Downscaling coarse observations to predict continuous species spatio-temporal distribution

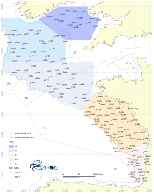
Going from coarse landings data to fine scale fish distribution

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

04/2022 - RESSTE

Ifremer L'INSTITUT agro Rennes Angers decod de la source à l'océan

Spatial data in ecology

	Survey data	Citizen science data	Declaration data
+	Standardized sampling plan High quality data	Access to more data Exact locations available	Mandatory declaration Massive data
−	Small sample size	Opportunistic (or even preferential) sampling	Aggregated at the scale of administrative units
Examples			
	EVHOE data, Bay of Biscay (marine ecology)	Ebird application (ornithology)	Harvest data, Wisconsin (hunting)

How to integrate all these datasources?

(especially when they do not have the same spatial resolution)

Change of support

Common issue in statistical literature

"Modifiable areal unit" problem (MAUP): aggregation of data over increasingly larger geographic scales (e.g. data collected at point level but regrouped/declared at coarse level)

Several fields of application: climatology, health science, ecology

But mainly **standard observational data (Poisson, Gaussian)**, while data may be more complex in ecological applications (e.g. zero-inflated lognormal data)

➠ **Objective of our work:** provide an approach that suits for complex data

➠ Base our model on an existing framework in the context of fishery science:

Alglave Baptiste, Rivot Etienne, Etienne Marie-Pierre, Woillez Mathieu, Thorson James T, Vermard 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)

Change of support

Common issue in statistical literature

"Modifiable areal unit" problem (MAUP): aggregation of data over increasingly larger geographic scales (e.g. data collected at point level but regrouped/declared at coarse level)

Several fields of application: climatology, health science, ecology

But mainly **standard observational data (Poisson, Gaussian)**, while data may be more complex in ecological applications (e.g. zero-inflated lognormal data)

➠ **Objective of our work:** provide an approach that suits for complex data

➠ Base our model on an existing framework in the context of fishery science:

Alglave Baptiste, Rivot Etienne, Etienne Marie-Pierre, Woillez Mathieu, Thorson James T, Vermard 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)

Change of support

Common issue in statistical literature

"Modifiable areal unit" problem (MAUP): aggregation of data over increasingly larger geographic scales (e.g. data collected at point level but regrouped/declared at coarse level)

Several fields of application: climatology, health science, ecology

But mainly **standard observational data (Poisson, Gaussian)**, while data may be more complex in ecological applications (e.g. zero-inflated lognormal data)

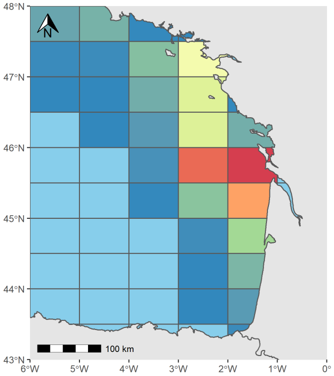
➠ **Objective of our work:** provide an approach that suits for complex data

➠ Base our model on an existing framework in the context of fishery science:

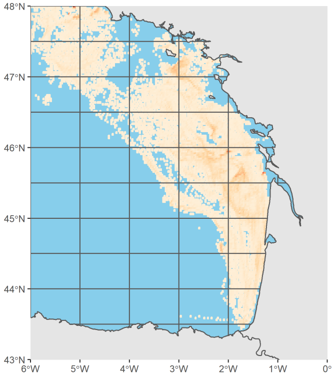
Alglave Baptiste, Rivot Etienne, Etienne Marie-Pierre, Woillez Mathieu, Thorson James T, Vermard Youen (2022). ***Combining scientific survey and commercial catch data to map fish distribution.*** *ICES Journal of Marine Science IN PRESS. https://doi.org/10.1093/icesjms/fsac032*

Commercial catch declarations data in fishery science

Logbook landings
(Sole, trawlers OTB-DEF, 2018)



Fishing locations (VMS)
(Trawlers OTB-DEF, 2018)



