# R-INLA: A FLEXIBLE TOOL FOR IMPLEMENTING BAYESIAN REGRESSION MODELS

A. Fuster Alonso*[1], S. Cerviño[2], D. Conesa[1], M. Cousido-Rocha[2], F. Izquierdo[2], M.G. Pennino[2]

[1]Universitat de València

[2]Instituto Español de Oceanografía (IEO-CSIC)

VNIVERSITAT DE VALÈNCIA

INSTITUTO ESPAÑOL DE OCEANOGRAFÍA

CSIC
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

VIII XORNADA DE USUARIOS DE R EN GALICIA

# Index

# Introduction

Fishing has been a very important source of food, but also a source of employment and economic benefits.

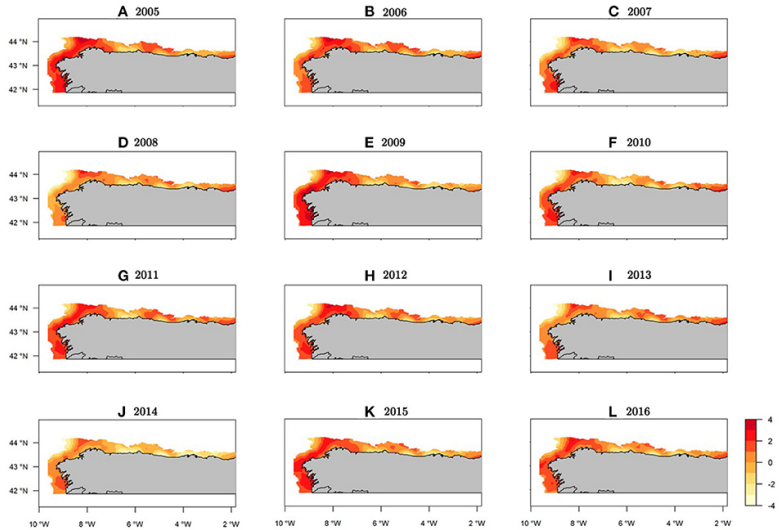
# Introduction

BUT This economic perspective has caused failures in fishery management systems. In 2020 a 34.2 % of all marine fish stocks monitores by FAO are currently being overfishing.

