Advances in Environmental Protection 环境保护前沿, 2026, 16(11), 1692-1704
Published Online November 2026 in Hans. https://www.hanspub.org/journal/aep
https://doi.org/10.12677/aep.2026.1611170

Hans汉斯

利用土地利用回归模型与集成机器学习方法评估中国$PM_{2.5}$污染时空分布

马健容，廖　鸿，杨昊林，崔　漾，马雪怡，张　平

西安工程大学环境与化学工程学院，陕西 西安

收稿日期：2026年9月24日；录用日期：2026年10月28日；发布日期：2026年11月17日

摘　要

随着社会经济快速发展与城市化、工业化进程加快推进，$PM_{2.5}$污染日益受到广泛关注，准确模拟$PM_{2.5}$污染空间格局，是制定大气污染防控措施和开展健康效应评估的重要基础。本研究基于随机森林(RF)、类别特征提升(CatBoost)和极端梯度提升(XGBoost)三种机器学习方法，构建土地利用回归－时空加权(LUR-SW)机器学习集成模型，融合2018~2020年气溶胶光学厚度(AOD)、ERA5气象数据、土地利用、污染源及植被指数等多源遥感数据，模拟中国$PM_{2.5}$浓度的时空分布，为健康暴露风险精确评估与城市可持续发展提供科学支撑。结果表明，LUR建模变量中AOD与$PM_{2.5}$浓度呈正相关，而气温和降水均与$PM_{2.5}$浓度呈负相关。CatBoost、XGBoost和RF模型的R^2分别为0.86、0.89和0.84，表明三种模型均具有良好的预测性能，其中XGBoost模型的预测性能最优。$PM_{2.5}$浓度呈现明显的季节性变化特征，冬季最高、夏季最低，且在空间上呈强正自相关性，空间分布以高－高聚集以及低－低聚集特征为主。研究结果可为$PM_{2.5}$污染精细化模拟、预警区识别与暴露风险量化提供可靠依据。

关键词

$PM_{2.5}$，土地利用回归模型，集成机器学习方法，空间自相关

Assessing the Spatiotemporal Distribution of $PM_{2.5}$ Pollution in China Using Land Use Regression Model and Ensemble Machine Learning Methods

Jianrong Ma, Hong Liao, Haolin Yang, Ying Cui, Xueyi Ma, Ping Zhang

School of Environmental and Chemical Engineering, Xi'an Polytechnic University, Xi'an Shaanxi

Received: September 24, 2026; accepted: October 28, 2026; published: November 17, 2026

文章引用: 马健容, 廖鸿, 杨昊林, 崔漾, 马雪怡, 张平. 利用土地利用回归模型与集成机器学习方法评估中国 $PM_{2.5}$ 污染时空分布[J]. 环境保护前沿, 2026, 16(11): 1692-1704. DOI: 10.12677/aep.2026.1611170

Abstract

With rapid socio-economic development and the accelerating pace of urbanization and industrialization, $PM_{2.5}$ pollution has increasingly come under the spotlight. Accurately modelling the spatial distribution of $PM_{2.5}$ pollution is a crucial foundation for formulating air pollution control measures and conducting health impact assessments. This study utilizes three machine learning methods—Random Forest (RF), Category Feature Boosting (CatBoost) and Extreme Gradient Boosting (XGBoost)—to construct a Land Use Regression-Spartoi-Temporal Weighting (LUR-SW) machine learning ensemble model. By integrating multi-source remote sensing data—including aerosol optical depth (AOD) from 2018 to 2020, ERA5 meteorological data, land use, pollution sources and vegetation indices, to simulate the spatiotemporal distribution of $PM_{2.5}$ concentrations in China, thereby providing scientific support for the precise assessment of health exposure risks and urban sustainable development. The results indicate that, among the LUR modelling variables, AOD is positively correlated with $PM_{2.5}$ concentrations, whilst both air temperature and precipitation are negatively correlated with $PM_{2.5}$ concentrations. The R^2 values of the CatBoost, XGBoost, and RF models were 0.86, 0.89, and 0.84, respectively, indicating that all three models exhibited good predictive performance, with XGBoost achieving the best performance. $PM_{2.5}$ concentrations exhibit distinct seasonal variations, peaking in winter and reaching their lowest levels in summer. Spatially, they display strong positive autocorrelation, with spatial distribution characterized primarily by "high-high" and "low-low" clustering. The findings of this study provide a reliable basis for the detailed modelling of $PM_{2.5}$ pollution, the identification of early-warning zones, and the quantification of exposure risks.

Keywords

$PM_{2.5}$, Land Use Regression Model, Ensemble Machine Learning Methods, Spatial Autocorrelation

Copyright © 2026 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

随着我国城镇化与工业化持续推进，高能耗产业生产活动、工业废气排放与机动车尾气持续叠加，使得区域性大气污染问题长期存在。颗粒物($PM_{2.5}$)是衡量区域大气质量的关键指标，因其粒径微小、滞留时间长、扩散范围广，可对居民健康产生显著影响。已有流行病学研究证实，$PM_{2.5}$ 与呼吸系统疾病、心血管疾病、肺癌等不良健康效应密切相关。同时，颗粒物还与各类气象因素之间存在复杂的相互作用，进而影响气候和环境[1]。因此，开展 $PM_{2.5}$ 相关研究具有重要的现实意义。当前，国内外对 $PM_{2.5}$ 监测的手段主要依托地面监测站点实现，但传统站点监测体系仍存在一定局限：监测站点空间分布不均衡，城乡站点布设密度差异显著，容易造成对污染空间分布特征刻画的偏差，难以准确捕捉 $PM_{2.5}$ 的精细空间梯度，从而制约对大气污染动态演化过程的精准解析[2] [3]。因此，构建高精度、时空连续的 $PM_{2.5}$ 浓度估算方法，可为科学制定大气污染防控措施和完善人群健康风险防控策略提供重要参考与决策支持。

