

# 应用哨兵2号卫星遥感影像数据和机器学习算法对锡林郭勒草原土壤表层有机碳及全氮的估算<sup>1)</sup>

杨珺婷 李晓松

(中国科学院空天信息创新研究院(中国科学院大学),北京,100094)

**摘要** 以锡林郭勒草原为研究区,应用谷歌地球引擎(GEE)云计算平台、哨兵2号卫星(Sentinel-2)遥感数据及其他辅助数据,选择随机森林(RF)、支持向量机(SVM)、多层感知器(MLP)模型3种机器学习算法,对土壤表层(0~20 cm)30 m分辨率的土壤有机碳(SOC)和土壤总氮(STN)质量分数进行估算,比较不同方法的精度差异,分析各影响因素的重要性。结果表明:①3种机器学习算法中,随机森林模型对土壤有机碳、土壤总氮估测效果最佳(土壤总氮模型,决定系数为0.67、均方根误差为0.024、百分比偏差为2.09、四分位数间距性能比(RPIQ)为2.14;土壤有机碳估测模型,决定系数为0.68、均方根误差为0.17、百分比偏差为2.26、四分位数间距性能比为2.37)。②各影响因素重要性评价结果表明,哨兵2号卫星遥感数据、植被指数,对土壤有机碳、土壤总氮估算贡献较大;在土壤有机碳估测模型中,光谱反射率信息贡献率为30.9%,植被指数贡献率为28.4%;在土壤总氮估测模型中,光谱反射率信息贡献率为23.2%,植被指数贡献率为46.5%。③锡林郭勒草原,土壤有机碳、土壤总氮质量分数整体偏低,空间分布趋势为自西向东逐渐增加。

**关键词** 土壤有机碳;土壤全氮;数字土壤制图;锡林郭勒草原

**分类号** S151.9

**Estimation of Topsoil Organic Carbon and Total Nitrogen in Xilin Gol Grassland Using Sentinel-2 and Machine Learning**//Yang Junting, Li Xiaosong(Aerospace Information Research Institute, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Beijing 100094, P. R. China)//Journal of Northeast Forestry University, 2022, 50(1): 64-71.

Taking Xilin Gol Grassland as the study area, we estimated soil organic carbon (SOC) and soil total nitrogen (STN) mass fraction of soil surface layer (0-20 cm) with 30 m resolution from the Google Earth Engine (GEE) cloud computing platform by employing Sentinel-2 data and ancillary factors with random forest (RF), support vector machine (SVM) and multi-layer perceptron (MLP) three kinds of machine learning algorithms, and compare the results of different methods. The results show: (1) Among the three machine learning algorithms, the random forest (RF) model has the best accuracy both on the estimation content of soil organic carbon ( $R^2$  of 0.68, the root mean square error of 0.17, the ratio of prediction deviation of 2.26, ratio of performance to interquartile distance of 2.37) and soil total nitrogen ( $R^2$  of 0.67, the root mean square error of 0.024, the ratio of prediction deviation of 2.09, and the ratio of performance to interquartile distance of 2.14). (2) The variable importance evaluation results show that the remote sensing data and vegetation index of Sentinel-2 had a large contribution to the estimation of the mass fraction of soil organic carbon and soil total nitrogen; for the soil organic carbon estimation model, the contribution rate of spectral reflectance information was 30.9%, and that of vegetation index was 28.4%; while in the soil total nitrogen estimation model, that of spectral reflectance information was 23.2%, and that of vegetation index was 46.5%. (3) In Xilin Gol Grassland, the contents of soil organic carbon and soil total nitrogen were generally low, and the spatial distribution trend was consistent with the trend of gradual increase from west to east.

**Keywords** Soil organic carbon; Soil total nitrogen; Digital soil mapping; Xilin Gol Grassland

DOI:10.13759/j.cnki.dlxb.2022.01.022

土壤有机碳(SOC)、土壤总氮(STN)是植物生长的必需养分,同时也是全球碳和氮循环的主要组成部分<sup>[1-3]</sup>。土壤有机碳、土壤总氮空间分布的高精度制图,对于准确评估土壤质量、精准农业和生态环境评价非常重要。锡林郭勒草原作为国家重要生态屏障和畜产品基地,草原退化和沙漠化面积已占全盟草原面积72.22%,在锡林郭勒盟草地开展土壤有机碳和土壤全氮的高分辨率制图对支撑退化监

控、修复工作非常重要<sup>[4-5]</sup>。

传统的方法预测土壤有机碳、土壤总氮的空间分布,通过大量的地面调查数据,根据土壤类型或土地利用类型分类统计后,将土壤有机碳、土壤总氮储量平均分配给每个土壤类型或土地利用类型的地图单位<sup>[6]</sup>。但是,受到采样频率的限制,昂贵且耗时,无法显示每个地图单元中土壤有机碳、土壤总氮的较大空间异质性和时间动态变化。徐彬彬等<sup>[7]</sup>通过分析南疆土壤的光谱反射特性与有机质含量之间的关系,发现土壤有机质含量与土壤反射光谱存在一定的相关关系,后续研究又进一步发现整个近红外(700~2500 nm)、可见光(400~700 nm)波长区域均受土壤有机质含量的控制,利用土壤反射光谱逐渐成为常用的土壤属性制图方法<sup>[8-10]</sup>。随着遥感数

1) 国家重点研发计划(2016YFC0500806)。

第一作者简介:杨珺婷,女,1996年3月生,中国科学院空天信息创新研究院(中国科学院大学),硕士研究生。E-mail: yangjt@aircas.ac.cn。

通信作者:李晓松,中国科学院空天信息创新研究院(中国科学院大学),研究员。E-mail: lixs@aircas.ac.cn。

收稿日期:2021年3月26日。

责任编辑:张玉。

据的时空分辨率不断提高,数据获取更容易,各种卫星传感器已成功应用于不同规模的数字土壤制图,其中最常用的是陆地卫星(Landsat)、中分辨率成像光谱仪(MODIS)<sup>[10-13]</sup>。例如,贾伟等<sup>[10]</sup>利用陆地卫星 8 号上携带陆地成像仪(Landsat 8 OLI)光谱数据,成功反演了三江源地区的土壤氮含量。研究证明,分辨率较高的传感器很适合捕获土壤特性的微小空间变化,且遥感的土壤制图取决于遥感图像的可用性和质量<sup>[14]</sup>。但是,中分辨率成像光谱仪数据只能提供较低的分辨率,而陆地卫星尽管具有较高的空间分辨率,但 16 d 的重访周期却增加了选择无云影像的难度哨兵 2 号卫星(Sentinel-2)数据具有很高空间分辨率(10~30 m)和能为研究提供足够数据量的 5 d 重访周期,已经在土壤属性估算中初步展现出充分的潜力<sup>[6,15-16]</sup>,但是在国内的相关应用还很少。

