ISRIC World Soil Information

Soil property mapping with the GSIF framework

2D versus 3D regression modelling

Report for the AfSIS transition phase, June 2013 – November 2014

Bas Kempen

Background

The GSIF framework allows two-dimensional as well as three-dimensional regression modelling of soil properties.

In the case of two-dimensional modelling, a regression model is fitted for a single soil layer, such layer can be, for example, one of the six standard depth intervals according to the GlobalSoilMap specifications. For model fitting, the soil data of the layer of interest is extracted from the soil profile description in a model calibration dataset (e.g. the Africa Soil Profiles Database). A regression model is then fitted to the selected subset, after which the model can be used to predict the soil property of interest across a prediction area. If one wants to predict the soil for each of the standard depth intervals with 2D modelling, then this would require six regression models: one for each depth interval.

Alternatively, instead of modelling soil layers individually to predict the three-dimensional variation of a specific soil property, one can also model the soil layers simultaneously by fitting a single regression model to the soil profile data (to all layers instead of depth-specific layers). This is what in GSIF is called a '3D regression model'. In a 3D model, the depth of a soil observation is added to the regression model as an independent (predictor) variable in addition to the environmental covariates. The depth of a single soil observations is taken as the midpoint of the soil profile layer the observations belong to.

The advantage of using a 3D regression model to predict the 3D distribution of soil properties is that it is computationally efficient since only one model needs to be fitted instead of multiple models, and that effect of depth on soil property distribution is modelled explicitly by including depth as an independent variable. In this way, the fitted regression model can be used to predict the soil property of interest at any depth. A major disadvantage, however, in case a *linear* regression model is used, is that the effects of the independent variables on the target soil property, quantified by the model coefficients, are assumed to be constant. The 3D linear regression model is the mere sum of lateral (environmental covariates) and vertical (depth) component. In many situations this is unrealistic. For example, the relationship between a vegetation index such as the Enhanced Vegetation Index (EVI) and soil organic carbon content is expected to be different for the topsoil than for the subsoil, as shown in the example below.

```
> summary(lm(lORCDRC~M13EVIALT,data=et2))

Call:
lm(formula = lORCDRC ~ M13EVIALT, data = et2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.853e+00  5.915e-02   31.33   <2e-16 ***
M13EVIALT   3.587e-04  2.182e-05   16.44   <2e-16 ***

Multiple R-squared:  0.2137, Adjusted R-squared:  0.213

> summary(lm(lORCDRC~M13EVIALT,data=et4))

Call:
lm(formula = lORCDRC ~ M13EVIALT, data = et4)

Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.734e+00  5.563e-02   31.17   <2e-16 ***
M13EVIALT   2.282e-04  2.041e-05   11.18   <2e-16 ***

Multiple R-squared:  0.1252, Adjusted R-squared:  0.1242
```

This example shows the results of a linear regression of soil organic carbon (on log-scale) on EVI for GSM standard depth 2 (5-15 cm) and GSM standard depth 4 (30-60 cm). The estimated model coefficients show that for depth 2 the organic carbon content increases with $3.587*10^{-4}$ log% for one unit of increase in EVI. For depth 4 this is $2.282*10^{-4}$ log%. The models also show that the correlation between organic carbon and EVI is much stronger for depth 2 than for depth 4: the model for the former explains 21.3% of the variation in the dataset whereas the model for the latter explains 12.4%.

Despite this disadvantage of the 3D linear regression model, it is unclear what the effect of the assumption of ‘constant model coefficients with depth’ is on the prediction accuracy. If an accuracy assessment through model validation shows that this assumption negatively affects map accuracy, it might be worthwhile to investigate how to extend the 3D linear regression model so that model coefficients can vary with depth.

