Species distribution modelling

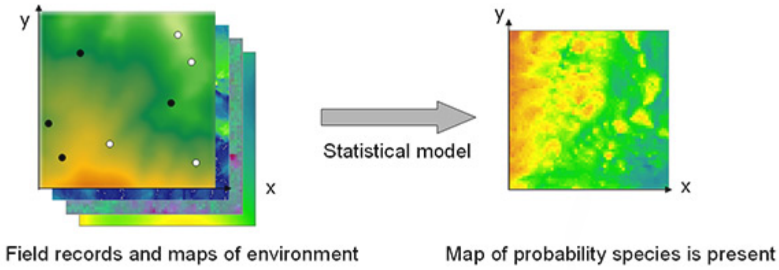
David V. Conesa Guillén

VAlencia BAyesian Research group

Universitat de València

joint work with X. Barber, A. López-Quílez, S. Lladosa, J. Martínez Minaya, F. Muñoz, I. Paradinas, M. G. Pennino

- Species distribution modelling 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.


y
x
Statistical model
y
x
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, prevalence of diseases, etc.

- Most of the available data come from the designed field-based biodiversity monitoring programmes.
- Different types of response variable:
  - Presence/absence of the species,
  - Abundance,
  - Proportion of the species.
- Two situations:
  - non-target species: assume independence between observation locations and the response variable.
  - Target species: only sampling where there is knowledge about the presence of the species (preferential sampling).
- Other extensions:
  - misalignment,
  - Spatio-temporal,
  - Functions of covariates.

1. Statistical modelling of the basic situation
2. Extensions of the basic case
3. Conclusions
4. References

1 Statistical modelling of the basic situation

Basic situation.

A model for the presence/absence of a species (abundance and proportion could also be done similarly)

1. Define a binary random variable as a response variable (presence (1) or absence (0) of the species):

$$Z_i \sim \text{Ber}(\pi_i), i = 1, \ldots, n$$

2. The probability of presence π_i is linked with the linear predictor and the spacial random effect:

$$\text{logit}(\pi_i) = X_i\beta + W_i \, , \; i = 1, \ldots, n$$

The spatial effect W is Gaussian distributed: it is the Latent Gaussian field.

3. A prior distribution for the hyperparameters of W is assigned.

- The resulting spatial model is a latent Gaussian model.
- We could compute approximations for the posterior distributions of the parameters with INLA.
- But this is a continuously indexed Gaussian Field 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

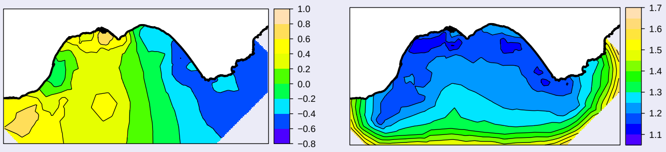
$$\begin{aligned} Z_i &\sim \text{Ber}(\pi_i),\ i = 1, \dots, n \\ \text{logit}(\pi_i) &= X_i \beta + W_i \\ \pi(\beta_j) &\sim N(\mu_{\beta_j}, p_{\beta_j}) \\ W_i &\sim N(0, \mathbf{Q}(\kappa, \tau)) \\ 2\log\kappa &\sim N(\mu_\kappa, p_\kappa) \\ \log\tau &\sim N(\mu_\tau, p_\tau) \end{aligned}$$

- Now the spatial effect depends on two different parameters: κ and τ, which determine the range of the effect and the total variance, respectively.
- Once the inference is performed, the model is used to predict the behaviour in unsampled places: maps of probability.
- Model comparison with two criterions:
  - DIC (Spiegelhalter et al., 2002)
  - CPO (Pettit, 1990)

Distribution of Mediterranean horse mackerel in Gulf of Almería (Muñoz et al., 2013)

- In spite of its low commercial value, mackerel plays an important role in the observed transition zone between the Mediterranean and Atlantic sea.
- Covariates in final model: log(Depth) (neg. effect) and chlorophyll-a (pos. effect).

