



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**

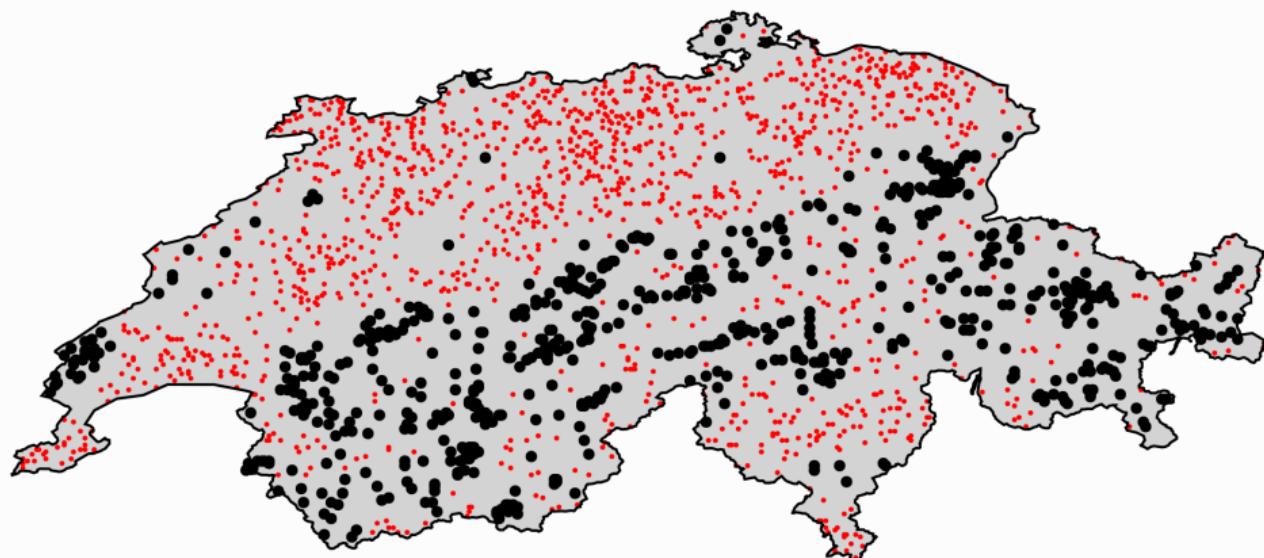


SOLUTION!

pseudo-absences

what are the assumptions we make

THE (INFLATED) OBSERVATION DATA



§ 2

Training the model

THE NAIVE BAYES CLASSIFIER

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

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

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



SETUP



CROSS-VALIDATION

Can we train the model

assumes parallel universes with slightly less data

is the model good?



NUL CLASSIFIERS

coin flip

no skill

constant

EXPECTATIONS

Model	MCC	PPV	NPV	DOR	Accuracy
noskill	0.0	0.338178	0.661822	1.0	0.552373
coinflip	-0.323643	0.338178	0.338178	0.261102	0.338178
constantpositive	0.0	0.338178	NaN	NaN	0.338178
constantnegative	0.0	NaN	0.661822	NaN	0.661822

CROSS-VALIDATION STRATEGY

k-fold

validation / training / testing

CROSS-VALIDATION RESULTS

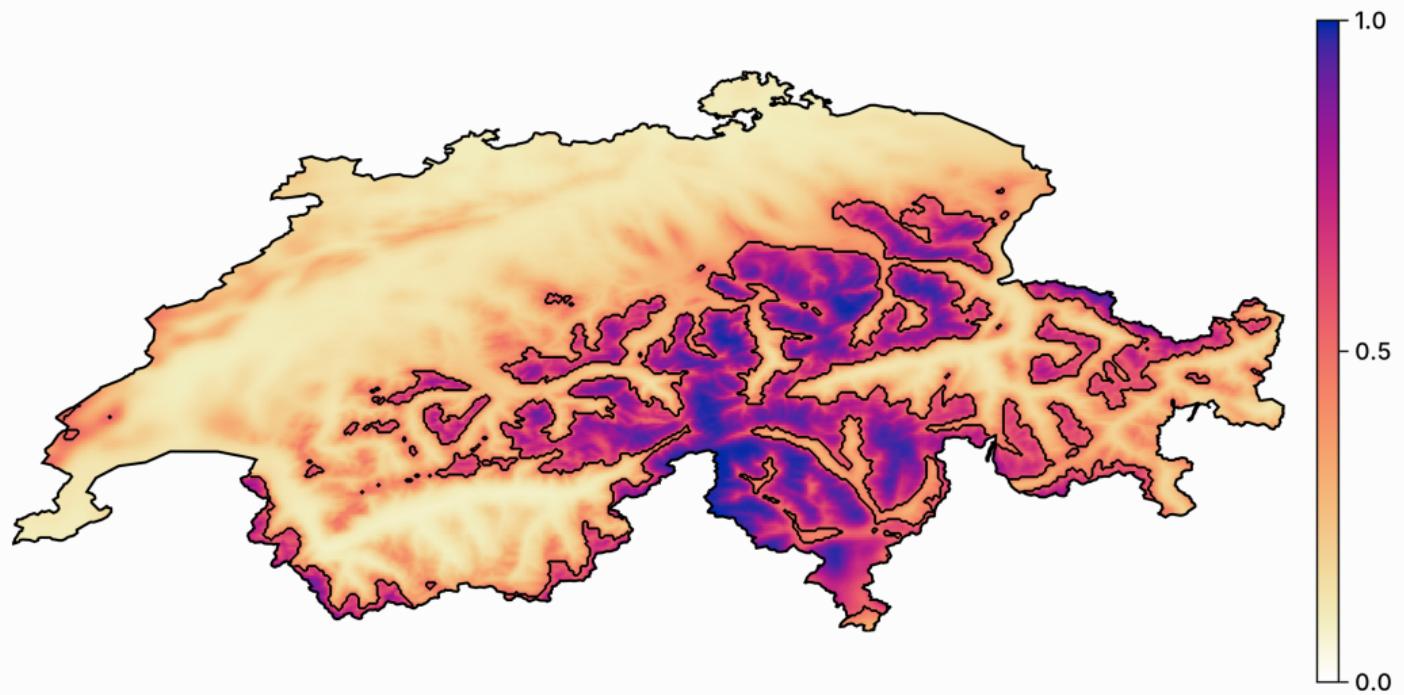
Model	MCC	PPV	NPV	DOR	Accuracy
noskill	0.0	0.338178	0.661822	1.0	0.552373
coinflip	-0.323643	0.338178	0.338178	0.261102	0.338178
constantpositive	0.0	0.338178	NaN	NaN	0.338178
constantnegative	0.0	NaN	0.661822	NaN	0.661822
Validation	0.285042	0.573912	0.739042	3.97394	0.698661
Training	0.287722	0.57633	0.73963	3.869	0.699451