The report describes the results of a study that investigates this. This study uses a soil dataset from Ethiopia that was obtained from the Africa Soil Profiles Database v1.1 (AfSPD) {Leenaars, 2013 #130}. In addition, the effect of using global versus local covariates on map accuracy was investigated, as well as use of non-linear random forest models {Strobl, 2009 #674} as an alternative to linear regression models. This was done for three soil properties: organic carbon content, clay content and pH. The prediction accuracy if the linear regression model was determined from the model residuals (internal validation), that of the random forest models by cross-validation (out-of-bag residuals). The organic carbon values were transformed to natural logarithms before modelling.

Data

The AfSPD contains 1842 profiles for Ethiopia (Fig. 1), that are made up of 6365 horizons. These horizons are distributed as follows over the six GSM standard depths (based on the midpoints of the horizons):

- 0-5 cm: 61
- 5-15 cm: 1148
- 15-30 cm: 345
- 30-60 cm: 1041
- 60-100 cm: 833
- 100-200 cm: 1305

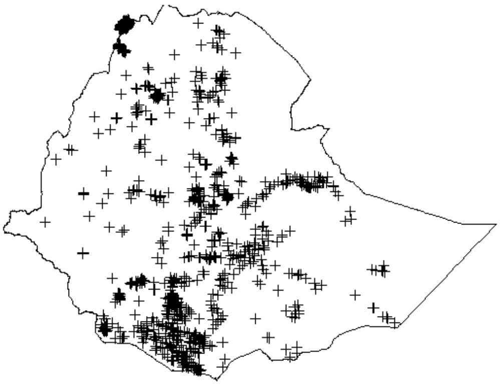

Fig. 1. Ethiopia soil profile locations in the AfSPD.

Methods

Two- and three-dimensional linear regression models were fitted using the covariates that were used to generate the Africa soil property maps at 250 m resolution for the AfSIS projects. These covariates included SoilGrids maps of organic carbon, pH, clay, bulk density, CEC and depth to bedrock, the AfSIS datasets M13EVIxxx (Enhanced Vegetation Index on monthly basis) and M13RB7xxx (Mid-infrared reflectance on monthly basis), and elevation and slope layers. This set of covariates is referred to as 'global covariates'.

For 2D modelling, a regression model was fitted for each of the GSM standard depths using the point data as presented above. For 3D modelling, one regression model was fitted using the complete Ethiopia dataset. For this model, depth was used as a covariate.

This exercise was repeated using i) random forest (RF) modelling, and ii) a set of 'local' covariates. This set included all AfSIS datasets, a landform, geology and soil map derived from the SOTER for northeastern Africa, DEM-derived and land cover layers from the WorldGrids repository, landform attribute layers derived using e-SOTER landform mapping methodology, and layers from the USGS Africa Ecosystems Mapping.

Results

2D versus 3D models

Table 1 shows the validation results of the 2D and 3D linear regression models fitted to the Ethiopian soil data using the global covariate set for organic carbon, clay and pH. The sd1,...,sd6 columns store the validation results of the models fitted for each of the six GSM standard depths. The *RMSE* is the root mean squared error, *cor* is the Pearson correlation coefficient, r^2 is the coefficient of determination which quantifies the fraction of the variance that is explained by the model, *n* is the number of point data used to calibrate the models. The '2D' validation results are computed from the pooled observed, predicted and residual values of the six standard depth models.

The validation results show that the difference in RMSE between the 2D and 3D models is negligible. This is consistent for the three soil properties considered here. These results is surprising since we would have expected a better performance by the 2D models given that these have depth-specific model coefficients. Reasons we can think of that explains these results is the limited number of covariates used, which might hamper the performance of the 2D models, and the fact that both models heavily rely on the SoilGrids soil property maps which have depth specific predictions (other covariates have values that are constant with depth).

Global versus local covariate data

To test the effect of the choice of covariates on the model accuracy we fitted 2D and 3D models with the local covariate set. The validation results are presented in Table 2. For the 3D models the improvement in RMSE compared to the global covariate set is small. For the 2D models the improvement in RMSE is somewhat larger (6.4% smaller RSME for carbon, 6.3% for clay and 3.7% for pH). The 2D models perform better than the 3D models but the difference still is not very large (11% for carbon, 6.3% for clay and 3.7% for pH).

