Spatial prediction of soil sand content at various sampling density based on geostatistical and machine learning algorithms in plain areas

Catena

Qu, Lili; Lu, Huizhong; Tian, Zhiyuan; Schoorl, J.M.; Huang, Biao et al

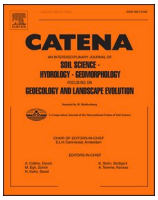
https://doi.org/10.1016/j.catena.2023.107572

Catena 234 (2024) 107572

Contents lists available at ScienceDirect

Catena

journal homepage: www.elsevier.com/locate/catena



Spatial prediction of soil sand content at various sampling density based on geostatistical and machine learning algorithms in plain areas

Lili Qu [a,b], Huizhong Lu [c], Zhiyuan Tian [a,*], J.M. Schoorl [d], Biao Huang [a], Yonghong Liang [e], Dan Qiu [e], Yin Liang [a]

[a] State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, Nanjing 210008, China
[b] College of Advanced Agricultural Sciences, University of Chinese Academy of Sciences, Beijing 100049, China
[c] National Key Laboratory of Water Disaster Prevention, Nanjing Hydraulic Research Institute, Nanjing 210029, China
[d] Soil Geography and Landscape, Wageningen University, P.O. Box 47, NL-6700 AA Wageningen, the Netherlands
[e] Jiangsu Province Station of Farmland Quality and Agro-environment Protection, Nanjing 210029, China

ARTICLE INFO

Keywords:
Digital soil mapping
Sampling density
Model comparison
Geographical detector
Soil sand content
Uncertainty assessment

ABSTRACT

Accurate prediction of the spatial distribution of soil sand content is a pre-requisite for land use management, soil quality evaluation and erosion control, as it determines the transport and movement of soil water, fertilizer, air and heat. Digital soil mapping (DSM) is extensively employed for predicting soil properties. However, practical research is required to address the challenge of selecting an optimal prediction model that is both cost-effective and accurate at a specific sampling density. In this study, topsoil samples were collected from 2,848 sampling points in the eastern plains of China (107,200 km^2). The performance of different prediction models for mapping soil sand content was compared at 12 levels of sampling density. Moreover, the geographical detector, a statistical method used to assess the spatial stratified heterogeneity of variables, was adopted to determine the major drivers of spatial variation in soil sand content. The results indicated that climate factors are the major drivers of the spatial variability in soil sand content. For the 100% sample size (26.57 samples/10^3 km^2), the geostatistical models that did not depend on environmental variables (ordinary kriging, sequential Gaussian simulation) performed best, followed by the machine learning models (random forest, cubist and support vector machine) and the geostatistical model with environmental variables (co-kriging). Sampling density had a considerable impact on model accuracy, and the advantages of machine learning models became apparent when sampling densities were below 20% (5.31 samples/10^3 km^2). Therefore, the best combination of prediction model and sampling density should be selected to obtain maps of soil sand content economically and accurately. This study provides a valuable reference for the selection of prediction methods in the practical application of DSM.

1. Introduction

Soil texture is the fundamental physical property of soil and has an important influence on other soil properties (Hu et al., 2021; Tang et al., 2022). Sand particles, representing large soil particles, play a crucial role in the transport and movement of water, fertilizer, air, heat, and microorganisms. Soil sand content directly influences soil fertility and soil quality and ultimately affects crop yield (Laborczi et al., 2019; Gupta et al., 2022; Yageta et al., 2019). Soil sand content is influenced by multiple factors, including climate, parent material, topography, and anthropogenic activities. As a result, it demonstrates a distinct spatial distribution pattern with certain regularity and zonality (Chen et al., 2022; Lamichhane et al., 2019; Zhang et al., 2017). Therefore, accurate prediction of the spatial distribution of sand content and an understanding of its spatial distribution pattern and influencing factors contribute to precise regional land management, soil quality evaluation and soil erosion control.

Traditional soil mapping methods rely on expert knowledge and a

Abbreviations: DSM, digital soil mapping; OK, ordinary kriging; SGS, sequential Gaussian simulation; COK, co-kriging; RF, random forest; SVM, support vector machine; RMSE, root-mean squared error; R^2, coefficient of determination; MAE, mean absolute error.

* Corresponding author.
E-mail address: tianzhiyuan@issas.ac.cn (Z. Tian).

https://doi.org/10.1016/j.catena.2023.107572
Received 14 June 2023; Received in revised form 27 September 2023; Accepted 30 September 2023
Available online 3 October 2023

large amount of sampling data, making soil surveys very laborious, time-consuming and expensive (Liu et al., 2022; Sanderman et al., 2017). In recent years, digital soil mapping (DSM) technology has been recognized as an innovative and effective way of expressing the spatial distribution of soil properties. It has demonstrated significant potential for updating maps of soil properties (Martin et al., 2021; Wadoux et al., 2020; McBratney et al., 2003). DSM technology effectively achieves a pattern of inferring soil information based on the soil-forming environment, overcoming the data limitation of traditional mapping methods and allowing for more accurate and high-resolution soil maps (Liu et al., 2020). Importantly, the results of DSM are reproducible and able to quantify uncertainties (Arrouays et al., 2020).

DSM methods include linear models, geostatistical models, machine learning models and hybrid models (Minasny et al., 2013; Zhang et al., 2017). The most representative of these are geostatistical models and machine learning models. The geostatistical method is an optimal linear unbiased interpolation valuation method based on a semi-covariance function that fully accounts for the spatially varying characteristics of the variables (Madenoglu et al., 2020). Ordinary kriging (OK) is the most representative geostatistical model. It had been widely used in early spatial mapping of soil properties due to its operability, but the method relies on the spatial autocorrelation of predictor variables and disregards the correlation between soil characteristics and environmental factors (Ma et al., 2022). Its derivatives, such as regression kriging and co-kriging (COK), overcome this problem. The inclusion of auxiliary variables can lead to more accurate predictions (Wan et al., 2021). Machine learning models learn patterns from data to identify relationships between soil characteristics and environmental factors, which are then used to predict soil properties (Tian et al., 2022; Zhang et al., 2017). They include support vector machine (SVM), cubist, artificial neural network model, Bayesian model and random forest (RF). Machine learning models have become an increasingly mainstream approach for DSM because they can handle a variety of complex nonlinear problems with no requirement for data distribution and perform well over large areas (Zhang et al., 2017; Lamichhane et al., 2019; Wadoux et al., 2020; Poggio et al., 2021).

So far, it is not clear from the literature which type of model performs best for soil particle composition prediction. Zhao et al. (2022) compared the predictive performance of seven different models for soil clay content mapping in Australia, where the sampling density was 170 samples/10^3 km^2 and found that the optimal prediction model was RF. Beguin et al. (2017) compared eight different approaches in Canada where the sampling density was 0.17 samples/10^3 km^2 and concluded that Bayesian inference with geostatistical model exhibited the best prediction performance for soil sand content mapping. Silva et al. (2020) compared three different mapping methods in Brazil, where the sampling density was 83 samples/10^3 km^2 and found that SVM was the optimal model for predicting sand content and clay content. These studies showed that the predictive performance of mapping is highly dependent on the model, and the appropriate choice of model determines the efficiency of the mapping and the reliability of the results (Sun et al., 2022; Veronesi and Schillaci, 2019). In addition to predictive models, sampling density in DSM has a significant impact on the accuracy of soil mapping (John et al., 2022; Lai et al., 2021). However, the sampling densities in these studies varied considerably, and it is worth investigating whether the discrepancies in conclusions are related to the differences in sampling densities.

The sampling density is a key factor in determining the cost-effectiveness of mapping, and its influence on mapping accuracy should therefore be given considerable attention. Guo et al. (2018) compared the performance of a partial least squares regression model at sampling densities ranging from 20 to 626 samples/km^2, while Yang et al. (2020) evaluated the performance of a random forest at sampling densities ranging from 2.59 to 32.47 samples/km^2. Both studies revealed that higher sampling densities led to better model accuracy. Long et al. (2018) investigated the effect of sampling density ranging from 9 to 10 samples/km^2 on OK accuracy and concluded that the prediction accuracy was more sensitive to sampling density in flat areas. However, many studies have solely focused on the impact of sampling density on the accuracy of a single prediction model, neglecting the need for a comprehensive evaluation of the multidimensional integration of sampling density and prediction models. Further research is needed to determine the optimal combination of sampling density and prediction model for obtaining accurate and cost-effective spatial information on soil properties in a specific area.

The eastern plain of China, where Jiangsu Province is located, is one of the regions in China with the most extensive plain and cultivated land area (Qu et al., 2023). The region includes wide spread area with a high sand content due to the alluvial deposits of rivers and the deposition of the sea. The region is characterized by a loose soil structure and abundant rainfall, which makes it a high-risk area for soil erosion (Qu et al., 2022). Consequently, there is an urgent need for comprehensive data on the distribution of soil sand content within the area for agricultural development and soil conservation.

In this study, we compared the performance of two main types of DSM models (geostatistical models and machine learning models) in a large plain area, not only in terms of overall accuracy and local uncertainty but also in terms of sensitivity to sampling density. More specifically, we compared increasing sampling densities for soil sand content mapping with OK, sequential Gaussian simulation (SGS), COK, RF, Cubist and SVM. The aims of this study were to i) identify the driving factors of the spatial distribution of soil sand content; ii) explore the most suitable mapping models for sand content in plain areas; and iii) explore the impact of sampling density on the accuracy of mapping models.

