基于偏振反射模型和随机森林回归的叶片氮含量反演

张子晗[1]，晏 磊[1,2]，刘思远[1]，付 瑜[1]，姜凯文[1]，杨 彬[3]，刘绥华[4]，张飞舟[1*]

1. 北京大学地球与空间科学学院，遥感与地理信息系统研究所空间信息集成与3S工程应用北京市重点实验室，北京 100871
2. 桂林航空工业学院，广西高校无人机遥测重点实验室，广西 桂林 541004
3. 湖南大学电气与信息工程学院，湖南 长沙 410082
4. 贵州师范大学地理与环境科学学院，贵州 贵阳 550001

摘 要 叶片氮含量极大程度上影响植被生物化学过程，有重要的研究意义。利用机载高光谱数据反演叶片氮含量在农业遥感领域有广泛应用，但其反演精度不能完全满足精细农业的需要，有一定提升空间。叶片氮含量遥感反演精度受机理误差和算法误差的影响，机理误差主要来源于叶片表面反射。传感器探测到的反射辐射既包含叶片内部多次散射，又包含叶片表面镜面反射部分，只有前者是携带叶片内部生化组分(如氮含量)信息的，由于后者是入射光在叶表蜡质层发生的直接反射，因此该部分并不携带叶片内部信息。根据菲涅尔定律，叶表镜面反射是部分偏振的，而内部散射是非偏振的，因而通过偏振反射建模可部分去除叶表镜面反射影响，以消除机理误差。算法误差主要来源于不同氮含量反演算法对于高光谱数据挖掘能力的差别。比较了偏最小二乘法、主成分回归、支持向量机、K-近邻算法和随机森林回归在高光谱叶片氮含量反演中的表现，在调整算法参数之后，选择使用随机森林回归算法以减少高光谱反演算法误差。以常绿针叶林、落叶阔叶林和针阔混交林为研究对象，利用多角度偏振卫星POLDER/PARASOL的多光谱数据库构建二向偏振反射模型，用以模拟和分析研究区森林的偏振反射率；从HySpex传感器系统获取的机载高光谱数据中去除偏振反射率带来的光谱机理误差，以实现叶片氮含量的精确反演。以均方根误差为主要指标评估精度变化可获得以下结论：在高光谱叶片氮含量反演中，消除偏振反射率带来的机理误差后，各算法反演精度均有提升，平均提升了4.244%。其中，随机森林回归可以最大程度减小反演算法误差(可决系数达到0.803，均方根误差达到0.252)，且对光谱偏振信息最为敏感，去除偏振后精度提高了13.103%。相比于广泛使用的偏最小二乘算法，去除光谱机理误差并减小反演算法误差后，叶片氮含量反演精度整体提高了32.440%。该研究实现了基于机载高光谱数据的叶片氮含量精确反演，证明了在叶片氮含量反演中去除偏振反射率的必要性，体现了在高光谱氮含量反演中随机森林算法的应用潜力。

关键词 遥感反演；偏振遥感；叶片氮含量；高光谱数据；随机森林回归；双向偏振分布函数
中图分类号：S127 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)09-2911-07

引 言

在科研人员利用高光谱遥感数据进行植被内部重要物质含量反演之后的30多年里，高光谱遥感技术已广泛应用于研究与植被相关的多种重要的生物化学物质，例如：水分、纤维素、蛋白质、叶绿素、氮元素、硫元素和磷元素等。叶片氮元素含量(leaf nitrogen concentration，LNC)与植物光合呼吸作用和其他的相关生物过程密切相关，其含量仅占叶片总质量的0.2%～6.4%，但却能极大程度上影响全球气候变化过程，并在广域碳氮循环过程中起到主导作用，对于解构生态环境规律有重大科学意义。氮元素在叶片内部的主要依托为蛋白质和叶绿素，这两种物质在可见光近红外(400～2 500 nm)波段有独特的吸收特性，因而可以利用高光谱数据进行氮元素含量反演。

利用高光谱数据进行氮含量反演的算法可以大致分为三

收稿日期：2020-09-26，**修订日期**：2021-01-08
基金项目：国家重点研发计划项目(2017YFC0210102，2017YFB0503004)，国家自然科学基金项目(41801227)，湖南省自然科学基金项目(2019JJ50047)，贵州省教育厅项目(喀斯特山地生态环境保护与资源利用协同创新中心)资助
作者简介：张子晗，1997年生，北京大学地球与空间科学学院博士研究生 e-mail：zzh_cytus@pku.edu.cn
*通讯作者 e-mail：zhangfz@pku.edu.cn

