

引用格式: Yang Lijuan, Zhang Jianxia, Lin Musheng. Research on Methods of Remotely Sensed $PM_{2.5}$ Concentrations Estimation in Four Provinces and One City along the East Coast of China[J]. Remote Sensing Technology and Application, 2021, 36(6): 1408-1415. [杨立娟, 张建霞, 林木生. 中国东部沿海四省一市 $PM_{2.5}$ 浓度遥感估算方法研究[J]. 遥感技术与应用, 2021, 36(6): 1408-1415.]
DOI: 10.11873/j.issn.1004-0323.2021.6.1408

中国东部沿海四省一市 $PM_{2.5}$ 浓度遥感估算方法研究

杨立娟, 张建霞, 林木生

(闽江学院 测绘工程系, 福建 福州 350018)

摘要: 卫星遥感反演的气溶胶光学深度(AOD)产品已被广泛应用于近地面 $PM_{2.5}$ 浓度的估算。已有研究表明通过构建AOD和 $PM_{2.5}$ 之间的高级统计模型—线性混合效应模型(LME)可以有效获取近地面 $PM_{2.5}$ 浓度的空间分布, 但由于引入了大量的气象和土地利用等因子, 使得模型对变量的解译能力有所降低。为此, 基于MODIS AOD(空间分辨率: 3 km), 以我国东部长江三角洲—福建—广东(YRD-FJ-GD)为研究区, 构建了两种非参数机器学习模型, 即支持向量机(SVM)和随机森林(RF)模型, 来估算2018年YRD-FJ-GD地区的近地面 $PM_{2.5}$ 浓度, 并将其与线性混合效应模型(LME)的估算结果进行对比。研究发现, 3种模型估算的 $PM_{2.5}$ 浓度与地面实测值之间的 R^2 均高于0.6, 其中, RF模型的估算精度最优, 模型拟合的 R^2 高达0.91, 比SVM模型($R^2=0.79$)和LME模型($R^2=0.64$)的估算结果分别提高了13%和30%; 且RMSE($\sim 9.07 \mu g/m^3$)也远低于LME($\sim 19.09 \mu g/m^3$)和SVM模型($\sim 17.29 \mu g/m^3$)。此外, 由随机森林(RF)模型估算的2018年YRD-FJ-GD地区的 $PM_{2.5}$ 空间分布显示, 长江三角洲(YRD)地区的年均 $PM_{2.5}$ 浓度最高($>46 \mu g/m^3$), 其次为广东省(GD), 福建地区(FJ)的年均 $PM_{2.5}$ 浓度最低($<37 \mu g/m^3$); 4个季节的平均 $PM_{2.5}$ 浓度则呈现冬季($46.32 \mu g/m^3$) $>$ 春季($38.80 \mu g/m^3$) $>$ 秋季($36.15 \mu g/m^3$) $>$ 夏季($30.16 \mu g/m^3$)的分布格局。研究结果表明: 与高级统计模型(LME)和机器学习(SVM)相比, 随机森林(RF)模型能更好地应用于YRD-FJ-GD地区的 $PM_{2.5}$ 浓度估算。

关键词: LME; SVM; RF; $PM_{2.5}$ 估算; YRD-FJ-GD

中图分类号: TP79 **文献标志码:** A **文章编号:** 1004-0323(2021)06-1408-08

1 引言

随着我国经济迅速发展和城市不断扩张, 空气污染已成为我国亟需解决的重要环境问题^[1-2]。研究表明, 悬浮在空气中且空气动力学直径小于 $2.5 \mu m$ 的细颗粒物(即: $PM_{2.5}$)不仅对大气有消光作用, 而且与人体负面健康效应相关^[3-6]。因此, 有必要对我国城市地区进行近地面 $PM_{2.5}$ 浓度的有效监测。传统的地面环境监测网络可以提供准确的空间和时间数据, 但是却无法获取连续的 $PM_{2.5}$ 浓度空间分布^[7-8]。卫星遥感反演的气溶胶光学深度(Aerosol

Optical Depth, AOD)产品已被广泛用于全球范围的 $PM_{2.5}$ 估算^[9-11]。

已有研究主要通过构建AOD和 $PM_{2.5}$ 之间的简单线性回归模型和高级统计模型来估算近地面的 $PM_{2.5}$ 浓度。由于AOD和 $PM_{2.5}$ 的相关性受大气边界层高度和空气湿度的影响, 因此, 只包含AOD的简单线性模型其估算精度较低^[12-13]。和简单线性模型相比, 高级统计模型引入了更多的辅助参数(如: 气象和土地利用等), 从而使得模型估算精度大大提高。例如, Liu等^[14]开发了两层的广义加和模型

收稿日期: 2020-09-15; 修订日期: 2021-11-06

基金项目: 闽江学院优秀引进人才科研启动项目(MJY20001), 福建省自然科学基金项目(2021J05204)。

作者简介: 杨立娟(1985—), 女, 福建三明人, 博士, 副教授, 主要从事遥感技术与应用研究。E-mail: subrinazhong@aliyun.com

