

叶昊天,邹春辉,田宏伟.一种基于机器学习和葵花 8 号数据的 $PM_{2.5}$ 浓度监测方法[J].气象与环境科学,2023,46(3):89-97.
Ye Haotian,Zou Chunhui,Tian Hongwei. A Method for $PM_{2.5}$ Concentration Monitoring Based on Machine Learning and Himawari-8 Satellite Data[J].
Meteorological and Environmental Sciences,2023,46(3):89-97.
doi:10.16765/j.cnki.1673-7148.2023.03.012

一种基于机器学习和葵花 8 号数据的 $PM_{2.5}$ 浓度监测方法

叶昊天^{1,2,3}, 邹春辉^{1,2}, 田宏伟^{1,2}

(1. 中国气象局·河南省农业气象保障与应用技术重点开放实验室, 郑州 450003; 2. 河南省气象科学研究所, 郑州 450003;
3. 安阳国家气候观象台, 河南 安阳 455000)

摘要: 基于卫星遥感数据和气象数据构建模型是实现 $PM_{2.5}$ 浓度监测的重要方法。利用 2017—2020 年河南区域的葵花 8 号 AOD 相关产品、地面气象观测数据及 $PM_{2.5}$ 浓度逐时地面观测数据, 建立样本数据库, 探究了 $PM_{2.5}$ 浓度与遥感反演参数、气象因子的相关关系, 提出一种基于空间距离的机器学习方法, 构建了卫星遥感与气象数据相结合的 $PM_{2.5}$ 监测模型, 并利用 $PM_{2.5}$ 观测数据对模型进行散点拟合和空间分布的验证。结果表明: $PM_{2.5}$ 浓度与能见度 (V)、气温 (T)、气溶胶光学厚度 (AOD) 有极为显著的相关关系; 从 $PM_{2.5}$ 浓度反演值与观测值的散点回归分析来看, $PM_{2.5}$ 浓度反演值与观测值相关系数 $R^2 = 0.73$; 从 $PM_{2.5}$ 浓度反演值与观测值的空间分布图来看, $PM_{2.5}$ 浓度反演值实现了空间格点的小时监测, 弥补了非晴空及夜间监测数据的缺失, 在空间分布上与观测值一致, 具有很高的精确性和可靠性。基于空间距离的机器学习方法构建 $PM_{2.5}$ 浓度监测模型能很好地应用于 $PM_{2.5}$ 浓度监测, 为 $PM_{2.5}$ 浓度监测提供了新的思路和方法。

关键词: $PM_{2.5}$ 浓度; 空间距离; 机器学习; 葵花 8 号

中图分类号: X87; X83; X513

文献标志码: B

文章编号: 1673-7148(2023)03-0089-09

引言

大气污染是大多数发展中国家普遍面临的一个难题。经济发展导致中国极大地加剧了对能源的需求, 工业化进程的加剧导致中国大部地区空气污染日趋严重。大气颗粒物 (PM) 是地球大气中存在的各种固态和液态颗粒状物质的总称, 即大气气溶胶^[1]。大气气溶胶粒子的空气动力学直径为 1~100 μm , 它主要来源于土壤扬尘、海盐和植物花粉等自然源和人类生产生活活动造成的人为源^[2-3]。其中可吸入颗粒物 $PM_{2.5}$, 即空气动力学直径小于 2.5 μm 的颗粒物, 大多含有重金属等有毒物质, 进入人体后会沉积

在支气管和肺泡中, 对人体健康的影响较大^[4-5]。

$PM_{2.5}$ 浓度监测以传统的地面监测为主, 通过建立环境监测网络为大气环境质量监测、评估等提供大量的地面监测数据。由于 $PM_{2.5}$ 地面监测站运营成本高, 数量有限, 致使地面站只能获取局地小范围内的 $PM_{2.5}$ 浓度信息, 很难满足全面准确的空气质量监测要求, 地面监测网不能全面、连续、实时地反映我国空气质量状况。因此, 建立一套高效、准确、快速的环境监测、调查与评价系统十分必要^[6]。随着遥感技术的快速发展, 具有覆盖范围广、监测成本低、处理速度快、可长时间持续动态监测等优点的遥感技术已被广泛地应用于大气监测领

收稿日期: 2022-09-19; 修订日期: 2022-12-31

基金项目: 国家自然科学基金项目 (41805090); 中国气象局·河南省农业气象保障与应用技术重点开放实验室应用技术研究基金项目 (KZ201911, KM202018); 安阳市国家气候观象台开放研究基金项目 (AYNCOF2022313)

作者简介: 叶昊天 (1995-), 男, 河南新乡人, 工程师, 硕士, 从事生态环境遥感研究. E-mail: 624864475@qq.com

通信作者: 邹春辉 (1974-), 男, 江西余江人, 正高级工程师, 硕士, 从事遥感数据应用研究. E-mail: z198@163.com

域。利用遥感技术可以揭示污染源位置及扩散状态,对大尺度的大气环境变化和大气环境污染情况进行快速、实时、动态地监测,并能对突发性大气环境污染事件进行追踪调查,可有效减少和降低大气污染造成的损失^[7-9]。