类：第一类方法建立在叶片反射率与叶片氮含量之间统计关系的基础上，以逐次线性拟合回归(stepwise multiple linear regression，SMLR)和偏最小二乘回归(partial least squares regression，PLSR)为代表[1]，在高光谱波段数较多时，这类算法的回归计算比较耗时，精度相比于后两类方法更低，但由于算法原理简单明晰，仍被广泛使用；第二类算法建立在高光谱植被指数的基础上，通过计算高光谱植被指数来反演叶片氮含量[2-3]，这类算法通过提取红边等深层次光谱特征来构建植被指数，对于高光谱数据的挖掘仍旧不够深入，但由于计算简单，常用于在农学相关的应用领域粗略估计叶片氮含量；第三类算法建立在机器学习的基础上，以支持向量机(support vector machine，SVM)、人工神经网络(artificial neural network，ANN)和随机森林(random forest，RF)为代表[4-5]，这类算法属于遥感领域与计算机领域学科交叉的范畴，具有极大的精度优势。不同的算法选择会产生算法固有误差，影响氮含量反演精度，为了尽可能减小算法误差，同时尽可能利用高光谱数据蕴含的信息，本研究中测试了多种氮含量反演方法，最终以反演精度为指标，选择了第三类算法中的 RF 算法作为本研究中叶片氮含量反演使用的主体框架。

上述三类氮含量反演算法的输入参数皆为高光谱反射率，但从定量遥感分析的角度来看，光谱测量得到的反射率包括两个部分的贡献：一部分来源于入射光在叶片内部的多次散射过程，另一部分来源于入射光在叶片表面的镜面反射过程[6]。在多次散射过程中，入射光大部分透过叶片表面进入叶片内部，与叶片内部结构中的叶绿素和蛋白质等含氮生物化学物质发生相互作用，在叶片内部多次散射后，携带氮元素信息离开叶片，这部分来自于叶片内部多次散射的漫反射率所携带的光谱信息是氮含量反演密切关注的。而在镜面反射过程中，入射光部分在叶表蜡质层直接发生镜面反射，这部分反射能量不包含与叶片内部叶绿素和蛋白质等含氮生物化学物质相关的信息，在氮含量反演中属于高光谱反射率自带的数据误差，如果使用包含镜面反射部分的高光谱反射率直接进行氮含量反演会产生机理误差，影响氮含量反演精度。依据菲涅尔原理，叶片反射光的镜面反射组分是部分偏振的，而叶片反射光的漫反射组分由于在叶片内部多次散射导致的消光作用是非偏振的。因此，通过偏振反射建模方法获得双向偏振分布函数(bidirectional polarization distribution function，BPDF)，可模拟计算偏振反射率，进而扣除测量获取的叶片高光谱反射率中与氮含量反演无关的镜面反射组分，从源头减少机理误差，提高氮含量反演精度。

综上所述，影响叶片氮含量反演精度的误差来源有两个：算法误差和机理误差。选择 RF 算法进行叶片氮含量反演算法可以尽可能减小算法误差，构建 BPDF 可以去除大部分机理误差，最终实现基于 BPDF 和 RF 的叶片氮含量精确反演。

1 实验部分

1.1 氮含量反演数据

本研究中用于氮含量反演的数据为主数据集。叶片氮元素反演的研究地点为德国巴伐利亚国家森林公园[7](49°3′19″N，13°12′9″E)。根据优势树种的类型，选择使用国际地圈生物圈计划(international geosphere-biosphere programme，IGBP)土地覆盖分类体系，优势树种与 IGBP 类别的对应情况见表 1。

表 1 巴伐利亚国家森林公园内优势树种对应的 IGBP 类别

Table 1 IGBP classes of dominant vegetation in Bavarian Forest National Park

树种	IGBP 类别	类名
欧洲云杉	IGBP01	常绿针叶林
欧洲山毛榉	IGBP04	落叶阔叶林
欧洲冷杉	IGBP01	常绿针叶林
假挪威槭	IGBP04	落叶阔叶林
欧洲花楸	IGBP04	落叶阔叶林

由于本研究区域内的优势树种中阔叶树木均为落叶阔叶品种，针叶树木都为常绿针叶品种，所以在研究区域中被简单标注为 3 种类型：落叶阔叶林(IGBP04)、常绿针叶林(IGBP01)和针阔混交林(IGBP05)，其点位空间分布如图 1 所示。

共有 26 个点位作为氮元素反演研究区(8 片落叶阔叶林，8 片常绿针叶林，10 片针阔混交林)。其中，每个采样区在实地占地为 30 m×30 m，点位范围用分层随机抽样的方法选择 8～9 棵树木采样，单颗树木至少采集 20 片叶片，妥善保存好后在实验室用凯式定氮法测量氮元素含量。高光谱数据的获取时间为 2013 年 7 月 22 日，利用 HySpex 传感器系统在 3 000 m 高度获取高光谱影像(HySpex 传感器系统包括 2 台高光谱成像光谱仪，一台覆盖 400～1 000 nm 的光谱范围，另一台覆盖 1 000～2 500 nm 的光谱范围)，高光谱数据共有 418 个窄波段，410～2 495 nm 内，每 5 nm 一个通道。氮含量反演使用的原始反射率是经过几何校正和大气校正处理后的 26 个样区内的 418 个波段的双向反射因子数据(bidirectional reflectance factor，BRF)，除了 BRF 数据之外无人机系统还记录了获取数据时的观测几何(太阳天顶角、观测天顶角、相对方位角)。

