Environ Monit Assess (2025) 197:367
https://doi.org/10.1007/s10661-025-13814-z

RESEARCH

Predicting surface soil pH spatial distribution based on three machine learning methods: a case study of Heilongjiang Province

Pu Huang · Qing Huang · Jingtian Wang · Yuhan Shi

Received: 17 December 2024 / Accepted: 27 February 2025 / Published online: 8 March 2025

Abstract Comprehensive and accurate acquisition of surface soil pH spatial distribution information is essential for monitoring soil degradation and providing scientific guidance for agricultural practices. This study focused on Heilongjiang Province in China, utilizing data from 125 soil survey sampling points. Key environmental covariates were identified as modeling inputs through Pearson correlation analysis and recursive feature elimination (RFE). Three machine learning models—support vector machine (SVM), random forest (RF), and extreme gradient boosting (XGBoost)—were employed to predict surface soil pH in the study area. The modeling outcomes and distinctions among these models were then thoroughly compared. The results showed that the mean monthly temperature maximum (MMTmax), mean monthly precipitation minimum (MMPmin), mean annual precipitation (MAP), drought index (DI), and mean monthly wind speed maximum (MMWSmax) were the most important environmental covariates for modeling. Climate variables are better suited to reflect the nonlinear relationships between soil properties and the environment in large and flat areas during mapping. Among the mapping models, XGBoost exhibited the highest prediction performance (R^2=0.705, RMSE=0.633, MAE=0.484), followed by RF (R^2=0.688, RMSE=0.656, MAE=0.497), while SVM was considered unstable in this study. For uncertainty maps, XGBoost demonstrated lower uncertainty primarily in high-altitude mountainous forest regions, whereas RF achieved higher prediction consistency mainly in low-altitude plain areas. Each prediction model had its advantages in different terrain regions, yet XGBoost was regarded as the optimal model. According to the optimal model, the typical black soil in Heilongjiang Province generally exhibited weak acidity, with an average pH of 6.42, showing a gradual increasing trend from east to west and from north to south. Soil acidification mainly occurred in the meadow black soil and albic black soil regions of Heilongjiang Province's eastern and northeastern parts. It is imperative to rigorously control the application of nitrogen fertilizers and to focus on improving the soil's acid–base buffering capacity.

Keywords Heilongjiang Province · Digital soil mapping · Surface soil pH · Machine learning · Model comparison · Uncertainty assessment

P. Huang · Q. Huang (✉) · J. Wang · Y. Shi
National Key Laboratory of Efficient Utilization of Arid and Semi-arid Arable Land in Northern China, Beijing 100081, China
e-mail: huangqing@caas.cn

P. Huang
e-mail: 2209913206@qq.com

J. Wang
e-mail: 2877129151@qq.com

Y. Shi
e-mail: 821012450685@caas.cn

P. Huang · Q. Huang · J. Wang · Y. Shi
Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences, Beijing 100081, China

1 Introduction

Soil is a crucial component of the earth's surface system and is among the planet's most valuable natural resources (Zhang et al., 2020; Zhu et al., 2015). It supports primary life processes on the earth, maintains the balance of ecosystems, and performs vital ecological functions such as carbon sequestration, nutrient cycling, and climate regulation (Li et al., 2022; Shen, 2018). However, in recent years, the burgeoning global population and the rapid advancement of industrialization have brought the issue of resource scarcity into sharp focus, particularly concerning soil quality and soil security (Hu et al., 2021). Human-induced overdevelopment has led to soil degradation, resulting in a myriad of problems including soil erosion, reduced fertility, salinization, acidification, and alkalization (Lal, 2015; Smith et al., 2016). Faced with these pressing challenges, it is imperative to monitor the extent of soil degradation and implement prompt remedial actions.

Traditional monitoring method follows a system that includes data collection, indoor preliminary prediction, field investigation, indoor interpretation, field verification, delineation, and mapping (Sun et al., 2011). However, this method is often hampered by time constraints and labor costs, making it challenging to achieve comprehensive and real-time monitoring capabilities. In contrast, digital soil mapping (DSM) uses sample data to establish correlations between soil and its environmental covariates (Zhang et al., 2017). It is an emerging and efficient method for expressing soil spatial distribution and has been widely applied in soil mapping research. For instance, Henderson et al. (2005) employed a decision tree (DT) algorithm to complete a nationwide soil properties mapping at a 250 m resolution in Australia; Wiesmeier et al. (2011) used random forest (RF) to elucidate the spatial distribution characteristics of soil organic carbon (SOC) in a semi-arid watershed area in Inner Mongolia; Heung et al. (2016) compared the mapping accuracy of 10 machine learning algorithms in a medium-scale area and found that the k-nearest neighbor (KNN) and the support vector machine (SVM) had the highest accuracy; Camera et al. (2017) compared soil maps generated by RF and logistic regression, deeming that RF had lower validation errors and prediction uncertainties; Petermann et al. (2021) used machine learning algorithms (MARS, RF, SVM) and spatial cross-validation methods to develop a radon concentration map of Germany, with results showing that RF performed significantly better than other algorithms; Nguyen et al. (2022) developed a SOC storage prediction model integrating multi-sensor data and the XGBoost algorithm, which outperformed RF and SVM models. From the above cases, it is apparent that both single models such as DT and SVM, as well as ensemble models like RF and XGBoost, are mainstream technological algorithms in the current DSM. However, each prediction model has its unique features, so when applying these models in practice, the most suitable mapping model should be selected based on different study areas and soil properties.

The black soil in Northeast China is among the world's most fertile, with a short cultivation history. It exhibits great production potential due to its excellent properties, high fertility, and suitable farming conditions. As the core region for grain production in Northeast China, Heilongjiang Province accounts for over 56% of the country's typical black soil farmland area. Therefore, conducting a comprehensive survey of the extent of soil acidification is particularly important. However, with a total area of 470,700 square kilometers, ranking sixth among China's provincial-level administrative regions, and its northern part covered by the vast Greater Khingan Range forests, Heilongjiang Province has relatively sparse soil sample data. The challenge lies in using these limited and unevenly distributed sample data to build an accurate spatial prediction model, which is where machine learning (ML) excels. By combining spatial distribution prediction results with uncertainty maps, users can quickly grasp soil conditions and conduct targeted supplementary surveys. Such research findings have significant practical value for monitoring soil degradation.

This study focused on the surface soil pH of Heilongjiang Province as the research subject. To optimize the feature subset for spatial modeling, we sequentially applied Pearson correlation analysis and recursive feature elimination (RFE) to a set of 35 environmental covariates. We then conducted a comprehensive comparison of the outcomes and distinctions among three machine learning (ML) models—SVM, RF, and XGBoost—to assess their prediction performance. Consequently, we generated grading maps and uncertainty maps, with the aim of providing a solid

scientific foundation for agricultural planning and soil management in the black soil region of Northeast China.

2 Materials and methods

2.1 Study area

Heilongjiang Province, situated in Northeast China, has a climate which is classified as cold temperate continental monsoon and temperate continental monsoon. Its geographical coordinates span from 121°11'E to 135°05'E and from 43°26'N to 53°33'N. The province is renowned for its numerous rivers and lakes, including the Heilongjiang, Songhua, Wusuli, and Suifen rivers, as well as the Xingkai Lake, Jingpo Lake, and Wudalianchi. The terrain is characterized by higher elevations in the northwest, north, and southeast, and lower elevations in the northeast and southwest (Fig. 1).

Heilongjiang Province is particularly significant for studying soil pH and acidification due to its unique soil composition and agricultural importance. The region is home to some of the most fertile black soils in the world, which are crucial for agricultural productivity. These black soils, known as "phaeozems" or "chernozems," are rich in organic matter and nutrients, making them highly productive for crops such as corn, soybeans, and wheat. However, in recent years, due to unreasonable farming practices and excessive application of nitrogen fertilizers (Ju & Gu, 2014), large areas of black soil farmland in Heilongjiang Province have become severely acidified. This acidification not only inhibits crop growth and reduces yields, but also easily leads to ecosystem imbalance, worsening agricultural production conditions, thus attracting widespread attention.