近年来,国内外学者直接或间接地建立基于卫星遥感反演的气溶胶光学厚度(AOD)与近地面颗粒物浓度间的相关模型,对地面颗粒污染物浓度和监测区域污染进行了估算。Li等^[10]利用MODIS AOD数据,依赖暗背景算法,对中国东部地区的大气污染进行了有效监测,发现AOD能很好地反映城市区域内污染程度的复杂变化,MODIS AOD与多波段太阳光度计测量结果对比显示,MODIS AOD均在估计的误差范围内。Engel等^[11]建立了MODIS AOD与PM_{2.5}地面站点监测数据之间的相关关系,其中与PM_{2.5}浓度日均值的 R^2 达到0.43,与PM_{2.5}浓度小时均值的 R^2 达到0.40。Wang等^[12]在假设气溶胶垂直分布为负指数衰减的前提下,首先给出MODIS AOD与近地面颗粒物浓度的相关模型,并通过实测数据分别验证了垂直订正和湿度订正方法的有效性。Liu等^[13]使用了一个通用线性回归模型,将卫星遥感的AOD、边界层高度和相对湿度作为主要因子,同时引入若干地理、环境因素作为辅助因子,得到PM_{2.5}浓度的反演模型。吴序鹏等^[14]利用MODIS卫星遥感资料,结合地面观测资料,分析了2013年12月发生在淮安的一次持续性大气污染事件,探讨了卫星遥感气溶胶光学厚度与地面污染物浓度的关系。基于AOD反演PM_{2.5}浓度在一些地区达到了较好的反演结果。但上述研究中采用的卫星遥感影像多为MODIS,分辨率普遍偏低,难以保证遥感影像数据和地面监测数据在空间和时间上匹配的一致性。

张猛等^[15]采用机器学习中常用的多层映射反向传播神经网络,构建了波段反射率与PM_{2.5}浓度之间的反演模型,Landsat 8卫星在中国大陆地区的重访周期为16天,时间分辨率较差。Wei J等^[16]利用PM_{2.5}地基观测数据、MODIS AOD等,基于卫星遥感技术和机器学习方法,考虑了影响PM_{2.5}污染丰富的自然(如气象、地表覆盖等)和人文(如人类分布和污染排放等)因素,及大气污染明显的时空异质特性,发展了时空-极端随机树(STET)模型,构建了稳健的PM_{2.5}-AOD转换关系,重构了中国近20年1 km分辨率的PM_{2.5}浓度历史数据集。在整

体精度、时间尺度、空间覆盖范围和空间分辨率方面总体上较好,模型能够较准确地估算中国不同年份的PM_{2.5}日浓度,同时能较准确预测历史PM_{2.5}月浓度。

日本地球观测研究中心(EORC)2015年发布基于葵花8(Himawari8)的全球5 km分辨率气溶胶光学厚度产品,基于葵花8的AOD产品时间分辨率为10 min,对于开展区域PM时空变化监测,具有较高的实际应用价值。李鹏等^[17]基于葵花8气溶胶光学厚度(AOD)产品,分析了天津市2018—2019年气溶胶光学厚度时空分布特征,结果表明,葵花8号AOD数据在区域大气污染监测领域有较好的应用潜力。牛晓君等^[18]提出利用葵花卫星可见光波段数据及地表反射率比值反演气溶胶光学厚度的新算法,结果表明,经新算法反演的葵花8号AOD相对于MODIS AOD,相对误差降低了30%以上,无论在反演精度、有效反演空间范围和有效反演时间范围和空间分辨率等方面均为更优。葛邦宇等^[19]基于葵花8号数据,分析了京津冀地区的大气空间分布和日变化情况,采用暗目标法,对葵花8号静止卫星陆地气溶胶光学厚度反演,能反映气溶胶的高时间变化信息,有望成为大气环境污染变化监测新的重要手段。

从国内外利用遥感卫星AOD数据监测地面PM_{2.5}浓度的研究来看,目前葵花8号数据的应用仍处于起步阶段,且单一的线性模型已经无法满足当前的研究需要,拟合模型的类型将趋于丰富,通过神经网络模型、地理加权回归模型、混合模型的应用,反演精度不断提高^[20]。此外,由于AOD等遥感数据极易受云、雾等因素影响,单纯利用遥感数据难以在业务中实现气溶胶逐时监测业务化应用。本文提出了一种基于空间距离的机器学习方法,利用葵花8号AOD相关产品和地面气象观测数据,构建了卫星遥感与气象数据相结合的PM_{2.5}监测模型,利用PM_{2.5}观测数据对模型进行散点拟合和空间分布的验证。研究结果实现了以PM_{2.5}浓度为代表的河南省大气气溶胶小时监测,可为PM_{2.5}监测提供新的思路和方法,具有较为重要的科学意义与广泛的应用价值。

1 数据与方法

1.1 研究区概况

河南省位于31.3°—36.5°N、110.3°—116.7°E,东邻山东和安徽省,西接陕西和山西省,北与河北省

相临,南面与湖北省接壤,总面积为 $1.67 \times 10^5 \text{ km}^2$;地势西高东低,中部和东部为黄淮海平原,西南部为南阳盆地。河南省大部分区域处于暖温带,南部跨亚热带,属北亚热带向暖温带过渡的大陆性季风气候,同时还具有自东向西由平原向丘陵山地气候过渡的特征,具有四季分明、雨热同期、复杂多样和气象灾害频繁的特点。全省由南向北年平均气温为 $10.5 \sim 16.7 \text{ }^\circ\text{C}$,年均降水量为 $407.7 \sim 1295.8 \text{ mm}$,降雨以6—8月的最多,年均日照时数为 $1285.7 \sim 2292.9 \text{ h}$,全年无霜期为201—285天。河南省作为中国能源消费大省,以煤炭为主要能源,同时河南省也是中国交通物流枢纽,汽车污染物排放严重,此外河南省农业体量大,其中耕地面积接近50%,种植业和养殖业也是重要污染来源,大气环境质量面临极大的挑战^[21]。

1.2 数据来源

葵花8号卫星的星下点位于 140.7°E 的赤道上空,距地高度约为 35800 km ,搭载有辐射成像仪(AHI)。该传感器共有3个可见光通道和13个近红外及红外通道,最高分辨率可达 0.5 km 。AHI全盘扫描周期为 10 min ,如选定时间对特定区域进行扫描,可达 2.5 min 观测周期。由于AHI传感器设置了对气溶胶较为敏感的可见光波段,因此在气溶胶反演方面潜力巨大。本研究选取了2017年1月—2020年12月的每日逐时的葵花8卫星AOD产品(空间分辨率为 0.05°),数据来源于JAXA(日本宇宙航空研究开发机构)官网(<https://www.eorc.jaxa.jp/tree/index.html>),数据产品由日本EROC(Earth Observation Research Center)制作。该AOD数据产品先后有2.0、2.1和3.0的版本,考虑到数据版本的连续性,研究中统一采用的是2.1的版本。AOD数据产品中,主要提供了AOD(Aerosol Optical Depth)、AE(Angstrom Exponent)、RF(Optical Depth Ratio Fine)、SSA(Single Scattering Albedo at 500 nm)四种遥感参数数据,本研究仅将AOD、AE、RF纳入研究。