WHAT TO DO IF THE MODEL IS TRAINABLE?

train it!

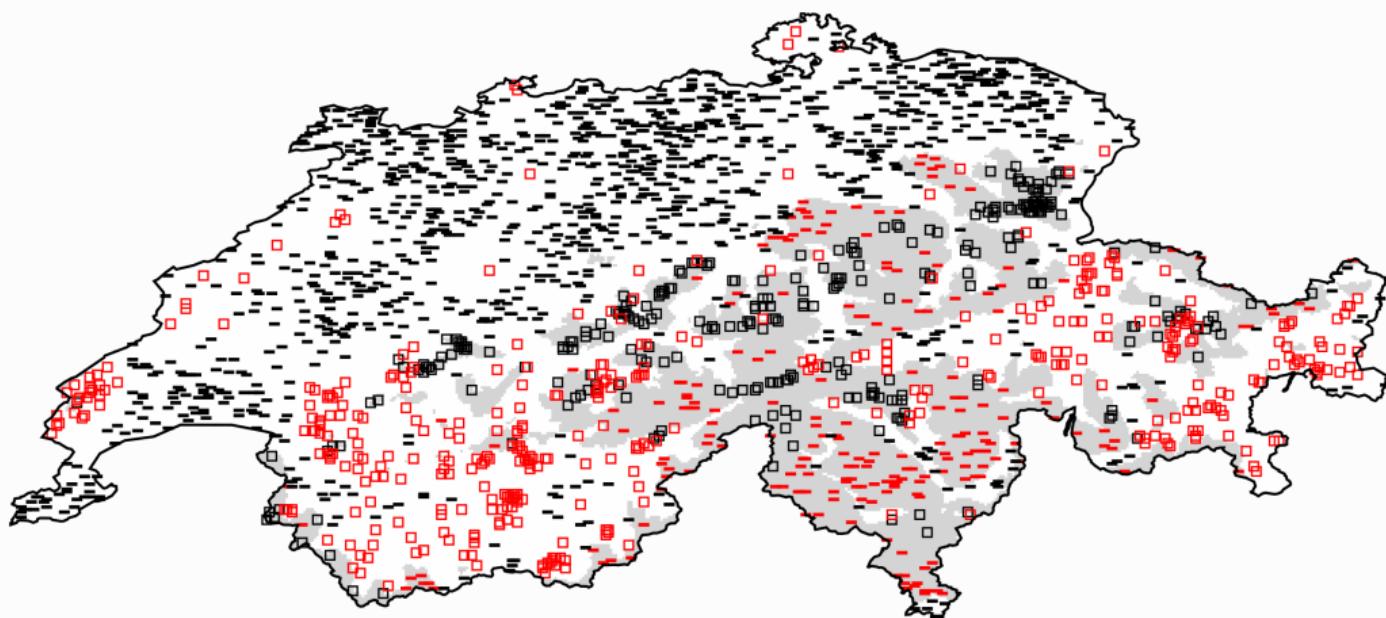
re-use the full dataset

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 THRESHOLD CLASSIFICATION

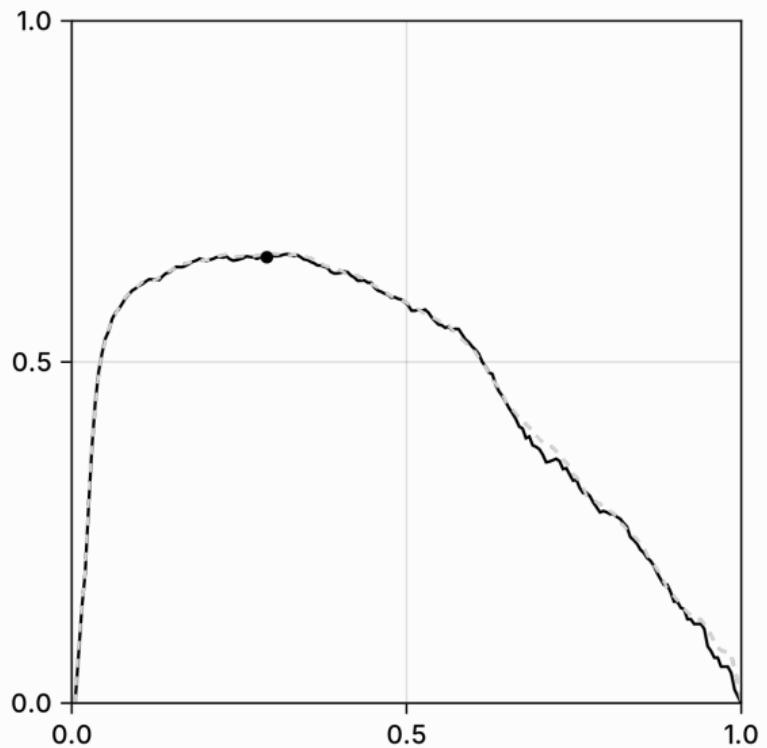
$p_{\text{plus}} > p_{\text{minus}}$ means threshold is 0.5

is it?

