DOI:10. 7522/j. issn. 1000-0240. 2022. 0026
SUN Xingliang, HAO Xiaohua, WANG Jian, et al. Research on retrieval of MODIS fraction snow cover based on spectral environmental random forest regression model[J]. Journal of Glaciology and Geocryology, 2022, 44(1):147-158. [孙兴亮, 郝晓华, 王建, 等. 基于光谱-环境随机森林回归模型的MODIS积雪面积比例反演研究[J]. 冰川冻土, 2022, 44(1):147-158.]

基于光谱-环境随机森林回归模型的MODIS积雪面积比例反演研究

孙兴亮[1,2,3,4], 郝晓华[2], 王 建[2], 赵宏宇[5], 纪文政[2]

(1. 兰州交通大学 测绘与地理信息学院,甘肃 兰州 730070; 2. 中国科学院 西北生态环境资源研究院,甘肃 兰州 730000;
3. 地理国情监测技术应用国家地方联合工程研究中心,甘肃 兰州 730070; 4. 甘肃省地理国情监测工程实验室,
甘肃 兰州 730070; 5. 北京师范大学 地表过程与资源生态国家重点实验室,北京 100875)

摘 要: 积雪面积比例(Fractional Snow Cover, FSC)数据能在亚像元尺度上定量的描述像元内积雪覆盖的程度,相比二值积雪面积数据可以更加精确地估计积雪覆盖的面积。基于机器学习的随机森林回归模型可以表示高维的非线性关系,可显著提高MODIS FSC的反演精度。采用随机森林回归模型结合光谱、环境信息构建了一个新的回归模型——光谱-环境随机森林回归(Spectral Environment Random Forest Regressor, SE-RFR)模型,用于MODIS数据反演中国区域的FSC。利用中国典型积雪区内由Landsat 8地表反射率数据获取的FSC数据作为参考值,对SE-RFR模型的反演精度进行评估。研究表明,利用"SE-RFR"获取的FSC数据RMSE、MAE分别为0. 160、0. 104,精度较高。此外,根据SE-RFR模型与未加入环境信息的随机森林回归(S-RFR)模型比较结果可知,加入环境信息的随机森林回归模型提高了FSC反演的精度,特别是在受环境信息影响较大的青藏高原地区,RMSE从0. 200降低到0. 181。最后,将SE-RFR模型与目前使用广泛的MODIS FSC反演模型FSC_NDSI、MODSCAG和SSEmod进行了比较,结果表明SE-RFR模型的RMSE与FSC_NDSI、MODSCAG和SSEmod模型的RMSE相比,平均RMSE分别提高了12. 0%、8. 3%和5. 5%。总体来说,SE-RFR模型可以准确地提取MODIS FSC,对于区域乃至全球FSC产品制备具有广泛的应用前景。

关键词: MODIS; 光谱信息; 环境信息; 积雪面积比例; FSC; 随机森林
中图分类号: P426. 63⁺5; TP75 **文献标志码:** A **文章编号:** 1000-0240(2022)01-0147-12

0 引言

积雪是冰冻圈重要的组成要素之一,是气候变化的指示器。积雪面积作为积雪的一个重要特征,对于区域水文、地表能量过程具有重要的意义[1-3]。遥感积雪面积数据由于覆盖面积大,时空分辨率较高,已被广泛应用在积雪面积的反演研究中[4]。遥感积雪面积数据主要分为二值积雪面积数据和积雪面积比例数据。二值积雪面积数据反演的精度与雪深、地形和地表类型密切相关,研究表明斑状分布的积雪、山区或林区分布的积雪,由于混合像元的影响,二值积雪面积数据很难反映积雪分布特征[5-7]。FSC数据用像元内积雪覆盖的比例来表示积雪覆盖的面积[8],可以在亚像元尺度上定量描述像元内积雪的覆盖程度,相比于二值积雪面积数据可以更加准确地估计积雪覆盖面积[9-10]。

MODIS FSC比例数据已经取代二值积雪面积数据作为许多水文和大气模型的重要输入参数[11-13]。目前,MODIS FSC的提取方法主要包括三种:线性回归模型、混合像元分解模型和机器学习

收稿日期: 2021-07-09; **修订日期:** 2021-10-08
基金项目: 国家重点研发计划项目(2019YFC1510503);国家自然科学基金项目(41971325;42171391);兰州交通大学优秀平台(201806)资助
作者简介: 孙兴亮,硕士研究生,主要从事积雪遥感研究. E-mail: 0219771@stu.lzjtu.edu.cn
通信作者: 郝晓华,研究员,主要从事积雪遥感、积雪与气候变化研究. E-mail: haoxh@lzb.ac.cn

模型。线性回归模型主要是利用FSC和与其相关的指数[如归一化植被指数(Normalized Difference Vegetation Index,NDVI)、归一化积雪指数(Normalized Difference Snow Index,NDSI)等]间的线性关系构建回归模型,许多学者都做了大量的研究[14-16]。代表性研究成果主要是Salomonson等[16]构建的线性回归模型(FSC_NDSI),该模型被NASA的MODIS全球积雪覆盖产品所采用。混合像元分解模型主要是选择图像端元,通过线性光谱混合分析模型(LSMA)进行解混以获取FSC。代表性研究主要包括:Painter等[17]结合LSMA模型和积雪辐射传输模型发展了一种针对MODIS数据的FSC提取的算法MODSCAG;施建成[18]发展了一种改进"多端元光谱混合分析"方法反演MODIS FSC,该算法通过对MOD09GA数据进行图像端元自动提取,并利用能够代表图像端元类的典型端元库进行"多端元光谱混合分析"反演FSC数据;Zhao等[19]考虑地表类型信息对FSC提取的影响,提出了一种基于空间光谱环境(SSE)信息的端元提取算法,并结合LSMA模型提取MODIS FSC的算法(SSEmod)。机器学习也是目前获取MODIS FSC的新方法,其中代表性研究包括:Dobreva等[20]首次利用人工神经网络(ANN)模型来反演MODIS FSC,取得了良好效果;Czyzowsk等[21]、Hou等[22-23]在此基础上考虑了地形、温度、海拔、地表覆盖类型等环境信息,有效地提高了MODIS FSC数据制备的精度。以上研究表明,机器学习方法能够有效地反演FSC,进一步结合环境信息,可以提高FSC的反演精度。

综合提取MODIS FSC的三种常用方法,线性回归模型物理意义明确,易于实现,但仅仅考虑NDSI与FSC之间的关系,忽略了地形、地表类型等环境信息对FSC提取的影响。混合像元分解模型通过考虑地表类型信息可以有效提高积雪识别精度,但在地形复杂、地表覆盖类型多样的地区仍然会高估或者低估积雪覆盖面积,需引入更多影响积雪识别的环境信息,使算法在估计积雪覆盖面积上有更好的精度[24]。相比线性回归模型和混合像元分解模型,机器学习模型结合环境信息(地形、地表覆盖类型)在高山区反演FSC具有更高的精度[22-23],但利用ANN模型处理高维数据的回归问题时收敛速度慢且易造成过拟合[25]。已有研究表明[26],相较于支持向量机(Support Vector Machine,SVM)和ANN模型,随机森林(Random Forest)在山区积雪面积提取中更加准确,具有良好的鲁棒性。在以往利用随机森林模型反演FSC的研究[26-27]中,特征数据的选择多集中于地表反射率、积雪指数、DEM等信息,忽略了地形、地表温度、地表覆盖类型等环境信息对FSC提取的影响。

