

基于卫星遥感和气象再分析资料的 北京市 $\text{PM}_{2.5}$ 浓度反演研究

邵琦, 陈云浩, 李京*

(北京师范大学环境遥感与数字城市北京市重点实验室, 北京 100875)

摘要:为获取城市尺度空间连续准确的 $\text{PM}_{2.5}$ 浓度, 该文以北京市为研究区, 充分考虑气溶胶光学厚度和气象因素的影响, 利用 2017 年 MODIS 3 km AOD 产品和 ECMWF-ERA5 气象再分析资料, 结合空气质量监测站点的 $\text{PM}_{2.5}$ 数据, 分别基于随机森林、多元线性回归、支持向量机和神经网络方法反演近地面的 $\text{PM}_{2.5}$ 浓度, 结果表明: 随机森林的反演精度最高、效果最好, 其平均绝对误差(MAE)为 $12.08 \mu\text{g}/\text{m}^3$ 、均方根误差(RMSE)为 $18.42 \mu\text{g}/\text{m}^3$; 对自变量进行重要性分析, 气溶胶光学厚度对 $\text{PM}_{2.5}$ 反演模型的影响最大, 其次为相对湿度、边界层高度和 2 m 气温。利用随机森林模型反演非采暖期(4—11月)的 $\text{PM}_{2.5}$ 浓度, 并进行时空分析, 结果表明: $\text{PM}_{2.5}$ 月均浓度差异显著; 空间分布呈现出中部、南部和东部较高, 北部、东北部和西南部较低的特点; 变化幅度呈现出中部和南部较大, 西北部和西南部较小的分布特点。

关键词: $\text{PM}_{2.5}$; 卫星遥感; 气象再分析资料; 机器学习; 时空分析

中图分类号: TP79; X513 **文献标识码:** A **文章编号:** 1672-0504(2018)03-0032-07

0 引言

近年来, 我国北方大中城市空气污染问题日益凸显, $\text{PM}_{2.5}$ 等颗粒物浓度不断上升, 成为空气污染的首要污染物^[1]。 $\text{PM}_{2.5}$ 化学成分复杂, 易吸附有毒有害物质, 悬浮时间长且难以自然沉降^[2], 其对可见光的消光作用(散射和吸收)会显著降低大气能见度。此外, 进入人体内的 $\text{PM}_{2.5}$ 大部分沉积在支气管和肺泡中, 会导致心血管和哮喘等疾病, 对人体健康造成极大的危害^[3,4]。目前, $\text{PM}_{2.5}$ 的监测方法主要有地面监测、模型模拟和遥感监测。地面监测虽然可精确获取实时的数据, 但监测成本昂贵、站点稀少且分布不均, 无法获取连续广域的时空分布数据, 难以及时、准确地反映 $\text{PM}_{2.5}$ 的时空演变趋势^[5,6]。模型模拟主要利用大气物理化学传输模型模拟 $\text{PM}_{2.5}$ 的扩散、迁移与转化过程, 进而得到其浓度分布结果。虽然模型的模拟预测能力强, 但需要昂贵的软硬件支持, 且预测精度受制于输入数据的准确性和完整性, 不利于大范围推广使用^[7]。卫星遥感技术具有空间覆盖范围广、观测时间连续和数据获取精确、低廉等优点, 常用于大范围 $\text{PM}_{2.5}$ 的反演, 可有效弥补站点监测和模型模拟的不足。

遥感反演 $\text{PM}_{2.5}$ 浓度主要是以气溶胶光学厚度

(AOD)和气象数据为自变量, 建立线性回归模型估算 $\text{PM}_{2.5}$ 的浓度^[8-13]。但 $\text{PM}_{2.5}$ 的形成、扩散、迁移及转化过程复杂多变, 非线性特征显著, 简单的线性模型难以准确估算其浓度。机器学习算法(如随机森林、支持向量机和神经网络等方法)实现简单, 计算速度快, 对异常值和噪声容忍度高, 非常适用于非线性问题的求解。近年来, 许多学者尝试使用机器学习算法对 $\text{PM}_{2.5}$ 浓度进行反演, 取得了良好的效果。Hu等^[14]利用美国大陆监测的 $\text{PM}_{2.5}$ 数据、MODIS 10 km 的 AOD 数据、NLDAS 气象模式资料和土地覆盖等数据, 建立了基于随机森林的 $\text{PM}_{2.5}$ 反演模型, 得到了美国大陆 2011 年的日均 $\text{PM}_{2.5}$ 浓度。李丽娟^[15]采用 MODIS 10 km 的 AOD 数据、GOES 气象模式数据及 $\text{PM}_{2.5}$ 地面监测数据, 应用随机森林算法反演得到京津冀地区春夏秋冬四季的 $\text{PM}_{2.5}$ 浓度。以上研究虽然可以获取空间连续的 $\text{PM}_{2.5}$ 浓度, 但反演结果的空间分辨率较粗, 仅适用于大尺度区域, 不适合城市尺度。

为获取城市尺度空间连续准确的 $\text{PM}_{2.5}$ 浓度, 本文充分考虑 AOD 和气象因素的影响, 采用 MODIS 3 km 的 AOD 产品、欧洲中期天气预报中心(ECMWF)的 ERA5 气象再分析资料以及空气质量监测站点的 $\text{PM}_{2.5}$ 数据, 分别基于随机森林、多元线性回归、