1.2 偏振反射建模数据

偏振反射建模数据为辅数据集。本研究通偏振反射建模获取 BPDF，进而估计偏振反射率。用于偏振反射建模的 BRDF-BPDF 数据库[8]由搭载在 PARASOL 卫星上的 POLDER 传感器获取。本研究中为了建立合适的 BPDF 模型所使用的数据来自于 POLDER 经过校正的产品，包含了数据获取时的经纬度和观测几何、地物的 IGBP 类别以及地物类型同质性、样区的归一化植被指数(normalized difference vegetation index，NDVI)、气溶胶情况、6 个波段(490，565，670，765，865 和 1 020 nm)的反射率和 865 nm 波段偏振反射率。

该 BRDF-BPDF 数据库从 POLDER 的连续 7 年的数据中挑选出质量最好的 1 年，即 2008 年，同时，在 POLDER 数据中挑选了质量最优秀(尽可能满足对全球范围的覆盖、

数据有效性高、样区 IGBP 类别匀质性高)的点位。最终数据库的时间跨度和点位选择定为：包括 2008 年 12 个月份的数据，每个月份采取 50 个均匀分布在全球的点位，在每个点位上按照不同的观测几何选择高质量 POLDER/PARASOL 体系的数据。本研究使用该数据库中 2008 年 7 月份欧洲范围内的数据。

