SDMS as Geostatistical models

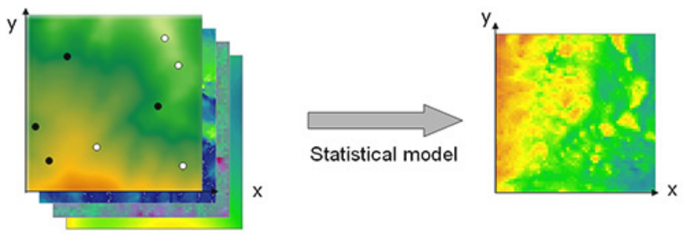
David V. Conesa Guillén

VAlencia BAyesian Research group

Universitat de València

Species distribution modelling (SDMs)

- SDMs links spatially referenced records of species occurrence with maps of environmental variables in order to create a statistical model of the relationship between a species and its environment.

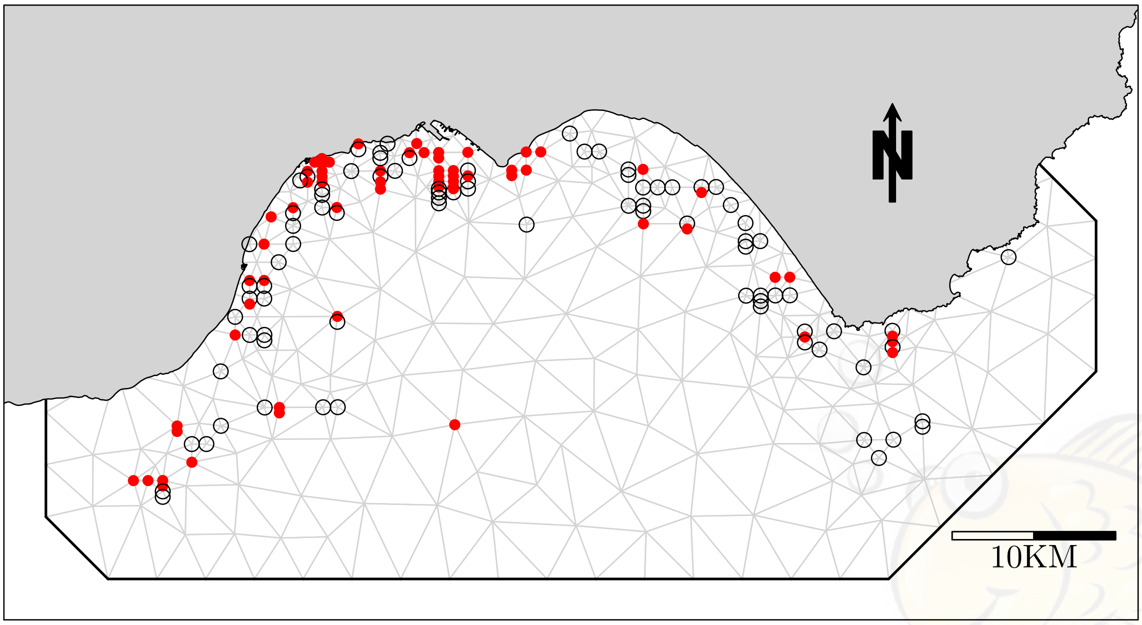


Field records and maps of environment   Map of probability species is present

- Typical examples: diseases, fish species, plants, animals, etc.
- Typical covariates: elevation, climate, vegetation, human disturbance, temperature, chlorophyll-a, etc.
- Applications: climate change, conservation of species, etc.
- Many options to deal with them. See for instance among many other reviews Rodrigues et al (2023a, 2023b). Our focus here will be deal with them as Geostatistical Bayesian Hierarchical models.

Occurrence of Mackerel in Gulf of Almería

The measures are presences (red) and absences (**white**)



1 Species Distribution Models as Geostatistical models

SDMs as Bayesian Hierarchical Geostatistical models.

- Different response variables (normal, Poisson, binomial, beta, etc.) depending of the situation.
- Data Model y $= (y_1, \ldots, y_n)$ represents the observed values of the corresponding variable of interest with mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ at n locations.
- Process model each μ_i can be easily linked to a structured additive predictor η_i and a geostatistical spacial random effect:

$$g(\mu_i) = X_i\beta + W_i \,,\; i = 1, \ldots, n.$$

  The spatial effect W is Gaussian: it is a Latent Gaussian field.
- This predictor can incorporate also smooth nonlinear effects of covariates, time trends and seasonal effects, random intercept and slopes as well as temporal random effects.
- Prior distributions for the parameters and hyperparameters need to be assigned.

