Postprint of Alpine Grassland Soil Salinization Inversion Based on Interpretable Machine Learning Models

Authors: Yang Mingxin, Ning Xiaochun, Liusheng Yang, Yafei Zhang, Shi Mingming, Yanbin Kang, Huang Qingdongzhi, Wang Shouxing, Zhou Huakun

Date: 2026-03-24T22:05:27+00:00

Abstract

Grassland soil salinization not only exacerbates the degradation of grassland vegetation and affects the performance of grassland ecological functions, but also restricts the ecological protection and restoration of degraded grasslands. In the alpine grassland soil salinization development areas of the Yellow River Source, an interpretable machine learning salinization inversion model based on feature selection was constructed through ground grid sampling combined with Sentinel-2 spectral indices. The influencing factors of the salinization model were revealed, and the spatial distribution characteristics of regional salinization were monitored.

The results show that: (1) The salinization inversion model constructed based on Stepwise Regression (STEP) feature selection combined with the Random Forest (RF) model achieved the optimal accuracy, with a coefficient of determination (R^2) of 0.64 and a root mean square error (RMSE) of 0.76 $g \cdot kg^{-1}$. (2) Analysis of the optimal model based on Shapley Additive Explanations (SHAP) indicated that among the nine feature variables, the Normalized Difference Water Index (NDWI) had the highest contribution to the model in this study and was positively correlated with soil salinization. (3) Monitoring showed that 47.13% of the study area was mildly salinized and 33.19% was moderately salinized, while severe salinization and saline soils were less developed in the study area and were mainly distributed along rivers and valley zones.

By exploring interpretable salinization monitoring models with different combinations of methods, this study provides a scientific basis and technical support for ecological protection and high-quality development in the Yellow River Basin.

Full Text

Preamble

Vol. 49, No. 3, March 2026

GEOGRAPHY

49 No. 3

Mar. 2026

Inversion of Soil Salinization in Alpine Grasslands Based on Interpretable Machine Learning Models

Yang Mingxin[1,2,3,4], **Ning Xiaochun**[2], **Yang Liusheng**[2], **Zhang Yafei**[2], **Shi Mingming**[2], **Kang Yanbin**[2], **Huang Qingdongzhi**[2], **Wang Shouxing**[2], **Zhou Huakun**[3]

([1]School of Geographical Sciences, Qinghai Normal University, Xining 810008, China; [2]Xining Center for Integrated Natural Resources Survey, China Geological Survey, Xining 810008, China; [3]Northwest Institute of Plateau Biology, Chinese Academy of Sciences, Xining 810008, China; [4]Academy of Plateau Science and Sustainability, Xining 810008, China)

Abstract

Soil salinization is a critical factor leading to the degradation of alpine grasslands. Accurate monitoring and inversion of soil salt content (SSC) are essential for the ecological protection and sustainable management of the Qinghai-Tibet Plateau. This study utilizes multi-source remote sensing data, including Sentinel-2 multispectral imagery and topographic factors, to construct an inversion model for soil salinization in alpine grasslands. By integrating various machine learning algorithms—such as Random Forest (RF), Extreme Gradient Boosting (XGBoost), and LightGBM—we developed a high-precision estimation framework. Furthermore, we employed SHapley Additive exPlanations (SHAP) to interpret the model, identifying the primary environmental drivers and spectral indices contributing to the salinization process. The results demonstrate that the ensemble learning models significantly outperform traditional linear regression models in terms of accuracy. This research provides a scientific basis for the dynamic monitoring of soil quality and the restoration of degraded grasslands in high-altitude regions.

1. Introduction

Soil salinization is a global environmental challenge that severely restricts agricultural productivity and threatens ecosystem stability. In the alpine grasslands of the Qinghai-Tibet Plateau, salinization is particularly sensitive to climate change and anthropogenic activities. The accumulation of soluble salts in the surface soil layer not only inhibits vegetation growth but also alters the physical and chemical properties of the soil, leading to land desertification and a reduction in biodiversity. Traditional field sampling and laboratory analysis methods, while

810021; 3. Qinghai Provincial Key Laboratory of Restoration Ecology in Cold Regions, Northwest Institute of Plateau Biology, Chinese Academy of Sciences,

Xining, Qinghai

Xining, the capital of Qinghai Province, serves as the political, economic, and cultural center of the region. Located on the eastern edge of the Qinghai-Tibet Plateau, it occupies a strategic position as a gateway to the plateau' s interior. The city is situated in the Huangshui River valley, characterized by its unique high-altitude continental climate and significant geographical importance within the "Belt and Road" initiative.

Geographical and Environmental Context

The topography of Xining is defined by its mountainous surroundings and river systems. As a typical plateau city, it maintains an average elevation of approximately 2,261 meters. This altitude contributes to its reputation as the "Summer Capital of China," offering a cool climate during the peak summer months when much of the country experiences extreme heat. The ecological environment of Xining is critical for the regional water security of Northwest China, as it sits within the drainage basin of the Yellow River' s major tributaries.

Socio-Economic Development

In recent years, Xining has undergone rapid urbanization and industrial transformation. The city' s economic structure has shifted from traditional manufacturing toward high-tech industries, green energy, and modern services. Key development areas include the production of lithium batteries, photovoltaic materials, and specialty plateau agriculture. Furthermore, Xining serves as a vital transportation hub, connecting the inland provinces with the Tibet Autonomous Region and Xinjiang via the Qinghai-Tibet Railway and an extensive network of highways.

Cultural and Scientific Significance

Xining is a multi-ethnic city where Han, Hui, Tibetan, and Tu populations coexist, creating a rich tapestry of cultural heritage. This diversity is reflected in the

city's architecture, religious sites, and local traditions. From a scientific perspective, the region provides a natural laboratory for research in plateau ecology, high-altitude medicine, and renewable energy systems. Academic institutions in Xining are increasingly focused on sustainable development strategies that balance economic growth with the preservation of the fragile plateau ecosystem.

810008; 4. Key Laboratory of Coupling Process and Effect of Natural Resources Elements, Beijing 100055)

摘要

Grassland soil salinization not only exacerbates the degradation of grassland vegetation and impairs essential ecological functions, but also significantly restricts the ecological restoration and sustainable development of degraded grasslands.

···protection and restoration. In the regions of the Source Area of the Yellow River where alpine grassland soil salinization is developing, an interpretable machine learning inversion model for salinization was constructed. This model utilizes ground-based grid sampling combined with Sentinel-2 spectral indices and incorporates feature selection techniques. The study reveals the primary influencing factors of the salinization model···

factors, and monitored the spatial distribution characteristics of regional salinization. The results indicate that: (1) The salinization inversion model constructed using stepwise regression (STEP) feature selection combined with the random forest (RF) model achieved the highest accuracy, with a coefficient of determination (R^2) of 0.64 and a root mean square error (RMSE) of···

0.76 $g \cdot kg^{-1}$. (2) Analysis of the optimal model based on SHAP (SHapley Additive exPlanations) values indicates that among the nine feature variables, the Normalized Difference Water Index (NDWI) was the most significant factor.

The Normalized Difference Water Index (NDWI) contributed most significantly to the model in this study and exhibited a positive correlation with soil salinization. (3) Monitoring results indicate that 47.13% of the study area···

The study area is characterized by varying degrees of soil degradation: 33.19% of the region is classified as moderately salinized, while mildly salinized areas also constitute a significant portion. In contrast, severely salinized soils and saline soils are less prevalent in the study area, primarily distributed along riverbanks and valley regions. By exploring interpretable salinization monitoring models through various methodological combinations, this research provides a scientific foundation and technical support for the ecological protection and high-quality development of the Yellow River Basin.

Keywords: Salinization; Machine Learning; SHAP; Alpine Grassland; Source Region of the Yellow River
Article ID: 1000-6060 (2026) 03-0549-10 (0549–0558)

Abstract

Soil salinization is a critical environmental challenge affecting the ecological stability and sustainable development of alpine grasslands. This study focuses on the Source Region of the Yellow River, utilizing machine learning algorithms integrated with SHAP (SHapley Additive exPlanations) values to analyze the spatial distribution and driving factors of soil salinity. By synthesizing multi-source geospatial data, including remote sensing indices, topographic features, and climatic variables, we developed a robust predictive model to map salinization levels across the region. Our results demonstrate that machine learning models, particularly ensemble learning methods, provide high accuracy in identifying salinized areas. Furthermore, the SHAP analysis reveals the non-linear contributions of various environmental factors, highlighting the dominant roles of evapotranspiration, groundwater depth, and vegetation cover in governing salt dynamics. These findings provide a scientific basis for land degradation assessment and the implementation of targeted ecological restoration strategies in high-altitude cold regions.