Spatial effect



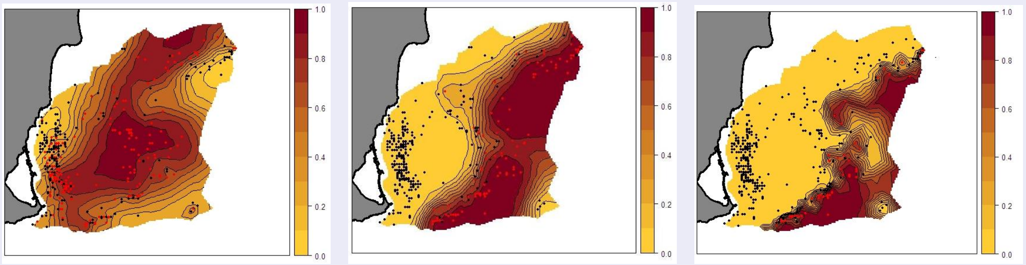
Posterior mean and standard deviation of the spatial effect

- 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 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.

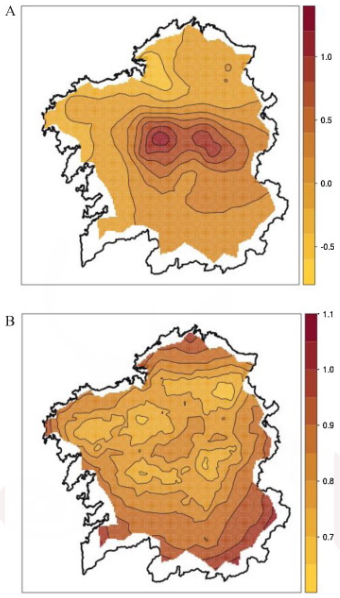
Median of the posterior probability of the presence of elasmobranch species



- 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.

Distribution of prevalence of Bovine paramphistomosis in Galicia (González-Warleta et al., 2013).

- Big concern in Galicia about paramphistomosis in cows and the possibility of infecting humans.
- Interest of detecting the zones of higher prevalence.
- Gain knowledge about the different covariates: most relevant are Temperature and log(slope).
- Spatial component of the fitted model in dairy cows throughout Galicia: Posterior mean (A) and standard deviation (B).



2 Extensions of the basic case

Target species

- The previous model 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 species is known or expected to be high:

  ⟶ fishermen fish in areas where they are likely to find fish.
- This is a clear example of preferential sampling.

Implementation of Preferential sampling as a marked point pattern.

- S unknown target spatial stationary Gaussian process
- The sampling design process (point pattern) depends on S. In our case, $X|S$ is a log-Gaussian Cox Process assuming that the intensity of the Point Process depends on the abundance of the species.
- The abundance observed at those locations, $Y|S$ is as a set of mutually independent Gaussian variates. They can be seen as a noisy version of S.

- The likelihood of the LGCP can be computed in an approximate way, modeling grid cell *counts*
- Both models (points and marks) can be seen as latent Gaussian models, and be jointly modelled with `INLA`.
- As in the non-target species scenario, the underlying spatial model is a continuously indexed Gaussian Field.
- Again, we use the SPDE approach to approximate it as a GRMF.

The preferential model

Point Process model

$$\boldsymbol{x}|\boldsymbol{\eta} \sim \mathcal{P}(\exp(\boldsymbol{\eta}))$$

$$\boldsymbol{\eta} = \mathbf{1}\beta_{0\eta} + \mathbf{X}_{\eta}\boldsymbol{\beta}_{\eta} + \boldsymbol{\theta}$$

$$\pi(\beta_0) \propto 1$$

$$\beta_j \overset{\text{iid}}{\sim} \mathrm{N}(0, 1\text{e-}05)$$

Model for the marks

$$y_i|\kappa_i \overset{\text{iid}}{\sim} \mathrm{N}(\kappa_i, \tau^2)$$

$$\kappa_i = \beta_{0\kappa} + \mathbf{X}_{i\kappa}\boldsymbol{\beta}_{\kappa} + \theta_i$$

