Ecological Informatics 80 (2024) 102542

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

A shiny R app for spatial analysis of species distribution models

Mario Figueira *, David Conesa, Antonio López-Quílez

University of Valencia, Spain

ARTICLE INFO

Keywords:
INLA
Species distribution models
Geostatistics
Preferential models
Shiny

ABSTRACT

In ecology, Species Distribution Models (SDMs) are a statistical tool whose use has expanded considerably over the last two decades. As their use has grown, so has the complexity of the data analysed and the structures of the models used.

This has led to the development of various tools to facilitate the incorporation and use of these new data and statistical methodologies, mostly embodied in new R packages and Shiny applications that allow different types of SDMs to be solved. However, the Integrated Nested Laplace Approximation (INLA) approach, which has become increasingly popular in the field of ecological sciences, has not yet been integrated into an application that can synthesise the complexity of its code into a user-friendly interface for continuous spatial modelling.

To overcome this shortcoming, we present in this paper a novel application, called BAYSPINS (BAYesian SPatial INla for SDMs), which allows the use of INLA for those who are not very experienced, or for those who are experienced and prefer a tool that allows them to carry out an initial analysis quickly, avoiding the process of writing code.

BAYSPINS allows both geostatistical and preferential modelling, as well as a mixture of the two. It integrates the complex and hard coded SPDE-FEM (Stochastic Partial Differential Equation, along with the Finite Elements Method) approach to perform continuous spatial analysis with a visual interface. It also allows the use of default settings that automate the process or the customisation of a large number of elements that drive the modelling process. In this way, quick initial evaluations or more rigorous studies of the data provided by the user can be carried out, depending on the user's skill and understanding of the fundamentals underpinning the application.

1. Introduction

In ecology, statistical tools are widely used, with particular emphasis on species distribution models (SDMs) for the spatial analysis of species-environment relationships. Due to their wide and intensive use, SDMs have become one of the main tools for the inference, prediction and projection of species occurrence and abundance distributions (Fletcher and Fortin, 2019; Guisan et al., 2017; Ovaskainen and Abrego, 2020).

The use of large databases (Pyšek et al., 2017), data from geographic information systems (Yu et al., 2019), and tracking systems (Abrahms et al., 2021) is becoming increasingly common. This leads to an increase in the complexity of computational tools and statistical methods for SDMs. To facilitate their implementation, a large number of R packages and Shiny applications have been developed to solve specific problems and support ecological studies. To provide a brief overview, here are some examples in the field of SDMs: the WALLACE and WALLACE 2 R packages for species niche/distribution modelling (Kass et al., 2018; Kass et al., 2023); the ESDM R package, which allows the creation of ensembles of SDM predictions onto a single base geometry (Woodman et al., 2019); and the NTBOX R package, which permits users to perform ecological niche modelling in a fast and straightforward manner (Osorio-Olvera et al., 2020), all of which are built as a Shiny application (Chang et al., 2023) to provide a graphical user interface. Other example of software implementation is Jung (2023), which allows to use different approaches to perform integrated species distribution models.

The Bayesian paradigm can be used to fit SDMs. In particular, one way to implement these models is through hierarchical Bayesian models (Martínez-Minaya et al., 2018), which can be solved using the Integrated Nested Laplace Approximation (INLA) methodology, (Rue et al., 2009, 2017). The Stochastic Partial Differential Equations and Finite Element Methods (SPDE-FEM) methodology (Bakka et al., 2018; Krainski et al., 2018; Lindgren et al., 2011) implemented within the R-INLA environment, is used to analyse the spatial process. Nevertheless, the use of the entire INLA methodology together with the SPDE-FEM approach can

* Corresponding author.
E-mail address: Mario.Figueira@uv.es (M. Figueira).

https://doi.org/10.1016/j.ecoinf.2024.102542
Received 22 September 2023; Received in revised form 22 February 2024; Accepted 22 February 2024
Available online 1 March 2024

sometimes be challenging for unfamiliar users.

In this paper, we present BAYSPINS, a Shiny application designed to address SDMs in a Bayesian spatial context. What distinguishes BAYSPINS from its counterparts is its unique integration of the INLA methodology, specifically tailored for geostatistical challenges. While there are applications that use the INLA approach, they predominantly cater to areal data analysis in epidemiology (Adin et al., 2019; Moraga, 2017). While the primary intent of BAYSPINS is to serve as a streamlined and flexible solution for SDM, its utility extends beyond this, encompassing any field where geostatistical, log-Gaussian Cox process (LGCP), preferential, or mixture models are relevant. This versatility allows it to be useful in disciplines such as environmental sciences (Huang et al., 2017), spatial econometrics, and geospatial health (Moraga et al., 2021), among others.

In fact, our motivation to develop an application based on INLA lays in its ability to accommodate covariate effects in the construction of all these complex models, together with the remarkable speed with which the data ensemble and fitting process are executed. Thanks to INLA's implementation of the SPDE-FEM methodology, geostatistical models (Diggle et al., 1998), preferential models (Diggle et al., 2010) and a mixture of both (Figueira et al., 2023b) can be fitted with high accuracy and computational power, as well as with LGCP models for point processes or presence-only data. BAYSPINS, with its simple visual interface, can be used to support research and for rapid spatial evaluation of ecological data distributed over a continuous space, avoiding the complexity of building such models in the R-INLA environment. In addition, the application enables the updating of prior distributions, allowing users to set up prior information on parameters and hyperparameters.

2. Application design

The design of BAYSPINS focuses on allowing the user to solve complex models without the need for in-depth knowledge of code or fundamentals, but also has advanced options that allow the user to set up more complicated configurations (the scheme of the application's structure can be seen in Fig. 1). To this end, the structure of the application (hereinafter app) is divided into the following tabs, which will be described in detail in the remaining sections of the paper:

1. **Presentation**. This tab provides a brief introduction to the app, the models that can be applied, and the type of data that can be analysed, together with a short glossary of terminology.
2. **Simulation**. From this tab, the user can set up simulations of data with the structure of an SDM. In particular, it is possible to simulate from Gaussian or Gamma distributions and to include a predictor consisting of an intercept, a spatially defined covariate that can have different structures for the effect, together with a spatially structured random effect. As usual, simulation allows the app to be tested and operated in a controlled environment.
3. **Uploading data**. This section allows the loading of the main data for the analysis as well as other data related to the covariates. This second data frame, allows users to establish additional information for submodels of the covariates or directly identify the required prediction locations together with the values of the explanatory variables in them.
4. **Model fitting**. This section presents the inference models for spatial data. The four available model structures are: (i) an *independent model*, (ii) a log-*Gaussian Cox process model*, (iii) a *preferential model* and (iv) a *mixture model*, which we will use depending on whether the data come from an independent, preferential sampling or a mixture of data from different samplers.
   (a) **Independent modelling**. The independent or geostatistical model allows the user to analyse data with a spatial dependency structure. This model is appropriate for situations where the data under analysis exhibits spatial structure and lacks a discernible dependence on the sampling process used to acquire the data. From this tab, users can set up the geostatistical model to be used to analyse their data, selecting the covariates from the loaded data frame, as well as whether or not to include other predetermined components: the intercept and the spatially structured random effect.
   (b) **Log-Gaussian Cox Process modelling**. The second tab for the models we examine focuses on the Log-Gaussian Cox Process (LGCP) model. The model analyses the spatial patterns of sampling locations and allows the examination of presence-only data. The LGCP model permits the user to infer the intensity of the underlying process responsible for generating presences or sampling locations. This modelling tab provides the user with the necessary tools to analyse the process responsible for the generation of sample locations. There are numerous elements that enable the user to tailor the statistical analysis, including the selection of explanatory variables to be included and their effect structure, as well as the prior distribution for these effects.
   (c) **Preferential modelling**. The preferential model allows the joint analysis of the response variable and the locations. This model is applicable when elements or variables that influence the data also affect the sampling process, thus sharing information between the two. This, in turn, improves the overall inference and prediction capabilities. Similar to the independent model tab, the preferential model tab allows the user to set the conditions

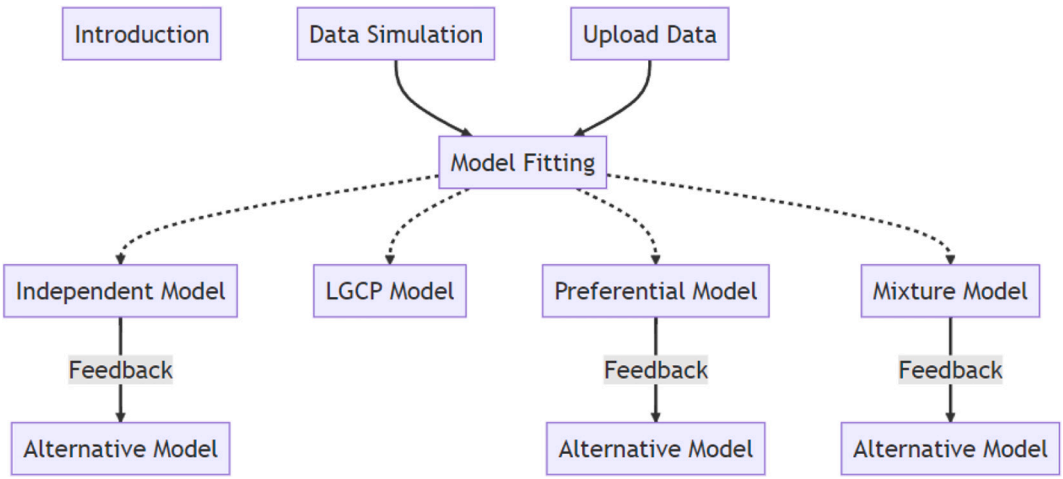


Fig. 1. General scheme of the app structure.

for the analysis by selecting the components of the model as well as the distribution for the likelihood.

