~~Performance of three machine learning algorithms for predicting soil organic carbon in German agricultural soil~~Spatial prediction of organic carbon in German agricultural topsoil using machine learning algorithms

Ali Sakhaee[1], Anika Gebauer[2], Mareike Ließ[2], Axel Don[1]

[1]Thünen Institute of Climate Smart Agriculture, Braunschweig, Germany
[2]Department Soil System Science, Helmholtz Centre for Environmental Research – UFZ, Halle (Saale), Germany

Correspondence to: Ali Sakhaee (a.sakhaee@thuenen.de)

Abstract

As the largest terrestrial carbon pool, ~~S~~soil organic carbon (SOC), ~~as the largest terrestrial carbon pool,~~ has the potential to influence and mitigate climate change~~ and mitigation~~, ~~and consequently~~hence the importance of SOC monitoring ~~is important~~ in the frameworks of ~~different~~ various international treaties. ~~There is therefore a need for h~~High resolution SOC maps are therefore required. Machine learning (ML) offers new opportunities to ~~do this~~ develop these due to its ~~cap~~ability for data mining of large datasets. The aim of this study~~, therefore,~~ was to ~~test~~ apply three ~~commonly used~~ algorithms commonly used in digital soil mapping – random forest (RF), boosted regression trees (BRT) and support vector machine for regression (SVR) – on the first German Agricultural Soil Inventory to model agricultural topsoil (0-30 cm) SOC content~~.~~ and develop a two-model approach to address the high variability of SOC in German agricultural soils. Model performance is often limited by the size and quality of the soil dataset available for calibration and validation. Therefore, the impact of enlarging the training data was tested by including data from the European Land Use/Land Cover Area Frame Survey for agricultural sites in Germany. Nested cross-validation was implemented for model evaluation and parameter tuning.~~ Moreover,~~ ~~G~~grid search and the differential evolution algorithm were also applied to ensure that each algorithm was appropriately tuned ~~and optimised suitably~~. The SOC content of the German Agricultural Soil Inventory was highly variable, ranging from 4 g kg^{-1} to 480 g kg^{-1}. However, only 4% of all soils contained more than 87 g kg^{-1} SOC and were considered organic or degraded organic soils. The results showed that SVR ~~provided~~ produced the best performance with an RMSE of 32 g kg^{-1} when the algorithms were trained on the full dataset. However, the average RMSE of all algorithms decreased by 34% when mineral and organic soils were modelled separately, with the best result from SVR with a RMSE of 21 g kg^{-1}. ~~Model performance is often limited by the size and quality of the available soil dataset for calibration and validation. Therefore, the impact of enlarging the training data was tested by including 1223 data points from the European Land Use/Land Cover Area Frame Survey for agricultural sites in Germany.~~ The model performance was enhanced ~~for maximum~~ by up to 1% for mineral soils and by 2% for organic soils. Despite the ~~cap~~ability of machine learning algorithms in general~~,~~ and ~~particularly~~ SVR in particular, ~~in~~ to model~~ling~~ SOC on a national scale, the study showed that the most important aspect ~~to~~ for improv~~ing~~e the model performance was to separate the modelling of mineral and organic soils.

1 Introduction

Soil organic carbon (SOC) is the largest terrestrial carbon pool (Wang et al., 2020) and plays an essential role in agriculture. Since SOC influences various physical, chemical and biological properties of soil (Reeves, 1997), numerous studies recognise it as a crucial indicator of soil quality (Castaldi et al., 2019; Meersmans et al., 2012a; Reeves, 1997) ~~(Castaldi et al., 2019; Meersmans et al., 2012; Reeves, 1997)~~ and therefore~~. Thus,~~ its decline is identified as a threat that leads to soil degradation (Castaldi et al., 2019; Poeplau et al., 2020). Moreover, when considering carbon sequestration, the SOC pool provides the option for climate change mitigation (Meersmans et al., 2012a; Ward et al., 2019) ~~(Meersmans et al., 2012; Ward et al., 2019)~~. ~~Consequently,~~ SOC monitoring is therefore important in the frameworks of various international treaties such as the European Union Soil Thematic Strategy and the United Nations Framework Convention on Climate Change (Meersmans et al., 2012b; Poeplau et al., 2020) ~~(Meersmans et al., 2012; Poeplau et al., 2020).~~and ~~T~~there is, ~~therefore~~ growing interest in understanding the spatial distribution of SOC at different scales in response to an increasing demand for a better assessment of SOC (Minasny et al., 2013). This is particularly important for agricultural land due to its potential for carbon sequestration (Lal, 2004).

In digital soil mapping (DSM), a soil attribute is ~~formulated as~~described by an empirical quantitative function of seven factors: soil properties, climate, organisms, topography, parent material, time, and spatial position (McBratney et al., 2003). ~~Therefore, t~~This function, known as the SCORPAN model, can be applied to spatially predict the soil ~~attribute~~ property of interest (Minasny et al., 2013). Within this framework, machine learning algorithms aim to automatically extract ~~the~~ information from the data for predictive purposes (Behrens et al., 2005). This is of particular~~ly~~ ~~intriguing~~ interest in view of the recent expansion of ~~databases at a different scale in soil science~~soil databases and the ~~the complexity of the covariates~~vast amount of data to approximate the soil forming factors ~~in recent years~~ (McBratney et al., 2003; Wadoux et al., 2020), thus making DSM cost-effective, time-efficient and applicable over large areas with good results (Behrens and Scholten, 2006; Camera et al., 2017).

Despite the advantages of DSM, it is crucial to ~~consider~~ note that its application requires soil databases of an adequate sample size for training and testing. Furthermore, consistent and quality-checked datasets are a prerequisite for DSM. Several soil inventories and monitoring networks for SOC have been ~~formed~~ established on a national scale in countries such as Sweden (Poeplau et al., 2015), France (Belon et al., 2012; Arrouays et al., 2002) ~~(Belon et al., 2012; Meersmans et al., 2012)~~ Denmark (Taghizadeh-Toosi et al., 2014) and Scotland (Chapman et al., 2013). ~~Nonetheless~~However, in Germany the most critical shortcomings of soil inventories ~~in Germany~~ concern ~~are~~ the lack of ~~a~~ large-scale, high-quality SOC ~~inventory~~ monitoring (Wiesmeier et al., 2012) with ~~a~~ periodic and standardised sampling focused on agricultural soils (Prechtel et al., 2009). These issues have now been ~~solved~~ addressed in the first German Agricultural Soil Inventory (Poeplau et al., 2020). This inventory was ~~conducted~~ carried out on a national scale ~~with a sampling depth down to 100 cm~~considering a sampling depth of 1 m at 3104 sampling sites covering agricultural land. Furthermore, on a European scale, the Land Use/Land Cover Area Frame Survey (LUCAS) undertaken in 2009 is the first harmonised topsoil survey with physico-chemical analyses of georeferenced topsoil samples ~~in~~ from 23 European states (Tóth et al., 2013). Therefore, by taking advantage of DSM and of both the German Agricultural Soil Inventory and ~~the~~ LUCAS survey, it is possible to regionalise from single-point measurements to obtain ~~complete~~ high-resolution cover soil data nationwide and thus provide a baseline for both SOC monitoring ~~as well as for~~and environmental and climatic modelling for Germany.

Boosted regression trees (BRT), random forest (RF) and support vector machine for regression (SVR) are among the most widely used algorithms in DSM (Padarian et al., 2020). For example, Martin et al. (2014) predicted topsoil SOC on a national scale for France using the BRT algorithm ~~and~~ comparing~~ed~~ its results when the same algorithm was coupled with a geostatistical approach. They concluded that due to the large distances between sampling sites, spatial autocorrelation is unlikely ~~since spatial autocorrelation is not feasible~~ in most national inventories, and the BRT algorithm alone is sufficient for this purpose. This algorithm has was also been used on a national scale in China for data from the 1980s and 2010s in order to predict topsoil SOC and its spatial-temporal change, as well as the main drivers of its variability (Wang et al., 2021). ~~Moreover,~~ RF has also become more popular in DSM due to its relative simplicity and performance. For example, this algorithm was implemented to map topsoil SOC on a national scale in Madagascar and ~~obtain~~ identify its main drivers (Ramifehiarivo et al., 2017). Ramifehiarivo et al. (2017) concluded ~~that~~ that the uncertainty of the map generated by RF model training~~uncertainty of the algorithm~~ was lower when compared with the maps that were formerly generated for the country. Moreover, this algorithm was compared with the Cubist ~~model~~ algorithm for mapping SOC at different resolutions on a regional scale in China and ~~could~~ was found to outperform~~ed~~ it (Li et al., 2021). Fewer studies have used SVR than RF to predict SOC~~ than RF~~. Studies have mainly implemented SVR on a regional scale with a limited number of samples (Forkuor et al., 2017; Were et al., 2015) or on a national scale (Switzerland) with very few samples (150 samples from the European LUCAS survey) (Zhou et al., 2021). However, in a study comparing different algorithms, including SVR and RF, on a continental scale and within each country in Latin America, the results indicated that the best-performing algorithm varied ~~in different countries~~from country to country (Guevara et al., 2018). The~~is~~ difference mainly depended on ~~sample~~ data density, quality, ~~dispersity and~~representativeness and ~~also~~ country size, which affect~~s~~ the heterogeneity of land use and environmental conditions.

Another important consideration when applying machine learning is the impact of the parameter-tuning strategy in algorithm performance. This is particularly crucial when the objective of the study is ~~the comparisons of~~to compare different machine learning algorithms. Although some algorithms are less sensitive to tuning, this step is more important for others, particularly those with a higher number of parameters (Tziachris et al., 2020; Wadoux et al., 2020). Furthermore, as algorithms differ by ~~the~~ type of ~~their~~ parameter~~s~~, continuous or discrete, the chosen strategy should be aligned ~~in accordance~~ with this difference (Ließ et al., 2021). ~~This is particularly more important for algorithms with continuous parameters.~~ For example, ~~it has been shown that~~ ~~that~~ the performance of SVR and BRT ~~is~~ has been shown to be better and more stable when optimised by a differential evolution (DE) algorithm than tuned by grid search (Zhang et al., 2011; Gebauer et al., 2020). Despite this importance, in a review of studies that have applied DSM, Wadoux et al. (2020) state that almost half of them implemented parameter tuning, with grid search the most common strategy applied for this purpose. This finding indicates that the role of parameter tuning and optimisation is unfortunately undermined in DSM. This is particularly evident when the application of machine learning in this field is compared with other fields, where various studies have shown the impact of parameter-tuning strategies on the performance of algorithms such as SVR and BRT (Liang et al., 2011; Santos et al., 2021; Bhadra et al., 2012; Deng et al., 2019).

The aims of the present study were~~was~~ therefore: i) to address the above-mentioned parameter-tuning issue and consequently provide a true comparison of the performance of BRT, RF, and SVR in modelling the SOC contents of German agricultural topsoils (0-30 cm), ii) to assess the impact of training data size by extending the data of the German Agricultural Soil Inventory with LUCAS data for model calibration, and iii) to develop a two-model approach to address the high variability of SOC in German agricultural soils and compare it with a single-model approach.

2 Materials and methods

2.1 Soil data

The models were built using SOC content data from two soil inventories. The first dataset was from the German Agricultural Soil Inventory, which ~~consists~~ comprise of 3104 sites ~~with a fixed~~collected along a grid of 8x8 km throughout Germany (Poeplau et al., 2020). The sites were sampled and analysed for different soil properties, including SOC content measured via dry combustion, for the upper 30 cm of the soil between 2012 and 2018. The second dataset was the European LUCAS survey that provides SOC content, similarly also measured via dry combustion, ~~for all EU countries,~~ with the sampling depth limited to 0-20 cm (Tóth et al., 2013). For Germany, data collected on agricultural soils cover 1223 sites. Therefore, in order to harmonise the depths of both datasets, the~~yse~~ were subdivided into two classes: mineral and organic soils ~~classes~~ according to a SOC threshold value of 87.0 g kg^{-1}. Accordingly, ~~considering~~ all soils above this threshold were considered as organic soils comprising peat soils and disturbed and degraded peat soils (Poeplau et al., 2020). Linear regression functions were derived for both mineral, Eq. 1, and organic, Eq. 2, soil classes on behalf of the data of the German Agricultural Soil Inventory to relate the SOC content of 0-30 cm to that of 0-20 cm.~~Linear correlation functions between 0-30 cm and 0-20 cm were derived for each soil class of the German Agricultural Soil Inventory separately.~~ These functions were then applied to the corresponding soil class from the LUCAS data in order to estimate 0-30 cm topsoil SOC. ~~With a slope of 0.881 for mineral soils and 1.02 for organic soils, they changed the mean of LUCAS data by less than 6%. The depth-extrapolated values of mineral and organic soils were then combined to form the complete dataset.~~ The 0-30 cm LUCAS data generated and the original 0-20 cm LUCAS data were then used by each algorithm to check the effect of depth extrapolation.

$$y = 1.01 + 0.881x \qquad (1)$$

$$y = 1.6 + 1.02x \qquad (2)$$

2.2 Covariates

Covariates from multiple sources were included to approximate the SCORPAN factors throughout Germany. In the case of multiple data products for one covariate, the one with the best quality (~~least~~ fewer artefacts)~~,~~ and the highest spatial resolution was added. These were then resampled in ArcGIS (ESRI, 2013) using the INSPIRE standard grid at 100 m resolution (Eurostat grid generation tool for ArcGIS). The resampling method was either the nearest neighbour for categorical covariates or bilinear interpolation for continuous covariates. The same INSPIRE grid was also used to rasterise the vector covariates~~ as well~~. Finally, they were stacked and overlaid on SOC databases in order to extract ~~the~~ values at the sampling points.

Following the SCORPAN framework, 24 covariates including x and y coordinates for spatial position~~s~~ were compiled. In order to ~~capture~~ represent the climate factor~~s~~ (C factor), precipitation (DWD, 2018c), sunshine duration (DWD, 2017), summer days (DWD, 2018b), and minimum temperature (DWD, 2018a) were ~~used~~ applied according to the study of Schneider et al. (2021). Using principal component analysis, these four covariates were ~~indicated~~ identified to be the most important ~~among~~ out of 34 available climate factors for SOC in the German Agricultural Soil Inventory dataset. Moreover, type of agricultural land use is one of the main drivers of SOC variability ~~at~~ on a national scale (Poeplau et al., 2020)~~. Thus~~, therefore the land ~~-~~use map from the official topographic-cartographic information system (BKG, 2019) with its corresponding classes according to the German Agricultural Soil Inventory was rasterised and included. This is a categorical covariate, representing the organism factor of SCORPAN (O factor), ~~that~~ which distinguishes croplands from grasslands and captures their spatial distribution throughout Germany.

