

# Bias correction in integrated multispecies distribution models

André Luza. UBx/BIOGECO, INRAE. [andre-luis.luz@u-bordeaux.fr](mailto:andre-luis.luz@u-bordeaux.fr)

Didier Alard, FAUNA's director, UBx/BIOGECO, INRAE

Frédéric Barraquand, CNRS/IMB, UBx

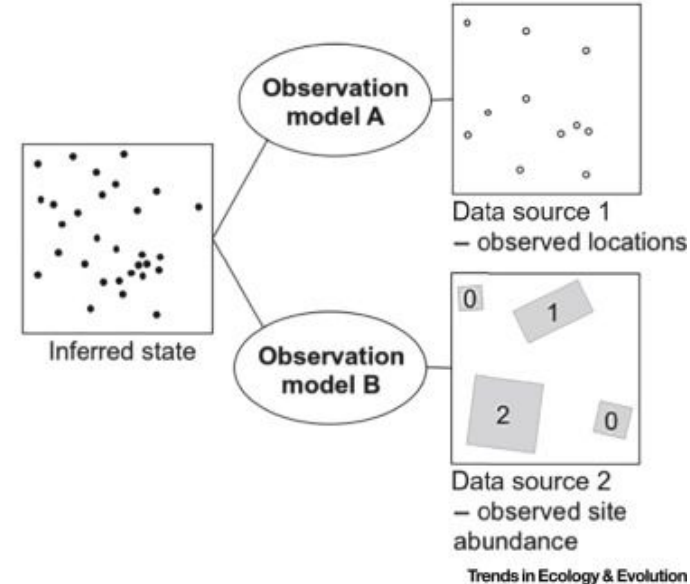
# Background

Developing integrated models that use jointly PA and PO data and overcome biases in PO data is an active area of species distribution modeling.

- Enable the correct use of the abundant and spatially extensive opportunistic data
- Improves model estimation (accuracy and precision) and predictive performance
- Inference on fine-scale intensity, at broad spatial extents

*Retain the strengths of different data types  
and correct as much as possible for their  
weaknesses*

(B) Model-based data integration



Isaac et al. 2020. Trends Ecol Evol

Trends in Ecology & Evolution

# Foundational work



Special Feature: New Opportunities at the Interface Between Ecology and Statistics

 Free Access

## **Bias correction in species distribution models: pooling survey and collection data for multiple species**

William Fithian  Jane Elith, Trevor Hastie, David A. Keith

First published: 10 October 2014 | <https://doi.org/10.1111/2041-210X.12242> | Citations: 320

One of the first Integrated SDM that jointly models PA and PO data for multiple species.

- Species PA and PO data seen as realizations of a **Poisson Point Process**: link between locations of individuals and abundance, occupancy, species richness, ...

# The model

$\mathcal{S}$  the point process that describes a random set of coordinates  $s$  representing the position/point process of individuals of one species in a two-dimensional domain  $\mathcal{D}$ , for  $k=1, \dots, \mathcal{M}$  species. The PA data model for site  $i$  is

$$\mathcal{S}_k \sim \text{IPP}(\lambda_k(s_i))$$

$$\log(\lambda_k(s_i)) = \alpha_k + \beta_k \times x_i$$

And the thinning process is modeled as a log-linear model

$$\log(\lambda_k(s) b_k(s)) = \alpha_k + \beta_k \times x(s) + \gamma_k + \delta \times z(s)$$

$$\mathcal{T}_k \sim \text{IPP}(\lambda_k(s) b_k(s))$$

Where  $\delta$  is the proportional-bias effect which is constant across the  $\mathcal{M}$  species. The joint likelihood function is

$$\mathcal{L}_k(\alpha_k, \beta_k, \gamma_k, \delta) = \mathcal{L}_k^{PA}(\alpha_k, \beta_k) + \mathcal{L}_k^{PO}(\alpha_k, \beta_k, \gamma_k, \delta)$$

Use of background (quadrature) points

One of the first Integrated SDM that jointly models PA and PO data for multiple species.

- Species PA and PO data seen as realizations of a **Poisson Point Process**: link between locations of individuals and abundance, occupancy, species richness, ...

PA data is assumed to be unbiased by design → used to estimate intensity and correct the spatial bias in PO data;

Multi-species data can provide relevant information ("borrow strength") to correct the spatial bias in PO data.

# Foundational work



Special Feature: New Opportunities at the Interface Between Ecology and Statistics

Free Access

## **Bias correction in species distribution models: pooling survey and collection data for multiple species**

William Fithian Jane Elith, Trevor Hastie, David A. Keith

First published: 10 October 2014 | <https://doi.org/10.1111/2041-210X.12242> | Citations: 320



Only 13 matches in Google Scholar using the “multispeciesPP” package (R)

Two evaluated the model through simulations:

- Fithian et al. (2014): *simulations not well documented*;
- Peel et al. (2019, MEE): *simulations initialized with their own data (mollusks, Antarctica)*.



# Objectives & Methods

## 1 - Replication of simulation-based results

UNDER REVIEW

RESCIENCE C

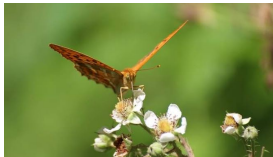
Replication / ecology

### **[~Re] Bias correction in species distribution models: pooling survey and collection data for multiple species**

Andre Luza<sup>1, ID</sup> and Frédéric Barraquand<sup>2, ID</sup>

<sup>1</sup>Biodiversité Gènes et Communautés (UMR 1202), University of Bordeaux, INRAE, Pessac, France – <sup>2</sup>Institut de Mathématiques de Bordeaux (UMR 5251), University of Bordeaux, CNRS, Bordeaux INP, Talence, France

## 2 - Application to butterfly data in the greater area of Bordeaux, SW France



# Objectives & Methods

## 1 - Replication of simulation-based results

- *Does including presence-only data improve model estimation?*
- They compared coefficients  $\beta$  of **species 1** using 5 different models (legend of the figure)
- Key result: The all species' ISDM produced accurate and more precise coefficients than other models

## “Detective work”

UNDER REVIEW

RESCIENCE C

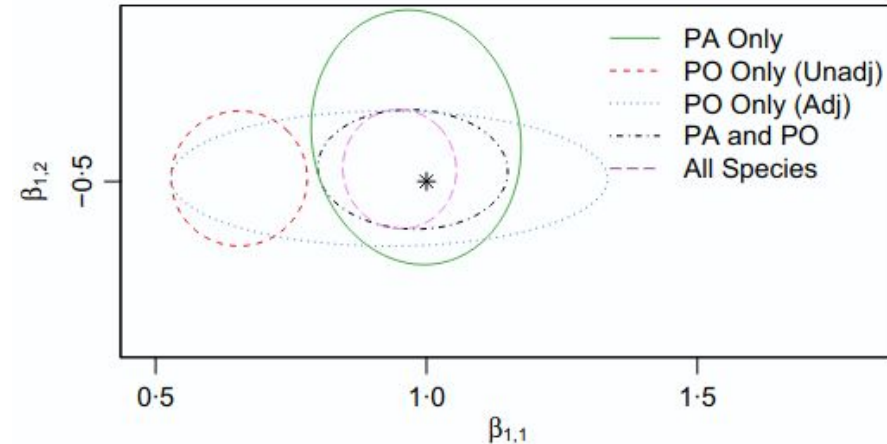
Replication / ecology

### **[Re] Bias correction in species distribution models: pooling survey and collection data for multiple species**

Andre Luza<sup>1,®</sup> and Frédéric Barraquand<sup>2,®</sup>

<sup>1</sup>Biodiversité Gènes et Communautés (UMR 1202), University of Bordeaux, INRAE, Pessac, France – <sup>2</sup>Institut de Mathématiques de Bordeaux (UMR 5251), University of Bordeaux, CNRS, Bordeaux INP, Talence, France

Simulation: Confidence Ellipses for  $\beta_1$



Ninety-five percent Wald confidence regions for  $\beta$ , the species distribution coefficients for species 1, obtained by using different methods. The plot illustrates the precision and accuracy with which the coefficients are estimated by each method. The black star denotes the true values of the parameters of interest.

# Results (simulations)

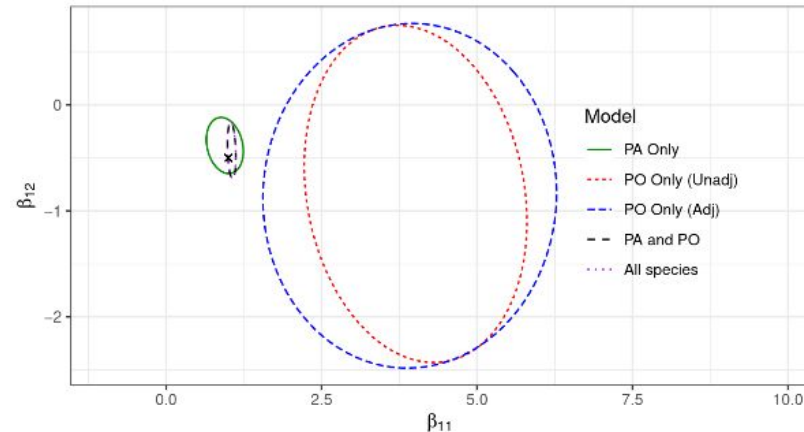
## Six scenarios tested

Table 1: Summary of the simulation settings used across the exercises. PO average and standard deviation SD are calculated across the  $M$  species.

