RESPONSE TO REVIEWER COMMENTS. REVIEWER #1

We thank the reviewer for their comments, which we have used to clarify the research strategy and generally improve the manuscript.

Comment-1: Samples are taken in clusters covering only 12 fields in the study area (containing presumably a few 102 fields), with rather a poor spatial distribution. This will most probably affect the distribution of the data in the multi-dimensional space, and hence, do not cover enough the associated landscape complexity within the study area. These 2 concerns (poor spatial distribution and poor distribution in multi-dimensional space) are a major problem when the model is used for extrapolation / predicting pixels elsewhere in study area, i.e. outside the variable range covered by the calibration dataset.

Response-1: Although the samples are grouped spatially in fields for logistical reasons, they were carefully located to capture the distribution of conditions across the feature space of the agricultural fields. Specifically, the samples include the variety of parent materials in the study area, form transects across topographic positions, as well as cover regional highs and lows. If the modelling technique relied on spatial autocorrelation, such as with kriging, the spatial position of these samples would indeed have been a problem. However, because the modelling method uses spatial regression, which relies on the principle of spatial association, the important space to cover was the feature space. Therefore, the samples were taken to encompass the variable range of the agriculture fields as best could be determined prior to sampling. Notably, only agricultural fields were sampled and thus non-agricultural areas are masked out of prediction maps because they were outside the variable range covered by the calibration dataset. Text has been added to emphasize these points.

Comment-2: As a consequence, it’s quite possible that the differences in SOC stock maps between the two methods are more the consequence of the fact that the 2 modelling approaches (i.e. direct versus indirect) are reacting differently on this shortcoming (inappropriate multidimensional data cover) then it is actually reflecting a real difference in model output just/purely caused by the fact that 2 different approaches were used.

Response-2: We understand this concern, but consider the variables used to calculate the SOC stock (indirect) and the SOC stock itself (direct) to be intertwined due to their definitional relationship. Because of this relationship, covering the feature space of one approach increases the probability that the feature space of the other approach is also covered.

Comment-3: Finally, it’s clear how the authors calculated errors on SOC stocks by using classical error propagation techniques for individual pixels (i.e. for both the direct and the indirect method (including error predictions on components)), but it’s not clear if/how spatial autocorrelation was taken into account when mapping these errors. It’s important to integrate this effect of spatial autocorrelation in order to make a fair comparison between the error maps obtained by the two methods.

Response-3: Spatial autocorrelation was used in the grouping of rule condition zones, on which error estimations are applied. Spatial autocorrelation minimizes how different unsampled areas within a condition zone could be from the sampled locations. This closely resembles the approaches of Shrestha and Solomatine (2006) and Malone et al. (2006) for taking autocorrelation into account when mapping estimated errors from spatial regression models.

RESPONSE TO REVIEWER COMMENTS. REVIEWER #2

We thank the reviewer for their comments and suggestions for improving the paper. The following is our response to the specific comments.

Comment-1: I wonder about the effect of the 10 closely-paired samples…in addition to ensuring the estimated errors capture random variation, could it induce a bias towards these sample points?

Response-1: The closely-paired samples were spread across the feature space, which should help them be representative of the random variation across the spectrum of conditions and minimize emphasis on a particular set of conditions. Some text was added to explain this part of the sampling strategy.

The Cubist software does have an option for dealing with data sets that are biased towards a grouping of target variable values. Our experimentation with this feature did not produce regression equations or estimated errors that were very different from those produced without that feature activated. This gives us confidence that the closely-paired samples were distributed in a way to not bias the models built.

Comment-2: I am a little bit worried about the large number of potential predictors in the pool compared to the number of data on the target variable (117). I appreciate that the authors use the discussion to suggest explanations of why particular predictors were selected, and thus partly validate their selection, but am still not totally convinced that the same could not be done even with junk data and this many potential predictors. I wonder if some acknowledgement of the potential of data-mining software to overfit should be included and commented on. I don't think Cubist does anything to deal with the size of predictor pool (in a multiple hypothesis testing kind of way): : :a comment on this issue could be useful.

Response-2: Text has been added to provide more detail about how we used Cubist to reduce the predictor pool. Specifically, multiple passes were used so that the final models were built from predictor pools equal to the size of the sets used in the models (3-19 predictors divided between multiple rules). This is explained in greater detail in the referenced paper Miller et al. (2015).

Comment-3: As you state in the methods section, for the propagation of error in the indirect approach, the variances and covariances should be those of the residuals from the fitted models, not of the data themselves. It seems that this is what was done, but lines 5 and 6 of page 770 made me wonder if the variances and covariances of the raw data had been used. Could you clarify this, as this could be an alternative explanation of the larger uncertainties resulting from the indirect approach?

Response-3: The text has been corrected to specify the use of the residual's covariance.

Comment-4: Define f in Equations 2 and 3, and explain exactly what |f| is.

Response-4: |f| is now defined with an explanation for why it is used. Equation 3 was also corrected.

Comment-5: In the cross validation (Table 5), I am not sure why the results for predictions of SOC stock by the indirect approach are omitted. I think the table should include these.

Response-5: Cross-validation was only conducted on the products of the Cubist models themselves. To cross-validate the SOC stock predictions from the indirect approach would be the cumulative cross-validations of the component models. We argue that cross-validations only offer a measure of the models' robustness (how much it would change if different points were used), but does not measure the quality of the best model's performance. For these reasons, calculating a cross-validation of the SOC stock from cross-validations of the component models would be intensive without a large gain in information.

Comment-6: I think it would also be good to provide some validation of the uncertainties…I appreciate the difficulties of validating with a small, clustered dataset such as this, but I think it would be worth including some measure of the adequacy of uncertainty assessment in the cross validation. One possibility is the mean of the theta-statistic, which should be close to 1 (see e.g. Lark, RM. 2000. A comparison of some robust estimators of the variogram for use in soil survey. European Journal of Soil Science, 51, 137-157).

Response-6: We contend that the cross-validation statistics are not really a validation of the model based on all of the sample points, especially for MLR models of soil systems. For this reason, our focus in this paper is to compare the results of the two modelling approaches with each other. Nonetheless, the change in the MAE between the models used (based on all points) and the cross-validation models are compared. Additional text has been added to point out that the stability of the MAEs suggests that the estimated uncertainties are also robust.

Comment-7: It is quite interesting that although the ME for all subsoil component variables was <1, the resulting predictions of SOC stock gave a ME of 1.67. This is worth commenting on in Section 3.1.1.

Response-7: We agree this is an intriguing result and text has been added to section 3.1.1. to highlight it.

Comment-8: I think that the residuals for all variables are assumed normal…however, depending on the dataset, it may be more appropriate to model log SOC % as normally distributed. Some comment about this, and about the effect that this could have on predictions and uncertainties in the indirect approach could be useful.

Response-8: This is indeed an important point. Text, along with a table of residual skewness coefficients, has been added to describe the distribution of the residuals and their potential effect on the results.

Comment-9: Is a conservative estimate of the spatial distribution the best thing? The most conservative would be to use the mean across the entire study area, but this would not be very useful. I am not sure whether the paper is recommending that the more conservative approach should be used, or just saying that the direct approach is more conservative than the indirect approach.

Response-9: Recognizing that different situations will have different needs, we were careful to chose the term 'conservative' to avoid judging which approach may be best for a given set of goals or purpose. Specifically, sometimes representing variation can be more useful than minimizing the amount of error, and vice versa. A sentence has been added to the conclusions to emphasize this point.

Comment-10: What exactly is meant by the 'spatial association approach'?

Response-10: Spatial association is a term parallel to spatial autocorrelation. Like the specific method of kriging is often described when applying a spatial autocorrelation approach, the term spatial regression is a popular method for applying the concept of spatial association. We use the term spatial association frequently to emphasize the difference between our study and a similar study that compared different spatial autocorrelation methods. Additional text has been added to clarify spatial association for those who may not be familiar with the term.

Comment-11: Were all soil profiles deeper than 2 m?

Response-11: Thank you for catching this omission. Indeed field logistics prevented the full 2 m from being sampled for some of the profiles and some assumptions needed to be made. Text explaining this has been added.

Comment-12: Page 767, sentence starting on line 27: 'correlation…of R2 = 0.59'. Correlation should be measured by r, not R2…reword this sentence.

Response-12: Wording has been corrected.

Comment-13: Page 768, line 6: direct R2 = 0.14, but in Table 4 is 0.19…is this correct?

Response-13: The typographical error has been fixed.

Comment-14: Figures 2, 3 and 4: I am not sure that the hillshade effect helps. I found it difficult to distinguish between the effect of the hillshade and the SOC stock differences. I would suggest removing this effect.

Response-14: Agreed. The hillshade effect has been removed from the respective figures.

MARKED-UP MANUSCRIPT VERSION

Comparison of spatial association approaches for landscape mapping of soil organic carbon stocks

Bradley A. Miller*, Sylvia Koszinski, Marc Wehrhan, and Michael Sommer

Leibniz Centre for Agricultural Landscape Research (ZALF) e.V., Institute of Soil Landscape Research, Eberswalder Straße 84, 15374 Müncheberg, Germany

*Corresponding author

Email addresses*:* miller@zalf.de (B.A. Miller), skoszinski@zalf.de (S. Koszinski), wehrhan@zalf.de (M. Wehrhan), sommer@zalf.de (M. Sommer)

Abstract

The distribution of soil organic carbon (SOC) can be variable at small analysis scales, but consideration of its role in regional and global issues demands the mapping of large extents. There are many different strategies for mapping SOC, among which are to model the variables needed to calculate the SOC stock indirectly or to model the SOC stock directly. The purpose of this research is to compare direct and indirect approaches to mapping SOC stocks from rule-based, multiple linear regression models applied at the landscape scale via spatial association. The final products for both strategies are high-resolution maps of SOC stocks (kg m^{-2}), covering an area of 122 km^{2}, with accompanying maps of estimated error. For the direct modelling approach, the estimated error map was based on the internal error estimations from the model rules. For the indirect approach, the estimated error map was produced by spatially combining the error estimates of component models via standard error propagation equations. We compared these two strategies for mapping SOC stocks on the basis of the qualities of the resulting maps as well as the magnitude and distribution of the estimated error. The direct approach produced a map with less spatial variation than the map produced by the indirect approach. The increased spatial variation represented by the indirect approach improved R^2 values for the topsoil and subsoil stocks. Although the indirect approach had a lower mean estimated error for the topsoil stock, the mean estimated error for the total SOC stock (topsoil + subsoil) was lower for the direct approach. For these reasons, we recommend the direct approach to modelling SOC stocks be considered a more conservative estimate of the SOC stocks' spatial distribution.

Keywords: digital soil mapping, organic carbon, spatial association, estimated error, uncertainty

Highlights

1. Spatial association methods for mapping SOC stock directly and indirectly were compared.
2. Data mining produced models that could be interpreted by expert knowledge.
3. The indirect approach map had greater spatial variation and higher R^2 values.
4. The direct approach map had less spatial variation and a lower total estimated error.

1. Introduction

The storage of carbon in soil is a critical point of information for several environmental issues. Globally, soil carbon, which is about 60% organic carbon, accounts for 3.3 times more carbon than that found in the atmosphere (Lal, 2004). The high amount of carbon stored in the soil~~,~~ makes soil carbon an important factor for understanding the carbon cycle and dynamics influencing global climate change (Grace, 2004; Johnston et al., 2004; Powlson et al., 2011). In addition, higher concentrations of soil organic carbon (SOC) are associated with better water storage capacity, regulation of nutrients, and stabilization of soil aggregates resulting in improved soil structure and resistance to erosion (Neemann, 1991; Angers and Carter, 1996; Rawls et al., 2003; Snyder and Vazquez, 2005; Johnston et al., 2009; Kay, 1998; Wilhelm et al., 2004). Each of these factors has important roles in issues of water management and crop productivity.

Although SOC management has far reaching implications, the distribution of SOC is highly variable and dynamic at the field-scale (Cambardella et al., 1994; McBratney and Pringle, 1999; Walter et al., 2003; Kravchenko et al., 2006b; Simbahan et al., 2006). Differing conditions, such as hydrology or management practices, greatly impact the SOC content (Kravchenko et al., 2006a). The combination of global implications and high spatial variability make high-resolution maps of SOC for large extents desirable for both policy decisions and land-owner response. This situation creates the need to accurately and efficiently assess the spatial distribution of SOC stocks at a high-resolution. High-resolution mapping captures information essential for assessing field-specific conditions, which can later be aggregated as need to provide summary information.

