Modelling higher complex Species Distribution Models

David V. Conesa Guillén

VAlencia BAyesian Research group

Universitat de València

The real world is not easy (1)

But in real life things are even more complicated:

- Sometimes data do not come from independent data because the sampling is not random but preferential. Like fishermen who go fishing were they think there is fish, so clever option (Pennino et al., 2019).
- Sometimes there is no available information in those places we want to make predictions: misalignment (Barber et al., 2016).
- In other situations, the spatial dependence structure is NOT the same throughout the time considered and different spatio-temporal structures can be considered (Paradinas et al., 2017).
- The above presented approach does not take into account that spatial data are often multivariate. Consequently, a multivariate geostatistical approach should be used (Barber et al., 2019; Barber et al., 2021).

The real world is not easy (2)

- When geographical elements or physical barriers are present, the spatial dependence structure could NOT be the same throughout the domain: non-stationarity (Martínez-Minaya et al., 2019; Cendoya et al., 2022).
- Sometimes information is presented aggregated and areal spatial data analysis are required (Carmezim et al., 2022; Sarzo et al., 2023).
- In some occassions, we can have information from varios sources and want to join them in one unique modelling (Figueira et al., 2024a; Figueira et al., 2025).
- An important situation is how to deal with only-presence data (Martino et al., 2021, Fernández et al., 2025).
- See Martínez-Minaya et al. (2018) for a review of some of these issues.
- See also Figueira et al. (2024b) for a Shiny app that allows to use most of the procedures here mentioned in an easy way.

1 Preferential sampling

Dealing with preferential sampling.

- Geoestatistical models usually assume that sampling locations and the process being modelled are stochastically independent.
- But, sampling locations are deliberately concentrated in areas where the abundance of the variable of interest is known or expected to be high

Implementation of Preferential sampling as a marked point pattern.

- The sampling design process (point pattern) depends on the unknown target spatial stationary Gaussian process. That is, the observed locations $(s_1, \ldots, s_n)$ come from a log-Gaussian Cox process with intensity

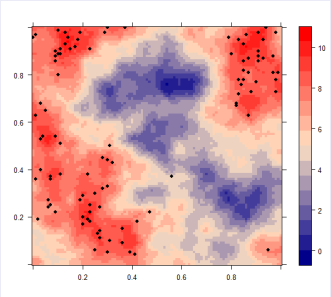
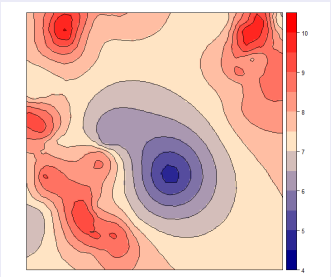
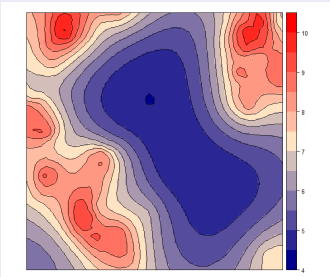
$$\Lambda_i = \exp\{\alpha_1 + \beta w_i\},$$

where β is the shared parameter between both likelihoods that express the preferential sampling.

- The species characteristic (usually abundance) y_i is assumed to follow an exponential family distribution whose mean is related with the spatial term:

$$\begin{aligned} y_i &\sim F(\mu_i, \gamma^2) \\ g(\mu_i) &= \alpha_2 + w_i \end{aligned}$$

Simulated example:

- Left: Simulated Gaussian field and point pattern representing the sampling locations.
- Middle: Posterior predictive mean maps of the simulated abundance process without preferential sampling correction ...
- Right: ... and with the preferential sampling correction.

Red shrimp near Alicante: the preferentiality appears in the covariate shared

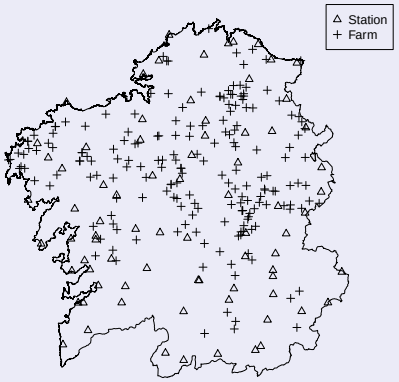
- Left: Abundance data at the sampling locations,
- Middle: posterior predictive mean of the red shrimp distribution without the preferential bathymetric correction ...
- Right: ... and with the preferential bathymetric correction,
- the non-preferential model prediction seems driven by the spatial effect while the preferential model map corrects the bathymetric effect providing a more natural pattern of the red shrimp distribution.

2 Misalingment

Dealing with misalignment

- It appears when measurement values of the covariates are not known at the observed locations nor at those locations where we are going to make predictions.

Example of misalignment: the 67 official weather stations in Galicia do not coincide with the farms where data were observed



Dealing with misalignment (2)

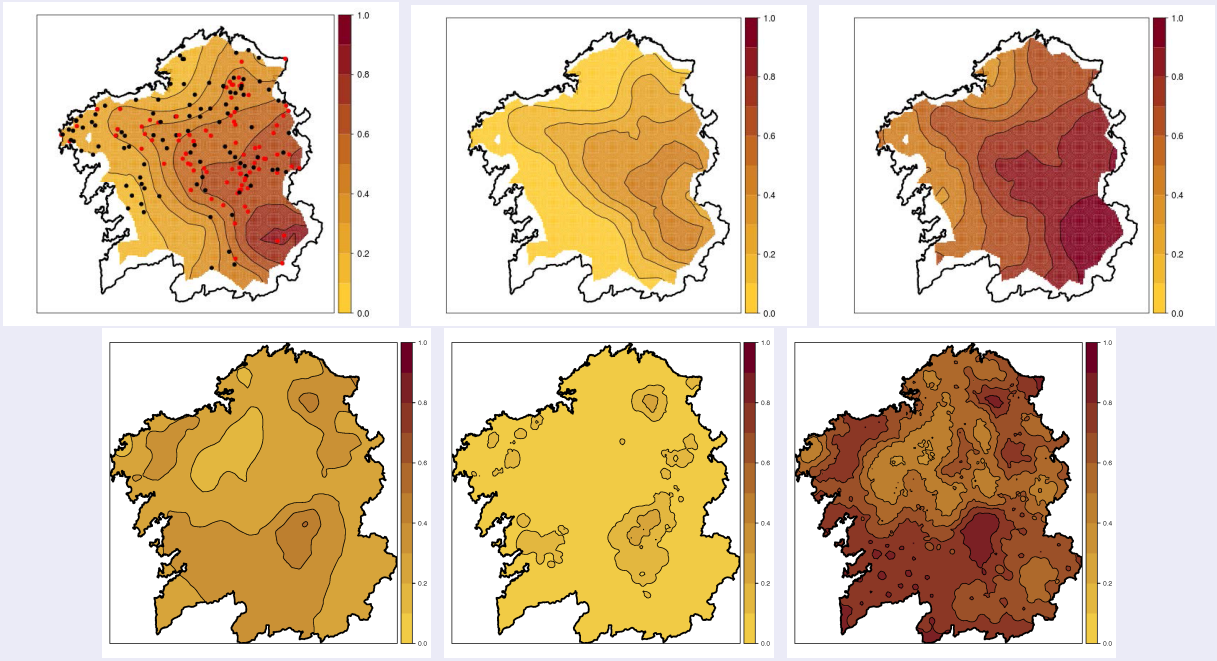
- A spatial geostatistical model is specified for the covariate, and it is estimated jointly with the species distribution models in a Bayesian context.
- The joint model is specified as follows

$$
\begin{aligned}
y_i &\sim F(\mu_i) \\
g(\mu_i) &= \beta_0 + \beta_1 \phi_i + w_i \\
w &\sim N(0, Q(\kappa, \tau)) \\
\\
x_i &\overset{iid}{\sim} N(\phi_i, \sigma_x^2) \\
\phi &\sim N(0, Q(\gamma, \delta))
\end{aligned}
$$

  - μ_i is the mean of the observed variable of interest at site s_i;
  - x_i is the covariate of interest whose spatial distribution is specified through its mean (a realization of the Matérn Gaussian process ϕ depending on the parameters γ and δ), and variance σ_x^2, which is introduced to express any possible measurement error; and
  - w is the spatial process for the response.

