Integrating massive and heterogeneous spatio-temporal data in environmental science.

Marine ecology as field of application

Baptiste Alglave

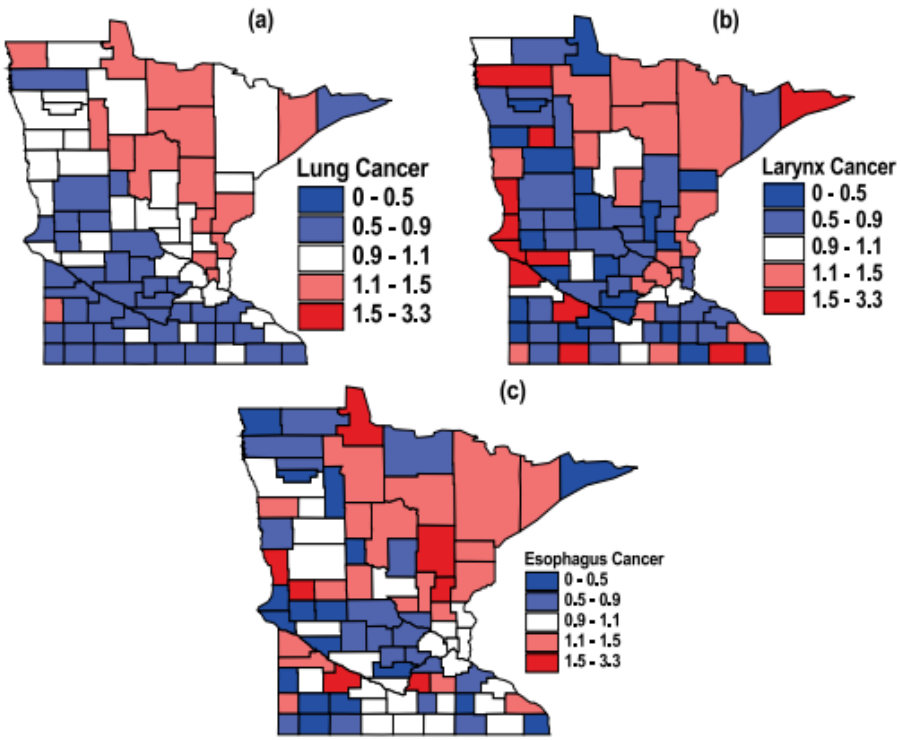
February 2024

Cursus

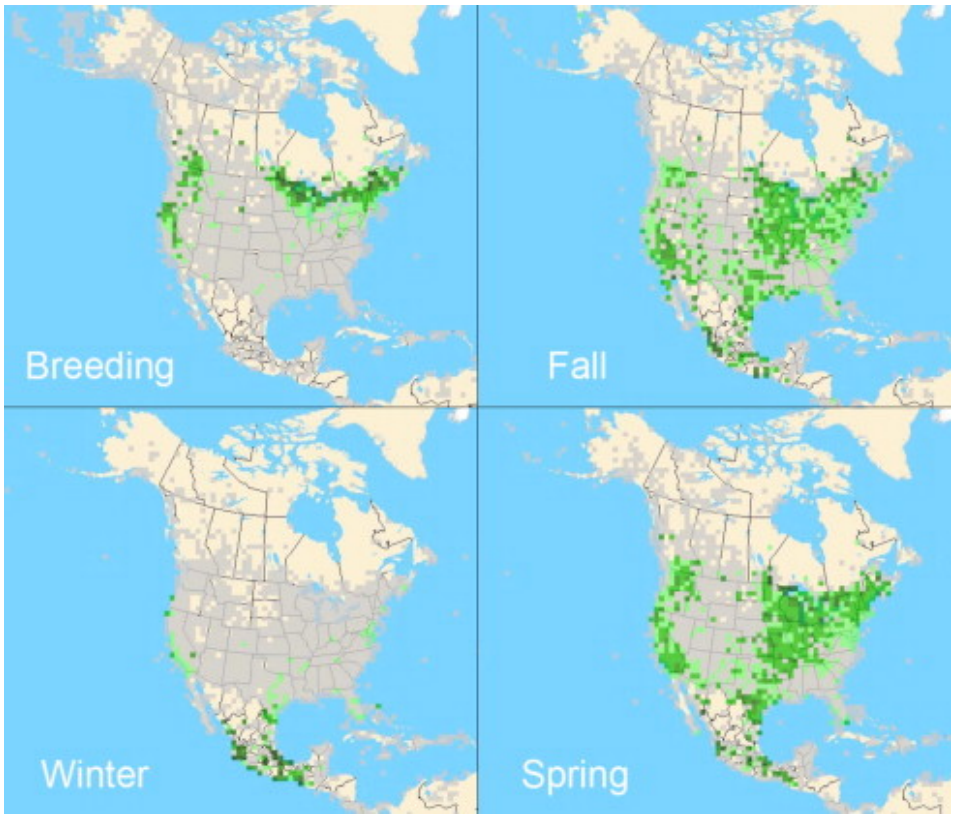
- MSc agronomy (Institut Agro, Rennes)
- PhD between Ifremer / Institut Agro (Rennes, Nantes, Brest)
- Postdoc at University of Washington (Seattle)

Massive sources of spatio-temporal data

Epidemiology
(Healthcare data)



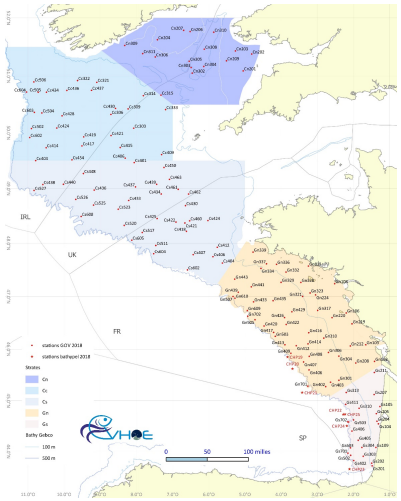
Ecology
(Ebirds data)



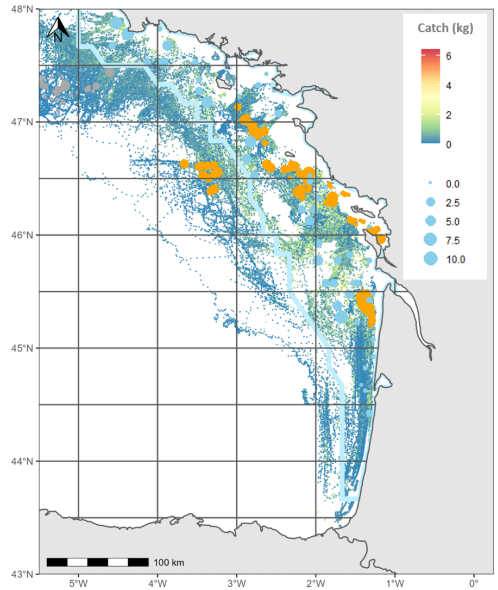
Now ubiquitous in many fields of application (agronomy, geography, climatology, oceanography, economy and **fisheries**)

	Standardized data	New massive data
+	Standardized sampling plan Exact locations	Both tight and wide spatio-temporal coverage
−	High costs Small sample size No backup when sampling issues	Opportunistic/preferantial sampling Diverse spatial scales, spatial aggregation issues

Examples



EVHOE survey data



"VMS x logbook" data

How to infer spatio-temporal processes from these data?
How to combine all these data sources?

Several methodological issues:

- Combine massive data sources with standardized data sources
- Non-standardized/preferential sampling
- Use aggregated data to infer fine-scale processes (change of support)

How to infer spatio-temporal processes from these data?
How to combine all these data sources?

Several methodological issues:

- Combine massive data sources with standardized data sources
- Non-standardized/preferential sampling
- Use aggregated data to infer fine-scale processes (change of support)

Hierarchical models as framework

Let's define observations Y:

$$Y|S,\theta \sim \mathcal{L}_Y(S,\theta_{obs})$$

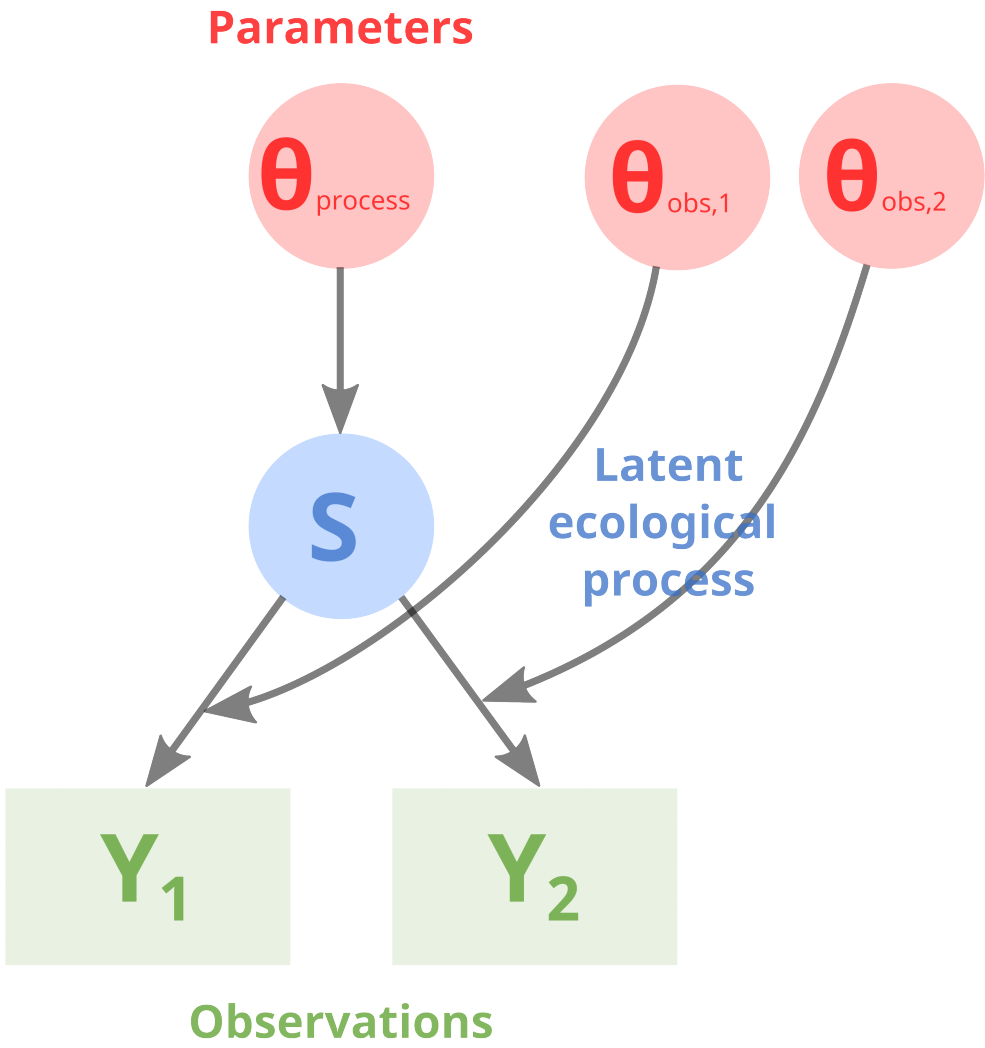
a latent field S:

$$S|\theta_{process} \sim \mathcal{L}_S(\theta_{process})$$

and parameters:

$$\theta = (\theta_{obs},\theta_{process})$$

There can be several data sources Y_1 and Y_2, in which case each has its own probability distribution ($\mathcal{L}_{Y_1}$,$\mathcal{L}_{Y_2}$) and observation parameters ($\theta_{obs,1}$, $\theta_{obs,2}$).

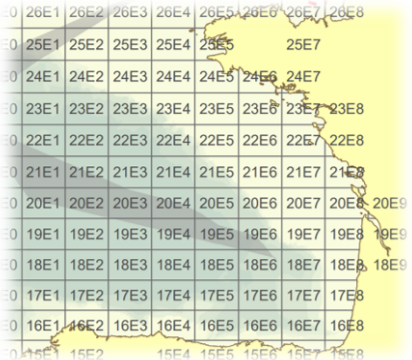


Application field: fisheries data

Scientific survey

Onboard observer

Catch declarations

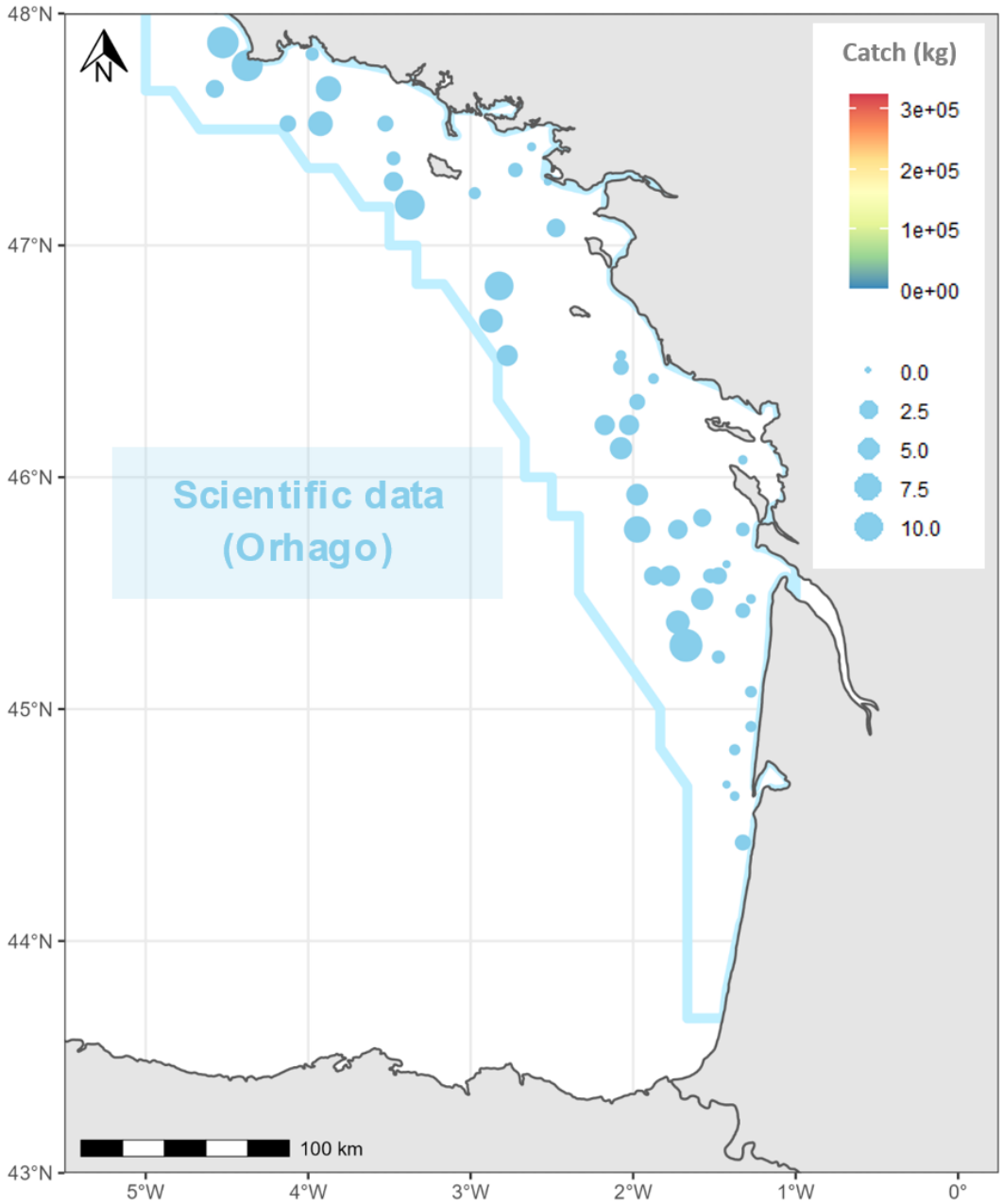


Vessel Monitoring System (VMS)

Scientific survey

Standardized
Exact locations

Low sampling volume



Onboard observer

Available over whole year

Preferential sampling

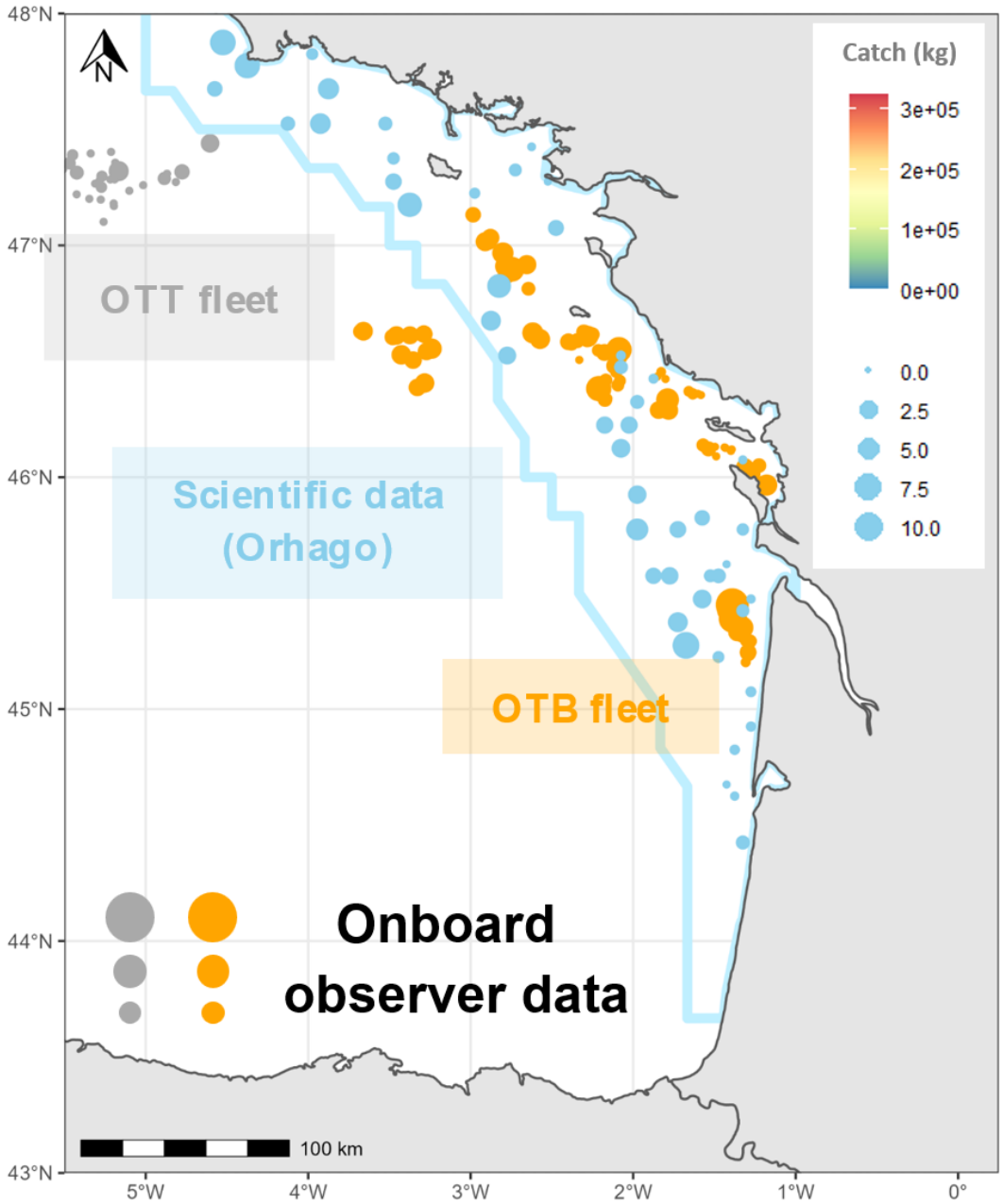
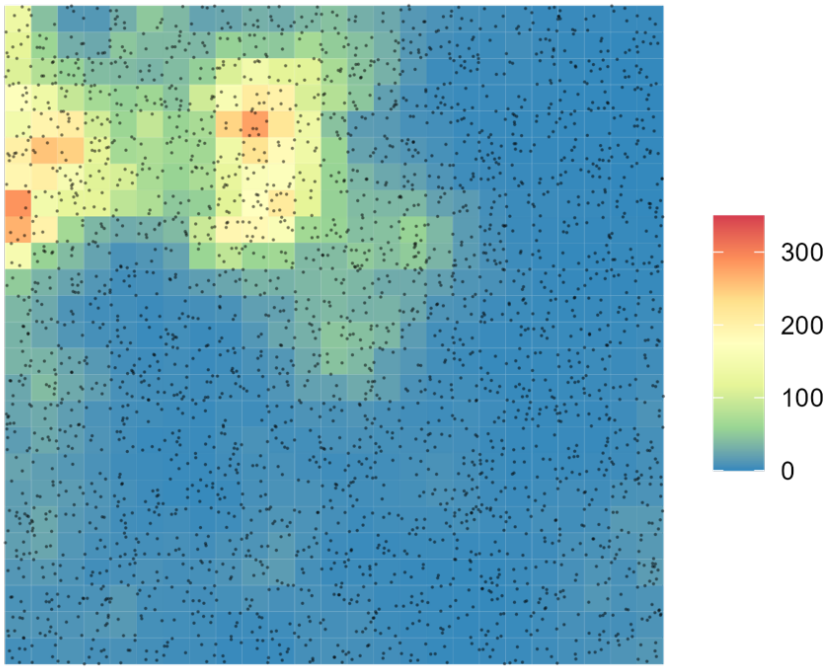


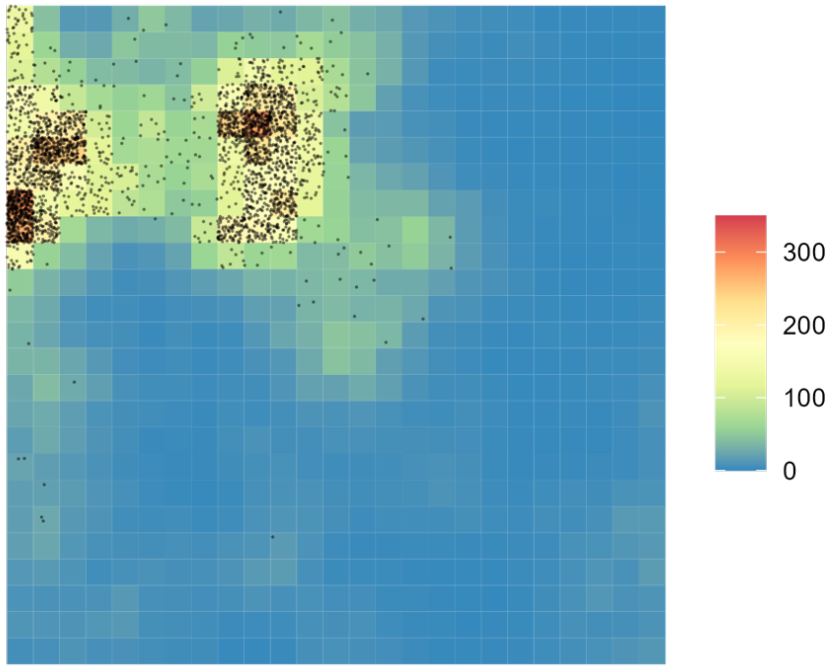
Table of Contents

What is preferential sampling?

