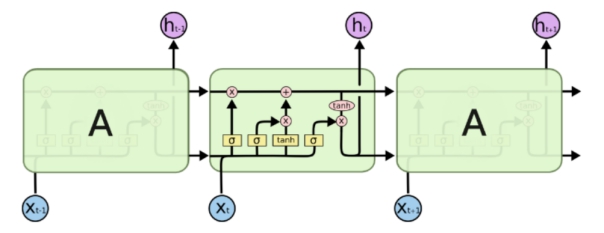
6）这三个矩阵是我们的模型的线性关系参数，它在整个RNN网络中是共享的，这点和DNN很不相同。 也正因为是共享了，它体现了RNN的模型的“循环反馈”的思想。

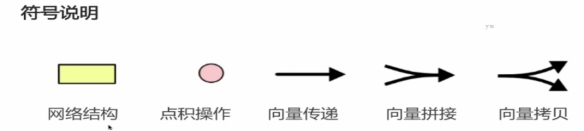
##### 2.长短期记性模型（LSTM）

如果我们略去RNN每层都有的,那么RNN模型可以简化为如下形式：



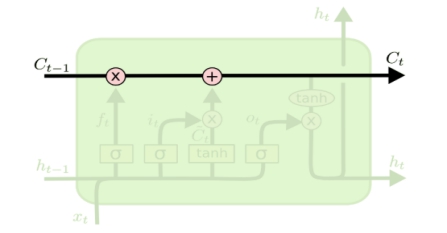
图中可以清晰地看出隐藏状态可以由和，得到的后，一则用于计算当前层的模型损失，二则用于计算下一层的。由于RNN模型存在梯度消失的问题，业内大牛对于序列索引位置的隐藏结构做了改进，通过一些技巧将隐藏结构复杂起来，从而不免梯度消失的问题。这样特殊处理的RNN就死LSTM。LSTM存在很多变种，本文采用最为常见的LSTM，如下图所示：





上面给出了LSTM模型的一般结构，下面我们就此LSTM模型一点点剖析在每个序列索引位置时刻的内部结构。

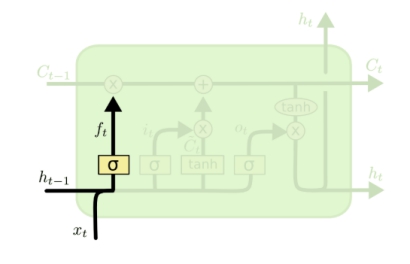
从上图可以不难看出，在每个序列索引位置时刻向前传播的除了和RNN一样的隐藏状态，还多了一个隐藏状态，如下图中的长横线。这个隐藏状态我们一般称为细胞状态（cell state）。记为



LSTM中除了细胞状态，还有一些其他结构，一般称之为门控结构（Gate）,其在每个序列索引位置的门常有遗忘门，输入门和输出门三种。下面我们就来讨论上中涉及的遗忘门，输入门和输出门。

* 1. LSTM的遗忘门

遗忘门（forget gate）即为控制是否遗忘，在LSTM中以一定概率来选择是否遗忘上一层的隐藏细胞状态，它的子结构图如下：



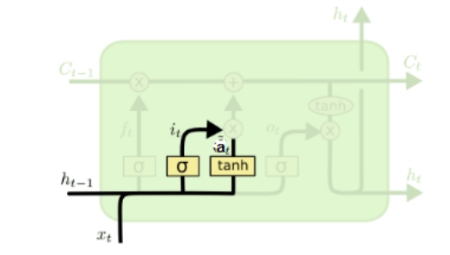
图中输入的有上一序列的隐藏状态和本序列数据，再通过一个激活函数，一般是sigmoid，得到遗忘门输出。由于sigmoid的输出在[0,1]之间，因此这里的输出代表了遗忘上一层隐藏细胞状态的概率。

数学表达式为：

其中：为线性关系的系数和偏置，此处和RNN中的类似，为激活函数

* 1. LSTM的输入门

输入门（input gate）负责处理当前序列位置的输入，输入门子结构图如下



由上图可以看到输入门由两部分组成，第一部分使用和sigmoid激活函数，输出为，第二部分使用了tanh激活函数，输出为 ，两者的结果后续会相乘来更新细胞状态。

数学表达式为：

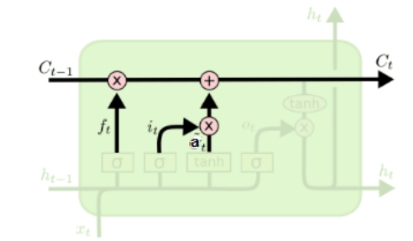




其中：，为线性关系的系数和偏置，此处和RNN中的相似，为sigmoid激活函数

* 1. LSTM的细胞状态更新

在讨论LSTM输出门之前，我们先看看LSTM的细胞状态。前面的遗忘门和输入门的结果都会作用与细胞状态。再看从细胞状态如何得到，如下图所示：



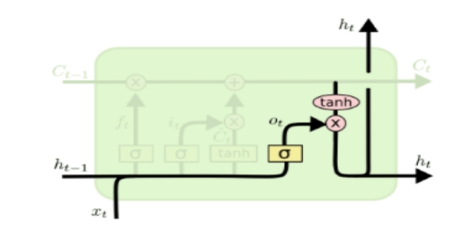
由上图可以看出，细胞状态由两个部分组成，第一部分是和遗忘门输出的乘积。第二部分是输入门和的乘积，数学公式为：



其中：为Hadamard积。

* 1. LSTM的输出门

有了新的隐藏细胞状态，我们就可以来看输出门了，输出门子结构如下：



由上图可以看出。隐藏状态的更新有两个部分组成，第一部分是，它由上一个序列的隐藏状态和本序列数据，以及激活函数sigmoid得到，第二部分有隐藏状态和tanh激活函数组成，即：





1. **LSTM前向传播算法**

LSTM模型有两个隐藏状态，模型参数几乎是RNN的4倍，因为现在多了

这些参数

前向传播过程中每个序列索引位置的过程为：

1. 更新遗忘门输出：



1. 更新输入门两部分输出：





1. 更新细胞状态：



1. 更新输出门输出：





1. 更新当前序列索引预测输出：



1. **LSTM反向传播算法推导关键点**

在有了LSTM前向传播算法，推导反向传播算法就容易了，处理思路与RNN的反向传播算法思路一致，也是通过梯度下降算法来更新迭代我LSTM模型中的所有参数。关键点在于计算损失函数的偏导数来获得所有参数。