土壤的形成过程复杂,地表信息难以分辨,完全依据遥感数据估算土壤属性有很大不确定性,所以环境数据(气候、地形、母质等)常和遥感数据一起使用,作为估算土壤属性的指标。McBratney et al.<sup>[17]</sup>根据土壤发生理论归纳总结出了土壤方程(SCORPAN),将生物、母质、地形、气候、其他土壤属性相关的指标,通过样点数据,利用数学模型建模估算特定土壤属性的方法定义为数字土壤制图,并成为现在主要的土壤制图方法。当前,从简单的线性统计模型,地统计学到高级和复杂的机器学习技术(随机森林、神经网络、分类与回归树等)等数字土壤制图方法,已被用于绘制土壤有机碳的空间分布图<sup>[10,13]</sup>。在众多建模方法中,随机森林(RF)<sup>[18]</sup>是一种新的数据挖掘方法,能够建模高维非线性关系,处理分类和预测评价目标,避免过拟合,具有较强的鲁棒性,可提供变量重要性评价且参数少,成为很多研究选择这种方法的原因<sup>[19-20]</sup>。也有一些研究证明,依据结构风险最小化发展的支持向量机(SVM)方法,可以为土壤属性估测提供更好的支撑<sup>[21]</sup>。带有反向传播算法的多层感知器神经网络(MLP 神经网络),因为极强的学习能力,也被成功应用到了土壤属性制图中<sup>[22-23]</sup>。在大量数字土壤制图研究中,对机器学习方法的选择各有不同,同时对不同机器学习方法的对比研究也没有定论。

为此,本研究以锡林郭勒草原为研究区,应用谷歌地球引擎(GEE)云计算平台、哨兵 2 号卫星遥感数据及其他辅助数据,选择随机森林、支持向量机、多层感知器人工神经网络模型 3 种机器学习算法,对土壤表层(0~20 cm) 30 m 分辨率的土壤有机碳和土壤总氮质量分数进行估算,比较不同方法的精度

差异,分析各影响因素的重要性。旨在为生产高精度高空间分辨率的锡林郭勒草地土壤有机碳、土壤总氮地图、掌握其空间分布、监控土地变化提供参考。

## 1 研究区概况

锡林郭勒盟位于内蒙古的东部北侧,地理位置是北纬 42°32′~46°41′、东经 111°59′~120°。根据 30 m 分辨率的 2010 年中国国家土地覆盖数据集,锡林郭勒盟土地覆盖类型,被分为草地、湿地、耕地、林地、人工表面和其他,通过裁剪确定了研究区内草地地区作为本研究区。锡林郭勒盟草地东界在大兴安岭西麓,南界在阴山的北麓,西北界为中蒙边界山脉,由 50 多片面积几百到几千平方千米的平地小草原以及夹于其间的 20 多片丘陵小草原组成。锡林郭勒盟草地是我国类型复杂、生物多样性丰富,在温带草地中具有代表性的草地,是隔断蒙古沙源与京津冀区域的重要植被沙障<sup>[24]</sup>,总面积 20.3×10<sup>4</sup> km<sup>2</sup>。研究区属于大陆性半干旱草原气候,四季分明,年均温-0.4℃,最冷月份(1 月份)平均气温-22.3℃,最热月份(7 月份)平均气温 18.8℃,气温由西南到东北递减,全年降水量 200~400 mm,降水量由西向东递增。

草地类型空间分布与水热条件密切相关,具有明显的水平地带性。由西南到东北依次为:以小针茅(*Stipa klemenzi*)建群的荒漠草原;以杂类草相对较少,以大针茅(*Stipa grandis*)、克氏针茅(*Stipa krylouii*)、糙隐子草(*Cleistogenes squarrosa*)建群的典型草原;杂类草比例较大,主要以线叶菊(*Filifolium sibiricum*)、贝加尔针茅(*Stipa baicalensis*)、羊草(*Leymus chinensis*)为建群种的草甸草原;同时中西部及南部沙质地段,还广泛分布着小叶锦鸡儿(*Caragana microphylla*)灌丛;发育在沟谷洼地及河流的芨芨草(*Achnatherum splendens*)草甸;分布在湖泊周围的盐化湿地上的以红砂(*Reaumuria songarica*)、盐爪爪(*Kalidium foliatum*)为主要植被的盐化低地草甸。土壤类型,以黑钙土、栗钙土、风化土、棕钙土为主。

## 2 研究方法

### 2.1 样点数据

2019 年 8 月份在锡林郭勒盟草地野外采样,共计采集土壤样点 140 个(见图 1),从东到西覆盖了研究区内的不同草地类型,并充分考虑了关于草地退化、土壤类型等信息。样方尺寸为 30 m×30 m,土壤剖面深度为 0~20 cm,水平和垂直方向上每隔 10 m 采集 1 次样品,共 9 个样品。将 9 个土壤样品充

分混合,在实验室进行土壤属性分析。每个样本的地理坐标是用手持设备记录的全球定位系统(GPS)匹配遥感图像(Trimble GPS,精度<1 m)。将采集的土样带回实验室自然风干、研磨并剔除土样中的杂质后过筛(筛孔直径2 mm)。采用重铬酸钾容量-外加热法(油浴)测定土壤有机碳质量分数,土壤全氮质量分数选择凯氏定氮法测定。

## 2.2 数据来源与处理

依据土壤方程<sup>[17]</sup>和相关研究成果<sup>[25-30]</sup>,本研究选择哨兵2号卫星反射率数据、归一化植被指数(NDVI)、干枯燃料指数(DFI)、红边叶绿素指数(CI)、倒红边叶绿素指数(IRECI)、高程、坡度、坡向、地形湿度指数(TWI)及年均温、年均降水作为环境评