(d) **Mixture modelling**. The mixture model is a combined framework that facilitates the examination of various dependency patterns in the location generation process. It bears a close resemblance in structure to the preferential model, offering the ability to distinguish between various sampling processes. Consequently, it allows the user to specify the dependence between the data and the specific sampling processes through which they were collected. This concluding tab gives the user the flexibility to configure a mixture model, enabling the analysis of data derived from both independent and preferential sampling scenarios.

Once the modelling is done, these results can be used to provide feedback or to perform a sequential learning process for a new dataset from the same geostatistical phenomenon, which could be modelled by one of the three proposed model structures, thanks to the control options available in INLA and configurable in the app.

Finally, for more advanced users, BAYSPINS allows the download of the structure of the data, the model formulae and the INLA fitting code. This is done through an input panel in the user interface, where it is possible to specify when the model information is to be saved and the path where it is be saved.

3. Presentation and GitHub repository

As can be seen in Fig. 2, the presentation tab briefly highlights the motivation and context for the development of the app, which is a summary of the abstract of this article.

The app itself is available in the following GitHub repository: https://github.com/MarioFigueiraP/ShinyAppSpatialModelFeedback. The repository contains valuable information for users, providing guidance on how to use the app and fostering a deeper understanding of the modelling structures inherent in the app. It also houses example data files, along with instructions for setting up the app with the required R package dependencies. The repository also contains data files specifically prepared for testing the app. The supplementary material provides a brief introduction to INLA and explains the structure of the models presented in the paper.

The GitHub repository contains three versions of the app: a stable or main version with the lastest INLA update (with the dependence on fmesher library for mesh creation), an older version and, finally, the latest available version for the app. Through the GitHub page, users can request changes, suggest new implementations or report bugs.

4. Simulation

In order to start learning about the app and also to allow users to try out different modellings, the simulation tab has been created to create simulated datasets. Below is a brief description of the simulation process.

SDM datasets usually consist of three types of variables: the response, the covariates, and the locations where the data have been observed (typically represented by their geographic coordinates). For the sake of simplicity, the app simulates from a model that includes an intercept, a function of a covariate (selected by the user), and a spatial structure, and is formulated as follows:

$$\begin{aligned} Y(s) &\sim \text{Dist}(\mu(s), \phi), \\ g(\mu(s)) &= \beta_0 + f(X(s)) + U(s), \end{aligned} \tag{1}$$

where $Y(s)$ represents the value of the response at location s following a probability distribution with mean $\mu(s)$ and dispersion ϕ (at the present moment the distribution can be either Gaussian or Gamma, depending on user preferences, but we plan to include others in the near future); the mean $\mu(s)$ is linked to the predictor by the link function g (the identity for the Gaussian and the *log* function for the Gamma distribution); β_0 is the intercept; $f(X)$ is a function representing the relationship with one covariate; and finally, $U(s)$ refers to the spatial structure.

Simulating from Eq. (1) is done by simulating from the spatial structure, treating it as a Gaussian Markov Random Field, taking the values set by the user for the intercept and the function of the covariate and finally simulating from the corresponding probability distribution with the dispersion (or variance, depending on the distribution chosen) also provided by the user.

Given that simulating an SDM is like mimicking the real distribution of a species, note that simulating locations from the simulated distribution imitates the usual sampling process (e.g. determining the locations where we would measure the species). From the app, it is possible to perform the independent and preferential sampling mentioned above, but also a mixture of the two. In the independent case, only the number of samples and the seed of the process are needed, but for the dependent sampling we also need the preferential factor of the sampling. In the case of the mixture, we also need the proportion of each type of sampling. The necessary panels are presented in Fig. 3, together with the output of the simulation.

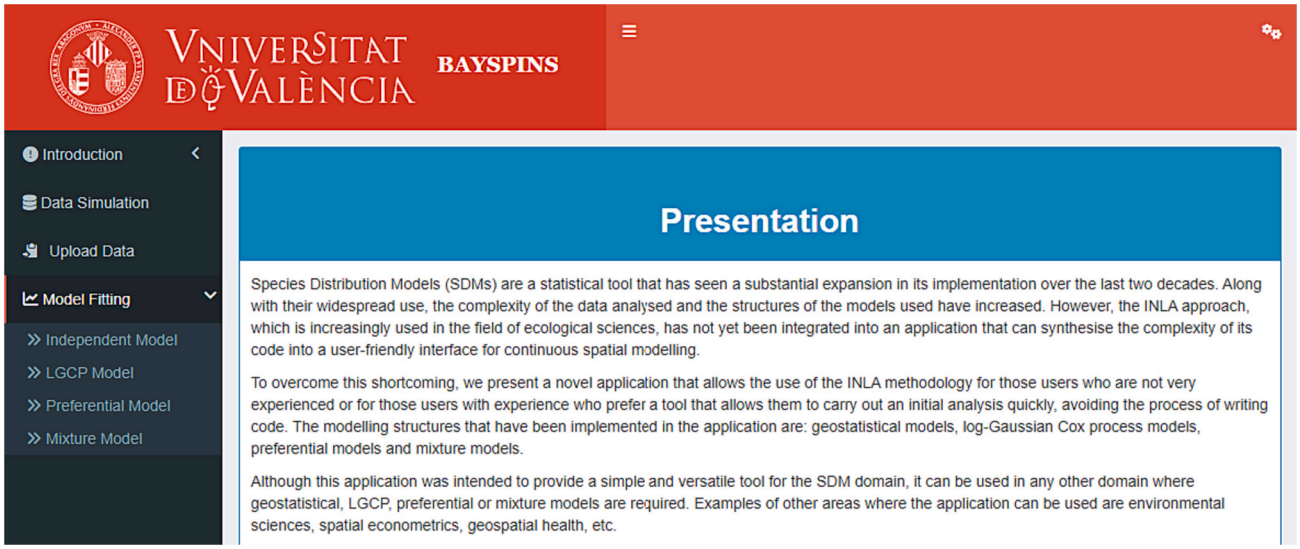
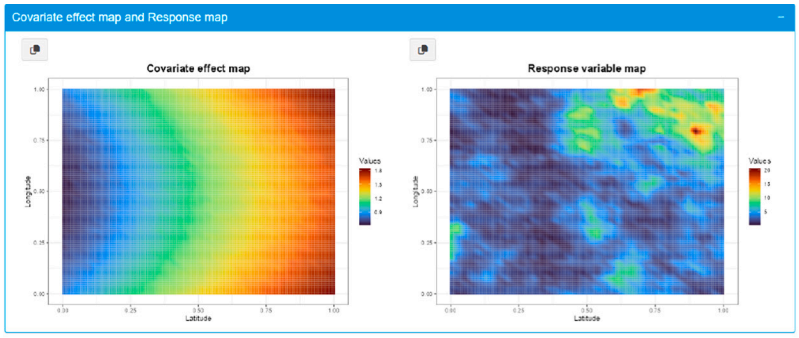
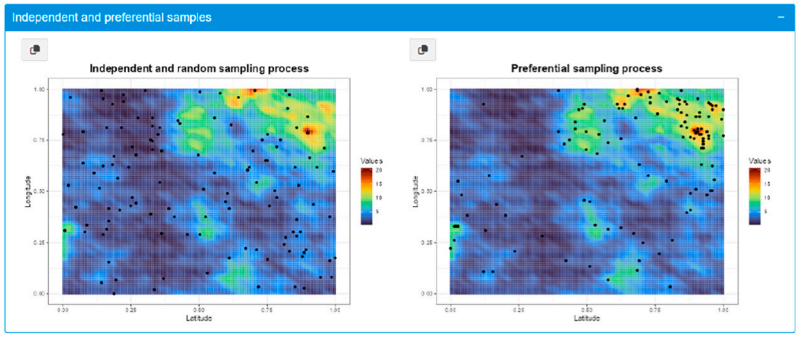


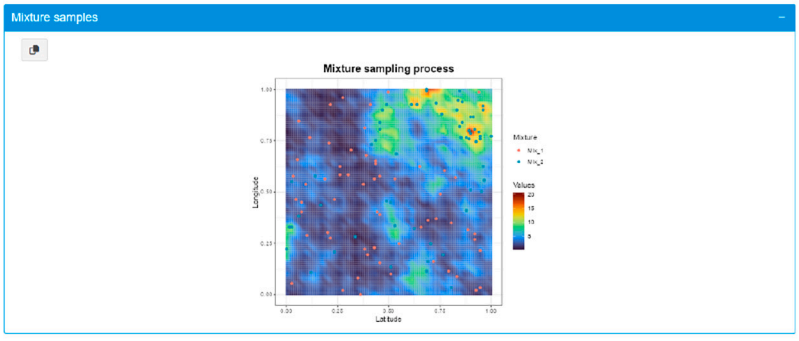
Fig. 2. Detail of the app presentation window.



