基于机器学习的兰州空气污染模拟研究

摘 要：空气污染是气象环境领域研究的重点问题之一。近年来我国已经开展了大量的高分辨率的空气污染自动观测，如何利用这些历史数据进行精确的空气质量预报已成为研究者和业务从事者的研究课题。机器学习近年来广泛运用于预报问题，多种预报模型在空气质量预报上体现出良好的效果，但不同模型在不同地区的预报效果研究中却有差异。本文针对兰州市 2016-2022 年空气污染数据和气象数据，利用随机森林算法（random forest，RF）和长短时记忆网络（long short-term memory，LSTM）对空气质量指数（AQI）进行回溯模拟预报，并用差异融合算法进行差异融合，对三种预测结果进行分析。逐季融合结果 MSE 指标相对 LSTM 下降了 2. 73%，R^2 相对 LSTM 提升了 2. 34%。逐月融合结果相对于逐季融合结果 MSE 指标下降了 12.82%，R^2 指标提升了 10.42%。对 2000 年至 2022 年空气质量指数进行统计分析，发现其中 2012 年相较于 2000 年 API 下降约 35. 62%，2022 年相较于 2014 年 AQI 下降约 28. 87%。在 API 和 AQI 转换年份间（2012 年和 2014 年），AQI 相较于 API 上升 3. 19%。若视 API 和 AQI 可比，则 2022 年相较于 2000 年空气质量指数下降达 52. 74%。对近 20 余年空气质量指数月平均和日平均数据进行 STL 加法分解，每 1000 天日平均污染指数趋势分量平均下降 9. 61，每 10 月月平均污染指数趋势分量下降 2. 89。

1. 引言

空气污染是全球重点研究的环境领域问题之一。已有研究显示，高浓度大气污染物（如 PM2.5）会引发呼吸道疾病、心脏病或其他心血管问题[1-2]，危害人民群众生命安全。在衡量空气污染的指标中，AQI 最为社会所熟知，其数值越大，表明空气污染越严重。为了客观的评价空气污染程度并对空气质量进行合理预测，我国政府现今已经建立了 2700 多个检测站，使用检测仪器 26.8 万余台[3]，积累了大量的空气质量数据。随着社会大众对空气污染的关注度逐渐增高，如何利用已有的大量观测数据进行更加精确的空气质量预报，近年来也受到越来越多研究者和相关业务从事者的重视。

早期人们进行空气质量预报所采用的手段主要是数值模拟，以动力学和大气化学为基础，根据排放源数据和气象数据，通过方程组的形式构建数学模型。数值模拟有其独特的优势，例如 WRF−Chem 模型建立的污染物预报系统，就可以有效地对 PM2.5 和 PM10 进行预报[4]，然而数值模拟也面临着计算成本过高的劣势。机器学习作为新兴的以数据驱动的手段，近年来被广泛运用于空气污染预报领域，如基于 RF 和气象参数构建的模型，在太原市和关中盆地 PM2.5 浓度预报上体现出良好的效果[5-6]。目前学界和业界对基于机器学习的空气污染预报进行了多种尝试，经过合理调参后的模型已经可以达到模式模拟的稳定性和准确度[7-9]，LSTM 空气质量预测模型在 AQI 低于 200 时的预测效果相对更优，且在不同的空气质量等级时选择不同的预测算法进行预测可能会有更优秀的模拟效果[10]。考虑到各个模型在同一变化趋势下的预测准确度不同，特别是空气质量数据在某一个时间范围内发生突发性改变的时候，模型的预测效果差异很大[11]，对模型结果进行比较分析，最后选取合适的结果输出，

就显得尤为重要。

《中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要》中对我国大气污染防治提出了更高的目标，本文面向兰州开展基于机器学习的空气污染模拟研究，有助于推动空气污染预测水平提升，更好保护人民群众生命健康。

2. 资料与研究区域

2.1 数据获取与预处理

课题使用数据从公开数据渠道获取，其中气象数据来自“慧聚数据”，采集了2013年1月1日至2022年12月31日逐日气象数据，数据内容包括日最低温度、日最高温度、日温度、日湿度、日风速、日风级、日能见度、日平均总云量、日气压。

空气污染数据(AQI部分)来自CnOpenData，采集了2016年1月1日0:00至2022年12月31日23:00的AQI值和PM2.5、PM10、SO_2、NO_2、O_3、CO六项污染物的分时数据。逐日数据根据分时数据取24小时平均计算得到。

空气污染数据(API部分)来自青阅数据，采集了2000年至2013年API的逐日数据。

对于数据中的缺失值，缺失一条的以前一条和后一条数据作平均值；连续缺失两条的，以前一条数据作为缺失的第一条数据，再将第一条数据和缺失条的后一条数据做平均作为缺失的第二条数据；连续缺失三条及以上的数据直接舍去。

2.2 研究区兰州市2016年-2020年空气污染特征分析

基于2016年1月1日至2020年12月31日逐日数据，对兰州空气质量变化进行说明。

兰州 2016-2020 年空气质量呈现出明显的季节性变化规律，重度及以上空气污染常常出现于春、冬两季，夏、秋两季污染情况较好。年际上来看，2020年兰州全年未发生重度及以上污染天气，2016-2019 年均有重度污染事件发生。五年期统计来看，80%以上天数处于空气质量良及以上，重度污染和严重污染占约1.8%。

