Deutsche
Forschungsgemeinschaft
DFG

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

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

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 challaging 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

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

Distance function

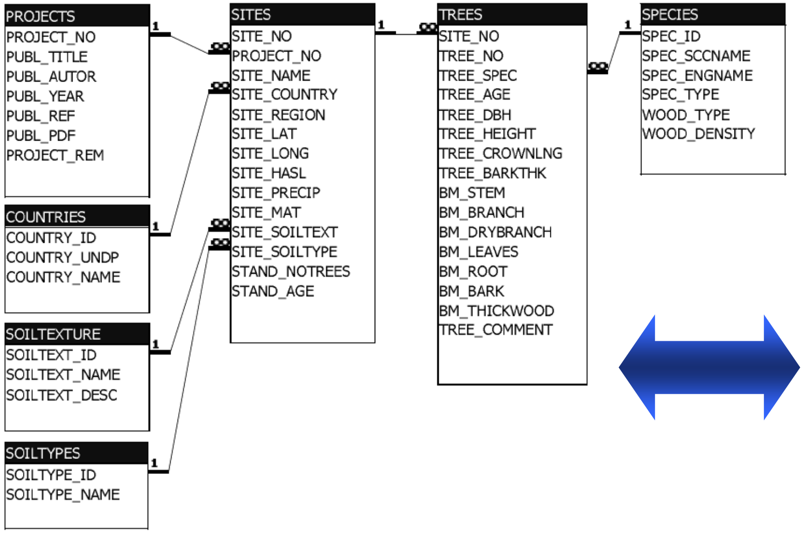
- As distance function given multivariate measures from cluster- or discriminant analyses can be used:

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

d_w	= weighted distance between two instances
n	= number of variables
w_r	= wheigh assigned to the variable r
r	= r^{th} variable of an instance
(x_i,x_j)	= instances
δr	= standardisation factor (range of variable or multiple of σ of variable r)
c	= >=1 Minkowski constant (2= euclidean distance)

Implementation

- To run the *k*-NN Algorithm a suitable software application and database is necessary



