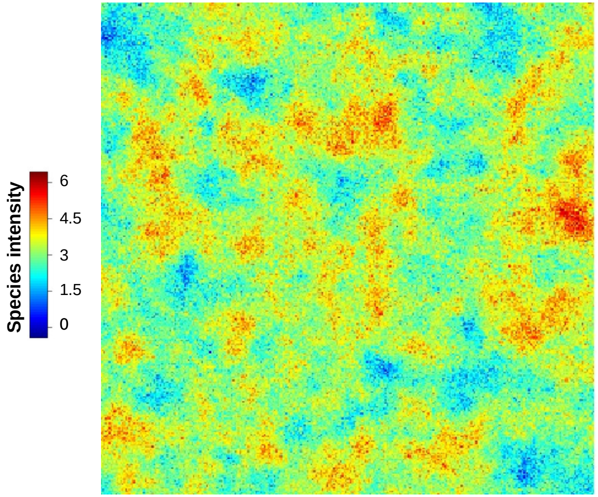
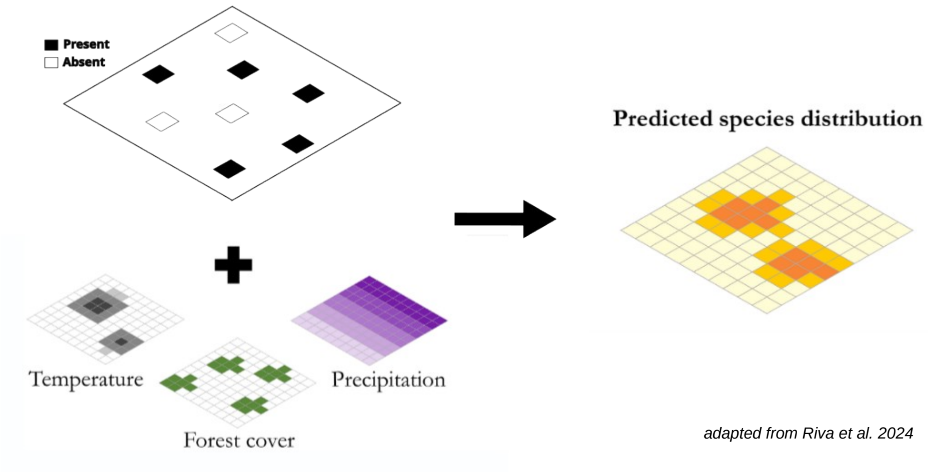
Best of Both Worlds?

Integrating standardized and opportunistic data for predictive modelling of plant pathogen distributions

***Bénard A**, Lasgorceux F, Papaïx J, Opitz T, Bunz Y, Combrisson D*

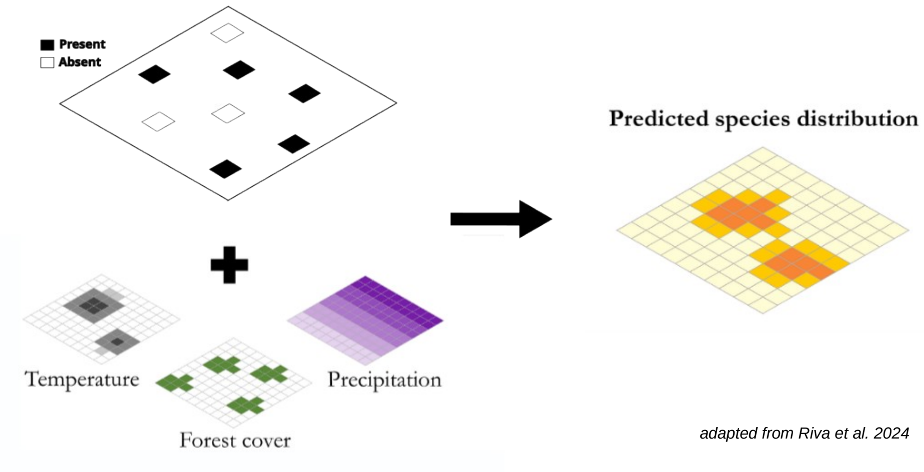


Species Distribution Models

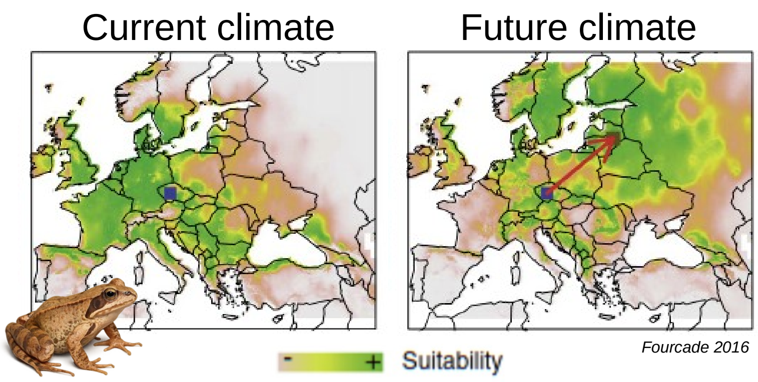


adapted from Riva et al. 2024

Species Distribution Models



adapted from Riva et al. 2024

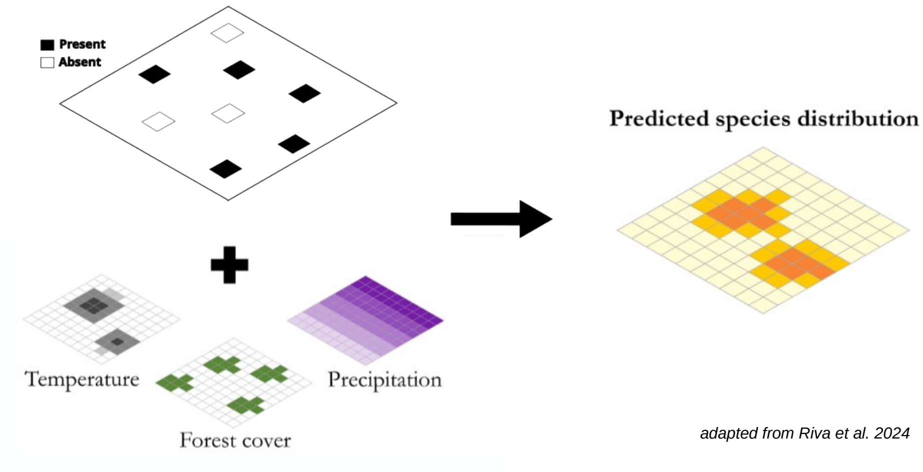


Fourcade 2016

MaxEnt, Bioclim, Random Forests, ANNs, GLM, GAM, …

- Forecast future shifts in species range

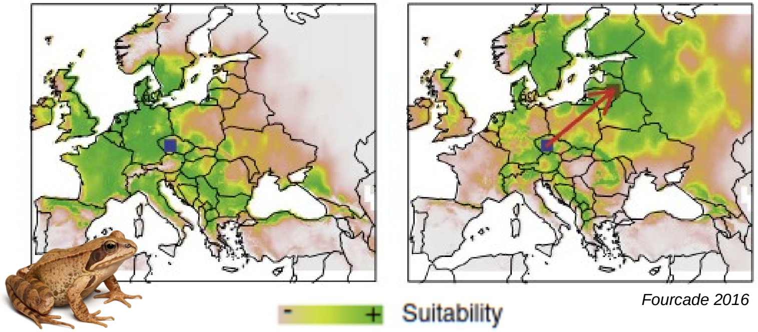
Species Distribution Models



adapted from Riva et al. 2024

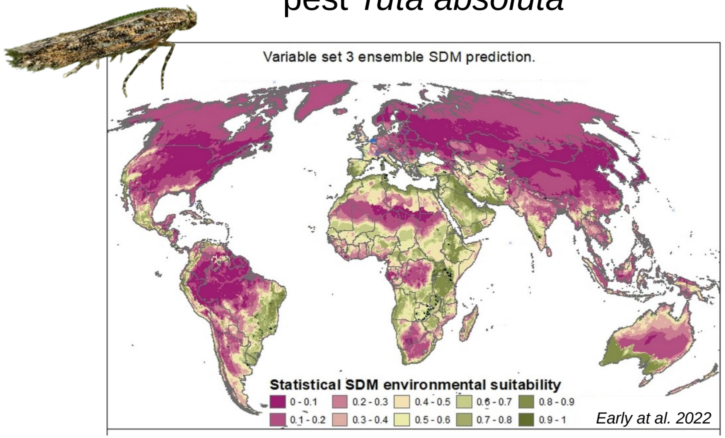
Current climate

Future climate



Fourcade 2016

Suitable habitat for invasive crop pest *Tuta absoluta*



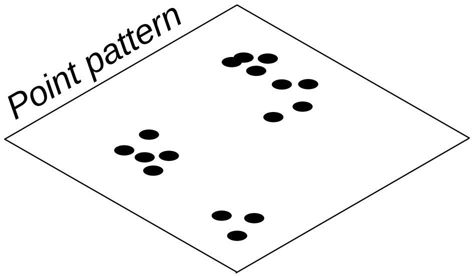
Early at al. 2022

MaxEnt, Bioclim, Random Forests, ANNs, GLM, GAM, …

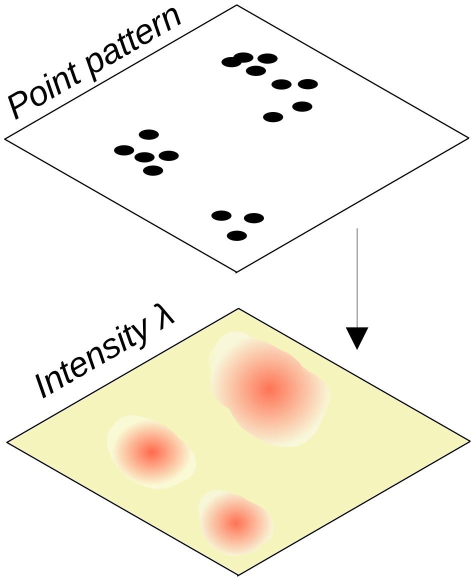
- Forecast future shifts in species range

