北京大学学报(自然科学版) 第 56 卷 第 1 期 2020 年 1 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 56, No. 1 (Jan. 2020)
doi: 10.13209/j.0479-8023.2019.110

基于空地协同采样的植被覆盖度随机森林估算方法

程俊毅 张显峰† 孙敏 罗鹏 杨婉婷

北京大学遥感与地理信息系统研究所, 北京 100871; † 通信作者, E-mail: xfzhang@pku.edu.cn

摘要 基于无人机高光谱影像, 建立地形复杂地区植被覆盖度的非参数随机森林回归估算模型。为获得构建随机森林模型所需的足够数量的训练样本, 利用低空无人机搭载的光学相机, 在从地面难以到达的山地、水域和植被茂密区, 通过垂直拍摄获得厘米分辨率的航拍影像, 作为对地面样方采样的补充。首先计算地面数码相机照片和无人机可见光影像的红绿蓝植被指数(red-green-blue vegetation index, RGBVI), 然后使用大津分割法提取样方的植被覆盖信息, 得到构建模型所需的训练样本。在此基础上, 基于2018年8月16—18日在内蒙古自治区察右中旗油娄沟矿区获取的GaiaSky-mini2无人机高光谱影像数据, 利用递归特征消除算法优选参与随机森林回归的特征变量集, 利用空地协同获取的训练样本构建植被覆盖度的随机森林回归估算模型。该模型在测试集上的确定系数R^2为0.923, 均方根误差为0.087, 优于常用的像元二分模型, 可用于矿区植被动态信息的精细化监测。

关键词 植被覆盖度; 随机森林; 空地协同采样; 无人机高光谱; 矿区

Random Forest Model for the Estimation of Fractional Vegetation Coverage Based on a UAV-Ground Co-Sampling Strategy

CHENG Junyi, ZHANG Xianfeng†, SUN Min, LUO Peng, YANG Wanting

Institute of Remote Sensing and GIS, Peking University, Beijing 100871; † Corresponding author, E-mail: xfzhang@pku.edu.cn

Abstract A nonparametric regression — random forest model for the estimation of fractional vegetation coverage (FVC) in a complex topographic area is presented based on low-altitude unmanned aerial vehicle (UAV) hyperspectral imagery. In order to collect a large number of sufficient training samples required for random forest algorithm, the UAV equipped with an optical camera was used to vertically capture the images of land covers in several inaccessible areas such as high mountains, water body and densely forested areas, to increase the density of the ground sampling. The RGBVI (red-green-blue vegetation index) was calculated first and then the Otsu method was adopted to extract the FVC values of the samples from the UAV optical images and ground photos. After that, the hyperspectral images captured by the UAV GaiaSky-mini2 hyperspectral imaging system in the Youlougou Mining area, Chayouzhong County, Inner Mongolia on August 16–18, 2018 were used to extract feature variables, and this feature set was filtered by recursive feature elimination algorithm based on the importance of the variables. On the basis of the optimized feature set and extended training samples using the proposed UAV-ground co-sampling approach, the random forest estimation model was constructed to estimate the FVC in the study area. Results indicated that the model achieved a determinant coefficient (R^2) of 0.923 and a RMSE of 0.087 on the testing sample set and outperformed the commonly used Pixel Dichotomy method. It can be used in the fast and accurate monitoring of vegetation dynamics in mining areas.

Key words fractional vegetation coverage; random forest; UAV-ground co-sampling; UAV hyperspectral remote sensing; mining area

内蒙古自治区科技厅重大专项(数字化矿区资源管理与矿区生态环境监测技术与应用, 2015—2018)和国家重点研发计划(2017YFC1500902)资助

收稿日期: 2019–01–24; 修回日期: 2019–05–21

植被覆盖度指植被在地面的垂直投影面积占统计面积的百分比[1], 是衡量地表植被状况的一个重要指标, 在生态平衡、水土流失和气候变化研究中有重要作用[2]。植被覆盖是控制土壤侵蚀的关键因素, 侵蚀量与植被覆盖度具有显著的负相关关系[3], 因此, 监测矿区植被覆盖度变化对了解采矿活动对矿区植被的影响十分重要。

传统的地表实测法难以在较大空间尺度上进行动态测量[4], 遥感技术具有大范围、多时相、多尺度数据获取与连续观测等特点, 已成为大范围地表植被覆盖度反演的重要手段。植被覆盖度的遥感估算方法目前主要有混合像元分解法、统计模型法和数据挖掘法[5]。

混合像元分解法一般应用于缺乏地面实测数据的情况[6], 主要有线性模型、概率模型、几何光学模型、随机几何模型和模糊分析模型等, 其中线性分解模型的应用最广泛[5,7], 但在干旱半干旱地区, 植被分布零散稀疏, 较难获得纯净像元的典型光谱, 导致估算误差较大。

因此, 在有地面样本的情况下, 多采用建立统计模型来估算植被覆盖度。统计模型法通常选择遥感数据的某一波段或几个波段的组合及植被指数, 与地面实测样方植被覆盖度进行统计分析, 建立两者之间的经验统计关系[8–9], 可分为线性回归模型和非线性回归模型, 其中前者应用较广泛[10–11]。然而, 许多研究未充分考虑参数回归较严格的假设条件和多元回归对变量之间非共线性的要求, 降低了反演模型的可靠性。解决此问题的途径之一是寻找预测效果达到甚至超过参数模型的非参数方法[12]。

数据挖掘法可以从大量数据中提取隐含在其中的潜在有用的信息和知识。目前, 人工神经网络、支持向量机和随机森林等非参数回归算法已在植被覆盖度反演中得到应用。人工神经网络具有很强的自组织、自学习、自适应能力和较强的非线性容错性, 成为植被覆盖度反演的重要方法之一[5,13–14]。支持向量机在解决小样本、非线性及高维模式识别等问题中表现出很好的泛化能力, 近年来也应用于植被覆盖度的估算[15]。

与参数回归等方法相比, 随机森林模型无需对变量的正态性和独立性等假设条件进行检验, 广泛应用于环境生态学等领域, 但在植被覆盖度估算中应用相对较少[12]。

支持向量机、神经网络和随机森林等机器学习算法都需要大量训练样本的支撑, 然而, 在诸如山区、水域和茂密森林等人员难以到达的区域, 通过地面采样获取训练样本十分困难, 甚至存在大量无法布设地面样方的区域, 导致模型建立困难和模型预测精度较低等问题。

鉴于上述背景, 本文针对植被覆盖度反演中难以满足线性回归的假设条件及变量共线性等统计假设的问题, 使用预测效果较好的随机森林回归算法, 建立矿山尺度的植被覆盖估算模型。同时, 针对模型需要大量训练样本的问题, 引入低空无人机可见光影像数据, 为地面样方进行加密, 探讨协同空(无人机遥感)地(地面样方)获取大量训练样本的可行性及随机森林模型在植被覆盖度反演中的适用性。

1 数据与方法

1.1 研究区位置

研究区位于内蒙古自治区乌兰察布市察哈尔右翼中旗的油篓沟金矿区(图 1), 该矿区地理坐标为 41.2039°—41.2141°N, 112.2986°—112.3079°E。研究区包含北侧的尾矿坝、南侧的露天矿及位于沟底的选矿区 3 个部分。尾矿坝南侧属阳坡, 植物的长势比阴坡的露天矿区域差, 植被覆盖相对较低。研究区有灌木、草本、农作物和乔木等多种类型植被, 且高中低植被覆盖度均有, 具有较好的代表性。