价指标参与建模。具体数据来源及数据处理见表1。

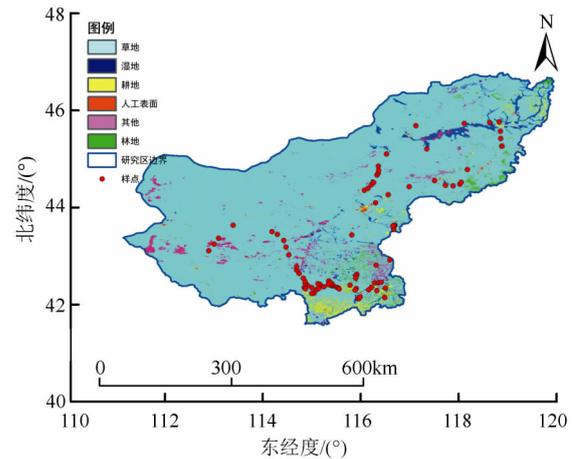


图1 研究区采样点设置

表1 数据来源及处理

数据集	数据ID/下载地址	获取时间	遥感数据处理及获取
Sentinel-2 地表反射率1级产品	COPERNICUS/S2_SR	2019年5月	影像筛选、去云、归一化植被指数最大值合成、裁剪、掩膜、提取植被指数和波段信息
SRTM 30 m 分辨率数字高程数据	USGS/SRTMGL1_003	2000年	提取高程,进行地形分析得到坡度、坡向
地形湿度指数	http://doi.org/10.5281/zenodo.1447210	2017年	
气象数据	ECMWF/ERA5/DAILY	2010—2019年	均值处理

### 2.2.1 遥感数据与预处理

遥感影像数据源选择高时空分辨率的哨兵2号卫星数据,最高空间分辨率可达10 m,重访周期为5 d。影像来自于谷歌地球引擎上的天顶反射率数据集,影像已经进行了辐射定标与地形校正。在谷歌地球引擎平台通过时间筛选函数进行选择2019年7—8月份生长季影像数据,用来提取地面反射率信息以及光谱指数。所有的影像利用质量控制波段(QA)进行去云处理,得到云量低于20%的影像数据后,再对影像进行归一化植被指数最大值合成,并根据研究区范围进行裁剪。选择常用的哨兵2号卫星的波段2~9和波段11、12,并计算常用的光谱指数,包括归一化植被指数、干枯燃料指数、红边叶绿素指数、倒红边叶绿素指数。

归一化植被指数<sup>[31]</sup>和几个红边指数,是植被覆盖度常用且易于计算的依据卫星图像的指示指标,也经常在土壤制图中作为生物指标被选择为评价指标参与建模<sup>[25,27,32]</sup>。干枯燃料指数,是Cao et al.<sup>[33]</sup>提出的一种依据中分辨率成像光谱仪生产的可以代表地表非光合植被信息的植被指数,被认为可以很好地反应地表非光合植被的覆盖情况,目前还没有被应用到土壤属性研究中。但是,根据Wang et al.<sup>[19]</sup>在澳大利亚的干旱牧场研究发现,非光合植被信息对土壤碳含量的估测起了至关重要的作用。因此,本研究将哨兵2号卫星波段与中分辨率成像光谱仪波段信息进行对应后,计算并引入了干枯燃料指数作为评价指标之一,拟提高模型的估测能力。

$$I_{NV} = (N_{IR} - R_{ed}) / (N_{IR} + R_{ed});$$

$$I_{DF} = 100(1 - I_{SWR1} / I_{SWR2}) / (R_{ed} / N_{IR});$$

$$I_C = (R_{ed,E3} / R_{ed,E1}) - 1;$$

$$I_{IREC} = (R_{ed,E3} - R_{ed,E2}) R_{ed,E2} / R_{ed,E1}。$$

式中: $I_{NV}$ 为归一化植被指数; $I_{DF}$ 为干枯燃料指数; $I_C$ 为红边叶绿素指数; $I_{IREC}$ 为倒红边叶绿素指数; $R_{ed}$ 、 $R_{ed,E1}$ 、 $R_{ed,E2}$ 、 $R_{ed,E3}$ 、 $N_{IR}$ 、 $I_{SWR1}$ 、 $I_{SWR2}$ 分别对应哨兵2号卫星数据的红光、红边和近红外波段的4、5、6、7、8、12、11。

### 2.2.2 辅助数据

土壤属性分布水平受环境因素及其复杂相互作用的影响。通常,气候决定了土壤属性分布的格局<sup>[13,34]</sup>。在这项研究中,从谷歌地球引擎获取了欧洲中期天气预报中心(ECMWF)对全球气候的第五代大气再分析数据(ERA5)产品2010—2019年10 a的地面2 m处的日均温、日均降水量信息,并计算10 a年均温、年均降水,作为气候评价指标参与模型训练。

在复杂的地形中,通常会随着高度、坡度、坡向、地表汇水能力的变化而观察到土壤属性的变化<sup>[25,35-36]</sup>,这些物理特征也已被用于预测土壤属性的空间分布<sup>[29,37-38]</sup>。本研究选择了高程、坡度、坡向、地形湿度指数作为地形因子。从谷歌地球引擎平台在线数据获取了航天飞机雷达地形测绘(SRTM V3)数据产品<sup>[39]</sup>,该数据产品由美国国家航空航天局(NASA)、美国国家地理空间情报局以及德国和意大利航天局共同生产,分辨率达到了

30 m。根据高程数据,通过谷歌地球引擎地形分析命令计算坡度、坡向;地形湿度指数数据,来自于 Hengl<sup>[40]</sup> 利用自动化地球科学分析系统(SAGA)生产的全球地形数据集。

### 2.3 模型选择

**随机森林:** 随机森林<sup>[18]</sup> 是一种可用于分类和回归的集成机器学习方法,在训练时构造大量决策树,每棵树都依赖于独立采样的随机向量的值;然后将这些决策树聚合在一起,对数据集中的预测目标(土壤属性)给出一个单独的预测;最后预测结果是所有单棵树输出的平均值<sup>[41]</sup>。模型以其在处理多元非线性数据方面的优势,而被尝试运用到土壤属性的预测之中<sup>[20]</sup>,在多项土壤属性研究中得到了非常好的模型表现<sup>[19-20,38]</sup>。且模型运行时袋装(OOB)样本可用于内部交叉验证准确性和辅助数据重要性评估<sup>[23]</sup>,这提供了对每个输入辅助数据对预测过程贡献的定量测量,即所谓的随机森林变量重要性。模型参数较少,只有子树个数、树深、最多叶子数等参数,使用方便。本研究的随机森林模型的实现与模型参数优化,均在谷歌地球引擎平台通过调用应用程序接口(API)完成。

