doi: 10.6046/gtzyyg.2020.04.18
引用格式: 杨立娟. 基于两层随机森林模型估算中国东部沿海地区的 $PM_{2.5}$ 浓度[J]. 国土资源遥感,2020,32(4):137 - 144. (Yang L J. Estimating $PM_{2.5}$ concentrations in eastern coastal area of China using a two - stage random forest model[J]. Remote Sensing for Land and Resources, 2020, 32(4): 137 - 144.)

基于两层随机森林模型估算中国东部沿海地区的 $PM_{2.5}$ 浓度

杨立娟
(闽江学院测绘工程系,福州 350018)

摘要: 基于暗像元算法反演的气溶胶光学厚度(aerosol optical depht, AOD)产品已被广泛应用于近地面 $PM_{2.5}$ 浓度估算,但该算法不能有效反演高反射率地表的 AOD 值。为此,本研究通过构建包含气象因子的随机森林模型来估算缺失的 AOD 值,并在此基础上,结合 AOD、气象、植被覆盖度和道路密度等参数构建第二层随机森林模型,以估算长江三角洲和珠江三角洲地区的近地面 $PM_{2.5}$ 浓度。研究结果表明,由随机森林模型反演的 AOD 值与 MODIS AOD 值高度相关($R^2 = 0.94$);且模型反演的 $PM_{2.5}$ 浓度与地面实测值之间的 R^2 高达 0.97,均方根误差仅为 5.57 μg/m^3。据此获得的 $PM_{2.5}$ 浓度空间分布显示,$PM_{2.5}$ 年均浓度的高值区域主要分布在地表高程较低的江苏省(≥40 μg/m^3)。研究表明,本研究所构建的包含 AOD 和其他辅助变量的 2 层随机森林模型可有效获取近地面 $PM_{2.5}$ 浓度的空间分布。

关键词: 随机森林模型; $PM_{2.5}$ 空间分布; AOD 反演; 长江三角洲; 珠江三角洲
中图法分类号: TP 79 **文献标志码**: A **文章编号**: 1001 - 070X(2020)04 - 0137 - 08

0 引言

随着我国经济的迅速发展和城市的急剧扩张,空气污染已成为了我国亟需解决的重要环境问题[1-4]。研究表明,悬浮在空气中的动力学直径小于 2.5 μm 的细颗粒物($PM_{2.5}$)是造成我国大部分城市地区雾霾的主要污染物,且其与人体的负面健康密切相关[5-8]。因此,有必要对我国城市地区的近地面 $PM_{2.5}$ 浓度进行有效的估算。传统的地面环境监测网络可以提供准确的监测站点周围的 $PM_{2.5}$ 浓度数据,但却无法获取连续的近地面 $PM_{2.5}$ 浓度空间分布。研究表明,卫星遥感方式可有效用于缺少地面监测网络区域的 $PM_{2.5}$ 浓度估算,其中,由卫星遥感反演的气溶胶光学厚度(aerosol optical depth, AOD)产品已被广泛应用于全球范围内的 $PM_{2.5}$ 估算[9-13]。

利用 AOD 产品来反演区域近地面 $PM_{2.5}$ 浓度的相关模型已经从简单的回归模型(如:线性回归模型[14-15])逐渐发展为高级统计模型(如:土地利用回归模型(land use regression, LUR)[16-17]、地理加权回归模型(geographically weighted regression, GWR)[18-19]、地理和时间加权回归模型(geographically and temporally weighted regression, GTWR)[20-21]以及线性混合效应模型(linear mixed effects model, LME)[22-24])。与简单的回归模型相比,这些高级统计模型通常获得较高的 $PM_{2.5}$ 浓度反演精度,但由于受气候条件的影响,同一模型在不同研究区的反演能力也有所差异。例如,Lee 等[22]利用 LME 模型对美国东北部地区的 $PM_{2.5}$ 浓度进行反演,结果表明,模型反演的 R^2 为 0.92,均方根误差(root mean square error, RMSE)为 2.45 μg/m^3;而 Sorek 等[25]将 LME 模型运用在以色列地区的 $PM_{2.5}$ 浓度反演时,其 R^2 仅为 0.45,*RMSE* 高达 12.06 μg/m^3。因此,为了提高模型反演的准确性,研究者们逐渐引入更多的辅助变量(如气象参数和土地利用信息)来构建 AOD - $PM_{2.5}$ 模型。例如,Ma 等[26]构建了包含 8 个变量的 GWR 模型来反演全国的 $PM_{2.5}$ 浓度,结果表明,GWR 模型反演的 R^2 为 0.64,*RMSE* 为 32.98 μg/m^3;He 等[21]通过考虑 AOD - $PM_{2.5}$ 的时间变化引入 5 个气象参数和 2 个土地利用变

收稿日期: 2020 - 02 - 03; **修订日期**: 2020 - 03 - 11
基金项目: 闽江学院引进人才科研启动项目"基于机器学习的中高空间分辨率 $PM_{2.5}$ 遥感估算模型研究"(编号: MJY20001)和闽江学院纵向校级项目"基于卫星遥感的 $PM_{2.5}$ 浓度时空分布研究"(编号: MYK19029)共同资助。
作者简介: 杨立娟(1985 -),女,博士,副教授,主要从事环境与资源遥感研究。Email: subrinarzhong@aliyun.com。

量,进一步将 GWR 模型扩展成 GTWR 模型来反演全国的 $PM_{2.5}$浓度,结果表明由 GTWR 模型反演的 R^2 提高至0.80,*RMSE* 也下降为 18.58 μg/m^3。