1.2 数据获取与预处理

1.2.1 无人机高光谱数据

2018 年 8 月 16—18 日, 本研究组利用大疆 M600 搭载 GaiaSky-mini2 高光谱成像仪获取研究区的高光谱影像。该成像仪获取波长范围为 391.3~

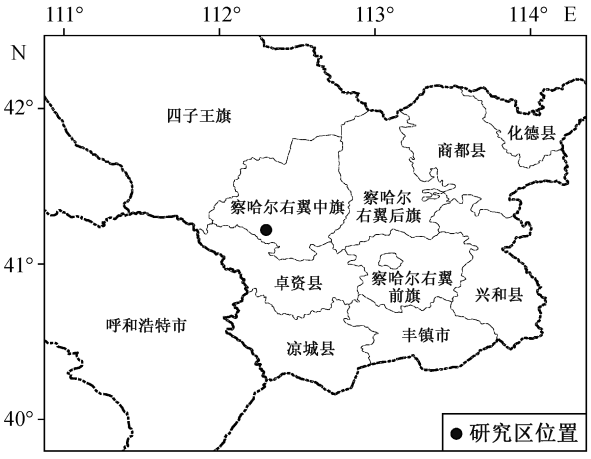


图 1 研究区地理位置

Fig. 1 Geographic location of the study area

997.6 nm, 光谱分辨率为3.5±0.5 nm, 共176个波段, 拍摄方式为悬停内置扫描, 即无人机每到一幅影像的中心点, 原地悬停约3.7 s, 进行推扫式成像。无人机变高飞行, 距地面高度为280 m, 影像空间分辨率约为13.7 cm。

1) 镜头校正。GaiaSky-mini2采用狭缝扫描方式, 镜头边缘与中心成像的亮度存在差异, 需要利用标准光源对进行校准。我们根据厂家提供的校准参数及预处理软件进行镜头校准, 消除畸变和亮度不匀。

2) 反射率校准。高光谱相机采集的被测物的影像亮度值会受到光源光谱、光源强度、镜头透过率、光谱仪的衍射效率、光谱响应效率以及被测物的反射率等因素影响, 无人机飞到一定高度后, 成像会受大气和水汽等因素影响。因此, 对高光谱数据的反射率校正包括亮暗法辐射校正和大气校正两个部分。亮暗法辐射校正指在采集数据的同时采集标准白板和盖上镜头后的暗背景DN值, 校正算法如下:

$$R_{\text{Sample}}=\frac{\text{DN}_{\text{Sample}}-\text{DN}_{\text{Dark}}}{\text{DN}_{\text{White}}-\text{DN}_{\text{Dark}}}, \tag{1}$$

式中, $\text{DN}_{\text{Sample}}$, DN_{Dark}和DN_{White}分别代表被测物、暗背景和白板的像元亮度值, R_{Sample}代表被测地物亮暗法辐射校正后的辐射亮度。采集暗背景数据时, 需关闭光源, 或用设备镜头盖将设备外光源封闭, 采集到的影像DN值代表相机的系统误差DN值。为完成高光谱影像的大气校正, 无人机飞行时, 在被拍摄区域放置3张经国家计量院标定的反射率分别为20%, 40%和60%的漫反射参考灰布, 同步获得参考灰布的高光谱影像数据, 通过参考灰布的已知反射率完成对被测物反射率的大气纠正, 即将影像辐射校正后被测物与参考灰布的辐射亮度比值与参考灰布的标定反射率的乘积作为被测物的地表反射率。

3) 高光谱影像拼接。首先计算全部176个波段的信噪比, 在红、绿、蓝3个波段区间分别挑选一个信噪比最高的波段。利用PIX4D商用软件包计算3个波段合成后每幅无人机高光谱影像的外方位元素和数字高程模型, 并将计算得到的外方位元素赋值给同一景影像所有的176个波段。对每个波段影像进行拼接融合, 得到整个实验区域的高光谱拼接影像(图2)。

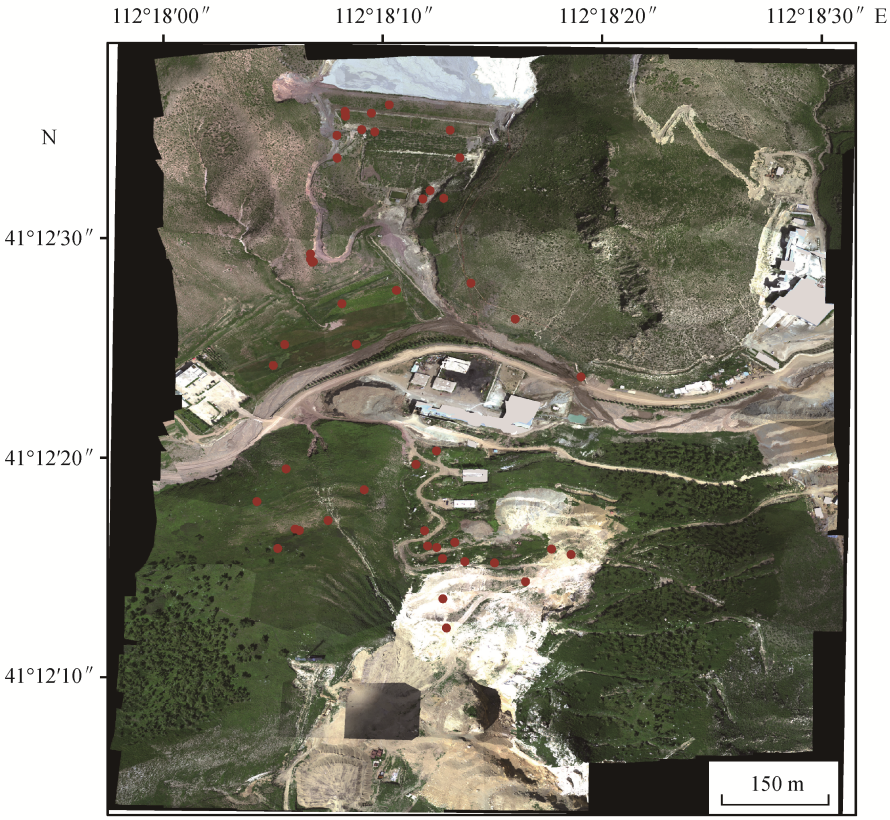
1.2.2 地面样方数据

地面采样工作与无人机飞行同步进行, 结合前期考察情况, 在Google地图上选择对不同植被类型以及不同植被覆盖度有代表性的区域作为观测样方。在研究区共选择36个有代表性的样地, 其中包括18个草本样方、16个灌木样方和2个农作物样方。样方分布如图2所示。灌木样方的面积为2 m×2 m, 草地和作物的样方面积为1 m×1 m。在布设样方时, 利用指南针和皮尺, 在样方的四角插入长约35 cm 的地钉, 保证样方的两边与南北方向平行, 使用红色尼龙绳将样方圈起。垂直向下拍摄照片, 拍摄时镜头距地面的高度约为1.5 m。拍摄后, 对照片进行畸变校正和裁剪, 得到样方的标准照片。