**支持向量机:** 支持向量机是由 Cortes et al.<sup>[42]</sup> 提出的常用的机器学习方法,已经被广泛应用于土地覆盖分类与土壤属性估测建模中,并且表现出了很好的估测效果。支持向量机方法原理,主要是通过核函数将数据映射到高维特征空间,然后依据结构化风险最小化原则,在新的空间中构建一个最佳的超平面。支持向量机共有 4 种核函数类型,线性函数、多项式函数、双曲线正切函数函数、径向基函数(RBF)<sup>[43]</sup>。内核函数及其参数的选择,会影响支持向量机模型分析结果的准确性,本研究选择已被广泛用于土壤制图研究的径向基函数核函数<sup>[43]</sup>。对于径向基函数核函数,需要定义 2 个参数,包括惩罚成本(C)和内核宽度(sigma);本研究中支持向量机参数选择格网搜索方法,即按照一定的步长遍历参数在一定范围内的所有取值,并根据五折交叉验证选择表现最好的模型对应的参数组合作为最终建模的参数取值。

**多层感知机(MLP):** 通过模拟人脑通过大量神经元互相作用对信息进行处理,是一种分布式并行的处理信息的抽象数学模型。在众多神经网络算法中,多层感知机神经网络模型带有反向传播算法,是常用的神经网络算法之一<sup>[22-23]</sup>。多层感知机神经网络的体系结构由输入层、隐藏层、输出层组成,每个层都有一组互连的节点(神经元)并行工作

(见图 2)。

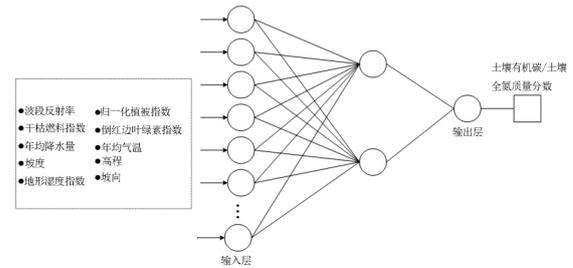


图 2 多层感知机模型(MLP)结构图

### 2.4 模型评价

利用谷歌地球引擎平台上的随机分割函数,将样本随机划分成 7 : 3 的比例,70% 的数据作为建模集、30% 数据作为验证集。选用决定系数 ( $R^2$ )、均方根误差 ( $E_{RMS}$ )、百分比偏差 ( $D_{RP}$ )、四分位数间距性能比 ( $R_{PIQ}$ ) 作为指标,衡量模型的精确度。 $R^2$  越大,同时  $E_{RMS}$  越小,模型效果越好。

$$R^2 = \{ \sum [(O_i - \bar{O})(P_i - \bar{P})] / [(\sum_i (P_i - \bar{P})^2)^{1/2} (\sum_i (O_i - \bar{O})^2)^{1/2}] \}^2;$$

$$E_{RMS} = [ (1/n) \sum_i (P_i - O_i)^2 ]^{1/2};$$

$$D_{RP} = D_s / E_{RMS};$$

$$R_{PIQ} = Q_1 / E_{RMS}。$$

式中:  $P_i$ 、 $O_i$  分别为预测和实测的土壤有机碳(或土壤全氮)质量分数;  $i = 1, 2, \dots, n$ ,  $n$  为验证样本的数量;  $\bar{P}$ 、 $\bar{O}$  为预测和实测的土壤有机碳(或土壤全氮)质量分数的均值;  $D_s$  为样本实测值的标准差;  $Q_1$  为样本实测值的四分位数中的第一分位数与第三分位数的差值。

好的模型预测应该对应高的  $R^2$ 、相对  $D_{RP}$  和  $R_{PIQ}$ <sup>[44]</sup> 值,较低的  $E_{RMS}$  值。本研究采用的模型分类标准是依据  $R_{PD}$  值,将  $R_{PD}$  分为优秀 ( $D_{RP} > 2.5$ )、非常好 ( $2.0 < D_{RP} \leq 2.5$ )、良好 ( $1.8 < D_{RP} \leq 2.0$ )、一般 ( $1.4 < D_{RP} \leq 1.8$ )、较差 ( $1.0 < D_{RP} \leq 1.4$ )、非常差 ( $D_{RP} \leq 1.0$ ) 模型 6 类<sup>[45-46]</sup>。

## 3 结果与分析

### 3.1 采样点基础数据统计

由表 2 可见: 研究区土壤有机碳质量分数均值为 0.82%, 土壤全氮质量分数均值为 0.089%。土壤有机碳、土壤全氮的变异系数相对较高 (60%、58%), 这是由于研究区内不同的草原类型、植被分布、土地利用等造成的土壤有机碳与土壤全氮的空间异质性,也进一步说明了生产高空间分辨率的土壤属性数字地图的必要性。且建模集与验证集的样本,土壤有机碳、土壤全氮质量分数分布较为均匀,均值相近,并与整体样本点值的分布相似,样本划分合理。

表2 采样点基础数据统计结果

样本集	土壤有机碳		土壤全氮	
	质量分数/1%	变异系数/1%	质量分数/1%	变异系数/1%
建模集	0.801±0.510	63	0.088±0.050	56
验证集	0.887±0.500	56	0.092±0.060	63
全集	0.831±0.500	60	0.089±0.050	58

注:质量分数数据为“平均值±标准差”。

### 3.2 评价指标间的相关性

由表3可见:遥感评价指标中波段反射率信息,与土壤有机碳质量分数、土壤全氮质量分数呈现出了显著的负相关( $P<0.05$ )。土壤有机碳与土壤全氮质量分数,与归一化植被指数、干枯燃料指数之间都表现出显著正相关,这是因为土壤碳氮往往控制着植被的生长,土壤碳氮含量越高,肥力越强,植被生长越旺盛,这也和很多研究的结果保持了一致。土壤属性与气候因子之间的相关关系中,土壤有机碳和土壤全氮,均与年均温呈负相关、与年均降水呈正相关;地形对土壤属性的影响不如其他因子显著;这是由于研究区(主要是锡林郭勒盟草地)地形较为平缓,不包含一些山地区域,因此地形对土壤属性的影响并未体现出来。此外,土壤有机碳质量分数与土壤全氮质量分数表现出了极强的相关性( $r=0.81$ )。