2.2 Soil sample data

The soil sample data utilized in this study were derived from *Soil Series of China: Heilongjiang Volume* (Zhai et al., 2020), with the primary collection period between 2009 and 2011. A total of 129 sampling sites were recorded, among which 125 surface soil samples (approximately 0–5 cm) contained pH information (Fig. 1). The determination and analysis of soil pH were conducted strictly in accordance with the guidelines

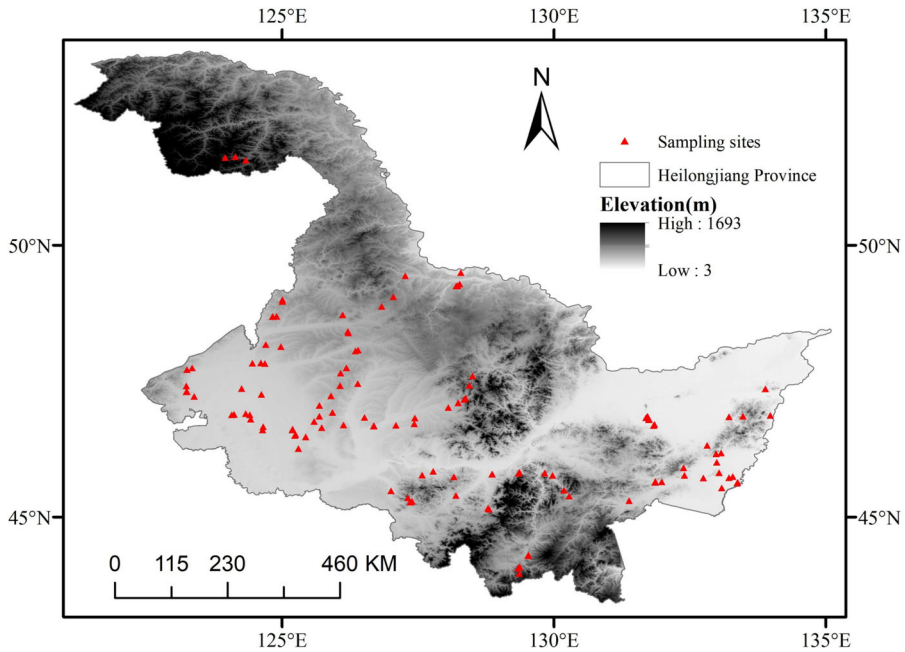

Fig. 1 Location map of the study area and the sampling sites

in *Soil Survey Laboratory Methods* (Zhang & Gong, 2012). For more details about the soil sample data, we refer to Liu et al. (2022a).

2.3 Environmental covariates

DSM is a mapping method based on the soil-landscape theory (Huggett, 1975), which posits that soil formation is primarily related to five key soil-forming factors: terrain, climate, parent material, organisms, and time. However, the comprehensive consideration of the time factor remains a scientific challenge, and no explicit breakthrough has been achieved yet. In this study, a total of 35 environmental covariates were selected from five aspects: terrain, climate, parent material, organisms, and soil factors (Table 1), to participate in the feature subset optimization of DSM. In terms of terrain, the "NASA/NASADEM_HGT/001" data product with a 30 m resolution provided by Google Earth Engine (GEE) was utilized, and the terrain-derived factors were calculated and obtained using ArcMap 10.8 software. Climate covariates (averages from the 1990s to 2010s) mainly originated from 13 datasets with a 30 m resolution, provided by the Fine Resolution Mapping of Mountain Environment (FRMM) of the Institute of Mountain Hazards and Environment, Chinese Academy of Sciences. The mean annual ground temperature data, averaging from the 1960s to the 2010s, was sourced from the Resources and Environmental Science Data Center (https://www.resdc.cn/), with a 1 km resolution. For parent material, studies by Böhner and Antonić (2009) and Liu et al. (2022b) were referenced, using Band 7 of Landsat 5 imagery as a substitute. Organism covariates were all from the year 2012, close to the sampling time. Soil data were also obtained from the Resources and Environmental Science Data Center of the Chinese Academy of Sciences, with a 1 km resolution. To meet the mapping requirements, all environmental covariates were resampled to a 100 m resolution using bilinear interpolation in ArcMap 10.8.

2.4 Optimization of environmental covariates subsets

2.4.1 Pearson correlation analysis

Pearson correlation analysis is a widely used method for assessing the degree of linear correlation among pairs of continuous variables (Guo et al., 2024). Values range from -1 to 1: values close to -1 indicate a strong negative correlation, values close to 0 indicate little to no correlation, and values close to 1 indicate a strong positive correlation. In this study, the corrplot package of the R language was employed for correlation mapping and for the elimination of environmental covariates that were not significantly correlated ($P < 0.05$) (Wei et al., 2013).

2.4.2 Recursive feature elimination

RFE is a greedy algorithm that iteratively eliminates the least important features based on model performance (e.g., R^2) (Demarchi et al., 2020; Kaya et al., 2022). In this study, RF was used as the base model, with the optimal feature subset identified through 10-fold cross-validation. The algorithm was implemented using the Caret and Random Forest packages in R language (Kuhn, 2015; Liaw & Wiener, 2015).

2.5 Spatial prediction models

2.5.1 Support vector machine

SVM is a machine learning algorithm developed based on the Vapnik-Chervonenkis (VC) dimension and the principle of structural risk minimization (Vapnik, 2013). When performing regression tasks, SVM continuously adjusts the hyperplane to minimize prediction error while maintaining the maximum margin with the data points, thereby finding the global optimal solution. In this study, the optimal modeling parameters for SVM were determined through grid search (sigma = 0.1, C = 10).

2.5.2 Random forest

RF is an ensemble machine learning algorithm based on decision trees. When applying RF, each tree is trained using a random subset of the dataset, and only a random subset of features is considered when splitting nodes, which significantly reduces the risk of model overfitting (Suleymanov et al., 2023). The final prediction result is the average of all decision tree predictions (Breiman, 2001). In this study, the optimal modeling parameters for RF were determined through grid search (ntree = 500, mtry = 2, nodesize = 5).

Table 1 Basic information of environmental covariates

Environmental covariates	Resolution/m	Data sources
Elevation	30	
Slope	30	
Aspect	30	
Plan Curvature (PLC)	30	Google Earth Engine
Profile Curvature (PRC)	30	
Topographic Relief (TR)	30	
Topographic Roughness (TRO)	30	
Topographic Wetness Index (TWI)	30	
Mean Annual Ground Temperature (MAGT)	1000	Resources and Environmental Science Data Center
Mean Annual Precipitation (MAP)	30	
Mean Month Precipitation Maximum (MMPmax)	30	
Mean Month Precipitation Minimum (MMPmin)	30	
Mean Annual Wind Speed (MAWS)	30	
Mean Month Wind Speed Maximum (MMWSmax)	30	
Mean Month Wind Speed Minimum (MMWSmin)	30	
Mean Annual Temperature (MAT)	30	Fine Resolution Mapping of Mountain Environment
Maximum Annual Temperature (ATmax)	30	
Minimum Annual Temperature (ATmin)	30	
Mean Month Temperature Maximum (MMTmax)	30	
Mean Month Temperature Minimum (MMTmin)	30	
Annual Temperature Range (ATR)	30	
Drought Index (DI)	30	
Landsat5 Band7	30	Google Earth Engine
Leaf Area Index Maximum (LAImax)	500	
Leaf Area Index Mean (LAImean)	500	
Normalized Difference Vegetation Index (NDVI)	30	Global Resources Data Cloud
Fraction Vegetation Coverage (FVC)	250	
Net Primary Production (NPP)	500	
Landsat5 Band3	30	
Landsat5 Band4	30	Google Earth Engine
Landsat5 Band5	30	
Land Use (LU)	30	International Research Center of Big Data for Sustainable Development Goals
Silt	1000	
Sand	1000	Resources and Environmental Science Data Center
Clay	1000	

2.5.3 Extreme gradient boosting

XGBoost is also an ensemble machine learning algorithm based on decision trees (Zhang et al., 2022). Unlike RF, XGBoost gradually reduces prediction errors by iteratively adding new decision trees, with each tree built on the residuals of the previous round. This approach ensures that the new trees effectively correct the prediction errors from the previous round, making it an advanced gradient-boosting algorithm. In this study, the optimal modeling parameters for XGBoost were determined through a grid search (n_estimators = 500, max_depth = 7, learning_rate = 0.1, gamma = 1, colsample_bytree = 0.7, min_child_weight = 4, subsample = 0.9).

