

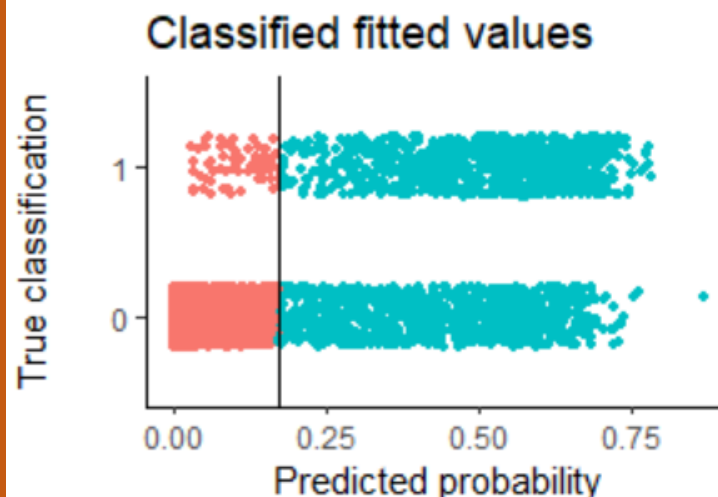
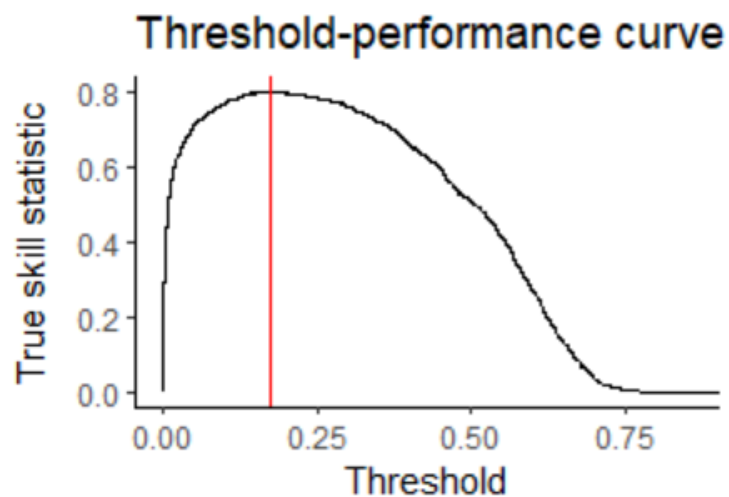
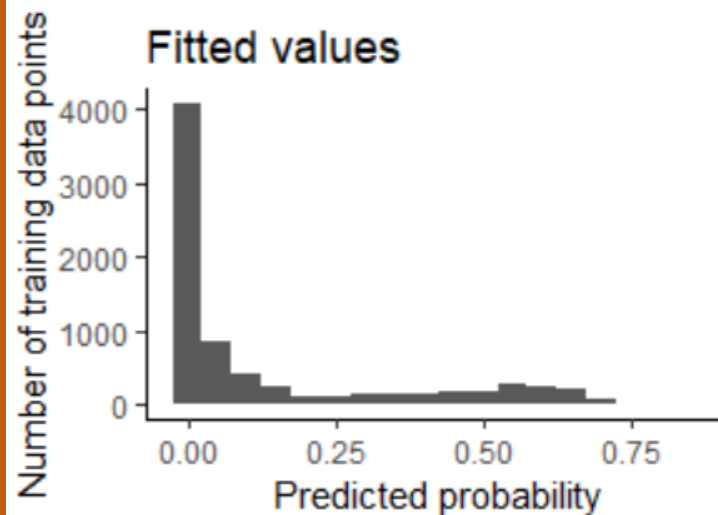
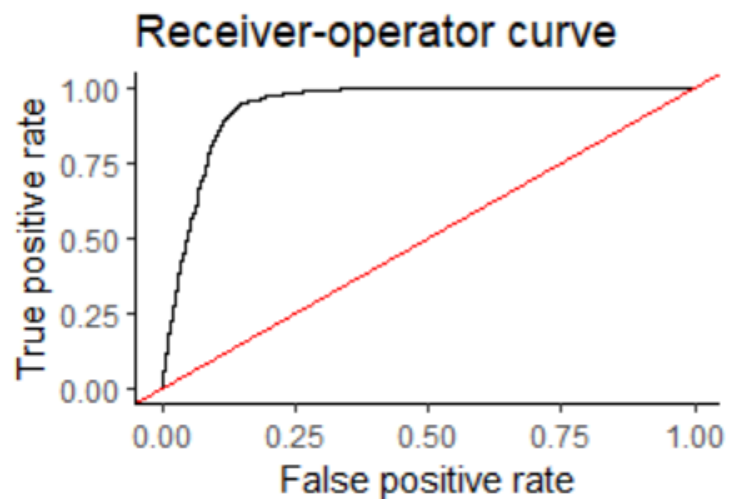
Model evaluation

