# 智能化空气环境监测与分析

石桂萌　吕笑语　李美莹
杭州师范大学　浙江杭州　310000

**摘　要:** 近年来，我国在经济高速发展的同时，也面临着空气污染问题。随着空气污染的日益严重，空气质量开始恶化，酸雨、雾霾等环境问题日益增多，对我国的经济和生态系统、社会的可持续发展、人们的身体健康造成了严重影响。我们需要深入分析空气质量现状，研究其影响因素，并制定相关对策和采取相应措施来缓解甚至解决空气污染与经济发展之间的矛盾。本研究利用机器学习神经网络模型、决策树的集成学习方法和空气质量指数（Air Quality Index，简称 AQI），进行空气质量的预测与分析，实现对未来空气质量的准确预测，并探索空气质量与环境因素之间的关联关系。

**关键词：**污染物；空气质量；神经网络；随机森林；AQI

## 1. 前言

近年来，我国在经济高速发展的同时，也面临着空气污染问题。随着空气污染的日益严重，空气质量开始恶化，酸雨、雾霾等环境问题日益增多，对我国的经济和生态系统、社会的可持续发展、人们的身体健康造成了严重影响。我们需要深入分析空气质量现状，研究其影响因素，并制定相关对策和采取相应措施来缓解甚至解决空气污染与经济发展之间的矛盾。

## 2. 数据

### 2.1 数据来源与收集方法

我们访问中国国家统计局的官方网站《中国统计年鉴》和中国碳核算数据库，下载相关行业和年份的统计数据。

### 2.2 数据集描述与特征

本数据集包含江浙沪地区 2010–2021 年的不同时间的气象数据和环境污染物数据。监测站点遍布江苏、浙江和上海，涵盖了城市、郊区和农村等多个地区。

数据集特征主要包括不同时间的温度、湿度、风速、风向和大气中的污染物浓度。

## 3. 利用神经网络模型进行空气质量预报

### 3.1 引言

空气质量是人类健康和环境质量的重要指标，因此准确预测和分析空气质量对于保护公众健康和实施环境管理措施至关重要。本研究利用机器学习神经网络模型，基于历史空气质量数据和相关环境因素，进行空气质量的预测与分析。通过构建和训练神经网络模型，实现对未来空气质量的准确预测，并探索空气质量与环境因素之间的关联关系。

### 3.2 模型设计

#### 3.2.1 模型原理

该模型采用循环神经网络 LSTM 进行空气质量时间序列预测。LSTM 是一种能够处理时间相关性的数据的循环神经网络。相比简单的 RNN，它引入了记忆单元，可以获取长期的时间依赖，解决简单 RNN 中的梯度消失和梯度爆炸问题。

LSTM 的计算公式如下：

记忆单元：

• 遗忘门 $\mathrm{f}_t$：控制前一个记忆单元状态 $C_{t-1}$ 中多少信息会被遗忘。

• 输入门 $\mathrm{i}_t$：决定有多少新的信息 $\overline{C_t}$ 将被输入到细胞状态 $C_{\mathrm{t}}$ 中。

• 待输入的新信息 $\overline{C_t}$：通过权重参数 $W_{\mathrm{c}}$ 和偏置参数 $\mathrm{b}_c$ 计算的出。

• 细胞状态 $C_{\mathrm{t}}$：通过组合遗忘门和输入门得到更新后的细胞状态。

• 输出门 $\mathrm{o}_t$：决定有多少细胞状态会被输出。

• 隐状态 $\mathrm{h}_t$：通过输出门 $\mathrm{o}_t$ 和细胞状态 $C_t$ 相乘后经过双曲正切函数处理得到。

#### 3.2.2 模型建立

（1）数据准备：首先，我们需要准备具有历史空气质量信息的数据集，此数据集中包括多个空气质量指标（如

$SO_2$、$NO_2$、$PM_{2.5}$、$PM_{10}$、$CO$、$O_3$等）以及与其相关的其他特征（日期）。根据预测的目标，我们选择需要预测的污染物列，并进行数据清洗和预处理，删除空值行和数据缩放。

（2）数据划分：将数据集划分为训练集和测试集。通常，我们会将大部分数据用于模型的训练，选择其中一部分作为测试集，用于评估模型的性能。在本章节中，我们使用 80% 的数据作为训练集，20% 的数据作为测试集。

（3）生成训练样本：定义一个函数生成训练样本，每个样本包含过去的观测和将来的预测目标。同时，我们将数据集转换为时间序列的格式，其中是输入序列，是预测的下一个时间步的输出。

（4）模型构建：我们使用模型构建神经网络。该模型包括一个层作为主要模型层，用于处理序列数据，以及一个全连接层用于输出预测结果。（长短期记忆网络）是一种循环神经网络的变体，适用于处理具有时间关系的序列数据。

