Predicting the risk of establishment of the invasive beetle *Popillia japonica* in Europe

Davide MARTINETTI – INRAE, UR BioSP – Avignon

Theory-driven Analysis of Ecological Data - FRB-CESAB

Montpellier, April 5th, 2023

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

INRAE

Biostatistique BIO/Π & Processus Spatiaux

IPM Popillia
Integrated Pest Management of Japanese Beetle

Outline

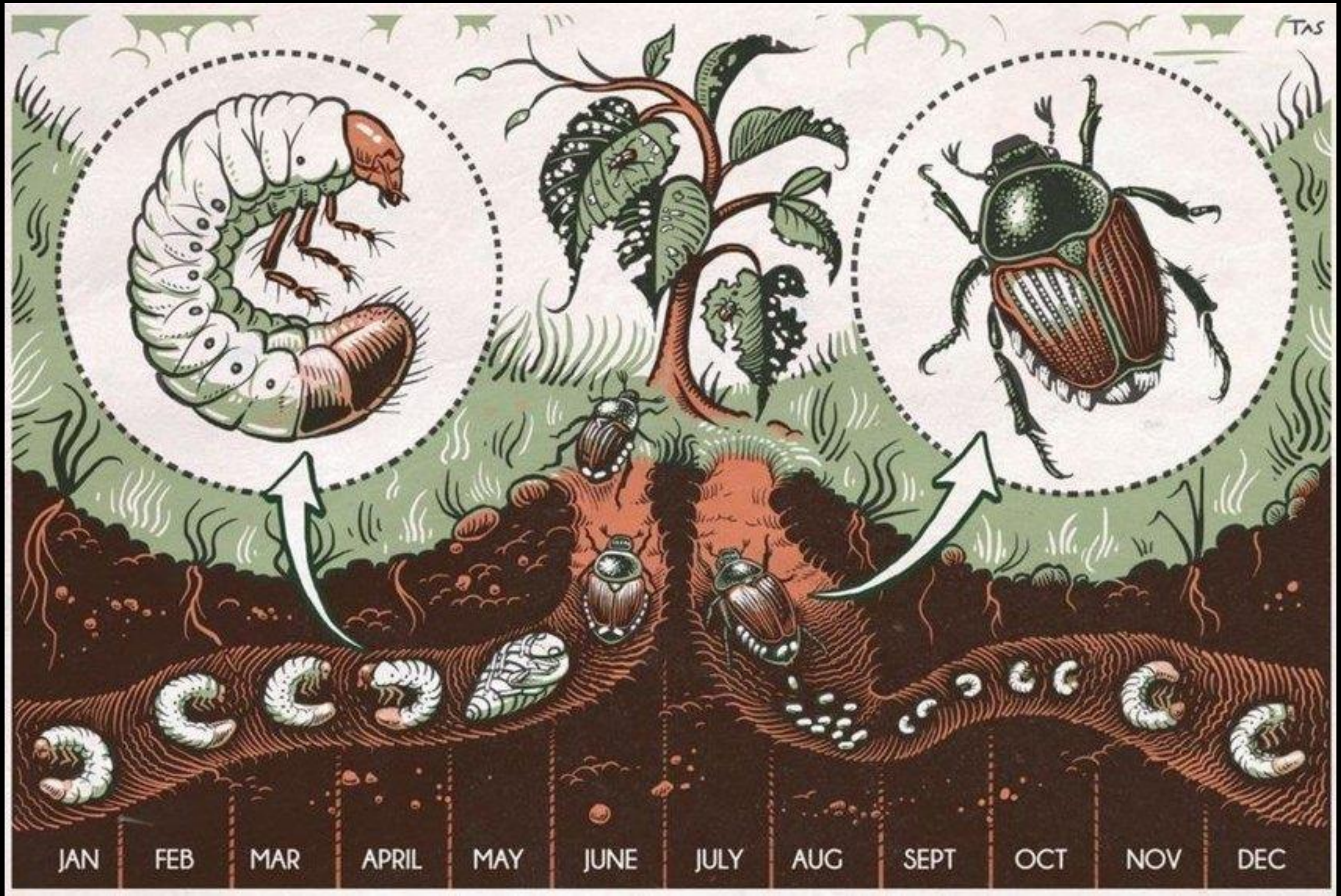
Theory-driven Analysis of Ecological Data

More serious outline

1. The 4 ”W” of *Popillia japonica*
    - Who? Where? When? Why?
2. Species distribution model with opportunistic citizen-science data
    - Presence-only data
    - Opportunistic data
    - SDM
    - Results
3. A reaction-diffusion model and its observation process
    - The mechanistic model
    - The statistical model

Who

Popillia japonica



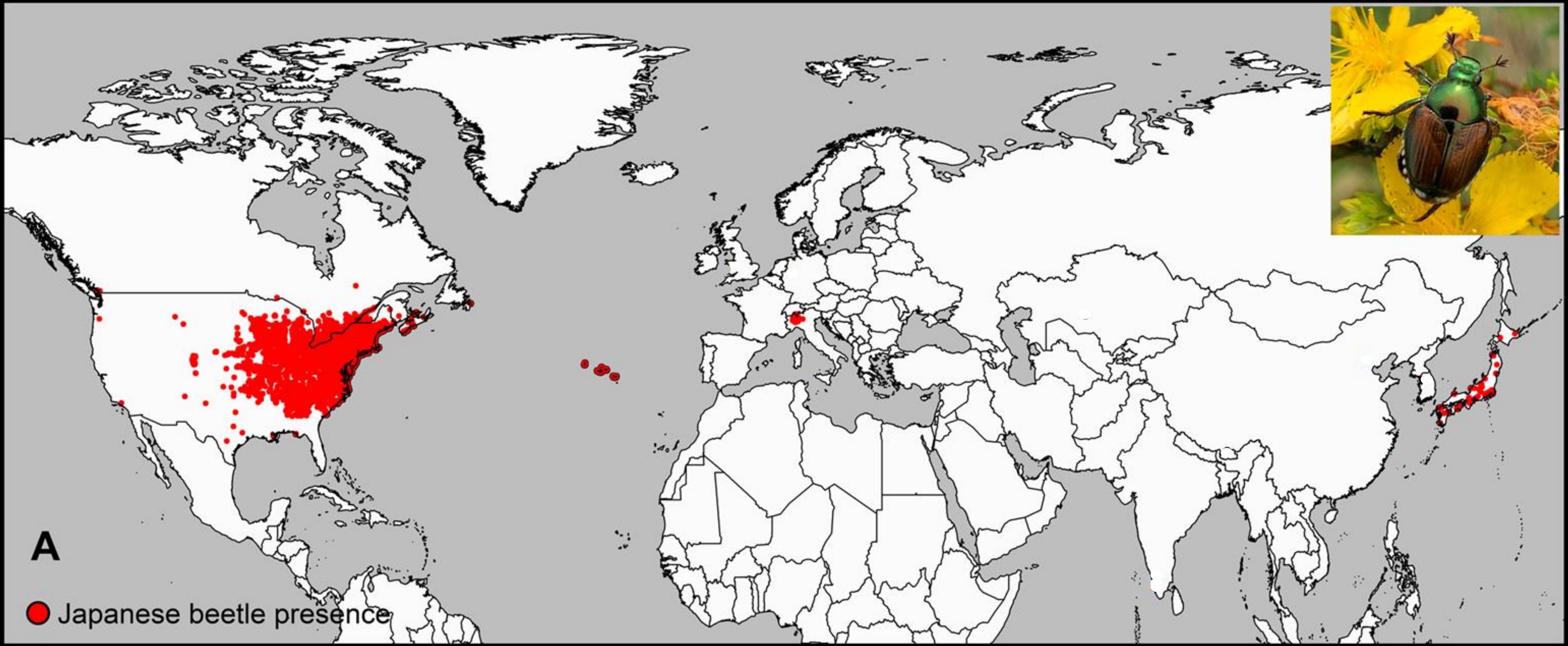
Japanese beetle

Scientific classification	
Kingdom:	Animalia
Phylum:	Arthropoda
Class:	Insecta
Order:	Coleoptera
Family:	Scarabaeidae
Genus:	*Popillia*
Species:	***P. japonica***

