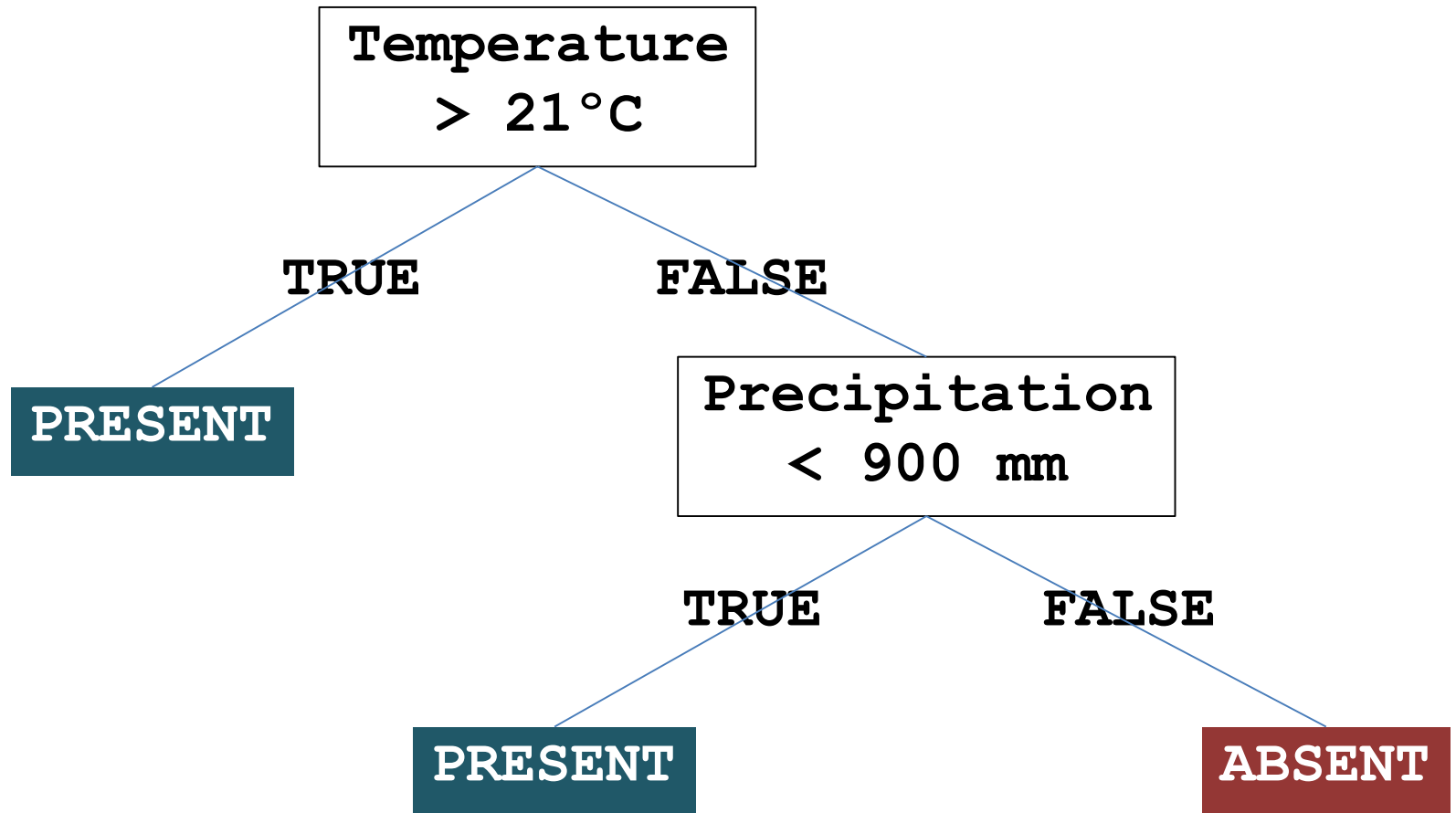


BART variable selection

A. Márcia Barbosa

<https://modtools.wordpress.com/barbosa>

Decision trees



BART: Bayesian regression approach
to a cumulative sum of simple trees

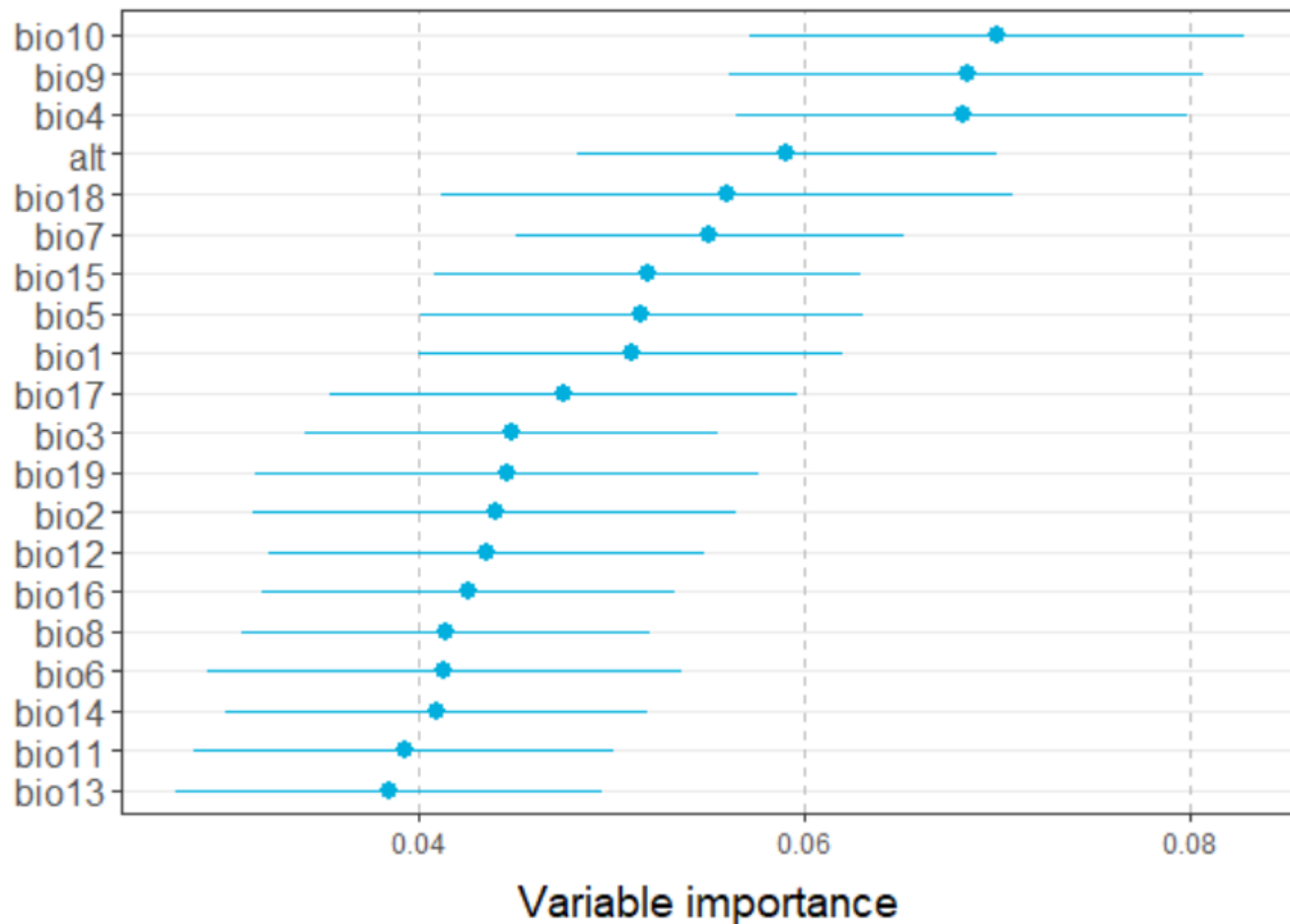
Variable importance

- each variable can be used more than once in the same tree, and much more than once within many trees
- variable importance can be measured as the number of times each variable is used in a branch split
- in models with many trees, differences in importance becomes less pronounced, as less informative variables still get in a high number of splits

