Interpretable ML for biodiversity

An introduction using species distribution models

Timothée Poisot Université de Montréal

September 29, 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

THE PROBLEM IN ECOLOGICAL TERMS

We have information about a species

THE PROBLEM IN OTHER WORDS

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

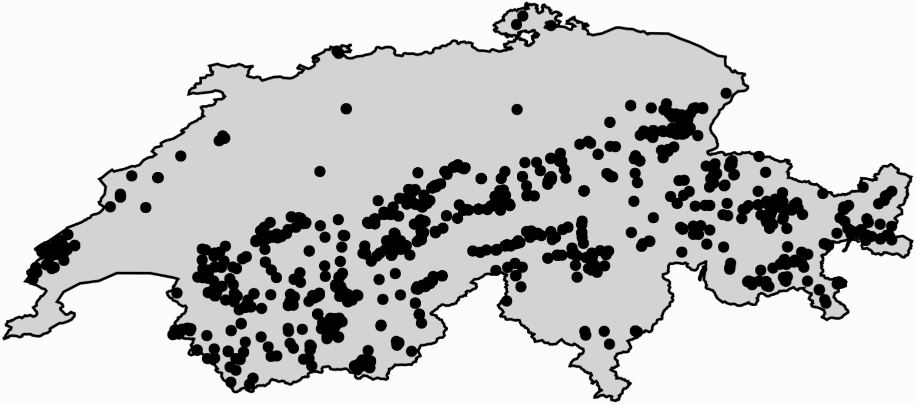
We want to find an algorithm $f(\mathbf{x}) = \hat{y}$ that results in the distance between $\hat{y}$ and y being *small*

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!

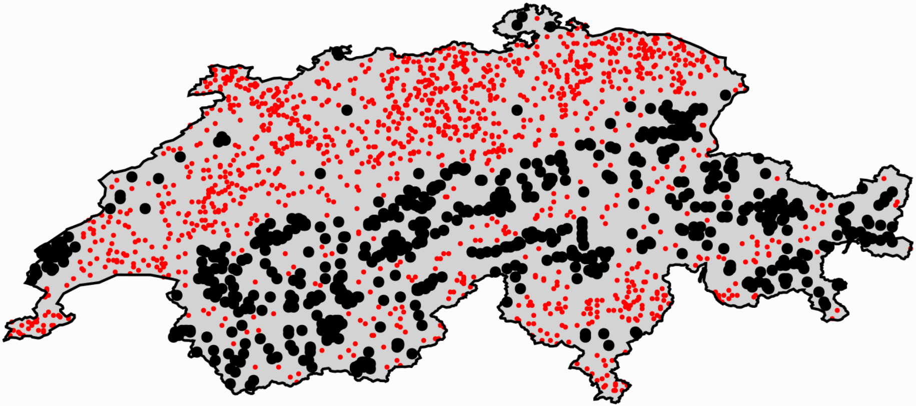
We want $\hat{y} \in \mathbb{B}$, and so far we are missing negative values

SOLUTION!

pseudo-absences

what are the assumptions we make

THE (INFLATED) OBSERVATION DATA



§ 2

Training the model

THE NAIVE BAYES CLASSIFIER

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

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

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

SETUP

Can we train the model

assumes parallel universes with slightly less data

is the model good?

NULL CLASSIFIERS

coin flip

no skill

constant

EXPECTATIONS

Model	MCC	PPV	NPV	DOR	Accuracy
noskill	0.0	0.339825	0.660175	1.0	0.551312
coinflip	-0.320351	0.339825	0.339825	0.264967	0.339825
constantpositive	0.0	0.339825	NaN	NaN	0.339825
constantnegative	0.0	NaN	0.660175	NaN	0.660175