1.3 空地协同采样获取训练样本

矿区位于较陡峭的山地, 山顶、悬崖、陡坡及尾矿坝等一些区域难以到达, 无法布设样方。由于机器学习模型需要大量的训练样本, 因此我们利用大疆精灵3无人机搭载光学相机在无法到达的重点区域进行拍摄, 获取高空间分辨率(3 cm左右)的可见光影像, 对地面样方进行加密和扩展。选择尾矿坝北侧、尾矿坝山顶、露天矿山顶和农作物田块中间4片区域, 在其对应的高光谱影像上分别裁剪大小为(m×15)×(n×15)个像素(对应地面m×n个2 m×2 m的样方)的区域, 并裁剪相同区域的可见光影像。将裁剪完的高光谱影像和可见光影像划分为m×n个网格(每个网格对应地面2 m×2 m的样方), 提取每个网格的植被覆盖度作为训练样本。

在研究区选择的样方主要由两部分组成: 地面相机拍摄的样方以及从无人机可见光影像上选择的样方, 均为三波段RGB影像。本研究利用阈值分割法[16]从样方影像提取植被覆盖度值, 包括影像变换和阈值确定两个步骤。影像变换主要有两种方法: 一是基于可见光谱的植被指数; 二是变换至另一色彩空间, 增强植被与非植被的差异。本研究将多篇文献中精度较高的方法进行实验和对照, 从中选择精度最高的方法来提取样方影像的植被覆盖度。共选择7种植被指数: 红绿比指数(red-green ratio index, RGRI)[17]、蓝绿比指数(blue-green ratio index, BGRI)[18]、红绿蓝植被指数(red-green-blue vegetation index, RGBVI)[18]、过绿指数(excess green, EXG)[19]、可见光波段差值植被指数(visible-



红光波长为 639.1 nm, 绿光波长为 551.2 nm, 蓝光波长为 458.9 nm; 红色圆点为地面样方位置

图 2 研究区无人机影像及地面样方分布

Fig. 2 UAV image of the study area and locations of the field samples

band difference vegetation index, VDVI)[20]、归一化绿红差值指数(normalized green-red difference index, NGRDI)[21]和归一化绿蓝差值指数(normalized green-blue difference index, NGBDI)[22]。变换色彩空间选择 CIELab 色彩空间[23], 该色彩空间能够更好地反映植被与非植被的差别。CIELab 模式也由 3 个通道组成, L 通道是明度, a 通道的颜色是从红色到深绿色, b 通道则是从蓝色到黄色。

利用上述 8 种方法对样方可见光影像进行变换后, 得到 8 种变换后的影像, 再采用大津法对变换后的影像进行阈值分割[24], 按图像的灰度特性, 将图像分成背景和前景两部分, 然后选定阈值, 使植被与非植被的类间方差达到最大值, 用来提取样方植被覆盖度值。有了足够数量的样方数据, 即可构建随机森林植被覆盖度反演模型。最终, 基于空地协同采样方法, 共获得 379 个样方的植被覆盖度值。随机选择 250 个样方作为训练集参与模型训练, 剩下的 129 个样方作为测试集, 用来验证模型估算精度。

1.4 随机森林反演模型的构建

1.4.1 特征变量的提取

虽然高光谱遥感影像可以提供丰富的地物光谱信息, 在植物的识别和长势监测等方面比多光谱遥感影像更具优势, 但也存在信息量大、信息冗余且波段之间相关性高的缺点, 因此需要对高光谱遥感数据进行降维处理、特征提取和选择[25]。本研究基于无人机高光谱影像 176 个波段的反射率, 分别进行相关性分析、植被指数计算以及主成分分析, 提取随机森林模型的回归特征变量。首先对 176 个波段与植被覆盖度进行相关性分析, 选择与植被覆盖度相关性最高的波段参与特征集的构建。同时, 为了避免特征之间的高相关性, 在红、绿、蓝和近红外 4 个波段区间分别选择一个相关性最高的波段作为特征变量。

我们选择 7 种在植被覆盖度提取中相关性较高的植被指数[26–29]: 归一化差值植被指数(normalized difference vegetation index, NDVI)、比值植被指数(simple ratio vegetation index, SR)、增强型植被指

数(enhanced vegetation index, EVI)、土壤调节植被指数(soil adjusted vegetation index, SAVI)、大气阻抗植被指数(atmospherically resistant vegetation index, ARVI)、改进的土壤调节植被指数(modified soil adjusted vegetation index, MSAVI)和红边归一化差值植被指数(red edge normalized difference vegetation index, RENDVI)。这 7 种植被指数的计算公式如下:

$$\mathrm{NDVI} = (\mathrm{NIR} - R) / (\mathrm{NIR} + R), \tag{2}$$

$$\mathrm{SR} = \mathrm{NIR} / R, \tag{3}$$

$$\mathrm{EVI} = (\mathrm{NIR} - R) / (\mathrm{NIR} + 6R - 7.5B), \tag{4}$$

$$\mathrm{SAVI} = (\mathrm{NIR} - R)(1 + L) / (\mathrm{NIR} + R + L), \tag{5}$$

$$\mathrm{ARVI} = (\mathrm{NIR} - 2R + B) / (\mathrm{NIR} + 2R - B), \tag{6}$$

$$\mathrm{MSAVI} = \left[(2\mathrm{NIR} + 1) - \sqrt{(2\mathrm{NIR} + 1)^2 - 8(\mathrm{NIR} - R)}\right] / 2, \tag{7}$$

$$\mathrm{RENDVI} = \frac{\rho_{750} - \rho_{705}}{\rho_{750} + \rho_{705}}, \tag{8}$$

式中, R, NIR 和 B 分别为红光、近红外光及蓝光波段的反射率; L 为土壤调整因子, 本文选取 0.5; ρ_{750} 和 ρ_{705} 分别代表中心波长为 750 nm 和 705 nm 的波段反射率。红光、近红外光及蓝光波段反射率的选择方法如下: 在这 3 种波段对应的波长范围(红光: 630~690 nm; 近红外光: 760~900 nm; 蓝光: 450~520 nm)内, 逐个取值进行组合, 计算每种组合与植被覆盖度的相关性, 从每种植被指数中选择最优的波段组合作为高光谱的植被指数值。有研究表明, 与通过求平均值得到的宽波段相比, 经这种方式选出的窄波段与植被覆盖度的相关性更强, 能更好地刻画植被相对较窄的吸收谷[30]。

通过主成分分析, 可以实现高光谱数据的降维, 大大减小数据量。本研究在 Matlab 环境下对样本的高光谱影像数据进行主成分分析, 第一主成分的贡献率为 65.99%, 前两个主成分的累积贡献率达到 98.79%。利用主成分分析, 将原 176 维的数据降到二维, 在大大减少数据量的同时, 能够保留原始数据的大部分信息。

