Interpretable ML for biodiversity

An introduction using species distribution models

Timothée Poisot Université de Montréal

October 1, 2024

MAIN GOALS

1. How do we produce a model?
2. How do we convey that it works?
3. How do we talk about how it makes predictions?
4. How do we use it to guide actions?

THE STEPS

1. Get data about species occurrences
2. Build a classifier and make it as good as we can
3. Measure its performance
4. Explain some predictions
5. Generate counterfactual explanations
6. Briefly discuss ensemble models

BUT WHY...

... think of SDM as a ML problem? Because they are! We want to learn a predictive algorithm from data

... the focus on explainability? We cannot ask people to *trust* - we must *convince* and *explain*

§ 1

Problem statement

We have information about a species

THE PROBLEM IN OTHER WORDS

We have a series of observations $y \in \mathbb{B}$, and predictors variables $\mathbf{X} \in \mathbb{R}$

We want to find an algorithm $f(\mathbf{x}) = \hat{y}$ that results in the distance between $\hat{y}$ and y being *small*

SETTING UP THE DATA FOR OUR EXAMPLE

The predictor data will come from CHELSA2 - we will start with the 19 BioClim variables

We will use data on observations of *Turdus torquatus* in Switzerland, downloaded from the copy of the eBird dataset on GBIF

THE OBSERVATION DATA

PROBLEM!

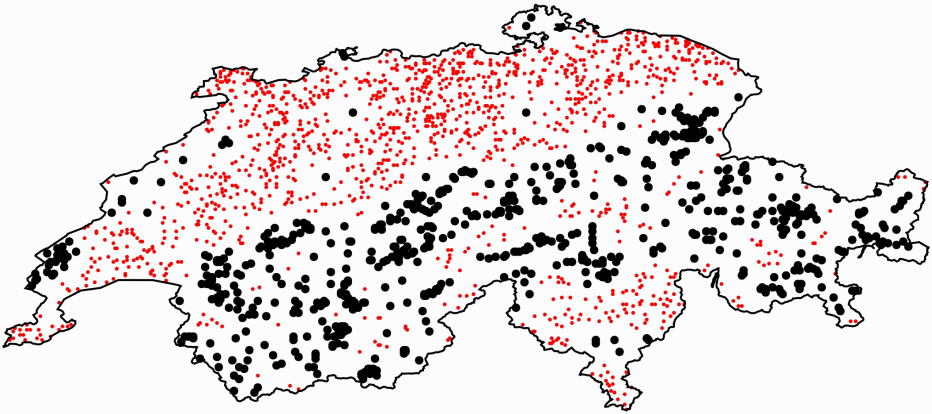
We want $\hat{y} \in \mathbb{B}$, and so far we are missing negative values

SOLUTION!

pseudo-absences

what are the assumptions we make

THE (INFLATED) OBSERVATION DATA



§ 2

Training the model

$$P(+|x) = \frac{P(+)}{P(x)} P(x|+)$$

$$\hat{y} = \text{argmax}_j \, P(\mathbf{c}_j) \prod_i P(\mathbf{x}_i|\mathbf{c}_j)$$

$$P(x|+) = \text{pdf}(x, \mathcal{N}(\mu_+, \sigma_+))$$

SETUP

Can we train the model

assumes parallel universes with slightly less data

is the model good?

NULL CLASSIFIERS

coin flip

no skill

constant

Model	MCC	PPV	NPV	DOR	Accuracy
noskill	-3.09497e-17	0.339373	0.660627	1.0	0.551602
coinflip	-0.321254	0.339373	0.339373	0.263902	0.339373
constantpositive	0.0	0.339373	NaN	NaN	0.339373
constantnegative	0.0	NaN	0.660627	NaN	0.660627

CROSS-VALIDATION STRATEGY

k-fold

validation / training / testing

CROSS-VALIDATION RESULTS

Model	MCC	PPV	NPV	DOR	Accuracy
noskill	-3.09497e-17	0.339373	0.660627	1.0	0.551602
coinflip	-0.321254	0.339373	0.339373	0.263902	0.339373
constantpositive	0.0	0.339373	NaN	NaN	0.339373
constantnegative	0.0	NaN	0.660627	NaN	0.660627
Validation	0.384895	0.647967	0.768219	6.60144	0.736062
Training	0.392684	0.652914	0.770902	6.33576	0.740233

WHAT TO DO IF THE MODEL IS TRAINABLE?

train it!

