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*

LEARNING/TEACHING GOALS

1. Validation and training
2. Data leakage and data transformation
3. Overfitting and bagging
4. Partial responses and Shapley values
5. 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 observed 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)

THE OBSERVATION DATA

PROBLEM!

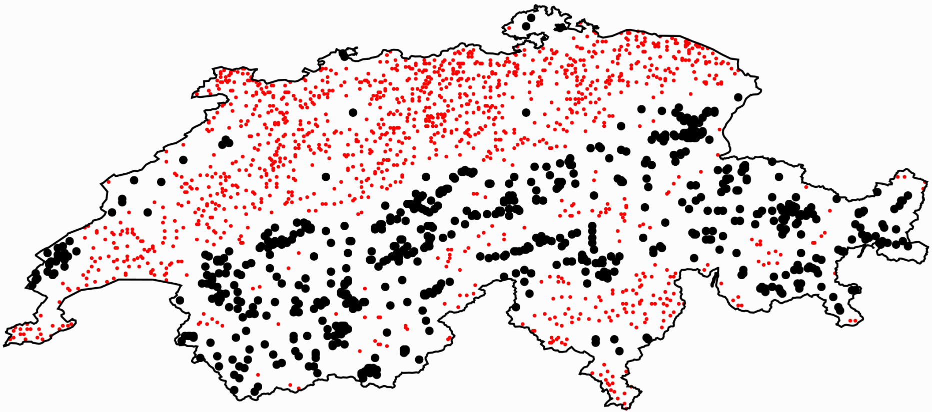
We want $\hat{y} \in \mathbb{B}$, and so far we are missing negative values

SOLUTION!

pseudo-absences

what are the assumptions we make

THE (INFLATED) OBSERVATION DATA



§ 2

Training the model

A SIMPLE DECISION TREE

SETUP

Can we train the model

assumes parallel universes with slightly less data

is the model good?

NULL CLASSIFIERS

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

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
Coin flip	-0.32	0.34	0.34	0.26	0.34
+	0.00	0.34			0.34
-	0.00		0.66		0.66
Validation	0.63	0.74	0.89	25.41	0.83
Training	0.66	0.76	0.90	33.66	0.84

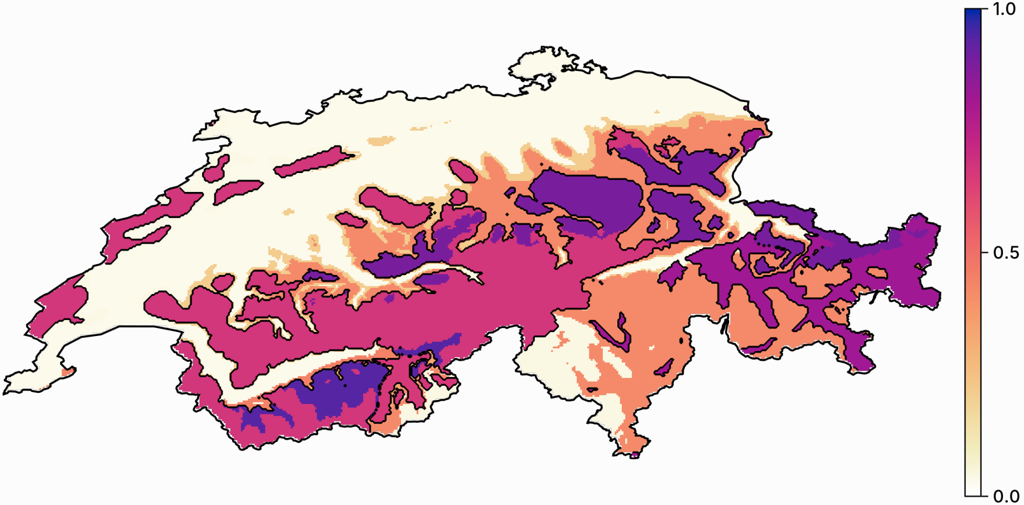
WHAT TO DO IF THE MODEL IS TRAINABLE?

train it!

