引 用 格 式：Shen Jie，Xin Xiaoping，Zhang Jing，*et al*.Reconstruction of SIF Remote Sensing Data of Vegetation in China based on Cubist[J].Remote Sensing Technology and Application，2022，37(1)：244-252.[沈洁，辛晓平，张景，等.基于Cubist的中国植被区域叶绿素荧光数据重建[J].遥感技术与应用，2022，37(1)：244-252.]
DOI：10.11873/j.issn.1004-0323.2022.1.0244

基于Cubist的中国植被区域叶绿素荧光数据重建

沈 洁[1]，辛晓平[1]，张 景[2]，苗 晨[2]，王 旭[1]，丁 蕾[1]，沈贝贝[1]
（1.中国农业科学院农业资源与农业区划研究所，北京 100081；
2.国家遥感中心，北京 100036）

摘要：日光诱导叶绿素荧光（Solar-Induced chlorophyll Fluorescence，SIF）是植物在太阳光照条件下，在光合作用过程中发射出的光谱信号（650～800 nm），SIF相比于植被指数等参数更能直接地反映植被光合作用的相关信息，为大尺度GPP估算带来了新的途径。但目前卫星SIF数据或存在分辨率较低的不足，或存在数据空间不连续的局限，对于应用到大尺度中连续GPP的估算中有一定难度。OCO-2 SIF数据拥有较高的空间分辨率，但却是空间离散数据。针对上述问题，着重研究对离散的OCO-2 SIF数据进行连续预测的方法，生成中国—蒙古草地生态系统的较高精度连续SIF数据集。结果如下：通过Cubist回归树算法，结合MODIS反射率数据，气象数据及土地利用类型，建立了每8 d的0.05°分辨率的连续SIF数据集，预测精度为R^2=0.65，RMSE=0.114。其中，对作物类SIF预测的精度最高，为R^2=0.71，RMSE=0.117；其次为对森林与草地的预测，两者的R^2和RMSE分别为0.64/0.123，0.60/0.112。

关 键 词：日光诱导叶绿素荧光；Cubist模型；数据重建
中图分类号：Q948；TP392 **文献标志码**：A **文章编号**：1004-0323(2022)01-0244-09

1 引 言

太阳诱导叶绿素荧光（Solar-Induced chlorophyll Fluorescence，SIF）指植被在光合作用中发射的一种光学信号，植物叶片在日光照射下将所吸收的一部分光能用于光合作用，另一部分激发叶绿素分子使其发生电子跃迁后，以长波形式发射荧光，或以热能形式向外散耗，叶绿素分子吸收光子，被激发的叶绿素重新发射光子而产生的一种光谱范围为650～800 nm的光信号，因此具有直接指示植被光合作用的巨大潜力，是进行监测GPP和植被光合作用的有效工具[1-3]。叶绿素荧光作为植被光合作用的副产品，相比于植被指数等参数更能直接地反映植被光合作用的相关信息，是植被生理状态的无损探针[4]。近年来，通过卫星遥感技术得到了大量探测全球尺度SIF的研究和产品。Frankenberg等[5-7]基于日本GOSAT卫星绘制了首张荧光地图。由此，引发了SIF研究的热潮，多种可用于探测荧光的卫星反演产品也应运而生。如GOME-2传感器（搭载于MetOp-A/B卫星）可提供740 nm附近，分辨率40 km×80 km的荧光峰值分布图[8]。2009年日本发射的“温室气体观测卫星”（GOSAT）提供直径为10 km分辨率，770 nm附近的叶绿素荧光。美国2014年发射的OCO-2可以提供比GOSAT更高空间分辨率的数据，足迹大小为1.3 km×2.25 km。Sentinel-5P搭载的TROPOMI传感器空间分辨率为7 km×7 km，2019年8月起为3.5 km×7.5 km，也能够获取更高分辨率的荧光产品。2016年中国发

收稿日期：2021-06-16；**修订日期**：2021-09-29
基金项目：国家重点研发计划项目“草地碳收支监测评估技术合作研究”（2017YFE0104500），国家自然科学基金“基于全生命周期分析的多尺度草甸草原经营景观碳收支研究”（41771205），财政部和农业农村部国家现代农业产业技术体系，中央级公益性科研院所基本科研业务费专项（Y2020YJ19，1610132021016）资助。
作者简介：沈 洁（1996－），女，宁夏中卫人，硕士研究生，主要从事草地生态遥感研究。E-mail：JShen_10@163.com
通讯作者：辛晓平（1970－），女，甘肃天水人，研究员，主要从事草地生态遥感研究。E-mail：xinxiaoping@caas.cn

射的 TanSat 卫星能够提供空间分辨率为 2 km×2 km 的数据,未来计划发射或即将发射的荧光探测卫星用 FLEX(FLORIS),美国 TEMPO 卫星与 OCO-3 卫星,以及 GeoCARB 卫星[9],将为 SIF 数据的研究与应用提供更多支持与可能。然而,当前已有的卫星所衍生的叶绿素荧光产品,或存在空间分辨率都较为粗略的问题,或其产品本身只是离散的脚点,不足以直接进行更精细的尺度或生态系统水平的分析[10]。目前,这个问题能通过从 OCO-2 卫星中获得更高分辨率的 SIF 产品得到部分解决[11]。同时,解决 SIF 数据集的空间不连续性,也成为近来全球 SIF 数据研究的重点问题,对于测量光合作用与更好地耦合不同时空尺度的 GPP 等信息将十分有价值[12-14],目前,也有许多学者从神经网络、机器学习等方面探索连续 SIF 数据集的构建[15-19]。基于此,研究将尝试解决 SIF 数据的空间不连续性,通过较高分辨率的 OCO-2 脚点数据,结合反射率、气象因子、土地利用类型等数据生成连续的 SIF 产品。

2　研究区域与数据

2.1　研究区域概况

研究区域为中国大陆内的植被覆盖区域,包含了常绿针叶林、常绿阔叶林、落叶针叶林、落叶阔叶林、混交林、郁闭灌丛、开放灌丛、多树草原、稀树草原、草原、永久湿地、作物共 12 种植被覆盖类型,如图 1,其经度范围为 73°～136° E,纬度范围为 17°～54° N。