收稿日期: 2018-03-01; 修回日期: 2018-04-22

基金项目: 住房和城乡建设部科学技术计划北京建筑大学北京未来城市设计高精尖创新中心开放课题(UDC2017030212、UDC201650100); 水资源安全北京实验室资助项目

作者简介: 邵琦(1994—), 女, 硕士研究生, 研究方向为遥感应用。* 通讯作者 E-mail: lijing@bnu.edu.cn

支持向量机和神经网络方法反演近地面的PM_{2.5}浓度;对比4种模型的反演结果,从中选取效果最优的模型反演北京市2017年非采暖期(4—11月)的PM_{2.5}浓度,并进行时空分析。

1 研究区及数据

1.1 研究区概况

北京市位于华北平原北部,地理坐标为115.7°~117.4°E,39.4°~41.6°N。西部、北部和东北部被太行山和燕山环绕,东南部是北京小平原,地势呈现为西北高、东南低。受蒙古高压的控制,北京市具有典型的温带大陆性季风气候特征^[3],夏季高温多雨,冬季寒冷干燥,春秋短促。多年平均气温为12℃,年平均风速为2.4 m/s,以偏北风为主^[16]。受冬季采暖、春季风沙、施工扬尘、机动车尾气以及周边重工业基地等污染源因素的影响,北京市空气污染天气频发^[11],引起社会公众的普遍关注。

1.2 数据

1.2.1 遥感数据 本文选用的遥感数据为美国NASA网站(<https://ladsweb.modaps.eosdis.nasa.gov/>)公开的MODIS Terra和Aqua Level 2级气溶胶产品数据——MOD04_3K/MYD04_3K。产品空间分辨率为3 km,版本为C006,采用的反演算法为暗像元法^[17]。该方法利用浓密植被(暗像元)在近红外与红、蓝通道的反射率呈线性关系的特性,借助6S辐射传输模型计算暗像元的地表反射率,进而反演气溶胶光学厚度^[18]。冬季植被覆盖率低,暗像元少,所以AOD值大部分缺失。选用北京市2017年1—12月的AOD产品数据,利用ENVI的MCTK插件对其进行几何校正、投影转换和裁剪等预处理。

1.2.2 气象再分析资料 本文选用的气象再分析资料为ECMWF公开的ERA5数据(<http://apps.ecmwf.int/datasets/>),ERA5是迄今最强大的全球气候监测数据集,与之前的ERA-Interim产品相比,ERA5具有更高的时空分辨率,时间分辨率为逐小时,空间分辨率高达0.125°。研究表明:ECMWF再分析资料的精度高于美国国家环境预报中心(NCEP)的再分析资料^[19,20]。本文选用北京市2017年1—12月0.125°的ECMWF-ERA5气象模式数据,气象变量包括边界层高度(BLH)、相对湿度(RH)、2 m气温(T2m)、地面气压(SP)、10 m经向风速和10 m纬向风速,进一步对10 m经向风速和10 m纬向风速做矢量化处理,得到10 m风速(WS10)

和10 m风向(WD10)。

1.2.3 PM_{2.5}数据 PM_{2.5}数据为北京市2017年1—12月空气质量监测站逐小时的污染物监测数据(<http://www.bjmemc.com.cn/>),研究区范围内共有35个站点,如图1所示。

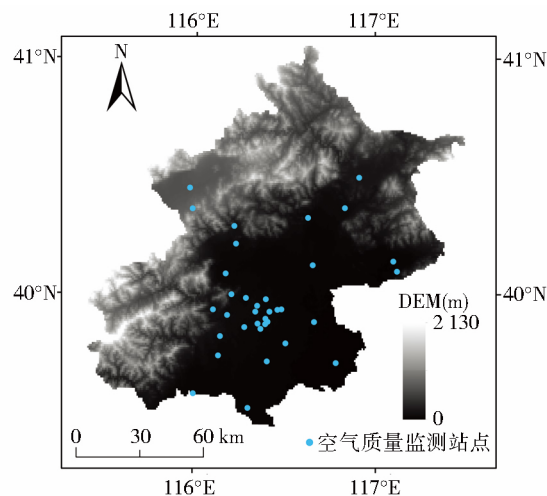


图1 北京市空气质量监测站点分布
Fig. 1 Distribution of air quality monitoring sites in Beijing

1.2.4 数据时空匹配 由于卫星过境时不是整点时刻,AOD数据的空间分辨率与ERA5数据不一致,需要对数据做时空匹配。以AOD数据为基准,取AOD过境时间前后2 h内的PM_{2.5}观测数据进行拉格朗日插值,得到AOD过境时间的PM_{2.5}浓度;取与AOD过境时间最邻近的ERA5数据作为AOD过境时间的气象数据;采用双线性内插方法对ERA5数据做降尺度处理,将其空间分辨率调整成与AOD数据一致。

