Eesti Maaülikool
Estonian University of Life Sciences

PREDICTION MODELS OF SOIL ORGANIC CARBON AND BULK DENSITY OF ARABLE MINERAL SOILS

MINERAALSETE PÕLLUMULDADE ORGAANILISE SÜSINIKU JA LASUVUSTIHEDUSE STATISTILISED PROGNOOSIMUDELID

ELSA PUTKU

A thesis
for applying for the degree of Doctor of Philosophy
in Agriculture

Väitekiri
filosoofiadoktori kraadi taotlemiseks
põllumajanduse erialal

Tartu 2016

EESTI MAAÜLIKOOL
ESTONIAN UNIVERSITY OF LIFE SCIENCES

Eesti Maaülikool
Estonian University of Life Sciences

PREDICTION MODELS OF SOIL ORGANIC CARBON AND BULK DENSITY OF ARABLE MINERAL SOILS

MINERAALSETE PÕLLUMULDADE ORGAANILISE SÜSINIKU JA LASUVUSTIHEDUSE STATISTILISED PROGNOOSIMUDELID

ELSA PUTKU

A thesis
for applying for the degree of Doctor of Philosophy
in Agriculture

Väitekiri
filosoofiadoktori kraadi taotlemiseks
põllumajanduse erialal

Tartu 2016

Institute of Agricultural and Environmental Sciences
Estonian University of Life Sciences

According to verdict No 6-14/3-6, May 3, 2016, the Doctoral Committee of the Agricultural Sciences of the Estonian University of Life Sciences has accepted the thesis for the defence of the degree of Doctor of Philosophy in Agriculture.

Opponent: Prof. **Thomas Kätterer**, PhD
Department of Ecology
Swedish University of Agricultural Sciences

Supervisors: Prof. **Alar Astover**, PhD
Institute of Agricultural and Environmental Sciences
Estonian University of Life Sciences

Assoc. Prof. **Christian Ritz**, PhD
Department of Nutrition, Exercise and Sports
University of Copenhagen

Defence of the thesis:
Estonian University of Life Sciences, room 2A1, Kreutzwaldi 5, Tartu on 14th of June, 2016, at 10:15

The English in the current thesis was revised by Christian Ritz and the Estonian by Kalle Hein.

Publication of this thesis is supported by the Estonian University of Life Sciences, Department of Soil Science and Agrochemistry.

© Elsa Putku, 2016
ISSN 2382-7076
ISBN 978-9949-569-27-4 (trükis)
ISBN 978-9949-569-28-1 (pdf)

CONTENTS

LIST OF ORIGINAL PUBLICATIONS 7
ABBREVIATIONS 8
1. INTRODUCTION 9
2. REVIEW OF THE LITERATURE 12
2.1. Overview of prediction approaches in soil science 12
2.1.1. Prediction of SOC 13
2.1.2. Soil bulk density and pedotransfer functions 16
2.2. Calculating soil organic carbon stock 20
2.3. SOC stocks globally and locally 20
2.4. Related studies in Estonia 21
3. HYPOTHESIS AND AIMS OF THE STUDY 23
4. MATERIALS AND METHODS 24
4.1. Description of the dataset 24
4.1.1. Monitoring scheme 24
4.1.2. Measured soil properties 25
4.2. Framework of prediction 26
4.3. Prediction models 27
4.3.1. Linear regression model 27
4.3.2. Linear mixed model 28
4.4. A dataset used for kriging 29
4.5. The Estonian large-scale soil map and area of case study 30
4.6. Software 31
5. RESULTS 32
5.1. Soil characteristics 32
5.2. Prediction models for D_b 32
5.3. Prediction models of SOC concentration 33
5.3.1. Using site spatial information for kriging 33
5.4. Combining predictions of separate SOC (%) and D_b models to calculate SOC stock 34
5.5. Model application: a case study of Tartu County 37

6. DISCUSSION 40
6.1. Comparison of prediction models 40
6.2. Limitations and perspectives 44
7. CONCLUSIONS 46
REFERENCES 48
SUMMARY IN ESTONIAN 60
ACKNOWLEDGEMENTS 65
ORIGINAL PUBLICATIONS 67
CURRICULUM VITAE 98
ELULOOKIRJELDUS 100
LIST OF PUBLICATIONS 102

LIST OF ORIGINAL PUBLICATIONS

I **Suuster, E.**, Ritz, C., Roostalu, H., Reintam, E., Kõlli, R. & Astover, A. 2011. Soil bulk density pedotransfer functions of the humus horizon in arable soils. Geoderma, 163, 74–82.

II **Suuster, E.**, Ritz, C., Roostalu, H., Kõlli, R. & Astover, A. 2012. Modelling soil organic carbon concentration of mineral soils in arable land using legacy soil data. European Journal of Soil Science, 63, 351–359.

III Ritz, C., **Putku, E.** & Astover, A. 2015. A practical two-step approach for mixed model-based kriging, with an application to the prediction of soil organic carbon concentration. European Journal of Soil Science, 66, 548–554.

Contributions:

Paper	Idea and study	Data collection	Data analysis	Manuscript preparation
I	AA, **EP**, CR	**EP**, AA	**EP**, CR	**EP**, AA, CR, HR, ER, RK
II	**EP**, AA, CR,	**EP**	**EP**, CR	**EP**, CR, AA, HR, RK
III	CR, **EP**	**EP**, AA	CR, **EP**	CR, **EP**, AA

AA – Alar Astover; CR – Christian Ritz; **EP** – Elsa Putku; ER – Endla Reintam; HR – Hugo Roostalu; RK – Raimo Kõlli

ABBREVIATIONS

ANCOVA	analysis of covariance
ANN	artificial neural network
C	carbon
CART	classification and regression tree
Cf	physical clay
CO_2	carbon dioxide
D_b	bulk density (g cm^{-3})
EBLUP	empirical best linear unbiased predictor
ESC	Estonian Soil Classification
FAO-UNESCO	Food and Agriculture Organization of the United Nations-United Nations Educational, Scientific and Cultural Organization
ha	hectare
LMM	linear mixed model
MLR	multiple linear regression
Mt	megatonne
NDVI	normalized difference vegetation index
NDWI	normalized difference wetness index
PTF	pedotransfer function
R^2	coefficient of determination
REML	restricted maximum likelihood
RF	random forests
RMSE	root mean square error
SMN	soil monitoring network
SOC	soil organic carbon
Wc	water content

1. INTRODUCTION

Collecting soil information is known as the formation of legacy data about soils. They include former soil survey/monitoring reports and related datasets, soil maps, soil profile descriptions etc. (Odeh *et al.*, 2012). However, often such data were inaccessible to data analytic processing and are nowadays rarely used in decision making processes. Therefore the "renewal" of these data through digitization and modelling is essential (Rossiter, 2008). The same applies for legacy soil data available in Estonia from past large-scale soil mapping, soil monitoring network (SMN) and agrochemical survey of the arable land.

Soil organic matter in general and carbon in it (SOC) are one of the most common indicators of soil quality and it has multiple interactions with ecosystem services. One of the soil-mediated ecosystem services is carbon storage and greenhouse gas regulation (Keesstra *et al.*, 2016). Soils represent the largest terrestrial carbon pool (Batjes, 1996) and recently their interactions with the global climate system have attracted much interest. Apart from carbon sequestration or emission potential of soils, the SOC status is essential for soil functioning and fertility which is the basis for sustainable biomass production. Thus, knowledge of soil SOC status is needed from local to global scale. However, in most regions there are not enough analytical data for SOC status, especially in finer spatial scales. Consequently, prediction of SOC has been carried out using different pedometric techniques (pedotransfer functions, digital soil mapping).

SOC is expressed by the proportion of organic carbon (SOC, %) or by area basis (SOC stock). Soil bulk density (D_b) is a parameter used in weight-to-volume conversion of SOC (%) to SOC stock. SOC (%) is a portion of the soil measured as carbon in organic form, excluding living macro- and mesofauna and living plant tissues, usually expressed as proportion of carbon in soil weight (i.e., g C per 100 g of soil, g C per kg of soil). It is one of the frequently used indicators in SMN reflecting soil-type-specific soil fertility and functioning of the soil, important physical and chemical properties of the soil. When the SMN provides a nationally consistent assessment of soil carbon conditions across major soil types and land uses (van Wesemael *et al.*, 2010), it offers a solid basis to estimate the status or changes of SOC (%) and stock. Stock (density) is SOC expressed on area basis (e.g., per hectare (ha) or g per m^2). The

calculation of SOC stock is based on SOC (%), soil D_b, the relative contribution of fine earth material to total soil mass and layer thickness or sampling depth. The unit of SOC stock is usually t C ha^{-1} (Mg C ha^{-1}) or g C m^{-2}.

D_b is an indicator of soil physical status. It is defined as the oven-dry weight of soil divided by its volume expressed by g cm^{-3} or Mg m^{-3}. D_b determines physical properties of the soil by dictating the water and solute movement through the soil. D_b is strongly negatively correlated with soil organic matter being lower at higher organic matter content (Arvidsson, 1998; Franzluebbers, 2002; Ruehlmann & Körschens, 2009). Consequently D_b is inversely related to soil depth as much of the organic matter is in the upper part of soil profile (Balland *et al.*, 2008). Another factor determining the D_b in soil profile is soil texture/particle-size distribution as soils with large amount of fine particles (silt and clay) have lower D_b compared to soils with more coarse particles (Benites *et al.*, 2007). Additionally, soil D_b is influenced by land use and mechanical stress introduced by different tillage practices (Nawaz *et al.*, 2012). Increased D_b is indicator of soil compaction which is caused mainly by anthropogenic operations (mechanical farm/forestry and military operations) but also by natural processes (rain, plant roots, foot traffic, soil fauna) rearranging the soils solid particles (Nawaz *et al.*, 2012). Increased D_b affects water infiltration, available water capacity, soil porosity, plant nutrient availability, rooting depth and soil microorganism activity, which all influence key soil processes and productivity (Lipiec & Hatano, 2003; Håkansson, 2005; Nawaz *et al.*, 2012).

Morvan *et al.* (2008) defined SMN as 'a set of sites/areas where changes in soil characteristics are documented through periodic assessment of an extended set of soil parameters'. Although SOC (%) is often measured in soil monitoring, the same cannot be said for D_b, which relies on time-consuming and expensive sampling of cores, especially in soil monitoring with many sites. Therefore, prediction of missing soil properties may to some extent compensate for incompletely observed data. One of the distinctive features of soil monitoring is that the samples are collected through stratified (random) or systematic sampling (by grids or transects) (Arrouays *et al.*, 2012). The sampling design defines the methods for analysing the data and detecting significant differences in the soil properties. The search for the model with the best prediction accuracy has been an ongoing research topic for years (e.g., Brus & de

Gruijter, 1997; Schrumpf *et al*., 2011; Martin *et al*., 2014). Already many prediction models have been developed however extending the models beyond the geomorphic region or soil types where they were designed in is not advisable (McBratney *et al*., 2002). In practical terms it's often even impossible because local legacy soil data and maps are mainly based on country-specific soil classification system. It means that there is need to develop models where predictor variables are compatible with existing large-scale soil map.

The present thesis addresses the question of developing statistical methods to predict SOC (%) and D_b from other available soil properties based on digitized legacy soil data available via SMN data and legacy soil map.

2. REVIEW OF THE LITERATURE

2.1. Overview of prediction approaches in soil science

In soil science, prediction procedures can be divided in two groups: digital soil mapping and pedotransfer functions/rules. Pedotransfer functions (PTFs) are functions that predict certain soil properties from other easily, routinely or inexpensively measured soil physical and chemical properties (McBratney *et al.*, 2002). Digital soil mapping serves the same goal with PTFs, however, it also includes other environmental variables and applies the model to geographic database to create a predictive map of specific soil properties (Scull *et al.*, 2003). Gradually the use of PTFs is declining and digital soil mapping techniques become more popular due to enhanced possibilities using GIS and spatial data (Minasny *et al.*, 2013), although in many cases scientists demonstrated the benefits of both approaches (e.g., Lettens *et al.*, 2005; Bonfatti *et al.*, 2016). Both approaches produce a static model, however for simulating the fluxes and turnover of soil carbon, dynamic simulation models have been developed, e.g., ICBM, RothC, Century, CANDY, C-TOOL (Andrén *et al.*, 2004; Rukhovich *et al.*, 2007; Dendoncker *et al.*, 2008; Lugato *et al.*, 2014; Taghizadeh-Toosi *et al.*, 2014). In current thesis only the static models are used.

PTFs and digital soil mapping rely on statistical and geostatistical modelling approaches for prediction of soil properties (further on referred as prediction models). Essentially, prediction models can be divided into two groups (Breiman, 2001):

1) Grey-box models: methods assuming a priori probability distributions that (are believed to) characterize the data structure (e.g., linear regression, logistic regression) and
2) Black-box models: distribution-free algorithmic methods that in various ways search the data for structures (e.g., classification and regression trees).

These two approaches may also be referred to as statistical modelling and statistical learning, respectively (Hastie *et al.*, 2009).

2.1.1. Prediction of SOC

Statistical relationships between SOC (%) and its potential predictor variables have been studied (e.g., Grimm *et al.*, 2008; Gray *et al.*, 2009; Oueslati *et al.*, 2013; Bonfatti *et al.*, 2016) but the main focus has been on predicting SOC stock compared to SOC (%). Thus, prediction methods and predictor variables for both SOC (%) and stock are introduced in the following subchapters.

2.1.1.1. Grey-box models

The most common way to estimate SOC stock used to be calculating a mean or median SOC stock by stratifying the data after soil type and/or land use (Arrouays *et al.*, 2001; Yu *et al.*, 2007; Meersmans *et al.*, 2008). However, the most popular method is multiple linear regression (MLR) (Meersmans *et al.*, 2011; Oueslati *et al.*, 2013; Gray *et al.*, 2015). The prediction accuracy of the MLR models has been generally higher than for the mean/median approach (Meersmans *et al.*, 2008; Gray *et al.*, 2009). Also the stratification of SOC (%) by soil types, land use, vegetation type, topographical parameters has shown better results in terms of explained variation in the data (Bell & Worall, 2009; Oueslati *et al.*, 2013), especially in topsoil (Meersmans *et al.*, 2009a).

Spatial variability

Although mean and regression approaches have the advantage of easy usage and interpretation they do not capture the spatial variability of soil properties, e.g., SOC varies already in small scales both horizontally and vertically (VandenBygaart, 2007; Saby *et al.*, 2008). Geostatistical methods are useful to capture the spatial variability of soil (Lark, 2012; Brevik *et al.*, 2016). Typically kriging techniques like ordinary kriging and regression kriging have been used in SOC stock prediction (Mishra *et al.*, 2009; Phachomphon *et al.*, 2010; Liu *et al.*, 2015; Bonfatti *et al.*, 2016). Generally, predictions using kriging techniques have achieved better prediction accuracy compared to MLR (Kumar *et al.*, 2012; Liu *et al.*, 2015).

A more complex grey-box model is linear mixed model (LMM) which additionally to measurable soil properties capture the systematic and spatial variation by including a set of random effects into the model (Lark

& Cullis, 2004). The random variation in the data can be introduced by the sampling design. Several studies have exploited the benefits of LMM in soil science (e.g., Haskard *et al.*, 2010; Maia *et al.*, 2010). Doetterl *et al.* (2013) fitted a LMM at a regional scale to predict SOC stock in different soil layers for cropland and identified large spatial variability of the stocks. As they used data from two different databases, the origin of the data was considered as a random effect and the spatial correlation was modelled using a Gaussian exponential correlation structure. The systematic variation of soil sampling was the reason for Maia *et al.* (2010) to use mixed modelling (site variable as random effect) for analyzing the effects of agricultural management systems on soil C stocks. Stevens *et al.* (2015) characterized the controlling factors of SOC (%) of croplands in field scale by LMM approach using Matérn model to describe spatial covariance structure.

Predictor variables in grey-box models

Pedologic parameters have been most frequently used in grey-box models, e.g., soil type, texture/particle size distribution, parent material. Grüneberg *et al.* (2010) demonstrated that the high variability of SOC stock in soil horizon can be attributed to the thickness of the horizon and suggested that the accuracy of SOC stock estimates can be improved by taking the samples by horizons instead of depth increments. Nevertheless, the thickness of A-horizon has rarely been used in models. Climatic variables (annual mean precipitation, annual mean temperature) have been identified as major predictors of carbon storage in croplands (Gray *et al.*, 2009; Meersmans *et al.*, 2011; Doetterl *et al.*, 2013) having a strong positive relationship with precipitation and negative relationship with temperature (Meersmans *et al.*, 2011; Gray *et al.*, 2015). Also land use has an important impact on the amount and distribution of SOC (Meersmans *et al.*, 2008; 2009a). Topographical variables were identified as the strongest variables with hydrological variables in the croplands of the Great Duchy of Luxembourg (Doetterl *et al.*, 2013). Additionally, Bell & Worall (2009) and Xin *et al.* (2016) emphasized that including altitude in models improve the prediction accuracy, Gray *et al.* (2015) demonstrated that aspect plays a significant role in deeper horizons.

2.1.1.2. Black-box models

Different data mining and machine learning methods have become increasingly popular in SOC studies. Different decision tree approaches: boosted regression trees (Martin *et al.*, 2014, Akpa *et al.*, 2016), classification and regression trees (CART) (Bou Kheir *et al.*, 2010), decision tree (Gray *et al.*, 2009; Stoorvogel *et al.*, 2009), multiple additive regression tree (Krishnan *et al.*, 2007; Lo Seen *et al.*, 2010); random forests (RF) (Grimm *et al.*, 2008; Wiesmeier *et al.*, 2014; Were *et al.*, 2015; Akpa *et al.*, 2016; Bonfatti *et al.*, 2016) and RF-CART (Wiesmeier *et al.*, 2011) have been used more frequently in black-box models. In comparative studies like Gray *et al.* (2009), classification tree method yielded in the second place for predictive accuracy compared to MLR and median approach. Were *et al.* (2015) compared RF, support vector machines for regression, and artificial neural networks (ANN) models to predict SOC stocks from 19 environmental variables and laboratory soil data. The best predictive performance was achieved with both support vector machines and ANN with total nitrogen as the best predictor in all three models. Akpa *et al.* (2016) achieved the best prediction accuracy with RF compared to Cubist and boosted regression trees.

Predictor variables in black-box models

In black-box models, the focus of predictors usually stems from the state factor approach where soil is viewed as a function of climate, organisms, topography, parent material and time (Jenny, 1941) and its derivate in spatial dimension with additional factors soil and time, the scorpan model (McBratney *et al.*, 2003). Topographic variables (e.g., slope, elevation, curvature) have been used the most in black-box models (Figure 1). Amongst climatic variables, they are mainly associated with temperature or precipitation (e.g., Krishnan *et al.*, 2007; Wiesmeier *et al.*, 2014) but also evapotranspiration has been used (Martin *et al.*, 2014). Pedologic parameters like parent material, soil type and textural classes/particle size distribution have been frequently used as input variables. Vegetation indices NDVI, NDWI were also important variables in modelling (Bou Kheir *et al.*, 2010; Were *et al.*, 2015; Akpa *et al.*, 2016). Large influence of land use/land cover has been shown in predicting SOC (e.g., Adhikari *et al.*, 2014).

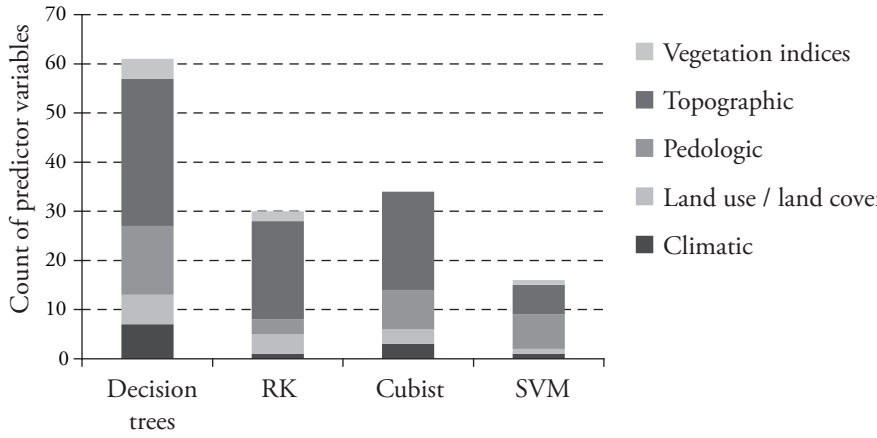

Figure 1. Predictor variables used in black-box SOC models (decision trees; RK – regression kriging; Cubist; SVM – support vector machine) grouped by predictor classes based on 15 publications (Krishnan *et al.*, 2007; Grimm *et al.*, 2008; Gray *et al.*, 2009; Stoorvogel *et al.*, 2009; Bou Kheir *et al.*, 2010; Lo Seen *et al.*, 2010; Wiesmeier *et al.*, 2011, 2014; Adhikari *et al.*, 2014; Martin *et al.*, 2014; Gray *et al.*, 2015; Liu *et al.*, 2015; Were *et al.*, 2015; Akpa *et al.*, 2016; Bonfatti *et al.*, 2016).

Due to the black-box nature of these models, the relationships between the SOC and predictor variables cannot easily be interpreted; only the importance of the variable can be assessed. Wiesmeier *et al.* (2011) proposed to use RF combined with CART to supplement the black-box procedure with some kind of useful interpretation. This approach relies on the assumption that the splits in the regression trees randomly generated when constructing the random forest are not in general deviating too much from the splits in the single regression tree fitted to the entire training data.

2.1.2. Soil bulk density and pedotransfer functions

The most frequently used methods in predicting D_b are similarly to SOC (%) grey-box models (Table 1). MLR has mainly been used to predict D_b from basic soil properties in different scales (e.g., Kätterer *et al.*, 2006; Hollis *et al.*, 2012). MLR-based predictions are more used in SOC stock studies than prediction models by other methods, especially more general PTFs like $D_b = 1.66 – 0.318*SOC(\%)^{1/2}$ (Manrique & Jones, 1991) requiring only one input variable (e.g., Meersmans *et al.*, 2008; Stevens *et al.*, 2014). Next to McBratney *et al.* (2002), other comparative studies

have also concluded that existing PTFs should be used with care in new conditions and more accurate results come from recalibrating existing PTFs or developing new ones (Kaur *et al.*, 2002; De Vos *et al.*, 2005; Nanko *et al.*, 2014). Therefore, the PTFs already developed in tropical/temperate regions or mineral/forest soils have been used in the same conditions for e.g., SOC stock assessment (PTF by Benites *et al.*, 2007 used in Bonfatti *et al.*, 2015).

On the contrary to easy interpretation and understanding of the underlying mechanism of the regression method, the black-box models provide an alternative way for developing PTFs. However, ANN-predicted D_b values have not been significantly better compared to MLR-predicted values (Tranter *et al.*, 2007; Wang *et al.*, 2013). Also, RF predicted D_b more effectively compared to ANN (although in subsoil ANN performed better), Cubist and boosted regression trees (Taalab *et al.*, 2013; Akpa *et al.*, 2016). Sequeira *et al.* (2014) introduced a novel approach to predict D_b incorporating information from surrounding horizons, mainly D_b, in the soil profile (without using SOC concentration) using RF method (root mean square error – RMSE 0.10 to 0.15 g cm^{-3}). Similar results where SOC was not included were obtained by Taalab *et al.* (2013; 2015) who applied RF, ANN and Bayesian network to predict D_b in a landscape scale using only landscape-derived attributes resulting in R^2 of 0.37–0.55 and RMSE of 0.17–0.19 g cm^{-3}, respectively. Other black-box models like multiple additive regression trees or k-ANN methods have shown more precise prediction of D_b compared to MLR models (Martin *et al.*, 2009; Botula *et al.*, 2015).

Studies have shown that SOC (%) is the most significant input variable to prediction models next to soil textural classes/particle size distribution (e.g., Manrique & Jones, 1991; Bernoux *et al.*, 1998; Kätterer *et al.*, 2006; Wang *et al.*, 2013; Botula *et al.*, 2015). Water content at sampling time is inversely related to D_b and has shown to improve prediction accuracy (Manrique & Jones, 1991; Heuscher *et al.*, 2005) but nevertheless this soil property is rarely used in D_b prediction. However, Yang *et al.* (2015) showed that field-measured D_b and water content were highly correlated and improved the prediction of D_b by co-kriging with kriged soil water content. Sampling depth has been used more frequently than water content (Table 1). Although, the models performance improved with depth as a variable, the predictive quality did not increase significantly (De Vos *et al.*, 2005; Heuscher *et al.*, 2005).

Similarly to SOC (%) the range of variables for D_b prediction is large, and some of them, such as SOC (%) and texture are more frequently used than others, e.g., water content. Also, the LMM has been rarely applied for D_b prediction. Only Lark *et al.* (2014) developed a PTF to predict D_b of mineral soils from organic carbon content with the aim of determining sampling requirements for D_b and SOC in soil monitoring.

Table 1. Summary of some studies in soil bulk density prediction.

Reference	Land use	Location	Method	Predictors
Manrique and Jones (1991)	Various	USA, Hawaii, Puerto Rico	Grey-box model	SOC, clay, silt, sand, water content
Bernoux *et al.* (1998)	Various	Brazil	Grey-box model	SOC, clay, sand, pH*
Calhoun *et al.* (2001)	Various	Ohio, USA	Grey-box model	SOC, depth, clay, sand*
Kaur *et al.* (2002)	Various	India	Grey-box model	SOC, clay, silt
De Vos *et al.* (2005)	Forest	Belgium	Grey-box model + recalibration with depth and texture variables	LOI, depth, clay, sand
Heuscher *et al.* (2005)	Various	Mainly USA	Grey-box model	SOC, depth, clay, silt, water content
Kätterer *et al.* (2006)	Arable land	Sweden	Grey-box model	particle size classes, textural classes, SOM and SOM classes
Benites *et al.* (2007)	Various	Brazil	Grey-box model	SOC, clay, sum of basic cations*
Krishnan *et al.* (2007)	Various	India	Black-box model	SOC, soil type, clay, silt, sand, horizon type, land-cover type
Tranter *et al.* (2007)	Mainly arable lands	Australia	Black-box and grey-box model	SOC, depth, clay, sand, silt,

Martin *et al.* (2009)	Mainly arable lands	France	Black-box model	SOC, depth, clay, silt, land cover, layer type
Brahim *et al.* (2012)	Various	Tunisia	Grey-box model	SOC, clay, f-sand, f-silt, pH
Hollis *et al.* (2012)	Various	Europe	Grey-box model	SOC, sand, clay, horizon mid-depth
Wang *et al.* (2013)		Loess Plateau, China	Black-box and grey-box model	Sand, silt, clay, SOC, altitude, vegetation coverage, growth age, slope gradient and aspect
Taalab *et al.* (2013; 2015)	Various	English Midlans	Black-box model	Soil group, land use, parent material, bedrock, topography, climate
Adhikari *et al.* (2014)	Various	Denmark	Grey-box model	18 environmental variables
Lark *et al.* (2014)	Arable	England	Grey-box model	SOC
Nanko *et al.* (2014)	Forest	Japan	Grey-box model	SOC, SOM, LOI
Sequeira *et al.* (2014)	Various	USA	Black-box model	designation, depth, textural class, thickness
Botula *et al.* (2015)	Various	Lower Congo	Black-box and grey-box model	Sand, silt, clay, SOC, pH, CEC, DCB-Fe, DCB-Al
Yang *et al.* (2015)	Various	(South-west) China	Grey-box model	Water content
Akpa *et al.* (2016)	Various	Nigeria	Black-box model	Topography, vegetation/ anthropogenic factors, climate, parent material

* predictors in all grey-box models, additional predictors in smaller data groups.
SOC – soil organic carbon; SOM – organic matter; LOI – loss-on-ignition;
CEC – cation exchange capacity; DCB-Fe – dithionite–citrate–bicarbonate-extractable Fe; DCB-Al – dithionite–citrate–bicarbonate-extractable Al.

2.2. Calculating soil organic carbon stock

Although, the simplest way to obtain SOC stock estimates is by direct calculation (Eq 1),

$$SOC\,(t\ C\ ha^{-1}) = (A(m) * area - coarse\,fraction) * D_b\,(t\ m^{-3}) * SOC\,(\%)\ (\mathbf{Eq\ 1})$$

it is more difficult to do so in large scale due to exhaustive input data (measurements of SOC, D_b, depth, classes of stoniness) requirements. Those requirements can be met in national soil inventory or legacy soil databases. In that case SOC stocks are calculated using different soil modelling and prediction methods. The prediction methods are used to fill the gap between existing and missing information in calculating the SOC stock by predicting SOC and D_b separately. For example, Meersmans *et al*. (2011) calculated the SOC stocks of cropland in Belgium by modelling SOC (%) values with MLR and used a PTF by Manrique and Jones (1991) to predict D_b. Also Brahim *et al*. (2011) combined regression models and a PTF to predict the missing values of SOC and D_b. In the Belgian database, missing D_b values were predicted with a calibrated PTF to geomatch SOC stocks with landscape units (Lettens *et al*., 2005). Milne *et al*. (2007) developed new PTFs for both SOC and D_b. An actual comparison of the associated variability and prediction accuracy of such combined methods was done by Taalab *et al*. (2013) where D_b was predicted using RF. The results showed that the SOC stock calculated with the RF-predicted D_b underestimated the D_b values compared to the SOC stock calculated with measured D_b.

2.3. SOC stocks globally and locally

Global and local estimation of past, current and future SOC stocks are important because of their importance to climate change and food security. Globally, soil carbon plays an important role in world's carbon cycle. This is relevant in terms of climate change where the CO_2 emissions have increased since 1750 and largely due to anthropogenic fossil CO_2 emissions (IPCC, 2014). The net emissions from agriculture, forestry and other land use activities increased 8% over the period of 1990–2010 whereas the main source of increase were emissions of agriculture by 8% (4.6 to 5.0 Gt CO_2 eq yr^{-1} in 1990s and 2000s; 5.3 Gt CO_2 eq yr^{-1} in 2011)(FAO, 2014).

Globally, Batjes (1996) estimated the SOC stock 1462–1548 Gt of C in the upper 0–100 cm calculating stock for each soil layer in the profile and per FAO-UNESCO soil subunit using the World Inventory of Soil Emission Potentials (WISE) database. Lower worldwide SOC stock estimates (1417 Gt of C for top 1 m) were reported by Hiederer & Köchy (2011) where the amended Harmonized World Soil Database (HWSD) was used. This is the most up to date spatial SOC stock map available.

In Europe, SOC stocks have been modelled by pedotransfer rule method (Jones *et al.*, 2005) and most recently by regression model approach (de Brogniez *et al.*, 2015). In pan-European scale, the total carbon stock has been estimated to be 80.3 Gt and for Estonia 1.5 Gt (Jones *et al.*, 2005). Lugato *et al.* (2014) used the Century model to simulate the SOC stock in pan-European level in agricultural soils and predicted that arable land stored 7.65 Gt of carbon, corresponding to 43% of total agricultural SOC stock in the region (17.63 Gt). Kõlli *et al.* (2010) estimated the SOC stock in mineral soils of Estonia to 0.3 Gt for the entire soil cover by direct calculations based on soil data from large-scale soil mapping (1 : 10 000).

2.4. Related studies in Estonia

SOC (%) and stocks in Estonian arable soils have been well studied. However, little work has been done on spatial patterns in the distribution of SOC. Several generalizations of the status and stocks of organic matter in Estonian soils have been made (Kõlli *et al.*, 2011). The studies have mainly focused on humus quality, composition of soil organic matter, SOC distribution in soil types and sources and fluxes of SOC (Kask, 1975; Kanal, 1996; Kõlli & Ellermäe, 2003; Reintam, 2007; Kõlli *et al.*, 2010 etc.). Kõlli *et al.* (2009) calculated SOC stocks for soil types on the topsoil, and the whole solum concluding, that the stocks depend mainly on soil carbonate and clay content, moisture regime, pedon fabric and land use type. 86.4 Mt of SOC is sequestered in Estonian arable soils with an average of 68.9 t ha^{-1} for humus cover and 20.1 t ha^{-1} in the subsoil (Kõlli & Ellermäe, 2003).