2.2　数据来源

轨道碳观测 2 号(Orbiting Carbon Observatory-2,OCO-2)是一颗由美国国家航空航天局(National Aeronautics and Space Administration, NASA)于 2014 年发射的探测大气中二氧化碳浓度的卫星。同时,该卫星也以的“夫琅禾费暗线填充原理”(Fraunhofer Line Discrimination, FLD),应用了 Frankenberg 等[5]作全球尺度 SIF 的算法,开发了自己的 IMAP-DOAS 算法以提取 758 nm 处和 770 nm 处的 O_2-A 波段的 SIF 值,反演得到 L2 级荧光产品。该产品有 3 个观测模式,即星下点观测(Nadir 模式)、闪烁观测(Glint 模式)和目标观测(Target 模式)。其中,Nadir 模式有较好的分辨率,闪烁观测有较高的信噪比,目标观测的单次观测数据点较多,适用于与地面站点的匹配验证。研究收集了 2018 年 5 月至 2019 年该版本的 SIF 数据,并从中提取了已校正的日值 SIF,通过云掩膜编码剔除了有云的脚点数据。

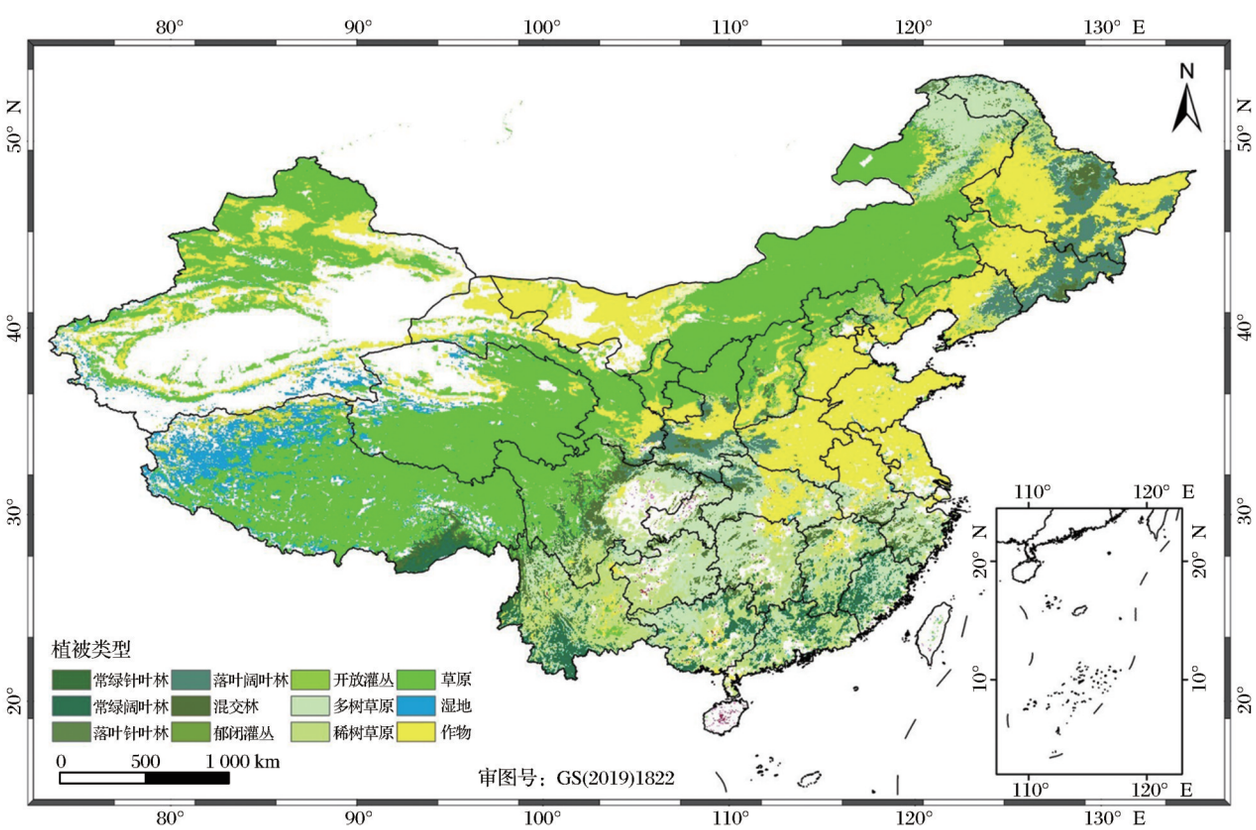


图 1　研究区域范围

Fig.1　Research area

反射率数据来源于中分辨率成像光谱仪MODIS（Moderate-resolution Imaging Spectroradiometer）的MCD43C4产品，此产品的生成过程中都使用了Terra和Aqua数据，从而为质量保证输入数据提供了最高的可能性。该产品包含在3级数据集中提供的16 d数据，该数据集投影到0.05°（5.6 km）的纬度—经度气候建模网格（Climate Model Grid，CMG），包括MODIS在目标日的当地太阳正午天顶角处前7个光谱带的星下反射率（Nadir_Reflectance），并且以地理投影（纬度-经度）形式提供全球每天的反射率数据。气象数据采用全球范围的MERRA-2数据集，该数据集是美国航天局利用戈达德地球观测系统模型5（GEOS-5）及其大气数据同化系统（ADAS）对卫星时代进行的大气再分析。其空间分辨率为0.5°×0.625°（纬向×经向），时间跨度为1980年至今。该数据集针对坐标点提供年统计，月统计，日变化统计和时间序列数据（如1 h分辨率），针对区域提供年平均数据，本文使用其日变化统计的气象数据集，从中获取1 h分辨率的太阳辐射数据用于计算光合有效辐射（Photosynthetically Active Radiation，PAR），以及每日的最大、最小、平均温度，以及降水率，以此来得到日均温度（Temperature，T），计算饱和水汽压差等参数（Vapor Pressure Deficit，VPD），并对该数据进行重采样以得到固定格网数据。

3 研究方法

3.1 数据预处理

日光诱导叶绿素荧光（SIF）与光合有效辐射以及植被状况有关，因此，研究将采用3类参数，即植被条件、气象条件、土地利用类型，来描述预测的SIF值。其中，由植被增强指数（Enhanced Vegetation Index，EVI）代表植被条件，EVI是被用于监测植被状况最广泛的一种植被指数[19]，其计算公式如下：

$$\mathrm{EVI}=\frac{\rho_{\mathrm{NIR}}-\rho_{\mathrm{R}}}{\rho_{\mathrm{NIR}}+C_1*\rho_R-C_2*\rho_{\mathrm{Blue}}+L}*G \tag{1}$$