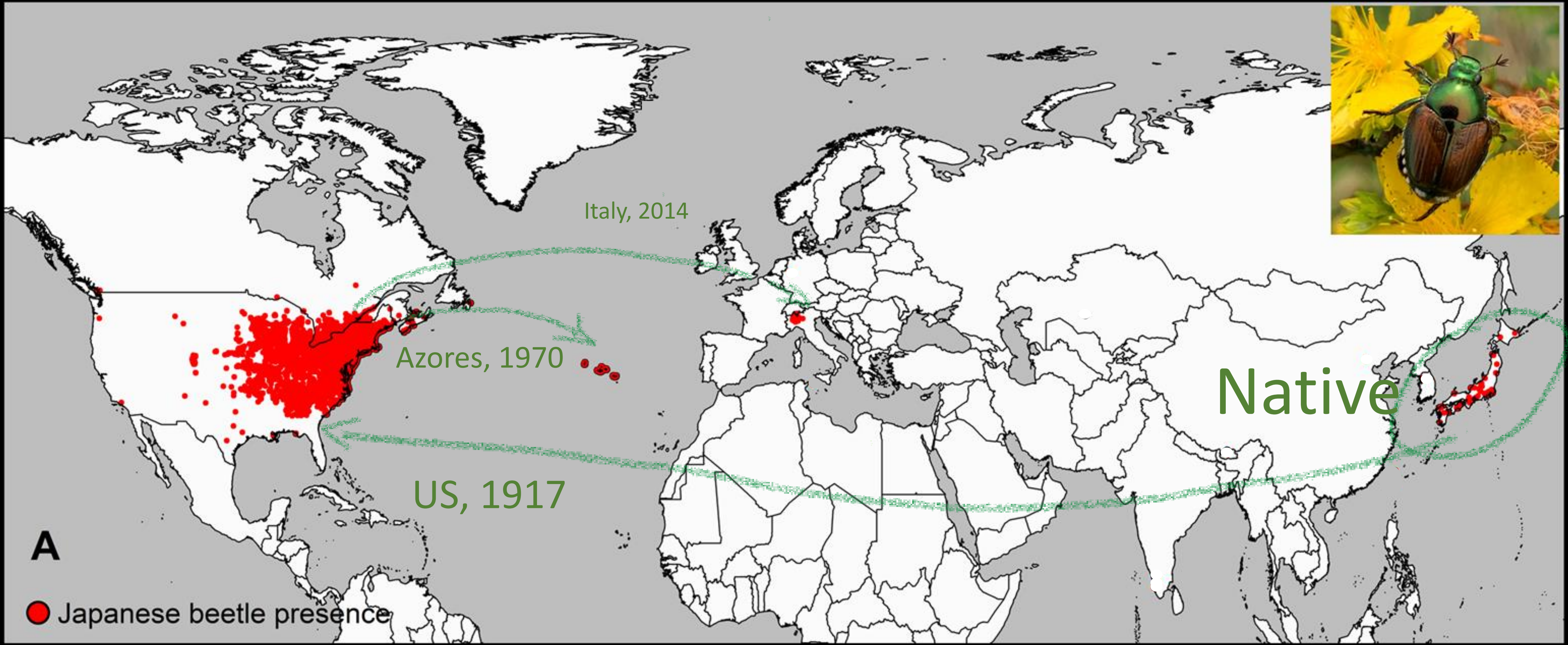
Binomial name

Popillia japonica
Newman, 1841

Where

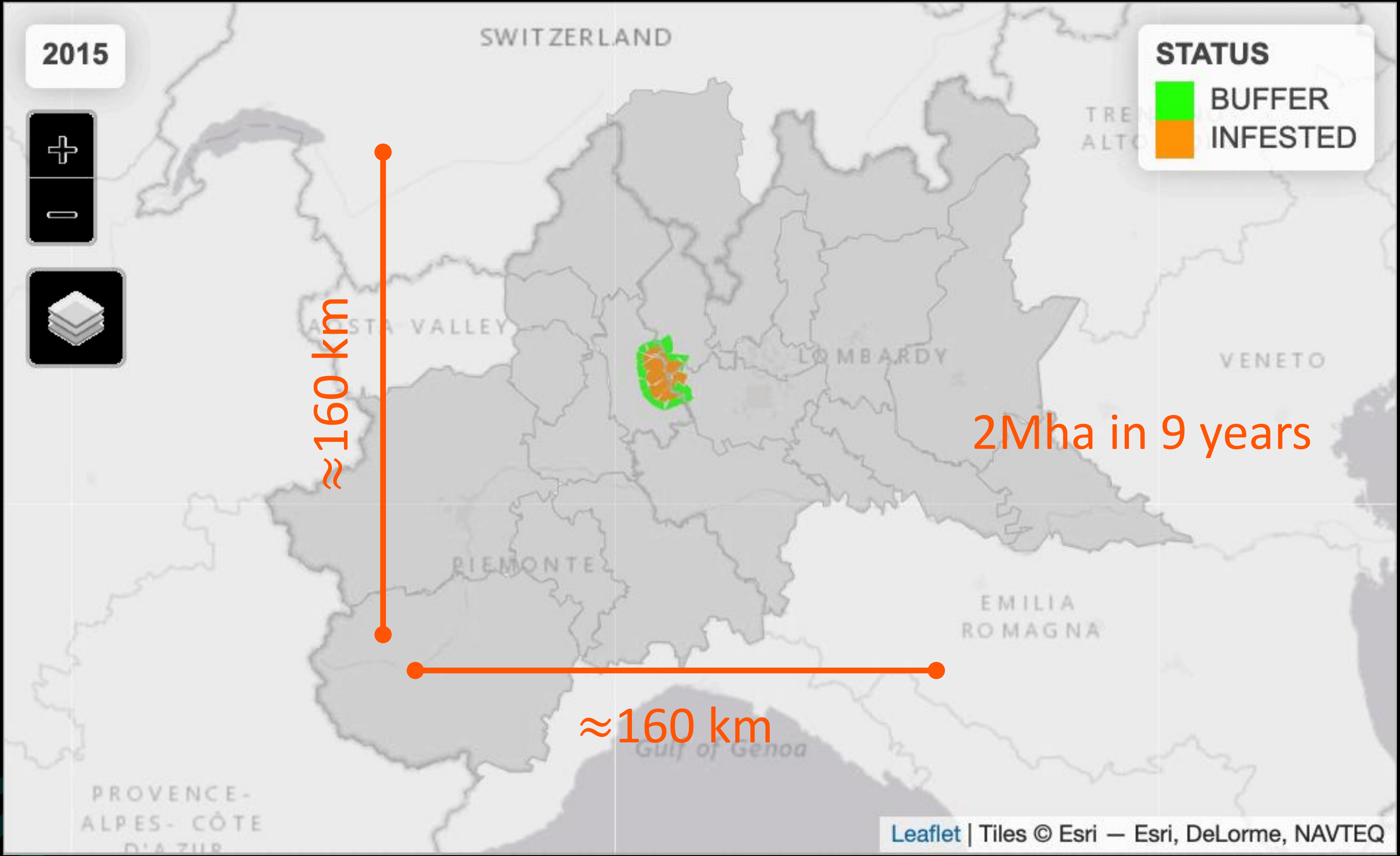


When


Italy, 2014
Azores, 1970
Native
US, 1917
A
Japanese beetle presence

Why

Italy, July 2021


2015
SWITZERLAND
STATUS
BUFFER
INFESTED
AOSTA VALLEY
LOMBARDY
VENETO
PIEMONTE
EMILIA ROMAGNA
Gulf of Genoa
PROVENCE-ALPES-CÔTE D'AZUR
≈160 km
≈160 km
2Mha in 9 years
Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ

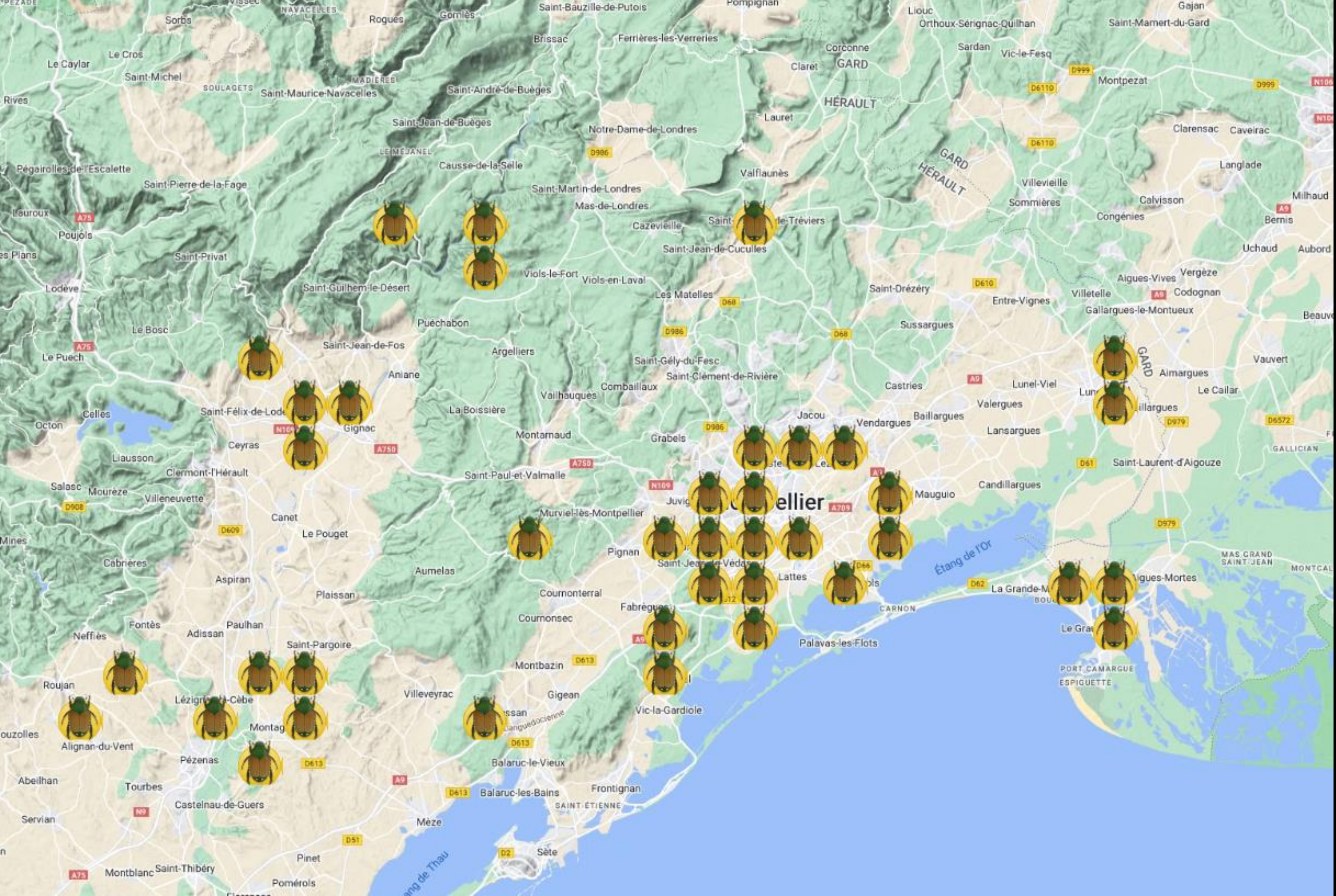
Surveillance & containment strategies

- React fast
- Detect as early as possible
- Eradicate when possible
- Constraints
    - Money
    - Time
    - Coverage