how do we check this

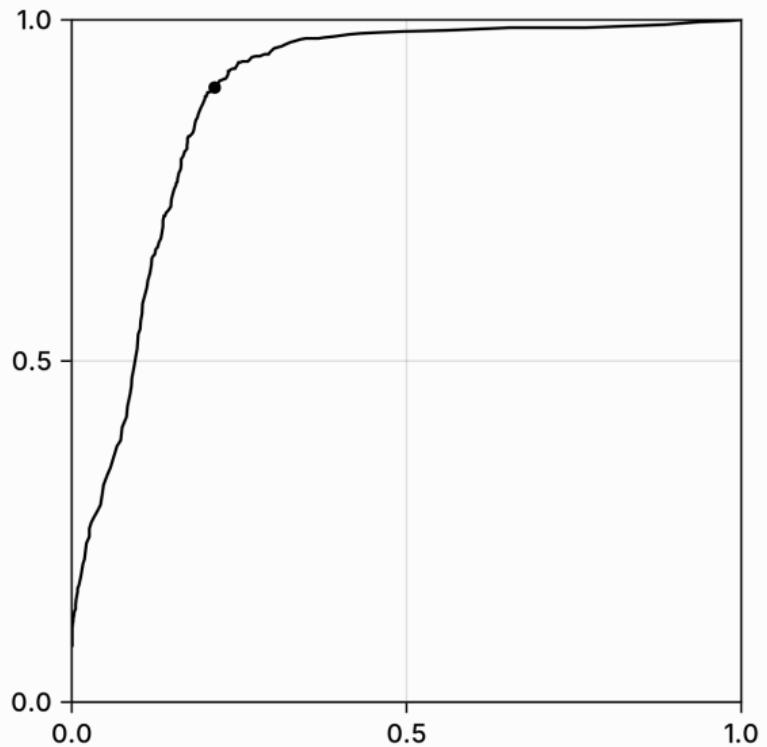


LEARNING CURVE FOR THE THRESHOLD



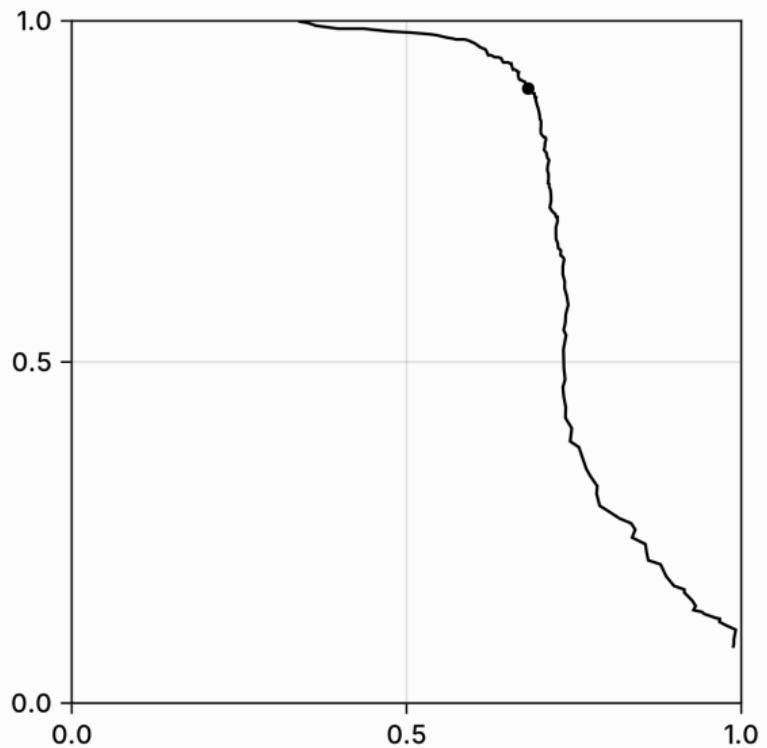


RECEIVER OPERATING CHARACTERISTIC





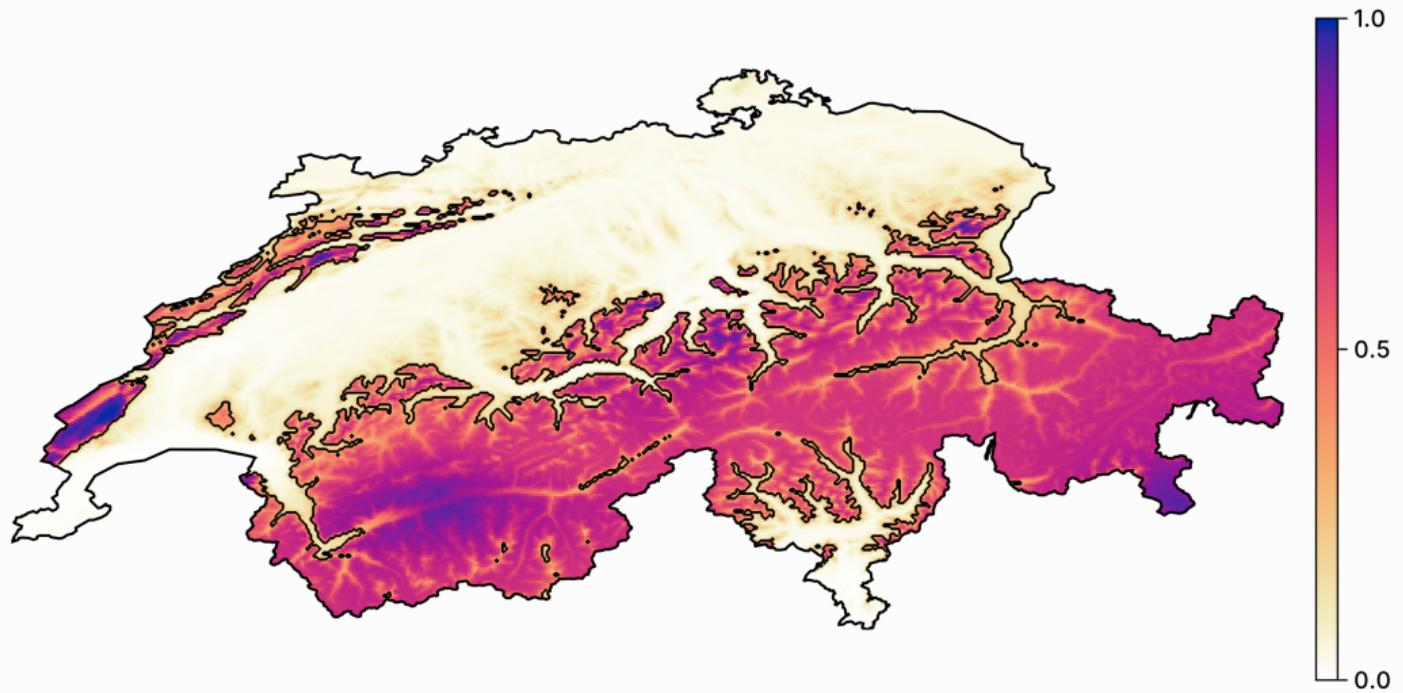
PRECISION-RECALL CURVE



REVISITING THE MODEL PERFORMANCE

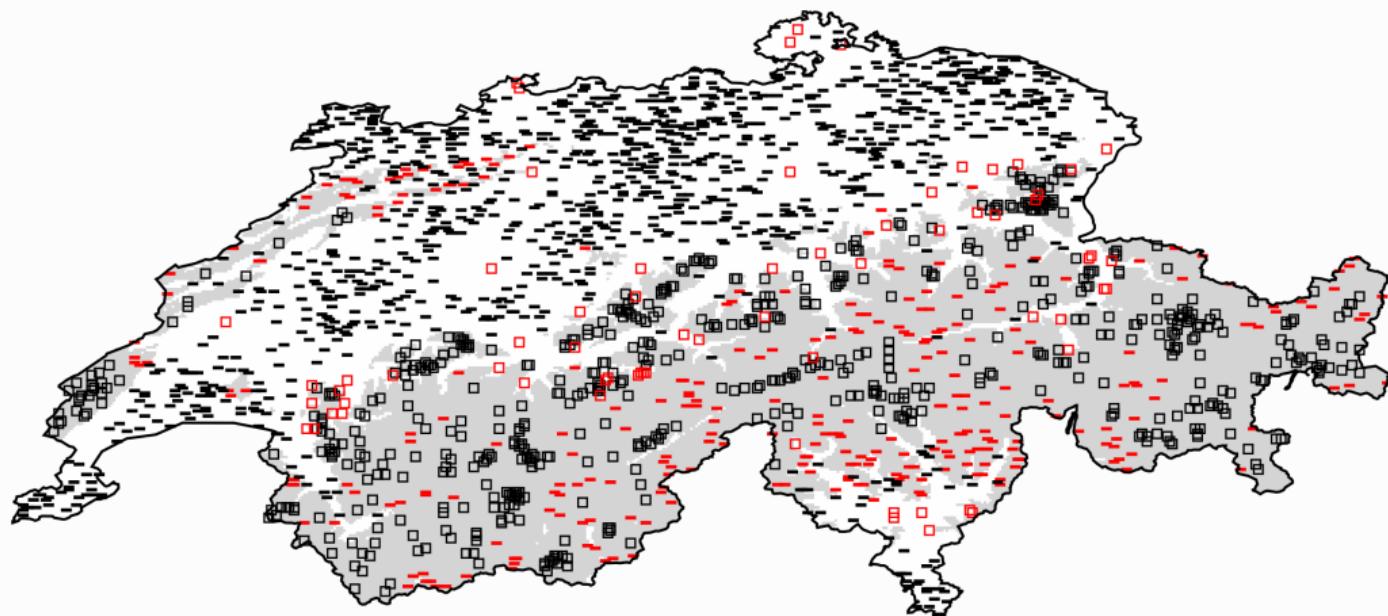
Model	MCC	PPV	NPV	DOR	Accuracy
noskill	0.0	0.338178	0.661822	1.0	0.552373
coinflip	-0.323643	0.338178	0.338178	0.261102	0.338178
constantpositive	0.0	0.338178	NaN	NaN	0.338178
constantnegative	0.0	NaN	0.661822	NaN	0.661822
Previous	0.285042	0.573912	0.739042	3.97394	0.698661
Validation	0.65469	0.682064	0.940772	38.1561	0.825594
Training	0.658932	0.686247	0.941019	34.9715	0.827627

 UPDATED PREDICTION





HOW IS THIS MODEL BETTER?