Many studies have tested a variety of strategies for predicting the spatial distribution of SOC (Minasny et al., 2013 and references therein). The various studies on SOC mapping have analyzed different soil depths, which has large implications for the consideration of the complete SOC stock (Richter and Markewitz, 1995; Batjes, 1996, Jobbágy and Jackson, 2000; Sombroek et al, 2000; Schwartz and Namri, 2002; Meersmans et al., 2009). For example, some have focused on spatially modelling the topsoil to depths of 20-30 cm (e.g. Ungaro et al., 2010; Zhang et al., 2010; Martin et al., 2011). Other variations of strategies for digital SOC mapping differ in which variables are modelled in order to predict SOC. For instance, some studies have modelled the SOC stock (e.g. kg m^{-2}, T ha^{-1}, kg m^{-3}) directly (Simbahan et al., 2006; Lufafa et al., 2008; Nyssen et al., 2008; Mishra et al., 2010; Phachomphon et al., 2010; Kempen et al., 2011), while others have separately modelled the variables needed to calculate the SOC stock and then combined them (Grimm et al., 2008; Khalil et al., 2013; Lacoste et al., 2014). The usual component variables are total bulk density (BD), particles > 2 mm (SK), SOC concentration (SOC$_{\%}$), and stock thickness (H), which are then combined by:

$$SOC_{stock} = \left(\frac{\mathrm{SOC}_{\%}}{100}\right) * (BD * 1000) * \left(\frac{100-SK}{100}\right) * H \qquad (1)$$

where, SOC$_{stock}$ is in kg m^{-2}, SOC$_{\%}$ is in percent, BD in g cm^{-3}, SK in percent, and H in m.

Irrespective of the approach used, an important output of digital soil mapping is a measure of uncertainty. Orton et al. (2014) compared uncertainties resulting from directly modelling the SOC stock (direct = calculate-then-model) with modelling component variables for calculating the SOC stock (indirect = model-then-calculate), based on geostatistical approaches that rely on spatial autocorrelation. In the present study, we made a similar assessment for rule-based, multiple linear regression (MLR) models, which rely on spatial association.

With the spatial association (i.e. spatial regression) approach to soil mapping, the empirical model error can be transferred along with the model itself (Lemercier et al., 2012). For digital soil mapping, Malone et al. (2011) adapted the Shrestha and Solomatine (2006) approach for empirically summarizing model error and extending that information to prediction areas. In those previous studies, areas expected to have similar errors were grouped by cluster analysis. Because similar sites are already grouped together in rule-based, MLR models, the estimated errors can be applied to the areas meeting the same rule conditions and thus mapped. The ability to map predictions of soil properties and the confidence in those predictions via spatial association is important for landscape to national extents because of the common limitation of sampling density (Martin et al., 2014).

The purpose of this study was to compare the maps of SOC stocks produced from direct and indirect modelling approaches, using rule-based MLR. The resulting maps were compared in terms of their predicted spatial patterns, coefficient of determination (R^2), as well as the magnitude and spatial distribution of the estimated errors. The predictors selected for the models via the data mining procedure were evaluated in the context of known landscape processes. In addition, the separate assessment of topsoil and subsoil stocks tested the models' ability to predict SOC storage at depths to two meters.

2. Methods

2.1. Study Area and Sampling

A dominantly agricultural area located near Wulfen, Saxony-Anhalt, Germany, which has been examined by several previous studies (Selige et al., 2006; Brenning et al., 2008; Kühn et al., 2009; Migdall et al., 2009), was selected for this research. The mapping area extends from 11.86°N, 51.74°E to 11.96°N, 51.90°E (Figure 1), covering a total area of 122 km^2. The landscape includes hummocky till plain, outwash plain, loess, and a broad floodplain (Königlich Preußische Geologische Landesanstalt, 1913a, b). The study area is dominated by Calcaric Cambisols and Luvic Phaeozems, while the depressional area in the floodplain is primarily Dystric Gleysols (European Commission, 2014). Between 2005 and 2006, 117 locations were sampled from a variety of landscape positions in 12 different agricultural fields, covering the known feature space for agricultural land in this area. Because all models were calibrated and validated on these samples, evaluation of the resulting maps focused on areas with similar land-use (i.e. water bodies and urban areas excluded). Ten of the sample points, also spread across the feature space, were of repeated locations (within 2 m of original), which helped to insure that random error was reflected in the assessment of estimated error.

Soil horizons identified in the field were sampled at each sampling location. To avoid biases from horizon classifications and to focus on the two major process zones for SOC, the soil profile of two meters was divided into topsoil and subsoil stocks. The division was defined by the largest decrease in $SOC_{\%}$, as determined by lab analysis, between field identified horizons. Not all profiles were able to be sampled to the full depth of two meters. In those cases, the properties of the sampled subsoil were assumed to be representative of the remaining depth. Data for the horizons within each stock were combined using a thickness-weighted mean, as appropriate. Descriptive statistics for these observation points are provided in Table 1.

2.2. Modelling

Models for each of the target variables were generated using the Cubist 2.08 software (Quinlan 1992, 1993, 1994). Previous studies have demonstrated the utility of this tool for digital soil mapping (Bui et al. 2006; Minasny and McBratney, 2008; Adhikari et al., 2013; Lacoste et al., 2014). Cubist uses a data mining algorithm to build two-tiered models. The top level consists of a series of conditional rules that can utilize both continuous and categorical predictors. For each rule, a MLR equation is produced for predicting the target variable. Cubist's process for selecting predictors and building the models is described in Quinlan (1993) and Holmes et al. (1999) and will not be repeated here. One advantage of this approach is the interpretability of the produced model, which allows the modeler to assess relationships between the model and physical processes (Bui et al., 2006).

The results of the data mining process are dependent upon the predictors made available to the data mining software. For this reason, we used the large predictor pool method described by Miller et al. (2015) to identify the optimal models for each of the respective target variables. That method includes a multiple pass test, which reapplies the Cubist algorithms to the limited pool selected by the previous run. This helps to insure that the selected predictors have been optimally reduced by the Cubist software, decreasing the concern of overfitting. The predictor pool for this study included 410 base maps covering the full extent of the study area (Table 2). These base maps consisted of a legacy geologic map, a variety of remote sensing/spectral products, and digital terrain analysis (DTA). The spectral products ranged from four bands of Ikonos data to a variety of Landsat data collected at different times in 2006. DTA was conducted on a 2 m resolution, digital elevation model (DEM), created from LiDAR data that was also collected in 2006. The DTA base maps included land-surface derivatives based on a wide range of analysis scales (a-scales) and a suite of hydrologic indicators. Land-surface derivatives were calculated in GRASS 6.4.3 (Geographic Resources Analysis Support System, grass.osgeo.org) and ArcGIS 10.1 (www.esri.com/software/arcgis). Hydrologic indicators were calculated using SAGA 2.1.0 (System for Automated Geoscientific Analysis, http://www.saga-gis.org/en/index.html).

The predictors selected by the Cubist software were then used as base maps to generate maps of SOC_{stock}. Using the raster calculator in ArcGIS 10.1, the base maps were combined according to the MLR equations produced by Cubist. When base maps of different resolutions were combined, the finest resolution was maintained. The respective MLR equations were only applied in the areas that met the conditions of the Cubist model's first tier. The first experimental approach used this method to directly map SOC_{stock} from the SOC_{stock} calculated at each sample point. The second experimental approach used this method to map each of the component variables. These modelled variables were then used as base maps to create a SOC_{stock} map. The raster calculator was then again used to combine the component variables, but this time according to equation 1. For both experimental approaches, the topsoil and subsoil were mapped separately. After the respective SOC_{stock} maps were produced, they were added together to create total SOC_{stock} maps.

Within the extent of the study area, there were a few areas with conditions outside the range observed in the point samples. In these limited cases, extreme predictor values produced model predictions of target variables either far below or above the ranges observed for the respective target variables. To address this issue, spatial predictions were limited to be within 10% of the observed target variable minimum and maximums.

2.3. Propagation of Error

For each of the model rules, estimated error was calculated based on the internal fit of the MLR to the data classified within that rule. This estimation provided a measure for the respective uncertainty under each rule. The conditions for the respective rules were used to spatially classify the base maps, thus allowing the estimated errors to be mapped. Measurement error, positional error, and limitations of the model to predict the target variable were all empirically encapsulated by the estimated error.

When the target variable was the end product, the uncertainty was simply represented by the estimated error. However, when multiple variables were modelled and subsequently used to calculate the final product, the estimated errors of the component variables propagated through the combination of those variables in the function. In order to map estimated error for the indirect approach of modelling SOC_{stock}, estimated error maps were produced for each of the component variables. These error estimation maps were then combined using standard equations for propagation of error (Mardia et al., 1979; Taylor, 1997; Weisstein, 2014). Although potentially biased by the approximation to a first-order Taylor series expansion, simplified equations for error propagation are more practical and are regularly used in engineering and physical science applications (Goodman, 1960; Ku, 1966). Because covariance between variables has the potential to impact the estimation of SOC_{stock} (Panda et al., 2008; Goidts et al., 2009), we did not assume the variables were independent. The observed residual covariance was thus used to modify the estimated error within the standard equations for propagation of error by multiplication,

$$\sigma_f \approx |f|\sqrt{\left(\frac{\sigma_A}{A}\right)^2 + \left(\frac{\sigma_B}{B}\right)^2 + 2\frac{cov_{AB}}{AB}} \qquad (2)$$

and by addition,

$$\sigma_f \approx \sqrt{{\sigma_A}^2 + {\sigma_B}^2 + 2cov_{AB}} \qquad (3)$$

where *f* is the result of the original function (to convert from relative to estimated error), A and B are the real variables, with estimated errors σ_A and σ_B, and their residuals' covariance cov_{AB}. In order to calculate a predicted relative error (e.g. $\frac{\sigma_A}{A}$) at unsampled locations, the predicted variable was assumed to accurately represent the variable's magnitude.

Locations with small ratios between estimated error and predicted values together with large, negative covariances had the potential to produce a calculation taking the square root of a negative. This issue was addressed by not considering the covariance in those limited circumstances. While this solution may have led to an overestimation of error, it provided a means to mathematically calculate estimated error without declaring it to be zero.

3. Results

3.1. Models

3.1.1. Model building and fitting performance

Explicit models were obtained for each of the component variables needed to calculate SOC_{stock} indirectly and for predicting SOC_{stock} directly. Models for predicting component variables used a higher quantity of predictors for each of the respective models than the direct modelling approach (Table 3). With the exception of $SOC_{\%}$, the models for component variables included a combination of DTA and spectral variables. The $SOC_{\%}$ models relied solely on DTA predictors for both stocks, but with additional spatial partitioning by geologic map units for the topsoil model. The models for directly predicting the SOC_{stock} used only three DTA predictors for the topsoil and only four Landsat predictors for the subsoil.

Fitting performances for the component variable models were better than the fitting performances for the direct modelling of SOC_{stock} (Table 4). For the component variables, R^2 values of subsoil models were only slightly less than the topsoil models. $SOC_{\%}$ was the exception by having the lowest fitting performance for the subsoil stock ($R^2 = 0.55$), while the model for the $SOC_{\%}$ topsoil was able to fit observations with an R^2 of 0.86. However, it was the aim of this research to examine if the performance of the models was maintained through the calculation of SOC_{stock}.

Comparison of the SOC_{stock} predictions by the indirect approach to observed values showed better performance for the topsoil stock (R^2 = 0.73) than for the subsoil stock (R^2 = 0.34). Fitting performance for directly modelling SOC_{stock} showed the same pattern, but was lower than the indirect approach for both stocks. Analysis of the direct approach's ability to fit observed values yielded an R^2 of 0.58 for the topsoil and 0.~~14~~ 19 for the subsoil.

In general, calculated model efficiencies (ME) showed that the respective models reduced the mean absolute error (MAE) to about half the MAE that would result from simply using the mean of all points as the prediction. The $SOC_{\%}$ model for the topsoil improved upon the mean model more than the other MLR models with a ME of 0.34. However, an intriguing result is the lack of model efficiency for the indirect modelling of the subsoil's SOC_{stock}. Despite the component models all having MEs well below one, the indirect approach did not improve upon the mean model for predicting the subsoil SOC_{stock}. Although the ME of the direct model for subsoil SOC_{stock} was also not as good as the other models, it was still an improvement over the mean model.

3.1.2. Model Robustness

It is common for digital soil mapping models to be evaluated by cross-validation procedures. However, in the context of this study, the meaning of such an analysis has less utility. Higher sample density increases the robustness of the model (Minasny et al, 2013); thus the popularity of cross-validation procedures over independent validation procedures in order to maintain more points in the calibration set. However, the model generated for each cross-validation run is different because of differences in calibration sets. The performance of each run is dependent on the randomly selected calibration points' ability to represent the variation in the remaining validation points. For a simple data trend, a single outlier would have minimal effect because only the runs in which it is included in the validation set – and not used in calibrating the model – would have lower performance values. However, in a complex landscape where similar soil properties can result from different combinations of factors, the concept of an outlier has many more dimensions (Johnson et al., 1990; Phillips, 1998). A point with a similar value can be an outlier by being a product of a different set of factors. In other words, the problem of induction continues to apply in predictive soil mapping. Further, in the context of error propagation, the error estimation from the actual model used seems more appropriate than the mean of error estimations from a series of less robust models.

