ICS 07.040
CCS B10

T/SSSC

中 国 土 壤 学 会 团 体 标 准

T/SSSC 046—2026

红壤长期植被覆盖区土壤有机质含量时空变化遥感监测技术规程

Technical Code of Practice for Remote Sensing Monitoring of Spatiotemporal Variations in Soil Organic Matter Content in Long-Term Vegetation-Covered Red Soil Regions

2026-11-27 发布　　2026-12-15 实施

中国土壤学会 发 布

目　　次

前　言 III
1 范围 1
2 规范性引用文件 1
3 术语、定义和缩略语 1
4 监测流程 3
5 数据获取 4
6 指标提取与筛选 6
7 有机质含量遥感监测模型构建及验证 7
8 有机质含量时空变化评价 8
9 监测报告 8
附录 A （资料性） 遥感指标计算 9
附录 B （资料性） 基于随机森林进行特征重要性排序算法 13
附录 C （资料性） 卷积神经网络算法 14
参考文献 17

前　言

本文件按照 GB/T 1.1—2020《标准化工作导则 第 1 部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由中国土壤学会提出并归口。

本文件起草单位：中国科学院南京土壤研究所、南京工业大学、南京大学。

本文件主要起草人：马利霞、于东升、解宪丽、刘杰、刘明、张乾、关廷宇、郑光。

红壤长期植被覆盖区土壤有机质含量时空变化遥感监测技术规程

1 范围

本文件确立了红壤长期植被覆盖区土壤有机质含量时空变化遥感监测的技术规程，包括术语和定义、缩略语、监测流程、数据获取、指标提取与筛选、有机质含量遥感监测模型构建及验证、有机质含量时空变化评价、监测报告。

本文件适用于区域尺度红壤长期植被覆盖区土壤有机质含量时空变化监测。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 1.1—2020 标准化工作导则 第 1 部分：标准化文件的结构和起草规则