OPPORTUNISTIC CITIZEN-SCIENCE DATA

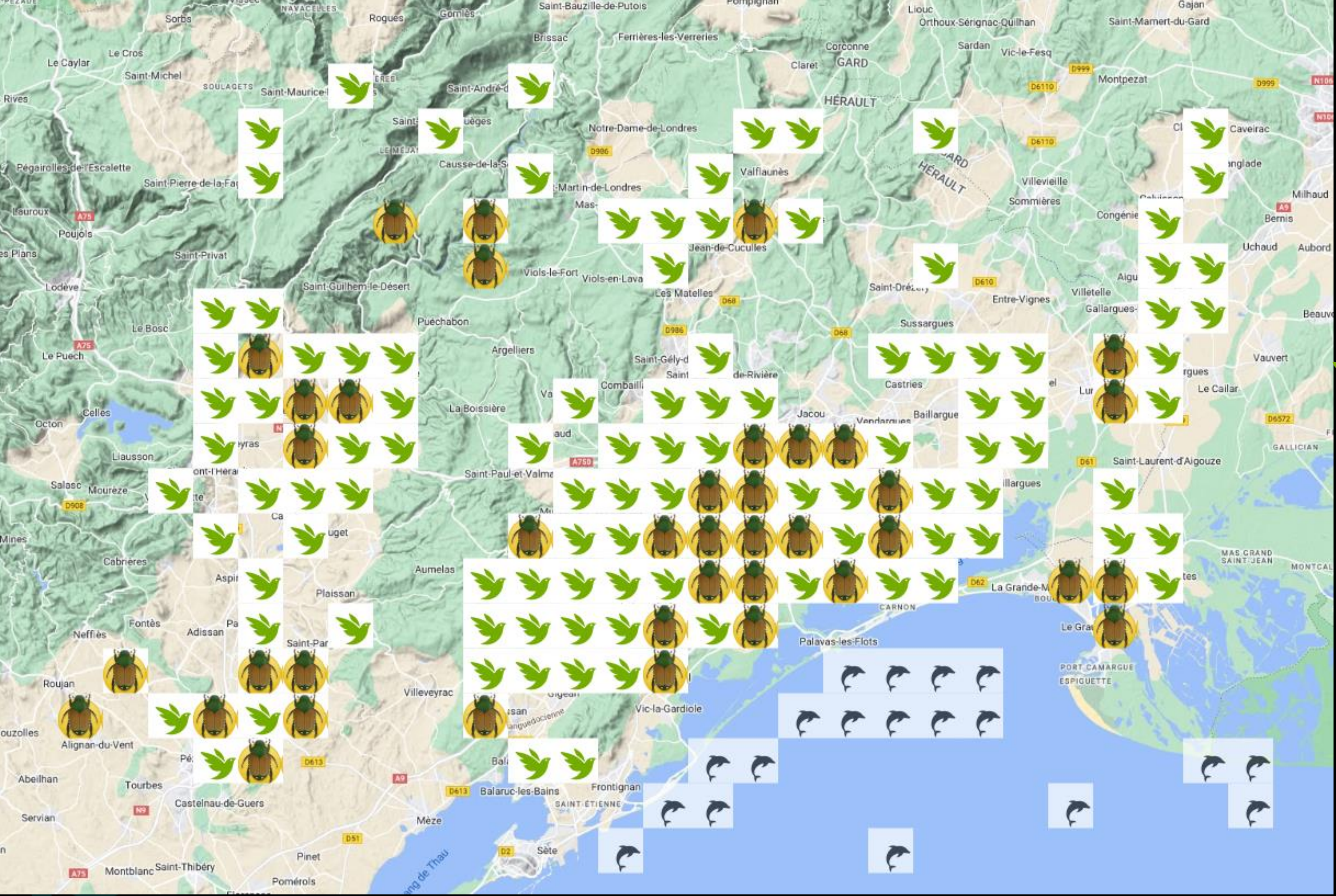
DEMO

OPPORTUNISTIC DATA



Legend

Presence

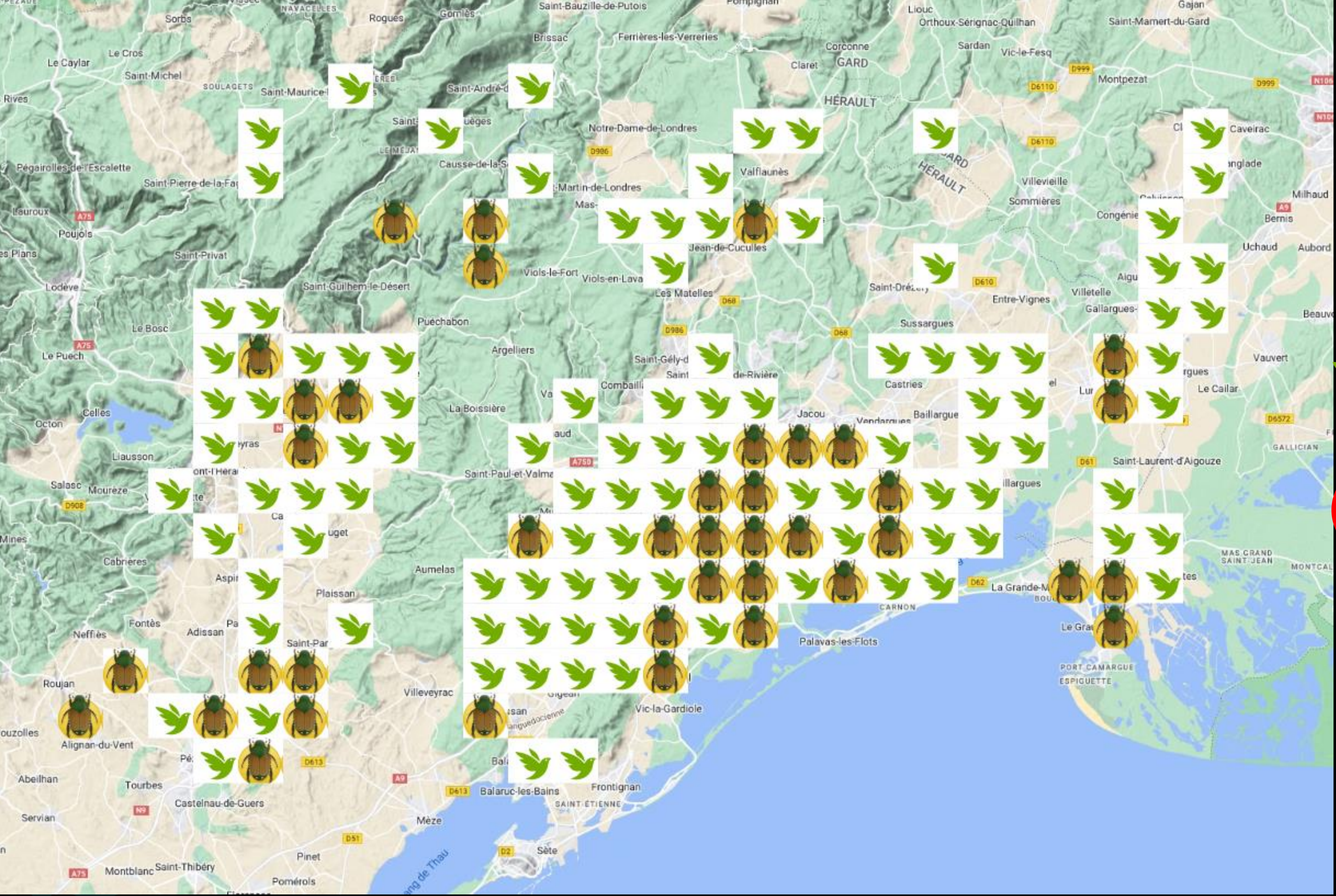


Legend

Presence

Terrestrial

Marine

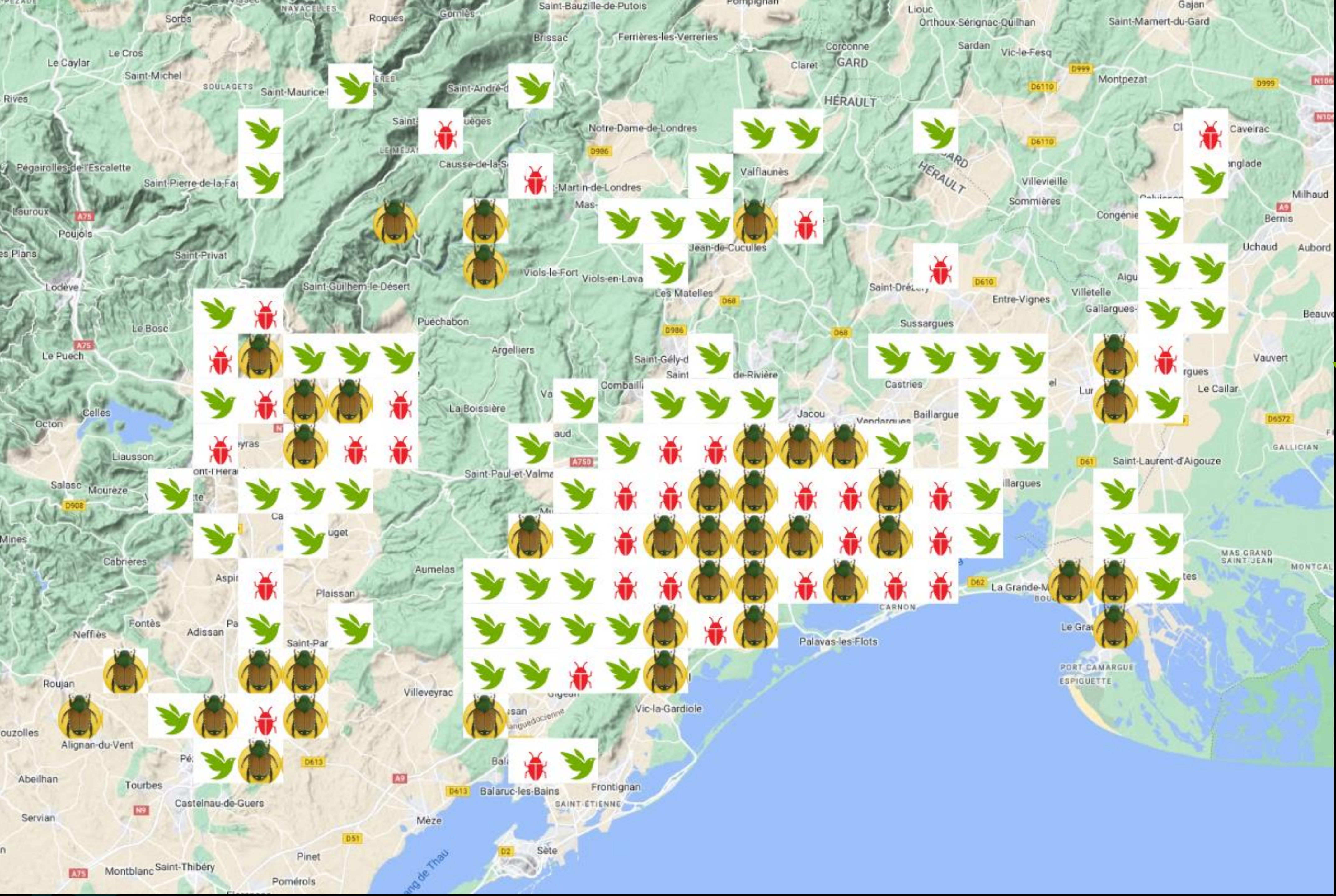


Legend

Presence

Terrestrial

Marine

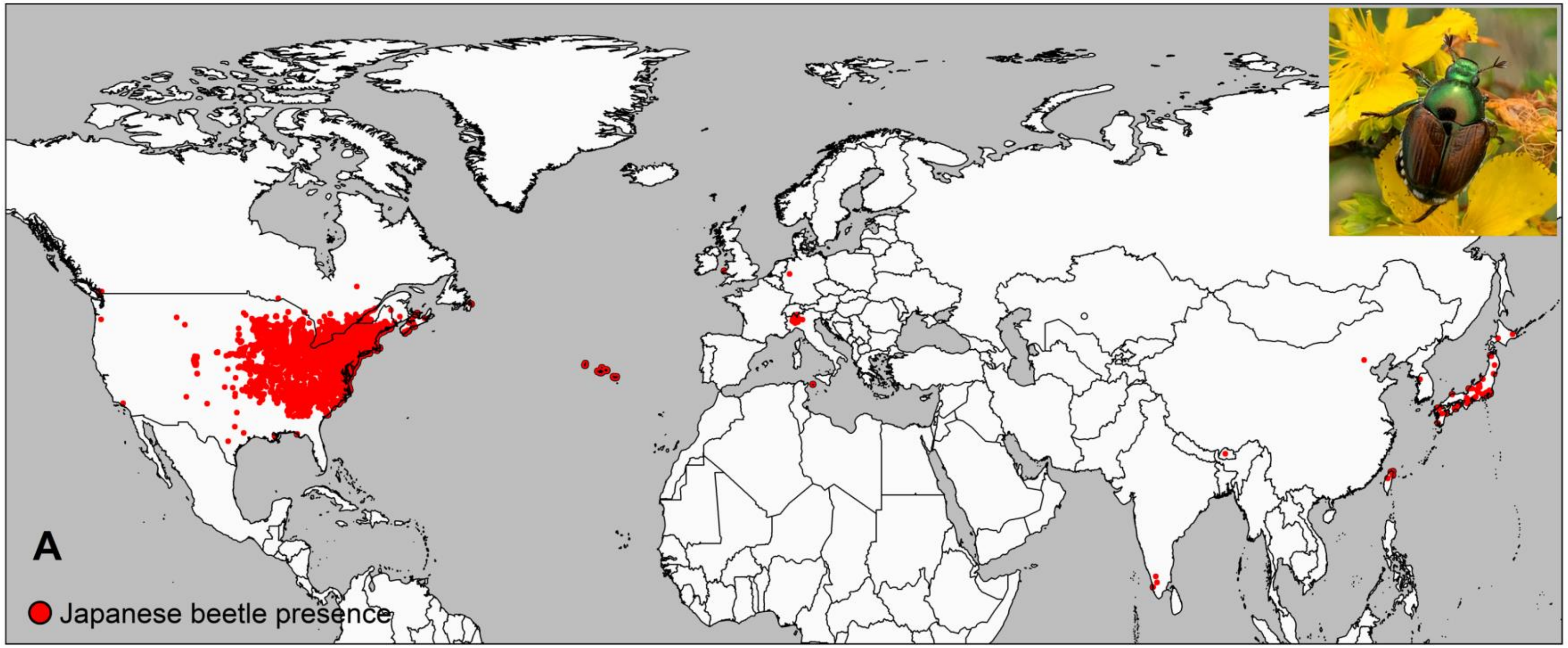


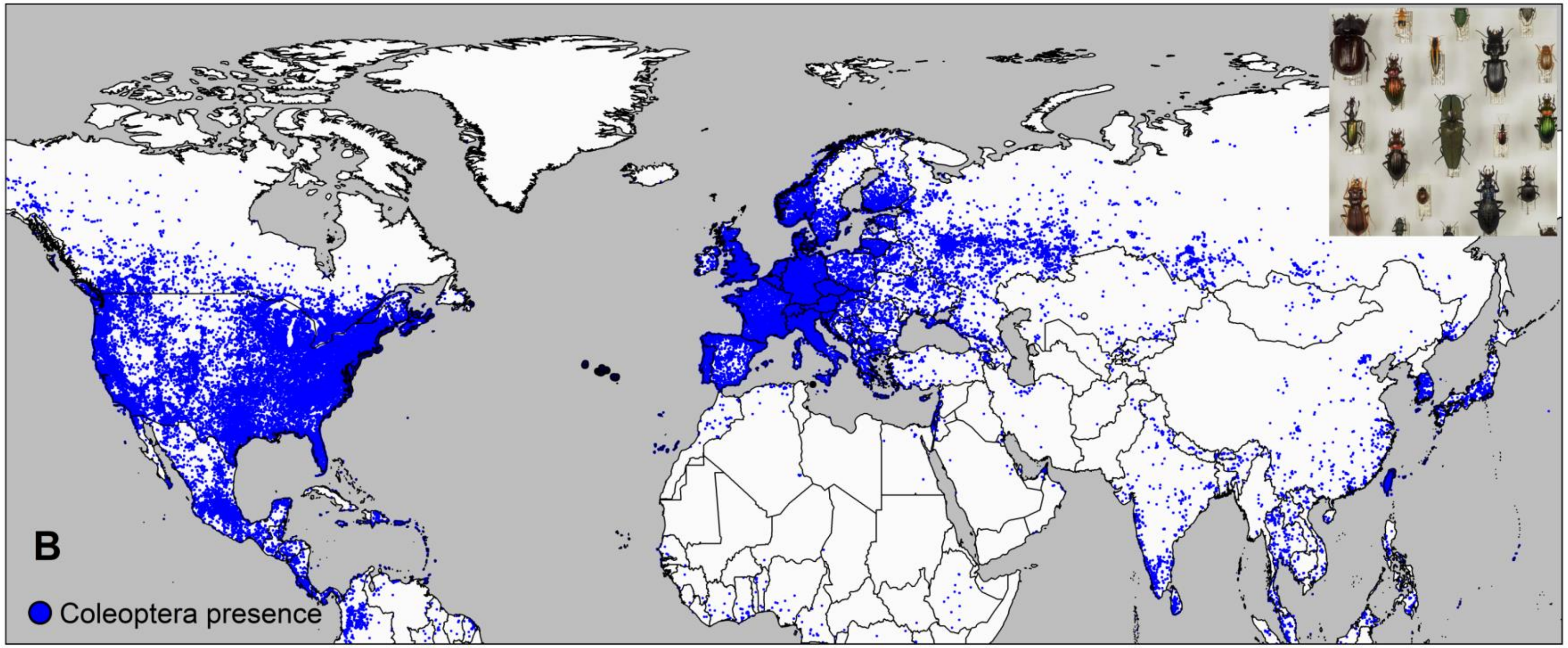
Legend

Presence

Terrestrial

Insects





Presences

(Popillia japonica)

***6844* cells**

much less then

Pseudo-absences

(Coleoptera)