Tree-based methods

Finally, we investigated the use of non-linear tree models for soil property modelling, in particular random forest models, as an alternative for the linear regression model. The main advantage of using tree models over linear regression models is that these can model non-linear relationships. The linear regression model, assumes a linear relationship between the target variable and predictor variable, which in reality might not be true. Another advantage is that tree models do not make any assumptions on the data, i.e. these are non-parametric models. The way tree models are constructed, by means of binary recursive partitioning of the input data, allows the modelling of depth-specific relationships between the target soil property and the environmental covariates. In other words, a tree model does not suffer from the constraint of having constant model coefficients with depth, like the 3D regression model.

Table 1. Validation results of the 2D and 3D linear regression models fitted to the Ethiopia soil data.

Carbon								
	sd 1	sd2	sd 3	sd 4	sd 5	sd 6	2D	3D
n	49	994	267	872	678	1066	3877	3877
RMSE	**8.4**	**12.0**	**9.5**	**6.5**	**4.7**	**3.7**	**7.8**	**8.1**
cor	0.900	0.511	0.558	0.481	0.421	0.339	0.656	0.613
r^2	0.810	0.261	0.311	0.231	0.178	0.115	0.430	0.376
Clay								
	sd 1	sd2	sd 3	sd 4	sd 5	sd 6	2D	3D
n	63	1031	284	931	715	1121	4082	4098
RMSE	**10.6**	**14.4**	**14.4**	**15.5**	**16.4**	**17.7**	**16.0**	**16.5**
cor	0.769	0.568	0.618	0.561	0.574	0.563	0.595	0.558
r^2	0.591	0.323	0.381	0.314	0.330	0.316	0.354	0.310
pH								
	sd 1	sd2	sd 3	sd 4	sd 5	sd 6	2D	3D
n	62	1032	284	930	713	1117	4076	4125
RMSE	**0.47**	**0.85**	**0.70**	**0.83**	**0.82**	**0.79**	**0.82**	**0.84**
cor	0.912	0.649	0.784	0.697	0.734	0.740	0.723	0.705
r^2	0.833	0.421	0.614	0.486	0.539	0.547	0.522	0.497

Table 2. Validation results of the 2D and 3D linear regression models fitted to the Ethiopia soil data using a local covariate dataset.

Carbon								
	sd 1	sd2	sd 3	sd 4	sd 5	sd 6	2D	3D
n	50	996	267	875	679	1066	3872	3896
RMSE	**10.7**	**11.5**	**8.8**	**5.8**	**4.4**	**3.5**	**7.3**	**8.2**
cor	0.706	0.569	0.639	0.611	0.516	0.434	0.702	0.615
r^2	0.498	0.324	0.408	0.374	0.266	0.188	0.493	0.378
Clay								
	sd 1	sd2	sd 3	sd 4	sd 5	sd 6	2D	3D
n	50	995	266	872	676	1063	3872	3885
RMSE	**12.6**	**14.1**	**13.7**	**15.1**	**15.8**	**15.8**	**15.0**	**16.0**
cor	0.661	0.571	0.666	0.590	0.611	0.656	0.642	0.577
r^2	0.436	0.326	0.444	0.348	0.373	0.430	0.400	0.333
pH								
	sd 1	sd2	sd 3	sd 4	sd 5	sd 6	2D	3D
n	50	995	266	874	677	1063	3875	3879
RMSE	**0.51**	**0.86**	**0.69**	**0.83**	**0.77**	**0.73**	**0.79**	**0.82**
cor	0.892	0.643	0.786	0.705	0.774	0.792	0.744	0.723
r^2	0.796	0.414	0.617	0.497	0.599	0.627	0.554	0.523

Table 3. Validation results of the 2D and 3D random forest models fitted to the Ethiopia soil data using global and local covariate data.