Exercise	nPA	nPO (unique cells)	PO average $\pm$ SD [range]	$\gamma$	$\delta$	$\mathcal{D}$
I	500	81 (71)	$3.85 \pm 2.41$ [1,8]	-4	-0.3	1,000
II	500	592 (451)	$28.19 \pm 10.03$ [10,45]	-2	-0.3	1,000
III	500	1632 (850)	$77.41 \pm 26.33$ [36,121]	-1	-0.3	1,000
IV	500	15596 (8519)	$742.67 \pm 273.39$ [412,1409]	-1	-0.3	10,000
V	500	771 (733)	$36.72 \pm 17.01$ [15,74]	-4	-0.3	10,000
VI	500	9057 (8553)	$431.29 \pm 339.31$ [176,1711]	-4	-0.3	50,000

## Key results

- ISDMs performed better than other models
- Multi-species ISDM  $\approx$  single-species ISDM
- Different results than Fithian et al. (2015), and original parameter values not found



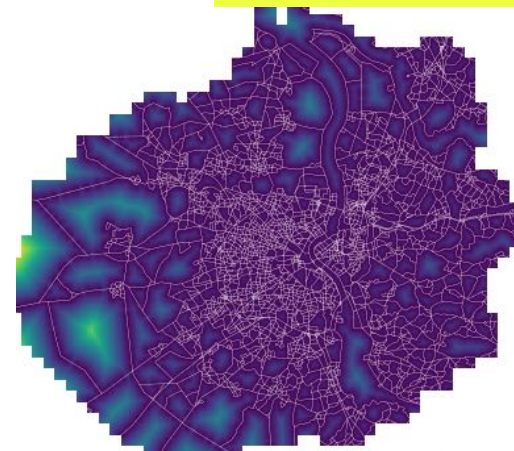
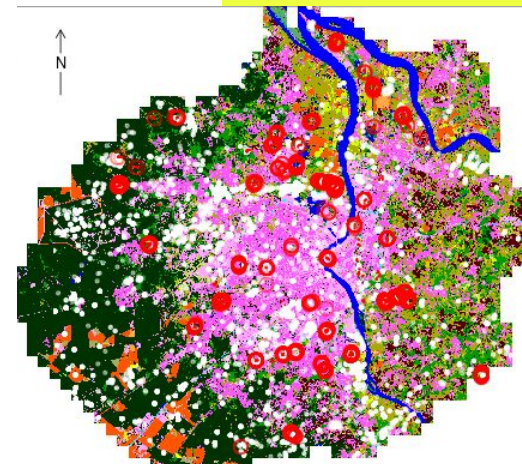
Ninety-five percent Wald confidence regions for  $\beta$ , the species distribution coefficients for species 1, obtained by using different methods. The plot illustrates the precision and accuracy with which the coefficients are estimated by each method. The black dot denotes the true values of the parameters of interest.



# Objectives & Methods

## 2 - Application of the Fithian's et al. ISDM to butterfly data in the greater Bordeaux area

- Compilation of butterfly records from 1998-2025 (n=30,647; 20,058 PO points, 10,589 records from protocolled samples) from different projects (125) and data sets (301) in Bordeaux + 10 km buffer.
- Cells of 100x100m. Environmental data from CESBIO, EUDEM, ROUTE 500®.
- Well-known area (projects), speciose butterfly community (61 spp + 30 records), ecological and bias gradients well covered.



# Objectives & Methods

## 2 - Application of the Fithian's et al. ISDM to butterfly data in the greater Bordeaux area

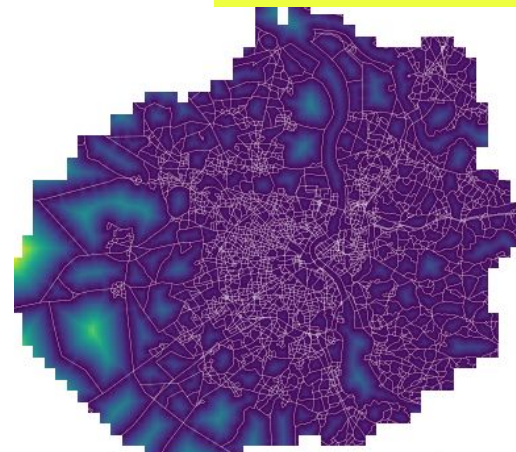
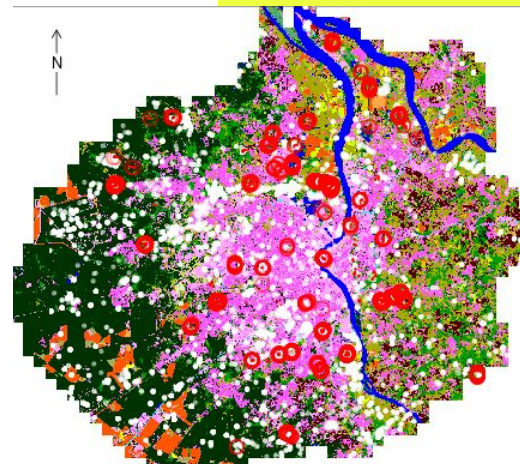
- Compilation of butterfly records from 1998-2025 (n=30,647; 20,058 PO points, 10,589 records from protocolled samples) from different projects (125) and data sets (301) in Bordeaux + 10 km buffer.
- Cells of 100x100m. Environmental data from CESBIO, EUDEM, ROUTE 500®.
- Well-known area (projects), speciose butterfly community (61 spp + 30 records), ecological and bias gradients well covered.

Is there evidence in empirical data in favor of the multispecies ISDM over other modeling options?

a) Are coefficients (grassland and urban cover effects) from the all species ISDM more precise for low-intensity species?

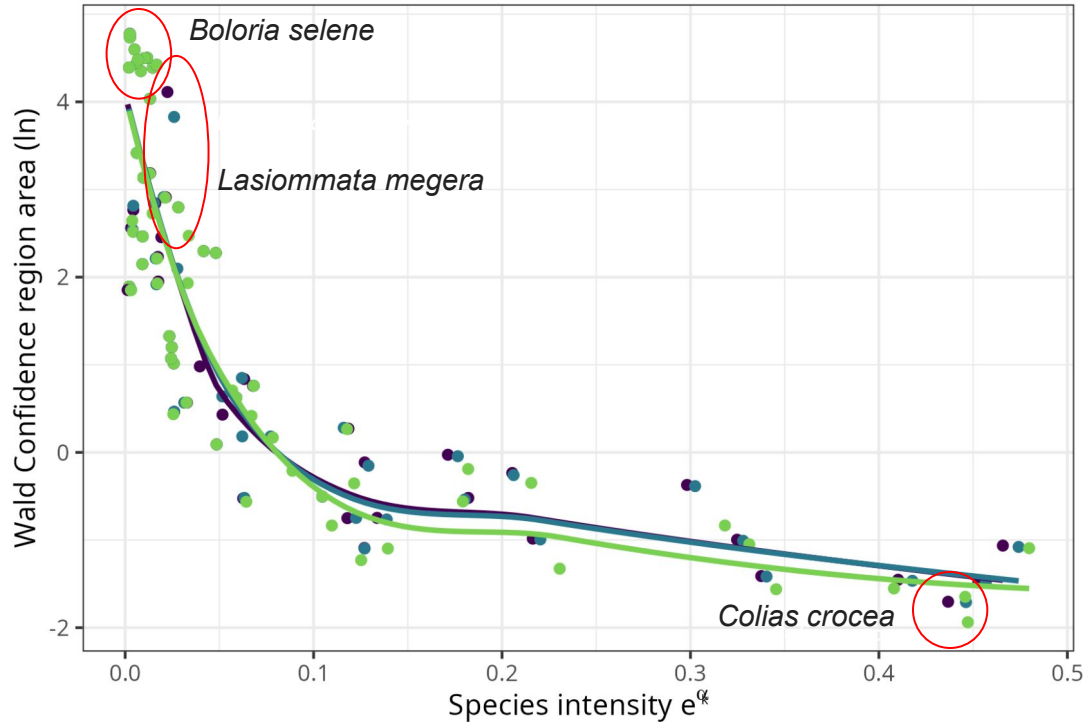
b) Are confidence regions smaller for high-intensity species ( $>e^{\alpha_k}$ )? For 'oversampled' species ( $>\gamma_k$ )? When PA and PO data overlap largely? When the n. cells of PA data > n. PO points?

c) Is predictive performance improved?



# Results (application)

a) Are coefficients from the all species ISDM more precise for low-intensity species?



For species with average intensity  $< 0.1$  (1 ind/10 ha) there was almost no difference between models

Beyond this intensity, the all species ISDM was more precise.

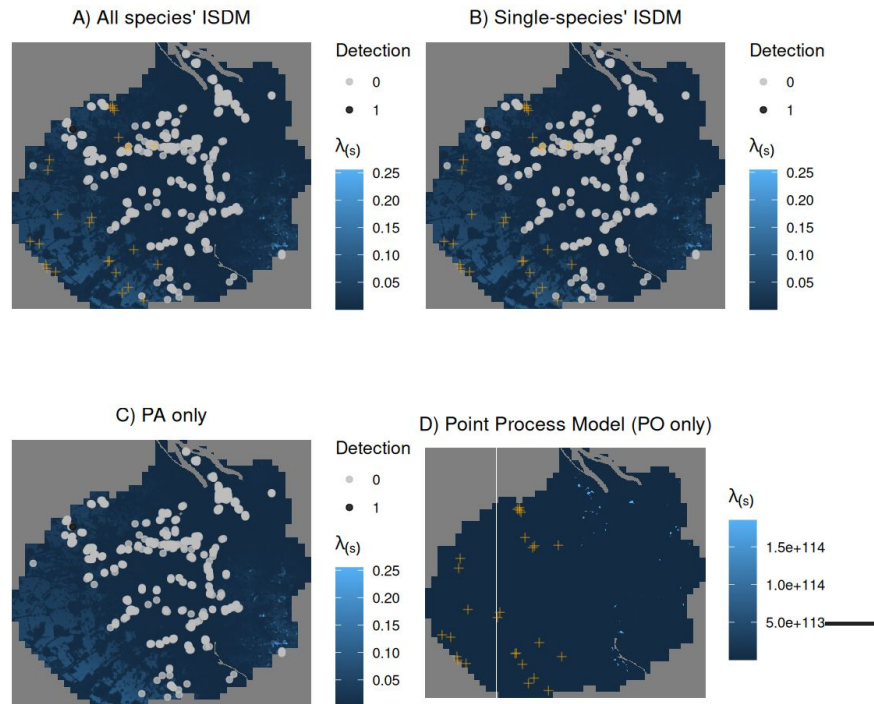
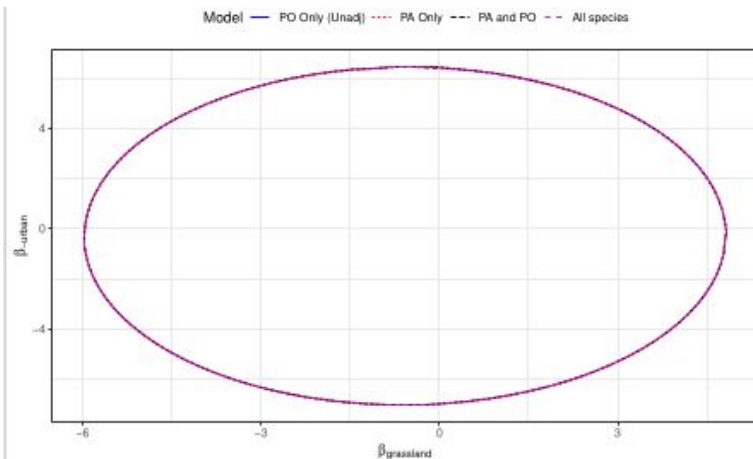
# Results (application)

