融合多源遥感与机器学习的太原市土壤全氮含量数据集（2020）研发

邵 馨[1,2]，杨 婷[1*]

1. 中国科学院地理科学与资源研究所黄河三角洲现代农业工程实验室，北京 100101；
2. 云南师范大学地理学部，昆明 650500

摘 要：土壤全氮含量是反映土壤养分水平与生态功能的重要指标，对农业生产、生态保护与环境安全具有重要意义。本研究依托 Google Earth Engine（GEE）云计算平台，融合多源遥感数据，选取了包括 MODIS NDVI 指数、Sentinel-2 近红外波段反射率、地表土壤水分、降水量、地表温度和数字高程模型（DEM）等关键环境因子作为输入变量，采用随机森林回归（Random Forest，RF）、分类回归树（Classification and Regression Tree，CART）和梯度提升回归树（Gradient Boosting Regression Tree，GBRT）三种机器学习模型，对土壤全氮含量进行反演，构建了 2020 年太原市土壤全氮含量数据。以国际土壤信息中心（ISRIC）通过 SoilGrids 项目提供的全球土壤全氮含量数据为参考，采用均方根误差（RMSE）和决定系数（R^2）作为评估指标，通过交叉验证，RF、CART、GBRT 的各层土壤全氮含量平均 RMSE 分别为 0.16 g/kg、0.21 g/kg、0.33 g/kg，平均 R^2 分别为 0.62、0.64、0.85。数据验证结果表明，该数据集具有较高的精度和可靠性，可为区域土壤养分评估、农业生产决策及生态环境管理提供科学支持。数据集内容为 2020 年太原市多层（包括 6 个深度层次：0–5 cm、5–15 cm、15–30 cm、30–60 cm、60–100 cm 与 100–200 cm）土壤全氮含量数据，空间分辨率为 30 m，以.tif 格式存储，共 18 个数据文件，数据总量为 1.52 GB（压缩为 1 个文件，219 MB）。

关键词：GEE；土壤全氮；多源遥感数据；机器学习模型

DOI: https://doi.org/10.3974/geodp.2025.03.08

CSTR: https://cstr.escience.org.cn/CSTR:20146.14.2025.03.08

数据可用性声明：

本文关联实体数据集已在《全球变化数据仓储电子杂志（中英文）》出版，可获取：https://doi.org/10.3974/geodb.2025.04.01.V1 或 https://cstr.escience.org.cn/CSTR:20146.11.2025.04.01.V1.

收稿日期：2025-05-06；**修订日期**：2025-07-21；**出版日期**：2025-09-25

基金项目：中华人民共和国科学技术部（2023YFD1701804）

***通讯作者**：杨婷，中国科学院地理科学与资源研究所，yangt@igsnrr.ac.cn

数据引用方式：[1] 邵馨，杨婷. 融合多源遥感与机器学习的太原市土壤全氮含量数据集（2020）研发[J]. 全球变化数据学报，2025，9(3)：323–330. https://doi.org/10.3974/geodp.2025.03.08. https://cstr.escience.org.cn/CSTR:20146.14.2025.03.08.

[2] 邵馨，杨婷. 融合多源遥感与机器学习的太原市多层土壤全氮含量数据集（2020年）[J/DB/OL]. 全球变化数据仓储电子杂志，2025. https://doi.org/10.3974/geodb.2025.04.01.V1. https://cstr.escience.org.cn/CSTR:20146.11.2025.04.01.V1.

1 前言

土壤是大多数陆地生命的基础，展现出独特的复杂性和动态特征，其营养成分对维持生态平衡和促进自然发展起到了重要作用[1]。氮是植物生长所必需的重要矿物质元素，对于土壤肥力和植物生长具有重要的影响。土壤全氮含量是衡量土壤氮储量的重要指标，是评价土壤肥力水平的重要指标之一，直接影响农作物的产量和质量[2–4]。

传统的土壤全氮监测方法主要依赖地面采样与化学分析，虽然精度较高，但由于样本数量、时间成本和空间代表性等因素的限制，难以满足大尺度、高分辨率动态监测的需求[5,6]。随着遥感技术的发展[7–9]，结合机器学习模型的应用为构建区域尺度土壤氮含量数据集提供了新的路径。通过整合多源遥感数据，利用随机森林、梯度提升树等非线性回归算法，能够实现对土壤氮含量的空间化反演[5,10]。这些方法不仅提高了土壤氮素监测的效率和精度，还为土壤管理和农业决策提供了科学依据。