# Introduction

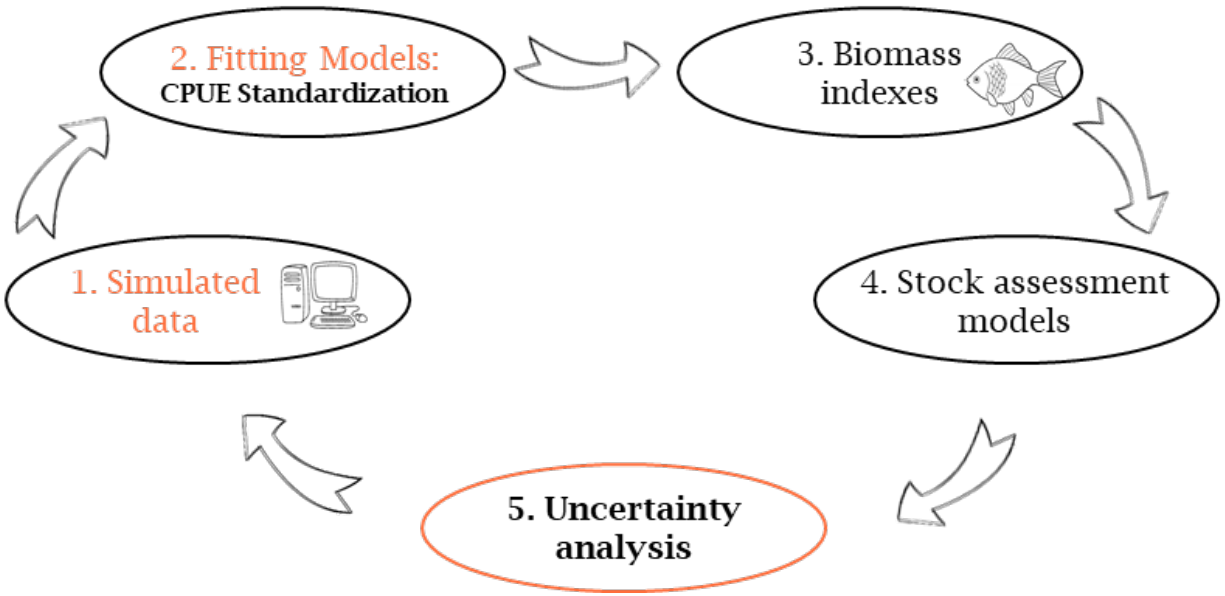
## Solution

From this problem arises the necessity of improving scientific advice quality for fisheries management, but how? By using statistical modeling to describe the spatio-temporal behavior of the target species biomass.



# Our project

This work is a part of an ongoing project that will focus on the analysis of uncertainty associated with relative biomass indices derived from the standardization of catch per unit effort (CPUE) indices.



# How to model CPUE indices

CPUE indices are considered as one of the main information sources used in fish stock assessment models (Zou et al., 2019). There are many ways for the standardization of these indices:

- Generalized lineal models (GLMs).
- Generalized additive models (GAMs).
- Geostatistical models.

## Bayesian inference

Also, when performing statistical inference there are two approaches: the frequentist and the Bayesian one. The main difference between them is how they interpret probability. In this work, we will use Bayesian inference to fit the models, but why?

# Bayesian methodology

Bayesian inference relies on Bayes' theorem (1) to estimate the probability of the model parameters ($\theta$) given the observed data:

$$\pi(\theta|\text{data}) \propto \pi(\text{data}|\theta)\,\pi(\theta) \tag{1}$$

- $\pi(\theta|\text{data})$ is the posterior distribution of the model parameters.
- $\pi(\text{data}|\theta)$ is the likelihood of the model.
- $\pi(\theta)$ is the prior distribution of the model parameters.

# Bayesian methodology

## Problem

Bayesian inference on complex models results in tricky analytical expressions to obtain a posterior distributions $\pi(\theta|\text{data})$.

$$\int_{\Theta} \pi(\text{data}|\theta)\pi(\theta)d\theta \tag{2}$$

## Solution

In order to estimate this expression computational approaches, as Markov chain Monte Carlo (MCMC) methods, and numerical approximations, as Integrated Nested Laplace Approximation (INLA) methodology, are needed.

# Bayesian methodology

Bayesian approach has become more popular the last decades. Some of the reasons underneath are:

- The posteriori distribution $\pi(\theta|data)$ provide all the information about the parameters.
- The ease with which prior information could be incorporated $\pi(\theta)$. When this knowledge is poor, vague priors are assumed so that the posterior.
- A extensive number of complex models could be fitted efficiently and fast compared to another classic techniques, because of the important advances in computational statistics in the last decades.

# Bayesian methodology

Bayesian approach has become more popular the last decades. Some of the reasons underneath are:

- The posteriori distribution $\pi(\theta|data)$ provide all the information about the parameters.
- The ease with which prior information could be incorporated $\pi(\theta)$. When this knowledge is poor, vague priors are assumed so that the posterior.
- A extensive number of complex models could be fitted efficiently and fast compared to another classic techniques, because of the important advances in computational statistics in the last decades.

# Bayesian methodology

Bayesian approach has become more popular the last decades. Some of the reasons underneath are:

- The posteriori distribution $\pi(\theta|data)$ provide all the information about the parameters.
- The ease with which prior information could be incorporated $\pi(\theta)$. When this knowledge is poor, vague priors are assumed so that the posterior.
- A extensive number of complex models could be fitted efficiently and fast compared to another classic techniques, because of the important advances in computational statistics in the last decades.