The European Digital Elevation Model (EUDEM) (European Union Copernicus Land Monitoring Service, 2016) with original resolution of 25 m was resampled to 100 m. ~~and s~~Six covariates derived from the ~~is layer~~resampled layer were also added to integrate the ~~topography and~~ relief parameters (R factor). Slope, plan curvature and profile curvature, generated ~~on~~ with SAGA (Conrad et al., 2015), were included to capture the slope's gradient, convexity-concavity and convergence-divergence. These factors influence the soil distribution throughout the landscape, e.g. affecting flow over the surface, thus impacting SOC and its dynamic (Ritchie et al., 2007). Moreover, ~~north-south and east-west aspects were~~slope exposition (aspect) was ~~obtained~~ calculated from the ~~from~~ EUDEM as ~~these~~ it influences soil development and subsequently affects SOC (Carter and Ciolkosz, 1991). The circular variable was then decomposed into northness and eastness. The Topographic Wetness Index (TWI), generated on SAGA ~~(Conrad et al., 2015)~~, was also added since it captures the soil moisture distribution of the landscape and some studies have shown its direct correlation ~~has a direct correlation~~ with SOC (Pei et al., 2010). A geomorphographic map of Germany (~~Federal Institute for Geosciences and Natural Resources (~~BGR~~)~~, 2007) ~~containing~~ featuring 25 geomorphic categories was also used to distinguish between four different landscape areas of the country: North German lowlands, highlands, Alpine foothills and the Alps.

Continuing with the framework, a large-scale soil landscape unit map ("Soil Scapes in Germany~~Bodengrosslandschaft~~") (~~Federal Institute for Geosciences and Natural Resources (~~BGR~~)~~, 2008) comprising 38 classes was used. This covariate divides Germany by various geo-factors that can be compiled into a map with 12 soil regions~~ representing mainly the parent materials~~. Similarly, ~~large-scale~~the soil-climate region map ~~("Bodenklima")~~ (Roßberg et al., 2007) with 50 classes was added. Moreover, the Hydrogeological unit according Hydrogeological map of Germany~~Germany's hydrogeological unit map~~ (BGR and S~~G~~DG, 2019). The hydrogeological map provides information about ~~-~~hydrogeologically relevant attributes including consolidation, type of porosity, permeability, type of rock and geochemical classification. ~~lithology and its hydrological characteristics.~~ These categorical maps were rasterised and applied to the model as the P factor of SCORPAN. Moreover, the soil factor of the framework (S factor) was captured by eight covariates that represent different aspects of its properties: the map of organic soils (Roßkopf et al., 2015) that distinguishes mineral soils from organic ones and explains their spatial distribution throughout the country, as well as the maps of nitrogen (Ballabio et al., 2019) and clay content (Ballabio et al., 2016) since they directly correlate with SOC. As nitrogen is a crucial component of soil organic matter, regions with higher total nitrogen have higher SOC (Ballabio et al., 2019). Also for clay content, different studies have shown that coarser soil textures tend to have a lower accumulation of SOC (Zhong et al., 2018; Hoyle et al., 2011). The ~~M~~map of pH from ~~(~~Ballabio et al., (2019) was included since soil pH directly impacts microbial activities that influence the turnover of soil organic matter, and consequently negatively correlates with SOC (Malik et al., 2018). Furthermore, the map of available water capacity (Ballabio et al., 2016) was used as this soil ~~properties~~ property is another interactive factor with SOC through plant productivity and soil texture (Burke et al., 1989; Yu et al., 2021). Soil erosion is also a key factor in the SOC cycle (Li et al., 2019), which was added through the ~~soil~~ map of Europe's net soil erosion and deposition rates ~~erosion~~

~~map of Europe~~ (Borrelli et al., 2018). Based on the WaTEM/SEDEM model, this map illustrates the potential spatial displacement and transport of soil sediments due to water erosion (Borrelli et al., 2018). Figure S1 provides a more detailed view for better visualisation of the covariates that were used in this study.

2.3 Boosted Regression Trees

Developed by Friedman et al. (2000), BRT is a tree-based algorithm that applies boosting ~~method~~ to improve accuracy. Boosting ~~method~~ relies on combining several approximate prediction models rather than obtaining one ~~single~~ highly accurate ~~one~~ model (Schapire, 2003). Thus, the decision trees are grown sequentially so that each decision tree predicts the residual of the previous one and therefore~~. Consequently,~~ the number of trees influences the performance of the algorithm and requires tuning. However, to incorporate randomness in~~to~~ the model and subsequently increase the robustness of performance, the trees are grown on a randomly selected data subset with no replacement (Friedman, 2002). The size of this subset is controlled by a parameter known as a bag fraction. Furthermore, the contribution of each new tree to the final model is regularised by learning rate, also known as shrinkage(Friedman et al., 2009). Finally, the number of splits in each tree that divides the response variable into subsets is optimised by interaction depth. The BRT model was built in R using the “gbm” package (Greenwell et al., 2019).

2.4 Random Forest

Similar to BRT, RF is another tree-based algorithm. RF uses bootstrap sampling of the dataset for growing a decision tree. Subsequently, by aggregating the results of a large number of decision trees, the bias and variance of the final model can be reduced (Breiman, 1999). The method of bootstrapping in conjunction with aggregating, known as bagging, increases the robustness and stability of RF. However, the trees from different bootstraps may form a similar structure if all covariates participate in a split of each node. Thus, the variance cannot be reduced optimally through the bagging process (Kuhn and Johnson, 2013). In order to avoid this tree correlation, a random subset of covariates, i.e. predictors, is selected at each split. The parameter m_{try} defines the number of predictors included in this subset and should be tuned (Kuhn and Johnson, 2013). The RF algorithm was implemented by setting the number of trees to 1000 and using the “Ranger” package (Wright and Ziegler, 2017) in R.

2.5 Support Vector Regression

SVR is a form of support vector machine adopted for regression. From all possible solutions, i.e. estimation function, for the problem, SVR tries to obtain an estimation function ~~with~~ that has at most ~~the maximum~~ $\boldsymbol{\varepsilon}$ ~~error~~ deviation from the response values of the training data while minimising model complexity (Smola and Schölkopf, 2004). Thus, a symmetrical tolerance threshold, $\boldsymbol{\varepsilon}$-insensitivity zone, is created around the estimation function ~~within which the vectors are not penalised~~ (Awad and Khanna, 2015). ~~However, t~~The data vectors of the samples that lie on the boundary of the $\boldsymbol{\varepsilon}$-insensitivity zone are called support vectors. The vector lying within the insensitivity zone are not penalized. ~~Therefore,~~ $\boldsymbol{\varepsilon}$ is an optimisable parameter that controls the width of $\boldsymbol{\varepsilon}$-insensitivity, alters the model complexity and inversely impacts the number of support vectors ~~inversely~~ (Cherkassky and Ma, 2004). Moreover, the trade-off between model complexity and tolerance of $\boldsymbol{\varepsilon}$ deviation is controlled by a parameter named C (Smola and Schölkopf, 2004; Cherkassky and Ma, 2004). Optimising the C parameter has a crucial impact on SVR performance since a high C can lead to overfitting, while a low C can cause under fitting (Kuhn and Johnson, 2013). The use of kernel functions makes SVR a powerful tool for nonlinear problems. By implementing these functions, SVR can map the data space ~~from its original dimension~~ to a higher dimensional space where a nonlinear problem can be solved linearly. In this study, the Radial Basis Function (RBF) kernel was used with gamma as its tuneable parameter. This parameter affects the generalisation performance of SVR by inversely controlling the influence of support vectors ~~inversely~~ (Battineni et al., 2019). SVR was implemented from the package e1071 in R (Hornik et al., 2021).

2.6 Performance evaluation

When training a predictive model, it is important to evaluate its generalisation performance on unseen data of the same type (Hawkins et al., 2003). However, as the number of available samples is usually a limiting factor, the evaluation process is often done by k-fold cross validation (CV). Therefore, the dataset is divided into k folds and k – 1 folds are used for training the model and one fold for testing. This process is repeated k times so each fold participate in train and test. ~~However, as the number of available samples is usually a limiting factor, the evaluation process is often done by randomly splitting the available dataset into training and testing sets multiple times, i.e. cross-validation (CV). Although this process is effective, it is not entirely immune from biased estimation of error. However, to ensure that the estimated error in model evaluation is as unbiased as possible,~~ However, to ensure the robustness of the model, ~~every~~ each model training step should be performed within the CV. This includes finding the best parameter sets for the chosen algorithm (Varma and Simon, 2006). Thus, the algorithms in this study were applied on a stratified nested CV.

First, to ensure that the SOC distribution was represented in the CV scheme, Germany was divided into 50 strata using a 100x100 km INSPIRE grid ~~into 50 strata~~. Random samples from each stratum were then taken and compiled into a fold. This procedure was continued to create five folds and was repeated five times, forming the outer loop of CV used for model evaluation. ~~Large~~ A long distance between neighboring samples, 8120 m on average, prevents train and test data from being spatially autocorrelated. Since the aim was to tune the algorithms' parameters ~~of the algorithms~~, the training set of the outer loop of CV was nested, creating five folds as the inner loop on which the parameter tuning was performed. To evaluate the performance of algorithms, root-mean-squared error (RMSE), Eq. 3~~1~~, mean absolute error (MAE), Eq. 4~~2~~, and mean absolute percentage error (MAPE), Eq. 5~~3~~, were used~~.~~. Furthermore, AIC, Eq. 6, BIC, Eq. 7, and %Bias, Eq. 8, are also included in Table S2 for more detailed comparison.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2}$$ (3~~1~~)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)$$ (4~~2~~)

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{P_i - O_i}{O_i}\right| \times 100$$ (5~~3~~)

$$AIC = -2ln(L) + 2k$$ (6)

$$BIC = -2ln(L) + log(n)\cancel{\log\log}\, k$$ (7)

$$\%BIAS = \frac{1}{n}\sum_{i=1}^{n}\frac{(P_i - O_i)}{O_i} \times 100$$ (8)

w~~W~~here n is the number of samples, L is likelihood, k is the number of parameters, and P_i and O_i are the predicted and observed values, respectively. ~~were~~

2.6.1 Parameter tuning

As mentioned ~~in Sect.1~~previously, choosing a suitable strategy for parameter tuning is a crucial step in machine learning particularly ~~for~~ when comparing the performance of ~~the~~ algorithms. Therefore, two strategies were applied depending on the algorithm: 1) a grid search for RF and 2) optimisation with the DE algorithm for BRT and SVR. ~~The first strategy was an exhaustive search over a defined space consisting of lower bound, upper bound and *n* steps in between for the target parameter~~One major problem with applying the grid search strategy for algorithms that comprise continuous parameters such as BRT and SVR is that it is impossible to consider the whole continuous parameter space. Thus, the parameter combination for testing should be determined. However, this is not problematic for tuning RF in the present case since m_{try} is a parameter with discrete values. The DE algorithm however, is a~~n~~ stochastic approach to solve an optimisation problem that can be applied to both continuous and discrete parameters~~. Therefore, the target parameters in this strategy should be discrete or discretised beforehand if they are continuous (Probst et al., 2019). This strategy was applied to RF since the tuning parameter is discrete. However, the second strategy is a stochastic approach of searching over a continuous space in order to solve an optimisation problem (Qin et al., 2009) and~~. This method is described in more detail by Storn & Price (1997). Therefore, SVR and BRT ~~were~~ are optimised by this strategy as the former algorithm has continuous parameters and the latter one has both continuous and discrete parameters~~they have continuous parameters~~. For the optimisation task in the present study, the R package "DEoptim" was applied (Peterson et al., 2021). Table S1 shows the parameters and their tuning range for each algorithm.

2.6.2 Variable importance

Variable importance was assessed by permutation (Ließ et al., 2021). ~~Therefore, t~~The values of ~~each~~ a particular covariate in the test set ~~was~~ were shuffled ~~10 times~~ and ~~on each occasion the~~ prior to applying the respective model to eliminate any predictor-response relationship present with regards to that predictor~~trained model corresponding to that test set was applied~~. The variable importance corresponds to the relative increase in the test set RMSE. This ~~process~~procedure was repeated 10 times for each covariate. The resulting values were averaged ~~and T~~ ~~the population of RMSE was averaged and its~~ ~~to be used for calculation of relative change to the RMSE of the original test set was calculated.~~ Thus, the variable importance of each covariate in terms of ~~percentage~~ relative change in RMSE was obtained.

2.7 Modelling approaches

~~Three approaches were designed~~We followed a two-by-two strategy resulting in four modelling approaches to test the performance of the algorithms (Table 1). ~~The models were built based on nested CV, while the train and test sets remained identical for the three algorithms to make the results comparable.~~

Table 1: Modelling approaches

	Dataset 1: German Agricultural Soil Inventory	Dataset 2: German Agricultural Soil Inventory + LUCAS
One-Model-Approach	AP1	AP1L
Two-Model-Approach	AP2	AP2L

On the one hand, we~~The first approach (AP1)~~ only used the SOC ~~content~~ data from the German Agricultural Soil Inventory and corresponding values from the covariates ~~were used~~ to ~~build~~ train the models (AP1). ~~Thus, the dataset was cross-validated and used by BRT, RF and SVR to predict the SOC content of German agricultural soils. The results of this approach served as a baseline on which the model improvement for each algorithm in the other two approaches was assessed.~~

Due to the high variability of SOC in the agricultural soils of Germany, we then trained two separate models for organic and mineral soils (AP2) ~~was developed and tested~~ to identify whether ~~it~~ this could improve model performance. Accordingly, the German Agricultural Soil Inventory was subdivided by the threshold 87 g kg^{-1} into mineral and organic soils. ~~and then two were used to train separate models were trained. This approach was named AP2. The same nested CV procedure was applied for both data subsets. The results of BRT, RF and SVR were compared to identify which one had better performance under mineral and organic soils separately. Finally, each algorithm's predicted SOC values from the two separate models was were combined, and the error metrics were calculated for the full data set to identify the impact of AP2 on model performance. The CV folds for this procedure match the one from the AP1 models.~~

The impact of enlarging the training set on model performance was then examined for both, AP1 and AP2 ~~approaches~~. Thus, 1223 depth-extrapolated samples of the LUCAS data were added to the training sets of AP1. The corresponding modelling approach was ~~and~~ named AP1L. Moreover, the same threshold (87 g kg^{-1}) was used to subdivide this dataset and each soil class was included to the training set of the corresponding soil class of AP2. This modelling approach was then ~~and~~ named AP2L.

The test sets for the model performance evaluation~~of the CV procedure~~ remained the same for all four approaches~~.~~ ~~The models were built based on nested CV, while the train and test sets remained identical for the three algorithms~~ to make the results comparable. ~~Thus, the dataset was cross-validated and used by BRT, RF and SVR to predict the SOC content of German agricultural soils.~~ The results of ~~this~~the AP1 approach served as a baseline on which the model improvement for each algorithm in the other ~~two~~ approaches were~~was~~ assessed.

