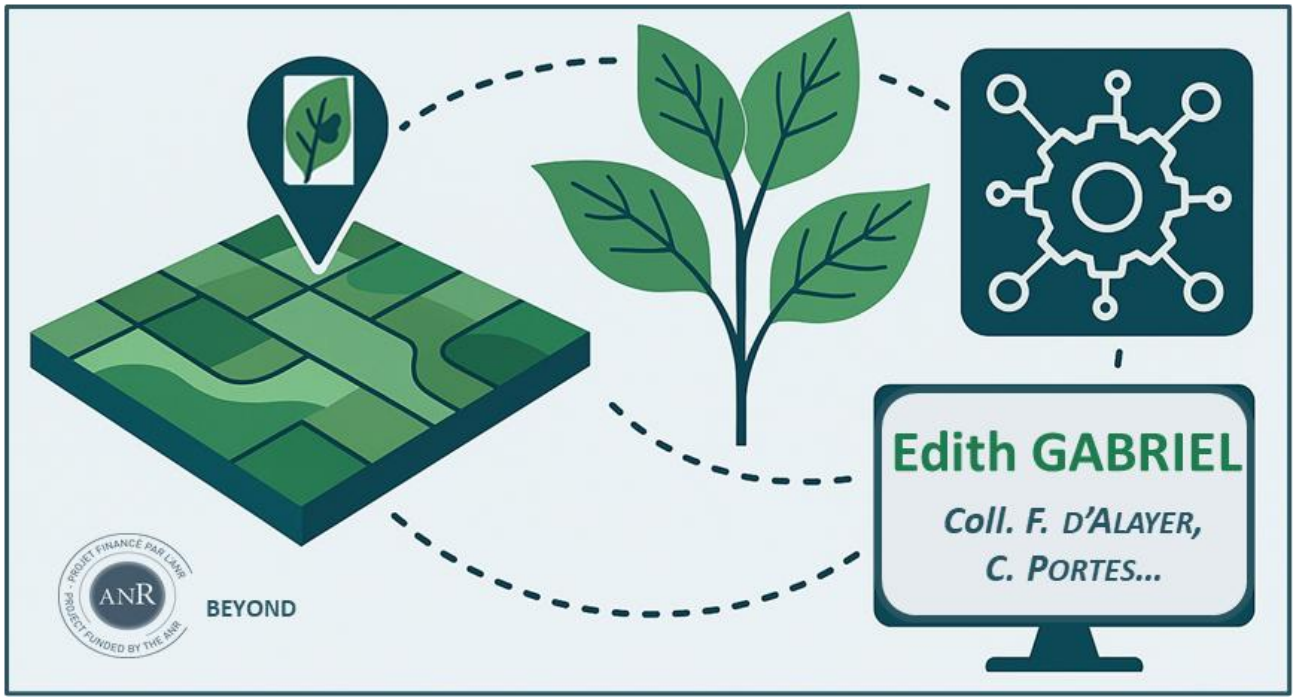
Integrating Spatial Modeling and Machine Learning for Plant Health Surveillance



MINES PARIS | PSL
INRAE
la science pour la vie, l'humain, la terre
Fondation MINES PARIS

GEOLEARNING

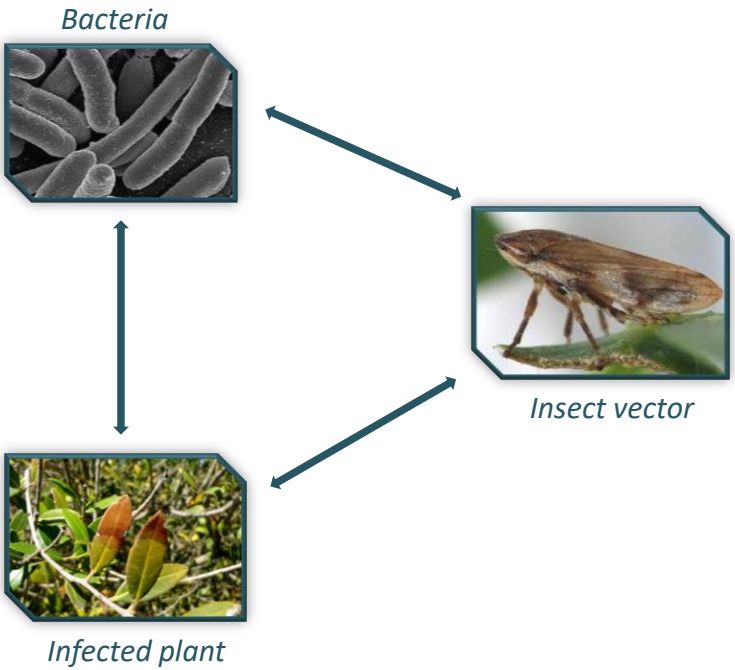
Fréjus, April 2025

ANDRA
CCR
BNP PARIBAS
SCOR
FONDATION POUR LA SCIENCE

Once upon a time *Xylella fastisiosa*

- Phytopathogenic bacterium
- Transmitted by insect vectors
- Decline or death of infected plants
- Not very characteristic symptoms

  ⇒ Analyses → healthy ✓ / contaminated ✗
- Long latency period
- More than 400 host species




