

Computing BART models

A. Márcia Barbosa

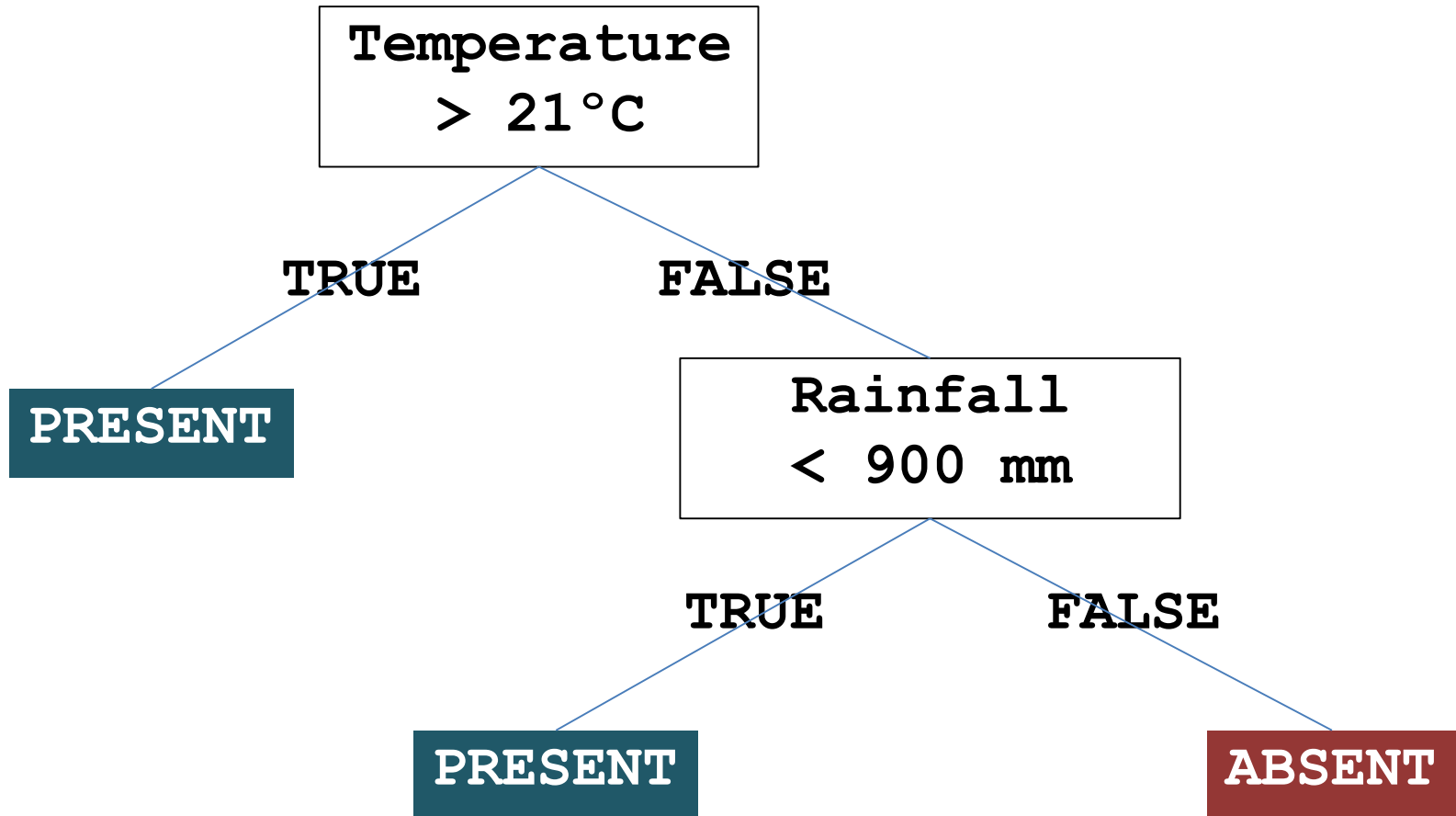
<https://modtools.wordpress.com/barbosa>

Species distribution models

- regression-based (e.g. GLM, GAM)
- machine learning (e.g. neural networks, boosted trees)

some methods combine the two approaches

Decision trees



Decision trees

- Occam's razor: simpler is better (avoid over-fitting)
- build small trees that accurately fit the data
- stopping criterion, pruning