HJ 166—2026 土壤环境监测技术规范 第 7 部分：样品采集

HJ 166—2026 土壤环境监测技术规范 第 8 部分：样品制备

GB/T 47361-2026 土壤有机质的测定

HJ 1068 土壤粒度的测定吸液管法和比重计法

3 术语、定义和缩略语

3.1 术语和定义

下列术语和定义适用于本文件。

3.1.1

砂粒含量 Sand content

土壤中粒径小于2㎜大于等于0.05㎜的矿物质颗粒所占的比例。

3.1.2

粉粒含量 Silt content

土壤中粒径小于0.05mm大于等于0.002mm的矿物质颗粒所占的比例。

3.1.3

黏粒含量 Clay content

土壤中粒径＜2 μm 的矿物质颗粒所占的比例。

3.1.4

训练样本 Training sample

用于遥感监测模型训练，通过遥感影像特征与对应的实测土壤属性建立或优化模型参数的样本。

3.1.5

验证样本 Validation sample

用于评价遥感监测模型的预测性能，通过比较模型预测的土壤有机质含量与对应的实测值，评估模型监测精度的样本。

3.1.6

红壤长期植被覆盖区 Long-term vegetation-covered regions

红壤区内，目标年份前连续5年，年内地表无植被覆盖期均不足2个月，导致难以找到有效的裸土期遥感影像。

3.1.7

长时序遥感影像 Long-term time series of remote sensing images

在较长时间范围内（例如≥10年），按照一定时间间隔获取或整理形成的多期遥感影像数据集合，用于分析地表参数的长期变化特征。

3.1.8

有机质含量时空变化 Spatiotemporal variation of soil organic matter content

在一定区域和时间范围内，基于不同时期的土壤有机质含量监测结果，表征土壤有机质含量的时间变化特征及其空间分布差异。

3.1.9

遥感监测模型 Remote sensing monitoring model

利用遥感影像及相关辅助数据，通过建立遥感特征与土壤有机质含量之间的定量关系，实现土壤有机质含量估算的数学模型或算法。

3.2 缩略语

下列缩略语适用于本文件。

GPS：Global Positioning System，全球定位系统

CNN：Convolutional Neural Network，卷积神经网络

MAE：Mean Absolute Error，平均绝对误差

SPXY：基于 X–Y 距离的样本集划分方法

R2：R square，决定系数

RMSE：Root-Mean-Square Error，均方根误差

4 监测流程

红壤长期植被覆盖区土壤有机质含量时空变化遥感监测主要包括数据获取、指标提取与筛选、有机质含量遥感监测模型构建及验证、有机质含量时空变化评价4个步骤（图1）。

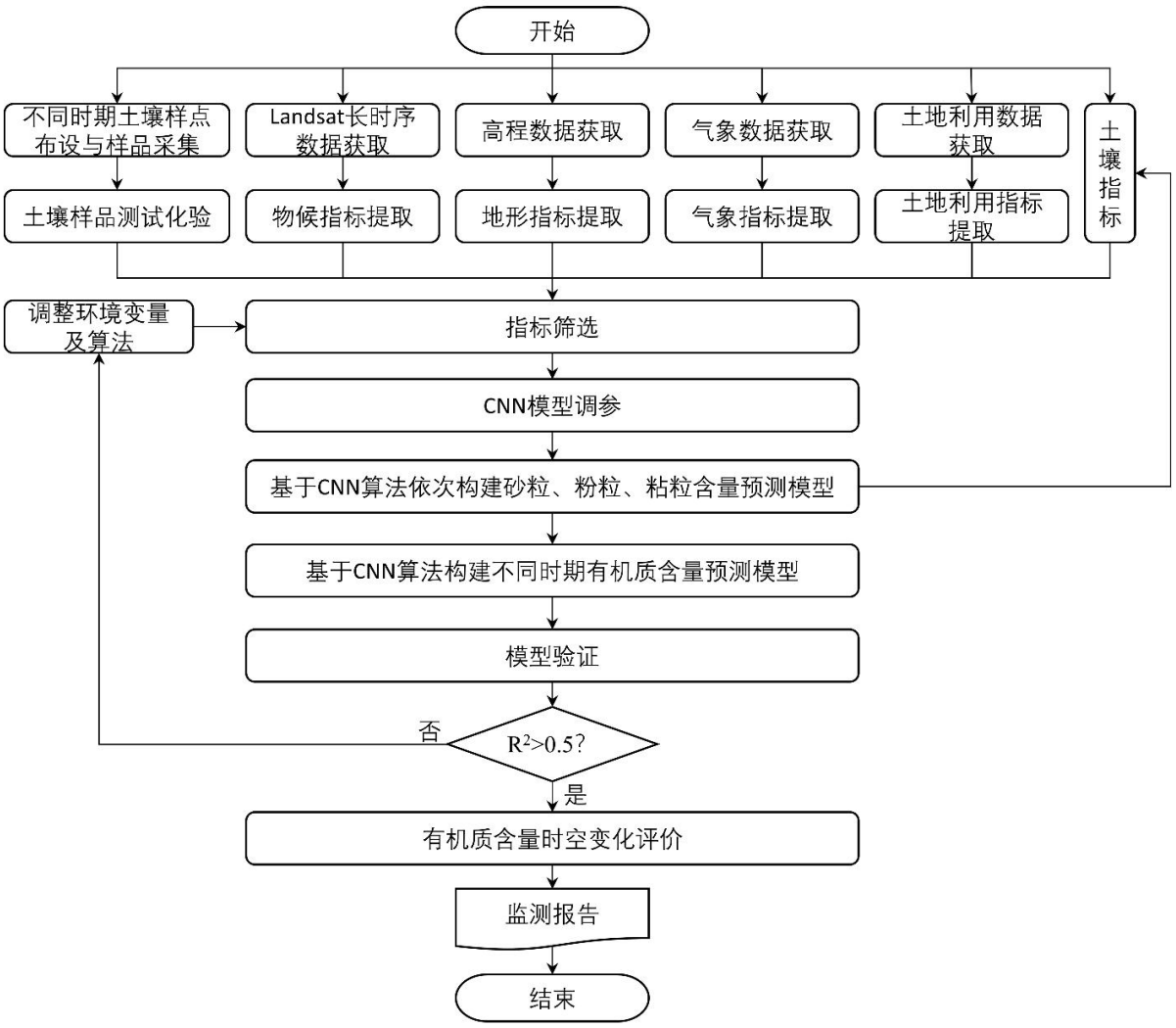


图1 有机质含量时空变化遥感监测流程图

5 数据获取

5.1 土壤样品采集与化验

5.1.1 土壤样点布设

应综合考虑土地利用类型（森林、农田、草地等）、土壤类型及其空间分布特征，采用分层布点方式确定样点位置。应根据各土地利用类型和土壤类型的面积占比，合理分配各类型的样点数量。对于面积占比较大或土壤有机质含量空间变异较强的类型，可适当增加样点数量。各类型样点应兼顾不同坡位的分布，避免样点过度集中于单一坡位。应结合监测区面积、样点总数及空间分布特征进行分区布点，提高样点的空间代表性和分布均匀性。样点应具有代表性，宜避开道路、沟渠、建筑物及其他可能影响土壤有机质含量代表性的特殊地段。