在LSTM中有两个隐藏状态和。这里需要定义两个，如下





为了便于推导，我们将损失函数分成两个部分，第一个部分是时刻位置的损失，第二个部分是时刻之后损失，即：



而在最后的序列索引位置的和为：





接下来我们由和反向推出和。

的梯度由本层时刻的输出梯度误差和大于时刻的误差两部分决定，即：



整个LSTM反向传播的难点就在于这部分的计算，由于，在第一项中，包含一个的递推关系，第二项就复杂了，函数里面又可以表示为 

函数的第一项中，包含一个的递推关系，在函数的第二项中，和都包含的递推关系。因此，最终这部分的计算有四部分组成。即：



而的反向梯度误差由前一层的梯度误差和本层的从传回来的梯度误差两部分组成，即：







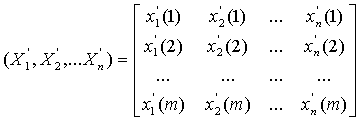
有了和，计算这一堆参数的梯度就容易了，这里给出了的梯度计算过程：



##### 3.灰色关联分析(GRA)

灰色关联分析适用于探究非线性相关性。灰色关联分析是指对一个系统发展变化态势的定量描述和比较的方法，其基本思想是通过确定参考数据列和若干个比较数据列的几何形状相似程度来判断其联系是否紧密，它反映了曲线间的关联程度。

3.1设n个数据序列形成如下矩阵：



其中m为指标的个数

3.2确定参考数据列

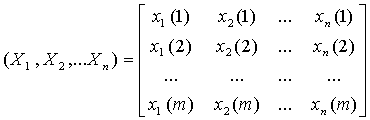
参考数据列应该是一个理想的比较标准，可以以各指标的最优值（或最劣值）构成 参考数据列，也可根据评价目的选择其它参照值．记作



3.3对指标数据进行无量纲化

由于系统中各因素的物理意义不同，导致数据的量纲也不一定相同，不便于比较，或在比较时难以得到正确的结论。因此在进行灰色关联度分析时，一般都要进行无量纲化的数据处理，本文拟采用Z-Score()标准化。

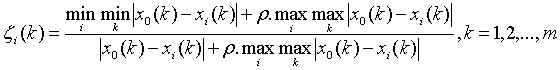
标准化后数据记作：



3.4逐个计算被评价指标序列与参考序列对应元素的绝对差值



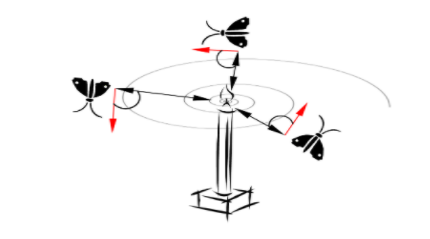
3.5计算相关系数



##### 4.飞蛾扑火算法(MFO)

飞蛾扑火优化(Moth-flame optimization,MFO)，由Seyedali Mirjalili在2015年提出， 为优化领域提供了一种新的启发式搜索范式：螺旋搜索。

飞蛾在夜间飞行有着特殊的导航方式：横向定向。它会和光源保持着一定角度飞行， 从而能够保持直线的飞行路径。但这种方式只在光源离光源较远的地方才生效。当有人 与光源的距离过近，它飞行的路径已经不再是直线，而是一种螺旋的路径。



受这种自然现象的启发，Seyedali Mirjalili将飞蛾绕着光源螺旋飞行过程抽象成了一个参数寻优的过程，飞蛾飞行的空间就是寻优问题的空间。一只飞蛾就是问题的一个解，而火焰（光源）即是问题的一个较优解，每一只飞蛾对应一个光源，避免了算法陷入局部最优；当飞蛾与火焰足够多时，飞蛾的飞行区域就能覆盖解空间的绝大部分区域，从而保证了算法的探索能力；而在寻优的过程中，火焰数随着迭代次数的增加而减少，使飞蛾能够充分搜索更优解的邻域空间，保证了算法的利用能力。

基于以上特点，MFO在探索与利用之间找到了平衡，从而使算法在优化问题上有个较好的结果。总的来说MFO也是一种基于种群的随机启发式搜索算法，它与PSO、GSA等算法最大的区别就在于其粒子搜索路径是螺旋形的，粒子围绕着更优解以一种螺旋的方式移动，而不是直线移动。