选取的河南省2017年1月—2020年12月的PM_{2.5}质量浓度逐时地面观测数据,来自环保部环境监测总站空气质量实时发布系统(<http://www.cnemc.cn/>)。其中,河南省内共有84个国控空气质量监测站,主要分布在城区。因数据中存在部分缺测或明显异常的记录,为保证数据有效性,剔除了无监测数据、部分要素缺测或数值明显异常的记录。

选取河南省2017年1月—2020年12月的121个标准地面气象观测站的逐时观测数据(图1),包括

气温、空气相对湿度、降水量、风速、最大风速和能见度,均为自动观测数据。数据来源于河南省大气探测中心CIMISS系统。

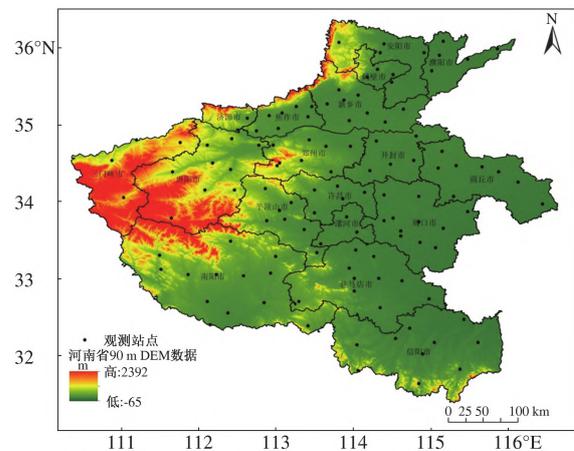


图1 河南省121个标准地面气象观测站分布图

1.3 研究方法

选取河南省2017年1月—2018年12月的PM_{2.5}浓度逐时观测数据与对应时次的葵花8卫星气溶胶光学厚度(AOD)、Angstrom指数(AE)、光学厚度精细比(RF)数据,以及气温(T)、空气相对湿度(U)、降水量(R)、风速(F)、最大风速(F_{\max})和能见度(V)等气象要素的观测数据,建立历史观测样本库。为了确保样本库记录的有效性,需要剔除无效及异常数据,例如葵花8卫星AOD数据产品只有白天、晴空区域的数值才是相对准确的,部分PM_{2.5}观测数据也存在缺测的情况。经过筛选,建立了有136790条记录的样本数据库。

在建立样本库和PM_{2.5}浓度反演计算过程中,与PM_{2.5}浓度相关的因子包括遥感参数和气象因子,其中气象因子为站点观测数据。由于PM_{2.5}观测站点和气象观测站点位置并不相同,因此需要将各气象因子数据进行空间插值,以便与PM_{2.5}观测值空间位置匹配,研究中采用了反距离权重法进行插值处理。由于遥感AOD数据产品分辨率为 0.05° ,为与PM观测值空间位置匹配,采用了双线性内插算法对AOD、AE和RF数据进行了插值。此外,通过对PM_{2.5}浓度和各相关因子的相关性研究显示,PM_{2.5}浓度和能见度的对数有显著的线性关系^[22-23],因此对能见度因子进行了对数处理,以 $\log(V)$ 表示。

本研究采用一种基于空间距离的机器学习方法,利用葵花8号遥感数据与地面气象观测数据,构

建 PM_{2.5} 浓度监测模型,引入遥感反演参数和气象要素等 9 个因子,首先计算空间距离,然后进行样本点选择,最后进行 PM_{2.5} 质量浓度的反演。

(1)空间距离计算

在样本库中,由逐时 PM_{2.5} 浓度和相关因子构成的记录可以用向量的形式来表示: $V_i = (X_{1i}, X_{2i}, \dots, X_{9i}, Y_i), i = 1, 2, \dots, n$ 。其中 X_1, X_2, \dots, X_9 , 对应 AOD、AE、RF、 T 、 U 、 R 、 F 、 F_{\max} 、 $\log(V)$ 等 9 个因子, Y_i 对应 PM_{2.5} 质量浓度。定义向量 V_i, V_j 之间的空间距离为

$$d_{ij} = \sum_{k=1}^9 a_k (X_{ki} - X_{kj})^2 \quad (1)$$

式中, d_{ij} 为 V_i, V_j 之间的空间距离; a_k 为第 k 个因子的距离权重系数,规定 a_k 的取值区间为 $[0, 1]$, X_{ki}, X_{kj} 为 V_i, V_j 中的各因子数值。

考虑到不同的因子数值单位和范围差异较大,在计算空间距离前,首先需要对各因子数值进行标准化处理。标准化处理公式为

$$x_i = (X_i - X_{i\min}) / (X_{i\max} - X_{i\min}) \quad (2)$$

其中, X_i, x_i 分别为第 i 个因子标准化前后的因子值, $X_{i\max}, X_{i\min}$ 为第 i 个因子样本数据的最大、最小值。

(2) 临近参考样本点选择

对于某一向量 $V_m (X_{1m}, X_{2m}, \dots, X_{9m}, Y_m)$, 其中 $X_{1m}, X_{2m}, \dots, X_{9m}$ 为已知量, Y_m 为未知量, 则可以根据公式(1) 在样本点库中找到若干与已知向量空间

