

Infering species distribution from aggregated spatial data

Going from coarse landings data to fine scale species distribution

Baptiste Alglave, Marie-Pierre Etienne, Kasper Kristensen, Youen Vermard, Mathieu Woillez, Etienne Rivot

04/2022 - GdR Ecostat



Spatial data in ecology

Survey data



Standardized sampling plan
High quality data



Small sample size



Examples

EVHOE data, Bay of Biscay (marine ecology)

Citizen science data

Access to more data
Exact locations available

Opportunistic (or even preferential) sampling



Ebird application (ornithology)

Declaration data

Mandatory declaration
Massive data

Aggregated at the scale of administrative units



Harvest data, Wisconsin (hunting)

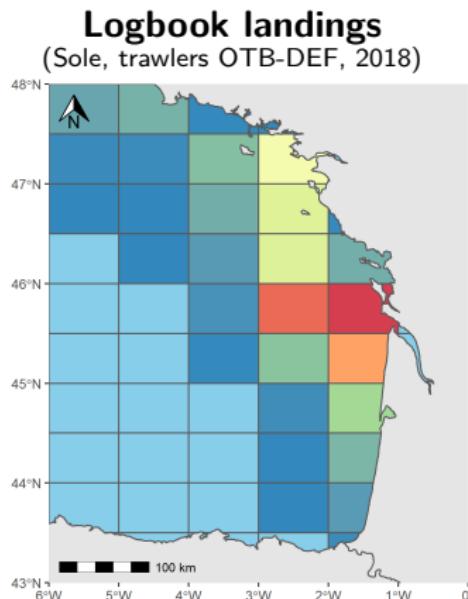
How to integrate all these datasources? (especially when they do not have the same spatial resolution)

- ➡ Some specific application in the context of fishery science:
Alglave Baptiste, Rivot Etienne, Etienne Marie-Pierre, Woillez Mathieu, Thorson James T, Vermaud Youen (2022). Combining scientific survey and commercial catch data to map fish distribution. ICES Journal of Marine Science IN PRESS. <https://doi.org/10.1093/icesjms/fsac032>

How to integrate all these datasources? (especially when they do not have the same spatial resolution)

- ➡ Some specific application in the context of fishery science:
Alglave Baptiste, Rivot Etienne, Etienne Marie-Pierre, Woillez Mathieu, Thorson James T, Vermaud Youen (2022). Combining scientific survey and commercial catch data to map fish distribution. ICES Journal of Marine Science IN PRESS. <https://doi.org/10.1093/icesjms/fsac032>

Commercial catch declarations data in fishery science



Spatial resolution

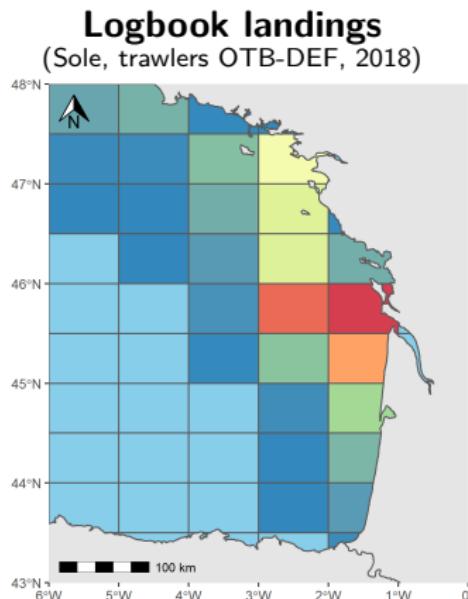
Catch are daily declared at the resolution of ICES rectangles