The National Soil monitoring data provides an extensive database of soil properties in mineral soils. However, these data has been mainly described only by descriptive statistics (Kokk & Rooma 1983; Lehtveer

& Kokk, 1995). Studies in soil bulk density are also related to presenting the descriptive statistics (Eesti NSV mullastik arvudes, 1989). However, Kask (1997) showed a negative linear relationship between soil D_b and SOC (%) where it is evident that the mineral soils with higher SOC (%), the D_b values are lower (R^2=0.86). Also, Roostalu (in Lauringson *et al.*, 2015) developed PTFs for D_b of mineral arable soils using humus concentration and stratifying the models by soil texture: soils with heavier texture and more humus have smaller D_b. Although modelling D_b has not been an a priori goal of soil scientists in Estonia, the studies including D_b were related to studying the compaction introduced by machinery and calculating the critical values of D_b regarding different crops (Nugis & Lehtveer, 1991; Lehtveer & Kokk, 1995; Reintam *et al.*, 2008; Krebstein *et al.*, 2014).

The Estonian legacy data of D_b and SOC (%) have been rarely used in actual decision making and data analysis. As the amount of data collecting was large at the time of soil monitoring, there was not much time left for deeper analysis on data, only the descriptive statistics after monitoring year were provided. Continuous soil monitoring was stopped because of the change in political regime in Estonia in 1991 as the financial support on soil monitoring significantly decreased. Considering the hot topics in soil science (e.g., compaction, soils as C reservoir) it is essential to analyze the spatial pattern of SOC and D_b in regional level to contribute in the wider scientific community.

3. HYPOTHESIS AND AIMS OF THE STUDY

The main hypothesis of this thesis was whether or not the use of linear mixed model could exploit the hierarchical structure in the National Soil Monitoring network data to improve prediction accuracy of soil properties. It was assumed that the Estonian legacy soil data (National Soil Monitoring and Estonian large-scale soil map) were suitable for developing prediction models. Also it was assumed that the developed models could be used for spatial extrapolation of new parameters (e.g., SOC stock) across mineral arable soils. Additionally, this would lead to new possibilities for using the soil legacy data by creating other soil-specific models and parameters.

The specific aims of the study were:

1. Digitize the National Soil Monitoring data from 1983–1994 stored on paper to computer and setting up the database structure.
2. Develop a prediction model for D_b and SOC (%) in Estonian arable soils with the compatibility of Estonian large-scale soil map (paper **I**, **II**).
3. Compare the prediction quality between the grey-box model (linear regression, median approach, linear mixed model) and black-box model (random forests) (paper **I, II**).
4. Extend the prediction models in paper **I, II** to incorporate spatial variability using a kriging approach (paper **III**).
5. Combine and compare predictions of separate SOC (%) and D_b models to calculate SOC stock and implement the results into large-scale Estonian Soil Map.

4. MATERIALS AND METHODS

4.1. Description of the dataset

Between 1983 and 1994, a consistent, complex and systematic monitoring of arable soils took place in Estonia under the framework of national SMN (Kokk & Lehtveer, 1995). The soil monitoring program was resumed in 2001 as part of the national environmental monitoring program and 26 monitoring sites were selected to be monitored (including sites previously in SMN) (KKM, 2016). The models in this thesis have been developed based on 79 monitoring sites studied in 1983–1994 under the national SMN program (these data were digitized), 15 monitoring sites (5 overlapping with national SMN) from 2008 under the newly commenced monitoring program (0.7% of data) (KKM, 2008) and one area from 1977 (0.1% of data) (to complement the dataset with finer textures) (Figure 1 in paper **I**). Monitoring sites have mainly been under cereal-based crop rotation with ploughing in autumn. As land-use information was very inconsistent in the database, crop or rotation history was not included in this study. The samples and data from the period 1983–1994 were archived at the Estonian Soil Museum at the Estonian University of Life Sciences and the samples from 2002 to 2008 at the Agricultural Research Centre. Now all the samples associated with the NSM are archived at the Agricultural Research Centre.

4.1.1. Monitoring scheme

Each monitoring site was generally with the size of 60×180 m and smaller sites 30×90 m. In each of the sites four parallel transects were established at the distance of 20 m. Each transect comprised of ten 1-m^2 monitoring plots; there were 40 monitoring plots for one monitoring site with the total of 3160 plots. The monitoring interval was three years. In the monitoring plots, the thickness of the humus horizon and average soil samples were taken. Some soil properties, i.e., D_b and water content were measured in four sampling depths: 3, 15, 25, and 40 cm (other depths e.g., 7.5 cm, were rare). In the fifteen monitoring sites studied in 2008, soil samples were taken from five random locations at 5 and 20 cm depth. Therefore, the database of D_b was larger than the database used for SOC (%) modelling (17 293 entries vs. 8 697 entries).

4.1.2. Measured soil properties

The data presented in this thesis was only for the humus horizon (A-horizon) representing all major Estonian soil types from arable land. Based on the Estonian Soil Classification (ESC), the A-horizon is a dark-colored topsoil layer that is formed in automorphic conditions by the accumulation of well-humified organic matter that is firmly bound to mineral soil particles (Kokk *et al.*, 1968). The established ESC limits of SOC (%) for A-horizon are 0.6–6%. The dataset contains sampling depth, oven-dry bulk density (D_b), water content (Wc) at sampling time, soil type, thickness of the A-horizon, SOC (%), texture and physical clay (Cf) content (Table 1 in Paper **I**).

Sampling of D_b was carried out in four replications by the core method using 50-cm^3-volume cylinders in 1983–1994; however, in 2008, a steel cylinder of 88.2-cm^3-volume (d = 5.3 cm, h = 4 cm) was used. The arithmetic mean from four replications in monitoring plot was used. Measured D_b values correspond to the equilibrium D_b because sampling was carried out at the end of the growing season. D_b values of below 0.8 g cm^{-3} or above 2.0 g cm^{-3} were excluded from this study, as they represent unreliable data for the A-horizon in mineral soils.

The thickness of the A-horizon ranged from 12 to 100 cm. The gravimetric water content was determined at 105 °C (% weight). The sampled soils were classified according to the ESC; where approximate soil groups are given, classifications were also based on the World Reference Base 2006 (WRB, 2006).

SOC (%) was determined by Tjurin method, based on wet digestion with acid potassium dichromate of carbon (Tjurin IV, 1937; Vorobeva, 1998). No soil-type-specific correction factor was applied because there was large soil type variation and, moreover, correction for incomplete oxidation is not recommended (Meersmans *et al.*, 2009b).

Particle-size distribution of fine earth was determined using the pipette method by Katschinski, in which physical clay (Cf) is defined as the mass fraction of particles smaller than 0.01 mm (Katschinski, 1956). According to the Katschinski classification, the soils were classified based on their physical clay (<0.01 mm) content with the approximate classes names after FAO (2006) system as follows: those with less than 10.0%

physical clay were classified as sand; from 10.1 to 20% as loamy sand; from 20.1 to 30.0% as sandy loam; from 30.1 to 40.0% as loam; from 40.1 to 50.0% as sandy clay loam; and soils with over 50.1% physical clay were classified as clay.

4.2. Framework of prediction

A prediction model provides individual predicted values for various configurations of predictor variables. The primary quality measure is the accuracy of predicted values. In contrast, the usual output of a statistical model (estimated standard errors, confidence intervals and p-values) are of no interest of prediction. Consequently, whether or not distributional assumptions are fulfilled is not important.

The first step is to define a number of candidate models based on well-defined criteria that reflect the intended applications of the models. In this case, such criteria must be compatible with the Estonian soil map; that is, soil-science-specific considerations replace statistical techniques of model selection. Evaluation and comparison of the candidate models is usually done by fitting them to *training* and *validation* datasets (in case of two or more models). The predictive performance of the best model is tested on a third independent *test* dataset (in case of only a single candidate model only training and validation datasets are needed). This way, the overly optimistic conclusions about the prediction accuracy and the underestimation of prediction error can be avoided by not using identical datasets in all steps (more details in paper **I**). The entire database was randomly split by following a common procedure: a 50% subset to a training dataset, a 25% subset to a validation and a 25% subset to a test dataset (Hastie *et al.*, 2009).

The comparison of the candidate models was based on comparing these predicted values with the corresponding observed values in the validation data by RMSE (the square root of the mean of the squared differences between predicted and observed values). RMSE is a measure that is easy to interpret – the scale is the same as for the soil property considered; the smaller the measure the better. One advantage of RMSE is that it is applicable to both grey-box models such as MLR and LMM and arbitrary black-box models and thus it enables comparisons between very different modelling approaches. Additionally, the agreement or disagreement

between the predicted and observed values was assessed visually. The best candidate model was then fitted to 75% of the national SMN dataset and assessed the prediction error on the new dataset, i.e., *test* dataset using same quality measures.

4.3. Prediction models

Four different prediction models were compared. Firstly, the median approach where the median value of a soil characteristic from the individual measurements in each stratum for each combination of the soil type and texture is calculated (paper **II**). Secondly, RF which is a black-box model being completely different from the median approach (more details in paper **II**). Soil type, Cf content, A-horizon thickness and year are used as predictor variables in RF approach for SOC (%) prediction model. RF were included in prediction models comparison as they have been used more in the last decade and also to assess their prediction accuracy compared to linear models. Two remaining models were statistical relationship-based approaches: linear regression and LMM.

4.3.1. Linear regression model

In a simple linear regression model with one predictor variable, there are two parameters: an intercept and a slope (Hastie *et al.*, 2009). In case of two or more quantitative variables the process is called multiple linear regression – MLR. The analysis of covariance (ANCOVA) approach is extending the MLR methodology to the situation where additionally categorical explanatory variables are available. These linear regression models are usually fitted using ordinary least-squares estimation. The ANCOVA approach possesses one major weakness in the context of prediction: the more explanatory variables that are included in the model, the greater is the risk of overfitting and the recommendation is to settle for a simple ANCOVA including a limited number of predictor variables in an additive fashion (very few interactions if any at all) that are selected based on experience or literature (Vittinghoff *et al.*, 2005) (more details in Paper **II**). The predictor variables were chosen in the model from the soil science perspective and their significance was tested comparing the model with and without the predictor variable using likelihood ratio test (Table 2).

Table 2. Untransformed predictor variables used in D_b and SOC (%) linear regression prediction models.

Predictor variable	Response	
	D_b (g cm^{-3})	SOC (%)
A-horizon thickness	Continuous	Continuous
Cf content	Continuous	Continuous
Sampling depth	Continuous	
SOC (%)	Continuous	
Wc	Continuous	
Soil type		Categorical

4.3.2. Linear mixed model

The sampling scheme in the national SMN database implies that there is a hierarchical structure in the data for each year: transects are nested within plots, which are again nested within sites. The hierarchical structure in the data introduces several layers of replication, in the sense that the data may exhibit unexplained variation not only in the individual point measurements, but also at transect, plot and site levels. The hierarchical structure is captured by the random effects reflecting the predictor variable levels of no primary interest. More details on LMM methodology can be found in paper **I.**

Considering the hierarchical structure of the database and the soil properties associated with D_b and SOC (%), the predictor variables shown in Table 3 were selected. The random effects related to the sampling scheme, that is site, transect, and plot, were the same in both models as they reflect the same hierarchical data structure. However, there were some additional random effects in both prediction models. More specifically, year-specific site effects were included in the D_b model anticipating that site effect might be different between years. Also the year-specific transect effects were also included in the SOC (%) LMM as were site-specific slope coefficients for the Cf content as the relationship between SOC (%) and Cf content was expected to vary from site to site. The fixed effects of the SOC (%) model were exactly the same as used in the linear regression approach (Model A). LMM-based prediction may utilize both the estimated fixed-effects parameters and the empirical best linear unbiased predictor (EBLUPs) as estimates for the unobserved random effects. Details on prediction using EBLUPs are given in paper **II** and **III**.

Table 3. Random and fixed effects included in hypothesis testing and prediction model. ("*" denotes combined terms where the effect of a quantitative variable is modified according to the level of a categorical variable).

Random effects	Fixed effects	
	$y = D_b$	
Site	Soil type	Soil type * Wc
Transect	Sampling depth	Cf
Plot	Sampling depth * SOC	Texture * SOC
Year	SOC	Soil type * SOC
Site by year	A-horizon thickness	Texture
	Sampling depth * Cf	Sampling depth * Wc
	Wc	Soil type * Cf
	Texture * moisture	
	y = SOC (%)	
Site	Soil type	
Transect	A-horizon thickness	
Plot	Cf content	
Year		
Site * Cf content		
Transect by year		

4.4. A dataset used for kriging

The data for mixed model-based kriging were sampled in a sub region of Tartu County as part of agrochemical survey in Estonia carried out in 2002–2004 (Agricultural Research Centre, 2016) (Figure 1 paper **III**). In total, SOC (%) data for 85 arable fields were available (Table 4). The survey scheme was to take one composite soil sample from a 3–5 ha field. The centroid coordinates were used to match these fields to soil map data. The soil type and texture data (for mean Cf content) were extracted from the soil map data as in the survey only SOC (%) was measured. The selected sub region for model validation contained three of the original 79 sites used in the second step.

Table 4. Descriptive statistics of the agrochemical survey dataset from Tartu County in 2002–2004 (Agricultural Research Centre, 2016).

Soil code by ESC	SOC, %	A (cm)	Physical clay content, %
Ko (n=4)[a]	1.62[b] ± 0.17	24.88 ± 1.6	27.5 ± 5
KI (n=15)	1.56 ± 0.25	24.5 ± 7.1	24.33 ± 5.94
LP (n=11)	1.67 ± 0.33	23.64 ± 8.03	17.37 ± 4.67
Lk (n=1)	1.62	28	7.5
Kog(n=8)	2.06 ± 0.48	25.31 ± 1.53	27.5 ± 4.63
KIg (n=13)	1.78 ± 0.43	25.65 ± 2.13	25 ± 8.16
LPg (n=11)	1.98 ± 0.83	25.23 ± 1.6	21.36 ± 8.09
G (n=22)	2.66 ± 0.97	25.52 ± 6.9	32.73 ± 6.12

[a] – count, [b] – mean ± standard error

4.5. The Estonian large-scale soil map and area of case study

The soil map of Estonia is based on a systematic large-scale mapping of soil cover that took place from 1954 to 1989 (Piho *et al.*, 1960, Kokk *et al.*, 1968, Metsamajandite…1978/1983, Põllumajanduslike…1989). The mapping was started in 1954 when by the leadership of Igna Rooma a specific soil group was formed under the Ministry of Agriculture (later known as RPI Põllumajandusprojekt). The soil map was digitized in 2001 at the scale of 1:10 000 and is the property of the Estonian Land Board (Estonian Land Board, 2001; 2016). The soil map covers the area of 43 300 km^2, which is the whole territory of Estonia except the land under towns and settlements. The Estonian soil map database contains information about the soil type (after ESC), soil texture, A-horizon thickness and classes of stoniness.

The first step was to distinguish arable areas from other areas on the digital soil map by using the layer of fields from Agricultural Registers and Information Board. However, the entire arable area is too large for carrying out the methods. Therefore, a specific case of Tartu County in south-east Estonia was considered as an example to present the results on a map. Tartu County covers 6.9% of Estonia's land surface and 35% of the counties area is in agricultural use (Astover *et al.*, 2010).

Calculating SOC stocks

The next step was to predict SOC (%) and D_b using three different approaches: median approach, linear regression and LMM (without kriging). Models were fitted to 75% of the data and tested on a 25% sub dataset. Then SOC stock was calculated based on Eq 1 while using predicted D_b and SOC (%) values in the national SMN database. Additionally, SOC stock was also calculated by Eq 1 using available soil information in the national SMN database to compare the predicted SOC stock. Then, the models were applied on legacy soil map data to predict SOC and D_b. Finally, the SOC stock was calculated based on these predictions for soil-type-based polygons. The results were visualized on the map using weighted averages for achieving a more general overview of the county: 1) weighted averages of SOC stock were calculated based on a grid (50*50 m^2); 2) a grid thematic map was produced by interpolating the weighted averages of SOC stock (inverse distance weighting interpolator). The GIS analysis was performed in MapInfo Pro (version 15).

4.6. Software

The environment for statistical computing R was used for all statistical analyses (R Core Team, 2015). In all of the papers, the extension package lme4 was used for the LMM analysis (Bates *et al.*, 2015). LMMs can be fitted using also other statistical software (SAS, SPSS, Stata, HLM) (West *et al.*, 2012). In paper **III** the extension packages nlme and spaMM were additionally used to extract the exponential and Matérn correlation parameters (Pinheiro *et al.*, 2015; Rousset & Ferdy, 2014). Spatial predictions based on LMM were obtained using custom-made R functions (paper **III**). In paper **II**, the prediction using random forests was done with extension package randomForest (Liaw and Wiener, 2002).

5. RESULTS

5.1. Soil characteristics

Mean observed D_b values at plot scale ranged between 1.31±0.16 and 1.65±0.10 g cm^{-3} depending on the soil type (Table 1 in paper **I**). Significantly higher soil D_b was in the sampling depth 21–30 cm and lower D_b in the upper layer of the soil (3–10 cm) compared to other depths (Figure 2 in paper **I**). The distribution of D_b among texture classes indicated that D_b was significantly higher in medium textures, i.e., loamy sand and sandy loam (Figure 3 in paper **I**). Plots with the sandy loam texture were the most represented in the dataset (n=5019; 49%) having the mean D_b of 1.47±0.15 g cm^{-3} (0.82…2.09). Similar, although significantly higher D_b values were observed in loamy sand with the average of 1.49±0.13 g cm^{-3} (0.82…2.01)

Mean SOC (%) ranged between 0.96±0.26% and 3.04±1.07% by soil type. More specifically, Gleysols and gleyic soils (G, Kg, Kog, KIg, LPg) had larger SOC (%) than automorphic soils (Kr, K, Ko, KI, LP, Lk) and soils formed on carbonate-rich parent material (Kr, K, Ko, KI) had greater SOC (%) compared to carbonate-free soils (LP, Lk) (Figure 1 in paper **II**).

5.2. Prediction models for D_b

Validation of the three candidate models suggested that the LMM had better prediction accuracy compared to two other models. However, for comparison purposes both the LMM (Eq 2 in paper **I**) and model A (Eq 1 in paper **I**) were further tested. The RMSE based on test dataset was the smallest for mixed model 0.09 g cm^{-3}. However, model A had slightly higher RMSE of 0.11 g cm^{-3} with the limited predictive potential, strongly underestimating D_b values above 1.6 g cm^{-3} and overestimating below 1.2 g cm^{-3}. The predictions based on combined mixed-effects prediction are more precise than the prediction without random effects (Figure 6 paper **I**). Among the random effects included in the LMM (Table 4 in paper **I**), year-specific site effects incorporate most variation (68% relative to the residual error). Most variation is associated with the residual error that is pertaining to the individual measurements within each survey plot.

5.3. Prediction models of SOC concentration

Four candidate models were fitted in the training data. Validation showed that the median approach had the poorest prediction accuracy as the observed and predicted values were not in good agreement (Figure 2a in paper **II**). The prediction accuracy for the ANCOVA was better than for the median approach with an RMSE of 0.48% (Table 1 in paper **II**). The random forests approach resulted in a RMSE of 0.42% (Figure 7c in paper **II**). The permutation importance showed soil type with the highest rank (84%) following Cf content 54%, year 41% and A-horizon thickness 28%. The visual assessment of the agreement between observed and predicted SOC (%) values was similar for these three approaches (Figure 7a–c in paper **II**). Accurate predictions for higher SOC (%) i.e., over 4% were not achievable by these methods as the predicted values were too large.

Soil type explained most of the variation (>80%) in SOC (%) distribution in all of the tested models compared to A-horizon thickness, Cf content and year. The results of visual assessment between observed and predicted SOC (%) indicate a two-group pattern for all relationships in Figure 2 in paper **II**. SOC (%) values over 3% were more scattered around the line whereas lower SOC (%) values were more densely grouped. The range of predicted values from the LMM were 0.61–4.88% (observed range was 0.27–5.98%), which was generally wider than the ranges of predicted values for the three other models (0.43–4.07%).

Based on the validation of four models, the LMM (Eq 1 in paper **III**) was fitted into 75% of the dataset resulting in a RMSE of 0.22%. Observed values agreed well with LMM predictions (Figure 3 in paper **II**), although the spread around the one–to–one line was slightly larger compared to the validation of the LMM (Figure 2d in paper **II**): some observations for the former deviated substantially from the line. The deviating observations were mainly from Gleysols with observed SOC (%) of more than 3.7%.

5.3.1. Using site spatial information for kriging

The sites in the national SMN database were provided with coordinate information and the LMM was fitted into the entire dataset for kriging purposes. The fixed effect part of the model was the same as above. The

mixed model-based kriging approach as proposed in paper **III** was not applicable for D_b prediction as there was no single random effect that accounted for the majority of the variation in D_b. The estimated variance components for the random effects are shown in Table 2 in paper **III**. Amongst the two tested correlation structures, the exponential correlation structure proved to be statistically significant with the estimated nugget and range equal to 0.23 and 10.5 km, respectively (paper **III**).

LMM-based kriging was used on a new dataset from Tartu County (section 4.1). The two-step approach for kriging resulted in a RMSE of 1.06%. There was a very modest gain in using the spatial information in the data as the RMSE for the predictions based only on fixed effects resulted in a RMSE of 1.07% reflecting that predictions on new sites are limited with this model.

5.4. Combining predictions of separate SOC (%) and D_b models to calculate SOC stock

In the median approach, the median values of SOC (%) and D_b were obtained according to unique soil type – texture combinations. The MLR model for soil D_b is model A (Eq 1 in paper **I**) and ANCOVA model for SOC (%) contained Cf content, soil type and A-horizon thickness as predictors. The LMMs were Eq 2 in paper **I** and Eq 1 in paper **III** for D_b and SOC (%), respectively. The smallest RMSE was achieved by the LMM for both soil properties, corresponding to an RMSE of 0.07 g cm^{-3} for D_b and 0.17% for SOC (%) (Figure 2).

The results are in accordance with the separate analysis of D_b and SOC prediction models: median approach and linear regression models produce relatively the same size of RMSE whereas the LMM predicts more accurately. The largest gain in prediction accuracy was achieved by the LMM for SOC (%). This can be explained by the LMM methodology incorporating more sources of variation which are left unexplained by the three selected predictor variables in ANCOVA model. In case of predicting D_b, the gain in prediction accuracy was not as large as was in SOC (%) because of larger inherent variation in these data.

Plotting the calculated SOC stock values against the predicted SOC stock values, the LMM produces a very good agreement between the values

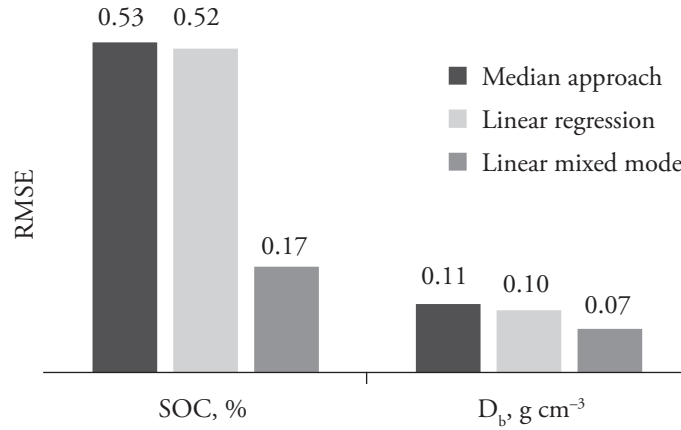

Figure 2. Comparison of root mean square error for predicting D_b and SOC (%) using three different prediction methods: median approach, linear regression and linear mixed model.

Table 5. Predicted mean and calculated root mean square error of SOC stock in three different prediction methods.

	Mean, t C ha^{-1}	RMSE, t C ha^{-1}
Median approach	65.7	19
Linear regression	64.3	22
Linear mixed model	70.8	7

indicating that the LMM provided the best prediction accuracy among selected methods (Figure 3; Table 5). The results for the median approach indicated that the predictions over 75 t C ha^{-1} are underestimated as the predictions have shifted below the one-to-one line. The linear regression results also showed a good fit. However, similarly to the median approach, SOC stock over 75 t C ha^{-1} was underestimated.

The spread around the one-to-one line was small for the LMM-based prediction (Figure 3) resulting in the best prediction accuracy with the smallest RMSE, 7 t C ha^{-1} (Table 5). For soils with potentially larger SOC stock the prediction based on the LMM seems to be promising.

The prediction accuracy and the mean predicted SOC stock per ha was the lowest for linear regression approach with the mean of 64.3 t C ha^{-1}

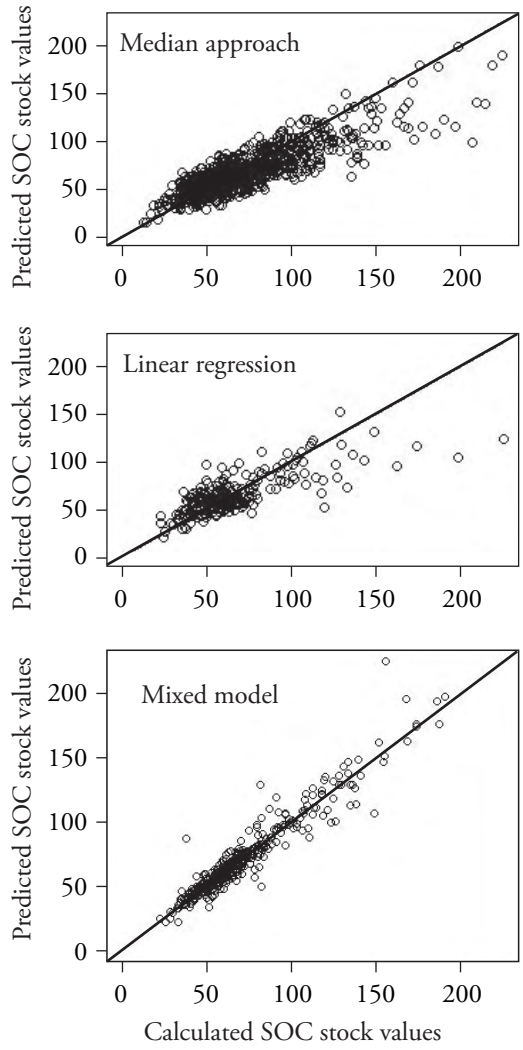

Figure 3. Calculated and predicted values of SOC stock (t C ha^{-1}).

and RMSE 22 t C ha^{-1}. Smaller RMSE was obtained for median approach compared to linear regression. Three times better prediction accuracy was achieved with LMM approach. It is worth noting that the kriging approach had not been incorporated and, hence, the prediction accuracy might even be improved more. Additionally, the predicted mean SOC stock was the highest by the latter method, 70.8 t C ha^{-1}, reflecting an underestimation of the mean by the median and linear regression model.

5.5. Model application: a case study of Tartu County

The LMM-based prediction model for SOC (%) (Eq 1 in paper **III**) fitted to 75% of the data was used for predicting the SOC (%) values in the mineral soils of Tartu County as the model yielded in the smallest RMSE – 0.22%. Predictions at new sites were based on neighboring sites (Eq. 5 in paper **II**) available in the national SMN dataset (Rooma, Kõpla, Vorbuse, Hünoneni, Metsakikka, Krõõla 1, Krõõla 2 and Tamme) (Figure 4).

Generally, the LMM-predicted SOC (%) values agree well with the known mean of SOC (%) by soil types as the soils formed on carbonate rich parent material have higher mean SOC (%) (Table 6). However, the predicted SOC (%) in gleyic soils is similar (KIg, LPg) or lower (Kog) compared to the same soil type in automorphic conditions. Gleysols have the highest SOC (%) as expected. Distinctively misleading values can be seen for eroded soils (slightly eroded E1 and moderately eroded E2 soils), as slightly eroded soils generally should have higher SOC (%) levels than moderately eroded soils.

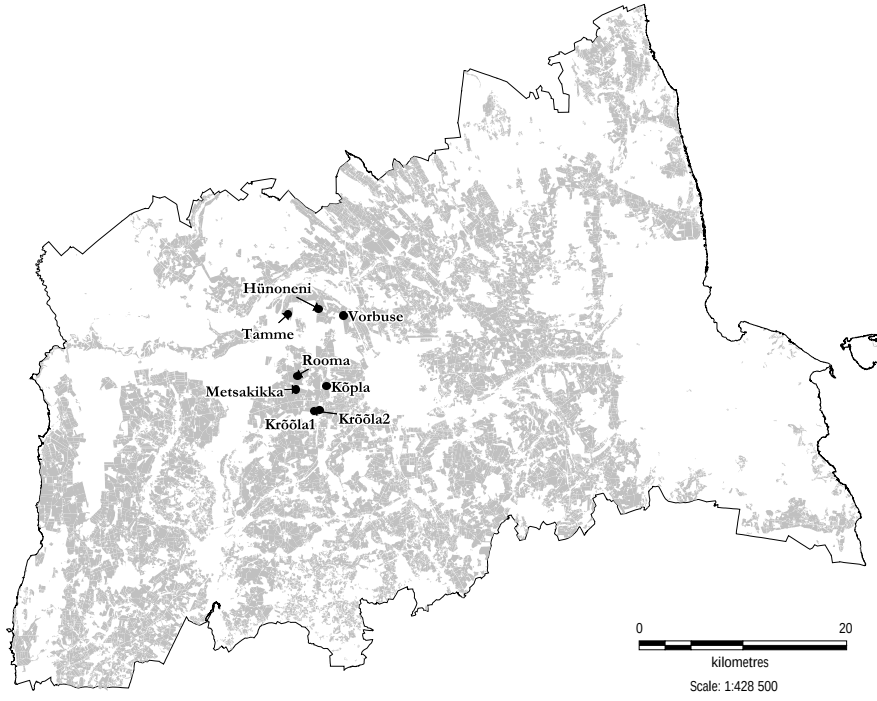

Figure 4. Arable fields (shown in grey color) and sites from the National Soil Monitoring network in Tartu County.

Table 6. Linear mixed model predicted topsoil SOC (%) values for mineral soils in Estonian Soil Map in Tartu County.

Soil type	RMSE	Mean	Median	Min	Max
Kr	0.28	1.47	1.56	0.91	1.77
K	0.16	1.27	1.30	1.05	1.50
Ko	0.19	1.24	1.25	1.12	1.41
KI	0.07	1.22	1.24	0.99	1.62
LP	0.03	1.16	1.15	0.99	1.46
Lk	0.16	1.09	1.11	0.86	1.18
Kog	0.59	1.22	1.21	1.03	1.39
KIg	0.25	1.41	1.4	0.96	1.60
LPg	0.15	1.19	1.15	0.95	1.96
G	0.92	3.24	3.35	2.41	4.88
E1	0.2	1.17	1.19	1.04	1.54
E2	0.55	1.25	1.24	0.52	2.44
D	0.48	1.44	1.24	0.79	3.14

The map with the predicted SOC (%) (weighted averages) on mineral soils show the local differences in SOC (%) (Figure 4 in paper **II**). The predicted SOC (%) ranged from 0.6 to 4.8%, indicating large soil cover heterogeneity. Most of the arable soils in Tartu County have 0.8–1.2% SOC in humus horizon however, soils with SOC (%) above 2% are very rare reflecting the absence of gleyic soils.

The calculated SOC stock based on LMM prediction for D_b and SOC (%) properties was also implemented in the Tartu County data (Figure 5). Average estimated SOC stock of Tartu County is 54.8 t C ha^{-1} and total topsoil SOC stock 1.8 Mt in humus horizon of arable mineral soils (Table 7). Most of the SOC is stored in gleysols, deluvial and pseudopodzolic soils in Tartu County. In addition, gleyic pseudopodzolic and gleyic eluvial soils are widespread in the County however they store 2.5–5 times less SOC. Gleysols have the highest average SOC stock 103.7 t C ha^{-1} and pebble rendzinas the smallest average SOC stock 35 t C ha^{-1}.

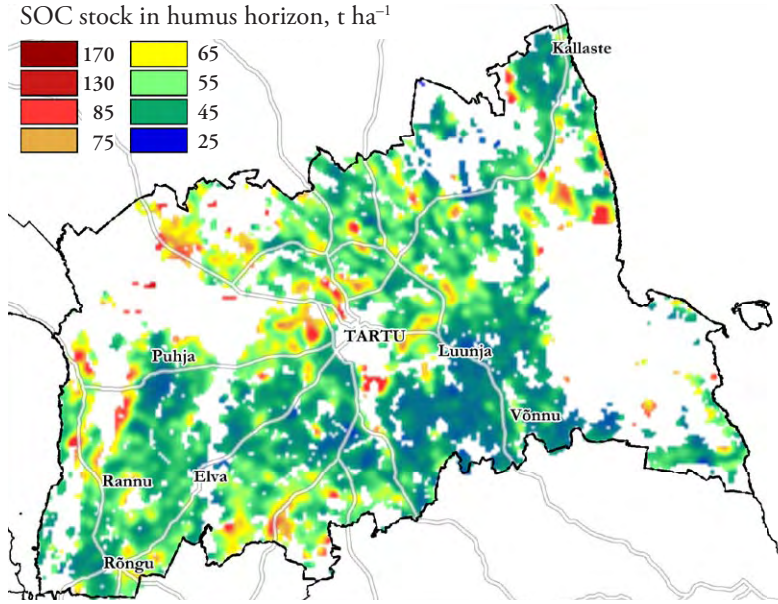

Figure 5. Estimated SOC stock (t ha^{-1}) based on linear mixed-model predictions in humus horizon of mineral soils (SOC concentration <6%) in Tartu County.