因此,本研究利用随机森林回归(Random Forest Regressor)模型易于架构、抗噪性能强、防止过拟合的优点,引入了成像角度(观测角度)、地形、地表覆盖类型、地表温度、降雪等环境信息,构建了的光谱-环境随机森林回归模型(Spectral Environment Random Forest Regressor,SE-RFR)并用于中国区域FSC反演。并利用Landsat 8地表反射率数据生成的FSC对其进行了精度评估,分析了环境信息的引入对随机森林回归模型提取FSC的作用,并且与三种MODIS FSC反演算法(FSC_NDSI、MODSCAG、SSEmod)获取的FSC数据进行了对比,客观地描述SE-RFR模型的反演精度。

1 数据及预处理

本研究中主要使用MOD09GA地表反射率数据、MCD12Q1地表类型数据、ERA5-Land再分析数据、SRTM数字高程数据和Landsat 8地表反射率数据。MOD09GA、MCD12Q1、ERA5-LAND和SRTM数据主要用于提取随机森林回归模型的输入数据。Landsat 8地表反射率数据用于制备"真值"FSC,一部分用作模型的输入数据,另一部分作为验证数据,来对模型进行精度评估。以上输入数据在输入SE-RFR模型前需采用min-max标准化法进行归一化处理,以避免方差过大的特征对机器学习算法造成影响[28],所有输入数据需选取与Landsat 8数据时间、空间范围一致的影像数据,并采用与Landsat 8影像一致的投影系统将其重投影。

1.1 MOD09GA

MOD09GA逐日地表反射率数据源于NASA(https://search.earthdata.nasa.gov),空间分辨率为500 m,正弦投影且已经过大气校正。该数据是本研究的主要数据源,输入数据包括七个通道地表反射率数据(b01-b07),太阳天顶角、太阳方位角、传感器天顶角、传感器方位角四个角度数据和NDVI、NDSI、归一化林地积雪指数(Normalized Difference Forest Snow Index,NDFSI)三个指数数据。在提取输入数据前,需利用MOD09GA的质量评估QA提供的云掩膜信息来去除云像元,以免对模型训练造成影响。

1.2　MCD12Q1

MCD12Q1地表覆盖类型数据[29]来源于NASA，空间分辨率为500 m，正弦投影，可以提供逐年全球地表覆盖类型数据，数据覆盖时间自2001年至2019年，包含13个科学数据集，5个分类标准（IGBP，UMD，LAI，BGC，PFT）。本研究中使用了国际地圈-生物圈计划（IGBP）分类标准的地表类型数据，共包含17种地表类型，从1到9的IGBP代码被视为代表林冠高度超过2 m且树木覆盖率高于30%的森林区域，而其他IGBP代码被归类为非森林区域。该数据是随机森林回归模型输入数据中的重要环境信息，用于区分森林与非森林区域，同时也用来评估FSC数据在不同地表覆盖类型下的精度。

1.3　ERA5-Land与SRTM数字高程数据

ERA5-Land再分析数据源于哥白尼气候数据库（Copernicus Climate Data Store），时间分辨率为1 h，空间分辨率为0.1 rad，GLL经纬度投影，数据覆盖时间自1981年1月至2021年5月。本研究主要利用该数据集中的地表温度和降雪数据作为随机森林回归模型的输入数据。MODIS Terra在当地上午过境，为了将再分析资料与卫星观测数据相匹配，本研究中地表温度数据为当日12:00前的平均地表温度，降雪数据为当日12:00前的累积降雪。SRTM数字高程数据源于NASA，空间分辨率为90 m，WGS 84投影，主要用于提取高程数据，并基于高程数据采用4邻域法计算坡度、坡向。

1.4　Landsat 8地表反射率数据

Landsat 8地表反射率数据由美国地质调查局（United States Geological Survey，USGS）提供，已经过大气校正，空间分辨率为30 m，时间分辨率为16 d，WGS84 UTM投影。本数据主要用于制备Landsat 8 FSC数据（L8-FSC）。制备L8-FSC时先根据Wang等[30]开发改进的SNOMAP算法从Landsat 8地表反射率数据中提取积雪二值影像，改进的SNOMAP算法采用NDVI、NDSI和NDFSI相结合的方法来提取积雪像元。然后将30 m的积雪二值数据聚合成分辨率为500 m的FSC数据[7]。聚合公式由式（1）给出。

$$\mathrm{FSC}_i=\frac{s}{n}=\frac{\sum_{j=1}^{s}1}{[500/30]^2} \tag{1}$$

式中：[ ]表示取整；n表示一个500 m分辨率像元内30 m分辨率像元的个数；s表示一个500 m分辨率像元内30 m分辨率积雪像元的个数。

本研究在2014—2020年积雪期（本年11月1日至次年3月31日）期间共选取了中国区域内的32景Landsat 8地表反射率影像数据来制备L8-FSC。选取原则：影像数据无云（云覆盖率小于2%）且积雪覆盖率在30%~90%之间。其中，20景影像用于SE-RFR模型的训练，约有230多万个有效像元；12景影像用于验证SE-RFR模型的准确性，约有130多万个有效像元。其中训练样本与验证样本相互独立，训练样本及验证样本主要选自东北-内蒙古、北疆、青藏高原三大积雪区，积雪区及样本的空间分布如图1所示。

2　研究方法

2.1　光谱-环境随机森林回归模型的构建

2.1.1　随机森林回归模型

随机森林回归模型[31]是一种基于回归决策树的集成学习模型，取各决策树$\{h(x,\theta_t)\}$的均值回归预测的结果：

$$\bar{h}(x)=\frac{1}{T}\sum_{t=1}^{T}\left\{h\left(x,\theta_t\right)\right\} \tag{2}$$

式中：x为自变量；θ_t为服从独立同分布的随机变量；T为决策树数量；$h(x,\theta_t)$为基于x和θ_t的输出。此外，随机森林回归算法引入了Bagging思想[32]，随机独立地抽取子样本集、独立地构建决策树进行计算，并且在构建决策树时，每个节点随机选取特征子集，从中选取最优特征进行分裂。这使得模型拥有更好的预测能力，对噪声、异常值有很好的容忍度，并在一定程度上避免过拟合。

2.1.2　构建光谱-环境随机森林回归模型

考虑环境信息对提取FSC数据的影响，本研究结合光谱信息（地表反射率、NDVI、NDSI、NDFSI）和环境信息（成像角度、地形、地表类型、地表温度及降雪）构建了SE-RFR模型。光谱-环境信息作为特征数据，详细信息如表1所示，L8-FSC作为“真值”数据，两者输入到随机森林回归模型中进行训练，进而优化参数获取性能较好的光谱-环境随机森林回归（SE-RFR）模型。

随机森林回归模型有放回的抽取样本数据（袋内样本）用于决策树的训练，其余数据（袋外样本，OOB）便可作为测试集数据与真值计算得到泛化分数（1和泛化误差的差），用于估计模型的精度，避免使用交叉验证等方法来评价模型精度，大大节省了模型训练花费的时间。影响随机森林回归模型精度的参数主要有两个：决策树数目（n_trees）和树的