Variable importance

get a plot of variable importance:

```
varimp(mod_BART, plots = TRUE)
```

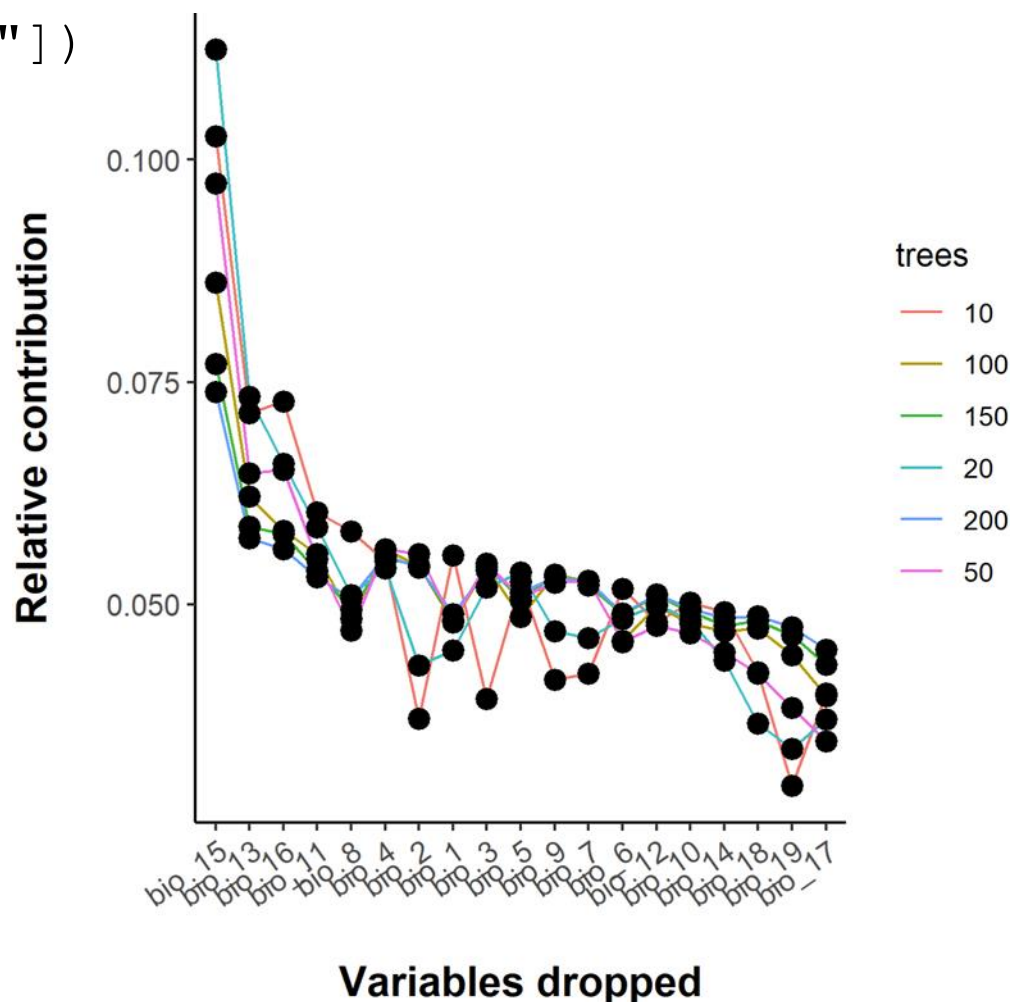


Variable importance

get a **diagnostic** plot of variable importance:

```
varimp.diag(x.data = dat[, var_cols],  
             y.data = dat[, "presence"])
```

- subjective, informal
- visualization purposes
- computationally intensive!



Variable selection: reducing collinearity

- BART models don't seem affected by using a large number of predictor variables
- correlations between variables also don't affect predictive power (within similar correlation structures)
- but correlations make it difficult for any model to distinguish the effects of different variables
- so, variable selection may be more unstable with highly correlated variables

Variable selection: reducing collinearity

```
library(collinear)
```

```
var_sel <- collinear(df = dat, response = spc_col,  
predictors = var_cols, max_cor = 0.75, max_vif = 5)
```

```
[1] "bio_12" "bio_15" "bio_18" "bio_19" "bio_3"  "bio_8"  
[7] "bio_2"  "bio_9"
```

see ?collinear help file for more info!