(Generalized Additive Model, GAM),并将其运用在美国东北部 Massachusetts 地区的 PM_{2.5}反演,结果表明由 GAM 估算的 PM_{2.5}浓度与地面观测值之间的 R^2 为 0.79。Lee 等^[15]提出了日校正的线性混合效应模型(Linear Mixed Effects model, LME)来估算美国东北部地区的 PM_{2.5}浓度,结果表明模型估算值和地面实测值高度相关,二者之间的 R^2 高达 0.92。Hu 等^[16]以美国东南部为研究区,开发了包含多变量的地理加权回归模型(Geographically Weighted Regression, GWR),结果表明模型拟合的 R^2 为 0.64。He 等^[17]在此基础上考虑了时间变化对 AOD-PM_{2.5}关系的影响,并构建了地理和时间加权回归模型(Geographically and Temporal Weighted Regression, GTWR)来估算我国近地面 PM_{2.5}浓度,结果表明 GTWR 模型可以解释我国 80% 的 PM_{2.5}浓度变异。张莹和李正强^[18]通过计算细颗粒物光学厚度占总光学厚度的比例(气溶胶细模态比例)来建立气溶胶细模态光学厚度 AOD_f与 PM_{2.5}的线性回归关系,结果表明该方法能够有效地估算灰霾期间的 PM_{2.5}浓度,其模型模拟的 R^2 可达 0.77。此外,也有研究者采用土地利用回归模型(Land Use Regression, LUR)^[19]和时间结构自适应模型(Timely Structure Adaptive Modeling, TSAM)^[20]来估算区域的 PM_{2.5}浓度,这些模型也获得了较高的估算精度($R^2 > 0.8$)。

上述研究均表明引入气象和土地利用等辅助因子的高级统计模型可以较好地应用于区域 PM_{2.5}浓度的估算,但同时也有研究表明辅助因子的增加将降低模型对变量的解译能力^[21]。与高级统计模型相比,机器学习能更好地处理高维数据集,且其泛化能力较强,为此,不少研究者开始探索利用机器学习算法来估算区域的 PM_{2.5}浓度。例如,Gupta 和 Christopher^[22]构建了 AOD 和 PM_{2.5}之间的人工神经网络(Artificial Neural Network, ANN)框架,结果表明由 ANN 估算的 PM_{2.5}浓度与地面观测值之间的 R^2 为 0.61。Vahid^[23]比较了 3 种机器学习算法(决策树(Decision Tree, DT),贝叶斯网络(Bayesian Network, BN)和支持向量机(Support Vector Machine, SVM)在伊朗德黑兰地区的 PM_{2.5}估算能力,结果表明 SVM 的估算精度优于 DT 和 BN。与 ANN 和 SVM 等算法相比,随机森林(Random Forest, RF)不仅能够处理高维度数据,且在建模时无需对数据进行规范化和特征选择,同时,随机森林针对多数据样本时可有效地减小预测误差,并且提

供了单个特征变量对模型性能的重要性指标,因此,RF 算法已被广泛应用于各研究领域。少数研究者尝试利用 RF 算法来估算近地面 PM_{2.5}浓度,结果表明 RF 算法具有较高的 PM_{2.5}估算能力。例如,Hu 等^[24]引入了 AOD 和 39 个气象及土地利用参数,构建了 RF 模型来估算美国的 PM_{2.5}浓度,结果表明 RF 模型拟合的 R^2 为 0.80。Cole 等^[25]以美国 Ohio 州为研究区,基于 AOD 和 11 个辅助气象参数构建了 RF 模型来估算该研究区的 PM_{2.5}浓度,结果表明模型估算值和地面实测值之间的相关性高达 0.91。目前,利用 RF 模型来估算近地面 PM_{2.5}浓度的研究还较少。

鉴于高级统计模型和机器学习均能较好地估算近地面 PM_{2.5}浓度,本研究以我国东部连续区域,即长江三角洲、福建省和广东省(YRD-FJ-GD)为研究区,构建了基于 AOD、气象和土地利用等参数的线性混合效应模型(LME)、支持向量机(SVM)和随机森林(RF)模型,并对比了 3 种模型在 YRD-FJ-GD 地区的 PM_{2.5}估算能力。在此基础上,通过选择估算精度最高的模型来估算 2018 年 YRD-FJ-GD 地区的 PM_{2.5}浓度空间分布。

2 数据和方法

2.1 研究区

以我国东部连续区域,即长江三角洲、福建省和广东省(YRD-FJ-GD)为研究区,整个研究区涵盖了 55 个城市,其中,长江三角洲包含了上海直辖市、江苏省和浙江省共 25 个城市,福建和广东省分别包含了 9 个和 21 个城市(图 1)。长江三角洲和珠江三角洲(包含广东省、香港和澳门)是我国经济最发达的两个城市群,随着近十几年经济的迅速发展和城市的急剧扩张,长江三角洲和珠江三角洲的空气污染也成为了首要的环境问题。资料显示,长江三角洲和珠江三角洲近几年的年均 PM_{2.5}浓度分别为 67 $\mu\text{g}/\text{m}^3$ 和 55 $\mu\text{g}/\text{m}^3$,超过了我国《环境空气质量标准》(GB3095-2012)的二级标准($\sim 35 \mu\text{g}/\text{m}^3$)。而福建省位于长江三角洲和珠江三角洲之间,区域内以山区和丘陵地区为主,气候温暖,降水丰富,近几年的年均 PM_{2.5}浓度为 35 $\mu\text{g}/\text{m}^3$,远低于长江三角洲和珠江三角洲的年均值。

2.2 数据

2.2.1 PM_{2.5}和卫星数据

实验采用的 PM_{2.5}浓度数据来源于 4 个省环境

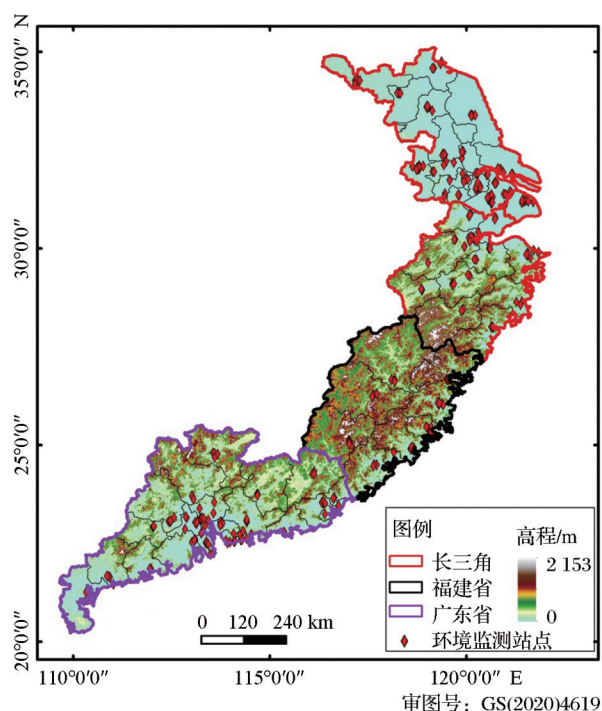


图1 研究区(YRD-FJ-GD)