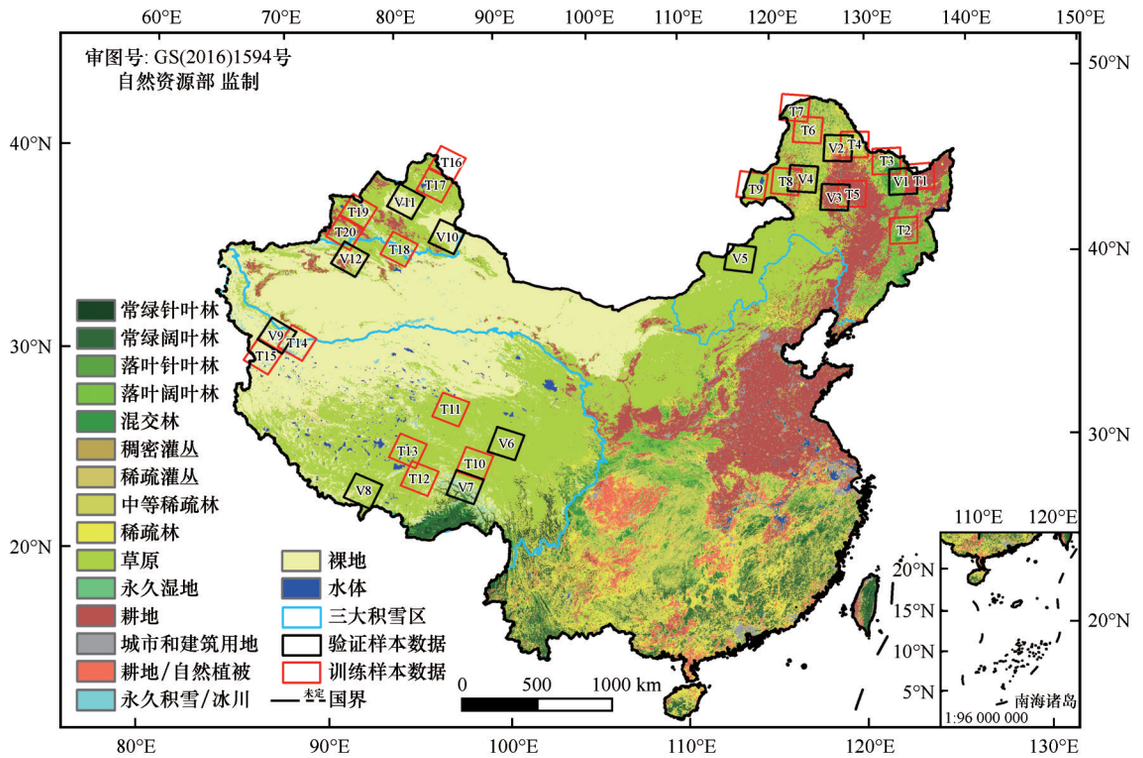

图 1 研究区概况及样本数据的空间分布

Fig. 1 The location of the study and the spatial distribution of samples data

表 1 特征数据的详细信息

Table 1 The detailed information of feature data

光谱信息		环境信息	
地表反射率数据	b01（620~670 nm）	角度数据	传感器天顶角
	b02（841~876 nm）		传感器方位角
	b03（459~479 nm）		太阳天顶角
	b04（545~565 nm）		太阳方位角
	b05（1 230~1 250 nm）	地形	DEM 高程
	b06（1 628~1 652 nm）		坡度
	b07（2 105~2 155 nm）		坡向
指数数据	NDVI	其他	地表覆盖类型
	NDSI		地表温度
	NDFSI		降雪

最大深度（max_depth），即决策树的最大节点数。因此采用 OOB 泛化分数（OOB_Score）为指标，选择最优的参数组合 n_trees 和 max_depth，SE-RFR 模型的实现及 FSC 的反演流程如图 2 所示。

训练过程主要分为两步，首先根据 OOB_Score 选取合适的参数 max_depth，再根据选好的参数 max_depth 选择合适的参数 n_trees。图 3（a）、3（b）分别为参数 max_depth、n_trees 的训练过程，随着树的增多，模型精度的增益会很小[33]，因此本研究中 SE-RFR 的 n_trees 和 max_depth 被设置为 1 500、40。

2.2 其他 MODIS FSC 反演算法

本研究中为了客观评价 SE-RFR 模型的精度，本文将其与三种常用的 MODIS FSC 反演算法（FSC_NDSI、MODSCAG、SSEmod）进行比较，这三种反演算法的模型介绍如下。

FSC_NDSI 线性回归模型，由 Salomonson 等[8]利用归一化积雪指数（NDSI）与 FSC 之间的线性关系构建的简单线性回归模型，该算法被 NASA 的 MODIS 全球积雪覆盖产品（MOD10A1）所采用，计算简单，但具有较大的不确定性，其计算公式如式（3）所示。

$$FSC=1.45NDSI-0.01 \tag{3}$$

MODSCAG 模型[17]是根据野外和实验室采集光谱获取非积雪端元光谱库，主要非积雪端元包含植被、岩石和土壤端元，对于积雪端元，通过辐射传输模型模拟不同粒径的积雪光谱建立光谱库。本研究通过渐进辐射传输模型（Asymptotic Radiative Transfer，ART）模拟了不同粒径的积雪光谱[34-35]，通过多端元线性光谱混合分析模型，根据误差最小迭代原则计算获取了最优的 FSC。该算法物理机制明确，但未考虑非积雪端元随影像动态变化，并且模型模拟的积雪光谱与实际积雪光谱存在差异。