在大气污染物空间模拟研究中，土地利用回归(LUR)模型是预测污染物空间分布的有效方法之一。该模型能够利用有限的监测站点数据，并结合土地利用信息及其他特征变量，评估监测站点稀疏地区的污染物浓度[4] [5]。自 1997 年 Briggs 等[6]在 SAVIAH 小区域空气质量与健康研究项目中利用多种变量构建回归模型并实现区域空气污染空间制图以来，LUR 模型得到了广泛应用。刘炳杰[7]等人以全国为研究

范围，验证了 LUR 模型在国家尺度污染物模拟中的适用性，剖析了 2015 年全国 $PM_{2.5}$ 时空演化规律及其与各类地理因子的关联性。杨明亮[8]等将土地利用回归(LUR)模型应用于安徽省污染较为严重的皖北地区，综合土地利用、道路、污染源、气象、高程及人口等共 105 个变量，建立了该地区四季和年均 LUR 模型，并通过留一交叉验证评估了模型精度。

然而，目前大多机器学习模型尚未充分考虑 $PM_{2.5}$ 的时空异质性，可能导致 $PM_{2.5}$ 浓度估算产生较大预测偏差[9] [10]。针对 $PM_{2.5}$ 时空非平稳特征，学界发展了多种处理时空异质性的模型：ST-kriging (时空克里金)主要依靠污染物浓度的时空自相关性进行插值，可实现 $PM_{2.5}$ 浓度的空间连续估计与可视化制图，在监测点位密集的局部城区具有较好的平滑估计优势。然而，在我国跨气候带、地貌类型多样的区域尺度污染模拟中，其性能可能受到监测站点分布、时空非平稳性及区域异质性的影响；GTWRF (时空地理加权随机森林)融合时空加权机制与随机森林的非线性拟合能力，可在一定程度上弥补传统全局模型难以刻画局部异质性的不足，并捕捉驱动因子作用的局部时空差异。与传统统计模型相比，该模型在特定数据和研究区域中可能取得更高的预测精度，但在全国尺度、跨地貌区域应用时，其泛化能力仍可能受到局部样本代表性、时空权重设置及区域异质性等因素的影响，进而出现局部过拟合或区域性估计偏差。为此，Li [11]等人提出地理时间加权树模型(GTW-Tree)，该模型能够输出各空间位置对应的变量重要性，进而分析各类预测因子对 $PM_{2.5}$ 的时空变化影响。现有大尺度模拟研究还存在对典型区域污染成因仅停留在现象描述、缺少模型不确定性分析等不足。这些研究为 $PM_{2.5}$ 浓度模拟提供了新的技术路径，同时表明基于 LUR 模型开展 $PM_{2.5}$ 空间分布模拟仍具有较大的拓展空间。

本研究基于 $PM_{2.5}$ 监测数据、气象因素、土地利用数据、社会经济因素和其他辅助变量，整合多源卫星遥感数据，建立了一种高效且准确的 LUR-时空加权机器学习集成模型。

2. 研究区与研究数据

2.1. 研究区概况

中国地理跨度大，国土面积广阔，地形地貌复杂多样，整体地势自西向东呈三级阶梯状分布，气候格局复杂，区域自然条件具有显著的空间异质性。$PM_{2.5}$ 浓度在空间维度上总体呈现北高南低、东高西低的分布特征，黄淮海平原、汾渭平原与四川盆地等区域受产业活动密集及地形阻滞扩散等因素影响，颗粒物污染问题较为突出。

2.2. 数据来源与处理

为系统探究中国 $PM_{2.5}$ 的时空分布特征及影响机制，本项目收集多源高精度环境与社会地理数据开展综合分析。在大气污染观测数据方面，采用 2018~2020 年全国 1267 个环境监测地面站点的 $PM_{2.5}$ 实测浓度数据(港澳台数据暂缺)，其中各年度有效站点数量分别为 2018 年 1500 个、2019 年 1508 个和 2020 年 1528 个。土地利用数据采用武汉大学杨杰和黄昕教授团队发布的中国土地覆盖数据集(China Land Cover Dataset, CLCD)，以准确表征不同土地利用类型的空间分布特征。气象数据来源于 ECMWF 欧洲气象资料网，选取气温、表压、总降水量、相对湿度、边界层高度和风速作为核心气象输入变量。研究采用的 250 m 分辨率月度归一化植被指数(NDVI)数据覆盖 2000~2022 年全国范围，该数据集基于 MOD13Q1 产品并结合土地利用信息，经过 SG 滤波、噪声重构、最大值合成与全国拼接等预处理获得。气溶胶光学厚度(AOD)数据选用 LGHAP 数据集，来源于华东师范大学团队公开成果，可为大范围 $PM_{2.5}$ 遥感反演提供可靠数据支撑。道路密度数据由 OpenStreetMap 矢量道路图层计算得到，通过统计不同尺度缓冲区内的道路总长度获取；地形地貌数据使用 90 m 分辨率的航天飞机雷达地形任务(SRTM)数字高程模型(DEM)

和坡度数据(SLOPE)。在社会地理数据方面，POI 兴趣点数据来源于高德开发平台，植被覆盖、GDP 和人口密度数据均来自中国科学院资源环境科学数据中心。

在完成多源数据预处理后，本研究融合土地利用回归 LUR 模型、时空加权信息和机器学习集成方法，构建混合预测框架，利用 R^2、RMSE 和 MAE 指标开展十折交叉验证，筛选综合预测性能最优的模型，实现中国地区逐日 $PM_{2.5}$ 浓度高精度模拟，并进一步揭示全国 $PM_{2.5}$ 的时空演变规律与空间聚集特征。