Fig.1 The study area (YRD-FJ-GD)

保护厅和上海市环境保护局发布的2018年共297个环境监测站点的 $PM_{2.5}$ 浓度。卫星数据主要来源于搭载在Terra星上的中分辨率成像光谱仪(Moderate Resolution Imaging Spectroradiometer, MODIS)提供的空间分辨率为3 km的AOD数据(以下简称MODIS 3 km AOD)。MODIS 3 km AOD产品主要采用暗像元算法进行反演,其数据可在美国国家航空航天局(NASA)网站下载(<https://ladsweb.modaps.eosdis.nasa.gov/search/>)。本研究主要利用IDL语言批量提取 $0.55\ \mu m$ 波段处的AOD值。

2.2.2 气象数据

实验的气象数据来源于戈达德地球观测系统数据同化系统(GEOS)提供的前向处理数据,其空间分辨率为 0.25° (纬度) $\times 0.3125^\circ$ (经度)。共下载了包含大气边界层高度(PBLH, m),地表温度(TS, K), 2 m气温(T2M, K), 10 m气温(T10M, K), 10 m东向风(U10M, m/s), 10 m北向风(V10M, m/s),相对湿度(RH, %), 风的纬向分量(U-component, m/s), 经向分量(V-component, m/s)以及气压(PS, hPa)等10个气象场数据。采用反距离权重插值法(Inverse Distance Weighted, IDW)将所有气象场数据插值到 $3\ km \times 3\ km$ 网格中,再提取覆盖整个研究区的气象数据,本研究的插值和提取过程在Python和IDL中完成。

2.2.3 土地利用数据

实验的土地利用数据包含植被覆盖度和道路密度数据。其中,植被覆盖度数据主要来源于空间分辨率为1 km的MODIS月植被指数产品(MOD13A3),为了便于后期数据建模,采用最临近插值法将植被指数产品重采样成 $3\ km \times 3\ km$ (与AOD的空间分辨率相同);道路矢量数据主要来源于百度地图,共获取了研究区包含高速、国道、省道和县道等在内的道路矢量文件。道路密度计算主要在ArcGIS中完成,首先在ArcGIS中生成覆盖研究区的 $3\ km \times 3\ km$ 网格,然后将道路矢量数据与该网格进行叠加,并计算每个网格的道路矢量长度与该网格面积的比值(即:道路密度)。

2.3 数据建模和验证

实验构建了基于AOD、气象和土地利用等参数的线性混合效应模型(LME)、支持向量机(SVM)和随机森林(RF)模型来估算YRD-FJ-GD地区的 $PM_{2.5}$ 浓度。图2给出了实验的技术框图。线性混合效应模型(LME)是一种高级统计模型,主要由固定效应部分和随机效应部分组成,其中,固定效应表示的是所有天数和所有站点的各变量与 $PM_{2.5}$ 之间的关系,而随机效应表示的是每天所有站点的各变量和 $PM_{2.5}$ 浓度之间的关系^[15]。支持向量机(Support Vector Machine, SVM)是一种基于数据的机器学习方法,它通过核函数将数据样本映射到更高维的空间里,然后构造类间边缘(Margin)的超平面(Hyperplane)以达到分类和回归的目的^[23]。本研究采用径向基函数核(Radial Basis Function Kernel, RBF kernel)为SVM的核函数,并通过计算最优的惩罚系数(cost)和RBF核的宽度(gamma)来对样本进行训练。

随机森林是一种集成的机器学习方法,它采用bagging思想和Bootstrap重抽样技术,从原始样本N中有放回地抽取N个样本组成新的训练样本集(即:随机子样本),然后基于随机子集特征选择方法为每个子样本选择一个预测子集,并将所有决策时输出的均值来作为最终的输出结果^[26]。随机森林通过确定每个节点的预测变量数(mtry)和每个随机森林(ntree)中决策树的数目来获取最优模型,且模型中的指标Mean Decrease Accuracy (IncMSE)可用于检查每个变量在 $PM_{2.5}$ 浓度变异中的重要性。上述3个模型可用以下公式简单表示:

$$\begin{aligned} \text{PM}_{2.5ij} = & (\alpha + u_j) + (\beta + v_j) \text{AOD}_{ij} + (\gamma_1 + x_{1j}) \text{PBLH}_{ij} + \\ & (\gamma_2 + x_{2j}) \text{TS}_{ij} + (\gamma_3 + x_{3j}) \text{T2M}_{ij} + (\gamma_4 + \\ & x_{4j}) \text{T10M}_{ij} + (\gamma_5 + x_{5j}) \text{U10M}_{ij} + (\gamma_6 + \\ & x_{6j}) \text{V10M}_{ij} + (\gamma_7 + x_{7j}) \text{RH}_{ij} + (\gamma_8 + \\ & x_{8j}) \text{U}_{\text{component}_{ij}} + (\gamma_9 + x_{9j}) \text{V}_{\text{component}_{ij}} + (\gamma_{10} + \\ & x_{10j}) \text{PS}_{ij} + \text{NDVI}_i + \text{RD}_i + \varepsilon_{ij} \end{aligned} \quad (1)$$

$$\text{PM}_{2.5} = \text{SVM}(\text{AOD}, \text{PBLH}, \text{TS}, \text{T2M}, \text{T10M}, \text{U10M}, \text{V10M}, \text{RH}, \text{Ucomponent}, \text{Vcomponent}, \text{PS}, \text{NDVI}, \text{RD}) \quad (2)$$

$$\text{PM}_{2.5} = \text{RF}(\text{AOD}, \text{PBLH}, \text{TS}, \text{T2M}, \text{T10M}, \text{U10M}, \text{V10M}, \text{RH}, \text{Ucomponent}, \text{Vcomponent}, \text{PS}, \text{NDVI}, \text{RD}) \quad (3)$$