Analysing the Fasciolosis in Galicia (Barber et al., 2016)

Posterior mean of the probability of occurrence (left) and the first (center) and third (right) quantiles. Red points are presences and black absences. Above without taking into account misalignment, below taking it into account.



3 Coregionalised approach for multivariate SDMs

Coregionalised approach for multivariate SDMs

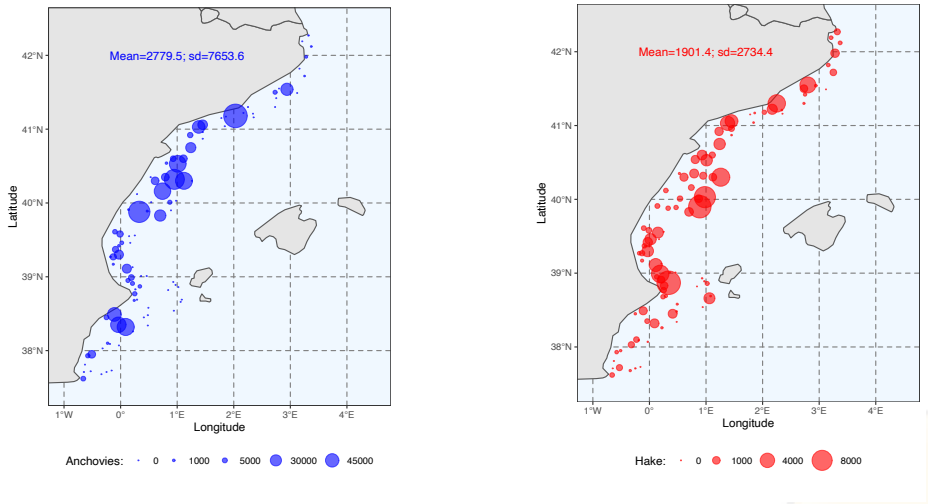
- SDMs are usually analized as univariate problems, not taking into account that spatial data are often multivariate.
- In order to explain the possible presence/absence or the abundance of a species, we must include in the modelling the biotic relationships with other species:
  - competition,
  - predator-prey,
  - parasitism, or
  - mutualism.
- Despite its enormous relevance, these factors are usually ignored in most classical SDM-based studies.
- Better than include these relationships via covariates, a great idea here is to use coregionalised models.

Coregionalised approach for multivariate SDMs (2)

- Information about the distribution of marine species:
  - fishery-independent data (scientific surveys at sea) and
  - fishery-dependent data (collection and sampling by observers in commercial vessels), preferential sampling.
- Scientific survey data are considered to be of higher quality (sampling designed to ensure unbiased estimations).
- But surveys may produce biased estimates:
  - for species with preferential habitats that are in strata only partially included in the survey sampling design,
  - or just for catchability issues.
- Again a good idea is to use multivariate geostatistical coregionalised models.
- Species well represented in the sampling could be used to infer the spatial behaviour of those correlated but poorly represented.

Describing interaction between anchovies and hake (Barber et al., 2021)

Application in fishery Ecology: a prey-predator example between anchovies and hake, one of its predator species, both collected during trawling surveys in the Mediterranean sea.



Data from EU-funded survey project MEDIterranean Trawl Survey (MEDITS): map shows the abundance at each sampling location of anchovies (**left**) and hake (**right**).

Coregionalised model for anchovies and hake

Coregionalised model that tries to reflect the predator-prey relationship between anchovies and hake species

$$\begin{aligned} \log(\text{Hake}) &= \beta_{10} + \beta_{11}\,\text{Bathymetry} + W_1 + \epsilon_1 \\ \log(\text{Anchovy}) &= \beta_{20} + \beta_{21}\,\text{Bathymetry} + \alpha W_1 + W_2 + \epsilon_2 \end{aligned}$$

where

- α represents the association between both species;
- W_1 and W_2 are the spatially structured effects;
- linear relationship between species and bathymetry (non-linear could also have been fitted by incorporating second-order random walk latent models);
- non-informative Gaussian priors for all fixed effects; and PC-priors to describe prior knowledge of hyperparameters of the spatial terms;
- no interacting univariate models $\alpha = 0$ were also analysed.
- implemented in INLA by sharing the spatial term in both likelihoods.

Coregionalised model for anchovies and hake: results

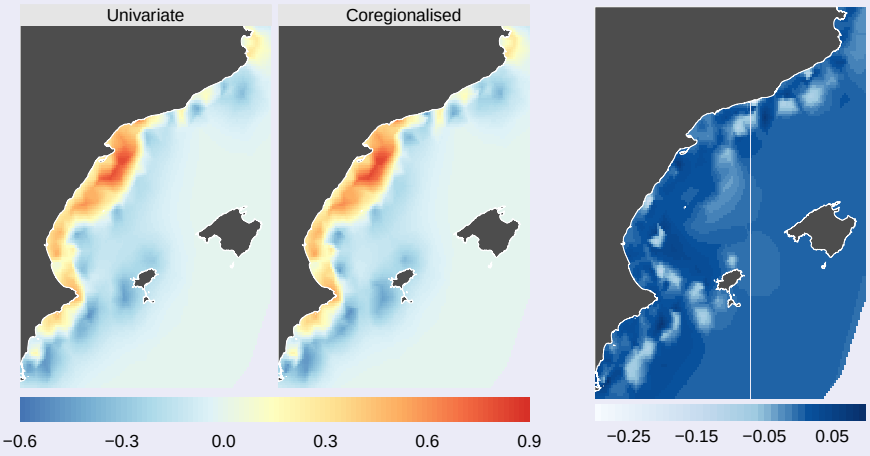
Coregionalised model that tries to reflect the predator-prey relationship between anchovies and hake species

	Mean	sd	$q_{0,025}$	$q_{0,5}$	$q_{0,975}$
β_{10}	6.5258	0.2759	5.9600	6.5343	7.0434
β_{20}	4.4426	0.3934	3.6091	4.4609	5.1692
β_{11}	−0.0038	0.0007	−0.0052	−0.0038	−0.0024
β_{21}	−0.0021	0.0008	−0.0036	−0.0021	−0.0006
ϕ_1	0.161	0.028	0.113	0.158	0.223
σ_1	2.148	0.152	1.864	2.143	2.463
ϕ_2	0.575	0.154	0.350	0.548	0.948
σ_2	1.544	0.194	1.212	1.527	1.971
α	0.143	0.073	−0.003	0.144	0.285

Association between hake and anchovy is positive: geographical patterns of abundance of both species are related.

Coregionalised model for anchovies and hake: results (2)

Maps of posterior mean of spatial effect in univariate model (left), coregionalised model (middle), and difference of two spatial effects (right) for anchovies



Although differences are not remarkable, they still show that coregionalised model provides different (better?) results than the univariate model.

Coregionalised model for anchovies and hake: results (3)

Predictive capacity of Univariate and coregionalised models obtained by cross-validation of real 2016 data with the previous adjusted model.

Model	Measure	Value
Univariate	MAE	4212.35
Coregionalisation	MAE	3842.35
Univariate	RMSE	8422.54
Coregionalisation	RMSE	7697.88

Coregionalised model shows lower values than the univariate model, both in terms of MAE and RMSE, indicating a better predictive capacity.

4 Spatio-temporal structures

Spatio-temporal structures for SDMs

- In SDMs, most studies have been repeated periodically for long periods of time and so, there is interest in knowing whether the spatial evolution of the system under study varies not only in space but also in time.
- Temporal correlation depends on the same principle as spatial correlation: temporally close observations tend to be more related than temporally distant ones.
- Consequently, model fitting and predictions improve when a temporal term is added.
- However, temporal and spatial scales are different and the spatio-temporal analysis is more complicated than the simple addition of an extra dimension to the continuous spatial domain.

