Interpretable ML for biodiversity

An introduction using species distribution models

Timothée Poisot Université de Montréal

October 3, 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?

BUT WHY...

... think of SDM as ML problems? 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*

WHAT WE WILL *NOT* DISCUSS

1. Image recognition
2. Sound recognition
3. Generative AI

LEARNING/TEACHING GOALS

- ML basics
  - cross-validation
  - hyper-parameters tuning
  - bagging and ensembles
- Pitfalls
  - data leakage
  - overfitting
- Explainable ML
  - partial responses
  - Shapley values
- Counterfactuals

§ 1

Problem statement

THE PROBLEM IN ECOLOGICAL TERMS

We have information about a species, taking the form of (lon, lat) for points where the species was observed

Using this information, we can extract a suite of environmental variables for the locations where the species was observed

We can do the same thing for locations where the species was not observed

Where could we observe this species?

THE PROBLEM IN ML TERMS

We have a series of labels $\mathbf{y}_n \in \mathbb{B}$, and features $\mathbf{X}_{m,n} \in \mathbb{R}$

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

An algorithm that does this job well is generalizable (we can apply it on data it has not been trained on) and makes credible predictions

SETTING UP THE DATA FOR OUR EXAMPLE

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

Two series of environmental layers

1. CHELSA2 BioClim variables (19)
2. EarthEnv land cover variables (12)

Now is *not* the time to make assumptions about which are relevant!

THE OBSERVATION DATA

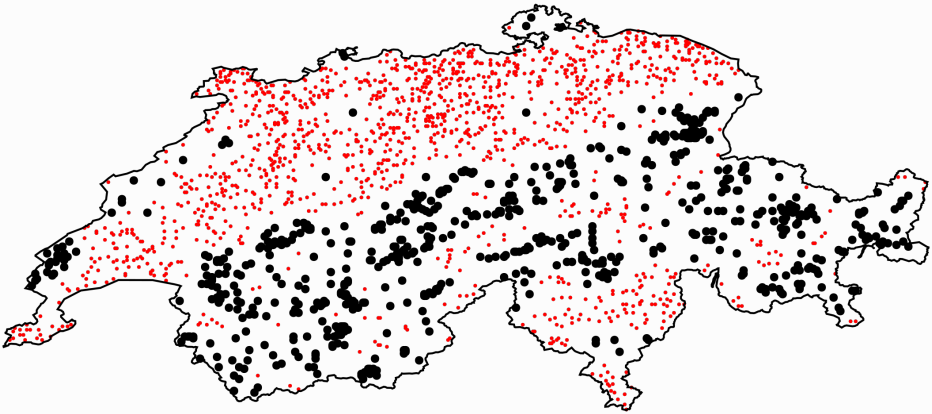
PROBLEM (AND SOLUTION)

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

We generate pseudo-absences with the following rules:

1. Locations further away from a presence are more likely
2. Locations less than 5km away from a presence are ruled out

THE (INFLATED) OBSERVATION DATA



§ 2

Training the model

A SIMPLE DECISION TREE

SETUP

Can we train the model?

More specifically – if we train the model, how well can we expect it to perform?

assumes parallel universes with slightly less data

is the model good?

coin flip

no skill

constant

The null classifiers tell us what we need to beat in order to perform better than random.

Model	MCC	PPV	NPV	DOR	Accuracy
No skill	-0.00	0.34	0.66	1.00	0.55
Coin flip	-0.32	0.34	0.34	0.26	0.34
+	0.00	0.34			0.34
-	0.00		0.66		0.66

In practice, the no-skill classifier is the most informative: what if we only know the positive class prevalence?

CROSS-VALIDATION STRATEGY

k-fold

validation / training / testing

CROSS-VALIDATION RESULTS

Model	MCC	PPV	NPV	DOR	Accuracy
No skill	-0.00	0.34	0.66	1.00	0.55
Dec. tree (val.)	0.64	0.77	0.87	26.59	0.84
Dec. tree (tr.)	0.66	0.78	0.88	28.70	0.85