其中:PM_{2.5ij}为第*i*个站点第*j*天的实测PM_{2.5}浓度;AOD_{ij}为第*i*个站点第*j*天的AOD值。PBLH_{ij}、TS_{ij}、T2M_{ij}、T10M_{ij}、U10M_{ij}、V10M_{ij}、RH_{ij}、Ucomponent_{ij}、Vcomponent_{ij}和PS_{ij}分别为上述对应的气象因子在第*i*个站点第*j*天的值;NDVI_i和RD_i分别为第*i*个站点的植被指数和道路密度值。ε_{ij}为第*i*个站点第*j*天

的误差项。

实验采用广泛使用的10折交叉验证方法(10-fold Cross Validation, CV)来评价3个模型在YRD-FJ-GD地区的PM_{2.5}估算能力。10折交叉验证法即将建模数据随机分为10份,选择其中9份用于模型训练,剩余1份用于预测,最终将10次验证的平均结果作为模型的估算精度。同时采用表征拟合度的决定系数(*R*²)和均方根误差(RMSE)来评价模型的估算值和实测值之间的拟合度和误差情况^[27-28]。实验的模型拟合和验证均在R-studio中完成。*R*²、RMSE的计算如下式所示:

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \quad (4)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (\hat{Y}_i - Y_i)^2}{n}} \quad (5)$$

其中: \hat{Y}_i 为模型预测的PM_{2.5}值; \bar{Y} 为地面PM_{2.5}浓度均值;*Y_i*为地面PM_{2.5}监测值;*n*为建模数据集数据总数。

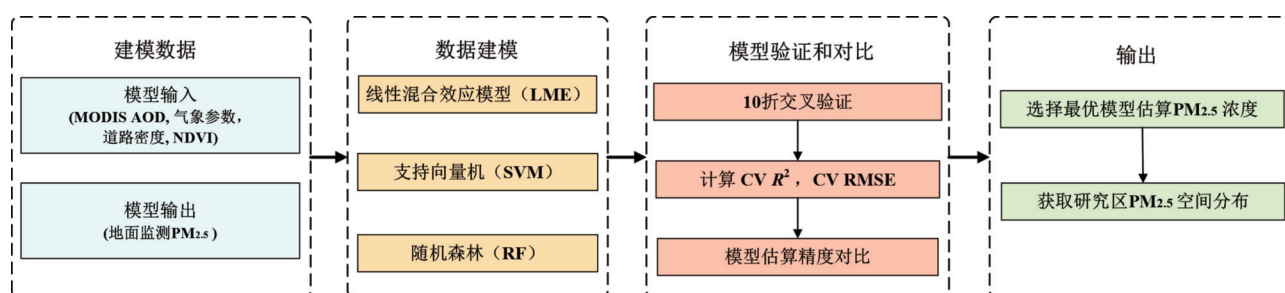


图2 技术路线框图

Fig.2 The framework of this study

3 结果与分析

3.1 模型拟合和验证

由PM_{2.5}时间序列值分析可知,2018年YRD-FJ-GD地区的PM_{2.5}浓度范围为1~377 μg/m³,各区域的PM_{2.5}年均浓度总体表现为YRD>GD>FJ。AOD与PM_{2.5}年均浓度呈现一致的区域分布,其中YRD地区的气溶胶分布为3个地区最高,其AOD值为2.19。分季节来看,3个地区的AOD与PM_{2.5}浓度均表现出冬春两季高于夏秋两季,这与该地区的气候条件相关。

实验利用14个变量分别构建了线性混合效应模型(LME),支持向量机回归模型(SVM)和随机森林模型(RF)来估算2018年YRD-FJ-GD地区的PM_{2.5}浓度。图3(a)~3(c)给出了3种模型的拟合结

果,结果显示由LME和SVM模型估算的PM_{2.5}浓度与地面实测值之间的*R*²分别为0.64和0.79, RMSE分别为19.09 μg/m³和17.29 μg/m³。3种模型中随机森林模型(RF)的估算精度最高,模型拟合的*R*²高达0.91,比LME和SVM模型的精度分别提高了30%和13%;RMSE也降低至9.07 μg/m³。3种模型的十折交叉验证(CV)结果(如图3(d)~图3(f))也表明RF模型的估算精度(CV *R*²=0.87; CV RMSE=12.09 μg/m³)高于LME(CV *R*²=0.62; CV RMSE=19.25 μg/m³)和SVM模型(CV *R*²=0.71; CV RMSE=18.47 μg/m³)。由于本研究引入了13个自变量来估算YRD-FJ-GD地区的PM_{2.5}浓度,且这些自变量的量纲均不同,因此在进行LME和SVM训练时,均需对数据进行规范化。而随机森林(RF)采用了Bagging思想和Bootstrap重抽样技术对数据

样本进行抽样并构建多棵决策树,它能够随机选择决策树节点划分特征,且当样本特征维度较高时,无需进行数据规范化和特征选择。与SVM相比,RF可有效地减小预测误差,并在一定程度上避免了过拟合。