Spatio-temporal structures

Assuming a geostatistical spatial term (w):

$$\mathrm{w} \sim N(0, \mathrm{Q}(\kappa, \tau))$$

There are different U_{st} structures to infer the fundamental temporal behaviour of the process. Among them and following Paradinas et al. (2017):

• Opportunistic spatial distribution

$$U_{st} = w_{s_t} \quad (1)$$

• $\simeq$ Persistent spatial pattern with uncorrelated intensities

$$\begin{aligned} U_{st} &= w_{st} + v_t \\ v_t &= N(0, \sigma_t) \end{aligned} \quad (3)$$

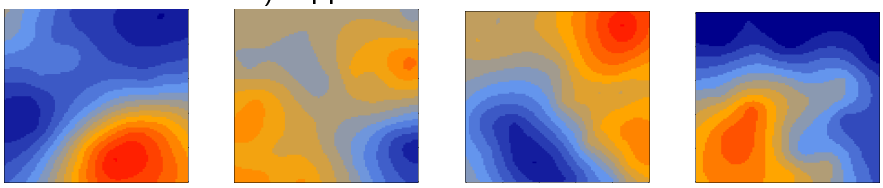
• Persistent spatial pattern with a temporal trend

$$U_{st} = w_{st} + f_t(t) \quad (2)$$

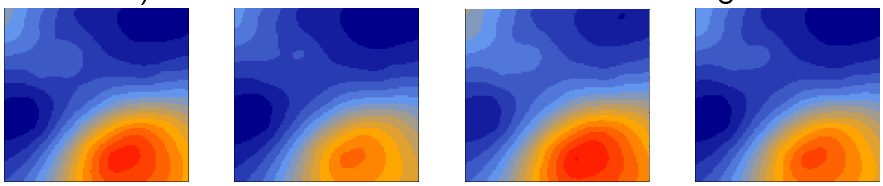
• Correlation among neighbouring years

$$\begin{aligned} U_{st} &= r_{st} + w_{st} \\ r_{st} &\sim \mathrm{N}\left(\sum_k \rho_k r_{st-k}, \tau_r^{-1}\right) \end{aligned} \quad (4)$$

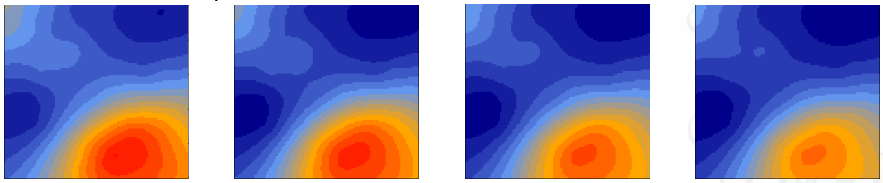
a) Opportunistic distribution



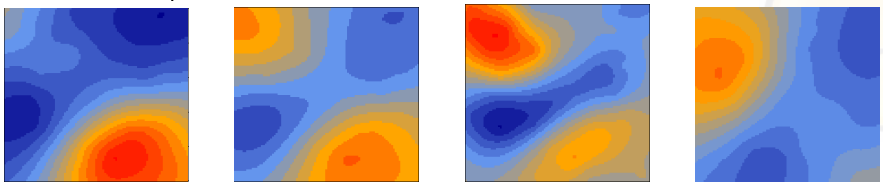
b) Persistent distribution with random changes



c) Persistent distribution with trend



d) Progressive spatio-temporal distribution



5 Dealing with excess of zeroes

How to deal with excess of zeroes

- A source of overdispersion caused by a disagreement between the data and the distribution assumed: there aremore zeros than the proposed distribution could reasonably explain.
- Different options:
  - Zero-inflated models: finite mixture of a degenerate distribution with all its mass at zero with a discrete distribution with support in $\mathbb{Z}^+ \cup \{0\}$.
  - Hurdle models for count data: a finite mixture of a degenerate distribution with all its mass at zero and a zero truncated discrete distribution. Unlike the previous, all observed zeros come from the zero-degenerate distribution.
  - Hurdle models for semi-continuous processes (rain, plant coverage, chemical concentrations, etc.) measured in the $[0, \infty)$ interval having high proportions of zero values are modeled as two independent sub-processes: one determines whether the response is zero, and the other determines the intensity when the response is non-zero using a continuous well known distribution (like Gamma).

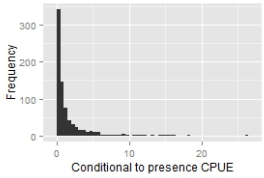
Analysis of hake recruitment (Paradinas et al., 2017)

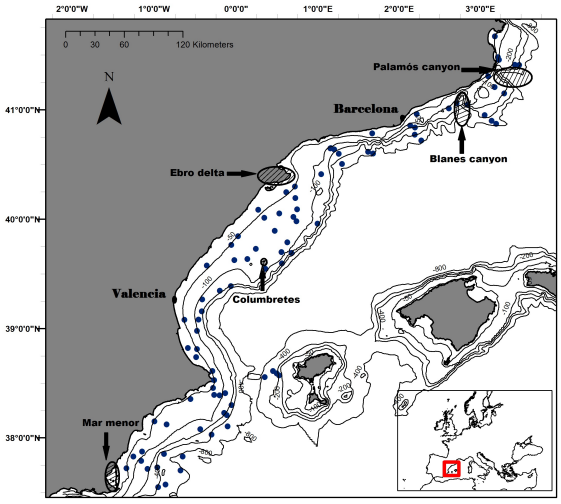
Data:

- 1048 observations
- from 2000 to 2012
- Occurrence

Presence	Absence
758	290

- Excess of zeroes
- Conditional-to-presence Abundance





Bathymetry is the main known driving factor of hake juveniles. Preference to 80-250 meters according to literature.

Two-part hurdle model for the abundance (continuous) of hake juveniles that allows different spatio-temporal structures in both distributions via shared components:

- Occurrence $\rightarrow Y_{st} \sim \text{Ber}(\pi_{st})$
$$\text{logit}(\pi_{st}) = \beta_o + f(depth) + u_{st}$$

- Abundance $\rightarrow Z_{st} \sim \mathcal{G}a(a_{st}, b_{st})$
$$\log(\mu_{st}) = \beta_a + \theta_f f(depth) + \theta_u u_{st}$$

  - u_{st} corresponding a spatio-temporal structure
  - θ_f and θ_f are scaling parameters that link the shared component in the bathymetry and in the spatial effect
  - The latent Gaussian field for the bathymetry is random walk of order 2

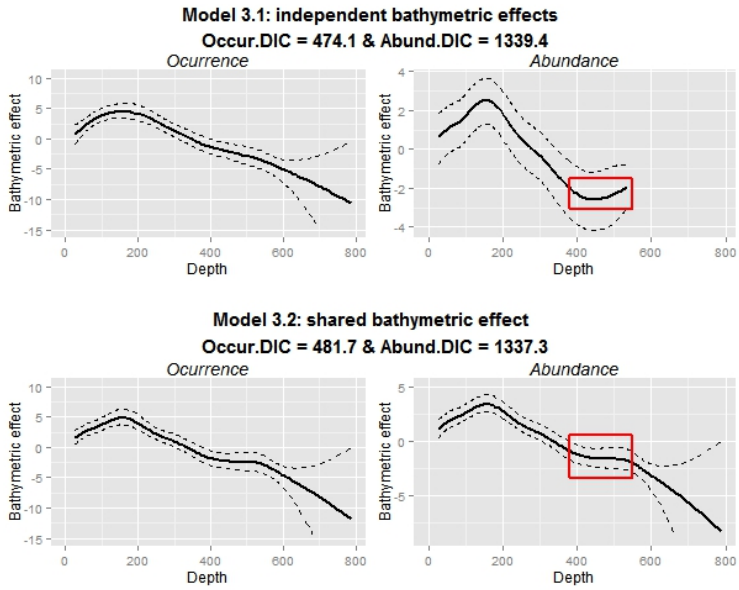
- Every combination of spatio-temporal structures and shared components were compared based on the deviance information criterion (DIC). Here's a summary:

Model	Structure	DIC	
		Occur	Abund
Model 0.1	I(b) + I(w) + I(iid.t)	466.8	1424.6
Model 0.2	I(b) + I(w) + I(trend.t)	475.9	1428.2
Model 1.1	I(b) + w_t	554.9	1487.6
Model 2.1	I(b) + S(w) + I(iid.t)	513.1	1432.8
Model 2.2	S(b) + I(w) + I(iid.t)	479.3	1425.4
Model 3.1	I(b) + I(w*t)	**474.1**	**1339.4**
Model 3.2	S(b) + I(w*t)	**481.7**	**1337.3**
Model 3.3	I(b) + S(w*t)	493.4	1573.7
Model 3.4	S(b) + S(w*t)	494.7	1573.7

b = bathymetry, w = spatial effect, t=temporal effect, w*t = spatio temporally structured effect. S() = shared components, I() = independent components.