距离相近的点, 然后可根据空间插值的方法计算 Y_m 。选择的相近点的数量对模型的运算速度和结果的精度都会有一定影响, 可以通过样本训练的方法来确定。

(3) PM_{2.5} 计算

若已知 n 个临近参考样本 V_1, V_2, \dots, V_n 和某个仅知 X 因子数值的 $V_m (X_{1m}, X_{2m}, \dots, X_{9m}, Y_m)$, 则其中的 Y_m 可以通过空间插值的方式计算得到。其计算公式为

$$Y_m = \frac{\sum_{k=1}^n \frac{Y_k}{d_{mk}}}{\sum_{k=1}^n \frac{1}{d_{mk}}} \quad (3)$$

式中, Y_k 为第 k 个样本点观测值, d_{mk} 为 V_m 与 V_k 之间的空间距离, Y_m 为待反演的 PM_{2.5} 的质量浓度。

在 PM 浓度计算式中, 需要针对 Y_m 分别确定各相关因子的距离权重系数 $a_k (k = 1, 2, \dots, 9)$ 和临近参考样本点数 n 。研究中利用 2017—2018 年的记录作为样本库, 2019 年的 PM_{2.5} 浓度与相关因子观测记录作为训练样本, 通过计算反演记录与观测记录的最小方差来确定各系数的最佳值。训练过程需要大量的计算, 需要在高性能计算机上进行。本文选用总样本库的一部分(1/17) 进行训练定参, 得到了 PM_{2.5} 浓度的权重系数 $a_1 \sim a_9$ 和参考样本点数 n (表 1)。

表 1 2017 年 1 月—2018 年 12 月河南省 PM_{2.5} 浓度的权重系数和参考样本点数

相关因子	AOD	AE	RF	T	U	R	F	F_{\max}	$\log(V)$	n
权重系数	0.49	0.11	0.10	0.10	0.49	0.31	0.11	0.10	0.83	20

2 结果与分析

2.1 PM_{2.5} 与遥感参数、气象因子的相关性分析

利用 2019 年河南省各环境监测站 PM_{2.5} 浓度观测数据及相应 AOD、AE、RF 等遥感及 T, U, R 等气象因子数据, 分析 PM_{2.5} 浓度与各因子之间的相关关系, 结果如表 2 所示 (R^2 为决定系数)。PM_{2.5} 浓度与 AOD、AE、RF、 T, U, F, F_{\max}, V 均有显著的相关关系, 其中 PM_{2.5} 浓度与 V, T, AOD 的相关性最为密切(均通过 0.001 的显著性水平检验)。此结果与成通^[24]、吕建华^[25]、于杰^[26] 等的研究结果一致。本研究基于此 3 个因子建立 PM_{2.5} 的多元回归方程, 用于与机器学习结果的对比(相关参数见表 3)。

表 2 2019 年 1—12 月河南省 PM_{2.5} 与遥感及气象因子相关性分析

因子	相关方程	R^2
AOD	$y = 18.201x + 36.265$	0.1056
AE	$y = 22.209x + 27.01$	0.0870
RF	$y = 46.551x + 24.556$	0.0868
T	$y = -2.3208x + 100.12$	0.2586
U	$y = 0.6809x + 20.205$	0.0930
R	$y = -1.2566x + 49.83$	0.0000
F	$y = -5.0702x + 63.631$	0.0292
F_{\max}	$y = -5.3953x + 67.602$	0.0388
$\log(V)$	$y = -113.43x + 514.8$	0.5729

注: 分析样本点数为 54572。

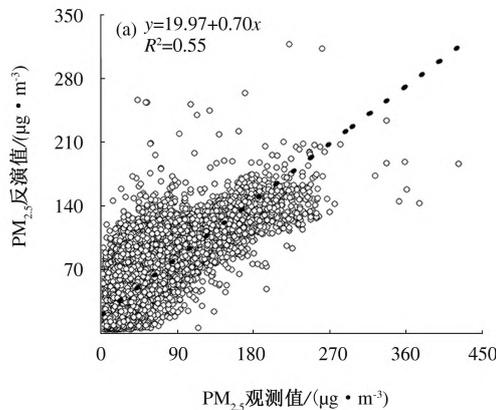
表 3 2019 年 1—12 月河南省基于多元回归的 PM_{2.5} 反演模型参数

因子	系数	标准误差	t-Stat
常数项	451.1573	1.9457	231.8741
AOD	5.0113	0.1627	30.7971
T	-0.9581	0.0140	-68.3569
log(V)	-93.7527	0.4933	-190.0594
回归统计	R ² = 0.745	SDE = 26.1590	n = 54572

注:t-Stat = 回归系数/系数标准误差,R² 为决定系数,SDE 为标准误差,n 为样本数。

2.2 PM_{2.5} 浓度反演精度验证

基于 2019 年训练数据、训练参数确定的模型,



对 2020 年遥感参数和气象因子样本记录(样本数 $n = 38682$),分别采用常规的多元回归法和基于空间距离的机器学习法进行 PM_{2.5} 反演,并与实际 PM_{2.5} 观测值进行对比检验,结果如图 2 所示。常规的多元回归法反演结果表明,相关系数 $R^2 = 0.55$,均方根误差为 $21.34 \mu\text{g} \cdot \text{m}^{-3}$;基于空间距离的机器学习法反演结果表明,相关系数 $R^2 = 0.73$,均方根误差为 $17.68 \mu\text{g} \cdot \text{m}^{-3}$ 。采用基于空间距离的机器学习法相较于常规的多元回归法反演的 PM_{2.5} 浓度的精度明显提高,其中决定系数 R^2 从 0.55 增加至 0.73,均方根误差 RMSE 由 $21.34 \mu\text{g} \cdot \text{m}^{-3}$ 下降至 $17.68 \mu\text{g} \cdot \text{m}^{-3}$ 。

