https://doi.org/10.5194/egusphere-2025-2360
Preprint. Discussion started: 19 June 2025
© Author(s) 2025. CC BY 4.0 License.

Reducing Temporal Uncertainty in Soil Bulk Density Estimation Using Remote Sensing and Machine Learning Approaches

Sunantha Ousaha[12], Zhenfeng Shao[1], Zeeshan Afzal[1]

[1]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, 430072, China

[2]Soil Resources Survey and Research Division, Land Development Department, Ministry of Agriculture and Cooperatives, Bangkok, 10900, Thailand

Correspondence to: Sunantha Ousaha (Sunantha.sun@hotmail.com)

Abstract. Soil bulk density (BD), a key physical property affecting soil compaction, porosity, and carbon stock estimation, exhibits considerable spatial and temporal variability. However, current BD estimation methods especially traditional pedotransfer functions (PTFs) are inherently static and not designed for temporal analysis. This presents a significant limitation for soil monitoring across large and heterogeneous regions. In this study, we developed a machine learning (ML) approach integrated with remote sensing data to map and monitor BD across Thailand from 2004 to 2009 at national scale. We used multispectral indices, topographic variables, climate data, and organic carbon content to train six ML models: Artificial Neural Networks (ANN), Deep Neural Networks, Random Forest, Support Vector Regression, XGBoost, and LightGBM. Model performance was evaluated using in-situ BD measurements from 236 soil samples collected in 2004. For benchmarking purposes, 76 published PTFs were also assessed on the same dataset. Results showed that the ANN model achieved the highest prediction accuracy ($R^2 = 0.986$; RMSE = 0.017 $g\ cm^{-3}$), outperforming both other ML models and all PTFs. Temporal analysis using the ANN model revealed a 7.27% increase in mean BD and a 41.23% reduction in standard deviation between 2004 and 2009, indicating increased soil compaction and reduced variability. Feature importance analysis identified organic carbon, vegetation indices, slope, and temperature as the most influential variables. The resulting high-resolution BD maps captured national-scale spatial and temporal trends and provide a robust foundation for soil quality monitoring, carbon accounting, and sustainable land use planning in tropical agroecosystems.

1 Introduction

Soil bulk density (BD) is a critical parameter in soil science, playing a pivotal role in the measurement of soil organic carbon (SOC) stocks and influencing a wide range of soil properties (Demir et al., 2022). Surface BD is dominated factor controlling soil porosity and compaction (Li et al., 2019), which in turn impact root penetration, water movement, and microbial activity. Despite its importance, obtaining reliable BD measurements remains a challenge. Traditional methods, such as core sampling,

https://doi.org/10.5194/egusphere-2025-2360
Preprint. Discussion started: 19 June 2025
© Author(s) 2025. CC BY 4.0 License.

are labor-intensive, time-consuming, and spatially limited, especially in large and heterogeneous landscapes. Consequently, indirect approaches like PTFs and advanced techniques incorporating remote sensing and machine learning have emerged as promising alternatives for large-scale BD estimation.

Pedotransfer Functions (PTFs) have long been used to estimate BD by predicting soil properties based on readily available soil attributes. PTFs are widely utilized to estimate SOC stocks across different scales (Manuel Rodríguez-Rastrero, 2022). Early work by Manrique and Jones (1991) established one of the first PTFs to estimate BD using organic carbon. PTF models have incorporated additional variables, including fine earth fractions, organic carbon (OC), organic matter (OM), and particle size fractions (Patil and Singh, 2016). Schillaci et al. (2021) explored using soil and environmental data for PTFs, achieving variable but promising results. While PTFs offer a cost-effective and accessible approach, they are inherently limited by the mathematical assumptions of their underlying models and the quality of the data used to derive them (Vasiliniuc and Patriche, 2015). This makes PTFs less effective in predicting BD across heterogeneous landscapes, as noted by Nasta et al. (2020) and Sevastas et al. (2018). Therefore, the integration of diverse data sources and advanced modelling techniques is crucial to overcome these limitations.

The advent of remote sensing technologies has significantly enhanced the potential for BD estimation. Multispectral, hyperspectral, LiDAR, and synthetic aperture radar (SAR) sensors, deployed on satellite and airborne platforms, provide extensive spatial coverage and enable detailed soil property mapping (Poggio and Gimona, 2017). These sensors, including Landsat 7, Landsat 8, Sentinel-1/2, Hyperion, and various spectroscopy techniques, offer new possibilities for improving BD prediction accuracy (Yang and Guo, 2019). Integrating multispectral and SAR data has proven particularly effective, as it captures complementary information on soil surface properties and structure (Hengl et al., 2017). Hyperspectral imagery and LiDAR-derived covariates have also shown strong correlations with soil properties and spatial distribution patterns (Guo et al., 2021; Pittman and Hu, 2020). Soil characteristics influenced by plant cover have shown strong correlations with hyperspectral spectra (Anne et al., 2014). In contrast, vis–NIR spectra from spectroscopy did not show significant differences in performance compared to PTFs-based models, but were still superior (Katuwal et al., 2020). However, while remote sensing data significantly contribute to BD prediction, studies suggest that combining these inputs with machine learning (ML) models can further enhance accuracy and efficiency

Recent machine learning (ML) algorithms outperform regression functions and geostatistical methods in BD prediction due to their ability to capture non-linear relationships and complex interactions among soil properties (Anne et al., 2014; Panagos et al., 2024). Traditional methods like regression functions and geostatistical techniques such as Kriging (Poggio and Gimona, 2017) and Variograms are often limited by their linear assumptions and inability to manage complex soil variable interactions (Padarian et al., 2020), leading to issues such as overestimation (Panagos et al., 2024). In contrast, ML algorithms excel in these areas by leveraging large datasets to identify intricate patterns and relationships. Studies have demonstrated that advanced ML models, such as random forest (Hengl et al., 2017), artificial neural networks (Aitkenhead and Coull, 2020), support vector Machines (Hateffard et al., 2023), and extreme gradient boosting (Salehi Hikouei et al., 2021) offer significant improvements in prediction accuracy. These models can efficiently account for spatial variability and autocorrelation, providing more fast, reliable (Kim et al., 2023), and precise BD predictions across heterogeneous landscapes (Guo et al., 2021). However, ML models have limitations which can be challenging to acquire and manage (James et al., 2013). ML models often act as "black boxes," making their predictions difficult to interpret and validate (Rudin, 2019). Models trained on specific regional data may not generalize well to other areas, necessitating periodic retraining with updated data to maintain accuracy (Panagos et al., 2024). Despite advances in BD prediction using remote sensing and machine learning, most studies focus on spatial accuracy, with limited attention to year-to-year variability and temporal uncertainty at national scales.

In this study, we develop a machine learning framework that integrates remote sensing, environmental variables, and in-situ organic carbon data to estimate soil bulk density (BD) at national scale. We train six machine learning models, Artificial Neural Networks (ANN), Deep Neural Networks (DNN), Random Forest (RF), Support Vector Regression (SVR), XGBoost, and LightGBM using soil and spectral data from 2004. The best-performing model is then applied to 2009 satellite and OC data to predict BD and analyze temporal changes. To benchmark performance, we evaluate 76 published pedotransfer functions (PTFs) using the same 2004 dataset. Our objectives are to (a) identify the most accurate and scalable ML approach for BD prediction using satellite-derived covariates, (b) quantify temporal changes and uncertainty in BD between 2004 and 2009, and (c) compare ML-based predictions against conventional PTFs. These findings enable accurate, scalable, and temporally responsive monitoring of soil bulk density for improved land management and resource planning.

2 Materials and methods

2.1 Soil Data Collection

Historical soil samples were collected by the Land Development Department (LDD), Ministry of Agriculture and Cooperatives in Thailand, were utilized in this study using random sampling strategy that cover soil and crop type in agriculture area. The dataset was prepared to explore and develop a robust ML model. Figure 1 illustrates a dataset of 236 soil samples collected in 2004, which were used for model development. Additionally, soil samples with OC data collected in 2009 were used for model implementation. These samples included measurements of soil bulk density (g cm^{-3}), organic carbon (OC) content, organic matter (OM), and particle size (sand, silt, clay). BD was collected by undisturbed soil core sampling method by a cylindrical core sampler for bulk density analysis (Fao, 2023; Holliday, 1990) based on the following equation (1). Additionally, Walkley and Black’s method was used to analyse the soil sample pits to determine the percentage of OC calculated following Eq. (1),

$$\text{Bulk density (BD)} = \frac{(\text{dry soil mass} + \text{cylinder mass}) - \text{cylinder}}{\text{volumn of the cylinder}} \times 100, \quad (1)$$

$$\text{BD} = \frac{\text{dry soil mass (g)}}{\text{volume of cylinder (cm}^3\text{)}} \times 100, \quad (2)$$

The volume of the cylinder is defined in Eq. (3),

$$\text{V}_s = \pi r^2 h, \quad (3)$$

https://doi.org/10.5194/egusphere-2025-2360
Preprint. Discussion started: 19 June 2025
© Author(s) 2025. CC BY 4.0 License.

where π is 3.1416, r is radius (half of the diameter) (cm), and h is height of the core (cm)

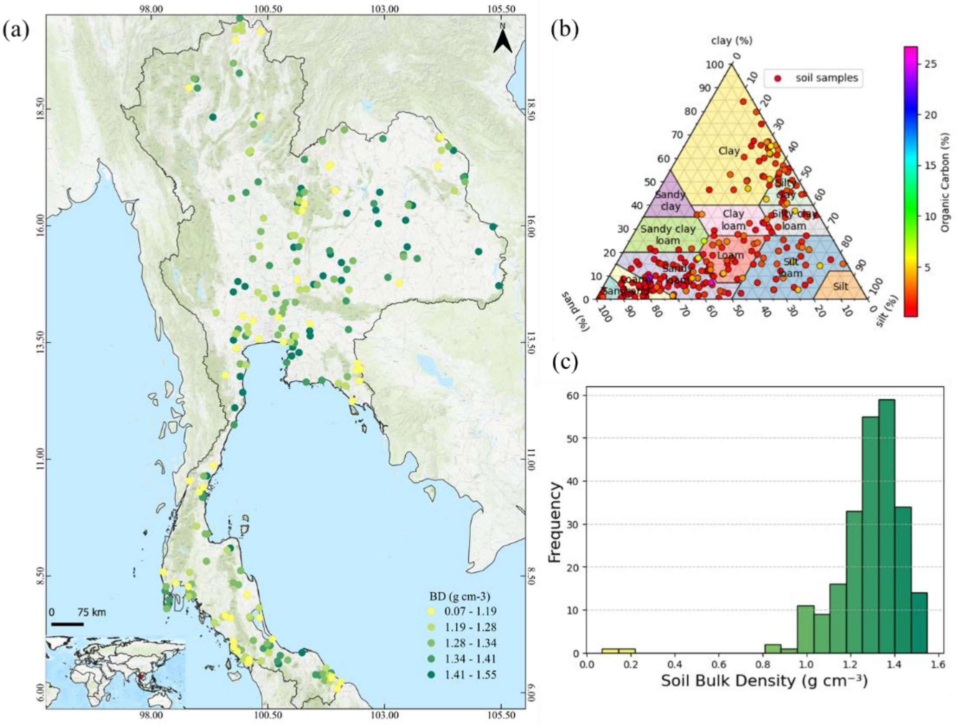


Figure 1 Study area map and location: (a) spatial distribution of soil sampling points across the study area, (b) USDA soil texture triangle diagram, and (c) histogram depicting the frequency distribution of soil bulk density (g cm⁻³).

2.2 Remote Sensing Data Collection and Preprocessing

All remote sensing data were obtained from Earth Engine data catalog, consisting of Landsat 5 thematic mapper (TM) imagery, topographic data, and climate data.

2.2.1 Landsat 5 Thematic Mapper (TM) and Pre-processing

The study area experiences frequent cloud cover due to its temperate climate, posing a significant challenge for satellite image analysis. To address this, Landsat 5 Thematic Mapper (TM) imagery for the years 2004 and 2009 was processed to correct the entire year's dataset and ensure reliable input for subsequent analysis. The imagery consists of six spectral bands spanning the visible, near-infrared (NIR), and shortwave infrared (SWIR) regions, each with a 30-meter spatial resolution. Landsat 5 TM Level 2 images were obtained from the Google Earth Engine (GEE) data catalog, leveraging its extensive repository and advanced tools for cloud correction and quality enhancement.

The cloud masking process utilized the QA band to identify and remove cloud-affected pixels across the entire dataset. Relevant bits in the QA band, specifically bit 3 for cloud shadows and bit 5 for clouds, were used to generate a comprehensive cloud mask. This ensured that only clear pixels were retained for analysis. To further mitigate the impact of residual clouds and anomalies, a weighted median composite image was created for each year, combining cloud-free pixels throughout the entire year. This composite approach minimized noise and temporal inconsistencies, providing a robust foundation for deriving key predictive indices. Vegetation indices, moisture indices, and soil indices were calculated from the composite images as covariates for BD estimation. These indices, tailored to capture the environmental variability of the study area, are detailed in Table 1.

Table 1. List of commonly used vegetation indices, soil indices, and moisture index relating to soil bulk density (BD)

Spectral Index	Abbreviate	Formular
Normalized Difference Vegetation Index	NDVI	(Band 4 – Band 3) / (Band 4 + Band 3)
Enhanced Vegetation Index	EVI	2.5 × ((Band 4 – Band 3) / (Band 4 + 6 × Band 3 – 7.5 × Band 1 + 1))
Bare Soil Index	BSI	(Band 5 + Band 3) - (Band 4 + Band 1) / (Band 5 + Band 3) + (Band 4 + Band 1)
Modified Soil Adjusted Vegetation Index	MSAVI	(2 × Band 4 + 1 – sqrt ((2 × Band 4 + 1)^2 – 8 × (Band 4 – Band 3))) / 2
Soil Adjusted Vegetation Index	SAVI	((Band 4 – Band 3) / (Band 4 + Band 3 + 0.5)) × (1.5)
Clay Index	CI	(SWIR1 - SWIR2) / (SWIR1 + SWIR2)
Normalized Difference Soil Index	NDSoI	(Band 7 – Band 2) / (Band 7 + Band 2)
Dry Bareness Soil Index	DBSI	(SWIR1-Green/SWIR1+Green)-NDVI
Normalized Difference Moisture Index	NDMI	(Band 4 – Band 5) / (Band 4 + Band 5)

2.2.2 Topography

Various terrain attributes were derived, including elevation, slope, and aspect, from the Shuttle radar topography mission SRTM Digital Elevation Data Version 4, which has an original pixel size of 90 meters. These data were used to obtain topographic data with a resampling at 30 m.

2.2.3 Climate data

For this study, TerraClimate data was utilized to obtain climate variables for the years 2004 and 2009. TerraClimate provides gridded datasets of precipitation accumulation and mean temperature at global scales, synthesized from weather station data. The data used in this study represents mean values at a specific spatial resolution of 4638.3 meters meters.

2.3 Existing 76 Pedotransfer Functions (PTFs)

This study evaluates 76 published PTFs using in-situ soil data from 2004 to identify the best-performing model for soil bulk density (BD) prediction, detailed in Table 2. The selection of these PTFs was guided by their reliance on accessible and practical soil properties, including sand, silt, and clay contents, organic carbon (OC), organic matter (OM), and soil depth. These properties are widely used due to their cost-effectiveness and ease of measurement. The PTFs are categorized into five primary groups based on their input variables: 26 rely exclusively on OM, 21 utilize OC as the primary predictor, 4 are dependent solely on particle size (PS) fractions, 6 combine particle size and OM (PSOM), and 18 integrate particle size and OC (PSOC). The PTF with the lowest Root Mean Square Error (RMSE) will be selected and applied to 2009 soil data for BD calculation, enabling a comparative analysis of temporal BD changes with values derived from the ANN model and remote sensing data.