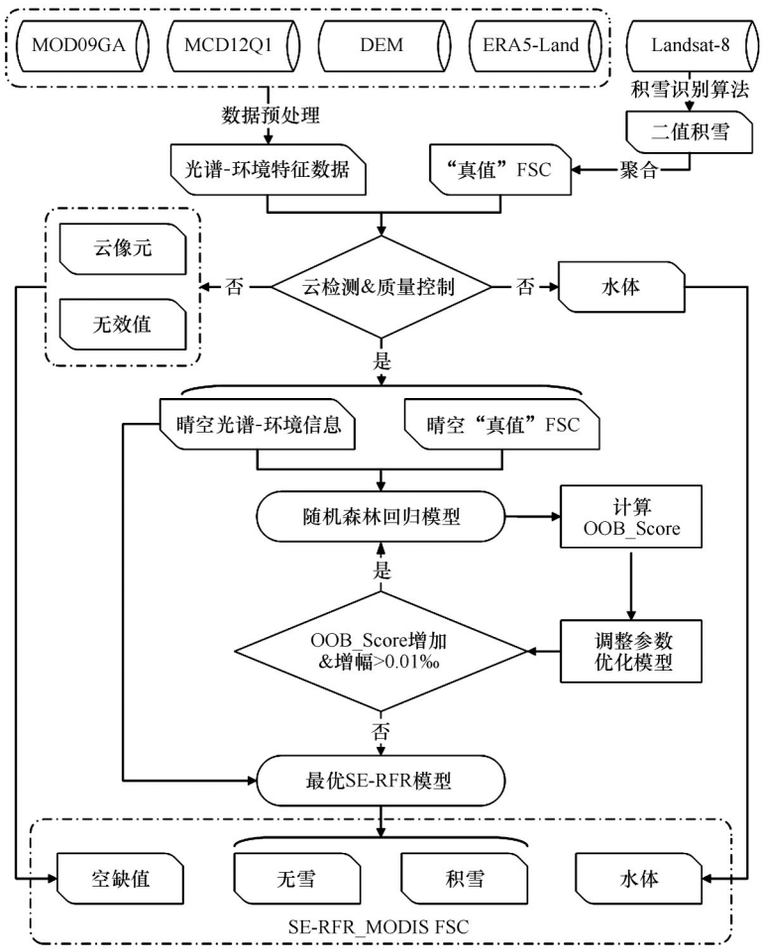

图2　SE-RFR模型的实现及FSC的反演流程

Fig. 2　Processing flowchart of SE-RFR model

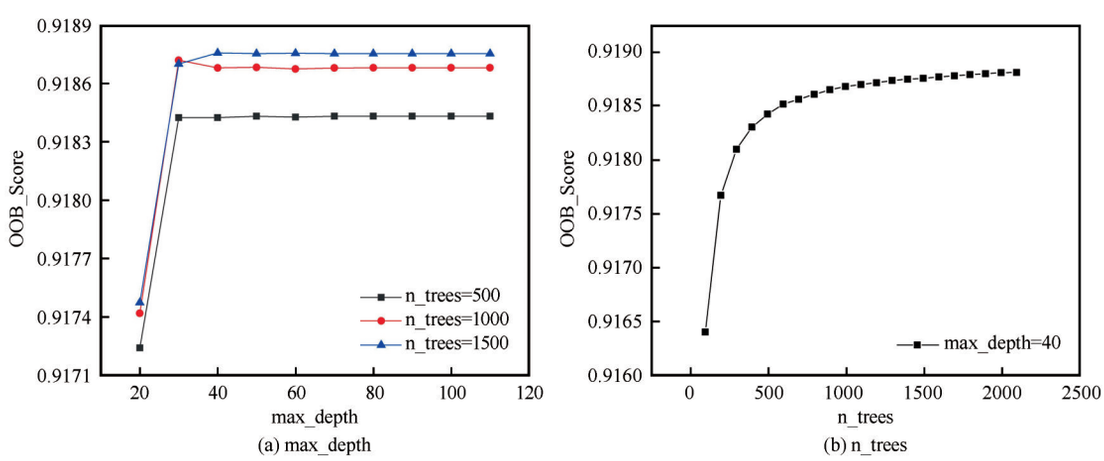

图3　OOB_Score值随参数max_depth、n_trees的变化情况

Fig. 3　The change of OOB_Score value with the change of parameters max_depth and n_trees

SSEmod模型[19]是考虑地表类型信息对FSC提取的影响，提出了一种基于空间光谱环境（SSE）信息的动态积雪和非积雪端元自动提取算法，并结合线性光谱混合分析模型来提取MODIS FSC。该模型主要特点是引入地表类型信息来初步估计端元的数量，减少候选端元的谱冗余。此外，在林区和非林区提取了不同数量和类型的积雪端元，通过动态阈值分割方法选择其他端元，并根据候选端元像素的光谱差异来调整最终的端元，从算法原理上具有较高的精度，主要受限于MODIS地表反射率产

品(MOD09GA)波段的数量,导致该算法在复杂地表类型条件下精度较低。

2.3 精度评估方法

12景Landsat 8地表反射率数据生成的L8-FSC作为真值来验证SE-RFR模型反演FSC的精度。采用均方根误差(Root Mean Square Error,RMSE)和平均绝对误差(Mean Absolute Error,MAE)作为模型精度的评价因子。RMSE、MAE可根据式(4)、(5)计算

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-y_i\right)^2} \tag{4}$$

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|x_i-y_i\right| \tag{5}$$

式中:x_i和y_i分别为FSC数据像元的真值和反演值;n为数据的样本个数。

3 结果与讨论

3.1 光谱-环境随机森林模型的精度评估

本文利用中国12景L8-FSC数据作为真值对SE-RFR模型进行精度评估,精度验证结果如表2所示。结果表明12景验证数据总体上RMSE和MEA分别为0.160、0.104,产品的精度较高。由于积雪特征存在显著的空间差异,本研究验证了模型在不同积雪区的精度,可以看到,模型在北疆积雪区RMSE和MAE分别为0.110、0.058,在东北-内蒙古积雪区RMSE和MEA分别为0.169、0.113,在青藏高原积雪区RMSE和MAE分别为0.181、0.129。仅看RMSE指标,北疆雪区精度最高,东北-内蒙古雪区次之,青藏高原雪区较差。模型精度的差异是由积雪区不同的积雪特征引起的,北疆雪区由于地势平坦,积雪大范围分布,混合像元相对较少;东北雪区由于森林分布广泛,林区内混合像元较多,导致精度略低;青藏高原雪区降雪较少且地形复杂,积雪多呈现斑状块分布,混合像元较多,因而相对来说精度最低。

表2 中国三大积雪区内SE-RFR模型的平均精度验证结果

Table 2 The average accuracy validation results of the SE-RFR model in three snow-covered regions of China

	RMSE	MAE
总体	0.160	0.104
北疆	0.110	0.058
东北-内蒙古	0.169	0.113
青藏高原	0.181	0.129

为了验证SE-RFR模型在不同地表覆盖类型条件下的反演精度,按1.2节中的地表覆盖类型数据将12景验证影像分为林区与非林区像元,对其进行精度评估。精度验证结果如表3所示,非林区的RMSE和MEA分别为0.139、0.085;林区的RMSE和MEA分别为0.235、0.192。SE-RFR模型在林区和非林区精度都较高,但非林区具有更高的精度。

表3 林区与非林区SE-RFR模型的平均精度验证结果

Table 3 The average accuracy validation results of the SE-RFR model in forest areas and non-forest areas

区域	RMSE	MAE
林区	0.235	0.192
非林区	0.139	0.085

为研究SE-RFR模型对FSC低值区、中值区、高值区的反演精度,本研究将FSC根据数值大小分为三级,第一级为(0.15,0.50],表示低值区;第二级为(0.50,0.80],表示中值区;第三级为(0.80,1.00],表示高值区。对于FSC值小于0.15的区间,由于数值太低,存在较大的不确定性,不参与精度评估。对SE-RFR模型反演的FSC进行验证,验证结果如表4所示。低值区RMSE和MAE分别为0.222、0.177,中值区RMSE和MAE分别为0.183、0.146,高值区RMSE和MAE分别为0.122、0.071,高值区精度最高,低值区最低。表明该模型对于中、高值区反演校准,而低值区精度略低,因此该模型具有较高的可靠性。图4进一步展示了不同分级FSC的反演值和真值的空间密度分布图,可以看到中高值区间内沿对角线分布的六边形颜色呈红色,表明像元分布较多,反演精度较高,特别是高值区内大部分呈红色,说明反演值与真值基本一致,表明了算法的稳定性和可靠性。

表4 各区间SE-RFR模型反演FSC的精度验证结果

Table 4 The average accuracy validation results of SE-RFR FSC in different sections

FSC分级	FSC数值区间	RMSE	MAE
低值区	(0.15,0.50]	0.222	0.177
中值区	(0.50,0.80]	0.183	0.146
高值区	(0.80,1.00]	0.122	0.071

