

中图分类号：P237

学校代码：10081

U D C:

密 级：公开



硕 士 学 位 论 文

基于 Sentinel 影像与地形因子的土壤盐分 遥感反演研究

论文作者： 李杰
学生类别： 全日制专业型
学科专业： 测绘工程
学位类别： 工程硕士
指导教师： 汪金花 教 授
协助导师： 张婷婷 副研究员
企业导师： 方 磊 高级工程师

唐山 华北理工大学
2023 年 10 月

**Remote sensing inversion of soil salinity based on
Sentinel images and terrain factors**

Dissertation Submitted to
North China University of Science and Technology
in partial fulfillment of the requirement
for the degree of
Master of Engineering (ME)

by
Li Jie
(Surveying and Mapping Engineering)

Professor Wang Jinhua
Supervisor: Zhang Tingting
Fang Lei

October, 2023

摘 要

土壤盐渍化是目前世界面临的最主要的环境问题之一，对全球的农业生产构成了严重地威胁，严重影响着当地的农业发展和生态建设。黄河三角洲流域的土壤盐渍化现象越来越严重，了解该区土壤盐分的空间分布信息可为该区的土地管理、生态修护和农业的可持续发展提供决策依据。基于此，以山东省东营市境内的黄河入海口区域为研究区，基于 Sentinel-1/2 影像数据以及东营市的 30m 分辨率的 ASTER GDEM 高程数据，结合野外与卫星同步获得的实测数据，采用逐步回归模型（Stepwise Regression, SR），随机森林（Random Forest, RF），支持向量机（Support Vector Regression, SVR）和偏最小二乘回归（Partial Least Squares Regression, PLSR）方法开展了土壤盐分含量反演的研究，并以最优特征组合和方法绘制研究区表层土壤盐分的空间分布图。主要研究成果如下：

1) 基于 Sentinel-1/2 影像数据和 DEM 高程数据提取的遥感影像特征和地形因子，建立了土壤盐度和特征数据集，采用随机森林模型的特征重要性方法来筛选研究区土壤盐分预测的最优特征。结果发现，光学影像中的 CRSI 指数在反演该地区土壤盐分的所有特征中贡献度最大，其重要性得分为 0.572，显著高于地形因子 SC (0.238) 和雷达特征 VV (0.165)，其可作为该地区土壤盐分反演的优秀指标。

2) 通过对比研究区基于不同数据源特征及其组合的情况下土壤含盐量反演精度的相对高低，结果发现，综合考虑遥感影像特征和地形因子来反演该地区的土壤盐分效果最好，其次为仅用遥感影像特征，预测精度最差的是仅用地形因子作为土壤盐分反演的参数。由于最佳的预测效果是综合考虑遥感影像特征和地形因子，表明基于多源特征监测土壤盐分具有很好的预测效果。

3) 基于上述最优数据源特征组合的基础上，通过使用随机森林、多元逐步回归、支持向量机和偏最小二乘 4 种机器学习算法构建了研究区土壤盐分反演模型，结果表明，PLSR 模型的预测效果最好 ($R^2=0.66$, RMSE=1.30)，其次为 RF 模型 ($R^2=0.63$, RMSE=1.33) 和 SR ($R^2=0.51$, RMSE=1.38)，预测效果最差的为 SVR ($R^2=0.40$, RMSE=0.29)。由此可见 PLSR 是最适合此区域土壤盐分反演的模型，可在对土壤盐分贡献度较好的特征筛选后实现研究区土壤盐分的快速反演。

图 28 幅；表 14 个；参 77 篇。

关键词：土壤盐分；数字高程模型；Sentinel-1/2；机器学习；黄河三角洲

分类号：P237

Abstract

Soil salinization is currently one of the main environmental problems facing the world, seriously affecting local agricultural development and ecological construction. The Yellow River estuary area in Dongying City, Shandong Province is used as the research area. Stepwise regression, random forest, support vector machine and partial least squares regression were used to study the inversion of soil salinity based on Sentinel-1/2 image and DEM data. The optimal feature combinations and method were used to map the spatial distribution of soil salt. The main research results are as follows:

1) Based on remote sensing image features and terrain factors extracted from Sentinel-1/2 image and DEM data, soil salinity and feature dataset was established. The feature importance method of random forest was used to screen the optimal features for soil salinity prediction in the study area. The results showed that the CRSI index contributed the most to the retrieval of all characteristics of soil salinity in the region.

2) Comparing the relative accuracy of soil salt content retrieval based on different data source and combinations, it is best to comprehensively consider remote sensing image features and terrain factors to retrieve soil salt content in the area, followed by using only remote sensing image features, the worst prediction accuracy is using only terrain factors as parameters for soil salt retrieval. The best prediction effect is to comprehensively consider remote sensing image features and terrain factors, indicating that monitoring soil salinity based on multi-source features has a good prediction effect.

3) The inversion results show that among the four modeling methods, the partial least squares regression has the best overall prediction effect ($R^2=0.66$, $RMSE=1.30$) based on the combination of radar features, salinity vegetation index and terrain factors extracted from multi-source image data. It can be seen that the PLSR method is the most suitable model for soil salt inversion in this study area, which can effectively quantitatively estimate soil salt in the study area. Finally, the spatial distribution of soil salinity in the study area was obtained using PLSR method based on the filtered features.

Figure 28; Table 14; Reference 77

Keywords: Yellow River Delta, soil salinity, DEM, sentinel-1/2, machine learning

Chinese books catalog: P237

目 录

第 1 章 绪论.....	1
1.1 研究背景和意义.....	1
1.2 国内外研究现状.....	2
1.2.1 土壤盐分的光学遥感监测.....	2
1.2.2 土壤盐分的雷达遥感监测.....	3
1.2.3 土壤盐分的地形因子监测.....	3
1.2.4 土壤盐分特征筛选算法.....	4
1.3 研究目标和内容.....	5
1.3.1 研究目标.....	5
1.3.2 研究内容.....	5
1.4 技术路线.....	6
第 2 章 理论基础.....	8
2.1 土壤雷达影像的极化目标分解.....	8
2.1.1 相干目标分解和非相干目标分解.....	8
2.1.2 H/A/ α 极化分解理论.....	10
2.2 土壤盐分的介电特性.....	11
2.3 土壤盐分的建模方法.....	12
2.3.1 逐步多元回归.....	12
2.3.2 支持向量机.....	13
2.3.3 随机森林.....	14
2.3.4 偏最小二乘回归.....	15
2.4 本章小结.....	16
第 3 章 研究区域概况和数据来源.....	17
3.1 研究区概况.....	17
3.1.1 地理位置.....	17
3.1.2 气候降水.....	18
3.1.3 地形地貌.....	18
3.1.4 土壤类型.....	19
3.2 数据源介绍.....	19
3.2.1 地面采样数据.....	20
3.2.2 Sentinel-1 数据.....	22
3.2.3 Sentinel-2 数据.....	22
3.2.4 DEM 数据.....	23
3.3 本章小结.....	23

第4章 监测信息的数据处理	24
4.1 研究区数据预处理	24
4.1.1 Sentinel-1 SAR 雷达数据	24
4.1.2 Sentinel-2 MSI 光学数据	29
4.1.3 DEM 数据	30
4.2 特征提取	31
4.2.1 Sentinel-1 SAR 雷达特征	31
4.2.2 Sentinel-2 MSI 光学特征	32
4.2.3 地形因子	34
4.3 土壤盐异常值的确定	35
4.4 本章小结	36
第5章 结果与分析	38
5.1 特征筛选	38
5.1.1 特征筛选的方法	38
5.1.2 数据集划分	38
5.1.3 筛选结果	39
5.2 模型配置与评估	41
5.2.1 模型配置	41
5.2.2 模型评估	43
5.3 基于遥感影像特征的土壤含盐量反演	44
5.3.1 回归模型结果	44
5.3.2 模型比较	45
5.4 基于地形因子的土壤含盐量反演	46
5.4.1 回归模型结果	46
5.4.2 模型比较	47
5.5 基于全参数的土壤含盐量反演	47
5.5.1 回归模型结果	47
5.5.2 模型比较	48
5.5.3 最优模型	49
5.6 土壤盐分制图	50
5.7 本章小结	52
结论	53
参考文献	55

第1章 绪论

1.1 研究背景和意义

土壤盐渍化是指由人为活动和环境因素综合影响而导致土壤中的可溶性盐的逐渐积累、最终形成具有高盐分土壤的过程，其中人为活动主要为不合理的灌溉及施肥等，而环境因素则包括当地的地形、母质、气候和水文等众多因素^[1]。土壤盐渍化作为一种常见的土壤退化现象，其不仅会影响农作物的生长环境，进而影响农作物的产量，限制当地的农业可持续发展，更会破坏当地的生态平衡，从而导致当地的生态环境恶化及引发生态危机的可能，已成为一个全球关注并亟待解决的环境问题^[2, 3]。

目前，全球已有约 $9.55 \times 10^8 \text{hm}^2$ 的土地资源受土壤盐渍化的影响，而全球的总耕地也才 $15 \times 10^8 \text{hm}^2$ 的面积，其中次生盐渍土就有 $0.77 \times 10^8 \text{hm}^2$ ，而且这个数字还在不断地增长^[4, 5]。其中受土壤盐渍化影响面积最大的为澳大利亚，其仅次生盐渍化就占了其约 10% 的土地。亚洲中东部次之，伊朗约有 $6.8 \times 10^6 \text{hm}^2$ 的土地受盐渍化的影响，而其总耕地面积约为 $2.2 \times 10^7 \text{hm}^2$ ，其中盐渍化土壤占其总耕地面积约 1/3；在中国有 $1.34 \times 10^8 \text{hm}^2$ 的土地用于耕地，而其受盐碱化威胁的土壤就有约 $9.33 \times 10^6 \text{hm}^2$ ，且其盐渍化土壤正在以每年 $1.35 \times 10^4 \text{hm}^2$ 的速度增加。亚洲南部和非洲北部也不同程度地受到土壤盐渍化的影响^[6, 7]。

我国是土壤盐渍化危害问题比较严重的国家之一，其主要分布在华北、东北、西北及沿海地区。其中，黄河三角洲是一个比较典型的滨海盐渍土分布区，深受土壤盐渍化的影响，其约有 $44.29 \times 10^4 \text{hm}^2$ 的土地受土壤盐渍化的影响，占其全部区域的一半以上。该地区由于其低海拔、蒸发量大，海水入侵后造成的地下水矿化问题严重，极易在地表堆积盐分，造成严重的土壤盐渍化威胁^[8, 9]。近些年来，黄河三角洲的盐渍化问题日益严重，对当地的生态环境、农业的发展和经济都造成了巨大的影响，大大地制约着当地的可持续发展。因此，面对这日益严重的土壤盐渍化问题，及时且有效地加强对黄河三角洲土壤盐渍化区域尺度的监测是很有必要的，为日后决策者科学地管理土地资源和改善土壤环境提供了一个有用的参考依据，也会对我国的经济发展起巨大的积极作用^[10]。

传统的土壤盐分测量方法费时费力，成本高，范围小，且不易于动态监测，难以满足当代要求日益严格的社会需求。随着遥感技术在监测土壤信息方面的快速发展，利用遥感技术来获取研究区土壤盐渍化的空间分布已成为了可能。目前，常用

于监测土壤盐分的遥感方法有光学遥感和微波遥感。光学遥感其以信息量丰富、监测周期短、成本低等优点，可及时且动态地监测大范围的土壤盐渍化的空间分布情况，因此光学遥感技术已被广泛地用于高效且准确地监测土壤盐渍化的信息。迄今为止，已有大量的研究通过光学遥感数据来对土壤盐渍化进行分析与预测，并取得了较好的结果。虽然光学遥感技术是目前最广泛用于土壤监测的方法，但其也具有很多局限性，如其易受云雾的影响且对地表不具有穿透性。近年来，微波遥感技术由于其具有良好的穿透性，能全天候地监测地表一定深度的土壤成分，已被越来越多的学者用于获取大范围的土壤盐分信息，这很好地弥补了光学数据的不足。虽然光学遥感技术和微波遥感技术都有其优缺点，但如果仅使用遥感数据来反演土壤盐分信息，有时候反演的精度也不太理想。土壤盐分与土壤的形成息息相关，而土壤的形成受到母质、气候、地形、生物与时间这 5 个成土因子的综合影响。因此本研究将在保留光学遥感数据和雷达遥感数据优势的基础上，并耦合从土壤成因的角度上考虑的环境因子来进行土壤盐渍化的反演研究，以期能在大尺度上较为准确地监测土壤盐渍化的空间分布和筛选最优参数组合，为黄河三角洲土壤盐渍化的治理及生态建设提供参考依据。

1.2 国内外研究现状

土壤盐渍化是由于大量人为活动和自然干扰造成的盐分在土壤中逐渐累积的一种土壤退化现象，目前已对世界上很多国家的农业生产和生态保护造成了严重的影响^[11, 12]。中国是盐渍化问题比较严重的国家之一，由于其分布面积大，危害时间长，显著地降低了当地的农业生产力和经济效益^[13, 14]。土壤盐分的含盐量直接反映了土壤的质量。因此，及时且有效地获取土壤盐分的信息具有十分重要的意义，目前已有许多研究使用光学数据、雷达数据或地形数据来反演研究区土壤盐渍化的分布情况^[15-17]。

1.2.1 土壤盐分的光学遥感监测

与传统的野外调查和实验室分析方法相比，遥感技术具有影像重访周期短，成本低，覆盖范围大，信息量丰富等优势，已成为一种广泛地土壤盐分动态监测和评价的方法，在土壤盐分预测等方面的研究已经表现出了巨大的潜力^[18, 19]。近几十年来，随着遥感技术和成熟与发展，光学遥感数据已被越来越多的学者相继应用于土地盐碱化的监测等研究。EL HARTI 等通过使用 Landsat TM/OLI 数据，在盐分指数（SI）的基础上加入了蓝波段，建立了一个新的土壤盐度指数（OLI-SI）用于

土壤盐度的估算，提高了摩洛哥罐区的土壤盐分反演精度^[20]。SCUDIERO 等观察到，冠层响应盐度指数（CRSI）与土壤盐度之间存在很强的相关性，结果显示来自 Landsat 7ETM+数据的盐度指数 CRSI 可以用于识别受土壤盐渍化影响的区域和作为一个获取和制作研究区土壤盐分空间分布图的指标^[21]。BANNARI 等发现，Sentinel-2 卫星影像的 SWIR 波段对土壤盐度更为敏感，其可以作为土壤盐度建模和监测的最佳候选波段^[22]。李晓燕等在基于 Landsat-TM 和 MSS 影像的基础上，结合 GIS 技术，对研究区盐碱化土壤做了动态变化预测及成因分析^[23]。王爽等在优选出的敏感波段基础上，使用 Landsat-TM 影像，建立最佳反演模型来获取土壤盐渍化信息^[24]。然而，他们大多使用光学遥感数据，很少使用极化 SAR 数据，SAR 影像包含更多的结构特征^[25]，采用雷达极化分解技术获取土地盐渍化信息的研究并不多，基本上还处于初步研究阶段^[26]。

1.2.2 土壤盐分的雷达遥感监测

微波遥感是一种常见的地表信息获取手段，具有良好的穿透性和全天候探测能力，弥补了光学遥感的不足，在土壤成分预测上具有一定的优势^[27]。TAGHADOSI 等^[28]对 Sentinel-1 SAR 影像数据应用了基于纹理的支持向量回归(SVR)分析来绘制盐度，研究了直接关联雷达强度与土壤盐分的方法，获得了较好的土壤盐分反演结果。刘全明等^[17]通过使用从 Radarsat-2 影像提取的后向散射系数与土壤盐分测量数据，构建了内蒙古河套罐区表层土壤含盐量的反演模型。LASNE 等^[29]研究表明，在微波频率 1~7GHz 范围内（Sentinel-1 影像数据的中心频率为 5.404GHz），虚部对土壤盐分敏感，而实部与水分含量关系更大。因此有效地利用极化 SAR 数据可为大范围的获取研究区的土壤盐分信息提供了一种新的途径，可以动态、及时地对研究区农业生产实践提供技术支持。

1.2.3 土壤盐分的地形因子监测

然而，仅使用雷达数据进行土壤含盐量的研究往往精度不高。由于土壤盐渍化成因复杂，受自然条件和人类活动方式的影响^[30]，与多种环境因子都有一定的关系，因此众多学者也开始从土壤发生学的角度上选择环境变量来进行研究区土壤含盐量的反演^[31]。地形参数作为最常用且具有代表性的环境因子，已被大量的研究作为反演土壤盐分含量的参数，并获得了极好的效果^[32, 33]。研究结果^[34]表明，地形因素对土壤盐度有最为决定性的贡献。SHAHRAYINI 等观察到，地形因素，如到渠道网络的垂直距离（VDCN）、分析性山丘遮蔽（AH）、流量累积（FA）和

地形湿度指数 (TWI)，对土壤盐碱化的预测有很大贡献^[35]。TAGHIZADEH MEHRJARDI 等通过使用回归树模型确定了地形指数 TWI 是在 60~100cm 深度区间中最为重要的参数，他们的结果证明，随着土壤深度的增加，地形参数变得更加重要^[36]。贺文君等利用样线法对不同高程下土壤水盐的信息进行了调查，结果表明，随着高程的增加，各土层土壤水含量呈现降低趋势，而表层土壤含盐量呈增加趋势^[37]。

1.2.4 土壤盐分特征筛选算法

为了提高土壤盐渍化反演的精度，本文将从土壤发生学的角度来考虑应选择的地形因子，并在此基础上，结合 Sentinel-1/2 数据，运用了多元逐步回归、支持向量机、随机森林和偏最小二乘 4 种机器学习算法对研究区进行土壤含盐量的研究。目前已有大量的研究使用了机器学习算法用于研究区土壤含盐量的建模及预测。SIDIKE 等用偏最小二乘回归法估算了中国平罗县的土壤盐度，得到了很好的反演精度，且其预测精度优于逐步回归方法^[38]。FARIFTEH 等也用偏最小二乘回归法对土壤 EC 测量值与反射光谱进行建模，得到了良好的预测^[39]。NURMEMET 等使用了支持向量机分类的方法和融合数据 (Landsat ETM+、PALSAR 和 Radarsat-2) 绘制了中国西北地区的土壤盐渍化图，研究结果表明，支持向量机分类是很好地监测土壤盐渍化的一种方法^[26]。WANG 等融合遥感数据和景观特性，采用了偏最小二乘回归 (PLSR)、卷积神经网络 (CNN)、支持向量机 (SVM) 和随机森林 (RF) 四种机器学习方法对中国新建南部地区的盐渍土进行监测，结果表明，随机森林 (RF) 模型是该地区最好的回归模型^[34]。

由于选取的参数较多且部分机器学习算法自身没有特征筛选的功能，会造成信息冗余，因此选择合适的特征筛选算法尤为重要。目前常用的变量筛选方法有皮尔森相关系数，灰度关联分析，岭回归和最优子集法，也有一些较小众的筛选方法，如遗传算法 (GA) 和随机森林 (RF)。ALLBED 等采用皮尔逊相关系数分析法来筛选特征，发现土壤调节植被指数 (SAVI) 在评估强烈盐渍化区域有巨大的潜力^[40]。赵庆展等采用最佳指数法 Optimum Index Factor (OIF) 对无人机多光谱影像特征进行组合分析，得出了研究区地物分类的最佳波段组合，分类精度有了明显提高。朱高飞采用 Landsat 5 数据和 Charles Horowitz 提出的最佳指数法选择最佳波段组合，得出波段 1 和波段 3 的组合与土壤盐分的相关系数最高^[41]。陈俊英等利用灰色关联法对构建的光谱指数进行筛选，结果表明，以优选盐分指数和光谱指数作为变量组构建的模型优于植被指数变量组^[42]。XU 等提出了一种基于自适应遗传算

法的支持向量机回归 (AGA-SVR) 的新方法, 并与其他模型进行比较, 结果表明基于遗传算法的支持向量回归 (GA-SVR) 与 AGA-SVR 以较少的特征参数获得了更为准确的土壤盐碱化的信息^[43]。GENURER 等研究了一些基于随机变量的变量重要性指数行为的实验见解, 提出了一种不依赖于特定的模型假设, 只基于数据驱动的阈值来进行决策的方法^[44]。

总的来说, 越来越多的学者使用光学或雷达数据来检索土壤盐度。然而, 很少有研究从土壤成因的角度出发选择环境变量, 并结合光学或雷达数据来监测土壤盐渍化信息。本研究将利用从 DEM 数据中提取的地形因子和从 Sentinel-1/2 遥感影像数据中提取的雷达特征和盐度植被指数, 基于随机森林特征重要性的方法, 筛选最佳的变量及其最佳组合, 并结合四种机器学习算法来绘制东营市土壤盐渍化的空间分布图。

1.3 研究目标和内容

1.3.1 研究目标

本文将山东省东营市内的黄河三角洲为研究对象, 利用从 Sentinel-1/2 遥感数据和 DEM 数据提取出的特征因子, 并使用了特征重要性的方法来评估和筛选出对该地区土壤含盐量贡献度最大的几个变量, 并筛选出最优的变量组合代入多元逐步回归, 随机森林, 支持向量机和偏最小二乘法 4 种机器学习算法中, 对模型的精度结果进行对比分析, 最后选出最优的变量组合和建模方法来对研究区的土壤盐渍化空间分布进行预测和制图, 并统计该地区各类盐渍土面积所占的比例, 为该地区土壤盐渍化的环境治理和生态保护提供了科学指导。

1.3.2 研究内容

本文以山东省东营市内的黄河三角洲区域为研究目标, 基于 Sentinel-1 SAR 雷达数据、Sentinel-2 MSI 光学数据和 30m 分辨率的 DEM 数据为数据源, 耦合四种机器学习算法 (随机森林模型、支持向量机、多元逐步回归和偏最小二乘法), 分别对不同的数据源提取的特征进行建模分析与比较, 筛选出最适应该地区的特征参数与模型, 实现该区域大面积的土壤含盐量的估计。本文主要研究内容包括:

1) 选取与土壤含盐量有高相关性的因子。基于大量采用光学特征反演土壤含盐量研究的基础上, 本文首先考虑了雷达影像特征与土壤含盐量从理论上具有一定的相关性, 且雷达影像对反演土壤含盐量具有一定的优势 (如可穿透云层和一