Table 2. List of the 76 published PTFs with references

TPFs	Year	Function	Reference
OM1	1964	$BD = 10^{[2.09963 - 0.00064(\log om) - 0.22302(\log om)2]}/100$	(Curtis and Post, 1964)
OM2	1966	$BD = 1.53–0.05 \cdot om$	(Saini, 1966)
OM3	1970	$BD = 1.482–0.6786 \cdot \log_{10} om$	(Jeffrey, 1970)
OM4	1973	$BD = 100/((om/K_1) + (100 - om/K_2))$	(Adams, 1973)
OM5	1973	$BD = 1/(0.6268 + 0.0361 \cdot om)$	(Drew, 1973)
OM6	1983	$BD = \exp(-2.314 - 1.0788 \times \log(om / 100) - 0.1132 \times (\log(om / 100))^2)$	(Federer, 1983)
OM7	1983	$BD = \exp(-2.314 - 1.0788 \times \log(om) - 0.1132 \times (np.\log(om))^2)$	(Federer, 1983)
OM8	1983	$BD = 0.111 \cdot 1.45/(1.45 \cdot om /100 + 0.111 \cdot (1 - om /100))$	(Federer, 1983)
OM9	1981	$BD = 1.558–0.728 \cdot \log_{10} om$	(Harrison and Bocock, 1981)
OM10	1982	$BD = 100/((om /K_1) + (100 - om)/K_3)$	(Rawls and Brakensiek, 1982)
OM11	1989	$BD = 0.075 + 1.301 \cdot \exp(-0.06 \cdot om)$	(Grigal et al., 1989)
OM12	1989	$BD = \exp(-2.39 - 1.316 \cdot \ln(om /100) - 0.167 \cdot (\ln(om /100))^2)$	(Huntington et al., 1989)
OM13	1989	$BD = 1/(0.564 + 0.0556 \cdot om)$	(Honeysett and Ratkowsky, 1989)
OM14	1989	$\ln BD = -2.39 - 1.316 \cdot \ln(om) - 0.167(\ln(om))^2$	(Huntington et al., 1989)
OM15	1994	$BD = 1.565–0.2298 \cdot om^{0.5}$	(Tamminen and Starr, 1994)
OM16	2000	$BD = 100/[(om/0.244) + ((100 - om)/MBD_1)]$	(Post and Kwon, 2000)
OM17	2002	$BD = (0.12 \times 1.4)/(om/100 \cdot 1.4 + (1-om/100) \times 0.12)$	(Tremblay et al., 2002)
OM18	2004	$BD = \exp(-1.81–0.892 \cdot \ln(om/100)-0.092 \cdot \ln(om/100)^2)$	(Prevost, 2004)
OM19	2004	$BD = \exp(-2.314–1.0788 \cdot \ln(om/100) - 0.1132 \cdot \ln(om /100)^2)$	(Dexter, 2004)
OM20	2005	$BD = 1.775–0.173 \cdot om^{0.5}$	(De Vos et al., 2005)
OM21	2006	$BD = 100 / (om / 0.244 + (100 - om) / 1.41)$	(Cienciala et al., 2006)
OM22	2008	$BD = (0.111 \cdot 1.767)/(1.767 \cdot om/100 + [(1-om/100) \cdot 0.111])$	(Perie and Ouimet, 2008)

https://doi.org/10.5194/egusphere-2025-2360
Preprint. Discussion started: 19 June 2025
© Author(s) 2025. CC BY 4.0 License.

OM23	2008	$BD = 100/(om/0.244 + (100\text{-}om)/1.41)$	(Perie and Ouimet, 2008)
OM24	2012	$BD = \exp(0.5379 - 0.0653 \cdot om^{0.5})$	(Han et al., 2012)
OM25	2014	$BD = 100/(OM/0.14 + (100\text{-}om)/1.153]$	(Nanko et al., 2014)
OM26	2023	$BD = 0.348 + 0.993 \times \exp(-0.0882 \times om)$	(Tao et al., 2023)
OC1	1970	$BD = 1.37–0.076 \cdot oc$	(Williams, 1971)
OC2	1980	$BD = 1.72 - 0.294 \times oc^{0.5}$	(Alexander, 1980)
OC3	1991	$BD = 1.510 - 0.113 \cdot oc$	(Manrique and Jones, 1991)
OC4	1986	$BD = 1.446–0.000645 \cdot depth - 0.344 \cdot \log_{10}(oc)$	(Zinke et al., 1986)
OC5	1989	$\ln BD = 0.263 - 0.147 \cdot \ln(oc) - 0.103(\ln(oc))^2$	(Huntington et al., 1989)
OC6	2003	$BD = 1.2901–0.1229 \cdot \ln(oc)$	(Wu et al., 2003)
OC7	2005	$BD = 1.608–0.0872 \cdot oc$	(Valzano F et al., 2005)
OC8	2005	$BD = 1.780–0.379 \cdot oc^{0.5} + 0.00123 \cdot depth$	(Heuscher et al., 2005)
OC9	2005	$BD = 1.3565 \cdot \exp(-0.0046 \cdot oc \cdot 10)$	(Song et al., 2005)
OC10	2005	$BD = 1.3770 \cdot \exp(-0.0048 \cdot oc \cdot 10)$	(Song et al., 2005)
OC11	2007	$BD = 0.29 + 1.2033 \cdot \exp(-0.075 \cdot oc)$	(Yang et al., 2007)
OC12	2009	$BD = 2.684 – 140.943 \times 0.006) \times \exp(–0.006 \cdot oc)$	(Ruehlmann and Körschens, 2009)
OC13	2009	$BD = (2.684–140.943 \cdot 0.008) \cdot \exp(-0.008 \cdot oc \cdot 10)$	(Ruehlmann and Körschens, 2009)
OC14	2011	$BD = 1.4842–0.1424 \cdot oc$	(Kobal et al., 2011)
OC15	2012	$BD = 1.4903–0.33293 \cdot \ln(oc)$	(Hollis et al., 2012)
OC16	2015	$BD = 0.701 + 0.952 \cdot \exp(-0.29 \cdot oc)$	(Hossain et al., 2015)
OC17	2018	$BD = 1.448\exp^{(-0.03(oc)}$	(Abdelbaki, 2018)
OC18	2016	$BD = 1.705925–0.342497 \cdot oc^{0.5}$	(Reidy et al., 2016)
OC19	2017	$BD = 1.197 \times oc^{-0.229}$	(Atwood et al., 2017)
OC20	2018	$BD = 1/0.733 + 0.0982 \times (oc/100)$	(Chen et al., 2018)
OC21	2018	$BD = 2.039 - 0.563 \cdot oc + 0.103 \cdot oc^2$	(Sevastas et al., 2018)
OC22	2024	$BD = 0.4527 + 1.0816 \times \exp(-0.2155 \cdot oc)$	(Do et al., 2024)
PS1	1998	$BD = 1.352 – 0.0045(cl)$	(Bernoux et al., 1998)
PS2	2007	$BD = 1.5224 – 0.0005(cl)$	(Benites et al., 2007)
PS3	2007	$BD = 1.35 + 0.0045 \cdot sa + (44.7 \text{ - } sa)^2 \cdot (-6 \cdot 10–5) + 0.060 \cdot \ln(depth)$	(Tranter et al., 2007)
PS4	2016	$BD = 1.177 + 0.00263 \cdot sa - 0.0439 \cdot \ln(si) + 0.00208 \cdot si$	(Akpa et al., 2016)
PSOM1	2004	$x = -1.2141 + 4.23123 \cdot sa/100;\ y = -1.70126 + 7.55319 \cdot cl/100$ $z = -1.55601 + 0.507094 \cdot om;\ w = -0.0771892 + 0.256629 \cdot x + 0.256704 \cdot x^2–0.140911 \cdot x^3–0.0237361 \cdot y–0.098737 \cdot x^2 \cdot y–0.140381 \cdot y^2 + 0.0140902 \cdot x \cdot y^2 + 0.0287001 \cdot y^3$ $BD = 1.36411 + 0.185628 \cdot (0.0845397 + 0.701658 \cdot w–0.614038 \cdot w^2–1.18871 \cdot w^3+0.0991862 \cdot y–0.301816 \cdot w \cdot y–0.153337 \cdot w^2 \cdot y–0.072242 \cdot y^2 + 0.392736 \cdot w \cdot y^2 +$	(Rawls et al., 2004)

		$0.0886315 \cdot y^3 - 0.601301 \cdot z + 0.651673 \cdot w \cdot z - 1.37484 \cdot w^2 \cdot z + 0.298823 \cdot y \cdot z - 0.192686 \cdot w \cdot z \cdot y + 0.0815752 \cdot y^2 \cdot z - 0.0450214 \cdot z^2 - 0.179529 \cdot w \cdot z^2 - 0.0797412 \cdot y \cdot z^2 + 0.00942183 \cdot z^3)$	
PSOM2	1957	$BD = 1.8014 - 0.8491 \cdot \log_{10}(om + 2) + 0.0026 \cdot cl$	(Eschner et al., 1957)
PSOM3	2004	$BD = 1/(0.59 + 0.00163 \cdot cl + 0.0253 \cdot om)$	(Dexter, 2004)
PSOM4	2010	$BD = 1.308 + 0.0119 \cdot cl + 0.0103 \cdot sa - 0.00018 \cdot cl^2 - 0.00008 \cdot sa^2$ $0.00062 \cdot si \cdot om - 0.00059 \cdot sa \cdot om$	(Keller and Håkansson, 2010)
PSOM5	2011	$BD = 100/(om/0.224 + (100-om)/(0.935 + 0.049 \cdot \log_{10}(depth) + 0.0055 \cdot sa + 0.000065 \cdot (sa-38.96)^2))$	(Minasny and Hartemink, 2011)
PSOM6	2013	$BD = 100/(om/0.224 + (100 - om)/[1.017 + 0.0032 \cdot sa + 0.054 \cdot \log_{10}(depth)])$	(Hong et al., 2013)
PSOC1	1998	$BD = 1.578 - 0.054 \times oc^{-0.006} \times si^{-0.004} \times cl$	(Tomasella and Hodnett, 1998)
PSOC2	1998	$BD = 1.398 - 0.0047 \cdot cl - 0.042 \cdot oc$	(Bernoux et al., 1998)
PSOC3	1998	$BD = 0.87 + 0.071 \cdot \ln(cl) + 0.093 \cdot \ln(sa) - 0.254 \cdot \ln(oc)$	(Hallett et al., 1998)
PSOC4	1998	$BD = 1.46 - 0.0254 \cdot \ln(cl) + 0.0279 \cdot \ln(sa) - 0.261 \cdot \ln(oc)$	(Hallett et al., 1998)
PSOC5	2000	$BD = 1.70398 - 0.00313 \cdot si + 0.00261 \cdot cl - 0.11245 \cdot oc$	(Leonaviciute, 2000)
PSOC6	2001	$BD = 1.673 - 0.0071 \cdot oc - 0.0017 \cdot si - 0.003 \cdot cl$	(Calhoun et al., 2001)
PSOC7	2002	$BD = \exp(0.313 - 0.191 \cdot oc + 0.02102 \cdot cl - 0.0004768 \cdot cl^2 - 0.00432 \cdot si)$	(Kaur et al., 2002)
PSOC8	2005	$BD = 1.711 - 0.0487 \cdot oc^2 + 0.0059 \cdot oc^3 + 0.002 \cdot cl$	(Heuscher et al., 2005)
PSOC9	2005	$BD = 1.674 - 0.310 \cdot oc^{0.5} + 0.015 \cdot cl - 2.41 \cdot 10^{-4} \cdot si^2$	(Heuscher et al., 2005)
PSOC10	2008	$BD = 1.5688 - 0.0005(cl) - 0.009(oc)$	(Benites et al., 2007)
PSOC11	2009	$BD = 1.386 - 0.078 \cdot oc + 0.001 \cdot si + 0.001 \cdot cl$	(Men et al., 2008)
PSOC12	2012	$BD = 0.80806 + 0.823844 \cdot \exp(-0.27993 \cdot oc) + 0.0014065 \cdot sa - 0.0010299 \cdot cl$	(Hollis et al., 2012)
PSOC13	2012	$BD = 0.69794 + 0.750636 \exp(-0.230550 \cdot oc) + 0.0008687 \cdot sa - 0.0005164 \cdot cl$	(Hollis et al., 2012)
PSOC14	2013	$BD = 1.228 - 0.155 \times \log(oc) + 0.008 \cdot sa$	(Al-Qinna and Jaber, 2013)
PSOC15	2015	$BD = 1.64581 - 0.00362(cl) - 0.0016 \cdot sa - 0.0158 \cdot oc$	(Botula et al., 2015)
PSOC16	2017	$BD = (1.6179 - 0.0180 \cdot (cl + 1)^{0.46} - 0.0398 \cdot oc^{0.55})^{.33}$	(Beutler et al., 2017)
PSOC17	2018	$BD = 2.268 - 0.179 \times \ln(sa) - 0.345 * \ln(oc)$	(Sevastas et al., 2018)
PSOC18	2024	$BD = 1.243 + 2.983 \times 10^{-3} (sa) + 4.187 \times 10^{-3} (sa) - 6.208 \times 10^{-2} (oc)$	(Huf Dos Reis et al., 2024)

MBD_1= mineral bulk density (1.64 g cm^{-3}), VMF = bulk density of the mineral fraction per texture class in g cm^{-3} according to the belgian texture triangle, depth = mean depth of the soil sample (cm), K_1 = 0.223 g cm^{-3}, K_2 = 1.27 g cm^{-3}, sa = sand (%), oc = organic carbon (%), om = organic matter (%), cl = clay (%), si = silt (%)

2.4 Machine Learning Algorithms

In this work, three types of machine learning algorithms were tested for BD estimation, neural networks (ANN, DNN), ensemble models (RF, XGBoost, LightGBM), and Support Vector Regression (SVR). Each was evaluated to identify the most accurate and robust predictive model.

2.4.1 Artificial neural network (ANN)

ANN is a machine learning algorithm that mimics the structure of biological neurons, enabling computers to learn similarly to human brains. The commonly used ANN type is the multi-layer perceptron, which operates on a feed-forward model, processing inputs through neurons to produce outputs (Ghaderi et al., 2019). It consists of three layers: an input layer, hidden layers, and an output layer (Liu et al., 2020). The performance of an ANN heavily depends on the careful tuning of its hyperparameters, which govern the model's structure and learning process. Key hyperparameters include the number of neurons in the hidden layers, which defines the model’s capacity to learn complex patterns, and the activation function, such as ReLU, which introduces non-linearity to enhance learning capabilities. The learning rate plays a critical role in controlling the speed of weight updates during training, balancing stability and convergence efficiency. Batch size determines how often the model updates weights during training, influencing both computational efficiency and generalization. Finally, the number of epochs dictates the number of times the model processes the entire dataset, requiring a balance between adequate training and computational resources. These hyperparameters, when appropriately tuned, significantly enhance the model’s ability to learn and generalize, making them essential for achieving optimal performance.