- Identify suitable habitats in new areas

Species Distribution Models

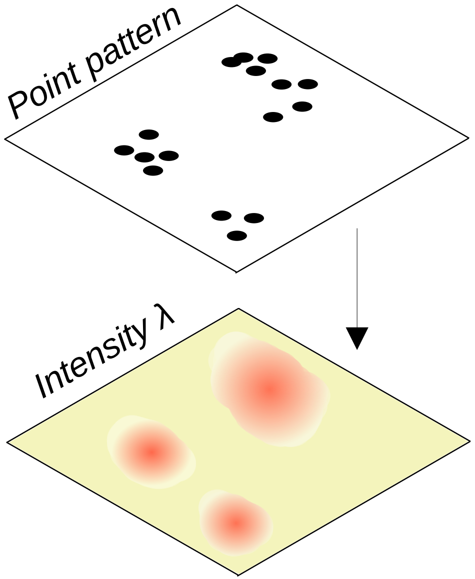


Species Distribution Models

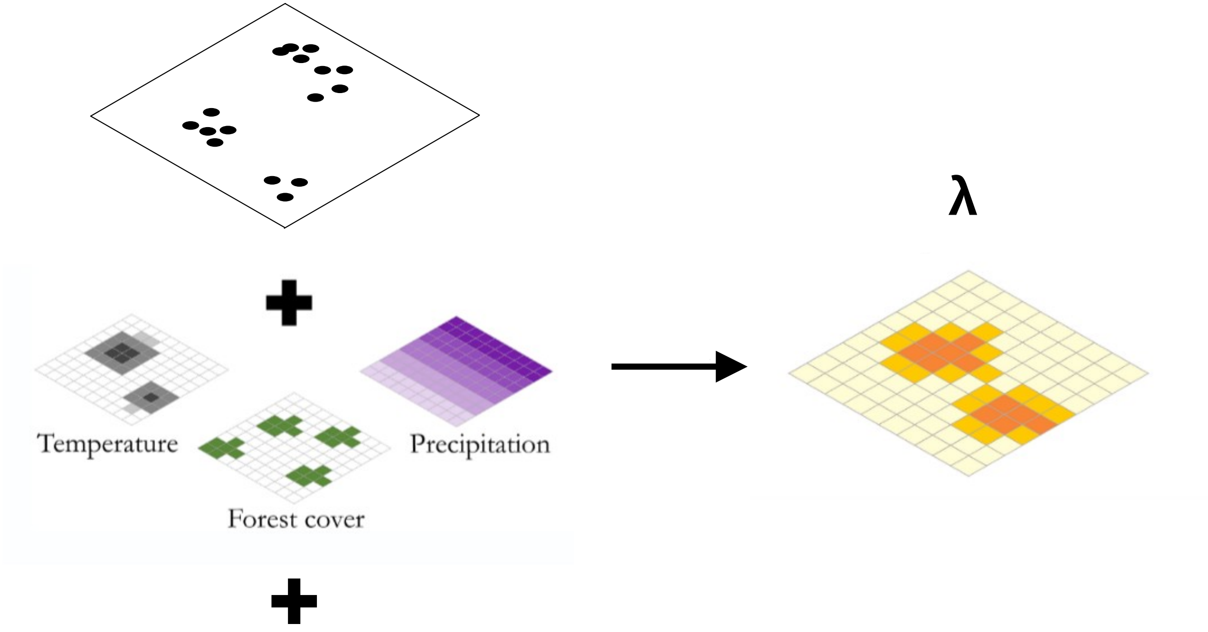


Inhomogeneous point process:
Log Gaussian Cox Process

Species Distribution Models



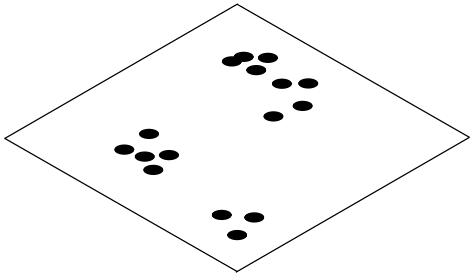
Inhomogeneous point process:
Log Gaussian Cox Process



Latent Gaussian Random Field
(spatial autocorrelation)

Species Distribution Models

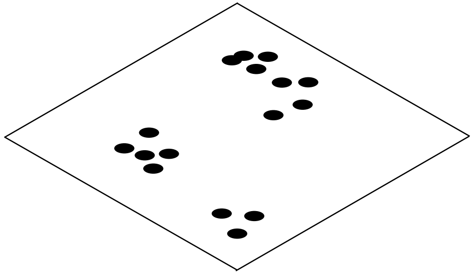
Presence-only data



$$\lambda(s)=\exp\left(\beta_0+\sum_{j=1}^{p}\beta_j x_j(s)+W(s)\right)$$

Species Distribution Models

Presence-only data

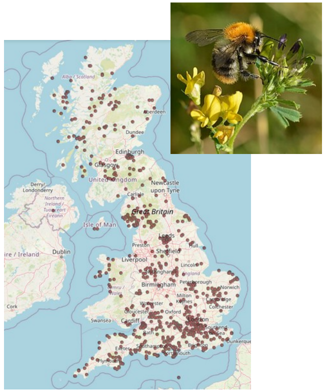


$$\lambda(s) = \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j(s) + W(s)\right)$$

Opportunistic species observations

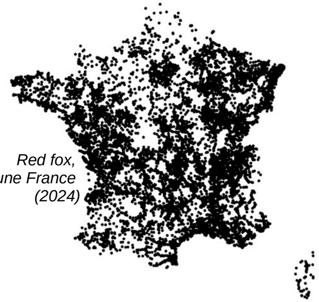
- Museum records & Atlases

Carder bee



Natural History Museum

- Citizen sciences



Red fox, Faune France (2024)

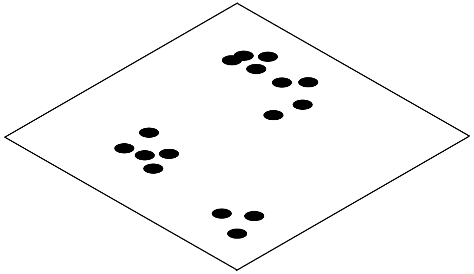
Faune France

iNaturalist

Pl@ntNet

Species Distribution Models

Presence-only data

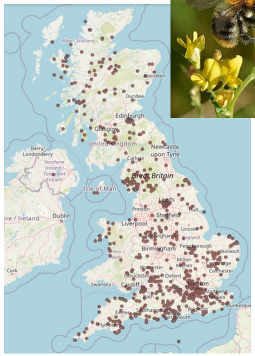


$$\lambda(s) = \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j(s) + W(s)\right)$$

Opportunistic species observations

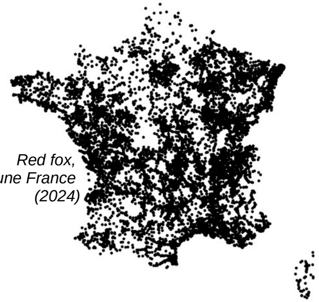
- Museum records & Atlases

Carder bee



Natural History Museum

- Citizen sciences



Red fox, Faune France (2024)