Table 7. SOC stock in humus horizon of mineral soils in Tartu County according to soil groups after Estonian Soil classification.

Soil group	Area (ha)	SOC stock (t ha^{-1})			Total SOC stock (kt)
		Min	Average	Max	
Kr	53	23.7	37.2	54.5	1
K	318	6	36	71	18
Ko	2 753	3.5	41.6	67.2	66
KI	8 124	3.5	42.1	68.3	82
LP	28 794	2.6	40.3	121.7	260
Lk	2 849	7.8	37.1	60.3	67
Kog	2 437	14.4	37.6	59.6	17
KIg	11 416	8.7	46.7	82.6	101
LPg	14 802	2.8	38.9	92.1	106
G	11 026	5.7	103.7	266	521
E1	2 825	7.1	36.3	131.8	131
E2	1 907	1.7	37	80.8	133
D	2 975	23	92.5	304	330
Total	90279				1834

6. DISCUSSION

6.1. Comparison of prediction models

The national SMN of Estonia is an example of a hierarchical data sampling scheme as each of the monitoring site was divided into four transects with ten sampling points on each transect. In the thesis, several types of prediction models were compared: median approach, linear regression models, LMM and random forests. Same methods have been used many times in similar studies in the soil literature (e.g., Grimm *et al.*, 2008; Gray *et al.*, 2009; Meersmans *et al.*, 2009a; Talaab *et al.*, 2013). However, LMMs have not been used that often. The main advantages of median approach and linear regression are ease of application, flexibility, interpretability of parameter and availability on statistical software's (Vittinghoff *et al.*, 2005). Additionally, they can outperform complicated models in case few data points are available (Hastie *et al.*, 2009). Although these benefits seem to be much less compelling as more sophisticated methods with more powerful computers are available, the approaches were retained for comparison with previous studies. LMM is a flexible method in case of correlated data. A major advantage compared to linear regression and median approach by the inclusion of random effects and allowing the modelling of variety of correlation patterns. Consequently, the LMM framework allows understanding of the soil processes by describing the underlying soil mechanism via fixed effects while including the spatial correlation of soil properties via random effects (Lark, 2012). In recent years random forests have become increasingly popular in prediction of soil properties (Grimm *et al.*, 2008; Wiesmeier *et al.*, 2011; Were *et al.*, 2015). Random forests represent supervised learning algorithm combining random decision trees with bagging (Breiman, 2001). The advantages of the method are the large number of predictor variables used in tree constructing, less overfitting producing higher prediction accuracy and fast runtime (Breiman, 2001). However, due to the black-box nature of the method the relationship between the response and the predictor variable is difficult to interpret.

Comparison of prediction models for soil D_b and SOC (%) showed that the best prediction accuracy was achieved by LMMs. However, the difference between RMSE was smaller between the D_b prediction models compared to SOC (%) prediction models as the relative improvement of

the RMSE was 25% for D_b and ~50% for SOC (%) comparing LMM to other studied models. Nevertheless, the observed values where in better accordance with the predicted values for LMMs showing that other approaches over- and underestimated lower and higher D_b or SOC (%) values. In view of the hierarchical structure in the Estonian national SMN database it is not surprising that the LMMs lead to improved prediction accuracy as the competitor models did not incorporate any nested structure. Ignoring the hierarchical structure or in other words neglecting the correlation structure present in the data and simply using a linear regression model will often lead to a misjudgment of the actual amount of information available in the data such as underestimation of the SOC stock.

The comparisons with similar studies in SOC (%) or D_b prediction are scarce in terms of prediction quality measures (RMSE, MSE) or stratification of their dataset to training and validation; although these shortcomings have decreased in recent years. Some studies have used stratification of their data and quality measure like RMSE, mean error or R^2 of their models for estimating the prediction accuracy (e.g., Gray *et al.*, 2009; Meersmans *et al.*, 2011; Botula *et al.*, 2015) however this is not a rule of thumb in validating prediction results (Bell & Worall, 2009). The usability of proposed models is sometimes difficult as the descriptive statistics of soil properties under prediction have not been given (e.g., in Gray *et al.*, 2009). It can be reported that random forests produced higher prediction accuracy on SOC (%) in this thesis (RMSE 0.42%) compared to other studies (Table 8). However, the MLR prediction accuracy in Belgian mineral soils (Meersmans *et al.*, 2011) was better compared to Estonian mineral soils (RMSE 0.48%). The MLR approach for soil D_b prediction yielded in a RMSE of 0.11 g cm^{-3} in Estonian mineral soils. Similar results were obtained by multiple studies (Table 8). Until 2007 most of the studies in D_b prediction used linear regression techniques but due to the data availability from satellite imagery and digital elevation models algorithmic approaches like RF-s, regression trees and ANN-s have become more used (Martin *et al.*, 2009; Taalab *et al.*, 2013, 2015; Adhikari *et al.*, 2014; Sequeira *et al.*, 2014). Therefore, landscape variables, digital elevation model and soil maps have been used instead of measured soil properties (Taalab *et al.*, 2013, 2015; Adhikari *et al.*, 2014). LMM has been used in D_b prediction for identifying the optimum sampling requirement (Lark *et al.*, 2014) resulting in this field-scale study with the RMSE smaller than reported in this thesis. Similar prediction

Table 8. Root mean square error of median approach, linear regression model, linear mixed model and random forests in other soil studies.

Response	Prediction method	RMSE	Reference	Remarks
SOC (%)	Random forests	0.33–1.72	Grimm *et al.*, 2008	RMSE from subsoil to topsoil; tropical soil
SOC (%)	MLR Decision tree Median	1.45 1.50 1.57	Gray *et al.*, 2009	Topsoil; global dataset: climate, topography and parent material as predictor variables
SOC (%)	MLR	0.39	Meersmans *et al.*, 2011	
SOC (%)	Random forests	0.56–0.80	Akpa *et al.*, 2016	RMSE from 15–30 cm to 0–5 cm depths
SOC (%)	MLR	0.35	Xin *et al.*, 2016	
D_b	MLR	0.10–0.17	Kätterer *et al.*, 2006	topsoil, different predictors
D_b	MLR	0.19	Benites *et al.*, 2007	
D_b	MLR	0.15	Tranter *et al.*, 2007	
D_b	MLR	0.11–0.17	Hollis *et al.*, 2012	RMSE from subsoils to topsoils of mineral horizon
D_b	LMM	0.05	Lark *et al.*, 2014	PTF for 2.5–7.5 cm soil depth
D_b	MLR	0.13	Wang *et al.*, 2014	
D_b	MLR	0.13	Botula *et al.*, 2015	Highly weathered soils, different land use

accuracy with Lark *et al.* (2014) was achieved by Martin *et al.* (2009) with gradient boosting algorithm.

The LMM approach was extended with the kriging option to utilize the spatial correlation structure available in the data. However, this was only done for the SOC (%) prediction model as the site effect was the dominant effect amongst other random effects (describing 86.8% of variation). Whereas in D_b prediction model, most of the variation was from the residual error associated with the individual measurements within

each survey plot. Nevertheless, only a small gain was achieved in prediction accuracy by the mixed model-based kriging as the validation data was not designed for the purpose and moreover the sampling scheme was different from the training dataset. This finding clearly emphasizes the need for using prediction models appropriate for the data structure.

It was important to choose the predictor variables in models according to the compatibility of Estonian Soil Map as one of the major benefits of having predicted SOC (%) and D_b values is to create new map layer with SOC stock. SOC stock is an important soil characteristic in soil quality and global climate aspects. At first the SOC stock were calculated based on predicted variables and data from the national SMN database. The estimated SOC stock was different between prediction methods. Again the LMM provided prediction with the smallest RMSE of 7 t ha^{-1} what can be expected as the separate models for D_b and SOC were also best with the LMM. Although no similar comparison cannot be made from the literature, other studies where SOC stock was directly predicted from a set of variables showed prediction accuracy with fairly the same magnitude (Wiesmeier *et al.*, 2011, 2014). However, Martin *et al.* (2014) achieved 2–3 times smaller RMSE with a black-box model predictions. Therefore the question whether to use a single model or combined approach to assess SOC stocks still remains but in the present case no large discrepancies can be noted. Taalab *et al.* (2013) studied the differences in SOC stock estimation when D_b is predicted with random forests or considered as point-based value and soil group mean D_b value. Their results showed that using RF-predicted D_b values the SOC stock was underestimated the D_b whereas stratified mean value of D_b overestimated the measured SOC stock. Meersmans *et al.* (2008) compared mean and models approach in SOC stock prediction and showed a 10.11% difference between them with mean approach producing smaller SOC stock. Meersmans *et al.* (2009a) concluded that SOC stock near the surface soil is determined by land use and soil type. Consequently, in case models are developed on various land uses, the prediction models generally include land use as a predictor.

To add value to the legacy soil data and make the predictions results more available they were implemented to Estonian Soil Map. Integrating prediction results to maps provides a tool for analyzing SOC stock spatial patterns for decision making (e.g., Meersmans *et al.*, 2009a). The predicted SOC and D_b values enabled to produce the first estimation

of Tartu County SOC storage in mineral soils of arable land. So far the gross total SOC stocks of Estonia under different land use types has been estimated (Kõlli *et al.*, 2010) but no detailed map in this large-scale soil map has been produced. SOC stock calculated on soil-type-based polygons show the regional differences and heterogeneity in soil cover over a relatively small region of Estonia. Nevertheless the framework starting from the prediction methodology to implementing the results into soil map can be used in entire Estonia to estimate the arable land SOC stock. Also this methodology can be applied in other similar pedo-ecological conditions.

6.2. Limitations and perspectives

One of the limitations of applying the same methodology in other conditions is certainly the accumulation of errors that are introduced by separately predicting D_b and SOC (%) values. Additionally, the measurement errors in the predictor variables can influence the outcome as in this case only laboratory and sampling data were used of which the latter are highly dependent on the expertise of the experimenter. Any of these factors can introduce uncertainties to the estimates.

Another limitation to the predictions comes from lack of representative data on specific soil types and textures. Specifically, soils with heavier texture and some soil types (e.g., eroded soils) were less represented in the national SMN database compared to others. Therefore, the predictions for the Soil Map database significantly overestimated SOC (%) and stock for slightly and moderately eroded soils. Similarly, organic soils (cultivated) were not used although the amount of carbon stored in Histosols is highly relevant in terms of climate change. More precise predictions could be achieved by using land management history in predictive modelling. Although these data were not available for national SMN database, the results in papers I, II, and III clearly indicated the necessity to include more variables as the site and year effect on SOC (%) was large in LMM and RF approaches. Therefore it can be assumed that crop rotation, fertilizers, soil management and also climate of the year could improve the predictions. For better understanding of climate effect and time (year effect) dynamic SOC models are useful. So it's a perspective to analyze SOC status of Estonian mineral soil using dynamic models instead of static one used now. Another perspective is to predict SOC (%) and stock

in mineral soils under permanent and temporary grassland, organic soil and also soils under forest as the data are available in Estonian conditions. Similarly, as soil store SOC (%) in deeper horizons, the prediction of SOC (%) and stock in entire soil profile will produce more accurate results. Additionally, summing up all the SOC stocks from different land uses would enable to estimate the SOC status of the entire country.

Spatial scale of the predictions has to be treated with care. As was shown, the kriging approach didn't improve the prediction substantially. This is not surprising as the monitoring sites are not scattered randomly over Estonia but form smaller regions which are very distant form each other. Therefore, the predictions at new sites are limited with this model. Perhaps a wider spatial cover of sites could improve the predictions. Another perspective can be including more recent national SMN data to modelling and propose a single model for predicting SOC stock from available soil parameters matching with the Soil Map database. Also, the inclusion of topographic and climatic data to LMM model could improve the predictions. Moreover, development of a black-box model for predicting SOC with measured soil properties and data from digital elevation model and climatic variables. Even in the other studies this approach has been rarely used and is worth of investigating.

SOC stock is only one of the possibilities to use the developed prediction models for SOC (%) and D_b. Other applications can be related to soil management decisions – compaction, modelling soil N content and through that yield models and optimization of fertilizer rates. SOC is related to soil nitrogen and is therefore directly related to yield models, optimization of nitrogen fertilizer rates which are important in crop and environmental science (e.g., Kukk *et al*. (2011) used soil nitrogen content derived from soil map to model biomass yield of reed canary grass). Astover *et al*. (2006) demonstrated how N fertilization can be optimized based on field-specific humus (organic carbon) concentration. Synthesizing the study by Astover *et al* (2006) with current thesis provides a basis for developing spatial decision support system for sustainable soil use.

7. CONCLUSIONS

The key findings and conclusions of the present thesis may be summarized as follows:

- Amongst the predictor variables in D_b prediction models simple regression model overestimated the importance of SOC (%). The linear mixed model approach revealed that after SOC (%) most of the systematic variation in the data can be attributed to sampling depth and Wc. However, both of these predictors are rarely used in D_b modelling studies and therefore the findings in the present thesis suggest that the inclusion of these parameters could significantly improve the predictions.

- In SOC (%) prediction models soil type had the controlling effect on predictions in linear regression, random forests and linear mixed model. Other tested variables (Cf content, A-horizon thickness) showed a marginal but significant impact on SOC (%) prediction.

- Large part of the random variation associated with D_b resulted from residual error related to the individual measurements within each survey plot indicating to the sampling methodology and the lack of information on land use history. However, in SOC (%) prediction most of the variation was found between sites showing that SOC (%) is more variable between sites.

- Comparison of different methods showed that the linear mixed model for SOC (%) and D_b resulted in the highest prediction accuracy (RMSE 0.22% and RMSE 0.09 g cm^{-3}). Random forests represented the black-box models and achieved two times lower prediction accuracy – RMSE 0.42%. However, the latter method amongst with the linear regression methods under- and overestimated soil properties with higher or lower values. The linear mixed model managed to incorporate different types of variations exposed by the monitoring design. Nevertheless, linear mixed model-based predictions are scarce in the literature; however, in some cases the data sampling scheme is suitable for using this method and then it is strongly advisable to use it as the predictions can be otherwise misleading.

- The linear mixed model approach enables to use kriging for utilizing the spatial information available in the data by coordinates of the sites. Mixed model-based kriging for SOC (%) had a very modest gain in prediction accuracy. This can be explained by the validation data not designed for the purpose of prediction and moreover the sampling scheme was different from the training data.

- The predicted D_b and SOC (%) using median approach, linear regression and linear mixed model were used to estimate the SOC stock in mineral soil of arable land in national SMN database. The best prediction accuracy was achieved by the mixed model-predicted values resulting three times higher RMSE of 7 t C ha^{-1} compared to other methods. Thus, the linear mixed model-based SOC stock estimation was implemented into Estonian Soil Map data to observe spatial pattern of SOC stock in mineral soil of arable land in Tartu County. Average estimated SOC stock of Tartu County is 54.8 t C ha^{-1} and total topsoil SOC stock of mineral arable soils 1.8 Mt.

- The methodology outlined in this thesis for improving prediction models is general and applicable to situations with more candidate models in similar data structures. The methodology can be successfully used to renew existing legacy soil data and use it efficiently addressing current scientific and practical soil use questions.

REFERENCES

Adhikari, K., Hartemink, A.E., Minasny, B., Bou Kheir, R., Greve, M.B. & Greve, M.H. 2014. Digital mapping of soil organic carbon contents and stocks in Denmark. PLoS ONE, 9(8), e105519.

Agricultural Research Centre [Põllumajandusuuringute Keskus]. 2016. GIS database for soil fertility.

Akpa, S.I.C., Odeh, I.O.A., Bishop, T.F.A., Hartemink, A.E. & Amapu, I.Y. 2016. Total soil organic carbon and carbon sequestration potential in Nigeria. Geoderma, 271, 202–215.

Andrén, O., Kätterer, T & Karlsson, T. 2004. ICBM regional model for estimations of dynamics of agricultural soil carbon pools. Nutrient Cycling in Agroecosystems, 70, 231–239.

Arrouays, D., Deslais, W. & Badeau, V. 2001. The carbon content of topsoil and its geographical distribution in France. Soil Use and Management, 17, 7–11.

Arrouays, D., Marchant, B.P., Saby, N.P.A., Meersmans, J., Orton, T.G., Martin, M.P., Bellamy, P.H., Lark, R.M. & Kibblewhite, M. 2012. Generic Issues on Broad-Scale Soil Monitoring Schemes: A Review. Pedosphere, 22, 456–469.

Arvidsson, J. 1998. Influence of soil texture and organic matter content on bulk density, air content, compression index and crop yield in field and laboratory compression experiments. Soil and Tillage Research, 49, 159–170.

Astover, A., Kõlli, R., Kukk, L. & Tamm, I. 2010. Tartumaa maaressurss (Land resource in Tartu County. p. 107. Estonian University of Life Sciences, Tartu, Estonia. [in Estonian]

Astover, A., Roostalu, H., Mõtte, M., Tamm, I., Vasiliev, N. & Lemetti, I. 2006. Decision support system for agricultural land use and fertilization optimisation: a case study on barley production in Estonia. Agricultural and Food Science, 15 (2), 77–88.

Balland, V., Pollacco, J.A.P. & Arp, P.A. 2008. Modeling soil hydraulic properties for a wide range of soil conditions. Ecological Modelling, 219, 300–316.

Bates, D., Maechler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1–48.

Batjes, N.H. 1996. Total carbon and nitrogen in the soils of the world. European Journal of Soil Science, 47, 151–163.

Bell, M.J. & Worrall, F. 2009. Estimating a region's soil organic carbon baseline: The undervalued role of land-management. Geoderma, 152(1–2), 74–84.

Benites, V.M., Machado, P.L.O.A., Fidalgo, E.C.C., Coelho, M.R. & Madari, B.E. 2007. Pedotransfer functions for estimating soil bulk density from existing soil survey reports in Brazil. Geoderma, 139, 90–97.

Bernoux, M., Cerri, C., Arrouays, D., Jolivet, C. & Volkoff, B. 1998. Bulk densities of Brazilian Amazon soils related to other soil properties. Soil Science Society of America, 62, 743–749.

Bou Kheir, R., Greve, M.H., Bųcher, P.K., Greve, M.B., Larsen, R. & McCloy, K. 2010. Predictive mapping of soil organic carbon in wet cultivated lands using classification-tree based models: The case study of Denmark. Journal of Environmental Management, 91, 1150–1160.

Bonfatti, B.R., Hartemink, A.E., Giasson, E., Tornquist, C.G. & Adhikari, K. 2016. Digital mapping of soil carbon in a viticultural region of Southern Brazil. Geoderma, 261, 204–221.

Botula, Y., Nemes, A., Van Ranst, E., Mafuka, P., De Pue, J. & Cornelis, W.M. 2015. Hierarchical pedotransfer functions to predict bulk density of highly weathered soils in Central Africa. Soil Science Society of America Journal, 79, 476–486.

Brahim, N., Bernoux, M. & Gallali, T. 2012. Pedotransfer functions to estimate soil bulk density for Northern Africa: Tunisia case. Journal of Arid Environments, 81, 77–83.

Breiman, L. 2001. Random Forests. Machine Learning, 45, 5–32.

Brevik, E.C., Calzolari, C., Miller, B.A., Pereira, P., Kabala, C., Baumgarten, A. & Jordįn, A. 2016. Soil mapping, classification, and pedologic modeling: History and future directions. Geoderma, 264, Part B, 256–274.

Brus, D.J. & de Gruijter, J.J. 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). Geoderma, 80, 1–44.

Calhoun, F.G., Smeck, N.E., Slater, B.L., Bigham, J.M. & Hall, G.F. 2001. Predicting bulk density of Ohio soils from morphology, genetic principles, and laboratory characterization data. Soil Science Society of America, 65, 811–819.

de Brogniez, D., Ballabio, C., Stevens, A., Jones, R.J.A., Montanarella, L. & van Wesemael, B. 2015. A map of the topsoil organic carbon content of Europe generated by a generalized additive model. European Journal of Soil Science, 66, 121–134.

De Vos, B., Van Meirvenne, M., Quataert, P., Deckers, J. & Muys, B. 2005. Predictive quality of pedotransfer functions for estimating bulk density of forest soils. Soil Science Society of America Journal, 69(2), 500–510.

Dendoncker, N., Van Wesemael, B., Smith, P., Lettens, S., Roelandt, C. & Rounsevell, M. 2008. Assessing scale effects on modelled soil organic carbon contents as a result of land use change in Belgium. Soil Use and Management, 24, 8–18.

Doetterl, S., Stevens, A., van Oost, K., Quine, T.A. & van Wesemael, B. 2013. Spatially-explicit regional-scale prediction of soil organic carbon stocks in cropland using environmental variables and mixed model approaches. Geoderma, 204–205, 31–42.

Eesti NSV mullastik arvudes VII, VIII. 1989. Tallinn, 88.

Estonian Land Board. 2001. Vabariigi digitaalse suuremõõtkavalise mullastiku kaardi seletuskiri. URL: http://geoportaal.maaamet.ee/docs/muld/ mullakaardi_seletuskiri.pdf?t=20091211092214

Estonian Land Board. 2016. URL: http://geoportaal.maaamet.ee/eng/Maps-and-Data/Estonian-Soil-Map-p316.html. Last visited: 13.01.2016.

FAO. 2006. Guidelines for soil description. Fourth edition. Rome.

FAO. 2014. Agriculture, Forestry and Other Land Use Emissions by Sources and Removals by Sinks. 1990–2011 analysis. Tubiello, F.N., Salvatore, M., Cóndor Golec, R.D., Ferrara, A., Rossi, S., Biancalani, R., Federici, S., Jacobs, H., Flammini, A. ESS Working Paper No. 2, Mar 2014

Franzluebbers, A.J. 2002. Water infiltration and soil structure related to organic matter and its stratification with depth. Soil and Tillage Research, 66, 197–205.

Gray, J.M., Bishop, T.F.A. & Yang, X. 2015. Pragmatic models for the prediction and digital mapping of soil properties in eastern Australia. Soil Research, 53, 24–42.

Gray, J.M., Humphreys, G.S. & Deckers, J.A. 2009. Relationships in soil distribution as revealed by a global soil database. Geoderma, 150, 309–323.

Grimm, R., Behrens, T., Märker, M. & Elsenbeer, H. 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island — Digital soil mapping using Random Forests analysis. Geoderma, 146, 102–113.

Grüneberg, E., Schöning, I., Kalko, E.K.V. & Weisser, W.W. 2010. Regional organic carbon stock variability: A comparison between depth increments and soil horizons. Geoderma, 155, 426–433.

Håkansson, I. 2005. Machinery-induced compaction of arable soils: incidence – consequences – counter-measures. Reports from the Division of Soil Management No. 109. Swedish University of Agricultural Sciences. SLU Service/Repro, Uppsala.

Haskard, K.A., Rawlins, B.G. & Lark, R.M. 2010. A LMM, with non-stationary mean and covariance, for soil potassium based on gamma radiometry. Biogeosciences Discussions, 7, 1839–1862.

Hastie, T., Tibshirani, R. & Friedman, J. 2009. The Elements of Statistical Learning. 2nd edition. Springer, New York.

Heuscher, S.A., Brandt, C.C. & Jardine, P.M. 2005. Using Soil Physical and Chemical Properties to Estimate Bulk Density. Soil Science Society of America, 69, 51–56.

Hiederer, R. & Köchy, M. 2011. Global soil organic carbon estimates and the harmonized world soil database. EUR 25225 EN. Publications Office of the European Union. 79 pp.

Hollis, J.M., Hannam, J. & Bellamy, P.H. 2012. Empirically-derived pedotransfer functions for predicting bulk density in European soils. European Journal of Soil Science, 63, 96–109.

IPCC, 2014. Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Edenhofer, O., R. Pichs-Madruga, Y. Sokona, E. Farahani, S. Kadner, K. Seyboth, A. Adler, I. Baum, S. Brunner, P. Eickemeier, B. Kriemann, J. Savolainen, S. Schlömer, C. von Stechow, T. Zwickel and J.C. Minx (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Jenny H. 1941. Factors of Soil Formation. McGraw-Hill, New York.

Jones, R.J.A., Hiederer, R., Rusco, E. & Montanarella, L. 2005. Estimating organic carbon in the soils of Europe for policy support. European Journal of Soil Science, 56, 655–671.

Kanal, A. 1996. Organic matter turnover in arable Podzoluvisols. PhD thesis. Estonian Agricultural University. Tartu.

Kask, R. 1975. Eesti NSV maafond ja selle põllumajanduslik kvaliteet [Estonian land resource and its agricultural quality]. Valgus, Tallinn.

Kask, R. 1997. Orgaanilise aine mõjust mulla tihedusele. [Effects of organic matter to soil bulk density] Akadeemilise Põllumajanduse Seltsi toimetised 4, Tartu, 21–24.

Katschinski, N.A. 1956. Die mechanische Bodenanalyse und die Klassifikation der Böden nach ihrer mechanischen Zusammensetzung. Rapports au Sixiéme Congrés de la Science du Sol. Paris, B, 321–327.

Kätterer, T., Andrén, O. & Jansson, P. 2006. Pedotransfer functions for estimating plant available water and bulk density in Swedish agricultural soils. Acta Agriculturae Scandinavica, Section B — Soil & Plant Science, 56, 263–276.

Kaur, R., Kumar, S. & Gurung, H.P. 2002. A pedo-transfer function (PTF) for estimating soil bulk density from basic soil data and its comparison with existing PTFs. Soil Research, 40, 847–58.

Keesstra, S. D., Bouma, J., Wallinga, J., Tittonell, P., Smith, P., Cerdą, A., Montanarella, L., Quinton, J. N., Pachepsky, Y., van der Putten, W. H., Bardgett, R. D., Moolenaar, S., Mol, G., Jansen, B. & Fresco, L. O. 2016. The significance of soils and soil science towards realization of the United Nations Sustainable Development Goals. SOIL, 2, 111–128.

KKM. 2008. Mullaseire 2008 a. URL: http://seire.keskkonnainfo.ee/index.php?option=com_content&view=article&id=3618:2008-a&catid=1034:mullaseire-2008&Itemid=3953

KKM. 2016. Riikliku keskkonnaseire mullaseire alaprogrammi aastaaruanded 2004–2015. URL: http://seire.keskkonnainfo.ee/ index.php?option=com_content&view=article&id=2083&Itemid=396

Kokk, R. & Rooma, I. 1983. Haritavad maad [Arable soils]. Estonian Soils in Figures III: 3–26.

Kokk, R., Rooma, I. & Valler. V. 1968. Mullastiku kaardistamise välitööde metoodika [Field works instructions for large scale mapping of soils]. EPA. Tartu.

Kõlli, R., Astover, A. & Reintam, E. 2011. A review about researches on soil science, Agraarteadus, 12, 21–27.

Kõlli, R. & Ellermäe, O. 2003. Humus status of postlithogenic arable mineral soils. Agronomy Research, 1, 161–174.

Kõlli, R., Köster, T., Kauer K. and Lemetti, I. 2010. Pedoecological Regularities of Organic Carbon Retention in Estonian Mineral Soils. International Journal of Geosciences, Vol. 1 No. 3, 139–148.

Kõlli, R., Ellermäe, O., Köster, T., Lemetti, I., Asi, E. & Kauer, K. 2009. Stocks of organic carbon in Estonian soils. Estonian Journal of Earth Sciences, 58 (2), 95–108.

Krebstein, K., von Janowsky, K., Kuht, J. & Reintam, E. 2014. The effect of tractor wheeling on the soil properties and root growth of smooth brome. Plant, Soil and Environment, 60 (2), 74–79.

Krishnan P., Bourgeon Gérard, Lo Seen Danny, Nair K.M., Prasanna R., Srinivas S., Muthusankar G., Laurent, D. & Ramesh B.R. 2007. Organic carbon stock map for soils of southern India: a multifactorial approach. Current Science, 93 (5), 706–710.

Kukk, L., Roostalu, H., Suuster, E., Rossner, H., Shanskiy, M. & Astover, A. 2011. Reed canary grass biomass yield and energy use efficiency in Northern European pedoclimatic conditions. Biomass & Bioenergy, 35(10), 4407–4416.

Kumar, S., Lal, R. & Liu, D. 2012. A geographically weighted regression kriging approach for mapping soil organic carbon stock. Geoderma, 189–190, 627–34.

Lark, R.M. & Cullis, B.R. 2004. Model-based analysis using REML for inference from systematically sampled data on soil. European Journal of Soil Science, 55, 799–813.

Lark, R.M. 2012. Towards soil geostatistics. Spatial Statistics, 1, 92–99.

Lark, R.M., Rawlins, B.G., Robinson, D.A., Lebron, I. & Tye, A.M. 2014. Implications of short-range spatial variation of soil bulk density for adequate field-sampling protocols: methodology and results from two contrasting soils. European Journal of Soil Science, 65, 803–814.

Lauringson, E., Astover, A., Roostalu, H., Kauer, K., Talgre, L., Penu, P. & Loide, V. 2015. Huumusbilansi mudel taimekasvatuse jätkusuutlikkuse hindamise töövahendina. Eesti Maaülikool, Põllumajandus- ja keskkonnainstituut; Põllumajandusuuringute Keskus.

Lehtveer, R & Kokk, R. 1995. Põllumuldade seire [National Soil Monitoring], Tallinn, käsikiri (unpublished).

Lettens, S., Van Orshoven, J., van Wesemael, B., De Vos, B. & Muys, B. 2005. Stocks and fluxes of soil organic carbon for landscape units in Belgium derived from heterogeneous data sets for 1990 and 2000. Geoderma, 127, 11–23.

Liaw, A. & Wiener, M. 2002. Classification and regression by randomForest. R News 2(3), 18–22.

Lipiec, J. & Hatano, R. 2003. Quantification of compaction effects on soil physical properties and crop growth. Geoderma, 116, 107–136.

Liu, Y., Guo, L., Jiang, Q., Zhang, H. & Chen, Y. 2015. Comparing geospatial techniques to predict SOC stocks. Soil and Tillage Research, 148, 46–58.

Lo Seen, D., Ramesh, B.R., Nair, K.M., Martin, M., Arrouays, D. & Bourgeon, G. 2010. Soil carbon stocks, deforestation and land-cover changes in the Western Ghats biodiversity hotspot (India). Global Change Biology, 16, 1777–1792.

Lugato, E., Panagos, P., Bampa, F., Jones, A. & Montanarella, L. 2014. A new baseline of organic carbon stock in European agricultural soils using a modelling approach. Global Change Biology, 20, 313–326.

Maia, S.M.F., Ogle, S.M., Cerri, C.C. & Cerri, C.E.P. 2010. Changes in soil organic carbon storage under different agricultural management systems in the Southwest Amazon Region of Brazil. Soil and Tillage Research, 106, 177–184.

Manrique, L.A. & Jones, C.A. 1991. Bulk density of soils in relation to soil physical and chemical properties. Soil Science Society of America, 55, 476–481.

Martin, M.P., Lo Seen, D., Boulonne, L., Jolivet, C., Nair, K.M., Bourgeon, G. & Arrouays, D. 2009. Optimizing pedotransfer functions for estimating soil bulk density using boosted regression trees. Soil Science Society of America Journal, 73, 485–493.

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

McBratney, A.B., Mendonça Santos, M.L. & Minasny, B. 2003. On digital soil mapping. Geoderma, 117, 3–52.

McBratney, A.B., Minasny, B., Cattle, S.R. & Vervoort, R.W. 2002. From pedotransfer functions to soil inference systems. Geoderma, 109, 41–73.

Meersmans, J., De Ridder, F., Canters, F., De Baets, S. & Van Molle, M. 2008. A multiple regression approach to assess the spatial distribution of Soil Organic Carbon (SOC) at the regional scale (Flanders, Belgium). Geoderma, 143, 1–13.

Meersmans, J., van Wesemael, B., De Ridder, F. & Van Molle, M. 2009a. Modelling the three-dimensional spatial distribution of soil organic carbon (SOC) at the regional scale (Flanders, Belgium). Geoderma, 152, 43–52.

Meersmans, J., van Wesemael, B. & Van Molle, M. 2009b. Determining soil organic carbon for agricultural soils: a comparison between the Walkley & Black and the dry combustion methods (north Belgium). Soil Use & Management, 25, 346–353.