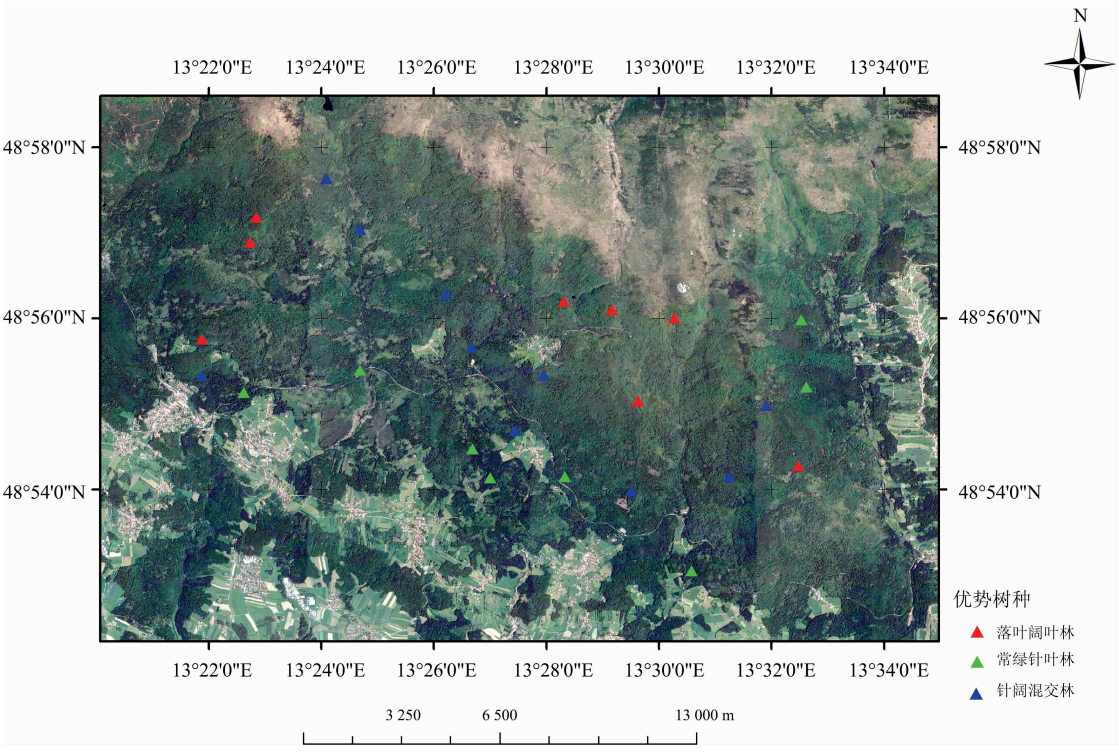

图 1　叶片氮元素反演点位空间分布

Fig. 1　Spatial distribution of foliar nitrogen content retrieval plot

1.3　数据一致性说明

使用两组相对独立的数据集，主数据集用于氮含量反演，辅数据集用于建立偏振反射模型，使用主数据集和辅数据集的流程图如图 2 所示。从数据获取的时间来看，考虑到偏振反射建模具有较强的季节周期性特征，为了与巴伐利亚国家森林公园的氮含量反演高光谱数据的获取时间契合，在进行偏振反射建模时仅使用 BRDF-BPDF 数据库数据中与机载数据同月份的数据，以保证时间一致性。从数据获取的地理位置来看，由于氮含量反演数据的研究位置在德国，同时考虑到在 BRDF-BPDF 数据库数据中获取的点位数量不能太少，最终根据 BRDF-BPDF 数据库数据的经纬度，筛选出欧洲范围内的数据进行偏振反射建模，以保证地理位置一致性。从两组数据集的空间分辨率来看，POLDER 数据的分辨率为 6.5 km，无人机高光谱数据的分辨率为 1.65 m(400～1 000 nm)和 3.3 m(1 000～2 500 nm)，但由于 BRDF-BPDF 数据库中收录的 POLDER 数据匀质性大于 75%，可认为在数据匹配时受空间尺度的影响较小。

1.3　方法

1.3.1　偏振反射建模

研究中获取的高光谱数据为 418 个窄波段的 BRF。根据菲涅尔定律，地表反射是部分偏振的，其中的偏振部分主要为线偏振，也是由观测几何(太阳天顶角、观测天顶角、相对方位角)决定的，可用 BPDF 表征其多角度偏振反射特性。有文献表明，在可见光近红外波段，BPDF 可视为观测几何稳定时的光谱不变量，在不同波段的偏振反射率为定值[9]。BPDF 的决定因素是地物类型和观测几何，为了去除地表反射率的偏振部分，需要利用 BRDF-BPDF 数据库中的 7 月份欧洲范围的数据，建立 3 个分别适应 IGBP01，IGBP04 和 IGBP05 这 3 种不同地表 IGBP 类别的 BPDF 模型以模拟估计偏振反射率(polarization bidirectional reflectance factor，PBRF)，进而从 BRF 中去扣除具有部分偏振特性的镜面反射组分——偏振反射率 PBRF，最终获得漫反射反射率(diffuse bidirectional reflectance factor，DBRF)，以消除机理误差的影响。

目前精度较高的 BPDF 模型几乎全部为半经验模型，纯粹的经验模型虽然能够在局部较好拟合偏振反射率，但由于缺乏物理机理，模型参数严重依赖训练数据，且过拟合明显。半经验模型通过较少的经验参数对包含物理意义的模型进行校正调整，其结果是更具有普适性。本研究选择了 5 个应用最广、精度最高的 BPDF 模型用于模拟 IGBP01，IGBP04 和 IGBP05 这 3 种不同地表 IGBP 类别下的偏振反射率 PBRF，分别为：Nadal 模型[10]、Waquet 模型[11]、Maignan 模型[12]、Litvinov 模型[13]和 Diner 模型[14]。

为了使公式简洁明了，在 BPDF 模型的公式表达中使用一些简记。其中，太阳光方向与天顶方向(其夹角称为太阳天顶角 θ_S，观测方向与天顶方向的夹角称为观测天顶角 θ_V，

太阳光方向在地表的投影与观测方向在地表的投影之间的夹角称为相对方位角 ϕ，镜面反射过程中的法线与天顶方向的夹角称为半天顶角 θ_h，入射方向与反射方向之间的夹角的一半为入射角 α_I，折射角为 α_T，入射方向与反射方向之间方向的变化为方向散射角 γ，折射率为 N。部分角度的余弦值简记为

$$\mu_S = \cos\theta_S \quad (1)$$

$$\mu_V = \cos\theta_V \quad (2)$$

$$\mu_I = \cos\alpha_I \quad (3)$$

$$\mu_T = \cos\alpha_T \quad (4)$$

由菲涅尔公式 $F_p(\gamma, N)$得到的变量之间的相关关系见式(5)

$$F_p(\gamma, N) = \frac{1}{2}\left[\left(\frac{N\mu_T - \mu_I}{N\mu_T + \mu_I}\right)^2 - \left(\frac{N\mu_I - \mu_T}{N\mu_I + \mu_T}\right)^2\right] \quad (5)$$