定深度的地表），从土壤发生学的角度上考虑了可能影响土壤盐分的环境因子。综合以上数据提取出的特性，利用了随机森林的特征重要性来评估这些特征因子的重要性得分并排序。最后筛选出在研究区域与土壤含盐量具有高相关性的遥感影像特征参数或环境因子。

2) 基于四种机器学习算法对从不同的数据源提取的特征进行建模与比较。本文采用以上四种机器学习算法，分别对从遥感影像特征提取的特征、从 DEM 数据提取出的地形因子和综合考虑所有因子进行土壤含盐量的建模与评估，筛选出精度表现最佳的一组数据源，基于适用于该研究区的特征，发展适用于该区域的土壤含盐量反演模型，并进行精度评价和野外验证。

3) 基于不同机器学习算法土壤含盐量反演的精度评估与比较。通过传统探钻的手段获取东营市地区土壤含盐量数据，再结合从 DEM 数据提取出的地形因子和从 Sentinel-1/2 遥感数据中提取出的特征参数，基于 4 种机器学习模型进行土壤含盐量与各特征参数的关系分析。在训练集上，对多种类型的特征，构建多种机器学习算法模型，构建模型的过程中采用的是基于粒子群算法并结合交叉验证的方法进行参数的调优。将优化好的模型用于测试集进行土壤含盐量的估算及验证，并选用测试集上表现最优的模型进行黄河三角洲区域的土壤含盐量预测研究并制图。

1.4 技术路线

本文以 Sentinel-1 SAR 雷达影像、Sentinel-2 MSI 光学影像和 30m 的 DEM 高程数据为基础，提取出雷达影像特征、盐度植被指数和地形因子，并结合同时期的野外土壤盐分测量数据，建立研究区土壤盐分特征数据集，首先对数据集中的离散值进行了提出，并根据交叉验证的方法将此数据集划分为训练集和验证集，随后采用了随机森林模型的特征重要性的方法对研究区土壤盐分的驱动因素进行排序与筛选，提取出对研究区土壤盐分贡献性最大的特征；接着在基于筛选的最优特征的基础上，进一步分析与对比研究区基于不同数据源特征及其组合的情况下土壤含盐量反演精度的相对高低，寻找出最适应于研究区土壤盐分反演的数据源的特征组合；在多源数据的最佳特征组合的基础上，分析并对比基于随机森林模型、支持向量回归、多元逐步回归和偏最小二乘回归方法的模型表现与预测效果，并基于最佳的模型用于研究区土壤盐分的预测，实现该区域大面积的土壤含盐量的估计，统计出各程度盐渍土的比例。本文按照如下数据处理及方法先后进行了实验，其主要阶段及研究技术路线如图 1 所示：

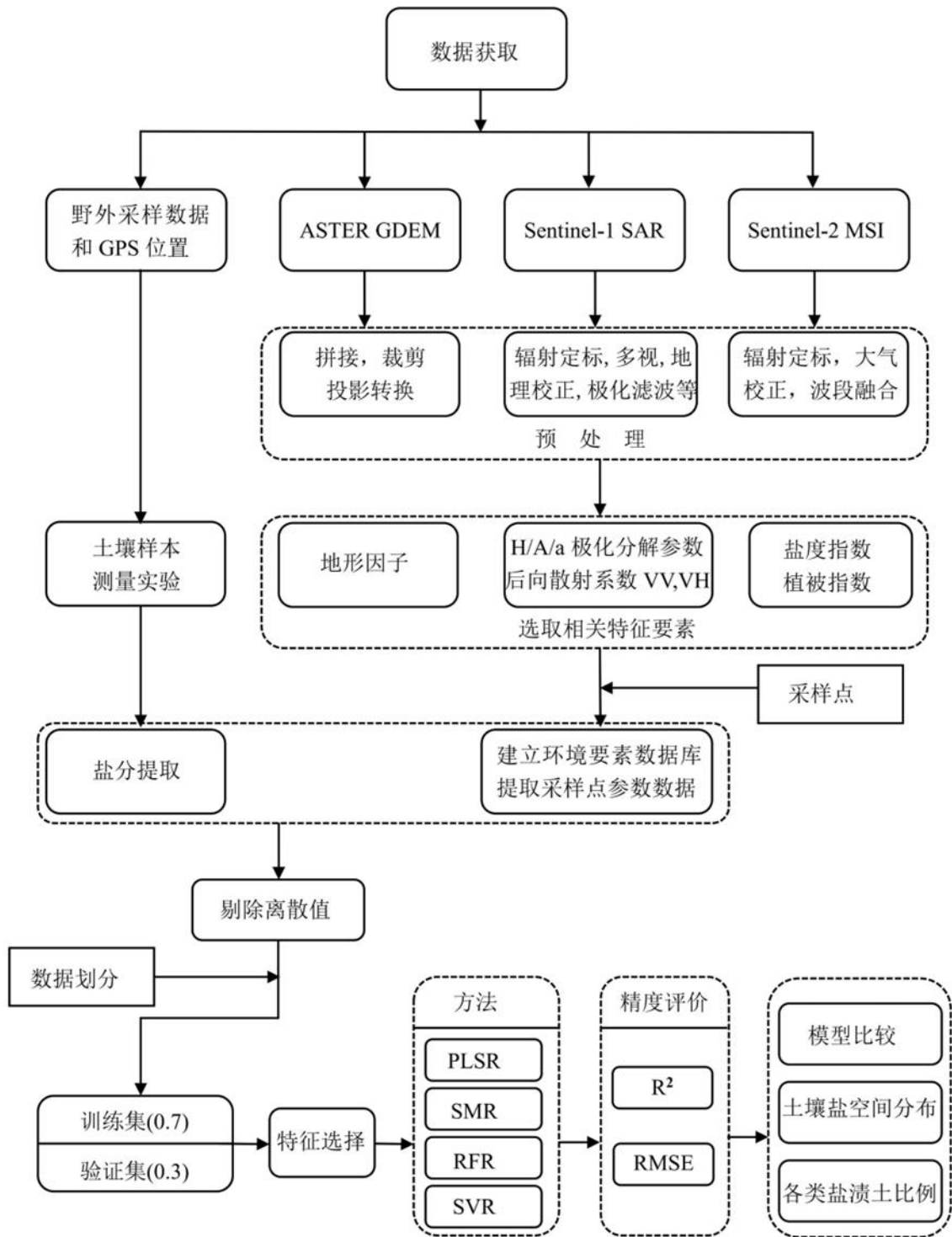


图 1 研究技术路线图

第 2 章 理论基础

SAR 经发射天线发射电磁波，再经地物目标对其的变极化反应后，将以极化散射矩阵的形式存储在数据存储中，这就有了 SAR 数据。为了充分利用 SAR 数据的极化信息，首先需要对原始的 SAR 数据进行分析，提取出更为精确的地物目标属性信息，其核心理论就是极化分解理论。

2.1 土壤雷达影像的极化目标分解

极化散射矩阵反应了地物目标在一定的姿态和观察频率下对电磁波的极化散射效应，可将其记为：

$$[S] = \begin{bmatrix} S_{VV} & S_{VH} \\ S_{HV} & S_{HH} \end{bmatrix} \quad (1)$$

式中：

S_{VV} —垂直极化 V 方式发射，垂直极化 V 接受时的目标后向散射系数；

S_{VH} —垂直极化 V 方式发射，水平极化 H 接受时的目标后向散射系数；

S_{HH} —水平极化 H 方式发射，水平极化 H 接受时的目标后向散射系数。

极化 SAR 分析每个地物目标单元内的散射回波来获得该目标的[S]矩阵，[S]矩阵可以统一目标散射回波的能量特性、相位特性和极化特性等信息，较好地表征了地物目标的电磁散射特性^[45]。其中地物目标的极化特性因与它自身的形状结构有关，所以可以反映其他雷达极化参数所不能探测的地物表面的粗糙程度、对称性等特征，因此，对目标物的极化特性提取就很有必要^[46]。

地物目标极化分解理论是由 Huynen 提出，该理论可以充分地利用极化特性的信息，其目的是通过解析地物自身的物理特性来进一步解释地物的散射特性。随着时代的发展，各种目标极化分解理论也相继诞生。Cloude^[47]将这些理论分为了以下两种。一种为相干目标分解，即当散射目标的散射特征是确定的或稳态的，散射回波是相干的时；另一种是非相干目标分解，即当目标散射是非确定的，随时间而变化的，散射回波是非相干的。

2.1.1 相干目标分解和非相干目标分解

相干极化分解方法是在极化散射矩阵[S]基础上进行的操作处理，其主要有 SDH 分解，SSCM 分解，Cameron 分解和 Pauli 分解等。相干分解的目的是将测得

的散射矩阵[S]表示为几个散射矩阵之和，从而确定一种某种确定的散射机理。

相干矩阵也提供了极化散射矩阵中相对相位及矩阵元素之间的相关性信息，主要包括极化通道间的相位差 Φ 与相关系数 ρ 。

$$\phi_{ij} = \text{Arg}(\langle S_i S_j^* \rangle), i, j \in \{h, v\} \quad (2)$$

$$\rho_{ij} = \frac{\langle S_i S_j^* \rangle}{\sqrt{\langle |S_i|^2 \rangle + \langle |S_j|^2 \rangle}}, i, j \in \{h, v\} \quad (3)$$

式中：

||—取模操作；

Arg—取相位操作。

同极化相关系数通常用来描述散射的随机性，值越小，散射随机性越高，具有特殊朝向的散射体一般同极化相关系数更高。在自然地物中，同极化与交叉极化散射矩阵元素是非相干的，对应的相位差一般为0，具有偶次散射机制的散射体理论上具有较高的相位差。

基于极化相干矩阵（**T**矩阵）、极化协方差矩阵（**C**矩阵）和 Stokes 矩阵来进行分解的叫做非相干目标分解。极化散射矩阵能对一定条件下的确定性目标的散射特性来进行描述，但是，当目标的散射特性由多个独立的散射体构成或具有变化时，散射回波表现出一定的散射随机性，这就需用高阶统计方法来描述分布式电磁散射特性^[48, 49]。目前常用二阶统计量来描述，包括协方差矩阵 **C** 与相干矩阵 **T**，其可由[S]矩阵转化得来，它们的公式如下：

$$[C] = \langle K_B * K_B^{*T} \rangle \quad (4)$$

$$[T] = \langle K_P * K_P^{*T} \rangle \quad (5)$$

式中，当满足互易原理的条件 $S_{HV} = S_{VH}$ 时，其 $K_B = [S_{VV}, \sqrt{2}S_{VH}, S_{HH}]^T$ ， $K_P = \frac{1}{2}[S_{VV} + S_{HH}, S_{VV} - S_{HH}, 2S_{HV}]^T$ 。

本文所研究的山东省东营市的土壤盐分的散射过程可看作是一个随机、不确定的过程，因此需要用非相干目标分解方法进行研究。目前，常用的非相干目标分解方法主要包括 H/A/ α 极化分解、Cloude-Pottier 极化分解、Yamaguchi 极化分解、和 Freeman-Durden 极化分解，本文主要采用了基于[C]矩阵的 H/A/ α 分解，该方法作为较常用的方法，可分解出大量的极化参数，为后续土壤盐分的建模及预测提供了数据支持。

2.1.2 H/A/α极化分解理论

1997年, CLOUDE 和 POTTIER 提出一种基于二阶统计量的 H/A/α极化分解方法, 用于计算样本平均参数^[50]。这种方法不受特定的统计分布假设和多变量模型物理约束条件的影响。它利用对 3×3 相干矩阵 T_3 的特征矢量分析, 可将相干矩阵分解为不同的散射过程类型 (特征矢量) 及其对应的相对幅度 (特征值), 且在分解的过程中, 其特征值不会改变。该极化分解方法能够提取出散射机制的平均参数, 这些极化参数于地表散射的物理机制相联系, 可在一定程度上反应出地表散射特性。因此本文主要采用 H/A/α极化分解产生的特征值和特征矢量及其参数组合来建立它们与土壤含盐量之间的关系。在 H/A/α极化分解过程中, 相干矩阵 T_3 可展开为三个相互独立的目标之和:

$$T_3 = \sum_{i=1}^3 \lambda_i T_i = \lambda_1 e_1 e_1^{*T} + \lambda_2 e_2 e_2^{*T} + \lambda_3 e_3 e_3^{*T} \quad (6)$$

式中:

λ_i —实数, 对应 T_3 的特征值, 分别描述 3 个 T_i 的统计权重;

T_i —秩为 1 的独立相干矩阵, 描述一定的散射机制;

e_i —特征向量。其特征矢量 e_i 可表示为:

$$e_i = e^{i\phi_i} [\cos \alpha_i, \sin \alpha_i \cos \beta_i e^{i\delta_i}, \sin \alpha_i \cos \beta_i e^{i\gamma_i}] \quad (7)$$

式中:

ϕ 、 δ 、 γ —特征矢量的相位角;

β —二倍的极化方位角;

α — T_i 对应的散射机制。

根据特征值和特征矢量可计算出 3 个基本的参数, 它们是平均散射角 $\bar{\alpha}$ 、极化熵 H、各向异性度 A, 它们的公式如下。

$$H = -\sum_{i=1}^3 p_i \log_3 p_i \quad (8)$$

$$\bar{\alpha} = \sum_{i=1}^3 \lambda_i \rho_i \alpha_i \quad (9)$$

$$A = (\lambda_2 - \lambda_3) / (\lambda_2 + \lambda_3) \quad (10)$$

式中:

H—极化散射熵;

A —各向异性度;

$\bar{\alpha}$ —平均散射角;

P_i —各特征值占总特征值和的比重。

极化散射熵 H 代表的是散射的随机性, 其取值范围在 $[0,1]$ 之间, 通常其值越大, 目标散射随机程度越高; $\bar{\alpha}$ 用于描述目标散射机制从奇次散射、体散射、再到偶次散射的变化过程, 当 $\alpha=0^\circ$ 时, 表征各向同性表面, 即奇次散射; 伴随 α 的增大会呈现出各向异性, 当 $\alpha=45^\circ$ 时, 为偶极子; 当 $\alpha>45^\circ$ 时, 表示为各向异性的二面角; 而当 $\alpha=90^\circ$ 时则表征为各向同性的二面角反射器。 A 为各向异性度, 当 H 的值较大时($H>0.7$), 表示其余不占优的两种散射机制对结果的影响效果。

除了上述的三个极化参数外, 在 $H/A/\alpha$ 极化分解的过程中还产生了其他大量的极化参数, 其中有的极化参数已研究出其物理意义, 但是有的极化参数的物理意义尚待研究, 本研究中采用了 $H/A/\alpha$ 极化分解方法所分解的全部参数。

2.2 土壤盐分的介电特性

介电常数是物体的物理特性之一, 用于描述物体的表面电学特性, 由物质的组成和温度决定。土壤介电常数的实部受到土壤含水量的影响, 其虚部则受土壤含盐量的影响, 而土壤中水溶液的含盐量的变化会影响土壤的导电性(即电导率值)。而介电常数的变化会影响雷达图像所观测的目标物的后向散射系数, 这就使微波遥感技术用于观测土壤盐分的分布提供了理论支持。邵云等人通过对含盐含水土壤进行了微波介电特性分析研究, 结果揭示了在特定土壤含水量的情况下, 介电常数虚部值随土壤盐分的增加而增大。由 RADARSAT-2 影像数据提取出来的后向散射系数 $\sigma(0)$ 与土壤样品测量的介电常数虚部的相关系数达到 0.70, 与实部只有 0.27, 与土壤样本的盐分的相关系数为 0.69^[51]。这表面土壤盐分对雷达图像记录的后向散射系数 $\sigma(0)$ 有重要的贡献, SAR 影像数据可用于较好地监测土壤盐分信息。为进一步分析本研究中样本点土壤导电性即土壤的电导率与土壤含盐量之间的关系, 将本次研究采样点测的土壤电导率与计算出的全盐含量进行了如下相关分析。

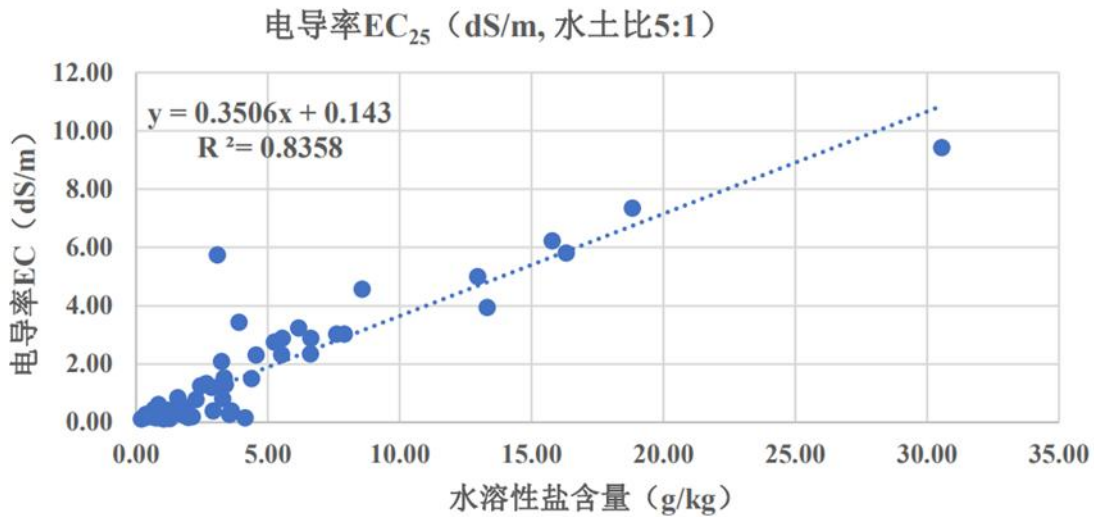


图2 研究区电导率与全盐含量的关系

从上图可知, 研究区电导率 Y 与水溶性盐含量 X 呈高度的线性关系, 它们的相关性 R^2 为 0.8358, 其公式为: $Y=0.3506X+0.143$ 。

因此, 当用雷达数据监测地面的土壤盐分时, 土壤含盐量的变化会影响土壤介电常数的变化, 因而会对雷达图像所观测的目标后向散射系数造成影响, 雷达影像可用于监测地面的土壤盐分。

2.3 土壤盐分的建模方法

为了将实测的土壤含盐量数据与所选的各种类型的特征进行分析及建模, 本文采用了机器学习的方法。目前, 常用于监测土壤含盐量的方法有逐步多元回归 (Stepwise multiple regression, SMR), 支持向量回归 (Support vector regression, SVR), 偏最小二乘回归 (Partial least squares regression, PLSR) 和随机森林 (Random forest regression, RFR) 等。本文采用了以上四种方法对该地区的土壤含盐量进行建模与分析。

2.3.1 逐步多元回归

逐步多元回归 (Stepwise multiple regression, SMR) 是一种具有筛选变量的多元线性回归模型, 它能较好地剔除变量中不重要的参数, 也能解决变量中具有高度共线性的问题。它的操作流程是逐个将自变量导入模型中, 每个导入的自变量都要经过 F 检验, 并对已选好的变量逐个做 T 检验由于导入的新变量使得旧变量不再显著, 就会自动地剔除旧变量, 这样就会得到最优的变量集模型。逐步回归常用方

法有向前法和向后法，这里用的是向前法，即变量由少到多，一个一个加入，直到没有可加入的变量。它的流程如下：

1) 将因变量 Y 与 n 个自变量 X_1, X_2, \dots, X_n 分别建立一元回归方程

$$Y = \beta_0 + \beta_i X_i + \varepsilon, i=1, 2, \dots, n \quad (11)$$

求出各变量 X_i 的回归系数的 F 检验统计量的值，记为 $F_1^{(1)}, \dots, F_n^{(1)}$ ，取其中最大的 $F_{i1}^{(1)}$ ，即

$$F_{i1}^{(1)} = \max \{F_1^{(1)}, F_2^{(1)}, \dots, F_n^{(1)}\} \quad (12)$$

在给定的显著性水平 α 下，记其对应的临界值为 $F^{(1)}$ ，若 $F_{i1}^{(1)} > F^{(1)}$ ，则 X_{i1} 其被选入模型中。

2) 建立 Y 与自变量 $\{X_{i1}, X_1\}, \dots, \{X_{i1}, X_{i-1}\}, \{X_{i1}, X_{i+1}\}, \dots, \{X_{i1}, X_n\}$ 的二元回归模型。

除去上面引入的变量 X_{i1} ，余下 $n-1$ 个自变量集，求出其中各变量回归系数的 F 检验统计量的值，记为 $F_1^{(2)}, \dots, F_n^{(2)}$ ，取其中最大的 $F_{i2}^{(2)}$ ，即

$$F_{i2}^{(2)} = \max \{F_1^{(2)}, F_2^{(2)}, \dots, F_{i-1}^{(2)}, F_{i+1}^{(2)}, \dots, F_n^{(2)}\} \quad (13)$$

集其临界值 $F^{(2)}$ ，若 $F_{i2}^{(2)} > F^{(2)}$ ，则 X_{i2} 被选入模型中，否则，终止选入变量。

3) 继续执行 Y 与变量集 $\{X_{i1}, X_{i2}, X_k\}$ 回归操作，重复 2 操作。

逐步回归分析可以用 SPSS 软件和 MATLAB 软件，本文选用软件 MATLAB 的 `stepwise` 函数对数据进行逐步回归操作，以土壤含盐量 Y 为因变量，17 个 SAR 极化分解参数、2 个后向散射系数、13 个盐度植被指数和 18 个 DEM 衍生因子共 50 个参数作为逐步回归的自变量。

2.3.2 支持向量机

支持向量回归 (Support vector regression, SVR) 是基于统计学习理论，作为支持向量机 (SVM) 的分支从而被提出的。与一般的线性回归相比，支持向量回归只有在预测值 $f(x)$ 和真实值 y 之间差距的绝对值大于阈值 ε 才计算损失，任何低于阈值 ε 都是可接受的。支持向量回归 (SVR) 将数据以一个非线性变换转换到一个高维的空间，可解决了“离散值多”和“过学习”问题，并以此构建回归模型来解决高维问题^[52]。SVM 有两个主要参数，它们是惩罚参数 C 和内核参数 γ ，其中参数 C 直接影响模型的稳定性， C 越高，说明越不能容忍出现误差，容易过拟合， C