- The resulting hierarchical spatial model do not yield to analytical expressions for the posterior distributions of the parameters.
- Posterior distributions of the parameters can be approached by means of MCMC but they are not computationally efficient in this context.
- The Integrated Nested Laplace Approximation (Rue et al., 2009), colloquially known as INLA, is a better and faster alternative to approximate the posterior marginals of interest.
- But geostatistics involves working with continuously indexed Gaussian Fields and INLA cannot be applied directly.
- Lindgren et al. (2011) proposed an explicit link between *Gaussian fields* and *Gaussian Markov random fields* (GMRF): the Stochastic Partial Differential Equation approach (**SPDE**).
- GMRFs are discretely indexed $\longrightarrow$ the Markov property makes the precision matrix involved sparse (allowing the use of faster numerical algorithms).

Final modelling using SPDE

$$\begin{aligned} Z_i &\sim \text{Ber}(\pi_i),\ i = 1, \ldots, n \\ \text{logit}(\pi_i) &= X_i\beta + w_i \\ \pi(\beta_j) &\sim N(\mu_{\beta_j}, p_{\beta_j}) \\ \text{w} &\sim N(0, \mathbf{Q}(\kappa, \tau)) \\ 2\log\kappa &\sim N(\mu_\kappa, p_\kappa) \\ \log\tau &\sim N(\mu_\tau, p_\tau) \end{aligned}$$

- The spatial effect depends on now two different parameters: κ (the range of the effect) and τ (the total variance). Prior information for them via pc-priors (Simpson et al., 2017).
- Once the inference is performed, the model is used to predict presence/abundance in unsampled places (providing a map of the species behaviour).
- As usual, model comparison can be done with criterions:
  - DIC (Spiegelhalter et al., 2002), WAIC (Watanabe, 2010)
  - LCPO (Gneiting and Raftery, 2007)

2 Modelling the Ocurrence of a species

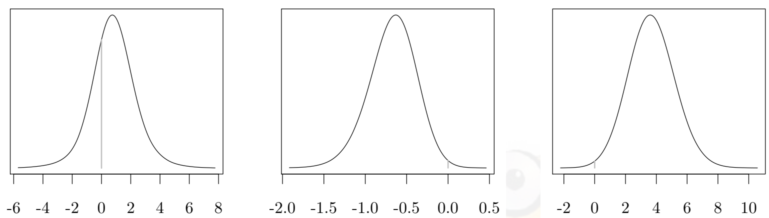
Logistic regression models for presences and absences

When the observed variable is the ocurrence (presence/absence), we obtain maps of the posterior predictive distribution of the probability of finding the species (or disease):

- Distribution of Mediterranean horse mackerel in Gulf of Almería (Muñoz et al., 2013).
- Distribution of prevalence of Bovine paramphistomosis in Galicia (González-Warleta et al., 2013).
- Distribution of prevalence of Citrus Black Spot (Martínez-Minaya et al., 2018).
- Distribution of vulnerable (to fishing pressure) elasmobranch species (Pennino et al., 2013).
- Distribution of prevalence of Xylella fastidiosa in Alicante (Cendoya et al., 2020).

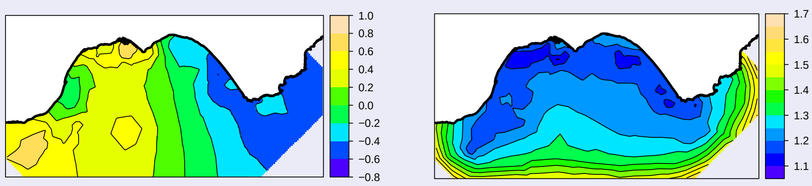
- In spite of its low commercial value, mackerel plays an important role in the observed transition zone between the Mediterranean and Atlantic sea.
- Covariates: geographical coordinates, bathymetry, monthly precipitation, sea surface temperature, chlorophyll-a concentration.

- We used DIC and CPO to compare models with different combinations of the available covariates.
- Final model includes log(Depth) (negative effect) and chlorophyll-a (positive effect).



