

DOI: 10.20103/j.stxb.202408101889

高子滢,王海燕,何潇,雷相东,仇皓雷.基于机器学习算法的吉林省长白落叶松林最大净初级生产力影响因素分析.生态学报,2025,45(10): 4927-4939.

Gao Z Y, Wang H Y, He X, Lei X D, Qiu H L. Analysis of influencing factors of the maximum of NPP (NPP_{max}) of *Larix olgensis* plantations in Jilin Province based on machine learning algorithm. Acta Ecologica Sinica, 2025, 45(10): 4927-4939.

基于机器学习算法的吉林省长白落叶松林最大净初级生产力影响因素分析

高子滢¹, 王海燕^{1,*}, 何 潇², 雷相东², 仇皓雷¹

1 北京林业大学林学院, 林木资源高效生产全国重点实验室, 森林培育与保护教育部重点实验室, 北京 100083

2 中国林业科学研究院资源信息研究所, 北京 100091

摘要:净初级生产力(NPP)是评估森林生态系统质量的重要指标,其最大值(NPP_{max})可以反映森林的生产潜力,研究植被 NPP_{max} 对合理利用森林资源和森林可持续经营具有重要意义。以吉林省长白落叶松(*Larix olgensis* A. Henry)林为对象,基于吉林省一类清查固定样地调查数据、2000—2022 年的 MOD17A3 整合提取的 NPP_{max} 数据、环境因子数据(地形、气候和土壤因子),采用相关性分析、机器学习算法、Shapley 加性解释(SHAP)等方法,分析 NPP_{max} 变化的驱动因子。4 种机器学习算法中,随机森林模型(RF)对落叶松林 NPP_{max} 的预测精度最高, R^2 为 0.540,对应的 RMSE 值为 $11.320 \text{ gC m}^{-2} \text{ a}^{-1}$,其它模型依次为增强回归树($R^2 = 0.523$)、人工神经网络($R^2 = 0.519$)和支持向量机模型($R^2 = 0.501$)。基于 RF 模型 SHAP 分析显示,影响落叶松林 NPP_{max} 变化的主导因素为 18℃ 以上年积温、全氮和海拔。年积温、全氮含量和海拔对 NPP_{max} 存在阈值效应,当 18℃ 以上年积温接近 200℃、全氮含量达到 2.0 g/kg 以及海拔约为 800 m 时, NPP_{max} 的增减变化明显。通过机器学习模型和 SHAP 分析处理 NPP_{max} -环境因子复杂非线性关系,较好地解释了环境变量对 NPP_{max} 的影响机制。研究表明环境变量和 NPP_{max} 之间存在明显的阈值效应,为落叶松林立地质量评价提供了基础。

关键词: NPP 最大值;长白落叶松林;机器学习;SHAP 分析

Analysis of influencing factors of the maximum of NPP (NPP_{max}) of *Larix olgensis* plantations in Jilin Province based on machine learning algorithm

GAO Ziying¹, WANG Haiyan^{1,*}, HE Xiao², LEI Xiangdong², QIU HaoLei¹

1 State Key Laboratory of Efficient Production of Forest Resources, Key Laboratory of Forest Cultivation and Protection Ministry of Education, College of Forestry, Beijing Forestry University, Beijing 100083, China

2 Institute of Forest Resource Information Techniques, Chinese Academy of Forestry, Beijing 100091, China

Abstract: Net primary productivity (NPP) is a key indicator for assessing the quality of forest ecosystems, and its maximum value (NPP_{max}) reflects the potential for forest productivity. Studying vegetation NPP_{max} is crucial for the rational utilization of forest resources and sustainable forest management. This study focuses on *Larix olgensis* A. Henry plantations in Jilin Province as the research subject. Using the national forest inventory data of permanent plots in Jilin Province, NPP_{max} data extracted by MOD17A3 from 2000 to 2022, and environmental factors (terrain, climate, and soil), this study applied correlation analysis, machine learning algorithms, and Shapley additive explanation (SHAP) methods to analyze the driving factors of NPP_{max} change. Among the four machine learning algorithms, the random forest model (RF) predicted the NPP_{max}

基金项目:十四五重点研发项目课题(2022YFD2200501)

收稿日期:2024-08-10; 网络出版日期:2025-03-19

* 通讯作者 Corresponding author. E-mail: haiyanwang72@aliyun.com

of *L. olgensis* plantations with the highest accuracy ($R^2 = 0.540$, $RMSE = 11.320 \text{ gC m}^{-2} \text{ a}^{-1}$), followed by the boosted regression tree model ($R^2 = 0.523$), artificial neural network ($R^2 = 0.519$), and support vector machine model ($R^2 = 0.501$). The SHAP analysis based on the RF model revealed that the main factors affecting the change in $NPP_{\max}$ were the annual accumulated temperature above 18°C , total nitrogen, and elevation. The annual accumulated temperature, total nitrogen content and elevation had a threshold effect on $NPP_{\max}$. The $NPP_{\max}$ significantly varied when the annual accumulated temperature above 18°C approached 200°C , the soil total nitrogen content reached 2.0 g/kg , and the elevation was approximately 800 m above sea level. By using a machine learning model and SHAP analysis to address the complex nonlinear relationship between $NPP_{\max}$ and the environment, we can better explain the mechanism between $NPP_{\max}$ and environmental predictors. An obvious threshold effect exists between the environmental variables and $NPP_{\max}$, which provides a basis for evaluating the site quality of *L. olgensis* plantations.

Key Words: NPP maximum; *Larix olgensis* plantation; machine learning; SHAP analysis

植被净初级生产力(Net Primary Productivity, NPP)指绿色植被在单位时间、单位面积内经光合作用积累并扣除自养呼吸消耗后的有机物质总量^[1],反映了植被的生产能力和生态系统的质量^[2]。植被 NPP 最大值($NPP_{\max}$)表示生态系统健康,可以充分利用光和水资源,不受自然灾害影响,能反映当地植被生产力潜力,也是衡量植被固碳潜力的重要指标。环境因子是植物生长潜力的主要限制因素之一,研究其与植被 $NPP_{\max}$ 的关系对合理利用自然环境资源、充分发挥植被生产潜力以及评价立地质量状况具有重要意义。

植被 NPP 的动态变化反映了植被对环境因子的综合响应。近年来,研究主要集中在气候、土壤、地形和人类活动等因素对不同地区植被 NPP 变化的驱动作用。Xi 等^[3]利用 MOD17A3 NPP 数据分析了气候、地形和土地利用变化对 NPP 空间异质性的影响,发现土地利用变化是影响 NPP 空间异质性的最显著因子。闫丰等^[4]利用偏相关和复相关方法对太行山-燕山植被 NPP 驱动因子进行分析,发现人类活动对 NPP 变化具有较大影响。何颖等^[5]通过地理探测器分析了气候、地形、土壤类型等因子对流域 NPP 空间异质性的影响,指出年均降水量和土壤类型对 NPP 变化解释力最强。随着人工智能技术的发展,机器学习算法在处理复杂非线性关系及适应多维数据方面具有明显优势^[6],其灵活性和泛化能力使其在实际应用中得到了广泛应用和认可,为分析植被生产力与环境因子的关系提供了新的方法。Yu 等^[7]发现随机森林模型在估算 NPP 方面具有良好前景,并可基于过程的模型 MOD17A3 NPP 相媲美。卫星遥感技术和大数据的进步为植被 NPP 的研究提供了丰富的数据支持,相关 NPP 估算模型和监测方法不断成熟。尽管前人的研究在 NPP 领域取得了显著成果,但多数研究仅聚焦于 NPP 本身^[8],而对 $NPP_{\max}$ 反映生态系统植被生长潜力的关键指标却鲜有讨论^[9]。 $NPP_{\max}$ 受多种环境因子的影响,然而这些因子与 $NPP_{\max}$ 的具体关系尚不明确,且量化分析仍较少,机器学习算法的发展为解析 $NPP_{\max}$ 的影响因子提供了新途径,能更深入地探讨植被生长潜力及其对环境变化的响应机制。