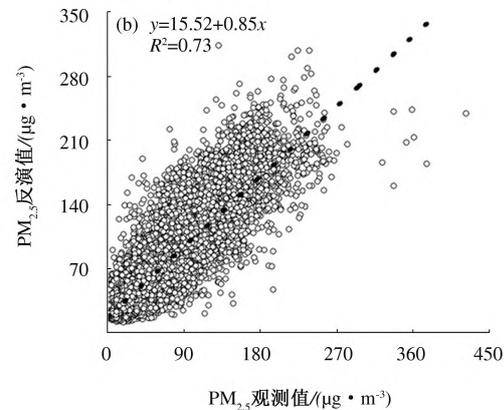


图 2 2020 年 1—12 月河南省 PM_{2.5} 浓度多元回归法 (a)、基于空间距离的机器学习法 (b) 的反演精度检验

2.3 PM_{2.5} 浓度反演值与观测值空间对比

为检验基于空间距离的机器学习法反演河南省 PM_{2.5} 浓度在空间上的反演效果,选取晴空条件下的 2020 年 12 月 27 日 14 时河南区域葵花 8 AOD 数据,并利用相应时刻河南省地面气象观测数据,采用基于空间距离的机器学习法,反演河南省 PM_{2.5} 浓度,并选取同时刻 PM_{2.5} 国控空气质量监测站实况监测数据,采用反距离权重法进行空间插值得到其空间分布。三者的空间分布如图 3 所示,三者在空间上的分布趋势基本一致,豫北新乡、安阳和濮阳等地的 AOD 及 PM_{2.5} 浓度较高,豫西、豫南部分区域的较低,整体呈现东北部的较高,西南部的较低分布。图 3(a) 仅利用葵花 8 AOD 数据只能获得晴空区域的监测数据,实际应用还会受到云的影响,经常出现缺测情况(空白区域);图 3(b) 利用站点观测数据插值获取的 PM_{2.5} 浓度空间分布,总体站点较为稀疏,利用插值法获取空间分布存在很大的局限性,无法进行较大空间范围的 PM_{2.5} 浓度的监测;图

3(c) 利用基于空间距离的机器学习法反演河南省 PM_{2.5} 浓度,不受天气的影响,可以弥补缺测值,在夜间也能提供 PM_{2.5} 浓度的反演数据,可以实现空间格点的观测,观测分辨率更高。

此外,为降低基于数据库直接反演 PM_{2.5} 浓度产生的误差,选取 2017—2020 年以外晴空条件下的数据,将反演结果与实况监测结果进行对比检验。根据河南省内 84 个国控空气质量监测站的站点位置,提取反演结果对应位置的像素值,进行对比检验,选取 2022 年 1—7 月共 13885 个样本,结果如图 4 所示。反演结果与实况监测结果的相关系数 R^2 为 0.75,均方根误差为 $15.84 \mu\text{g} \cdot \text{m}^{-3}$,反演精度较高。

此外,选取 2022 年 1 月 31 日 11 时河南区域葵花 8 AOD 数据、PM_{2.5} 监测浓度、PM_{2.5} 反演浓度(图 5),探讨了非建模数据反演的河南省 PM_{2.5} 浓度与观测值在空间分布上的关系。结果表明,基于空间距离的机器学习法反演的河南省 PM_{2.5} 浓度在空间分布上与观测值的一致,弥补了非晴空、插值法产生的误差,具有明显的优势。