PROJECTS
PROJECT_NO
PUBL_TITLE
PUBL_AUTOR
PUBL_YEAR
PUBL_REF
PUBL_PDF
PROJECT_REM

COUNTRIES
COUNTRY_ID
COUNTRY_UNDP
COUNTRY_NAME

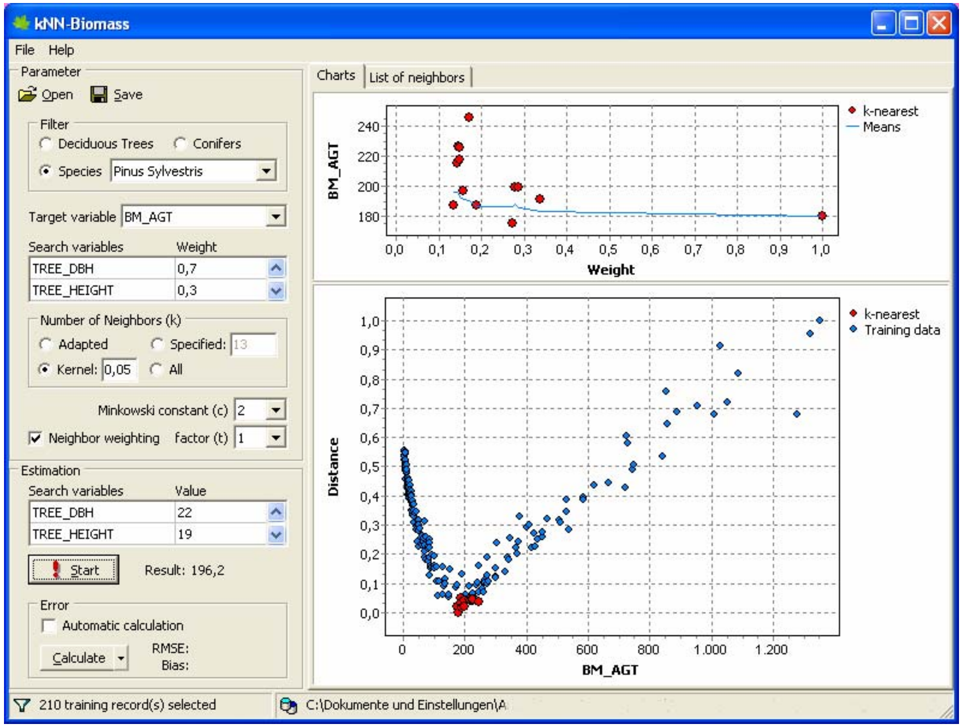
SOILTEXTURE
SOILTEXT_ID
SOILTEXT_NAME
SOILTEXT_DESC

SOILTYPES
SOILTYPE_ID
SOILTYPE_NAME

SITES
SITE_NO
PROJECT_NO
SITE_NAME
SITE_COUNTRY
SITE_REGION
SITE_LAT
SITE_LONG
SITE_HASL
SITE_PRECIP
SITE_MAT
SITE_SOILTEXT
SITE_SOILTYPE
STAND_NOTREES
STAND_AGE

TREES
SITE_NO
TREE_NO
TREE_SPEC
TREE_AGE
TREE_DBH
TREE_HEIGHT
TREE_CROWNLNG
TREE_BARKTHK
BM_STEM
BM_BRANCH
BM_DRYBRANCH
BM_LEAVES
BM_ROOT
BM_BARK
BM_THICKWOOD
TREE_COMMENT