Nonetheless, the models in this study were cross-validated using the k-fold method with 10 iterations. The R^2 was naturally reduced in the cross-validation analysis, but the ~~mean absolute error~~ ~~(~~MAE~~)~~ was not as severely affected (Table 5). The R^2 values for the respective models all decreased greatly in the cross-validation, except for the topsoil $SOC_{\%}$ and the subsoil SOC_{stock} models. The subsoil SOC_{stock} model already had a low R^2 value for the internal fit. In contrast, the MAEs for the cross-validation of the models were not increased enough to present a practical problem. The relative stability of the MAEs also suggests that the estimated uncertainties are also robust. For example, the MAE for both stocks of BD only increased 0.03 g cm^{-3}. Also, the MAE for $SOC_{\%}$ only increased 0.13% and 0.03% for the topsoil and subsoil, respectively. Similarly, the MAE for the direct SOC_{stock} model increased 0.67 kg m^{-2} and 0.05 kg m^{-2} for the topsoil and subsoil, respectively. The MAE for the models of stock H and SK did increase more in cross-validation. However, they had a minor impact on the indirect modelling of SOC_{stock}. The increase of 5.9 cm for the topsoil H MAE was only a shift of the depth estimated by topsoil or subsoil models. The larger MAE for SK was more of an issue for the subsoil. However, the majority of the samples had SK below 5%, leaving most of the error due to the difficulty in predicting the limited areas of high SK. While it was possible that a different sampling design could have improved the R^2 values for cross-validation, they are not always practical for landscape-scale mapping.

3.1.3. Comparison with previous studies

It is difficult to compare results between SOC mapping studies due to differences in study areas and strategies for defining SOC_{stock} (i.e. map extent and resolution, sampling density, and consideration of depth). Further, the differences between and variability within methods for estimating component variables for calculating SOC_{stock} can have a large impact on results, especially bulk density (Liebens and VanMolle, 2003; Schrumpf et al., 2011) and $SOC_{\%}$ (Lowther et al., 1990; Soon and Abboud, 1991; Sutherland, 1998; Bowman et al., 2002). Also, because model performance is dependent upon the provided predictors, results of different studies can vary based on the predictors available to and derived by the modeller (Miller et al., 2015). However, because the area in this study has been used for several previous studies, some comparisons between methods can be made.

Kühn et al. (2009) examined many of the same samples used in this study and found a ~~correlation~~ coefficient of determination between soil electrical conductivity and soil organic matter to a 1 m depth (kg m^{-2}) of $R^2 = 0.59$. Although a slightly different calculation, that ~~correlation~~ coefficient of determination is similar to this study's direct model of topsoil SOC_{stock} ($R^2 = 0.58$), which used three DTA predictors. However, for the topsoil, the indirect approach in this study produced a SOC_{stock} model with less estimated error and an R^2 of 0.73. The Kühn et al. (2009) study usually included depths that this study defined as subsoil, where the models in this study did not perform as well (direct R^2 = 0.~~14~~19, indirect $R^2 = 0.34$).

For the same area as this study, Selige et al. (2006) compared MLR and partial least-square regression for predicting $SOC_{\%}$ from hyperspectral data with a 6 m spatial resolution. Although the study by Selige et al. (2006) utilized a higher spectral resolution, the MLR models produced by both that study and the present study had R^2 of 0.86 for the topsoil $SOC_{\%}$. In the present study, Cubist was able to compensate for the limited spectral information by utilizing several DTA predictors that were available at a high spatial resolution.

3.2. SOC_{stock} maps

Application of the obtained models and aggregation of the component variable maps by equation 1 produced maps of predicted SOC_{stock} for the topsoil and subsoil (Figures 2 and 3). The respective topsoil and subsoil maps were added together to produce a total SOC_{stock} map to a depth of 2 m (Figure 4). Although some field boundaries were observed, the dominant pattern appeared to be associated with terrain features. This interpretation was supported by the number of DTA predictors selected by Cubist for many of the models. However, it would not have been safe to assume this pattern from the list of selected predictors alone. Certain predictors (i.e. spectral data reflecting land use patterns) could have dominated calculations without being the most frequently selected category of predictors.

The map derived from the direct approach for modelling the topsoil SOC_{stock} emphasizes drainageways. Whereas the map derived by the same approach for the subsoil SOC_{stock} reflects more patterns of land use, especially in the uplands in the southern part of the study area. The topsoil SOC_{stock} map based on the indirect approach has similar overall patterns to the direct approach's map. However, both the topsoil and subsoil maps produced by the indirect approach display greater spatial variation.

Patterns in the topsoil SOC_{stock} map, based on the indirect approach, mostly coincide with terrain features, but do contain some transitions that align with field boundaries. The corresponding map for the subsoil reflects patterns of microtopography and slope gradient. Larger values for the subsoil SOC_{stock} are predicted by the indirect approach for local lows in elevation (smaller a-scales). Predictions of larger subsoil SOC_{stock} on steeper slopes result from the modelling of thinner topsoil stocks in these areas and the consistent calculation of a 2 m profile. Consequently, the subsoil is calculated to be thicker in these areas, substantially increasing the subsoil SOC_{stock} prediction compared to other areas of the subsoil.

Maps derived by both approaches for the total SOC_{stock} primarily reflected patterns from the topsoil maps because of the higher concentration of SOC that defined the topsoil stock.

Nonetheless, modelled storage for the subsoil stock contributed about one-third of the prediction of total SOC_{stock} and recognized additional complexity in the SOC landscape. Despite the greater variation in the indirect approach's prediction of SOC_{stock}, the difference between estimates of total SOC_{stock} by the two approaches were within 5 kg m^{-2} for the majority of the map area (Figure 5). Also, the summed SOC_{stock} for the study area was only 6% more for the indirect (1.9 Mt) versus the direct (1.8 Mt) approach. The mean SOC_{stock} estimate for the study area by the direct approach was 14.7 kg m^{-2}, whereas the indirect approach estimated 15.7 kg m^{-2}.

These aggregated landscape estimates agreed with those made by the Harmonized World Soil Database (HWSD; FAO/IIASA/ISRIC/ISSCAS/JRC, 2012) for this area. The HWSD estimated several soil properties from taxonomic pedotransfer functions for static topsoil (0-30 cm) and subsoil (30-100 cm) depth zones. Within the area of the present study, the HWSD has a cell resolution of approximately 765 m. Calculating SOC_{stock} from that data yielded a mean of 8.8 kg m^{-2}. Assuming the characteristics of the subsoil to 100 cm extended to 200 cm, the mean SOC_{stock} would be 15.3 kg m^{-2}.

3.3 Error estimations

The mapping of estimated errors based on the conditions of rules generated by Cubist resulted in a spatial representation of uncertainty (Figure 6). In order to calculate the final estimated errors for the indirect approach, estimated errors for models of component variables were combined spatially by equations ~~3~~ 2 and ~~4~~3. Due to the known covariance of component variables, the observed covariance of the residuals was included in the calculation of error propagation through the calculation of the total SOC_{stock}. Inclusion of covariance reduced relative error estimates in the topsoil because increases in residuals for BD coincided with decreases in the residuals for percent fine-earth, increases in fine-earth BD residuals coincided with decreases in $SOC_{\%}$ residuals, and increases in SOC content (kg m^{-3}) residuals coincided with decreases in stock thickness residuals. The influence of covariance was ~~not~~mostly the same in the subsoil calculations. The exception was a positive covariance between the residuals for modelling BD and the percent fine-earth. ~~With the~~

~~exception of the covariance between fine-earth BD and SOC$_{\%}$, which was very small, subsoil covariances were positive.~~ ~~However, overall~~Nonetheless, the covariances were relatively small with respect to the estimated errors and therefore had a minimal impact on the final calculation of estimated error.

The application of error estimates based on the full range of predicted values in a rule zone to small values in that zone yielded extremely high relative error values. Although the areal extent for this type of situation was very limited, the issue needed to be addressed in order to maintain the readability of the attribute scale. Therefore relative error was capped at one for the original relative error grids, but not thereafter for the calculation of error propagation.

Despite not having as strong of a fitting performance as the indirect approach, the direct approach had lower estimated errors for greater extents of the study area. The mean estimated error for the total SOC$_{stock}$ map derived by the direct approach was 2.81 kg m^{-2}, compared to 8.17 kg m^{-2} for the indirect approach. This behavior in the models may be explained by the negative covariance between the residuals for many of the variables influencing the SOC$_{stock}$. The observed covariances did reduce the calculation of error through propagation. However, they did not reduce the estimated error for the indirect approach to as low as the estimated error based on the direct modelling approach. It is also useful to note that the residuals for modelling SK and SOC$_{\%}$ ~~were~~had a negative and positive skewed, respectively, for both stocks (Table 6). ~~However, for the~~Of the residuals ~~of~~for the final prediction of SOC$_{stock}$, regardless of approach or stock, only the indirect model for the subsoil had strongly skewed residuals. This suggests that ~~the~~ error for the indirect model of the subsoil SOC$_{stock}$ may have been overestimated.

The spatial distribution of model rules was an important factor in the resulting maps' estimated error. The models for the direct approach used fewer rules than the component variable models, resulting in less spatial variation of the estimated error. However, variation in predicted values did introduce additional spatial variation to the mapping of relative error. Nonetheless, the map of relative error from the indirect approach was more complex than that resulting from the direct approach. In addition to using more rules for each model, the combined relative estimated error for the indirect approach was further tessellated by the unique intersections of the different spatial distributions of the rules for each component variable model.

4. Discussion

4.1. Predictor selection

4.1.1. Review of relationships between predictors and environmental conditions

Spectral predictors from satellites such as Ikonos and Landsat have most commonly been used to detect characteristics of land use, vegetation, and soil water content (Bannari et al., 1995; Xie et al., 2008). However, they have also been used to detect mineralogy on sparsely vegetated areas (Mulder et al., 2011). Although Ikonos has a finer spatial resolution, it is limited to three bands (band 1 = blue, band 2 = green, and band 3 = red) in the visible spectrum, plus a near infrared band (band 4 = NIR). Landsat provides additional bands in the shortwave infrared (band 5 = SWIR-1; band 7 = SWIR-2) and thermal infrared (band 6 = TIR). The relative reflectance of a single band can be used to distinguish landscape conditions. For example, the green band can be used to distinguish different vegetation from bare soil. However, combinations of bands - particularly including the red and NIR bands - have been even more useful for distinguishing the spectral signature of different land uses (Richards, 2006) and the condition of the vegetation (Ashley and Rea, 1975; Myneni et al., 1995; Rasmussen, 1998; Daughtry, 2001; Hatfield et al., 2008). Additional use of TIR emission would resemble methods such as the Surface Temperature/Vegetation Index for estimating soil moisture (Bartholic et al., 1972; Heilman et al., 1976; Carlson et al., 1994; Li et al., 2009; Petropoulus et al., 2009). Similarly, use of SWIR wavelengths in concert with red and infrared ~~red~~ bands would be a way of compensating for the changing effect of soil reflection in dry to wet conditions (Huete, 1988; Lobell and Asner, 2002). Relationships between bands in the visible to SWIR range have also been used to predict SOC$_{\%}$ and its biochemical composition (Bartholomeus et al., 2008; Gomez et al., 2008; Stevens et al., 2010).

Spectral predictors have been used for both classification of discrete phenomenon and quantification of continuous phenomenon on the landscape. Because of the rule-based MLR structure of the Cubist models, spectral predictors used for conditional rules were more likely to be distinguishing discrete features (e.g. vegetation/land use type) than when used within an MLR equation. Continuous features (e.g. vegetation health) were more likely to be represented in MLR equations.

DTA predictors in this study were all derived from the LiDAR data for elevation. The land-surface derivatives (e.g. slope gradient, relative elevation) described the surface geometry with which the climate interacts. For example, aspect has been shown to influence the amount of solar insolation a hillslope receives (Hunckler and Schaetzl, 1997; Beaudette and O'Geen, 2009). The surface geometry is also known to direct water flow, which affects erosion processes and groundwater recharge (Huggett, 1975; Zevenbergen and Thorne, 1987). Hydrologic predictors (e.g. flow accumulation, catchment slope) provided additional information about the relative volume and energy that the water flow may have (Moore et al., 1991; Wilson and Gallant, 2000).

4.1.2. Topsoil model predictors

All of the topsoil models generated by Cubist relied on DTA predictors the most. Of those predictors, different a-scales of relative elevation, topographic position index (TPI), and aspect were the most commonly used. With the exception of the direct SOC$_{stock}$ model, every topsoil model also included one or two predictors indicative of flow accumulation (i.e. flow path length, SAGA wetness index, or modified catchment area).

Aspect at different a-scales influenced predictions for three of the indirect topsoil models. The Cubist generated model identified decreasing topsoil SOC$_{\%}$ on more north facing slopes (155 m a-scale), which corresponds with a potential decrease in plant productivity due to less solar insolation. Aspect (215 m a-scale) was also used to predict higher topsoil BD on south to west facing slopes, especially on topographic (2000 m a-scale) and micro-topographic (20 m a-scale) highs. Additionally, aspect at a variety of a-scales was used to predict decreasing topsoil SK for low TPI areas facing southeast to southwest. Together, these models suggested a pattern of increased erosion and deposition along the southern sides of hillslopes. This type of pattern has been observed before in other landscapes and has been attributed to topo-climatic differences such as exposure to storms, differences in temperature regime, rainfall effectiveness, or vegetation density (Kennedy, 1976; Churchill, 1981; Cuff, 1985; Weaver, 1991).