Uniform sampling



Strong preferential sampling



- When sampling locations depend on the process under study

Following Diggle *et al.* (2010), let's define a spatial process $\boldsymbol{S} = \{S(\boldsymbol{X}) : \boldsymbol{X} \in \mathbb{R}^2\}$, sampling locations $\boldsymbol{X} = (X_1, ..., X_n)$ and observations $\boldsymbol{Y} = (Y_1, ..., Y_n)$.

A complete model needs to specify the joint distribution of $\boldsymbol{S}$, $\boldsymbol{X}$ and $\boldsymbol{Y}$:

$$\begin{aligned} P[\boldsymbol{S}, \boldsymbol{X}, \boldsymbol{Y}] &= P[\boldsymbol{S}] \cdot P[\boldsymbol{X}, \boldsymbol{Y} | \boldsymbol{S}] \\ &= P[\boldsymbol{S}] \cdot P[\boldsymbol{X} | \boldsymbol{S}] \cdot P[\boldsymbol{Y} | \boldsymbol{X}, \boldsymbol{S}] \end{aligned}$$

Sampling is independent from the latent process

$$P[\boldsymbol{S}, \boldsymbol{X}] = P[\boldsymbol{S}] \cdot P[\boldsymbol{X}]$$

Sampling is dependent from the latent process

$$P[\boldsymbol{S}, \boldsymbol{X}] \neq P[\boldsymbol{S}] \cdot P[\boldsymbol{X}]$$

Assuming independence between S and X:

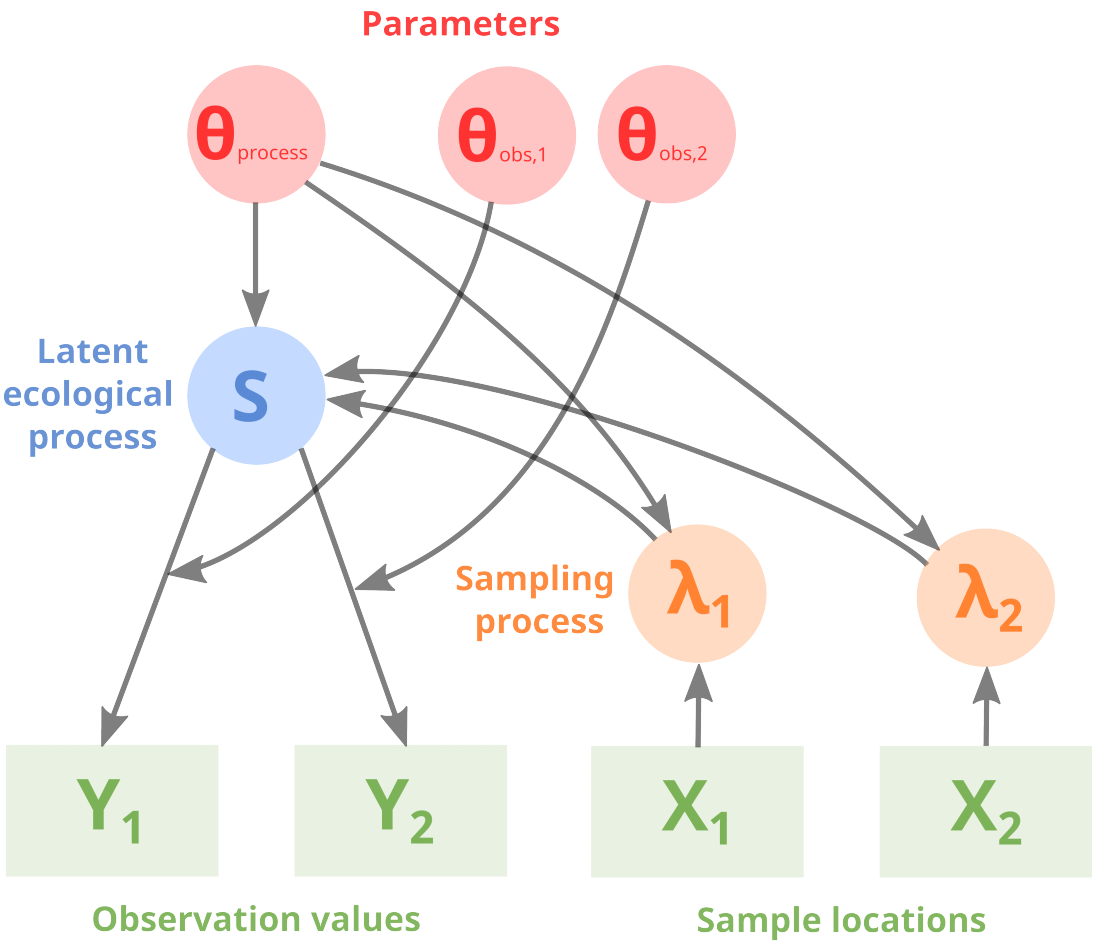
$$\begin{aligned} P[S, X, Y] &= P[S] \cdot P[X|S] \cdot P[Y|X, S] \\ &= P[S] \cdot P[X] \cdot P[Y|X, S] \end{aligned}$$

By conditioning on X, we find back the equation on which standard geostatistics is based:

$$P[S, Y] = P[S] \cdot P[Y|S]$$

However, under **preferential sampling**, such assumption does not hold and one has to consider the dependence between S and X in inference.

To model the dependence between X and S, one solution is to account for X in inference and relate these to S through an extra layer λ.



Diggle et al., (2010)

$$S(x) = \mu_S + \delta(x)$$
$$\delta \sim GF(0, \mathcal{C}(x, x'))$$

- - - - - - - - - - -

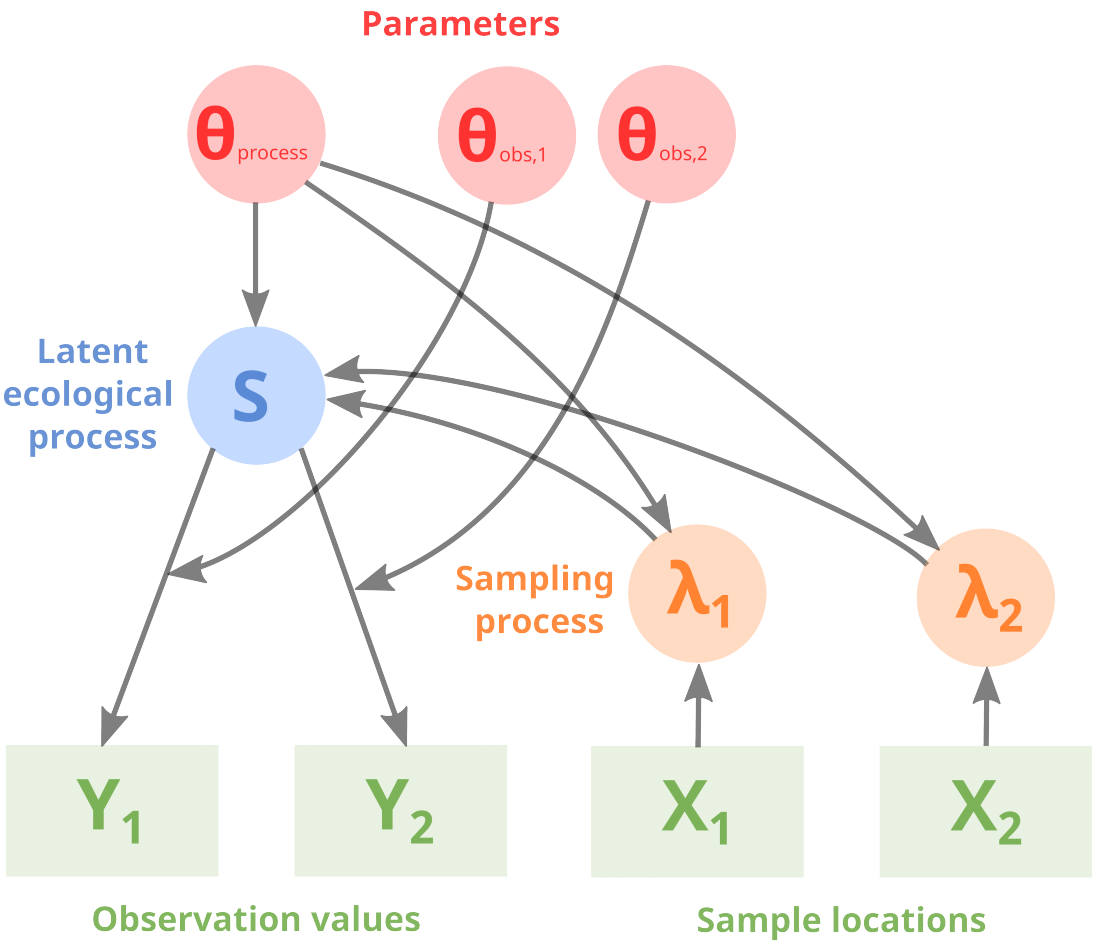
$$X \sim \mathcal{IPP}(\lambda(x))$$
$$\log \lambda(x) = \mu_X + b \cdot S(x)$$

- - - - - - - - - - -

$$Y_i \sim \mathcal{N}(S(x_i), \sigma^2)$$

➠ **Still some limits:** only spatial, only one sampling process, Gaussian observations and latent field, parameterization of preferential sampling (often X is rather a mixture of preferential sampling and other processes).

To model the dependence between X and S, one solution is to account for X in inference and relate these to S through an extra layer λ.



Diggle et al., (2010)

$$S(x) = \mu_S + \delta(x)$$
$$\delta \sim GF(0, \mathcal{C}(x, x'))$$

- - - - - - - - - - -

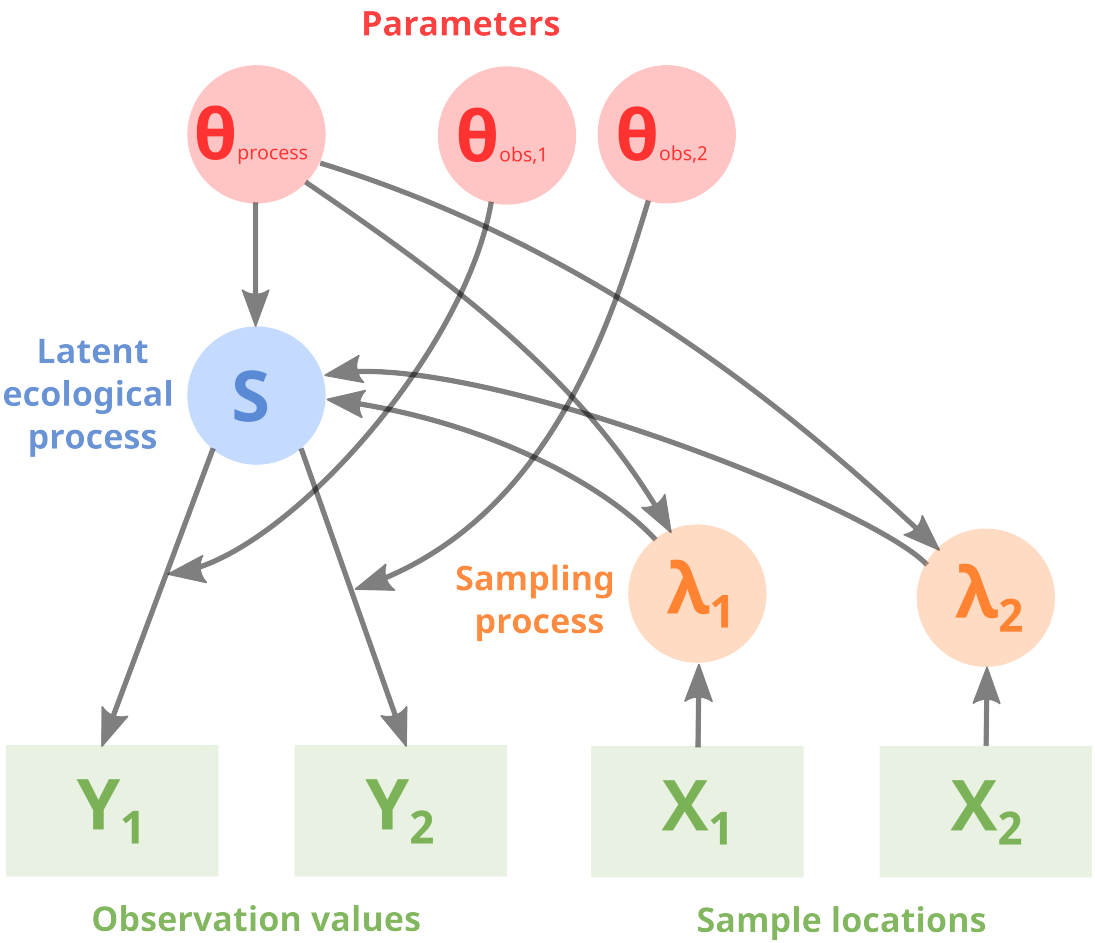
$$X \sim \mathcal{IPP}(\lambda(x))$$
$$\log \lambda(x) = \mu_X + b \cdot S(x)$$

- - - - - - - - - - -

$$Y_i \sim \mathcal{N}(S(x_i), \sigma^2)$$

➠ **Still some limits:** only spatial, only one sampling process, Gaussian observations and latent field, parameterization of preferential sampling (often X is rather a mixture of preferential sampling and other processes).

To model the dependence between $\boldsymbol{X}$ and S, one solution is to account for $\boldsymbol{X}$ in inference and relate these to S through an extra layer λ.



Alglave et al., (2022)

$$f(S(x)) = \mu_S + \delta(x)$$
$$\delta \sim GF(0, \mathcal{C}(x, x'))$$

- - - - - - - - - - -

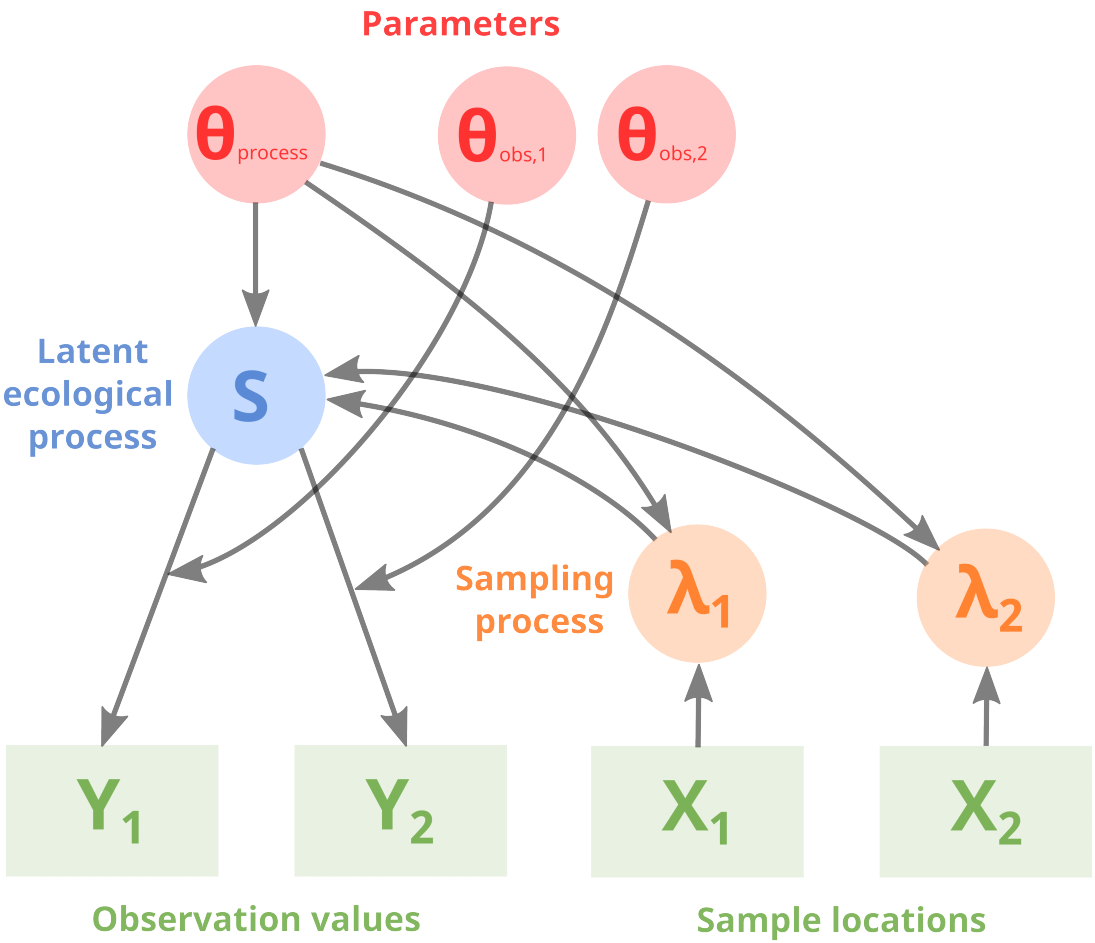
$$\boldsymbol{X} \sim \mathcal{IPP}(\lambda(x))$$
$$\log \lambda(x) = \mu_X + b \cdot f(S(x))$$

- - - - - - - - - - -

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

➠ **Still some limits:** only spatial, only one sampling process, **Gaussian observations and latent field**, parameterization of preferential sampling (often $\boldsymbol{X}$ is rather a mixture of preferential sampling and other processes).

To model the dependence between $\boldsymbol{X}$ and S, one solution is to account for $\boldsymbol{X}$ in inference and relate these to S through an extra layer λ.



Alglave et al., (2022)

$$f(S(x)) = \mu_S + \delta(x)$$
$$\delta \sim GF(0, \mathcal{C}(x, x'))$$

- - - - - - - - - - - -

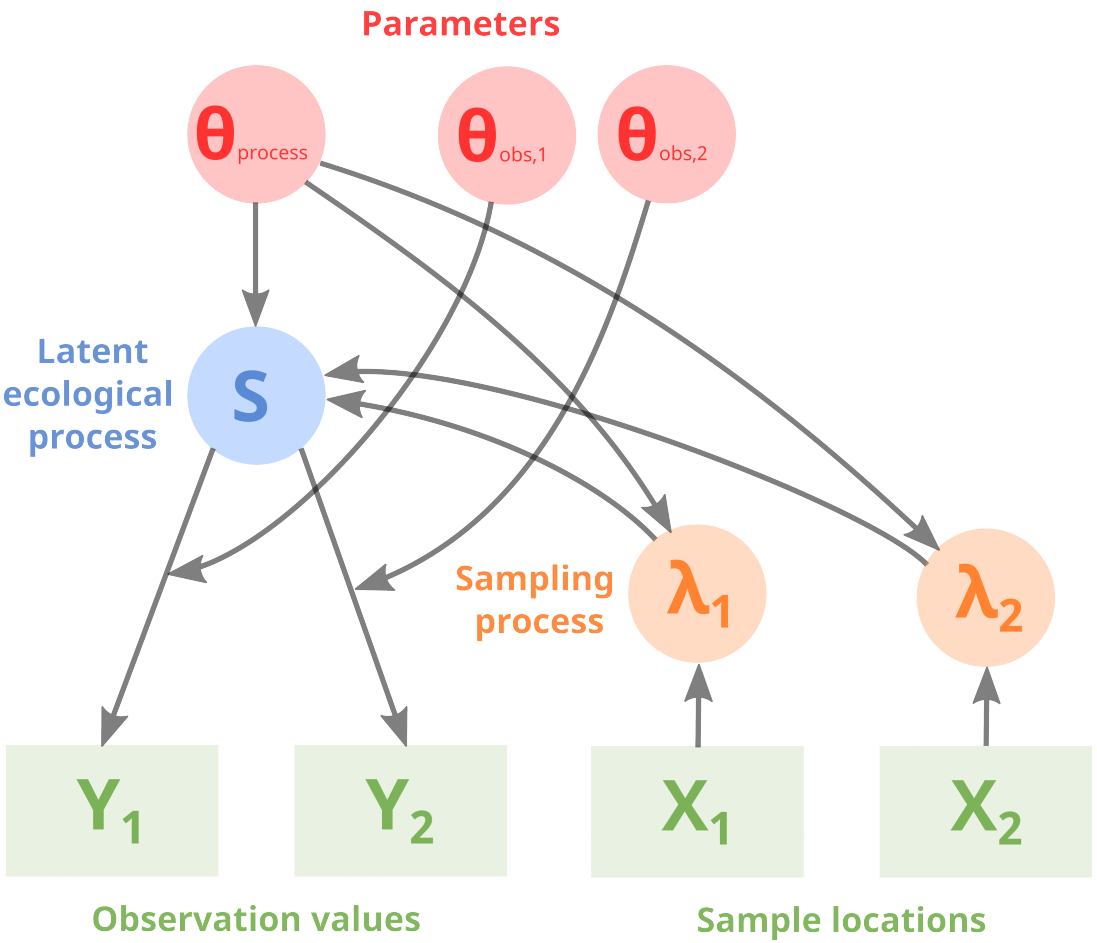
$$\boldsymbol{X}_j \sim \mathcal{IPP}(\lambda_j(x))$$
$$\log \lambda_j(x) = \mu_{X,j} + b_j \cdot f(S(x))$$

- - - - - - - - - - - -

$$Y_i \sim \mathcal{L}(q_j \cdot S(x_i), \sigma_j^2)$$

➠ **Still some limits:** only spatial, **only one sampling process**, Gaussian observations and latent field, parameterization of preferential sampling (often $\boldsymbol{X}$ is rather a mixture of preferential sampling and other processes).