SPECIES
SPEC_ID
SPEC_SCCNAME
SPEC_ENGNAME
SPEC_TYPE
WOOD_TYPE
WOOD_DENSITY



kNN-Biomass

Parameter

Filter
○ Deciduous Trees ○ Conifers
● Species: Pinus Sylvestris

Target variable: BM_AGT

Search variables	Weight
TREE_DBH	0,7
TREE_HEIGHT	0,3

Number of Neighbors (k)
○ Adapted ○ Specified: 13
● Kernel: 0,05 ○ All

Minkowski constant (c): 2
☑ Neighbor weighting factor (t): 1

Estimation

Search variables	Value
TREE_DBH	22
TREE_HEIGHT	19

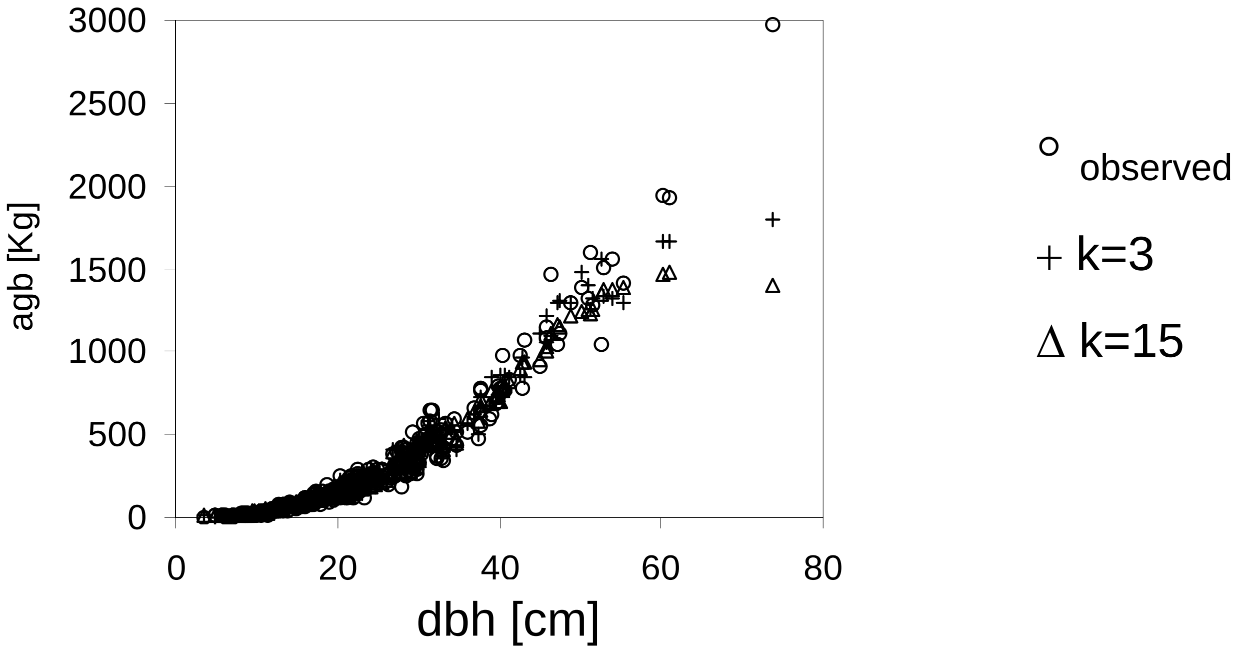
Result: 196,2

Error
□ Automatic calculation
RMSE:
Bias:

210 training record(s) selected
C:\Dokumente und Einstellungen\A

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



GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

Cross validation

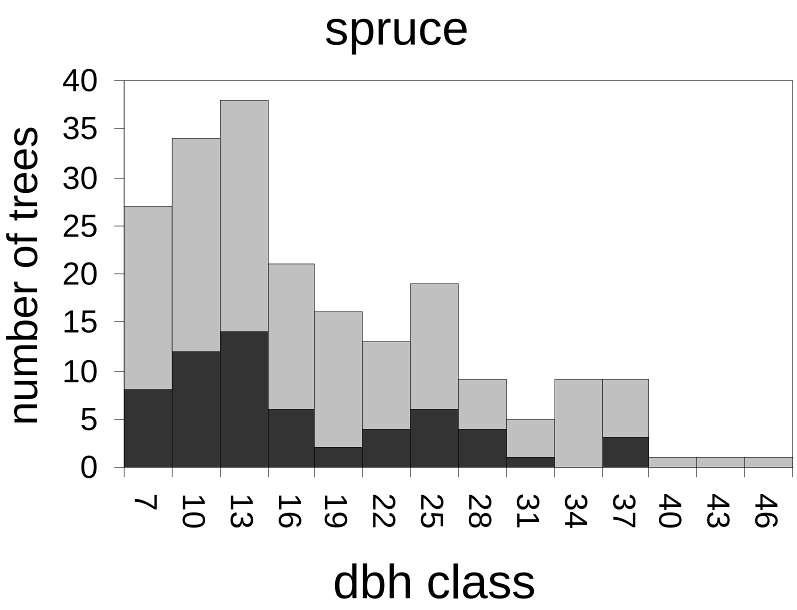
- To determine the parameters for the distance- and weighting function as well as *k* cross-validation methods are suitable
  - Therefore an estimation for every tree is derived based on the remaining N-1 trees of the training data.
  - The definition of optimal weighting factors, the size of the neighbourhood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