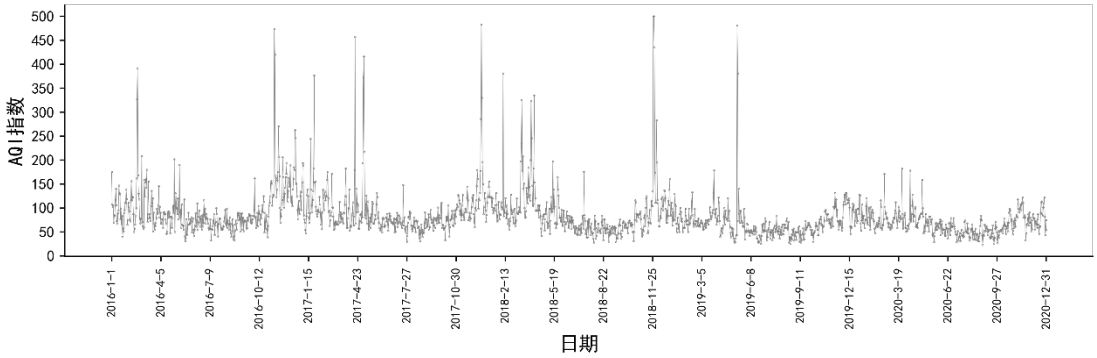

图 1 2016-2020年兰州AQI逐日变化

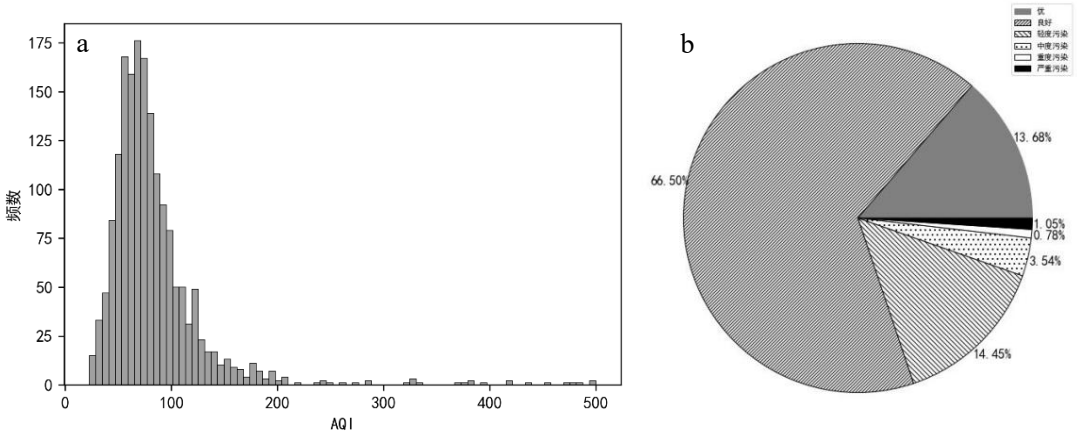

图 2 2016-2020 兰州空气质量分布情况

（图 a 为 AQI 频数分布，图 b 为等级分布）

空气污染监测数据取平均得到其季节和月度变化情况。AQI 指数呈现春季至秋季逐渐减少，冬季反弹至最高的变化趋势，PM2.5、PM10、SO_2、NO_2、CO 与 AQI 变化趋势类似，O_3 则呈现相反的变化态势，春季至夏季 O_3 浓度呈现升高态势至顶峰，随后秋冬两季逐渐减少。AQI 指数在八月达到最低，十二月最高，而臭氧则在八月达到最高，十二月达到最低。

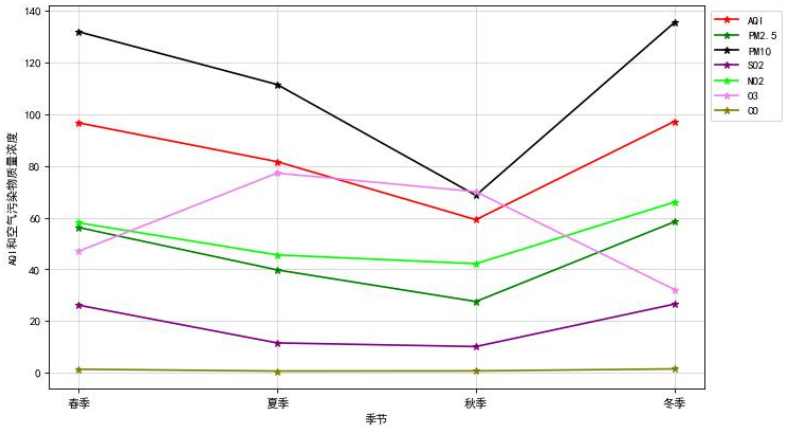

图 3 2016-2020 兰州 AQI 季变化

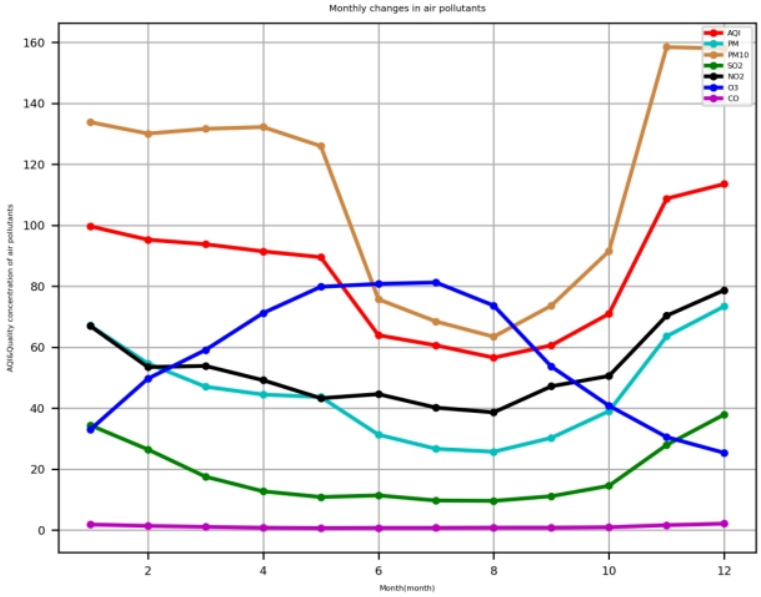

图 4 2016-2020 年兰州 AQI 月变化

3. 兰州市近 20 余年空气质量分析

为未来进一步改进预报模型，并寻找统计规律订正模型的预测结果，根据兰州市 2000 年-2012 年 API 数据和 2014 年-2020 年 AQI 数据，对兰州市空气质量变化情况进行研究。

3.1 兰州市 2000 年至 2012 年首要污染物情况

对兰州市 2000 年至 2012 年首要污染物情况进行分析，在空气质量指数改革以前，API 污染指数仅包含二氧化硫，氮氧化物和可吸入颗粒物（PM10）。兰州市空气质量指数的变化主要由可吸入颗粒物（PM10）构成。

