SPECIES DISTRIBUTION MODELS

MATIS BREILLAD | FLORIAN DUBOIS | ALVIN VEDEL

SUMMARY

PROJECT MANAGEMENT

Maximilien Servajean

Manager

Alvin Vedel

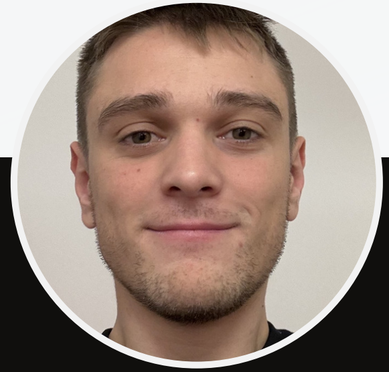
Team Leader

SVM models & Clustering

Matis Breillad

ACP

Random Forest



Florian Dubois

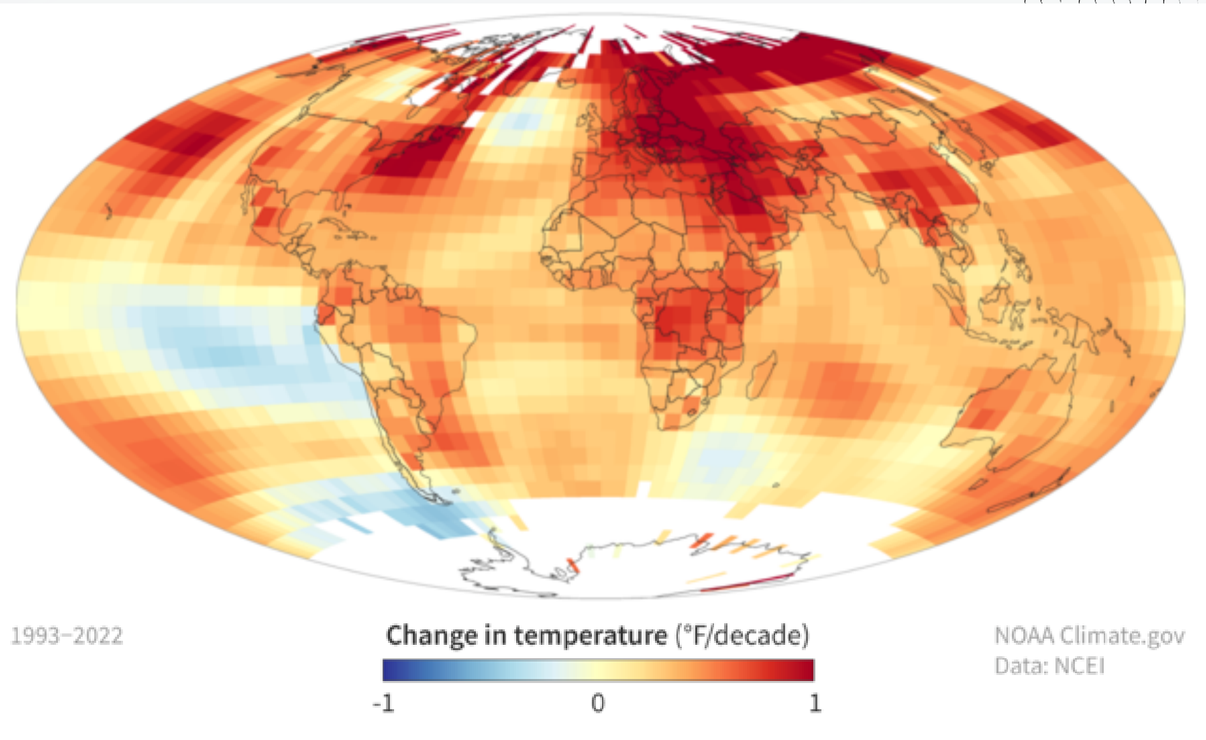
KNN

PO data integration

ENVIRONNEMENTAL CONTEXT

- **40%** of plants are threatened with **extinction**

- **2022** temperature **1.06 ˚C** warmer than the pre-industrial period



https://www.climate.gov/news-features/understanding-climate/climate-change-global-temperature

PROJET KAGGLE

kaggle

GeoLifeCLEF 2023 - LifeCLEF 2023 x FGVC10

Location-based species presence prediction

- The distribution of species models holds the answer.

- Kaggle Leaderboard

$$F_1 = \frac{1}{N} \sum_{i=1}^{N} \frac{\mathrm{TP}_i}{\mathrm{TP}_i + (\mathrm{FP}_i + \mathrm{FN}_i)/2}$$

Private Score	Public Score
0.15816	0.16249

OUR DATA

- **Presence / Absence data**

  Data of individual

  1 OBSERVATION = 1 SPECIESID + 1 PATCHID

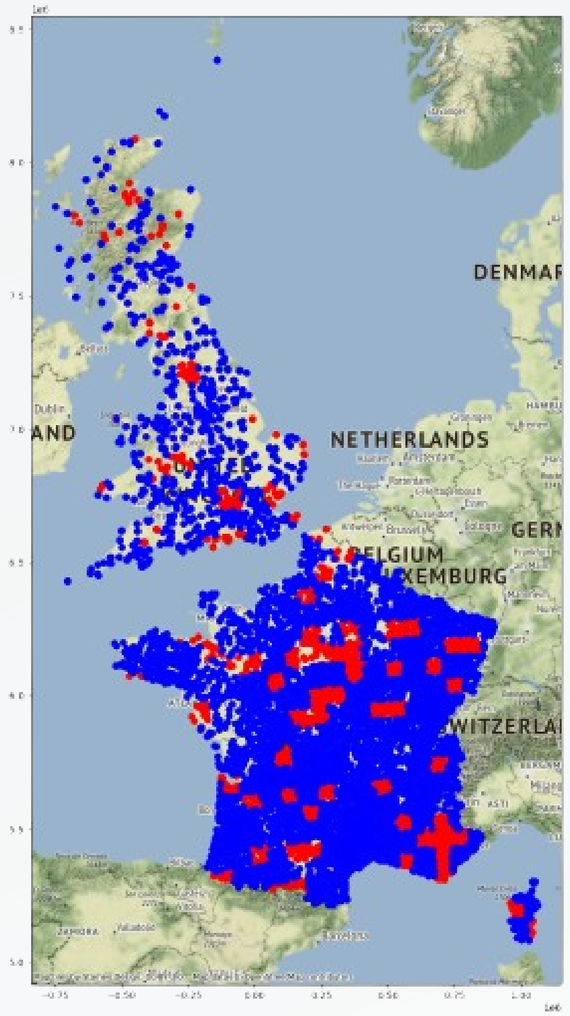
- **Abiotic data**

  Co-variables

  bio1, ..., bio19

  soil composition (clay, nitrogen, ...)