2. Materials and methods

2.1. Study area

The study was carried out in Jiangsu Province (30°45′-35°08′N, 116°21′-121°56′E), situated in the eastern coast of China, with a total area of 107,200 km^2 (Fig. 1). The northern regions are mostly characterized by a warm temperate humid, semi-humid monsoon climate, whereas the southern regions are predominantly marked by a subtropical humid monsoon climate. The mean annual precipitation is 800–1,200 mm. Precipitation is concentrated in summer, accounting for half of the annual precipitation, and precipitation in winter is the lowest, accounting for approximately one-tenth of the annual precipitation. The mean annual temperature is 13–16 °C, with a decreasing distribution from south to north. Jiangsu is one of the provinces with the lowest elevation in China, with most of the area below 50 m above sea level. The low hills predominantly cluster in the southwest region, whereas the remaining approximately 85% of the area is composed of plains with elevations less than 20 m above sea level. Wheat, rice, and oilseed rape are the main crops. The most frequent soils are Inceptisols, Alfisols and Entisols according to the USDA classification (Xie et al., 2022). Sandy soils in the study area were widely distributed along the abandoned Yellow River in the north, the coastal area in the east and the alluvial area of the Yangtze River in the south (Qu et al., 2023).

2.2. Sampling and data acquisition

2.2.1. Soil data

A systematic soil survey of the study area was conducted through the 2017 Cultivated Land Quality Survey Project in Jiangsu Province. A grid system was used to collect a total of 2,848 topsoil samples (0–20 cm) in cultivated land (Fig. 1). The sampling points were situated in predominantly flat plains within the cultivated land. To obtain representative soil samples, a composite soil sample was created by collecting and thoroughly mixing five individual soil samples within a 10 m radius of each sampling location (Rawlins et al., 2009). The soil samples were air-

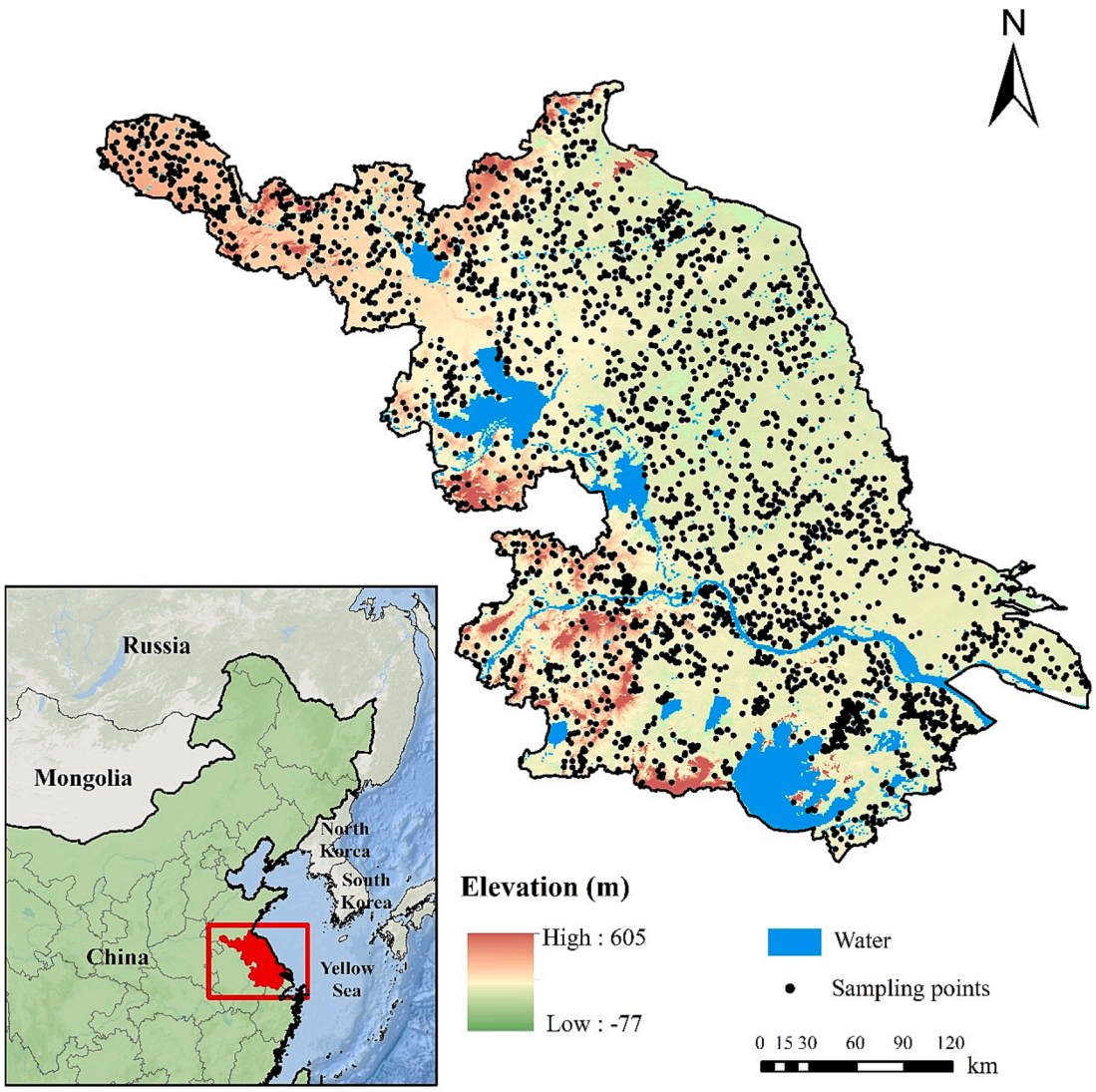


Fig. 1. Location of the study area in China and distribution of sampling points (n = 2,848).

dried and then softly ground to pass through a 2-mm sieve for particle size measurement by a laser scattering particle analyzer (Beckman Coulter, LS13320).

2.2.2. Environmental covariates

According to the SCORPAN model (McBratney et al., 2003), soil formation is influenced by various factors, including soil data (S), climate (C), organisms (O), relief (R), parent material (P), age (A) and spatial position (N). We therefore selected a series of environmental factors associated with them (Table 1).

The climate variables were obtained from the WorldClim database (Hijmans et al., 2005) with a spatial resolution of 1 km. Topographic variables were calculated using SAGA GIS software (https://www.saga-gis.org) based on ASTER GDEM data provided by the USGS (https://earthexplorer.usgs.gov) at 30 m spatial resolution. The vegetation variables were obtained from Liu et al. (2022). The soil type variables were derived from the classification of soil types in the 1:1 million digital soil map of China (https://www.resdc.cn). All the environmental covariates were resampled to a raster cell size of 90 m by bilinear interpolation, as the selected environmental covariates had different spatial resolutions due to the different data sources. Among the aforementioned environmental variables, 13 variables were selected for mapping, which were significantly correlated with the spatial distribution of soil sand content ($P < 0.01$) in the factor detector (Table S1). A detailed explanation of the factor detector is given in Section 2.4.1.

2.3. Sampling density

We employed the 10-fold cross validation method to assess the accuracy of the models. To investigate their performance under varying sampling densities, we randomly resampled each independent fold with a ratio ranging from 3% to 100%. The number of sample points and sampling densities corresponding to the ratios are shown in Table 2. Moreover, we computed the statistical average of the 10-fold cross validation to provide a measurement of the accuracy across different sampling densities for the models.

The semivariance analysis serves as a fundamental tool in geostatistics for quantifying the spatial variability of regionalized variables (Zhu et al., 2021). Here, we employed the semivariance function to analyze the spatial variation in the soil sand content at different sampling densities. Detailed calculations of the semivariance were given in Qu et al. (2023). The semivariance analysis was conducted in GS$^+$ (version 9.0).

2.4. Modelling and mapping

2.4.1. Geographical detector

Geographical detector (SI Materials and Methods) is a new statistical

Table 1
The description of the environmental variables for soil sand content prediction.

Category	Variable	Description	Spatial resolution	Selected for modelling
Climate	MAT	Mean annual air temperature (°C)	1000 m	Yes
	TempRange	Mean annual temperature range (°C)	1000 m	Yes
	TempSeason	Air temperature seasonality (°C)	1000 m	No
	SolarRad	Mean annual solar radiation ($Jm^{-2}\,yr^{-1}$)	1000 m	No
	MAP	Mean annual precipitation (mm)	1000 m	Yes
	PrecSeason	Precipitation standard deviation (mm)	1000 m	No
	WindSpeed	Wind speed ($m\,s^{-1}$)	1000 m	Yes
	VaporPress	Water vapor pressor (kPa)	1000 m	Yes
Topography	Elev	Elevation above sea level (m)	30 m	Yes
	Slpp	Slope gradient (%)	30 m	Yes
	Curpln	Planform curvature	30 m	Yes
	Curprf	Profile curvature	30 m	Yes
	TWI	Topographic wetness index	30 m	Yes
	TPI	Topographic position index	30 m	No
	OpenNeg	Negative terrain openness	30 m	No
	OpenPos	Positive terrain openness	30 m	No
Vegetation	NDVImean	Mean NDVI	90 m	Yes
	NDVIstd	Standard deviation of NDVI	90 m	Yes
Soil types	SoilOrd	Soil types of soil order	1000 m	No
	SoilGrp	Soil types of soil group	1000 m	Yes

method based on the principle of spatial differentiation of geographical phenomena to reveal the driving factors of a target variable (Wang et al., 2010). The core assumption of this method is that if an independent variable has a significant impact on a dependent variable, then the spatial distribution of these two variables should exhibit similarity (Liu et al., 2021). The model consisted of four sub-models: factor, interaction, risk, and ecological detectors. In this study, soil sand content was taken as variable Y, while environmental variables were taken as factor X. The factor detector was used to detect the importance of environmental variables in explaining the spatial distribution of sand content. The interaction detector was used to detect how these variables interacted with each other. The relevant calculations are made using Geodetector software (https://www.geodetector.cn).