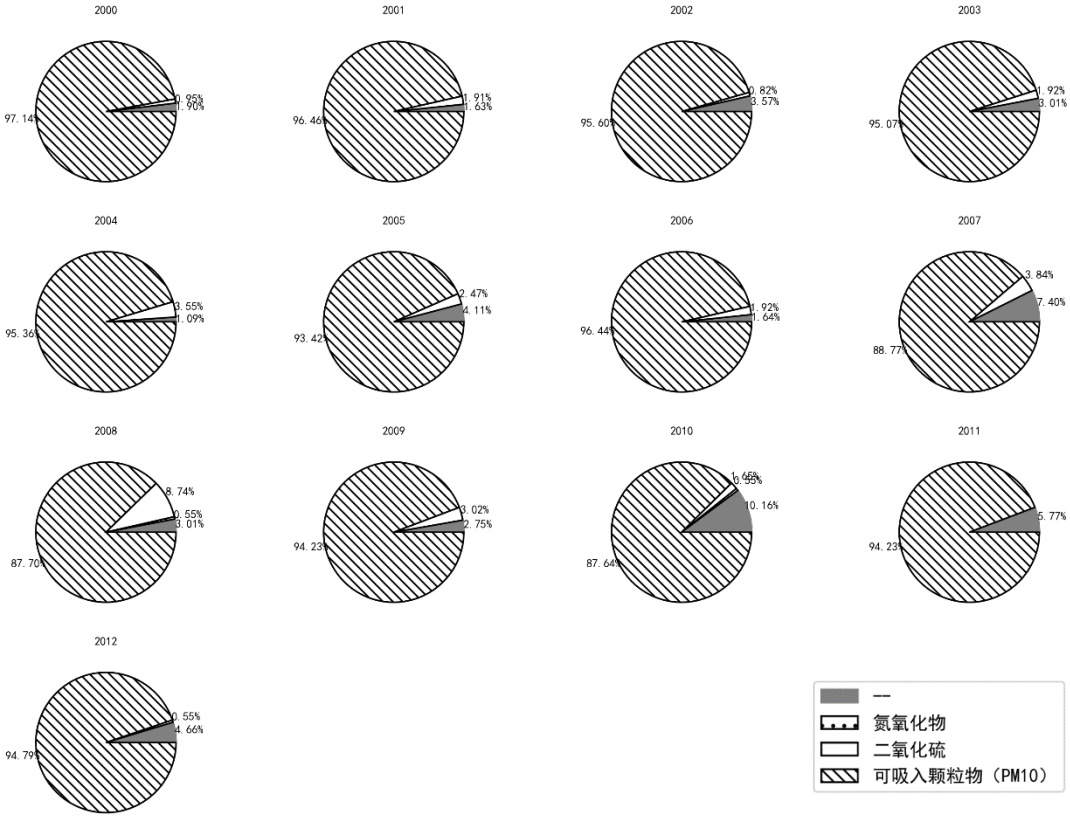

图 5 2000-2012 年兰州首要污染物

据上图可知，兰州市 2000 年至 2012 年使用 API 作为空气质量指数期间，兰州市首要污染物为 PM10，占比全年天数除 2007、2008、2010 三年之外均达到 93%以上。2008 年 SO_2 首要污染物占比天数相对占 13 年期较多。

根据陈雪和陈瑞等[12-13]关于兰州大气污染物特征及变化的研究，兰州市在采取 AQI 作为空气质量指数之后，兰州市首要污染物主要由 PM2.5 和 PM10 共同组成，少数天数存在 O3、SO2 和氮氧化物的首要贡献，但随年份增加 O3 作为首要污染物的天数逐年增加。PM10 和 PM2.5 的分担率合计占比超过 50%。在 2013 年至 2019 年不达标天气及首要污染物天气中，PM10 和 PM2.5 作为首要污染物天数仍超过半数。根据这一特性，我们将兰州市 API 数据和 AQI 数据合并在一起进行研究。

3.2 兰州市空气质量指数近 20 余年统计及分解

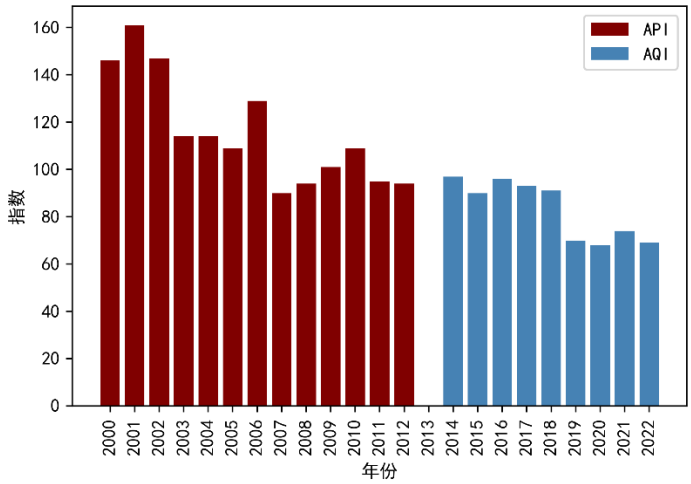

图 6 兰州市空气质量年变化

从兰州市空气质量年数据来看，整体 20 年呈现下降趋势。其中 2012 年相较于 2000 年 API 下降约 35.62%，2022 年相较于 2014 年 AQI 下降约 28.87%。在 API 和 AQI 转换年份间（2012 年和 2014 年），AQI 相较于 API 上升 3.19%。若视 API 和 AQI 可比，则 2022 年相较于 2000 年空气质量指数下降达 52.74%。从下降速度来看，2003 年、2007 年、2019 年是三个空气质量相较于前一年下降较快的年份。