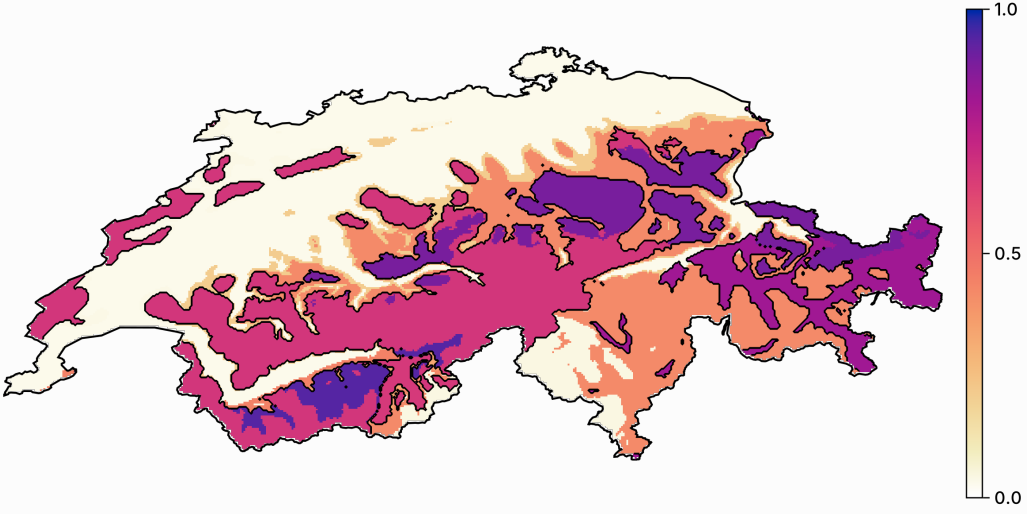
WHAT TO DO IF THE MODEL IS TRAINABLE?

train it!

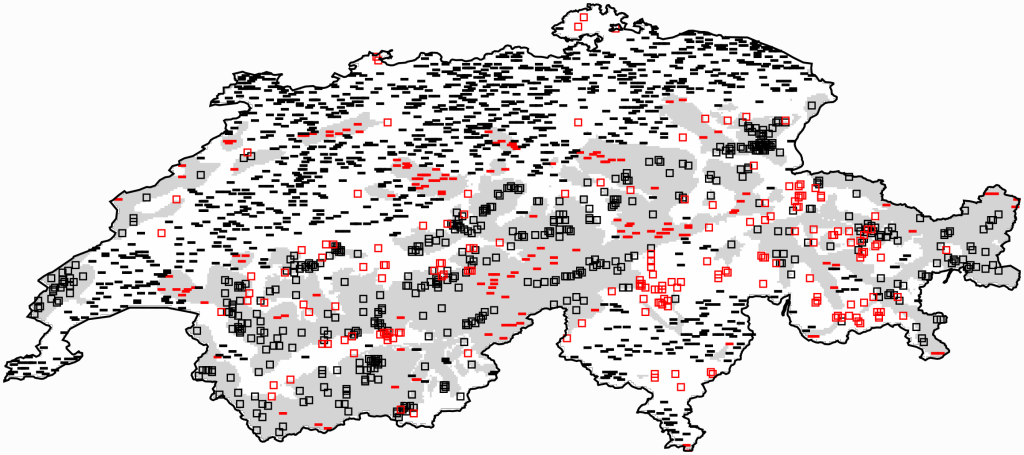
re-use the full dataset

THE MODEL TRAINING PIPELINE

INITIAL PREDICTION



HOW IS THIS MODEL WRONG?



CAN WE IMPROVE ON THIS MODEL?

