Hindawi Publishing Corporation
Applied and Environmental Soil Science
Volume 2014, Article ID 809495, 12 pages
http://dx.doi.org/10.1155/2014/809495

Research Article

Comparison of Three Supervised Learning Methods for Digital Soil Mapping: Application to a Complex Terrain in the Ecuadorian Andes

Martin Hitziger[1] and Mareike Ließ[2]

[1] *Natural and Social Science Interface Group, Department of Environmental Systems Science, ETH Zürich, Universitätsstrasse 22, CHN J71, 8092 Zurich, Switzerland*
[2] *Soil Physics Group, Department of Geosciences, Bayreuth University, Universitätsstrasse 30, 95447 Bayreuth, Germany*

Correspondence should be addressed to Mareike Ließ; mareike.liess@uni-bayreuth.de

Received 14 February 2014; Revised 9 April 2014; Accepted 11 April 2014; Published 20 May 2014

Academic Editor: Ryusuke Hatano

A digital soil mapping approach is applied to a complex, mountainous terrain in the Ecuadorian Andes. Relief features are derived from a digital elevation model and used as predictors for topsoil texture classes sand, silt, and clay. The performance of three statistical learning methods is compared: linear regression, random forest, and stochastic gradient boosting of regression trees. In linear regression, a stepwise backward variable selection procedure is applied and overfitting is controlled by minimizing Mallow's Cp. For random forest and boosting, the effect of predictor selection and tuning procedures is assessed. 100-fold repetitions of a 5-fold cross-validation of the selected modelling procedures are employed for validation, uncertainty assessment, and method comparison. Absolute assessment of model performance is achieved by comparing the prediction error of the selected method and the mean. Boosting performs best, providing predictions that are reliably better than the mean. The median reduction of the root mean square error is around 5%. Elevation is the most important predictor. All models clearly distinguish ridges and slopes. The predicted texture patterns are interpreted as result of catena sequences (eluviation of fine particles on slope shoulders) and landslides (mixing up mineral soil horizons on slopes).

1. Introduction

The most prominent conceptual model for explaining and interpreting the spatial distribution of soils is the fundamental equation of soil-forming factors, known from Jenny [1]. This conceptualization points at five very generally formulated factors influencing soil development: climate, organisms, relief, parent material, and time. According to the initial letters of these factors, the model is referred to as "*clorpt*" model.

This model can thus be applied for predicting the spatial development of soils as a function of one or several clorpt factors, an approach that is referred to as digital soil mapping (DSM). Within the scope of DSM, a considerable number of studies that span a large range of theoretical and applied goals, methodological approaches, prediction factors, and data sources were conducted. McBratney et al. [2] review 67 studies in which soil classes and/or soil attributes were spatially predicted and Grunwald [3] provides a multicriteria characterization of 90 studies conducted no earlier than 2007 in Geoderma and Soil Science Society of America Journals alone.

Among the factors of the clorpt model, Schaetzl and Anderson [4] attribute the relief factor the highest explanatory power for short scale variability of soil. According to them, this is due to it providing potential and kinetic energy on water movement and thus conditioning the redistribution of energy and matter. The same observation of the relief having a large impact on soil patterns is also manifest in the catena concept, originally defined by Milne [5] as the sequence of soils between a hilltop and the adjacent valley bottom.

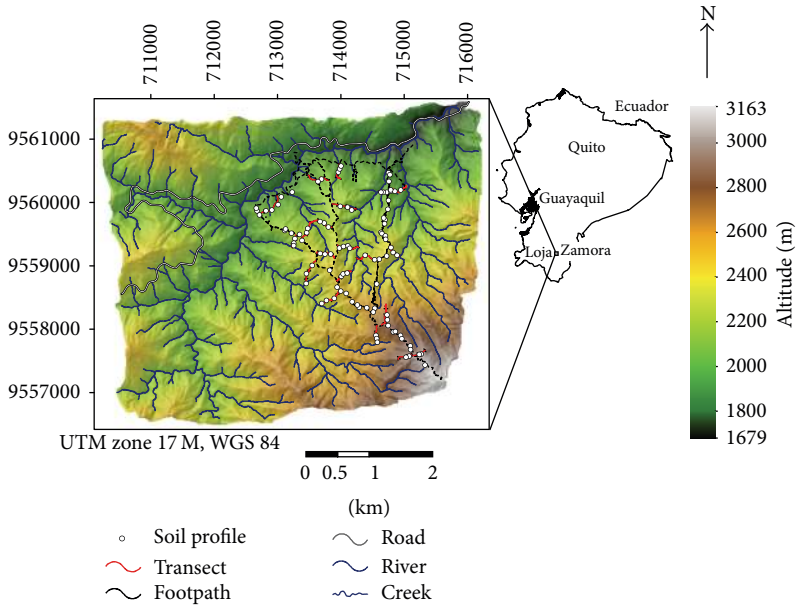

Figure 1: Research area with soil profile locations. Overlaid hillshading with light source from north (adapted from [17]).

This paper presents a methodology to compare linear regression, random forest, and stochastic gradient boosting of regression trees (subsequently referred to as boosting), applied to predicting topsoil texture in a complex terrain in the Ecuadorian Andes. Despite the large number of studies and the importance of relief for soil development, studies applying DSM to extensive, complex mountainous terrains are scarce (e.g., [6, 7]). Whilst linear regression is a commonplace method in a multitude of DSM applications, boosting has not been applied as frequently. In fact, neither McBratney et al. [2] nor Grunwald [3] review any application of both boosted regression trees and random forest to the same data. One study that employed both techniques is Viscarra Rossel and Behrens [8] who predicted clay content, SOC, and pH from remote sensing data.

2. Material and Methods

2.1. Research Area. The research area is located in the "Reserva de Biósfera San Francisco" (RBSF) at around 3°58′S and 79°4′W. It is situated on the eastern escarpment of the southern Ecuadorian Andes, in the valley of the San Francisco River. The area extends across 26 km^2 and spans the altitudinal range from 1800 m to 3200 m a.s.l. (Figure 1).

It is affected by the tropical trade wind regime with strong easterlies all over the year. Combined with the location at the eastern Andean range, this results in >6000 mm annual precipitation, reaching its maximum intensity from April to August. Westerly winds occur only in the somewhat drier November [18]. The lower parts of the area (<2200 m a.s.l.) belong to the "tierra templada" with temperatures around 13–19°C and relatively lower rainfall intensities. The upper parts (>2200 m) belong to the "*tierra fria*" with temperatures around 6–13°C and a maximum observed precipitation rate of 36.2 mm h^{-1} [18, 19].

The NNW-facing slopes of the valley up to the tree line are mostly covered by evergreen broadleaved mountain forests, followed by subpáramo shrubland above the tree line. Homeier and Werner [20] estimate that the RBSF harbours some 1500–1700 species of seed plants, or about 10% of the entire spermatophyte flora of Ecuador. A clear differentiation can be observed between dense forests in the ravines and more open crest woodlands [21]. Canopy height rarely exceeds 20 m and bigger trees are rare [22]. The forests on the SSE facing slopes of the valley were largely converted to pasture land 12–30 years ago [23].

The area is part of the Chiguinda unit of the Zamora Series. Parent material is highly weathered and comprises metasiltstones, siltstones, and quartzites, intermixed with layers of phyllite and clay schists [24]. The area is rich in sloping mires [17], thus leading to a prevalence of soils with stagnic properties and thick organic horizons. These were described as, for example, Humaquepts [25] (according to the USDA Soil Taxonomy [26]) and as Histosols and Stagnosols, associated with Umbrisols, Cambisols, Leptosols, and Regosols [27] (according to the World Reference Base for Soil Resources [28]).