Variable selection with {embarcadero}

1. Fit a model with all variables and a small ensemble of trees (default 10) a number of times (default 50)
2. Remove the least informative variable across the 50
3. Rerun model (again, 50 times) recording RMSE
4. Repeat two previous steps until 3 variables remain
5. Select the model (of 3 or more variables) with the smallest average RMSE

this procedure is implemented automatically

Variable selection with {embarcadero}

```
# select minimal subset of relevant variables:
```

```
set.seed(123)
```

```
varsel_emb <- variable.step(x.data = dat[, var_sel],  
                             y.data = dat[, "presence"])
```

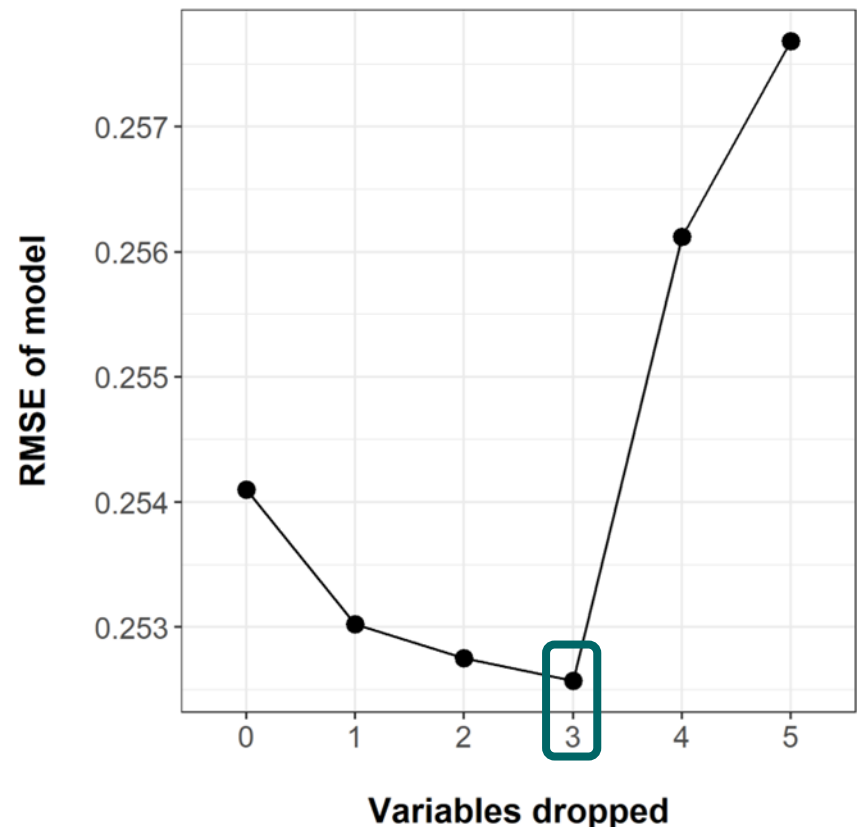
*the model excluding the
first 3 removed variables
had the smallest error*



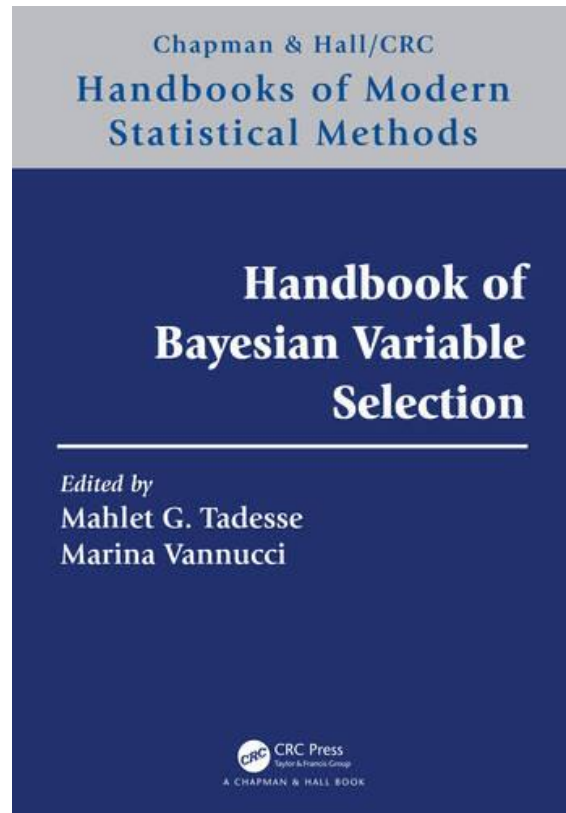
```
varsel_emb  
# "bio_12" "bio_15" "bio_18"  
"bio_8"   "bio_9"
```



*so, 5 variables selected out of
the 8 candidates (8 - 3)*



Variable selection with `{nonlinvarsel}`



Chapter 17:

Carvalho C.M., George E.I., Hahn PR & **McCulloch R.E.**
(2021) Variable Selection and Interaction Detection with
Bayesian Additive Regression Trees.