To model the dependence between X and S, one solution is to account for X in inference and relate these to S through an extra layer λ.



Alglave et al., (2022)

$$f(S(x)) = \mu_S + \delta(x)$$
$$\delta \sim GF(0, \mathcal{C}(x, x'))$$

- - - - - - - - - - -

$$X_j \sim \mathcal{IPP}(\lambda_j(x))$$
$$\log \lambda_j(x) = \mu_{X,j} + b_j \cdot f(S(x)) + \Gamma_X(x)^T \cdot \beta_{X,j} + \eta_j(x)$$

- - - - - - - - - - -

$$Y_i \sim \mathcal{L}(q_j \cdot S(x_i), \sigma_j^2)$$

➠ **Still some limits:** only spatial, only one sampling process, Gaussian observations and latent field, **parameterization of preferential sampling (often X is rather a mixture of preferential sampling and other processes)**.

Inference method

For both the Bayesian and the frequentist framework, the computation burden comes from the integration of the likelihood over the random effects:

$$L_M(\boldsymbol{\theta}) = P[\boldsymbol{Y}|\boldsymbol{\theta}] = \int_{\mathbb{R}^q} P[\boldsymbol{Y}, \boldsymbol{\delta}|\boldsymbol{\theta}] d\boldsymbol{\delta}$$

The Laplace approximation allows to strongly improve speed in both cases.

Frequentist framework

(Template Model Builder - Kristensen et al., 2014)

Simplify the expression of $L_M(\boldsymbol{\theta})$

$$L_M(\boldsymbol{\theta}) \approx L_M^*(\boldsymbol{\theta}) = (2\pi)^{q/2} |\boldsymbol{H}(\boldsymbol{\theta})|^{-1/2} \exp[-f_{nll}(\boldsymbol{\theta}, \boldsymbol{\delta}^*(\boldsymbol{\theta}))]$$

➠ Search for $\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left(L_M^*(\boldsymbol{\theta})\right)$

Bayesian framework

(R-INLA - Rue et al., 2015)

Approximate the marginal distribution

$$\tilde{P}(\boldsymbol{\theta}|\boldsymbol{Y}) \propto \left. \frac{P(\boldsymbol{\delta}, \boldsymbol{\theta}, \boldsymbol{Y})}{\tilde{P}_{\boldsymbol{G}}(\boldsymbol{\delta}|\boldsymbol{\theta}, \boldsymbol{Y})} \right|_{\boldsymbol{\delta}=\boldsymbol{\delta}^*(\boldsymbol{\theta})}$$

$$\tilde{P}(\delta_i \mid \boldsymbol{Y}) = \int \tilde{P}(\delta_i \mid \boldsymbol{\theta}, \boldsymbol{Y}) \tilde{P}(\boldsymbol{\theta} \mid \boldsymbol{Y}) d\boldsymbol{\theta}$$

with $\mathbf{H}(\boldsymbol{\theta})$ the Hessian of the negative joint log-likelihood $f_{nll}(\boldsymbol{\theta}, \boldsymbol{\delta}^*(\boldsymbol{\theta}))$, q the size of the latent effect, $\boldsymbol{\delta}^*(\boldsymbol{\theta})$ the conditional mode of $\boldsymbol{\delta}$ relatively to fixed parameters $\boldsymbol{\theta}$.

Inference method

For both the Bayesian and the frequentist framework, the computation burden comes from the integration of the likelihood over the random effects:

$$L_M(\boldsymbol{\theta}) = P[\boldsymbol{Y}|\boldsymbol{\theta}] = \int_{\mathbb{R}^q} P[\boldsymbol{Y}, \boldsymbol{\delta}|\boldsymbol{\theta}] d\boldsymbol{\delta}$$

The Laplace approximation allows to strongly improve speed in both cases.

Frequentist framework

(Template Model Builder - Kristensen et al., 2014)

Simplify the expression of $L_M(\boldsymbol{\theta})$

$$L_M(\boldsymbol{\theta}) \approx L_M^*(\boldsymbol{\theta}) = (2\pi)^{q/2} |\boldsymbol{H}(\boldsymbol{\theta})|^{-1/2} \exp[-f_{nll}(\boldsymbol{\theta}, \boldsymbol{\delta}^*(\boldsymbol{\theta})]$$

➠ Search for $\underset{\boldsymbol{\theta}}{\text{argmax}}\,(L_M^*(\boldsymbol{\theta}))$

Bayesian framework

(R-INLA - Rue et al., 2015)

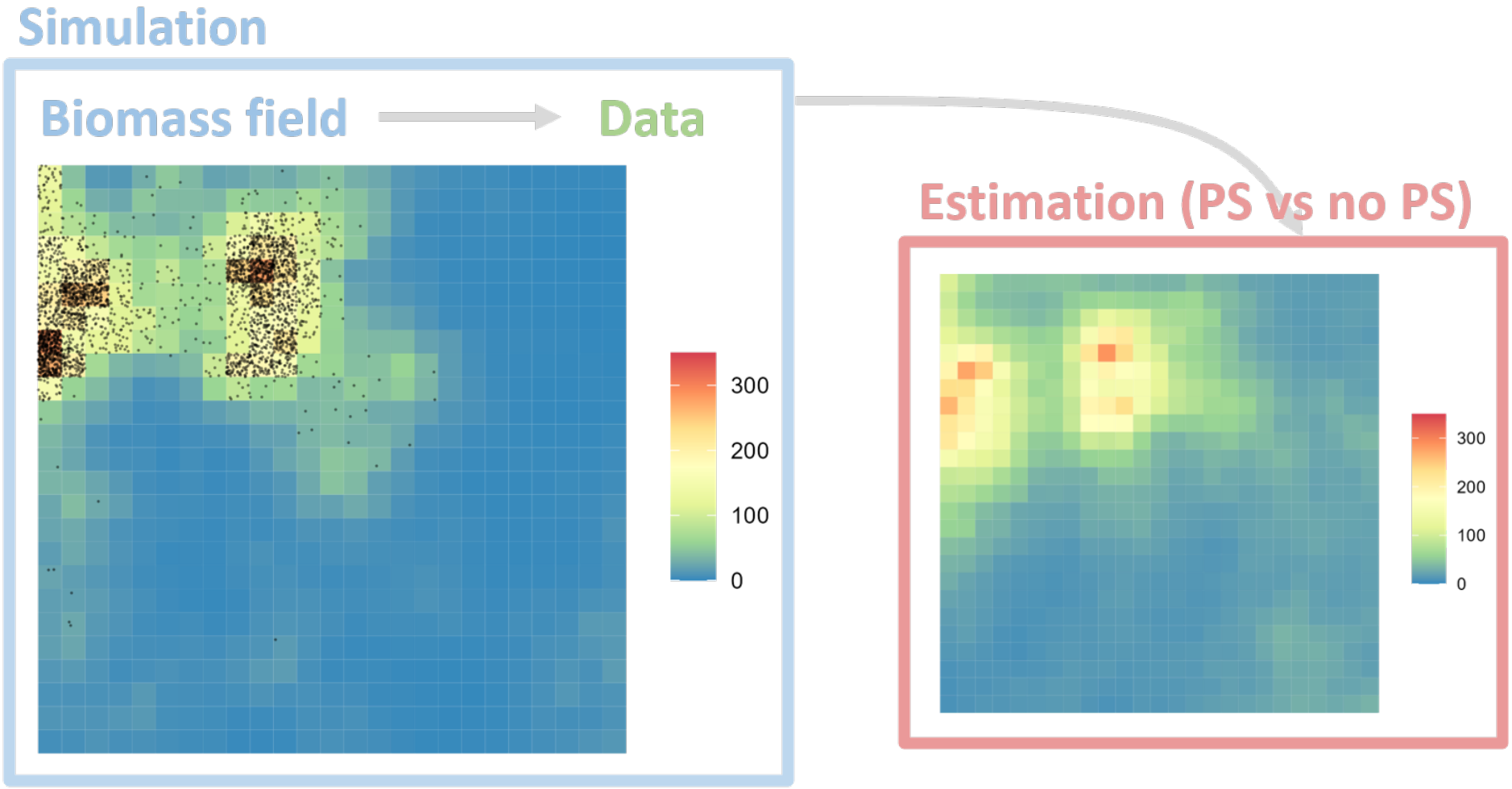
Approximate the marginal distribution

$$\tilde{P}(\boldsymbol{\theta}|\boldsymbol{Y}) \propto \left. \frac{P(\boldsymbol{\delta}, \boldsymbol{\theta}, \boldsymbol{Y})}{\tilde{P}_{\boldsymbol{G}}(\boldsymbol{\delta}|\boldsymbol{\theta}, \boldsymbol{Y})} \right|_{\boldsymbol{\delta}=\boldsymbol{\delta}^*(\boldsymbol{\theta})}$$

$$\tilde{P}(\delta_i \mid \boldsymbol{Y}) = \int \tilde{P}(\delta_i \mid \boldsymbol{\theta}, \boldsymbol{Y}) \tilde{P}(\boldsymbol{\theta} \mid \boldsymbol{Y}) d\theta$$

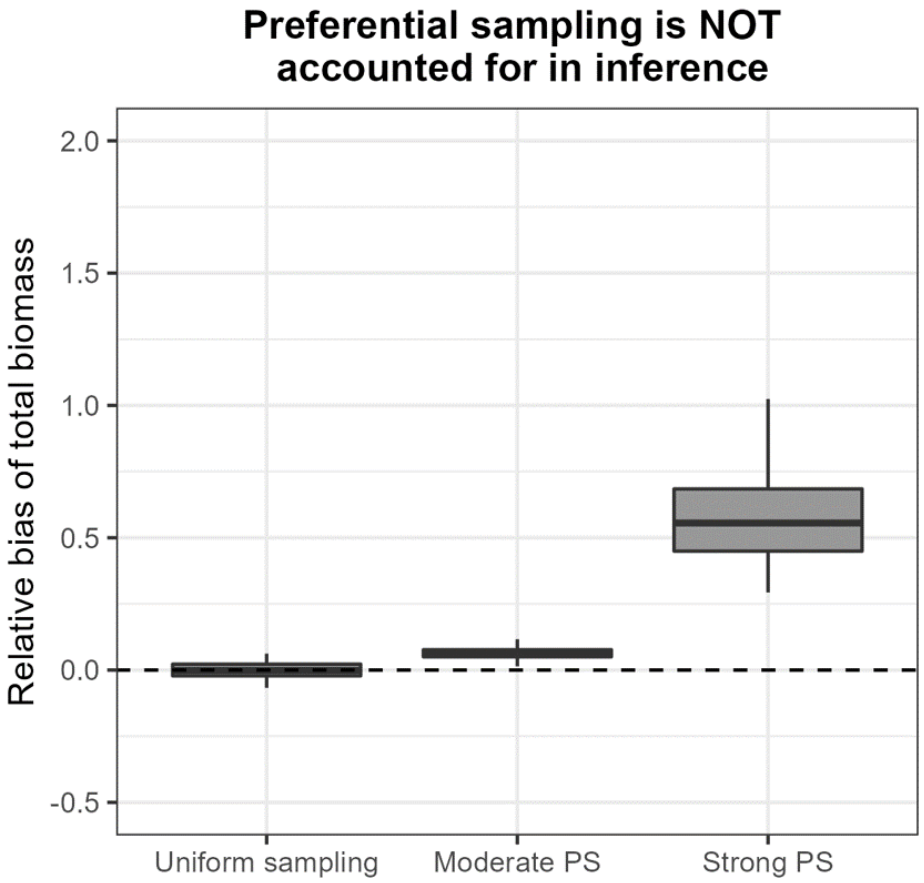
with $\mathbf{H}(\boldsymbol{\theta})$ the Hessian of the negative joint log-likelihood $f_{nll}(\boldsymbol{\theta}, \boldsymbol{\delta}^*(\boldsymbol{\theta}))$, q the size of the latent effect, $\boldsymbol{\delta}^*(\boldsymbol{\theta})$ the conditional mode of $\boldsymbol{\delta}$ relatively to fixed parameters $\boldsymbol{\theta}$.

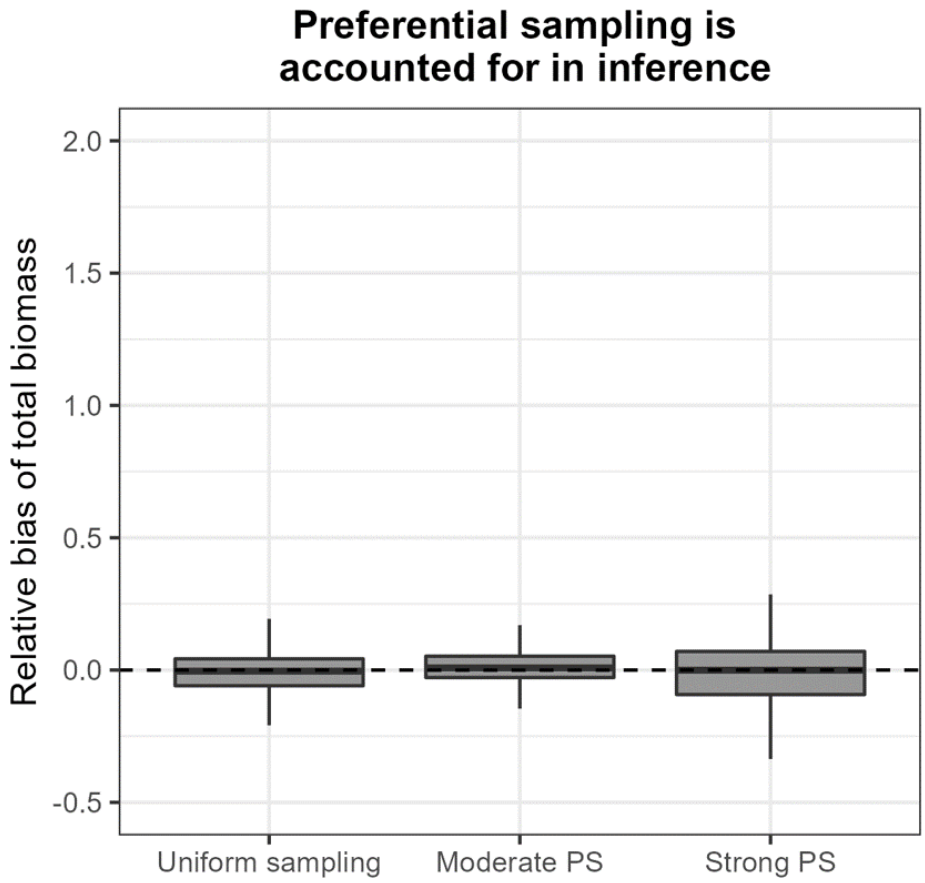
Let's do some simulations to assess the effect of PS on inference.



Comparison metric: Bias of biomass $= \frac{\hat{B}-B}{B}$ with $B = \int_{\mathcal{D}} S(x)dx$

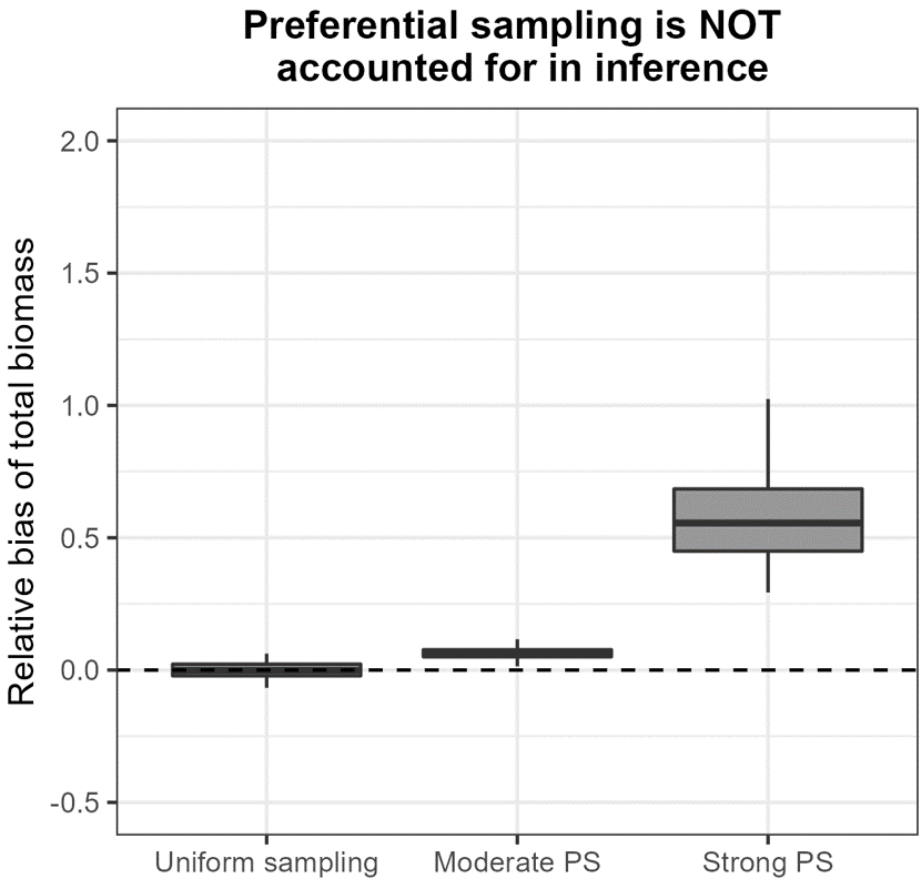
What is the problem with ignoring preferential sampling?

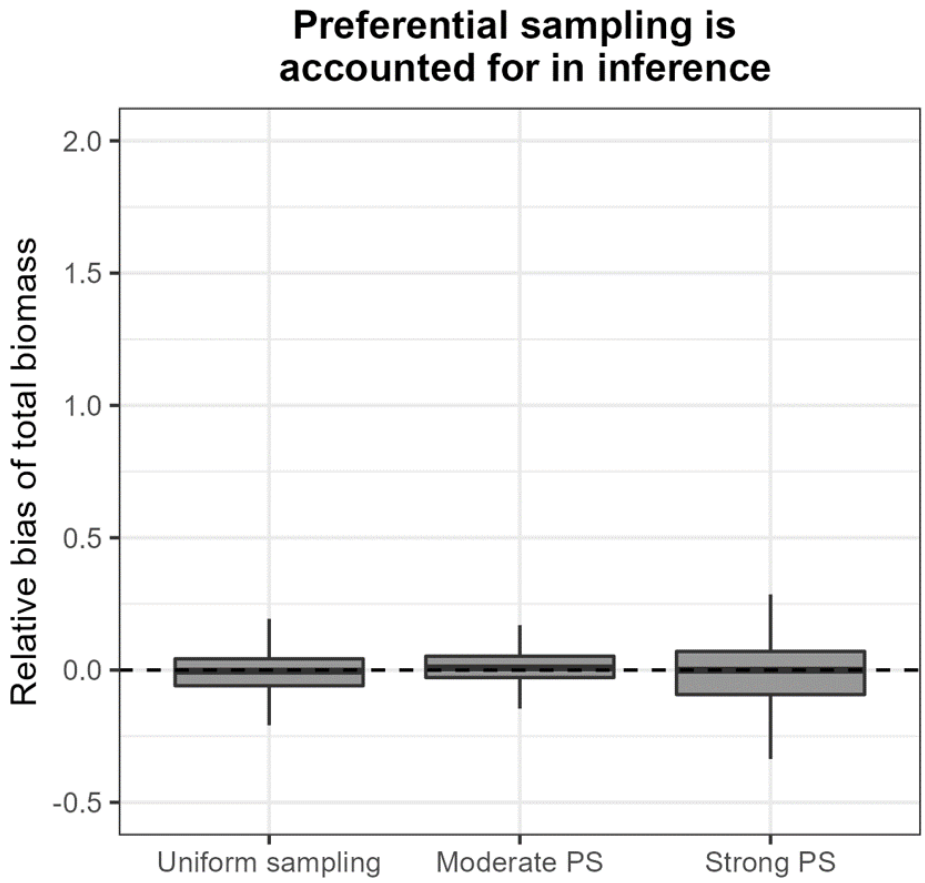




Ignoring PS leads to strong over-estimation of species distribution when PS is strong ➠ it needs to be accounted for in inference

What is the problem with ignoring preferential sampling?