2.4.2 Deep neural network (DNN)

DNN model was chosen for its ability to capture complex relationships in environmental and spectral features (Kim et al., 2023). DNNs typically consist of an input layer, multiple hidden layers, and an output layer. The increased depth of hidden layers enables DNNs to learn hierarchical feature representations, making them highly effective for tasks requiring advanced modeling capabilities. The performance of a DNN is significantly influenced by its hyperparameters, which control the architecture and training process. Key hyperparameters include the number of dense units in each hidden layer, which defines the network's capacity to learn from data, and the activation function, such as ReLU, which enhances the model's ability to capture non-linear relationships. Dropout rates are critical for regularization, reducing the risk of overfitting by randomly deactivating neurons during training. Batch normalization is another important technique that normalizes layer inputs, stabilizing and accelerating the training process. Additionally, the learning rate determines the step size during weight updates, balancing convergence speed and stability. Callbacks, such as early stopping and learning rate reduction, play a pivotal role in dynamically adjusting the training process to avoid overfitting and optimize performance. These hyperparameters collectively ensure that the DNN effectively learns and generalizes, making them essential to achieving accurate and reliable predictions for soil bulk density estimation.

2.4.3 Random Forest (RF)

RF algorithm, developed by Breiman (2001) and discussed by Hengl et al. (2018), builds numerous decision trees during training. RF effectively manages outliers, handles unbalanced data distributions, accommodates non-linear patterns, and captures complex relationships (Ao et al., 2019). The performance of RF is governed by several important hyperparameters. The number of trees (*n_estimators*) controls the size of the forest, with more trees generally improving stability and accuracy at the cost of computation time. The maximum depth of each tree (*max_depth*) determines the level of detail captured by the model, while the minimum samples required to split a node (*min_samples_split*) and the minimum samples required to be at a leaf node (*min_samples_leaf*) control the granularity of tree splitting. Another critical parameter is the *mtry* value, which specifies the number of features to consider at each split, balancing model performance and computational efficiency (Kesbi et al., 2016). The final predictions are the weighted average of the individual tree's outputs.

2.4.4 Extreme gradient boosting (XGBoost)

XGBoost utilizes an ensemble learning approach and a gradient boosting framework, demonstrating robust performance in handling complex datasets (Liu et al., 2024). This makes it a suitable choice for predicting BD with high-dimensional features. XGBoost works by sequentially adding decision trees to minimize the residual errors of the previous models, effectively improving prediction accuracy with each iteration. Key hyperparameters play a crucial role in optimizing XGBoost's performance. The learning rate (η) determines the contribution of each tree to the final prediction, with smaller values ensuring stable convergence. The number of trees (n estimators) balances the trade-off between model complexity and computation time. The maximum depth of each tree (maxdepth) controls the level of granularity in capturing data patterns, while the minimum child weight parameter prevents overfitting by limiting the number of observations a leaf node can have. Subsampling parameters, such as subsample and colsample_bytree, reduce overfitting by introducing randomness in the training process. Collectively, these hyperparameters allow XGBoost to efficiently capture intricate patterns in high-dimensional data while maintaining model robustness.

2.4.5 Light gradient boosting machine (LightGBM)

LightGBM model was utilized for predicting BD based on environmental and multispectral indices. The LightGBM model demonstrated promising performance in predicting soil BD, leveraging its gradient-based boosting technique to handle complex datasets efficiently and accurately. As well as, it can significantly outperform XGBoost in terms of computational speed and memory consumption (Ke et al., 2017). The model's performance is driven by key hyperparameters. The learning rate (η) governs the pace of weight updates during training, ensuring stable convergence. The number of trees (n estimators) controls the model's overall complexity and accuracy. The maximum number of leaves (*num_leaves*) determines the granularity of tree splits, directly influencing the model's ability to capture complex patterns. The minimum child samples parameter ensures that leaves do not overfit small, noisy subsets of data. These hyperparameters allow LightGBM to deliver accurate predictions while maintaining computational efficiency, making it highly suitable for large-scale predictive tasks.

2.4.6 Support vector regression (SVR)

SVR was developed in was developed in the mid-1990s as an extension of the Support Vector Machine (SVM) algorithm (Müller et al., 1997). The goal of SVR is to approximate the relationship between input and output variables while minimizing prediction error (Yan et al., 2018). SVR performance relies on key hyperparameters: the kernel function captures non-linear patterns, the regularization parameter (C) balances training error and model simplicity, the kernel coefficient (γ) determines data point influence, and the epsilon parameter sets the tolerance margin for prediction errors. To ensure consistency, features were standardized using a StandardScaler to normalize input data. This combination of hyperparameter tuning and preprocessing enables SVR to deliver accurate and robust predictions while effectively managing noise and data variability.

2.5 Hyperparameter Tuning and Model Optimization

In this study, the Expected Improvement (EI) acquisition function was employed as part of the Bayesian Optimization (BO) process to fine-tune hyperparameters for all machine learning models, including ANN, DNN, RF, XGBoost, LightGBM, and SVR. The EI function, defined as Eq. (4),

$$EI(X) = \mathbb{E}[max(f(X_{best}) - f(X_{next}), 0)], \tag{4}$$

where $EI(X)$ is the Expected Improvement, $\mathbb{E}$ is the expected value, X_{best} is the hyperparameter configuration with the best-observed objective function value so far, $f(X_{next})$ is the predicted value of the objective function at a new hyperparameter configuration X_{next}. By maximizing the EI, the BO process effectively identifies promising regions in the hyperparameter space where the Gaussian Process (GP) predictions exhibit high uncertainty (exploration) and simultaneously refines regions already known to perform well (exploitation) (Zhao et al., 2024). This iterative approach minimizes the risk of being trapped in local minima, ensuring a systematic and efficient improvement of the objective function. BO combined with k-fold cross-validation (k-FCV) was utilized as the hyperparameter tuning strategy to ensure optimal performance across all machine learning models (Mockus, 2005). BO efficiently explores the hyperparameter space by constructing a probabilistic surrogate model, to approximate the objective function (Sreenivasulu and Rayalu, 2024). The objective is to iteratively identify the hyperparameter configuration that minimizes the model validation loss, as defined in Eq. (5),

$$f(\theta) = \frac{1}{k}\sum_{i=1}^{k} \mathcal{L}_i(\theta), \tag{5}$$

where $f(\theta)$ is objective function, k is the number of folds in the k-fold cross-validation, $\mathcal{L}_i$ is the loss function for the i-th fold of cross-validation, θ is the hyperparameters.

To investigate the impact of training methods, batch size versus entire dataset usage, two neural network models were designed and compared. The ANN model utilized mini-batches during training, where BO fine-tuned key hyperparameters, including hidden layer configurations ([128, 64], [64, 32]), learning rates (0.001–0.01), and batch sizes (32, 64, 128). Each configuration was validated using 10-fold cross-validation to ensure robustness. In contrast, DNN model was trained using the entire dataset without batching. For DNN, BO optimized dense layer configurations (64 units per layer), dropout rates (0.1–0.5), and learning rates (0.001–0.01). Similar to ANN, k-FCV was employed to validate the hyperparameter configurations. This experimental setup allowed for a systematic evaluation of the effects of training methods on the predictive performance of the models.

For ensemble models, BO fine-tuned RF parameters, including the number of trees ($n\_estimators$ from 50-500) and splitting criteria (max_depth from 5 to None, $min\_samples\_split$ from 2-10), validated through k-FCV. XGBoost, BO optimized *learning rates* (0.01–0.1), *maximum depths* (5–20), and *subsampling ratios* (0.6–0.9). Similarly, LightGBM hyperparameters were tuned for *learning rates* (0.01–0.1), *number of leaves* (20–80), and *min_child_sample* (10–30).

For SVR, BO refined kernel-specific parameters, including the regularization parameter (C from 1-1000), kernel coefficient ($gamma$: 0.001-0.1), and *epsilon* (0.01–0.1). The integration of BO with k-FCV rigorously validated each configuration, enhancing the generalizability and predictive performance of all models.

Each hyperparameter configuration was evaluated using 10-fold cross-validation, where the dataset was partitioned into 10 equal folds. For each fold, the model was trained on $k-1$ folds and validated on the remaining fold, iteratively cycling through all folds to ensure a comprehensive evaluation. This process allowed the network to be trained and tested ten times, ensuring robust performance assessment and reliable optimization of the hyperparameters (Wong and Yeh, 2019; Rodriguez et al., 2009).

2.6 Validation and Model Assessment

The validation and model assessment process are divided into two parts: machine learning model development and the validation of 76 PTFs. For the machine learning model development, the dataset is split into 80% for training and 20% for testing. This split ensures that the model is trained on a substantial portion of the data while reserving an independent test set for unbiased evaluation of its predictive performance. The training phase employs 10-fold cross-validation to tune model parameters and assess the model's generalizability to unseen data. Evaluation metrics including the coefficient of determination (R^2), root mean squared error (RMSE) and mean absolute error (MAE) are applied as defined in Eq. (6)-(9). For the validation of the 76 PTFs, the same metrics (R^2, RMSE, and MAE), along with the unbiased root mean square deviation (ubRMSD), are utilized as defined in Eq. (6)-(10),

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(X_i - Y_i)^2}{\sum_{i=1}^{n}(\bar{Y} - Y_i)^2}, \tag{6}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)^2 , \tag{7}$$

$$RMSE = \sqrt{MSE} , \tag{8}$$

$$MAE = \frac{1}{2}\sum_{i=1}^{n}|X_i - Y_i|, \tag{9}$$

https://doi.org/10.5194/egusphere-2025-2360
Preprint. Discussion started: 19 June 2025
© Author(s) 2025. CC BY 4.0 License.

$$\text{ubRMSD} = \sqrt{\frac{1}{2}\sum_{i=1}^{n}[(X_i - \bar{X}) - (Y_i - \bar{Y})]^2}\ , \tag{10}$$

where X_i is the predicted values from the model, Y_i is the observed values, $\bar{X}$ is the mean of the predicted values, $\bar{Y}$ is the mean of the observed values, n is the total number of observations.

2.7 Model Implementations and Comparative Analysis

The model implementation focused on selecting the approach that provided the highest accuracy, lowest RMSE, and balanced feature importance using remote sensing data. The selected model was then applied to new datasets collected in 2009, utilizing only OC data as the sole predictor for BD, as no ground-truth BD measurements were available for validation in that year. This approach allowed us to evaluate the model's robustness in estimating BD using limited inputs while ensuring its generalizability to new datasets. The 2009 dataset comprised 76,089 soil samples, containing OC percentages at a depth of 30 cm, which were integrated with remote sensing data to generate BD predictions and analyse temporal changes across diverse soil and land use conditions.

2.8 Uncertainty and Variability Quantification

The temporal uncertainty and variability of BD predictions between 2004 and 2009 were quantified using statistical techniques, including descriptive statistics, percentage change analysis, and hypothesis testing. Descriptive statistics were employed to calculate the mean and standard deviation for both years, providing insights into central tendencies and dispersion. Temporal changes were assessed by computing the percentage change in mean values and the percentage change in standard deviation, while uncertainty was quantified as the absolute difference in standard deviations between the two periods (Faber, 1999). To evaluate the statistical significance of mean differences, Welch's t-test was applied as a robust method to compare the two years (Delacre et al., 2017). Below is a detailed description of the steps and formulas used following Eq. (9)-(12),

$$\Delta\mu\ (\%) = \frac{\mu_{2009} - \mu_{2004}}{\mu_{2004}} \times 100, \tag{9}$$

$$\Delta\sigma(\%) = \frac{\sigma_{2009} - \sigma_{2004}}{\sigma_{2004}} \times 100, \tag{10}$$

$$U\ =\ |\sigma_{2009} - \sigma_{2004}|, \tag{11}$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}\ , \tag{12}$$

where μ_{2009} and μ_{2004} are the mean BD in 2009 and 2004, μ_{2009} and σ_{2004} are the standard deviation BD in 2009 and 2004, $\bar{X}_1$ and $\bar{X}_2$ are the means of the two samples, s_1^2 and s_2^2 are the variances of the two samples, N_1 and N_2 are the sample sizes of the two samples.

3 Results

3.1 Descriptive Statistics

Table 3 summarizes the soil dataset. BD values range from 0.07 to 1.55 g cm⁻³, with an average of 1.28 g cm⁻³ and a standard deviation of 0.01 g cm⁻³. The data is negatively skewed (-2.91) with high kurtosis (16.30), indicating most values are clustered around the mean with a few lower extremes. OC content varies from 0.12% to 26.66%, averaging 1.83% and a standard deviation of 2.51%. High skewness (7.26) and kurtosis (65.88) suggest a distribution dominated by lower values with a few high outliers. The table details the 2004 soil properties: sand (0.30%-98%), silt (0.50%-83.50%), and clay (0.50%-84%). Sand has a near-normal distribution, while silt and clay show moderate skewness and kurtosis. High variability in OC and OM highlights the need for advanced modelling to capture these patterns accurately.

Figure 1(b) shows the USDA Soil Texture Triangle for the 2004 soil samples, revealing a predominance of clay and clay loam textures. Most samples are concentrated in high-clay content regions, indicating a significant presence of clay-rich soils, while smaller clusters in the sandy clay loam and loam categories reflect the textural variability across the study area.

Table 3 Description of statistical analysis of soil data sampling in 2004 for 76 PTFs validation and developing BD predictive model

Soil properties	Unit	Min	Mean	Max	SD	Median	Skewness	Kurtosis
sand	%	0.30	42.91	98	28.79	44.85	0.04	-1.35
silt	%	0.50	34.42	83.50	17.22	33.00	0.36	-0.34
clay	%	0.50	22.68	84.00	20.35	15.00	0.99	-0.17
OC	%	0.12	1.83	26.66	2.51	1.31	7.32	65.72
OM	%	0.21	3.13	45.96	4.31	2.25	7.32	65.72
BD	g cm^{-3}	0.07	1.28	1.55	0.01	1.31	-2.96	16.55

The correlation analysis between the features and the target variable reveals that OC has the strongest correlation with BD, showing a very strong negative relationship (r = -0.92), indicating higher OC is associated with significantly lower BD, as shown in Figure 2. Other features exhibit much weaker correlations, both positive and negative. Positive correlations include CI (0.19), NDVI (0.16), and EVI (0.16), suggesting higher values in these indices are associated with higher BD. Negative correlations are observed with rainfall (-0.02), temperature (-0.02), aspect (-0.04), and elevation (-0.09), though their impact is minimal compared to OC. This highlights that while multiple features influence BD, OC is the most significant factor.

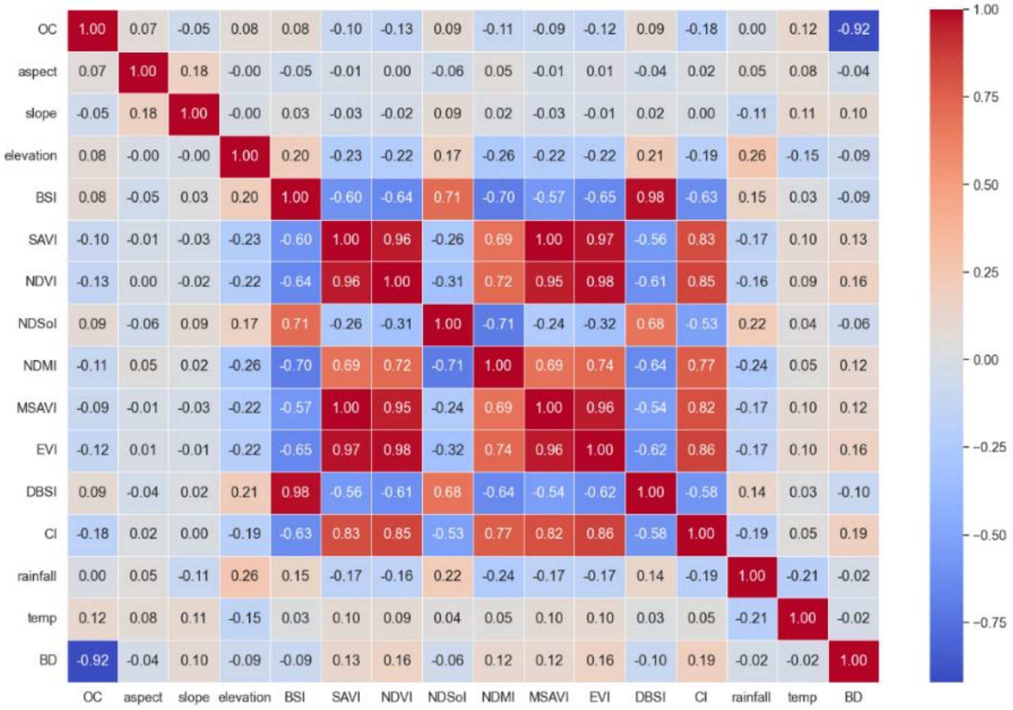


Figure 2 Correlation matrix of BD predictors (remote sensing and environment variables) and BD.