Carbon				
	2D		3D	
	global	*local*	*global*	*local*
n	3877	3877	3890	3896
RMSE	**7.2**	**7.3**	**6.9**	**6.8**
cor	0.709	0.708	0.741	0.753
r^2	0.503	0.501	0.550	0.567
Clay				
	2D		3D	
	global	*local*	*global*	*local*
n	4082	3915	4098	3885
RMSE	**15.6**	**15.0**	**13.4**	**12.8**
cor	0.620	0.647	0.741	0.760
r^2	0.384	0.418	0.548	0.577
pH				
	2D		3D	
	global	*local*	*global*	*local*
n	4059	3918	4074	3888
RMSE	**0.76**	**0.73**	**0.60**	**0.57**
cor	0.765	0.792	0.866	0.878
r^2	0.585	0.627	0.750	0.771

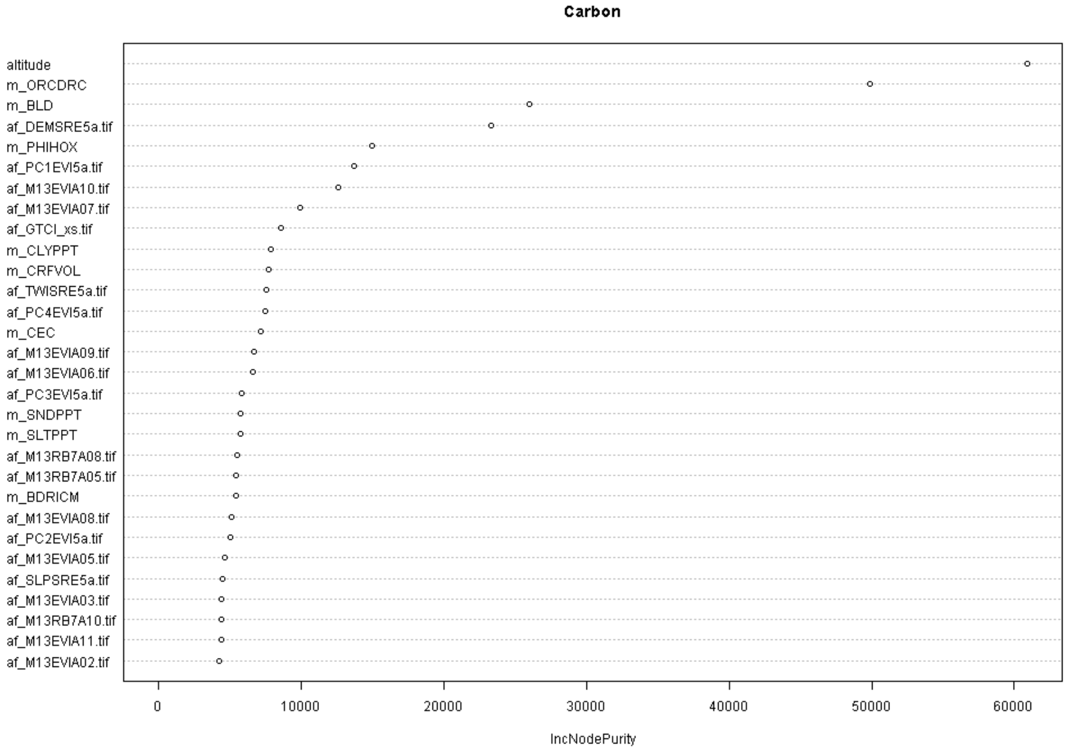

Fig. 2. Variable Importance Plot for the carbon random forest model. The y-axis represents the mean difference in the residual sum of squares as a result of randomly permutating the values of the covariate.