表3 各评价指标间相关系数

相关因子	相关系数	
	土壤全氮	土壤有机碳
B <sub>2</sub>	-0.181*	-0.250*
B <sub>3</sub>	-0.251*	-0.336**
B <sub>4</sub>	-0.267*	-0.346**
B <sub>5</sub>	-0.294**	-0.374**
B <sub>6</sub>	-0.305**	-0.388**
B <sub>7</sub>	-0.301**	-0.382**
B <sub>8</sub>	-0.266*	-0.348*
B <sub>9</sub>	-0.240*	-0.321*
B <sub>11</sub>	-0.203*	-0.278*
B <sub>12</sub>	-0.265*	-0.352*
红边叶绿素指数	0.141	0.184*
干枯燃料指数	0.331**	0.423**
倒红边叶绿素指数	0.004	0.063
归一化植被指数	0.468**	0.537**
地形湿度指数	0.002	0.134
坡度	0.038	0.091
坡向	-0.079	-0.071
高程	0.135*	0.120
年均温	-0.186*	-0.262*
年降水量	0.284*	0.358**
土壤全氮	1.000	0.809**
土壤有机碳	0.809**	1.000

注: B<sub>2</sub>~B<sub>12</sub>为哨兵2号影像的各个波段的反射率值。\*表示显著相关( $P<0.05$ ); \*\*表示极显著相关( $P<0.01$ )。

### 3.3 2个模型各评价指标的重要性综合排序

利用随机森林模型计算各个评价指标的重要

性,图3显示了2个模型之间不同的主要环境特征。图3中已剔除重要性小于1%的评价指标,因为重要性评价得分小于1%,在统计学中没有任何意义。在土壤有机碳估算模型中,地形湿度指数重要性排序第一(得分12%),其次是年降水量、年均温(得分10.7%、9.6%)。土壤颜色指数重要性排序第四,然后是干枯燃料指数、哨兵2号卫星的2个短波红外波段。而在土壤全氮模型中,归一化植被指数重要性排序排在第一(得分25.2%),其次依次为年降水量(得分8.9%)、干枯燃料指数(得分8.3%)、红边叶绿素指数(得分6.7%);地形因子中,高程(得分6.3%)、地形湿度指数(得分4.6%)是重要性较高的因子。总体看,在2个模型中遥感评价指标(哨兵2号卫星波段与光谱指数)贡献较大,尤其是土壤全氮模型中(光谱反射率23.2%,光谱指数46.5%),地形和气候对土壤有机碳、土壤全氮的空间分布也有一定程度的贡献。

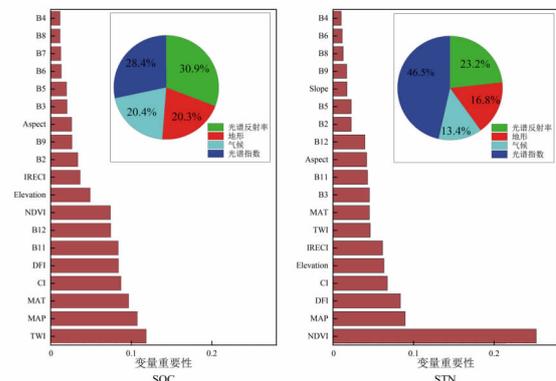


图3 2个模型各评价指标的重要性

SOC为土壤有机碳; STN为土壤全氮; Elevation为高程; Slope为坡度; Aspect为坡向; MAP为年降水量; MAT为年均温; NDVI为归一化植被指数; DFI为干枯燃料指数; CI为红边叶绿素指数; IRECI为倒红边叶绿素指数; TWI为地形湿度指数; B<sub>1</sub>~B<sub>12</sub>分别对应哨兵2号遥感影像的各个波段反射率。

### 3.4 3种算法构建的模型精度检验

由图4可见:随机森林模型拟合效果最佳,其次是多层感知机算法。随机森林算法构建的土壤全氮模型,决定系数为0.67、均方根误差为0.024、百分比偏差为2.09、四分位数间距性能比为2.14。随机森林算法构建的土壤有机碳模型,决定系数为0.68、均方根误差为0.17、百分比偏差为2.26、四分位数间距性能比为2.37。由散点图可见:在土壤全氮模型中,虽然随机森林模型的整体精度最好,但是在土壤全氮的高值区,多层感知神经网络模型估测效果更好,而支持向量机模型在土壤全氮估测中高估和低估现象都比较明显。在土壤有机碳模型中,随机森林模型在低值区存在一定的高估,但是模型整体表现优异,支持向量机模型、多层感知机算法模型