值越小，容易欠拟合；参数 γ 是选择径向基核 RBF 作为 kernel 后，该函数自带的一个参数。 γ 越大，支持向量越少， γ 值越小，支持向量越多^[53]。支持向量的个数影响训练与预测的速度。SVR 有很多个版本，如 Segmentation Algorithms, Decomposition Algorithms, C-SVR, V-SVR 和 Sequential Minimal Optimization (SMO) 等^[54]。在本研究的土壤制图中，采用的是顺序最小优化算 (SMO)，因为它能推导出高性能模型^[55]。SMO-SVR 回归模型可表示为如下公式：

$$f(x) = \sum_{i=1}^m (\hat{\alpha}_i - \alpha_i) k(x_i^T x) + b \quad (14)$$

式中：

$\hat{\alpha}_i, \alpha_i$ —拉格朗日乘子；

$k(x_i^T x)$ —RBF 核函数。

本文利用 MATLAB 软件，采用台湾大学林智仁教授等开发的 SVM 模式识别与回归软件包 LIBSVM 进行支持向量回归模型的建立与预测。针对如何选择最优参数问题，常用的支持向量回归参数优化的方法有网格寻优 (GS)，遗传算法 (GA) 和粒子群算法 (PSO)，本研究使用的是支持向量机的 ε -回归形式，并使用基于粒子群的方法寻找 SVR 中最佳的 C 值和 γ 参数。

2.3.3 随机森林

随机森林(Random forest regression, RFR)是由 BREIMAN 于 2001 年把分类树组合在一起的一种集成学习算法，该算法具有非线性挖掘能力，良好的抗噪能力，数据的分布可不满足任何假设，对数据集适应能力强，训练速度快等优势^[56]。它可随机迭代地对数据和特征信息进行采样，来得到一大组分类和回归树或森林，目前已被广泛地应用于土壤成分的监测上，包括预测土壤有机质、土壤含水量、土壤养分和土壤含盐量。随机森林采用 bootstrap 聚合算法将原始训练集随机采样成 n 个训练子集，并从每个训练集中随机选取 K 个特征 ($K < n$)，反复根据这 K 个特征建立起 m 颗子决策树并求出它们的预测结果，最后，对分类模型进行投票，选出得票数最高的模型作为最终决策^[57]。具体流程如下图：

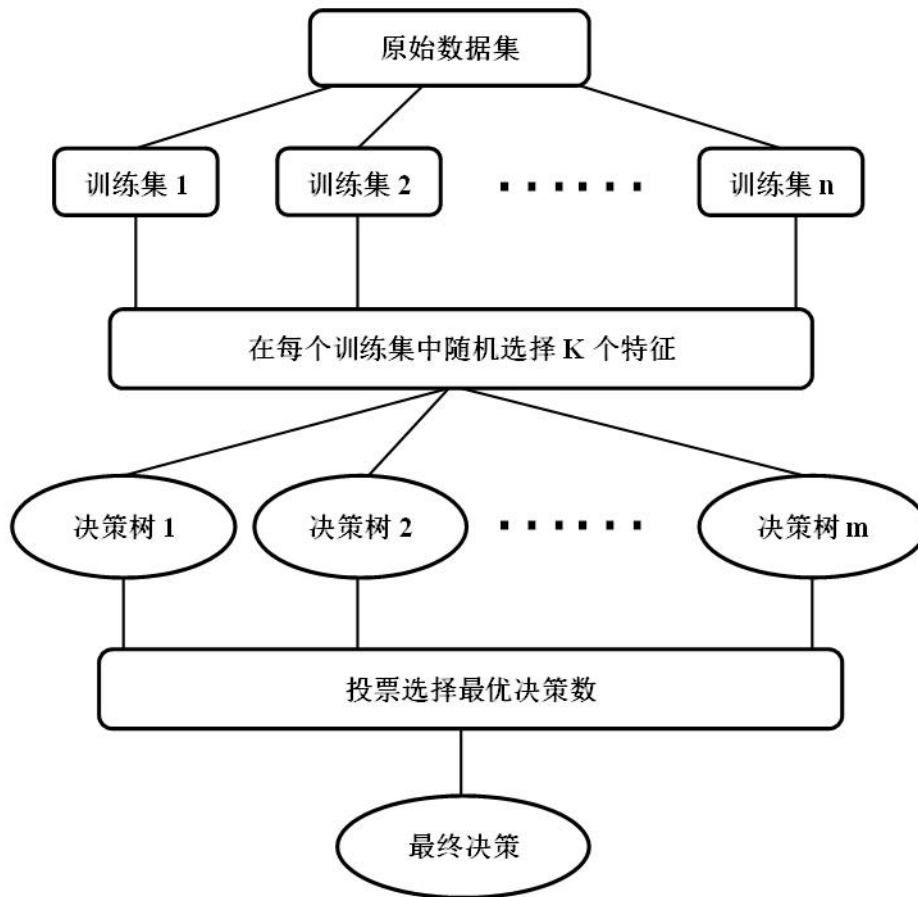


图 3 随机森林流程图

从图 3 的随机森林流程图中可以看出，一般随机森林模型有 3 个关键参数。决策树的数量（Num-Trees），每棵树的每个节点随机抽取的特征数（Mtry）以及每片树叶的最小观察次数（Min-Leaf）^[58]。本研究中使用的是 MATLAB 软件自带的 Tree-Bagger 函数，需要用户定义 Num-Trees 和 Min-Leaf 参数，一般默认情况下，对于回归，Min-Leaf 应设置为 5，为 1 时用作分类，而 Num-Tree 的默认值不确定它是否有可靠的结果，因此需要遍历确定 Num-Trees 的最优值。

2.3.4 偏最小二乘回归

偏最小二乘回归(Partial least squares regression, PLSR)是研究两个矩阵关系的一种方法，即可以研究多个自变量与多个因变量的关系，也能够在当自变量间存在多重共线性的情况下仍能得到较好的模型^[59]。目前，偏最小二乘回归已广泛地用在了遥感监测土壤成分上，如土壤含盐量^[60]，土壤养分^[61]，土壤有机碳^[62]和土壤含水量^[63]等。偏最小二乘回归是将预测矩阵 X 和观察矩阵 Y 重新投影到一个新空

间，来建立一个新的回归模型，具有很大地降低噪声的作用，筛选了一些解释能力强的信息，剔除了一些无用的信息影响^[64]。其一般的多元底层方程如下：

$$X = TP^T + E \quad (15)$$

$$Y = UQ^T + F \quad (16)$$

式中：

Y— $n \times p$ 的响应矩阵；

X— $n \times m$ 的建模矩阵；

T—X 的投影，即因子矩阵；

U—Y 的投影；

P, Q—正交载荷矩阵；

E, F—误差项。

偏最小二乘回归根据上述的估计因子 T 和 U，与载荷矩阵 P 和 Q 最终可建 Y 与 X 的线性模型，即：

$$Y = Xb + e \quad (17)$$

式中：

b—偏最小二乘回归的系数；

e—误差向量。

本研究将使用 PLSR 来讨论土壤含盐量与环境因子和雷达参数之间的关系，并建立土壤盐与各参数的回归模型。

2.4 本章小结

本章介绍了雷达影像数据的极化目标分解理论，并阐述了本文选用的极化分解方法 H/A/ α 及该极化分解的过程，并对这个极化分解的方法提取出的参数进行物理意义的研究与分析。接着对 Sentinel-1 SAR 影像数据进行 H/A/ α 极化分解操作，Sentinel-1 是双极化，只能生成 C2 矩阵，得到 17 个极化分解参数。同时，为了有效的利用数据反演土壤含盐量信息，本文引进了 4 种机器学习的方法来建模特征参数与土壤盐分的模型，并详细地介绍了各个模型的基本理论，为之后合理地运用这些模型来定量分析与预测研究区土壤含盐量的空间分布提供了理论依据和方法。

第3章 研究区域概况和数据来源

考虑到土壤盐碱化的成因是由多种因素的影响，如研究区的气候、降水、地形地貌、地下水和土壤类型等因素。本章简要介绍研究区的各种可能影响土壤盐分的因素情况及用于预测土壤含盐量的数据和所选用的特征。

3.1 研究区概况

3.1.1 地理位置

研究区位于山东省东营市，东营市位于山东省北部黄河三角洲地区，其地理位置为东经 $118^{\circ} 07' \sim 119^{\circ} 10'$ ，北纬 $36^{\circ} 55' \sim 38^{\circ} 10'$ 。东部濒临黄海，东北濒临渤海，东与垦利区，南部与淄博市、泰安市接壤，西部接壤滨州市，北与河口区相交。县境呈西南至东北向狭长带状，南北纵距 102.5 千米，东西 8.5~25 千米，总面积 1665.6 平方千米^[65]。

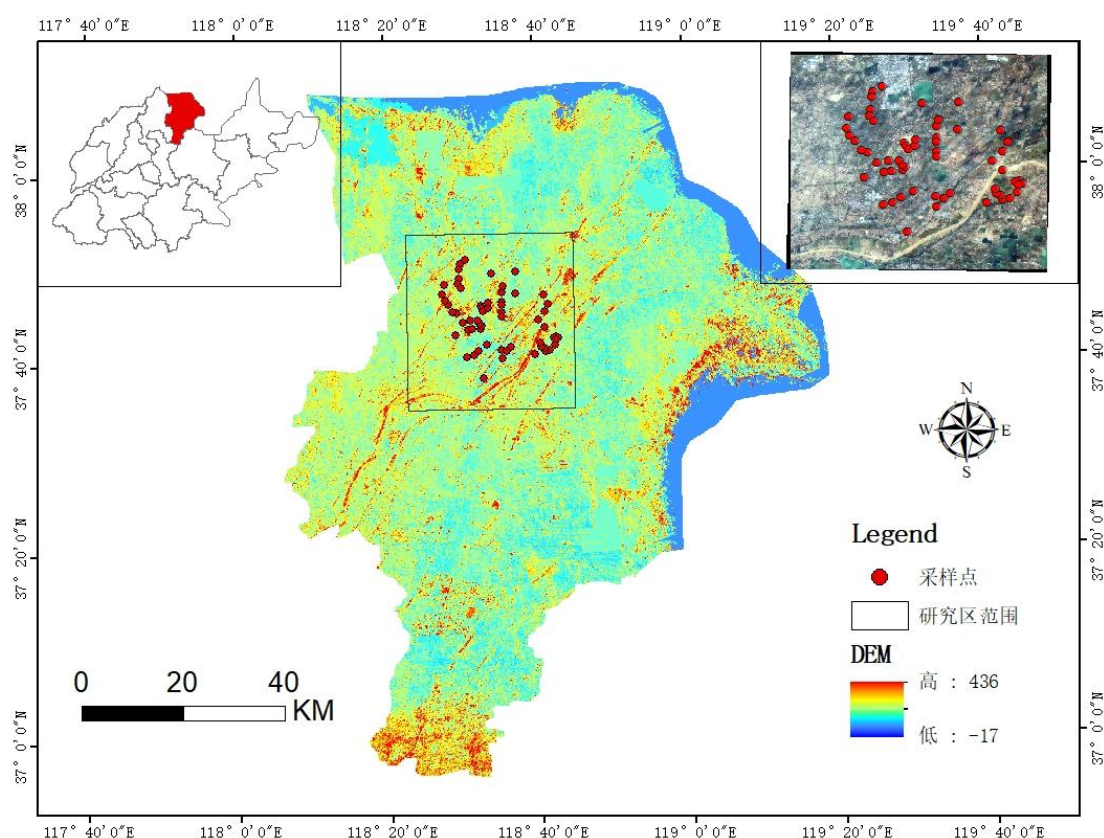


图4 研究区采样点分布示意图

3.1.2 气候降水

研究区地处温带大陆性季风气候区，大陆性气候占主导地位，气候温和，有丰富的光照资源且年辐射总量，为 $128.1\text{kw}/\text{cm}^2$ ，年均温为 12.8°C ，大于等于 10°C 的积温为 4300°C ，四季分明，伴随着降雨的季节性变化，主要降雨集中在夏季且温度是全年最高：春季降雨量较少，且伴随着大风，使其气候较为干燥；冬季降水稀少，蒸发作用强烈，极易形成季节性干旱和地表积盐的现象发生。年平均降水量为 550.9mm ，年蒸发量为 1944mm ，是降水总量的 3.55 倍，因此年降水量远远小于年蒸发量。由于存在季节性蒸散作用，使得地下水蒸发过程中带动着盐分上升到地表表层，土壤水分转化成大气水分，将土壤盐分留存于地表。

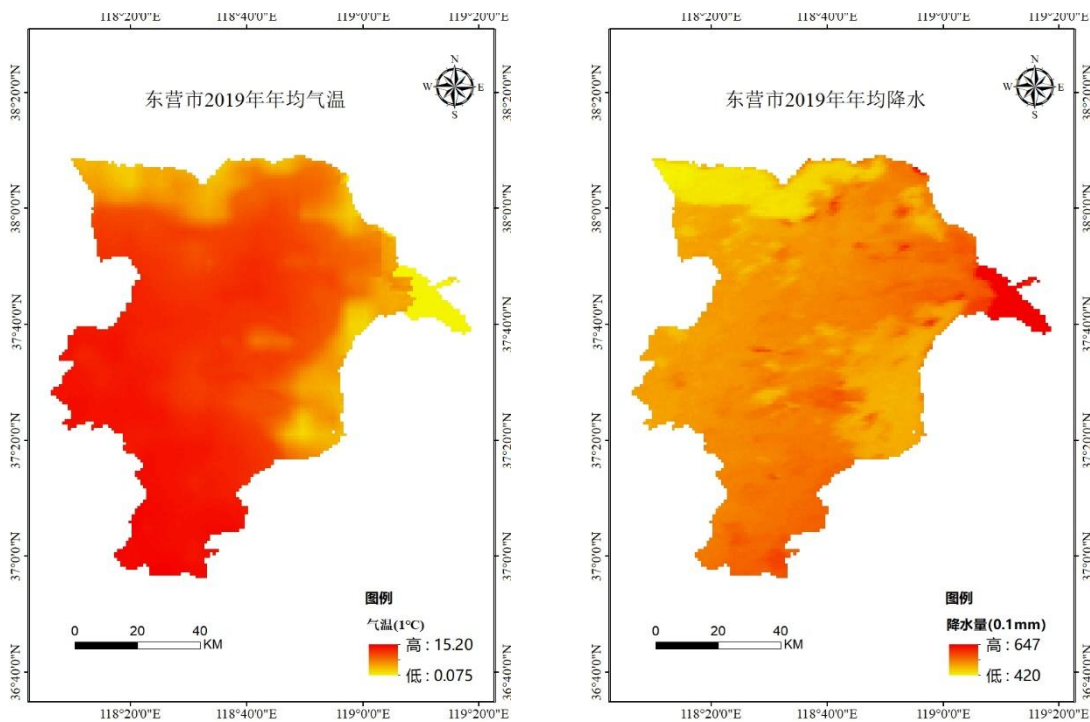


图 5 东营市 2019 年年均温度与降水量图

3.1.3 地形地貌

研究区地势沿黄河走向自西南向东北倾斜。西南部最高高程为 28 米，东北部最低高程 1 米，最高平均海拔 11.5m，沿海滩涂地面平均海拔为 2m。东部地区陆地形成的时间与西部地区差异较大，其中东部区域是由于黄河冲击形成的，形成时间较晚。研究区主要有 5 种微地貌：古河滩高地、河滩高地、微斜平地、浅平洼地和海滩地。其中占东营市总面积最多的是微斜平地，约占 54.54%，是岗、洼过渡地带；其次是海滩地，占市总面积的 27.05%，其与海岸线平行呈带状分布；接着

是浅平洼地，其占市总面积的 10.68%，在小清河南方主要分布在古河滩高地间，在小清河北边则分布于微斜平地之中、缓岗之间和黄河故道低洼处；随后是古河滩高地，占市总面积的 4.15%，主要分布于黄河决口扇面上游；最后，占市总面积最小的是河滩高地，占市总面积的 3.58%，主要分布于黄河河道至大堤之间。

3.1.4 土壤类型

东营市的土壤主要有褐土、砂姜黑土、潮土、盐土及水稻土五类。东营市约有 4%左右的土壤为褐土，其主要分布在小清河以南区域，适合种粮棉菜。约有 0.6%的土壤为砂姜黑土，其主要分布于小清河南褐土区的低洼处，适合农业种植。有 59%的土壤为潮土，占全市一半以上，适合种浅根作物。36%的土壤为盐土，主要在沿海区域，这些区域有很多芦苇、蒿子、茅草、黄须菜等自然植被，可开发水、牧养殖（植）。只有在利津和垦利老稻区有少量的水稻土，其只占 0.2%的土壤。

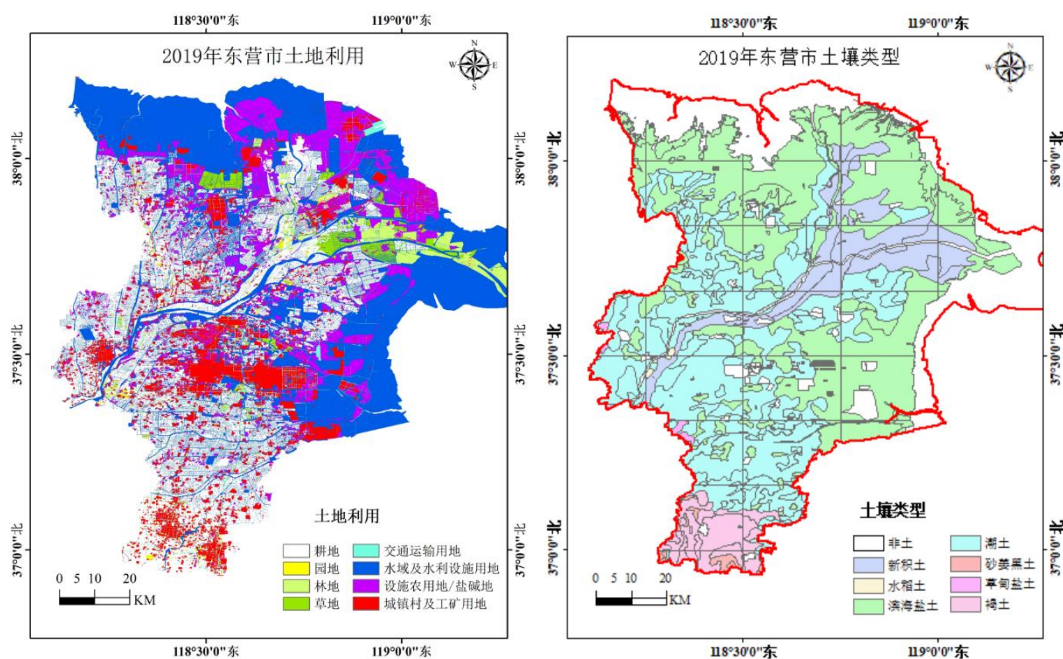


图6 东营市土地利用和土壤类型图

3.2 数据源介绍

土壤的含盐量与土壤的形成息息相关，土壤成土学说认为土壤的形成是受母质（parent material）、气候（climate）、生物（organism）、地形（relief）和时间（time）5个成土因子综合的影响^[31]。因而综合考虑可能影响土壤含盐量的因子会对实验模型的精度有很大的提高。前面已经介绍雷达数据的后向散射系数与介电常

数有一定的关系，而介电常数的变化又会影响土壤含盐量的变化。本文将在保留雷达数据优点的基础上，再结合各环境因子来反演该地区的土壤盐分信息。地形因子是最常用的环境因子，本文的地形因子数据由地理空间数据云平台提供的 30M 分辨率的 DEM 数据提取得出。生物要素如归一化植被指数（NDVI）等由 Sentinel-2 影像数据提取出各植被指数，母质及土壤要素也根据 Sentinel-2 影像数据计算各盐度指数。因此研究总共选取了 17 个 H/A/ α 极化分解参数，2 个后向散射系数，13 个盐度植被指数和 18 个地形指数作为土壤盐分预测的特征变量。具体选择的参数见表 1。

表 1 基于 Sentinel-1/2 和 DEM 衍生的特征变量

特征类型 Feature types	参数数量 Number	参选变量 Variable selection
H/A/ α 极化分解参数	17	Alpha, Entropy, Anisotropy, delta, lambda, Alpha1, Alpha2, delta1, delta2, L1, L2, p1, p2, HA, H(1-A), (1-H)A, (1-H)(1-A)
后向散射系数	2	Sigma0_VV, Sigma0_VH
盐度植被指数	13	土壤调节植被指数 SAVI, 植被土壤盐度指数 VSSI, 归一化盐度差异指数 NDSI, 盐度比 SI, 归一化植被指数 NDVI, 冠层盐度响应植被指数 CRSI, 亮度指数 BI, 盐度指数 SI1, 盐度指数 SI2, 盐度指数 SI3, 盐度指数 SI4, 盐度指数 SI5, 盐度指数 SI6
DEM 衍生因子	18	山体阴影 AH, 纵向曲率 LC, 切向曲率 TC, 收敛指数 CI, 总流域面积 TCA, 地形湿度指数 TWI, 坡长因子 LSF, 河网基准面 CNBL, 与河网距离 CND, 谷深 VD, 相对坡长位置 RSP, 高程 DEM, 坡向 AS, 坡度 S, 起伏度 QFD, 粗糙度 RS, 平面曲率 SC, 剖面曲率 PC

3.2.1 地面采样数据

采样设计综合考虑研究区土壤类型、植被类型、景观特征以及土地利用方式等因素。首先，需要开始在野外采集土壤样本。在现场采样过程中，随机选择了 60 个采样点，这些采样点均匀分布，采样深度为 0~20 cm，采样时间为 2019 年 10 月 9 日至 2019 年 12 月 14 日。所有的采样点尽量远离道路和建筑，考虑到这个区域有一些高压线，在采样过程中也应该远离高压线。因此，在野外采样时，为了安全和方便，一些采样点的选择在与道路相同的方向上，且选取的试验区之间的间距保持在 2km 以上，这个距离足以避免样本之间的空间依赖性。接着，进行了实验室的分析。土壤样本点的盐分的提取实验是采用电导率法，首先需要进行土样品的制备，具体方法是将土样带回实验室进行自然风干以去除杂质，并筛选小于 2mm 的土壤颗粒。然后，通过称量 20g 的土壤样品与 100mL 的去离子水配制成水土比为

5:1 的土壤溶液，采用电导率仪测定土壤溶液的 $EC_{1:5}$ ($EC_{1:5}$ 的值是以水土比为 5:1 进行提取的土壤溶液)，并利用电导法提取土壤中各离子组分的构成和组成含量，最后通过土壤含盐量与电导率之间的关系提取出土壤的全盐含量。

