Comparison of a k-NN approach and regression techniques for single tree biomass estimation

8th FIA Symposium of the USDA Forest Service, October 16-19 Monterey CA

Lutz Fehrmann & Christoph Kleinn
Introduction

• On the way to more general biomass estimation approaches on single tree level a compilation of readily available datasets is required and useful.
  – This might be very challenging because the willingness to share data is not always well developed

• Once a comprehensive enough database is given, also instance based methods like the $k$-NN approach can be applied
The $k$-NN approach

- The $k$-NN method is based on a non-parametric pattern recognition algorithm
- Basic idea is to classify an unknown feature of an instance according to its similarity to other known instances stored in a database
  - Based on a calculated distance the $k$ nearest (most similar) neighbours to a certain query point are identified and under the assumption that they are also similar concerning their target values, used to derive an estimation
The $k$-NN approach

- Different to regression analysis or process model approaches no functional relationships between the variables have to be formulated.
- The estimations are derived as local approximations, not as a global function.

$$\hat{f}(x_q) \leftarrow \frac{\sum_{i=1}^k w_k f(x_k)}{\sum_{i=1}^k w_k}$$
Distance function

- As distance function given multivariate measures from cluster- or discriminant analyses can be used:

\[
d_w(x_i, x_j) = \sum_{r=1}^{n} \left( w_r \cdot \frac{|x_{ir} - x_{jr}|}{\delta_r} \right)^c \left[ \frac{1}{c} \right]
\]

- \(d_w\) = weighted distance between two instances
- \(n\) = number of variables
- \(w_r\) = weight assigned to the variable \(r\)
- \(r\) = \(r^{th}\) variable of an instance
- \(x_i, x_j\) = instances
- \(\delta_r\) = standardisation factor (range of variable or multiple of \(\sigma\) of variable \(r\))
- \(c\) = \(\geq 1\) Minkowski constant (2 = euclidean distance)
Implementation

- To run the $k$-NN Algorithm a suitable software application and database is necessary
Size of the Neighbourhood

• Instance based methods come along with a typical bias-variance dilemma that is in parts influenced by asymmetric neighbourhoods at the edges of the feature space of the training data.
Cross validation

• To determine the parameters for the distance- and weighting function as well as \( k \) cross-validation methods are suitable
  – Therefore an estimation for every tree is derived based on the remaining N-1 trees of the training data.
  – The definition of optimal weighting factors, the size of the neighbourhood and parameters of the distance function can be approximated by an iterative process or by means of optimization algorithms.
Example

• A large dataset of Norway spruce and Scots pine trees (provided by the METLA) was used to evaluate the $k$-NN approach in comparison to regression models
  – Datasets where split into „modelling“ ($n=143$ for spruce, $n=145$ for pine) and „test“ ($n=60$ each) subsets
  – Modelling subsets where used to estimate regression coefficients and as training data for the $k$-NN algorithm (independent variables are dbh and height)
Example

• Predictions for the „test“ datasets were used to compare the performance of both approaches by means of different error criterions

![Graphs showing number of trees by dbh class for spruce and pine](image)

![Legend for spruce and pine graphs](image)
Example

- Multiple cross-validation was used to minimize the RMSE and bias by an approximation of optimal feature weights and parameter settings.
Example

- Alternative to a fixed number of neighbours also a kernel-method was applied
  - In this case neighbours are considered up to a defined standardized distance
Example

- Linear mixed effect models and simple linear models were used as reference.
Example results

- The RMSE could be reduced in comparison to regression models for both species:

<table>
<thead>
<tr>
<th>Regression models / Approach</th>
<th>RMSE</th>
<th>rMSE%</th>
<th>MAPE</th>
<th>ME</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scots pine</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \ln \alpha + \beta \ln d_k + \chi \ln h_k + e_k )</td>
<td>20.68</td>
<td>15.79</td>
<td>9.67</td>
<td>-2.562</td>
</tr>
<tr>
<td>( \ln \alpha + \ln a_k + \beta \ln d_k + \chi \ln h_k + e_k )</td>
<td>19.76</td>
<td>15.00</td>
<td>9.21</td>
<td>-1.718</td>
</tr>
<tr>
<td><strong>k-NN</strong></td>
<td>19.41</td>
<td>14.54</td>
<td>12.61</td>
<td>0.009</td>
</tr>
<tr>
<td><strong>Norway Spruce</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \ln \alpha + \beta \ln d_k + \chi \ln(h/d) + e_k )</td>
<td>22.91</td>
<td>19.85</td>
<td>13.80</td>
<td>-1.630</td>
</tr>
<tr>
<td>( \ln \alpha + \ln a_k + \beta \ln d_k + \chi(h/d) + e_k )</td>
<td>20.31</td>
<td>17.36</td>
<td>13.73</td>
<td>-0.398</td>
</tr>
<tr>
<td><strong>k-NN</strong></td>
<td>19.19</td>
<td>16.42</td>
<td>13.98</td>
<td>-0.493</td>
</tr>
</tbody>
</table>
Outlook

- The $k$-NN method offers the possibility to include additional variables (for example meta information about sites or tree species) without knowledge about the cause-and-effect relationships.

- In case of using multiple search variables the implementation of optimization approaches, like the genetic algorithm (Tomppo and Halme, 2004), for feature weighting is required and useful.
• Thank you!

This study was conducted in close collaboration with the Finnish Forest Research Institute (METLA). We thank Errki Tomppo and Aleksi Lehtonen!