Some course information:

Seminar based on pair work, Wednesday 29.11.

Final exam 12-3pm, Thursday 30.11.

Generalized boosted models

Modelling in physical geography, 5cr, 30.10.-30.11.2017

1. Theory

GBMs are also known as **BRTs** (boosted regression trees), even more flexible than GAMs!

A form of advanced regression using both statistical and machine learning traditions. Combines strengths of two algorithms:

- **Regression trees** (models that relate a response to their predictors by recursive binary splits)
- **Boosting** (builds and combines a collection of models to give improved predictive performance). Similar to bagging, stacking and model averaging, but boosting is sequential → fitting one model after another, seeking to minimize residuals weighted by the previous model's errors

The final BRT model can be understood as an additive regression model in which individual terms are simple trees, fitted in a forward, stagewise fashion

- Does not aim to fit a single parsimonious model

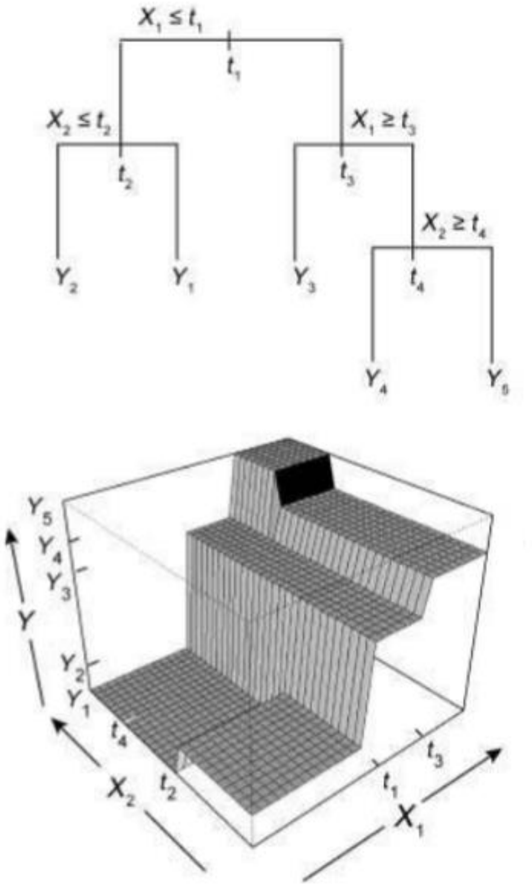


Fig. 1. A single decision tree (upper panel), with a response Y, two predictor variables, X_1 and X_2 and split points t_1, t_2, etc. The bottom panel shows its prediction surface (after Hastie *et al.* 2001)

2. Applications

Generally good predictive performance

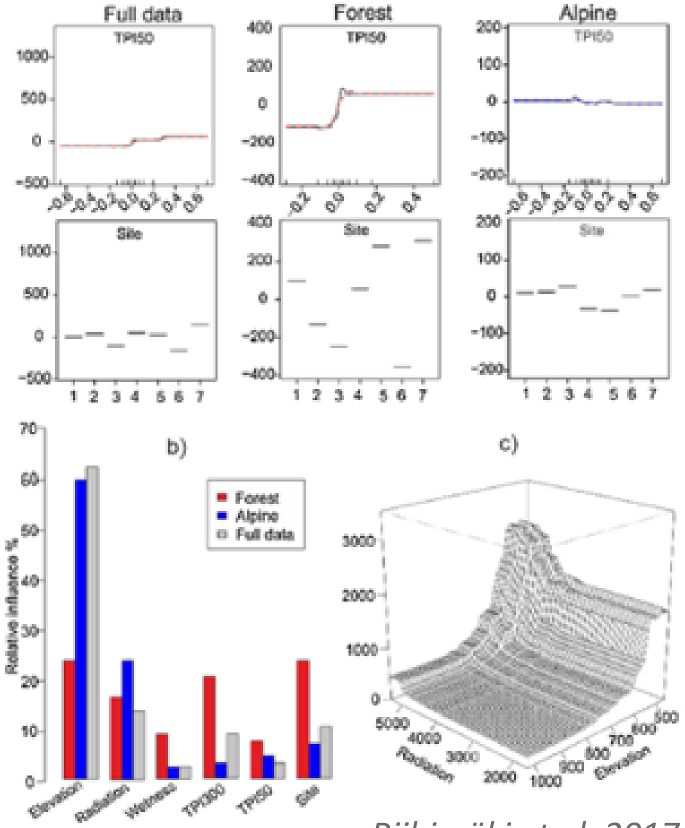
Complex, but can be summarized to give powerful insight

No need for *a priori* data alteration or elimination of outliers

- Can fit complex nonlinear relationships
- Can handle different types of predictor variables
- Can handle correlated variables
- Can handle NAs

Useful for variable selection

- Reliable identification of relevant variables
- Reliable identification of possible interactions between variables



Riihimäki et al. 2017

3. Examples in R – GBM model building

> library(gbm) #opens gbm library

Input: ***gbm()*** expects a model formula to be supplied, specifying model structure to fit

> gbm1 <- gbm(Y~ X1 + X2 + X3, distribution = "gaussian", n.trees = 3000, shrinkage = 0.001, interaction.depth = 4, data = theData)

where:

Y = response variable

Xi = predictor variable/s

distribution = same as the families in GLM and GAM; Gaussian, bernoulli (binomial in gbm), poisson

n.trees = number of trees to fit; i.e. number of iterations; default = 100, larger preferable (dismo's "gbm.step" function used to determine optimal n.trees)

shrinkage = learning rate; determines impact of each tree on final outcome; lower values preferred as it is better to improve a model by taking many small steps than by taking fewer large steps→ allow model to generalize; but lower values require higher n.trees to model all relations and will be slow

interaction.depth = number of leaves; maximum depth of variable interactions. 1 = additive model (default); 2 = up to 2-way interactions; etc.

data = data used

The best parameters depend on the data. Other models parameters that you can set:

Output: a fitted ***gbm*** object to be analyzed

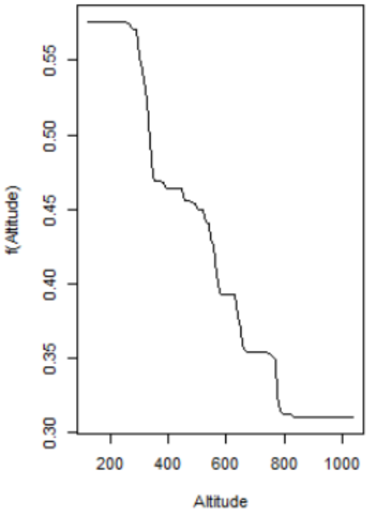
```
> gbm1 <- gbm(empher~altitude + wab, distribution=”bernoulli")
> gbm2 <- gbm(totalspr~altitude + wab, distribution =”poisson")
> gbm3 <- gbm(gdd~altitude + wab, distribution =”gaussian")
```

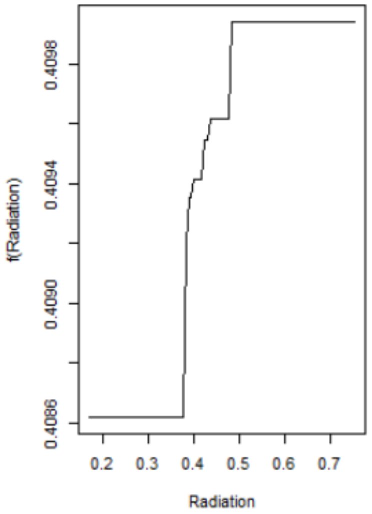
3. Examples in R – response curves

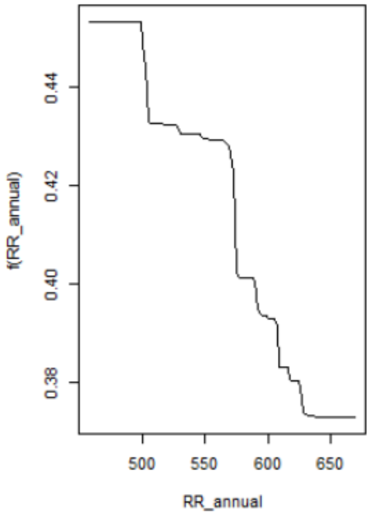
Example: GBM of NDVI ~altitude, radiation, rainfall, July temperature

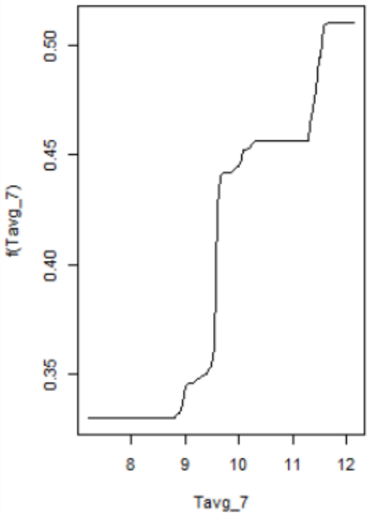
```
> set.seed(0)
# build GBM model
> gbm1 <- gbm(NDVI ~ Altitude + Radiation + RR_annual + Tavg_7,
distribution="gaussian", n.trees = 3000, data=RastiData)

# estimate optimal number of iterations with "OOB"
> best.iter <- gbm.perf(gbm1, method="OOB", plot.it=TRUE)

# plot response curves
> par(mfrow=c(2,3)) # to plot in one window
> plot.gbm(gbm1, 1, best.iter) # plots *explvar1
> plot.gbm(gbm1, 2, best.iter) # plots *explvar2..
> plot.gbm(gbm1, 3, best.iter)
> plot.gbm(gbm1, 4, best.iter)
```

- X-axis = variation in predictor variable
- Y-axis = the modelled values of the response variable plotted against an even distribution of the predictor, after accounting for the average effects from other predictors
- Of interest: the shape, direction, and thresholds of the response curve









```
# plot response curves separately
> plot.gbm(gbm, *n, best.iter)
```

3. Examples in R – variable relative influence

With GBMs, the model summary gives a dataframe and a bar graph of the summary of variable relative importance

For `distribution="gaussian"` it returns the reduction of squared error attributable to each variable. For other loss functions it returns the reduction attributable to each variable in sum of squared error in predicting the gradient on each iteration (i.e. tree)

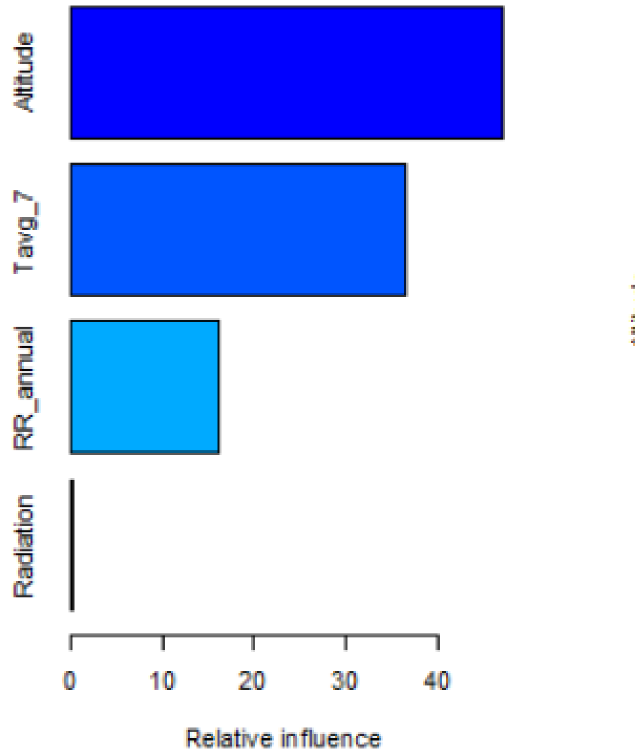
Calculated from the:

- frequency a variable is selected as a model predictor
- improvement resulting from the inclusion of the variable

The *relative* influence (or contribution) of each variable is scaled so that the sum adds to 100

- Higher numbers indicate a stronger influence on the response variable

Relative influence of altitude, radiation, rainfall, temperature on NDVI



```
# dataframe and graph of var relative influence
> summary(gbm)
```

3. Examples in R – variable selection with GBM

NDVI ~altitude, radiation, rainfall, temperature

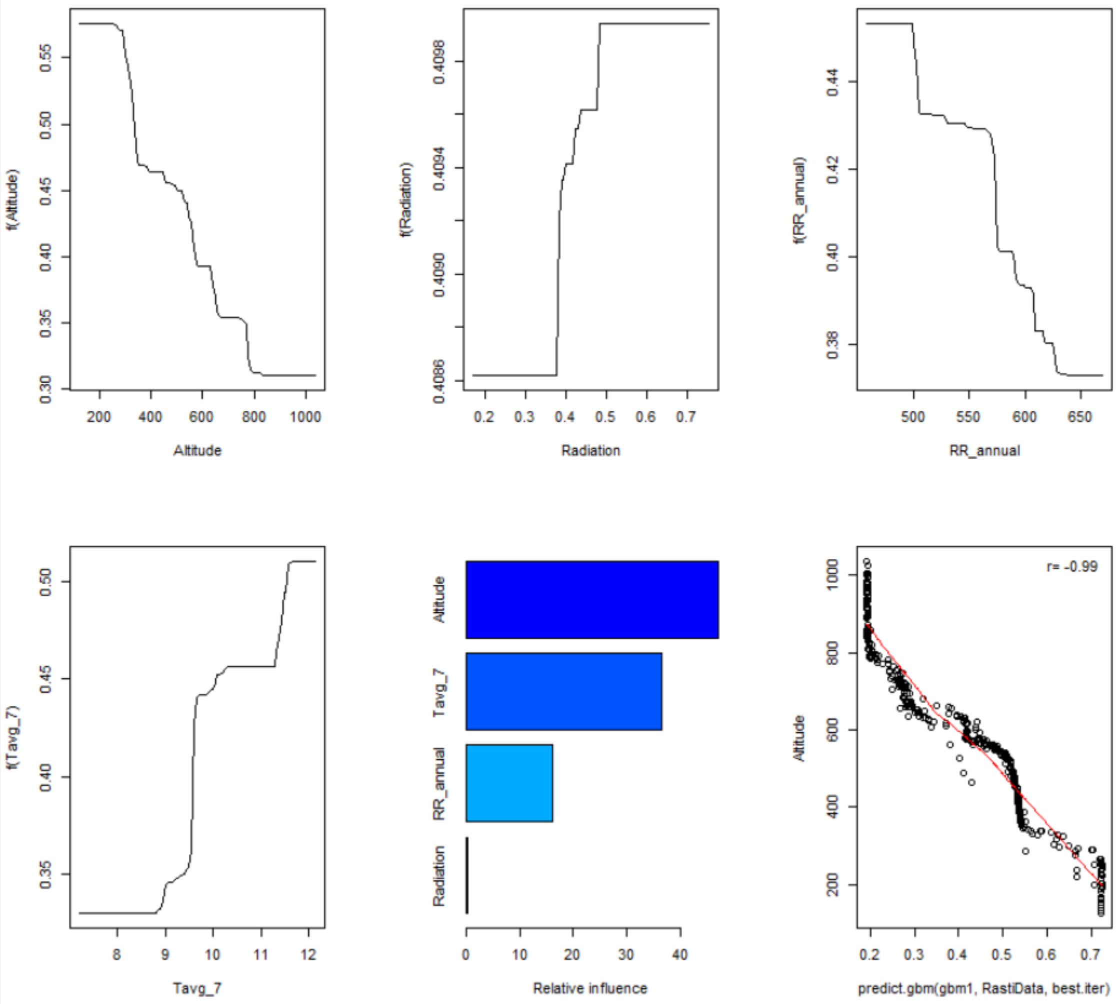
GBMs utilize many weakly important variables to improve on predictions

Non-informative predictors are largely ignored when fitting trees

No need to remove variables from model, unless:

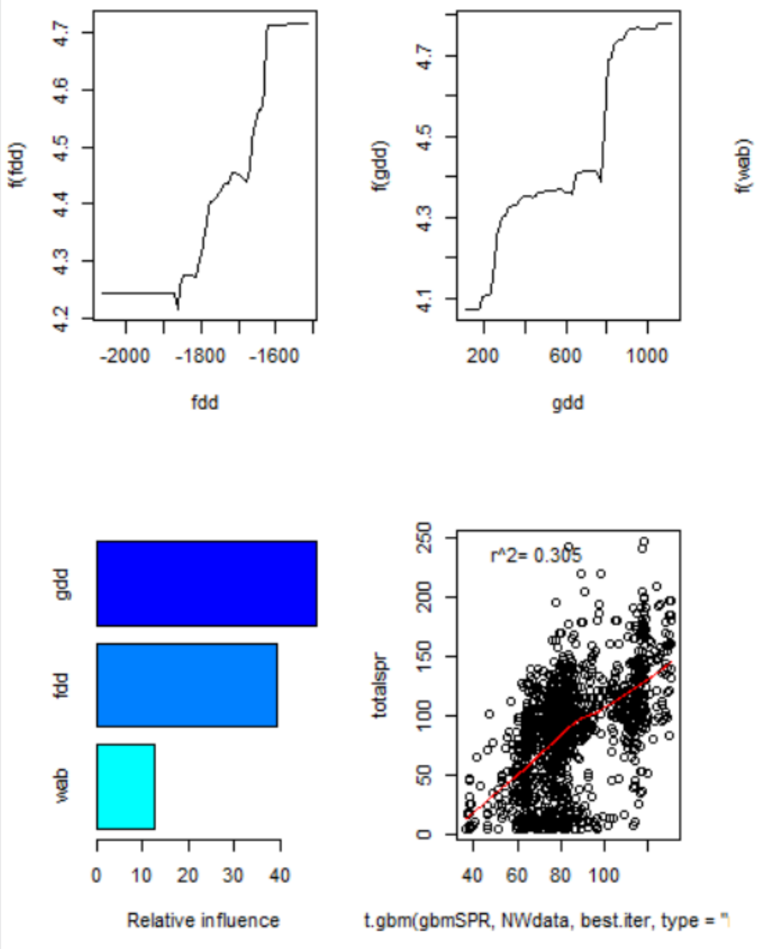
- To ease implementation (faster modelling)
- Using small data sets (redundant predictors can degrade performance by increasing variance)