## Confidence ellipses and intensity map of the mall pearl-bordered fritillary (*Petit collier argenté*)

*Boloria selene*



- Low-density species, no difference between models
- PO not reliable (not shown)
- Species with low thinning-model intercept  $\gamma = -13.31$





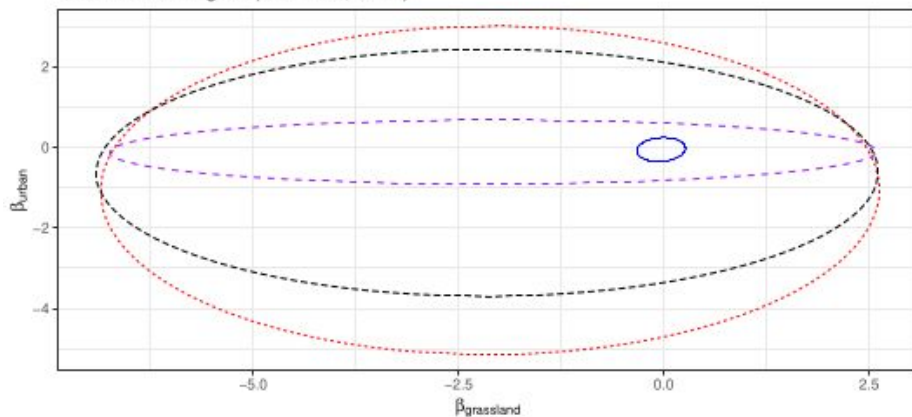
# Results (application)

## Confidence ellipses and intensity map of the wall brown (La Mègère)



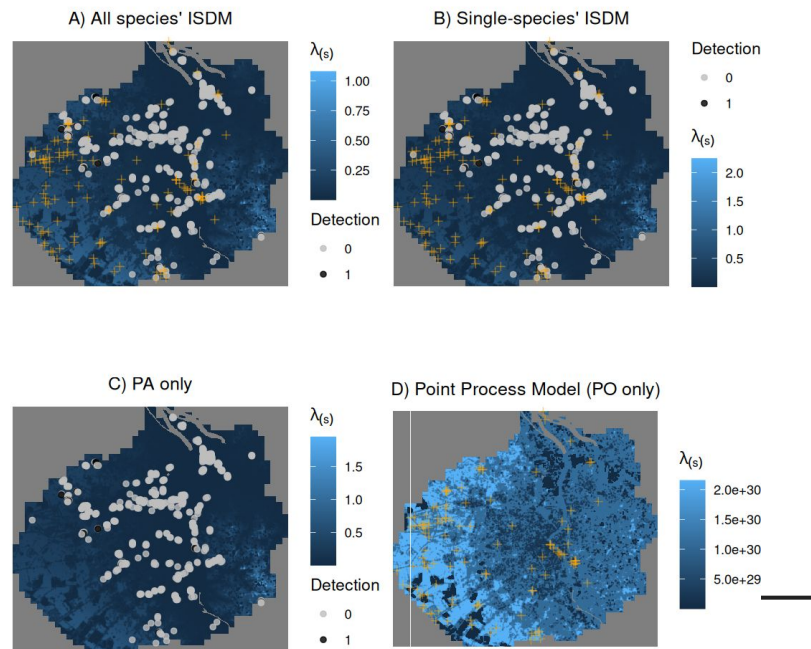
- Low-density species
- All species ISDM was better
- Species with the highest  $\gamma$  ( $\gamma = -7.35$ )

*Lasiommata megera* (Linnaeus, 1767)



Model — PO Only (Unadj) ··· PA Only - - - PA and PO - - - All species

*Lasiommata megera*



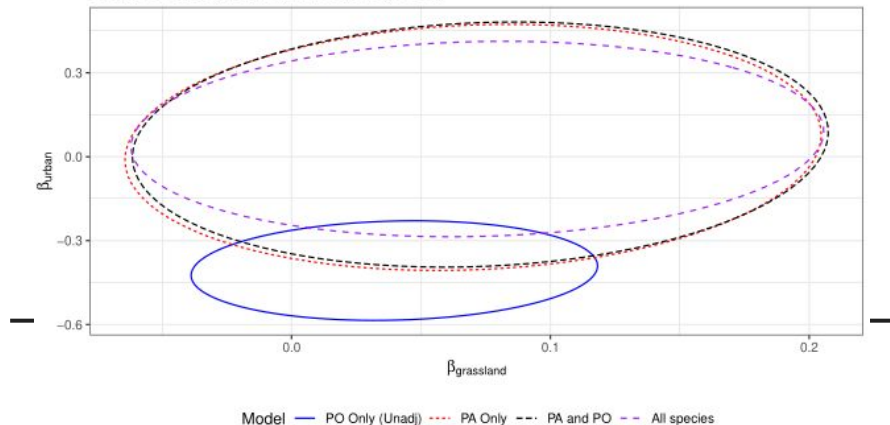
# Results (application)

## Confidence ellipses and intensity map of the clouded yellow (Le Souci)

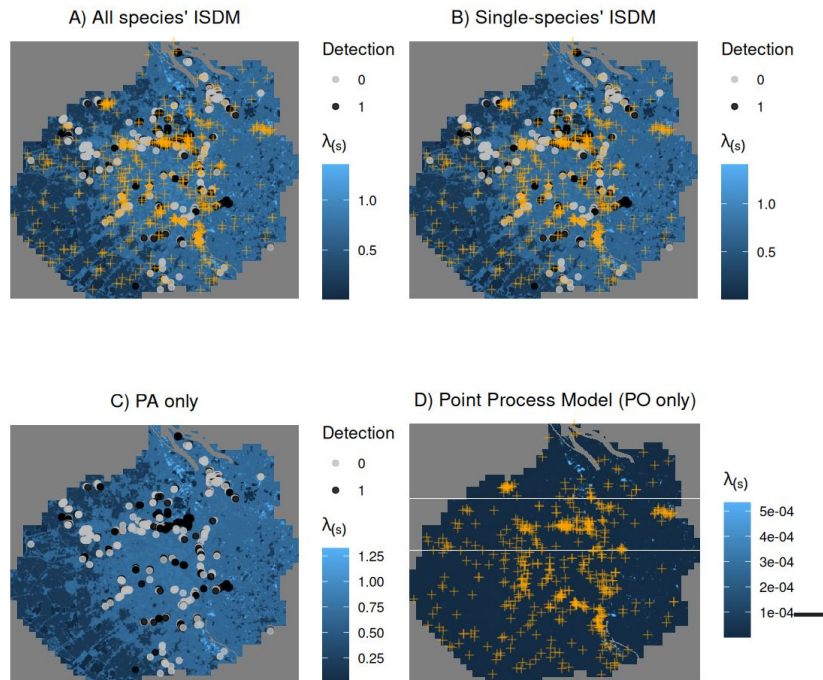


- High-density species
- All species ISDM was better
- Thinning-model intercept  $\gamma = -10.11$

*Colias crocea* (Geoffroy in Fourcroy, 1785)



*Colias crocea*





# Results (application)

## b) Smaller confidence regions for the all species ISDM?

Linear regression to test the influence of data characteristics (Z) on 95% Wald confidence region area (Y)

$$Y = (\gamma_0 + \gamma_{MS}) + (\theta_0 + \theta_{MS}) Z + \epsilon$$

1-Small difference in Wald CR between ISDMs

2-Smaller 95% Wald CR (disregarding the ISDM) when:

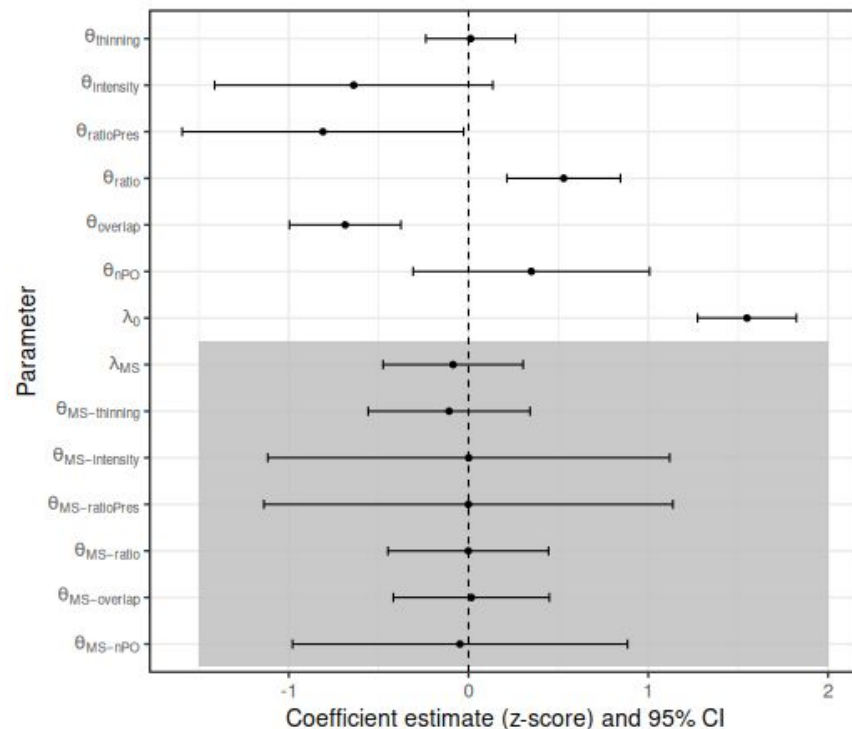
High *overlap* between PA and PO data

Species have high *intensity*  $\exp(\alpha_k)$

High number of presences relative to the number of PA sites (*ratioPres*)

Number of PA sites is larger than the number of PO points (*ratio*)

3-Species-specific thinning intercept  $\gamma_k$ : weaker thinning → CI area of the all species' ISDM decreases