# MCMC methods

- Tradicionally, some computational approaches have been proposed to estimate the posterior distributions with Markov chain Monte Carlo (MCMC) methods, implemented in softwares such as WinBUGS (Lunn et al., 2000).
- MCMC yields simulations from the set of model parameters, i.e., a multivariate distribution. Consequently, we obtain the joint posterior distribution.
- It could require several simulations for valid inference.
- We may be concerned with a single parameter or a subset of parameters.
- In addition, we have to verify that the burn-in period is finished, i.e. that we have achieved the posterior distribution.

# Integrated Nested Laplace Approximation

- Could happened that we only needed marginal inference on some parameters, i.e., we could need $\pi(\theta_i|\text{data})$.
- Rue, Martino and Chopin (2009) provided a new way to approximate marginal posterior distributions in the Bayesian context. This is the situation here, as we have to deal with many univariate distributions.
- INLA is computationally faster due to because we do not need to simulate the posteriori, only approximate it numerically.
- The only assumption needed is that the statistical model is a Latent Gaussian Model (LGM). In fact, most statistical models are LGM's, e.g., spatial models, temporal models, spline models, GLM, GLZ, survival models, joint models, etc.

# R-INLA package

**R-INLA**

available from https://www.r-inla.org/ is a package in R that do approximate Bayesian inference for Latent Gaussian Models.

# R-INLA resources

- https://becarioprecario.github.io/ - Virgilio Gómez Rubio, Titular professor at the University of Castilla-La Mancha.
- Blangiardo, M. and Cameletti, M. (2015). Spatial and spatio-temporal Bayesian models with R-INLA. John Wiley and Sons.
- McElreath, R. (2018). Statistical rethinking: A Bayesian course with examples in R and Stan. Chapman and Hall/CRC.
- Wang, X., Yue, Y. and Faraway, J. J. (2018). Bayesian regression modeling with INLA. Chapman and Hall/CRC.
- Zuur, A. F., Ieno, E. N. and Saveliev, A. A. (2017). Spatial, temporal and spatial-temporal ecological data analysis with r-inla. Highland Statistics Ltd, 1.

# R-INLA resources

- A gentle INLA tutorial
- Haakon Bakka's website
- Julian Faraway website
- R-INLA discussion group

# Generalized additive model

CPUE indices can be modeled using smoothing functions:

## GAM

$$\begin{aligned} \text{CPUE}_i &\sim \text{Gamma}(\mu_i, \phi) \quad i = 1, .., n\,, \\ log(\mu_i) &= \beta_0 + f(B_i)\,, \end{aligned} \tag{3}$$

In `R-INLA`, these smooting functions are considered as latents fields:

- rw1, random walk of order 1.
- rw2, random walk of order 2.

# Spatial model

CPUE indices can also be modeled using a spatial term in order to incorporate spatial correlation.

## Spatial model

$$\begin{aligned} \text{CPUE}(s_i) &\sim \text{Gamma}(\mu(s_i), \phi) \quad i = 1, .., n\,, \\ log(\mu(s_i)) &= \beta_0 + f(B(s_i)) + u(s_i)\,, \\ u(s_i) &\sim \text{GMRF}(0, \Sigma)\,, \end{aligned} \tag{4}$$

Where $u(s_i)$, note againt that, is a latent field. PROBLEM: $u(s_i)$ is a GRF but not a GMRF as required by `R-INLA`. Then, Stochastic partial differential equation (SPDE) approach link both

# Hurdle model

Hurdle model consists of modeling two process: (1) a binary part to fit the presence/absence of the target specie and (2) a continuous part to model the intensity when the response is non-zero (Izquierdo et al., 2021):