Note: use ***> set.seed(0)*** *before calling gbm to ensure that relative influence is the same if you (or someone else) need to run the same gbm again*



3. Examples in R – predict and model fit

Total species richness explained by climate



R^2 = 0.305 → climate explains ~30% of the variability of total species richness around its mean

```
> gbmSPR <- gbm(totalspr ~ fdd + gdd + wab,
distribution="poisson", n.trees = 3000,
interaction.depth = 4, data=NWdata)
> best.iter <-
gbm.perf(gbmSPR,method="OOB",plot.it=TRUE)
> par(mfrow=c(2,3))
> plot.gbm(gbmSPR,1,best.iter)
> plot.gbm(gbmSPR,2,best.iter)
> plot.gbm(gbmSPR,3,best.iter)
> summary(gbmSPR,n.trees=best.iter)
> var  rel.inf
gdd gdd 47.88899
fdd fdd 39.37178
wab wab 12.73923

# plot predicted values
> plot(predict.gbm(gbmSPR, NWdata,
best.iter, type="response"), totalspr)
> lines(lowess(predict.gbm(gbmSPR, NWdata,
best.iter, type="response"), totalspr), col
= "red")
> r2data <- cor.test(predict.gbm(gbmSPR,
NWdata, best.iter), totalspr)
> SPRr2 <- (r2data$estimate)^2
> legend("topleft",paste
("r^2=",round(SPRr2, 3)), bty="n") # r^2
```

GBMs in R – main functions

- *library(gbm)* = opens the gbm library
- *?gbm =* get help on gbm
- *gbm() =* for fitting the model
- *gbm.perf()* = for estimating the best model performance, estimates optimal number of boosting iterations for gbm object (three methods available: an independent test set (test); out-of-bag estimation (OOB); and v-fold cross validation (cv)
- *summary() =* for extracting relative influence of predictors
- *plot.gbm() =* for viewing response curves, plotted individually
- *predict.gbm() =* use the fitted model to predict *Yi* at *Xi,*

  remember → (gam, type=”response”) inside the function

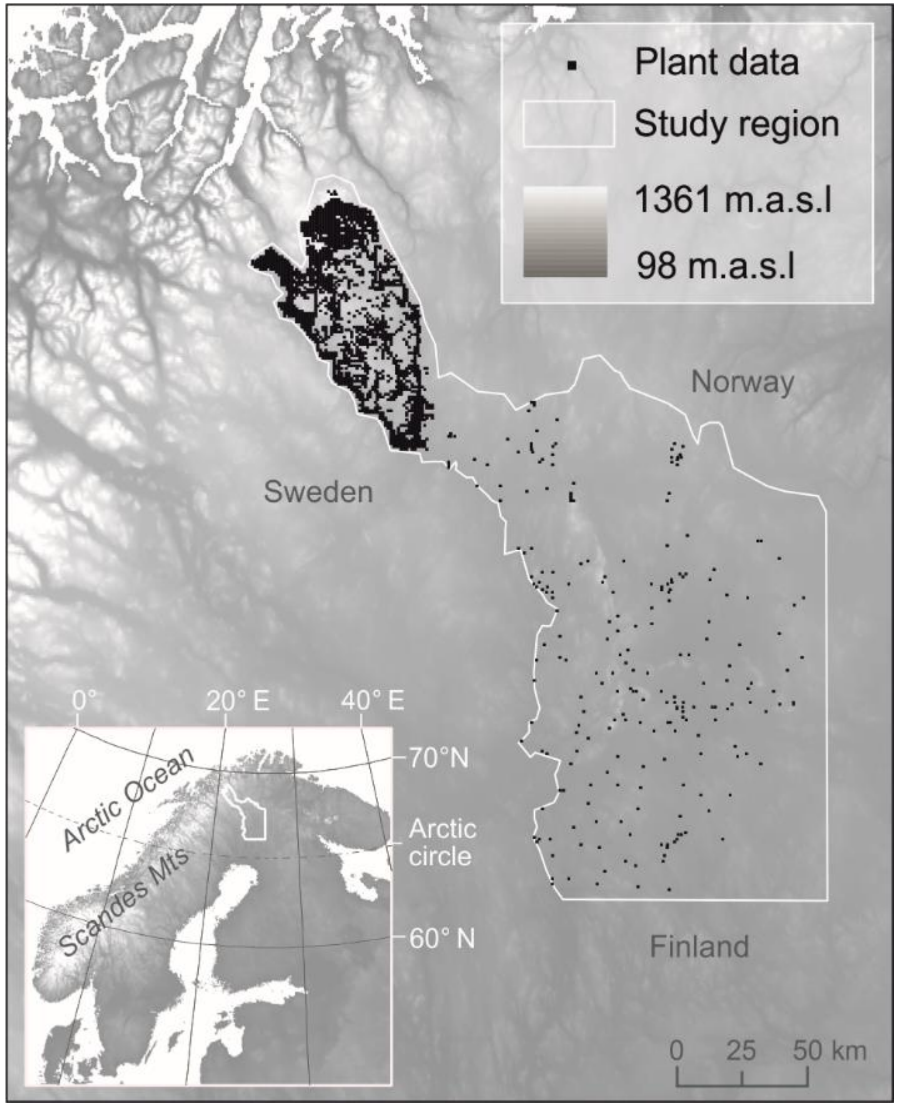
Practical – modelling species richness in Northwestern Lapland with GBM