VMS pings are vessels GPS locations emitted each hour

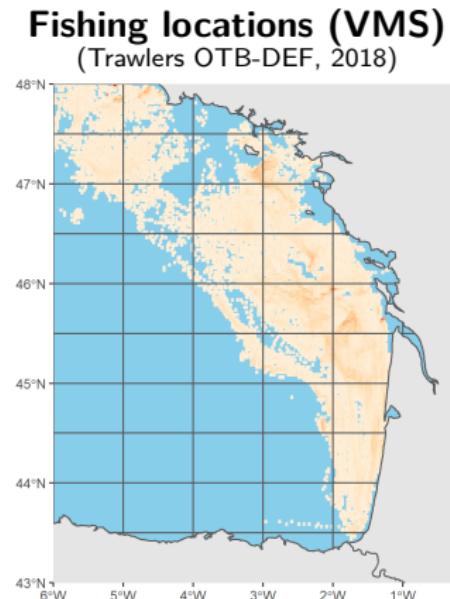
→ Refine landings spatial resolution

Commercial catch declarations data in fishery science



Spatial resolution

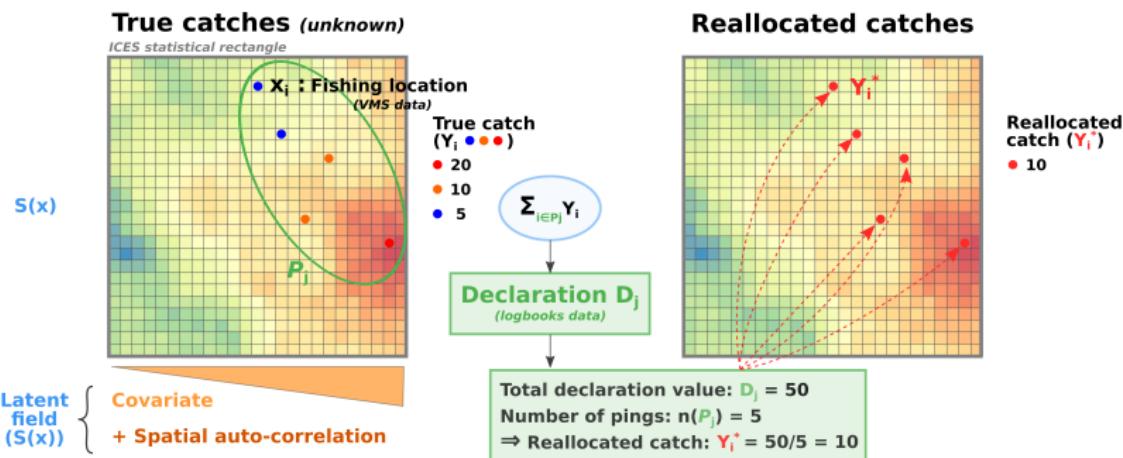
Catch are daily declared at the resolution of ICES rectangles



VMS pings are vessels GPS locations emitted each hour

➡ Refine landings spatial resolution

Two alternative procedures to reallocate catches



Actual situation

$$Y_i | S(x_i), x_i \sim \mathcal{L}_Y(S(x_i), \xi, \sigma^2)$$

$$Y_i = \frac{D_j}{n(P_j)} = Y_i^*$$

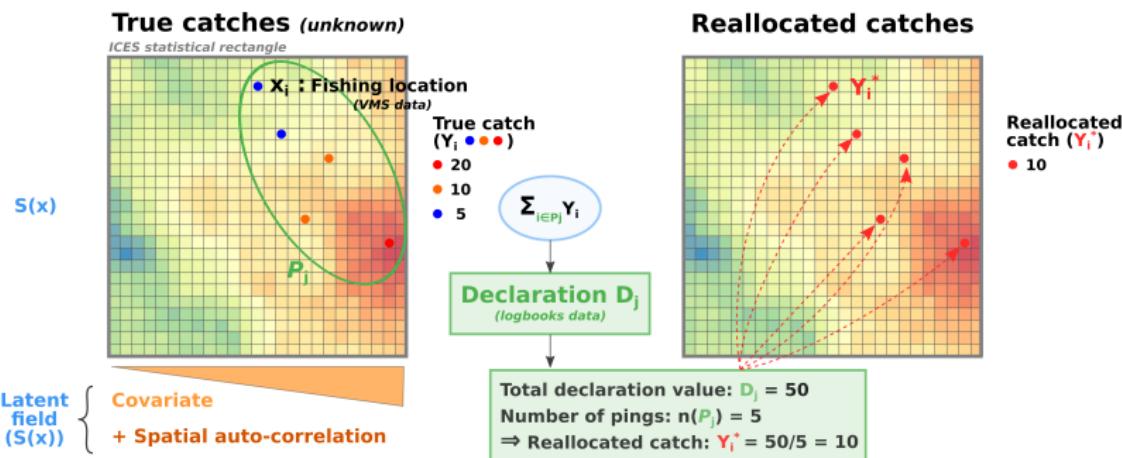
Alternative solution

$$D_j = \sum_{i \in P_j} Y_i$$

$$D_j | S_{P_j}, P_j \sim \mathcal{L}_D(S_{P_j}, \xi, \sigma^2)$$

Match \mathcal{L}_D and \mathcal{L}_Y moments (appendix)