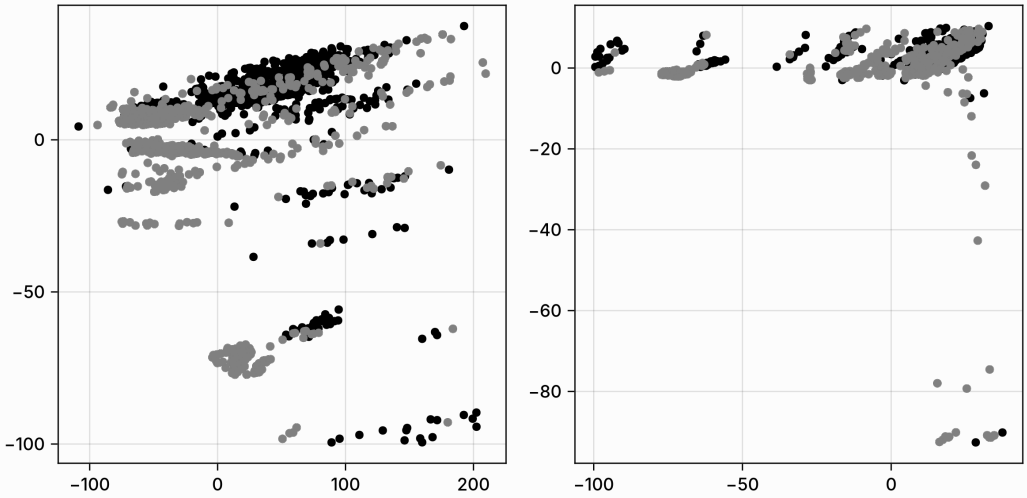
variable selection

data transformation

hyper-parameters tuning

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

A NOTE ON PCA



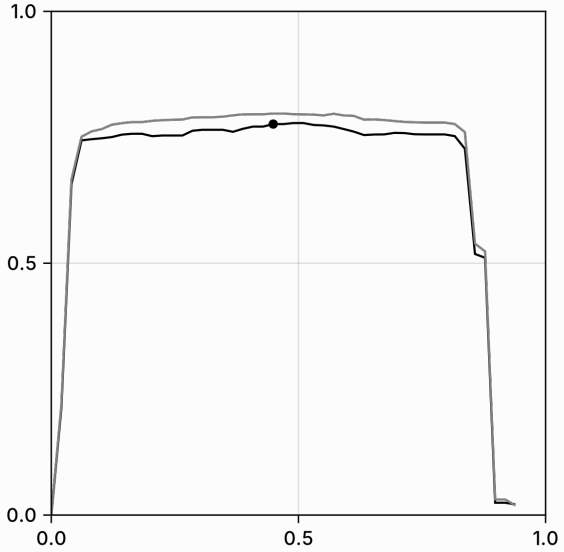
MOVING THESHOLD CLASSIFICATION

p plus > p minus means threshold is 0.5

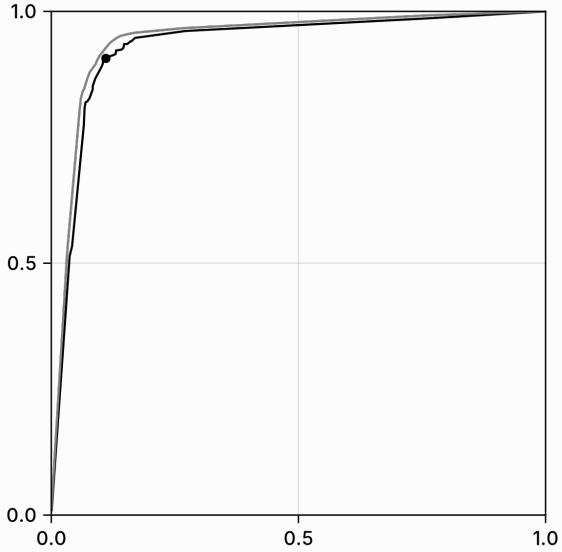
is it?