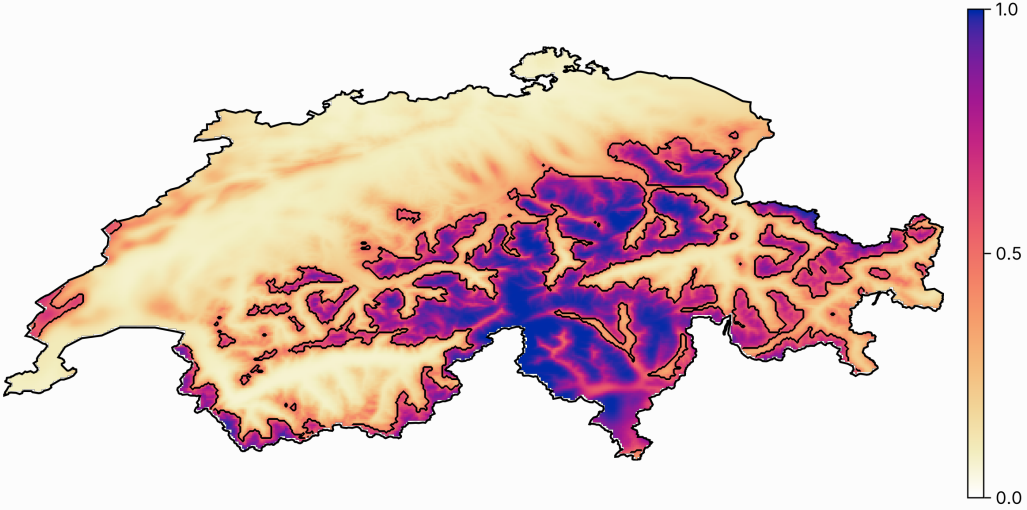
re-use the full dataset

A NOTE ON DATA LEAKAGE

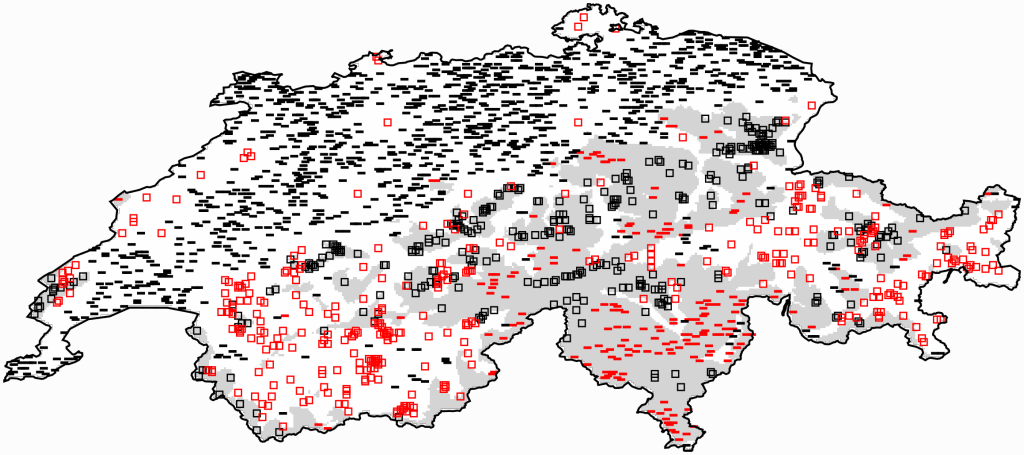
DATA TRANSFORMATION USING PCA

THE MODEL TRAINING PIPELINE

INITIAL PREDICTION



HOW IS THIS MODEL WRONG?



CAN WE IMPROVE ON THIS MODEL?

variable selection

data transformation

hyper-parameters tuning

will focus on the later (same process for the two above)