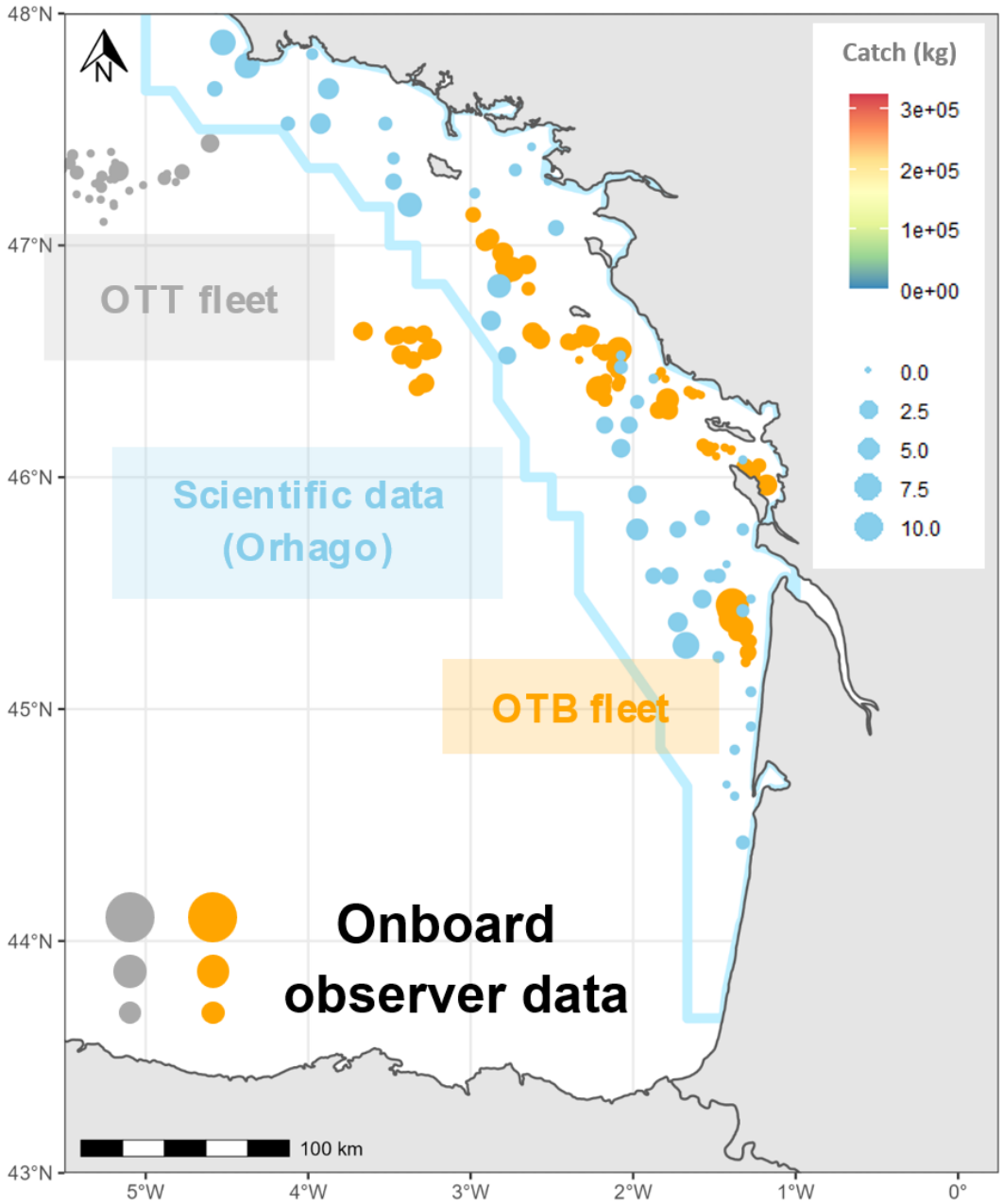


Ignoring PS leads to strong over-estimation of species distribution when PS is strong ➠ it needs to be accounted for in inference

Onboard observer

Available over whole year

Preferential sampling



Catch declarations

Large amount of data

Aggregated over rough scale

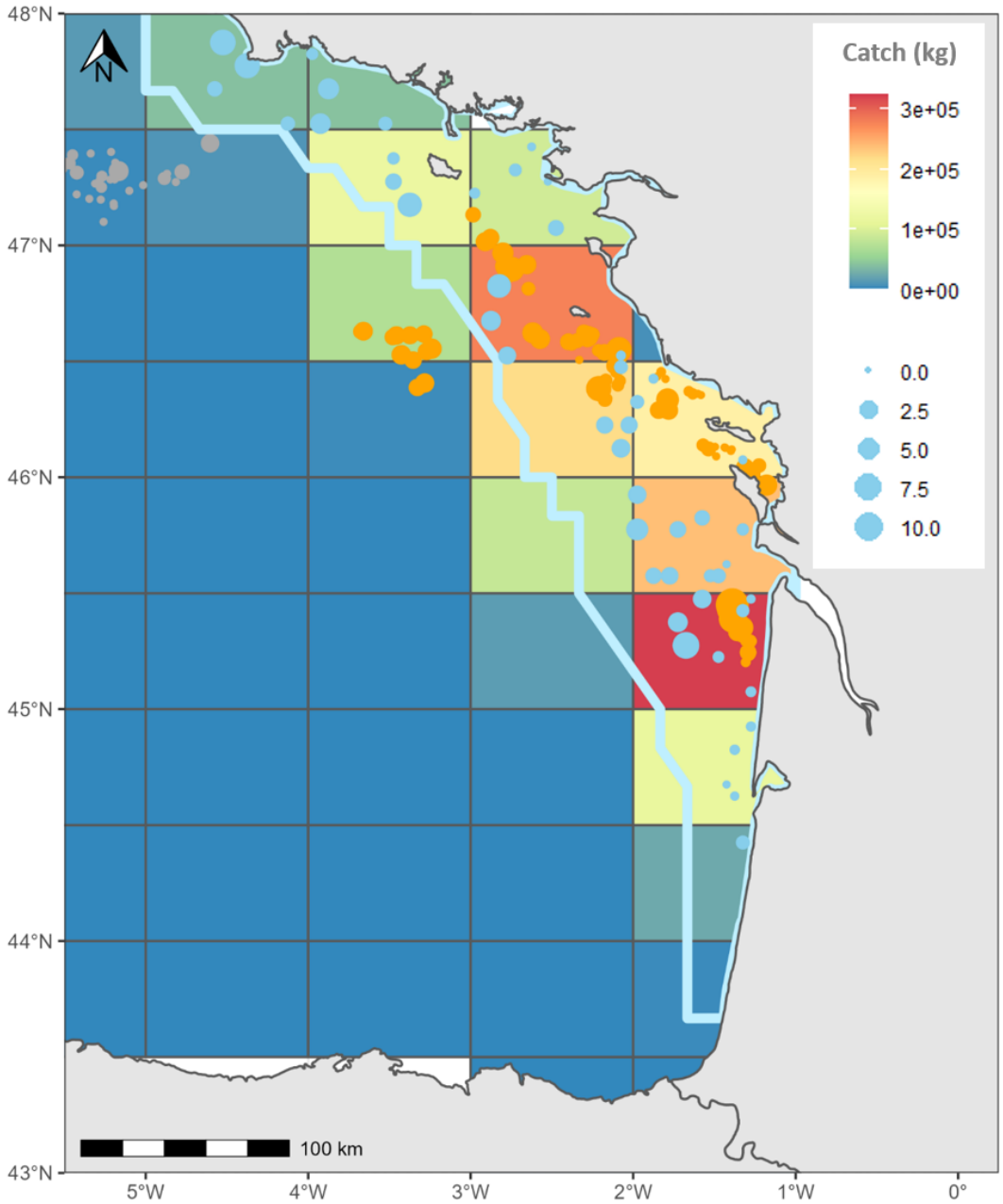


Table of Contents

Big challenge

Combine:

- Punctual observations

$$Y_i | S(x_i), \theta_{obs}$$

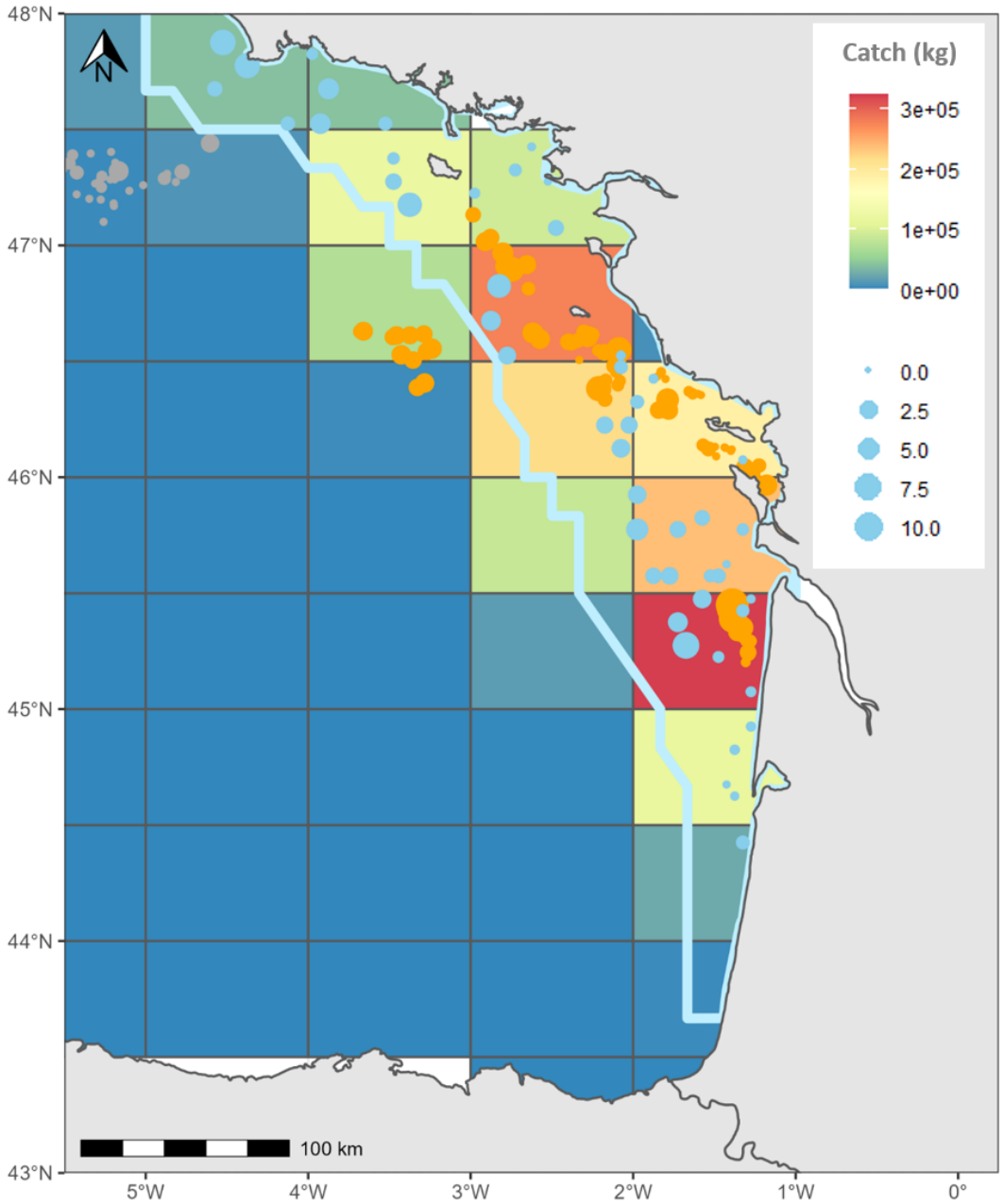
with

- Aggregated observations

$$D_a | S_a, \theta_{obs}$$

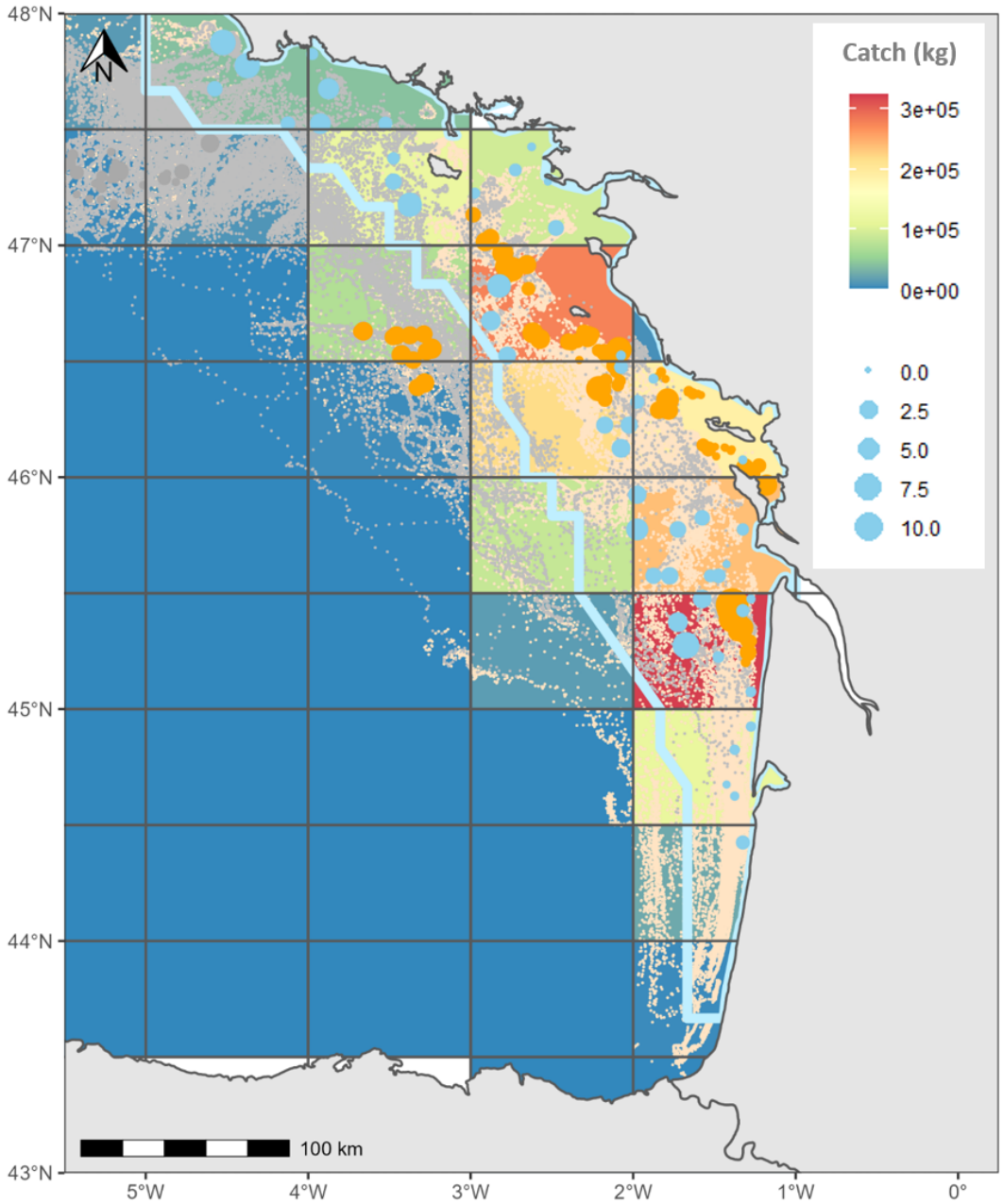
with a being the spatial aggregation unit level

while observations are possibly complex (i.e. zero-inflated positive continuous)

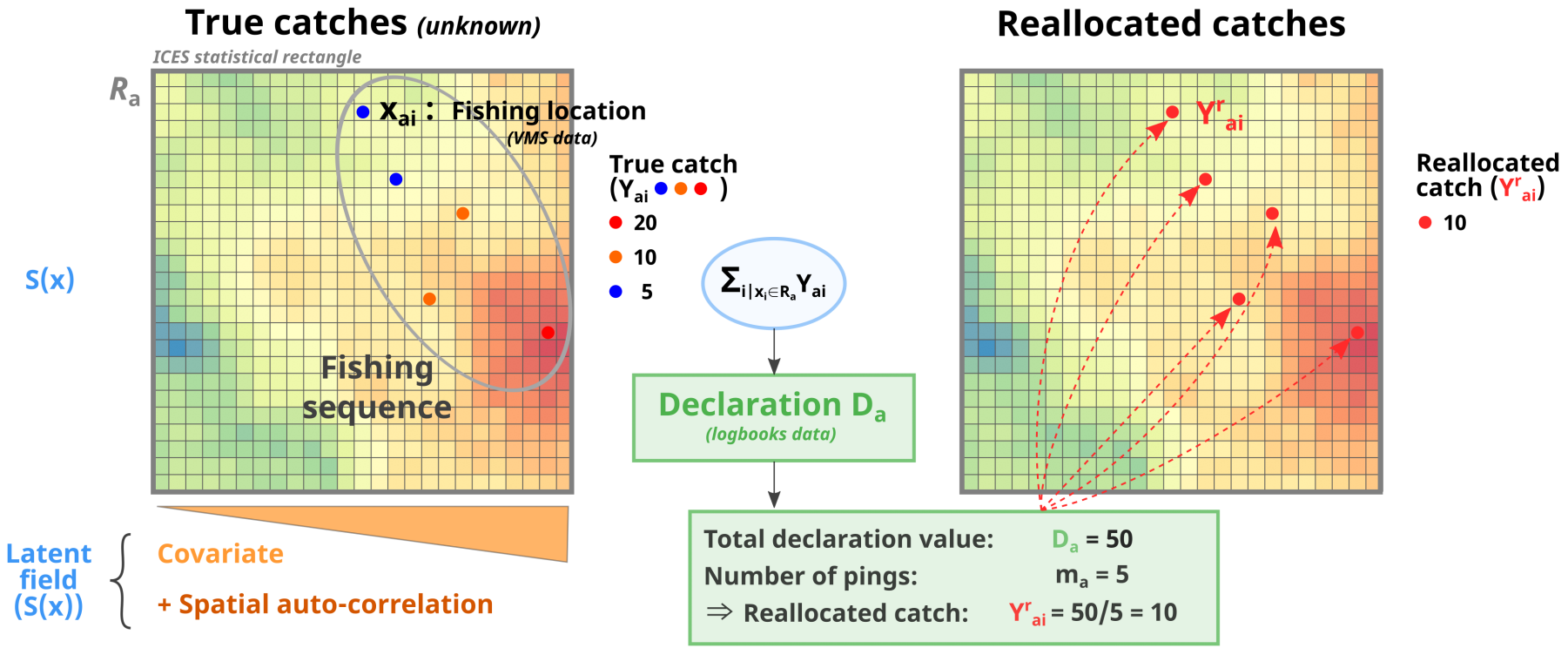


VMS

Locations of fishing positions



How to handle change of support?

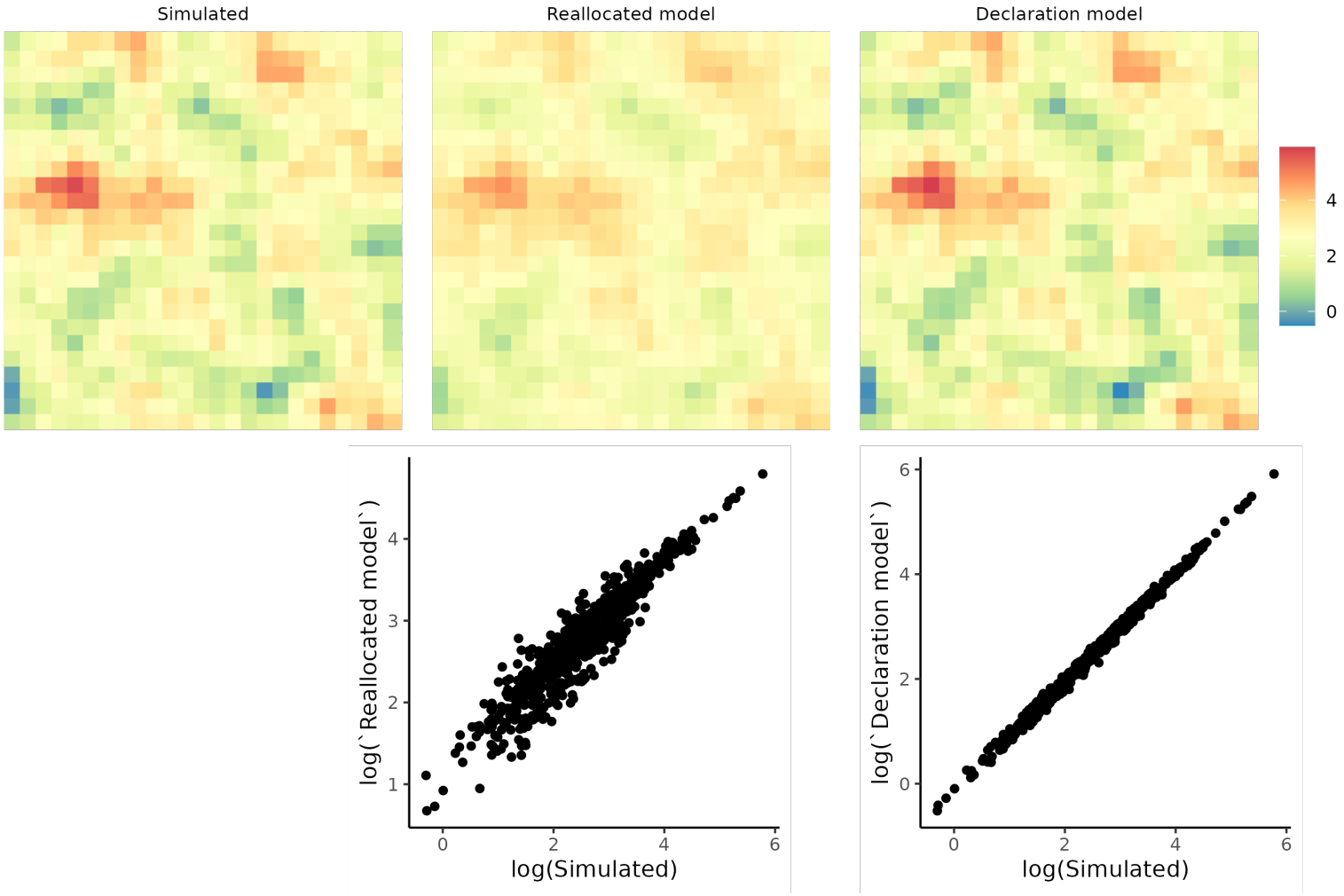


Standard way is rough (➠ **Reallocated approach**)

Another option (➠ **Declaration model**):

- define $\mathcal{L}_Y$ the probability distribution for (unobserved) punctual observation Y_{ai}
- consider $D_a = \sum_{i|x_i \in \mathcal{R}_a} Y_{ai}$
- Accordingly, define $\mathcal{L}_D$ the distribution of D_a by matching the moments of D_a and Y_{ai}

Simulation testing



➠ **The Reallocated approach** predicts rough smoothed maps, while the **Declaration model** predicts more accurate maps

