基于随机森林的高分辨率 $PM_{2.5}$ 遥感反演——以广东省为例

申 原 陈朝亮 钱 静 刘 军

（中国科学院深圳先进技术研究院 深圳 518055）

摘 要 细颗粒物（$PM_{2.5}$）监测是大气污染治理的重要手段，受限于地面观测点的数量，从遥感反演 $PM_{2.5}$ 是常规地面观测的有效补充，是当前的研究热点。通常遥感反演 $PM_{2.5}$ 的思路是先反演大气气溶胶光学厚度，然后基于统计关系由大气气溶胶光学厚度反演 $PM_{2.5}$。该方法容易造成误差传递，从而导致反演模型的不稳定。该文提出了一种基于随机森林算法（一种机器学习算法）的 $PM_{2.5}$ 遥感反演方法，直接建立中分辨率成像光谱仪（Moderate Resolution Imaging Spectroradiometer，MODIS）影像与地面实测 $PM_{2.5}$ 的关系，可以避免传统反演 $PM_{2.5}$ 时先反演大气气溶胶光学厚度带来的误差，最终得到精度更高的 $PM_{2.5}$ 反演结果。该方法先用随机森林算法对 MODIS 影像和经过克里金插值后的地面监测站 $PM_{2.5}$ 数据进行训练和测试；然后，根据测试的均方根误差从多个模型中选取最优（均方根误差最小）的模型；最后，将此模型用于整幅 MODIS 影像，得到整个区域的 $PM_{2.5}$ 反演结果。实验选取了广东省四个季节多幅 MODIS 影像数据进行验证，并通过决定系数和均方根误差两个表现指标进行对比和分析，验证了所提算法的优越性。

关键词 随机森林；机器学习；$PM_{2.5}$ 遥感反演；克里金插值；均方根误差

中图分类号 S 127 **文献标志码** A

High Resolution $PM_{2.5}$ Estimation Using Remote Sensing Data Based on Random Forest— a Case Study of Guangdong, China

SHEN Yuan CHEN Chaoliang QIAN Jing LIU Jun

(*Shenzhen Institutes of Advanced Technology*, *Chinese Academy of Sciences*, *Shenzhen* 518055, *China*)

Abstract $PM_{2.5}$ monitoring is an important means of air pollution control. Limited by the number of ground observation points, $PM_{2.5}$ estimation from remote sensing data is an effective complement to conventional ground observation. The key idea of remote sensing estimation of $PM_{2.5}$ is to retrieve aerosol optical depth firstly, and subsequently to reverse $PM_{2.5}$ by aerosol optical depth based on the statistical relationship. This approach however is highly possible to cause error transmission, leading to instability of the inversion model. In this paper, we propose a $PM_{2.5}$ remote sensing estimation method based on random forest algorithm to

收稿日期：2018-02-24 **修回日期**：2018-03-15

基金项目：深圳市科技创新委员会基础研究项目（JCYJ20150831194441446）

作者简介：申原，硕士研究生，研究方向为遥感影像处理；陈朝亮，硕士研究生，研究方向为环境遥感；钱静，博士，副研究员，研究方向为环境遥感；刘军（通讯作者），博士，副研究员，研究方向为环境遥感与高性能地学计算，E-mail：jun.liu@siat.ac.cn。

directly establish the relationship between moderate resolution imaging spectroradiometer (MODIS) image and ground measured $PM_{2.5}$, so as to avoid the inversion error of atmospheric aerosol optical depth, finally obtain the $PM_{2.5}$ estimation result with high precision. The method first uses random forest to train and test the MODIS image and ground monitoring station $PM_{2.5}$ data after kriging interploation, and then selects the best model from multiple models according to the root mean square error (RMSE) of test index. Finally, the approach uses this model in the whole MODIS image to obtain the $PM_{2.5}$ estimation result of the whole area. This experiment selects many MODIS image data from four seasons in Guangdong province to verify and compare the two performance indicators of R^2 and RMSE. The results show that the proposed approach outperforms other approaches significantly.

Keywords random forest; machine learning; PM_{25} estimation using remote sensing data; kriging interpolation; root-mean-square error

1 引 言

随着我国工业化、城市化进程不断加快，环境污染越来越严重，与居民生活密切相关的空气质量指数成为社会关注的热点话题。细颗粒物($PM_{2.5}$)是衡量空气质量的重要参数指标，采用科学的方法监测 $PM_{2.5}$ 的分布和浓度，对研究其本身的理化特性、揭示雾霾成因及采取正确的防治措施具有重要意义。

目前，$PM_{2.5}$ 的监测方法主要包括地面监测和卫星遥感监测两种[1]。虽然地面监测结果较精确，但由于成本较高、地面监测站数量少，导致监测结果时空不连续，无法获得足够多的数据来研究整个区域 $PM_{2.5}$ 的扩散方式和传输特性[2,3]。卫星遥感监测具有数据获取方便、监测范围广等优势，能很好地弥补地面监测的不足。现有的 $PM_{2.5}$ 监测反演方法都是先反演大气气溶胶光学厚度(Aerosol Optical Depth，AOD)，再对气溶胶光学厚度与地面实测 $PM_{2.5}$ 的关系进行统计分析，用统计得到的关系来推出无地面监测点区域的 $PM_{2.5}$ 值。国内外很多学者用此方法进行了大量的研究，其基本假设是，AOD 与 $PM_{2.5}$ 具有良好的、稳定的统计关系。因此，现有大量研究都集中在提高 AOD 反演精度上，如引入各种订正、加入更多辅助数据、结合数值预报模式等，也在特定研究区域内取得了比较好的结果。如陈辉等[4]利用地理加权回归模型建立了我国冬季的 AOD-$PM_{2.5}$ 模型；王子峰[5]系统地研究了卫星遥感估算近地面颗粒物浓度的算法；王中挺等[6]利用暗目标和高分一号卫星 16 m 相机数据反演了京津地区的气溶胶光学厚度；王静等[7]研究了北京市中分辨率成像光谱仪(Moderate Resolution Imaging Spectroradiometer，MODIS)气溶胶光学厚度和 $PM_{2.5}$ 质量浓度的特征及其相关性；Song 等[8]利用 MODIS C5 AOD 产品反演出了珠江三角洲地区的 $PM_{2.5}$ 浓度；张天棋[9]则验证了基于 MODIS 影像的中国地区气溶胶产品与 $PM_{2.5}$ 反演的关系；Chu 等[10]基于 MODIS AOD 产品，反演出中国台湾北部的 $PM_{2.5}$ 浓度等。但由于在反演气溶胶光学厚度 AOD 过程中会产生误差，用带有误差的 AOD 与地面实测 $PM_{2.5}$ 建立统计关系时会导致误差的传递，从而影响最终 $PM_{2.5}$ 的反演精度。