2.4.1.1. Factor detector. The q-value was used to measure the contribution rate of a given variable. The higher the value is, the more spatial variation in the given variable is explained by the driver, which can be expressed as follows (Wang et al., 2010):

$$q = 1 - \frac{\sum_{h=1}^{L} N_h \sigma_h^2}{N\sigma^2}, q \in [0, 1] \quad (1)$$

where $h = 1, \ldots, L$, L is the strata of X, N_h and N are the number of sample units in strata h and the total region, respectively. σ_h^2 and σ^2 are the variances of the dependent variable in strata h and the total region, respectively.

2.4.1.2. Interaction detector. The interaction detector was used to evaluate whether factors X1 and X2 when acting together enhance or weaken the explanatory power of the dependent variable Y. First, the explanatory power of factors X1 and X2 is separately calculated for Y, q (X1) and q(X2), respectively. Then, $q(X1 \cap X2)$ is calculated. Finally, the q-values of the three variables are compared to evaluate the interaction effect. There are four categories in which the detected interactions can be classified: nonlinear weakening, single factor nonlinear weakening, double factors enhancement and nonlinear enhancement (Table S2).

2.4.2. Geostatistical models

2.4.2.1. Ordinary kriging (OK). OK is the most common format of kriging. The method aims to provide unbiased and optimal estimates of the variables by considering the semivariance in the spatial relationships between data points within the region being analyzed (Zhu et al., 2021). The algorithm only requires the target variable and is relatively simple to use and interpret. Similar to the SGS and COK models, the "gstat" package (2.1–1) (Gräler et al., 2016) in R software was used for prediction.

2.4.2.2. Sequential Gaussian simulation (SGS). SGS is an effective geostatistical method for the stochastic simulation of continuous variables (Shen et al., 2021). The SGS principle involves simulating the probability distribution function of a point whose data are unknown based on available sample data. From the simulated distribution, one result is randomly selected to represent the value of the unknown point (Ma et al., 2022). Each time a simulation value is obtained, it is replaced with the original data for the next point in the simulation. SGS can therefore generate multiple possible spatial distribution patterns, effectively avoiding the smoothing effect of OK (Zhu et al., 2021).

2.4.2.3. Co-kriging (COK). COK is an interpolation method derived from OK that makes full use of additional information from environmental variables to interpolate target variables. The autocorrelation of the target variables and the cross-correlation between the target variables and other environmental variables all contribute to a better prediction of the result (Zeng et al., 2023). However, the model is limited in that it can only be calculated using three environmental factors (Shen et al., 2019). Here, we used the mean annual temperature range, mean annual air temperature and water vapor pressor, which had the highest q-value in the factor detector, as covariates in the mapping.

2.4.3. Machine learning models

2.4.3.1. Random forest (RF). RF provides a prediction for the variables based on ensembles of regression trees (Breiman, 2001). RF is widely used in soil mapping because it is very inclusive of noise and outliers generated during processing, making the classification results more accurate. The RF model was implemented through the "randomForest" package (4.6–14) (Liaw and Wiener, 2002) in R. There are two parameters in RF modelling that need to be defined by the user: the number of input variables (m_{try}) in each tree and the number of trees (n_{tree}). Similar

Table 2
Ratio corresponding to the number of sample points and sampling density.

Proportion, %	3	5	10	20	30	40	50	60	70	80	90	100
Number	85	142	285	570	854	1139	1424	1709	1994	2278	2563	2848
Density, samples/10^3 km^2	0.80	1.33	2.66	5.31	7.97	10.63	13.28	15.94	18.60	21.25	23.91	26.57

to the cubist and SVM models, the grid search method of the "caret" package (6.0–88) (Kuhn, 2015) in R software was used to select the optimal model. The final model was configured using the parameters that resulted in the lowest prediction error.

2.4.3.2. Cubist. Cubist is a robust method for continuous variable classification that utilizes rule-based regression decision trees to make predictions with high accuracy. It fits a separate multivariate linear model to each leaf node of the regression tree based on a set of conditional rules, addressing the shortcomings associated with a single model and improving model prediction accuracy (Zhao et al., 2022). The cubist model was implemented through the "Cubist" package (0.4.2.1) (Kuhn et al., 2012) in R.

2.4.3.3. Support vector machine (SVM). SVM is a popular supervised learning technique for classification and regression based on statistical learning theory (Smola and Scholkopf, 2004). The SVM model transforms the original data into a new hyperspace using kernel functions. In this new hyperspace, the SVM algorithm searches for a hyperplane that effectively divides the classes while maximizing the margin between them. The "kernlab" package (0.9–29) (Karatzoglou et al., 2004) of R software was used to develop the SVM model.

2.5. Accuracy and uncertainty estimation

Model performance was evaluated by the mean coefficient of determination (R^2), root-mean squared error (*RMSE*) and mean absolute error (*MAE*), as shown in equations (2)-(4). Smaller *RMSE* and *MAE* values are associated with smaller errors and higher prediction accuracy. The larger the R^2 value is, the closer the fit is to the 1:1 line.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2}{\sum_{i=1}^{n}(\widehat{y_i} - \overline{y_i})^2} \tag{2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2}{n}} \tag{3}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|(\widehat{y_i} - y_i)| \tag{4}$$

where $\widehat{y_i}$ is the predicted soil sand content of sample i; y_i is the in-situ measurement of sample i; $\overline{y_i}$ is the mean value of soil sand content of all in-situ measurements, and n is the total number of samples.

Uncertainty analysis allows quantitative spatial analysis of soil property mapping results (Shrestha and Solomatine, 2006). The width of the 90% prediction interval (PI) is a useful measure of the prediction uncertainty. We computed the 90% PI in this study at every pixel. The limits were identified using the 0.05 and 0.95 quantiles of 100 replications.

3. Results and discussion

3.1. Descriptive statistics

The descriptive statistics of the sand content of the total soil samples (2,848) are shown in Table 3. The sand content of the topsoil in the study area showed a wide range from 3.2% to 95%. The mean value of the sand content was 51.6% with moderate variability (CV = 34.4%). Since the datasets were normally distributed (Fig. S1), no data transformation was applied.

3.2. Driving factors of soil sand content

Based on the analysis of the factor detector, the contribution of the main factors that affect the spatial distribution of the soil sand content is shown in Fig. 2. The first three main contributing factors were mean annual range (q = 32.8%), mean annual air temperature (q = 24.2%) and water vapor pressor (q = 23.7%). They were all climatic indicators, suggesting that long-term climatic conditions have the greatest influence on the formation and modification of the spatial distribution patterns of soil sand content.

For the topographic indicators, the topographic wetness index had the strongest explanatory power for soil sand content (q = 23.0%), followed by elevation (q = 13.3%) and slope (q = 8.5%). The topographic wetness index is a physical indicator of the influence of regional topography on runoff flow and storage which facilitates the identification of catchment areas (Grabs et al., 2009). The high explanatory power of the topographic wetness index for soil sand content indicates that the spatial distribution of soil sand content was inextricably linked to the scouring action of water flow. Similar conclusions were drawn by Mello et al. (2022), who found that drainage patterns were closely related to soil formation; therefore, hydrological attributes represented by drainage networks showed great importance for soil prediction. The driving effect of all vegetation indicators on the spatial distribution of sand content was not considerable.

The additive interactions between the factors were further detected with an interaction detector. As shown in Fig. 3, the 13 factors formed 91 pairs of interactions when superimposed in two-by-two interactions. Approximately 74% of the factor interaction combinations showed a nonlinear enhanced relationship, 26% demonstrated a double factor enhanced relationship, and no independent relationship was found. The q-values of the influence of the interaction between the two factors were all greater than the q-values of a single factor, indicating that the influence of the interaction between the factors was higher than the influence of a single factor (Fang et al., 2021). The results of the interaction detector emphasized that the factors influencing the spatial distribution of soil sand content were not the result of a simple additive relationship of individual factors, but the result of the superposition of the spatial distribution characteristics of different factors. For example, the impact of wind speed as a single factor on the spatial variation of soil sand content is minimal. However, when combined with the soil type factor, the interaction effect of these two factors, as indicated by the q-value of 0.17, falls into the category of double factors enhancement. This indicates that wind speed has a statistically meaningful influence on soil sand content in specific soil types.

3.3. Model performance

Table 4 shows the summary statistics of the prediction accuracy of each model based on a 100% sample size. The performance of the six models in predicting the soil sand content in the study area was found to be acceptable. Among the geostatistical models, OK performed best with the lowest value of *RMSE* (9.78) and the highest value of R^2 (0.72). SGS was slightly inferior, with an R^2 of 0.68, and COK was the worst, with an R^2 of 0.38. Among the machine learning models, RF had a higher

Table 3
Descriptive statistics of soil sand content in the study area.

	Min (%)	Median (%)	Max (%)	Mean (%)	SD (%)	CV (%)	Skewness	Kurtosis
Sand content	3.2	55.0	95.0	51.6	17.8	34.4	−0.49	−0.36

SD, standard deviation;
CV, coefficient of variation.

