European soil bulk density and organic carbon stock database using machine learning based pedotransfer function

Songchao Chen[1,2#], Zhongxing Chen[1,2#], Xianglin Zhang[2], Zhongkui Luo[2], Calogero Schillaci[3], Dominique Arrouays[4], Anne C. Richer-de-Forges[4], Zhou Shi[2]

[1]ZJU-Hangzhou Global Scientific and Technological Innovation Center, Zhejiang University, Hangzhou, 311215, China
[2]College of Environmental and Resource Sciences, Zhejiang University, Hangzhou, 310058, China
[3]European Commission, Joint Research Centre, Ispra, 21026, Italy
[4]INRAE, Info&Sols, Orléans, 45075, France

# These authors contributed equally.

Correspondence to: Zhou Shi (Email: shizhou@zju.du.cn)

Abstract. Soil bulk density (BD) serves as a fundamental indicator of soil health and quality, exerting a significant influence on critical factors such as plant growth, nutrient availability, and water retention. Due to its limited availability in soil databases, the application of pedotransfer functions (PTFs) has emerged as a potent tool for predicting BD using other easily measurable soil properties, while the impact of these PTFs' accuracy on soil organic carbon (SOC) stock calculation has been rarely explored. In this study, we proposed an innovative local modelling approach for predicting BD across Europe using the recently released BD data from the LUCAS Soil 2018 (0-20 cm). Our approach involved a combination of neighbour sample search, Forward Recursive Feature Selection (FRFS) and Random Forest (RF) model (local-RF$_{\text{FRFS}}$). The results showed that local-RF$_{\text{FRFS}}$ had a good performance in predicting BD (R^2 of 0.58, RMSE of 0.19 g cm^{-3}), surpassing the traditional PTFs (R^2 of 0.40-0.45, RMSE of 0.22 g cm^{-3}) and global PTFs using RF with and without FRFS (R^2 of 0.56-0.57, RMSE of 0.19 g cm^{-3}). Interestingly, we found the best traditional PTF (R^2=0.84, RMSE=1.39 kg m^{-2}) performed close to the local-RF$_{\text{FRFS}}$ (R^2=0.85, RMSE=1.32 kg m^{-2}) in SOC stock calculation using BD predictions. However, the local-RF$_{\text{FRFS}}$ still performed better (ΔR^2>0.2 and ΔRMSE>0.1 g cm^{-3}) for soil samples with low SOC stock (<3 kg m^{-2}). Therefore, we suggest that the local-RF$_{\text{FRFS}}$ is a promising method for BD prediction while traditional PTFs would be more efficient when BD is subsequently utilized for calculating SOC stock. Finally, we produced two BD and SOC stocks datasets (18,945 and 15,389 soil samples) for LUCAS Soil 2018 using the best traditional PTF and local-RF$_{\text{FRFS}}$, respectively. This dataset is archived from the Zenodo platform at https://zenodo.org/records/10211884 (Chen et al., 2023).The outcomes of this study present a meaningful advancement in enhancing the predictive accuracy of BD, and the resultant BD and SOC stock datasets across the Europe enable more precise soil hydrological and biological modelling.

1 Introduction

Soil plays a pivotal role in supporting ecosystems and sustaining life on our planet (Rabot et al., 2018). Its physical properties are crucial for various disciplines such as agriculture, environmental science, and land management. Among these properties, soil bulk density (BD) holds particular significance as it serves as a fundamental indicator of soil health, structure, and water holding capacity. BD directly influences vital factors like plant growth, nutrient availability, and overall soil quality (Dam et al., 2005; Chen et al., 2018; Schillaci et al., 2021). Additionally, BD plays a crucial role in calculating SOC storage, making it even more essential in soil studies. Nonetheless, the uncertainty in SOC stock estimates arises due to the variations in methods used to substitute for missing BD data (Benites et al., 2007; Dawson and Smith, 2007; Wiesmeier et al., 2012). It is important to acknowledge that BD exhibits considerable variations across different geographical regions due to factors like diverse soil types, climate conditions, vegetation cover, and land cover patterns (Hollis et al., 2011; Lark et al., 2014; Li et al., 2019). These regional disparities underscore the need for a comprehensive understanding of BD in soil research and its implications for various aspects of ecosystem functioning and management.

Characterizing the spatial distribution of BD across a diverse and extensive continent like Europe presents a complex challenge (Chen et al., 2018; Nasta et al., 2020; Palladino et al., 2022). Traditional soil sampling and laboratory analyses have been time-consuming, costly, and impractical at a large scale (Makovníková et al, 2017). In response to this challenge, the development of pedotransfer functions (PTFs) has emerged as a powerful approach (Van Looy et al., 2017). PTFs are mathematical models that estimate soil properties, such as BD, based on readily available and easily measurable soil data (e.g., SOC, clay, silt, sand). These functions serve as invaluable tools for predicting soil properties at unvisited locations, facilitating regional soil mapping, and enhancing our understanding of soil dynamics across vast areas (Chen et al., 2018; Schillaci et al., 2021; Palladino et al., 2022). Furthermore, the incorporation of globally available covariates, such as topography and land cover, showed a promise in enhancing the effectiveness and applicability of PTFs for gap-filling of BD (Ramcharan et al., 2017; Bondi et al., 2018; Patton et al., 2019).

In the early stage, PTFs predominantly employed regression techniques due to their simplicity (Gupta and Larson, 1979; Rawls and Brakensiek, 1985). However, with advancements in science and technology, a wide range of models have been developed for deriving PTFs, particularly for continuous predicted variables. These methods encompass linear regression (LM), generalized linear models (GLM), generalized additive models (GAM), regression trees, neural networks, support vector machines, and random forests (RF) (Van Looy et al., 2017). The utilization of these advanced techniques has substantially improved the accuracy of BD prediction (Table 1).

Table 1: Summary of previous studies on using PTFs for BD prediction

ID	Scale	Sample size	Model	R^2	Reference
1	Landscape	164	Naive-BN	0.26	Taalab et al. (2015)
			Hierarchical-BN	0.42	
2	National	2,462	MLR	0.41	Katuwal et al. (2020)

			RF	0.62	
			RR	0.60	
			ANN	0.61	
3	National	1,357	GBM	0.53	Chen et al. (2018)
4	Regional	169	k-NN	0.32	Ghehi et al. (2012)
			BRT	0.30	
5	National	485	GBM	0.67	Jalabert et al., 2010)
6	Regional	495	ANN	0.71	Yi et al. (2016)
			MLR	0.63	
7	National	188	MLR	0.21	Schillaci et al. (2021)
			MLR-BS	0.38	
			ANN	0.48	

MLR, multiple linear regression; RF, random forest; RR, regression rules; ANN, artificial neural networks; GBM, generalized boosted models; BRT, boosted regression trees; BN, Bayesian network; k-NN, k-nearest neighbour; MLR-BS, multiple linear regression (stepwise variable selection)

PTFs have emerged as an alternative approach to address the scarcity of BD data (Van Looy et al., 2017). They have been implemented and tested in diverse regions and countries, providing a practical and cost-effective means for predicting BD using readily available soil properties. However, it is noteworthy that the majority of previous studies utilizing machine learning (ML) and PTFs for BD prediction have been conducted at regional or national scales, with limited research focusing on the intercontinental scale (Taalab et al., 2015; Shiri et al., 2017; Katuwal et al., 2020). Despite the accomplishments of PTFs in BD estimation, a gap emerges when transitioning to global modelling (a fixed model to predict all the unknown samples) endeavours. The reliance on global models, while useful in capturing broad patterns, often faces constraints in delivering accurate predictions at finer scales (Gupta et al., 2018). These global models may fail to account for the nuanced spatial and environmental variations that play a pivotal role in determining BD across different landscapes. While numerous studies have harnessed ML based PTFs (ML-PTFs) to improve the model performance for BD at national and regional levels, the expansion of these methodologies to encompass continental contexts remains relatively limited (Nasta et al., 2020; Schillaci et al., 2021; Palladino et al., 2022). This gap underscores the need for a modelling approach that bridges the gap between broad-scale global modelling and context-specific requirements of diverse regions and ecosystems. This is where the concept of local modelling steps in. The local model adopts a dynamic modelling strategy: it firstly selected a part of similar samples close to each unknown sample in the predictor space, then it fits a predictive model using the selected similar samples (not the whole data). Since the selected similar samples vary for each unknown sample, the corresponding local model is different from others. Local modelling strategy enables the consideration of environmental relevance by clustering data under comparable environmental conditions, which aids in constructing specialized PTFs that capture soil-environment relationships (Nocita et al., 2014; Chen et al., 2018). Thus, there is a compelling need for further investigations and developments in local modelling techniques to improving BD predictions. Furthermore, despite of the widely use of PTFs for predicting BD in SOC stock calculation from continental to global scales, how the accuracy of PTFs based BD prediction impacts the quality of SOC stock remains poorly explored (Cotrufo et al., 2019; Augusto and Boč et al., 2022; Luo et al., 2022; De Rosa et al., 2023).