采用 GEE（Google Earth Engine）平台的提取可以极大提高遥感影像处理的计算效率和时间效率[11]，为海量遥感大数据的快速处理提供了机遇[12]。在此背景下，本文依托 GEE 云计算平台，融合多源遥感数据与主流机器学习算法，构建了 2020 年太原市土壤全氮含量空间分布数据集。该数据集覆盖 0–200 cm 深度范围的 6 个土层，空间分辨率 30 m，为高质量耕地资源调查与区域农业信息化管理提供了基础支撑。

2 数据集元数据简介

《融合多源遥感与机器学习的太原市多层土壤全氮含量数据集（2020 年）》[13]的名称、作者、地理区域、数据年代、时间分辨率、空间分辨率、数据集组成、数据出版与共享服务平台、数据共享政策等信息见表 1。

3 数据研发方法

3.1 数据来源

本研究所需相关数据及来源包括：（1）NDVI 指数（AVHRR NDVI 长时间序列数据集，16 天合成，约 5.1 km 分辨率）[15]；（2）Sentinel-2 近红外波段反射率（Level-2A 产品，b8 波段，10 m 分辨率）[16]；（3）地表土壤水分（OpenLandMap 土壤水分-33 kPa（b10 波段），约 250 m 分辨率）[17]；（4）降水量（CHIRPS 数据集，0.05° 分辨率，约 5.6 km）[18]；（5）地表温度（MOD11A1 数据集，白天地表温度 LST_Day_1 km 波段，1 km 分辨率）[19]；（6）数字高程模型（SRTM DEM 数据集，30 m 分辨率）[20]；（7）表层氮含量数据（SoilGrids）[21]。

3.2 算法原理

3.2.1 随机森林回归

随机森林回归（Random Forest，RF）是一种集成学习方法，通过构建多个决策树并结合其输出结果来提高预测精度[22,23]。RF 的基本思想是利用“投票”机制，选择多个随机抽样的子集进行训练，从而减少单棵决策树的过拟合风险。在本研究中，RF 模型通过对整合

表 1　《融合多源遥感与机器学习的太原市多层土壤全氮含量数据集（2020 年）》元数据简表

条　目	描　述
数据集名称	融合多源遥感与机器学习的太原市多层土壤全氮含量数据集（2020 年）
数据集短名	TY_SoilN2020
作者信息	邵馨，云南师范大学地理学部，2323130115@ynnu.edu.cn 杨婷，中国科学院地理科学与资源研究所，yangt@igsnrr.ac.cn
地理区域	太原市
数据年代	2020 年
时间分辨率	年
空间分辨率	30 m
数据格式	.tif
数据量	1.52 GB（压缩后为 219 MB）
数据集组成	2020 年太原市多层土壤全氮含量数据
基金项目	中华人民共和国科学技术部（2023YFD1701804）
数据计算环境	GEE、ArcGIS
出版与共享服务平台	全球变化科学研究数据出版系统 http://www.geodoi.ac.cn
地址	北京市朝阳区大屯路甲 11 号 100101，中国科学院地理科学与资源研究所
数据共享政策	（1）“数据”以最便利的方式通过互联网系统免费向全社会开放，用户免费浏览、免费下载；（2）最终用户使用“数据”需要按照引用格式在参考文献或适当的位置标注数据来源；（3）增值服务用户或以任何形式散发和传播（包括通过计算机服务器）“数据”的用户需要与《全球变化数据学报（中英文）》编辑部签署书面协议，获得许可；（4）摘取“数据”中的部分记录创作新数据的作者需要遵循 10%引用原则，即从本数据集中摘取的数据记录少于新数据集总记录量的 10%，同时需要对摘取的数据记录标注数据来源[14]
数据和论文检索系统	DOI，CSTR，Crossref，DCI，CSCD，CNKI，SciEngine，WDS，GEOSS，PubScholar，CKRSC

的多源遥感数据进行训练，自动学习不同环境因子与土壤氮含量之间的复杂关系，最终输出土壤氮含量的预测值。

3.2.2　分类回归树

分类回归树（Classification and Regression Tree，CART）是一种非参数的统计方法，采用二叉树结构，根据具体的规则对节点进行分枝，同时为了获得更高的精度在决策树生长的时候进行必要的剪枝，通过对子树的评估，获得平均误分代价最小的最终决策树[24,25]。因为自身方法的实现快速、简单、分类准确，现已在遥感影像分类中获得了大规模的应用。