# Results (application)

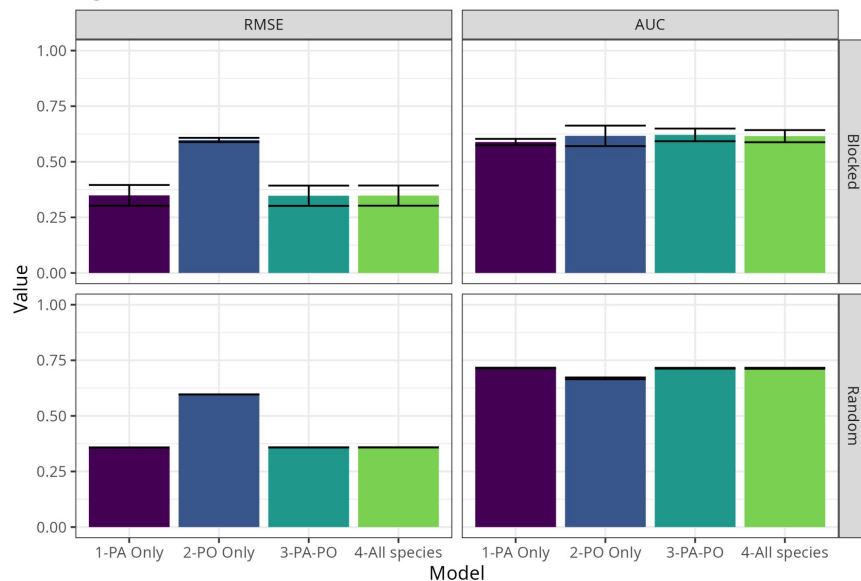


## c) Is predictive performance improved?

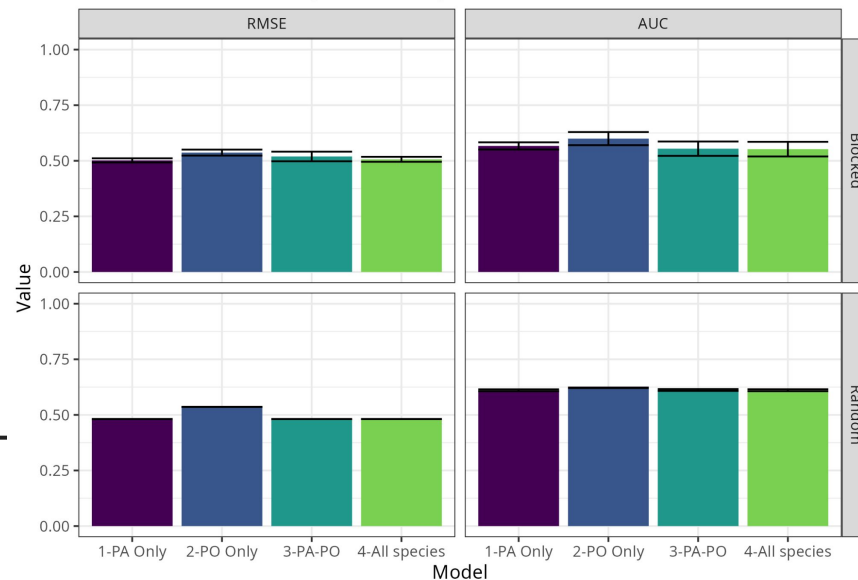
Different CV techniques led to the same conclusion: similar predictive performance across models;



*Aglais io* (Linnaeus, 1758)



*Colias crocea* (Geoffroy in Fourcroy, 1785)

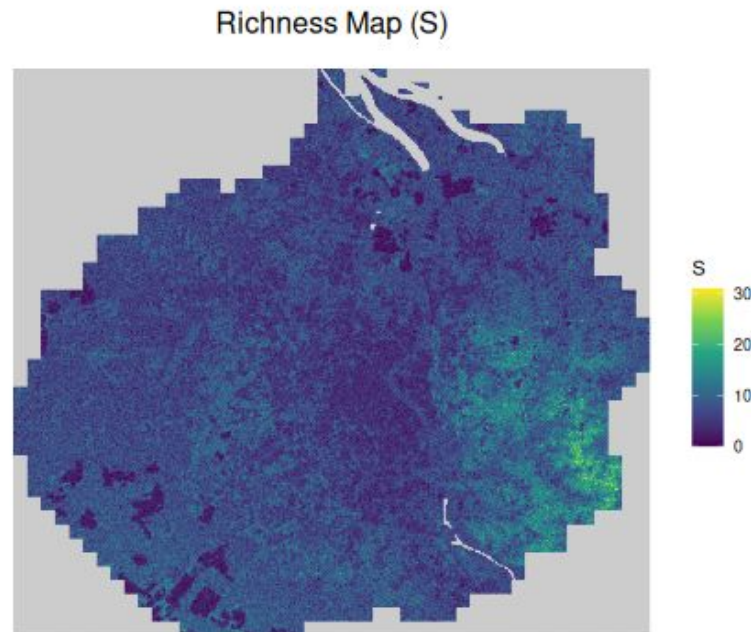


## 5x Cross Validation

- Blocked CV: controlling for spatial autocorrelation
- Random: no control
- Average(SE) of RMSE and AUC across folds
- Six species evaluated

# Take home message

- In simulations, the all species' ISDM did not perform better than single species (PA-PO) ISDM.
- ISDMs were better than models of single data sets (as in Dorazio (2014, GEB), Koshkina et al. (2017, MEE), Peel et al. (2019, MEE))
- In empirical analyses, for not so rare species, the all species' ISDMs performed better than the other models to estimate coefficients
  - Improvements on predictive performance were minimal (all spp vs. single spp).
  - Is such a small difference enough to favor the all species integrated model?
- We expect to discuss the results in light of urban ecology questions.



# Work in more advanced stages ...

## Article 1 & Rescience C

arXiv > stat > arXiv:2510.08151

Statistics > Applications

[Submitted on 9 Oct 2025]

**Evaluating multi-season occupancy models with autocorrelation fitted to heterogeneous datasets**

André Luís Luza, Didier Alard, Frédéric Barraquand

UNDER REVIEW

RESCIENCE C

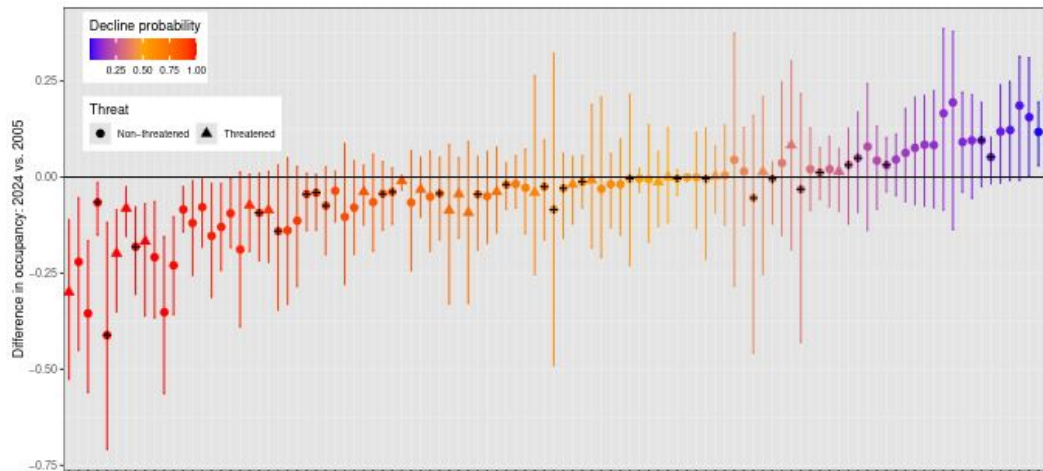
Replication / ecology

**[~Re] Bias correction in species distribution models: pooling survey and collection data for multiple species**

André Luza<sup>1,\*</sup> and Frédéric Barraquand<sup>2,\*</sup>

<sup>1</sup>Biodiversité Gènes et Communautés (UMR 1202), University of Bordeaux, INRAE, Pessac, France – <sup>2</sup>Institut de Mathématiques de Bordeaux (UMR 5251), University of Bordeaux, CNRS, Bordeaux INP, Talence, France

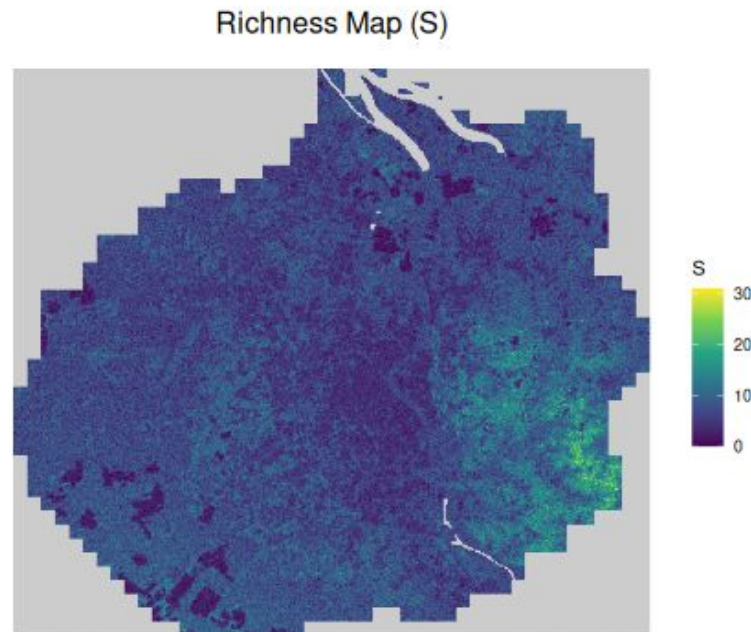
## Article 2: A two-decade analysis of butterfly occupancy trends in SW France reveals more declining losers than winners



Avignon, February / 2026

# Take home message

- In simulations, the all species' ISDM did not perform better than single species (PA-PO) ISDM.
- ISDMs were better than models of single data sets (as in Dorazio (2014, GEB), Koshkina et al. (2017, MEE), Peel et al. (2019, MEE))
- In empirical analyses, for not so rare species, the all species' ISDMs performed better than the other models to estimate coefficients
  - Improvements on predictive performance were minimal (all spp vs. single spp).
  - Is such a small difference enough to favor the all species integrated model?
- We expect to discuss the results in light of urban ecology questions.



