Pure Mathematics 理论数学, 2024, 14(6), 447-459
Published Online June 2024 in Hans. https://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2024.146263

Hans汉斯

基于MAIAC AOD时空补值数据的$PM_{2.5}$浓度估算研究

熊英杰，杜　宁，王　莉，王　耀

贵州大学矿业学院，贵州　贵阳

收稿日期：2024年5月7日；录用日期：2024年6月7日；发布日期：2024年6月30日

摘　要

气溶胶光学厚度被广泛应用于$PM_{2.5}$浓度估算中，受极端气候影响以及卫星传感器影响，AOD数据存在大量缺失，本文提出Prophet-LSTM + P-Bshade时空补值模型对MAIAC AOD数据进行补值并使用Catbooost模型结合AOD数据以及ERA5气象数据对中国2020年陆地区域的$PM_{2.5}$浓度进行估算。结果表明：① Prophet-LSTM + P-Bshade时空补值模型精度明显优于传统补值方法，R、MASE和MAE分别为0.891、0.275和0.183。② Catboost模型在$PM_{2.5}$浓度估算中比常用的其他机器学习等模型显示更高的估算精度，R、MASE和MAE分别为0.93、15.89 μg·m^{-3}和10.54 μg·m^{-3}。③ 中国陆地区域2020年的$PM_{2.5}$浓度在季节尺度分布上明显，整体呈现冬季 > 春季 > 秋季 > 夏季的季节分布特点。在空间分布上，$PM_{2.5}$浓度整体呈现东部地区较高，塔里木盆地区域局部较高的特点。

关键词

MAIAC AOD，Prophet-LSTM + P-Bshade，时空补值，$PM_{2.5}$，Catbooost

Research on $PM_{2.5}$ Concentration Estimation Based on MAIAC AOD Spatiotemporal Supplement Data

Yingjie Xiong, Ning Du, Li Wang, Yao Wang

Mining College of Guizhou University, Guiyang Guizhou

Received: May 7th, 2024; accepted: Jun. 7th, 2024; published: Jun. 30th, 2024

Abstract

Aerosol optical thickness is widely used in $PM_{2.5}$ concentration estimation, due to the influence of

文章引用: 熊英杰, 杜宁, 王莉, 王耀. 基于MAIAC AOD时空补值数据的$PM_{2.5}$浓度估算研究[J]. 理论数学, 2024, 14(6): 447-459. DOI: 10.12677/pm.2024.146263

extreme climate and satellite sensors, there are a large number of missing AOD data, this paper proposes the Prophet-LSTM + P-Bshade spatiotemporal compensation model to supplement the MAIAC AOD data, and uses the Catbooost model combined with AOD data and ERA5 meteorological data to estimate the $PM_{2.5}$ concentration in the land area of China in 2020. The results show that: (1) The accuracy of the Prophet-LSTM + P-Bshade spatiotemporal compensation model is significantly better than that of the traditional compensation method, with R, MASE and MAE of 0.891, 0.275 and 0.183, respectively. (2) The Catboost model showed higher estimation accuracy than other commonly used machine learning models in $PM_{2.5}$ concentration estimation, with R, MASE and MAE of 0.93, 15.89 μg·m^{-3} and 10.54 μg·m^{-3}, respectively. (3) $PM_{2.5}$ concentrations in China's land areas in 2020 were significantly distributed on a seasonal scale, showing the seasonal distribution characteristics of winter > spring > autumn > summer. In terms of spatial distribution, $PM_{2.5}$ concentrations were higher in the eastern region and higher in the Tarim Basin.

Keywords

MAIAC AOD, Prophet-LSTM + P-Bshade, Spatiotemporal Supplement, $PM_{2.5}$, Catbooost

1. 引言

我国经济水平发展迅速，城镇化水平不断提升，人们生活水平日益提高，包括 SO_2、CO_2、粉尘等污染物的排放也逐渐增多，加重了我国大气生态环境污染情况，导致霾污染天气频发，雾和霾均属于大气气溶胶系统的组成部分，但是霾的核心是一些粉尘微粒，是空气动力学直径为 2.5 微米或更小的微粒($PM_{2.5}$)大量聚积的过程[1] [2]。由于 $PM_{2.5}$ 直径极其微小，导致其可以通过人的呼吸进入人们的呼吸道以及肺泡中，在空气中携带的有毒物质和重金属污染物也会附着在这些细小颗粒物上随之进入人体参与血液循环，对人体造成严重损伤[3] [4]。对 $PM_{2.5}$ 的浓度进行估算是有重要意义的研究。

$PM_{2.5}$ 主要通过地面空气质量监测网站获取，由于我国监测网络运行较晚，自 2013 年才开始运行，且地面监测站点分布不均，在一定程度上影响了对 $PM_{2.5}$ 污染影响程度评估的能力[5]，无法支撑我国大范围的 $PM_{2.5}$ 浓度监测需求，所以越来越多的学者选择使用气溶胶光学厚度来进行 $PM_{2.5}$ 浓度估算，能有效弥补地面站点不足的缺点[6] [7]，气溶胶光学厚度(Aerosol Optical Depth, AOD)为介质的消光系数在垂直方向上的积分，用来描述气溶胶对光的削减作用，是气溶胶最重要的参数之一，AOD 是一个无量纲正值。卫星遥感 AOD 数据可通过不同的数据模型进行 $PM_{2.5}$ 的浓度估算，证明了 $PM_{2.5}$ 与 AOD 之间会随着时间和空间的变化而相应产生变化[8]，卫星遥感气溶胶光学厚度已经被广泛应用于 $PM_{2.5}$ 的浓度估算中[9] [10]，在吴宇宏等[11]的研究中，AOD 因子与 $PM_{2.5}$ 的相关因系数高达 0.65，在吴迪[12]的研究中，AOD 因子与 $PM_{2.5}$ 浓度的相关系数为 0.50。通过卫星获取的 AOD 有着大范围高精度的优点，但是当遇到多云、多雨或者雪覆盖导致地面反射率增高时，卫星遥感获取的 AOD 数据会存在大范围的缺失。缺失的 AOD 数据对 $PM_{2.5}$ 浓度估算精度有一定的影响，所以在对 PM2.5 进行估算时，对 AOD 数据进行补值是必要的处理。

传统 AOD 补值方法包括克里金插值、泛克里金插值和反距离加权等[13]，在研究 $PM_{2.5}$-AOD 的时空关系中，常用的统计方法有线性回归模型、多元线性回归模型[14] [15]、线性混合效应(LME)模型[16]、