3 Results and Discussion

3.1 Comparison of algorithms on the data from the German Agricultural Soil Inventory (AP1)

The range of the topsoil SOC content ~~of topsoil~~ for the German Agricultural Soil Inventory dataset was 4 g kg^{-1} to 480 g kg^{-1}, with a mean of 27 g kg^{-1} and a median of 16 g kg^{-1}. Figure 1 shows the spatial distribution of the ~~implemented~~ data. For the first approach (AP1), BRT, RF, and SVR were applied to model SOC using data from German Agricultural Soil Inventory. The RMSE and MAPE indicated that SVR had a better general performance than the ~~two~~ other two algorithms (Fig.~~,~~ 2). In this respect, the RMSE of SVR was 5% lower than that from RF and 4% lower than that from BRT. Furthermore, its MAPE was 3% and 7% lower than that from RF and BRT respectively. However, despite the difference in overall performance, the spatial distribution of relative residuals indicated that all three algorithms were less accurate in ~~the~~ northern of Germany compared with the centre and south of the country (Fig. 3A). This can be explained by the characteristics of this region and its higher SOC variability. The northern part of Germany is~~ a~~ lowland dominated by a sandy soil texture from pleistoceen sedimentation with geomorphological structures such as ground moraines, terminal moraines and aprons (Roßkopf et al., 2015). Despite general geomorphological and pedological similarities throughout the region, 1) organic soils ~~in~~ under agricultural use ~~Germany~~ are mainly located in the north and 2) mineral soils with the lowest and ~~the~~

highest SOC contents are also located in the northeast and northwest respectively. Therefore, this region has the ~~highest~~ widest SOC range ~~on agricultural soils~~.

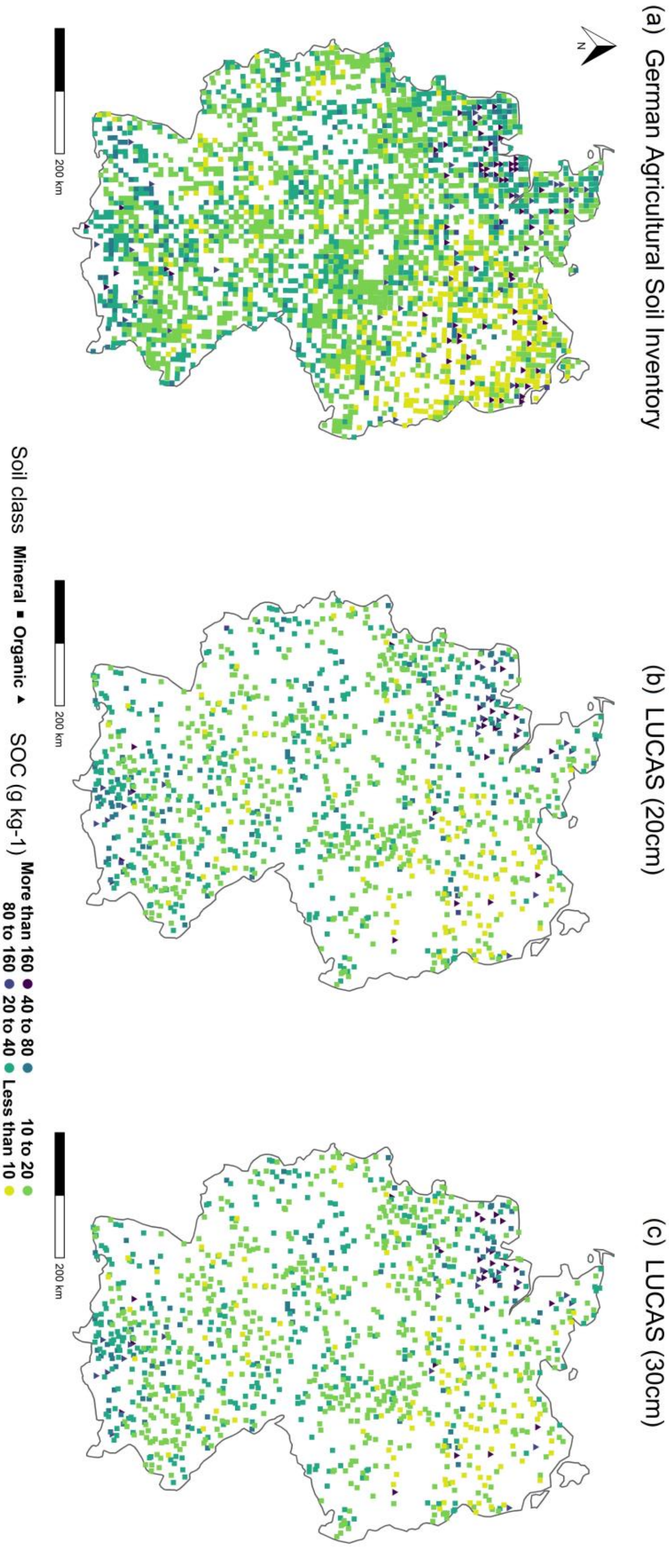

Figure 1: Soil organic carbon content in the topsoil of two soil inventories:~~.~~ A) German Agricultural Soil Inventory (0-30 cm), B) LUCAS at its original sampling depth (0-20 cm) and~~,~~ C) LUCAS after depth extrapolation (0-30 cm)


1
2
3
4
5
6
(a)
(b)
(c)
One-model approach
Two-model approach
BRT
RF
SVR
RMSE (g kg-1)
MAE (g kg-1)
MAPE (%)
AP1
AP1L
AP2
AP2L
36
34
32
30
28
26
24
22
20
13
12
11
10
9
8
54
52
50
48
46
44
42
40
38
36
34
32
30

Figure 2: Performance indicators of the three algorithms. One-model approach (w~~W~~ithout LUCAS data AP1 and with LUCAS data AP1L) versus the two-model approach (AP2 and AP2L) for A) RMSE (g kg^{-1}), B) MAE (g kg^{-1}) and C) MAPE (%). The whiskers of boxplots show 1.5 times the interquartile range. Please note that the y-axis is shortened for better visibility and does not display a zero. BRT = boosted regression trees, RF = random forest, and SVR = support vector regression.

Consequently, the variable importance (Fig. 4A) indicated that the map of organic soils ~~contains the highest available information among all~~was the most important covariate~~s for the algorithms~~. The value ~~for~~ of the variable importance for this covariate was 65% in SVR, 72% in RF and 84% in BRT. These values firstly show the crucial role of the map of organic soils for the algorithms in explaining the variability of SOC and, secondly, the comparatively greater importance of this predictor and the lower variable importance of other predictors in the BRT model compared with the SVR model.~~how BRT mainly relies on the map of organic soils to predict SOC compared with SVR.~~ Despite the importance of the organic soil map, the scatterplots (Fig. 5A) show that all three algorithms underpredicted the SOC of ~~the~~ organic soils and had similar heteroscedasticity patterns in their residuals. Thus, while most residuals from mineral soils followed the 1:1 line, they became more scattered in soils with a higher SOC content. The underprediction of SOC in organic soils can be explained by their ~~low~~ small sample size, resulting in a dataset with a ~~high~~ wide SOC range and a unimodal distribution that leaves these soils in the tail. Consequently, the organic soils were underrepresented and the results were systematically pulled towards mineral soils, ~~regardless~~ irrespective of the choice of algorithm. Different studies have shown that predicting soil properties with mineral and organic soils combined can lead to underprediction or overprediction of one soil class, depending on the distribution of the dataset (Brogniez et al., 2015; Guio Blanco et al., 2018; Mulder et al., 2016).

Although the map of organic soils was able to distinguish between the two soil classes, i.e. between mineral and organic soil, it could not separate the mineral soils with a low SOC content in the northeast from those with a high SOC content in the northwest. The spatial distribution of the residuals (Fig. 6A) showed~~s~~ that SVR and BRT generally underpredicted the mineral soils in the northwest part of Germany, while RF overpredicted them. Furthermore, unlike RF and SVR, BRT ~~distinctively~~ appreciably overpredicted SOC of ~~the~~ north-east Germany's mineral soils with the lowest SOC content (<10 g kg^{-1}). This result indicates that the algorithms differed in their performance in mineral soils. This difference was mainly due to the information they obtained from the land use map. As the second most important covariate for all three algorithms (Fig. 4 A), the value for variable importance for this covariate was 22% in SVR, but just 11% in RF and 9% in BRT. Thus, SVR exploits more information from this covariate than RF and particularly BRT. Land use is one of the main drivers of SOC variability on a national scale due to the higher SOC content in grasslands than in croplands (Poeplau et al., 2020). Therefore, this covariate was able to differentiate between the soils of the northeast, which are under cropland, and those in the northwest as they are more under grassland. Consequently, the reliance of BRT on the map of organic soils at the ~~cost~~ expense of land use could explain why this algorithm overpredicted SOC in croplands in the northeast.

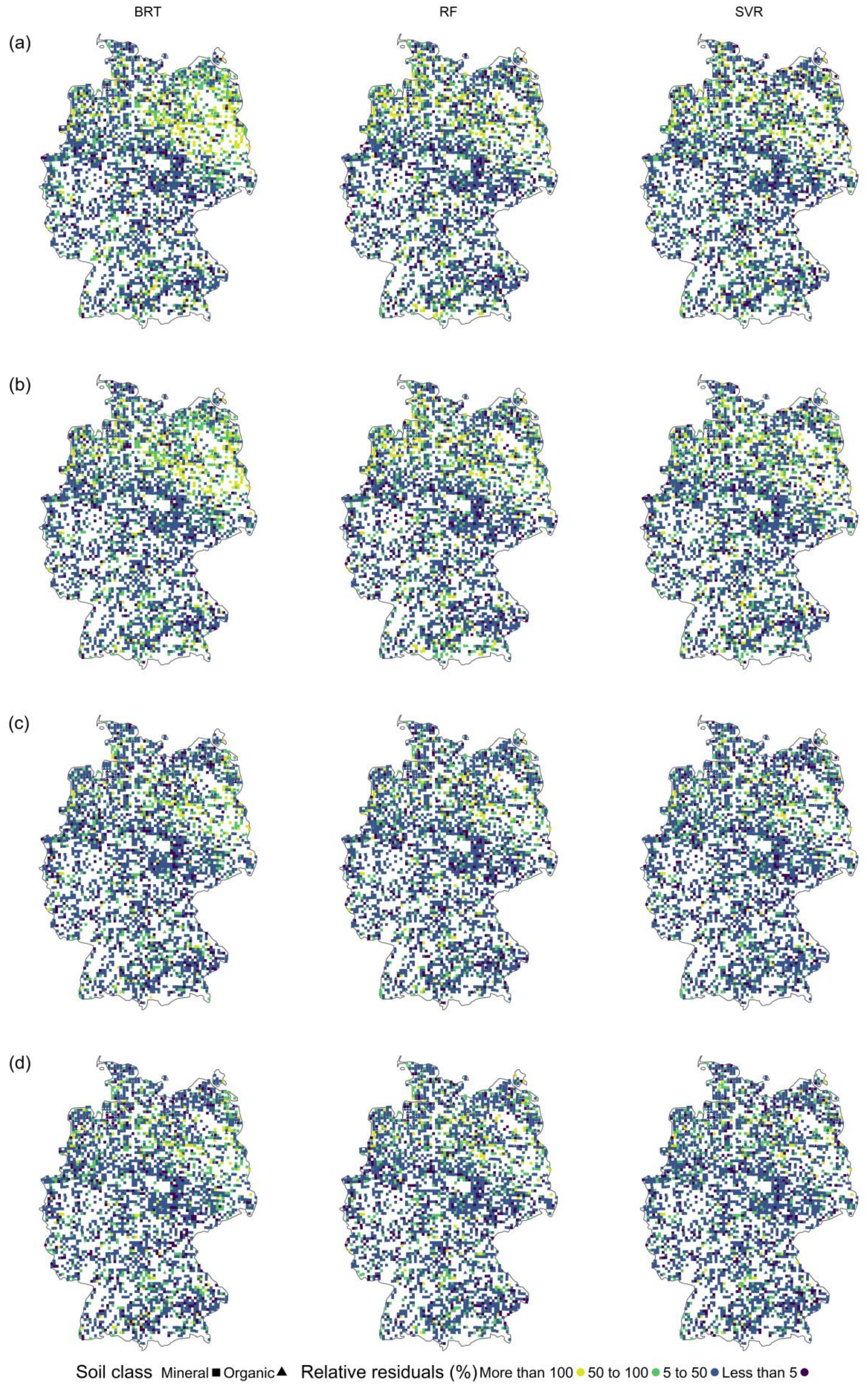

Figure 3: Spatial distribution of relative residuals. A) AP1 approach, B) AP1L approach, C) AP2 approach and D) AP2L approach. BRT = boosted regression trees, RF = random forest, and SVR = support vector regression.

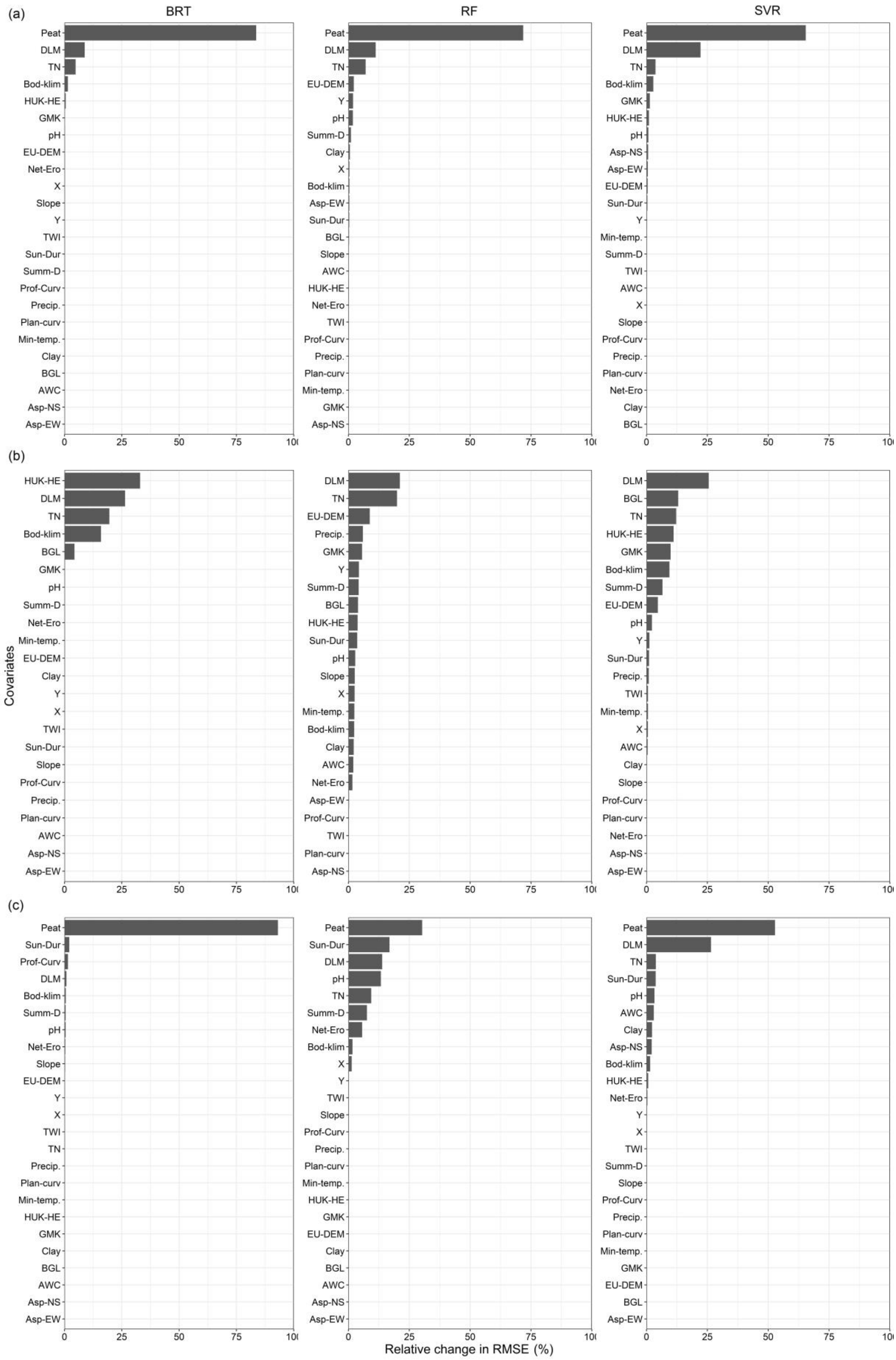

Figure 4: Variable importance in terms of average relative change (%) in RMSE. A) AP1, B) mineral soil subset of AP2 and C) organic soil subset of AP2. The full name for each abbreviation is presented in Table S~~4~~3. BRT = boosted regression trees, RF = random forest, and SVR = support vector regression.