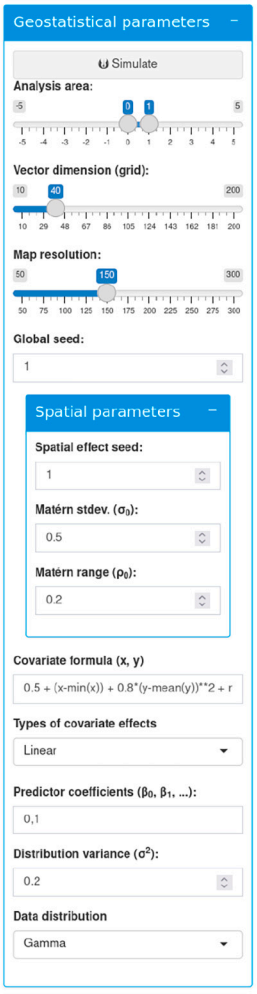
(a) Geostatistical simulation.



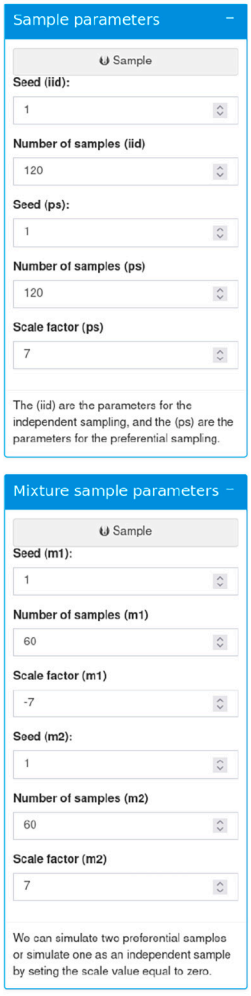
(b) Independent and preferential sampling



(c) Mixture sampling



(d) Spatial panel



(e) Sampling panels

Fig. 3. Panels and outputs of the geostatistical and sampling process simulations performed by the app.

5. Uploading data

This section deals with data uploading. It provides information on the structure of the user interface and how the results are displayed using the two tabs: *main analysis data uploading* and *auxiliary data frame* (see Fig. 4).

As the name suggests, the first interface input allows users to upload the data required for the main analysis. In order to upload the data correctly, it should be noted that the first two columns are reserved for the locations, with the longitude and latitude (or x and y axis) being the first and second columns, respectively. The remaining columns will be those relating to the explanatory variables. It is worth noting that the app is prepared to deal with any sort of misaligned data, both for the response and the covariates. Once the data frame has been uploaded, a table with the data and a plot of the distribution of the observations with their values will (automatically) be displayed.

The second interface input of the uploading tab enables the user to introduce additional covariate data, providing functionality to upgrade the covariate data frame (again handling misaligned or incomplete data), improving prediction accuracy or fix prediction points.

1. *Additional data*: for this functionality we only need to add the data frame with the values of each covariate in a column to the right of the coordinates with the same name as they appear in the *main analysis data frame*. The integration of new information for the same covariates will re-structure the data frame, to include both the available values and those that are NAs.

   Since the prediction process requires the values of the covariates at the prediction locations, there is another feature for the additional data design, which is to incorporate more information in the prediction process.
2. *Fixing prediction points*: this feature allows the user to define each row as a prediction point, with its location information and the values of the required covariates. Therefore, all the covariates have to be

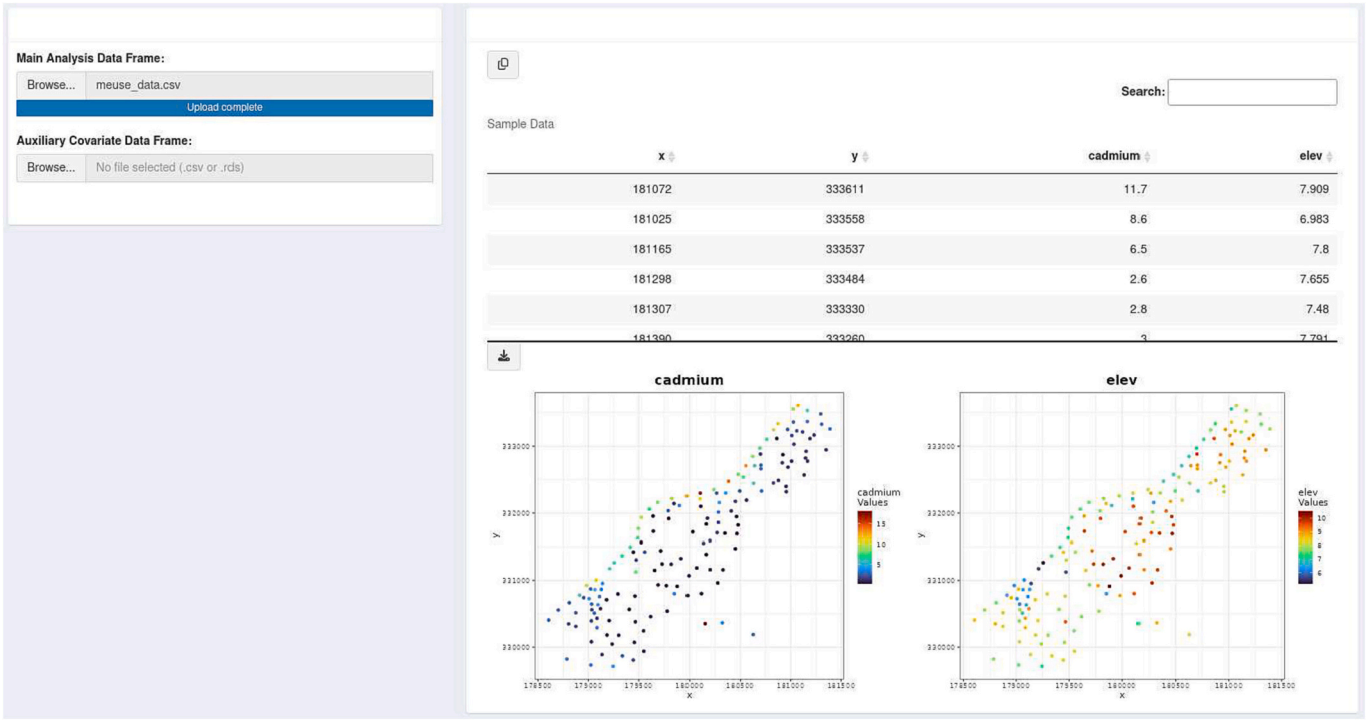


Fig. 4. The Uploading data tab allows data be imported for analysis, including sample locations, the response variable and explanatory variables. It is also possible to upload an auxiliary data frame, which improves the inferential and predictive performance of the model.

included with the same names as they had when introduced into the main analysis data frame. This allows a prediction grid to be built according to the information available in the new dataframe, avoiding the use of model covariates. However, this in turn implies that if we use this feature, there should be no NAs, since this option has been explicitly designed for prediction.

As before, the second interface (*auxiliary data frame*) also displays tables and graphs with the coordinates and variables of this new data frame.

6. Model fitting

In this section we describe the model fitting tab, which is at the heart of our app. In particular, we outline the underpinnings of the inference and prediction processes for the four available statistical models: *independent or geostatistical model* (Diggle et al., 1998), *LGCP model* (Simpson et al., 2016), *preferential model* (Diggle et al., 2010), and *mixture model* (Figueira et al., 2023b). Further details about the models can be found in the supplementary material. We devote a subsection to each model, describing in detail each component necessary to set up the analysis and integrate it into the inference process. Many elements that are essential for configuring the analysis are either identical or very similar across all models. Consequently, these common components are discussed in detail in the first subsection and briefly reviewed in subsequent ones.

6.1. Geostatistical modelling

The geostatistical modelling tab, often referred to as the independent model, allows users to fine-tune the process that determines the values of the response variable. It includes an interceptor, a spatial effect, and any of the covariates present in the uploaded (or simulated) database. As these covariates can be included as linear or non-linear effects (or even as a factor with random effects, as explained below), the resulting model used can be succinctly represented by the following formula for the response variable data $y = (y_1, \ldots, y_n)$ based on the locations $s = (s_1, \ldots, s_n)$:

$$\begin{aligned} y_i | s_i &\sim \text{Dist}(\mu(s_i), \phi), \\ g(\mu(s_i)) &= \beta_0 + \mathbf{X}\boldsymbol{\beta} + \sum_j f_j(z_j) + u_i(\rho, \sigma), \end{aligned} \tag{2}$$

where Dist$(\cdot)$ stands for any of the available likelihood distributions (Bernoulli, binomial, beta, exponential, Gaussian, Gamma, Poisson); $\boldsymbol{\beta}$ are the linear effects, $f_j(\cdot)$ are the non-linear and random effects; and $u_i(\rho, \sigma)$ is the spatial effect with Matérn covariance. If necessary, a transformation of the response variable could also be performed outside the app, enabling the application of some of the available distributions. In addition, if there are NAs due to misaligned or missing data, the app internally contains a model for each covariate with this problem. In particular, and in line with Barber et al. (2016); Krainski et al. (2018), the covariates of these models consist of an intercept and a spatial random effect

$$\begin{aligned} x_i | s_i^* &\sim \text{N}(\mu(s_i^*), \phi^*), \\ \mu(s_i^*) &= \beta_0^* + u_i^*(\rho^*, \sigma^*). \end{aligned} \tag{3}$$

However, there may be factors or categorical variables available as explanatory variables for which the process described above would not be valid. For each factor available in the dataset, a checkbox appears allowing the user to select the level of the factor. In particular, the user can select a reference level or an alternative process consisting of using the level of the factor associated with the closest location to the prediction location.