则都表现不佳。按照百分比偏差 ( $D_{RP}$ ) 的模型分类标准,随机森林算法构建的土壤有机碳模型、土壤全氮模型都是非常好的模型;支持向量机算法、多层感

知机算法构建的土壤有机碳模型与土壤全氮模型,虽然存在一些较为明显的局部低估和高估现象,但也不失为一个良好的模型。

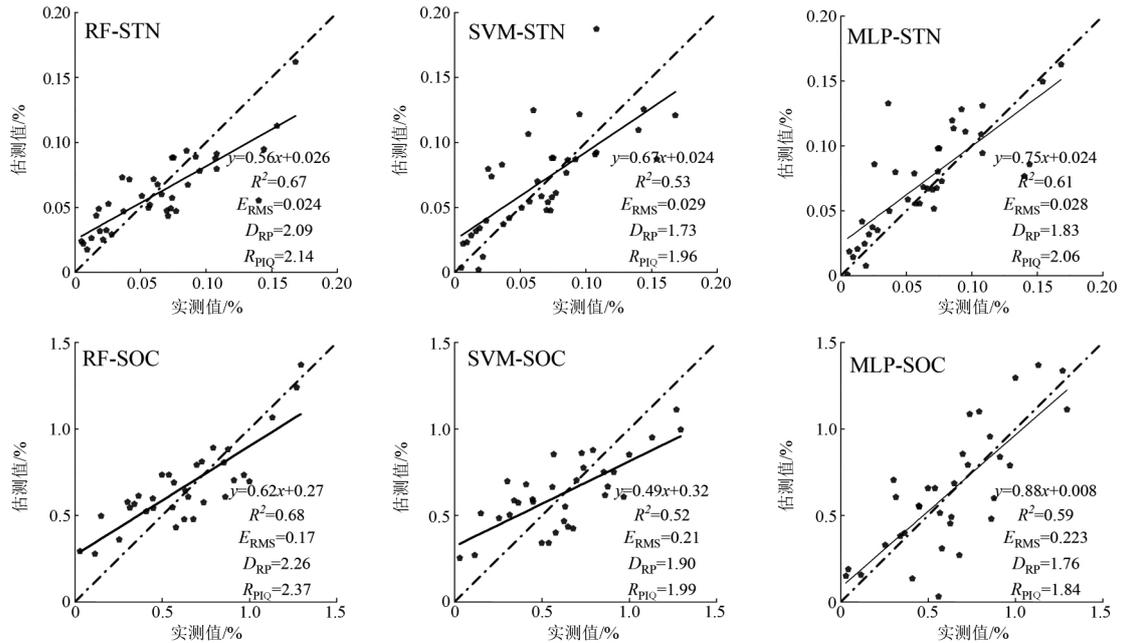


图 4 模型精度验证散点图

### 3.5 土壤属性空间分布特征

由效果最好的随机森林模型预测土壤有机碳、土壤全氮的空间分布(见图 5,土壤有机碳和土壤全氮的空间分布表现出了很大的一致性,与相关性分析中两种土壤属性的表现一样。土壤有机碳与土壤全氮,从西向东增加,在西部西南地区分布高于西部

其他区域,这和草地类型的变化规律呈现出一致性。这显示出草地类型对土壤属性的影响,自西向东草原类型依次为温性荒漠草原、典型草原、草甸草原;这种现象在不同草地类型的土壤属性统计分析中得到证实(见表 4)。总体看,锡林郭勒盟草地土壤有机碳和土壤全氮质量分数很低,处于缺乏状态。

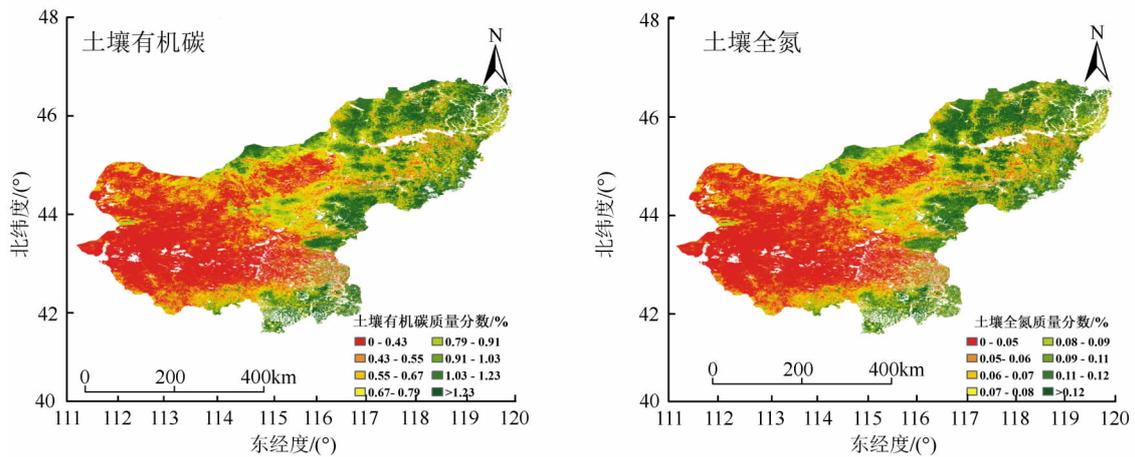


图 5 土壤属性空间分布

表 4 不同草原类型土壤属性特征

草原类型	土壤有机碳质量分数/%	土壤全氮质量分数/%
草甸草原	0.92±0.22	0.096±0.032
典型草原	0.69±0.27	0.076±0.036
荒漠草原	0.42±0.19	0.047±0.019

注:表中数据为“平均值±标准差”。

## 4 讨论

本研究中不同机器学习模型的精度对比表明,

机器学习算法的选择对土壤属性预测模型的性能有很大的影响(见图 4)。总体看,随机森林模型预测效果最好,其次为多层感知机模型。这与 Wang et al.<sup>[19]</sup>用不同机器学习方法在澳大利亚干旱牧场的土壤有机碳估测研究结果相似。但是,也有在不同机器学习的土壤属性制图研究中,出现不同的结论,例如:Were et al.<sup>[23]</sup>在肯尼亚东部森林区域对比支持向量机、多层感知机、随机森林方法,用陆地卫星

8号(Landsat 8)数据预测土壤有机碳时,支持向量回归(SVR)建模优于随机森林建模、人工神经网络(ANN)建模。

根据随机森林模型中的评价指标重要性评价(见图3),整体看,遥感指标对评价土壤有机碳质量分数、土壤全氮质量分数的空间分布贡献最多,其总重要性分别为59%、69%。已有研究观察到了类似的结果,遥感数据通过波段反射率和植被指数以及一些土壤指数,可以提供与植被生长和土壤状况有关的生物物理特性<sup>[6,15-16]</sup>。植被是土壤中有机碳和总氮的重要来源,它与表土中土壤碳和氮的空间格局高度相关<sup>[25,27,32]</sup>。本研究中植被指数的选择,除了常用的归一化植被指数之外,还选择了被广泛应用于非光合植被信息提取的干枯燃料指数和使用了哨兵2号卫星数据的红边信息的倒红边叶绿素指数;在因子重要性评价中,重要性在土壤有机碳模型中分别排到了第四、第十,在土壤全氮模型中分别排到了第三、第六。这充分证明了,在土壤有机碳、土壤全氮估测中,代表非光合植被信息干枯燃料指数的重要性以及哨兵2号卫星红边信息对土壤属性建模的贡献。

环境指标中的地形指标被确定为本研究的土壤预测模型(尤其是土壤有机碳预测模型)的重要预测因子。地形是可以控制景观尺度水文和土壤过程的关键因素,对土壤形成有重要影响,进而影响土壤性质的空间分布。在所有地形指标中,地形湿度指数是土壤有机碳预测最重要的指标,地形湿度指数捕获土壤水分分布的能力,它经常被用作绘制土壤特性图的关键因子<sup>[23]</sup>。气候也是影响土壤形成过程的5个基本要素之一,并且已经充分证明了其对土壤碳和氮的影响<sup>[13,34]</sup>。在随机森林模型中,年降水量和年均温分别被确定为土壤有机碳预测模型的第二和第三重要的指标,是土壤全氮预测模型中的第二和第八重要的指标。很多研究都一致表明了,温度和降水量是控制土壤碳和氮循环的最重要的气候指标,它们通过生物或非生物途径影响土壤的碳和氮库。例如,温度和降水会影响净初级生产力(NPP)、土壤中碳和氮的输入,并影响生物活性、凋落物的积累和分解速率,从而影响土壤有机碳、土壤全氮空间分布<sup>[34]</sup>。