5 种模型的偏振反射建模公式表达见表 2。

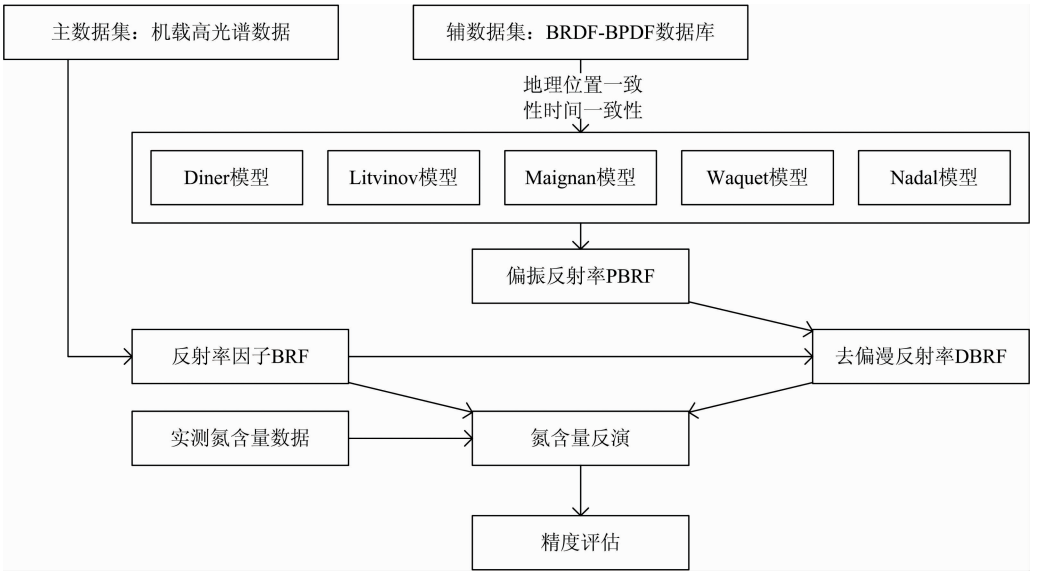

图 2 数据使用流程图

Fig. 2 Flowchart of data utilization

表 2 5 种偏振反射模型的公式表达

Table 2 Formulas of 5 BPDF models

模型名称	公式	应用场景	自由参数
Nadal 模型	$\mathrm{PBRF} = A\left[1 - \exp\left(-B\frac{F_p(\gamma, N)}{\mu_S + \mu_V}\right)\right]$	森林、灌木林和低植被地表覆盖类型	A, B
Waquet 模型	$\mathrm{PBRF} = AF_p(\gamma, N)S(\theta_S)S(\theta_V)$ [其中，$S(\theta)$ 为后向散射中最大值的遮蔽函数]	森林、作物种植地和城市地表覆盖类别	A, B
Maignan 模型	$\mathrm{PBRF} = A\frac{\exp(-\tan\alpha_I)\exp(-\mathrm{NDVI})F_p(\gamma, N)}{4(\mu_S + \mu_V)}$	植被地表覆盖类型	A
Litvinov 模型	$\mathrm{PBRF} = A\frac{\pi F_p(\gamma, N)}{4(\mu_S + \mu_V)\cos\theta_H}S(\gamma, C)f(B, \theta_H)$ [其中，$f(B, \theta_H)$ 表征的是地表在空间中的高斯分布，$S(\gamma, C)$ 是遮蔽函数]	植被和土壤地表覆盖类型	A, B, C
Diner 模型	$\mathrm{PBRF} = A\frac{F_p(\gamma, N)}{8\pi\mu_S\mu_V\cos\theta_H}$	草地地表覆盖类型	A

利用辅数据 BRDF-BPDF 数据库中的数据分别建立 IGBP01，IGBP04 和 IGBP05 这 3 种地表覆盖类型下的 5 种 BPDF 模型，分别拟合得到自由参数。通过十折交叉验证的数据分组策略，以均方根误差(root mean square error，RMSE)和可决系数(coefficient of determination，RSQ)为指标，挑选 3 种地表覆盖类型下的最优 BPDF 模型来模拟偏振反射率 PBRF。

1.3.2 氮含量反演

氮含量反演研究基于德国巴伐利亚国家森林公园的实测氮含量主数据、无人机高光谱主数据和偏振反射建模获得模拟仿真结果，研究主要包括三个步骤：首先，通过 BPDF 模拟计算偏振反射率 PBRF，进而计算漫反射率 DBRF，获取氮含量反演的算法输入对照组(原始反射率 BRF 和去偏漫反射率 DBRF)；然后，设置 RF 算法参数，利用整体散射系数和

去偏后的漫反射散射系数作为算法输入对照组进行叶片氮含量反演；最后，横向对比 RF 算法和其他算法的表现，并对偏振反射率在氮含量反演中的影响进行定量评估。

1.3.3　精度评估

偏振反射建模使用 POLDER/PARASOL 体系的数据库中数据量较大，采用十折交叉验证的思路进行分组后验证精度，重复 10 次十折交叉验证，取均方根误差 RMSE 和可决系数 RSQ 的均值作为 BPDF 模型精度的度量。氮含量反演使用 26 个点位的高光谱数据受到数据量的限制，采用留一法交叉验证的思路进行数据分组后验证精度，同样以均方根误差 RMSE 和可决系数 RSQ 作为叶片氮含量反演精度的度量，见式(6)和式(7)。

$$\mathrm{RMSE}=\sqrt{\frac{\sum(\hat{x}-\tilde{x})^2}{n}} \tag{6}$$

$$\mathrm{RSQ}=r^2,\ r=\frac{\mathrm{cov}(\hat{x},\tilde{x})}{\sqrt{\mathrm{var}(\hat{x})\mathrm{var}(\tilde{x})}} \tag{7}$$