3.2 光谱-环境随机森林回归模型对环境信息的依赖性

为了评估环境信息对于随机森林回归模型的

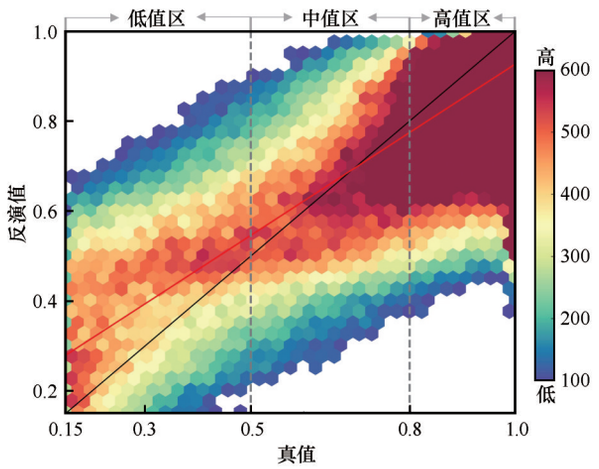

图4　不同分级FSC反演值和真值的六边形分箱图

Fig. 4　The spatial density distribution map of the inversion value and true value of different grades of FSC

表5　中国不同积雪区S-RFR、SE-RFR模型的平均精度验证结果

Table 5　The average accuracy validation results of the S-RFR and SE-RFR model in three snow-covered regions of China

积雪区	模型	RMSE	MAE
总体	S-RFR	0. 171	0. 107
	SE-RFR	0. 160	0. 104
北疆	S-RFR	0. 125	0. 069
	SE-RFR	0. 110	0. 058
东北-内蒙古	S-RFR	0. 172	0. 112
	SE-RFR	0. 169	0. 112
青藏高原	S-RFR	0. 200	0. 131
	SE-RFR	0. 181	0. 129

重要性，本文分别对引入环境信息（成像角度、地形、地表温度、地表覆盖类型、降雪等）前后的随机森林回归模型进行比较分析（引入环境信息前的随机森林回归模型本文简称为S-RFR）。本研究同样用12景L8-FSC验证数据对S-RFR和SE-RFR模型进行精度评估，表5展示了精度验证结果。S-RFR和SE-RFR模型RMSE分别为0. 171、0. 160，MAE分别为0. 107、0. 104，加入环境信息后，RMSE降低了0. 011，MAE降低了0. 003。北疆与东北-内蒙古积雪区精度提高较少，RMSE分别从0. 125、0. 172降低到0. 110、0. 169，青藏高原积雪区精度提高较大，RMSE从0. 200降低到0. 181，降低了0. 019。结果表明地形、地表温度、地表覆盖类型等环境信息的引入，可以有效提高随机森林回归模型对青藏高原山区斑状积雪的识别精度。图5进一步展示了青藏高原山区斑状积雪的反演结果，可以明显看出S-RFR模型反演的FSC对斑状积雪高估，尤其在地形

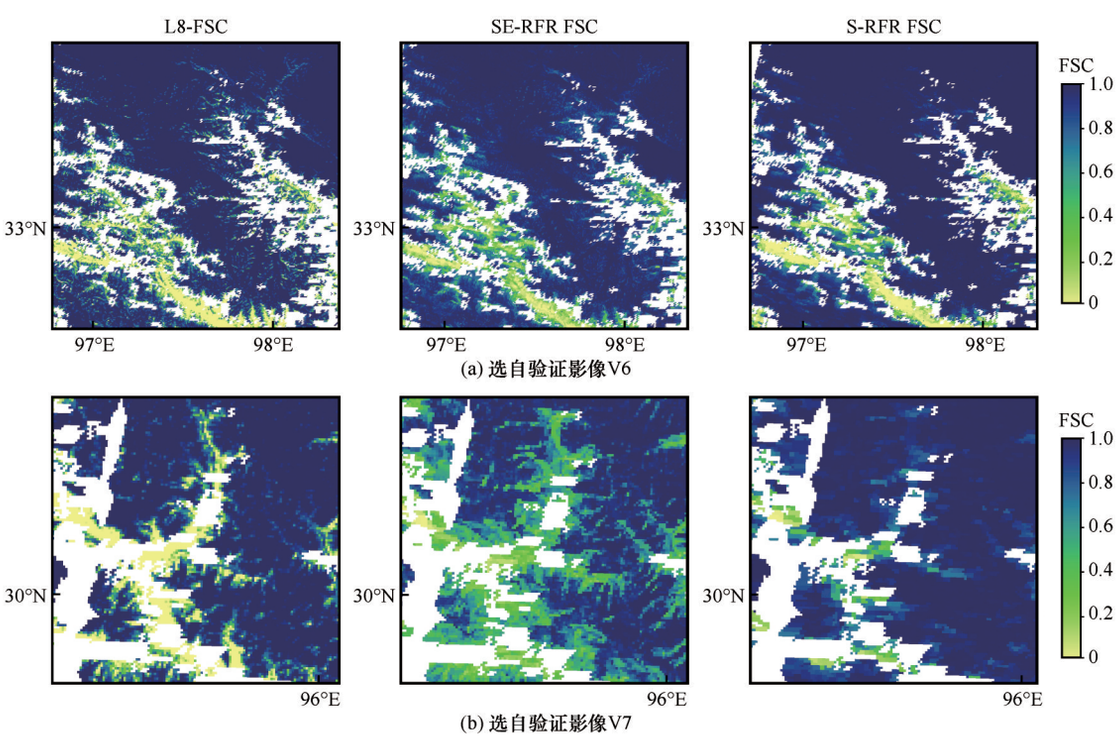

图5　L8-FSC、SE-RFR FSC、S-RFR FSC斑状积雪的反演结果

Fig. 5　The inversion result of patchy snow cover in L8-FSC, SE-RFR FSC, S-RFR FSC

[(a), (b) respectively selected from the verification images V6 and V7]

起伏变化较大的山区，SE-RFR模型反演的FSC与真值更为接近，说明引入了环境信息的SE-RFR模型有效地提高了青藏高原山区斑状积雪的识别精度。

3.3 与其他MODIS FSC反演算法的比较

为了客观评价SE-RFR模型的精度，我们又将SE-RFR模型与线性回归模型FSC_NDSI，混合像元分解模型MDOSCAG、SSEmod进行比较分析。在此使用相同的12景L8-FSC验证数据对各模型进行精度验证。

图6展示了由FSC_NDSI、MODSCAG、SSEmod和SE-RFR模型反演的12景验证数据的平均精度（RMSE），可以明显看出与其他模型相比SE-RFR模型的反演精度最高，且具有较好的准确性与稳定性。表6进一步统计了各模型的平均RMSE和平均MAE，可以看到FSC_NDSI、MODSCAG、SSEmod和SE-RFR模型的平均RMSE分别为0.280、0.243、0.215和0.160，平均MAE分别为0.208、0.136、0.117和0.104。结果表明，相较于FSC_NDSI、MODSCAG和SSEmod模型，SE-RFR模型的平均RMSE提高了12.0%、8.3%和5.5%，平均MAE分别提高了10.4%、3.2%和1.3%。总体来说，SE-RFR模型的精度最高，SSEmod模型次之，其次是MODSCAG模型，FSC_NDSI模型精度最差。图7展示了使用SE-RFR、SSEmod、MODSCAG和FSC_NDSI模型在三大积雪区获取的FSC影像，可以明显看出SE-RFR模型反演的FSC更接近于真值。结果表明，在提取MODIS FSC时，基于物理机制的混合像元分解模型要优于基于统计关系的FSC_NDSI模型；考虑动态光谱库的混合像元分解模型SSEmod要优于端元固定的混合像元分解模型MODSCAG；在目前MODIS地表反射率产品（MOD09GA）仅有7个波段的条件限制下，考虑物理过程约束（光谱、环境信息）的SE-RFR模型具有更高的FSC提取精度。