3.2.3　梯度提升回归树

梯度提升回归树（Gradient Boosting Regression Tree, GBRT）是一种基于决策树的提升算法，通过逐步构建弱分类器并结合其结果来提高模型的整体性能[26]。GBRT 通过最小化损失函数（如均方误差）来优化模型参数，逐步调整预测结果以提高准确性。

3.3　技术路线

本研究的技术路线如图 1 所示。基于收集到的 2020 年多源数据，对收集的数据进行预处理，包括数据清洗、格式转换和空间分辨率统一。随后，从数据中提取相关特征，并使

用 RF、CART 和 GBRT 方法建立模型，以选定因子作为训练数据进行训练，最后，输出不同深度的土壤全氮含量数据集。

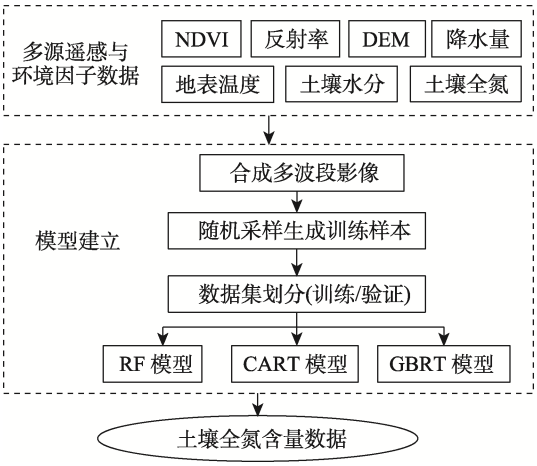

图 1 土壤全氮含量数据集研发技术路线图

4 数据结果与验证

4.1 数据集组成

数据集存储为.tif 格式，由 18 个数据文件组成，分别是 2020 年随机森林回归、分类回归树和梯度提升回归树 3 种机器学习模型对应的 6 个深度层次，包括 0–5 cm、5–15 cm、15–30 cm、30–60 cm、60–100 cm 和 100–200 cm 的土壤全氮含量。数据集的空间分辨率为 30 m。

4.2 数据结果

图 2 展示了 2020 年太原市多层土壤全氮含量的空间分布情况。总体来看，土壤全氮含量随土层加深逐渐减少，高值主要集中在 0–5 cm 表层，而 100–200 cm 深层土壤全氮含量普遍较低，低于 0.5 g/kg，体现了地表有机质输入与养分积累的规律。

在空间分布上，土壤全氮含量较高的区域主要分布在阳曲北部丘陵、古交矿区及西部吕梁山区等地形起伏大、有植被覆盖或人为干预较少的区域。这些地区具有较多的枯落物堆积和植被残体，是有机质和氮素积累的重要来源。尤其是在古交矿区，煤炭资源开发虽造成局部土地破坏，但植被恢复区肥力输入相对较高。相反，土壤全氮含量较低的区域集中在太原盆地南部及汾河冲积平原等地，这些区域多为集约农业区，耕作强度大、氮素流失严重，并受限于冲积物母质贫氮性与人为扰动等因素，全氮水平偏低。

4.3 数据结果验证

为验证土壤全氮数据集的精度与可靠性，本研究选取国际土壤信息中心（ISRIC）SoilGrids 项目提供的全球土壤全氮含量数据作为对照基准，采用交叉验证方法，并以均方根误差（Root Mean Square Error，RMSE）与决定系数（Coefficient of Determination，R^2）作为评估指标，系统比较不同模型在不同土层深度下的预测表现，具体结果见表 2。