Spatial resolution

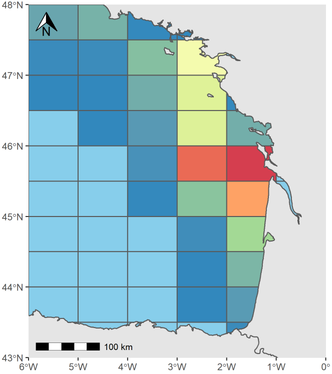
Catch are daily declared at the resolution of ICES rectangles

VMS pings are vessels GPS locations emitted each hour

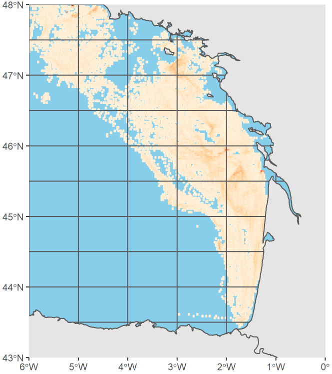
➠ Refine landings spatial resolution

Commercial catch declarations data in fishery science

Logbook landings
(Sole, trawlers OTB-DEF, 2018)



Fishing locations (VMS)
(Trawlers OTB-DEF, 2018)



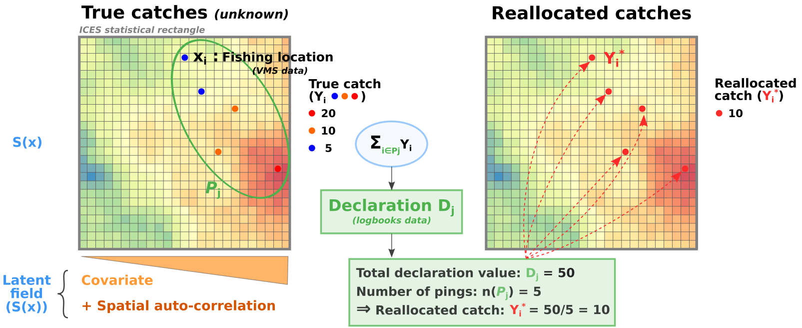
Spatial resolution

Catch are daily declared at the resolution of ICES rectangles

VMS pings are vessels GPS locations emitted each hour

➠ **Refine landings spatial resolution**

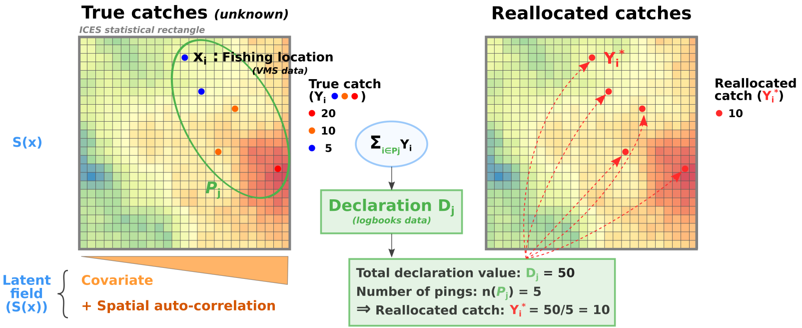
Two alternative procedures to reallocate catches



Current situation $Y_i | S(x_i), x_i \sim \mathcal{L}_Y(S(x_i), \xi, \sigma^2)$ $Y_i = \frac{D_j}{n(\mathcal{P}_j)} = Y_i^*$

Alternative solution $D_j = \sum_{i \in \mathcal{P}_j} Y_i$ $D_j | S_{\mathcal{P}_j}, \mathcal{P}_j \sim \mathcal{L}_D(S_{\mathcal{P}_j}, \xi, \sigma^2)$ ➠ Match $\mathcal{L}_D$ and $\mathcal{L}_Y$ moments