2.2. Dataset. Transects were defined by Ließ et al. [27] to cover a representative yet accessible selection of different terrain forms. Along each of the transects, three plots were selected in the field aiming at (1) equidistance from each other, (2) coverage of full transect length, (3) coverage of different terrain forms, and (4) suitability for pit construction and soil sampling. 107 soil pits were excavated at the positions displayed in Figure 1. Topsoil texture was determined by wet

Table 1: Terrain parameter calculation with SAGA GIS software.

	Parameter	Module	Reference	Saga author/year
1	Elevation (El)	Fill sinks	[9]	Wichmann 2003
2	Slope (α)	Slope, aspect, curvature	[10]	Conrad 2001
3	Aspect			
4	Valley depth	Relative heights and slope positions	—	Böhner and Conrad 2008
5	Wind	Wind effect	—	Boehner and Ringeler 2008, Conrad 2011
6	CI	Convergence index	[11]	Conrad 2003
7	Specific catchment area	Flow tracing/kinematic routing algorithm (KRA)	[12]	Conrad 2001
8	OFD	Strahler order ≥5	[13]	Olaya 2004
		Overland flow distance to channel network	[14]	Conrad 2001 & 2011
9	SWI	Saga wetness index	[15]	Böhner and Conrad 2001
10	LS factor	Length of slope factor	[16]	Conrad 2003

sieving and pipette analysis according to Köhn [29]. Due to some missing data, 99 samples entered the soil texture models.

Spatial information on terrain parameters for the regionalisation of the sampled point data was derived from a digital elevation model (DEM) of 10 m raster cell resolution provided by the research unit's database [30]. Ten terrain features were selected for the set of initial predictors and calculated with the terrain analysis modules of the open-source software SAGA [31] (Table 1). Most of these terrain features or very similar indices were already used for predicting responses related to soil texture [32–36].

2.3. Supervised Learning Methods. Linear regression is very simple, fast, and efficient. Therefore, it is commonly used in many applications [8, 34, 35]. However, it follows assumptions such as linear relations between predictor and response variables, normally distributed and independent data points, and constant variance [37].

The properties of decision tree methods are almost contrary to those of linear regression. They have to be solved iteratively and approximately. But they are free of the rigid assumptions, and they can accommodate all sorts of variable scales, intervariable relations, and distributions. Random forest decorrelates the individual trees and decreases the prediction variance in comparison to simple regression trees [33]. This method was repeatedly applied to DSM [8, 38, 39]. Stochastic gradient boosting was developed by Freund and Schapire [40] and modified by Friedman [41, 42]. Additionally to the variance reduction already implemented in random forest, boosting also reduces the bias of the prediction [43]. It was praised as the "best off-the-shelf classifier in the world" by Breiman [44]. Recent regression applications to DSM were described by Viscarra Rossel and Behrens [8] and classification applications by Grinand et al. [45] and Lacoste et al. [46]. All modelling was done within the *R* software environment, version 2.14.1 [47].

2.3.1. Linear Regression. A stepwise backward variable selection was conducted (*R* package: *leaps*). The criterion for selecting the final model and to prevent overfitting was minimizing Mallows's Cp [48]. Collinearity among the predictors would violate the assumption of independent and identically distributed data, inflate the variance of parameters, and bias predictor selection. During the stepwise variable selection, this error exacerbates as variables "wrongly" skipped would change the trajectory of the subsequent selection decisions [49]. Therefore, collinearity was checked with a correlation analysis using Pearson's correlation coefficient >0.7 as threshold criterion for excluding predictors in accordance with Dormann et al. [49].

Outliers were detected with fitted values versus standardized residuals plots. Exclusion criterion was a standardized residual that exceeds the 3-fold standard deviation. To check for excessively influential data points, leverage versus standardized residuals plots were used. Leverage describes the influence of observed values on the fitted values for the same data points [50]. Cook's Distance derives from a combination of standardized residuals and the leverage of data points. It measures the effect that deleting an observation has on the prediction parameters of a model [51]. Exclusion criterion for influential data points was Cook's Distance exceeding 0.5.

The normality of the random component was checked with *q-q* normal plots. The violation of independence and identically distributed random components would bias parameter estimates and increase the risk of type *I* errors (falsely rejecting the null hypothesis of no effect) [52]. This assumption has got two facets. On the one hand, constant standard deviation was checked by residuals versus fitted values plots. On the other hand, spatial autocorrelation of the residuals was checked by spatial plots of the residuals and Moran's *I*. Moran's *I* calculates the correlation of the data points within certain distance classes. Similarly to Pearson's correlation coefficient, positive values indicate positive spatial autocorrelation and vice versa, with values of zero indicating no spatial autocorrelation [53]. For the largest distance classes, Moran's *I* is large due to small sample size—for these classes Moran's *I* cannot reliably detect spatial autocorrelation and has thus to be discarded [49]. This analysis calculated

Moran's I for distance classes of 250 m and considered absolute values >0.3 as indicator of autocorrelation. Spatial autocorrelation plots and Moran's I were calculated with the *R* package *ncf*.

2.3.2. Random Forest. Random forest repeatedly fits trees to bootstrap samples of the predictor data of randomly selected predictors and then averages the predictions [43]. Due to the bootstrapping procedure, only a subset of the data is used to fit every particular tree. For fitting the random forest models, the *R* package *randomForest* with a squared error loss function was used. Random forest contains several tuning parameters, some of which control internal random processes: number of randomly selected predictors used to fit each tree ("*mtry*"), minimum node size ("*nodesize*"), size of the bootstrap sample ("*sampsize*"), and number of trees fitted ("*ntree*").

Despite these tuning parameters, it has been argued that random forest does not need much tuning effort, because decision tree methods are said to be robust, if not immune against noise variables [37]. Therefore, many authors argue that no predictor selection was required [38, 54]. However, some predictors in the preliminary random forest models were assigned negative importance scores. Secondly, it is frequently said that random forest does not overfit [8, 38, 55, 56]. However, Hastie et al. [37] state that this conclusion was misconceived. Indeed random forest would not overfit when increasing *ntree* but would approach the expectation of the prediction. Yet, they emphasize that the expectation of the prediction can nevertheless overfit the data.

Therefore, several modelling procedures were employed for comparison.

(1) No predictor selection and no tuning (*R* default values for *mtry*, *nodesize*, and *sampsize*).

(2) Predictor selection, but no tuning. For predictor selection, the *R* package *Boruta* was employed. Boruta compares the importance scores of original predictors with the importance scores of their counter variables. Boruta renders three importance classes of variables, "*confirmed,*" "*tentative*" (decision algorithm nonconverging), and "*rejected*" [57]. All *confirmed* and *tentative* predictors were used.

(3) No predictor selection, but tuning of *sampsize*. Sampsize was tuned by fitting 18 models with different *sampsizes* $(5, 10, \ldots, 90)$ and choosing the parameter value that renders the lowest prediction error.

(4) No predictor selection, but tuning of *mtry*. *mtry* is suggested as a potentially sensitive parameter by Breiman and Cutler [58] and thus is used for regular tuning [38, 55, 56]. It was tuned by using the function *tuneRF*.

The rank order of procedure complexity is thus (1) < (2) = (3) = (4). Procedures (2) to (4) are of a similar level of complexity, yet both are more complex than (1). Throughout the analysis *ntree* has been set to 2000 in order to make sure it is large enough. As there are inherently stochastical processes in building random forest models, each of these four fitting procedures was conducted in 100-fold replication. A 5-fold cross-validated RMSE was used as a measure of the prediction error.

Rather than testing the RMSE distributions for significant differences, it was decided to boxplot them and to visually determine the superior fitting procedure according to the following criteria: median and variance of the RMSE distributions and complexity of the modelling procedure. The following decision rule was applied to select the best procedure: *the model with the lowest median prediction error was selected unless the upper hinge of its boxplot was worse than the upper hinge of the boxplot of the model with the second-lowest median prediction error. If there were several procedures with lowest yet about equal median prediction errors, the one with the better upper boxplot hinge was selected. In case of both similar median prediction errors and similar upper hinges, the model with the lower complexity rank was selected.*

2.3.3. Gradient Boosting. Boosting draws bootstrap samples of the predictor data, fits a tree, and subtracts the prediction from the original data. The trees are iteratively fitted to the residuals and the predictions summed up [37]. Measures to prevent overfitting are thus crucial because the sequential nature of boosting (in contrast to merging models as in random forest) allows trees to be added until the model is completely overfitted [52].