其中：ρ_{NIR}、ρ_{R}、ρ_{Blue}分别表示近红外波段、红波段、蓝波段的反射率；C_1和C_2为气溶胶阻抗系数；L为土壤调节参数；G为常数。EVI使用蓝波段的数据去校正红波段的气溶剂影响，加入土壤调节参数，使得EVI对大多数冠层背景不敏感（带有雪背景的除外）[20-21]。通常取$C_1=6$，$C_2=7.5$，$L=1$，$G=2.5$。EVI全体取值范围为-2.5～2.5[22]，植被中的EVI取值范围通常在0.2～0.8。从MCD43C4中提取诸天的0.05°反射率数据，由式（2）计算EVI，后计算每8 d的平均EVI值，得到0.05° 8 d分辨率的EVI数据集。

气象条件选取光合有效辐射，气温，饱和水汽压差3个参数来表示可能影响叶绿素荧光发射的环境条件，如太阳辐射，温度、水胁迫等。光合有效辐射为MERRA-2的1 h分辨率的太阳辐射数据集中获取漫射辐射（Diffuse PAR）和直接辐射（Direct PAR）之和，在该数据集中，直接辐射名为光束辐射（Beam PAR），并累积到1 d成为逐天数据，单位：MJ·m^{-2}。气温和饱和水汽压差由逐天的MERRA-2数据集获得，该逐天数据集能够直接获得平均温度，单位：K，此外能够获取每日最高、最低温度和降水率。饱和水汽压差指一定温度下，饱和水汽压与空气中的实际水汽压间的差值，表示实际空气距离水汽饱和状态的程度，由最高、最低温度和水汽压计算方法如下：

$$\mathrm{VPD}=E_{\mathrm{sat}}-\mathrm{VP} \tag{2}$$

$$E_{\mathrm{sat}}=\frac{E(T_{\max})+E(T_{\min})}{2} \tag{3}$$

$$E(T_{\max})=0.6108*\mathrm{e}^{\frac{17.27*(T_{\max}-273.15)}{T_{\max}}} \tag{4}$$

$$E(T_{\min})=0.6108*\mathrm{e}^{\frac{17.27*(T_{\min}-273.15)}{T_{\min}}} \tag{5}$$

其中：E_{sat}是饱和水汽压；VP是水汽压；由降水率而得，降水率单位为kg·m^{-2}·s^{-1}，乘以时间，并换算为水汽压差，单位：KPa，$T_{\max}$和$T_{\min}$分别是日最高、最低温，单位：K。

由于气象数据集经纬向的分辨率不同，需要重采样到0.05°，再聚合为8 d的累积PAR，平均温度和饱和水汽压数据集。土地利用类型由MCD12C1（分辨率：0.05°）中的国际地理圈-生物圈计划（IGBP）分类方案获得。

日光诱导叶绿素荧光由OCO-2 SIF（1.3 km × 2.25 km）的Nadir模式脚点数据获取，该模式近似垂直观测，由此而受到的测量角度影响可忽略不计。获取每个OCO-2 SIF脚点中心位置的经纬度，并将将所有OCO-2 SIF汇总到每个8 d间隔的0.05°×0.05°网格单元，若每个网格单元内捕获到超过5个的SIF脚点值，即以该网格内所有SIF脚点的均值作为其SIF值，单位：W·m^{-2}·μm^{-1}·sr^{-1}。经汇总后，在每8 d的间隔内，这些网格单元占总陆地面积约0.3%。

对于每个具有SIF数据的网格单元，从网格化EVI、PAR、气温、VPD和土地覆盖类型提取对应单元，组成从2018年5月至2019年12月共20个月的

每8 d 0.05°分辨率数据集，共含有约21.5万条记录，将2018年5月至2019年6月约14.0万条数据作为训练集，2019年7月至12月约7.5万条数据作为测试集。集合所有EVI、PAR、气温、VPD和土地覆盖类型数据作为对2018年5月至2019年12月期间连续SIF的预测集。

3.2　Cubist回归树预测算法

OCO-2 SIF数据具有较高的分辨率，但该数据为离散的脚点数据，且呈条带状分布，使得一般的线性插值或重采样方法难以得到较为准确且分辨率高的连续SIF数据，使得OCO-2 SIF数据常被整合为1°分辨率每16 d的数据集，掩盖了该数据的高分辨率优势[23]。因此，需要从非线性或机器学习等方法出发，结合多种连续的遥感数据估计连续且精度较高的SIF数据集。研究使用机器学习方法中的Cubist回归树，以数据驱动的方法开发预测0.05°分辨率的连续日光诱导叶绿素荧光(SIF)。树型算法的基础决策树因其易理解、易构建、运算快等特性，被广泛应用于统计学及数据挖掘等领域。决策树可以分为两类：分类决策树与回归决策树，处理离散型数据时主要用分类决策树，处理连续型数据就会用到回归决策树，后者常被用于预测连续变化的值。经典回归树是由BREIMAN等提出的分类与回归树(Classification And Regression Trees，CART)方法[24]，CART由特征选择、树的生成及剪枝组成，通过不断将数据分为两组，分组时通过穷举每一个特征的每一个阈值来寻找最优切分特征与最优切分点，衡量方法为整体误差平方和最小化。同时，通过类似交叉验证法进行剪枝，避免回归树增长过长。但CART对样本的预测都采用最终叶子节点处所有训练集训练结果的均值，由此导致对新的样本集预测偏差较大，难以达到理想预测效果[25]。