2.6 Prediction evaluation and uncertainty estimation

10-fold cross-validation is a specialized method for evaluating the accuracy of ML models, particularly suitable for small data sample sets (Liu et al., 2022b; Wong, 2015). It divides the dataset evenly into 10 parts and then performs 10 rounds of training and validation. In each iteration, one part serves as the validation set, while the remaining nine parts constitute the training set. The final model performance evaluation is the average of the validation results from these 10 rounds. This study employed the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE) from cross-validation to assess the prediction accuracy of different models. To evaluate the uncertainty of the prediction results, we conducted 100 iterations using the model and calculated the variance of the predictions for each location, thereby creating an uncertainty map. This approach will help identify areas where the predictions may have significant errors.

3 Results

3.1 Descriptive statistical analysis of surface soil pH

The histograms of frequency of surface soil pH (Fig. 2) showed that the pH values ranged from 5.100 to 10.100. The median pH was 6.570, while the mean pH was 6.897. The standard deviation was 1.089, indicating a moderate level of variation. The skewness coefficient was 0.700, indicating a slight positive skew in the data distribution. This suggests that the distribution has a longer tail on the right side, meaning there are a few higher pH values pulling the mean slightly above the median. The kurtosis coefficient was −0.324, which indicates a flatter peak compared to a normal distribution. This platykurtic distribution implies that the data are more dispersed and have fewer extreme values than a normal distribution, which could reflect a more uniform spread of soil pH across the study area. The variation coefficient was 0.158, further confirming a moderate degree of spatial variation.

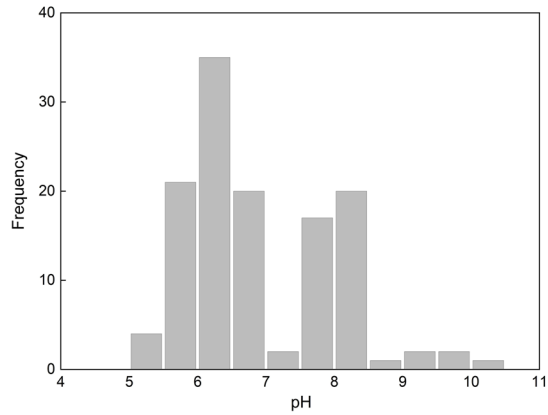

Fig. 2 Histograms of frequency of surface soil pH obtained from field survey

3.2 Optimization of environmental covariates

From Fig. 3, we can observe a positive correlation between surface soil pH and a range of variables, including Band3, Band7, MAWS, DI, MAGT, ATmin, ATmax, MAT, ATR, Sand, MMWSmax, MMTmax, and MMTmin ($P < 0.05$), with correlation coefficients ranging from 0.202 to 0.630. Notably, the strongest positive correlations were found with MMTmax, MAT, MMWSmax, and MAWS, with coefficients of 0.630, 0.448, 0.443, and 0.441 respectively, indicating a moderate level of association. Conversely, surface soil pH exhibited a negative correlation with Elevation, LAImax, Band4, TRO, TR, TWI, Silt, MAP, Slope, PRC, MMPmax, and MMPmin ($P < 0.05$), with correlation coefficients ranging from −0.625 to −0.184. Among these, the highest negative correlations were observed with MAP, MMPmin, LAImax, and

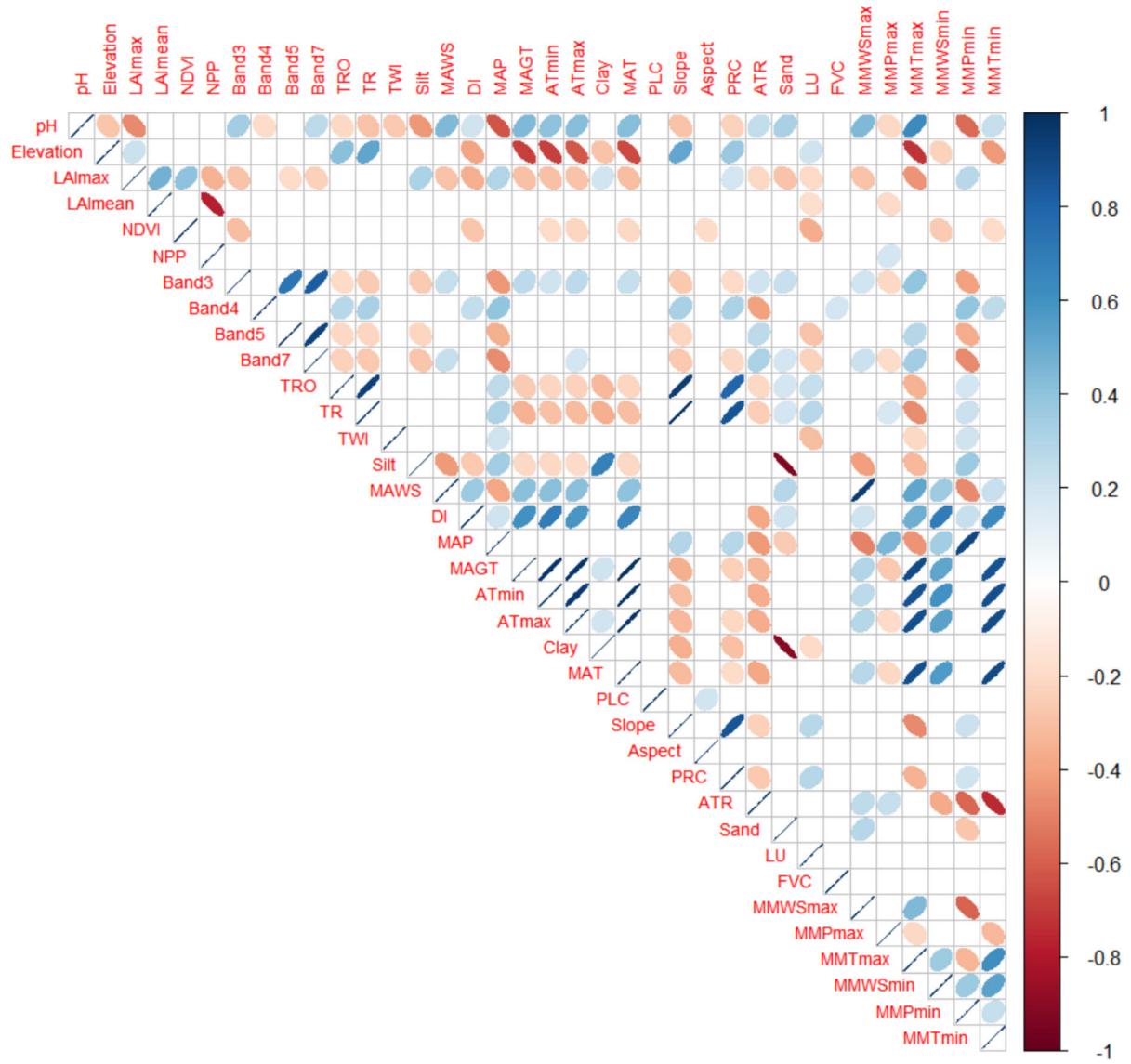

Fig. 3 Heatmap of the correlation between surface soil pH and environmental covariates

Silt, with coefficients of −0.625, −0.560, −0.471, and −0.431 respectively, also indicating a moderate degree of correlation. Additionally, it is worth noting that the pH level had no significant correlation with LAImean, NDVI, NPP, Band5, Clay, PLC, Aspect, LU, FVC, and MMWSmin ($P > 0.05$). Consequently, these variables were excluded from further consideration.