# Additional slides





# References

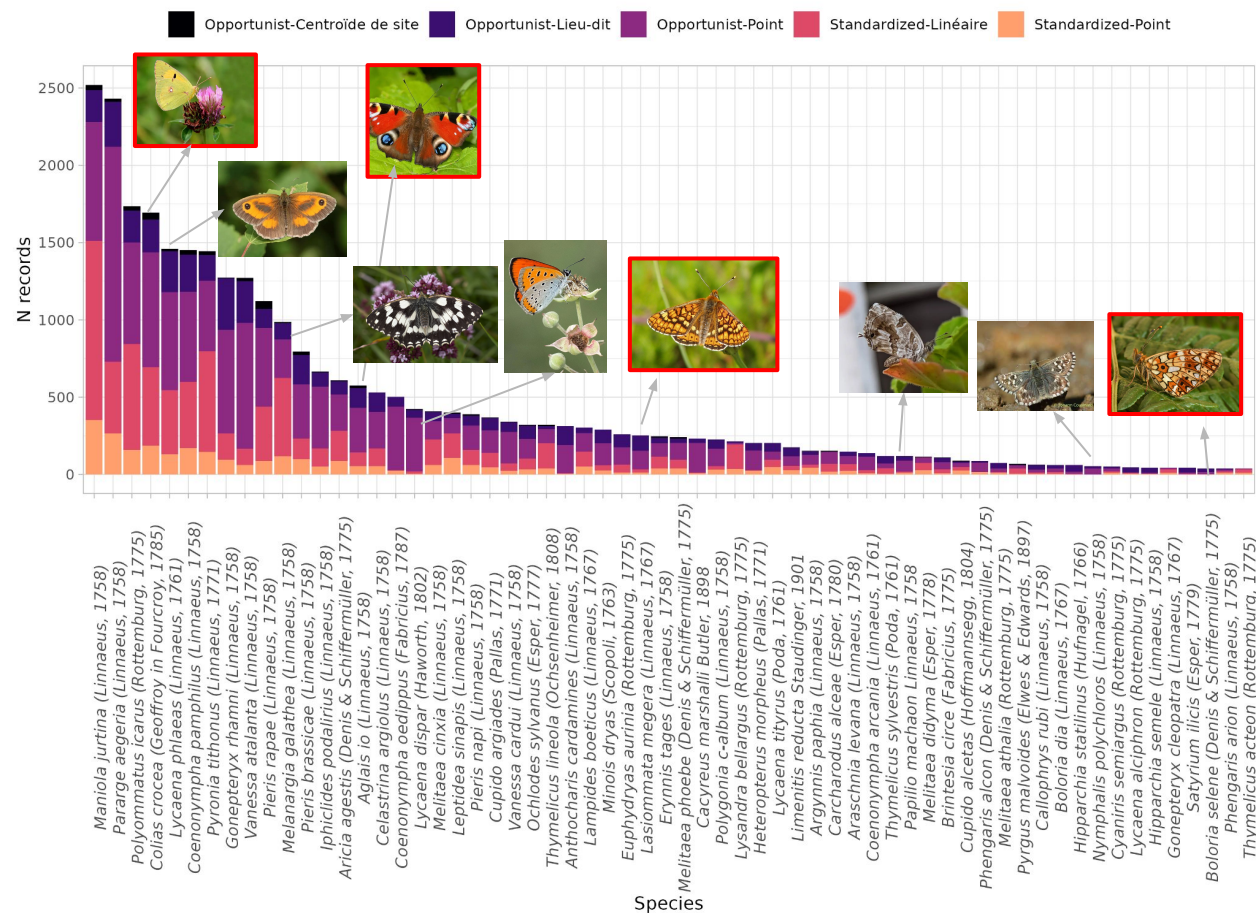
- Barbosa, W. L., & Alves-Souza, S. N. (2025). Data quality issues in data used in species distribution models: A systematic literature review. *Ecological Informatics*, 103378.
- Rapacciuolo, G., Young, A., & Johnson, R. (2021). Deriving indicators of biodiversity change from unstructured community-contributed data. *Oikos*, 130(8), 1225-1239.
- Fletcher Jr, R. J., Hefley, T. J., Robertson, E. P., Zuckerberg, B., McCleery, R. A., & Dorazio, R. M. (2019). A practical guide for combining data to model species distributions. *Ecology*, 100(6), e02710.
- Warton, D.I. & Shepherd, L.C. (2010). Poisson point process models solve the pseudo-absence problem" for presence-only data in ecology. *The Annals of Applied Statistics*, pp. 1383-1402.

## ISDMs

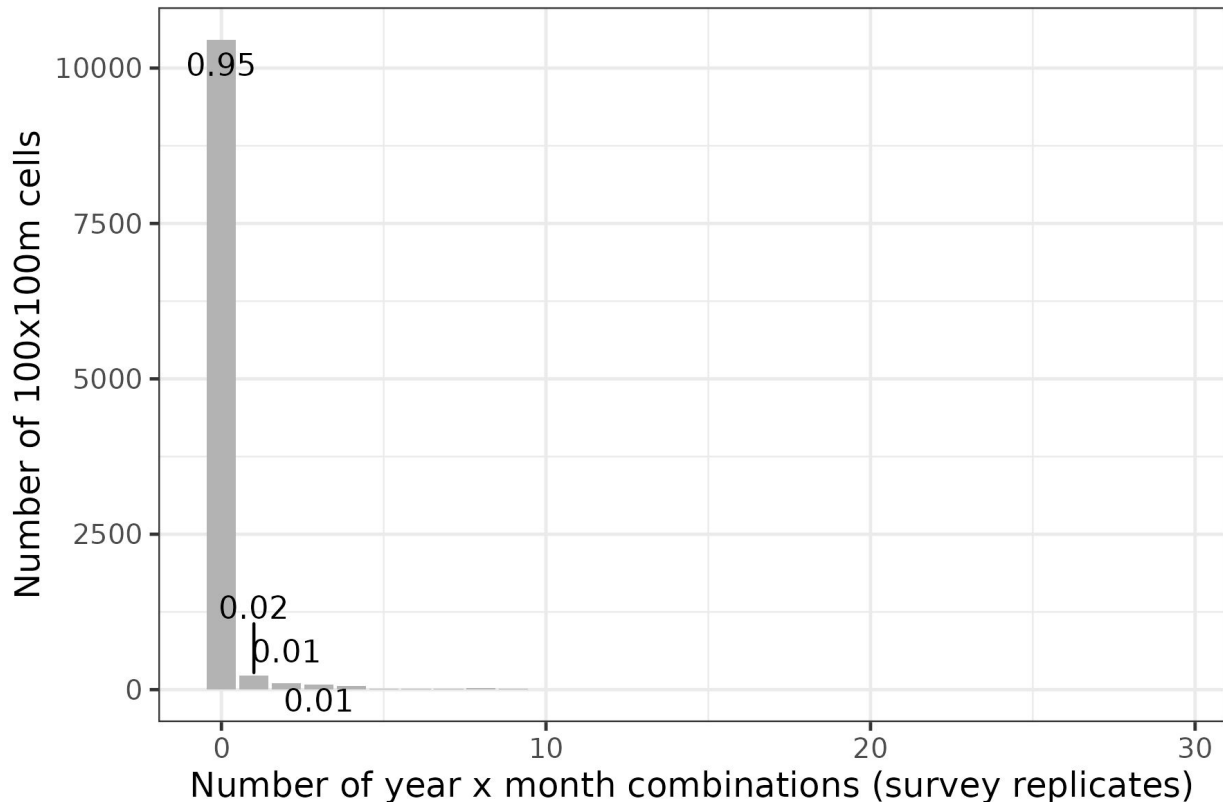
- Dorazio, R.M. (2014). Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography*, 23, 1472-1484
- Doser, J. W., Leuenberger, W., Sillett, T. S., Hallworth, M. T. & Zipkin, E. F. (2022). Integrated community occupancy models: A framework to assess occurrence and biodiversity dynamics using multiple data sources. *Methods in Ecology and Evolution*, 13, 919-932. <https://doi.org/10.1111/2041-210X.13811>
- Fithian, W., Elith, J., Hastie, T. & Keith, D.A. (2015). Bias correction in species distribution models: Pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, 6, 424-438
- Fletcher, R.J., McCleery, R.A., Greene, D.U. et al. Integrated models that unite local and regional data reveal larger-scale environmental relationships and improve predictions of species distributions. *Landscape Ecol* 31, 1369-1382 (2016). <https://doi.org/10.1007/s10980-015-0327-9>
- Isaac, N.J.B., et al. (2020). Data Integration for Large-Scale Models of Species Distributions. *Trends in Ecology & Evolution*, 35, 56-67.
- Miller, D.A.W., Pacifici, K., Sanderlin, J.S. & Reich, B.J. (2019). The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution*, 10, 22-37



