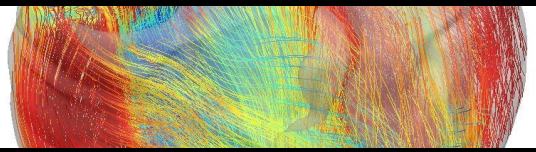
Modeling species spatial distributions in ecology

André Luís Luza (andre.luis-luza@u-bordeaux.fr)

23/01/2025

Numerics – 2024/2025 – Semaine « Numérique et transitions »



Topics

- Introduction
- Sources of Distribution Information
- Models and Data
- Site Occupancy Models and Data
- Interactive Activity
- Wrap-Up and Q&A
- Concluding remarks

Introduction

Fundamental quantities/state variables:

- Species occurrence (z, z_i)
- Species abundance (N, N_i)

Involve field observations and some sort of spatial model (SDMs)

Introduction

Distribution: manner in which a biological taxon is spatially arranged

- individual property

The way we perceive distribution can change in function of taxonomic, temporal, and spatial aspects.

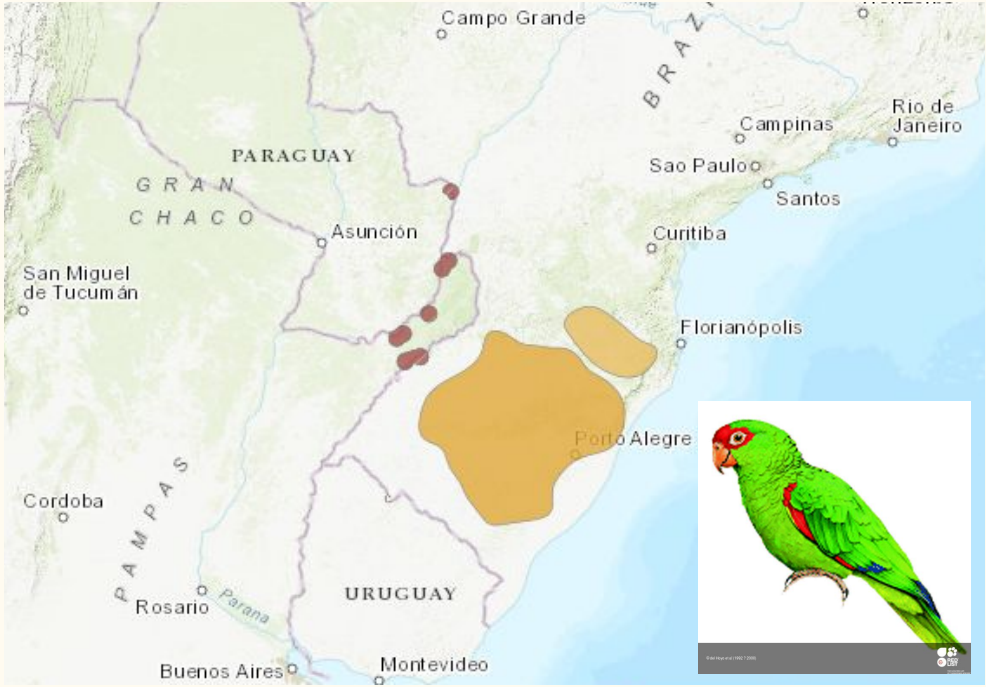
The Greater rhea (*Rhea americana*) range map (Source: IUCN, International Union for Conservation of Nature)

Introduction

Distribution: manner in which a biological taxon is spatially arranged

The way we perceive distribution can change in function of taxonomic, temporal, and spatial aspects.

Distribution can be large or small (endemic species)



The Red-spectacled Amazon (*Amazona pretrei*) range map (Source: IUCN)

Introduction

Distribution: manner in which a biological taxon is spatially arranged

The way we perceive distribution can change in function of taxonomic, temporal, and spatial aspects.

Distribution can be large or small (micro endemic species)

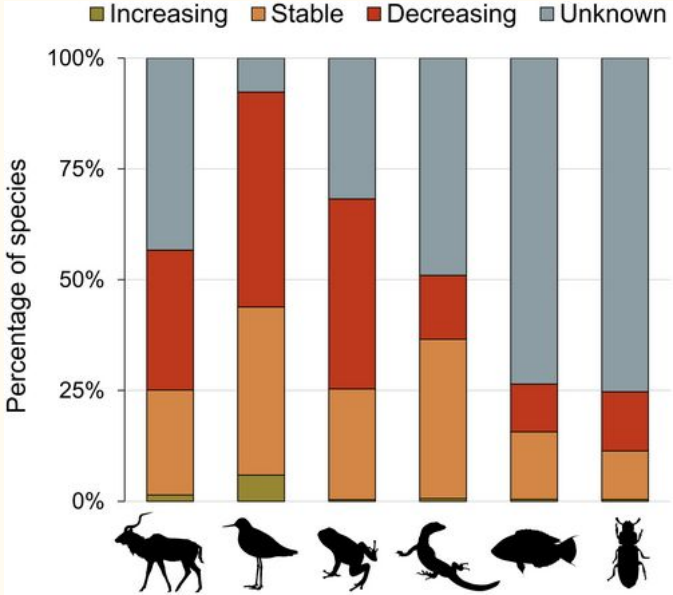
The Admirable-Redbelly-toad (*Melanophryniscus admirabilis*) range map (Source: IUCN)

Introduction

We're in the midst of a biodiversity crisis

- Fast extinction rates
- Widespread and rapid land use change: *suitable habitats are becoming rarer, smaller and isolated*
- Disease outbreaks and spread of invasive species
- Lost of unknown populations

Knowledge about species distribution can help to provide solutions
Inform conservation (IUCN)



Source: Finn et al. 2023. 10.1111/brv.12974

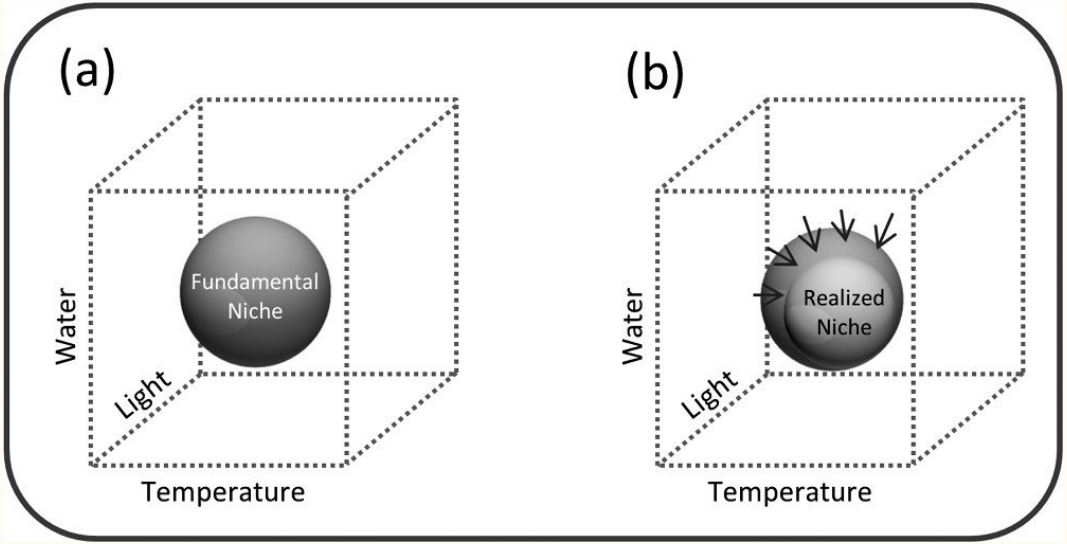
Sources of Distribution Information

Early assessments of species distribution:

- Non-spatial **niche** projections based on empirical (presence) data

Ecological niche: species' responses to a suite of environmental conditions

Non-spatial, non-statistical



(Source: Guisan et al. 2017)

Sources of Distribution Information

Early assessments of species distribution:

- Expert knowledge
  - Hand made range/distribution maps
  - What's inside? And beyond?
  - Do the edges really exist?

Spatial, non-statistical

The Greater rhea (*Rhea americana*) range map (Sourc
IUCN)

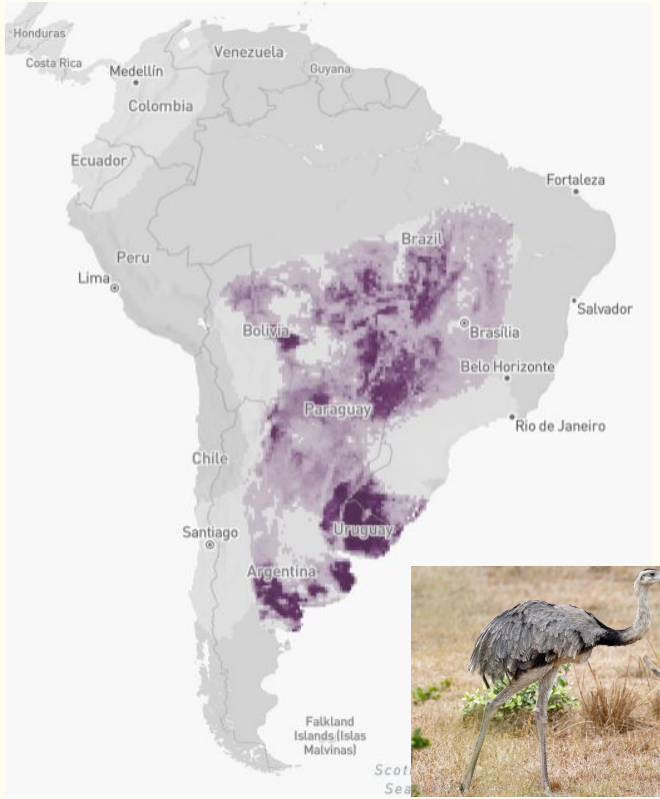
Sources of Distribution Information

Modern assessments (80's) of species distribution:

- Statistical models used to estimate ψ_i based on the relationships between environmental values and species presence and absence (in some cases).
    - *Species distribution models, bioclimatic models, climate envelopes, ecological niche models, habitat suitability models*

- Abundance maps, occurrence probability maps

Spatial, statistical



Greater rhea abundance map (Source: eBird)

Principles of Species Distribution Models

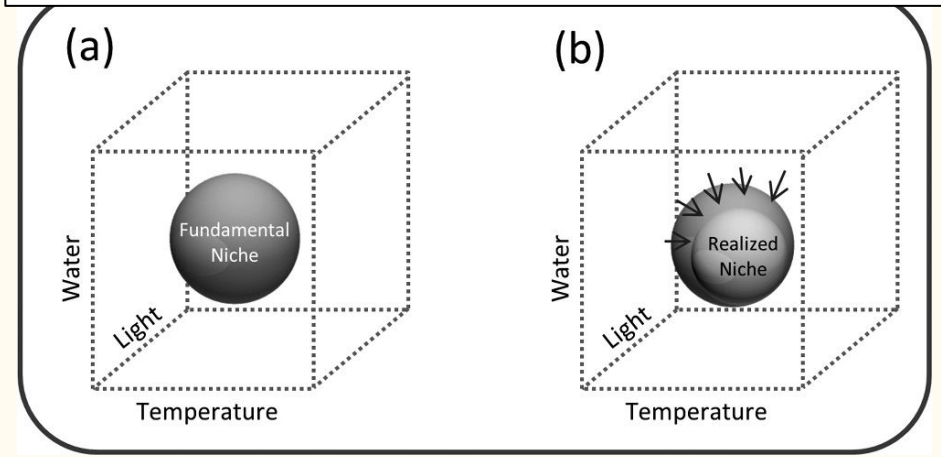
Let $z_i \in \{0,1\}$ be the (latent) realized/true presence of one focal species in $i=1, \ldots, i=M$ sites.