***49010* cells**

Take-home message

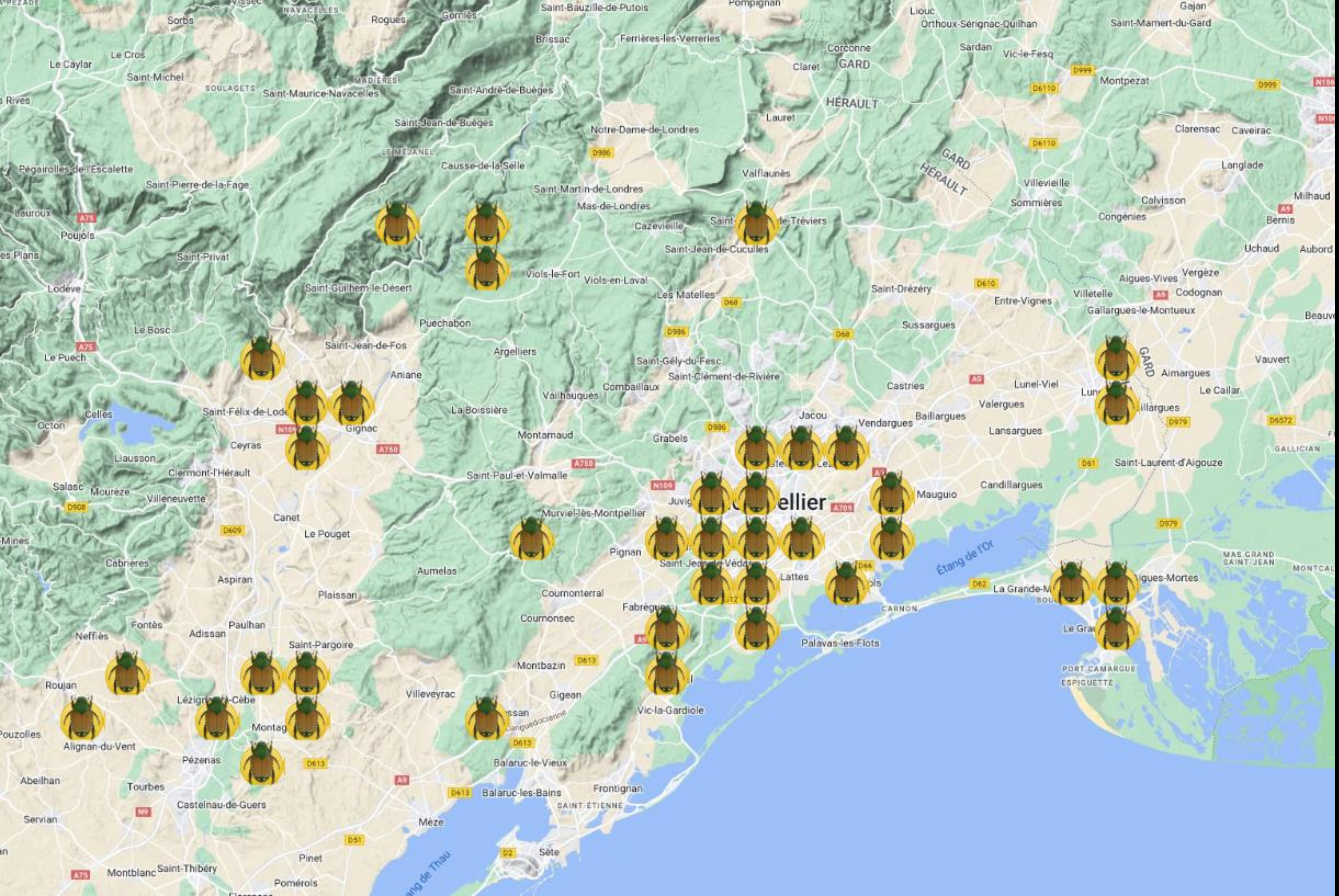
- **Opportunistic data are abundant and ready to use…**
- **… but suffer from sampling bias**

Solution: Pseudo-absences using **target-group**[1] strategy

- Higher taxonomic level
- Same observers
- Same dates/period

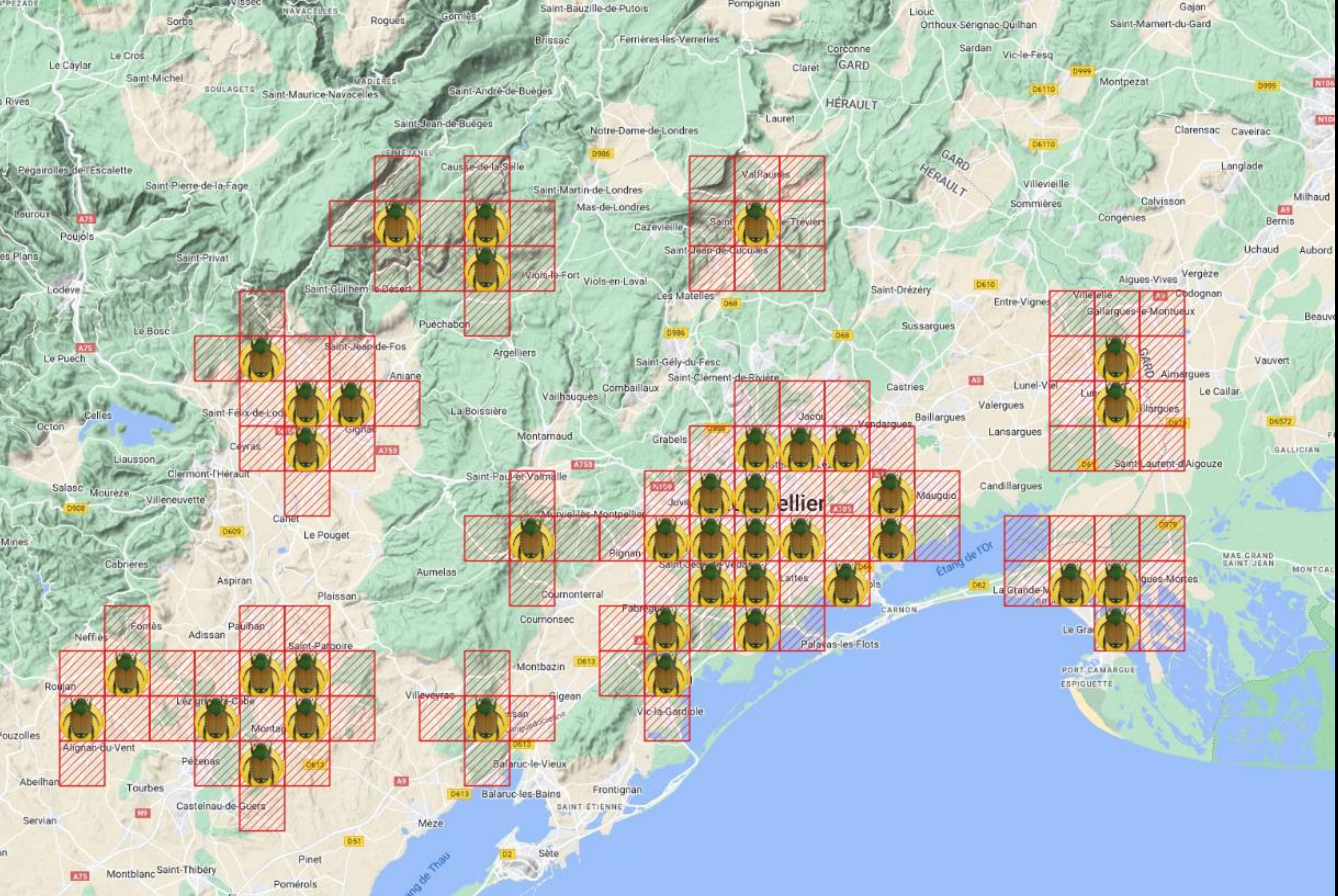
[1]Phillips *et al*. (2009)

PRESENCE-ONLY DATA



Legend

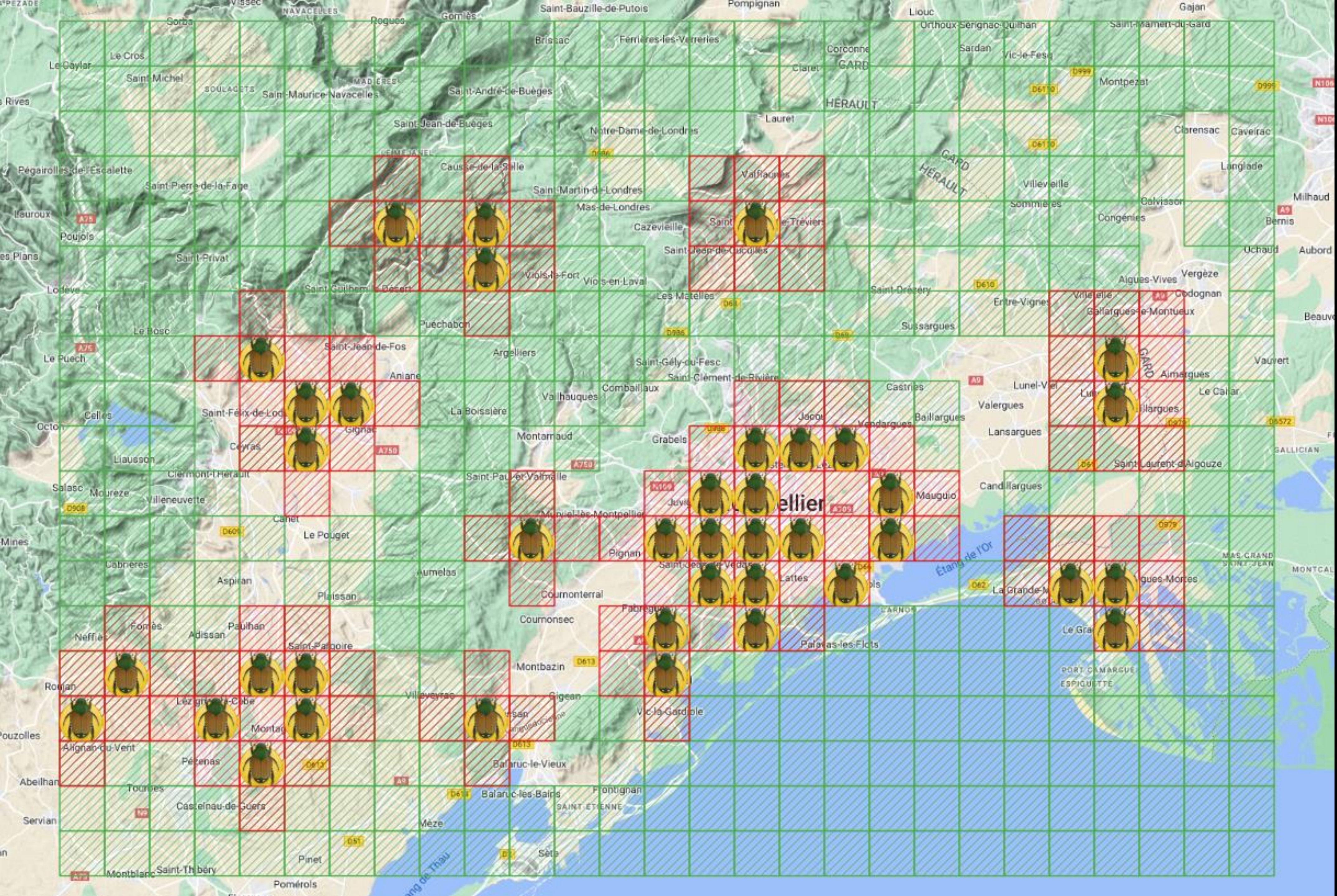
Presence



Legend

Presence

Neighbour



Legend

Presence

Neighbour

Pseudo absence

Take-home message

1. **You may trust presence data…**
2. **…but generate pseudo-absences wisely**

Barbet-Massin *et al.* (2012)
Valavi *et al.* (2022)

SPECIES DISTRIBUTION MODEL

Species Distribution Models

$$Y = f(X, \epsilon)$$

- $\mathrm{Y} \in \{0,1\}$: presence or (pseudo-)absence of a certain species
- $\mathrm{X} \in \mathbb{R}^n$: covariates
- ε : some kind of error
- $f : \mathbb{R}^n \to [0,1]$: some kind of function

Covariates

Climatic variables

Land use

Human footprint

Vegetation

Soil

133 variables

All my data

133 variables

55854 observations

Presence	Var_1	Var_2	…	…	…	Var_132	Var_133
Yes							
No							
Yes							
…							
…							
…							
Yes							
No							