Two alternative procedures to reallocate catches



Actual situation $Y_i | S(x_i), x_i \sim \mathcal{L}_Y(S(x_i), \xi, \sigma^2)$

$$Y_i = \frac{D_j}{n(P_j)} = Y_i^*$$

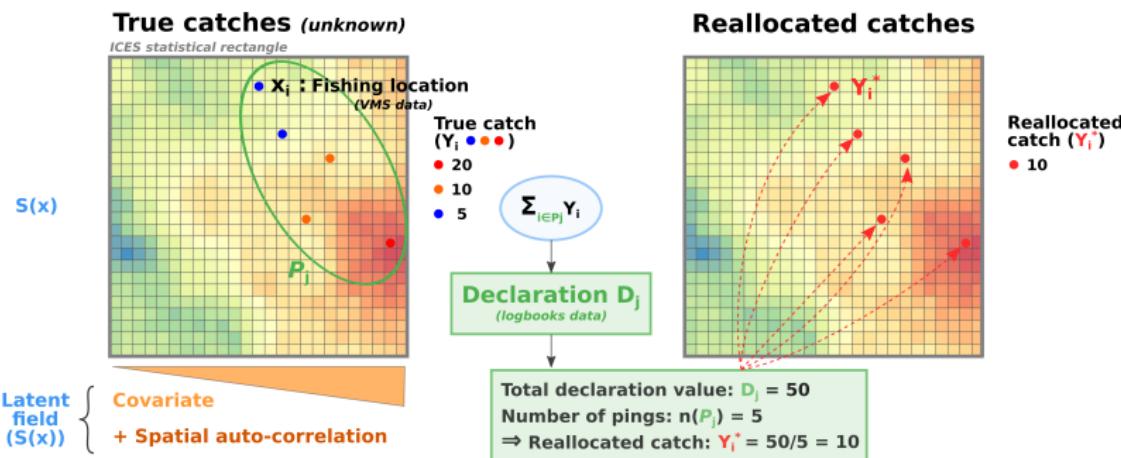
Alternative solution

$$D_j = \sum_{i \in P_j} Y_i$$

$$D_j | S_{P_j}, P_j \sim \mathcal{L}_D(S_{P_j}, \xi, \sigma^2)$$

Match \mathcal{L}_D and \mathcal{L}_Y moments (appendix)

Two alternative procedures to reallocate catches



Actual situation $Y_i | S(x_i), x_i \sim \mathcal{L}_Y(S(x_i), \xi, \sigma^2)$

$$Y_i = \frac{D_j}{n(P_j)} = Y_i^*$$

Alternative solution

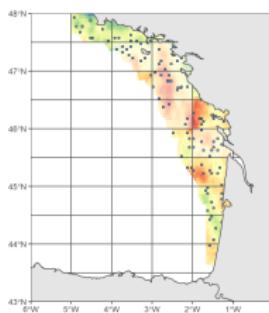
$$D_j = \sum_{i \in P_j} Y_i$$

$$D_j | S_{P_j}, P_j \sim \mathcal{L}_D(S_{P_j}, \xi, \sigma^2)$$

⇒ Match \mathcal{L}_D and \mathcal{L}_Y moments (appendix)

Simulation-estimation and case study

Simulation-estimation



Simulation

- **Latent field** (covariate + spatial random effect)
- **Commercial data** (3000 samples over 2/3 of the area)
- **Reallocation process** (10 locations per declaration)
- **Scientific data** (100 samples over the whole the area)

Estimation

Comparison of 3 model configurations:

- 1/ Model fitted to scientific data only
- 2/ Integrated model (= scientific + commercial data) with commercial likelihood built on Y_i^*
- 3/ Integrated model with commercial likelihood built on D_j

Estimation realized through TMB (Template Model Builder)
100 runs of simulation-estimation

Model evaluation

1/ Mean square prediction error

$$MSPE = \frac{\sum_{x=1}^n (S(x) - \hat{S}(x))^2}{n}$$

2/ Covariate effect (or species-habitat relationship):

$$\beta s = 2 \text{ versus } \hat{\beta}s$$

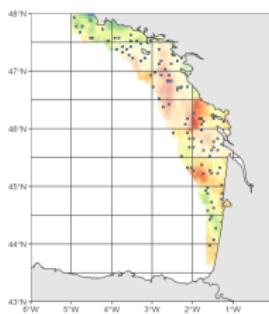
Case study: Sole in the Bay of Biscay



Survey data: Orhago
Commercial data: OTB-DEF trawlers (to ease convergence onboard observer data were integrated in the fit)
Fitted models: same as simulations
Covariate: substrate

Simulation-estimation and case study

Simulation-estimation



Simulation

- **Latent field** (covariate + spatial random effect)
- **Commercial data** (3000 samples over 2/3 of the area)
- **Reallocation process** (10 locations per declaration)
- **Scientific data** (100 samples over the whole the area)

Estimation

Comparison of 3 model configurations:

- 1/ Model fitted to **scientific data** only
- 2/ Integrated model (= **scientific** + **commercial data**) with **commercial likelihood built on Y_i^***
- 3/ Integrated model with **commercial likelihood built on D_j**

Estimation realized through TMB (Template Model Builder)
100 runs of simulation-estimation

Model evaluation

1/ Mean square prediction error

$$MSPE = \frac{\sum_{x=1}^n (S(x) - \hat{S}(x))^2}{n}$$

2/ Covariate effect (or species-habitat relationship):

$$\beta s = 2 \text{ versus } \hat{\beta}s$$

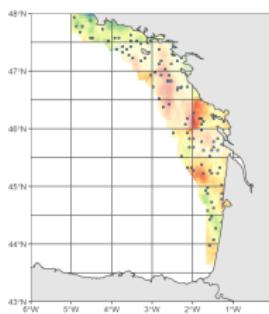
Case study: Sole in the Bay of Biscay



Survey data: Orhago
Commercial data: OTB-DEF trawlers (to ease convergence onboard observer data were integrated in the fit)
Fitted models: same as simulations
Covariate: substrate

Simulation-estimation and case study

Simulation-estimation



Simulation

- **Latent field** (covariate + spatial random effect)
- **Commercial data** (3000 samples over 2/3 of the area)
- **Reallocation process** (10 locations per declaration)
- **Scientific data** (100 samples over the whole the area)

Model evaluation

1/ Mean square prediction error

$$MSPE = \frac{\sum_{x=1}^n (S(x) - \hat{S}(x))^2}{n}$$

2/ Covariate effect (or species-habitat relationship):

$$\beta_s = 2 \text{ versus } \hat{\beta}_s$$

Estimation

Comparison of 3 model configurations:

- 1/ Model fitted to **scientific data** only
- 2/ Integrated model (= **scientific** + **commercial data**) with **commercial likelihood built on Y_i^***
- 3/ Integrated model with **commercial likelihood built on D_j**

Estimation realized through TMB (Template Model Builder)
100 runs of simulation-estimation

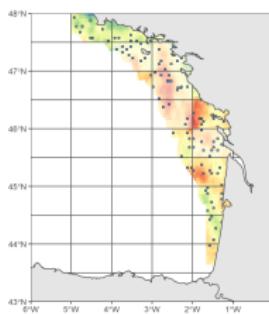
Case study: Sole in the Bay of Biscay



Survey data: Orhago
Commercial data: OTB-DEF trawlers (to ease convergence onboard observer data were integrated in the fit)
Fitted models: same as simulations
Covariate: substrate

Simulation-estimation and case study

Simulation-estimation



Simulation

- **Latent field** (covariate + spatial random effect)
- **Commercial data** (3000 samples over 2/3 of the area)
- **Reallocation process** (10 locations per declaration)
- **Scientific data** (100 samples over the whole the area)