其中，x 为研究对象，$\hat{x}$ 为变量预测值，$\tilde{x}$ 为变量实际测量值。在评估 BPDF 模型精度时 x 代表偏振反射率 PBRF，在评估氮含量反演精度时 x 代表叶片氮含量 LNC。

2　结果与讨论

2.1　BPDF 模型选择结果与分析

首先，按照氮含量反演研究时间、区域进行辅数据集 BRDF-BPDF 数据库的筛选，选择了 POLDER 数据库中 2008 年 7 月采集的欧洲范围内的数据；然后，利用十折交叉验证的数据划分思路将数据库中的数据随机划分为训练集和验证集，利用训练集数据拟合获得 5 个 BPDF 模型的自由参数；获得参数组后利用中位数方法求得单样区最优模型参数组，根据该模型参数组预测偏振反射率 PBRF；将预测得到的偏振反射率与包含于数据库中的实测 865 nm 偏振反射率数据进行对比，最后利用均方根误差 RMSE 和可决系数 RSQ 来评价 BPDF 模型的效果，5 种模型的评价结果见表 3。

表 3　IGBP01/04/05 地表覆盖类型下 5 种 BPDF 模型精度评价

Table 3　Accuracy assessment of 5 BPDF models on IGBP01/04/05

类别	Nadal 模型		Waquet 模型		Maignan 模型		Litvinov 模型		Diner 模型	
	RSQ	RMSE	RSQ	RMSE	RSQ	RMSE	RSQ	RMSE	RSQ	RMSE
IGBP01	0.697	0.003	0.690	0.003	0.695	0.003	0.688	0.003	0.655	0.003
IGBP04	0.875	0.002	0.857	0.002	0.871	0.002	0.878	0.002	0.821	0.002
IGBP05	0.812	0.002	0.789	0.002	0.818	0.002	0.808	0.002	0.759	0.002

最终，以可决系数 RSQ 为主要评判指标，均方根误差 RMSE 为辅助评判指标，在 RSQ 尽可能大的情况下选择 RMSE 尽量小的 BPDF 模型。从整体来看，在同一地表覆盖类型下不同 BPDF 模型的模拟精度差异不大(保留 3 位小数后，均方根误差 RMSE 差异极小，可决系数 RSQ 差异相对明显)，而在不同地表覆盖类型下，IGBP04 地表覆盖类型的 BPDF 模型表现显著优于 IGBP01 和 IGBP05，说明阔叶树木的偏振效应相较于针叶树木而言更易模拟，推测可能与起偏介质的面积大小有关。模型选择结果为：在 IGBP01 地表覆盖类型下最优 BPDF 模型为 Nadal 模型，在 IGBP04 地表覆盖类型下最优 BPDF 模型为 Litvinov 模型，在 IGBP05 地表覆盖类型下最优 BPDF 模型为 Maignan 模型。由于不同地表覆盖类型的起偏特性不同，最优 BPDF 模型并不统一，分别利用选出的 3 种 BPDF 模型模拟对应的 3 种地表覆盖类型的偏振反射率。

2.2　RF 模型参数设置与叶片氮反演精度横向对比分析

在偏振反射建模的基础上利用原始反射率 BRF 和漫反射率 DBRF 作为 RF 算法输入进行叶片氮含量反演。RF 的核心参数为决策树数量，本研究分析比对了去偏前后决策树数量为 5，10，50 和 100 的 RF 算法的表现，见表 4。研究结果说明，在决策树数量一定时，使用去偏数据进行叶片氮含量反演普遍能获得精度提升且在 RF 决策树数量设定为 10 时获得最优氮含量预测模型。利用 RF 算法进行氮含量反演能获取的最高精度为：可决系数达到 0.803，均方根误差达到 0.252。

表 4　不同决策树数量的 RF 算法精度评价

Table 4　RF algorithm accuracy assessment of different estimator numbers

决策树数量	去偏前		去偏后	
	RSQ	RMSE	RSQ	RMSE
5	0.722	0.307	0.747	0.300
10	0.769	0.290	0.803	0.252
50	0.732	0.303	0.746	0.292
100	0.720	0.309	0.729	0.303

2.3　偏振反射率占比统计分析

在 26 个采样点位，偏振反射率 PBRF 在整体反射率 BRF 中的平均占比为 3.580%～11.649%，各个点位内的偏振反射率占比箱形图见图 3，箱形图形象地展示了偏振反射率 PBRF 占比的算数均值、最小值、下四分位数、中位数、上四分位数和最大值。该分布表明偏振反射率 PBRF 在整体反射率 BRF 中平均占比为 3.350%，最高占比 16.519%，说明偏振反射率的影响是不可忽略的，机理误差足以影响反演精度，在高光谱叶片氮含量反演中有去除偏振反射率 PBRF 的必要性。