Although DTA parameters dominated the topsoil models, their predictions were often modified by spectral variables. For example, the primary distinction for predicting topsoil H was between low and high relative elevations. Low relative elevations had a mean topsoil H that was about 20 cm thicker than high relative elevations (1,100 m a-scale). Within most MLR equations, however, predictions were increased by less blue and more green reflectance in early July. This combined use of blue and green bands indicated increasing topsoil H with more productive vegetation on wetter soils. In summary, the dominant pattern identified by the model was between high-low ground (Bushnell, 1943; Sommer et al., 2008), but the degree of topsoil thinning or thickening was predicted by the vegetation’s response to soil conditions.

Cubist selected a much simpler combination of only DTA predictors to directly model the topsoil $\mathrm{SOC}_{\mathrm{stock}}$. In general, the model predicted increasing $\mathrm{SOC}_{\mathrm{stock}}$ with decreasing vertical distance to channel. Areas low in relative elevation (1,100 m a-scale) and not far above the channel network were predicted to have the largest $\mathrm{SOC}_{\mathrm{stock}}$. However, for areas low in relative elevation, but sufficiently above the DEM based channel network, the model predicted the opposite trend of the $\mathrm{SOC}_{\mathrm{stock}}$ *decreasing* with decreasing vertical distance to channel. This pattern identified by the model may be explained by a corresponding pattern observed in the model for the topsoil H. In that model, areas low in relative elevation (1,100 m a-scale) were predicted to have some of the thickest topsoil stocks. However, within a few of those zones the modelled topsoil H decreased with decreasing relative elevation and TPI. This trend in the observed data, as detected by Cubist, was potentially caused by an eroding out of topsoil sediments closer to the center of drainageways. In which case, the vertical distance to channel – used in the topsoil SOC_{stock} model - may have been more an indicator of proximity to the channel than wetness; the threshold was only 0.5 m above the channel modelled from the DEM. Predictors related to surface flow energy would have been expected to be better predictors of this kind of process. However, the upslope drainage network for much of the map area extended beyond the boundaries of the available data. Thus the use of local elevation data may have been a better proxy in this case, compared to the predictors calculated from truncated watersheds.

4.1.3. Subsoil model predictors

With the exception of $SOC_{\%}$, the subsoil models all used several predictors from Landsat. Selection of Landsat predictors for subsoil models suggested that vegetation characteristics or surface soil moisture at different times of the year indicated subsoil conditions. In contrast, the subsoil $SOC_{\%}$ model's complete dependence on DTA predictors suggested that soil property was mostly related to hydrology and that vegetation had little response to or effect on the SOC content in the subsoil.

An example of spectral predictors detecting vegetation characteristics that likely reflected subsoil conditions was the subsoil SK model. All of the MLR equations were strongly influenced by the predictors of stream power, catchment slope, or SAGA wetness index. However, the ~~skeleton~~ SK predictions were modified by green reflectance in June and additional Landsat predictors collected at different times of the year that related to the vigor of the vegetation. The weaker or drier the vegetation appeared, the higher the prediction of SK content in the subsoil. Assuming soil moisture conditions did not reach detrimental levels that year, these patterns fit known relationships between particle size, soil drainage, and timing to crop maturity (Day and Intalap, 1970; Rawls et al., 1982).

The generated model for subsoil BD most likely utilized a relationship with soil moisture as detected by spectral predictors. In all areas, the MLR equations decreased predictions of subsoil BD with increasing reflectance in the blue and SWIR-1 bands along with increasing emission in the TIR band. Increases in the normalized difference vegetation index (NDVI) were used to slightly increase predictions of subsoil BD. The use of the NDVI to offset the decreasing BD predicted by the other Landsat predictors suggested those variables were indicating soil moisture conditions. Locations that are wetter due to surface runoff would have a greater potential for organic material to be translocated deeper in the soil profile (Schaetzl, 1986; Schaetzl, 1990). Also, the association of wetter environments with cooler temperatures and anaerobic conditions would also inhibit decomposition (Gates, 1942; Krause et al., 1959; Frazier and Lee, 1971).

The subsoil $SOC_{\%}$ model was different than the other subsoil models generated. Instead of selecting spectral predictors, the subsoil $SOC_{\%}$ model relied solely on DTA predictors. The model predicted the highest subsoil $SOC_{\%}$ on steeper mid-slopes. The pattern of increasing subsoil $SOC_{\%}$ from the upper to middle slope fit the landscape translocation model proposed by Sommer et al. (2000). In that study, the $SOC_{\%}$ in the Bh horizon increased from the upper slope to the midslope due to lateral translocation. Different than the pattern identified in the present study, the data in Sommer et al. (2000) showed a continued increase in the $SOC_{\%}$ of Bh horizons in the downslope position. However, this contradiction may be partially explained by aggradation where the slope gradient declines and the topsoil stock has been overthickened by developmental upbuilding (McDonald and Busacca, 1990; Almond and Tonkin, 1999). Also, lateral flow would be expected to return closer to the surface at downslope positions. In Sommer et al. (2000), while the upslope and midslope profiles had E horizons separating the Bh from A horizons, the downslope Bh horizons were exceptionally thick with little to no division between them and the A horizon. In that situation, the definition of topsoil used in the present study would have grouped the downslope Bh horizons into the topsoil stock. Therefore, the Cubist generated model may have been a simplification of the complex interaction between topography and lateral flow depth and direction.

The rule groups for subsoil $SOC_{\%}$ also differentiated for the plan curvature where the slope gradient was not too high and the stream power index (SPI) was not too low. Concave plan curvatures (138 m a-scale) were predicted to have increasingly higher and convex plan curvatures were predicted to have increasingly lower subsoil $SOC_{\%}$. This relationship with plan curvature matches patterns of water movement identified to be important to soil formation by Huggett (1975), where convergent footslopes have the highest deposition rates (Pennock and De Jong, 1987). Assuming the absence of any restrictive layer below, areas with the highest sediment deposition rates would be expected to also have the highest volume of water infiltration.

The Cubist generated model for predicting the subsoil SOC_{stock} was simpler than any of the indirect component models. It used only one MLR equation to relate red and infrared predictors to subsoil SOC_{stock}. This model predicted more SOC_{stock} storage with increasing reflectance in the red and SWIR-2 bands along with increasing emission in the TIR band – primarily captured on 6 July. Of these variables, model predictions were dominated by increasing reflectance in the red band increasing the estimated subsoil SOC_{stock}. This suggested less productive vegetation corresponding with larger subsoil SOC_{stock}. This trend was counter to the patterns observed in the topsoil models, but was sensible in the context of how the subsoil stock was defined for this study. Although the *total* SOC_{stock} was less in areas with lower plant productivity, the subsoil SOC_{stock} was larger relative to other subsoil areas due to the inverse relationship between topsoil and subsoil H used in this study. A thicker topsoil stock would mean a thinner subsoil stock – and vice versa – due to the 2 m depth limit. Regarding the other predictors in this model, increases in SWIR-2 reflectance could have indicated more plant productivity. However, its use with the TIR band suggested that together they were indicators of wetter soil conditions.

4.2. Unconventional predictor selections

The Cubist software made some intriguing selections in regards to predictors that were calculated using alternative approaches. One example of this was the selection of alternative types of aspect predictors. The conversion of aspect to northness and eastness is generally considered to be the preferred method for addressing the circular problem of using aspect as a predictor. In our approach of including many different predictors in the available pool, we also experimented with simply rotating the central angle (position of 0°) to each cardinal direction for creating different aspect predictors. In the models generated for this study, northness and eastness were only selected for the topsoil $SOC_{\%}$ model. In contrast, rotated versions of aspect were selected for the topsoil $SOC_{\%}$, topsoil BD, as well as the topsoil and subsoil SK models.

Another example of an intriguing predictor selection by Cubist was the use of bands from the LandsatLook products. These images were limited to four bands (SWIR-1, NIR, red, and TIR) and were smoothed by an algorithm to facilitate image selection and visual interpretation. Although the USGS does not recommend the use of these files for data analysis, the Cubist data mining found them to be more useful than the data without LandsatLook processing. Most of these selections can be explained by the greater variety of LandsatLook dates provided in the predictor pool. However, there were a few instances where Cubist chose LandsatLook data over the unprocessed version of the same Landsat data.

4.3. Error propagation

Although both the direct and indirect modelling approaches had base maps with a 2 m resolution available to them, the direct modelling approach produced a more generalized SOC_{stock} map. In terms of predicted error, the cost of trying to account for the variation in all of the variables related to the SOC_{stock} appeared to be larger relative errors. The SOC_{stock} model from the direct approach, on the other hand, did not attempt to predict as many variations occurring at small phenomenon scales. Because these very local variations were difficult to predict, the estimated error for the direct approach was less than for the indirect approach for most of the map area. Therefore, it may be appropriate to consider the direct modelling approach to be a conservative approach for estimating the SOC_{stock} for landscapes.

Possible sources of error in the base maps included atmospheric conditions for the satellite data and the estimation of bare earth elevation under dense vegetation for the DEM. Several spectral capture dates were made available in the predictor pool to enable Cubist to not only select the optimal changes in seasonal vegetation characteristics, but to also select the image with minimal noise from atmospheric effects such as clouds. Fewer options were available for DTA predictors, because all DTA predictors needed to be derived from the same high-resolution DEM. The effect of anomalies in the elevation data was more pronounced for larger a-scales. For example, a small forest plot – located roughly between the two larger cities in the center of the map area – had not been fully filtered out by the bare-earth algorithm. Any DTA calculation that included this area in its analysis neighborhood was incorrectly influenced by those elevation values. The impact on this study's models was an increased prediction of SOC_{stock} in the surrounding area.

The error propagation method used in this study could not directly account for errors in the base maps. Instead, it could only quantify the combined model, base map, and target variable error observed at sample locations. Although none of the sample points were in proximity to the before mentioned error in the DEM, this phenomenon of elevation error affecting scale-dependent predictors would have applied universally, even where the error was less obvious. The higher relative error for both mapping approaches in the area surrounding the known problem in the DEM suggested this potential source of error was at least partially accounted for.

5. Conclusions

This study demonstrated the use of spatial association to predict the SOC_{stock} and the estimated error at unsampled locations within a 122 km^2 landscape at a high-resolution. The Cubist data mining software detected patterns in the observed soil data, which was used to predict soil properties in the greater map region. The ability of the available base maps to predict the variation of those soil properties was quantified for each conditional rule of the respective models. The spatial characteristics of the model rules allowed the uncertainty to be mapped along with the target variable prediction.

There were two main advantages to using data mining software to produce relatively simple model structures. First, patterns between the predictors and target variables were objectively identified. Second, the resulting models were simple enough to be interpreted by the user and related to known processes in the soil system. A relationship between selected predictors and known processes provided confidence that their use in the model was not coincidental. The separate modelling of topsoil and subsoil stocks identified a general division between useful predictors for predicting soil properties at different depths. The data mining in this study suggested DTA predictors tend to be most useful for topsoil properties, while spectral characteristics of vegetation and soil moisture tend to be more useful for indicating subsoil properties.

Direct and indirect approaches were tested for predicting the SOC_{stock} with the rule-based, MLR spatial modelling method. Although the spatial patterns in the two maps were generally similar, the indirect approach produced a map with more spatial variation. While attempting to account for more sources of variability resulted in less estimated error for the topsoil (indirect MAE = 1.69, direct MAE = 2.27), the indirect approach had a higher potential for error in the subsoil (indirect MAE = 2.75, direct MAE = 1.37). Because the direct approach accounts for less variation (topsoil: direct R^2 = 0.58, indirect R^2 = 0.73; subsoil: direct R^2 = 0.14, indirect R^2 = 0.34), but also results in a lower total MAE (direct MAE = 3.64, indirect MAE = 4.44), it should be considered a more conservative prediction of the SOC_{stock}'s spatial distribution. The choice of which approach is best will likely depend on a given situation's need to prioritize the representation of spatial pattern or to minimize estimated error.

Acknowledgements

Data used in this research was collected as part of the Preagro project, funded by the German Federal Ministry of Education and Research (BMBF), under grant reference number 0339740/2. We thank Carsten Hoffmann for his suggestions during the development of this study.

References

Adhikari, K., Kheir, R.B., Greve, M.B., and Greve, M.H.: Comparing kriging and regression approaches for mapping soil clay content in a diverse Danish landscape, Soil Science, 178(9), 505-517, doi:10.1097/SS.0000000000000013, 2013.

Almond, P.C. and Tonkin, P.J.: Pedogenesis by upbuilding in an extreme leaching and weathering environment, and slow loess accretion, south Westland, New Zealand, Geoderma, 92(1-2), 1-36, doi: 10.1016/S0016-7061(99)00016-6, 1999.

Angers, D.A. and Carter, M.R.: Aggregation and organic matter storage in cool, humid agricultural soils, in: Structure and Organic Matter Storage in Agricultural Soils, Carter, M.R. and Stewart, B.A. (Eds.), CRC Press, Boca Raton, 193-211, 1996.

Ashley, M.D. and Rea, J.: Seasonal vegetation differences from ERTS imagery, Journal of American Society of Photogrammetry, 41(6), 713-719, 1975.

Bannari, A., Morin, D., Bonn, F., and Huete, A.R.: A review of vegetation indices, Remote Sensing Reviews, 13(1-2), 95-120, doi:10.1080/02757259509532298, 1995.