Soil salinization is one of the most prominent global issues regarding land degradation [**?**].

Studies have been conducted in various regions to investigate the distribution of grassland soil salinization, its ecological effects, and soil characteristics. These research efforts aim to understand the spatial patterns of salt accumulation and how these processes impact the surrounding environment. By analyzing the physical and chemical properties of the soil, researchers have been able to identify the primary drivers of degradation in these sensitive ecosystems.

Furthermore, the ecological consequences of increasing salinity levels have been a focal point of recent investigations. These studies highlight the significant shifts in plant community composition and the reduction in biodiversity that often accompany soil salinization. Understanding these dynamics is crucial for developing effective management strategies and restoration techniques to mitigate the negative impacts on grassland productivity and sustainability.

Soil salinization is developed to varying degrees across the Yellow River Basin in China. In recent years, with the continuous advancement of the ecological protection and high-quality development strategy for the Yellow River Basin, the monitoring and management of soil salinization have become critical scientific issues. Soil salinization not only restricts the sustainable development of regional agriculture but also poses a significant threat to the stability of the ecological environment. Therefore, accurately and efficiently obtaining information on the spatial distribution and dynamic changes of soil salinity is of great significance for the scientific utilization of land resources and the restoration of the ecological environment.

Traditional methods for monitoring soil salinity primarily rely on field sampling and laboratory analysis. Although these methods provide high measurement

accuracy, they are time-consuming, labor-intensive, and difficult to apply to large-scale, continuous spatial monitoring. With the rapid development of remote sensing technology, multispectral and hyperspectral remote sensing have become important tools for monitoring soil salinization due to their advantages of wide coverage, high timeliness, and low cost. By establishing quantitative inversion models between remote sensing spectral information and measured soil salinity data, researchers can achieve rapid mapping of soil salinity over large areas.

In the process of remote sensing inversion for soil salinity, the selection of characteristic variables and the construction of inversion models are key factors determining the accuracy of the results. Current research often utilizes original spectral bands, spectral indices (such as the Salinity Index, SI, and Vegetation Index, VI), and terrain factors as input variables. Furthermore, the application of machine learning algorithms, such as Random Forest (RF), Support Vector Machines (SVM), and Artificial Neural Networks (ANN), has significantly improved the non-linear modeling capabilities and prediction accuracy of soil salinity inversion compared to traditional linear regression methods.

However, due to the complexity of the surface environment in the Yellow River Basin—including variations in soil type, vegetation cover, and moisture content—the spectral characteristics of saline soils are often subject to interference from multiple factors. This leads to the "same object with different spectra" or "different objects with the same spectrum" phenomenon, which limits the stability and universality of inversion models. Consequently, exploring how to integrate multi-source data and optimize machine learning algorithms to improve the precision of soil salinity inversion remains a frontier topic in current remote sensing and soil science research.

Introduction

Research has been conducted on soil improvement and related fields [**?**, **?**]. However, specifically regarding the salinity of alpine grassland soils, current studies remain limited. Understanding the mechanisms of soil salinization in these high-altitude ecosystems is critical for developing effective restoration strategies and maintaining ecological stability.

In recent years, extensive research has been conducted in regions such as the middle reaches of the Yellow River, the lower reaches, and the Yellow River Delta.

Research on soil salinization in grasslands remains relatively limited. Traditional monitoring of grassland soil salinization primarily relies on field sampling and laboratory analysis. While these methods provide high accuracy, they are labor-intensive, time-consuming, and difficult to implement over large geographical areas. Consequently, they cannot meet the demands for rapid, large-scale monitoring required for modern ecological management.

In recent years, remote sensing technology has emerged as a powerful tool for monitoring soil salinization due to its wide coverage, periodicity, and cost-effectiveness. By leveraging the spectral response characteristics of saline soils, researchers have developed various spectral indices and inversion models to estimate soil salt content (SSC). However, the complex vegetation cover and diverse soil types in grassland ecosystems pose significant challenges to the accuracy of remote sensing inversions. Vegetation often masks the spectral signals of the underlying soil, leading to uncertainties in the relationship between surface reflectance and actual salt concentrations.

To address these challenges, machine learning and deep learning techniques have been increasingly integrated into salinization research. These advanced computational methods can capture non-linear relationships between multi-source environmental variables and soil properties more effectively than traditional linear regression models. By combining high-resolution satellite imagery with topographic data, climatic factors, and field observations, these models offer a promising pathway for improving the precision and spatial resolution of grassland soil salinity mapping.

A significant amount of work has been conducted regarding the investigation, monitoring, improvement, and remediation of cropland soil salinization.

Monitoring is primarily based on manual ground surveys; however, these manual investigations are time-consuming and labor-intensive.

However, in the Source Region of the Yellow River, the localized effects of drought and freeze-thaw cycles significantly impact the hydrological processes. These environmental stressors lead to complex variations in soil moisture and vegetation cover, which in turn influence the regional water balance. Understanding these dynamics is crucial for accurate climate modeling and water resource management in this ecologically sensitive area.

Due to factors such as high labor intensity, high survey costs, and limited coverage area, traditional monitoring methods face significant challenges in large-scale ecological assessments. These conventional approaches often struggle to provide the temporal and spatial resolution required for comprehensive environmental management. Consequently, there is an increasing need for more efficient, automated, and cost-effective solutions to supplement or replace manual field surveys. Recent advancements in remote sensing and automated data collection technologies offer promising alternatives to overcome these inherent limitations.

In many regions, varying degrees of grassland soil salinization have developed.

methods struggle to meet practical requirements [**?**, **?**]. Satellite remote sensing, due to its high

Currently, this issue has not yet received widespread attention [**?**]. Soil salinization in grasslands not only restricts the growth and development of vegetation but also leads to a significant decline in ecosystem productivity and biodiversity.

Furthermore, the accumulation of soluble salts alters the physical and chemical properties of the soil, such as increasing soil pH and bulk density while reducing soil porosity and nutrient availability. These changes create a feedback loop that further exacerbates land degradation, posing a severe threat to regional ecological security and sustainable agricultural development.

Due to their high temporal and spatial resolution and extensive monitoring coverage, these methods have been widely adopted in current research.

···exacerbates the degradation of grassland vegetation, thereby impacting regional water conservation and soil and water retention.

widely applied in soil salinization monitoring [**?**]. Remote sensing monitoring of soil salinization···

ecosystem functions [**?**], while simultaneously restricting the ecological restoration of degraded alpine grasslands.

Monitoring is primarily achieved through the integration of ground-based monitoring data with remote sensing spectral information.

Therefore, it is essential to carry out monitoring of soil salinization in the alpine grasslands of the Yellow River Source Region.

The construction of relational models for the inversion of surface salinization has received widespread attention in recent years. Soil salinization is a major environmental challenge that threatens agricultural productivity and ecological stability, particularly in arid and semi-arid regions. To address this, researchers have increasingly turned to remote sensing technologies and advanced modeling techniques to monitor and quantify salt content across vast landscapes.

By integrating multi-source geospatial data—including satellite imagery, topographic indices, and climatic variables—scientists are developing sophisticated mathematical and statistical frameworks to map salinity levels. These relational models range from traditional linear regressions to more complex machine learning algorithms, such as random forests and neural networks, which can capture the non-linear relationships between surface reflectance and soil salt concentration. The ultimate goal of these inversion models is to provide accurate, high-resolution spatial data that can inform land management strategies and mitigation efforts.

The research holds significant importance for the strategy of ecological protection and high-quality development in the Yellow River Basin.

The primary focus of current research is the accuracy of remote sensing monitoring models for soil salinization [**?**].

The development of advanced methodologies in this field holds significant importance. By integrating machine learning and deep learning techniques, researchers can achieve higher levels of precision and efficiency in data analysis.

