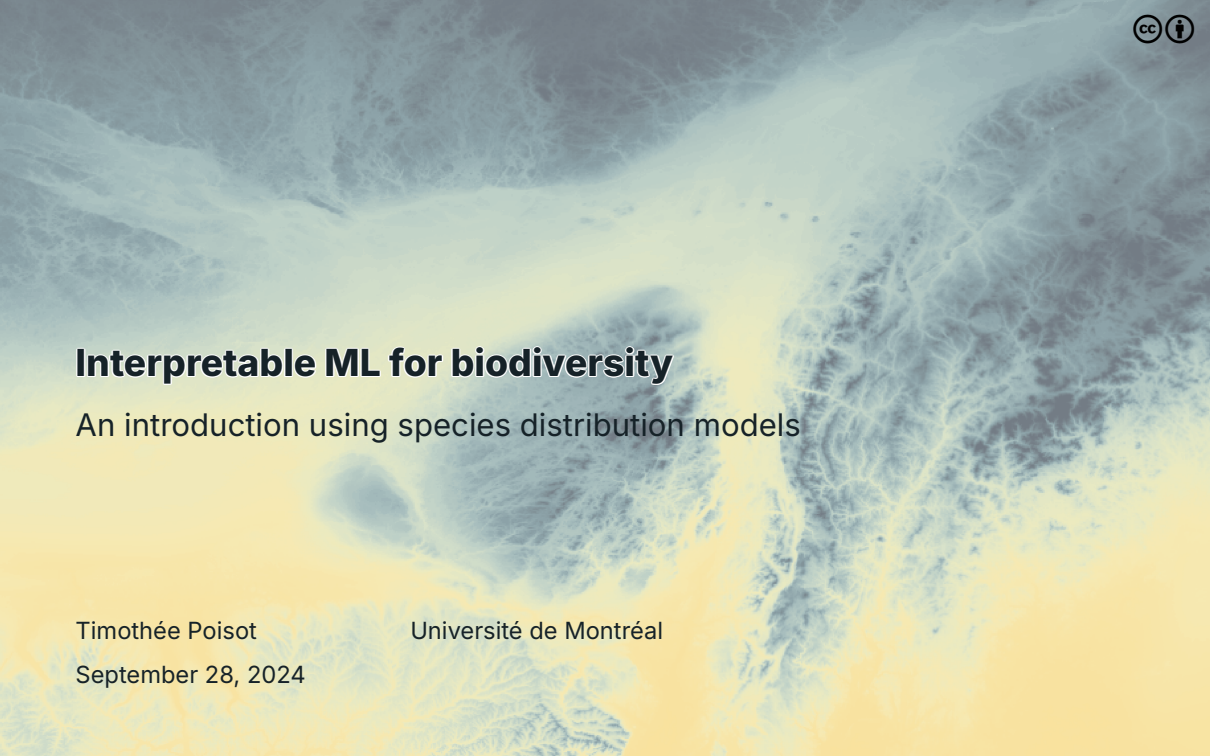


Interpretable ML for biodiversity

An introduction using species distribution models

Timothée Poisot
Université de Montréal

September 28, 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

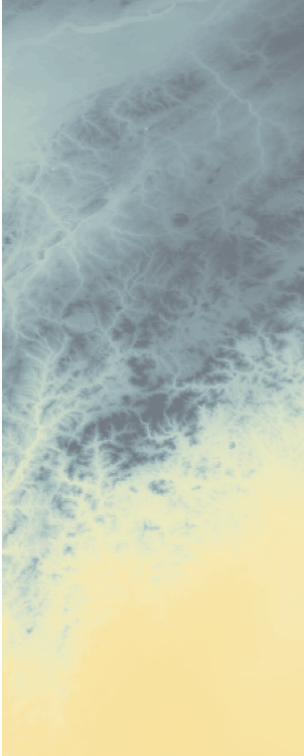
BUT WHY...

... 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

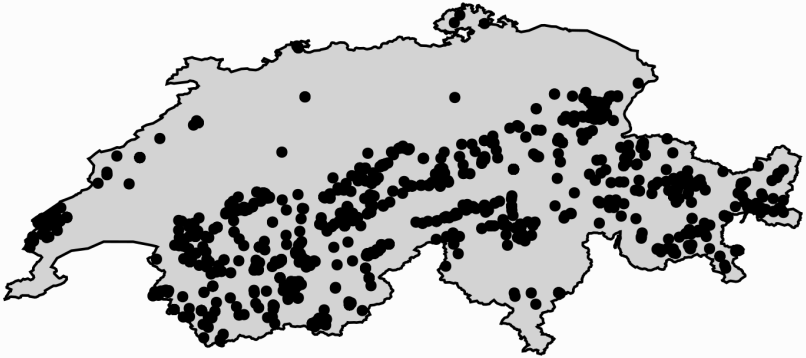
We have a series of observations $y \in \mathbb{B}$, and predictors variables $\mathbf{X} \in \mathbb{R}$

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



PROBLEM!