在反演 AOD 方面，一般用 MODIS 数据。基本思路是通过假设不同的气溶胶模式和观测几何状况，再计算气溶胶光学厚度与大气下界的半球反射率、大气反射率和卫星天顶角与太阳天顶角的 cos 余弦值之间的对应关系。据此建立查找

表，通过动态气溶胶模式的输入来查算气溶胶光学厚度。Griggs[11]在大气层顶平行且无云等假设前提下，根据模拟大气辐射传输模型，发现了 AOD 与红外波段和可见光波段的相关关系。Levy 等[12]将中红外波段的气溶胶信息加入到反演过程中，将红蓝两波段独立反演改进为红、蓝、中红外三波段同时反演，并更新了原有的大气气溶胶模式和 AOD 反演查找表。除了 MODIS 数据，也可借助于其他数据反演气溶胶光学厚度，如 Holben 等[13]用先进超高分辨率辐射计数据和对比法反演了马里萨赫勒地区的 AOD，反演误差在 0.1 左右；Isakov 等[14]利用机载可见红外成像光谱仪数据，反演了美国俄克拉荷马州和拉皮德城两个地点的 AOD，结果发现，当地表反射率差异大于 0.5 时，AOD 反演精度可达 0.9。上述反演方法在实际运用中均取得了较好的成果，但由于用到了很多辅助数据，使得计算精度难以控制。另外，AOD 数据在反演过程中本身就存在误差，用上述方法无法避免误差的传递[15]。因此，如何减小误差、获得更高的反演精度，一直是近年的研究热点[16,17]。

但是，目前 $PM_{2.5}$ 遥感反演方法存在以下三个方面的问题：

(1) AOD 与 $PM_{2.5}$ 关系的稳定性。这是通过 AOD 反演 $PM_{2.5}$ 的基本假设，大量研究结果表明，AOD 与 $PM_{2.5}$ 存在一定的统计相关性，但是这个相关性在不同的区域有所不同，在同一区域的不同时间也可能不同。因此，针对特定区域、特定时间的 $PM_{2.5}$ 反演，该关系的稳定性起着至关重要的作用。

(2) 误差传递过程。通过建立各种精细的物理模型，提高 AOD 反演精度，从而能够更精确地反演 $PM_{2.5}$，但仍然存在一个误差传递的过程。反演 AOD 是有误差的，从带误差的 AOD 反演 $PM_{2.5}$ 仍然会存在误差，因此，误差传递过程可能会导致某些区域的 $PM_{2.5}$ 反演精度偏低。现在也有大量学者通过日平均、月平均、季平均、年平均等尺度研究 AOD 与 $PM_{2.5}$ 的关系，这在一定程度上能够抵消误差传递过程带来的偏差。但在某一时刻尺度上，这个误差传递过程造成的影响可能更大。

(3) 模型的适用性。通过引入各种订正、加入更多的辅助数据、结合数值模式等能够提高 AOD 反演的精度。但是，加入更多的因子，就意味着引入了更多的不确定性，对模型的适用性提出了更严格的要求。因此，同样的方法，换一个研究区，效果可能会变得很差。

针对上述问题，本文提出了一种基于随机森林机器学习法与 MODIS 影像相结合的 $PM_{2.5}$ 遥感反演方法。从 MODIS 遥感数据出发，通过机器学习的手段，直接建立遥感影像本身与实测 $PM_{2.5}$ 的关系，以避免误差的传递。初步实验结果表明，反演的结果与地面实测 $PM_{2.5}$ 具有较好的相关性。

2　MODIS 的气溶胶光学厚度反演算法

TERRA 和 AQUA 是美国地球观测系统计划中的两颗重要卫星，它们搭载的 MODIS 扫描宽度为 2 330 km，具有 36 个光谱波段，波长范围 0.14～14 μm，空间分辨率 0.25～1 km。MODIS 以其高时空分辨率、多通道、覆盖范围广等优点被广泛应用于气溶胶光学厚度的反演。

MODIS AOD 算法自问世以来经过多次改进，现已更新到 C6 版本。在 2008 年发布的 C5 版本中，暗目标法(Dark Target Algorithm，DT)，又称暗像元法，只用于暗目标地区的反演；深蓝算法(Deep Blue Algorithm，DB)只用于反演亮目标区域。DT 与 DB 反演结果不做融合，只提供分辨率 10 km 的气溶胶产品。在 2012 年发布的 C6 版本中，DT 与 DB 反演结果进行融合，并且基于 DT 算法反演的 AOD 产品

分辨率可达 3 km。

反演气溶胶光学厚度的基本原理[1]是，假定观测表面是均匀的朗伯面，建立大气顶层辐射亮度值与表面反射率关系，在不考虑大气吸收情况下，卫星接收的辐射值为：

$$\rho_a(\tau,\sigma,\phi)=\rho_b(\tau,\sigma,\phi)+\frac{T_\tau T_\sigma \rho(\tau,\sigma,\phi)}{\left[1-\rho_S(\tau,\sigma,\phi)S\right]} \quad (1)$$