# Number of records per butterfly species

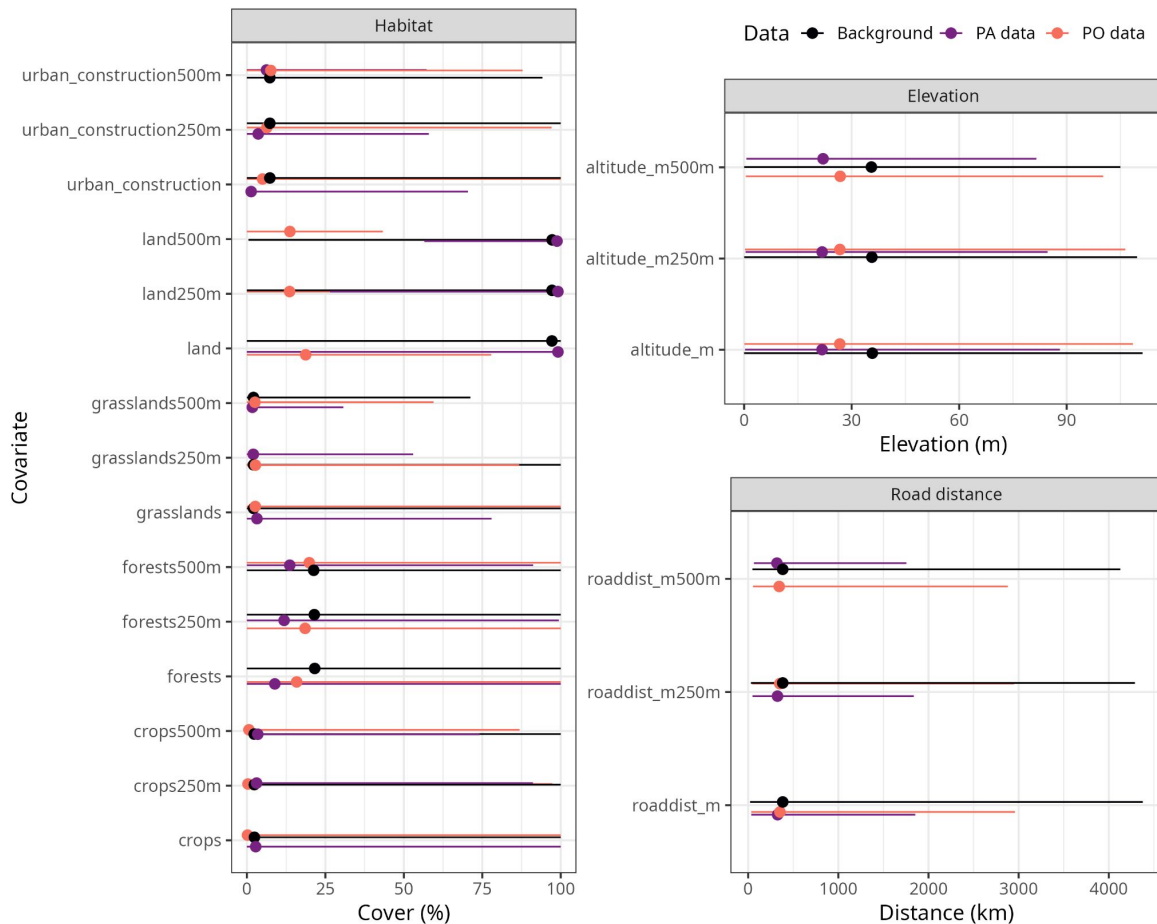


# Levels of sampling replication



- We don't have enough survey replication to correct for imperfect detection using occupancy-like ISDMs (e.g. Koshkina et al. 2017, Doser et al. 2022)

# Coverage of environmental and bias gradients

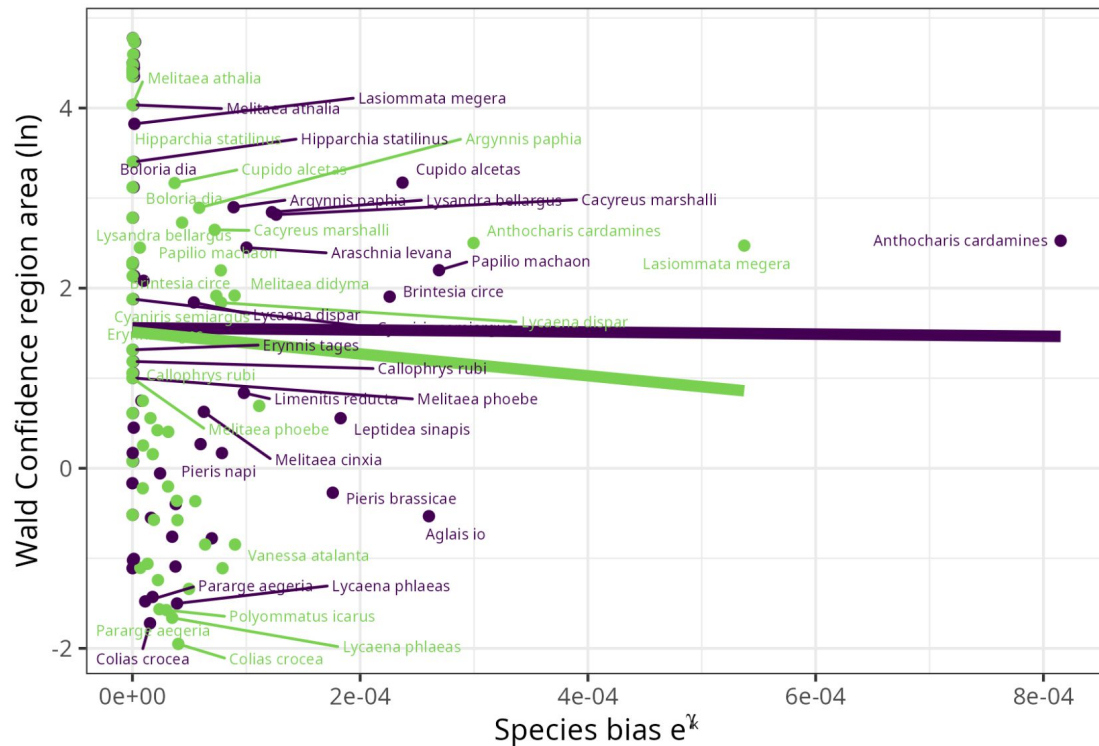


Environmental and bias gradients well covered by the data sets

- Larger discrepancy for land cover (100 - % water cover) in the PO data

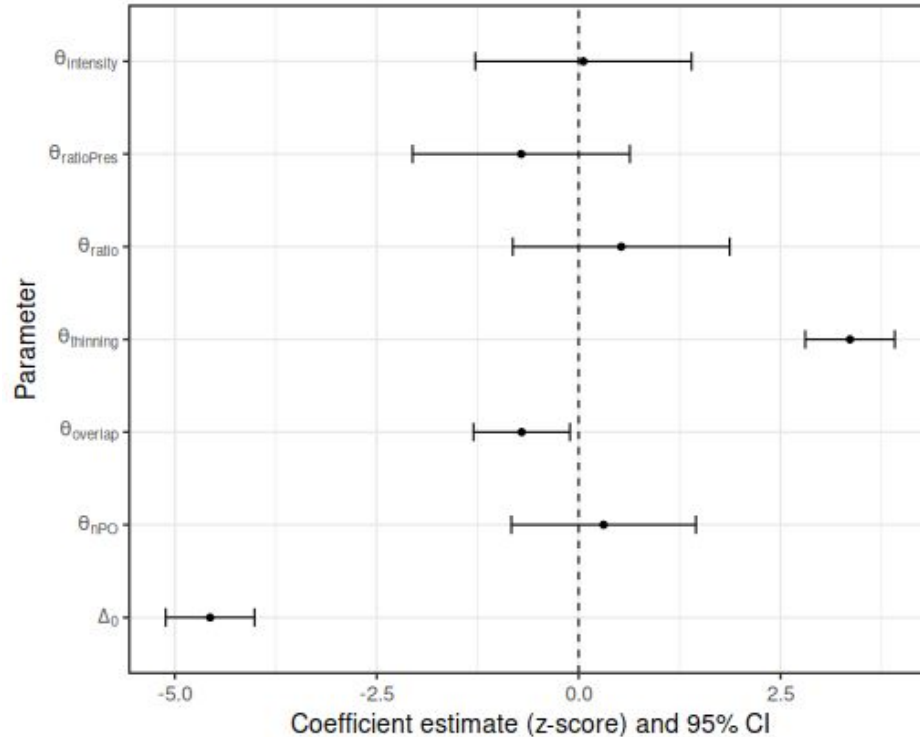
# Bias estimates

- Relationship between estimated bias (x-axis) and 95% Wald CI area



# Analysis of the difference of CI between PA-PO and all species model

- The species-specific thinning intercept ( $\gamma_k$ ) and overlap explain the difference





# Map of the bias

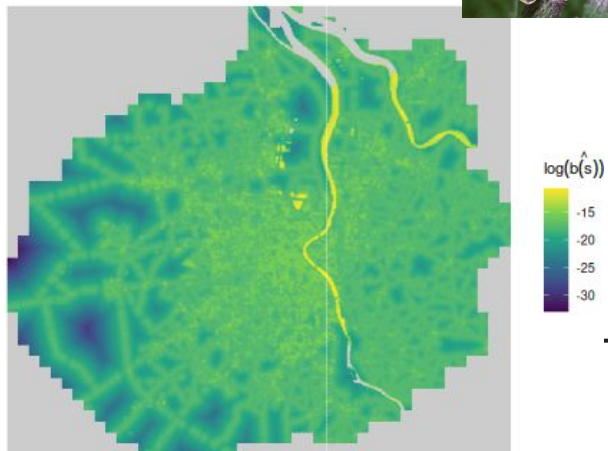
- The bias model has two components:
  - The species-specific intercept ( $\gamma_k$ ) not related to covariates
  - The proportional-bias coefficient ( $\delta$ ) constant across species, which depicts the effect of the covariates on the spatial thinning.

Low-intensity species for which the all species ISDM performed better than other models

Lower thinning

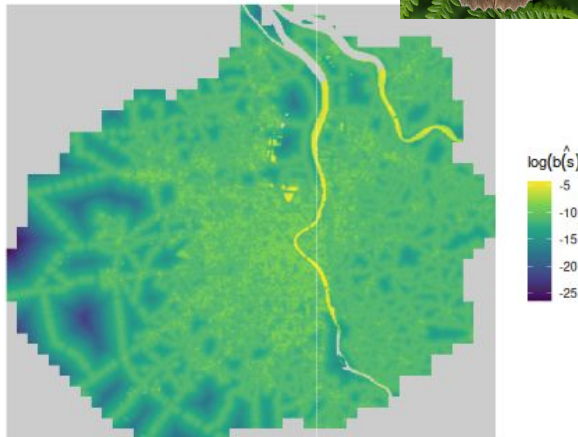
$\gamma = -18.36$

*Thymelicus lineola* (Ochsenheimer, 1808)



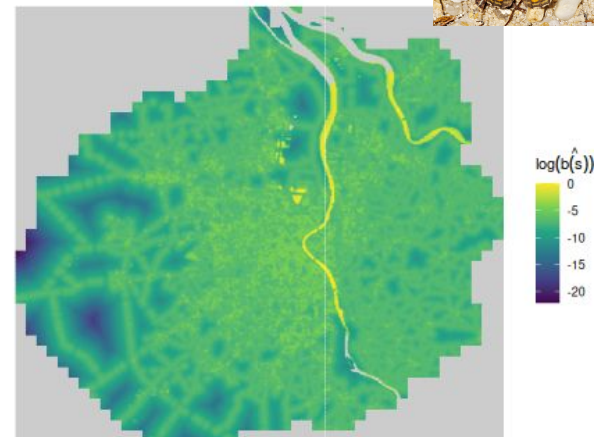
$\gamma = -11.87$

*Maniola jurtina* (Linnaeus, 1758)