此外,也有研究者采用机器学习算法来反演区域的近地面 $PM_{2.5}$浓度。例如,Gupta 等[14]使用人工神经网络(artificial neural network, ANN)算法来减少由 AOD 带来的估算误差,其结果表明由 ANN 反演的 $PM_{2.5}$浓度与实测值之间的 R^2 为 0.61; Mehdipour 等[27]比较了 3 种机器学习算法:决策树(decision tree, DT)、批量归一化(batch normalization, BN)和支持向量机(support vector machine, SVM)在伊朗德黑兰的 $PM_{2.5}$浓度反演能力,结果证实 SVM 的反演精度最高。与高级统计模型类似,机器学习算法也引入了辅助变量来提高模型的反演精度,但这些多参数反演模型在建模前均需要对各参数和 $PM_{2.5}$浓度之间的相关性以及各参数之间的自相关性进行验证,且模型的结果不能体现各输入变量影响 $PM_{2.5}$浓度变异的重要性[28-29]。因此,本研究利用一种集成的机器学习算法(随机森林)来反演区域的 $PM_{2.5}$浓度。随机森林不仅可以通过调整 2 个参数(即 m_{try}和 n_{tree})来获得模型的最优估计,同时还能提供各变量影响 $PM_{2.5}$浓度变异的重要性指标,从而比其他机器学习算法更合理地解释了近地面 $PM_{2.5}$浓度的变化特征。

本研究的主要目的是利用随机森林模型来估算城市地区的近地面 $PM_{2.5}$浓度。其中,用于构建模型的数据主要有来自中分辨率成像光谱仪(Moderate-Resolution Imaging Spectroradiometer, MODIS)空间分辨率为 3 km 的 AOD 产品(以下简称为 MODIS 3 km AOD)、气象因子、植被覆盖度和道路密度等 4 类参数。本研究选择 2 个东部沿海地区作为研究区域,即:长江三角洲(YRD)和珠江三角洲(PRD)地区,并构建包含多参数的随机森林模型来估算该区域的近地面 $PM_{2.5}$浓度。由于 MODIS 3 km AOD 产品是采用暗像元算法反演而得,这将导致高反射率的地表(建筑密集的城区和道路等)无有效的 AOD 值。为此,本研究提出了一种 2 层的随机森林估算模型,其中第一层模型主要用来估算高反射率地表的 AOD 值,并结合 MODIS 3 km AOD 产品来获取能够覆盖 YRD 和 PRD 区域的 AOD 全空间覆盖分布;在此基础上,结合 AOD、气象因子、植被覆盖度和道路密度等参数来构建第二层随机森林模型,以估算 2018 年 YRD 和 PRD 地区的近地面 $PM_{2.5}$浓度。

1　研究区概况及数据源

1.1　研究区概况

本研究选取了 2 个东部沿海地区,研究区范围如图 1 所示。整个研究区涵盖了上海市、江苏省、浙江省和广东省 4 个区域,其中,YRD 地区包含了上海市、江苏省和浙江省在内的 25 个城市,PRD 地区包含了 21 个城市。随着经济的迅速发展和城市的急剧扩张,YRD 和 PRD 地区已成为我国面积最大的、经济最发达的 2 个城市群,伴随而来的是这 2 个地区的空气质量也在不断下降。在过去的十几 a 中,YRD 和 PRD 地区 $PM_{2.5}$年均浓度分别高达 67 μg/m^{3}和 55 μg/m^3,都超过了我国环境空气质量的二级标准(~35 μg/m^3)[13, 30]。

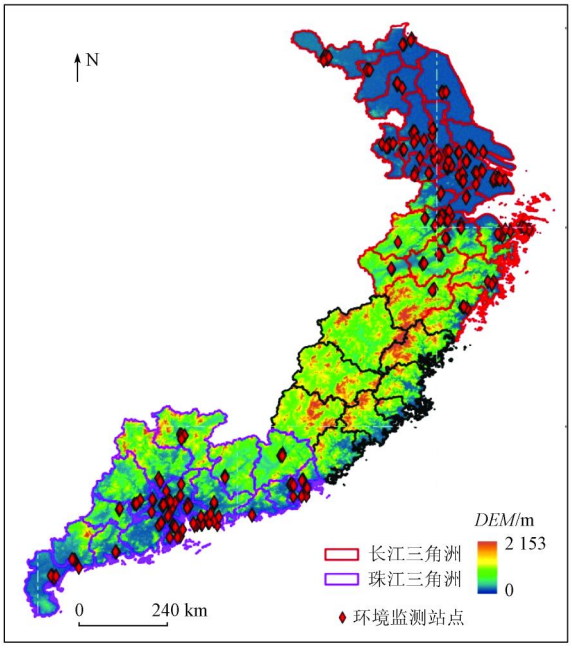


图 1　研究区示意图

Fig. 1　The study area

1.2　数据源

1.2.1　$PM_{2.5}$和 MODIS AOD 数据

本研究的 $PM_{2.5}$浓度数据主要来自 2018 年上海市、江苏省、浙江省和广东省环境保护厅网站提供的共 260 个地面监测站点的每小时 $PM_{2.5}$浓度。建模前需先对所有站点的 $PM_{2.5}$浓度异常值进行剔除,如 $PM_{2.5}$小于 0 μg/m^{3}和 $PM_{2.5}$为 NA(not available,即该值不可用)等。AOD 数据主要由 Terra 星上搭载的 MODIS 传感器所提供的空间分辨率为 3 km 的 AOD 产品,该数据下载于美国航空航天局网站(https://ladsweb.modaps.eosdis.nasa.gov/search/),然后利用 IDL 语言提取 0.55 μm 处的 AOD 值。

1.2.2　气象数据

气象数据主要来源于戈达德地球观测系统数据同化系统提供的空间分辨率为 0.25° × 0.312 5°的前向处理数据(GEOS-5 FP)。本研究共下载了 10 个气象参数:行星边界层高度(PBLH, m)、地表温