此外,实验还利用随机森林中的 MeanDe-

creaseAccuracy (IncMSE) 指标来获取各变量对YRD-FJ-GD地区 $PM_{2.5}$ 浓度变异的重要性,当IncMSE值越高,表明该变量对 $PM_{2.5}$ 浓度的解译能力越强。研究结果表明,PBLH,AOD和RH是解释YRD-FJ-GD区域中 $PM_{2.5}$ 浓度变化的前3个最重要变量(图4)。

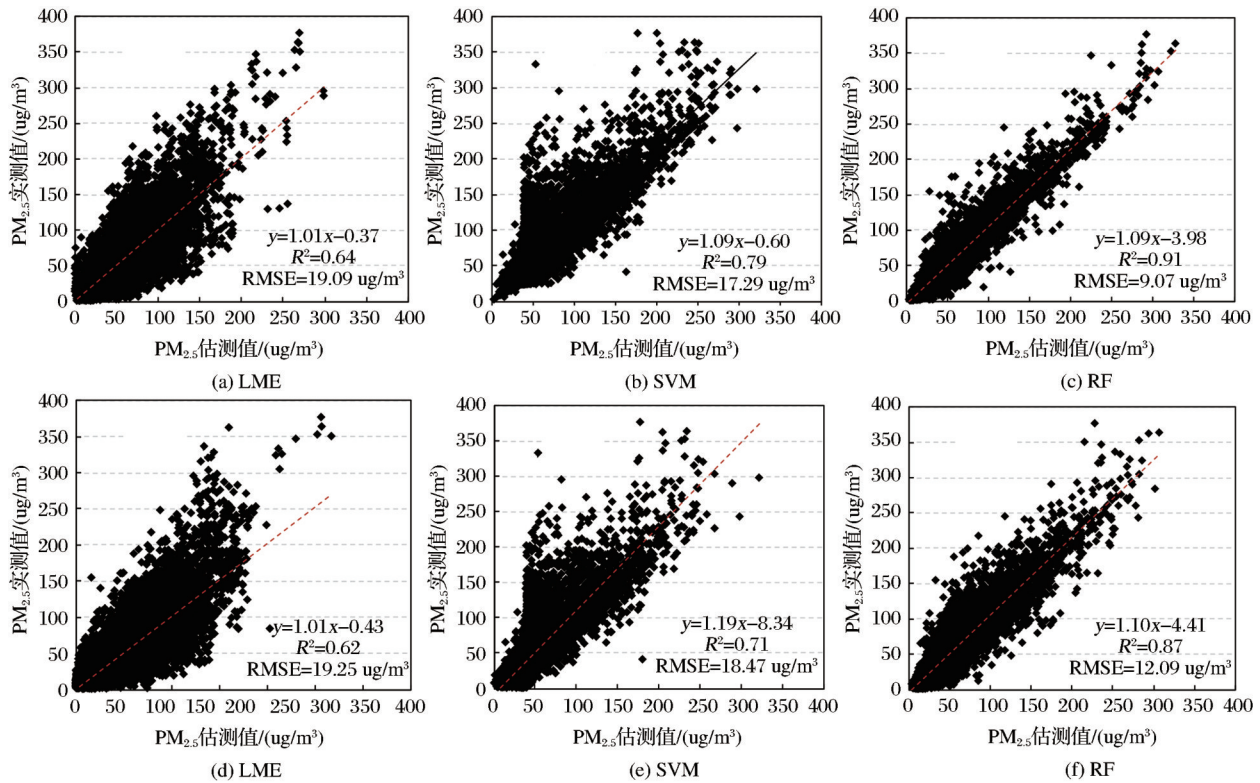


图3 模型拟合和验证结果

Fig.3 The results of model fitting and validation

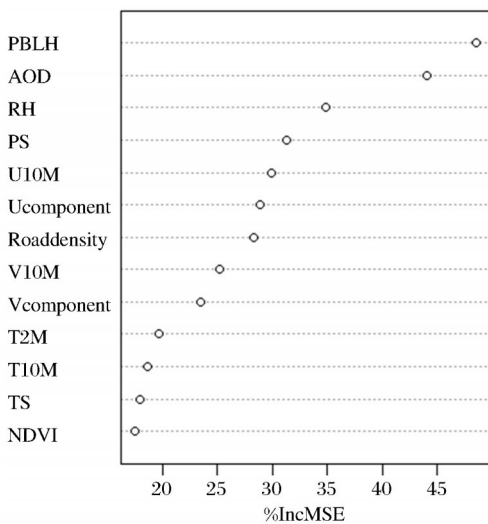


图4 各变量对 $PM_{2.5}$ 浓度变异的重要性

Fig.4 The importance of each variable in explaining the $PM_{2.5}$ variability

实验进一步将建模数据按照春、夏、秋和冬4个季节进行分类,分别探讨了3种模型在YRD-FJ-GD地区的 $PM_{2.5}$ 估算能力。表1给出了3种模型在4个季节的模型估算能力对比。总体来看,RF模型在四季的估算能力均优于LME和SVM模型,其模型拟合的 R^2 均在0.9左右, RMSE则呈现夏季 ($\sim 6.11 \mu g/m^3$) < 秋季 ($\sim 7.84 \mu g/m^3$) < 春季 ($\sim 8.49 \mu g/m^3$) < 冬季 ($\sim 12.53 \mu g/m^3$) 的分布趋势,且4个季节的RMSE值远低于LME模型。进一步探讨模型在4个季节的性能差异,研究发现冬春两季的建模数据样本量远高于夏秋两季,且夏秋两季的AOD- $PM_{2.5}$ 相关性呈现较为明显的日差异性($r=0.1\sim 0.5$),经日差异校正后,LME模型在夏秋两季的估算精度提高至0.55,略低于冬春两季。与LME模型相比,两种机器学习算法在4个季节的估算精度均更优,这是由于机器学习算法具有高泛化能

力,能更好地处理高维数据集(如:实验引入了13个变量),但SVM在针对多数据样本时(如:本研究包含29 874个数据样本),其训练过程耗时更长,且算法精度要低于RF模型。