To address aforementioned issues, we investigated RF model in combination with variable selection and local modelling strategy, to evaluate the potential of different PTFs in BD prediction as well as SOC stock calculation. The main objectives of this study are as follows:

- Compare the predictive accuracy of traditional PTFs (T-PTFs) and ML-PTFs for BD prediction;
- Evaluate the potential of local modelling strategies for BD prediction;
- Investigate the impact of PTFs-based BD prediction on the accuracy of SOC stocks calculation.

2 Materials and methods

2.1 Soil data

The soil data were compiled from the Land Use and Coverage Area Frame Survey Soil (LUCAS Soil) campaigns conducted in 2009, 2015 and 2018 (Fernández-Ugalde et al., 2022; Panagos et al., 2022). The survey encompassed a stratified random sampling approach, which identified approximately 20,000 soil sampling locations across the European Union and the United Kingdom for each campaign. At each sampling site (circle of 4 m diameter plot), 5 soil samples (0–20 cm) were collected after the removal of the litter layer, and the land cover (LC) was recorded. These samples were then combined into a bulked composite soil sample for analysis. Subsequently, all soil samples underwent air-drying and sieving to less than 2 mm. Standard laboratory analysis was conducted in a single laboratory, including particle size fractions (clay, silt, sand, %), coarse fragments (CF, %), BD (g cm-3), pH (in water), SOC (g kg-1), carbonates (CaCO3, g kg-1), total nitrogen (N, g kg-1), extractable potassium (K, mg kg-1), cation exchange capacity (CEC, cmol(+) kg-1). For more comprehensive information about LUCAS Soil 2009/2015/2018, we refer to Orgiazzi et al. (2022). In the LUCAS Soil 2018 survey, soil sampling was conducted across all EU Member States and UK, employing the identical set of 25,947 locations that were targeted during the 2015 survey (Fernández-Ugalde et al., 2022). However, due to the absence of particle-size fractions in LUCAS Soil 2018, we resorted to use the data from LUCAS Soil 2009/2015 by the unique identifier soil ID (Panagos et al., 2022). To ensure the reliability of the data, we excluded samples with soil particle fractions recorded as 0. Finally, 5,163 samples were retained for further analysis (Fig. 1).

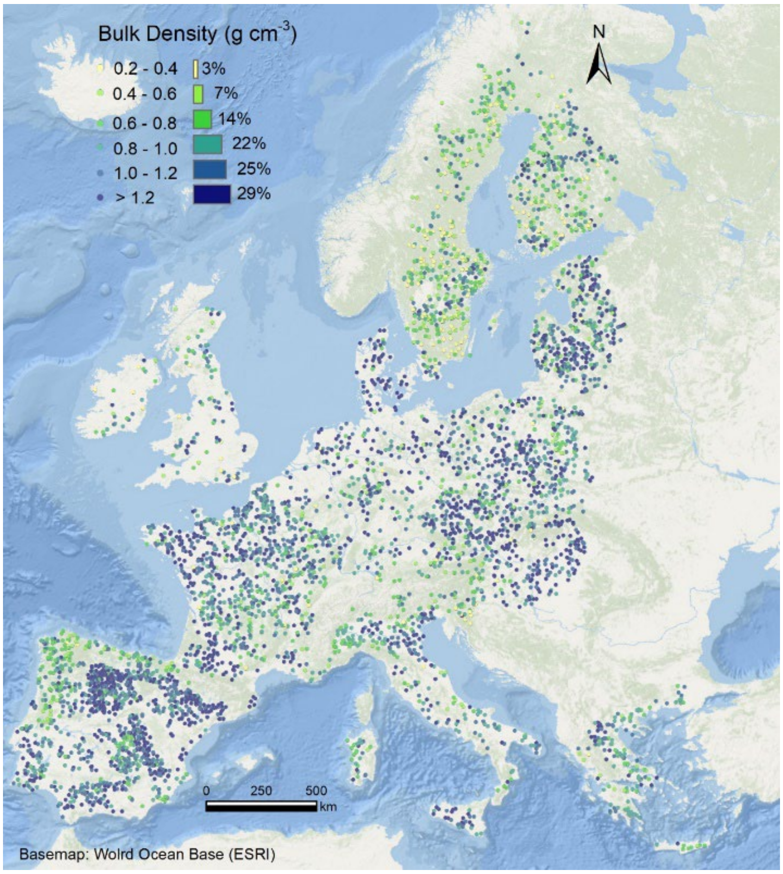

Figure 1: Spatial distribution of 5,163 soil samples from LUCAS Soil 2018.

Since BD was measured for the whole mass (BDsample) and CF was measured for the mass fraction (CFmassfraction) in the LUCAS Soil 2015/2018, they cannot be used to calculate SOC stock which requires BD of the fine fraction (BDfine) and CF calculated by volume fraction (CFvolumefraction) (Poeplau et al., 2017). Therefore, we used a recently released dataset for BDfine and CFvolumefraction by (Pacini et al., 2023) based on BDsample and CFmassfraction. Please note that the BD and CF mentioned hereafter all refer to BDfine and CFvolumefraction.

2.2 Predictor variables related to relief and climate

The elevation was derived from Shuttle Radar Topography Mission (SRTM) 1-km Digital Elevation Model (DEM) (Farr et al., 2007). Climatic data, including mean annual precipitation (MAP) and average annual temperature (MAT), were acquired from the WorldClim Version 2 at 1 km resolution (Fick and Hijmans, 2017). The Global-PET dataset at 1 km resolution was used to extract Potential evapotranspiration (PET) and Aridity index (AI) (Zomer et al., 2022).

2.3 Traditional PTFs

We evaluated four traditional PTFs (T-PTFs) that have been widely used to estimated BD in previous studies at a broad scale (Atwood et al., 2017; Chen et al., 2018; Sun et al., 2020; Tao et al., 2023). The parameters in these T-PTFs were refitted by the Levenberg-Marquardt non-linear least-square method available in the minpack.lm R package based on our data (Bates and Watts, 1981; Zhu et al., 1997; Elzhov et al., 2015). These refitted parameters of T-PTFs are presented in Table 2.

Table 2: Summary of four traditional PTFs defined in previous research

Model	Function	Refitted coefficients			Reference	R^2
		a	b	c		
1	$BD = a * \%SOC^b$	1.197	-0.229		Atwood et al. (2017)	0.397
2	$BD = \frac{1}{a + b * \%SOC}$	0.733	0.0982		Chen et al. (2018)	0.445
3	$BD = \frac{100}{\frac{\%SOM}{0.244} - \frac{(100 - \%SOM)}{a}}$	1.231			Sun et al. (2020)	0.405
4	$BD = a + b \times \exp(-c \times \%SOM)$	0.348	0.993	0.0882	Tao et al., (2023)	0.445

BD, bulk density (g cm^{-3}); *SOM*, soil organic matter (%); *SOC*, soil organic carbon (%).

2.4 Global ML-PTFs

To compare with T-PTFs, we used random forest (RF) to construct global ML-PTFs for predicting BD. RF is an ensemble learning method that aggregates predictions from multiple decision trees to obtain the final estimates of the target variable. In growing a decision tree, a random subsample of data is selected from the verification dataset, and a set of random predictor variables is used for splitting the subsampled data. Two parameters, ntree and mtry, were optimized by 10-fold cross-validation. Here, 16 predictor variables, such as sand, silt, clay, SOC, elevation (Table 3), were used to build the global RF model.