1.4.2 特征集的选择

前面得到的特征集可能存在信息冗余或泛化性较差的问题, 本文采用两步处理方法: 第一步, 分析测试样本的特征与植被覆盖度的相关性, 进行初选; 第二步, 利用随机森林方法计算每个特征的重要性, 采用递归消除法逐个去除重要性最低的特征。在随机森林模型的构建中, 可通过给某个特征随机加入噪声, 利用前后袋外数据(OOB)的误差值计算某个特征的重要性。本研究利用递归特征消除算法对特征集进行选择。通过特征选择, 一是找出与植被覆盖度高度相关的特征变量, 二是选出能够充分预测植被覆盖度且数目最少的特征变量集。递归特征消除算法的步骤如下: 1) 利用全部特征建立随机森林模型, 计算特征集中每个变量的重要性并排序; 2) 删除重要性最低的特征; 3) 用新的特征变量集建立新的随机森林模型, 再次计算每个变量的重要性并排序。若与前一个特征集精度差异较大则不能删掉该特征, 最终的特征集为删除之前的特征集; 若精度差异较小, 则重复步骤 2。

1.4.3 随机森林回归模型的建立

随机森林是一种监督学习算法, 结合了 Breiman 的 Boost aggregating 集成思想与贝尔实验室提出的特征随机选取思想, 已被 Breiman[31]证明其稳定性和泛化性比单棵的决策树更好。简单地说, 随机森林采用随机有放回的方法选择训练数据, 构造多个决策树分类器, 通过随机选择特征来构建最优分割, 最后将学习到的模型进行组合来增加整体的效果, 是目前预测效果最好的非参数回归模型之一。与参数回归等方法相比, 随机森林模型无需对变量的正态性和独立性等假设条件进行检验, 也不需要考虑变量的贡献, 运算高效, 结果准确[32–33]。随机森林模型的具体算法参见文献[31,33]。

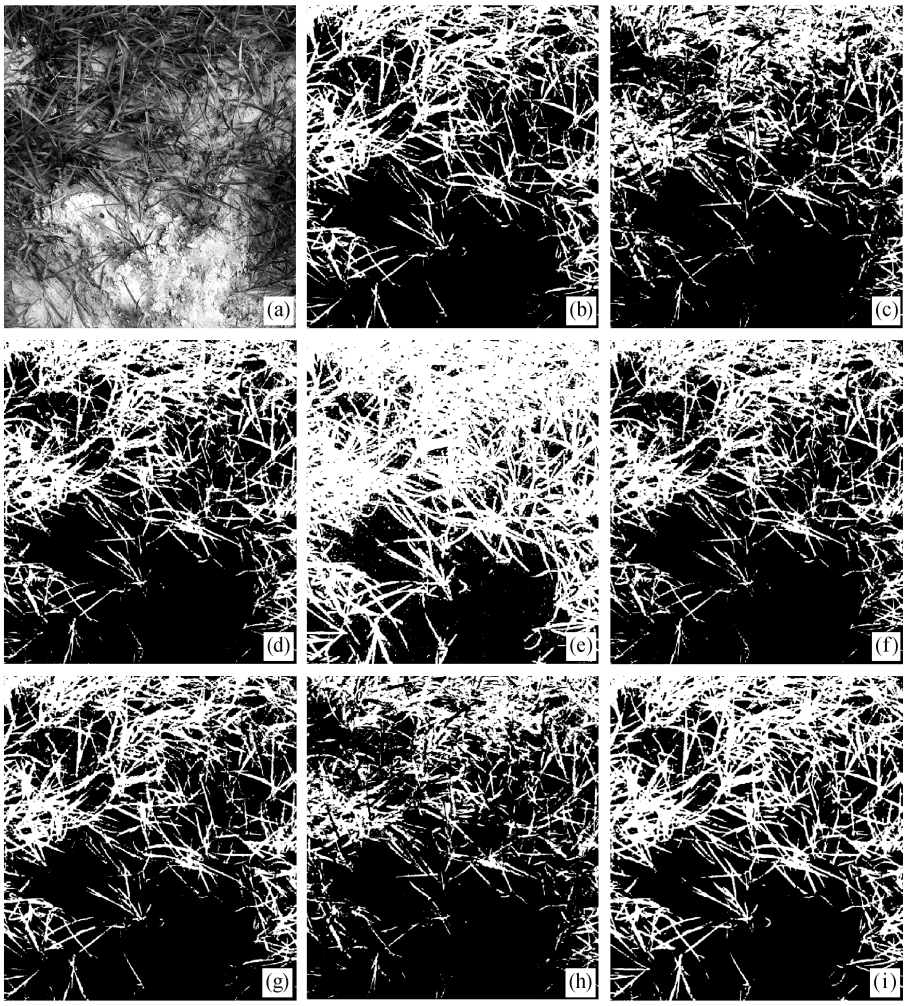
2 结果与分析

2.1 基于随机森林的植被覆盖度估算模型

2.1.1 样方植被覆盖度的提取与比较

基于 1.3 节描述的几种利用可见光 RGB 影像提取植被覆盖度的方法, 分别对地面拍摄和无人机拍摄的 RGB 影像进行植被覆盖度提取。编号为 2606 的地面样方照片以及无人机获取的农作物种植区中某个样方的植被覆盖度提取结果分别如图 3 和图 4 所示。

为验证上述几种方法对植被覆盖度的估算精度, 结合地面实地考察情况, 以样方照片和无人机影像选取的植被与非植被点作为真值验证数据, 计算混淆矩阵(表 1)。



(a) 编号为 2606 的地面样方照片; (b) RGRI 提取结果; (c) BGRI 提取结果; (d) RGBVI 提取结果; (e) EXG 提取结果; (f) VDVI 提取结果; (g) NGRDI 提取结果; (h) NGBDI 提取结果; (i) 转换到 CIELab 颜色空间的提取结果。下同

图 3 地面样方影像及 8 种方法的植被覆盖度提取结果

Fig. 3 Image of the field sample and corresponding FVC estimations using eight methods

对空间分辨率极高的地面数码照片来说, 除 BGRI 和 NGBDI 外, 利用其余 5 种植被指数和转换至 CIElab 颜色空间后, 均能较好地区分植被与非植被, 并且转换至 CIELab 颜色空间和利用 RGBVI 指数两种方法得到的植被边缘更加清晰, 能够更好地反映植被的形状。对于空间分辨率为 3 cm 左右的无人机可见光影像, RGBVI 与 CIELab 颜色空间的精度相差不大, RGBVI 颜色空间的精度稍高, 对比样方影像, 可以发现 RGBVI 能够反映植被的纹理和间隙, 更贴近真实情况。最终选择大津分割法来分割 RGBVI 影像, 实现从 379 个样方的地面数码照片和无人机 RGB 影像提取植被覆盖度, 作为随机森林模型的训练和检验样本。

2.1.2 特征变量的提取与选择

使用上面提到的特征变量提取与选择方法, 提取 4 个特定波长反射率(分别为 475.2, 501.4, 673.5 和 897.7 nm)、最优波段组合计算的 7 种植被指数以及两个主成分共 13 个特征变量。然后, 提取 129 个测试样本的上述 13 个特征值, 并与相应样方的植被覆盖度进行相关性分析(表 2)。

可以看出, 897.7 nm 处的近红外光波段与第一主成分与植被覆盖度的相关性不大, 即这两个特征的泛化能力较差, 因此在特征集中被去掉。利用依据变量重要性递归消除的方法, 对剩余 11 个特征变量进一步进行选择(表 3)。在逐步去掉特征的过程中, 前 5 个模型的 R^2 相差较小, 在第 6 个模型中去