3. 研究方法

3.1. 土地利用回归(LUR)模型

LUR (Land Use Regression)模型最早由 Briggs 等[6]提出，是用于模拟城市尺度大气污染物浓度空间分布特征的常用模型。该模型将监测点周边的土地利用、交通流量、工业排放、地形、人口分布、气象及社会经济因素等数据作为自变量，结合数十个监测点的大气污染物浓度数据作为因变量，构建回归模型，分析各变量对大气污染物浓度空间分布的贡献[12] [13]。本文采用 2018~2020 年全国国控监测站点 $PM_{2.5}$ 浓度作为因变量，选取相应预测因子作为自变量，构建多元线性回归模型，模型表达式为[8]：

$$y_i = \alpha_0 + \sum_{k=1}^{n} \alpha_k \cdot X_{ik} + \varepsilon_i \tag{1}$$

式中，y_i 为因变量(第 i 站点的 $PM_{2.5}$ 四季和年均值)，α_0 为截距常数(模型拟合得到的研究区域 $PM_{2.5}$ 背景浓度)，α_k 为变量系数，X_{ik} 为自变量，ε_i 为随机误差。

3.1.1. 预测变量筛选

预测变量通过双变量相关性分析进行筛选，以评估变量之间的关联程度与变化趋势。本研究使用 SPSS 软件，将 AOD、DEM、NO_2 等 22 个变量输入有监督逐步变量选择程序，开展逐步线性回归分析[4]。通过 Pearson 相关分析计算自变量与因变量之间的相关系数 r，评估 $PM_{2.5}$ 与各自变量之间的相关性，选取同类变量中相关性最强且符合预期关系的预测变量纳入 LUR 模型。本研究中模型所采用的自变量数据详见表 1。

Table 1. Table of independent variable data
表 1. 自变量数据

数据类别	名称	缩写	计量单位	数据处理方式
土地利用数据	耕地面积	AAL	m^2	缓冲区提取(100、300、500、1000、2000、3000、4000、5000)分区统计
	建设用地面积	ACL	m^2	
	林地面积	AFL	m^2	
	草地面积	AGL	m^2	
	未利用地面积	AUL	m^2	
	水域面积	AWL	m^2	
气溶胶数据	气溶胶光学厚度	AOD	/	值提取至点
气象数据	边界层高度	BLH	m	值提取至点
	相对湿度	RH	%	
	表面气压	SP	Pa	
	温度	TEMP	K	
	总降水量	TP	m	
	风速	WIND	$m \cdot s^{-1}$	

续表

社会经济与人为排放因子	国内生产总值	GDP	yuan	缓冲区提取(100、300、500、1000、2000、3000、4000、5000)分区统计
	夜间灯光指数	NL	/	
	人口	POP	Number	
	道路长度	ROAD	m	
地形因素	高程	DEM	m	值提取至点
	坡度	SLOPE	°	
大气污染指标	二氧化硫	SO_2	$\mu g/m^3$	值提取至点
	二氧化氮	NO_2	$\mu g/m^3$	
植被遥感指标	归一化指数	NDVI	/	缓冲区提取(100、300、500、1000、2000、3000、4000、5000)分区统计

3.1.2. 逐步回归与 LUR 模型构建

采用 SPSS 软件对初步筛选的变量进行逐步线性回归分析，构建 LUR 模型，并依据显著性水平($P < 0.1$)筛选变量。同时，进行多重共线性诊断，剔除方差膨胀因子(VIF) > 3 的预测变量，最终确定预测变量数据集[14]。随后，将 LUR 模型筛选出的重要预测变量分别输入四种机器学习方法，比较不同 LUR-机器学习组合模型的预测性能。

3.2. 集成机器学习方法

本研究利用 LUR 模型识别的关键变量构建预测模型。为提高污染物浓度预测的准确性，我们采用了多种回归算法，包括上述 RF、CatBoost、XGBoost 及集成模型。

堆叠(Stacking)，又称堆叠泛化，是一种集成学习方法，通过整合多个机器学习模型的预测结果构建综合模型，以提升整体预测性能[15]。其核心思想是利用基础模型(第 0 层)的折叠预测结果训练元模型(第 1 层)，从而生成集成预测结果。本研究通过评估训练集和测试集上的模型性能，从上述三种机器学习模型中选取表现最佳的两个模型作为第 0 层基础模型。随后，将这两个基础模型的预测结果分别输入三种机器学习算法，作为第 1 层元模型，以评估不同堆叠组合的集成效果。

3.3. 模型精度验证与评价

本研究采用十折交叉验证方法评估模型的可靠性与预测精度，并选用三个评价指标，包括决定系数(R^2)、平均绝对误差(*MAE*)和均方根误差(*RMSE*)。各指标计算公式如下：

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|^2 \tag{4}$$

式中，y_i是监测值，$\bar{y}_i$是监测值的均值，$\hat{y}_i$是模型预测值，n为预测样本数量。

3.4. 空间自相关性

空间自相关分析是评估地理要素空间集聚特征的重要方法，可揭示$PM_{2.5}$浓度在不同空间单元之间的依赖关系与分布规律。本研究采用全局空间自相关(Global Moran's *I*)与局部空间自相关(Local Indicators of Spatial Association, LISA)相结合的方法，刻画中国$PM_{2.5}$浓度在整体和局部尺度上的空间集聚特征。

3.4.1. 空间全局自相关

全局莫兰指数(Global Moran's *I*)用于表征研究区域内$PM_{2.5}$浓度的整体空间关联程度，判断其在空间上是否存在显著的集聚特征。计算公式如下：

$$I=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}W_{ij}\left(x_i-\bar{x}\right)\left(x_j-\bar{x}\right)}{\sum_{i=1}^{n}\sum_{j=1}^{n}W_{ij}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2} \tag{5}$$

式中：I为全局莫兰指数；n为监测城市总数；x_i、x_j分别为第i、j城市的$PM_{2.5}$浓度；$\bar{x}$为研究区$PM_{2.5}$浓度均值；W_{ij}为要素间i、j之间的空间权重。