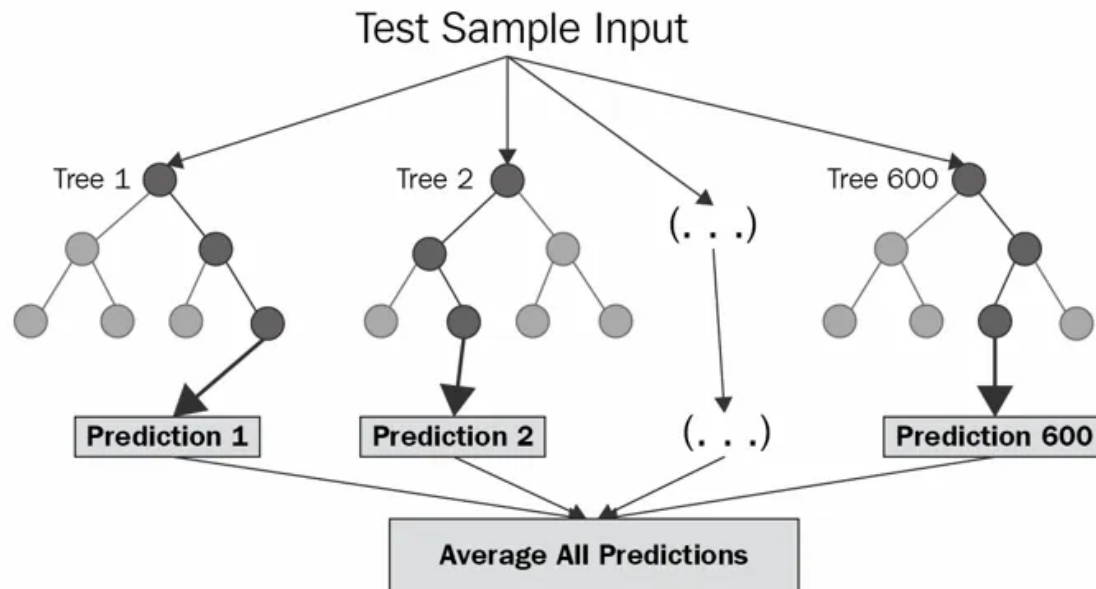
but...

single trees have poor predictive capacity!

Decision tree ensembles

RF (Random Forests):

- ensemble of trees computed independently from random subsets of the data
- average prediction of all trees

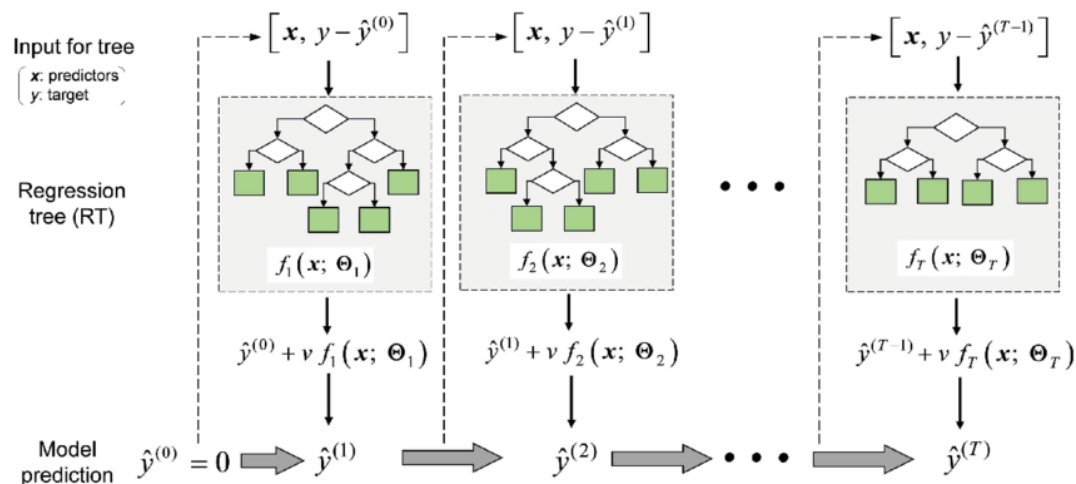


Decision tree ensembles

BRT (**B**oosted **R**egression **T**rees), a.k.a.