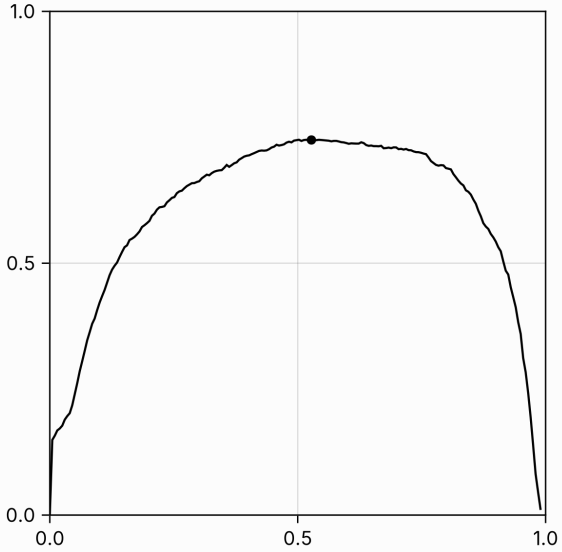
MOVING THESHOLD CLASSIFICATION

p plus > p minus means threshold is 0.5

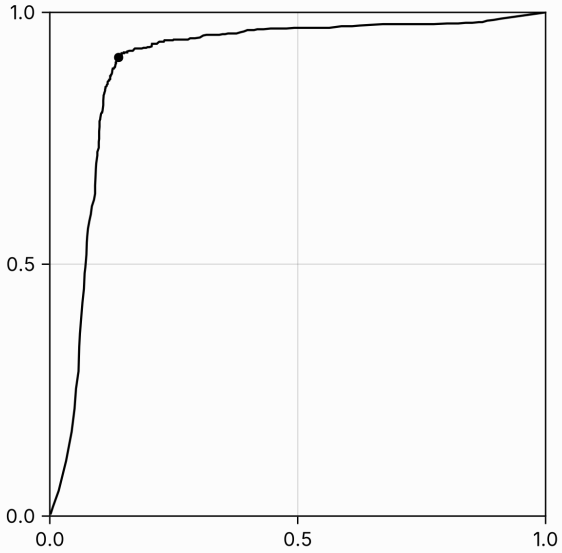
is it?