Table 3: The variables used in RF_{Full} and RF_{FRFS}. For RF_{FRFS}, the order of variables is listed by the descending importance.

Model	Selected predictors	Number of predictors
RF_{Full}	clay, silt, sand, pH, SOC, $CaCO_3$, N, K, CEC, AI, PET, Elevation, MAP, MAT, LC	15
RF_{FRFS}	SOC, N, pH, PET, MAP, LC, AI, MAT, ELE, EC, clay, silt	12

Furthermore, we adopted a recently proposed variable selection method, namely forward recursive feature selection (FRFS) to reduce the number of predictor variables while not losing model performance. FRFS employs a forward selection strategy, involving the following sequential steps: (1) initially, a RF model is fitted using all the n predictors, and their variable importance is calculated; (2) the most important predictor (only one) is selected to create an initial model, and its performance is assessed using k-fold cross-validation with a single predictor in the pool; (3) subsequently, a series of models are constructed using two predictors, where the first predictor is chosen from the pool, and the second predictor is selected from the remaining predictors. The model performances are evaluated, and the model with the best performance is recorded; (4) the pool of predictors is then updated based on the predictors from the best-performing model in the previous step; (5) The process is iteratively repeated, progressively increasing the number of predictors from 3 to n. Ultimately, the predictors used in the model with the best performance are selected to form the final predictive model, as detailed in the work of Xiao et al. (2022). The R script for implementing FRFS is accessible at https://doi.org/10.5281/zenodo.7141020. In this study, FRFS was applied to select the most relevant predictors constructing the predictive models (Table 3).

For clarity, in global modelling, we refer to the RF model using the full variables as global-RF$_{\mathrm{FULL}}$, and the combination of RF with FRFS as global-RF$_{\mathrm{FRFS}}$.

2.5 Local ML-PTFs

The development of local ML-PTFs consists of four steps: 1) use the Mahalanobis distance to calculate the distances of predictor variables between each sample to be predicted and all the samples in the database. 2) select k nearest neighbour samples to fit a RF model for each unknown sample; 3) predict the BD for each unknown sample using relevant RF models. Since the number of nearest neighbour samples (k) is an important parameter in the local model, we evaluated its effect on the model performance by testing k from 20 to 700 (20, 40, 60, 80, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600, 650, 700).

For clarity, we refer to the local modelling using the full variables as local-RF$_{\mathrm{FULL}}$, and for the combined use of RF and variables selected by global-RF$_{\mathrm{FRFS}}$, we refer it as local-RF$_{\mathrm{FRFS}}$.

2.6 Model evaluation

Due to the large sample size, single random split is stable compared to k-fold cross-validation or repeated random split (Chen et al., 2021). Therefore, we used randomly split (80% for calibration and 20% for validation) to assess the model performance of T-PTFs and ML-PTFs. The root mean square error (RMSE) and determination coefficient (R^2) were used as performance indicators on the validation set (Chen et al., 2022). These indices are defined as following Eq. (1) and (2):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2} \qquad (1)$$

$$R^2 = \frac{\sum_i^n (O_i - P_i)^2}{\sum_i^n (O_i - \bar{O})^2} \qquad (2)$$

where n represents the number of observations, O_i and P_i are the observed and predicted BD for observation i, and $\bar{o}$ is the mean of the observed BD. A good model has a $RMSE$ close to 0, and also a higher R^2 close to 1.

2.7 The build-up of extended BD and SOC stock datasets in Europe

Since only part of LUCAS 2015/2018 had soil particle fractions and $CaCO_3$, we used the unique samples ID to link the missing soil particle fractions and $CaCO_3$ using LUCAS Soil 2009 for the same sampling sites. This operation is reasonable since soil particle fractions and $CaCO_3$ will not have a notable change within a decade. The SOC stock (kg m^{-2}) at a depth of 0-20 cm for LUCAS Soil 2018 was calculated by the BD_{fine} (g cm^{-3}), $CF_{volumefraction}$ and SOC content (g kg^{-1}) (Poeplau et al., 2017) as Eq. (3).

$$SOCS = SOC \times BD_{fine} \times 20\ cm\ \times (1 - CF_{volumefraction})/100 \quad (3)$$

3 Results

3.1 Statistics of BD and its correlation with predictor variables

Figure 2 illustrates the histogram of BD and their distribution in a ternary soil texture triangle. The dataset consists of 5,163 soil samples with BD ranging from 0.20 to 1.89 g cm^{-3}. Approximately half of the soil samples exhibited BD between 0.8 and 1.4 g cm^{-3}, while less than 10% of the soil samples had BD exceeding 1.4 g cm^{-3}. The selected soil samples covered a wide range of soil texture classes, with the exception of clay soils.

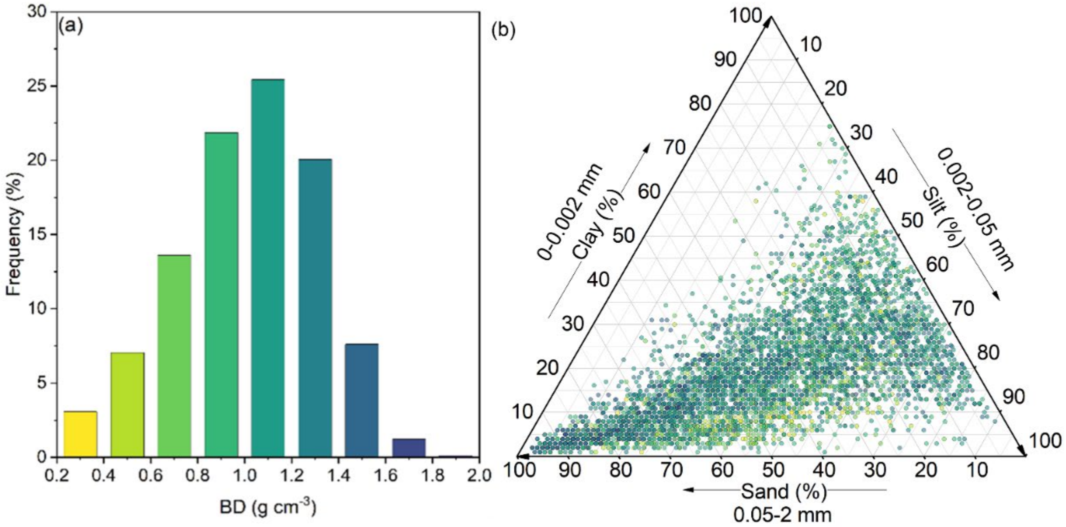

Figure 2: Histogram of BD (a) and USDA soil texture triangle (b). The point colors shown in the texture triangle correspond to the colors present in the left histogram. The percentage of each bin is indicated over the bin in the histogram.

Figure 3 depicts the correlation matrix between BD and the 15 predictor variables. BD exhibited positive correlations with pH and MAT, with correlation coefficients (r) greater than 0.2. On the other hand, BD showed notably high negative correlations with most of the other predictors. Among these, the most influential negative predictor was SOC (r=-0.62). Additionally, we observed BD was negatively correlated to N and AI.

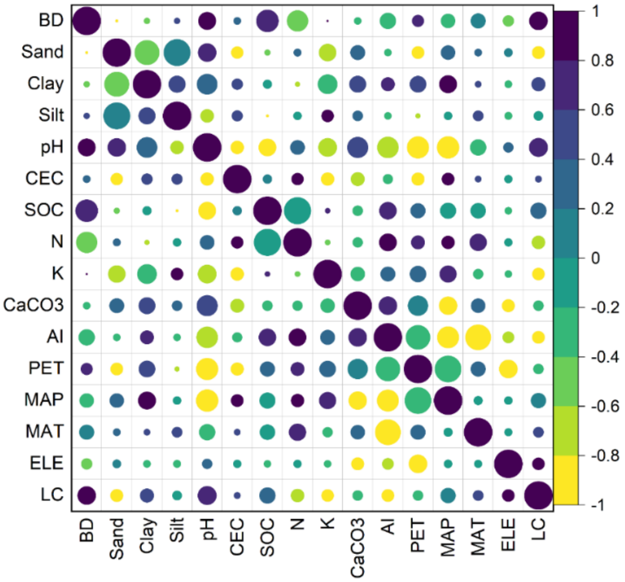

Figure 3: Correlation plot among BD and predictors.