Bartholic, J.F., Namken, L.N., and Wiegand, C.L.: Aerial thermal scanner to determine temperature of soils and of crop canopies differing in water stress, Agronomy Journal, 64(5), 603-608, 1972.

Bartholomeus, H.M., Schaepman, M.E., Kooistra, L., Stevens, A., Hoogmoed, W.B., and Spaargaren, O.S.P.: Spectral reflectance based indices for soil organic carbon quantification, Geoderma, 145(1-2), 28-36, doi:10.1016/j.geoderma.2008.01.010, 2008.

Batjes, N.H.: Total carbon and nitrogen in the soils of the world, European Journal of Soil Science, 47, 151–163, doi:10.1111/j.1365-2389.1996.tb01386.x, 1996.

Beaudette, D.E. and O’Geen, A.T.: Quantifying the aspect affect: an application of solar radiation modeling for soil survey, Soil Science Society of America Journal, 73(4), 1345-1352, 2009.

Bowman, R.A., Reeder, J.D., and Wienhold, B.J.: Quantifying laboratory and field variability to assess potential for carbon sequestration, Communications in Soil Science and Plant Analysis, 33(9-10), 1629-1642, doi:10.1081/CSS-120004304, 2002.

Brenning, A., Koszinski, S., and Sommer, M.: Geostatistical homogenization of soil conductivity across field boundaries, Geoderma, 143, 254-260, doi:10.1016/j.geoderma.2007.11.007, 2008.

Bui, E.N., Henderson, B.L., and Viergever, K.: Knowledge discovery from models of soil properties developed through data mining, Ecological Modelling, 191, 431-446, doi:10.1016/j.ecolmodel.2005.05.021, 2006.

Bushnell, T.M.: Some aspects of the soil catena concept, Soil Science Society Proceedings, 7(C), 466-476, 1943.

Cambardella, C.A., Moorman, T.B., Novak, J.M., Parkin, T.B., Karlen, D.L., Turco, R.F., and Konopka, A.E.: Field-scale variability of soil properties in central Iowa soils, Soil Science Society of America, 58, 1501-1511, doi:10.2136/sssaj1994.03615995005800050033x, 1994.

Carlson, T.N., Gilles, R.R., and Perry, E.M.: A method to make use of thermal infrared temperature and NDVI measurements to infer surface soil water content and fractional vegetation cover, Remote Sensing Reviews, 9(1-2), 161-173, doi:10.1080/02757259409532220, 1994.

Churchill, R.R.: Aspect-related differences in badlands slope morphology, Annals of the Association of American Geographers, 71(3), 374-388, doi:10.1111/j.1467-8306.1981.tb01363.x, 1981.

Cuff, J.R.I.: Quantifying erosion-causing parameters in a New Zealand watershed, in: Soil Conservation, El-Swaify, S.A., Moldenhauer, W.C., Lo, A. (Eds.), Soil Conservation Society of America, Ankeny, 99-112, 1985.

Daughtry, C.S.T.: Discriminating crop residues from soil by shortwave infrared reflectance, Agronomy Journal, 93, 125-131, doi:10.2134/agronj2001.931125x, 2001.

Day, A.D. and Intalap, S.: Some effects of soil moisture stress on the growth of wheat (Triticum aestivum L. em Thell), Agronomy Journal, 62(1), 27-29, doi:10.2134/agronj1970.00021962006200010009x, 1970.

European Commission: European Soil Database v2, European Soil Data Centre, http://eusoils.jrc.ec.europa.eu, last access: 7 October 2014.

FAO/IIASA/ISRIC/ISSCAS/JRC: Harmonized World Soil Database (version 1.2), FAO, Rome, Italy and IIASA, Laxenburg, Austria, 2012.

Frazier, B.E. and Lee, G.B.: Characteristics and classification of three Wisconsin Histosols, Soil Science Society of America Journal, 35(5), 776-780, doi:10.2136/sssaj1971.03615995003500050040x, 1971.

Gates, F.C.: The bogs of northern lower Michigan, Ecological Monographs, 12, 216-254, doi:10.2307/1943542, 1942.

GLCF: Global Land Cover Facility, University of Maryland, http://glcf.umd.edu/data, last access: 19 February 2014.

Goidts, E., Van Wesemael, B., and Crucifix, M.: Magnitude and sources of uncertainties in soil organic carbon (SOC) stock assessments at various scales, European Journal of Soil Science, 60(5), 723-739, doi:10.1111/j.1365-2389.2009.01157.x, 2009.

Gomez, C., Viscarra Rossel, R.A., and McBratney, A.B.: Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: An Australian case study, Geoderma, 146(3-4), 403-411, doi:10.1016/j.geoderma.2008.06.011, 2008.

Goodman, L.A.: On the exact variance of products, Journal of the American Statistical Association, 55, 708-713, 1960.

Grace, J.: Understanding and managing the global carbon cycle, Journal of Ecology, 92(2), 189-202, doi:10.1111/j.0022-0477.2004.00874.x, 2004.

Grimm, R., Behrens, T., Märker, M., and Elsenbeer, H.: Soil organic carbon concentrations and stocks on Barro Colorado Island – digital soil mapping using random forests analysis, Geoderma, 146(1-2), 102-113, doi:10.1016/j.geoderma.2008.05.008, 2008.

Hatfield, J.L., Gitelson, A.A., Schepers, J.S., and Walthall, C.L.: Application of spectral remote sensing for agronomic decisions, Agronomy Journal, 100(3), S-117 – S-131, doi:10.2134/agronj2006.0370c, 2008.

Heilman, J.L., Kanemasu. E.T., Rosenberg, N.J., and Blad, B.L.: Thermal scanner measurement of canopy temperatures to estimate evapotranspiration, Remote Sensing of Environment, 5(C), 137-145, doi:10.1016/0034-4257(76)90044-4, 1976.

Holmes, G., Hall, M., and Frank, E.: Generating rule sets from model trees, Advanced Topics in Artificial Intelligence, Lecture Notes in Computer Science, 1747, 1-12, doi:10.1007/3-540-46695-9_1, 1999.

Huete, A.R.: A soil-adjusted vegetation index (SAVI), Remote Sensing of Environment, 25, 295-309, doi:10.1016/0034-4257(88)90106-X, 1988.

Huggett, R.J.: Soil landscape systems: a model of soil genesis, Geoderma, 13, 1-22, doi:10.1016/0016-7061(75)90035-X, 1975.

Hunckler, R.V. and Schaetzl, R.J.: Spodosol development as affected by geomorphic aspect, Baraga County, Michigan, Soil Science Society of America Journal, 61(4), 1105-1115, doi:10.2136/sssaj1997.03615995006100040017x, 1997.

Jobbágy, E.G. and Jackson, R.B.: The vertical distribution of soil organic carbon and its relation to climate and vegetation, Ecological Applications, 10(2), 423–436, doi:10.1890/1051-0761(2000)010[0423:TVDOSO]2.0.CO;2, 2000.

Johnson, D.L., Keller, E.A., and Rockwell, T.K.: Dynamic pedogenesis: new views on some key concepts, and a model for interpreting quaternary soils, Quaternary Research, 33(3), 306-319, doi:10.1016/0033-5894(90)90058-S, 1990.

Johnston, A.E., Poulton, P.R., and Coleman, K.: Soil organic matter: its importance in sustainable agriculture and carbon dioxide fluxes, Advances in Agronomy, 101, 1-57, doi:10.1016/S0065-2113(08)00801-8, 2009.

Johnston, C.A., Groffman, P., Breshears, D.D., Cardon, Z.G., Currie, W., Emanuel, W., Gaudinski, J., Jackson, R.B., Lajtha, K., Nadelhoffer, K., Nelson, D., Jr., Mac Post, W., Retallack, G., and Wielopolski, L.: Carbon cycling in soil, Frontiers in Ecology and the Environment, 2(10), 522-528, doi:10.2307/3868382, 2004.

Kay, B.D.: Soil structure and organic carbon: a review, in: Soil Processes and the Carbon Cycle, Lal, R., Kimble, J.M., Follett, R.F., and Stewart, B.A. (Eds.), CRC Press, Boca Raton, 169-197, 1998.

Kempen, B., Brus, D.J., and Stoorvogel, J.J.: Three-dimensional mapping of soil organic matter content using soil type-specific depth functions, Geoderma, 162, 107-123, doi:10.1016/j.geoderma.2011.01.010, 2011.

Kennedy, B.A.: Valley-side slopes and climate, in: Geomorphology and Climate, Derbyshire, E. (Ed.), John Wiley, London, 171-201, 1976.

Khalil, M.I., Kiely, G., O'Brien, P., and Müller, C.: Organic carbon stocks in agricultural soils in Ireland using combined empirical and GIS approaches, Geoderma, 193-195, 222-235, doi:10.1016/j.geoderma.2012.10.005, 2013.

Königlich Preußische Geologische Landesanstalt: Geologische Karte von Preußen und benachbarten Bundesstaten, 1:25,000 (Geological Map of Prussia and adjacent Federal States, 1:25,000), Landesamt f. Geologie und Bergwesen, Halle, Sachsen-Anhalt, Germany, Sheet Wulfen 4137, 1913a.

Königlich Preußische Geologische Landesanstalt: Geologische Karte von Preußen und benachbarten Bundesstaten, 1:25,000 (Geological Map of Prussia and adjacent Federal States, 1:25,000), Landesamt f. Geologie und Bergwesen, Halle, Sachsen-Anhalt, Germany, Sheet Cöthen 4237, 1913b.

Krause, H.H., Rieger, S., and Wilde, S.A.: Soils and forest growth on different aspects in the Tanana watershed of interior Alaska, Ecology, 40, 492-495, doi:10.2307/1929773, 1959.

Kravchenko, A.N., Robertson, G.P., Hao, X., and Bullock, D.G.: Management practice effects on surface total carbon: differences in spatial variability patterns, Agronomy Journal, 98, 1559-1568, doi:10.2134/agronj2006.0066, 2006a.

Kravchenko, A.N., Robertson, G.P., Snap, S.S., and Smucker, A.J.M.: Using information about spatial variability to improve estimates of total soil carbon, Agronomy Journal, 98, 823-829, doi:10.2134/agronj2005.0305, 2006b.

Ku, H.H.: Notes on the use of propagation of error formulas, Journal of Research of the National Bureau of Standards – C. Engineering and Instrumentation, 70C(4), 263-273, 1966.

Kühn, J., Brenning, A., Wehrhan, M., Koszinski, S., and Sommer, M.: Interpretation of electrical conductivity patterns by soil properties and geological maps for precision agriculture, Precision Agriculture, 10, 490-507, doi:10.1007/s11119-008-9103-z, 2009.

Lal, R.: Soil carbon sequestration impacts on global climate change and food security, Science, 304(5677), 1623-1627, doi:10.1126/science.1097396, 2004.

Lacoste, M., Minasny, B., McBratney, A., Michot, D., Viaud, V., and Walter, C.: High resolution 3D mapping of soil organic carbon in a heterogeneous agricultural landscape, Geoderma, 213, 296-311, doi:10.1016/j.geoderma.2013.07.002, 2014.

Lemercier, B., Lacoste, M., Loum, M., and Walter, C.: Extrapolation at regional scale of local soil knowledge using boosted classification trees: a two-step approach, Geoderma, 171-172, 75-84, doi:10.1016/j.geoderma.2011.03.010, 2012.

Li, Z.L., Tang, R., Wan, Z., Bi, Y., Zhou, C., Tang, B., Yan, G., and Zhang, X.: A review of current methodologies for regional evapotranspiration estimation from remotely sensed data, Sensors 9(5), 3801-3853, doi:10.3390/s90503801, 2009.

Liebens, J. and VanMolle, M.: Influence of estimation procedure on soil organic carbon stock assessment in Flanders, Belgium, Soil Use and Management, 19(4), 364-371, doi:10.1111/j.1475-2743.2003.tb00327.x, 2003.

Lobell, D.B. and Asner, G.P.: Moisture effects on soil reflectance, Soil Science Society of America Journal, 66, 722-727, doi:10.2136/sssaj2002.7220, 2002.

Lowther, J.R., Smethurst, P.J., Carlyle, J.C., and Nambiar, E.K.S.: Methods for determining organic carbon in podzolic sands, Communications in Soil Science and Plant Analysis, 21(5-6), 457-470, doi:10.1080/00103629009368245, 1990.

Lufafa, A., Diédhiou, I., Samba, S.A.N., Séné, M., Khouma, M., Kizito, F., Dick, R.P., Dossa, E., and Noller, J.S.: Carbon stocks and patterns in native shrub communities of Senegal’s Peanut Basin, Geoderma, 146, 75-82, doi:10.1016/j.geoderma.2008.05.024, 2008.

Malone, B.P., McBratney, A.B., and Minasny, B.: Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes, Geoderma, 160, 614-626, doi:10.1016/j.geoderma.2010.11.013, 2011.

Mardia, K.V., Kent, J.T., and Bibby, J.M.: Multivariate Analysis, Academic Press, London, United Kingdom, 521 pp., 1979.