在实验之前，需要对采样点的一些土壤特性进行了描述性统计。表 2 描述了基于最小值、最大值、平均值、标准差和可变系数 (CV) 的四种土壤特性，包括土壤含盐量、EC 值、土壤含水量和 PH 值。从表中可以看出，整个采样区的土壤含盐量变化很大，在 0.20 至 30.56 g/kg 之间，CV 值大于 1。土壤中的 EC 值在 0.12 至 9.43 ds/m 之间变化，而 CV 接近土壤含盐量。土壤含水量 (SMC) 变化不大，为 6.71%~26.51%，而 CV 值为 31.31%。pH 值在 7.53~8.98 之间，CV 值最小。

表 2 土壤特征的描述性统计

土壤特征 Soil Properties	最小值 Minimum	最大值 Maximum	平均值 Mean	标准差 Standard Deviation	方差 Variance	可变系数 Variable Coefficient/%
EC/(ds·m ⁻¹)	0.12	9.43	1.52	2.02	4.09	132.89
Salinity/(g·kg ⁻¹)	0.20	30.56	4.09	5.38	27.64	131.54
SMC/%	6.71	26.51	16.80	5.26	27.64	31.31
PH value	7.53	8.98	8.14	0.31	0.098	3.81

注：EC 表示电导率，单位 ds/m；SMC 表示土壤含水量，%。

上表列出了四种土壤特征的特征统计数据，与以往研究的范围类似，表明本试验的采样点具有代表性。根据土壤含盐量的分级标准，土壤盐度可分为以下五个等级：无、轻度、中度、高度和极端^[66]。表 3 根据 EC 值的取值范围显示了土壤盐碱化标准分类，主要有如下 5 类：非盐碱化、轻度盐碱化、中度盐碱化、高度盐碱化和极度盐碱化。

表 3 土壤盐分分类标准

土壤盐分分类 Salinity Class	饱和土壤浸提液 EC (ds/m)	各程度盐分百分比 Proportion/%	植物反应 Plant Response
非盐碱化	< 2	70.2	对作物不产生盐害
轻度盐碱化	2~4	19.2	对盐分极敏感的作物产量可能受到影响
中度盐碱化	4~8	8.5	对盐分敏感作物产量受影响，但对耐盐植物无多大影响
重度盐碱化	8~16	2.1	只有耐盐作物有收成，但影响产量
极重盐碱化	> 16	0	只有少数耐盐植物生长

上表的结果表明，约 70.2% 的采样点属于非盐渍土，19.2% 的采样点属于轻度盐渍土，8.5% 的采样点属于中度盐渍土，2.1% 的采样点属于高度盐渍土，采样区

域内没有极度盐渍土。

3.2.2 Sentinel-1 数据

在本文中，使用的是从欧洲航天局（ESA）免费在线门户网站获取的 C 波段的 Sentinel-1 SAR 产品。Sentinel-1 由两颗极轨卫星 A 星和 B 星组成，这两颗卫星搭载的传感器为合成孔径雷达（SAR），属于主动微波遥感卫星，可穿透云层和地表。Sentinel-1 为每颗卫星提供 12 天重访周期的有源 SAR 数据。本研究使用了从欧洲航天局数据中心下载的 2019 年 10 月 8 号的 Sentinel-1 数据的单视复数影像（SLC）。具体参数见表 4 所示。

表 4 研究区 Sentinel-1 数据主要参数

参数 Parameter	值 Value	参数 Parameter	值 Value
Date	8 October 2019	Processing level	Level 1
Polarization	VV, VH	Frequency	5.4 GHz
Relative orbit	69	Range pixel spacing	2.32 m
Orbit direction	Ascending	Azimuth pixel spacing	13.93 m
Mode	IW	Ground resolution (Range × Azimuth)	≈5 m × 20 m
File format	SAFE	Incidence angle	39.15°

3.2.3 Sentinel-2 数据

哨兵 2 号卫星（Sentinel-2A 和 Sentinel-2B）携带一枚多光谱成像仪(MSI)，可覆盖 13 个光谱波段，从 786km 的高度对地球表面进行成像，确保时间分辨率为 5 天，其也具有更高的空间分辨率，其最高波段的分辨率可达 10 米。与 Landsat TM、ETM+（重访周期为 16 天）等遥感影像相比，Sentinel-2 影像具有更高的空间分辨率、光谱分辨率及更短的重访周期，其可见光波段的地面分辨率达 10m。本文选择的 Sentinel-2A 影像数据的云覆盖百分比小于 10%，选取的时间是 2019 年 10 月 19 日，与土壤含盐量野外样本采集的时间是最接近最符合的。影像的预处理工作包括辐射校正、大气校正、重采样、影像拼接与裁剪等。如果是 L1A 级产品数据，需要利用欧空局提供的 Sen2cor 数据处理软件对 Sentinel-2A 遥感影像进行辐射校正与大气校正，本文选用的是已经过辐射定标和大气校正的地表反射率产品 L2A，因此不需要再进行辐射定标和大气校正处理。本文选取的 Sentinel-2A 影像数据的具体参数如表 5 所示。

表5 研究区 Sentinel-2 数据主要参数

参数 Parameter	值 Value	参数 Parameter	值 Value
Date	Oct 19th, 2019	Processing level	Level-2A
Relative orbit	132	Cloud cover percentage	3.269213
Orbit direction	Descending	Resolution	10m、20m、60m
File format	SAFE	Time resolution	10 d

3.2.4 DEM 数据

土壤盐渍化的形成是一个复杂的过程，如母质、气候、地形、生物等因素都会对土壤盐渍化形成产生重要作用。为了综合考虑土壤盐渍化的影响因子，本研究也选取了地形衍生因子进行土壤盐渍化的反演。最终选取了18个地形因子作为参选变量来反演该地区的土壤盐分。其中DEM数据在地理空间数据云下载的ASTER GDEM V2数据，分辨率为30m，将下载后的DEM数据先在ARCGIS上作预处理，如图像拼接，裁剪，投影等，再将预处理后的数据运用SAGA GIS进行处理。在处理前最好先作填凹处理，可以得到14个基础地形参数，运用ARCGIS从投影的30m数字高程模型计算平面曲率，剖面曲率，起伏度和粗糙度等。

3.3 本章小结

本章主要介绍了研究区的概况和预测土壤含盐量所使用的数据源。研究区概况介绍了研究的地理位置，该研究区位于东营市内黄河三角洲，这里的农作物和植被深受土壤盐碱化的影响，因此选此作为研究区，具有代表性和实际意义。此外，还介绍了该研究区的气候降水、地形地貌和土壤类型等，它们都是潜在可能影响研究区土壤含盐量的环境因素或人为因素，了解此研究区的这些信息对土壤含盐量的反演及分析具有一定的必要性；紧接着介绍了本文用于反演研究区土壤盐分所使用的数据，主要包括Sentinel-1 SAR 雷达影像、Sentinel-2 MSI 光学影像和DEM高程数据，介绍了这些数据源的选择原则及其一些参数的信息，此外也介绍了地面土壤含盐量实测数据的测量方法及其采样点数据的空间分布情况，并制作了采样数据的土壤特征的描述性统计及其各类盐渍化土壤的分类标准。因此本章主要介绍了关于土壤含盐量反演的所使用的数据源。

第 4 章 监测信息的数据处理

为了后续研究能较好地分析与比较研究区的土壤含盐量的建模及预测结果，首先需要对所选的遥感影像数据和 DEM 数据进行预处理，提取本文所需要的雷达特征、盐度指数和地形因子。为了保证土壤盐分实测数据的有效性，也需要对野外采样的土壤盐分的样本数据进行异常值的确定并剔除它以使模型能得到更好地预测表现。因此本章详细介绍了本文所使用的各类数据源的预处理、特征的提取及土壤盐分的异常值的剔除。

4.1 研究区数据预处理

由于原始的影像或数据不能直接提供有用的参数，因此，需要先对遥感影像和 DEM 数据做预处理操作，以便后续能较好地提供本文所需要用到的参数。

4.1.1 Sentinel-1 SAR 雷达数据

将从欧空局下载后 Sentinel-1 SAR 影像数据在 SNAP 软件上进行预处理，并提取该地区影像的后向散射系数。本文采用的是 Sentinel-1 GRD 格式的数据提取后向散射系数，因为 GRD 数据是同一景影像的 SLC 数据体量的 1/4 左右，方便进行后续操作。单数复数影像 SLC 数据（Single Look Complex）可经多视处理及 WGS84 椭球投影至地距后得到地距多视影像 GRD 数据（Ground Range Detected）。以下是获得后向散射系数的操作步骤。

1) 轨道校正

将覆盖研究区的 Sentinel-1 IW GRD 雷达影像数据导入 SNAP 软件中进行轨道校正，可 SNAP 软件直接从欧空局下载轨道文件。Sentinel-1 元数据里包括轨道相关参数，但精度不是很高。轨道校正可把元数据.xml 文件里描述轨道的参数值，如星历数据，以更精确的轨道文件替换掉。若轨道数据不准确，将会给后续数据处理带来不少误差。



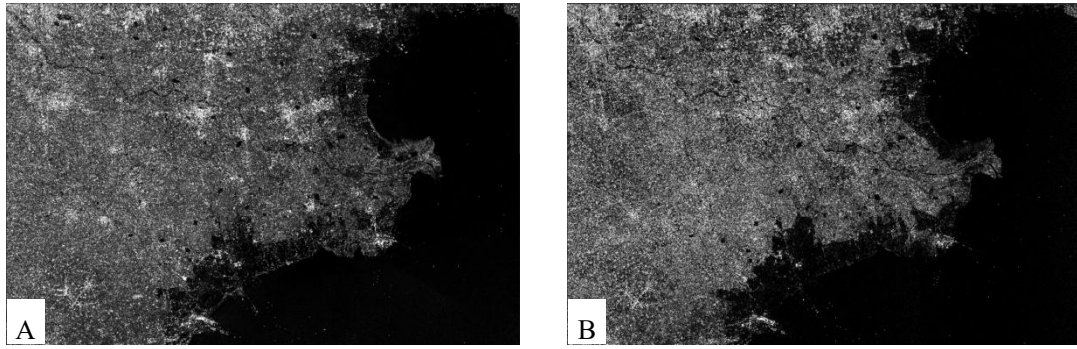
图 7 Sentinel-1 原始影像图

2) 热噪声去除

热噪声是 SAR 卫星系统自带的噪声，即背景噪声，所获得的 SAR 影像一般都应有热噪声，且由于 SAR 天线发射机需要发射出很强的功率，造成的内部的热量是不可忽视的，它会对后续需要获得的后向散射系数的精度造成一定程度的影响，因此在对影像数据的处理过程中需要消除热噪声的影响。本文利用 SNAP 软件对 Sentinel-1 影响的热噪声进行了处理。

3) 辐射定标

辐射定标是把接收的能量，即后向散射信号转化为有单位的物理量或一些无单位的比例值。由于 SAR 影像数据本身具有穿透性，可穿透云雾的遮挡，不受大气云层的影响，因此只需要对 SAR 影像做辐射定标就行，不需要对其进行光学影像上的大气校正处理。在 SNAP 软件中，有三种类型的后向散射系数可供选择，如 σ_0 , γ_0 , β_0 ，一般研究选择的都是归一化的雷达后向散射系数（雷达散射截面） σ_0 即可，后两种是在第一种的基础上做了一些修正得到的，本文选择的是常用的 σ_0 ，结果如下。

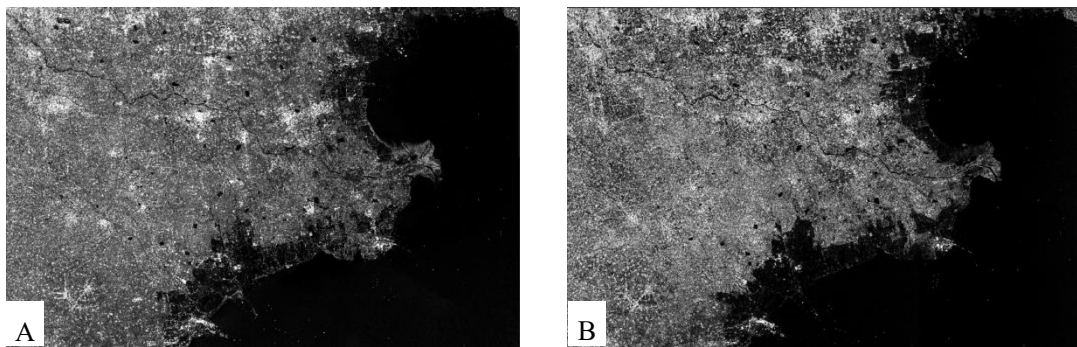


A—Sigma0_VV; B—Sigma0_VH.

图 8 雷达影像后向散射系数图

4) 斑点滤波

相干斑是 SAR 影像常见的现象，对于 SAR 分类等应用来说，它是噪声。为了提高土壤盐分的反演精度，本文选用的是比较常用的 Refined Lee 滤波（改进的 Lee 滤波器），它是一种自适应的滤波器，处理效果比较好，可在相除相干斑的影响下较好地保留极化信息。滤波窗口选用滤波效果较好的 7*7 窗口。可以从图中看出，经滤波处理后影像的斑点噪声减少了。



A—Sigma0_VV; B—Sigma0_VH.

图 9 滤波后 Sentinel-1 灰度值图

对比滤波前后两个后向散射系数的灰度值图，可以看出滤波后的两幅后向散射系数的灰度值影像图中的斑点噪声减少了。

5) 地形校正

对极化 SAR 数据做地形校正，不仅对影像赋予了实际坐标信息，还进行了地形辐射校正处理。本文在 SNAP 软件中使用的是距离多普勒地形校正法，数据重采样采用的是默认的双线性插值法。

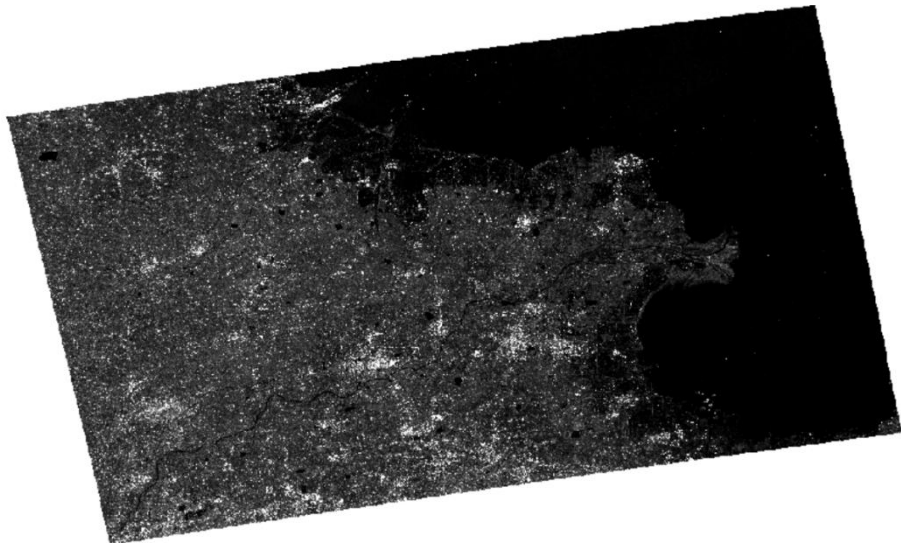


图 10 地形校正后研究区 Sentinel-1 影像图

6) 分贝化

通过上面处理后得到的是线性比例单位的后向散射系数，其值一般为较小的正数，但很小，它是由于传输的距离较远，接收器收到的功率较小，即雷达后向散射很小。为后续数据分析及可视化等操作的方便，需要将后向散射系数分贝化，其公式如下：

$$\sigma_0(dB) = 10 * \log_{10}(\sigma_0) \quad (18)$$

将影像分贝化后的结果如下：

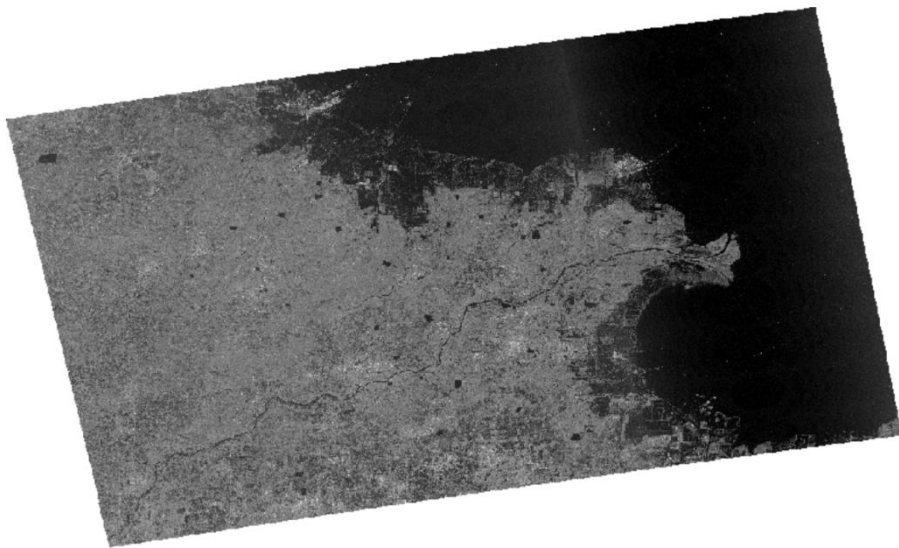


图 11 分贝化后 Sentinel-1 影像图

对比分贝化前后的影像图可以看出，分贝化后的影像更易于看出很多细节的地

方，图中的纹理也更清晰了，如图中的黄河区域线条更加明显。

接着需要对 Sentinel-1 IW SLC 影像数据进行预处理，SLC 包含相位等信息，可通过处理 SLC 数据得到极化协方差矩阵，进而得到极化特征，而 GRD 无法得到极化协方差矩阵，也就无法获得极化特征，只能获得一些纹理特征或几何特征，其步骤主要包括轨道校正、辐射定标、DEBURST、多视、地形校正和极化滤波等预处理操作。由于上述有些操作已经介绍过，这里重点介绍不一样的操作，如辐射定标、多视、极化滤波和极化分解操作。

1) 辐射定标

极化 SAR 数据的辐射定标不同于光学数据的，对其需要进行复定标处理，也不是上面选择的 σ_0 这种，它的处理过程是对复数的实部和虚部都进行定标处理。

2) 多视

多视处理的目的是为了获得近似正方形的像素（方位像与距离向），它不仅可消除或减弱相干斑噪声，而且也大大地减少了后续的数据量。本文利用 SNAP 软件的 Multi-look 工具，根据斜距的方位向和距离向的分辨率及入射角，可计算除多视处理的视数，本文多视操作选取的视数为 1×5 。下图是多视后以 C11 为红色，C22 为绿色，C11/C22 为蓝色合成的影像图。



图 12 多视后的 Sentinel-1 影像

3) 极化滤波

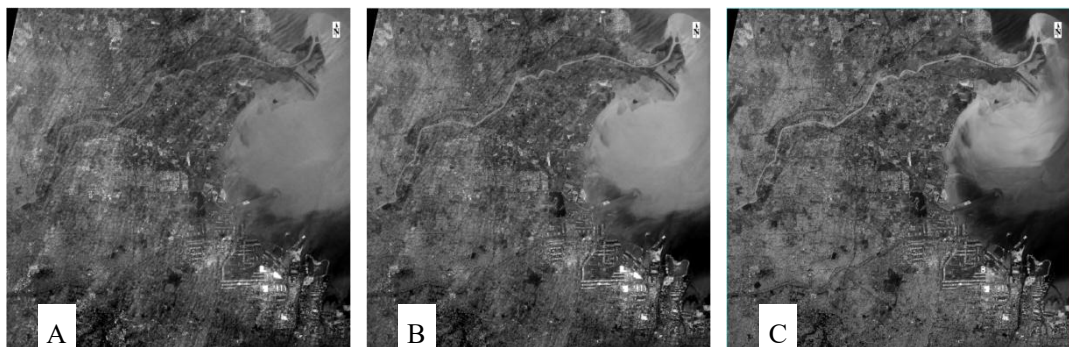
本文在 POLSARPRO 软件上进行极化滤波处理，采用滤波器为 Lee Refined Filter，即改进型 Lee 滤波或叫精致型 Lee 滤波，选择这个滤波器的原因是其可以在相除相干斑的影响下较好地保留极化信息，一般情况下，做极化分解所用的滤波器都是这个。这里的滤波窗口选择默认的 7×7 窗口。

4.1.2 Sentinel-2 MSI 光学数据

为了能提取出研究需要的盐度植被指数，需要先对原始的 Sentinel-2 光学影像进行预处理操作。本文使用的是已经经过辐射定标和大气校正的 L2A 级产品，因此本文不需要再进行辐射定标和大气校正这些预处理步骤，只需要进行重采样、格式转化和波段融合。本文对 Sentinel-2 影像的预处理步骤如下。

1) 重采样是一种影像数据重新组织过程中的灰度处理方法。常用的重采样方法有最邻近内插法、双线性内插法和三次卷积法内插。重采样可分为上采样和下采样，如果是大于原信号就是上采样，反之则为下采样，上采样也就是内插或插值。本文将从欧空局下载后 Sentinel-2 MSI 影像数据在 SNAP V8.0 软件上进行预处理。首先需要将 Sentinel-2 L2A 级产品导入到 SNAP V8.0 软件中，对其进行重采样，采样分辨率以光学波段 B2（分辨率为 10 m）为主，本文重采样方法使用的是最邻近法。

2) 格式转换。由于目前的格式并不能在 ENVI 中打开，因此需要将其格式转换为 ENVI 格式文件。导出后的波段信息是一段一段分开的。部分波段信息如下。



A—Blue; B—Green; C—Red; .