（5）模型训练：使用均方误差作为损失函数，优化算法进行模型优化。

（6）反向缩放预测结果：由于之前对数据进行了缩放处理，需要将预测结果反向缩放，还原到原始数据范围。

（7）模型评估：使用测试集对训练好的模型进行评估。计算均方误差、均方根误差和平均绝对误差等指标，以评估模型的性能。

3.3 结果与分析

经模型求解我们得到均方误差、均方根误差和平均绝对误差的值，如下表所示：

**表一 模型评估表**

| 均方误差 (MSE) | 均方根误差 (RMSE) | 平均绝对误差 (MAE) |
|---|---|---|
| 0.06538105010986328 | 0.25569718439956135 | 9.768716524442036 |

根据这些指标，可以看出模型在预测目标污染物时的误差比较小，均方根误差和平均绝对误差都在可接受范围内。这意味着我们所构建的神经网络模型相对准确地预测了目标污染物的值。

**4. 利用随机森林算法进行空气质量预报**

4.1 引言

在过去的几十年里，随机森林方法逐渐成为气象学、环境科学和大气污染物研究中一种常用的预测工具，是一种基于决策树的集成学习方法，它能够有效处理多变量和非线性的关系，并具有较强的泛化能力。在这里通过利用历史气象数据和污染物浓度数据，结合其他相关因素，我们构建出针对浙江省空气质量的预测模型，为空气污染物控制和监测提供有价值的参考。

4.2 模型设计

随机森林是一个包含多个决策树的分类器，并且其输出的类别是由个别树输出的类别的众数而定。利用相同的训练数搭建多个独立的分类模型，然后通过投票的方法，以少数服从多数原则作出最终的分类决策。

1. 数据集的随机采样：从原始训练集中随机选择一定数量的样本（有放回地采样），形成新的采样集。这样做可以保证每个样本有机会被选中，并且每个决策树的训练集会有一部分相同的样本。

2. 特征集的随机选择：从所有特征中随机选择一部分特征，形成新的特征子集。这样做可以减少特征的相关性，增加模型的多样性。

3. 构建决策树：使用上述的采样集和特征子集构建决策树。在每个节点上，根据某个准则选择最优的特征，将节点分裂成多个子节点。不断递归地构建子节点，直到满足停止条件为止。具体的分裂过程和剪枝可以使用以下公式表示：

分裂准则：

$$J(\theta,t_f)=\frac{m_{left}}{m}*G_{left}+\frac{m_{right}}{m}*G_{right}$$ 公式一

剪枝准则：

$$C_\alpha(T)=C(T)+\alpha|T|$$ 公式二

其中，$J(\theta,\mathrm{t}_f)$ 表示节点分裂的评估准则，$\mathrm{m}_{left}$ 表示左子节点的样本数，$G_{\mathrm{left}}$ 表示左子节点的不纯度，$\mathrm{m}_{\mathrm{right}}$ 表示右子节点的样本数，$G_{right}$ 表示右子节点的不纯度，$C_\alpha(T)$ 表示剪枝准则，$C(T)$ 表示叶子节点的不纯度，$\alpha$ 是一个调节参数。

4. 预测：对于分类问题，使用投票的方式来确定最终的预测结果。每棵决策树都对新样本进行分类，最终选择票数最多的类别作为预测结果。对于回归问题，使用平均值或中位数等方式汇总每棵决策树的预测结果，得到最终的预测值。可以使用以下公式表示：

$$\hat{y} = \frac{1}{B}\sum_{i=1}^{B} f_i(x)$$

其中，$\hat{y}$ 表示模型的预测值，$B$ 表示模型中决策树的个数，$f_i(x)$ 表示第 i 棵决策树的预测结果。

4.3 结果与分析

经过模型求解得到均方误差为 0.033，表示随机森林回归模型在测试集上的预测结果与实际值的平均差异较小。这个值越小，通常意味着模型的预测性能越好。

## 5. 利用模型进行空气质量评价

5.1 引言

为了更好地评价空气质量，目前广泛使用的指标之一是空气质量指数（Air Quality Index，简称 AQI）。AQI 是根据不同污染物的浓度水平，针对特定的健康影响区间进行分类，从而对空气质量进行等级评价的指标。AQI 的引入使得公众能够更直观地了解空气质量，并根据不同级别的 AQI 采取适当的防护措施。

本研究拟借助先进的机器学习技术，结合气象和污染物数据建立 AQI 模型，并通过模型的应用实现空气质量的实时预测和监测。通过该模型，可以根据实时收集的气象数据和污染物浓度数据，准确地预测出当地的 AQI 值。