The boosting models were fitted by using the code published by Elith et al. [43], which is based on the package *gbm*. As in random forest there is a range of tuning parameters: *interaction depth* determines the number of splits in each tree and *shrinkage* reduces the contribution of each individual tree to the final model. The smaller the latter is, the lower the prediction risk and the more trees and calculation time are required [37]. Ridgeway [59] recommends setting *shrinkage* to 0.01–0.001. Elith et al. [43] recommend setting it small enough to allow at least for 1000 trees. Hence, for all procedures, *shrinkage* was set to the lower end of the recommendations (0.001), as this parameter can apparently not be set too low except for computational reasons. This generally allowed for more than a thousand trees. The *subsampling rate* determines the size of the bootstrap sample: Elith et al. [43] recommend 0.5–0.75; the *R* default is 0.75. The *number of trees* (*ntree*) is more relevant than for random forest, as gradient boosting overfits if *ntree* is excessive. Hence, *ntree* has to be determined for each individual application. However, the function *gbm.step* does so automatically. Therefore, tuning *ntree* is of no concern in the subsequent steps.

Regarding tuning of other parameters, Elith et al. [43] provide data that tuning *interaction depth* and *shrinkage* has some (yet for small datasets not much) effect on the prediction error. Regarding variable selection, Elith et al. [43] argue that, for small datasets in particular, redundant predictors would degrade the prediction by increasing its variance. Therefore, a scheme was developed to compare several modelling procedures.

(1) No variable selection and no tuning.

(2) The optimal model of procedure (1), but checked whether variable selection with the function *gbm.simplify* improves the prediction.

(3) The optimal model from procedure (2), but with a *subsampling rate* of 0.5.

(4) The optimal model from procedure (2), but with an *interaction depth* of 2.

The rank order of procedure complexity thus is (1) < (2) < (3) = (4). The RMSE of a 5-fold cross-validation was used to evaluate the runs. This scheme was run in 100-fold repetition to account for the variation resulting from the involved stochastic processes. The superior modelling procedure was identified by the decision rule described for random forests.

2.3.4. Model Comparison and Validation. In order to compare model performance and to estimate modelling uncertainty of the three applied methods, a 5-fold cross-validation scheme was computed. The cross-validation covers the complete modelling procedures selected in the previous sections, including predictor selection. The cross-validation was conducted in 100-fold repetition to account for the effect of external (e.g., sample attribution to cross-validation groups) and internal (e.g., bootstrapping) random events and to derive a measure of the variance of the modelling procedures. Therefore, a root mean square error (RMSE) distribution of the complete modelling procedure was obtained which covers a large sample of random configurations. To allow absolute judgement of model performance, the RMSE distribution of the mean of the data was calculated by the same scheme.

Rather than assessing the model's performance by using common indicators such as the coefficient of determination (R^2) or the Nash-Sutcliffe Model Efficiency [60], a graphical approach was selected because of its higher information content. Therefore, the RMSE distributions resulting from the 100-fold repetitions of the cross-validation were box-plotted, adding the boxplot parameters of the RMSE distribution of the mean as baseline. The method rendering the most useful predictions was determined by the same decision rule that was used for determining the best modelling procedure, exempting the complexity criterion, which does not apply here.

2.4. Prediction. From the three modelled texture classes, the one with the smallest reduction of the median RMSE prediction error is dropped and the pertinent class is derived as the difference between 100% and the sum of the remaining two classes. Therefore, the three classes always add up to 100%, and no scaling or logit transformation is required.

3. Results and Discussion

3.1. Data Overview. The interquartile ranges clearly differentiate the rather low clay contents 8.3–18.7% from the higher sand contents 22.8–41.7% and the even higher silt contents 47.0–59.9% (Figure 2). Thus texture data showed that most analysed soils are rather loamy which is in accordance with Schrumpf et al. [25].

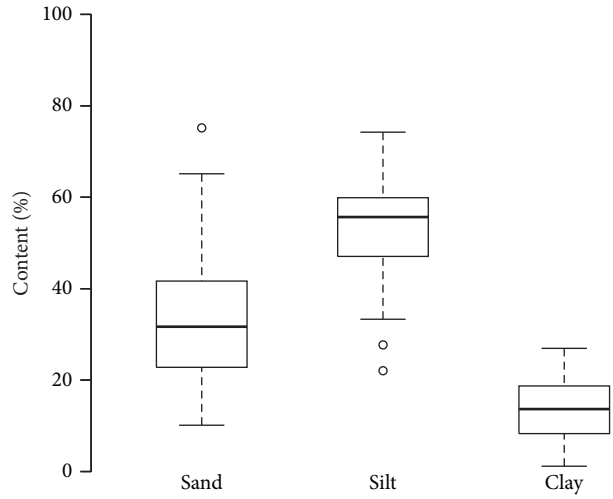

Figure 2: Boxplots of soil texture.

3.2. Model Adaptation

3.2.1. Linear Regression. According to the minimal Mallow's Cp, the best performing models for sand (S) and clay content (C) contained two predictors each, while silt (S_i) was best explained by three predictors. The full formulas of the selected models are displayed in

$$S = -13.131 + 0.016 \cdot \mathrm{EL} + 0.616 \cdot \mathrm{LS} + \mathrm{Error},$$

$$S_i = 96.729 - 0.008 \cdot \mathrm{EL} - 11.989 \cdot \alpha - 2.744 \cdot \mathrm{SWI} + \mathrm{Error},$$

$$C = 37.945 - 0.009 \cdot \mathrm{EL} - 7.959 \cdot \alpha + \mathrm{Error}. \quad (1)$$

From calculated Pearson's correlation coefficients of the predictors none exceeded 0.7. Therefore, the models were unbiased by collinearity and the assumption of independent predictors was not violated. The residuals showed no heteroscedasticity; the random components were normally distributed and no outliers were determined. The maps of the residuals (Figures 3(a)–3(c)) seem to show some spatial autocorrelation. However, Moran's I (Figures 3(d)–3(f)) did not exceed the threshold criterion of 0.3 for any of the distance classes.

3.2.2. Random Forest. Figure 4 displays the prediction error distributions of the four modelling procedures. Predictor selection only had a slightly positive effect on the prediction error distribution for silt content, without a marked increase in the variance of the prediction error. Tuning reduced the prediction error of the silt model even further. Despite literature recommendations and the frequent use of *mtry* as the main tuning parameter, the effect of tuning *sampsize* was larger than of tuning *mtry* (Figure 4(b)). Regarding the sand and clay models, predictor selection as well as tuning impaired model performance (Figures 4(a) and 4(c)).