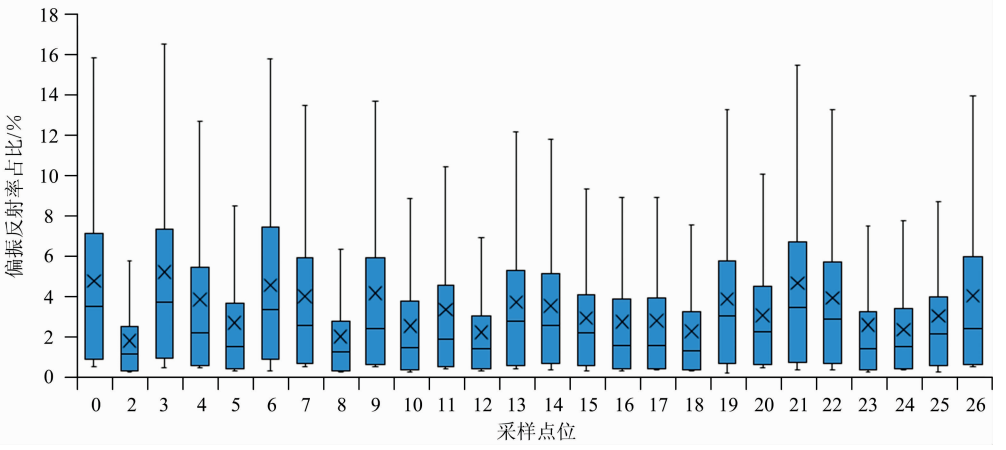

图 3 偏振反射率在整体反射率中占比的分布

图中箱体内的×表示算数均值，箱体内的黑色横线表示中位数

Fig. 3 Ratio of PBRF to BRF

In this figure, the × in the box represents the mean value, and the transverse line in the box represents the median value

2.4 叶片氮反演精度横向对比分析

横向对比了 PLSR 算法、主成分回归(principal component regression，PCR)算法、支持向量回归(support vector regression，SVR)算法、K-近邻(K-nearest neighbor，KNN)回归算法和 RF 回归算法在氮含量反演中的表现，利用留一法进行交叉验证，精度评价结果见表 5。其中，由于偏最小二乘算法应用最为广泛[1]，将其作为基线算法进行整体精度评估。研究结果说明，在利用上述方法进行氮含量反演的过程中，去除偏振后叶片氮含量反演的精度普遍都有提升(KNN 算法在选择的最邻近元数目 K 足够大时，最邻近元的选择在去除偏振前后相同，因而模型预测结果相同，无精度提升)。以均方根误差 RMSE 为核心指标评估提升幅度，平均提升幅度为 4.244%，其中 RF 算法去除偏振后叶片氮含量反演精度提升幅度最大，为13.103%，且精度最优(达到了上述 5 种反演算法中的最高可决系数 0.803 和最低均方根误差 0.252)，充分说明了 RF 算法利用高光谱数据进行叶片氮含量反演的潜力。相较于作为基线算法的 PLSR(未进行光谱去偏)，通过偏振反射建模去除大部分光谱机理误差并通过选择随机森林模型减小算法误差后，叶片氮含量反演精度整体提高 32.440%，证明了去除机理误差和减少算法误差的必要性，实现了高精度叶片氮含量反演。

表 5 5 种叶片氮含量反演算法精度评价

Table 5 Accuracy assessment of 5 LNC retrieval algorithms

算法名称	去偏前		去偏后	
	RSQ	RMSE	RSQ	RMSE
PLSR	0.595	0.373	0.624	0.359
PCR	0.678	0.329	0.680	0.328
SVR	0.732	0.304	0.735	0.302
KNN	0.748	0.291	0.748	0.291
RF	0.769	0.290	0.803	0.252

3 结 论

研究旨在提高叶片氮含量反演精度，从机理误差和算法误差两个误差源入手，通过偏振反射率建模去除了大部分机理误差，通过选择精度最优的随机森林算法并调整参数尽可能降低了算法误差，最终实现了叶片氮含量精确反演，相比于基线方法整体精度提升为 32.440%(可决系数 RSQ 达到 0.803，均方根误差 RMSE 达到 0.252)。同时，在讨论中定量评估了偏振反射率 PBRF 在整体反射率 BRF 中的占比分布并对比了去除偏振前后多种氮含量反演算法的精度变化，体现了随机森林算法在高光谱信息挖掘中的优势，也证明了去除不携带氮含量信息的偏振反射率的必要性。

致谢：本研究使用的德国巴伐利亚国家森林公园叶片氮含量实测数据和无人机高光谱数据由威斯康星大学麦迪逊分校的王智慧博士提供，在此表示衷心的感谢！

References

[1] Kokaly R F, Asner G P, Ollinger S V, et al. Remote Sensing of Environment, 2009, 113: S78.

[2] Li F, Mistele B, Hu Y, et al. European Journal of Agronomy, 2014, 52: 198.

[3] Yao X, Zhu Y, Tian Y, et al. International Journal of Applied Earth Observation and Geoinformation, 2010, 12(2): 89.