$$\pi(\beta_0) \propto 1$$

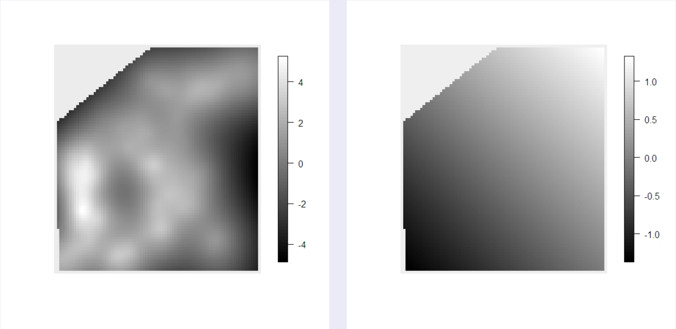
$$\beta_j \overset{\text{iid}}{\sim} \mathrm{N}(0, 1\text{e-}05)$$

⟶ where $\boldsymbol{\theta}$ is a GMRF with the same characteristics of that used in the non-target scenario.

Example: Distribution of European hake in the Gulf of Alicante (Spain).

- Studies about European hake don't take into account preferential sampling.
- CPUE represents the marks and the fishing locations are the point pattern.
- No covariates were included but the possibility of fishermen error was included.

Posterior mean of the spatial effect of the preferential sampling model with fisherman error (left) and (right) the non-preferential sampling model

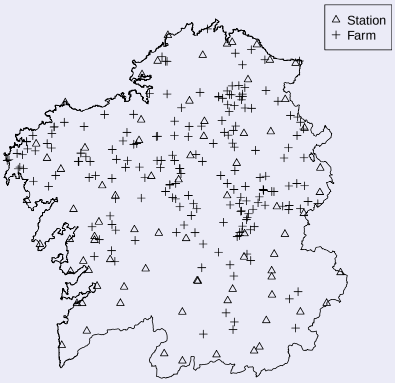


- In the first situation hot spots of high-CPUE density can be appreciated.
- In the other case the spatial component surface is smoother: no particular hot spot is marked.

Misalignment.

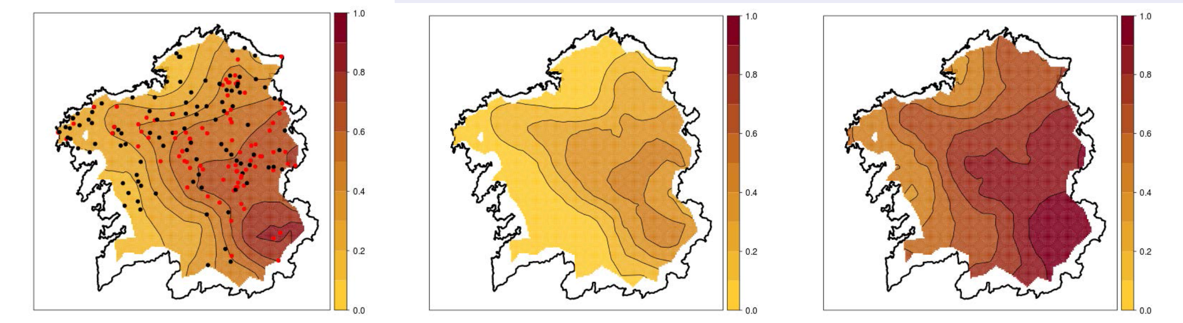
- It appears when measurement values of the covariates are not known at the observed locations nor at those locations where we are going to make predictions.