GBM (**G**eneralized **B**oosted **M**odels):

- sequence of simple trees (weak learners) for previous prediction residuals
- sum of trees x learning rate



Decision tree ensembles

BART (**B**ayesian **A**dditive **R**egression **T**rees)

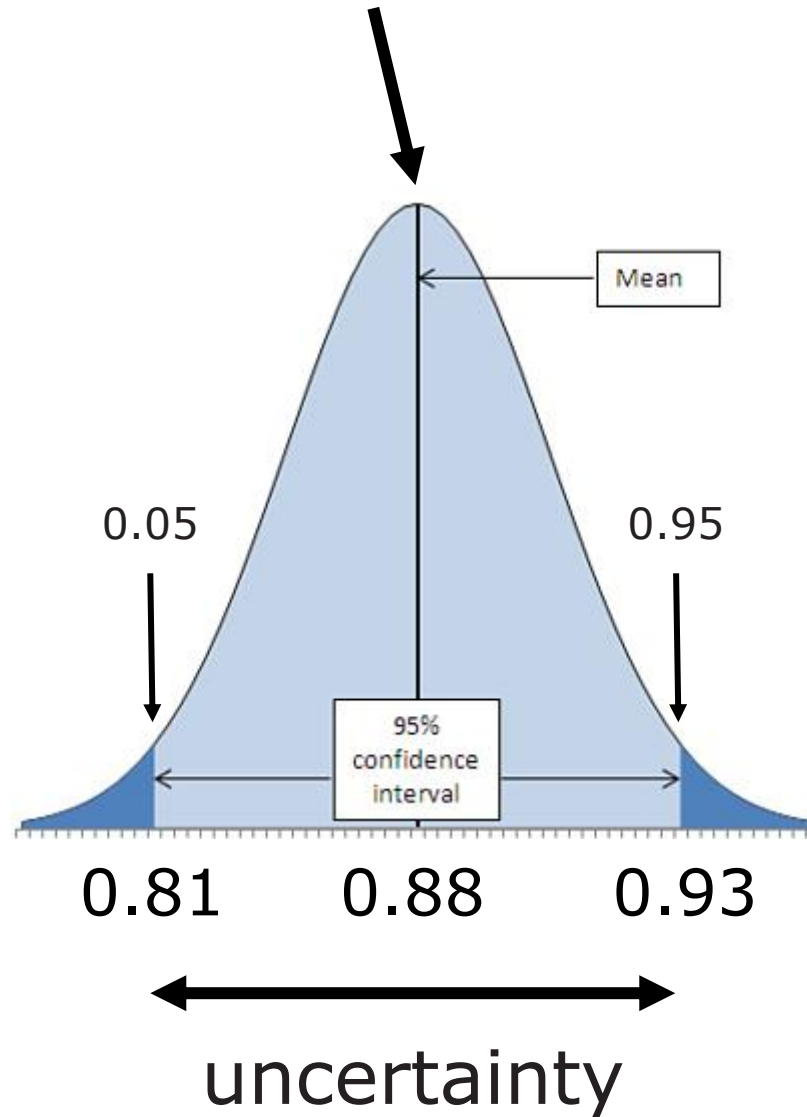
- non-parametric Bayesian regression approach to an ensemble of trees
- cutting-edge alternative to other tree-based methods (RF, BRT/GBM)
- built-in complexity penalty
- inherently handles prediction uncertainty

BART advantages

- it's Bayesian: for each site, the **baseline expected value (prior)** is not a single estimate (e.g. prevalence), but a **probability distribution** representing all possible estimates with associated levels of uncertainty
- the estimate is then updated to get a predicted (posterior) distribution based on the new evidence (presence or absence of species observation records)

BART advantages

posterior mean probability



BART advantages

- with the built-in complexity penalty, **default parameters** are sensible and generally **work very well**
- anyway, there's also an **automated cross-validation** approach to prior selection
- advanced users can go more in depth within the BART literature to set better priors, though that's usually not recommended

BART advantages

- BART provides an "out-of-the-box" procedure that is statistically robust, provides excellent performance, and is also robust to changes in data or parameter choices
- it's an example of how the **probabilistic framework** of Bayesian statistics can avoid many of the problems with classical approaches

BART:

[2010]

Bayesian Additive Regression Trees

Hugh A. Chipman, Edward I. George, Robert E. McCulloch *

APPLICATION

Methods in Ecology and Evolution  BRITISH
ECOLOGICAL
SOCIETY

embarcadero: Species distribution modelling with Bayesian additive regression trees in R

Colin J. Carlson^{1,2} 

[2020]

Papers using BART

Associations between habitat quality and body size in the Carpathian land snail *Vestia turgida*: species distribution model selection and assessment of performance

