引文格式：谢永强. 集成多源数据与XGBoost算法京津冀地区土壤水分空间反演[J]. 地理空间信息, 2024, 22(12): 20-24. doi:10. 3969/j.issn. 1672-4623.2024.12.005

集成多源数据与XGBoost算法京津冀地区土壤水分空间反演

谢永强[1]
（1. 中国四维测绘技术有限公司，北京 100025）

摘　要：为提升土壤水分（soil moisture，SM）遥感反演时效性及其空间分辨率，以京津冀地区为研究区，集成Sentinel-1 Synthetic Aperture Radar （SAR）、Sentinel-2 Multi-Spectral Imaging （MSI）、Digital Elevation Model （DEM）、气候数据与地面同步实测SM资料，引入Extreme Gradient Boosting （XGBoost）算法构建SM预测模型进而反演区域100 m分辨率水平上SM空间分布，最后通过决定系数（R^2）、平均绝对误差（mean absolute error，MAE）和均方根误差（root mean square error，RMSE）评估反演精度。结果表明，京津冀地区SM空间异质性受地形因素控制，不同类型的环境变量在反演SM分布中的贡献率依次为：地形因子（39.84%）>MSI（22.05%）>气候因素（19.66%）>SAR（18.45%）。XGBoost算法具有较强拟合能力，其反演的SM精度R^2达0.72，MAE、RMSE分别为0.030、0.036，反演结果显示京津冀地区SM呈现自东北向周边地带性减少分布。该研究展示了集成多源数据结合XGBoost算法在异质性景观区SM监测方面的潜力。

关键词：土壤水分；XGBoost算法；多源数据；空间反演
中图分类号：P208　　**文献标志码：**B　　**文章编号：**1672-4623（2024）12-0020-05

Spatial Inversion of Soil Moisture in the Beijing-Tianjin-Hebei Region Using Integrated Multi-source Data and XGBoost Algorithm

XIE Yongqiang[1]
(1. China Four Dimensions Surveying & Mapping Technology Co., Ltd., Beijing 100025, China)

Abstract: For improving the timeliness and spatial resolution of inverting soil moisture (SM) from remote sensing, taking the Beijing-Tianjin-Hebei region as the study area, we integrated Sentinel-1 synthetic aperture radar (SAR), Sentinel-2 multi-spectral imaging (MSI), digital elevation model (DEM), climate data and simultaneous ground-based SM data, and used the XGBoost algorithm to construct an SM prediction model to invert the spatial distribution of SM at the 100 m resolution level in the region. We evaluated the inversion accuracy by coefficient of determination (R2), mean absolute error (MAE) and root mean square error (RMSE). The results show that the spatial heterogeneity of SM in the Beijing-Tianjin-Hebei region is controlled by topographic factors, and the contribution of different source environmental variables to the SM inversion distribution is in the order of topographic factors(39.84%) > MSI (22.05%) > climatic factors (19.66%) > SAR (18.45%). The XGBoost algorithm has a strong fitting ability, and its R^2 reaches 0.72, MAE and RMSE are 0.030 and 0.036, respectively. The inversion results show that SM in the Beijing-Tianjin-Hebei region shows a zonal decrease in distribution from the northeast to the periphery. This study demonstrates the potential of multi-source data integration combined with the XGBoost algorithm for SM monitoring in heterogeneous landscape areas.

Key words: SM, XGBoost algorithm, multi-source data, spatial inversion

当前被广泛用于SM反演的众源遥感资料主要包括多光谱（multi-spectral imaging，MSI）[1-2]、合成孔径雷达（synthetic aperture radar，SAR）[3-4]和微波影像[5-6]。前人已经在不同景观尺度上开发了支持向量机[7-10]、随机森林（random forest，RF）[8]、决策树[9]、梯度提升树[11]、高斯过程回归[12]等SM反演模型[13]。极限梯度提升树（extreme gradient boosting，XGBoost）是当前一种流行的树模型，其通过一系列迭代残差和正则化来修正预测偏差，较好解决了不平衡样本条件下收敛速度慢、欠拟合问题[14]。当前XGBoost算法被成功应用于径流模拟、地质灾害识别、遥感影像分类、水质识别等领域，并展示了非线性回归拟合方面的潜力，然而该算法在SM反演研究中并不多见。