其中，ρ_a 为卫星观测表面的反射率；ρ_b 为大气反射率；$\tau=\cos\varphi$，$\sigma=\cos\omega$，φ 和 ω 分别为卫星天顶角和太阳天顶角；ϕ 为卫星与太阳之间的相对方位角；S 为大气下界的半球反射率；T 为大气透过率。

当地表反射率很小时，卫星观测反射率主要取决于大气分子和气溶胶散射发生的反射率。在反演过程中，先假设不同的气溶胶模型和不同的观测几何状况，再计算气溶胶光学厚度与大气下界的半球反射率、大气反射率及 T_τ、T_σ 之间的对应关系。据此建立查找表，通过动态气溶胶模式的输入来查算气溶胶光学厚度。

陆地上的植被、湿土壤和水体在可见光波段发射率都很低，在卫星图像上被称为暗像元。在无云的暗像元上空区域，卫星观测反射率与气溶胶光学厚度之间呈正比例关系，利用这种关系反演 AOD 的算法被称为暗像元法。暗像元法根据遥感图像的归一化植被指数(Normalized Difference Vegetation Index，NDVI)值和短波红外通道(2.13 μm 和 3.8 μm)观测值进行暗像元识别，再依据一定的关系假定这些暗像元在可见光红蓝通道的地表反射率，之后基于表观反射率的大气贡献项和大气辐射传输模型建立气溶胶查找表，以此来反演气溶胶光学厚度。该方法基于表观反射率的大气贡献项及利用卫星观测的路径辐射反演气溶胶光学厚度。利用暗像元法反演 AOD 时，对茂密的绿色植被、湿土壤和水体等低地表反射率区域反演效果明显，但对中高纬度的冬季和干旱地区等高反射率区域，不能采用暗像元法反演 AOD。因为当地表反射率增大时，传感器接收到的辐射值与气溶胶厚度的正比例值将减小，甚至随着反射率的增加，辐射值与气溶胶厚度之间不再存在比例关系。

在中高纬度的冬季、裸地和沙漠等区域，晴天无云时地表反射率很高，但蓝光波段对高亮地表具有低反射率的特征。深蓝算法原理就是利用蓝光波段这一特性，构建高亮地表反射率数据库，再通过查找表来构建与最优卫星观测辐射值的对应关系，以此确定气溶胶光学厚度。由于该算法只针对反射率高的地表物体，对海洋等低反射率地区不能达到很好的反演效果，因此开发了融合 DT/DB 算法的融合 AOD 产品。

由于暗像元法不能反演高反射率区域的气溶胶厚度，而深蓝算法无法反演海洋区域的 AOD，鉴于此种弊端，C6 版本将二者融合得到了 DT/DB 融合算法。该算法的核心思想是：对于海洋区域，选用暗目标法反演 AOD；对于陆地上的高亮区域用深蓝算法反演，陆地上的暗地表则用暗目标法反演。陆地上用 NDVI 值来划分暗地表和高亮区域。

3 本文的方法

3.1 随机森林

随机森林(Random Forest，RF)是并行式集成学习法的一个扩展变体。RF 在以决策树为基学习器构建并行式集成学习法的基础上，进一步在决策树的训练过程中引入随机属性选择。它通过自助法重采样技术，从原始训练样本集 N 中有放回地重复随机抽取 k 个样本生成新的训练样本集合，然后根据自助样本集生成 k 个分类树组成随机森林，新数据的分类结果按分类树投票多少形成的分数而定。其实质是对决策树算法的一种改进，将多个决策树合并在一起，每棵树的建立依赖于一个独立抽取的样品，森林中的每棵树具

有相同的分布，分类误差取决于每一棵树的分类能力和它们之间的相关性。特征选择采用随机的方法去分裂每一个节点，然后比较不同情况下产生的误差。能够检测到的内在估计误差、分类能力和相关性决定选择特征的数目。单棵树的分类能力可能很小，但在随机产生大量的决策树后，一个测试样品可以通过每一棵树的分类结果经统计后选择最可能的分类。具体来说，传统决策树在选择划分属性时是在当前结点的属性集合(假定有 d 个属性)中选择一个最优属性。而在 RF 中，对基决策树的每个结点，先从该结点的属性集合中随机选择一个包含 k 个属性的子集，然后再从这个子集中选择一个最优属性用于划分。这里的参数 k 控制了随机性的引入程度，一般取值为 $k=\log_2 d$。

随机森林是以 K 个决策树 $\{h(X,\theta_k), k=1,2,\cdots,K\}$ 为基本分类器，进行集成学习后得到一个组合分类器。当输入待分类样本时，随机森林输出的分类结果由每个决策树的分类结果简单投票决定。这里的 $\theta_k(k=1,2,\cdots,K)$ 是一个随机变量序列，它是由随机森林的两大随机化思想决定的：

(1) Bagging 思想。从原样本集 X 中有放回地随机抽取 K 个与原样本集同样大小的训练样本集 $T_k, k=1,2,\cdots,K$ (每次约 37% 的样本未被抽中)，每个训练样本集构造一个对应的决策树。

(2) 特征子空间思想。在对决策树每个节点进行分裂时，从全部属性中等概率随机抽取一个属性子集(通常取 $\log_2 M+1$ 个属性，M 为特征总数)，再从该子集中选择一个最优属性来分裂节点。

由于构建每个决策树时，随机抽取训练样本集和属性子集的过程都是独立的，且总体都是一样的，因此是一个独立同分布的随机变量序列。训练随机森林的过程就是训练各个决策树的过程，由于各个决策树的训练是相互独立的，因此随机森林的训练可以通过并行处理来实现，这将大大提高生成模型的效率。将以同样的方式训练得到 K 个决策树组合起来，就可以得到一个随机森林。当输入待分类的样本时，随机森林输出的分类结果由每个决策树的输出结果进行投票决定。