how do we check this

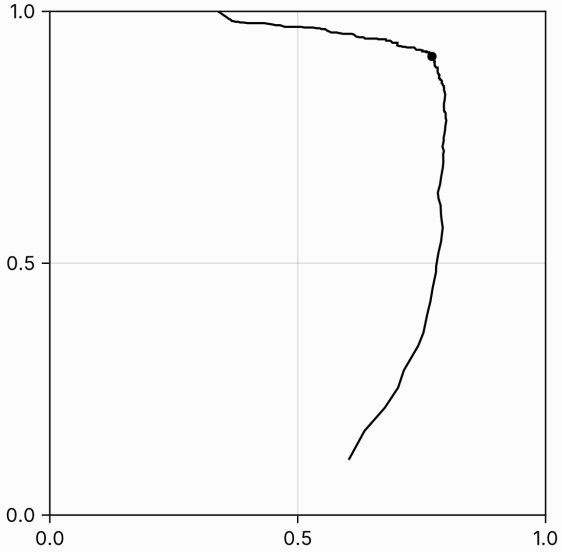
LEARNING CURVE FOR THE THRESHOLD



RECEIVER OPERATING CHARACTERISTIC



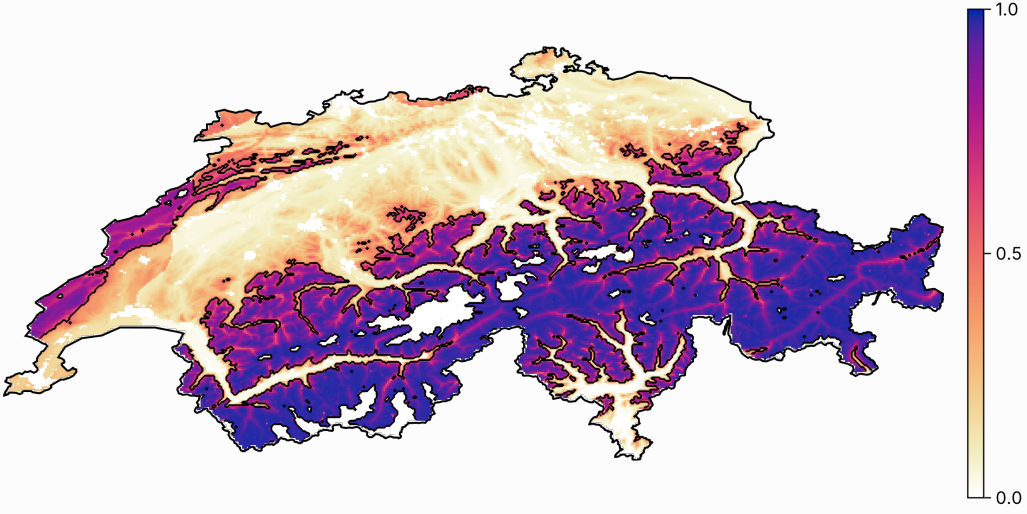
PRECISION-RECALL CURVE



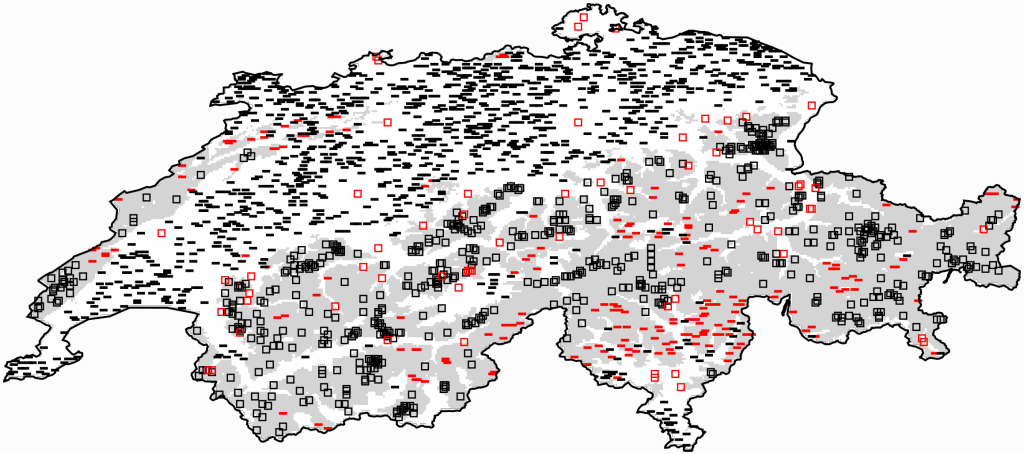
REVISITING THE MODEL PERFORMANCE

Model	MCC	PPV	NPV	DOR	Accuracy
noskill	-3.09497e-17	0.339373	0.660627	1.0	0.551602
coinflip	-0.321254	0.339373	0.339373	0.263902	0.339373
constantpositive	0.0	0.339373	NaN	NaN	0.339373
constantnegative	0.0	NaN	0.660627	NaN	0.660627
Previous	0.384895	0.647967	0.768219	6.60144	0.736062
Validation	0.744385	0.771569	0.947628	79.99	0.876604
Training	0.73678	0.764253	0.947113	59.0436	0.873436

UPDATED PREDICTION



HOW IS THIS MODEL BETTER?



REVISITING ASSUMPTIONS

- pseudo-absences
- not just a statistical exercise

VARIABLE IMPORTANCE

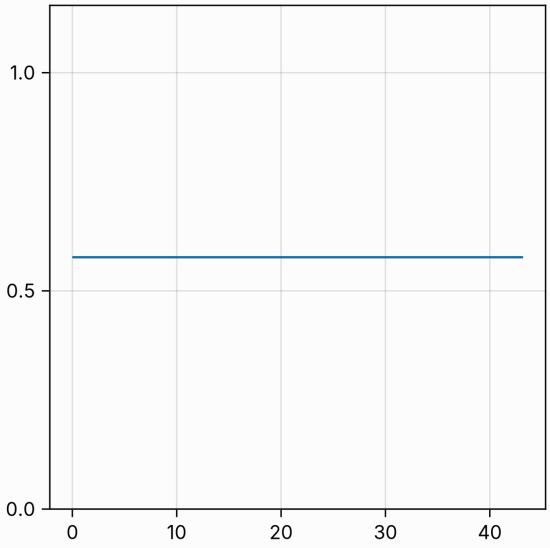
Layer	Variable	Import.
1	BIO1	0.59825
5	BIO5	0.233834
8	BIO8	0.104139
28	Urban/Built-up	0.0412146
29	Snow/Ice	0.0225633

§ 3

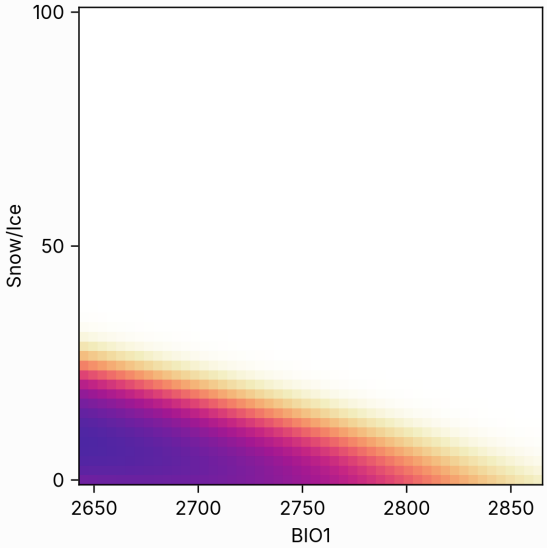
But why?