5.1.2 土壤样品采集

参考HJ 166—2026进行土壤样品采集，不同时期通过多点混合采样方法采集0 - 20 cm深度中间部分的土壤样品，质量约500 g。样品密封于塑料自封袋并标注点位信息，同时使用全球定位系统GPS记录采样坐标。

5.1.3 土壤样品测试化验

按照HJ 166—2026规定的方法进行土壤样品的处理和贮存，按照HJ 1068-2019规定的方法化验土壤砂粒、粉粒、黏粒含量，按照GB/T 47361-2026规定的方法化验土壤有机质含量。

5.2 遥感数据获取

5.2.1 光学遥感数据

基于完成辐射定标、大气校正、几何校正等预处理后的光学遥感影像，选取目标年份及前连续三年的影像，并依据监测区边界进行裁剪和掩膜处理。影像需包含蓝光、绿光、红光、近红外、短波红外、热红外波段，空间分辨率不低于30米，至少每16天一景有效影像。综合考虑到长时序遥感影像的可获得性、时间覆盖和空间分辨率，推荐采用Landsat数据集[1]。

5.2.2 高程数据

选取空间分辨率≤30米的高程数据。

5.3 其他数据

5.3.1 气象数据

通过中国国家地球系统科学数据中心（https://www.geodata.cn/data/）下载目标年份逐月温度和降雨数据集。

5.3.2 土地利用数据

可通过中国国家地球系统科学数据中心（https://www.geodata.cn/data/）下载基于Landsat数据通过目视解译得到的土地利用空间分布。也可自行通过目标年份的Landsat数据集进行相关目视解译。提取不同土地利用（林地、草地、水田、旱地）的空间分布数据。

6 指标提取与筛选

6.1 指标提取

基于遥感数据等提取相关指标，包括气象、地形、土地利用、物候及土壤指标。

6.1.1 气象指标

针对目标年份，提取温度和降水的月均数据作为气象指标。

6.1.2 地形指标

基于高程数据提取14个基础地形指标，包括高程值、坡度、坡向、地形湿度指数、坡长坡度因子、分析性山体阴影、河道网络基准面、河道网络距离、闭合洼地、汇流指数、平面曲率、剖面曲率、相对坡度位置、山谷深度。具体参见附录A。

6.1.3 土地利用指标

基于样点土壤属性变量大小对土地利用进行编号（例如旱地1、草地2、林地3、水田4），作为土地利用指标。

6.1.4 物候指标

基于光学反射率数据提取两类物候指标，即年内物候指标和年际物候指标。具体参见附录A。

6.1.5 土壤指标

预测有机质含量的土壤指标包括土壤颗粒组成（模型预测的砂、粉、黏粒含量）。为保证颗粒组成之间的一致性，砂、粉、黏粒量采用逐步建模策略：首先，预测粉粒含量；其次，以预测粉粒为土壤特征预测砂粒含量；随后，计算初始黏粒含量（即1-粉粒-砂粒），并结合预测的粉粒和砂粒含量为主要土壤特征进一步预测黏粒含量。最终，将预测的粉粒、砂粒、以及初始和预测的黏粒含量作为土壤特征共同用于有机质含量预测模型。其中，砂粒、粉粒和黏粒含量的预测顺序可根据各组分的预测精度及不同预测顺序下的整体建模效果进行调整，优先预测精度较高的组分，并将其预测结果作为后续组分预测模型的输入特征。

6.2 特征筛选

为了减少冗余以及多重共线性，通过两步来对特征进行选择。首先，通过对所有变量进行皮尔逊相关分析，对于两个变量相关性较高（相关性超过0.98），选择删除其中一个。对于剩余的所有变量通过随机森林算法进行重要性排序，具体参见附录B。进而计算重要性累计曲线，设置阈值选择用于土壤属性建模的特征。其中阈值作为一个参数通过与模型参数共同调参确定。