To $z_i = 1$, three general conditions need to be met

- Dispersal
- **Habitat suitability (ψ)**
- Biotic environment

$z_i \sim \text{Bernoulli}(\psi)$

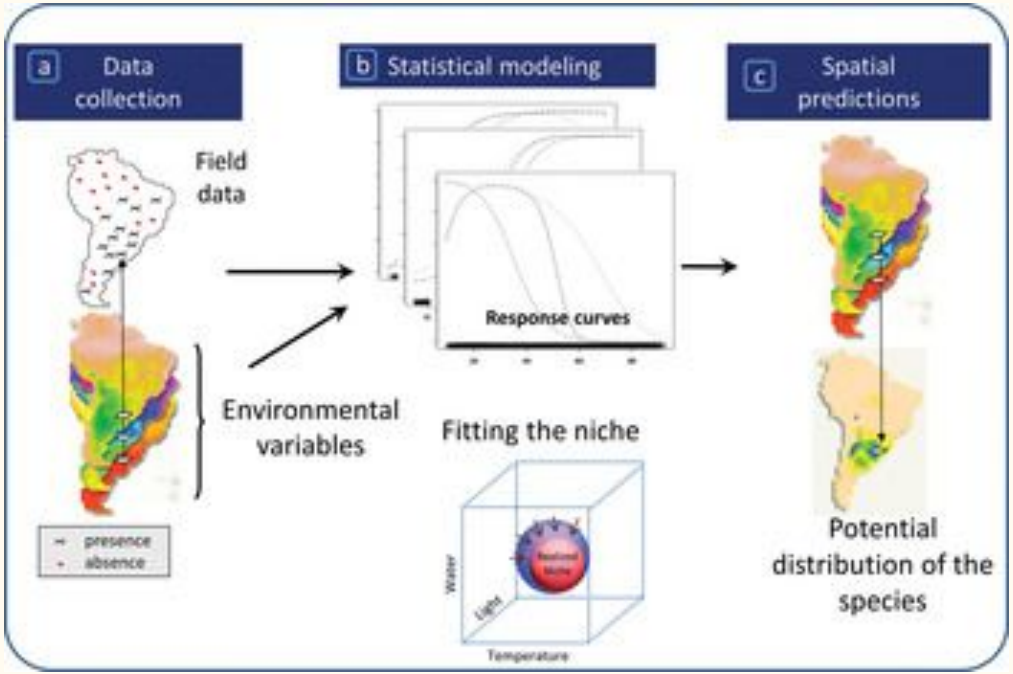
Project the niche into the geographic space



(Source: Guisan et al. 2017: 10.1017/9781139028271)

Principles of Species Distribution Models

Combine knowledge from ecology, natural history, remote sensing (GIS) and statistics



(Source: Guisan et al. 2017: 10.1017/9781139028271)

Once established, the model can be used to make:

- Inference about determinants
- Predictions (interpolation, extrapolation, hindcast, forecast)

(Hierarchical)Models and Data

Let ψ be the probability the site is occupied, p the detection probability in truly occupied sites (assumed to be $p=1$ for now), and z_i as before

$$z_i \sim \text{Bernoulli}(\psi) \text{ (ecological model)}$$

$$y_i \mid z_i \sim \text{Bernoulli}(z_i p) \text{ (observation model)}$$

Let $y_i = (y_1, y_2, y_3, \ldots, y_M)$ be the set of observations taken at the M sites, which could be

$$y_i = (0, 1, 0, 0, 1, 1, 0, 1) \quad = z_i$$

$$L(y_i) = \begin{cases} \psi, & \text{if } y_i = 1 \\ 1 - \psi, & \text{if } y_i = 0 \end{cases}$$

Pigmy rice rat (*Oligoryzomys*)

Kéry & Schaub 2012, 10.1016/C2010-0-68368-4

(Hierarchical)Models and Data

Let ψ be the probability the site is occupied, p the detection probability in truly occupied sites (assumed to be $p=1$ for now), and z_i as before

$$z_i \sim \text{Bernoulli}(\psi) \text{ (ecological model)}$$

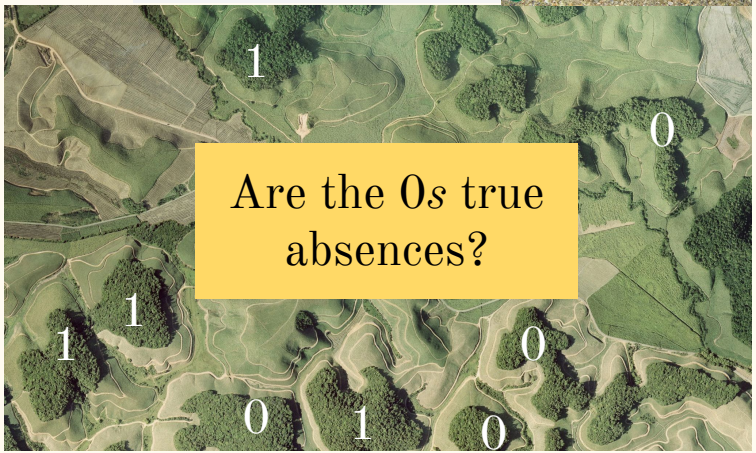
$$y_i \mid z_i \sim \text{Bernoulli}(z_i p) \text{ (observation model)}$$

Let $y_i = (y_1, y_2, y_3, \ldots, y_M)$ be the set of observations taken at the M sites, which could be

$$y_i = (0, 1, 0, 0, 1, 1, 0, 1) = z_i$$

$$L(y_i) = \begin{cases} \psi, & \text{if } y_i = 1 \\ 1 - \psi, & \text{if } y_i = 0 \end{cases}$$

Pigmy rice rat (*Oligoryzomys*)



Kery & Schaub 2012, 10.1016/C2010-0-68368-4

(Hierarchical)Models and Data

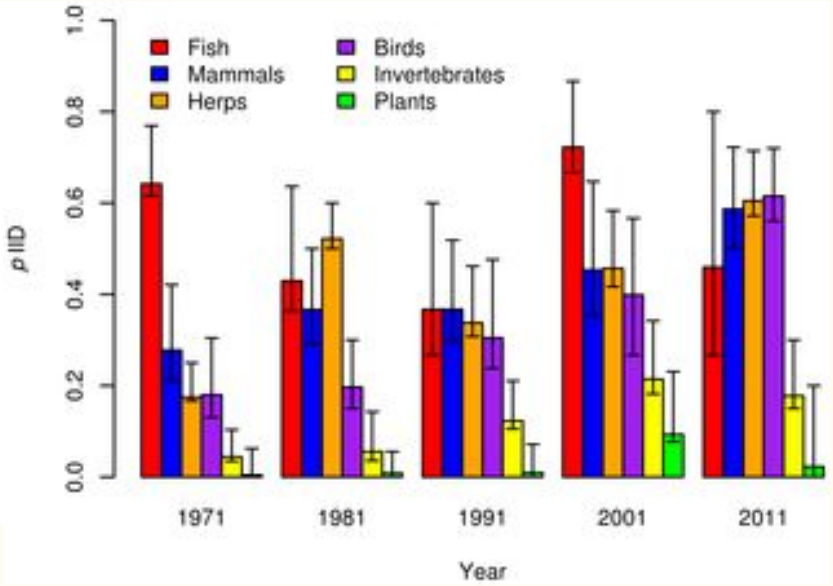
The predominant literature on SDMs (GLM, Maxent, RF, GAM) assumes **perfect** and **constant detection:**

Kellner & Swihart (2014): Review about the consideration of imperfect detection in ecology:

- 23%±1.8 (123/537) of articles addressed imperfect detection

- Range of single-survey detectability were 0.29 and 0.71

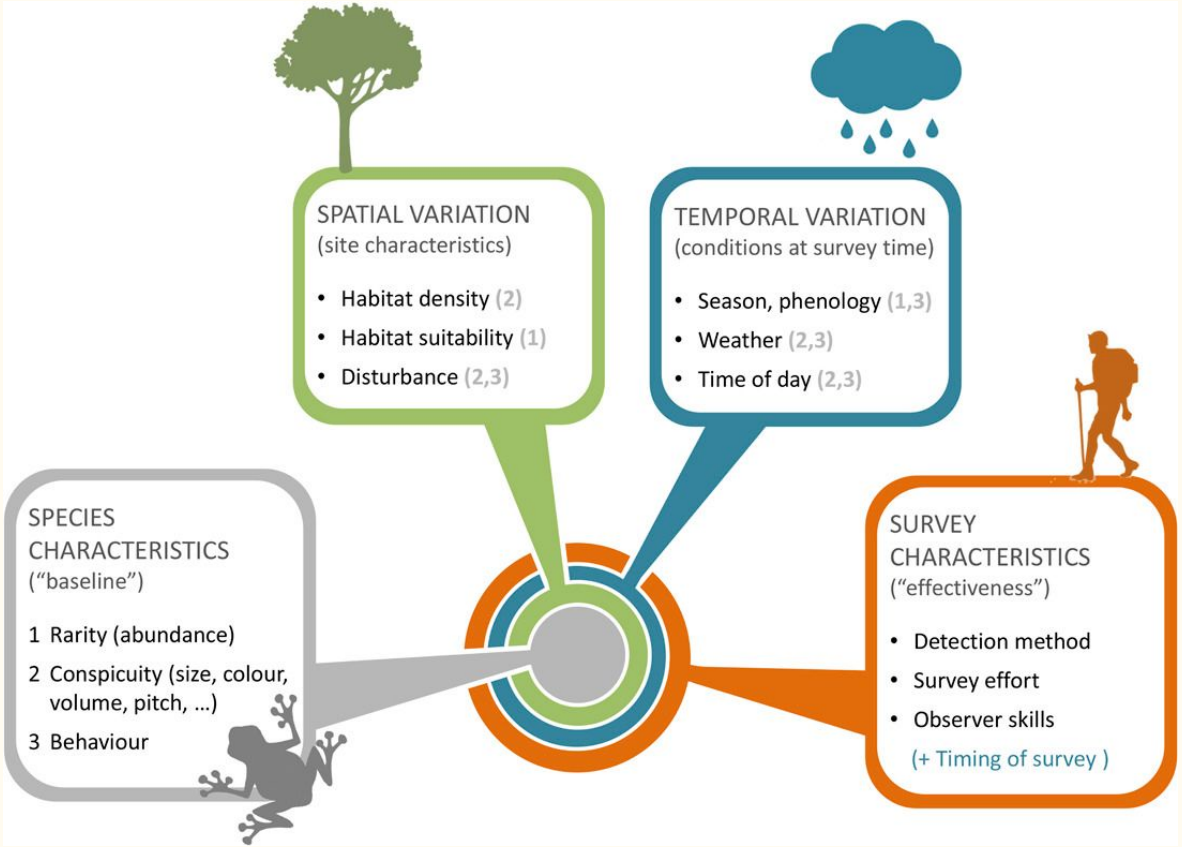
- 70% of the studies had at least one estimated detection probability per survey < 0.5



Kellner & Swihart 2014. 10.1371/journal.pone.0111436

(Hierarchical)Models and Data

Which factors affect species detection in truly occupied sites?



Guillera-Arroita 2016 (Ecography, DOI: 10.1111/ecog.02445)

(Hierarchical)Models and Data

Let ψ, the probability the site is occupied, p the detection probability in truly occupied sites (now acknowledge that $p < 1$), and z_i as before

$$z_i \sim \text{Bernoulli}(\psi) \text{ (ecological model)}$$

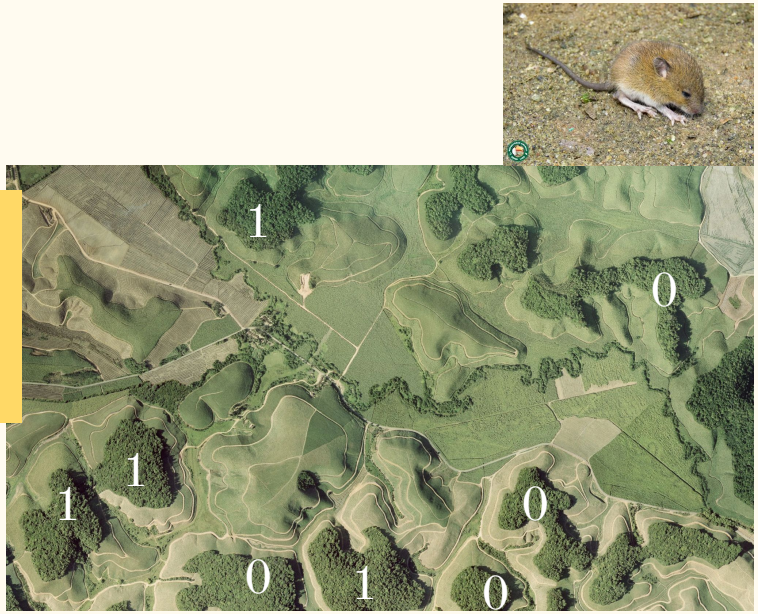
$$y_i \mid z_i \sim \text{Bernoulli}(z_i p) \text{ (observation model)}$$

Let $y_i = (y_1, y_2, y_3, \ldots, y_M)$ be the set of observations taken at the M sites, which could be

$$y_i = (0, 1, 0, 0, 1, 1, 0, 1)$$

ψ and p not identifiable (only ψp)
Apparent species distribution

$$L(y_i) = \begin{cases} \psi \cdot p, & \text{if } y_i = 1 \\ \psi(1 - p) + (1 - \psi), & \text{if } y_i = 0 \end{cases}$$



Kery & Schaub 2012

(Hierarchical)Models and Data

To make ψ and p identifiable we must conduct $j=1, \ldots, j=J$ surveys to all / subset of the sites. Consider now a **detection/non-detection occupancy data**, with $J=4$ replicated surveys (days)

$$z_i \sim \text{Bernoulli}(\psi) \text{ (ecological model)}$$

$$y_{ij} | z_i \sim \text{Bernoulli}(z_i p) \text{ (observation model)}$$

Let $y_{ij} = (y_{1,1}, y_{1,2}, y_{1,2}, \ldots, y_{M,J})$ be the set of observations taken at the M sites at J surveys. For a single site it could be

$y_{i=1,J} = (1, 0, 0, 1)$,

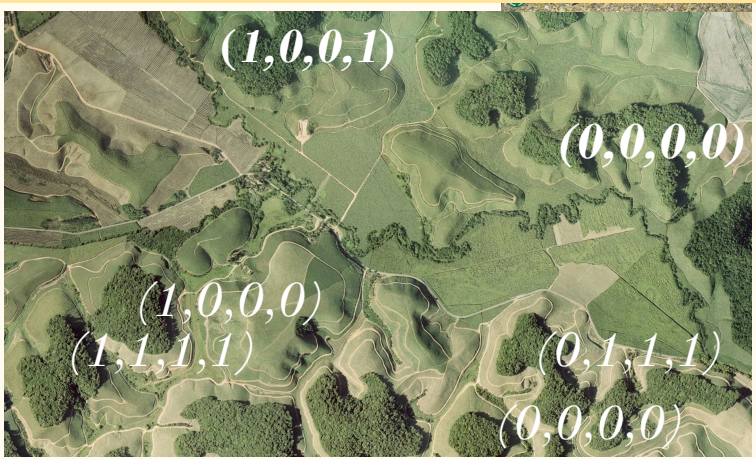
$$L(y_{i=1,J} | z_{i=1} = 1) = \psi \cdot p^J \cdot (1-p)^J$$

$y_{i=2,J} = (0, 0, 0, 0)$,

$$L(y_{i=2,J} | z_{i=2} = 1) = \psi \cdot (1-p)^J$$

$$L(y_{i=2,J} | z_{i=2} = 0) = 1-\psi$$

For any given survey budget, there is a trade-off between sampling more sites, and applying more effort per site.



(Hierarchical)Models and Data

Heterogeneous occupancy data

Let $y_{ij} = (y_{1,1}, y_{1,2}, y_{1,2}, \ldots, y_{M,J})$ be the set of observations taken at the M sites at J surveys. For a single site it could be

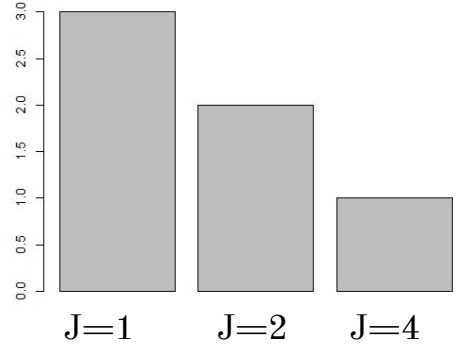
$y_{i=1,J} = (1, 0, -, -)$,

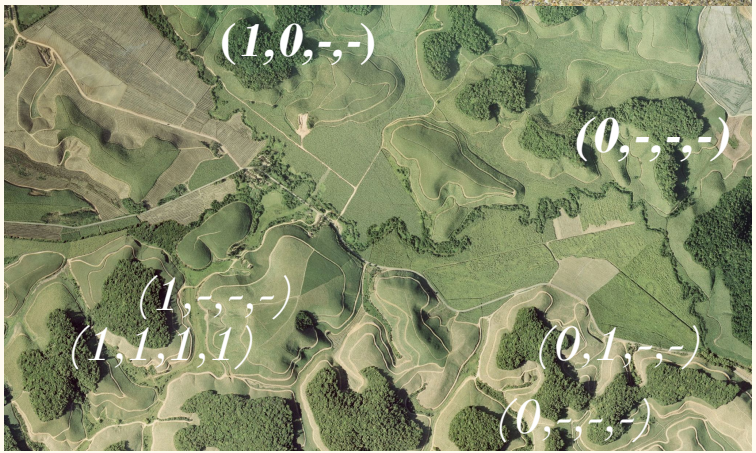
$$L(y_{i=1,J}|z_{i=1}=1) = \psi \cdot p \cdot (1-p)$$

$y_{i=2,J} = (0, -, -, -)$,

$$L(y_{i=2,J}|z_{i=2}=1) = \psi \cdot (1-p)$$

$$L(y_{i=2,J}|z_{i=2}=0) = 1-\psi$$



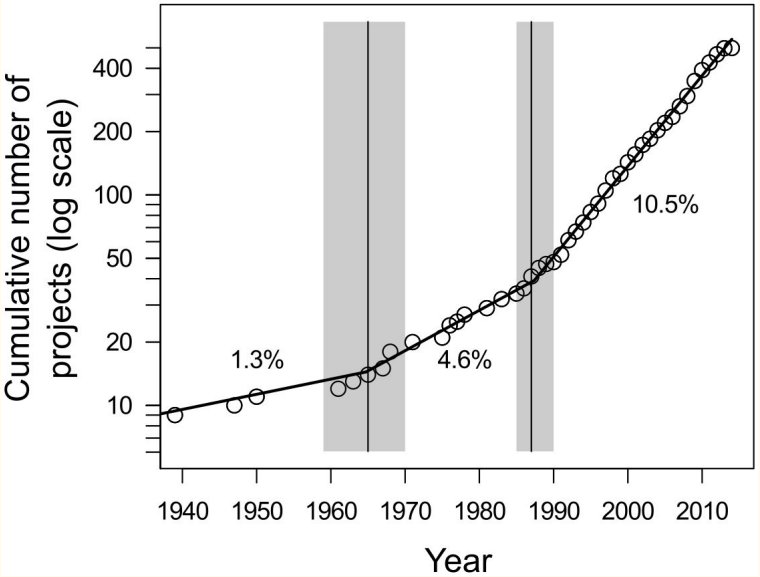


(Hierarchical)Models and Data

To model spatial distribution we need more data, right?