how do we check this

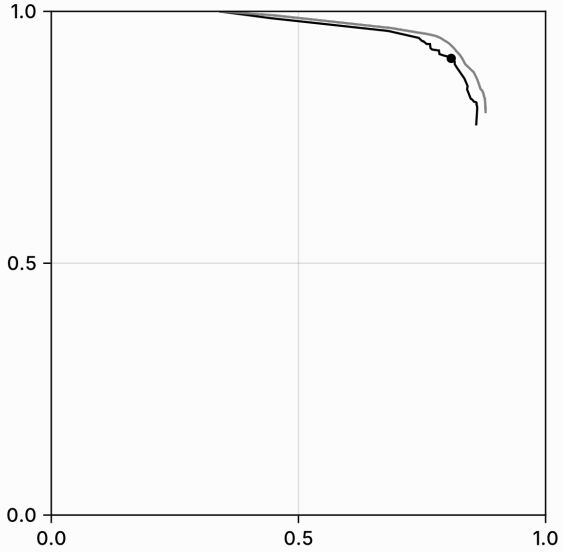
LEARNING CURVE FOR THE THRESHOLD



RECEIVER OPERATING CHARACTERISTIC



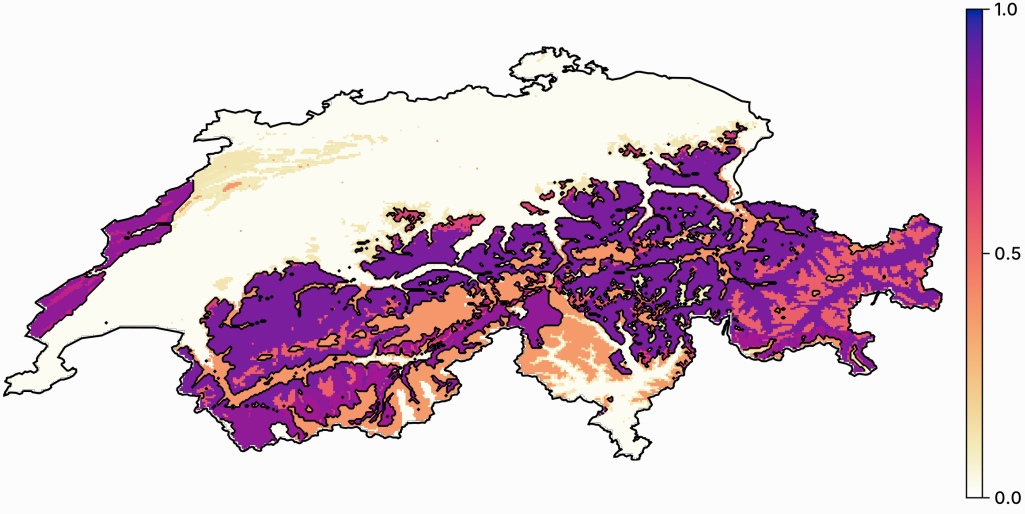
PRECISION-RECALL CURVE



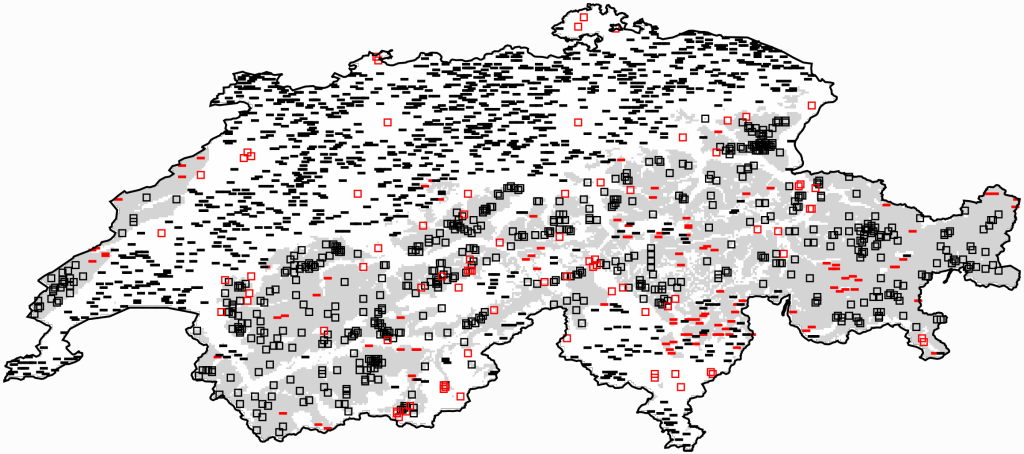
REVISITING THE MODEL PERFORMANCE

Model	MCC	PPV	NPV	DOR	Accuracy
No skill	-0.00	0.34	0.66	1.00	0.55
Dec. tree (val.)	0.64	0.77	0.87	26.59	0.84
Dec. tree (tr.)	0.66	0.78	0.88	28.70	0.85
Tuned tree (val.)	0.77	0.79	0.95	113.44	0.89
Tuned tree (tr.)	0.80	0.81	0.96	114.37	0.90

UPDATED PREDICTION

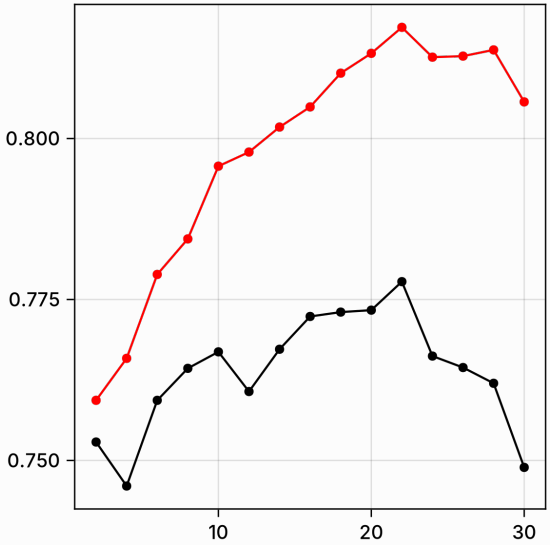


HOW IS THIS MODEL BETTER?



BUT WAIT!