<https://www.routledge.com/Handbook-of-Bayesian-Variable-Selection/Tadesse-Vannucci/p/book/9780367543761>

Variable selection with {nonlinvarsel}

1. Compute models using the prediction of the full model as the response variable, each time using a different subset of the predictor variables
2. Identify minimal subset that provides a more than satisfactory fit (e.g. $R^2 > 95\%$)
3. Forward, backward and all-subsets selection are implemented: selected variables may vary (same as with stepwise regression)

Variable selection with {nonlinvarsel}

```
# forward selection:
```

```
varsel_fwd <- vsf(dat[complete_rows, var_cols],  
dat[complete_rows, "BART_P"])
```

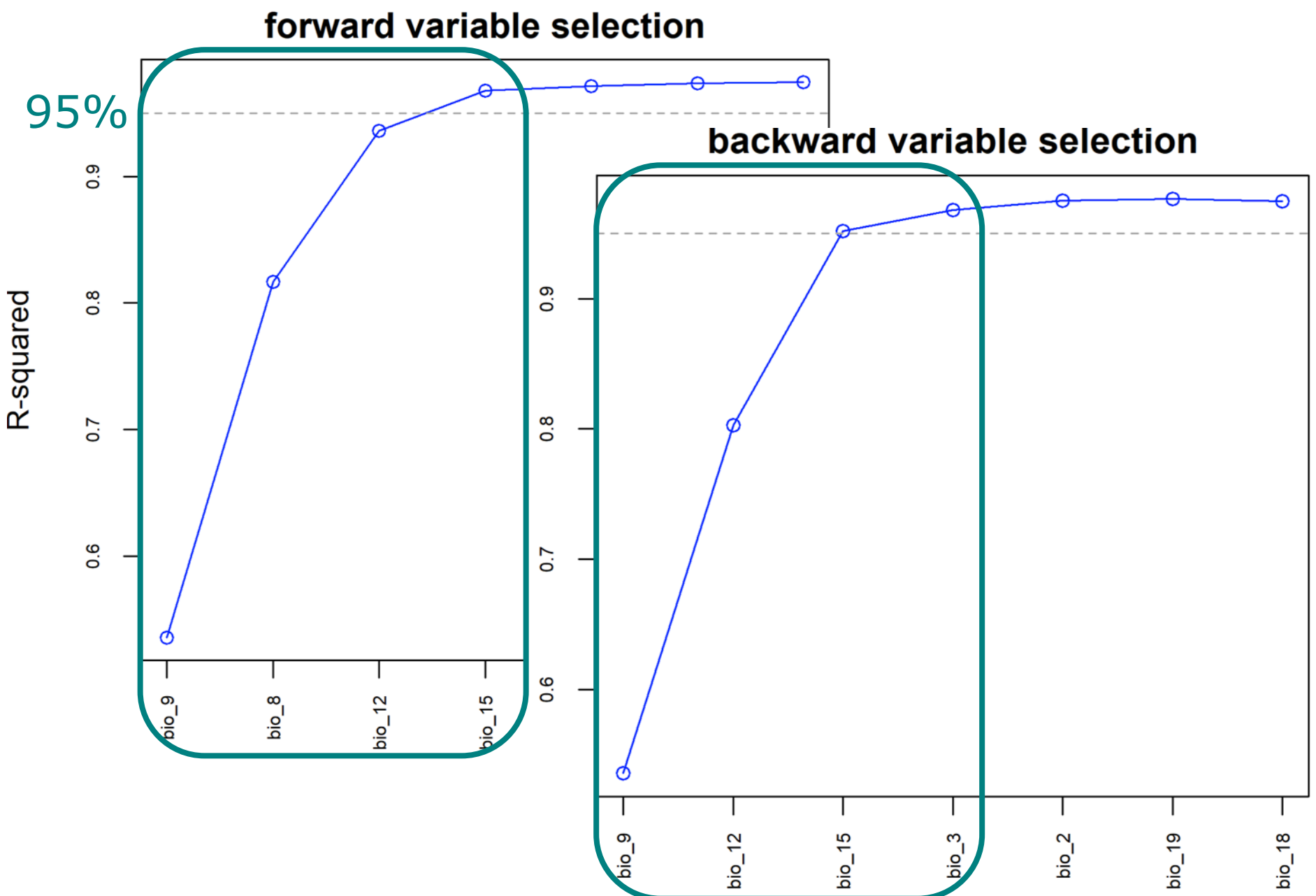
```
# backward selection:
```

```
varsel_bwd <- vsb(dat[complete_rows, var_cols],  
dat[complete_rows, "BART_P"])
```

```
# all-subsets selection:
```

```
varsel_nonlin_all <- vsa(dat[complete_rows, var_cols],  
dat[complete_rows, "BART_P"]) # very computationally  
intensive!
```