落叶松是我国东北地区人工造林的主要树种之一,分别占吉林省森林面积和蓄积量的 8.20% 和 5.60% ^[10],对木材生产和防护等生态系统服务具有重要作用。本文以吉林省长白落叶松(*Larix olgensis* A. Henry)林为研究对象,基于 $NPP_{\max}$ 、气候变量、地形因子、土壤因子等数据,利用相关性分析、机器学习模型和 Shapley 加性解释等方法,构建 $NPP_{\max}$ 和环境因子之间的关系模型,揭示影响长白落叶松林 $NPP_{\max}$ 的主导因素,以期为提高落叶松人工林生长潜力和立地质量评价提供科学依据。

1 研究区概况

吉林省地跨 $121^\circ38'—131^\circ19'E$ 、 $40^\circ50'—46^\circ19'N$ 。属温带大陆性季风气候,四季分明,雨热同季。全省年平均气温在 $-2—6^\circ\text{C}$ 之间,年平均降水量为 $500—600 \text{ mm}$ 。地貌形态差异明显,地势由东南向西北倾斜,呈现明显的东南高、西北低的特征。土壤类型主要包括黑土、黑钙土、草甸土、棕壤等。森林资源丰富,类型多

样,主要集中分布在东部山区和中东部低山丘陵地区,主要树种有红松(*Pinus koraiensis*)、云杉(*Picea asperata*)、长白落叶松、樟子松(*Pinus sylvestris* var. *mongolica*)、蒙古栎(*Quercus mongolica*)、白桦(*Betula platyphylla*)、水曲柳(*Fraxinus mandshurica*)、胡桃楸(*Juglans mandshurica*)和杨树(*Populus L.*)等^[11]。研究区固定样地分布见图 1。

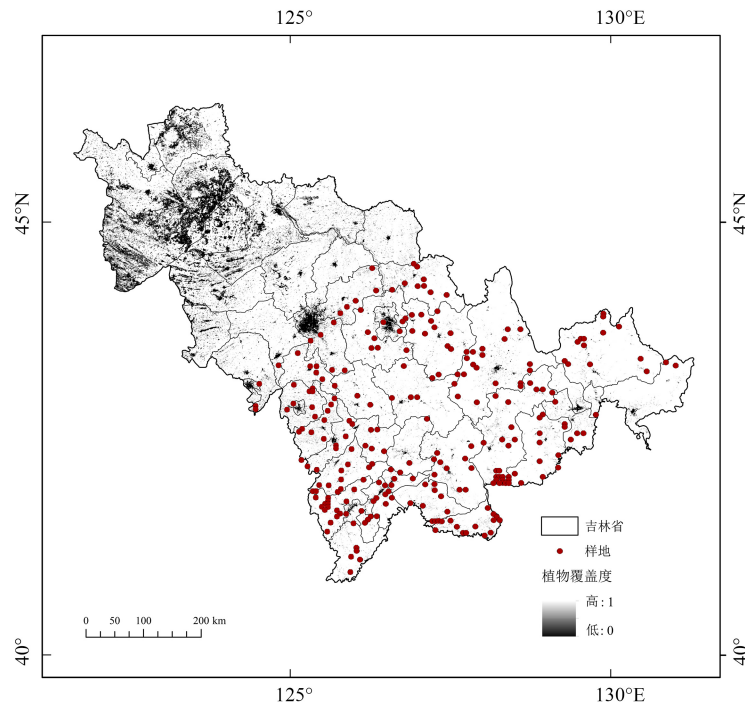


图 1 研究区固定样地分布

Fig.1 Distribution of permanent plots in the study area

2 材料与方法

2.1 数据来源

本研究采用吉林省第 5—9 次森林资源连续清查的落叶松人工纯林固定样地数据,即样地中落叶松断面积占比超过 65%的样地,样地面积为 0.06 hm²。样地调查因子包括经纬度坐标、海拔、坡度、坡向、坡位、土壤类型、土壤厚度、枯枝落叶厚度、腐殖质厚度等立地因子。

NPP 数据来源于美国国家航空航天局网站 (<https://urs.earthdata.nasa.gov/>) 提供的 2000—2022 年 MOD17A3 NPP 数据,空间分辨率为 500 m,时间分辨率为 1 a。使用 MODIS Reprojection Tool (MRT) 工具对数据产品进行拼接、重采样和重投影等操作;并基于研究区矢量数据批量裁剪得到 2000—2022 年 NPP 的时间序列图像,从而获得 2000—2022 年间 NPP_{max} 数据。利用样地经纬度坐标提取 NPP_{max},数据的逻辑检查主要包括:删除株数少于 20 株的样地,共 262 块固定样地;以 NPP_{max} 为目标变量,利用箱线图法剔除异常值,得到 252 块有效样地数据。

气候数据的获取选自亚太地区气候软件 ClimateAP^[12],其可以根据经纬度与海拔提供亚太地区范围的气候数据,本研究每个样地参与建模的气候变量为 16 个,每个样地的气候因子为该样地在 2000—2022 年间的气候变量平均值。

土壤数据取自 SoilGrid 全球土壤数据集(<https://www.soilgrids.org/>),空间分辨率为 250 m,包含阳离子交换量、全氮、有机碳含量等 11 个土壤属性变量。本研究的 NPP_{max} 及环境因子统计量如表 1 所示。

表 1 吉林省落叶松人工林 NPP 的环境变量

Table 1 Environmental variables for NPP of larch plantations in Jilin Province

变量类型 Variable type	基本变量 Basic variable	变量代码 Variable code	单位 Unit	平均值±标准差(或分类变量的水平) Mean ± Standard deviation (or Levels of a categorical variable)
最大植被生产力 Maximum vegetation productivity	植被净初级生产力	NPP _{max}	gC m ⁻² a ⁻¹	616.06±85.64
地形因素 Topographic factor	海拔	Elevation	m	600.79±289.15
	坡向	Aspect	—	(1)北坡 (2)东北坡 (3)东坡 (4)东南坡 (5)南坡 (6)西南坡 (7)西坡 (8)西北坡 (9)无坡向
	坡位	position	—	(1)脊部 (2)上坡 (3)中坡 (4)下坡 (5)山谷 (6)平地
	坡度	Slope	°	9.74±7.48
气候变量 Climatic variable	年均温	MAT	℃	4.35±1.31
	最暖/冷月均温	MWMT MCMT	℃	21.35±1.69 -15.72±0.85
	最暖最冷月温差	TD	℃	37.07±1.48
	年均降水量	MAP	mm	748.02±110.49
	湿热指数	AHM	—	19.78±3.89
	18℃以上/下年积温	DD18 DD_18	℃	243.27±99.67 5132.91±372.85
	0℃以下年积温	DD_0	℃	1446.36±141.89
	5℃以上年积温	DD5	℃	2029.04±297.77
	无霜期天数	NFFD	d	177.97±11.49
	前一年 8 月至当年 7 月的降雪 降水量	PAS	mm	75.05±27.30
	近 30 年的极端最高/低温	EXT EMT	℃	33.02±1.34 -30.11±1.36
	哈格里夫斯相对蒸发	Eref	—	703.85±40.14
	哈格里夫斯水分亏缺	CMD	—	136.54±51.53
土壤变量 Soil variable	细土粒密度	bdod	kg/dm ³	1.34±0.04
	石砾体积分数	cfvo	cm ³ /100cm ³	21.55±4.52
		sand		31.61±5.78
	土壤质地(土壤机械组成)	slit clay	—	44.00±4.22 24.37±3.36
	土壤 pH	pH	—	6.81±0.26
	阳离子交换量	CEC	cmol/kg	23.65±0.26
	土壤有机碳含量	SOC	g/kg	24.62±5.96
	全氮	TN	g/kg	1.71±0.41
	有机碳密度	OCD	kg/dm ³	15.52±2.09
	有机碳储量	OCS	t/hm ²	63.17±8.81
	土壤类型	Type	—	(113)暗棕壤 (151)沼泽土 (162)白浆土 (167)草甸土 (186)火山灰土
	土壤厚度	ST	cm	41.99±10.38
	腐殖质厚度	HT	cm	4.25±3.41
	枯枝落叶厚度	LT	cm	3.35±2.14