Faune France

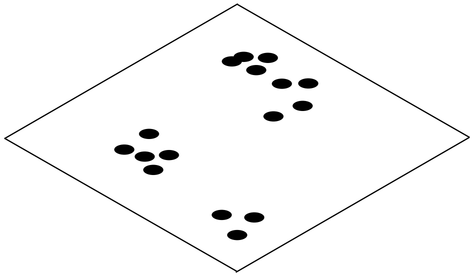
iNaturalist

Pl@ntNet

→ Sampling bias

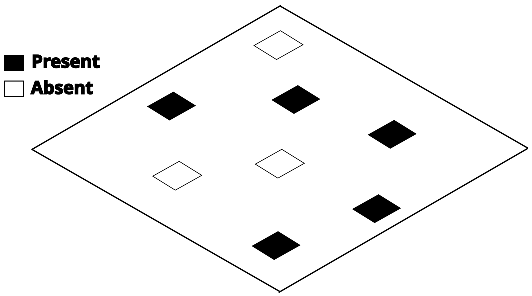
Species Distribution Models

Presence-only data



$$\lambda(s)=\exp\left(\beta_0+\sum_{j=1}^{p}\beta_j x_j(s)+W(s)\right)$$

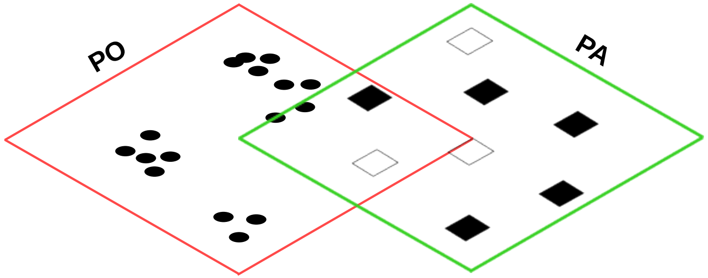
Presence/Absence data



$$cloglog(p_i)=\beta_0+\sum_{j=1}\beta_j x_j(s)+W(s)$$

Species Distribution Models

Joint modelling



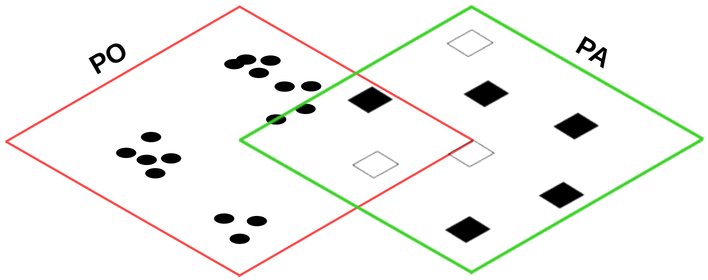
- $\lambda(s) = \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j(s) + W(s)\right)$
- $cloglog(p_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j(s) + W(s)$

joint likelihood

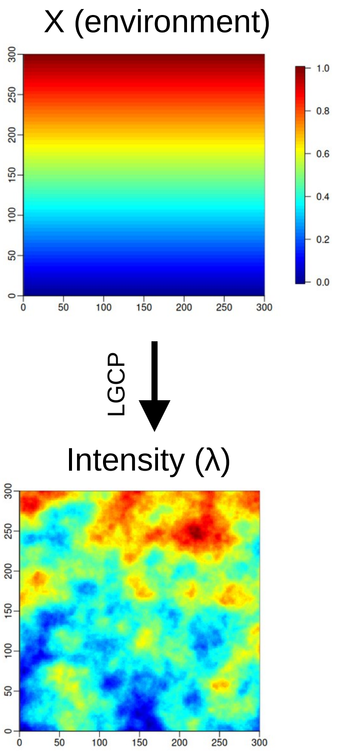
<u>Simulation study of joint SDMs</u>

Simmonds et al. (2020), *Ecography*

Joint modelling



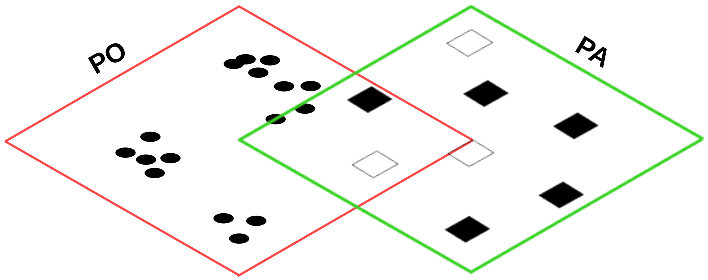
- $\lambda(s) = \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j(s) + W(s)\right)$
- $cloglog(p_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j(s) + W(s)$



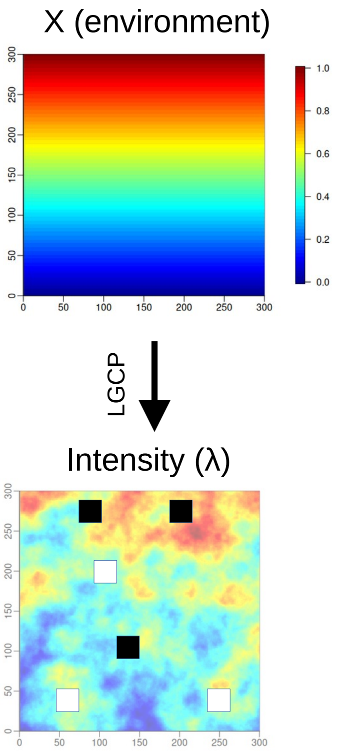
Simulation study of joint SDMs

Simmonds et al. (2020), *Ecography*

Joint modelling



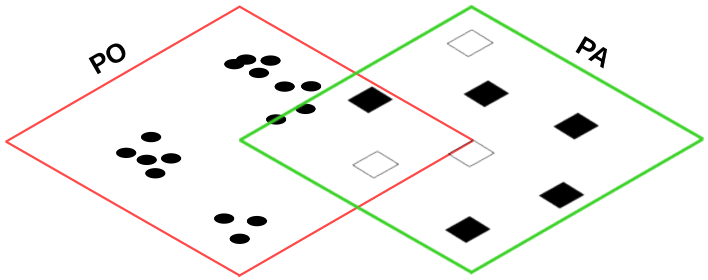
- $\lambda(s) = \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j(s) + W(s)\right)$
- $cloglog(p_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j(s) + W(s)$



Presence/Absence data

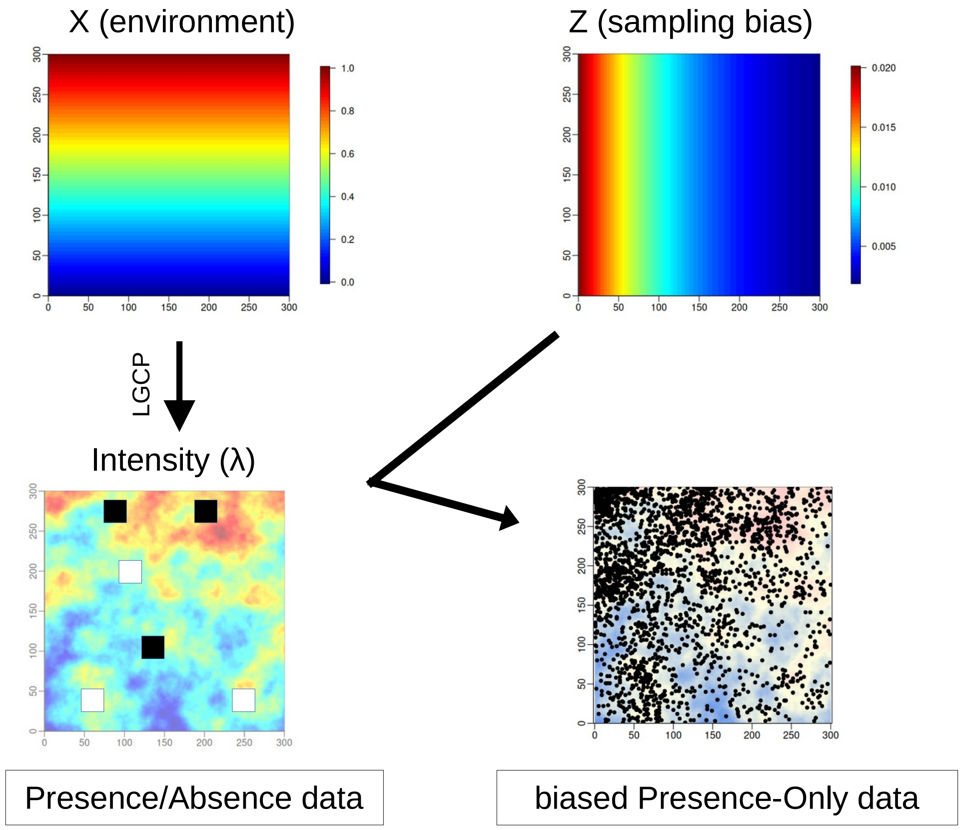
Simulation study of joint SDMs

Joint modelling



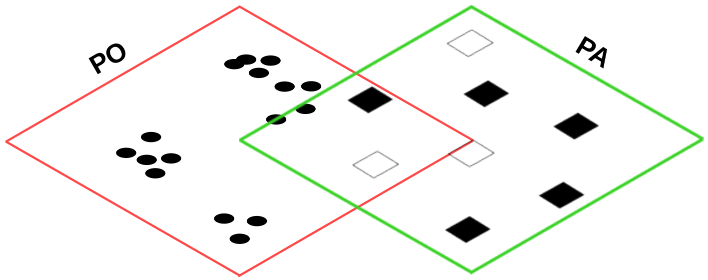
- $\lambda(s) = \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j(s) + W(s)\right)$
- $cloglog(p_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j(s) + W(s)$