图 13 Sentinel-2 影像导出波段

3) 由于 SNAP 软件对 Sentinel-2 影像数据重采样后输出的各个光谱波段信息分开储存的，因此只需选择研究所需要的六种波段（Blue, Green, Red, NIR, SWIR1 和 SWIR2）进行波段融合，以便后续更方便盐度植被指数的提取。实验使用 ENVI Classic 软件对重采样后的影像进行波段融合。波段融合所使用的投影坐

标系为 WGS_1984_UTM_Zone_50N，空间分辨率为 10m*10m。将波段融合的 Sentinel-2 影像在 ENVI 上显示其真彩色影像图（图 14）。



图 14 研究区 Sentinel-2 真彩色影像图

4.1.3 DEM 数据

本文的 DEM 数据是从地理空间数据云平台下载的，选取的数据集为 30 米分辨率的 ASTER GDEM 数字高程数据。选取的空间位置是按照行政区选取的，结果显示研究区需要 4 副 DEM 数据。因此，为后续更方便的进行数据处理及出图，首先需要对 DEM 数据在 ARCGIS 上作预处理。本文的预处理步骤主要有图像拼接，裁剪和投影。

1) 图像拼接是指将两幅或多幅影像拼在一起，构成一幅整体影像的技术过程。本文通过使用 ARCGIS 软件的镶嵌至新栅格，将下载的 4 副 DEM 影像拼接。

2) 图像裁剪。本文使用 ARCGIS 软件分析工具里的按掩模提取功能，并根据研究区的范围对拼接的 DEM 数据进行裁剪。结果如下图所示。

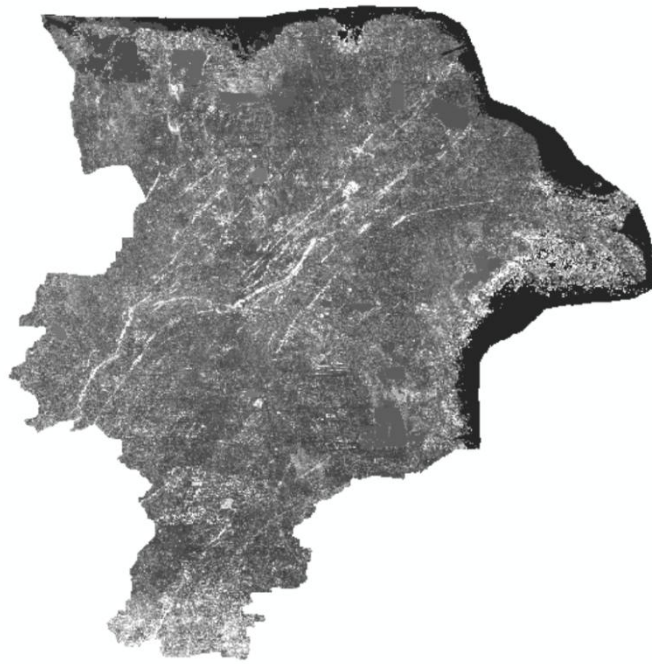


图 15 研究区 DEM 采集

3) 投影。本文使用 ARCGIS 软件投影和变换里的投影栅格对研究区的 DEM 数据进行投影, 输出坐标系为 WGS_1984_UTM_Zone_50N。

4.2 特征提取

基于上一节对研究区影像数据的预处理后, 接着就可以开始对从不同的数据源的雷达影像特征、盐度植被指数和地形因子这些参数进行提取, 为后续对研究区土壤含盐量的反演提供了数据准备。

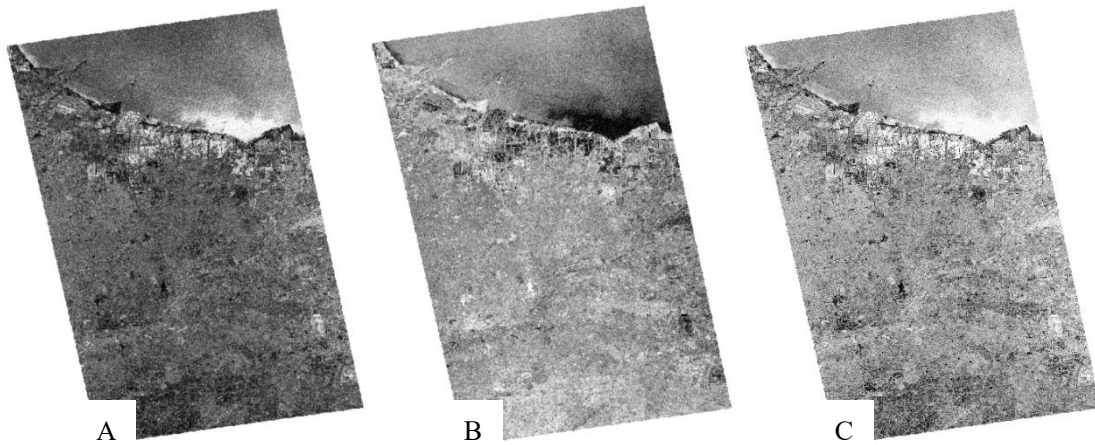
4.2.1 Sentinel-1 SAR 雷达特征

雷达数据分解的不同极化特征在对地物的识别上有所区别, 因此有效地获取不同的极化特征在准确地识别地物特性上有极为重要的作用。本文在基于前述对研究区 Sentinel-1 SAR 数据进行轨道校正、辐射定标、生成 C2 协方差矩阵、多视、地理校正和极化滤波等预处理的基础上, 最后通过使用 POLSARPRO 软件对 Sentinel-1 数据滤波后的影像进行 $H/A/\alpha$ 极化分解, 从而获取了研究区的极化特征信息, 最终得到的 $H/A/\alpha$ 极化分解参数共 17 个和 2 个后向散射系数共 19 个特征参数。具体参数见表 6。

表 6 H/A/α 极化分解参数

特征类型 Feature types	描述 Description	参数 Parameters	来源 Source
原始后向散射 Original Backscatter features	归一化后向散射系数 Normalized backscatter coefficient	Sigma0_VV	SNAP
		Sigma0_VH	SNAP
	散射熵 Entropy		POLSARPRO
	各向异性 Anisotropy		POLSARPRO
极化分解参数 Polarization decomposition parameter	平均散射角 Alpha		POLSARPRO
	特征值 1 L1		POLSARPRO
	特征值 2 L2		POLSARPRO
	H/A/α极化分解参数 Delta1, Delta2, p1, p2, Alpha1 Alpha2, Delta lambda, HA, (1-H) (1-A), H(1-A), (1-H) A		POLSARPRO

接着，展示了极化分解处理后得到的部分极化特征的灰度图。部分 $H/A/\alpha$ 极化分解参数的结果如图 16 所示。



A—Alpha; B—Anisotropy; C—Entropy.

图 16 H/A/α极化分解参数

最后需要将上述提取的各影像特征在 ENVI Classic 软件中进行波段融合，以便后续可批量提取各特征影像在采样点的灰度值，接着将其导出为 TIFF 格式，再在 ARCGIS 处理软件上通过使用空间分析工具的提取分析功能，并根据采样点的经纬度坐标，来批量提取出各采样点对应的各个类型的雷达极化特征的灰度值。

4.2.2 Sentinel-2 MSI 光学特征

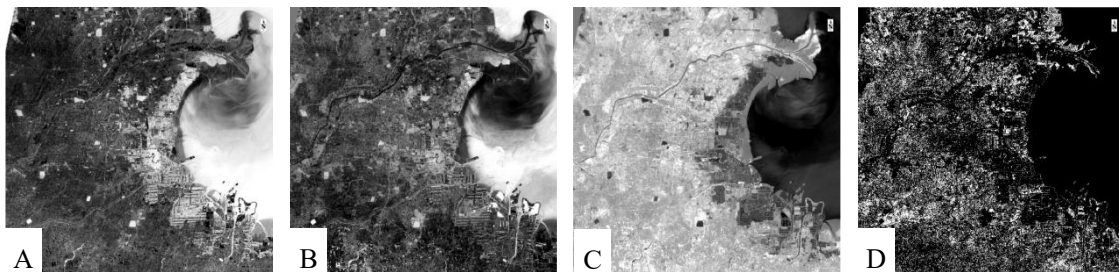
在基于 Sentinel-2 光学影像预处理的基础上，导入了本文所需要的 6 个 Sentinel-2 影像波段，随后在 ENVI 软件中使用 Band Math 功能，按如下公式计算

出各植被指数和盐度指数，如表 7 所示。

表 7 研究区植被盐度指数

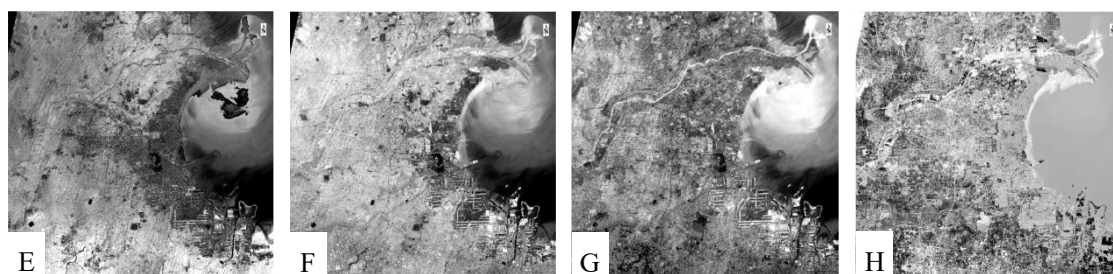
特征类型 Feature types	所选参数 Parameters	所选公式 Formulations	参考文献 References
植被指数 Vegetation Index	NDVI	$\frac{F_{NIR} - F_R}{F_{NIR} + F_R}$	[33]
	SAVI	$\frac{(F_{NIR} - F_R) * (1 + L)}{(F_{NIR} + F_R + L)}$	[33]
	NSI	$\frac{F_{SWIR1} - F_{SWIR2}}{F_{SWIR1} + F_{NIR}}$	[67]
	VSSI	$2 * F_G - 5 * (F_R + F_{NIR})$	[68]
	NDSI	$\frac{F_R - F_{NIR}}{F_R + F_{NIR}}$	[69]
	SR	$\frac{F_G - F_R}{F_B + F_R}$	[67]
	CRSI	$\sqrt{\frac{F_{NIR} * F_R - F_G * F_B}{F_{NIR} * F_R + F_G * F_B}}$	[70]
	BI	$\sqrt{F_G^2 + F_R^2 + F_{NIR}^2}$	[69]
	SI1	$\sqrt{(F_R * F_G)}$	[70]
	SI2	$\sqrt{(F_R * F_B)}$	[40]
盐度指数 Salinity Index	SI3	$\sqrt{(F_R^2 * F_G^2)}$	[40]
	SI4	$\frac{F_{NIR} * F_{SWIR1} - F_{SWIR1}^2}{F_{NIR}}$	[67]
	SI5	F_B / F_R	[44]
	SI6	$\frac{F_R * F_{NIR}}{F_G}$	[44]

最后，在 ENVI 上可得到以上 8 个植被指数和 6 个盐度指数特征图，并展示处理后得到的部分特征影像图。部分植被指数和盐度指数的结果如图 17 所示。



A—VSSI; B—SR; C—SAVI; D—NSI; E—CRSI; F—BI; G—SI1; H—SI4.

图 17 研究区盐度植被指数图



A—VSSI; B—SR; C—SAVI; D—NSI; E—CRSI; F—BI; G—SI1; H—SI4.

图 17 研究区盐度植被指数图（续）

最后将上述提取的各影像特征在 ENVI Classic 软件中进行波段融合，以便后续可批量提取各个特征影像在采样点的灰度值，接着将其导出为 TIFF 格式，再在 ARCGIS 软件中使用空间分析工具的提取分析，并根据采样点的经纬度坐标，来提取出各采样点对应的各植被指数和盐度指数的灰度值。

4.2.3 地形因子

本文的地形因子根据可能影响土壤盐分的基础上选择了如表 8 所示地形因子。

表 8 DEM 衍生的地形参数

特征类型 Feature types	描述 Description	参数 Parameters	参考文献 References
地形指数 Topographic index	高程	DEM	[36]
	山体阴影	AH	[35]
	坡向	AS	[71]
	坡度	S	[71]
	纵向曲率	LC	[71]
	切向曲率	TC	[71]
	河网距离	CND	[72]
	河网基准面	CNBL	[72]
	总流域面积	TCA	[72]
	地形湿度	TWI	[35]
	坡长因子	LSF	[35]
	收敛指数	CI	[35]
	相对坡度位置	RSP	[35]
	谷深	VD	[73]
	起伏度	RS	[71]
	粗糙度	CCD	[71]
	平面曲率	SC	[71]
	剖面曲率	PC	[71]

在基于上一节对 DEM 影像数据的预处理的基础上，使用 ARCGIS 软件的 3D Analyst Tools 里的功能，该功能支持对坡度，坡向，平面曲率，剖面曲率和总曲率等的计算。接着分别将投影后的 30m 分辨率的数字高程模型中放入不同的功能中可计算出坡度、坡向、平面曲率、剖面曲率、起伏度和粗糙度；运用 SAGA GIS 软件对预处理后的数据进行处理，可以得到 14 个基础地形参数（在处理前最好先作填凹处理）。

图 18 展示了部分计算的地形因子可视化图，其结果如下所示：

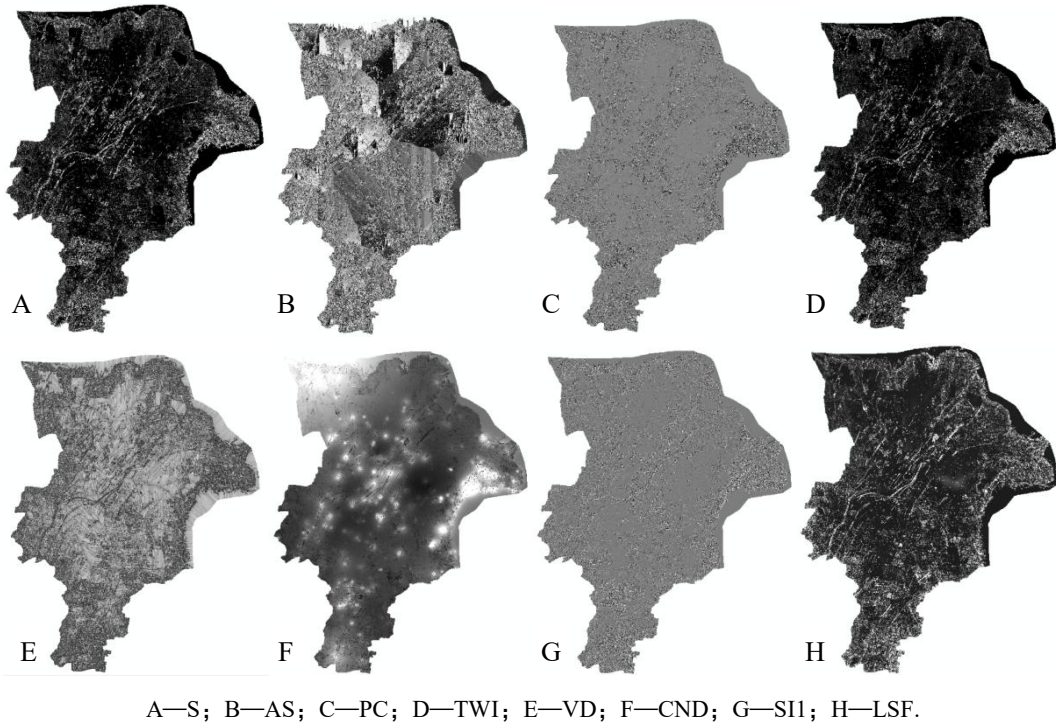


图 18 研究区地形因子图

最后将上述提取的各个影像特征在 ARCGIS 软件中使用空间分析工具里的多值提取至点功能，并根据采样点的经纬度坐标矢量文件，来批量提取出各采样点对应的各个类型的地形因子的灰度值。

4.3 土壤盐异常值的确定

为了确保土壤盐分实测数据的有效性，需要对野外采样的土壤盐分的样本数据进行异常值的剔除。本文采用了数据处理软件的 RCOPLLOT 函数来识别并剔除离散值。在将雷达特征、盐度植被指数和地形因子这三类特征在采样点对应的灰度值进行提取的基础上，建立了土壤盐分与这些特征的土壤盐分特征数据集。将其导入 RCOPLLOT 函数后，可根据作图后如果存在红颜色的竖条，判断该点为离群点。该

方法可简单地用 3 倍标准偏差来判断可疑数据是否应该舍去。原理由于在多次试验中，根据随机变量的正态分布规律，其测量值落在 $X-3S$ 与 $X+3S$ 之间的概率为 99.73%，出现在此外的概率只有 0.27%，这种情况出现的可能性极小，几乎不可能的事。因此在实验中，如果出现这个情况，就认为该测量数据是不可靠的，应将其舍弃。在本实验过程中，将野外实测的土壤盐分值作为测量值 Y ，将提取的雷达特征、盐度植被指数和地形因子作为预测参数 X_i 输入。当测量数据与测量结果的算术平均值的差超过 3 倍标准偏差时，则应舍弃。

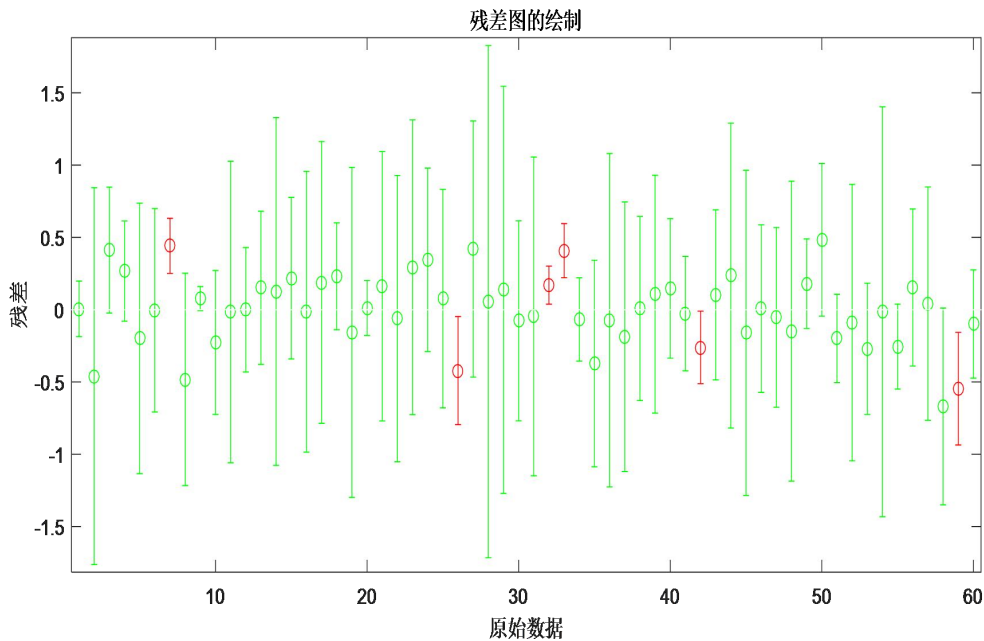


图 19 异常值的分布

图 19 中 X 轴是 60 组采样点的原始数据，Y 轴是残差的分布情况。根据上图的异常点的分布，删除了图中红色所示的 6 组土壤盐分的实测数据，最终得到了 54 组有效样本点，用于后续研究区土壤盐分的建模及验证。

4.4 本章小结

本章主要描述了本文实践中用于推导研究区域土壤盐度图的数据处理流程和特征提取过程，并详细介绍了在运用本文 4 种模型对研究区土壤含盐量进行反演研究前的数据准备及处理过程，主要介绍了对 Sentinel-1 雷达数据、Sentinel-2 光学影像数据和 DEM 高程数据的预处理工作及其相应参数的提取流程。其中 Sentinel-1 SAR 数据的预处理步骤包括了轨道校正、辐射定标、多视、地理校正、滤波处理等预处理过程；Sentinel-1 光学影像预处理主要包括辐射定标、大气校正、重采样

和盐度植被指数的提取，为之后的土壤含盐量的预测提供了数据支持。因此本章的主要内容是正式开始土壤含盐量的反演计算的数据预处理的阶段。接着也将实测土壤盐分数据与这些提取的参数建立数据集，并对这些特征参数进行划分数据集。最后，为确保数据的有效性，剔除了土壤盐分实测数据的异常值。因此，本章内容主要介绍了土壤含盐量定量反演的数据准备和特征提取及筛选的介绍。

第 5 章 结果与分析

5.1 特征筛选

本文综合考虑了从各种类型的数据源提取出的 17 个 H/A/ α 极化分解参数, 2 个后向散射系数, 13 个盐度植被指数和 18 个地形指数(具体参数见表 1), 以期能筛选出最适应于该地区土壤含盐量反演的特征参数, 因此本文采用了随机森林的特征重要性得分来评估和筛选出该区域具有最大贡献度的参数。此外, 不是使用的特征越多对模型的表现越好, 相反一些没有贡献性的特征输入模型中, 反而可能会降低模型的表现, 因此在建模前, 首先对众多的特征进行筛选, 也有利于提高各种模型的预测表现。

5.1.1 特征筛选的方法

本次研究中, 由于选取的参数众多, 首先需要排除一些不重要的参数的干扰, 本文采用了随机森林的特征重要性来对变量进行重要性度量并将它们从有最大贡献性到有最小贡献性进行了排序, 筛选其中具有贡献性最大的特征, 并剔除其它重要性较低的变量。它的基本思想为, 在单棵决策树 t 下, 对输入每个袋外样本(out-of-bag, OOB)记录其 OOB_t 对应的误差 $errOOB_t$, 接着, 随机排列 OOB_t 中 X^j 的值, 可得到一组新的扰动样本 \widetilde{OOB}_t^j , 并计算其误差 $err\widetilde{OOB}_t^j$, X^j 的特征重要性就是在求取单个决策树上 $err\widetilde{OOB}_t^j$ 和 $errOOB_t$ 两误差的减少量的基础上, 再求取所有决策树减少量的平均值^[57, 74, 75]。其公式如下:

$$VI(X^j) = \frac{1}{ntree} \sum_{t=1}^{ntree} (err\widetilde{OOB}_t^j - errOOB_t) \quad (19)$$

式中:

t —单颗决策树;

$ntree$ —RF 中树的数量。

可以从公式中看出, 当 $err\widetilde{OOB}_t^j$ 和 $errOOB_t$ 没有差异时, 输入的变量没有预测值, 说明 VI 越大, 其相应变量也就越重要。

5.1.2 数据集划分

根据上述处理好的各个参数影像, 提取出各个采样点的以上的各类特征, 建立

采样点实测土壤盐分与各个特征的土壤数据集。将此数据集随机分为训练集和验证集两部分，用训练集来对模型进行训练，用验证集来评估模型的好坏及最优模型的选择。

为了充分利用本文少量的采样数据，本文使用的是 K 折交叉验证 (K-fold Cross Validation, K-CV) 的方法来对数据集进行划分，这是比较常用的交叉验证方法。其基本思想是将数据集随机地划分成 K 大小相等的样本子集，然后依次选择其中的一份作为验证集，将其余剩下的 K-1 份子集数据作为训练集对模型进行训练。最终结果以这 K 次模型表现的平均值作为评估指标。K 的取值大于等于 2，Taghizadeh-Mehrjardi 等^[76]建议将 K 设为 5，其优势在于计算所获结果具备无偏估计和稳定可靠的特征。本文将 K 设为 5，其划分数据的流程如下图所示。