R-script: *”GBM_Practical1.pdf”*

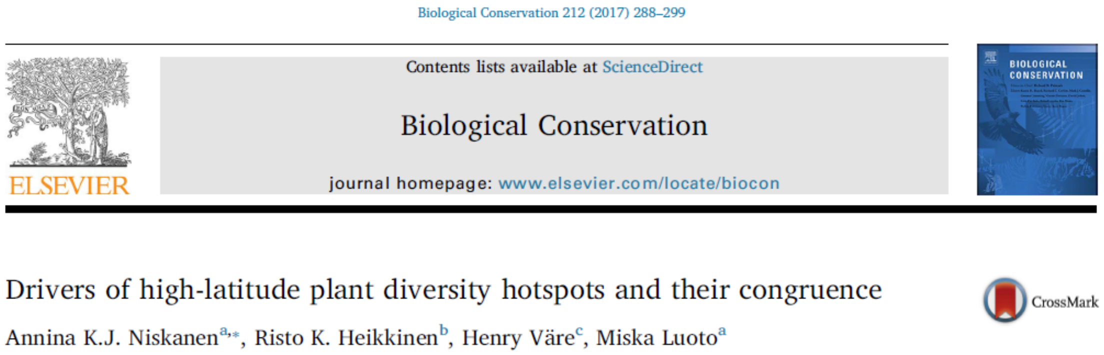
Data: *”NW_Lapland_data.csv”* (same as with GLM and GAM practicals)

- totalspr = total vascular plant species richness
- rarespr = rare vascular plant species richness
- fdd = freezing degree days; overwintering temp conditions
- gdd = growing degree days; growing season temp conditions
- wab = water balance; moisture conditions
- calc = calcareousness of soil ; indicates soil pH
- relalt = relative altitude; topographic roughness
- altitude = mean altitude
- *betnan*
- *dryoct*
- *empher*
- *gersyl*
- *linbor*
- *phycae*
- *rangla*
- *vacmyr*
- *vacvit*

distributions of vascular plants (presence/absence) within a given cell



Practical – modelling species richness in Northwestern Lapland with GBM



Biological Conservation 212 (2017) 288–299

Contents lists available at ScienceDirect

Biological Conservation

journal homepage: www.elsevier.com/locate/biocon

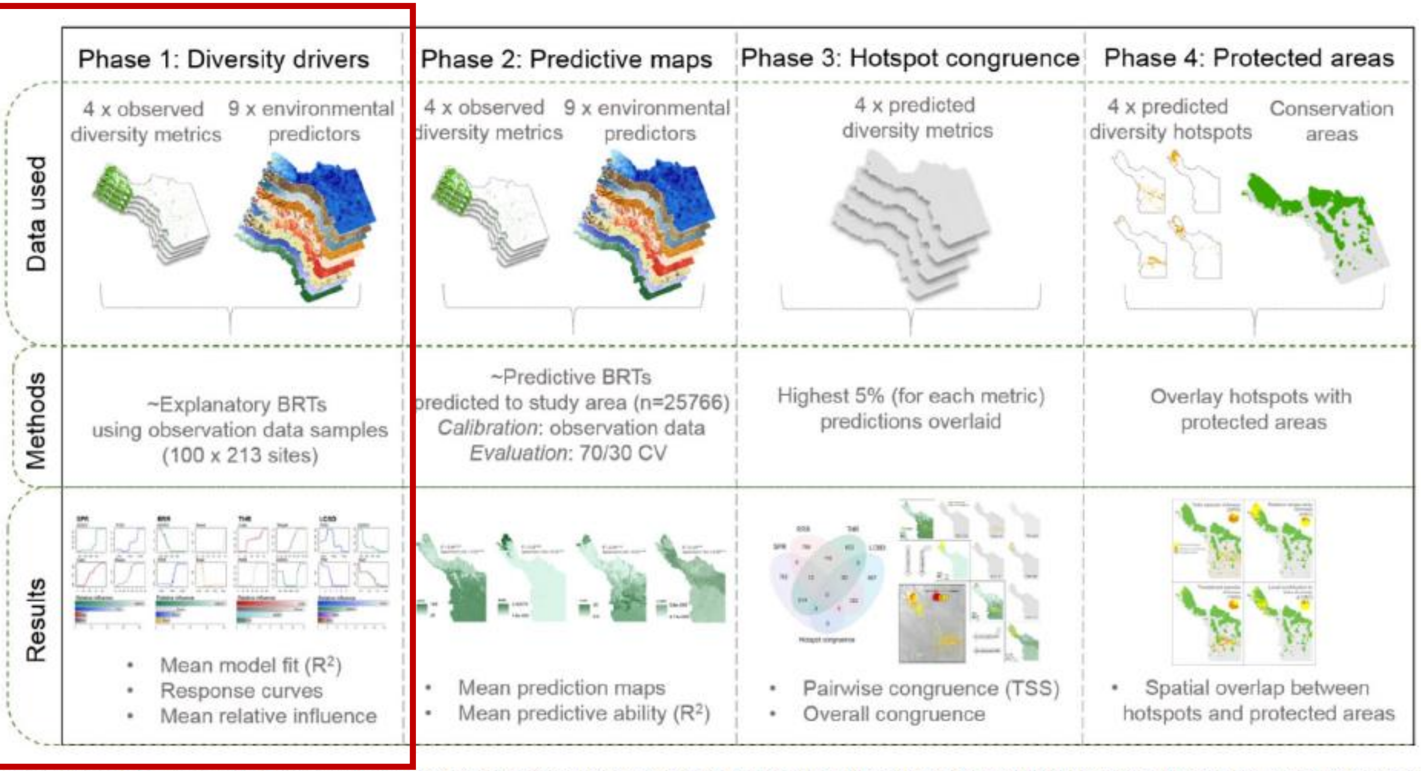
Drivers of high-latitude plant diversity hotspots and their congruence

Annina K.J. Niskanen[a,*], Risto K. Heikkinen[b], Henry Väre[c], Miska Luoto[a]

We are using a subset of the same data, but focusing on one diversity metric and five environmental predictors

We will build a GBM (BRT) for the metric ”totalspr”….

