Predicting density of Swiss forest soils by component wise gradient boosting
Soil density: Why?

Important soil parameter e. g. for
- calculations of stocks (nutrients, soil organic carbon [SOC])
- water holding capacity / soil porosity
- evaluation of soil compaction

My presentation
- how is it measured..
- what if no measurements available..
- why did we work on it right now..
- what data and method we used..
- how is this method working..
- what we found..
Soil density measurements

Method

- sampling of volumetric cores in soil pits
- drying at 105° C
- for density of fine soil fraction (< 2mm): sieving

Problems

- large work input (cost)
- large variability (compaction at sampling)
- bias (cores preferentially taken at spots with less gravel)

→ repeated measurements per horizon

Pedotransfer rules / functions

- soil density is correlated with other soil properties that are easier to measure
- attempts have been made to give general transfer function
- parameters mostly used: soil organic carbon, organic matter, grain size (clay etc.)

**Drawbacks**

- transfer to other study regions is limited!
- negative bias: underestimation of density (de Vos, 2005)
- recalibration necessary
Project at WSL: Forests soils and water balance in a changing climate

Background: Modeling of water retention curve for Swiss forests soils
Required soil parameters: soil texture, soil density, organic matter

Available data
- Measured densities at only 210 forest sites
- Other soil properties available from ~ 3,000 sites

Objectives
- develop density pedotransfer function for Swiss forest soil horizons
- evaluation of its performance
soil data (1/3) – what was measured?

- Density **measurements**: ~ 839 horizon (210 profile sites)
- 3 repeated measurements → median used
- For same horizons **partially** data on:
  - slope, soil depth, sample depth, horizon thickness
  - field estimate of organic matter
    - of soil color (hue, value, chroma)
    - of gravel content
    - of density (5 classes)
  - share of sand, silt, clay fraction
  - pH, SOC content
  - cation exchange capacity, base saturation, total nitrogen content
  - wetness characteristics (9 categories),
    soil depth limitation by rock or ground water
soil data (2/3) – missing values in covariates

**Problem:** data set of covariates (soil properties) incomplete in calibration set, but also in prediction data.

- prediction set incomplete: *soil color*

1) **Imputation** (e.g. with „missforest“ using Random Forest)
   - precondition: subset with NA are not biased
   - check of descriptive statistics → imputation not possible here

2) **Eliminate missing values**
   - a) omit soil horizon with NA
     - *SOC (127 horizons)*
   - b) omit covariate with large amount of NA
     - *total nitrogen, cation exchange capacity, basic saturation*
soil data (3/3) – profile location

- calibration set (134 sites with 559 horizon)
- validation set (34 sites with 131 horizon)
Method (1/3) – componentwise gradient boosting

- "weak" learner algorithm (small step size $\nu$)
  - base-learners, e.g.
    - linear
    - smooth nonlinear
    - trees
  - base-learners summed up for prediction

\[ Y(s) = \sum_j f(x(s))_j + \epsilon(s) \]

1. Compute residuals (negative gradient)
2. Fit base-learners and select best fitting
3. Update parameters stepwise as sum of previously fitted estimates of the base-learners
   \[ f^m = f^{m-1} + \nu u^m \]
Method (2/3) – advantages

- non-linearity
- smooth spatial surface
- exclusion of non-relevant covariates
- good predictive power expected
- robust loss functions
- interpretability (depending on base-learners)

overfitting?
Method (3/3) – model building and selection

Model
\[ Y = X\beta + f(x_1) + f(x_1 \times x_2) + \ldots + \epsilon \]

Model building
A. transformation of response and covariates

1. Box-Cox transformation / log / sqrt

B. selection of relevant covariates

2. selection of linear covariates by 10fold cross-validation

3. merge categories of selected covariates with partial residual plots

4. fit boosting with smooth covariates on residuals

5. select final model with residual plots and cross-validation
Resulting Model
resulting model (1/2) – linear parameters

linear parameters in best-fit model:

$\uparrow$ sample depth (sqrt)
$\downarrow$ slope angle (sqrt)
$\uparrow$ field estimate of density (3 classes)
$\uparrow\uparrow$ 8 aggregated soil map units (1:200'000)

linear interactions in best-fit model:

$\uparrow\uparrow$ profile depth for 5 soil map units
$\uparrow\uparrow$ share of silt and clay / soil organic carbon (SOC) for unlimited, ground water or rock limited soils
resulting model (2/2) – smooth elements
Model Performance

Validation with measurements of 131 horizon not used for calibration.

<table>
<thead>
<tr>
<th>Method</th>
<th>RMSE / MADE</th>
<th>$R^2$ (robust)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>0.205 / 0.235</td>
<td>0.669</td>
</tr>
<tr>
<td>Linear + Smooth</td>
<td>0.178 / 0.188</td>
<td>0.771</td>
</tr>
</tbody>
</table>

* root mean squared error, median absolute deviation error
Summary / Conclusion

- **expanded pedotransfer function** with 8 input covariates and spatial position
- model improvement by including **non-linear terms** with boosting algorithm
- **successful model selection** without overfitting
- **satisfactory model performance** (robust $R^2 = 0.77$ with independent validation set)

**Outlook**

- Investigate in final fit with robust methods
- Investigate in predictive distribution to compute standard errors to the predictions