为克服CART等简单的回归树的局限，Quinlan[26]提出了由M5模型树发展而来的Cubist回归树。Cubist回归树的特点在于模型树的叶子节点上是一个线性回归模型，一系列的分段线性模型组合为Cubist回归树，能够很好地解决非线性问题[27]。Cubist树训练规则简单、有效，速度快，对输入空间的分割由算法自动进行，能够处理高维属性的问题。模型树将输入的数据集样本空间划分为不同的长方形区域，其边缘互相平行，在每一层模型树中，选择识别力最强的属性成为子树的根节点，将样本根据该根节点划分为若干个子集。为防止树过度增长，对节点增长设置多个停止条件：节点样本的目标属性标准差与总体样本的目标属性标准差的比例或差值小于某一阈值，或节点的样本数低于某一阈值[27]。建立初步的模型树之后，还需对树进行剪枝，即归并某些冗杂的子树并用叶子节点代替，从而提高模型树的效率与简洁程度，最后需要使用平滑方法对剪枝后叶子节点的不连续行进行补偿，具体需要参考叶子节点的父节点来使用平滑方法，将父节点与叶子节点重新拟合为一个新的线性方程[28]。目前，该模型已被广泛用于估计生物物理变量与碳通量中，如对叶面积指数[29]与净生态系统交换量等的估计[28]。

研究采用有Kuhn等[30]针对R语言开发的Cubist包进行建模，目前该程序包已经更新到0.2.3版本。Cubist在建模和预测时分别通过设置规则数(Committees)和实例数(Instances)进行优化，规则数指在树模型中需要使用的模型数量，取值范围为0～100，实例数表示在预测时需要参考来修正结果的样本数量，取值范围为0～9。然而，规则数越高不一定就代表模型的精度越高，也可能出现过拟合的情况。为避免这类情况发生，提高模型的稳定性和准确性，在训练模型前需要先进行参数优化，以最能精简模型的同时，有较高的模拟预测精度。模型拟合的精度由计算训练集中预测值和观测值的平均绝对误差(Mean Absolute Error，MAE)，相对误差(Relative Error，RE)，和相关系数R(Correlation Coefficient)来衡量。其计算方法如下：

$$\mathrm{MAE}=\frac{1}{N}\sum_{i=1}^{N}\left|y_i-\hat{y}_i\right| \tag{6}$$

$$\mathrm{RE}=\frac{\mathrm{MAE}_T}{\mathrm{MAE}_\mu} \tag{7}$$

$$R=\frac{\mathrm{Cov}(Y,\hat{Y})}{\sqrt{\mathrm{VAR}(Y)\,\mathrm{VAR}(\hat{Y})}} \tag{8}$$

其中：N为数据总数；y_i为观测值；$\hat{y}_i$为模型预测值；MAE_T为模型当前MAE；MAE_μ为预测平均值的MAE。

模型验证精度由计算测试集中预测值与观测值的决定系数R^2(Coefficient of Determination)，和均方根误差(Root Mean Squared Error，RMSE)来衡量。其计算方法如下：

$$R^2=1-\frac{\sum_{i=1}^{N}\left(y_i-\hat{y}_i\right)^2}{\sum_{i=1}^{N}\left(y_i-\bar{y}\right)^2} \tag{9}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i-\hat{y}_i\right)^2} \tag{10}$$

其中：N为数据总数；y_i为观测值；$\hat{y}_i$为模型预测值；$\bar{y}$为观测值的平均值。

4 结果与分析

4.1 基于Cubist回归树的预测模型

4.1.1 模型训练与拟合

Cubist回归树算法的树模型(即规则集)会被自动修剪或合并，因此可以不需要事先进行子集的划分。因此，研究直接通过样本数据对Cubist模型进行参数调优。由于本研究样本数量较大(样本量为万级)，故采用常用于参数调优的经典方法十折交叉验证法[31]，该方法最大的优势在于能重复运用随机产生的子样本进行训练和验证。研究使用训练集数据，以十折交叉验证训练模型中规则数(Committees)和实例数(Neighbor)两个参数，训练结果如图2。

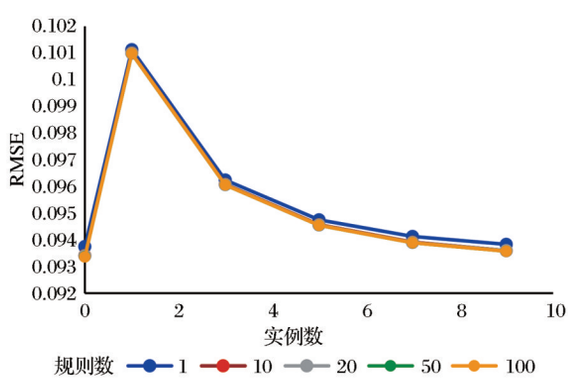


图2 规则数与实例数参数优化

Fig.2 Parameter optimization of Committees and Instances

其中，RMSE最小时规则数为100，实例数为9，其次规则数为10，实例数为9。在实际建模过程中，当规则数为100时模型拟合的时间与预测时间迅速增大，考虑到模型拟合与预测速度，选择规则数为10。在实际预测过程中。实例数为9或0的结果差别不大，因此在预测时的实例数选为0。

研究还考虑了样本数量大小对模型精度的影响，以及土地利用类型信息的影响。同时，为得到较大样本量，加入同时段蒙古区域数据，采用随机抽样，以样本量为10 000、50 000、87 000、139 200、174 000，并分考虑土地利用类型信息与不考虑土地利用类型信息两种情况分别建模，每一次建模都使用验证集所有数据进行验证。计算描述拟合精度的平均绝对误差(MAE)、相对误差(RE)、相关系数(R)、描述拟合精度的决定系数(R^2)和均方根误差RMSE，结果如表1所示。

表1 模型拟合精度与验证精度统计

Table 1 model fitting accuracy and verification accuracy statistics

Land Use=TURE					
	Fitting			Validation	
样本量	MAE	RE	R	R^2	RMSE
10 000	0.077	0.439 7	0.790	0.632 7	0.104 7
50 000	0.077	0.435 6	0.792	0.637 7	0.103 9
87 000	0.077	0.429 6	0.797	0.639 4	0.103 7
139 200	0.076	0.428 9	0.799	0.651 1	0.103 5
174 000	0.069	0.356 0	0.817	0.662 1	0.097 0
Land Use=FALSE					
	Fitting			Validation	
样本量	MAE	RE	R	R^2	RMSE
10 000	0.078	0.446 1	0.781	0.616 8	0.106 9
50 000	0.078	0.439 5	0.786	0.617 4	0.106 9
87 000	0.077	0.433 5	0.791	0.621 5	0.106 3
139 200	0.077	0.432 1	0.794	0.618 1	0.106 7
174 000	0.070	0.359 0	0.813	0.641 6	0.100 1