广义加法模型(GAM) [17]和地理加权回归模型(GWR) [18] [19]等，KIM [15]等人利用 MODIS AOD 数据和部分 ERA5 气象再分析数据建立了一个多元线性回归(MLR)模型对 $PM_{2.5}$ 浓度进行估算，总体相关系数大于 0.8。这些实验中使用的 AOD 补值方法均为传统补值方法，对 AOD 的时空分布会产生一定的影响，进而影响 $PM_{2.5}$ 浓度估算的精度。为进一步提高精度，有学者在混合效应模型中加入与 $PM_{2.5}$ 相关的时间和空间辅助变量，包括气温、降水、REA5 气象再分析资料、土地利用数据和人口数据等变量，相较于普通的线性回归模型精度有所提高。也有学者用不同的方法对 AOD 进行补值得到更好的 $PM_{2.5}$ 浓度估算结果，Liu [20]等人将 MERRA-2 AOD 进行重采样处理为 Himawari-8 AOD 进行数据填补，并将填补好的 Himawari-8 AOD 和 NDVI、路网、人口密度和坐标等作为预测因子使用随机森林模型对地面 $PM_{2.5}$ 进行浓度估算，结果显示 R^2 为 0.88。机器学习也常用于 $PM_{2.5}$ 的浓度估算中，Li [21]等人利用 MAIAC AOD 数据，结合其他时空预测因子和空间自相关性构建了一个每日混合效应空间模型，利用嵌入的结构化和非结构化空间随机效应来解释空间自相关性。然后基于自举聚合(Catboost)对基本模型的点估计进行平均，以减少预测中的方差。然后，开发了约束优化，对 $PM_{2.5}$ 浓度的完整时间序列进行了预测。

在 $PM_{2.5}$ 浓度估算的众多常用变量中，AOD 与 $PM_{2.5}$ 浓度的相关性较强，但卫星遥感 AOD 数据存在大面积缺失，传统的 AOD 补值方法只顾及了该数据的空间相关性或者时空相关性，没有综合考虑两者的互相关关系，对补值结果会产生一定的影响，从而影响 $PM_{2.5}$ 浓度估算的精度。因此，本文提出一种新的 AOD 补值方法对 MAIAC AOD 进行补值，并将 AOD 补值结果结合 ERA5 气象再分析数据使用 Catboost 算法进行 $PM_{2.5}$ 浓度进行估算。

2. 研究概况

2.1. 研究区概况

本次研究区域为中华人民共和国的陆地部分，中国陆地总面积约为 960 万平方千米。目前中国有 34 个省级行政区，包括 23 个省、5 个自治区、4 个直辖市和两个特别行政区(https://www.gov.cn/guoqing/index.htm)。按行政区主要划分为七个大区，分别是东北地区、华北地区、华中地区、华南地区、华东地区、西北地区和西南地区，研究区域如图 1 所示。

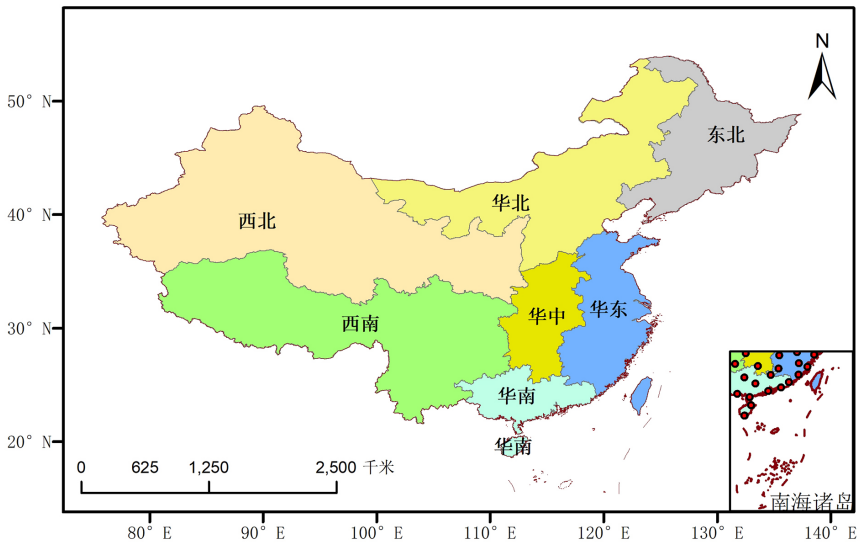

Figure 1. Overview of the study area and distribution of $PM_{2.5}$ stations
图 1. 研究区概况及 $PM_{2.5}$ 站点分布

2.2. 数据来源及处理

本研究使用数据包括 MAIAC AOD 数据、AERONET AOD 数据、ERA5 气象再分析数据和 $PM_{2.5}$ 站点数据。其详细信息及来源介绍如下。

2.2.1. MAIAC AOD 数据

美国国家宇航局(NASA)分别于 1999 年和 2002 年发射了 Terra 卫星和 Aqua 卫星，搭载于两个卫星上的 MODIS 传感器提供了 AOD 数据[22]，多角度大气校正算法(Multiangle Implementation of Atmospheric Correct, MAIAC) AOD 是 Lyapustin 等人发布的全球覆盖高空间分辨率(1 km)每日 AOD 数据集(MCD19A2)，MAIAC AOD 属于 MODIS 的 L2A 级产品[23]。MCD19A2 Version 6 的数据在 2023 年 7 月 31 日之后弃用，官方更新了 MCD19A2 Version 6.1 数据[24]，该数据将 AOD 值的有效范围从第六版本的 0-3 更新到了现在的 0-6，数据可以从 NASA 官网网站进行获取(https://ladsweb.modaps.eosdis.nasa.gov/)。根据最新的 MAIAC 数据操作指南对该数据进行下载中国陆地区域 2016 年 3 月~2021 年 2 月的 AOD 数据文件进行预处理并质量控制，选取 AOD_QA 为“0000”代表 Best quality 的数据并将 MCD19A2 中 Terra 和 Aqua 卫星的 Optical_Depth_055 波段数据融合作为实验数据。

2.2.2. AERONET AOD 数据

AERONET AOD 常作为卫星遥感 AOD 的验证数据，可以从全球布站的气溶胶特性地基观测网(http://aeronet.gsfc.nasa.gov/)进行获取。MODIS 过境时间为上午 10 点 30 分(Terra 卫星过境时间)和下午 13 点 30 分(Aqua 卫星过境时间)，以 AERONET 站点为中心采用 50 km × 50 km 的空间窗口计算卫星的 AOD 空间平均值，匹配地面观测点时间 ± 30 min 的 AERONET AOD 平均值[25]，以此进行验证分析。

MAIAC AOD 数据是 550 nm 波段处的数据，AERONET 站点提供的 AOD 数据并不包括 550 nm 波段的 AOD 数据，因此，需要通过 AERONET AOD 440 nm 和 675 nm 两个波段的 AOD 插值出 550 nm 波段的 AOD 值[26]，波长计算公式如下：

$$\alpha_{\lambda_1\sim\lambda_2} = -\frac{\ln\left(\tau_{\lambda_1}/\tau_{\lambda_2}\right)}{\ln\left(\lambda_1-\lambda_2\right)} \tag{2.1}$$

$$\tau_{\lambda_3} = \tau_{\lambda_1}\times\left(\frac{\lambda_3}{\lambda_1}\right)^{-\alpha_{\lambda_1\sim\lambda_2}} \tag{2.2}$$