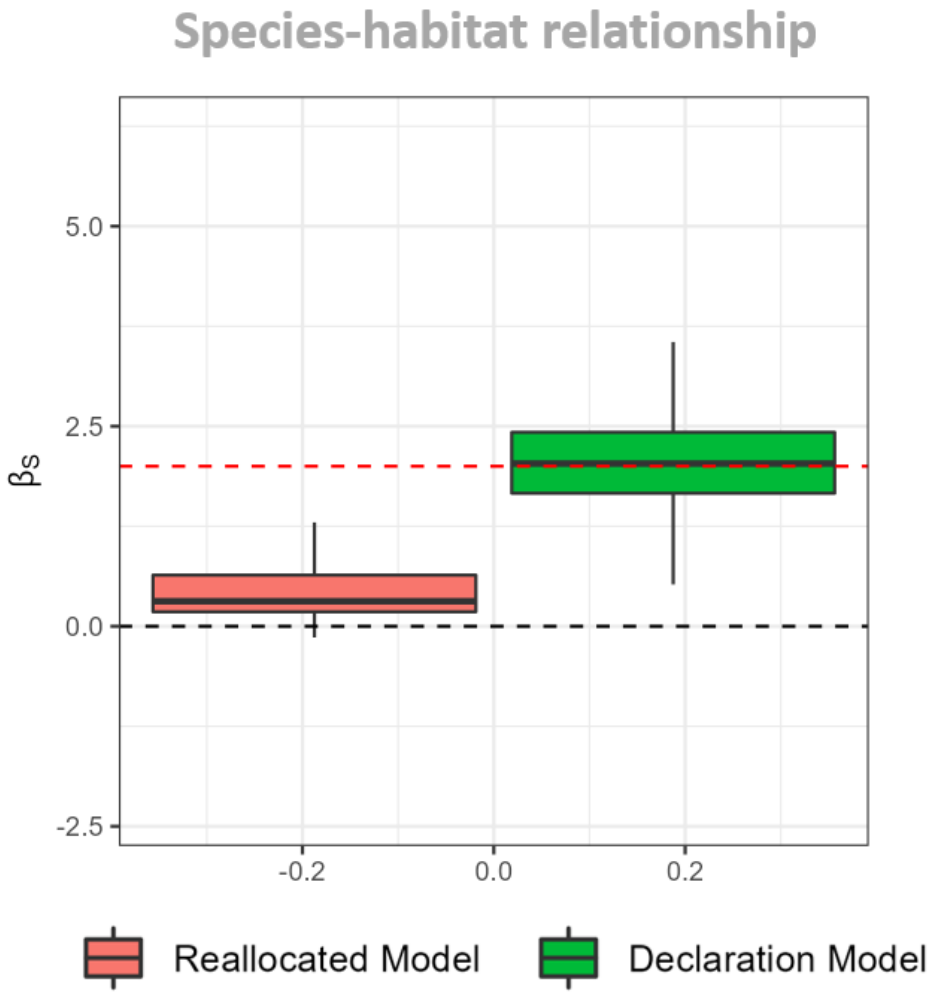
Look at the parameter of the species habitat relationship β

$$log(S(x)) = \mu_S + \mathbf{\Gamma}_S(x)^T \cdot \boldsymbol{\beta} + \delta(x)$$

Compare $\beta = 2$ to estimated $\hat{\beta}$

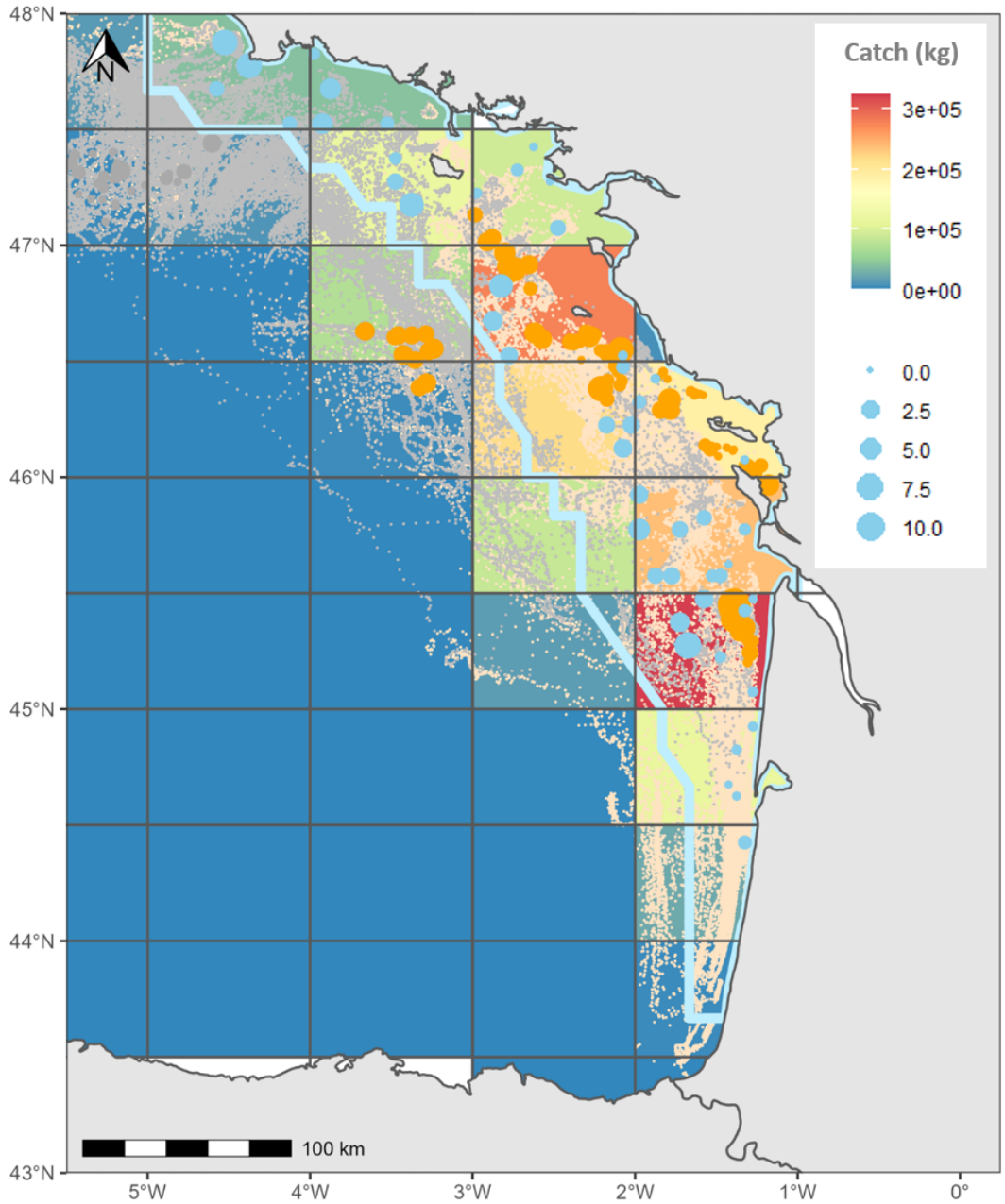
Reallocated approach ➠ Biased $\hat{\beta}$, loss of the species-habitat relationship.

Declaration model ➠ Unbiased $\hat{\beta}$, recover the species-habitat relationship.



Let's combine the data sources!

➠ Extend the framework in time



Let's combine the data sources!

➠ Extend the framework in time

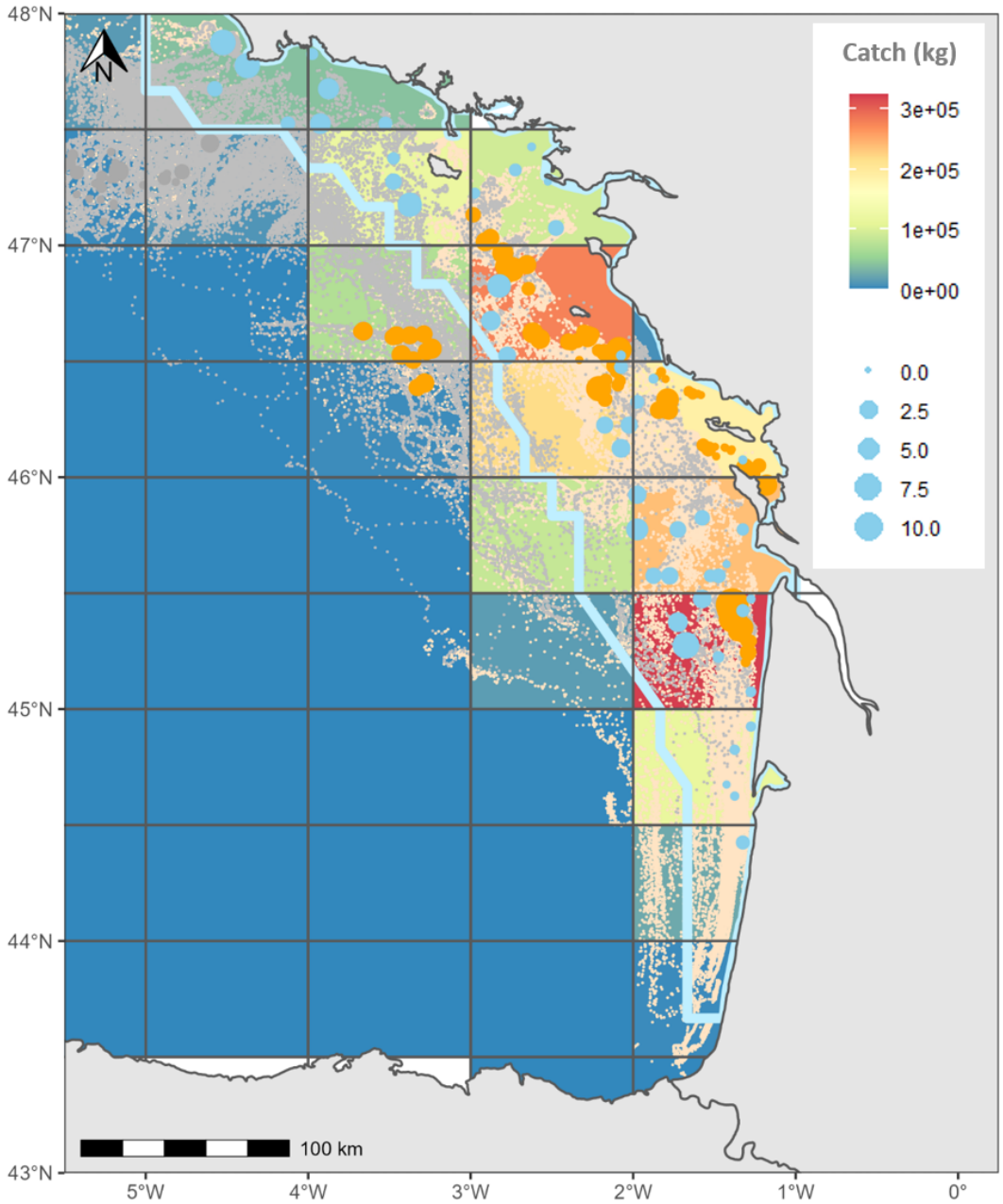
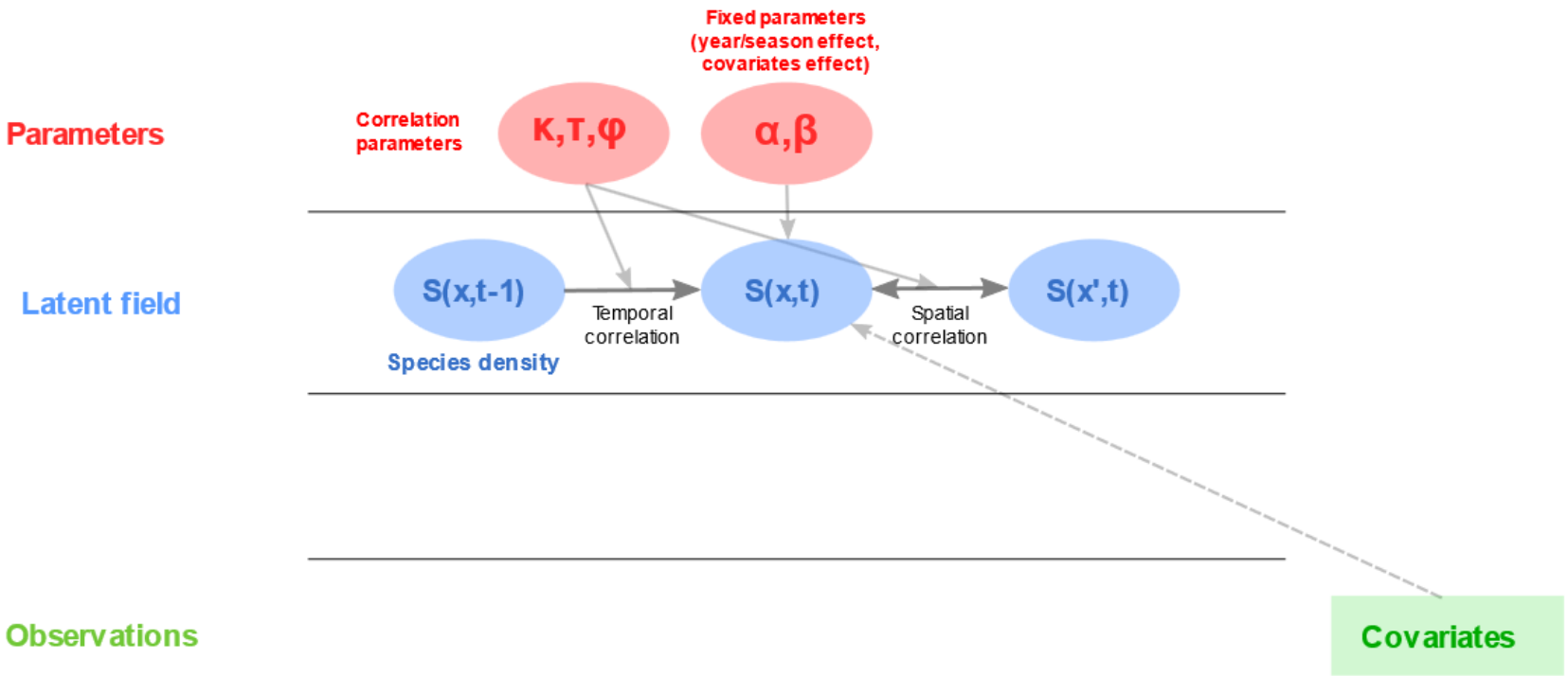
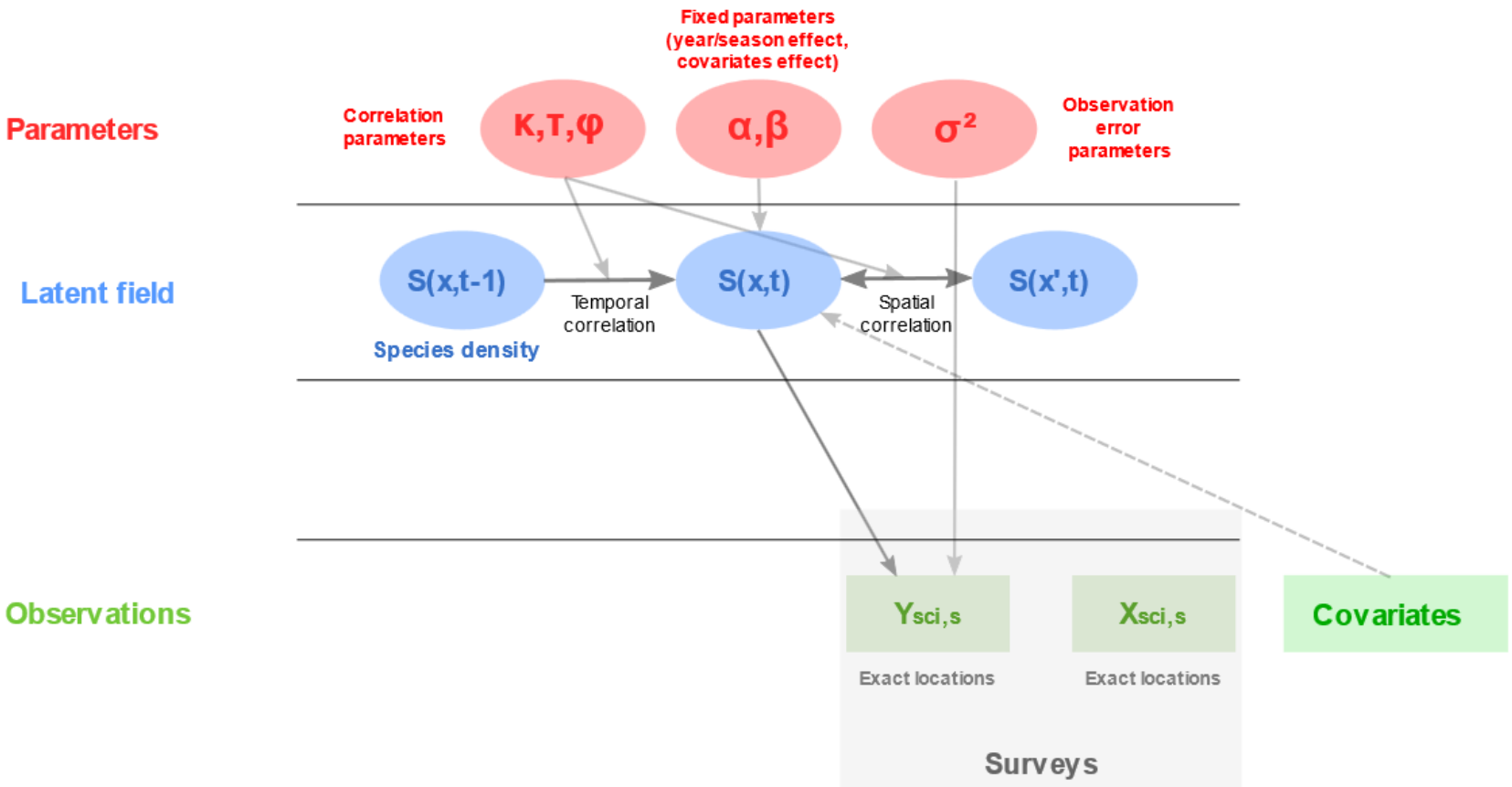


Table of Contents

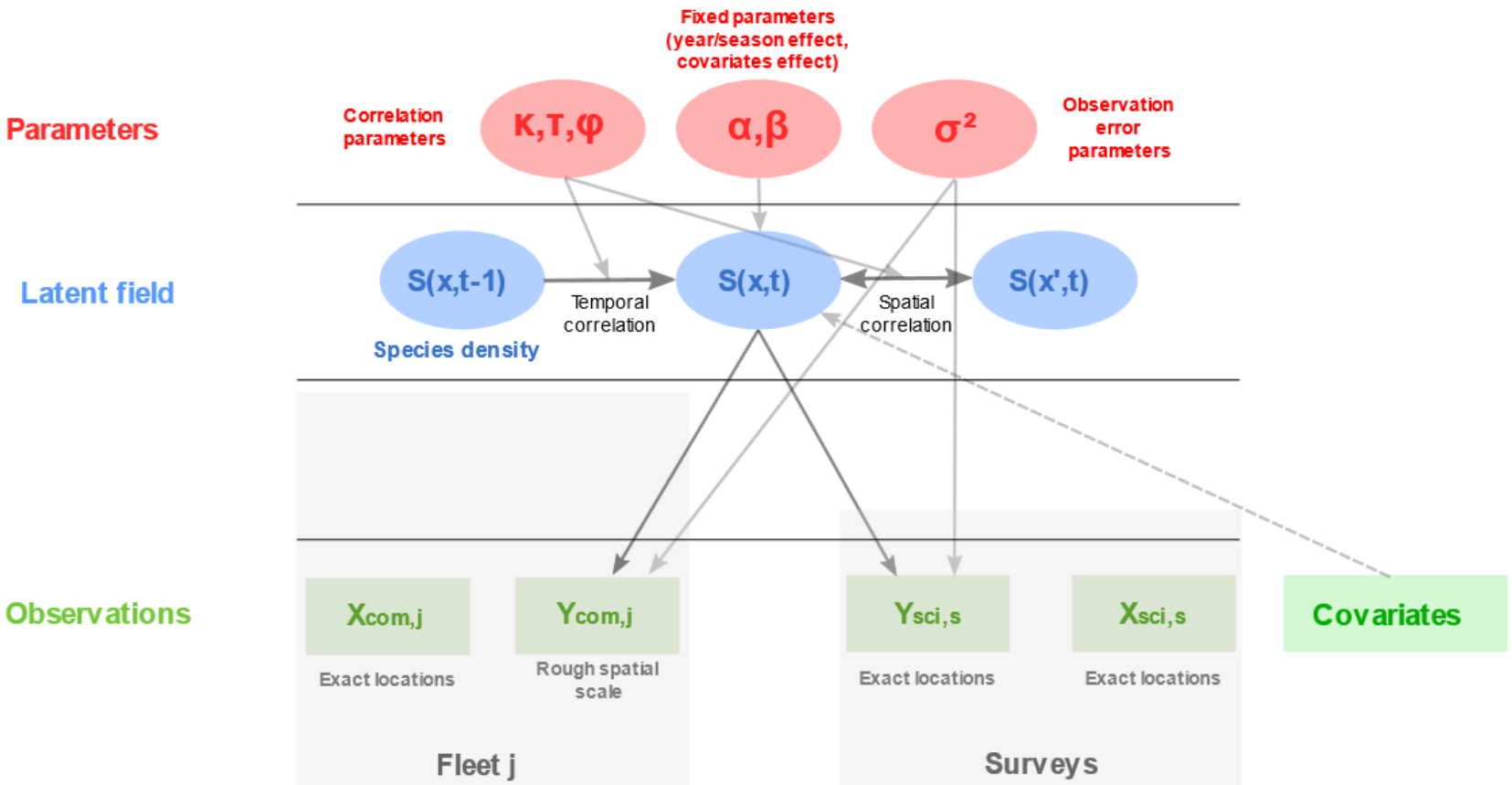
$$log(S(x,t)) = \mu_S(t) + \Gamma_S(x,t)^T \cdot \beta + \delta(x,t)$$
$$\delta \sim GF(0, \mathcal{C}(x,x';t,t'))$$