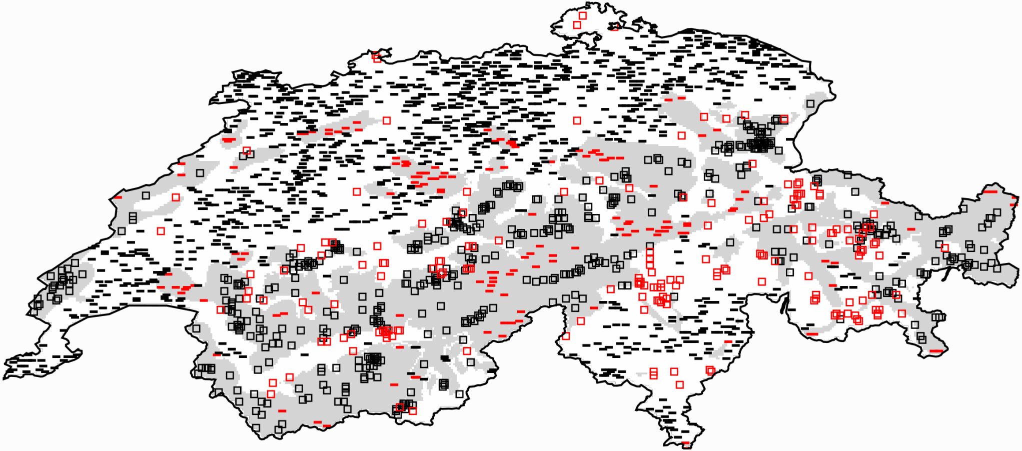
re-use the full dataset

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)

DATA LEAKAGE

A NOTE ON PCA

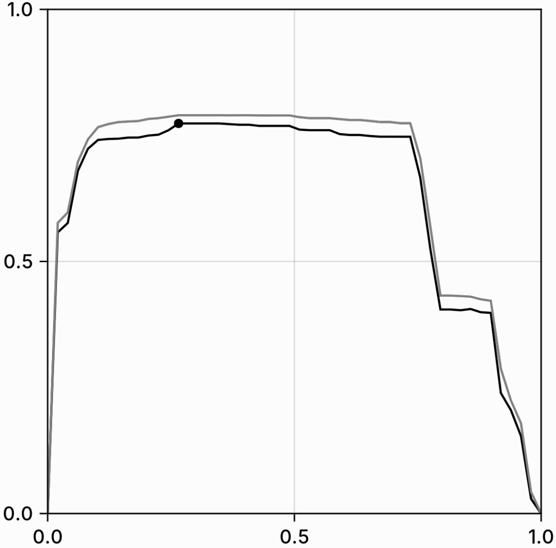
MOVING THESHOLD CLASSIFICATION

p plus > p minus means threshold is 0.5

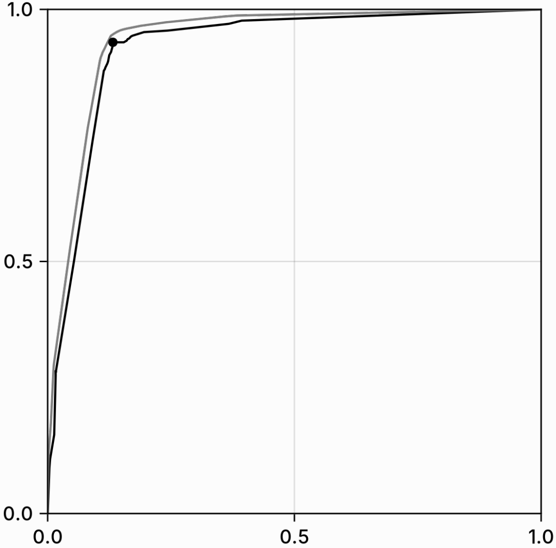
is it?