MFO过程如下：

1.初始化飞蛾种群

2.对飞蛾种群进行适应度评价

3.重复如下过程指导达到停止标准

3.1自适应更新火焰个数n,当迭代次数为1时，飞蛾个数即为火焰个数

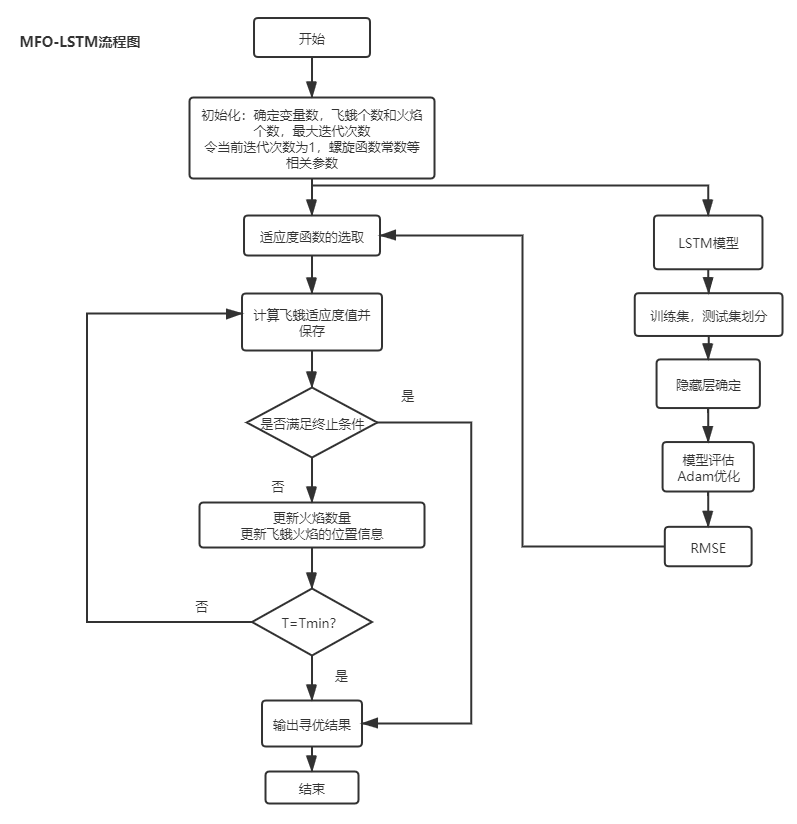
3.2对飞蛾种群适应度进行排序，取出适应度较好的n个飞蛾作为火焰

3.3更新飞蛾的搜索参数

3.4根据每只飞蛾对应的火焰与飞行参数更新飞蛾的位置

4.输出所得最优解（火焰）

##### 5.技术路线



## 第2章：理论基础与文献综述

### 2.1理论基础

#### 2.1.1空气质量指数介绍

空气质量指数(Air Quality Index 缩写为AQI)是描述空气质量状况好坏的无量纲指数，其中参与评价的大气污染物为[SO2](https://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=582152&ss_c=ssc.citiao.link)、NO2、PM10、PM2.5、O3、CO，AQI值越大说明大气污染越严重。

通常空气质量划分为六个等级，如下表：

|  |  |  |
| --- | --- | --- |
| AQI等级 | AQI | 对于人们生活的影响 |
| 优 |  | 无影响 |
| 良 |  | 可能对敏感人群有影响 |
| 轻度污染 |  | 加重敏感人群症状，健康人群出现不适 |
| 中度污染 |  | 对于敏感人群影响很大，并影响健康人群呼吸系统 |
| 重度污染 |  | 健康人群会出现相应的呼吸道症状 |
| 严重污染 |  | 健康人群耐受下降，出现明显不良症状导致某些疾病 |

#### 2.1.2空气中主要污染物介绍

我们在对空气质量进行预测和分析之前，需要了解空气污染物的主要构成。空气中主要有六大污染物即：硫氧化物、氮氧化物、一氧化碳、臭氧、细颗粒物和可吸入颗粒物。

##### （1）硫氧化物

硫氧化物（SOx）主要是指空气中 S02 和 SO3 等污染物气体，硫氧化物主要产生于以煤和石油作为燃料的工业生产。硫氧化物具有很强的刺激性，在空气中遇水会形成酸雨，二氧化硫和粉尘还会发生作用产生有害的物质，具有很强的致癌性，给人们的身体健康造成了很大的危害[29]。

##### （2）氮氧化物

氮氧化物（NOx）主要是指空气中 NO 和 NO2两种污染物。NO 危害比较小，但是其易于臭氧结合生成二氧化氮，二氧化氮主要来自工业生产中化石燃料的燃烧和汽车尾气的排放，二氧化氮是一种刺激性很强的有毒气体，人体吸入少许的二氧化氮有可能导致支气管炎和肺水肿等呼吸道疾病[30]。

##### （3）一氧化碳

一氧化碳（CO）主要是含碳的燃料在不充分燃烧的情况下产生的一种无色无味的有毒气体，一氧化碳非常易于人体血液中的血红蛋白进行结合，从而使人缺氧甚至导致死亡[31]。

##### （4）臭氧

臭氧（O3）是一种具有刺激性、可燃性的淡蓝色气体，主要是由矿物燃料和汽车尾气排放产生的氮氧化物（NOx）与氧气结合生成。臭氧在紫外线的照射下极易与碳氢化合物发生反应从而导致化学烟雾污染，臭氧浓度过高会对人体的免疫系统造成很大的伤害[32]。

##### （5）细颗粒物

细颗粒物（PM2.5）是指空气中当量直径小于或者等于的 2.5um 的颗粒物，这种颗粒物非常细微，容易吸附重金属和有机污染物，极易通过呼吸道进入人体，对人体的呼吸道和肺造成极大的伤害，严重影响人体的身心健康[33]。

##### （6）可吸入颗粒物

可吸入颗粒物（PM10）是指空气中当量直径小于或者等于的 10um 的颗粒物，这种颗粒物容易通过呼吸道进入人体，对人体的呼吸道、肺部和心血管系统造成很大的危害。PM10经常会随风飘到其他地方，导致二次污染[34]。