Lavender


Citrus spp

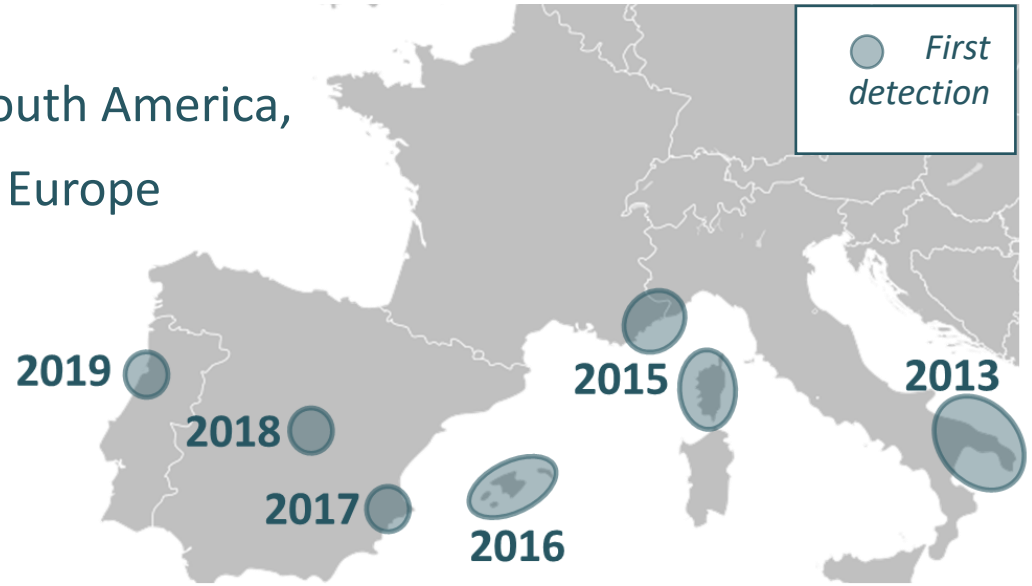

Grapevine



Olive tree

Xylella fastisiosa in EU

- After North America, South America, and Asia, *Xf* conquered Europe

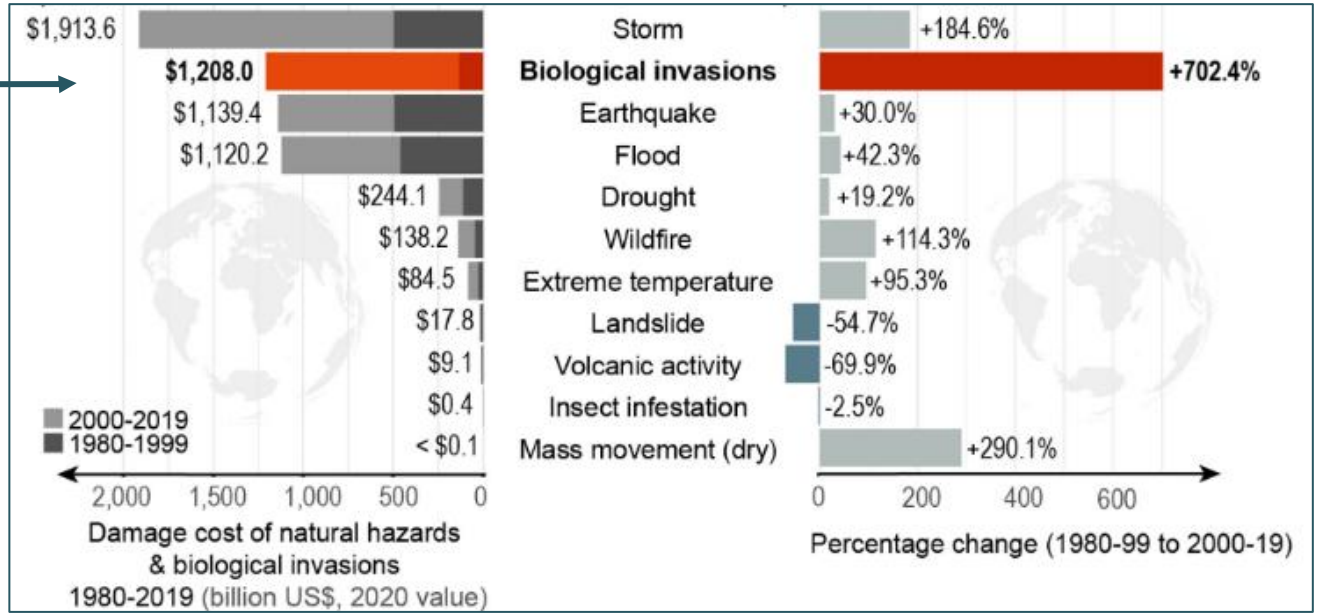


- Devastating economic impact: *Xf* has the potential to affect

70% of the production of older Olive trees (>30 y/o)
35% of the younger ones
11% of citrus
13% of almond
1-2% of grape production

} production loss of 5.5 billion €

Xylella fastisiosa in EU



Turbelin et al. *Perspectives in Ecology and Conservation 21 (2023) 143–150*

- **Devastating economic impact**: *Xf* has the potential to affect

70% of the production of older Olive trees (>30 y/o)

35% of the younger ones

11% of citrus

13% of almond

1-2% of grape production

} production loss of 5.5 billion €

How to fight?

- No effective curative method ⇒ { **surveillance** / prophylaxis / destruction } are the key strategies
- Quarantine pest under regulation
- Control measures to prevent the spread
  - Establisment of demarcated areas

    Infected zone

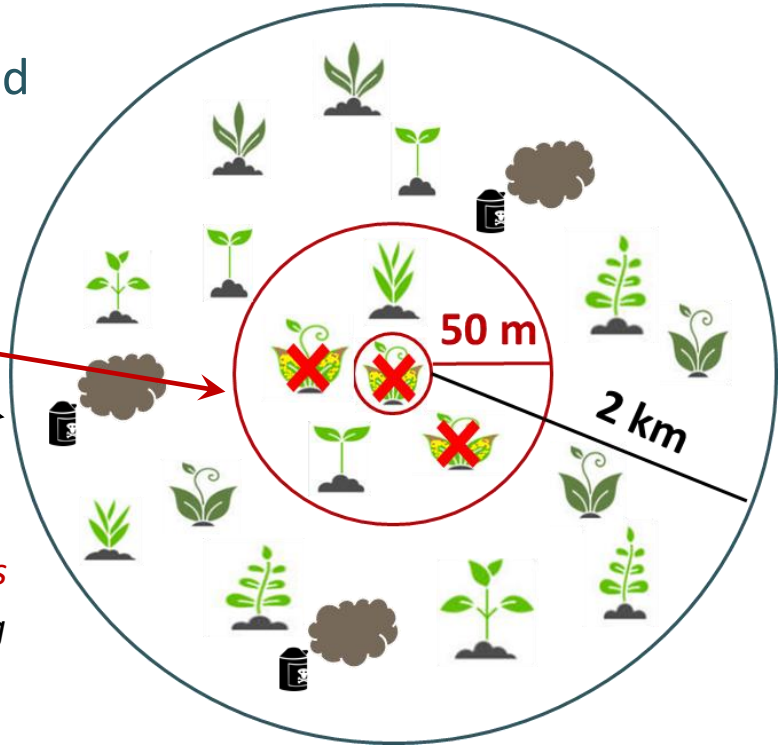
    Buffer zone

  - Eradication measures

    Remove infected or symptomatic plant

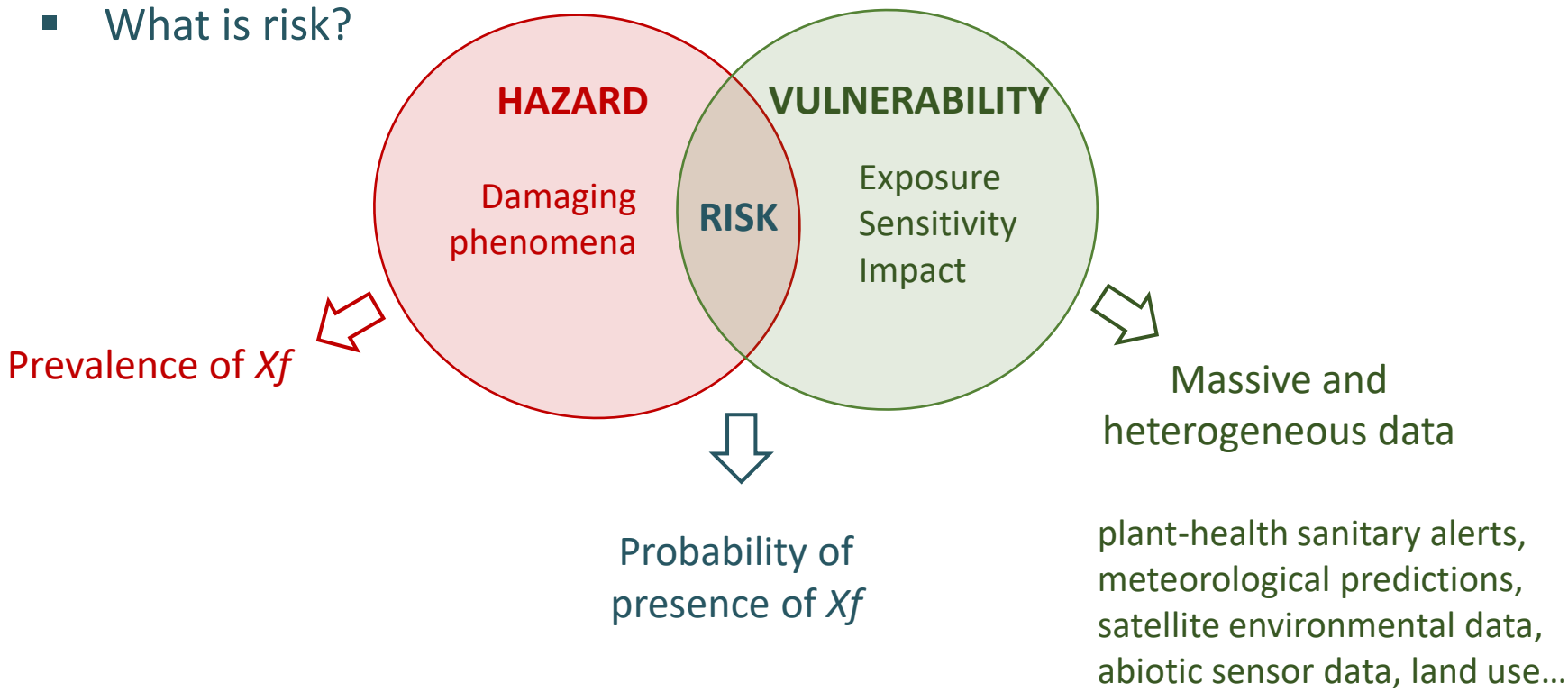
    and plants which belong to the same species

    Expanded surveillance: sampling and testing



Risk-based surveillance of *Xylella fastisiosa*

- Strategic allocation of sampling efforts across time, space and populations considering various risk factors
- What is risk?