NPP_{max}: 植被净初级生产力最大值 Maximum vegetation net primary productivity; elevation: 海拔 elevation; Aspect: 坡向 Slope aspect; position: 坡位 Slope position; Slope: 坡度 slope; MAT: 年均温 Mean annual temperature; MWMT: 最暖月均温 Mean warmest month temperature; MCMT: 最冷月均温 Mean coldest month temperature; TD: 最暖最冷月温差 Temperature difference between MWMT and MCMT; MAP: 年均降水量 Mean annual precipitation; AHM: 湿热指数 Annual heat; moisture index; DD_18: 18℃以下年积温 Degree-days below 18℃; DD18: 18℃以上年积温 Degree-days above 18℃; DD_0: 0℃以下年积温 Degree-days below 0℃; DD5: 5℃以上年积温 Degree-days above 5℃; NFFD: 无霜期天数 Number of frost-free days; PAS: 前一年 8 月至当年 7 月的降雪降水量 Precipitation as snow between August in previous year and July in current year; EXT: 超过 30 年的极端最高温度 Extreme

maximum temperature over 30 years; EMT: 超过 30 年的极端最低温度 Extreme minimum temperature over 30 years; Eref: 哈格里夫斯相对蒸发 Hargreaves reference evaporation; CMD: 哈格里夫斯水分亏缺 Hargreaves climatic moisture deficit; bdod: 细土粒密度 Fine soil grain density; cfvo: 石砾体积分数 Gravel volume fraction; sand: 砂粒含量 Sand content; silt: 粉粒含量 Silt content; clay: 黏粒含量 Clay content; pH: 土壤 pH Soil pH; CEC: 阳离子交换量 Cation exchange capacity; SOC: 有机碳含量 Soil organic carbon content; TN: 全氮 Total nitrogen; OCD: 有机碳密度 Organic carbon density; OCS: 有机碳储量 Organic carbon stock; Type: 土壤类型 Soil type; ST: 土壤厚度 Soil thickness; HT: 腐殖质厚度 Humus thickness; LT: 枯枝落叶厚度 Litter thickness

2.2 研究方法

2.2.1 变量筛选

为避免多重共线性及模型过拟合对预测精度的影响,对 NPP_{max} 与环境变量之间进行相关性分析和方差分析,以筛除冗余特征,提高模型泛化能力。本研究按以下规则进行变量筛选:对于分类变量(坡位、坡向、土壤类型):使用方差分析进行变量筛选,选择对 NPP_{max} 有显著影响的变量;对于连续变量:首先选择与 NPP_{max} 相关系数绝对值大于 0.2 且具有显著性的环境变量,然后以 0.6 作为阈值,剔除环境因子之间相关系数大于 0.6 的变量。

2.2.2 机器学习模型

本研究选择随机森林模型(RF)、支持向量机模型(SVM)、增强回归树模型(BRT)和人工神经网络(ANN)4种机器学习方法进行落叶松林 NPP_{max} 模型的构建,模型的参数设置及拟合过程通过 R 语言^[13]实现。为获得稳健的预测结果,使用网格搜索法(Grid search)对模型进行调参,并利用十折交叉验证来评价模型的性能。

RF^[14]由传统的分类和回归树(CART)发展而来,利用多棵决策树的集成学习方法,具有较高的模型精度、泛化能力和不易过拟合的优点。RF 根据 Bootstrap 重采样方法生成多个不同的训练数据集,并在每个子集上构建决策树;最终在随机抽样的特征子集中选取子样本中的最优变量进行节点分割^[15]。研究中通过指定 CART 决策树的数目 ntree 进行反复训练,从而寻找估测模型的最优参数 mtry。

SVM^[16]是应用最广泛的监督类机器学习算法之一,利用线性数据的超平面映射不同数据类型的边界,并将非线性数据转换为可分割的线性数据,可用于处理分类和回归问题^[17]常用的核函数如线性、多项式、S 形核和径向基核函数等。本研究选择最优的多项式函数为核函数,通过反复训练和调试 cost、gamma 和 degree 这 3 个参数最终确定最优结果。

BRT^[18]是一种基于集成式的有监督算法。该方法是对 CART 的提升,通过整合 CART 和提升(boosting)算法,利用多轮迭代改进模型预测能力^[19]。具体方法包括通过调整 BRT 训练集的收缩参数(shrinkage)、回归树模型数量(n.trees)和树的深度(interaction.depth)等关键参数,以达到优化模型的目的。

ANN^[20]是一种受到生物神经元启发的计算模型,由输入层、隐藏层和输出层组成,可通过自适应学习来产生输入与输出之间的非线性对应关系。在训练神经网络前需要对数据进行标准化,如 z 标准化(z-normalization)、最小-最大标准化(Min-Max normalization)等。本研究使用最小-最大标准化使所有标准化后的数据落在[0,1]之间,设置输出层的激活函数为 Sigmoid 函数,将隐藏层 size 设定为 1—16,权重衰减率分别为{0,0.05,0.1,0.3}并通过反复训练该模型以获得最佳效果。

2.2.3 模型评价指标

机器学习模型的估测精度和预测能力可通过决定系数(R^2)、均方根误差(RMSE)、相对均方根误差(RRMSE)和平均绝对误差(MAE)反映。 R^2 表示模型对数据的拟合优度, R^2 越接近于 1,预测模型精度越高。RMSE、RRMSE 和 MAE 可以衡量模型预测值和实际值的误差,RMSE、RRMSE 和 MAE 值越小,模型的模拟能力越好。

$$\text{决定系数: } R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (1)$$

$$\text{均方根误差:} \quad \text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2)$$

$$\text{相对均方根误差:} \quad \text{RRMSE} = \frac{\text{RMSE}}{\bar{y}} \quad (3)$$

$$\text{平均绝对误差:} \quad \text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (4)$$

式中: y_i 为第 i 个样地 $\text{NPP}_{\max}$ 观测值 ($\text{gC m}^{-2} \text{a}^{-1}$); $\hat{y}_i$ 为第 i 个样地 $\text{NPP}_{\max}$ 预测值 ($\text{gC m}^{-2} \text{a}^{-1}$); $\bar{y}$ 为所有样地 $\text{NPP}_{\max}$ 观测值的平均值 ($\text{gC m}^{-2} \text{a}^{-1}$); n 为样地数。

2.2.4 基于 SHAP 的可解释性方法

SHAP 是基于博弈论构建的加性解释方法,通过 Shapley 值量化每个输入的特征变量对单个样本预测的边际贡献,与仅依赖于预测模型本身的特性相比,Shapley 值能更好地解释指标的变化趋势和特征的重要性^[21]。

$$y_i = y_{\text{base}} + f(x_{i,1}) + f(x_{i,2}) + \cdots + f(x_{i,j}) + \cdots + f(x_{i,k}) \quad (5)$$

式中: y_i 为第 i 个样地的 $\text{NPP}_{\max}$ 预测值; y_{base} 为所有样地的 $\text{NPP}_{\max}$ 预测平均值; x_i 为第 i 个样地的 $\text{NPP}_{\max}$ 观测值; $f(x_{i,j})$ 为 x_i 样地第 j 个特征的 SHAP 值; k 为输入特征的数量。