3.2 Performance evaluation of 76 PTFs using in-situ data

The comparative evaluation of 76 PTFs was carried out using RMSE, using 236 in-situ soil datasets collected in 2004. The RMSE are illustrated in Fig. 3, showcasing the significant variability in predictive accuracy among the PTFs evaluated. The RMSE values ranged from 0.051 g cm^{-3} to 6.273 g cm^{-3}, with a median of 0.171 g cm^{-3}. The mean RMSE was 0.425 g cm^{-3}, indicating that while many models performed adequately, there was considerable deviation among them. About 25% of the models achieved RMSE values below 0.101 g cm^{-3}, showcasing strong predictive performance. The OC16 model exhibited the highest performance, achieving an RMSE of 0.021 g cm^{-3}, a MAE of 0.014 g cm^{-3}, and an R^2 of 0.985, reflecting its superior accuracy and consistency in soil bulk density prediction. In contrast, the poorest-performing model, PSOC8, exhibited an RMSE of 6.273 g cm^{-3}, highlighting significant predictive errors (Fig. 4). The MAE values, which measure the average magnitude of errors, ranged from 0.014 g cm^{-3} to 2.105 g cm^{-3}, with a median value of 0.136 g cm^{-3}. This suggests that half of the models had error magnitudes below this threshold. The ubRMSD values, representing unbiased random errors, varied from 0.007 g cm^{-3} to 6.206 g cm^{-3}, with a median of 0.0945 g cm^{-3}, indicating that some models exhibited very low random errors while others had significant discrepancies.

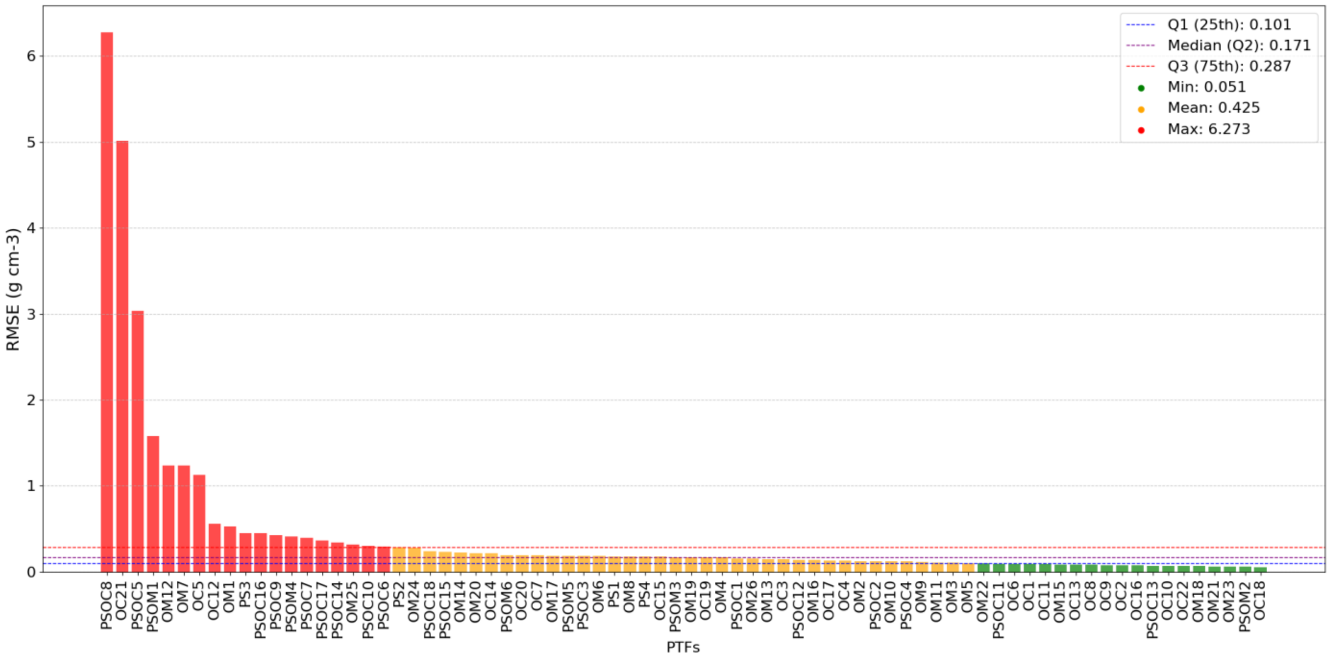


Figure 3 RMSE comparison of 76 PTFs methods for soil bulk density estimation in 2004

The performance of predictive PTFs models for BD, as categorized by the parameters, exhibited significant variability across different groups (Table 4). The OC18 model demonstrated exceptional predictive capability, achieving the lowest RMSE of 0.051 g/cm³ and the highest R² of 0.885, making it the most accurate model overall. Conversely, the PSOC8 model showed the poorest performance, with the highest RMSE of 6.273 g/cm³ and an R² of -1328.22, reflecting substantial prediction errors and uncertainty. Among OM-based models, OM21 and OM23 stood out with RMSE values of 0.062 g/cm³, highlighting their superior accuracy within the group. For PS-based models, PS4 the most reliable, with an RMSE of 0.179 g/cm³, though PS models generally exhibited higher variability and lower predictive reliability. In the PS_OM group, PSOM1 achieved the best performance, with an RMSE of 0.061 g/cm³, while in the PSOC group, PSOC13 performed strongly, with an RMSE of 0.072 g/cm³. Overall, models incorporating OC predictors, particularly OC18, consistently exhibited the highest accuracy and lowest uncertainty, whereas particle-size-based models were more prone to variability and less reliable in predicting BD.

Table 4 Top three performing PTFs for BD estimation categorized by predictor groups

PTFs	PTF Method	RMSE	ubRMSD	MAE	R^2
Organic carbon based PTFs	OC18	0.051	0.041	0.014	0.885
	OC22	0.072	0.053	0.055	0.824
	OC10	0.072	0.071	0.056	0.824
Organic matter based PTFs	OM21	0.063	0.052	0.046	0.867

	OM23	0.063	0.052	0.046	0.867
	OM18	0.070	0.053	0.050	0.835
Particle size based PTFs	PS4	0.179	0.165	0.134	-0.078
	PS1	0.180	0.177	0.125	-0.094
	PS2	0.285	0.170	0.230	-1.753
Particle size and organic carbon based PTFs	PSOC13	0.072	0.067	0.041	0.824
	PSOC11	0.094	0.092	0.067	0.701
	PSOC4	0.120	0.118	0.088	0.514
Particle size and organic matter based PTFs	PSOM2	0.061	0.057	0.041	0.874
	PSOM1	0.162	1.567	0.111	0.110
	PSOM3	0.171	0.067	0.159	0.008

3.3 Performance of ML-BD Predictive Model

To determine the most effective machine learning model for BD prediction, six algorithms were evaluated: ANN, DDN, RF, XGBoost, SVR, and LightGBM. The performance of each model was measured using the RMSE, MAE, and R² on both training and testing datasets, with hyperparameters tuned to achieve the highest accuracy (Table 6). Among the models, the ANN model achieved the highest predictive accuracy, with a testing R² of 0.986, the lowest RMSE of 0.017 g cm^{-3}, and an MAE of 0.012 g cm^{-3}, indicating its robustness in capturing the complex patterns in the dataset. The RF model also demonstrated strong performance, with a testing R² of 0.965, RMSE of 0.029 g cm^{-3}, and a minimal MAE of 0.010 g cm^{-3}, making it another reliable option for BD prediction. The XGBoost model, despite its superior performance on the training set (R² = 0.998, RMSE = 0.005 g cm^{-3}), showed a drop in accuracy on the testing set, with a testing R² of 0.936 and a higher RMSE of 0.040 g cm^{-3}. This suggests potential overfitting, where the model performs exceptionally well on the training data but struggles to generalize to new data. Similarly, LightGBM exhibited satisfactory performance, with a testing R² of 0.894 and RMSE of 0.053 g cm^{-3}, making it a viable alternative. The DNN model had a testing R² of 0.859, RMSE of 0.049 g cm^{-3}, and MAE of 0.035 g cm^{-3}, indicating moderate predictive power. In contrast, the SVR model recorded the lowest predictive accuracy, with a testing R² of only 0.549, RMSE of 0.116 g cm^{-3}, and the highest MAE of 0.074 g cm^{-3}, making it less effective for BD estimation in this context.

Table 5. The optimal hyperparameters

Model	Architecture/Parameters	Optimizer	Loss Function	Training Parameters
ANN	Input: 128 neurons; Hidden: [64, 32] neurons; ReLU activation; Output: 1 neuron	Adam	MSE	1000 epochs, batch size = 64
DNN	Dense layers: [64, 64] neurons; ReLU activation; Batch normalization; Dropout regularization	Adam (Ir-0.01)	MSE	1000 epochs, early stopping, learning rate reduction
RF	*n estimators* =100; *max depth* = None; *min samples_split* = 4; *min samples_leaf*=1	-	-	-

© Author(s) 2025. CC BY 4.0 License.

XGBoost	*n estimators* =1000; *Max Depth*= 20; *Learning Rate* (η) = 0.01; *min child_weight*=2; *Subsample* = 0.8; *col sample* bytree=0.8	-	-	-
LightGBM	*n estimators* =1000; *Learning Rate* = 0.01; *Number of Leaves* = 40	-	-	-
SVR	Kernel: RBF; C=100; γ=0.1; ϵ=0.1	-	-	StandardScaler applied

Table 6. Performance comparison of machine learning models for BD predictions

ML models	Hyperparameters	Training			Testing		
		R^2	RMSE	MAE	R^2	RMSE	MAE
ANN		0.995	0.012	0.008	0.986	0.017	0.012
DNN		0.900	0.030	0.050	0.859	0.049	0.035
RF		0.986	0.020	0.004	0.965	0.029	0.010
XGBoost		0.998	0.005	0.001	0.936	0.040	0.021
SVR		0.922	0.048	0.041	0.549	0.116	0.074
LightGBM		0.879	0.063	0.019	0.894	0.053	0.022

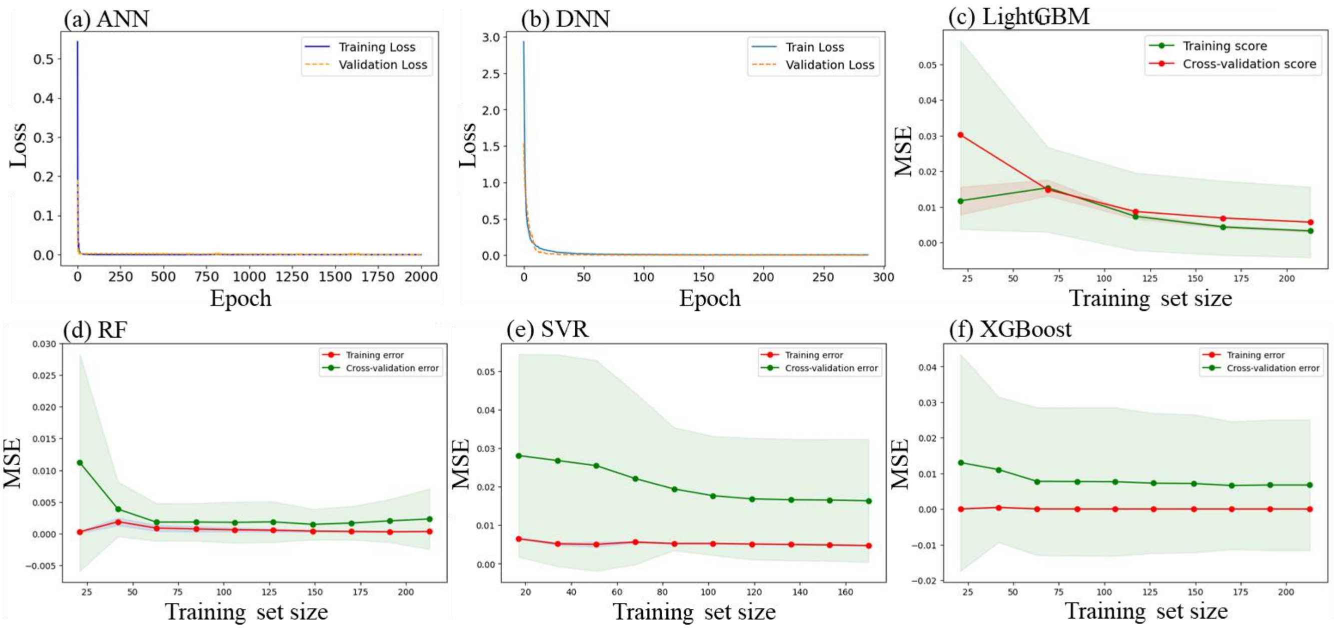


Figure 4. Loss function curves for neural network regression models: (a) Artificial Neural Network (ANN), (b) Deep Neural Network (DNN), and learning curves for other machine learning models: (c) LightGBM, (d) Random Forest, (e) Support Vector Regression (SVR), and (f) XGBoost for soil bulk density prediction.