Meersmans, J., van Wesemael, B., Goidts, E., Van Molle, M., De Baets, S. & De Ridder, F. 2011. Spatial analysis of soil organic carbon evolution in Belgian croplands and grasslands, 1960–2006. Global Change Biology, 17, 466–479.

Metsamajandite maade mullastiku kaardistamise välitööde juhend. 1978/1983. Tallinn.

Milne, E., Adamat, R.A., Batjes, N.H., Bernoux, M., Bhattacharyya, T., Cerri, C.C., Cerri, C.E.P., Coleman, K., Easter, M., Falloon, P., Feller, C., Gicheru, P., Kamoni, P., Killian, K., Pal, D.K., Paustian, K., Powlson, D.S., Rawajfih, Z., Sessay, M., Williams, S. & Wokabi, S. 2007. National and sub-national assessments of soil organic carbon stocks and changes: The GEFSOC modelling system. Agriculture, Ecosystems & Environment, 122, 3–12.

Minasny, B., McBratney, A.B., Malone, B.P. & Wheeler, I. 2013. Chapter One – Digital Mapping of Soil Carbon. Advances in Agronomy, 118, 1–47.

Mishra, U., Lal, R., Slater, B., Calhoun, F., Liu, D. & Van Meirvenne, M. 2009. Predicting soil organic carbon stock using profile depth distribution functions and ordinary kriging. Soil Science Society of America Journal, 73, 614–621.

Morvan, X., Saby, N.P.A., Arrouays, D., Le Bas, C., Jones, R.J.A., Verheijen, F.G.A., Bellamy, P.H., Stephens, M. & Kibblewhite, M.G. 2008. Soil monitoring in Europe: A review of existing systems and requirements for harmonization. Science of the Total Environment, 391, 1–12.

Nanko, K., Ugawa, S., Hashimoto, S., Imaya, A., Kobayashi, M., Sakai, H., Ishizuka, S., Miura, S., Tanaka, N., Takahashi, M. & Kaneko, S. 2014. A pedotransfer function for estimating bulk density of forest soil in Japan affected by volcanic ash. Geoderma, 213, 36–45.

Nawaz, M.F., Bourrié, G. & Trolard, F. 2012. Soil compaction impact and modelling. A review. Agronomy for Sustainable Development, 33, 291–309.

Nugis, E & Lehtveer, E. 1991. On spreading of soil degradation by machines in Estonia. In: Mullakaitse probleeme Eestis. Valgus, Tallinn.

Odeh, I.O.A., Leenaars, J., Hartemink, A.E. & Amapu I. 2012. The challenges of collating legacy data for digital mapping of Nigerian soils. Digital Soil Assessments and Beyond: Proceedings of the 5th Global Workshop on Digital Soil Mapping 2012, Sydney, Australia. Eds. B. Minasny, B. P. Malone, A. B. McBratney. CRC Press

Oueslati, I., Allamano, P., Bonifacio, E. & Claps, P. 2013. Vegetation and topographic control on spatial variability of soil organic carbon. Pedosphere. 23(1), 48–58.

Phachomphon, K., Dlamini, P. & Chaplot, V. 2010. Estimating carbon stocks at a regional level using soil information and easily accessible auxiliary variables. Geoderma, 155, 372–380.

Piho, A., Rooma, I. & Rõõs. O. 1960. Maafondi mullastiku kaardistamise välitööde juhend. Tartu

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & R Core Team (2015). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-121, URL: http://CRAN.R-project.org/package=nlme.

Põllumajanduslike ettevõtete looduslike maade mullastiku kaardistamine. 1989. Tallinn.

Agricultural Research Centre. 2016. GIS database for soil fertility.

R Core Team, 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Reintam, L. 2007. Soil formation on reddish-brown calcareous till under herbaceous vegetation during forty years. Estonian Journal of Earth Sciences, 56(2), 65–84.

Reintam, E., Trükmann, K., Kuht, J., Toomsoo, A., Teesalu, T., Köster, T., Edesi, L., Nugis, E. 2008. Effect of Cirsium arvense L. on soil physical properties and crop growth. Agricultural and Food Science, 17, 153–164.

Rossiter, D.G. 2008. Digital soil mapping as a component of data renewal for areas with sparse soil data infrastructures. In: A.E. Hartemink, A.B. McBratney & M.L. Mendonēa-Santos (Editors), Digital Soil Mapping with Limited Data. Springer, pp. 69–80.

Rousset, F. & Ferdy, J-B. 2014. Testing environmental and genetic effects in the presence of spatial autocorrelation. Ecography, 37(8), 781–790.

Ruehlmann, J. & Körschens, M. 2009. Calculating the effect of soil organic matter concentration on soil bulk density. Soil Science Society of America Journal, 73, 876–885.

Rukhovich, D.I., Koroleva, P.V., Vilchevskaya, E.V., Romanenkov, V.A. & Kolesnikova, L.G. 2007. Constructing a spatially-resolved database for modelling soil organic carbon stocks of croplands in European Russia. Regional Environmental Change, 7, 51–61.

Saby, N.P.A., Arrouays, D., Antoni, V., Lemercier, B., Follain, S., Walter, C. & Schvartz, C. 2008. Changes in soil organic carbon in a mountainous French region, 1990–2004. Soil Use and Management, 24, 254–262.

Schrumpf, M., Schulze, E.D., Kaiser, K. & Schumacher, J. 2011. How accurately can soil organic carbon stocks and stock changes be quantified by soil inventories? Biogeosciences, 8, 1193–1212.

Scull, P., Franklin, J., Chadwick, O.A. & McArthur, D. 2003. Predictive soil mapping: a review. Progress in Physical Geography, 27, 171–97.

Sequeira, C.H., Wills, S.A., Seybold, C.A. & West, L.T. 2014. Predicting soil bulk density for incomplete databases. Geoderma, 213, 64–73.

Stevens, F., Bogaert, P. & van Wesemael, B. 2015. Detecting and quantifying field-related spatial variation of soil organic carbon using mixed-effect models and airborne imagery. Geoderma, 259–260, 93–103.

Stevens, F., Bogaert, P., van Oost, K., Doetterl, S. & van Wesemael, B. 2014. Regional-scale characterization of the geomorphic control of the spatial distribution of soil organic carbon in cropland. European Journal of Soil Science, 65, 539–552.

Stoorvogel, J.J., Kempen, B., Heuvelink, G.B.M. & de Bruin, S. 2009. Implementation and evaluation of existing knowledge for digital soil mapping in Senegal. Geoderma, 149, 161–170.

Taalab, K.P., Corstanje, R., Creamer, R. & Whelan, M.J. 2013. Modelling soil bulk density at the landscape scale and its contributions to C stock uncertainty. Biogeosciences, 10, 4691–4704.

Taalab, K.P., Corstanje, R., Zawadzka, J., Mayr, T., Whelan, M.J., Hannam, J.A. & Creamer, R. 2015. On the application of Bayesian Networks in Digital Soil Mapping. Geoderma, 259–260, 134–148.

Taghizadeh-Toosi, A., Christensen, B.T., Hutchings, N.J., Vejlin, J., Kätterer, T., Glendining, M. & Olesen, J.E. 2014. C-TOOL: A simple model for simulating whole-profile carbon storage in temperate agricultural soils. Ecological Modelling, 292, 11–25.

Tjurin IV. 1937. Soil organic matter and its role in pedogenesis and soil productivity. Study of soil humus. Moskva, Seľskozgiz [in Russian].

Tranter, G., Minasny, B., Mcbratney, A.B., Murphy, B., Mckenzie, N.J., Grundy, M. & Brough, D. 2007. Building and testing conceptual and empirical models for predicting soil bulk density. Soil Use and Management, 23, 437–443.

van Wesemael, B., Paustian, K., Andrén, O., Cerri, C.E.P., Dodd, M., Etchevers, J., Goidts, E., Grace, P., Kätterer, T., McConkey, B.G., Ogle, S., Pan, G. & Siebner, C. 2010. How can soil monitoring networks be used to improve predictions of organic carbon pool dynamics and CO2 fluxes in agricultural soils? Plant and Soil, 338, 247–259.

VandenBygaart, A.J., Gregorich, E.G., Angers, D.A. & McConkey, B.G. 2007. Assessment of the lateral and vertical variability of soil organic carbon. Canadian Journal of Soil Science, 87, 433–444.

Vittinghoff, E., Glidden, D.V., Shiboski, S.C. & McCulloch, C.E. 2005. Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models. Springer, New York.

Vorobeva, L.A. 1998. Khimicheskij analiz pochv [Chemical analysis of soils]. Moscow University Press, Moscow, 272 pp.

Wang, Y., Shao, M., Liu, Z. & Zhang, C. 2013. Prediction of bulk density of soils in the Loess Plateau region of China. Surveys in Geophysics, 35, 395–413.

Were, K., Bui, D.T., Dick, Ų.B. & Singh, B.R. 2015. 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. Ecological Indicators, 52, 394–403.

West, B. T., & Galecki, A. T. 2012. An overview of current software procedures for fitting LMMs. The American Statistician, 65(4), 274–282.

Wiesmeier, M., Barthold, F., Blank, B. & Kögel-Knabner, I. 2011. Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. Plant and Soil, 340, 7–24.

Wiesmeier, M., Barthold, F., Spörlein, P., Geuß, U., Hangen, E., Reischl, A., Schilling, B., Angst, G., von Lützow, M. & Kögel-Knabner, I. 2014. Estimation of total organic carbon storage and its driving factors in soils of Bavaria (southeast Germany). Geoderma Regional, 1, 67–78.

WRB. 2006. World reference base for soil resources 2006. World Soil Resources Reports No. 103. FAO, Rome

Xin, Z., Qin, Y. & Yu, X. 2016. Spatial variability in soil organic carbon and its influencing factors in a hilly watershed of the Loess Plateau, China. Catena, 137, 660-9.

Yang, Q., Luo, W., Jiang, Z., Li, W. & Yuan, D. 2015. Improve the prediction of soil bulk density by cokriging with predicted soil water content as auxiliary variable. Journal of Soils and Sediments, 16, 77–84.

Yu, D.S., Shi, X.Z., Wang, H.J., Sun, W.X., Chen, J.M., Liu, Q.H. & Zhao, Y.C. 2007. Regional patterns of soil organic carbon stocks in China. Journal of Environmental Management, 85, 680–689.

SUMMARY IN ESTONIAN

Mineraalsete põllumuldade orgaanilise süsiniku ja lasuvustiheduse statistilised prognoosimudelid

Mullateaduses on viimastel aastatel hakatud suurt tähelepanu pöörama pärandandmetele. Nendeks peetakse eelmiste põlvkondade kogutud andmeid mullaomaduste kohta. Sellised andmed võivad esineda mitmesugusel kujul: mullaseire/uurimisalade aruanded või andmebaasid, mullastikukaardid, mullaprofiili kirjeldused jm (Odeh *et al.*, 2012). Mulla pärandandmete digiteerimine ja andmeanalüüs võimaldaks seda teavet senisest tõhusamalt otsustusprotsessidesse kaasata.

Ka Eestis on aastakümnete jooksul rohkelt mullaandmeid kogutud, alates riiklikust põllumuldade seirest, muldade suuremõõtkavalisest kaardistamisest ja agrokeemilisest seirest ning lõpetades püsiuurimisalade ja põldkatsetega. Nende mahukate andmekogude puhul on peamiselt piirdutud vaid kirjeldava statistika koostamisega, kuid need sobiksid ka uute andmekihtide modelleerimiseks ja aitaksid nii kaasa nüüdisaegse Eesti mullainfosüsteemi loomisele.

Mulla orgaanilise süsiniku sisaldus ja lasuvustihedus on olulised mulla seisundi näitajad, eriti põllumajanduse seisukohast. Mulla orgaanilise süsiniku sisaldus mõjutab mulla viljakust, huumusseisundit ja selle kaudu ka saagipotentsiaali. Orgaaniline süsinik parandab mulla struktuursust ja veehoiuvõimet, suurendab toiteelementide neelamisvõimet, varustab taimi toiteelementidega ning on mullaelustiku orgaanilise toidu allikas. Üleilmsel tasandil talletab muld kaks korda rohkem süsinikku kui atmosfäär ja kolm korda rohkem kui taimed, olles seega tähtis tegur kliimamuutuste kontekstis (Batjes, 1996). Mulla lasuvustihedus on mulla talitlusi reguleeriv füüsikaline omadus, mille järgi hinnatakse mulla tihenemist ja põllumajandustehnika tekitatud masindegradatsiooni. Peale selle, et mulla orgaanilise süsiniku sisaldus ja lasuvustihedus on iseenesest olulised mullanäitajad, on neid vaja ka mulla süsinikuvaru arvutamiseks.

Alati ei ole otstarbekas ega majanduslikult võimalikki kõiki mulla näitajad otseselt mõõta. Seetõttu on kasutusele võetud mulla seosefunktsioonid, mille abil leitakse mulla näitajad teiste, lihtsamini mõõdetavate näitajate kaudu (McBratney *et al.*, 2002). Maailmas on mitmesugused

pedomeetrilised uurimissuunad (matemaatilised ja statistilised meetodid mullateaduses) väga populaarsed, sest nüüdisaegsete (geo)infosüsteemide võimalused loovad aluse tõhusaks andmetöötluseks ja -esituseks.

Mullaseires toimub andmete kogumine kindla skeemi järgi ja jaguneb üldjoontes kaheks: juhuslik või mittetõenäosuslik (nt. kindlaksmääratud võrgustiku või transektide järgi). Andmete kogumise meetod määrab statistilise andmeanalüüsi meetodi. Kuigi on olemas arvestatav kogus erinevaid prognoosimudeleid nii lasuvustihedusele kui orgaanilise süsiniku sisaldusele, siis paraku ei ole soovitav ega isegi võimalik neid mudeleid kohaldada teistes pedo-klimaatilistes tingimustes. Sageli on see tingitud sellest, et rahvuslikud pärandandmed tuginevad kohalikule mulla klassifitseerimise süsteemile. Seetõttu on jätkuvalt vajalik väljatöötada regioonispetsiifilisi mudeleid, mis ühilduksid kohaliku mullastikukaardiga.

Doktoritöö eesmärk oli välja töötada metoodika, mille abil prognoosida põllumuldade huumushorisondi orgaanilise süsiniku sisaldust (artikkel **II, III**) ja lasuvustihedust (**I**) lihtsamini mõõdetavate ja kättesaadavamate näitajate põhjal. Samuti oli tähtis, et mudelid ühilduksid Eesti suuremõõtkavalise mullastikukaardiga, mis võimaldaks üldistada prognoosimudelite tulemusi ja teha näiteks üle-eestilisi maakasutusotsustusi. Ühtlasi hinnati mudelite põhjal arvutatud süsinikuvaru. Doktoritöös kasutati põllumuldade seire (Kokk & Lehtveer, 1995; KKM, 2008) ja Eesti suuremõõtkavalise mullastikukaardi (1 : 10 000) andmestikku (Maa-amet, 2016) ning loodi sel moel lisandväärtust olemasolevatele pärandandmetele.

Prognoosimudelite koostamiseks kasutati Eesti põllumuldade seire andmestikku peamiselt aastatest 1983–1994. Andmestikus on 90 vaatlusväljakut, mille sees on neli paralleelset transekti kümne vaatluslapiga. Seega on põllumuldade seire proovid kogutud süstemaatilise mittetõenäosusliku meetodiga ning saadud andmestik on hierarhilise struktuuriga. Igal vaatluslapil mõõdeti huumushorisondi tüsedus ja võeti keskmine mullaproov. Seirealad olid peamiselt künnipõhises teravilja külvikorras. Modelleerimisel kasutati ainult põllumuldade huumushorisondi andmeid ning algandmetest võeti valimisse orgaanilise süsiniku sisalduse näitajad vahemikus 0,6–6% (Tjurini meetod). Paralleelselt võeti lasuvustiheduse ja niiskusesisalduse proovid neljast sügavusest. Mulla mehaanilised osakesed jaotati lõimiserühmadesse Katšinski süsteemi järgi, milles füüsikaline savi on määratletud kui osakesed suurusega alla 0,01 mm. Krigingu meetodi kontrolliks kasutati ka Tartu maakonna agrokeemilise seire andmestikku

aastatest 2002–2004 (Põllumajandusuuringute Keskus, 2016). Prognoosimudelite tulemused seostati Eesti suuremõõtkavalise mullastikukaardiga Tartu maakonna näitel.

Prognoosimudeliks nimetatakse statistilist mudelit, mis ennustab uuritava tunnuse väärtusi, kasutades argumenttunnuseid ehk faktoreid. Mullaseire andmestik jagati prognoosimudelite loomiseks juhumeetodil kolme ossa: 50% mudeli loomiseks, 25% mudeli kontrollimiseks ja 25% mudeli katsetamiseks. Lõplik mudel leiti 75% andmestiku põhjal ning ruutkeskmine viga ennustatud ja mõõdetud väärtuste vahel esitati 25% andmestikule. Kasutades mitut andmestikku, saab vältida niihästi liiga optimistlikke järeldusi prognoosi täpsuse kohta kui ka prognoosi täpsuse alahindamist. Parim prognoosimudel oli väikseima ruutkeskmise veaga. Ruutkeskmist viga on lihtne tõlgendada, sest ühik kattub uuritava tunnuse ühikuga.

Töös kasutati nelja staatilist mudelit. Mediaanipõhise meetodi järgi arvutati mediaan orgaanilise süsiniku sisaldusele ja lasuvustihedusele (süsinikuvaru arvutamisel) olenevalt mullaliigist ja lõimisest. Lineaarse regressioonanalüüsi puhul kasutati mitmest regressiooni lasuvustiheduse mudelis ja kovariatsioonanalüüsi orgaanilise süsiniku sisalduse mudelis. Kolmas, mullateaduses vähem kasutatud mudel oli lineaarne segamudel, milles faktorid jaotatakse fikseerituteks ja juhuslikeks. Selles saab mõõdetud faktorite kõrval arvestada ka otseselt mittemõõdetud faktoreid ning andmete korduvmõõtmisi. Viimane prognoosimudel, juhumets, kuulub andmekaevandamise rühma. Juhumets koosneb otsustuspuude kogumist ning need saadakse argumenttunnustest moodustatud valimitest. Juhumetsa puhul otsitakse ja modelleeritakse seaduspärasusi ning otseseid seoseid tunnuste vahel kirjeldada ei saa.

Analüüsi tulemusena selgus, et täpseima prognoosiga on nii orgaanilise süsiniku kui ka lasuvustiheduse puhul segamudel. Lasuvustiheduse segamudeli puhul oli ruutkeskmine viga 0,09 g cm^{-3} olles seega 25% väiksem võrreldes lineaarse regressiooniga. Orgaanilise süsiniku segamudeli prognoos paranes võrreldes teiste mudelitega veelgi rohkem ehk ca 50% ruutkeskmise veaga 0,22%. Segamudeli eelis teiste lineaarsete meetodite ees oli korduvmõõtmistega arvestamine, mis tulenes mullaseire andmestiku hierarhilisest struktuurist: vaatlusväljaku sees oli neli transekti ja igal transektil kümme vaatluslappi.

Lasuvustiheduse segamudelis olid fikseeritud faktoriteks proovivõtusügavus, niiskusesisaldus, füüsikalise savi sisaldus, orgaanilise süsiniku sisaldus, lõimis ja huumushorisondi tüsedus ning nende koosmõjud. Segamudelis kirjeldas proovivõtusügavus kõige suuremat osa lasuvustiheduse süstemaatilisest varieeruvusest. Kirjanduses avaldatud varasemates uuringutes on proovivõtusügavust kasutatud faktorina harva, ent käesoleva doktoritöö põhjal võib väita, et see on vajalik täpsemate prognoositulemuste saamiseks. Samuti on vähe kasutatud proovivõtmisaegset mulla niiskusesisaldust, kuid segamudelis oli see statistiliselt olulise ja suure mõjuga ning sarnase tulemuse on saanud ka Yang *et al.* (2015) uuring. Juhuslikest faktoritest (vaatlusväljak, transekt, vaatluslapp, aasta, aasta ja vaatlusväljaku koosmõju) oli suurima mõjuga koguvarieeruvusele mudeli jääk. Järelikult on vaatluslappidelt võetud keskmiste proovide varieeruvus suur. See võib olla tingitud proovivõtu metoodikast ning faktoritest, mida pole olemasolevate juhuslike faktoritega arvestatud (nt. maakasutus).

Orgaanilise süsiniku sisalduse segamudelis oli kolm fikseeritud faktorit (mullaliik, füüsikalise savi sisaldus ja huumushorisondi tüsedus) ning kuus juhuslikku faktorit (vaatlusväljak, transekt, vaatluslapp, aasta ja nende kombinatsioonid). Kõige suurema mõjuga fikseeritud faktor oli mullaliik (84% varieeruvusest) ning suurim juhuslik varieeruvus oli vaatlusväljakute vahel. Segamudelis on võimalik kasutada juhusliku faktori ehk siinsel juhul vaatlusväljaku parima lineaarse nihketa hinnangu prognoosi (EBLUP) prognoositäpsuse suurendamiseks (**III**). Ruumilise autokorrelatsiooni kirjeldamiseks kasutati eksponentsiaalset variogrammi mudelit, mille ehe variatsioon oli 0,23 ja ulatus 10,5 km. Paraku segamudeli ja krigingu kasutamine oluliselt ei suurendanud prognoositäpsust. Oluliseks limiteerivaks teguriks ruumilise analüüsi juures oli mullaseire vaatlusväljakute väike hajuvus kogu riigi territooriumil; nt Tartu maakonnas olid kõik mullaseire vaatlusväljakud samas piirkonnas.

Viimases etapis arvutati põllumuldade süsinikuvaru, kasutades lineaarse regressiooni, mediaani ja segamudeli ennustatud lasuvustiheduse ja orgaanilise süsiniku sisalduse väärtusi. Oodatult oli täpseima prognoosiga segamudeli ennustatud väärtustel põhinev süsinikuvaru (RMSE = 7 t C ha^{-1} ja keskmine süsinikuvaru 70,8 t C ha^{-1}). Järgmisena leiti Eesti mullakaardile tuginedes Tartu maakonna põllumuldade huumushorisondi süsinikuvaru – 1,8 Mt. Kitsaskohana võib siinjuures välja tuua, et kuigi üldiselt ennustas segamudel hästi, polnud teatud mullaliikide puhul (näiteks erodeeritud mullad) tulemused ootuspärased. Nimelt hindab

mudel erodeeritud muldade süsinikusisaldust üle ja see on tingitud nende muldade vähesest esindatusest seireandmebaasis. Et juhuslikest faktoritest oli orgaanilise süsiniku sisaldusele suurima mõjuga vaatlusväljak (st seirekoht), siis on tõenäoliselt võimalik prognoositäpsust parandada, kaasates andmeanalüüsi täpsemaid andmeid konkreetse põllu maakasutuse kohta, nagu külvikord, väetiste kasutamine, mullaharimine jne. Ühtlasi oleks vaja ennustada uuritud näitajaid kogu mullaprofiili ulatuses, sest süsinikku on talletunud ka sügavamatesse kihtidesse. Huvitava võrdluse saaks rohumaade ja põllukultuuride prognoosimudelite tulemuste kõrvutamisel.

Mulla orgaanilise süsiniku sisalduse ja lasuvustiheduse mudeleid saab kasutada teistegi mulla omaduste hindamiseks, süsinikuvaru on vaid üks rakendusvõimalus. Näiteks saab neid rakendada maakasutuse hindamisel, mulla tihenemise uurimisel ja mulla lämmastikusisalduse modelleerimisel, mis on omakorda aluseks saagimudelitele ning väetamisnormide kujundamisele.

Mullaseire andmestikku kasutades andis parima lasuvustiheduse ja orgaanilise süsiniku sisalduse prognoosi lineaarne segamudel, mis arvestas andmete struktuuri kõige paremini. Väljatöötatud metoodika võimaldab väärtustada mulla pärandandmeid ning kasutada neid tõhusalt aktuaalsete teaduslike ja praktiliste mullaspetsiifiliste küsimuste lahendamiseks.

ACKNOWLEDGEMENTS

The greatest thanks go of course to my brilliant supervisors Professor Alar Astover and Associate Professor Christian Ritz. For sure I would not be here today if it was not for you both. Alar has to be credited for his superpowers in convincing that the work you do is important and applying a PhD thesis is absolutely the thing to do. Also thank you for the guidance in soil science which cannot be underestimated for a young starting scientist. I still have a lot to learn. Thank you Christian for leading me into the mysterious world of statistics and making it possible to dive into the world of R. I finally have realised that statistics is the way of thinking and more of a philosophical thing without right or wrong solutions. Thank you Alar and Christian for your guidance, patience, enthusiasm and encouragement! And of course for all the valuable comments on the articles and the thesis!

I would also like to thank the colleagues from the Department of Soil Science and Agrochemistry from Eerika period: Raja, Imbi and Triin for the relaxing coffee breaks, Merrit, Tõnu, Indrek and Avo for always being there; Liia, Diego and Karin for the discussions about the pros and cons of being a PhD student and of science. Also Hugo, Raimo and Endla for their valuable comments in article preparations. Finally Helis – the best roommate one can ever have and the best friend at all times!

This work could not have been done without legacy soil data. Therefore I want to thank all of the soil surveyors for collecting such valuable data. Especially Igna Rooma as the mapping of the soils and also monitoring of arable land started under his leadership. Also Ragnar Kokk, Rein Lehtveer and Toomas Teras for their hard work during fieldwork and behind the desk. Additionally, Priit Penu and the Agricultural Research Centre for continuing the soil monitoring and collecting soil data from various studies and for allowing me to use it.

I also want to give credit to the financial supporters for enabling a young scientist to visit inspiring and motivating conferences and courses: Estonian Ministry of Education and Research, SA Archimedes (DoRa mobility programme and Kristjan Jaak) and Doctoral School of Earth Sciences and Ecology (the project was financed from European Social Fund through Estonian Operational Programme for Human Resource

Development). Additionally, I am grateful for the financial support that enabled to develop these models: this study was funded by the Ministry of the Agriculture of Estonia project No. 8-2/T10037PKML “Implementation of soil maps and databases for sustainable land use and agricultural production”, and by Estonian Research Council through KESTA programme “Geoinformatic development of biodiversity, soil and earth data systems”. Digitilized data are included in collection Soil Museum (MUMU) at Estonian Unversity of Life Sciences and stored also at geocollections database SARV (http://ermas.geokogud.info/).

Finally, I want to thank my family and friends. Especially Margus for always encouraging and believeing in me and for showing there is always so much to learn. I hope I can be an inspiration for You. Also I want to thank my mother for the support by all means all the time. I hope my two beloved daughters Laura and Grete will one day appreciate the work and studies I have accomplished and follow the steps of their parents for pursuing their dream and the need for new knowledge. Last but not least I want to thank all of my friends, especially the ones in Tartu as these thirteen years of studies from bachelor to PhD would not be the same if it were not for you: Mats, Veiko, Katre, Katrin, Madli, Mart, Sten, Triinu, Kaido, Keiu, Triin. Also as I am a team player I want to acknowledge my dear friends from the volleyball team.

I

ORIGINAL PUBLICATIONS

Suuster, E., Ritz, C., Roostalu, H., Reintam, E., Kõlli, R. & Astover, A. 2011. Soil bulk density pedotransfer functions of the humus horizon in arable soils.

Geoderma, 163, 74–82.

Geoderma 163 (2011) 74–82

Contents lists available at ScienceDirect

Geoderma

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

Soil bulk density pedotransfer functions of the humus horizon in arable soils

Elsa Suuster [a,*], Christian Ritz [b], Hugo Roostalu [a], Endla Reintam [a], Raimo Kõlli [a], Alar Astover [a]

[a] Department of Soil Science and Agrochemistry, Institute of Agricultural and Environmental Sciences, Estonian University of Life Sciences, Kreutzwaldi 1a, Tartu 51014, Estonia
[b] Department of Basic Sciences and Environment, Faculty of Life Sciences, University of Copenhagen, Thorvaldsensvej 40, DK-1871 Frederiksberg C, Denmark

ARTICLE INFO

Article history:
Received 2 July 2010
Received in revised form 29 March 2011
Accepted 4 April 2011

Keywords:
Soil bulk density
Organic carbon
Sampling depth
Water content
Mixed model
Prediction quality

ABSTRACT

Soil bulk density (D_b) is an important indicator of soil quality, site productivity and soil compaction and is widely used as an input variable for various models. However, D_b is often not determined during actual soil surveys, and many attempts have been made to predict it from other more easily and routinely measured variables. The aim of this study was to develop a predictive model for D_b estimation for the humus horizon of arable soils. In addition, we hypothesised that the mixed model approach would enable more accurate predictions than the multiple regression model. Estonian National Soil Monitoring data (1983–2008) were used for D_b modelling. The dataset contains 17,293 entries for the humus horizon of Estonian arable soils. Mixed model methodology and linear multiple regression analysis were used to develop prediction models. Soil sampling depth, soil organic carbon and water content had the greatest effects on D_b estimation. D_b is determined through interactions of explanatory variables, and their contribution to D_b variation in mixed model differs from variable contribution indicated by the simple regression model. We found that the regression model overestimates the importance of SOC. The mixed model prediction was an improvement compared to multiple regression models (MSE = 0.009 compared to MSE = 0.014). In addition, the predicted values in the mixed model showed good agreement with observed values, while the multiple regression model underestimated D_b values above 1.6 g cm^{-3} and overestimated D_b values below 1.2 g cm^{-3}. We conclude that the mixed model approach enables higher prediction accuracy than multiple regression analysis. We propose a mixed model with 20 variables as a prediction model for the humus horizon of Estonian arable soils. The methodology is also applicable for other soil conditions.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

Soil bulk density (D_b) is widely used as an input variable in various predictive and descriptive soil models. It is the primary soil physical property and is used for weight-to-volume conversion, essential for the assessment of soil nutrient pools and carbon stocks. D_b affects site productivity and is an indicator of soil compaction and porosity. Soil compaction is studied due to its effects on several agricultural issues including crop yields and quality reduction (Batey, 2009). Soil organic carbon (SOC) has become increasingly interesting in the context of climate studies because carbon sequestration into, and its release from, soil carbon pools may influence global warming (Batjes, 1998; Smith, 2004). Therefore, precise estimates of D_b are essential for reliably calculating SOC stocks at various scales.

Despite the importance of D_b, it is often not determined in soil surveys. Direct measurement of D_b is considered to be labour-intensive, time-consuming, tedious, and expensive. Thus, several empirical functions have been developed to predict D_b from more easily collected physical and chemical soil data (e.g. Benites et al., 2007; Heuscher et al., 2005; Kaur et al., 2002; Tranter et al., 2007). In soil science, these predictive functions are usually termed pedotransfer functions (PTFs), and they predict certain soil properties from other easily, routinely, or inexpensively measured properties (McBratney et al., 2002).

The systematic variation of D_b is described mostly by SOC and texture (Benites et al., 2007; Bernoux et al., 1998; Heuscher et al., 2005; Huntington et al., 1989; Kaur et al., 2002; Manrique and Jones, 1991; Tomasella and Hodnett, 1998). Soil chemical properties have also been shown to be of great importance for the determination of D_b (Benites et al., 2007). Some studies have investigated the relevance of soil sample depth and the suborder of the soil profile to D_b estimation (De Vos et al., 2005; Harrison and Bocock, 1981; Heuscher et al., 2005; Tranter et al., 2007).

Methods used for constructing PTFs for D_b vary from basic statistical methods to multivariate spatial statistics. Stepwise multiple linear regression has been commonly used (e.g. Benites et al., 2007; Bernoux et al., 1998), although various other techniques are now available. Tranter et al. (2007) concluded in their study that artificial neural network and single regression tree models do not show improved performance of models compared with the multiple linear regression. On the other hand, Martin et al. (2009) found that multiple additive regression trees improved the accuracy of D_b predictions over multiple regression methods. PTFs for D_b have been developed for

* Corresponding author. Tel.: +372 7313545; fax: +372 7313 539.
E-mail address: elsa.suuster@emu.ee (E. Suuster).