… and extract the same information for our results



	Phase 1: Diversity drivers	Phase 2: Predictive maps	Phase 3: Hotspot congruence	Phase 4: Protected areas
Data used	4 x observed diversity metrics 9 x environmental predictors	4 x observed diversity metrics 9 x environmental predictors	4 x predicted diversity metrics	4 x predicted diversity hotspots Conservation areas
Methods	~Explanatory BRTs using observation data samples (100 x 213 sites)	~Predictive BRTs predicted to study area (n=25766) *Calibration*: observation data *Evaluation*: 70/30 CV	Highest 5% (for each metric) predictions overlaid	Overlay hotspots with protected areas
Results	• Mean model fit (R^2) • Response curves • Mean relative influence	• Mean prediction maps • Mean predictive ability (R^2)	• Pairwise congruence (TSS) • Overall congruence	• Spatial overlap between hotspots and protected areas

Fig. 2. An overview of the data and methods used in each of the four phases of the modelling framework employed here, from building the boosted regression trees (BRT) to overlaying predictions with existing protected areas, and their results.

Practical – modelling species richness in Northwestern Lapland with GBM

TSR = Total species richness

Results from explanatory GBM (BRT)
> gbm()

Results from GBM spatial predictions →

We will be looking at spatial predictions next week!

A.K.J. Niskanen et al.   Biological Conservation 212 (2017) 288–299

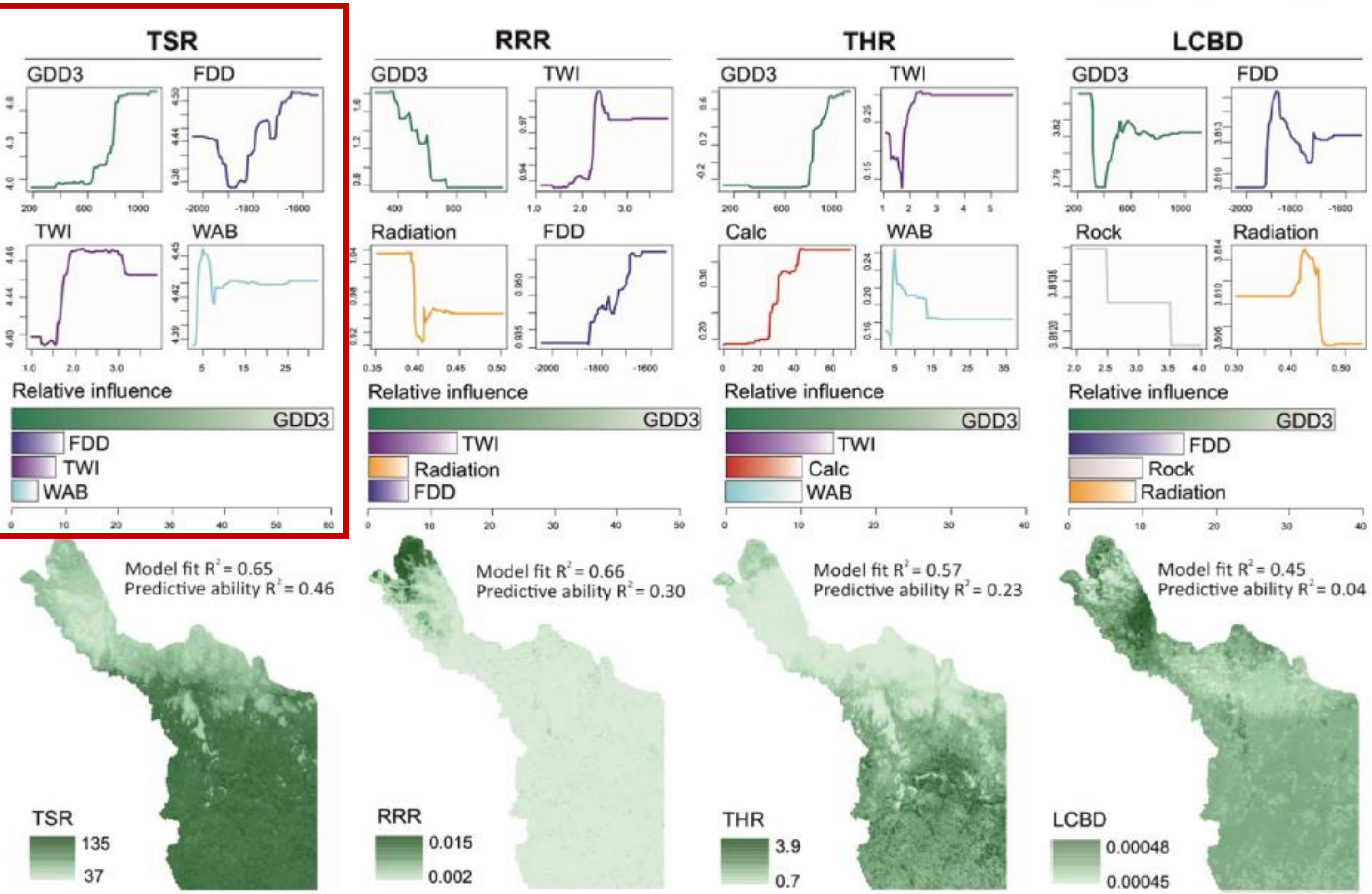
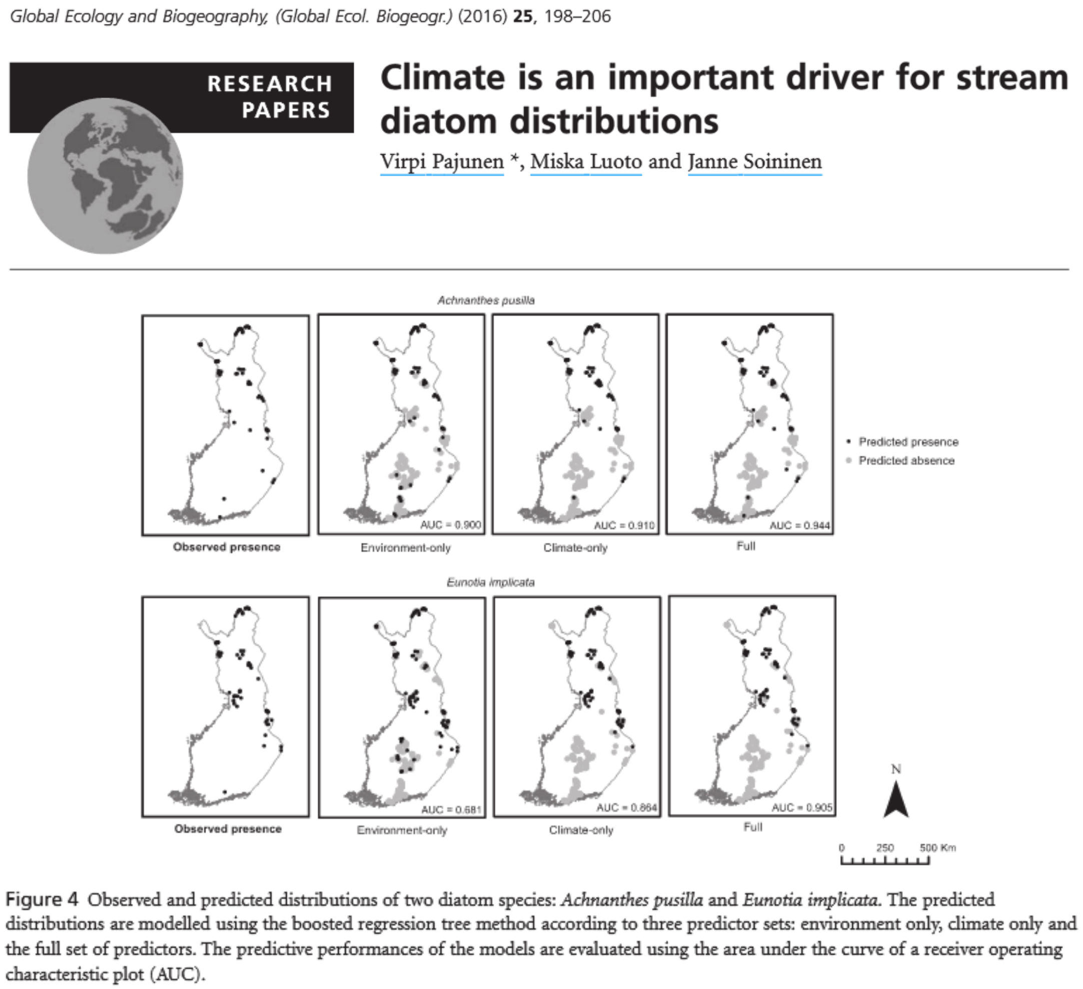