本文以该地为案例，通过集成星地众源资料构建XGBoost回归模型反演大区域尺度SM空间分布特征，以期为复杂景观区SM监测提供新技术手段。

1　研究区与研究方法

1.1　研究区概况

京津冀地处华北北部、渤海湾西侧，地理坐标

收稿日期：2023-05-18。
作者简介：谢永强（1985—），本科，工程师，从事多源遥感信息融合提取与定量反演工作，E-mail：start.net@163.com。

介于113°27′~119°50′E，36°05′~42°40′N，总面积21.8万km²。地区跨燕山太行山地、坝上高原与黄淮海平原，地势自西北两侧向东、南部倾斜，海拔介于0~2 882 m（图1）。

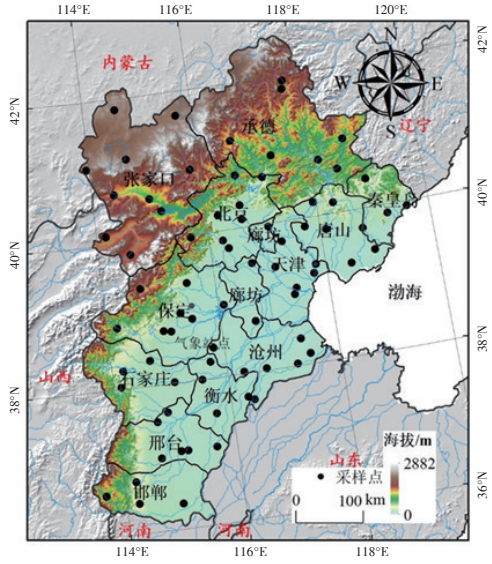


图1　土壤水分采样点分布图

审图号：GS(2024)4468号

受海陆位置与气候环流系统影响形成温带季风气候，气候季节性差异明显但雨热同步，其多年平均气温12℃，降水量在400~800 mm之间，近年来常受干旱威胁。该地汇聚海、滦、辽等水系，水文特征平稳且常年处于枯水状态。地带性土壤为黄壤、潮土、盐土等，质地轻而肥力质量轻、保水性差而墒情低。作为北方经济、文化和人口发展中心，京津冀地区土壤水源涵养面临严峻压力。

1.2　数据收集与处理

1.2.1　地面人工实测数据

考虑卫星遥感器成像时间，于2022年5月2日—5月7日进行地面SM采样。为确保样本代表性，以土地利用类型、DEM数据为分层要素，借鉴Blanco [5]等分层布点法进行样点设计，最后共采集175个表层土壤样本（0~20 cm）。每一样方内具有均质景观，其尺寸为30×30 m网格，于样方四角与中心位置采集土壤样品，经充分混合后装袋封存并现场称重。同时利用手持GPS仪记录样方中心点空间坐标。将封存的土壤样品带回实验室，利用烘干法测定SM[5,8-9]。

1.2.2　Sentinel-1&2遥感数据

Sentinel-1&2是欧空局哥白尼计划重要对地观测卫星之一（https://scihub.copernicus.eu/s2/#/home），在陆地生态领域具有广泛应用。本文选用2020年5月1日-12日Sentine-1B的IW数据，。对每景SAR图像场使用SNAP（version 6.0, ESA）软件先逐次进行斑点过滤、辐射校准、地形校正和拼接裁剪处理；然后基于3×3像素窗口生成HV（水平发射-垂直通道）、HH（水平发射-水平通道）的后项散射系数[8]。

使用的Sentinel-2B MSI数据时间域为5月2日—15日（https://scihub.Copernicus.eu/s2 /#/home），产品等级为L1C（已经过几何校正），其包含13个波段，存在10、20、60 m不同空间分辨率[6]。先利用Sen2Cor程序进行大气校正处理，然后使用SNAP软件实施重采样、镶嵌、裁剪等预处理，并利用Vegetation sepctral index模块提取相关光谱指数。

1.2.3　地形因子

地形因子解释了地表产流、汇聚、渗流、蒸散等具体径流过程[6,8]。选用Aster DEM数据为地形因子数据源，该数据由美国宇航局ASTER（Advanced Spaceborne Thermal Emission and Reflection Radiometer）传感器联合日本宇航研发的TerraSar-X卫星进行激光雷达测量系统制作而成，空间分辨率达30 m，可描述微地貌形态。利用SAGAGIS（http://www.saga-gis.org/en/index.html）软件提取7个与地表水文密切相关的地形因子。