3.2　本文方法

如前所述，本文试图越过反演 AOD 的过程，基于随机森林的机器学习方法，直接建立 $PM_{2.5}$ 与 MODIS 影像本身的关系。具体而言，分为以下几个步骤。

3.2.1 时空匹配预处理

本文使用的数据是 MOD021KM，空间分辨率为 1 km，包含 16 个发射率波段、22 个辐射率波段和 22 个反射率波段。AOD 是 MODIS 提供的 3 km 产品，地面实测 $PM_{2.5}$ 数据使用了 102 个站点的每小时观测数据。

在时间匹配方面，由于 MODIS Terra 卫星过境时间是上午十点半，因此选取过境当天上午十点与十一点的 $PM_{2.5}$ 监测数据，并计算其平均值，作为卫星过境时的地面观测值。

在空间匹配方面，由于 AOD 数据空间分辨率为 3 km，MODIS 数据的空间分辨率为 1 km。因此，通过地面监测站点的经纬度实现监测站点与影像数据的空间匹配。同时，为了直观显示 $PM_{2.5}$ 的真实空间分布，将所有站点的 $PM_{2.5}$ 监测值通过克里金插值法，插值成空间分辨率为 1 km 的数据。另外，在研究区内，受云层及其他因素的影响，AOD 数据经常出现大量数据缺失，因此，采用克里金插值法将缺失的数据进行插值。

3.2.2 样本选择

本文研究区内有 102 个 $PM_{2.5}$ 地面监测站点，大致按照 7∶3 的比例，随机将 70 个站点用于训练，32 个站点用于测试。在生成训练样本时需要考虑云层的影响。将云产品叠加到 MODIS 数据上，如果站点位置有云，则该站点对应位置的像素不作为训练样本。对于第 i 个站点，如

果其对应的 MODIS 影像像素不为云，则该像素为有效像素，其对应的训练样本格式如公式(2)所示。

$$\langle x_1, x_2, \cdots, x_{16}, x_{17}, \cdots, x_{38}, x_{39}, \cdots, x_{60} \rangle_i, y_i \quad (2)$$

其中，x_1～x_{16} 为 16 个波段的发射率值；x_{17}～x_{38} 为 22 个波段的辐射率值；x_{39}～x_{60} 为 22 个波段的反射率值；y_i 为当前站点对应的 $PM_{2.5}$ 实测值。

为了提高模型的预测能力，本文对训练样本做了一定的增强处理，即除了选取当前站点对应的 MODIS 影像像素外，同时也选取了该像素 5×5 邻域内的所有像素，连同这些像素对应于插值后的 $PM_{2.5}$ 实测数据，一起构成新的训练样本。这样做的理由是，根据地理学第一定律，对于插值后的 $PM_{2.5}$ 实测数据，本文认为站点附近 5×5 邻域内的插值数据具有很大的可信度，可以认为是真实值。通过这种方式，在没有云的情况下，一个站点最多可以生成 25 个训练样本。

对于测试样本，则只选取当前站点对应的 MODIS 影像像素的值。因此，在没有云的情况下，最多可以有 32 个测试样本。

3.2.3 模型训练

一般而言，训练样本的分布越均匀，训练的模型越具有代表性。由于 102 个站点是按照大约 7∶3 的比例随机分配，因此，为了达到训练样本分布的均匀性，本文将此随机分配过程重复 150 次，得到 150 组训练样本及其配套的测试样本。将每一组样本输入到随机森林算法中，得到 150 个训练模型。然后将每组样本中的测试样本输入到对应的模型中，得到对应的预测值。选择预测值表现指标最优的模型作为最终的模型，其对应的训练样本和测试样本作为最终选出的样本。

表现指标以均方根误差(Root Mean Square Error，RMSE)最小来判定，即：

$$RMSE_j = \sqrt{\frac{\sum_{n=1}^{N_j} \left(y'_{j,n} - y_{j,n} \right)^2}{N_j}} \quad (3)$$

其中，j 为训练的组数；N_j 为该组的有效测试样本数；$y'_{j,n}$ 和 $y_{j,n}$ 分别为预测值和实际观测值。选择 RMSE 最小的一个模型为最优模型。

3.2.4 模型测试

将上一步中选出的最优模型用于整幅 MODIS 影像的无云区域，对于影像中有云区域 $PM_{2.5}$ 的值以 0 代替，从而得到整幅 MODIS 影像的 $PM_{2.5}$ 反演结果。

4 结果与分析

4.1 研究区

广东省地处中国大陆最南部，东邻福建，北接江西、湖南，西连广西，南临南海，珠江口东西两侧分别与香港、澳门特别行政区接壤，西南部雷州半岛隔琼州海峡与海南省相望。全境位于北纬 20°13′～25°31′ 和东经 109°39′～117°19′，东西跨度约 800 km，南北跨度约 600 km。全省陆地面积为 179 800 km^2。广东省属于东亚季风区，从北向南分别为中亚热带、南亚热带和热带气候，是中国光、热和水资源最丰富的地区之一。以广州为核心的珠三角地区是中国城市化进程最快的区域之一，伴随而来的大气污染问题也比较突出。本文的研究区域如图 1 所示，图中三角形标示点为 102 个环境监测站，逐小时发布 $PM_{2.5}$ 监测数据。

为了验证本文方法的有效性，采用了 MODIS 的 L2 级 1 km 数据(MOD021KM)对 $PM_{2.5}$ 进行反演，数据时间分别为 2015.04.15、2015.04.17、2015.08.08、2015.08.26、2015.10.15、2015.12.20、2016.02.06、2016.02.09、2016.03.20，时间跨越 2 个年份，包含了 4 个季节。数据来源于美国国家航天宇航局(https://ladsweb.modaps.eosdis.nasa.gov/)。该数据包含 16 个波段的发射率数据、22 个波段的反射率数据和 22 个波段的辐射率数据。作为对比，