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

Martin, M.P., Wattenbach, M., Smith, P., Meersmans, J., Jolivet, C., Boulonne, L., and Arrouays, D.: Spatial distribution of soil organic carbon stocks in France, Biogeosciences, 8, 1053-1065, doi:10.5194/bg-8-1053-2011, 2011.

McBratney, A.B. and Pringle, M.J.: Estimating average and proportional variograms of soil properties and their potential use in precision agriculture, Precision Agriculture, 1, 125-152, doi:10.1023/A:1009995404447, 1999.

McDonald, E.V. and Busacca, A.J.: Record of pre-late Wisconsin giant floods in the Channeled Scabland interpreted from loess deposits, Geology, 16, 728-731, doi:10.1130/0091-7613(1988)0162.3.CO;2, 1988.

Meersmans, J., Van Wesemael, B., De Ridder, F., Fallas Dotti, M., De Baets, S., and Van Molle, M.: Changes in organic carbon distribution with depth in agricultural soils in northern Belgium, 1960-2006, Global Change Biology, 15(11), 2739-2750, doi:10.1111/j.1365-2486.2009.01855.x, 2009.

Migdall, S., Bach, H., Bobert, J., Wehrhan, M., and Mauser, W.: Inversion of a canopy reflectance model using hyperspectral imagery for monitoring wheat growth and estimating yield, Precision Agriculture, 10, 508-524, doi:10.1007/s11119-009-9104-6, 2009.

Miller, B.A., Koszinski, S., Wehrhan, M., and Sommer, M., Impact of multi-scale predictor selection for modeling soil properties, Geoderma, 239-240, 97-106, doi:10.1016/j.geoderma.2014.09.018, 2015.

Minasny, B. and McBratney, A.B.: Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy, Chemometrics and Intelligent Laboratory Systems, 94(1), 72-79, doi:10.1016/j.chemolab.2008.06.003, 2008.

Minasny, B., McBratney, A.B., Malone, B.P., and Wheeler, I.: Digital mapping of soil carbon, Advances in Agronomy, 118, 1-47, doi:10.1016/B978-0-12-405942-9.00001-3, 2013.

Mishra, U., Lal, R., Liu, D., and Van Meirvenne, M.: Predicting the spatial variation of the soil organic carbon pool at a regional scale, Soil Science Society of America Journal, 74, 906-914, doi:10.2136/sssaj2009.0158, 2010.

Moore, I.D., Grayson, R.B., and Ladson, A.R.: Digital terrain modelling: a review of hydrological, geomorphological, and biological applications, Hydrological Processes, 5(1), 3-30, doi:10.1002/hyp.3360050103, 1991.

Mulder, V.L., de Bruin, S., Schaepman, M.E., and Mayr, T.R.: The use of remote sensing in soil and terrain mapping – a review, Geoderma, 162, 1-19, doi:10.1016/j.geoderma.2010.12.018, 2011.

Myneni, R.B., Hall, F.G., Sellers, P.J., and Marshak, A.L.: The interpretation of spectral vegetation indexes, IEEE Transactions on Geoscience and Remote Sensing, 33(2), 481-486, doi:10.1109/36.377948, 1995.

Neemann, W.: Bestimmung des Bodenerodierbarkeitsfaktors für winderosionsgefährdete Böden Norddeutschlands (Determination of soil erodibility factors for wind-erosion endangered soils in Northern Germany), Geologisches Jahrbuch Reihe F, 25, 131 pp., 1991.

Nyssen, J., Temesgen, H., Lemenih, M., Zenebe, A., Haregeweyn, N., and Haile, M.: Spatial and temporal variation of soil organic carbon stocks in a lake retreat area of the Ethiopian Rift Valley, Geoderma, 146, 261-268, doi:10.1016/j.geoderma.2008.06.007, 2008.

Orton, T.G., Pringle, M.J., Page, K.L., Dalal, R.C., and Bishop, T.F.A.: Spatial prediction of soil organic carbon stock using a linear model of coregionalisation, Geoderma, 230-231, 119-130, doi:10.1016/j.geoderma.2014.04.016, 2014.

Panda, D.K., Singh, R., Kundu, D.K., Chakraborty, H., and Kumar, A.: Improved estimation of soil organic carbon storage uncertainty using first-order Taylor series approximation, Soil Science Society of America Journal, 72, 1708-1710, doi:10.2136/sssaj2007.0242N, 2008.

Pennock, D.J. and De Jong, E.: The influence of slope curvature on soil erosion and deposition in hummock terrain, Soil Science, 144(3), 209-217, doi:10.1097/00010694-198709000-00007, 1987.

Petropoulus, G., Carlson, T.N., Wooster, M.J., and Islam, S.: A review of Ts/VI remote sensing based methods for the retrieval of land surface energy fluxes and soil surface moisture, Progress in Physical Geography, 33(2), 224-250, doi:10.1177/0309133309338997, 2009.

Phachomphon, K., Dlamini, P., and Chaplot, V.: Estimating carbon stocks at a regional level using soil information and easily accessible auxiliary variables, Geoderma, 155(3-4), 372-380, doi:10.1016/j.geoderma.2009.12.020, 2010.

Phillips, J.D.: On the relations between complex systems and the factorial model of soil formation (with discussion), Geoderma, 86, 1-21, doi:10.1016/S0016-7061(98)00054-8, 1998.

Powlson, D.S., Whitmore, A.P., and Goulding, K.W.T.: 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, doi:10.1111/j.1365-2389.2010.01342.x, 2011.

Quinlan, J.R. Learning with continuous classes, Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, 343-348, 1992.

Quinlan, J.R.: Combining instance-based and model-based learning, in: Proceedings of the Tenth International Conference on Machine Learning, Kaufmann, M. (Ed.), 236-243, 1993.

Quinlan, J.R.: C4.5: Programs for machine learning, Machine Learning, 16, 235-240, 1994.

Rasmussen, M.S.: Developing simple, operational, consistent NDVI-vegetation models by applying environmental and climatic information: Part I. Assessment of net primary production, International Journal of Remote Sensing, 19(1), 97-117, doi:10.1080/014311698216459, 1998.

Rawls, W.J., Brakensiek, D.L., and Saxton, K.E.: Estimation of soil water properties, Transactions of the American Society of Agricultural Engineers, 25(5), 1316-1320, 1982.

Rawls, W.J., Pachepsky, Y.A., Ritchie, J.C., Sobecki, T.M., and Bloodworth, H.: Effect of soil organic carbon on soil water retention, Geoderma, 116(1-2), 61-76, doi:10.1016/S0016-7061(03)00094-6, 2003.

Richards, J.A.: Remote Sensing Digital Image Analysis, Springer, 439 pp., 2006.

Richter, D.D. and Markewitz, D.: How deep is soil?, Bioscience, 45(9), 600-609, doi:10.2307/1312764, 1995.

Schaetzl, R.J.: Complete soil profile inversion by tree uprooting, Physical Geography, 7(2), 181-189, doi:10.1080/02723646.1986.10642290, 1986.

Schrumpf, M., Schulze, E.D., Kaiser, K., and Schumacher, J.: How accurately can soil organic carbon stocks and stock changes be quantified by soil inventories?, Biogeosciences, 8(5), 1193-1212, doi:10.5194/bg-8-1193-2011, 2011.

Schwartz, D. and Namri, M.: Mapping the total organic carbon in the soils of the Congo, Global and Planetary Change, 33(1–2), 77–93, doi:10.1016/S0921-8181(02)00063-2, 2002.

Selige, S. Böhner, J., and Schmidhalter, U.: High resolution topsoil mapping using hyperspectral image and field data in multivariate regression modeling procedures, Geoderma, 136(1-2), 235-244, doi:10.1016/j.geoderma.2006.03.050, 2006.

Simbahan, G.C., Dobermann, A., Goovaerts, P., Ping, J., and Haddix, L.: Fine-resolution mapping of soil organic carbon based on multivariate secondary data, Geoderma, 132, 471-489, doi:10.1016/j.geoderma.2005.07.001, 2006.

Snyder, V.A. and Vazquez, M.A.: Structure, in: Encyclopedia of Soils in the Environment, Hillel, D., Hatfield, J.L., Powlson, D.S., Rozenweig, C., Scow, K.M., Singer, M.J., and Sparks, D.L. (Eds.), Elsevier Academic Press, 54-68, 2005.

Shrestha, D.L. and Solomatine, D.P.: Machine learning approaches for estimation of prediction interval for the model output, Neural Networks, 19(2), 225-235, doi:10.1016/j.neunet.2006.01.012, 2006.

Sombroek, W.G., Fearnside, P.M., and Cravo, M.: Geographic assessment of carbon stored in Amazonian terrestrial ecosystems and their soils in particular, in: Global Climate Change and Tropical Ecosystems, Lal, R., Kimble, J.M., and Stewart, B.A. (Eds.), CRC Lewis, Boca Raton, 375–389, 2000.

Sommer, M., Gerke, H.H., and Deumlich, D.: Modelling soil landscape genesis – a “time split” approach for hummocky agricultural landscapes, Geoderma, 145, 480-493, doi:10.1016/j.geoderma.2008.01.012, 2008.

Sommer, M., Halm, D., Weller, U., Zarei, M., and Stahr, K.: Lateral podzolization in a granite landscape, Soil Science Society of America Journal, 64(4), 1434-1442, doi:10.2136/sssaj2000.6441434x, 2000.

Soon, Y.K. and Abboud, S.: A comparison of some methods for soil organic carbon determination, Communications in Soil Science and Plant Analysis, 22(9-10), 943-954, doi: 10.1080/00103629109368465, 1991.

Stevens, A., Udelhoven, T., Denis, A., Tychon, B., Lioy, R., Hoffmann, L., and van Wesemael, B.: Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy, Geoderma, 158(1-2), 32-45, doi:10.1016/j.geoderma.2009.11.032, 2010.

Sutherland, R.A.: Loss-on-ignition estimates of organic matter and relationships to organic carbon in fluvial sediments, Hydrobiologia, 389(1-3), 153-167, doi: 10.1023/A:1003570219018, 1998.

Taylor, J.R.: An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements, 2nd ed. University Science Books, Sausalito, California, USA, 1997.

Ungaro, F., Staffilani, F., and Tarocco, P.: Assessing and mapping topsoil organic carbon stock at regional scale: a scorpan kriging approach conditional on soil map delineations and land use, Land Degradation & Development, 21, 565-581, doi:10.1002/ldr.998, 2010.

USGS: Earth Explorer, U.S. Geological Survey, http://earthexplorer.usgs.gov, last access 19 February 2014.

Walter, C., Viscarra Rossel, R.A., and McBratney, A.B.: Spatio-temporal simulation of the field-scale evolution of organic carbon over the landscape, Soil Science Society of America Journal, 67, 1477-1486, doi:10.2136/sssaj2003.1477, 2003.

Weaver, A. van Breda, The distribution of soil erosion as a function of slope aspect and parent material in Ciskei, Southern Africa, GeoJournal, 23(1), 29-34, doi:10.1007/BF00204406, 1991.

Weisstein, E.W.: Error Propagation, Wolfram MathWorld, http://mathworld.wolfram.com/ErrorPropagation.html, last access: 25 August 2014.

Wilhelm, W.W., Johnson, J.M.F., Hatfield, J.L., Voorhees, W.B., and Linden, D.R.: Crop and soil productivity response to corn residue removal: a literature review, Agronomy Journal, 96, 1-17, doi:10.2134/agronj2004.1000, 2004.

Wilson, J.P. and Gallant, J.C. (Eds.): Terrain Analysis: Principles and Applications, John Wiley & Sons, 2000.

Xie, Y., Sha, Z., and Yu, M.: Remote sensing imagery in vegetation mapping: a review. Journal of Plant Ecology, 1(1), 9-23, doi:10.1093/jpe/rtm005, 2008.

Zevenbergen, L.W. and Thorne, C.R.: Quantitative analysis of land surface topography, Earth Surface Processes and Landforms, 12(1), 47-56, doi:10.1002/esp.3290120107, 1987.

Zhang, Z., Yu, D., Shi, X., Warner, E., Ren, H., Sun, W., Tan, M., and Wang, H.: Application of categorical information in the spatial prediction of soil organic carbon in the red soil area of China, Soil Science and Plant Nutrition, 56, 307-318, doi:10.1111/j.1747-0765.2010.00457.x, 2010.