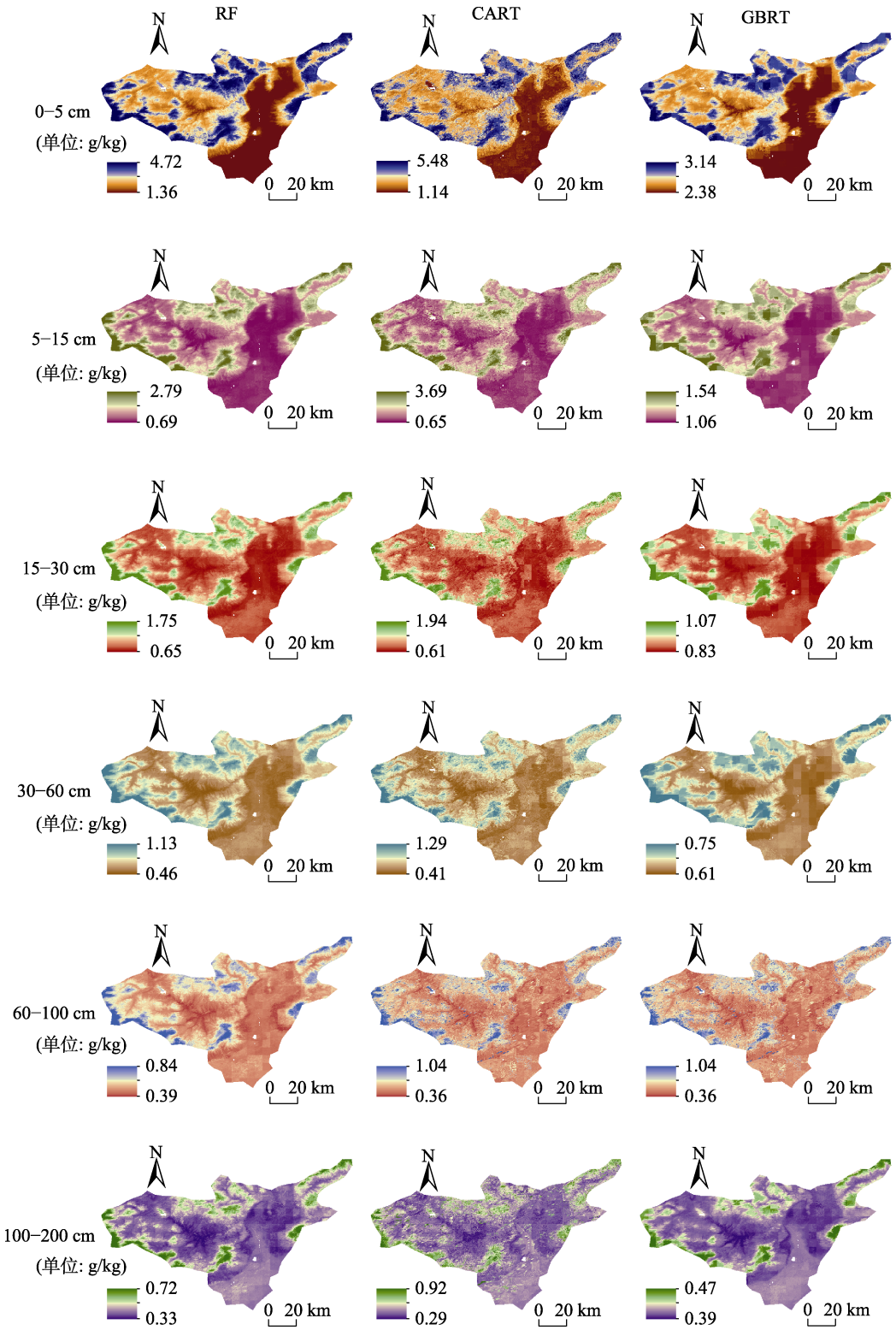

图 2　太原市多层土壤全氮含量空间分布图（2020）

表 2 太原市土壤全氮含量数据不同模型预测精度评价结果统计表

土壤深度（cm）	RF		CART		GBRT	
	RMSE (g/kg)	R^2	RMSE (g/kg)	R^2	RMSE (g/kg)	R^2
0–5	0.40	0.75	0.52	0.75	0.85	0.91
5–15	0.21	0.79	0.28	0.80	0.50	0.91
15–30	0.12	0.73	0.15	0.78	0.26	0.90
30–60	0.08	0.75	0.10	0.77	0.16	0.89
60–100	0.08	0.39	0.10	0.38	0.11	0.76
100–200	0.07	0.32	0.09	0.35	0.10	0.73

从整体趋势看，模型在浅层土壤（0–60 cm）中的预测性能优于深层（60–200 cm），表现在较高的 R^2 值。例如，在 0–60 cm 层，模型的 R^2 均超过 0.73，表明对该层氮素空间分布的拟合效果较好；而在 100–200 cm 层，最低 R^2 降至 0.32，显示出深层预测误差显著增大，可能与土壤异质性增强等因素有关。