## Hurdle model

$$
\begin{aligned}
\mathrm{Y}(\mathrm{s}_i) &\sim \mathrm{Bernoulli}(\pi(\mathrm{s}_i)) \quad \mathrm{i} = 1, .., \mathrm{n}\,, \\
\mathrm{CPUE}(\mathrm{s}_i) &\sim \mathrm{Gamma}(\mu(\mathrm{s}_i), \phi) \quad \mathrm{i} = 1, .., \mathrm{n}\,, \\
\mathrm{logit}(\pi(\mathrm{s}_i)) &= \beta_{0\mathrm{Y}} + f(B(\mathrm{s}_i)) + u_{\mathrm{Y}}(\mathrm{s}_i)\,, \\
u_{\mathrm{Y}}(s_i) &\sim \mathrm{GMRF}(0, \Sigma_{\mathrm{Y}})\,, \\
\log(\mu(\mathrm{s}_i)) &= \beta_{0\mathrm{CPUE}} + \alpha f(B(\mathrm{s}_i)) + u_{\mathrm{CPUE}}(\mathrm{s}_i)\,, \\
u_{\mathrm{CPUE}}(s_i) &\sim \mathrm{GMRF}(0, \Sigma_{\mathrm{CPUE}})\,,
\end{aligned}
\tag{5}
$$

# Code hurdle models

## Formula

```
formula.hurdle <- y ~ -1 + Intercept +

    f(bath.ber, model = "rw1") +

    f(bath.con, copy="bath.ber",fixed = F) +

    f(i.ber, model = spde) +

    f(i.con, model = spde)
```

# Code hurdle models

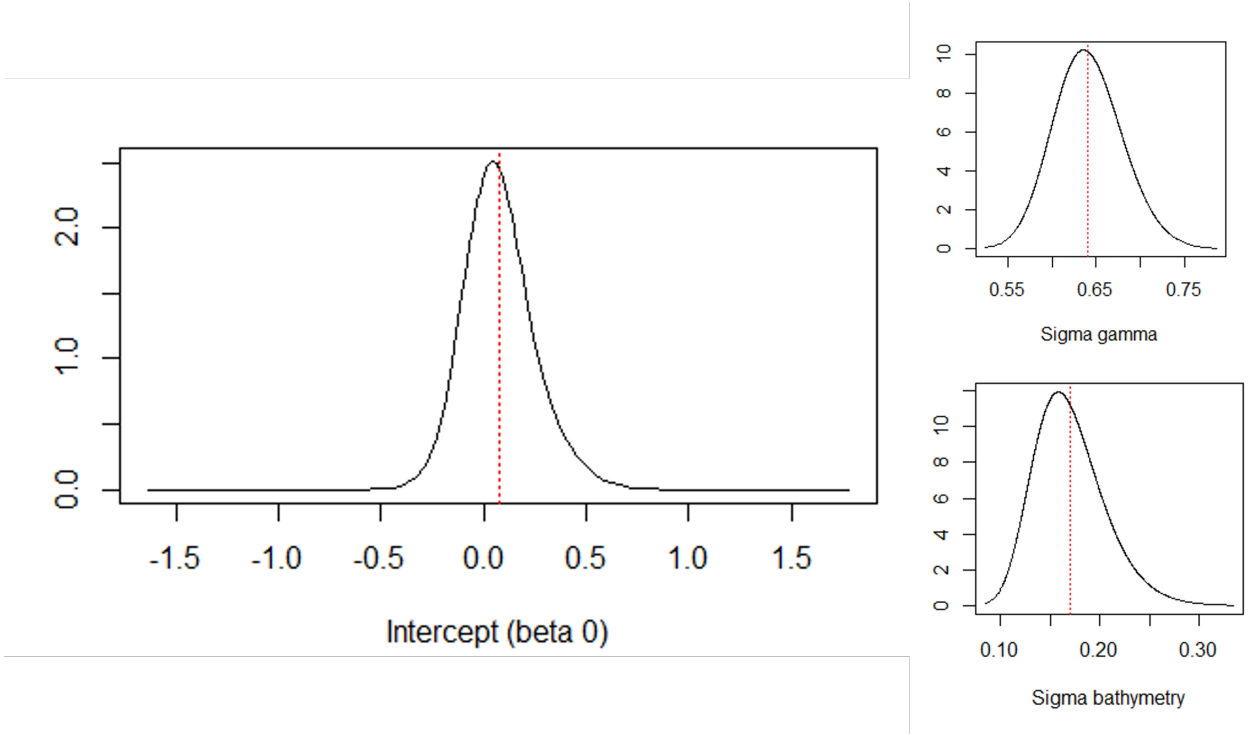
## Run the model: *inla*

```
set.seed(12345)
model.hurdle <- inla(
  formula.hurdle,
  family = c('binomial', "gamma"),
  data = inla.stack.data(Stack.final),
  control.compute = list(
    dic = TRUE,
    cpo = TRUE,
    waic = T,
    config = FALSE),
  control.predictor = list(A = inla.stack.A(
    Stack.final), compute = TRUE))

summary(model.hurdle)
```