度(T_s,K)、2 m 分辨率气温($T_{2\,m}$,K)、10 m 分辨率气温($T_{10\,m}$,K)、10 m 分辨率东向风($U_{10\,m}$,m/s)、10 m 分辨率北向风($V_{10\,m}$,m/s)、相对湿度(RH,%)、风的纬向分量(U - component,m/s)和经向分量(V - component,m/s)以及气压(PS,hPa)。采用最临近插值法将所有气象参数插值到 3 km ×3 km 网格中,插值过程在 Python 软件中完成。

1.2.3　土地利用数据

本研究的土地利用数据包含植被覆盖度和道路密度数据。其中,植被覆盖度数据来源于 MODIS 提供的空间分辨率为 1 km 的植被指数产品(MOD13A3),该产品已被广泛用于全球植被状况的监测。道路数据来源于百度地图提供的包含高速公路、国道和省道等主要道路的矢量文件,并在 ArcGIS 软件中计算每个 3 km ×3 km 网格单元的道路密度值。

2　模型开发和验证

随机森林是一种集成的机器学习方法,最初由 Breiman[31] 开发,目前已被广泛用于市场营销、生物分类和医学等各个领域。随机森林使用 bootstrap 重采样方法选择随机子样本,然后基于随机子样本的特征选择方法为每个子样本选择一个预测子集,并将多数选票结果作为随机森林的最终预测结果[32]。随机森林通过确定每个节点的预测变量数(m_{try})和每个决策树的数目(n_{tree})这 2 个重要参数来获得最优估计[33]。本研究旨在构建包含 MODIS 3 km AOD、气象因子、植被覆盖度和道路密度等参数的随机森林模型,以估算 YRD 和 PRD 地区的近地面 $PM_{2.5}$ 浓度。由于 MODIS 3 km AOD 主要采用暗像元算法来反演,因此在高反射率的地表(建筑密集的城区和道路等)无有效的 AOD 值。已有的研究主要利用简单的克里格插值法或将多个传感器反演的 AOD 进行融合来获取连续的 AOD 空间分布[28,34],但这 2 种方法的估算精度均较低。为此,本研究通过构建随机森林模型,利用相关气象因子(如 PBLH 和 RH 等)来估算缺失的 AOD 值。在此基础上,结合 AOD、气象、植被覆盖度以及道路密度等参数,构建第二层随机森林模型来估算 2018 年 YRD 和 PRD 地区的近地面 $PM_{2.5}$ 浓度。另一方面,随机森林模型中的指标——增长的错误率平方均值(increased in mean squared error, IncMSE)可用于验证各变量在 $PM_{2.5}$ 浓度变异中的重要性,因此,相比于其他机器学习算法,随机森林算法的应用更广泛。IncMSE 值越大,代表该变量的重要性越大。本研究所提出的 2 层随机森林模型可简写为:

$$AOD = RF(PBLH, T_s, T_{2\,m}, T_{10\,m}, U_{10\,m}, V_{10\,m}, RH), \quad (1)$$

$$PM_{2.5} = RF(AOD, meteorologicalfields, vegetationcover, roaddensity), \quad (2)$$

式中:*AOD* 为第一层模型的估算值;*meteorologicalfields* 主要包含了 *PBLH*, T_s, $T_{2\,m}$, $T_{10\,m}$, $U_{10\,m}$, $V_{10\,m}$, *RH*, *U - component*, *V - component* 和 *PS* 等 10 个参数;*vegetationcover* 为植被覆盖度,*roaddensity* 为道路密度。

本研究采用 10 折交叉验证方法来评价 2 层随机森林模型估算 $PM_{2.5}$ 浓度的能力。10 折交叉验证法是指将建模数据集随机分为 10 个部分,90% 的数据用于模型训练,剩余 10% 用于模型预测。另外,本研究使用决定系数(R^2)和 RMSE 这 2 项指标来评估模型估算的 $PM_{2.5}$ 浓度和地面实测值之间的相关性。

3　结果与分析

3.1　数据统计

表 1 给出了参与建模的 14 个参数的统计数据。其中,2018 年 YRD 和 PRD 地区的 $PM_{2.5}$ 浓度范围为 1 ~ 377 μg/m^3,总体表现为 YRD 地区较高,PRD 地区略低;AOD 的分布和 $PM_{2.5}$ 浓度类似,2018 年 AOD 平均值呈现 YRD 高于 PRD 的格局。就 AOD - $PM_{2.5}$ 关系而言,二者在夏秋 2 季的相关性约为 0.25,而在冬春 2 季则下降至 0.10。造成这种关系差异的主要原因是 AOD - $PM_{2.5}$ 之间的相关性易受不同气候条件的影响。PBLH 与 $PM_{2.5}$ 浓度的季节性特征呈现明显的相反趋势,即在 $PM_{2.5}$ 浓度较高(低)的冬春(夏秋)2 季,PBLH 较低(高)。MODIS NDVI 的季均值分别为 0.32(春)、0.40(夏)、0.40(秋)和 0.30(冬)。

表 1　建模参数的统计数据

Tab. 1　Statistics of parameters for model fitting

变量	最小值	最大值	均值	标准差
$PM_{2.5}$/(μg·m^{-3})	1.00	377.00	41.53	32.00
AOD	0.01	2.20	0.26	0.18
PBLH/m	63.38	2 227.65	940.37	941.88
PS/hPa	918.60	1 034.00	1 003.20	1 007.00
RH/%	13.50	100.00	62.30	64.30
$T_{2\,m}$/K	269.80	310.20	294.90	296.00
$T_{10\,m}$/K	269.40	309.30	294.20	295.30
T_s/K	271.90	320.80	297.70	298.90
$U_{10\,m}$/(m·s^{-1})	-11.44	9.04	-0.79	-0.95
U - component/(m·s^{-1})	-15.35	14.36	-1.07	-1.38
$V_{10\,m}$/(m·s^{-1})	-17.95	11.59	-0.23	-0.18
V - component/(m·s^{-1})	-24.14	18.07	-0.40	-0.42
vegetationcover	0.00	0.87	0.35	0.33
roaddensity/(km·km^{-2})	0.11	2.31	1.13	1.05