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

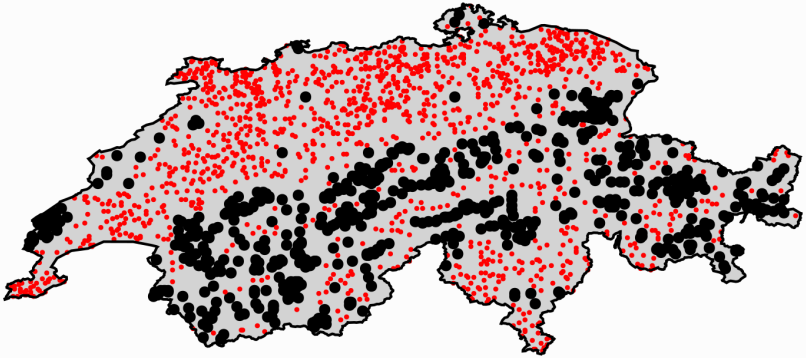
SOLUTION!

pseudo-absences

what are the assumptions we make

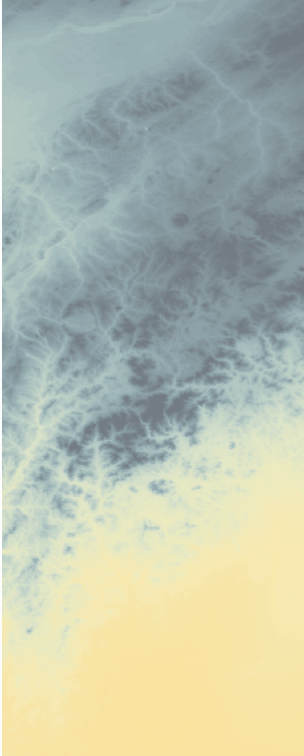
```
SDM Layer with 45336 Bool cells
    Proj string: +proj=longlat +datum=WGS84 +no_defs
    Grid size: (239, 543)
```

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

```
SDeMo.MultivariateTransform{MultivariateStats.PCA} → SDeMo.NaiveBayes → P(x) ≥ 0.5
```

Can we train the model

assumes parallel universes with slightly less data

is the model good?

CROSS-VALIDATION STRATEGY

k-fold

validation / training / testing

`0.33164125784542653`

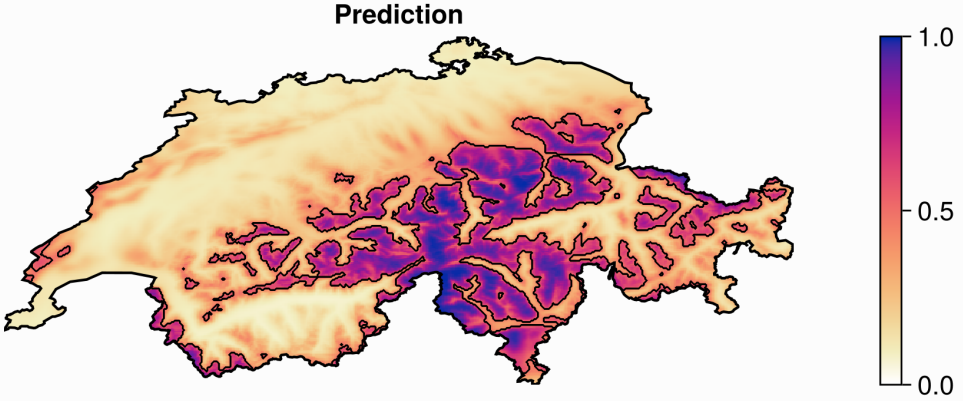
WHAT TO DO IF THE MODEL IS TRAINABLE?

train it!

re-use the full dataset

```
SDM Layer with 69967 Bool cells
    Proj string: +proj=longlat +datum=WGS84 +no_defs
    Grid size: (239, 543)
```

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

OPTIMIZING THE THRESHOLD