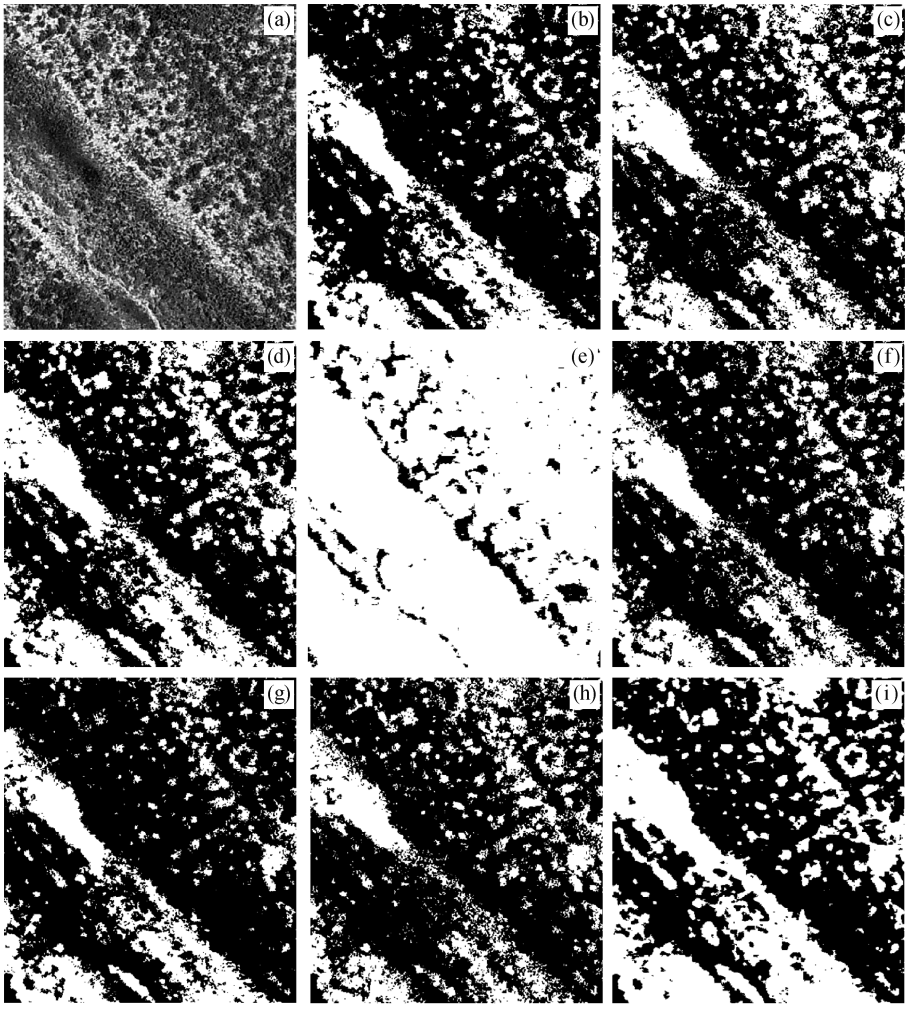


图 4　农作物区无人机样方影像及 8 种方法的植被覆盖度提取结果

Fig. 4　Image of the FVC sample in the crop area and corresponding FVC estimations using eight methods

表 1　地面样方(编号为2606)和农作物区内无人机影像样方的植被覆盖度估算精度(%)

Table 1　Accuracy of FVC estimation of the field sample (No. 2606)image and the UAV image in the crop area (%)

样方	RGRI	BGRI	RGBVI	EXG	VDVI	NGRDI	NGBDI	CIELab
地面	98.32	76.10	98.65	94.97	98.06	99.63	73.33	98.92
无人机	89.70	92.66	98.92	64.97	89.92	85.62	87.16	96.60

表 2　测试样本的特征变量与植被覆盖度相关性分析

Table 2　Correlation analysis between the feature variables and FVC values of the testing samples

特征	相关系数	特征	相关系数
NDVI	0.92	$\rho_{475.2}$	0.75
SR	0.88	$\rho_{501.4}$	0.74
EVI	0.89	$\rho_{673.5}$	0.76
SAVI	0.91	$\rho_{897.7}$	0.42
ARVI	0.92	PC1	0.30
MSAVI	0.91	PC2	0.84
RENDVI	0.91		

掉 EVI 后，测试集的 R^2 和 RMSE 均明显下降。因此，保留 EVI，最终得到重要性依次降低的 7 个特征为 SAVI, NDVI, ARVI, RENDVI, MSAVI, SR 和 EVI。

2.1.3 随机森林模型参数的确定

随机森林模型在 WEKA 开源智能分析环境下搭建。将训练样本划分为训练集和验证集，通过网格搜参并比较不同参数下验证集精度的方法寻找最优参数，最终确定随机森林模型中几个重要的参数值：决策树的个数为 100，深度为 11，每棵树使用

表 3 特征集选择过程

Table 3 Process of feature set selection

模型编号	特征集	R^2
1	NDVI, SR, EVI, SAVI, ARVI, MSAVI, RENDVI, $\rho_{475.2}$, $\rho_{510.4}$, $\rho_{673.5}$, PC2	0.932
2	NDVI, SR, EVI, SAVI, ARVI, MSAVI, RENDVI, $\rho_{510.4}$, $\rho_{673.5}$, PC2	0.931
3	NDVI, SR, EVI, SAVI, ARVI, MSAVI, RENDVI, $\rho_{673.5}$, PC2	0.928
4	NDVI, SR, EVI, SAVI, ARVI, MSAVI, RENDVI, PC2	0.926
5	NDVI, SR, EVI, SAVI, ARVI, MSAVI, RENDVI	0.923
6	NDVI, SR, SAVI, ARVI, MSAVI, RENDVI	0.901

的特征数为4。

2.2 植被覆盖度估算结果

利用所构建的随机森林模型对整个研究区的植被覆盖度进行估算, 结果如图5所示。从整体上来, 北边处于阳坡的尾矿坝地区植被覆盖度明显低于南边阴坡地区, 与半干旱地区阴坡长势优于阳坡的实地考察情况相吻合。阴坡大部分为灌木和乔木, 草本较少, 低植被覆盖度样本较少。阳坡草本与灌木相间分布, 高中低植被覆盖度样本均有。将阳坡估算结果图局部放大(图6), 发现植被覆盖度最高的地区对应长势好的大丛灌木, 与高空间分辨率可见光影像目视解译发现的灌木生长规律相同, 且大丛灌木的边缘较清晰。同时看出, 在道路中生长的稀薄的草层能与道路土石面区分开, 与利用卫星影像的传统方法相比, 利用无人机影像反演植被覆盖度具有明显的优势。并且, 高光谱影像能识别地面照片和无人机可见光影像中相对难以识别的泛黄的草本植物, 体现出光谱信息量大的优势。

2.3 精度评价

为了验证随机森林模型估算的精度, 利用129个测试样本进行植被覆盖度的估算, 并与植被覆盖度估算中广泛应用的像元二分法进行对比。像元二分法中, 除通常使用的NDVI外, 还利用ARVI, RENDVI, MSAVI, SAVI, SR和EVI等指数进行植被覆盖度估算。将研究区植被指数累积分布为95%和5%处的值作为相应像元二分模型中纯净植被和纯净土壤像元的植被指数值, 结果表明, NDVI估算的植被覆盖度精度最高, 与白彦[34]的研究结论相同。因此, 本文只对比使用NDVI的像元二分法与随机森林模型的估算精度。