表1 3个模型在四季的PM_{2.5}估算对比

Table 1 The comparison of three models in four seasons

模型	春		夏		秋		冬	
	R ²	RMSE ($\mu\text{g}/\text{m}^3$)	R ²	RMSE ($\mu\text{g}/\text{m}^3$)	R ²	RMSE ($\mu\text{g}/\text{m}^3$)	R ²	RMSE ($\mu\text{g}/\text{m}^3$)
LME	0.60	18.82	0.50	13.82	0.55	17.33	0.64	29.01
SVR	0.74	13.12	0.81	9.01	0.78	13.32	0.80	20.05
RF	0.86	8.49	0.92	6.11	0.92	7.84	0.91	12.53

3.2 PM_{2.5}浓度估算

鉴于随机森林(RF)模型的估算精度优于线性混合效应模型(LME)和支持向量机模型(SVM),本研究利用RF模型来估算2018年YRD-FJ-GD地区的年均和季均PM_{2.5}浓度。图5给出了2018年YRD-FJ-GD地区的季均和年均PM_{2.5}浓度空间分布。总体上,2018年YRD-FJ-GD地区的季均PM_{2.5}浓度呈现:冬季($46.32 \mu\text{g}/\text{m}^3$)>春季($38.80 \mu\text{g}/\text{m}^3$)>秋季($36.15 \mu\text{g}/\text{m}^3$)>夏季($30.16 \mu\text{g}/\text{m}^3$)的趋势。分地区来看,长江三角洲(YRD)地区的季均PM_{2.5}

浓度均为最高,其中,苏州,无锡和宿迁这3个城市的冬季PM_{2.5}平均浓度分别达到71.38、68.91和68.82 $\mu\text{g}/\text{m}^3$;广东省次之,其中,广州,东莞和佛山的季均PM_{2.5}浓度均高于该地区的其他城市。进一步分析造成YRD-FJ-GD地区的PM_{2.5}平均浓度呈现季节性变化特征的原因可知,夏季的边界层高度(PBLH)总体上比冬季高,而气压(PS)却低于冬季,较低的边界层高度以及较高的气压不利于颗粒物扩散,它将使颗粒物浓度迅速增加。2018年YRD-FJ-GD地区的年均PM_{2.5}浓度范围为29~64 $\mu\text{g}/\text{m}^3$,其中,位于长江三角洲(YRD)地区的大部分城市的年均PM_{2.5}浓度均超过46 $\mu\text{g}/\text{m}^3$;而福建(FJ)的年均PM_{2.5}浓度则较低($\leq 37 \mu\text{g}/\text{m}^3$)。3个地区的年均PM_{2.5}浓度空间分布呈YRD>GD>FJ的格局,其中,PM_{2.5}年均浓度的高值区域主要分布在江苏省的所有城市($\text{PM}_{2.5} \geq 40 \mu\text{g}/\text{m}^3$),其中,苏州、无锡和宿迁是污染最严重的3个地区,其年均PM_{2.5}浓度超过50 $\mu\text{g}/\text{m}^3$,而年均PM_{2.5}浓度最低位于福建省的南平、莆田和浙江省的丽水市。