$I\in[-1,1]$：当I为正数时，$PM_{2.5}$浓度呈现空间正相关性，主要表现为高–高聚集和低–低相邻聚集；当I为负数时，$PM_{2.5}$浓度呈现空间负相关性，主要表现为高–低或低–高分布；而$I=0$时，$PM_{2.5}$浓度呈随机空间分布；当标准化检验统计量$Z_i>1.96$且显著性概率$p<0.05$时，表明空间自相关在统计学上具有显著性。

3.4.2. 空间局部自相关

局部空间自相关(LISA)可用于识别各监测单元与其邻近单元之间的局部空间关联模式，从而定位高值集聚区、低值集聚区以及空间异常区域。计算公式如下：

$$I_i=\frac{\left(x_i-\bar{x}\right)}{S^2}\sum_{j=1,j\neq i}^{n}W_{ij}\left(x_j-\bar{x}\right) \tag{6}$$

式中：S^2为$PM_{2.5}$浓度的方差。

3.5. 时空加权集成模型

从时间维度来看，空气污染浓度并非恒定不变，而是呈现明显的季节性变化和日际波动。相邻日期的污染状况通常具有较强的相似性，但这种相似性会随着时间间隔的增加而逐渐减弱。同时，$PM_{2.5}$浓度在空间上也存在显著差异，不同地点间的污染水平差异通常随空间距离的增加而扩大，且空间影响关系可能具有非对称性。鉴于空气污染复杂的时空非平稳特征，不同于以往主要关注单一效应关系的研究，本研究引入时空加权机制，将不同时间点的时空权重信息融入原有的机器学习集成模型，以更准确地刻画和捕捉空气污染的时空变化规律，进一步提高$PM_{2.5}$浓度估算的精度。

3.6. 技术路线

本研究基于土地利用、气象、NDVI、AOD、道路密度、地形及社会经济数据，通过相关性分析、逐步回归和共线性诊断筛选预测变量。随后，比较四种机器学习方法及LUR组合、集成模型的预测性能，并结合时空加权项分析$PM_{2.5}$空间分布及其空间自相关特征(图 1)。

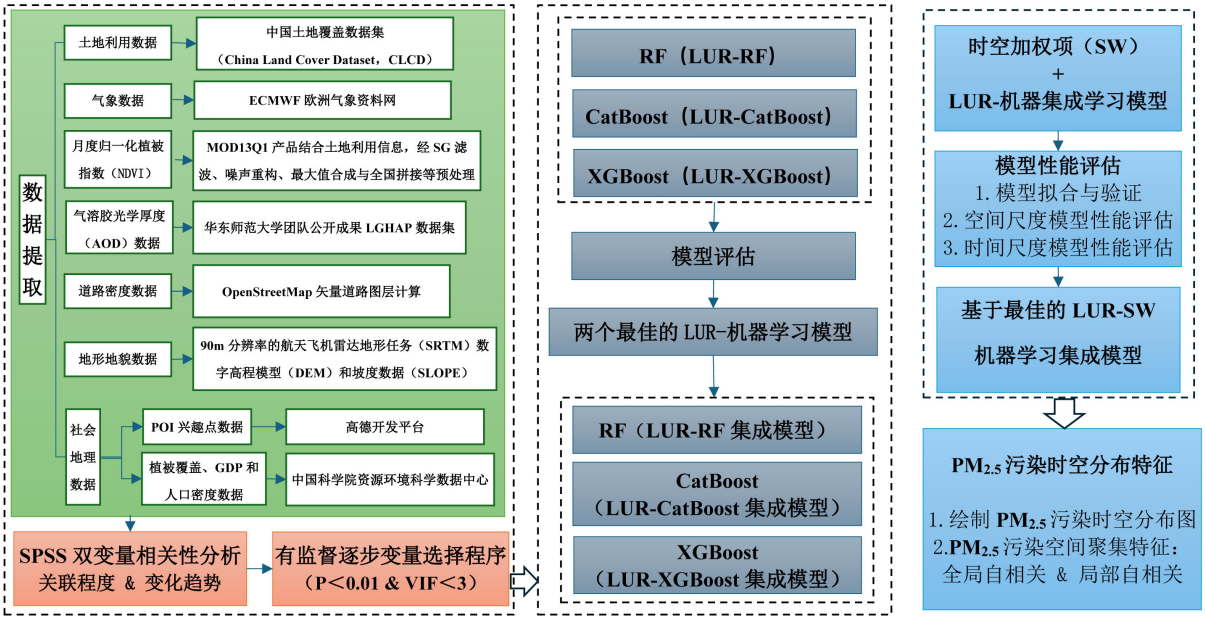


Figure 1. Technological route
图 1. 技术路线图

4. 结果与分析

4.1. LUR 模型构建与变量选取

经过相关性分析筛选自变量后，以研究区 $PM_{2.5}$ 浓度为因变量，剔除 VIF > 3 的变量，采用多元线性回归方法构建 LUR 模型。研究区 2018~2020 年 $PM_{2.5}$ 的 LUR 模型结果分别见表 2、表 3 和表 4。

Table 2. Results of variable selection for the LUR model in 2018
表 2. 2018 年 LUR 模型变量筛选结果

	B	Beta	t	显著性	Partial R^2	VIF	
常量	11.523		53.471	0.000			
NO_2	0.797	0.429	315.983	0.000	0.566	1.766	
AOD	42.550	0.416	357.277	0.000	0.767	1.304	
SO_2	0.335	0.130	108.394	0.000	0.722	1.386	
TEMP	−0.285	−0.095	−79.664	0.000	0.738	1.355	
TP	−0.696	−0.081	−71.841	0.000	0.827	1.209	
GDP	-8.293×10^{-5}	−0.087	−75.147	0.000	0.781	1.281	R = 0.716
RH	−0.116	−0.073	−51.260	0.000	0.518	1.930	R^2 = 0.513
Road	−0.005	−0.041	−26.656	<0.001	0.441	2.266	调整后 R^2 = 0.513
AUL	0.000	0.034	32.496	<0.001	0.945	1.059	
SLOPE	−0.150	−0.016	−14.473	<0.001	0.883	1.132	
Pop	0.000	−0.024	−17.595	<0.001	0.567	1.763	
AWL	-1.648×10^{-6}	−0.011	−10.470	<0.001	0.970	1.031	
AFL	-3.275×10^{-8}	−0.011	−8.814	<0.001	0.719	1.390	