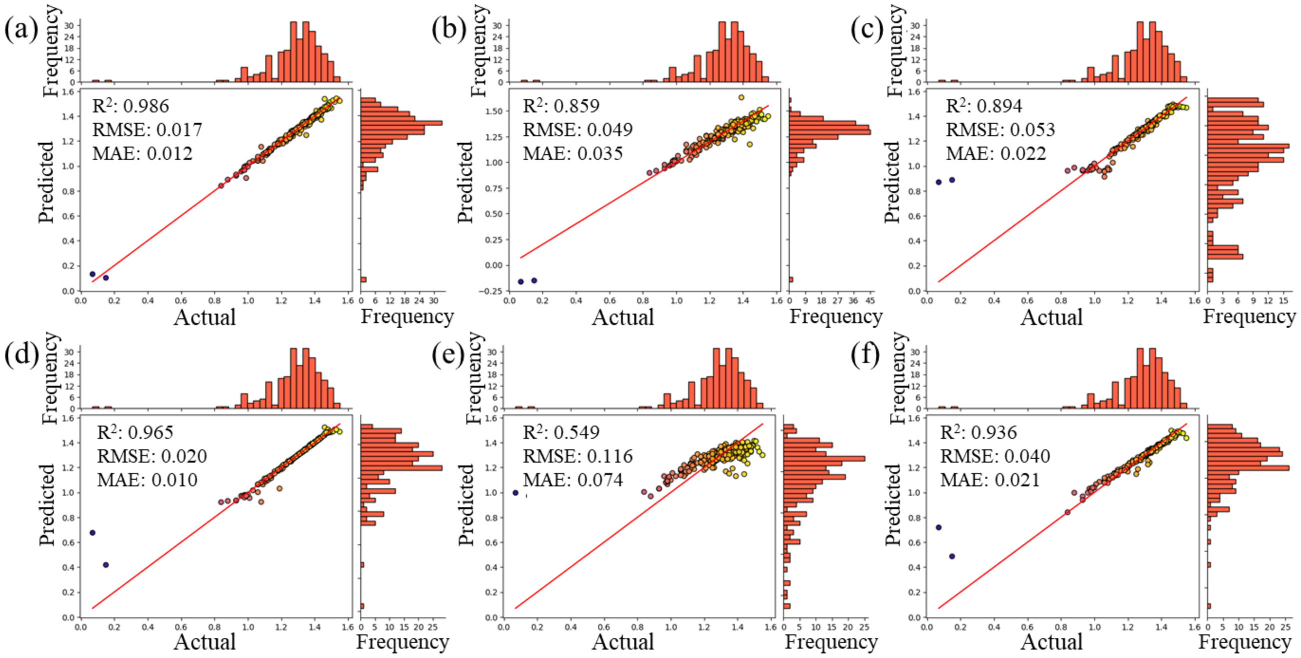
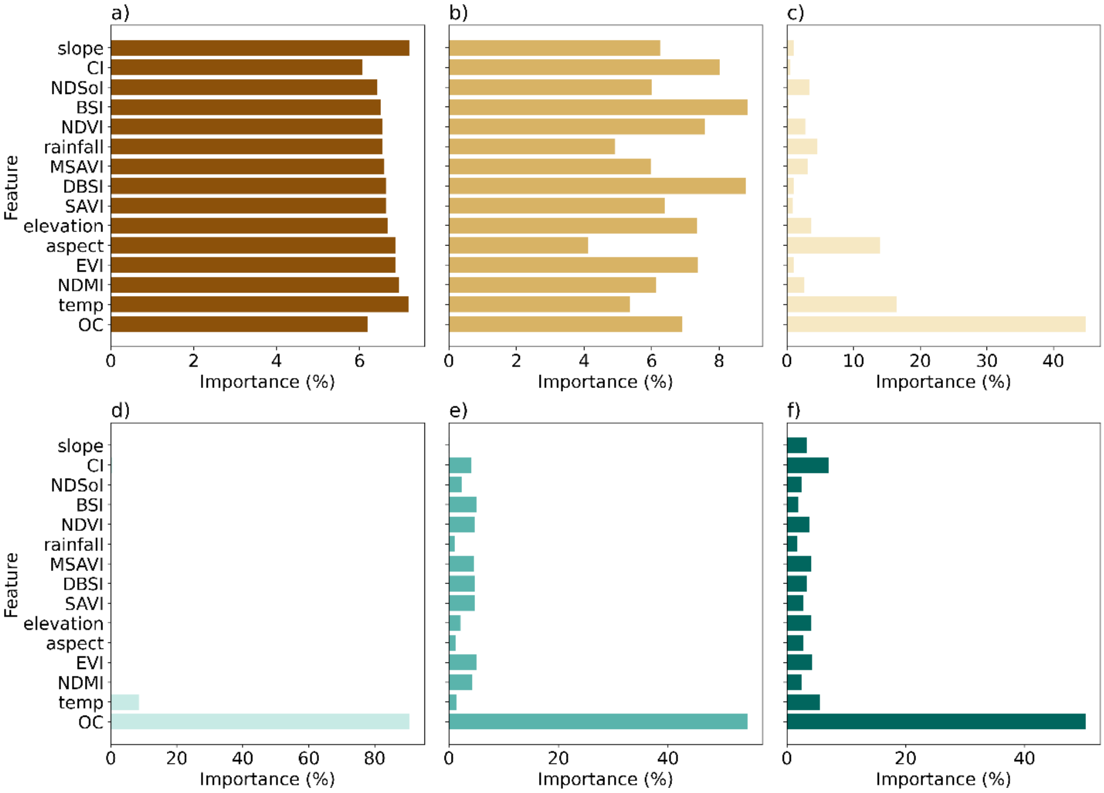


Figure 5. Scatterplot of predicted and actual BD using remote sensing data across difference machine learning: (a) ANN, (b) DNN, (c) LightGBM, (d) Random forest, (e) SVR, (f) XGboost

3.4 Importance of BD Predictors

The feature importance analysis was performed on six machine learning models (ANN, DNN, LightGBM, RF, SVR, and XGBoost) to evaluate the relevance of various remote sensing and environmental features in predicting BD. The analysis revealed distinct patterns in feature utilization across the models, highlighting the varying predictive strategies employed by each algorithm. The ANN model exhibited a relatively even distribution of feature importance, with temperature (7.18%) and slope (7.21%) being the most significant predictors. This suggests that the ANN model effectively integrates a broad spectrum of remote sensing and environmental data without heavily relying on a single feature. Similarly, the DNN model demonstrated a balanced feature utilization, identifying Bare Soil Index (BSI, 8.84%) and Dry Bare Soil Index (DBSI, 8.78%) as the primary predictors, indicating that it leverages a diverse set of inputs to achieve accurate BD predictions. In contrast, the LightGBM model showed a strong dependence on a limited set of key variables, with OC emerging as the dominant feature, accounting for 44.80% of its predictive power. Other relevant features included temperature (16.49%) and aspect (13.94%), reflecting a narrower focus compared to the ANN and DNN models. The RF model displayed an even higher dependency on OC, which contributed to 90.54% of its predictive power, followed by temperature (8.52%). This heavy reliance on OC suggests that the RF model is highly sensitive to variations in organic carbon content, making it less effective in scenarios with limited or heterogeneous OC data. The SVR model, while also prioritizing OC at 54.49%, showed a more balanced distribution of 385 secondary predictors, such as BSI (5.02%) and DBSI (4.80%), compared to the RF model. This indicates that while SVR is influenced by OC, it also considers other spectral indices, making it slightly more adaptable. The XGBoost model followed a similar pattern, with OC (50.25%) as the most important predictor, but also placed considerable weight on Clay Index (CI, 6.97%) and temperature (5.56%), reflecting a more diversified utilization of remote sensing data compared to LightGBM and RF.



390

Figure 6. Variable importance for the six machine learning models in predicting BD across the study area: (a) ANN, (b) DNN, (c) LightGBM, (d) Random Forest, (e) SVR, (f) XGBoost.

3.5 Temporal Trends and Variability (2004–2009)

The temporal changes in BD between 2004 and 2009 using a robust ANN model, focusing on identifying significant shifts and 395 evaluating associated uncertainties over time. Table 6 provides a summary of the descriptive statistics for the ANN model,

© Author(s) 2025. CC BY 4.0 License.

highlighting the changes in BD values, while Figure 9 visually represents these variations through boxplots and histograms. The results reveal a consistent increase in BD values from 2004 to 2009, suggesting a trend towards denser soils. The mean BD value rose from 1.28 g cm^{-3} in 2004 to 1.38 g cm^{-3} in 2009, reflecting a 7.27% increase. This increase may be attributed to factors such as intensified land management practices, reduced soil organic matter, or increased soil compaction over time. The minimum BD values also showed a substantial increase, rising from 0.12 g cm^{-3} in 2004 to 0.95 g cm^{-3} in 2009. This shift indicates a decline in the prevalence of low bulk density soils, possibly due to reduced loose soil structures or enhanced soil consolidation. Additionally, the standard deviation of BD decreased significantly from 0.17 to 0.10 g cm^{-3}, representing a 41.23% reduction in variability, which suggests that soil conditions became more uniform and stable over the study period. This trend may be associated with stabilized land use practices, reduced erosion, or more consistent soil management. Further supporting these observations, the skewness improved from -2.81 in 2004 to -0.58 in 2009, while kurtosis decreased from 15.37 to -0.41, indicating a shift from a highly skewed and heavy-tailed distribution to a more symmetrical, normal-like distribution. This transformation suggests a reduction in the occurrence of extreme BD values and a more balanced distribution of BD by 2009.

The negative t-statistic of -8.4213 means that the mean BD in 2004 is lower than in 2009, showing a substantial difference between the two years. The large absolute value indicates strong evidence that this difference is not due to random chance. The high degrees of freedom (235.5038) reflect that Welch's t-test correctly accounts for the unequal sample sizes and variances between the two datasets, ensuring reliable results.

Figure 8 provide additional insights into the frequency distribution of BD. In 2004, BD values displayed a broader spread with a peak around 1.30 g cm^{-3}, while in 2009, the distribution narrowed and shifted towards 1.40 g cm^{-3}, suggesting increased soil compaction. The 2009 distribution also shows fewer low-density values, supporting the trend of more compacted soils over time. These temporal changes reflect a general increase in BD values and reduced variability, suggesting a transition towards denser and more consistent soil conditions across the study area. The lower uncertainty value in 2009 compared to 2004 indicates greater stability in the BD predictions, underscoring the robustness of the ANN model in capturing temporal dynamics in soil bulk density.

Table 7. Comparison of predicted BD values in 2004 and 2009 using ANN model compare with PTFs

Year	Method	Min	Mean	Max	SD	Skewness	Kurtosis	Mean Change (%)	SD Change (%)	Uncertainty	Welch's t-test
2004	ANN	0.12	1.28	1.55	0.17	-2.83	15.73	7.274	-41.23	0.07	-8.42
2009	ANN	0.95	1.38	1.56	0.10	-0.58	-0.41				(235.50)*
2004	OC18	-0.06	1.29	1.59	0.19	-2.96	16.44	10.34	-41.35	0.08	-10.64
2009	OC18	0.96	1.42	1.71	0.11	-0.39	-0.12				(235.50)*

* is degree of freedom

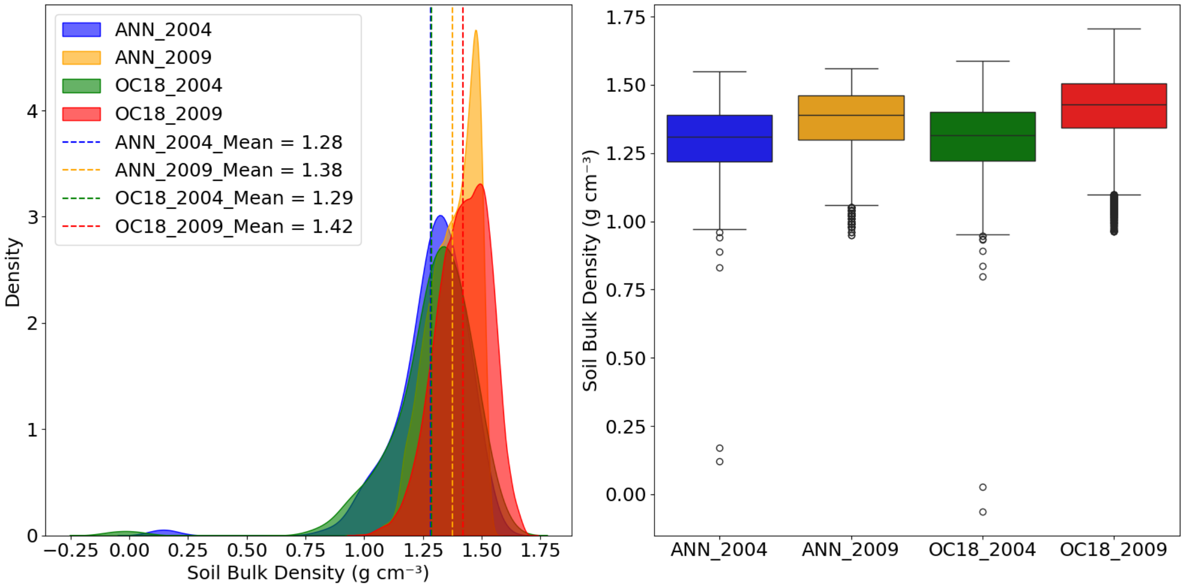


Figure 7. Histogram of predicted soil bulk density (BD) in 2004 and 2009 between a robust ANN model combined remote sensing and PTFs (OC18) method.

3.6 Spatial distribution of soil bulk density (BD) in Thailand

Figure 9 illustrates the spatial distribution of BD across Thailand in 2009, categorized into six distinct BD classes ranging from 0.95 g cm^{-3} to 1.56 g cm^{-3}. The distribution patterns show a clear spatial heterogeneity, with a notable concentration of lower BD values (0.95–1.27 g cm^{-3}) predominantly located in the northern, central, and southern regions. These lower bulk density values are often associated with less compacted soils, which could be linked to a combination of lower OC content and reduced land management pressures.

Conversely, higher BD values (1.44–1.56 g cm^{-3}) are primarily found in the northeastern part of Thailand, indicating more compacted soils. This pattern suggests that the northeastern region might have undergone more intensive agricultural activities or experienced land degradation processes, leading to increased bulk density. The presence of higher BD values in these areas could pose challenges for crop productivity due to potential soil compaction and decreased soil porosity, which can impact root penetration and water infiltration.

https://doi.org/10.5194/egusphere-2025-2360
Preprint. Discussion started: 19 June 2025
© Author(s) 2025. CC BY 4.0 License.