Simmonds et al. (2020), *Ecography*



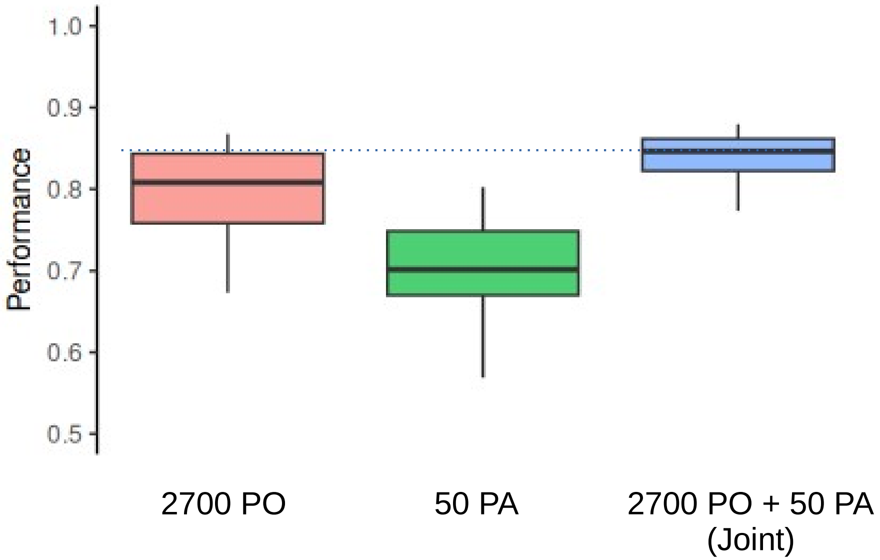
Simulation study of joint SDMs

Joint modelling



- $\lambda(s) = \exp\left(\beta_0 + \beta_1 x(s) + \beta_{bias} z(s) + W(s)\right)$
- $cloglog(p_i) = \beta_0 + \beta_1 x(s) + W(s)$

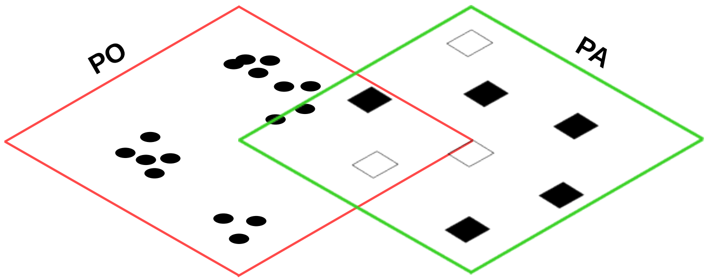
Joint modelling > Single-source model



Simmonds et al. (2020), *Ecography*

Simulation study of joint SDMs

Joint modelling

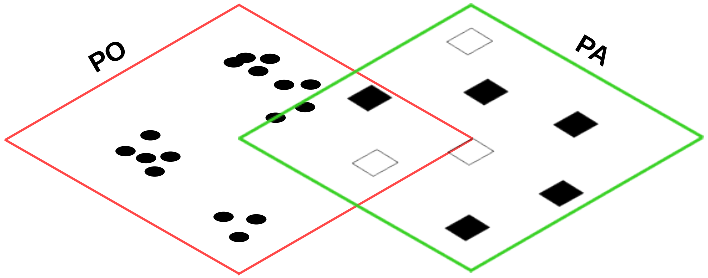


- $\lambda(s) = \exp\left(\beta_0 + \beta_1 x(s) + \beta_{bias} z(s) + W(s)\right)$
- $cloglog(p_i) = \beta_0 + \beta_1 x(s) + W(s)$

- Do **PO** and **PA** data contribute equally to intensity estimation?
- Does the relative power of each source shift with sample sizes?
- Can targeted sampling of **PA** data enhance joint model performance even with few quadrats?

Simulation study of joint SDMs

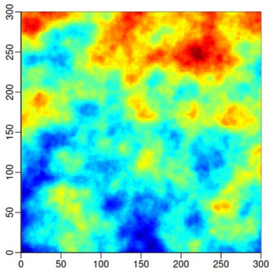
Joint modelling



- $\lambda(s) = \exp\left(\beta_0 + \beta_1 x(s) + \beta_{bias} z(s) + W(s)\right)$
- $cloglog(p_i) = \beta_0 + \beta_1 x(s) + W(s)$

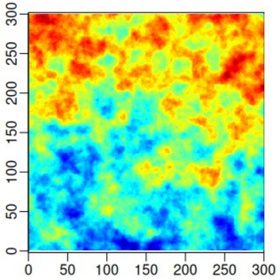
- Do **PO** and **PA** data contribute equally to intensity estimation?
- Does the relative power of each source shift with sample sizes?
- Can targeted sampling of **PA** data enhance joint model performance even with few quadrats?

Predictive performance



Intensity (λ)
Truth

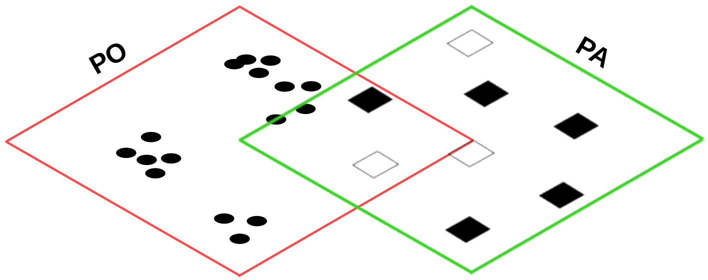
Pearson correlation



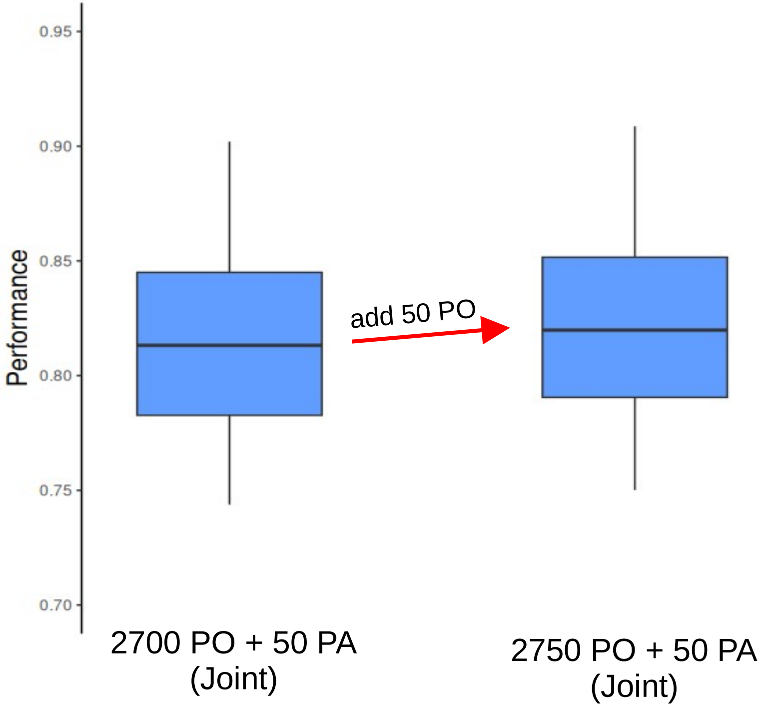
Model prediction

Simulation study of joint SDMs

Joint modelling

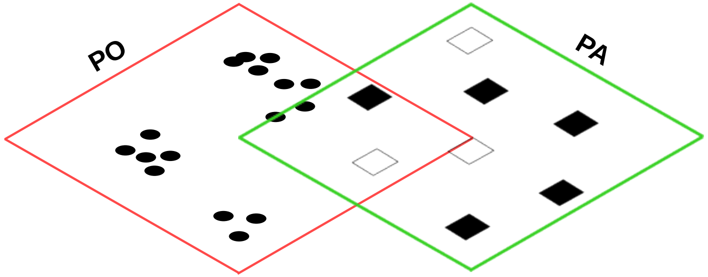


- $\lambda(s)=\exp\left(\beta_0+\beta_1 x(s)+\beta_{bias} z(s)+W(s)\right)$
- $cloglog(p_i)=\beta_0+\beta_1 x(s)+W(s)$

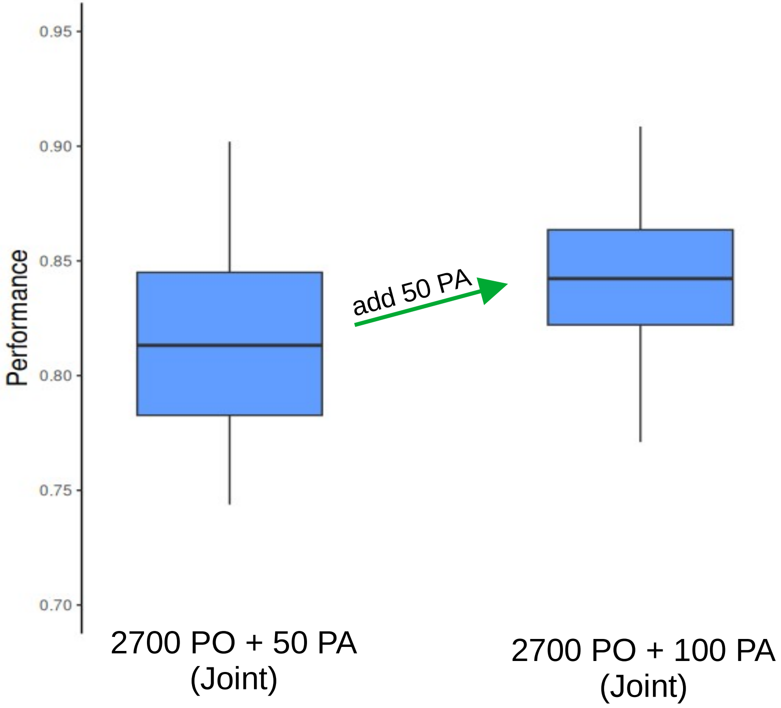


Simulation study of joint SDMs

Joint modelling

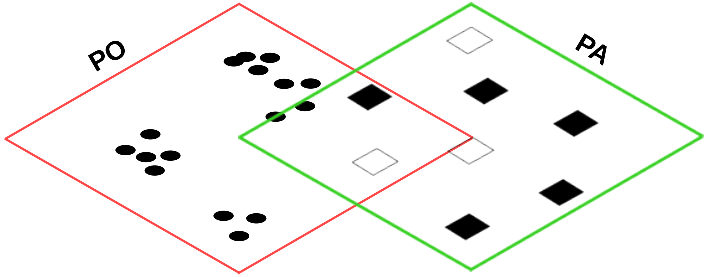


- $\lambda(s) = \exp\left(\beta_0 + \beta_1 x(s) + \beta_{bias} z(s) + W(s)\right)$
- $cloglog(p_i) = \beta_0 + \beta_1 x(s) + W(s)$