3.2　随机森林建模和验证

本研究共使用了前文所说的 14 个变量来参与建模。通过对随机森林模型的训练，最终将 m_{try} 和 n_{tree} 分别设为 4 和 500，以达到最优估计。图2给出了 13 个自变量在 $PM_{2.5}$ 浓度变异中的重要性评价。结果表明，*AOD*，*PBLH* 和 *RH* 是解释 YRD 和 PRD 区域中 $PM_{2.5}$ 浓度变化的前 3 个最重要变量。此外，研究发现，在获取的 29 873 组建模数据中，9 871 组（33%）的数据具有完整的 *AOD* 和 $PM_{2.5}$ 值，剩余的 20 002 组（67%）只有 $PM_{2.5}$ 值（缺少 *AOD*）。本文首先选择 33% 的数据集（同时拥有 *AOD* 和 $PM_{2.5}$ 值）对随机森林模型进行训练，结果表明模型估算的 $PM_{2.5}$ 浓度和地面实测值之间的 R^2 为 0.95，*RMSE* 为 5.73 μg/m^3。为了获取完整的 AOD 值空间分布，本研究构建了包含 *PBLH*，*RH*，*PS* 和 T_s 等气象参数的随机森林模型来预测缺失的 AOD 值，结果表明，由随机森林估算的 *AOD* 与 MODIS 3 km *AOD* 高度相关，二者之间的 R^2 高达 0.94（图 3(a)）。在此基础上，进一步利用第一层模型估算的 AOD 值，并结合气象因子以及植被覆盖度和道路密度等参数来构建第二层随机森林模型，结果显示模型估算的 2018 年 YRD 和 PRD 地区的 $PM_{2.5}$ 浓度值和实际测量值之间的 R^2 达到了 0.97，*RMSE* 为 5.57 μg/m^3（图 3(b)）。

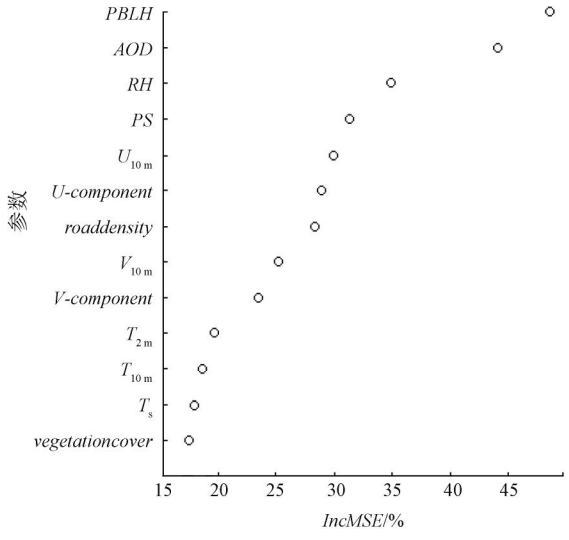
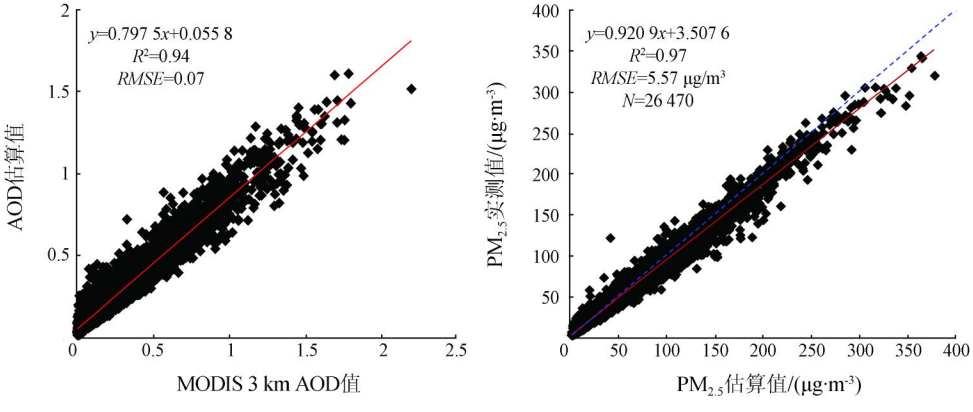