7 有机质含量遥感监测模型构建及验证

7.1 模型调参

通过一维卷积神经网络（CNN）算法构建基于遥感指标预测目标土壤属性的模型。CNN模型架构包括两个卷积层和丢弃层，后接一个全局平均池化层和全连接输出层。每个卷积层都需设置滤波器、卷积核，丢弃层需设定丢弃率参数。所有层均采用ReLU激活函数。模型通过Adam优化器进行编译，需设置学习率参数。此外，为平衡训练速度和稳定性，需设定批量处理大小参数。模型所有参数通过贝叶斯优化进行调优，具体参数和调优区间见附录C。参数的性能可通过交叉验证进行评估，重点关注较低的训练集平均绝对误差（*MAE*）。

7.2 模型构建

先对土壤砂粒、粉粒、黏粒含量进行模型构建。将筛选特征作为输入变量，采用CNN算法依次建立土壤砂粒、粉粒、黏粒含量的模型。将其模型预测值作为特征变量加入到环境变量中，再次采用CNN算法分别建立T1和T2不同时期土壤有机质含量预测模型。CNN算法具体参见附录C。

7.3 模型验证

通过基于 X–Y 距离的样本集划分方法（SPXY）抽取70%的训练样本和30%的验证样本，具体参见附录D。基于验证样本对各属性建模结果进行精度评价，选择决定系数（R^2）、均方根误差(RMSE)作为精度评价指标。按照公式（1）计算R^2，按照公式（2）计算RMSE。原则上土壤有机质含量的决定系数R^2应大于0.5，精度验证不合格的，需要调整输入变量或算法，直至满足精度要求。

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

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad \text{.....(2)}$$

式中，

R^2—决定系数；

RMSE—均方根误差；

n—采样点数量；

$\hat{y}_i$ —土壤样点i的预测值；

y_i —土壤样点i的实测值；

$\bar{y}$ —土壤样点验证样本实测值的平均值。

8 有机质含量时空变化评价

根据不同时期样点构建有机质含量预测模型，进而分别生成各个时期有机质含量空间分布图，通过后一时期（T2）结果减去前一时期（T1）结果得到有机质含量在目标期间的变化结果（ΔRMSE）。考虑两个时期制图误差RMSE设定分级阈值σ：

$$\sigma = 1.96 \times \sqrt{RMSE_{T1}^2 + RMSE_{T2}^2} \quad \text{.....(3)}$$

将有机质含量变化分为三个级别，包括“显著降低”（ΔRMSE<-σ）、“基本不变”（-σ≤ΔRMSE≤σ）、“显著增加”（ΔRMSE>σ）。

9 监测报告

有机质含量时空变化监测报告包括采用的遥感数据和样点数据、监测技术流程、有机质含量制图结果和精度评价、有机质含量时空变化评价等信息。监测结果形式宜采用统计表格和图表等，统计表格包括样点特征等信息。图片包括土壤样点分布和有机质含量变化分级分布图片等。

附录 A

（资料性）

遥感指标计算

表A.1 地形指标

指标	含义
高程值	高程数据的值Z，直接由DEM提取
坡度	$Slope = \arctan(\sqrt{(\frac{\partial Z}{\partial x}) + (\frac{\partial Z}{\partial y})})$
坡向	$Aspect = \arctan 2(\frac{\partial Z}{\partial y}, \frac{\partial Z}{\partial x})$
地形湿度指数	$TWI = \ln(\frac{SCA}{\tan(\beta)})$ SCA为汇水面积，β为局部坡度
坡长坡度因子	$LS = (\frac{SCA}{22.13})^{0.4} \times (\frac{\sin\beta}{0.0896})^{1.3}$
分析性山体阴影	$Hillshade = 255 \times (cos(slope) \times cos(90° - altitude) + sin(slope) \times sin(90° - altitude) \times cos(azimuth - aspect))$
河道网络基准面[2]	基于河道网络栅格，采用张力样条插值对河道单元格的高程值进行插值，生成整个区域的基准面高程。
河道网络距离	$D_{channel} = Z - Z_{baselevel}$ $Z_{baselevel}$表示插值的河道基准面高程。
闭合洼地[3]	通过洼地填充算法识别，原始DEM与填充后DEM的差值即为洼地深度。
汇流指数[4]	采用D8算法计算每个栅格单元的汇流累积量，表示上游汇流区域内的栅格数量。
平面曲率	$Plan_curvature = \frac{\partial^2 Z}{\partial x^2}(\frac{\partial Z}{\partial y})^2 - 2\frac{\partial^2 Z}{\partial x \partial y}\frac{\partial Z}{\partial x}\frac{\partial Z}{\partial y} + \frac{\partial^2 Z}{\partial y^2}(\frac{\partial Z}{\partial x})^2$ 正值表示汇聚流，负值表示发散流。
剖面曲率	$Plan_curvature = \frac{\frac{\partial^2 Z}{\partial x^2}(\frac{\partial Z}{\partial x})^2 + 2\frac{\partial^2 Z}{\partial x \partial y}\frac{\partial Z}{\partial x}\frac{\partial Z}{\partial y} + \frac{\partial^2 Z}{\partial y^2}(\frac{\partial Z}{\partial y})^2}{(\frac{\partial Z}{\partial x})^2 + (\frac{\partial Z}{\partial y})^2}$ 沿坡度方向的地表曲率，正值表示凸坡，负值表示凹坡。
相对坡度位置	$RSP = \frac{HO - HU}{2}$ HO为坡高，HU为谷深
山谷深度	$Valley_Depth = Z_{ridge} - Z$