REVISITING ASSUMPTIONS

- pseudo-absences
- not just a statistical exercise



VARIABLE IMPORTANCE

BIO	Import.
11.0	0.342392
5.0	0.292015
10.0	0.254812
1.0	0.110781

§ 3

But why?



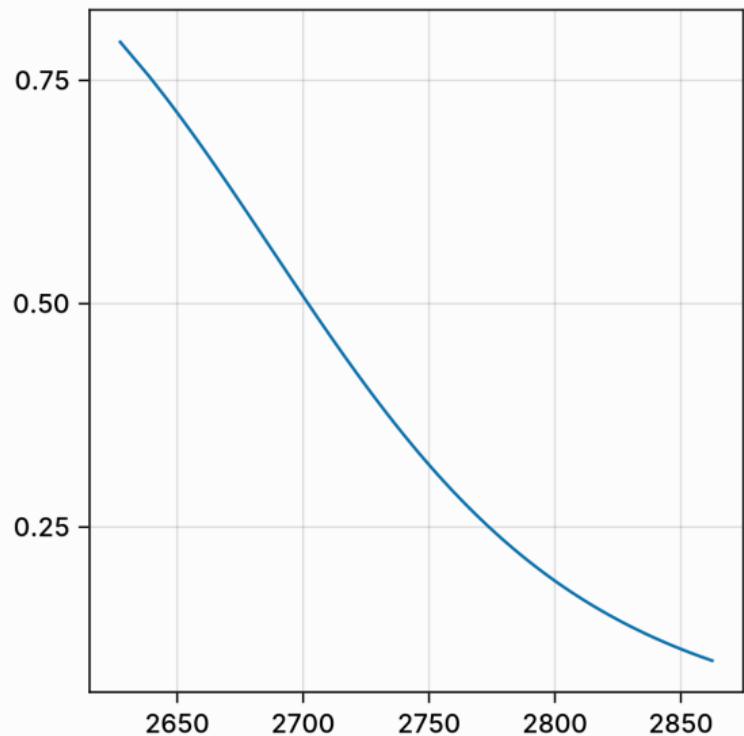


INTRO EXPLAINABLE

AN ECOLOGY TOOL: PARTIAL RESPONSE CURVES