3.2 Enlarging the dataset with additional soil inventories (AP1L)

A larger soil dataset may provide additional information and consequently improve model performance. This possibility was explored in the AP1L approach ~~with~~ by adding ~~the~~ LUCAS data. The SOC content of LUCAS data at its original depth ranged from 4 g kg^{-1} to 500 g kg^{-1}~~.~~, with a mean of 30 g kg^{-1} and a median of 18 g kg^{-1}. After extrapolating the depth to 30 cm, the new range was from 5 g kg^{-1} to 512 g kg^{-1}~~.~~, with a mean of 28 g kg^{-1} and a median of 17 g kg^{-1}. The spatial distribution of LUCAS data at their original and extrapolated depth is shown in Figure 1.

A statistical test was performed on the residuals of models built on LUCAS data with the original and extrapolated depths. That was done to identify whether extrapolating the depth of LUCAS data to that of the German Agricultural Soil Inventory would significantly affect model performance after their inclusion in the training set. With the Shapiro-Wilk test rejecting the normality assumption of residuals of all corresponding algorithms at 20 cm and 30 cm, the non-parametric Kruskal-Wallis test showed no significant difference between the residuals at ~~both~~ either depth~~s~~. Thus, the extrapolation of ~~the~~ soil depth had no significant impact on ~~the~~ data quality to regionalis~~z~~e SOC. As a result, any further change in the performance of the algorithms after adding LUCAS data was due to enlargement of the training set ~~being enlarged~~. The result of the algorithms at both depths can be found in the supplementary information (Fig. S3~~1~~).

After enlarging the training set from 2278 to 3501 sampling points, BRT obtained the lowest RMSE (Fig. 2A1) and MAE among the algorithms (Fig. 2B1). A comparison of the error metrics of corresponding algorithms from the AP1 approach with those from the AP1L approach showed that BRT had the highest error reduction at 7% in the MAPE and 5% in the RMSE and MAE. Furthermore, although the error metrics of RF did not improve as much as those of BRT, additional training points were still beneficial for this algorithm. However, SVR did not follow any systematic change under the AP1L. Despite a 2% decrease in MAPE, the RMSE increased by 3% and MAE remained unchanged. To explore the potential explanation for this behaviour by SVR, the residuals of mineral soils were separated from those of organic soils. Additional samples reduced the RMSE in mineral soils for all algorithms by between 9% and 13%. However, this error increased by 9% in the organic subset for SVR, while it increased by just 1% for RF and even decreased by 1% for BRT. This indicated that enlarging the training set by data with similar characteristics had a greater influence on systematic error of the underrepresented soil class in SVR. This influence is understandable when considering the higher optimised ε in the AP1L approach compared with that of ~~the~~ AP1 approach. The higher value of ε means that the hyperplane for the training set is less complex (Cherkassky and Ma, 2004) and more suitable for predicting most soil samples, i.e. mineral soils. Thus, when this hyperplane was fitted to the test set identical to the AP1, the generalisation performance was hindered because it could not capture the variability of samples with higher SOC values, i.e. organic soils.

Further evaluation revealed that regardless of the change in error metrics, the relative residuals of the three algorithms had a similar spatial pattern to their counterpart from ~~the~~ AP1. Thus, they all showed lower accuracy in the northern region of Germany for similar reasons (Fig. 3B). Moreover, the scatterplots had a similar pattern with underpredicted organic soils (Fig. 5B). This confirmed~~s~~ that when organic soils are modelled with mineral soils, enlarging the training set does not provide enough information for BRT or RF to capture the high variability of SOC, particularly in the north of Germany.

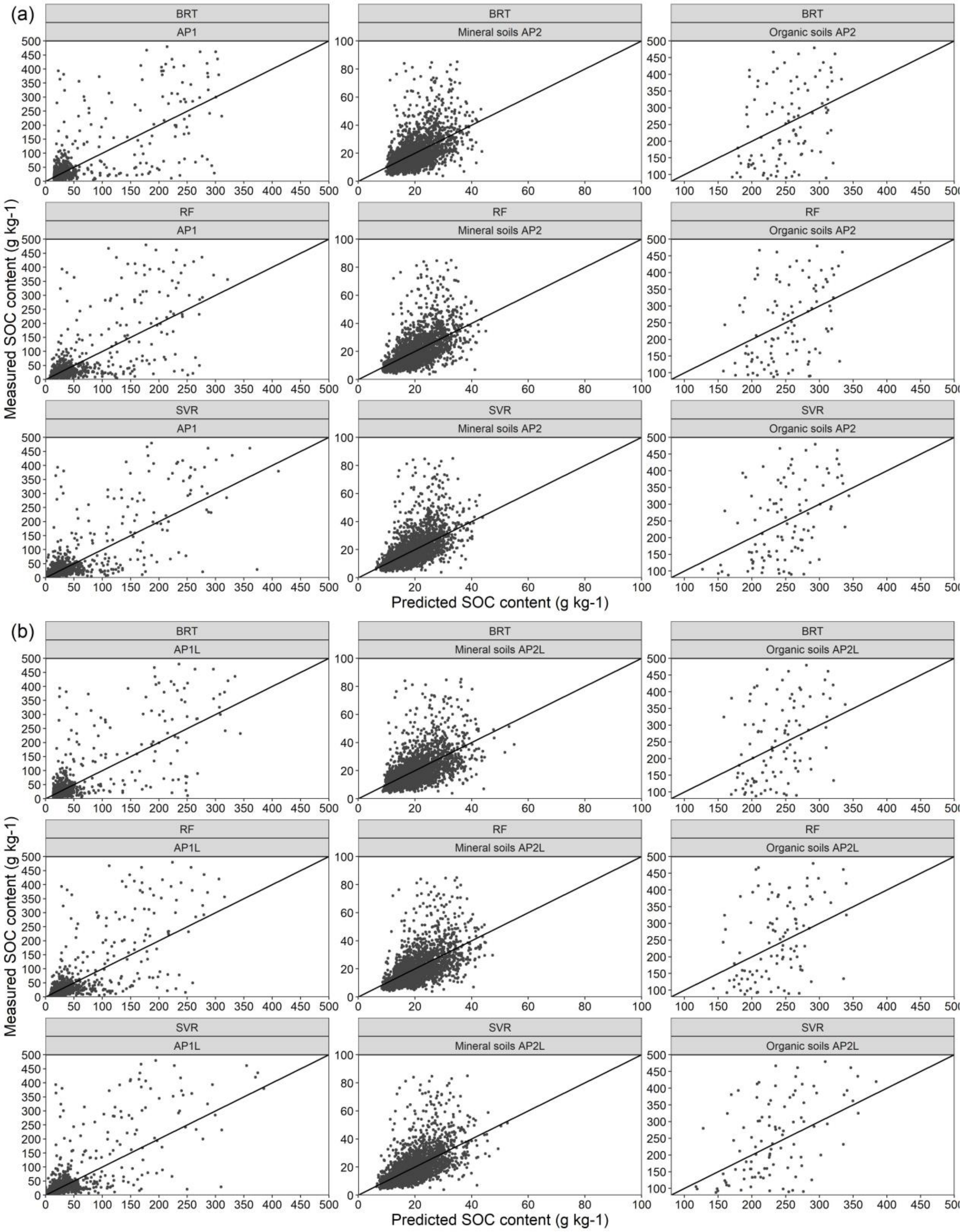

Figure 5: Scatterplot of residuals. A) AP1 approach and mineral and organic soils of AP2 and B) AP1L approach and mineral and organic soils of AP2L. BRT = boosted regression trees, RF = random forest, and SVR = support vector regression.

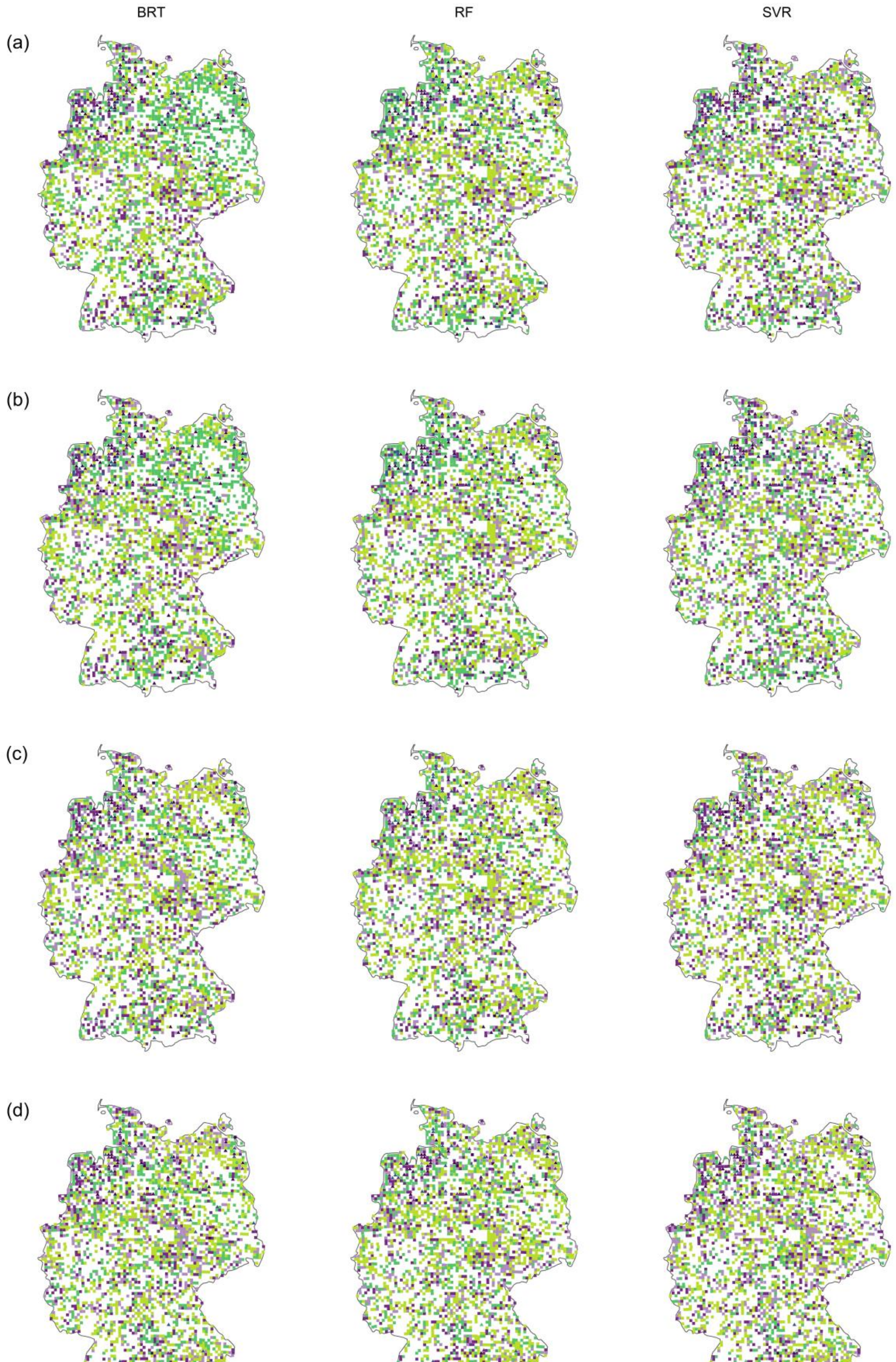

Figure 6: Spatial distribution of residuals. A) AP1 approach, B) AP1L approach, C) AP2 approach and D) AP2L approach. BRT = boosted regression trees, RF = random forest, and SVR = support vector regression.

3.3 Subdividing soil inventories into mineral and organic subsets (AP2 and AP2L)

As ~~presented~~ outlined in the sections above, the modelling of SOC content when mineral and organic soils were combined led to a systematic underprediction of soils with higher SOC values by all three algorithms, ~~regardless~~ irrespective of the number of training samples. Therefore, by implementing the AP2 approach with two models one for mineral soils and one for organic soils, a noticeable improvement in the performance of all algorithms was ~~observed,~~observed (Table S3B), with SVR showing the best error metrics (Fig. 2A6, Fig. 2B6, Fig. 2C6). This meant 34% lower RMSE, 30% lower MAE, and 32% lower MAPE than when this algorithm was trained under the AP1 approach with one model for all soils. As the high variability of SOC was initially hard to capture, the subdivision of the dataset provided a range that better represented each soil class. This was particularly beneficial for mineral soils (ranging from 4 g kg^{-1} to 85 g kg^{-1}) since the number of samples did not reduce drastically (only by 99 samples). Thus, the algorithms could better capture the relationship between SOC and covariates. Consequently, the overall performance improved when the underrepresented soil class was modelled separately. This is in line with the study of Rawlins et al. (2009) ~~which~~ that recommends ~~the~~ separate modelling of mineral and organic soils.

Nonetheless, following the AP2L approach with additional data, the RMSE and MAPE of the algorithms improved by less than 2% compared with AP2 (Table S3E). However, the greatest change was observed in the MAE of SVR with a 2% improvement. Therefore, additional training samples did not ~~considerably~~ greatly influence the performance since the majority of these samples were in mineral soils, while the limiting factor was the high variability of organic soils combined with its low number of samples. ~~Nevertheless~~However, an improvement was noted in relation to ~~the~~ all error metrics of SVR in the AP2L approach. This ~~was in~~ contrasted ~~to~~ with when the training set was enlarged without subdividing the data, i.e. AP1L. Therefore, it further confirmed that it is more important for SVR than for BRT and RF to model the soil classes separately when ~~its~~ the training set is enlarged by datasets with similar characteristics.