how do we check this

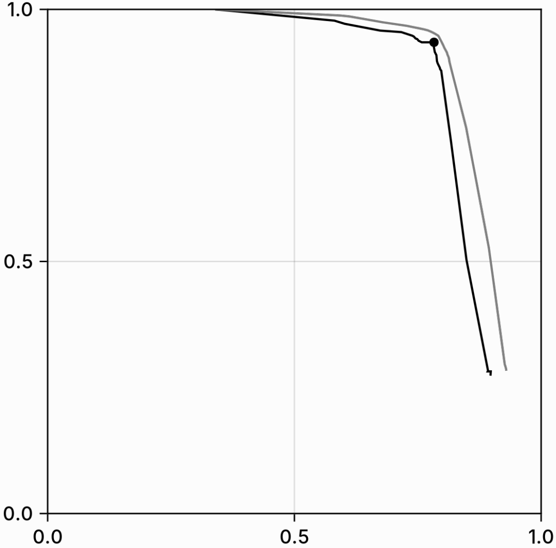
LEARNING CURVE FOR THE THRESHOLD



RECEIVER OPERATING CHARACTERISTIC



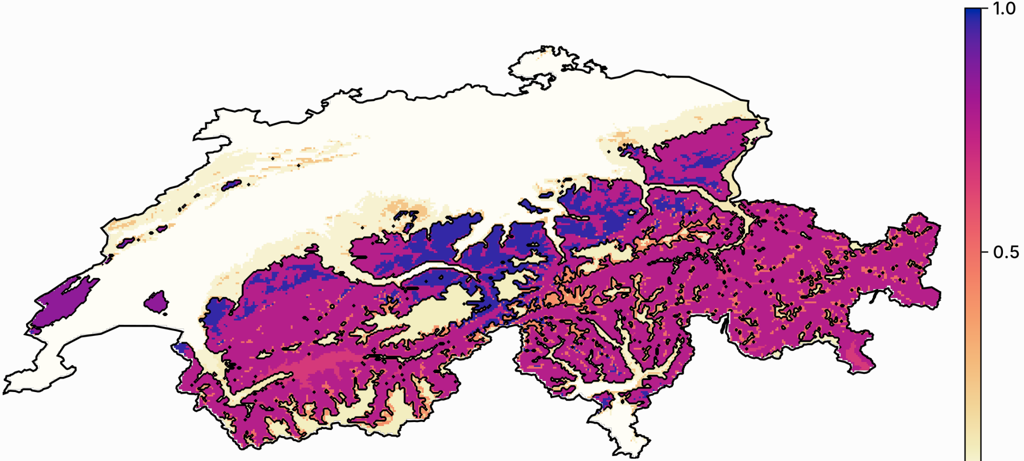
PRECISION-RECALL CURVE



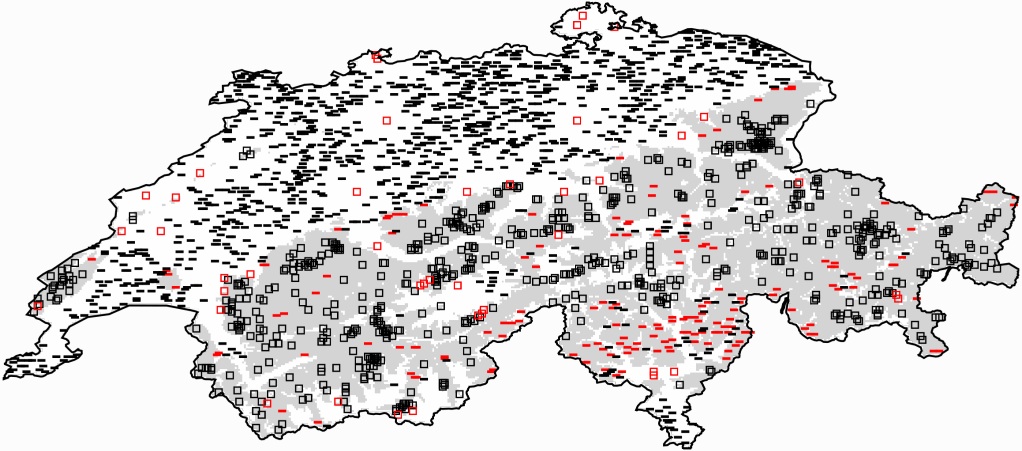
REVISITING THE MODEL PERFORMANCE

Model	MCC	PPV	NPV	DOR	Accuracy
No skill	-0.00	0.34	0.66	1.00	0.55
Coin flip	-0.32	0.34	0.34	0.26	0.34
+	0.00	0.34			0.34
-	0.00		0.66		0.66
Validation	0.63	0.74	0.89	25.41	0.83