2 研究方法

综合国内外研究^[21-24],选取AOD、BLH、RH、T2m、SP、WS10、WD10和月份(Month)作为PM_{2.5}反演模型的自变量。剔除数据集中AOD、气象因子和PM_{2.5}的缺失值及异常值,最终得到3 932个样本。将数据集按照7:3的比例随机划分为训练集和测试集,训练集用于建立模型,测试集用于精度评价。基于相同的训练集和测试集,选取随机森林、多元线性回归、支持向量机和神经网络方法对PM_{2.5}浓度进行建模反演,计算真实值和预测值的平均绝对误差(MAE)和均方根误差(RMSE),判断模型的优劣;从中选择最优模型反演北京市的PM_{2.5}浓度,并进行时空分析。技术路线如图2所示。

(1)随机森林(Random Forest, RF)是一种分类与回归技术,同时也是一种组合式的自学习技术^[25],其可通过对大量回归决策子树的汇总提高模型的预

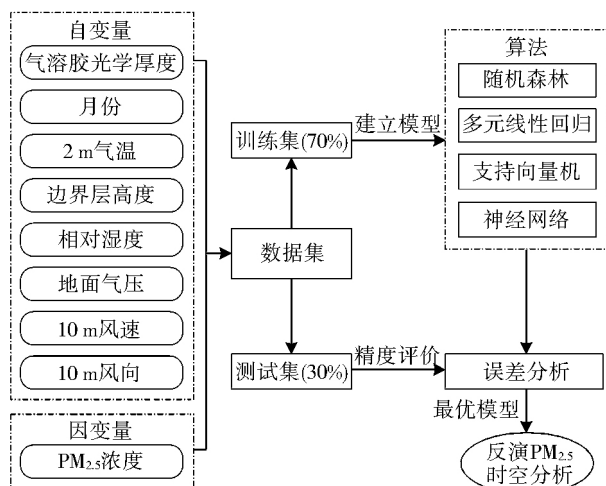


图 2 技术路线
Fig. 2 Flow chart of technical route

测精度^[26],同时避免了过拟合等问题;随机森林实现简单,计算开销小,对异常值和噪声容忍度高,非常适用于非线性数据建模,并可对变量进行重要性分析^[27]。随机森林模型的泛化能力和预测精度取决于决策树个数和特征选择个数。对训练集进行 10 折 10 次交叉验证,计算袋外数据错误率(Out of Bag Error),根据错误率最小准则确定模型参数。

(2)多元线性回归(Multiple Linear Regression, MLR)是以多个解释变量的给定值为条件的回归分析,利用线性方程拟合多个自变量和一个因变量的关系^[28],回归系数由最小二乘法确定。

(3)支持向量机(Support Vector Machine, SVM)是建立在统计学习理论和结构风险最小化原理基础上的机器学习方法,适用于非线性和高维数据建模^[29],其泛化能力和预测精度由核函数和参数 γ 确定,对训练集采用 10 折 10 次交叉验证,计算平均反演精度,根据平均精度最大化准则确定模型参数。

(4)神经网络(Neural Networks, NN)是由大量简单处理单元(称为神经元)互相连接而形成的复杂网络系统^[30],是一个高度复杂的非线性动力系统,适合处理多因素、多条件、不精确和模糊的问题。神经网络通过多层网络的传播和反馈,根据梯度下降最大化原则修正网络权重,求解最优解;其泛化能力和预测精度由神经元个数和隐藏层数量确定,对训练集采用 10 折 10 次交叉验证,计算平均反演精度,根据平均精度最大化准则确定模型参数。

3 结果与讨论

3.1 结果验证

对测试集进行精度评价,计算真实值和预测值的 MAE 和 RMSE,判断模型的优劣(图 3)。由图 3

可知,随机森林对 $PM_{2.5}$ 的反演精度最高、效果最好,支持向量机和神经网络次之,传统的多元线性回归反演精度最低、效果最差。多元线性回归大部分测试集在 1:1 线两侧分布略微散乱,在 $PM_{2.5}$ 真实值超过 $200 \mu g/m^3$ 的部分,低估现象严重,所有预测值均在 1:1 线下方,且距离较远,主要是因为多元线性回归对极端值敏感,极端值的存在会影响预测精度。支持向量机和神经网络在 $PM_{2.5}$ 真实值超过 $200 \mu g/m^3$ 的部分,预测均出现低估现象。随机森林的大部分测试集聚集在 1:1 线周围,有较高的拟合度。以 $100 \mu g/m^3$ 为界线,低于 $100 \mu g/m^3$ 的 $PM_{2.5}$ 比高于 $100 \mu g/m^3$ 的更加贴近 1:1 线,表明随机森林在 $PM_{2.5}$ 值较低时反演精度更好;在 $PM_{2.5}$ 真实值超过 $200 \mu g/m^3$ 的部分,随机森林的预测值大部分位于 1:1 线附近,少部分在 1:1 线下方,出现低估现象,预测误差比 $PM_{2.5}$ 低值部分大。此外,在 $PM_{2.5}$ 值较高时,无论是机器学习方法还是多元线性回归方法,预测值和真实值都相差较大,说明在 $PM_{2.5}$ 值较高时,反演误差相对较高。