Variable selection with {nonlinvarsel}



Models with selected variables

build a model with only the selected variables:

```
mod <- bart(y.train = dat[, spc_col],  
            x.train = dat[, varsel_fwd],  
            keeptrees = TRUE)
```



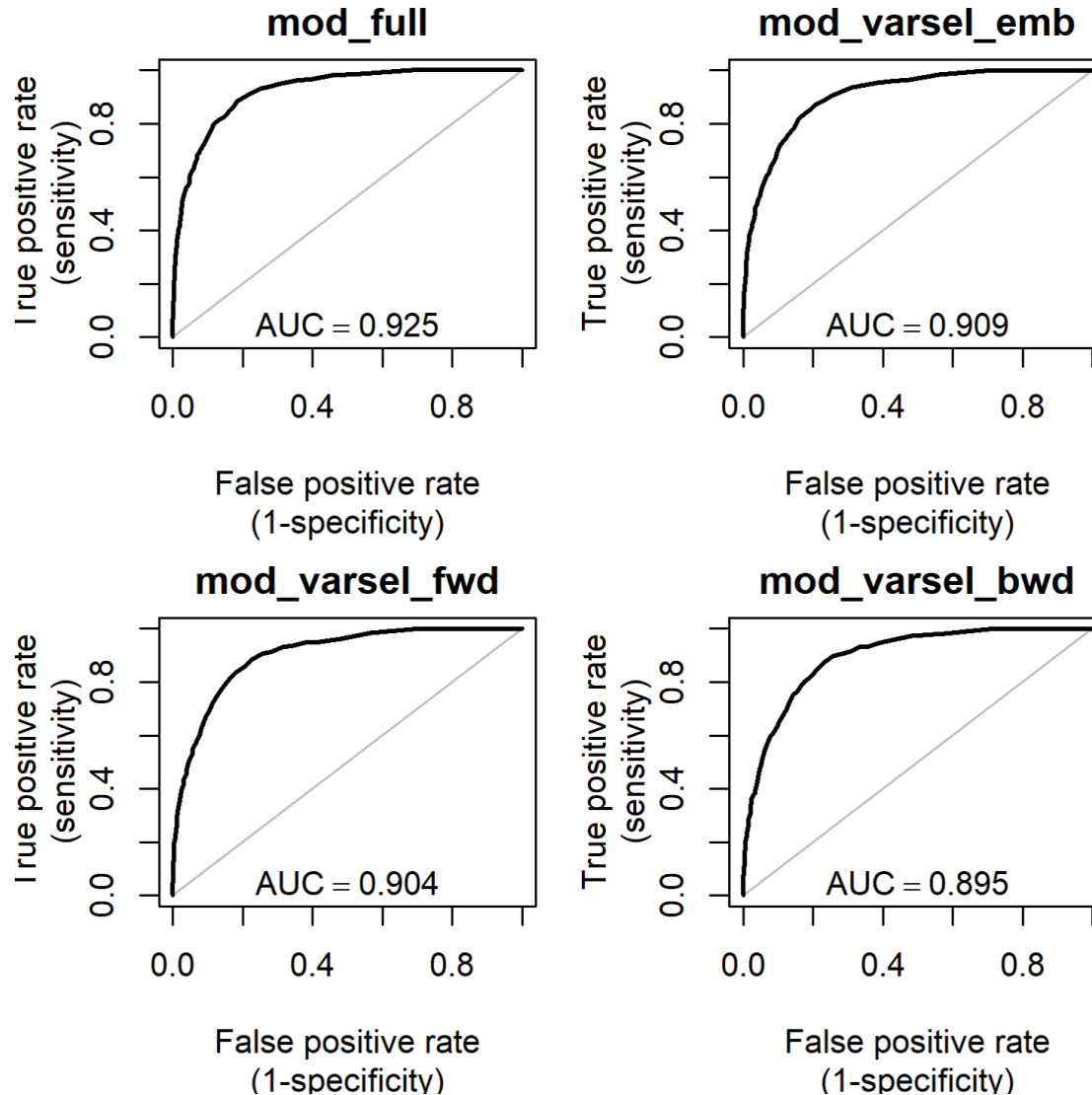
this is necessary if you
want to get predictions and
partial dependence plots
from this model afterwards