Objectives

Learning and predicting risk, and optimizing surveillance

Learning the risk
and its determinants

- ✓ identify risk factors
- ✓ produce risk maps

Machine learning model
adapted to spatial extrapolation

RISK = HAZARD × VULNERABILITY

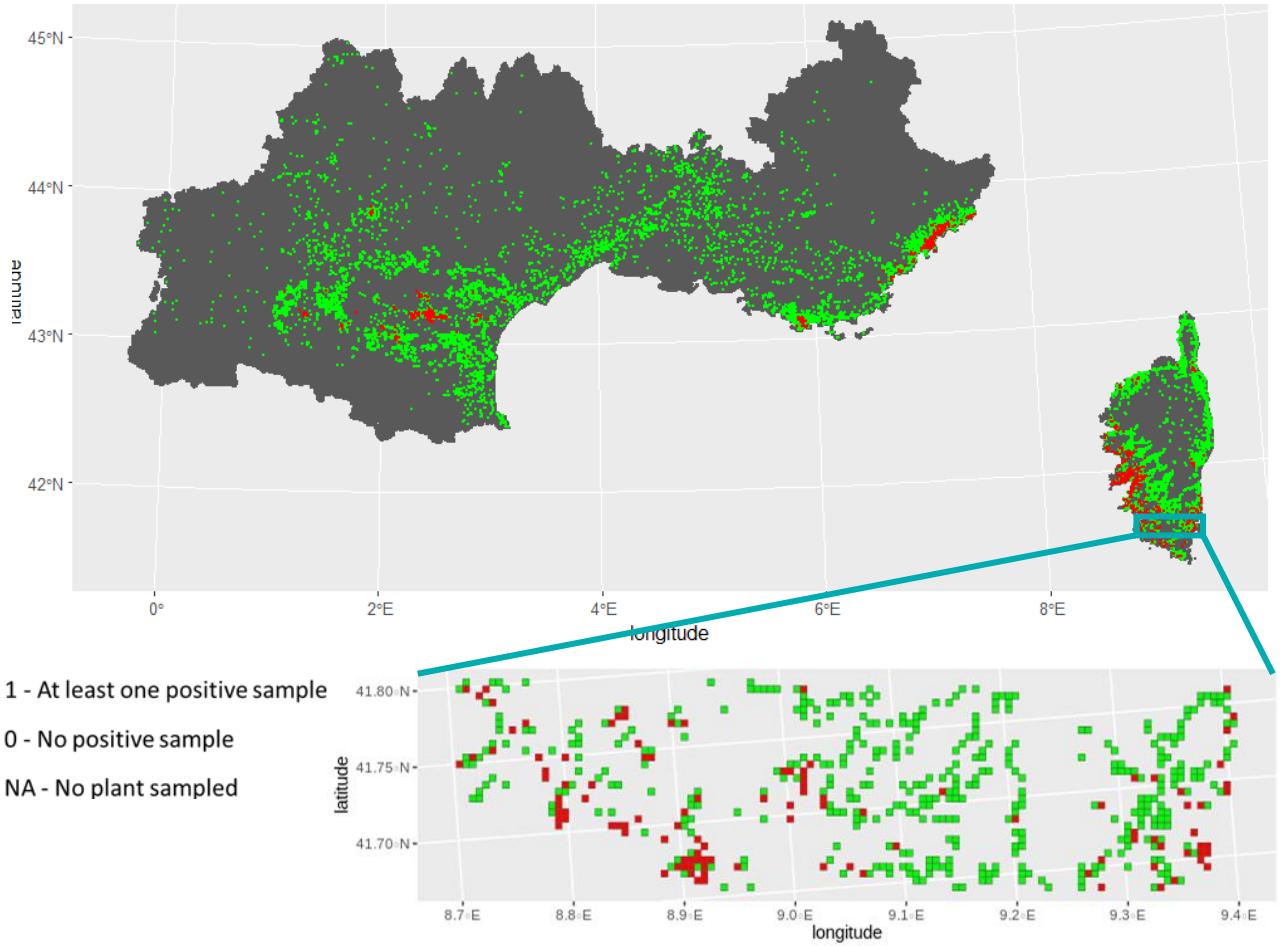
Point process models

Planning data collection
from the risk evaluation

- ✓ design surveillance models
- ✓ propose sequential sampling strategies conditional on the risk

Xylella fastisiosa in France

- Presence/absence of *Xf* 2015 – 2023

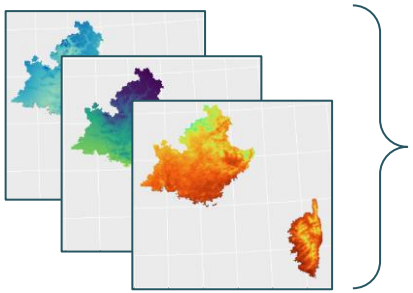


Infected	Non infected
913	9714

Learning and predicting the risk

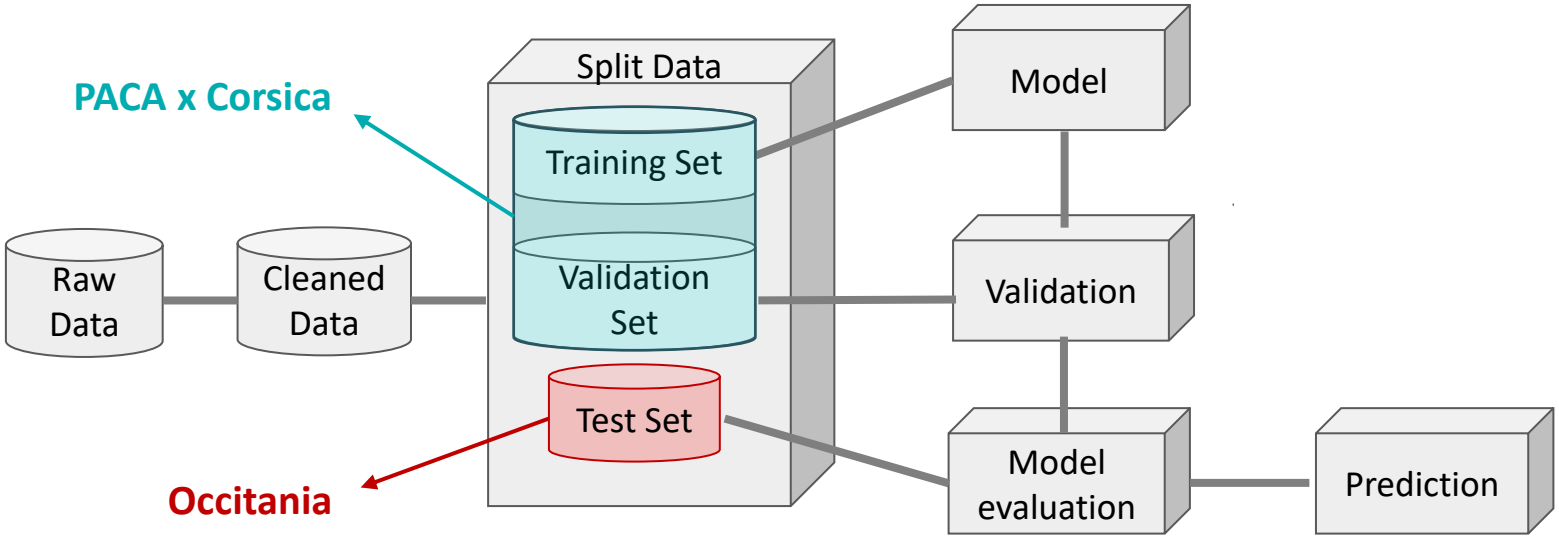
- 112 environmental and climatic factors