5.2 模型设计

具体的计算公式可能因地区和评估机构的不同而有所差异。下面是一种基本的计算公式示例，用于计算空气质量指数（AQI）：

首先，计算各个污染物的污染指数（PI）：

$$P = \frac{C}{I} * 100\%$$

其中，C 表示测量的污染物浓度，I 表示该污染物的标准或阈值。

然后根据每个污染物的污染指数（PI）计算空气质量指数：

$$AQI = \max(P_1, P_2, P_3, ..., P_n)$$

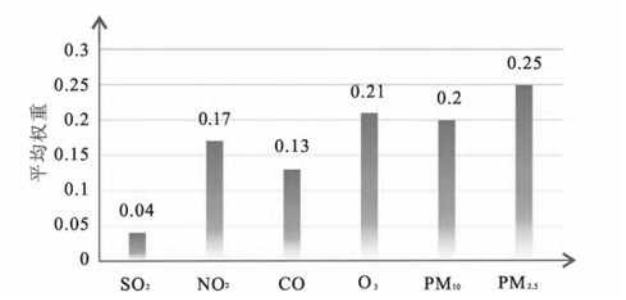
其中，PI1、PI2、PI3 等表示各个污染物的污染物指数，n 表示用于计算 AQI 的污染物个数，max 表示取最大值。

表二 中国环境保护部 AQI 评价标准

| AQI 值 | 空气质量 | 影响程度 |
|---|---|---|
| 0–50 | 空气质量优 | 对健康影响较小 |
| 51–100 | 空气质量良好 | 一般人群可正常活动 |
| 101–150 | 轻度污染 | 敏感人群可能出现轻度症状 |
| 151–200 | 中度污染 | 敏感人群可能出现明显症状，一般人群可能有轻度影响 |
| 201–300 | 重度污染 | 敏感人群可能出现较重症状，一般人群可能有明显症状 |
| 301–500 | 严重污染 | 敏感人群可能出现严重症状，一般人群可能有更严重症状 |

5.3 结果与分析

根据主要污染物占比的权重，我们可以得出结论：二氧化硫（$SO_2$）浓度对空气质量指数（AQI）的影响相对较小，而细颗粒物（PM2.5）、可吸入颗粒物（PM10）和臭氧（$O_3$）对空气污染的影响更为重要。这三项污染物的影响相较其他污染物更突出，其所占的权重达到总体 AQI 的 66%。




图一 污染物占比权重图

## 6. 结论

空气污染是一种全球性问题，对环境和人类健康造成广泛而严重的影响。为了有效地进行空气质量预报和评价，全面评估空气污染对环境的影响，我们需要解决当前空气污染研究中存在的核心问题。

本文利用人工智能理论方法对空气质量进行分析，通过详细的模拟对比验证分析，我们确认了这些模型的有效性。

（1）在预测空气质量方面，我们可以使用 LSTM 模型来学习历史空气质量数据的模式，并预测未来的空气质量趋势。通过将过去一段时间的空气质量数据作为输入，在 LSTM 模型中进行训练，可以使模型学习到时间序列数据中的动态变化和周期性。然后，我们可以利用已训练好的模型，输入当前的环境数据，预测未来的空气质量水平。通过使用 LSTM 模型，我们可以更准确地预测空气质量的变化趋势，及时采取措施来改善和管理空气质量。

（2）随机森林决策树模型通过结合多个决策树的预测

结果来进行最终的预测。每个决策树使用不同的样本和特征进行训练，通过投票或者平均的方式得到最终的预测结果。预测时需要考虑数据的质量和特征选择的合理性，同时，利用均方根误差（RMSE）、平均绝对误差（MAE）等评估指标对模型的性能进行评估。

（3）AQI 模型可以帮助人们了解当前和未来的空气质量状况，为公众和决策者提供有关健康风险和污染管控的信息。然而，需要注意的是，AQI 模型仅是一种综合指数，无法提供详细的污染物浓度和成分信息。在实际应用中，还需要结合其他数据和模型，例如气象模型、空气动力学模型等，来进一步完善空气质量的预报和管理工作。

作者简介：石桂萌（2003.8—），女，汉族，河南省信阳市人，学生，本科在读，学校：杭州师范大学浙江 杭州 310000，研究方向：大数据管理与应用

吕笑语（2003.9–），女，汉族，河北省邯郸市，学生，本科在读，学校：杭州师范大学，浙江 杭州 310000，研究方向：大数据管理与应用

李美莹（2002.11–），女，汉族，河南省新乡人，学生，本科在读，学校：杭州师范大学，浙江 杭州 310000，研究方向：数学。

**参考文献：**

[1] 龙尧水 . 基于神经网络模型和血常规指标的孕妇地中海贫血预测研究 [J]. 国际检验医学杂志 , 2023,44(20)

[2] 刘梦 . 基于流的时间相关特征的 VPN 加密流量识别 [J] 海南师范大学学报 ( 自然科学版 ). 2023,36(03)

[3] 李吉 . 基于数据挖掘的陶瓷艺术品价格预测研究 [D]. 景德镇陶瓷大学 .2023

[4] 闫云凤 . 基于决策森林的回归模型方法研究及应用 [D]. 浙江大学 .2019

[5] 张珉铨 . 雾霾天气下 PM2.5 污染物分布特征研究 [J] 环境科学与管理 . 2023,48(10)

[6] 刘杰 . 北京大气污染物时空变化规律及评价预测模型研究 [D]. 北京科技大学 .2015