式中 λ_1 和 λ_2 为波长，$\alpha_{\lambda_1\sim\lambda_2}$ 为波长 λ_1 和 λ_2 之间的波长指数，τ_{λ_1}、τ_{λ_2} 和 τ_{λ_3} 分别对应波长为 λ_1、λ_2 和 λ_3 时的 AOD 数据。

2.2.3. ERA5 气象再分析数据

本次研究所用的 ERA5 再分析数据可以从欧洲中期天气预报中心(ECMWF)获取(https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5)，下载 2016~2020 年中国陆地区域的气象数据，包括温度(TEMP)、风速(WS)、边界层高度(BLH)和相对湿度(RH)，气象数据的空间分辨率为 25 km。

现有研究表明，温度、风速、边界层高度和相对湿度等气象数据与 $PM_{2.5}$ 有高相关性，作为预测 $PM_{2.5}$ 的辅助变量可以有效提高预测精度[27]。地面气象监测点虽然可以提供高精度的气象数据，但是由于地面站点的空间分布有限，难以提供大面积的气象监测数据，且站点可供下载的气象参数较少加上时间跨度较大，相对而言，选择气象再分析数据能提供更多的参数选择和较高的时空分辨率和时间跨度[28]。

ERA5 气象再分析数据的风速(WS)变量可以由 10 m 风速 u 变量和 10 m 风速 v 分量计算得到。计算

公式如下：

$$\mathrm{WS}=\sqrt{u_{10}^2+v_{10}^2} \tag{2.3}$$

式中 WS 为最终使用的风速变量；u_{10} 为 10 m 风速 u 变量；v_{10} 为 10 m 风速 v 变量。

2.2.4. $PM_{2.5}$ 站点监测数据

$PM_{2.5}$ 的数据可以从中国环境监测总站的全国城市空气质量实时发布平台(https://air.cnemc.cn:18007/)进行获取，本次研究下载了 2016~2020 年的 $PM_{2.5}$ 地面监测数据。在研究时间范围内的 $PM_{2.5}$ 监测站点有 1500 个左右，相对于本文的研究区域来说，站点数据冗余过多，经过筛选以及剔除(数据缺失和异常)多余站点后剩余 138 个站点。$PM_{2.5}$ 监测站点分布见图 1。

3. 分析方法

3.1. MAIAC AOD 时空补值模型

AOD 数据是具有时空属性的时空数据，在进行 AOD 数据的补值时，需要同时兼顾时间和空间的影响。本文对 MAIAC AOD 数据分别使用 Prophet-LSTM 时序组合模型进行时间维度补值和 P-Bshade 模型进行空间维度补值，然后将两个维度的补值结果进行线性融合，最终得到顾及时空影响的 AOD 补值结果。总流程如图 2 所示。

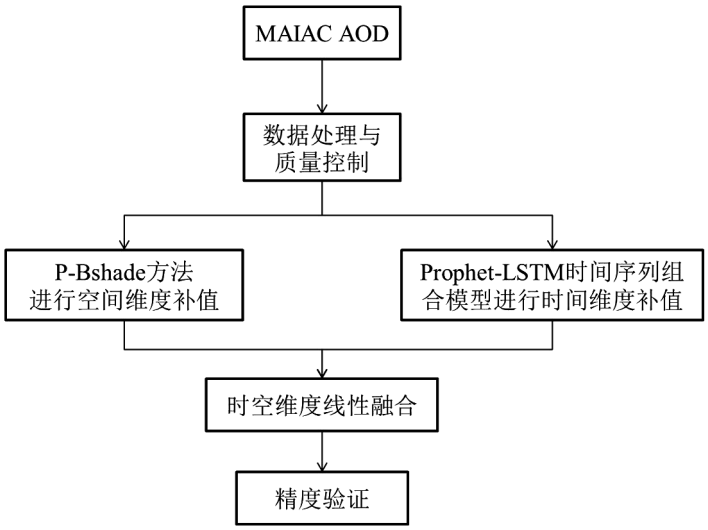

Figure 2. Flow chart of the spatiotemporal compensation model
图 2. 时空补值模型流程图

按下式将时间维和空间维补值结果进行线性融合。

$$Y_0=A\hat{y}_0+B\hat{t}_0$$

式中，A 表示空间维度权重；B 表示时间维度权重；Y_0 表示最终补值结果；$\hat{y}_0$ 表示空间维补值结果；$\hat{t}_0$ 表示时间维补值结果。本文所用的时间维和空间维的插值方法中，涉及到计算空间或时间协方差的影响，顾及到时空平衡性，本文的 A、B 取值均为 0.5。

3.1.1. Prophet-LSTM 时间序列补值组合模型

Prophet 是一种基于加法模型的时间序列数据预测，有具体的数学模型，能快速地进行时间预测，在

建模过程中考虑了趋势线、季节性、周期性，以及外生变量等因素的影响，预测效果好，相对于传统时序模型有很大优势。Prophet 对于异常值、丢失的数据具有健壮性，可以对杂乱的数据进行合理的预测[29]。LSTM 模型是一种特殊的循环神经网络(RNN)模型，LSTM 模型的“门”机制使得信息可以在时间序列中正确地流动，“门”机制可以限制流量信息，从而使得梯度在反向传播中不会消失或者爆炸。LSTM 模型对不同类型的数据有很强的适应性，可以捕获长时间序列数据的非线性关系[30] [31]，LSTM 让循环神经网络具备更强更好的记忆性能，可以有效地处理较长的时间序列数据。

Prophet-LSTM 时间序列补值组合模型由 Prophet 模型和 LSTM 模型两个部分组成，Prophet 模型负责为 LSTM 模型提供完整的 AOD 时间序列，LSTM 模型负责对 AOD 数据进行时间维度补值。具体模型构建如图 3。

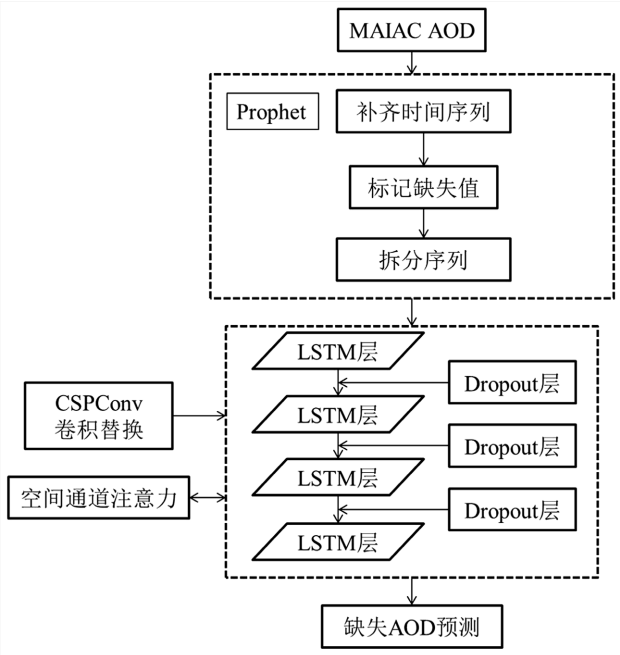

Figure 3. Prophet-LSTM combined timing model
图 3. Prophet-LSTM 组合时序模型

Prophet-LSTM 算法流程描述：

① 数据预处理：选择用 Prophet 补齐 MAIAC AOD 数据的时间序列，并用 mask 标记缺失值。对于数据中的缺失值，将其替换为 NaN (Not a Number)，以便在后续步骤中进行处理。同时，创建掩码来标记原始序列中的缺失值位置，将缺失值位置的掩码设置为 1，非缺失值位置的掩码设置为 0。