Pooling opportunistic data: Planned data + Museum data + Citizen-science projects (broad niche coverage)

- GBIF, iNaturalist, eBird
- Prevalence of non-sampled sites y_i=NA
- Presence-only data (lacking y_i=0)
- When y_i=1:
  - Limited temporal replication (J visits)
  - Biased (non-random) environmental sampling
  - Limited data about the observation process (covariates, V_{ij})
  - Big and heterogeneous data (varied detection histories)



Cumulative number of ecological and environmental citizen science projects as revealed by a systematic global search (Pocock et al. 2017, 10.1371/journal.pone.0172579)

(Hierarchical)Models and Data

There are three conditions for $y=1$ (ie. record coordinates in a database)

- The species must occur ($z=1$)
- **The site must be sampled/visited**
- **The species must be detected**

TABLE 1. List of challenges for the use of citizen science data for monitoring biodiversity. There are four broad categories and 10 individual challenges

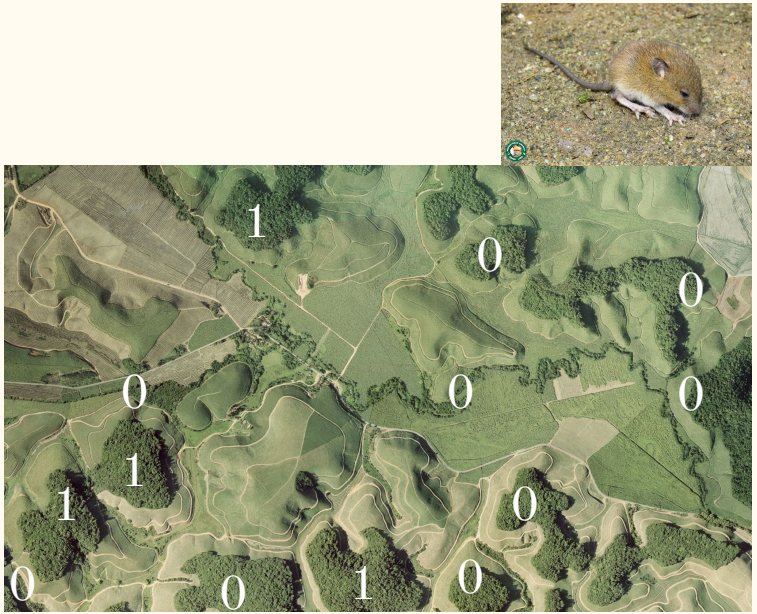
Category	Challenge
Observer behaviour	Spatial bias
	Observer differences
	Reporting preferences
	False positive errors

Johnson et al. 2023, 10.1111/2041-210X.13834

(Hierarchical)Models and Data

How to obtain absences from presence-only data?

- Background absences (proportional)
- Absences in the neighborhood



Johnson et al. 2023, 10.1111/2041-210X.13834

(Hierarchical)Models and Data

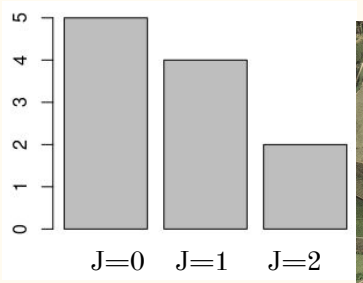
How to obtain absences from presence-only data?

- Background absences
- Absences in the neighborhood
- Community data (target-group)

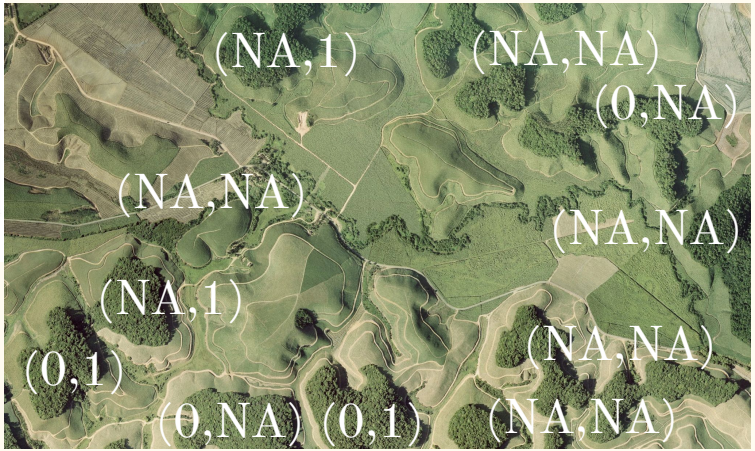
Month1-ObsB

Month1-ObsA

Johnson et al. 2023, 10.1111/2041-210X.13834



Focal species



Month1-ObsB

Month1-ObsA

van Strien et al. 2013. 10.1111/1365-2664.12158

Wrap-up of data

Available data sets can be:

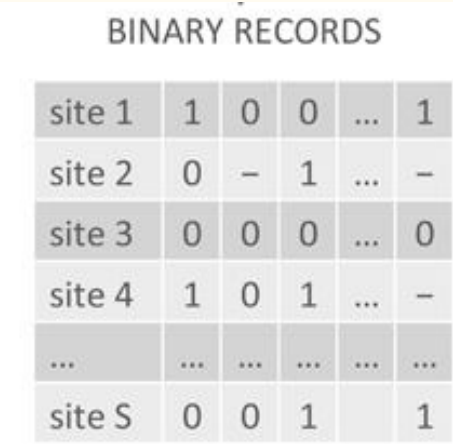
Spatial coverage
Availability
Bias

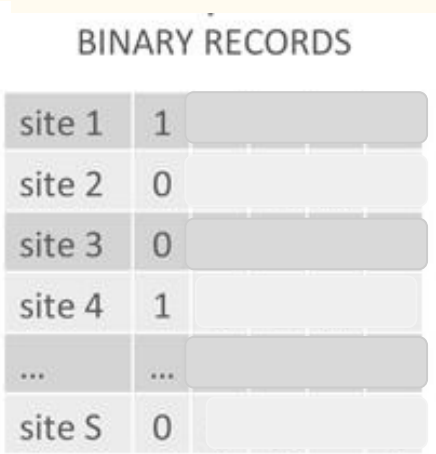
(Gold) Standard (planned repeats)

Heterogeneous (replicates in some sites)

Single-visit (no repeat)

Presence-only





Johnson et al. 2023, 10.1111/2041-210X.13834

(Hierarchical)Models and Data

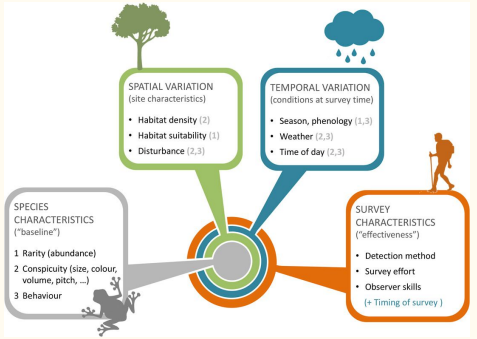
Instead of just estimating ψ and p, we often want to estimate the effect of site/visit covariates on these parameters.

The logistic regression is a natural choice for variables whose values are [0,1],

$$\log(\psi_i / (1 - \psi_i)) = X_i'\boldsymbol{\beta}$$
$$\log(p_{ij} / (1 - p_{ij})) = V_{ij}'\boldsymbol{\alpha}$$

$\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$: vectors of coefficients

X_i and V_{ij}: site-level and site-/survey-level covariates



Latimer et al. 2006. doi.org/10.1890/04-0609

(Hierarchical)Models and Data

Instead of just estimating ψ and p, we often want to estimate the effect of covariates on occupancy and detection per site and/or visit.

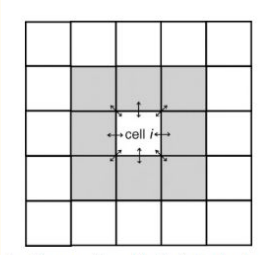
$$z_{it} \sim \text{Bernoulli}(\psi_{it}) \text{ (ecological model)}$$

$$y_{itj} \mid z_{it} \sim \text{Bernoulli}(z_{it} p_{itj}) \text{ (observation model)}$$

$$\log(\psi_{it} / (1 - \psi_{it})) = X_{it}'\boldsymbol{\beta} + \omega_{(s)} + \eta_t$$

$$\log(p_{itj} / (1 - p_{itj})) = V_{itj}'\boldsymbol{\alpha}$$

Spatial neighbors of site i

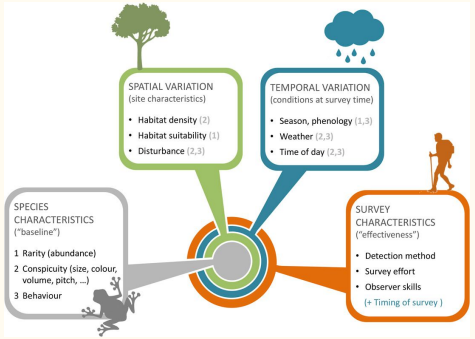


$\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$: vectors of coefficients

$\omega_{(s)}$: spatial random effect

η_t : temporal random effect

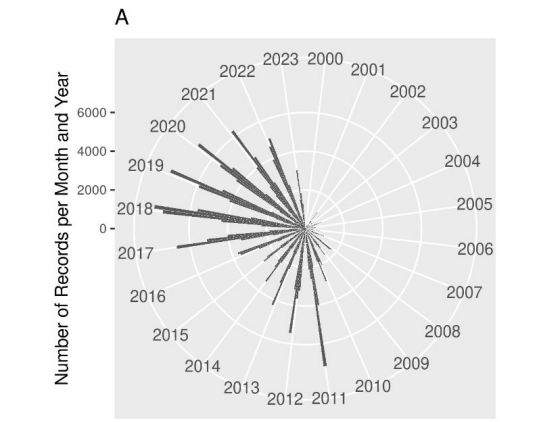
X_{it} and V_{itj}: site/year/survey-level covariates