190

3.2 Selection of predictor variables

Table 3 presents the predictor variables utilized in the RF model for predicting BD. In the global-RF_{FULL} model, 15 predictor variables were included, namely clay, silt, sand, pH, SOC, $CaCO_3$, N, K, CEC, AI, PET, Elevation, MAP, MAT, and LC. On the other hand, the global-RF_{FRFS} identified a subset of 8 variables by FRFS that were deemed most important for BD prediction. 195 These selected predictors, ranked in descending order of importance, were SOC, N, pH, PET, MAP, LC, AI, and MAT.

3.3 Comparison of ML-PTFs and T-PTFs in BD prediction

In this study, we compared ML-PTFs with four T-PTFs (Fig. 4 and 5). The model performance (R^2) of T-PTFs ranged from 0.40 and 0.45. The global-RF models had higher model performance with R^2 around 0.57 for global-RF_{FULL} and global-RF_{FRFS} where the later performed slightly better. As for local models, it was clear that the model performance showed an increasing 200 trend when the number of neighbour samples increased and some fluctuations were observed after the model performance reached a plateau. The number of neighbour samples were optimized at 350 and 400 for local-RF_{FRFS} and local-RF_{FULL}, respectively. Compared to global modelling, the best local-RF_{FRFS} and local-RF_{FULL} performed slightly better with R^2 around 0.58.

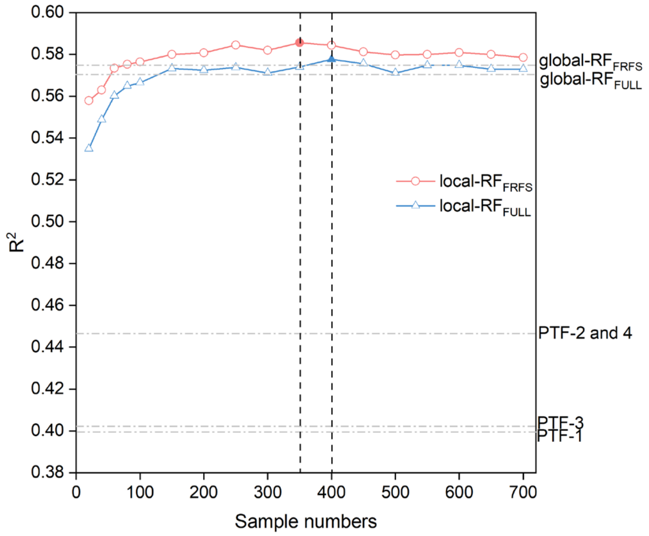

Figure 4: Model performance of the eight PTFs.

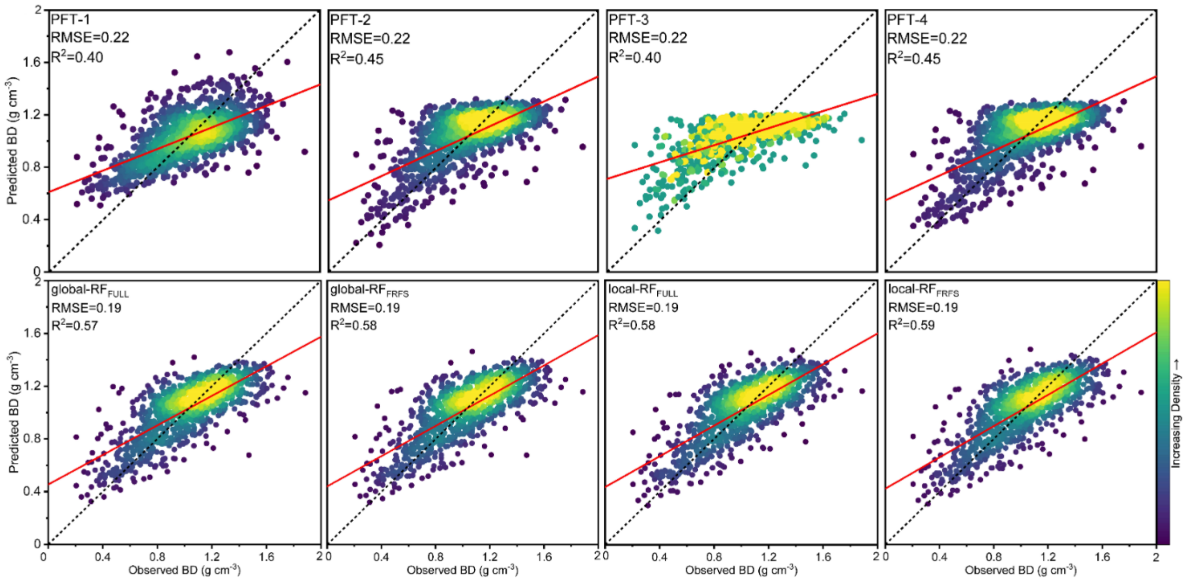

Figure 5: The performance of BD by using eight PTFs.

3.4 Comparison of ML-PTFs and T-PTFs in SOC stock calculation

We investigated how using BD estimated by PTFs impacted the accuracy of SOC stock calculation (Fig. 6). We found that SOC stock calculation using BD predictions from four T-PTFs resulted in a good performance with RMSE and R^2 ranging

1.39-1.89 kg m^{-2} and 0.70-0.84, respectively. Meanwhile, SOC stock calculation using BD predictions from four ML-PTFs performed similar (RMSE ranging 1.32-1.36 kg m^{-2}, R^2 ranging 0.84-0.85) to the best T-PTF (T-PTF-4, p=0.19-0.29).

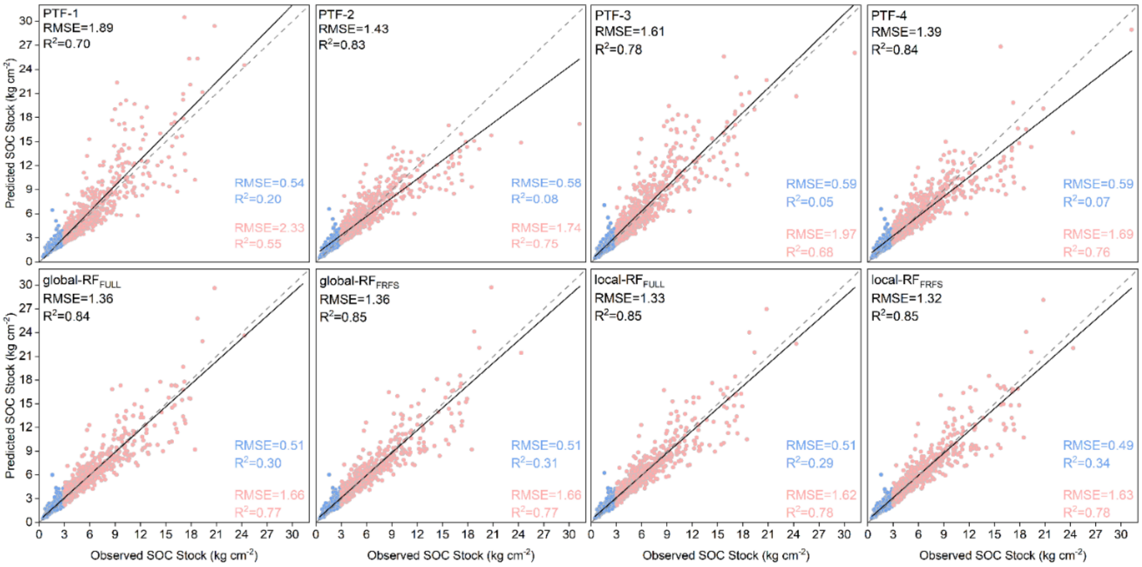

Figure 6: The performance of SOC stock prediction by using eight PTFs.