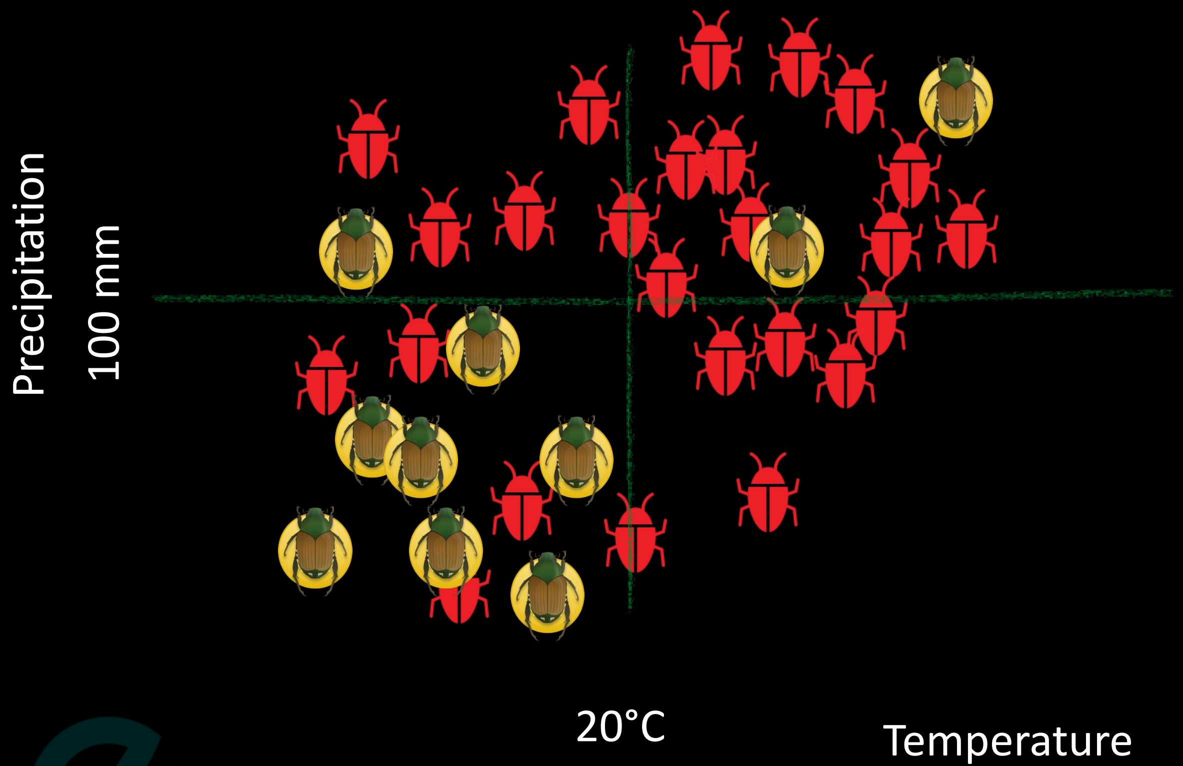
Choice of the algorithm

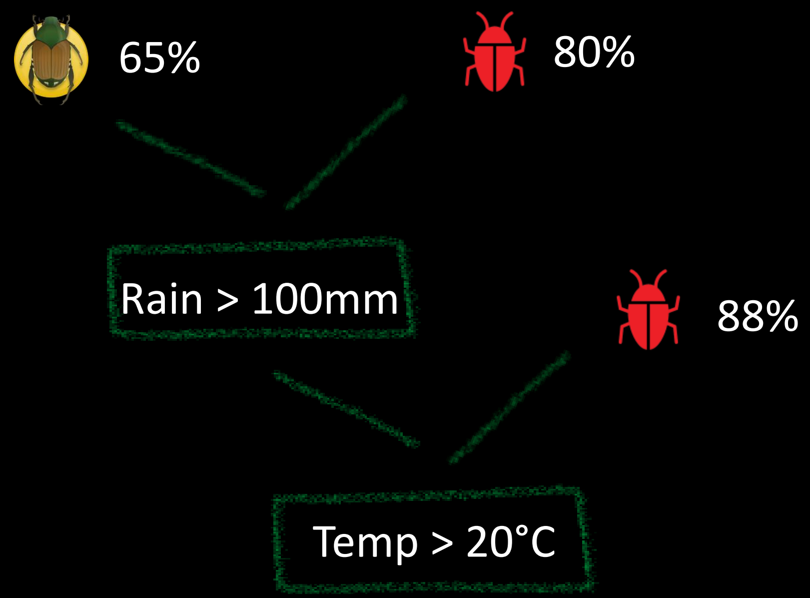
BIOCLIM = Bioclimatic Analysis
GLM = Generalized Linear Model
GAM = Generalized Additive Model
MARS = Multivariate Adaptive Regression Splines
BRT = Boosted Regression Tree
RF = Random Forest

Good for unbalanced datasets [1]
Estimation of variable importance [2]
Robust against multicollinearity [3]

[1] Barbet-Massin *et al*. (2012)
[2] Genuer *et al.* (2010)
[3] Freeman *et al.* (2016)

Random forest in a nutshell...



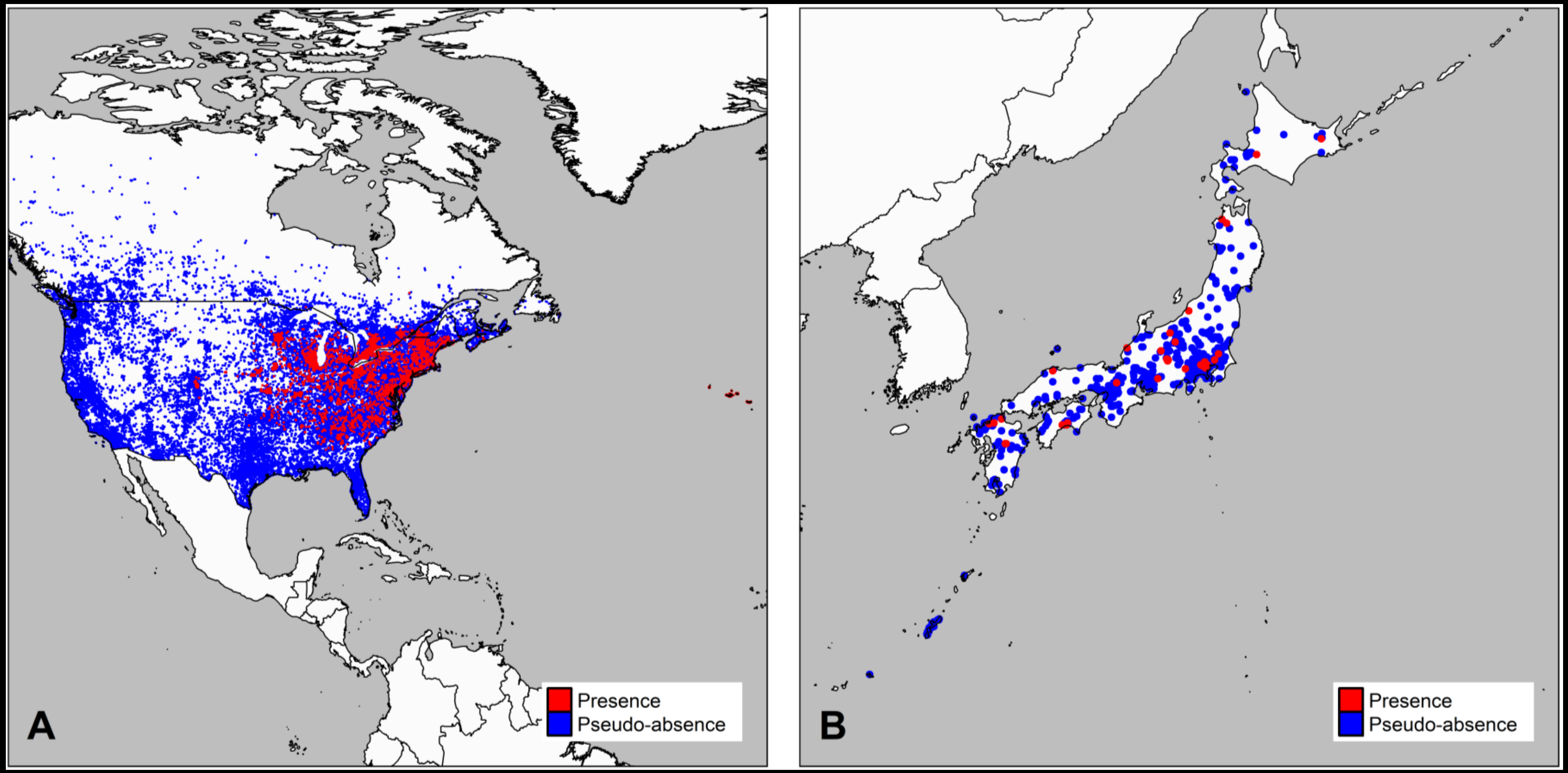


Random forest in a nutshell…

Presence	Var_1	Var_2	…	…	…	Var_132	Var_133
Yes							
No							
Yes							
…							
…							
…							
Yes							
No							

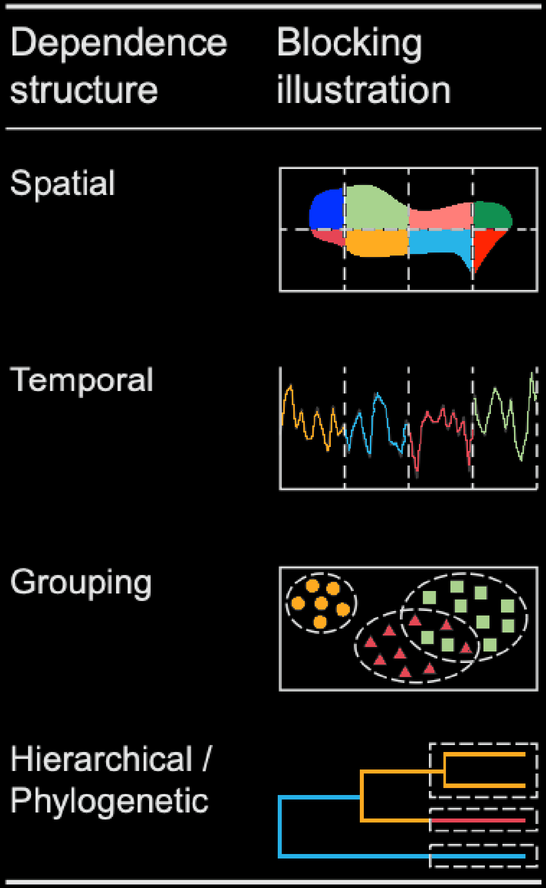
Model training

Train data from native and long-invaded regions since **newly invaded** regions may reflect dispersal limitations rather then real unsuitability



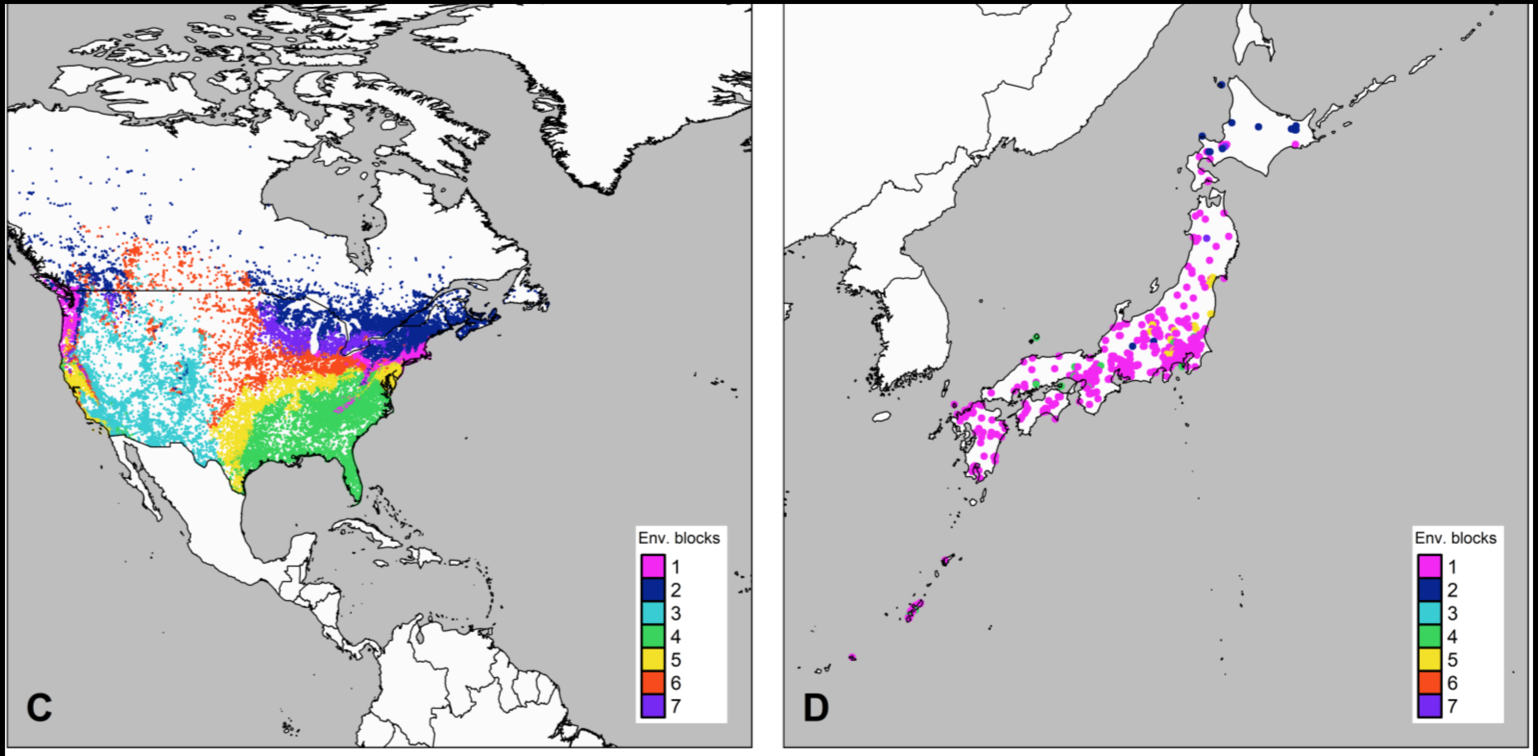
Elith *et al.* (2010)