As can be seen in Fig. 5, from the modelling tab we can set the following elements: the type of data to be analysed (simulated or uploaded), the distribution of the response variable (the ones mentioned above), whether the auxiliary covariate data can be used to improve the inference process (the former functionality) and whether they are used to fix the locations and covariates at the prediction points (the latter functionality). In addition, we can also select the dimensions of the prediction grid, which will be re-evaluated internally in order to select only those points that are contained within the inner mesh boundary. More information on mesh building can be found in the next subsection,

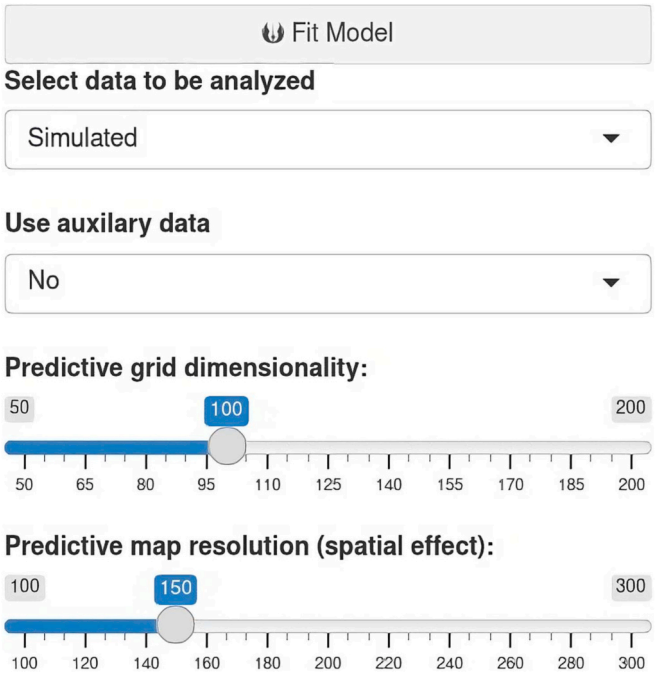


Fig. 5. Interface elements for determining which data to analyse (simulated or uploaded) and whether or not to use auxiliary data, with the definition of the spatial resolution.

but for a more detailed explanation the reader is referred to Lindgren et al. (2011); Krainski et al. (2018); Gómez-Rubio (2020).

The remaining elements of the modelling tab are used to configure the mesh for the spatial effect, the covariates and their effects along with the corresponding priors, and other advanced settings for INLA. All of these are explained in the following subsections.

6.1.1. *Mesh building*

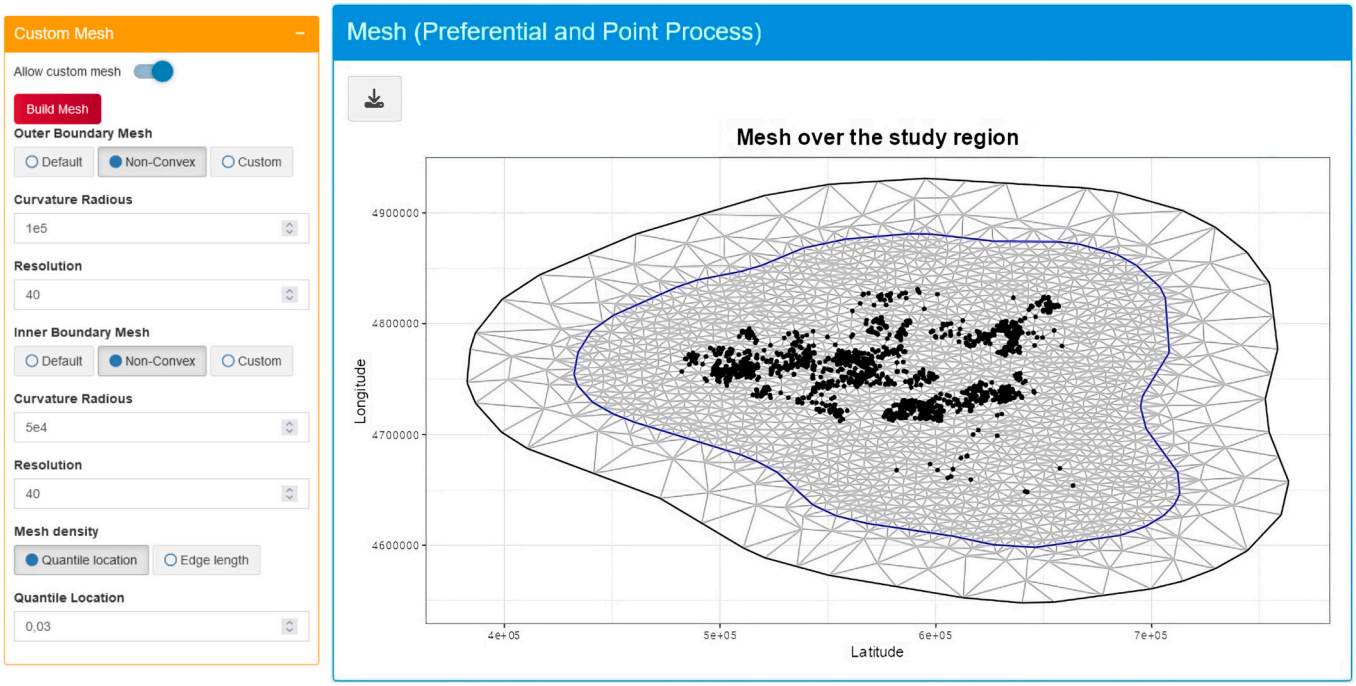
The analysis of the spatial effect is the component that forces us to use the Stochastic Partial Differential Equations and Finite Element Methods (SPDE-FEM) approach implemented in INLA (Lindgren, 2012; Lindgren et al., 2011; Lindgren and Rue, 2015). This approach also requires the definition of a mesh in whose nodes the SPDE is solved and which will be interpolated to the observation or prediction locations using the FEM methodology. That is, the observations and predictions must be contained within this mesh.

The mesh is constructed based on the locations of the observations. Specifically, a distance for the edges of the triangles is set according to an automatice algorithm (based on the quantile distribution of the distance between at least 50 random points from the observations) or by setting it manually. In addition, the boundaries can be modified by selecting a non-convex shape or by importing a boundary vertex location from a csv or rds file, enabling the user to define the boundaries to be used. Fig. 6 shows the mesh generation interface available in each modelling tab.

6.1.2. *Setting up the linear predictor with the prior distributions*

The components of the model are of two kinds, the default components (intercept and spatial effect) mentioned above, and the user defined components. The latter can be selected from the *main analysis data frame* using a checkbox. From here it is possible to select one of the following available effects: a linear effect (or reference level for factor variables), a first or second order random walk (rw1 or rw2), a one-dimensional SPDE (spde1) or an iid effect for factor variables.

As we are working within the Bayesian framework, priors must be elicited for all the parameters and hyperparameters governing the model. For each selected covariate, the user can specify among all the available distributions: normal distribution (for linear effects), log-gamma distribution, uniform distribution, flat uniform distribution, and the more recent and suitable-for-INLA penalised complexity priors (PC-prior, Simpson et al., 2017). However, for spatial effects, there is a dedicated field in the interface to configure one's own prior distribution, specifically choosing between base prior (Lindgren, 2012) and PC-prior



(a) Mesh panel (b) Mesh example

Fig. 6. Control panel for configuring the construction of the mesh, together with an example of a mesh over a study region.

distributions (Fuglstad et al., 2019).

6.1.3. Advanced INLA features

Once all the parameters and elements for the modelling process are defined, the Bayesian inference and prediction within the INLA approach is automatically performed jointly. Common performance measures for model selection (DIC, Spiegelhalter et al., 2002; WAIC, Watanabe, 2013 and CPO, Pettit, 1990) are part of the results. The final items in the modelling tab allow advanced users to fine-tune this process. In particular, the user can select the INLA mode (*compact*, *experimental* or *classic*, where *compact* is the default mode) and also specify the values of the modes for the hyperparameters.