The remaining 25 environmental covariates were subjected to RFE for feature subset optimization. The results (Fig. 4) showed that when the number of environmental covariates was reduced to 5, the 10-fold cross-validation R^2 reached its maximum value of 0.709, while the R^2 for other numbers of environmental covariates consistently below 0.700. Consequently, the final input covariates were determined to be 5. The corresponding optimal subset is displayed in Fig. 5, including MMTmax, MMPmin, MAP, DI, and MMWSmax, with their %IncMSE values being 0.549, 0.364, 0.277, 0.263, and 0.257 respectively, IncNodePurity values being 35.290, 28.973, 26.989, 25.196, and 18.794 respectively. These indicators demonstrated their decisive role in the modeling process.

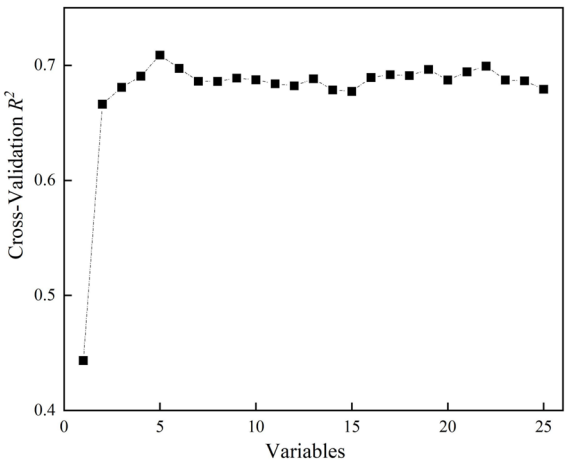

Fig. 4 Coefficient of determination changes with the number of environmental covariates in RFE process

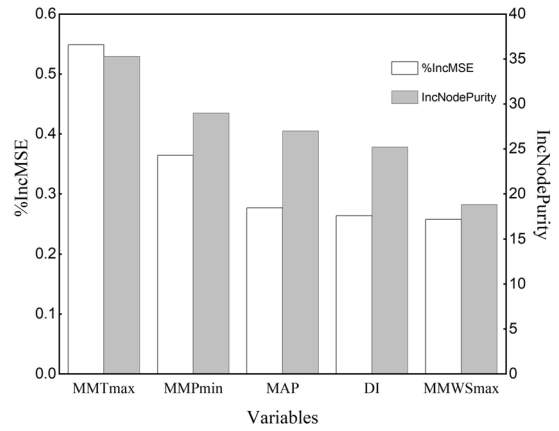

Fig. 5 The relative importance indicators of covariates in the optimal subset

In this study, the covariates selected in the optimal subsets were exclusively climate variables, indicating that in a vast area like Heilongjiang Province—with a total area of 470,700 square kilometers and relatively small elevation variations (a height difference of 1,690 ms)—terrain factors, especially PLC and Aspect, contribute less to the prediction of soil properties. This finding is consistent with the research of Scarpone et al. (2016), Chen et al. (2019), and Zhang et al. (2021). Furthermore, due to the flat terrain, the farming altitude in Heilongjiang Province is generally below 800 ms (Liu et al., 2022c). In recent years, the intensity of agriculture in Heilongjiang Province cultivation area has been gradually increasing, causing changes in vegetation and soil factors to be more regulated by human factors. However, in the Greater Khingan Range area, where human activities are less disruptive and the region boasts high forest coverage, resulting in greater variation in pH values among homogeneous soils at different elevations. Thus, biological factors and soil factors do not adequately reflect the spatial differentiation characteristics of surface soil pH in the study area.

MMTmax reflects the impact on soil microbial activities and the rate of organic matter decomposition, MMPmin and MAP reflect the promotion of soil chemical weathering and leaching processes, DI reflects the balance between precipitation and evaporation, and MMWSmax may increase wind erosion and promote soil water evaporation. Collectively, climate variables can better reveal the complex nonlinear relationship between surface soil pH and environmental covariates during the mapping process. This is consistent with the experience of Luo et al. (2024) in mapping soil properties in the black soil region of Northeast China.

3.3 Comparison of different models

Spatial modeling of surface soil pH using MMTmax, MMPmin, MAP, DI, and MMWSmax, the prediction performance of different models is detailed in Table 2. The results showed that the prediction performance of the three models was very similar, with XGBoost exhibiting the highest prediction accuracy ($R^2 = 0.705$, RMSE = 0.633, MAE = 0.484), followed by RF ($R^2 = 0.688$, RMSE = 0.656, MAE = 0.497), and the lowest prediction accuracy was SVM ($R^2 = 0.681$, RMSE = 0.656, MAE = 0.480).

The R^2 values indicate that XGBoost explains approximately 70.5% of the variability in surface soil pH, suggesting it captures the underlying patterns in the data more effectively than RF and SVM. The RMSE and MAE values further support this, as XGBoost has the lowest errors, indicating better precision in its pre-

Table 2 Prediction performance of different models

Model	R^2	RMSE	MAE
SVM	0.681	0.656	0.480
RF	0.688	0.656	0.497
XGBoost	0.705	0.633	0.484

dictions. The relatively small differences in RMSE and MAE between the models suggest that all three models perform reasonably well, but XGBoost's slightly lower errors highlight its superior predictive capability.

From the case of 10-fold cross-validation (Fig. 6), the R^2 trend for RF and XGBoost was essentially the same, both reaching their lowest at the third fold (R^2 being 0.507 and 0.542 respectively), but still above 0.500, indicating that the RF and XGBoost models had good robustness in predicting surface soil pH in the study area. In contrast, SVM achieved the highest R^2 at the ninth fold ($R^2 = 0.903$) and the lowest at the third fold ($R^2 = 0.146$), with a R^2 range as high as 0.757. This indicated that SVM exhibited significant volatility in 10-fold cross-validation, suggesting a potential risk of overfitting or underfitting when predicting results. The high variability in SVM's performance may be due to its sensitivity to the choice of kernel and hyperparameters, which can lead to overfitting on specific folds or underfitting when the model fails to capture the underlying data structure. This volatility underscores the importance of careful parameter tuning and model validation when using SVM for spatial predictions.

Overall, the high accuracy of XGBoost can be explained by its ability to iteratively correct errors from previous models, leading to improved predictions with each iteration. This structural advantage allows XGBoost to maintain high accuracy and stability, even when faced with varying subsets of data during cross-validation. The consistent performance and lower error metrics of XGBoost make it the most reliable model for spatial modeling of surface soil pH in this study.

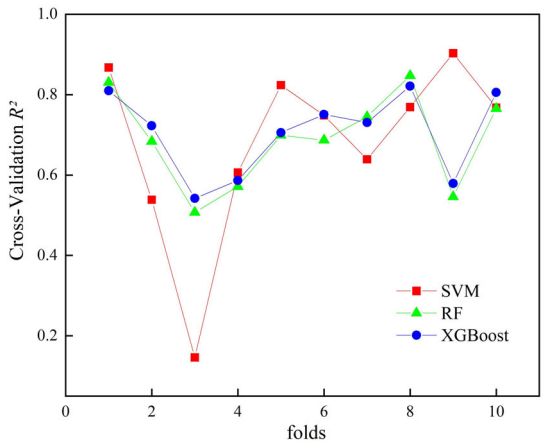

Fig. 6 Coefficient of determination of 10-fold cross-validation for each model

3.4 Comparison of grading maps

The surface soil pH prediction results from the three models were graded with a gradient of 0.5 for mapping (Fig. 7). It can be observed that the predictions of RF and XGBoost exhibited very similar spatial distribution patterns. They both predicted high pH values in the southwestern and central-eastern parts of the study area, with these high values decreasing towards the surrounding areas. The areas with low pH values were primarily concentrated in the Greater Khingan Range region and the central mountainous areas with an average altitude above 800 ms. In contrast, the SVM prediction map showed significant differences from those of RF and XGBoost, particularly in predicting the decreasing pH trend in the southwestern area and identifying the high pH value areas in the central-eastern part. This suggested that SVM demonstrated certain limitations in capturing the complex nonlinear relationships between soil properties and their environmental covariates in this region.