图 2　各变量在 $PM_{2.5}$ 浓度变异中的重要性

Fig. 2　Importance of each parameter in $PM_{2.5}$ variability

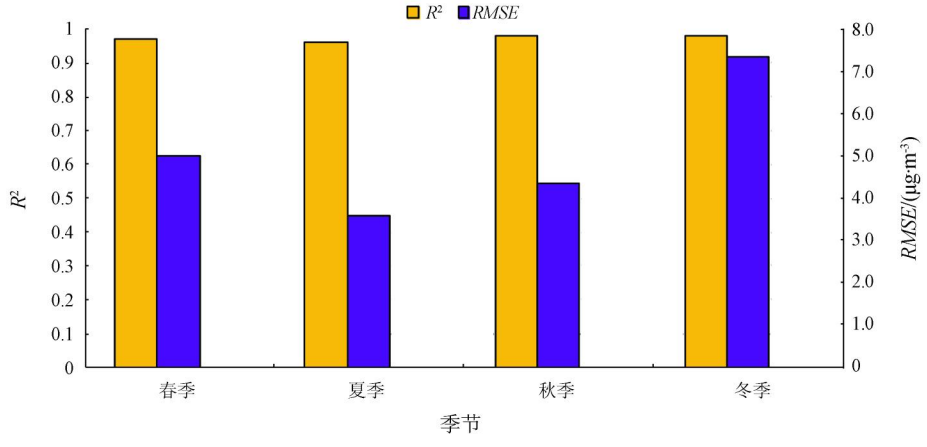


（a）MODIS 估算值　　（b）$PM_{2.5}$ 估算值

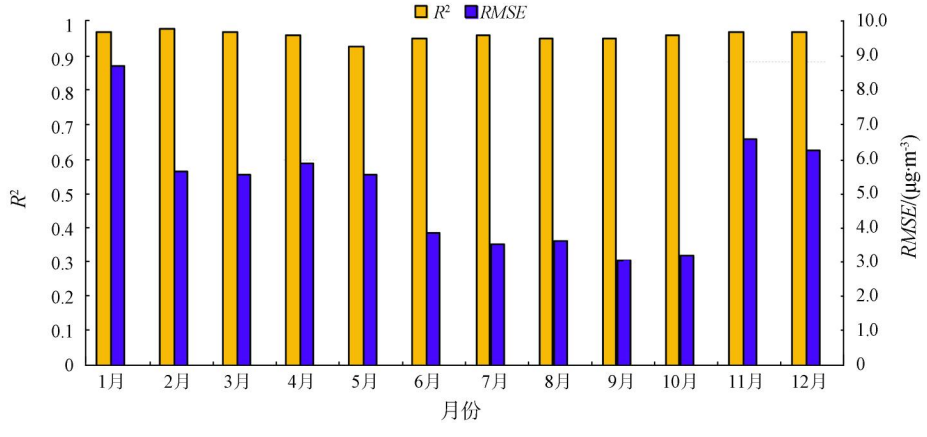
图 3　2 层随机森林模型的估算结果

Fig. 3　Estimated results of random forest model

进一步分地区对模型性能进行验证，结果发现由随机森林模型估算的 YRD 和 PRD 地区的 $PM_{2.5}$ 浓度与地面实测值之间的 R^2 分别为 0.98 和 0.97；*RMSE* 分别为 5.85 μg/m^3 和 4.67 μg/m^3。Ma 等[13] 和 Song 等[30] 分别利用 LME 和 GWR 模型对 YRD 和 PRD 地区开展了 $PM_{2.5}$ 浓度遥感估算，结果表明模型估算值和地面实测值之间的 R^2 仅为 0.67（YRD）和 0.73（PRD）。图 4 显示了 4 个季节和 12 个月份的模型反演结果对比，结果表明春季、夏季、秋季和冬季模型的 R^2 分别为 0.97，0.96，0.98 和 0.98；4 个季节的 *RMSE* 总体表现为：冬季（7.34 μg/m^3）> 春季（5.00 μg/m^3）> 秋季（4.35 μg/m^3）> 夏季（3.60 μg/m^3）。而 12 个月份中模型拟合的 R^2 均在 0.93 以上，6—10 月份的 *RMSE* 略低于其他月份。分区域和分季节的模型估算结果表明了本研究所提出的 2 层随机森林模型在 YRD 和 PRD 区域中具有较高的 $PM_{2.5}$ 估算能力。图 5 和表 2 显示了利用 10 折交叉验证法（cross validation，CV）对 2 层随机森林模型进行验证的结果。可以看出，全年和 4 个季节的模型交叉验证结果和拟合结果均表现出良好的一致性，模型 CV 估算的 R^2 均大于0.95，且 4 个季节的 RMSE 也呈现出冬春 2 季高于夏秋 2 季的特点。



（a）分季节模型估算



（b）分月份模型估算

图 4　分季节和分月份的模型估算结果

Fig. 4　Estimated results of random forest model for four seasons and twelve months

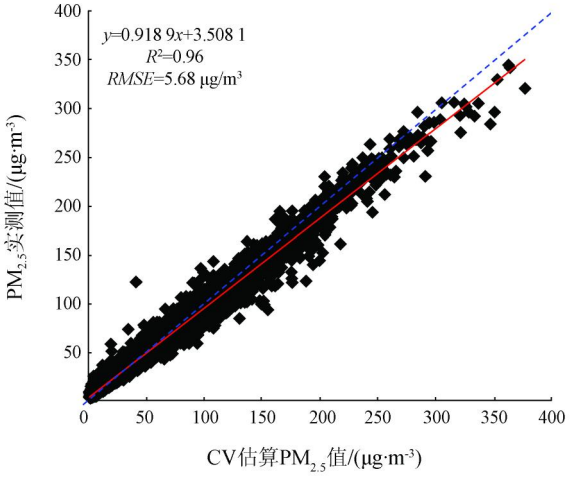


图 5　模型 CV 估算结果

Fig. 5　CV results of random forest model

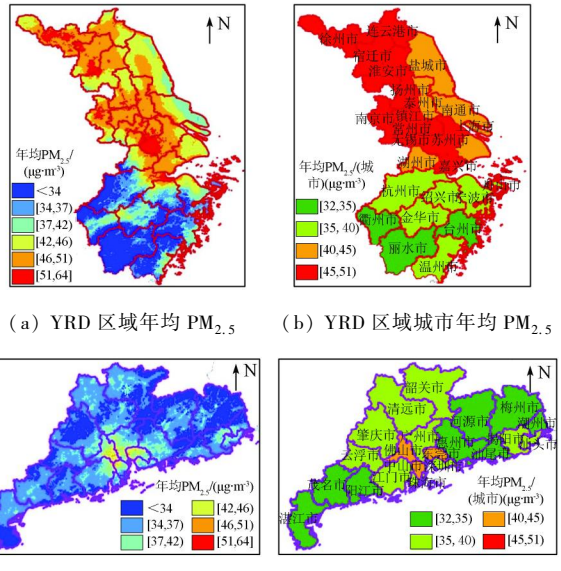
表 2　全年和 4 个季节的模型 CV 估算结果

Tab. 2　CV results of random forest model for the entire period and four seasons

时间	R^2	$RMSE/(\mu g \cdot m^{-3})$
全年	0.97	5.73
春季	0.97	5.99
夏季	0.95	3.99
秋季	0.96	4.62
冬季	0.96	7.66