Estimation

Comparison of 3 model configurations:

- 1/ Model fitted to **scientific data** only
- 2/ Integrated model (= **scientific** + **commercial data**) with **commercial likelihood built on Y_i^***
- 3/ Integrated model with **commercial likelihood built on D_j**

Estimation realized through TMB (Template Model Builder)
100 runs of simulation-estimation

Model evaluation

1/ Mean square prediction error

$$MSPE = \frac{\sum_{x=1}^n (S(x) - \hat{S}(x))^2}{n}$$

2/ Covariate effect (or species-habitat relationship):

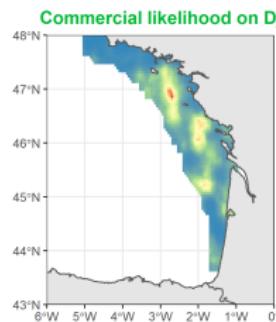
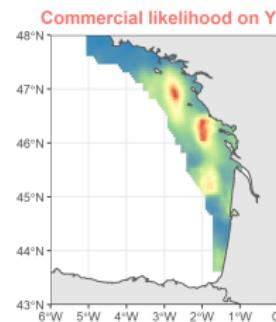
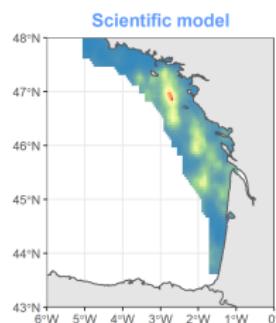
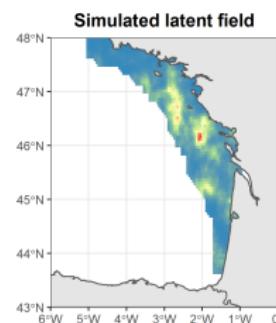
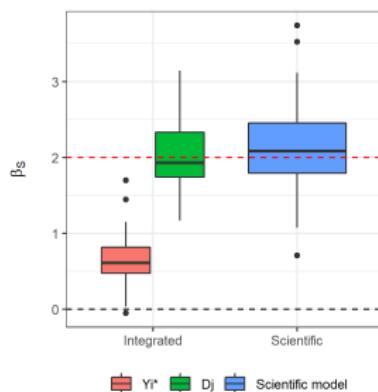
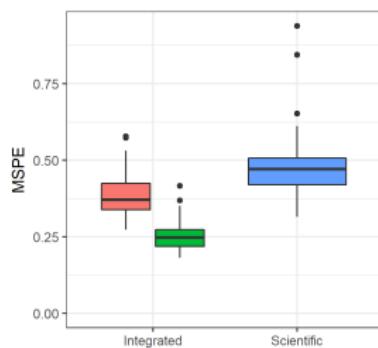
$$\beta s = 2 \text{ versus } \hat{\beta}s$$

Case study: Sole in the Bay of Biscay



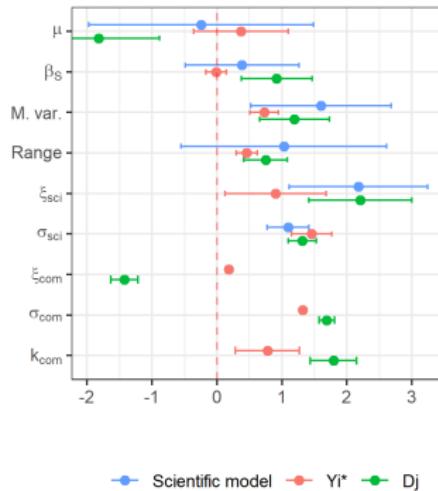
Survey data: Orhago
Commercial data: OTB-DEF trawlers (to ease convergence onboard observer data were integrated in the fit)
Fitted models: same as simulations
Covariate: substrate