A. Márcia Barbosa

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

BART predictive accuracy

summary(mod)



BART predictive accuracy

- generally good **discrimination** capacity with less overfitting than other powerful methods
- generally good also with other aspects of model performance, e.g. **calibration**

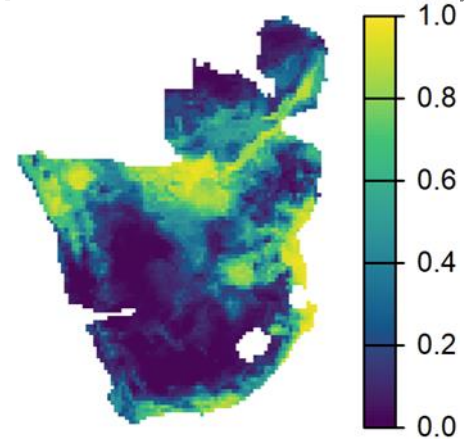
Two facets of model accuracy

- discrimination, classification

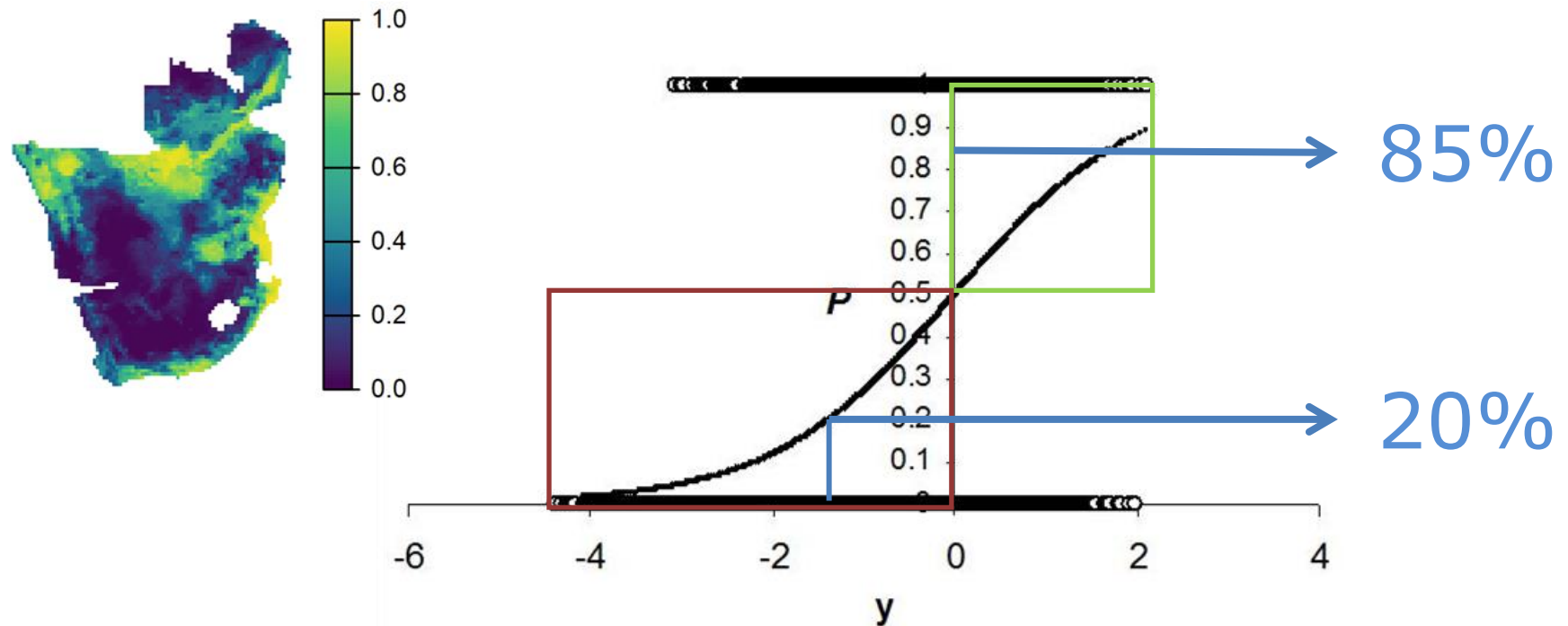
how well the model distinguishes **presence**
from absence (e.g. sensitivity, AUC, TSS, Kappa)