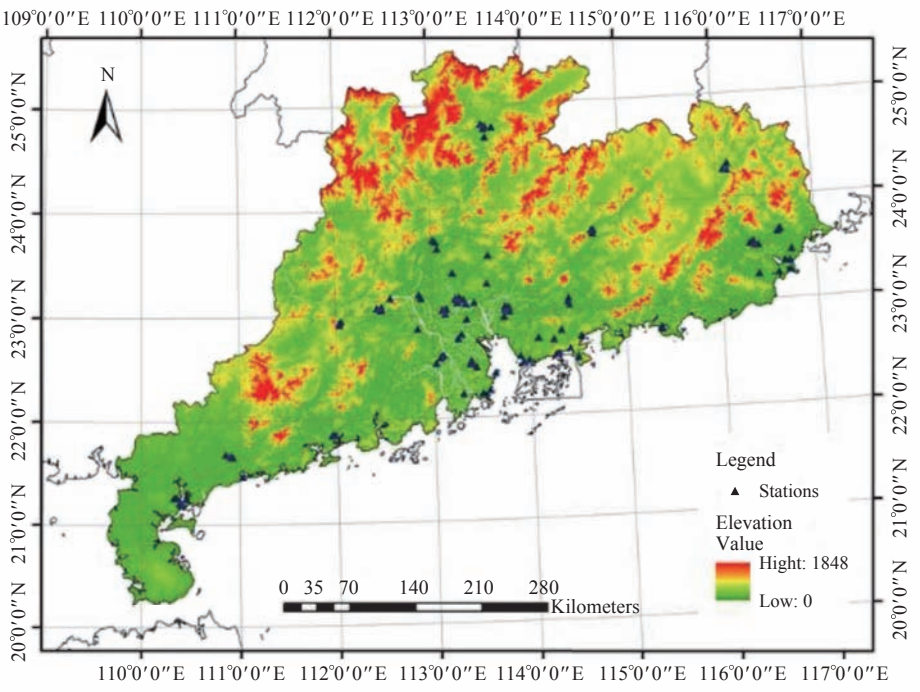


图 1　研究区域

Fig. 1　Study area

同时采用 MODIS 产品中分辨率最高的 3 km 气溶胶产品(AOD)进行试验分析(http://modis-atmos.gsfc.nasa.gov/products.html)，该产品采用最新 C6 版本中的 DT 与 DB 融合算法。本文使用的随机森林算法通过 Weka[18]来实现，网址为 https://www.cs.waikato.ac.nz/ml/weka/index.html。

受广东省气候环境的影响，MODIS 数据经常被大量云层覆盖，导致 AOD 产品上经常出现大面积数据缺失。因此，在使用时，需要使用插值算法弥补这些数据缺失。本文采用克里金插值方法。也有很多研究者自行反演 AOD，但是其精度往往取决于引入的更多辅助数据和特别操作。因此，为了消除其他因素的影响，本文仅仅使用 MODIS 发布的最高分辨率的 AOD 产品，通过最经典的线性回归方法反演 $PM_{2.5}$。线性回归模型基本形式如公式(4)所示。

$$Y=x_i^T\alpha+\beta_i,\quad i=1,2,\cdots,n \tag{4}$$

其中，Y 为因变量；x 为解释变量；i 为第 i 个解释变量；α 为待估系数；β_i 为常数项。由于研究区域经常被云层覆盖，本文选择 2015 年云量相对较少的几天来进行实验验证。实验的 $PM_{2.5}$ 地面监测数据来源于广东省 102 个环境监测站，随机选择其中的 70 个站点用作训练，剩下的 32 个站点做测试。同时用决定系数(R^2)和均方根误差(RMSE)作为评价指标来对反演效果进行对比分析。

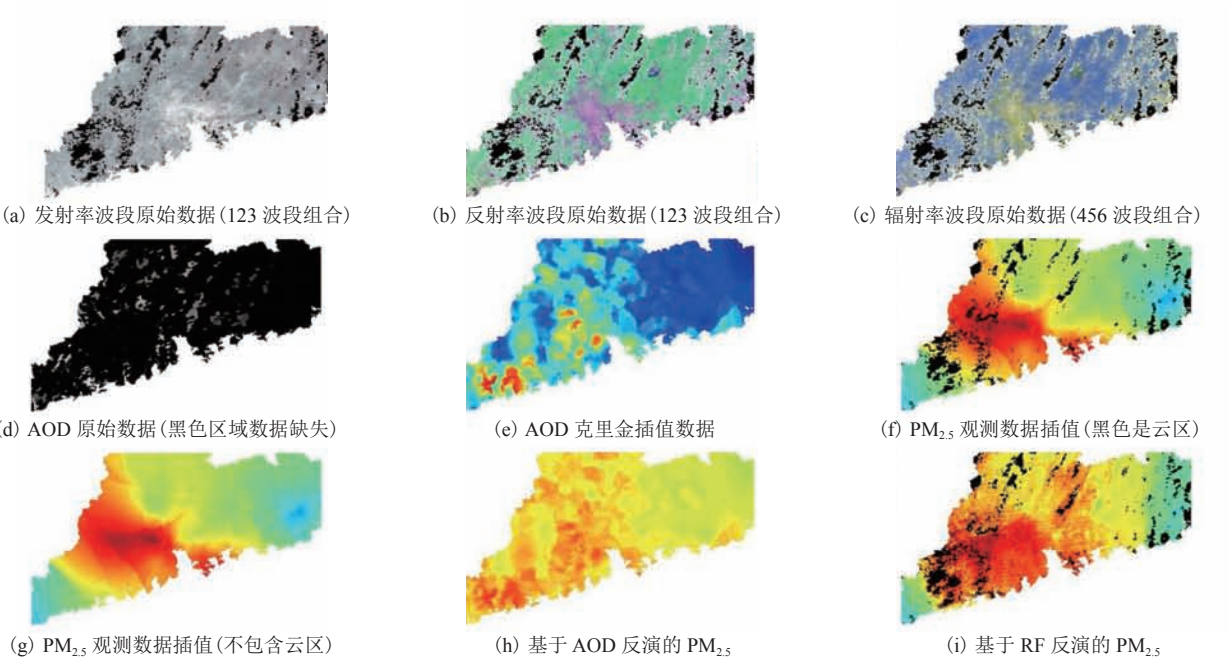
决定系数是指在表征因变数的总平方和中，由自变数引起的平方和所占的比例，称为 R 平方，记为 R^2。由于 $R^2<R$ 可以防止对相关系数所表示的相关做夸张的解释，因此决定系数的大小决定了相关的密切程度。当 R^2 越接近 1 时，表示相关方程式参考价值越高；相反，越接近 0 时，表示参考价值越低。表达公式为：

$$R^2=\frac{\sum_{i=1}^{n}\left(y_i'-\bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2} \tag{5}$$