表A.2 年内物候指标

类型	指标	含义
统计值	蓝光波段反射率 Blue	**统计值:** 最小值 最大值 第二小值 第二大值 平均值 最小值与第一四分位数之间的平均值 第三四分位数与最大值之间的平均值 第一四分位数与第三四分位数之间的平均值 除去最大值和最小值后的平均值 中值 标准差 各数据区间的差异值 最后一次观测值
	绿光波段反射率 Green	
	红光波段反射率 Red	
	近红外波段反射率 Nir	
	短波红外波段反射率 1 Swir1	
	短波红外波段反射率 2 Swir2	
	归一化植被指数 NDVI=(Nir-Red)/(Nir+Red)	
	归一化指数 S1N=(Nir-Swir1)/(Nir+Swir1)	
	归一化指数 S2N=(Nir-Swir2)/(Nir+Swir2)	
	归一化指数 GN=(Green-Red)/(Green+Red)	
	归一化指数 S1S2=(Swir1-Swir2)/(Swir1+Swir2)	
	光谱变异指数 SVVI=SD（Blue, Green, Red, Nir, Swir1, Swir2）-SD（Nir, Swir1, Swir2）	
	缨帽变换绿度 TCG=-0.1603×Blue+-0.4934×Red+ 0.2819×Green+0.7940×Nir+-0.0002×Swir1+-0.1446×Swir2	
特定指标统计值对应日期的光谱数据	蓝光波段反射率 Blue	**由以下指标决定的日期:** 最小 NDVI 最大 NDVI 第二小 NDVI 第二大 NDVI NDVI 中位数 最小值与第一四分位数之间的 NDVI 均值 第三四分位数与最大值之间的 NDVI 均值 最小 LST 最大 LST 第二小 LST 第二大 LST LST 中值 最小值与第一四分位数之间的 LST 均值 第三四分位数与最大值之间的 LST 均值 最小 S2N 最大 S2N 最小值与第一四分位数之间的 S2N 均值 第三四分位数与最大值之间的 S2N 均值
	绿光波段反射率 Green	
	红光波段反射率 Red	
	近红外波段反射率 Nir	
	短波红外波段反射率 1 Swir1	
	短波红外波段反射率 2 Swir2	
NDVI 相关的季节特征数据	RNph_sos	生长季开始的 NDVI
	RNph_eos	生长季结束的 NDVI
	RNph_sos_slope	从生长季开始到最大值的 NDVI 增长斜率
	RNph_eos_slope	从最大值到生长季结束的 NDVI 下降斜率
	RNph_sos_amp	从生长季开始到最大值的 NDVI 幅度
	RNph_eos_amp	从最大值到生长季结束的 NDVI 幅度