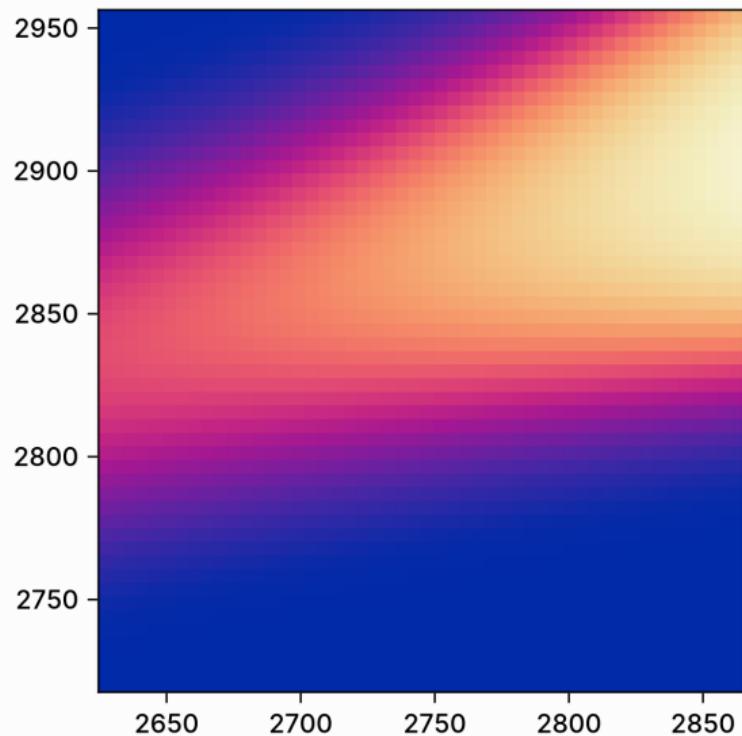


EXAMPLE WITH TEMPERATURE



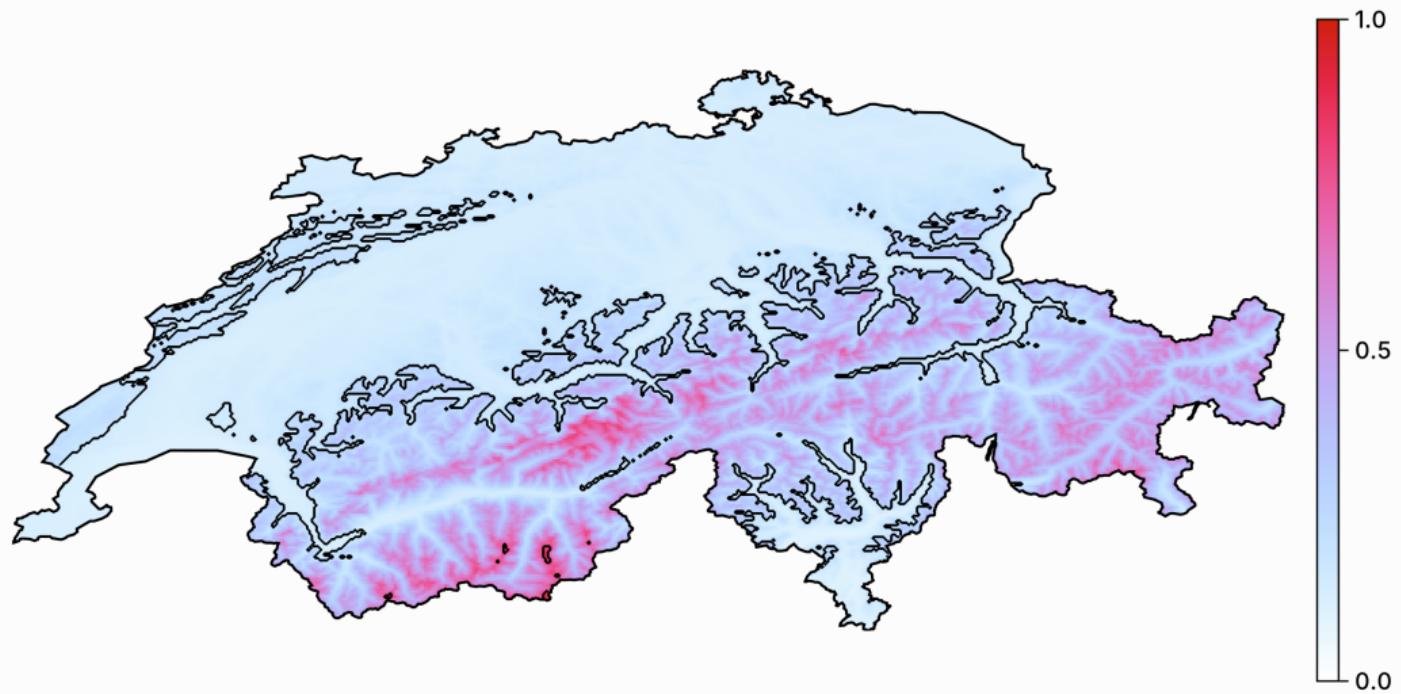


EXAMPLE WITH TWO VARIABLES



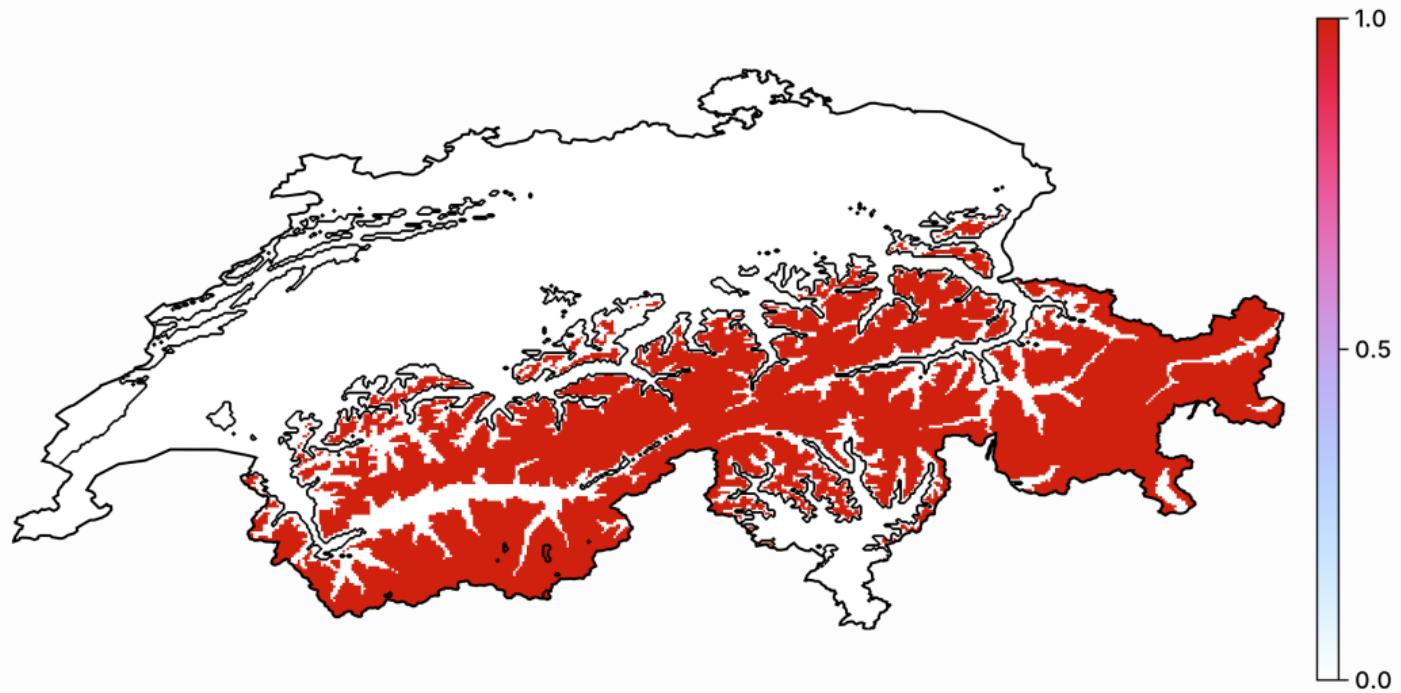


SPATIALIZED PARTIAL RESPONSE PLOT





SPATIALIZED PARTIAL RESPONSE (BINARY OUTCOME)