由表1能够明显看出，随着样本量的增大，模型拟合时的误差，即MAE和RE都在逐渐减小，R在逐渐增大，表示模型拟合的精度越来越好。验证精度中，R^2与RMSE也均随着样本量的增大而变好，但由于RMSE始终较大，使得R^2的提高有限。同时，有土地利用类型的结果均优于不使用土地利用类型的结果。最终选择用全部训练样本训练，使用土地利用类型信息建立的Cubist模型。

4.1.2 模型精度验证

使用测试集数据对训练好的Cubist模型进行验证，验证结果如图3。

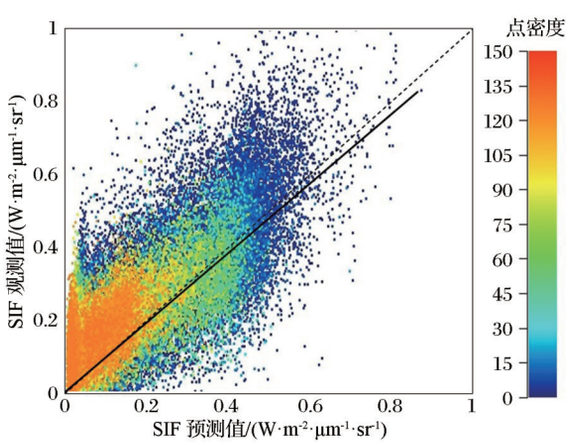


图3 Cubist模型预测SIF与观测SIF验证

Fig.3 Verification of Cubist model prediction SIF and observation SIF

由图 3 可得,R^2为 0.67,RMSE= 0.097,说明SIF 预测值对SIF 观测值有较好的拟合效果。通过与 1∶1 线的对比可得,整体上 SIF 预测值对 SIF 观测值有有一定的低估。均方根误差 RMSE 较大,为 0.097 W·m^{-2}·μm^{-1}·sr^{-1}。

为进一步分析误差的分布,以 0.05 W·m^{-2}·μm^{-1}·sr^{-1}为步长,统计了每个阶段SIF 观测值的数量(图 4(a)),并计算每个阶段处的累积误差(图4(b))。可以看出SIF值主要分布在0～0.2 W·m^{-2}·μm^{-1}·sr^{-1}中,累积误差显示了误差来源与SIF值的分布基本一致,由于SIF值的分布,其误差也主要集中在0～0.3 W·m^{-2}·μm^{-1}·sr^{-1}内,这一结果与XIAO 等(2019)的全球SIF 连续预测研究结果相同。但当SIF 观测值超过0.5 W·m^{-2}·μm^{-1}·sr^{-1}时的累积误差并未饱和,又呈现出增加的趋势。在观测值超过0.5 W·m^{-2}·μm^{-1}·sr^{-1}后,才又逐渐趋于平缓。造成这一现象的原因可能是由于卫星获取SIF 数据时,超过 0.5 W·m^{-2}·μm^{-1}·sr^{-1}的荧光值反演误差变大,同时,本研究仅针对于中国及蒙古国的区域而进行,样本的范围和数量较为局限,也可能在预测过程中导致累积的误差较大,使得模型的RMSE较大,R^{2}较小。

此外,根据土地利用类型对不同生态群落进行验证,将常绿针叶林,常绿阔叶林,落叶针叶林,落叶阔叶林,和混交林归为森林一类,多树草原,稀树草原,和草原归为草地类,作物类,以及永久湿地

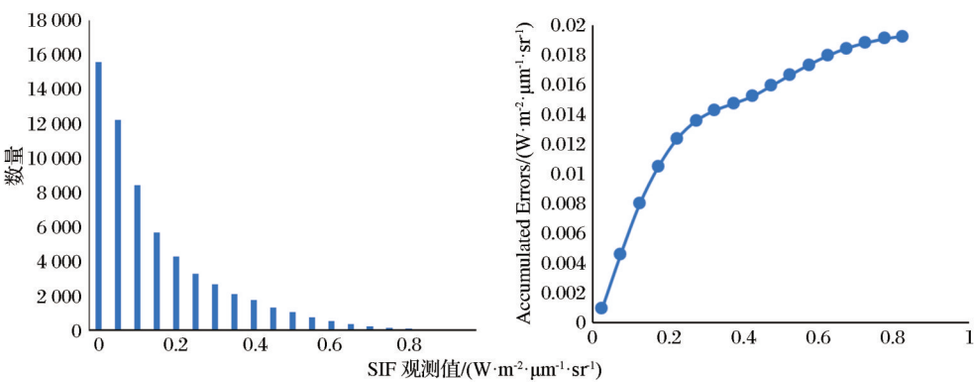


图 4　误差统计(a)SIF 观测值分布与(b)SIF 观测值的累积误差

Fig.4　error statistics (a) distribution of SIF observations and (b) Accumulative Error of SIF observations