  Popdensity, Lights, Built, Railways, Roads



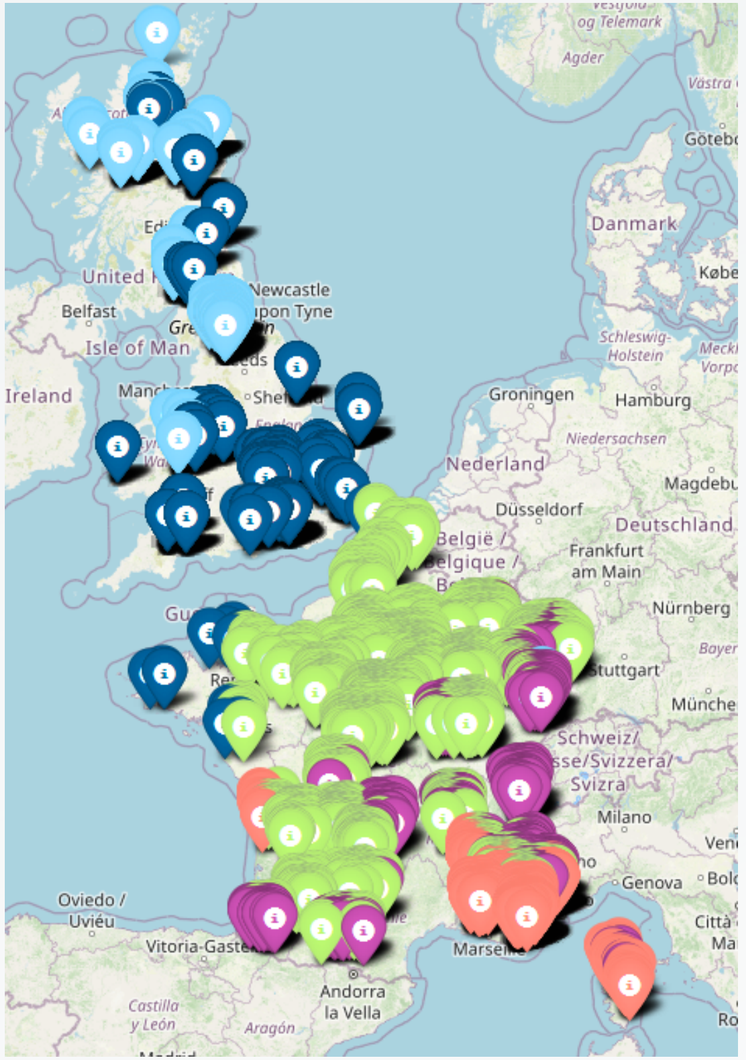
distribution of observations

OUR DATA

- Clustering of patchIds based on environmental variables

A patern appears like “ecoregions”