V. M. Tytar

Institute of Zoology, NASU

<https://orcid.org/0000-0002-0864-2548>

O. Baidashnikov

National Museum of Natural History, NASU

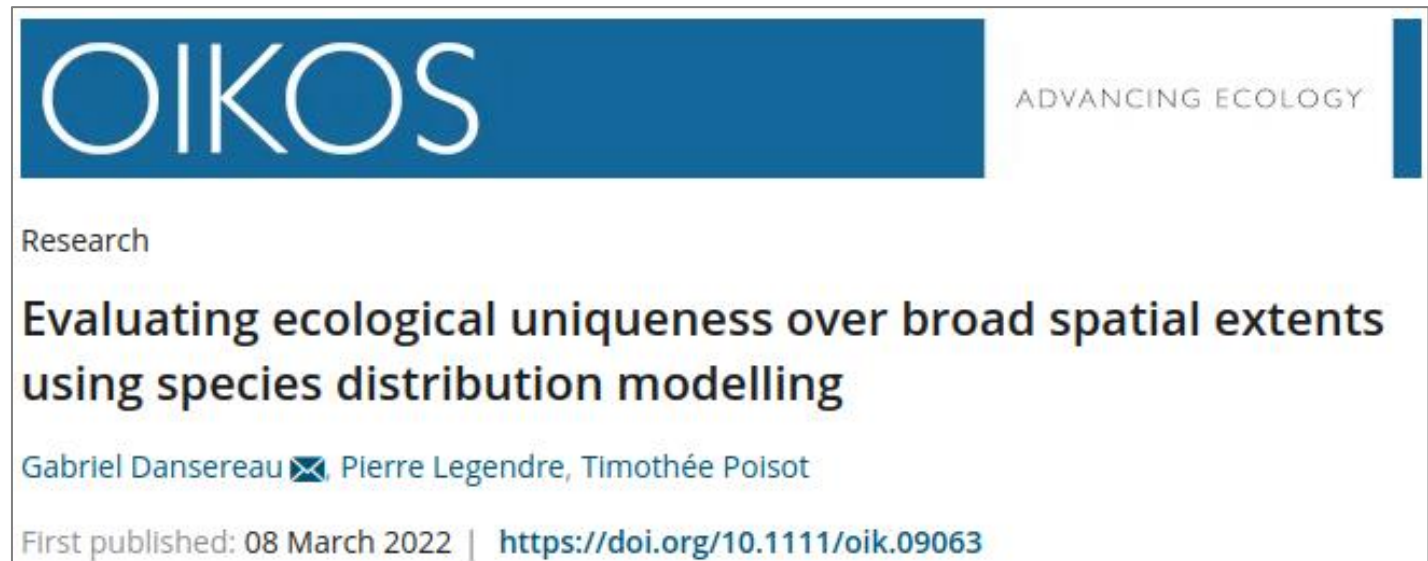


Plague risk in the western United States over seven decades of environmental change

 Colin J. Carlson,  Sarah N. Bevins,  Boris V. Schmid

doi: <https://doi.org/10.1101/2021.02.26.433096>

Papers using BART



more examples at the ['embarcadero' paper](#)
Citing literature section

Papers comparing BART to others

Diversity and Distributions

Open Access

A Journal of
Conservation
Biogeography

RESEARCH ARTICLE |  Open Access |  

Potential distributions of invasive vertebrates in the Iberian Peninsula under projected changes in climate extreme events

Rocío A. Baquero , A. Márcia Barbosa, Daniel Ayllón, Carlos Guerra, Enrique Sánchez, Miguel B. Araújo, Graciela G. Nicola

Results

Selected models showed high consensus and good predictive capacity on block cross-validation areas. Generalized Linear Models and Generalized Additive Models scored highest in reliability (calibration), but Bayesian Additive Regression Trees provided the best balance between calibration and discrimination capacity. Forecasts include visible changes in environmental favourability, with losses generally outweighing the gains, but with some areas becoming more favourable for several species.

Papers comparing BART to others



Ecological Informatics

Volume 61, March 2021, 101198



Statistical comparison of additive regression tree methods on ecological grassland data

Emily Plant^a , Rachel King^a, Jarrod Kath^b

Highlights

- BART models explain grassland variability under drought and wet conditions well.
- BART explained more variation than BRT, linear regression and random forest models.
- BART should be used in conjunction with other ecological modelling approaches.

Papers recommending single methods

ECOGRAPHY

Open Access

A JOURNAL OF SPACE
AND TIME IN ECOLOGY

Research



Open Access



Testing whether ensemble modelling is advantageous for maximising predictive performance of species distribution models

Tianxiao Hao ✉, Jane Elith, José J. Lahoz-Monfort, Gurutzeta Guillera-Arroita

performing better than ensembles. This study shows **no particular benefit to using ensembles over individual tuned models**. It also suggests that further robust testing of performance is required for situations where models are used to predict to distant places or environments.