These advancements not only facilitate a deeper understanding of complex systems but also provide robust frameworks for addressing long-standing challenges in scientific research. Furthermore, the application of these computational tools enables the processing of large-scale datasets, leading to more reliable predictions and innovative insights that were previously unattainable through traditional analytical methods. Ultimately, this progress contributes to the broader scientific community by establishing new standards for technical accuracy and methodological rigor.

The accuracy and reliability of the model depend on two primary factors: first, the sample size of the ground monitoring data, and second, the quality and resolution of the remote sensing data. Ground-based observations provide the essential "truth" data required for calibration and validation, while remote sensing provides the spatial coverage necessary for large-scale analysis. Ensuring a robust integration of these two data sources is critical for minimizing uncertainty in the final estimates.

[3-4]

Currently, the problem of soil salinization in the arid grasslands of China is highly prominent. Previous researchers have conducted extensive studies in regions such as the desert oases of Xinjiang and the Songnen Plain in Northeast China.

Remote sensing spectral variables currently include a wide array of salinity indices, vegetation indices, water indices, and soil indices [**?**]. Because these variables differ in their sensitivity to surface conditions, their selection is critical for accurate estimation.

The accuracy of predictive modeling is significantly influenced by the size of the ground-truth sample set and the specific remote sensing spectral variables incorporated into the model. Furthermore, while sensors such as MODIS and Landsat are widely utilized, their performance can vary across different landscapes.

Permafrost is widely distributed in this region, and the characteristics of the freeze-thaw cycle across the four seasons are distinct [**?**].

Remote sensing data from sources such as the Sentinel satellites and Unmanned Aerial Vehicles (UAVs) are increasingly favored due to their high spatial and temporal resolution.

1.2 数据来源与处理

Differences in resolution also affect the accuracy of salinization inversion.

1.2.1 土壤样品采集与处理本研究土壤样品采集

Model methods: in contrast to traditional linear models, those based on machine learning in recent years···

From August 20 to September 10, 2024, and August 2, 2025, respectively···

[18-19]

The accuracy of non-parametric machine learning models, such as eXtreme Gradient Boosting (XGBoost) and Random Forest (RF), has improved significantly.

Machine Learning Models

Machine learning models represent a core component of modern artificial intelligence, providing the mathematical frameworks necessary for systems to learn from data and make predictions or decisions without being explicitly programmed for specific tasks. These models function by identifying complex patterns and statistical regularities within training datasets, which are then generalized to process unseen information. In the context of contemporary research, machine learning models are typically categorized into supervised learning, unsupervised learning, and reinforcement learning, each serving distinct analytical purposes ranging from classification and regression to clustering and dimensionality reduction.

The efficacy of a machine learning model is fundamentally determined by its architecture, the quality of the input data, and the optimization algorithms used during the training phase. Recent advancements in deep learning have significantly expanded the capabilities of these models, particularly through the use of multi-layered neural networks that can automatically extract hierarchical features from raw data. As these models continue to evolve, ensuring their interpretability, robustness, and computational efficiency remains a primary focus of academic inquiry and industrial application.

Machine learning is capable of incorporating a vast array of spectral indices and iteratively training them to develop an ideal model. However, previous research has demonstrated that an increase in the number of variables involved in modeling does not necessarily lead to higher model accuracy. Selecting appropriate feature selection methods is essential, as it not only addresses the issue of multicollinearity among variables but also enhances the overall precision of the model. Consequently, feature screening during the modeling process is a critical step.

The average annual precipitation ranges from 300 to 500 mm, while the average evaporation ranges from 1000 to 1500 mm.

The sampling process was conducted sequentially. To ensure that the sampling points covered the entire study area as uniformly as possible, a 2 km × 2 km grid was utilized to deploy the sampling points in 2024. These sampling points ···

The study area spatially covers diverse geomorphological units, including mountains, plains, and river valleys. A total of 103 soil samples were initially collected across these regions. Building upon this foundation, an additional 26 samples were strategically collected in 2025, focusing specifically on areas with prominent distributions of saline-alkali patches.

Soil samples. Due to constraints such as terrain and road conditions during the actual sampling process, a total of 129 soil samples were ultimately collected [Figure 1: see original paper]. The specific sampling procedure···

During the process, geographical coordinates, surface vegetation coverage, and other relevant parameters were recorded at each sampling point.

This is particularly important [**?**, **?**]. Furthermore, because machine learning models function as "black boxes," the specific contributions of the feature variables involved in the modeling process cannot be clearly expressed, resulting in a lack of model interpretability.

Shapley Additive Explanations (SHAP) is a method used to interpret the predictions of machine learning models. Based on the concept of Shapley values from cooperative game theory, SHAP assigns each feature an importance value for a particular prediction. By calculating the contribution of each feature to the difference between the actual prediction and the average prediction, SHAP provides a mathematically rigorous way to understand how individual features influence the model' s output. This approach ensures consistency and local accuracy, making it a widely adopted framework for enhancing the transparency and explainability of complex "black-box" models.

The SHAP (SHapley Additive exPlanations) interpretability method [**?**] can quantify the contribution of different feature variables to the model' s output.

Information regarding the dominant vegetation species and the presence of saline-alkaline patches was recorded. Subsequently, soil samples were collected from a depth of 0–20 cm using a soil auger with an internal diameter of 7 cm. To minimize accidental errors caused by micro-topography and other local variations, a triangular sampling method was employed within a 1 m^2 area at each sampling site. Three soil cores were collected and thoroughly mixed to form a composite sample, which was then passed through a 2 mm sieve.

The contribution of these models is increasingly being integrated with machine learning models for application in various fields.

The samples were sieved to remove large soil clods, stones, and vegetation roots, then placed into sample bags and labeled accordingly.

In various modeling studies [**?**], interpretability methods have been widely applied. For instance, Jia et al. [**?**] utilized the SHAP algorithm to analyze their results.

Finally, the samples were brought back to the laboratory for air-drying, and the soil water-soluble salt content was determined using the gravimetric method.

Abstract

This study elucidates the optimal models for soil salinity inversion in arid and coastal regions. By leveraging advanced remote sensing techniques and machine learning algorithms, we analyze the spectral characteristics of saline soils across diverse geographical contexts. The research identifies key environmental covariates and spectral indices that enhance the accuracy of salinity estimation. Our findings demonstrate that while specific model architectures may vary in performance depending on regional soil properties, certain integrated approaches consistently provide superior predictive capabilities for monitoring soil degradation and managing land resources in these vulnerable ecosystems.

1. Introduction

Soil salinization represents a critical environmental challenge, particularly in arid and coastal landscapes where it threatens agricultural productivity and ecosystem stability. Arid regions are characterized by high evaporation rates and limited leaching, leading to the accumulation of salts in the surface soil layers. Conversely, coastal areas face salinization primarily through seawater intrusion and tidal influences. Accurate mapping and monitoring of soil salinity are essential for implementing effective mitigation strategies.

Recent advancements in satellite remote sensing have provided powerful tools for large-scale soil salinity assessment. However, the complexity of soil matrices and the non-linear relationship between spectral reflectance and salt content necessitate the development of robust inversion models. This paper evaluates various modeling frameworks, ranging from traditional linear regressions to sophisticated deep learning architectures, to determine the most effective methods for salinity quantification in these distinct environments.

2. Methodology and Data Processing

The inversion process begins with the acquisition of multi-source geospatial data, including multispectral satellite imagery and ground-truth soil samples. To account for the unique characteristics of arid and coastal soils, we employ specific preprocessing techniques to enhance the signal-to-noise ratio of the spectral data.

2.1 Spectral Index Construction

We utilize a suite of Salinity Indices (SI) derived from different spectral bands. These indices are designed to capture the unique absorption features of salt minerals. In addition to standard indices, we incorporate vegetation indices such as the Normalized Difference Vegetation Index (NDVI) as proxies for soil health, particularly in areas where salt stress inhibits plant growth.

2.2 Machine Learning Algorithms

Several machine learning models were evaluated for their inversion performance: - **Random Forest (RF):** An ensemble learning method that handles high-dimensional data and non-linear relationships effectively. - **Support Vector Machines (SVM):** Useful for classification and regression tasks in high-dimensional spaces. - **Gradient Boosting Decision Trees (GBDT):** An iterative approach that minimizes loss functions to improve prediction accuracy. - **

The total amount of soluble salts was used as the total soil salt content for each sample [**?**].