3.5 Summary of the extended European BD and SOC stock database

To enlarge the BD and SOC stock database for the Europe, we refitted best ML-PTF (local-RF_{FRFS}) and T-PTF (T-PTF-4) using all the 5,163 soil samples to predict soil samples without BD and then calculated SOC stock, which resulting in 15,389 and 18,945 soil samples for the extended database respectively (less soil samples had all the required variables for the use of local-RF_{FRFS}). As shown in Fig. 7, these extended BD and SOC stock databases are more regularly distributed across EU and UK compared to the points in Fig. 1. In EU and UK, BD was primarily distributed within 1.0-1.2 g cm^{-3} (46-47%) while the SOCS was mainly comprised between 2 and 4 kg m^{-2}.

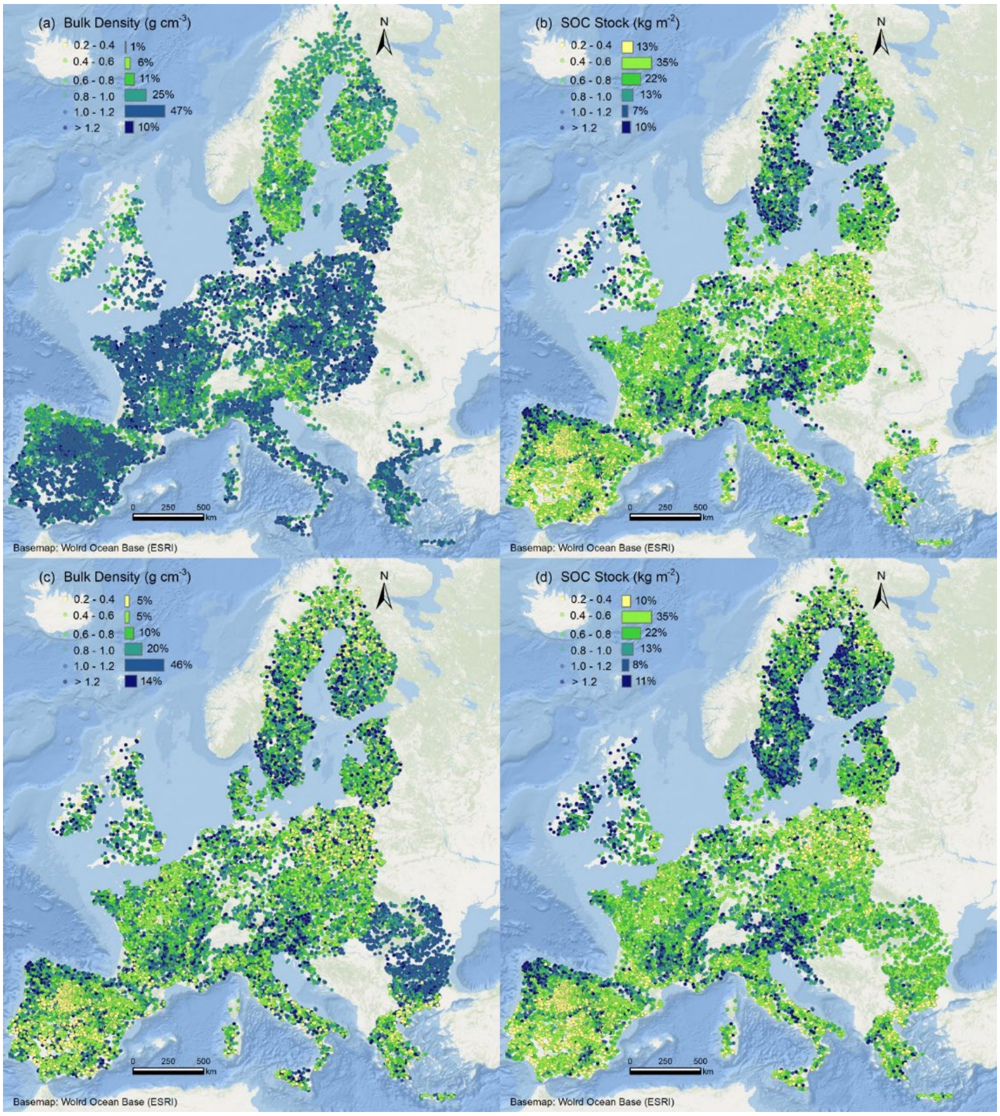

Figure 7: Spatial distributions of 15,389 soil samples with BD (a) and SOC stock (b) from LUCAS 2018 Soil using local-RF$_{\mathrm{FRFS}}$, and 18,945 soil samples with BD (c) and SOC stock (d) from LUCAS 2018 Soil using T- PTF-4.

As shown in Fig. 8, in the database created by local-RF$_{\mathrm{FRFS}}$ (15,389 soil samples), the soil samples in 0-20 cm under cropland had the highest median BD of 1.11 g cm^{-3}, while woodland exhibited the lowest median BD at 0.84 g cm^{-3}. Conversely, woodland had the highest median SOC stock at 6.21 kg m^{-2}, while cropland showed the lowest median SOC stock at 3.06 kg m^{-2}. As for the database built on T-PTF-4 (18,945 soil samples), cropland also had the highest median BD at 1.14 g cm^{-3} while woodland exhibited the lowest median BD at 0.86 g cm^{-3}. In contrast, the SOC stock under woodland presented the highest median SOC stock at 6.96 kg m^{-2} while cropland had the lowest median SOC stock at 3.17 kg m^{-2}.

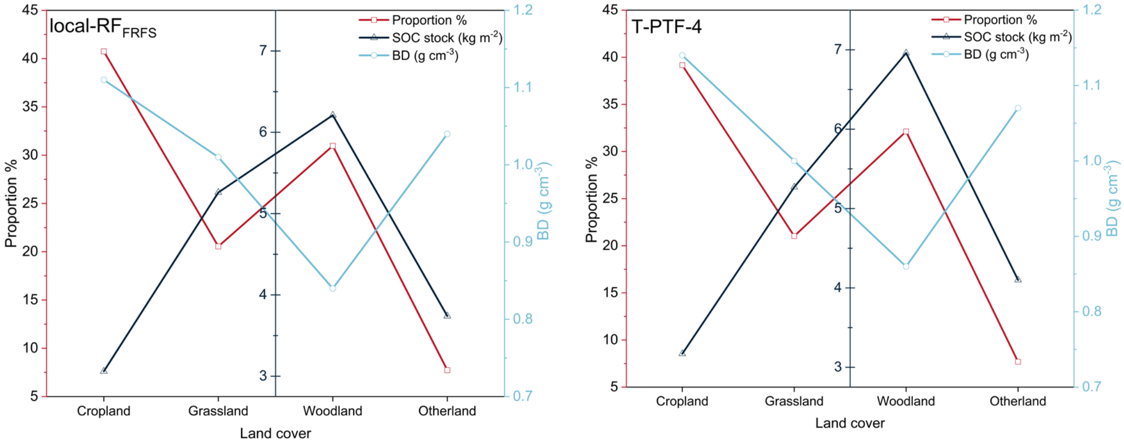

Figure 8: Variations of BD and SOC stock under different land uses using local-RF$_{FRFS}$ and T-PTF-4.

4 Discussion

4.1 The superiority of ML-PTFs in BD prediction

In this study, using the LUCAS Soil and 15 predictor variables, we compared the model performance of four commonly used T-PTFs and four ML-PTFs for BD. Four T-PTFs showed a moderate model performance with R^2 of 0.40-0.45, which is close to a recent developed Hollis-type T-PTF (R^2 of 0.41, Hollis et al., 2012) that refitted by LUCAS Soil 2018 data (De Rosa et al., 2023). Our results underscored the efficacy of ML-PTFs in successfully predicting BD at a continental scale, yielding a substantial R^2 ranging from 0.57 to 0.59. It indicates that when adding more relevant predictor variables (e.g., N, pH, PET, MAP) in the soil database, ML-PTFs is a better choice for improving BD prediction. Otherwise, T-PTFs are still the best choice to impute the missing data due to their simplicity (Van Looy et al., 2017).