## 5 结论

制图中使用高分辨率的哨兵2号卫星及依据哨兵2号卫星数据计算的光谱指数,是土壤有机碳模型、土壤全氮模型中的重要预测因子,尤其是代表非光合植被信息的干枯燃料指数和倒红边叶绿素指

数。

不同的机器学习方法的选择会影响土壤属性的估测效果,随机森林模型在本研究中的土壤有机碳、土壤全氮估测中估测效果最佳:(土壤全氮模型,决定系数为0.67、均方根误差为0.024、百分比偏差为2.09、四分位数间距性能比为2.14;土壤有机碳模型,决定系数为0.68、均方根误差为0.17、百分比偏差为2.26、四分位数间距性能比为2.37)。

锡林郭勒草地土壤有机碳质量分数、土壤全氮质量分数整体偏低,土壤有机碳与土壤全氮之间具有一致的空间分布规律和极强的相关性( $r=0.81$ ),空间分布保持一致,自西向东逐渐增加。

## 参 考 文 献

- [1] PLAZA C, GASCÓ G, MÉNDEZ M A, et al. Chapter 2—soil organic matter in dryland ecosystems [J]. *The Future of Soil Carbon*, 2018: 39–70. doi: 10.1016/B978-0-12-811687-6.00002-X.
- [2] LAL R. Soil carbon sequestration impacts on global climate change and food security [J]. *Science*, 2004, 304: 1623–1627. doi: 10.1126/science.1097396.
- [3] TAO W. Aeolian desertification and its control in Northern China [J]. *International Soil and Water Conservation Research*, 2014, 2(4): 34–41.
- [4] 高露,张圣微,赵鸿彬,等.退化草原土壤理化性质空间异质性及其对土壤水分的影响[J].*干旱区研究*, 2020, 37(3): 607–617.
- [5] 吴昕,徐凯,张继吟,等.中国锡林郭勒草原沙漠化的沙源及其地质学成因[J].*中国沙漠*, 2018, 38(1): 92–100.
- [6] LIN C, ZHU A X, WANG Z F, et al. The refined spatiotemporal representation of soil organic matter based on remote images fusion of Sentinel-2 and Sentinel-3 [J]. *International Journal of Applied Earth Observation and Geoinformation*, 2020, 89. doi: 10.1016/j.jag.2020.102094.
- [7] 徐彬彬,戴昌达.新疆土壤光谱反射特性与有机质含量的相关分析[J].*科学通报*, 1980(6): 282–284.
- [8] 姜赛平,张怀志,张认连,等.基于三种空间预测模型的海南岛土壤有机质空间分布研究[J].*土壤学报*, 2018, 55(4): 1007–1017.
- [9] WULF H, MULDER V L, SCHAEPMAN M, et al. Remote sensing of soils [R]. University of Zurich, Switzerland, 2015. doi: 10.13140/2.1.1098.0649.
- [10] 贾伟,高小红,杨扬,等.基于Landsat 8 OLI影像的三江源区表层土壤全氮空间格局反演[J].*干旱区研究*, 2015, 32(5): 890–896.
- [11] ZHOU T, GENG Y J, JI C, et al. Prediction of soil organic carbon and the C:N ratio on a national scale using machine learning and satellite data: A comparison between Sentinel-2, Sentinel-3 and Landsat-8 images [J]. *Science of The Total Environment*, 2021, 755. doi: 10.1016/j.scitotenv.2020.142661.
- [12] 贾萍萍,孙媛,尚天浩,等.基于高光谱和Landsat-8 OLI影像的盐渍化土壤水盐估算模型构建[J].*生态学杂志*, 2020, 39(7): 2456–2466.
- [13] 庞龙辉,刘峰,赵霞,等.青海省表层土壤属性数字制图[J].*土壤通报*, 2019, 50(3): 505–513.
- [14] WANG S, ZHUANG Q L, WANG Q B, et al. Mapping stocks of soil organic carbon and soil total nitrogen in Liaoning Province of China [J]. *Geoderma*, 2017, 305: 250–263. doi: 10.1016/j.geoderma.2017.05.048.
- [15] VAUDOUR E, GOMEZ C, FOUAD Y, et al. Sentinel-2 image capacities to predict common topsoil properties of temperate and Mediterranean agroecosystems [J]. *Remote Sensing of Environment*, 2019, 223: 21–33. doi: 10.1016/j.rse.2019.01.006.
- [16] CASTALDI F, HUENI A, CHABRILLAT S, et al. Evaluating the capability of the Sentinel 2 data for soil organic carbon prediction in croplands [J]. *ISPRS Journal of Photogrammetry and Re-*