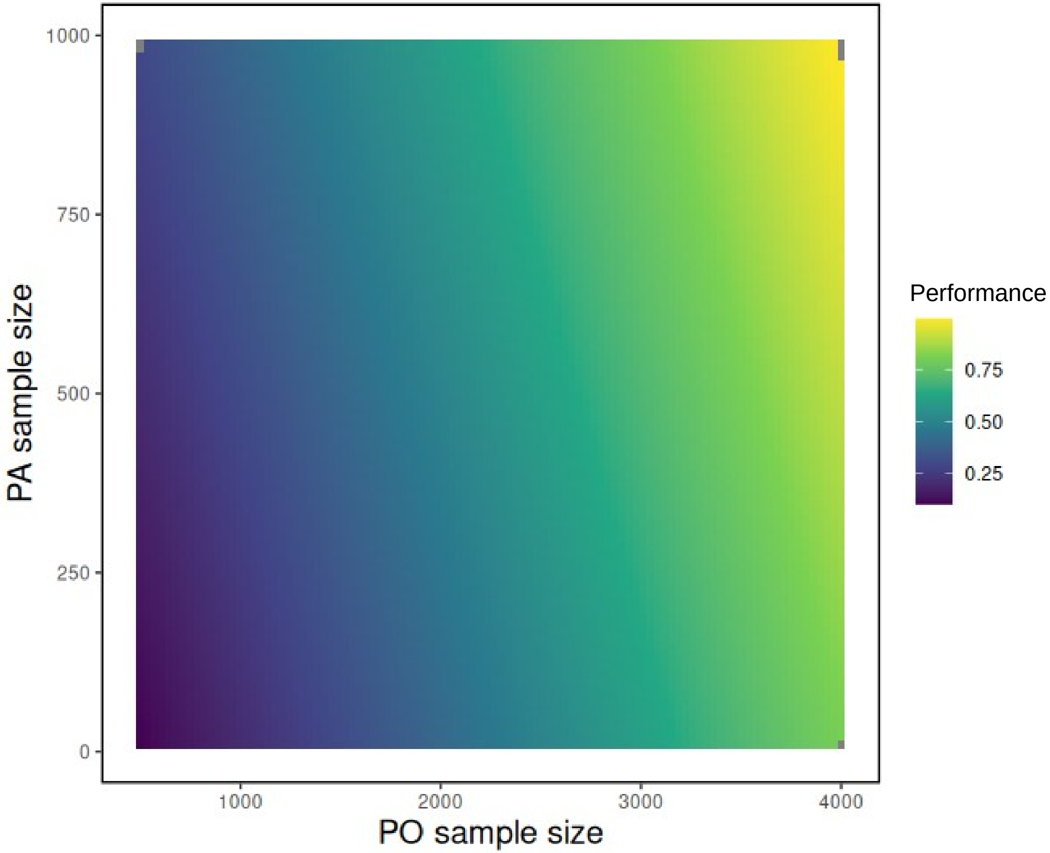
Simulation study of joint SDMs

Joint modelling



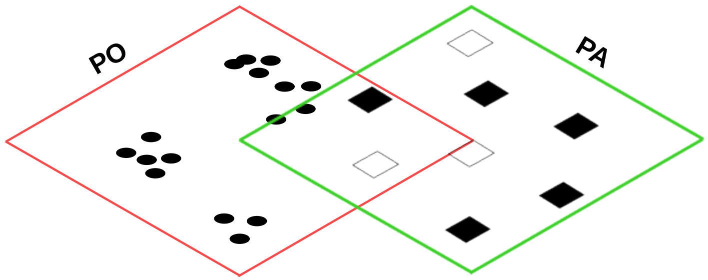
- $\lambda(s) = \exp\left(\beta_0 + \beta_1 x(s) + \beta_{bias} z(s) + W(s)\right)$
- $cloglog(p_i) = \beta_0 + \beta_1 x(s) + W(s)$

Example heatmap of performance



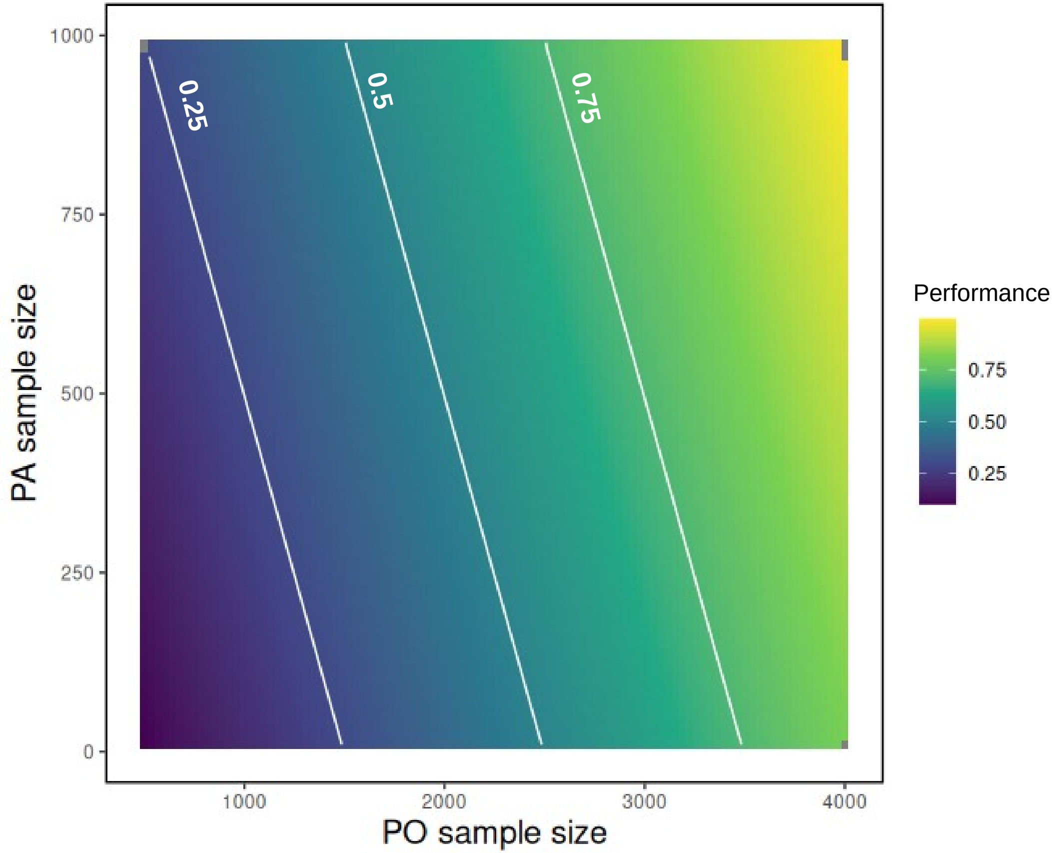
Simulation study of joint SDMs

Joint modelling



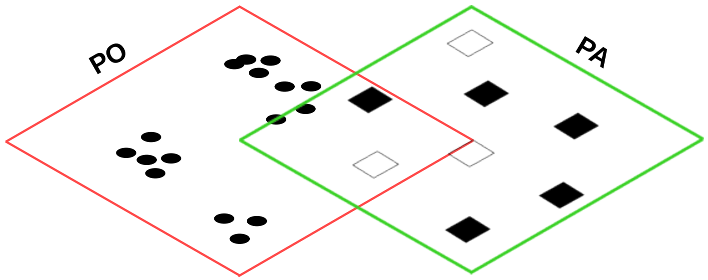
- $\lambda(s) = \exp(\beta_0 + \beta_1 x(s) + \beta_{bias} z(s) + W(s))$
- $cloglog(p_i) = \beta_0 + \beta_1 x(s) + W(s)$