In addition to global PTFs that use all the samples, we introduced the local modelling strategy in PTFs which searches similar samples first and then builds the relevant PTF for each unknown sample dynamically. Generally, the model performance of local PTFs (local-RF$_{FULL}$ and local-RF$_{FRFS}$) for BD prediction continuously improved with the number of neighbour samples increased, and then it reached a plateau when number of neighbour samples reached approximately 350 to 400 (Fig. 4). Compared to the global PTFs (4,500 soil samples), the size of local PTFs were much smaller (350-400 soil samples) with slightly better model performance. Therefore, local PTFs can be an efficient tool for imputing the missing data using a large soil database (Padarian et al., 2019; Sanderman et al., 2020).

Comparing with the commonly used T-PTFs that refitted by our data, the local-FR$_{FRFS}$ model greatly improved model performance in BD prediction (ΔR^2 around 0.2). Our results suggest that ML-PTFs performed much better for BD prediction than T-PTFs. This is resulted from the fact that most of ML models are able to handle non-linear and complex relationships between the predictor variables and the response variable so as to improve predictions than T-PTFs (Katuwal et al., 2020;

Palladino et al., 2022). Meanwhile, the T-PTFs typically rely solely on SOC or SOM for BD prediction. This approach maintains model simplicity but overlooks readily available predictor variables such as particle size fractions, MAT and MAP, which are also pertinent to BD prediction (Abdelbaki, 2018). Despite of the high diversity in landscapes and climates at a continental scale, the proposed local-FR$_{\text{FRFS}}$ model demonstrated similar or even superior performance compared to the ML-PTFs conducted at regional and national scales (Table 1).

4.2 Performance of FRFS and variable importance in BD prediction by ML-PTFs

We reduced the number of predictor variables in RF model from 15 to 8 using the FRFS algorithm, and the model performance of global-RF$_{\text{FRFS}}$ for BD using FRFS selected variables was higher than global-RF$_{\text{FULL}}$ using full variables (Table 3). Though the local-RF$_{\text{FRFS}}$ (R^2 of 0.59) only had marginal superiority over the local-RF$_{\text{FULL}}$ model (R^2 of 0.58), it facilitated the reduction of variables, consequently enhancing prediction efficiency (Fig. 4 and 5). This outcome validates the capacity of FRFS to simplify the model complexity while concurrently enhancing predictive accuracy (Xiao et al., 2022, Zhang et al., 2023). Being a useful tool for gap-filling the missing data, an ideal PTF require both high parsimony and good fit. If the developed PTF needs too many predictors variables, its practical applicability would be limited, as much fewer soil samples have all the required predictors variables.

4.3 The build-up of extended BD and SOC stock datasets in Europe

We used the BD predictions from eight PTFs as input to calculate the SOC stock. The result showed that the model performance of SOC stock (R^2 of 0.70-0.85) was much higher than those of BD (R^2 of 0.40-0.59) (Fig. 5 and 6). It can be explained by the interdependence between BD and SOC. For instance, a soil sample with a higher SOC commonly has a larger pore space due to the greater presence of SOC, leading to a lower BD (Perie and Ouimet, 2008; Chen et al., 2018). As shown in Fig. 6, high SOC and BD were always underestimated while the low SOC and BD were overestimated. By multiplying these two negatively correlated variables, the predicted SOC stock could be closer to the observed SOC stock as the overestimation (underestimation) of BD can counterbalance the underestimation (overestimation) of SOC, resulting in improved model performance than BD. It is interesting to note that the model performance of best T-PTFs (T-PTF-4, R^2 of 0.84) and ML-PTFs (local-RF$_{\text{FRFS}}$, R^2 of 0.85) was quite close in SOC stock prediction. This indicate that the improvement of BD prediction by ML-PTFs did not impact the accuracy of SOC stock prediction. Looking into the scatter plots shown in Fig. 5, we can observe that the ML-PTFs performed much better than T-PTFs for soil samples with high BD (low SOC) while limited difference was found for soil samples with low BD. Compared to T-PTFs, ML-PTFs tended to predict SOC stock better for soil samples with low SOC stock (<3 kg cm^{-2}) while similar model performance can be found in soil samples with high SOC stock (>3 kg cm^{-2}), which is evident in Fig. 5. As a result, the best T-PTFs performed quite similar to the best ML-PTFs when considering the soil samples with wide range of SOC stock. Our result indicated that the T-PTFs would be accurate enough to estimate BD which is subsequently used for SOC stock calculation due to their simplicity. Otherwise, ML-PTFs are suggested for more accurate BD prediction, especially for regions with low SOC stock such as dry land regions in Spain and Italy (Maestre et al., 2021; De Rosa et al., 2023; Wang et al., 2023).

4.4 The build-up of extended BD and SOC stock datasets in Europe

It is essential to acknowledge that our developed PTFs for BD prediction was constructed based on LUCAS Soil data (0-20 cm), confining its applicability to topsoil within the EU and UK (Orgiazzi et al., 2022, Panagos et al., 2022). However, the potential of their extrapolation capability to other regions or deep soil (>20 cm) necessitates further evaluation. As more soil database becomes available from diverse regions as well as deep soil (Lal, 2018; Tautges et al., 2019; Batjes et al., 2020; Yost et al., 2020;), the proposed methodology can be used to update the PTFs, thereby broadening its area of applicability (Chen et al., 2018; Meyer and Pebesma, 2021).

When using PTFs based BD prediction to detect SOC stock changes, the impact of the accuracy of BD PTFs on the accuracy of SOC stock calculation remains unclear since the equivalent soil mass approach also require BD as input (Schrumpf et al., 2011; Wendt and Hauser, 2013). Therefore, this issue should be investigated in future studies.

5 Data availability

All the soil data used in this article are available at the following data sources: (1) Land Use and Coverage Area Frame Survey Soil (LUCAS Soil) 2009 via https://esdac.jrc.ec.europa.eu/content/lucas-2009-topsoil-data (Panagos et al., 2022), (2) LUCAS Soil 2015 via https://esdac.jrc.ec.europa.eu/content/lucas2015-topsoil-data (Fernández-Ugalde et al., 2022), (3) LUCAS Soil 2018 via https://esdac.jrc.ec.europa.eu/content/lucas-2018-topsoil-data (Panagos et al., 2022), (4) the European soil BD and SOC stock dataset in this paper is available at https://zenodo.org/records/10211884 (Chen et al., 2023).

6 Conclusions

Using the largest extendable soil dataset for Europe, we have developed ML-PTFs for predicting BD across the EU and UK. In comparison with four T-PTFs, the best ML-PTFs, namely local-RF_{FRFS}, exhibited superior performance for BD prediction with percentage increase in R^2 at 31.1-47.5% and percentage decrease in RMSE at 13.6%. When the predicted BD is subsequently used for SOC stock calculation, we found that the best T-PTFs preformed quite similar to the best ML-PTFs, indicating the fact that T-PTFs would be enough for BD prediction when targeting in SOC stock calculation. However, for regions with low SOC stock (<3 kg m^{-2}), ML-PTFs are still recommended due to its high accuracy in SOC stock calculation. Finally, we established two comprehensive pan-European soil BD and SOC stock databases including 15,389 and 18,945 soil samples in LUCAS Soil 2018 using the best ML-PTF and T-PTF, respectively. Our study proposed a potential model to improve the predictive accuracy of BD, and the resultant BD and SOC stock datasets across the EU and UK enable more precise soil hydrological and biological modelling at a continental scale.

7 Author contributions

SC, ZC and XZ compiled the data. SC and ZC performed the analysis and drafted the manuscript. ZL, DA, ACRF and ZS validated the results and revised the manuscript. SC acquired of the financial support ZS supervised this work.

8 Competing interests

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

9 Disclaimer

Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

10 Financial support

This study is funded by the National Natural Science Foundation of China (No. 42201054).

11 References

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

Atwood, T. B., Connolly, R. M., Almahasheer, H., Carnell, P. E., Duarte, C. M., Ewers Lewis, C. J., and Lovelock, C. E.: Global patterns in mangrove soil carbon stocks and losses, Nat. Clim. Chang., 7, 523-528, https://doi:10.1038/nclimate3326, 2017.

Augusto, L., and Boča, A.: Tree functional traits, forest biomass, and tree species diversity interact with site properties to drive forest soil carbon, Nat. Commun., 13, 1097, https://doi.org/10.1038/s41467-022-28748-0, 2022.