从模型表现来看，3 种机器学习模型（RF、CART、GBRT）在不同土层和区域中的预测能力存在差异：RF 整体稳定，尤其在 0–60 cm 范围内表现优异（R^2 为 0.73–0.79），反映其对异常值的鲁棒性与对多维特征的综合感知能力；但在深层（100–200 cm）预测性能明显下降（R^2 降至 0.32），显示其泛化能力受限。CART 在表层（0–5 cm）某些区域存在预测偏高的现象，尽管 R^2 可达 0.75，但 RMSE 为 0.52 g/kg，且存在过拟合风险。这可能与 CART 对输入变量组合高度敏感、易受样本分布不均或极端值干扰有关。GBRT 在各层土壤中整体 R^2 值高于其他模型，在表层表现最优（R^2 高达 0.91），但对应 RMSE 高达 0.85 g/kg，表明其对高变异性区域存在“过度响应”，易高估氮含量峰值。

此外，不同模型对特征变量的响应机制差异亦显著影响其预测表现：CART 对高频扰动变量（如 NDVI、地表温度）较为敏感，易引发极值偏差；RF 对局部异常值具有较强容忍性，但可能低估局地高值；GBRT 依赖残差迭代机制捕捉复杂非线性结构，对模型参数与地形因子（如 DEM）较为敏感，在地形起伏较大区域表现更依赖 DEM 等变量的精度。

综上所述，不同模型在不同深度与区域的适用性存在差异，提示研究者在开展土壤全氮反演时，应结合区域特征与预目标选择最合适的模型。本研究构建的数据集在 0–60 cm 深度内已达到较高的预测精度（R^2>0.70，RMSE<0.5 g/kg），具有良好的科学适用性与推广价值。

5 讨论和总结

本研究基于 GEE 平台，融合 6 类遥感数据构建了太原市土壤全氮含量的高分辨率空间分布数据集。该数据集采用多源遥感协变量驱动的回归建模方法，在 30 m 空间分辨率下显著提升了对农田边界区（如汾河平原水稻田）和矿区复垦地（如古交煤矿区）土壤氮含量的表达能力。与 SoilGrids 全球数据集相比，本研究成果更准确地刻画了区域尺度下土壤全氮的空间梯度变化，尤其在土地利用结构复杂、人为干扰强烈的异质景观中表现出更高的

空间精度与代表性，验证了区域尺度多源数据融合建模的可行性与必要性。

在土壤剖面尺度上，太原市土壤全氮含量垂向分布呈显著表聚特征：0–30 cm 层富集明显，主要归因于地表有机质输入、高强度人为干预及物理-生物过程耦合作用；而深层氮含量随深度递增而减少的趋势，则受控于有机质输入衰减、微生物活性分异、淋溶-黏土屏障作用及根系-人为干扰深度限制等多尺度机制的共同驱动。该垂向分异规律为农业面源污染管控和耕作区氮素精准施用提供了理论支持。

尽管如此，本数据集仍存在一定局限性：其建模体系主要受制于单一年度观测和表层遥感协变量的主导性，尚难以全面反映深层土壤理化属性（如 pH、CEC、黏粒含量）及其年际动态变化。未来应结合原位传感器网络、氮循环过程模型与多时相遥感信息，构建时空连续、深浅兼顾的土壤氮监测体系，并进一步引入具物理约束的深度学习模型，以增强在矿区、冲积平原等异质地貌中的泛化能力与可转移性。

作者分工：杨婷对数据集的开发做了总体设计；邵馨收集和处理了多源遥感数据并撰写了数据论文。

利益冲突声明：本研究不存在研究者以及与公开研究成果有关的利益冲突。

参考文献

[1] 庞煜龚, 张孟豪, 姜敏等. 广东省肇庆市高要区耕地土壤理化性质和微生物特征的空间异质性及综合质量评价[J]. 华南农业大学学报, 2025, 46(2): 151–163.

[2] Chapin, F. S., Matson, P. A., Mooney, H. A. Principles of Terrestrial Ecosystem Ecology [M]. Berlin: Springer, 2011.

[3] Htwe, N. M. P. S., Ruangrak, E. A review of sensing, uptake, and environmental factors influencing nitrate accumulation in crops [J]. *Journal of Plant Nutrition*, 2021, 44(3):1–12.

[4] 刘丽琪, 魏广源, 周萍. 基于机器学习优化建模的 GF-5 影像土壤总氮量预测填图[J]. 智慧农业, 2024, 6(5): 61–73.

[5] 宋雪, 张民, 周洪印等. 基于土壤优化光谱参数估测太湖地区土壤全氮含量[J]. 农业资源与环境学报, 2020, 37(1): 43–50. https://doi.org/10.13254/j.jare.2018.0365.