3.3　区域 $PM_{2.5}$ 浓度估算

图 6 分别显示了 2018 年 YRD 和 PRD 区域的年均 $PM_{2.5}$ 浓度分布和 46 个城市的年均 $PM_{2.5}$ 浓度



（a）YRD 区域年均 $PM_{2.5}$　（b）YRD 区域城市年均 $PM_{2.5}$

（c）PRD 区域年均 $PM_{2.5}$　（d）PRD 区域城市年均 $PM_{2.5}$

图 6　YRD 和 PRD 区域的年均 $PM_{2.5}$ 浓度空间分布

Fig. 6　Spatial distribution of annual $PM_{2.5}$ concentrations in YRD and PRD

分布。从图 6 可以看出,整个研究区的$PM_{2.5}$年均浓度在 29 ~64 μg/m^{3}之间,其中,位于 YRD 的大部分地区 $PM_{2.5}$年均浓度均超过 46 μg/m^3。46 个城市中,$PM_{2.5}$年均浓度的高值区域主要分布在江苏省的所有城市($PM_{2.5}$≥40 μg/m^3),其中,徐州市、无锡市和宿迁市是空气污染最严重的 3 个地区,年均 $PM_{2.5}$浓度值超过 50 μg/m^3;而浙江省南部地区的年均 $PM_{2.5}$浓度均低于 35 μg/m^3,其中,$PM_{2.5}$年均浓度的最低值位于浙江省的丽水市。PRD 地区的 $PM_{2.5}$年均浓度总体比 YRD 的低,其中,广州市、东莞市和佛山市等城市的 $PM_{2.5}$年均浓度高于其他省内城市。对比 YRD 和 PRD 区域的地表高程可以看出,海拔较低的 YRD 北部地区的 $PM_{2.5}$年均浓度较高,而海拔较高的 PRD 北部部分区域则表现为 $PM_{2.5}$年均浓度相对较低。这是由于频繁的人类活动主要集中在海拔较低的平原,由人类活动带来的汽车尾气和工业排放的废气会使空气中的颗粒物浓度急剧升高。

图 7 显示了 YRD 和 PRD 区域 4 季的 $PM_{2.5}$平均浓度。4 个季节中,冬季的 $PM_{2.5}$浓度最高(~46.32 μg/m^3),其次是春季(~38.80 μg/m^3)和秋季(~36.15 μg/m^3);夏季的 $PM_{2.5}$平均值最低,仅为 30.16 μg/m^3,比冬季低 35%。分地区来看,YRD 地区 4 个季节的平均 $PM_{2.5}$浓度要高于 PRD。从图 7 还可以看出,$PM_{2.5}$季均浓度的高值区域全部位于 YRD 的北部,其中,徐州市、无锡市和宿迁市等城市的冬季 $PM_{2.5}$平均浓度分别达到 71.38 μg/m^3,68.91 μg/m^{3}和 68.82 μg/m^3。此外,位于 PRD 地区的广州市、东莞市和佛山市的 $PM_{2.5}$季均浓度要略高于广东省的其他城市。本文的研究成果与 Ma 等[13]和 Song 等[30]的研究结果基本一致,均体现出了位于 YRD 的徐州市、无锡市和宿迁市以及位于 PRD 的广州市、东莞市和佛山市等城市的空气污染状况较严重。进一步对 YRD 和 PRD 区域的 $PM_{2.5}$浓度季节性变化进行分析。总体来看,该区域的边界层高度(PBLH)夏季较高,冬季较低;而气压(PS)与 PBLH 相比则呈现相反的趋势。较低的大气边界层高度以及较高的气压等不利条件均易使空气中的颗粒物浓度迅速增加,因此,这也是导致 YRD 和 PRD 区域的 $PM_{2.5}$浓度呈现冬春 2 季高于夏秋 2 季的重要原因。

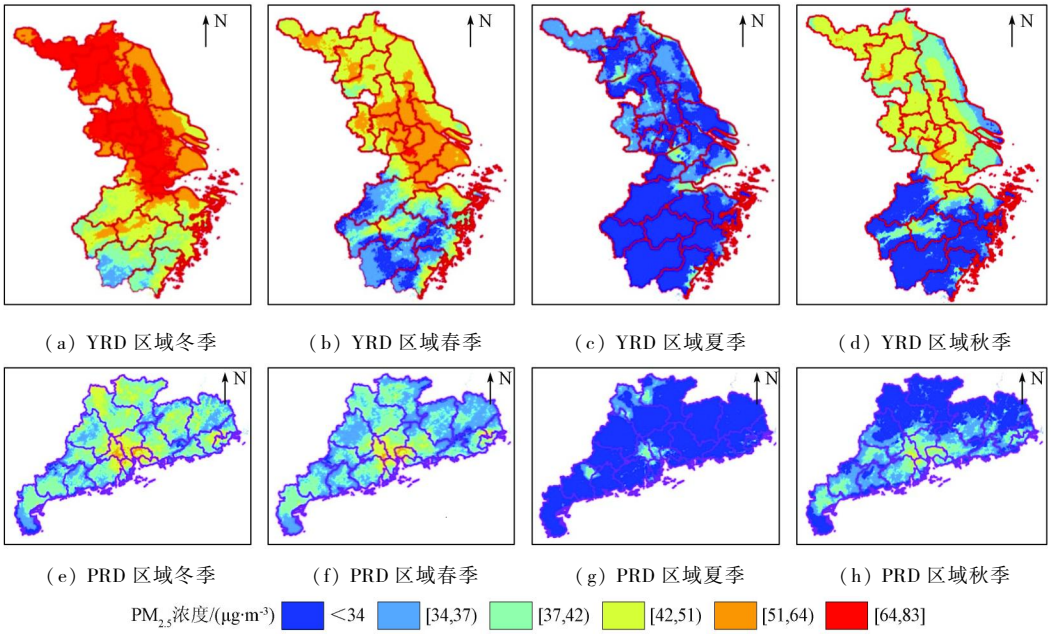


图 7　YRD 和 PRD 区域 4 季的 $PM_{2.5}$空间分布

Fig. 7　Spatial distribution of seasonal $PM_{2.5}$ concentrations in YRD and PRD

4　结论

本研究基于 MODIS 3 km AOD、气象因子、植被覆盖度和道路密度等多类参数,构建了 2 层随机森林模型来估算 YRD 和 PRD 地区的近地面 $PM_{2.5}$浓度,得到以下结论:

1) AOD 是解释 YRD 和 PRD 地区 $PM_{2.5}$浓度变化的最重要变量之一,利用随机森林模型可有效填补高反射率地表的 AOD 值,从而获得连续的 AOD 空间分布。

2) 由 2 层随机森林模型反演的 YRD 和 PRD 地区的 $PM_{2.5}$浓度与地面实测值高度相关;分季节的模型性能对比结果也表明了本研究的 2 层随机森林模型在 YRD 和 PRD 地区具有较高的 $PM_{2.5}$反演能力。

3）已有研究结果表明 YRD 和 PRD 地区的冬春 2 季 $PM_{2.5}$ 浓度均高于夏秋 2 季，这与本文的结论相似。此外，已有研究主要利用 AOD 来反演 YRD 和 PRD 地区的 $PM_{2.5}$ 浓度，并未获取连续空间分布的 AOD，结果有失准确性。本文利用 2 层随机森林模型获取了研究区全空间覆盖的 $PM_{2.5}$ 浓度空间分布，从而更清晰地揭示了 YRD 和 PRD 地区的 $PM_{2.5}$ 污染的时空分异趋势。

本研究还存在几点不足：首先，本研究没有考虑地面环境监测站点的地理位置差异，因此，模型的估算精度是否会受站点地理位置的影响还需要进一步验证；其次，本研究提出的模型虽然获得了较高的 R^2，但模型的 RMSE 也略高于一些应用在欧美国家的模型。因此，今后的研究将进一步考虑地理位置以及人为因素（如：人口密度和工业污染源）等对区域近地面 $PM_{2.5}$ 浓度的影响。

志谢：本研究所用的 MODIS 数据和气象数据由美国国家航空航天局（NASA）提供，在此表示感谢。

参考文献（References）：

[1] Ma J Z, Chen Y, Wang W, et al. Strong air pollution causes widespread haze – clouds over China[J]. Journal of Geophysical Research – Atmospheres, 2010, 115(D18204).

[2] Deng J J, Du K, Wang K, et al. Long – term atmospheric visibility trend in Southeast China, 1973—2010[J]. Atmospheric Environment, 2012, 59: 11 – 21.

[3] Fang X, Zou B, Liu X P, et al. Satellite – based ground PM2.5 estimation using timely structure adaptive modeling[J]. Remote Sensing of Environment, 2016, 186: 152 – 163.

[4] 陈 辉，厉 青，王中挺，等. MERSI 和 MODIS 卫星监测京津冀及周边地区 PM2.5 浓度[J]. 遥感学报，2018，22(5)：822 – 832.
Chen H, Li Q, Wang Z T, et al. Utilization of MERSI and MODIS data to monitor PM2.5 concentrations in Beijing – Tianjin – Hebei and its surrounding areas[J]. Journal of Remote Sensing, 2018, 22(5): 822 – 832.

[5] Zhao P S, Zhang X L, Xu X F, et al. Long – term visibility trends and characteristics in the region of Beijing, Tianjin, and Hebei, China[J]. Atmospheric Research, 2011, 101(3): 711 – 718.

[6] Cao J J, Shen Z X, Chow J C, et al. Winter and summer PM2.5 chemical compositions in fourteen Chinese cities[J]. Journal of the Air and Waste Management Association, 2012, 62(10): 1214 – 1226.

[7] Wang S, Li G G, Gong Z Y, et al. Spatial distribution, seasonal variation and regionalization of PM2.5 concentrations in China[J]. Science China – Chemistry, 2015, 58(9): 1435 – 1443.

[8] Xu H, Guang J, Xue Y, et al. A consistent aerosol optical depth (AOD) dataset over mainland China by integration of several AOD products[J]. Atmospheric Environment, 2015, 114: 48 – 56.

[9] van Donkelaar A, Martin R V, Park R J. Estimating ground – level PM2.5 using aerosol optical depth determined from satellite remote sensing[J]. Journal of Geophysical Research – Atmospheres, 2006, 111(D21).

[10] Fan X H, Chen H B, Lin L F, et al. Retrieval of aerosol optical properties over the Beijing area using POLDER/PARASOL satellite polarization measurements[J]. Advances in Atmospheric Sciences, 2009, 26(6): 1099 – 1107.

[11] Livingston J M, Redemann J, Shinozuka Y, et al. Comparison of MODIS 3 km and 10 km resolution aerosol optical depth retrievals over land with airborne sunphotometer measurements during ARCTAS summer 2008[J]. Atmospheric Chemistry and Physics, 2014, 14(4): 2015 – 2038.

[12] 贾松林，苏 林，陶金花，等. 卫星遥感监测近地表细颗粒物多元回归方法研究[J]. 中国环境科学，2014(3)：565 – 573.
Jia S L, Su L, Tao J H, et al. A study of multiple regression method for estimating concentration of fine particulate matter using satellite remote sensing[J]. China Environmental Science, 2014(3): 565 – 573.

[13] Ma Z W, Liu Y, Zhao Q Y, et al. Satellite – derived high resolution PM2.5 concentrations in Yangtze River Delta region of China using improved linear mixed effects model[J]. Atmospheric Environment, 2016, 133: 156 – 164.

[14] Gupta P, Christopher S A. 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.

[15] 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.