Example heatmap of performance



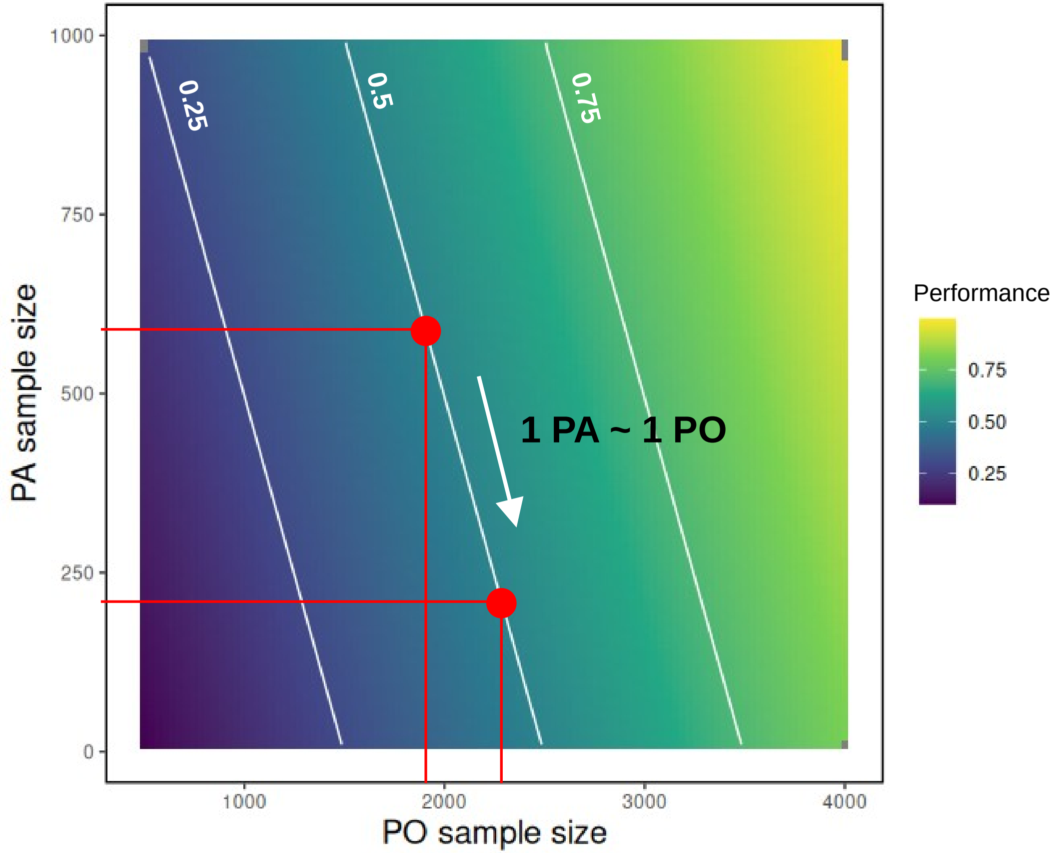
Simulation study of joint SDMs

Joint modelling



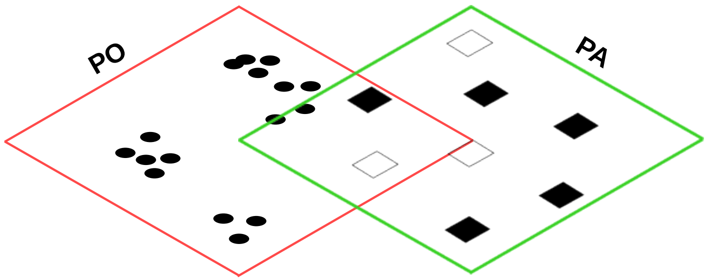
- $\lambda(s) = \exp\left(\beta_0 + \beta_1 x(s) + \beta_{bias} z(s) + W(s)\right)$
- $cloglog(p_i) = \beta_0 + \beta_1 x(s) + W(s)$

Example heatmap of performance



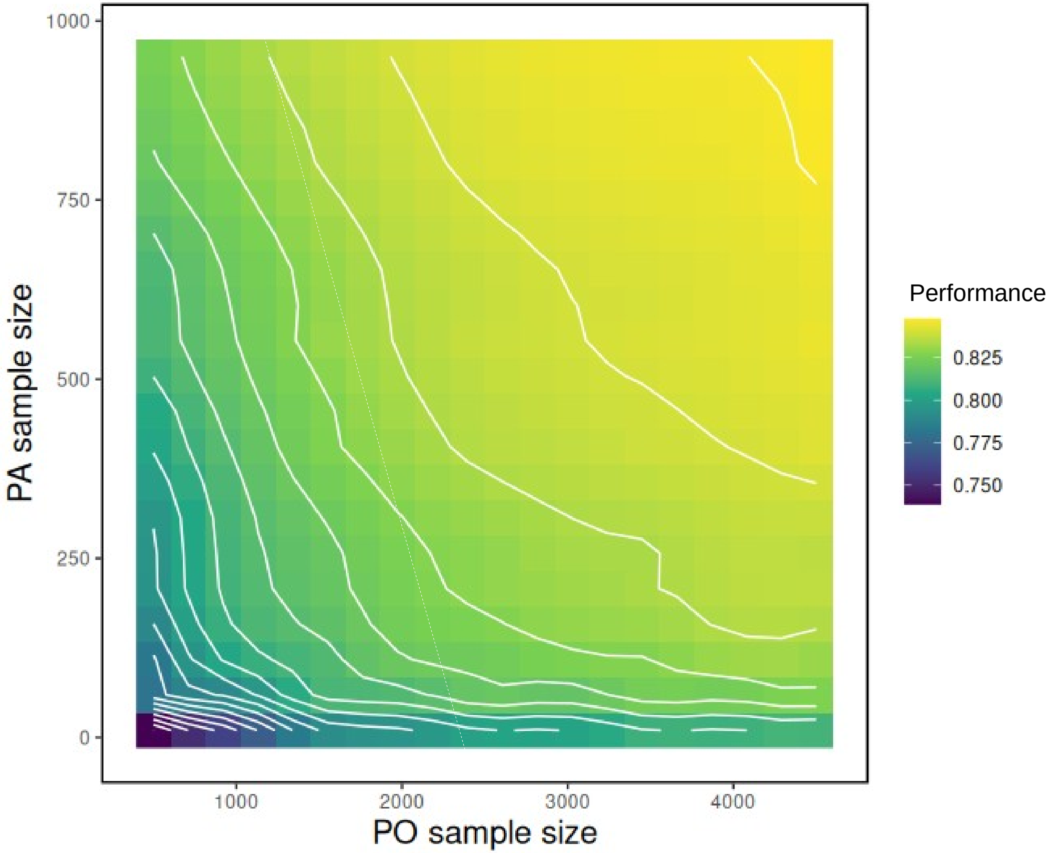
Simulation study of joint SDMs

Joint modelling



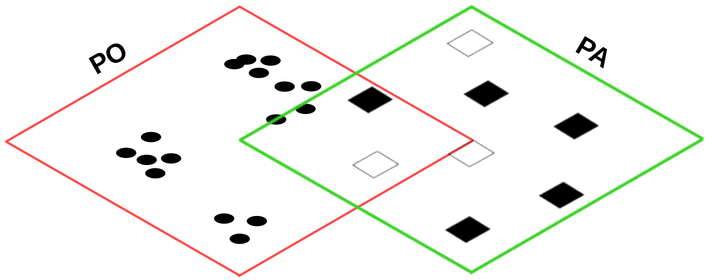
- $\lambda(s) = \exp\left(\beta_0 + \beta_1 x(s) + \beta_{bias} z(s) + W(s)\right)$
- $cloglog(p_i) = \beta_0 + \beta_1 x(s) + W(s)$