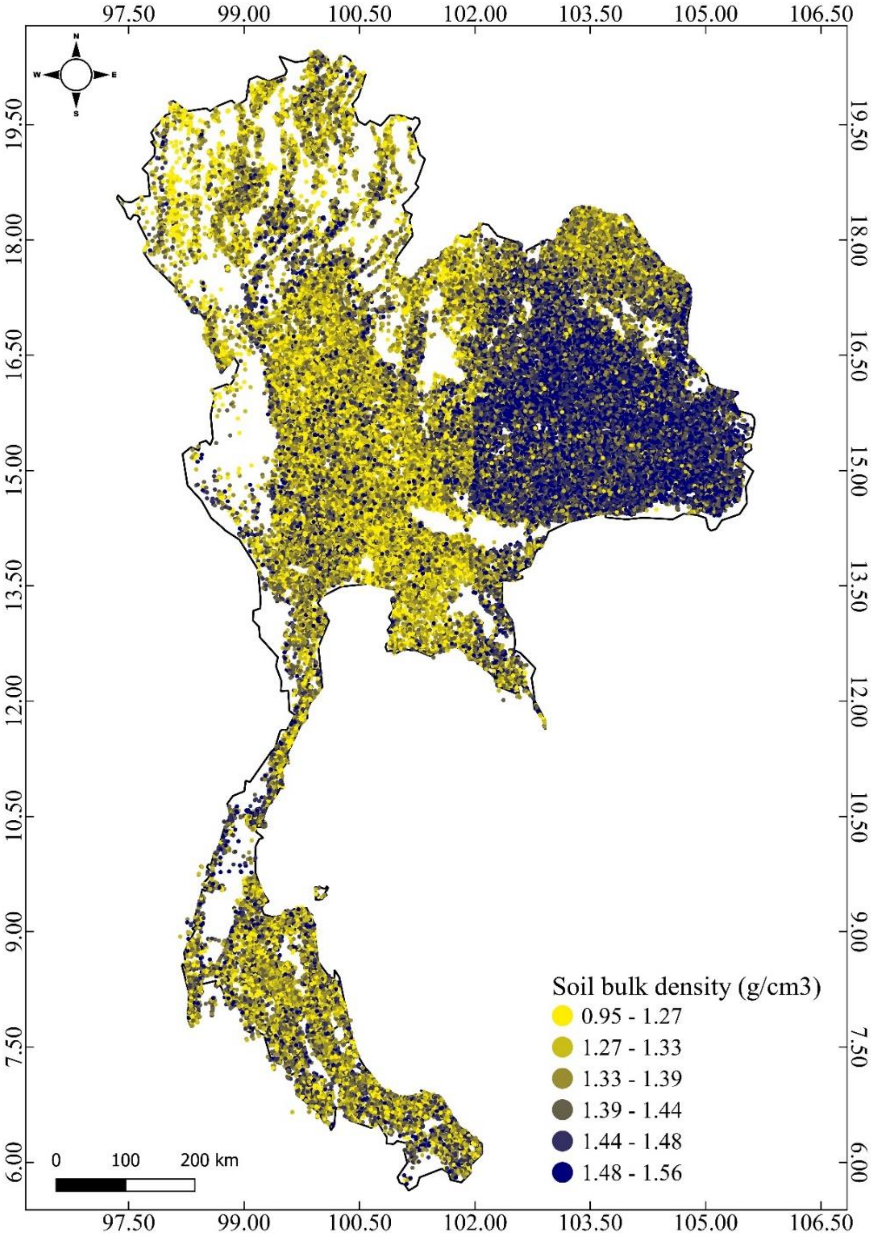


Figure 8 Spatial distribution of BD in 2009

4 Discussion

4.1 Limitations of PTF-Based BD Models

The evaluation of BD predictions using 76 published PTFs revealed several limitations, primarily related to the narrow focus on specific soil properties. These models often rely heavily on OC or Particle Size (PS), which restricts their applicability across diverse landscapes. While OC-based PTFs have been shown to perform well under stable and homogeneous soil conditions (Alexander, 1980; Sevastas et al., 2018; Yi et al., 2016), their over-reliance on OC can result in significant deviations when soil properties are influenced by additional factors, such as soil texture, moisture, and topographic variations (Lupikis et al., 2017; Nasta et al., 2020; Tsui et al., 2013). Our findings show that OC-based PTFs, while exhibiting strong alignment with the RS-ANN model in mean BD values, displayed higher variability and greater prediction uncertainty, particularly in regions with fluctuating organic matter content. The heavy dependence on OC as the primary predictor of BD was evident in PTFs such as Manrique and Jones (1991), Reidy et al. (2016), and Do et al. (2024), which showed higher uncertainty. This aligns with studies by Rawls et al. (2004) and Honeysett and Ratkowsky (1989), who found that OC-based models tend to perform poorly in soils where other factors, such as soil compaction, soil texture, or land management practices, have a stronger influence on bulk density. Additionally, Saini (1966) and Williams (1971) demonstrated significant deviations even under low OC conditions (<2%), further emphasizing the limitations of OC-reliant PTFs in complex and heterogeneous landscapes (Vasiliniuc and Patriche, 2015). High sensitivity to extreme OC or OM values (>20%) was another key limitation observed in several PTF models, including Heuscher et al. (2005), Valzano F et al. (2005), and Tomasella and Hodnett (1998), which often resulted in negative BD predictions. This phenomenon is particularly problematic in soils with high organic matter content, where extreme BD values can lead to unreliable estimates (Sevastas et al., 2018). Our study identified 12 PTFs that were highly sensitive to OC values above 20%, frequently producing unrealistic or negative BD predictions. This issue has been highlighted in previous research, where Minasny and Hartemink (2011) and Jeffrey (1970) reported that PTFs derived from small datasets or regions with specific soil properties tend to perform poorly when applied outside their calibration range.

The analysis also revealed that PTFs based solely on PS or those incorporating PS and OC (PS-OC) showed substantial prediction errors and high variability compared to OC-based and remote sensing-integrated models. This finding is consistent with Al-Qinna and Jaber (2013), who found that PS-based PTFs perform well in sandy soils but struggle in regions with mixed soil textures. Models such as Akpa et al. (2016) and Bernoux et al. (1998), which were heavily reliant on particle size, displayed significant discrepancies in minimum and mean BD values when applied to heterogeneous landscapes. This variability suggests that PS-based models may not capture the complex interactions between soil properties, leading to reduced reliability in diverse regions. Our findings are supported by Nasta et al. (2020), who emphasized that PTFs based on static soil properties fail to capture temporal variability, leading to high uncertainty when applied over extended periods. Many of the evaluated PTFs displayed significant uncertainty when applied outside their calibration regions. Models such as Prevost (2004) and Hollis et al. (2012) showed relatively low uncertainty in regions with homogeneous soil properties but failed to perform reliably in more complex environments with varying soil textures and climatic conditions. This is consistent with findings by Xu et al. (2016) and Vasiliniuc and Patriche (2015), who reported that PTFs developed using small or region-specific datasets tend to produce biased estimates when applied at larger scales. The limited generalizability of these models suggests that their use should be confined to areas with similar soil properties and environmental conditions to those in the calibration dataset.

4.2 Impact of BD Predictors and Organic Carbon for Model Prediction

The feature importance analysis for BD prediction revealed distinct patterns in predictor usage across various machine learning models, highlighting the critical role of different environmental and remote sensing variables in model performance. Each of the six models evaluated (ANN, DNN, LightGBM, RF, SVR, and XGBoost) exhibited unique dependencies on OC and other predictors, demonstrating the diverse strategies and strengths of these algorithms. The analysis identified three distinct patterns in feature utilization across models: (1) balanced use of diverse predictors in ANN and DNN models, (2) heavy reliance on OC in tree-based models.

The integration of remote sensing data significantly enhances the predictive performance of machine learning models for BD estimation. Memon et al. (2019) reported the ability of neural network algorithms to capture complex, non-linear relationships in high-dimensional data. The balanced utilization of temperature, NDMI, EVI, and slope by the ANN model indicates its strength in leveraging diverse environmental information for accurate BD predictions, supported by Schillaci et al. (2021), Yi et al. (2016), and Jalabert et al. (2010). This approach contrasts with other models like RF and SVR, which predominantly rely on a single predictor (OC), reducing their effectiveness in heterogeneous landscapes. Studies by (Schillaci et al., 2021) and (Yang et al., 2007) corroborate these findings, showing that tree-based models, such as RF and LightGBM, tend to overfit specific features, making them less robust in regions with complex soil properties. This reliance limits their capacity to generalize and highlights the advantage of ANN models in utilizing a comprehensive set of predictors for robust soil property estimation across diverse environmental conditions. The feature importance analysis of the ANN model revealed that slope, temperature, NDMI, and EVI were the top predictors, highlighting their significant influence on BD estimation. Each of these variables interacts with soil properties and land surface dynamics. Tsui et al. (2013) reported steeper slopes are often linked to lower SOC due to higher erosion rates and reduced water infiltration, leading to increased bulk density in topsoil layers. Additionally, slope affects soil compaction, root penetration, and the overall structure of the soil profile, influencing the spatial distribution of soil properties (Lupikis et al., 2017). Consistent with the Results of Yang et al. (2007) and Schillaci et al. (2021). Davidson and Janssens (2006), Jalabert et al. (2010) and Yi et al. (2016) demonstrated the effects of temperature on organic matter decomposition and soil compaction, while Gao (1996) validated the use of moisture indices in capturing soil moisture dynamics, which are closely linked to BD changes. Finally, Huete et al. (2002) and Galle et al. (2021) confirmed the utility of EVI and other vegetation indices in reflecting soil structure and health, supporting their inclusion in BD prediction models.

In contrast, models like Random Forest (RF) and Support Vector Regression (SVR), which relied predominantly on OC, showed less balanced feature importance and produced less generalizable outcomes. Previous studies have consistently reported a strong negative correlation between OC and BD (Yang et al., 2007; Tsui et al., 2013; Lupikis et al., 2017), emphasizing the role of OC as a key driver of soil bulk density. However, over-reliance on OC can lead to significant predictive errors in regions where other soil and environmental factors, such as texture, moisture, and topographic variations, play a larger role. Perie and Ouimet (2008) and Minasny and Hartemink (2011) highlighted that OC-based models may perform poorly in soils with varying mineral compositions or where land management practices strongly influence soil properties. This is consistent with the findings of Al-Qinna and Jaber (2013) who reported significant prediction errors when applied to diverse soil textures. The sensitivity of models to specific soil properties resulted in high variability and lower stability. The study by Xu et al. (2016) supports this observation, suggesting that incorporating particle size without a robust framework for integrating other soil characteristics can lead to increased uncertainty and limited applicability across diverse landscapes.

4.3 Assessing the adaptability of robust ANN model

The study evaluated the adaptability of an ANN model using BD data from 2004 to predict BD in 2009. Four key factors underscore the effectiveness of ANN models in BD prediction: 1) ANN autonomously learns and extract pertinent features from input data during training. This process allows ANN to capture complex patterns and relationships in data without explicit feature engineering from large datasets (Lecun et al., 2015), regularization techniques to bolster model generalization (Zhang et al., 2016). This approach allows ANN to effectively handle multidimensional data such as environmental and remote sensing indices (Figure 7(a)). Contrarily, Katuwal et al. (2020) encountered challenges and did not achieve satisfactory results when attempting to utilize only vis–NIR spectra for predicting BD.; 2) ANN adeptly capture nonlinear relationships inherent in BD data (Negiş, 2024; Erzin et al., 2008), crucial for accurate predictions across varying environmental conditions (Dragović, 2022) and agricultural lands (Abbaspour-Gilandeh et al., 2023).; 3) ANN ability to predict in the year 2004 and 2009 in the study area significantly enhances their reliability in BD prediction. Studies demonstrate ANN consistent performance over time (Abiodun et al., 2018), adapting well to datasets from different years or regions (Zhang, 2010). Unfortunately, the study by Hateffard et al. (2023) reported a poor result (R^2 = -0.746) when integrating ANN with NDVI and spectral bands.; 4) ANN demonstrate notable resilience to data variability and outliers, essential for maintaining stable predictions of BD (Khemis et al., 2022). This resilience enables effective handling of variations in soil characteristics and environmental conditions, enhancing reliability in long-term soil studies (Ünal et al., 2023). Despite previous studies by (Katuwal et al., 2020) and (Hateffard et al., 2023) failing to achieve accurate BD predictions, ANN have shown significant advancements in their predictive capabilities. In this study, we developed a robust ANN model for BD prediction using open-source remote sensing data, achieving high accuracy across different years. The effectiveness of ANN performance is supported by studies by (Li et al., 2013) and (Zhao et al., 2009).

© Author(s) 2025. CC BY 4.0 License.

4.4 Uncertainty and Variability in BD Prediction

This study evaluated the uncertainty and variability in BD predictions between 2004 and 2009 using the RS-based ANN model and traditional PTFs. Our findings revealed five key sources of uncertainty and variability that significantly impact the reliability of BD predictions, influencing their applicability across diverse landscapes. 1) Dependence on OC as the primary predictor emerged as a major source of uncertainty in both traditional PTFs and several ML models. Studies have shown that PTFs exhibit high sensitivity to OC leading to inconsistent predictions in soils with varying OC levels. This sensitivity can result in overfitting (Mcbratney et al., 2003) and poor generalizability when applied across different soil types and landscapes (Hou et al., 2024). Similarly, ML models such as Random Forest (RF) and XGBoost, which also depend heavily on OC (Chen et al., 2024), demonstrated increased variability and instability under varying conditions. The integration of diverse predictors, such as spectral indices and topographic data, as employed in the RS-ANN model, reduced uncertainty, indicating higher stability and improved prediction reliability (Jain and Zongker, 1997). 2) Temporal variations were another key source of uncertainty. While the RS-ANN model showed a moderate increase in mean BD, PTF models showed more drastic changes (11.01% and 10.34%, respectively). This discrepancy suggests that models relying heavily on OC tend to overestimate temporal changes in BD, resulting in less stable predictions over time. 3) Changes in standard deviation (SD) between 2004 and 2009 highlighted variations in spatial prediction accuracy. While the RS-ANN model showed a 41.23% reduction in SD, indicating more consistent BD values, other models experienced sharper decreases, raising concerns about potential overfitting. Spatial variability in BD predictions can be influenced by differences in soil structure and management practices, which are not adequately captured by traditional PTFs. 4) Skewness and kurtosis analyses revealed that the RS-ANN model improved from a highly skewed distribution in 2004 (skewness = -2.81, kurtosis = 15.37) to a more balanced distribution in 2009 (skewness = -0.58, kurtosis = -0.41). In contrast, PTFs continued to show high skewness and kurtosis, indicating persistent prediction errors for outliers. This reflects a lack of robustness in handling extreme BD values, which are critical for accurate soil assessments. 5) The RS-ANN model demonstrated broader applicability across diverse soil types and land uses compared to traditional PTFs and OC-dominant ML models. This adaptability is crucial in regions with heterogeneous soil properties, where multiple factors (e.g., soil moisture, texture, and topography) influence BD. Studies have shown that models integrating diverse input variables perform better in capturing complex soil dynamics and reducing prediction uncertainty.

5 Conclusion

This study developed a reliable BD prediction model by combining remote sensing data, environmental factors, and several machine learning techniques. The ANN model showed better performance than traditional PTFs and other machine learning models. The ANN model’s balanced use of predictors effectively captured spatial and temporal variability in BD, offering more stable and reliable predictions with lower uncertainty compared to the OC-dependent PTFs and ensemble-based models. The temporal analysis from 2004 to 2009 revealed a consistent increase in mean BD values with reduced variability, indicating the model's robustness for long-term monitoring. Additionally, the ANN model successfully represented the spatial heterogeneity of BD across Thailand, providing critical insights for soil management and sustainable land-use planning. ANN model, leveraging remote sensing data, proves to be a valuable tool for national-scale BD estimation, enabling more precise soil health monitoring, carbon accounting, and sustainable land management. Future research should focus on refining the temporal dynamics of the model would allow for more accurate monitoring of long-term soil changes, especially under varying climatic conditions and land-use practices. Integrating time-series analysis with climate projections could provide more accurate predictions of how BD and other soil properties evolve over time.

Data availability

The data and analyses that support these findings will be made available in response to a reasonable request but are not hosted in an online repository at this time in order to protect the privacy of growers.

Author contributions

SO: conceptualization, investigation, data curation, formal analysis, visualization, writing (original draft). ZS: writing (review and editing). AP: resources, writing (review and editing). ZA: resources, writing (review and editing).

Competing interests

The contact author has declared that none of the authors has any competing interests.

Acknowledgements

We thank to the Soil Research Survey and Research Division, Land Development Department, Thailand, for providing the valuable soil sampling data utilized in the development and validation of the soil bulk density estimation model.

References

Abbaspour-Gilandeh, Y., Abbaspour-Gilandeh, M., Babaie, H. A., and Shahgoli, G.: Modeling agricultural soil bulk density using artificial neural network and adaptive neuro-fuzzy inference system, Earth Science Informatics, 16, 57-65, 10.1007/s12145-022-00920-6, 2023.

Abdelbaki, A. M.: Evaluation of pedotransfer functions for predicting soil bulk density for U.S. soils, Ain Shams Engineering Journal, 9, 1611-1619, https://doi.org/10.1016/j.asej.2016.12.002, 2018.

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., and Arshad, H.: State-of-the-art in artificial neural network applications: A survey, Heliyon, 4, e00938, https://doi.org/10.1016/j.heliyon.2018.e00938, 2018.

Adams, W. A.: The effect of organic matter on the bulk and true densities of some uncultivated podzolic soils, Journal of Soil Science, 24, 10-17, https://doi.org/10.1111/j.1365-2389.1973.tb00737.x, 1973.

Aitkenhead, M. and Coull, M.: Mapping soil profile depth, bulk density and carbon stock in Scotland using remote sensing and spatial covariates, European Journal of Soil Science, 71, 553-567, https://doi.org/10.1111/ejss.12916, 2020.

Akpa, S. I. C., Ugbaje, S. U., Bishop, T. F. A., and Odeh, I. O. A.: Enhancing pedotransfer functions with environmental data for estimating bulk density and effective cation exchange capacity in a data-sparse situation, Soil Use and Management, 32, 644-658, https://doi.org/10.1111/sum.12310, 2016.