The range of SHAP values represents the importance of a feature, where positive and negative values indicate the direction and magnitude of the feature' s influence on the model' s prediction. Specifically, a positive SHAP value suggests that the feature increases the predicted outcome relative to the baseline (the average prediction across the dataset), while a negative SHAP value indicates that the feature decreases the predicted outcome. The absolute value of the SHAP score reflects the strength of this impact; the further the value is from zero, the more significant the feature' s contribution to that specific prediction. By aggregating these values across all instances, researchers can determine global feature importance and understand how individual variables drive the behavior of complex machine learning models.

1.2.2 遥感光谱指数获取与处理本研究遥感影像

values represent the promotion and hindrance of features toward the model' s predictions, respectively. The model

The SHAP interpretability analysis not only reveals the underlying variable-driven mechanisms of the model but also provides a valuable reference for gaining a deeper understanding of the factors influencing soil salinization [**?**].

This study focuses on a representative region of the Yellow River Source Area characterized by the development of soil salinization in arid grasslands. By integrating ground-based grid sampling with the spectral characteristics of Sentinel remote sensing data, we employed various variable selection methods in conjunction with interpretable machine learning models. The objectives were to monitor the distribution characteristics of soil salinization within the study area and to reveal the underlying factors influencing regional soil salinization inversion. This research aims to provide a scientific reference for the ecological protection, restoration, and management of alpine grasslands.

Data was obtained from the Google Earth Engine (GEE) cloud platform. Utilizing the GEE platform' s high-performance parallel computing capabilities and its extensive archive of multi-source geospatial data, we processed and analyzed the required datasets. This approach allows for efficient large-scale environmental monitoring and data extraction, bypassing the limitations of traditional local

processing methods.

The editing platform invoked data from August to September in both 2024 and 2025.

Using Sentinel-2 data, a multispectral remote sensing image of the research area was obtained through maximum value composite (MVC) synthesis. A total of 22 spectral indices were subsequently calculated and downloaded based on these multispectral bands.

indices, which include 16 salinity indices and 6 indices reflecting information such as surface vegetation, water bodies, and soil (Table 1). The spatial resolution of the image data is 10 m, and the data format is TIFF. Subsequently, within the ArcGIS software environment···

In this study, spectral indices were extracted through spatiotemporal matching, yielding 103 samples from 2024 and 26 samples from 2025. These two datasets were subsequently merged for further analysis.

All sample data collected throughout the year were used to construct the modeling dataset, which served as the basis for the subsequent modeling process.

1.1 研究区概况

multicollinearity among variables and improve model accuracy. Therefore, this study···

1.3.1 特征变量筛选特征筛选可以有效降低变量

The study area is located in Maduo County, Golog Tibetan Autonomous Prefecture, Qinghai Province.

Using Pearson Correlation (PC), Genetic Algorithms (GA), and Recursive Feature Elimination (RFE), we performed feature selection on the initial feature set to identify the optimal subset of features.

Feature Selection Methodology

To improve model performance and reduce computational complexity, we employed a multi-stage feature selection process. First, Pearson Correlation (PC) was utilized to analyze the linear relationship between individual features and the target variable, allowing for the removal of redundant or irrelevant variables. Subsequently, a Genetic Algorithm (GA) was implemented to perform a global search across the feature space, identifying combinations of features that maximize predictive accuracy through heuristic optimization. Finally, Recursive Feature Elimination (RFE) was applied to iteratively prune the least significant features based on model weights, ensuring that only the most robust predictors remained in the final model.

Located on the northern shore of Ngoring Lake within the region, this area serves as a critical ecological functional zone at the source of the Yellow River.

Variable Selection Methods

In this study, we employ four distinct variable selection methods, including Recursive Feature Elimination (RFE) and Stepwise Regression (STEP), to identify the most significant predictors for our model.

Recursive Feature Elimination (RFE)

Recursive Feature Elimination (RFE) is a backward selection algorithm that aims to find the best performing subset of variables by repeatedly constructing a model and removing the least important features. In each iteration, the importance of each feature is calculated using a base estimator (such as a Support Vector Machine or Random Forest). The feature with the lowest importance score is pruned, and the process is repeated on the remaining set until the desired number of features is reached. This approach effectively accounts for feature dependencies and interactions, ensuring that the final subset provides the highest predictive power.

Stepwise Regression (STEP)

Stepwise Regression (STEP) is an automated procedure used to select the most relevant independent variables for a regression model. This method typically combines forward selection and backward elimination. Starting with an empty model, variables are added one by one based on their statistical significance (usually determined by p-values or the Akaike Information Criterion, AIC). After each addition, the algorithm re-evaluates the existing variables to determine if any should be removed due to a loss of significance when new variables are introduced. This iterative process continues until no more variables can be added or removed according to the predefined entry and exit criteria, resulting in a parsimonious model that retains only the most impactful predictors.

The study area is located within the core zone of the Yellow River Source Sector of the Sanjiangyuan National Park. The topography is characterized by higher elevations in the north and south and lower elevations in the center, with the central region primarily consisting of piedmont proluvial-alluvial plains. The average elevation of the area is 4,500 m, and the Yellow River traverses the study site. The surface vegetation is dominated by alpine steppe, which is characterized by sparse coverage and ecological fragility; distinct saline-alkali patches are visible in localized areas. The climate of the study area is cold and dry, with a long-term annual mean temperature of -3 °C.

The 23 initially constructed spectral variable features were screened using specific methods. This process aimed to identify the most significant predictors while reducing dimensionality and eliminating redundant information.

[Figure 1: see original paper]

By applying these feature selection techniques, we ensured that the resulting model maintains high computational efficiency and robust predictive performance. The selection process focused on variables that demonstrated the strongest correlation with the target biochemical properties, thereby enhancing the overall accuracy of the machine learning framework.

In the field of optimization, the Genetic Algorithm (GA) is a method used to search for optimal solutions by simulating the processes of natural evolution.

method, which is characterized by its capacity for stochastic global optimization [**?**]. By iteratively training the model, RFE can eliminate unimportant variable features to select the optimal feature subset [**?**]. STEP functions by incrementally adding···

Inversion of Alpine Grassland Soil Salinization Based on Interpretable Machine Learning Models

Abstract

Soil salinization is a critical factor limiting the sustainable development of animal husbandry and the ecological stability of alpine grasslands. Accurate monitoring and inversion of soil salt content (SSC) are essential for land degradation assessment and ecological restoration. This study focuses on the alpine grasslands of the Qinghai-Tibetan Plateau, utilizing multi-source remote sensing data, including Sentinel-2 multispectral imagery and topographic factors. We developed a soil salinity inversion framework by integrating various machine learning algorithms, such as Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM). To address the "black box" nature of these models, we employed SHapley Additive exPlanations (SHAP) to interpret the contribution of different environmental variables to the inversion results. Our findings indicate that the integrated machine learning models significantly outperform traditional linear regression models in terms of accuracy. Specifically, the XGBoost model achieved the highest precision, with an R^2 of 0.64 and a Root Mean Square Error (RMSE) of 0.76 $g \cdot kg^{-1}$. The SHAP analysis revealed that the Normalized Difference Vegetation Index (NDVI) and elevation are the most influential predictors for soil salinity in this region. This research provides a robust methodological framework for high-precision monitoring of soil salinization in alpine ecosystems and offers a scientific basis for regional ecological management.

---

1. Introduction

Alpine grasslands are among the most sensitive ecosystems to climate change and human activities. In recent years, the synergistic effects of global warm-

ing and overgrazing have led to varying degrees of land degradation, with soil salinization emerging as a prominent issue. Soil salinization not only alters the physical and chemical properties of the soil but also inhibits vegetation growth, reduces biodiversity, and threatens the livelihood of local herders [**?**, **?**]. Therefore, achieving rapid and accurate spatial mapping of soil salt content (SSC) is of great practical significance for the protection and sustainable use of alpine grassland resources.

Traditional soil salinity monitoring relies heavily on field sampling and laboratory analysis. While these methods provide high accuracy at specific points, they are time-consuming, labor-intensive, and difficult to apply across large geographical scales. Remote sensing technology, characterized by its wide coverage and multi-temporal observation capabilities, has become the primary tool for regional soil salinity monitoring [**?**]. Previous

The map is produced based on the standard map with the approval number GS(2019)1822 from the Standard Map Service website of the Ministry of Natural Resources. The boundaries of the base map have not been modified.