  Most of them are spatially autocorrelated

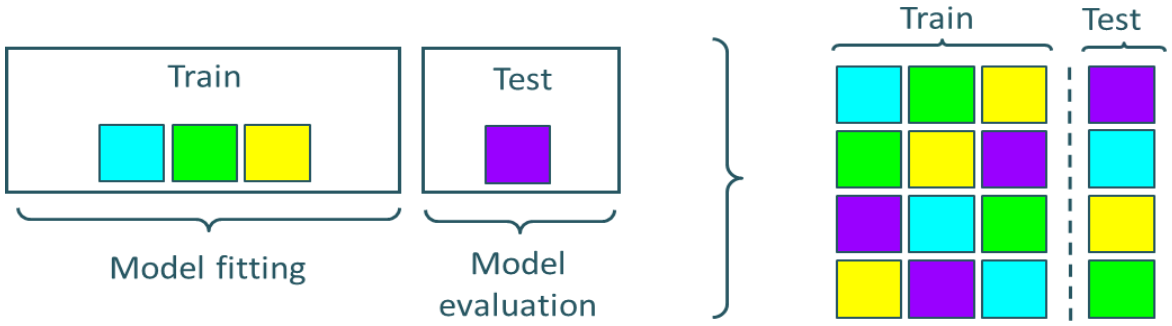


Bioclimatic
Soil composition
Land type
Land orientation
Altitude

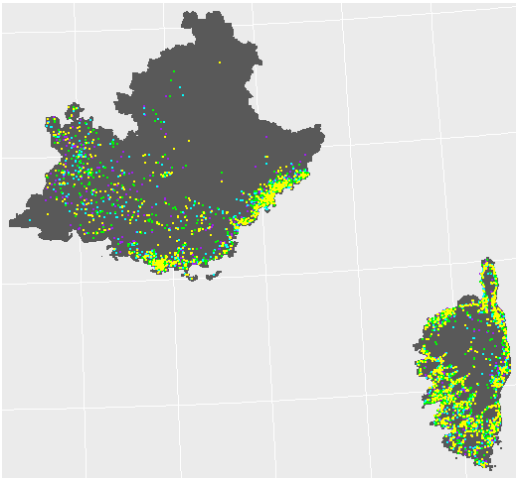
- Supervised learning workflow



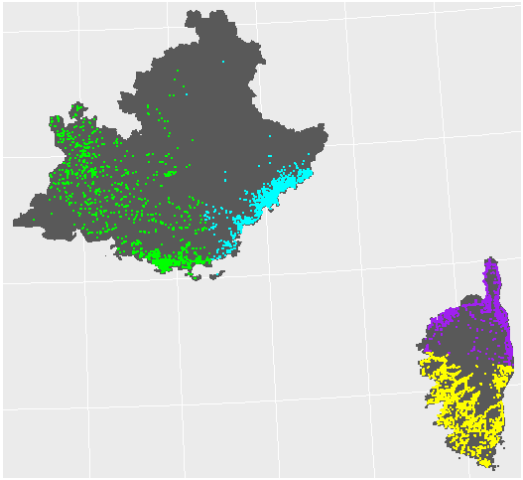
k-fold cross-validation



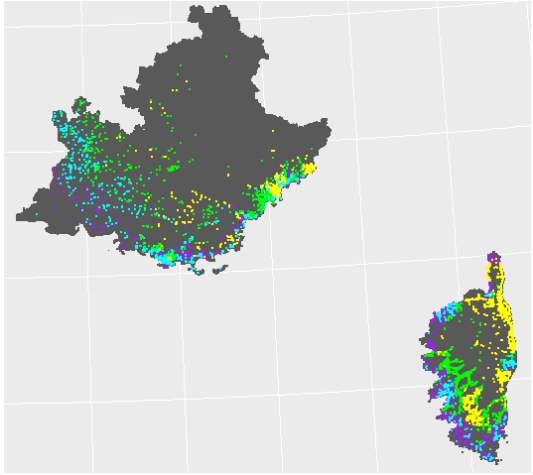
Random CV



Spatial CV



Environmental block CV



Folds 1 2 3 4

Overly optimistic model assessment for spatial dependent data

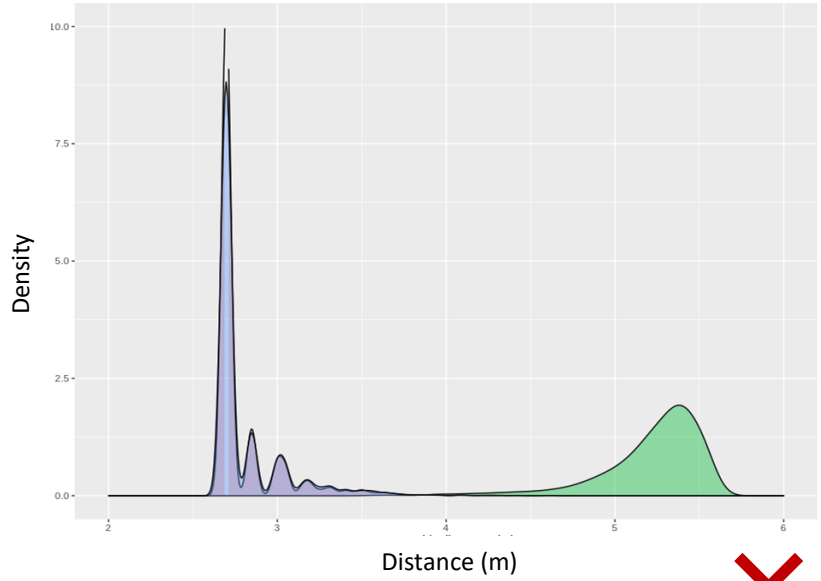
Cross-validation in spatial extrapolation

- Learning in PACA and Corsica, predicting in Occitania

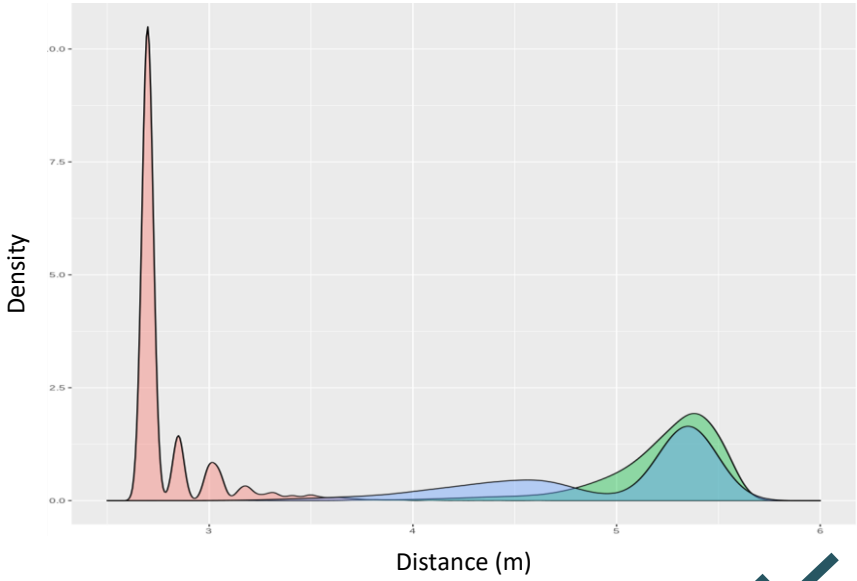
Nearest neighbor distance distribution

Train to train *Train to test* *CV*

Random CV



Spatial/block CV



The edf of nearest neighbour distances found during prediction is matched during the spatial CV process