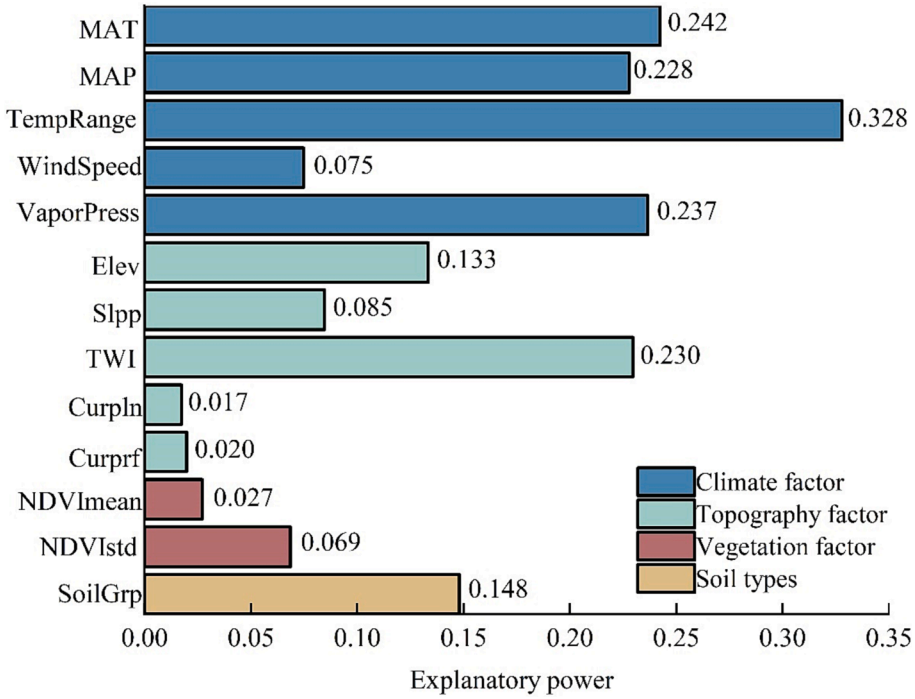


Fig. 2. Statistics of q value for single factors affecting sand content.

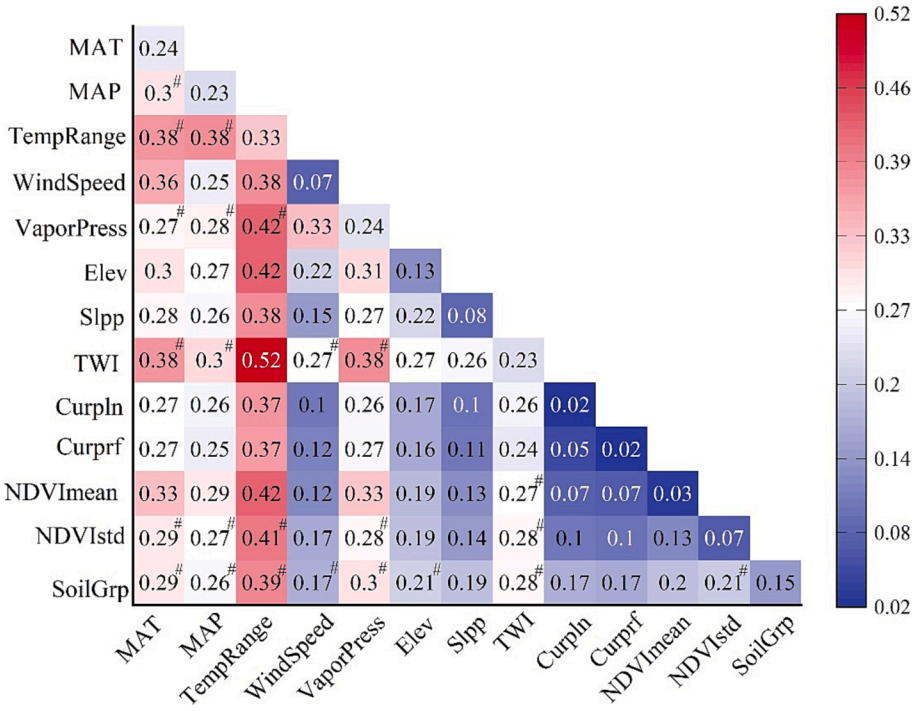


Fig. 3. Statistics of q-value for interaction factors affecting sand content (#indicates double factors enhancement and others indicate nonlinear enhancement).

accuracy (R^2 = 0.56)followed by Cubist (R^2 = 0.49), and SVM had a lower accuracy (R^2 = 0.30).

According to our results, OK and SGS exhibited the best performance in terms of mapping accuracy. However, this contradicts the findings of many studies that considered machine learning models to be better suited to soil property prediction (Szatmári and Pásztor, 2019; Sharififar, 2022; Veronesi and Schillaci, 2019; Zhang et al., 2020). This is because machine learning models were thought to have several advantages. On the one hand, since most machine learning models are based on regression tree algorithms, the aggregation of multiple trees will improve the stability of the model. On the other hand, machine learning models make full use of additional information from environmental variables, so the model's accuracy will be improved. However, the machine learning models lost its superiority in our study area. This was because in areas of gentle topography, where environmental gradients were small, the degree of spatial synergy between environmental factors and soils was usually low, resulting in a reduction in the effectiveness of using environmental variables to predict soil properties (Pouladi et al.,

Table 4
Accuracy of the sand contents predicted with OK, SGS, COK, RF, Cubist and SVM, evaluated as R^2, *MAE* and *RMSE*.

	Geostatistical models			Machine learning models		
	OK	SGS	COK	RF	Cubist	SVM
R^2	0.72	0.68	0.38	0.56	0.49	0.30
MAE	7.30	7.60	9.83	8.54	9.67	11.45
RMSE	9.78	9.98	13.65	11.24	12.87	15.03

2019; Sun et al., 2022). Among geostatistical models, OK and SGS, which relied entirely on the spatial autocorrelation of the target variables, contained valuable information that could not be captured by environmental covariates, and the inclusion of spatial information improved the prediction performance (Beguin et al., 2017). COK, a derivative of OK, considered both the spatial autocorrelation of the target variable and additional environmental variables, but the forced addition of environmental variables to the operation reduced the prediction accuracy due to the low synergy between environmental variables and target variables in our study area. Hence, in the absence of highly synergistic or highly accurate environmental variables, COK exhibits limited performance in accurately mapping soil sand content in the plains.

The performance of the predictive model is also correlated with other factors. Among them, different target variables for prediction can lead to different performances of the models. Wu et al. (2022) found that OK performed poorly when mapping soil organic carbon, which was attributed to the fact that the spatial distribution of soil organic carbon was subject to multiple influences from natural and human activities. Lu et al. (2023) found that nonlinear model outperformed linear model when mapping soil pH and carbonates. This is due to its high spatial variability and sensitivity to local environmental factors. These examples demonstrated how different target variables in digital soil mapping can present unique challenges and require tailored approaches to achieve accurate predictions. Furthermore, the topographical conditions of the study area could critically affect the accuracy of different mapping methods. A myriad of studies have considered machine learning models to be more effective for mapping in hilly mountainous areas with undulating terrain (Behrens et al., 2010; Huang et al., 2022; Long et al., 2018; Zhang et al., 2019). Pouladi et al. (2019) conducted research in areas with flat terrain and demonstrated that OK performed best and that no additional environmental variables were considered necessary for mapping.

Machine learning has gained widespread usage in recent years due to its great flexibility in establishing relationships between dependent and independent variables. However, it does not account for spatial correlations and is effectively a “black box”. Thus, spatial statistics should not be neglected. The future advancement of DSM capabilities can be greatly enhanced by effectively integrating machine learning and spatial statistics (Heuvelink and Webster, 2022).