#### 2.1.3预测流程与评价指标

1.预测流程

空气质量的预测流程包括原始数据的搜集、预处理，构建空气质量的预测模型并对结果进行评价和分析。

1）原始数据的搜集、预处理

2）构建空气质量预测模型

3）模型的评价与分析

2.评价指标

本文构建了SVM,MFO-LSTM两个神经网络空气质量预测模型并进行对比

1）MAE

2）RMSE

### 2.2文献综述

#### 2.2.1国内研究文献综述

黄进,张金池以苏州市2002～2007年各月API值，采用ARIMA模型进行预测分析[1] 。丁卉,刘永红,曹生现提出灰色聚类和模糊评判的方法建立了城市空气质量评价模型[2]。司志娟,孙宝盛,李小芳将灰色GM(1,1)模型与人工神经网络模型组合并改进,对天津市空气质量状况建立改进型灰色神经网络组合模型，具有较好的预测效果[3]。艾洪福,石莹,提出了采用BP人工神经网络的雾霾天气预测系统[4]。Jie Zhou, Ying Cao等人采用LSTM算法解决机器翻译问题，并取得突破性进展[5]。Xiang Li, Dawei Song使用卷积神经网络（CNN）提取与任务相关的特征，此外还使用循序神经网络（RNN），以集成序列的上下文信息, 实验结果表明，就价和情感的情感方面而言，该框架优于经典方法[6]。Fei-yan Zhou, Lin-peng Jin提出了一种结合深度神经网络和规则推理进行PVC检测的新方法, 将LCNN和LSTM结合在一起用于PVC检测[7]。郭利进,井海明等人利用自回归滑动平均模型(ARMA)为卡尔曼滤波建立数学模型,提出了将径向基函数(RBF)神经网络融合于卡尔曼滤波,实现了新的融合算法对AQI进行预测[8]。范竣翔,李琦等人对空气污染物时序数据设计了3种缺失值处理算法并搭建含有全连接层与LSTM层的深度循环神经网络(DRNN)用于时空预报[9]。付亚丽,洪亚针对传统极限学习机的隐含层节点数、及其阈值和权值随机生成造成网络学习能力下降的问题,提出利用PSO算法并行优化极限学习机的隐含层节点数和隐含层的阈值、权值[10]。尹琪,胡红萍等人采用太原市2014年至2016年的空气污染物监测数据,分别将改进的粒子群算法(IPSO)和遗传算法(GA)与支持向量机(SVM)相结合,通过参数寻优构建新模型完成对空气质量指数(AQI)的预测[11]。杨其,陈水忠等人针对传统的基于时间序列的ARMA建模方法和深度学习LSTM网络进行了适用性和实时性对比分析，结果表明实时在线预测情况下LSTM网络优势明显,可以在一定程度上反映噪声变化的趋势并据此对整体控制系统的滤波算法进行优化[12]。李泽龙,杨春节等人针对高炉炼铁是一个动态过程,具有大延迟,工况复杂的特性,采用LSTM-RNN模型进行硅含量预测,充分发挥了其处理时间序列时挖掘前后关联信息的优势[13]。白盛楠,申晓留以气象、大气污染物因素作为PM2.5预测指标,提出基于LSTM循环神经网络的PM2.5预测模型[14]。石晓文,蒋洪迅通过与应用支持向量机、反向传播神经网络、多元线性回归所建立的模型进行了的多重对比实验,表明LSTM预测模型在空气质量指数低于200时具有更高的预测精度[15]。郑洋洋, 白艳萍，侯宇超利用keras的LSTM对太原空气质量指数进行仿真预测[16]。Qitao Zhang, Chenji Wei等人利用长期短期记忆（LSTM）建立了用于预测水饱和度分布的预测模型。从监视和模拟实际油藏得到的数据集被用于模型训练和测试[17]。张澍一,陈松蹊等人提出消除气象干扰的时空调整方法,对北京地区的空气质量评估[18]。邓丽,邬群勇等人提出一种（Stack Sparse Auto-Encoder, SSAE）和（LongShort Term Memory,LSTM）的PM2.5小时浓度预测模型分析了京津冀地区的空气和气象数据[19]。于凯阳,黄志炯等人以珠三角为研究区域,利用HDDM-SRSM不确定性诊断方法量化了清单排放（SO2、NOx、VOCs和NH3）、边界条件浓度和气象（风速和温度）等模型输入参数不确定性对空气质量模拟的影响[20]。程蓉,钱雪忠为解决人工神经网络等方法在预测空气质量指数中存在的容易过拟合、网络结构复杂等问题,提出一种优化的随机森林方法来预测空气质量指数[21]。蒋洪迅,田嘉,孙彩虹创新性地提出了递归随机森林特征消除多层神经网络（RRFMLP）的空气质量预测模型[22]。刘厚凤,徐薇等人利用2020年初新冠疫情对人类生产、生活的一次自然冲击及由此导致的大气污染状况变化，研究了空气质量对污染物减排及气象因素的响应关系[23]。李亚男引入灰色聚类检验模型,给出了三维灰色投影面积关联度的面板聚类步骤.最后通过对珠三角四市的空气质量问题的聚类分析,结果表明各类别的差异明显,层次划分清楚[24]。Yong'an Zhang, Binbin Yan等人提出了一种深度学习预测的新方法，并在此基础上构建了股票市场的深度学习混合预测模型-CEEMD-PCA-LSTM[25]。赵明珠,王丹等人以北京市某地铁车站的温度实测数据为例,采用小波去噪、数据窗口化处理以及时序数据建模方法,分别建立了车站温度的差分自回归移动平均(ARIMA)预测模型、长短时记忆(LSTM)神经网络预测模型、双向长短时记忆(BiLSTM)神经网络预测模型[26]。

#### 2.2.2 国外研究文献综述

Thomas Fischer,ChristopherKrauss利用长期短期记忆网络深入学习对环境质量进行预测f351[27]。Zheng B，NairA，Wh Q等利用数据局部优化实现LSTM和RNN的融合[28]。A.Suleiman，M.R和Tight，A.D.Quinn等人 针对罗马尼亚的空气质量和主要空气污染物，应用机器学习方法管理城市 交通相关颗粒物浓度(PM2.5、PM10)，模拟结果表明预测精度较高[29]。 Deyun Wang等人基于两相分解技术和改进极限学习机的空气质量指数混 合预测模型，预测结果与实测结果相关性较好[30]。

JiDegang和XieXiaoxian等人基于FCM和BP神经网络的空气质量预测与分析，结果与实测结果吻合度较高[31]。Yoichi Inai和Fujiwara等人使用BP神经网络预测对PM2.5, PM10浓度进行了实验，预测结果显示BP神经网络的方法能够准确的模拟上述污染物的浓度[32]。E.Esposito和O.Popoola等人研 究了指示性低成本空气质量传感系统现场随机标定的动态神经网络体系结构[33]。Alimissis A,Philippopoulos K,Tzanis通过人工神经网络对城市空气污染指数进行预测[34]。Nicholas E. Johnson, Bartosz Bonczak, Constantine E等人提出了一种基于Shinyei PPD42气溶胶监测仪的低成本空气质量监测平台的设计，并研究了传感器在密集异质城市环境中的适用性[35]。Mengyuan TANG,Degang JI.Fuzzy充分利用WRF-Chem模型对污染，天气，化学成分的预测作为输入特征，设计了一个综合评估框架来提高预测性能[36]。A.Alimissis和D.Deligiorgi等人研究了基于人工神 经网络模型的城市空气污染空间估算[37]。Sagar V Belavadi, Sreenidhi Rajagopal, Ranjani R,在时间序列数据方面获得成功的成功记录，因此选择了长期短期记忆（LSTM）递归神经网络（RNN）模型来执行空气质量预测任务[38]。Ricardo Navares, José L. Aznarte在马德里不同地区的案例研究中采用了一种预测CO，NO2，O3，PM10，SO2和花粉浓度的方法该方法介绍并比较了循环神经网络中长期短期记忆的不同拓扑，使用非参数假设检验来备份选择最准确和最可靠配置的决策，与用作基准的两种传统方法相比，结果表明准确性有所提高[39]。Aboli N. Londhe, Mithilesh Atulkar根据连续的原始噪声信号对心电图波进行逐样本分类，混合卷积处理铅到铅的变化，通过考虑向前和向后时间戳来自动提取和分类样本的混合模型，对嘈杂的数据集进行实验以提高鲁棒性，可靠性和泛化性[40]。