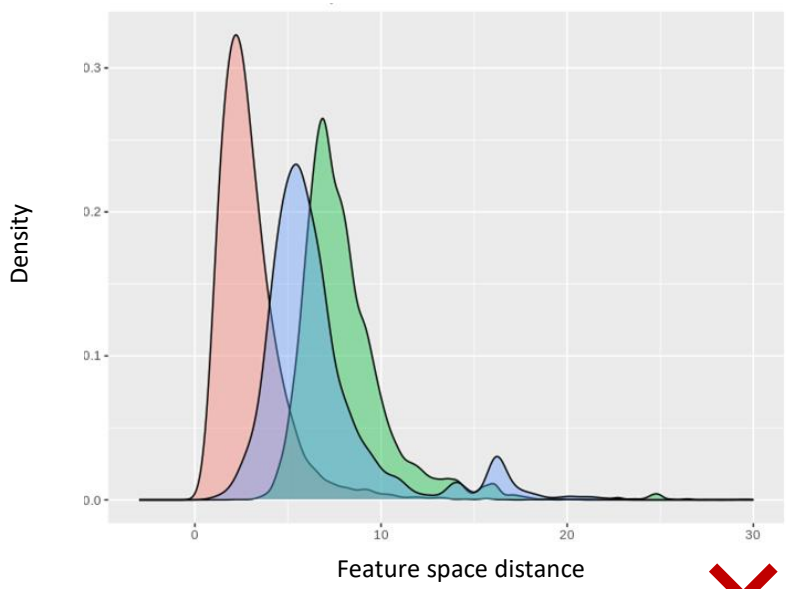
Cross-validation in spatial extrapolation

- Learning in PACA and Corsica, predicting in Occitania

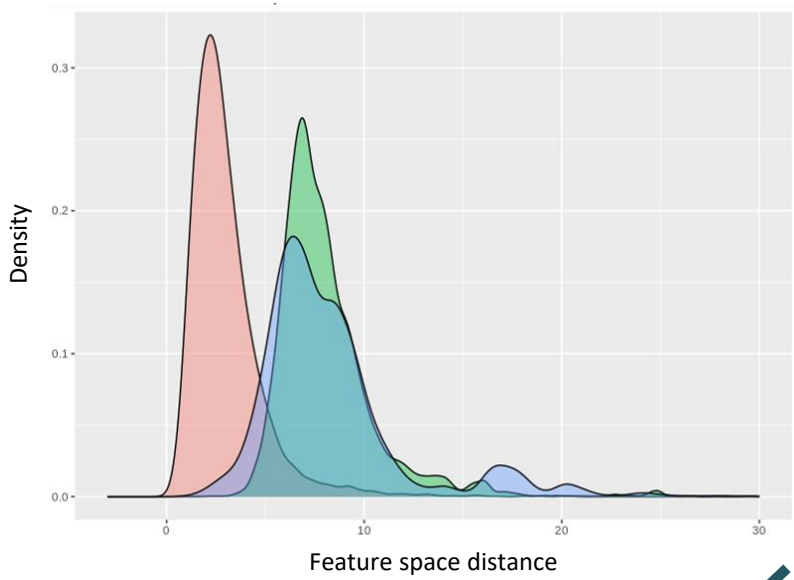
Distance distribution in feature space

Train to train *Train to test* *CV*

Spatial CV



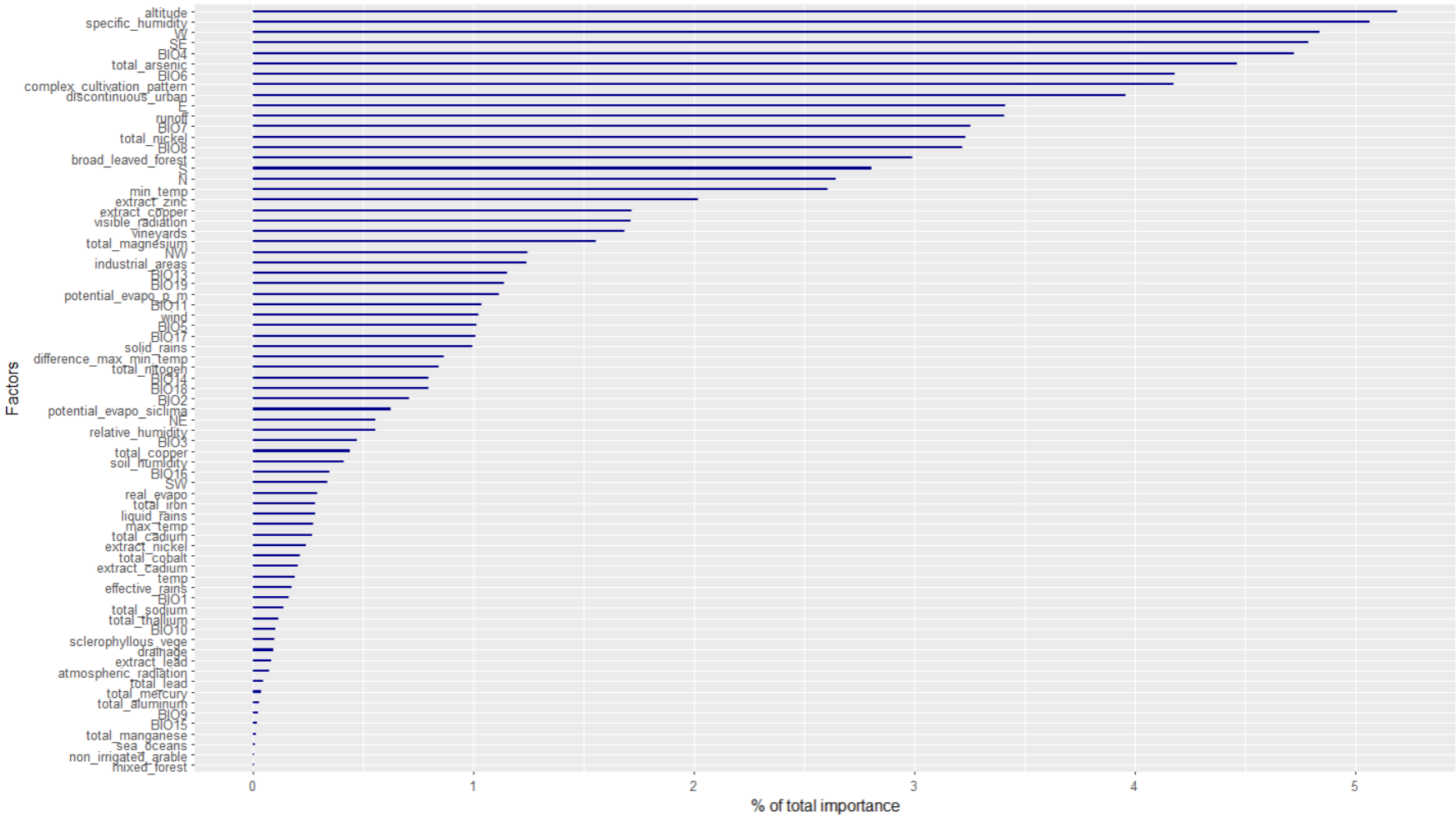
Environmental block CV



The distance of training data is rather small, compared to what is required during prediction.
Environmental block CV is doing a better job