以129个测试样本为参照, 分别计算随机森林模型和像元二分法估算结果的确定系数(R^2)、均方根误差(RMSE)和平均绝对误差(MAE), 结果表明, 像元二分法的预测结果中, 虽然R^2也较高, 但整体上存在较大偏差, RMSE和MAE分别为0.130和0.109(均大于0.1), 偏离真实结果的程度较大; 随机森林算法在提高模型拟合度的同时, 将误差降低, R^2达到0.923, RMSE和MAE分别降低至0.087和0.069。对比两个模型预测结果的分布(图7)可以看出, 随机森林模型对高中低植被覆盖度的估算误差相对均衡, 像元二分法则高估低植被覆盖区的植被覆盖度。

从误差分布直方图(图8(a))可见, 随机森林模型的估算误差基本上满足均值为0的正态分布, 而像元二分法的误差均值大于0(约为0.06)。同时, 随机森林回归的误差分布集中性更明显, 其值分布在0附近, 而像元二分法的误差分布较分散。利用二者误差分布的箱型图(图8(b))做进一步比较, 随机森林模型的误差中位线和均值接近0, 而像元二分法的误差中位线和均值均大于0, 并且, 随机森林模型的四分位间距远小于像元二分法, 50%以上的样本误差区间为[−0.05, 0.05]。这与随机森林回归的特性有关。无人机影像在拍摄时, 可能因姿态不稳造成地物反射率异常, 使用单一植被指数或多种植被指数的线性组合时, 若单点的植被指数出现波动, 会对结果造成显著影响。随机森林算法利用随机特征建立决策树, 最终根据多棵树的结果投票预测结果, 单一特征或几个特征的异常对最终结果的影响相对较小。作为预测效果最好的非参数模型之一, 随机森林模型对样本及特征的分布无严格假设, 不易出现估计结果整体上发生偏离的情况。

综上所述, 通过几个定量评价参数与同像元二分法误差分布的对比, 证明本文构建的植被覆盖度随机森林估算模型具有较高的精度, 尤其是针对无

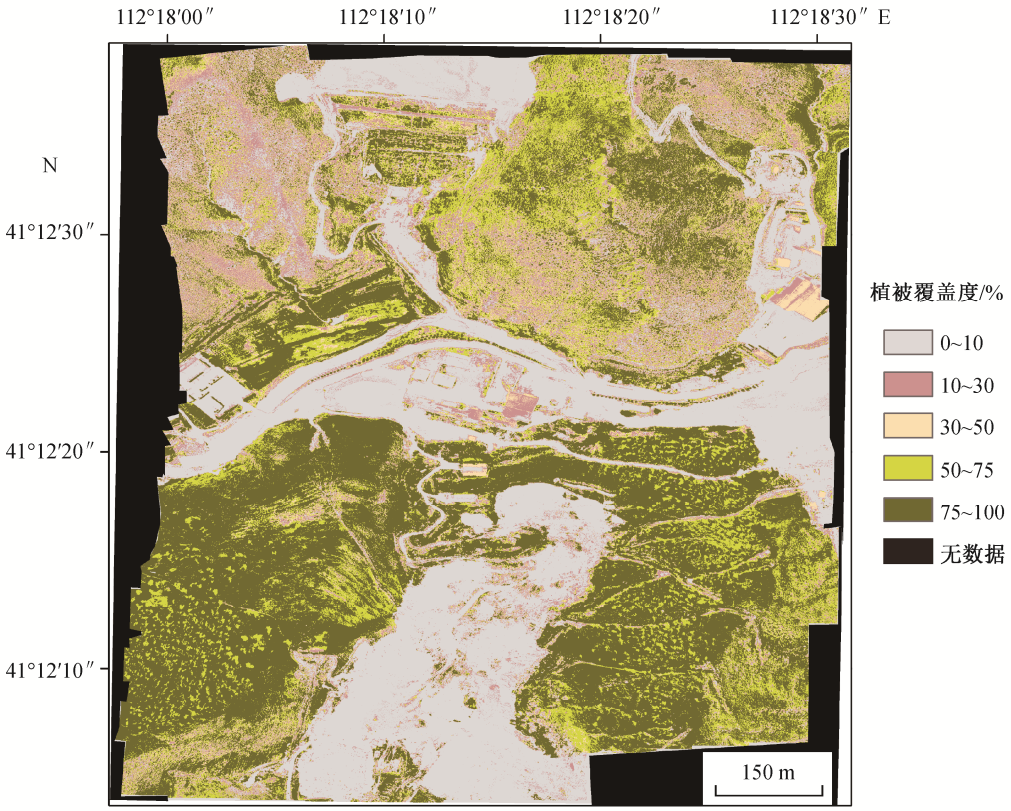


图 5　研究区植被覆盖度估算结果

Fig. 5　Estimation of fractional vegetation cover in the study area

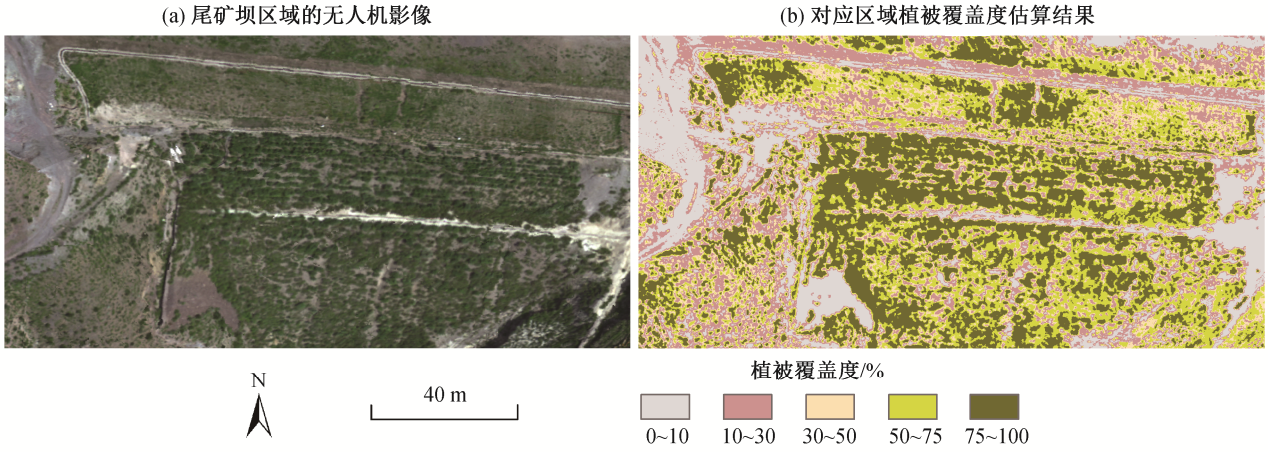


图 6　尾矿坝区域影像与植被覆盖度估算结果

Fig. 6　Image and estimation of FVC in the tailings dam area

人机姿态不稳定造成的反射率波动等问题，有明显的鲁棒性。

2.4　问题与讨论

本研究建模时采用的样方基本上为 2 m×2 m，而进行研究区植被覆盖度估算时使用的无人机高光谱影像的空间分辨率为 13.7 cm，它们之间的尺度存在差异，需对其影响进行探讨。首先，在研究区域内的选矿区选取一块包含高、中、低不同植被覆盖度的影像(约 60 m×80 m)，然后利用影像中每个像元的光谱反射率计算植被指数，将其输入模型中，得到该影像每个像元的植被覆盖度。然后，将影像划分为 2.05 m×2.05 m 的网格(对应 15×15 个像素)，计