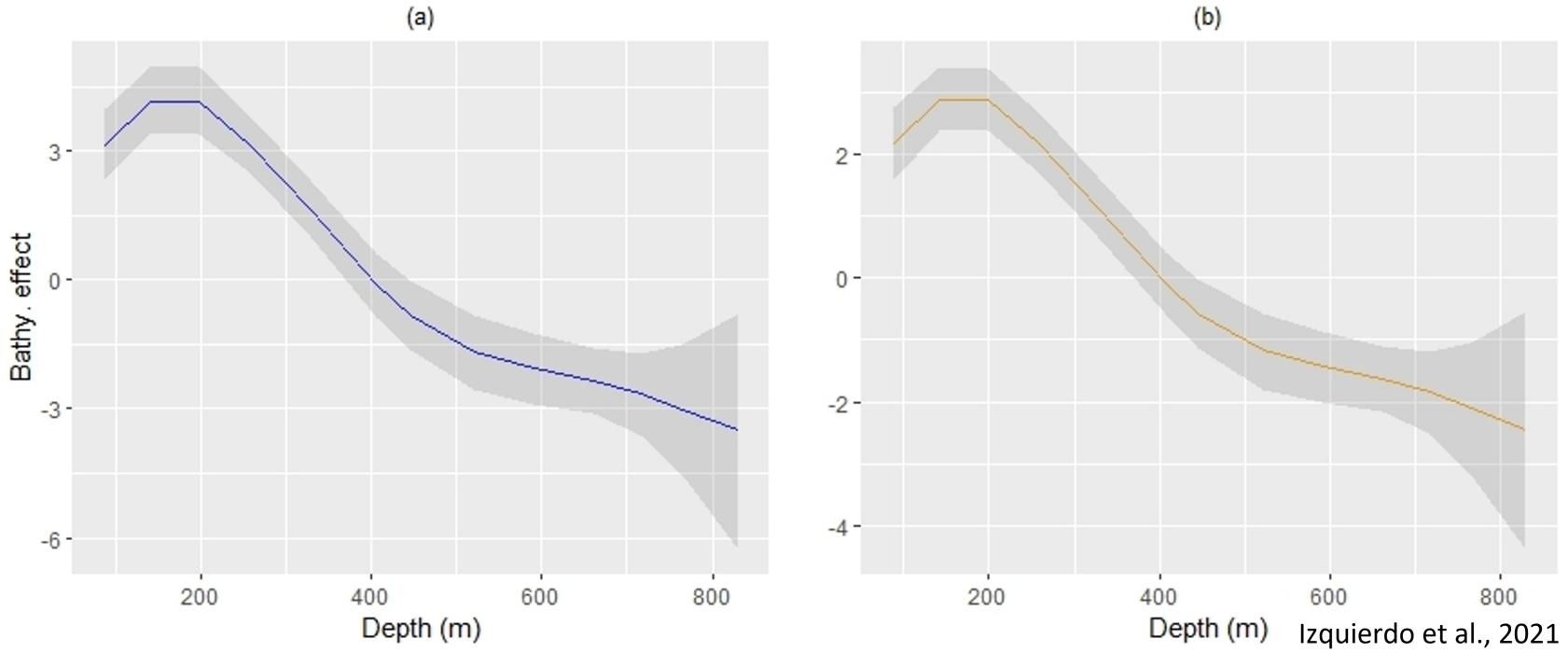
# Examples of results

## Posterior distributions fixed effects and hyperparameters



# Examples of results

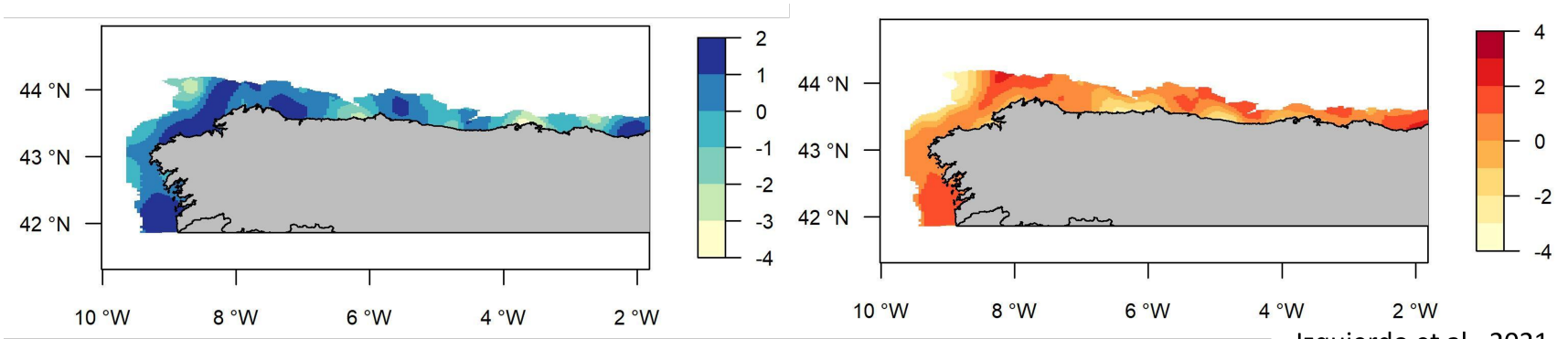