(Hierarchical)Models and Data

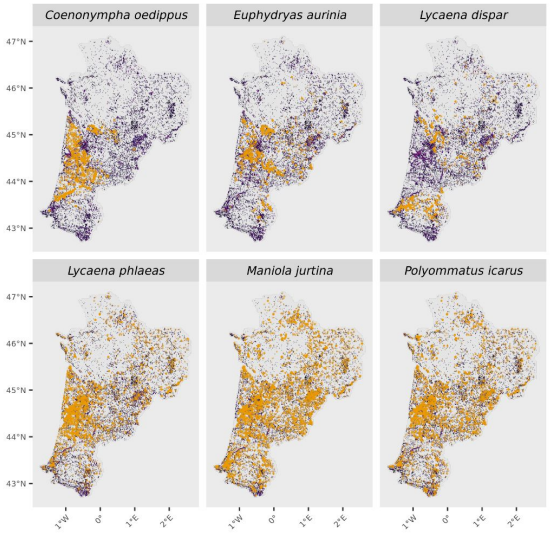
Example: Nouvelle Aquitaine Butterflies

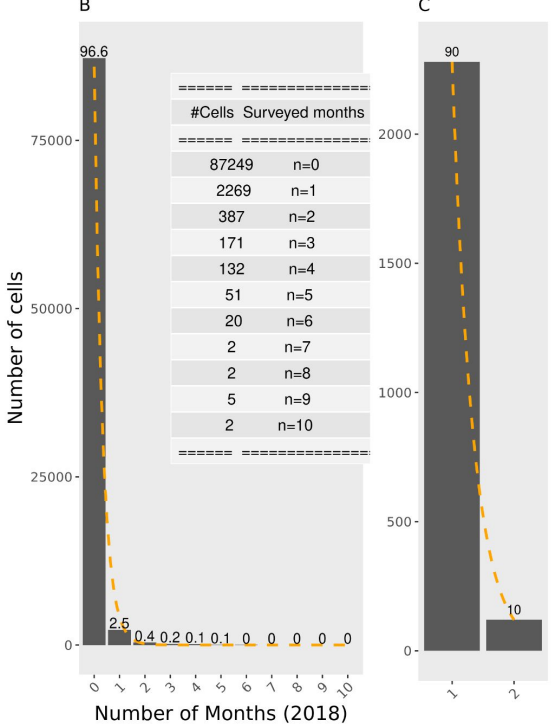
A) Distribution of records over time
B) Empirical distribution of visits to sites
C) Distribution of visits to sites often used in simulations
D) Spatial distribution of effort and records
E) Distribution of records across years and months, for six species

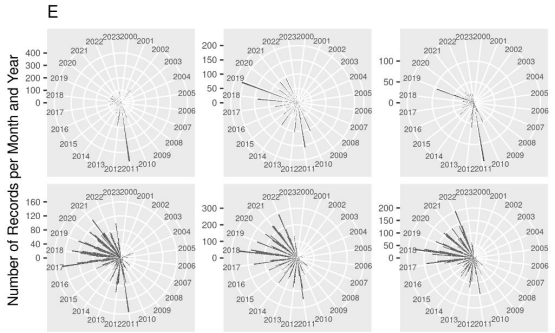


FAUNA
OBSERVATOIRE DE LA FAUNE SAUVAGE DE NOUVELLE-AQUITAINE

log(Records)
0 2 4 6 8



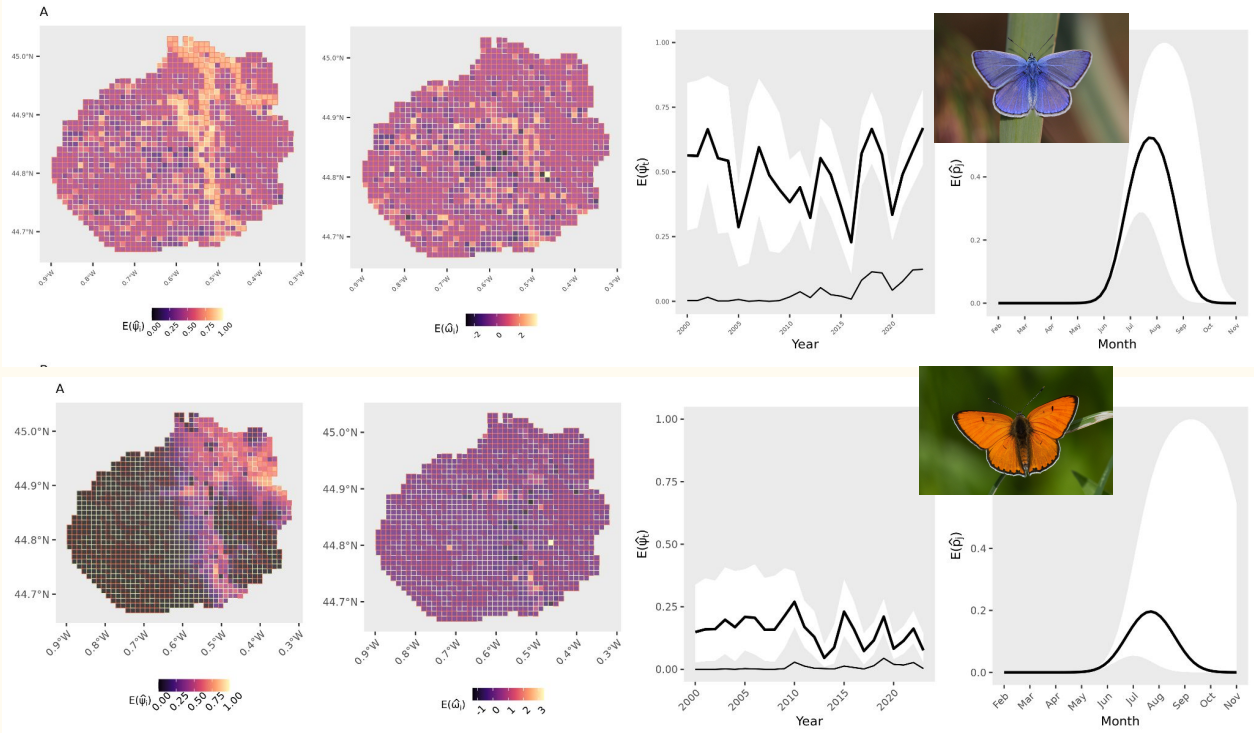




Model application (results)

$$\log(\psi_{it} / (1 - \psi_{it})) = X_{it}'\boldsymbol{\beta} + \omega_{(s)} + \eta_t$$

$$\log(p_{itj} / (1 - p_{itj})) = V_{itj}'\boldsymbol{\alpha}$$



A) Predicted occupancy probability in a buffer of 10km^2 around Bordeaux

B) Spatial random effect

C) Temporal occupancy trends (since 2000)

D) Detection probability over months

Interactive Activity

We will use R Programming Environment to:

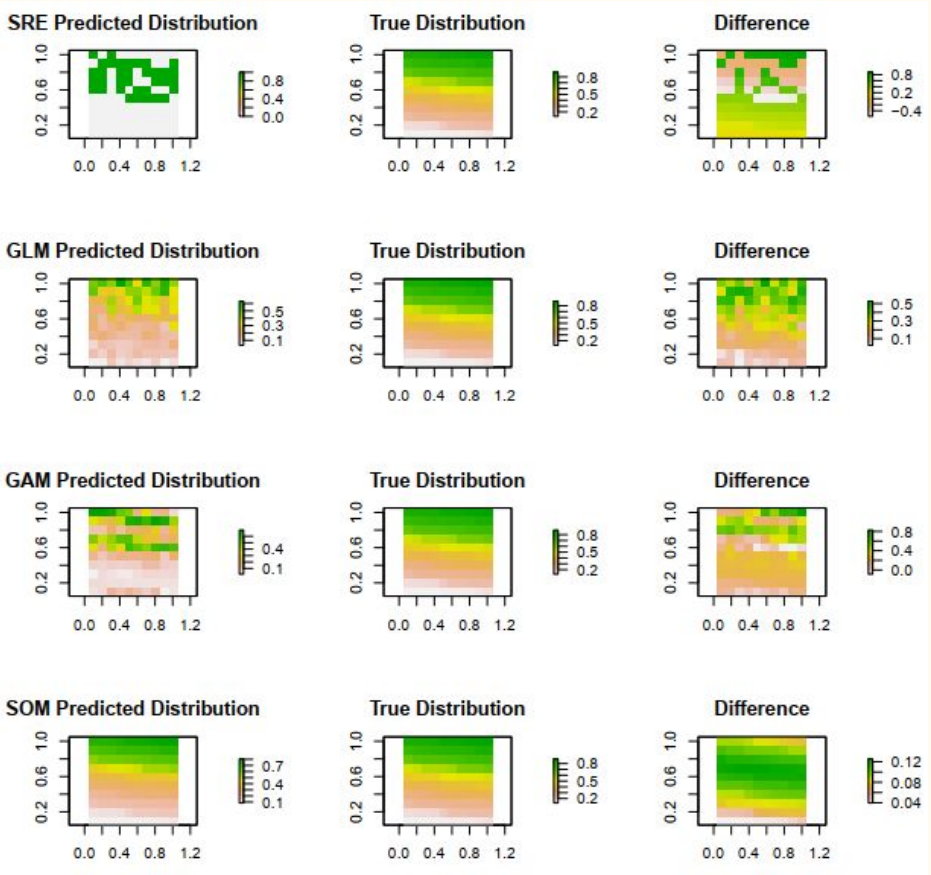
- Simulate some occupancy data (gold-standard data, single-visit/aggregated data)
- Analyze the data using:
  - BIOMOD, GLM, GAM, Hierarchical (Occupancy) Model
  - Create models without and with covariates
- Compare the performance of the different models

Wrap-Up and Q&A

Did you understand that the amount of information in data will determine the most suitable model?

Why is it important to account for imperfect detection?