- Mountains
- Littoral
- Plains
- Armoric (Bretagne)
- Northern UK

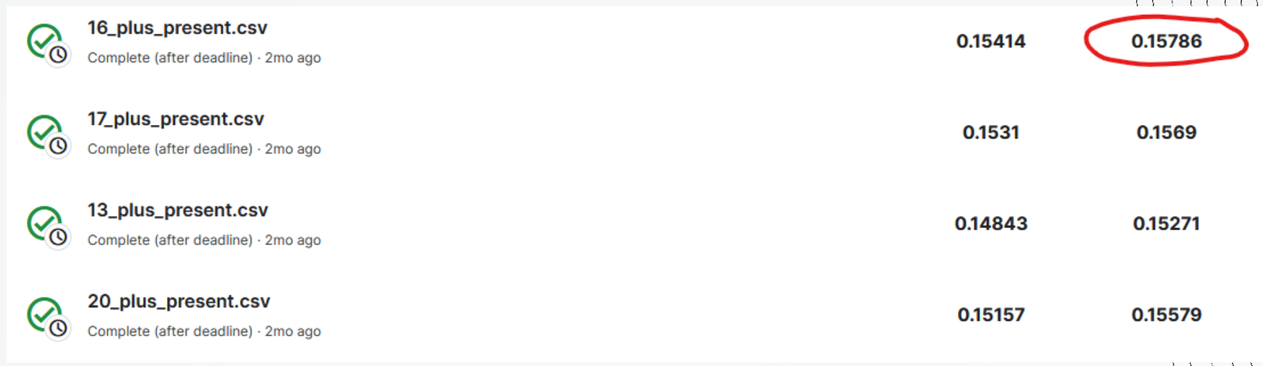


distribution of abiotic data

NAIVES MODELS AND RESULTS

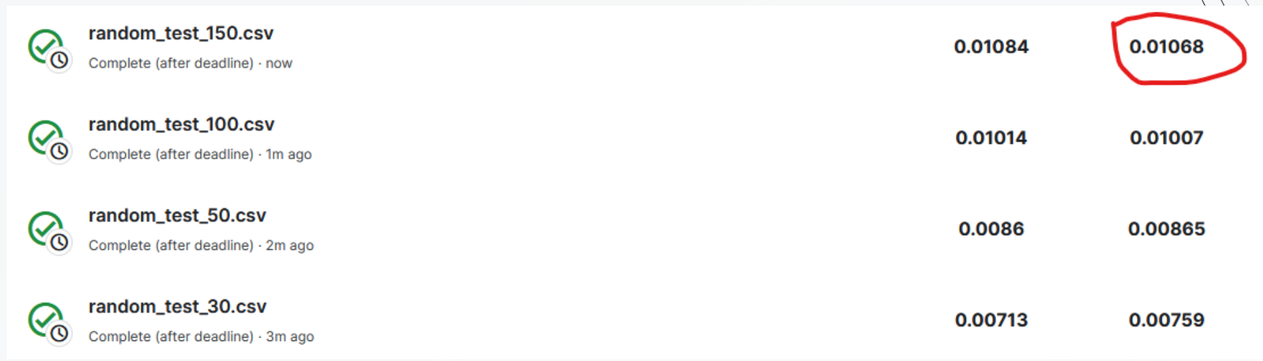
2 Kinds of naives models

- K most presents
- Random



16_plus_present.csv Complete (after deadline) · 2mo ago	0.15414	0.15786
17_plus_present.csv Complete (after deadline) · 2mo ago	0.1531	0.1569
13_plus_present.csv Complete (after deadline) · 2mo ago	0.14843	0.15271
20_plus_present.csv Complete (after deadline) · 2mo ago	0.15157	0.15579

K most presents model kaggle result



random_test_150.csv Complete (after deadline) · now	0.01084	0.01068
random_test_100.csv Complete (after deadline) · 1m ago	0.01014	0.01007
random_test_50.csv Complete (after deadline) · 2m ago	0.0086	0.00865
random_test_30.csv Complete (after deadline) · 3m ago	0.00713	0.00759

random model kaggle result

SVM MODEL ATTEMPT

- **The SVM principle**

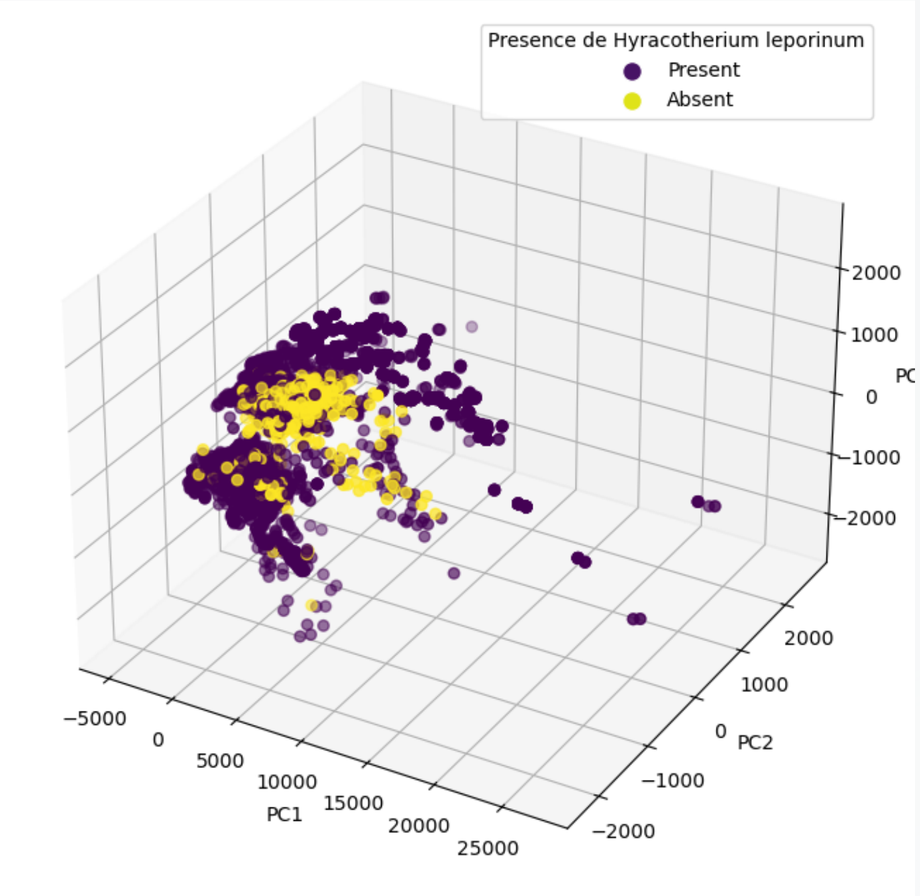
  Kernel : polynomial

- **Pre-processing of data**

  Quantitative variables : bio1, ..., bio19 ; soil composition

  ACP : 98% of explained variance in 3 dimensions

- Results



3-dimension PCA

SVM MODEL ATTEMPT

- The SVM principle
- Pre-processing of data
- **Results**

modeles_svm_d4_p01_p3.csv Complete (after deadline) · 2h ago	0.17112	0.17581
modeles_svm_d3_p01_p3.csv Complete (after deadline) · 4h ago	0.17465	0.17909
modeles_svm_d3_p02_p3.csv Complete (after deadline) · 4h ago	0.12772	0.13112

SVM Model Result

K-NEAREST NEIGHBORS

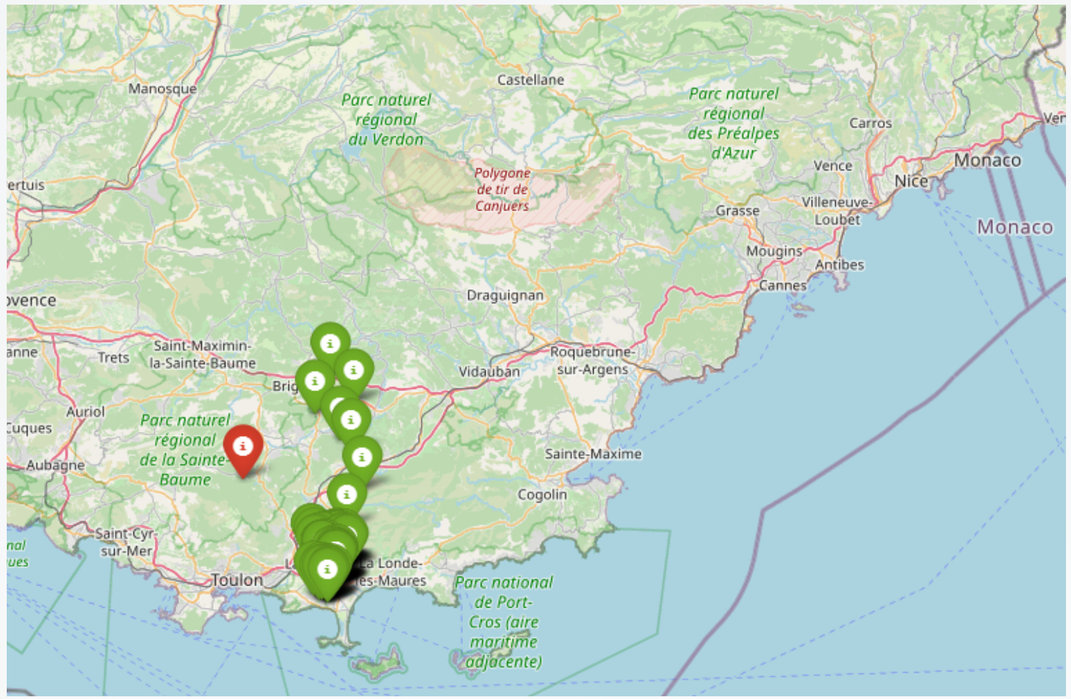
2 hyperparameters:

- K: number of neighbors
- P: numbers of occurences / K (threshold)

variables used

- latitude and longitude

Geographic model



Spatial K-nearest neighborhood

K-NEAREST NEIGHBORS

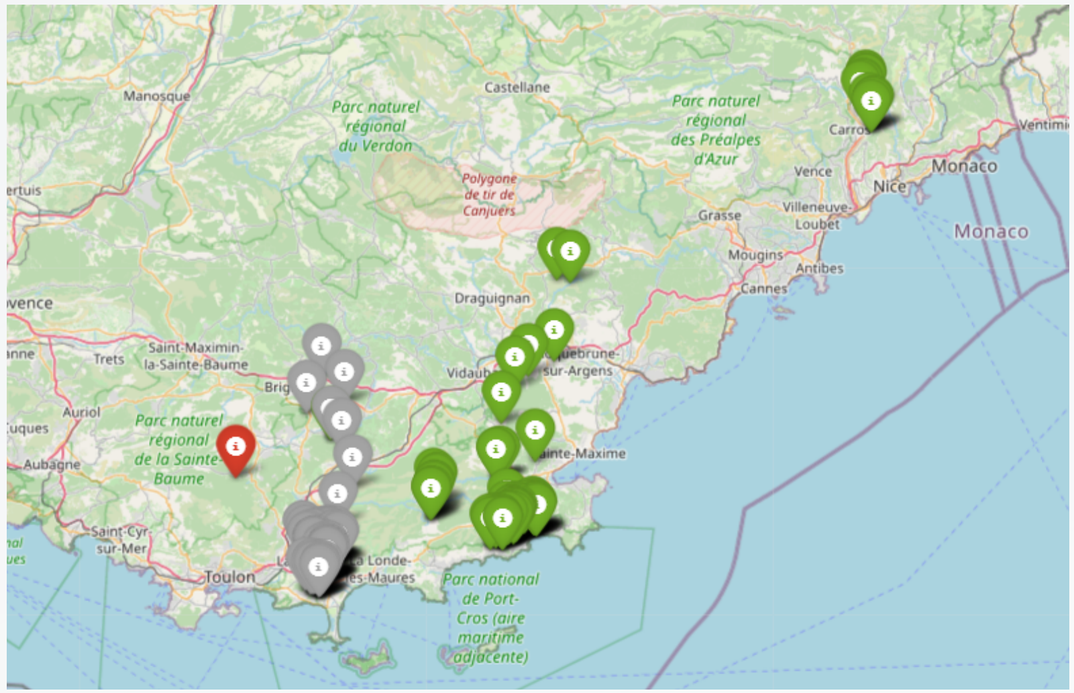
2 hyperparameters:

- K: number of neighbors
- P: numbers of occurences / K (threshold)