其中，y 为待拟合数值；$\bar{y}$ 为其均值；y' 为其拟合值。

4.2　2015 年 8 月 8 日反演结果对比

图 2 给出了 2015 年 8 月 8 日的 MOD021KM



(a) 发射率波段原始数据(123 波段组合) (b) 反射率波段原始数据(123 波段组合) (c) 辐射率波段原始数据(456 波段组合)

(d) AOD 原始数据(黑色区域数据缺失) (e) AOD 克里金插值数据 (f) $PM_{2.5}$ 观测数据插值(黑色是云区)

(g) $PM_{2.5}$ 观测数据插值(不包含云区) (h) 基于 AOD 反演的 $PM_{2.5}$ (i) 基于 RF 反演的 $PM_{2.5}$

图 2 2015 年 8 月 8 日实验结果

Fig. 2 Experimental results on 2015.8.8

数据。为显示方便，发射率和辐射率采用 123 波段合成，反射率采用 456 波段合成。当日的数据中有部分云层，采用 MODIS 的云检测产品构建掩膜，如图 2(f)所示的黑色区域即为云区。当日的 AOD 数据存在大量缺失，如图 2(d)中的黑色区域。将此数据进行克里金插值，并用假彩色显示，具体如图 2(e)所示，表现出明显的区块效应。$PM_{2.5}$ 的地面观测值是点状数据，本文利用克里金插值将点状数据插值为面状数据，如图 2(f)、(g)所示，其中图 2(f)加了云掩膜。图 2(h)是基于克里金插值后的 AOD 数据经过线性回归反演得到的 $PM_{2.5}$。图 2(i)是本文方法得到的 $PM_{2.5}$ 结果，其中颜色越红，表示 $PM_{2.5}$ 浓度越大；颜色越蓝，表示 $PM_{2.5}$ 浓度越低。

从图 2(f)、(g)所示的地面观测值可以看出，中间区域的 $PM_{2.5}$ 浓度很高，东北和西南两个区域的浓度较低。AOD 反演的结果与地面观测结果差异很大，这是由 AOD 数据缺失导致的。而本文方法在整体趋势上与地面观测结果非常一致，表现出中间高、东北和西南低的趋势。

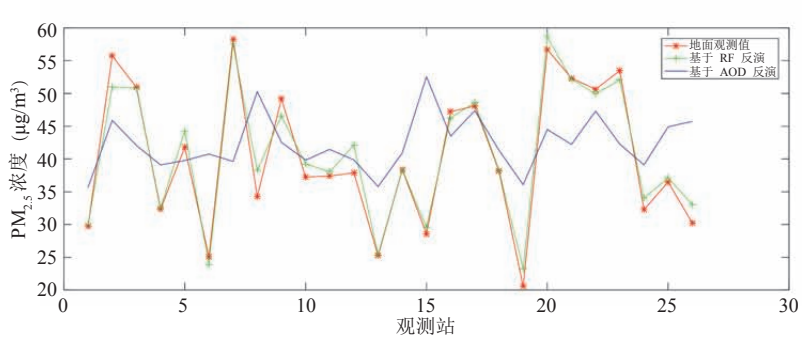
受云层影响，32 个验证站中只有 26 个站有数据，因此给出了 26 个地面观测站的统计结果。

由图 3(a)可以看出，本文方法在各个观测站上的预测值都能与实际观测值有较好的匹配，而 AOD 方法匹配度较差。由散点图和线性拟合结果(图 3(b)、(c))可以看出，本文方法的 R^2 达到 0.97，RMSE 小于 2，表现出了极强的相关性；而 AOD 方法表现非常差，这也说明 AOD 的数据缺失对 $PM_{2.5}$ 的反演有极大的负面影响。

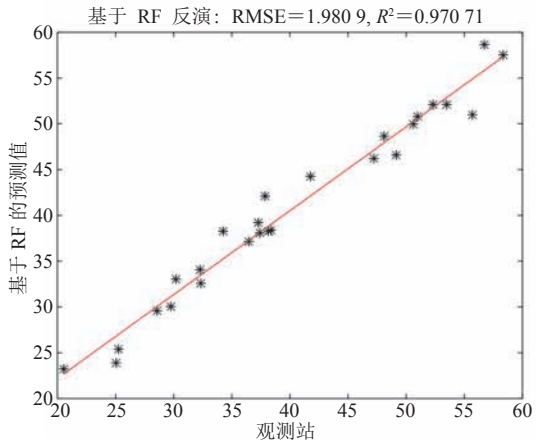
4.3 2015 年 4 月 17 日反演结果对比

另外选择云量更少的一天的数据进行实验，结果如图 4 所示。图中少量的黑色区域即为 MODIS 云检测产品提供的云掩膜。AOD 产品的数据缺失程度较上一个实验有了明显改善，采用克里金插值后没有表现出明显的区块效应。AOD 和本文方法反演的结果如图 4(h)、(i)所示。