- The extent of species distribution will be underestimated if $p<1$;
- Estimates of covariate relationships will be biased to zero/low when $p<1$;
- Factors that affect the observation may end up being in predictive models of occurrence or may mask the effect of factors that do affect species occurrence



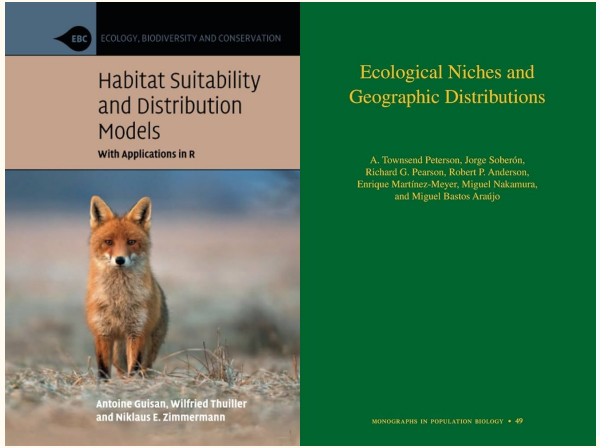
Kéry & Schaub 2012, 10.1016/C2010-0-68368-4

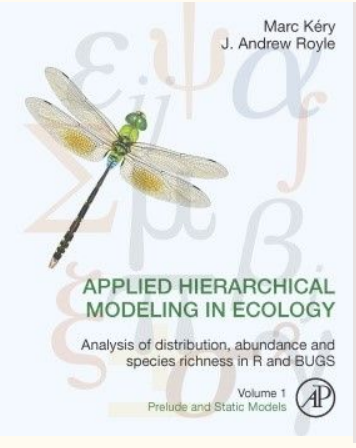
Concluding remarks

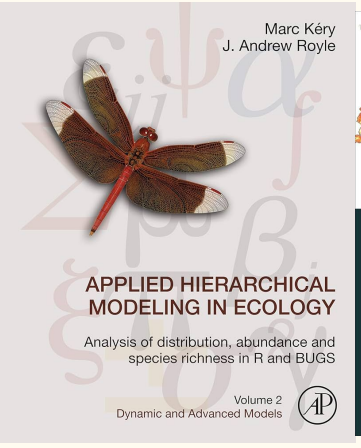
- Modern models acknowledge that the data can be generated by ecological and observation processes
    - Hierarchical formulation of SDMs enable to estimate occupancy/habitat suitability and observation probability
-
- If absences are not available, use the most suitable model and acknowledge that you're estimating observation probability/apparent distribution

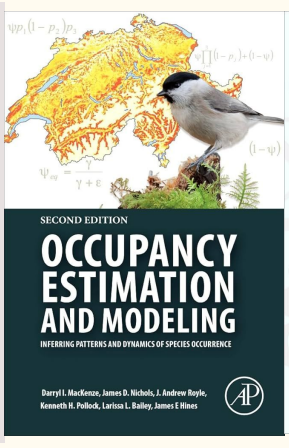
- SOM enable to deal with several sources of bias in opportunistic data sets→ output useful for conservation proposes (IUCN AOO, extent)

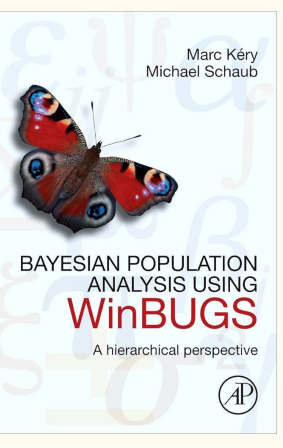
Resources for learning











Tutorials/implementations:

https://ebird.github.io/ebird-best-practices/intro.html

https://ebird.org/explore

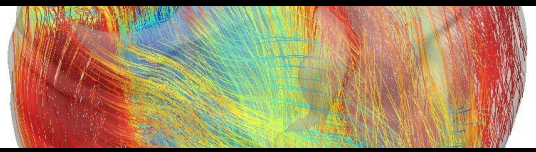
R Code: https://github.com/andreluza/SDM_talk

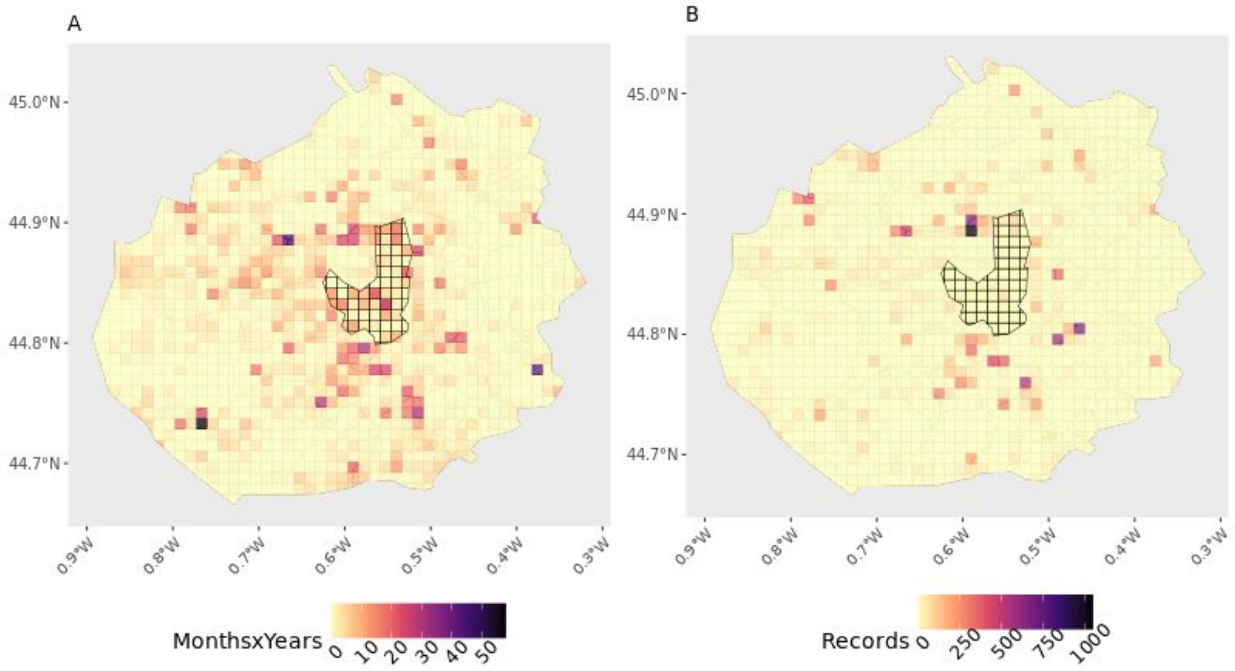
Modeling species spatial distributions in ecology

André Luís Luza (andre.luis-luza@u-bordeaux.fr)

23/01/2025

Numerics – 2024/2025 – Semaine « Numérique et transitions »




A
45.0°N
44.9°N
44.8°N
44.7°N
0.9°W
0.8°W
0.7°W
0.6°W
0.5°W
0.4°W
0.3°W
MonthsxYears
0
10
20
30
40
50
B
45.0°N
44.9°N
44.8°N
44.7°N
0.9°W
0.8°W
0.7°W
0.6°W
0.5°W
0.4°W
0.3°W
Records
0
250
500
750
1000

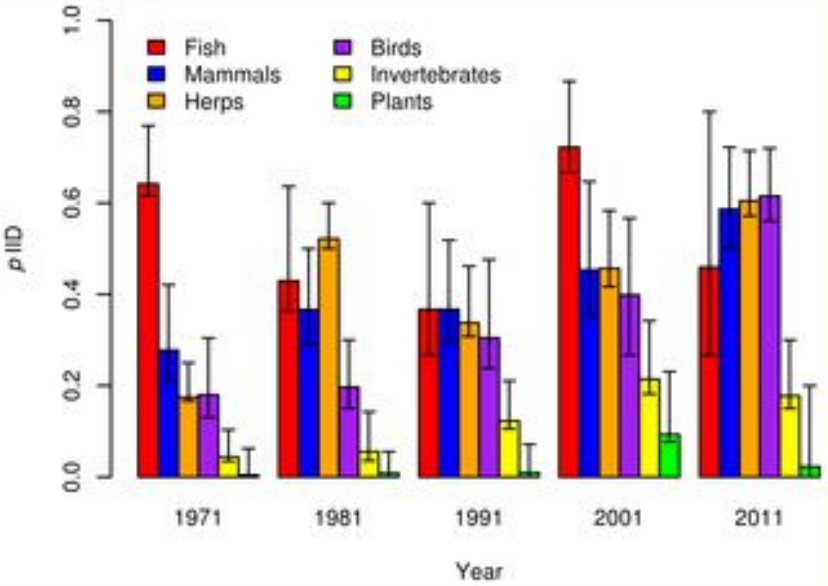
Species distribution models (SDMs)

The predominant literature on classic SDMs assumes perfect and constant detection:

$$z_i = y_i$$

Perfect detection is a very strong assumption - observations are not a perfect description of the reality

Classic SDMs model apparent distribution (product of ψ_i and p)



Kellner & Swihart 2014, Guillera-Arroita et al. 2015

Site occupancy model: An hierarchical SDM

Assumptions

- *Closure*: over the duration of the j visits, z_i must not change
  - If temporary emigration happens, ψ_i = proportion of sites used
  - If colonization and extinction happen, we can “open” the model (multi-season model with closure within seasons)
  - Some advocate the single-visit data guarantee closure (Lele et al. 2012)
- *No false positives* (z_i=1 is true)
  - 100% sure the species is there (no misidentification)
- *Homogeneous detection across sites*
  - If heterogeneous, covariates capture variation on p_{ij} (otherwise ψ_i will be underestimated)

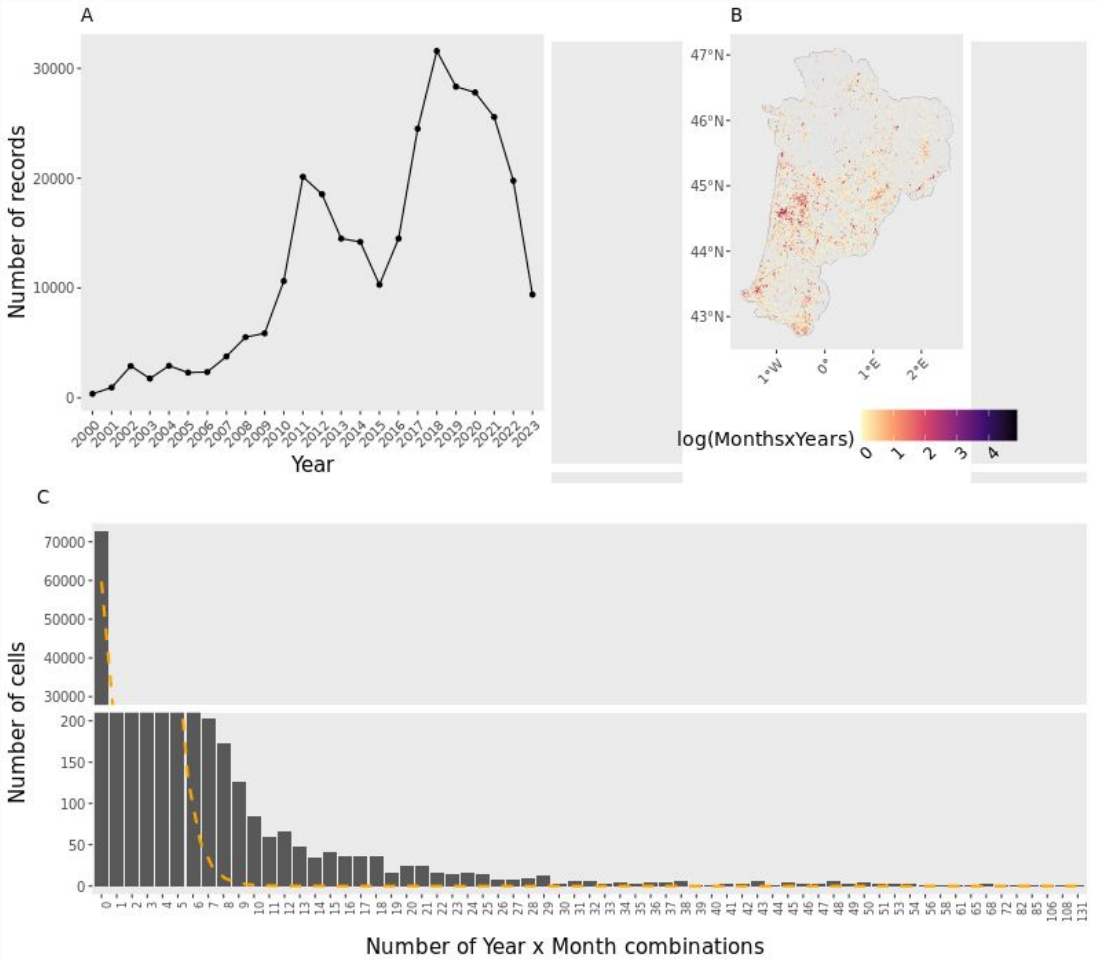
Guillera-Arroita 2016, Mackenzie et al. 2002