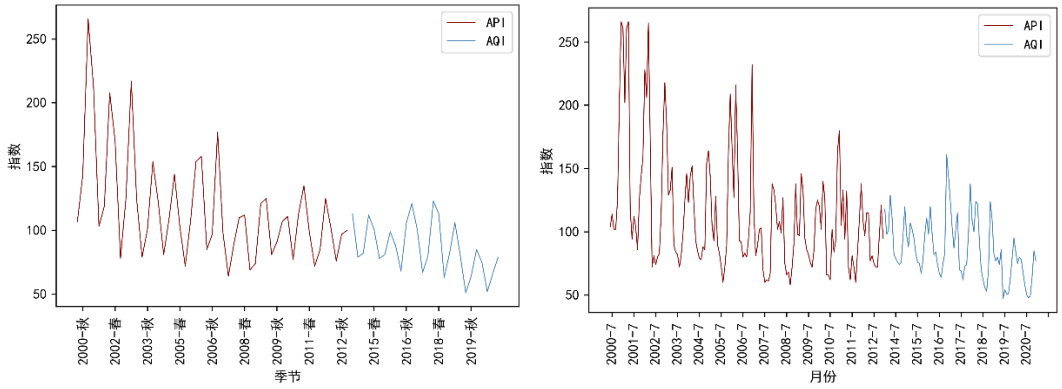

图 7 兰州市空气质量季月变化

从季节和月份变化来看，采用 Mann-Kendall 趋势检验对兰州市季节和月份趋势进行分析，兰州市 API 和 AQI 仍然呈现波动下降趋势。其振幅和极大值、极小值均出现了下降的现象。

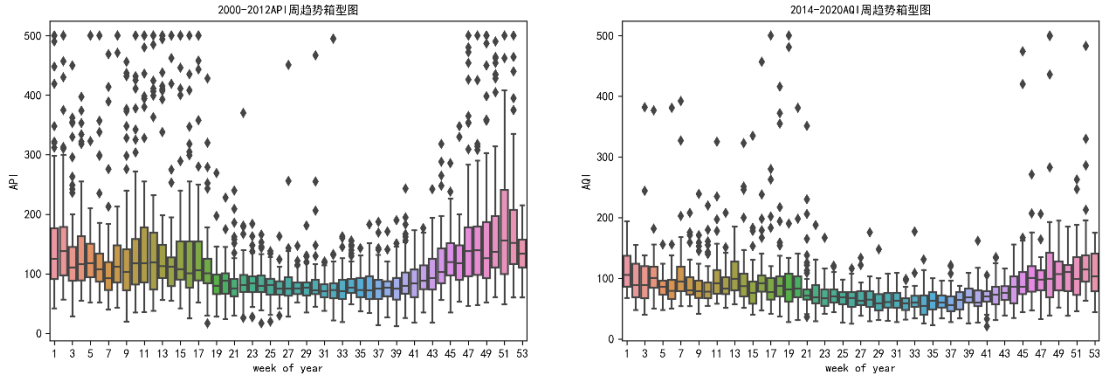

图 8 兰州市空气质量周变化

将全年划分为 53 个周，对兰州市 API 和 AQI 周趋势变化进行分析，可见污染较强的周集中于第 1 周后、第 51 周前后、第 13 周前后。对上述图分析还可发现，AQI 历史数据记载的极值信息相对较少，这一方面说明污染情况在好转，另一方面可能会导致在基于 AQI 模型建立的预报网络中对极值的预报偏弱。尤其在第 21 周至 31 周，对极值数据记载差异相对明显。

原始数据可视为趋势项、季节项和残差项的加法组合，通过对趋势、季节和残差的分析，可以更好的预测未来的情况。若视 API 和 AQI 为可比项，则数据时间长度可拉长至 20 年左右，有利于分析长期的趋势作用。