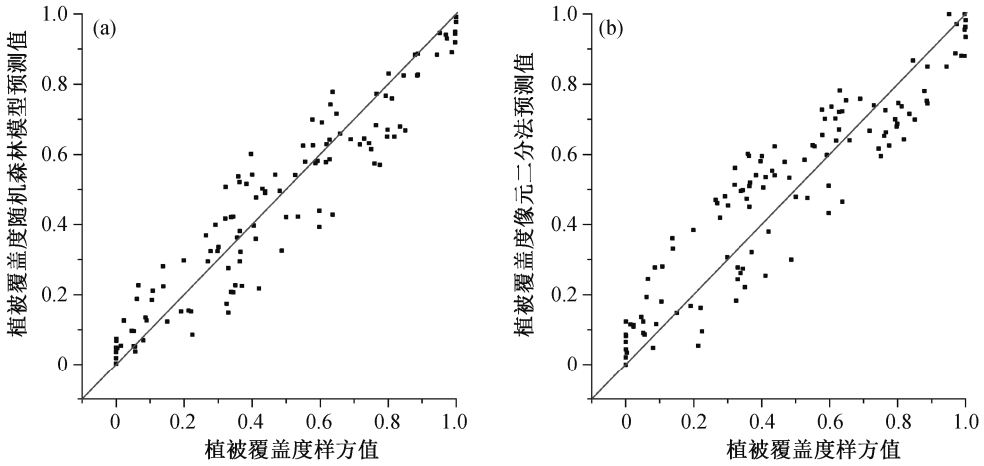


图 7 随机森林模型与像元二分法估算结果精度比较

Fig. 7 Accuracy of the FVC estimations using the Random Forest model and the Pixel Dichotomy method

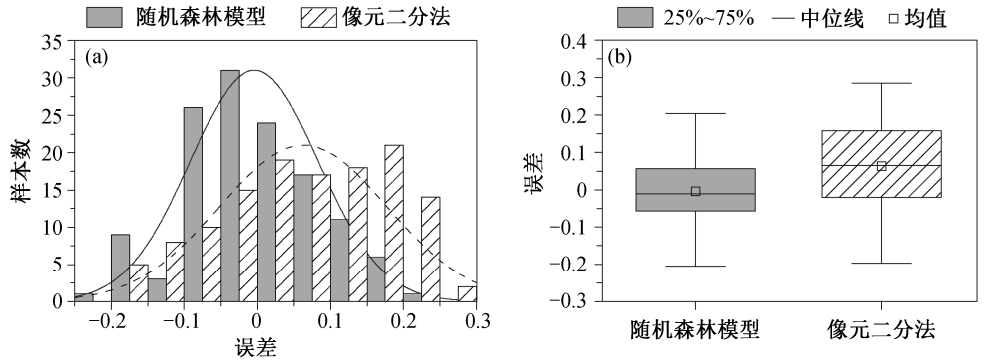


图 8 随机森林模型与像元二分法估算结果的误差分布

Fig. 8 Error distribution of random forest and pixel dichotomy models

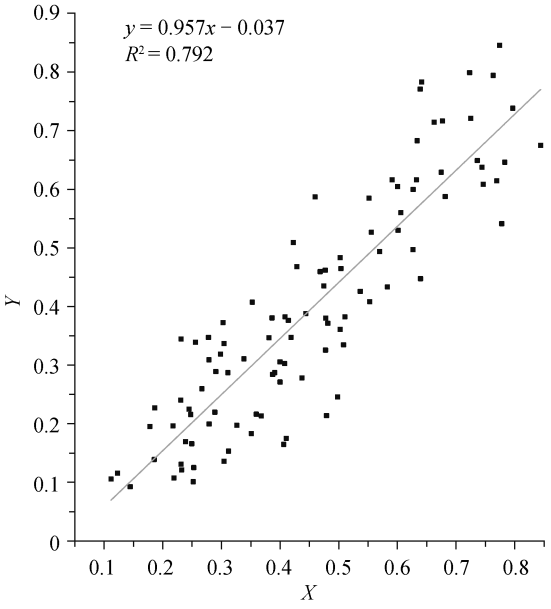


图 9 基于两种尺度样方数据的模型估算结果之间的相关性

Fig. 9 Correlation of FVC estimated by the models based on two scale quadrat training samples

算每个网格的平均光谱反射率, 利用模型得到每个网格的植被覆盖度。最后, 将各个网格中的像元植被覆盖度相加后, 除以各个网格对应的像元数, 转换为对应网格的植被覆盖度。将上一步中对网格的所有像元植被覆盖度求平均得到的网格植被覆盖度与利用网格平均光谱得到的植被覆盖度进行相关性分析, 结果如图 9 所示。图 9 中, X 轴表示 2 m×2 m 样方平均光谱估算的植被覆盖度, Y 轴表示 13.7 cm ×13.7 cm 像元估算的植被覆盖度在对应的 2 m×2 m 样方内的平均值。

利用像元尺度估算的植被覆盖度转换得到的 2 m 样方植被覆盖度与 2 m 网格内平均光谱估算的植被覆盖度的相关系数 R 为 0.89, 可认为二者具有强相关性, 说明在植被覆盖度的估算中, 基本上可以将光谱混合视为线性混合, 本文涉及的地面样方和高光谱影像间的尺度差异对所构建的植被覆盖度估算模型影响不大。

除随机森林回归模型外，也有一些学者使用支持向量机、神经网络以及多元线性回归等方法对植被覆盖度进行反演。本研究基于2.1.2节选择的含7个特征的特征集，相应地构建基于支持向量机、BP神经网络和多元线性回归的植被覆盖度估算模型。基于测试样本集进行估算精度评价后发现，与传统的像元二分法相比，几种机器学习模型的估算精度都有一定程度的提高。其中，随机森林模型的3种评价指标(R^2，RMSE和MAE)均为最优。像元二分法存在明显的高估现象，支持向量机模型存在一定程度的低估现象，而随机森林模型无明显的高估或低估。

3 结论

本文基于低空无人机获取的高光谱遥感影像，利用随机森林回归分析方法建立小尺度范围植被覆盖度的非参数估算模型。通过利用低空无人机获取的厘米空间分辨率可见光影像来加密地面植被样方数据，尤其对地面难以进行采样的山地、水域和植被茂密区域，解决了随机森林等机器学习方法需要大量训练样本的难题。同时，基于随机森林模型特征变量的重要性及递归特征消除算法，优化选取光谱波段、植被指数和主成分等3类特征变量，构建了具有较强鲁棒性的随机森林模型，以内蒙古自治区察右中旗油娄沟矿区为研究区，进行植被覆盖度的估算实验，得出如下结论。