3 结果与分析

3.1 环境变量的筛选

分类环境变量与落叶松 $\text{NPP}_{\max}$ 方差分析如表 2 所示,不同的土壤类型 (type)、坡向 (aspect)、坡位 (position) 对 $\text{NPP}_{\max}$ 值影响均不显著 ($P > 0.05$)。连续型环境变量与落叶松 $\text{NPP}_{\max}$ 相关性分析如图 2 所示,以气候因子为例,说明变量筛选过程:气候变量中除最冷月均温 (MCMT) 与 $\text{NPP}_{\max}$ 相关性不显著且相关系数小于 0.2 外,其余变量与落叶松林 $\text{NPP}_{\max}$ 都达到极显著水平 ($P < 0.01$),其中与 $\text{NPP}_{\max}$ 相关性最大的气候变量是 18°C 以上年积温 (DD18),相关系数为 -0.61,最终筛选得到相关性系数小于 0.6 的变量为年均降水量 (MAP)。最终结果保留了 11 个环境变量作为 $\text{NPP}_{\max}$ 模型的输入变量,分别为: 18°C 以上年积温、年均降水量、全氮 (TN)、石砾体积分数 (cfvo)、有机碳储量 (OCS)、pH、阳离子交换量 (CEC)、细土粒密度 (bdod)、枯枝落叶厚度 (LT)、腐殖质厚度 (HT) 和海拔 (elevation)。

表 2 $\text{NPP}_{\max}$ 与分类环境变量的方差分析结果

Table 2 Analysis of variance between $\text{NPP}_{\max}$ and categorical environment variables

指标 Index	组间平方和 Sum of squares between classes	<i>df</i>	均方 Mean square	<i>F</i>	<i>P</i>
土壤类型 Soil type	55844.623	4	13961.156	1.932	0.106
坡向 Slope aspect	87688.011	8	10961.001	1.519	0.151
坡位 Slope position	24456.895	5	4891.379	0.662	0.652

$P < 0.05$ 代表显著, $P < 0.01$ 表示极显著

3.2 不同机器学习模型预测精度及检验

利用 RF、SVM、BRT、ANN 这 4 种机器学习算法,基于 11 个环境协变量,建立吉林省落叶松林 $\text{NPP}_{\max}$ 的估测模型,交叉验证结果如表 3 所示。不同模型的拟合 R^2 均大于 0.50,误差统计量 $\text{RMSE} < 12 \text{ gC m}^{-2} \text{a}^{-1}$, $\text{RRMSE} < 10\%$, MAE 在 $40.2 \text{ gC m}^{-2} \text{a}^{-1}$ 以内,其中,RF 预测效果最佳, R^2 为 0.540,对应的 RMSE、MAE、RRMSE 分别为 $11.320 \text{ gC m}^{-2} \text{a}^{-1}$ 、 $37.555 \text{ gC m}^{-2} \text{a}^{-1}$ 和 9.22%, BRT ($R^2 = 0.523$, $\text{RMSE} = 11.987 \text{ gC m}^{-2} \text{a}^{-1}$) 与 RF 模型的性能接近,预测结果较为可靠,ANN 和 SVM 拟合效果略差, R^2 分别为 0.519 和 0.501。总体上,4 种机器学习算法建立的 $\text{NPP}_{\max}$ 的估测模型效果排序为: RF > BRT > ANN > SVM, RF 算法估测精度最高,具有更好的稳定性。

和预测能力,因此,选择 RF 模型作为模拟研究区落叶松人工林 NPP_{max} 的最佳估测模型,并进行可解释性分析。

表 3 基于机器学习算法的估测模型精度评价(平均值±标准差)

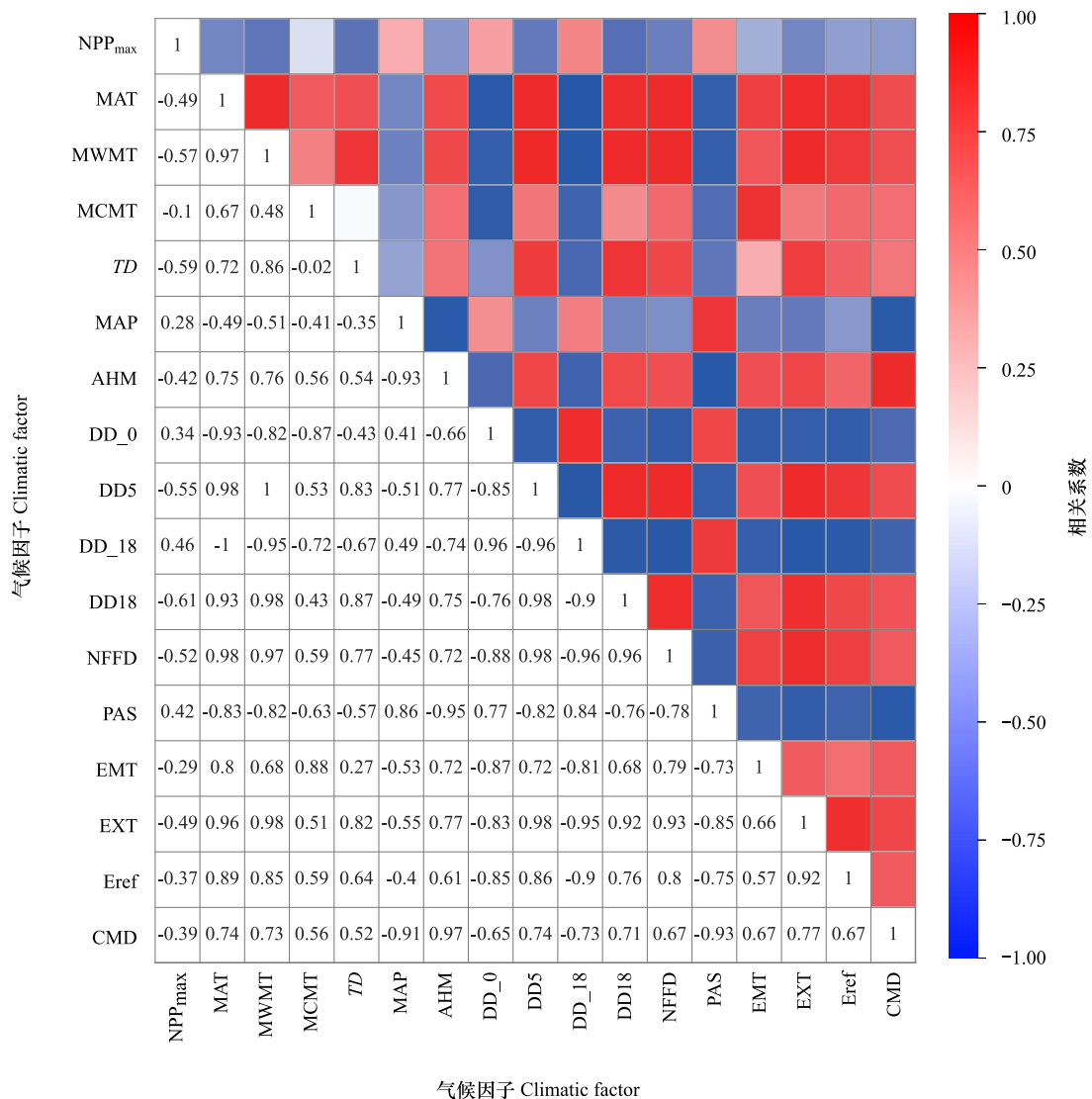
Table 3 Accuracy evaluation of estimation models based on machine learning algorithms (mean±standard deviation)