AN ECOLOGY TOOL: PARTIAL RESPONSE CURVES

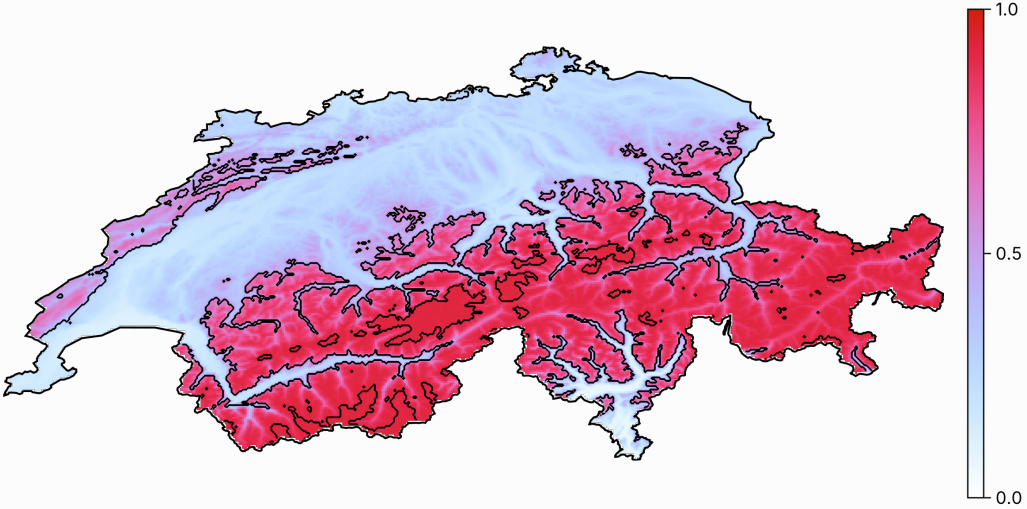
EXAMPLE WITH TEMPERATURE



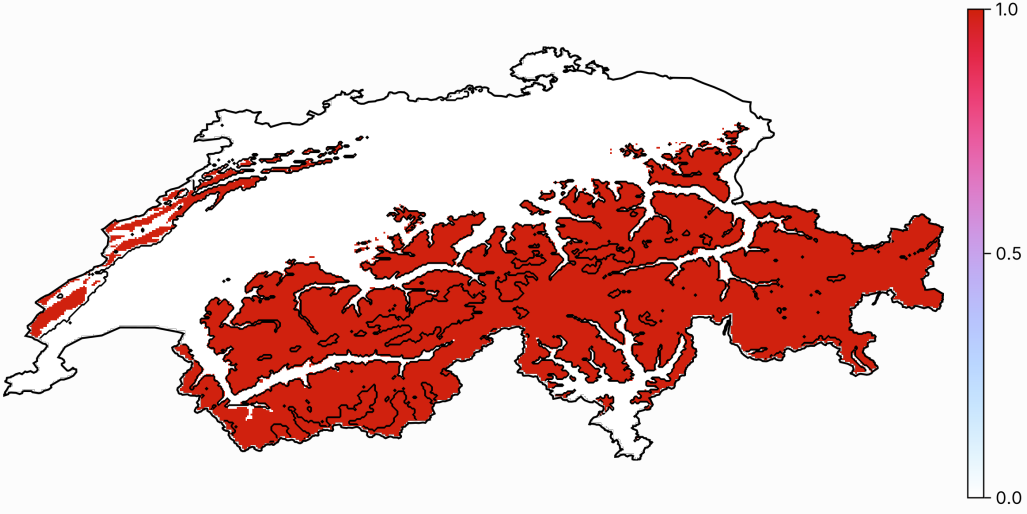
EXAMPLE WITH TWO VARIABLES



SPATIALIZED PARTIAL RESPONSE PLOT



SPATIALIZED PARTIAL RESPONSE (BINARY OUTCOME)



INFLATED RESPONSE CURVES

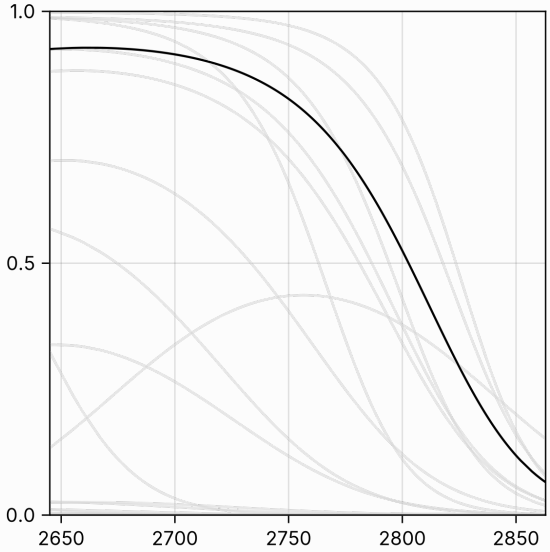
Averaging the variables is masking a lot of variability!

Alternative solution:

1. Generate a grid for all the variables
2. For all combinations in this grid, use it as the stand-in for the variables to replace

In practice: Monte-Carlo on a reasonable number of samples.