3.4. Effects of sampling density on mapping accuracy

Fig. 4 shows the spatial prediction accuracy of the soil sand content

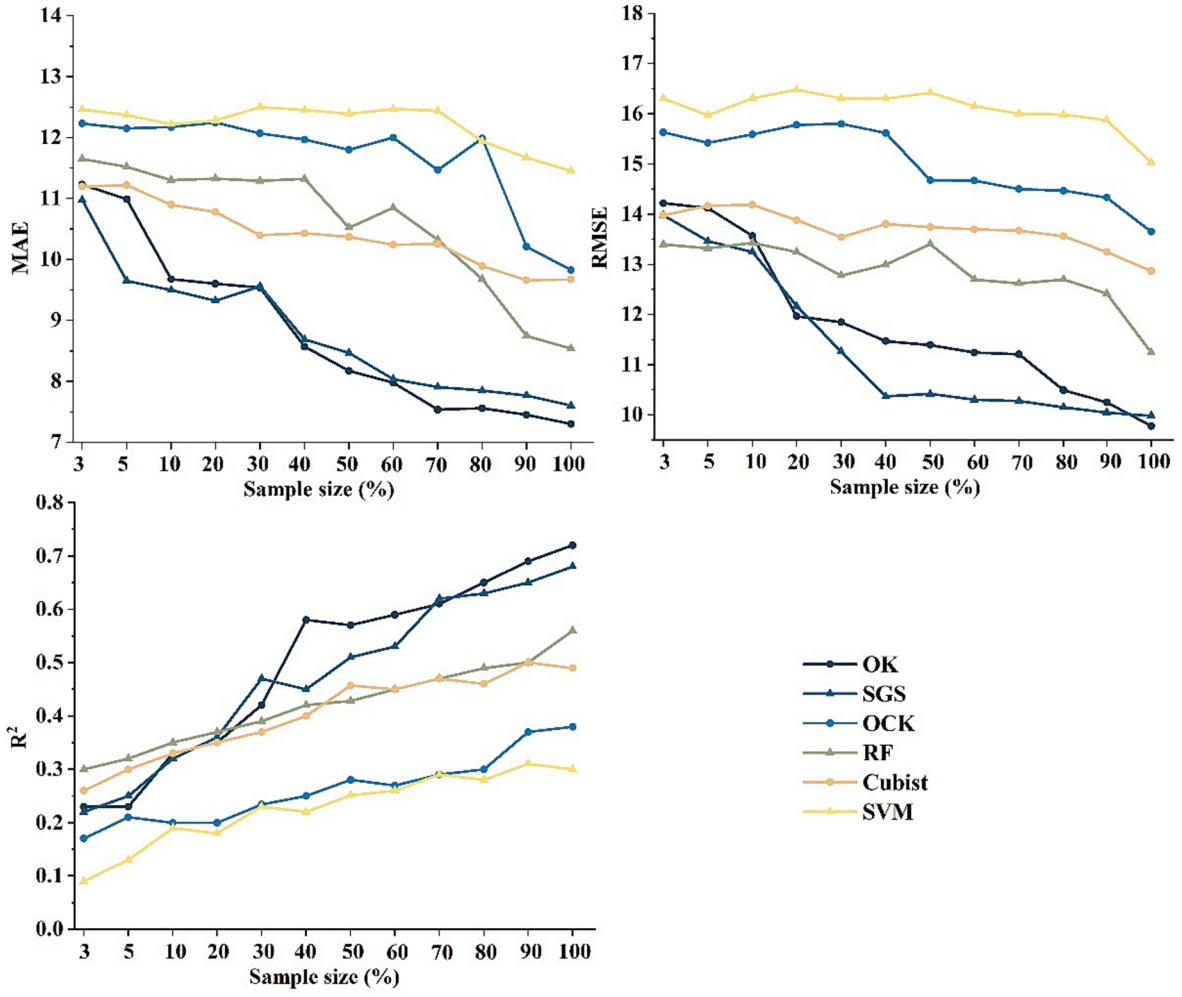


Fig. 4. Comparison of the accuracy of sand content prediction with different sampling densities.

in the study area with the change in sampling density. The study used a total of 12 sampling densities from 3% to 100% of the calibration dataset (Table 2) for soil sand content prediction. As shown in Fig. 4, the prediction accuracy of the models was sensitive to the density of sample points, and the pattern of overall variation was largely consistent. R^2, *MAE* and *RMSE* fluctuated considerably for small sample sizes (<50%), with R^2 gradually increasing and *MAE* and *RMSE* gradually decreasing as the number of sample points increased. When the sampling density continued to increase, the prediction accuracy of each model tended to stabilize, indicating that the method of improving prediction accuracy by increasing the sampling density had a marginal effect, in line with the previous assessment of Lai et al. (2021). This boils down to the increased density of sample points leading to higher spatial aggregation and information saturation. In this case, even if the sampling density is increased, no additional information can be provided.

No model always maintained an absolute accuracy advantage over other models at different sampling densities. Similar conclusions have been drawn in many studies (Shao et al., 2021; Sun et al., 2022). The geostatistical model and the machine learning model both have strengths and weaknesses in terms of prediction accuracy at different point densities. In the variation in R^2 with sampling density in Fig. 4, there was an intersection between the machine learning model and geostatistical model. At high sampling densities (>20%, 4.26 samples/10^3 km^2), geostatistical models (OK and SGS) had an advantage over other models in terms of accuracy, with higher R^2 and smaller *RMSE* and *MAE*. Pouladi et al. (2019) also found that OK could achieve effective prediction of soil properties with a high density of sample points. When the sampling density was less than 20% (4.26 samples/10^3 km^2), the machine learning models, RF and Cubist showed superiority, with higher R^2 and relatively lower *RMSE* and *MAE* compared to other models. This was because the spatial autocorrelation relied upon by geostatistical models was difficult to capture at low sampling density.

To confirm this point, we computed semivariograms for soil sand content at six sampling densities (Fig. 5). The optimal model for each sampling density was selected based on the lowest residual sum of squares and the highest R^2. The range (a) represents the maximum distance at which spatial autocorrelation occurs, indicating that a variable loses its spatial autocorrelation beyond that distance (Webster and Oliver, 2001). As the sampling density decreases, the range becomes shorter. At the same time, the fitting accuracy of the semivariogram, on which the geostatistical model relies, gradually decreases. This might be an important factor leading to poor mapping performance of geostatistical models under low-density conditions. Whereas machine learning models made full use of the additional information provided by environmental covariates to aid prediction and therefore had higher prediction accuracy than geostatistical models. In the process of solving practical problems, to save costs, we should reasonably choose the most suitable model to improve the prediction accuracy according to the density of sample points (Arrouays et al., 2014).

3.5. Spatial distribution and uncertainty analysis

Fig. 6 shows the predicted maps of soil sand content with various models based on a 100% sample size. Overall, the spatial distribution pattern of sand content obtained by the six models was generally similar, with the high value areas distributed in the northwest and coastal areas of the study area and the low value areas distributed in the central and southwest areas. Actually, there are historical reasons for the formation of sandy areas in the study area. The high value area in the northwest is associated with the accumulation of river sediment. The second largest river in China, the Yellow River, used to flow there for over 600 years, and the deposition of river sediment has created large areas of sandy soil in the area. The high values in coastal areas are associated with marine sediments of the Yellow Sea.

This scenario is similar to the results of the second national soil survey in China in the 1980s (Qu et al., 2023), demonstrating that soil texture is a fairly stable soil property that changes less over time. However, there have been discrepancies between the soil maps produced from the sample points in the China National Soil Series Survey and Compilation (2009–2019) and our mapping results. The mapping results of Liu et al. (2022) indicated that the sand content of the topsoil layer was higher in the southwest and northwest, which corresponded well with the distribution of hills in Jiangsu Province. The reasons for this discrepancy could be the different mapping methods, a lower density of sampling points (0.50 samples/10^3 km^2) than ours and a high dependence on environmental variables. Moreover, the scope of their study was national and much larger than our study area; therefore, a lower accuracy of the mapping result may have contributed to the discrepancy.

The prediction maps of soil sand content produced by the three

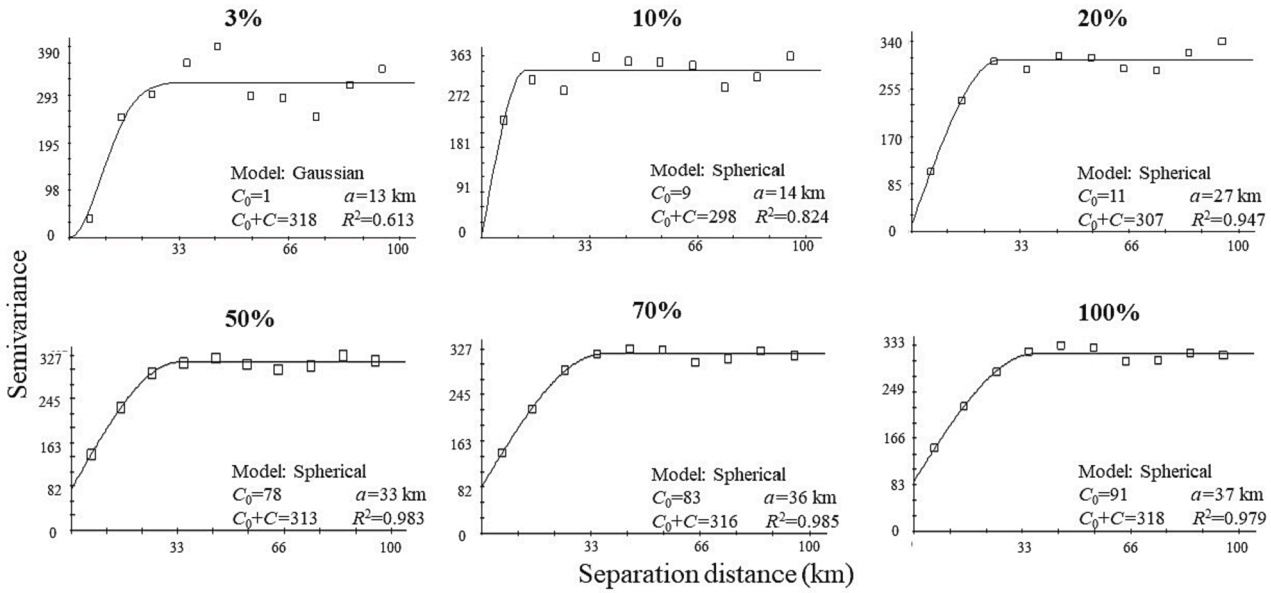


Fig. 5. Semivariograms for soil sand content at different sampling densities.