Cross-validation strategy



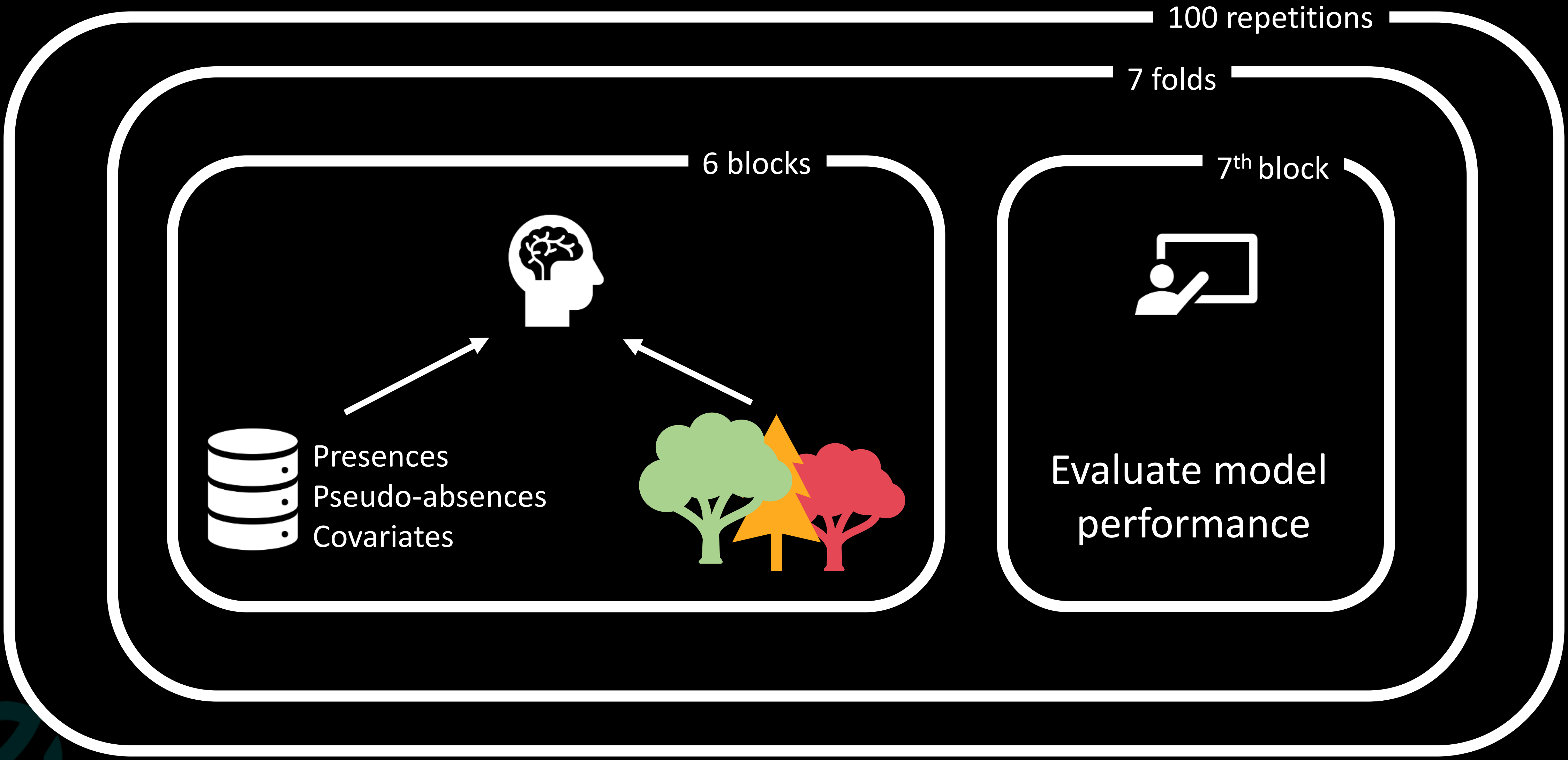
Roberts *et al*. (2017)

7 blocks according to environmental distance



Ploton *et al*. (2020)
Valavi *et al*. (2019)

Machine learning



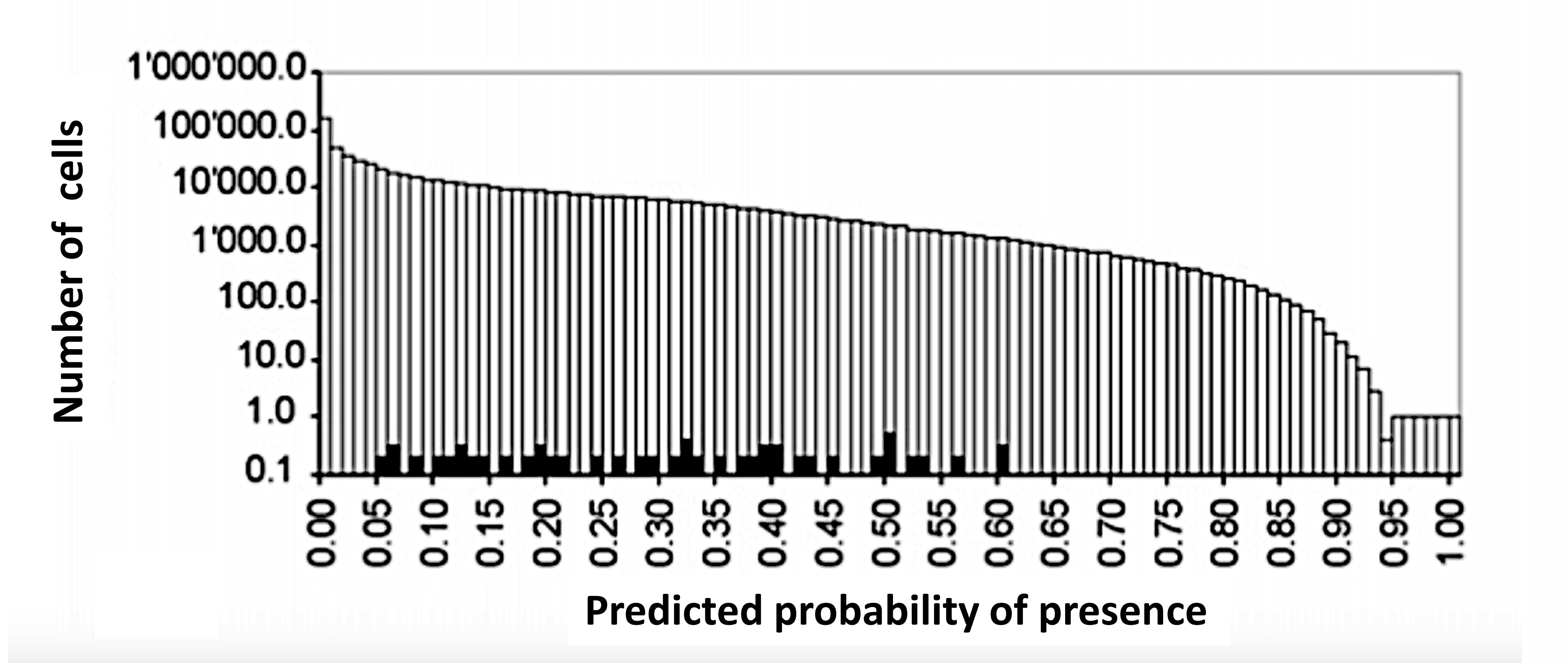
Predictions

How to go from probability in [0,1] to binary {0,1}?