variables
selected before:
then do same for
varsel_bwd
varsel_emb, etc.

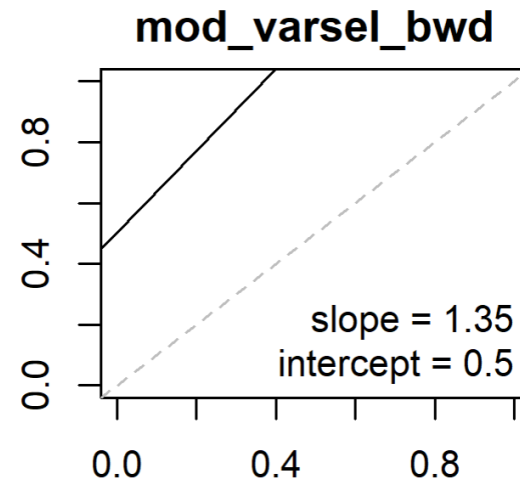
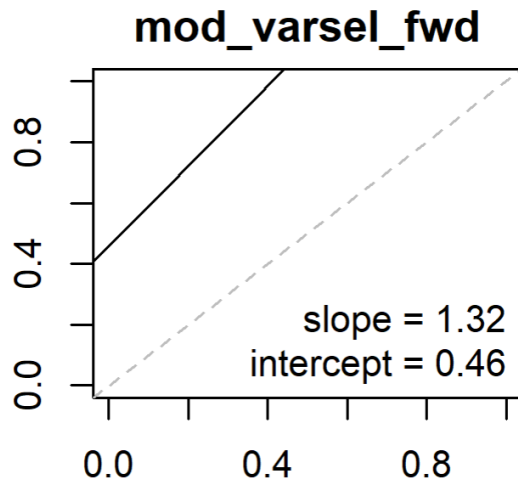
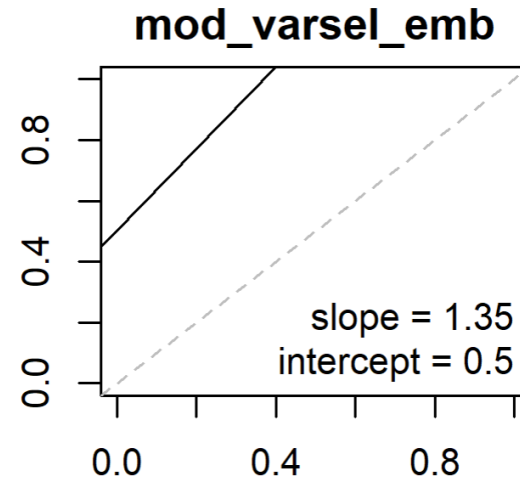
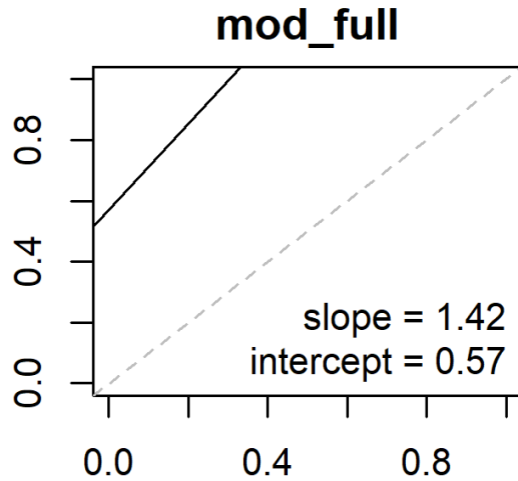
Compare model predictions