② 为了进行训练，将完整的序列按照长度为 64 的窗口进行循环拆分，并将每个窗口作为一个样本输入到模型中，保证模型可以对时间序列的不同部分进行学习和预测，同时也便于训练和批量处理。在拆分序列的过程中，只使用已有的数据作为输入，补齐的数据仅用于填充缺失部分，不参与损失函数的计算。这样可以确保模型在训练时只利用真实的数据进行学习，而不会受到补齐数据的影响。

③ 在 LSTM 模型中，为了提高模型表达能力，在 LSTM 网络中堆叠了 4 个 LSTM 层，每个 LSTM 层都具有相同的隐藏状态大小(hidden_size)，以确保信息的传递和记忆能力。并且将 LSTM 中的普通卷积换成了 CSPConv 并添加空间通道注意力(Spatial Channel Attention)，空间通道注意力是一种自适应地调整通道权重的方法，可以使模型更关注重要的特征通道。在 CSPConv Block 的最后一个残差块作为空间通

道注意力模块，以增强模型的特征表达能力。

④ 为了避免过拟合问题，在每个 LSTM 层之间添加了一个 Dropout 层，Dropout 层可以在训练过程中随机丢弃一部分神经元的输出，这样可以减轻网络对某些局部特征的依赖，减少过拟合风险。

⑤ 在计算损失函数时，使用带有掩码(mask)的损失函数，只计算非缺失部分的损失，对于每个时间切片，只计算掩码为 0 部分的损失。对于每个时间切片，根据掩码来选择是否计算该时间切片的损失。以预测序列与真实序列之间的均方差(Mean Squared Error, MSE)作为损失函数。使用带有掩码的损失函数时，只计算非缺失部分的预测值和真实值之间的方差。

⑥ 在训练阶段，通过反向传播和优化算法对模型进行训练，直到模型收敛并达到最佳性能。训练收敛后，使用构建的网络，为序列中缺失的部分预测其数值。对于连续缺失的情况，我们逐步迭代地预测序列的缺失部分，先预测第一个缺失值，然后将其用于下一个缺失值的预测，以此类推。

⑦ 模型在训练的时候同时训练正反方向的序列，避免序列头的缺失；对于中段缺失的部分，使用正反模型预测值的均值；对于头端的缺失，使用对侧方向的预测值。训练完成后，我们可以使用已经训练好的网络来预测序列中缺失部分的数值。对于连续缺失的情况，我们逐步迭代地预测序列的缺失部分，先预测第一个缺失值，然后将其用于下一个缺失值的预测，以此类推。

3.1.2. P-Bshade 模型空间维度补值

P-Bshade 方法是在空间维度进行的插值方法，计算原理如下：

$$\hat{y}_0 = \sum_{i=1}^{n} \omega_i y_i \tag{3.1}$$

式中，y_i 表示第 i 个空间周围采样数据的观测值；ω_i 表示第 i 个空间周围采样数据对缺失数据的空间贡献权重；ω_i 可以用下式计算求得：

$$\begin{bmatrix} C(y_1,y_2) & \cdots & C(y_1,y_n) & b_1 \\ \vdots & \ddots & \vdots & \vdots \\ C(y_n,y_1) & \cdots & C(y_n,y_n) & b_n \\ b_1 & \cdots & b_n & 0 \end{bmatrix} \begin{bmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_n \\ \mu \end{bmatrix} = \begin{bmatrix} C(y_1,y_0) \\ C(y_2,y_0) \\ \vdots \\ C(y_n,y_0) \\ 1 \end{bmatrix} \tag{3.2}$$

式中，方程中间的矩阵为待求矩阵；μ 为拉格朗日系数。方程左边的矩阵中 $C(y_i,y_{i'})$ 为第 i 个空间附近采样点的时间序列与第 i' 个空间附近采样点的时间序列的协方差，b_i 为第 i 个空间附近采样点的时间序列与缺失数据点的时间序列的期望比。方程右边的矩阵中 $C(y_i,y_0)$ 为第 i 个空间周围采样点的时间序列与缺失数据点的时间序列的协方差，并满足 $1 \le i \le n$，$1 \le i' \le n$。

首先选取每天的 MAIAC AOD 缺失数据附近 n 个相关性最大的空间采样数据进行插值计算，采用相关系数 R 来说明相关性的强弱，对于一个 AOD 缺失序列，计算其附近空间点的 AOD 数据序列和缺失 AOD 数据点的数据序列的相关系数 R，R 越大则表示相关性越强，反之则越弱。之后在缺失点附近找到非空 AOD 序列且相关系数 R 最大的十组序列，构建拉格朗日方程组之后求解权重，最后得出缺失值。

3.2. Catboost 模型构建

Catboost 模型[32]是一种梯度提升术算法，是基于一种对称二叉树为基学习器的 GBDT 框架，专门用于处理类别型特征的机器学习问题，还解决了模型训练和预测过程中的梯度偏移和预测偏移问题，从而达到减少过拟合风险的目的，进而提高模型算法的准确性和泛化能力。该模型对 $PM_{2.5}$ 浓度显示出较好的估算能力[33] [34]。在进行类别型特征处理的时候，常用的方法是用整个数据集的均值代替类别特征的值，

并进行平滑处理，这样的方法对于单个样本的估计量是有偏的，Catboost 算法通过使用除去该样本的数据子集而不是整个数据集来进行估计，公式如下：

$$m_{i,k}=\frac{\sum_{m_{i,k}\in D_k}\left[m_{i,k}=m_{i,j}\right]n_j+ap}{\sum_{m_{i,k}\in D_k}\left[m_{i,k}=m_{i,j}\right]+a} \tag{3.3}$$

式中 $m_{i,k}$ 为第 k 个样本的第 i 个样本特征，n_j 为第 j 个样本的标签特征值，$m_{i,j}$ 为第 k 个样本前的第 j 个样本的第 i 个类别特征，D_k 为随机序列中在第 k 个样本前的数据集，a 通常为大于 0 的参数，p 为先验项。