Feature selection

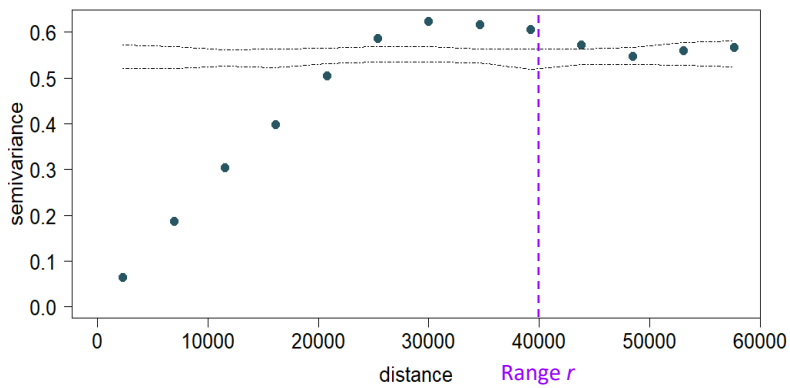
Scores based on importance obtained from 29 machine learning methods

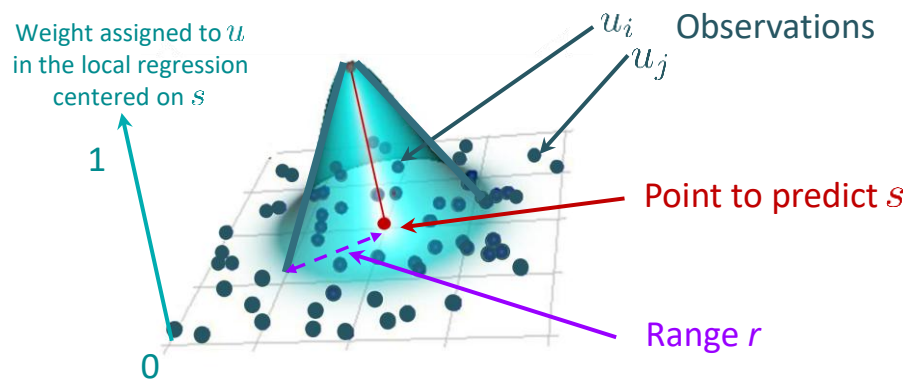


Predictive model 1

XGBoost x spatially weighted factors

- We define a weight matrix W_k for spatially structured factors X_k

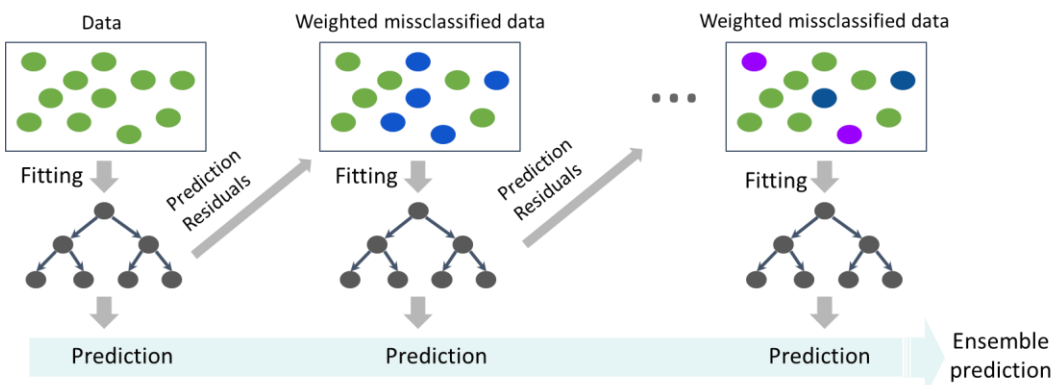




- We get a new set of factors: $X_W = (WX_{spat}, X_{aspat})$

- eXtreme Gradient Boosting

$$Y = f(X_W)$$



Predictive model 2

Spatial logistic regression model

- Presence/absence: $Y(s) \sim \mathcal{B}er(p(s))$

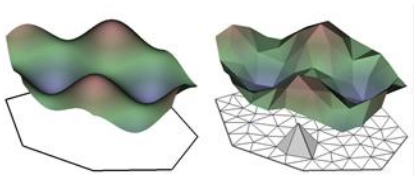
- Probability of presence $p(s)$, such that

$$log\frac{p(s)}{1-p(s)} = \beta X(s) + Z(s)$$

Factors

Gaussian random field

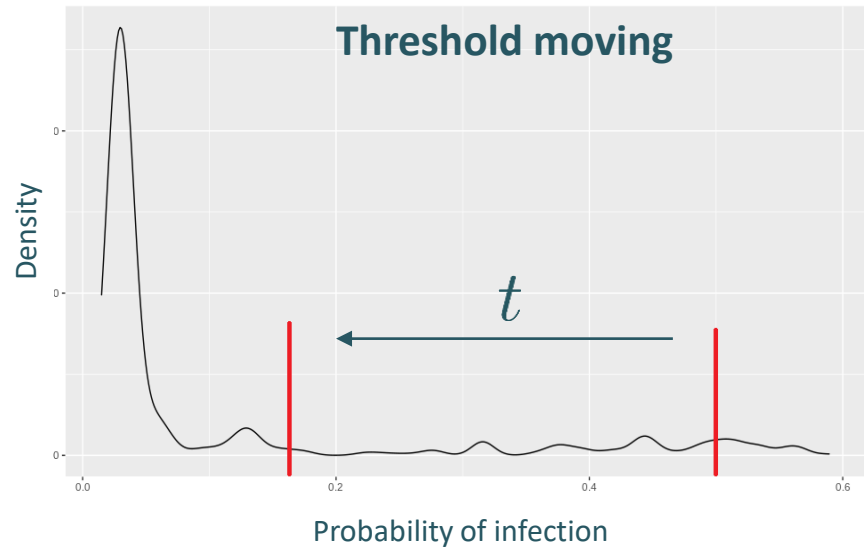
- Inference INLA/SPDE



Validation

- **Presence/absence:**

$$Y(s) = \begin{cases} 1, & \text{if } p(s) \geq t, \\ 0, & \text{if } p(s) < t \end{cases}$$



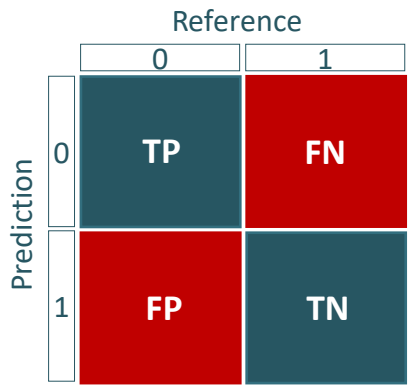
- **Metrics**

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

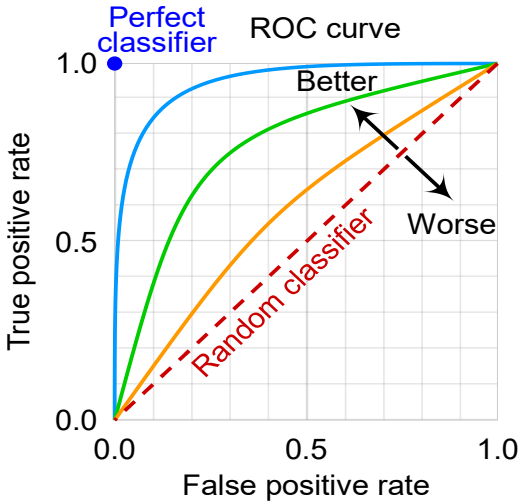
$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Balanced accuracy} = \frac{\text{Sens.} + \text{Spec.}}{2}$$