EXAMPLE



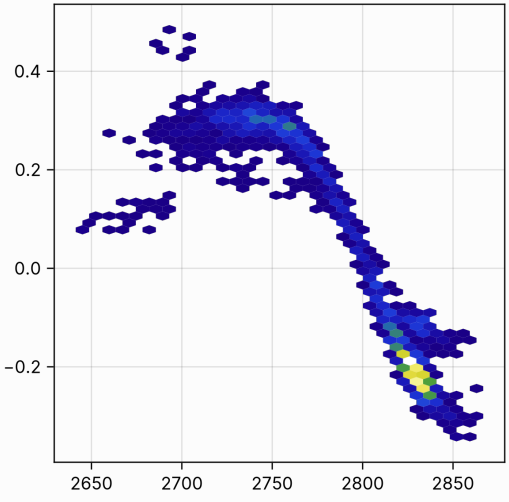
LIMITATIONS

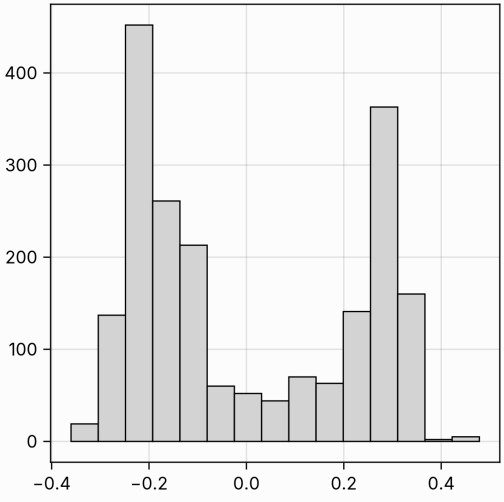
- partial responses can only generate model-level information
- they break the structure of values for all predictors at the scale of a single observation
- their interpretation is unclear

SHAPLEY

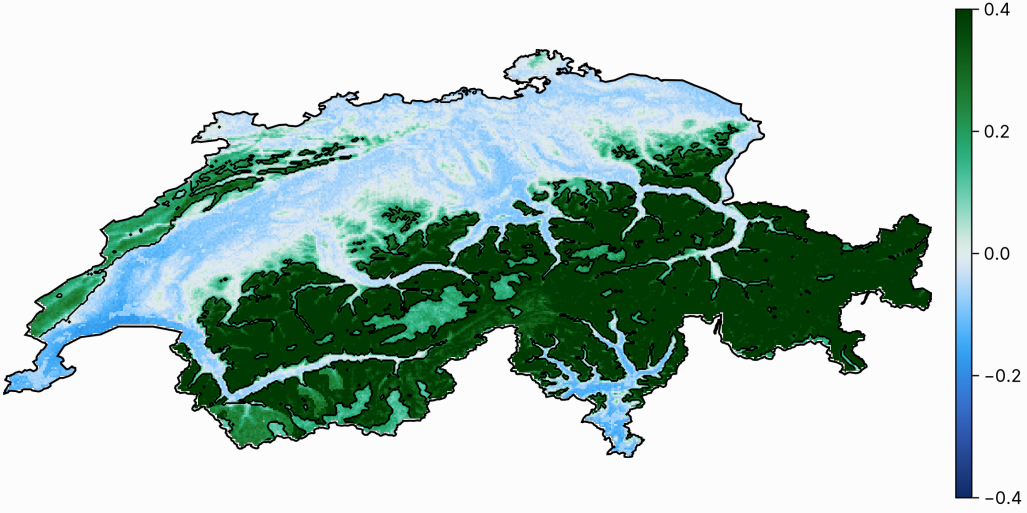
EXAMPLE

RESPONSE CURVES REVISITED





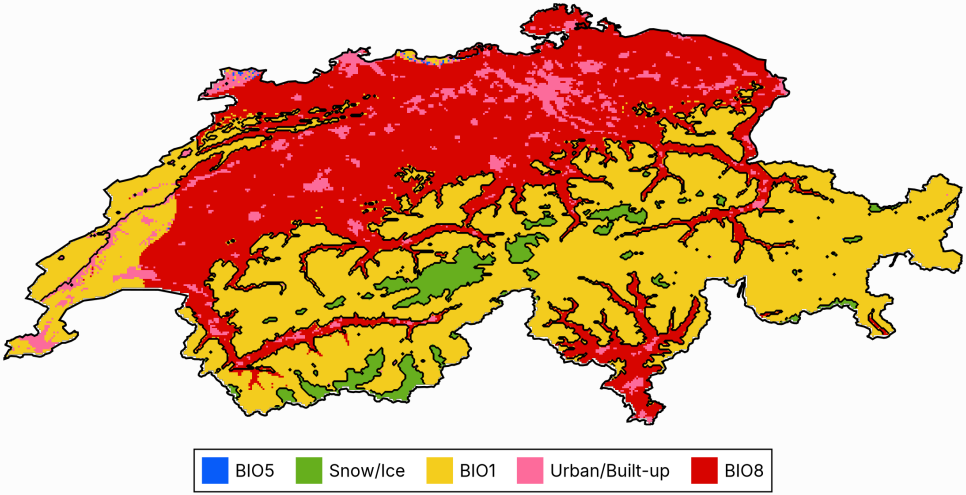
ON A MAP



VARIABLE IMPORTANCE REVISITED

Layer	Variable	Import.	Shap. imp.
1	BIO1	0.59825	0.415164
8	BIO8	0.104139	0.261453
5	BIO5	0.233834	0.137676
28	Urban/Built-up	0.0412146	0.125304
29	Snow/Ice	0.0225633	0.0604023