Joint model with bias covariate



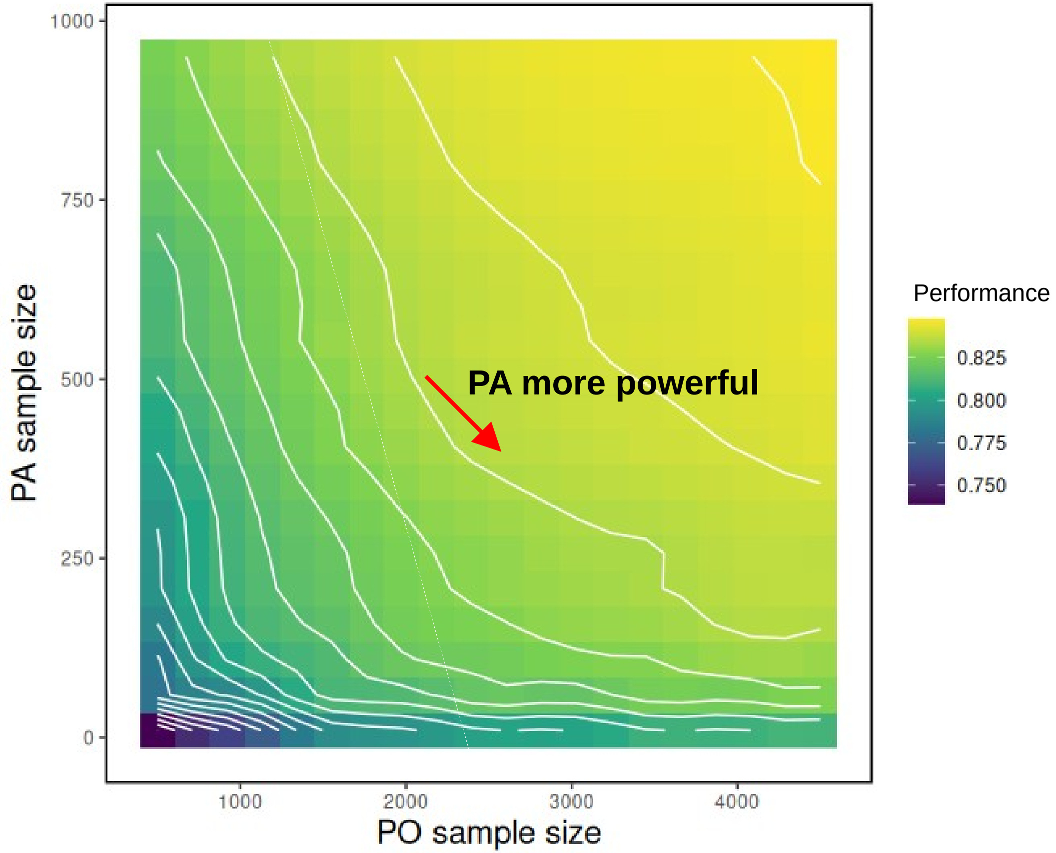
Simulation study of joint SDMs

Joint modelling



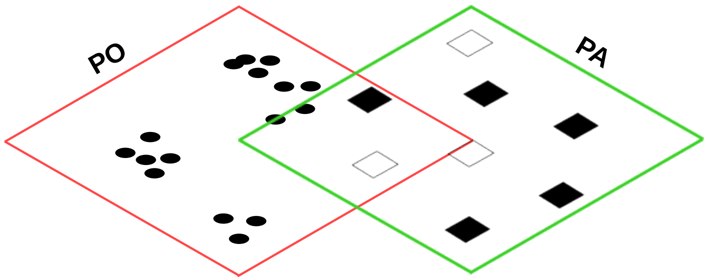
- $\lambda(s) = \exp\left(\beta_0 + \beta_1 x(s) + \beta_{bias} z(s) + W(s)\right)$
- $cloglog(p_i) = \beta_0 + \beta_1 x(s) + W(s)$

Joint model with bias covariate



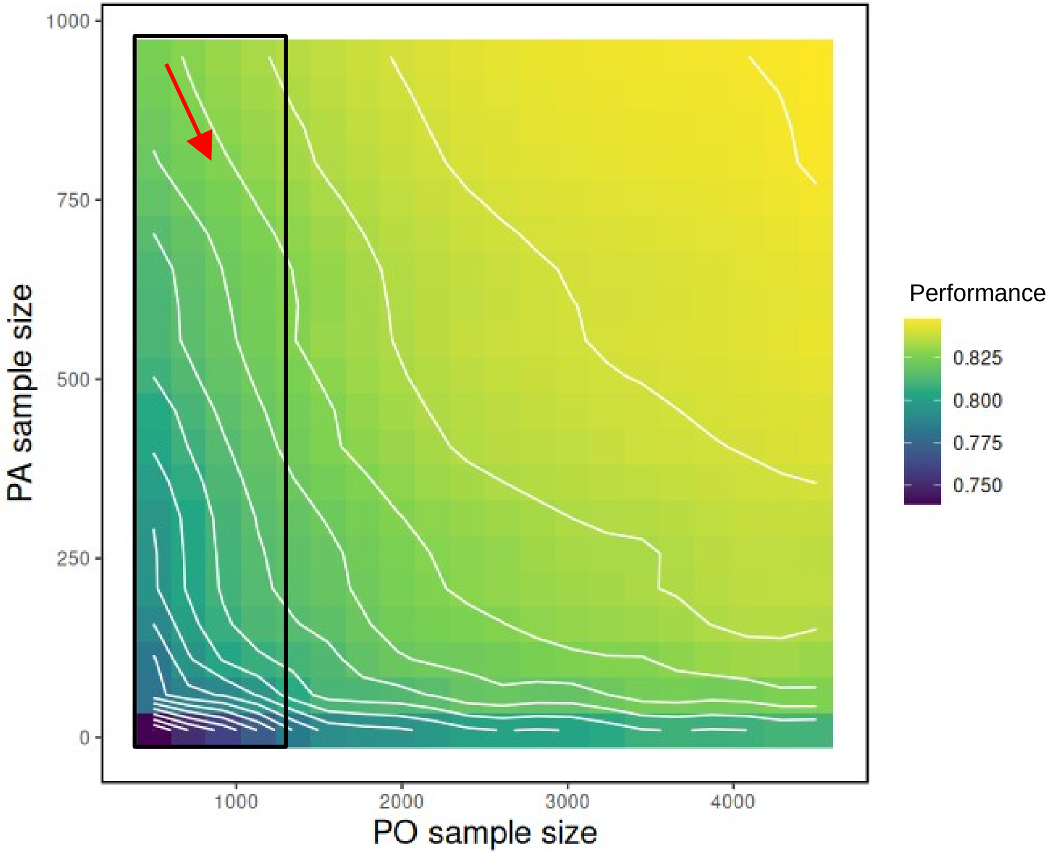
Simulation study of joint SDMs

Joint modelling



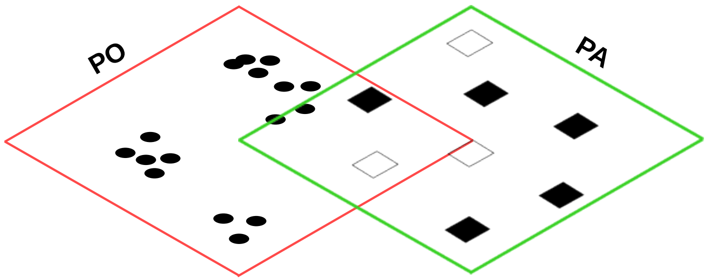
- $\lambda(s) = \exp\left(\beta_0 + \beta_1 x(s) + \beta_{bias} z(s) + W(s)\right)$
- $cloglog(p_i) = \beta_0 + \beta_1 x(s) + W(s)$

Joint model with bias covariate



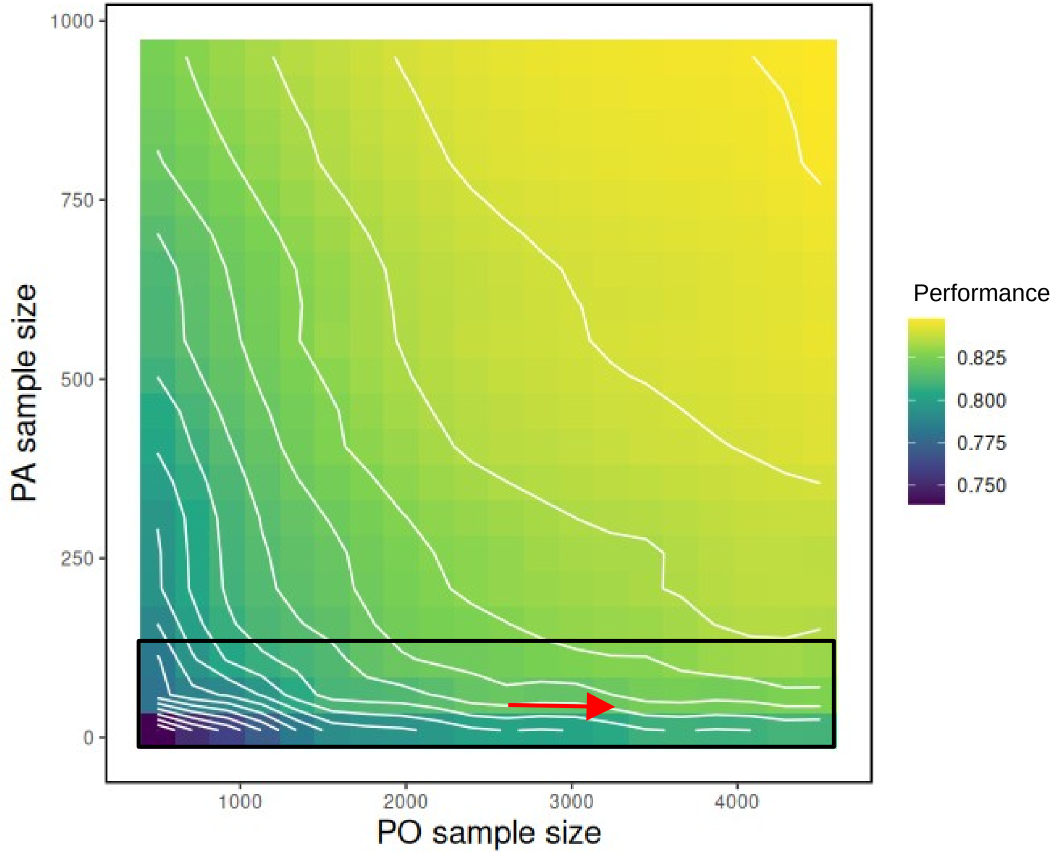
Simulation study of joint SDMs

Joint modelling



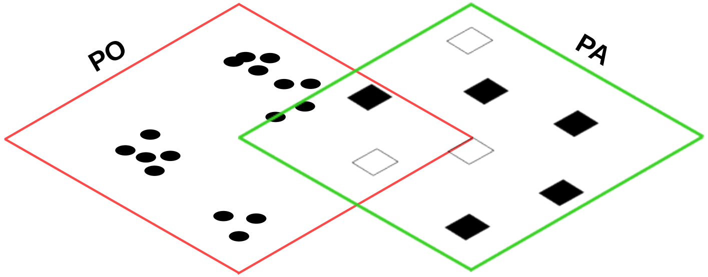
- $\lambda(s) = \exp\left(\beta_0 + \beta_1 x(s) + \beta_{bias} z(s) + W(s)\right)$
- $cloglog(p_i) = \beta_0 + \beta_1 x(s) + W(s)$