Statistically, the mean of SVM predictions was 6.402, with a standard deviation of 0.841. RF predictions had a mean of 6.459 and a standard deviation of 0.779. XGBoost predictions had a mean of 6.420 and a standard deviation of 0.774. These statistical figures further substantiated that RF and XGBoost demonstrated greater consistency in predicting surface soil pH in the study area.

3.5 Uncertainty assessment and comparison

Based on the comparison results of prediction performance, we conducted further uncertainty assessment and comparison for the superior models RF and XGBoost. We employed a strategy for uncertainty assessment that involved running the prediction model 100 times and recording the prediction results of each iteration in detail. We then used the variance of all predictions to create an uncertainty map. According to the Bootstrap sampling principle, prediction errors are generated randomly. Through multiple iterations, this ran-

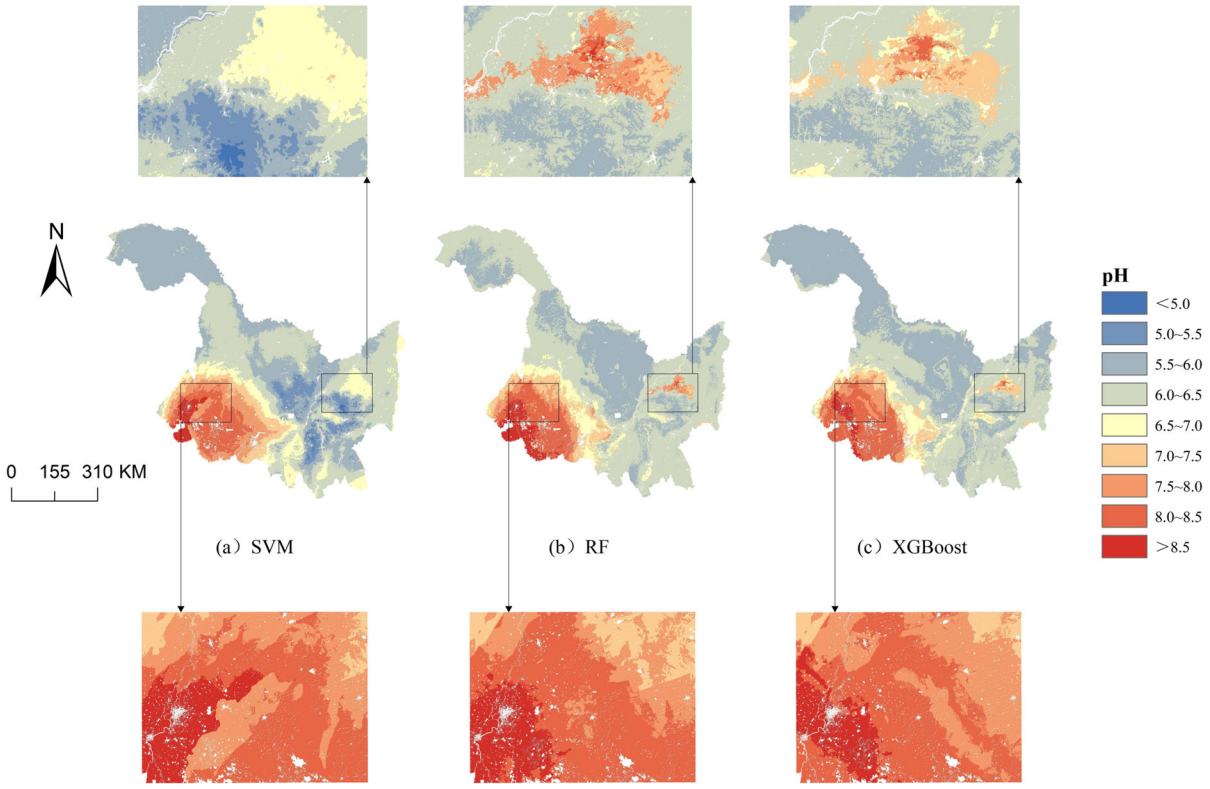

Fig. 7 Grading map of surface soil pH for each model

domness can be captured, enabling a relatively precise estimation of uncertainty.

The uncertainty of the two models was assessed and compared in Fig. 8. The results showed that both RF and XGBoost exhibited lower uncertainty in mountain forest areas with average elevations above 800 ms (e.g., the Greater Khingan Range area, Heihe City, Hegang City, Jiamusi City, Jixi City, Mudanjiang City). These regions are characterized by relatively stable environmental conditions, with minimal human disturbance

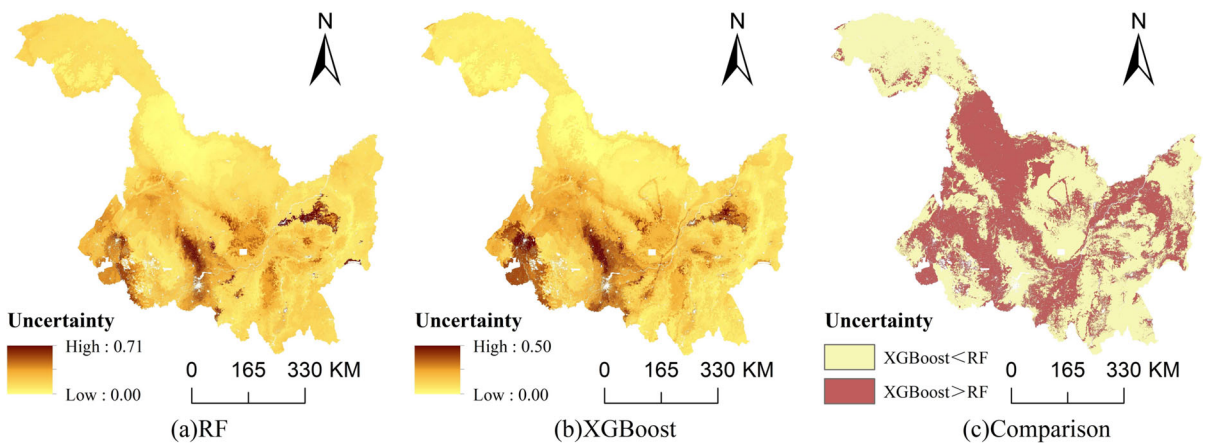

Fig. 8 Uncertainty comparison of the prediction results for RF and XGBoost models

and homogeneous land use, such as dense forests and natural vegetation. The soil in these areas tends to be less affected by external factors, resulting in lower natural variation in soil properties, including pH. As a result, even with a limited number of soil samples, the models were able to produce consistent and reliable predictions. The stability of the environmental conditions in these high-elevation areas contributes to the lower uncertainty observed in the model predictions.

In contrast, uncertainty was relatively higher in lowland plains where human activities are more frequent (e.g., the central-southern region of Qiqihar City, the western region of Daqing City, the central region of Suihua City, the western region of Harbin City, the northwestern region of Shuangyashan City). These areas are predominantly agricultural or urban, with significant human interventions such as farming practices, industrial activities, and urban development. Agricultural cultivation, in particular, involves the application of fertilizers, pesticides, and irrigation, which can significantly alter soil pH and other properties. Industrial emissions and urban construction further contribute to soil variability by introducing pollutants and changing land use patterns. These frequent and diverse human disturbances lead to greater soil heterogeneity, making it more challenging for the models to predict soil pH accurately. Consequently, the higher variability in soil properties in these lowland plains results in increased uncertainty in the model predictions.

Therefore, we concluded that the uncertainty of DSM in this study was primarily attributed to soil variation. In the mountain forest areas, the natural stability of the environment and minimal human impact result in lower soil variability, allowing the models to achieve more consistent predictions even with fewer samples. In contrast, the lowland plains experience significant human-induced changes, leading to greater soil heterogeneity and higher uncertainty in predictions. This highlights the importance of considering land use and human activities when assessing soil variability and model performance. In future studies, it is recommended to increase the density of sample point layouts in these lowland plain areas to better capture the spatial variability of soil properties.