INFLATED RESPONSE CURVES

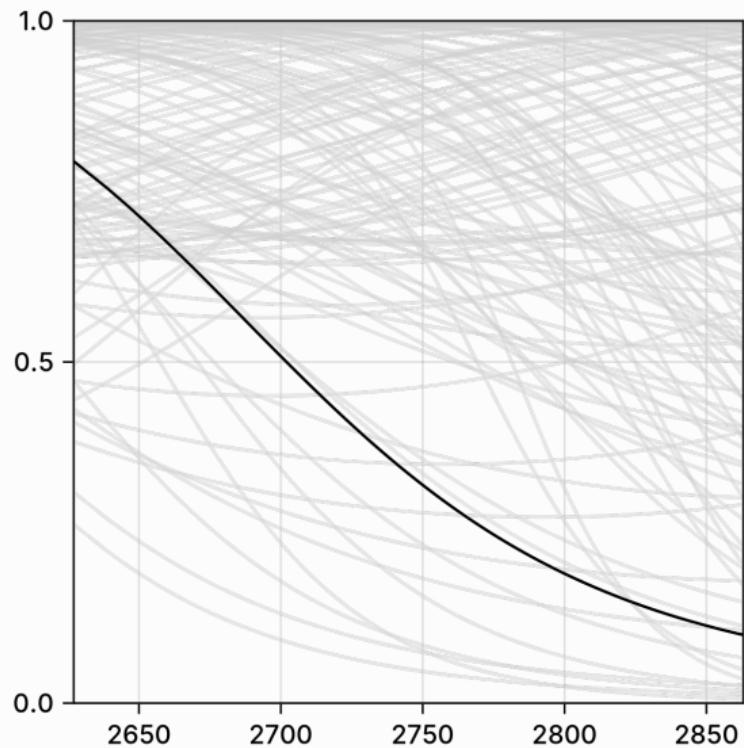
Averaging the variables is **masking a lot of variability!**

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

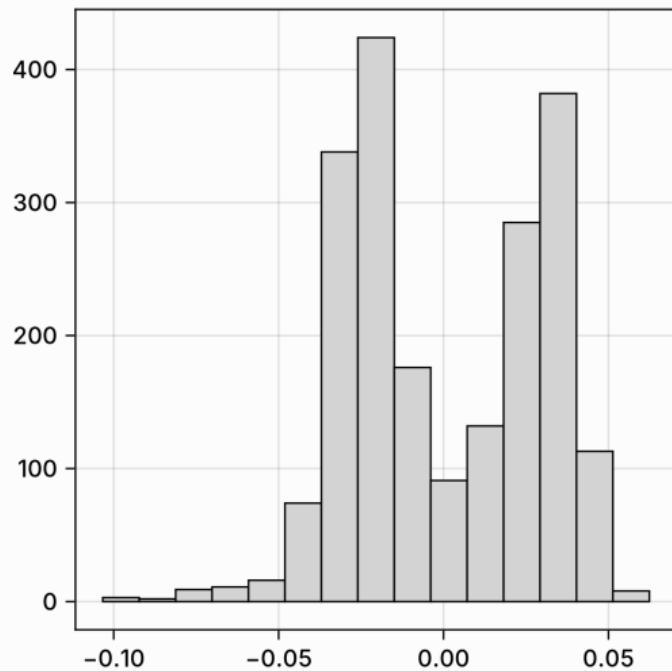
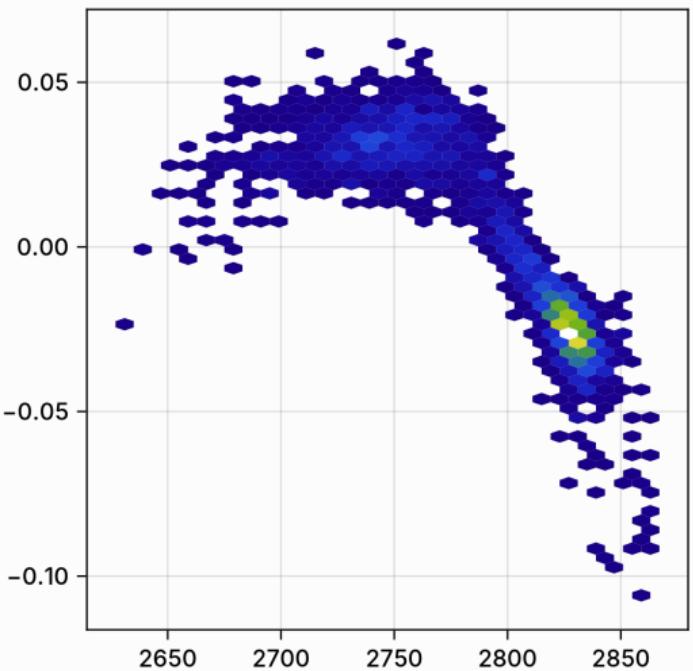