Intercept   log(Depth)   chlorophyll-a

Spatial effect



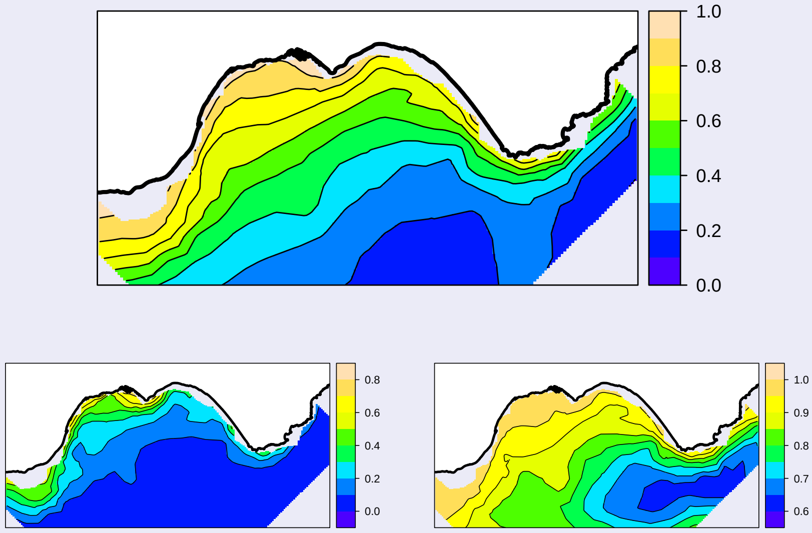
Posterior mean and standard deviation of the spatial effect

Interpretation

- There seems to be a east-west effect.
- This was associated a posteriori by local experts with the fact that the western area of the bay is a protected coastline with favourable conditions for the species.
- The model provided (unexpectedly) a quantification of the impact of this protective action on the Mediterranean horse mackerel.

Distribution of Mediterranean horse mackerel in Gulf of Almería (3)

Posterior probability of ocurrence

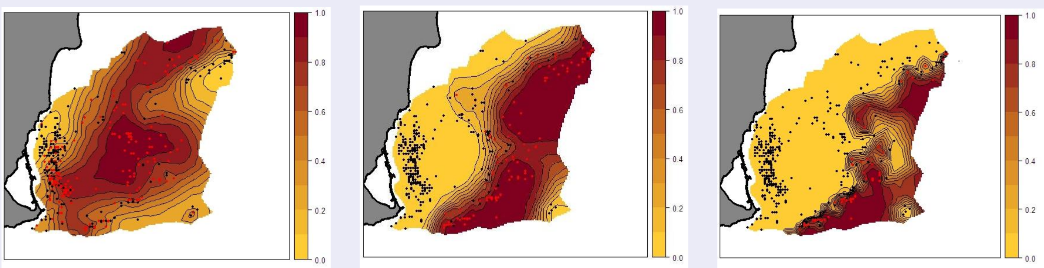


Median (top) and quartiles 1 and 3 of π_i|data

Distribution of three elasmobranch species (Pennino et al., 2013).

- There is an increasing concern over elasmobranch species because they are highly vulnerable to fishing pressure.
- Main predictors of elasmobranch habitats are depth, slope of seabed and type of substrate, followed by temperature and chlorophyll-a.

Median of the posterior probability of the presence of elasmobranch species



- Species show different optimum depths: could indicate a sort of fine-tuned bathymetric segregation, though they coexist on shelf and slope bottoms.
- These maps can be used to identify sensitive habitats, with a final aim to improve management and conservation of these vulnerable species.