续表

WIIND	0.155	0.011	8.370	<0.001	0.633	1.581
BLH	−0.001	−0.07	−5.252	<0.001	0.529	1.891
NL	0.010	0.004	3.002	0.003	0.574	1.743

Table 3. Results of variable selection for the LUR model in 2019
表 3. 2019 年 LUR 模型变量筛选结果

变量	B	Beta	t	显著性	Partial R^2	VIF	
常量	16.389		43.301	0.000			
NO_2	0.697	0.380	271.453	0.000	0.556	1.798	
AOD	36.889	0.363	299.598	0.000	0.742	1.348	
TEMP	−0.512	−0.162	−105.954	0.000	0.466	2.148	
GDP	-8.393×10^{-5}	−0.091	−76.969	0.000	0.772	1.295	
SO_2	0.302	0.095	78.229	0.000	0.732	1.367	
TP	−0.576	−0.070	−61.997	0.000	0.857	1.167	
SP	0.013	0.065	55.727	0.000	0.811	1.232	
AUL	0.000	0.028	25.527	<0.001	0.905	1.105	
NDVI	−28.947	−0.095	−55.608	0.000	0.372	2.685	R = 0.689 R^2 = 0.475 调整后 R^2 = 0.475
Road	−0.002	−0.018	−11.338	<0.001	0.412	2.427	
RH	−0.049	−0.032	−24.041	<0.001	0.605	1.652	
AFL	1.095×10^{-7}	0.036	25.860	<0.001	0.573	1.746	
Pop	0.000	−0.024	−16.654	<0.001	0.539	1.856	
AGL	2.176×10^{-5}	0.026	22.313	<0.001	0.782	1.279	
NL	−0.055	−0.023	−16.839	<0.001	0.560	1.786	
AWL	-2.550×10^{-6}	−0.017	−15.636	<0.001	0.951	1.051	
AAL	4.321×10^{-6}	0.017	14.892	<0.001	0.840	1.190	
BLH	−0.002	−0.013	−11.041	<0.001	0.748	1.338	
SLOPE	−0.108	−0.011	−9.977	<0.001	0.852	1.174	

Table 4. Results of variable selection for the LUR model in 2020
表 4. 2020 年 LUR 模型变量筛选结果

变量	B	Beta	t	显著性	Partial R^2	VIF	
常量	26.101		93.897	0.000			
NO_2	0.680	0.374	293.145	0.000	0.590	1.695	
AOD	37.846	0.392	350.839	0.000	0.769	1.301	
TEMP	−0.413	−0.140	−101.965	0.000	0.506	1.975	R = 0.721 R^2 = 0.520 调整后 R^2 = 0.520
SO_2	0.479	0.130	113.074	0.000	0.726	1.378	
TP	−0.534	−0.081	−75.963	0.000	0.841	1.189	
GDP	-6.639×10^{-5}	−0.070	−63.941	0.000	0.801	1.248	

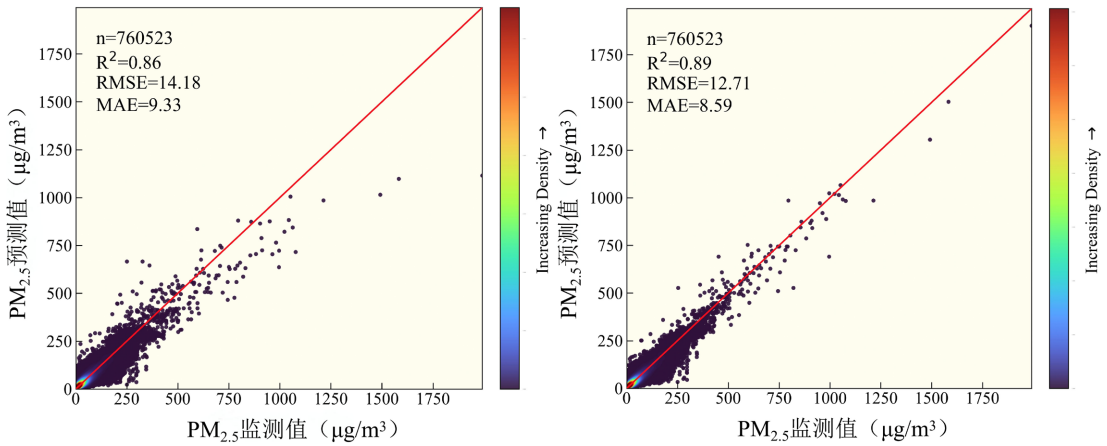
续表

NDVI	−31.840	−0.113	−71.475	0.000	0.381	2.624
Road	−0.002	−0.018	−12.159	<0.001	0.425	2.353
AUL	0.000	0.033	32.339	<0.001	0.929	1.076
AFL	1.208×10^{-5}	0.043	33.494	<0.001	0.593	1.687
AAL	6.401×10^{-6}	0.028	26.326	<0.001	0.846	1.181
RH	−0.071	−0.047	−34.432	<0.001	0.508	1.967
BLH	−0.004	−0.038	−25.544	<0.001	0.440	2.274
Pop	-9.735×10^{-5}	−0.024	−18.407	<0.001	0.550	1.819
AWL	-2.298×10^{-6}	−0.016	−16.427	<0.001	0.956	1.046
NL	−0.018	−0.009	−6.625	<0.001	0.550	1.817
WIND	−0.062	−0.005	−3.607	<0.001	0.563	1.777