...and with BART you don't even need the tuning!

BART software



packages:

```
[BART, BayesTree, bartMachine]
```

- **dbarts**
- **embarcadero** (wrapper for **dbarts**
in a SDM context)

Set up the data

example of a data frame that can be used for computing a model:

presence	x	y	cells	bio_1	bio_2	bio_3	bio_4
0	26.125	-25.125	5954	19.64711	16.02806	55.41713	482.3851
0	26.375	-25.125	5955	20.27168	16.56423	55.74219	487.0897
0	26.625	-25.125	5956	20.01786	16.05557	55.92709	470.6123
1	26.875	-25.125	5957	19.65874	15.92198	56.15314	465.0204
1	27.125	-25.125	5958	19.52998	15.71729	55.63602	465.7927
0	27.375	-25.125	5959	19.54671	15.34325	54.76509	470.6145
0	27.625	-25.125	5960	19.77601	15.25366	54.40098	475.5373

*species occurrence data:
1 = presence,
0 = no recorded presence*

*values of the
predictor variables*

Compute BART model

```
library(embarcadero)

mod <- bart(x.train = dat[, var_cols],
             y.train = dat[, "presence"],
             keeptrees = TRUE, seed = 123)
```

this makes the resulting object heavier,
but it is necessary if you want to get
predictions from this model afterwards

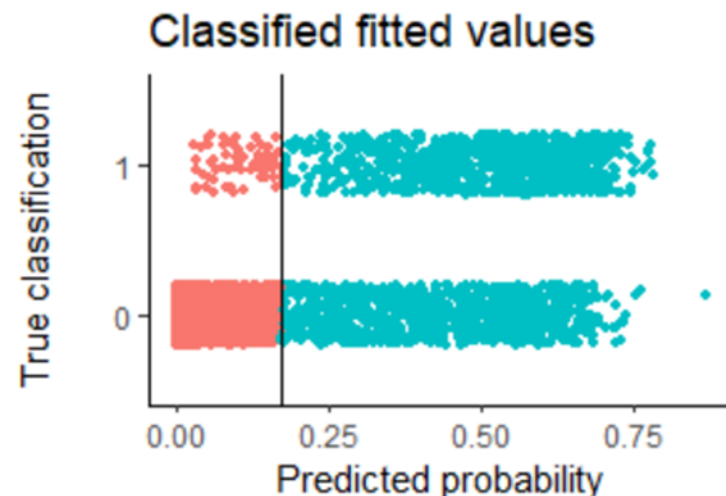
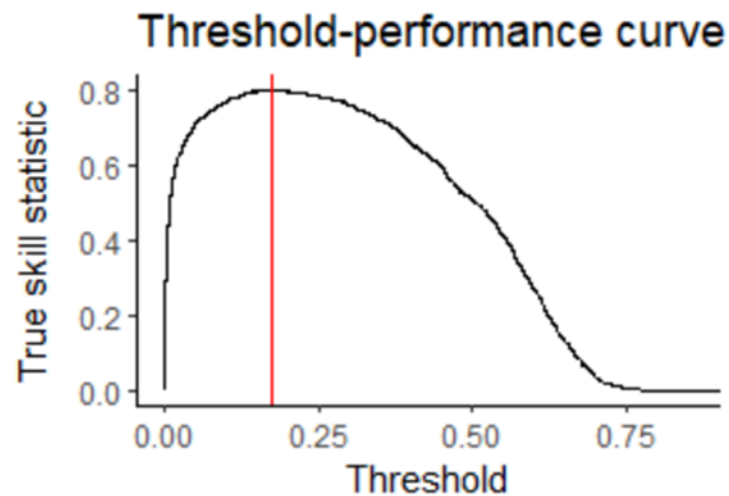
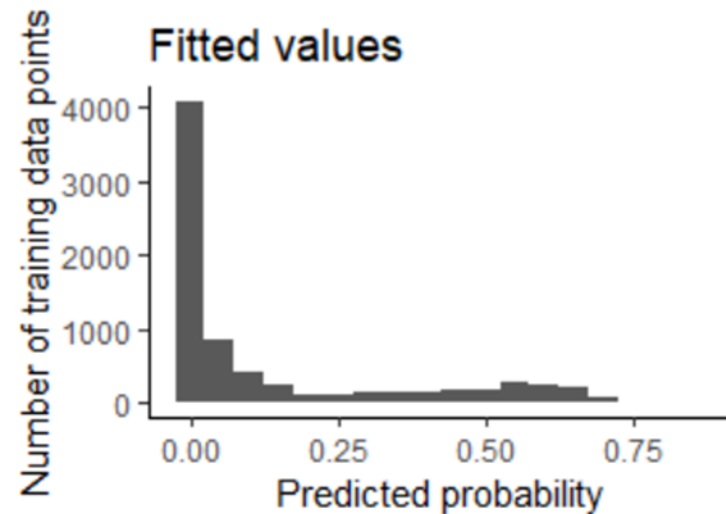
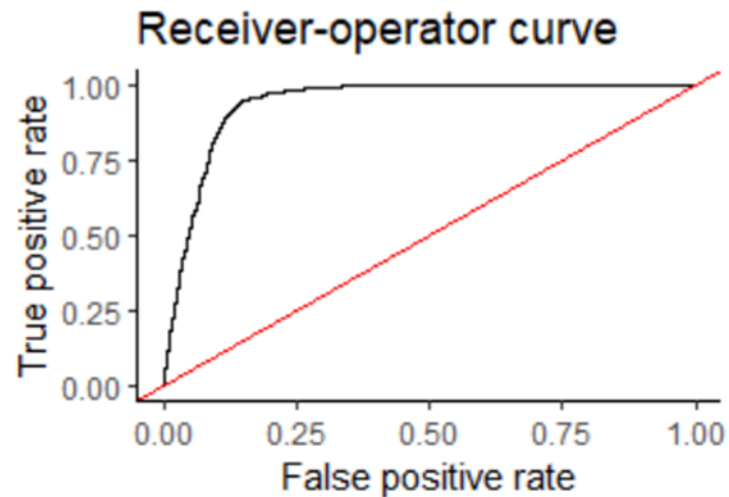