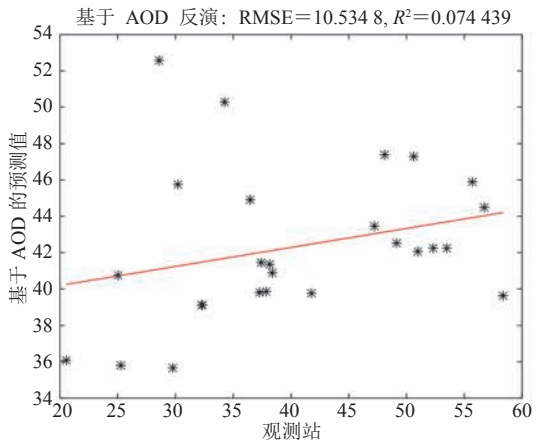
从图 4(f)、(g)所示的地面观测值插值结果可以看出，中部 $PM_{2.5}$ 浓度较大，东北角区域次之，西南角区域最小。AOD 反演结果在中部区域跟地面观测值较一致，但是西南角区域明显偏



(a) 反演值与观测值的比较(26 个有效观测站，另外 6 个观测站被云层覆盖)



(b) 本文方法的散点图及统计指标



(c) 基于 AOD 方法的散点图及统计指标

图 3　2015 年 8 月 8 日实验结果

Fig. 3　Experimental results on 2015.8.8

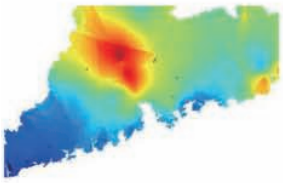
(a) 发射率波段原始数据(123 波段组合)

(b) 反射率波段原始数据(123 波段组合)

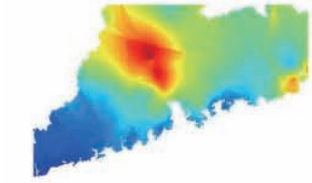
(c) 辐射率波段原始数据(456 波段组合)

(d) AOD 原始数据(黑色区域数据缺失)

(e) AOD 克里金插值数据



(f) $PM_{2.5}$ 观测数据插值(黑色是云区)



(g) $PM_{2.5}$ 观测数据插值(不包含云区)

(h) 基于 AOD 反演的 $PM_{2.5}$

(i) 基于 RF 反演的 $PM_{2.5}$

图 4　2015 年 4 月 17 日实验结果

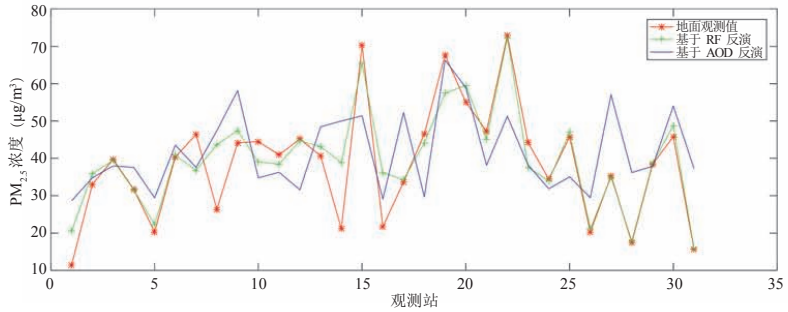
Fig. 4　Experimental results on 2015.4.17

大。而本文方法依然与地面观测结果表现出明显的一致性。由于云量较少，32 个验证站点中有 31 个属于有效站点。AOD 和本文方法反演的结果与地面观测结果如图 5(a)所示，可以看出，大部分站点上本文方法与地面观测结果吻合度非常好，而 AOD 则有明显的偏差。

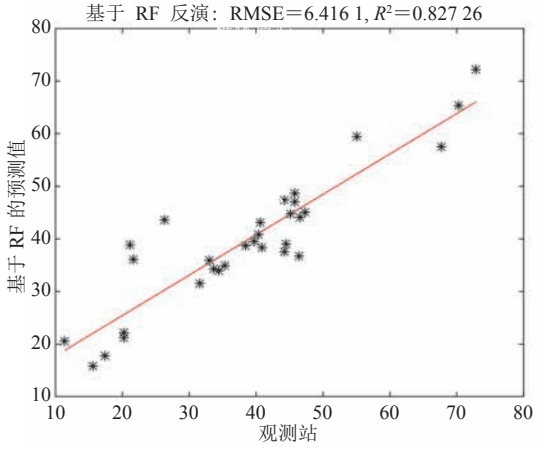
由散点图和线性拟合统计结果(图 5(b)、(c))可以看出，本文方法的 R^2 远比 AOD 反演的高，RMSE 更低，表现出更强的线性关系；而 AOD 反演结果则表现得比较离散。由于云量比上一个实验少，所以 AOD 反演的 R^2 有了明显提高。

用同样方法对其他几个日期的数据进行了实验，计算 R^2 和 RMSE 的平均值，结果如表 1 所示。

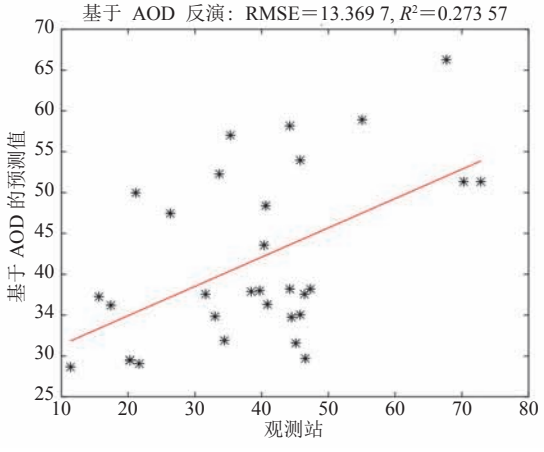
由表 1 可以看出，本文方法(基于 RF 反演)的 $PM_{2.5}$ 均值比基于 AOD 法要高，而 RMSE 更低。从 RMSE 来看，本文方法也具有比较明显的优势。R^2 和 RMSE 波动的主要因素是云量的影响，以及 AOD 数据本身的缺失问题。因为在本研究区，AOD 数据缺失现象有时比较严重，通过克里金插值补齐的数值并不能完全反映真实的