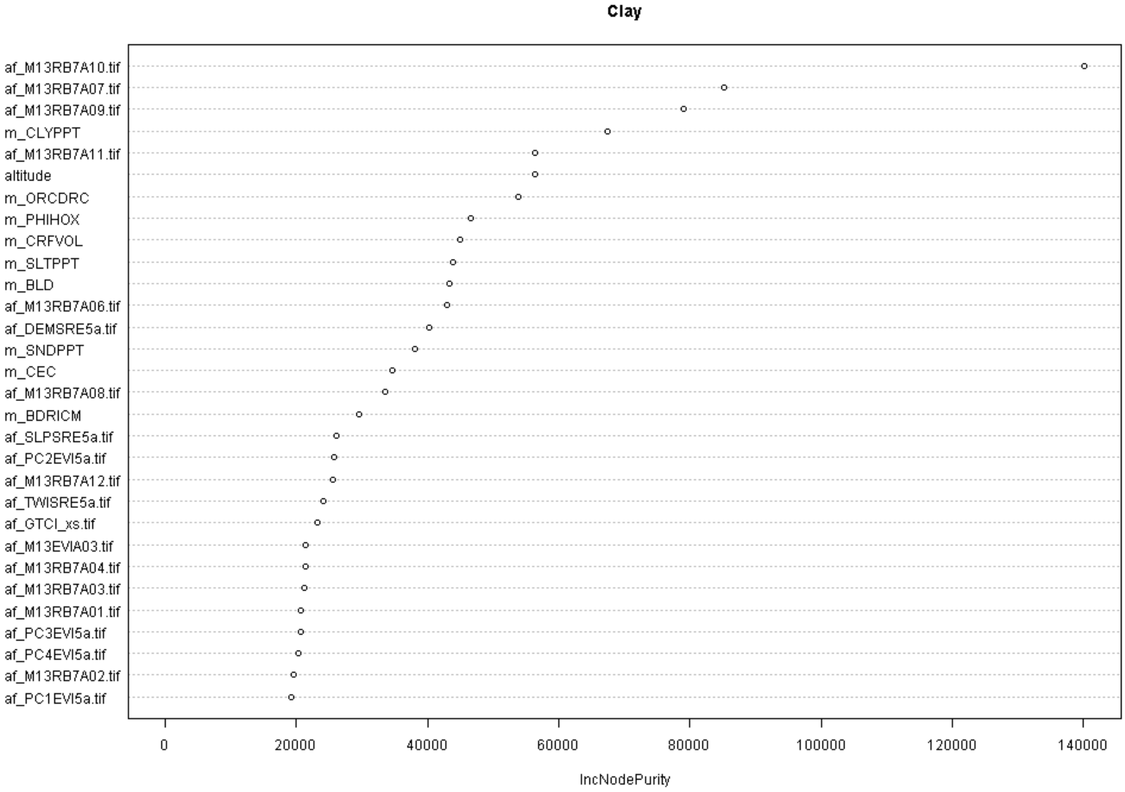

Fig. 3. Variable Importance Plot for the clay random forest model. The y-axis represents the mean difference in the residual sum of squares as a result of randomly permutating the values of the covariate.