0016-7061/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.geoderma.2011.04.005

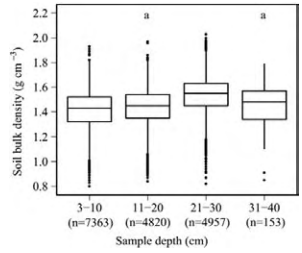

Fig. 2. Variation of soil bulk density by sampling depth (similar letters indicate statistically insignificant differences, P>0.05).

however, in 2008, a steel cylinder of 88.2-cm^3-volume (d = 5.3 cm, h = 4 cm) was used. Measured D_b values correspond to the equilibrium D_b because sampling was carried out at the end of the growing season. D_b values below 0.8 g cm^{-3} or above 2.0 g cm^{-3} were excluded from this study, as they represent unreliable data for the A-horizon in mineral soils. The gravimetric water content was determined at 105 °C (% weight). The sampled soils were classified according to the ESC; where approximate soil groups are given, classifications were also based on the World Reference Base 2006 (WRB, 2006). The thickness of the A-horizon ranged from 12 to 100 cm; data were carefully investigated to ensure that the sampling depth would not exceed the A-horizon thickness. The SOC concentration was determined by Tjurin, based on wet digestion with acid dichromate of carbon (Vorobeva, 1998), which is similar to the Walkley–Black procedure (Walkley and Black, 1934). SOC values above 6% (60 g kg^{-1}) were excluded to eliminate samples in the database that contained histic materials. Particle-size distribution of fine earth was determined using the pipette method by Katchinski, where physical clay is defined as the mass fraction of particles smaller than 0.01 mm (Katschinski, 1956). According to the Katschinski classification, the soils were classified based on their physical clay (<0.01 mm) content as follows: those with less than 10.0% Cf were classified as sand; from 10.1 to 20% as loamy sand; from 20.1 to 30.0% as light loam; from 30.1 to 40.0% as medium loam; from 40.1 to 50.0% as heavy loam; and soils with over 50.1% Cf were classified as clay. The distribution of D_b among texture classes indicates that D_b decreases in heavier textures (Fig. 3). These texture classes were used to ensure that the developed models would be compatible with the encodings used for the large-scale soil map of Estonia.

2.2. Statistical procedures

2.2.1. Mixed model methodology

The sampling scheme described in section 2.1 implies that there is a hierarchical structure present in the data for each year: plots are nested within transects, which are again nested within sites. We used a mixed model approach to accommodate this hierarchical structure in the statistical analysis.

As measurements from within the same transect are likely to be more closely correlated than measurements from different transects, we may need to use a model that allows for correlation between measurements from the same transect. The presence of correlation may reflect that various minor, unknown transect-specific factors influence all measurements from a given transect. However, such effects are not usually captured by any of the available explanatory variables that have actually been recorded for this particular transect (typically, the underlying factors may be entirely unknown to the experimenter or may be very difficult to measure or quantify).

Table 1
Descriptive statistics of soil properties used in modelling by soil groups.

Soil or soil association (*WRB 2006* — approximate translation)	Soil code by ESC	Property				
		D_b (g cm^{-3})	Wc (%)	A (cm)	SOC (%)	Cf (%)
Haplic Regosol (Calcaric, Hyperskeletic) — pebble rich rendzinas	Kr	1.45 ± 0.16[a]	16.1 ± 5.7	26.6 ± 4.6	2.13 ± 0.66	26.3 ± 5.2
	(n = 266)[c]	1.25…1.64[b]	8.6…22.2	21…32	1.48…3.21	18.3…31
Haplic Cambisol and Regosol (Calcaric, Endosceletic) — pebble rendzinas	K	1.48 ± 0.15	17.4 ± 5.2	27.8 ± 5.1	1.97 ± 0.61	26.9 ± 5.2
	(n = 3393)	1.30…1.66	10.9..22.9	22…34	1.33…2.66	21.1…34.2
Haplic Phaeozem (Calcaric) — typical brown soils	Ko	1.48 ± 0.14	17.5 ± 4.8	27.4 ± 4.0	1.6 ± 0.37	24.0 ± 5.2
	(n = 1594)	1.30…1.66	11.3…23.3	23…32	1.16…2.11	17.3…31.6
Cutanic Luvisol (Humic) — eluviated brown soils	Kl	1.49 ± 0.12	16.7 ± 5.3	28.5 ± 5.1	1.49 ± 0.37	21.1 ± 4.8
	(n = 2443)	1.33…1.64	10…22.9	24…33	1.03…1.92	15.8…27
Umbric Stagnic Albeluvisol — pseudopodzolic soils	LP	1.49 ± 0.13	13.9 ± 5.5	28.6 ± 4.0	1.15 ± 0.26	16.8 ± 3.7
	(n = 3661)	1.33…1.65	6.4…20.5	25…33	0.88…1.50	13.1…19.9
Umbric Carbic Podzol — sod-podzolic soils	Lk	1.53 ± 0.1	15.2 ± 3.6	29.3 ± 4.5	1.01 ± 0.18	10.7 ± 3.2
	(n = 107)	1.41…1.64	10.5…17.9	26…33	0.85…1.17	8.1…15
Endogleyic Phaeozem — gleyic typical brown soils	Kog	1.38 ± 0.18	16.5 ± 7.8	27.1 ± 3.4	2.14 ± 0.77	26.5 ± 3.5
	(n = 756)	1.15…1.63	8.5…29.3	23…31	1.29…3.34	21.8…31.4
Endogleyic Cutanic Luvisol (Humic) — gleyic eluvial soils	Klg	1.42 ± 0.18	18.8 ± 8.9	29.1 ± 4.6	2.05 ± 0.84	25.2 ± 4.0
	(n = 1497)	1.17…1.65	8.3…31.8	25…35	1.22…3.42	21.5…29.9
Umbric Stagnic Endogleyic Albeluvisol — gleyic pseudopodzolic soils	LPg	1.47 ± 0.15	18.3 ± 7.0	29.2 ± 4.8	1.5 ± 0.74	22.2 ± 6.7
	(n = 943)	1.27…1.65	11.1…26.7	25…34	0.90…2.29	12.9…30.3
Luvic Mollic Spodic Gleysol — gleysols	G	1.31 ± 0.16	26.3 ± 11.1	28.4 ± 7.8	3.04 ± 1.07	29.6 ± 18.9
	(n = 2088)	1.09…1.49	12.3…40.5	21…36	1.64…4.47	6.9…60.4
Haplic Cambisol. Cutanic Luvisol. Umbric Albeluvisol slightly eroded	E1	1.59 ± 0.13	14.5 ± 4.2	21.8 ± 3.1	1.01 ± 0.46	19.2 ± 4.3
	(n = 185)	1.46…1.73	9…19.1	18…26	0.59…1.44	13.4…25
Slightly eroded soils						
Aric Regosols (Eutric) moderately eroded — moderately eroded soils	E2	1.65 ± 0.10	14.5 ± 2.4	18.8 ± 3.2	0.96 ± 0.26	21.3 ± 3.2
	(n = 93)	1.50…1.78	11.6…17.8	14…22	0.60…1.30	18.4…24.9
Colluvic Regosol (Cumulihumic) — deluvial soils	D	1.50 ± 0.14	17.6 ± 4.7	60.1 ± 21.3	1.48 ± 0.81	19.1 ± 3.5
	(n = 233)	1.29…1.66	12.9…23.6	32…98	0.85…2.20	14.7…24.2

ESC — Estonian Soil Classification, D_b — bulk density, Wc — soil moisture content in the moment of soil sampling, A — thickness of humus horizon, SOC — soil organic carbon content, Cf — physical clay content (particles with diameter less than 0.01 mm).

[a] Mean ± standard error.
[b] 0.1 decile…0.9 decile.
[c] Count.

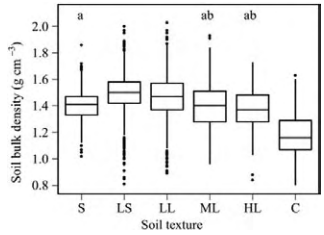

Fig. 3. Soil texture classes by D_b and the multiple comparison results (similar letters indicate statistically insignificant differences, P>0.05). S — sand (n=272); LS — loamy sand (n=3615); LL — light loam (n = 5019); ML — medium loam (n=1037); HL — heavy loam (n=92); C — clay (n=170).

Similarly, we anticipate that correlation could also be present at the plot and site levels. Moreover, we can even anticipate correlation that is related to differences in agricultural practices or year-to-year and site-to-site (within year) climate variations. Again, it may be very difficult (or even impossible) to seek to quantify the factors that cause such discrepancies. In other words, the hierarchical structure in the data introduces several layers of replication and, as a result, the data may exhibit unexplained variation not only in the individual D_b measurements, but also at transect, plot, and site levels. This phenomenon is often referred to as pseudo-replication.

Ignoring the hierarchical structure (in other words neglecting the correlation structure) present in the data, and simply using a multiple linear regression model, will often lead to a misjudgement of the actual amount of information available in the data and this may be problematic for results based on estimation and hypothesis testing. Therefore, in many situations, mixed models are the most suitable statistical methodology for analysing complex data structures.

The mixed model approach handles correlation by assuming that each transect, plot, or site (or any other grouping level) in the data correspond to some random, but unobservable quantity, a so-called *random effect* (e.g. if there are 90 sites in the database, there will be 90 site-specific random effects in the model). As a consequence, all measurements within a specific transect, plot, or site share this random quantity and therefore they are no longer independent of each other but are correlated (Goidts et al., 2009). We note that within the above framework, the correlation between any pair of measurements is not adapted to any geographical or spatial information about the measurements (i.e., the correlation is the same for a pair of neighbouring sites as it is for two sites located far apart).

Another important consequence of introducing random effects is that we assume that transects, plots, and sites present in the database were chosen at random (or were sampled randomly) from representative collections (or distributions) of transects, plots, and sites in Estonia. In practice, this assumption may be an idealisation that is only satisfied approximately because of practical limitations in how the sampling was carried out. However, it still means that the conclusions and results based on the mixed model analysis are valid and can be generalised to the entire populations from which samples in the database were obtained. Such generalisations would not be warranted outside of the mixed model framework with the above assumptions implicitly being in force.

Apart from the above considerations (which are related to the data structure, that is, the experimental design), a mixed model can be thought of in much the same way as multiple linear regression and analysis-of-variance models. Therefore, the mixed model can be used to describe average trends in the database as explained by a number of categorical, and/or quantitative explanatory variables (which in the context of mixed models are called *fixed effects* because they are related to systematic or fixed patterns in the data that are believed to be controlled by some available explanatory variable), but at the same time, the mixed model makes appropriate adjustments for pseudo-replication.

Ideally, the mixed model would also utilise coordinate information by imposing a spatial correlation structure on the random effects (Lark and Cullis, 2004). Thus far, we have assumed sites to be independent, ignoring the fact that neighbouring sites may have more in common (e.g. they may share geological properties or land-use characteristics) than sites that are further apart. Such similarities between sites may decline or even disappear completely as the distance between sites increases. For the present study, the assumption of independence between sampling sites may be justified by the widespread distribution of sites across Estonia. On a more practical note, coordinate information is not yet available for the data reported in this paper. Without coordinate information, no spatial correlation structure can be included in the mixed model; this fact precludes us from using kriging methods (Lark et al., 2006) and will have some bearing on the way we derive predictions. Below, we discuss prediction in more detail.

2.2.2. Estimation and hypothesis testing

Our initial mixed model includes the following explanatory variables: SOC, Wc, Cf, A-horizon thickness, texture, sampling depth, and soil type as well as their combinations as fixed effects. Site, transect, plot, year, and the site-year interaction were considered random effects (Table 2). This model was fitted using maximum-likelihood estimation, and the validity of the standard assumptions about correct mean structure, variance homogeneity, and normality of the all involved random effects was assessed using residual plots and quantile–quantile plots of the raw residuals and the estimated random effects (the so-called *EBLUP*s). Likelihood ratio tests were used to simplify the fixed-effects part of the model as much as possible by leaving out non-significant terms (e.g. stepwise backward reduction). For the final model, parameter estimates are reported by means of partial models that summarise the effects of the quantitative explanatory variables as the categorical explanatory variables (e.g. soil type and texture) change from one level to another. P-values based on approximate t-tests were used to evaluate the statistical significance of parameters of interest (such as slope parameters) that were associated with the quantitative explanatory variables (e.g. SOC and Wc) and to make pair-wise comparisons between the levels of the categorical explanatory variables left in the final model. To provide additional insight about how the systematic variation is distributed across the explanatory variables, analysis of variance was used for calculating the proportions of variation explained.

2.2.3. Prediction

By a prediction model, we mean a statistical model capable of providing individual predicted D_b values (point estimates of the D_b) for various configurations of explanatory variables. In contrast to estimation and hypothesis testing explained above, we are no longer concerned with average trends in the data (these may or may not be useful for predicting individual measurements). Instead, the accuracy

Table 2
Random and fixed effects included in hypothesis testing and prediction model ("*" denotes combined terms where the effect of a quantitative variable is modified according to the level of a categorical variable).

Fixed effects			Random effects
Soil type	Sampling depth	Sampling depth * SOC	Site
SOC	Thickness of A horizon	Sampling depth * Cf	Transect
Wc	Texture * moisture	Soil type * Wc	Plot
Cf	Texture * SOC	Soil type * SOC	Year
Texture	Sampling depth * Wc	Soil type * Cf	Site * year

of the predicted D_b values is the primary quality measure by which to assess a prediction model. Classical features of a statistical model (such as estimated standard errors, confidence intervals, and p-values) are of no interest in the context of prediction. Moreover, it is possible that significant terms may not contribute appreciably to the predictions, whereas non-significant terms may improve predictions substantially.

Consequently, to establish an appropriate prediction model, a different approach is needed. The first step is to define a number of *candidate models* based on well-defined criteria that reflect the intended applications of the models. In our case, such criteria must be compatible with the Estonian soil map; that is, soil-science-specific considerations replace statistical techniques. The second step is to evaluate and compare the candidate models to select the best one, and the third step is to assess the performance of the chosen model.

The initial mixed model was our first candidate for a good prediction model. However, we also considered a few other models that involved fewer explanatory variables and that did not include random effects. The second model considered (model A) contained the variables sampling depth, 1/Wc, Cf and Wc interaction, A-horizon thickness, SOC, Cf, and Cf squared. We included transformed version of Wc and Cf because their effects on D_b were not linear in the original scale and thus needed to be transformed. The third model (model B) had the simplest structure. It only involved three explanatory variables (sampling depth, texture classes, and SOC) and would therefore be widely applicable. The latter two candidate models are multiple linear regression models; these were fitted using ordinary least-squares estimation. The rationale for considering these models is that they may be applicable for geographical areas where no detailed soil information is available (that is, only a few explanatory variables are recorded). Moreover, compatibility with the Estonian digital soil map is also important. Thus, three candidate models were considered, although the procedure we describe below will work for larger collection of candidate models.

To select the best model among our candidates, we fitted each candidate model to common *training* data (a subset of the entire database), and subsequently we compared the models using *validation* data (another smaller subset of the database). More specifically, the models fitted to the training data were used to obtain predicted D_b values for the configurations of explanatory variables found in the validation data. The comparison of the candidate models was then based on comparing these predicted values with the corresponding observed values in the validation data by means of summary measures (e.g. mean square error MSE, the mean of the squared differences between predicted and observed values, or RMSE, the square root of MSE) (Goidts et al., 2009); for both measures, smaller is better. Additionally, the agreement or disagreement between the predicted and observed values was assessed visually (e.g. Meersmans et al., 2009).

Two final models were chosen based on predictive quality (as evaluated using the above procedures) and taking into consideration compatibility with the soil survey information and the soil map. Once we had decided on the final prediction models, they were re-fitted to the combined training and validation datasets and could be used for future prediction. Finally, the predictive performance was evaluated and reported using a *test* dataset (yet another smaller subset of the entire database). By using a completely different subset at this stage, we ensured that we were not underestimating the prediction error (which could result from using the same data for both choosing and assessing the best model). We reported the resulting MSE and plotted the predicted values versus the observed values.

We followed the rule of thumb of dividing the initial database randomly into three parts: 50% constituted the training dataset, 25% constituted the validation dataset, and 25% constituted the test dataset (Hastie et al., 2009, pp. 219–223).

Mixed model based prediction relies on both the estimated fixed-effects parameters and the EBLUPs as estimates for the unobserved random effects. Moreover, for prediction purposes, the mixed model exhibits a desirable feature called *shrinkage* that results in more accurate predictions by utilising the EBLUPs. Essentially, shrinkage is a way to ensure that an observed D_b value that is extreme (for some configuration of the explanatory variables) does not necessarily result in an equally extreme future prediction for the same configuration of explanatory variables (Steyerberg, 2009, p. 87). However, for prediction involving a new site, corresponding EBLUPs are not directly available as this site was not part of the data used for establishing the original model. There are two possible solutions to handle this issue: 1) to use an EBLUP equal to 0 for any new site, which will give results close to those based on the corresponding multiple linear regression models (obtained by leaving out the site-specific random effects) or 2) to devise some extrapolation scheme for estimating EBLUPs for a new site based on neighbouring EBLUPs present in the model (*kriging*). For reasons already outlined in section 2.2.1, we use the first option.

All statistical analyses were carried out using the statistical software R version 2.12.0 (R Development Core Team, 2010). Specifically, for the mixed model analysis, the function lmer() in the extension package lme4 was used (Bates and Maechler, 2010).

3. Results

3.1. Average trends in the database

Most variation results from the residual error associated with the individual measurements within each survey plot. Among the random effects included in the mixed model, year-specific site effects are responsible for the most variation (68% relative to the residual error). Site and year effects each roughly contribute half as much variation as does the residual error. The variation resulting from transects and survey plots is marginal.

Soil type is statistically significant, but is responsible for a relatively small part of the variation in D_b (1.7%, which is the same amount as the variation caused by the A-horizon thickness). Soil type effects on D_b estimation have a similar magnitude (except for Deluvial soils), but this difference is not statistically significant. The influence of Cf on D_b does not depend on soil type (Table 3). SOC decreases D_b in studied soil types (except for soil type Lk), but no specific pattern could be discerned (Table 4). Water content has a decreasing effect on D_b, and this effect is highly dependent on soil type. The soil type specific slope parameters for Wc and SOC were not statistically significant different.

There is an overall borderline-significant effect for sampling depth, which contributes the most in describing the systematic variation in D_b. Moreover, D_b increases to a sampling depth of 25 cm and then decreases for the greatest depth. No statistically significant differences were found between a sampling depth of 40 cm and sampling depths of 15 and 25 cm (data not presented). Effects of SOC, Cf, and Wc significantly influence D_b with sampling depth. SOC slopes are not statistically different between sampling depths, but an increase in sampling depth increases the effect of SOC. Higher water content

Table 3
Mixed models fixed effects and their *p*-values.

Fixed effects	P-value
Sampling depth * Wc	<0.001
Sampling depth * SOC	<0.0001
Sampling depth * Cf	0.002
Soil type * Wc	<0.0001
Soil type * SOC	0.006
Soil type * Cf	0.066
Texture * Wc	<0.001
Texture * SOC	0.990
Thickness of A horizon	<0.001

Table 4
Partial models showing the effect of soil type, sampling depth, and soil texture on the D_b level. Bold numbers indicate statistically significant slopes ($p<0.05$).

	Estimate[b]	SE[c]	Wc	SOC	Cf
Soil type[a]					
Kr	1.42	0.04	0.003	−0.08	
K	1.43	0.03	0.0001	−0.08	
Ko	1.43	0.03	−0.001	−0.09	
KI	1.43	0.03	−0.001	−0.08	
LP	1.43	0.04	−0.002	−0.07	
Lk	1.45	0.17	−0.007	0.02	
Kog	1.40	0.04	−0.008	−0.05	
KIg	1.41	0.04	−0.007	−0.04	
LPg	1.41	0.04	−0.007	−0.03	
G	1.43	0.04	−0.007	−0.03	
E1	1.44	0.04	−0.012	−0.07	
E2	1.44	0.07	−0.005	−0.11	
D	1.57	0.04	−0.011	−0.01	
Sampling depth (cm)					
3	1.57	0.04	**−0.011**	−0.005	**0.003**
15	1.59	0.04	**−0.012**	0.005	**0.002**
25	1.65	0.04	**−0.012**	0.012	**0.002**
40	1.64	0.05	−0.014	0.072	−0.005
Soil texture					
Sand	1.57	0.04	−0.01		
Loamy sand	1.56	0.01	−0.02		
Light loam	1.56	0.02	−0.02		
Medium loam	1.54	0.02	−0.02		
Heavy loam	1.42	0.03	−0.02		
Clay	1.57	0.04	−0.01		

[a] Soil type by ESC classification (references to the WRB shown in Table 1).
[b] Scaled estimate.
[c] Standard error of intercept estimate.

significantly decreases the effect of sampling depth on D_b; however, clay content has the opposite effect.

Soil texture classes are important components for D_b estimation and have similar effects, except for heavy loam. The slope coefficients for water content for different soil texture classes indicate that for all of the soil textures classes considered, a change in the Wc results in a similar reduction in D_b both for lighter and heavier textures. An exception is the texture class "sand", for which the Wc slope is smaller (but still negative). There are several statistically significant differences between texture classes (data not presented).

Water content describes 17% of the variation found in the D_b estimation and has the second strongest influence on D_b. SOC explains two times less variation than does Wc and ten times less than does sampling depth. A thicker A-horizon decreases D_b estimates.

3.2. Prediction models

One of the objectives of this paper was to propose a prediction model to calculate D_b in the A-horizon of Estonian arable soils. Thus, three candidate models were fitted to the training data: model A (sampling depth, 1/Wc, A-horizon thickness, SOC, Cf and Wc interaction, Cf and squared Cf), model B (SOC, texture and sampling depth), and a mixed model (Table 2). Validation of the three candidate models indicated that the mixed model produced a very good agreement between observed and predicted values, whereas models A (adjusted-$R^2=0.43$; MSE = 0.014) and B (adjusted-$R^2=0.35$; MSE = 0.015) underestimated values above 1.6 g cm^{-3} (Fig. 4). This is in agreement with calculated mean square error for the validation dataset, in which the mixed model had the smallest MSE (0.009).

Based on these results, we tested the performance of model A (Eq. (1)) and the mixed model (Eq. (2)). The mixed model predicted more accurately than model A, but model A is comparable with the methods used in previous field studies (e.g. Benites et al., 2007).

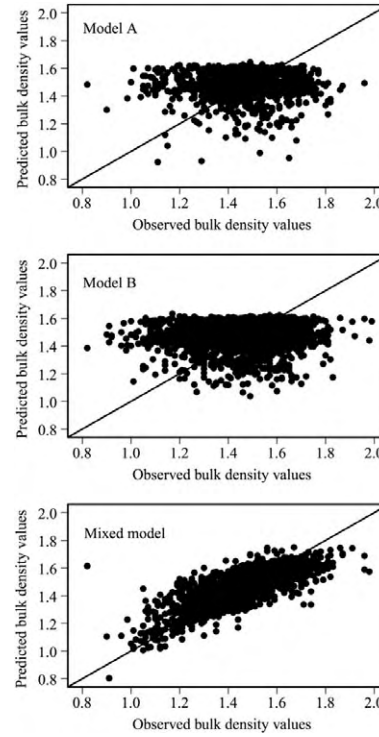

Fig. 4. Validation of the candidate models by comparing observed and predicted D_b values (g cm^{-3}) of the validation sub-dataset.

Model A parameter estimates agree with the mixed model results (estimates shown in Eq. (1)). Wc, A-horizon thickness, SOC, and interaction between Wc and Cf had a negative effect on D_b in model A. In addition, the slope of the transformed Wc had the greatest magnitude of all of the slope parameters, but for higher Wc levels, the effect on D_b estimation decreased. In model A, SOC accounted for 23% of the variation in D_b and led to a decrease in D_b with increasing SOC in the A-horizon. The mixed model parameter estimates showed the same trends as described in section 3.1. Sampling depth parameter estimates did not show the same effect in model A as in the mixed model. In model A, sampling depth parameter estimates increased up to 40 cm depth, but in the mixed model, the highest estimate was for a sampling depth of 25 cm.

Model A.

$$D_b = \beta_0 - 0.45 / Wc - 0.004 * A(\text{cm}) - 0.08 * SOC + 0.00004 * Cf^2 + 0.01 * Cf - 0.0002 * Wc : Cf, \quad (1)$$

where β_0 is the intercept, which is 1.65 for sampling depth 3 cm, 1.68 for 15 cm, 1.76 for 25 cm and 1.77 for 40 cm.

The mixed model is shown here with the actual parameter estimates shown in Table 4:

$$D_b = \beta 1_{depth} + \beta 2_{soil\ type} + \beta 4_{A(cm)} + \beta 6_{texture} + \beta 7_{texture^* Wc} + \beta 8_{depth^* Wc} + \beta 9_{soil\ type^* Wc} + \beta 10_{depth^* SOC} + \beta 11_{soil\ type^* SOC} + \beta 12_{soil\ type^* Cf} + \beta 13_{depth^* Cf} + A_{site} + B_{site^* year} + C_{year} + D_{transect} + E_{point} \quad (2)$$

The calculated MSE based on the test dataset for the two models remains 0.013 for model A and 0.009 for the mixed model, indicating that both models work well for different data. The plot of observed D_b against predicted values indicates good agreement for the mixed model (Fig. 5). In contrast, model A has limited predictive potential, strongly underestimating D_b values above 1.6 g cm^{-3} and overestimating values below 1.2 g cm^{-3}. Results indicate that the combined mixed-effects prediction is more precise than the prediction without random effects (Fig. 6). Prediction based on fixed effects only is more scattered and more in disagreement with the reference line as compared to the mixed-effects predictions. This indicates less accurate results, which lead to both under- and overestimating D_b values.

4. Discussion

The mixed model approach identified salient and interesting patterns and trends that describe the variation in D_b. The results indicate that a large part of the variation results from the individual measurements within each survey plot. This can be explained by the sampling methodology: D_b is measured by taking cores, and thus the results are highly dependent on the sample collector. The small volume of cores (most 50-cm^3-volume) used for sampling could also be the reason for some of the unexplained variation. Another reason for the large residual variation could be that some effects are not taken into account in the model. Inclusion of land-use history in the mixed model approach may improve the results significantly because cropping patterns and tillage practices influence soil D_b (Lampurlanés and Cantero-Martinez, 2003; Lauringson, 2003; Reintam et al., 2008). Our database precluded applying cropping and tillage history as explanatory variables. However, these effects could still be present in the model indirectly as part of the random year effect. The random effects of site, site within year, and year incorporated equal amounts of variation; this indicates substantial differences between sites due to diverse pedo-climatic conditions. Transect and survey plot contribute less as variation at these finer levels of sampling was low, reflecting a high degree of homogeneity in this part of the sampling procedure. Thus, when examining the different levels of sampling, it appears that most of the unexplained variation is between sites, implying the presence of local soil structures that the model in general and the included explanatory variables in particular cannot describe or capture adequately.

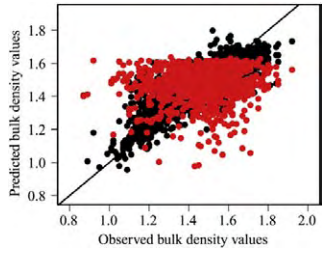

Fig. 5. Performance of soil bulk density prediction models based on test sub-dataset comparing model A (prediction showed in red colour) and mixed model (prediction showed in black colour).

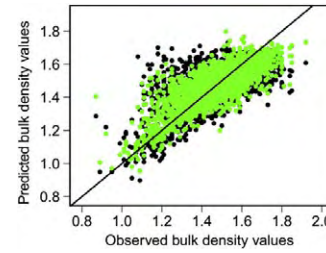

Fig. 6. Performance of soil bulk density prediction model based on test sub-dataset comparing mixed model (prediction showed in green) and mixed model without random effect (prediction showed in black colour).

SOC content is strongly and negatively correlated with D_b. This effect can be seen in all models considered, meaning that an increase in SOC leads to a decrease in D_b. This finding is in good agreement with previous studies (e.g. Benites et al., 2007; Bernoux et al., 1998; Heuscher et al., 2005). Tranter et al. (2007) suggested that this negative correlation is the result of soil aggregation. For the simple regression models, SOC describes approximately 1/4 of the variation in D_b estimation, but in the mixed model, its contribution is marginal. A plausible explanation may be that the mixed model incorporates various types of variation (explained or unexplained in terms of explanatory variables) that reduces the effect of SOC (e.g. effects are no longer hidden in the effect of SOC). Moreover, the effect of SOC is constant across the different soil textures, but it varies significantly with sampling depth and soil type. Contrary to our results from the mixed model approach, Kätterer et al. (2006) concluded that soil texture class modified the effect of SOC significantly; however, inclusion of texture class in addition to SOC in the topsoil D_b regression model decreased RMSE only from 0.13 to 0.12. Our study indicates that the simple regression model overestimates the importance of SOC in D_b variation.

Soil sampling depth describes most of the variation in D_b in the mixed model. In the simple regression model, its effect is smaller, but it still accounts for 1/5 of the variation. In most D_b prediction models described in the literature, sampling depth was not considered as a possible variable for predicting D_b. Some studies have concluded that sample depth can be used as a predictor, but that it describes a small part of the variability in D_b (Calhoun et al., 2001; De Vos et al., 2005; Heuscher et al., 2005). Heuscher et al. (2005) suggest that sampling depth can be used to predict oven-dry D_b (although it is not claimed to be a very strong predictor). They argue that because sampling depth is related to mechanical stress caused by overburdened soil, it could be expected to affect D_b. The main difference between previous studies and the one presented here is that previous studies have treated sampling depth as a continuous variable (Calhoun et al., 2001; De Vos et al., 2005; Heuscher et al., 2005; Tranter et al., 2007). Because most of the soil samples were taken at certain depths, it is not warranted to extend conclusions to arbitrary depths; therefore, we handled sampling depth as a categorical variable. Tranter et al. (2007), amongst others, find that at deeper sampling depths, D_b values increase. Our results agree with this finding; however, at a sampling depth of 40 cm, D_b starts to decrease. The highest D_b value (at a depth

of 25 cm) can be explained by the formation of a plough-pan layer (Alakukku et al., 2003; Hammel, 1994), and this is probably the reason why below that depth D_b is again slightly lower. Although Calhoun et al. (2001) concluded that genetic horizons contribute more to D_b prediction than does sampling depth, we suggest that sampling depth should be incorporated in any modelling strategy because previous studies have not examined one horizon in sufficient detail to predict D_b.

Soil water content is statistically significant in all of the developed models that we tested and is inversely related to D_b. Water content has rarely been used as an input variable in D_b prediction models, and to our knowledge, only studies based on data from the United States have used water content as a predictive factor (Heuscher et al., 2005; Manrique and Jones, 1991). These studies conclude, as does this study, that water content has a significant, decreasing effect on D_b. However, due to differences in D_b and Wc measurement methods, the results are not directly comparable. Heuscher et al. (2005), Manrique and Jones (1991) used a "coated clod" method for obtaining D_b measurements and Wc at −1500 kPa soil water potential. In reviewing the literature, no published PTFs were found for dry D_b determined by the core method and in which Wc at the time of sampling was included as a predictive variable. Several studies have emphasised that Wc at the time of measurement has a great impact on the resulting D_b values, and the results are comparable with soil sampled at standardised moisture content (Håkansson and Lipiec, 2000; Logsdon and Karlen, 2004). Under field conditions, and especially in extensive soil monitoring programmes, it is not practical to standardise soil water content. In our dataset, water content is not standardised, but inclusion of this variable (e.g. with a value corresponding to the field capacity) in proposed models ensures better comparability of predicted D_b values across the soils. Water content is not the strongest predictor of D_b, but it is still statistically significant and has effects that are modified differently by different levels of sampling depth, soil type, and texture. Water content could be expected to have an impact mainly on heavy-textured soils with high shrink–swell potential; however our results showed that it also had a decreasing effect on D_b for light- and medium-textured soils. As our study was restricted to the A-horizon, it can be assumed that, in addition to the clay particles, also the swelling and shrinkage of organic complexes (humus) also contributes significantly to D_b changes. Kuznetsova and Danilova (1989) concluded that soil loosening during swelling and shrinkage is also dependent on the content and quality of humus. Gray and Allbrook (2002) found that the majority of studies showed no significant relationship between SOC and shrinkage, and contrary results are seldom.

Various techniques can be used to construct prediction models. In this study, we used mixed model and multiple linear regression analysis. Two final models were chosen to compare two different approaches to predict D_b values and to provide results that are comparable with previous studies (because multiple linear regression has been the most common technique for developing D_b PTFs). The results indicate that D_b prediction is more accurate using the mixed models approach than it is using regression analysis. The latter tends to underestimate values above 1.6 g cm^{-3} (also found by Martin et al. (2009)) and to overestimate D_b values below 1.2 g cm^{-3}. This indicates that the model works within a very limited range. If we consider that in arable soils, D_b values can range much more widely, then this type of model is of limited practical use. However, models with fewer variables remain useful for cases in which limited data are available. It is also important to use the same variables as are present in the Estonian large-scale soil map. Compatibility with available soil map data allows us to use the developed models for the calculation of SOC stocks in arable Estonian soils.