[16] Yang X F, Zheng Y X, Geng G N, et al. Development of PM2.5 and NO_2 models in a LUR framework incorporating satellite remote sensing and air quality model data in Pearl River Delta region, China[J]. Environmental Pollution, 2017, 226: 143 – 153.

[17] 阳海鸥，陈文波，梁照凤. LUR 模型模拟的南昌市 PM2.5 浓度与土地利用类型的关系[J]. 农业工程学报，2017，33(6)：232 – 239.
Yang H O, Chen W B, Liang Z F. Relationship of PM2.5 concentration and land use type in Nanchang City based on LUR simulation[J]. Transactions of the Chinese Society of Agricultural Engineering, 2017, 33(6): 232 – 239.

[18] Zhang T H, Gong W, Wang W, et al. Ground level PM2.5 estimates over China using satellite – based geographically weighted regression (GWR) models are improved by including NO_2 and enhanced vegetation index (EVI)[J]. International Journal of Environmental Research and Public Health, 2016, 13(12): 1215.

[19] Xiao L, Lang Y, Christakos G. High – resolution spatiotemporal mapping of PM2.5 concentrations at mainland China using a combined BME – GWR technique[J]. Atmospheric Environment, 2018, 173: 295 – 305.

[20] Bai Y, Wu L X, Qin K, et al. A Geographically and temporally weighted regression model for ground – level PM2.5 estimation from satellite – derived 500 m resolution AOD[J]. Remote Sensing, 2016, 8(3): 262.

[21] He Q Q, Huang B. Satellite – based mapping of daily high – resolution ground PM2.5 in China via space – time regression modeling

[J]. Remote Sensing of Environment, 2018, 206: 72 – 83.

[22] Lee H J, Liu Y, Coull B A, et al. A novel calibration approach of MODIS AOD data to predict PM2.5 concentrations[J]. Atmospheric Chemistry and Physics, 2011, 11(15): 7991 – 8002.

[23] 杨立娟,徐涵秋,金致凡. MODIS 卫星遥感估计福州地区近地面 PM2.5 浓度[J]. 遥感学报, 2018, 22(1): 64 – 75.
Yang L J, Xu H Q, Jin Z F. Estimation of ground – level $PM_{2.5}$ concentrations using MODIS satellite data in Fuzhou, China[J]. Journal of Remote Sensing, 2018, 22(1): 64 – 75.

[24] Yang L J, Xu H Q, Jin Z F. Estimating ground – level PM2.5 over a coastal region of China using satellite AOD and a combined model[J]. Journal of Cleaner Production, 2019, 227: 472 – 482.

[25] Sorek H M, Kloog I, Koutrakis P, et al. Assessment of PM2.5 concentrations over bright surfaces using MODIS satellite observations [J]. Remote Sensing of Environment, 2015, 163: 180 – 185.

[26] Ma Z W, Hu X F, Huang L, et al. Estimating ground – level PM2.5 in China using satellite remote sensing[J]. Environmental Science and Technology, 2014, 48(13): 7436 – 7444.

[27] Mehdipour V, Stevenson D S, Memarianfard 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.

[28] Hu X F, Belle J H, Meng X, et al. Estimating PM2.5 concentrations in the conterminous united states using the random forest approach[J]. Environmental Science and Technology, 2017, 51(12): 6936 – 6944.

[29] Brokamp C, Jandarov R, Hossain M, et al. Predicting daily urban fine particulate matter concentrations using a random forest model [J]. Environmental Science and Technology, 2018, 52(7): 4173 – 4179.

[30] Song W Z, Jia H F, Huang J F, et al. A satellite – based geographically weighted regression model for regional PM2.5 estimation over the Pearl River Delta region in China[J]. Remote Sensing of Environment, 2014, 154: 1 – 7.

[31] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5 – 32.

[32] Khosravi I, Alavipanah S K. 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.

[33] Huang K Y, Xiao Q Y, Meng X, et al. Predicting monthly high – resolution PM2.5 concentrations with random forest model in the North China Plain[J]. Environmental Pollution, 2018, 242: 675 – 683.

[34] 谢志英,刘 浩,唐新明. 北京市 MODIS 气溶胶光学厚度与 PM10 质量浓度的相关性分析[J]. 环境科学学报, 2015(10): 3292 – 3299.
Xie Z Y, Liu H, Tang X M. Correlation analysis between MODIS aerosol optical depth and PM10 concentration over Beijing[J]. Acta Scientiae Circumstantiae, 2015(10): 3292 – 3299.

Estimating $PM_{2.5}$ concentrations in eastern coastal area of China using a two – stage random forest model

YANG Lijuan

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

Abstract: The aerosol optical depth (AOD) derived via dark – target algorithm has been widely used as an effective tool for estimating $PM_{2.5}$ concentrations. However, this algorithm cannot effectively retrieve AOD on the bright surface. Therefore, the authors used a random forest model incorporating meteorological parameters to predict the missing AOD values, and then employed a second – stage random forest model combining the retrieved AOD with meteorological parameters, vegetation cover and road density to estimate the $PM_{2.5}$ concentrations in two districts of eastern coastal zone of China, i.e., YRD and PRD. The result shows that the proposed model performed very well, achieving R^2 of 0.94 for AOD predictions and MODIS AOD and an overall R^2 of 0.97 with RMSE being only 5.57 $\mu g/m^3$ between the estimated and observed $PM_{2.5}$ concentrations. The spatial distribution of $PM_{2.5}$ concentrations suggests that the high values are mainly located in Jiangsu Province with low elevation ($\geqslant$40 $\mu g/m^3$). The results indicate that the proposed two – stage random forest model incorporated with satellite AOD and other variables could be effectively used for estimating the ground – level $PM_{2.5}$ concentrations.

Keywords: random forest model; $PM_{2.5}$ distribution; AOD retrieval; YRD; PRD

(责任编辑: 李 瑜)