Good model

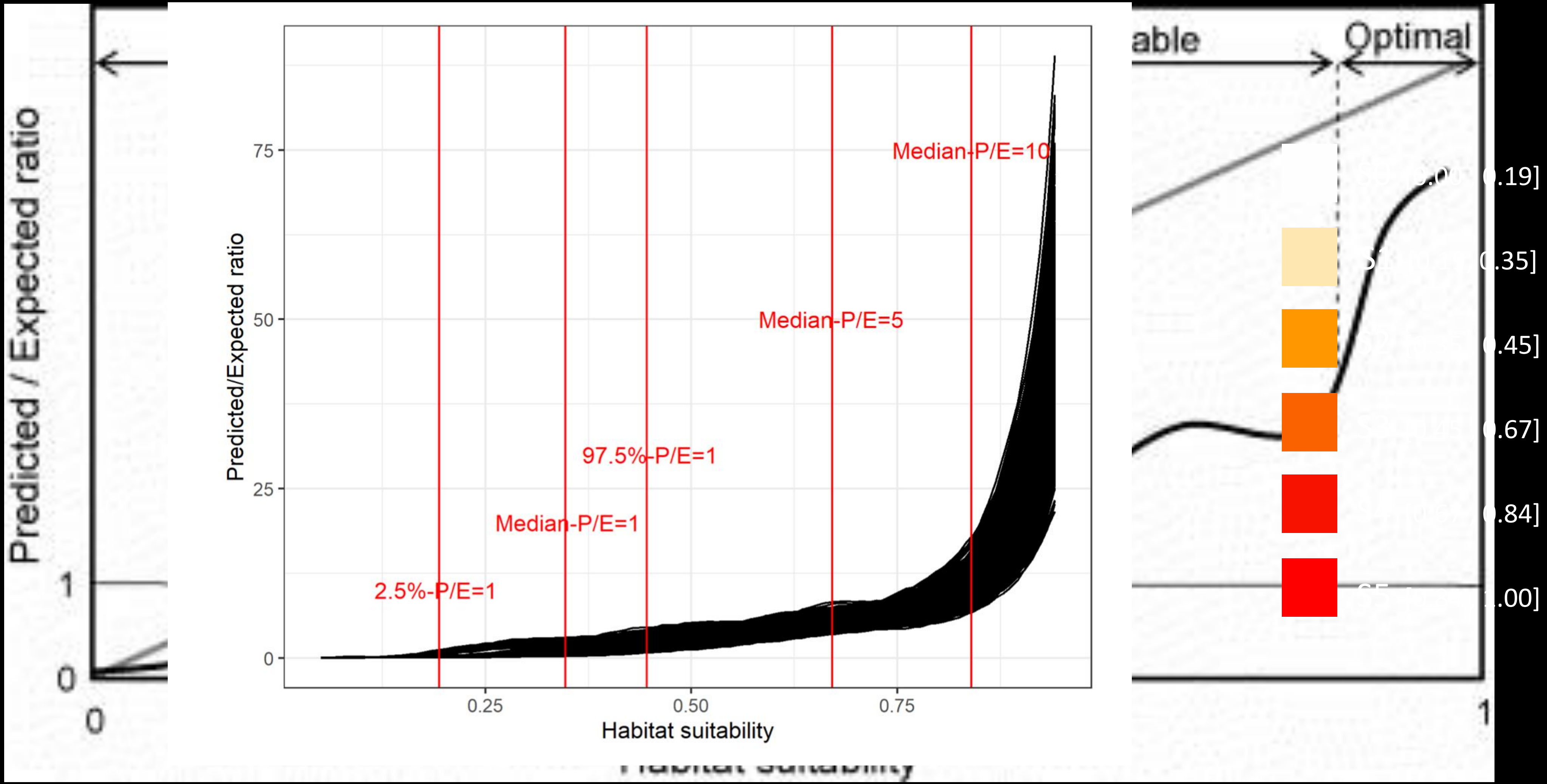
Random model

Bad model

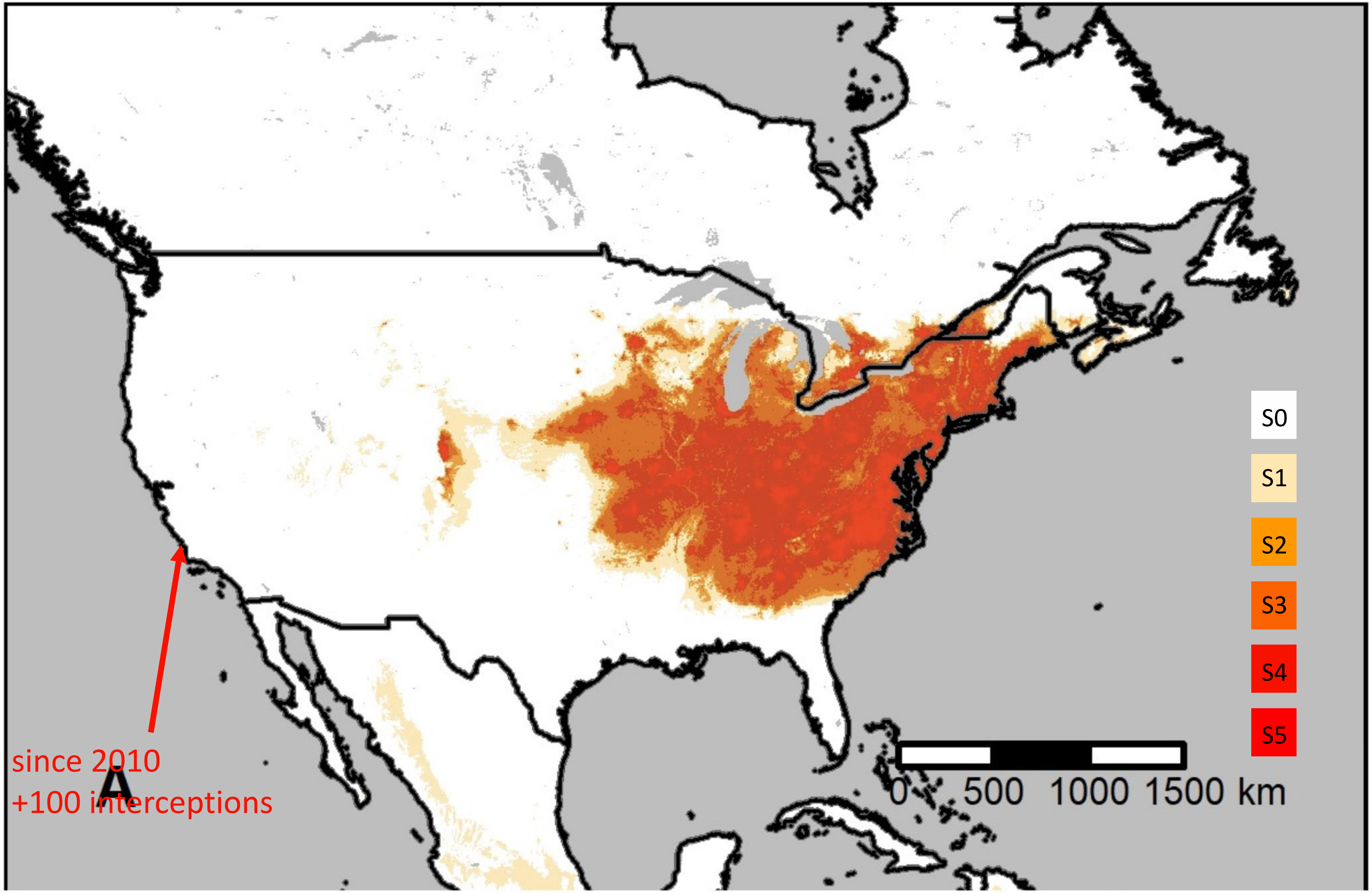


Hirzel *et al.* (2016)

Boyce Predicted to Expected ratio (P/E ratio)



Boyce *et al.* (2002)



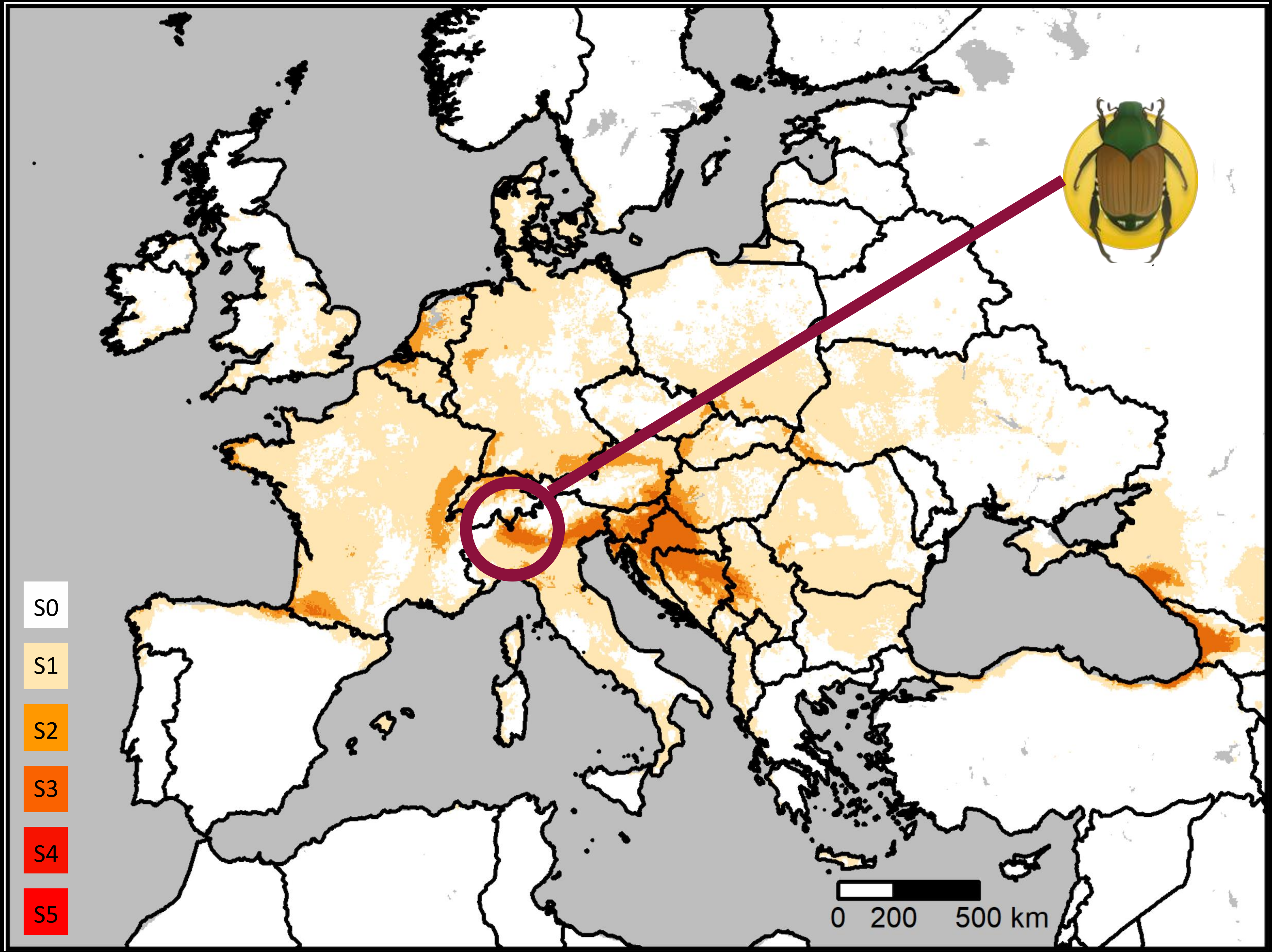
US official classification

× = Highly Infested

+ = Infested

⚠ = Quarantine

○ = Uninfested


S0
S1
S2
S3
S4
S5
0 200 500 km

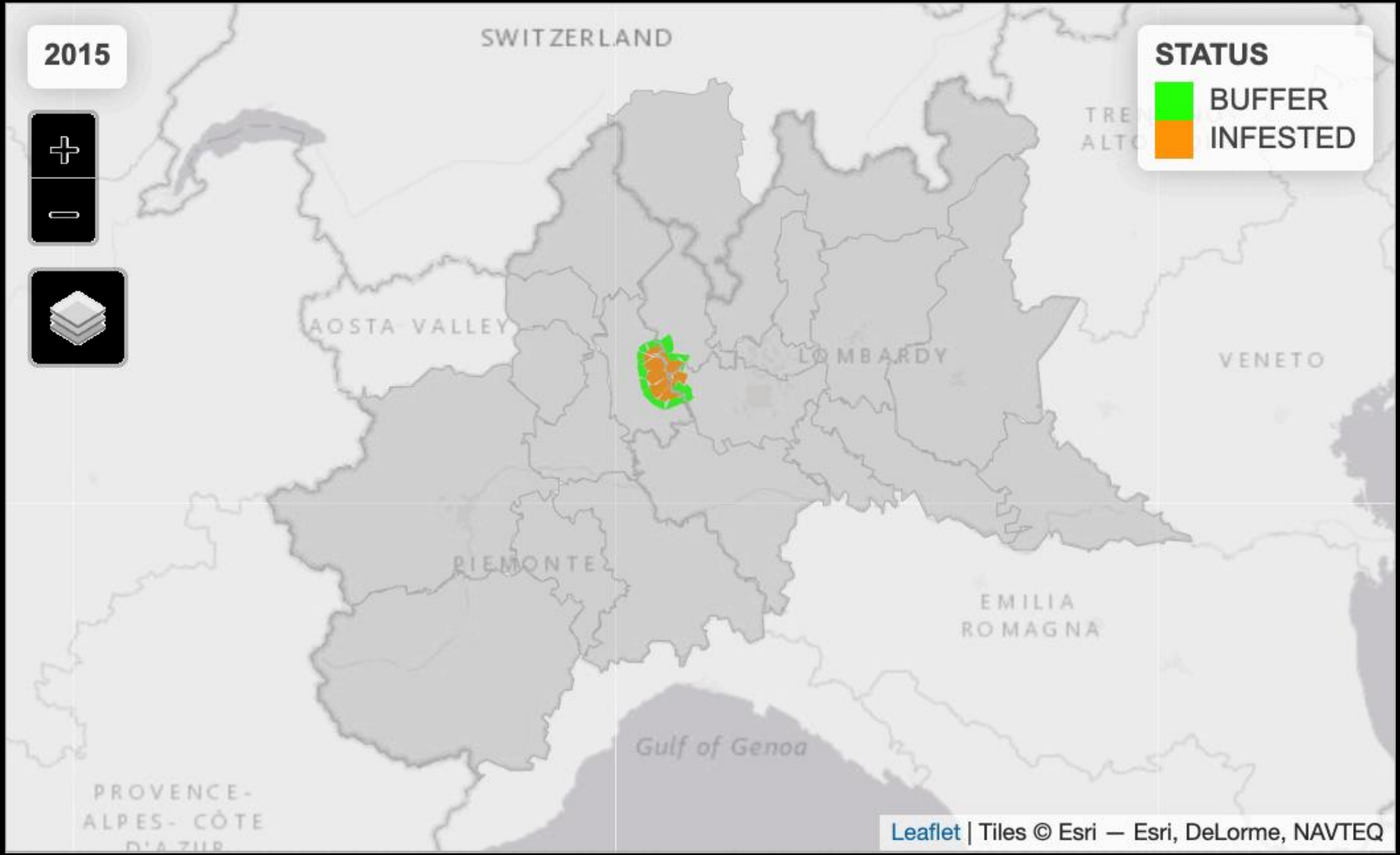
REACTION-DIFFUSION MODEL & OBSERVATION PROCESS

The reaction-diffusion equation

$$\frac{\partial V(x,y,t)}{\partial t} = DV(x,y,t) + R(x,y)V(x,y)$$
$$V(x,y,0) = I_{2015}$$

- $V(x,y,t)$ = concentration of PJ in (x,y) at time t
- D = diffusion coefficient
- $R(x,y) = -\frac{1}{\mu} + \sum_{i=0}^{5} \beta_i \, \mathbf{1}_i(x,y)$:
    - μ = life expectancy
    - β_i = birth rate depend on suitability class at location β_i