(a) 反演值与观测值的比较(31 个有效观测站，另外 1 个观测站被云层覆盖)



(b) 本文方法的散点图及统计指标



(c) 基于 AOD 方法的散点图及统计指标

图 5 2015 年 4 月 17 日实验结果

Fig. 5 Experimental results on 2015.4.17

表 1 评价指标的平均值

Table 1 The average values of assessment indices

指标	R^2		RMSE	
	基于 RF 反演	基于 AOD 反演	基于 RF 反演	基于 AOD 反演
平均值	0.879 2	0.170 1	3.619 3	7.493 9

AOD 空间分布。

5 结　论

MODIS AOD 算法的不断改进，目的是得到结果更为精确的 AOD 产品。但在反演 AOD 过程中不可避免地会有误差的存在，因此现在常用的 AOD 反演 $PM_{2.5}$ 方法无法避免误差。本文结合随机森林的机器学习算法，从遥感影像本身数据出发，直接建立遥感影像与实测 $PM_{2.5}$ 的关系，从而避免了误差传递。选取了 MODIS 数据分辨率 3 km 的 AOD 产品和广东省 102 个环境监测站点的 $PM_{2.5}$ 数据进行试验。试验结果表明，本方法能够取得更好的 $PM_{2.5}$ 反演效果，同时将 $PM_{2.5}$ 反演的空间分辨率提高到 1 km。下一步研究将扩大研究区域，采用更多的数据，进一步提高算法的可用性。另外，采用其他更好的机器学习方法来确定反演模型也是今后研究的重点。

参 考 文 献

[1] Lin CQ, Li Y, Yuan ZB, et al. Using satellite remote sensing data to estimate the high-resolution distribution of ground-level $PM_{2.5}$ [J]. Remote Sensing of Environment, 2015, 156: 117-128.

[2] Gong W, Huang S, Zhang TH, et al. Impact and suggestion of column-to-surface vertical correction scheme on the relationship between satellite AOD and ground-level $PM_{2.5}$ in China [J]. Remote Sensing, 2017, 9(10): 1038.

[3] Chen ZY, Chen DL, Zhuang Y, et al. Examining the influence of crop residue burning on local $PM_{2.5}$ concentrations in Heilongjiang province using ground observation and remote sensing data [J]. Remote Sensing, 2017, 9(10): 971.

[4] 陈辉, 厉青, 张玉环, 等. 基于地理加权模型的我国冬季 $PM_{2.5}$ 遥感估算方法研究 [J]. 环境科学学报, 2016, 36(6): 2142-2151.

[5] 王子峰. 卫星遥感估算近地面颗粒物浓度的算法研究 [D]. 北京: 中国科学院遥感应用研究所, 2010.

[6] 王中挺, 辛金元, 贾松林, 等. 利用暗目标法从高分一号卫星 16 m 相机数据反演气溶胶光学厚度 [J]. 遥感学报, 2015, 19(3): 530-538.

[7] 王静, 杨复沫, 王鼎益, 等. 北京市 MODIS 气溶胶光学厚度和 $PM_{2.5}$ 质量浓度的特征及其相关性 [J]. 中国科学院研究生院学报, 2010, 27(1): 10-16.

[8] Song WZ, Jia HF, Huang JF, et al. A satellite-based geographically weighted regression model for regional $PM_{2.5}$, estimation over the pearl river delta region in China [J]. Remote Sensing of Environment, 2014, 154: 1-7.

[9] 张天棋. 基于 MODIS 影像的中国地区气溶胶产品验证与 $PM_{2.5}$ 反演 [D]. 成都: 西南交通大学, 2017.

[10] Chu DA, Tsai TC, Chen JP, et al. Interpreting aerosol lider profiles to better estimate surface $PM_{2.5}$, for columnar AOD measurements [J]. Atmospheric Environment, 2013, 79(11): 172-187.

[11] Griggs M. Measurements of atmospheric aerosol optical thickness over water using ERTS-1 data [J]. Journal of the Air Pollution Control Association, 1975, 25(6): 622-626.

[12] Levy RC, Kozak GM, Wadsworth CB, et al. Towards a long-term global aerosol optical depth record: applying a consistent adrosol retrieval algorithm to MODIS and VIIRS-observed reflectance [J]. Atmospheric Measurement Techniques, 2015, 10(8): 4083-4110.

[13] Holben B, Vermote E, Kaufman YJ, et al. Aerosol retrieval over land from AVHRR data application for atmospheric correction [J]. IEEE Transactions on Geoscience and Remote Sensing, 1992, 30(2): 212-222.

[14] Isakov VY, Feind RE, Vasilyev OB, et al. Retrieval of aerosol spectral optical thic-kness from AVIRIS data [J]. International Journal of Remote Sensing, 1996, 17(11): 2165-2184.

[15] Shi YS, Matsunaga T. Long-term trends and sptial patterns of satellite-retrieved $PM_{2.5}$ concentrations in South and Southeast Asia from 1999 to 2014 [J]. Science of the Total Environment, 2018, 615: 177-186.

[16] Jung CR, Hwang BF, Chen WT. Incorporating long-term satellite-based aerosol optical depth, localized land use data, and meteorological variables to estimate ground-level $PM_{2.5}$ concentrations in Taiwan from 2005 to 2015 [J]. Environmental Pollution, 2017, doi: 10.1016/j.envpol.2017.11.016.

[17] Mao X, Shen T, Feng X. Prediction of hourly ground-level $PM_{2.5}$ concentrations 3 days in advance using neural networks with satellite data in eastern China [J]. Atmospheric Pollution Research, 2017, 8(6): 1005-1015.

[18] The WEKA Workbench. Data mining: practical machine learning tools and techniques [EB/OL]. [2018-3-12]. https://www.cs.waikato.ac.nz/ml/weka/citing.html.