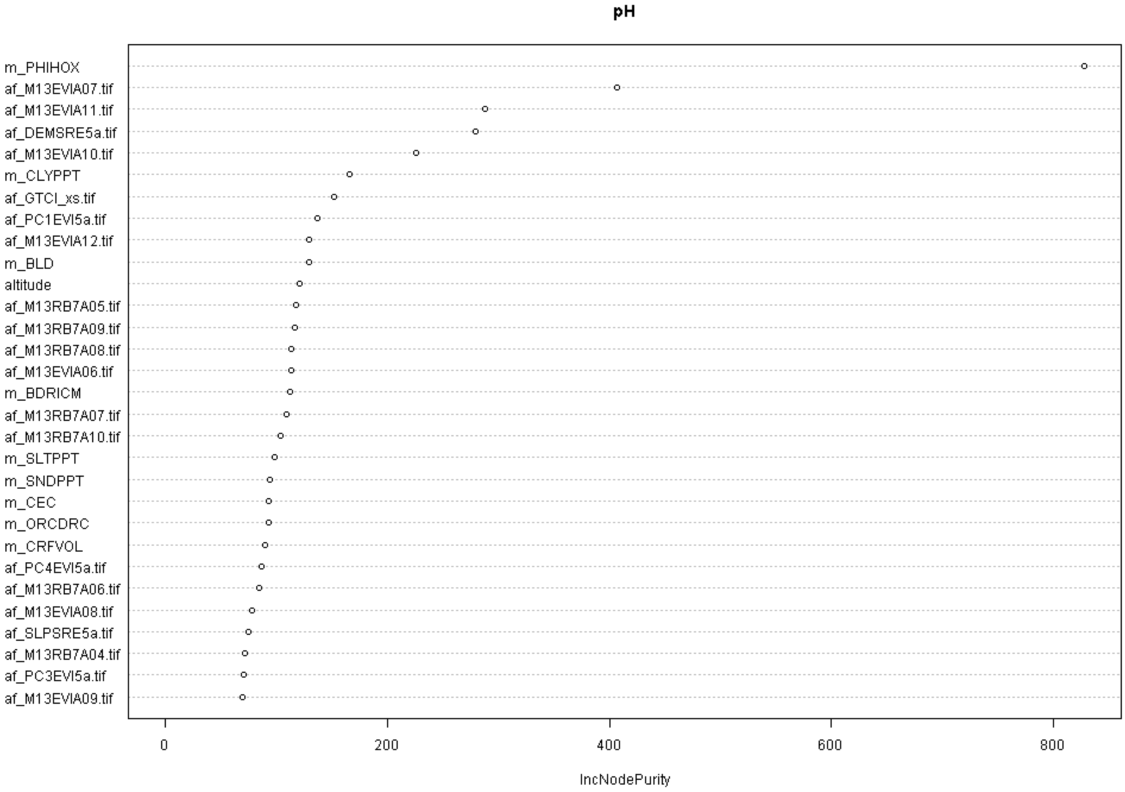

Fig. 4. Variable Importance Plot for the pH random forest model. The y-axis represents the mean difference in the residual sum of squares as a result of randomly permutating the values of the covariate.

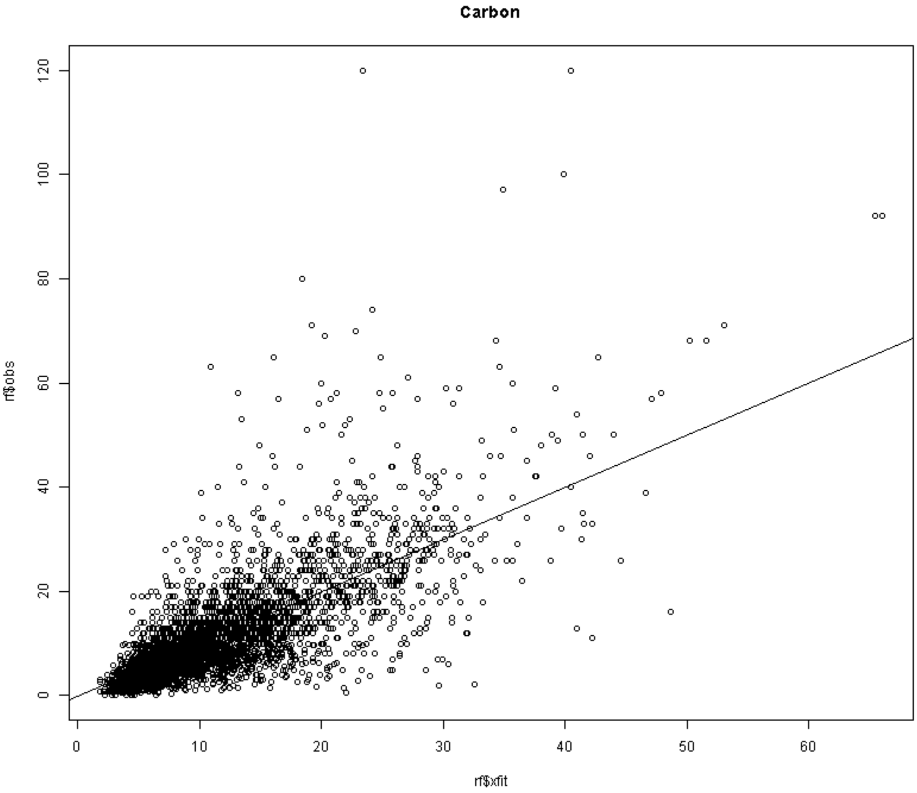

Fig. 5. Observed versus predicted organic carbon content values. Predictions by 3D RF models with global covariate data. The solid line is the 1:1 line. Predicted values were back-transformed from log-scale.