Bates, D. M., and Watts, D. G.: Nonlinear regression analysis and its applications: Nonlinear regression analysis and its applications, 1981.

Batjes, N. H., Ribeiro, E., and Van Oostrum, A.: Standardised soil profile data to support global mapping and modelling (WoSIS snapshot 2019), Earth Syst. Sci. Data, 12, 299-320, https://doi.org/10.5194/essd-12-299-2020, 2020.

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:https://doi.org/10.1016/j.geoderma.2007.01.005, 2007.

Bondi, G., Creamer, R., Ferrari, A., Fenton, O., & Wall, D.: Using machine learning to predict soil bulk density on the basis of visual parameters: Tools for in-field and post-field evaluation, Geoderma 318, 137-147, https://doi:https://doi.org/10.1016/j.geoderma.2017.11.035, 2018.

Chen, S., Arrouays, D., Mulder, V. L., Poggio, L., Minasny, B., Roudier, P., Libohova, Z., Lagacherie, P., Shi, Z., Hannam, J., Meersmans, J., Richer-de-Forges, A. C., and Walter, C.: Digital mapping of GlobalSoilMap soil properties at a broad scale: A review, Geoderma 409, 115567, https://doi:10.1016/j.geoderma.2021.115567, 2022.

Chen, S., Chen, Z., Zhang, X., Luo, Z., Schillaci, C., Arrouays, D., Richer-de-Forges, A. C., Shi, Z. European soil bulk density and organic carbon stock database using LUCAS Soil 2018 [Data set]. Zenodo. https://doi.org/10.5281/zenodo.10211884, 2023

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:10.1016/j.geoderma.2017.10.009, 2018.

Chen, S., Xu, H., Xu, D., Ji, W., Li, S., Yang, M., Hu, B., Zhou, Y., Wang, N., Arrouays, D., and Shi, Z.: Evaluating validation strategies on the performance of soil property prediction from regional to continental spectral data, Geoderma 400, 115159, https://doi.org/10.1016/j.geoderma.2021.115159, 2021.

Cotrufo, M. F., Ranalli, M. G., Haddix, M. L., Six, J., and Lugato, E.: Soil carbon storage informed by particulate and mineral-associated organic matter, Nat. Geosci., 12, 989-994, https://doi.org/10.1038/s41561-019-0484-6, 2019.

Dam, R. F., Mehdi, B. B., Burgess, M. S. E., Madramootoo, C. A., Mehuys, G. R., and Callum, I. R.: Soil bulk density and crop yield under eleven consecutive years of corn with different tillage and residue practices in a sandy loam soil in central Canada, Soil Till. Res., 84, 41-53, https://doi:10.1016/j.still.2004.08.006, 2005.

Dawson, J. J. C., and Smith, P.: Carbon losses from soil and its consequences for land-use management, Sci. Total Environ., 382, 165-190, https://doi:https://doi.org/10.1016/j.scitotenv.2007.03.023, 2007.

De Rosa, D., Ballabio, C., Lugato, E., Fasiolo, M., Jones, A., and Panagos, P.: Soil organic carbon stocks in European croplands and grasslands: How much have we lost in the past decade?, Glob. Chang. Biol., 30, e16992, https://doi.org/10.1111/gcb.16992, 2023.

Elzhov, T. V., Mullen, K. M., Spiess, A. N., and Bolker, B.: minpack.lm: R Interface to the Levenberg-Marquardt Nonlinear Least-Squares Algorithm Found in MINPACK, Plus Support for Bounds, 2015.

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., and Alsdorf, D.: The shuttle radar topography mission, Rev. Geophys., 45, RG2004, https://doi.org/10.1029/2005RG000183, 2007.

Fernández-Ugalde, O., Orgiazzi, A., Marechal, A., Jones, A., Scarpa, S., Panagos, P., and Van Liedekerke, M.: LUCAS 2018 soil module: presentation of dataset and results: Publications Office of the European Union, 2022.

Fick, S. E., and Hijmans, R. J.: WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas, Int. J. Climatol., 37, 4302-4315, https://doi:10.1002/joc.5086, 2017.

Ghehi, N. G., Nemes, A., Verdoodt, A., Van Ranst, E., Cornelis, W. M., and Boeckx, P.: Nonparametric techniques for predicting soil bulk density of tropical rainforest topsoils in Rwanda, Soil Sci. Soc. Am. J., 76, 1172-1183, https://doi:10.2136/sssaj2011.0330, 2012.

Gupta, A., Vasava, H. B., Das, B. S., and Choubey, A. K.: Local modeling approaches for estimating soil properties in selected Indian soils using diffuse reflectance data over visible to near-infrared region, Geoderma 325, 59-71, https://doi:https://doi.org/10.1016/j.geoderma.2018.03.025, 2018.

Gupta, S. C., and Larson, W. E: Estimating soil-waster retention characteristics from particle-size distribution, organic-matter percent, and bulk-density, Water Resour. Res., 15, 1633-1635, https://doi:10.1029/WR015i006p01633, 1979.

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

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

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:10.1016/j.geoderma.2019.114080, 2020.

Lal, R.: Digging deeper: A holistic perspective of factors affecting soil organic carbon sequestration in agroecosystems, Glob. Chang. Biol., 24, 3285-3301, https://doi: 10.1111/gcb.14054, 2018.

Lark, R. M., Rawlins, B. G., Robinson, D. A., Lebron, I., and Tye, A. M.: Implications of short-range spatial variation of soil bulk density for adequate field-sampling protocols: methodology and results from two contrasting soils, Eur. J. Soil Sci., 65, 803-814, https://doi:10.1111/ejss.12178, 2014.

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

Maestre, F. T., Benito, B. M., Berdugo, M., Concostrina-Zubiri, L., Delgado-Baquerizo, M., Eldridge, D. J., Guirado, E., Gross, N., Kéfi, S., Bagousse-Pinguet, Y. L., Ochoa-Hueso, R., and Soliveres, S.: Biogeography of global drylands, New Phytol., 231, 540-558, https://doi.org/10.1111/nph.17395, 2021.

Makovníková, J., Širáň, M., Houšková, B., Pálka, B., and Jones, A.: Comparison of different models for predicting soil bulk density. Case study–Slovakian agricultural soils, Int. Agrophys., 31, 491-498, https://doi:10.1515/intag-2016-0079, 2017.

Meyer, H., and Pebesma, E.: Predicting into unknown space? Estimating the area of applicability of spatial prediction models, Methods Ecol. Evol., 12, 1620-1633, https://doi:10.1111/2041-210X.13650, 2021.

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

Nocita, M., Stevens, A., Toth, G., Panagos, P., van Wesemael, B., and Montanarella, L.: Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach, Soil Biol. Biochem., 68, 337-347, https://doi.org/10.1016/j.soilbio.2013.10.022, 2014.

Orgiazzi, A., Panagos, P., Fernández-Ugalde, O., Wojda, P., Labouyrie, M., Ballabio, C., Franco, A., Pistocchi, A., Montanarella, L and Jones, A.: LUCAS Soil Biodiversity and LUCAS Soil Pesticides, new tools for research and policy development, Eur. J. Soil Sci., 73, e13299, https://doi.org/10.1111/ejss.13299, 2022.

Pacini, L., Yunta, F., Jones, A., Montanarella, L., Barrè, P., Saia, S., Chen, S., and Schillaci, C.: Fine earth soil bulk density at 0.2 m depth from Land Use and Coverage Area Frame Survey (LUCAS) soil 2018, Eur. J. Soil Sci., 74(4), e13391, https://doi:10.1111/ejss.13391, 2023.

Padarian, J., Minasny, B. and McBratney, A.B.: Transfer learning to localise a continental soil vis-NIR calibration model, Geoderma 340, 279-288, https://doi.org/10.1016/j.geoderma.2019.01.009, 2019.

Palladino, M., Romano, N., Pasolli, E., and Nasta, P.: Developing pedotransfer functions for predicting soil bulk density in Campania, Geoderma 412, 115726, https://doi:10.1016/j.geoderma.2022.115726, 2022.