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



Fasciolosis in Galicia solved not taking into account misalignment

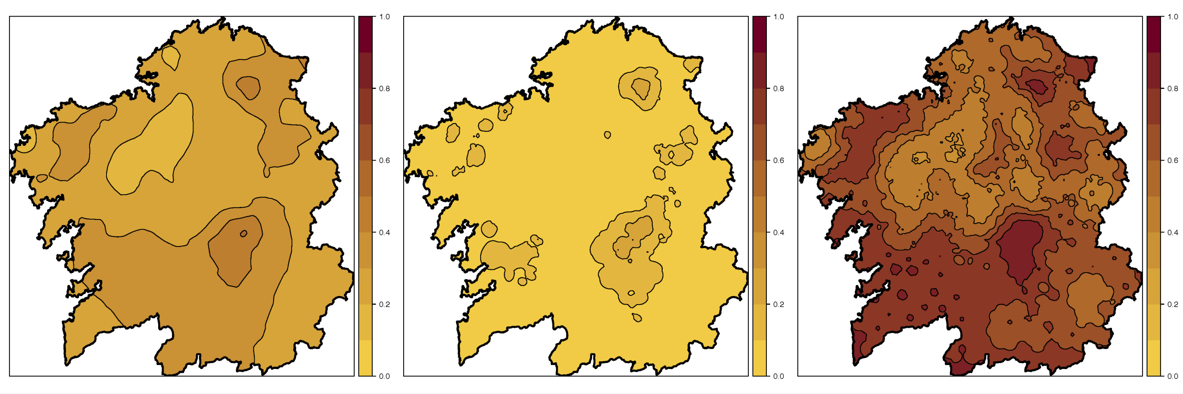
Posterior mean of the probability of occurrence (left) and the first (center) and third (right) quantiles. Red points mean Presence and black points mean Absence



Fasciolosis in Galicia solved taking into account misalignment

- Modelling under uncertainty in the covariates can still be performed using the SPDE approach
- Barber et al. (2015) have used this approach to analyze the presence *fasciolosis* in Galicia.

Posterior mean of the probability of occurrence (left) and the first (center) and third (right) quantiles



Incorporating functions of effects

- Effect of affected locations can modeled through dispersal kernels (non linear functions depending on distances).
- Different kernels (exponential, Cauchy, etc.) can be choosen and the bandwidth must be defined. Exponential:

$$K(d_{il}, h) = \exp\left\{\frac{-|d_{ij}|}{h}\right\}$$

- Another option is using a reverse sigmoidal transformation for the nearest location.

$$f(d_i) = \gamma \frac{\alpha^k}{d_i^{*k} + \alpha^k} \ .$$

- More about this on yesterday poster by Martínez-Minaya et al.
- Smoothing splines can also be incorporated.

Spatio-temporal analysis

- It is also possible to incorporate temporal components (autorregresive, random walks, etc.) to the previous model.
- Example: the persistence over time of abundance hot-spots is key in order to identify nursery areas in fisheries.
- Paradinas et al. (2015) have analyzed persistance by comparing different spatio-temporal models.

Combining extensions

- Two-part spatio-temporal model with shared components:
  - Occurrence → $Y_{st} \sim \text{Ber}(\pi_{st})$
    $\text{logit}(\pi_{st}) = Intc_o + f(depth) + u_{st}$
  - Abundance → $Z_{st} \sim \mathcal{G}a(a_{st}, b_{st})$
    $log(\mu_{st}) = Intc_a + \theta_f f(depth) + \theta_u u_{st}$
    - ★ u_{st} is the spatio-temporal structure
    - ★ θ is a scaling parameter that link the shared component between different linear predictors
    - ★ $f() = RW(2) = x_d - 2x_{d+1} + x_{d+1}$

Spatio-temporal structures

Assuming a geostatistical spatial term (W):

$$W \sim N(0, Q(\kappa, \tau)) \qquad (1)$$

Propose 4 U_{st} structures to infer the fundamental temporal behaviour of the process:

- ↑ Yearly differences

$$U_{st} = w_{s_t} \qquad (2)$$

- ≃ Persistent spatial pattern with uncorrelated intensities

$$\begin{aligned} U_{st} &= w_{st} + v_t \\ v_t &= N(0, \sigma_t) \end{aligned} \qquad (4)$$