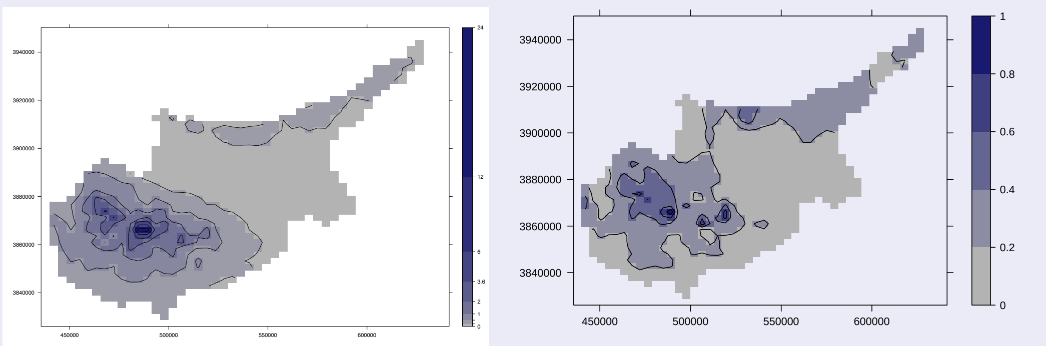
3 Modelling the Abundance of a species

Dealing with positive continuous distributions

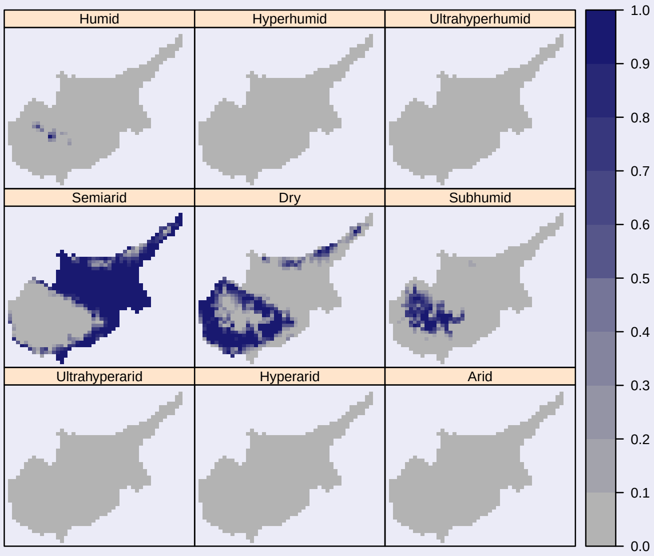
The observed variable can be also abundances (normal, gamma, lognormal, etc.):

- Abundance of hake recruitment (Paradinas et al., 2015, 2017; Izquierdo et al., 2021), amount of discards (Pennino et al., 2014), abundace of red mullet (Paradinas et al, 2020), CPUE biomass indices (Fuster-Alonso et al. 2024).
- Milk protein content of cows with fasciolosis (González-Warleta et al., 2024).
- Spatial distribution of bioclimatic indices (Barber et al., 2017).

Mean (left) and standard deviation (right) of the posterior predictive distribution of the Ombrothermic Index



- Bayesian statistics allows us to go further with the bioclimatic indices. We can obtain maps of the spatial bioclimatic probability distribution by showing the posterior probability of an index belonging to each climate subtype.
- The map contains the posterior probability of the nine possible ombrotypes (categories of the Ombrothermic Index) that can be observed in the Mediterranean bioclimate at each location.



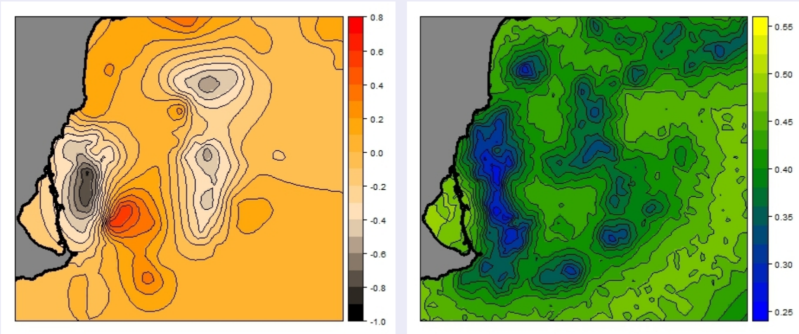
4 Dealing with proportions

Beta spatial regression for proportions

The observed variable can also be expressed in terms of proportions:

- Paradinas et al. (2016, 2018) have applied spatial Beta regression models to analyse the proportion of discards, and identified two main fishing-suitable areas based on the proportion of discarded regulated fish proportions.