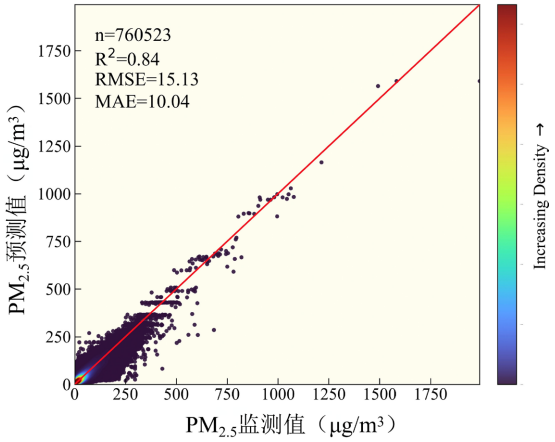
标准化回归系数 Beta 的正负反映各变量对研究区 $PM_{2.5}$ 浓度的作用方向，其绝对值大小表征变量对 $PM_{2.5}$ 空间分布的相对影响程度。以 2018 年结果为例，AOD 的标准化回归系数 Beta 为 0.416，表明 AOD 与 $PM_{2.5}$ 浓度呈正相关，即气溶胶光学厚度增加会促进 $PM_{2.5}$ 浓度升高。2018~2020 年模型调整后 R^2 分别为 0.513、0.475 和 0.520，整体在 0.475~0.520 范围内小幅波动，模型解释力相对稳定，其中 2020 年的拟合效果最佳。筛选后各年份变量的 VIF 均小于 3，表明模型不存在明显的多重共线性问题。根据 2018~2020 年的回归结果，NO_2 与 AOD 是影响 $PM_{2.5}$ 空间分布的核心驱动因子，温度和相对湿度对 $PM_{2.5}$ 污染的影响方向较为稳定；而 SO_2、NDVI、风速以及地形和气压等变量的入选情况与贡献程度存在一定的年际差异。

4.2. 集成机器学习方法性能评价

图 2 展示了集成机器学习模型的拟合效果。三种模型的 R^2 均达到 0.84 以上，表明其均具有较好的拟合能力。其中，CatBoost、XGBoost 和 RF 模型的 R^2 分别为 0.86、0.89 和 0.84，表明 XGBoost 模型的拟合性能最优，而 CatBoost 与 RF 模型的拟合性能较为接近。RF 模型的 RMSE 最高，为 15.13 μg/m³，而 XGBoost 模型的 RMSE 最低，为 12.71 μg/m³。在 MAE 方面，RF 模型最高，为 10.04 μg/m³，XGBoost 模型最低，为 8.59 μg/m³。三种模型的 RMSE 均低于 20 μg/m³，MAE 均低于 15 μg/m³，整体表明三种模型均具有较好的预测精度和拟合效果。



(a) Catboost 模型拟合图　　(b) XGBoost 模型拟合图



(c) RF 模型拟合图

Figure 2. Ensemble machine learning illustration of the fitting results
图 2. 集成机器学习拟合效果图

4.3. $PM_{2.5}$ 时空分布特征

本研究基于 LUR 模型预测的 $PM_{2.5}$ 浓度计算中国 $PM_{2.5}$ 平均浓度，如表 5 所示。结果揭示了中国 $PM_{2.5}$ 浓度显著的季节差异，其中夏季浓度最低，春季和秋季浓度相对适中，冬季浓度最高，尤其在北疆和华北平原中部地区表现较为突出。年度空间分布结果显示，部分南疆地区 $PM_{2.5}$ 浓度为 90~120 μg/m³，华北平原中部和部分西北部地区为 60~90 μg/m³，其他区域总体低于 60 μg/m³。南疆地区 $PM_{2.5}$ 浓度较高可能与地形相对封闭，气候干旱少雨等因素密切相关。全国 $PM_{2.5}$ 浓度呈现显著的时空异质性，整体表现为冬高夏低、北高南低、东高西低的分布特征，华北平原、黄淮海平原及南疆盆地是主要的高污染区域。结合核心变量 AOD 及已有研究可知，华北平原的 $PM_{2.5}$ 污染主要受工业、交通等人为源排放驱动；南疆地区沙尘天气频发，自然尘源是导致 $PM_{2.5}$ 浓度偏高的主要因素，人为排放则起到叠加作用。全国 $PM_{2.5}$ 浓度展示出较强的空间集聚性，整体以高－高集聚和低－低集聚格局为主。

Table 5. Average $PM_{2.5}$ concentrations in China, 2018~2020
表 5. 2018~2020 年中国 $PM_{2.5}$ 平均浓度

地区	春	夏	秋	冬	年	地区	春	夏	秋	冬	年
北京市	50.1396	37.8183	41.4614	46.3154	43.9113	湖南省	31.8248	21.139	35.8585	63.2166	38.803
天津市	52.3648	39.974	52.3609	69.78	53.3612	广东省	23.8573	14.3266	28.5126	31.4368	24.1268
上海市	38.3985	24.896	26.8863	46.0603	34.1343	海南省	15.5102	9.716	16.0681	18.9125	14.5962
重庆市	39.5431	21.231	27.255	60.8415	35.5519	四川省	32.33	19.4371	25.3561	58.8004	34.5155
河北省	51.3994	34.979	53.1378	80.6086	54.5622	贵州省	26.3722	13.4315	18.587	32.2306	22.9067
山西省	50.8393	34.5215	49.7538	84.1906	55.4604	云南省	28.6137	13.0847	18.0053	27.7249	21.7667
辽宁省	39.7688	23.0243	33.1573	55.0326	36.9084	陕西省	43.5433	25.3482	43.0558	89.8075	53.7862
吉林省	38.3965	16.2278	21.4769	54.4482	30.4861	甘肃省	35.406	22.3676	24.9848	49.5969	34.7832
黑龙江省	28.3052	12.3241	16.6216	47.8855	24.3497	青海省	28.7112	17.7768	21.3846	37.8098	28.1251
江苏省	38.2182	30.152	41.4663	73.0061	48.5546	内蒙古自治区	32.433	21.6051	25.4687	38.6105	29.1922
浙江省	29.7209	18.762	25.8059	37.3028	28.2423	广西壮族自治区	26.0617	17.3737	28.1799	34.3015	26.5834