- Temporal trend

$$U_{st} = w_{st} + f_t(t) \qquad (3)$$

- Correlation among neighbouring years

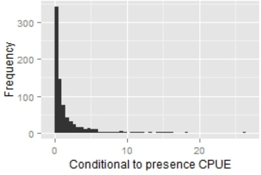
$$\begin{aligned} U_{st} &= r_{st} + w_{st} \\ r_{st} &= \rho_t U_{st-1} \end{aligned} \qquad (5)$$

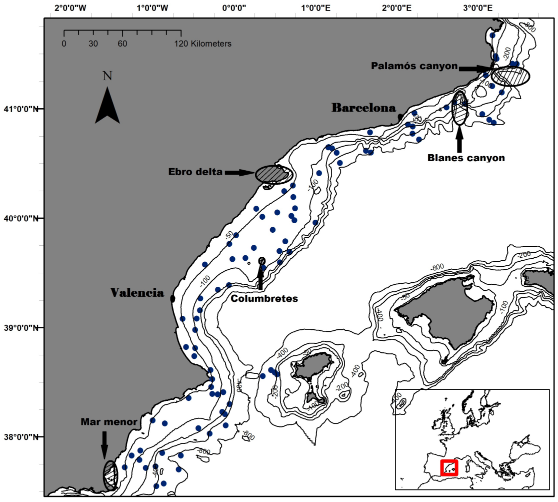
A case study about hake recruitment

- Data:
  - 1048 observations
  - from 2000 to 2012
  - Occurrence

Presence	Absence
758	290

  - Conditional-to-presence Abundance





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

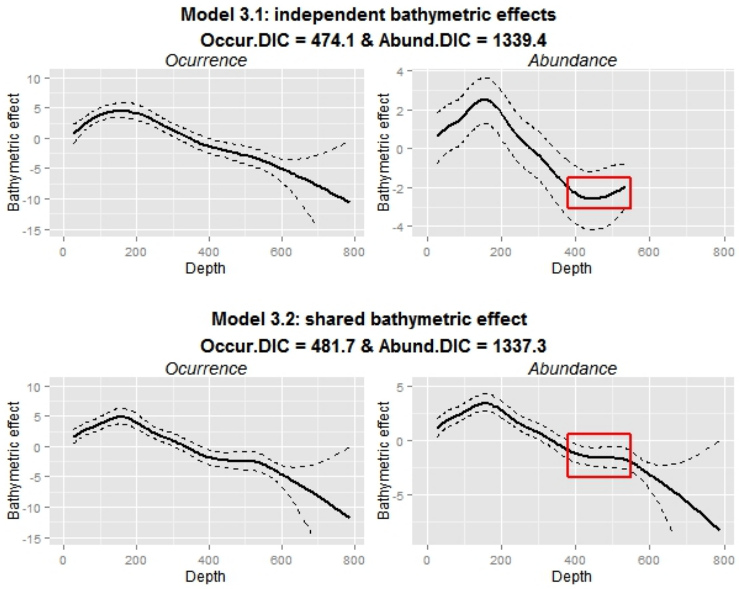
- Every combination of spatio-temporal structures and shared components were compared based on the deviance information criterion (DIC). Here's a summary:

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.