1 Geographic location of the study area and spatial distribution of sampling sites

Calculation Formulas of Salinity and Vegetation Indices

The following table details the calculation formulas for the salinity and vegetation indices utilized in this study:

- **Salinity Index 1 (S1):** $\sqrt{B \times R}$
- **Salinity Index 2 (S2):** $\sqrt{G \times R}$
- **Salinity Index 3 (S3):** $\sqrt{G^2 + R^2 + NIR^2}$
- **Salinity Index 4 (S4):** $\sqrt{G^2 + R^2}$
- **Salinity Index 5 (S5):** $B \times R$
- **Salinity Index 6 (S6):** $G \times R$
- **Salinity Index 7 (S7):** $(G \times R)/B$
- **Salinity Index 8 (S8):** $(B + R)/G$
- **Salinity Index 9 (S9):** $(B + R)/(G + R)$
- **Salinity Index (SI):** $\sqrt{B \times R}$
- **Salinity Index T (SI_T):** $R/NIR \times 100$
- **Salinity Index 1 (SI1):** $\sqrt{B^2 + G^2}$
- **Salinity Index 2 (SI2):** $\sqrt{G^2 + R^2}$
- **Salinity Index 3 (SI3):** $\sqrt{R^2 + NIR^2}$
- **Salinity Index 4 (SI4):** $\sqrt{G^2 + R^2 + NIR^2}$
- **Normalized Difference Salinity Index (NDSI):** $(R - NIR)/(R + NIR)$
- **Normalized Difference Vegetation Index (NDVI):** $(NIR - R)/(NIR + R)$
- **Enhanced Vegetation Index (EVI):** $2.5 \times \frac{NIR-R}{NIR+6\times R-7.5\times B+1}$

- **Normalized Difference Phenology Index (NDPI):** $\frac{NIR-(0.74\times R+0.26\times SWIR1)}{NIR+(0.74\times R+0.26\times SWIR1)}$

参考文献

Blue Red

(Blue - Red) (Blue + Red) Green × Red Blue

Blue × NIR

Blue × Red Green

Red × NIR Green

(Swir1 - Swir2) (Swri1 + Swri2) (Green + Red) 2 Blue + Red

Green + Red + NIR

Green2 + Red2 Swir1 NIR

(Red - NIR) (Red + NIR) (NIR - Red) (NIR + Red)

2.5 × [(NIR - Red)/(NIR + 6 × Red - 7.5 × Blue + 1)]

[NIR -(0.74 × Red + 0.26 × Swir1)] [NIR +(0.74 × Red + 0.26 × Swir1)] NIR (NIR + Red)

(NIR - Swir1) (NIR + Swir1)

1.5 ×(NIR - Red) (NIR + Red + 0.5)

Note: Red, Green, Blue, NIR, Swir1, and Swir2 represent the reflectance values for the Red, Green, Blue, Near-Infrared, Short-wave Infrared 1, and Short-wave Infrared 2 bands, respectively.

Red × NIR × 100

(Green + Red + NIR) 2

Green × Red

Features are added or removed to construct the model, aiming to identify the optimal feature subset while simultaneously avoiding overfitting.

The four variable selection methods mentioned above are

the degree of influence on total soil salt content. A larger absolute value indicates a greater impact on the prediction of soil

methods commonly used in recent years. Through these four variable selection methods,

total salt content.

model performance. All variable selection procedures were implemented within the RStudio environment.

2 结果与分析

Modeling the results obtained from these methods allows for a comparison of model accuracy to evaluate the performance of different approaches.

1.3.2 机器学习模型构建与精度评价不同的机器

Learning methods exhibit varying performance across different studies and datasets. In this research, we selected two machine learning algorithms that have demonstrated superior performance in recent years: XGBoost and Random Forest (RF).

These machine learning methods were utilized for model training and accuracy comparison to identify the optimal precision.

The SHAP (SHapley Additive exPlanations) values of the spectral variables were calculated, where the absolute value of the SHAP score reflects the contribution of each variable to the soil prediction.

2.1 实测土壤含盐量

Based on the statistical analysis of 129 measured soil salinity data points, the study area

exhibits an average soil salinity of 3.57 $g \cdot kg^{-1}$, with values ranging from a minimum of 0.57 $g \cdot kg^{-1}$ to a maximum of 11.65 $g \cdot kg^{-1}$. The standard deviation is 1.45 $g \cdot kg^{-1}$, and the coefficient of variation

and the model method with the best robustness was selected as the optimal machine learning model for soil salinity

is 40.61%, indicating that the measured soil salinity data possesses a high degree of spatial dispersion

inversion mapping. Among the models evaluated, XGBoost is

classified according to established soil salinization standards [**?**]. Statistical results indicate that the study area overall

an ensemble of many tree models that forms a powerful classifier. By introducing the number of subtrees and the values of leaf nodes into the loss function, it fully accounts for regularization and effectively avoids overfitting [**?**]. Random Forest (RF) is a non-parametric machine learning algorithm that trains samples using multiple decision trees and integrates their predictions. When addressing regression prediction problems, it aggregates the results by averaging the predicted values of multiple decision trees. Compared to individual decision tree algorithms, Random Forest exhibits stronger anti-interference capabilities and superior model generalization [**?**]. In this study, the measured total soil salt content is used as the dependent variable, while spectral indices selected from different feature variables serve as

high and strong variability. According to the classification standards, the area is characterized as mildly salinized. Specifically, there are 18 non-salinized measured samples

with an average salinity of 0.80 $g \cdot kg^{-1}$. There are 86 mildly salinized samples with an average salinity of 1.44 $g \cdot kg^{-1}$. Moderately salinized samples

total 17, with an average salinity of 2.66 $g \cdot kg^{-1}$. Severely salinized soil and saline soil account for only 4 samples each, with average salinities of 4.90 $g \cdot kg^{-1}$ and 8.07 $g \cdot kg^{-1}$, respectively (Table 2).

4 种特征变量筛选结果表明，不同的方法筛选

Machine learning modeling was performed using the independent variables, with 70% of the data allocated for model training and the remaining 30% reserved for model validation.

The results obtained from different methods exhibited significant variation. Specifically, the Pearson Correlation (PC) method ($P < 0.01$) identified 8 specific features.

The accuracy of each model was evaluated through a process involving model training and hyperparameter optimization.

Vegetation indices, including NDWI, SAVI, and EVI, were utilized. The stepwise regression (STEP) method selected 9 features.

Model training and validation were conducted using a 70/30 data split, supplemented by 10-fold cross-validation.

All procedures, including model training, hyperparameter optimization, and accuracy calculations, were implemented within the RStudio environment.

Model performance was assessed using three metrics: the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE).

1.3.3 最优机器学习模型解释 SHAP 可以针对机

Methodology and Analysis

Machine learning black-box models are utilized to isolate the marginal contribution of each independent variable to the dependent variable. By integrating the optimal machine learning model with SHAP (SHapley Additive exPlanations) analysis methods, we can achieve a high degree of interpretability. This approach allows for a granular decomposition of model predictions, quantifying how much each feature influences the final output.

The SHAP framework provides a mathematically rigorous way to assign importance values to features based on cooperative game theory. Unlike traditional global importance metrics, SHAP values offer local explanations, showing how

specific inputs drive the model' s decision-making process for individual observations. This transparency is crucial for validating the reliability of complex models and ensuring that the underlying physical or economic relationships are accurately captured.

method, quantitatively explaining the importance of all spectral feature variables in the optimal machine learning model and how variations in these spectral features influence the prediction of total soil salt content. In RStudio, the "Shap" package was utilized to calculate all SHAP values.

feature variables, including 4 salinity indices (S1, S2, S4, SI4) and 3 vegetation variables, including 6 salinity indices (S1, S2, S5, SI, SI1, SI2) and 3

vegetation indices (EVI, NDPI, NDWI). The Genetic Algorithm (GA) selected four characteristic variables, specifically four salinity indices (S2, S6, S7, SI4). Recursive Feature Elimination (RFE) screened···

Five feature variables were selected for the study, consisting of four salinity indices ($S1$, $S7$, $S6$, and $SI4$) and one vegetation index ($NDPI$).

