

基于 MAIAC AOD 的长株潭城市群 PM_{2.5} 浓度时空分布模拟

邵光达 韩亮 姜柳琦 方新*

(湖南城市学院市政与测绘工程学院, 湖南 益阳 413000)

摘要: 本文基于 PM_{2.5} 浓度、气溶胶光学厚度、气象数据、土地利用占比, 以及景观指数构建随机森林模型, 并通过模型预测 PM_{2.5} 浓度形成一个高分辨率全覆盖的长株潭城市群的 PM_{2.5} 浓度数据集。随机森林模型评价结果 R² 为 0.9617, MSE 为 11.0563, MAE 为 2.1864, 解释度为 0.9617。2022 年长株潭城市群 AOD 季节分布上表现为冬季 (0.481) > 春季 (0.451) > 秋季 (0.416) > 夏季 (0.313), 空间分布上呈现为各城市市区 AOD 值偏高, 向四周延伸逐渐降低; 整体上呈现为北部向南部逐渐减低趋势。

关键词: PM_{2.5}; 气溶胶光学厚度; 时空分布; 随机森林建模

随着工业化和城镇化的高速发展, 空气污染已成为全球非常严峻的环境问题之一, 而现阶段我国的各类空气污染问题尤为突出, 其中首要污染物为大气细颗粒物 (Particle Matter, PM_{2.5})。PM_{2.5} 又称细颗粒物, 指的是空气动力学当量直径小于等于 2.5 微米的颗粒物。随着工业化和城镇化的高速发展, 空气污染已成为全球非常严峻的环境问题之一, 而现阶段我国的各类空气污染问题尤为突出, 其中首要污染物为大气细颗粒物。PM_{2.5} 又称细颗粒物, 指的是空气动力学当量直径小于等于 2.5 微米的颗粒物。AOD (Aerosol Optical Depth) 是指通过在大气柱中的气溶胶的散射和吸收来测量消光, 它与细颗粒的负载量有关。与传统的地面监测相比, 通过卫星监测可以突破地面监测站点选址的限制, 在时间和空间覆盖度上具有更大的优势。

卫星遥感反演的 AOD 已被普遍应用于近地面 PM_{2.5} 预测, 人们通过构建基于 AOD 的模型实现了对 PM_{2.5} 的估算, 如郭建平等人以 2007–2008 年 AOD 的数据为基础, 结合多种气象因素, 建立了 BP 神经网络模型来估计 PM_{2.5} 的浓度。杨晓辉、肖登攀等为研究京津冀地区 PM_{2.5} 浓度时空变化趋势, 利用多角度大气校正 (MAIAC) 气溶胶光学厚度 (AOD) 产品, 结合气象和土地利用等数据, 构建线性混合效应 (LME) 和地理加权回归 (GWR) 组成的两阶段统计回归模型, 建立了 2013–2020 年 1km 空间分辨率的 PM_{2.5} 浓度数据集。杨云等利用宝鸡市监测站的每小时监测数据, 建立了遗传算法改进后的 PM_{2.5} 小时浓度的预测模型, 并与 BP 神经网络的预测结果进行了比较, 结果显示, 改进后的 BP 神经网络的预测结果在精度上得到了提高。孙津利用与 PM_{2.5} 浓度存在强关联的卫星气溶胶光学厚度 (AOD), 使用多模型综合的随机森林方法预测了全覆盖的高时空分辨率 PM_{2.5} 浓度分布并结合北京市的个体时空轨迹数据进行了个体短期暴露评估试验。

一、研究数据与方法

(一) 研究数据

1. 地面 PM_{2.5} 监测数据

PM_{2.5} 监测数据作为随机森林回归的“因变量”, 用在模型

的训练和检验当中。本文采用的地面监测数据为中国环境监测总站发布的逐时浓度数据, 其中研究区共有 24 个国控监测站, 包括长沙市 10 个, 湘潭市 5 个, 株洲市 9 个。