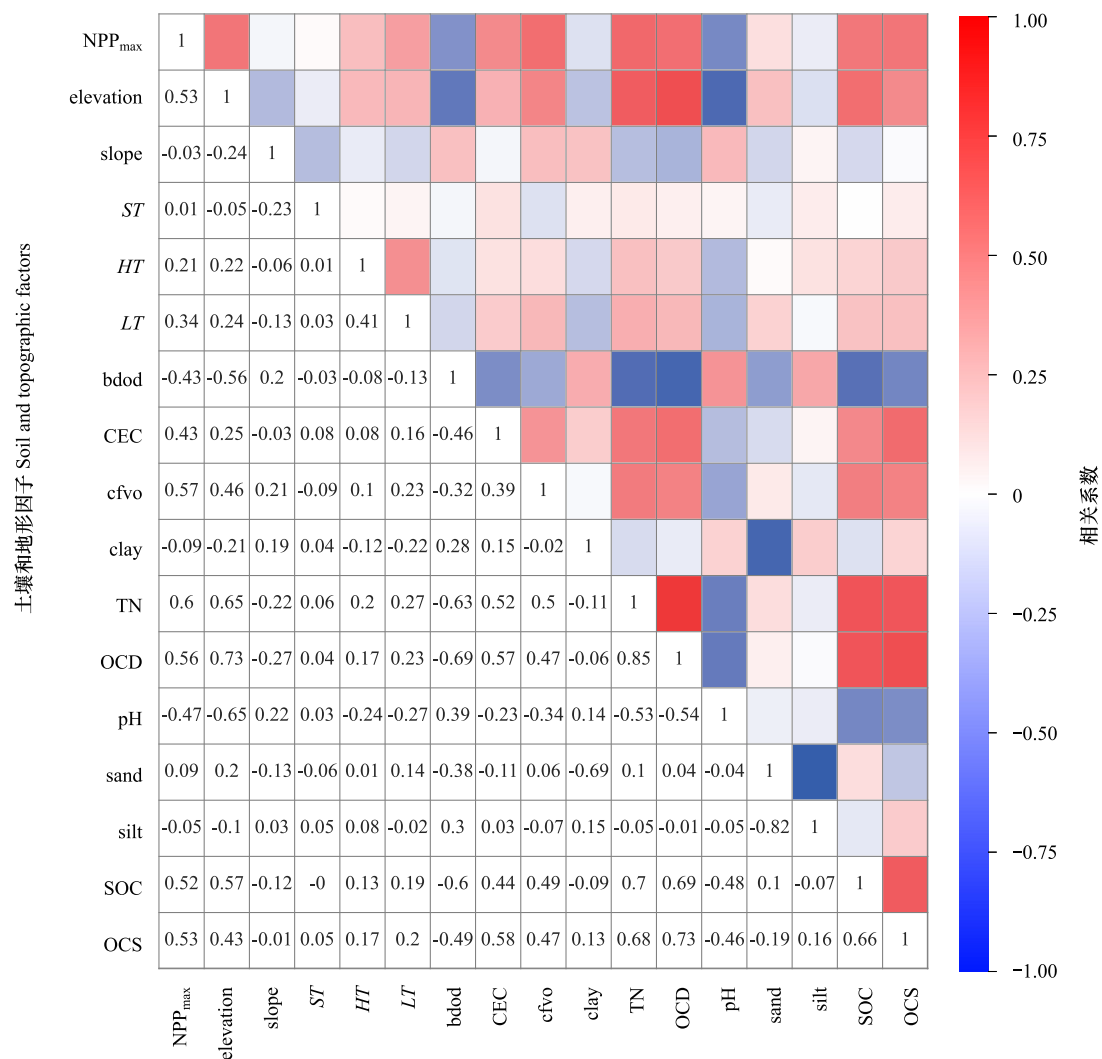
模型 Model	评价指标 Evaluation index			
	决定系数 R^2	均方根误差 RMSE/($gC\ m^{-2}\ a^{-1}$)	相对均方法误差 RRMSE/%	平均绝对误差 MAE/($gC\ m^{-2}\ a^{-1}$)
随机森林 Random forest (RF)	0.540±0.164	11.320±3.643	9.22±2.94	37.555±11.722
支持向量机 Support vector machine (SVM)	0.501±0.197	11.689±4.228	9.44±3.12	39.292±11.051
增强回归树 Boosted regression tree (BRT)	0.523±0.146	11.987±2.363	9.74±1.74	40.190±5.867
人工神经网络 Artificial Neural Network (ANN)	0.519±0.158	11.488±3.744	9.29±2.76	39.933±9.388

R^2 : 决定系数 Coefficient of determination; RMSE: 均方根误差 Root mean square error; RRMSE: 相对均方根误差 Relative root mean squared error; MAE: 平均绝对误差 Mean absolute error

3.3 基于 SHAP 值的 NPP 影响因素

选取 RF 最优模型,绘制各环境协变量对 NPP_{max} 的重要性排序柱状图,使用 SHAP 平均值作为量化指标,表示环境变量对预测结果的贡献程度。图 3 为每个环境变量对 NPP_{max} 的平均重要性,结果表明土壤变量对





土壤和地形因子 Soil and topographic factors

图2 NPP_{max}与连续性环境变量的相关系数矩阵

Fig.2 Correlation coefficient matrix between NPP_{max} and continuous environment variables

NPP_{max}: 植被净初级生产力最大值 Maximum vegetation net primary productivity; MAT: 年均温 Mean annual temperature; MWMT: 最暖月均温 Mean warmest month temperature; MCMT: 最冷月均温 Mean coldest month temperature; TD: 最暖最冷月温差 Temperature difference between MWMT and MCMT; MAP: 年均降水量 Mean annual precipitation; AHM: 湿热指数 Annual heat moisture index; DD_0: 0℃以下年积温 Degree-days below 0℃; DD5: 5℃以上年积温 Degree-days above 5℃; DD_18: 18℃以下年积温 Degree-days below 18℃; DD18: 18℃以上年积温 Degree-days above 18℃; NFFD: 无霜期天数 Number of frost-free days; PAS: 前一年8月至当年7月的降雪降水量 Precipitation as snow between August in previous year and July in current year; EMT: 超过30年的极端最低温度 Extreme minimum temperature over 30 years; EXT: 超过30年的极端最高温度 Extreme maximum temperature over 30 years; Eref: 哈格里夫斯相对蒸发 Hargreaves reference evaporation; CMD: 哈格里夫斯水分亏缺 Hargreaves climatic moisture deficit; elevation: 海拔 elevation; slope: 坡度 slope; ST: 土壤厚度 Soil thickness; HT: 腐殖质厚度 Humus thickness; LT: 枯枝落叶厚度 Litter thickness; bdod: 细土粒密度 Fine soil grain density; CEC: 阳离子交换量 Cation exchange capacity; cfvo: 石砾体积分数 Gravel volume fraction; clay: 黏粒含量 Clay content; TN: 全氮 Total nitrogen; OCD: 有机碳密度 Organic carbon density; pH: 土壤 pH Soil pH; sand: 砂粒含量 Sand content; silt: 粉粒含量 Silt content; SOC: 有机碳含量 Soil organic carbon content; OCS: 有机碳储量 Organic carbon stock

NPP_{max} 的相对贡献率最大为 63.22%, 其次是气候因子相对贡献率为 24.14%, 地形因子的相对贡献率最小为 12.64%。通过归一化处理每个变量类别的贡献率, 发现地形因子占主导地位, 其归一化后的比重为 38.76%, 其次是气候因子, 占 37.01%, 而土壤因子的相对贡献率较低, 为 24.23%。研究发现, 影响 NPP_{max} 的最大气候

变量为 18℃ 以上年积温 (DD18), 相对贡献率为 17.94%; 土壤变量中全氮对 NPP_{max} 的相对贡献率最大, 为 13.76%, 地形因子中海拔对 NPP_{max} 的相对贡献率达到 12.64%。各影响因素重要性排序由大到小依次为 DD18、TN、elevation、cfvo、OCS、pH、bdod、MAP、CEC、LT 和 HT, 其中 18℃ 以上年积温、TN、海拔是影响 NPP_{max} 的主要驱动因子。

SHAP 值能够反映各特征重要程度及正负影响, SHAP 值的零线(垂直黑线)与预测平均值(基准值)相比,明确区分了变量对 NPP_{max} 的正负驱动作用方向。如图 4 所示, 18℃ 以上年积温中的高值样本点聚焦于 SHAP 的负值, 即 18℃ 以上年积温越高, 则 NPP_{max} 值越小; 相反 18℃ 以上年积温越低, NPP_{max} 值越高。同样呈相似趋势的环境因素还有 pH、细土粒密度, 均在高值中对 NPP_{max} 的负向影响更明显。而 TN、海拔、石砾体积分数、有机碳储量和阳离子交换量, 在高值中对 NPP_{max} 的正向影响更明显。