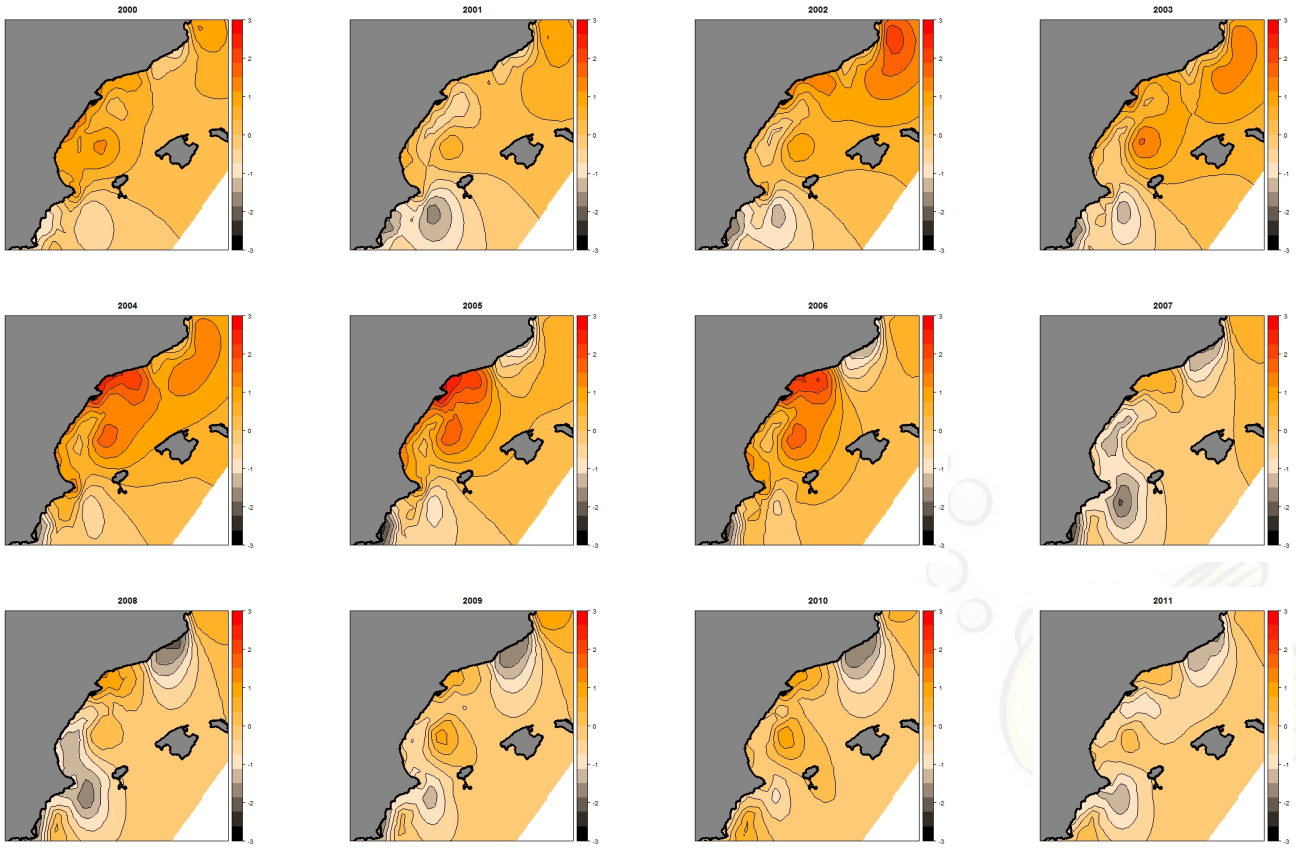
Results: shared bathymetric effect

- Finally **S(b) + I(w*t)** was selected



- Model 3.1 sligthly overfit the data (box). Still abundance DIC better in Model 3.2
- Model 3.2 fits a more natural bathymetric effect
- Model 3.2 allow the model predict deeper

Results: posterior distribution of the spatial effect for all years in the study



6 The effect of barriers

Dealing with barriers in SDMs

- When geographical elements or physical barriers are present, the spatial dependence structure could NOT be the same throughout all the domain: non-stationarity.
- SDMs usually assume stationarity in the spatial random effect, that is, the spatial autocorrelation only depends on the distance between locations, not on the direction nor the coordinates.
- Can be incorrect when physical barriers are present.

Dealing with non-stationarity

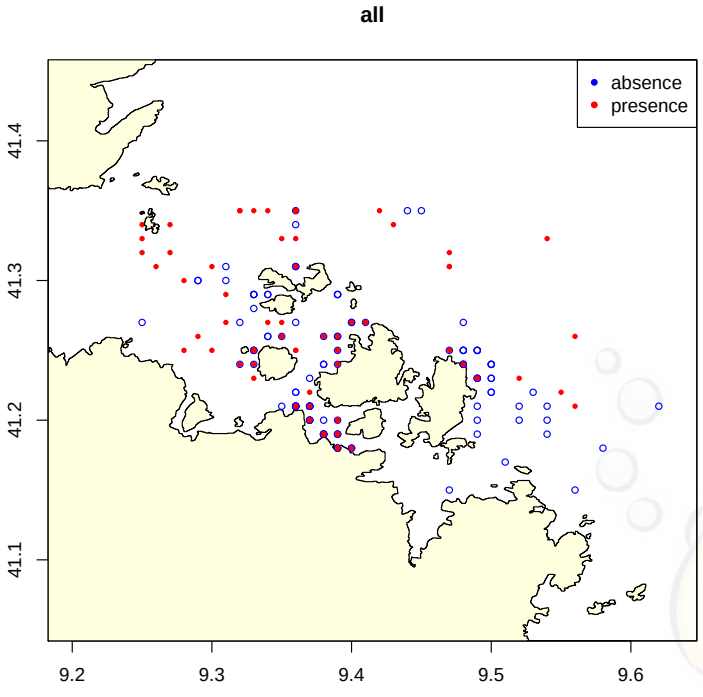
- Bakka et al. (2019) proposed a way to construct a GMRF locally (with one governing equation for the normal area, and another for the barrier area) as the solution to the system of SPDEs:

$$
\begin{aligned}
u(s) - \nabla \cdot \frac{r^2}{8} \nabla u(s) &= r \sqrt{\frac{\pi}{2}} \sigma_u W(s), \text{ for } s \in \Omega_n, \\
u(s) - \nabla \cdot \frac{r_b^2}{8} \nabla u(s) &= r_b \sqrt{\frac{\pi}{2}} \sigma_u W(s), \text{ for } s \in \Omega_b
\end{aligned}
$$

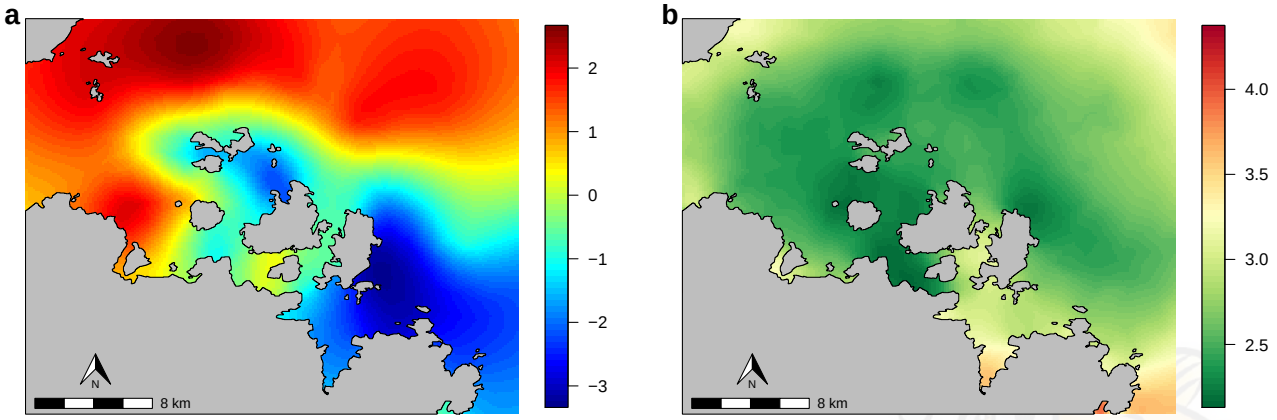
  where
  - Ω_n is the normal area; Ω_b is the barrier area; and
  - the range r_b in the barrier area is fixed close to zero.
- Martínez-Minaya et al. (2019) and Panunzi et al. (2024) have analyzed fisheries barriers with this approach.