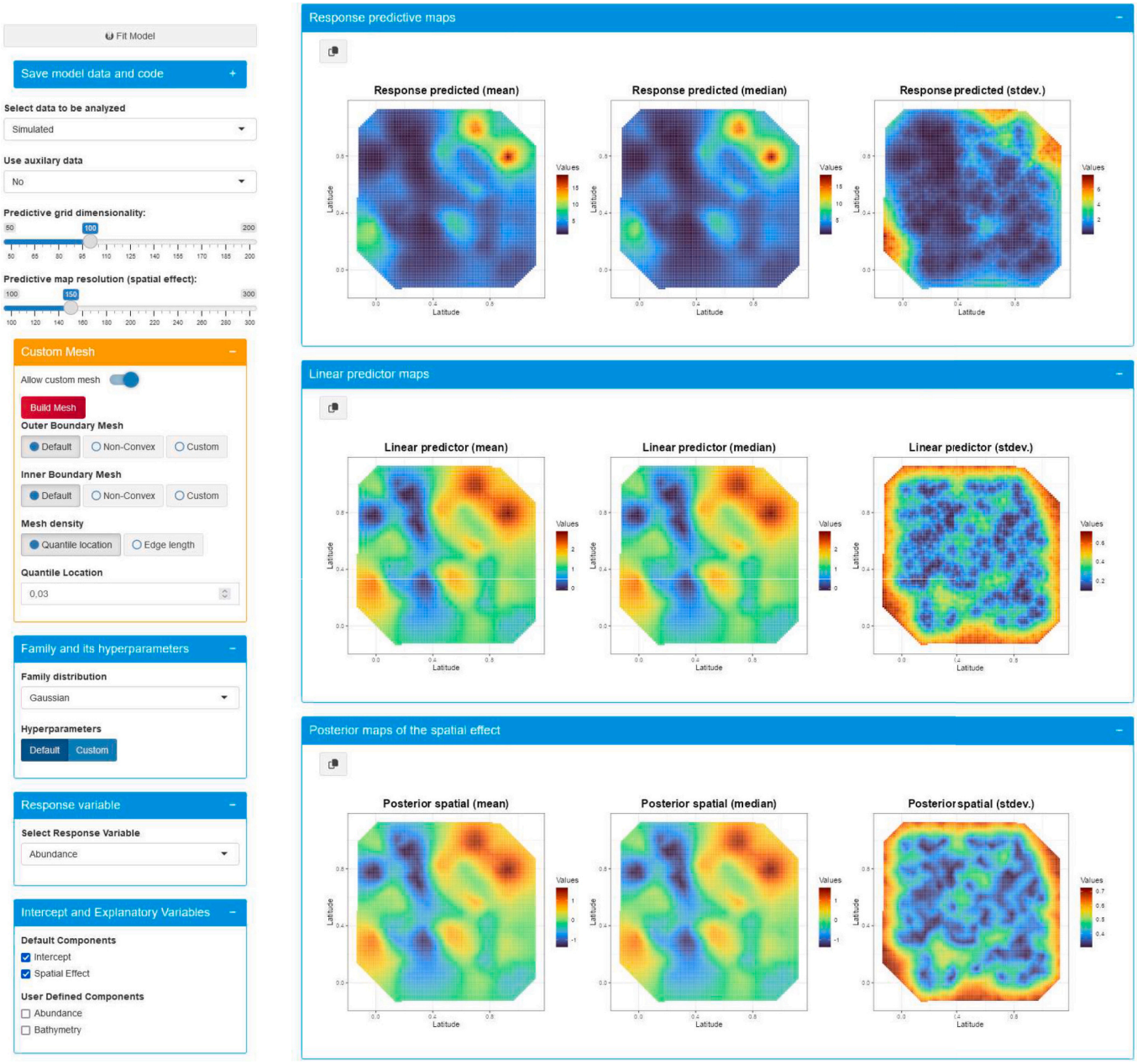
6.1.4. Geostatistical example

In what follows, we make use of the simulated dataset already shown in Fig. 3b to demonstrate how BAYSPINS works. We start by choosing a Gamma distribution for the response variable of these data, selecting the intercept, the bathymetry (with a linear effect among those available) as the only covariate, and also the spatial effect.

Fig. 7 shows the control panel settings and the resulting model maps for the response variable, the linear predictor and the spatial effect. Note that although not shown in Fig. 7, the app also provides other results such as the distributions of the parameters and hyperparameters, or DIC, WAIC and CPO.

6.2. LGCP modelling

The second modelling tab is dedicated to Log Gaussian Cox Process (LGCP) models. From here, users can use a Bayesian hierarchical model to analyse the process that generates the locations of a sample (Illian et al., 2012; Simpson et al., 2016). These data are very common in certain fields, such as ecology, and are also known as presence-only



(a) Panel example (b) Modelling results

Fig. 7. Geostatistical example panels and modelling output results.

data.

A LGCP model is usually defined as follows

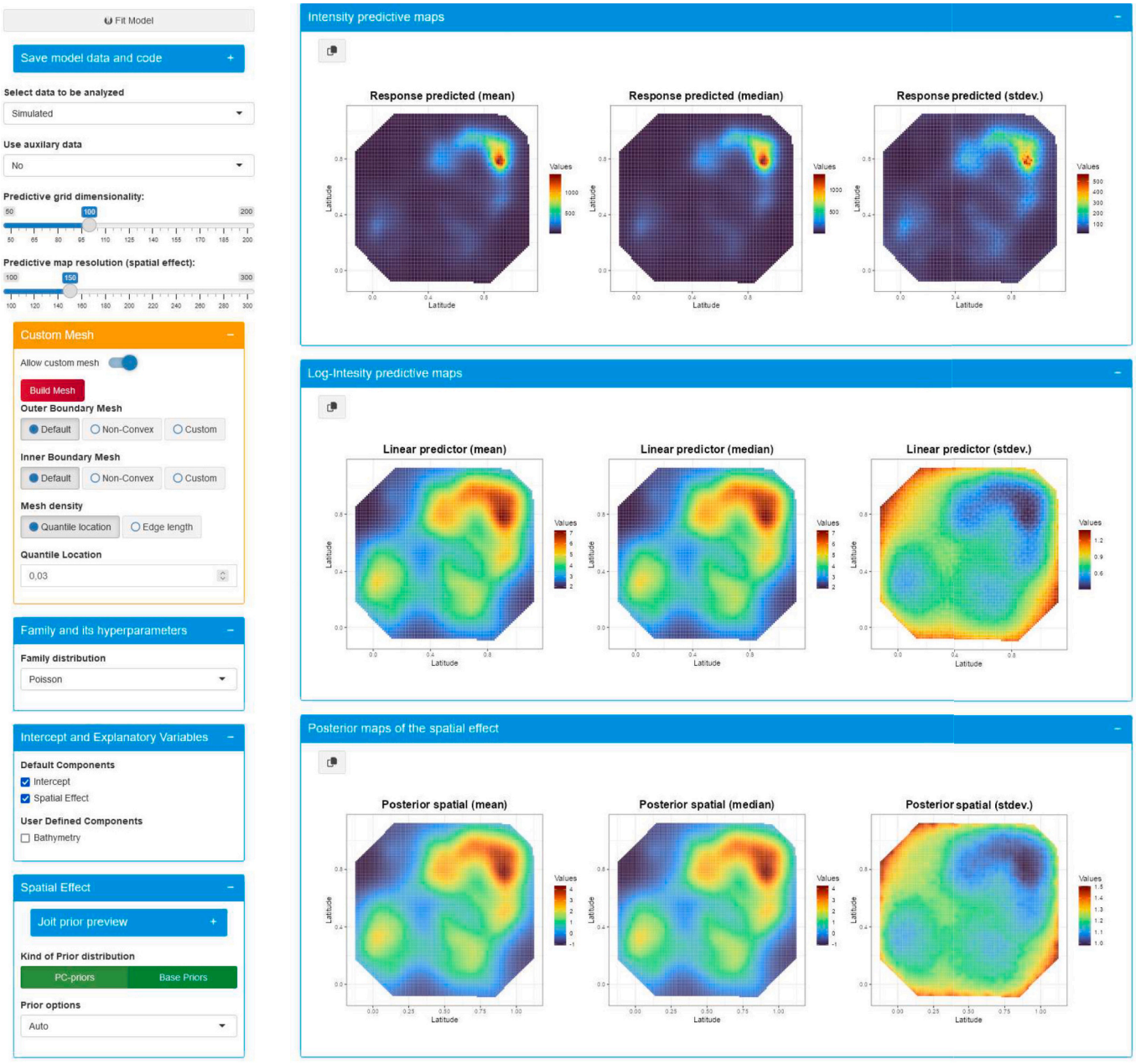
$$\begin{aligned} s_i &\sim \mathrm{LGCP}(\lambda(s_i)), \\ log(\lambda(s_i)) &= \beta_0 + \mathbf{X}\boldsymbol{\beta} + \sum_j f_j(z_j) + u_i(\rho, \sigma), \end{aligned} \tag{4}$$

where the linear predictor has the same structure as Eq. (2). The main feature of the LGCP model is the evaluation of the spatial dependence of the intensity function $\lambda(s)$. However, it is important to note that the likelihood function is exclusively Poisson because it is characterised as a non-homogeneous Poisson process. Apart from that, other elements such as the mesh configuration, the choice of explanatory variables, and the setting of priors for parameters and hyperparameters can be arranged in a similar way to the options in the geostatistical tab.

Using the simulated example shown in Fig. 3b for preferred samples, we now present in Fig. 8 the results obtained using this model. In particular, we have selected all the available elements for the linear predictor: intercept, bathymetry and spatial effect. We have left most of the settings at their default values, with the exception of the mesh. We have adjusted the quantile location parameter slightly upwards to reduce the density of the mesh.

6.3. *Preferential modelling*

For those scenarios in which the sampling process is driven by some sort of preferentiality associated with the location marks (Diggle et al., 2010; Pennino et al., 2019), BAYSPINS also allows the use of a preferential model (third item of the modelling tab). This is a joint model of the response variable (also called marks) and the sampling locations, in which the user can decide (via a check box) which effects are related, with the default being that all effects are shared (Bakka et al., 2018;



(a) Panel example

(b) Modelling results

Fig. 8. LGCP example panels and modelling output results.

Krainski et al., 2018) except the linear ones.