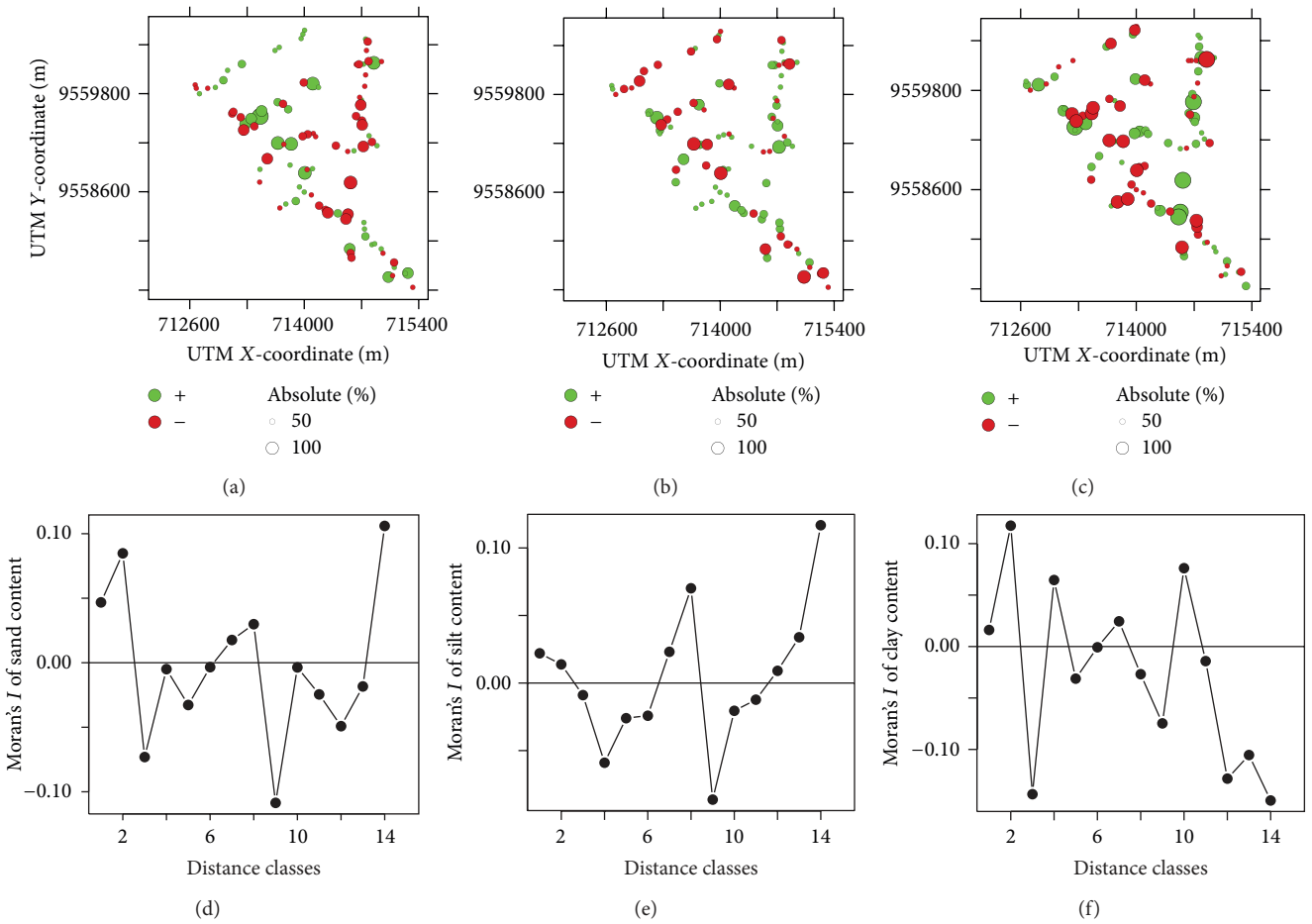

FIGURE 3: Spatial distribution of residuals in % of the maximum absolute residual value (red = negative values, green = positive values, top) and Moran's I (bottom). Models from linear regression for ((a) + (d)) sand content, ((b) + (e)) silt content, and ((c) + (f)) clay content.

In general, the impact of the variable selection and the tuning steps was relatively small for all responses. The application of the decision rule led to the selection of the following modelling procedures: for sand and clay content, modelling procedure (1) was selected; concerning silt content, modelling procedure (3) was selected.

Figure 5 displays the maps of the residuals and Moran's I for the selected random forest models. The autocorrelation of the residuals did not exceed Moran's $I > 0.3$.

3.2.3. Gradient Boosting. Figure 6 shows a comparison of the cross-validated prediction error distributions of the four boosting procedures. The improvement of the median RMSE was rather small. The variance remained constant throughout all four procedures but was relatively high in relation to the small differences in median prediction error. Predictor selection had a positive effect, whilst additional tuning did rather not improve or even impaired the predictions.

The application of the decision rule led to the identification of the following modelling procedures: for sand and silt content procedure (2) was chosen; for clay content procedure (4) was chosen. The selected procedures and the usage of the full dataset to fit the final models resulted in 4050 trees for sand content, 2250 trees for silt content, and 950 trees for clay content. Hence, Elith et al.'s [43] recommendation to set shrinkage low enough to allow for at least 1000 trees was accounted for except in the case of clay that required slightly less than 1000 trees, presumably due to the model allowing first-order interactions. The maps of the residuals (Figures 7(a)–7(c)) and Moran's I (Figures 7(d)–7(f)) did not show any autocorrelation.

3.3. Model Comparison and Validation. Figure 8 compares the 5-fold cross-validated prediction error distributions of the selected linear regression, random forest, and gradient boosting models to the mean.

Linear regression predictions for sand and clay and random forest predictions for clay were generally better than the mean but still overlapped partly with the RMSE distributions of the mean. The exception is the linear regression for silt, which has been clearly overfitting. The boosting prediction was the superior method according to the decision rule, was in all responses better than the mean, and showed hardly any overlap with its RMSE distributions except for silt, even though absolute differences to the other methods remained small. This contrasts with the finding of Viscarra Rossell

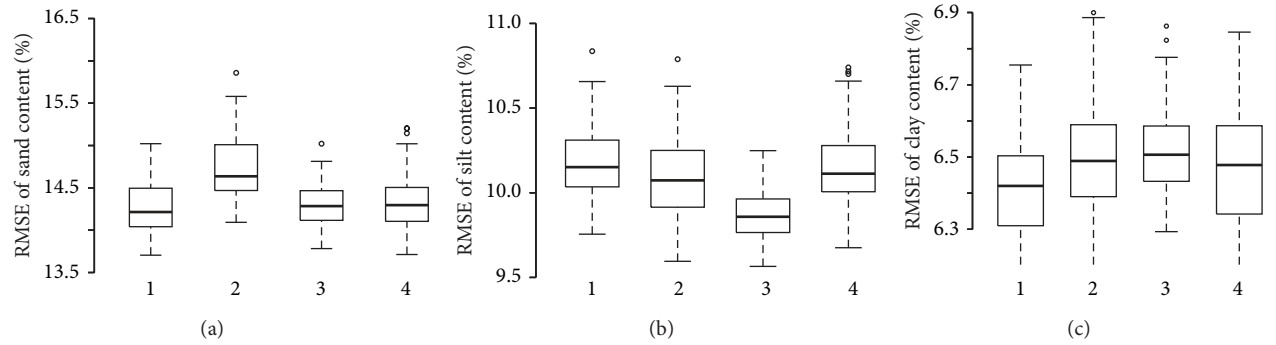

Figure 4: Prediction error distributions as RMSE of the four random forest modelling procedures.