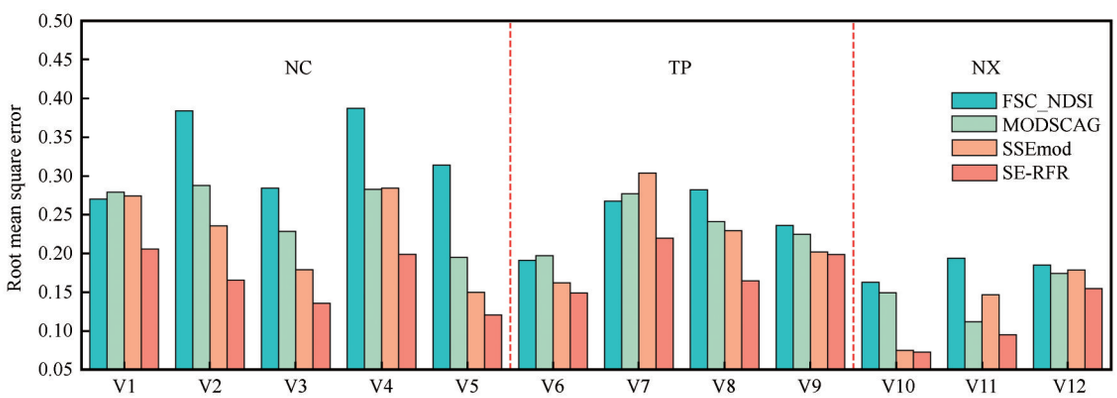

图6 FSC_NDSI、MODSCAG、SSEmod和SE-RFR FSC的精度验证结果（NC表示东北地区-内蒙古积雪区，TP表示青藏高原积雪区，NX表示北疆积雪区）

Fig. 6 The accuracy validation result of FSC_NDSI, MODSCAG, SSEmod, and SE-RFR FSC (NC, TP, NX respectively represent the Northeast China-Inner Mongolia snow area, the Qinghai-Tibet Plateau snow area, and the northern Xinjiang snow area)

表6 中国区域不同FSC反演算法的平均精度验证结果

Table 6 The average accuracy validation results of different methods for FSC retrieval in China

方法	RMSE	MAE
FSC_NDSI	0.280	0.208
MODSCAG	0.243	0.136
SSEmod	0.215	0.117
SE-RFR	0.160	0.104

SE-RFR模型充分考虑了光谱和环境信息，并且训练样本具有很好的代表性，随机森林回归算法随机独立地选取特征子集构建决策树，可以充分利用最优的特征数据进行FSC反演，模型拥有更好的鲁棒性，并在一定程度上避免过拟合。混合像元分解模型的精度很大程度上依赖于端元的选择，通过改进端元提取的方法可以提高FSC估计的精度[10]。SSEmod模型针对每一幅影像通过动态阈值分割法自动地提取端元，通过线性光谱混合分析模型获取FSC，受制于MOD09GA影像的端元数量不足，导致算法的精度不高；MODSCAG模型虽然考虑了不同粒径的积雪端元，但其非雪端元是固定的，除受制于MOD09GA影像的端元数量不足外，对于不同区域的影像，其端元存在着不确定性和不一致性，导致算法的精度不高。FSC_NDSI算法仅仅利用了NDSI与FSC之间的统计关系构建了经验模型，普

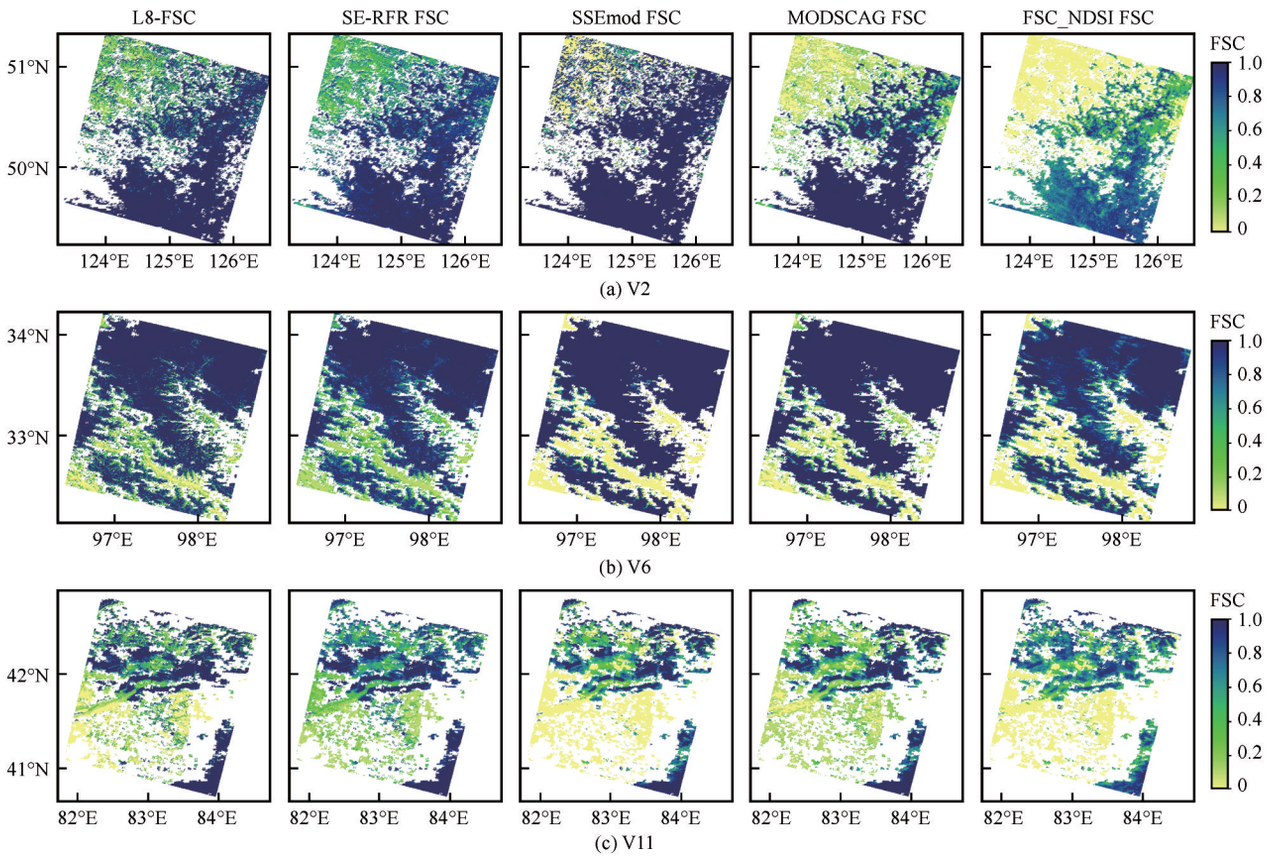

图7　L8-FSC、SE-RFR FSC、SSEmod FSC、MODSCAG FSC和FSC_NDSI FSC在中国不同积雪区的结果

Fig. 7　The result of L8-FSC, SE-RFR FSC, SSEmod FSC, MODSCAG FSC and FSC_NDSI FSC in different snow cover regions of China