1.2.4　降水和气温数据

与地面采样时间同步的降水量数据来自Global Precipitation Measurement（GPM）降水产品，其空间分辨率为0.1°，时间分辨率为0.5 h，从Google Earth Engine开源平台（https://urs.earthdata.nasa.gov/）收集研究区2020年5月1日—9日降水量。气温数据从气象数据共享服务网（https://urs.earthdata.nasa.gov/）获取，并利用Oridinary Kriging方法获取观测期内平均温度。

为与地面SM采样数据空间信息匹配，将多源环境变量重采样至30 m空间分辨率，重采样方法为cubic，并设置其投影系统为WGS-UTM-1984。

1.3　研究方法

1.3.1　基于XGBoost算法的土壤水分反演方法

从多源地理空间数据平台收集土壤水分影响因子后，首先，利用Pearson分析法明确环境因子与SM之间相关性及其有效性相关；其次，构建基于XGBoost回归的SM预测模型，并自适应模型参数以确保其稳健性；然后，利用XGBoost模型中变量对模型精度增益，对环境因子进行重要性排序；最后，利用独立样本验证模型精度，并将模型代入环境变量空间数据集进行SM空间反演。本文技术流程见图2。

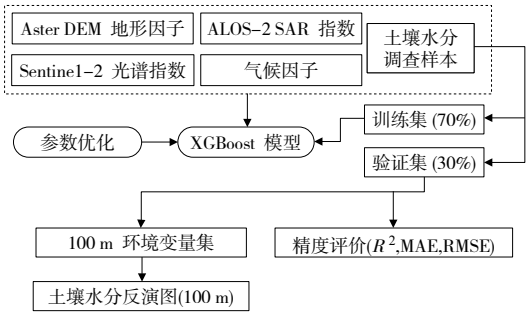


图2 本研究技术流程图

1.3.2 XGBoost回归模型

XGBoost（extreme gradient boosting）是由Chen[14]等基于梯度提升学习器的一种改进集成回归树（classification and regression tree，CART）算法，主要是通过迭代地添加弱CART建立强学习器[7-8]，即在训练过程中不断将前一CART输出结果作为一新变量添加至后一棵CART上，并利用梯度提升对之前预测结果的残差进行拟合[14]。为避免过拟合，XGBoost引入最小化目标函数，允许预测和正则化相结合。XGBoost最终预测定义为每一CART的平均输出。其通用函数如下：

$$f_i^p = \sum_{k=1}^{l} f_k(x_i) = f_i^{p-1} + f_i(x_i)$$

式中，$f_p(x_i)$为第 p 次迭代过程中CART对 i 个输入样本 x_i 产生的预测结果。对优化目标Obj，XGBoost对损失函数定义如下：

$$\mathrm{Obj}^{(p)} = \sum_{k=1}^{n} l(\bar{y}_i, y_i) + \sum_{k=1}^{p} \delta(f_i)$$

式中，n 为样本数量；δ 为训练阶段；y_i 为第 i 个样本集里的因变量；y_i 为其平均值。鉴于XGBoost模型参数复杂、自适应收敛困难，采用人为干预法对模型超参数（gamma、alpha、lanbda、bosting iterations）进行优化[14-15]。

1.4 土壤水分预测精度评价

将175个样本数据随机划分成两组：训练集（70%，123个）和验证集（30%，52个），前者用于模型训练，后者用以对模型精度验证[5,8]。基于真实测量值与预测值，计算得到决定系数（R^2）、平均绝对误差（mean absolute error, MAE）和均方根误差（root mean square error, RMSE）进行精度评价。

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i')^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$\mathrm{MAE} = \sum_{i=1}^{n} \frac{|y_i - y_i'|}{n}$$