Presence of dolphins in the Madalena

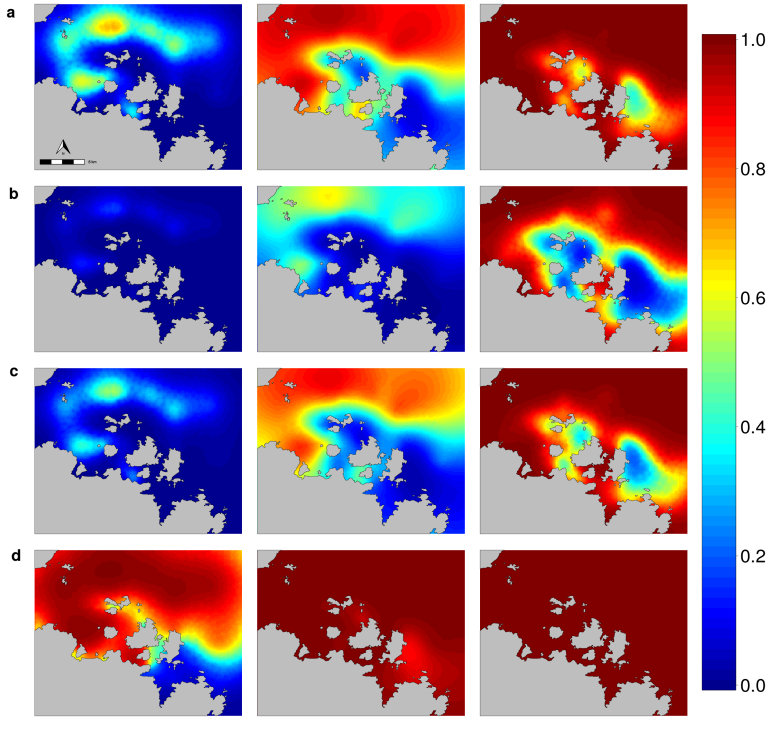
- Martínez-Minaya et al. (2019) have used this approach to analyse the presence of bottlenose dolphins in the North of Sardinia.



- Mean and standard deviation for posterior distribution of the spatial effect w.



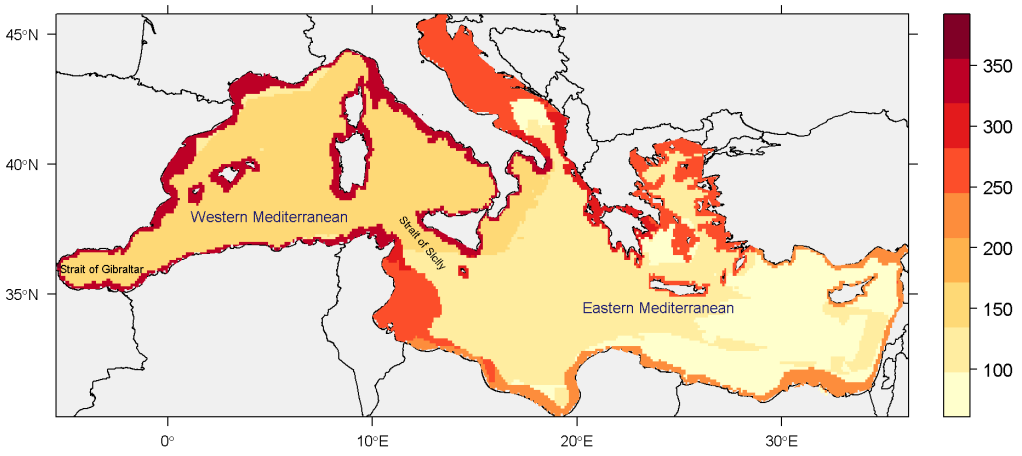
- Posterior predictive distribution of the probability of presence: 95 % credible intervals (First and third panel respectively) and the median (central panel) for the different seasons. a: autumn, b: summer, c: spring and d: winter.



7 Analysing areal data

Areal data

The quantity of interest is only defined for regions/quadrats .... It is measured/reported for certain fixed regions.



The way of modeling these quantities only defined at certain fixed regions is recognizing the fact that near quadrats should have similar information.

Hierarchical Bayesian Autocorrelated models

- Spatial Markovian assumption: value of the random variable at a given site only depends on the values at a specified set of neighboring sites.
- Data Model

$$Y(s_i) \sim Distribution(\theta_i)$$

where the distribution of response variable can be Normal (abundance), Poisson (number of species), Binomial, Beta (proportion of discards), etc.

- Process model

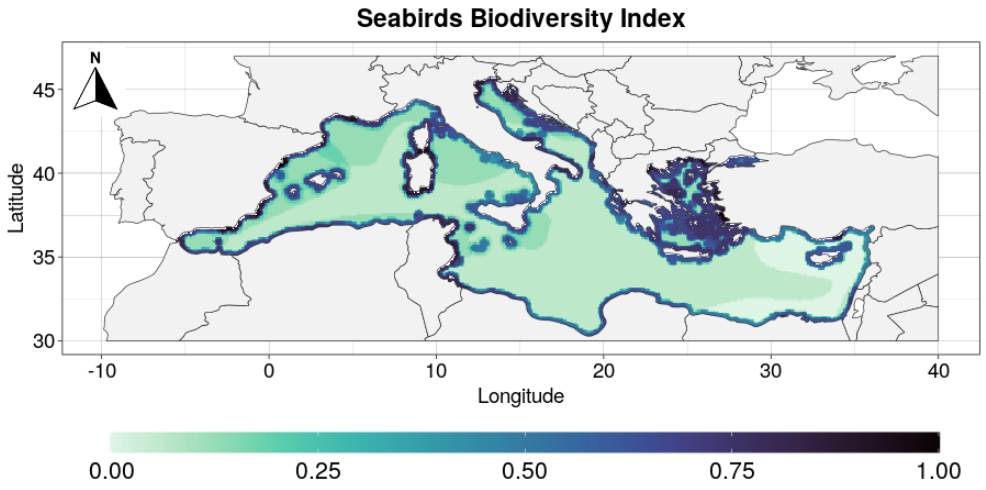
$$g(\theta_i) = X(s_i)\beta + Z_i$$

  - $X(s_i)$ known covariates
  - β fixed effects parameters
  - Spatial random effect has an Intrinsic Conditional Auto-Regressive distr.

$$[Z_i|Z_j, i \neq j, \tau_Z^2] \sim N\left(\bar{Z}_i = \frac{1}{\sum_j w_{ij}} \sum_j Z_j w_{ij}, \tau_i^2 = \frac{\tau_u^2}{\sum_j w_{ij}}\right)$$

    where $w_{ij} = 1$ if adjacent (0 otherwise).
  - Predictor can incorporate smooth nonlinear effects of covariates, time trends, seasonal effects, random intercept and slopes and temporal random effects.
- The prior for the hyperparameter τ_u must be assigned.