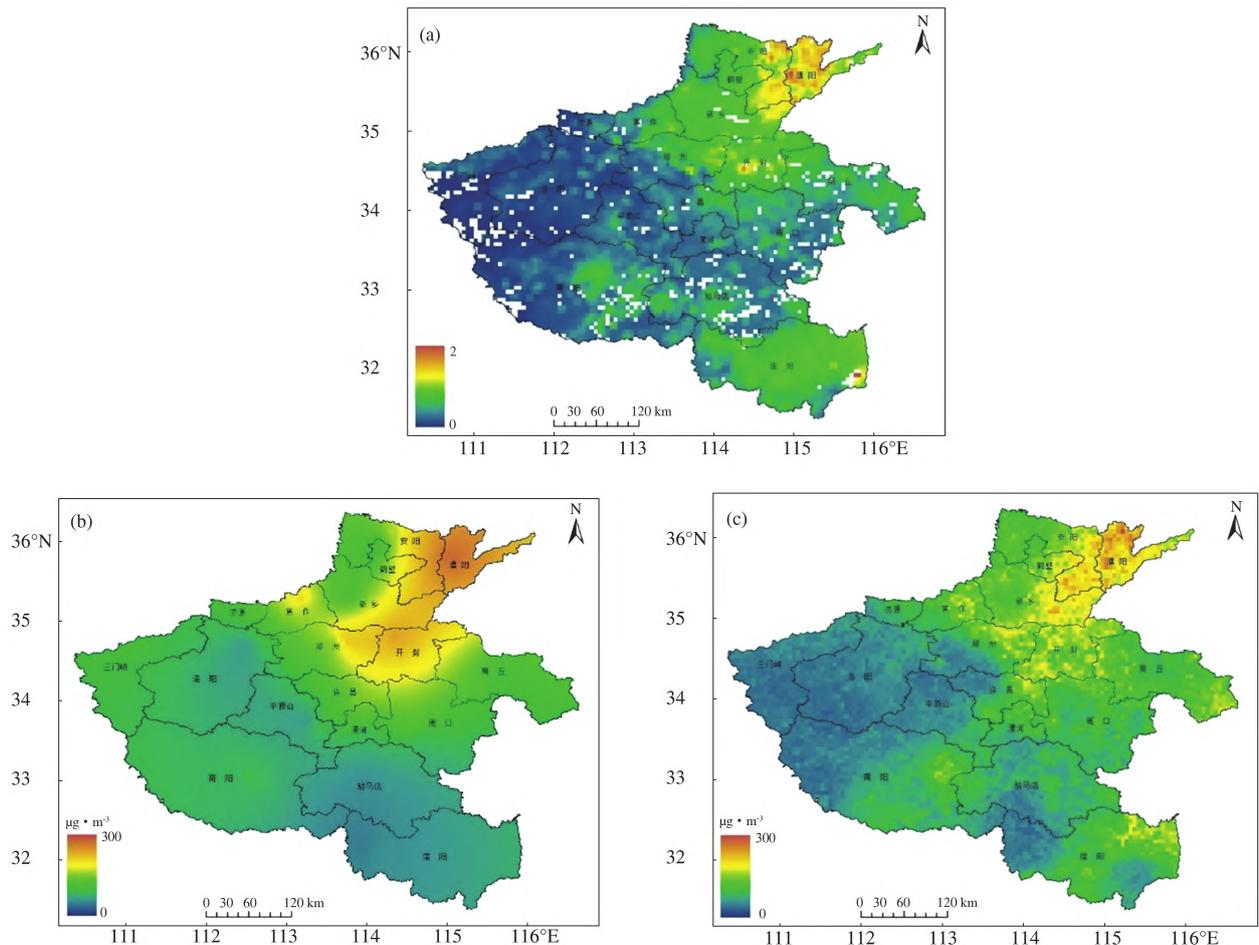


图 3 2020 年 12 月 27 日 14 时河南省 AOD (a)、PM_{2.5} 监测浓度 (b)、PM_{2.5} 反演浓度 (c) 空间分布

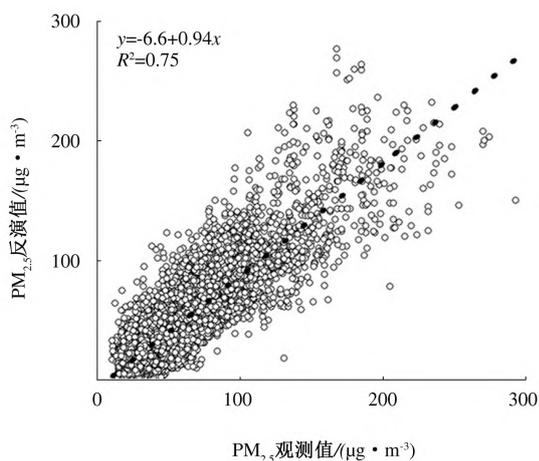


图 4 2022 年 1—7 月河南省 PM_{2.5} 浓度反演精度检验

3 讨论与结论

3.1 讨论

近年来,大气环境监测研究引起国内外学者越来越多的重视,相对于传统的地面监测手段,卫星遥

感技术以其独特的优势,已成为大气环境监测的重要手段。本文采用的卫星遥感数据为葵花 8 卫星数据,从时间分辨率及空间分辨率上,比 Engle^[11]、高星星^[14]等基于 MODIS AOD 反演 PM_{2.5} 浓度,有明显的优势,可以对研究区域开展长时间、高频次的监测。郭建平等^[7]基于 BP 网络框架利用 MODIS 气溶胶光学厚度产品估算中国东部的 PM_{2.5} 浓度,以 5 个大气成分站的 PM_{2.5} 观测数据进行地面验证,其中反演效果最好的庐山站的决定系数 $R^2 = 0.68$ 。wei 等^[16]基于 MODIS AOD 采用机器学习方法,发展了时空-极端随机树 (STET) 模型,重构了中国近 20 年 1 km 分辨率的 PM_{2.5} 浓度历史数据集,从时间尺度上模型能够较准确估算中国不同年份的 PM_{2.5} 日浓度。张猛等^[15]基于 Landsat 8 OLI 卫星遥感影像数据,采用机器学习的方法,反演了北京地区的 PM_{2.5} 浓度,具有较高的空间分辨率 (30 m),但在中国地区的重访周期为 16 天,不能支持 PM_{2.5} 浓度的长时间动态监测。上述研究结果表明,基于卫星遥感技术和机器学习方法可以较好地反演 PM_{2.5} 浓度。