SHAP 特征依赖图分析了影响因素的趋势和阈值效应, 可以量化各特征对模型预测结果的贡献, 并展示特征与预测结果之间的非线性关系。如图 5 所示, 当 18℃ 以上年积温大于 200℃ 时, SHAP 值由平稳状态逐渐下降, 表明年积温的增加在一定范围内对 NPP_{max} 正向贡献较大, 但随着年积温超过某一阈值后, NPP_{max} 由平稳状态逐渐减少。在年积温达到 390℃ 时 SHAP 值降至最低, 表示此时年积温对 NPP_{max} 负向影响最大, 随着年积温进一步增加, SHAP 值有所上升, 即 NPP_{max} 有所提

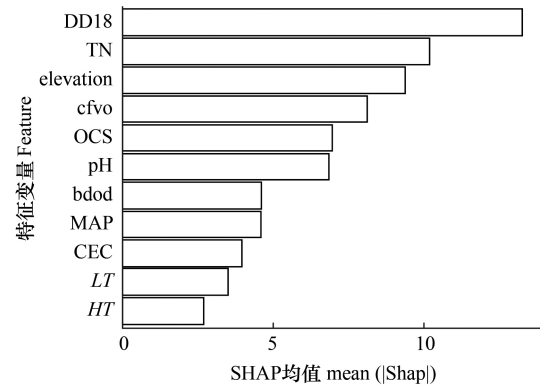


图 3 基于 SHAP 理论的环境因子重要性排序

Fig. 3 Importance ranking of environmental factors based on SHAP theory

DD18: 18℃ 以上年积温 Degree-days above 18℃; TN: 全氮 Total nitrogen; elevation: 海拔 Elevation; cfvo: 石砾体积分数 Gravel volume fraction; OCS: 有机碳储量 Organic carbon stock; pH: 土壤 pH Soil pH; bdod: 细土粒密度 Fine soil grain density; MAP: 年均降水量 Mean annual precipitation; CEC: 阳离子交换量 Cation exchange capacity; LT: 枯枝落叶厚度 Litter thickness; HT: 腐殖质厚度 Humus thickness

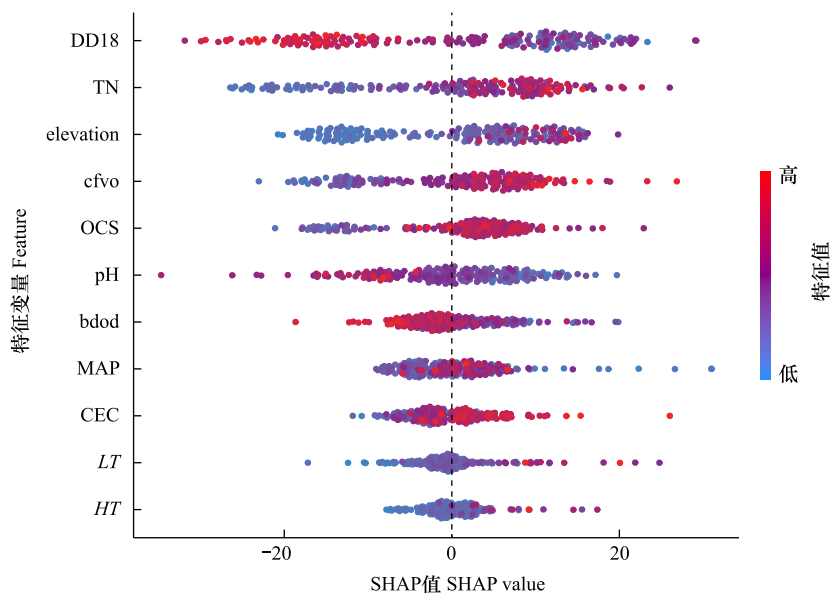


图 4 基于 RF 模型 SHAP 摘要图

Fig. 4 SHAP summary diagram based on RF model

每一个点代表一个样本, 点的颜色表示对应特征值的大小, 蓝色表示低值, 红色表示高值

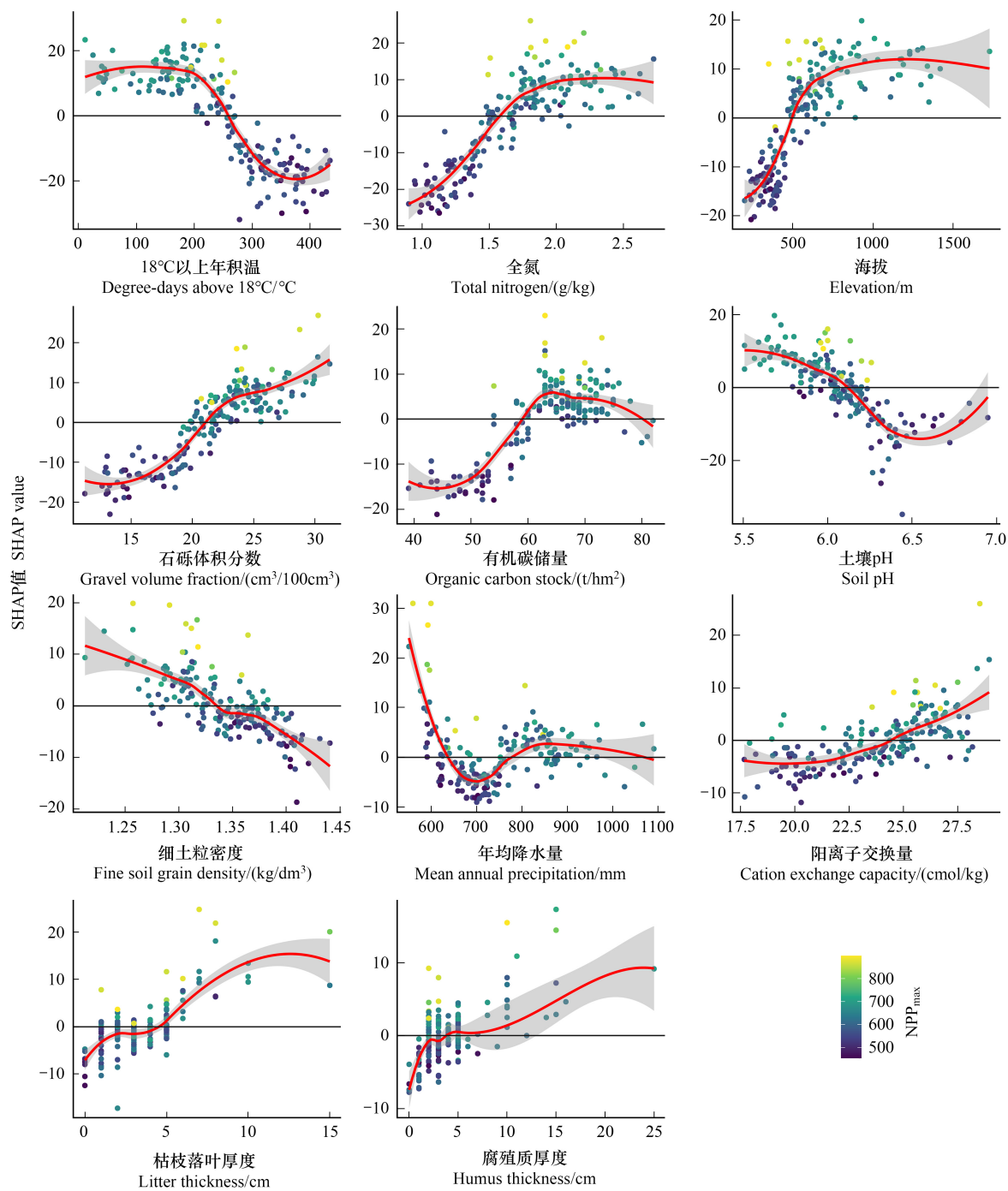


图5 NPP_{max} 影响因子的特征依赖图

Fig.5 Feature dependency graph of NPP_{max} impact factor

灰色阴影表示 SHAP 值的 95% 置信区间

高。SHAP 值随着全氮含量的增加先升高后趋于平稳,在全氮含量小于 2.0 g/kg 时,SHAP 值显著增加,即全氮含量的增加显著促进了 NPP_{max} 的升高,超过 2.0 g/kg 时 SHAP 值达到相对稳定状态,表示全氮含量的继续增加对 NPP_{max} 的贡献逐渐减少,边际效应趋于平稳。海拔、石砾体积分数、阳离子交换量、枯枝落叶厚度和腐殖质厚度对 NPP_{max} 的影响趋势与全氮含量趋势类似,均与 NPP_{max} 呈非线性正相关关系;与之相反,细土砾密度对 NPP_{max} 的影响呈非线性负相关关系。除此之外,有机碳储量、pH 和年均降水量对 NPP_{max} 的影响较为复