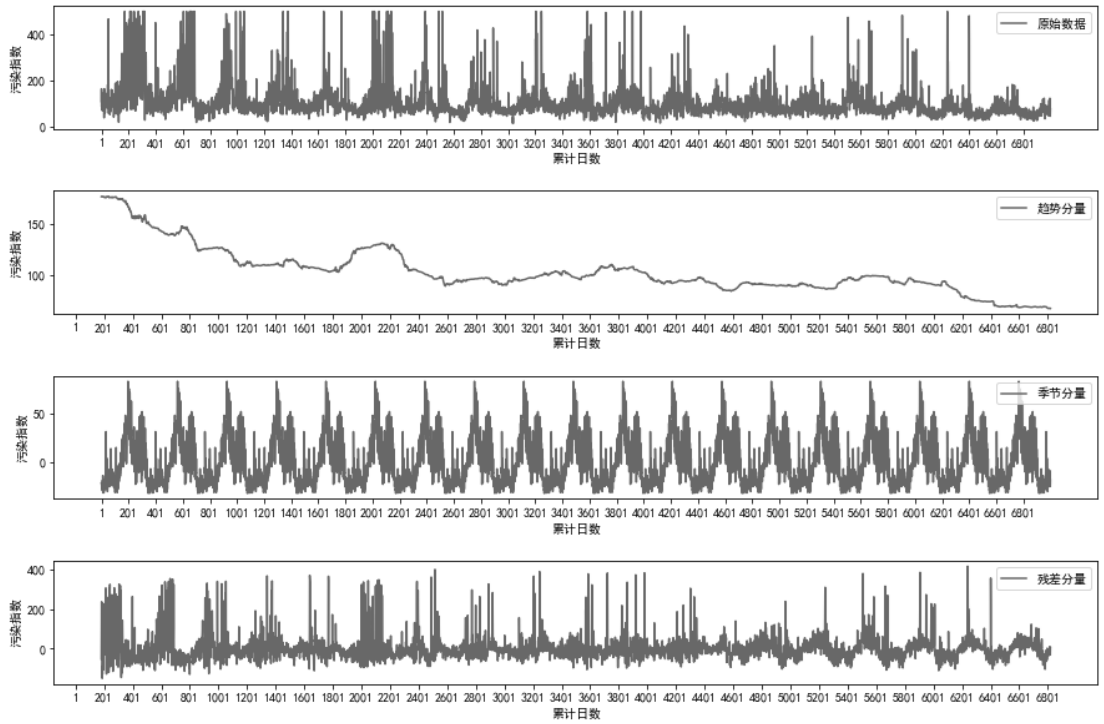

图 9 兰州市日数据 STL 分解

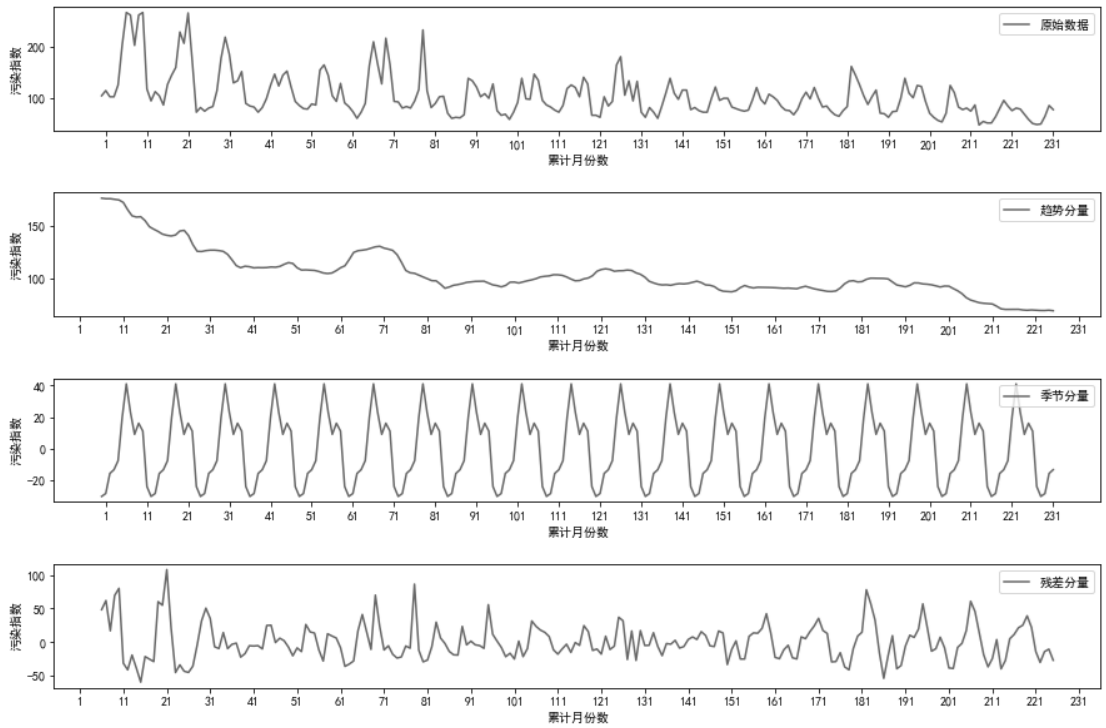

图 10 兰州市月数据 STL 分解

对月平均数据和日数据进行分解后，对趋势分量进行多项式拟合，拟合曲线分别为

$$y = -2.89x + 137.48 \qquad y = -9.61 \times 10^{-3}x + 1.38 \times 10^{2}$$

由拟合曲线系数可知，每 1000 天日平均污染指数趋势分量平均下降 9. 61，每 10 月月平均污染指数趋势分量下降 2. 89。可据此判断未来一定时间内的趋势分量变化。

4. 模型的构建与评价

4. 1 算法简介

随机森林算法：随机森林是由 Breiman 和 Cutler 于 2001 年提出的一种基于决策树的机器学习算法，其结合 Bagging 算法和随机子空间的思想，利用大量决策树对样本进行训练，用于解决分类或回归预测问题。对数据采用 Bootstrap 抽样法，有放回地从 n 个样本中每次抽取一个，抽取 n 次形成样本量为 n 的训练子集。重复 T 次，生成 T 个训练子集。每个训练子集单独构建一棵决策树，构建过程中，在每个节点进行随机特征变量的随机选取，并基于分裂规则比较信息属性进行节点分割。最终以这 T 棵决策树的预测结果的平均值或众数作为最终预测值。

长短时记忆网络算法：LSTM（ Long Short-Term Memory ）算法作为深度学习方法的一种，主要用于处理时间序列数据，最早由 Sepp Hochreiter 和 Jürgen Schmidhuber 与 1997 年提出并发表于《LONG SHORT-TERM MEMORY》这篇文章中。其是一种基于递归神经网络并结合适当梯度学习的算法。具体而言，LSTM 算法在 SimpleRNN 算法的基础上，增

加了相隔多个 TimeSteps 来传递信息的方法，保证在网络训练中，每一个时间节点的信息都能被获取或更新或抛弃，因此能保存较长时间之前的信息，解决了 RNN 算法中梯度消失的问题，从而可以处理 long sequences/timeseries 问题。但同时相比于 SimpleRNN 算法，其计算复杂度高，训练时间更久。

差异融合算法：差异融合算法主要由阈值搜寻法和差异融合法两种算法相互配合进行工作。高嵩所使用的差异融合法有所设阈值和 AQI 真实值的固定差值两个衡量指标。其利用 RF 模型在 AQI 数值较大，高于某个波动阈值时预测结果较优；改进 LSTM 模型在 AQI 数值波动较小，低于某个波动阈值时预测效果较优的特点，对两个衡量指标进行比较。若固定差值大于所设阈值，采用 RF 模型结果；若固定差值小于所设阈值，采用改进 LSTM 模型。最后通过循环得到预测结果。

阈值搜寻法辅助差异融合法选取最佳的融合阈值。阈值搜寻法的衡量指标主要为 MSE 均方差。首先利用 RF 模型和改进 LSTM 模型结合差异融合算法计算出差异融合模型（DFA）的预测值，并得到 DFA、RF 和改进 LSTM 三种模型的 MSE 值。再将 DFA 模型的 MSE 值与 RF 模型和改进 LSTM 模型的 MSE 值的最小值进行比较。若 DFA 模型的 MSE 值小于最小值，则将 DFA 模型的 MSE 值和所得阈值分别装入两个集合之中，以集合中最小的 MSE 值所对应的阈值为阈值搜寻法得到的最佳融合阈值。

4.2 因子及超参数选择

4.2.1 因子选择

结合前人经验，引入常见气象因子温度、湿度、风速、能见度。经过指标挑选，剔除相关系数大于 0.7 的因子以避免多重共线性对模型预报性能的影响，还将日变温、日平均总云量加入预报。考虑气象因子的滞后性和污染物的迟滞，因子选择时记录前一日和前两日的上述因子以及前一日、前两日的 AQI 值。由于使用因子全部为真实值，以该因子序列建立的预报模型可视为提前一日的污染潜势预报。在实际运用中，将真实输入改为 WRF 模式预报值和 AQI 预测值，可进行中长期预报。