Training	0.66	0.76	0.90	33.66	0.84
Validation	0.77	0.78	0.96	140.95	0.89
Training	0.79	0.79	0.97	125.70	0.90

UPDATED PREDICTION



HOW IS THIS MODEL BETTER?



BUT WAIT!

slide on overfitting

§ 3

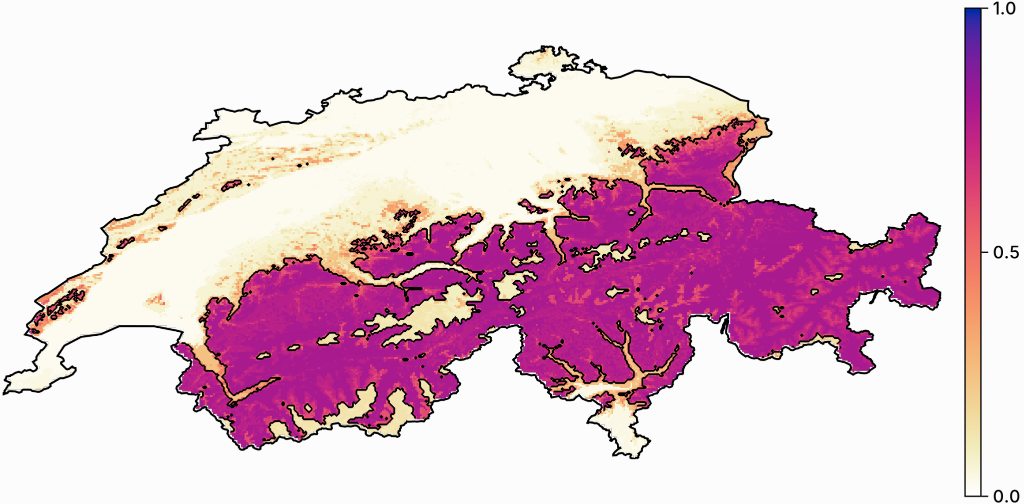
Ensemble models

LIMITS OF A SINGLE MODEL

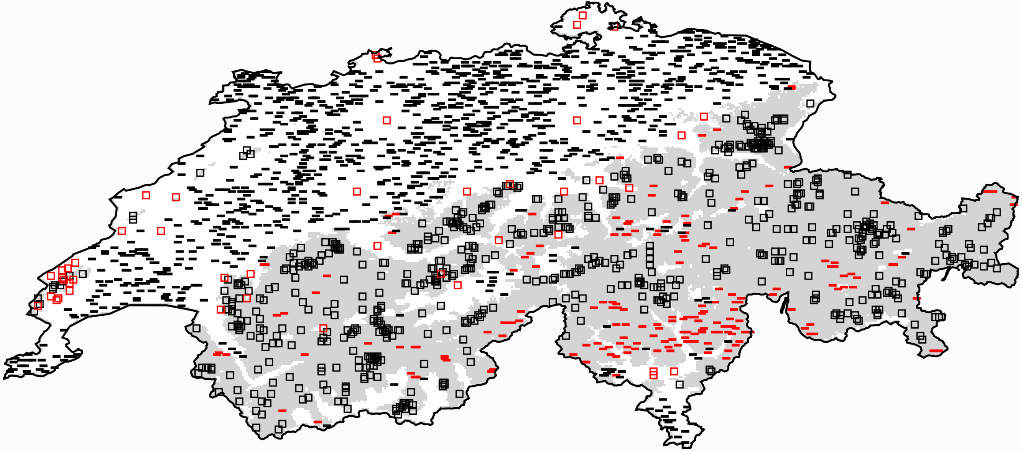
- a single model
- different parts of data may have different signal
- do we need all the variables all the time?
- bias v. variance tradeoff
- limit overfitting