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(y_i - y_i')^2}{n}}$$

式中，y 、y' 分别为XGBoost模型预测土壤水分含量、地面实际测量值；$i \in 1, 2, \cdots, n$ 为独立验证样本数量。

2 结果与分析

2.1 研究区土壤含水率描述性统计

研究区175个土壤水分（SM）样本范围为0.14~0.64，标准差、平均值为0.04、0.26，呈现中等程度变异性（离差系数72.62%）。华北异质性景观区表层SM与下垫面属性、气候、土壤-水-植被之间的交互活动密切相关，因此环境因子的离散梯度影响着SM的反演效果。研究区100 m空间分辨率环境因子统计特征见表1，除Eas、Ele、HC、MC、Nor、Slo、VC的离散性呈强烈程度外（0<CV<1）外，其他因子具有中度离散性（0.1>|CV|>1）。其中MSI光谱指数DVI、EVI、IBI、NDBI、NDVI、RVI和SAVI的离差系数仅为0.27~1.21，说明这些环境特征在该区域内均一性好；类似的变量气候因子MAP和MAT的在域内空间差异不大。从SAR影像指数来看，HV、HVs、NDHV对研究区地物信息差异性的描述相近（CV介于-0.27~0.44），而H_V、HH探测的地表信息差异性高于前者。

2.2 土壤水分与环境变量之间统计关系分析

将基于地面SM样点位置提取的全部环境变量与实测SM数据进行相关性分析，得到二者间线性统计关系。研究区内SM与降水量（MAP）之间呈现出极显著正相关关系（$r=63$，$P<0.01$），这是由于前期降水数量分布直接影响土壤含水率、湿度等，区内MAP介于134.6~284.6 mm（表1），在降水-植被-土壤复杂生态过程影响下，导致SM发生局部变异。地形是引起有SM空间变异的重要因子，研究显示SM与向东度（Eas）、海拔（Ele）和垂直曲率（HC）呈显著线性相关（$P<0.05$），这些地形属性描述了地表水文、产流、入渗等特征。SM与其他环境因子相关性不显著，可能其间存在非线性关系，对此仍需进一步探究。

2.3 XGBoost建模与验证分析

超参数是影响XGBoost模型结构、运行速率和收敛性的关键配置。在构建XGBoost模型时以地面实测SM为因变量、环境因子为自变量，XGBoost算法进行拟合二者间非线性关系。为确保模型可移植性，使用

表1　环境变量描述性统计特征

环境变量	最大值	最小值	平均值	标准差	离差系数
SM	0.46	0.14	0.26	0.04	0.16
Asp/(°)	357.42	1.29	163.39	102.68	0.63
Eas	1.00	−1.00	0.11	0.64	5.56
Ele/m	1 419.91	1.23	173.59	314.52	1.81
HC	0.00	0.00	0.00	0.00	11.69
MC	0.00	0.00	0.00	0.00	−31.62
Nor	1.00	−1.00	−0.04	0.76	−18.85
Slo/(°)	27.83	0.03	1.15	2.77	2.42
VC	0.00	0.00	0.00	0.00	−9.11
HH	6.63	−14.35	−4.79	3.57	−0.74
HV	−6.54	−24.26	−13.85	2.97	−0.21
HVs	11.12	1.40	3.04	1.33	0.44
H_V	27 587.00	1 059.00	5 834.40	3 571.95	0.61
NDHV	0.83	0.17	0.47	0.12	0.27
DVI	1 996.00	0.00	445.83	321.30	0.72
EVI	5.07	0.00	0.40	0.47	1.17
IBI	0.16	−0.18	0.06	0.06	0.94
NDBI	0.22	−0.23	0.05	0.07	1.28
NDVI	0.58	0.00	0.15	0.10	0.64
RVI	3.76	1.00	1.39	0.37	0.27
SAVI	0.87	0.00	0.22	0.14	0.64
MAP/mm	284.60	134.60	130.47	52.72	0.40
MAT/℃	15.64	4.09	12.32	2.63	0.21

GridSearch（格网搜索）结合交叉验证方法进行超参数寻优，其调试最佳取值如图3所示。可见该模型性能受超参数影响较大，当参数gamma、alpha、lanbda、bosting iterations分别取值0.01、0.1、0.1、325时，模型精度R^2达到最高值（0.843），故而该数组为文中XGBoost模型反演SM的最优配置。基于以上优化过程此训练XGBoost模型，并分别利用测试集、训练集计算精度评价指标R^2、MAE和RMSE，结果如图4所示。利用30%的测试样本验证得到XGBoost模型的R^2为0.72，相比于训练样本集里R^2减小了9.2%；相应地，其MAE和RMSE则为0.030、0.036，相对于建模集分别降低了6.14%、25.0%。显然，XGBoost算法反演的SM值与实测值分布于1∶1线两侧，并具有局部离散特征，表明该模型在训练过程中并未完全拟合环境—SM之间关系（图3a），而将其移植应用于验证集后该离散性有一定增大但依然可接受（图3b）。基于线性分析与XGBoost算法模拟结果来看（图3），XGBoost模型R^2始终高于简单线性统计关系，表明XGBoost模型能挖掘SM环境因素间显性、隐式关系进而更稳健拟合SM。综合而言，使用多源环境变量与