## 第3章：空气质量指数预测研究

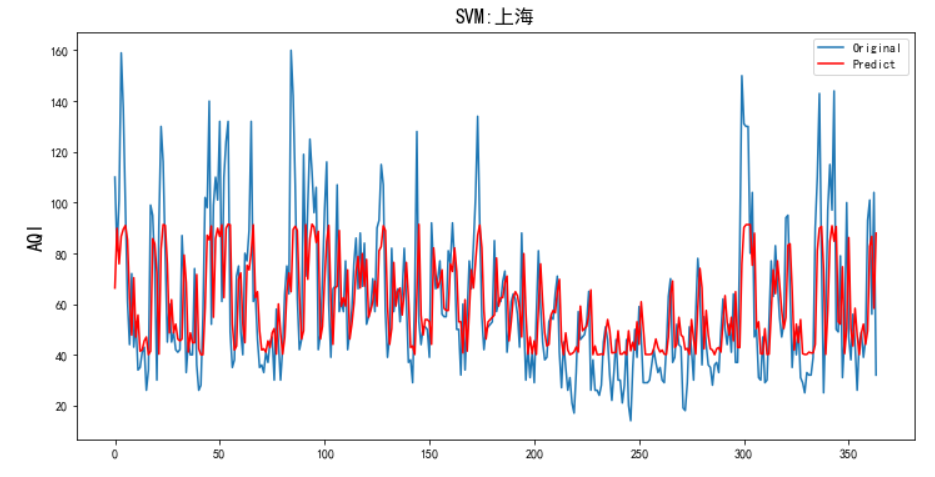
### 3.1数据获取

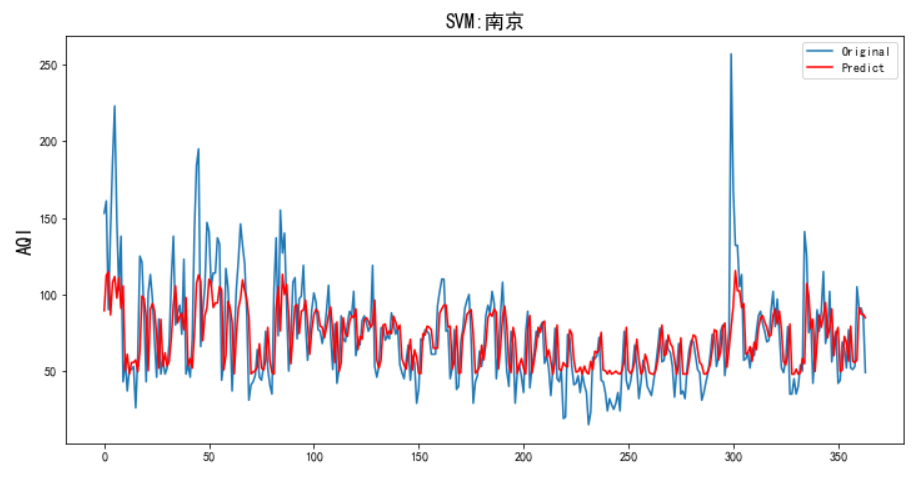
本文数据是从天气后报网站([http://www.tianqihoubao.com](http://www.tianqihoubao.com/)) 通过Python爬虫获取的从2015年01月01日至2019年12月31日的长三角地区26个城市的日数据，包括（AQI,SO2,PM2.5,NO2,CO,PM10,O3指数数据的日数据）。

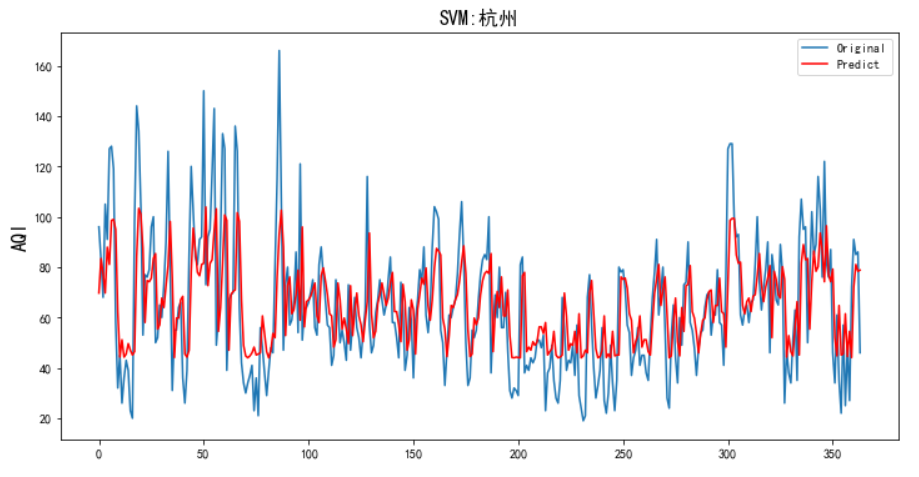
### 3.2 SVM预测

支持向量分类的方法可以扩展来解决回归问题，此方法成为支持向量回归。通过支持向量分类生成的模型仅取决于训练数据的子集，因为构建模型的成本函数并不关心超出裕度的训练点。类似地，由支持向量回归产生的模型仅取决于训练数据的子集，因为成本函数会忽略预测接近其目标的样本。

下面展示SVM预测模型的上海、南京、杭州2019年AQI预测值和原值对比的折线图，蓝线为原始值，红线为预测值，其余图见附录。







### 3.3 MFO-LSTM神经网络预测

MFO-LSTM神经网络的AQI预测方法的步骤如下;