33 variables used

- Differents temperature measurements
- Soil composition
- Population density

Environnemental model

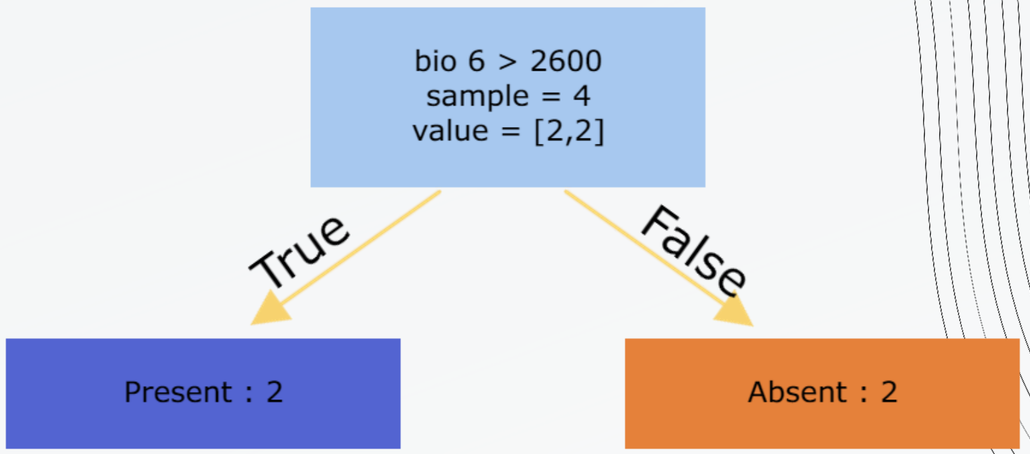


distribution of abiotic data

RANDOM FOREST

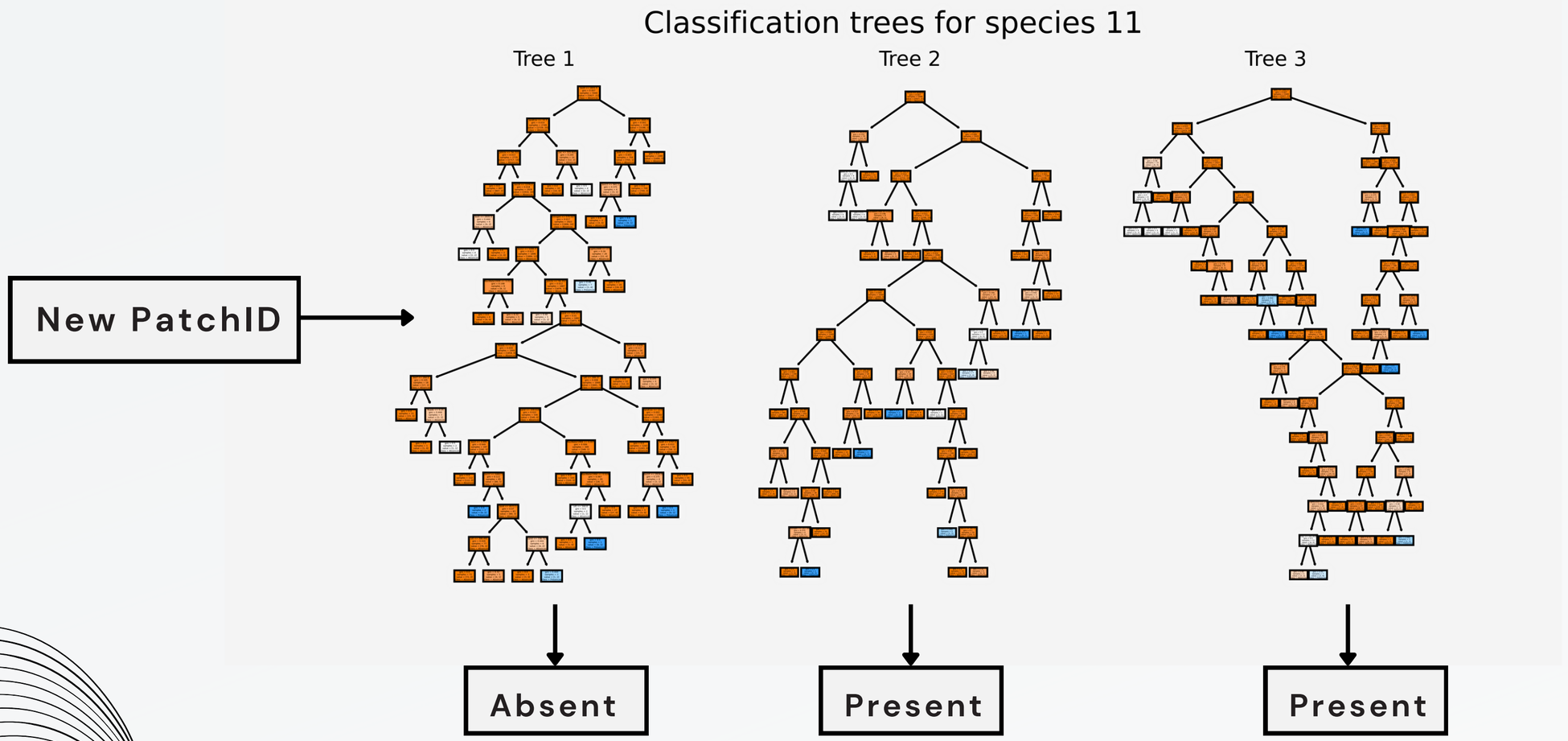
patchID	bio3	bio4	bio5	bio6	NavWater2009	Popdensity1990	Popdensity2010	11
3010676	2.880000	4422	2935	2760	0.008707	7.000000	7.000000	1
3010677	2.880000	4422	2935	2500	0.008707	7.000000	7.000000	0
3010677	2.880000	4422	2935	2500	0.008707	7.000000	7.000000	0
3010683	2.880000	4422	2935	2760	0.008707	7.000000	7.000000	1

extract from our data



classification tree (example)