Analyzing seabirds biodiversity index (Sarzo et al., 2023)



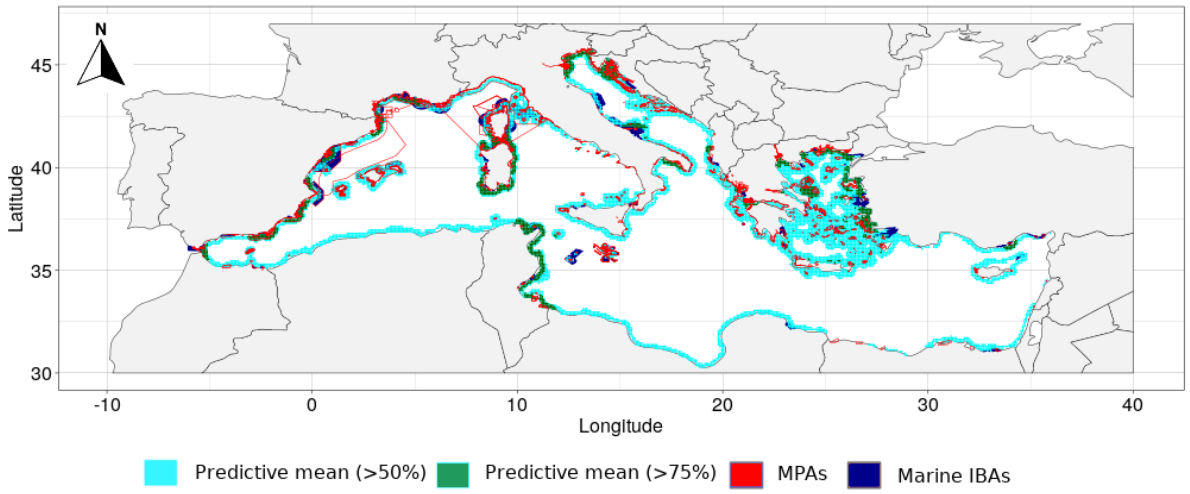
Seabirds biodiversity index for the Mediterranean Sea

Analyzing seabirds biodiversity index (Sarzo et al., 2023) (2)

As biodiversity index is scaled in the unit interval (0,1), a Bayesian Spatial Beta regression model is needed to analyze it:

$$
\begin{aligned}
Y_i &\sim \text{Beta}(\mu_i, \phi)\,, \\
\text{logit}(\mu) &= X_i\boldsymbol{\beta} + u_i\,, \\
\phi &= \exp(\theta)\,, \\
\\
\beta_j &\sim \text{N}(0, \tau = 10^{-3})\,, \\
u_i | u_j, i \neq j, \tau &\sim \text{N}\left(\frac{1}{\mathcal{N}_i}\sum_{i\sim j} u_j, \frac{1}{\mathcal{N}_i \tau}\right)\,, \\
\tau &\sim \text{PC-prior}(5, 0{,}1)\,, \\
\theta &\sim \text{LogGamma}(1, 0{,}01)\,.
\end{aligned}
$$

Analyzing seabirds biodiversity index (Sarzo et al., 2023) (3)



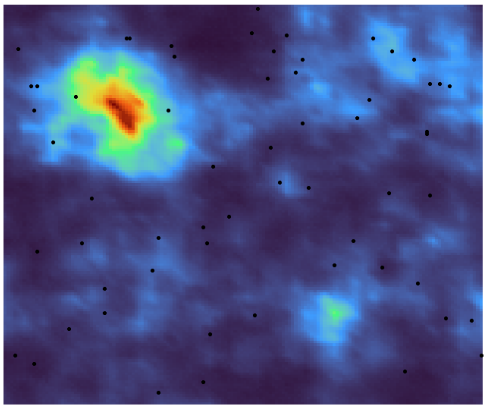
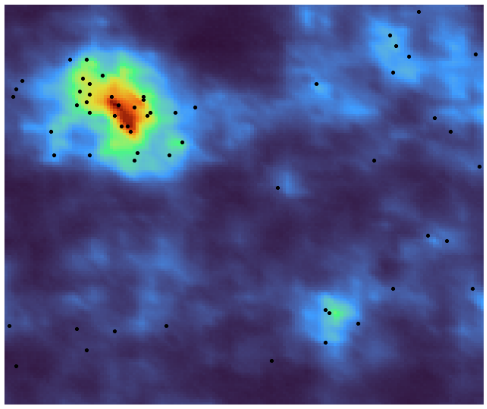
Overlap among posterior predictive mean values over 0.50 (cyan), 0.75 (green), MPAs (red), and marine IBAs (dark blue) in the Mediterranean Sea.

8 Combining different sources of data for SDMs

Combining information in Environmental sciences.

- Information can be provided by means of different sources in order to explain the same phenomenon. For example:
  - Dependent data: the sampling is preferential, influenced by the process under study. Usually in on-board observation, citizen data.
  - Independent data: it's a random sampling (information not shared with the geostatistical process). Usually in scientific surveys.
- Combining these diverse datasets has become an essential tool for managing the increasing complexity and volume of ecological data.
- As data complexity and volume grow, computational demands of previous proposals for data integration are also higher, creating significant challenges for practical implementation.
- Combining information in complex Spatio-temporal models is a clear example of this resulting computational burden.

Example of two differente sources of information

- Left side represents independent data collected from scientific surveys with a completely random sample pattern,
- while the right hand side represents dependent data coming from opportunistic or preferential sampling (e.g., whale watching or fishery commercial vessels),
- in which observers tend to look for a specific species or to fish in those areas where they expect to find it.

How to combine information

- There are different ways for combining sources of information:
  - Data pooling, aggregating data without explicitly accounting for their diversity;
  - Ensemble modelling, where multiple diverse models are created to predict a single outcome, either by applying various individual models to the same dataset or by using a single modelling setup across different datasets<,
  - Integrated models, combining data by modelling various datasets simultaneously.

Integrated models.

- Integrated models allow different sources of information to be combined by sharing components:
  - used to combine samples with different structures, such as completely random samples, stratified random samples or preferential samples; or
  - can also be used to combine information from different types of data on variables that share components in the latent structure.
- Examples:
  - combining the number of catches of a species with other information on the abundance of the same species;
  - combining information on the presence/absence of a toxin along a river with other measures of its concentration at different locations
- By combining the two sources of information in a joint model, we can analyse both variables simultaneously and to use common elements of the latent field for a more accurate and robust estimation.

Integrated models (2).

- In the previous examples, we have two different likelihoods whose linear predictors share a spatial component up to a parameter α:

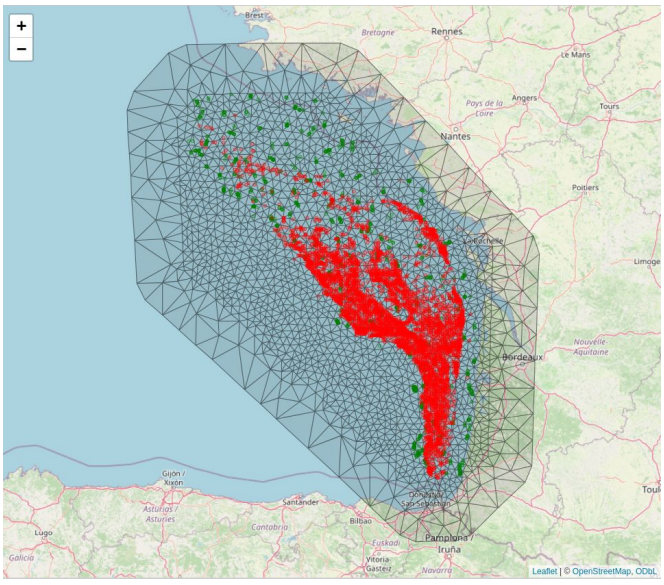
$$
\begin{aligned}
y_{1i} \mid \eta_{1i}, \boldsymbol{\theta}_1 &\sim \ell_1(y_{1i} \mid \eta_{1i}, \boldsymbol{\theta}_1), \\
y_{2j} \mid \eta_{2j}, \boldsymbol{\theta}_2 &\sim \ell_2(y_{2j} \mid \eta_{2j}, \boldsymbol{\theta}_2), \\
g_1(\mu_{1i}) = \eta_{1i} &= \beta_{10} + \mathsf{A}_{1i}\boldsymbol{\beta}_1 + \textstyle\sum_{j=1}^{J} f_{1j}(z_{1ij}) + u(s_i), \\
g_2(\mu_{2j}) = \eta_{2j} &= \beta_{20} + \mathsf{A}_{2j}\boldsymbol{\beta}_2 + \textstyle\sum_{j=1}^{J} f_{2j}(z_{2ij}) + \alpha \cdot u(s_j),
\end{aligned}
$$

  where ℓ_1 and ℓ_2 are the likelihood functions for y_1 and y_2, and $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are the set of hyperparameters associated with each likelihood.
- Note the flexibility of the integrated modelling approach to analyze multiple sources of information together.

Inference and prediction

- Resulting spatial models can be analysed with the INLA-SPDE approach:
  - Inference on the parameters;
  - Model selection; and
  - Prediction on new locations (map of the quantity of interest all over the space).
- Although INLA is fast, still the computational burden is huge.
- Figueira et al. (2024a) have presented a way carry out a Bayesian sequential learning process of a model for new data on the same phenomenon by using information from a previous spatial model to feed another one.
- Figueira et al, (2025a) have also presented a sequential consensus Bayesian inference procedure designed to offer the flexibility of integrated models while significantly reducing computational costs, in the context of spatio-temporal species distributions models.