适性强，但精度较低。

3.4　光谱-环境随机森林模型的不确定性及展望

地形、地表温度、地表类型等环境信息是影响积雪检测的重要因素[36]。在东北-内蒙古雪区森林资源丰富，尽管与线性回归模型相比，混合像元分解模型对林区积雪提取有了一定改进，但也会低估FSC，主要原因如下，由于林区树冠遮挡造成阴影，致使产生一系列的暗像元，削减卫星接收的辐射能量，而这些暗像元一般为雪。而SE-RFR模型在引入地形、地表类型、地表温度等通用的环境信息外，又引入成像角度、降雪信息来反演FSC，提高了精度。当然这种问题在引入环境信息后不可能完全解决，故SE-RFR FSC也存在一些高估或低估现象。同样，青藏高原受地形影响严重，山区阴影也对积雪提取造成影响，降低了混合像元分解模型、线性回归模型的精度。北疆雪区地势较为平坦，地表多为裸土、草原，各模型对其区域内积雪低估程度较小。

相较于混合像元分解模型，利用随机森林模型结合环境信息反演FSC，使得模型易于构建。本研究中SE-RFR模型共输入了20种特征数据，包括光谱信息与环境信息，其中三种指数数据是由地表反射率波段信息计算而来，这造成了一定的冗余信息。在后续研究工作中，需要进一步提高模型的计算效率，使其适于制备产品。

4　结论

本研究利用MODIS数据，构建了一个考虑光谱信息、环境信息的光谱-环境随机森林回归模型（SE-RFR）来反演中国区域的FSC。利用中国典型积雪区的Landsat 8 FSC数据作为参考值验证了SE-RFR模型的反演精度，评估了SE-RFR模型对环境信息的依赖性，同时与FSC_NDSI、MODSCAG和SSEmod等国内外常用的MODIS FSC反演模型进行了比较，得到以下结论：

（1）利用SE-RFR模型反演的MODIS FSC在中国区域精度较高，平均RMSE、MAE分别为0.160、0.104。北疆积雪区精度最高，RMSE为0.110；东北-内蒙古积雪区次之，RMSE为0.172；青藏高原

积雪区较差,RMSE为0.181。

(2)对引入环境信息前后的随机森林回归模型获取的MODIS FSC进行了对比,发现成像角度、地形、地表类型、地表温度、降雪等环境信息的引入可以在一定程度上提高FSC的反演精度。特别是在积雪受地形影响较大的青藏高原地区,RMSE从0.200降低到0.181,提高了1.9%,有效解决了斑状积雪的高估问题。

(3)将SE-RFR模型与线性回归模型(FSC_NDSI)、混合像元分解模型(MODSCAG、SSEmod)进行了对比,表明SE-RFR模型的精度最高。对于所有积雪区的平均RMSE,SE-RFR模型为0.160,与FSC_NDSI、MODSCAG和SSEmod模型的平均RMSE(0.280、0.243、0.215)相比,分别提高了12.0%、8.3%、5.5%。

总体而言,SE-RFR模型算法可以更准确地反演MODIS FSC,并且模型结构简单易于构建,鲁棒性强,对于区域乃至全球MODIS FSC产品制备具有广泛的应用前景,从而为区域水文、气候模型提供更准确的输入数据。

参考文献(References):

[1] Zhang Tingjun. Influence of the seasonal snow cover on the ground thermal regime: an overview[J]. Reviews of Geophysics, 2005, 43(4): RG4002.

[2] Sturm M, Holmgren J, McFadden J P, et al. Snow-shrub interactions in Arctic tundra: a hypothesis with climatic implications [J]. Journal of Climate, 2001, 14(3): 336-344.

[3] Robinson D A, Dewey K F, Heim R R. Global snow cover monitoring: an update[J]. Bulletin of the American Meteorological Society, 1993, 74(9): 1689-1696.

[4] König M, Winther J G, Isaksson E. Measuring snow and glacier ice properties from satellite[J]. Reviews of Geophysics, 2001, 39(1): 1-27.

[5] Dozier J, Painter T H. Multispectral and hyperspectral remote sensing of alpine snow properties[J]. Annual Review of Earth and Planetary Sciences, 2004, 32: 465-494.

[6] Hall D K, Benson C S, Field W O. Changes of glaciers in glacier bay, Alaska, using ground and satellite measurements[J]. Physical Geography, 1995, 16(1): 27-41.

[7] Zhao Hongyu, Hao Xiaohua, Zheng Zhaojun, et al. A new algorithm of fractional snow cover basing on FY-3D/MERSI-Ⅱ [J]. Remote Sensing Technology and Application, 2018, 33(6): 1004-1016. [赵宏宇, 郝晓华, 郑照军, 等. 基于FY-3D/MERSI-Ⅱ的积雪面积比例提取算法[J]. 遥感技术与应用, 2018, 33(6): 1004-1016.]

[8] Salomonson V V, Appel I. Development of the Aqua MODIS NDSI fractional snow cover algorithm and validation results [J]. IEEE Transactions on Geoscience and Remote Sensing, 2006, 44(7): 1747-1756.

[9] Hao Shirui, Jiang Lingmei, Wang Gongxue, et al. The effect of scale and snow fragmentation on the accuracy of fractional snow cover data over the Tibetan Plateau[C]//2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). Fort Worth, TX, USA. Piscataway, NJ: IEEE, 2017: 4250-4253.

[10] Lei Huajin, Li Hongyi, Wang Jian, et al. MODIS fractional snow cover products preparing on Tibetan Plateau based on environmental information and regression model [J]. Remote Sensing Technology and Application, 2020, 35(6): 1303-1311. [雷华锦, 李弘毅, 王建, 等. 基于环境信息和回归模型的青藏高原MODIS积雪面积比例产品制备[J]. 遥感技术与应用, 2020, 35(6): 1303-1311.]

[11] Liston G E. Interrelationships among snow distribution, snowmelt, and snow cover depletion: Implications for atmospheric, hydrologic, and ecologic modeling[J]. Journal of Applied Meteorology, 1999, 38(10): 1474-1487.

[12] Carey C J, Hart S C, Aciego S M, et al. Microbial community structure of subalpine snow in the sierra Nevada, California [J]. Arctic, Antarctic, and Alpine Research, 2016, 48(4): 685-701.

[13] Hao Xiaohua, Wang Jie, Wang Jian, et al. Observations of snow mixed pixel spectral characteristics using a ground-based spectral radiometer and comparing with unmixing algorithms [J]. Spectroscopy and Spectral Analysis, 2012, 32(10): 2753-2758. [郝晓华, 王杰, 王建, 等. 积雪混合像元光谱特征观测及解混方法比较[J]. 光谱学与光谱分析, 2012, 32(10): 2753-2758.]

[14] Barton J S, Hall D K, Riggs G A. Remote sensing of fractional snow cover using Moderate Resolution Imaging Spectroradiometer (MODIS) data[C]// Proceedings of the 57th Eastern Snow Conference. 2000: 171-183.

[15] Kaufman Y J, Kleidman R G, Hall D K, et al. Remote sensing of subpixel snow cover using 0.66 and 2.1 μm channels[J]. Geophysical Research Letters, 2002, 29(16): 28-1.