- calibration / reliability

how the continuous predicted values reflect
frequency of presence (e.g. Miller slope,
Hosmer-Lemeshow goodness-of-fit, Boyce index)



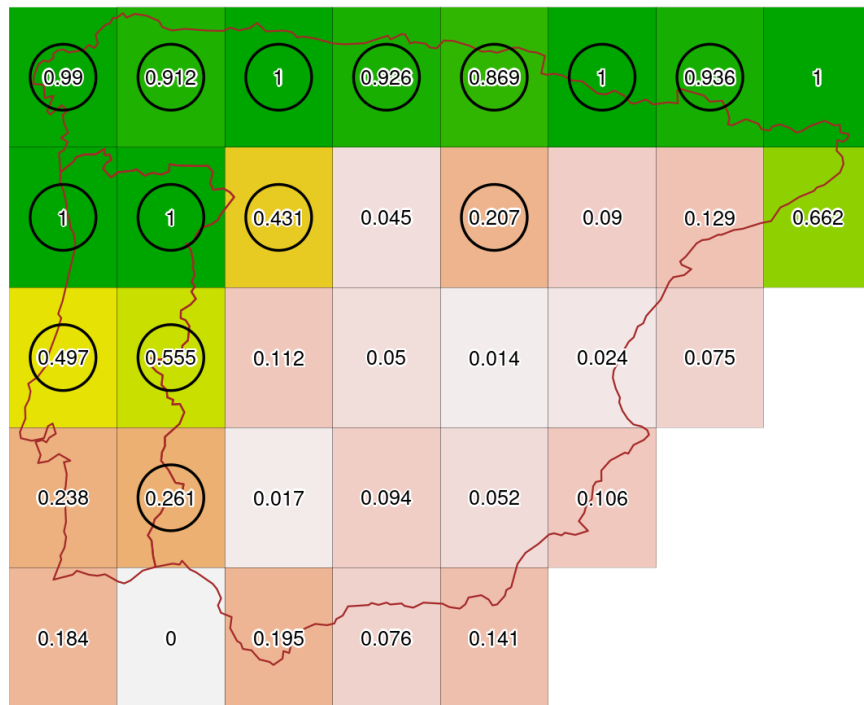
Two facets of model accuracy



model predictions are continuous values, not
binary "presence" or "absence" – like
species presence in nature is also gradual!