杂,呈现出波动趋势,不确定性较大。

4 讨论

4.1 基于机器学习算法的落叶松 NPP_{max} 估算精度比较

不同地区植被 NPP 受不同反馈机制的影响,土壤资源的变化、气候条件和地形因子共同作用于植被生长,从而解释了植被 NPP 的空间差异^[22]。本文对多个影响因子与 NPP_{max} 进行方差分析和相关性分析,发现 11 个敏感因子,分别为气候变量:18℃ 以上年积温、年均降水量;土壤因子:全氮、石砾体积分、有机碳储量、pH、细土粒密度、阳离子交换量、腐殖质厚度、枯枝落叶厚度以及地形因素海拔,这与已有研究分析的影响 NPP 的主要环境因子基本一致^[22-24]。本研究采用多种机器学习方法对 NPP_{max} 进行反演,发现 RF 模型的预测精度优于其他三种模型。RF 模型可进行高维非线性数据建模,具有模型精度较高、泛化能力和鲁棒性较强的优点^[25]。先前的研究发现,RF 模型对森林 NPP 预测精度优于其他机器学习模型和传统线性回归模型^[26-27],使用 RF 模型的预测结果能够真实反映落叶松潜在 NPP 实际情况^[28]。本研究构建了 4 种机器学习模型,依据 RMSE 最小的原则筛选出的模型其 R^2 为 0.540,比基于 CART 决策树算法发现的气候因素解释 65% 的 NPP_{max} ($R^2=0.65$)^[8] 低。产生这种差异的原因可能是存在数据量、数据分布和模型自身等因素导致的不确定性,各模型敏感性和模型解释量有待提高^[17]。另外,其他影响落叶松林 NPP_{max} 的因素仍需进一步研究,从而进一步探讨机器学习方法在 NPP_{max} 预测中的有效性。

4.2 影响落叶松 NPP_{max} 的关键因子

植被 NPP_{max} 驱动因素比较复杂,不仅受植被本身生理特性制约,还受到气候、地形、土壤及人类活动直接或间接的影响^[29-30]。在气候变量中,18℃ 以上年积温是落叶松林 NPP_{max} 变化的最重要驱动因素,可以显著影响树冠层的光合效率和蒸腾作用促进落叶松生长,但过高的年积温 ($>200^{\circ}\text{C}$) 引起的水分胁迫和矿物质吸收困难会降低光合效率从而使落叶松林 NPP_{max} 下降,这与 Li 等^[31] 的研究结果相同。 NPP_{max} 与年均降水量呈非线性关系,在年均降水量的影响下, NPP_{max} 呈先下降、后增加、再次下降,这可能是随着水分有效性的增加,年均降水量对落叶松林 NPP_{max} 的正向作用逐渐减弱所致,与 He 等^[32] 研究结果一致。

土壤是植物根系生长的基质,为植物提供所必需的生长介质、水分和养分等,土壤性状对植物的正负反馈效应影响植被 NPP_{max} 的变化。在土壤变量中,对长白落叶松林 NPP_{max} 预测结果贡献最大的因素是全氮,接着依次是石砾体积分、有机碳储量、pH、细土粒密度、阳离子交换量、枯枝落叶厚度和腐殖质厚度。全氮含量的增加能有效提高植物叶片的光合作用,增加植物碳固定,促进植物的生长^[33]。本研究发现 TN 含量与 NPP_{max} 之间存在阈值效应,即引起植被 NPP_{max} 变化的范围区间相对固定,更多的氮输入 ($>2.0\text{ g/kg}$) 并不会引起落叶松林 NPP_{max} 的增加,土壤中多余的氮素以 N_2O 和 NO_3^- 的形式排出^[34]。石砾体积分的增加可以改善土壤结构及水分状况,有利于植物生长^[35]。枯枝落叶厚度和腐殖质厚度均对落叶松林 NPP_{max} 起正向影响,通过增加土壤有机质含量改善土壤质地,使植物有更好的生长环境,从而促进 NPP_{max} 增加。 NPP_{max} 与细土粒密度呈负相关,与前人的研究一致^[36],其原因可能是细土粒密度增加会降低土壤孔隙度,使土壤紧实,土壤通透性和水源涵养的能力降低,直接影响植物的正常生长发育。阳离子交换量反映土壤保肥供肥和缓冲阳离子 (K^+ 、 NH_4^+) 的能力,落叶松林 NPP_{max} 与阳离子交换量呈正相关。本研究中 NPP_{max} 与有机碳储量、pH 均存在拐点,即达到某个值时 NPP_{max} 改变上升或下降的趋势,与 NPP_{max} 存在不确定性,对 NPP_{max} 的影响还有待进一步确认。地形因子中,海拔通过直接影响研究区水热条件和土壤肥力,从而影响落叶松林 NPP_{max} 的变化^[37],本研究中海拔达到约 800 m 后 NPP_{max} 基本稳定。

确定环境变量对 NPP_{max} 影响对于改善森林生态系统对全球气候变化的响应及其服务和功能的维持至关重要。SHAP 分析表明, NPP_{max} 与 18℃ 以上年积温、全氮含量、海拔之间存在非线性关系,且具有阈值效应。因此,有必要加强对吉林省落叶松林生态系统的监测,为应对气候变化提供科学依据。未来的研究需进一步探索阈值效应的机理和过程。在不同空间尺度下,阈值条件可能会相应变化,这需要通过多尺度分析进行验证。

4.3 不足与展望

本研究选取了 2000—2022 年间的 NPP 最大值作为长白落叶松林的 NPP_{max} , 然而, 由于森林 NPP 的监测通常依赖于长期数据的支持, 而短时间序列的使用限制了对森林 NPP 变化规律的深入探讨。未来研究应重点关注长时序数据的收集与分析, 以便更全面地揭示落叶松林 NPP_{max} 的变化趋势及其对环境因子的长期响应机制。鉴于实地调查的难度及人类活动难以量化等因素, 本研究主要考虑了气候、土壤和地形等环境因子的影响, 尚未涉及人类活动和生物多样性等因素。后续应加强对多因素共同作用的探讨, 构建更为完善的 NPP 综合评估体系。

5 结论

(1) RF、BRT、ANN、SVM 4 种机器学习算法对落叶松 NPP_{max} 预测表现良好, 其中 RF 模型精度最高, R^2 达到 0.540, 对应的 RMSE 值最小为 $11.320 \text{ gC m}^{-2} \text{ a}^{-1}$, 其次为 BRT、ANN、SVM 模型。

(2) SHAP 可视化分析表明, 18°C 以上年积温、土壤全氮、海拔是影响落叶松 NPP_{max} 的主要因子, 研究进一步揭示了 NPP_{max} 与多个环境变量之间的非线性关系: 全氮、海拔、石砾体积分数、阳离子交换量、枯枝落叶厚度和腐殖质厚度与落叶松 NPP_{max} 呈非线性正相关, 18°C 以上年积温、细土粒密度与落叶松 NPP_{max} 呈非线性负相关。

(3) 年积温、全氮含量和海拔对 NPP_{max} 表现出明显的阈值效应。当 18°C 以上年积温接近 200°C 、全氮含量达到 2.0 g/kg 以及海拔约为 800 m 时, NPP_{max} 的增减变化明显。

本研究通过精确识别年积温、全氮和海拔等关键因素的作用, 揭示了长白落叶松林在特定环境条件下的生长潜力与资源利用效率, 不仅为立地质量评价提供了科学依据, 也有助于提升森林碳汇功能和森林生态系统的可持续性。

参考文献 (References):