Al-Qinna, M. and Jaber, S.: Predicting Soil Bulk Density Using Advanced Pedotransfer Functions in an Arid Environment, Transactions of the ASABE (American Society of Agricultural and Biological Engineers), 56, 963-976, 10.13031/trans.56.9922, 2013.

Alexander, E. B.: Bulk Densities of California Soils in Relation to Other Soil Properties, Soil Science Society of America Journal, 44, 689-692, https://doi.org/10.2136/sssaj1980.03615995004400040005x, 1980.

Anne, N. J. P., Abd-Elrahman, A. H., Lewis, D. B., and Hewitt, N. A.: Modeling soil parameters using hyperspectral image reflectance in subtropical coastal wetlands, International Journal of Applied Earth Observation and Geoinformation, 33, 47-56, https://doi.org/10.1016/j.jag.2014.04.007, 2014.

Ao, Y., Li, H., Zhu, L., Ali, S., and Yang, Z.: The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling, Journal of Petroleum Science and Engineering, 174, 776-789, https://doi.org/10.1016/j.petrol.2018.11.067, 2019.

Atwood, T. B., Connolly, R. M., Almahasheer, H., Carnell, P. E., Duarte, C. M., Ewers Lewis, C. J., Irigoien, X., Kelleway, J. J., Lavery, P. S., and Macreadie, P. I.: Global patterns in mangrove soil carbon stocks and losses, Nature Climate Change, 7, 523-528, 2017.

Benites, V. M., Machado, P. L. O. A., Fidalgo, E. C. C., Coelho, M. R., and Madari, B. E.: Pedotransfer functions for estimating soil bulk density from existing soil survey reports in Brazil, Geoderma, 139, 90-97, https://doi.org/10.1016/j.geoderma.2007.01.005, 2007.

Bernoux, M., Cerri, C., Arrouays, D., Jolivet, C., and Volkoff, B.: Bulk Densities of Brazilian Amazon Soils Related to Other Soil Properties, Soil Science Society of America Journal, 62, 743-749, https://doi.org/10.2136/sssaj1998.03615995006200030029x, 1998.

Beutler, S., Pereira, M., Tassinari, W., Duarte de Menezes, M., Valladares, G., and Anjos, L.: Bulk Density Prediction for Histosols and Soil Horizons with High Organic Matter Content, Revista Brasileira de Ciência do Solo, 41, 10.1590/18069657rbcs20160158, 2017.

© Author(s) 2025. CC BY 4.0 License.

Botula, Y.-D., Nemes, A., Van Ranst, E., Mafuka, P., De Pue, J., and Cornelis, W. M.: Hierarchical Pedotransfer Functions to Predict Bulk Density of Highly Weathered Soils in Central Africa, Soil Science Society of America Journal, 79, 476-486, https://doi.org/10.2136/sssaj2014.06.0238, 2015.

Breiman, L.: Random Forests, Machine Learning, 45, 5-32, 10.1023/A:1010933404324, 2001.

Calhoun, F. G., Smeck, N. E., Slater, B. L., Bigham, J. M., and Hall, G. F.: Predicting bulk density of Ohio Soils from Morphology, Genetic Principles, and Laboratory Characterization Data, Soil Science Society of America Journal, 65, 811-819, https://doi.org/10.2136/sssaj2001.653811x, 2001.

Chen, S., Richer-de-Forges, A. C., Saby, N. P. A., Martin, M. P., Walter, C., and Arrouays, D.: Building a pedotransfer function for soil bulk density on regional dataset and testing its validity over a larger area, Geoderma, 312, 52-63, https://doi.org/10.1016/j.geoderma.2017.10.009, 2018.

Chen, S., Chen, Z., Zhang, X., Luo, Z., Schillaci, C., Arrouays, D., Richer-de-Forges, A., and Shi, Z.: European soil bulk density and organic carbon stock database using machine learning based pedotransfer function, 10.5194/essd-2023-493, 2024.

Cienciala, E., Exnerova, Z., Macku, J., and Henzlik, V.: Forest topsoil organic carbon content in Southwest Bohemia region, J. For. Sci.(Prague), 52, 387-398, 2006.

Curtis, R. O. and Post, B. W.: Estimating Bulk Density from Organic-Matter Content in Some Vermont Forest Soils, Soil Science Society of America Journal, 28, 285-286, https://doi.org/10.2136/sssaj1964.03615995002800020044x, 1964.

Davidson, E. A. and Janssens, I. A.: Temperature sensitivity of soil carbon decomposition and feedbacks to climate change, Nature, 440, 165-173, 10.1038/nature04514, 2006.

De Vos, B., Van Meirvenne, M., Quataert, P., Deckers, J., and Muys, B.: Predictive Quality of Pedotransfer Functions for Estimating Bulk Density of Forest Soils, Soil Science Society of America Journal, 69, 500-510, https://doi.org/10.2136/sssaj2005.0500, 2005.

Delacre, M., Lakens, D., and Leys, C.: Why Psychologists Should by Default Use Welch's t-test Instead of Student's t-test, International Review of Social Psychology, 30, 92, 10.5334/irsp.82, 2017.

Demir, Z., Özdemir, N., and Bülbül, E.: Relationships between some soil properties and bulk density under different land use, Soil Studies, 11, 43-50, 10.21657/soilst.1218353, 2022.

Dexter, A. R.: Soil physical quality: Part I. Theory, effects of soil texture, density, and organic matter, and effects on root growth, Geoderma, 120, 201-214, https://doi.org/10.1016/j.geoderma.2003.09.004, 2004.

Do, M.-T. T., Van, L. N., Le, X.-H., Nguyen, G. V., Yeon, M., and Lee, G.: National variability in soil organic carbon stock predictions: Impact of bulk density pedotransfer functions, International Soil and Water Conservation Research, https://doi.org/10.1016/j.iswcr.2024.04.002, 2024.

Dragović, S.: Artificial neural network modeling in environmental radioactivity studies – A review, Science of The Total Environment, 847, 157526, https://doi.org/10.1016/j.scitotenv.2022.157526, 2022.

Drew, L. A.: Bulk density estimation based on organic matter content of some Minnesota soils, 1973.

© Author(s) 2025. CC BY 4.0 License.

Erzin, Y., Rao, B. H., and Singh, D. N.: Artificial neural network models for predicting soil thermal resistivity, International Journal of Thermal Sciences, 47, 1347-1358, https://doi.org/10.1016/j.ijthermalsci.2007.11.001, 2008.

Eschner, A. R., Jones, B., and Moyle, R.: Physical properties of 134 soils in six northeastern states, Station Paper NE-89. Upper Darby, PA: US Department of Agriculture, Forest Service, Northeastern Forest Experiment Station. 11 p., 89, 1957.

Faber, N. M.: Estimating the uncertainty in estimates of root mean square error of prediction: application to determining the size of an adequate test set in multivariate calibration, Chemometrics and Intelligent Laboratory Systems, 49, 79-89, https://doi.org/10.1016/S0169-7439(99)00027-1, 1999.

FAO: Standard operating procedure for soil bulk density by cylinder method, Food and Agriculture Organization of the United Nations Rome, Italy, https://doi.org/10.4060/cc7568en, 2023.

Federer, C. A.: Nitrogen Mineralization and Nitrification: Depth Variation in Four New England Forest Soils, Soil Science Society of America Journal, 47, 1008-1014, https://doi.org/10.2136/sssaj1983.03615995004700050034x, 1983.

Galle, N. J., Brinton, W., Vos, R., Basu, B., Duarte, F., Collier, M., Ratti, C., and Pilla, F.: Correlation of WorldView-3 spectral vegetation indices and soil health indicators of individual urban trees with exceptions to topsoil disturbance, City and Environment Interactions, 11, 100068, https://doi.org/10.1016/j.cacint.2021.100068, 2021.

Gao, B.-C.: A Normalized Difference Water Index for Remote Sensing of Vegetation Liquid Water from Space, Remote Sensing of Environment, 58, 257-266, 10.1016/S0034-4257(96)00067-3, 1996.

Ghaderi, A., Abbaszadeh Shahri, A., and Larsson, S.: An artificial neural network based model to predict spatial soil type distribution using piezocone penetration test data (CPTu), Bulletin of Engineering Geology and the Environment, 78, 4579-4588, 10.1007/s10064-018-1400-9, 2019.

Grigal, D. F., Brovold, S. L., Nord, W. S., and Ohmann, L. F.: Bulk density of surface soils and peat in the north central united states, Canadian Journal of Soil Science, 69, 895-900, 10.4141/cjss89-092, 1989.

Guo, L., Sun, X., Fu, P., Shi, T., Dang, L., Chen, Y., Linderman, M., Zhang, G., Zhang, Y., Jiang, Q., Zhang, H., and Zeng, C.: Mapping soil organic carbon stock by hyperspectral and time-series multispectral remote sensing images in low-relief agricultural areas, Geoderma, 398, 115118, https://doi.org/10.1016/j.geoderma.2021.115118, 2021.

Hallett, S. H., Hollis, J. M., and Keay, C. A.: Derivation and Evaluation of a set of Pedogenically-based Empirical Algorithms for Predicting Bulk Density in British Soils., 1998.

Han, G., Zhang, G.-L., Gong, Z.-T., and Wang, G.-F.: Pedotransfer Functions for Estimating Soil Bulk Density in China, Soil Science, 177, 158–164, 10.1097/SS.0b013e31823fd493, 2012.

Harrison, A. F. and Bocock, K. L.: Estimation of Soil Bulk-Density from Loss-on-Ignition Values, Journal of Applied Ecology, 18, 919-927, 10.2307/2402382, 1981.

Hateffard, F., Szatmári, G., and Tibor, N.: Applicability of machine learning models for predicting soil organic carbon content and bulk density under different soil conditions, Soil Science Annual, 74, 1-11, 10.37501/soilsa/165879, 2023.

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., and Gräler, B.: Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables, PeerJ, 6, e5518, 10.7717/peerj.5518, 2018.

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., and Kempen, B.: SoilGrids250m: Global gridded soil information based on machine learning, PLOS ONE, 12, e0169748, 10.1371/journal.pone.0169748, 2017.

Heuscher, S. A., Brandt, C. C., and Jardine, P. M.: Using Soil Physical and Chemical Properties to Estimate Bulk Density, Soil Science Society of America Journal, 69, 51-56, https://doi.org/10.2136/sssaj2005.0051a, 2005.

Holliday, V. T.: Methods of soil analysis, part 1, physical and mineralogical methods (2nd edition), A. Klute, Ed., 1986, American Society of Agronomy, Agronomy Monographs 9(1), Madison, Wisconsin, 1188 pp., $60.00, Geoarchaeology, 5, 87-89, https://doi.org/10.1002/gea.3340050110, 1990.

Hollis, J. M., Hannam, J., and Bellamy, P. H.: Empirically-derived pedotransfer functions for predicting bulk density in European soils, European Journal of Soil Science, 63, 96-109, https://doi.org/10.1111/j.1365-2389.2011.01412.x, 2012.

Honeysett, J. L. and Ratkowsky, D. A.: The use of ignition loss to estimate bulk density of forest soils, Journal of Soil Science, 40, 299-308, https://doi.org/10.1111/j.1365-2389.1989.tb01275.x, 1989.

Hong, S. Y., Minasny, B., Han, K. H., Kim, Y., and Lee, K.: Predicting and mapping soil available water capacity in Korea, PeerJ, 1, e71, 10.7717/peerj.71, 2013.

Hossain, M. F., Chen, W., and Zhang, Y.: Bulk density of mineral and organic soils in the Canada's arctic and sub-arctic, Information Processing in Agriculture, 2, 183-190, https://doi.org/10.1016/j.inpa.2015.09.001, 2015.

Hou, C.-J., Lu, Y.-H., Tseng, Y.-C., Tsai, Y.-C., Huang, W.-L., and Juang, K.-W.: Using a comprehensive model for cropland types in relationships between soil bulk density and organic carbon to predict site-specific carbon stocks, Journal of Soils and Sediments, 24, 2584-2598, 10.1007/s11368-024-03829-3, 2024.

Huete, A., Didan, K., Miura, T., Rodriguez, E., Gao, X., and Ferreira, L. G.: Overview of the Radiometric and Biophysical Performance of the MODIS Vegetation Indices, Remote Sensing of Environment, 83, 195-213, 10.1016/S0034-4257(02)00096-2, 2002.

Huf dos Reis, A., Teixeira, W., Fontana, A., Barros, A., Victoria, D., Vasques, G., Samuel-Rosa, A., Vasconcelos, M., and Monteiro, J.: Hierarchical pedotransfer functions for predicting bulk density in Brazilian soils, Scientia Agricola, 81, 10.1590/1678-992x-2022-0255, 2024.

Huntington, T. G., Johnson, C. E., Johnson, A. H., Siccama, T. G., and Rran, D. F.: CARBON, ORGANIC MATTER, AND BULK DENSITY RELATIONSHIPS IN A FORESTED SPODOSOL, Soil Science, 148, 380-386, 1989.

Jain, A. and Zongker, D.: Feature selection: evaluation, application, and small sample performance, IEEE Transactions on Pattern Analysis and Machine Intelligence, 19, 153-158, 10.1109/34.574797, 1997.

Jalabert, S., Martin, M., Renaud, J.-P., Boulonne, L., Jolivet, C., Montanarella, L., and Arrouays, D.: Estimating forest soil bulk density using boosted regression modeling, Soil Use and Management, 26, 516-528, 10.1111/j.1475-2743.2010.00305.x, 2010.

James, G., Witten, D., Hastie, T., Tibshirani, R., and Taylor, J.: An introduction to statistical learning : with applications in R, New York : Springer, [2013] ©20132013.

Jeffrey, D. W.: A Note on the use of Ignition Loss as a Means for the Approximate Estimation of Soil Bulk Density, Journal of Ecology, 58, 297-299, 10.2307/2258183, 1970.

Katuwal, S., Knadel, M., Norgaard, T., Moldrup, P., Greve, M. H., and de Jonge, L. W.: Predicting the dry bulk density of soils across Denmark: Comparison of single-parameter, multi-parameter, and vis–NIR based models, Geoderma, 361, 114080, https://doi.org/10.1016/j.geoderma.2019.114080, 2020.

Kaur, R., Kumar, S., and Gurung, H.: A pedo-transfer function (PTF) for estimating soil bulk density from basic soil data and its comparison with existing PTFs, Soil Research, 40, 847-858, 10.1071/SR01023, 2002.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y.: LightGBM: A Highly Efficient Gradient Boosting Decision Tree, Neural Information Processing Systems,

Keller, T. and Håkansson, I.: Estimation of reference bulk density from soil particle size distribution and soil organic matter content, Geoderma, 154, 398-406, https://doi.org/10.1016/j.geoderma.2009.11.013, 2010.

Kesbi, F. G., Mianji, G. R., Honarvar, M., and Javaremi, A. N.: Tuning and Application of Random Forest Algorithm in Genomic Evaluation,

Khemis, C., Abrougui, K., Mohammadi, A., Karim, G., Dorbolo, S., Mercatoris, B., Mutuku, E., Cornelis, W., and Chehaibi, S.: Development of Artificial Neural Networks to Predict the Effect of Tractor Speed on Soil Compaction Using Penetrologger Test Results, Processes, 10, 1109, 10.3390/pr10061109, 2022.

Kim, D., Kim, T., Jeon, J., and Son, Y.: Soil-Surface-Image-Feature-Based Rapid Prediction of Soil Water Content and Bulk Density Using a Deep Neural Network, Applied Sciences, 13, 4430, 2023.