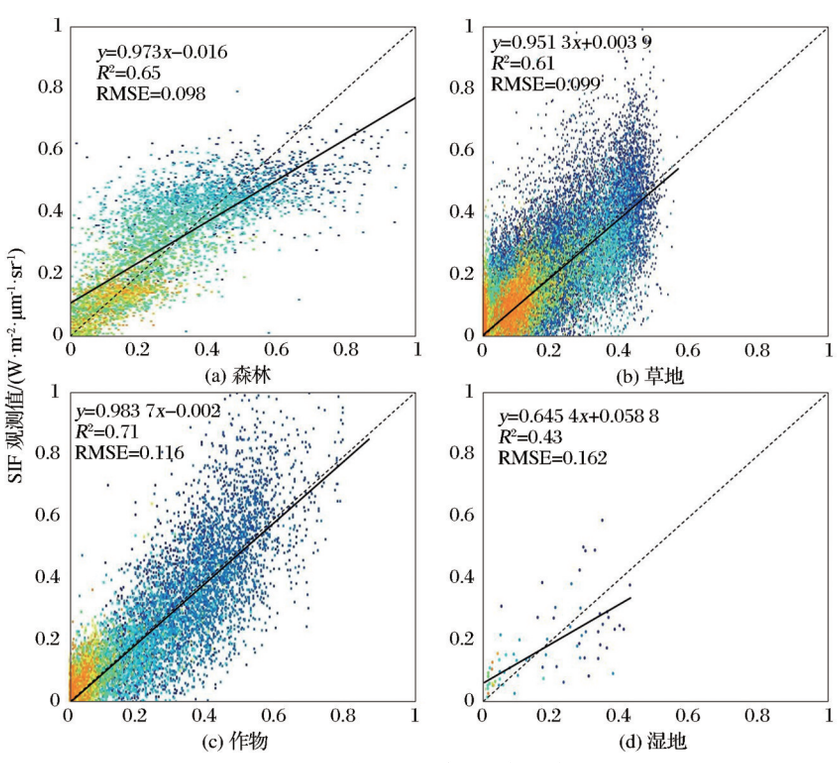


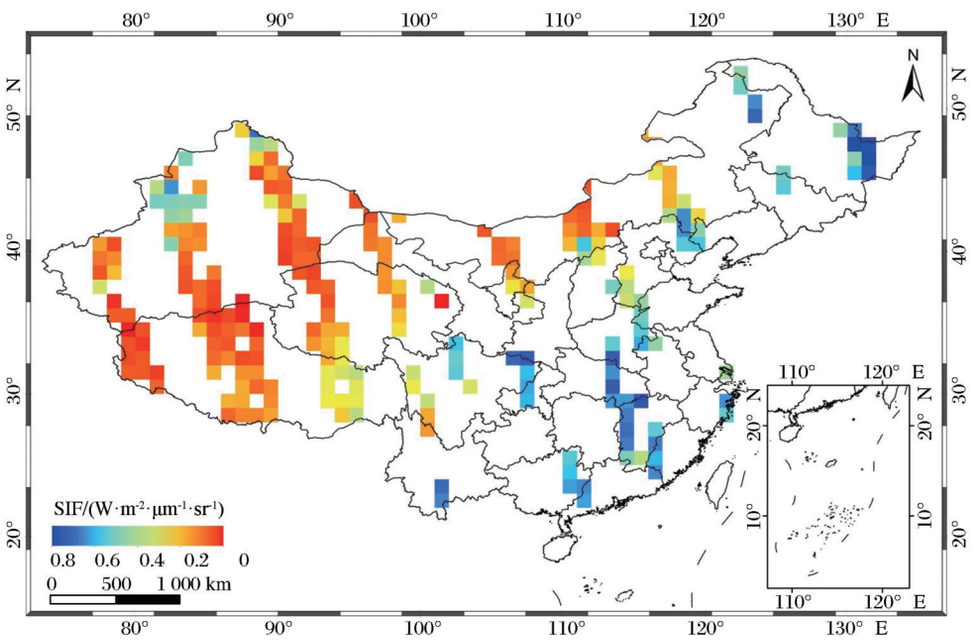
图 5　不同生态群落中 Cubist 模型预测 SIF 与观测 SIF 验证

Fig.5　Verification of Cubist model prediction SIF and observation SIF in different ecological communities

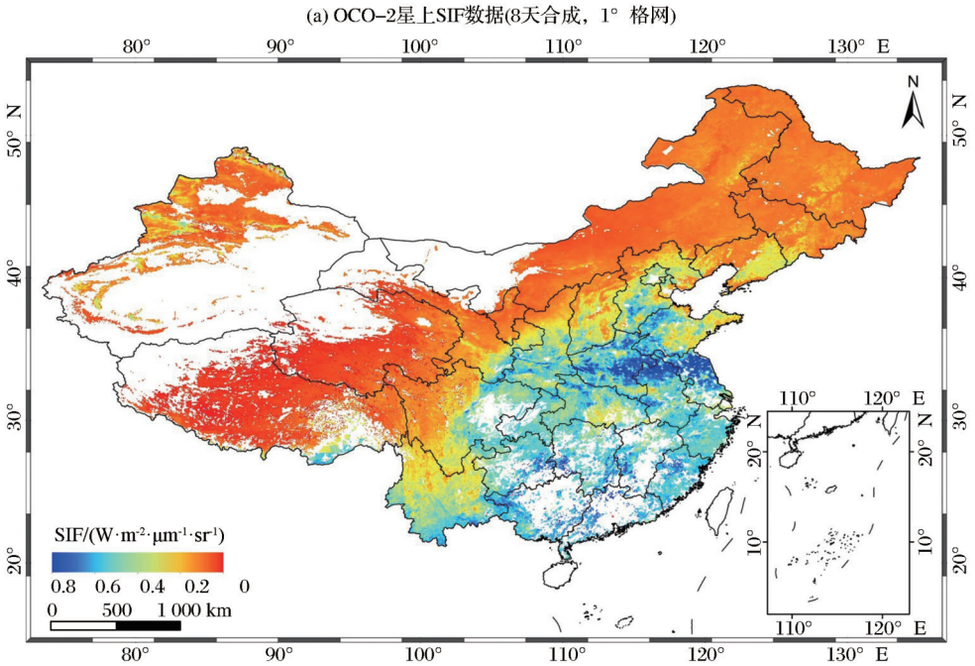
类，如图5，分别得到Cubist模型在预测不同生态群落中的精度。其中，Cubist回归树模型在对作物类图5(c)中的SIF模拟精度最高，R^2为0.71，RMSE=0.116 W·m^{-2}·μm^{-1}·sr^{-1}。此外，模型在森林类图5(a)和草地类图5(b)的SIF模拟都具有较好的表现，R^2分别为0.65和0.61，RMSE分别为：0.098 W·m^{-2}·μm^{-1}·sr^{-1}和0.099 W·m^{-2}·μm^{-1}·sr^{-1}。验证集中永久湿地类图5(d)的点较少，其验证结果也较差，R^2为0.43，RMSE=0.162 W·m^{-2}·μm^{-1}·sr^{-1}。

4.2 高分辨率连续日光诱导叶绿素荧光数据集

研究将每8 d的EVI，光合有效辐射(PAR)，平均温度，饱和水汽压差(VPD)和土地利用类型逐像元对应，建立空间连续的预测数据集。采用在3.1.1中经样本训练好的Cubist回归树模型，预测每8 d的0.05°分辨率的连续SIF。图6以2018年7月4日至11日为例，展示了聚合到每8 d的1°分辨率OCO-2 SIF网格数据和本研究做出的由Cubist预测的连续SIF产品。可以看到Cubist模型预测的SIF具有较