	RNph_ave	生长季开始与结束之间的 NDVI 平均值
	RNph_sum	生长季开始与结束之间的 NDVI 总和

LST：指地表温度

表A.3 年际物候指标

类型	指标	含义
目标年份前三年的不同波段和指标指数的统计值	蓝光波段反射率 Blue	最小值 最大值 第二小值 第二大值 中位数 标准差 全部数值的均值 除去最小值和最大值后的均值 最后一次观测值
	绿光波段反射率 Green	
	红光波段反射率 Red	
	近红外波段反射率 Nir	
	短波红外波段反射率 1 Swir1	
	短波红外波段反射率 2 Swir2	
	归一化植被指数 NDVI=(Nir-Red)/(Nir+Red)	
	归一化指数 S1N=(Nir-Swir1)/(Nir+Swir1)	
	归一化指数 S2N=(Nir-Swir2)/(Nir+Swir2)	
目标年以及前三年对应指标的线性拟合坡度和标准差	蓝光波段反射率 Blue	数值与观测日期线性回归的斜率 标准差
	绿光波段反射率 Green	
	红光波段反射率 Red	
	近红外波段反射率 Nir	
	短波红外波段反射率 1 Swir1	
	短波红外波段反射率 2 Swir2	
	归一化植被指数 NDVI=(Nir-Red)/(Nir+Red)	
	归一化指数 S1N=(Nir-Swir1)/(Nir+Swir1)	
	归一化指数 S2N=(Nir-Swir2)/(Nir+Swir2)	
目标年和前三年的差值统计值	蓝光波段反射率 Blue	最小值 最大值 第二小值 第二大值 最大值之后的数值 最小值之后的数值 全部数值的均值 去除最小值和最大值后的均值
	绿光波段反射率 Green	
	红光波段反射率 Red	
	近红外波段反射率 Nir	
	短波红外波段反射率 1 Swir1	
	短波红外波段反射率 2 Swir2	
	归一化植被指数 NDVI=(Nir-Red)/(Nir+Red)	
	归一化指数 S1N=(Nir-Swir1)/(Nir+Swir1)	
	归一化指数 S2N=(Nir-Swir2)/(Nir+Swir2)	
特定指标（例如 NDVI）统计值对应日期的光谱指标	蓝光波段反射率 Blue	最小 NDVI 最大 NDVI 第二小 NDVI 第二大 NDVI NDVI 中位数 最小 LST 最大 LST 第二小 LST 第二大 LST LST 中位数
	绿光波段反射率 Green	
	红光波段反射率 Red	
	近红外波段反射率 Nir	
	短波红外波段反射率 1 Swir1	
	短波红外波段反射率 2 Swir2	

LST：指地表温度

附录 B

（资料性）

基于随机森林进行特征重要性排序算法

随机森林是一种集成学习方法，由大量决策树组成。每棵决策树在训练过程中，随机抽取样本和特征进行节点分裂，以减少预测误差（如均方误差MSE）。集成多棵树后，随机森林能够有效提高预测精度并增强模型的鲁棒性。通过分析每个特征在树中分裂节点所带来的目标变量误差减小量来评估其重要性。具体步骤为：

（1）节点分裂误差减少量

对第t棵树的第j个节点，如果特征X_i被用于分裂，则该节点的误差减少量为：

$$\Delta MSE_{i,j}^{(t)} = MSE_{parent} - (\frac{N_{left}}{N_{parent}} MSE_{left} + \frac{N_{right}}{N_{parent}} MSE_{right}) \tag{B.1}$$

MSE_{parent}分裂前节点的均方误差,MSE_{left}和MSE_{right}为分裂后左右子节点的均方误差,N_{parent}，N_{left}和N_{right}分别为父节点和左右子节点的样本数。

（2）单棵树特征总贡献

对第t棵树中，特征Xi在所有节点的误差减少量进行累加：

$$imp_i^{(t)} = \sum_{j \in nodesX_i} \Delta MSE_{i,j}^{(t)} \tag{B.2}$$

（3）随机森林的总体重要性

对随机森林所有T棵树，将每棵树中特征贡献取平均，得到特征X_i的总体重要性：

$$imp(X_i) = \frac{1}{T} \sum_{t=1}^{T} imp_i^{(t)} \tag{B.3}$$

（4）特征排序

对所有特征重要性归一化到0-1的范围，进而对特征从高到低进行排序。

附录 C

（资料性）

卷积神经网络算法

卷积神经网络算法（CNN）是一种常用于处理高维数据的深度学习模型。与传统神经网络相比，CNN 能够通过卷积操作自动提取局部特征，并通过多层网络进行特征组合与抽象。其核心原理和操作流程如下：