[6] 章海亮, 谢潮勇, 田彭等. 基于可见/近红外光谱和数据驱动的机器学习方法测量土壤有机质和全氮[J]. 光谱学与光谱分析, 2023, 43(7): 2226–2231.

[7] 赵春江. 农业遥感研究与应用进展[J]. 农业机械学报, 2014, 45(12): 277–293.

[8] 聂鹏程, 钱程, 覃锐苗等. 天空地一体化信息感知与融合技术发展现状与趋势[J]. 智能化农业装备学报, 2023, 4(2): 1–11.

[9] Zhang, S., Zhang, J. H., Bai, Y., *et al*. Evaluation and improvement of the daily Boreal Ecosystem Productivity Simulator in simulating gross primary productivity at 41 flux sites across Europe [J]. *Ecological Modelling*, 2018, 368: 205–232. https://doi.org/10.1016/j.ecolmodel.2017.11.023.

[10] 杨子, 潘鑫, 袁洁等. 基于随机森林算法的卫星监测太湖蓝藻数据集(2019)[J]. 全球变化数据学报, 2023, 7(3): 321–326. https://doi.org/10.3974/geodp.2023.03.11. https://cstr.escience.org.cn/CSTR:20146.14.2023.03.11.

[11] 潘霞. 基于 Google Earth Engine 云平台下地物覆被类型的遥感影像智能分类方法研究[D]. 呼和浩特: 内蒙古农业大学, 2021.

[12] 遥感云计算平台发展及地球科学应用[OL]. https://d.wanfangdata.com.cn/Periodical/ygxb202101014

(accessed on 13 April 2025).

[13] 邵馨, 杨婷. 融合多源遥感与机器学习的太原市多层土壤全氮含量数据集(2020年)[J/DB/OL]. 全球变化数据仓储电子杂志, 2025. https://doi.org/10.3974/geodb.2025.04.01.V1. https://cstr.escience.org.cn/CSTR:20146.11.2025.04.01.

[14] 全球变化科学研究数据出版系统. 全球变化科学研究数据共享政策[OL]. https://doi.org/10.3974/dp.policy.2014.05 (2017年更新).

[15] NOAA National Climatic Data Center. NOAA Climate Data Record (CDR) of AVHRR NDVI, Version 5 [DB/OL]. 2020. https://developers.google.com/earth-engine/datasets/catalog/NOAA_CDR_AVHRR_NDVI_V5.

[16] 欧洲航天局. Sentinel-2 [OL]. https://scihub.copernicus.eu/dhus/#/home.

[17] Hengl, T., Gupta, S. OpenLandMap soil moisture at 33 kPa [DB/OL]. 2017. https://developers.google.com/earth-engine/datasets/catalog/OpenLandMap_SOL_SOL_WATERCONTENT-33KPA_USDA-4B1C_M_v01.

[18] Funk, C., Peterson, P., Landsfeld, M., *et al*. CHIRPS daily precipitation data [DB/OL]. 2015. https://developers.google.com/earth-engine/datasets/catalog/UCSB_CHG_CHIRPS_DAILY.

[19] NASA. MODIS terra land surface temperature and emissivity daily L3 global 1 km SIN grid V006 (MOD11A1) [DB/OL]. 2020. https://developers.google.com/earth-engine/datasets/catalog/MODIS_006_MOD11A1.

[20] USGS. SRTMGL1 global 30 m DEM (Version 003) [DB/OL]. 2000. https://developers.google.com/earth-engine/datasets/catalog/USGS_SRTMGL1_003.

[21] ISRIC—World Soil Information. SoilGrids: global gridded soil information (Nitrogen) [DB/OL]. 2020. https://developers.google.com/earth-engine/datasets/catalog/projects_soilgrids-isric_nitrogen_mean.

[22] Prasad, A. M., Iverson, L. R., Liaw, A. Newer classification and regression tree techniques: bagging and random forests for ecological regression [J]. *Ecosystems*, 2006, 9(2): 181–199.

[23] Breiman, L. Random forests [J]. *Machine Learning*, 2001, 45(1): 5–32.

[24] 王大鹏, 王周龙, 李德一等. 综合非光谱信息的荒漠化土地 CART 分类[J]. 遥感学报, 2007, 11(4): 487–492.

[25] Breiman, L., Friedman, J. H., Olshen, R. A., *et al*. Classification and Regression Trees [M]. Belmont: Wadsworth International Group, 1984.

[26] Friedman, J. H. Greedy function approximation: a gradient boosting machine [J]. *Annals of Statistics*, 2001, 29(5): 1189–1232.