From the comparison results (Fig. 8c), XGBoost predominantly exhibited lower uncertainty in high-elevation mountain forest areas, while RF demonstrated higher prediction consistency primarily in the lowland plains. Each prediction model had its unique advantages in different terrain areas. Overall, the area where XGBoost had less uncertainty than RF accounted for approximately 59.65%, and the area where uncertainty was greater in XGBoost than in RF accounted for about 40.35%. Under the specific conditions of this study, XGBoost emerged as the optimal model for predicting the spatial distribution of surface soil pH in Heilongjiang Province.

3.6 Model interpretability

To further explore the relationship between environmental covariates and surface soil pH, we utilized the XGBoost model to generate accumulated local effects (ALE) plots (Fig. 9). These plots offered a visual representation of the impact each predictor has on surface soil pH predictions. We can observe that high pH values were mainly influenced by high MMTmax values, high MMWSmax values, and low MAP values, and vice versa. This is largely consistent with the results of the Pearson correlation analysis. The ALE plots also showed MMPmin exhibited a trend of fluctuating changes, suggesting a more complex relationship with pH. Furthermore, the effects of DI, MMWSmax, and MMTmax on pH predictions were observed to surge sharply after surpassing a specific threshold. This indicated that the influence of these covariates on soil pH was significantly amplified once they exceeded a certain level, highlighting their critical role in determining soil pH dynamics.

4 Discussion

4.1 Spatiotemporal characteristics and causes of soil acidification

The study by Wu et al. (2018) indicated that the soil pH in the black soil region of Northeast China has remained relatively stable since around the year 2010. Therefore, the results of this study can be considered representative of the current spatial distribution characteristics of surface soil pH in Heilongjiang Province. According to the optimal prediction results, the typical black soil in

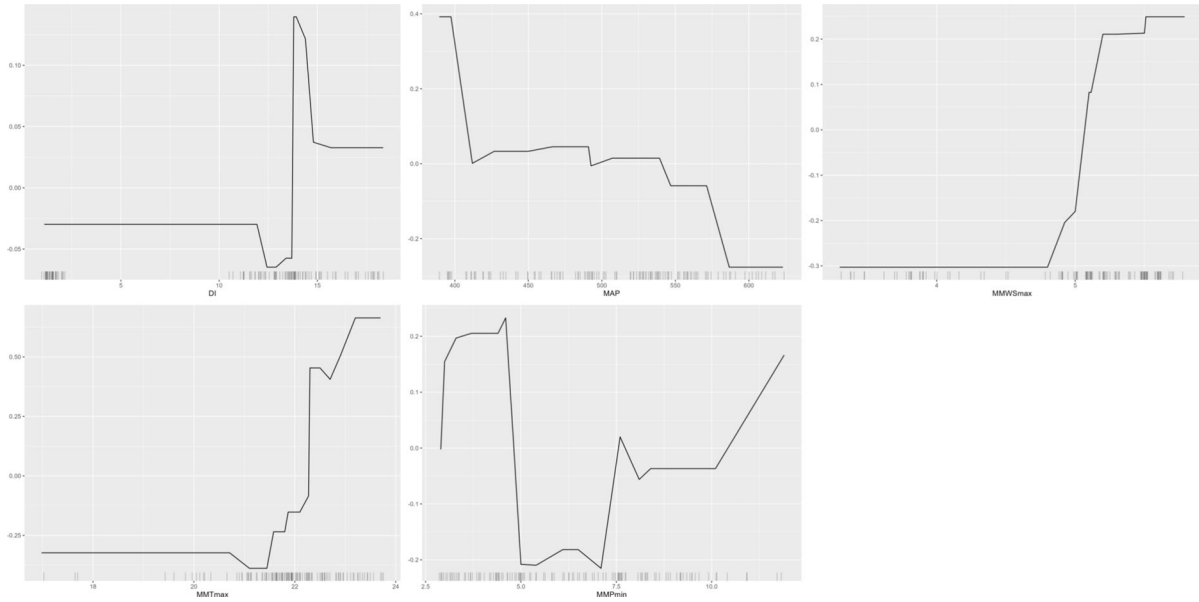

Fig. 9 Accumulated local effect (ALE) plots using the XGBoost model for model interpretability

Heilongjiang Province is generally weakly acidic, with an average pH of 6.42. The pH values increase gradually from east to west and from north to south in space, which is largely consistent with the findings of Liu et al. (2019). When compared with historical data (Wu et al., 2018)—the average surface soil pH from 1988 to 1991 was 7.08, and from 2002 to 2006 was 6.58—the surface soil pH in Heilongjiang Province has decreased by 0.66 and 0.16 units respectively, with an average decrease of 0.13 units every 5 years. Particularly in the meadow black soil and albic black soil regions of eastern and northeastern Heilongjiang Province, varying degrees of soil acidification have repeatedly occurred. In these regions, the specific processes contributing to acidification include the leaching of base cations (e.g., calcium, magnesium) due to high rainfall and the accumulation of acidic cations (e.g., aluminum, hydrogen) in the soil profile. These regions are characterized by concentrated heavy precipitation and more intensive agricultural activities, which accelerate the leaching of nutrients and the accumulation of acidic substances. Additionally, the historical overuse of nitrogen fertilizers in these areas has further exacerbated the acidification process.

Fundamentally, the excessive application of nitrogen fertilizer is the key factor leading to soil acidification (Ju & Gu, 2014). When ammonium-based nitrogen fertilizers are applied to the soil, nitrification reactions occur, producing H^+ which causes acidification. Moreover, if NO_3^- is leached away by rainfall, all the H^+ will remain in the soil, ultimately leading to increased acidification. However, intensive agricultural practices (e.g., continuous monocropping and insufficient crop rotation) also contribute to soil acidification. It can exacerbate soil acidification by depleting soil organic matter and reducing its buffering capacity. Organic matter plays a crucial role in maintaining soil pH by providing cation exchange capacity and buffering against acidification. In regions with low organic matter content, soils are more susceptible to acidification due to reduced buffering capacity.

The ecological impacts of soil acidification are profound. After soil acidification occurs, a large amount of acid accumulates in the solid phase layer of the soil, and the demand for lime and other alkaline substances needed to neutralize this acid also increases correspondingly, making improvement extremely challenging (Zhao et al., 2023). Therefore, a reasonable adjustment of nitrogen fertilizer application strategies, combined with the practice of returning straw to the field to reduce excessive nitrogen fertilizer use, is a key measure to control soil acidification in the black soil region of Northeast China. Additionally, adopting sustainable land management practices, such as crop

rotation, cover cropping, and the application of organic amendments, can help improve soil organic matter content and enhance soil buffering capacity. The implementation in specific areas can be guided by the prediction results and uncertainty maps from this study.

4.2 Limitations and future research

Three advanced ML models—SVM, RF, and XGBoost—were combined with soil environmental covariates optimization based on Pearson correlation analysis and RFE, to predict the spatial distribution of surface soil pH in Heilongjiang Province. Additionally, we explored the uncertainty of the prediction results from each model and identified areas where sampling density needs to be increased in future surveys. Overall, XGBoost was considered the best model for predicting surface soil pH in the study area, and the experience of this study can provide some guidance for subsequent mapping research in the black soil region of Northeast China.

In terms of environmental covariate selection, this study primarily chose terrain, climate, and biological factor data, which are easily accessible. The results showed that terrain factors and biological factors contributed less to the prediction of surface soil pH in Heilongjiang Province, while climate factors contributed more significantly. The inclusion of multi-year climate average condition indicators was the key to achieving high prediction accuracy in this study. However, it is worth noting that other soil factors (e.g., organic matter content, soil texture) were not included in this study. These factors could potentially provide additional insights into soil pH variability, especially in regions with complex land use patterns or diverse soil types. In future mapping explorations, a combination of multiple climate factor variables along with soil property data can be considered to improve the accuracy of DSM in large and flat areas.