为消除不同特征之间不同量纲的影响，使用 Max−min 归一化对训练集和测试集进行处理，使数值落在[0, 1]区间。归一化时训练集和测试集分开。

4.2.2 超参数选择

RF 模型超参数采用随机搜索（Random Search），利用三折交叉验证的方式，以均方根误差为取优标准获得模型最佳参数，模型以训练集上最优模型进行保存。本研究使用的 RF 部分最优超参数为：

n_estimators=600, max_depth=15， min_samples_split=2, min_samples_leaf=1

LSTM 使用的部分最优超参数为：

num_epochs = 1100，learning_rate = 0.001，节点数 128

4.3 模型的结果

根据上述方法，将 2016 年 1 月 1 日至 2020 年 12 月 31 日数据作为训练集，基于 RF 和 LSTM 建立模型。以 2021 年全年日数据作为测试集，得到 RF 和 LSTM 的预报结果。

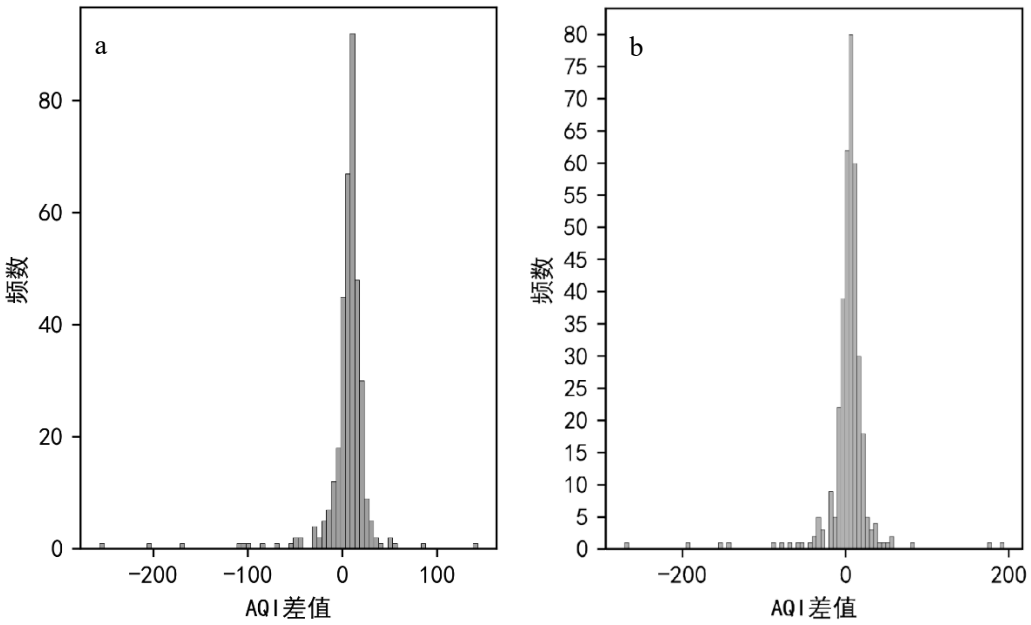

图 11 模型绝对误差频数分布

（图 a 为 RF 预报结果，图 b 为 LSTM 预报结果）

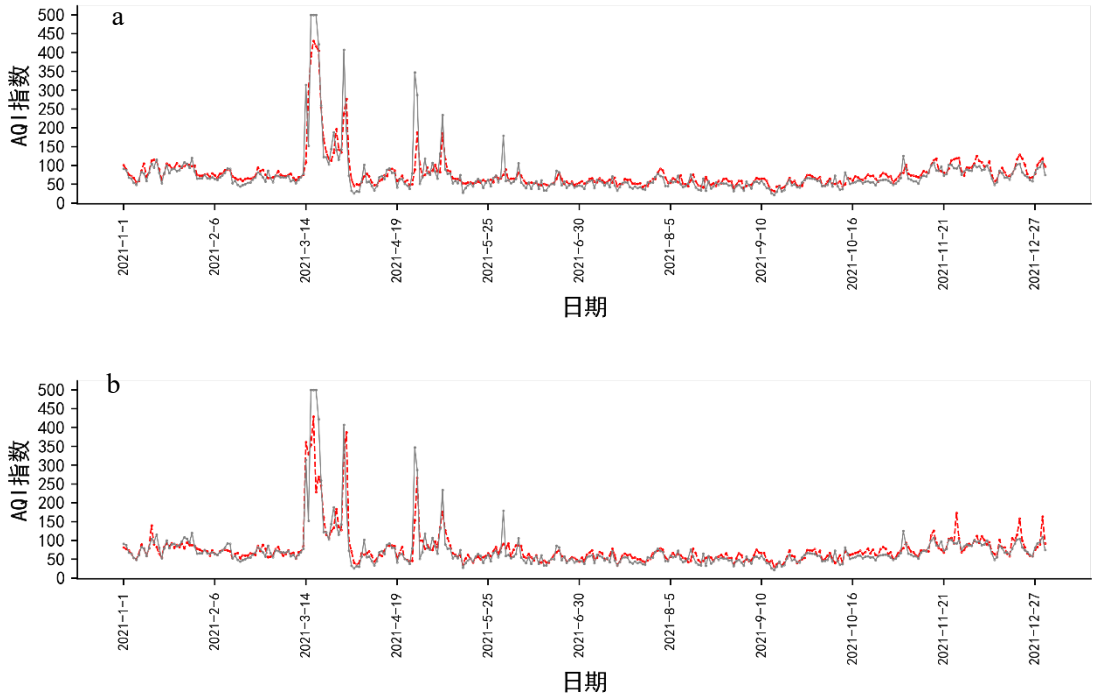

图 12 模型预报结果

（图 a 为 RF 结果，图 b 为 LSTM 结果）

图 11 表示模型的预测值与真值的绝对误差，其中组距为 5。两模型的预报误差分布均为单峰形分布，RF 模型预测绝对误差小于 5 的频数占总频数的 21.64%，绝对误差小于 10 的频数占总频数的 47.40%；LSTM 模型绝对误差小于 5 的频数占总频数的 18.90%，绝对误差小于 10 的频数占总频数的 36.99%。