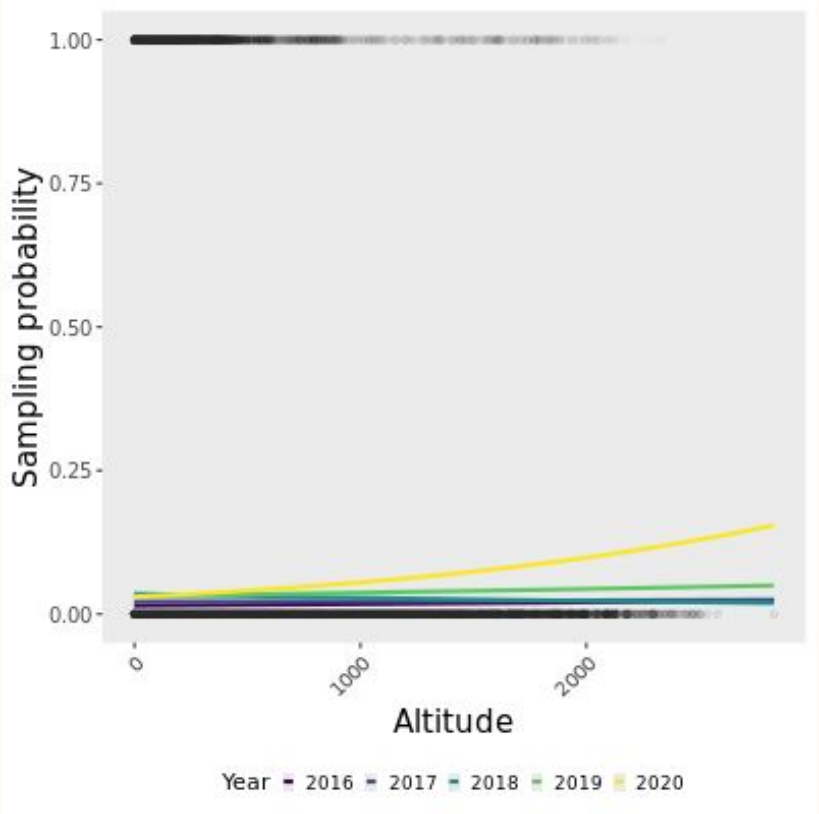
Use of site occupancy models to map geographic distributions

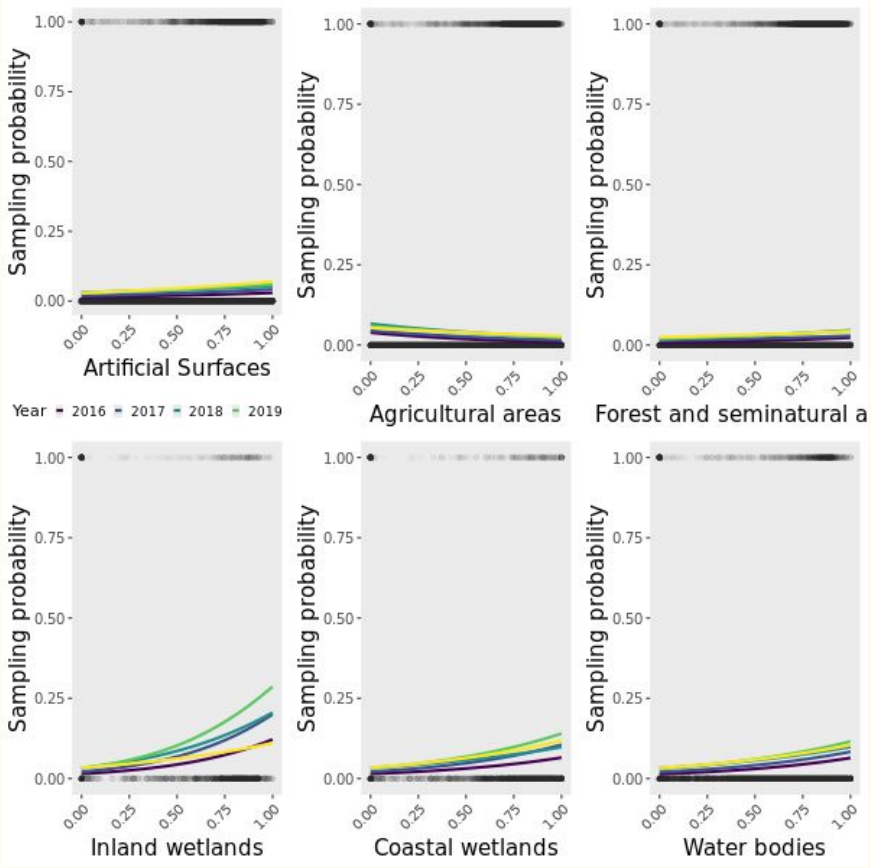
	Project elements	Unstructured	Semistructured	Structured
Why What	**Objectives** • Clear objectives • Clear planned data analysis			
How When Where	**Survey design** • Target sample size • Locations selected to fit design • Trained or expert observers			
How When Where	**Rigorous protocol** • Preselected locations • Ability to estimate detectability • Specifications for date, time of day, etc. • Specialized equipment required			
Who	**Open and flexible recruitment** • Open • Accessible • Appealing to wide range of observers			
How When Where Who	**Observation process recorded** • Time of day, date, location • Number of observers • Distance traveled • Style/type of survey			
	Variation accounted for by analytical techniques			
	Variation controlled by restricted protocol			

Note: The size of the dots relates approximately to the extent or frequency with which projects of different types meet the various criteria.

Review about the characteristics of unstructured, semistructured and structured citizen science projects (Kelling et al. 2019, BioScience)


A
Number of records
30000
20000
10000
0
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
Year
B
47°N
46°N
45°N
44°N
43°N
1°W
0°
1°E
2°E
log(MonthsxYears)
0
1
2
3
4
C
Number of cells
70000
60000
50000
40000
30000
200
150
100
50
0
Number of Year x Month combinations


Sampling probability
1.00
0.75
0.50
0.25
0.00
0
1000
2000
Altitude
Year 2016 2017 2018 2019 2020


Sampling probability
Artificial Surfaces
Agricultural areas
Forest and seminatural a
Year 2016 2017 2018 2019
Inland wetlands
Coastal wetlands
Water bodies

Why does this matter?

- Allocating efforts of conservation and population monitoring/management
- Definition of suitable habitats
- Control of invasive species, disease vectors
  - Underestimation of distribution in time (p ignored, ε is overestimated)
  - Underestimation of the uncertainty

Temporal dynamics in the occupancy of two *Aedes* species in Rio Grande do Sul

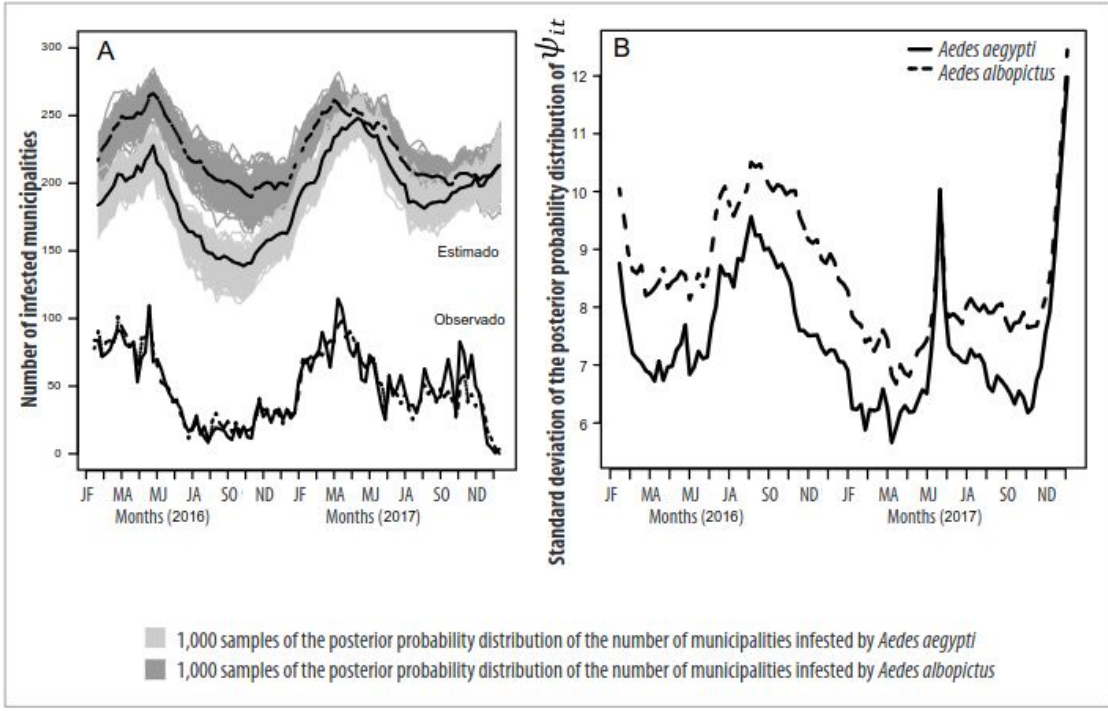


Figure 1 – Temporal variation on the number of infested municipalities (A) and on the uncertainty about infestation (B) by *Aedes aegypti* and *Aedes albopictus*, Rio Grande do Sul, 2016-2017

Why does this matter?