4 结 语

实验对比了线性混合效应模型(LME)、支持向量机(SVM)和随机森林(RF)模型在YRD-FJ-GD

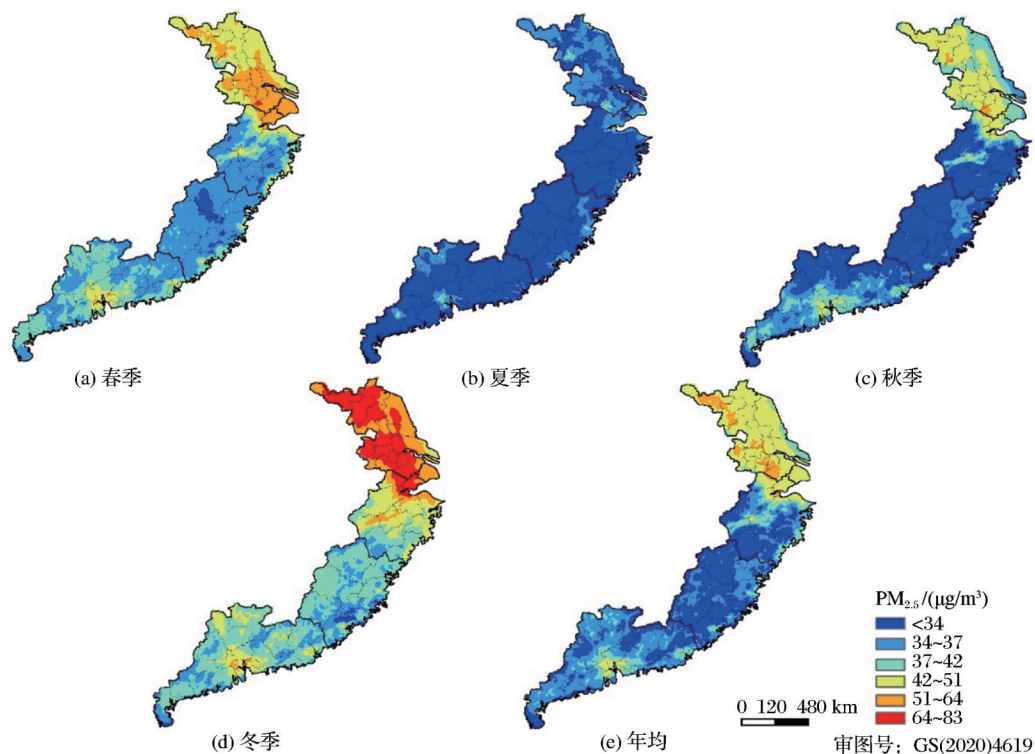


图5 2018年YRD-FJ-GD地区季均和年均PM_{2.5}浓度空间分布

Fig.5 The distribution of seasonal/annual-mean PM_{2.5} in YRD-FJ-GD in 2018

地区的PM_{2.5}估算能力,研究表明,统计模型(LME)和机器学习(SVM和RF)均能较好地应用于YRD-FJ-GD地区的PM_{2.5}估算,且RF模型的PM_{2.5}估算能力优于SVM和LME模型。由RF模型估算出的YRD-FJ-GD地区的PM_{2.5}浓度分布能清晰地展示出该地区的环境污染情况。此外,实验的结果还有一些局限性。首先,只对比了3种模型的估算能力,随机森林模型是否优于其他模型(如:地理加权回归模型和土地利用回归模型等)还未进一步验证;其次,并未考虑人为因素(如:人口密度、工业污染源等)以及监测站点的地理位置差异(如:站点分别位于工业污染源较多的城区和森林覆盖率较高的山区)对区域PM_{2.5}浓度变异的影响。因此,未来的研究将从以下两个方面展开:①利用其他高级统计模型和机器学习方法来探讨YRD-FJ-GD地区的PM_{2.5}估算,以寻求更高精度的估算方法;②加入人口密度、工业污染源分布情况以及监测站点的地理位置等,探讨其是否能进一步提高区域PM_{2.5}浓度的估算能力。

参考文献(References):

- [1] Li Z B, Roy David, Zhang H K, *et al.* Evaluation of Landsat-8 and Sentinel-2A aerosol optical depth retrievals across Chinese cities and implications for medium spatial resolution urban aerosol monitoring[J]. *Remote Sensing*, 2019, 11(2): 122. DOI: 10.3390/rs11020122.
- [2] Lu X M, Wang J J, Yan Y T, *et al.* Estimating hourly PM_{2.5} concentrations using Himawari-8 AOD and a DBSCAN-modified deep learning model over the YRDUA, China[J]. *Atmospheric Pollution Research*, 2021, 12(2): 183-192. DOI: 10.1016/j.apr.2020.10.020.
- [3] Wang Z T, Gao S L, Xie J F, *et al.* Identification of multiple dysregulated metabolic pathways by GC-MS-based profiling of lung tissue in mice with PM_{2.5}-induced asthma[J]. *Chemosphere*, 2019, 220: 1-10. DOI: 10.1016/j.chemosphere.2019.06.063.
- [4] Rojas-Rueda David, Vrijheid Martine, Robinson Oliver, *et al.* Environmental burden of childhood disease in Europe[J]. *International Journal of Environmental Research and Public Health*, 2019, 16(6): 1084. DOI: 10.3390/ijerph16061084.
- [5] Guo B, Wang X X, Pei L, *et al.* Identifying the spatiotemporal dynamic of PM_{2.5} concentrations at multiple scales using geographically and temporally weighted regression model across China during 2015-2018[J]. *Science of the Total Environment*, 2021, 751: 141765. DOI: 10.1016/j.scitotenv.2020.141765.
- [6] Zhang P, Ma W J, Wen F, *et al.* Estimating PM_{2.5} concentration using the machine learning GA-SVM method to improve the land use regression model in Shaanxi, China[J]. *Ecotoxicology and Environmental Safety*, 2021, 225: 112772. DOI: 10.1016/j.ecoenv.2021.112772.
- [7] Wang Y, Yuan Q Q, Li T W, *et al.* Full-coverage spatiotemporal mapping of ambient PM_{2.5} and PM₁₀ over China from Sentinel-5P and assimilated datasets: Considering the precursors and chemical compositions[J]. *Science of the Total Environment*, 2021, 793: 148535. DOI: 10.1016/j.scitotenv.2021.148535.
- [8] Mhawish Alaa, Banerjee Tirthankar, Sorek-Hamer Meytar, *et al.* Comparison and evaluation of MODIS Multi-angle Implementation of Atmospheric Correction (MAIAC) aerosol product over South Asia[J]. *Remote Sensing of Environment*, 2019, 224: 12-28. DOI: 10.1016/j.rse.2019.01.033.
- [9] Huang Keyong, Xiao Qingyang, Meng Xia, *et al.* Predicting monthly high-resolution PM_{2.5} concentrations with random forest model in the North China Plain[J]. *Environmental Pollution*, 2018, 242: 675-683. DOI: 10.1016/j.envpol.2018.07.016.
- [10] Goldberg Daniel, Gupta Pawan, Wang Kai, *et al.* Using gap-filled MAIAC AOD and WRF-Chem to estimate daily PM_{2.5} concentrations at 1 km resolution in the Eastern United States[J]. *Atmospheric Environment*, 2019, 199: 443-452.
- [11] Chen Chuchih, Wang Yinru, Yeh Hungyi, *et al.* Estimating monthly PM_{2.5} concentrations from satellite remote sensing data, meteorological variables, and land use data using ensemble statistical modeling and a random forest approach[J]. *Environmental Pollution*, 2021, 291: 118159. DOI: 10.1016/j.envpol.2021.118159.
- [12] Jill E C, Christopher H, Basil C, *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.
- [13] Wang Z F, Chen L F, Tao J H, *et al.* Satellite-based estimation of regional Particulate Matter (PM) in Beijing using vertical-and-RH correcting method[J]. *Remote Sensing of Environment*, 2010, 114(1): 50-63.
- [14] Liu Y, Paciorek Christopher and Koutrakis Petros. Estimating regional spatial and temporal variability of PM_{2.5} concentrations using satellite data, meteorology, and land use information[J]. *Environmental Health Perspectives*, 2009, 117(6): 886-892.
- [15] Lee H J, Liu Y, Brent C, *et al.* A novel calibration approach of MODIS AOD data to predict PM_{2.5} concentrations[J]. *Atmospheric Chemistry and Physics*, 2011, 11(15): 7991-8002.
- [16] Hu X F, Lance W, Al-Hamdan Mohammad, *et al.* Estimating ground-level PM_{2.5} concentrations in the southeastern US using geographically weighted regression[J]. *Environmental Research*, 2013, 121: 1-10. DOI: 10.1016/j.envres.2012.11.003.
- [17] He Q Q, Huang B. Satellite-based mapping of daily high-resolution ground PM_{2.5} in China via space-time regression modeling[J]. *Remote Sensing of Environment*, 2018, 206: 72-83. DOI: 10.1016/j.rse.2017.12.018.