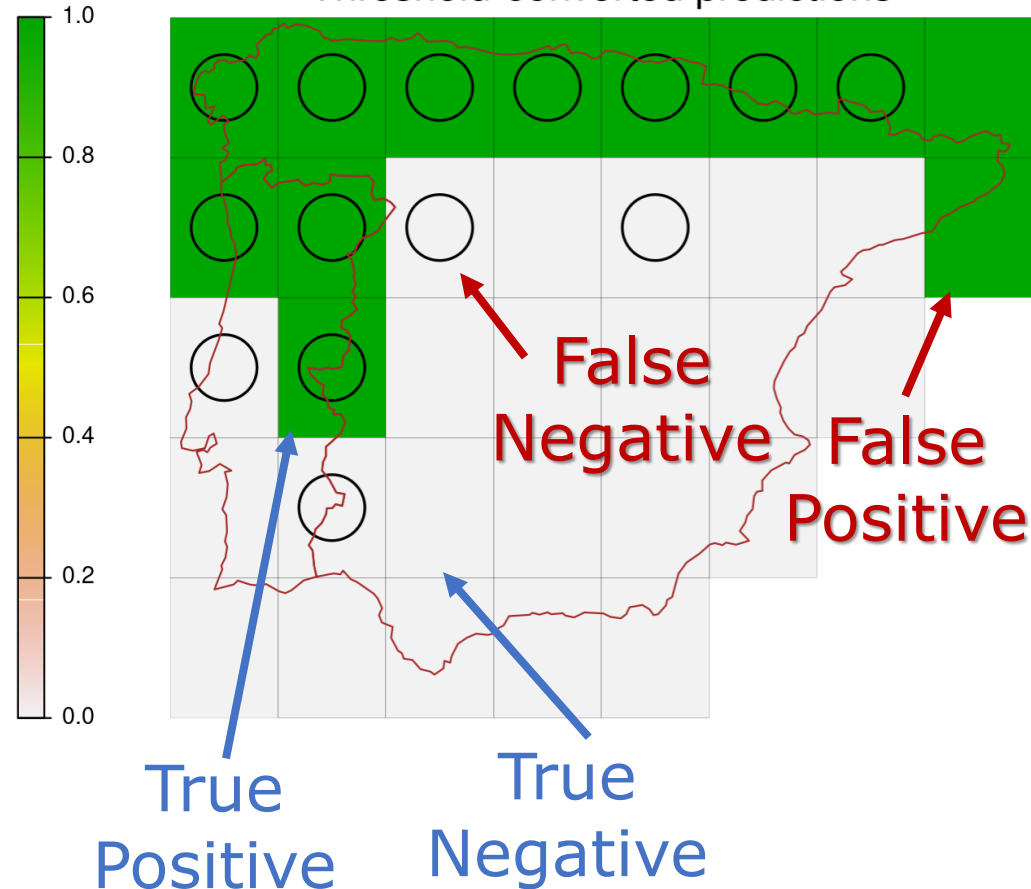
Discrimination and classification

Original predictions



○ species presence

Threshold-converted predictions



for interactive threshold choice effects:
<https://ambiogeo.shinyapps.io/modevapp/>

Discrimination and classification

- **overall accuracy** / correct classification rate
- **sensitivity** / **recall** / true positive rate: % correct prediction of presences
- **specificity** / true negative rate: % correct prediction of absences
- **precision** / positive pred. value: % predicted presences that are observed as such

[0, 1]

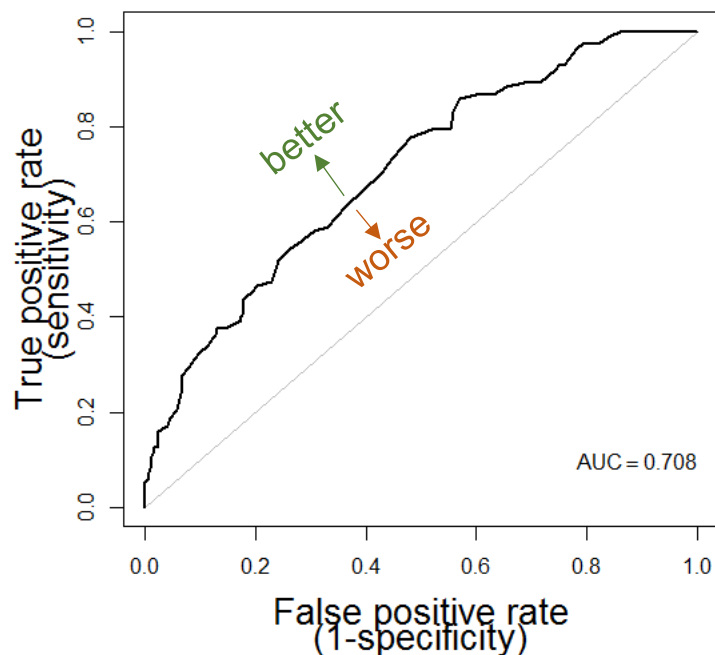
Discrimination and classification

- **Cohen's kappa** correct classification over chance
- **TSS (true skill statistic)** combines sensitivity and specificity

$[-1, 1]$

Discrimination and classification

- AUC (Area Under the ROC Curve)