- remote Sensing, 2019, 147: 267–282. doi: 10.1016/j.isprsjprs.2018.11.026.
- [17] MCBRATNEY A B, SANTOS M L M, MINASNY B. On digital soil mapping [J]. *Geoderma*, 2003, 117: 3–52. doi: 10.1016/S0016-7061(03)00223-4.
- [18] BREIMAN L. Random forests [J]. *Machine Learning*, 2001, 45 (1): 5–32.
- [19] WANG B, WATERS C, ORGILL S, et al. High resolution mapping of soil organic carbon stocks using remote sensing variables in the semi-arid rangelands of eastern Australia [J]. *Science of The Total Environment*, 2018, 630: 367–378. doi: 10.1016/J.SCI-TOTENV.2018.02.204.
- [20] ZHANG H, WU P B, YIN A J, et al. Prediction of soil organic carbon in an intensively managed reclamation zone of eastern China: A comparison of multiple linear regressions and the random forest model [J]. *Science of the Total Environment*, 2017, 592: 704–713. doi: 10.1016/j.scitotenv.2017.02.146.
- [21] BALLABIO C. Spatial prediction of soil properties in temperate mountain regions using support vector regression [J]. *Geoderma*, 2009, 151(3/4): 338–350.
- [22] MALONE B P, MCBRATNEY A B, MINASNY B, et al. Mapping continuous depth functions of soil carbon storage and available water capacity [J]. *Geoderma*, 2009, 154(1/2): 138–152.
- [23] WERE K, BUI D T, DICK Ø B, et al. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afrotropical landscape [J]. *Ecological Indicators*, 2015, 52: 394–403. doi: 10.1016/j.ecolind.2014.12.028.
- [24] 张玥,许端阳,王子玉,等.2000—2015 年锡林郭勒盟防风固沙服务功能变化驱动因素分析 [J]. *生态学报*, 2021, 41(2): 603–614.
- [25] 朱阿兴,杨琳,樊乃卿,等.数字土壤制图研究综述与展望 [J]. *地理科学进展*, 2018, 37(1): 66–78.
- [26] OBA G, WELADJI R B, MSANGAMENO D J, et al. Scaling effects of proximate desertification drivers on soil nutrients in northeastern Tanzania [J]. *Journal of Arid Environments*, 2008, 72(10): 1820–1829.
- [27] LAMICHHANE S, KUMAR L, WILSON B. Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review [J]. *Geoderma*, 2019, 352: 395–413. doi: 10.1016/j.geoderma.2019.05.031.
- [28] 孙孝林,赵玉国,刘峰,等.数字土壤制图及其研究进展 [J]. *土壤通报*, 2013, 44(3): 752–759.
- [29] ZERAATPISHEH M, AYOUBI S, JAFARI A, et al. Digital mapping of soil properties using multiple machine learning in a semi-arid region, central Iran [J]. *Geoderma*, 2019, 338: 445–452. doi: 10.1016/j.geoderma.2018.09.006.
- [30] CHAGAS C DA S, JUNIOR W DE C, BHERING S B, et al. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions [J]. *CATENA*, 2016, 139: 232–240. doi: 10.1016/j.catena.2016.01.001.
- [31] TUCKER C J. Red and photographic infrared linear combinations for monitoring vegetation [J]. *Remote Sensing of Environment*, 1979, 8(2): 127–150.
- [32] WANG X Y, LI Y Q, GONG X W, et al. Storage, pattern and driving factors of soil organic carbon in an ecologically fragile zone of northern China [J]. *Geoderma*, 2019, 343: 155–165. doi: 10.1016/j.geoderma.2019.02.030.
- [33] CAO X, CHEN J, MATSUSHITA B, et al. Developing a MODIS-based index to discriminate dead fuel from photosynthetic vegetation and soil background in the Asian steppe area [J]. *International Journal of Remote Sensing*, 2010, 31(6): 1589–1604.
- [34] DAVIDSON E A, JANSSENS I A. Temperature sensitivity of soil carbon decomposition and feedbacks to climate change [J]. *Nature*, 2006, 440: 165–173. doi: 10.1038/nature04514.
- [35] 任建新,唐亮,何明珠,等.陇西县土壤氮储量的空间分布特征及其影响因素研究 [J]. *水土保持学报*, 2020, 34(6): 318–324.
- [36] 孙文义,郭胜利.黄土丘陵沟壑区小流域土壤有机碳空间分布及其影响因素 [J]. *生态学报*, 2011, 31(6): 1604–1616.
- [37] KESKIN H, GRUNWALD S, HARRIS W G. Digital mapping of soil carbon fractions with machine learning [J]. *Geoderma*, 2019, 339: 40–58. doi: 10.1016/j.geoderma.2018.12.037.
- [38] ZHOU T, GENG Y J, CHEN J, et al. High-resolution digital mapping of soil organic carbon and soil total nitrogen using DEM derivatives, Sentinel-1 and Sentinel-2 data based on machine learning algorithms [J]. *Science of the Total Environment*, 2020, 729(8). doi: 10.1016/j.scitotenv.2020.138244.
- [39] FARR T G, ROSEN P A, CARO E, et al. The shuttle radar topography mission [J]. *Reviews of Geophysics*, 2007, 45(2). doi: 10.1029/2005RG000183.
- [40] HENGL T. Global DEM derivatives at 250 m, 1 km and 2 km based on the MERIT DEM(1.0) [Data set]. Zenodo, 2018. doi: org/10.5281/zenodo.1447210.
- [41] LIAW A, WIENER M. Classification and regression by random-forest [J]. *R News*, 2002, 2/3: 18–22.
- [42] CORTES C, VAPNIK V. Support-vector networks [J]. *Machine Learning*, 1995, 20: 273–297. doi: 10.1023/A:1022627411411.
- [43] 陈黔,李晓松,修晓敏,等.基于 Google Earth Engine 与机器学习的大尺度 30 m 分辨率沙地灌木覆盖度估算 [J]. *生态学报*, 2019, 39(11): 4056–4069.
- [44] MCBRATNEY A B, FERNÁNDEZ-AHUMADA E, PALAGOS B, et al. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy [J]. *TrAC Trends in Analytical Chemistry*, 2010, 29(9): 1073–1081.
- [45] CIPULLO S, NAWAR S, MOUAZEN A M, et al. Predicting bio-availability change of complex chemical mixtures in contaminated soils using visible and near-infrared spectroscopy and random forest regression [J]. *Scientific Reports*, 2019, 9(1). doi: 10.1038/s41598-019-41161-w.
- [46] VISCARRA ROSSEL R A, WALVOORT D J J, MCBRATNEY A B, et al. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties [J]. *Geoderma*, 2006, 131(1/2): 59–75.
- [38] RABBI S M F, TIGHE M K, WARREN C R, et al. High water availability in drought tolerant crops is driven by root engineering of the soil micro-habitat [J]. *Geoderma*, 2020. doi: 10.1016/j.geoderma.2020.114738.
- [39] YUAN Z, CHEN H Y H. Global trends in senesced-leaf nitrogen and phosphorus [J]. *Global Ecology & Biogeography*, 2009, 18(5): 532–542.
- [40] 白小芳.施肥对华北落叶松养分含量及生态化学计量特征影响研究 [D].北京:中国科学院研究生院(教育部水土保持与生态环境研究中心), 2016.

### (上接 63 页)

- [34] 宋影.板栗剩余物堆肥及其对板栗幼苗生长和果实提质增产作用机理研究 [D].北京:北京林业大学, 2019.
- [35] OKADA K, AIBA S, KITAYAMA K. Influence of temperature and soil nitrogen and phosphorus availabilities on fine-root productivity in tropical rainforests on Mount Kinabalu, Borneo [J]. *Ecological Research*, 2017, 32(2): 145–156.
- [36] 李峰卿,王秀花,楚秀丽,等.缓释肥 N/P 养分配比及加载量对 3 种珍贵树种大规模容器苗生长的影响 [J]. *林业科学研究*, 2017, 30(5): 743–750.
- [37] 李永涛,杨庆山,王莉莉,等.滨海盐碱地不同林龄白蜡人工林根系分布及土壤特性变化 [J]. *东北林业大学学报*, 2020, 48