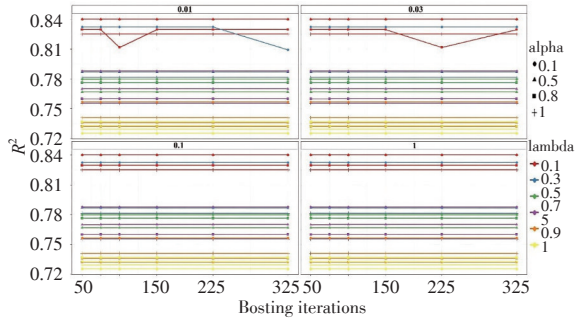


图3　使用网格搜索法XGBoost模型超参数优化过程

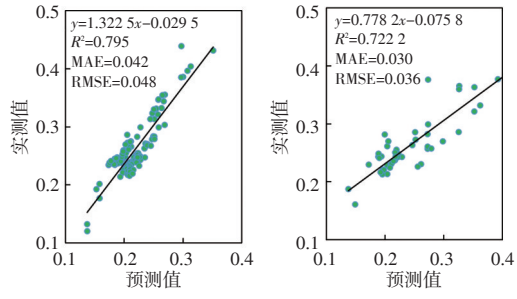


图4　土壤水分反演与实测值关系图

XGBoost算反演大区域尺度SM空间分布过程中不存在过拟合而具有良好的适用性。

2.4　土壤水分影响因素分析

图5为XGBoost算法根据环境协变量对模型性能的平均降低，进而无偏量化环境变量对SM的影响。本研究中MAP是第一重要变量，已有研究从不同景观尺度均证实MAP是SM的载体与驱动因素[6,12]。Ele、HC通过控制地表径流、地表景观变化来影响土SM，Ele、HC决定了景观类型、蒸散发速率、径流剪应力等，能够捕获不同景观位置上地形、水分信息。在京津冀尺度研究表明MAT是SM的重要预测因子，可能由于研究区MAT受海拔、坡向、下垫面属性影响呈现一定空间分异，进而影响了SM的局部变化。其中，MAP、Ele、HC和Eas这4个变量的重要性分值介于6.31~13.72%，显著大于其他变量重要性分值，解释了SM变异性的55.64%。而变量HVs、NDHV、HV、CI、Slo、DVI、H_V、VC的相对重要性介于3.67%~4.86%，解释了32.63%的信息，是影响SM的次要变量。其他变量的相对重要性介于1.33%~3.46%之间，其占SM总信息的11.73%，因而是不重要变量，但并非意味着这些变量对不能解释SM的变异性。

进一步归类分析发现影响京津冀地区SM空间分布性最重要的变量是地形变量，其相对重要性为39.84%；气候、MSI、SAR依次解释了19.66%、22.05%、18.45%的信息。这是由于来自窄带、红边和红外光谱区域的光谱指数对绿色植被很敏感，而研究区农用地类分布范围广而易对太阳辐射产生遮挡而造

成MSI数据饱和。地形变量的贡献率高于SAR变量，因为本研究区地形与SM的关系比SAR变量更密切（表1），地形因子反映了可及性人类干扰和微气候变化，进而影响SM分布。SAR衍生的后向散射系数来自地物体积散射，而体积散射通常捕捉地表纹理、湿度饱信息。此外Sentinel-1卫星*C*波段SAR数据来自长波脉冲，这导致了斑点噪声的高度污染而不易提取精细的土壤介电属性和纹理特征。