MOST IMPORTANT PREDICTOR



all in a single model so we can ask effect of variable instead of effect of PC1 or whatever

§ 4

What if?

what they are

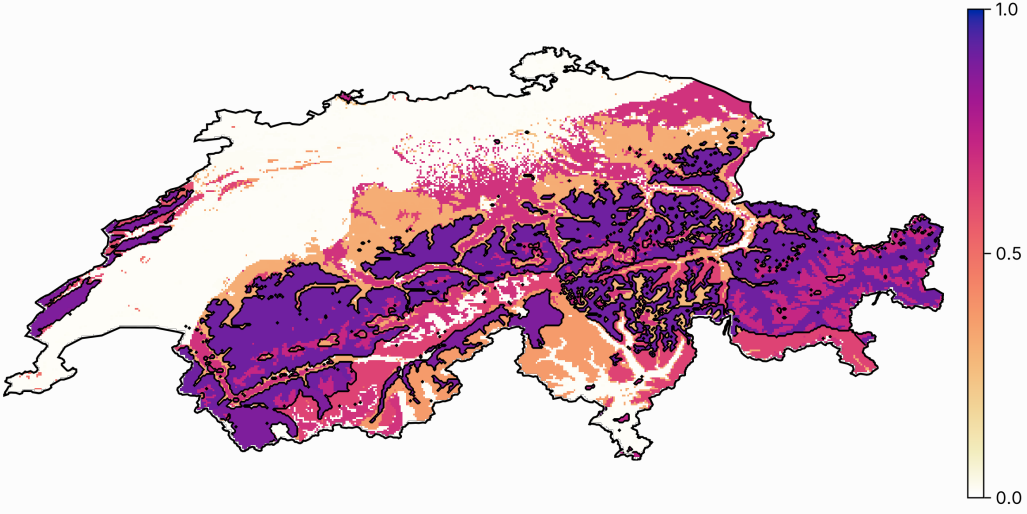
SETTING UP A NEW PROBLEM

- land use
- decision tree - very easy to overfit
- at most 18 nodes of depth at most 7
- same process

VARIABLE IMPORTANCE

Layer	Variable	Relative importance
1	BIO1	0.637614
5	BIO5	0.248376
8	BIO8	0.0824429
28	Urban/Built-up	0.0190532
29	Snow/Ice	0.0125144

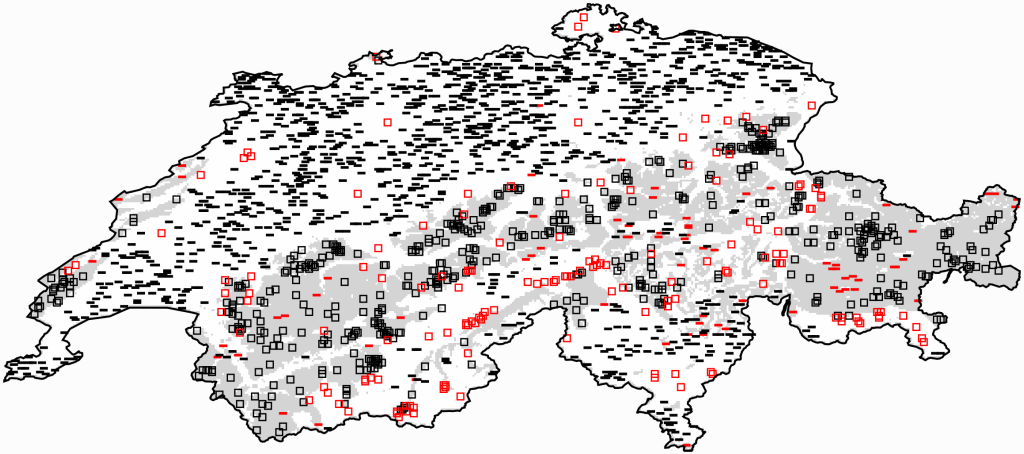
VISUALIZING THE PREDICTION



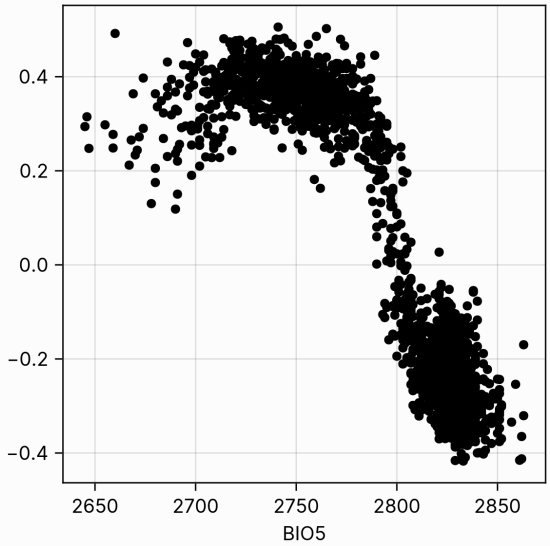
THE RASHOMON EFFECT

- different but equally likely alternatives
- happens at all steps in the process
- variable selected, threshold used, model type