Simulation-estimation

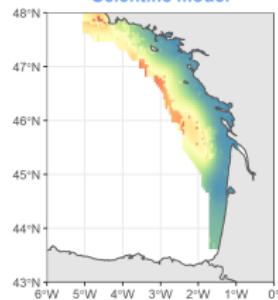


Case study: Sole in the Bay of Biscay

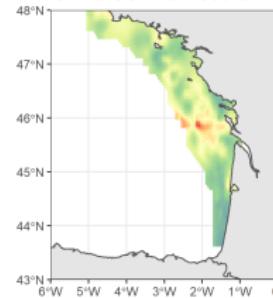
Parameters estimates



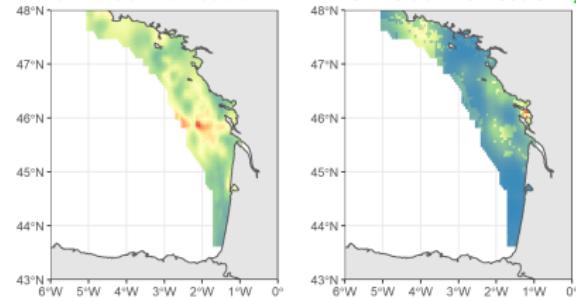
Scientific model



Commercial likelihood on Yi^*



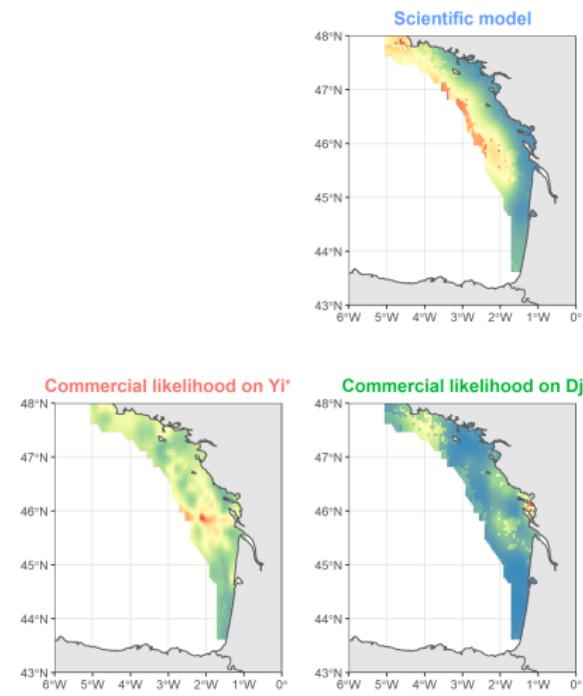
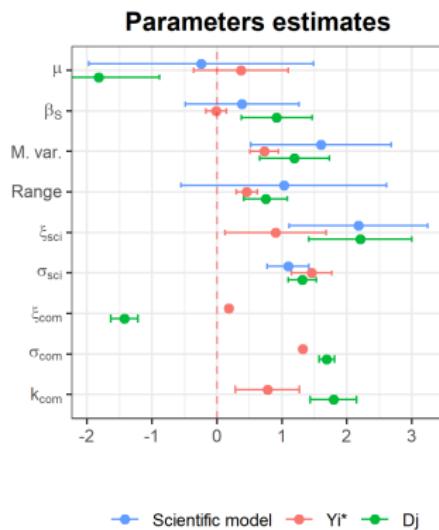
Commercial likelihood on Dj



The integrated model fitted to Dj:

- Recover the species-habitat relationship (β_s)
- Modifies the contrasts of the map (shape and intensity of the hotspots/coldspots)

Case study: Sole in the Bay of Biscay



The integrated model fitted to D_j :
➡ Recovers the species-habitat relationship (β_s)
➡ Modifies the contrasts of the map
(shape and intensity of the hotspots/coldspots)

Discussion

- Integrated framework that combines **catch declarations data** (rough resolution) and **scientific data** (exact locations)
 - ➡ Allows to estimate the **habitat effect** through commercial data
 - ➡ Modifies the **contrasts of the map** (hotspots vs. coldspots)
- Some limits:
 - ➡ How to ease convergence ?
 - ➡ Need to make the hypothesis that fishing locations (P_j) are known
- Is it a generic framework ?
 - ➡ The overall approach is,
(i.e. modelling observed aggregated observations as a sum of latent punctual observations)
 - ➡ But need to adapt the observation model to the data
(here zeroinflated positive continuous data)
- Change of support applications
 - ➡ Increasing amount of available data that may face such issues
 - ➡ Here is one proposition to integrate aggregated data in SDM