## Smoothing bathymetry




Izquierdo et al., 2021

# Examples of results

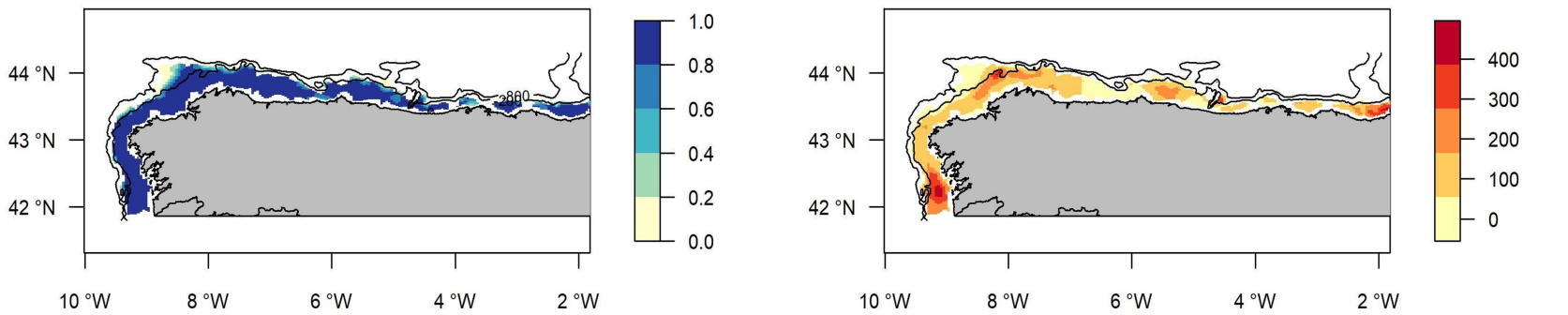
**Mean of the posteriori distribution for the spatial effect**