Three data sets of info for the hake in the bay of Biscay



Case study of the distribution of hake in the Bay of Biscay during the years 2003 to 2021.

Three data sets of info for the hake in the bay of Biscay

- Information about the distribution of hake in the Bay of Biscay during the years 2003 to 2021 comes via three datasets:
  - the scientific EVHOE trawl survey which collected discrete abundance data;
  - one commercial fishing fleet, sampled by on-board observers and targeting hake, collected continuous biomass data; and
  - another commercial fishing fleet, sampled also by on-board observers but not targeting hake, recorded presence-absence data.
- Note that the first commercial fleet targeting hake carried out a preferential exploration of the sea in order to maximise the hake biomass catch.

Combining the different sources of information (3)

The joint model that takes into account all this information:

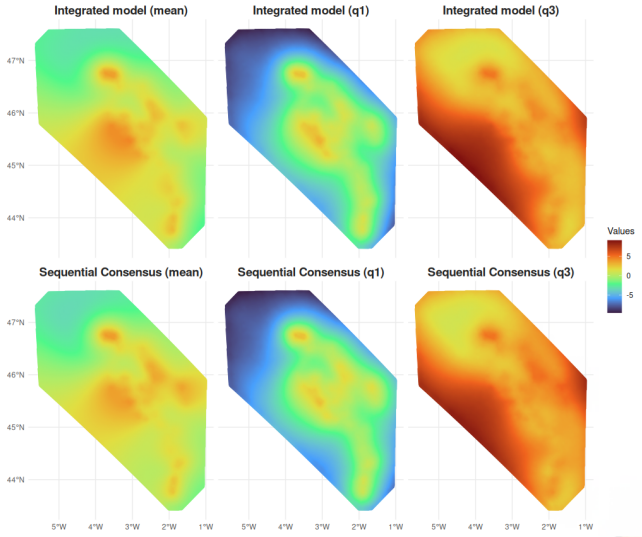
$$
\begin{aligned}
y_{1i} \mid \eta_{1i} &\sim \text{Po}(\lambda_i) \\
\log(\lambda_i) &= \beta_{10} + f_{1d}(z_{di}) + f_{1y}(z_{yi}) + \alpha_{s1} \cdot u_i, \\
y_{2j} \mid \eta_{2j}, \tau &\sim \text{Gamma}(y_i \mid \eta_{2j}, \tau), \\
\log(\mu_j) &= \beta_{20} + f_{2d}(z_{dj}) + f_{2y}(z_{yj}) + u_j, \\
y_{3j} \mid \eta_{3j} &\sim \text{Ber}(\pi_j), \\
\text{logit}(\pi_j) &= \beta_{30} + \alpha_{d3} \cdot f_{2d}(z_{dj}) + f_{3y}(z_{yj}) + \alpha_{s3} \cdot u_j, \\
y(s_j) \mid \lambda(s_j) &\sim \text{LGCP}(\lambda(s_j)), \\
\log(\lambda(s_j)) &= \beta_{40} + \alpha_{d4} \cdot f_{2d}(z_{dj}) + u_j^*,
\end{aligned}
$$

where:

- the $\boldsymbol{\alpha}$ components represent scaling parameters for the shared effects;
- f_d represents a structured random effect related to the depth covariate (a second-order random walk for f_{1d} and a one-dimensional SPDE for f_{2d});
- f_y is a first-order random walk for years; and
- u is a separable type III spatio-temporal effect (Knorr-Held, 2000), with a precision matrix derived from a two-dimensional SPDE with Mátern covariances for the spatial part and an iid precision matrix for the temporal part.

Results of the distribution of the hake in the bay of Biscay

- Spatial-temporal effect for the first temporal node of the spatio-temporal component for the LGCP:



- Even with three small datasets, our sequential consensus inference proposal yields similar results to those obtained from the joint model.
- But, in terms of computational time, the integrated model takes 62,12 minutes, while our proposal takes 13,81 minutes (around 4 times less).

9 Dealing with presence-only data

How to handle presence-only data

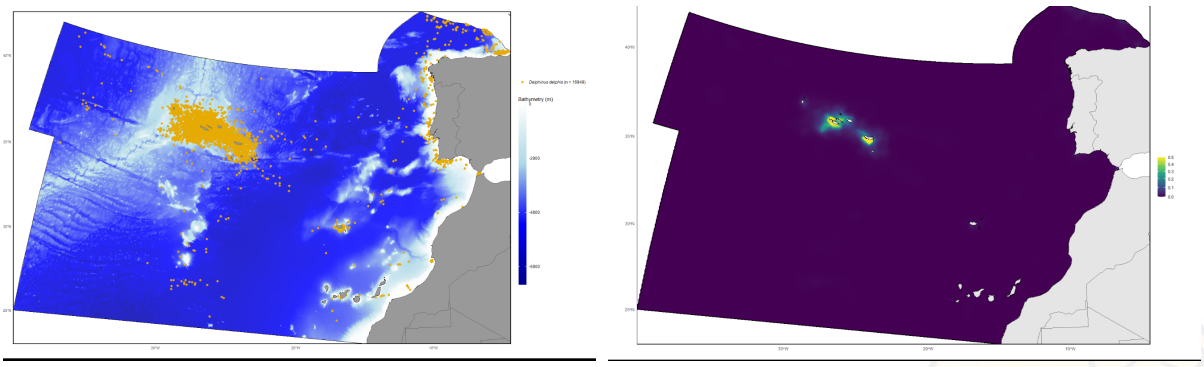
- In many situations our data are just a mere representation of places where we have seen the species of interest.
- We cannot say that it is not in the remaining places because we do not know whether the species is present or not. So there is no information about where the species is not.
- The usual approach is to generate some pseudo-absences (Fuster et al., 2024).
- This approach has many drawbacks: where should they be (although random, not clear), how many of them, in which proportion.
- Dorazio (2014) highlighted various statistical models developed to integrate presence-only data from different research protocols. But these models often overlook the impact of imperfect detectability and survey bias.
- A better choice is to consider the observed locations as a point pattern, that is, point events coming from Point process model (Warton and Shepherd, 2010; Renner et al., 2015).

How to handle presence-only data in fisheries context

- Yuan et al. (2017) introduced many useful tools for analyzing point patterns in INLA in the context of spatio-temporal distance sampling data from a large-scale survey of blue whales.
- After this work, Martino et al. (2021) demonstrated that correcting for detectability issues can mitigate biases in estimates caused by varying detection mechanisms during data collection, thereby allowing for the integration of multiple information sources.
- Pace et al. (2022) went further by providing a novel approach to refine the presence-only detection function, particularly using social media data sources, and by exploring potential seasonal effects, supported by a larger dataset and a broader study area.
- Still, the relatively low number of sightings, dispersed over time, limits the ability to identify detailed temporal patterns. Not space-time analysis.

Assessing cetacean spatial distribution from opportunistic sighting data

Fernández et al. (2025) have used this approach to model the spatial distribution of common dolphin (Delphinus delphis) and sperm whale (Physeter macrocephalus) using Log-Gaussian Cox Process with presence-only data and integrating a proxy for sampling effort.

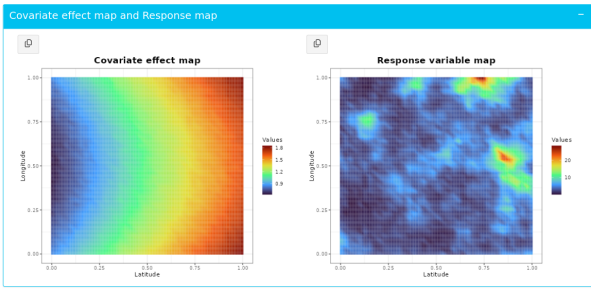


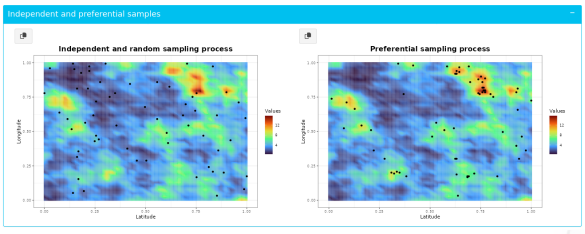
Left: Study area with Doklphins sightseens in summer; right: prediction of the intensity (ind/km^2).