Discussion

- Integrated framework that combines **catch declarations data** (rough resolution) and **scientific data** (exact locations)
 - ➡ Allows to estimate the **habitat effect** through commercial data
 - ➡ Modifies the **contrasts of the map** (hotspots vs. coldspots)
- Some limits:
 - ➡ How to ease convergence ?
 - ➡ Need to make the hypothesis that fishing locations (P_j) are known
- Is it a generic framework ?
 - ➡ The overall approach is,
(i.e. modelling observed aggregated observations as a sum of latent punctual observations)
 - ➡ But need to adapt the observation model to the data
(here zeroinflated positive continuous data)
- Change of support applications
 - ➡ Increasing amount of available data that may face such issues
 - ➡ Here is one proposition to integrate aggregated data in SDM

Discussion

- Integrated framework that combines **catch declarations data** (rough resolution) and **scientific data** (exact locations)
 - ➡ Allows to estimate the **habitat effect** through commercial data
 - ➡ Modifies the **contrasts of the map** (hotspots vs. coldspots)
- Some limits:
 - ➡ How to ease convergence ?
 - ➡ Need to make the hypothesis that fishing locations (P_j) are known
- Is it a generic framework ?
 - ➡ The overall approach is,
(i.e. modelling observed aggregated observations as a sum of latent punctual observations)
 - ➡ But need to adapt the observation model to the data
(here zeroinflated positive continuous data)
- Change of support applications
 - ➡ Increasing amount of available data that may face such issues
 - ➡ Here is one proposition to integrate aggregated data in SDM

Discussion

- Integrated framework that combines **catch declarations data** (rough resolution) and **scientific data** (exact locations)
 - ➡ Allows to estimate the **habitat effect** through commercial data
 - ➡ Modifies the **contrasts of the map** (hotspots vs. coldspots)
- **Some limits:**
 - ➡ How to ease convergence ?
 - ➡ Need to make the hypothesis that fishing locations (P_j) are known
- **Is it a generic framework ?**
 - ➡ The overall approach is,
(i.e. modelling observed aggregated observations as a sum of latent punctual observations)
 - ➡ But need to adapt the observation model to the data
(here zeroinflated positive continuous data)
- **Change of support applications**
 - ➡ Increasing amount of available data that may face such issues
 - ➡ Here is one proposition to integrate aggregated data in SDM

Thank you for your attention!



Punctual observation model (Y_i):

$L(y, \mu, \sigma^2)$ is the lognormal likelihood for observation y , mean μ and variance σ^2

Y and D are supposed conditional on S and x

$$P(Y_i = y_i) = \begin{cases} p_i & \text{if } y_i = 0 \\ (1 - p_i) \cdot L\left(y_i, \mu_i = \frac{S(x_i)}{(1-p_i)}, \sigma^2\right) & \text{if } y_i > 0 \end{cases}$$

$$p_i = \exp(-e^\xi \cdot S(x_i))$$

Declaration model ($D_j = \sum_{i \in \mathcal{P}_j} Y_i$):

$$P(D_j = 0) = \prod_{i \in \mathcal{P}_j} P(Y_i = 0) = \exp \left\{ - \sum_{i \in \mathcal{P}_j} e^\xi \cdot S(x_i) \right\} = \pi_j.$$

$$P(D_j = d_j | d_j > 0) = L\left(d_j, \mu_j = E(D_j | d_j > 0), \sigma_j^2 = \ln\left(\frac{\text{Var}(D_j | d_j > 0)}{E(D_j | d_j > 0)^2}\right) + 1\right)$$

$$E(D_j | d_j > 0) = \frac{\sum_{i \in \mathcal{P}_j} S(x_i)}{1 - \pi_j}$$

$$\text{Var}(D_j | d_j > 0) = \frac{\sum_{i \in \mathcal{P}_j} \text{Var}(Y_i)}{1 - \pi_j} - \frac{\pi_j}{(1 - \pi_j)^2} E(D_j)^2$$

$$\text{Var}(Y_i) = \frac{S(x_i)^2}{1 - p_i} (e^{\sigma^2} - (1 - p_i))$$