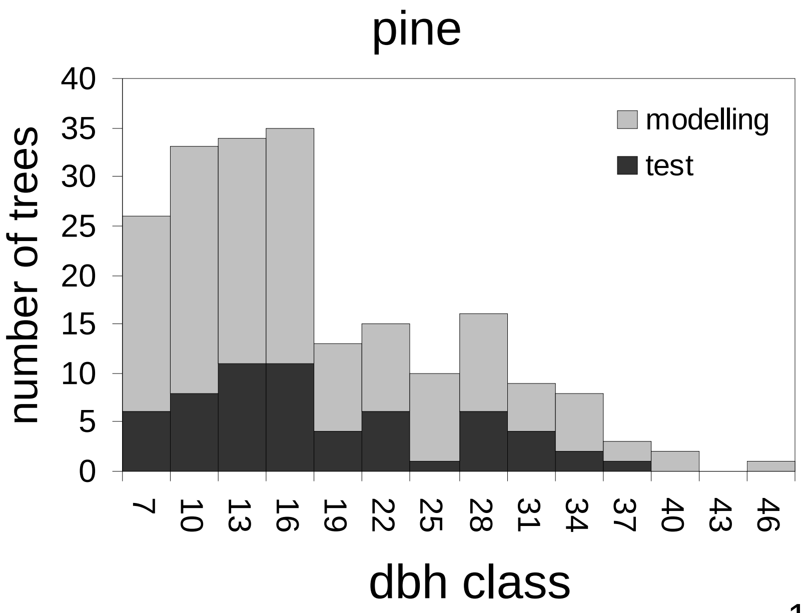
Example

- 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 coefficiants and as training data for the *k*-NN algorithm (independent variables are dbh and height)

Example

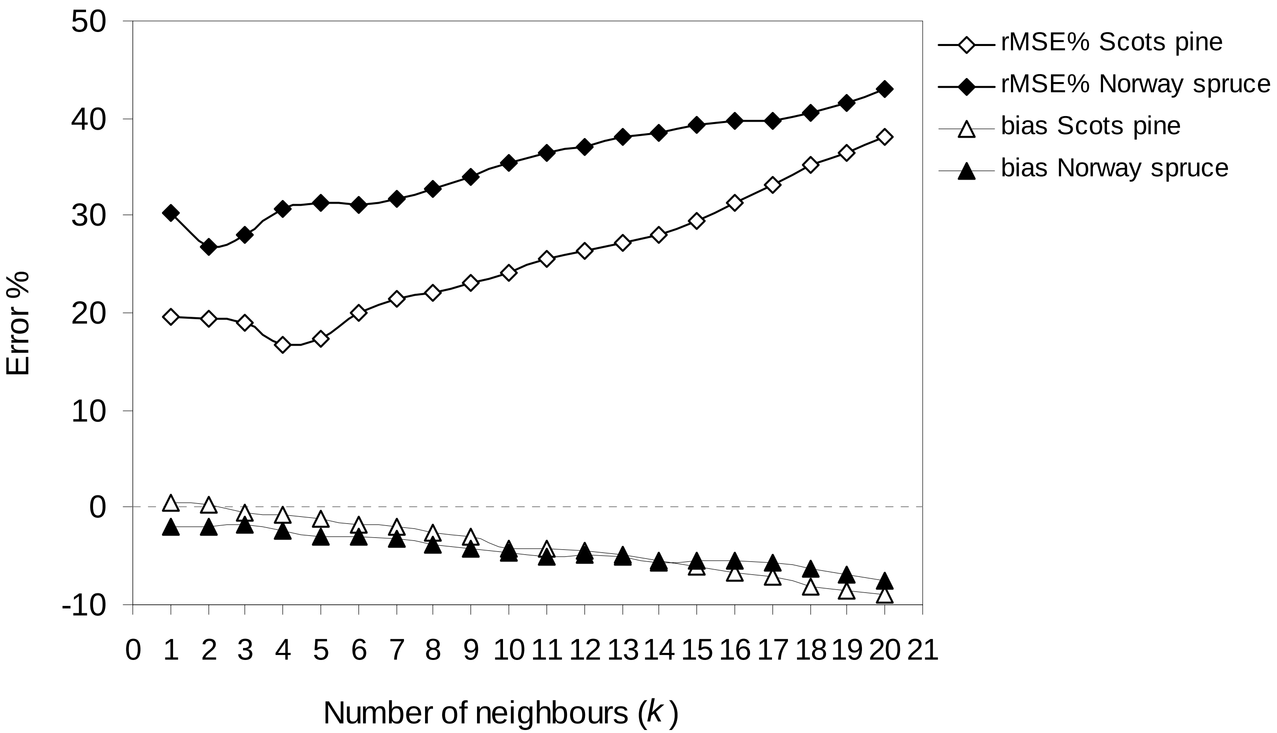
- Predictions for the „test“ datasets were used to compare the performance of both approaches by means of different error criterions