1）搜集长三角地区26个城市的AQI数据，按比例划分训练集和测试集

2）采用MFO算法优化LSTM神经网络预测模型的各个参数

3）输入MFO算法优化好的参数、训练集，进行LSTM神经网络预测模型的迭代优化

4）利用已训练好的LSTM神经网络模型对测试集进行预测，并评估模型误差

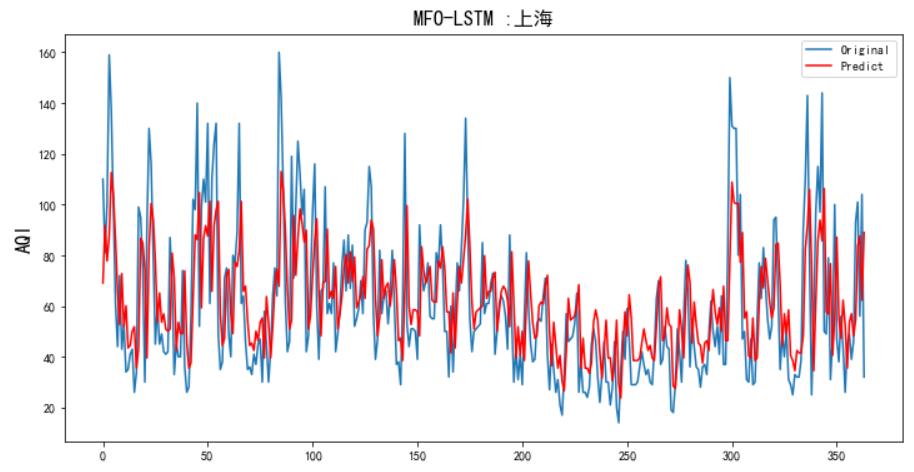
本文中的参数空间选取为时间序列的滞后阶数n以及LSTM神经网络模型的神经元的个数cells，以2015-2018年的日数据为训练集，2019年的日数据为测试集，长三角地区26个城市AQI预测模型训练结果如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 城市 | 神经元个数 | 滞后阶数 | RMSE | MAE |
| 安庆 | 57 | 1 | 24.3507 | 16.9489 |
| 常州 | 26 | 2 | 21.5150 | 15.9616 |
| 池州 | 80 | 1 | 26.4806 | 16.8052 |
| 滁州 | 80 | 1 | 27.9749 | 19.6415 |
| 杭州 | 43 | 1 | 21.2054 | 16.0199 |
| 合肥 | 52 | 1 | 26.8604 | 19.4795 |
| 淮安 | 33 | 2 | 26.3696 | 20.5057 |
| 嘉兴 | 80 | 1 | 21.9191 | 17.0457 |
| 金华 | 10 | 1 | 16.4633 | 12.9613 |
| 泰州 | 58 | 1 | 27.2908 | 21.0714 |
| 马鞍山 | 80 | 1 | 24.8325 | 18.1065 |
| 南京 | 57 | 2 | 25.3621 | 18.4732 |
| 南通 | 52 | 1 | 25.9194 | 19.8214 |
| 宁波 | 69 | 1 | 19.7890 | 15.0043 |
| 衢州 | 73 | 1 | 17.9384 | 13.6242 |
| 上海 | 47 | 1 | 23.3075 | 18.4686 |
| 苏州 | 31 | 1 | 23.8186 | 18.9691 |
| 台州 | 80 | 1 | 15.3966 | 12.5401 |
| 铜陵 | 72 | 1 | 25.4565 | 17.1629 |
| 无锡 | 10 | 2 | 21.5340 | 17.1943 |
| 芜湖 | 74 | 1 | 27.0580 | 18.7792 |
| 宣城 | 80 | 1 | 24.1703 | 17.0799 |
| 盐城 | 42 | 1 | 23.9844 | 18.3831 |
| 扬州 | 10 | 1 | 20.4168 | 14.8537 |
| 镇江 | 10 | 2 | 27.0375 | 21.0613 |
| 舟山 | 65 | 5 | 15.9411 | 11.3676 |

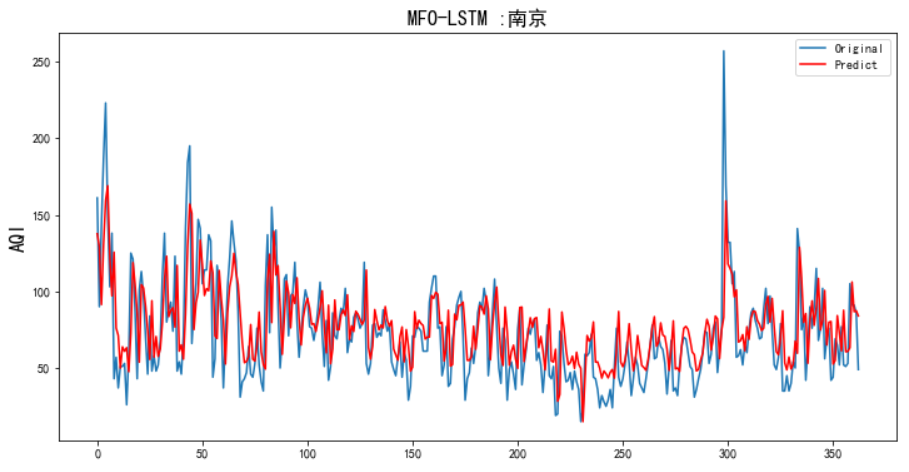
本文使用RMSE,MAE来评价模型，可以看出MFO-LSTM模型的结果要优于SVM模型。