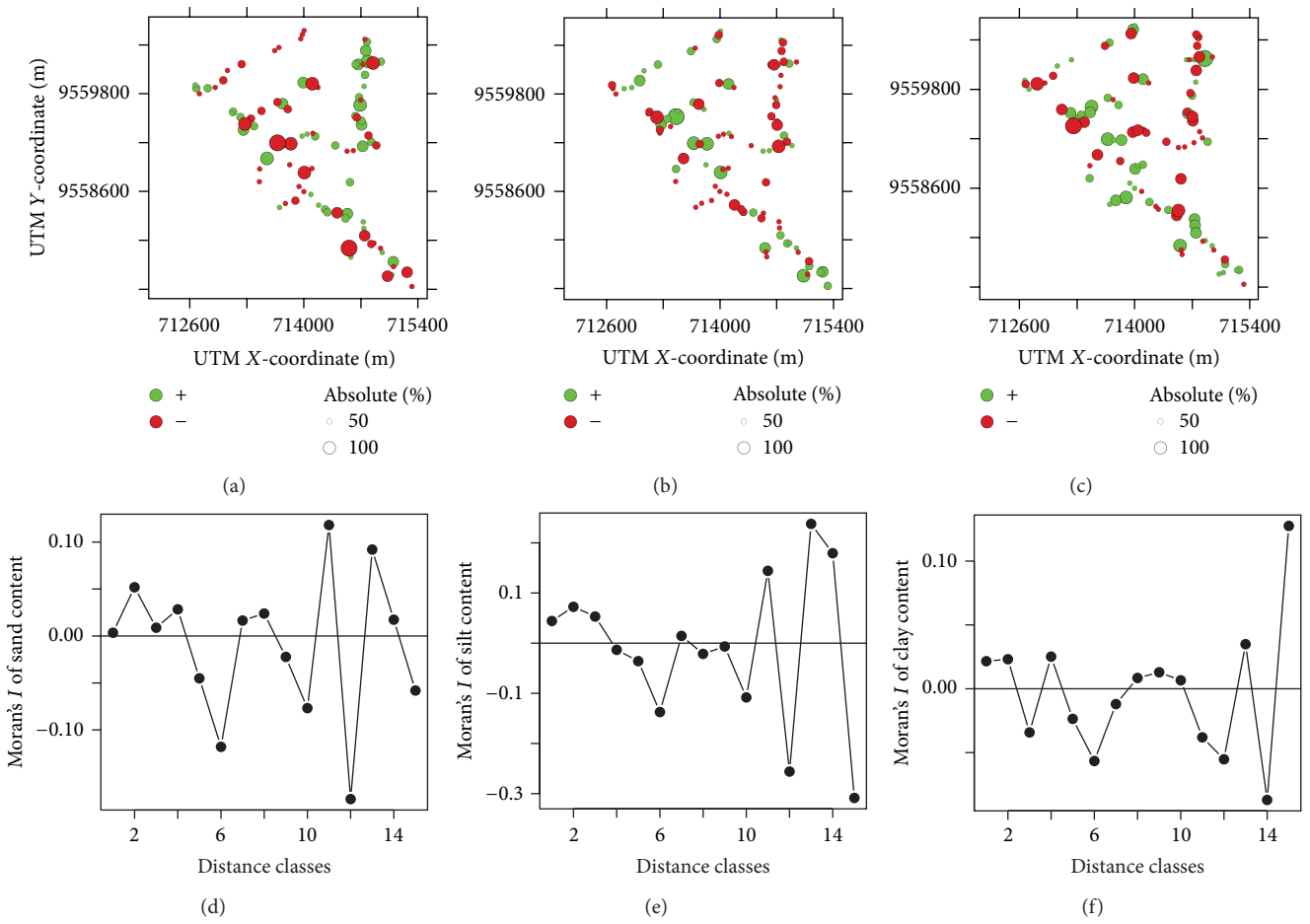

Figure 5: Spatial distribution of residuals in % of the maximum absolute residual value (red = negative values, green = positive values, top) and Moran's I (bottom). Models from random forest for ((a) + (d)) sand content, ((b) + (e)) silt content, and ((c) + (f)) clay content.

and Behrens [8] who reported boosting to perform worse than random forest in predicting soil properties from remote sensing data. However, the median reduction of the cross-validated RMSE of the boosting procedures compared to the mean was only 5%.

3.4. Prediction. The boosting models were used to predict the response variable for the full extent of the DEM. The silt model was dropped as it provides the smallest reduction of the median RMSE prediction error. Instead, the silt content was derived as the difference between 100% and the sum of the sand and the clay prediction for each cell. Maps of the three soil texture classes are displayed in Figures 9(a)–9(c). Despite relying on different predictors, two effects can be detected visually, on the one hand elevation dependencies (a finding was also stated by Wilcke et al. [61] and Ließ et al. [7])

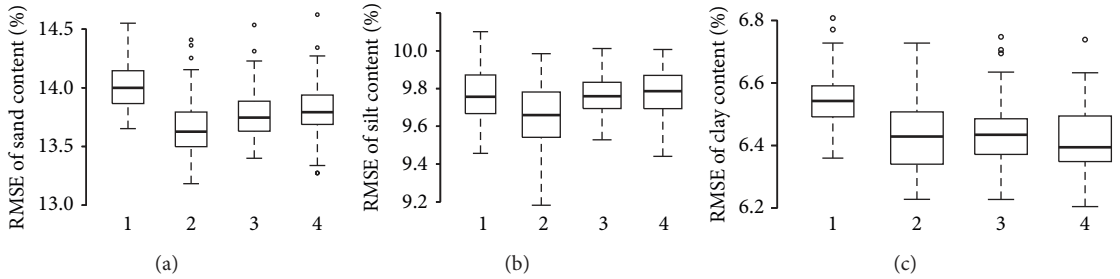

FIGURE 6: Cross-validated prediction error distributions as RMSE of the four boosting procedures.

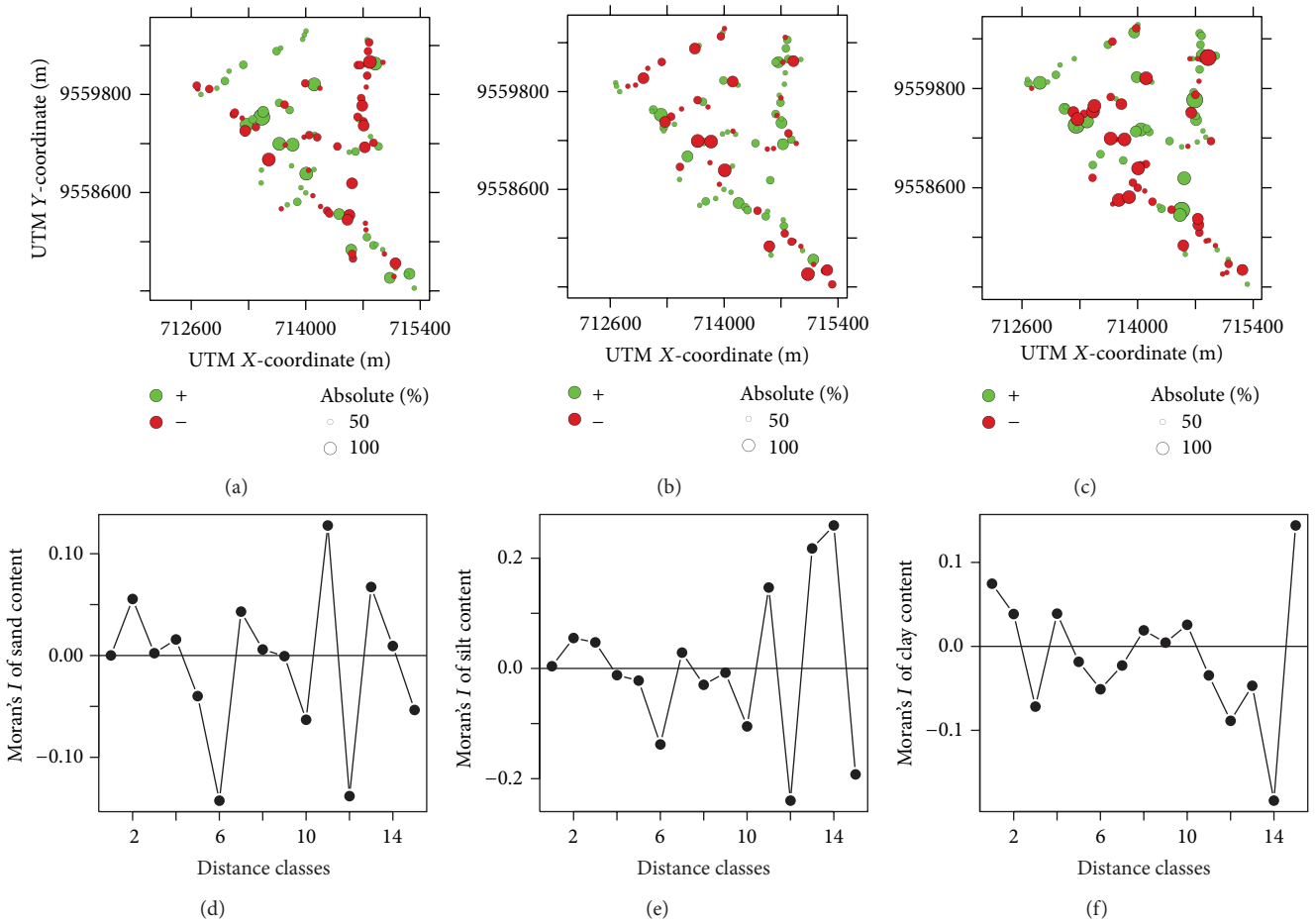

FIGURE 7: Spatial distribution of residuals in % of the maximum absolute residual value (red = negative values, green = positive values, top) and Moran's I (bottom). Models from gradient boosting for ((a) + (d)) sand content, ((b) + (e)) silt content, and ((c) + (f)) clay content.

and on the other hand clear differentiations between ridges and hillslopes.