[16] Salomonson V V, Appel I. Estimating fractional snow cover from MODIS using the normalized difference snow index[J]. Remote Sensing of Environment, 2004, 89(3): 351-360.

[17] Painter T H, Rittger K, McKenzie C, et al. Retrieval of subpixel snow covered area, grain size, and albedo from MODIS[J]. Remote Sensing of Environment, 2009, 113(4): 868-879.

[18] Shi Jiancheng. An automatic algorithm on estimating sub-pixel snow cover from modis[J]. Quaternary Sciences, 2012, 32(1): 6-15. [施建成. MODIS亚像元积雪覆盖反演算法研究——纪念杰出的地理学家、冰川学家施雅风先生逝世一周年[J]. 第四纪研究, 2012, 32(1): 6-15.]

[19] Zhao Hongyu, Hao Xiaohua, Wang Jian, et al. The spatial-spectral-environmental extraction endmember algorithm and application in the MODIS fractional snow cover retrieval[J]. Remote Sensing, 2020, 12(22): 3693.

[20] Dobreva I D, Klein A G. Fractional snow cover mapping through artificial neural network analysis of MODIS surface reflectance [J]. Remote Sensing of Environment, 2011, 115(12): 3355-3366.

[21] Czyzowska-Wisniewski E H, van Leeuwen W J D, Hirschboeck K K, et al. Fractional snow cover estimation in complex alpine-forested environments using an artificial neural network[J]. Remote Sensing of Environment, 2015, 156: 403-417.

[22] Hou Jinliang, Huang Chunlin. An application of ANN for mountainous snow cover fraction mapping with MODIS and ancillary topographic data[C]//2013 IEEE International Geoscience and Remote Sensing Symposium - IGARSS. Melbourne, VIC, Australia. Piscataway, NJ: IEEE, 2013: 1186-1189.

[23] Hou Jinliang, Huang Chunlin. Improving mountainous snow cover fraction mapping via artificial neural networks combined with MODIS and ancillary topographic data[J]. IEEE Transactions on Geoscience and Remote Sensing, 2014, 52(9): 5601-5611.

[24] Liang Hui, Huang Xiaodong, Sun Yanhua, et al. Fractional snow-cover mapping based on MODIS and UAV data over the Tibetan Plateau[J]. Remote Sensing, 2017, 9(12): 1332.

[25] Du Xu, Feng Jingyu, Lü Shaoqing, et al. $PM_{2.5}$ concentration prediction model based on random forest regression analysis [J]. Telecommunications Science, 2017, 33(7): 66-75. [杜续, 冯景瑜, 吕少卿, 等. 基于随机森林回归分析的$PM_{2.5}$浓度预测模型[J]. 电信科学, 2017, 33(7): 66-75.]

[26] Liu Changyu, Huang Xiaodong, Li Xubing, et al. MODIS fractional snow cover mapping using machine learning technology in a mountainous area [J]. Remote Sensing, 2020, 12(6): 962.

[27] Liang Hui. Fractional snow-cover mapping based on MODIS data over the Tibetan Plateau[D]. Lanzhou: Lanzhou University, 2019. [梁慧. 青藏高原MODIS积雪面积比例制图算法研究[D]. 兰州: 兰州大学, 2019.]

[28] Zhao Hongyu. Long time series of cloud-free fractional snow cover products in China[D]. Beijing: University of Chinese Academy of Sciences, 2020. [赵宏宇. 中国区域长时间序列积雪面积比例产品的制备[D]. 北京: 中国科学院大学, 2020.]

[29] Sulla-Menashe D, Friedl M A. User guide to collection 6 MODIS land cover (MCD12Q1 and MCD12C1) product [J]. USGS: Reston, VA, USA, 2018: 1-18.

[30] Wang Xiaoyan, Chen Siyong, Wang Jian. An adaptive snow identification algorithm in the forests of northeast China [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 13: 5211-5222.

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

[32] Breiman L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123-140.

[33] Tyralis H, Papacharalampous G, Tantanee S. How to explain and predict the shape parameter of the generalized extreme value distribution of streamflow extremes using a big dataset[J]. Journal of Hydrology, 2019, 574: 628-645.

[34] Kokhanovsky A A, Zege E P. Scattering optics of snow[J]. Applied Optics, 2004, 43(7): 1589.

[35] Hao Xiaohua, Wang Jie, Wang Jian, et al. The measurement and retrieval of the spectral reflectance of different snow grain size on northern Xinjiang, China[J]. Spectroscopy and Spectral Analysis, 2013, 33(1): 190-195. [郝晓华, 王杰, 王建, 等. 北疆地区不同雪粒径光谱特征观测及反演研究[J]. 光谱学与光谱分析, 2013, 33(1): 190-195.]

[36] Hall D K, Kelly R E J, Riggs G A, et al. Assessment of the relative accuracy of hemispheric-scale snow-cover maps [J]. Annals of Glaciology, 2002, 34: 24-30.

Research on retrieval of MODIS fraction snow cover based on spectral environmental random forest regression model

SUN Xingliang[1,2,3,4], HAO Xiaohua[2], WANG Jian[2], ZHAO Hongyu[5], JI Wenzhen[2]

(*1. Faculty of Geomatics, Lanzhou Jiaotong University, Lanzhou 730070, China; 2. Northwest Institute of Ecology and Environmental Resources, Chinese Academy of Sciences, Lanzhou 730000, China; 3. National-Local Joint Engineering Research Center of Technologies and Applications for National Geographic State Monitoring, Lanzhou 730070, China; 4. Gansu Provincial Engineering Laboratory for National Geographic State Monitoring, Lanzhou 730070, China; 5. State Key Laboratory of Earth Surface Processes and Resource Ecology, Beijing Normal University, Beijing 100875, China*)

Abstract: The fractional snow cover (FSC) data can quantitatively describe the extent of snow cover in a pixel on the sub-pixel scale, and can estimate the area of snow cover more accurately than binary snow area data. The random forest regression model based on machine learning can represent high-dimensional nonlinear relationships, which can significantly improve the inversion accuracy of MODIS FSC. In this study, a new regression model, Spectral Environment Random Forest Regressor (SE-RFR) model, was constructed using random forest regression model combined with spectral and environmental information, which was used to retrieve the FSC from MODIS data in China. We used the FSC obtained from Landsat 8 surface reflectance data in a typical snow area in China as a reference value to evaluate the inversion accuracy of the SE-RFR model. Research shows that the RMSE and MAE of FSC data obtained by SE-REF are 0.160 and 0.104, respectively, which has high accuracy. The SE-RFR model is compared with the Spectral Random Forest Regressor (S-RFR) without environmental information. It shows that the random forest regression model with environmental information improves the accuracy of FSC inversion, especially in the Qinghai-Tibet Plateau region, which is influenced by environmental information, and the RMSE decreased from 0.200 to 0.181. Finally, the SE-RFR model was compared with the currently widely used MODIS FSC inversion models FSC_NDSI, MODSCAG and SSEmod. The results showed that the average RMSE of the SE-RFR model is increased by 12.0%, 8.3% and 5.5%, respectively, compared with the RMSE of the FSC_NDSI, MODSCAG and SSEmod models. In general, the SE-RFR model can accurately extract MODIS FSC, which has wide application prospects for the preparation of regional and even global FSC products.

Key words: MODIS; fractional snow cover; spectrum information; environmental information; random forest

(责任编委：戴礼云)