图 12 展示模型的预测值和真值在 2021 年上的分布情况，由图可见，除少数极值，预测值能够反映大部分时段的趋势变化情况。

模型结果的评价采用决定系数 R^2，平均绝对误差 MAE，均方误差 MSE，均方根误差 RMSE 进行评价。RF 模型预报得到 MAE：14.77，MSE：783.96，RMSE：28.00，R^2：0.7718；LSTM 模型预报 MAE：14.68，MSE：920.74，RMSE：30.34，R^2：0.7320。从评价指标上来看，在 4.2.2 给定参数条件下的 RF 模型的整体全年日预报性能要相对较优，但由于存在调参的误差问题，这里并不重点比较 LSTM 和 RF 在全年日预报中的优劣，而是重点关注其预报结果的差异问题。

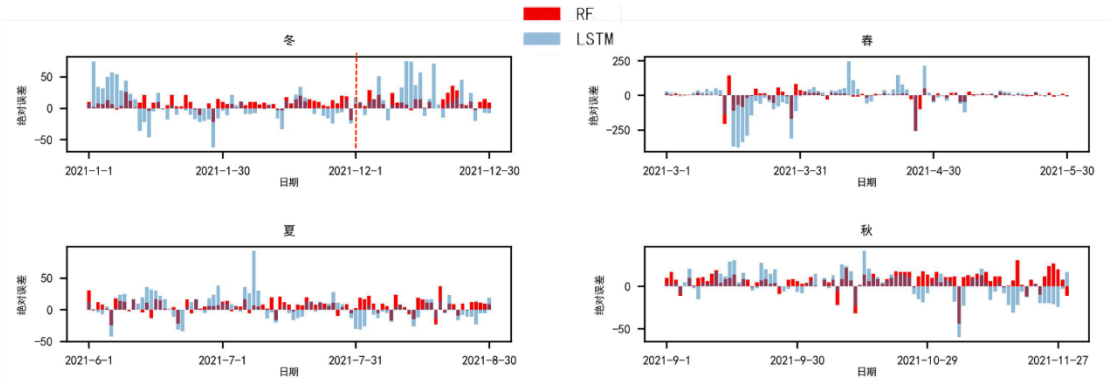

图 13 2021 年模型预测绝对误差

对 RF 和 LSTM 预测结果逐季节分析，发现四季节两模型预报结果存在一定的差异性，春季 RF 和 LSTM 差异性比较显著。上图中冬季结果非连续，虚线前结果为 2021 年 1 月 1 日至 2021 年 2 月 28 日，虚线后结果为 2021 年 12 月 1 日至 12 月 31 日。

本文选取预测日前两日 AQI 差值和预测日 RF 和 LSTM 的预测差值作为差异融合算法的融合阈值的考虑项，最终选取预测日前两日 AQI 差值作为融合阈值。利用 2021 年全年两模型预测数据和真实 AQI 数据对春、夏、秋、冬四季节分别独立进行阈值搜寻，以最小 MSE 作为结束条件，得到春季阈值为 0，夏季阈值为 0，秋季阈值为 4，冬季阈值为 2.

对模型的拓展性进行验证，仍然使用原模型的超参数和阈值，模型对 2022 年全年预报结果如下：

RF 模型结果：MAE：13.36，MSE：451.42，RMSE：21.25，R^2：0.5491

LSTM 结果：MAE：12.96，MSE：461.36，RMSE：21.48，R^2：0.5391

融合结果：MAE：13.22，MSE：448.76，RMSE：21.18，R^2：0.5517

结果显示，使用 2021 年模型上搜寻得到的阈值对 2022 年两模型预报结果进行阈值融合，仍然能够提升模型的预报性能，其中 MSE 指标相对 RF 预报下降了 0.59%，相对 LSTM 下降了 2.73%；RMSE 指标相对 RF 预报下降了 0.33%，相对 LSTM 下降了 1.40%；R^2 指标相对 RF 提升了 0.47%，相对 LSTM 提升了 2.34%。

考虑不同时间尺度差异，利用 2021 年全年两模型预测数据和真实 AQI 数据对每月份分别独立进行阈值搜寻，以最小 MSE 作为结束条件，得到 1-12 月阈值分别为:2,5,0,0,2,0,3,32,2,6,1,2。

对模型的拓展性进行验证，在逐月阈值融合后，融合模型预报性能提升显著，2022 年融合结果如下：MAE: 12.86，MSE: 391.21，RMSE: 19.78，R^2: 0.6092。与 RF 模型对比，MSE 指标相对下降了 13.34%，RMSE 指标相对下降了 6.92%，R^2 指标相对提升了 10.95%；与 LSTM 模型相比，MSE 指标相对下降了 15.21%，RMSE 指标相对下降了 7.91%，R^2 指标相对提升了 13.00%；与使用季节阈值的融合模型相比，MSE 指标相对下降了 12.82%，RMSE