Posterior mean (left) standard deviation (right) of the spatial component of the regulated species discard proportion



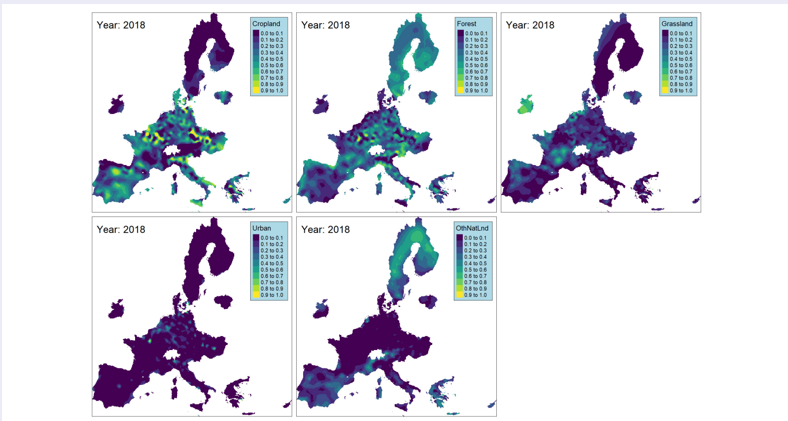
- Martínez-Minaya et al. (2019) have analysed the proportion of arapbidosis thaliana to study the effect of climate change.
- Castilla et al. (2020) have also analysed the ecological, genetic and evolutionary drivers of regional genetic differentiation in Arabidopsis thaliana.

Dirichlet regression for vectors of proportions

The observed variable can be expressed in terms of proportios of different groups:

- Compositional data $Y = \{y_{c \times i}\}$ are defined as a vector of values in the interval (0,1) that satisfies $\sum_{c=1}^{C} y_{ci} = 1 : c \in \{1, .., C\}$ and $n \in \{1, ..., N\}$.
- In other words, data consisting in proportions (of categories) adding up to 1.
- Martínez-Minaya and Rue (2024) have re-analysed the proportions of *Arapdidosis Taliana* in Peninsula Iberica to study the effect of climate change.
- Figueira et al (2025) have analysed proportions of land-use in Europe.

Posterior predictive distribution of proportions of land use.



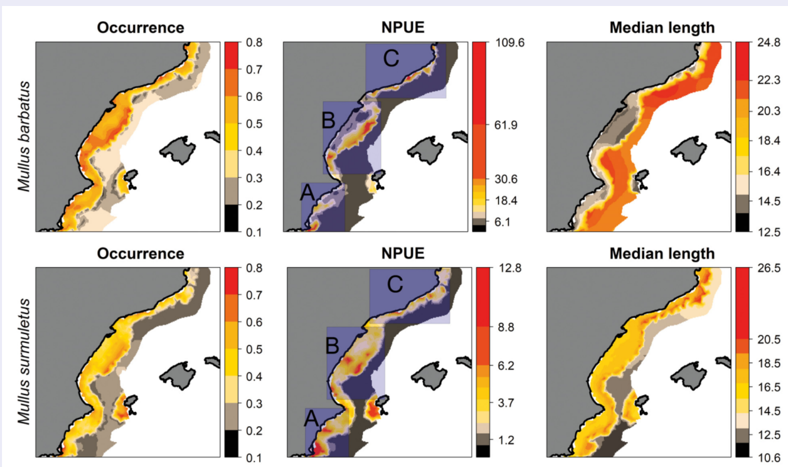
5 Dealing with counts

Regression models for counts

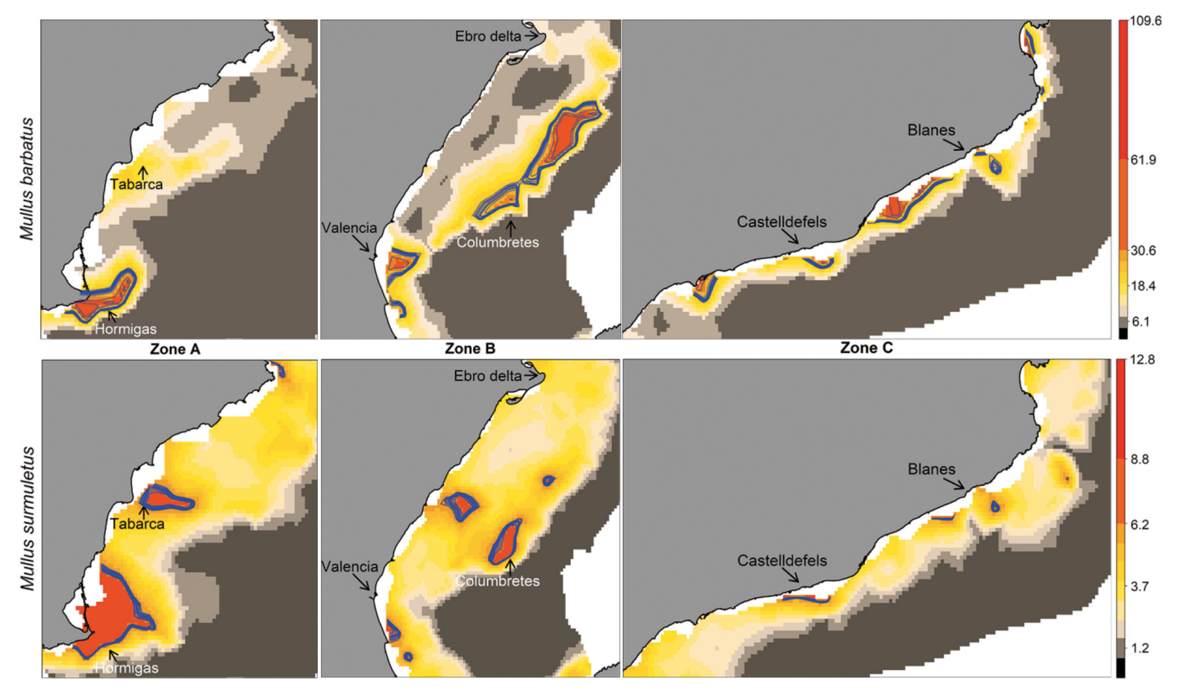
The observed variable can also be expressed in terms of counts:

- Paradinas et al. (2020) have considered spatial Negative Binomial regression models to analyse the number of catchs per 30 minutes (NPUE) in order to understand the spatiotemporal persistence of fish distributions and so help to define fish hotspots and effective fisheries-restricted areas.

Average spatial distributions of Mullus barbatus and M. surmuletus occurrence, NPUE and median length (cm) between 2000 and 2016. Blue rectangles represent zoomed sections in next slide.



Zoomed spatial distribution of Mullus barbatus and M. surmuletus NPUE in Murcia (Zone A), Valencia (Zone B) and Catalonia (Zone C). Hotspots, identified as areas with NPUE above the 90 th percentile, are highlighted in blue



6 References

References

1. X. Barber, J. Morales, A. López-Quílez, D. Conesa, A. Martínez-Mayoral and A. Barber (2017). Bayesian hierarchical models for analysing the spatial distribution of bioclimatic indices. *Statistics and Operations Research Transactions*, **41**(2), 277–296.
2. A. Castilla, B. Méndez-Vigo, A. Marcer, J. Martínez-Minaya, D. Conesa, X. Picó and C. Alonso-Blanco (2020). Ecological, genetic and evolutionary drivers of regional genetic differentiation in *Arabidopsis thaliana*. *BMC Evolutionary Biology*, **20:71**.
3. M. Cendoya, J. Martínez-Minaya, V. Dalmau, A. Ferrer, M. Saponari, D. Conesa, A. López-Quílez, A. Vicent (2020). Spatial Bayesian modeling applied to the surveys of *Xylella fastidiosa* in Alicante (Spain) and Apulia (Italy). *Frontiers in Plant Science*, **11**:1204.
4. M. Figueira, C. Guarner, D. Conesa, A. López-Quílez, T. Krisztin (2025). Unveiling land use dynamics: Insights from a hierarchical Bayesian spatio-temporal modelling of Compositional Data. *Journal Agricultural, Biological and Env. Statistics*, **in press**.
5. T. Gneiting and A.E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* 102 (477), 359–378.
6. M. González-Warleta, S. Lladosa, J.A. Castro-Hermida, A.M. Martínez-Ibeas, D. Conesa, F. Muñoz, A. López-Quílez, Y. Manga-González and M. Mezo, (2013). Bovine paramphistomosis in Galicia (Spain): Prevalence, intensity, aetiology and geospatial distribution of the infection. *Veterinary Parasitology*, 191(3-4): 252-263.