本文使用 Catboost 模型来建立 MAIAC AOD-$PM_{2.5}$ 之间的关系，并加入温度(TEMP)、风速(WS)、边界层高度(BLH)和相对湿度(RH)作为相关预测因子，与 AOD 一起建立模型对 $PM_{2.5}$ 浓度进行估算，模型建立流程如图 4：

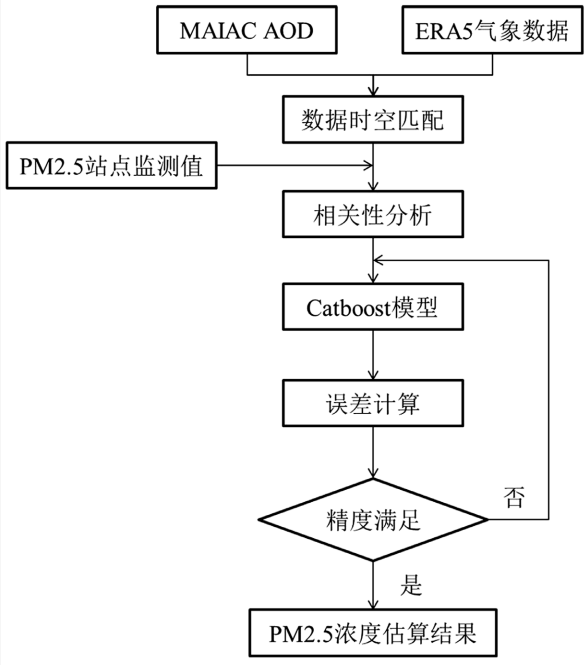

Figure 4. Flow chart of $PM_{2.5}$ estimation by Catboost
图 4. Catboost 估算 $PM_{2.5}$ 浓度流程图

3.3. 评价指标

AOD 补值和 $PM_{2.5}$ 的浓度估算都采用平均绝对误差(MAE)、均方根误差(RMSE)和皮尔逊相关系数(R)作为评价指标来进行精度验证，AOD 补值精度验证使用地面站点 AERONET AOD 数据与补值结果进行验证，$PM_{2.5}$ 的估算精度验证使用 $PM_{2.5}$ 站点与估算结果进行验证。各评价指标计算见下式：

$$\begin{cases}\mathrm{MAE}=\dfrac{\sum_{i=1}^{n}\left|\hat{y}_i-y_i\right|}{n}\\ \mathrm{RMSE}=\sqrt{\dfrac{\sum_{i=1}^{n}\left(\hat{y}_i-y_i\right)^2}{n}}\\ \mathrm{R}=\dfrac{\mathrm{COV}\left(\hat{y}_i,y_i\right)}{\sigma_{\hat{y}_i}\sigma_{y_i}}\end{cases} \tag{3.4}$$

式中 $\hat{y}_i$ 和 y_i 分别是模型计算和站点监测的 AOD 值($PM_{2.5}$ 值)，n 是样本的个数， $\mathrm{COV}(\hat{y}_i, y_i)$ 为模型计算值和站点监测 AOD 值($PM_{2.5}$ 浓度)的协方差，$\sigma_{\hat{y}_i}$ 为模型计算结果的标准差，σ_{y_i} 为站点检测值的标准差。

4. 结果分析

4.1. 精度指标评价

据气象学标准定义的实际季节来定义季节：春季为 3、4 和 5 月，夏季为 6、7 和 8 月，秋季为 9、10 和 11 月，冬季为 12 月、次年的 1 月和 2 月。本文对 2020 年 3 月~2021 年 2 月的 MAIAC AOD 进行时空补值并对 $PM_{2.5}$ 浓度进行估算。图 5 为 AOD 时空补值数据与地面站点数据的线性拟合结果对比图，图 6 为 $PM_{2.5}$ 浓度估算结果与地面站点数据的线性拟合结果图。

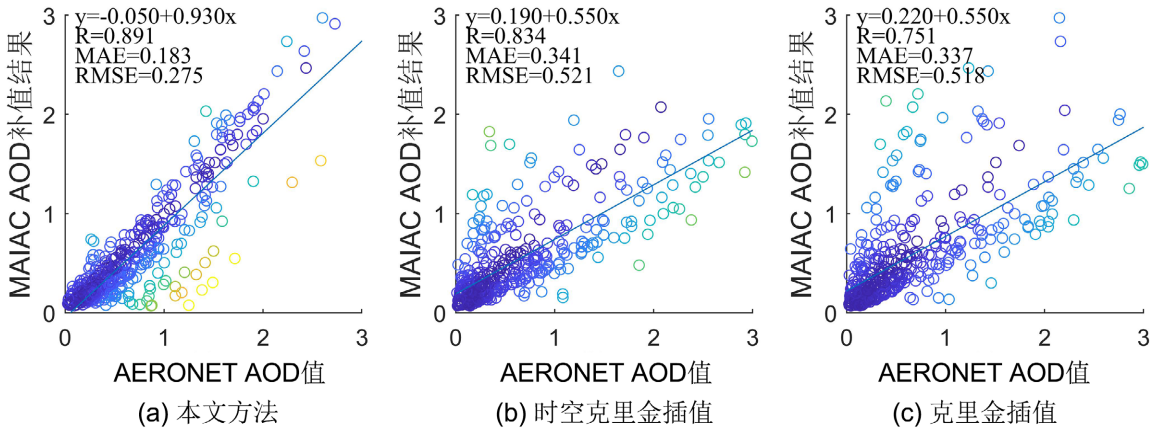

Figure 5. Fitting diagram of AOD top-up results
图 5. AOD 补值结果拟合图

本文所用的 Prophet-LSTM + P-Bshade 时空补值模型相对于常用的经典克里金模型和时空克里金模型有较高的拟合精度，R、MASE 和 MAE 分别为 0.891、0.275 和 0.183，经典克里金补值模型的 R、MASE 和 MAE 分别为 0.751、0.518 和 0.337 空克里金补值模型的 R、MASE 和 MAE 分别为 0.834、0.521 和 0.341。从各模型补值的结果和站点监测值的收敛来看，本文提出的补值方法收敛性最好。经典克里金模型是基于数据空间相关性进行的补值方法，未考虑数据的时间相关性，因而模型补值结果较差，时空克里金模型是经典克里金模型在空间维度上的延伸，在补值时需要多个时间点的空间截面数据进行时空连续性插值，在进行大范围的补值研究时，计算量十分庞大，且 AOD 的补值中存在低估现象。本文提出的 Prophet-LSTM + P-Bshade 时空补值模型基于 AOD 数据的时空相关性对其进行了补值，且相对于对比模型来说有最好的补值效果。

将 ERA5 气象数据和经过 Prophet-LSTM + P-Bshade 补值后的 AOD 数据作为估算变量输入到 Catboost 模型中，即可得到 $PM_{2.5}$ 浓度的空间分布数据，图 6 为模型估算结果与 $PM_{2.5}$ 站点拟合结果图。