(a) OCO-2星上SIF数据(8天合成，1° 格网)



(b) OCO-2星上SIF重建数据(8天合成，0.05° 格网)

审图号：GS(2019)1822号

图6 每8天聚合到1°SIF格网数据，Cubist模型预测SIF数据对比(以2018年7月4日~11日为例)

Fig.6 Aggregate to 1° SIF grid data every 8 days, cubist model forecast SIF data comparison (take July 4, 2018-july 11, 2018 as an example)

高的空间分辨率与连续性,能够用于进一步的科学研究,其中,中国-蒙古区域整体呈现出东部、南部较高,向西、向北逐渐降低的规律。

5　结语

本文主要研究了对OCO-2 SIF数据的连续预测方法,生成了每8 d 0.05°分辨率的SIF产品。预测连续SIF过程中,本研究基于Cubist回归树模型,结合由MODIS反射率数据计算所得,代表植被生长状况的EVI数据,代表气候条件的平均温度、PAR、VPD、以及土地利用类型数据进行回归树建模。通过调优参数选择Cubist回归树中规则数(Committees)为10进行模型拟合。考虑样本大小的变化对拟合结果的影响,结果发现在目前的样本数(17.4万)中,模型精度一直随样本数量的增加而增加,同时考虑土地利用类型信息对模型拟合的影响,结果表明使用土地利用类型的模型精度略高于不使用该因子的模型精度。最终,采用全部训练样本并加入土地利用类型因子进行建模,模型的验证精度较好,为R^2=0.67,RMSE=0.097,表明加入的土地利用类型因子提供了更多信息。Cubist能够应用于大尺度及长时序的SIF数据重建中,然而对比于龙龙等[23]用卷积神经网络(Convolutional Neural Networks, CNN)对其小范围的感兴趣区域内OCO-2 SIF建模的验证精度不算很高,主要是由于在大尺度的重建中,会引入较多的噪声,可以通过像元植被覆盖度的大小对SIF数据进行进一步地筛选,提高算法精度。

该模型对不同生态群落(森林,草地,作物,永久湿地)的预测精度不同,其中,农田的验证精度最高,决定系数R^2为0.71,主要是由于农田的景观较为一致,使得该模型能较准确地进行预测。此外,对林地、草地的验证结果也较好,R平方分别为0.65和0.61。研究表明,日光诱导叶绿素荧光(SIF)的值主要集分布在0~0.2 $W\cdot m^{-2}\cdot \mu m^{-1}\cdot sr^{-1}$,该模型预测的误差主要集中在SIF低于0.3 $W\cdot m^{-2}\cdot \mu m^{-1}\cdot sr^{-1}$时。本研究由Cubist模型预测出的更为精细的每8 d 0.05°分辨率SIF数据集,能够较好地表现植被变化的季节周期,描述植被生长状况。

参考文献(Reference):

[1] Zhang Y, Guanter L, Berry J A, *et al*. Model-based analysis of the relationship between Sun-Induced chlorophyll Fluorescence and gross primary production for remote sensing applications [J]. Remote Sensing of Environment, 2016, 187 : 145-155.

[2] Albert P C, Esa T, Jon A, *et al*. Linking chlorophyll a fluorescence to photosynthesis for remote sensing applications: mechanisms and challenges [J]. Journal of experimental botany, 2014, 65(15) : 4065-4095.

[3] Frankenberg C, Berry J. Solar Induced chlorophyll Fluorescence: origins, relation to photosynthesis and retrieval [J]. Comprehensive Remote Sensing, 2018, 3:143-162.

[4] Meroni M, Rossini M, Guanter L, *et al*. Remote sensing of Solar-Induced chlorophyll Fluorescence: Review of methods and applications [J]. Remote Sensing of Environment, 2009, 113(10):2037-2051.

[5] Frankenberg, Fisher, J B, *et al*. New global observations of the terrestrial carbon cycle from GOSAT: Patterns of plant fluorescence with gross primary productivity [J]. Geophysical Research Letters, 2011, 38(17): 351-365.

[6] Guanter L, Frankenberg C, Dudhia A, *et al*. Retrieval and global assessment of terrestrial chlorophyll fluorescence from GOSAT space measurements [J]. Remote Sensing of Environment, 2012, 121(none):236-251.

[7] Joiner J, Yoshida Y, Vasilkov A P, *et al*. Filling-in of far-red and near-Infrared solar lines by terrestrial and atmospheric effects: simulations and space-based observations from SCIAMACHY and GOSAT [J]. Atmospheric Measurement Techniques Discussions, 2012, 5(1):163-210.

[8] Köhler P, Guanter L, Joiner J. A linear method for the retrieval of Sun-Induced chlorophyll Fluorescence from GOME-2 and SCIAMACHY data [J]. Atmospheric Measurement Techniques, 8, 6(2015-06-26), 2015, 8:2589-2608.

[9] Ji Menghao, Tang Bohui, Li Zhaoliang. Review of Solar-Induced chlorophyll Fluorescence retrieval methods from satellite data [J]. Remote sensing Technology And Application, 2019, 34(3):455-466.[纪梦豪,唐伯惠,李召良.太阳诱导叶绿素荧光的卫星遥感反演方法研究进展[J].遥感技术与应用,2019,34(3):455-466.]

[10] Li X, Xiao J, He B. Chlorophyll fluorescence observed by OCO-2 is strongly related to gross primary productivity estimated from flux towers in temperate forests [J]. Remote Sensing of Environment, 2018, 204:659-671.

[11] Frankenberg C, O'Dell C, Berry J, *et al*. Prospects for chlorophyll fluorescence remote sensing from the Orbiting Carbon Observatory-2 [J]. Remote Sensing of Environment, 2014, 147:1-12.

[12] Duveiller G, Cescatti A. Spatially downscaling Sun-Induced chlorophyll Fluorescence leads to an improved temporal correlation with gross primary productivity [J]. Remote Sensing of Environment, 2016, 182:72-89.

[13] Yu L, Wen J, Chang C Y, *et al*. High-resolution global contiguous SIF of OCO-2 [J]. Geophysical Research Letters, 2019, 46:1449-1458.