7. M. González-Warleta, J.A. Castro-Hermida, M. Figueira, J. López, D. Conesa, A. López-Quílez, F.M. Ubeira, M. Mezo (2024). Bayesian hierarchical modelling of the geospatial distribution of fasciolosis in dairy cattle and the impact on production: Application to the main milk-producing region (Galicia) in Spain. *Veterinary Parasitology*, **325**, 110091.
8. F. Izquierdo, I. Paradinas, S. Cerviño, D. Conesa, A. Fernández, F. Velasco, I. Preciado, A. Punzón, F. Saborido-Rey and M. G. Pennino (2021). Spatio-temporal assessment of the European hake (*Merluccius merluccius*) recruits in the northern Iberian Peninsula. *Frontiers in Marine Science, section Marine Conservation and Sustainability*, **8**: 614675.
9. F. Lindgren, H. Rue and J. Lindstrom, (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the SPDE approach (with discussion). *Journal of the Royal Statistical Society, Series B*, 73: 423-498.
10. J. Martínez-Minaya, D. Conesa, A. López-Quílez and A. Vicent (2018). Spatial and climatic factors associated with the geographical distribution of citrus black spot disease in South Africa. A Bayesian latent Gaussian model approach. *European Journal of Plant Pathology*, **151**(4), 991–1007.
11. J. Martínez-Minaya, D. Conesa, M.J. Fortin, C. Alonso-Blanco, F.X. Picó and A. Marcer (2019). A hierarchical Bayesian Beta regression approach to study the effects of geographical genetic structure and spatial autocorrelation on species distribution range shifts. *Molecular Ecology Resources*, **19**: 929–943.
12. J. Martínez-Minaya and H. Rue (2024). A flexible Bayesian tool for CoDa mixed models: logistic-normal distribution with Dirichlet covariance. *Statistics and Computing*, **34**, 116.
13. F. Muñoz, M. G. Pennino, D. Conesa, A. López-Quílez and J.M. Bellido (2013). Estimation and prediction of the spatial occurrence of fish species using Bayesian latent Gaussian models. *Stoch Environ Res Risk Assess*, 27: 1171–1180.

15 I. Paradinas, D. Conesa, M.G. Pennino, F. Muñoz, A.M. Fernández, A. López-Quílez and J.M. Bellido (2015). Bayesian spatio-temporal approach to identifying fish nurseries by validating persistence areas. *Marine Ecology Progress Series*, **528**: 245–255.

16 I. Paradinas, M. Marín, M.G. Pennino, A. López-Quílez, D. Conesa, D. Barreda, M. González and J.M. Bellido (2016). Identifying the best fishing-suitable areas under the new European discard ban. **Selected as Editor's choice**, *ICES Journal of Marine Science*, **73**(10): 2479–2487.

17 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.

18 I. Paradinas, M. G. Pennino, M. Marín, A. López-Quílez, J. M. Bellido and D. Conesa (2018). Modelling spatialy sampled proportion processes. *RevStat*, **16**(1): 71–86.

19 I. Paradinas, D. Conesa, A. López-Quílez, A. Esteban, L. Martín, J.M. Bellido and M.G. Pennino (2020). Identifying Mullidae species (*M. surmuletus* and *M. barbatus*) persistent hot-spots in the Western Mediterranean Sea. *Marine Ecology Progress Series*, **644**: 173–185.

20 M.G. Pennino, F. Muñoz, D. Conesa, A. López-Quílez, J.M. Bellido (2013). Modelling sensitive elasmobranch habitats. *Journal of Sea Research*, **83**: 209–218.

21 M.G. Pennino, F. Muñoz, D. Conesa, A. López-Quílez, J. M. Bellido (2014). Bayesian spatio-temporal discard model in a demersal trawl fishery. *Journal of Sea Research*, **90**: 44–53.

22 L. Rodrigues, E. Kikuchi, P. Kinas, D. Conesa, M.G. Pennino, F. Barbosa and L. Cardoso (2023). Modelling the distribution of marine fishery resources: where are we? *Fish and Fisheries*, **24**(1), 159–175

23 L. Rodrigues, N. W. Daudt, L.G. Cardoso, P. G. Kinas, D. Conesa and M.G. Pennino (2023). Species distribution modelling in the Southwestern Atlantic Ocean: trends and perspectives. *Ecological Modelling*, **486**, 110514.

24 H. Rue, S. Martino and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B*, 71(2): 319-392.

25 D. Simpson, H. Rue, A. Riebler, T.G. Martins and S.H. Sørbye (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Stat. Sci.* 2017, 32, 1–28.

26 D.J. Spiegelhalter, N.G. Best, B.P. Carlin and A. van der Linde (2002). Bayesian measures of model complesity and fit. *Journal of the Royal Statistical Society: Series B*, 64: 583-616.

27 S. Watanabe (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.*, 11, 3571–3594.