Two alternative procedures to reallocate catches



Current situation

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

$$Y_i = \frac{D_j}{\mathrm{n}(\mathcal{P}_j)} = Y_i^*$$

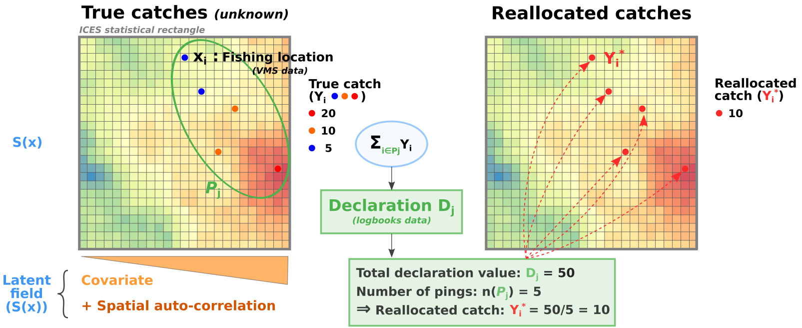
Alternative solution

$$D_j = \sum_{i \in \mathcal{P}_j} Y_i$$

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

➠ Match $\mathcal{L}_D$ and $\mathcal{L}_Y$ moments

Two alternative procedures to reallocate catches



Current situation $Y_i | S(x_i), x_i \sim \mathcal{L}_Y(S(x_i), \xi, \sigma^2)$ $Y_i = \frac{D_j}{\text{n}(\mathcal{P}_j)} = Y_i^*$

Alternative solution $D_j = \sum_{i\in\mathcal{P}_j} Y_i$ $D_j | S_{\mathcal{P}_j}, \mathcal{P}_j \sim \mathcal{L}_D(S_{\mathcal{P}_j}, \xi, \sigma^2)$ ➠ Match $\mathcal{L}_D$ and $\mathcal{L}_Y$ moments

Two alternative procedures to reallocate catches

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

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

$$p_i = \exp(-e^{\xi}.S(x_i))$$

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

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

$$\mathrm{P}\left(D_j = d_j | d_j > 0\right) = \quad ?$$

Two alternative procedures to reallocate catches

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

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

$$p_i = \exp(-e^{\xi}.S(x_i))$$

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

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

$$P(D_j = d_j | d_j > 0) = \quad ?$$

Specifying $\mathrm{P}\left(D_j = d_j | d_j > 0\right)$

Compute the moments of $\mathbf{D_j|d_j > 0}$

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

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

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

Consider $\mathbf{D_j|d_j > 0}$ is Lognormal too

$$\mathrm{P}\left(D_j = d_j | d_j > 0\right) =$$

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

Specifying $\mathrm{P}\left(D_j = d_j | d_j > 0\right)$

Compute the moments of $\mathbf{D_j | d_j > 0}$

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

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

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

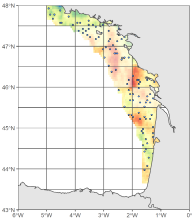
Consider $\mathbf{D_j | d_j > 0}$ is Lognormal too

$$\mathrm{P}\left(D_j = d_j | d_j > 0\right) =$$

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

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

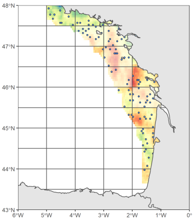
Estimation realized through TMB (Template Model Builder)
100 runs of simulation-estimation

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)

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 $\mathbf{Y_i^*}$

3/ **Integrated model** with commercial likelihood built on $\mathbf{D_j}$

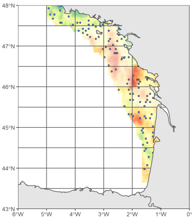
Estimation realized through TMB (Template Model Builder)
100 runs of simulation-estimation

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)

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 $\mathbf{Y}_{\mathbf{i}}^{*}$

3/ **Integrated model** with commercial likelihood built on $\mathbf{D}_{\mathbf{j}}$

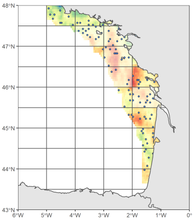
Estimation realized through TMB (Template Model Builder)
100 runs of simulation-estimation

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)

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 $\mathbf{Y}_i^*$

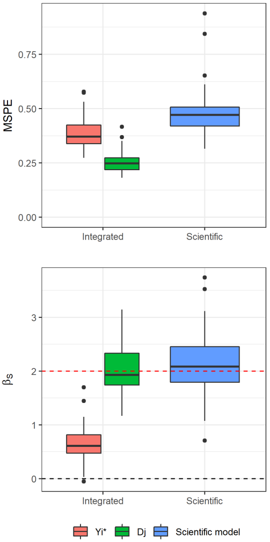
3/ Integrated model with commercial likelihood built on $\mathbf{D}_j$