Furthermore, the improvement of the algorithms in AP2 and AP2L was particularly noticeable in their relative residuals. By comparing these results with those from AP1 and AP1L, it was evident that the greatest improvement was observed in the northern region and the spatial distribution of relative residuals was more homogenous throughout the country for all algorithms, but particularly for RF and SVR (Fig. 3 C and D). This is understandable since by subdividing the data, the algorithms can no longer exploit any information from the map of organic soil for spatial variability of SOC in mineral soils. Thus, they obtain information from other covariates for this soil class (Fig. 4 B). Although land use and total nitrogen were still among the most important variables for the algorithms in mineral soils, the importance of the predictors representing the SCORPAN C and P factors increased in the absence of a soil organic map. This ~~could~~ was to be expected because ~~the~~ north-east ~~of~~ Germany, for example, has a continental climate (Roßkopf et al., 2015) and young moraine landscapes, while the north-west has a more oceanic climate (Roßkopf et al., 2015) with old moraine landscapes.

It is unsurprising that all the algorithms still relied on the map of organic soil to explain SOC in organic soil class. However, while SVR and RF ~~still~~ obtained information from other covariates, the value for variable importance of this map alone ~~is~~ was 93% in BRT (Fig. 4 C),~~.~~ ~~That~~ which makes this algorithm prone to greater errors, as can be seen in its error metrics (Table S2). Similar to mineral soils, the order of covariates was different between the algorithms in organic soils. In other words, in AP1 the three algorithms obtained almost all the information from the map of organic soil, land -use and total nitrogen ~~with~~ in that ~~similar~~ order of importance. In contrast, after subdividing the data, the algorithms differed~~ntiated~~ from each other by the order of covariates in their variable importance (Fig.~~ure~~ 4).

A comparison of the error metrics of each soil class in AP2 with its counterpart in AP2L revealed that the additional 1177 samples had a minor influence on the performance (from zero to a maximum of 2%) of the algorithms in mineral soils (Table S2). These results indicated that the German Agricultural Soil Inventory offers a good representation of the spatial variability of SOC in mineral soil under agricultural use throughout the country and that the inclusion of ~~including~~ more sample points ~~do~~ did not provide additional information about SOC variability in this soil class.

However, 46 additional organic soil samples from the LUCAS dataset improved the MAPE and MAE by 12% and 6% for SVR, by 10%, and 4% for RF, and by 7% and 2% for BRT, respectively, but the RMSE of the three algorithms was improved by less than 2%. Thus, additional organic samples mainly influenced the average magnitude of the error. This could be explained by organic soils having a wide range of SOC and the number of samples ~~was~~ being limited. Thus, the addition of LUCAS data to the training set ~~offered~~ gave the algorithms more information about spatial variability of SOC in this soil class. Despite this limitation, SVR had the best overall performance among the algorithms in AP2 and AP2L. It should be noted that training samples must span the complexity of the parameter space in order for the model to be able to ~~effectively~~ match the training data effectively and ~~to~~ generalis~~z~~e unseen data. A s~~S~~mall sample size can therefore negatively influence the predictive power of the algorithms. This complexity can be addressed by structural risk minimisation (SRM) (Al-Anazi and Gates, 2012). Implementation of SRM makes SVR capable of performing well in such datasets. Other studies have compared the performance of algorithms on different sample sizes ~~for~~ in predicting soil properties and shown that SVR is one of the best choices, if not the best, when the number of samples is a limiting factor (Al-Anazi and Gates, 2012; Khaledian and Miller, 2020). In contrast, in a study by Zhou et al. (2021), 150 samples with different sets of covariates at different resolutions were used to compare RF, BRT and SVR to predict SOC content in Switzerland. Their results showed that the best-performing algorithm varied depending on the resolution and covariates. However, the best performance throughout all scenarios was obtained by BRT. The discrepancy between their results and the results of the present study may be due to the parameter-tuning method of the algorithms, as they only used grid search, or other factors, including the spatial distribution of samples or the chosen set of covariates.

Table 2: Mean of error metrics of the three models for each approach.

Approach	Mean RMSE (g kg^{-1})	Mean MAE (g kg^{-1})	Mean MAPE (%)
AP1	32.6	12.3	49.0
AP1L	32.1	12.1	46.9
AP2	21.6	8.8	34.4
AP2L	21.3	8.7	34.3

Overall, the change in performance across different sample sizes, different algorithms and different approaches (Table S3) indicated that the most important aspect of modeling SOC content of German agricultural topsoil is a two-model approach. Although combining soil inventories for more training samples can possibly improve model performance, the effect was not noticeable compared to when each soil class was predicted by its dedicated model

(Table S3B and Table S3D). The advantage of two-model approach can also be seen in the average error metrics of the three models (Table 2). While the average RMSE of the models reduces by less than 1 g kg^{-1} after enlarging the training set, the same error metrics reduces by more than 10 g kg^{-1} in AP2 and AP2L (Table 2). Therefore, it is also recommended to consider the two-model approach in soil-landscape settings similar to Germany or situations where one-model approach cannot have good predictive performance.

The map of organic soil was used to spatially distinguish each soil class to map the SOC content of the class by its corresponding model. Figure S5 shows the spatial distribution of SOC content using the AP2L approach for the three algorithms. Although SVR captured a wider range of SOC, 2 g kg^{-1} to 371.5 g kg^{-1}, than BRT, 8 g kg^{-1} to 341.1 g kg^{-1}, and RF, 7.7 g kg^{-1} to 354.6 g kg^{-1}, all three algorithms showed a relatively similar distribution of SOC content across the country particularly in mineral soils. As shown in Figure S5, organic soils are mainly distributed in the north. These soils are mostly bogs in the northwest and fens in the northeast (Roßkopf et al., 2015). There is also a small distribution of organic soil in the foothills of Alps in the south. In mineral soils, a higher SOC content is mainly found in northwest and south of the country. As explained in the previous sections, one of the main reasons for this distribution is land use since these regions are mainly under grassland while low SOC content regions are found under cropland.

4 Conclusions

The three ~~most commonly used~~ algorithms most commonly used in DSM were ~~implemented~~ applied to predict the SOC content of German agricultural soils under different approaches. Suitable tuning strategies for each algorithm ensured optimum parameter tuning and made their performance truly comparable. Machine learning ~~algorithms~~ was shown to be powerful ~~in~~ at modelling SOC on a national scale. However, the study showed that separate modelling of mineral and organic soils was a better approach for modelling SOC compared ~~to~~ with ~~using~~ just one model. Thus, this approach ~~has~~ takes priority ~~to~~ over the choice of algorithm and number of training samples. ~~We recommend~~Further testing of this approach ~~to be further tested~~is recommended in countries and regions that cover both of these soil classes. Nonetheless, SVR had a better performance than RF and BRT, except when the number of samples in training was increased by additional dataset. This was disadvantageous for SVR and advantageous for BRT unless mineral and organic soils were modelled separately. In general, increasing the number of training samples led to limited improvement of performance. Therefore, this approach should be ~~done~~ adopted ~~with~~ giving consideration of the algorithm and the characteristics of the data. Furthermore, the better performance of SVR ~~over~~ compared with that of RF and BRT was particularly highlighted when predicting SOC in organic soils. ~~Thus, this~~ The good performance of ~~algorithm~~ SVR suggests that this algorithm should therefore be taken into greater account in DSM ~~when the number of samples is limited~~.

Data availability

The soil data used in this study are publicly available via: https://doi.org/10.3220/DATA20200203151139 and https://esdac.jrc.ec.europa.eu/content/lucas-2009-topsoil-data

Author contribution

AS and AD conceptualised and developed the methodology of the presented work, with input from ML.AS gathered the predictors with contributions from AD. AS executed the programming, testing of existing code components, formal analysis and ~~visualization~~visualisation. AG contributed to the programming. The preparation of the paper was done by all authors.

Competing interests

The authors declare that they have no conflict of interest except the author AD is a member of the journal's editorial board ~~of the journal~~.

Acknowledgements

This work is part of the SoilSpace3D-DE project. The LUCAS topsoil dataset used in this work was made available by the European Commission through the European Soil Data Centre managed by the Joint Research Centre (JRC);~~,~~ http://esdac.jrc.ec.europa.eu/".

References

Al-Anazi, A. F. and Gates, I. D.: Support vector regression to predict porosity and permeability: Effect of sample size, Comput. Geosci., 39, 64–76, https://doi.org/10.1016/j.cageo.2011.06.011, 2012.

Arrouays, D., Jolivet, C., Boulonne, L., Bodineau, G., Saby, N., & Grolleau, E.: A new projection in France: a multi-institutional soil quality monitoring network, Comptes Rendus l'Académie d'Agriculture Fr., 88, 93–103, 2002.

Awad, M. and Khanna, R.: Support Vector Regression, in: Efficient Learning Machines, Apress, Berkeley, CA, 67–80, https://doi.org/10.1007/978-1-4302-5990-9_4, 2015.

Ballabio, C., Panagos, P., and Monatanarella, L.: Mapping topsoil physical properties at European scale using the LUCAS database, Geoderma, 261, 110–123, https://doi.org/10.1016/j.geoderma.2015.07.006, 2016.

Ballabio, C., Lugato, E., Fernández-Ugalde, O., Orgiazzi, A., Jones, A., Borrelli, P., Montanarella, L., and Panagos, P.: Mapping LUCAS topsoil chemical properties at European scale using Gaussian process regression, Geoderma, 355, 113912, https://doi.org/10.1016/j.geoderma.2019.113912, 2019.

Battineni, G., Chintalapudi, N., and Amenta, F.: Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (SVM), Informatics Med. Unlocked, 16, 100200, https://doi.org/10.1016/j.imu.2019.100200, 2019.

Behrens, T. and Scholten, T.: Digital soil mapping in Germany - A review, J. Plant Nutr. Soil Sci., 169, 434–443, https://doi.org/10.1002/jpln.200521962, 2006.

Behrens, T., Förster, H., Scholten, T., Steinrücken, U., Spies, E. D., and Goldschmitt, M.: Digital soil mapping using artificial neural networks, J. Plant Nutr. Soil Sci., 168, 21–33, https://doi.org/10.1002/jpln.200421414, 2005.

Belon, E., Boisson, M., Deportes, I. Z., Eglin, T. K., Feix, I., Bispo, A. O., Galsomies, L., Leblond, S., and Guellier, C. R.: An inventory of trace elements inputs to French agricultural soils, Sci. Total Environ., 439, 87–95, https://doi.org/10.1016/j.scitotenv.2012.09.011, 2012.

Bhadra, T., Bandyopadhyay, S., and Maulik, U.: Differential Evolution Based Optimization of SVM Parameters for Meta Classifier Design, Procedia Technol., 4, 50–57, https://doi.org/10.1016/j.protcy.2012.05.006, 2012.

Borrelli, P., Van Oost, K., Meusburger, K., Alewell, C., Lugato, E., and Panagos, P.: A step towards a holistic assessment of soil degradation in Europe: Coupling on-site erosion with sediment transfer and carbon fluxes, Environ. Res., 161, 291–298, https://doi.org/10.1016/j.envres.2017.11.009, 2018.

Breiman, L.: Randon Forests, 1–35, 1999.

de Brogniez, D., Ballabio, C., Stevens, A., Jones, R. J. A., Montanarella, L., and van Wesemael, B.: A map of the topsoil organic carbon content of Europe generated by a generalized additive model, Eur. J. Soil Sci., 66, 121–134, https://doi.org/10.1111/ejss.12193, 2015.

Burke, I. C., Yonker, C. M., Parton, W. J., Cole, C. V., Flach, K., and Schimel, D. S.: Texture, Climate, and Cultivation Effects on Soil Organic Matter Content in U.S. Grassland Soils, Soil Sci. Soc. Am. J., 53, 800–805, https://doi.org/10.2136/sssaj1989.03615995005300030029x, 1989.

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

Carter, B. J. and Ciolkosz, E. J.: Slope gradient and aspect effects on soils developed from sandstone in Pennsylvania, Geoderma, 49, 199–213, https://doi.org/10.1016/0016-7061(91)90076-6, 1991.

Castaldi, F., Hueni, A., Chabrillat, S., Ward, K., Buttafuoco, G., Bomans, B., Vreys, K., Brell, M., and van Wesemael, B.: Evaluating the capability of the Sentinel 2 data for soil organic carbon prediction in croplands, ISPRS J. Photogramm. Remote Sens., 147, 267–282, https://doi.org/10.1016/j.isprsjprs.2018.11.026, 2019.

Chapman, S. J., Bell, J. S., Campbell, C. D., Hudson, G., Lilly, A., Nolan, A. J., Robertson, A. H. J., Potts, J. M., and Towers, W.: Comparison of soil carbon stocks in Scottish soils between 1978 and 2009, Eur. J. Soil Sci., 64, 455–465, https://doi.org/10.1111/ejss.12041, 2013.

Cherkassky, V. and Ma, Y.: FL Methods New Genetic Technology.pdf, 17, 113–126, 2004.

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., and Böhner, J.: System for Automated Geoscientific Analyses (SAGA) v. 2.1.4, Geosci. Model Dev., 8, 1991–2007, https://doi.org/10.5194/gmd-8-1991-2015, 2015.

Deng, S., Wang, C., Wang, M., and Sun, Z.: A gradient boosting decision tree approach for insider trading identification: An empirical model evaluation of China stock market, Appl. Soft Comput. J., 83, 105652, https://doi.org/10.1016/j.asoc.2019.105652, 2019.

DWD Climate Data Center (CDC): Multi-annual grids of annual sunshine duration over Germany 1981-2010,version v1.0, 2017.

DWD Climate Data Center (CDC): Multi-annual grids of monthly averaged daily minimum air temperature (2m) over Germany, version v1.0, 2018a.

DWD Climate Data Center (CDC): Multi-annual grids of number of summer days over Germany, version v1.0, 2018b.

DWD Climate Data Center (CDC): Multi-annual grids of precipitation height over Germany 1981-2010, version v1.0, 2018c.

Environmental Systems Research Institute (ESRI): ArcGIS 10.2 for Desktop, 2013.

European Union Copernicus Land Monitoring Service, and E. E. A. (EEA): European Digital Elevation Model (EU-DEM), Version 1.1, 2016.

Eurostat grid generation tool for ArcGIS: https://www.efgs.info/information-base/best-practices/tools/eurostat-grid-generation-tool-arcgis/.

Federal Agency for Cartography and Geodesy (BKG): Digitales Basis-Landschaftsmodell (Basis-DLM), Leipzig, 2019.

Federal Institute for Geosciences and Natural Resources (BGR): Geomorphographic Map of Germany (GMK1000), Hanover, 2007.

Federal Institute for Geosciences and Natural Resources (BGR): Soil scapes in Germany 1:5,000,000 (BGL5000), Hanover, 2008.

Federal Institute for Geosciences and Natural Resources (BGR) and German State Geological Surveys (SDG): Hydrogeological Map of Germany 1:250,000 (HÜK250), Hanover, 2019.