- [1] 赵芳, 张久阳, 刘思远, 王增艳, 王黎欢, 顾浩婷, 李万隆. 秦巴山地 NPP 及对气候变化响应的多维地带性与暖温带-亚热带界线. 生态学报, 2021, 41(1): 57-68.
- [2] Yuan Q Z, Wu S H, Dai E F, Zhao D S, Ren P, Zhang X R. NPP vulnerability of the potential vegetation of China to climate change in the past and future. Journal of Geographical Sciences, 2017, 27(2): 131-142.
- [3] Xi Z L, Chen G Z, Xing Y Q, Xu H D, Tian Z Y, Ma Y, Cui J, Li D. Spatial and temporal variation of vegetation NPP and analysis of influencing factors in Heilongjiang Province, China. Ecological Indicators, 2023, 154: 110798.
- [4] 闫丰, 李晨阳, 王靖, 鲁志雪, 黄雪寒, 王敏丽, 王文迪, 李若茜, 庞娇, 陈亚恒. 京津冀生态屏障区植被 NPP 时空演变驱动力定量解析. 环境科学, 2025, 46(1): 327-339.
- [5] 何颖, 马松梅, 张林, 张云玲, 贺凌云. 玛纳斯河流域 NPP 时空变化及其生物多样性维护功能重要性评价. 生态学报, 2023, 43(11): 4664-4673.
- [6] Yi Z Q, Wu L H. Identification of factors influencing net primary productivity of terrestrial ecosystems based on interpretable machine learning—evidence from the county-level administrative districts in China. Journal of Environmental Management, 2023, 326: 116798.
- [7] Yu B, Chen F, Chen H Y. NPP estimation using random forest and impact feature variable importance analysis. Journal of Spatial Science, 2019, 64(1): 173-192.
- [8] Zhou B H, Liao Z Z, Chen S R, Jia H Y, Zhu J Y, Fei X H. Net primary productivity of forest ecosystems in the southwest Karst region from the perspective of carbon neutralization. Forests, 2022, 13(9): 1367.
- [9] Cao D, Zhang J H, Xun L, Yang S S, Wang J W, Yao F M. Spatiotemporal variations of global terrestrial vegetation climate potential productivity under climate change. Science of the Total Environment, 2021, 770: 145320.
- [10] 李芸, 王铁夫, 孙玉军, 雷渊才, 邵卫才, 李杰. 吉林省落叶松林净初级生产力时空特征及其对气候变化的响应. 生态学报, 2022, 42(3): 947-959.
- [11] He X, Lei X D, Liu D, Lei Y C. Developing machine learning models with multiple environmental data to predict stand biomass in natural coniferous-broad leaved mixed forests in Jilin Province of China. Computers and Electronics in Agriculture, 2023, 212: 108162.

- [12] Wang T L, Wang G Y, Innes J L, Seely B, Chen B Z. ClimateAP: an application for dynamic local downscaling of historical and future climate data in Asia Pacific. *Frontiers of Agricultural Science and Engineering*, 2017, 4(4): 448.
- [13] R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2022, [2024-12-19]. URL <https://www.R-project.org/>.
- [14] Breiman L. Random forests. *Machine Learning*, 2001, 45: 5-32.
- [15] Shi T Z, Hu X J, Guo L, Su F Z, Tu W, Hu Z W, Liu H Z, Yang C, Wang J Z, Zhang J, Wu G F. Digital mapping of zinc in urban topsoil using multisource geospatial data and random forest. *Science of the Total Environment*, 2021, 792: 148455.
- [16] 吴喜之. 应用回归及分类: 基于 R. 北京: 中国人民大学出版社, 2016.
- [17] 沈琛琛, 肖文发, 朱建华, 曾立雄, 陈吉臻, 黄志霖. 基于机器学习算法的华中天然林土壤有机碳特征与关键影响因子. *林业科学*, 2024, 60(3): 65-77.
- [18] De'ath G. Boosted trees for ecological modeling and prediction. *Ecology*, 2007, 88(1): 243-251.
- [19] 雷相东. 机器学习算法在森林生长收获预估中的应用. *北京林业大学学报*, 2019, 41(12): 23-36.
- [20] 王星. 大数据分析: 方法与应用. 北京: 清华大学出版社, 2013.
- [21] Aas K, Jullum M, Løland A. Explaining individual predictions when features are dependent: more accurate approximations to shapley values. *Artificial Intelligence*, 2021, 298: 103502.
- [22] Sun H Z, Chen Y B, Xiong J N, Ye C C, Yong Z W, Wang Y, He D, Xu S C. Relationships between climate change, phenology, edaphic factors, and net primary productivity across the Tibetan Plateau. *International Journal of Applied Earth Observation and Geoinformation*, 2022, 107: 102708.
- [23] 刘欢欢, 陈印, 刘悦, 刚成诚. 基于随机森林模型的黄土高原草地净初级生产力时空格局及未来演变趋势模拟. *干旱区研究*, 2023, 40(1): 123-131.
- [24] Wei X D, Yang J, Luo P P, Lin L G, Lin K L, Guan J M. Assessment of the variation and influencing factors of vegetation NPP and carbon sink capacity under different natural conditions. *Ecological Indicators*, 2022, 138: 108834.
- [25] 仇皓雷, 王海燕. 机器学习在土壤性质预测研究中的应用进展. *生态学杂志*, 2025, 44(1): 283-294.
- [26] Xu M, Hu C M, Najjar R G, Herrmann M, Briceno H, Barnes B B, Roger Johansson J O, English D. Estimating estuarine primary production using satellite data and machine learning. *International Journal of Applied Earth Observation and Geoinformation*, 2022, 110: 102821.
- [27] Li T, Li M Y, Ren F, Tian L. Estimation and spatio-temporal change analysis of NPP in subtropical forests: a case study of Shaoguan, Guangdong, China. *Remote Sensing*, 2022, 14(11): 2541.
- [28] Cheng R M, Zhang J, Wang X Y, Zhang Z D. Growth suitability evaluation of *Larix principis-rupprechtii* Mayr based on potential NPP under different climate scenarios. *Sustainability*, 2023, 15(1): 331.
- [29] 李金珂, 杨玉婷, 张会茹, 黄铝文, 高义民. 秦巴山区近 15 年植被 NPP 时空演变特征及自然与人为因子解析. *生态学报*, 2019, 39(22): 8504-8515.
- [30] 王川, 王丽莎, 张勇勇, 赵文智, 冯相艳. 2000—2020 年祁连山植被净初级生产力时空变化及其驱动因素. *生态学报*, 2023, 43(23): 9710-9720.
- [31] Li L Y, Zeng Z Z, Zhang G, Duan K, Liu B J, Cai X T. Exploring the individualized effect of climatic drivers on modis net primary productivity through an explainable machine learning framework. *Remote Sensing*, 2022, 14(17): 4401.
- [32] He X, Lei X D, Zeng W S, Feng L Y, Zhou C F, Wu B Y. Quantifying the effects of stand and climate variables on biomass of larch plantations using random forests and national forest inventory data in North and Northeast China. *Sustainability*, 2022, 14(9): 5580.
- [33] 郭洁芸, 王雅歆, 李建龙. 氮添加对中国陆地生态系统植物-土壤碳动态的影响. *生态学报*, 2022, 42(12): 4823-4833.
- [34] 祝元丽, 冯向阳, 闫庆武, 吴子豪. 基于 CGBDT 的望奎县农田土壤有机碳主控因子研究. *中国环境科学*, 2024, 44(3): 1407-1417.
- [35] 胡慧, 包维楷, 张强, 李芳兰. 土壤石砾含量变化对根系及植物生长的影响研究进展. *应用与环境生物学报*, 2021, 27(2): 303-308.
- [36] 杜志勇, 丛楠. 植被与土壤特征对青藏高原不同程度退化草地的响应. *生态学报*, 2024, 44(6): 2504-2516.
- [37] 肖晶, 饶良懿. 2001—2020 年乌梁素河流域植被 NPP 时空变化及驱动因素分析. *环境科学*, 2024, 45(8): 4744-4755.