discrimination: area under the ROC curve

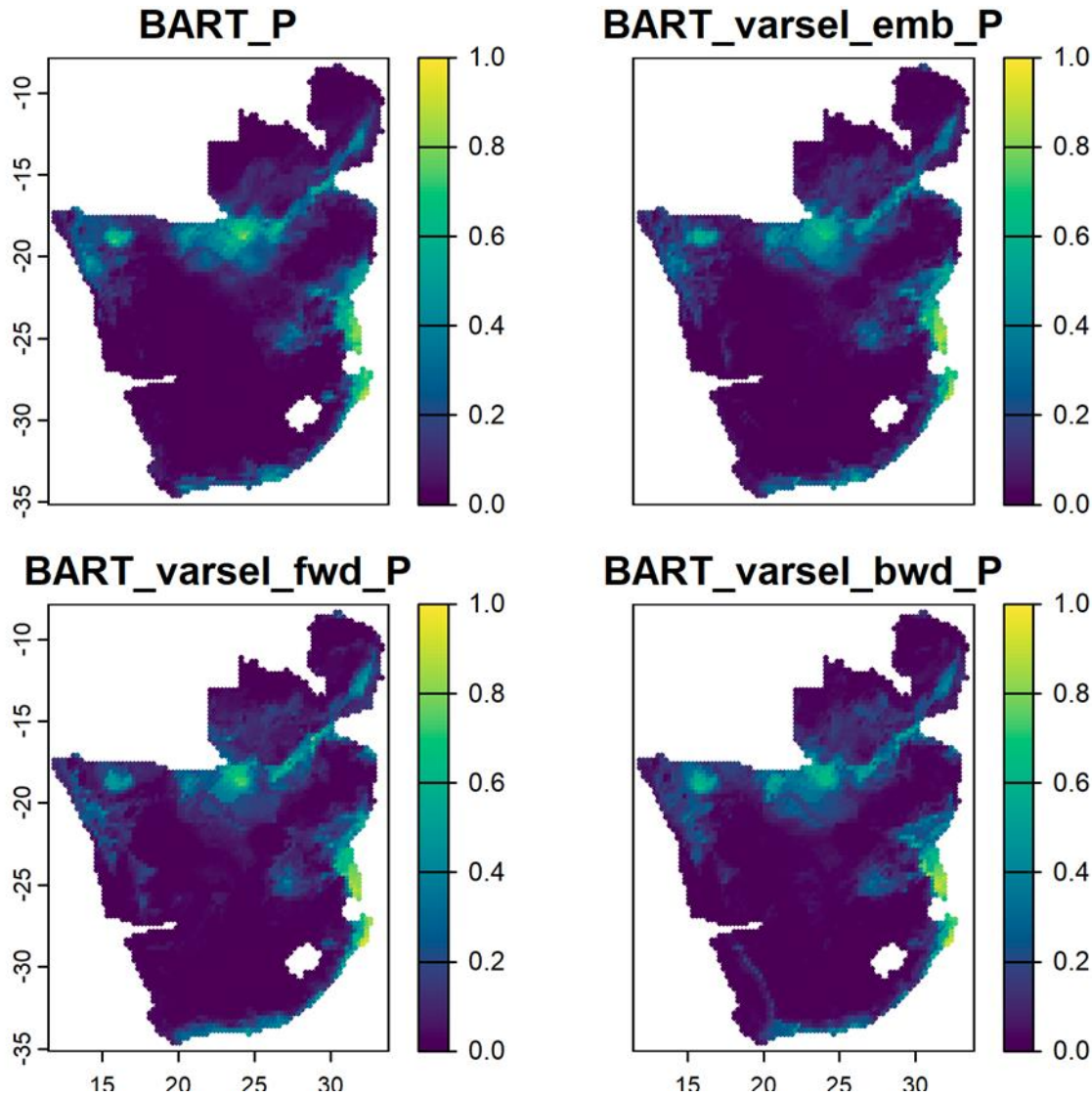


Compare model predictions

calibration: Miller line



Compare model predictions



correlations:

	BART_P
BART_emb_P	0.967
BART_fwd_P	0.917
BART_bwd_P	0.896