Figures

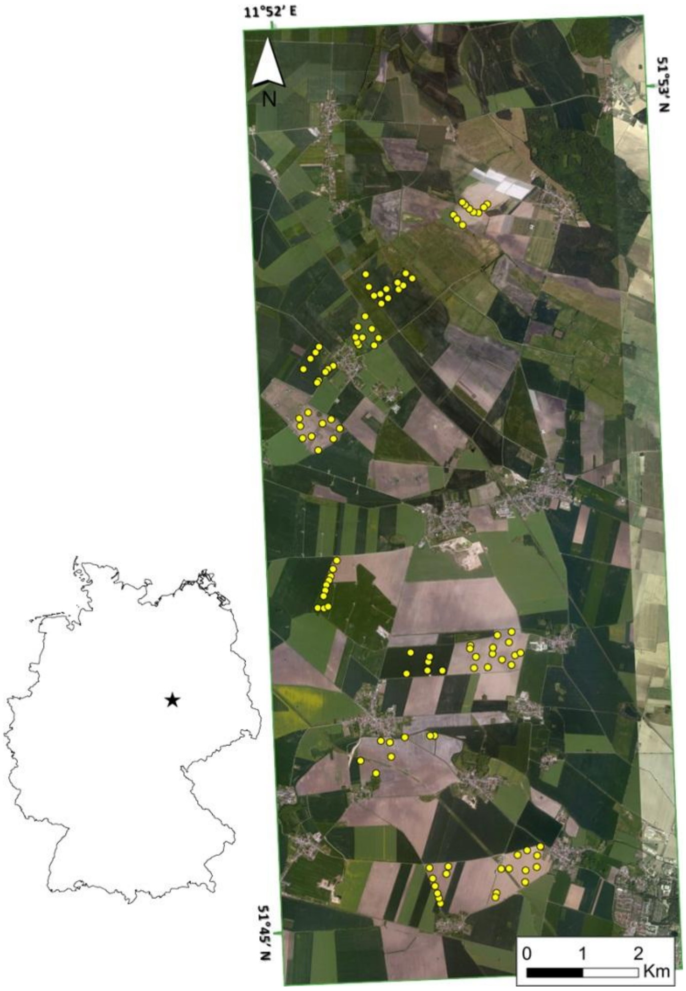

Figure 1. Locations of sample points and study area within Germany.

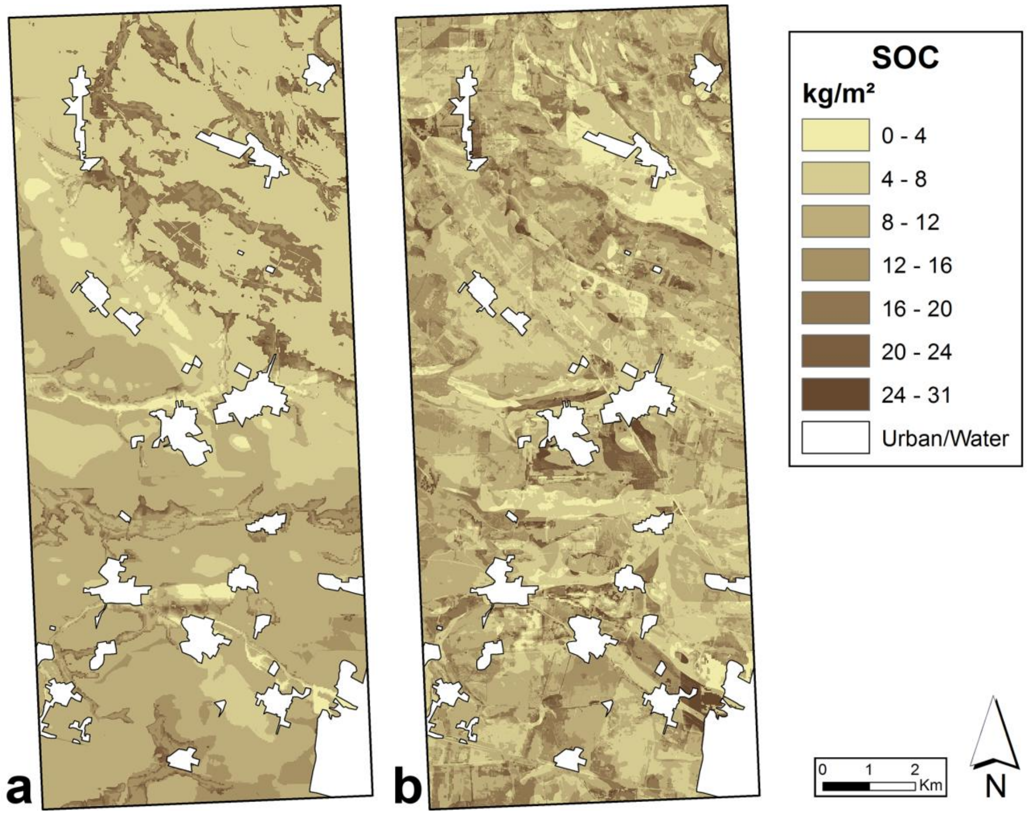

Figure 2. Topsoil SOC_{stock} modelled by a) the direct approach and b) the indirect approach. Overlaid on a hillshade to show relationship with relief and field boundaries.

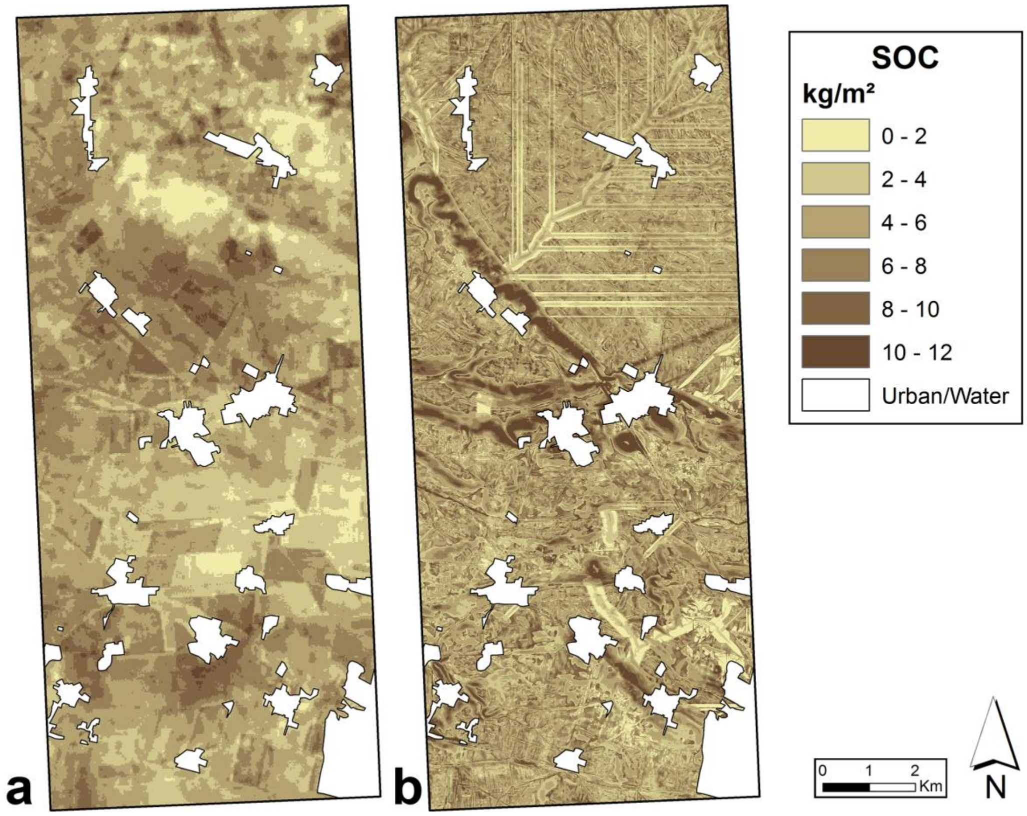

Figure 3. Subsoil SOC_{stock} modelled by a) the direct approach and b) the indirect approach. Overlaid on a hillshade to show relationship with relief and field boundaries.

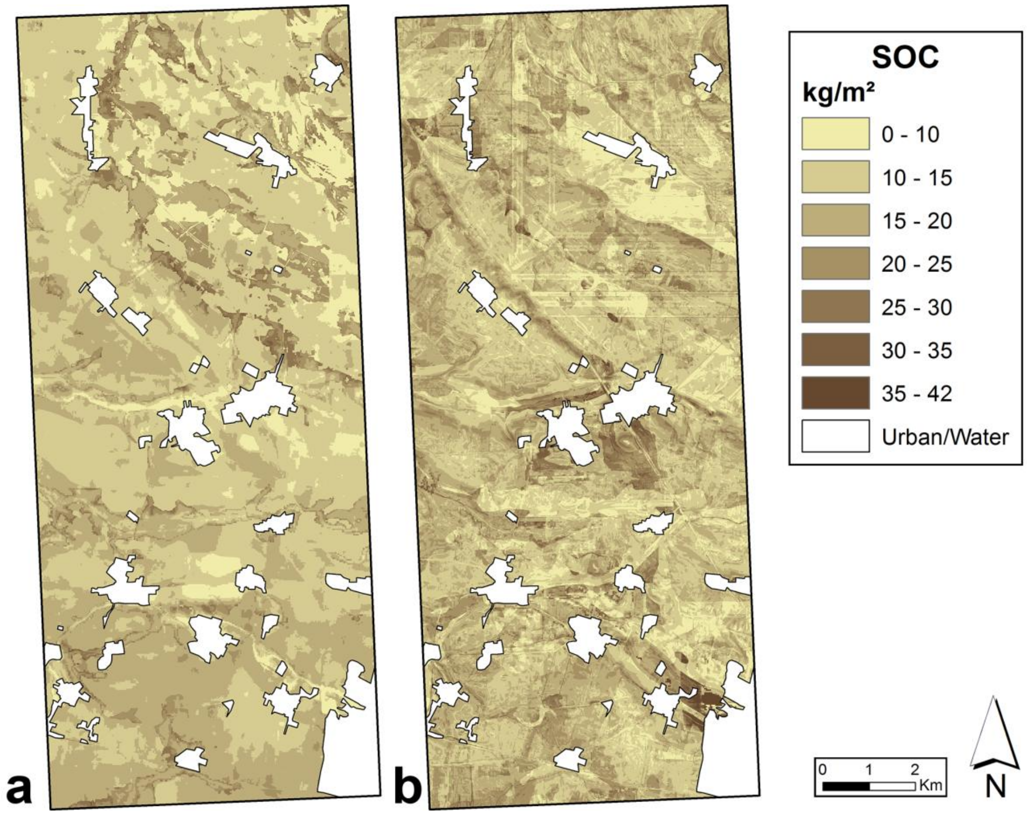

Figure 4. Total SOC$_{stock}$ (topsoil + subsoil) modelled by a) the direct approach and b) the indirect approach. Overlaid on a hillshade to show relationship with relief and field boundaries.

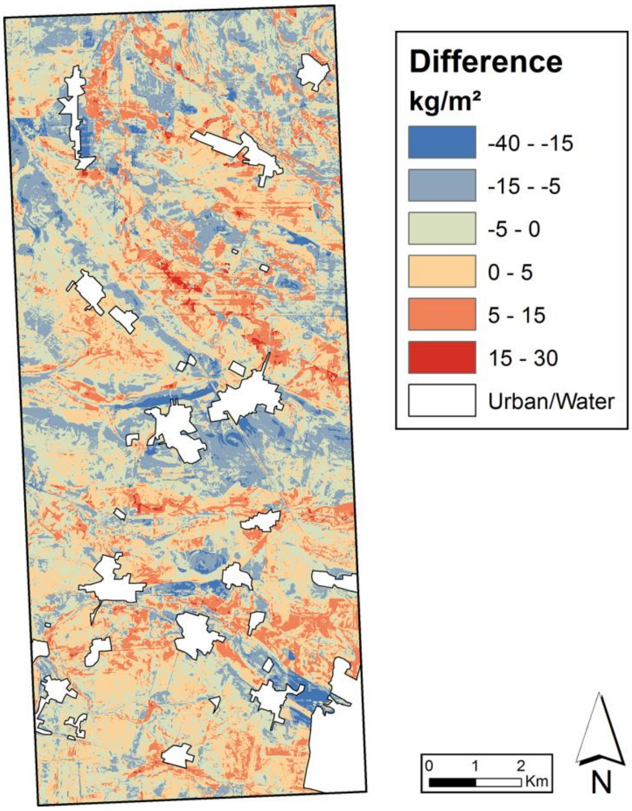

Figure 5. Calculated difference between the direct and indirect approaches of modelling the total SOC$_{stock}$. Negative values are where the indirect approach predicted more SOC$_{stock}$ than the direct approach and positive values are where the indirect approach predicted less.