Regarding model selection, while XGBoost demonstrated excellent prediction performance in this study, each model has its inherent limitations. For instance, SVM, despite its ability to handle high-dimensional data, is computationally intensive and less interpretable, making it challenging to scale to larger datasets or different regions. RF, on the other hand, is more interpretable and robust to overfitting but can be computationally expensive when dealing with large datasets. XGBoost, while efficient and accurate, may struggle with scalability in very large datasets due to its iterative nature. Future research could experiment with more types of ML models, such as generalized additive models (Mosleh et al., 2016; Xu et al., 2017), which offer flexibility in modeling non-linear relationships, or deep learning models (Amirian Chakan et al., 2017; Padarian et al., 2019; Yang et al., 2021), which can capture complex patterns in large datasets. Additionally, other tree-based models (He et al., 2024; Salmanpour et al., 2023; Sharififar, 2022) could be explored to further understand their adaptability in predicting the spatial distribution of surface soil pH. A comparison of model performance under different conditions, including varying sample sizes and spatial distributions, could provide deeper insights into the adaptability and limitations of each model.

On the mapping strategy, the method of partition modeling has been preliminarily explored in DSM (Luo et al., 2024), and the results showed that fully considering the differences between different regions can significantly improve mapping accuracy. Based on the relationship between the uncertainty assessment results and the terrain, we had concluded that RF and XGBoost models had their respective advantages in different terrain areas. Therefore, a method of spatial modeling separately for each terrain zone can be considered, which is expected to further improve mapping accuracy. This approach could be particularly useful in regions with significant environmental heterogeneity, where a single model may not capture the full range of soil pH variability.

5 Conclusions

This study demonstrates that using ML models combined with a large amount of environmental covariates for feature selection can attain high-precision prediction maps, even in large and flat areas with sparse samples. In the mapping process, multi-year climate average condition indicators were found to better reveal the nonlinear relationships between soil properties and the environment than terrain, parent material, biological, and soil factors. The resulting surface soil pH grading map and uncertainty map enable users to quickly understand soil conditions and devise precise soil management strategies. Future research could focus on conducting additional surveys in areas with high prediction

uncertainty and adopting a terrain-zone-based modeling approach to more comprehensively reflect spatial variations. The findings of this study are applicable to mapping large and flat areas within black soil regions.

Author contributions Conceptualization, Shi Y.; methodology, Huang Q.; software, Huang P.; validation, Wang J.; formal analysis, Huang P.; investigation, Huang P.; resources, Huang Q.; data curation, Huang P.; writing—original draft preparation, Huang P.; writing—review and editing, Huang P.; visualization, Huang P.; supervision, Huang Q.; project administration, Huang Q.; funding acquisition, Huang Q.

Funding Information This research was funded by the National Key Research and Development Program of China with grant number 2023YFD1500102 and the Fundamental Research Funds for Central Non-profit Scientific Institution with grant number 1610132024007.

Data availability No datasets were generated or analyzed during the current study.

Declarations

Ethical approval All authors have read, understood, and have complied as applicable with the statement on "Ethical responsibilities of Authors" as found in the Instructions for Authors.

Consent to participate Not applicable

Consent for publication Not applicable

Competing interests The authors declare no competing interests.

References

Amirian Chakan, A., Taghizadeh-Mehrjardi, R., Kerry, R., Kumar, S., Khordehbin, S., & Yusefi Khanghah, S. (2017). Spatial 3d distribution of soil organic carbon under different land use types. *Environmental monitoring and assessment, 189*, 1–16. https://doi.org/10.1007/s10661-017-5830-9

Böhner, J. & Antonić, O. (2009). Chapter 8 land-surface parameters specific to topo-climatology. *Developments in soil science, 33*, 195–226. https://doi.org/10.1016/S0166-2481(08)00008-1

Breiman, L. (2001). *Random forests. Machine learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

Camera, C., Zomeni, Z., Noller, J. S., Zissimos, A. M., Christoforou, I. C., & Bruggeman, A. (2017). A high resolution map of soil types and physical properties for Cyprus: A digital soil mapping optimization. *Geoderma, 285*, 35–49. https://doi.org/10.1016/j.geoderma.2016.09.019

Chen, S., Mulder, V. L., Martin, M. P., Walter, C., Lacoste, M., Richer-de Forges, A. C., Saby, N. P., Loiseau, T., Hu, B., & Arrouays, D. (2019). Probability mapping of soil thickness by random survival forest at a national scale. *Geoderma, 344*, 184–194. https://doi.org/10.1016/j.geoderma.2019.03.016

Demarchi, L., Kania, A., Ciężkowski, W., Piórkowski, H., Oświecimska-Piasko, Z., & Chormański, J. (2020). Recursive feature elimination and random forest classification of natura 2000 grasslands in lowland river valleys of Poland based on airborne hyperspectral and lidar data fusion. *Remote Sensing, 12*(11), 1842. https://doi.org/10.3390/rs12111842

Guo, J., Liu, F., Xu, S., Gao, L., Zhao, Z., Hu, W., Yu, D., & Zhao, Y. (2024). Comparison of digital mapping methods for the thickness of black soil layer of cultivated land in typical black soil area of Songnen Plain. *Journal of Geo-information Science, 26*, 1452–1468. https://doi.org/10.12082/dqxxkx.2024.230682

He, W., Xiao, Z., Lu, Q., Wei, L., & Liu, X. (2024). Digital mapping of soil particle size fractions in the loess plateau, China, using environmental variables and multivariate random forest. *Remote Sensing, 16*(5), 785. https://doi.org/10.3390/rs16050785

Henderson, B. L., Bui, E. N., Moran, C. J., & Simon, D. (2005). Australia-wide predictions of soil properties using decision trees. *Geoderma, 124*(3–4), 383–398. https://doi.org/10.1016/j.geoderma.2004.06.007

Heung, B., Ho, H. C., Zhang, J., Knudby, A., Bulmer, C. E., & Schmidt, M. G. (2016). An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. *Geoderma, 265*, 62–77. https://doi.org/10.1016/j.geoderma.2015.11.014

Hu, W., Tao, T., Tian, K., Zhao, Y., Huang, B., & Luo, Y. (2021). Status and prospect of farmland soil environmental quality management in China. *Acta Pedologica Sinica, 58*(5), 1094–1109. https://doi.org/10.11766/trxb202009220533

Huggett, R. J. (1975). Soil landscape systems: A model of soil genesis. *Geoderma, 13*(1), 1–22. https://doi.org/10.1016/0016-7061(75)90035-X

Ju, X., & Gu, B. (2014). Status-quo, problem and trend of nitrogen fertilization in China. *Journal of Plant Nutrition and Fertilizers, 20*(4), 783–795. https://doi.org/10.11674/zwyf.2014.0401

Kaya, F., Keshavarzi, A., Francaviglia, R., Kaplan, G., Başayiğit, L., & Dedeoğlu, M. (2022). Assessing machine learning-based prediction under different agricultural practices for

digital mapping of soil organic carbon and available phosphorus. *Agriculture, 12*(7), 1062. https://doi.org/10.3390/agriculture12071062

Kuhn, M. (2015). Caret: Classification and regression training. *Astrophysics Source Code Library* (pp. ascl–1505). https://doi.org/10.32614/CRAN.package.caret

Lal, R. (2015). Restoring soil quality to mitigate soil degradation. *Sustainability, 7*(5), 5875–5895. https://doi.org/10.3390/su7055875

Li, Y., Zhang, J., Jia, J., Fan, F., Zhang, F., & Zhang, J. (2022). Research progresses on farmland soil ecosystem multifunctionality. *Acta Pedologica Sinica, 59*(5), 1177–1189. https://doi.org/10.11766/trxb202109290532

Liaw, A., & Wiener, M. (2015). randomforest: Breiman and cutler's random forests for classification and regression. *R package version, 4*(1.1). https://doi.org/10.32614/CRAN.package.randomForest

Liu, F., Wu, H., Zhao, Y., Li, D., Yang, J., Song, X., Shi, Z., Zhu, A., & Zhang, G. (2022a). Mapping high resolution national soil information grids of China. *Science Bulletin, 67*(3), 328–340. https://doi.org/10.1016/j.scib.2021.10.013