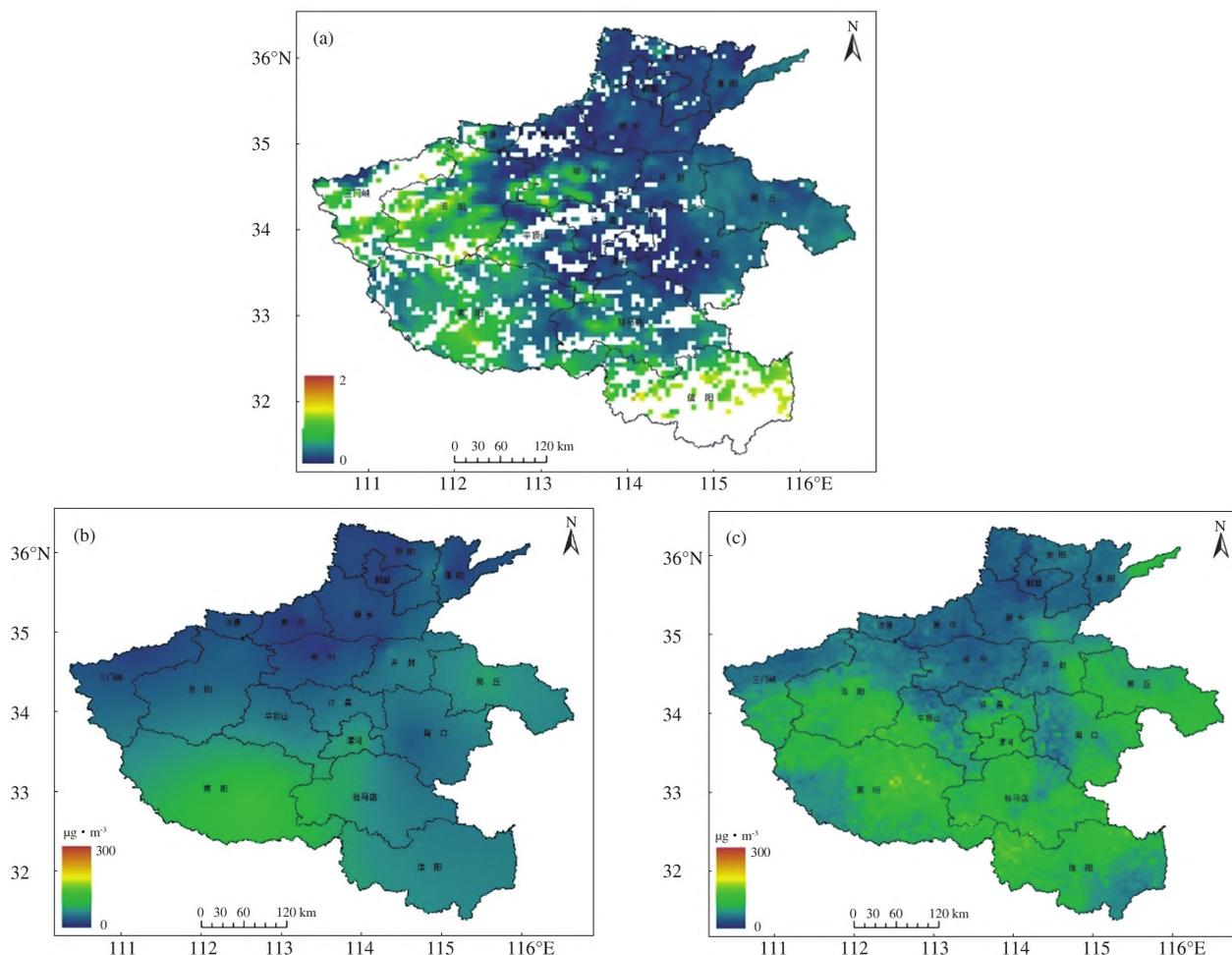


图5 2022年1月31日11时河南省AOD(a)、PM_{2.5}监测浓度(b)、PM_{2.5}反演浓度(c)空间分布

葵花8静止卫星数据具有高精度、高频次对地观测的优势,非常适合用于区域大气环境的动态监测。葛邦宇等^[19]采用暗目标法对2016年5月02:00—08:00葵花8静止卫星陆地气溶胶光学厚度进行反演,通过与地基观测网AERONET 5个站点实测对比,所有站点相关系数 R^2 均大于0.88,但是存在地面验证站点较少且地表类型缺少多样化、葵花8数据积累不足等问题。

本研究提出了一种基于空间距离的机器学习方法构建模型来反演河南省PM_{2.5}浓度,在建立样本数据库时充分考虑了样本的数量与多样性,其中包括2017—2020年每日逐时的葵花8卫星遥感反演AOD数据、84个国控空气质量监测站逐时地面观测数据、河南省121个标准地面气象观测站的逐时观测数据,比前人的研究有明显的优势,研究结果更具可靠性。基于样本库构建了基于空间距离的机器学习法的PM_{2.5}浓度反演模型,通过对PM_{2.5}浓度和遥感参数、气象因子的相关性分析,筛选、确立了反

演模型的最优化输入参数,在河南省取得了较为理想的反演效果($R^2=0.73$)。

3.2 结论

(1)PM_{2.5}浓度与遥感参数AOD、AE、RF及气象因子 T 、 U 、 F 、 F_{\max} 、 V 均有显著的相关关系(通过0.01的显著性水平检验),其中与AOD、 T 、 V 的相关性最为密切。

(2)基于空间距离的机器学习方法构建的监测模型,具有很强地揭示可吸入颗粒物PM_{2.5}浓度的能力,根据PM_{2.5}浓度反演值与观测值的散点回归分析,PM_{2.5}浓度反演值与观测值相关关系很好($R^2=0.73$),相较于常规的多元回归模型模拟结果($R^2=0.55$),有非常明显的提高。

(3)从PM_{2.5}浓度反演值与观测值的空间分布图来看,基于空间距离的机器学习方法构建模型反演PM_{2.5}浓度实现了空间格点的小时监测,弥补了非晴空及夜间监测数据的缺失,在空间分布上与观测值一致,减小了基于观测数据插值法产生的空间

分布上的误差,具有很高的精确性和可靠性。

(4)对地面站点观测 $PM_{2.5}$ 数据与基于空间距离的机器学习方法构建模型反演的 $PM_{2.5}$ 值进行散点拟合和空间分布对比分析结果表明,该方法估算的 $PM_{2.5}$ 浓度精度较高,具有推广到大范围空气污染监测的现实价值,同时基于空间距离的机器学习方法参数配置会随着样本库的增大,进一步提升 $PM_{2.5}$ 浓度估算精度。该方法有望成为大气污染变化监测与区域分析的一个新的重要手段,同时也为利用其他卫星遥感影像数据进行 PM 反演运算提供了有效的借鉴与参考。