本文选择 LightGBM 等八种常用机器学习、随机森林方法与 Catboost 模型进行对比，通过图 6 可知，Catboost 模型的拟合精度最高，R、MASE 和 MAE 分别为 0.93、15.89 $\mu g \cdot m^{-3}$ 和 10.54 $\mu g \cdot m^{-3}$，这些指数均优于其他八个对比模型。

其中较为传统的 Bagging 模型和 KNN 模型相对来说拟合精度较差，LightGBM、XGBoost 和 Catboost 模型是 GBDT 的三大主流模型，都是在 GBDT 算法框架下进行了改进。这三种模型在 $PM_{2.5}$ 浓度估算中都显示有较高的拟合精度。其中 Catboost 模型主要使用了 Ordered Target Statistics 方法将类别特征转化为

数值特征、基于贪心策略的特征组合方法、使用 Orcdered boostng 避免梯度偏移问题和使用对称二叉树作为基模型，其拟合效果最好。

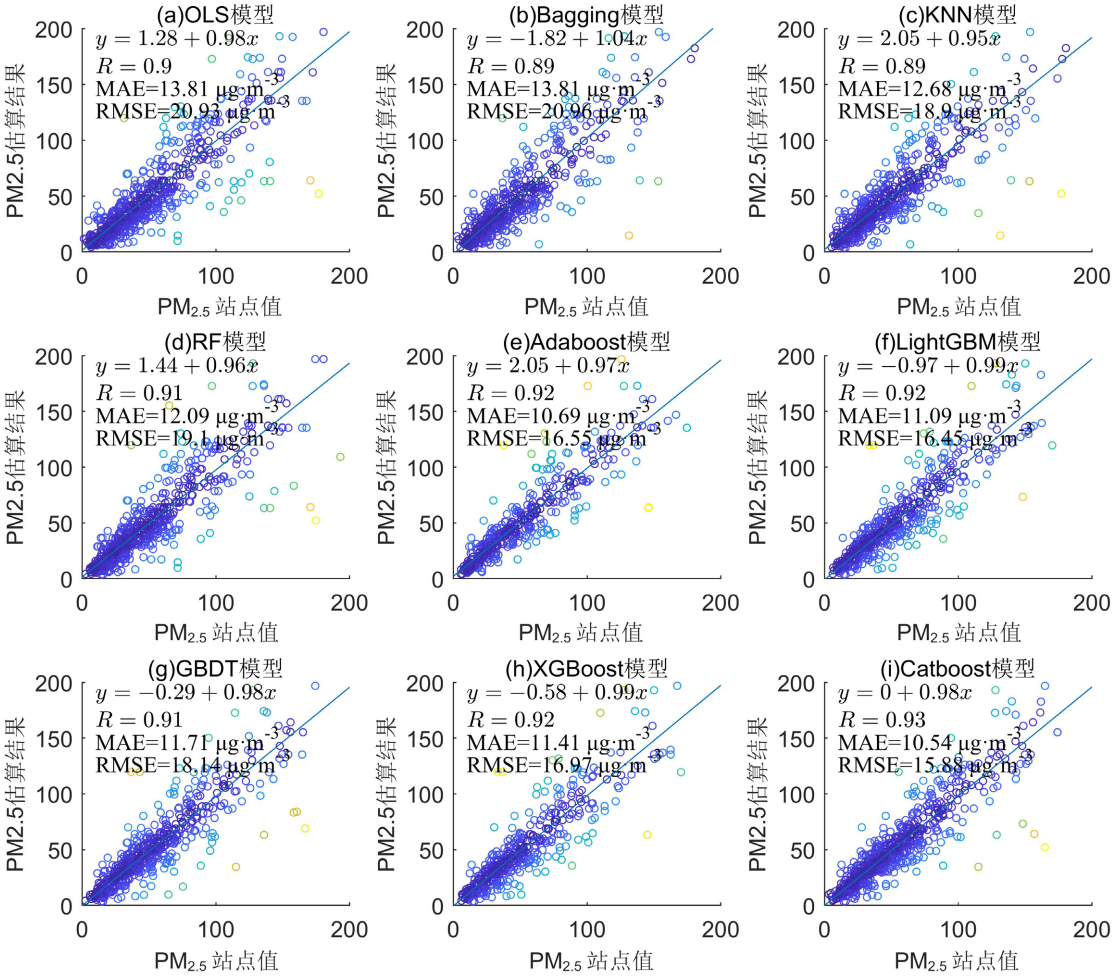

Figure 6. $PM_{2.5}$ concentration fitting results
图 6. $PM_{2.5}$ 浓度拟合结果图

4.2. $PM_{2.5}$ 时空分布分析

将补值后的 AOD 数据与 ERA5 气象数据输入 Catboost 模型中可以得到 2020 年 3 月至 2021 年 2 月的每日 $PM_{2.5}$ 浓度估算空间分布图，再按 4.1 节方法划分季节，按季节取均值可以得到 $PM_{2.5}$ 季均空间分布图，如图 7。

从季节分布上来看，2020 年四个季节 $PM_{2.5}$ 浓度分布特征较为明显，整体呈现冬季(51.37 μg·m^{-3}) > 春季(37.40 μg·m^{-3}) > 秋季(24.86 μg·m^{-3}) > 夏季(21.95 μg·m^{-3})的季节分布特点。从空间分布上来看，我国陆地区域 $PM_{2.5}$ 浓度呈现东高西低的特点。东部地区经济发达，我国三大经济圈(环渤海经济圈、长江三角洲经济圈和珠江三角洲经济圈)主要城市基本都在中国东部，高速城市化、工业化带来经济发展的同时，城市生态环境空间被大量蚕食、大量的流动人口朝着经济发达的地方聚集，污染排放的强度和密度剧增，使得这些经济发达的地方成为大气环境污染的重灾区[35]。西部地区经济相对落后，$PM_{2.5}$ 浓度整体浓度偏低，但是塔里木盆地区域 $PM_{2.5}$ 浓度局部较高，中国最大的沙漠——塔克拉玛干沙漠位于塔里木盆地附

近，长期的沙尘暴以及风蚀现象，会将高浓度的气溶胶颗粒物带到不同的地方从而影响当地气候，包括整个西北地区和华北地区等，不同程度上都会受到沙尘的影响，沙漠活动等自然因素影响是导致塔里木盆地区域的 $PM_{2.5}$ 浓度较高的主要原因[36]。