Using depth of 7 and up to 20 nodes



§ 3

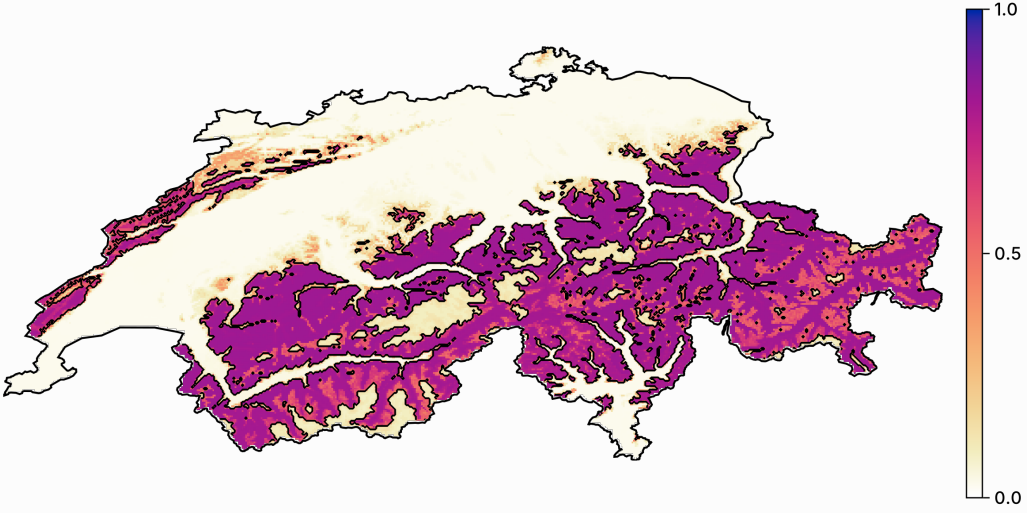
Ensemble models

LIMITS OF A SINGLE MODEL

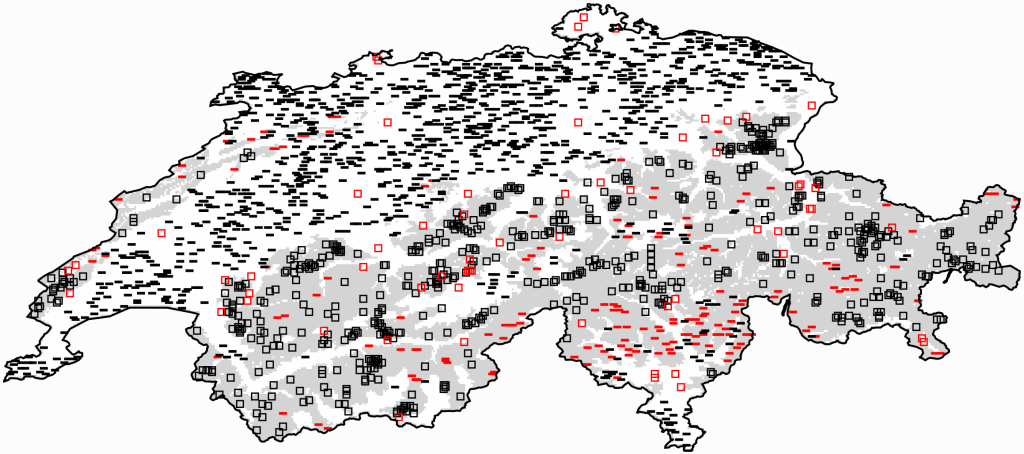
- a single model
- different parts of data may have different signal
- do we need all the variables all the time?
- bias v. variance tradeoff
- limit overfitting

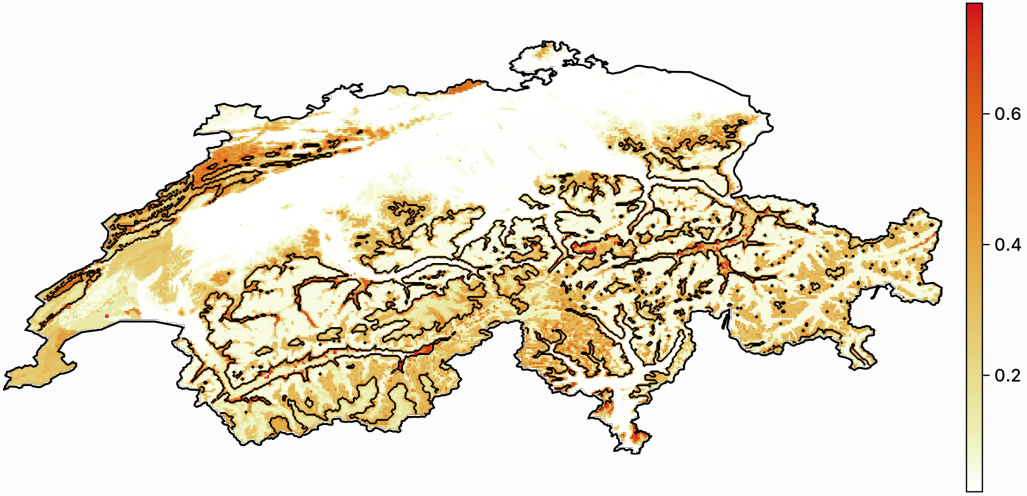
Model	MCC	PPV	NPV	DOR	Accuracy
No skill	-0.00	0.34	0.66	1.00	0.55
Dec. tree (val.)	0.64	0.77	0.87	26.59	0.84
Dec. tree (tr.)	0.66	0.78	0.88	28.70	0.85
Tuned tree (val.)	0.77	0.79	0.95	113.44	0.89
Tuned tree (tr.)	0.80	0.81	0.96	114.37	0.90
Forest (val.)	0.76	0.79	0.95	103.14	0.89
Forest (tr.)	0.77	0.79	0.95	73.86	0.89