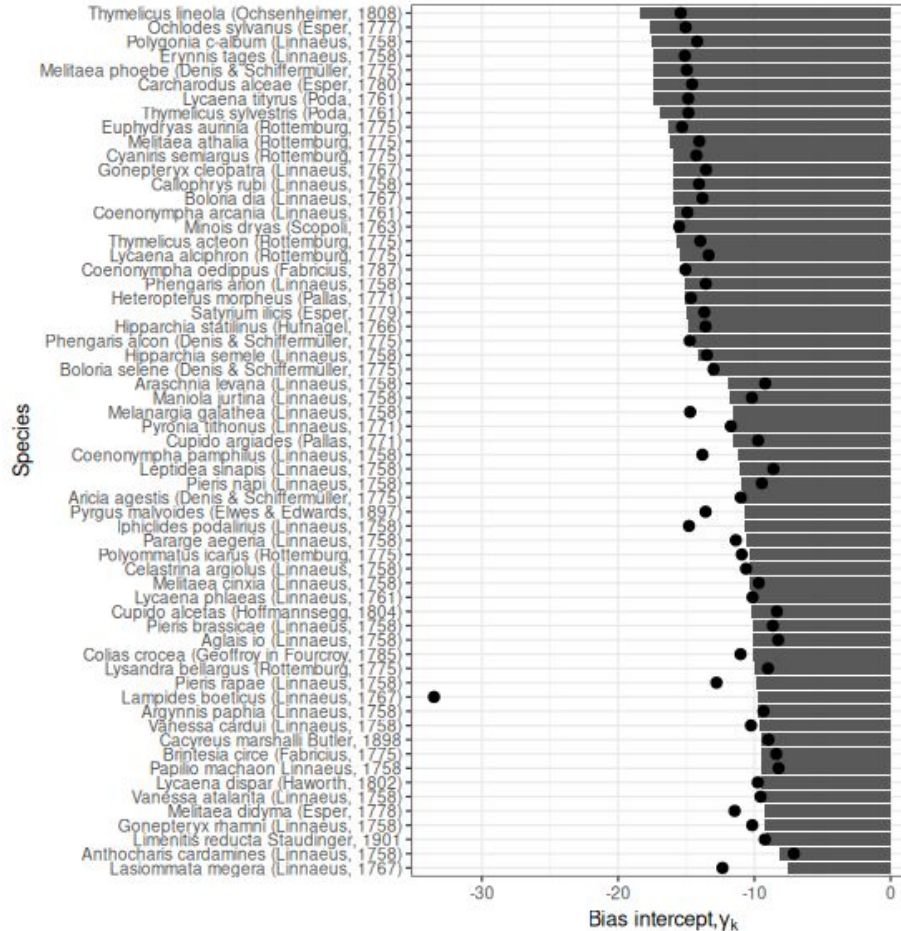


$\gamma = -7.54$

*Lasiommata megera* (Linnaeus, 1767)



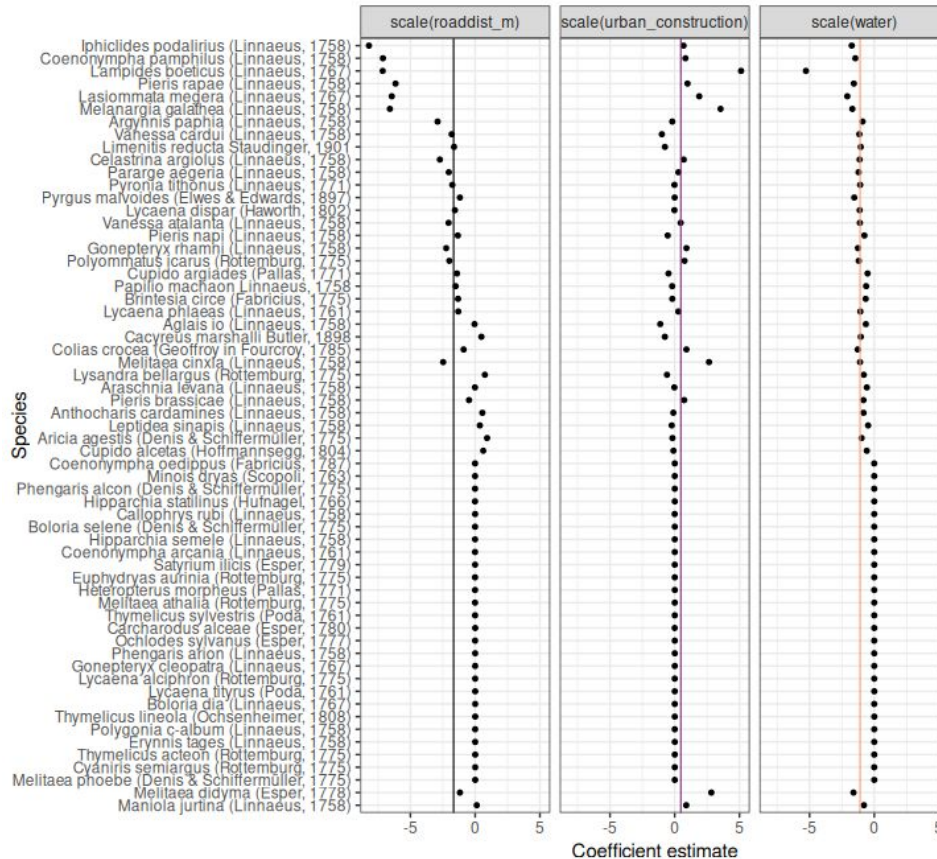
# Species-specific intercept ( $\gamma_k$ )



Estimates of the bias intercept

- Horizontal bars → all species ISDM
- Points → PA-PO model

# Proportional-bias effect ( $\delta$ )



Estimates of the proportional bias effect  $\delta$  (vertical bars)

Estimates of the proportional effect per species  $\delta_k$  (points), which is obtained through the PA-PO model

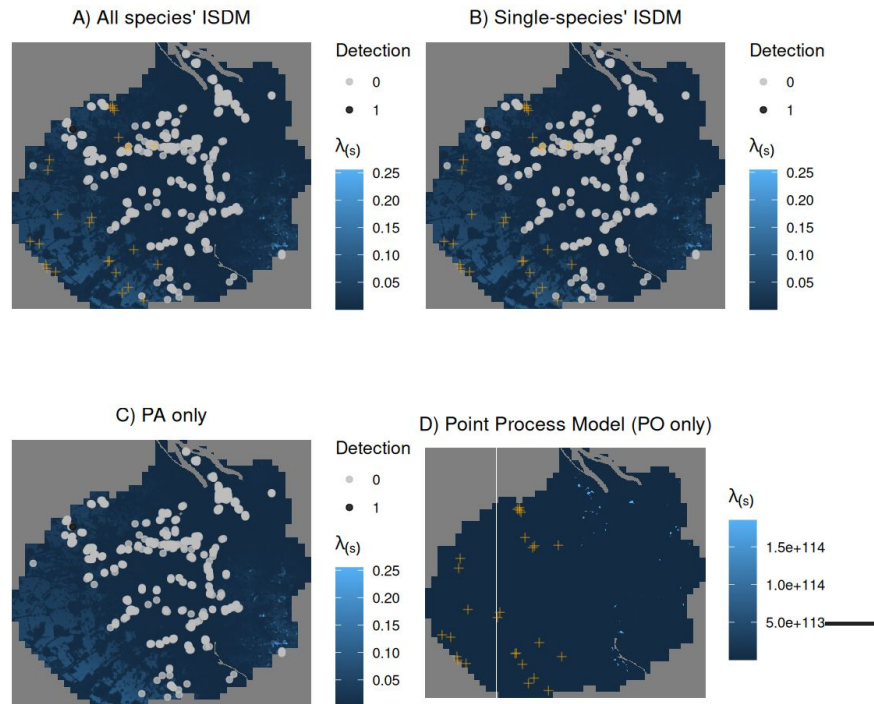
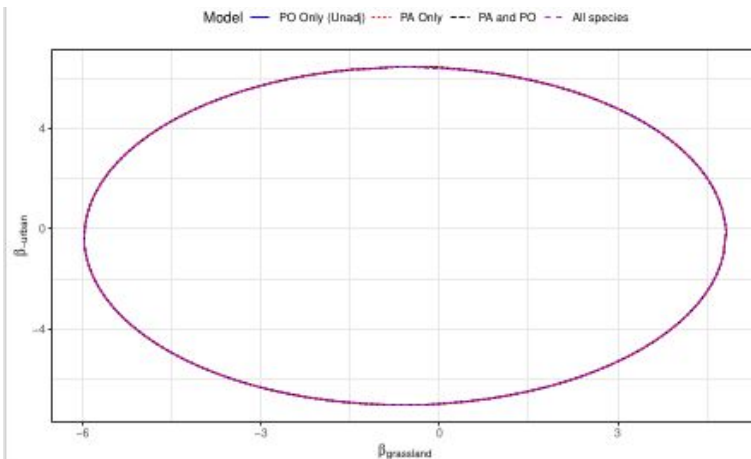
# Results (application)

## Confidence ellipses and intensity map of the mall pearl-bordered fritillary (*Petit collier argenté*)

*Boloria selene*



- Low-density species, no difference between models
- PO not reliable (not shown)
- Species with low thinning-model intercept  $\gamma = -13.31$