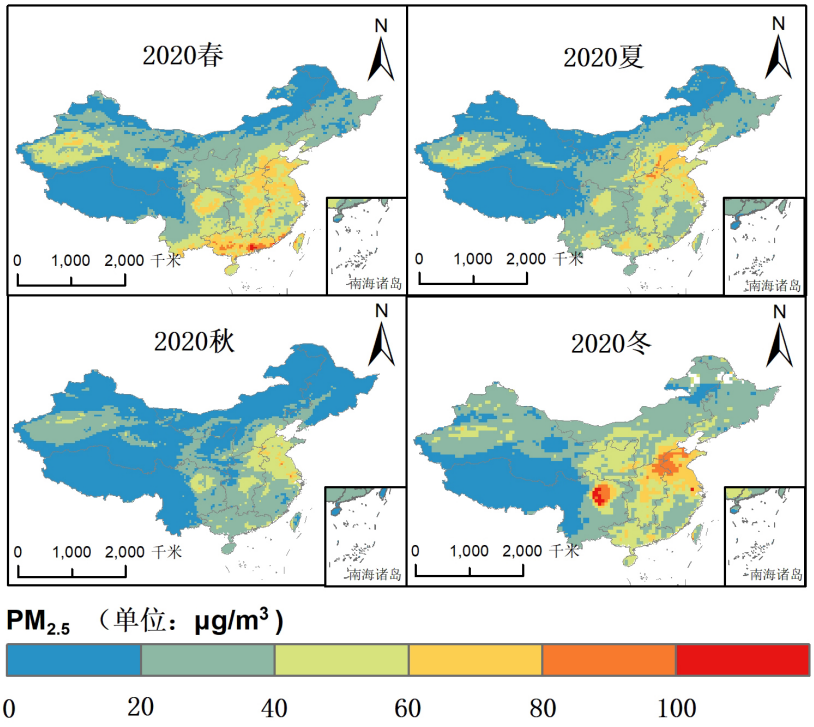

Figure 7. Spatial distribution of seasonal average $PM_{2.5}$ concentrations in 2020
图 7. 2020 年 $PM_{2.5}$ 浓度季均空间分布图

5. 结论

本文使用建立的 Prophet-LSTM + P-Bshade 时空补值模型对 MAIAC AOD 数据进行时空补值，并将其与 ERA5 气象再分析数据通过 Catboost 模型进行 $PM_{2.5}$ 浓度估算，为我国大气环境污染治理提供了数据支撑。对 2020 年 3 月~2021 年 2 月的全国陆地区域进行 $PM_{2.5}$ 浓度估算，分析得到以下结论。

(1) Prophet-LSTM + P-Bshade 时空补值模型精度明显优于传统补值方法，R、MASE 和 MAE 分别为 0.891、0.275 和 0.183。

(2) Catboost 模型在 $PM_{2.5}$ 浓度估算中相较于其他八个常用模型显示更高的估算精度，R、MASE 和 MAE 分别为 0.93、15.89 $\mu g \cdot m^{-3}$ 和 10.54 $\mu g \cdot m^{-3}$。

(3) 中国陆地区域 2020 年的 $PM_{2.5}$ 浓度在季节尺度分布上明显，整体呈现冬季 > 春季 > 秋季 > 夏季的季节分布特点。在空间分布上，$PM_{2.5}$ 浓度整体呈现东部地区较高，塔里木盆地区域局部较高的特点。

(4) 提出的 Prophet-LSTM + P-Bshade 时空补值模型在将 AOD 数据进行时间维和空间维补值结果线性融合时，本文顾及时空平稳性将两个维度的权重都取值为 0.5，未来可以尝试其他的方法来设定时空维度权重，以此达到更好的补值效果。

参考文献

[1] 徐栋夫, 曹萍萍, 王源程. 成都一次重污染过程的气溶胶光学特性垂直分布[J]. 气象, 2020, 46(7): 948-958.

[2] 任秀龙, 胡伟, 吴春苗, 等. 华北南部重污染城市周边区域二次气溶胶的化学特征及来源解析[J]. 环境科学, 2022, 43(3): 1159-1169.

[3] 王跃思, 张军科, 王莉莉, 等. 京津冀区域大气霾污染研究意义、现状及展望[J]. 地球科学进展, 2014, 29(3): 388-396.

[4] Ramanathan, V., Crutzen, P.J., Kiehl, J.T. and Rosenfeld, D. (2001) Aerosols, Climate, and the Hydrological Cycle. *Science*, **294**, 2119-2124. https://doi.org/10.1126/science.1064034

[5] Peng, J., Han, H., Yi, Y., Huang, H. and Xie, L. (2022) Machine Learning and Deep Learning Modeling and Simulation for Predicting $PM_{2.5}$ Concentrations. *Chemosphere*, **308**, Article ID: 136353. https://doi.org/10.1016/j.chemosphere.2022.136353

[6] van Donkelaar, A., Martin, R.V., Levy, R.C., da Silva, A.M., Krzyzanowski, M., Chubarova, N.E., *et al*. (2011) Satellite-Based Estimates of Ground-Level Fine Particulate Matter during Extreme Events: A Case Study of the Moscow Fires in 2010. *Atmospheric Environment*, **45**, 6225-6232. https://doi.org/10.1016/j.atmosenv.2011.07.068

[7] Lin, J., Nielsen, C.P., Zhao, Y., Lei, Y., Liu, Y. and McElroy, M.B. (2010) Recent Changes in Particulate Air Pollution over China Observed from Space and the Ground: Effectiveness of Emission Control. *Environmental Science & Technology*, **44**, 7771-7776. https://doi.org/10.1021/es101094t

[8] Engel-Cox, J.A., Holloman, C.H., Coutant, B.W. and Hoff, R.M. (2004) Qualitative and Quantitative Evaluation of MODIS Satellite Sensor Data for Regional and Urban Scale Air Quality. *Atmospheric Environment*, **38**, 2495-2509. https://doi.org/10.1016/j.atmosenv.2004.01.039

[9] van Donkelaar, A., Martin, R.V., Brauer, M., Hsu, N.C., Kahn, R.A., Levy, R.C., *et al*. (2016) Global Estimates of Fine Particulate Matter Using a Combined Geophysical-Statistical Method with Information from Satellites, Models, and Monitors. *Environmental Science & Technology*, **50**, 3762-3772. https://doi.org/10.1021/acs.est.5b05833

[10] 吴迪, 杜宁, 王莉, 等. 基于 GTWR-XGBoost 模型的四川省 PM_(2.5)小时浓度估算[J]. 环境科学, 2023, 44(7): 3738-3748.

[11] 吴宇宏, 杜宁, 王莉等. 基于 iLME+Geoi-RF 模型的四川省 PM_(2.5)浓度估算[J]. 环境科学, 2021, 42(12): 5602-5615.

[12] 吴迪. 基于 GTWR-XGBoost 模型的 PM_(2.5)浓度估算及与城市化关系研究[D]: [硕士学位论文]. 贵阳: 贵州大学, 2023.

[13] Chatterjee, A., Michalak, A.M., Kahn, R.A., Paradise, S.R., Braverman, A.J. and Miller, C.E. (2010) A Geostatistical Data Fusion Technique for Merging Remote Sensing and Ground-Based Observations of Aerosol Optical Thickness. *Journal of Geophysical Research*: *Atmospheres*, **115**, D20207. https://doi.org/10.1029/2009jd013765

[14] 许悦蕾, 刘延安, 施润和, 等. 气象要素对气溶胶光学厚度估算 PM_(2.5)的影响[J]. 环境科学学报, 2018, 38(10): 3868-3876.