3.2 变量重要性分析

随机森林算法采用袋外数据错误率对各个特征的重要性进行度量,得到各特征的重要性评分(图 4)。由图 4 可知,AOD 是最重要的输入参数,由于其反映了大气气溶胶粒子对入射辐射的散射和吸收作用,与颗粒物的浓度有密切关系,因此对模型的影响最大。RH、BLH 和 T2m 也是影响 $PM_{2.5}$ 浓度的重要因子,这 3 个参数主要影响气溶胶粒子的垂直分布和吸收差异。研究区内包含山地和平原,地形的影响导致气象因素空间差异显著,因此其重要性次于 AOD。Month 主要反映了不同时期各类污染源排放量及其对空气污染的贡献变化。WS10、WD10 和 SP 主要影响大气层的流动速度和方向,进而影响污染物的传输和扩散过程,相对而言,重要性低于 AOD、RH、BLH 和 T2m。

3.3 北京市非采暖期 $PM_{2.5}$ 分布

(1) $PM_{2.5}$ 时空分布。基于上文构建的随机森林模型,利用北京市 2017 年的 AOD 数据和 ERA5 气象要素数据,反演得到每日的 $PM_{2.5}$ 浓度,进而求得月均 $PM_{2.5}$ 浓度。由于 1 月、2 月、3 月和 12 月的 AOD 值缺失严重,无法得到这 4 个月的 $PM_{2.5}$ 浓度,将其剔除,得到北京市非采暖期(4—11 月)的 $PM_{2.5}$ 浓度(图 5)。从图 5 可以看出,5 月、6 月和 7 月 $PM_{2.5}$ 浓度的空间分布大致相同,高值区主要位于平原地区和延庆区南部,平原地区 $PM_{2.5}$ 浓度空间差异