this makes the model
reproducible, even if it uses
random number generation



Examine model

summary (mod)



Fine-tune prior parameters

```
mod <- retune(mod)
```

```
summary(mod)
```

- `embarcadero::retune` is a wrapper for `dbarts::xbart` – a cross-validation approach to the selection of parameters for the priors
- **computer-intensive** and usually not necessary (as default parameters perform well), but may slightly improve predictive capacity

Map model predictions

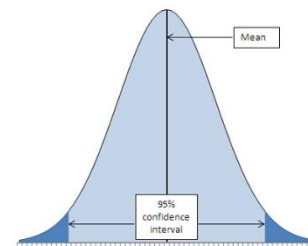
```
BART_P <- predict(mod, layers,
quantiles = c(0.025, 0.975))
```

← for raster maps

```
bart_pred <- predict_bart_df(mod, dat,
quantiles = c(0.025, 0.975))
```

← for data frames

pred	q0025	q0975	uncertainty
0.14932001	0.04895702	0.2978955	0.2489384
0.38387245	0.18238137	0.5964788	0.4140975
0.10360921	0.02981563	0.2106086	0.1807930
0.11268104	0.03936088	0.2172980	0.1779371
0.09813812	0.03002463	0.1836438	0.1536191
0.10752773	0.03991112	0.2091774	0.1692663



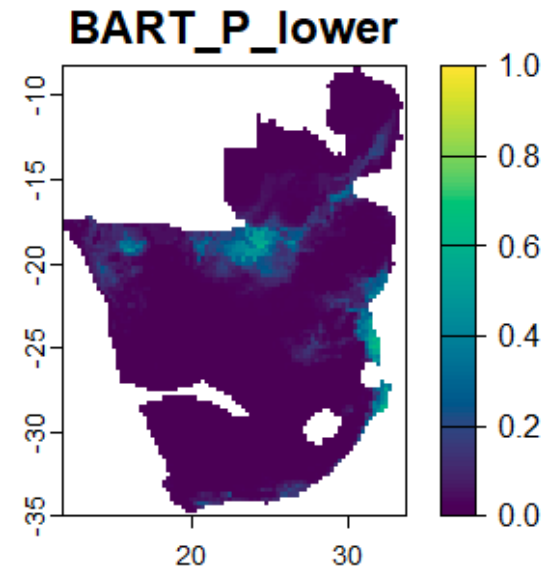
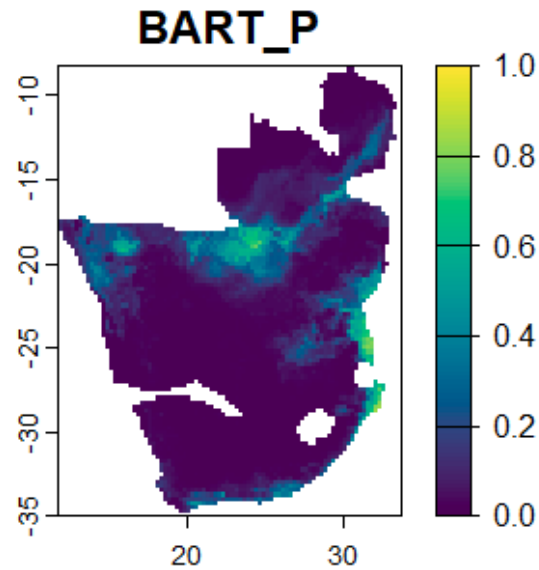
```
# add uncertainty (credible interval width):
```

```
bart_pred$uncertainty <- bart_pred[ , 3] -  
                           bart_pred[ , 2]
```