[15] Kim, S., Koo, J., Lee, H., Mok, J., Choi, M., Go, S., *et al*. (2021) Comparison of $PM_{2.5}$ in Seoul, Korea Estimated from the Various Ground-Based and Satellite AOD. *Applied Sciences*, **11**, Article No. 10755. https://doi.org/10.3390/app112210755

[16] Ma, Z., Hu, X., Sayer, A.M., Levy, R., Zhang, Q., Xue, Y., *et al*. (2016) Satellite-Based Spatiotemporal Trends in $PM_{2.5}$ Concentrations: China, 2004-2013. *Environmental Health Perspectives*, **124**, 184-192. https://doi.org/10.1289/ehp.1409481

[17] Liu, Y., Paciorek, C.J. and Koutrakis, P. (2009) Estimating Regional Spatial and Temporal Variability of $PM_{2.5}$ Concentrations Using Satellite Data, Meteorology, and Land Use Information. *Environmental Health Perspectives*, **117**, 886-892. https://doi.org/10.1289/ehp.0800123

[18] He, Q. and Huang, B. (2018) Satellite-Based Mapping of Daily High-Resolution Ground $PM_{2.5}$ in China via Space-Time Regression Modeling. *Remote Sensing of Environment*, **206**, 72-83. https://doi.org/10.1016/j.rse.2017.12.018

[19] Liu, N., Zou, B., Li, S., Zhang, H. and Qin, K. (2021) Prediction of $PM_{2.5}$ Concentrations at Unsampled Points Using Multiscale Geographically and Temporally Weighted Regression. *Environmental Pollution*, **284**, Article ID: 117116. https://doi.org/10.1016/j.envpol.2021.117116

[20] Liu, Z., Xiao, Q. and Li, R. (2023) Full Coverage Hourly $PM_{2.5}$ Concentrations' Estimation Using Himawari-8 and MERRA-2 AODs in China. *International Journal of Environmental Research and Public Health*, **20**, Article No. 1490. https://doi.org/10.3390/ijerph20021490

[21] Li, L., Zhang, J., Meng, X., Fang, Y., Ge, Y., Wang, J., *et al*. (2018) Estimation of $PM_{2.5}$ Concentrations at a High Spatiotemporal Resolution Using Constrained Mixed-Effect Bagging Models with MAIAC Aerosol Optical Depth. *Remote Sensing of Environment*, **217**, 573-586. https://doi.org/10.1016/j.rse.2018.09.001

[22] 金团囡, 杨兴川, 晏星, 等. 京津冀及周边 MAIAC AOD 和 PM_(2.5)质量浓度特征及相关性分析[J]. 环境科学, 2021, 42(6): 2604-2615.

[23] Lyapustin, A., Wang, Y., Korkin, S. and Huang, D. (2018) MODIS Collection 6 MAIAC Algorithm. *Atmospheric Measurement Techniques*, **11**, 5741-5765. https://doi.org/10.5194/amt-11-5741-2018

[24] Lyapustin, A. and Wang, Y. (2022) MODIS/Terra+Aqua Land Aerosol Optical Depth Daily L2G Global 1 km SIN Grid V061. MODIS/Terra+Aqua Land Aerosol Optical Depth Daily L2G Global 1 km SIN Grid LAADS DAAC (nasa.gov).

[25] Ichoku, C., Chu, D.A., Mattoo, S., Kaufman, Y.J., Remer, L.A., Tanré, D., *et al*. (2002) A Spatio-Temporal Approach for Global Validation and Analysis of MODIS Aerosol Products. *Geophysical Research Letters*, **29**, MOD1. https://doi.org/10.1029/2001gl013206

[26] Kahn, R.A., Gaitley, B.J., Martonchik, J.V., Diner, D.J., Crean, K.A. and Holben, B. (2005) Multiangle Imaging Spectroradiometer (MISR) Global Aerosol Optical Depth Validation Based on 2 Years of Coincident Aerosol Robotic Network (AERONET) Observations. *Journal of Geophysical Research*: *Atmospheres*, **110**, D10S04. https://doi.org/10.1029/2004jd004706

[27] Liu, J., Weng, F. and Li, Z. (2019) Satellite-Based $PM_{2.5}$ Estimation Directly from Reflectance at the Top of the Atmosphere Using a Machine Learning Algorithm. *Atmospheric Environment*, **208**, 113-122. https://doi.org/10.1016/j.atmosenv.2019.04.002

[28] 岳书平, 闫业超, 张树文, 等. 基于 ERA5-LAND 的中国东北地区近地表土壤冻融状态时空变化特征[J]. 地理学报, 2021, 76(11): 2765-2779.

[29] Taylor, S.J. and Letham, B. (2018) Forecasting at Scale. *The American Statistician*, **72**, 37-45. https://doi.org/10.1080/00031305.2017.1380080

[30] Wang, Z., Zhou, Y., Zhao, R., Wang, N., Biswas, A. and Shi, Z. (2021) High-Resolution Prediction of the Spatial Distribution of $PM_{2.5}$ Concentrations in China Using a Long Short-Term Memory Model. *Journal of Cleaner Production*, **297**, Article ID: 126493. https://doi.org/10.1016/j.jclepro.2021.126493

[31] Liang, L., Daniels, J., Biancardi, M. and Zhou, Y. (2023) Reconstructing Aerosol Optical Depth Using Spatiotemporal Long Short-Term Memory Convolutional Autoencoder. *Scientific Data*, **10**, Article No. 842. https://doi.org/10.1038/s41597-023-02696-w

[32] Prokhorenkova, L., Gusev, G., Vorobev, A., *et al*. (2019) CatBoost: Unbiased Boosting with Categorical Features.

[33] Guo, Z., Wang, X. and Ge, L. (2023) Classification Prediction Model of Indoor $PM_{2.5}$ Concentration Using Catboost Algorithm. *Frontiers in Built Environment*, **9**, Article No. 6174. https://doi.org/10.3389/fbuil.2023.1207193

[34] Shahriar, S.A., Kayes, I., Hasan, K., Hasan, M., Islam, R., Awang, N.R., *et al*. (2021) Potential of ARIMA-ANN, ARIMA-SVM, DT and Catboost for Atmospheric $PM_{2.5}$ Forecasting in Bangladesh. *Atmosphere*, **12**, Article No. 100. https://doi.org/10.3390/atmos12010100

[35] 李欣, 曹建华, 孙星. 空间视角下城市化对雾霾污染的影响分析——以长三角区域为例[J]. 环境经济研究, 2017, 2(2): 81-92.

[36] Filonchyk, M., Yan, H., Zhang, Z., Yang, S., Li, W. and Li, Y. (2019) Combined Use of Satellite and Surface Observations to Study Aerosol Optical Depth in Different Regions of China. *Scientific Reports*, **9**, Article ID: 1207193. https://doi.org/10.1038/s41598-019-42466-6