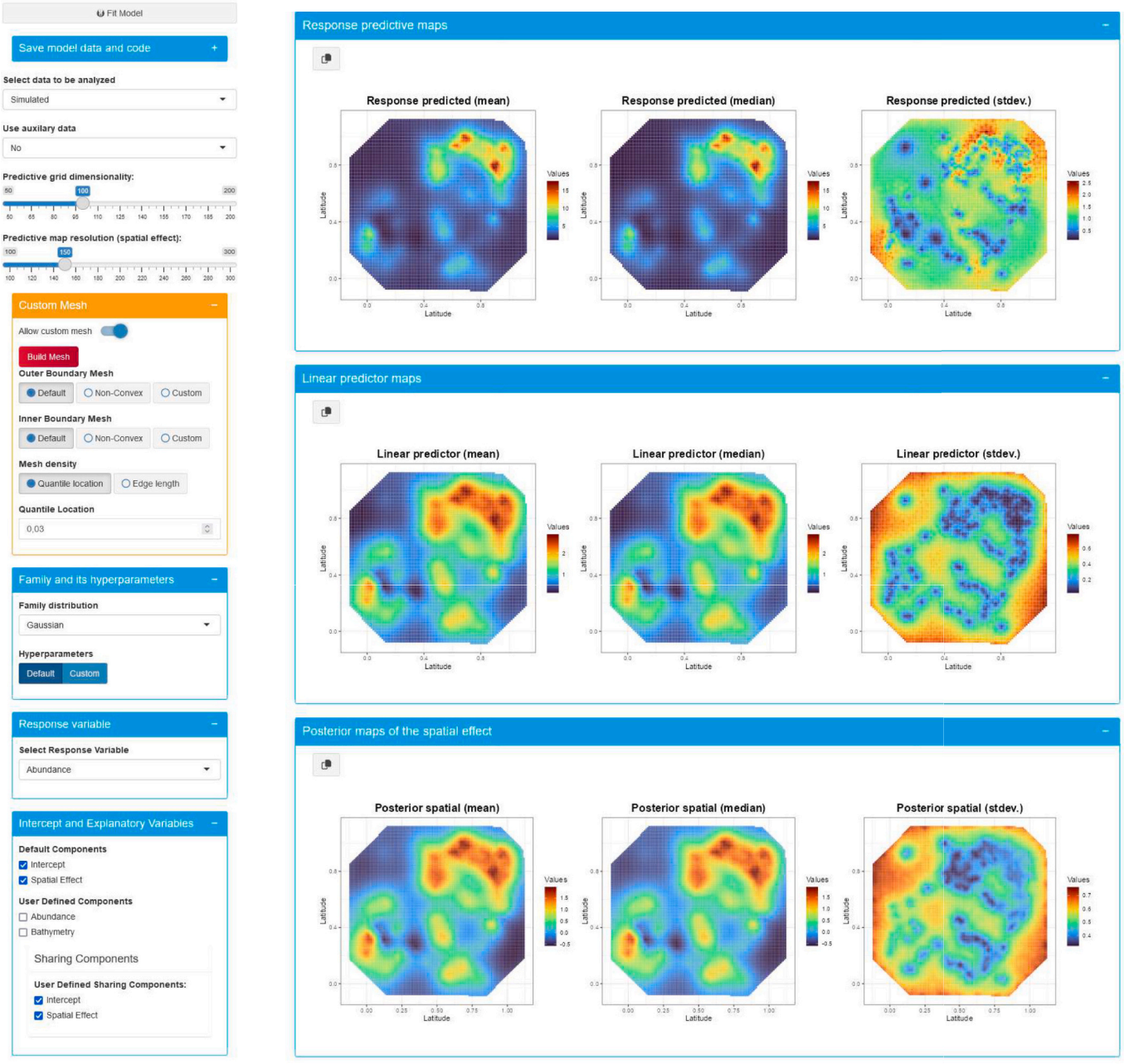
The basic structure of this joint model, for the marks $\mathbf{y} = (y_1, \ldots, y_n)$ and the locations $\mathbf{s} = (s_1, \ldots, s_n)$, can be described as follows

$$\begin{aligned} y_i|s_i &\sim \text{Dist}(\mu(s_i), \phi), \\ g(\mu(s_i)) &= \beta_0 + X\beta + \sum_j f_j(z_j) + u_i(\rho, \sigma), \\ s_i &\sim \text{LGCP}(\lambda(s_i)), \\ log(\lambda(s_i)) &= \beta_0' + X\beta' + \sum_j f_j(z_j) + u_i(\rho', \sigma'), \end{aligned} \tag{5}$$

where the elements of the LGCP latent field are shared with those of the geostatistical model by means of a scale parameter. For instance, the spatial effect would be $u_i' = \alpha \cdot u_i$, and thus, the u_i effect would be common to both the geostatistical and the point processes: while the overall shape of the effect remains consistent, it would be scaled by the α parameter. Note that most of the elements of 5 are identical to those shown for the geostatistical model, except that there are two likelihoods. It is also worth noting that in most cases only the spatial effect is shared (Bakka et al., 2018; Krainski et al., 2018; Martínez-Minaya et al., 2018), although other components of the predictor could also be shared (Paradinas et al., 2017).

As mentioned above, the preferential tab has a check box with the names of the variables present in the dataframe, from which the user can select which effect to assign (linear, random walk, spde) and which of them will have the effect shared between the geostatistical and the point process. If selecting one of the covariates chosen to share its effect is not selected, it is implemented in the point process with the selected effect, but its estimation is independent between the two processes.

The sharing or copying process means that the selected variable effect θ_i is copied by means of a very small transformation of it:



(a) Panel example (b) Modelling results

Fig. 9. Preferential example panels and modelling output results.

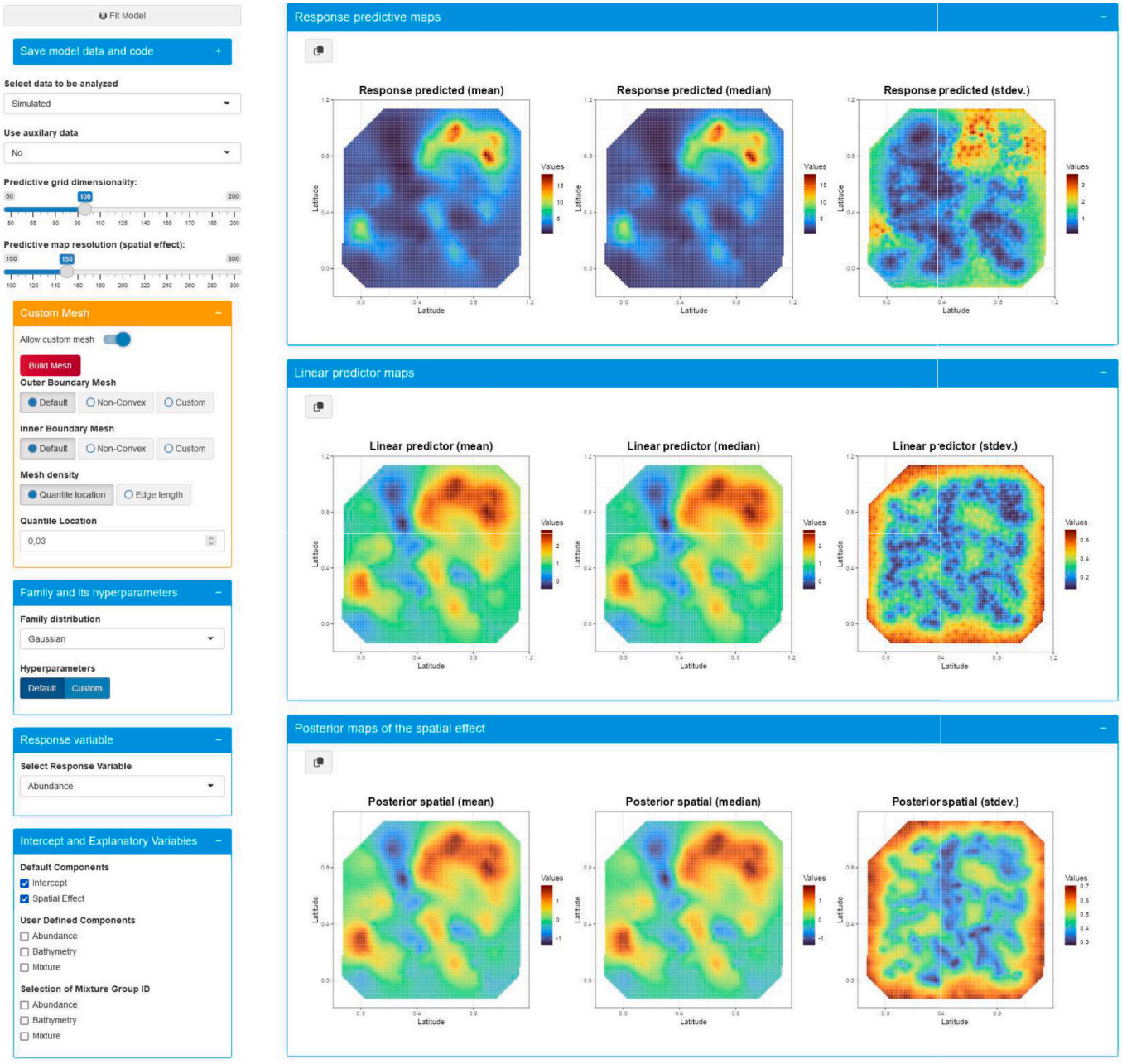
$$\theta_i^{'} = \alpha \cdot \theta_i + \varepsilon_i,$$

where $\theta_i^{'}$ is the copied effect of θ_i and ε is a tiny error component added for computational reasons. The factor α, viewed as a scaling factor, is estimated using a normal prior distribution. However, in the *advanced INLA configuration*, the user has the option to specify the mean and precision for these normal prior distributions.

Fig. 9 shows the modelling configuration panel and some of the results displayed by the app for the preferential simulated example in Fig. 3b) when using the preferential modelling tab. It is clear from this figure that the configuration elements are the same as those required for the geostatistical and LGCP models, apart from the shared effects. This similarity underscores our intention to build an app that prioritises simplicity and comprehensibility.