[0, 1] but 0.5 means random prediction

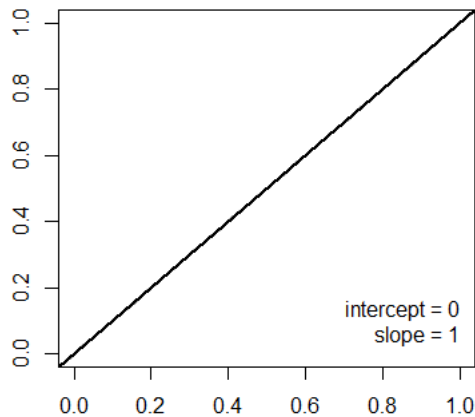
equates Wilcoxon test comparing predicted values for presences vs. absences

no single threshold required, but still measures **discrimination** (threshold implicit)

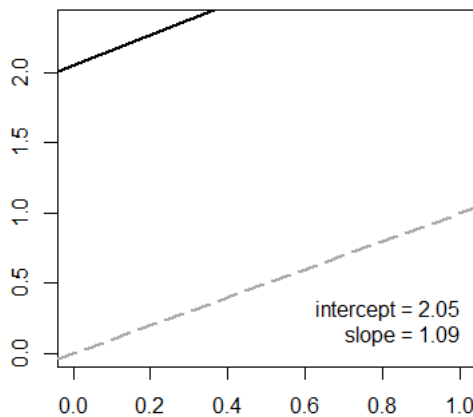
Calibration / reliability

- **Miller calibration** slope and intercept of a GLM of the observed values against the *logit* of predicted probabilities

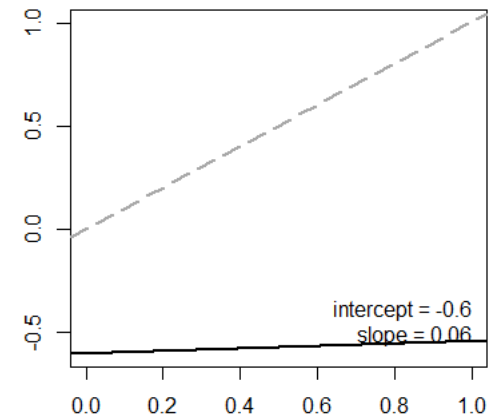
perfect



semi-good



not good



Practical: Evaluate model performance

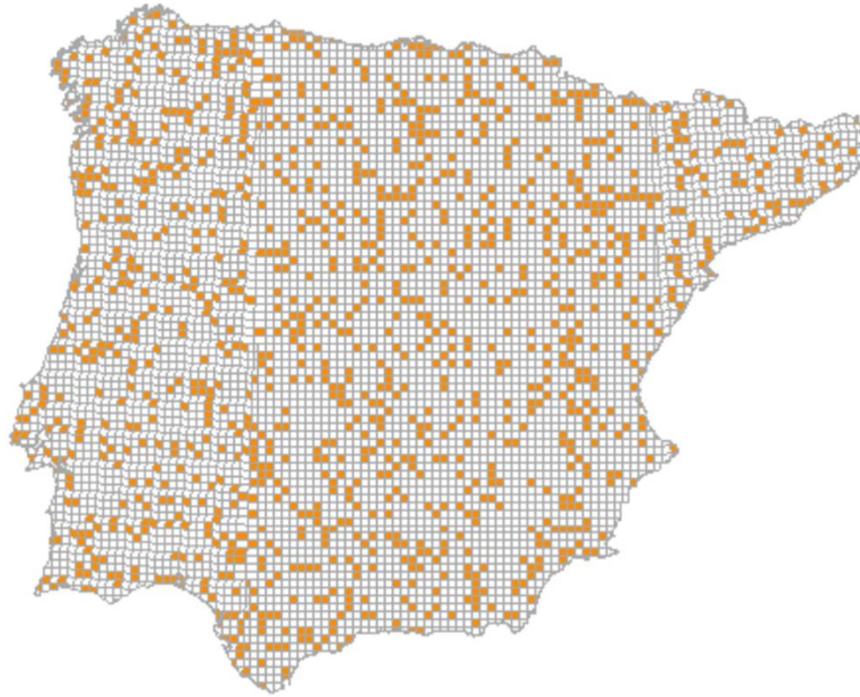
```
library(modEvA)
```

```
AUC(obs, pred) # ROC curve
```

```
threshMeasures(obs, pred, thresh = "maxTSS")
```