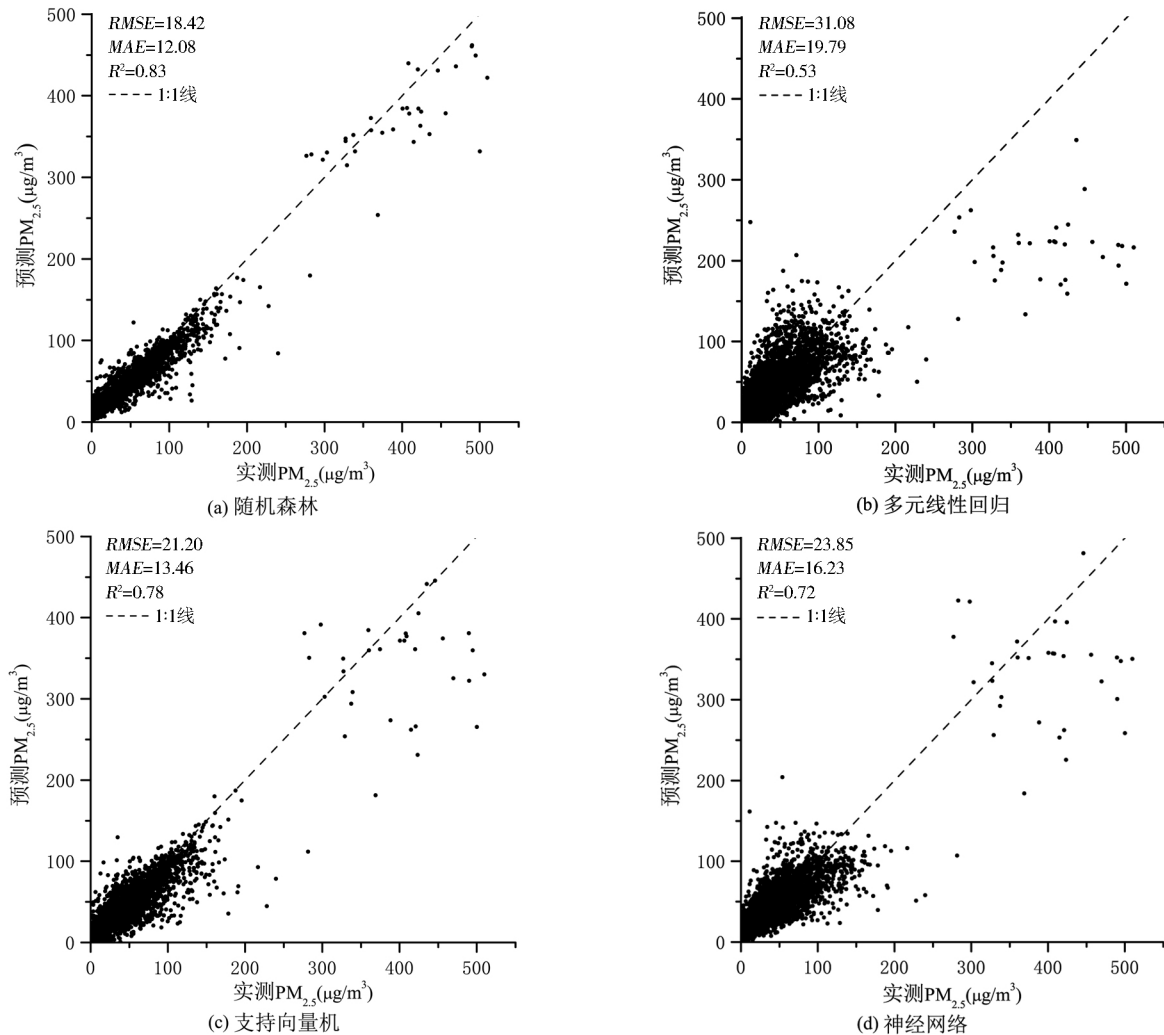


图3 模型反演结果对比
Fig. 3 Comparison of model inversion results

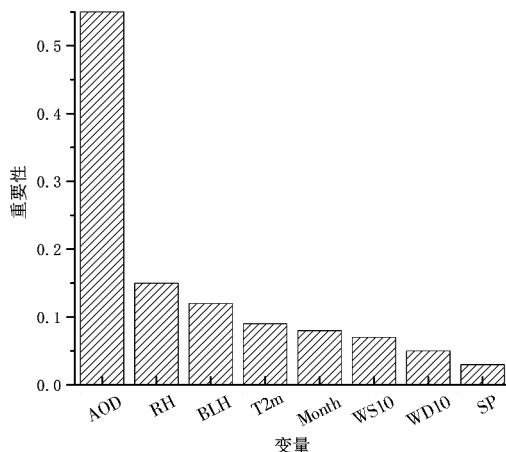
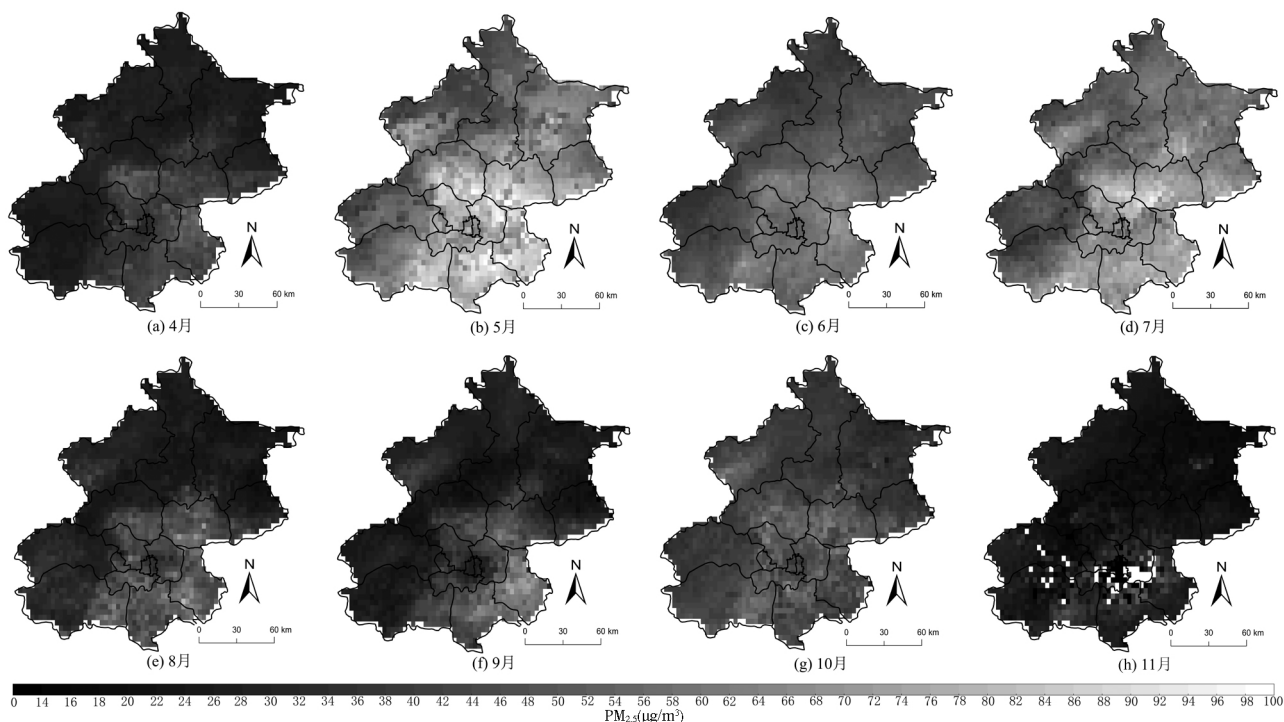


图4 随机森林算法变量重要性分析
Fig. 4 The importance of random forest variables

不大,分布较为均一;低值区主要位于西南部和北部的山地地区,空间分布差异显著,PM_{2.5}浓度由边缘向中心逐渐升高。4月、8月和9月PM_{2.5}浓度的高值区主要位于平原地区,延庆区南部有局部高值区,平原地区PM_{2.5}浓度空间差异显著,东城区和西城区

浓度较周边各区低,形成高值环绕、中心较低的空间布局;低值区主要位于北部、东北部和西南部的山地地区,空间分布较为均一。10月PM_{2.5}浓度的高值区主要位于平原地区和延庆区南部,低值区主要位于北部、东北部和西南部的山地地区以及东南部部分平原地区,高、低区PM_{2.5}浓度空间分布差异不大。11月的PM_{2.5}浓度最低,中心城区有所缺失,空间分布差异不大。逐月对比PM_{2.5}浓度可知,5月PM_{2.5}浓度最高,平原地区的PM_{2.5}浓度基本在80 μg/m³以上,山区的PM_{2.5}浓度基本在50 μg/m³以上。6月、7月和10月次之,8月和9月较低,4月和11月最低,是空气质量的优良时段。