续表

安徽省	45.8361	29.5799	48.0722	80.2196	52.4344	西藏自治区	28.404	14.994	14.0839	19.6458	17.7458
福建省	23.5493	14.3955	21.6992	25.8405	21.7126	宁夏回族自治区	37.8987	23.7357	27.8001	45.8847	35.694
江西省	30.7096	19.8653	32.895	42.57	32.2577	新疆维吾尔自治区	70.146	14.994	39.7899	86.844	59.3224
山东省	49.1644	31.9462	52.9135	88.0684	55.6072						
河南省	55.8437	34.2797	66.7868	111.3529	69.2283						
湖北省	36.0397	24.6272	42.0262	76.9485	47.1216						

4.4. 空间全局自相关(Global Moran's *I*)检验

由表 6 可知，春季、夏季、秋季、冬季及全年 Moran's *I* 的 Zi 值分别为 25.7684、25.8342、26.3007、26.1774 和 26.3137，且 *p* 值均为 0.001，表明研究区 $PM_{2.5}$ 空间聚集特征具有较高的统计显著性。Moran's *I* 指数分别为 0.8849、0.8207、0.8498、0.892 和 0.8937，表明中国春季、夏季、秋季、冬季及全年 $PM_{2.5}$ 浓度均呈现较强的空间正自相关特征，其空间自相关程度表现为全年 > 冬季 > 春季 > 秋季 > 夏季。

Table 6. Moran's *I* index table
表 6. Moran's *I* 指数

季节	Moran's I 指数	Zi	*p*
春季	0.8849	25.7684	0.001
夏季	0.8207	25.8342	0.001
秋季	0.8498	26.3007	0.001
冬季	0.892	26.1774	0.001
全年	0.8937	26.3137	0.001

4.5. 空间局部自相关(Local Indicators of Spatial Association, LISA)检验

由表 7 可见，四季及全年的 LISA 聚类结果中，多数监测单元表现为高－高(H-H)、低－低(L-L)两类集聚模式，表明 $PM_{2.5}$ 浓度具有显著的空间集聚特征。高－高聚集区主要分布于我国新疆南疆盆地与华北、黄淮等中部平原地区；低－低集聚区主要集中于青藏高原及西南、东北地区，区域整体污染水平较低。少部分区域呈现高－低(H-L)或低－高(L-H)的空间异常类型，表明局部空间异质性相对有限。

Table 7. Table of local Moran's indices
表 7. 局部莫兰指数

地区	Moran's *I* 指数					地区	Moran's *I* 指数				
	春	夏	秋	冬	年		春	夏	秋	冬	年
北京市	0.000031	0.813814	0.462226	−0.000021	0.000025	湖南省	0.000012	0.024822	0.056593	0.000016	0.000006
天津市	0.000053	0.988093	1.243595	0.000038	0.000099	广东省	0.000069	0.377584	0.377584	0.000139	0.000113
上海市	0	0.022264	−0.00561	0.000003	−0.000001	海南省	0.000241	0.889894	1.150217	0.000416	0.000427
重庆市	−0.000002	0.022276	0.110497	0.000004	0.000001	四川省	0.000005	0.079004	0.266787	0.000006	0.000004
河北省	0.000082	0.750857	1.801064	0.000233	0.000225	贵州省	0.000022	0.325902	0.60936	0.000055	0.000058

续表

山西省	0.000078	0.720192	1.531835	0.000236	0.000241	云南省	0.000011	0.476598	0.936073	0.000063	0.000054
辽宁省	0.000001	0.022073	0.018609	0.000001	0.000002	陕西省	0.000014	0.077392	0.460579	0.000165	0.000117
吉林省	0	0.191203	0.456844	0.000001	0.000013	甘肃省	0.000001	0.076933	0.244263	0.000005	0.000004
黑龙江省	0.000005	0.56364	1.142335	0.000003	0.000024	青海省	0.000004	0.123954	0.607948	0.000014	0.00001
江苏省	0.000038	0.272104	0.53431	0.000114	0.000115	内蒙古自治区	0	0.078192	0.222743	0.000006	0.000003
浙江省	0.00001	0.027842	0.101517	0.00004	0.000022	广西壮族自治区	0.000044	0.232231	0.15009	0.000099	0.000077
安徽省	0.000053	0.277333	1.099648	0.000125	0.000221	西藏自治区	0.000013	2.15489	1.320589	0.000015	0.000015
福建省	0.000041	0.306937	0.38469	0.000122	0.000087	宁夏回族自治区	0.000002	0.021921	0.054796	−0.000001	−0.000003
江西省	0.000015	0.031042	0.005421	0.000027	0.000012	新疆维吾尔自治区	0.000207	7.213769	1.016313	0.000072	0.000072
山东省	0.00009	0.579735	2.019324	0.000324	0.000296						
河南省	0.000152	0.664486	3.652222	0.00074	0.000664						
湖北省	0.000001	0.036919	0.444325	0.000112	0.000076						

相较于传统单一 LUR 模型，本研究构建的时空加权集成模型能够更有效地捕捉 $PM_{2.5}$ 复杂的时空非线性变化特征，具有更优的模拟性能，可为全国 $PM_{2.5}$ 空间制图、区域差异化污染治理及健康风险评估提供科学参考。

本研究仍存在一定的不确定性与局限性。模型模拟性能呈现明显的区域差异，在南方城市区域的适用性较好，而在北方乡村及荒漠区域的预测误差相对较高。一方面，全国监测站点空间分布不均，偏远地区的数据覆盖不足；另一方面，研究尚未纳入城市形态等精细化污染扩散指标，对局部污染特征的刻画仍存在局限。未来可进一步优化变量体系与数据集，以提高模型在全国范围内的模拟精度。