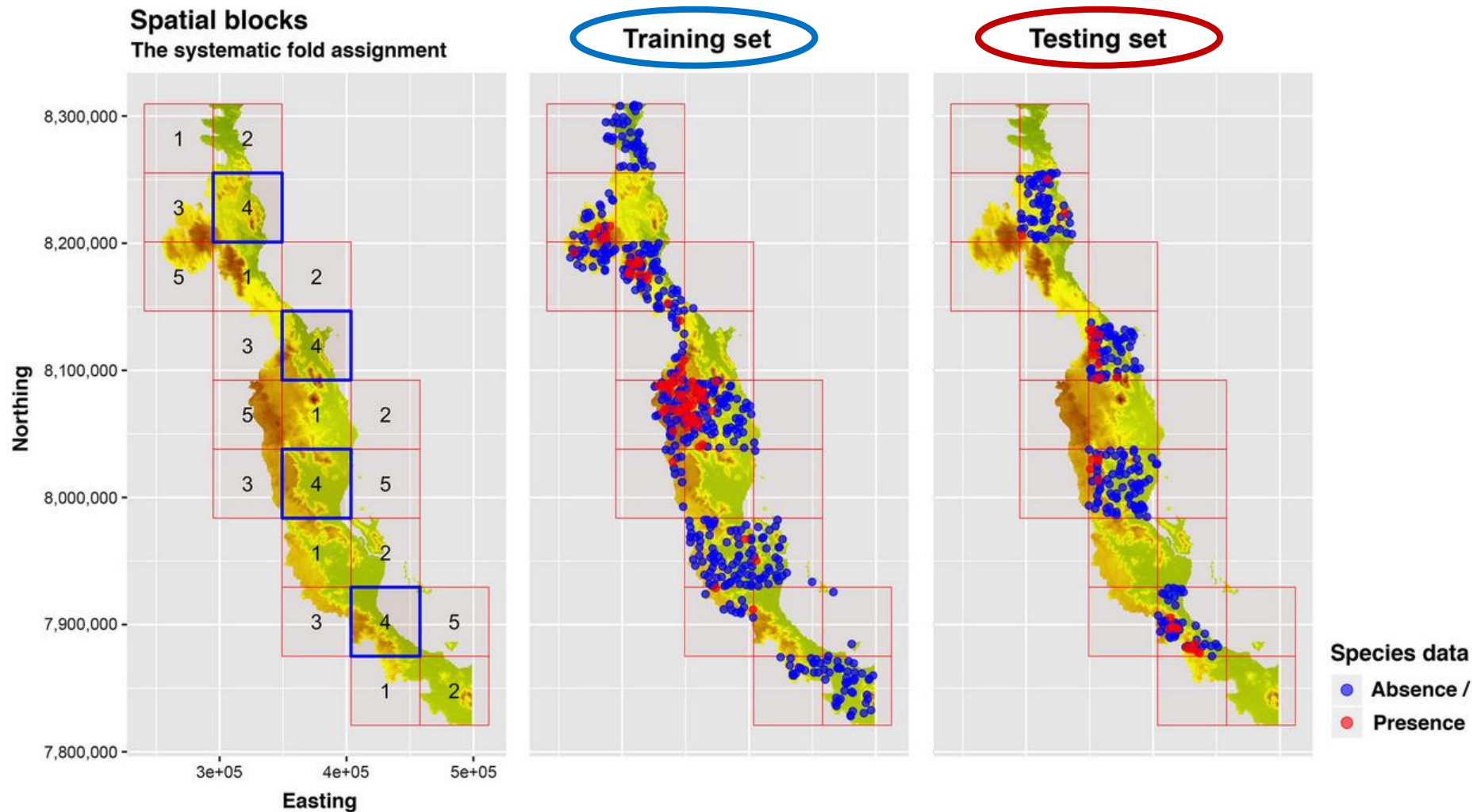
```
MillerCalib(obs, pred)
```

Random cross-validation

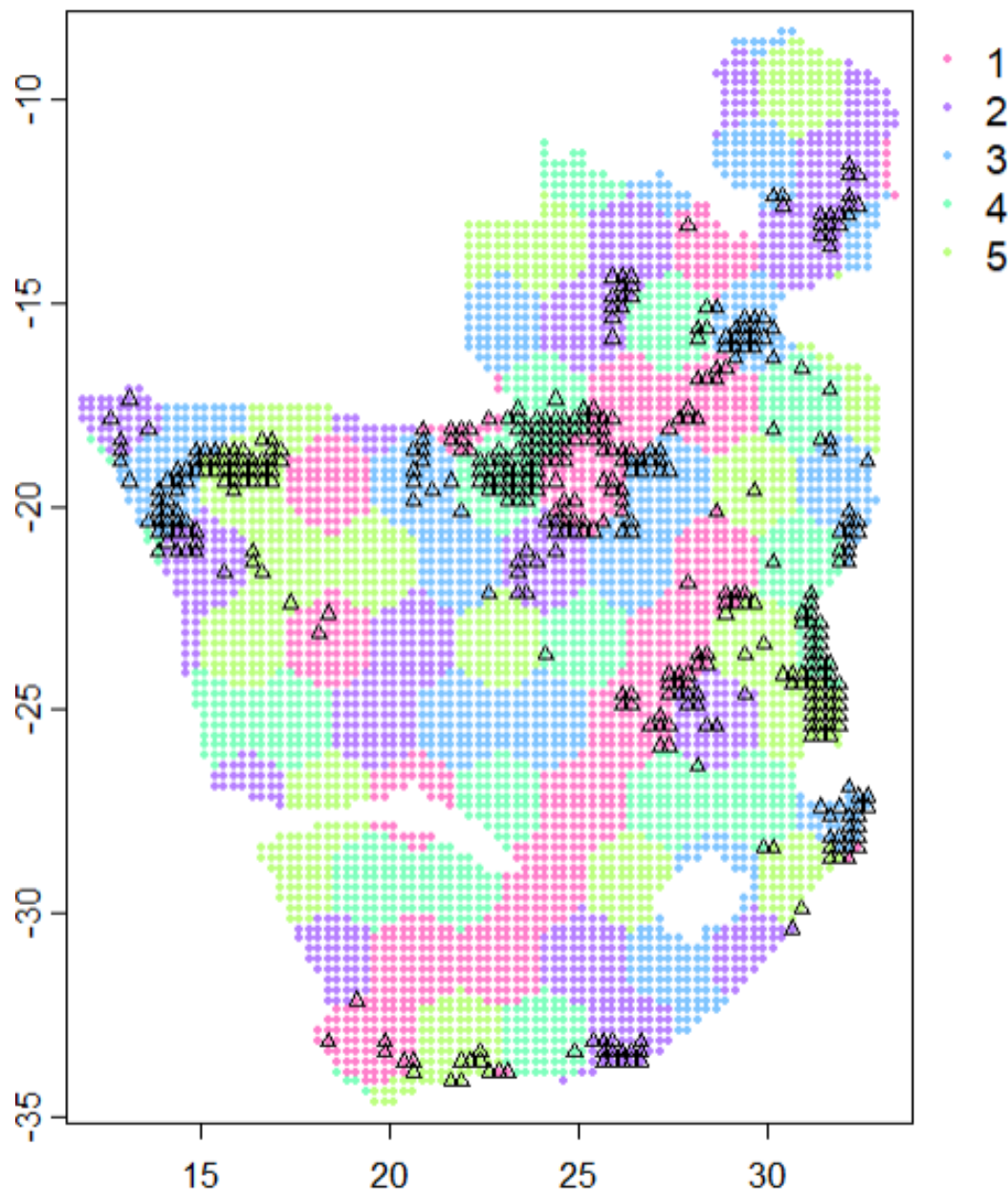


- each fold leaves out a random sample of sites
- but spatial autocorrelation makes it easier to predict near modelled sites
- so, random cross-validation can overestimate predictive capacity

Block cross-validation



Block cross-validation



assess model
performance
both in nearby and
more distant areas
than those used
for model fitting

Practical: Block cross-validation

```
library(blockCV)
```

```
blocks <- cv_spatial(dat, column = "presence",  
r = vars_rast[[1]], k = 5, size =  
sqrt(expanse(mod_region) / 20))
```

```
dat$fold <- blocks$folds_ids
```

```
for (f in folds) { ... }
```



*loops to model outside
each fold and evaluate
the model on that fold*

Practical: Block cross-validation

model cross-validation performance