PREDICTION OF THE ROTATION FOREST



PREDICTION OF THE ROTATION FOREST




0.6
0.4
0.2

REVISITING ASSUMPTIONS

- pseudo-absences
- not just a statistical exercise

VARIABLE IMPORTANCE

Layer	Variable	Import.
1	BIO1	0.910314
8	BIO8	0.0462913
29	Snow/Ice	0.0209557
24	Shrubs	0.018254
3	BIO3	0.00418499

§ 4

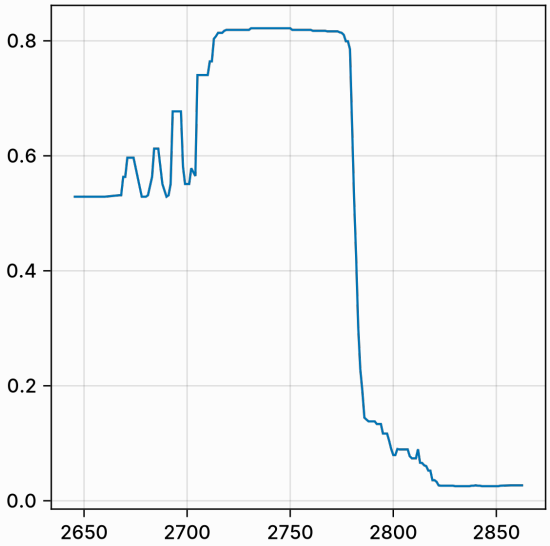
But why?

PARTIAL RESPONSE CURVES

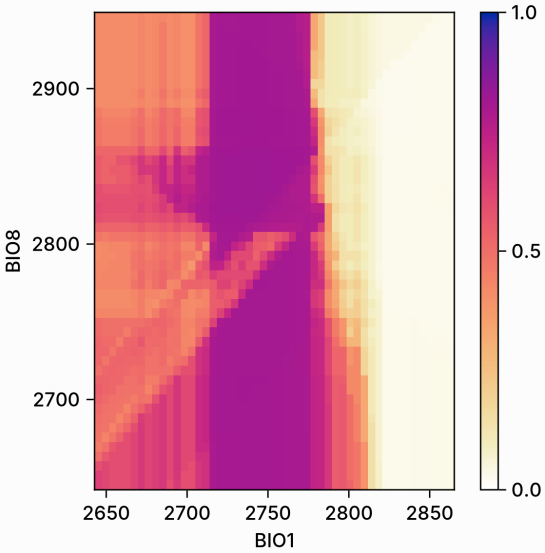
If we assume that all the variables except one take their average value, what is the prediction associated to the value that is unchanged?