Liu, F., Yang, F., Zhao, Y., Zhang, G., & Li, D. (2022b). Predicting soil depth in a large and complex area using machine learning and environmental correlations. *Journal of Integrative Agriculture, 21*(8), 2422–2434. https://doi.org/10.1016/S2095-3119(21)63692-4

Liu, Y., Pei, J., & Wang, J. (2019). Spatial distribution and relationship between organic matter and pH in the typical black soil region of northeast China. *Journal of Agricultural Resources and Environment, 36*, 738–743. https://doi.org/10.13254/j.jare.2018.0292

Liu, Y., Wu, K., Li, X., & Li, X. (2022c). Classification of land types at provincial level based on the goal of black land protection: A case study of Heilongjiang Province. *Scientia Geographica Sinica, 42*, 1348–1359. https://doi.org/10.13249/j.cnki.sgs.2022.08.003

Luo, C., Zhang, W., Meng, X., Yu, Y., Zhang, X., & Liu, H. (2024). Mapping the soil organic matter content in northeast China considering the difference between dry lands and paddy fields. *Soil and Tillage Research, 244*. https://doi.org/10.1016/j.still.2024.106270

Mosleh, Z., Salehi, M. H., Jafari, A., Borujeni, I. E., & Mehnatkesh, A. (2016). The effectiveness of digital soil mapping to predict soil properties over low-relief areas. *Environmental monitoring and assessment, 188*, 1–13. https://doi.org/10.1007/s10661-016-5204-8

Nguyen, T. T., Pham, T. D., Nguyen, C. T., Delfos, J., Archibald, R., Dang, K. B., Hoang, N. B., Guo, W., & Ngo, H. H. (2022). A novel intelligence approach based active and ensemble learning for agricultural soil organic carbon prediction using multispectral and SAR data fusion. *Science of The Total Environment, 804*. https://doi.org/10.1016/j.scitotenv.2021.150187

Padarian, J., Minasny, B., & McBratney, A. B. (2019). Using deep learning for digital soil mapping. *Soil, 5*(1), 79–89. https://doi.org/10.5194/soil-5-79-2019

Petermann, E., Meyer, H., Nussbaum, M., & Bossew, P. (2021). Mapping the geogenic radon potential for Germany by machine learning. *Science of The Total Environment, 754*. https://doi.org/10.1016/j.scitotenv.2020.142291

Salmanpour, A., Jamshidi, M., Fatehi, S., Ghanbarpouri, M., & Mirzavand, J. (2023). Assessment of macronutrients status using digital soil mapping techniques: a case study in Maru'ak area in Lorestan Province, Iran. *Environmental monitoring and assessment, 195*(4), 513. https://doi.org/10.1007/S10661-023-11145-5

Scarpone, C., Schmidt, M. G., Bulmer, C. E., & Knudby, A. (2016). Modelling soil thickness in the critical zone for Southern British Columbia. *Geoderma, 282*, 59–69. https://doi.org/10.1016/j.geoderma.2016.07.012

Sharififar, A. (2022). Accuracy and uncertainty of geostatistical models versus machine learning for digital mapping of soil calcium and potassium. *Environmental monitoring and assessment, 194*(10), 760. https://doi.org/10.1007/S10661-022-10434-9

Shen, R. (2018). Development, status and prospect of soil science. *Journal of Agriculture, 8*(1), 53–58. https://doi.org/10.11923/j.issn.2095-4050.cjas2018-1-053

Smith, P., House, J. I., Bustamante, M., Sobocká, J., Harper, R., Pan, G., West, P. C., Clark, J. M., Adhya, T., Rumpel, C., et al. (2016). Global change pressures on soils from land use and management. *Global change biology, 22*(3), 1008–1028. https://doi.org/10.1111/gcb.13068

Suleymanov, A., Gabbasova, I., Komissarov, M., Suleymanov, R., Garipov, T., Tuktarova, I., & Belan, L. (2023). Random forest modeling of soil properties in saline semi-arid areas. *Agriculture, 13*(5), 976. https://doi.org/10.3390/agriculture13050976

Sun, F., Lei, Q., Liu, Y., Li, H., & Wang, Q. (2011). The progress and prospect of digital soil mapping research. *Chinese Journal of Soil Science, 42*, 1502–1507. https://doi.org/10.19336/j.cnki.trtb.2011.06.041

Vapnik, V. (2013). *The nature of statistical learning theory*. New York: Springer science & business media.

Wei, T., Simko, V., Levy, M., Xie, Y., Jin, Y., Zemla, J., et al. (2013). corrplot: Visualization of a correlation matrix. *R package version 0.73, 230*(11), 1–26. https://doi.org/10.32614/CRAN.package.corrplot

Wiesmeier, M., Barthold, F., Blank, B., & Kögel-Knabner, I. (2011). Digital mapping of soil organic matter stocks using random forest modeling in a semi-arid steppe ecosystem. *Plant and soil, 340*, 7–24. https://doi.org/10.1007/s11104-010-0425-z

Wong, T.-T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern recognition, 48*(9), 2839–2846. https://doi.org/10.1016/j.patcog.2015.03.009

Wu, H., Wang, S., Huai, S., Yan, Z., Ma, C., Xue, Y., Xu, M., & Lu, C. (2018). Evolutionary characteristics of fertility and productivity of typical black soil in recent 30 years. *Journal of Plant Nutrition and Fertilizers, 24*(6), 1456–1464. https://doi.org/10.11674/zwyf.18238

Xu, Y., Smith, S. E., Grunwald, S., Abd-Elrahman, A., Wani, S. P., & Nair, V. D. (2017). Spatial downscaling of soil prediction models based on weighted generalized additive models in smallholder farm settings. *Environmental monitoring and assessment, 189*, 1–16. https://doi.org/10.1007/s10661-017-6212-z

Yang, L., Cai, Y., Zhang, L., Guo, M., Li, A., & Zhou, C. (2021). A deep learning method to predict soil organic carbon content at a regional scale using satellite-based phenology vari-

ables. *International Journal of Applied Earth Observation and Geoinformation, 102*, 102428. https://doi.org/10.1016/j.jag.2021.102428

Zhai, R., Xin, G., & Zhang, Z. (2020). *Soil series of China: Heilongjiang volume*. Beijing: Science Press.

Zhang, G., & Gong, Z. (2012). *Soil survey laboratory methods*. Beijing: Science Press.

Zhang, G., Liu, F., & Song, X. (2017). Recent progress and future prospect of digital soil mapping: A review. *Journal of integrative agriculture, 16*(12), 2871–2885. https://doi.org/10.1016/S2095-3119(17)61762-3

Zhang, G., Shi, Z., Zhu, A., Wang, Q., Wu, K., Shi, Z., Zhao, Y., Zhao, Y., Pan, X., Liu, F., & Song, X. (2020). Progress and perspective of studies on soils in space and time. *Acta Pedologica Sinica, 57*(5), 1060–1070. https://doi.org/10.11766/trxb202004270199

Zhang, S., Liu, G., Chen, S., Rasmussen, C., & Liu, B. (2021). Assessing soil thickness in a black soil watershed in northeast China using random forest and field observations. *International Soil and Water Conservation Research, 9*(1), 49–57. https://doi.org/10.1016/j.iswcr.2020.09.004

Zhang, W., Wan, H., Zhou, M., Wu, W., & Liu, H. (2022). Soil total and organic carbon mapping and uncertainty analysis using machine learning techniques. *Ecological Indicators, 143*, 109420. https://doi.org/10.1016/j.ecolind.2022.109420

Zhao, X., Pan, X., Ma, H., Dong, X., Che, J., Wang, C., Shi, Y., Liu, K., & Shen, R. (2023). Scientific issues and strategies of acid soil use in China. *Acta Pedologica Sinica, 60*(5), 1248–1263. https://doi.org/10.11766/trxb202307250290

Zhu, Y., Li, G., Zhang, G., & Fu, B. (2015). Soil security: From earth's critical zone to ecosystem services. *Acta Geographica Sinica, 70*(12), 1859–1869. https://doi.org/10.11821/dlxb201512001