UNIVERSITAT DE VALÈNCIA

Programa de Doctorado en ESTADÍSTICA Y OPTIMIZACIÓN

Spatio temporal modeling of species distribution

Óscar Rodríguez de Rivera Ortega

Supervisors:
Antonio López-Quílez
Marta Blangiardo

Octubre, 2018

Index

Acknowledgements

First of all, I would like to acknowledge my two outstanding supervisors, Professor Antonio López Quílez and Professor Marta Blangiardo. Since the first time I contacted Antonio when I was applying to the MSc, he has been a supportive teacher and mentor. It was Antonio who gave me the confidence to pursue a Doctorate, and guided me in the choice of this topic, which I have found fascinating. I have greatly valued his knowledge and attention to detail. Marta has also been a most inspiring mentor. Since I sent her my first email asking for a research period at the Imperial College London with “several conditions”, she showed me confidence and support. I have enjoyed our many discussions about projects and future plans. Also, she gave me the opportunity to participate in practical lessons as tutor, which has deepened my understandings. Both of them have been caring, wise, friendly and supportive, and my debt to them is enormous.

I highly appreciate the valuable contributions of the different researchers and technicians, who not only gave of their time generously, but also, provided me with the different datasets that made possible the different projects. I am also most grateful to other researchers who also gave time and data, but at the end -due to different circumstances- the projects could not be finished.

Colleagues from the University of Oviedo and Polytechnic University of Madrid have had a huge influence on my career, which is reflected in this study. In particular, I wold like to acknowledge Pablo and Aitor, who interested me in Ecological research in different ways. Asun, Marcos and researchers at Uniovi who showed me that other type of Professor is possible. And finally Nacho and Ruben for their wisdom, support and friendship over the years- it has been incredibly important.

I deeply look up on my parents, Paloma and Carlos. They had always been an example of “can do” attitude, effort and hard work. My deepest thanks go to Isa, Manu and Elvira, who have encouraged me through my years of study, even without knowing.

All these people and more have assisted me in many ways to successfully finish my PhD. They have made sacrifices and effort to allow me to pursue my dreams, and for that , I will always be grateful. I dedicate this work to them.

Last but not least, and never forgotten, I would like to remember all the researchers who rejected me, making possible this amazing project. And also, the Spanish Government's support to Research and Science that gave me important reasons to move to UK, and consequently, have the opportunity to work and develop my PhD during these years.

Summary

Species distribution models (SDMs) are numerical tools that combine observations of species occurrence or abundance with environmental estimates. The aim of these tools is to gain ecological and evolutionary insights and to predict distributions across landscapes, sometimes requiring extrapolation in space and time.

During the last years, decision support tools for species selection in Spain have been based on species distribution models (also called ecological niche models), that estimate the probability of occurrence of the species as a function of environmental predictors (e.g., climate, soil). The choice of the statistical method may have a dramatic effect on model performance; therefore comparisons of methods have received much interest in the last decades.

Modelling patterns of the presence/absence of species using local environmental factors has been a growing problem in Ecology in the last few years. This kind of modelling has been extensively used to address several issues, including the identification of essential fauna habitats in order to classify and manage conservation areas, and predicting the response of species to environmental features. Different approaches and methodologies have been proposed in this perspective during this time, most of them based in regression models from a classical perspective.

Several projects have focused on comparing these different methods. Most of these applications consist of explanatory models that seek to assess the relationship between environmental variables. Moreover, the theory of these methods is based on the fact that the observations are independent, while spatial autocorrelation is common in georeferenced ecological data. Spatial autocorrelation should be taken into account in the species distribution models, even if the data were collected through a standardized sampling scheme, since the observations are often close and subject to similar environmental features. In addition observer error, gaps in the sampling, missing data, and spatial mobility of the species can also affect the models.

As the statistical understanding of applied scientists increases and new techniques deliver larger, more complicated data sets, applied statisticians are faced with increasingly complex models. Naturally, as the complexity of these models increase, it becomes harder and harder to perform inference. Appropriately, a great deal of

effort has been expended on constructing numerical methods for performing approximate Bayesian inference. Undoubtedly, the most popular family of approximate inference methods in Bayesian statistics is the class of Markov Chain Monte Carlo (MCMC) methods. These methods, which exploded into popularity in the mid 1980's and have remained at the forefront of Bayesian statistics ever since, with the basic framework being extended to cope with increasingly more complex problems.