BOOTSTRAPPING AND AGGREGATION

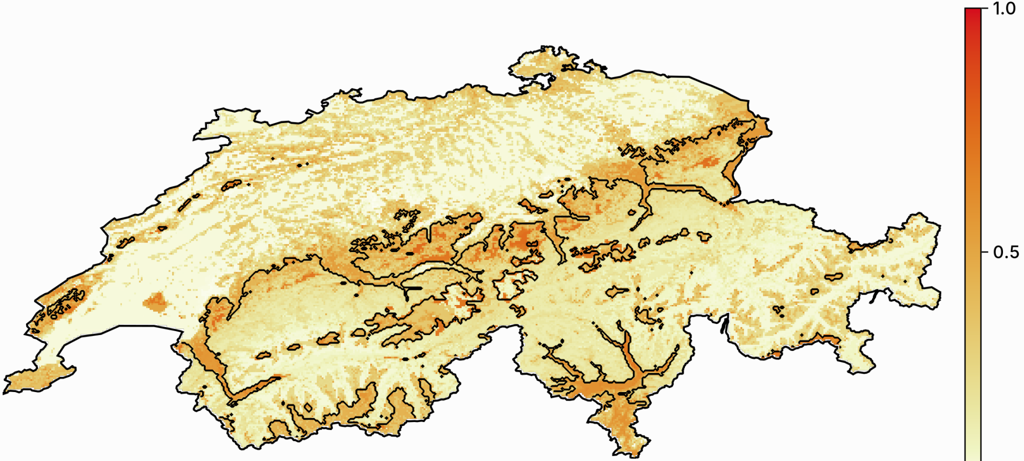
PREDICTION OF THE ROTATION FOREST



PREDICTION OF THE ROTATION FOREST



UNCERTAINTY



REVISITING ASSUMPTIONS

- pseudo-absences
- not just a statistical exercise

VARIABLE IMPORTANCE

Layer	Variable	Import.
6	BIO6	0.805809
7	BIO7	0.151936
23	Mixed/Other Trees	0.0225518
29	Snow/Ice	0.0192739
27	Regularly Flooded Vegetation	0.000429548

§ 4

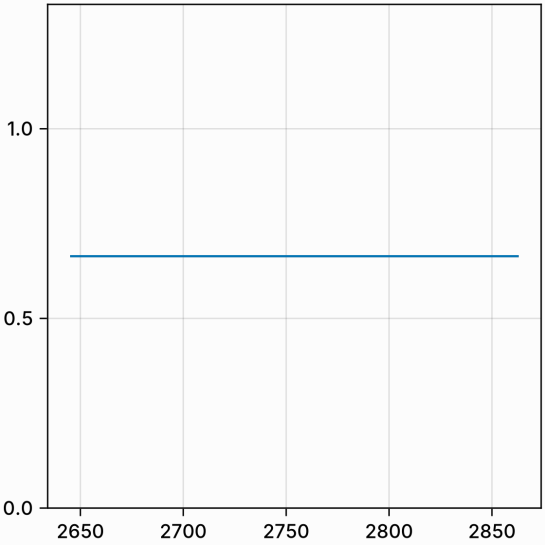
But why?