Several studies that developed D_b prediction models have used RMSE as one measure of predictive performance (Benites et al., 2007; Heuscher et al., 2005; Kaur et al., 2002; Martin et al., 2009; Tranter et al., 2007). However, Bernoux et al. (1998), Calhoun et al. (2001), and Manrique and Jones (1991), did not validate the developed PTFs and simply examined the R^2 value, which is not an indicator of predictive performance. An important aspect of prediction models is the number of samples in the soil databases. Some models are based on smaller datasets (e.g. n = 200–2000) (Benites et al., 2007; Bernoux et al., 1998; Calhoun et al., 2001; De Vos et al., 2005; Kaur et al., 2002; Tranter et al., 2007) and some use larger datasets (Heuscher et al., 2005; Manrique and Jones, 1991). Additionally, the size of databases gives an indication of how many different locations and soil types have been accounted for. Some models work only for a very refined area and under certain conditions (or even within one geographic location). Model working area is also limited by the method used; multiple regression models, for instance, apply only at the locations where samples are taken and cannot be easily expanded; however, similar soil properties from different geographic regions are viable for use in an already developed PTF. Parameters included in modelling also vary, and often the methods used for analysing the samples differ. Comparing the predictive quality of these models shows that models based on smaller databases do not differ in predictive performance, RMSE within the range of 0.15 to 0.19 g cm^{-3}, and indeed the RMSE for the largest dataset (n = 47,015) is also 0.19 (Heuscher et al., 2005). In the current study, we used a dataset of 17,293 samples and a mixed model method that enabled us to extend the prediction to include all of Estonia, as the dataset incorporates many different and widespread arable soil conditions in Estonia. The results show a significant improvement in prediction, as the RMSE for the regression model A is 0.11 g cm^{-3}, and for the mixed model, it is 0.09 g cm^{-3} (based on the test dataset). This indicates that, in addition to database size, prediction error can be improved by using more sophisticated models that utilise more detailed soil information. Previous studies have used 4–6 parameters, but we used 20 in the mixed model and seven in model A. Considering the database structure, size, available parameters, modelling techniques, and prediction quality, we suggest using the mixed model to predict oven-dry D_b.

The methodology outlined in this paper for finding the best prediction model is general, and therefore it is also applicable to situations where 10–20 candidate models are initially considered. In this paper, we used a mixed model approach assuming independence between sampling sites because of their widespread distribution across Estonia. However, the mixed model methodology may be extended to allow kriging by additionally assuming a spatial correlation structure between sites (possibly stratified within regions) as pointed out by Lark et al. (2006). Such an extension would require coordinate information (GIS data) as correlations between sites must be calculated using some distance measure (metric).

Finally, the present study demonstrates the power of the mixed model methodology for analysing soil data by not only utilising the information captured in available explanatory variables (as does the multiple linear regression approach), but also capturing a substantial part of the random variation that is attributable to the different sources of variation often found in databases. This finding has two important implications. Firstly, it means that the use of mixed model depends crucially on identifying the hierarchical data structure present in the database under study. Some databases have similar structure to those used in this study. An example is the National Cooperative Soil Characterization Database, in which the hierarchical structure with soil survey areas nested into counties nested into states nested into countries as well as the pedon information (e.g. observation date) can be treated as random effects; land use and analysed soil properties are treated as fixed effects (National Cooperative Soil Survey, 2011). However, other databases may reveal different structures with layers that need to be addressed through mixed effects modelling (e.g. the North South Wales Soil and Land Information System (SALIS) in Australia, in which variation in site location, soil profile, and profile details could be accommodated by means of random effects; NSW Government, 2011). Thus, prior

understanding of the major components of variation expected in a database will substantially improve on the insights gained from the applied mixed model. Secondly, this study provides empirical underpinning of the fact that prediction of soil properties can indeed be improved through the use of kriging by exploiting the random variation captured by the predicted random effects (EBLUPs). This perspective will be explored in future work.

5. Conclusions

The results of this study are based on a large and comprehensive database. They demonstrate that D_b for the A-horizon of arable soils is mainly determined by sampling depth and SOC. Additionally, water content should be considered for prediction of D_b even in the case of light-textured soils. In general, to enable an accurate prediction of D_b, it is advisable to evaluate every horizon separately and to take into account as many systematic and random effects as the study design and database content permit. The mixed model showed that D_b is determined through several combined effects involving qualitative and quantitative explanatory variables, and that their contribution to the variation in D_b is different from evidence shown by simple regression models. We found that the regression model overestimated the importance of SOC. The results also underpin the fact that increased model complexity significantly improves the accuracy of D_b prediction. Thus, using advanced techniques for constructing prediction models should be considered. The mixed model approach enables incorporation of various types of variation sources and therefore produces a more accurate prediction model. Multiple regression analysis enables D_b to be predicted when few variables are available, but then the prediction error increases. Thus, we suggest using the mixed model approach when the required data are available and when the goal is to achieve maximum prediction accuracy, which should be the case when developing PTFs.

Acknowledgements

The study was supported by the Estonian Science Foundation grant no. 7622. We would like to thank the Estonian Agricultural Project for initiating and carrying out the National Soil Monitoring Programme since 1983. In addition, we thank Priit Penu and Tiina Köster from the Agricultural Research Centre for their fruitful co-operation.

References

Alakukku, L., Weisskopf, P., Chamen, W.C.T., Tijink, F.G.J., van der Linden, J.P., Pires, S., Sommer, C., Spoor, G., 2003. Prevention strategies for field traffic-induced subsoil compaction: a review: Part 1. Machine soil interactions. Soil and Tillage Research 73 (1–2), 145–160.

Bates, D., Maechler, M., 2010. lme4: Linear Mixed-effects Models using S4 Classes. R Package Version 0.999375-32. http://cran.r-project.org/package=lme4 2010.

Batey, T., 2009. Soil compaction and soil management – a review. Soil Use and Management 25 (4), 335–345.

Batjes, N.H., 1998. Mitigation of atmospheric CO_2 concentrations by increased carbon sequestration in the soil. Biology and Fertility of Soils 27 (3), 230–235.

Benites, V.M., Machado, P.L.O.A., Fidalgo, E.C.C., Coelho, M.R., Madari, B.E., 2007. Pedotransfer functions for estimating soil bulk density from existing soil survey reports in Brazil. Geoderma 139 (1–2), 90–97.

Bernoux, M., Arrouays, D., Cerri, C., Volkoff, B., Jolivet, C., 1998. Bulk densities of Brazilian Amazon soils related to other soil properties. Soil Sci Soc Am J 62, 743–749.

Calhoun, F.G., Smeck, N.E., Slater, B.L., Bigham, J.M., Hall, G.F., 2001. Predicting bulk density of Ohio soils from Morphology, Genetic Principles, and Laboratory Characterization data. Soil Sci Soc Am J 65, 811–819.

De Vos, B., Van Meirvenne, M., Quataert, P., Deckers, J., Muys, B., 2005. Predictive quality of pedotransfer functions for estimating bulk density of forest soils. Soil Sci Soc Am J 69 (2), 500–510.

Goidts, E., van Weselmael, B., Van Oost, K., 2009. Driving forces of soil organic carbon evolution at the landscape and regional scale using data from a stratified soil monitoring. Global Change Biology 15 (12), 2981–3000.

Gray, C.W., Allbrook, C., 2002. Relationships between shrinkage indices and soil properties in some New Zealand soils. Geoderma 108 (3–4), 287–299.

Håkansson, I., Lipiec, J., 2000. A review of the usefulness of relative bulk density values in studies of soil structure and compaction. Soil and Tillage Research 53, 71–85.

Hammel, J., 1994. Effect of high-axle load traffic on subsoil physical properties and crop yields in the Pacific Northwest USA. Soil and Tillage Research 29 (2–3), 195–203.

Harrison, A.F., Bocock, K.L., 1981. Estimation of soil bulk-density from loss-on-ignition values. J Appl Ecol 18 (3), 919–927.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Springer, New York.

Heuscher, S.A., Brandt, C.C., Jardine, P.M., 2005. Using soil physical and chemical properties to estimate bulk density. Soil Sci Soc Am J 69 (1), 51–56.

Huntington, T.G., Johnson, Ś.E., Johnson, A.H., Siccama, T.G., Ryan, D.F., 1989. Carbon, organic matter, and bulk density relationships in a forested spodosol. Soil Sci 148 (5), 380–386.

Katschinski, N.A., 1956. Die mechanische Bodenanalyse und die Klassifikation der Böden nach ihrer mechanischen Zusammensetzung. Rapports au Sixiéme Congrés de la Science du Sol, Paris, B, pp. 321–327.

Kätterer, T., Andrén, O., Jansson, P.E., 2006. Pedotransfer functions for estimating plant available water and bulk density in Swedish agricultural soils. Acta Agriculturae Scandinavica, Section B - Soil & Plant Science 56 (4), 263–276.

Kaur, R., Kumar, S., Gurung, H.P., 2002. A pedo-transfer function (PTF) for estimating soil bulk density from basic soil data and its comparison with existing PTFs. Australian Journal of Soil Research 40, 847–857.

Kokk, R., Rooma, I., Valler, V., Reintam, L., 1968. Field Works Instructions for Large Scale Mapping of Soils. EPA, Tartu. [in Estonian].

Kuznetsova, I.V., Danilova, V.I., 1989. Loosening of soils by swelling and shrinkage. Soviet Soil Science 20 (6), 108–120.

Lampurlanés, J., Cantero-Martinez, C., 2003. Soil bulk density and penetration resistance under different tillage and crop management systems and their relationship with barley root growth. Agron J 95 (3), 526–536.

Lark, R.M., Cullis, B.R., 2004. Model-based analysis using REML for inference from systematically sampled data on soil. European Journal of Soil Science 55, 799–813.

Lark, R.M., Cullis, B.R., Welham, S.J., 2006. On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML. European Journal of Soil Science 57 (6), 787–799.

Lauringson, E., 2003. The Effect of Agrotechnical Methods on Soil Physical Properties, Weediness and Crop Yield. D.Sc. Thesis. Estonian Agricultural University, Tartu.

Logsdon, S.D., Karlen, D.L., 2004. Bulk density as a soil quality indicator during conversion to no-tillage. Soil and Tillage Research 78, 143–149.

Manrique, L.A., Jones, C.A., 1991. Bulk density of soils in relation to soil physical and chemical properties. Soil Sci Soc Am J 55, 476–481.

Martin, M.P., Lo Seen, D., Boulonne, L., Jolivet, C., Nair, K.M., Bourgeon, G., Arrouays, D., 2009. Optimizing pedotransfer functions for estimating soil bulk density using boosted regression trees. Soil Sci Soc Am J 73, 485.

McBratney, A.B., Minasny, B., Cattle, S.R., Vervoort, R.W., 2002. From pedotransfer functions to soil inference systems. Geoderma 109 (1–2), 41–73.

Meersmans, J., van Wesemael, B., De Ridder, F., Van Molle, M., 2009. Modelling the three-dimensional spatial distribution of soil organic carbon (SOC) at the regional scale (Flanders, Belgium). Geoderma 152 (1–2), 43–52.

National Cooperative Soil Survey, 2011. National Cooperative Soil Characterization Database. Available online at http://ssldata.nrcs.usda.gov. Last accessed: 29.03.2011.

NSW Government, 2011. NSW Natural Resource Atlas website, The NSW Soil and Land Information System. Available online at http://nratlas.nsw.gov.au/. Last accessed: 29.03.2011.

Prevost, M., 2004. Predicting soil properties from organic matter content following mechanical site preparation of forest soils. Soil Sci Soc Am J 68, 943–949.

R Development Core Team, 2010. R: A Language and Environment for Statistical Computing. ISBN 3-900051-07-0 R Foundation for Statistical Computing, Vienna, Austria URL:http://www.r-project.org.

Reintam, L., Kull, A., Palang, H., Rooma, I., 2003. Large-scale soil maps and a supplementary database for land use planning in Estonia. J Plant Nutr Soil Sci 166 (2), 225–231.

Reintam, E., Trukmann, K., Kuht, J., Toomsoo, A., Teesalu, T., Koster, T., Edesi, L., Nugis, E., 2008. Effect of *Cirsium arvense* L. on soil physical properties and crop growth. Agricultural and Food Science 17 (2), 153–164.

Smith, P., 2004. Soils as carbon sinks – the global context. Soil Use and Management 20, 212–218.

Steyerberg, E.W., 2009. Clinical Prediction Models. Springer, New York.

Teras, T., 1987. Soil Bulk Density in Estonian Arable Soils. Estonian Soils in Figures VI, Tallinn, pp. 17–28. [in Estonian].

Tomasella, J., Hodnett, M.G., 1998. Estimating soil water retention characteristics from limited data in Brazilian Amazonia. Soil Sci 163, 190–202.

Tranter, G., Minasny, B., McBratney, A.B., Murphy, B., McKenzie, N.J., Grundy, M., Brough, D., 2007. Building and testing conceptual and empirical models for predicting soil bulk density. Soil Use and Management 23 (4), 437–443.

Vorobeva, L.A., 1998. 272 pp Chemical Analysis of Soils. Moscow University Press, Moscow. [in Russian].

Walkley, A., Black, I.A., 1934. An examination of the Degtjareff Method for determining soil organic matter, and a proposed modification of the chromic acid titration method. Soil Sci 37 (1).

WRB, 2006. World reference base for soil resources 2006. World Soil Resources Reports No. 103. FAO, Rome.

II

Suuster, E., Ritz, C., Roostalu, H., Kõlli, R. & Astover, A. 2012. Modelling soil organic carbon concentration of mineral soils in arable land using legacy soil data.

European Journal of Soil Science, 63, 351–359.

European Journal of Soil Science

European Journal of Soil Science, June 2012, **63**, 351–359 doi: 10.1111/j.1365-2389.2012.01440.x

Modelling soil organic carbon concentration of mineral soils in arable land using legacy soil data

E. Suuster[a], C. Ritz[b], H. Roostalu[a], R. Kõlli[a] & A. Astover[a]

[a]*Department of Soil Science and Agrochemistry, Institute of Agricultural and Environmental Sciences, Estonian University of Life Sciences, Kreutzwaldi 1a, Tartu 51014, Estonia, and* [b]*Faculty of Life Sciences, Department of Basic Sciences and Environment, University of Copenhagen, Thorvaldsensvej 40, DK-1871 Frederiksberg C, Denmark*

Summary

Soil organic carbon (SOC) concentration is an essential factor in biomass production and soil functioning. SOC concentration values are often obtained by prediction but the prediction accuracy depends much on the method used. Currently, there is a lack of evidence in the soil science literature as to the advantages and shortcomings of the different commonly used prediction methods. Therefore, we compared and evaluated the merits of the median approach, analysis of covariance, mixed models and random forests in the context of prediction of SOC concentrations of mineral soils under arable management in the A-horizon. Three soil properties were used in all of the developed models: soil type, physical clay content (particle size <0.01 mm) and A-horizon thickness. We found that the mixed model predicted SOC concentrations with the smallest mean squared error ($0.05\%^2$), suggesting that a mixed-model approach is appropriate if the study design has a hierarchical structure as in our scenario. We used the Estonian National Soil Monitoring data on arable lands to predict SOC concentrations of mineral soils. Subsequently, the model with the best prediction accuracy was applied to the Estonian digital soil map for the case study area of Tartu County where the SOC predictions ranged from 0.6 to 4.8%. Our study indicates that predictions using legacy soil maps can be used in national inventories and for up-scaling estimates of carbon concentrations from county to country scales.

Introduction

Soil organic carbon (SOC) determines soil-type-specific soil fertility, functioning, physical and chemical properties. Also, soils represent the largest terrestrial carbon pool (Batjes, 1996). Thus, the status and spatio-temporal changes in SOC concentration and stocks in different land-use systems have been examined with regard to soils' potential to sequester carbon (Smith, 2008). The potential for additional carbon sequestration in agricultural systems has been argued to be small (Powlson *et al*., 2011) and, similarly, favouring conservation tillage over no-till systems only yields a small gain in sequestration (Baker *et al*., 2007). Apart from carbon sequestration in soil, the real focus should remain on SOC being essential for soil functioning and the basis for biomass production. Thus, nationwide predictions of SOC concentrations are still necessary to convert predictions from local to global scale. However, the precision of predictions is dependent on the model chosen.

Gray *et al*. (2009) compared categorical median, multiple linear regression and decision tree approaches to develop quantitative relationships between soil properties and environmental factors. Meersmans *et al*. (2008) used mean and multiple regression methods to predict and assess the spatial distribution of SOC stocks from soil texture, moisture and land use. Gray *et al*. (2009) used root mean squared error and mean error to prove their predictions to be broadly equivalent between the two approaches, and Meersmans *et al*. (2008) assessed their results by standard errors. Thus, a more comprehensive comparison of available prediction methods, using both the knowledge and experience from soil science, would be desirable.

Mixed models have not been used widely in predicting SOC concentration, but they have been used for modelling SOC stocks (Goidts *et al*., 2009; Maia *et al*., 2010) and other soil properties (Haskard *et al*., 2010; Suuster *et al*., 2011). The mixed model is an extension of analysis of covariance (ANCOVA) where random effects are included to describe hierarchical structures in the data. The random forests approach, a black box model, is also an effective tool in predicting SOC concentrations (Grimm *et al*., 2008; Wiesmeier *et al*., 2011). These modern methods, as well as the more traditional median and ANCOVA approaches, were considered in the current study.

Prediction models are usually derived from legacy soil data and can be applied to legacy soil maps, adding value through the

Correspondence: E. Suuster. E-mail: elsa.suuster@emu.ee

Received 20 October 2011; revised version accepted 7 February 2012

© 2012 The Authors
Journal compilation © 2012 British Society of Soil Science

predicted SOC concentrations (Bell & Worrall, 2009). Legacy soil data from soil monitoring networks are suitable for understanding soil properties and their spatio-temporal changes, including SOC concentration, if they are well designed, including, for example, replications within different land-use types and with resampling after 5–10 years (van Wesemael *et al*., 2011). Data from the Estonian National Soil Monitoring on arable lands have been analysed and published mainly at the level of descriptive statistics for general soil groups (Kokk & Rooma, 1983). The studies in Estonia have mainly focused on humus quality, composition of soil organic matter, SOC distribution in soil types, and sources and fluxes of SOC (Kask, 1975; Kõlli & Ellermäe, 2003; Kõlli *et al*., 2010). Additionally, the Estonian soil map provides input data for SOC concentration modelling (Reintam *et al*., 2003). Nevertheless, previous attempts to estimate SOC concentration in Estonia did not consider spatial distribution, and in general little work has been done on predictive modelling for SOC concentration.

This study seeks to compare some of the methods used in the past and some new methods in order to evaluate differences in accuracy of prediction of SOC concentration. Moreover, we want to demonstrate in a case study how the best prediction method augments our knowledge about the spatial distribution of SOC concentrations, thus providing added value to the existing Estonian legacy soil data.

Materials and methods

Study design

A soil monitoring network of arable land was established in Estonia in 1983 and it ran until 1994 and involved 79 different sites. In 2002 the monitoring of soil quality observations was resumed at 21 sites. We used data collected during 1983 to 2008 from 90 sites to develop prediction models for SOC concentrations in mineral soils in the A-horizon. The samples and data from the period 1983–1994 are archived at the Estonian Soil Museum at the Estonian University of Life Sciences and the samples from 2002 to 2008 at the Agricultural Research Centre. The monitoring sites were mainly under cereal-based crop rotations with ploughing in the autumn. Each site comprised four transects and within each transect 10 observation plots were established.

The dataset contains 8697 entries representing the A-horizon and all major soil types found on arable land. At every plot A-horizon thickness, SOC concentration, texture and physical clay (Cf; particle size <0.01 mm) content (%) were measured. A-horizon thickness ranged from 12 to 100 cm. All chemical soil analyses were carried out on the fine-earth fraction (<1 mm). SOC concentration was determined by wet digestion with acid potassium dichromate solution (Vorobeva, 1998). No soil-type-specific correction factor was applied because there was large soil type variation and general correction factors for incomplete oxidation are not recommended (Meersmans *et al*., 2009a). The established Estonian Soil Classification (ESC) range for SOC concentrations in the A-horizon is 0.6–6%. The particle-size distribution was determined using the pipette method, whereas the Cf was defined after Katschinski (1956) as the mass fraction of particles smaller than 0.01 mm, which as well as 'clay' particles also includes particles of 'fine silt'. Texture classes were sand, loamy sand, light loam, medium loam, heavy loam and clay. The soils were classified according to ESC; where approximate soil groups are given, classifications were also based on the World Reference Base (WRB) for Soil Resources (IUSS Working Group WRB, 2006).

Prediction models

In this study we focus on predicting SOC concentrations by using four different approaches: (i) robust prediction by using medians, (ii) ANCOVA, (iii) mixed-model analysis and (iv) random forests. Each of these approaches is described in more detail below.

Any statistical model is fitted to data in such a way that discrepancies between the data and the model become as small as possible (according to some criterion such as least squares). However, there may be a problem if we want to use the fitted model for prediction when using another dataset: the fitted model may simply follow the original data too closely to be useful for prediction using other datasets with slightly different features. This phenomenon is referred to as over-fitting. To avoid such overly optimistic results when evaluating a prediction model, the standard approach is to train the model on one dataset and to evaluate it on another independent test set. Where a number of models are initially considered (as in our case) a third validation dataset, which is independent of the training and test datasets, is therefore often needed for the intermediate step of determining the best model among the candidates. In this way different datasets are used for selecting and reporting the performance of the best approach, avoiding overly optimistic conclusions about the prediction accuracy. Following Hastie *et al*. (2009), we divided our database randomly into three parts, a training dataset (50%), a validation dataset (25%) and a test dataset (25%), because the entire dataset was large enough to ensure that each subset is representative.

Median approach

In order to apply the median approach, the training dataset is stratified according to the chosen explanatory variables, which are grouping factors such as soil type and texture. For each combination of soil type and texture the median value of the soil property is calculated from the individual measurements in each stratum of soil. For prediction, the median values are assigned to corresponding polygons/grid cells on the soil maps according to the configurations of the explanatory variables in those. Because of the coarse grouping used for obtaining the predictions, data from several map types (soil, land use) may be combined (Meersmans *et al*., 2008).

The main advantages of the median approach are ease of application and suitability for situations where limited soil information is available, and the main disadvantage is that stratification may lead to groups containing few or no data at all.

© 2012 The Authors
Journal compilation © 2012 British Society of Soil Science,

ANCOVA approach

In a simple linear regression model with one explanatory variable, there are two parameters: an intercept and a slope. In practice such linear relationships are often not the same under different conditions and therefore there is often a need to assume different intercepts or slopes for different conditions. The resulting model, analysis of covariance (ANCOVA), is a way to describe how the co-variation of explanatory variables influences the response.

The ANCOVA approach extends the multiple linear regression methodology to the situation where additional categorical explanatory variables are available. Some of the advantages of the ANCOVA approach are its flexibility, interpretability of parameters in terms of changes in the values of the explanatory variables and availability in major statistical software packages (Vittinghoff *et al*., 2005). The ANCOVA approach possesses one major weakness, that the more explanatory variables that are included in the model the greater is the risk of over-fitting. The additional variables may improve the fit (reduce the bias) by capturing minor and unimportant features that may not be present in other datasets, thus increasing the variance. For the same reason, the coefficient of determination, R^2, is not a good measure for assessing the predictive ability as it increases every time another explanatory variable is included in the model (Vittinghoff *et al*., 2005). Instead a measure based on the prediction error for a test dataset has to be used. As the ANCOVA model in our context is intended for prediction alone, model checking is reduced to checking whether or not there are single observations or subsets of observations that are not appropriately described by the model (as such observations may result in biased estimates and consequently biased predictions). Whether or not the distributional assumptions are fulfilled is not important at all.

We consider an ANCOVA model that involves three explanatory variables, Cf content (quantitative), soil type (categorical) and A-horizon thickness (quantitative), resulting in a linear relationship between SOC and Cf content with intercepts and slopes that depend on soil type, and also assume a linear relationship between SOC concentration and A-horizon thickness.

Similar ANCOVA approaches have been used to predict the distribution of SOC mass density by depth with land-use and soil type as explanatory variables (Meersmans *et al*., 2009b) and to predict general SOC mass density for each land-use–soil type combination (Meersmans *et al*., 2008). Bell & Worrall (2009) used an ANCOVA model to determine the forms of stratification (soil series, land use and soil group) that could improve the estimates of baseline SOC concentrations.

Mixed-model approach

The Estonian National Soil Monitoring data are hierarchically structured with plots nested within transects, which are again nested within sites and most sites have been visited repeatedly over several years: therefore a mixed-model approach can be considered. More specifically, the mixed model contains the fixed effects of soil type, Cf content and A-horizon thickness. Because of the hierarchical structure in the data, we included site, transect, plot and year as random effects. Additionally, year-specific effects along a transect were also included, as were site-specific slope coefficients for Cf content as the relationship between SOC and Cf content was expected to vary from site to site.

The mixed model formulation leads to a partition of the total random variation seen in the SOC measurements into a number of components that reflect contributions from various sources of systematic sampling imposed by the study design. For a new site the mixed model yields the following predicted SOC concentration (which we denote as y_{new}):

$$y_{new} = f(x_{new}) + B_{site} + B_{site,\ Cfcontent} \times Cf\ content_{new} + B_{plot} + B_{transect-year} + B_{transect} + B_{year} \quad (1)$$

where the fixed-effects component is denoted $f(x_{new})$ and all the random effects are denoted by B with subscripts. The resulting partitioning of the variance of the predicted SOC concentration is:

$$var(y_{new}) = \sigma^2_{site} + \sigma^2_{site,\ Cfcontent} \times Cf\ content^2_{new} + \sigma^2_{plot} + \sigma^2_{transect-year} + \sigma^2_{transect} + \sigma^2_{year} + \sigma^2_{residual} \quad (2)$$

where $Cf\ content^2_{new}$ denotes the square of the Cf content at the new site. As all random effects are assumed to be unknown, they all contribute to the uncertainty of the prediction (together with the residual error).

Similarly, the predicted SOC concentration for a known site, that is a site present in the training data, may be written as:

$$y_{new} = f(x_{new}) + b_{site} + b_{site,\ Cfcontent} \times Cf\ content_{old} + B_{plot} + B_{transect-year} + B_{transect} + B_{year}, \quad (3)$$

where lower-case b denotes the estimated site effects (Empirical Best Linear Unbiased Predictors - EBLUPs), which are obtained from the training data and $Cf\ content_{old}$ denotes the Cf content of the site, which is also available from the training dataset. The site-specific EBLUPs are averages per site that have been shrunk towards 0 depending on the amount of variation observed among the sites. By assuming that the site effect is known, the site-specific variation vanishes in the corresponding variance:

$$var(y_{new}) = 0 + 0 + \sigma^2_{plot} + \sigma^2_{transect-year} + \sigma^2_{transect} + \sigma^2_{year} + \sigma^2_{residual}. \quad (4)$$

In other words, the components σ^2_{site} and $\sigma^2_{site,\ Cf\ content}$ are now set to zero in the partitioning of the variance, reflecting that the prediction exploits the information available (in the training data) concerning this particular site. Consequently, the prediction is biased away from the overall population (corresponding to EBLUPs equal to 0) and towards the specific site of interest and the prediction becomes more accurate. At the same time the variance is reduced, yielding a more precise prediction (Hastie *et al*., 2009).We note

© 2012 The Authors

that by taking the above argument one step further we find that the minimum mean squared error (MSE) achievable is equal to $\sigma^2_{\text{residual}}$, which would occur in the unrealistic case where plot, site, transect and year effects were known but not the residual error attached to each individual soil sample.

A related approach is to use neighbouring sites for the prediction at a new site:

$$y_{\text{new}} = f(x_{\text{new}}) + (\Sigma_{i=1}^{n} b_{\text{site},i} + b_{\text{site, Cf content},\,i} \times Cf\ content_{\text{old},i})/n + B_{\text{plot}} + B_{\text{transect}-\text{year}} + B_{\text{transect}} + B_{\text{year}} \quad (5)$$

where the EBLUPs of n neighbouring sites, denoted $b_{\text{site},i}, \ldots, b_{\text{site},n}$ and $b_{\text{site, Cf content},\,i}, \ldots, b_{\text{site, Cf content},\,n}$, have been averaged. The improvement in the prediction achieved by adding the average of the EBLUPs (adding a bias) will depend crucially on the neighbouring sites being relatively similar to the new site; otherwise the accuracy of the prediction is not improved. This approach may only be reasonable for geographical regions that are relatively small, such as Tartu County.

Random forests approach

The random forests approach improves the prediction accuracy of regression tree methodology by reducing the variance without increasing the bias too much. By averaging a large number of regression trees that are generated from the training data, a random forest is constructed. For each regression tree only a subset of the explanatory variables is considered at each division (Liaw & Wiener, 2002). The generated regression trees will involve different combinations of explanatory variables, but in general none of these will provide better prediction models than that based on the entire training data. However, each individual tree may still capture some particular or unique feature in the data and therefore each of these trees is said to be a weak learner. By combining the generated trees into a weighted fit we can predict many of the features in the data accurately as a consequence of the accumulated knowledge. The random selection of subsets of explanatory variables provides a convenient way to handle datasets with many variables without having to decide on which to include or omit. The random forests approach is ideal for constructing prediction models from large datasets with many explanatory variables. In contrast, for the ANCOVA and to some extent the mixed model approaches, the handling of many explanatory variables often requires careful choice.

The random forests approach results in a smaller prediction error (reduced variance) and much less over-fitting (Breiman, 2001). The price paid for the improved prediction is that the relationship between the response and explanatory variables is difficult to interpret because the approach is a black-box one, and it is difficult to link the predicted values to the specific values or configurations of explanatory variables. The relative importance of the explanatory variables included may be assessed by using variable importance measures. For each explanatory variable the importance measure is obtained by averaging tree-specific differences between MSE for the original model and for the model where the values of the explanatory variable have been randomly reshuffled, thus destroying the effect of the explanatory variable (Liaw & Wiener, 2002). Analysis of variance was used to provide additional insight into how the variation is distributed across the explanatory variables in ANCOVA and the mixed-model approach.

The random forests approach has been used previously in studies for predicting the spatial distribution of SOC (Grimm *et al*., 2008; Wiesmeier *et al*., 2011) and is a powerful tool. We used soil type, Cf content, A-horizon thickness and year as explanatory variables in a random forests approach.

Software

All prediction models were fitted using the statistical environment R version 2.13.1 (R Development Core Team, 2011). The extension package lme4 was used for the mixed-model analysis (Bates *et al*., 2011) and the extension package randomForest for the random forests approach (Liaw & Wiener, 2002).

The Estonian large-scale soil map and area of case study

The soil map of Estonia is based on a systematic large-scale mapping of soil cover that took place from 1954 to 1992. The soil map was digitized in 2001 at the scale of 1:10 000 and is the property of the Estonian Land Board (2012). The soil map covers the area of 43 300 km^2, which is the whole territory of Estonia excluding towns and settlements. The Estonian soil map database contains information about the soil type (after ESC), soil texture, A-horizon thickness and stoniness. We considered the specific case of Tartu County in southeast Estonia as an example for predicting SOC concentrations from the model with the most accurate results. Tartu County covers 6.9% of Estonia's land surface and 35% of its area is in agricultural use (Astover *et al*., 2010).

The soil map covers the entire county, including both arable and forest regions. Thus, as a first step we used a field layer (1:5000–1:10 000) derived from the Agricultural Registers and Information Board (ARIB) to distinguish arable areas from others on the digital soil map as the prediction models were developed for mineral arable soils only. The soil data on arable areas from the modified soil map (35% of the county) provided the basis for predicting SOC concentrations for soil-type-based polygons. Secondly, the model with the greatest prediction accuracy was applied to the soil map data, whereby the site-specific effects were averaged based on neighbouring sites, as described previously. As the soil database contained soil texture classes, not Cf content, we used a mean approach to get approximate Cf content values (sand, 7.5%; loamy sand, 15%; light loam, 25%; medium loam, 35%; heavy loam, 45%; clay, 60%). To illustrate the predictions by means of a generalized thematic map we (i) calculated the weighted averages of predicted SOC concentrations based on a grid (50 × 50 m^2) and (ii) created a grid thematic map produced by interpolating the weighted averages of the predicted SOC

© 2012 The Authors

(inverse distance weighting interpolator). The GIS analysis was performed using MapInfo Professional (version 9.5).

Results

Gleysols and gleyic soils (G, Kg, Kog, KIg and LPg) had larger SOC concentrations than soils without reducing conditions (Kr, K, Ko, KI, LP and Lk) and those formed on carbonate-rich parent material (Kr, K, Ko and KI) had greater SOC concentrations than carbonate-free soils (LP and Lk) (Figure 1).