Hierarchical Bayesian models have traditionally relied on MCMC simulation techniques, which are computationally expensive and technically challenging, consequently limiting their use. However, a new statistical approach is now readily available, namely integrated nested Laplace approximations (INLA) via the R-INLA package. INLA methodology and its powerful application to modelling complex datasets has recently been introduced to a wider nontechnical audience. As opposed to MCMC simulations, INLA uses an approximation for inference and hence avoids the intense computational demands, convergence, and mixing problems sometimes encountered by MCMC algorithms. It can only be used for Gaussian models but this includes the class of models which we consider here for species distribution. Moreover, R-INLA can be compiled with the stochastic partial differential equations (SPDE) approach which through a discretisation of a continuous Gaussian field can cope efficiently with variables characterised by a complex spatial structure. This is the case of environmental inventories, since environmentalists or field workers start the inventory to target particular species, resulting in clustered spatial patterns and large regions without any values. Together, these new statistical methods and their implementation in R allow scientists to fit complex spatio-temporal models considerably faster and more reliably.

This Thesis presents species distribution models through different approaches, showing an evolution from classical models applied during the last years as MaxEnt, MARS or GAM, most of them based in regression models, to state-of-the-art methods currently used in several disciplines such as epidemiology or public health (INLA).

1 Introduction

1.1 Motivation

Understanding spatio-temporal dynamics of species is one of the main issues in many research areas (Martínez-Minaya et al. 2018). Species distribution models (SDM) commonly used in Ecology consist of numerical tools that combine observations of species occurrence or abundance with environmental covariates. They are used to gain ecological and evolutionary insights and to predict distributions across landscapes, sometimes requiring extrapolation in space and time (Elith & Leathwick, 2009). In SDM, the following steps are usually taken: (1) locations of occurrence of a species (or other phenomenon) are compiled; (2) values of environmental predictor variables (such as climate) at these locations are extracted from spatial databases; (3) the environmental values are used to fit a model to estimate similarity to the sites of occurrence, or another measure such as abundance of the species; (4) the model is used to predict the variable of interest across the study region (and perhaps for a future or past climate) (Hijmans & Ellith, 2015).

Modelling patterns of the presence/absence of the species using local environmental factors has been a growing problem in Ecology in the last few years (Chakraborty et al. 2010). This kind of modelling has been extensively used to address several issues, including the identification of essential fauna habitats in order to classify and manage conservation areas (Pressey et al. 2007), and predicting the response of species to environmental features (Midgley and Thuiller 2007; Loarie et al. 2008). Different approaches and methodologies have been proposed in this perspective (see for instance Guisan and Thuiller 2005; Hijman and Graham 2006; Wisz et al. 2008), with generalized linear and additive models (GLM and GAM) (Guisan et al. 2002), species envelope models such as BIOCLIM (Busby 1991) and the multivariate adaptive regression splines (MARS) (Leathwick et al. 2005) being some of the most commonly used (Muñoz et al. 2013).

Even though the limits of SDM for climate change impact assessments on complex ecological systems, it has been identified that species distribution models are conceptually well suited for simpler practical tasks: for instance in leading climate change adaptation strategies that involve habitat restoration or species selection for reforestation or forest management (Gray & Hamann, 2011, 2013; Hamann & Aitken, 2013; Schelhaas et al., 2015). For such management applications, the main task is to match source and target ambient. However, it is uncertain whether subsequent long-term forest growth and forest health are well described by species distribution models that may be used to guide initial decisions on species choice for a general geographic region (Maaten et al., 2016).

Currently the statistical understanding of applied scientists is increasing and new techniques can cope with larger, more complex data sets, so applied statisticians are faced with the need to specify sophisticated models. Logically, as the complexity of these models increase, it becomes harder to perform inference. The Bayesian approach is particularly appropriate as it is flexible and can deal with complex models, for instance including hierarchical structure or including missing data. Undoubtedly, the most popular family of approximate inference methods in Bayesian statistics is the class of Markov Chain Monte Carlo (MCMC) methods. These methods, which exploded into popularity in the mid 1980s have remained at the forefront of Bayesian statistics ever since, with the basic framework being extended to cope with increasingly more complex problems (Simpson et al. 2011).

Hierarchical models can simplify complex interactions by allowing parameters to vary at more than one level via the introduction of random effects. The expected value of the response is then expressed conditional on these random effects (Cosandey-Godin et al. 2014). The advantages of using hierarchical Bayesian models emerge more so as complexity increases, when, for example, spatio-temporal variability needs to be modelled explicitly (Cressie et al. 2009). The Bayesian framework also offers the advantage of providing full inference, such that model parameters and uncertainty can be quantified, which has great use in applied conservation (Wade 2000; Wintle et al. 2003).