Equivalent to a mean-field approximation

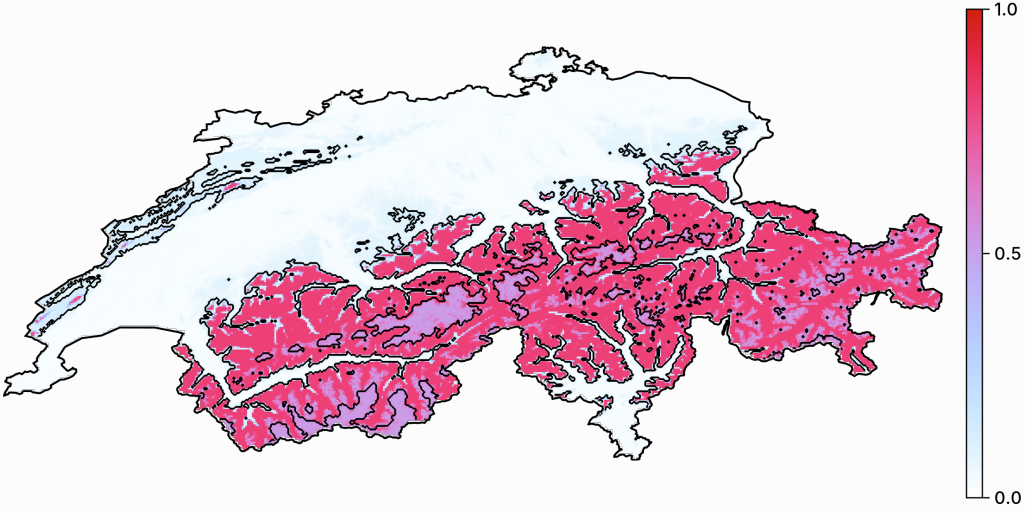
EXAMPLE WITH TEMPERATURE



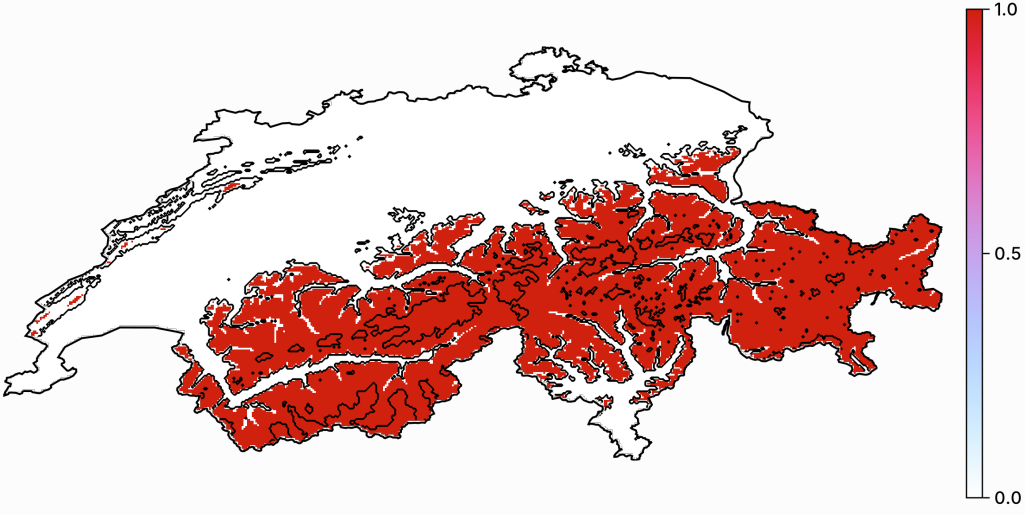
EXAMPLE WITH TWO VARIABLES



SPATIALIZED PARTIAL RESPONSE PLOT



SPATIALIZED PARTIAL RESPONSE (BINARY OUTCOME)



INFLATED RESPONSE CURVES

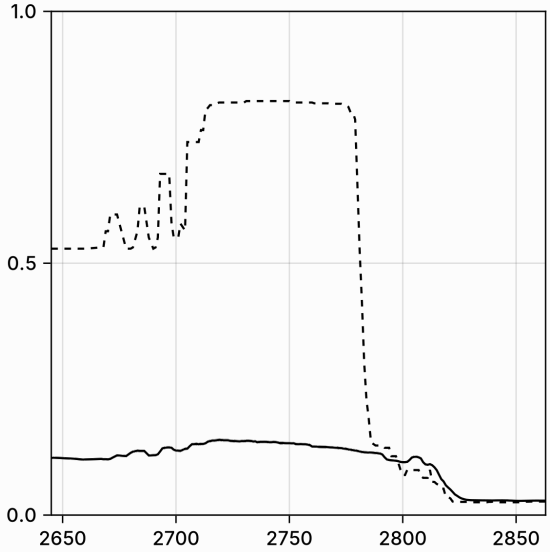
Averaging the variables is masking a lot of variability!

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



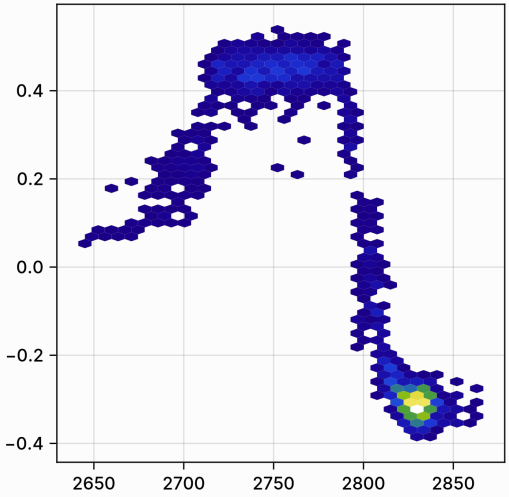
LIMITATIONS

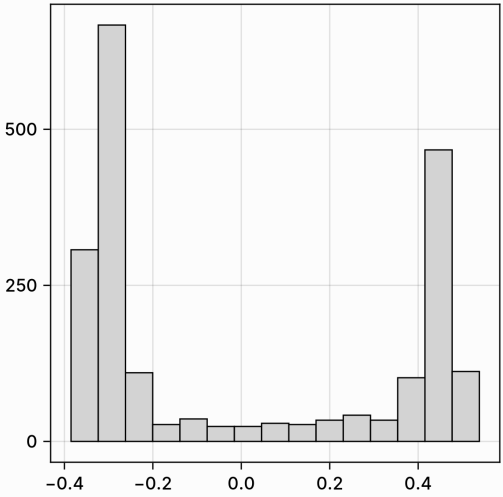
- partial responses can only generate model-level information
- they break the structure of values for all predictors at the scale of a single observation
- their interpretation is unclear