RANDOM FOREST

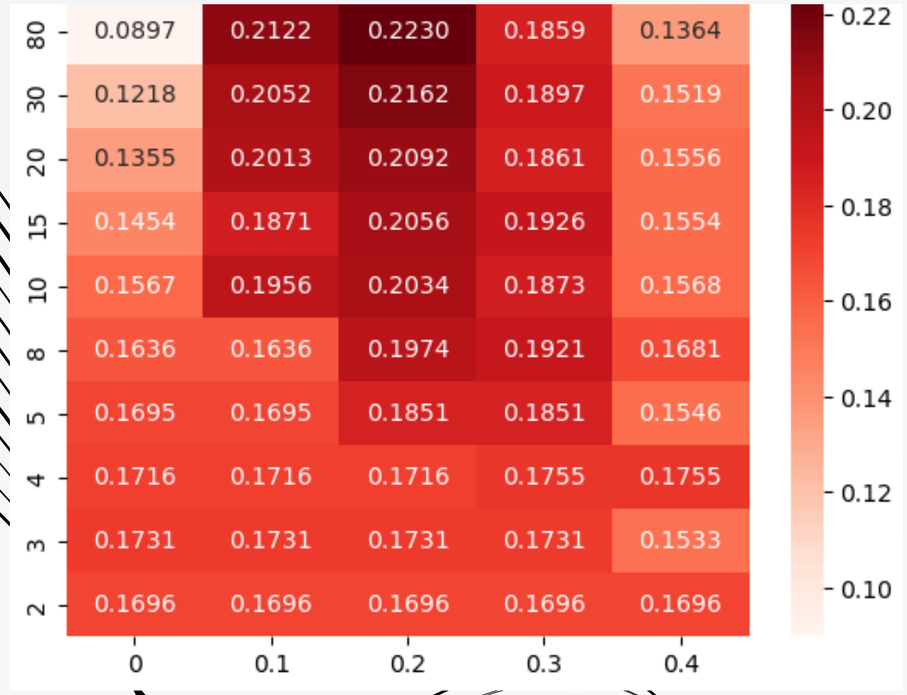


Random Forest with 3 trees

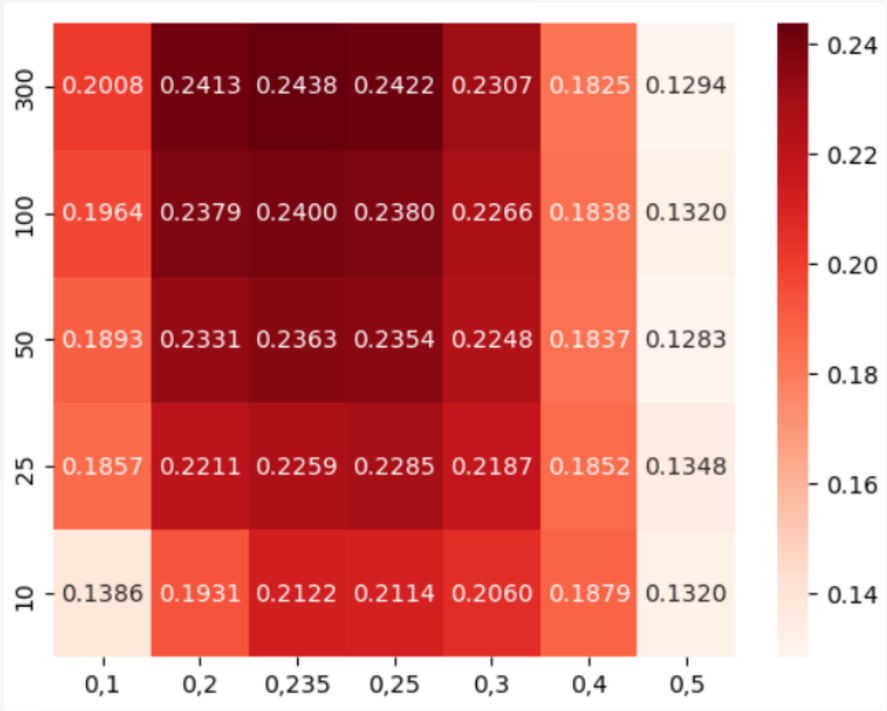
PROBABILITY SELECTION

We only consider values with probabilty over a threshold

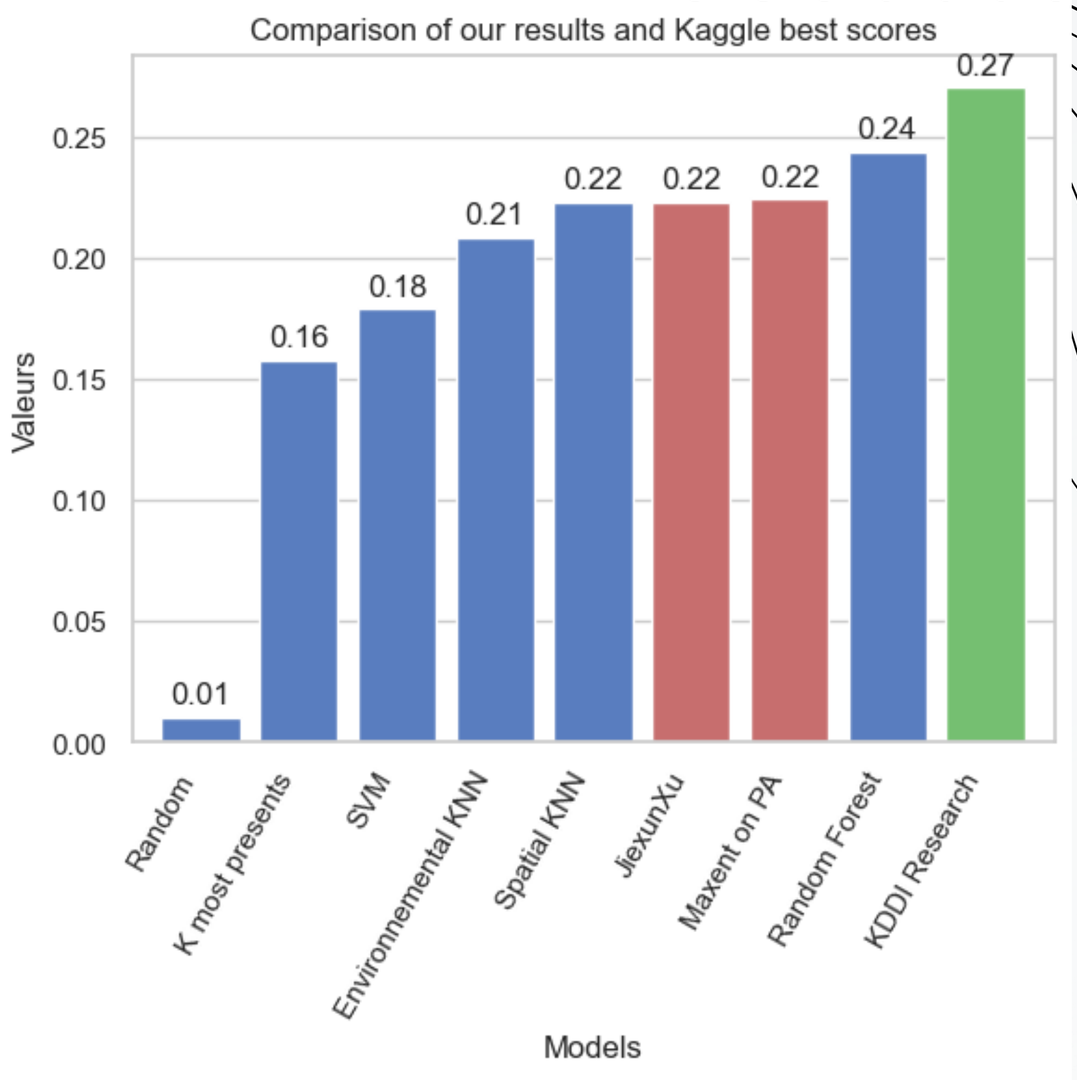
K- neighbours results



RandomForest results