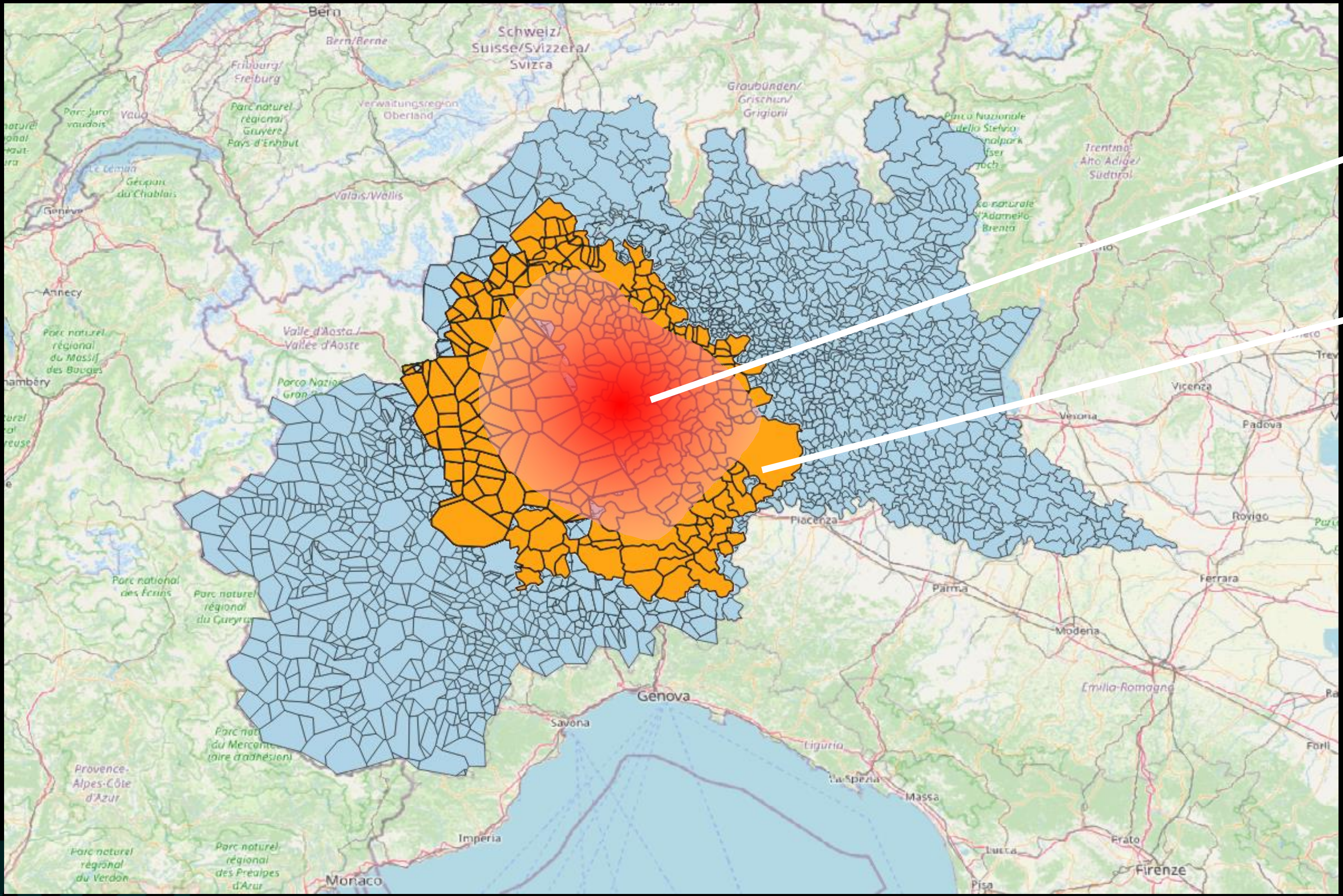
Observation process



Legend

- Administrative boundary
- Infested = at least 1 PJ found
- Buffer = <15km from infested

Parameter estimation



$V(\theta, t)$ for parameter θ at time t

$\theta = (D, \beta_i) =$ diffusion & birth rate

$\mathrm{O}(t)$ = observed presences at time t

Likelihood of θ

=

agreement btw $V(\theta, t)$ and $\mathrm{O}(t)$

Roquet & Bonnefon (2016)

Thanks

https://www.popillia.eu/

Leyli Borner
PostDoc, INRAE, UR IGEPP, Rennes

Sylvain Poggi
Researcher, INRAE, UR IGEPP, Rennes

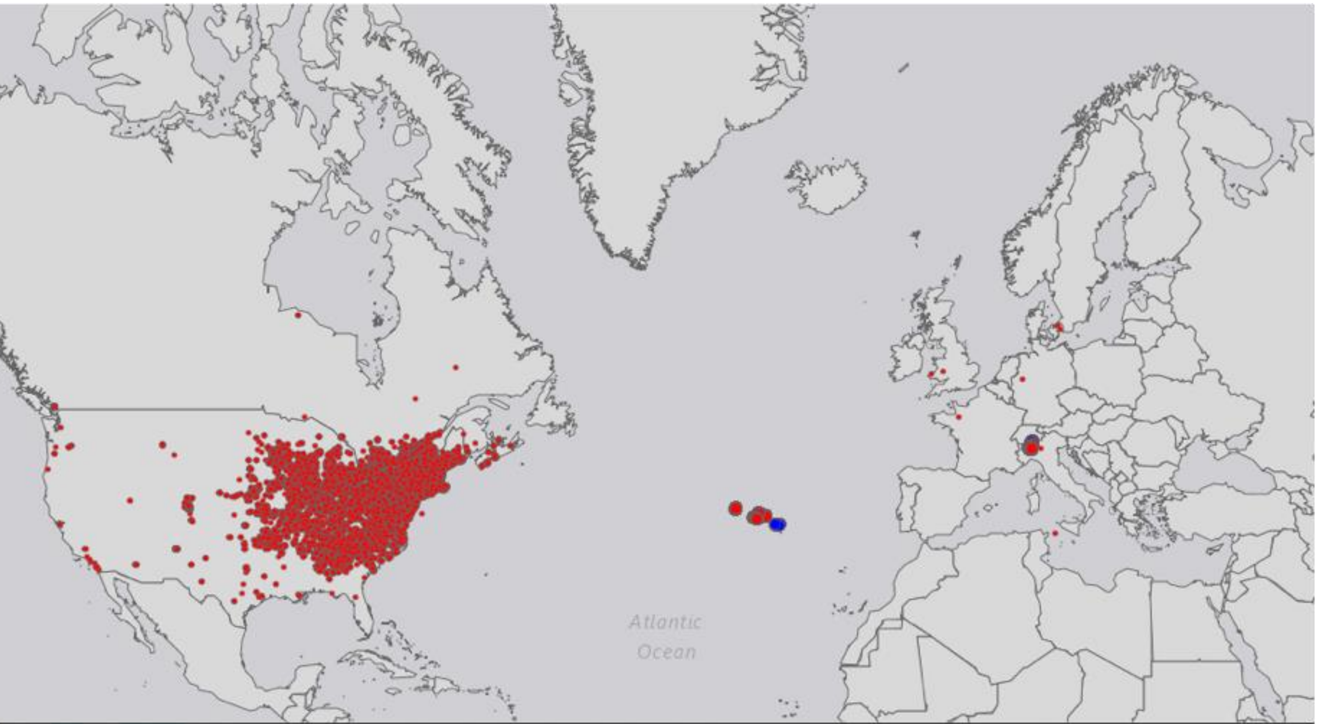
IPM Popillia
Integrated Pest Management of Japanese Beetle

References

1. **Barbet-Massin *et al*. (2012)**. *Selecting pseudo-absences for species distribution models: How, where and how many?*
2. **Boyce *et al. (2002)***. *Evaluating resource selection functions. Ecological modelling.*
3. **Elith *et al.* (2010)**. *The art of modelling range-shifting species.*
4. **Freeman *et al.* (2016)**. *Random forests and stochastic gradient boosting for predicting tree canopy cover: comparing tuning processes and model performance.*
5. **Genuer *et al.* (2010)**. *Variable selection using random forests.*
6. **Hirzel *et al*. (2006)**. *Evaluating the ability of habitat suitability models to predict species presences.*
7. **Phillips *et al*. (2009)**. *Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data*.
8. **Ploton *et al*. (2020)**. *Spatial validation reveals poor predictive performance of large-scale ecological mapping models.*
9. **Roberts *et al*. (2017)**. *Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure.*
10. **Roques & Bonnefon *(2016)***. *Modelling population dynamics in realistic landscapes with linear elements: A mechanistic-statistical reaction-diffusion approach.*
11. **Valavi *et al*. (2018)**. *blockCV: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models*.
12. **Valavi *et al.* (2021)**. *Predictive performance of presence-only species distribution models: a benchmark study with reproducible code.*

Presence data

	Official surveillance[1]	Citizen Science[2]	TOTAL
Europe	11,777	2,845	14,622
USA & Canada	962	29,498	30,460
TOTAL	**12,739**	**32,343**	**45,082**



Type of data	Count
Presence of PJ	4,206
No observation	9,126,667
TOTAL	**9,134,770**

Aggregated 4km

[1] From Italy, Switzerland, Portugal, Canada and US

[2] Including GBIF & iNaturalist web platforms *(as of November 2020)*

Pseudo-absence data: the target-group method

How to create absence data with the same sampling bias as presence data

Sampling bias in presence-only data from citizen science

- Bias towards of eye-catching, emblematic or newly-introduced species
- Positive bias towards urban & recreational areas and negative bias towards remote areas
- Lack of transect w.r.t. relevant bio-physical factors

Target group method (Ponder *et al.* 2001, Anderson 2003, Phillips *et al.* 2009)

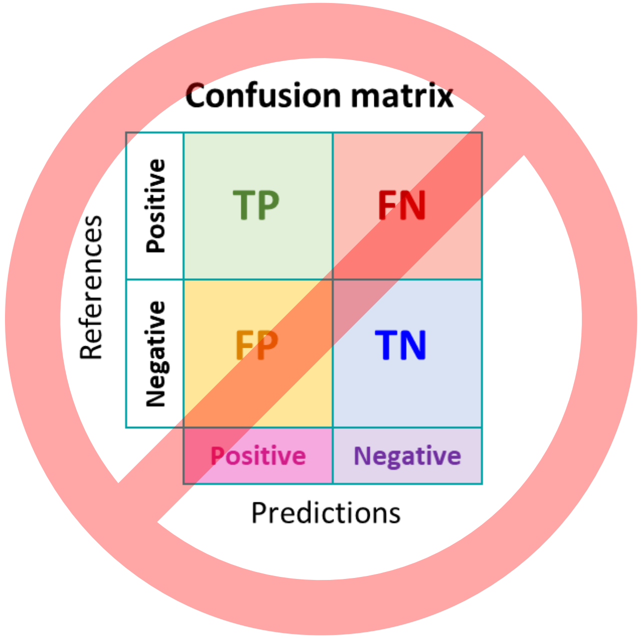
Create pseudo-absences from a set of species that may have the same sampling bias => the target group

For the case of *Popillia japonica*, we used the broader order of *Coleoptera*

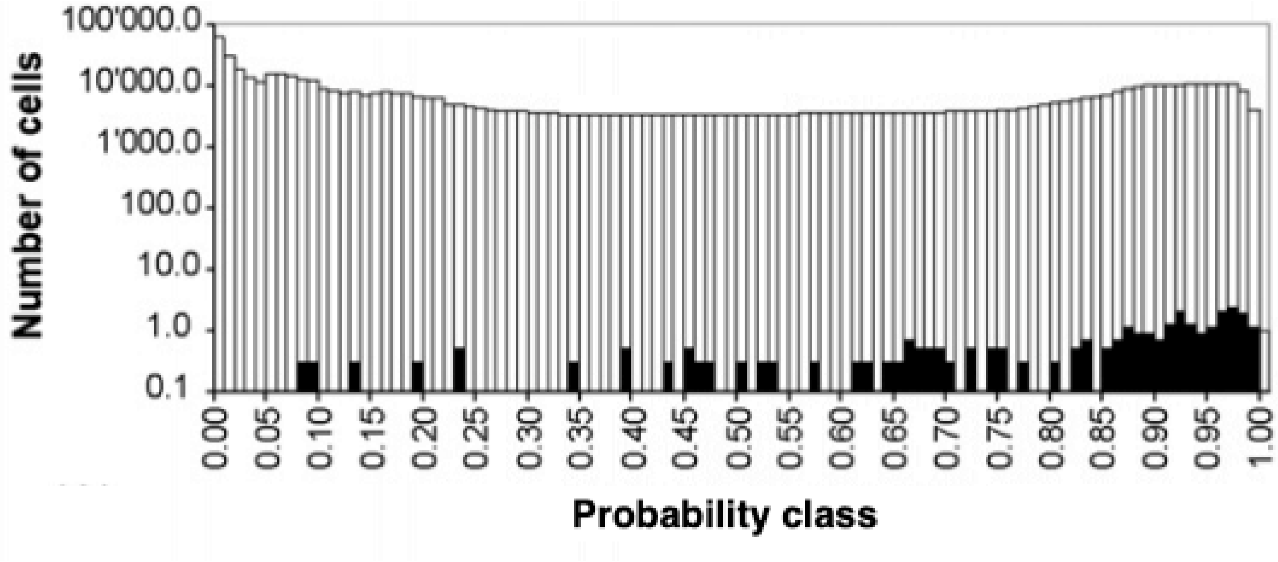
Type of data	Count
Popillia japonica	4,206
Coleoptera	49,000
No observation	9,126,667

Validation

No validation measures based on **confusion matrix:**
problems with true negative and false positive



Predicted/expected ratio



Boyce *et al.* 2002, Hirzel *et al.* 2006