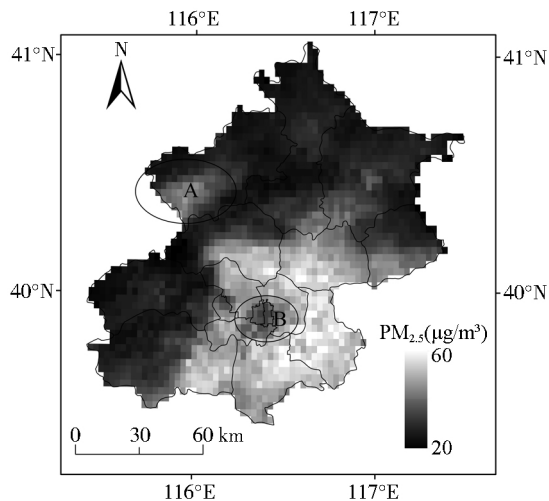
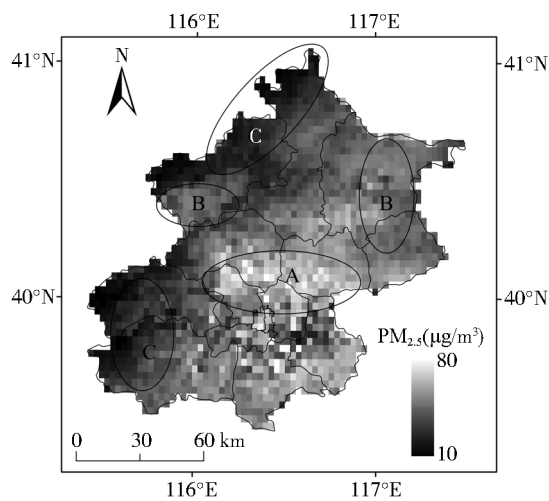
(2)PM_{2.5}均值分布。北京市非采暖期PM_{2.5}浓度均值的空间分布(图6)呈现出中部、南部和东部较高,北部、东北部和西南部较低的特点,高、低区的分界线大致与山地和平原的分界线相一致。中部、南部和东部主要为平原地区,人口、工业生产、交通较为密集,污染气体排放较多,加之建筑物及山地地形

图 5 2017 年北京市非采暖期 $PM_{2.5}$ 浓度空间分布Fig. 5 Distribution of monthly mean concentration of $PM_{2.5}$ in Beijing during the non-heating period in 2017

的阻挡,空气流动性差,污染物不易扩散, $PM_{2.5}$ 浓度较高;例如,大兴区、通州区、顺义区和昌平区分布着大量的重工业工厂,污染气体排放严重,北部的燕山阻挡了南下和北上的大气流动,导致污染物无法扩散,形成高浓度区域。北部、东北部和西南部主要为山地地区,植被覆盖率高,人口稀疏,地表较为空旷,空气流动性好,污染物排放少且扩散较快, $PM_{2.5}$ 浓度较低,如延庆区、密云区、怀柔区、门头沟区和房山区。另外,延庆区南部的井庄镇和大榆树镇(图 6 椭圆形区域 A)海拔较高,人口稀疏,经济发展水平相对滞后,以轻工业为主,污染气体排放较

多,山地地形和气候的影响容易形成山谷风环流和辐射逆温,不利于污染物扩散,导致 $PM_{2.5}$ 的浓度高于周边地区,形成局部高值区;东城区和西城区(图 6 椭圆形区域 B)的 $PM_{2.5}$ 浓度低于周边地区,形成高值环绕、中心较低的空间布局。

(3) $PM_{2.5}$ 变化幅度分析。北京市 $PM_{2.5}$ 变化幅度空间差异显著(图 7):中部和南部地区 $PM_{2.5}$ 浓度季节性变化明显,变化幅度大多在 $45 \mu g/m^3$ 以上,如顺义区和昌平区(图 7 椭圆形区域 A);东北部和西部次之,变化幅度多在 $35 \sim 45 \mu g/m^3$ 之间,如密云区、平谷区和延庆区南部(图 7 椭圆形区域 B);西

图 6 2017 年北京市非采暖期 $PM_{2.5}$ 浓度平均值
Fig. 6 Distribution of the mean concentration of $PM_{2.5}$ in Beijing during the non-heating period in 2017图 7 2017 年北京市非采暖期 $PM_{2.5}$ 浓度变化幅度
Fig. 7 The concentration variation of $PM_{2.5}$ in Beijing during the non-heating period in 2017

北部和西南部变化幅度较小,均在 $35 \mu\text{g}/\text{m}^3$ 以下,如门头沟区、房山区、延庆区和怀柔区(图7椭圆形区域C)。由 PM_{2.5} 浓度最大值分布月份(图8)可知,北京市大部分地区 PM_{2.5} 浓度最大值出现在5月,延庆区、怀柔区和平谷区(图8椭圆形区域A)最大值主要出现在7月。