CROSS-VALIDATION STRATEGY

k-fold

validation / training / testing

CROSS-VALIDATION RESULTS

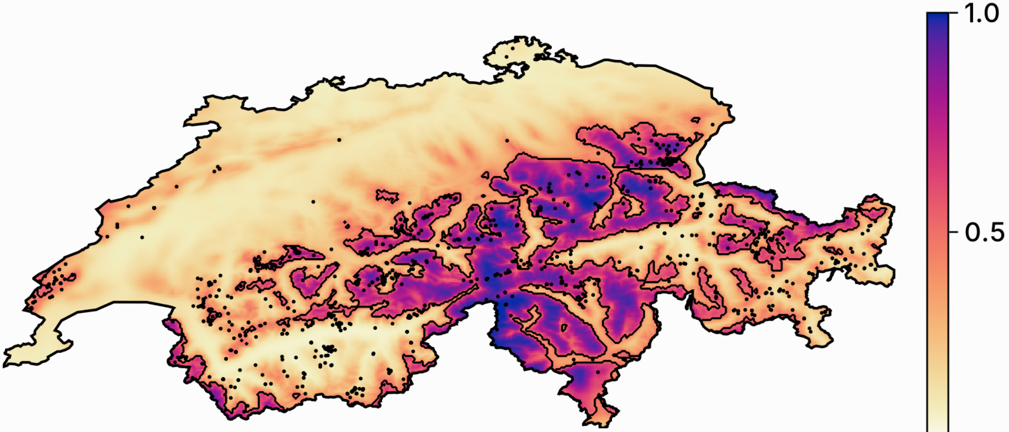
Model	MCC	PPV	NPV	DOR	Accuracy
noskill	0.0	0.339825	0.660175	1.0	0.551312
coinflip	-0.320351	0.339825	0.339825	0.264967	0.339825
constantpositive	0.0	0.339825	NaN	NaN	0.339825
constantnegative	0.0	NaN	0.660175	NaN	0.660175
Validation	0.305111	0.594557	0.742387	4.5723	0.706929
Training	0.316304	0.60115	0.746074	4.4334	0.710267

WHAT TO DO IF THE MODEL IS TRAINABLE?

train it!

re-use the full dataset

INITIAL PREDICTION



CAN WE IMPROVE ON THIS MODEL?

variable selection

data transformation

hyper-parameters tuning

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

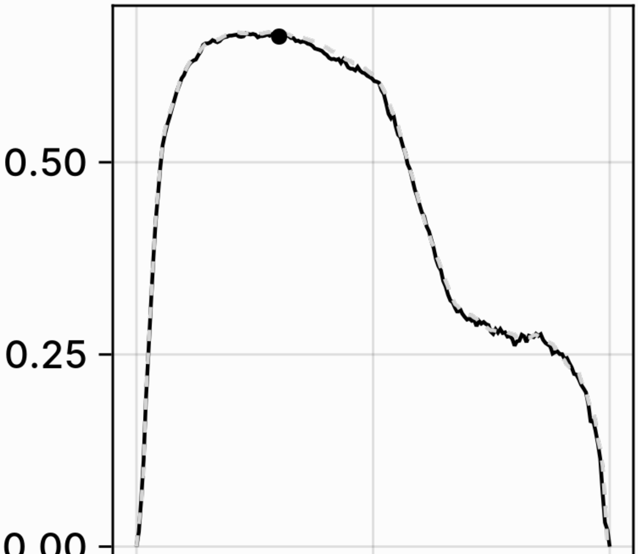
MOVING THESHOLD CLASSIFICATION

p plus > p minus means threshold is 0.5

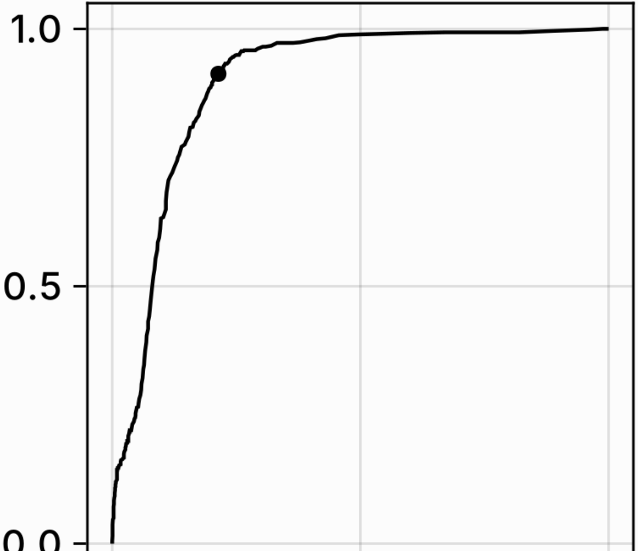
is it?