Confusion matrix



AU(RO)C



Model comparison: test in Occitania

0 4520 1 206

XGBoost (X_W)

AUC: **0.523**

Balanced accuracy: **0.437**

Prediction \ Reference	0	1
0	2740	151
1	1780	55

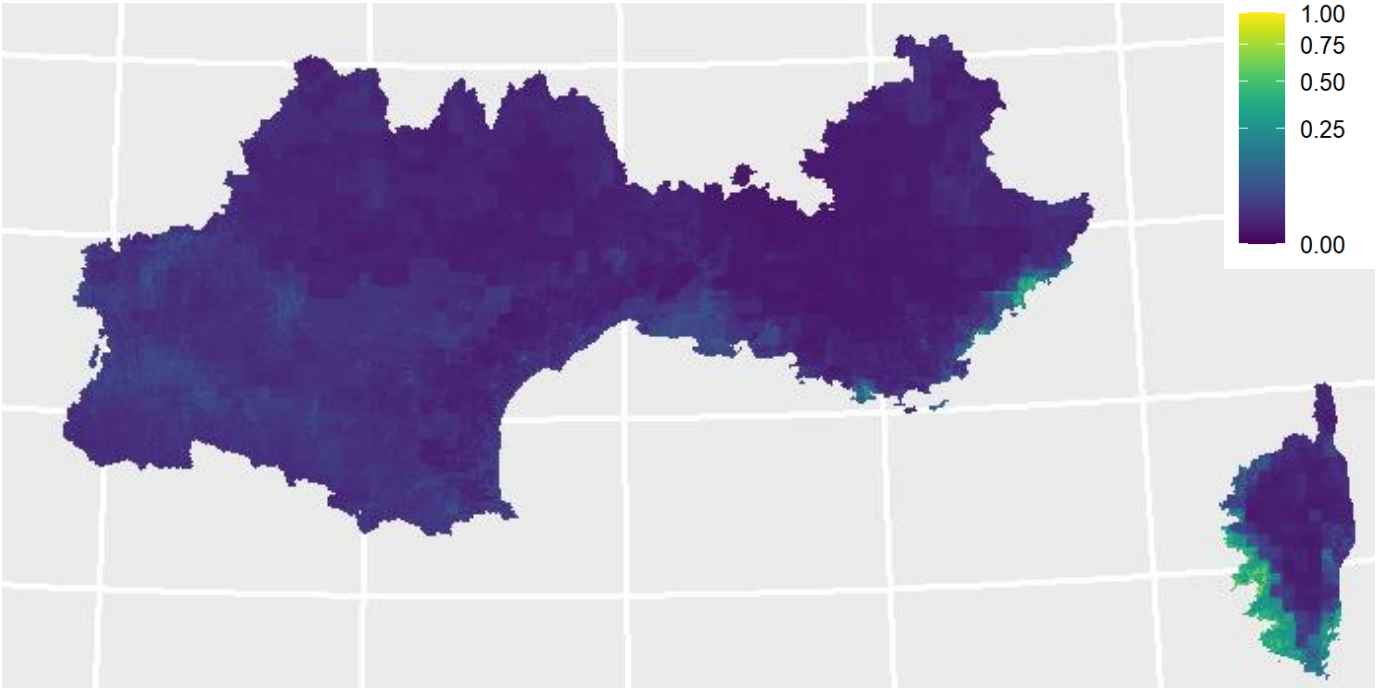
Spatial logistic regression model

AUC: **0.605**

Balanced accuracy: **0.397**

Prediction \ Reference	0	1
0	3280	192
1	1240	14

Spatial prediction - XGBoost(X_W)



Model comparison

- Models' performances are questionable
  - Hard to set the threshold in extrapolation
  - Existing areas of applicability*?

- But surveillance aims to improve it

Learning the risk and its determinants ⟹
- ✓ identify risk factors
- ✓ produce risk maps

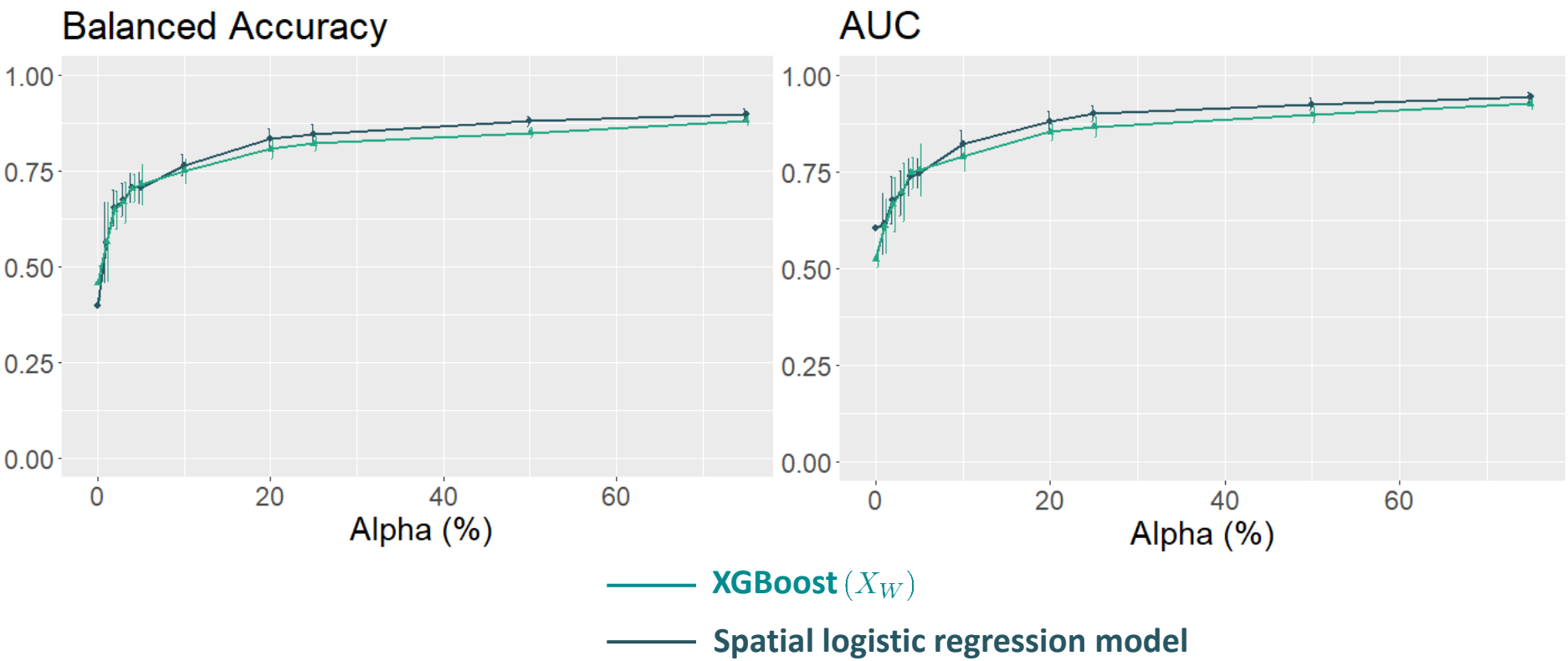
RISK = HAZARD × VULNERABILITY

Planning data collection from the risk evaluation ⟹
- ✓ design surveillance models
- ✓ propose sequential sampling strategies conditional on the risk (*remember François' talk*)