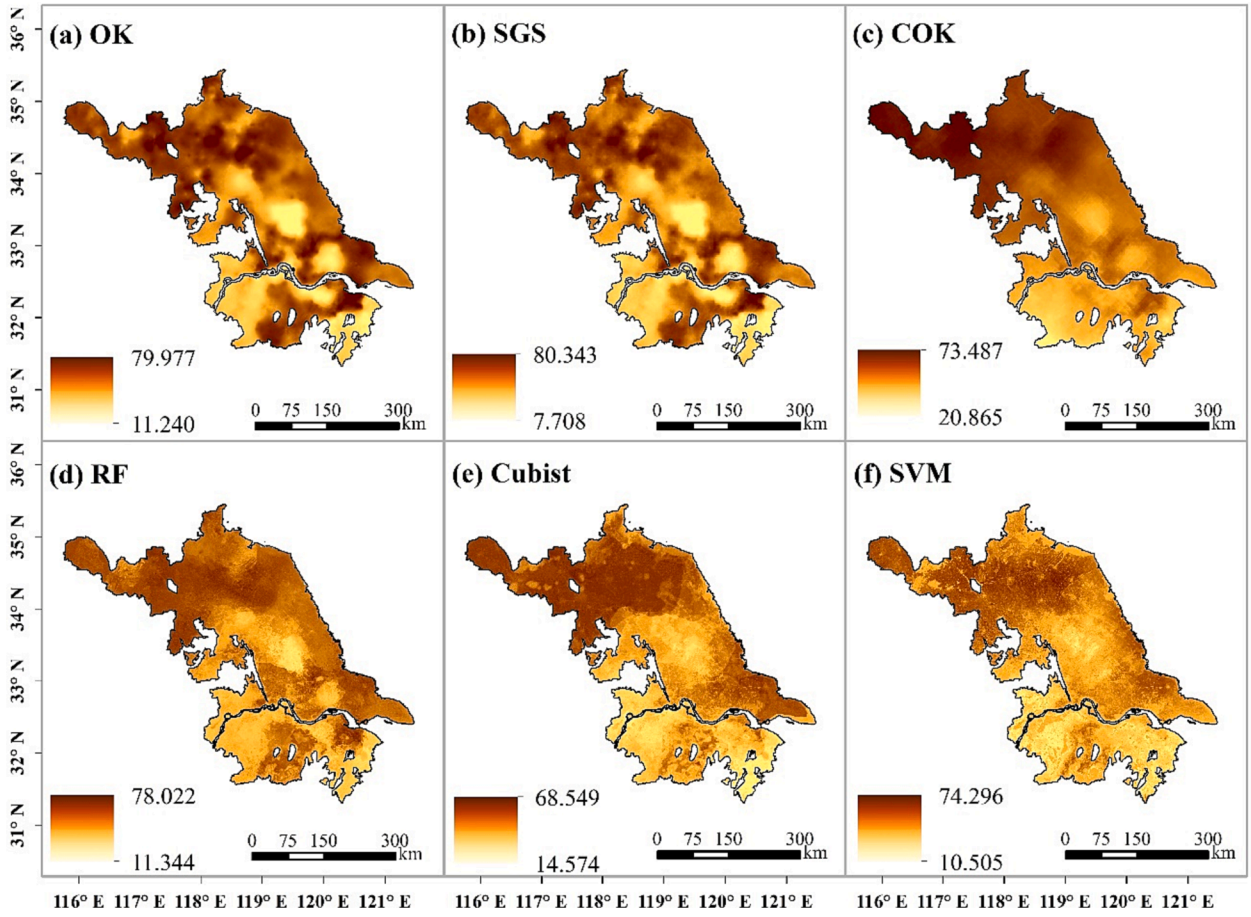


Fig. 6. The predicted maps of sand content with different models.

geostatistical models were fuzzier and blurrier than those produced by the machine learning models (Fig. 6). This was because geostatistical models rely principally on spatial distances. The predictions of soil sand content generated by OK and SGS were broadly similar (Fig. 6a and Fig. 6b). However, in terms of prediction range, OK exhibited a smoothing effect due to the filtering of information, i.e., the maximum and minimum values of the original data were removed. SGS had a wider range of predicted sand content that was closer to the measured values, as it emphasized the volatility of the raw data and could mitigate the smoothing of the data to some extent. COK considered environmental variables with small variability, yielding fuzzy mapping results and reducing the prediction range considerably (Fig. 6c). The mapping results of the machine learning models were more detailed. Among them, the mapping results of cubist and SVM showed segmentation characteristics with sporadic distribution of patched low-value areas that resembled the urban distribution of the region (Fig. 6e and Fig. 6f). This occurred because both cubist and SVM utilized linear functions in their underlying logic, and the input of NDVI among the environmental variables led to abrupt changes in the predictions. In terms of prediction range, RF had the widest range, followed by SVM and cubist.

Fig. 7 displays the uncertainties with different models of soil sand content, using lower and upper prediction limits at a 90% prediction interval. The uncertainty distributions of OK and SGS were smaller, indicating that these two models were more stable since they only consider the autocorrelations of the target variables. Their distribution was irregular throughout the study area. COK had a similar uncertainty distribution to the three machine learning models due to the inclusion of the calculation of environmental variables. The high uncertainty for COK, RF, Cubist and SVM was distributed in the northwest and southeast areas, and the low uncertainty was distributed in the central and southwest areas. Combined with the spatial distribution map of soil sand content (Fig. 6), the large uncertainty was distributed in the region with high soil sand content values. This could be attributed to the fact that in the high-value areas, the data were more discrete, which increased the uncertainty of the mapping. Fig. 8 displays the 90% prediction interval (PI) of each model for sand content prediction. This indicates that for each prediction model, there is a 90% probability that the true value falls within the prediction interval. OK yielded the smallest mean of 90% PI (20.98%), followed by SGS (21.75%), whose mean was smaller than that of RF (23.32%), cubist (23.45%), COK (23.57%), and SVM (23.79%).

In this study, the mapping results of several models were evaluated in terms of accuracy and stability. These two aspects were crucial factors in the evaluation of DSM results and could directly affect the reliability of the mapping results. However, in practice, the evaluation of digital soil mapping also includes other aspects, such as the interpretability and applicability of the map. These indicators are equally important but were not discussed in detail in this article. Therefore, in future studies, researchers should comprehensively consider multiple evaluation indicators to assess the quality of DSM mapping results to better suit practical applications.

3.6. *Limitations and further research*

There were certain limitations to our study. First, the selection of environmental variables was relatively conventional and less relevant to sand content, and advanced environmental covariates such as high-resolution remote sensing data (Padarian et al., 2022; Liu et al., 2020) and UAV imagery (Biney et al., 2023) were not included. This limitation may have restricted the ability to capture fine-scale variations in the soil

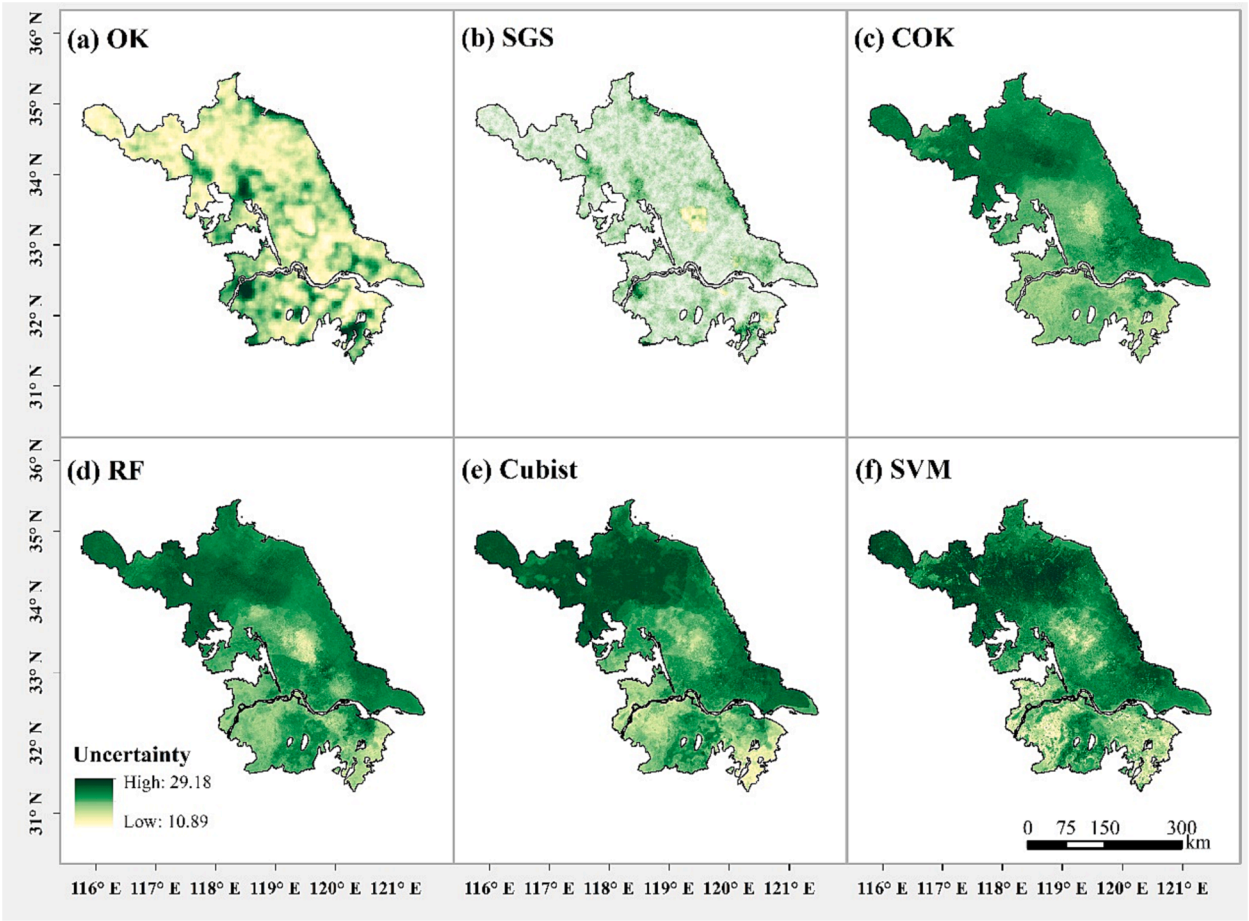


Fig. 7. Maps of the uncertainty of sand content prediction with different models.