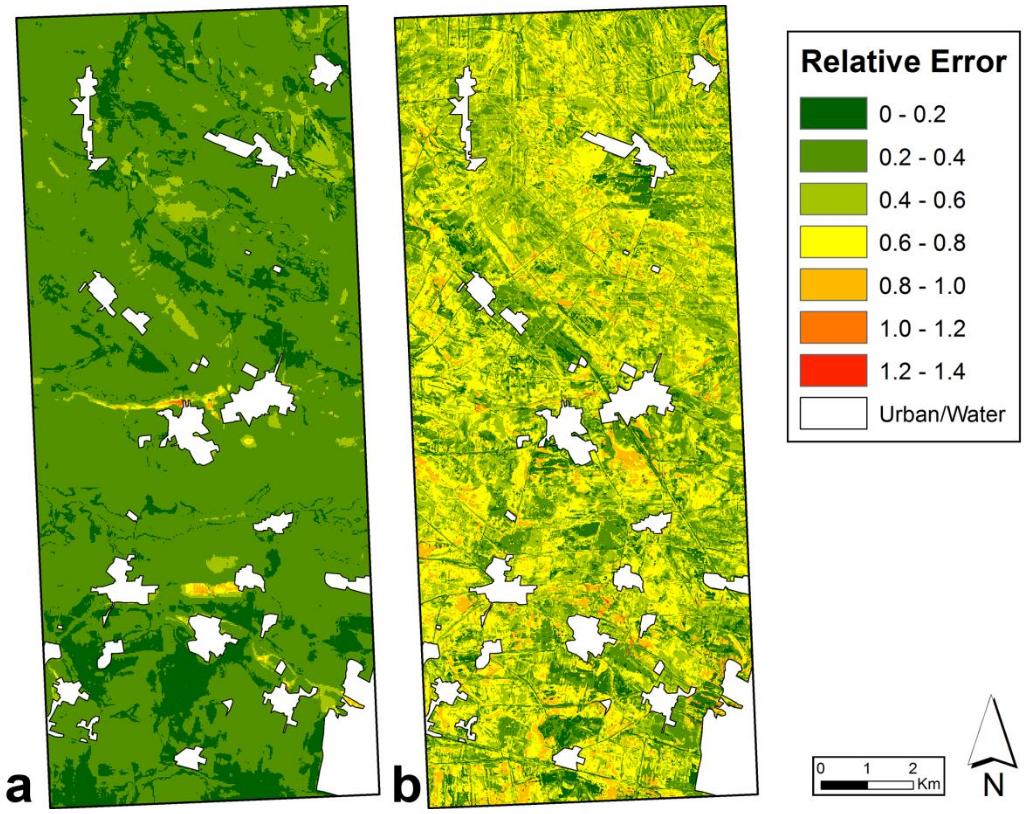

Figure 6. Estimated relative error for the total SOC_{stock} modelled by a) the direct approach and b) the indirect approach.

Tables

Table 1. Descriptive statistics for the observed target variables. BD = total bulk density (g cm^{-3}), SK = particles > 2 mm (%), $SOC_{\%}$ = SOC concentration (%), H = stock thickness (cm), and SOC_{stock} = mass of organic carbon per unit area of soil (kg m^{-2}).

Topsoil	**BD**	**SK**	**H**	**$SOC_{\%}$**	**SOC_{stock}**
Min.	1.18	0.00	10	0.75	1.80
Median	1.50	1.30	40	1.46	9.27
Mean	1.51	3.15	43.61	1.56	9.82
Max.	1.85	44.70	105	4.03	28.03
Std. Dev.	0.11	5.50	15.35	0.53	4.49
Subsoil					
Min.	1.33	0.00	18	0.02	0.07
Median	1.63	4.07	86	0.23	3.10
Mean	1.63	8.99	86.66	0.26	3.37
Max.	1.96	63.36	155	0.71	9.86
Std. Dev.	0.13	12.28	32.60	0.13	2.04

Table 2. Predictor variables considered in this study.

Predictor	Software	Analysis Scale
Elevation (LiDAR, bare-earth)	n/a	2 m
Slope gradient	GRASS	6 - 195 m
Profile curvature	GRASS	6 - 195 m
Plan curvature	GRASS	6 - 195 m
Aspect -west {rotated for N, E, and S}	GRASS	6 - 345 m
Aspect (8 classes)	ArcGIS (raster calculator)	6 - 345 m
Northness	transformed from aspect	6 - 345 m
Eastness	transformed from aspect	6 - 345 m
Longitudinal curvature	SAGA	10 m
Cross-section curvature	SAGA	10 m
Convexity	SAGA	10 m
Relative elevation - rect. neighborhood	ArcGIS toolbox	6 - 4000 m
Relative elevation - circ. neighborhood	ArcGIS toolbox	6 - 4000 m
Topographic position index (TPI)	ArcGIS toolbox	6 - 4000 m
TPI - slope position	ArcGIS toolbox	multiple
TPI - landform classification	ArcGIS toolbox	multiple
Hillslope position	ArcGIS toolbox	multiple
Catchment area	SAGA	n/a
Catchment slope	SAGA	n/a
Channel network base level	SAGA	n/a
Convergence index	SAGA	n/a
Flow accumulation	SAGA	n/a
Flow path length	SAGA	n/a
Length-slope factor	SAGA	n/a
Modified catchment area	SAGA	n/a
Relative slope position	SAGA	n/a
SAGA wetness index	SAGA	n/a
Stream power	SAGA	n/a
Vertical distance to channel	SAGA	n/a
Wetness index	SAGA	n/a
Geology (1:25,000 legacy map)	n/a	423 ha (mean)

Table 2 (cont'd).

Predictor	Resolution	Date
AVIS - LAI-green leaf area	5m	21 Jun. 2005
AVIS - LAI-brown leaf area	5m	21 Jun. 2005
Ikonos	4 m, 4 bands	4 Jul. 2006
Ikonos - panchromatic	1 m	4 Jul. 2006
Ikonos - LAI	5m	4 Jul. 2006
Ikonos - dry matter	5m	4 Jul. 2006
Landsat 5 NDVI (USGS, 2014)	30m	11 Jun. 2006
Landsat 5 NDVI (USGS, 2014)	30m	22 Jul. 2006
Landsat 5 LandsatLook (USGS, 2014)	30m, 3+1 band	20 Jun. 2006
Landsat 5 LandsatLook (USGS, 2014)	30m, 3+1 band	6 Jul. 2006
Landsat 5 LandsatLook (USGS, 2014)	30m, 3+1 band	22 Jul. 2006
Landsat 5 LandsatLook (USGS, 2014)	30m, 3+1 band	15 Sep. 2006
Landsat 5 LandsatLook (USGS, 2014)	30m, 3+1 band	17 Oct. 2006
Landsat 5 TM (USGS, 2014)	30m, 6 bands; 60m, 1 band	11 Jun. 2006
Landsat 5 TM (USGS, 2014)	30m, 6 bands; 60m, 1 band	22 Jul. 2006
Landsat 5 SR (GLCF, 2014)	30m, 7+2 bands	11 Jun. 2006
Landsat 5 SR (GLCF, 2014)	30m, 7+2 bands	22 Jul. 2006

Table 3. Relative use (%) of predictors in models derived by Cubist for the topsoil and subsoil stocks. BD = total bulk density (g cm^{-3}), SK = particles > 2 mm (%), SOC$_{\%}$ = SOC concentration (%), H = stock thickness (cm), and SOC$_{stock}$ = mass of organic carbon per unit area of soil (kg m^{-2}).

Topsoil			**Subsoil**		
Rules	*MLR*	*Predictor*	*Rules*	*MLR*	*Predictor*
BD			**BD**		
100%	100%	Relative elev. - circ. (2000 m)	100%	0%	Geology map units
51%	100%	Landsat5 SR, band 7 (6 Jun. 2006)	68%	100%	LandsatLook, band 5 (6 Jul. 2006)
17%	100%	Relative elev. - rect. (20 m)		100%	Landsat5 NDVI (22 Jul. 2006)
	96%	LandsatLook, band 5 (17 Oct. 2006)		100%	LandsatLook, band 6 (6 Jul. 2006)
	87%	Relative elev. - rect. (10 m)		100%	Landsat5 TM, band 1 (11 Jun. 2006)
	87%	Aspect, N central angle (215 m)		68%	Landsat5 SR, band 7 (22 Jul. 2006)
	83%	Landsat5 SR, band 2 (6 Jun. 2006)		32%	Landsat5 SR, band QA (6 Jun. 2006)
	34%	SAGA wetness index		32%	Landsat5 SR, band 1 (22 Jul. 2006)
	13%	Relative elev. - circ. (800 m)		32%	Landsat5 SR, band 6 (22 Jul. 2006)
SK			**SK**		
100%	100%	TPI (70 m)	100%	3%	Stream power
94%	0%	Aspect class (70 m)	76%	76%	Landsat5 SR, band 2 (11 Jun. 2006)
39%	16%	Relative elev. - rect. (550 m)	21%	0%	Profile Curvature (118 m)
37%	14%	LandsatLook, band 6 (17 Oct. 2006)	15%	79%	Landsat5 SR, band 4 (6 Jun. 2006)
	94%	Relative elev. - rect. (1800 m)		85%	Catchment slope
	84%	Landsat5 NDVI (11 Jun. 2006)		76%	LandsatLook, band 3 (20 Jun. 2006)
	80%	Aspect, N central angle (50 m)		56%	Landsat5 NDVI (11 Jun. 2006)
	78%	Landsat5 TM, band 4 (20 Jun. 2006)		56%	LandsatLook, band 4 (20 Jun. 2006)
	78%	Relative elev. - circ. (3000 m)		56%	Aspect, W central angle (70 m)
	64%	Aspect, N central angle (130 m)		21%	SAGA wetness index
	64%	Aspect, S central angle (345 m)			
	64%	Flow path length			
	37%	Aspect, N central angle (295 m)			
H			**H**		
100%	93%	Relative elev. - rect. (1100 m)			
39%	100%	LandsatLook, band 5 (15 Sept. 2006)			*Cubist not used*
34%	34%	LandsatLook, band 5 (22 Jul. 2006)			*(based on 2 m - topsoil thickness)*
25%	93%	Ikonos, band 2 (4 Jul. 2006)			
18%	7%	LandsatLook, band 4 (17 Oct. 2006)			
	100%	Relative elev. - rect. (1200 m)			
	93%	Ikonos, band 1 (4 Jul. 2006)			
	93%	Relative elev. - rect. (1300 m)			
	74%	LandsatLook, band 4 (15 Sept. 2006)			
	74%	TPI (1800 m)			
	74%	TPI (2600 m)			
	74%	Flow path length			
	28%	Relative elev. - circ. (650 m)			
	7%	Landsat5 TM, band 6 (11 Jun. 2006)			

Table 3 (cont'd).

Topsoil			Subsoil		
Rules	*MLR*	*Predictor*	*Rules*	*MLR*	*Predictor*
SOC$_{\%}$			**SOC$_{\%}$**		
100%	0%	Geology map units	100%	100%	Slope gradient (98 m)
49%	39%	Relative elev. - rect. (3200 m)	74%	74%	Stream power
39%	69%	Relative elev. - rect. (2000 m)	55%	55%	Plan curvature (138 m)
33%	74%	Flow path length		74%	Slope gradient (90 m)
21%	62%	Northness (155 m)		74%	Slope gradient (138 m)
	81%	TPI (1200 m)		74%	Slope gradient (185 m)
	80%	Relative elev. - rect. (250 m)		74%	Relative elev. - rect. (3400 m)
	80%	Northness (345 m)		55%	Plan curvature (90 m)
	74%	Aspect, W central angle (90 m)		19%	TPI (950 m)
	69%	Relative elev. - circ. (1600 m)		19%	Vertical distance to channel
	69%	TPI (1100 m)			
	62%	TPI (550 m)			
	62%	Northness (215 m)			
	62%	Eastness (345 m)			
	62%	Modified catchment area			
	32%	Aspect, W central angle (110 m)			
	21%	TPI (250 m)			
	21%	Aspect, W central angle (175 m)			
	12%	Northness (6 m)			
SOC$_{stock}$			**SOC$_{stock}$**		
100%	48%	Relative elev. - rect. (1100 m)		100%	LandsatLook, band 5 (6 Jul. 2006)
48%	100%	Vertical distance to channel		100%	LandsatLook, band 3 (6 Jul. 2006)
	80%	Channel network base level		100%	LandsatLook, band 6 (6 Jul. 2006)
				100%	Landsat5 TM, band 7 (11 Jun. 2006)

Table 4. Fitting performance for the respective models. The model's efficiency (ME) is the ratio between the model's mean absolute error (MAE) and the MAE that would result from only using the mean value as the model. Cubist reports the ME as relative error, but it is renamed here to avoid confusion with the more common definition of relative error. An ME of greater than one indicates that the model is not performing well.

Topsoil models	**BD**	**SK**	**H**	**SOC$_{\%}$**	**Indirect - SOC$_{stock}$**	**Direct - SOC$_{stock}$**
MAE	0.05	1.36	5.90	0.14	1.69	2.27
ME	0.52	0.41	0.47	0.34	0.49	0.66
R^2	0.69	0.85	0.71	0.86	0.73	0.58
Subsoil models						
MAE	0.06	3.77	5.90	0.06	2.75	1.37
ME	0.58	0.42	0.47	0.59	1.67	0.83
R^2	0.67	0.79	0.71	0.55	0.34	0.19

Table 5. Cross-validation performance for the respective models. Note that although the R^2 was severely reduced for most models, the MAE was generally only increased a small amount.

Topsoil models	**BD**	**SK**	**H**	**$SOC_{\%}$**	**Direct - SOC_{stock}**
MAE	0.08	2.70	11.80	0.27	2.94
ME	0.86	0.82	0.93	0.66	0.85
R^2	0.26	0.08	0.12	0.61	0.27
Subsoil models					
MAE	0.09	7.18	11.80	0.09	1.42
ME	0.80	0.80	0.93	0.98	0.86
R^2	0.36	0.26	0.12	0.05	0.17

Table 6. Skewness coefficients for the residuals of each model.

	BD	**SK**	**H**	**$SOC_{\%}$**	**Indirect - SOC_{stock}**	**Direct - SOC_{stock}**
Topsoil models	-0.25	-1.15	0.17	1.04	0.10	0.37
Subsoil models	0.11	-0.74	-0.17	1.18	-1.61	-0.16