[14] Bishop C M. Neural networks for pattern recognition [M]. Oxford: Oxford University Press, 1995.

[15] Zhang Y, Joiner J, Alemohammad S H, *et al*. A global spatially contiguous Solar-Induced Fluorescence (CSIF) dataset using neural networks [J]. Biogeosciences, 2018, 15(19): 5779-5800.

[16] Gentine P, Alemohammad S H. Reconstructed Solar-Induced Fluorescence: a machine learning vegetation product based on MODIS surface reflectance to reproduce GOME-2 solar-induced fluorescence [J]. Geophysical Research Letters, 2018, 45(7): 3136-3146.

[17] Duveiller G, Filipponi F, Walther S, *et al*. A spatially downscaled Sun-Induced Fluorescence global product for enhanced monitoring of vegetation productivity[J]. Earth System Science Data, 2020, 12(2): 1101-1116.

[18] Ma Y, Liu L, Chen R, *et al*. Generation of a global spatially continuous TanSat Solar-Induced chlorophyll Fluorescence product by considering the impact of the solar radiation intensity[J]. Remote Sensing, 2020, 12(13): 2167.

[19] Huete A, Didan K, Miura T, *et al*. Overview of the radiometric and biophysical performance of the MODIS vegetation indices[J]. Remote Sensing of Environment, 2002, 83(1-2): 195-213.

[20] Zhengxing W, Chuang L, Alfredo H, *et al*. From AVHRR-NDVI to MODIS-EVI: advances in vegetation index research [J]. Acta Ecologica Sinica, 2003, 23(5):979-987.[王正兴, 刘闯, Alfredo H. 植被指数研究进展:从AVHRR-NDVI到MODIS-EVI[J]. 生态学报, 2003,23(5):143-151.]

[21] Zhang C. Using MODIS vegetation index to study urban expansion and change[J]. Meteorological, 2006,32(10):20-26.[张春桂. 用MODIS植被指数研究福州城区空间扩展变化[J]. 气象, 2006, 32(10):20-26.]

[22] Ma Rui. Research on enhanced vegetation index algorithm and its application in the ecological environmental remote sensing production subsystem[D]. Kaifeng: Henan University, 2015.[马瑞. 增强植被指数算法的研究及其在生态环境遥感产品生产分系统的应用[D]. 开封:河南大学,2015.]

[23] Yu Longlong, Luo Ze, Yan Baoping. Reconstruction framework of high resolution sif remote sensing dataset in regions of interest[J]. Computer Systems & Applications, 2019, 28(9): 133-139.[于龙龙, 罗泽, 阎保平. 兴趣区域高分辨率叶绿素荧光遥感数据集重建框架[J]. 计算机系统应用, 2019, 28(9):133-139.]

[24] Breiman L. Classification and regression trees [M]. New York:Chapman and Hall,1984.

[25] Ma Ziqiang. Downscaling satellite-based precipitation estimates over the Oinghai-Tibetan Plateau at different temporal scales[D]. Hangzhou:Zhejiang University, 2017.[马自强. 青藏高原地区卫星降水数据时空降尺度研究[D]. 杭州:浙江大学, 2017.]

[26] Quinlan J R.Simplifying decision trees[J]. International Journal of Man-Machine Studies, 1987,27(3):221-234.

[27] Dai S, Yingchun F U, Zhao Y, *et al*. The remote sensing model for estimating urban impervious surface percentage based on the cubist model tree[J]. Journal of Geo-Information Science, 2016,18(10):1399-1409.[戴舒 付迎春 赵耀龙. 基于Cubist模型树的城市不透水面百分比遥感估算模型[J]. 地球信息科学学报, 2016,18(10):1399-1409.]

[28] Zhuang X J. Estimation of net ecosystem carbon exchange for the conterminous United States by combining MODIS and AmeriFlux data[J]. Agricultural & Forest Meteorology Amsterdam Elsevier, 2008,148(11):1827-1847.

[29] Gao F, Anderson M C, Kustas W P, *et al*. Retrieving Leaf Area Index from Landsat using MODIS LAI products and field measurements[J]. IEEE Geoscience & Remote Sensing Letters, 2014, 11(4):773-777.

[30] Kuhn M, Weston S, Keefer C. Cubist: rule- and instance-based regression modeling. http://ftp.ussg.ju.edu/CRAN/web/packages/cubist/,2014.

[31] Kuhn M, Johnson K. Applied Predictive Modeling[M]. New York:Springer,2013.

Reconstruction of SIF Remote Sensing Data of Vegetation in China based on Cubist

Shen Jie[1], Xin Xiaoping[1], Zhang Jing[2], Miao Chen[2], Wang Xü[1], Ding Lei[1], Shen Beibei[1]

(1.*Institute of Agricultural Resources and Agricultural Regional Planning, Chinese Academy of Agricultural Sciences, Beijing* 100081, *China*;

2.*National Remote Sensing Center of China, Beijing* 100036, *China*)

Abstract: Solar-Induced Chlorophyll Fluorescence (SIF) is the spectral signal (650~800 nm) emitted by plants in the process of photo-synthesis under sunlight conditions. SIF is more direct than vegetation index and other parameters. Reflecting the relevant infor-mation of vegetation photosynthesis, it brings a new way for large-scale Gross Primary Productivity(GPP)estimation. However, the current satellite SIF data may have insufficient resolution or discontinuity in the data space, which is difficult to apply to the estimation of continuous GPP on a large scale. OCO-2 SIF data has high spatial resolution, but it is spatially discrete data. In response to the above problems, this paper focuses on the method of con-tinuous prediction of discrete OCO-2 SIF data to generate a high-precision continuous SIF data set of the China-Mongolia grassland ecosy-stem. The results are as follows: Through the Cubist regression tree algorithm, combined with MODIS reflectance data, meteorologi-cal data and land use types, a continuous SIF data set with a resolution of 0.05° every 8 days is established, and the prediction accuracy is R^2= 0.65 and RMSE = 0.114. Among them, the accuracy of crop SIF prediction is the highest, with R^2= 0.71 and RMSE= 0.117; the second is the prediction of forest and grassland, with R^2 and RMSE of 0.64/0.123 and 0.60/0.112 respectively.

Key words: Solar-Induced chlorophyll fluorescence; Cubist model; Data reconstruction