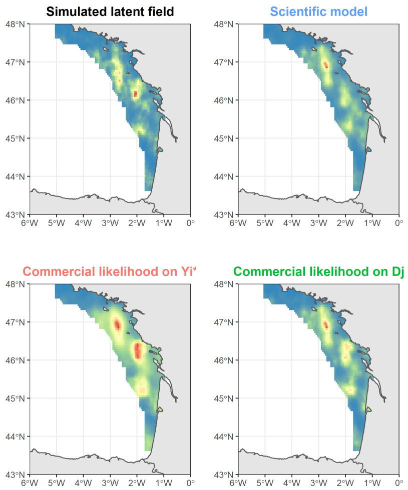
Estimation realized through TMB (Template Model Builder)
100 runs of simulation-estimation

Case study: Sole in the Bay of Biscay

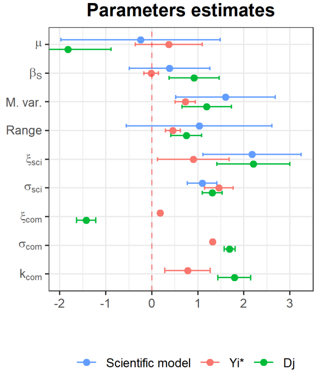
Survey data: Orhago
Commercial data: OTB-DEF trawlers (to ease convergence onboard observer data were integrated in the fit)
Fitted models: same as simulations
Covariate: substrate

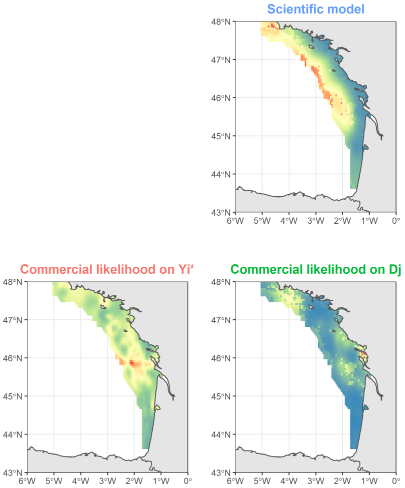
Simulation-estimation





Case study: Sole in the Bay of Biscay



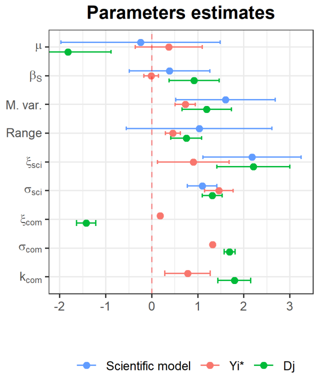


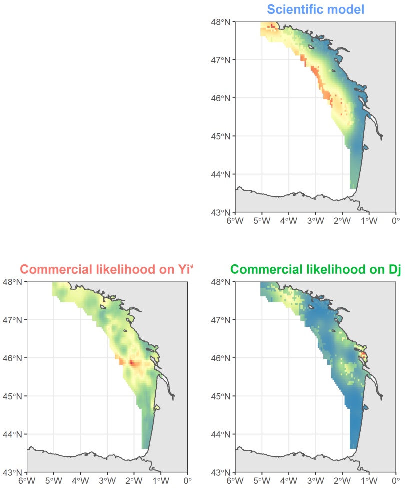
The **integrated model** fitted to D_j:

- Recovers the species-habitat relationship (β_S)
- Modifies the **contrasts of the map**

(shape and intensity of the hotspots/coldspots)

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 ($\mathcal{P}_j$) 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)

- **Moving to space-time ?**
  - ➠ Extending the observation model to account for **temporal misalignment**
  - ➠ Including spatio-**temporal covariance** in the latent field

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 ($\mathcal{P}_j$) 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)

- **Moving to space-time ?**
  ➠ Extending the observation model to account for **temporal misalignment**
  ➠ Including spatio-**temporal covariance** in the latent field

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 ($\mathcal{P}_j$) 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)
- **Moving to space-time ?**
  - ➠ Extending the observation model to account for **temporal misalignment**
  - ➠ Including spatio-**temporal covariance** in the latent field

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 ($\mathcal{P}_j$) 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)

- **Moving to space-time ?**
  ➠ Extending the observation model to account for **temporal misalignment**
  ➠ Including spatio-**temporal covariance** in the latent field

Thank you for your attention!