Forkuor, G., Hounkpatin, O. K. L., Welp, G., and Thiel, M.: High resolution mapping of soil properties using Remote Sensing variables in south-western Burkina Faso: A comparison of machine learning and multiple linear regression models, PLoS One, 12, 1–21, https://doi.org/10.1371/journal.pone.0170478, 2017.

Friedman, J., Tibshirani, R., and Hastie, T.: Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors), Ann. Stat., 28, 337–407, https://doi.org/10.1214/aos/1016120463, 2000.

Friedman, J., Hastie, T., and Tibshirani, R.: The Elements of Statistical Learning, second., Springer New York, New York, NY, 158-61 pp., https://doi.org/10.1007/b94608, 2009.

Friedman, J. H.: Stochastic gradient boosting, Comput. Stat. Data Anal., 38, 367–378, https://doi.org/10.1016/S0167-9473(01)00065-2, 2002.

Gebauer, A., Ellinger, M., M. Brito Gomez, V., and Ließ, M.: Development of pedotransfer functions for water retention in tropical mountain soil landscapes: Spotlight on parameter tuning in machine learning, 6, 215–229, https://doi.org/10.5194/soil-6-215-2020, 2020.

Greenwell, B., Boehmke, B., and Cunningham, J.: Package “gbm” - Generalized Boosted Regression Models, CRAN Repos., 39, 2019.

Guevara, M., Olmedo, G. F., Stell, E., Yigini, Y., Aguilar Duarte, Y., Arellano Hernández, C., Arévalo, G., Arroyo-Cruz, C. E., Bolivar, A., Bunning, S., Bustamante Cañas, N., Cruz-Gaistardo, C. O., Davila, F., Dell Acqua, M., Encina, A., Figueredo Tacona, H., Fontes, F., Hernández Herrera, J. A., Ibelles Navarro, A. R., Loayza, V., Manueles, A., Mendoza Jara, F., Olivera, C., Osorio Hermosilla, R., Pereira, G., Prieto, P., Alexis Ramos, I., Rey Brina, J. C., Rivera, R., Rodríguez-Rodríguez, J., Roopnarine, R., Rosales Ibarra, A., Rosales Riveiro, K. A., Schulz, G. A., Spence, A., Vasques, G., Vargas, R., and Vargas, R.: No Silver Bullet for Digital Soil Mapping: Country-specific Soil Organic Carbon Estimates across Latin America, SOIL Discuss., 1–20, https://doi.org/10.5194/soil-2017-40, 2018.

Guio Blanco, C. M., Brito Gomez, V. M., Crespo, P., and Ließ, M.: Spatial prediction of soil water retention in a Páramo landscape: Methodological insight into machine learning using random forest, Geoderma, 316, 100–114, https://doi.org/10.1016/j.geoderma.2017.12.002, 2018.

Hawkins, D. M., Basak, S. C., and Mills, D.: Assessing model fit by cross-validation, J. Chem. Inf. Comput. Sci., 43, 579–586, https://doi.org/10.1021/ci025626i, 2003.

Hornik, K., Weingessel, A., Leisch, F., and Davidmeyerr-projectorg, M. D. M.: Package 'e1071,' 2021.

Hoyle, F. C., Baldock, J. A., and Murphy, D. V: Soil Organic Carbon – Role in Rainfed Farming Systems, Rainfed Farming Syst., https://doi.org/10.1007/978-1-4020-9132-2, 2011.

Khaledian, Y. and Miller, B. A.: Selecting appropriate machine learning methods for digital soil mapping, Appl. Math. Model., 81, 401–418, https://doi.org/10.1016/j.apm.2019.12.016, 2020.

Kuhn, M. and Johnson, K.: Applied predictive modeling, 1st ed., Springer-Verlag, New York, 1–600 pp., https://doi.org/10.1007/978-1-4614-6849-3, 2013.

Lal, R.: Soil carbon sequestration impacts on global climate change and food security, Science (80-.)., 304, 1623–1627, https://doi.org/10.1126/science.1097396, 2004.

Li, T., Zhang, H., Wang, X., Cheng, S., Fang, H., Liu, G., and Yuan, W.: Soil erosion affects variations of soil organic carbon and soil respiration along a slope in Northeast China, Ecol. Process., 8, https://doi.org/10.1186/s13717-019-0184-6, 2019.

Li, X., Ding, J., Liu, J., Ge, X., and Zhang, J.: Digital mapping of soil organic carbon using sentinel series data: A case study of the ebinur lake watershed in xinjiang, Remote Sens., 13, 1–19, https://doi.org/10.3390/rs13040769, 2021.

Liang, W., Zhang, L., and Wang, M.: The chaos differential evolution optimization algorithm and its application to support vector regression machine, J. Softw., 6, 1297–1304, https://doi.org/10.4304/jsw.6.7.1297-1304, 2011.

Ließ, M., Gebauer, A., and Don, A.: Machine Learning With GA Optimization to Model the Agricultural Soil-Landscape of Germany: An Approach Involving Soil Functional Types With Their Multivariate Parameter Distributions Along the Depth Profile, Front. Environ. Sci., 9, 1–24, https://doi.org/10.3389/fenvs.2021.692959, 2021.

Malik, A. A., Puissant, J., Buckeridge, K. M., Goodall, T., Jehmlich, N., Chowdhury, S., Gweon, H. S., Peyton, J. M., Mason, K. E., van Agtmaal, M., Blaud, A., Clark, I. M., Whitaker, J., Pywell, R. F., Ostle, N., Gleixner, G., and Griffiths, R. I.: Land use driven change in soil pH affects microbial carbon cycling processes, Nat. Commun., 9, 1–10, https://doi.org/10.1038/s41467-018-05980-1, 2018.

Martin, M. P., Orton, T. G., Lacarce, E., Meersmans, J., Saby, N. P. A., Paroissien, J. B., Jolivet, C., Boulonne, L., and Arrouays, D.: Evaluation of modelling approaches for predicting the spatial distribution of soil organic carbon stocks at the national scale, Geoderma, 223–225, 97–107, https://doi.org/10.1016/j.geoderma.2014.01.005, 2014.

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

Meersmans, J., Martin, M. P., Lacarce, E., De Baets, S., Jolivet, C., Boulonne, L., Lehmann, S., Saby, N. P. A., Bispo, A., and Arrouays, D.: A high resolution map of French soil organic carbon, Agron. Sustain. Dev., 32, 841–851, https://doi.org/10.1007/s13593-012-0086-9, 2012a.

Meersmans, J., Martin, M. P., De Ridder, F., Lacarce, E., Wetterlind, J., De Baets, S., Bas, C. Le, Louis, B. P., Orton, T. G., Bispo, A., and Arrouays, D.: A novel soil organic C model using climate, soil type and management data at the national scale in France, Agron. Sustain. Dev., 32, 873–888, https://doi.org/10.1007/s13593-012-0085-x, 2012b.

Minasny, B., McBratney, A. B., Malone, B. P., and Wheeler, I.: Digital Mapping of Soil Carbon, Elsevier, 1–47 pp., https://doi.org/10.1016/B978-0-12-405942-9.00001-3, 2013.

Mulder, V. L., Lacoste, M., Richer-de-Forges, A. C., Martin, M. P., and Arrouays, D.: National versus global modelling the 3D distribution of soil organic carbon in mainland France, Geoderma, 263, 16–34, https://doi.org/10.1016/J.GEODERMA.2015.08.035, 2016.

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

Pei, T., Qin, C. Z., Zhu, A. X., Yang, L., Luo, M., Li, B., and Zhou, C.: Mapping soil organic matter using the topographic wetness index: A comparative study based on different flow-direction algorithms and kriging methods, Ecol. Indic., 10, 610–619, https://doi.org/10.1016/j.ecolind.2009.10.005, 2010.

Peterson, B., Ulrich, J., and Boudt, K.: Package ‘ DEoptim ,’ https://doi.org/10.18637/jss.v040.i06, 2021.

Poeplau, C., Bolinder, M. A., Eriksson, J., Lundblad, M., and Kätterer, T.: Positive trends in organic carbon storage in Swedish agricultural soils due to unexpected socio-economic drivers, 12, 3241–3251, https://doi.org/10.5194/bg-12-3241-2015, 2015.

Poeplau, C., Jacobs, A., Don, A., Vos, C., Schneider, F., Wittnebel, M., Tiemeyer, B., Heidkamp, A., Prietz, R., and Flessa, H.: Stocks of organic carbon in German agricultural soils—Key results of the first comprehensive inventory, J. Plant Nutr. Soil Sci., 183, 665–681, https://doi.org/10.1002/jpln.202000113, 2020.

Prechtel, A., Von L̈tzow, M., Schneider, B. U., Bens, O., Bannick, C. G., Kögel-Knabner, I., and Hüttl, R. F.: Organic Carbon in soils of Germany: Status quo and the need for new data to evaluate potentials and trends of soil carbon sequestration, J. Plant Nutr. Soil Sci., 172, 601–614, https://doi.org/10.1002/jpln.200900034, 2009.

Ramifehiarivo, N., Brossard, M., Grinand, C., Andriamananjara, A., Razafimbelo, T., Rasolohery, A., Razafimahatratra, H., Seyler, F., Ranaivoson, N., Rabenarivo, M., Albrecht, A., Razafindrabe, F., and Razakamanarivo, H.: Mapping soil organic carbon on a national scale: Towards an improved and updated map of Madagascar, Geoderma Reg., 9, 29–38, https://doi.org/10.1016/j.geodrs.2016.12.002, 2017.

Rawlins, B. G., Marchant, B. P., Smyth, D., Scheib, C., Lark, R. M., and Jordan, C.: Airborne radiometric survey data and a DTM as covariates for regional scale mapping of soil organic carbon across Northern Ireland, Eur. J. Soil Sci., 60, 44–54, https://doi.org/10.1111/j.1365-2389.2008.01092.x, 2009.

Reeves, D. W.: The role of soil organic matter in maintaining soil quality in continuous cropping systems, Soil Tillage Res., 43, 131–167, https://doi.org/10.1016/S0167-1987(97)00038-X, 1997.

Ritchie, J. C., McCarty, G. W., Venteris, E. R., and Kaspar, T. C.: Soil and soil organic carbon redistribution on the landscape, 89, 163–171, https://doi.org/10.1016/j.geomorph.2006.07.021, 2007.

Roßberg, D., Michel, V., Graf, R., Neukampf, R.: Definition von Boden-Klima-Räumen für die Bundesrepublik Deutschland, Nachrichtenblatt des Dtsch. Pflanzenschutzdienstes, 59, 155–161, 2007.

Roßkopf, N., Fell, H., and Zeitz, J.: Organic soils in Germany, their distribution and carbon stocks, 133, 157–170, https://doi.org/10.1016/j.catena.2015.05.004, 2015.

Santos, C. E. da S., Sampaio, R. C., Coelho, L. dos S., Bestarsd, G. A., and Llanos, C. H.: Multi-objective adaptive differential evolution for SVM/SVR hyperparameters selection, Pattern Recognit., 110, 107649, https://doi.org/10.1016/j.patcog.2020.107649, 2021.

Schapire, R. E.: The Boosting Approach to Machine Learning: An Overview, 149–171, https://doi.org/10.1007/978-0-387-21579-2_9, 2003.

Schneider, F., Amelung, W., and Don, A.: Origin of carbon in agricultural soil profiles deduced from depth gradients of C:N ratios, carbon fractions, δ13C and δ15N values, Plant Soil, 460, 123–148, https://doi.org/10.1007/s11104-020-04769-w, 2021.

Smola, A. J. and Schölkopf, B.: A tutorial on support vector regression, Stat. Comput., 14, 199–222, https://doi.org/10.1023/B:STCO.0000035301.49549.88, 2004.

Storn, R. and Price, K.: Differential Evolution - A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces, J. Glob. Optim., 11, 341–359, https://doi.org/10.1023/A:1008202821328, 1997.

Taghizadeh-Toosi, A., Olesen, J. E., Kristensen, K., Elsgaard, L., Østergaard, H. S., Lægdsmand, M., Greve, M. H., and Christensen, B. T.: Changes in carbon stocks of Danish agricultural mineral soils between 1986 and 2009, Eur. J. Soil Sci., 65, 730–740, https://doi.org/10.1111/ejss.12169, 2014.

Tóth, G., Jones, A., and Montanarella, L.: LUCAS Topsoil Survey: Methodology, Data, and Results, https://doi.org/10.2788/97922, 2013.

Tziachris, P., Aschonitis, V., Chatzistathis, T., Papadopoulou, M., and Doukas, I. J. D.: Comparing machine learning models and hybrid geostatistical methods using environmental and soil covariates for soil pH prediction, ISPRS Int. J. Geo-Information, 9, https://doi.org/10.3390/ijgi9040276, 2020.

Varma, S. and Simon, R.: Bias in error estimation when using cross-validation for model selection, BMC Bioinformatics, 7, 1–8, https://doi.org/10.1186/1471-2105-7-91, 2006.

Wadoux, A. M. J. C., Minasny, B., and McBratney, A. B.: Machine learning for digital soil mapping: Applications, challenges and suggested solutions, Earth-Science Rev., 210, https://doi.org/10.1016/j.earscirev.2020.103359, 2020.

Wang, S., Xu, L., Zhuang, Q., and He, N.: Investigating the spatio-temporal variability of soil organic carbon stocks in different ecosystems of China, Sci. Total Environ., 758, https://doi.org/10.1016/j.scitotenv.2020.143644, 2021.

Wang, X., Zhang, Y., Atkinson, P. M., and Yao, H.: Predicting soil organic carbon content in Spain by combining Landsat TM and ALOS PALSAR images, Int. J. Appl. Earth Obs. Geoinf., 92, 102182, https://doi.org/10.1016/j.jag.2020.102182, 2020.

Ward, K. J., Chabrillat, S., Neumann, C., and Foerster, S.: A remote sensing adapted approach for soil organic carbon prediction based on the spectrally clustered LUCAS soil database, Geoderma, 353, 297–307, https://doi.org/10.1016/j.geoderma.2019.07.010, 2019.

Were, K., Bui, D. T., Dick, Ø. B., and Singh, B. R.: A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape, Ecol. Indic., 52, 394–403, https://doi.org/10.1016/j.ecolind.2014.12.028, 2015.

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.

Wright, M. N. and Ziegler, A.: Ranger: A fast implementation of random forests for high dimensional data in C++ and R, J. Stat. Softw., 77, 1–17, https://doi.org/10.18637/jss.v077.i01, 2017.

Yu, D., Hu, F., Zhang, K., Liu, L., and Li, D.: Available water capacity and organic carbon storage profiles in soils developed from dark brown soil to boggy soil in Changbai Mountains, China, Soil Water Res., 16, 11–21, https://doi.org/10.17221/150/2019-SWR, 2021.

Zhang, J., Niu, Q., Li, K., and Irwin, G. W.: Model selection in SVMs using Differential Evolution, IFAC, 14717–14722 pp., https://doi.org/10.3182/20110828-6-IT-1002.00584, 2011.

Zhong, Z., Chen, Z., Xu, Y., Ren, C., Yang, G., Han, X., Ren, G., and Feng, Y.: Relationship between soil organic carbon stocks and clay content under different climatic conditions in Central China, 9, 1–14, https://doi.org/10.3390/f9100598, 2018.