2.3 模型精度评价

Based on the results of the screening of four types of characteristic variables, combined with Random Forest (RF) and XGBoost (XG-Boost) algorithms, this study constructs a predictive model for the target variable. By integrating these feature selection methods, we aim to enhance the model' s interpretability and predictive accuracy. The selected features undergo rigorous validation to ensure they capture the essential underlying patterns within the dataset, providing a robust foundation for the subsequent machine learning analysis.

Two machine learning methods were employed to construct eight distinct salinization estimation models. Accuracy validation using the test set demonstrated that different feature combinations significantly influenced model performance.

2 Statistical Analysis of Measured Soil Salinity

The statistical analysis of the measured soil salinity data provides a foundational understanding of the study area' s characteristics. Based on the degree of salinization, the collected samples were categorized into three primary levels: light salinization, moderate salinization, and severe salinization. This classification allows for a more nuanced evaluation of the spatial distribution and severity of soil degradation across the region.

As shown in , the descriptive statistics for all measured data points reveal significant variability in salt content. For the light salinization category, the soil salinity levels remain relatively low, representing the initial stages of salt accumulation. In contrast, the moderate salinization samples show a marked increase in salt concentration, indicating a more advanced state of soil degradation. The severe salinization category represents the most extreme cases, where

high salt concentrations pose significant challenges to agricultural productivity and ecosystem health.

The comprehensive analysis of all measured data indicates that the study area is characterized by a wide range of salinity values. The mean and standard deviation across these categories highlight the heterogeneous nature of soil salinity distribution. These statistical insights are crucial for calibrating the machine learning models and ensuring the accuracy of the subsequent mapping and predictive analysis. By distinguishing between these different levels of salinization, we can better understand the environmental drivers contributing to soil salinity in the region.

Minimum value / $g \cdot kg^{-1}$

Maximum value / $g \cdot kg^{-1}$

Mean value / $g \cdot kg^{-1}$

Standard deviation / $g \cdot kg^{-1}$

Coefficient of Variation /%

Grading Standard / $\mathrm{g \cdot kg^{-1}}$

Inversion of Soil Salinization in Alpine Grasslands Based on Interpretable Machine Learning Models

Abstract

Soil salinization is a critical factor limiting the sustainable development of animal husbandry and ecological security in the alpine grasslands of the Qinghai-Tibet Plateau. Accurate monitoring and inversion of soil salt content (SSC) are essential for land degradation assessment and ecological restoration. This study utilizes multi-source remote sensing data, including Sentinel-2 multispectral imagery and topographic factors, to construct an inversion model for soil salinization in alpine grasslands. To address the "black box" nature of traditional machine learning models, we introduce an interpretable machine learning framework. By combining the Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM) algorithms with SHapley Additive exPlanations (SHAP), we quantify the contribution of different environmental variables to SSC. The results indicate that the integrated models significantly outperform traditional linear regression models in terms of accuracy. Furthermore, the SHAP analysis reveals that vegetation indices and elevation are the primary drivers of soil salinity distribution in the study area. This research provides a scientific basis and technical support for the precise management and ecological protection of alpine grassland soils.

1. Introduction

The alpine grasslands of the Qinghai-Tibet Plateau constitute a unique and fragile ecosystem that plays a vital role in global climate regulation and water conservation. However, in recent years, due to the dual pressures of climate change and intensified human activities, soil salinization has become increasingly prominent in certain regions. Soil salinization not only leads to a decline in grassland productivity and biodiversity loss but also threatens the regional ecological balance. Therefore, achieving rapid and accurate monitoring of soil salt content (SSC) is of great significance for the scientific management of these ecosystems.

Remote sensing technology has become the primary means of monitoring soil salinization at large scales due to its advantages of wide coverage and high temporal resolution. Previous studies have demonstrated that multispectral data, particularly indices derived from the visible and near-infrared bands, are highly sensitive to soil salinity. However, the relationship between SSC and remote sensing variables is often complex and non-linear, making traditional statistical methods insufficient for high-precision inversion.

Machine learning (ML) algorithms, such as Random Forest (RF) and Gradient Boosting Decision Trees (GBDT), have shown excellent performance in handling high-dimensional data and non-linear relationships. Despite their predictive power, these models are often criticized as "black

The accuracy of the models constructed using the variable selection results exhibits significant differences. Specifically, the models built using Random Forest (RF) outperformed those constructed with XGBoost. For the RF-based models, the R^2 values ranged from 0.54 to 0.64, while the Root Mean Square Error (RMSE) values were between 0.76 and 1.42 $g \cdot kg^{-1}$, and the Mean Absolute Error (MAE) values···

The model demonstrates strong predictive performance for soil salinity levels between 1 and 6 $g \cdot kg^{-1}$. However, it exhibits a notable tendency to overestimate salinity in non-salinized soils (less than 1 $g \cdot kg^{-1}$) and significantly underestimate salinity in saline soils exceeding 6 $g \cdot kg^{-1}$.

is 0.53–0.79 $g \cdot kg^{-1}$, while the R^2 of the model constructed using XGBoost is

2.4 基于 SHAP 值的模型解释

0.82 $g \cdot kg^{-1}$ [Figure 2: see original paper]. Based on the results of the STEP variable selection, the model demonstrated the highest accuracy across both machine learning algorithms.

The contribution of each independent variable to the modeling process was ranked according to the mean SHAP (SHapley Additive exPlanations) importance values.

By comprehensively comparing the three accuracy evaluation metrics across all models, the results indicate that the model based on RF-STEP achieved the highest performance.

According to the importance ranking, the NDWI (Normalized Difference Water Index) contributed the most to the model.

The R^2 values ranged from 0.43 to 0.57, with RMSE (Root Mean Square Error) values between 0.81 and 1.57 $g \cdot kg^{-1}$ and MAE (Mean Absolute Error) values between 0.53 and 1.12 $g \cdot kg^{-1}$.

The highest model accuracy was consistently observed across the machine learning models.

The model constructed using the STEP method achieved the optimal precision. On the test set, the evaluation metrics for this model were $R^2 = 0.64$, RMSE $= 0.76$ $g \cdot kg^{-1}$, and MAE $= 0.53$ $g \cdot kg^{-1}$ [Figure 3: see original paper].

0.53 $g \cdot kg^{-1}$ [Figure 3a: see original paper]. Simultaneously, this study analyzed the predictive capability of the optimal model [Figure 3b: see original paper].

The results demonstrate the model' s performance regarding total soil salinity.

By combining the RF-STEP model with SHAP analysis, the variables were evaluated.

The variable importance was visualized through a ranking of mean SHAP values [Figure 4a: see original paper] and a SHAP value beeswarm plot [Figure 4b: see original paper]. Among the variables, NDWI had the highest mean SHAP value of 0.359, followed by S2, S1, EVI, and SI2, respectively.

The S5 index had the lowest contribution, with a mean SHAP value of only 0.037.

The SHAP value beeswarm plot for each variable indicates that NDWI and SI2 are positively correlated with soil salinization.

This suggests that higher values for these two characteristic variables lead to higher model predictions for soil salinity.

PC represents Pearson correlation; STEP denotes Stepwise Regression; GA refers to Genetic Algorithm; RFE stands for Recursive Feature Elimination; RF represents the Random Forest model; XGBoost denotes Extreme Gradient Boosting; R^2 is the coefficient of determination; RMSE is the Root Mean Square Error; and MAE is the Mean Absolute Error.

2 Comparison of the accuracy of different modeling methods

3 Model accuracy validation

Note: S1, S2, S5, SI, SI1, and SI2 represent Salinity Indices; NDWI denotes the Normalized Difference Water Index; EVI refers to the Enhanced Vegetation Index; and NDPI represents the Normalized Difference Phenology Index.

4 Explanation of SHAP value model

The better the effectiveness of soil salinization detection, the more significant the roles of the three characteristic variables $S2$, $S1$, and EVI become.

to select the optimal subset of features. Furthermore, a comparison was conducted between two machine learning methods.

soil salinization. Additionally, the four characteristic variables $NDPI$, $SI1$, SI, and $S5$

The model constructed using RF-STEP achieved the highest precision ($R^2 = 0.64$), which is consistent with the findings of Guo Jia.

The variables are negatively correlated with soil salinization, and their low values exert a significant negative impact. The SHAP values are primarily concentrated around zero, indicating that they do not have a substantial effect on the model.