Comparison of prediction models

The results reported here were obtained for validation data (see Table 1 for additional results for the training data). The median approach resulted in the poorest prediction accuracy (Table 1). Additionally, the observed and predicted values were not in good agreement; the range of predicted values was between 0.75 and 4.07% (Figure 2a). The prediction accuracy for the ANCOVA was better than the median approach with an MSE of $0.23\%^2$ (see also Figure 2b). Soil type explained 91% of the variation in SOC concentration, whereas Cf content, A-horizon thickness and soil type–Cf interaction contributed in total 2%, leaving 6% unexplained.

The random forests approach was successful with an MSE of $0.17\%^2$, similar to the ANCOVA results (Figure 2c). The measure of the importance of explanatory variables revealed soil type as the most important in explaining the variation in SOC concentration distribution. Although the MSE for the random forests model was smaller than the median and ANCOVA approaches (Table 1), the visual agreement between observed and predicted SOC concentrations was similar for the three approaches (Figure 2a–c). Predictions for SOC concentrations larger than 4% were not achievable by these methods (Figure 2a–c) as the predicted values underestimated the observed values.

Inclusion of random effects in the model substantially increased the prediction accuracy, producing the smallest MSE for the mixed model from all candidate models (Table 1; Figure 2d). Again, soil type was dominant in explaining variation in SOC concentration (84%). Thus, the contributions of A-horizon thickness and soil type–Cf content interaction are marginal (9.7 and 6.7%, respectively) in explaining variation in SOC concentration. Among the random effects most of the variation was from the sites (49%), whereas the plots were responsible for 18%: the remaining effects each contributed less than 10% of the variation, except for the residual variation (within each survey plot), which was slightly larger (12%). Thus, the variation between the sites provided the largest contribution to the random variation

The results of visual assessment of observed and predicted SOC concentrations indicate a two-group pattern for all relationships in Figure 2. SOC concentrations greater than 3% were more scattered around the line whereas smaller values were more densely grouped. These larger SOC concentrations are usually for the hydromorphic soils such as gleyic soils and gleysols.

The model with the greatest prediction accuracy

The validation of the four models (Table 1) indicated that the mixed model had the greatest prediction accuracy. Thus, the mixed model was used for obtaining predictions for the test dataset, resulting in an MSE of $0.05\%^2$. Observed values agreed well with predictions (Figure 3), although the spread around the one-to-one line was slightly larger than results from the mixed model (Figure 2d): some observations for the former deviated markedly from the line. The deviating observations were mainly from Gleysols with observed SOC concentrations of more than

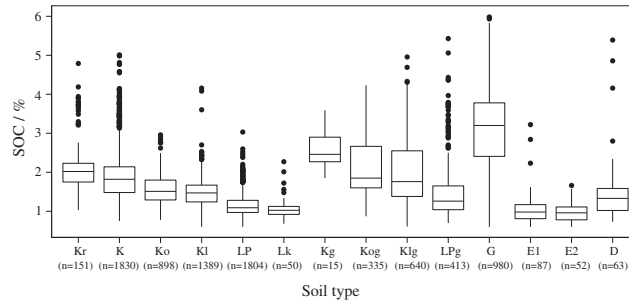

Figure 1 Variation in SOC concentration (%) by soil type in the A-horizon according to Estonian Soil Classification: D = *Colluvic Regosol (Cumulihumic)*[a]; E1 = *Haplic Cambisol*, *Cutanic Luvisol*, *Umbric Albeluvisol slightly eroded*; E2 = *Aric Regosol (Eutric) moderately eroded*; G = *Luvic Mollic Spodic Gleysol*; K = *Haplic Cambisol and Regosol*; KI = *Cutanic Luvisol*; KIg = *Endogleyic Cutanic Luvisol (Humic)*; Ko = *Haplic Phaeozem*; Kog = *Endogleyic Phaeozem*; Kr = *Haplic Regosol*; Lk = *Umbric Carbic Podzol*; LP = *Umbric Stagnic Albeluvisol*; LPg = *Umbric Stagnic Endogleyic Albeluvisol.* [a]Soil or soil association by World Reference Base (IUSS Working Group WRB, 2006). n = number of entries in the database.

© 2012 The Authors

Table 1 Comparison of the four candidate modelling approaches by means of mean error (ME, %), mean squared error (MSE, $\%^2$) and root mean squared error (RMSE, %) for the training and validation datasets

	Training dataset			Validation dataset		
Prediction approach	ME	MSE	RMSE	ME	MSE	RMSE
Median	0.07	0.29	0.54	0.10	0.32	0.56
ANCOVA	-1.59×10^{-18}	0.22	0.47	0.02	0.23	0.48
Random forests	0.001	0.09	0.30	0.03	0.17	0.42
Mixed model	-6.50×10^{-7}	0.01	0.09	0.01	0.06	0.24

3.7%. The actual range of predictions was 0.61–4.88%; thus, the mixed-model approach was capable of predicting larger SOC concentrations than the three other approaches. The estimated standard deviations also indicate the minimum precision that can be achieved for predictions as follows: site, 0.65%; plot, 0.22%; both transect and transect-specific year, 0.08%; year, 0.05%; and site-specific slope coefficient for Cf content, 0.02 (the latter quantifies the site-specific variation around the estimated slope of 0.0003 of the linear relationship between SOC concentration and Cf content). Thus, the largest gain in prediction accuracy can be achieved if site information is available.

Model application: a case study of Tartu County

The mixed-model approach as described above was applied to the data for Tartu County. The map with the predicted SOC concentrations (weighted averages) on mineral arable soils shows the local differences (Figure 4) (the white spots on the map

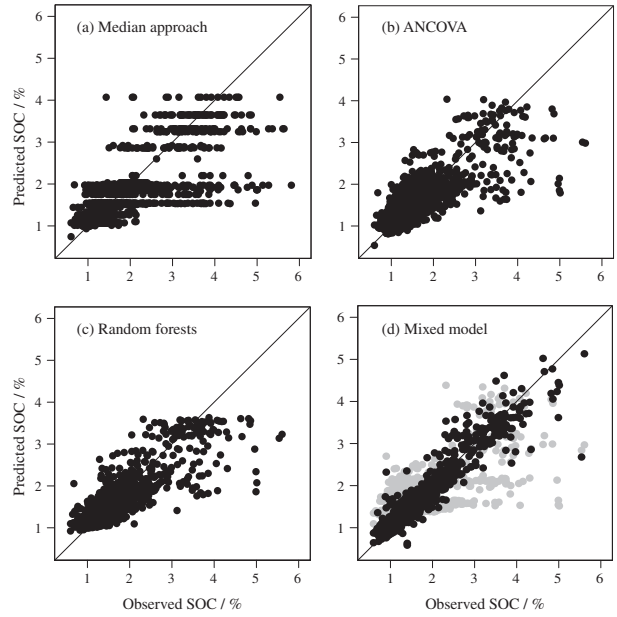

Figure 2 Observed and predicted SOC concentrations are displayed for the four approaches considered, the median approach (a), ANCOVA (b), random forests approach (c) and the mixed model (d), by comparing observed and predicted SOC (%) values for the validation dataset. The grey dots show the mixed model predictions obtained without using EBLUPs: the diagonal line is the 1:1 line.

© 2012 The Authors
Journal compilation © 2012 British Society of Soil Science,

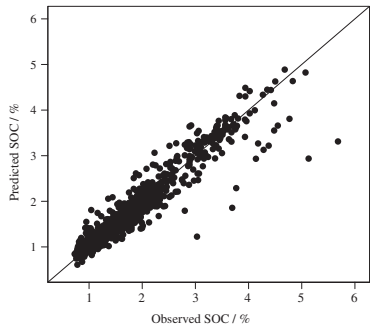

Figure 3 Performance of the best performing SOC prediction model (the mixed model) based on test dataset: the diagonal line is the 1:1 line.

correspond to towns and cities and other areas with no agricultural activity). The predicted SOC concentrations vary substantially in Tartu, from 0.6 to 4.8%, indicating large soil heterogeneity. Most of the arable soils in Tartu County have SOC concentration in the 0.8–1.2% class, especially in the south of the region, where the dominant soil type is LP (*Umbric Stagnic Albeluvisol*). The occurrence of SOC concentrations above 2% was very rare and reflected the absence of gleyic soils.

Discussion

Comparison of models

Prediction models for SOC concentration in mineral soils of arable land with conventional tillage that are based on legacy soil data are scarce in the literature. The mixed-model approach that we used had the best prediction accuracy, because it was able to use

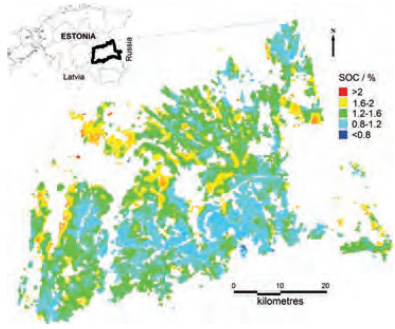

Figure 4 Mixed-model-based predicted SOC concentrations of mineral soils in the arable areas in Tartu County.

the hierarchical data structure. Among the models not exploiting the hierarchical structure, the random forests approach provided much greater prediction accuracy than the median or ANCOVA approaches. Thus, the random forests model should be preferred over median and ANCOVA models even in cases where a few explanatory variables are available. As the explanatory variables used in this study, particularly Cf content and soil type, are universal and readily available in most soil monitoring networks, soil databases and legacy soil maps in the world (Arrouays *et al*., 2008), our findings may not be limited only to the database that we have used. Suuster *et al*. (2011) obtained a similar result when comparing mixed models and ANCOVA models and this may be explained by the fact that the mixed model approach shrinks fixed-effects estimates towards zero and thus reduces the influence of explanatory variables that are not contributing much to the prediction (Vittinghoff *et al*., 2005).

We found large differences in prediction accuracy between methods, as did Meersmans *et al*. (2008). On the other hand, Gray *et al*. (2009) concluded that multiple linear regression, decision trees and the median approach give broadly the same predictions (the MSE was largest for the median approach and smallest for the multiple linear regression). The different results, however, in part result from differences in input variables: Gray *et al*. (2009) used environmental factors (climate, parent material and topography), while Meersmans *et al*. (2008) used soil texture, drainage and land use information.

In a similar way to Gray *et al*. (2009), we found that the median approach gave the least satisfying results for SOC concentration prediction. The MSE for the median approach was the largest among those considered and the predictions were only within a limited range (0.75–4.07%) for the median approach. However, no pattern in the predictions could be detected when the approach over- or underestimated SOC concentrations. Thus, ease of application and lack of assumptions on the SOC distribution when using medians may result in unsatisfactory SOC concentration predictions. The reason for the poor fits may be the complex data structure, which was not captured entirely. In contrast, the mixed model coped better with the variation imposed by sampling design. Therefore, one of the major benefits of using mixed-model procedures for predicting SOC concentrations from a soil monitoring network is the ability to reflect hierarchical data structures through random effects. Moreover, dependencies in the data, which can be used to link training and validation/test data, can be exploited in the mixed model setting to improve accuracy. Likewise, geographical information can be incorporated in the mixed model through spatial correlation and will introduce a dependency structure that extends from the training data to the validation/test data. Kriging is one way to use the spatial correlation structure across sites (Lark *et al*., 2006; Minasny & McBratney, 2007; Kempen *et al*., 2010). The mixed-model approach presented in this study could be further improved by model-based kriging (Mishra *et al*., 2009, 2010). This option will be addressed in a future study.

We now try to understand the underlying soil-specific mechanisms exploited in the models. The different methods

© 2012 The Authors

included almost the same range of explanatory variables of soil type, Cf content, soil type–Cf content interaction and A-horizon thickness. Soil type has often been used in SOC concentration prediction as it encompasses many different properties and variations through one summarizing code. Soil type is one of the most important variables for predicting SOC next to land management (Meersmans *et al*., 2008, 2009b; Bell & Worrall, 2009; Wiesmeier *et al*., 2011). Our results support this conclusion as soil type was the most important explanatory variable in all methods examined. Nevertheless, including soil type as the only explanatory variable may not be sufficient (Bell & Worrall, 2009).

Land management practices influence the amount of carbon sequestered into soil and loss of carbon from the soil (Powlson *et al*., 2011). We were not able to use land management information, but we can still see the influence on our predictions. For the mixed model the effects of site and plot were found to have very large estimated variances compared with other random effects, indicating their importance in capturing major components of the variation present in SOC. A tentative conclusion is that these site-specific effects reflect agricultural management (fertilizers, liming and drainage intensity) and topographical information such as curvature and slope. Likewise, we note that overall climate and crop-rotation effects contribute to the year-to-year variation although its estimated variance is smaller than for site.

It is well established that soils with larger Cf contents usually have larger SOC concentrations because of the stabilizing effect of clay particles. Cf content was relatively important in all of the models. The estimated fixed-effect slopes for soil-type-specific Cf content in the mixed model were positive, confirming that soils with greater Cf content (>50.1%) have a greater SOC concentration than sandy soils. Thus, the mixed-model approach appears to combine soil-specific relationships (captured in the fixed-effects part of the model) and sampling-specific structures (captured in the random-effects part).

Limitations

McBratney *et al*. (2002) pointed out that the prediction cannot readily be generalized beyond the specific geomorphic region. This criticism may also to some extent apply to the present study. However, some of the soil type compositions occurring in the Estonian database are very similar to structures found in other boreal pedoclimatic conditions. The measurement methods may also vary between countries. Having such similarities and discrepancies in mind, this study may provide some guidance on how to choose between different approaches for predicting SOC concentrations. It may still be necessary to adjust the models according to specific data configurations, as reflected in the available explanatory variables as well as in the data structure imposed by the applied sampling scheme.

Conclusions

Four different methods were used to predict SOC concentration in mineral arable soils in Estonia: these were the median, ANCOVA and mixed-model approaches and random forests. The results showed that the mixed model gave the most accurate predictions. The main reason for this finding is the ability of the mixed model to capture the variation introduced by the sampling design. The median and ANCOVA models both had MSE values larger than the mixed-model approach and random forests. Although median and ANCOVA models had merits over the latter methods, such as easy interpretation and application, and required less input data, they exhibited difficulties in capturing the entire range of SOC concentrations, especially at large concentrations. Random forests were an acceptable alternative, with prediction errors smaller than those from the median and ANCOVA models. The application of prediction models to the legacy soil map data highlighted the advantages of having accurate prediction models. Finally, we note that the entire framework for predicting SOC concentrations based on legacy soil data introduced in this paper can easily be up-scaled to country level.

Acknowledgements

We are grateful to the Editors and two anonymous reviewers for their numerous comments and suggestions, which have led to a much improved paper.

References

Arrouays, D., Richer de Forges, A., Morvan, X., Saby, N.P.A., Jones, A.R. & Le Bas, C. (eds) 2008. *Environmental Assessment of Soil for Monitoring: Volume IIb Survey of National Networks*, EUR 23490 EN/2B. Office for the Official Publications of the European Communities, Luxembourg.

Astover, A., Kõlli, R., Kukk, L. & Tamm, I. 2010. *Tartumaa maaressurss (Land resource in Tartu County)*. Estonian University of Life Sciences, Tartu.

Baker, J.M., Ochsner, T.E., Venterea, R.T. & Griffis, T.J. 2007. Tillage and soil carbon sequestration – what do we really know *Agriculture Ecosystems & Environment*, **118**, 1–5.

Bates, D., Maechler, M. & Bolker, B. 2011. *lme4: Linear Mixed-Effects Models Using S4 Classes*, R package version 0.999375-41 [WWW document]. URL http://cran.r-project.org/package=lme4 [accessed on 10 October 2011].

Batjes, N.H. 1996. Total carbon and nitrogen in the soils of the world. *European Journal of Soil Science*, **47**, 151–163.

Bell, M.J. & Worrall, F. 2009. Estimating a region's soil organic carbon baseline: the undervalued role of land-management. *Geoderma*, **152**, 74–84.

Breiman, L. 2001. Statistical modeling: the two cultures. *Statistical Science*, **16**, 199–231.

Estonian Land Board 2012. [WWW document]. URL http://geoportaal.maaamet.ee/eng/Maps-and-Data/Estonian-Soil-Map-p316.html [accessed on 2 January 2012].

Goidts, E., van Wesemael, B. & Van Oost, K. 2009. Driving forces of soil organic carbon evolution at the landscape and regional scale using data from a stratified soil monitoring. *Global Change Biology*, **15**, 2981–3000.

Gray, J.M., Humphreys, G.S. & Deckers, J.A. 2009. Relationships in soil distribution as revealed by a global soil database. *Geoderma*, **150**, 309–323.

© 2012 The Authors

Grimm, R., Behrens, T., Märker, M. & Elsenbeer, H. 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island – digital soil mapping using Random Forests analysis. *Geoderma*, **146**, 102–113.

Haskard, K.A., Rawlins, B.G. & Lark, R.M. 2010. A linear mixed model, with non-stationary mean and covariance, for soil potassium based on gamma radiometry. *Biogeosciences*, **7**, 2081–2089.

Hastie, T., Tibshirani, R. & Friedman, J. 2009. *The Elements of Statistical Learning*, 2nd edn. Springer, New York.

IUSS Working Group WRB 2006. *World Reference Base for Soil Resources 2006*, 2nd edn. World Soil Resources Reports No 103, FAO, Rome.

Kask, R. 1975. *Eesti NSV maafond ja selle põllumajanduslik kvaliteet [Estonian land resource and its agricultural quality]*. Valgus, Tallinn.

Katschinski, N.A. 1956. Die mechanische bodenanalyse und die klassifikation der böden nach ihrer mechanischen zusammensetzung. In: *Rapports au sixieme congres de la science du sol*, *Paris*, *B*. Laboureur & Cie, Paris. pp. 321–327.

Kempen, B., Heuvelink, G.B.M., Brus, D.J. & Stoorvogel, J.J. 2010. Pedometric mapping of soil organic matter using a soil map with quantified uncertainty. *European Journal of Soil Science*, **61**, 333–347.

Kokk, R. & Rooma, I. 1983. *Haritavad maad [Arable soils]. Estonian Soils in Figures III.* Eesti NSV Põllumajandusministeeriumi Informatsiooni ja Juurutamise Valitsus, Tallinn.

Kõlli, R. & Ellermäe, O. 2003. Humus status of postlithogenic arable mineral soils. *Agronomy Research*, **1**, 161–174.

Kõlli, R., Ellermäe, O., Kauer, K. & Köster, T. 2010. Erosion-affected soils in the Estonian landscape: humus status, patterns and classification. *Archives of Agronomy & Soil Science*, **56**, 149–164.

Lark, R.M., Cullis, B.R. & Welham, S.J. 2006. On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML. *European Journal of Soil Science*, **57**, 787–799.

Liaw, A. & Wiener, M. 2002. Classification and regression by randomForest. *R News*, **2**, 18–22.

Maia, S.M.F., Ogle, S.M., Cerri, C.C. & Cerri, C.E.P. 2010. Changes in soil organic carbon storage under different agricultural management systems in the Southwest Amazon Region of Brazil. *Soil & Tillage Research*, **106**, 177–184.

McBratney, A.B., Minasny, B., Cattle, S.R. & Vervoort, R.W. 2002. From pedotransfer functions to soil inference systems. *Geoderma*, **109**, 41–73.

Meersmans, J., De Ridder, F., Canters, F., De Baets, S. & Van Molle, M. 2008. A multiple regression approach to assess the spatial distribution of Soil Organic Carbon (SOC) at the regional scale (Flanders, Belgium). *Geoderma*, **143**, 1–13.

Meersmans, J., van Wesemael, B. & Van Molle, M. 2009a. Determining soil organic carbon for agricultural soils: a comparison between the Walkley & Black and the dry combustion methods (north Belgium). *Soil Use & Management*, **25**, 346–353.

Meersmans, J., van Wesemael, B., De Ridder, F. & Van Molle, M. 2009b. Modelling the three-dimensional spatial distribution of soil organic carbon (SOC) at the regional scale (Flanders, Belgium). *Geoderma*, **152**, 43–52.

Minasny, B. & McBratney, A.B. 2007. Spatial prediction of soil properties using EBLUP with the Matérn covariance function. *Geoderma*, **140**, 324–336.

Mishra, U., Lal, R., Slater, B., Calhoun, F., Liu, D. & Van Meirvenne, M. 2009. Predicting soil organic carbon stock using profile depth distribution functions and ordinary kriging. *Soil Science Society of America Journal*, **73**, 614–621.

Mishra, U., Lai, R., Liu, D. & Van Meirvenne, M. 2010. Predicting the spatial variation of the soil organic carbon pool at a regional scale. *Soil Science Society of America Journal*, **74**, 906–914.

Powlson, D.S., Whitmore, A.P. & Goulding, K.W.T. 2011. Soil carbon sequestration to mitigate climate change: a critical re-examination to identify the true and the false. *European Journal of Soil Science*, **62**, 42–55.

R Development Core Team 2011. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. ISBN 3-900051-07-0 [WWW document]. URL http://www.r-project.org [accessed on 10 October 2011].

Reintam, L., Kull, A., Palang, H. & Rooma, I. 2003. Large-scale soil maps and a supplementary database for land use planning in Estonia. *Journal of Plant Nutrition & Soil Science*, **166**, 225–231.

Smith, P. 2008. Land use change and soil organic carbon dynamics. *Nutrient Cycling in Agroecosystems*, **81**, 169–178.

Suuster, E., Ritz, C., Roostalu, H., Reintam, E., Kõlli, R. & Astover, A. 2011. Soil bulk density pedotransfer functions of the humus horizon in arable soils. *Geoderma*, **163**, 74–82.

van Wesemael, B., Paustian, K., Andrén, O., Cerri, C.E.P., Dodd, M., Etchevers, J. *et al*. 2011. How can soil monitoring networks be used to improve predictions of organic carbon pool dynamics and CO_2 fluxes in agricultural soils? *Plant & Soil*, **338**, 247–259.

Vittinghoff, E., Glidden, D.V., Shiboski, S.C. & McCulloch, C.E. 2005. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. Springer, New York.

Vorobeva, L.A. 1998. *Khimicheskij analiz pochv [Chemical analysis of soils]*. Moscow University Press, Moscow.

Wiesmeier, M., Barthold, F., Blank, B. & Kögel-Knabner, I. 2011. Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. *Plant & Soil*, **340**, 7–24.

© 2012 The Authors
Journal compilation © 2012 British Society of Soil Science,

III

Ritz, C., **Putku, E.** & Astover, A. 2015.
A practical two-step approach for mixed model-based kriging, with an application to the prediction of soil organic carbon concentration.

European Journal of Soil Science, 66, 548–554.

European Journal of Soil Science

European Journal of Soil Science, May 2015, **66**, 548–554 doi: 10.1111/ejss.12238

A practical two-step approach for mixed model-based kriging, with an application to the prediction of soil organic carbon concentration

C. Ritz[a], E. Putku[b] & A. Astover[b]

[a]*Department of Nutrition, Exercise and Sports, Faculty of Science, University of Copenhagen, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark, and* [b]*Department of Soil Science and Agrochemistry, Institute of Agricultural and Environmental Sciences, Estonian University of Life Sciences, Kreutzwaldi 1a, EE-51014 Tartu, Estonia*

Summary

Soil scientists often use prediction models to obtain values at unsampled locations. The spatial variation in the soil is best captured by using the empirical best linear unbiased predictor (EBLUP) based on a restricted maximum likelihood (REML) approach that efficiently exploits available data on both mean trends and correlation structures. We proposed a practical two-step implementation of the REML approach for model-based kriging, exemplified by predicting soil organic carbon (SOC) concentrations in mineral soils in Estonia from the large-scale digital soil map information and a previously established prediction model. The prediction model was a linear mixed model with soil type, physical clay content (particle size < 0.01 mm) and A-horizon thickness as fixed effects and site, transect, plot, year, year-transect random intercepts and site-specific random slopes for clay content. We used only the site-specific intercept EBLUPs for estimating spatial correlation parameters as they described most of the variation in the random effects (86.8%). Fitting an exponential correlation model to these EBLUPs resulted in an estimated range of 10.5 km and the estimated proportion of the variance from the nugget effect was 0.23. The results of a simulation study showed a downwards bias that decreased with sample size. The results were validated through an external dataset, resulting in root mean square errors (RMSE) of 1.06 and 1.07% for the two-step approach for kriging and the model with only fixed effects (no kriging), respectively. These results indicate that using the two-step approach for kriging may improve prediction.

Introduction

Predicting soil properties is an intrinsic part of many soil scientists' routine work. However, available data are often too scarcely distributed to be used without initially applying a kind of prediction model to extrapolate from available database information to soil maps with an adequate resolution for supporting decision making as required in agricultural management (Bocchi & Castrignanò, 2007). As the content of such databases is usually limited with regard to the number of sampling locations or sites, it is important to use data-analytical approaches that condense as much database information as possible into operational prediction models.

Soil monitoring networks provide data retrieved from design-based and model-based sampling strategies for carrying out field surveys. The advantages and disadvantages of both strategies are thoroughly discussed by Brus & De Gruijter (1997).

Correspondence: E. Putku. E-mail: elsa.putku@emu.ee

Received 26 December 2013; revised version accepted 12 January 2015

A prerequisite for using design-based methods is that samples are collected through a random sampling scheme. In practice, however, sampling is not completely at random because this would be too costly and labour intensive. Instead, the samples are collected, at least partly, through systematic sampling by means of grids or transects. As pointed out by Lark & Cullis (2004), linear mixed model methodology offers a flexible and powerful approach for modelling systematic variation in soil properties (through so-called fixed effects) while accounting for the spatial variation and variation introduced through systematic sampling schemes (through so-called random effects). Linear mixed models have been used for predicting soil properties in a number of different contexts (Haskard *et al*., 2010; Maia *et al*., 2010; Suuster *et al*., 2011, 2012; Doetterl *et al*., 2013).

Lark *et al*. (2006) introduced the REML-EBLUP (restricted maximum likelihood-empirical best linear unbiased predictors) approach for predicting soil properties, an approach that exploits both mean trends and correlation structures in the spatial distributions across the sampled region in order to produce predictions. Linear mixed

© 2015 British Society of Soil Science

models form the backbone of the approach and they are fitted with restricted maximum likelihood (REML) estimation rather than ordinary maximum likelihood estimation to avoid downwards-biased estimates of the variance parameters. The EBLUPs provide the link between the linear mixed model framework and kriging as they are the key quantities for capturing the spatial distribution in the data. The REML-EBLUP approach has been used in a number of applications in soil science (Minasny & McBratney, 2007; Chai *et al*., 2008; Santra *et al*., 2012). Moreover, Suuster *et al.* (2011, 2012) demonstrated that, even without using all available spatial information through model-based kriging, linear mixed model methodology and the use of EBLUPs still offers a powerful approach for obtaining predictions when compared with commonly used machine learning and regression techniques. These authors demonstrated improvements in prediction although they did not fully exploit the spatial data structure by assuming a spatial correlation structure.

However, in the literature there is a lack of concrete and transparent examples of how all the steps in the data analysis are actually carried out in practice using statistical software. Only a few publications provide appendices or supplementary material that actually show how analyses were carried out. Also, for open source software such as **R** (R Core Team, 2014) there seems to be a shortcoming when it comes to functionality for fitting linear mixed models with spatially correlated random effects. The packages gstat, geoR and geoRglm provide a wide range of methods and tools for carrying out kriging and model-based geostatistics (Ribeiro & Diggle, 2001; Christensen & Ribeiro, 2002; Pebesma, 2004). However, they do not provide functionality for analysing hierarchically structured data by means of linear mixed models with random effects that are spatially correlated. Specifically, when considering relevant mixed model functionality in **R**, the model fit function *lme()* in the package nlme (Pinheiro *et al.,* 2014) allows for spatially or temporally correlated residual terms but it is not possible to define a spatial correlation structure between random effects. However, such a correlation structure would be needed if repeated measurements were available for each of the sampling sites in a particular study.

The aim of this paper is to propose a flexible two-step implementation of the REML-EBLUP approach in open-source statistical software. The proposed approach addresses the shortcomings that currently exist in such software, but also, on a more general note, provides an alternative estimation procedure for handling analyses of complex hierarchically structured spatial data, an approach that may also be of interest in other contexts. We believe that the availability of such an open-source software infrastructure could promote a more widespread and routine use of state-of-the-art statistical methods by practitioners of soil science. We demonstrate the usefulness of the approach in a case study where the spatial distribution of soil organic carbon (SOC) concentrations in mineral soils in Estonia is predicted from available soil map data and an established prediction model.

Materials and methods

SOC concentration data

Data used for fitting the linear mixed model were collected by the Estonian national arable soil monitoring network during two periods: 1983–1994 and 2002–2008. In total, 79 representative sites all over Estonia were monitored. As the monitoring sites were located only at mineral soils, peaty and peat soils have to be neglected from the prediction model. Each sampling site had four transects with 10 plots ($1\,m^2$) in each transect. At every plot, A-horizon thickness, SOC concentration (determined by wet digestion with acid dichromate of carbon) and physical clay (Cf; particle size < 0.01 mm) content (%) were measured.

In the case study, the SOC values were predicted on the Estonian soil map (Estonian Land Board, 2013). This large-scale soil map (1:10 000) is in the Lambert-EST coordinate system and contains information about the soil type (according to the Estonian Soil Classification), texture, stoniness and A-horizon thickness. Texture classes available from the map database were used to get approximate Cf content values for the model by using average Cf content (sand, 7.5%; loamy sand, 15%; light loam, 25%; medium loam, 35%; heavy loam, 45%; clay, 60%). As the soil map covers all of Estonia, we had to select the arable land and mineral soils for the prediction model. A field map derived from the Estonian Agricultural Registers and Information Board was used to identify the arable soils suitable for the prediction model. For additional details on the study design and the data collection we refer to Suuster *et al*. (2012).

The validation data were sampled in a sub-region of Tartu County as part of an agrochemical survey in Estonia carried out in 2002–2004. In total, SOC concentration data for 85 arable fields were available for validation. The survey scheme was to take one composite soil sample from a 3–5 ha field. We used centroid coordinates to match these fields to soil map data. We extracted soil type and texture data (for mean Cf content) from the soil map data as in the survey only SOC concentration was measured. The selected sub-region for model validation contained 3 of the original 79 sites used in the second step. The study area for the validation dataset with the sites used for fitting the mixed model is shown in Figure 1.

Model-based kriging in two steps

We used the linear mixed model formulated for our case study to provide an illustration of the methodology. This model involved six levels of random effects, which were partly nested and partly crossed, and it may be formulated as follows:

$$\begin{aligned} SOC =& \beta_1\,(\text{soil type}) + \beta_2 \times A\,(cm) + \beta_3\,(\text{soil type}) \\ & \times Cf\,\text{content} + \mathrm{A_{site}} + \mathrm{B_{site}} \times Cf\ \text{content} + \mathrm{C_{plot}} \\ & + \mathrm{D_{transect}} + \mathrm{E_{year}} + \mathrm{F_{transect-year}} + \varepsilon. \end{aligned} \quad (1)$$

The fixed effects structure (the β_1, β_2 and β_3 parameters) corresponds to linear relationships between SOC and Cf content with

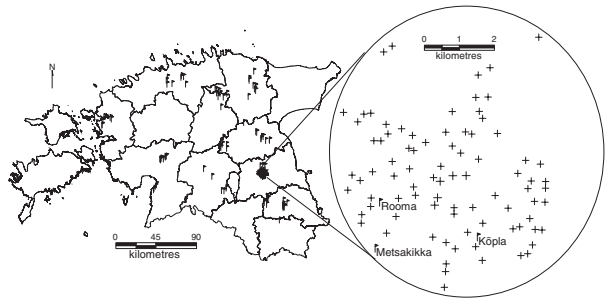

Figure 1 Map of sites (flags) for fitting the linear mixed model used in the first step of the proposed approach. In the circle the sites in the validation dataset (crosses) with the original sites (flags with site names) are zoomed in.

intercepts and slopes depending on the soil type, and additionally there is a separate gradient effect of the A-horizon thickness. The random effects structure includes random intercepts for site, plot, transect, year and transect by year as well as site-specific random slopes for Cf content. The random intercepts and slopes for sites (the A and B) and random intercepts for plots and transects within sites (the C and D) and years and transects within years (the E and F) as well as the residual errors (the ε) were assumed to follow independent normal distribution with mean 0 and unknown variance components. In the full formulation, we model A_{site} as normally distributed with mean zero and with correlation depending on the separation distance between two sites. However, we propose a flexible two-step approach in which we initially considered the linear mixed model but without the spatial correlation structure, thus containing only fixed effects and non-spatial random effects as listed above. Subsequently, spatial correlation was estimated using a generalized least squares procedure applied to the predicted random effects obtained from the initial linear mixed model.