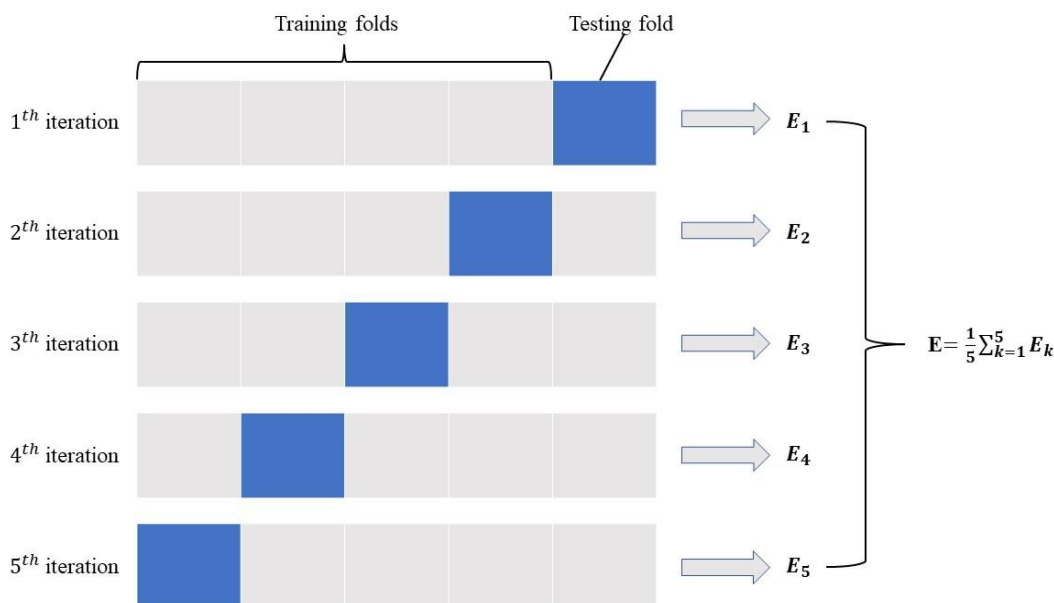


图 20 K 折交叉验证图

从上图可知，该交叉验证方法是多次利用已有的数据，将数据随机地化分为了 5 份子集，依次拿其 1/5 用于模型验证。剩余的 4/5 样本数据用于模型训练。它是对模型进行了多次的训练及验证过程，最终计算这 5 此模型表现的平均值作为最终的模型结果。

5.1.3 筛选结果

本次实验选用了 19 个雷达参数，13 个盐度植被指数和 18 个地形因子共 50 个参数作为自变量输入，然而，并不是每个参数对土壤盐分的反演精度起作用，甚至

一些参数会存在着噪声，有信息冗余和过拟合的可能，这类参数反而会降低回归模型预测的精度，这类参数就是多余的，需要被剔除^[28]。本次研究针对这一问题，采用了随机森林的特征重要性方法来筛选一些重要性的参数，剔除一些不太重要的参数，并且该算法不仅能计算出自变量对因变量的重要性，也能计算出所有变量之间的相对作用^[77]。该操作的运行平台是在 MATLAB 环境下，将所有的参选变量输入随机森林模型中，调用了 MATLAB 自带的 `TreeBagger` 函数计算出每个参数的相对重要性，并按降序排序。本文将按阈值来筛选变量，设置阈值为 0.05，即 $\text{threshold} \geq 0.05$ ，将阈值大于等于 0.05 的特征参数筛选出来，剔除了一些重要性不明显的特征^[75]。因此本文选择了前 10 个参数，通过计算前 10 个特征的累积重要性值与所有特征的累积重要值之比超过了 0.9，最后将这筛选的 10 个特征整理成如下表 9 所示。

表 9 特征变量重要性排序

重要性排序 Importance ranking	参数 Parameters	重要性得分 Importance score	对应波段 Band
1	CRSI	0.572	3
2	SI6	0.524	5
3	SC	0.238	8
4	SI5	0.226	4
5	AH	0.176	6
6	VV	0.165	1
7	VD	0.152	10
8	DEM	0.072	7
9	Entropy	0.069	2
10	TWI	0.053	9

当变量重要性（VI）越高，表明该特征与土壤 EC 关系越密切。因此，可以根据 VI 值来挑选出与 EC 值关系密切的一些特征。

将上述筛选的 10 个特征参数，在 ARCGIS 环境中显示了这些特征的灰度值的视觉图像，像元的对比值可揭示土壤盐分特性的响应，如图 21 所示，在该研究区的这 10 个特征中，极化分解参数 Entropy，后向散射系数 VV，光学指数 CRSI，SI5 和 SI6 有更多的空间和纹理信息，视觉信息更丰富，而剩余的地形因子视觉信息不太丰富。

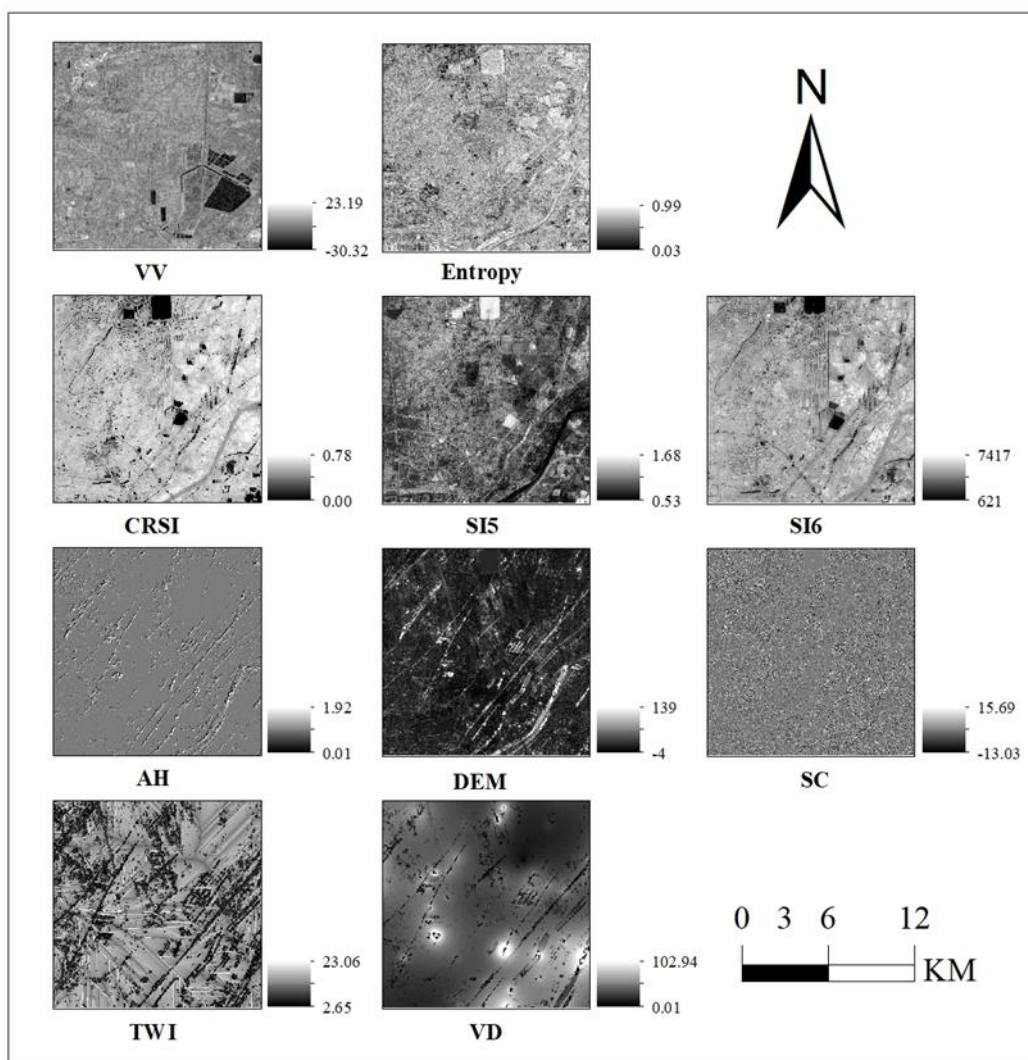


图 21 特征可视化图

5.2 模型配置与评估

5.2.1 模型配置

在本文使用的 4 种土壤含盐量预测的模型中，支持向量机和随机森林这两种模型需要对其的一些关键参数进行设置，以期能获得对于本模型表现的最佳参数。首先对于其他模型，支持向量回归模型（SVR）需要确定最佳的参数 C 和 γ 的值。 C 值过大对训练集数据的模型拟合程度越高，但预测阶段的精度很低，对误差有较低的容忍程度； C 值太小，则训练的模型精度较低，不能令人满意。 γ 值过大模型会过拟合，过小则模型的拟合效果不足。因此， C 和 γ 对 SVM 有很大的影响，为了提高模型精度就要考虑如何选择最合适的 C 和 γ 值。实

验采用的是基于粒子群优化的方法确定了其最优值。粒子群优化算法最初是模拟鸟群迁移和群聚行为而提出的基于群体智能的全局随机搜索算法。对于解决优化问题时，假如问题的解对应每一个粒子在搜索空间中的位置，每一个粒子由速度 V_i ，位置 X_i 和适应度三种属性决定，用适应度来衡量粒子的优劣。假定群体规模为 n 的粒子在空间中运动，其有粒子自身当前最优位置 P_i 和所有粒子当前最优位置 P_g 两个属性，每个粒子通过下面公式更新自己的速度和位置。

$$\begin{cases} v_i(k+1) = wv_i(k) + c_1r_1[P_i(k) - x_i(k)] + c_2r_2[P_g(k) - x_i(k)] \\ x_i(k+1) = x_i(k) + V_i(k+1) \end{cases} \quad (20)$$

式中：

- w —惯性因子；
- k —当前迭代次数；
- c_1, c_2 —学习因子；
- r_1, r_2 —分布在 $[0,1]$ 间的随机数。

本实验将参数 c_1 的初始值设为 2， c_2 的初始值为 0.6，最大进化数量即终止代数为 200，种群最大数量为 20。在 MATLAB 环境中其运行结果如图 22 所示：

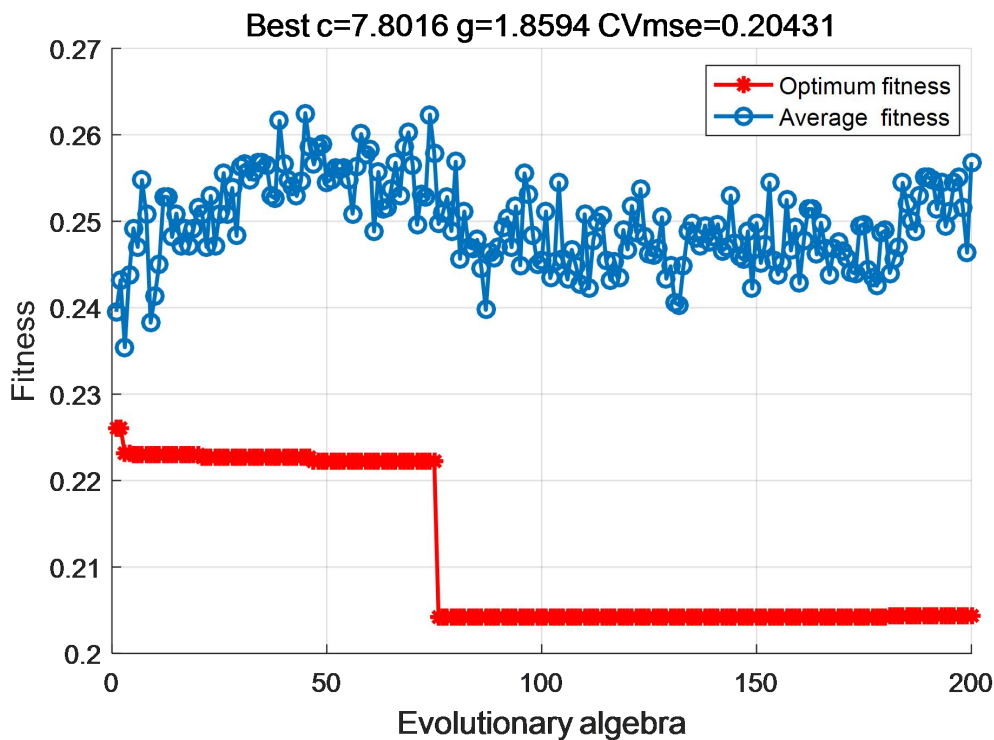


图 22 最佳参数 c 和 g 的确定

采用粒子群进行参数优化时，将结果的均方误差 MSE 作为粒子群算法的适应度函数。从上图可以得出，对于 SVR 模型，其 CVMSE 值趋于稳定时，它的值为 0.204，此时最佳的参数 c 为 7.802，最佳的参数 g 为 1.859，此时的 SVR 模型是最优的。

对于随机森林模型的最优决策树数量的选择，实验在 MATLAB 平台上实验，将 leaf 的初始值设为 5，初始值 Ntrees 设置为 500。将野外实测的土壤盐分数数据作为测量值 Y，将筛选的特征参数作为自变量 X 作为输入，通过使用 TreeBagger 函数来确定模型的最优参数 Ntrees，其结果如图 23 所示。

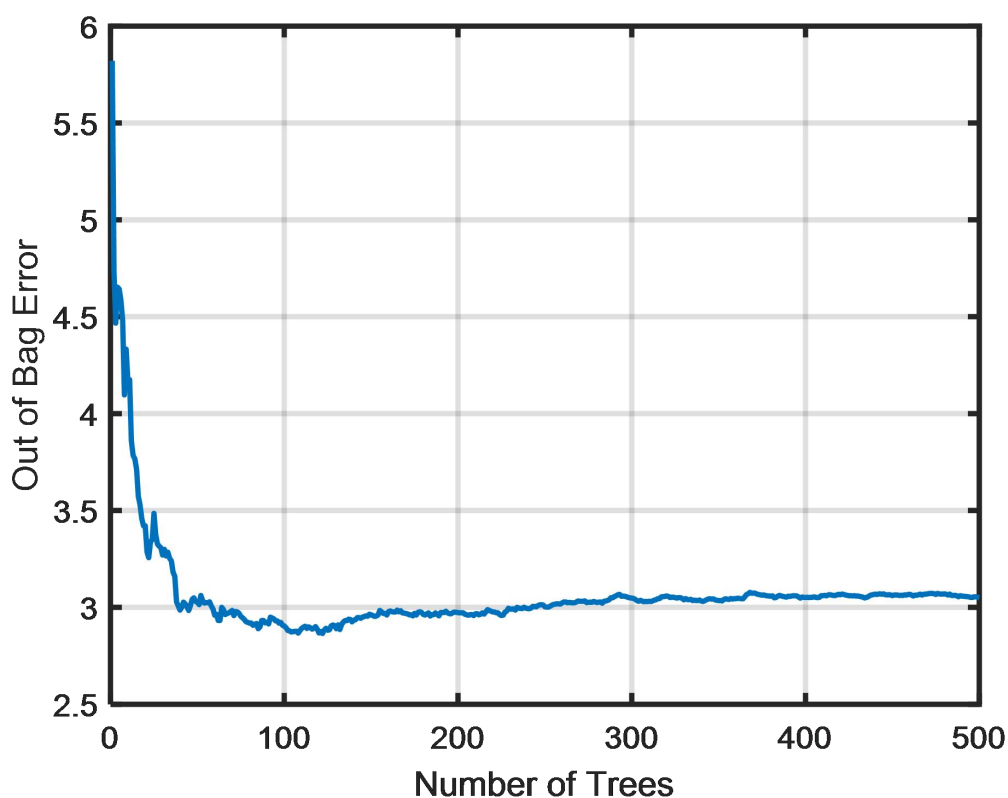


图 23 随机森林模型的决策树数量

从图 23 可以看出，当决策树的数量超过一定值时，其袋外误差基本收敛，再增加决策树的数量对模型的效果也基本没有提升，只会增加代码运行的负担。因此选择一个较大的决策树棵树即可，实验最后选定 RF 模型的参数 Ntrees 为 200。

5.2.2 模型评估

本文研究使用均方根误差(RMSE)和决定系数(R^2)来量化由训练集建立的模型精度和由验证集来表现该模型最终的预测精度。其计算公式如下：

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y}_i)^2} \quad (21)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - f_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (22)$$

式中：

y_i —真实数据；

f_i —预测数据；

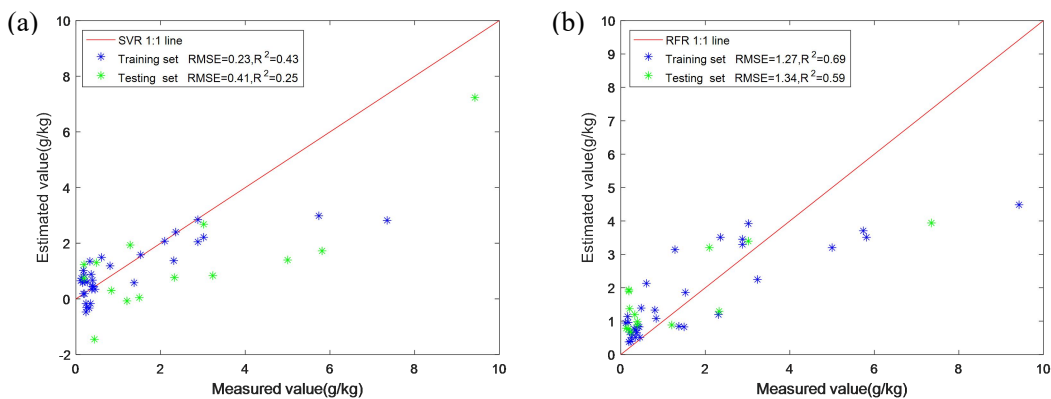
\bar{y} —测量数据的平均值。

当 R^2 越趋近 1，RMSE 值越趋向 0 时，则判定模型趋于最佳。

5.3 基于遥感影像特征的土壤含盐量反演

5.3.1 回归模型结果

从上述实验特征筛选的结果中，得到了 2 个雷达特征（极化分解参数 Entropy 和后向散射系数 VV）和 3 个盐度植被指数（冠层盐度响应植被指数 CRSI、盐度指数 SI5 和盐度指数 SI6）。当仅使用从遥感影像数据（包括光学影像数据和雷达影像数据）提取出的以上 5 个特征作为输入，以训练集中土壤的含盐量测量值作为输出，将其分别导入本文的四种模型中（SMR、SVR、RFR 和 PLSR），来预测该区域的土壤含盐量信息时，并用测试集数据来分别验证这四种模型的预测表现。这四种模型的训练集和测试集的性能表现如图 24 所示。



(a) —SVR； (b) —RFR； (c) —PLSR； (d) —SMR.

图 24 基于遥感影像特征的四中模型的表现

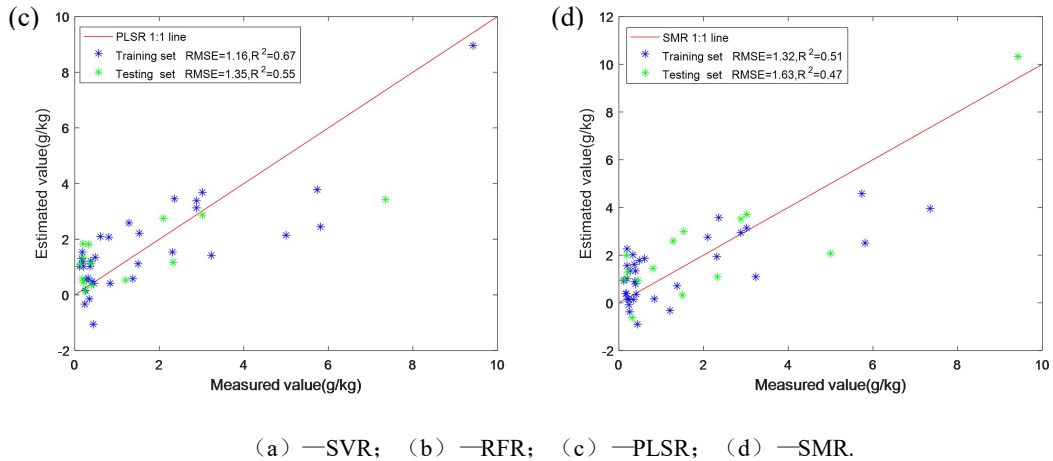


图 24 基于遥感影像特征的四种模型的表现 (续)

5.3.2 模型比较

为了较好地评估与比较这四种模型的性能表现，将图 24 的模型结果整理成表 10 所示。最低的 RMSE 和最高的 R^2 表示最高模型拟合特性。从表 10 中可清晰地看出，在训练集中 RFR 算法的性能表现最好 ($R^2=0.69$, $RMSE=1.27ds/m$)，紧接着的是 PLSR 算法 ($R^2=0.67$, $RMSE=1.16ds/m$) 和 SMR 算法 ($R^2=0.51$, $RMSE=1.32ds/m$)，性能表现最差的为 SVR 算法 ($R^2=0.43$, $RMSE=0.23ds/m$)。在测试集中，RFR 算法的性能表现也是最好的 ($R^2=0.59$, $RMSE=1.34ds/m$)，紧接着是 PLSR 算法 ($R^2=0.55$, $RMSE=1.36ds/m$) 和 SMR 算法 ($R^2=0.47$, $RMSE=1.63ds/m$)，性能表现最差的为 SVR 算法 ($R^2=0.25$, $RMSE=0.41ds/m$)。

表 10 基于遥感影像特征的 4 种模型性能的比较

模型 Model	训练集 Training set		测试集 Testing set	
	R^2	RMSE/($ds.m^{-1}$)	R^2	RMSE/($ds.m^{-1}$)
SMR	0.51	1.32	0.47	1.63
SVR	0.43	0.23	0.25	0.41
RFR	0.69	1.27	0.59	1.34
PLSR	0.67	1.16	0.55	1.35