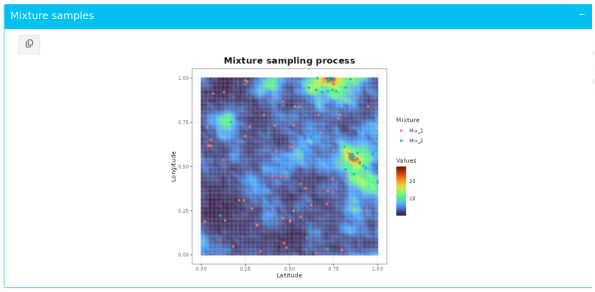
10 R shiny app

A Shiny app for implementing SDMs (Figueira et al., 2024b).

- BAYSPINS (BAYesian SPatial INla for SDMs) has been created to allow practicioners to perform:
  - geostatistical and preferential modeling;
  - a mixture of both modelings;
  - simulation of models to understand its behaviour;
  - feedback analysis between models.
- Can be used for applied scientists INLA beginners: the use of default settings automates the process or the customisation of a large number of elements that drive the modelling process.
- Allows from quick initial evaluations to more rigorous studies depending on the user's skill and understanding of the fundamentals underpinning the application.
- More info at `https://github.com/MarioFigueiraP/ShinyAppSpatialModelFeedback`.



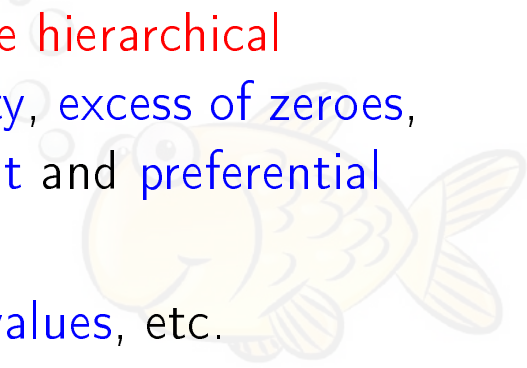




11 Conclusions

Conclusions

- Hierarchical Bayesian modelling can be a really useful tool for analysing SDMs.
- INLA + SPDE can also be very convenient as they are fast and provides good results (probably not the best ones, but the first ones in your analysis).
- Situations in real life bring us many other statistical issues that can be incorporated in the hierarchical framework presented: Non-stationarity, excess of zeroes, spatio-temporal models, misalignment and preferential sampling among others.
- Still many open problems! Extreme values, etc.



12 References

References

1. H. Bakka, J. Vanhatalo, J.B. Illian, D. Simpson and H. Rue (2019). Non-stationary Gaussian models with physical barriers. *Spatial Statistics*, 29, 268–288.
2. X. Barber, D. Conesa, S. Lladosa and A. López-Quílez (2016). Modelling the presence of disease under spatial misalignment using Bayesian latent Gaussian models. *Geospatial Health*, **11**, 469.
3. X. Barber, D. Conesa, A. López-Quílez and J. Morales (2019). Multivariate Bioclimatic indices modelling: A coregionalised approach. *Journal of Agricultural, Biological and Environmental Statistics*, **24**(2): 225–244.
4. X. Barber, D. Conesa, A. López-Quílez, J. Martínez-Minaya, I. Paradinas and M.G. Pennino (2021). Incorporating Biotic Information In Species Distribution Models: A Corregionalised Approach. *Mathematics*, **9**(4), 417.
5. J. Carmezim, M. G. Pennino, J. Martinez-Minaya, D. Conesa, M. Coll (2022). A mesoscale analysis of relations between fish species richness and environmental and anthropogenic pressures in the Mediterranean Sea. *Marine Environmental Research*, **180**, 105702.
6. M. Cendoya, A. Hubel, D. Conesa and A. Vicent (2022). Modeling the spatial distribution of *Xylella fastidiosa*. A non-stationary approach with dispersal barriers. *Phytopathology*, **112**, 1036–1045.
7. R. Dorazio (2014). Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography*, **23**(12), 1472–1484.
8. D. Fernández et al. (2025). Applying Bayesian preferential sampling models to assess cetacean spatial distribution from opportunistic sighting data. Poster at the 36th European Cetacean Conference.

9. M. Figueira, X. Barber, D. Conesa, A. López-Quílez, J. Martínez-Minaya, I. Paradinas, M.G. Pennino (2024a). Bayesian feedback in the framework of ecological sciences. *Ecological Informatics*, **84**, 102858.
10. M. Figueira, D. Conesa and A. López-Quílez (2024b). A Shiny R app for spatial analysis of Species Distribution Models. *Ecological Informatics*, **80**, 102542.
11. M. Figueira, D. Conesa, A. López-Quílez, I. Paradinas (2025). A computationally efficien.t procedure for combining ecological datasets by means of a sequential consensus inference. *Journal of Environmental and Ecological Statistics*, **in press**.
12. A. Fuster-Alonso J. Mestre-Tomás, J.C. Baez, M.G. Pennino, X. Barber, J.M. Bellido, D. Conesa, A. López-Quílez, J. Steenbeek, V. Christensen, M. Coll (2025). Machine learning applied to global scale species distribution models. Submitted.
13. J. Martínez-Minaya, M. Cameletti, D. Conesa and M.G. Pennino (2018). Species distribution modeling: a statistical review with focus in spatio-temporal issues. *Stochastic Environmental Research and Risk Assessment*, **32**, 3227–3244.
14. J. Martínez-Minaya, D. Conesa, H. Bakka and M.G. Pennino (2019). Dealing with physical barriers in bottlenose dolphin Tursiops truncatus distribution. *Ecological Modelling*, **406**: 44–49.
15. S. Martino, D.S. Pace, S. Moro, E. Casoli, D. Ventura, A. Frachea, M. Silvestri, A. Arcangeli, G. Giacomini, G. Ardizzone, G.J. Lasinio (2021). Integration of presence-onl data from several sources: a case study on dolphins' spatial distribution. *Ecography*, **44**(10), 1533–1543.
16. D. S. Pace, G. Panunzi, A. Arcangeli, S. Moro, G.J. Lasinio, and S. Martino (2022). Seasonal distribution of an opportunistic apex predator (Tursiops truncatus) in marine coastal habitats of the Western Mediterranean Sea. *Frontiers in Marine Science*, **9**, 939692.

17. G. Panunzi, S. Moro, I. Marques, S. Martino, F. Colloca, F. Ferretti, and G.J. Lasinio (2024). Estimating the spatial distribution of the white shark in the Mediterranean Sea via an integrated species distribution model accounting for physical barriers. *Environmetrics*, **36**(1), e2876.
18. I. Paradinas, D. Conesa, A. López-Quílez, J.M. Bellido (2017). Spatio-Temporal model structures with shared components for semi-continuous species distribution modelling. *Spatial Statistics*, **22**, 434-450.
19. M.G. Pennino, I. Paradinas, J.B. Illian, F. Muñoz, J.M. Bellido, A. López-Quílez and D. Conesa (2019). Accounting for preferential sampling in species distribution models. *Ecology and Evolution*, **9**(1): 653-663.
20. I.W. Renner, J. Elith, A. Baddeley, W. Fithian, T. Hastie, S.J. Phillips, G. Popovic, and D.I. Warton (2015). Point process models for presence-only analysis. *Methods in Ecology and Evolution*, **6**(4), 366–379.
21. B. Sarzo, J. Martínez-Minaya, M.G. Pennino, D. Conesa and M. Coll (2023). Modelling seabirds biodiversity through Bayesian Spatial Beta regression models: A proxy to inform marine protected areas in the Mediterranean Sea. *Marine Environmental Research*, **185**, 105860.
22. D.I. Warton and L.C. Shepherd (2010). Poisson point process models solve the pseudo-absence problem for presence-only data in ecology. *The Annals of Applied Statistics*, 1383–1402.
23. Y. Yuan, F.E. Bachl, F. Lindgren, D.L. Borchers, J.B. Illian, S.T. Buckland, H. Rue, T. Gerrodette (2017). Point process models for spatio-temporal distance sampling data from a large-scale survey of blue whales. *Ann. Appl. Stat.*, **11**(4): 2270–2297.