6.4. Mixture modelling

The final modelling tab facilitates the analysis of data from different sampling structures by allowing the user to integrate two types of likelihoods, one for the response variable and several others for generating locations, in line with (Figueira et al., 2023b). Accordingly, if we possess a dataset assembled from multiple sample data, denoted as $Y = (Y_1, \ldots, Y_K)$, each exhibiting a unique interdependency between the mark-generating and sampler processes, the mixture model allows a consolidated analysis. It integrates a single geostatistical sub-model for all the marks, while dedicating a separate LGCP sub-model to each set of locations associated with a unique dependent sampling structure:



(a) Panel example (b) Modelling results

Fig. 10. Mixture example panels and modelling output results.

$$
\begin{aligned}
y_i|s_i &\sim \text{Dist}(\mu(s_i), \phi), \\
g(\mu(s_i)) &= \beta_0 + X\beta + \sum_j f_j(z_j) + u_i(\rho, \sigma), \\
s_i^{(k)} &\sim \text{LGCP}\left(\lambda^{(k)}\left(s_i^{(k)}\right)\right), \\
log\left(\lambda^{(k)}\left(s_i^{(k)}\right)\right) &= \beta_0^{(k)} + X^{(k)}\beta^{(k)} + \sum_j f_j^{(k)}\left(z_j^{(k)}\right) + u_i^{(k)}\left(\rho^{(k)}, \sigma^{(k)}\right),
\end{aligned} \tag{6}
$$

where the superscript $k \in \{1, \ldots, K\}$ denotes the association with the dataset Y_k. Shared effects will then only be manifested manifest between the geostatistical process and the K sub-models corresponding to the location of each sampler. The other components of the model remain consistent with the terminology used in the preceding models.

The mixture tab retains the user interface elements of the analysis tabs described above, with an additional feature: the ability to select the column name in the data frame that pinpoints the identifiers for the sample structure associated with each data point, and the ability to select the shared effects between the geostatistical process and the point process for the different preferred sampling structures.

If data are assumed to be from independent sampling, the analysis of the mixture model will use only the geostatistical layer. Thus, user-supplied data must include a column that specifies the origin of the different preferential samplings, namely an identifier that differentiates the data based on their respective samplers. Samples derived from structures that are independent of the response variable generation should be labelled "Ind" in this column.

Fig. 10 shows the results of a simulated mixture model example coming from data outlined in Fig. 3c. These data stem from two distinct sampling structures. Their identifiers are contained in the Mix column name, with the labels Mix_1 and Mix_2 representing each sample structure. In the modelling of this example, we use the default settings for mesh and prediction grid. We choose the two default effects (intercept and spatial effect) and the bathymetry, the only explanatory variable available in the user defined components. Then, we select the Mixture variable name for the Selection of Mixture Group ID, leaving the rest of the elements with their default settings.

6.5. Feedback protocols

The app has a built-in utility that enables the redefinition of prior distributions for all model components, as already described in the corresponding sections, which allows the user to establish feedback protocols or sequential learning approaches. With this capability, when we have inferential results from a model associated with the same phenomenon and with the same components, we can use the characteristic moments of the posterior distributions to redefine the prior distributions of the new model (Figueira et al., 2023a). This can be done using the checkbox available to customise the prior distribution of the effects: the intercept, the linear or non-linear components, and for the spatial effect itself, if they are considered to have similar hyperparameter distributions to the one estimated with the previous data.

This may lead to a better identification of effects, better predictions and improvements in the computational time required for adjustment, together with an increase in the robustness of the inferential process.

7. Conclusions

The complexity of Species Distribution Models (SDMs) in ecology has lead to the creation of various tools and R packages to facilitate their implementation. However, the INLA methodology has not yet been integrated into a user-friendly app for continuous spatial modelling. To address this, the paper presents a novel app called BAYSPINS that allows users to perform geostatistical, LGCP, preferential and mixture modelling using the INLA methodology with a visual interface. In addition, the app provides default settings for quick evaluations or customisation options for more rigorous studies, depending on the user's skill level.

The paper delves into the many features embedded in the app. It illustrates the capabilities of the tool and also addresses potential pitfalls associated with applications of this type. Such applications often conceal much of their code and foundational methodology in order to improve user-friendliness, especially for newcomers to the field of statistics and computing. In recognition of these challenges, and due to space constraints, the basic concepts associated with INLA are appended in the supplementary material. Importantly, the entire codebase is accessible on an open-source GitHub repository. This platform also welcomes users to report issues, highlight bugs and suggest potential enhancements to BAYSPINS.

The paper also emphasis the adaptability of BAYSPINS. Beyond SDMs, it is equally applicable in areas such as environmental science, spatial econometrics, epidemiology, and more. This malleability paves the way for future expansion of the app to encompass global terrestrial models, space-time models, downscaling frameworks, data fusion models, hurdle-models (e.g. for rainfall data) and any other integration that users deem essential.

Nevertheless, it is crucial to recognise that in the field of spatial statistics, models that integrates structured spatial effects may pose inference challenges, including issues related to spatial confounding in its various interpretations (Gilbert et al., 2021; Urdangarin et al., 2023). In the domain of species distribution models, a particularly intriguing aspect of spatial confounding is the accurate discrimination of spatially structured covariates and the spatially random effect (Mäkinen et al., 2022). Various approaches have been proposed to overcome this challenge, including restricted spatial regression (Reich et al., 2006) and spatial+ (Dupont et al., 2022). To address this, BAYSPINS' users could download the data and code from the model and modify the spatial effect according to the specific procedure they wish to perform.

In summary, BAYSPINS is adept at processing diverse data to solve geostatistical, LGCP, preferential, and mixture models. It also supports inter-model feedback techniques. This app effectively bridges the gap between INLA (with SPDE-FEM) and non-expert users, crafting a user-centered design tool rich in options to refine and enhance both inference and prediction processes.

CRediT authorship contribution statement

Mario Figueira: Writing – review & editing, Writing – original draft, Software, Conceptualization. **David Conesa:** Writing – review & editing, Writing – original draft, Supervision, Conceptualization. **Antonio López-Quílez:** Writing – review & editing, Supervision, Conceptualization.

Data availability

The link to the code and the data used is in the document itself.

Acknowledgements

This paper is part of the project PID2022-136455NB-I00, funded by Ministerio de Ciencia, Innovación y Universidades of Spain (MCIN/AEI/10.13039/501100011033/FEDER, UE) and the European Regional Development Fund. DC also acknowledges Grant CIAICO/2022/165 funded by Generalitat Valenciana.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2024.102542.

References

Abrahms, B., Aikens, E.O., Armstrong, J.B., Deacy, W.W., Kauffman, M.J., Merkle, J.A., 2021. Emerging perspectives on resource tracking and animal movement ecology.