Joint model with bias covariate



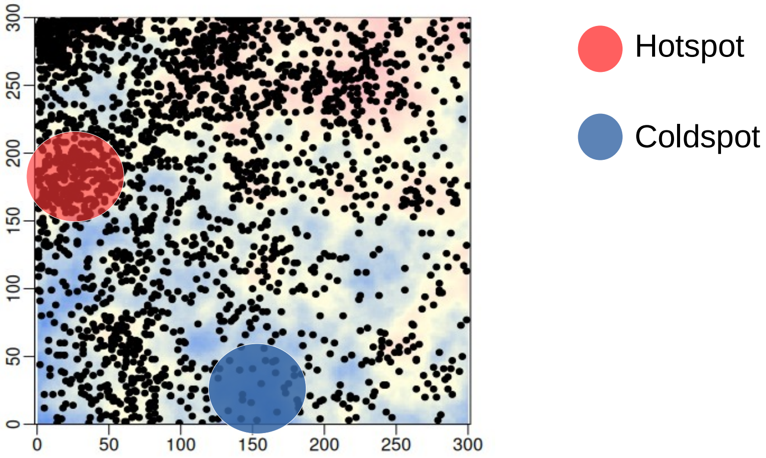
Simulation study of joint SDMs

Joint modelling



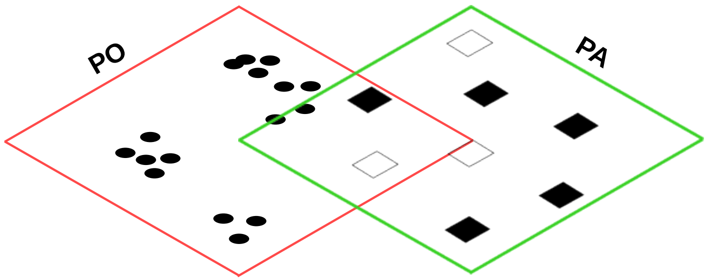
- $\lambda(s) = \exp\left(\beta_0 + \beta_1 x(s) + \beta_{bias} z(s) + W(s)\right)$
- $cloglog(p_i) = \beta_0 + \beta_1 x(s) + W(s)$

Presence-Only data



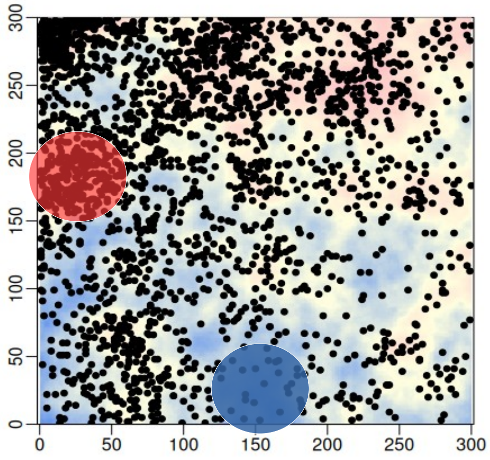
Simulation study of joint SDMs

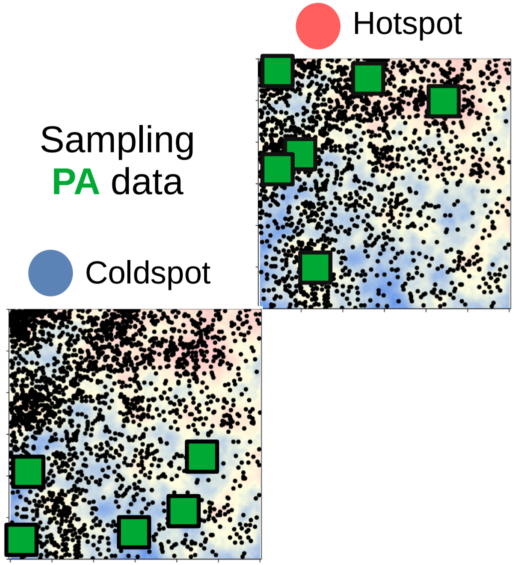
Joint modelling

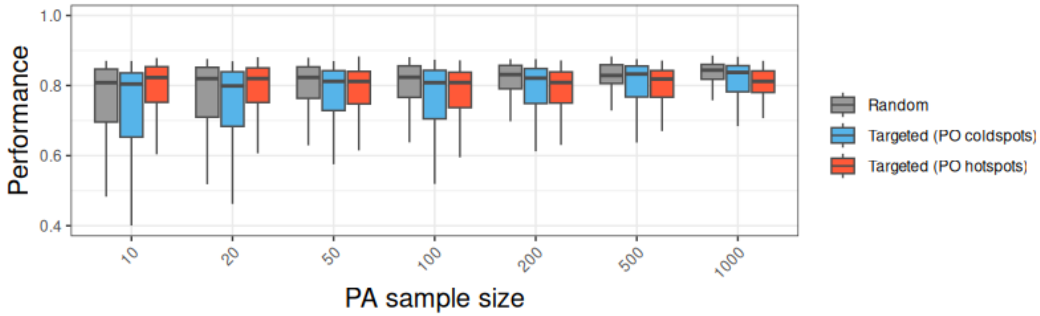


- $\lambda(s) = \exp\left(\beta_0 + \beta_1 x(s) + \beta_{bias} z(s) + W(s)\right)$
- $cloglog(p_i) = \beta_0 + \beta_1 x(s) + W(s)$

Presence-Only data

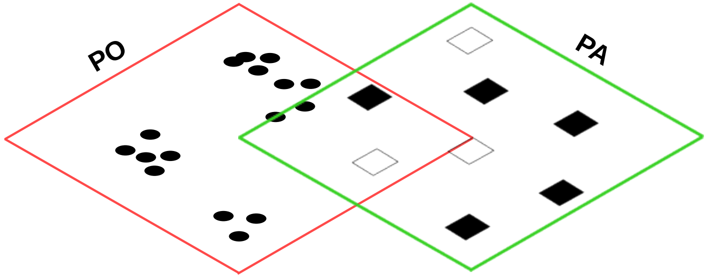






Simulation study of joint SDMs

Joint modelling



- $\lambda(s) = \exp\left(\beta_0 + \beta_1 x(s) + \beta_{bias} z(s) + W(s)\right)$
- $cloglog(p_i) = \beta_0 + \beta_1 x(s) + W(s)$

✗ Do **PO** and **PA** data contribute equally to intensity estimation?

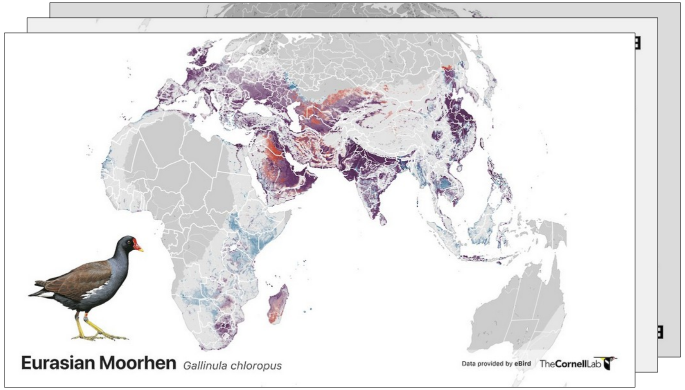
✓ Does the relative power of each source shift with sample sizes?

> 1 PA quadrat is roughly as useful as 20 PO points.
> This relative power shifts especially when one source of data is scarce or saturated.

✗ Can targeted sampling of **PA** data enhance joint model performance even with few quadrats?

> Presence-Only data adds spatial coverage (W), Presence/Absence data anchors covariate slopes
>
> Presence/Absence is most useful when spread across the **full range of environmental conditions**

Joint models offer a promising path for leveraging large-scale, biased opportunistic datasets.



Opportunistic records with the eBird project

→ Joints models where sampling bias cannot be (fully) captured by covariates.

→ Joints models with more than 2 sources

→ Account for imperfect species detection in standardized data

Targeted sampling

→ Other metrics to define where to sample Presence/Absence data?

→ Could Presence-Only data be target sampled instead?

Using citizen-science: real-time, adaptive sampling by defining high-priority areas for project participants

Thank you!

Bénard A,
Lasgorceux F,
Papaïx J,
Opitz T,
Bunz Y,
Combrisson D

Contact: annaelle.benard@inrae.fr

Statistiques pour les Sciences Participatives

Risques, Extrêmes et Statistique Spatio-TEmporelle

Seminar
"Approaches to Heterogeneous Data Modeling"

2 & 3 Feb. 2026 | Avignon, France

Abstract submission open
sciencesconf.org : cisstats2026