SHAPLEY

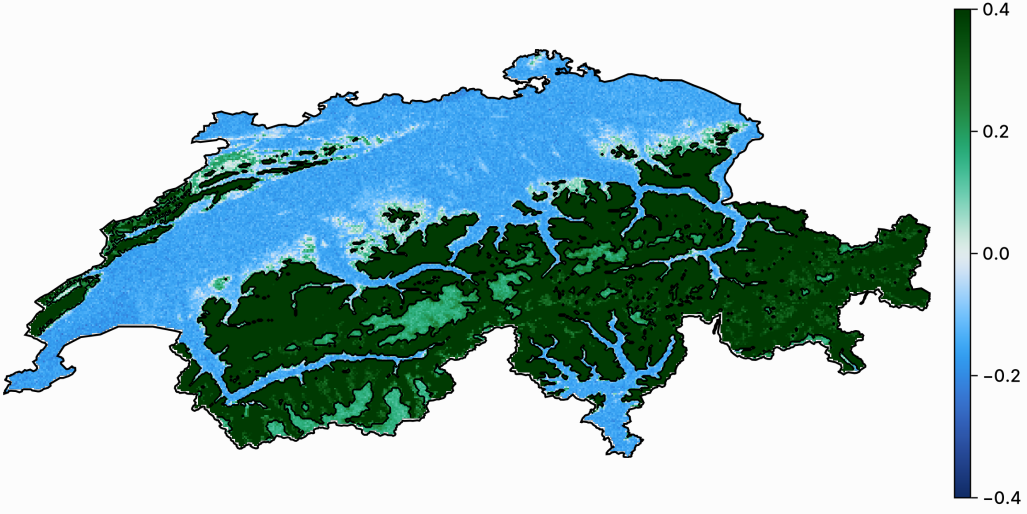
EXAMPLE

RESPONSE CURVES REVISITED





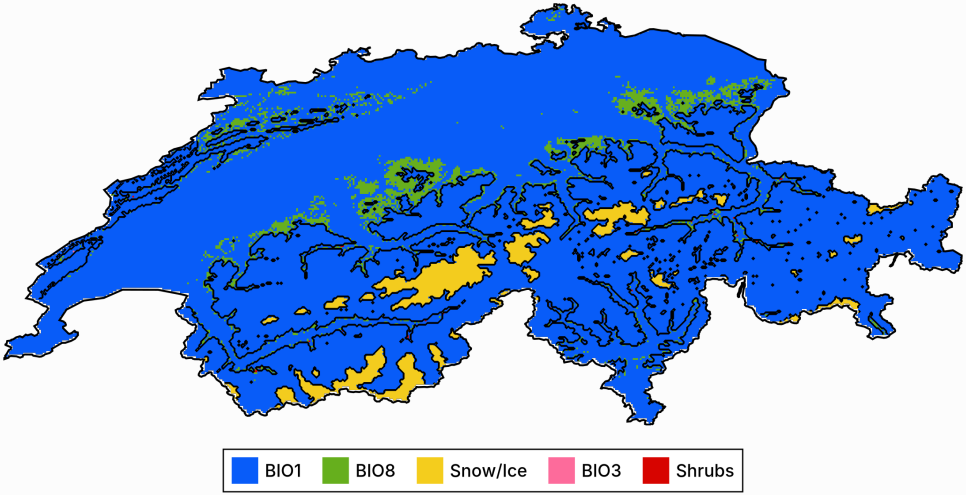
ON A MAP



VARIABLE IMPORTANCE REVISITED

Layer	Variable	Import.	Shap. imp.
1	BIO1	0.910314	0.846428
8	BIO8	0.0462913	0.0735605
29	Snow/Ice	0.0209557	0.0500842
24	Shrubs	0.018254	0.0193275
3	BIO3	0.00418499	0.0105997

MOST IMPORTANT PREDICTOR



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

§ 5

What if?

what they are

THE RASHOMON EFFECT

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

GENERATING A COUNTERFACTUAL

learning rate and loss function

use on prediction score and not yes/no!

§ 6

Conclusions