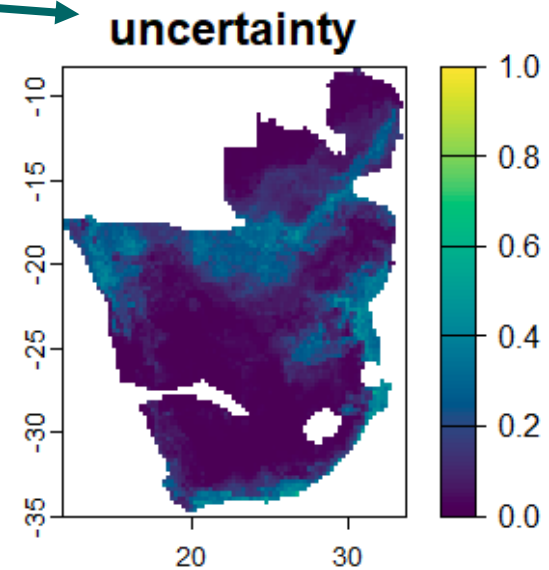
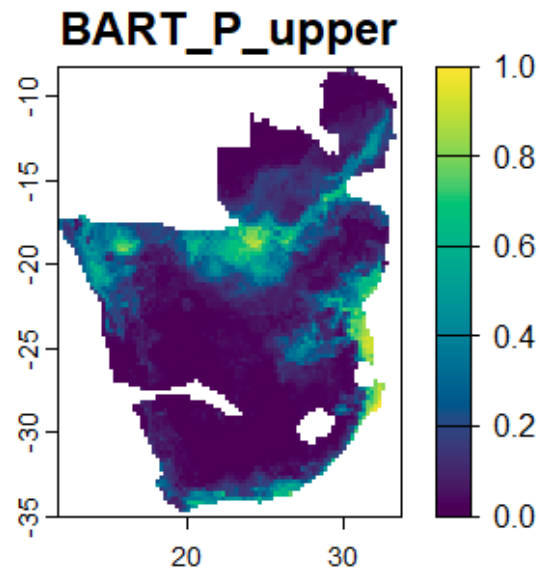
Map model predictions

`plot(BART_P)`

"posterior
mean"



width of the
credible interval



Probability vs. favourability

Presence probability (given by BART and other presence-absence models): likelihood of finding the species at a given environment, which depends on:

- the general likelihood of finding the species in the region, given how rare or common it is (**prevalence**)
- the **favourability** of that particular environment to the species, which may make it more likely to be there than in general across the region

Probability vs. favourability

Favourability can be obtained directly from probability by removing the effect of species prevalence, e.g. with the `fuzzySim::Fav()` function.

Favourability does **not necessarily** reflect areas where the species is **likely** to be present, but rather areas that are **favourable** to the species, even if it's absent.

For more details, see e.g.