Izquierdo et al., 2021

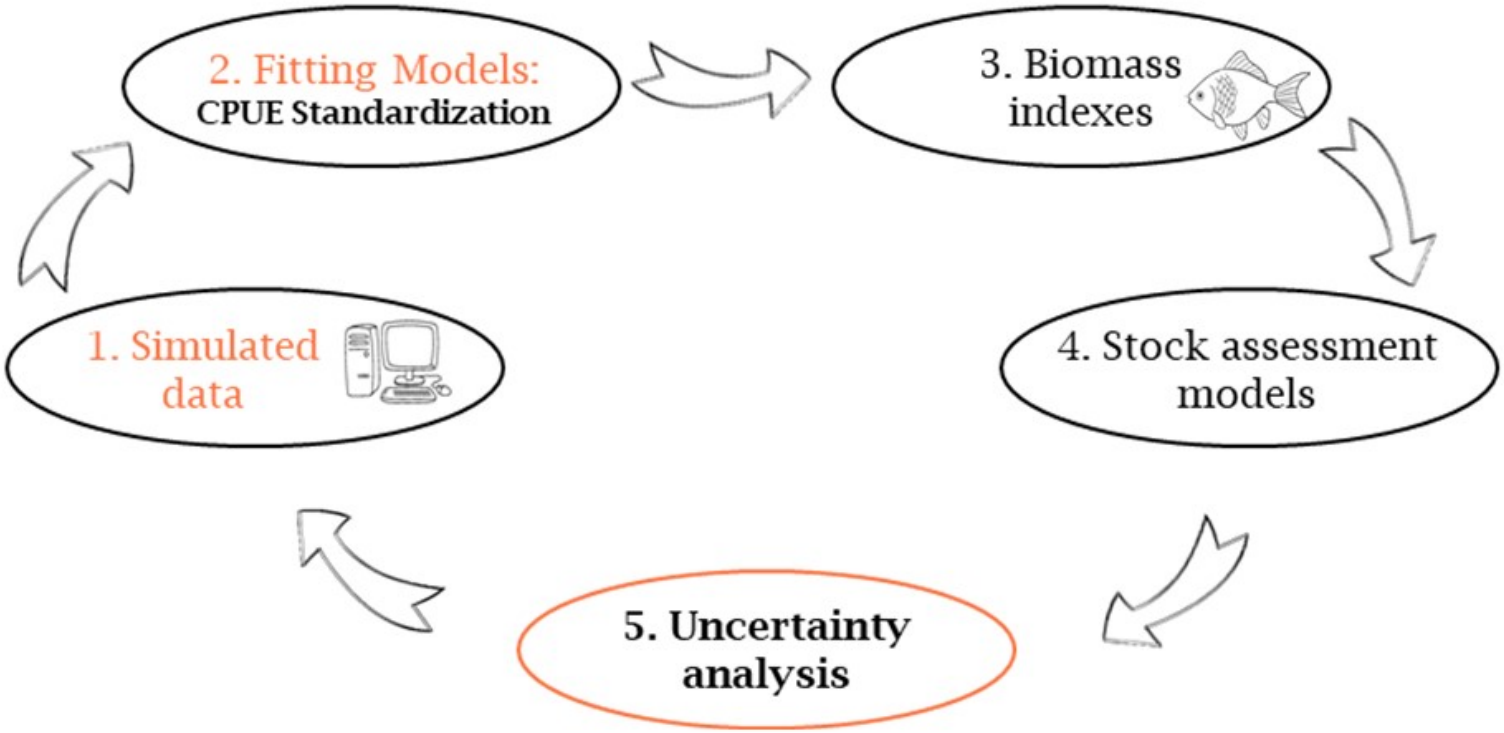
# Examples of results

## Median of the posterior predictive distribution




Izquierdo et al., 2021

# Our project


2. Fitting Models:
CPUE Standardization
3. Biomass indexes
4. Stock assessment models
5. Uncertainty analysis
1. Simulated data