Panagos, P., Van Liedekerke, M., Borrelli, P., Köninger, J., Ballabio, C., Orgiazzi, A., Lugato, E., Liakos, L., Hervas, J., Jones, A and Montanarella, L.: European Soil Data Centre 2.0: Soil data and knowledge in support of the EU policies, Eur. J. Soil Sci., 73, e13315, https://doi.org/10.1111/ejss.13315, 2022.

Patton, N. R., Lohse, K. A., Seyfried, M., Will, R., and Benner, S. G.: Lithology and coarse fraction adjusted bulk density estimates for determining total organic carbon stocks in dryland soils, Geoderma 337, 844-852, https://doi.org/10.1016/j.geoderma.2018.10.036, 2019.

Perie, C., and Ouimet, R.: Organic carbon, organic matter and bulk density relationships in boreal forest soils, Can. J. Soil Sci., 88, 315-325, https://doi:10.4141/cjss06008, 2008.

Poeplau, C., Vos, C., and Don, A.: Soil organic carbon stocks are systematically overestimated by misuse of the parameters bulk density and rock fragment content, Soil, 3, 61-66, https://doi:10.5194/soil-3-61-2017, 2017.

Rabot, E., Wiesmeier, M., Schlüter, S., and Vogel, H. J.: Soil structure as an indicator of soil functions: A review, Geoderma 314, 122-137, https://doi:10.1016/j.geoderma.2017.11.009, 2018.

Ramcharan, A., Hengl, T., Beaudette, D., and Wills, S.: A soil bulk density pedotransfer function based on machine learning: A case study with the ncss soil characterization database, Soil Sci. Soc. Am. J., 81, 1279-1287, https://doi:10.2136/sssaj2016.12.0421, 2017.

Rawls, W. J., and Brakensiek, D. L.: Prediction of soil water properties for hydrologic modeling, ASCE, 293-299, 1985.

Sanderman, J., Savage, K., and Dangal, S. R.: Mid-infrared spectroscopy for prediction of soil health indicators in the United States, Soil Sci. Soc. Am. J., 84(1), 251-261, https://doi.org/10.1002/saj2.20009, 2020.

Schillaci, C., Perego, A., Valkama, E., Märker, M., Saia, S., Veronesi, F., Lipani, A., Lombardo, L., Tadiello, T., Gamper, G. 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, Sci. Total Environ., 780, 146609, https://doi:10.1016/j.scitotenv.2021.146609, 2021.

Schrumpf, M., Schulze, E. D., Kaiser, K., and Schumacher, J.: How accurately can soil organic carbon stocks and stock changes be quantified by soil inventories?, Biogeosciences, 8, 1193-1212, https://doi.org/10.5194/bg-8-1193-2011, 2011.

Shiri, J., Keshavarzi, A., Kisi, O., Karimi, S., and Iturraran-Viveros, U.: Modeling soil bulk density through a complete data scanning procedure: Heuristic alternatives, J. Hydrol., 549, 592-602, https://doi.org/10.1016/j.jhydrol.2017.04.035, 2017.

Sun, W., Canadell, J. G., Yu, L., Yu, L., Zhang, W., Smith, P., Fischer, T., and Huang, Y.: Climate drives global soil carbon sequestration and crop yield changes under conservation agriculture, Glob. Chang. Biol., 26, 3325-3335, https://doi.org/10.1111/gcb.15001, 2020.

Taalab, K., Corstanje, R., Mayr, T. M., Whelan, M. J., and Creamer, R. E: The application of expert knowledge in Bayesian networks to predict soil bulk density at the landscape scale, Eur. J. Soil Sci., 66, 930-941, https://doi:10.1111/ejss.12282, 2015.

Tao, F., Huang, Y., Hungate, B. A., Manzoni, S., Frey, S. D., Schmidt, M. W. I., Reichstein, M., Carvalhais, N., Ciais, P., Jiang, L., Lehmann, J., Wang, Y., Houlton, B. Z., Ahrens, B., Mishra, U., Hugelius, G., Hocking, T. D., Lu, X., Shi, Z., Viatkin, K., Vargas, R., Yigini, Y., Omutom C., Malik, A. A., Peralta, G., Cuevas-Corona, R., Paolo, L. E. D., Luotto, I., Liao, C., Liang, Y., Saynes, V. S., Huang, X., and Luo, Y.: Microbial carbon use efficiency promotes global soil carbon storage, Nature, 618, 981-985, https://doi:10.1038/s41586-023-06042-3, 2023.

Tautges, N. E., Chiartas, J. L., Gaudin, A. C., O'Geen, A. T., Herrera, I., and Scow, K.M.: Deep soil inventories reveal that impacts of cover crops and compost on soil carbon sequestration differ in surface and subsurface soils, Glob. Chang. Biol., 25, 3753-3766, https://doi.org/10.1111/gcb.14762, 2019.

Van Looy, K., Bouma, J., Herbst, M., Koestel, J., Minasny, B., Mishra, U., Montzka, C., Nemes, A., Pachepsky, Y. A., Padarian, J., Schaap, M. G., Tóth, B., Verhoef, A., Vanderborght, J., van der Ploeg, M. J., Weihermüller, L., Zacharias, S., Zhang, Y., and Vereecken, H.: Pedotransfer functions in Earth system science: Challenges and perspectives, Rev. Geophys., 55, 1199-1256, https://doi.org/10.1002/2017RG000581, 2017.

Wang, M., Guo, X., Zhang, S., Xiao, L., Mishra, U., Yang, Y., Zhu, B., Wang, G., Mao, X., Qian, T., Jiang, T., Shi, Z., and Luo, Z.: Global soil profiles indicate depth-dependent soil carbon losses under a warmer climate, Nat. Commun., 13, 5514, https://doi.org/10.1038/s41467-022-33278-w, 2022.

Wang, Y., Luo, G., Li, C., Ye, H., Shi, H., Fan, B., Zhang, W., Zhang, C., Xie, M., and Zhang, Y.: Effects of land clearing for agriculture on soil organic carbon stocks in drylands: A meta-analysis, Glob. Chang. Biol., 29, 547-562, https://doi.org/10.1111/gcb.16481, 2023.

Wendt, J. W., and Hauser, S.: An equivalent soil mass procedure for monitoring soil organic carbon in multiple soil layers, Eur. J. Soil Sci., 64, 58-65, https://doi.org/10.1111/ejss.12002, 2013.

Wiesmeier, M., Spörlein, P., Geuß, U., Hangen, E., Haug, S., Reischl, A., Schilling, B., von Lützow, M., and Kögel-Knabner, I.: Soil organic carbon stocks in southeast Germany (Bavaria) as affected by land use, soil type and sampling depth, Glob. Chang. Biol., 18, 2233-2245, https://doi.org/10.1111/j.1365-2486.2012.02699.x, 2012.

Xiao, Y., Xue, J., Zhang, X., Wang, N., Hong, Y., Jiang, Y., Zhou, Y., Teng, H., Hu, B., Lugato, E., Richer-de-Forges, A. C., Arrouays, D., Shi, Z., and Chen, S.: Improving pedotransfer functions for predicting soil mineral associated organic carbon by ensemble machine learning, Geoderma 428, 116208, https://doi:10.1016/j.geoderma.2022.116208, 2022.

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.

Yost, J. L. and Hartemink, A. E.: How deep is the soil studied–an analysis of four soil science journals, Plant Soil, 452, 5-18, https://doi.org/10.1007/s11104-020-04550-z, 2020.

Zhang, X., Chen, S., Xue, J., Wang, N., Xiao, Y., Chen, Q., Hong, Y., Zhou, Y., Teng, H., Hu, B., Zhuo, Z., Ji, W., Huang, Y., Gou, Y., Richer-de-Forges, A. C., Arrouays, D., and Shi, Z.: Improving model parsimony and accuracy by modified greedy feature selection in digital soil mapping, Geoderma 432, 116383, https://doi.org/10.1016/j.geoderma.2023.116383, 2023.

Zhu, C., Byrd, R. H., Lu, P., and Nocedal, J.: Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization, Acm T. Math. Software, 23, 550-560, https://doi.org/10.1145/279232.279236, 1997.

Zomer, R. J., Xu, J., and Trabucco, A.: Version 3 of the global aridity index and potential evapotranspiration database, Sci. data, 9, 409-409, https://doi:10.1038/s41597-022-01493-1, 2022.