1) 利用高空间分辨率的无人机可见光影像数据获取样方植被覆盖度具有较高的准确性，通过空地协同，可以快速获取大量样方数据，为人员难以进入地区的样方数据获取提供了新的方法。

2) 利用随机森林回归模型能较好地估算地形复杂的矿区植被覆盖度，确定系数(R^2)达到0.923，均方根误差和平均绝对误差分别为0.087和0.069。随机森林模型中计算的特征重要性表明，植被指数在反演植被覆盖度中具有重要作用。

3) 与传统的像元二分模型及支持向量机等其他机器学习模型相比，随机森林模型反演植被覆盖度具有更高的精度和更强的鲁棒性，避免了明显的高估或低估现象。

对本文提出的协同低空无人机采样和地面采样的训练样本获取方法的可靠性，尚需补充不同地貌和不同植被覆盖情景下的实例验证，在未来的研究中应进一步开展常规地面样方尺度与无人机可见光影像采样窗口大小对植被覆盖度估算精度的敏感性分析。在特征选择方面，可考虑将纹理信息及多时相高光谱信息加入模型特征集，进一步提高模型的估算精度。

参考文献

[1] 梁顺林，李小文，王锦地. 定量遥感：理念与算法. 北京：科学出版社, 2013

[2] Lin Z S, Qi X Z. Vegetation evolution with degenerating soil ecology under unequal competition. Pedosphere, 2004, 14(3): 355–361

[3] 张清春，刘宝元，翟刚. 植被与水土流失研究综述. 水土保持研究, 2002, 9(4): 96–101

[4] 程红芳，章文波，陈锋. 植被覆盖度遥感估算方法研究进展. 国土资源遥感, 2008, 20(1): 13–18

[5] 赵健赟. 地表植被覆盖度遥感估算及其气候效应研究进展. 测绘与空间地理信息, 2015, 38(8): 77–80

[6] 吕长春，王忠武，钱少猛. 混合像元分解模型综述. 遥感信息, 2003, 18(3): 55–58

[7] 陈彦兵. 基于混合像元分解的鄱阳湖湿地植被覆盖度提取. 科技创新与应用, 2017(29): 71, 73

[8] Xiao J, Moody A. A comparison of methods for estimating fractional green vegetation cover within a desert-to-upland transition zone in central New Mexico, USA. Remote Sensing of Environment, 2005, 98(2): 237–250

[9] Shoshany M, Kutiel P, Lavee H. Monitoring temporal vegetation cover changes in Mediterranean and arid ecosystems using a remote sensing technique: case study of the Judean Mountain and the Judean Desert. Journal of Arid Environments, 1996, 33(1): 9–21

[10] 马中刚，孙华，王广兴，等. 基于Landsat 8-OLI的荒漠化地区植被覆盖度反演模型研究. 中南林业科技大学学报, 2016, 36(9): 12–18

[11] 宋清洁，崔霞，张瑶瑶，等. 基于小型无人机与MODIS数据的草地植被覆盖度研究——以甘南州为例. 草业科学, 2017, 34(1): 40–50

[12] 陈妍，宋豫秦，王伟. 基于随机森林回归的草场植被盖度反演模型研究——以新疆阿勒泰地区布尔津县为例. 生态学报, 2018, 38(7): 2384–2394

[13] Baret F, Morissette J T, Fernandes R A, et al. Evaluation of the representativeness of networks of sites for the global validation and intercomparison of land biophysical products: proposition of the CEOS-BELMANIP. IEEE Transactions on Geoscience and Remote Sensing, 2006, 44(7): 1794–1803

[14] Jia K, Liang S, Gu X, et al. Fractional vegetation cover estimation algorithm for Chinese GF-1 wide field view data. Remote Sensing of Environment, 2016, 177(5): 184–191

[15] 蔡宗磊, 包妮沙, 刘善军. 国产高分一号数据估算草地植被覆盖度方法研究——以呼伦贝尔草原露天煤矿区为例. 地理与地理信息科学, 2017, 33(2): 32–44

[16] Coy André, Dale R, Michael T, et al. Increasing the accuracy and automation of fractional vegetation cover estimation from digital photographs. Remote Sensing, 2016, 8(7): 1–14

[17] Gamon J A, Surfus J S. Assessing leaf pigment content and activity with a reflectometer. New Phytologist, 2010, 143(1): 105–117

[18] Sellaro R, Crepy M, Trupkin S A, et al. Cryptochrome as a sensor of the blue/green ratio of natural radiation in Arabidopsis. Plant Physiology, 2010, 154(1): 401–409

[19] Neto J C. A combined statistical-soft computing approach for classification and mapping weed species in minimum-tillage systems [D]. Lincoln, NE: University of Nebraska, 2006

[20] 汪小钦, 王苗苗, 王绍强, 等. 基于可见光波段无人机遥感的植被信息提取. 农业工程学报, 2015, 31(5): 152–159

[21] Meyer G E, Neto J C. Verification of color vegetation indices for automated crop imaging applications. Computers and Electronics in Agriculture, 2008, 63 (2): 282–293

[22] Hunt E R, Cavigelli M, Daughtry C S T, et al. Evaluation of digital photography from model aircraft for remote sensing of crop biomass and nitrogen status. Precision Agriculture, 2005, 6(4): 359–378

[23] Song W, Mu X, Yan G, et al. Extracting the green fractional vegetation cover from digital images using a shadow-resistant algorithm (SHAR-LABFVC). Remote Sensing, 2015, 7(8): 10425–10443

[24] 郭震冬, 顾正东, 许盛, 等. 利用无人机技术进行社区植被覆盖率调查. 北京测绘, 2017, 24(5): 88–91

[25] 包刚, 包玉海, 覃志豪, 等. 高光谱植被覆盖度遥感估算研究. 自然资源学报, 2013, 28(7): 1243–1254

[26] 陈明华, 柴鹏, 陈文祥, 等. 不同植被指数估算植被覆盖度的比较研究. 亚热带水土保持, 2016, 28(1): 1–4

[27] 李晓松, 李增元, 高志海, 等. 基于Hyperion植被指数的干旱地区稀疏植被覆盖度估测. 北京林业大学学报, 2010, 32(3): 95–100

[28] 徐爽, 沈润平, 杨晓月. 利用不同植被指数估算植被覆盖度的比较研究. 国土资源遥感, 2012, 24(4): 95–100

[29] 魏秀红, 靳瑰丽, 范燕敏, 等. 基于高光谱遥感的退化伊犁绢蒿荒漠草地群落盖度估算. 中国草地学报, 2017, 39(6): 35–41

[30] Liu N, Budkewitsch P, Treitz P. Examining spectral reflectance features related to Arctic percent vegetation cover: implications for hyperspectral remote sensing of Arctic tundra. Remote Sensing of Environment, 2017, 192: 58–72

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

[32] 李欣海. 随机森林模型在分类与回归分析中的应用. 应用昆虫学报, 2013, 50(4): 1190–1197

[33] Tian S, Zhang X. Random forest classification of land cover information of urban areas in arid regions based on TH-1 data. Remote Sensing for Land & Resources, 2016, 28(1): 43–49

[34] 白彦. 呼伦贝尔沙地植被覆盖度变化遥感监测研究 [D]. 呼和浩特: 内蒙古农业大学, 2013