2. MAIAC AOD 数据

MAIAC 是一种处理 MODIS 数据的新算法, MAIAC 通过时间序列分析和影像处理法对亮地表区和暗地表区进行大气校正和气溶胶反演, 在暗地区 MAIAC AOD 与传统的 MODIS AOD 算法的反演精度相似在亮地表区 MAIAC AOD 精度更高, 且其分辨率和有效观测范围得到了很大的提升。本文选取 2022 年覆盖长株潭城市群的 MCD19A2 产品, 共 730 期, 包括 h27v06、h28v06 这 2 个条带的 AOD 数据, 空间分辨率为 1 km, 时间分辨率为 1 d。

3. 土地利用数据

土地利用数据可以用于分析判断各类型土地利用数据与其他因素之间的相关性, 揭示其他要素的空间变化特征。本文使用的土地利用数据产品来自于中国自主研制的 30 米空间分辨率的全球地表覆盖数据。

4. 气象数据

ERA5 是 ECMWF (欧洲中期天气预报中心) 对 1950 年 1 月至今全球气候的第五代大气再分析数据集, 由 ECMWF 的哥白尼气候变化服务 (C3S) 生产。本研究使用的气象数据为 ERA5 DAILY 所提供的 2022 年 3 个 ERA5 气候再分析参数的汇总值: 2m 空气温度、2m 露点温度、表面气压。此数据集可通过 ERA5 官网 (表 2.5) 下载, 时间分辨率为 Hourly、空间分辨率为 0.1° × 0.1°。

(二) 研究方法

1. 随机森林模型基本原理

随机森林是 Breiman 提出的一种简洁有效的决策树类机器学习方法, 决策树是一种通过各种特征变量预测目标类别或值的树形结构, 可以用于分类或回归等相关研究。决策树的每个结点是基于一种特征变量对样本的一次分裂, 最优分裂导向最佳预测效果。但是单棵决策树是一个弱学习器, 容易造成过度学习样本本身而非总体趋势的过拟合现象。而随机森林的思想就是通过多棵

决策树集成“森林”来增强它的估测效力，并且通过注入“随机性”使得树与树之间差别增大，减小趋同可能带来的过拟合现象，整体更加健壮。

因为使用了集成算法，随机森林算法的精度要比大部分单个算法要好，它的精度非常高，在测试集上有很好的表现，同时它引入了随机性，不会轻易地陷入过拟合，可以更好地处理离散型数据，即便是在异常值和缺失值的问题上，它也可以得到很好的效果。

2. 随机森林模型构建

本文利用 python 编写程序构建随机森林模型，使用的样本数据包含长株潭城市群范围内的 24 个监测站点的 PM_{2.5} 浓度、AOD 数据、温度、气压、露点温度、建设用地占比、草林地占比、耕地占比、湿地水体占比、PD 指数、LSI 指数以及 AI 指数。

在建模工作的初始阶段，需要把数据分成训练和测试两个部分，这对于获得模型性能的实际估计至关重要，其中测试集部分只被用来评估模型的训练效果，以及调整模型的超参数 (hyperparameter)。首先，需要将目标 (y) 特征与输入 (X) 特征分开，我们可以利用 Scikit-Learn 的 train_test_split 功能将 20% 的数据作为测试集来评估我们的模型。模型筛选完所用估测变量后，可以对模型调参以达到最优效果。本实验所用调参方法是使用网格搜索 + 交叉验证进行参数的优化，在 scikit-learn 中，RF 的分类器是 RandomForestClassifier，回归器是 RandomForestRegressor。经过网格搜索 + 交叉验证，模型调优后本实验使用的参数设置为 criterion=absolute_error、bootstrap=True、max_features=auto、max_depth=70、min_samples_split=2、n_estimators=600、min_samples_leaf=2。

3. 模型验证与评价

综合前人研究的验证方法，我通过用 4 个指标对随机森林估测效果进行评价。4 个指标分别是估测值和实际值之间的可决定系数 R²、均方误差 MSE、平均绝对误差 MAE 以及可解释方差，通过模型所反馈的各项指标数值可对模型的优劣进行评估。