Zhou, T., Geng, Y., Ji, C., Xu, X., Wang, H., Pan, J., Bumberger, J., Haase, D., and Lausch, A.: Prediction of soil organic carbon and the C:N ratio on a national scale using machine learning and satellite data: A comparison between Sentinel-2, Sentinel-3 and Landsat-8 images, Sci. Total Environ., 755, https://doi.org/10.1016/j.scitotenv.2020.142661, 2021.

Spatial prediction of organic carbon in German agricultural topsoil using machine learning algorithms~~Performance of three machine learning algorithms for predicting soil organic carbon in German agricultural soil~~

Ali Sakhaee[1], Anika Gebauer[2], Mareike Ließ[2], Axel Don[1]
[1]Thünen Institute of Climate Smart Agriculture, Braunschweig, Germany
[2]Department Soil System Science, Helmholtz Centre for Environmental Research – UFZ, Halle (Saale), Germany

Correspondence to: Ali Sakhaee (a.sakhaee@thuenen.de)

Supplements

Table S1: ~~The range of parameters for tuning in full dataset (AP1 and AP1L) and mineral and organic soil subsets (AP2 and AP2L)~~Tuning parameter ranges corresponding to the models trained by different algorithms~~.~~ The ranges differ considering the one-model (full dataset, AP1 & AP1L) or two-model approach (mineral and organic data subset, AP2 & AP2L). BRT = boosted regression trees, RF = random forest, and SVR = support vector regression.

Algorithm	Tuning p~~P~~arameter	Full dataset	Mineral soil data subset	Organic soil data subset
SVR	C	1-100	1-50	1-200
	epsilon	0-5	0-1	0-5
	gamma	0.001-1	0.001-1	0.001-1
RF	mtry	3-13	3-13	3-13
BRT	number of trees	100-3000	100-3000	100-3000
	shrinkage	0.001-0.1	0.001-0.1	0.001-0.1
	interaction depth	1-5	1-5	1-5
	bag fraction	0.5-0.9	0.5-0.9	0.5-0.9

Table S2: ~~Error metrics of the algorithms~~Predictive model performance of the models trained with different machine learning algorithms and datasets: A) built on the German Agricultural Soil Inventory, B) including LUCAS data in the training set. BRT = boosted regression trees, RF = random forest, and SVR = support vector regression.

	Algorithm	RMSE	MAE	%MAPE	%Bias	AIC	BIC	Approach
A	BRT	32.9	12.4	50.9	-32	14865	14889	AP1
	RF	33.2	12.3	48.6	-30	14913	14919	AP1
	SVR	31.6	12.3	47.4	-20	14643	14661	AP1
	BRT	9.5	6.2	35.9	-20	7500	7524	Mineral
	RF	9.1	5.9	34	-20	7288	7294	Mineral
	SVR	9.2	5.8	31.8	-10	7331	7349	Mineral
	BRT	107	90.4	48.5	-26	757	768	Organic
	RF	106.1	89.3	48.2	-28	750	753	Organic
	SVR	101.7	86.9	45.6	-22	746	754	Organic
	BRT	22	9.1	36.3	-20	12578	12602	AP2
	RF	21.7	8.8	34.5	-20	12496	12502	AP2
	SVR	21	8.6	32.3	-10	12310	12328	AP2

	Algorithm	RMSE	MAE	%MAPE	%Bias	AIC	BIC	Approach
B	BRT	31.3	11.8	47.4	-30	14568	14592	AP1L
	RF	32.5	12.1	46.8	-30	14754	14759	AP1L
	SVR	32.6	12.3	46.4	-20	14775	14792	AP1L
	BRT	9.4	6.2	35.6	-20	7429	7453	Mineral
	RF	9.1	6	34.6	-20	7268	7274	Mineral
	SVR	9.1	5.8	31.7	-10	7275	7293	Mineral
	BRT	105.4	88.4	45	-20	754	765	Organic
	RF	104.1	86.2	43.5	-20	745	748	Organic
	SVR	100.2	81.7	40.2	-12	741	749	Organic
	BRT	21.7	9	36	-20	12486	12510	AP2L
	RF	21.4	8.7	34.9	-20	12379	12385	AP2L
	SVR	20.7	8.4	31.9	-10	12191	12209	AP2L

Table S3: Percent change in predictive model performance comparing models trained with different machine learning algorithms and data sets: A) and B) comparison of models trained by using data from the –German Agricultural Soil Inventory, C) and D) comparison of models trained by using data from the German Agricultural Soil Inventory and LUCAS, A) and C) comparison of models trained with~~comparison with regards to the~~ different machine learning algorithms, B) and D) comparison of the one-model approach (AP1) to the two-model approach (AP2), E) comparison of the approaches before and after including LUCAS. BRT = boosted regression trees, RF = random forest, and SVR = support vector regression.

	Algorithm	RMSE (%)	MAE (%)	MAPE (%)	Approach
A	BRT to RF	0.9	-0.8	-4.5	AP1
	RF to SVR	-4.8	0.0	-2.5	AP1
	BRT to SVR	-4.0	-0.8	-6.9	AP1
	BRT to RF	-4.2	-4.8	-5.3	Mineral
	RF to SVR	1.1	-1.7	-6.5	Mineral
	BRT to SVR	-3.2	-6.5	-11.4	Mineral
	BRT to RF	-0.8	-1.2	-0.6	Organic
	RF to SVR	-4.1	-2.7	-5.4	Organic
	BRT to SVR	-5.2	-4.0	-6.0	Organic
	BRT to RF	-1.4	-3.3	-5.0	AP2
	RF to SVR	-3.2	-2.3	-6.4	AP2
	BRT to SVR	-4.5	-5.5	-11.0	AP2
	Algorithm	RMSE	MAE	MAPE	Approach
B	BRT	-33.1	-26.6	-28.7	AP1 to AP2
	RF	-34.6	-28.5	-29.0	AP1 to AP2
	SVR	-33.5	-30.1	-31.9	AP1 to AP2
	Algorithm	RMSE	MAE	MAPE	Approach

C	BRT to RF	3.8	2.5	-1.3	AP1L
	RF to SVR	0.3	1.7	-0.9	AP1L
	BRT to SVR	4.2	4.2	-2.1	AP1L
	BRT to RF	-3.2	-3.2	-2.8	Mineral
	RF to SVR	0.0	-3.3	-8.4	Mineral
	BRT to SVR	-3.2	-6.5	-11.0	Mineral
	BRT to RF	-1.2	-2.5	-3.3	Organic
	RF to SVR	-3.7	-5.2	-7.6	Organic
	BRT to SVR	-5.2	-8.2	-10.7	Organic
	BRT to RF	-1.4	-3.3	-3.1	AP2L
	RF to SVR	-3.3	-3.4	-8.6	AP2L
	BRT to SVR	-4.6	-6.7	-11.4	AP2L
	Algorithm	RMSE	MAE	MAPE	Approach
D	BRT	-30.7	-23.7	-24.1	AP1L to AP2L
	RF	-34.2	-28.1	-25.4	AP1L to AP2L
	SVR	-36.5	-31.7	-31.3	AP1L to AP2L
	Algorithm	RMSE	MAE	MAPE	Approach
E	BRT	-4.9	-4.8	-6.9	AP1 to AP1L
	RF	-2.1	-1.6	-3.7	AP1 to AP1L
	SVR	3.2	0.0	-2.1	AP1 to AP1L
	BRT	-1.1	0.0	-0.8	Mineral
	RF	0.0	1.7	1.8	Mineral
	SVR	-1.1	0.0	-0.3	Mineral
	BRT	-1.5	-2.2	-7.2	Organic
	RF	-1.9	-3.5	-9.8	Organic
	SVR	-1.5	-6.0	-11.8	Organic
	BRT	-1.4	-1.1	-0.8	AP2 to AP2L
	RF	-1.4	-1.1	1.2	AP2 to AP2L
	SVR	-1.4	-2.3	-1.2	AP2 to AP2L

Table S4~~3~~: List of covariates, their abbreviations and ~~reference~~their SCORPAN ID.

SCORPAN ID	Covariates	Abbreviation
S	Net erosion	Net-Ero
	Available water capacity	AWC
	Total nitrogen	TN
	pH	pH
	Soil organic map	Peat

	Clay content	Clay
C	Multi-annual grid of annual sunshine duration over Germany	Sun-Dur
	Multi-annual grids of number of summer days over Germany	Summ-D
	Multi-annual grids of monthly averaged daily minimum air temperature (2m) over Germany	Min-temp
	Multi-annual grids of precipitation height over Germany	Precip
O	Landuse	DLM
R	Digital elevation model	EU-DEM
	Slope	Slope
	Aspect north south direcion	Aspect-NS
	Aspect east west direction	Aspect-EW
	Plan Curvature	Plan-Curv
	Profile curvature	Prof-Curv
	Topographic wetness index	TWI
	Geomorphographic map	GMK
P	Large-scale landscape unit map (Bodengrosslandschaft)	BGL
	Large-scale soil climate region map (Bodenklima)	Bod-klim
	Hydrological unit	HUK-HE
N	X coordination	x
	Y coordination	y

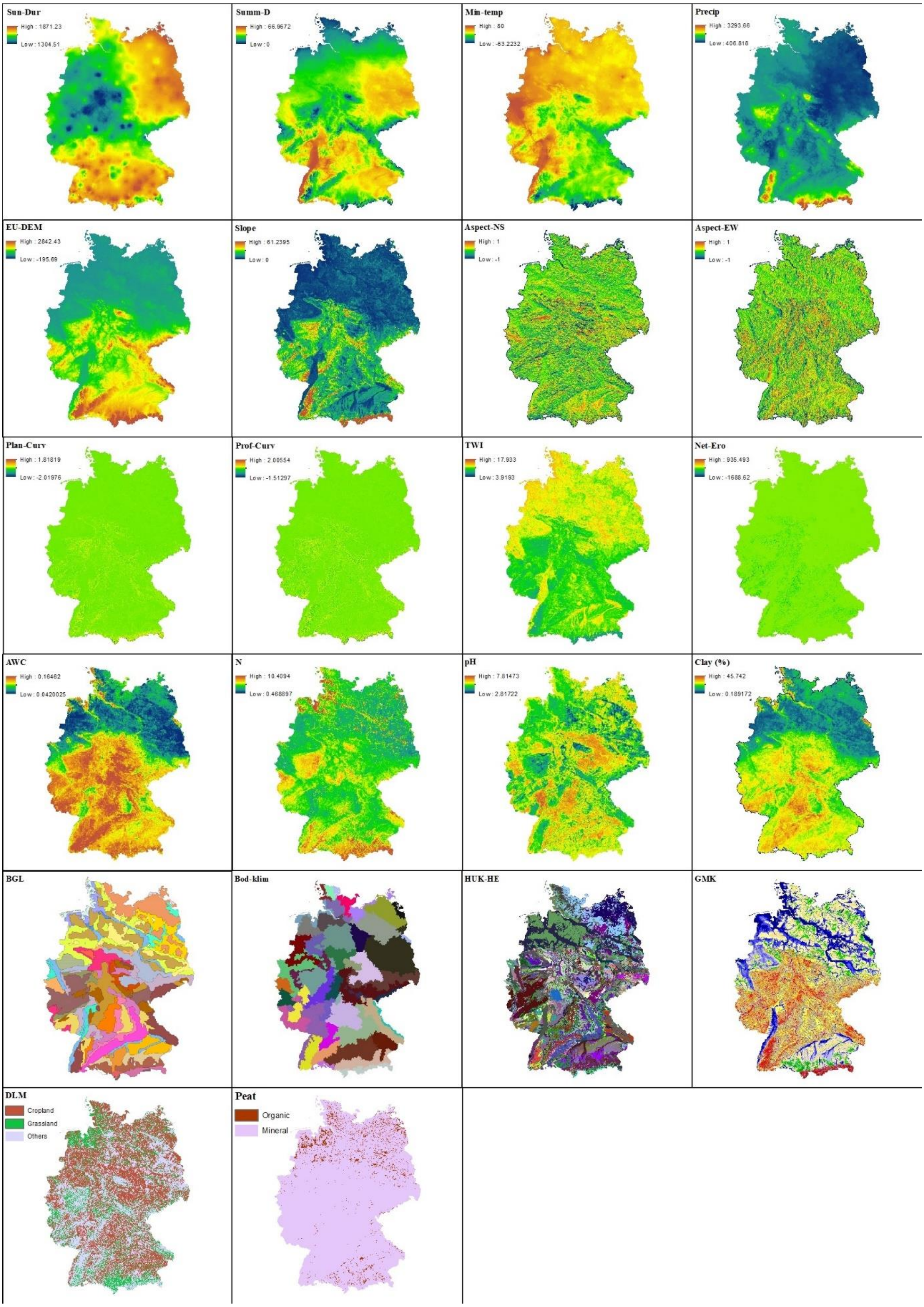

Figure S1: Selected covariates: Sun-Dur) sunshine duration (DWD, 2017), Summ-D) summer days (DWD, 2018b), Min-temp) minimum temperature (DWD, 2018a), Precip) precipitation (DWD, 2018c), EU-DEM) digital elevation model (European Union Copernicus Land Monitoring Service, 2016), Net-Ero) net soil erosion and deposition rates (Borrelli et al., 2018), AWC) available water capacity (Ballabio et al., 2016), N) total nitrogen (Ballabio et al., 2019), pH) map of pH (Ballabio et al., 2019), %Clay) % Clay (Ballabio et al., 2016), BGL) soil scapes unit (BGR, 2008) [Legend], Bod-

Klim) soil-climate region (Roßberg et al., 2007), HUK-HE) hydrogeological unit of hydrogeological map (BGR, SDG, 2019), GMK) geomorphographic map of Germany (BGR, 2007) [Legend], DLM) Land use (BKG, 2019), Peat) Organic soils (Roßkopf et al., 2015).