PARTIAL RESPONSE CURVES

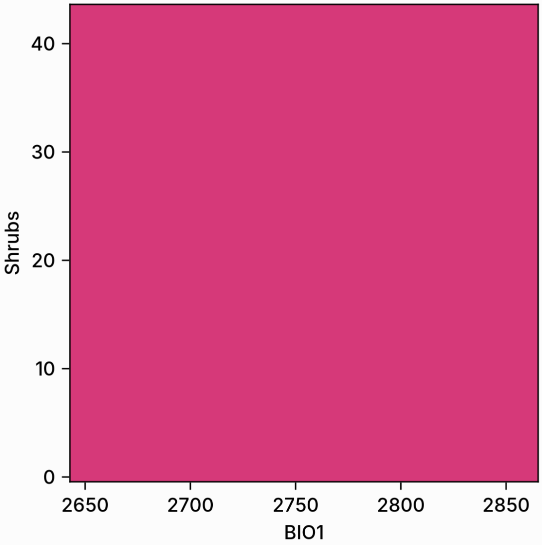
If we assume that all the variables except one take their average value, what is the prediction associated to the value that is unchanged?

Equivalent to a mean-field approximation

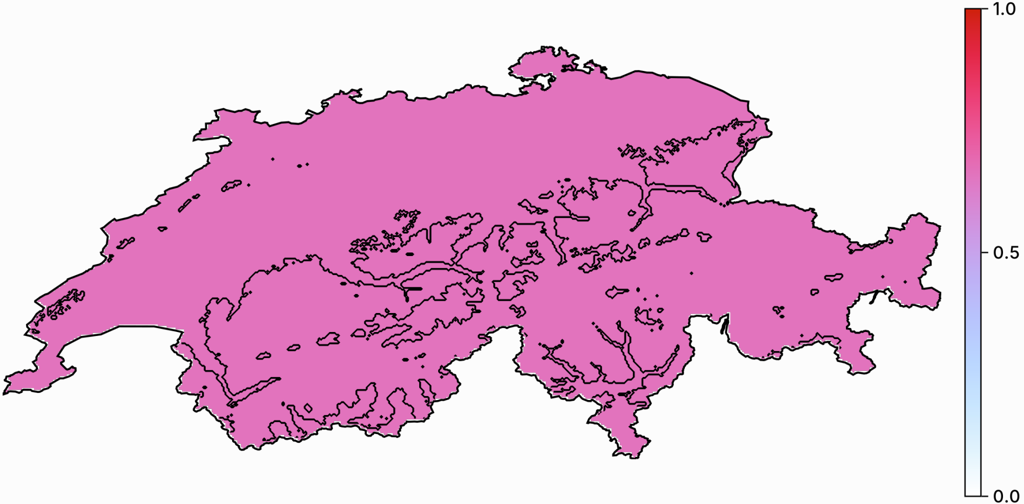
EXAMPLE WITH TEMPERATURE



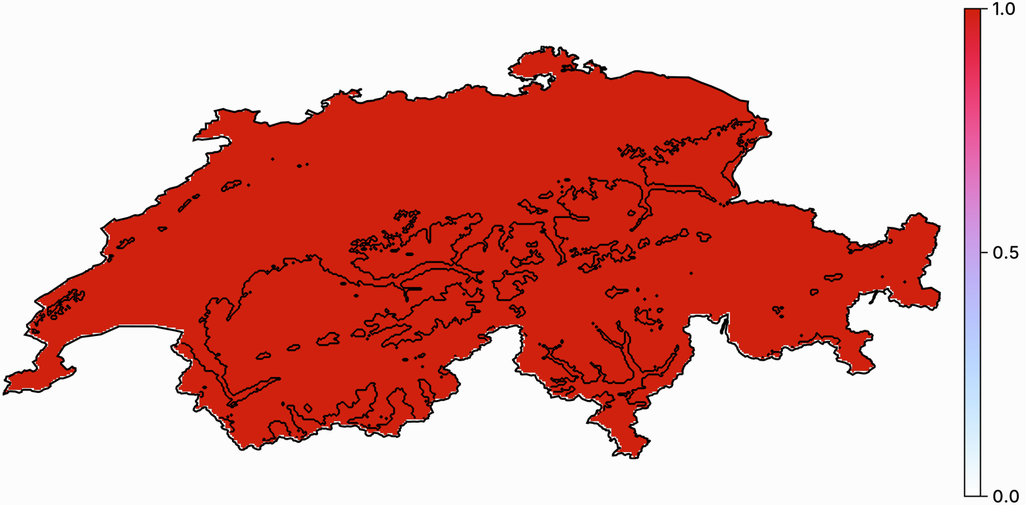
EXAMPLE WITH TWO VARIABLES



SPATIALIZED PARTIAL RESPONSE PLOT



SPATIALIZED PARTIAL RESPONSE (BINARY OUTCOME)