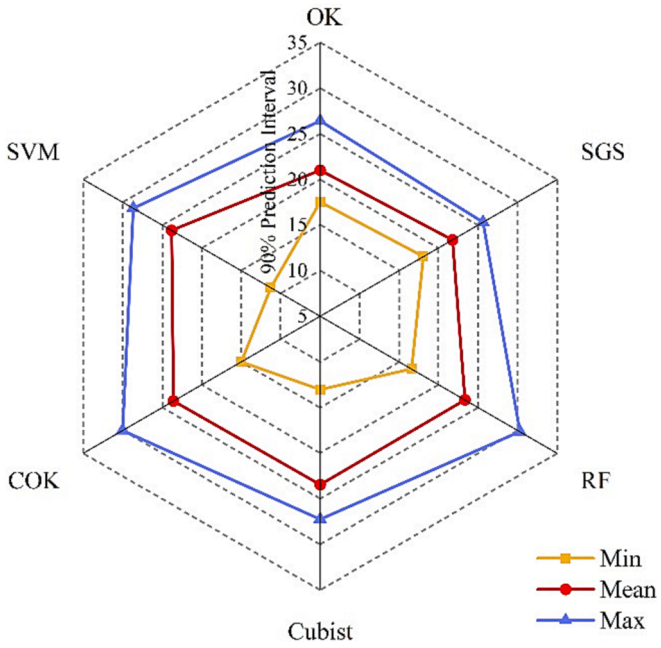


Fig. 8. 90% prediction interval (PI) of each model for sand content prediction.

mapping process. Second, the spatial resolution of the covariates was relatively low. For instance, environmental variables such as climate data were obtained at a resolution of 1 km, and soil information was also characterized on a large scale. These coarse data might contribute to the limited shared variation between the covariates and the soil mapping target variable. Third, to ensure the relevance of the selected environmental variables to soil sand content, we employed the geographical detector to screen the covariates. However, the results were not compared with the model without the screening process, which could potentially yield different outcomes.

Meanwhile, numerous studies have indicated that the sampling method and structure also have a considerable impact on the accuracy of the DSM (Du et al., 2021). Nevertheless, our study was based on a sample set collected from farmland, and the sample structure was not laid out at the diversity of the relief since the terrain in most study areas is very flat. To obtain more accurate soil maps, the next step would be to conduct an in-depth study of the influence of sampling methods and layout on the accuracy of the mapping.

4. Conclusion

In this study, we compared the effect of sampling density on the performance of two common types of soil prediction models (geostatistical models and machine learning models) for mapping the sand content of topsoil in plain areas. Moreover, the geographical detector was used to identify the main drivers of spatial variation in soil sand content. The results showed that the spatial variability of the soil sand content in this study was largely attributed to climate factors, and the dominant factors were the mean annual range, mean annual air temperature and water vapor pressor, with q-values of 32.8%, 24.2% and 23.7%, respectively. In terms of the prediction accuracy, the geostatistical models OK and SGS, which did not depend on environmental variables, had higher accuracy based on a 100% sample size, with R^2

values of 0.72 and 0.68, respectively. COK, which incorporated environmental variables, performed poorly, with an R^2 of 0.38. Among the machine learning models, RF had the highest accuracy, followed by cubist and SVM, with R^2 values of 0.56, 0.49 and 0.30, respectively. The sampling density had a considerable impact on the model accuracy. As the number of samples decreased, the model accuracy decreased (smaller R^2 and larger *RMSE* and *MAE*). The advantages of machine learning models became apparent when the sampling density was below 20% (5.31 samples/10^3 km^2) for soil sand content mapping in plain areas. The geostatistical models OK or SGS should be chosen when the sampling density is greater than 5.31 samples/10^3 km^2, and RF should be chosen when the sampling density is less than that. This study confirms that an optimal combination of model and sampling density should be chosen for the spatial prediction of soil properties in an area. If possible, further consideration needs to be given to the spatial structure of the target variables and their relationship with environmental covariates.

CRediT authorship contribution statement

Lili Qu: Methodology, Software, Formal analysis, Visualization, Writing – original draft, Writing – review & editing. **Huizhong Lu:** Supervision, Writing – review & editing. **Zhiyuan Tian:** Supervision, Conceptualization, Investigation, Writing – review & editing. **J.M. Schoorl:** Supervision, Writing – review & editing. **Biao Huang:** Investigation. **Yonghong Liang:** Investigation. **Dan Qiu:** Investigation. **Yin Liang:** Funding acquisition, Supervision, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgements

This study was supported by the Jiangsu Science and Technology Department (No. 2021059 and No.2019039) and the Soil and Water Conservation and Ecological Environment Monitoring Station of Jiangsu Province, China (JSSW201911005). We are grateful to the anonymous reviewers and editors for their professional and meticulous evaluation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.catena.2023.107572.

References

Arrouays, D., McBratney, A.B., Minasny, B., Hempel, J.W., Heuvelink, G.B.M., MacMillan, R.A., McKenzie, N.J., 2014. The GlobalSoilMap project specifications. GlobalSoilMap: Basis of the global spatial soil information system 9–12.
Arrouays, D., McBratney, A., Bouma, J., Libohova, Z., Richer-de-Forges, A.C., Morgan, C. L., Mulder, V.L., 2020. Impressions of digital soil maps: The good, the not so good, and making them ever better. Geoderma Reg. 20, e00255.
Beguin, J., Fuglstad, G.A., Mansuy, N., Paré, D., 2017. Predicting soil properties in the Canadian boreal forest with limited data: Comparison of spatial and non-spatial statistical approaches. Geoderma 306, 195–205.
Behrens, T., Zhu, A.X., Schmidt, K., Scholten, T., 2010. Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma. 155 (3–4), 175–185.
Biney, J.K.M., Houška, J., Volánek, J., Abebrese, D.K., Cervenka, J., 2023. Examining the influence of bare soil UAV imagery combined with auxiliary datasets to estimate and map soil organic carbon distribution in an erosion-prone agricultural field. Environ. Sci. Technol. 870, 161973.
Breiman, L., 2001. Random forests. Machine Learn. 45, 5–32.
Chen, S., Arrouays, D., Mulder, V.L., Poggio, L., Minasny, B., Roudier, P., Libohova, Z., Lagacherie, P., Shi, Z., Hannam, J., Meersmans, J., Richer, A.C., Walter, C., 2022. Digital mapping of GlobalSoilMap soil properties at a broad scale: A review. Geoderma 409, 115567.
Du, L., McCarty, G.W., Li, X., Rabenhorst, M.C., Wang, Q., Lee, S., Zou, Z., 2021. Spatial extrapolation of topographic models for mapping soil organic carbon using local samples. Geoderma 404, 115290.
Fang, L., Wang, L., Chen, W., Sun, J., Cao, Q., Wang, S., Wang, L., 2021. Identifying the impacts of natural and human factors on ecosystem service in the Yangtze and Yellow River Basins. J. Clean. Prod. 314, 127995.
Grabs, T., Seibert, J., Bishop, K., Laudon, H., 2009. Modeling spatial patterns of saturated areas: A comparison of the topographic wetness index and a dynamic distributed model. J. Hydrol. 373 (1–2), 15–23.
Gräler, B., Pebesma, E., Heuvelink, G., 2016. Spatio-temporal interpolation using gstat. R J. 8 (1), 204.
Guo, L., Linderman, M., Shi, T., Chen, Y., Duan, L., Zhang, H., 2018. Exploring the sensitivity of sampling density in digital mapping of soil organic carbon and its application in soil sampling. Remote Sens. 10 (6), 888.
Gupta, S., Bonetti, S., Lehmann, P., Or, D., 2022. Limited role of soil texture in mediating natural vegetation response to rainfall anomalies. Environ. Res. Lett. 17 (3), 034012.
Heuvelink, G.B., Webster, R., 2022. Spatial statistics and soil mapping: A blossoming partnership under pressure. Spat. Stat. 50, 100639.
Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25 (15), 1965–1978.
Hu, J., Xie, C., Xu, L., Qi, X., Zhu, S., Zhu, H., Zhou, Z., 2021. Direct analysis of soil composition for source apportionment by laser ablation single-particle aerosol mass spectrometry. Environ. Sci. Technol. 55 (14), 9721–9729.
Huang, H., Yang, L., Zhang, L., Pu, Y., Yang, C., Cai, Y., Zhou, C., 2022. A review on digital mapping of soil carbon in cropland: progress, challenge, and prospect. Environ. Res. Lett. 17, 123004.
John, K., Bouslihim, Y., Ofem, K.I., Hssaini, L., Razouk, R., Okon, P.B., Isong, I.A., Agyeman, P.C., Kebonye, N.M., Qin, C., 2022. Do model choice and sample ratios separately or simultaneously influence soil organic matter prediction. Int. Soil. Water. Conse. 10 (3), 470–486.
Karatzoglou, A., Smola, A., Hornik, K., 2004. kernlab-an S4 package for kernel methods in R. J. Stat. Softw. 11, 1–20.
Kuhn, M., 2015. Caret: classification and regression training. Astrophysics Source Code Library. ascl 1505, 003.
Kuhn M., Weston S., Keefer C., 2012. Cubist models for regression. R package Vignette R package version 0.0, 18, 480.
Laborczi, A., Szatmari, G., Kaposi, A.D., Pasztor, L., 2019. Comparison of soil texture maps synthetized from standard depth layers with directly compiled products. Geoderma 352, 360–372.
Lai, Y., Wang, H., Sun, X., 2021. A comparison of importance of modelling method and sample size for mapping soil organic matter in Guangdong, China. Ecol. Indic. 126, 107618.
Lamichhane, S., Kumar, L., Wilson, B., 2019. Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review. Geoderma 352, 395–413.
Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R news 2 (3), 18–22.
Liu, Y., Chen, Y., Wu, Z., Wang, B., Wang, S., 2021. Geographical detector-based stratified regression kriging strategy for mapping soil organic carbon with high spatial heterogeneity. Catena 196, 104953.
Liu, F., Rossiter, D.G., Song, X., Zhang, G.L., Wu, H., Zhao, Y., 2020. An approach for broad-scale predictive soil properties mapping in low-relief areas based on responses to solar radiation. Soil Sci. Soc. Am. J. 84 (1), 144–162.
Liu, F., Wu, H., Zhao, Y., Li, D., Yang, J., Song, X., Shi, Z., Zhu, A.X., Zhang, G., 2022. Mapping high resolution National Soil Information Grids of China. Sci. Bull. 67 (3), 328–340.
Long, J., Liu, Y., Xing, S., Qiu, L., Huang, Q., Zhou, B., Shen, J., Zhang, L., 2018. Effects of sampling density on interpolation accuracy for farmland soil organic matter concentration in a large region of complex topography. Ecol. Indic. 93, 562–571.
Lu, Q., Tian, S., Wei, L., 2023. Digital mapping of soil pH and carbonates at the European scale using environmental variables and machine learning. Environ. Sci. Technol. 856, 159171.
Ma, R., Zhu, X., Tian, Z., Qu, L., He, Y., Liang, Y., 2022. Spatial distribution and scale-specific controls of soil water-stable aggregates in southeastern China. J. Clean. Prod. 369, 133305.
Madenoglu, S., Atalay, F., Erpul, G., 2020. Uncertainty assessment of soil erodibility by direct sequential Gaussian simulation (DSIM) in semiarid land uses. Soil Till. Res. 204, 104731.
Martin, M.P., Dimassi, B., Roman Dobarco, M., Guenet, B., Arrouays, D., Angers, D.A., Blache, F., Huard, F., Soussana, J.F., Pellerin, S., 2021. Feasibility of the 4 per 1000 aspirational target for soil carbon: A case study for France. Global Change Biol. 27 (11), 2458–2477.
McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52.
Mello, F., Dematte, J., Rizzo, R., de Mello, D., Poppiel, R., Silvero, N., Sousa, G.P., 2022. Complex hydrological knowledge to support digital soil mapping. Geoderma 409, 115638.
Minasny, B., McBratney, A.B., Malone, B.P., Wheeler, I., 2013. Digital mapping of soil carbon. Adv. Agron. 118, 1–47.
Padarian, J., Stockmann, U., Minasny, B., McBratney, A.B., 2022. Monitoring changes in global soil organic carbon stocks from space. Remote Sens. Environ. 281, 113260.