Marginal dependence plots of the predictors selected by the boosting procedure (Figure 9(d)) help to specify the predicted patterns. While keeping all other predictors constant the sand content rises stepwise with elevation from 30 to 43%, which is clearly visible in the map (Figure 9(a)). However, the ridges did not comply with this simple description; their sand content was lower than of surrounding sites, despite their higher altitude. From a valley depth of 0 to 30 m, sand contents show a steep rise before settling around 38% for deeper valley structures. North and northeast exposed hillslopes show higher sand contents than all other expositions, 38% compared to about 33%.

The relation between clay content and elevation (Figure 9(e)) is opposite to that between sand content and elevation as can be also observed in the map (Figure 9(c)). For low to medium inclinations, clay contents are predicted with ca. 15% and sharply drop to ca. 13% for medium to high inclinations (Figure 9(e)). The effect of slope resulted in increased predicted clay contents for the platform-like areas of low inclination along the ridges. The combination of the effects of

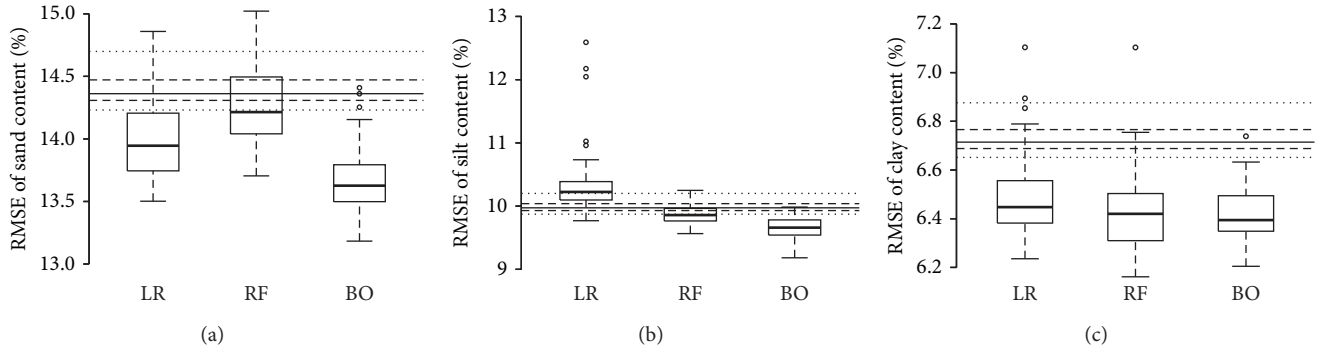

Figure 8: 5-fold cross-validated RMSE distributions of the selected modelling procedures. LM: linear regression; RF: random forest; BO: gradient boosting. The lines refer to the boxplot parameters of the RMSE distribution of the mean of the data as prediction model. Solid line: median; dashed line: upper/lower hinges; dotted line: upper/lower whiskers.

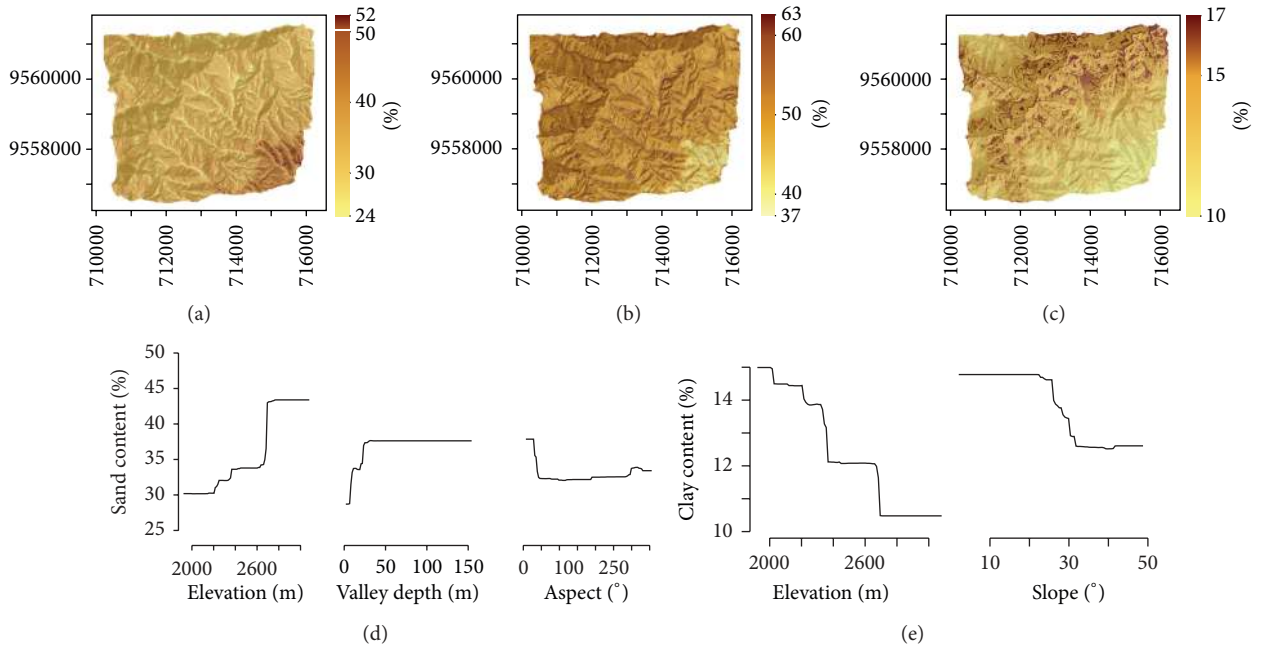

Figure 9: Soil texture maps predicted by gradient boosting with overlaid hillshading from north, (a) sand content, (b) silt content, and (c) clay content and marginal dependence plots displaying the importance of predictors for the sand content (d) and the clay content (e).

these two predictors and their interaction resulted in overall clay content predictions between 10% and 17%.

The conceptual model proposed by Simonson [62] conceives soil development as a function of four classes of processes: addition, removal, translocation, and transformation. In the case of relief-induced processes, particle transformations and translocations are most obviously relevant. Translocation processes according to Simonson's [62] classification have strong ties to the concept of soil catenas, perceiving the relief as driver for soil development by causing the redistribution of energy and matter along a slope [4]. Accordingly, the longer the slope length is, the more the colluvium tends to accumulate [4]. This process explains the overall pattern of increasing sand and decreasing clay content with elevation. The positive correlation of precipitation intensity and elevation [18] as well as Bauer's [63] conclusion that considerable amounts of precipitation are translocated by shallow subsurface water flow further strengthen this finding. The low predicted sand contents on ridges stand in contrast to this explanation. However, Ruhe [64] presented a modified catena sequence of five slope elements in which summit positions experience minimal erosion or accretion and mostly chemical weathering, thus not suffering from clay removal and actually being rich in fine-grained material. This extension of the catena model fits quite well to the predicted patterns.