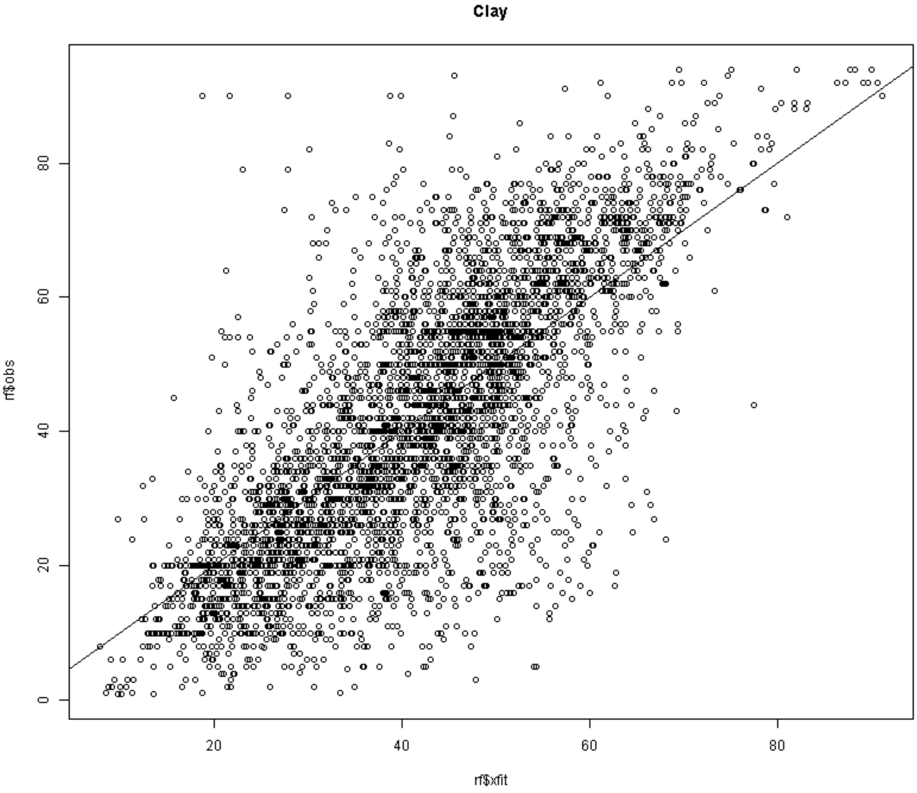

Fig. 6. Observed versus predicted clay content values. Predictions by 3D RF models with global covariate data. The solid line is the 1:1 line.