（1）输入层

输入层包括每个样本的所有特征，为提高训练稳定性和收敛速度，需对输入数据进行标准化或归一化处理。

算法中，需要指定批量大小，即在一次参数更新过程中，网络所使用的训练样本数量。样本梳理可以控制内存占用，平衡训练稳定性与收敛速度。

（2）卷积层

基于数据自动提取局部特征。使用一组可训练的卷积核（滤波器）在输入数据上滑动，通过卷积运算生成特征图。需指定滤波器和核的大小。这个过程中，可引入激活函数，使得网络可拟合复杂的函数关系。

（3）池化层

可降低特征图的空间维度，减少参数数量和计算量，同时保留主要特征。

（4）丢弃层

在训练过程中，按照设定概率随机“丢弃”部分神经元，以减少过拟合，提高模型泛化能力，防止某些神经元对预测过度依赖。

（5）多层卷积、池化和丢弃层组合

卷积层和池化层可重复堆叠，逐层提取更抽象、更高层次的特征。

（6）全连接层

将最终卷积/池化得到的特征图展开为一维向量，输入全连接层。通过矩阵乘法和激活函数，将抽取到的特征映射到预测目标空间。

（7）输出层

输出模型预测值。

（8）损失函数与训练

定义损失函数衡量预测结果与真实值的差距，回归模型常通过均方误差（MSE）表征。使用优化算法（如 Adam）更新网络参数。其中，学习率是其中一个重要参数。

（9）模型评估与应用

使用验证集或测试集评估 CNN 模型的性能，常用指标包括 R^2、RMSE 等。

针对一个两层网络的 CNN，其重要参数及调参区间见表 C.1.

表C.1 卷积神经网络参数及调参区间

类型	参数	调参区间
第一层卷积层	滤波器个数	[64, 128, 256]
	核尺寸	[3, 7]
第一层丢弃层	丢弃率	[0, 0.5]
第二层卷积层	滤波器个数	[64, 128, 256, 512]
	核尺寸	[3, 7]
第二层丢弃层	丢弃率	[0, 0.5]
损失函数	学习率	[0.01, 0.0001]
模型表现	批量大小	[8, 16, 32, 64, 128]

附录 D

（资料性）

基于 X－Y 距离的样本集划分方法（SPXY）

基于 X–Y 距离的样本集划分方法（SPXY）旨在根据特征空间（X）和响应变量空间（Y）的分布合理选择训练样本和测试样本，使模型训练和评估更加稳定和代表性。其核心思想是选择既能覆盖样本特征空间，又能反映目标变量变化范围的样本用于训练。生成的训练集和测试集在空间分布上更均匀，有利于建模稳定性和预测性能。其计算步骤为：

（1）计算样本间 X-Y 距离

对每个样本，构建包含输入特征向量 Xi 和目标变量 Yi 的联合向量[Xi,Yi]。

计算样本之间的欧氏距离或加权距离：

$$D_{ij} = \sqrt{\| X_i - X_j \|^2 + \alpha \| Y_i - Y_j \|^2} \tag{D.1}$$

其中∝为权重系数，用于平衡特征空间和响应变量空间的重要性，默认可取 1.

（2）选择初始样本

从样本集中选择一距离最远的样本作为初始训练样本，该步骤保证训练集覆盖特征和响应变量的边界。

（3）逐步添加样本到训练集

对剩余样本，计算其到已选训练样本的最小距离

$$d_i = \min_{j \in \text{训练集}} D_{ij} \tag{D.2}$$

选择距离最远的样本加入训练集，重复该过程，直到训练集达到预设比例或与其样本数量。

（4）划分测试集

剩余样本作为测试样本，保证测试集覆盖特征空间和响应变量空间的未选区域，增强模型评估的代表性。

参考文献

[1] Potapov, P., Hansen, M.C., Kommareddy, I., Kommareddy, A., Turubanova, S., Pickens, A., Adusei, B., Tyukavina, A., & Ying, Q. (2020). Landsat Analysis Ready Data for Global Land Cover and Land Cover Change Mapping. Remote Sensing, 12, 426

[2] Conrad, O. (2002). Vertical Distance to Channel Network. SAGA-GIS Tool Library Documentation.

[3] Conrad, O. (2007). Sink Removal (Fill Sinks). SAGA-GIS Tool Library.

[4] Tarboton, D.G. (1997). A new method for the determination of flow directions and upslope areas in grid digital elevation models. Water Resources Research, 33, 309-319