参考文献

- [1] 田宏伟,师丽魁,李梦夏. 两种地面 $PM_{2.5}$ 质量浓度遥感反演方法适用性比较[J]. 气象与环境科学, 2020, 43(3): 59-64.
- [2] 李成才,毛节泰,刘启汉. 利用 MODIS 遥感大气气溶胶及气溶胶产品的应用[J]. 北京大学学报(自然科学版), 2003, 39(1): 108-117.
- [3] 郝明途,林天佳,刘焱. 我国 $PM_{2.5}$ 的污染状况和污染特征[J]. 环境科学与管理, 2006, 31(2): 58-61, 67.
- [4] 孙志豪,崔燕平. $PM_{2.5}$ 对人体健康影响研究概述[J]. 环境科技, 2013, 26(4): 75-78.
- [5] Laden F, Neas L M, Dockery D W, et al. Association of Fine Particulate Matter from Different Sources with Daily Mortality in Six U. S. Cities. [J]. Environmental Health Perspectives, 2000, 108(10): 941-947.
- [6] 王桥. 中国环境遥感监测技术进展及若干前沿问题[J]. 遥感学报, 2021, 25(1): 25-36.
- [7] 郭建平,吴业荣,张小曳,等. BP 网络框架下 MODIS 气溶胶光学厚度产品估算中国东部 $PM_{2.5}$ [J]. 环境科学, 2013, 34(3): 817-825.
- [8] Hidy G M, Brook J R, Chow J C, et al. Remote sensing of particulate pollution from space: Have we reached the promised land? [J]. Journal of the Air & Waste Management Association, 2009, 59(10): 1130-1139.
- [9] Tao J H, Zhang M G, Chen L F, et al. A method to estimate concentrations of surface-level particulate matter using satellite-based aerosol optical thickness [J]. Science China Earth Science, 2013, 56(8): 1422-1433.
- [10] Li C C, Mao J T, Lau K-HA, et al. Characteristics of distribution and seasonal variation of aerosol optical depth in eastern China with MODIS products [J]. Chinese Science Bulletin, 2003, 48(22): 2488-2495.
- [11] Engle-Cox J A, Yong C H, Coutant B W, et al. Qualitative and quantitative evaluation of MODIS satellite sensor data for regional and urban scale air quality [J]. Atmospheric Environment, 2004, 38(16): 2495-2509.
- [12] Wang J, Christopher A. Intercomparison between satellitederived aerosol optical thickness and $PM_{2.5}$ mass: Implications for air quality studies [J]. Geophysica Research Letters, 2003, 30(21): 1-4.
- [13] Liu Y, Park R J, Jacob D, et al. Mapping annual mean ground-level $PM_{2.5}$ concentrations using multiangle imaging spectro radiometer aerosol optical thickness over the contiguous united states [J]. Journal of Geophysical Research, 2004, 109(22): D22206.
- [14] 吴序鹏,刘端阳,谢真珍,等. 江苏淮安地区大气污染变化特征及其与气象条件的关系 [J]. 气象与环境科学, 2018, 41(1): 31-38.
- [15] 张猛,张博. 基于机器学习和卫星遥感的 $PM_{2.5/10}$ 空间连续分布反演方法研究 [J]. 地球环境学报, 2020, 11(4): 447-455.
- [16] Wei J, Li Z Q, Al C, et al. Reconstructing 1-km-resolution high-quality $PM_{2.5}$ data records from 2000 to 2018 in China: spatiotemporal variations and policy implications [J]. Remote Sensing of Environment, 2021, 252(1): 1-17.
- [17] 李鹏,毕温凯,杨宁,等. 基于葵花-8 号气象卫星的天津地区气溶胶光学厚度分布特征 [J]. 天津科技, 2021, 48(8): 96-98.
- [18] 牛晓君,唐家奎,张自力,等. 基于葵花 8 号卫星数据的气溶胶反演算法及其在雾霾过程监测中的应用 [J]. 中国科学院大学学报, 2019, 36(5): 671-681.
- [19] 葛邦宇,杨磊库,陈兴峰,等. 暗目标法的 Himawari-8 静止卫星数据气溶胶反演 [J]. 遥感学报, 2018, 22(1): 38-50.
- [20] 吴健生,王茜. 基于 AOD 数据反演地面 $PM_{2.5}$ 浓度研究进展 [J]. 环境科学与技术, 2017, 40(8): 68-76.
- [21] 杜彦彦,黄青. 河南省 $PM_{2.5}$ 时空分布特征及其与植被覆盖度的关系 [J]. 生态环境学报, 2019, 28(11): 2257-2265.
- [22] 王京丽,刘旭林. 北京市大气细粒子质量浓度与能见度定量关系初探 [J]. 气象学报, 2006, 64(2): 221-228.
- [23] 王继康,张恒德,桂海林,等. 能见度与 $PM_{2.5}$ 浓度关系及其分布特征 [J]. 环境科学, 2019, 40(7): 2985-2993.
- [24] 成通,祝善友,张桂欣,等. 2018 年京津冀地区 $PM_{2.5}$ 季节变化及其与地表温度的关系分析 [J]. 遥感技术与应用, 2020, 35(6): 1457-1466.
- [25] 吕建华,彭岩波,谢刚. 2013 年济南市大气能见度与相对湿度和 PM_{10} 、 $PM_{2.5}$ 的关系 [J]. 气象与环境科学, 2016, 39(4): 93-97.
- [26] 于杰,车慧正,陈权亮,等. 2010-2012 年我国西北地区沙尘个例气溶胶特征分析 [J]. 气象与环境科学, 2016, 39(2): 33-40.

A Method for PM_{2.5} Concentration Monitoring Based on Machine Learning and Himawari-8 Satellite Data

Ye Haotian^{1,2,3}, Zou Chunhui^{1,2}, Tian Hongwei^{1,2}

(1. CMA · Henan Key Laboratory of Agrometeorological Support and Applied Technique, Zhengzhou 450003, China;

2. Henan Institute of Meteorological Sciences, Zhengzhou 450003, China;

3. Anyang National Climate Observatory, Anyang 455000, China)

Abstract: Modeling based on satellite remote sensing data and meteorological data is an important method for PM_{2.5} concentration monitoring. Based on the AOD related products of Himawari-8, meteorological observation data, hourly observation data of PM_{2.5} concentration in Henan from 2017 to 2020, we established a sample database in this research, explored the correlation between the PM_{2.5} concentration and the remote sensing retrieval parameters and meteorological factors, proposed a machine learning method based on spatial distance, constructed a PM_{2.5} monitoring model combining satellite remote sensing and meteorological data, and used PM_{2.5} observation data in verifying the scattered point fitting and spatial distribution of the model. The results show that PM_{2.5} has a very significant correlation with visibility (V), air temperature (T) and aerosol optical thickness (AOD). From the scatter regression analysis of the PM_{2.5} concentration retrieval and observation, the correlation between PM_{2.5} retrieval and observation is very good ($R^2 = 0.73$). From the spatial distributions of PM_{2.5} retrieval and observation, the retrieved PM_{2.5} has realized the hourly monitoring of spatial grid points, making up for the lack of non-clear sky and night monitoring data, which is consistent with the observed values in spatial distribution, so its accuracy and reliability are very high. Therefore, the PM_{2.5} monitoring model constructed by the machine learning method based on spatial distance could be well applied to PM_{2.5} monitoring, which is a new idea and method for PM_{2.5} monitoring.

Key words: PM_{2.5} concentration; spatial distance; machine learning; Himawari-8