Example

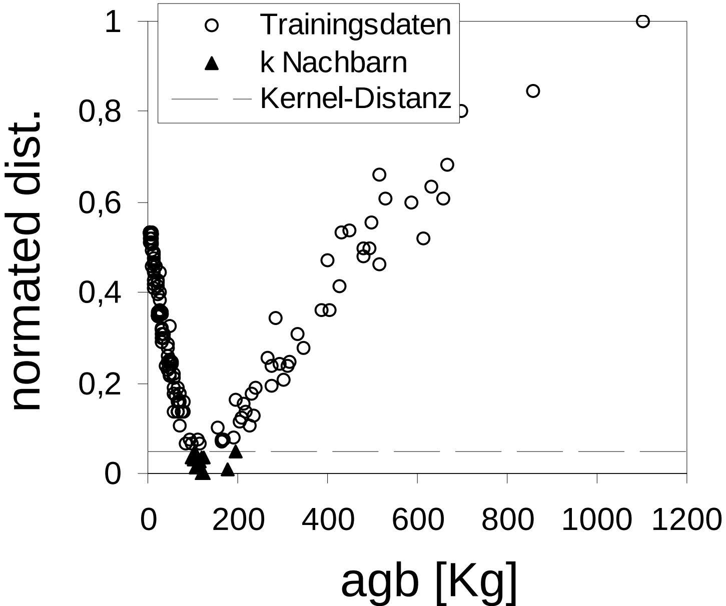
- Multiple cross-validation was used to minimize the RMSE and bias by an approximation of optimal feature weights and parameter settings.