Trends Ecol. Evol. 36, 308–320. URL. https://www.sciencedirect.com/science/article/pii/S0169534720303104. https://doi.org/10.1016/j.tree.2020.10.018.
Adin, A., Goicoa, T., Ugarte, M.D., 2019. Online relative risks/rates estimation in spatial and spatio-temporal disease mapping. Comput. Methods Prog. Biomed. 172, 103–116. URL. https://www.sciencedirect.com/science/article/pii/S0169260718318030. https://doi.org/10.1016/j.cmpb.2019.02.014.
Bakka, H., Rue, H., Fuglstad, G.A., Riebler, A., Bolin, D., Illian, J., Krainski, E., Simpson, D., Lindgren, F., 2018. Spatial modeling with R-INLA: a review. WIREs Comput. Stat. 10, e1443. URL: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.1443 https://doi.org/10.1002/wics.1443.
Barber, X., Conesa, D., Lladosa, S., López-Quílez, A., 2016. Modelling the presence of disease under spatial misalignment using Bayesian latent Gaussian models. Geospat. Health 11. URL. https://geospatialhealth.net/index.php/gh/article/view/415. https://doi.org/10.4081/gh.2016.415.
Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., Borges, B., 2023. Shiny: Web Application Framework for R. URL. https://shiny.posit.co/. R package version 1.7.4.9002.
Diggle, P.J., Tawn, J.A., Moyeed, R.A., 1998. Model-based geostatistics. J. R. Stat. Soc.: Ser. C: Appl. Stat. 47, 299–350. https://doi.org/10.1111/1467-9876.00113.
Diggle, P.J., Menezes, R., Su, T.L., 2010. Geostatistical inference under preferential sampling. J. R. Stat. Soc.: Ser. C: Appl. Stat. 59, 191–232. https://doi.org/10.1111/j.1467-9876.2009.00701.x.
Dupont, E., Wood, S.N., Augustin, N.H., 2022. Spatial+: a novel approach to spatial confounding. Biometrics 78, 1279–1290. https://doi.org/10.1111/biom.13656.
Figueira, M., Barber, X., Conesa, D., López-Quílez, A., Martínez-Minaya, J., Paradinas, I., Pennino, M.G., 2023a. Bayesian feedback in the framework of ecological sciences arXiv:2305.17922.
Figueira, M., Conesa, D., López-Quílez, A., Paradinas, I., 2023b. How to perform modeling with independent and preferential data jointly? arXiv:2307.07094.
Fletcher, R., Fortin, M.J., 2019. Spatial ecology and conservation modeling: Applications with R. https://doi.org/10.1007/978-3-030-01989-1.
Fuglstad, G.A., Simpson, D., Lindgren, F., Rue, H., 2019. Constructing priors that penalize the complexity of Gaussian random fields. J. Am. Stat. Assoc. 114, 445–452. https://doi.org/10.1080/01621459.2017.1415907.
Gilbert, B., Datta, A., Casey, J.A., Ogburn, E.L., 2021. A Causal Inference Framework for Spatial Confounding.
Gómez-Rubio, V., 2020. Bayesian Inference with INLA. Chapman & Hall/CRC Press. https://doi.org/10.1201/9781315175584.
Guisan, A., Thuiller, W., Zimmermann, N.E., 2017. Habitat Suitability and Distribution Models: With Applications in R. Ecology, Biodiversity and Conservation. Cambridge University Press. https://doi.org/10.1017/9781139028271.
Huang, J., Malone, B.P., Minasny, B., McBratney, A.B., Triantafilis, J., 2017. Evaluating a Bayesian modelling approach (INLA-SPDE) for environmental mapping. Sci. Total Environ. 609, 621–632. URL: https://www.sciencedirect.com/science/article/pii/S0048969717319046 https://doi.org/10.1016/j.scitotenv.2017.07.201.
Illian, J.B., Sørbye, S.H., Rue, H., 2012. A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA). Ann. Appl. Stat. 6, 1499–1530. https://doi.org/10.1214/11-AOAS530.
Jung, M., 2023. An integrated species distribution modelling framework for heterogeneous biodiversity data. Eco. Inform. 76, 102127. URL. https://www.sciencedirect.com/science/article/pii/S1574954123001565. https://doi.org/10.1016/j.ecoinf.2023.102127.
Kass, J.M., Vilela, B., Aiello-Lammens, M.E., Muscarella, R., Merow, C., Anderson, R.P., 2018. Wallace: a flexible platform for reproducible modeling of species niches and distributions built for community expansion. Methods Ecol. Evol. 9, 1151–1156. https://doi.org/10.1111/2041-210X.12945.
Kass, J.M., Pinilla-Buitrago, G.E., Paz, A., Johnson, B.A., Grisales-Betancur, V., Meenan, S.I., Attali, D., Broennimann, O., Galante, P.J., Maitner, B.S., Owens, H.L., Varela, S., Aiello-Lammens, M.E., Merow, C., Blair, M.E., Anderson, R.P., 2023. Wallace 2: a shiny app for modeling species niches and distributions redesigned to facilitate expansion via module contributions. Ecography 2023, e06547. https://doi.org/10.1111/ecog.06547 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecog.06547.
Krainski, E., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., Rue, H., 2018. Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA. https://doi.org/10.1201/9780429031892.
Lindgren, F., 2012. Continuous domain spatial models in R-INLA. ISBA Bull. 19, 14–20.
Lindgren, F., Rue, H., 2015. Bayesian spatial modelling with r-inla. J. Stat. Softw. 63, 1–25. URL. https://www.jstatsoft.org/index.php/jss/article/view/v063i19. 10.18637/jss.v063.i19.
Lindgren, F., Rue, H., Lindström, J., 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. J. R. Stat. Soc. Ser. B Stat Methodol. 73, 423–498. https://doi.org/10.1111/j.1467-9868.2011.00777.x.
Mäkinen, J., Numminen, E., Niittynen, P., Luoto, M., Vanhatalo, J., 2022. Spatial confounding in bayesian species distribution modeling. Ecography 2022, e06183. https://doi.org/10.1111/ecog.06183 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecog.06183.
Martínez-Minaya, J., Cameletti, M., Conesa, D., Pennino, M.G., 2018. Species distribution modeling: a statistical review with focus in spatio-temporal issues. Stoch. Env. Res. Risk A. 32, 3227–3244. https://doi.org/10.1007/s00477-018-1548-7.
Moraga, P., 2017. SpatialEpiApp: a shiny web application for the analysis of spatial and spatio-temporal disease data. Spatial Spatio-temporal Epidemiol. 23, 47–57. URL. https://www.sciencedirect.com/science/article/pii/S187758451730062X. https://doi.org/10.1016/j.sste.2017.08.001.
Moraga, P., Dean, C., Inoue, J., Morawiecki, P., Noureen, S.R., Wang, F., 2021. Bayesian spatial modelling of geostatistical data using inla and spde methods: a case study predicting malaria risk in Mozambique. Spatial Spatio-temporal Epidemiol. 39, 100440. URL: https://www.sciencedirect.com/science/article/pii/S1877584521000393 https://doi.org/10.1016/j.sste.2021.100440.
Osorio-Olvera, L., Lira-Noriega, A., Soberón, J., Peterson, A.T., Falconi, M., Contreras-Díaz, R.G., Martínez-Meyer, E., Barve, V., Barve, N., 2020. ntbox: an r package with graphical user interface for modelling and evaluating multidimensional ecological niches. Methods Ecol. Evol. 11, 1199–1206. URL. https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13452. https://doi.org/10.1111/2041-210X.13452.
Ovaskainen, O., Abrego, N., 2020. Joint species distribution modelling: with applications in R. In: Ecology, Biodiversity and Conservation. Cambridge University Press. https://doi.org/10.1017/9781108591720.
Paradinas, I., Conesa, D., López-Quílez, A., Bellido, J.M., 2017. Spatio-Temporal model structures with shared components for semi-continuous species distribution modelling. Spatial Statistics 22, 434–450. URL. https://www.sciencedirect.com/science/article/pii/S2211675316300872. https://doi.org/10.1016/j.spasta.2017.08.001. Spatio-temporal Statistical Methods in Environmental and Biometrical Problems.
Pennino, M.G., Paradinas, I., Illian, J.B., Muñoz, F., Bellido, J.M., López-Quílez, A., Conesa, D., 2019. Accounting for preferential sampling in species distribution models. Ecol. Evolut. 9, 653–663. https://doi.org/10.1002/ece3.4789 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/ece3.4789.
Pettit, L.I., 1990. The conditional predictive ordinate for the normal distribution. J. R. Stat. Soc. Ser. B Methodol. 52, 175–184. URL: http://www.jstor.org/stable/2345658.
Pyšek, P., Pergl, J., Essl, F., Lenzner, B., Dawson, W., Kreft, H., Weigelt, P., Winter, M., Kartesz, J., Nishino, M., Antonova, L.A., Barcelona, J.F., Cabesaz, F.J., Cárdenas, D., Cárdenas-Toro, J., Castaño, N., Chacón, E., Chatelain, C., Dullinger, S., Ebel, A.L., Figueiredo, E., Fuentes, N., Genovesi, P., Groom, Q.J., Henderson, L., Inderjit Kupriyanov, A., Masciadri, S., Maurel, N., Meerman, J., Morozova, O., Moser, D., Nickrent, D., Nowak, P.M., Pagad, S., Patzelt, A., Pelser, P.B., Seebens, H., Sheng Shu, W., Thomas, J., Velayos, M., Weber, E., Wieringa, J.J., Baptiste, M.P., van Kleunen, M., 2017. Naturalized alien flora of the world. Preslia 89, 203–274. https://doi.org/10.23855/preslia.2017.203.
Reich, B.J., Hodges, J.S., Zadnik, V., 2006. Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. Biometrics 62, 1197–1206. https://doi.org/10.1111/j.1541-0420.2006.00617.x.
Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J. R. Stat. Soc. Ser. B Stat Methodol. 71, 319–392. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2008.00700.x https://doi.org/10.1111/j.1467-9868.2008.00700.x.
Rue, H., Riebler, A., Sørbye, S.H., Illian, J.B., Simpson, D.P., Lindgren, F.K., 2017. Bayesian computing with INLA: a review. Ann. Rev. Stat. Appl. 4, 395–421. https://doi.org/10.1146/annurev-statistics-060116-054045.
Simpson, D., Illian, J.B., Lindgren, F., Sørbye, S.H., Rue, H., 2016. Going off grid: computationally efficient inference for log-Gaussian Cox processes. Biometrika 103, 49–70. https://doi.org/10.1093/biomet/asv064.
Simpson, D., Rue, H., Riebler, A., Martins, T.G., Sørbye, S.H., 2017. Penalising model component complexity: a principled, practical approach to constructing priors. Stat. Sci. 32, 1–28. https://doi.org/10.1214/16-STS576.
Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. J. R. Stat. Soc. Ser. B Stat Methodol. 64, 583–639. https://doi.org/10.1111/1467-9868.00353.
Urdangarin, A., Goicoa, T., Ugarte, M.D., 2023. Evaluating recent methods to overcome spatial confounding. Revista Matemática Complutense 36, 333–360. https://doi.org/10.1007/s13163-022-00449-8.
Watanabe, S., 2013. A widely applicable Bayesian information criterion. J. Mach. Learn. Res. 14, 867–897.
Woodman, S.M., Forney, K.A., Becker, E.A., DeAngelis, M.L., Hazen, E.L., Palacios, D.M., Redfern, J.V., 2019. esdm: a tool for creating and exploring ensembles of predictions from species distribution and abundance models. Methods Ecol. Evol. 10, 1923–1933. https://doi.org/10.1111/2041-210X.13283.
Yu, H., Liu, X., Kong, B., Li, R., Wang, G., 2019. Landscape ecology development supported by geospatial technologies: a review. Eco. Inform. 51, 185–192. URL: https://www.sciencedirect.com/science/article/pii/S1574954119300329 https://doi.org/10.1016/j.ecoinf.2019.03.006.