2.5 土壤盐渍化空间分布

Based on the optimal model, soil salinization in the study area was inverted and classified. From a spatial distribution perspective, the majority of the study area is characterized by mild salinization, accounting for 47.13% of the total area. This is followed by moderate salinization, which accounts for 33.19% and is primarily distributed along rivers and valley zones.

Non-salinized areas represent the third largest category at 18.28%, mainly distributed in the central part of the study area. In contrast, severe salinization and saline soils are less prevalent, accounting for only

1.13% and 0.27%, respectively. These areas are primarily located in the northern part of the study area and near the central primary

The accuracy of the model constructed using the Random Forest (RF) method is slightly superior to that of XGBoost, and it is similar to the accuracy of the RF-based model constructed by Li et al. [**?**] in July.

The R^2 values obtained by Ze et al. [**?**] and Tian Yichao et al. [**?**] for soil salinization inversion using XGBoost models were 0.54 and 0.57, respectively, which are comparable to the accuracy of the XGBoost modeling in this study.

Yang Lianbing et al. [**?**] conducted salinization inversion using different feature variable selection methods combined with RF models. Their results

indicated that the inversion model constructed using a Bayesian optimization algorithm combined with RF achieved the highest accuracy ($R^2 = 0.75$). That study incorporated 56 feature variables, including salinity indices, vegetation indices, feature space, topography, and temperature. This comprehensive consideration of factors

is likely the primary reason why their accuracy exceeds that of the present study. Furthermore, because the study area in this research is relatively small, there

was a lack of suitable spatial resolution

drainage systems (Figure 5 [Figure 5: see original paper]).

factors.

3 讨论

In the process, only vegetation indices and salinity indices were considered, while factors such as topography and surface characteristics were neglected. This limitation may affect the overall accuracy of the model, as environmental variables often play a significant role in the spatial distribution of soil properties. Future research should incorporate a broader range of auxiliary data to enhance the robustness of the predictive analysis.

data such as topography, temperature, and groundwater depth. Consequently, during the modeling process, factors such as temperature and humidity have demonstrated significant importance in previous studies of salinization modeling.

Machine learning is currently a focal point in the remote sensing modeling of soil salinization.

highlighted in [**?**, **?**], which may also account for the relatively lower accuracy of the model in this study.

Model applicability remains the most significant challenge in current research [**?**].

Due to these factors, future research will require a more comprehensive integration of feature variables.

This study utilizes four feature selection methods combined with two machine learning algorithms to conduct its analysis.

1. Introduction

In recent years, the rapid development of machine learning has provided powerful tools for processing complex datasets in scientific research. Feature selection, as a critical step in the data preprocessing phase, aims to identify the most relevant variables to improve model performance and reduce computational complexity. This study systematically evaluates the effectiveness of different feature selection strategies in enhancing the predictive accuracy of machine learning models.

2. Methodology

2.1 Feature Selection Methods

We implemented four distinct feature selection techniques to identify the optimal subset of variables. These methods include filter-based approaches, which

evaluate features based on statistical properties, and wrapper-based approaches, which utilize specific algorithms to determine the importance of each variable. By comparing these four methods, we aim to mitigate the risk of overfitting and ensure the robustness of the selected features.

2.2 Machine Learning Algorithms

Following the feature selection process, two machine learning algorithms were employed to construct predictive models. These algorithms were selected based on their proven performance in handling high-dimensional data and their ability to capture non-linear relationships within the dataset. The integration of these algorithms with the aforementioned feature selection methods allows for a comprehensive assessment of model stability and accuracy.

3. Results and Discussion

The experimental results indicate that the combination of specific feature selection methods and machine learning algorithms significantly impacts the final model performance. As shown in [Figure 1: see original paper], the reduction in feature dimensionality did not lead to a loss in predictive power; rather, it improved the interpretability of the models.

[Figure 1: see original paper]

Furthermore, the comparative analysis reveals that certain feature subsets are more sensitive to the choice of the machine learning algorithm. We observed that $\mathcal{F}$ values remained consistent across different iterations when utilizing the optimized feature sets. The relationship between the input variables and the target output can be expressed as:

$$y = f(x_i, \beta) + \epsilon$$

where x_i represents the selected feature vector and β denotes the model parameters. This approach ensures that the resulting models are both efficient and generalizable to unseen data.

Moreover, it is essential to explore the performance and characteristics of multimodal remote sensing models across various spatial scales.

machine learning models were employed to conduct remote sensing modeling of salinization, aiming to explore the performance and characteristics of different models.

applicability [**?**]. Furthermore, based on the accuracy verification of the optimal model developed in this study,

Applicability. Based on the accuracy comparisons across various models, the effectiveness of different variable selection methods varies significantly.

[Figure 1: see original paper]

The results indicate that the integration of specific feature engineering techniques with machine learning algorithms directly impacts the predictive performance of the models. While some models demonstrate high robustness across diverse datasets, others are more sensitive to the dimensionality of the input features. Therefore, selecting an appropriate variable screening strategy is crucial for optimizing model precision and ensuring the generalizability of the findings in practical applications.

As shown, the model performs well in predicting light, moderate, and severe salinization; however, it exhibits lower accuracy when identifying non-salinized or extreme salinization levels. This discrepancy may be attributed to the spectral similarity between certain soil types and the limited availability of training samples for extreme environmental conditions. To improve the robustness of the classification, further refinement of the feature selection process and the inclusion of multi-temporal remote sensing data may be necessary.

The methodology significantly impacts model accuracy. In this study, we conduct a comparative analysis of model performance as follows:

The prediction of non-salinized soils with a salt content of less than 1 $g \cdot kg^{-1}$ exhibits significant overestimation.

The results indicate that the feature selection based on the STEP method performed exceptionally well across both machine learning models. This finding is consistent with previous studies, such as [**?**, **?**], suggesting that the STEP method effectively constructs models by incrementally adding or removing features.

The prediction of saline soils with a salt content exceeding 6 $g \cdot kg^{-1}$ exhibits a significant underestimation phenomenon.

This may be attributed to the common phenomenon in random forest models where low values tend to be overestimated and high values tend to be underestimated [**?**].

Inversion of Soil Salinization in Alpine Grasslands Based on Interpretable Machine Learning Models

Abstract

Soil salinization is a critical factor leading to the degradation of alpine grasslands. Rapid and accurate monitoring of soil salt content (SSC) is essential for the ecological protection and sustainable development of these regions. This study focuses on the alpine grasslands of the Qinghai-Lake basin. By integrating multi-source remote sensing data, including Sentinel-2 multispectral imagery and topographic information, we extracted spectral bands, vegetation indices, and salinity indices as environmental covariates. We employed four machine

learning algorithms—Random Forest (RF), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and CatBoost—to construct SSC inversion models. To address the "black box" nature of these models, we utilized SHapley Additive exPlanations (SHAP) to quantify the contribution and influence of each environmental variable on soil salinization. The results indicate that the CatBoost model achieved the highest accuracy, with an R^2 of 0.78 and a Root Mean Square Error (RMSE) of 0.42 g/kg. SHAP analysis revealed that the Elevation, the B11 band (Short-Wave Infrared), and the Salinity Index (SI) were the most significant predictors. Furthermore, the study demonstrates that the relationship between environmental factors and SSC is non-linear and characterized by specific threshold effects. This research provides a scientific basis and technical support for the precise management and restoration of alpine grassland ecosystems.

1 Introduction

Alpine grasslands are a vital component of the global terrestrial ecosystem, playing an indispensable role in carbon sequestration, water conservation, and biodiversity maintenance. However, in recent years, the dual pressures of climate change and anthropogenic activities have intensified soil salinization in these regions, posing a severe threat to grassland productivity and ecological stability [**?**]. Traditional methods for monitoring soil salt content (SSC) rely on field sampling and laboratory analysis, which are time-consuming, labor-intensive, and difficult to implement over large spatial scales. Remote sensing technology, with its advantages of wide coverage and multi-temporal observation, has become a primary tool for regional soil salinity mapping [**?**].