5. 结论

本研究整合 RF、CatBoost 与 XGBoost 三类机器学习算法，构建 LUR-时空加权集成机器学习模型，以弥补传统土地利用回归(LUR)模型难以捕捉变量间非线性关系的不足，实现 2018~2020 年中国 $PM_{2.5}$ 浓度的高精度时空模拟，主要结论如下：

(1) 传统 LUR 模型的拟合精度相对有限，调整后 R^2 仅为 0.475~0.520。集成机器学习模型的预测性能显著提升，其中 XGBoost 模型表现最优，R^2 达到 0.89；CatBoost 模型次之($R^2 = 0.86$)；RF 模型的预测精度相对较低($R^2 = 0.84$)，误差最大。NO_2 和 AOD 是影响 $PM_{2.5}$ 空间分布的核心变量，而气象、生态及社会因素对 $PM_{2.5}$ 的影响呈现不同程度的年际差异。

(2) 中国 $PM_{2.5}$ 浓度存在显著的季节波动特征，各季节及全年全局莫兰指数 Moran's I 均大于 0.82，表明 $PM_{2.5}$ 浓度具有较强的空间正自相关性，空间聚集强度排序为：全年 > 冬季 > 春季 > 秋季 > 夏季。局部 LISA 结果表明，$PM_{2.5}$ 主要表现为高－高和低－低集聚为主要模式。

基金项目

陕西省大学生创新创业训练计划项目(编号：S202510709108)。

参考文献

[1] Li, R., Ma, T., Xu, Q. and Song, X. (2018) Using MAIAC AOD to Verify the $PM_{2.5}$ Spatial Patterns of a Land Use Regression Model. *Environmental Pollution*, **243**, 501-509. https://doi.org/10.1016/j.envpol.2018.09.026

[2] Wei, J., Huang, W., Li, Z., Xue, W., Peng, Y., Sun, L., *et al*. (2019) Estimating 1-Km-Resolution $PM_{2.5}$ Concentrations across China Using the Space-Time Random Forest Approach. *Remote Sensing of Environment*, **231**, Article 111221. https://doi.org/10.1016/j.rse.2019.111221

[3] Wu, C., Zeng, Y. and Lung, S.C. (2018) A Hybrid Kriging/Land-Use Regression Model to Assess $PM_{2.5}$ Spatial-Temporal Variability. *Science of The Total Environment*, **645**, 1456-1464. https://doi.org/10.1016/j.scitotenv.2018.07.073

[4] Huang, L., Zhang, C. and Bi, J. (2017) Development of Land Use Regression Models for $PM_{2.5}$, SO_2, NO_2 and O_3 in Nanjing, China. *Environmental Research*, **158**, 542-552. https://doi.org/10.1016/j.envres.2017.07.010

[5] Han, L., Zhao, J., Gao, Y., Gu, Z., Xin, K. and Zhang, J. (2020) Spatial Distribution Characteristics of $PM_{2.5}$ and PM_{10} in Xi'an City Predicted by Land Use Regression Models. *Sustainable Cities and Society*, **61**, Article 102329. https://doi.org/10.1016/j.scs.2020.102329

[6] Briggs, D.J., Collins, S., Elliott, P., Fischer, P., Kingham, S., Lebret, E., *et al*. (1997) Mapping Urban Air Pollution Using GIS: A Regression-Based Approach. *International Journal of Geographical Information Science*, **11**, 699-718. https://doi.org/10.1080/136588197242158

[7] 刘炳杰, 彭晓敏, 李继红. 基于 LUR 模型的中国 $PM_{2.5}$ 时空变化分析[J]. 环境科学, 2018, 39(12): 5296-5307.

[8] 杨明亮, 朱宗玖. 基于 LUR 模型下 $PM_{2.5}$ 浓度的空间分布模拟分析[J]. 大气与环境光学学报, 2022, 17(3): 347-359.

[9] Wang, Z., Li, R., Chen, Z., Yao, Q., Gao, B., Xu, M., *et al*. (2022) The Estimation of Hourly PM2.5 Concentrations across China Based on a Spatial and Temporal Weighted Continuous Deep Neural Network (STWC-DNN). *ISPRS Journal of Photogrammetry and Remote Sensing*, **190**, 38-55. https://doi.org/10.1016/j.isprsjprs.2022.05.011

[10] Wei, J., Liu, S., Li, Z., Liu, C., Qin, K., Liu, X., *et al*. (2022) Ground-Level NO_2 Surveillance from Space across China for High Resolution Using Interpretable Spatiotemporally Weighted Artificial Intelligence. *Environmental Science & Technology*, **56**, 9988-9998. https://doi.org/10.1021/acs.est.2c03834

[11] Li, T., Wang, Y. and Wu, J. (2024) Deriving $PM_{2.5}$ from Satellite Observations with Spatiotemporally Weighted Tree-Based Algorithms: Enhancing Modeling Accuracy and Interpretability. *npj Climate and Atmospheric Science*, **7**, Article No. 138. https://doi.org/10.1038/s41612-024-00692-4

[12] Ho, C., Chan, C., Cho, C., Lin, H., Lee, J. and Wu, C. (2015) Land Use Regression Modeling with Vertical Distribution Measurements for Fine Particulate Matter and Elements in an Urban Area. *Atmospheric Environment*, **104**, 256-263. https://doi.org/10.1016/j.atmosenv.2015.01.024

[13] 吴健生, 谢舞丹, 李嘉诚. 土地利用回归模型在大气污染时空分异研究中的应用[J]. 环境科学, 2016, 37(2): 413-419.

[14] 张平, 张凤倩, 朱慧敏, 等. 基于 FA-SVM 优化 LUR 模型的汾渭平原 $PM_{2.5}$ 时空格局模拟[J]. 西安工程大学学报, 2025, 39(3): 89-101.

[15] Yu, W., Li, S., Ye, T., Xu, R., Song, J. and Guo, Y. (2022) Deep Ensemble Machine Learning Framework for the Estimation of $PM_{2.5}$ Concentrations. *Environmental Health Perspectives*, **130**, Article 37004. https://doi.org/10.1289/ehp9752