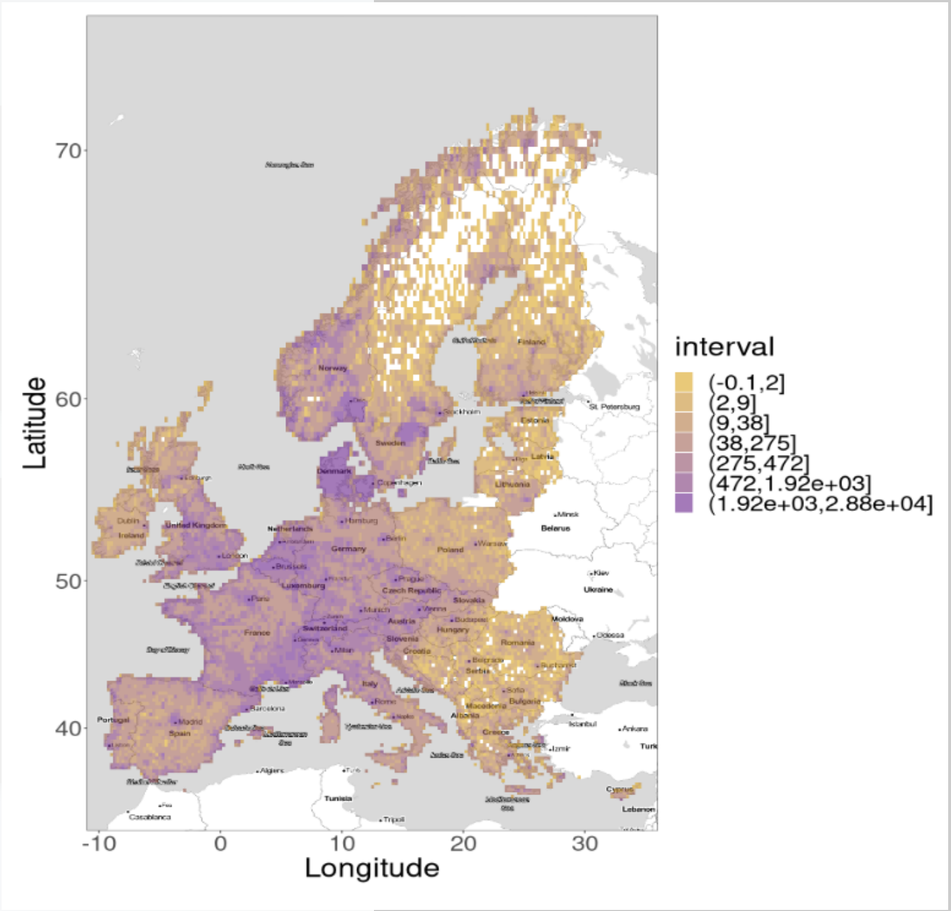
ALL RESULTS


Comparison of our results and Kaggle best scores
Valeurs
0.00
0.05
0.10
0.15
0.20
0.25
0.01
0.16
0.18
0.21
0.22
0.22
0.22
0.24
0.27
Random
K most presents
SVM
Environnemental KNN
Spatial KNN
JiexunXu
Maxent on PA
Random Forest
KDDI Research
Models