Joint likelihood: $[\boldsymbol{Y}, \delta|\theta] = ([\boldsymbol{X}_{com}|\theta, \delta] \cdot [\boldsymbol{Y}_{com}|\theta, \delta] \cdot [\boldsymbol{Y}_{sci}|\theta, \delta]) \cdot [\delta|\theta]$

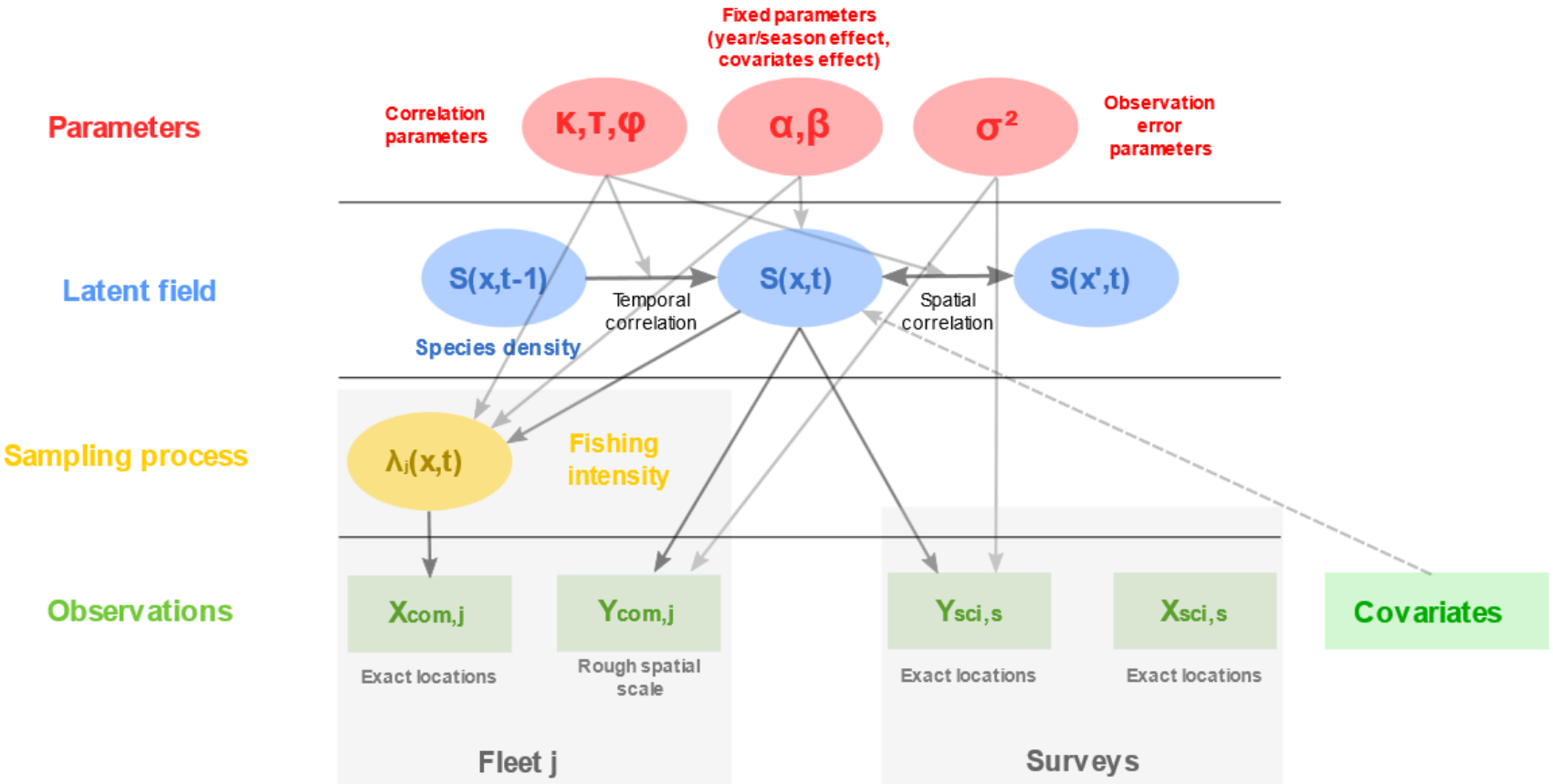


Joint likelihood: $[\boldsymbol{Y}, \delta|\theta] = ([\boldsymbol{X}_{com}|\theta, \delta] \cdot [\boldsymbol{Y}_{com}|\theta, \delta] \cdot [\boldsymbol{Y}_{sci}|\theta, \delta]) \cdot [\delta|\theta]$



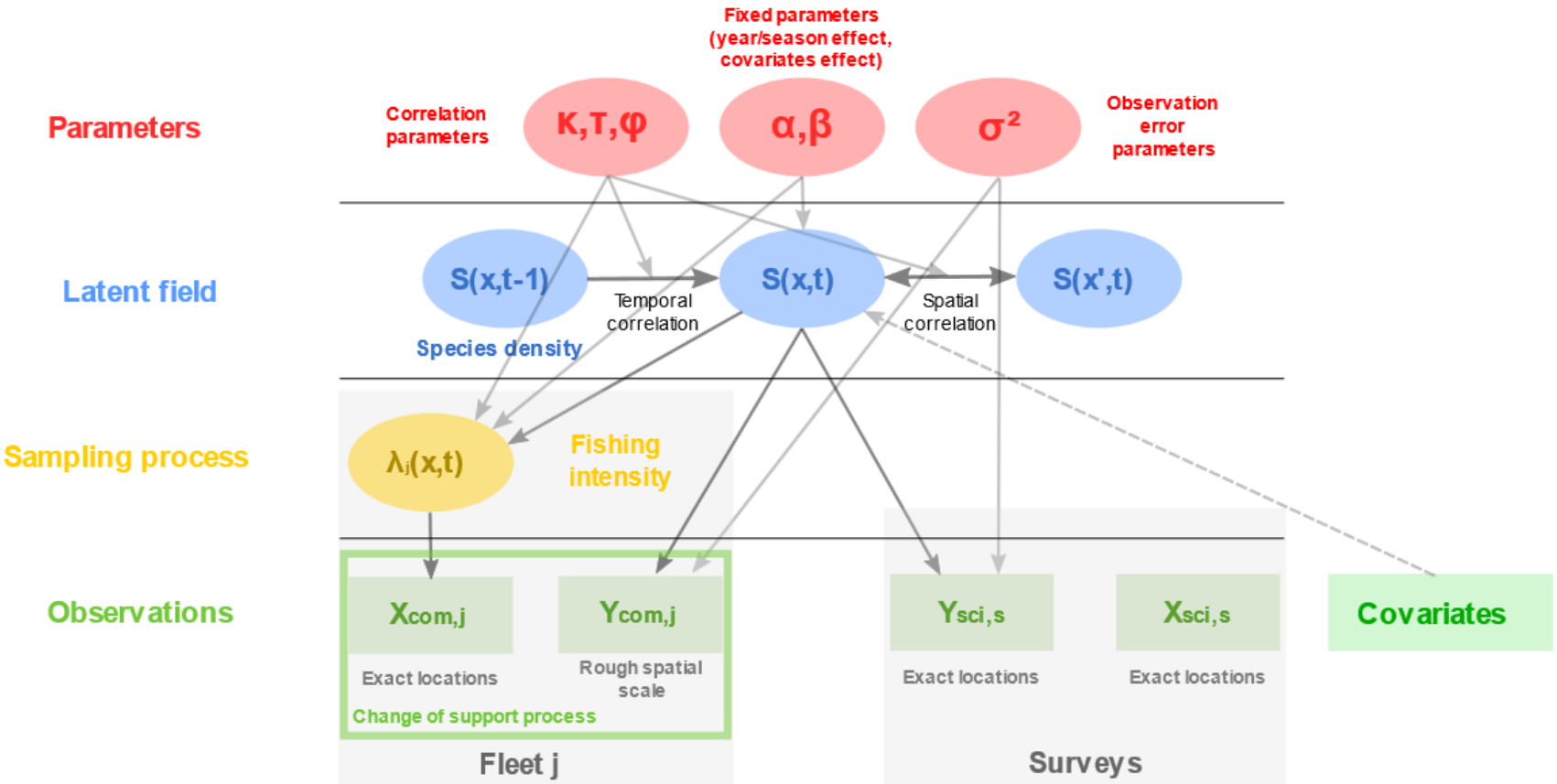
Joint likelihood: $[\boldsymbol{Y}, \boldsymbol{\delta}|\boldsymbol{\theta}] = ([\boldsymbol{X}_{com}|\boldsymbol{\theta}, \boldsymbol{\delta}] \cdot [\boldsymbol{Y}_{com}|\boldsymbol{\theta}, \boldsymbol{\delta}] \cdot [\boldsymbol{Y}_{sci}|\boldsymbol{\theta}, \boldsymbol{\delta}]) \cdot [\boldsymbol{\delta}|\boldsymbol{\theta}]$

$\boldsymbol{X}_{com} \sim \mathcal{IPP}(\lambda(x, t))$



Joint likelihood: $[\boldsymbol{Y}, \delta|\theta] = ([\boldsymbol{X}_{com}|\theta, \delta] \cdot [\boldsymbol{D}_{com}|\theta, \delta] \cdot [\boldsymbol{Y}_{sci}|\theta, \delta]) \cdot [\delta|\theta]$

$\boldsymbol{D}_{com} = \sum \boldsymbol{Y}_{com}$



How to assess the consistency between the datasets?

How to assess if the integrated model is consistent with the scientific model (=reference)?

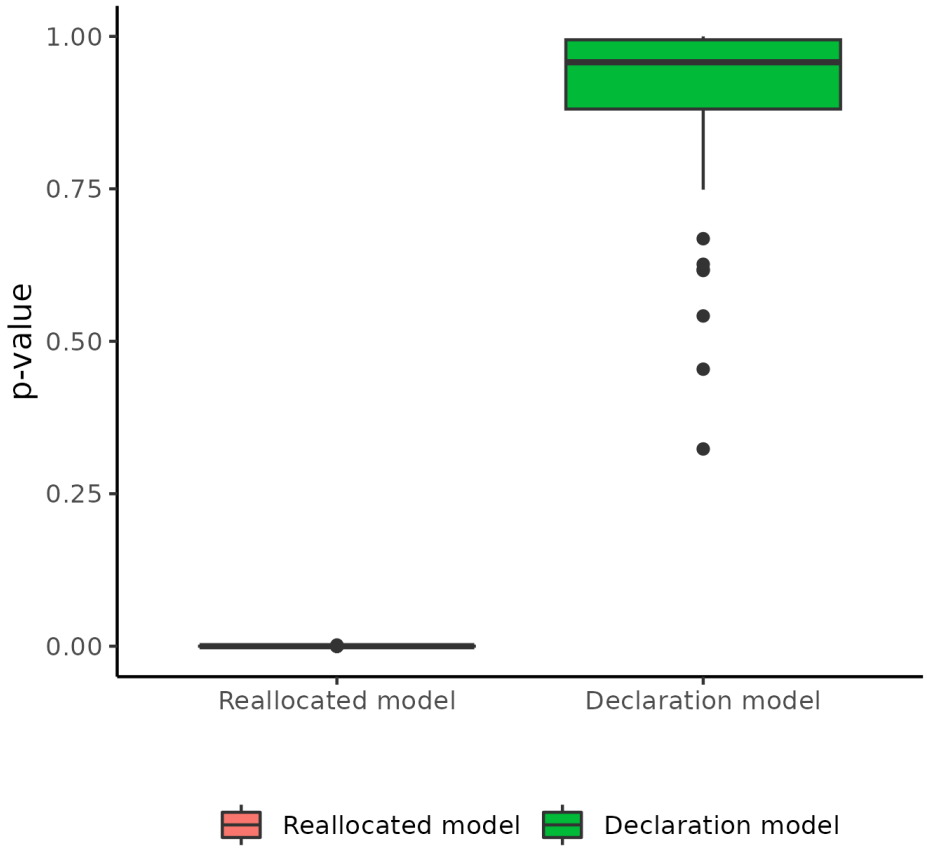
Likelihood ratio test

H0: There is no significant difference between both models

$$\frac{L_{sci}(\hat{\theta}_{sci})}{L_{sci}(\hat{\theta}_{int})} = 1$$

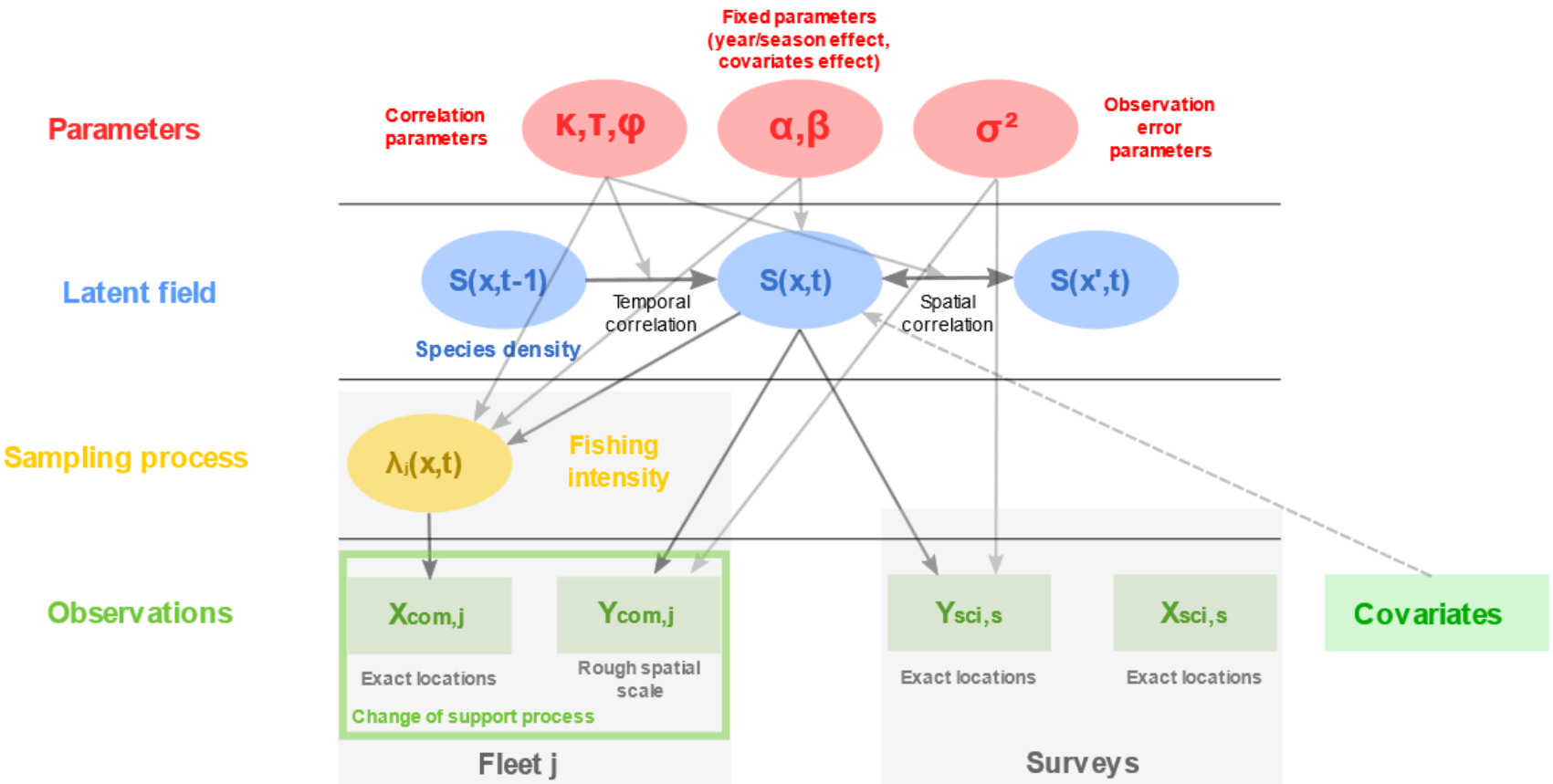
H1: The integrated model is inconsistent with the scientific model

$$\frac{L_{sci}(\hat{\theta}_{sci})}{L_{sci}(\hat{\theta}_{int})} \neq 1$$



- L_{sci} the likelihood of the scientific model
- $\hat{\theta}_{sci}$ the estimated parameters of the scientific model
- $\hat{\theta}_{int}$ the estimated parameters of the integrated model

But change of support faces strong convergence issues



At this point, the operational framework is built on reallocated declarations

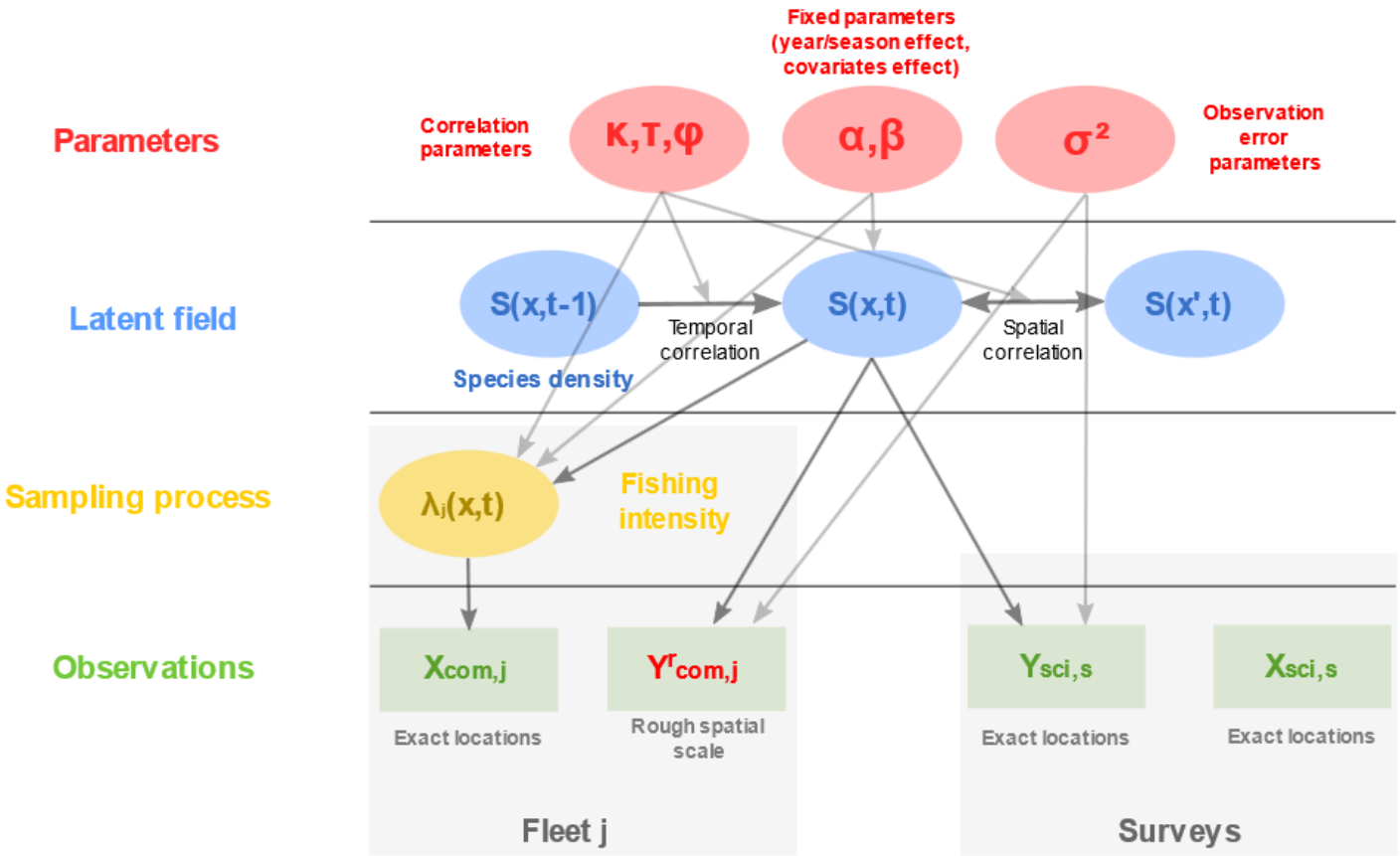
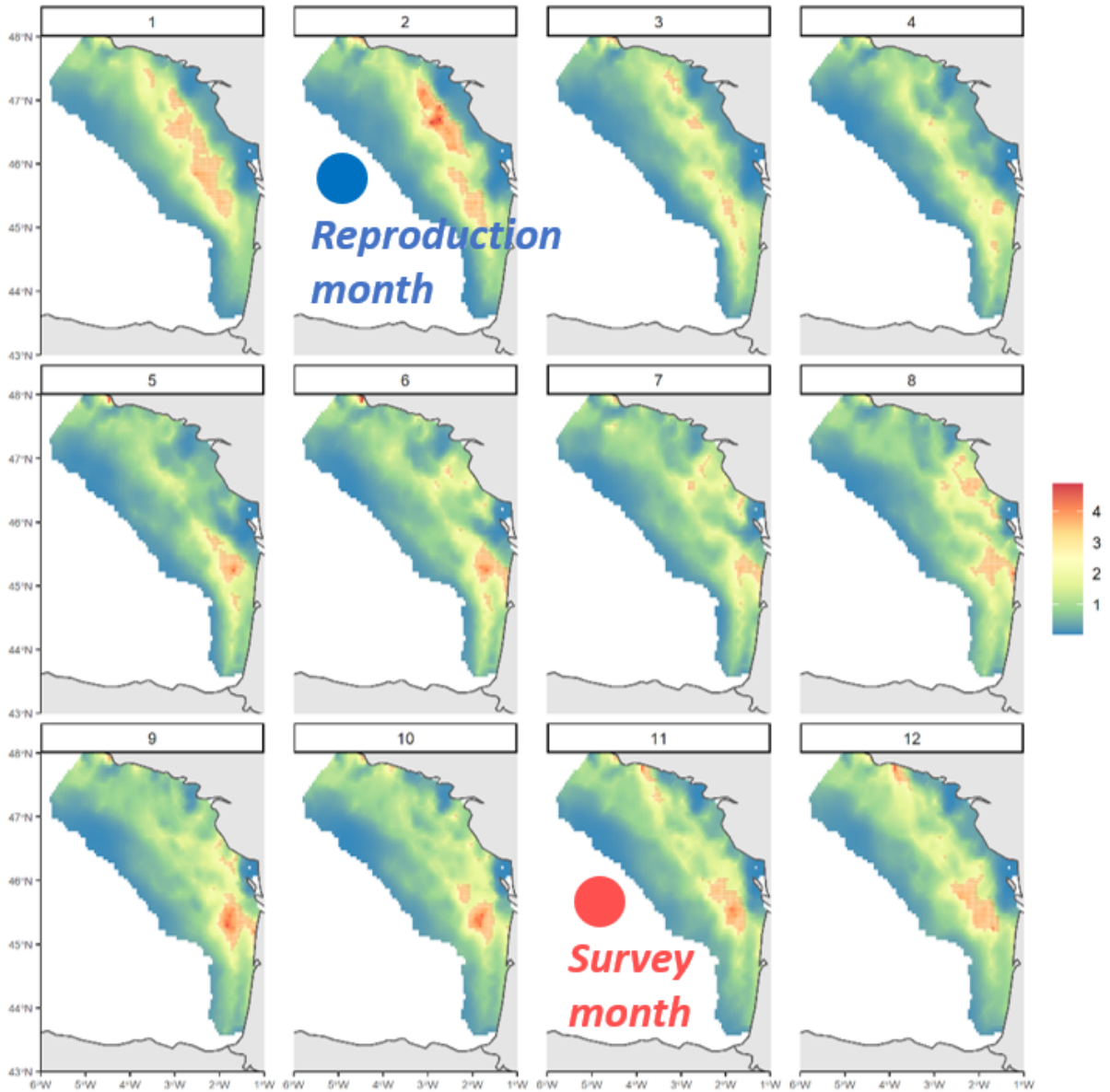