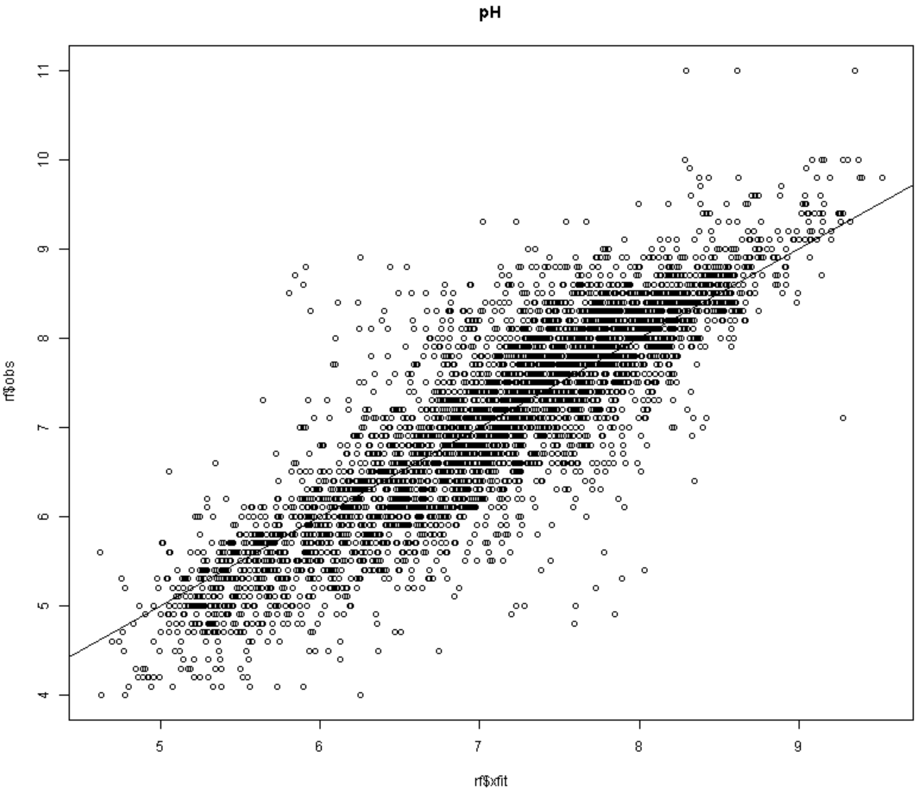

Fig. 7. Observed versus predicted clay content values. Predictions by 3D RF models with global covariate data. The solid line is the 1:1 line.

Table 3 shows the validation results for the 2D and 3D RF models fitted with global and local covariate data. Use of the non-linear random forest model resulted in a substantial improvement in prediction accuracy for the 2D and 3D models for both global and local covariate sets compared to the linear regression models. 3D RF models outperform 2D random forest models. Using full profile information and the addition of depth of observation as a covariate greatly benefits prediction accuracy.

The RMSE of the 3D RF models for carbon fitted with global covariate data was 14.8% smaller than the RMSE of the 3D linear regression model, for clay this was 22.4% and for pH 32.1%. The 2D RF models also performed better than the 2D linear models though the reductions in RMSE were not so large as for 3D: 6.4% for carbon, 6.3% for clay and 11.0% for pH. Like for the linear regression model, the models fitted with local covariates performed slightly better (clay, pH) or were as good as (carbon) the models fitted with global covariates.

Figures 2, 3 and 4 show the variable importance plots for the three 3D RF models. Depth (altitude) is by far the most important variable for predicting carbon, followed by the carbon (m_ORCDRC), bulk density (m_BLD) and pH (m_PHIHOX) SoilGrids maps, and elevation (af_DEMSRE5a). The most important variables for predicting clay are the mid-infrared reflectance images (af_M13RB7Axx), the SoilGrids clay map (m_CLYPPT) and depth (altitude). The most important variables for predicting pH are the SoilGrids pH map (m_PHIHOX), vegetation (af_M13EVIAxx) and elevation (af_DEMSRE5a).

Figures 5, 6 and 7 show scatterplots of the observed versus the predicted soil property values by the 3D RF models using global covariate data.

Conclusions

- 3D models perform as good as or slightly worse than 2D models. The difference in accuracy, however, is not large enough to invest efforts in expanding 3D *linear* regression models so that these allow for depth-specific coefficients for environmental covariates. Also because random forest modelling allows modelling of depth-specific effects of environmental covariates on the target soil property.
- Random forest models perform consistently better than linear regression models indicating that the relationship between the soil properties and the environmental covariates can better be modelled with a non-linear model than with a linear model.
- Models calibrated with local covariates perform slightly better than models calibrated with global covariates. The difference in accuracy, however, is small so that a global covariate datasets presents a good alternative for a local covariate set (which is more time-consuming to prepare).
- The findings regarding the accuracy of 2D versus 3D model predictions merit further research, preferably with a synthetic dataset.