PRESENCES ONLY DATA

The presence only data : similarities and differences

Biased data to take into account

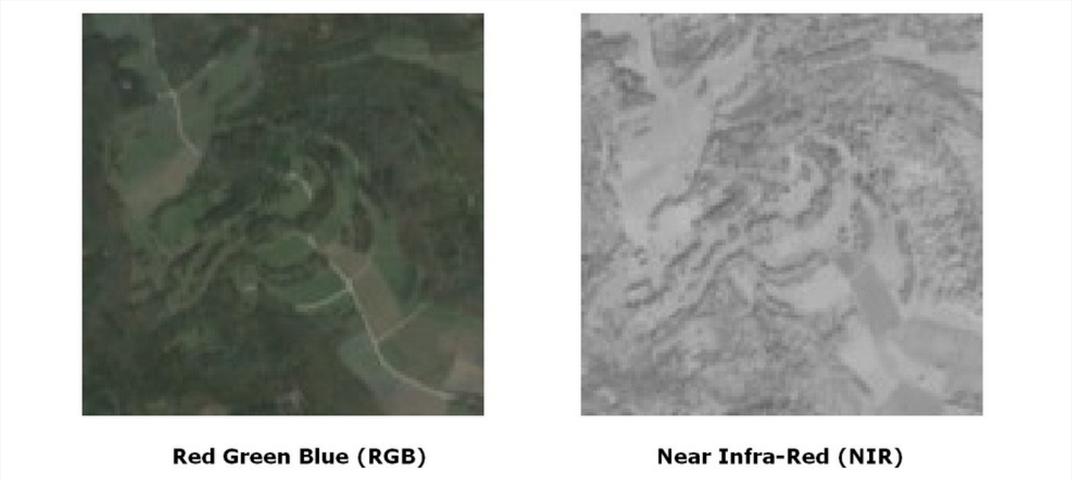


https://arxiv.org/ftp/arxiv/papers/2308/2308.05121.pdf

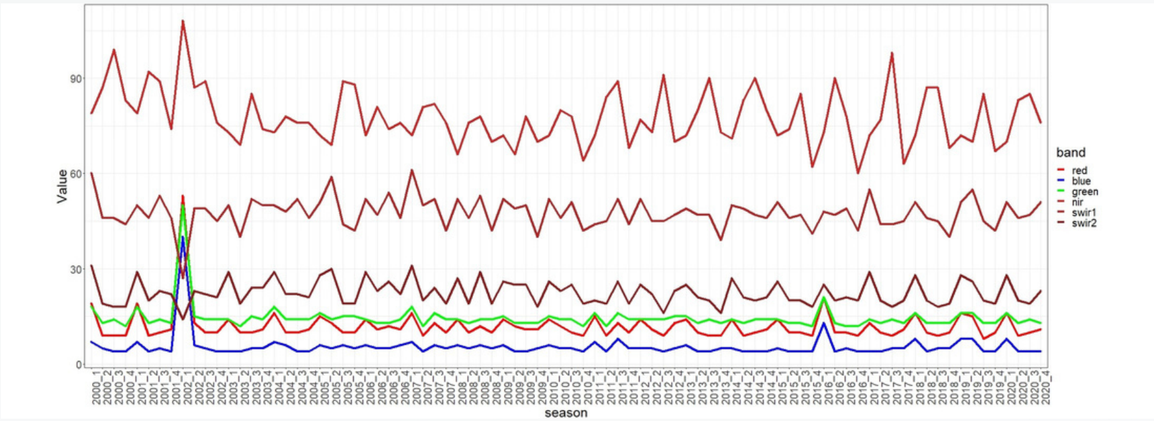
Pl@ntNet

UNUSED INFORMATION

- **Satellite images and time series**
- Raster images



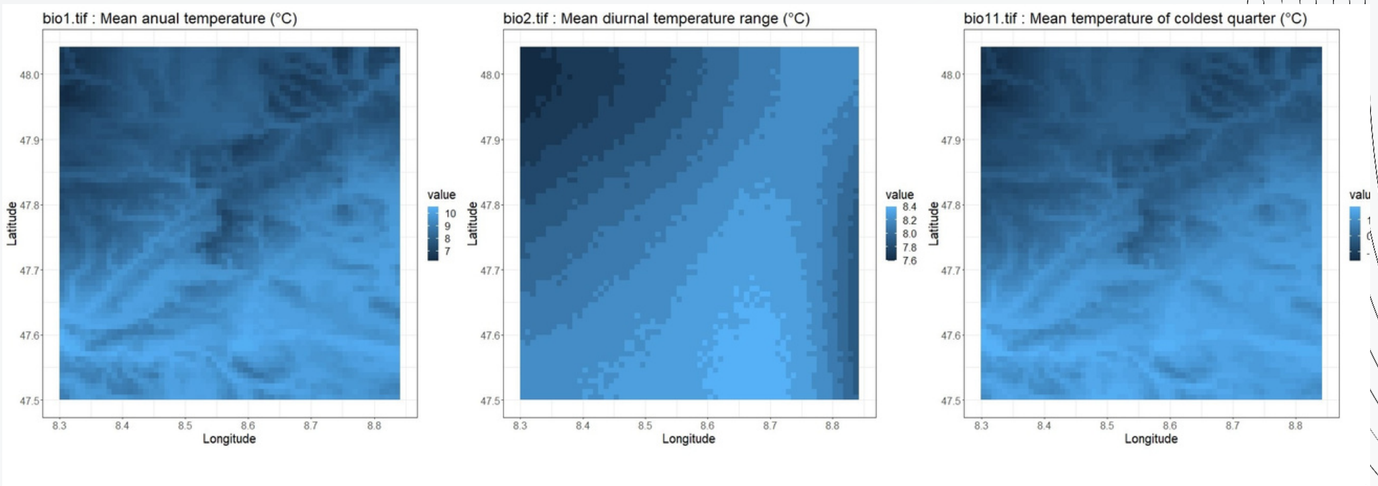
Satellite images



Satellite time series

UNUSED INFORMATION

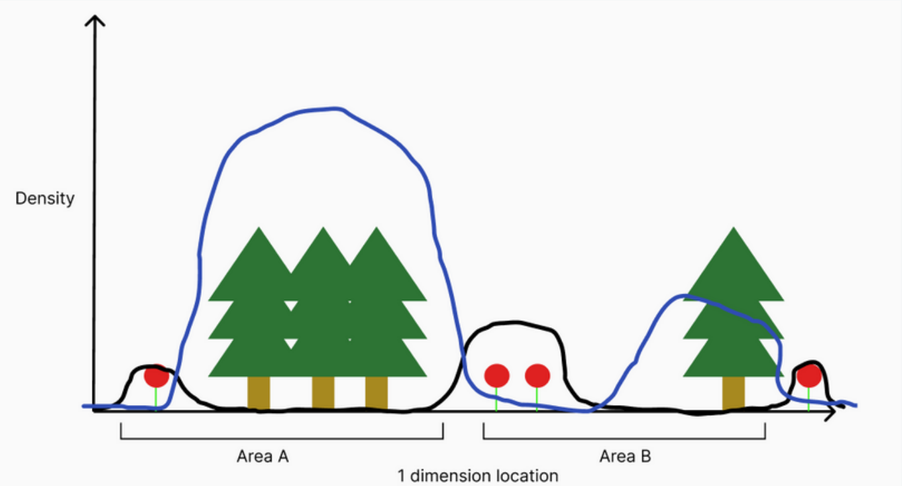
- Satellite images and time series
- **Raster images**



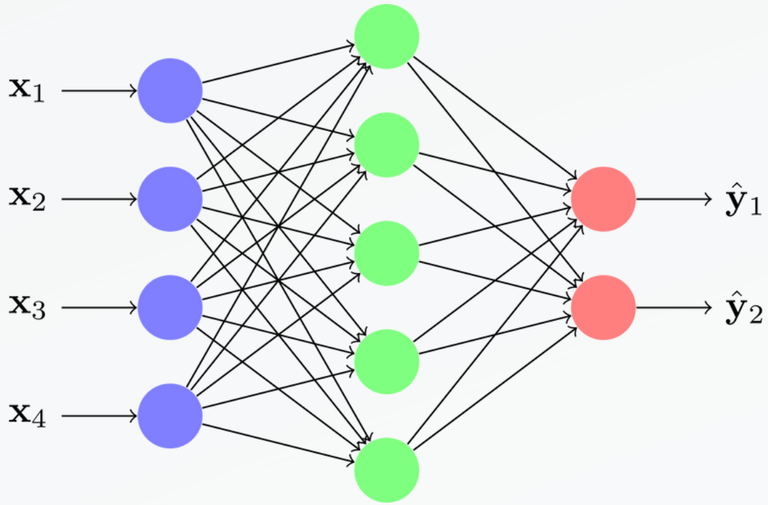
Environmental rasters

FUTURE MODELS

- Poisson point process



- Integration of other data
- Neuronal network



CONCLUSION

- Encouraging first results
- Understanding of the subject

- Get better scores with new models
- Come first in the 2024 Kaggle