# Conclusions

- Statistical modeling can help to solve problems related to fisheries management.
- Our project proposes the standardization of CPUE indices to obtain relative biomass indices, which requires the adjustment of different statistical models.
- There are many ways to fit these models, in our case Bayesian inference is performed on the model parameters with the software `R-INLA` (Rue et al., 2009).

# Conclusions

- Statistical modeling can help to solve problems related to fisheries management.
- Our project proposes the standardization of CPUE indices to obtain relative biomass indices, which requires the adjustment of different statistical models.
- There are many ways to fit these models, in our case Bayesian inference is performed on the model parameters with the software `R-INLA` (Rue et al., 2009).

# Conclusions

- Statistical modeling can help to solve problems related to fisheries management.
- Our project proposes the standardization of CPUE indices to obtain relative biomass indices, which requires the adjustment of different statistical models.
- There are many ways to fit these models, in our case Bayesian inference is performed on the model parameters with the software `R-INLA` (Rue et al., 2009).

# Conclusions

- R-INLA is a fast, robust and accurate tool for approximate Bayesian inference, making it a good alternative to MCMC methods.
- Some advantages of R-INLA are the low computational cost and the great variety of statistical models available.
- This work review the use of R-INLA through some examples (GAM, spatial model and hurdle model).

# Conclusions

- R-INLA is a fast, robust and accurate tool for approximate Bayesian inference, making it a good alternative to MCMC methods.
- Some advantages of R-INLA are the low computational cost and the great variety of statistical models available.
- This work review the use of R-INLA through some examples (GAM, spatial model and hurdle model).

# Conclusions

- `R-INLA` is a fast, robust and accurate tool for approximate Bayesian inference, making it a good alternative to MCMC methods.
- Some advantages of `R-INLA` are the low computational cost and the great variety of statistical models available.
- This work review the use of `R-INLA` through some examples (GAM, spatial model and hurdle model).

# References


- https://becarioprecario.github.io/ - Virgilio Gómez Rubio, Titular professor at the University of Castilla-La Mancha.
- Izquierdo, F., Paradinas, I., Cerviño, S., Conesa, D., Alonso-Fernández, A., Velasco, F., Preciad I., Punzón A., Saborido-Rey F. and Pennino, M. G. (2021). Spatio-temporal assessment of the European hake (Merluccius merluccius) recruits in the northern Iberian Peninsula. Frontiers in Marine Science, 8, 1.
- Lunn, D. J., Thomas, A., Best, N. and Spiegelhalter, D. (2000). WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and computing, 10(4), 325-337.
- Rue, Havard, Sara Martino, and Nicolas 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): 319–92.

# References


- Krainski, E., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D. and Rue, H. (2018). Advanced spatial modeling with stochastic partial differential equations using R and INLA. Chapman and Hall/CRC.
- Zhou, S., Campbell, R. A. and Hoyle, S. D. (2019). Catch per unit effort standardization using spatio-temporal models for Australia's Eastern Tuna and Billfish Fishery. ICES Journal of Marine Science, 76(6), 1489-1504.
- Zuur, A. F., Ieno, E. N. and Saveliev, A. A. (2017). Spatial, temporal and spatial-temporal ecological data analysis with r-inla. Highland Statistics Ltd, 1.
- GitHub: Models in `R-INLA`

# Thanks.

Contact e-mail: **alba.fuster1398@gmail.com**