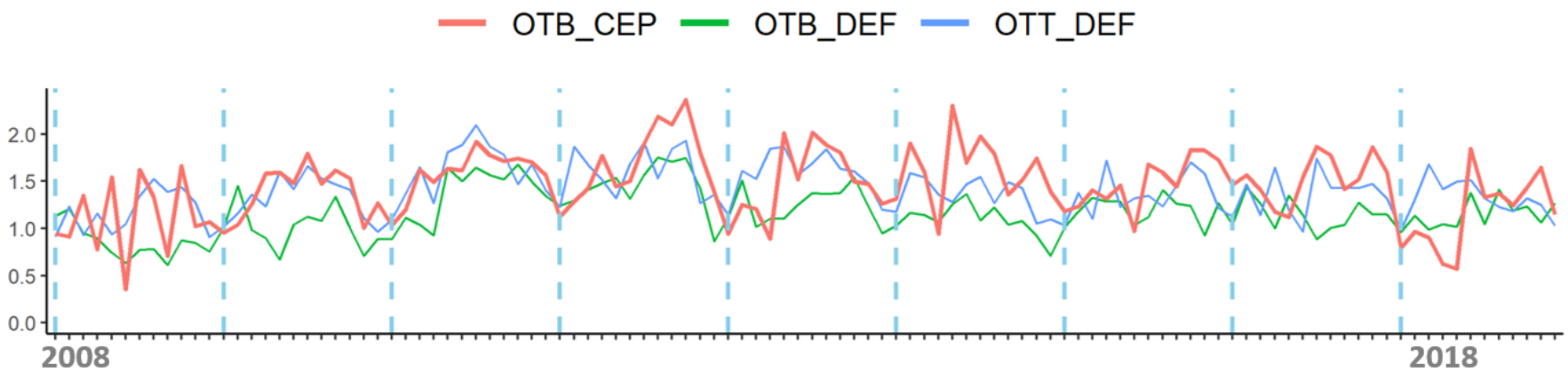
Table of Contents

Mapping species distribution at a fine temporal scale

Monthly maps of sole distribution
(average over 2008 - 2018)



Study the temporal variability of preferential sampling $b_j(t)$



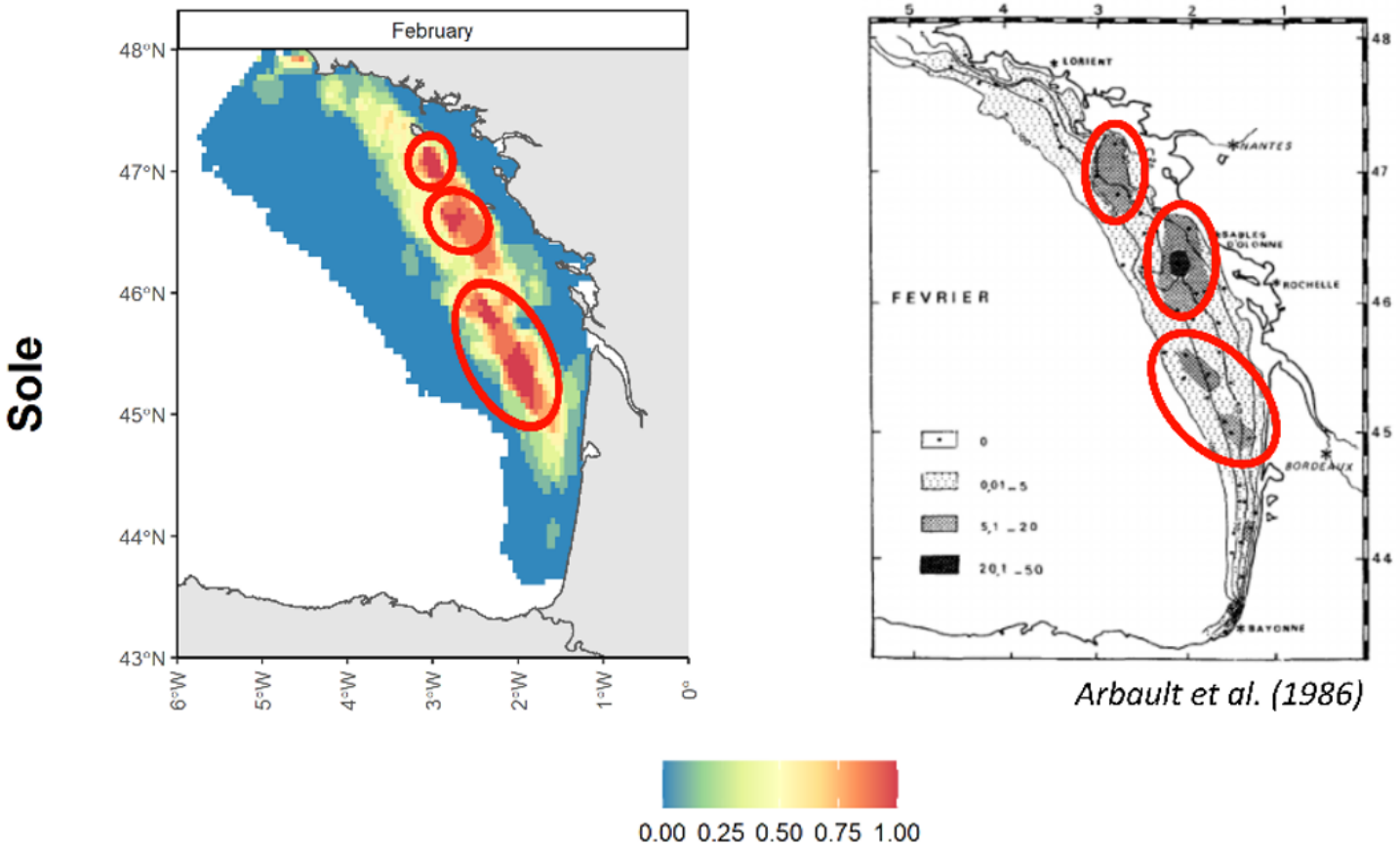
- Several fleets in inference each related to distinct sampling processes
- Several time series of preferential sampling parameters
- Variability between fleets + temporal variability

Identifying spawning areas

Based on indices of local aggregation

e.g. Getis and Ord index

- Significance of a cell x to be within an aggregation area
- How often is a cell within an aggregation area?



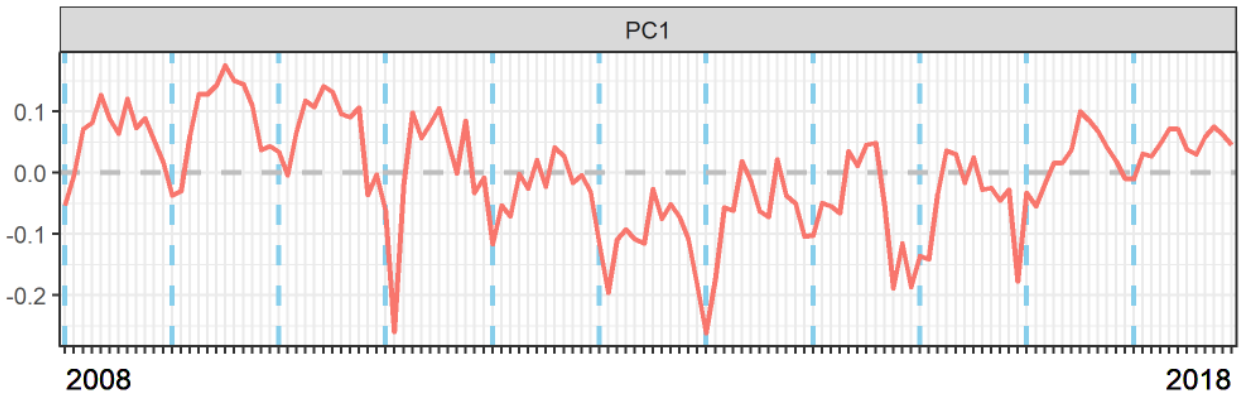
Arbault et al. (1986)

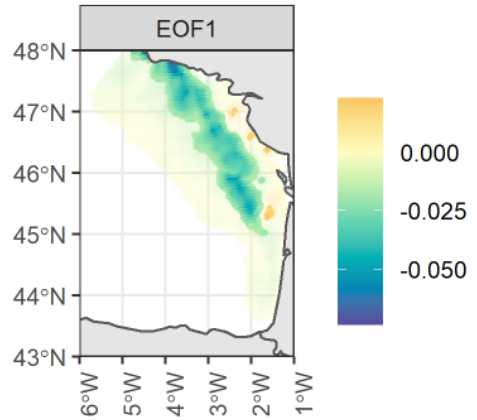
Synthesizing model spatial predictions

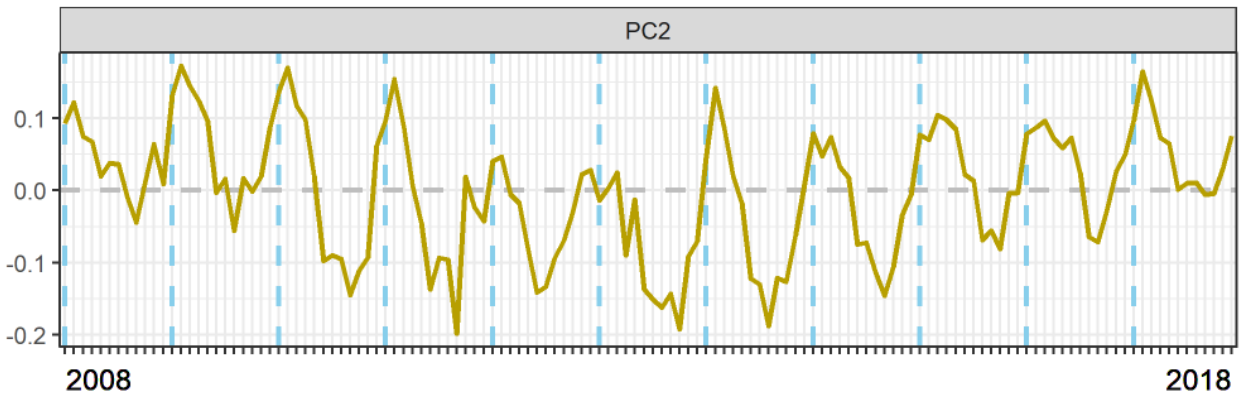
Empirical Orthogonal Functions

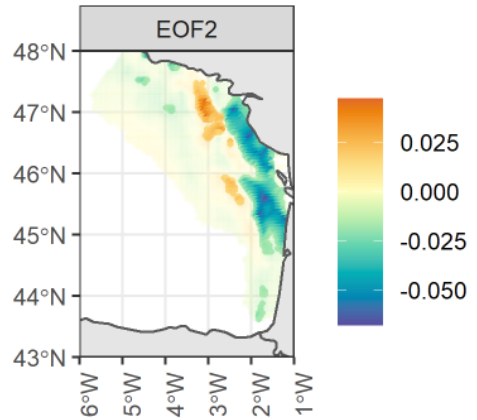
$$\Rightarrow \mathbf{S}_t = \sum_{k=1}^{K} \alpha_k(t)\mathbf{p}^k + \mathbf{n}_t$$

Minimize $\mathbf{n}_t$, orthogonality between $\mathbf{p}^k$







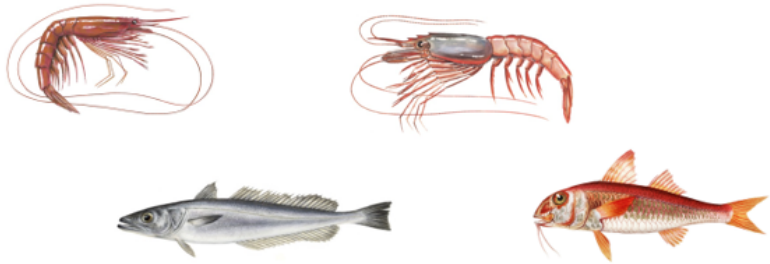


Operational applications

MACCO project

Mapping data-rich and data-poor species

Bay of Biscay



Med. Sea

CSTEP WG
Closure areas

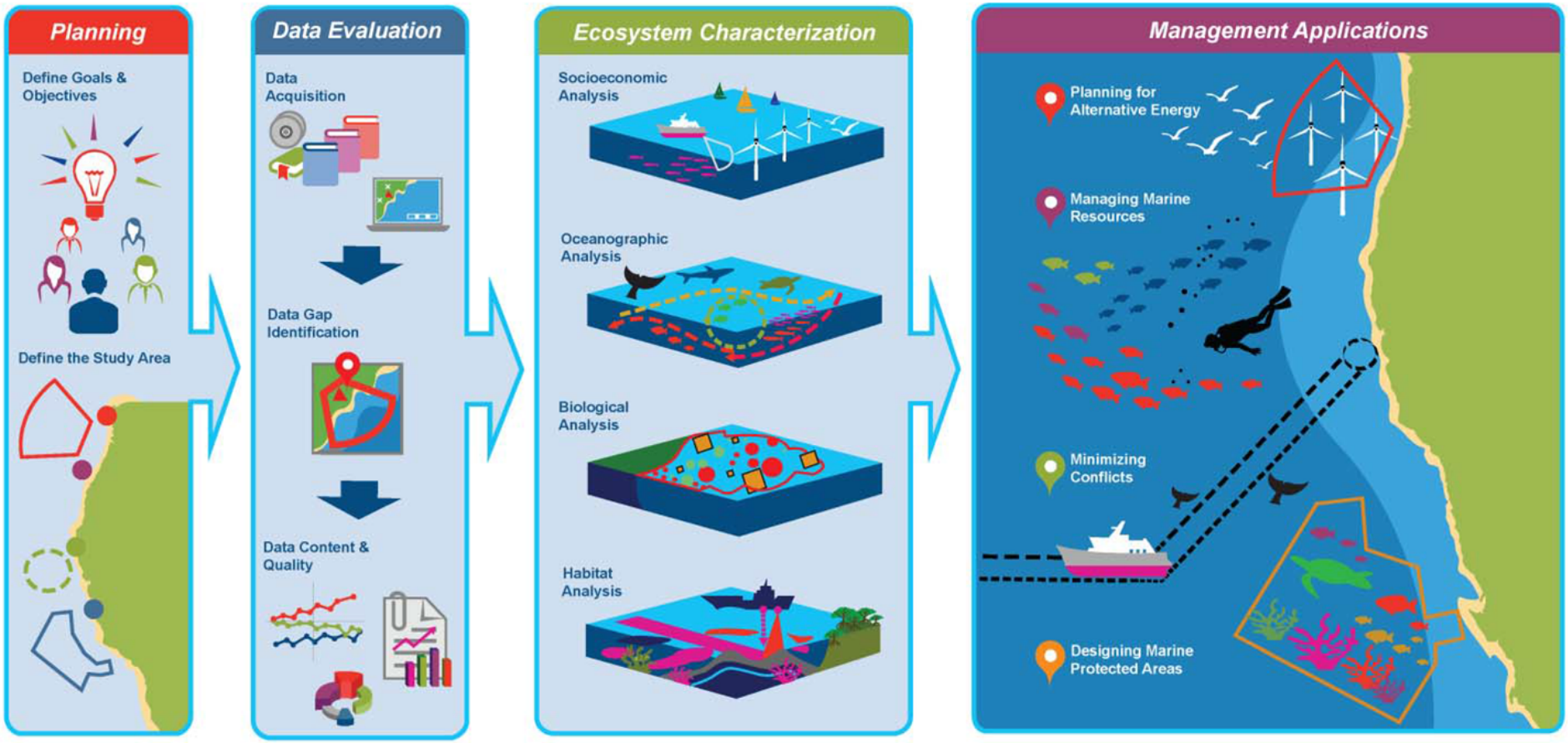
Use for marine spatial planning

Identification of
potential closure areas
+
Model parameterization
to assess these closure areas

ISIS-fish

MARXAN
conservation solutions

➠ use these outputs for **Marine Spatial Planning**

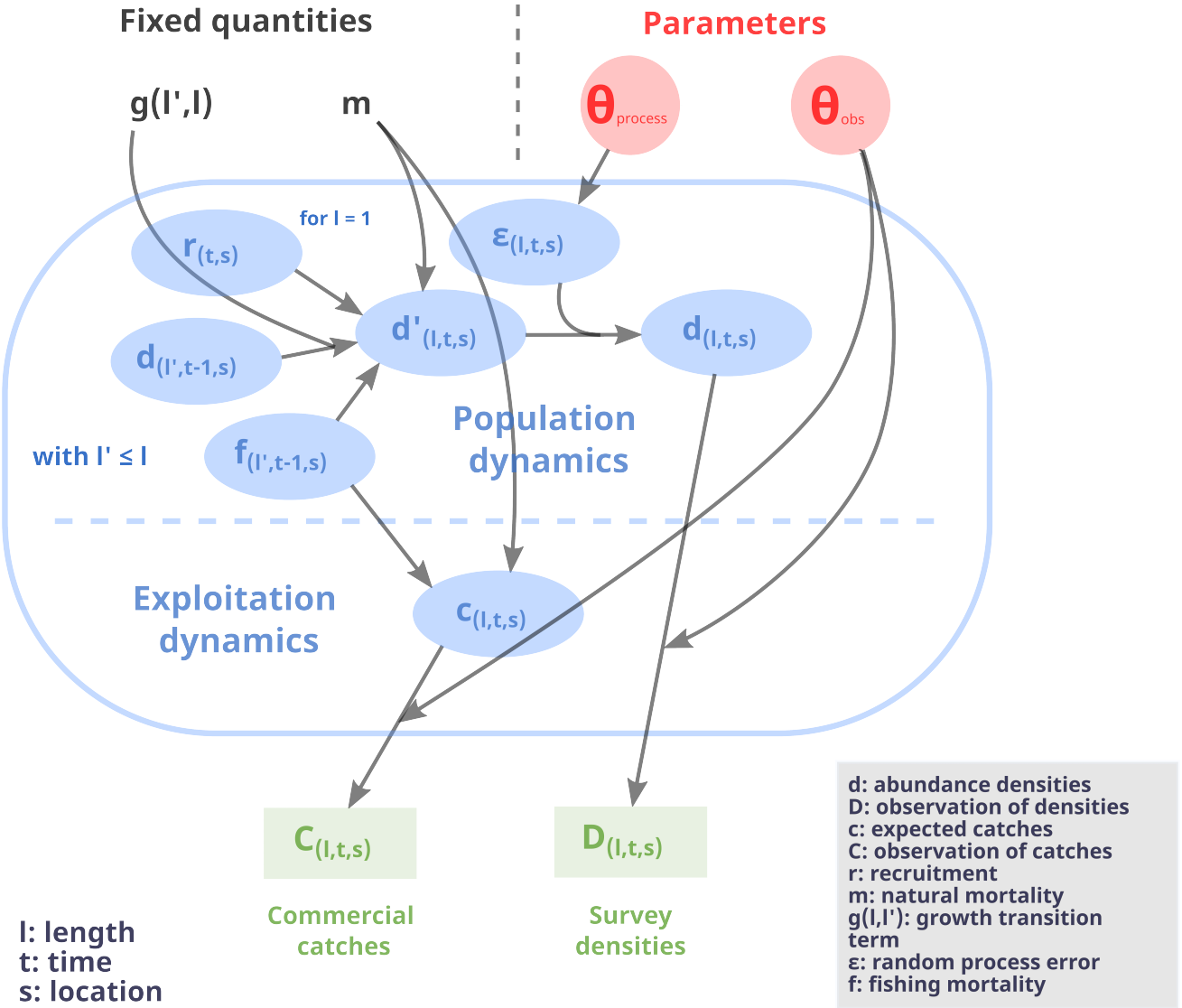


Including mechanistic processes in the latent field

Population dynamics model

$$\boldsymbol{d}_{(t,s)} = g\left(\boldsymbol{d}_{(t-1,s)}\right) \cdot e^{\boldsymbol{\varepsilon}_{(t,s)}}$$

with:
$\boldsymbol{d}_{(t,s)} = \{d_{(1,t,s)}, \ldots, d_{(L,t,s)}\}$,
$\boldsymbol{\varepsilon}_{(t,s)} = \{\varepsilon_{(1,t,s)}, \ldots, \varepsilon_{(L,t,s)}\}$,
$l \in \{1, \ldots, L\}$ the size classes,
$t \in \{1, \ldots, T\}$ the time steps,
$s \in \mathbb{R}^2$ the locations,
$g : \mathbb{R}^L \to \mathbb{R}^L$ the function modeling population dynamics.