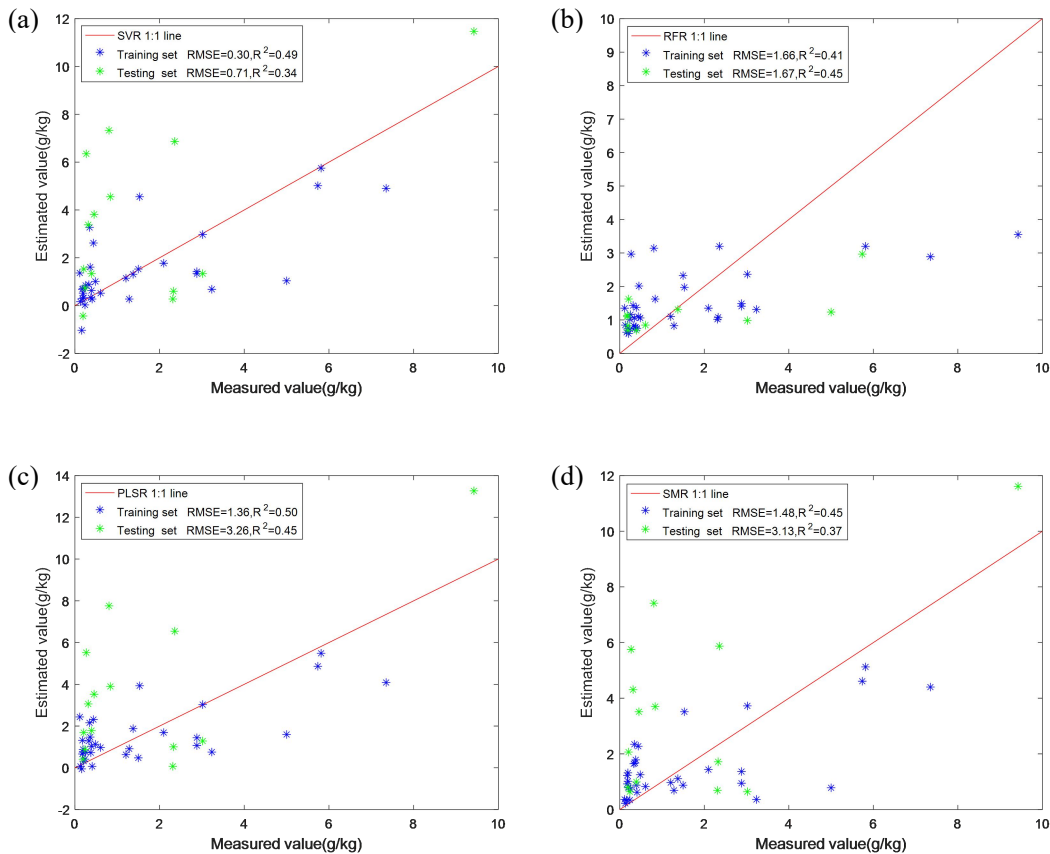
且从表 10 的结果中可明显地看出，以上四种模型它们的训练集的 R^2 都略微大于它们的测试集的 R^2 ，其原因可能为本实验所使用的样本量较少导致的；此外，SVR 模型训练集和测试集的 R^2 都比其他三种模型小，但它们的 RMSE 却更低，表现却比其他三种模型好，它可能是由模型本身的参数设置和样本值多集中在低数值

上所导致的结果（即 EC 多在 0~1 之间）。因此，综合考虑训练集和测试集的性能表现来看，RFR 算法是最适合利用遥感影像特征来反演该区域土壤盐分信息的模型。

5.4 基于地形因子的土壤含盐量反演

5.4.1 回归模型结果

在上述实验特征筛选的结果中，得到了 5 个筛选出的地形因子（高程 DEM、山体阴影 AH、平面曲率 SC、谷深 VD 和地形湿度指数 TWI）。当仅使用从 DEM 数据提取出的以上 5 个地形因子作为输入，以训练集中土壤的含盐量测量值作为输出，将其分别导入本文的四种模型中（SMR、SVR、RFR 和 PLSR），来预测该区域的土壤含盐量信息时，并用测试集数据来分别验证这四种模型的预测表现。这四种模型的训练集和验证集的性能表现如图 25 所示。



(a) —SVR; (b) —RFR; (c) —PLSR; (d) —SMR.

图 25 基于地形因子的四种模型的表现

5.4.2 模型比较

为了较好地评估与比较这四种模型的性能表现，将图 25 的模型结果整理成表 11 所示。最低的 RMSE 和最高的 R^2 表示最高模型拟合特性。从表 11 中可清晰地看出，在训练集中 PLSR 算法的性能表现最好 ($R^2=0.50$, $RMSE=1.36ds/m$)，紧接的是 SVR 算法 ($R^2=0.49$, $RMSE=0.30ds/m$) 和 SMR 算法 ($R^2=0.45$, $RMSE=1.48ds/m$)，性能最差的为 RFR 算法 ($R^2=0.41$, $RMSE=1.66ds/m$)。在验证集中，RFR 算法的性能表现是最好的 ($R^2=0.45$, $RMSE=1.67ds/m$)，紧接的是 PLSR 算法 ($R^2=0.45$, $RMSE=3.26ds/m$) 和 SMR 算法 ($R^2=0.37$, $RMSE=3.13ds/m$)，性能表现最差的为 SVR 算法 ($R^2=0.34$, $RMSE=0.71ds/m$)。

表 11 基于地形因子的 4 种模型性能的比较

模型 Model	训练集 Training set		测试集 Testing set	
	R^2	RMSE/(ds.m ⁻¹)	R^2	RMSE/(ds.m ⁻¹)
SMR	0.45	1.48	0.37	3.13
SVR	0.49	0.30	0.34	0.71
RFR	0.41	1.66	0.45	1.67
PLSR	0.50	1.36	0.45	3.26

从上表中，也可明显地看出，相较于使用遥感影像特征来反演该区域的土壤含盐量信息，使用地形因子来反演土壤盐分时精度整体较低。此外，综合训练集和测试集的性能表现来看，PLSR 算法是最适应于利用地形因子来反演该区域土壤盐分信息的模型。

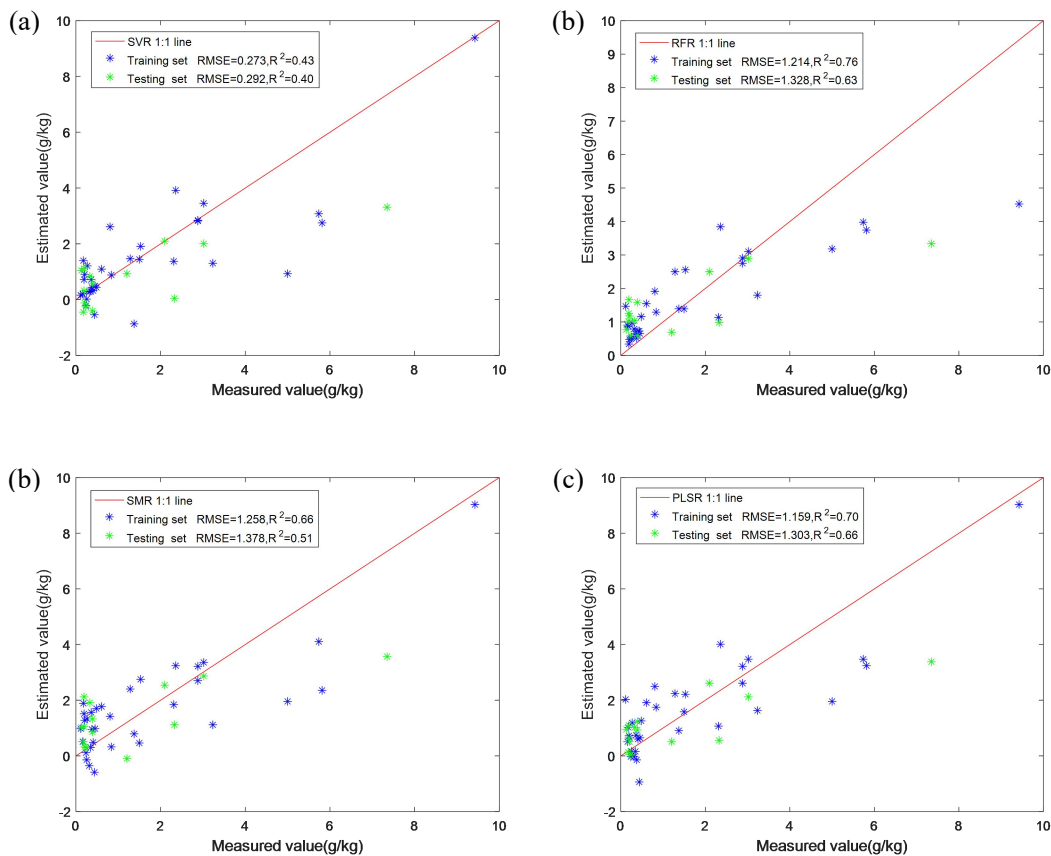
5.5 基于全参数的土壤含盐量反演

考虑到土壤的形成是一个复杂的过程，它受母质、气候、生物、地形和时间等因素的综合影响，而土壤的含盐量与土壤的形成息息相关。因此，综合考虑所有因子，在保留雷达影像和光学影像数据的优点的基础上，再加入比较常用的地形因子来反演该地区的土壤盐分信息。

5.5.1 回归模型结果

在特征筛选的结果中，得到了 2 个雷达特征（极化分解参数 Entropy 和后向散射系数 VV）、3 个盐度植被指数（冠层盐度响应植被指数 CRSI、盐度指数 SI5 和盐度指数 SI6）和 5 个地形因子（高程 DEM、山体阴影 AH、平面曲率 SC、谷深

VD 和地形湿度指数 TWI) 共计 10 个参数。当综合考虑及使用从遥感影像数据中提取的特征参数和从 DEM 数据中提取出的地形因子, 将以上筛选出的 10 个特征作为输入, 以训练集中土壤的含盐量测量值作为输出, 将其分别导入本文的四种模型中 (SMR、SVR、RFR 和 PLSR), 来预测该区域的土壤含盐量信息时, 并用测试集数据来分别验证这四种模型的预测表现。这四种模型的训练集和验证集的性能表现如下图所示, 图 26 显示了这 4 种模型的散点图和模型精度。



(a) —SVR; (b) —RFR; (c) —PLSR; (d) —SMR.

图 26 选用四种模型实测值与预测值的关系

从上图基于全参数结合四种模型对土壤盐分的预测值与测量值的分析, 可以看出其整体精度是优于前两种单独使用遥感影像特征或单独使用地形因子来反演该区域土壤盐分的精度。

5.5.2 模型比较

为了较好地评估与比较这四种模型的性能表现, 将上述训练的 4 种土壤盐渍度模型的性能结果整理成如表 12 所示。最低的 RMSE 和最高的 R^2 表示最高模型拟

合特性。在本文的4个模型的验证集中，PLSR算法的性能表现最好（ $R^2=0.66$ ， $RMSE=1.30ds/m$ ），其次为RFR模型（ $R^2=0.63$ ， $RMSE=1.33ds/m$ ），接着是SMR模型（ $R^2=0.51$ ， $RMSE=1.38ds/m$ ），性能最差的为SVR算法（ $R^2=0.40$ ， $RMSE=0.29ds/m$ ）。PLSR模型的测试集的RMSE为1.30ds/m略大于训练集的RMSE（1.16ds/m）。PLSR模型相对其他模型效果还不错，其测试集的 R^2 （0.66）也是本实验4种模型中最高的。

表12 基于全参数的4种模型的性能比较

模型 Models	训练集 Training set		测试集 Testing set	
	R^2	RMSE/(ds.m ⁻¹)	R^2	RMSE/(ds.m ⁻¹)
SMR	0.66	1.26	0.51	1.38
SVR	0.43	0.27	0.40	0.29
RFR	0.76	1.21	0.63	1.33
PLSR	0.70	1.16	0.66	1.30

5.5.3 最优模型

根据上表中的4种模型精度的比较，可以看出虽然随机森林模型在训练集中的模型精度最高（ $R^2=0.76$ ， $RMSE=1.21ds/m$ ），相比于偏最小二乘回归算法在训练集的模型精度（ $R^2=0.70$ ， $RMSE=1.16ds/m$ ）更优，但其在测试集上的模型表现（ $R^2=0.63$ ， $RMSE=1.33ds/m$ ）则低于偏最小二乘回归法的模型表现（ $R^2=0.66$ ， $RMSE=1.30ds/m$ ），且由于随机森林在训练集上的模型精度显著大于其在测试集上的预测精度，这可能是由于该模型有一定程度的过拟合。因此需要综合考虑训练集和测试集的 R^2 和RMSE，最后，选用测试集中对训练的模型预测性能表现最佳的偏最小二乘回归模型（PLSR）来进行后续的土壤盐渍化的分布研究。因此接下来这里主要介绍PLSR模型的一些基本参数的实验情况。在用偏最小二乘回归法对土壤盐分数据进行模型训练的过程中，也可以得出其基本的各土壤盐分特征参数对土壤盐分的相关性，图27展示了其标准化回归系数结果。偏最小二乘回归模型的回归系数结果表明，在土壤盐分与各特征参数的回归系数中，波段3（CRSI）和波段5（SI5）对土壤盐分估算贡献最高，其次是波段10（VD），波段7（DEM），波段8（SC）和波段1（VV），其余的波段对土壤的盐分的估算贡献较低。

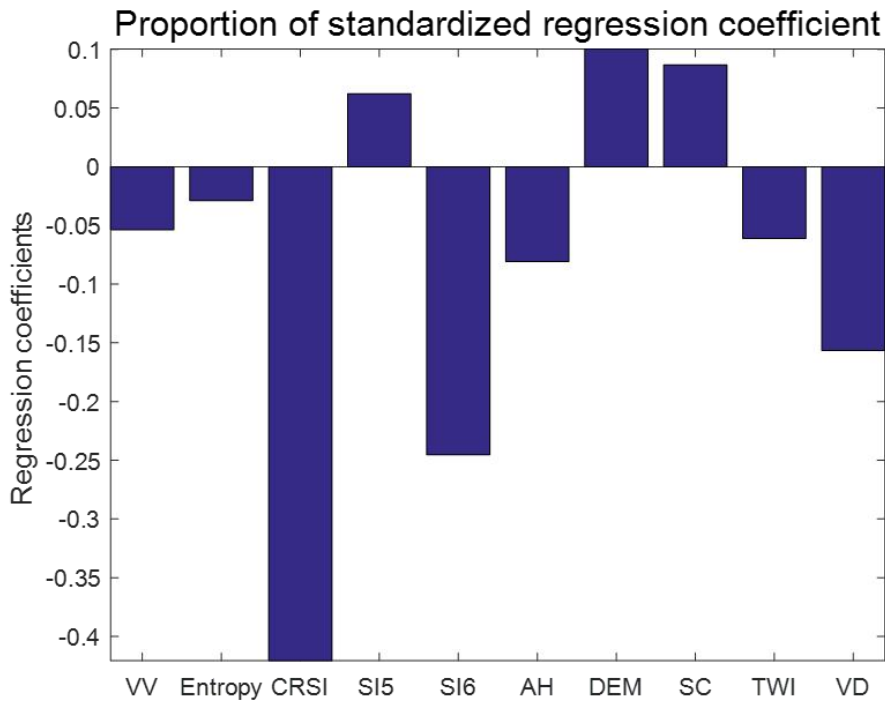


图 27 PLSR 模型中盐分与各参数的回归系数

最后，根据 PLSR 模型预测的结果，可求出该模型土壤盐分与各变量之间的回归方程，以便后续该地区土壤盐分空间分布图的制作。土壤盐分与各参数变量的对应关系如下表 13 所示：

表 13 土壤盐分预测模型

模型编号 Models No	EC 预测模型 Model in EC (ds/m)	R ²	RMSE/(ds.m ⁻¹)
标准数据方程 Standard data equation	Y=-0.054×VV-0.029×Entropy-0.421×CRSI+0.062×SI5-0.245×SI6-0.081×AH+0.10×DEM+0.087×SC-0.061×TWI-0.157×VD	0.66	1.30
原始数据方程 Original data equation	Y=13.733-0.054×VV-0.506×Entropy-12.669×CRSI+1.519×SI5-0.002×SI6-2.708×AH+0.027×DEM+0.372×SC-0.037×TWI-0.051×VD		

5.6 土壤盐分制图

从上面的分析中，可以得出全参数选择的预测表现大于遥感影像特征的预测表现和地形因子的预测表现，且在全参数选择的预测中，PLSR 模型是该研究区最佳的模型，因此，选用此模型来获取研究区土壤盐渍化的空间信息。以选取的 10 个特征影像为输入，利用训练好的 PLSR 模型，使用 ARCGIS 软件计算黄河三角洲各像元的土壤盐分，并生成土壤盐分的空间分布图(如图 28 所示)。

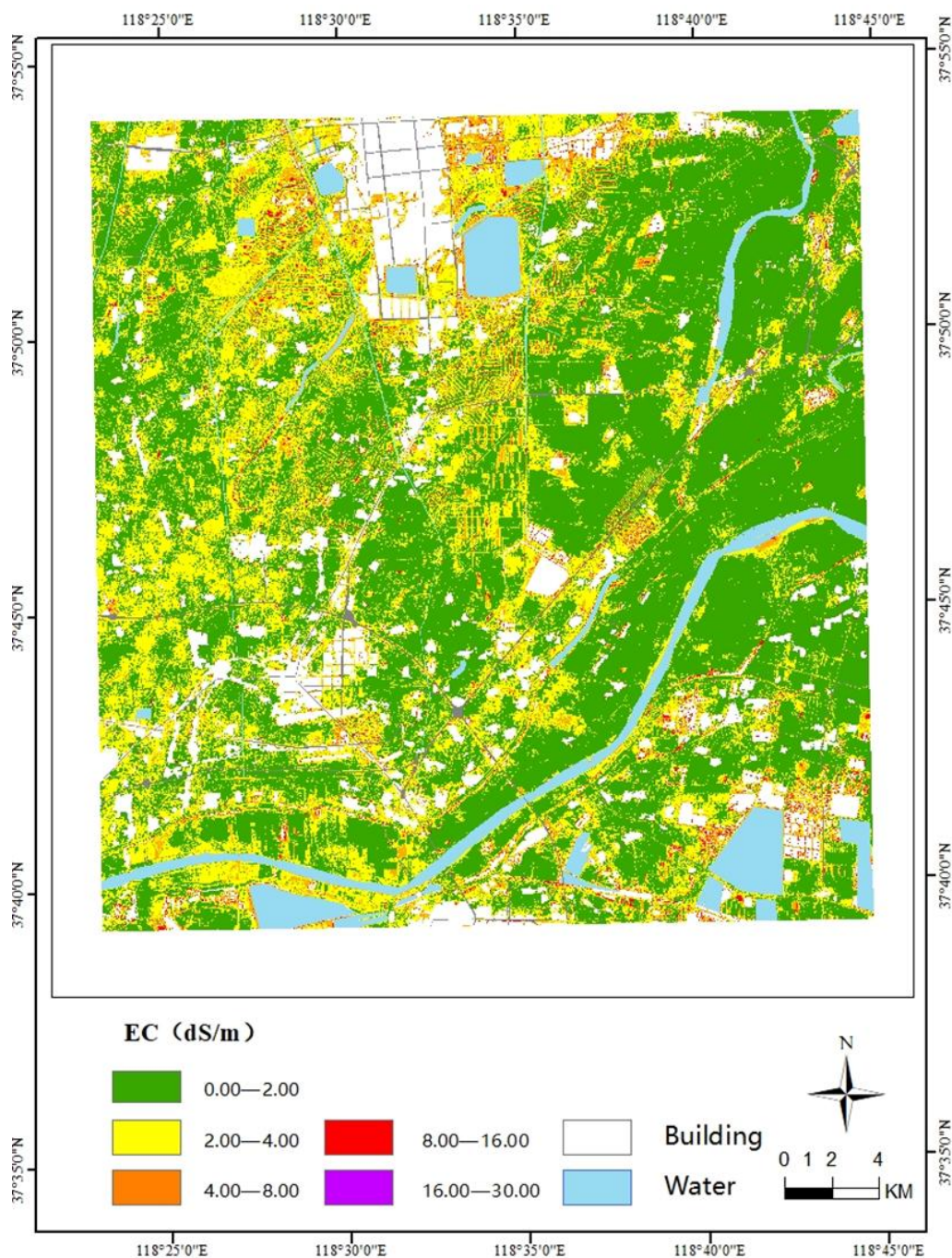


图 28 土壤含盐量的空间分布

统计上游研究区土壤盐渍化各类严重程度的占比，结果如表 14 所示，这里的比例是以各盐渍化程度像元的数量占整个研究区所有像元的百分比。从 2019 年黄河三角洲盐渍化程度分类图和各盐渍化程度统计表中可以看出，研究区主要为非盐渍化土壤，其面积过半，主要分布在研究区东部，占总面积的 64.2%，占比最大；轻度盐渍化区域主要分布在西部和北部，约占总面积的 29.2%；中度盐渍化区域呈零散分布于研究区，约占总面积的 6.2%；重度盐渍化区域仅少量分布在东南部和

北部，还有一些河流边上，仅占总面积的 0.4%，占比最小；研究区没有极重度盐渍化区域。各盐渍化程度区域占比与采样样本数据也基本吻合。结果也表明，从地理位置上上，土壤盐渍化程度由东向西逐渐递增，且沿水体向内陆边缘区域盐渍化程度最重，由于地形低，地下水浅，河口区一带盐碱度较高。此外，黄河及水库水面沿岸的土壤盐渍化程度普遍较严重，一般在 4 至 16 ds/m 范围内，咸水侵入引起土壤盐碱化。相反，由于地形较高，研究区东北部的土壤含盐量较低，其含量在约在 0~2 ds/m 内。研究区域测试数据集的总体分类精度为 85.71%，kappa 系数为 0.58。这是一个可以接受的结果，在现场测量时，温度和降水即使采取一些步骤（如重复测量、仔细选择采样等）对测量结果进行校准，也会产生误差。

表 14 研究区各盐渍化程度占比统计表

分类统计 Classified statistics	非盐渍化 Non-saline	轻度盐渍化 Slightly	中度盐渍化 Moderately	高盐渍化 Highly	极重盐渍化 Extremely
百分比 Proportion/%	64.2	29.2	6.2	0.4	0