指标相对下降了 6.61%，R^2 指标相对提升了 10.42%。相较于逐季融合结果，逐月融合效果提升相对明显。

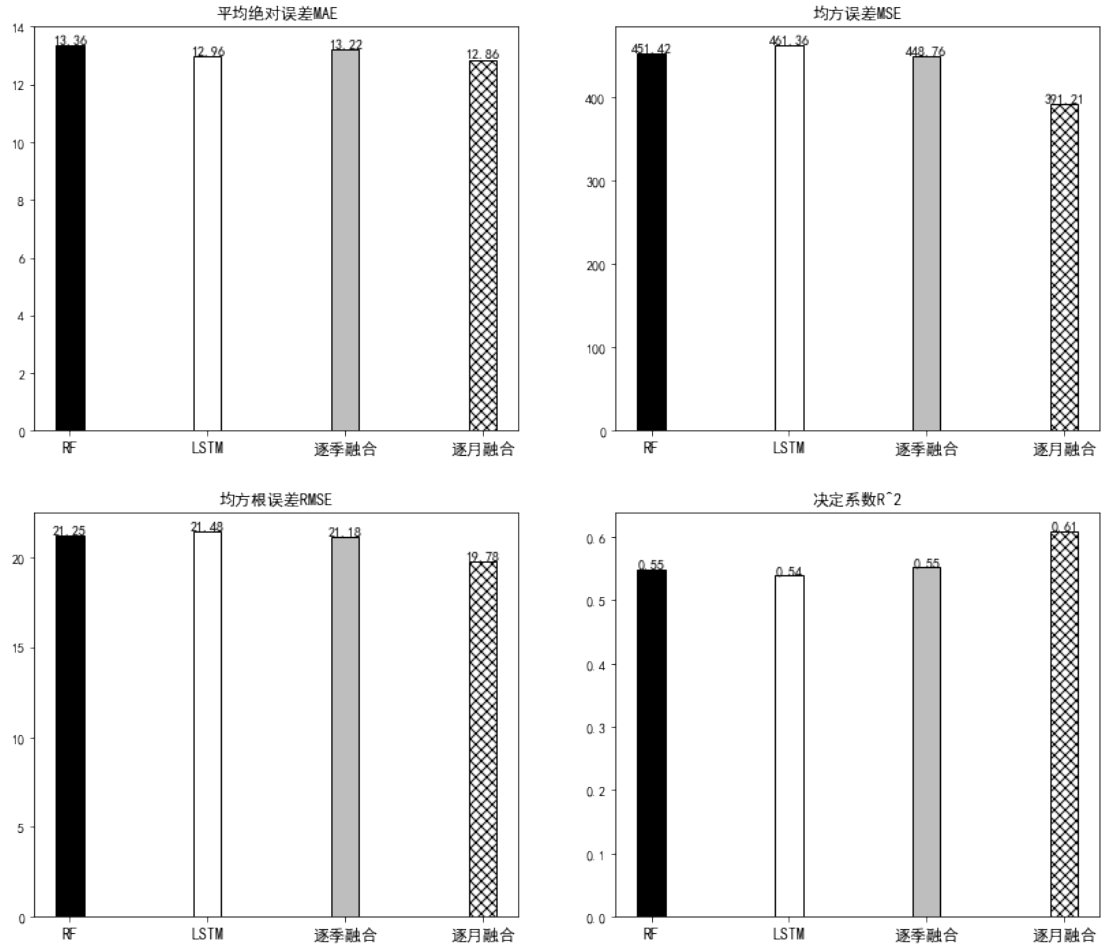

图 14 不同模型 AQI 预报效果对比

5. 小结

对兰州市 2000 年至 2020 年空气污染情况进行长时间分析。得到整体 20 年呈现下降趋势。若视 AQI 和 API 可比，相较 2000 年 2020 年指数下降达 53.42%。对空气质量指数月平均和日平均进行 STL 分解，对趋势分量进行拟合，得到每 1000 天日平均污染指数趋势分量平均下降 9.61，每 10 月月平均污染指数趋势分量下降 2.89。

基于 RF 和 LSTM 算法经过调参后建立的模型对 2021 年预报效果均较好，RF 模型的 MAE：14.77，MSE：783.96，RMSE：28.00，R^2：0.7718；LSTM 模型预报 MAE：14.68，MSE：920.74，RMSE：30.34，R^2：0.7320。且模型预报效果四季节分别存在不同的差异性。

根据差异融合方法取前两日 AQI 差值进行寻优，得到四季融合阈值，经过逐季融合后预报效果有一定上升。对 2022 年进行模型拓展性验证，2022 年预报结果相对 RF 提升较为微弱，对 LSTM 提升相对明显。相对 RF 提升微弱，可能是由于调参导致的 RF 和 LSTM 在模型预报精确度上 RF 全年整体明显好于 LSTM 造成的。这揭示了该算法使用前必须控制所需融合的模型的预报性能。相较于逐季融合模型，逐月融合模型预报效果提升明显，R^2 提升和 MSE 下降相对逐季融合模型均大于 10%，体现兰州市融合阈值方法需要精细化分析。

参考文献

[1] Qiu H, Tian LW, Pun VC, et al. Coarse particulate matter associated with increased risk of emergency hospital admissions for pneumonia in Hong Kong[J]. Thorax, 2014, 69(11): 1027–1033.

[2] 齐爱, 张亚娟, 杨惠芳. 大气 PM2.5 对心血管系统影响及其作用机制研究进展[J]. 环境与健康杂志, 2016, 33(5): 465–469.

[3] Zhongshan Yang,Jian Wang. A new air quality monitoring and early warning system: Air quality assessment and air pollutant concentration prediction[J]. Environmental Research,2017,158:

[4] Pablo E. Saide,Gregory R. Carmichael,Scott N. Spak,Laura Gallardo,Axel E. Osses,Marcelo A. Mena-Carrasco,Mariusz Pagowski. Forecasting urban PM10 and PM2.5 pollution episodes in very stable nocturnal conditions and complex terrain using WRF–Chem CO tracer model[J]. Atmospheric Environment,2011,45(16):

[5] 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.

[6] 苏雨萌. 基于机器学习方法的关中盆地 PM_(2.5)浓度的模拟和预报[D].兰州大学,2021.

[7] Tuan V. Vu,Zongbo Shi,Jing Cheng,Qiang Zhang,Kebin He,Shuxiao Wang,Roy M. Harrison. Assessing the impact of clean air action on air quality trends in Beijing using a machine learning technique[J]. Atmospheric Chemistry and Physics,2019,19(17):

[8] Unjin Pak,Jun Ma,Unsok Ryu,Kwangchol Ryom,U. Juhyok,Kyongsok Pak,Chanil Pak. Deep learning-based PM2.5 prediction considering the spatiotemporal correlations: A case study of Beijing, China[J]. Science of the Total Environment,2020,699(C):

[9] Junshan Wang,Guojie Song. A Deep Spatial-Temporal Ensemble Model for Air Quality Prediction[J]. Neurocomputing,2018,314:

[10] 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.

[11] 高嵩,何卓骏,刘子岳,刘家明,王刚,李登柯.基于机器学习的差异融合分析在空气质量预

测中的应用[J].电子测量技术,2021,44(18):85-92.

[12] 陈瑞,孙建云,魏巧珍等.2014—2020 年兰州市大气污染物特征及变化趋势分析[J].卫生研究,2021,50(05):769-774.

[13] 陈雪. 2013-2019 年兰州市城市环境空气质量变化趋势研究[D].兰州大学,2021.