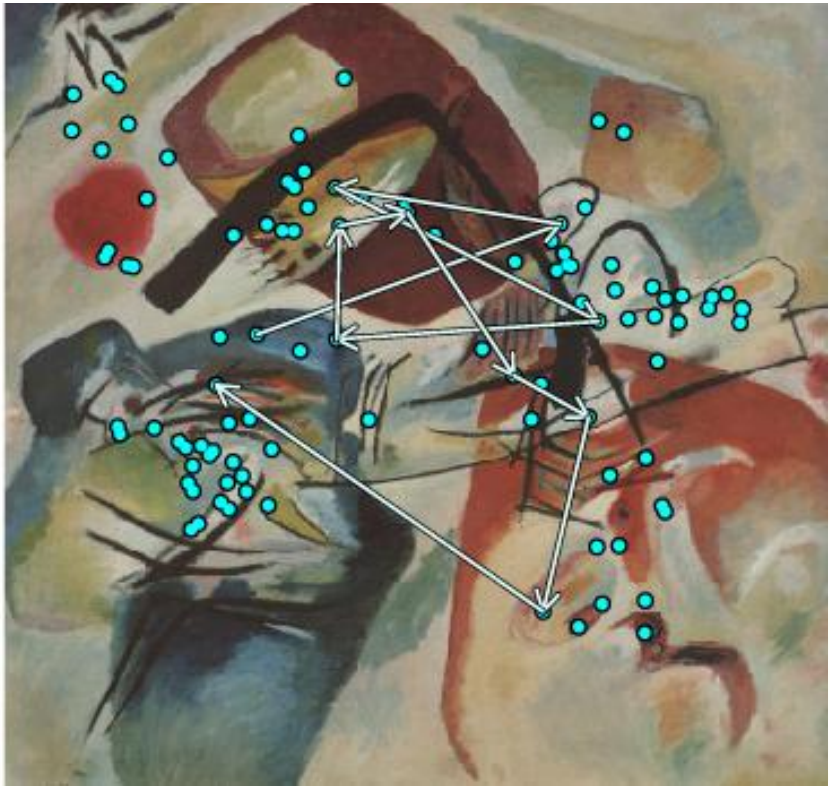
*Meyer & Pebesma (2021) *Methods in Ecology and Evolution*

Model comparison

- As « surveillance model » we randomly select α% of the observations in Occitania
- We test on the remaining set in Occitania



Modeling the surveillance process



First 100 fixation points

Movements during the first 3 Seconds

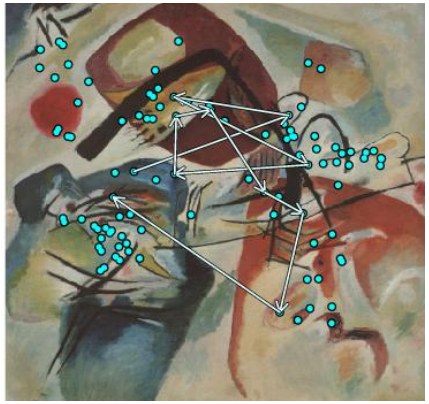
Wassily Kandinsky (1912)

The observer first takes in the entire scene, then gradually shifts focus to specific details.

Sequential Surveillance Scheme

Eye movement modeling

- **Spatial heterogeneity of fixation points**
  linked to the features of the target space
- **Dynamic contextuality**
  length of the jump between two points
- **Learning effect**
  time-dependent behavior reflecting self-interaction



Penttinen & Ylitalo (2016)

Surveillance design modeling

- **Spatial heterogeneity**
  underlying spatial risk of disease
- **Short term dependency**
  location of the last sample taken
- **Self-interaction**
  information already collected in the surrounding area

Sequential point process

For a sequence $\overrightarrow{s_k} = (s_1, \ldots, s_k)$ of ordered points in $W \subset \mathbb{R}^2$, each point s_i depends on all preceding points $\overrightarrow{s_{i-1}}$.

The density f of a sequence is defined as: $f(\overrightarrow{s_k}) = f_1(s_1) \prod_{i=1}^{k-1} f_{i+1}(s_{i+1}|\overrightarrow{s_i})$.

with

$$f_{i+1}(s|\overrightarrow{s_i}) \propto \alpha_r(s) K(s_i, s) \pi(\iota(s))$$

Risk map

prior knowledge or estimation of disease risk

Proposal kernel

favors nearby points in the sequence

$$K(s_i, s) \propto e^{-\frac{1}{2\sigma^2}\|s_i - s\|^2}$$

Reweighting function

modifies the sampling intensity based on the local information level *ι(s)*

$$\pi(p) \propto p^a(1-p)^b, a, b \in \mathbb{R}$$

Marked sequential point process

We have a **marked ordered sequence** $\overrightarrow{P_k} = ((s_1, m_1), \ldots, (s_k, m_k))$,

with binary marks $M(s_{k+1}) = \mathcal{B}(\alpha(s_{k+1}))$ given the status of the disease,

α being the true prevalence map.

Its **density** is: $g(\overrightarrow{P_k}) = g_1(P_1) \prod_{i=1}^{k-1} g_{i+1}(s_{i+1}, m_{i+1} | \overrightarrow{P_i})$,

with $g_{i+1}(s_{i+1}, m_{i+1} | \overrightarrow{P_i}) = f_{i+1}(s_{i+1} | \overrightarrow{P_i}) M(s_{i+1})$ and $f_{i+1|\overrightarrow{P_i}}(s) \propto \alpha_r(s) k(s_i, s) \pi(\iota_{\overrightarrow{P_i}}(s))$.

The **information map**

$$\iota_{\overrightarrow{P_i}}(s) = \frac{1}{2} + \frac{\sum_{k=1}^{i} (2m_k - 1) \exp\left(-\frac{||s_k - s||^2}{2h^2}\right)}{2 \sum_{k=1}^{i} \exp\left(-\frac{||s_k - s||^2}{2h^2}\right)}$$

is a kernel-weighted average of past samples' infection statuses, smoothed over space.

Optimal design for prevalence estimation

The **optimal surveillance design** P_k is defined by

$$\underset{P_k \in \mathcal{S}}{\operatorname{argmin}} \left\{ \mathrm{IBV}(\iota_{P_k}) \text{ while maximizing } \sum_{i=1}^{k} m_i \right\}$$

where $\mathcal{S}$ is the set of surveillance schemes and $\mathrm{IBV}(\iota_{P_k}) = \int_W \iota_{P_k}(s)(1 - \iota_{P_k}(s))du$ is the Integrated Bernoulli Variance (IBV) of the information map

The function $\pi(\iota(s))$ is tuned to prioritize high-uncertainty areas: $\pi_{opt}(p) = \mathbb{I}_{1/2}(p)$.

Then, we get the **prevalence estimate** as

$$\alpha_{P_k}(s) = \frac{\sum_{i=1}^{n} K(s_i, s) m_i}{\sum_{i=1}^{n} K(s_i, s)}, \quad \forall s \in W.$$

What next?

Combine the two!

Algorithm 1 Sequential Surveillance with Adaptive Risk Estimation

```
Initialize dataset D ← ∅
Initialize ML model M
Choose initial locations s⃗_0
for k = 1 to N do
    Step 1: Fit ML model on current data D
        M ← fit(D)
        α_r(s) ← M.predict(s) for all s ∈ W
    Step 2: Compute information map ι(s) from D
    Step 3: Compute Integrated Bernoulli Variance (IBV)
        IBV ← ∫_W ι(s)(1 − ι(s)) ds
    Step 4: Optimize re-weighting function π(ι(s))
    Step 5: Define sampling density
        f(s) ∝ α_r(s) · K(s_{k−1}, s) · π(ι(s))
    Step 6: Select next locations s_k
        s_k ← sample(f(s))
    Step 7: Observe infection status m_k
        D ← D ∪ {(s_k, m_k)}
end for
```

Learning the risk
and its determinants

Planning data collection
from the risk evaluation