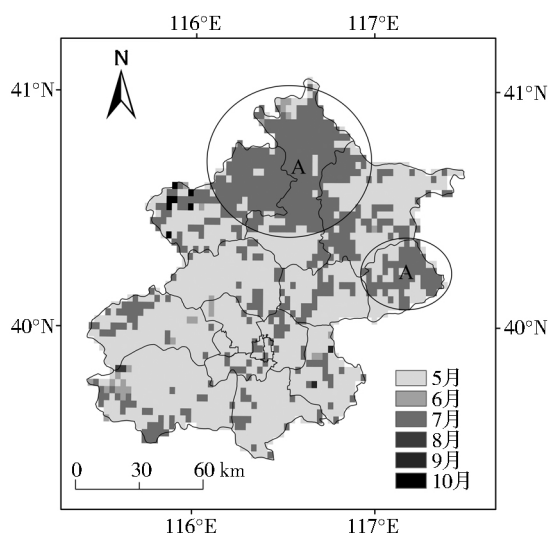


图8 2017年北京市非采暖期 PM_{2.5} 浓度最大值月份
Fig. 8 The month with maximum concentration of PM_{2.5} in Beijing during the non-heating period in 2017

4 结论与展望

本文利用 MODIS AOD 数据和 ERA5 气象再分析资料反演城市尺度近地面的 PM_{2.5} 浓度,结果表明:1)随机森林反演 PM_{2.5} 浓度的 MAE 为 $12.08 \mu\text{g}/\text{m}^3$, RMSE 为 $18.42 \mu\text{g}/\text{m}^3$,与多元线性回归(MAE 为 $19.79 \mu\text{g}/\text{m}^3$, RMSE 为 $31.08 \mu\text{g}/\text{m}^3$)、支持向量机(MAE 为 $13.46 \mu\text{g}/\text{m}^3$, RMSE 为 $21.20 \mu\text{g}/\text{m}^3$)和神经网络(MAE 为 $16.23 \mu\text{g}/\text{m}^3$, RMSE 为 $23.85 \mu\text{g}/\text{m}^3$)相比,随机森林的反演精度最高、效果最好。2)模型自变量中,AOD 是最重要的影响因素,其次是 RH、BLH 和 T2m。3)北京市非采暖期 PM_{2.5} 浓度差异显著,5月浓度最高,4月和11月浓度最低;PM_{2.5} 平均浓度空间分布呈现出中部、南部和东部较高,北部、东北部和西南部较低的特点;PM_{2.5} 浓度变化幅度呈现出中部和南部较大,东北部和西部次之,西北部和西南部较小的分布特点;PM_{2.5} 浓度最大值主要出现在5月和7月。

MODIS 的 AOD 产品冬季缺失严重,无法反演得到该时期的 PM_{2.5} 浓度,并且卫星每天仅过境2次,应用的时间局限性较大。Lee 等^[31]采用多角度大气校正算法(Multiangle Implementation of Atmospheric Correction, MAIAC)反演得到 1 km 的

AOD 数据,具有更高的空间分辨率,但该算法所需参数众多,计算复杂,尚未业务化;中国气象局研发的“CLDASV2.0”1 km 的气象再分析资料在中国区域的质量优于国际同类产品。如果将两者作为自变量引入随机森林模型,将会得到时空分辨率更高、更加精确的 PM_{2.5} 浓度,有望对城市空气质量进行近实时精确的监测。

参考文献:

- [1] 霍彦峰,杨关盈,白庆梅,等.基于 MODIS 气溶胶产品的安徽省颗粒物浓度反演研究[J].环境科学与技术,2017,40(2):59-64.
- [2] 王嫣然,张学霞,赵静瑶,等.北京地区不同季节 PM_{2.5} 和 PM₁₀ 浓度对地面气象因素的响应[J].中国环境监测,2017,33(2):34-41.
- [3] 林海峰,辛金元,张文煜,等.北京市近地层颗粒物浓度与气溶胶光学厚度相关性分析研究[J].环境科学,2013,34(3):826-834.
- [4] 李成才,毛节泰,刘启汉,等. MODIS 卫星遥感气溶胶产品在北京市大气污染研究中的应用[J].中国科学 D 辑(地球科学),2005(35):177-186.
- [5] 于雪,赵文吉,孙春媛,等.大气 PM_{2.5} 遥感反演研究进展[J].环境污染与防治,2017,39(10):1153-1158.
- [6] 薛文博,武卫玲,许艳玲,等.基于 WRF 模型与气溶胶光学厚度的 PM_{2.5} 近地面浓度卫星反演[J].环境科学研究,2016,29(12):1751-1758.
- [7] 吕明明.基于 MODIS 数据的 PM_{2.5} 浓度遥感反演——以京津冀城市群为例[D].北京:中国地质大学,2017.
- [8] TIAN J, CHEN D M. A semi-empirical model for predicting hourly ground-level fine particulate matter (PM_{2.5}) concentration in southern Ontario from satellite remote sensing and ground-based meteorological measurements[J]. Remote Sensing of Environment, 2010, 114(2): 221-229.
- [9] LIU Y, SARNAT J A, KILARU V, et al. Estimating ground-level PM_{2.5} in the eastern United States using satellite remote sensing[J]. Environment Science & Technology, 2005, 39(9): 3269-3278.
- [10] DONKELAAR A V, MARTIN R V, PARK R J. Estimating ground-level PM_{2.5} using aerosol optical depth determined from satellite remote sensing[J]. Journal of Geophysical Research: Atmospheres, 2006, 111(D21): 5049-5066.
- [11] 李军,孙春宝,刘威德,等.气象因素对北京市大气颗粒物浓度影响的非参数分析[J].环境科学研究,2009,22(6):663-669.
- [12] 李宗恺,潘云仙,孙澜桥.空气污染气象学原理及应用[M].北京:气象出版社,1985.557-569.
- [13] 赵佳楠,徐建华,卢德彬,等.基于 RF-LUR 模型的 PM_{2.5} 空间分布模拟——以长江三角洲地区为例[J].地理与地理信息科学,2018,34(1):18-23.
- [14] HU X F, BELLE J H, 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: 6936-6944.
- [15] 李丽娟.基于卫星遥感的京津冀地区 PM_{2.5} 反演研究[D].北