how do we check this

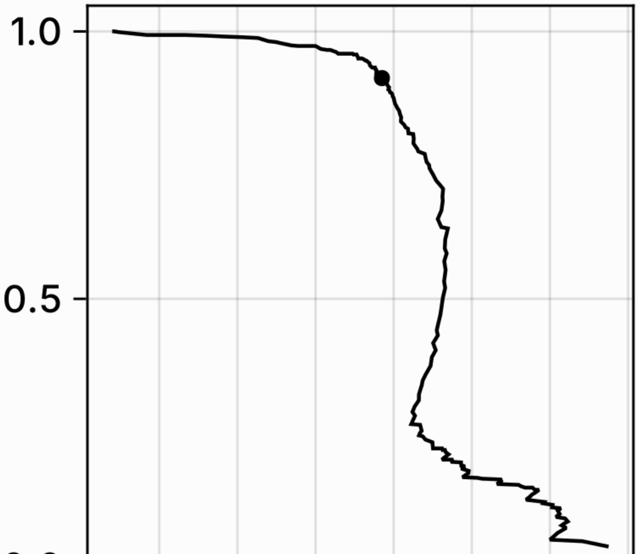
LEARNING CURVE FOR THE THRESHOLD



RECEIVER OPERATING CHARACTERISTIC

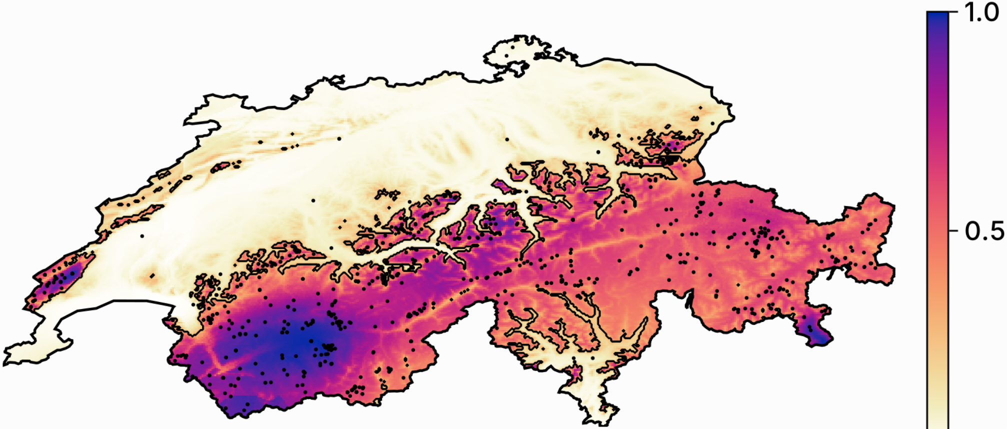


PRECISION-RECALL CURVE



REVISTING THE MODEL PERFORMANCE

UPDATED PREDICTION



VARIABLE IMPORTANCE

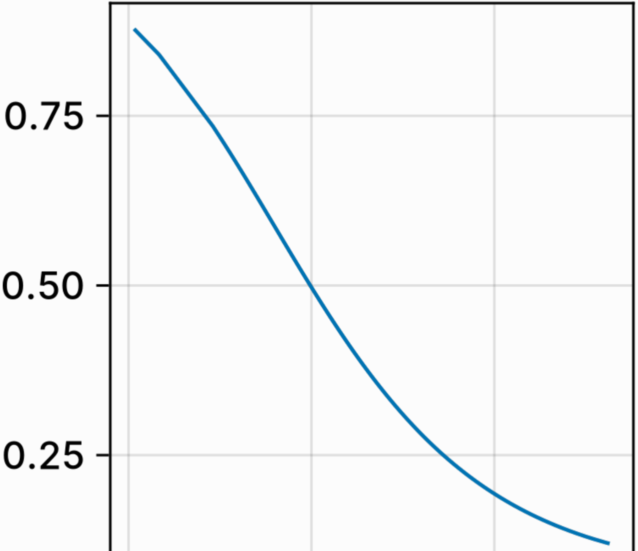
§ 3

But why?