Kobal, M., Urbancic, M., Potocic, N., De Vos, B., and Simoncic, P.: Pedotransfer functions
for bulk density estimation of forest soils, Forestry Soc 19–27, 2011.

LeCun, Y., Bengio, Y., and Hinton, G.: Deep learning, Nature, 521, 436-444, 10.1038/nature14539, 2015.

Leonaviciute, N.: Predicting soil bulk and particle densities by pedotransfer functions from existing soil data in Lithuania, Geografijos metraštis, 33, 7-330, 2000.

Li, Q.-q., Yue, T.-x., Wang, C.-q., Zhang, W.-j., Yu, Y., Li, B., Yang, J., and Bai, G.-c.: Spatially distributed modeling of soil organic matter across China: An application of artificial neural network approach, CATENA, 104, 210-218, https://doi.org/10.1016/j.catena.2012.11.012, 2013.

Li, S., Li, Q.-q., Wang, C.-q., Li, B., Gao, X.-s., Li, Y.-d., and Wu, D.-y.: Spatial variability of soil bulk density and its controlling factors in an agricultural intensive area of Chengdu Plain, Southwest China, Journal of Integrative Agriculture, 18, 290-300, https://doi.org/10.1016/S2095-3119(18)61930-6, 2019.

Liu, E., Liu, J., Yu, K., Wang, Y., and He, P.: A hybrid model for predicting spatial distribution of soil organic matter in a bamboo forest based on general regression neural network and interative algorithm, Journal of Forestry Research, 31, 1673-1680, 10.1007/s11676-019-00980-3, 2020.

https://doi.org/10.5194/egusphere-2025-2360
Preprint. Discussion started: 19 June 2025
© Author(s) 2025. CC BY 4.0 License.

Liu, S., Lu, L., Wang, F., Han, B., Ou, L., Gao, X., Luo, Y., Huo, W., and Zeng, Q.: Building a predictive model for hypertension related to environmental chemicals using machine learning, Environmental Science and Pollution Research, 31, 4595-4605, 10.1007/s11356-023-31384-w, 2024.

Lupikis, A., Bardule, A., Lazdiņš, A., Stola, J., and Butlers, A.: Carbon stock changes in drained arable organic soils in Latvia: Results of a pilot study, Agronomy Research, 15, 788-798, 2017.

Manrique, L. A. and Jones, C. A.: Bulk Density of Soils in Relation to Soil Physical and Chemical Properties, Soil Science Society of America Journal, 55, 476-481, https://doi.org/10.2136/sssaj1991.03615995005500020030x, 1991.

Manuel Rodríguez-Rastrero, A. O.-M.: Carbon Stock Assessment in Gypsum-Bearing Soils: The Role of Subsurface Soil Horizons, Earth, MDPI, 3, 1, 2022.

McBratney, A. B., Mendonça Santos, M. L., and Minasny, B.: On digital soil mapping, Geoderma, 117, 3-52, https://doi.org/10.1016/S0016-7061(03)00223-4, 2003.

Memon, N., Patel, S. B., and Patel, D. P.: Comparative Analysis of Artificial Neural Network and XGBoost Algorithm for PolSAR Image Classification, Pattern Recognition and Machine Intelligence, Cham, 2019//, 452-460,

Men, M., Peng, Z., Hao, X., and Yu, Z.: Investigation on Pedotransfer function for estimating soil bulk density in Hebei province, Chin. J. Soil Sci, 1, 20, 2008.

Minasny, B. and Hartemink, A. E.: Predicting soil properties in the tropics, Earth-Science Reviews, 106, 52-62, https://doi.org/10.1016/j.earscirev.2011.01.005, 2011.

Mockus, J.: The Bayesian approach to global optimization, System Modeling and Optimization: Proceedings of the 10th IFIP Conference New York City, USA, August 31–September 4, 1981, 473-481,

Müller, K. R., Smola, A. J., Rätsch, G., Schölkopf, B., Kohlmorgen, J., and Vapnik, V.: Predicting time series with support vector machines, Artificial Neural Networks — ICANN'97, Berlin, Heidelberg, 1997//, 999-1004,

Nanko, K., Ugawa, S., Hashimoto, S., Imaya, A., Kobayashi, M., Sakai, H., Ishizuka, S., Miura, S., Tanaka, N., Takahashi, M., and Kaneko, S.: A pedotransfer function for estimating bulk density of forest soil in Japan affected by volcanic ash, Geoderma, 213, 36-45, https://doi.org/10.1016/j.geoderma.2013.07.025, 2014.

Nasta, P., Palladino, M., Sica, B., Pizzolante, A., Trifuoggi, M., Toscanesi, M., Giarra, A., D'Auria, J., Nicodemo, F., Mazzitelli, C., Lazzaro, U., Di Fiore, P., and Romano, N.: Evaluating pedotransfer functions for predicting soil bulk density using hierarchical mapping information in Campania, Italy, Geoderma Regional, 21, e00267, https://doi.org/10.1016/j.geodrs.2020.e00267, 2020.

Negiş, H.: Using Models and Artificial Neural Networks to Predict Soil Compaction Based on Textural Properties of Soils under Agriculture, Agriculture, 14, 47, 2024.

Padarian, J., Minasny, B., and McBratney, A. B.: Machine learning and soil sciences: a review aided by machine learning tools, SOIL, 6, 35-52, 10.5194/soil-6-35-2020, 2020.

Panagos, P., De Rosa, D., Liakos, L., Labouyrie, M., Borrelli, P., and Ballabio, C.: Soil bulk density assessment in Europe, Agriculture, Ecosystems & Environment, 364, 108907, https://doi.org/10.1016/j.agee.2024.108907, 2024.

Patil, N. G. and Singh, S. K.: Pedotransfer Functions for Estimating Soil Hydraulic Properties: A Review, Pedosphere, 26, 417-430, https://doi.org/10.1016/S1002-0160(15)60054-6, 2016.

Perie, C. and Ouimet, R.: Organic carbon, organic matter and bulk density relationships in boreal forest soils, Canadian Journal of Soil Science, 88, 315-325, 10.4141/cjss06008, 2008.

Pittman, R. and Hu, B.: ESTIMATION OF SOIL BULK DENSITY AND CARBON USING MULTI-SOURCE REMOTELY SENSED DATA, ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, V-3-2020, 541-548, 10.5194/isprs-annals-V-3-2020-541-2020, 2020.

Poggio, L. and Gimona, A.: Assimilation of optical and radar remote sensing data in 3D mapping of soil properties over large areas, Science of The Total Environment, 579, 1094-1110, https://doi.org/10.1016/j.scitotenv.2016.11.078, 2017.

Post, W. M. and Kwon, K. C.: Soil carbon sequestration and land-use change: processes and potential, Global Change Biology, 6, 317-327, https://doi.org/10.1046/j.1365-2486.2000.00308.x, 2000.

Prevost, M.: Predicting Soil Properties from Organic Matter Content following Mechanical Site Preparation of Forest Soils, Soil Science Society of America Journal, 68, 943-949, https://doi.org/10.2136/sssaj2004.9430, 2004.

Rawls, W. J. and Brakensiek, D. L.: Estimating Soil Water Retention from Soil Properties, Journal of the Irrigation and Drainage Division, 108, 166-171, doi:10.1061/JRCEA4.0001383, 1982.

Rawls, W. J., Nemes, A., and Pachepsky, Y.: Effect of soil organic carbon on soil hydraulic properties, in: Developments in Soil Science, Elsevier, 95-114, https://doi.org/10.1016/S0166-2481(04)30006-1, 2004.

Reidy, B., Simo, I., Sills, P., and Creamer, R. E.: Pedotransfer functions for Irish soils – estimation of bulk density (ρb) per horizon type, SOIL, 2, 25-39, 10.5194/soil-2-25-2016, 2016.

Rodriguez, J. D., Perez, A., and Lozano, J. A.: Sensitivity analysis of k-fold cross validation in prediction error estimation, IEEE transactions on pattern analysis and machine intelligence, 32, 569-575, 2009.

Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence, 1, 206-215, 10.1038/s42256-019-0048-x, 2019.

Ruehlmann, J. and Körschens, M.: Calculating the Effect of Soil Organic Matter Concentration on Soil Bulk Density, Soil Science Society of America Journal, 73, 876-885, https://doi.org/10.2136/sssaj2007.0149, 2009.

Saini, G. R.: Organic Matter as a Measure of Bulk Density of Soil, Nature, 210, 1295-1296, 10.1038/2101295a0, 1966.

Salehi Hikouei, I., Kim, S. S., and Mishra, D. R.: Machine-Learning Classification of Soil Bulk Density in Salt Marsh Environments, Sensors, 21, 4408, 2021.

Schillaci, C., Perego, A., Valkama, E., Märker, M., Saia, S., Veronesi, F., Lipani, A., Lombardo, L., Tadiello, T., Gamper, H. A., Tedone, L., Moss, C., Pareja-Serrano, E., Amato, G., Kühl, K., Dămătîrcă, C., Cogato, A., Mzid, N., Eeswaran, R., Rabelo, M., Sperandio, G., Bosino, A., Bufalini, M., Tunçay, T., Ding, J., Fiorentini, M., Tiscornia, G., Conradt, S., Botta, M., and Acutis, M.: New pedotransfer approaches to predict soil bulk density using WoSIS soil data and environmental covariates in Mediterranean agro-ecosystems, Science of The Total Environment, 780, 146609, https://doi.org/10.1016/j.scitotenv.2021.146609, 2021.

https://doi.org/10.5194/egusphere-2025-2360
Preprint. Discussion started: 19 June 2025
© Author(s) 2025. CC BY 4.0 License.

Sevastas, S., Gasparatos, D., Botsis, D., Siarkos, I., Diamantaras, K. I., and Bilas, G.: Predicting bulk density using pedotransfer functions for soils in the Upper Anthemountas basin, Greece, Geoderma Regional, 14, e00169, https://doi.org/10.1016/j.GEODRS.2018.e00169, 2018.

Song, G., Li, L., Pan, G., and Zhang, Q.: Topsoil organic carbon storage of China and its loss by cultivation, Biogeochemistry, 74, 47-62, 10.1007/s10533-004-2222-3, 2005.

Sreenivasulu, T. and Rayalu, G. M.: Enhanced PM2.5 prediction in Delhi using a novel optimized STL-CNN-BILSTM-AM hybrid model, Asian Journal of Atmospheric Environment, 18, 25, 10.1007/s44273-024-00048-7, 2024.

Tamminen, P. and Starr, M.: Bulk density of forested mineral soils, Silva Fennica. 1994. 28(1): 53–60., 28, 10.14214/sf.a9162, 1994.

Tao, F., Huang, Y., Hungate, B., Manzoni, S., Frey, S., Schmidt, M., Reichstein, M., Carvalhais, N., Ciais, P., Jiang, L., Lehmann, J., Wang, Y., Houlton, B., Ahrens, B., Mishra, U., Hugelius, G., Hocking, T., Lu, X., Shi, Z., and Luo, Y.: Microbial carbon use efficiency promotes global soil carbon storage, Nature, 618, 1-5, 10.1038/s41586-023-06042-3, 2023.

Tomasella, J. and Hodnett, M.: Estimating soil water retention characteristics from limited data in Brazilian Amazonia, Soil Science, 163, 10.1097/00010694-199803000-00003, 1998.

Tranter, G., Minasny, B., Mcbratney, A. B., Murphy, B., Mckenzie, N. J., Grundy, M., and Brough, D.: Building and testing conceptual and empirical models for predicting soil bulk density, Soil Use and Management, 23, 437-443, https://doi.org/10.1111/j.1475-2743.2007.00092.x, 2007.

Tremblay, S., Ouimet, R., and Houle, D.: Prediction of organic carbon content in upland forest soils of Quebec, Canada, Canadian Journal of Forest Research, 32, 903-914, 10.1139/x02-023, 2002.

Tsui, C.-C., Tsai, C.-C., and Chen, Z.-S.: Soil organic carbon stocks in relation to elevation gradients in volcanic ash soils of Taiwan, Geoderma, 209-210, 119-127, 10.1016/j.geoderma.2013.06.013, 2013.

Ünal, İ., Kabaş, Ö., and Sözer, S.: Comparison of two different artificial neural network models for prediction of soil penetration resistance, Journal of Agricultural Engineering, 10.4081/jae.2023.1550, 2023.

Valzano F, Murphy BW, and T., K.: The impact of tillage on changes in soil carbon density with special emphasis on Australian conditions. , National carbon Accounting System, Australian Greenhouse Office, Canberra., 2005.

Vasiliniuc, I. and Patriche, C.: Validating Soil Bulk Density Pedotransfer Functions Using A Romanian Dataset, Carpathian Journal of Earth and Environmental Sciences, 10, 225-236, 2015.

Williams, R.: Relationships between the composition of soils and physical measurements made on them, Rothamsted Experimental Station Report, 1970, 5-35, 1971.

Wong, T.-T. and Yeh, P.-Y.: Reliable accuracy estimates from k-fold cross validation, IEEE Transactions on Knowledge and Data Engineering, 32, 1586-1594, 2019.

Wu, H., Guo, Z., and Peng, C.: Distribution and storage of soil organic carbon in China, Global biogeochemical cycles, 17, 2003.

https://doi.org/10.5194/egusphere-2025-2360
Preprint. Discussion started: 19 June 2025
© Author(s) 2025. CC BY 4.0 License.

Xu, L., He, N., and Yu, G.: Methods of evaluating soil bulk density: Impact on estimating large scale soil organic carbon storage, CATENA, 144, 94-101, https://doi.org/10.1016/j.catena.2016.05.001, 2016.

Yan, C., Shen, X., and Guo, F.: An improved support vector regression using least squares method, Structural and Multidisciplinary Optimization, 57, 2431-2445, 10.1007/s00158-017-1871-5, 2018.

Yang, R.-M. and Guo, W.-W.: Modelling of soil organic carbon and bulk density in invaded coastal wetlands using Sentinel-1 imagery, International Journal of Applied Earth Observation and Geoinformation, 82, 101906, https://doi.org/10.1016/j.jag.2019.101906, 2019.

Yang, Y., Mohammat, A., Feng, J., Zhou, R., and Fang, J.: Storage, patterns and environmental controls of soil organic carbon in China, Biogeochemistry, 84, 131-141, 10.1007/s10533-007-9109-z, 2007.

Yi, X., Li, G., and Yin, Y.: Pedotransfer functions for estimating soil bulk density: A case study in the three-river headwater region of qinghai province, China, Pedosphere, 26, 362-373, https://doi.org/10.1016/S1002-0160(15)60049-2, 2016.

Zhang, L., Zhang, L., and Du, B.: Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art, IEEE Geoscience and Remote Sensing Magazine, 4, 22-40, 10.1109/MGRS.2016.2540798, 2016.

Zhang, W.: Computational Ecology: Artificial Neural Networks and Their Applications (World Scientific, Singapore, 2010), 10.1142/7436, 2010.

Zhao, Y., Lin, L., and Schlarb, A. K.: Long Short-Term Memory Networks for the Automated Identification of the Stationary Phase in Tribological Experiments, Lubricants, 12, 423, 2024.

Zhao, Z., Chow, T. L., Rees, H. W., Yang, Q., Xing, Z., and Meng, F.-R.: Predict soil texture distributions using an artificial neural network model, Computers and Electronics in Agriculture, 65, 36-48, https://doi.org/10.1016/j.compag.2008.07.008, 2009.

Zinke, P. J., Millemann, R. E., and Boden, T. A.: Worldwide organic soil carbon and nitrogen data, Carbon Dioxide Information Center, Environmental Sciences Division, Oak …1986.