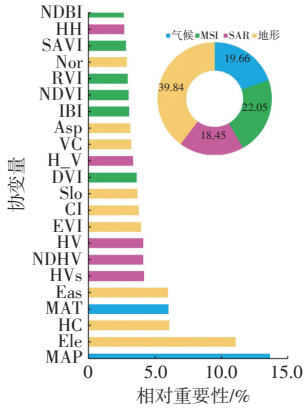


图5 环境变量对XGBoost模型的贡献

2.5 土壤水分反演结果分析

采用上述XGBoost模型反演得到研究区100 m像素水平的SM分布图（图6a）。反演结果显示京津冀地区SM出表现地带性分布，呈斑块状沿着经纬度、地形分布，总体为东高西低的格局。东部唐山、秦皇岛和天津等地SM含量较高，局部可达0.35以上，由于受到土地利用复杂变异以及地表侵蚀等干扰，导致空间连续性弱。区域中、南、北、西部地区SM含量为相对较低，多介于0.2~0.3，局部在0.1以下；该区地形由平原向土石山、丘陵、高原、荒漠变化，地表景观极具异质性，相应地引起了SM空间分布较大变异。本研究结果与Zhang[16]等综合多源SM产品反演得到的研究区同期1 km分辨率的SM数据不仅在空间具有相似性（图6b），而且在数值上接近（https://essd.copernicus.org/preprints/essd-2022-348/），这进一步证实了本研究反演的SM真实可信。

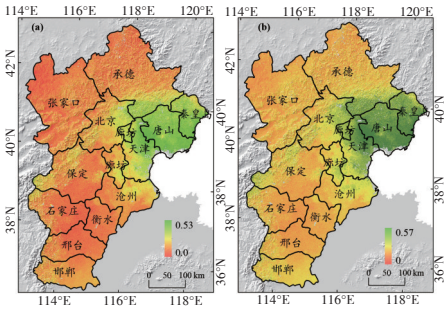


图6 京津冀地区土壤水分空间分布图

审图号：GS(2024)4468号

3 结 论

本文以175个地面实测土壤水分（SM）样点数据和经由Sentinel-1&2、DEM和气候等多源数据集提取得到的22个环境参数为基础，利用超参数优化后的XGBoost模型拟合SM—环境关系，反演京津冀地区SM空间分布。结论如下：

1）XGBoost模型可适用于SM监测应用中，构建的京津冀地区SM反演模型不存在过拟合，模型精度R^2为0.80（建模）和0.72（验证），RMSE为0.048（建模）、0.036（验证）。

2）作为结构参数复杂数据驱动模型，使用GridSearch能较好配置XGBoost模型超参数，进而确保模型拟合能力和稳健性。

3）京津冀地区SM与环境要素之间存在非线性关系，其中降水量和高程是SM空间分布主控因素。

参考文献

[1] Blanco C M G, Gomez V M B, Crespo P, et al. Spatial Prediction of Soil Water Retention in a Páramo Landscape: Methodological insight Into Machine Learning Using Random Forest[J]. Geoderma, 2018, 316:100-114

[2] 吴善玉，鲍艳松，李叶飞，等. 基于神经网络算法的Sentinel-1和Sentinel-2遥感数据联合反演土壤湿度研究[J]. 大气科学学报，2021，44(4):636-644

[3] Bhogapurapu N, Dey S, Mandal D, et al. Soil Moisture Retrieval over Croplands Using Dual-pol L-band GRD SAR data[J]. Remote Sensing of Environment, 2022, 271:112 900

[4] Wang Z, Zhao T, Shi J, et al. 1 km Soil Moisture Retrieval Using Multi-temporal Dual-channel SAR Data from Sentinel-1 A/B Satellites in a Semi-arid Watershed[J]. Remote Sensing of Environment, 2023, 284:113 334

[5] 王井利，高天娇，李荣平，等. 基于光学遥感和微波遥感的土壤水分反演模型研究[J]. 气象与环境学报，2023，39(1):73-82

[6] 艾小童. 基于ELM模型的极化SAR土壤水分降尺度研究[J]. 地理空间信息，2023，21(7):11-15

[7] Nativel S, Ayari E, Rodriguez-fernandez N. et al. Hybrid Methodology Using Sentinel-1/Sentinel-2 for Soil Moisture Estimation [J]. Remote Sensing. 2022, 14:2 434

[8] 王思楠，李瑞平，吴英杰，等. 基于环境变量和机器学习的土壤水分反演模型研究[J]. 农业机械学报，2022，53(5):332-341

[9] Han J, Mao K, Xu T, et al. A Soil Moisture Estimation Framework Based on the CART Algorithm and Its Application in China[J]. Journal of Hydrology, 2018, 563:65-75

[10] Ahmad S, Kalra A, Stephen H. Estimating Soil Moisture Us-

（下转第29页）