Olmos, M., Cao, J., Thorson, J. T., Punt, A. E., Monnahan, C. C., Alglave, B., & Szuwalski, C. (2023). A step towards the integration of spatial dynamics in population dynamics models: Eastern Bering Sea snow crab as a case study. Ecological Modelling, 485, 110484.

Table of Contents

Key points

How to infer spatio-temporal processes from these data?
How to combine all these data sources?

Several methodological issues:

- Non-standardized/preferential sampling ✔
- Combine massive data sources with standardized data sources ✔
- Use aggregated data to infer fine-scale processes (change of support) ∼

- Computation time !!!
- Alternative dimension reduction methods
- Investigate other learning methods ?
  (*e.g.* deep, auto-encoders)

Key points

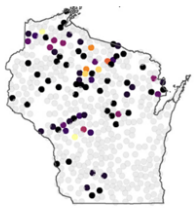
How to infer spatio-temporal processes from these data?
How to combine all these data sources?

Several methodological issues:

- Non-standardized/preferential sampling ✔
- Combine massive data sources with standardized data sources ✔
- Use aggregated data to infer fine-scale processes (change of support) ∼
- Computation time !!!
- Alternative dimension reduction methods
- Investigate other learning methods ?
  (*e.g.* deep, auto-encoders)

➠ Overall, methodological developments that are relevant for other fields of application

Terrestrial ecology



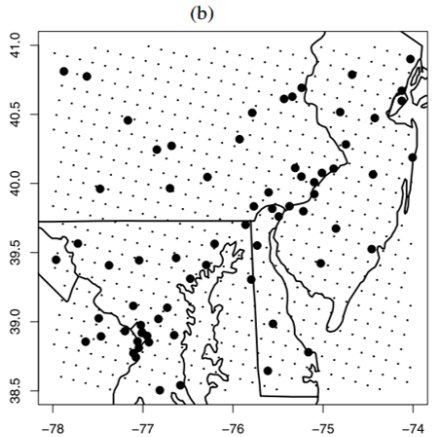
Camera submodel
$Y_{1i} \sim f(\lambda_i)$

Harvest submodel
$Y_{2i} \sim f(\lambda_i)$

Gilbert et al. (2021)

Harvest records
(aggregated data)
x
Camera-trap data
(exact locations data)

Air pollution



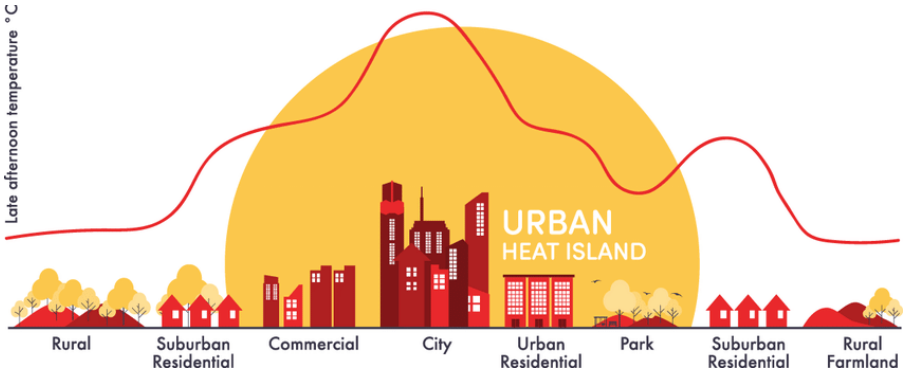
Berrocal et al. (2010)

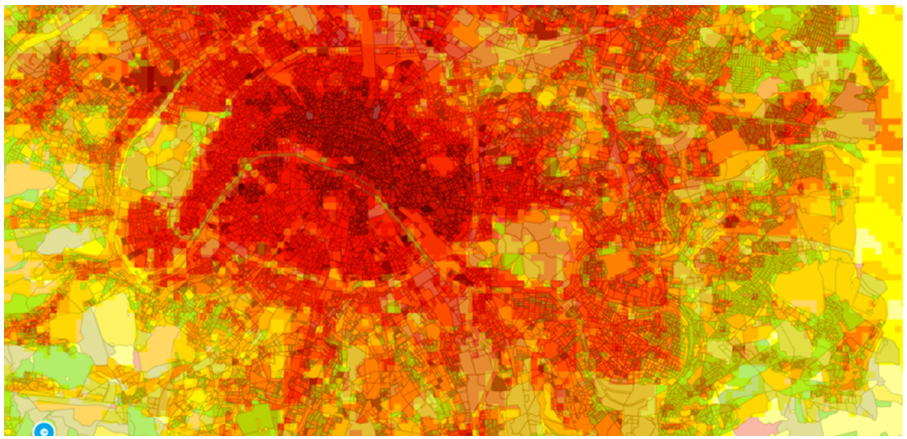
Outputs of numerical model
(massive rough data)
x
Monitoring networks data
(sparse high quality data)

And others: **geography, epidemiology, climate science...**

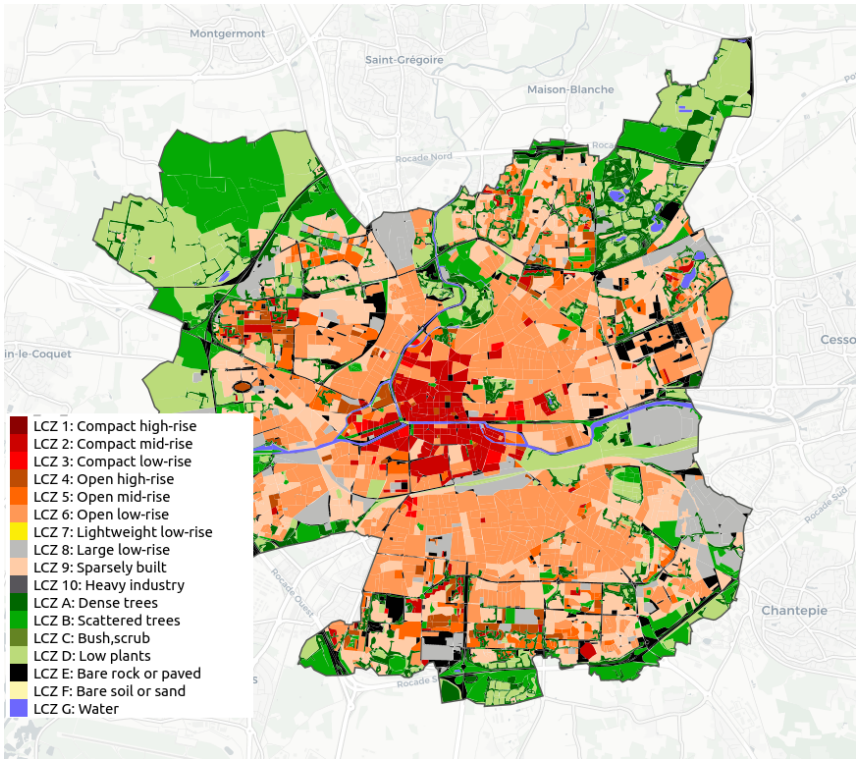
Research focus

Predicting Urban Heat Island (UHI)





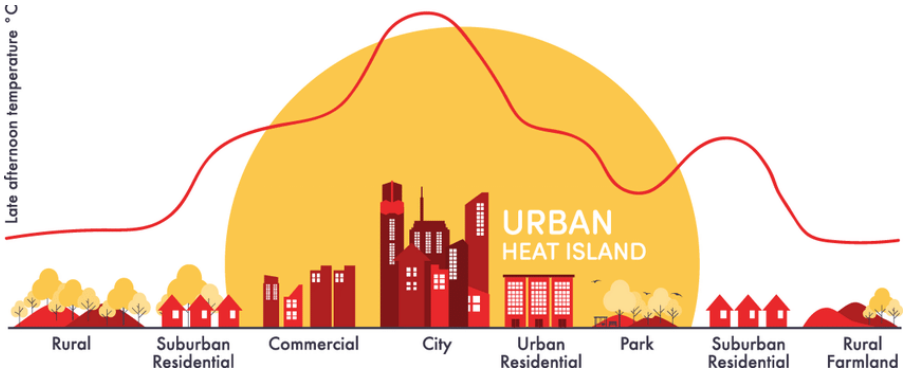
Mapping and analysing Local Climate Zones (LCZ)

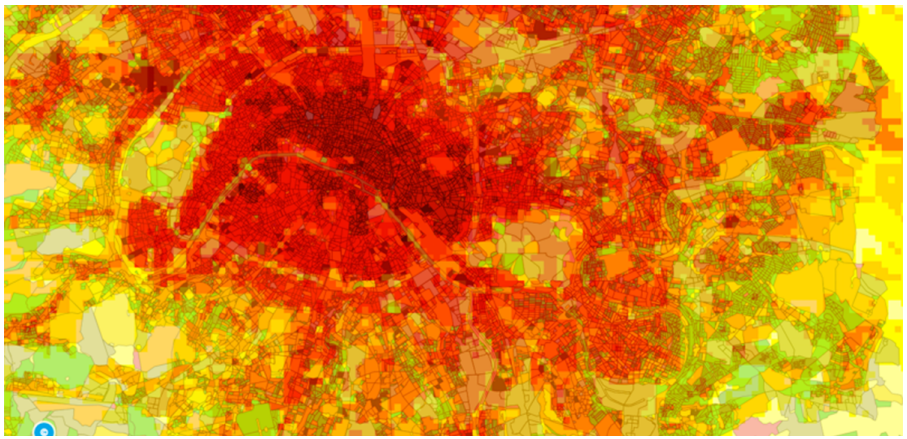


Making projections (long-term/short-term) to help decision makers to plan urban areas in a context of ecological transition and global warming.

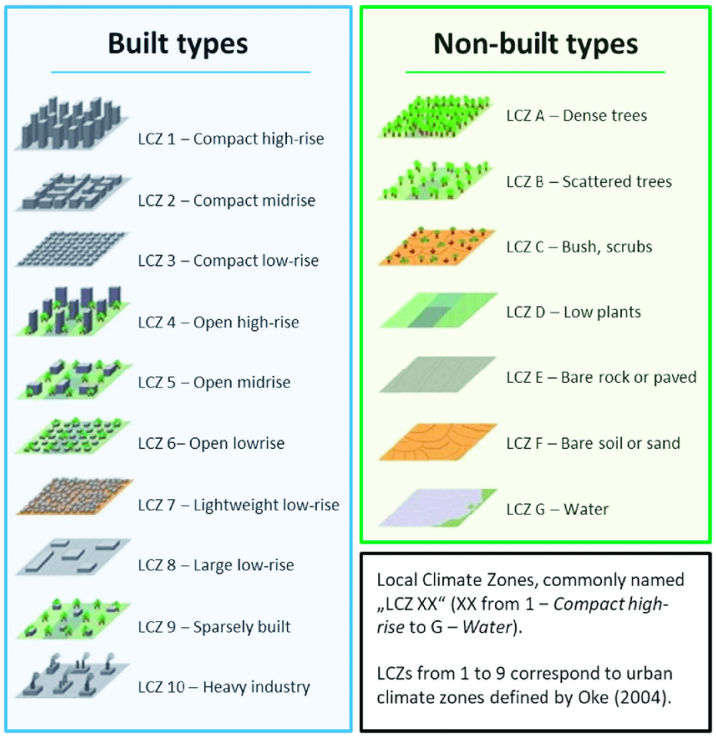
Research focus

Predicting Urban Heat Island (UHI)





Mapping and analysing Local Climate Zones (LCZ)



Making projections (long-term/short-term) to help decision makers to plan urban areas in a context of ecological transition and global warming.

Thank you for your attention!

Punctual observation model (Y_{ai})

When making the sum or the product over $i|x_i \in \mathcal{R}_a$, we simply denote $\prod_i$ and $\sum_i$

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

Y_{ai} and D_a are supposed conditional on $\boldsymbol{S}$ and $\boldsymbol{X}$.

$$\mathrm{P}\left(Y_{ai} = y_{ai}\right) = \begin{cases} p_{ai} & \text{if } y_{ai} = 0 \\ (1 - p_{ai}) \cdot \mathrm{L}\left(y_{ai}, \mu_{ai} = \frac{S(x_{ai})}{(1-p_{ai})}, \sigma^2\right) & \text{if } y_{ai} > 0 \end{cases}$$

$$p_{ai} = \exp(-e^{\xi}.S(x_{ai}))$$

Declaration model $(D_a = \sum_i Y_{ai})$

$$P(D_a = 0) = \prod_i P(Y_{ai} = 0) = \exp\left\{-\sum_i e^{\xi}.S(x_{ai})\right\} = \pi_a$$

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

Punctual observation model (Y_{ai})

When making the sum or the product over $i|x_i \in \mathcal{R}_a$, we simply denote $\prod_i$ and $\sum_i$

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

Y_{ai} and D_a are supposed conditional on $\boldsymbol{S}$ and $\boldsymbol{X}$.

$$\mathrm{P}\left(Y_{ai} = y_{ai}\right) = \begin{cases} p_{ai} & \text{if } y_{ai} = 0 \\ (1 - p_{ai}) \cdot \mathrm{L}\left(y_{ai}, \mu_{ai} = \frac{S(x_{ai})}{(1-p_{ai})}, \sigma^2\right) & \text{if } y_{ai} > 0 \end{cases}$$

$$p_{ai} = \exp(-e^{\xi}.S(x_{ai}))$$

Declaration model $(D_a = \sum_i Y_{ai})$

$$P(D_a = 0) = \prod_i P(Y_{ai} = 0) = \exp\left\{-\sum_i e^{\xi}.S(x_{ai})\right\} = \pi_a$$

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

Compute the moments of $D_a|d_a > 0$

$$E(D_a|d_a > 0) = \frac{\sum_i S(x_{ai})}{1 - \pi_a}$$

$$Var(D_a|d_a > 0) = \frac{\sum_i Var(Y_{ai})}{1 - \pi_a} - \frac{\pi_a}{(1 - \pi_a)^2} E(D_a)^2$$

$$Var(Y_{ai}) = \frac{S(x_{ai})^2}{1 - p_{ai}}(e^{\sigma^2} - (1 - p_{ai}))$$

Consider $D_a|d_a > 0$ is Lognormal too

$$\mathrm{P}\left(D_a = d_a|d_a > 0\right) =$$

$$\mathrm{L}\left(d_a, \mu_a = E(D_a|d_a > 0), \sigma_a^2 = ln(\frac{Var(D_a|d_a > 0)}{E(D_a|d_a > 0)^2} + 1)\right)$$

Compute the moments of $D_a|d_a > 0$

$$E(D_a|d_a > 0) = \frac{\sum_i S(x_{ai})}{1 - \pi_a}$$

$$Var(D_a|d_a > 0) = \frac{\sum_i Var(Y_{ai})}{1 - \pi_a} - \frac{\pi_a}{(1 - \pi_a)^2} E(D_a)^2$$

$$Var(Y_{ai}) = \frac{S(x_{ai})^2}{1 - p_{ai}}(e^{\sigma^2} - (1 - p_{ai}))$$

Consider $D_a|d_a > 0$ is Lognormal too

$$\mathrm{P}\left(D_a = d_a|d_a > 0\right) =$$

$$\mathrm{L}\left(d_a, \mu_a = E(D_a|d_a > 0), \sigma_a^2 = ln(\frac{Var(D_a|d_a > 0)}{E(D_a|d_a > 0)^2} + 1)\right)$$

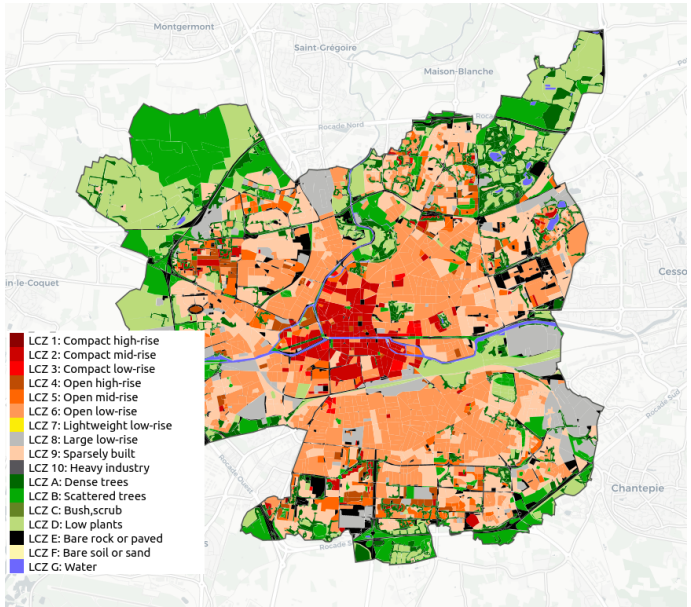
Few words about my current (new) activities

Overall context

➠ GIS to manage urban space in a context of global warming

- map climatic and environmental risks
- identify sensitive areas

in order to help decision making to plan urban areas



e.g. **GeoClimate** software

➠ platform integrating heterogeneous and massive datasets (buildings, roads, vegetation)

➠ produce geographic indicators to characterize the specific features of a given region (in particular map **Local Climate Zones**).

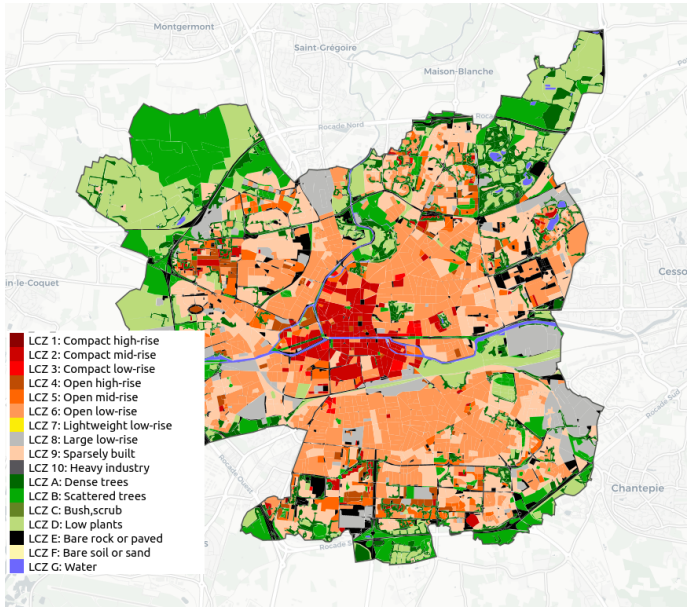
Few words about my current (new) activities

Overall context

➠ GIS to manage urban space in a context of global warming

- map climatic and environmental risks
- identify sensitive areas

in order to help decision making to plan urban areas



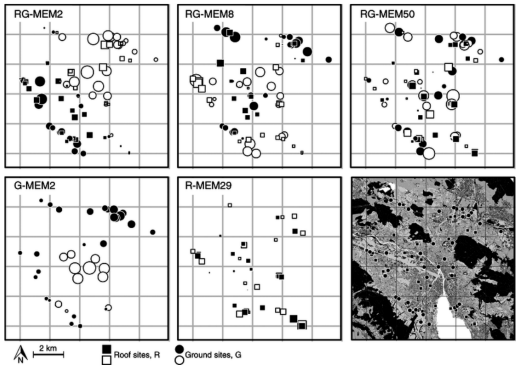
e.g. **GeoClimate** software
➠ platform integrating heterogeneous and massive datasets (buildings, roads, vegetation)
➠ produce geographic indicators to characterize the specific features of a given region (in particular map **Local Climate Zones**).

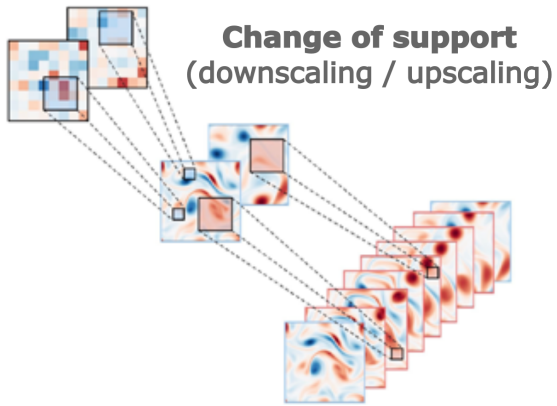
Research focus

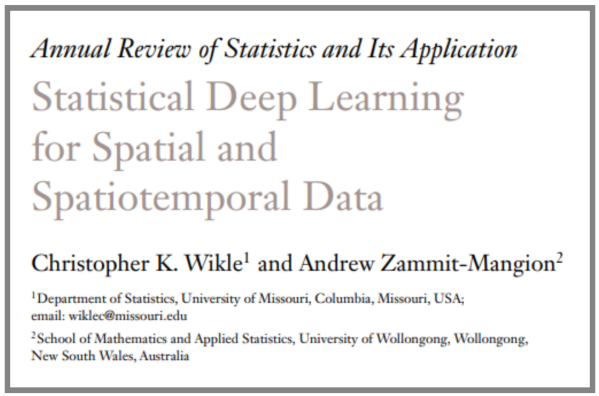
- **Clustering analysis** on GeoClimate outputs to identify city profiles (and possibly identify the ones where the risks related to climate change are higher)
  ➠ Spatial clustering on GeoClimate outputs (Master Thesis)

- Relating LCZ with other variables (economic, ecological indicators)
  ➠ **Change of support** issues because the datasets do not have the same spatial resolution

- Developing methods to predict urban heat islands (linear model is the approach in development)
  ➠ Towards **spatio-temporal deep-learning** methods






Annual Review of Statistics and Its Application

Statistical Deep Learning for Spatial and Spatiotemporal Data

Christopher K. Wikle[1] and Andrew Zammit-Mangion[2]

[1]Department of Statistics, University of Missouri, Columbia, Missouri, USA; email: wiklec@missouri.edu

[2]School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, New South Wales, Australia

Diggle, P. J., Menezes, R., and Su, T. 2010. Geostatistical inference under preferential sampling. Journal of the Royal Statistical Society: Series C (Applied Statistics), 59: 191–232.