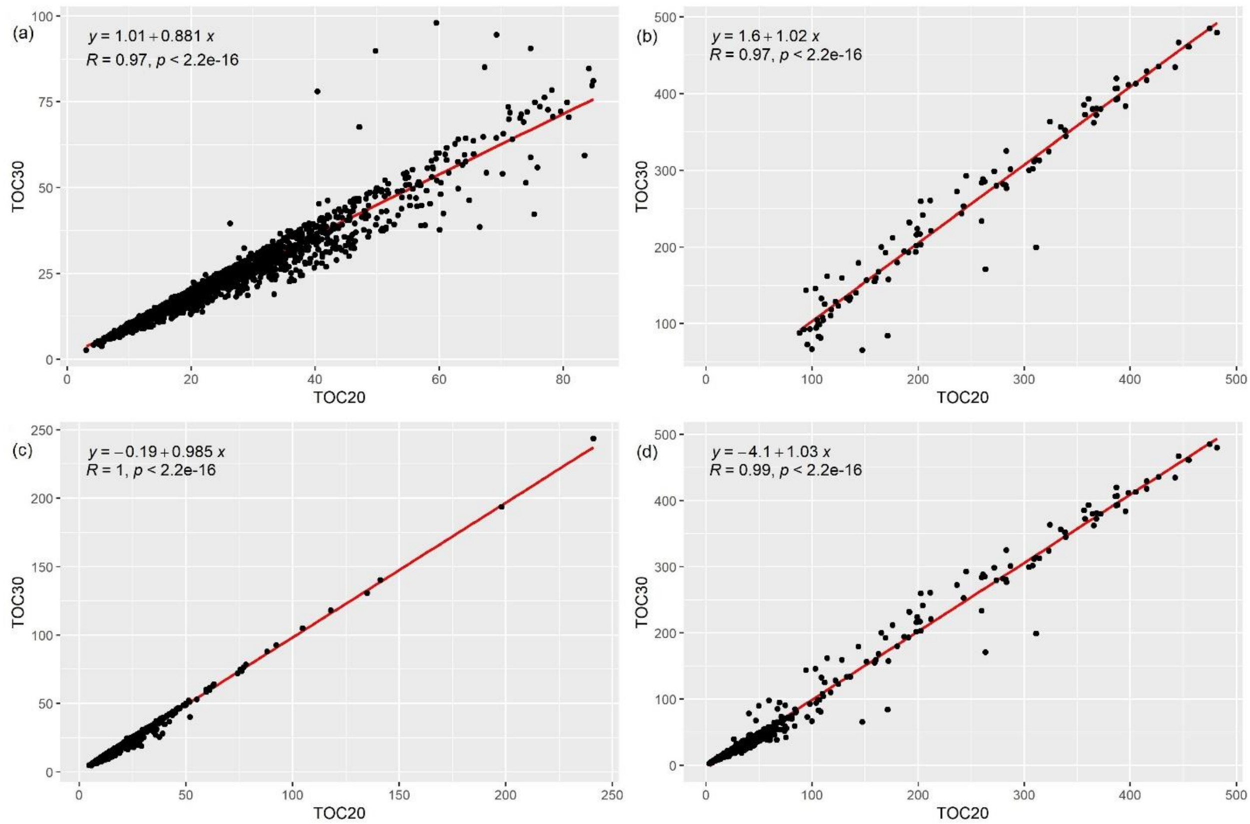

Figure S2: Regression plot for SOC depth extrapolation in A) Mineral soils, B) Organic soils, C) Cropland, D) Grassland.

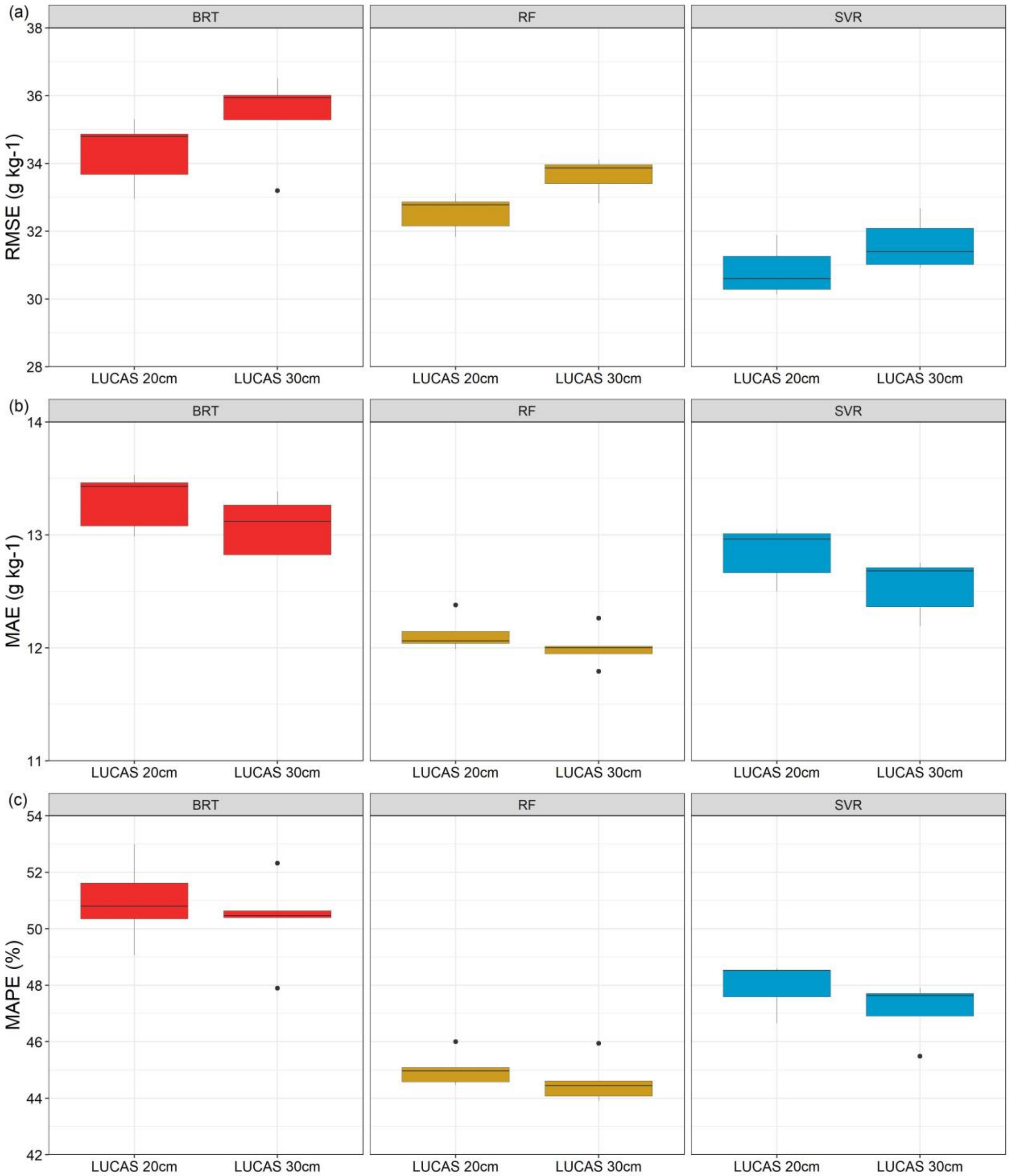

Figure S3~~1~~: Boxplots comparing ~~algorithm~~ model performance with regards to the three machine learning algorithms considering~~.~~ LUCAS at the original sampling depth (20 cm) versus LUCAS with depth extrapolated (30 cm): A) RMSE (g kg^{-1}), B) MAE (g kg^{-1}) and C) MAPE (%). BRT = boosted regression trees, RF = random forest, and SVR = support vector regression.

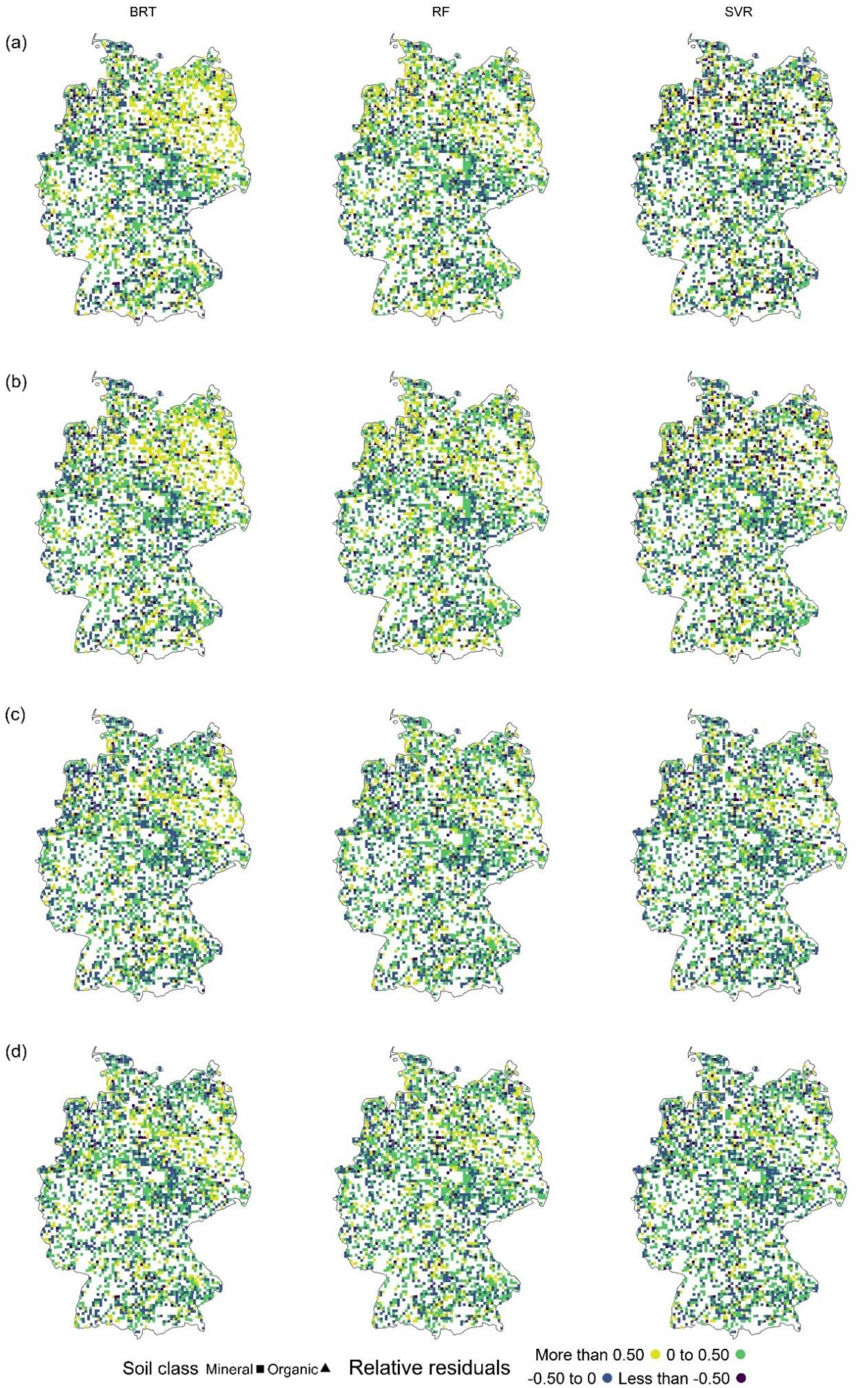

Figure S4: Spatial distribution of relative residuals from the models trained with the different machine learning algorithms. A) AP1 approach, B) AP1L approach, C) AP2 approach and D) AP2L approach. BRT = boosted regression trees, RF = random forest, and SVR = support vector regression.

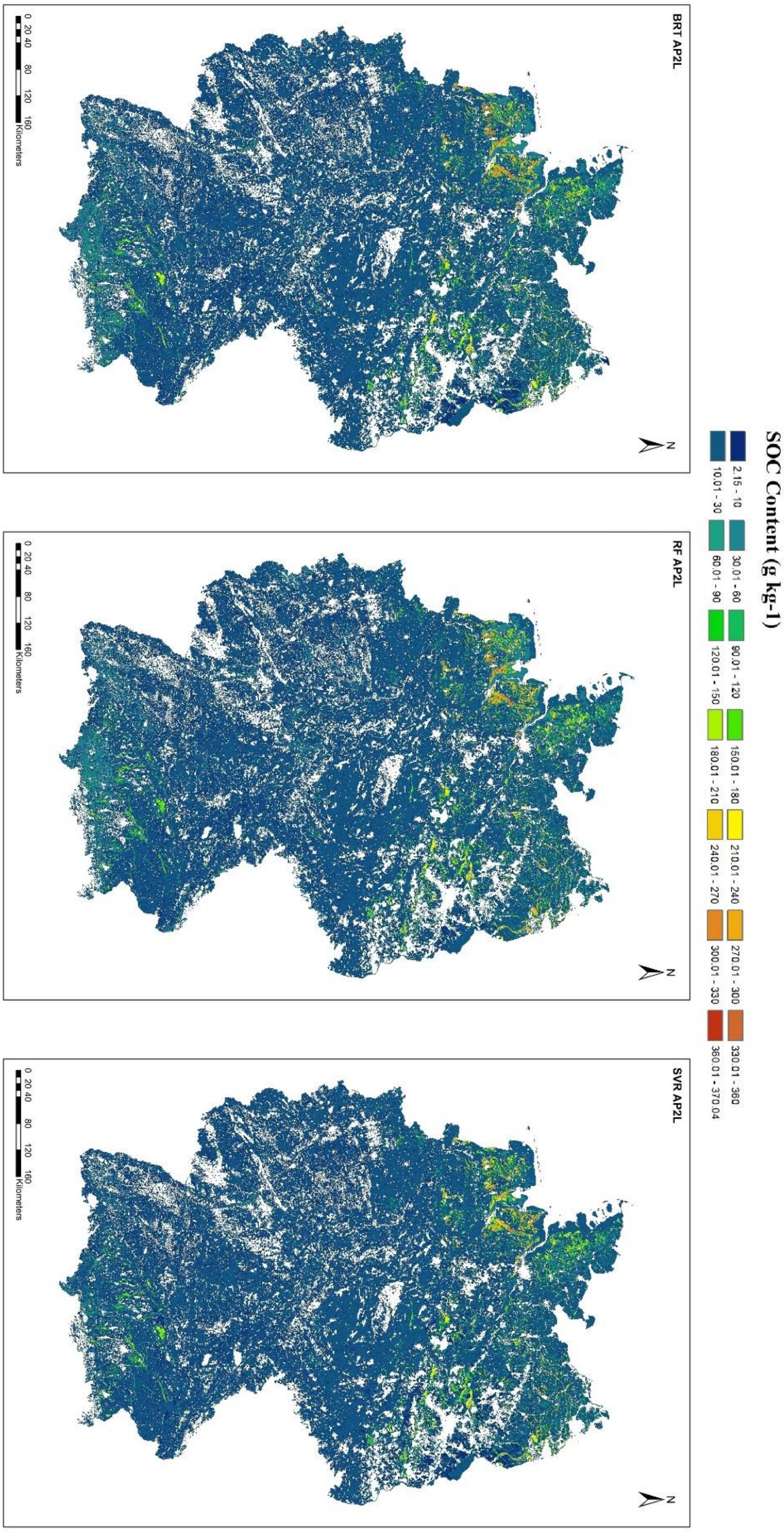

Figure S5: Spatial prediction of SOC content (g kg-1) of German agricultural soils based on the two-model approach for the three algorithms (BRT AP2L, RF AP2L, SVR AP2L). BRT = boosted regression trees, RF = random forest, and SVR = support vector regression.

Disclaimer: It is important to note that the provided spatial prediction of SOC content must not be used to identify the organic soils of Germany or to determine their spatial distribution. One reason is low sample size of organic soils and the systematic underestimation of their SOC content, which leads to an underestimation of their spatial extent. Furthermore, organic soils might have been mixed with mineral soil, i.e. due to deep ploughing, or feature a mineral soil cover. Thus, organic soils might be present despite having a mineral topsoils. Therefore, this study cannot nor intents to identify or classify organic soils.