表 1 随机森林回归模型评估结果表

评估	训练集	验证集
R ²	0.9617	0.9136
均方差	11.0563	26.0129
绝对差	2.1864	3.7303
解释度	0.9617	0.9136

根据构建的随机森林模型评价结果表明，随机森林模型对长株潭城市群 PM_{2.5}-AOD 之间的时空变化模拟能够展现优异的性能。

二、长株潭全域 PM_{2.5} 浓度时空特征分析

长株潭城市群 PM_{2.5} 浓度四季均浓度分别为 29.218 μg/m³、15.86 μg/m³、31.618 μg/m³、37.436 μg/m³，在时间分布上呈现“冬季高夏季低，春秋过渡”的季节变化特征。由于春季气温

的变化起伏不定，有“倒春寒”的说法，春季冷空气活动较弱，天气时晴时雨，温度波动大，使得空气中的污染物不易扩散，且随着温度上升，污染逐渐积累；同时春季天气多变，不稳定，加上长株潭城市群春季空气湿度比较高，对大气污染物有很强的吸附力，对流不强，颗粒物很难散失；重工业尾气的排放以及建筑用地的影响等原因导致春季的局部地区 PM_{2.5} 浓度居高不下；受到季风气候的影响，伴随着北方季风而来的污染物使得 PM_{2.5} 浓度比较高。

在四个季节中，夏季是 PM_{2.5} 浓度最小的季节，其最高值仅为 26.218 μg/m³。由于长株潭城市群的雨季比较长，从三月份开始，一直持续到七八月份，夏天的气候变化很大，降雨很多，对大气中的细颗粒物有很大的影响，加速了细颗粒物的沉降消失；而且夏天的温度很高，阳光很强，导致地表温度和大气温度都很高，大气中的细颗粒物很不稳定，导致了空气中的细颗粒物在夏天的扩散速度很快；同时由于夏季植被茂盛，净化作用加强也促使 PM_{2.5} 的浓度在夏天表现为最低。

长株潭城市群冬季降水少，空气相对湿度较低，加之冬季植被稀疏，对大气颗粒物的吸收、沉降能力弱；同时长株潭城市群冬季污染物排放量高，包括取暖活动中燃煤量增加，和气温骤减促使汽车出行的频率增加，加之春节期间鞭炮齐鸣，使得长株潭城市群冬季 PM_{2.5} 的平均浓度高达 37.436 μg/m³。

三、结论

(1) 随机森林模型评估结果 R² 为 0.96，MSE 为 11.05，MAE 为 2.18，解释度为 0.96，表明随机森林模型能较好地预测长株潭城市群全域的 PM_{2.5} 浓度。

(2) 2022 年长株潭城市群 AOD 季节分布表现为冬季 (0.481) > 春季 (0.451) > 秋季 (0.416) > 夏季 (0.313)，存在明显的季节变化特征；长株潭城市群 AOD 在空间分布上与土地利用类型有着较强关联，各城市市区 AOD 值偏高，向四周延伸逐渐降低；针对长株潭城市群全域 AOD 分析可知，北部地区 AOD 明显高于南部地区，呈现由北部向东南部逐渐降低趋势。

参考文献：

[1] 杨晓辉, 肖登攀, 柏会子, 唐建昭, 王卫, 郭风华, 刘剑锋. 2013-2020 年京津冀地区 PM_{2.5} 浓度时空变化模拟及趋势分析 [J]. 地理与地理信息科学, 2022, 38 (04): 58-67.

[2] 孙津. 结合气溶胶光学厚度的北京市 PM_{2.5} 小时浓度分布估计与个体暴露评估 [D]. 中国科学院大学 (中国科学院空天信息创新研究院), 2021.

资助信息：国家大学生创新训练项目 (No. S202211527013)。

通讯作者：方新 (博士、教授)，E-mail: xinfang@hncu.edu.cn，主要从事时空地理信息建模研究。