<https://modtools.wordpress.com/2023/07/11/weighted-probability-vs-favourability>

Probability vs. favourability

```
BART_F <- fuzzySim::Fav(pred = BART_P,  
sample.preval = prevalence(model = mod_full))
```

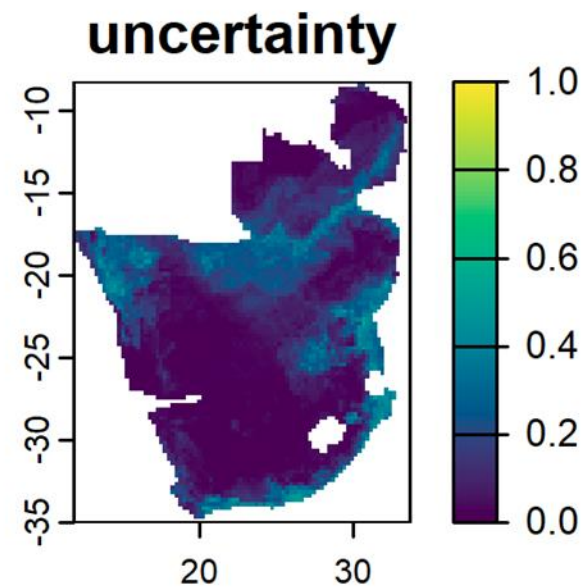
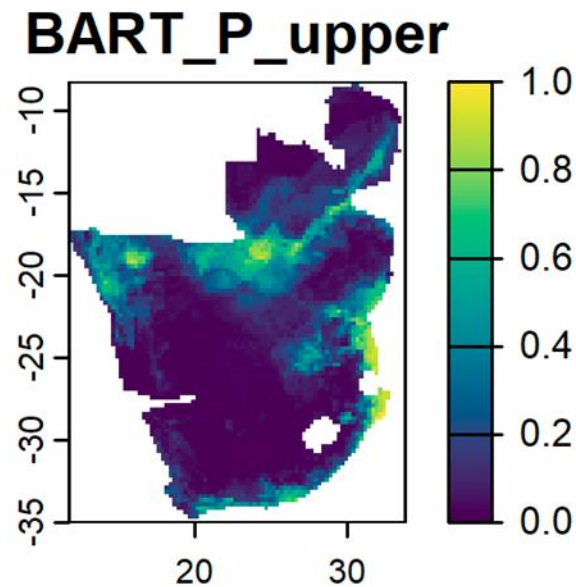
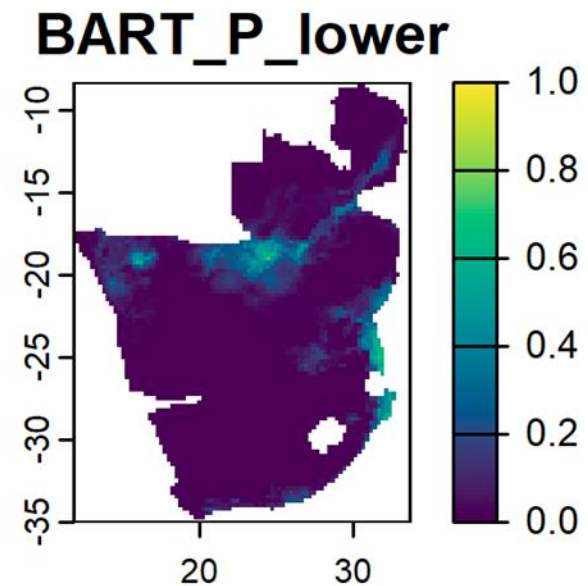
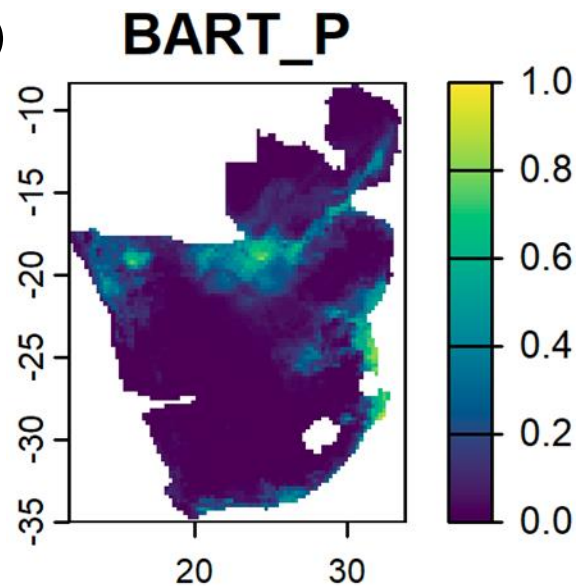
```
BART_F_lower <- Fav(pred = BART_P_lower,  
sample.preval = prevalence(model = mod_full))
```

```
BART_F_upper <- Fav(pred = BART_P_upper,  
sample.preval = prevalence(model = mod_full))
```

```
F_uncert <- BART_F_upper - BART_F_lower
```

Probability vs. favourability

```
plot(BART_P)
```



Probability vs. favourability

`plot(BART_F)`