Poggio, L., De Sousa, L.M., Batjes, N.H., Heuvelink, G., Kempen, B., Ribeiro, E., Rossiter, D., 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. Soil 7 (1), 217–240.
Pouladi, N., Møller, A.B., Tabatabai, S., Greve, M.H., 2019. Mapping soil organic matter contents at field level with Cubist, Random Forest and kriging. Geoderma 342, 85–92.
Qu, L., Guo, H., Li, M., Wu, F., Liang, Y., Zhu, X., Yuan, J., 2022. Experimental study on soil erosion of typical riverbank in coastal plain sandy area of Jiangsu Province. J. Soil Water Conserv. 36 (42–48), 56 in Chinese with English abstract.
Qu, L., Zhu, X., Liang, Y., Qiu, D., Zhang, Q., Liang, Y., 2023. Spatial variation of soil properties and evaluation of the risk of soil erodibility on a river alluvial and marine sedimentary plain in eastern China. J. Soils Sediments 23 (5), 2106–2119.
Rawlins, B.G., Scheib, A.J., Lark, R.M., Lister, T.R., 2009. Sampling and analytical plus subsampling variance components for five soil indicators observed at regional scale. Eur. J. Soil Sci. 60 (5), 740–747.
Sanderman, J., Hengl, T., Fiske, G.J., 2017. Soil carbon debt of 12,000 years of human land use. Proc. Natl. Acad. Sci. 114 (36), 9575–9580.
Shao, S., Zhang, H., Fan, M., Su, B., Wu, J., Zhang, M., Yang, L., Gao, C., 2021. Spatial variability-based sample size allocation for stratified sampling. Catena 206, 105509.
Sharififar, A., 2022. Accuracy and uncertainty of geostatistical models versus machine learning for digital mapping of soil calcium and potassium. Environ. Monit. Assess. 194 (10), 760.
Shen, W., Hu, Y., Zhang, J., Zhao, F., Bian, P., Liu, Y., 2021. Spatial distribution and human health risk assessment of soil heavy metals based on sequential Gaussian simulation and positive matrix factorization model: A case study in irrigation area of the Yellow River. Ecotoxicol. Environ. Saf. 225, 112752.
Shen, Q., Wang, Y., Wang, X., Liu, X., Zhang, X., Zhang, S., 2019. Comparing interpolation methods to predict soil total phosphorus in the Mollisol area of Northeast China. Catena 174, 59–72.
Shrestha, D.L., Solomatine, D.P., 2006. Machine learning approaches for estimation of prediction interval for the model output. Neural networks 19 (2), 225–235.
Silva, S.H.G., Weindorf, D.C., Pinto, L.C., Faria, W.M., Junior, F.W.A., Gomide, L.R., Curi, N., 2020. Soil texture prediction in tropical soils: A portable X-ray fluorescence spectrometry approach. Geoderma 362, 114136.
Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Stat. Comput. 14, 199–222.
Sun, X.L., Lai, Y.Q., Ding, X., Wu, Y.J., Wang, H.L., Wu, C., 2022. Variability of soil mapping accuracy with sample sizes, modelling methods and landform types in a regional case study. Catena 213, 106217.
Szatmári, G., Pásztor, L., 2019. Comparison of various uncertainty modelling approaches based on geostatistics and machine learning algorithms. Geoderma 337, 1329–1340.
Tang, L., Gudda, F.O., Wu, C., Ling, W., El-Ramady, H., Mosa, A., Wang, J., 2022. Contributions of partition and adsorption to polycyclic aromatic hydrocarbons sorption by fractionated soil at different particle sizes. Chemosphere 301, 134715.
Tian, Z., Liu, F., Liang, Y., Zhu, X., 2022. Mapping soil erodibility in southeast China at 250 m resolution: Using environmental variables and random forest regression with limited samples. Int. Soil. Water. Conse. 10 (1), 62–74.
Veronesi, F., Schillaci, C., 2019. Comparison between geostatistical and machine learning models as predictors of topsoil organic carbon with a focus on local uncertainty estimation. Ecol. Indic. 101, 1032–1044.
Wadoux, A.M.C., Minasny, B., McBratney, A.B., 2020. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth Sci. Rev. 210, 103359.
Wan, H., Li, J., Shang, S., Rahman, K., 2021. Exploratory factor analysis-based co-kriging method for spatial interpolation of multi-layered soil particle-size fractions and texture. J. Soils Sediments 21, 3868–3887.
Wang, J., Li, X., Christakos, G., Liao, Y., Zhang, T., Gu, X., Zheng, X., 2010. Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun Region, China. Int. J. Geogr. Inf. Sci. 24 (1), 107–127.
Webster, R., Oliver, M.A., 2001. Geostatistics for Environmental Scientists. John Wiley & Sons, Chichester.
Wu, Z., Chen, Y., Yang, Z., Zhu, Y., Han, Y., 2022. Mapping Soil Organic Carbon in Low-Relief Farmlands Based on Stratified Heterogeneous Relationship. Remote Sens. Basel. 14 (15), 3575.
Xie, E., Zhang, X., Lu, F., Peng, Y., Zhao, Y., 2022. Spatiotemporal changes in cropland soil organic carbon in a rapidly urbanizing area of southeastern China from 1980 to 2015. Land Degrad. Dev. 33 (9), 1323–1336.
Yageta, Y., Osbahr, H., Morimoto, Y., Clark, J., 2019. Comparing farmers' qualitative evaluation of soil fertility with quantitative soil fertility indicators in Kitui County, Kenya. Geoderma 344, 153–163.
Yang, L., Li, X., Shi, J., Shen, F., Qi, F., Gao, B., Zhou, C., 2020. Evaluation of conditioned Latin hypercube sampling for soil mapping based on a machine learning method. Geoderma 369, 114337.
Zeng, W., Wan, X., Gu, G., Lei, M., Yang, J., Chen, T., 2023. An interpolation method incorporating the pollution diffusion characteristics for soil heavy metals-taking a coke plant as an example. Sci. Total Environ. 857, 159698.
Zhang, G.L., Feng, L., Song, X.D., 2017. Recent progress and future prospect of digital soil mapping: A review. J. Integr. Agr. 16 (12), 2871–2885.
Zhang, Y., Guo, L., Chen, Y., Shi, T., Luo, M., Ju, Q., Zhang, H., Wang, S., 2019. Prediction of soil organic carbon based on Landsat 8 monthly NDVI data for the Jianghan Plain in Hubei Province, China. Remote Sens. 11 (14), 1683.
Zhang, Y., Ji, W., Saurette, D.D., Easher, T.H., Li, H., Shi, Z., Adamchuk, V.I., Biswas, A., 2020. Three-dimensional digital soil mapping of multiple soil properties at a field-scale using regression kriging. Geoderma 366, 114253.
Zhao, D., Wang, J., Zhao, X., Triantafilis, J., 2022. Clay content mapping and uncertainty estimation using weighted model averaging. Catena 209, 105791.
Zhu, X., Liang, Y., Tian, Z., Zhang, Y., Zhang, Y., Du, J., Wang, X., Li, Y., Qu, L., Dai, M., 2021. Simulating soil erodibility in southeastern China using a sequential Gaussian algorithm. Pedosphere 31 (5), 715–724.