[4] Chlingaryan A, Sukkarieh S, Whelan B. Computers and Electronics in Agriculture, 2018, 151: 61.

[5] Yao X, Huang Y, Shang G, et al. Remote Sensing, 2015, 7(11): 14939.

[6] Yang B, Zhao H, Chen W. Journal of Quantitative Spectroscopy and Radiative Transfer, 2017, 202: 13.

[7] Wang Z, Skidmore A K, Wang T, et al. International Journal of Applied Earth Observation and Geoinformation, 2017, 54: 84.
[8] Bréon F M, Maignan F. Earth System Science Data, 2017, 9(1): 31.
[9] Knyazikhin Y, Martonchik J V, Myneni R B, et al. Journal of Geophysical Research: Atmospheres, 1998, 103(D24): 32257.
[10] Nadal F, Bréon F M. IEEE Transactions on Geoscience and Remote Sensing, 1999, 37(3): 1709.
[11] Waquet F, Leon J F, Cairns B, et al. Applied Optics, 2009, 48(6): 1228.
[12] Maignan F, Bréon F M, Fédèle E, et al. Remote Sensing of Environment, 2009, 113(12): 2642.
[13] Litvinov P, Hasekamp O, Cairns B. Remote Sensing of Environment, 2011, 115(2): 781.
[14] Diner D, Xu F, Martonchik J, et al. Atmosphere, 2012, 3(4): 591.

Leaf Nitrogen Concentration Retrieval Based on Polarization Reflectance Model and Random Forest Regression

ZHANG Zi-han[1], YAN Lei[1,2], LIU Si-yuan[1], FU Yu[1], JIANG Kai-wen[1], YANG Bin[3], LIU Sui-hua[4], ZHANG Fei-zhou[1*]

1. Beijing Key Lab of Spatial Information Integration and 3S Application, Institute of Remote Sensing and Geographic Information System, School of Earth and Space Science, Peking University, Beijing 100871, China
2. Guangxi Key Lab of UAV Remote Sensing, Guilin University of Aerospace Technology, Guilin 541004, China
3. College of Electrical and Information Engineering, Hunan University, Changsha 410082, China
4. School of Geography and Environmental Science, Guizhou Normal University, Guiyang 550001, China

Abstract Leaf nitrogen concentration is of great significance in the vegetation biochemistry process. Airborne hyperspectral data is widely utilized to retrieve leaf nitrogen concentration. Since the current algorithms cannot completely fulfill the accuracy requirement of precision agriculture, it is urgent to improve the retrieval accuracy of leaf nitrogen concentration. The accuracy of leaf nitrogen concentration retrieval is restricted by principle error and algorithm error. The principle error is generated in the process of specular reflection at the leaf surface. The radiant energy detected by sensors consists of a specular components and multiple scattering components. Solely the multiple scattering component carries vegetation biochemistry information (leaf nitrogen concentration, for instance). The specular component represents the energy reflected directly at the foliar wax layer, thus carries no inner information of the leaf. Based on the Fresnel formula, the specular component is partially polarized, and the multiple scattering component is unpolarized. Therefore, the principle error can be eliminated by the specular reflectance estimate, particularly with the aid of polarization reflectance modelling. The algorithm error is derived from the difference of airborne hyperspectral data mining capability between different algorithms. The performance of Partial Least Squares Regression, Principal Component Regression, Support Vector Regression, K-Nearest Neighbor Regression and Random Forest Regression are systematically compared in this research, and ultimately Random Forest Regression is chosen to reduce the algorithm error. In this research, in order to estimate the polarization reflectance of broadleaf and needle vegetation, multispectral data gained by POLDER/PARASOL satellite (equipped with multi-angle polarization sensors) are used to establish Bidirectional Polarization Distribution Function model. Hyperspectral data gained by the HySpex sensor system is used to conduct high-precision retrieval of leaf nitrogen concentration. Root Mean Square Error is taken as a major evaluation index. The conclusion is: After eliminating polarization reflectance in hyperspectral data, an average accuracy improvement of 4.244% is achieved among the above algorithms. Random Forest Regression is rather competitive by reaching 13.103% improvement in accuracy (RSQ 0.803, RMSE 0.252), which indicates that Random Forest is sensitive to polarization information. Compared to the basic method (Partial Least Squares Regression), the accuracy is improved by 32.440% after eliminating principle error and reducing algorithm error. In our research, the high-accuracy retrieval of leaf nitrogen concentration is realized, proving the significance of eliminating polarization reflectance and indicates the potential of random forest regression in hyperspectral remote sensing retrieval.

Keywords Remote sensing retrieval; Polarization remote sensing; Leaf nitrogen concentration; Hyperspectral data; Random forest regression; Bidirectional polarization distribution function

* Corresponding author

(Received Sep. 26, 2020; accepted Jan. 8, 2021)