In the first step the appropriate linear mixed model with relevant fixed and random effects was fitted. We may view the fixed effects as a way to use available (recorded) information on soil properties (Lark, 2012). In a sense, fixed effects reflect the variation in the soil data that may be explained by existing knowledge, quantifying it into a number of explanatory variables. In contrast, the random effects capture the unexplained variation in the soil data. Linear mixed models may be fitted by using maximum likelihood or, to reduce downwards bias in the estimated variance parameters, REML estimation (Lark & Cullis, 2004). Following standard practice we used REML.

In the second step the additional information in the data regarding the spatial correlation structure between sites was extracted through a separate statistical analysis that takes the site-specific predicted random effects obtained from the above linear mixed model fitted in the first step as input data. These predicted random effects are usually referred to as EBLUPs and they may be viewed as reflecting most of the spatial information in the data (Robinson, 1991). Moreover, they are standard output from most statistical software packages that allow the fitting of linear mixed models.

For the random effects at the highest level in the hierarchical structure, it is well known that the corresponding predicted random effects (the site-specific EBLUPs in our example) will, in general, be less variable than the unobserved or assumed random effects, a fact often referred to as shrinkage (Fitzmaurice *et al*., 2011). However, it may be shown that, by ignoring the dependence of EBLUPs on parameter estimates, the spatial correlation between EBLUPs will be identical to the assumed spatial correlation of the unobserved random effects in the model (for example by using the equation (3.92) in McCulloch *et al*., 2008). This is an approximation for small datasets but the dependency will vanish asymptotically, that is for large sample sizes. To our knowledge this fact about EBLUPs has not previously been exploited for inferential purposes. So based on this result we derived the following statistical model for the site-specific intercept EBLUPs (the u_i in Equation (2)): a simple linear regression model only involving an intercept parameter denoted β_0 (assuming the slope parameter is equal to 0):

$$u_i = \beta_0 + \varepsilon_i, \tag{2}$$

assuming that the residual errors (the εs) are mean-zero normally distributed with standard deviation σ_{site}, which was already estimated in the first step, but is also estimated in this second step. However, for model-based kriging we will use the estimate from the first step as the estimate from the second step will reflect the shrinkage of the EBLUPs.

We considered two spatial correlation structures. The exponential correlation structure with a nugget assumed the following correlation between sites:

$$\mathrm{cor}\left(u_i,\ u_{i'}\right) = (1-\eta)\ \times \exp\left(-\frac{d\left(s_i, s_{i'}\right)}{\rho}\right), \tag{3}$$

for predicted random effects u_i and $u_{i'}$ corresponding to sites with coordinates s_i and $s_{i'}$, respectively. This correlation structure corresponds to an isotropic, exponentially decreasing correlation as a function of the distance $d(s_i,\ s_{i'})$ between sites s_i and $s_{i'}$.We used the Euclidean metric. The parameter η is the proportion

of the total variance (and is a variance ratio between 0 and 1) that may be attributed to a nugget effect. The parameter ρ is the distance or range parameter (in km), which controls how rapidly the correlation vanishes: the smaller the value the more rapid is the decay (Haskard *et al.*, 2010). For distances between sites larger than 3ρ the corresponding random effects are effectively uncorrelated with $\text{cor}\left(u_i,\ u_{i'}\right) < 0.05$ (Lark, 2009). As an alternative, following the recommendation by Stein (1999), it may be more appropriate to consider an isotropic Matérn correlation structure, which is defined by means of the gamma function and modified Bessel function of the second kind, denoted Γ and K_ν, respectively, as follows (Haskard *et al.,* 2007):

$$\text{cor}\left(u_i, u_{i'}\right) = \frac{\frac{d\left(s_i, s_{i'}\right)}{\rho} \times K_\nu\left(\frac{d\left(s_i, s_{i'}\right)}{\rho}\right)}{2^{\nu-1} \times \Gamma(\nu)}. \quad (4)$$

This correlation structure involves two parameters, ρ and ν, which control rate of decay and smoothness, respectively. Additionally, as with the exponential model, a parameter for the nugget effect could have been included (Sherman, 2011). Specifically, the parameter ρ controls the rate of decay of the correlation (the smaller ρ, the more rapidly the correlation vanishes) but although ρ has the same unit as used for the eastings/northings there is no general way to scale up the parameter to get distances that result in specific percentage reductions in the correlation. The larger the value of ν the smoother is the underlying random field characterizing the spatial correlation. For $\nu = 0.5$ the above exponential correlation structure in Equation (3) is recovered as a special case and for ν tending to infinity (in an appropriate way) the Gaussian correlation structure is obtained. The main advantage of the Matérn model over the exponential model is that it allows estimation of the degree of smoothness of the underlying random process (instead of an *a priori* assumption of a certain degree of smoothness). As the model defined in Equation (4) contained no random effects but only the spatial correlation structure, inference was based on generalized least squares REML estimation following Davidian & Giltinan (1995) and Pinheiro & Bates (2000). For smaller datasets we propose to estimate the smoothness parameter robustly using a profile likelihood approach over a grid of values of ν, using generalized least squares only to estimate the range parameter for each (fixed) ν value.

The estimated fixed-effects parameters, the estimated variance component of the site-specific random intercepts (from the analysis in the first step) and the estimated range parameter ρ (from the second step) were used to obtain predicted values and their standard errors. Specifically, we inserted these estimates into standard mixed model-based kriging formulas, such as equations (1.7) and (1.8) given by Lark *et al.* (2006).

The statistical environment **R** version 3.1.2 (R Core Team, 2014) was used for the statistical analyses. In particular, the **R** extension packages lme4 and nlme were used for carrying out the first and second steps, respectively (Bates *et al.*, 2014; Pinheiro *et al.*, 2014). The Matérn correlation structure was available through the extension package spaMM (Rousset & Ferdy, 2014). Note that the correlation parameters in nlme are usually estimated on an unrestricted logarithmic scale and subsequently back-transformed using the exponential function. Spatial predictions were obtained using custom-made **R** functions. Both data for the site-specific EBLUPs only and **R** script for carrying out the two-step approach for linear mixed model-based kriging reported in the Results section are provided in Appendix S1. For the simpler situation where only a single measurement was obtained per site the first step of our approach may be skipped entirely. However, in this case customized functionality is also available in the **R** package geoR.

Results

Simulation study

We generated a dataset from a simplified version of the linear mixed model, which we needed for the case study reported on in the next sub-section. We assumed zero mean and plot- and site-specific random intercepts (two plots per site, each with two measurements), which were assumed to have variances 0.5 and 5, respectively. Additionally, the site random effects were assumed to be spatially correlated according to an exponential correlation structure (using the Euclidean metric). The residual variance was set to 1. We considered three different values of the range parameter (1, 2, 5 and 10 km) in combination with a realistic range for the number of sites (25, 50 and 100, compared with 79 sites in the actual case study) and assuming pseudo-replication over 3 years. We generated 1000 datasets for each combination using randomly located sites (each coordinate sampled randomly from a uniform distribution) within sampling areas (squares) of widths 5.5, 11, 31 and 45 for the considered range parameters 1, 2, 5 and 10, respectively. The resulting median estimated range parameters along with the corresponding interquartile ranges are shown in Table 1. The proposed two-step approach produced a downwards-biased estimated range parameter. The bias decreased as the number of sites increased.

Case study

As a first step we considered the linear mixed model for predicting SOC concentrations previously established by Suuster *et al.* (2012). The resulting estimated variance components are shown in Table 2. It was evident that the main source of the variation in SOC data was the variation between the sites, which reflects both natural spatial variation and differences in agronomic practices such as fertilizer application, ploughing depth and liming. This finding was not surprising because the sites were the sampling units with the largest spatial variation by far whereas the remaining sampling units (plots and transects) were nested within the sites (Suuster *et al.,* 2012). This result also supported the assumption that most of the spatial information in the data was captured by the site-specific intercept EBLUPs and, consequently, estimation of spatial correlation parameters without much loss of information may be based solely on these EBLUPs.

Table 1 Simulation results based on 1000 datasets generated from a linear mixed model with plot- and site-specific random intercepts; the latter are assumed to be spatially correlated according an exponential correlation structure

	Number of sites		
	25	50	100
True range / km	Median range / km		
1	0.7 (0.3, 1.4)[a]	0.8 (0.4, 1.5)	0.9 (0.5, 1.4)
2	1.3 (0.5, 2.9)	1.5 (0.8, 2.9)	1.8 (1.2, 2.9)
5	3.3 (1.4, 7.2)	3.5 (1.8, 7.2)	3.9 (1.8, 6.8)
10	5.1 (1.6, 11.3)	6.0 (2.3, 13.5)	8.4 (5.1, 14.8)

[a]Interquartile range in parentheses. Median and interquartile range of the estimated range parameters from the proposed two-step approach applied to each of the generated datasets are shown for combinations of the true range of the generated data and number of sites.

Table 2 Estimated variance components from the linear mixed model fitted in the first step, shown on the squared SOC scale and as percentages of the total estimated variance in the model

Random effect	Estimated variance	Percentage of total variance
Site (intercept)	0.548	86.8
Site (slope on Cf content)	0.001	0.1
Plot (nested within transects)	0.048	7.7
Transect (nested within sites)	0.004	0.7
Transect-year (nested within year)	0.009	1.4
Year	0.002	0.3
Residual error	0.019	3.0

In the second step for the exponential correlation structure we obtained the following estimated nugget (as a proportion) and range with corresponding approximate 95% confidence intervals: 0.23 (0.05–0.66) and 10.5 (2.4–46.1) km, respectively. Moreover, the estimated site-specific variance component was 0.50%, which was because the shrinkage of the EBLUPs was somewhat smaller than the value reported in Table 2 for the analysis carried out in the first step. For the Matérn correlation structure we estimated the smoothness parameter to be 0.13 (0.05–0.58), indicating a very rough underlying process, which is not significantly different from 0.5 (at a 5% significance level), supporting the assumption that the exponential correlation structure with a nugget more adequately describes the lack of smoothness: Haskard (2007) reached a similar conclusion. Moreover, we obtained a very imprecisely determined estimated range with the entire interval $(0, \infty)$ as the 95% confidence interval; this also indicates that there is not sufficient information in the data to support a Matérn model.

For the validation dataset we calculated the root mean square error (RMSE) for the proposed two-step approach. As a comparison we also calculated RMSE based on the linear mixed model obtained in the first step so that predictions were based on the fixed-effects estimates only. Note that the approach described by Suuster *et al.* (2012) was not applicable as there was no overlap between the sites in the validation dataset and the sites used in the fitting of the linear mixed model (see Figure 1). The exponential model (using the two-step approach for kriging) resulted in an RMSE of 1.06%, compared with an RMSE of 1.07% for the model only using the fixed-effects estimates. There was in this example, therefore, only a very modest gain in using the two-step approach.

Discussion

By exploiting the fact that that correlation structures are preserved when replacing assumed (unknown) random effects by EBLUPs (not affected by shrinkage), and linking existing functionalities in the open-source statistical programming environment **R**, we have implemented a novel and flexible two-step approach for carrying out model-based kriging for complex hierarchical database structures. The under-estimation (downwards bias) in the simulation results might be explained by the additional non-spatial dependency between site-specific predicted random effects (EBLUPs) introduced through insertion of parameter estimates. Therefore the EBLUPs become more correlated than the unobserved site-specific random effects, which were only assumed in the model to be spatially correlated. This additional correlation will vanish as the number of sites increases. It is worth noting that the study areas were increased in size as the considered range parameter was increased (Stein, 1999). For any specific range value considered, we used the same size study area regardless of the number of sites considered to ensure comparable results. However, a more efficient design would require that more sites should be spread out to capture more of the exponentially decaying correlation (Lark, 2002; Zhu & Stein, 2005). This limitation may also in part explain the downwards bias.

The validation study showed a small gain in the accuracy of the predictions. This finding is probably a result of the validation dataset not being designed for the purpose of prediction. Validation data only covered a small part of the entire study area used for fitting the linear mixed model in the first place and only in a few cases were sites in the validation data close to sites in the training data in terms of the estimated range for the spatial correlation. Moreover, the training dataset and external validation dataset were obtained using different sampling schemes. The training data were from point sampling whereas for the validation data one sample covered 3–5 ha. In the latter case some information was lost because of 'generalization': in most cases the 3–5-ha plots included several soil mapping units but we use only centroid point data for prediction. It will also be useful to evaluate the predictive performance of the proposed approach for datasets specifically designed for obtaining predictions.

Our procedure may be extended to allow using other metrics such as the city block metric as well as stationary and non-stationary spatial correlation structures (Haskard *et al.*, 2007; Lark, 2009). In principle, the two-step approach should also be applicable for spatial binomial and count data through the function *glmer()* in the **R** package lme4.

Limitations

The flexibility of the proposed two-step procedure comes at the cost of losing some of the variation in the data because it is not propagated from the first to the second step. Thus we did not incorporate the uncertainty in the EBLUPs but simply assumed that they were equally variable, an assumption that may be questionable if large imbalances exist in sampling schemes between sites.

The predicted random effects were only asymptotically (for very large datasets) identical to the unobserved random effects assumed for the model, except for a scaling factor (McCulloch *et al*., 2008). Therefore the proposed approach will result in under-estimates for smaller sample sizes, but the bias will reduce as the sample size increases. The method relies on the assumption that most variation is captured by the random effects at the top level of the hierarchy, corresponding to the crudest grouping in the experimental design. Thus we only used the site-specific intercept EBLUPs available from a linear mixed model for retrieving spatial information, but not those for any of the other random effects in the model (one of which was site-specific random slope). For the SOC data we considered that this was motivated by the fact that the site-specific intercepts random effects explained most of the variation and the remaining random effects only provided very small contributions to the total variation in the data. However, we believe that in many applications in soil science it will be possible to identify a single random effects variable that corresponds to sampling units (locations or sites for example) with the largest spatial distribution (Lark & Cullis, 2004; Haskard *et al*., 2010).

Conclusions

We proposed a novel, two-step modelling methodology for carrying out model-based kriging. First, a linear mixed model was used to capture non-spatial correlation that is independent of the spatial distribution of sampling sites in the study. Then, in the second step, spatial correlation between sampling sites was modelled. Through simulations and a case study we have demonstrated that the proposed methodology provides an attractive and operational alternative approach to obtain predictions based on model-based kriging. The methodology is flexible in that it allows combined complex hierarchical models with arbitrary spatial correlation structures as they are essentially dealt with through separate models. Thus the approach is both conceptually and from an implementation perspective separating spatial and non-spatial components of correlation. This point of view may lead to a better understanding and appreciation of various sources of variation often observed in practice. In addition, as open source statistical software may be used, the methodology may more readily be adopted by practitioners.

Supporting Information

The following supporting information is available in the online version of this article:

Appendix S1. A practical two-step approach for mixed model-based kriging, with an application to the prediction of soil organic carbon concentration.

Acknowledgements

The study is supported by the Estonian Research Council from the Environmental Protection and Technology R&D Programme (KESTA) under activity 5: 'Geoinformatic development of biodiversity, soil and earth data system (ERMAS)'. We thank Priit Penu from the Agricultural Research Centre for providing the validation data. We are grateful to the editors and reviewers for their valuable comments. The authors have no conflict of interests to declare.

References

Bates, D., Maechler, M., Bolker, B. & Walker, S. 2014. *lme4: Linear Mixed-effects Models using Eigen and S4. R Package Version 1.1-7* [WWW document]. URL http://CRAN.Rproject.org/ package=lme4 [accessed on 15 January 2015].

Bocchi, S. & Castrignanò, A. 2007. Identification of different potential production areas for corn in Italy through multitemporal yield map analysis. *Field Crops Research*, **102**, 185–197.

Brus, D.J. & de Gruijter, J.J. 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). *Geoderma*, **80**, 1–44.

Chai, X., Shen, C., Yuan, X. & Huang, Y. 2008. Spatial prediction of soil organic matter in the presence of different external trends with REML-EBLUP. *Geoderma*, **148**, 159–166.

Christensen, O.F. & Ribeiro, P.J. Jr. 2002. geoRglm: a package for generalised linear spatial models. *R News*, **2**, 26–28.

Davidian, M. & Giltinan, D.M. 1995. *Nonlinear Models for Repeated Measurement Data*. Chapman & Hall, New York.

Doetterl, S., Stevens, A., van Oost, K., Quine, T.A. & van Wesemael, B. 2013. Spatially-explicit regional-scale prediction of soil organic carbon stocks in cropland using environmental variables and mixed model approaches. *Geoderma*, **204–205**, 31–42.

Estonian Land Board 2013. [WWW document]. URL http://geoportaal.maaamet.ee/eng/Maps-and-Data/Estonian-Soil-Map-p316.html [accessed on 7 December 2013].

Fitzmaurice, G.M., Laird, N.M. & Ware, J.H. 2011. *Applied Longitudinal Analysis*, 2nd edn. John Wiley & Sons, Hoboken, NJ.

Haskard, K.A. 2007. *An anisotropic matern spatial covariance model: REML estimation and properties.* PhD thesis, University of Adelaide, Adelaide.

Haskard, K.A., Cullis, B.R. & Verbyla, A.P. 2007. Anisotropic Matérn correlation and spatial prediction using REML. *Journal of Agricultural, Biological, & Environmental Statistics*, **12**, 147–160.

Haskard, K.A., Rawlins, B.G. & Lark, R.M. 2010. A linear mixed model, with non-stationary mean and covariance, for soil potassium based on gamma radiometry. *Biogeosciences Discussions*, **7**, 1839–1862.

Lark, R.M. 2002. Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood. *Geoderma*, **105**, 49–80.

Lark, R.M. 2009. Kriging a soil variable with a simple nonstationary variance model. *Journal of Agricultural, Biological, & Environmental Statistics*, **14**, 301–321.

Lark, R.M. 2012. Towards soil geostatistics. *Spatial Statistics*, **1**, 92–99.

Lark, R.M. & Cullis, B.R. 2004. Model-based analysis using REML for inference from systematically sampled data on soil. *European Journal of Soil Science*, **55**, 799–813.

Lark, R.M., Cullis, B.R. & Welham, S.J. 2006. On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML. *European Journal of Soil Science*, **57**, 787–799.

Maia, S.M.F., Ogle, S.M., Cerri, C.C. & Cerri, C.E.P. 2010. Changes in soil organic carbon storage under different agricultural management systems in the Southwest Amazon Region of Brazil. *Soil & Tillage Research*, **106**, 177–184.

McCulloch, C.E., Searle, S.R. & Neuhaus, J.M. 2008. *Generalized, Linear, and Mixed Models*, 2nd edn. John Wiley & Sons, Hoboken, NJ.

Minasny, B. & McBratney, A.B. 2007. Spatial prediction of soil properties using EBLUP with the Matérn covariance function. *Geoderma*, **140**, 324–336.

Pebesma, E.J. 2004. Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, **30**, 683–691.

Pinheiro, J.C. & Bates, D.M. 2000. *Mixed-Effects Models in S and S-plus*. Springer Verlag, New York.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & R Core Team. 2014. *nlme: Linear and Nonlinear Mixed Effects Models. R Package Version 3.1-118* [WWW document]. URL http://CRAN.R-project.org/package=nlme [accessed on 15 January 2015].

R Core Team 2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna [WWW document]. URL http://www.R-project.org [accessed on 15 January 2015].

Ribeiro, P.J. Jr. & Diggle, P.J. 2001. geoR: a package for geostatistical analysis. *R News*, **1**, 15–18.

Robinson, G.K. 1991. That BLUP is a good thing: the estimation of random effects. *Statistical Science*, **6**, 15–32.

Rousset, F. & Ferdy, J.B. 2014. Testing environmental and genetic effects in the presence of spatial autocorrelation. *Ecography*, **37**, 781–790.

Santra, P., Das, B. & Chakravarty, D. 2012. Spatial prediction of soil properties in a watershed scale through maximum likelihood approach. *Environmental Earth Sciences*, **65**, 2051–2061.

Sherman, M. 2011. *Spatial Statistics and Spatio-Temporal Data*. Wiley, Chichester.

Stein, M.L. 1999. *Interpolation of Spatial Data*. Springer-Verlag, New York.

Suuster, E., Ritz, C., Roostalu, H., Reintam, E., Kõlli, R. & Astover, A. 2011. Soil bulk density pedotransfer functions of the humus horizon in arable soils. *Geoderma*, **163**, 74–82.

Suuster, E., Ritz, C., Roostalu, H., Kõlli, R. & Astover, A. 2012. Modelling soil organic carbon concentration of mineral soils in arable land using legacy soil data. *European Journal of Soil Science*, **63**, 351–359.

Zhu, Z. & Stein, M.L. 2005. Spatial sampling design for parameter estimation of the covariance function. *Journal of Statistical Planning & Inference*, **134**, 583–603.

CURRICULUM VITAE

First name: Elsa
Last name: Putku
Date of birth: 19.10.1984
Employment: Agricultural Research Centre,
Kohtu 10A, Kuressaare 93812,
Department of Agricultural Research and Monitoring,
Soil Monitoring Bureau
Phone: +372 453 8262
E-mail: elsa.putku@pmk.agri.ee
Position: chief specialist

Education:

2008–2016	Estonian University of Life Sciences, Institute of Agricultural and Environmental Sciences, PhD studies
2006–2008	Estonian University of Life Sciences, master studies in natural sciences
2003–2006	Estonian University of Life Sciences, bachelor studies in natural sciences
1992–2003	Saaremaa Co-Educational Gymnasium

Career:

2015–…	Agricultural Research Centre, specialist
2010–2015	Estonian University of Life Sciences, Institute of Agricultural and Environmental Sciences, Department of Soil Science and Agrochemistry, specialist

Scientific degree: MSc in Natural Sciences
Institute and year of issuing degree:
Estonian University of Life Sciences, 2008

Administrative responsibilities:

2009–…	Estonian Soil Science Society

Research interest: pedometrics

Participation in research projects:

2010–2013	Ministry of the Agriculture of Estonia project No. 8-2/T10037PKML "Implementation of soil maps and databases for sustainable land use and agricultural production", principal investigator
2007	Rural Development Foundation project No.8-2/T7073MIMI "Landresource", principal investigator

ELULOOKIRJELDUS

Eesnimi: Elsa
Perekonnanimi: Putku
Sünniaeg: 19.10.1984
Töökoht: Põllumajandusuuringute Keskus,
Kohtu 10A, Kuressaare 93812,
Põllumajandusseire ja uuringute osakond,
Mullaseire büroo
Telefon: +372 453 8262
E-post: elsa.putku@pmk.agri.ee
Ametikoht: peaspetsialist

Haridustee:

2008–2016	Eesti Maaülikool, põllumajandus- ja keskkonnainstituut, mullateaduse ja agrokeemia osakond, doktorantuur
2006–2008	Eesti Maaülikool, loodusteaduse magistriõpingud
2003–2006	Eesti Maaülikool, loodusteaduse bakalaureuseõpingud
1992–2003	Saaremaa Ühisgümnaasium

Teenistuskäik:

2015–…	Põllumajandusuuringute Keskus, peaspetsialist
2010–2015	Eesti Maaülikool, põllumajandus- ja keskkonnainstituut, mullateaduse ja agrokeemia osakond, peaspetsialist

Teaduskraad: MSc (loodusteadus)
Teaduskraadi väljaandnud asutus, aasta:
Eesti Maaülikool, 2008

Teadusorganisatsiooniline ja -administratiivne tegevus:
2009–… Eesti Mullateaduse Selts

Teadustöö põhisuunad:
Pedomeetria, rakendusmudelid mullateaduses ja põllumajanduses

Osalemine uurimisprojektides:

2010–2013	EV Põllumajandusministeeriumi projekt nr. 8-2/T10037PKML „Mullastikukaartide- ja andmebaaside rakendused jätkusuutlikuks maakasutuseks ja põllumajandustootmiseks“, põhitäitja
2007	Maaelu Edendamise Sihtasutuse projekt nr. 8-2/T7073MIMI „Maaressurss“, põhitäitja

LIST OF PUBLICATIONS

1.1. Scholarly articles indexed by Thomson Reuters Web of Science

Ritz, C., **Putku, E.,** Astover, A. 2015. A practical two-step approach for mixed model-based kriging, with an application to the prediction of soil organic carbon concentration. European Journal of Soil Science, 66, 548–554.

Suuster, E., Ritz, C., Roostalu, H., Kõlli, R., Astover, A. 2012. Modelling soil organic carbon concentration of mineral soils in arable land using legacy soil data. European Journal of Soil Science, 63, 351–359.

Vasiliev, N., **Suuster, E.,** Luik, H., Värnik, R., Matveev, E., Astover, A. 2011. Productivity of Estonian dairy farms has declined after accession to the European Union. Agricultural Economics (AGRICECON), 57 (9), 457–463.

Kukk, L., Roostalu, H., **Suuster, E.,** Rossner, H., Shanskiy, M., Astover, A. 2011. Reed canary grass biomass yield and energy use efficiency in Northern European pedoclimatic conditions. Biomass & Bioenergy, 35 (10), 4407–4416.

Suuster, E., Ritz, C., Roostalu, H., Reintam, E., Kõlli, R., Astover, A. 2011. Soil bulk density pedotransfer functions of the humus horizon in arable soils. Geoderma, 163, 74–82.

Kukk, L., Astover, A., Muiste, P., Noormets, M., Roostalu, H., Sepp, K., **Suuster, E.** 2010. Assessment of abandoned agricultural land resource for bio-energy production in Estonia. Acta Agriculturae Scandinavica, Section B – Soil & Plant Science, 60 (2), 166–173.

3.4. Articles/presentations published in conference proceedings

Kukk, L., **Suuster, E.,** Astover, A., Noormets, M., Roostalu, H. 2009. Soil optimized suitability analysis for allocation of energy grasses. In: Van Ittersum, M.K., Wolf, J., Van Laar, H.H. (Ed.). Proceedings of the conference on Integrated Assessment of Agriculture and Sustainable Development: Setting the Agenda for Science and Policy (AgSap 2009). Egmond aan Zee, The Netherlands, 10–12 March 2009. (372–373). Wageningen University and Research Centre.

Suuster, E., Kukk, L., Astover, A., Noormets, M. 2009. Spatial assessment of abandoned agricultural land for bio-energy production. In: Van Ittersum, M.K., Wolf, J., Van Laar, H.H. (Ed.). Proceedings of the conference on Integrated Assessment of Agriculture and Sustainable Development: Setting the Agenda for Science and Policy (AgSap 2009). Egmond aan Zee, The Netherlands, 10–12 March 2009. (386–387). Wageningen University and Research Centre.

3.5. Articles/presentations published in local conference proceedings

Suuster, E., Astover, A., Roostalu, H., Penu, P. 2011. Suuremõõtkavalisest mullastikukaardist maakasutuse otsusteni. Kadaja, Jüri (Toim.). Agronoomia 2010/2011 (223–230). Saku: Eesti Maaviljeluse Instituut.

Suuster, E., Kukk, L., Astover, A., Roostalu, H., Noormets, M., Muiste, P. 2008. Kasutamata põllumajandusmaade potentsiaal bioenergia tootmiseks Saare maakonnas. Jõudu, J.; Noormets, M.; Viiralt, R.; Michelson, M.. Agronoomia 2008 (176–179). Tartu: Eesti Maaülikool.

5.2. Conference abstracts that do not belong to section 5.1

Putku, E., Ritz, C., Astover, A. 2016. Prediction of soil organic carbon concentration and soil bulk density of mineral soils for soil organic carbon stock estimation. In: Geophysical Research Abstracts, 18: European Geosciences Union General Assembly 2016 Vienna, Austria, 17–22 April. (EGU2016-12958).

Suuster, E., Ritz, C., Astover, A. 2012. Mixed model methodology is suitable for analyzing soil monitoring data. In: EUROSOIL 2012: 4th International Congress EUROSOIL 2012, Soil Science for the Benefit of Mankind and Environment, Fiera del Levante, Bari, Italy, 2–6 July 2012. Ed. Nicola Senesi. Bari, Italy, 1032.

Suuster, E., Ritz, C., Roostalu, H., Astover, A. 2011. Evaluation of different statistical approaches for predicting soil organic carbon stock. In: Abstract Proceedings of the 6th International Congress of the European Society for Soil Conservation "Innovative Strategies and

Policies for Soil Conservation": Thessaloniki, Greece 9–14 May. Ed. Karyotis, Th; Gabriels, D. National Agricultural Research Foundation, 143.

Reintam, E., Köster, T., Krebstein, K., Sanchez de Cima, D., Kuht, J., **Suuster, E.,** Selge, A., Astover, A. 2011. Soil productivity losses due to compaction in Estonia. In: Keesstra, S.; Mol, G. (Ed.). Program and abstract book: Soil Science in a Changing World (197). Wageningen, the Netherlands: Wageningen UR, Communication Services.

Suuster, E., Astover, A., Kõlli, R., Roostalu, H., Reintam, E., Penu, P. 2010. Spatial patterns of soil organic carbon stocks in Estonian arable soils. In: Geophysical Research Abstracts, 12: European Geosciences Union General Assembly 2010 Vienna, Austria, 02–07 May. (EGU2010-722).

Kukk, L., Astover, A., Roostalu, H., **Suuster, E.**, Noormets, M. 2009. Energy and land use efficiency of reed canary grass (Phalaris arundinacea L.) depending on nitrogen fertilisation. In: NJF Report Vol 3 No 3: NJF Seminar 428 "Energy conversion from biomass production – EU-AgroBiogas", Viborg, Denmark, 9–10 September 2009. 61–62.

6.3. Popular science articles

Roostalu, H., Astover, A., Kukk, L., **Suuster, E.** 2008. Bioenergia tootmise võimalustest põllumajanduses: I osa. Maamajandus, 32–35.

Roostalu, H., Astover, A., Kukk, L., **Suuster, E.** 2008. Bioenergia tootmise võimalustest põllumajanduses: II osa. Maamajandus, 42–45.

Roostalu, H.; Astover, A., Kukk, L., **Suuster, E.** 2008. Bioenergia tootmise võimalustest põllumajanduses: III osa. Maamajandus, 34–36.

Astover, A., Roostalu, H., Kukk, L., Muiste, P., Padari, A., **Suuster, E.**, Ostroukhova, A. 2008. Potentsiaalne maaressurss bioenergia tootmiseks. Eesti Põllumees, 38, 17.

VIIS VIIMAST KAITSMIST

BERIT TEIN

THE EFFECT OF DIFFERENT FARMING SYSTEMS ON POTATO TUBER YIELD AND QUALITY

VILJELUSSÜSTEEMIDE MÕJU KARTULI MUGULASAAGILE JA KVALITEEDILE

Dotsent **Are Selge**, teadur **Viacheslav Eremeev**, vanemteadur **Evelin Loit**

4. detsember 2015

HIIE IVANOVA

RESPONSES OF RESPIRATORY AND PHOTORESPIRATORY DECARBOXYLATIONS TO INTERNAL AND EXTERNAL FACTORS IN C_3 PLANTS

RESPIRATOORSE JA FOTORESPIRATOORSE DEKARBOKSÜÜLIMISE VASTUSED SISEMISTE JA VÄLISTE FAKTORITE TOIMEL C_3 TAIMEDELE

professor **Ülo Niinemets**, vanemteadur **Tiit Pärnik**, vanemteadur **Olav Keerberg**

08. jaanuar 2016

DIEGO SANCHEZ DE CIMA

SOIL PROPERTIES AFFECTED BY COVER CROPS AND FERTILIZATION IN A CROP ROTATION EXPERIMENT

VAHEKULTUURIDE JA VÄETAMISE MÕJU MULLA OMADUSTELE KÜLVIKORRAKATSES

dotsent **Endla Reintam**, emeriitprofessor **Anne Luik**

11. veebruar 2016

KRISTI PRAAKLE

CAMPYLOBACTER SPP. AND *LISTERIA MONOCYTOGENES* IN POULTRY PRODUCTS IN ESTONIA

CAMPYLOBACTER SPP. JA *LISTERIA MONOCYTOGENES* LINNULIHATOODETES EESTIS

professor **Mati Roasto**, professor **Marja-Liisa Hänninen** (Helsingi Ülikool), professor **Hannu Korkeala** (Helsingi Ülikool)

04. märts 2016

TARMO KALL

VERTICAL CRUSTAL MOVEMENTS BASED ON PRECISE LEVELLINGS IN ESTONIA

MAAKOORE VERTIKAALLIIKUMISED EESTIS TÄPPISNIVELLEERIMISTE ANDMETEL

Emeriitprofessor **Jüri Randjärv**, dotsent **Aive Liibusk**

29. aprill 2016

ISSN 2382-7076
ISBN 978-9949-569-27-4 (trükis)
ISBN 978-9949-569-28-1 (pdf)

roheline trükis
Trükitud taastoodetud paberile looduslike trükivärvidega. ©Ecoprint