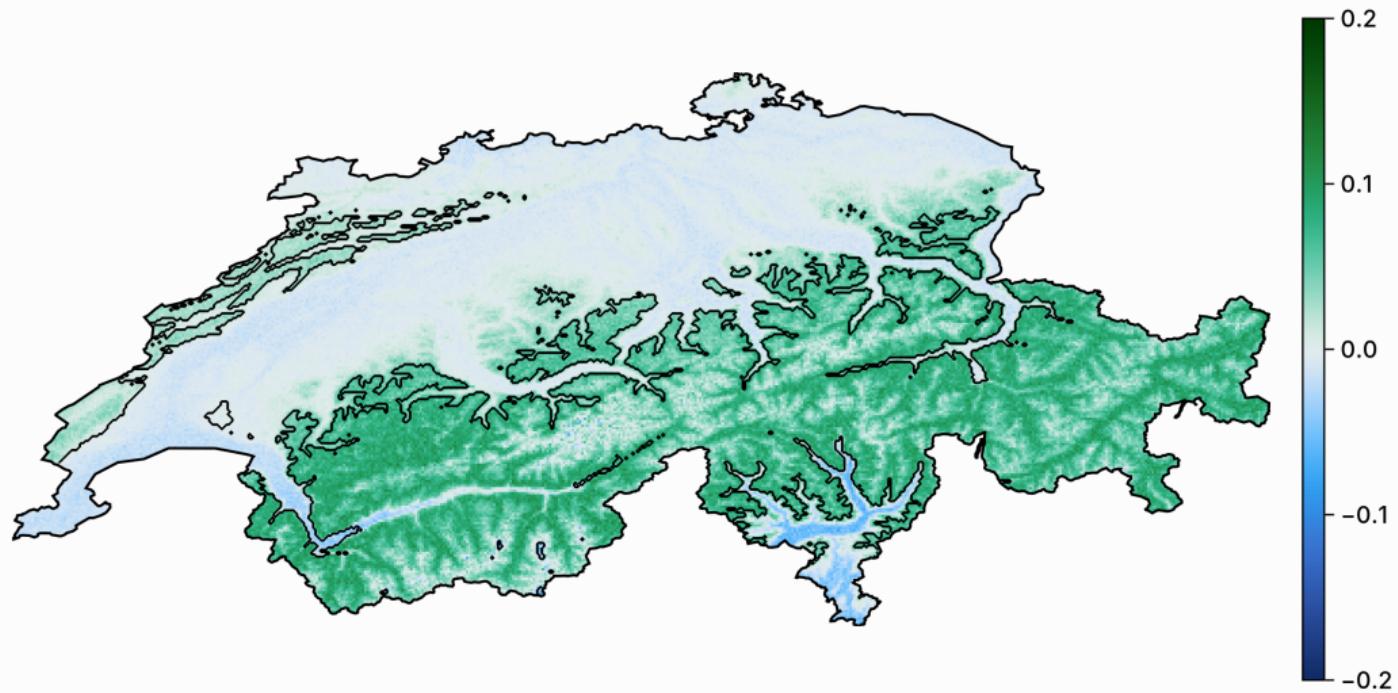
EXAMPLE



RESPONSE CURVES REVISITED



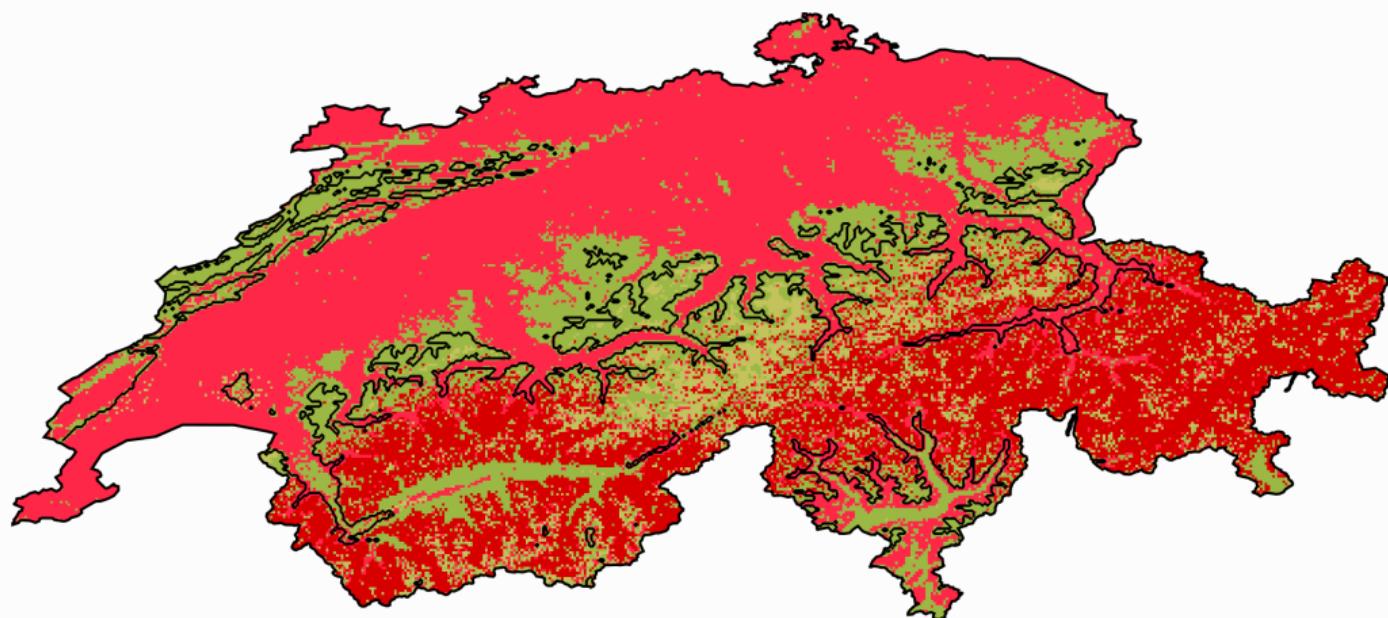
ON A MAP



VARIABLE IMPORTANCE REVISITED

BIO	Import.	Shap. imp.
5.0	0.292015	0.358287
11.0	0.342392	0.317696
1.0	0.110781	0.163154
10.0	0.254812	0.160862

 MOST IMPORTANT PREDICTOR



§ 4

What if?

INTRO TO COUNTERFACTUALS

what they are

§ 5

Ensemble models

§ 6

Conclusions