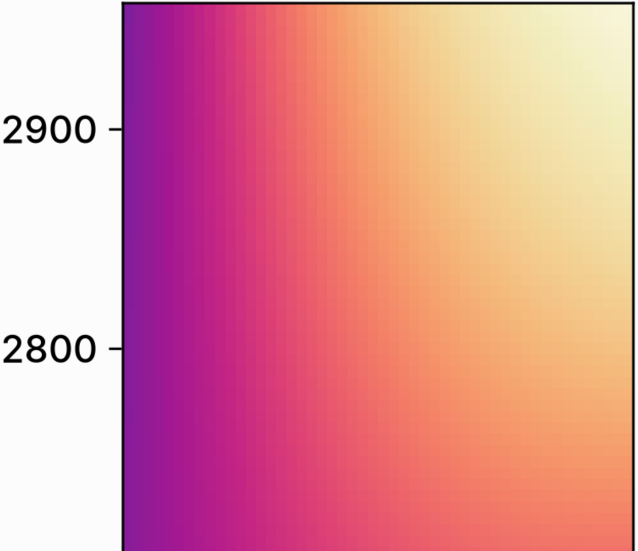
INTRO EXPLAINABLE

AN ECOLOGY TOOL: PARTIAL RESPONSE CURVES

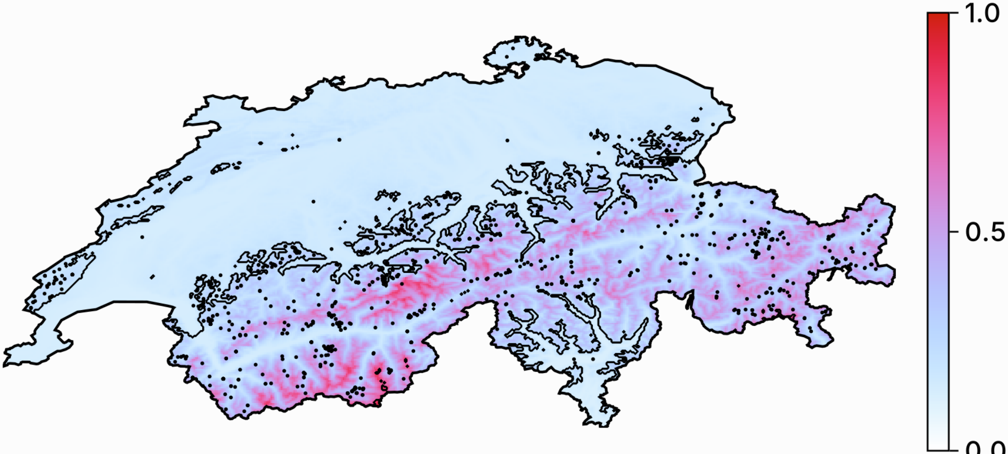
EXAMPLE WITH TEMPERATURE



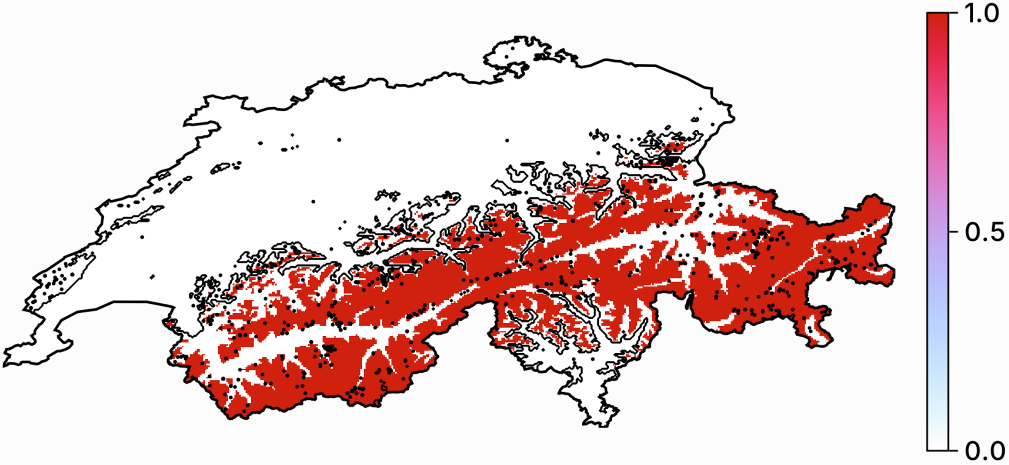
EXAMPLE WITH TWO VARIABLES



SPATIALIZED PARTIAL RESPONSE PLOT



SPATIALIZED PARTIAL RESPONSE (BINARY OUTCOME)