VISUALIZING THE ERRORS



PARTIAL RESPONSE (SHAPLEY)



GENERATING A COUNTERFACTUAL

learning rate and loss function

use on prediction score and not yes/no!

§ 5

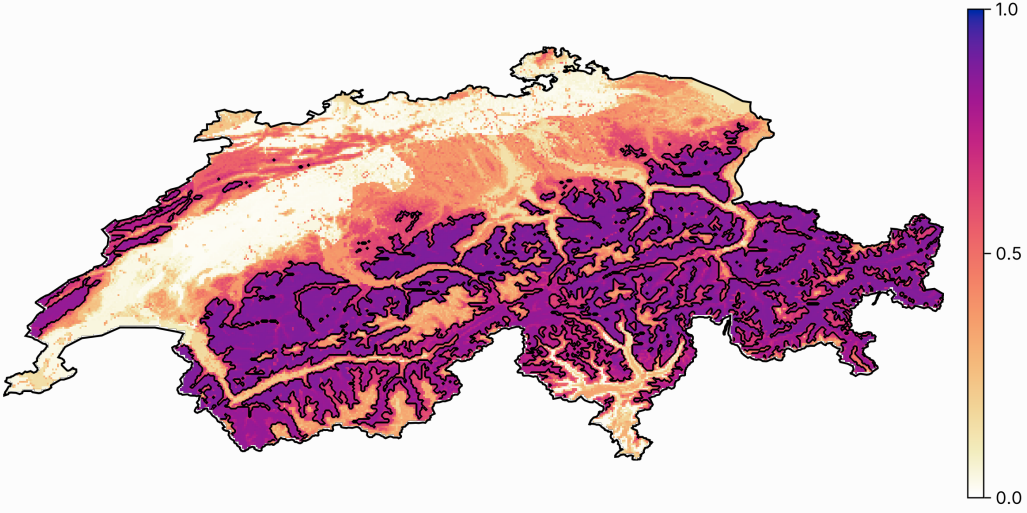
Ensemble models

LIMITS OF A SINGLE MODEL

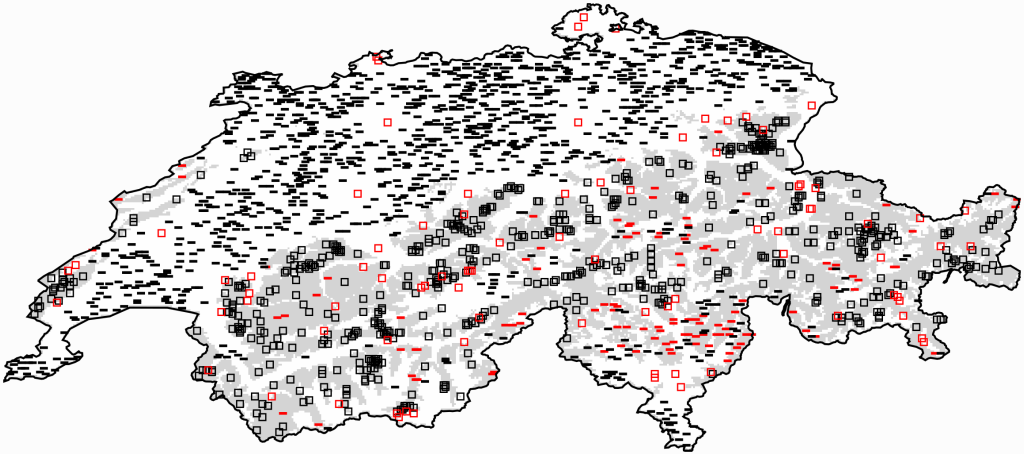
- a single model
- different parts of data may have different signal
- do we need all the variables all the time?
- bias v. variance tradeoff
- limit overfitting

AN EXAMPLE OF BAGGING: ROTATION FOREST

PREDICTION OF THE ROTATION FOREST



PREDICTION OF THE ROTATION FOREST



UNCERTAINTY

SETTING UP AN HETEROGENEOUS ENSEMBLE

§ 6

Conclusions