- 京:中国地质大学,2017.
- [16] 于淑秋,林学椿,徐祥德.北京市区大气污染的时空特征[J].应用气象学报,2002,13(特刊):92—99.
- [17] KAUFMAN Y J, SENDRA C. Algorithm for automatic atmosphere corrections to visible and near-IR satellite imagery[J]. International Journal of Remote Sensing, 1988, 9(8): 1357—1381.
- [18] 周寅.长三角地区大气气溶胶光学厚度遥感及时空变化规律研究[D].南京:南京信息工程大学,2014.
- [19] 陈钦明,宋淑丽,朱文耀.亚洲地区 ECMWF/NCEP 资料计算 ZTD 的精度分析[J].地球物理学报,2012,55(5):1541—1548.
- [20] DECKER M, BRUNKE M A, WANG Z, et al. Evaluation of the reanalysis products from GSFC, NCEP, and ECMWF using flux tower observations[J]. Journal of Climate, 2012, 25(25): 1916—1944.
- [21] 周丽,徐祥德,丁国安,等.北京地区气溶胶 $PM_{2.5}$ 粒子浓度的相关因子及其估算模型[J].气象学报,2003,61(6):761—768.
- [22] 苏维,张帅珺,赖新云,等.南昌市空气 $PM_{2.5}$ 和 PM_{10} 的时空动态及其影响因素[J].应用生态学报,2017,28(1):257—265.
- [23] TAI A P K, MICKLEY L J, JACOB D J. Correlations between fine particulate matter ($PM_{2.5}$) and meteorological variables in the United States: Implications for the sensitivity of $PM_{2.5}$ to climate change[J]. Atmospheric Environment, 2010, 44(32): 3976—3984.
- [24] 孟晓艳,魏桢,王瑞斌,等.灰霾试点城市 $PM_{2.5}$ 浓度特征及其影响因素分析[J].环境科学与技术,2013,36(9):76—80.
- [25] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5—32.
- [26] 李亨,田原,邬伦,等.基于随机森林方法的滑坡灾害危险性区划[J].地理与地理信息科学,2014,30(6):25—30.
- [27] 张雷,王琳琳,张旭东,等.随机森林算法基本思想及其在生态学中的应用——以云南松分布模拟为例[J].生态学报,2014,34(3):650—659.
- [28] 张景阳,潘光友.多元线性回归与 BP 神经网络预测模型对比与运用研究[J].昆明理工大学学报(自然科学版),2013,38(6):61—67.
- [29] 丁世飞,齐丙娟,谭红艳.支持向量机理论与算法研究综述[J].电子科技大学学报,2011,40(1):2—10.
- [30] 焦李成,杨淑媛,刘芳,等.神经网络七十年:回顾与展望[J].计算机学报,2016,39(8):1697—1716.
- [31] LEE M, KLOOG I, CHUDNOVSKY A, et al. Spatiotemporal prediction of fine particulate matter using high-resolution satellite images in the Southeastern US 2003—2011[J]. Exposure Science and Environmental Epidemiology, 2016, 26(4): 377—384.

Inversion of $PM_{2.5}$ Concentration in Beijing Based on Satellite Remote Sensing and Meteorological Reanalysis Data

SHAO Qi, CHEN Yun—hao, LI Jing

(Beijing Key Laboratory of Environmental Remote Sensing and Digital City,
Beijing Normal University, Beijing 100875, China)

Abstract: In order to obtain continuous and accurate spatial $PM_{2.5}$ concentration, this paper takes Beijing as the study area to retrieve $PM_{2.5}$ concentration based on a variety of methods including the random forest (RF), the multiple linear regression (MLR), the support vector machine (SVM) and the neural network (NN), using MODIS 3-km AOD products, ECMWF-ERA5 meteorological reanalysis data and $PM_{2.5}$ data at air quality monitoring stations. The results show that RF method has the highest inversion accuracy, whose mean absolute error (MAE) and root mean square error (RMSE) are $12.08 \mu\text{g}/\text{m}^3$ and $18.42 \mu\text{g}/\text{m}^3$, respectively. Based on the analysis of the significance of the variables, aerosol optical depth has the greatest impact on the model, followed by the relative humidity, the boundary layer height and the air temperature. The RF model was used to retrieve the $PM_{2.5}$ concentration during the non-heating period (April–November) in 2017, and the spatial-temporal analysis shows that the monthly average concentrations of $PM_{2.5}$ are significantly different. The spatial distribution shows higher $PM_{2.5}$ concentration in central, southern and eastern areas, and lower in north, northeast and southwest areas. In addition, the change magnitude of $PM_{2.5}$ concentration is larger in central and southern areas, smaller in northwest and southwest areas.

Key words: $PM_{2.5}$; satellite remote sensing; meteorological reanalysis data; machine learning; spatial-temporal analysis