5.7 本章小结

本章分别使用了 4 种机器学习算法对从不同的数据源提取的特征变量进行建模分析与比较，并在对研究区土壤含盐量的建模与预测的过程中，为了提高模型预测的精度，在基于使用前两种特征参数进行土壤含盐量的建模与预测中，综合考虑了所有因子进行预测，结果表明，综合考虑从遥感影像数据提取的遥感特征与从 DEM 高程数据提取的地形因子来预测研究区的土壤盐分的模型的整体表现优于分别使用遥感特征或地形因子预测下的土壤含盐量的精度，且其中使用遥感特征预测下的模型表现又要优于使用地形因子预测下的模型表现。在使用全参数的土壤含盐量的预测过程中，最佳的模型表现为 PLSR 算法，其次为 RFR 算法和 SMR 算法，SVR 算法在此过程中对土壤含盐量的预测效果最弱。最后，根据此最佳的特征参数组合和最佳的表现模型生成了该研究区的土壤盐分的空间分布图，并对其土壤盐分的分布规律进行了详细地概况与分析。

结 论

本文以山东省东营市境内的黄河入海口区域为研究区，基于 Sentinel-1/2 影像数据以及东营市的 30m 分辨率的 ASTER GDEM 高程数据，结合野外与卫星同步获得的实测数据，采用 4 种机器学习的方法分析了影响研究区土壤盐分的驱动因子及其组合，并对比不同模型的表现选取最优的方法提取研究区的土壤盐渍化信息。主要得到以下结论：

1) 基于随机森林模型的特征重要性方法可以有效筛选研究区土壤盐分的敏感特征。通过使用特征重要性的方法对比所有参选特征的重要性得分，结果表明从光学影像提取的 CRSI 指数在反演该地区土壤盐分的所有特征中贡献度最大，其重要性得分为 0.572，显著高于地形因子 SC (0.238) 和雷达特征 VV (0.165)。地形因子和雷达参数对该地区的贡献性较小，这说明了有光学数据提取出来的参数是最适应本区域土壤盐分反演建模的特征。

2) 多源特征组合能够显著提升土壤盐分模型反演的精度。本文基于常用于土壤盐分反演的 Sentinel-1/2 遥感数据，结合 DEM 提取的地形因子，通过对比不同数据源特征及其组合的情况下土壤含盐量反演精度的相对高低，结果发现，综合考虑遥感影像特征和地形因子来反演该地区土壤盐分的整体预测效果最好 (PLSR 的 $R^2=0.66$ ；RF 的 $R^2=0.63$ ；SR 的 $R^2=0.51$ ；SVR 的 $R^2=0.40$)，其次为仅用遥感影像特征 (PLSR 的 $R^2=0.55$ ；RF 的 $R^2=0.59$ ；SR 的 $R^2=0.47$ ；SVR 的 $R^2=0.25$)，预测精度最差的是仅用地形因子作为土壤盐分反演的参数 (PLSR 的 $R^2=0.45$ ；RF 的 $R^2=0.45$ ；SR 的 $R^2=0.37$ ；SVR 的 $R^2=0.34$)。最佳的预测效果是综合考虑遥感影像特征和地形因子，表明基于多源特征监测土壤盐分具有很好的预测效果。

3) 文中 4 种土壤盐分回归模型 (SMR, SVR, RFR 和 PLSR) 均取得了较高的精度，其中 PLSR 模型预测精度最好。在基于最优数据源特征组合的基础上，通过使用文中 4 种机器学习算法构建了研究区土壤盐分反演模型。研究结果显示，PLSR 模型的验证集精度最高 ($R^2=0.66$, $RMSE=1.30ds/m$)，是最适应该地区土壤盐分反演的模型。与 PLSR 相比，其他 3 种模型的预测精度较低，分别为 RFR ($R^2=0.63$, $RMSE=1.33ds/m$)，SMR 模型 ($R^2=0.51$, $RMSE=1.38ds/m$) 和最低精度的 SVR 模型 ($R^2=0.40$, $RMSE=0.29ds/m$)。从研究结果中也可看出，RFR 模型的训练集的 R^2 达到了 0.76，但其测试集的预测效果较低 ($R^2=0.63$)，这说明这个模型存在一定程度过拟合的可能，这是由于本次研究的样本数据相对较少而产生的。

4) 通过对研究区土壤盐渍化区域分布的预测和研究, 结果显示, 研究区主要为非盐渍化土壤, 其面积过半, 主要分布在研究区东部, 约占总面积的 64.2%, 占比最大; 轻度盐渍化区域主要分布在西部和北部, 约占总面积的 29.2%; 中度盐渍化区域呈零散分布于研究区, 约占总面积的 6.2%; 重度盐渍化区域仅少量分布在东南部和北部, 还有一些河流边上, 仅占总面积的 0.4%, 占比最小; 研究区没有极重度盐渍化区域, 各盐渍化程度区域占比与采样样本数据和参考数据情况也基本吻合。从中也可以看出盐渍化程度较重的区域大多在居民区周围和水体周围, 它们是由于地形较低, 地下水较浅导致的, 研究区东北部区域由于地形较高的原因其土壤含盐量呈现较低的程度。

虽然本文尝试在基于光学数据和雷达数据优势的基础上, 耦合地形因子进行土壤盐分的反演, 模型的精度有了显著的提高, 但从整体表现结果来看, 模型拟合的精度也受到一定的限制, 其可能是样本的数据量较少和大量采集的样本的 EC 值多集中在较小的区域 (EC 值多在 0~1ds/m 间), 即非盐渍土, 中度盐渍土及之上的盐渍土样本量太少。因此, 在今后的研究中, 应当尽可能地增加采集的样本量以及加强样本的代表性与充分性, 即应在研究区内均匀地采集各程度类型的盐渍土样本点, 以便后续更好的建模与评估模型的表现。

参考文献

- [1] 吴家林. 阿拉尔垦区棉田土壤盐渍化的遥感监测与植棉效益分析[D]. 阿拉尔:塔里木大学, 2021.
- [2] 范王涛. 土壤盐碱化危害及改良方法研究[J]. 农业与技术, 2020, 40(23):114-116.
- [3] 赵学伟. 基于 GPR 的黄河三角洲滨海盐渍土水盐含量探测模型研究[D]. 泰安:山东农业大学, 2019.
- [4] 孙铭岳. 基于多源遥感的黄河三角洲耕地土壤盐分和有机质含量反演[D]. 泰安:山东农业大学, 2022.
- [5] 李星佑, 张飞, 王箬. 土壤盐渍化遥感监测模型构建方法现状与发展趋势[J]. 自然资源遥感, 2022, 34(4):11-21.
- [6] 黄晶, 孔亚丽, 徐青山, 等. 盐渍土壤特征及改良措施研究进展[J]. 土壤, 2022, 54(1): 18-23.
- [7] 宫燕. 盐渍土壤微波介电模型研究[D]. 北京:中国科学院大学(中国科学院东北地理与农业生态研究所), 2020.
- [8] 王君. 基于高分遥感的黄河三角洲土壤盐渍化变化特征分析[D]. 济南:济南大学, 2020.
- [9] 李光超. 黄河三角洲土壤盐渍化研究综述[J]. 安徽农学通报, 2020, 26(Z1):113-115.
- [10] GUO B, ZANG W, ZHANG R. Soil Salinization Information in the Yellow River Delta Based on Feature Surface Models Using Landsat 8 OLI Data[J]. IEEE Access, 2020, 8(2):394-403.
- [11] KHADIM F K, SU H, XU L, et al. Soil Salinity Mapping in Everglades National Park Using Remote Sensing Techniques and Vegetation Salt Tolerance[J]. Physics and Chemistry of the Earth, Parts A/B/C, 2019, 110(2):31-50.
- [12] GAO Y, LIU X, HOU W, et al. Characteristics of Saline Soil in Extremely Arid Regions: A Case Study Using GF-3 and ALOS-2 Quad-Pol SAR Data in Qinghai, China[J]. Remote Sensing, 2021, 13(3):417-435.
- [13] 杨练兵, 郑宏伟, 罗格平, 等. 基于遗传算法优化 BP 神经网络的土壤盐渍化反演[J]. 地理与地理信息科学, 2021, 37(2):12-21+37.
- [14] 王飞, 杨胜天, 丁建丽, 等. 环境敏感变量优选及机器学习算法预测绿洲土壤盐分[J]. 农业工程学报, 2018, 34(22):102-110.
- [15] 郝启勇. 山东黄河北矿区土壤盐渍化特征分析[J]. 中国地质灾害与防治学报, 2021, 32(1):65-69.

- [16] 冯娟, 丁建丽, 魏雯瑜. 基于雷达数据的区域土壤盐渍化监测[J]. 国土资源遥感, 2019, 31(1):195-203.
- [17] 刘全明, 成秋明, 王学, 等. 河套灌区土壤盐渍化微波雷达反演[J]. 农业工程学报, 2016, 32(16):109-114.
- [18] HOA P, GIANG N, BINH N, et al. Soil Salinity Mapping Using SAR Sentinel-1 Data and Advanced Machine Learning Algorithms: A Case Study at Ben Tre Province of the Mekong River Delta (Vietnam)[J]. Remote Sensing, 2019,11(2):128-149.
- [19] ZHANG Q, LI L, SUN R, et al. Retrieval of the Soil Salinity From Sentinel-1 Dual-Polarized SAR Data Based on Deep Neural Network Regression[J]. IEEE Geo-science and Remote Sensing Letters, 2022, 19(2):1-5.
- [20] EL HARTI A, LHISSOU R, CHOKMANI K, et al. Spatiotemporal Monitoring of Soil Salinization in Irrigated Tadla Plain (Morocco) Using Satellite Spectral Indices[J]. International Journal of Applied Earth Observations Geoinformation, 2016, 50(2):64-73.
- [21] SCUDIERO E, SKAGGS T H, CORWIN D L. Regional-scale Soil Salinity Assessment Using Landsat ETM Plus Canopy Reflectance[J]. Remote Sensing of Environment, 2015, 169(2):335-343.
- [22] BANNARI A, EL-BATTAY A, BANNARI R, et al. Sentinel-MSI VNIR and SWIR Bands Sensitivity Analysis for Soil Salinity Discrimination in an Arid Landscape[J]. Remote Sensing, 2018, 10(6):855-875.
- [23] 李晓燕, 张树文. 吉林省大安市近 50 年土地盐碱化时空动态及成因分析[J]. 资源科学, 2005, (3):92-97.
- [24] 王爽, 丁建丽, 王璐, 等. 基于地表光谱建模的区域土壤盐渍化遥感监测研究[J]. 干旱区地理, 2016, 39(1):190-198.
- [25] LEE J S, GRUNES M R, POTTIER E. Quantitative Comparison of Classification Capability: Fully Polarimetric Versus Dual and Single-polarization SAR[J]. IEEE Transactions Geoscience and Remote Sensing, 2001, 39(11):2343-2351.
- [26] NURMEMET I, GHULAM A, TIYIP T, et al. Monitoring Soil Salinization in Keriya River Basin, Northwestern China Using Passive Reflective and Active Microwave Remote Sensing Data[J]. Remote Sensing, 2015, 7(7):8803-8829.
- [27] WU W, MUHAIMEED A S, AL-SHAFIE W M, et al. Using L-band Radar Data For Soil Salinity Mapping—A Case Study in Central Iraq[J]. Environmental Research Communications, 2019,1(8):4-18.

- [28] TAGHADOSI M M, HASANLOU M, EFTEKHARI K. Soil Salinity Mapping Using Dual-polarized SAR Sentinel-1 Imagery[J]. *International Journal of Remote Sensing*, 2019, 40(1):237-252.
- [29] LASNE Y, PAILLOU P, FREEMAN A, et al. Effect of Salinity on the Dielectric Properties of Geological Materials: Implication For Soil Moisture Detection By Means of Radar Remote Sensing[J]. *IEEE Transactions Geoscience and Remote Sensing*, 2008, 46(6):1674-1688.
- [30] DALIAKOPOULOS I N, TSANIS I K, KOUTROULIS A, et al. The Threat of Soil Salinity: A European Scale Review[J]. *Science of the Total Environment*, 2016, 573(2):727-739.
- [31] 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.
- [32] 何宝忠, 丁建丽, 王飞, 等. 基于物候特征的盐渍化信息数据挖掘研究[J]. *生态学报*, 2017, 37(9):3133-3148.
- [33] 蒙莉娜, 丁建丽, 王敬哲, 等. 基于环境变量的渭干河-库车河绿洲土壤盐分空间分布[J]. *农业工程学报*, 2020, 36(1):175-181.
- [34] WANG N, XUE J, Peng J, et al. Integrating Remote Sensing and Landscape Characteristics to Estimate Soil Salinity Using Machine Learning Methods: A Case Study from Southern Xinjiang, China[J]. *Remote Sensing*, 2020, 12(24):21-38.
- [35] SHAHRAYINI E, NOROOZI A A. Modeling and Mapping of Soil Salinity and Alkalinity Using Remote Sensing Data and Topographic Factors: a Case Study in Iran[J]. *Environmental Modeling and Assessment*, 2022, 27(5):901-913.
- [36] TAGHIZADEH-MEHRJARDI R, MINASNY B, SARMADIAN F, et al. Digital Mapping of Soil Salinity in Ardakan Region, Central Iran[J]. *Geoderma*, 2014, 213(2):15-28.
- [37] 贺文君, 韩广轩, 颜坤, 等. 微地形对滨海盐碱地土壤水盐分布和植物生物量的影响[J]. *生态学杂志*, 2021, 40(11):3585-3597.
- [38] SIDIKE A, ZHAO S H, WEN Y M. Estimating soil salinity in Pingluo County of China using QuickBird data and soil reflectance spectra[J]. *International Journal of Applied Earth Observation and Geoinformation*, 2014, 26(6):156-175.
- [39] FARIFTEH J, VANDERMEER F D, ATZBERGER C, et al. Quantitative Analysis of Salt-affected Soil Reflectance Spectra: A Comparison of Two Adaptive Methods (PLSR and ANN)[J]. *Remote Sensing of Environment*, 2007, 110(1):59-78.
- [40] ALLBED A, KUMAR L, ALDAKHEEL Y Y. Assessing Soil Salinity Using Soil Salinity and Vegetation Indices Derived from IKONOS High-spatial Resolution Imageries: Applications in

- A Date Palm Dominated Region[J]. *Geoderma*, 2014, 230(5):1-8.
- [41] 朱高飞. 农田盐渍化土壤光谱特征及其遥感反演与分类研究[D]. 乌鲁木齐:新疆农业大学, 2013.
- [42] 陈俊英, 姚志华, 张智韬, 等. 大田葵花土壤含盐量无人机遥感反演研究[J]. *农业机械学报*, 2020, 51(07):178-191.
- [43] XU H T, CHEN C B, ZHENG H W, et al. AGA-SVR-based Selection of Feature Subsets and Optimization of Parameter in Regional Soil Salinization Monitoring[J]. *International Journal of Remote Sensing*, 2020, 41(12):4470-4495.
- [44] GENUER R, POGGI J M, TULEAU-MALOT C. Variable Selection Using Random Forests[J]. *Pattern Recognition Letters*, 2010, 31(14):2225-2236.
- [45] 高元科. 基于微波遥感技术探测森林地表土壤含水率[D]. 哈尔滨:东北林业大学, 2016.
- [46] 贾殿纪. 环青海湖流域土壤盐分含量极化 SAR 反演研究[D]. 北京:中国地质大学, 2019.
- [47] CLOUDE S R, POTTIER E. A Review of Target Decomposition Theorems in Radar Polarimetry[J]. *IEEE Transactions Geoscience and Remote Sensing*, 1996, 34(2):498-518.
- [48] 兰利花. 基于极化雷达黄河三角洲区域土壤水盐含量反演[D]. 北京:中国地质大学, 2021.
- [49] 牛朝阳, 马德宝, 张向峰. SAR 目标极化分解方法研究[J]. *微计算机信息*, 2005, (23):103-105.
- [50] CLOUDE S R, POTTIER E. An Entropy Based Classification Scheme For Land Applications of Polarimetric SAR[J]. *IEEE Transactions Geoscience and Remote Sensing*, 1997, 35(1):68-78.
- [51] SHAO Y, HU Q R, GUO H D, et al. Effect of Dielectric Properties of Moist Salinized Soils on Backscattering Coefficients Extracted from RADARSAT Image[J]. *IEEE Transactions Geoscience and Remote Sensing*, 2003, 41(8):1879-1888.
- [52] SCHUG F, OKUJENI A, HAUER J, et al. Mapping Patterns of Urban Development in Ouagadougou, Burkina Faso, Using Machine Learning Regression Modeling With Bi-seasonal Landsat Time Series[J]. *Remote Sensing of Environment*, 2018, 210(2):217-228.
- [53] FEILHAUER H, ASNER G P, MARTIN R E. Multi-method Ensemble Selection of Spectral Bands Related to Leaf Biochemistry[J]. *Remote Sensing of Environment*, 2015, 164(2):57-65.
- [54] HUANG W C, LIU H Y, ZHANG Y, et al. Railway Dangerous Goods Transportation System Risk Identification: Comparisons Among SVM, PSO-SVM, GA-SVM and GS-SVM[J]. *Applied Soft Computing*, 2021, 109(1):39-55.
- [55] ZHANG H R, WANG X D, WU J B, et al. A New SMO Algorithm for Support Vector

- Machines[C], Proceedings of the International Symposium on Intelligence Computation and Applications, 2005. Wuhan: China University of Geosciences Press, 2005.
- [56] LOOZEN Y, REBEL K T, DE JONG S M, et al. Mapping Canopy Nitrogen in European Forests Using Remote Sensing and Environmental Variables with the Random Forests Method[J]. *Remote Sensing of Environment*, 2020, 247(7):34-57.
- [57] HOA P V, GIANG N V, BINH N A, et al. Soil Salinity Mapping Using SAR Sentinel-1 Data and Advanced Machine Learning Algorithms: A Case Study at Ben Tre Province of the Mekong River Delta (Vietnam)[J]. *Remote Sensing*, 2019, 11(2):21-42.
- [58] FRIEDMAN J H, MEULMAN J J. Multiple Additive Regression Trees with Application in Epidemiology[J]. *Statistics in Medicine*, 2003, 22(9):1365-1381.
- [59] ZENG W Z, ZHANG D Y, FANG Y H, et al. Comparison of Partial Least Square Regression, Support Vector Machine, and Deep-learning Techniques for Estimating Soil Salinity from Hyperspectral Data[J]. *Journal of Applied Remote Sensing*, 2018, 12(2):4-6.
- [60] FAN X W, LIU Y B, TAO J M, et al. Soil Salinity Retrieval from Advanced Multi-Spectral Sensor with Partial Least Square Regression[J]. *Remote Sensing*, 2015, 7(1):488-511.
- [61] RODRIGUEZ-FEBEREIRO M, DAFONTE J, FANDINO M, et al. Evaluation of Spectroscopy and Methodological Pre-Treatments to Estimate Soil Nutrients in the Vineyard[J]. *Remote Sensing*, 2022, 14(6):16-26.
- [62] WANG K, QI Y B, GUO W J, et al. Retrieval and Mapping of Soil Organic Carbon Using Sentinel-2A Spectral Images from Bare Cropland in Autumn[J]. *Remote Sensing*, 2021, 13(6):72-91.
- [63] HUANG X, SHI Z H, ZHU H D, et al. Soil Moisture Dynamics Within Soil Profiles and Associated Environmental Controls[J]. *Catena*, 2016, 136(6):189-196.
- [64] WANG J, LI Z J, QIN X B, et al. Hyperspectral Predicting Model of Soil Salinity in Tianjin Costal Area Using Partial Least Square Regression; proceedings of the IEEE Joint International Geoscience and Remote Sensing Symposium (IGARSS) / 35th Canadian Symposium on Remote Sensing, Quebec City, Canada, July 13-18, 2014[C]. IEEE: New York, 2014.
- [65] SHEN J Q, SHUAI Y M, LI P X, et al. Extraction and Spatio-Temporal Analysis of Impervious Surfaces over Dongying Based on Landsat Data[J]. *Remote Sensing*, 2021, 13(18):3666-3688.
- [66] IVUSHKIN K, BARTHOLOMEUS H, BREGT A K, et al. Global Mapping of Soil Salinity

- Change[J]. *Remote Sensing of Environment*, 2019, 231(6):1260-1320.
- [67] OMUTO C, VARGAS R, ABDELMAGID E, et al. Mapping of Salt-affected Soils[M]. 2021.
- [68] AHMED Z, AMBINAKUDIGE S. Does Land Use Change, Waterlogging, and Salinity Impact on Sustainability of Agriculture and Food Security? Evidence from Southwestern Coastal Region of Bangladesh[J]. *Environment Monitoring and Assessment*, 2023, 195(1):74-102.
- [69] KHAN N M, RASTOSKUEV V V, SATO Y, et al. Assessment of Hydrosaline Land Degradation by Using A Simple Approach of Remote Sensing Indicators[J]. *Agricultural Water Management*, 2005, 77(1-3):96-109.
- [70] RAMOS T B, CASTANHEIRA N, OLIVEIRA A R, et al. Soil Salinity Assessment Using Vegetation Indices Derived from Sentinel-2 Multispectral Data. Application to Leziria Grande, Portugal[J]. *Agricultural Water Management*, 2020, 241(3):24-36.
- [71] ALKHASAWNEH M S, NGAH U K, TAY L T, et al. Determination of Importance for Comprehensive Topographic Factors on Landslide Hazard Mapping Using Artificial Neural Network[J]. *Environmental Earth Sciences*, 2014, 72(3):787-799.
- [72] TAGHIZADEH-MEHRJARDI R, SCHMIDT K, TOOMANIAN N, et al. Improving the Spatial Prediction of Soil Salinity in Arid Regions Using Wavelet Transformation and Support Vector Regression Models[J]. *Geoderma*, 2021, 383(21):16-61.
- [73] GALLANT J C, DOWLING T I. A Multiresolution Index of Valley Bottom Flatness for Mapping Depositional Areas[J]. *Water Resources Research*, 2003, 39(12):12-26.
- [74] 王芳, 马素霞, 王河. 基于随机森林变量选择的飞灰含碳量预测模型[J]. *热力发电*, 2018, 47(11):89-95.
- [75] GENUER R, POGGI J M, TULEAU-MALOT C. Variable Selection Using Random Forests[J]. *Pattern Recognition Letters*, 2010, 31(14):2225-2236.
- [76] TAGHIZADEH-MEHRJARDI R, NABIOLLAHI K, KERRY R. Digital Mapping of Soil Organic Carbon at Multiple Depths Using Different Data Mining Techniques in Baneh Region, Iran[J]. *Geoderma*, 2016, 266(3):98-110.
- [77] BIAU G, SCORNET E. A Random Forest Guided Tour[J]. *Test*, 2016, 25(2):197-227.