- [18] Zhang Ying and Li Zhengqiang. Estimation of PM_{2.5} from fine-mode aerosol optical depth [J]. *Journal of Remote Sensing*, 2013, 17(4): 929–943. [张莹, 李正强. 利用细模态气溶胶光学厚度估计 PM_{2.5} [J]. *遥感学报*, 2013, 17(4): 929–943.]
- [19] Drew M, Jessie C, Joseph T B, *et al.* A hybrid land use regression/AERMOD model for predicting intra-urban variation in PM_{2.5} [J]. *Atmospheric Environment*, 2016, 131: 307–315. DOI: 10.1016/j.atmosenv.2016.01.045.
- [20] Fang X, Zou B, Liu X P, *et al.* Satellite-based ground PM_{2.5} estimation using timely structure adaptive modeling [J]. *Remote Sensing of Environment*, 2016, 186: 152–163. DOI: 10.1016/j.rse.2016.08.027.
- [21] Yang L J, Xu H Q, Yu S D. Estimating PM_{2.5} concentrations in Yangtze River Delta region of China using random forest model and the Top-of-Atmosphere reflectance [J]. *Journal of Environmental Management*, 2020, 272: 111061. DOI: 10.1016/j.jenvman.2020.111061.
- [22] Pawan G, Sundar C. Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: 2. a neural network approach [J]. *Journal of Geophysical Research-Atmospheres*, 2009, 114: D20205. DOI: 10.1029/2008JD011497.
- [23] Vahid M, David S, Mahsa M, *et al.* Comparing different methods for statistical modeling of particulate matter in Tehran, Iran [J]. *Air Quality Atmosphere and Health*, 2018, 11(10): 1155–1165.
- [24] Hu X F, Jessica B, Meng X, *et al.* Estimating PM_{2.5} Concentrations in the Conterminous United States Using the random forest approach [J]. *Environmental Science & Technology*, 2017, 51(12): 6936–6944.
- [25] Cole B, Roman J, Monir H, *et al.* Predicting daily urban fine particulate matter concentrations using a random forest model [J]. *Environmental Science & Technology*, 2018, 52(7): 4173–4179.
- [26] Iman K, Kazem A S. A random forest-based framework for crop mapping using temporal, spectral, textural and polarimetric observations [J]. *International Journal of Remote Sensing*, 2019, 40(18): 7221–7251.
- [27] Chai T F, Draxler R. Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? –arguments against avoiding RMSE in the literature [J]. *Geoscientific Model Development*, 2014, 7(3): 1247–1250.
- [28] Cort W, Kenji M. Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in assessing average model performance [J]. *Climate Research*, 2005, 30(1): 79–82.

Research on Methods of Remotely Sensed PM_{2.5} Concentrations Estimation in Four Provinces and One City along the East Coast of China

Yang Lijuan, Zhang Jianxia, Lin Musheng

(Department of Surveying and Mapping Engineering of Minjiang University, Fuzhou 350118, China)

Abstract: The Aerosol Optical Depth (AOD) derived from remote sensing imageries has been widely used in estimating ground-level PM_{2.5} concentrations in large areas. Previous studies that focused on PM_{2.5} estimation have reported high predictability of PM_{2.5} concentrations when using AOD and the advanced statistical model (i.e., Linear Mixed Effects model (LME)). However, the interpretation ability of the LME model was lowered, as it introduced many meteorological and land use variables in the model, and the importance of each variable to PM_{2.5} concentrations was hard to interpret. Therefore, this study developed two nonparametric machine learning methods, i.e., Support Vector Machine (SVM) and Random Forest (RF), to estimate the ground-level PM_{2.5} concentrations. The eastern Yangtze River Delta-Fujian-Guangdong (i.e., YRD-FJ-GD) region in China was employed as our study case, and we also compared the predictability of these two models with the LME model. The results showed that the overall R^2 between estimated and observed PM_{2.5} concentrations exceeded 0.6 for three models, where RF received a R^2 of 0.9, i.e., 13% and 30% higher than SVM ($R^2=0.79$) and LME ($R^2=0.64$) model, respectively. The RMSE values were 9.07, 17.29 and 19.09 $\mu\text{g}/\text{m}^3$ for RF, SVM and LME model, respectively. In addition, the spatial distribution of PM_{2.5} concentrations estimated from the optimal model (i.e., RF) illustrated high annual PM_{2.5} in YRD ($>46 \mu\text{g}/\text{m}^3$), and GD ranked the second. FJ exhibited a relatively low annual PM_{2.5} ($<37 \mu\text{g}/\text{m}^3$). The seasonal PM_{2.5} concentrations presented a distribution pattern as winter ($6.32 \mu\text{g}/\text{m}^3$) $>$ spring ($38.80 \mu\text{g}/\text{m}^3$) $>$ autumn ($36.15 \mu\text{g}/\text{m}^3$) $>$ summer ($30.16 \mu\text{g}/\text{m}^3$). Our results revealed that the AOD and RF model could be a good proxy for estimating PM_{2.5} concentrations in YRD-FJ-GD region.

Key words: LME; SVM; RF; PM_{2.5} estimation; YRD-FJ-GD