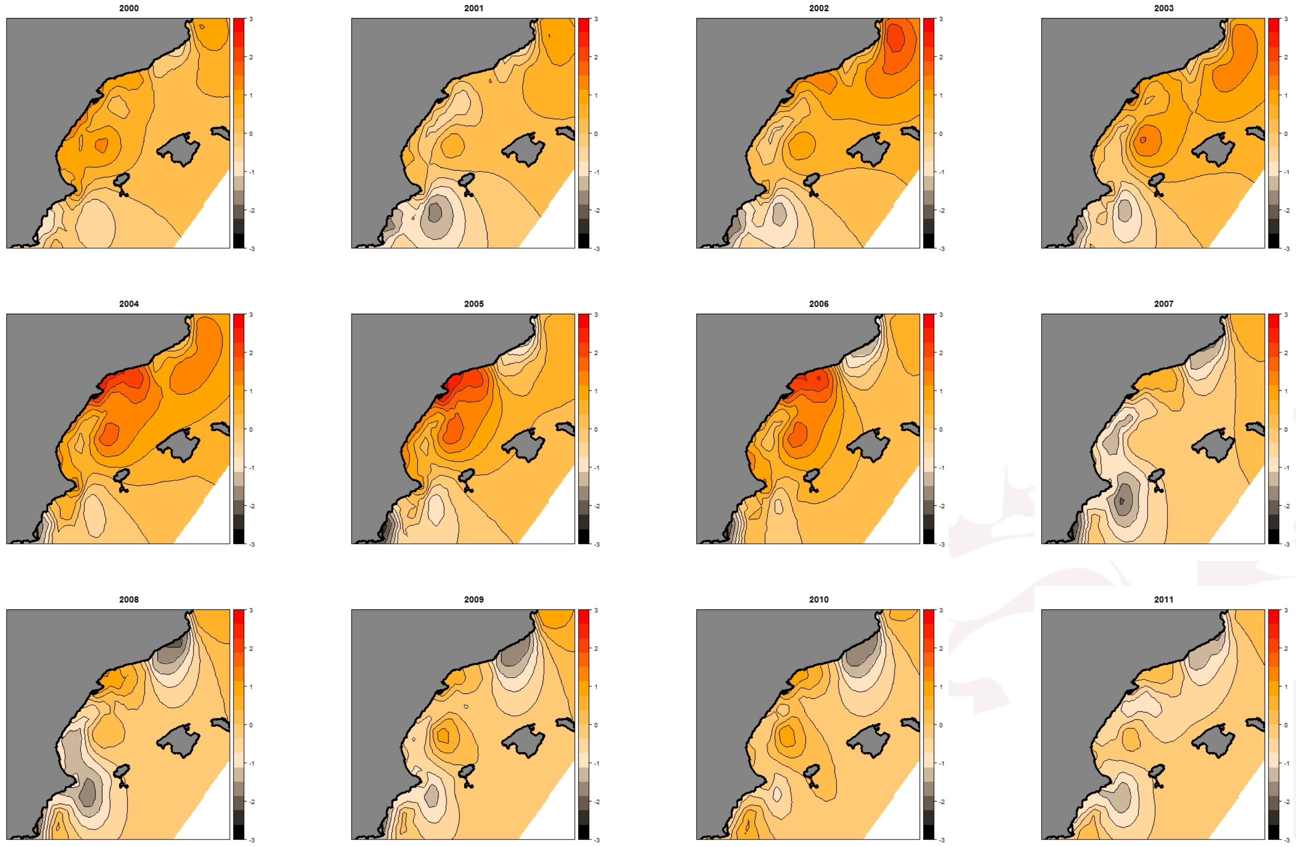
Results: shared bathymetric effect

- 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 spatial effects



3 Conclusions

Conclusions

- Hierarchical Bayesian modelling can be a really useful tool for analysing species distribution models.
- INLA can also be very convenient as it is fast and provides good results (probably not the best ones, but the first ones in your analysis).
- Many extensions can be handled, but many others will need other methods like MCMC or ABC.
- Still room for improvement.

4 References

1. 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*, in press.
2. 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.
3. 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.
4. J. Martínez-Minaya, D. Conesa, A. López-Quílez and A. Vicent (2015). Climatic distribution of citrus black spot caused by Phyllosticta citricarpa. A historical analysis of disease spread in South Africa. *European Journal of Plant Pathology*, **143**, 69–83.
5. 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.

6. I. Paradinas, M. G. Pennino, F. Muñoz, D. Conesa, A. M. Fernández, A. López-Quílez, J. M. Bellido (2015). A Bayesian approach to identifying fish nurseries. *Marine Ecology Progress Series*, **528**: 245–255.
7. M.G. Pennino, F. Muñoz, D. Conesa, A. López-Quílez, J.M. Bellido (2014). Modelling sensitive elasmobranch habitats. *Journal of Sea Research*, **83**: 209–218.
8. 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.
9. 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.
10. 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.

Thank you