下面展示MFO-LSTM模型的上海、南京、杭州2019年AQI预测值和原值对比的折线图，蓝线为原始值，红线为预测值，其余图见附录。

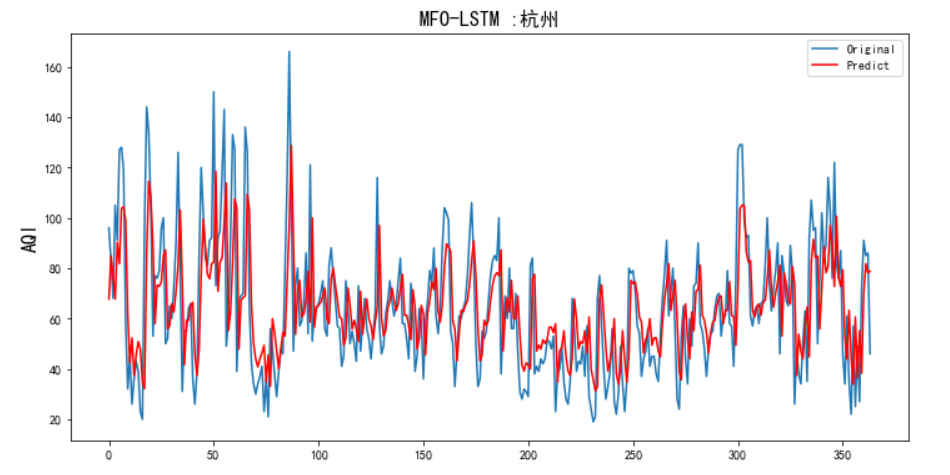
上海：



南京：



杭州：



## 第4章：空气质量指数影响因素分析

### 4.1灰色关联分析的介绍

### 4.2 基于灰色关联分析的空气质量指数影响因素分析

下面展示2019年26个城市的灰色关联分析的结果,并用蓝色标出了城市首要关联的污染物。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **PM2.5** | **PM10** | **SO2** | **N02** | **CO** | **O3** |
| 安庆 | 0.76 | 0.78 | 0.99 | 0.54 | 0.85 | 0.43 |
| 常州 | 0.72 | 0.76 | 0.51 | 0.54 | 0.66 | 0.67 |
| 池州 | 0.80 | 0.98 | 0.75 | 0.98 | 0.75 | 0.98 |
| 滁州 | 0.47 | 0.50 | 0.62 | 0.97 | 0.64 | 0.45 |
| 杭州 | 0.83 | 0.87 | 0.67 | 0.59 | 0.72 | 0.43 |
| 合肥 | 0.66 | 0.80 | 0.59 | 0.57 | 0.82 | 0.71 |
| 淮安 | 0.62 | 0.50 | 0.79 | 0.66 | 0.93 | 0.64 |
| 嘉兴 | 0.72 | 0.74 | 0.70 | 0.91 | 0.92 | 0.97 |
| 金华 | 0.59 | 0.67 | 0.79 | 0.93 | 0.76 | 0.48 |
| 泰州 | 0.63 | 0.67 | 0.82 | 0.61 | 0.91 | 0.56 |
| 马鞍山 | 0.70 | 0.70 | 0.47 | 0.88 | 0.50 | 0.89 |
| 南京 | 0.88 | 0.63 | 0.65 | 0.54 | 0.61 | 0.99 |
| 南通 | 0.89 | 0.75 | 0.60 | 0.79 | 0.51 | 0.82 |
| 宁波 | 0.48 | 0.53 | 0.92 | 0.84 | 0.83 | 0.80 |
| 衢州 | 0.60 | 0.60 | 0.73 | 0.75 | 0.91 | 0.45 |
| 上海 | 0.69 | 0.49 | 0.68 | 0.56 | 0.66 | 0.70 |
| 苏州 | 0.98 | 0.89 | 0.56 | 0.52 | 0.58 | 0.62 |
| 台州 | 0.44 | 0.54 | 0.45 | 0.44 | 0.44 | 0.81 |
| 铜陵 | 0.52 | 0.54 | 0.34 | 0.77 | 0.47 | 0.59 |
| 芜湖 | 0.64 | 0.76 | 0.75 | 0.57 | 0.46 | 0.69 |
| 无锡 | 0.96 | 0.64 | 0.98 | 0.59 | 0.50 | 0.76 |
| 宣城 | 0.78 | 0.58 | 0.83 | 0.67 | 0.86 | 0.96 |
| 盐城 | 0.98 | 0.81 | 0.61 | 0.48 | 0.55 | 0.48 |
| 扬州 | 0.91 | 0.48 | 0.74 | 0.86 | 0.51 | 0.85 |
| 镇江 | 0.61 | 0.61 | 0.83 | 0.97 | 0.93 | 0.80 |
| 舟山 | 0.33 | 0.38 | 0.50 | 0.38 | 0.47 | 0.55 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 主污染源 | 城市 |  |  |  |  |  |  |
| SO2 | 安庆 | 宁波 | 无锡 |  |  |  |  |
| PM10 | 常州 | 杭州 | 芜湖 |  |  |  |  |
| NO2 | 池州 | 滁州 | 金华 | 铜陵 | 镇江 |  |  |
| CO | 合肥 | 淮安 | 衢州 | 泰州 |  |  |  |
| O3 | 嘉兴 | 马鞍山 | 南京 | 上海 | 宣城 | 舟山 | 台州 |
| PM2.5 | 南通 | 苏州 | 盐城 | 扬州 |  |  |  |

安庆，宁波，无锡需要关注SO2主要污染源；常州，杭州，芜湖需要关注PM10主要污染源；池州，滁州，金华，铜陵，镇江需要关注NO2主要污染；合肥，淮安，衢州，泰州需要关注CO污染源；嘉兴，马鞍山，南京，上海，宣城，舟山，台州，需要关注O3主要污染源；南通，苏州，盐城，扬州需要关注PM2.5主要污染源。

## 第5章：总结与展望

### 5.1总结

### 5.2 展望

## 参考文献

[1] 黄进,张金池.苏州市空气质量的时间序列变化过程研究[J].环境科学与技术,2009,32(06):49-52.

[2] 丁卉,刘永红,曹生现.基于模糊-灰色聚类方法的城市空气质量评价研究[J].环境科学与技术,2013,36(S2):374-379.

[3] 司志娟,孙宝盛,李小芳.基于改进型灰色神经网络组合模型的空气质量预测[J].环境工程学报,2013,7(09):3543-3547.

[4] 艾洪福,石莹.基于BP人工神经网络的雾霾天气预测研究[J].计算机仿真,2015,32(01):402-405+415.

[5] Jie Zhou, Ying Cao, Xuguang Wang, et al. Deep Recurrent Models with Fast-Forward Connections for Neural Machine Translation. 2016, 4:371-383.

[6] Xiang Li, Dawei Song, Peng Zhang,et al.Deep fusion of multi-channel neurophysiological signal for emotion recognition and monitoring. 2017, 18(1):1-27.

[7] Fei-yan Zhou, Lin-peng Jin, Jun Dong. Premature ventricular contraction detection combining deep neural networks and rules inference. 2017, 79:42-51.

[8] 郭利进,井海明,南亚翔,修春波.基于卡尔曼滤波融合算法的空气质量指数预测[J].环境污染与防治,2017,39(04):388-391.