- Allocating efforts of conservation and population monitoring
- Definition of suitable habitats
- Control of invasive species, disease vectors
    - Underestimation of the spatial distribution

Spatiotemporal variation in distribution of two *Aedes* species in Rio Grande do Sul

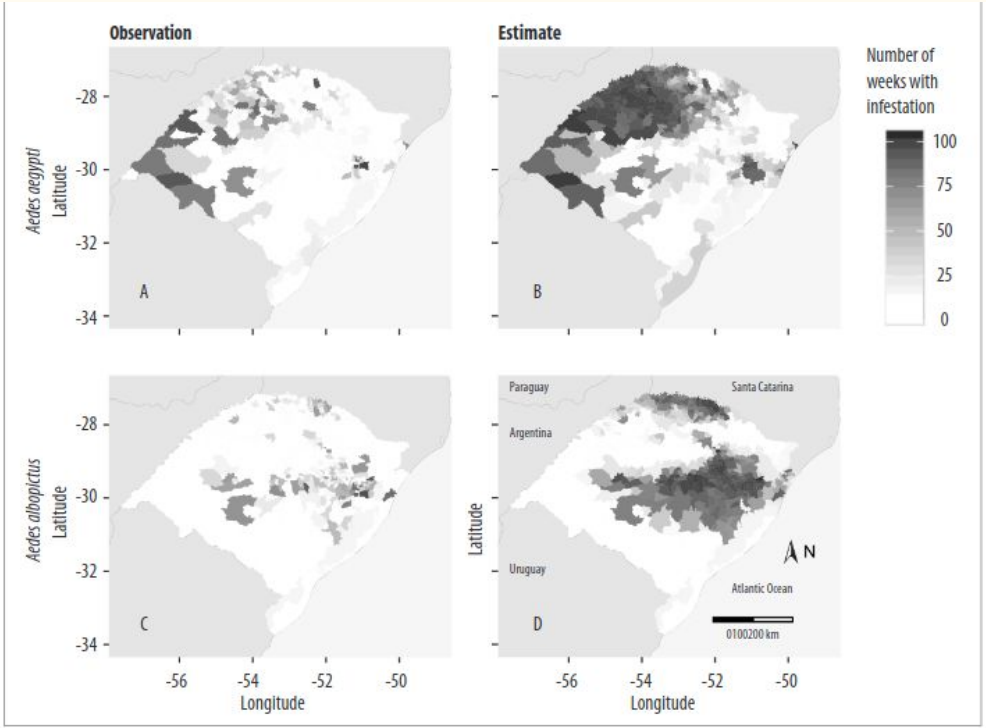
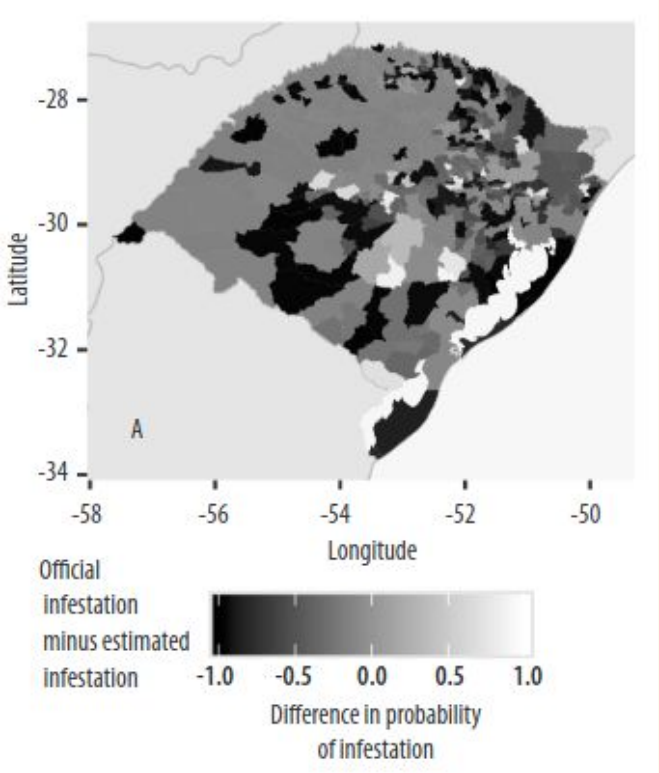


Figure 2 – Geographic distribution of *Aedes aegypti* (A and B) and *Aedes albopictus* (C and D), based on number of weeks with at least one detection of the species (A and C) and with presence of the species according to the model (B and D), Rio Grande do Sul, 2016-2017

Why does this matter?

- Allocating efforts of conservation and population monitoring
- Definition of suitable habitats
- Control of invasive species, disease vectors
  - Attribute low disease risk to municipalities with high ψ_i
  - Autochthonous dengue cases in municipalities not officially infested

Comparison between binary classification and probabilistic classification of infestation by *Aedes aegypti*

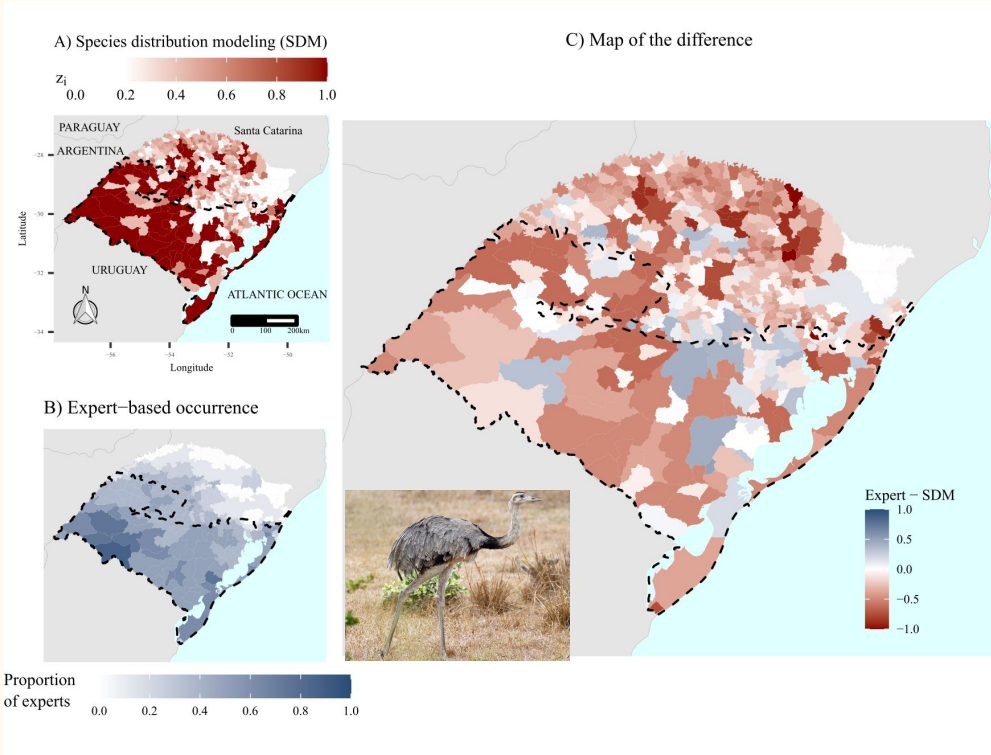


Combination of Sources of Distribution Information

Comparison of sources of distribution data can be informative:

- Prior knowledge
- Where + research is needed
- Suitable vs. unsuitable regions
- Where we need to improve the models

Example: The greater rhea distribution in Rio Grande do Sul, BR, using expert knowledge and SDMs fitted to citizen-science data (Luza et al. 2023. 10.1007/s43388-023-00143-3)

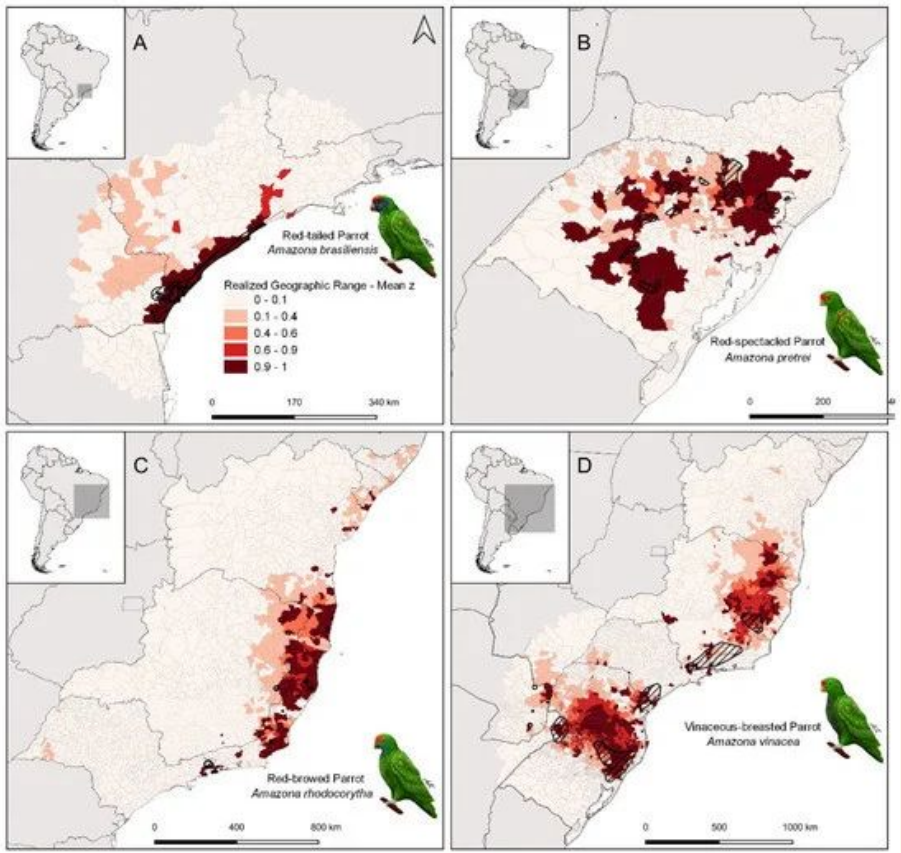


Information might not match (which is interesting)

Statistical tool (regression model) used to fill gaps in knowledge about species distribution and abundance (+demographic and derived parameters)

Often used to make predictions at broad scales (region +).

Parameter of interest for SD modelers: ψ_i or ψ_{it}, the site occupancy probability in site i and/or time t



Distribution of parrot species in Southern Brazil - estimates VS. IUCN (Zulian et al. 2021)