In recent years, machine learning (ML) algorithms have been widely applied to SSC inversion due to their ability to handle high-dimensional data and complex non-linear relationships [**?**]. While models such as Random Forest and Gradient Boosting Decision Trees (

5 Spatial distributions of salinity classes

This study utilizes the SHAP (SHapley Additive exPlanations) method to interpret the contributions of various characteristic variables within the

distribution. Furthermore, moderate and severe salinization, as well as saline soils, are primarily distributed zonally along rivers

optimal model. Through importance ranking and variable bee-swarm plots, the results

and valley regions. These river and valley areas serve as catchment

indicate that the NDWI (Normalized Difference Water Index) provides the highest contribution to the model, and its high values positively

zones and represent key topographic factors for salt accumulation [**?**, **?**]. Consequently, these locations

influence soil salinization. NDWI represents surface moisture;

exhibit relatively severe soil salinization. From the perspective of the model' s overall global prediction,

the stronger the soil water-salt transport process, the more conducive it is to the formation of salinization.

there may be uncertainties in the predicted distribution areas of severe salinization and saline soils.

[19,38]

The spatial distribution of salinization levels ([Figure 5: see original paper]) further indicates that the areas with severe salinization in the study area are primarily distributed along rivers and valleys. These specific regions···

This region also exhibits a high concentration of moisture. Furthermore, low values of the Enhanced Vegetation Index (EVI) show a significant negative correlation.

···impact soil salinization. Previous studies have also demonstrated that among various surface vegetation indices, the Enhanced Vegetation Index (EVI) is particularly effective at reflecting salinization levels and exhibits a negative correlation with soil salinity [**?**, **?**]. EVI serves as a proxy for surface vegetation health; higher soil salinity creates an environment increasingly detrimental to vegetation growth. At the same time, the two salinity indices, S1 and S2···

Due to the adoption of grid sampling in this study, the systematic nature of the data collection process has been significantly enhanced. This approach ensures a more uniform distribution of sample points across the study area, thereby reducing potential spatial bias and improving the representativeness of the results. By employing a grid-based framework, we are able to capture the spatial variability of the target variables more effectively than traditional random sampling methods.

The limited number of salt-affected sampling points has resulted in a lack of sufficient training data for the model [**?**]. In future investigations of grassland soil salinization, intensified sampling should be conducted in low-lying valley areas to ensure the global representativeness of the dataset.

4 结论

The partial indices contributed significantly to the salinization modeling in this study, and their low values

(1) Selection of different characteristic variables and different machine learning models

partially exerted a negative impact on the model. Previous studies have also demonstrated that S1 and S2 play a role in salinity

significantly influence the accuracy of salinization modeling. In this study, the model constructed using Random Forest (RF)

Monitoring results from this study indicate that the research area is primarily characterized by slight and moderate

The model constructed using RF achieved the highest accuracy, with an R^2 of 0.64, an RMSE of

performed prominently among the indices and exhibited a negative correlation with soil salinity [**?**, **?**].

salinization. Regarding spatial distribution, non-salinized areas are mainly located in the central part of the study area, while slight salinization is distributed throughout the entire region.

The modeling accuracy was slightly superior to that of XGBoost. Based on the STEP variable selection results, the model achieved an RMSE of 0.76 $g \cdot kg^{-1}$ and a MAE of 0.53 $g \cdot kg^{-1}$.

(2) Model interpretation based on SHAP values indicates that the NDWI index had the highest contribution to this

study' s model, followed by S2, S1, EVI, and SI2. Among these, NDWI and SI2 were positively correlated with soil salinization, while EVI,

Xinjiang[J]. Geological Review, 2024, 70(5): 1857-1872.]

S1 and S2 are negatively correlated with soil salinization.

[10] Metternicht G I, Zinck J A. Remote sensing of soil salinity: Poten

The spatial distribution of soil salinity indicates that the majority of the soil in the study area exhibits a state of slight salinization.

[11] Wang Jingzhe, Ding Jianli, Ge Xiangyu, et al. Soil salinization monitoring based on satellite-terrestrial sensing technology.

(3) Soil salinization mapping of the study area based on the inversion of the optimal model.

The results indicate that the study area is primarily characterized by slight salinization (47.13%) and moderate salinization (33.19%). In contrast, severe salinization and saline soils are less developed in the region and are mainly distributed along rivers and valleys.

tials and constraints[J]. Remote Sensing of Environment, 2003, 85 (1): 1-20.

Progress and Prospects in Monitoring [J]. Journal of Remote Sensing, 2024, 28(9): 2187-2208.

1. Introduction

Remote sensing technology has become an indispensable tool for Earth observation, providing critical data for understanding environmental changes, resource management, and disaster response. In recent years, the rapid development of sensor technology and computational power has significantly advanced our ability to monitor complex terrestrial and atmospheric processes. This paper reviews the current progress in remote sensing monitoring and discusses future directions in the field.

2. Current State of Remote Sensing Monitoring

The integration of multi-source data, including optical, thermal, and microwave remote sensing, has greatly enhanced the precision of environmental monitoring. High-resolution satellite constellations now provide frequent revisit times, allowing for near real-time observation of dynamic phenomena.

2.1 Machine Learning and Deep Learning Applications The adoption of machine learning and deep learning has revolutionized data processing workflows. These algorithms are particularly effective at handling the high dimensionality and non-linearity of remote sensing data. For instance, convolutional neural networks (CNNs) are widely used for land cover classification and object detection, while recurrent neural networks (RNNs) are employed for time-series analysis of vegetation indices.

[Figure 1: see original paper]

As shown in [Figure 1: see original paper], the workflow for deep learning-based monitoring typically involves data preprocessing, feature extraction, and model training. The accuracy of these models often surpasses traditional statistical methods, especially when dealing with large-scale datasets.

2.2 Multi-Scale and Multi-Platform Integration Modern monitoring systems increasingly rely on the synergy between satellite, aerial (UAV), and ground-based observations. This multi-scale approach allows researchers to bridge the gap between local-scale details and global-scale patterns. For example, UAVs provide high-resolution imagery for precision agriculture, which can be used to calibrate and validate coarser satellite data.

3. Key Methodologies and Mathematical Frameworks

The quantitative analysis of remote sensing data relies on robust physical and statistical models. Radiative transfer models are essential for atmospheric correction and the retrieval of surface parameters.

The relationship between the measured radiance and the surface reflectance can be expressed as:

$$L_{total} = L_p + \frac{\rho \cdot T \cdot E_{down}}{\pi(1 - s \cdot \rho)}$$

where L_{total} is the total radiance

Jingzhe, Ding Jianli, Ge Xiangyu, et al. Monitoring soil saliniza

Zonal distribution.

tion on the basis of remote sensing and proximal soil sensing: Prog

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 55, no. 6, pp. 84–90, 2017.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S.

28(9): 2187-2208.]

ation zone of arid oasis areas: A case study of the Yanqi Basin,

ress and prospective[J]. National Remote Sensing Bulletin, 2024, [12] Abuelgasim A, Ammad R. Mapping soil salinity in arid and semi-

Butcher K, Wick A F, Desutter T, et al. Soil salinity: A threat to

global food security[J]. Agronomy Journal, 2016, 108(6): 21892200.

Current Status and Future Research Hotspots of Soil Salinization

Li Jianguo, Pu Lijie, Zhu Ming, et al. *Journal of Geographical Sciences, 2012, 67(9): 1233-1245.*

Abstract

Soil salinization is a global environmental problem that severely restricts agricultural development and ecological stability, particularly in arid and semi-arid regions. This paper systematically reviews the current status of soil salinization research, analyzing the mechanisms of salt movement, monitoring techniques, and remediation strategies. By synthesizing domestic and international literature, we identify key trends in the field, including the integration of remote sensing (RS) and geographic information systems (GIS) for large-scale monitoring, the development of multi-scale simulation models, and the shift toward ecological restoration. Finally, the paper proposes future research hotspots, emphasizing the need for interdisciplinary approaches to address the challenges posed by climate change and anthropogenic activities on soil salinity dynamics.

1 Introduction

Soil salinization is one of the most significant land degradation processes worldwide, affecting approximately 7% of the Earth' s total land area. It not only reduces crop yields and threatens food security but also leads to the deterioration of soil physical and chemical properties, loss of biodiversity, and overall ecosystem instability. In China, saline-alkali soils are widely distributed, covering approximately 3.6×10^7 hectares, primarily in the Northwest, North, and Northeast regions.

As global climate change intensifies and human activities—such as improper irrigation and land use—

Note: Figure translations are in progress. See original paper for figures.