[9] 范竣翔,李琦,朱亚杰,侯俊雄,冯逍.基于RNN的空气污染时空预报模型研究[J].测绘科学,2017,42(07):76-83+120.

[10]付亚丽,洪亚.基于IPSO-ELM算法的空气质量预测[J].环境科学与技术,2017,40(S1):324-328.

[11]尹琪,胡红萍,白艳萍,王建中.基于GA-SVM的太原市空气质量指数预测[J].数学的实践与认识,2017,47(12):113-120.

[12]杨其,陈水忠,沈淑梅,朱振华.LSTM网络和ARMA模型对惯性器件随机误差预测适应性分析[J].电光与控制,2018,25(03):68-72+91.

[13]李泽龙,杨春节,刘文辉,周恒,李宇轩.基于LSTM-RNN模型的铁水硅含量预测[J].化工学报,2018,69(03):992-997.

[14]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+104.

[15]石晓文,蒋洪迅.面向高精度与强鲁棒的空气质量预测LSTM模型研究[J].统计与决策,2019,35(16):49-53.

[16]郑洋洋, 白艳萍, 侯宇超. 基于Keras的LSTM模型在空气质量指数预测的应用[J]. 数学的实践与认识, 2019, 049(007):138-143.

[17]Qitao Zhang,Chenji Wei,Yuhe Wang,et al.Potential for Prediction of Water Saturation Distribution in Reservoirs Utilizing Machine Learning Methods.2019,12(19)

[18]张澍一,陈松蹊,郭斌,王恒放,林伟.气象调整下的区域空气质量评估[J].中国科学:数学,2020,50(04):527-558.

[19]邓丽,邬群勇,杨水荣.融合SSAE深度特征学习和LSTM网络的PM\_(2.5)小时浓度预测[J/OL].环境科学学报:1-13[2020-09-30].

[20]于凯阳,黄志炯,史博文,郑传增,白莉,黄江荣,郑君瑜.珠三角空气质量模拟关键不确定性来源识别[J].环境科学学报,2020,40(08):2952-2961.

[21]程蓉,钱雪忠.基于神经随机森林的局部空气质量预测模型[J].计算机工程与设计,2020,41(07):1958-1966.

[22]蒋洪迅,田嘉,孙彩虹.面向PM\_(2.5)预测的递归随机森林与多层神经网络集成模型研究[J/OL].系统工程,2020(05):1-18[2020-09-30].

[23]刘厚凤,徐薇,魏敏,隋潇,许鹏举,李明燕,张美根.2020年初疫情管控对山东省空气质量影响的模拟[J/OL].环境科学:1-17[2020-09-30].

[24]李亚男.基于灰色投影面积关联度的面板数据聚类方法及应用[J].数学的实践与认识,2020,50(14):69-75.

[25]Yong'an Zhang,Binbin Yan,Memon Aasma.A novel deep learning framework: Prediction and analysis of financial time series using CEEMD and LSTM. 2020,159

[26]赵明珠,王丹,方杰,李岩,毛军.基于LSTM神经网络的地铁车站温度预测[J/OL].北京交通大学学报:1-8[2020-10-01].

[27]Thomas Fischer,Christopher Krauss.Deep learning with long short-term memory networks for air quality forecast.European Journal of Operational Research,Volume 270,Issue 2,2018,PP.654-669

[28]Zheng B,Nair A,Wu Q,et a1．EcoRNN：Fused LSTM RNN Implementation with Data Layout Optimization[J].2018．

[29]A.Suleiman,M.R.Tight,A.D. Quinn．Applying machine learning methods in managing urban concentrations oftraffic-related particulate matter(PM10 and PM2.5)[J].Atmospheric Pollution Research,2018．

[30]Deyun Wang,Shuai Wei,Hongyuan Luo,Chenqiang Yue,Olivier Grunder.A novel hybrid model for air quality index forecasting based on two-phase decomposition technique and modified extreme learning machine[J]． Science of the Total Environment，2016．

[31]Ji Degang,Xu Ao, Xie Xiaoxian.Prediction and Analysis of Air Quality Based on FCM and BP Neural Network[J].Meteorological and Environmental Research,2018,9(03):72-74．

[32]Yoichi Inai,Shuji Aoki,Hideyuki Honda,Hiroshi Furutani,Yutaka Matsumi,Mai Ouchi，Satoshi Sugawara,Fumio Hasebe,Mitsuo Uematsu,Masatomo Fujiwara.Balloon-borne tropospheric CO 2 observations over the equatorial eastern and western Pacific[J].Atmospheric Environment,2018,1 84．

[33]Esposito E,De Vito S,Salvato M,et a1.Dynamic neural network architectures for on field stochastic calibration of indicative low cost air quality sensing systems[J].Sensors and Actuators B:Chemical,2016, 231:701-713．

[34]Alimissis A,Philippopoulos K,Tzanis C G,et a1.Spatial estimation of urban air pollution with the use of artificial neural network models[J].Atmospheric Environment, 2018:S1352231018305119-

[35]Nicholas E.Johnson,Bartosz Bonczak,Constantine E.Kontokosta．Using a gradient boosting model to improve the performance of low-cost aerosol monitors in a dense,heterogeneous urban environment[J].Atmospheric Environment,20l8,l84．

[36]Mengyuan TANG,Degang JI.Fuzzy Comprehensive Evaluation and Prediction of Air Quality in Baoding City[J].Agricultural Biotechnology,20l8,7(03): 110-115．

[37]A.Alimissis,K.Philippopoulos,C.G.Tzanis，D.Deligiorgi．Spatial estimation of urban air pollutionwith the use of artificial neural network models[J]．Atmospheric Environment,2018

[38]Sagar V Belavadi,Sreenidhi Rajagopal,Ranjani R,et al.Air Quality Forecasting using LSTM RNN and Wireless Sensor Networks.2020,170:241-248.

[39]Ricardo Navares, José L. Aznarte. Predicting air quality with deep learning LSTM: Towards comprehensive models. 2020, 55

[40]Aboli N. Londhe,Mithilesh Atulkar.Semantic segmentation of ECG waves using hybrid channel-mix convolutional and bidirectional LSTM. 2021,63

## 致谢

## 附录