Fig. 3. Boosted regression tree (BRT) based results and predicted diversity maps for total species richness (TSR), range-rarity richness (RRR), threatened species richness (THR), and local contribution to β-diversity (LCBD). For each metric, a five-panel plot shows the BRT results: the four topmost panels are partial dependency plots showing the BRT-modelled responses of the metric to its most influential drivers according to the model with the highest explanatory power (R^2); the Relative influence panel shows the mean relative contributions (%) of the four most influential variables in predicting richness values. The map below the plot panels shows mean model predictions for each metric separately. See Table 2 for variable descriptions and abbreviations.

4. Other examples from the literature



Global Ecology and Biogeography, (Global Ecol. Biogeogr.) (2016) **25**, 198–206

RESEARCH PAPERS

Climate is an important driver for stream diatom distributions

Virpi Pajunen *, Miska Luoto and Janne Soininen

Figure 4 Observed and predicted distributions of two diatom species: *Achnanthes pusilla* and *Eunotia implicata*. The predicted distributions are modelled using the boosted regression tree method according to three predictor sets: environment only, climate only and the full set of predictors. The predictive performances of the models are evaluated using the area under the curve of a receiver operating characteristic plot (AUC).

Journal of Animal Ecology

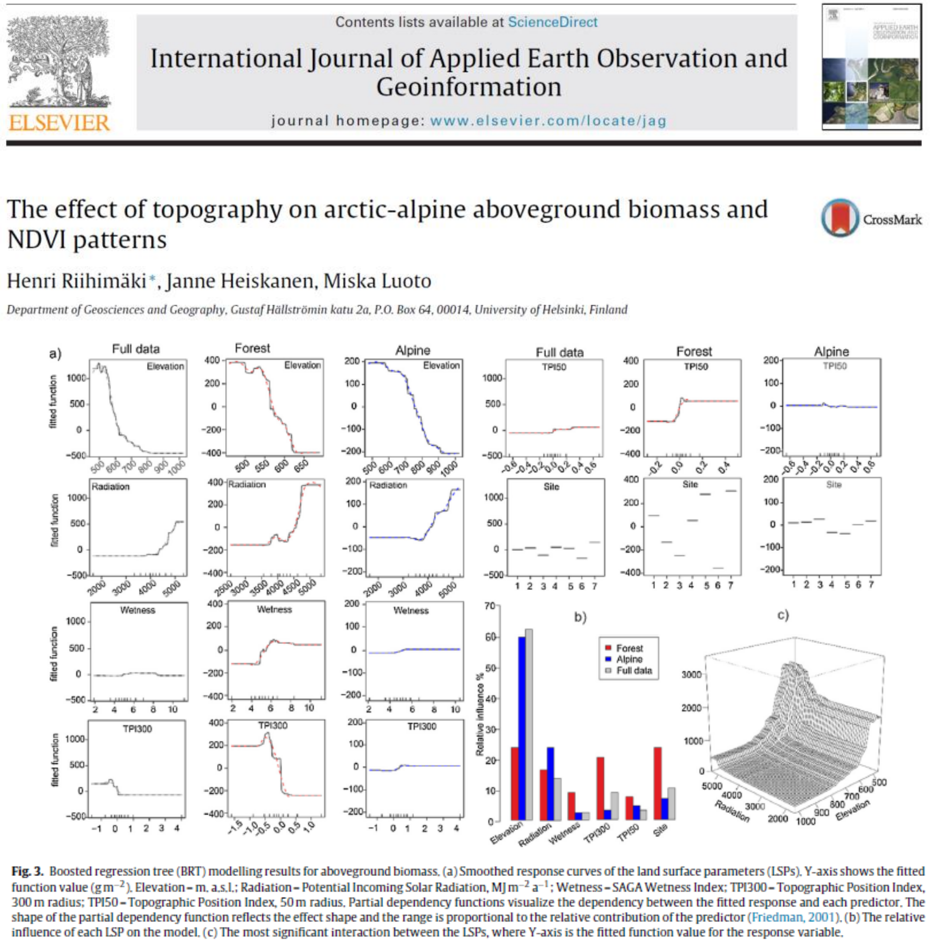
British Ecological Society

Journal of Animal Ecology 2008, **77**, 802–813 doi: 10.1111/j.1365-2656.2008.01390.x

A working guide to boosted regression trees

J. Elith[1]*, J. R. Leathwick[2] and T. Hastie[3]

[1]*School of Botany, The University of Melbourne, Parkville, Victoria, Australia 3010;* [2]*National Institute of Water and Atmospheric Research, PO Box 11115, Hamilton, New Zealand; and* [3]*Department of Statistics, Stanford University, CA, USA*



Contents lists available at ScienceDirect

ELSEVIER

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag

The effect of topography on arctic-alpine aboveground biomass and NDVI patterns

CrossMark

Henri Riihimäki *, Janne Heiskanen, Miska Luoto

Department of Geosciences and Geography, Gustaf Hällströmin katu 2a, P.O. Box 64, 00014, University of Helsinki, Finland

Fig. 3. Boosted regression tree (BRT) modelling results for aboveground biomass. (a) Smoothed response curves of the land surface parameters (LSPs). Y-axis shows the fitted function value (g m^{-2}). Elevation – m. a.s.l.; Radiation – Potential Incoming Solar Radiation, MJ m^{-2} a^{-1}; Wetness – SAGA Wetness Index; TPI300 – Topographic Position Index, 300 m radius; TPI50 – Topographic Position Index, 50 m radius. Partial dependency functions visualize the dependency between the fitted response and each predictor. The shape of the partial dependency function reflects the effect shape and the range is proportional to the relative contribution of the predictor (Friedman, 2001). (b) The relative influence of each LSP on the model. (c) The most significant interaction between the LSPs, where Y-axis is the fitted function value for the response variable.

Thursday's exercise:

Questions: ”Exercise_16.11.2017” in Moodle
R-script: Write your own :)
Data: ”*saana.csv*”; vascular plant species and abiotic condtions were surveyed on the Saana massif in north-western Finland (69°3′N 20°51′E) at ca. 700 m a.s.l., ca. 100–200 m above the birch treeline. Grid data of 1m x 1m quadrats. Within each quadrat:

mesotopo = mesotopography
soil_moist = soil moisture
soil_temp = soil temperature
soil_ph = soil pH
veg_height = vegetation height
vasc_spr = species richness of vascular plants
Empher_cover = cover of *Empetrum hermpaphroditum* (crowberry; dominant species)
Betnan = presence/absence of *Betula nana* (dwarf birch)
Cashyp = presence/absence of *Cassiope hypdnoides* (moss heather)
Empher = presence/absence of *Empetrum hermpaphroditum*
Gersyl = presence/absence of *Geranium sylvaticum* (woodland geranium)
Salret = presence/absence of *Salix reticulata* (snow willow)

Exercise 16.11.2017

1. Divide the saana.csv data randomly into two different datasets: model calibration data (70%) and model evaluation data (30%). Build the models based on calibration data and test the models using evaluation data. What is the predictive performance of the GLM, GAM and GBM models for *Betnan*, *Cashyp*, *Empher* and *Salret* based on AUC-values of the model evaluation data? Use mesotopo, soil_moist, soil_temp and soil_ph as predictors. Report the results in one short paragraph (max 5 sentences). Note, you can use *sample*-function to divide the data, e.g.

```
ncal <- 0.7
s <- sample(nrow(data1), ncal)
data_cal <- data1[s,]
data_eval <- data1[-s,]
```

2. What is the predictive performance of the GLM, GAM and GBM models for veg_height and vasc_spr? Build the models using calibration data and test the models using evaluation data. Use Spearman correlation -values as the evaluation metrics. Use the same set of predictors that you used in 1). Report the results in one short paragraph (max 5 sentences).

3. Characterize soil_moist, soil_temp, soil_ph, veg_height and vasc_spr conditions along the mesotopographic gradient using GAM. Model the values of these five responses at the valley bottom (mesotopo 1), mid-slope (mesotopo 5) and ridge-top (mesotopo 10). Present the results as an informative figure. Report the results in one short paragraph (max 5 sentences).

4. Does the cover of Empher (Empher_cover) has an effect on the vasc_spr when all other predictors are controlled for? Use the same set of predictors as used in 1). Use all three modelling frameworks to test the hypothesis. Report the results in one short paragraph (max 5 sentences). The main idea in this question: as a dominant species Empher_cover might have a strong influence on the vegetation properties – can we see the effect? Please, test it!

Deadline 21.11.2017