INFLATED RESPONSE CURVES

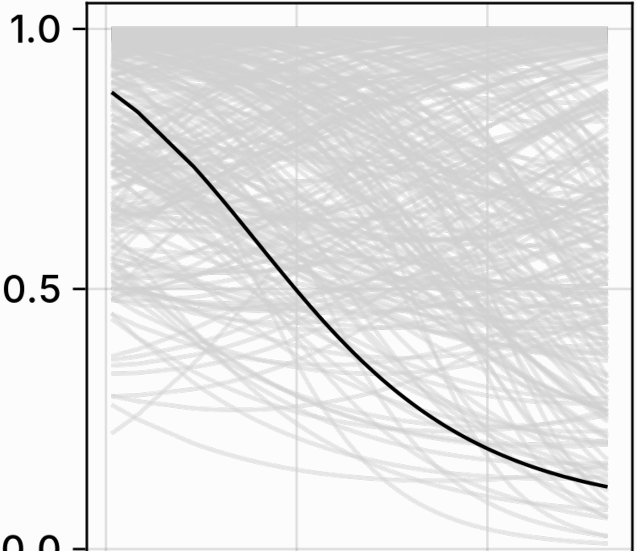
Averaging the variables is masking a lot of variability!

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.

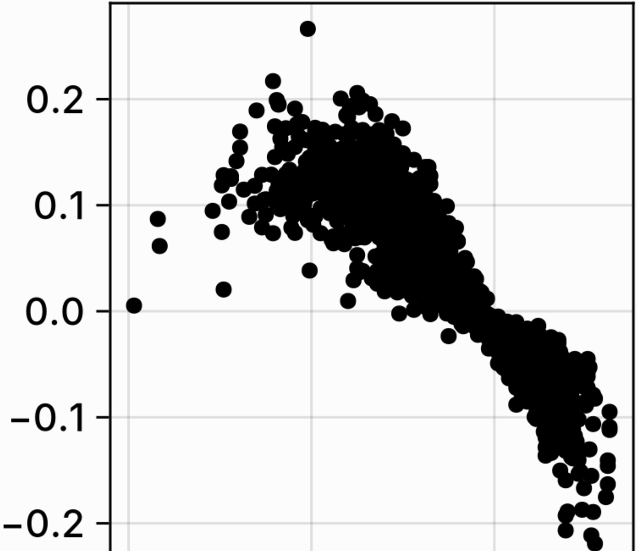
EXAMPLE



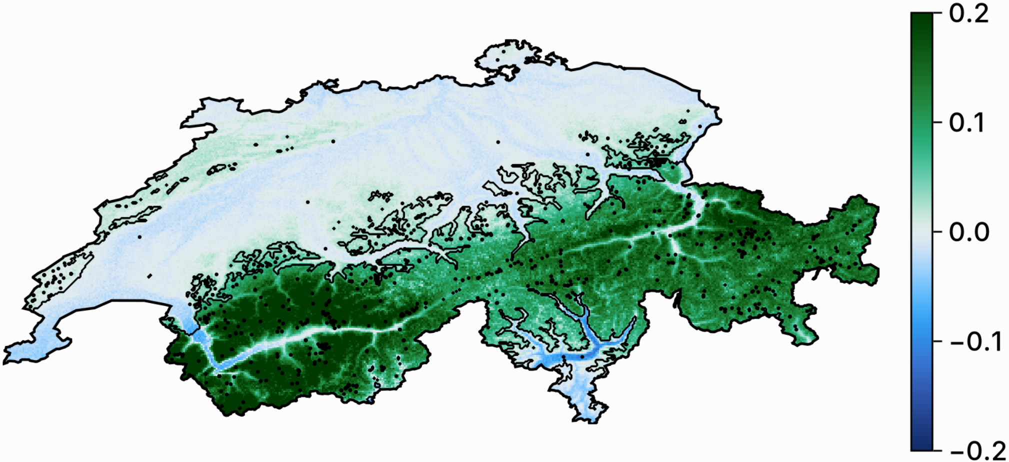
LIMITATIONS

- 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

EXAMPLE


0.2
0.1
0.0
−0.1
−0.2

ON A MAP



VARIABLE IMPORTANCE REVISITED

with shapley

MOST IMPORTANT PREDICTOR

mosaic map

§ 4

What if?

what they are

§ 5

Ensemble models

§ 6

Conclusions