Furthermore, the frequency of landslides could also contribute to coarser grain sizes on slopes compared to the ridges, as they tend to leave the ridges unaffected, to originate on the slope shoulders, bringing the lower horizons of the backslope area to daylight and mixing the horizons in the downslope area. Another reason explaining the rather low

clay contents at higher altitudes might be the predominating sand stones of this area. However, Ließ et al. [7] checked the influence of parent material on soil texture prediction. They found that it is not a relevant predictor for topsoil texture, most probably due to the small-scale variation in parent material and the frequent translocation of soil material by landslides [65].

4. Conclusions

The applied modelling design is useful, even though the applied cross-validation does not allow generalizations to other areas. Whilst all three methods performed similarly in absolute terms, boosting showed superior performance for all three response variables, predicting more precisely than the mean across almost all repetitions. Linear regression performed well for sand and clay but overfitted the silt response. It is thus recommended to test modified linear regression modelling designs that include interactions and use other indicators than Mallow's Cp for model selection. Random forest did not overfit the expectation of the data, yet only the clay model exceeded the prediction performance of the mean. However, even the variance explained by the boosting model reduced the cross-validated RMSE of the mean by only around 5%. Therefore, DSM applications in tropical mountain areas remain a challenge, even within the area used for model calibration, and it is recommended to extend the suite of predictors to factors of soil development other than relief features.

The most important predicted patterns were elevation dependencies and contrasts between ridges and slopes. Topsoil texture tended to be coarsest at high elevations, medium at low elevations, and finest at ridges. The predicted texture patterns can be interpreted as catena sequence, resulting from down slope eluviation of fine grain sizes initiating at the slope shoulder. According to this model the ridge areas were left unaffected by such sorting processes due to a lack of translocating water flow. Landslides are assumed to increase grain size at slope sites due to mixing the mineral soil horizons.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] H. Jenny, *Factors of Soil Formation*, McGraw-Hill, New York, NY, USA, 1941.

[2] A. B. McBratney, M. L. Mendonça Santos, and B. Minasny, "On digital soil mapping," *Geoderma*, vol. 117, no. 1-2, pp. 3–52, 2003.

[3] S. Grunwald, "Multi-criteria characterization of recent digital soil mapping and modeling approaches," *Geoderma*, vol. 152, no. 3-4, pp. 195–207, 2009.

[4] R. J. Schaetzl and S. Anderson, *Soils: Genesis and Geomorphology*, Cambridge University Press, Cambridge, UK, 2005.

[5] G. Milne, *Provisional Soil Map of East Africa*, African Agricultural Research Station, Amani, Tanganyika Territory, 1936.

[6] M. Ließ, B. Glaser, and B. Huwe, "Making use of the World Reference Base diagnostic horizons for the systematic description of the soil continuum—application to the tropical mountain soil-landscape of southern Ecuador," *Catena*, vol. 97, pp. 20–30, 2012.

[7] M. Ließ, B. Glaser, and B. Huwe, "Uncertainty in the spatial prediction of soil texture. Comparison of regression tree and Random Forest models," *Geoderma*, vol. 170, pp. 70–79, 2012.

[8] R. A. Viscarra Rossel and T. Behrens, "Using data mining to model and interpret soil diffuse reflectance spectra," *Geoderma*, vol. 158, no. 1-2, pp. 46–54, 2010.

[9] O. Planchon and F. Darboux, "A fast, simple and versatile algorithm to fill the depressions of digital elevation models," *Catena*, vol. 46, no. 2-3, pp. 159–176, 2002.

[10] L. W. Zevenbergen and C. R. Thorne, "Quantitative analysis of land surface topography," *Earth Surface Processes & Landforms*, vol. 12, no. 1, pp. 47–56, 1987.

[11] R. Köthe and F. Lehmeier, *SARA—System zur Automatischen Relief-Analyse: User Manual*, Department of Geography, University of Goettingen, 2nd edition, 1996.

[12] N. L. Lea, "An aspect-driven kinematic routing algorithm," in *Overland Flow: Hydraulics and Erosion Mechanics*, A. Parsons and A. Abrahams, Eds., pp. 393–407, Chapman and Hall, New York, NY, USA, 1992.

[13] A. N. Strahler, "Quantitative analysis of watershed geomorphology," *Transactions of the American Geophysical Union*, vol. 8, no. 6, pp. 913–920, 1957.

[14] J. F. O'Callaghan and D. M. Mark, "The extraction of drainage networks from digital elevation data," *Computer Vision, Graphics, & Image Processing*, vol. 28, no. 3, pp. 323–344, 1984.

[15] J. Boehner, R. Koethe, O. Conrad, J. Gross, A. Ringeler, and T. Selige, "Soil regionalisation by means of terrain analysis and process parameterisation," in *Soil Classification 2001*, E. Micheli, F. Nachtergaele, and L. Montanarella, Eds., pp. 213–222, European Soil Bureau, Luxembourg, 2002.

[16] P. J. J. Desmet and G. Govers, "A GIS procedure for automatically calculating the USLE LS factor on topographically complex landscape units," *Journal of Soil and Water Conservation*, vol. 51, no. 5, pp. 427–433, 1996.

[17] M. Ließ, M. Hitziger, and B. Huwe, "The sloping mire soil-landscape of Southern Ecuador: influence of predictor resolution and model tuning on random forest predictions," *Applied and Environmental Soil Science*, vol. 2014, Article ID 603132, 10 pages, 2014.

[18] J. Bendix, R. Rollenbeck, M. Richter, P. Fabian, and P. Emck, "Climate," in *Gradients in a Tropical Mountain Ecosystem of Ecuador*, E. Beck, J. Bendix, I. Kottke, F. Makeschin, and R. Mosandl, Eds., vol. 198, Springer, Berlin, Germany, 2008.

[19] E. Beck, F. Makeschin, F. Haubrich, M. Richter, J. Bendix, and C. Valerezo, "The Ecosystem (Reserva Biológica San Francisco)," in *Gradients in a Tropical Mountain Ecosystem of Ecuador*, E. Beck, J. Bendix, I. Kottke, F. Makeschin, and R. Mosandl, Eds., vol. 198, Springer, Berlin, Germany, 2008.

[20] J. Homeier and F. Werner, "Preliminary checklist of the spermatophytes of the reserva San Francisco (Province Zamora-Chinchipe, Ecuador)," in *Provisional Checklist of Flora and Fauna of the San Francisco Valley and Its Surroundings (Reserva Biológica San Francisco, Province Zamora-Chinchipe, Southern Ecuador)*, L. S. Breckle and S. Breckle, Eds., vol. 4, pp. 15–58, Ecotropical Monographs, 2007.

[21] J. Homeier, H. Dalitz, and S. W. Breckle, "Waldstruktur und Baumarten im montanen Regenwald der Estación Científica

San Franscisco in Südecuador," in *Berichte der Reinhold-Tüxen-Gesellschaft*, vol. 14, pp. 109–118, 2002.

[22] K. Müller-Hohenstein, A. Paulsch, D. Paulsch, and R. Schneider, "Vegetations- und Agrarlandschaftsstrukturen in den Bergwäldern Südecuadors," *Geographische Rundschau*, vol. 56, pp. 48–55, 1994.

[23] F. Werner, J. Homeier, and S. Gradstein, "Diversity of vascular epiphytes on isolated remnant trees in the montane forest belt of southern Ecuador," *Ecotropica*, vol. 11, pp. 21–40, 2005.

[24] M. Litherland, J. A. Aspen, and R. A. Jemielita, *The Metamorphic Belts of Ecuador*, vol. 11 of *Overseas Memoir*, British Geological Survey, 1994.