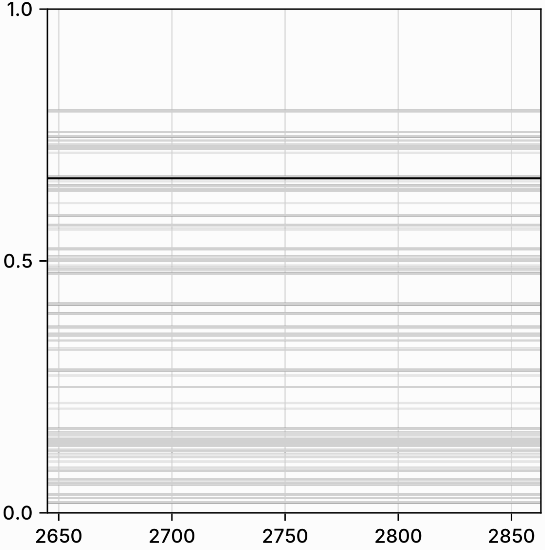
Averaging the variables is masking a lot of variability!

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

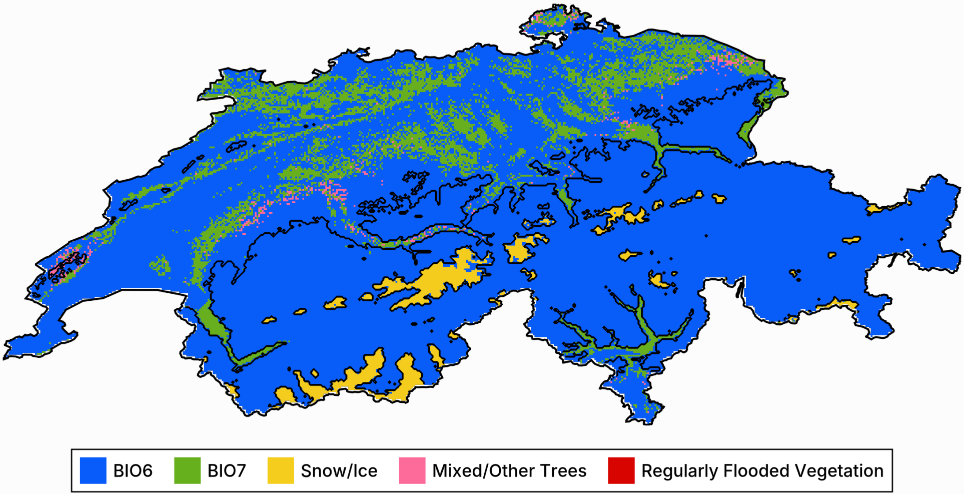
RESPONSE CURVES REVISITED

ON A MAP

VARIABLE IMPORTANCE REVISITED

Layer	Variable	Import.	Shap. imp.
6	BIO6	0.805809	0.796968
7	BIO7	0.151936	0.143859
29	Snow/Ice	0.0192739	0.0296059
23	Mixed/Other Trees	0.0225518	0.0292317
27	Regularly Flooded Vegetation	0.000429548	0.000336003

MOST IMPORTANT PREDICTOR



REVISITING THE DATA TRANSFORMATION

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

EVALUATING THE COUNTERFACTUALS

WHAT IS A GOOD COUNTERFACTUAL

learning rate and loss function

use on prediction score and not yes/no!

§ 6

Conclusions