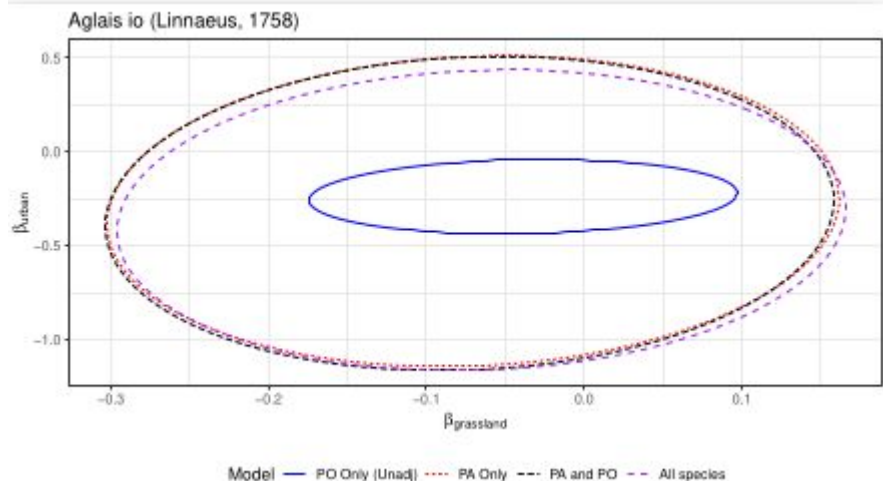


# Results (application)

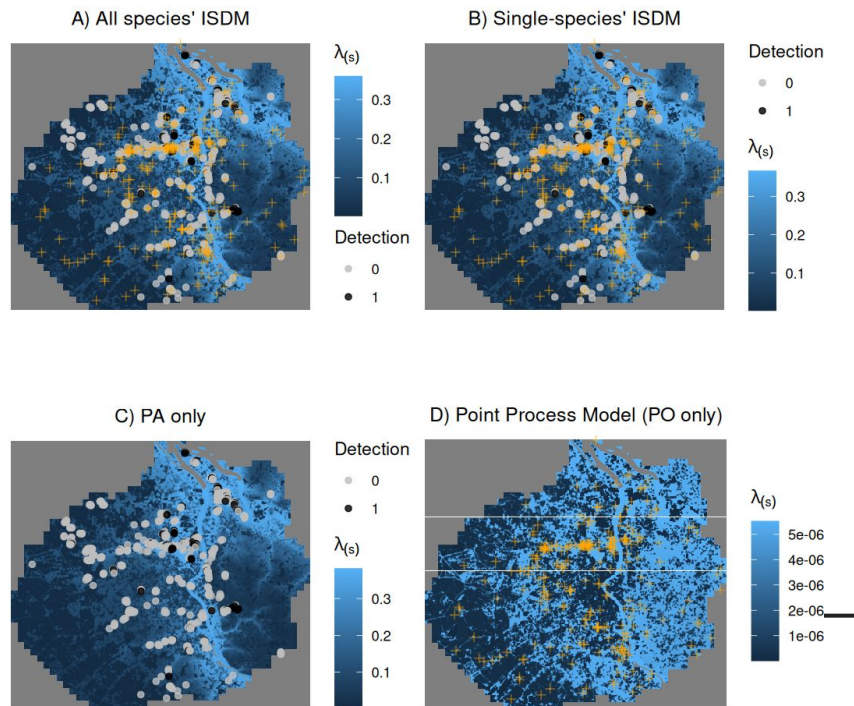
## Confidence ellipses and intensity map of the Peacock butterfly (Paon-du-jour)



- Low-density species
- All species ISDM was subtle better
- Thinning-model intercept  $\gamma = -10.13$

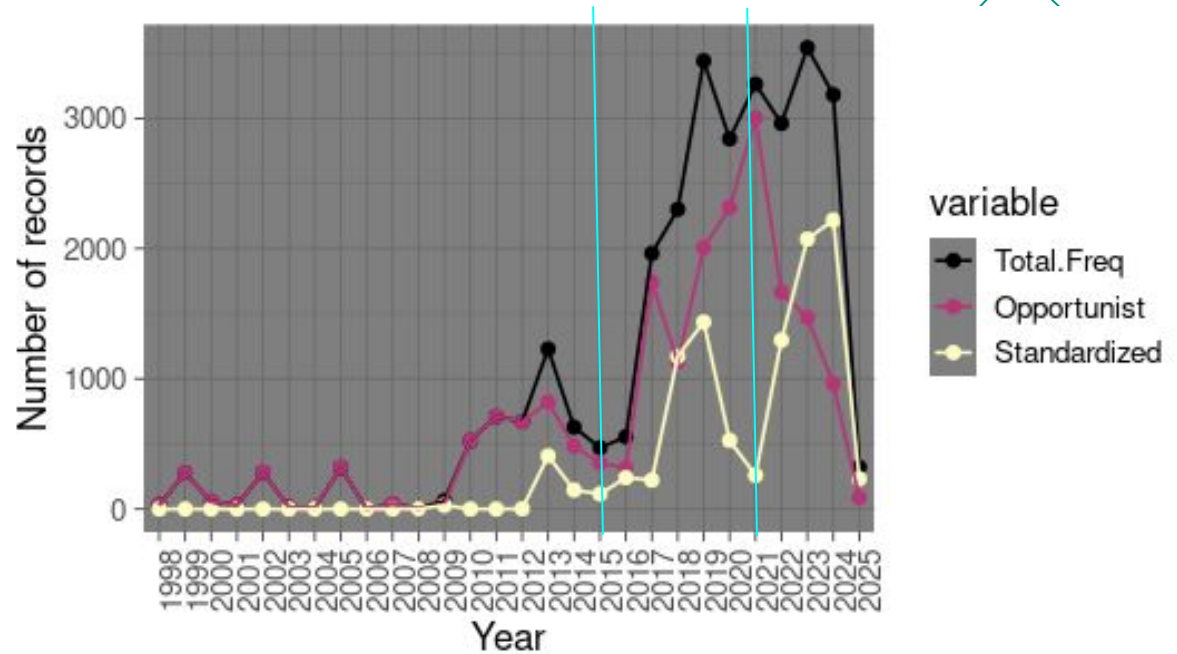


### *Aglais io*





# Amount of data over time



Number of butterfly records over time (total and per data type)

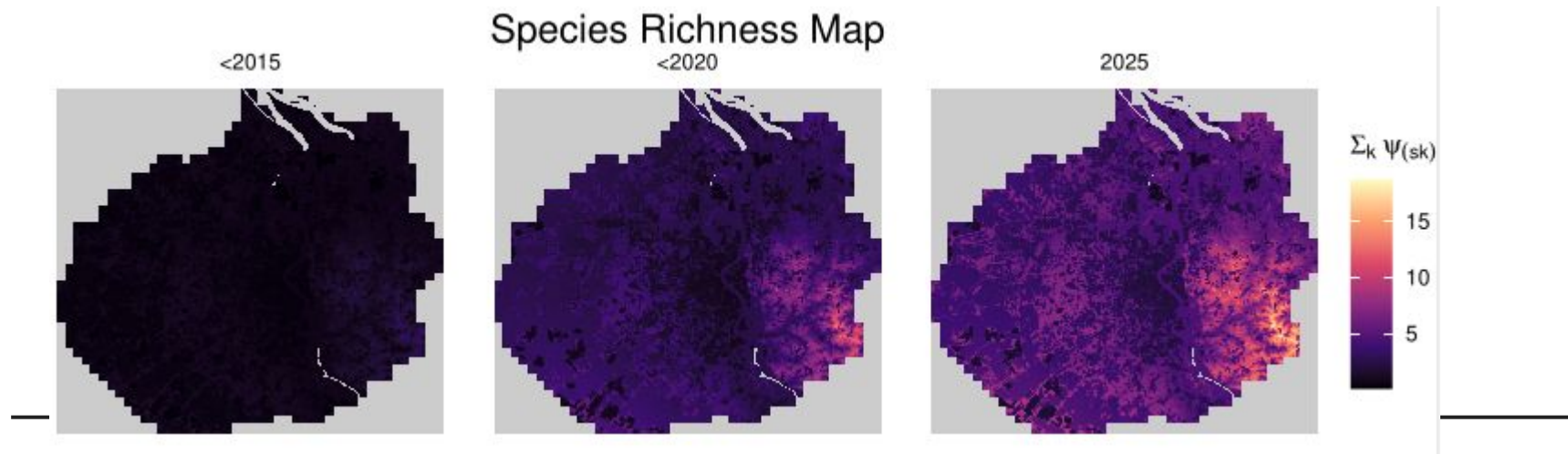


# Assessment over time

Data splitted in three periods.

Occupancy  $\psi_{(s)k}$  obtained from the estimated intensity using the cloglog link function, and summed across species to obtain richness per cell and period

Results reflect the amount of data in each period



# Cross-validation

5x cross validation (blockCV R package)

Blocked

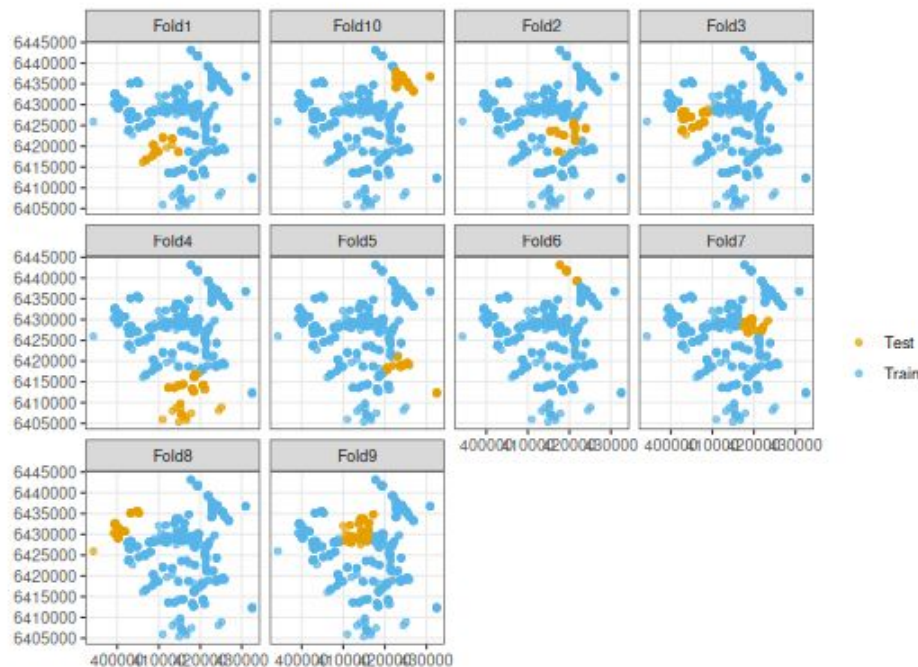
Folds based on lat-long coordinates of PA cells

PO data and background data inside the convex hull of test data were held out

Random

Random sort of PA cells and PO points in each fold (20% test, 80% train)

PO data and background data inside the test-data convex hull were held out



Use for  
longer talks



# Outline

- Context about PPM and ISDMs
- The model
- Objectives & Methods
- Results
  - **Simulations**
  - **Application to butterflies**
- Take home message



development in SDMs → foundation of integrated species  
distribution models (Warton & Shepherd 2010, Royle et al. 2012,  
**Context** Yackulic et al. 2013, Renner et al. 2013, Fithian et al. 2014, 2015).

*Presence-only data* = point locations

Opportunity for modeling species distribution

*Warning: Nonprobability samples*

The PPP treats the number and location of discrete points  
as random quantities governed by a Poisson distribution  
and a continuous intensity field  $\lambda$ :

$$\mathcal{S} \sim \text{PPP}(\lambda)$$