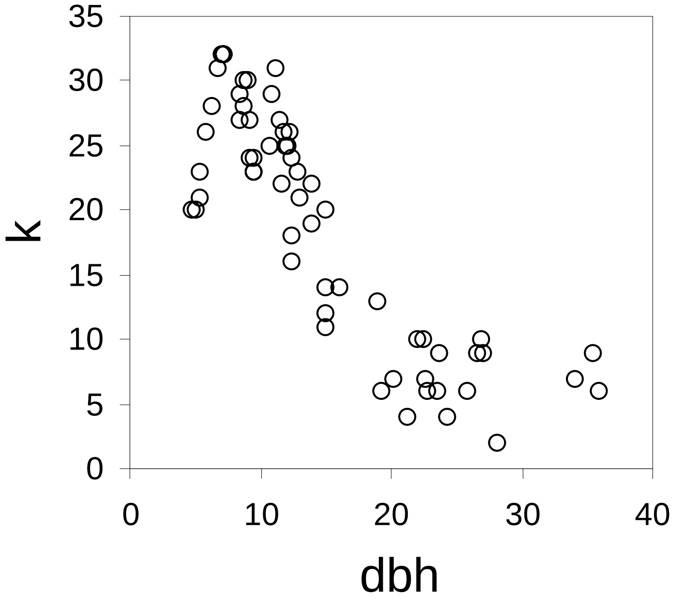


GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

Example

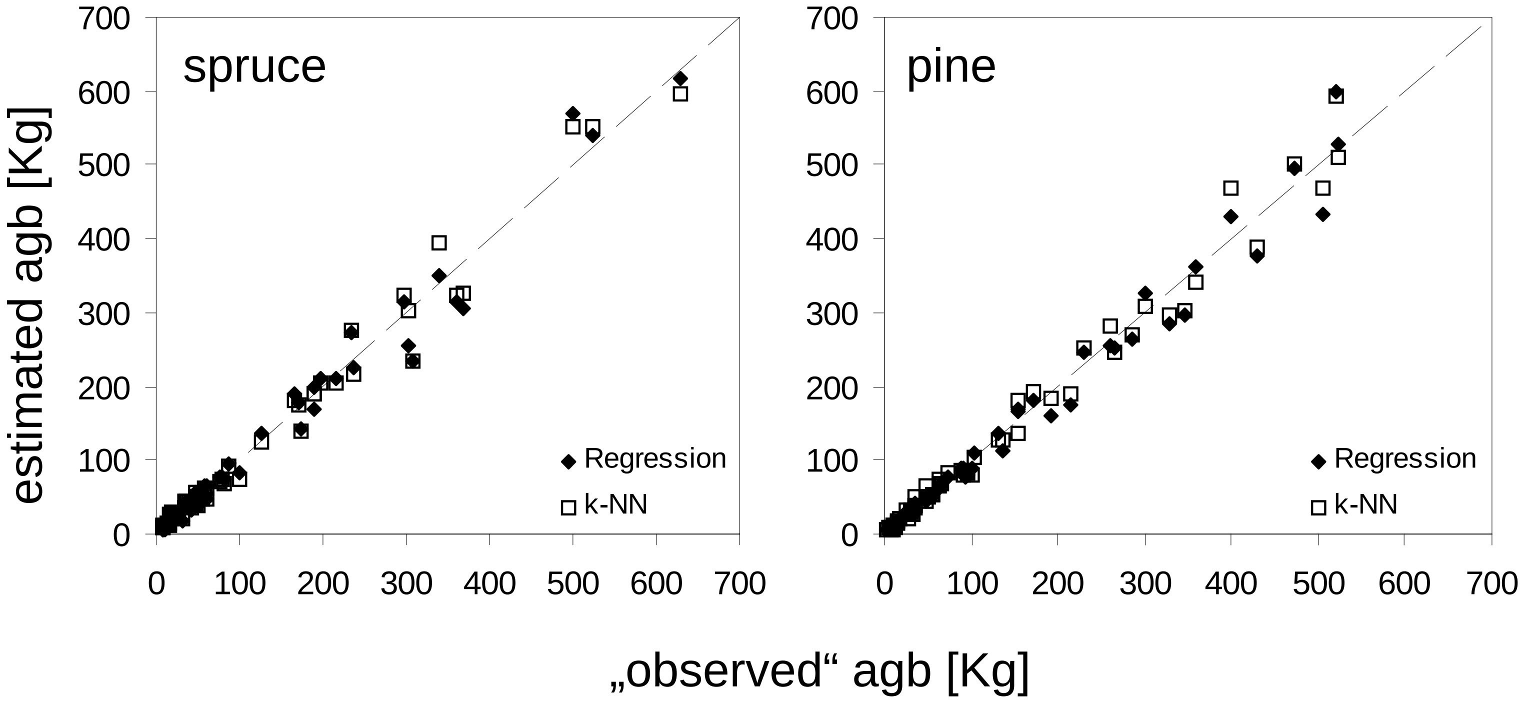
- Alternative to a fixed number of neighbours also a kernel- method was applied
  - In this case neighbours are considered up to a defined standardized distance





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:

Regression models / Approach	RMSE	rMSE%	MAPE	ME
Scots pine				
$\ln agb_{ki} = \ln\alpha + \beta\ln d_{ki} + \chi\ln h_{ki} + e_{ki}$	20.68	15.79	9.67	-2.562
$\ln agb_{ki} = \ln\alpha + \ln a_k + \beta\ln d_{ki} + \chi\ln h_{ki} + e_{ki}$	19.76	15.00	9.21	-1.718
k-NN	19.41	14.54	12.61	0.009
Norway Spruce				
$\ln agb_{ki} = \ln\alpha + \beta\ln d_{ki} + \chi\ln\left(h/d\right)_{ki} + e_{ki}$	22.91	19.85	13.80	-1.630
$\ln agb_{ki} = \ln\alpha + \ln a_k + \beta\ln d_{ki} + \chi\left(h/d\right)_{ki} + e_{ki}$	20.31	17.36	13.73	-0.398
k-NN	19.19	16.42	13.98	-0.493

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 optimazation 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!

Deutsche
Forschungsgemeinschaft
DFG