[25] M. Schrumpf, G. Guggenberger, C. Valarezo, and W. Zech, "Tropical montane rain forest soils. Development and nutrient status along an altitudinal gradient in the South Ecuadorian Andes," *Die Erde*, vol. 132, no. 1, pp. 43–59, 2001.

[26] USDA, NRCS, *Keys to Soil Taxonomy*, Soil Survey Staff, Washington, DC, USA, 10th edition, 2006.

[27] M. Ließ, B. Glaser, and B. Huwe, "Digital soil mapping in Southern Ecuador," *Erdkunde*, vol. 63, no. 4, pp. 309–319, 2009.

[28] FAO, IUSS Working Group WRB, *World Reference Base for Soil Resources*, ISRIC, Rome, Italy, 2007.

[29] M. Köhn, "Bemerkungen zur mechanischen Bodenanalyse III. Ein neuer Pipettapparat," *Zeitschrift für Pflanzenernährung, Düngung, Bodenkunde*, vol. 11, no. 1, pp. 50–54, 1928.

[30] T. Nauss, D. Göttlicher, M. Dobbermann, and J. Bendix, "Central data services in multidisciplinary environmental research projects," 2007, http://www.preagro.de/ezai/index.php/eZAI/article/view/28/28.

[31] SAGA User Group Association, *System for Automated Geoscientific Analyses*, Version 2.0.7, SAGA, Hamburg, Germany, 2011.

[32] I. O. A. Odeh, D. J. Chittleborough, and A. B. McBratney, "Elucidation of soil-landform interrelationships by canonical ordination analysis," *Geoderma*, vol. 49, no. 1-2, pp. 1–32, 1991.

[33] S. de Bruin and A. Stein, "Soil-landscape modelling using fuzzy c-means clustering of attribute data derived from a digital elevation model (DEM)," *Geoderma*, vol. 83, no. 1-2, pp. 17–33, 1998.

[34] A. B. McBratney, I. O. A. Odeh, T. F. A. Bishop, M. S. Dunbar, and T. M. Shatar, "An overview of pedometric techniques for use in soil survey," *Geoderma*, vol. 97, no. 3-4, pp. 293–327, 2000.

[35] A. Gobin, P. Campling, and J. Feyen, "Soil-landscape modelling to quantify spatial variability of soil texture," *Physics and Chemistry of the Earth B: Hydrology, Oceans and Atmosphere*, vol. 26, no. 1, pp. 41–45, 2001.

[36] D. J. Brown, M. K. Clayton, and K. McSweeney, "Potential terrain controls on soil color, texture contrast and grain-size deposition for the original catena landscape in Uganda," *Geoderma*, vol. 122, no. 1, pp. 51–72, 2004.

[37] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York, NY, USA, 2nd edition, 2009.

[38] R. Grimm, T. Behrens, M. Märker, and H. Elsenbeer, "Soil organic carbon concentrations and stocks on Barro Colorado Island—digital soil mapping using random forests analysis," *Geoderma*, vol. 146, no. 1-2, pp. 102–113, 2008.

[39] M. Wiesmeier, F. Barthold, B. Blank, and I. Kögel-Knabner, "Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem," *Plant and Soil*, vol. 340, no. 1, pp. 7–24, 2011.

[40] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[41] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[42] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics and Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[43] J. Elith, J. R. Leathwick, and T. Hastie, "A working guide to boosted regression trees," *Journal of Animal Ecology*, vol. 77, no. 4, pp. 802–813, 2008.

[44] L. Breiman, "Arcing classifiers," *Annals of Statistics*, vol. 26, no. 3, pp. 801–849, 1998.

[45] C. Grinand, D. Arrouays, B. Laroche, and M. P. Martin, "Extrapolating regional soil landscapes from an existing soil map: sampling intensity, validation procedures, and integration of spatial context," *Geoderma*, vol. 143, no. 1-2, pp. 180–190, 2008.

[46] M. Lacoste, B. Lemercier, and C. Walter, "Regional mapping of soil parent material by machine learning based on point data," *Geomorphology*, vol. 133, no. 1-2, pp. 90–99, 2011.

[47] R Development Core Team, *R: A Language and Environment for Statistical Computing (Version 2.14.1)*, R Foundation for Statistical Computing, Vienna, Austria, 2011.

[48] C. L. Mallows, "Some comments on CP," *Technometrics*, vol. 42, no. 1, pp. 87–94, 2000.

[49] C. F. Dormann, J. Elith, S. Bacher et al., "Collinearity: a review of methods to deal with it and a simulation study evaluating their performance," *Ecography*, vol. 35, no. 1, pp. 1–20, 2012.

[50] D. C. Hoaglin and R. E. Welsch, "The hat matrix in regression and ANOVA," *The American Statistician*, vol. 32, no. 1, pp. 17–22, 1978.

[51] R. D. Cook, "Detection of influential observations in linear regression," *Technometrics*, vol. 19, no. 1, pp. 15–18, 2000.

[52] C. F. Dormann, J. M. McPherson, M. B. Araújo et al., "Methods to account for spatial autocorrelation in the analysis of species distributional data: a review," *Ecography*, vol. 30, no. 5, pp. 609–628, 2007.

[53] M. Fortin and M. Dale, *Spatial Analysis—A Guide for Ecologists*, Cambridge University Press, Cambridge, UK, 2005.

[54] R. Díaz-Uriarte and S. Alvarez de Andrés, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol. 7, article 3, 2006.

[55] J. Peters, B. D. Baets, N. E. C. Verhoest et al., "Random forests as a tool for ecohydrological distribution modelling," *Ecological Modelling*, vol. 207, no. 2–4, pp. 304–318, 2007.

[56] A. M. Prasad, L. R. Iverson, and A. Liaw, "Newer classification and regression tree techniques: bagging and random forests for ecological prediction," *Ecosystems*, vol. 9, no. 2, pp. 181–199, 2006.

[57] W. Rudnicki and B. Kursa, "Boruta—a tool for finding significant attributes in information systems," CRAN 535 Reference Manual, 2012, http://cran.r-project.org/web/packages/Boruta/Boruta.pdf.

[58] L. Breiman and A. Cutler, "Random Forest—manual," 2004 http://www.stat.berkeley.edu/~breiman/RandomForests/ .

[59] G. Ridgeway, "Generalized boosted models: a guide to the gbm package," 2007, http://cran.r-project.org/web/packages/gbm/gbm.pdf.

[60] J. E. Nash and J. V. Sutcliffe, "River flow forecasting through conceptual models part I—a discussion of principles," *Journal of Hydrology*, vol. 10, no. 3, pp. 282–290, 1970.

[61] W. Wilcke, Y. Oelmann, A. Schmitt, C. Valarezo, W. Zech, and J. Homeier, "Soil properties and tree growth along an altitudinal transect in Ecuadorian tropical montane forest," *Journal of Plant Nutrition and Soil Science*, vol. 171, no. 2, pp. 220–230, 2008.

[62] R. W. Simonson, "Outline of a generalized theory of soil genesis," *Soil Science Society of America Proceedings*, vol. 23, pp. 152–156, 1956.

[63] F. Bauer, *Water flow paths in soils of an undisturbed and landslide affected mature montane rainforest in South Ecuador [Ph.D. thesis]*, University of Bayreuth, Bayreuth, Germany, 2010.

[64] R. V. Ruhe, "Elements of the soil landscape," in *Transactions of the 7th International Congress of Soil Science*, vol. 4, pp. 165–170, International Society of Soil Science, Madison, Wis, USA, 1960.

[65] R. W. Bussmann, W. Wilcke, and M. Richter, "Landslides as important disturbance regimes—causes and regeneration," in *Gradients in a Tropical Mountain Ecosystem of Ecuador*, E. Beck, J. Bendix, I. Kottke, F. Makeschin, and R. Mosandl, Eds., vol. 198, pp. 319–330, Springer, Berlin, Germany, 2008.