Hierarchical Bayesian models have traditionally relied on MCMC simulation techniques, which are computationally expensive and technically challenging, consequently limiting their use. However, a new statistical approach is now readily available, namely integrated nested Laplace approximations (INLA) via the R-INLA package (http://www.r-inla.org) (Cosandey-Godin et al. 2014). INLA methodology and its powerful application to modelling complex datasets has recently been introduced to a wider nontechnical audience (Illian et al. 2013). As opposed to MCMC simulations, INLA uses an approximation for inference and hence avoids the intense computational demands, convergence, and mixing problems sometimes encountered by MCMC algorithms (Rue and Martino 2007). It can only be used for Gaussian models but this includes the class of models which we consider here for species distribution. Moreover, R-INLA can be compiled with the stochastic partial differential equations (SPDE) approach (Lindgren et al. 2011) which trough a discretisation of a continuous Gaussian field can cope efficiently with variables characterised by a complex spatial structure. This is the case of environmental inventories, since environmentalists or field workers start the inventory to target particular species, resulting in clustered spatial patterns and large regions without any values. Together, these new statistical methods and their implementation in R allow scientists to fit complex spatio-temporal models considerably faster and more reliably (Rue et al. 2009).

Some of the latest species distribution models only use the presences of the species in the modelling process. Other methods use presence/absence data or pseudo-absences. Logistic regression is the traditional approach to analysing presence/absence data (Hijmans & Ellith, 2015). Currently the statistical understanding of applied scientists is increasing and new techniques can cope with larger, more complex data sets, so applied statisticians are faced with the need to specify sophisticated statistical models. Logically, as the complexity of these models increase, it becomes harder to perform inference. The Bayesian approach is particularly appropriate as it is flexible and can deal with complex models, for instance naturally accounting for a hierarchical structure which could characterize the data or allowing for missing data imputation. Undoubtedly, the most popular family of approximate inference methods in Bayesian statistics is the class of Markov Chain Monte Carlo (MCMC) methods. These methods, which exploded into popularity in the mid-1980s have remained at the forefront of Bayesian statistics ever since, with the basic framework being extended to cope with increasingly more complex problems (Simpson et al. 2011).

1.2 Objectives

The aim of this thesis is to approach spatial distribution of different groups from different perspectives in order to analyse the different approaches to this problem.

This work is a trip from the classical approach, commonly used by ecologists, to more complex solutions, already applied in several disciplines.

We are focused in applying advanced modelling techniques in order to understand species distribution and species behaviour and the relationships between them and environmental factors and have used first the most common models applied in ecology to move then to more advanced and complex perspectives.

The aim of the first project is twofold. We present and explain, from a mathematical point of view, several common tools designed for species distribution modeling. We are going to be focused on how these tools develop models based on regressions, and explore the advantages and disadvantages of each model. Then, we are going to compare these models and decide which is the most accurate according to easily understood indicators.

During the second paper we are going to be focused in to build a spatial model to predict the spatial distribution of several species characterised by a low level of presences, which leads to data sparsity. We will use real data on five species of amphibians obtained from inventories developed in Las Tablas de Daimiel National Park (TDNP-Spain) in 2011-2012 supported with environmental variables. Our approach is to specify a Bayesian hierarchical geostatistical modelling framework accounting for spatial dependency.

Finally, during the third project, the aim is to build spatial and spatio-temporal models to predict the distribution of four different species present in the Spanish Forest Inventory. We want to compare the different models and show how accounting for dependencies in We will generate distribution models for each species. We will specify a Bayesian hierarchical geostatistical modelling framework accounting for spatial dependency.

1.3 Data

We can define two data sources according to the projects generated during the PhD. For the first and the third project we have used data from the National Forest Inventory of Spain and for the second project data from a Herpetological inventory developed in the Tablas de Daimiel National Park. We can define the dataset used and analysed as follow.

The main data used during the projects dedicated to Forest Inventories were obtained from the Spanish Ministerio de Agricultura, Pesca y Alimentacion. Some of them (data from IV Inventory) were requested as for the moment are not available due to the inventory was not finished.

For the first paper published, “Development and Comparison of Species Distribution Models for Forest Inventories” (de Rivera and López-Quílez, 2017), we have analysed data from the II Spanish National Forest Inventory. Dataset comprises a systematic grid with 91,889 plots. From the data available, we have analysed presence and absence of 17 forest trees species.

Also, we have analysed 10 climatic predictors commonly used in tree species autecology in Spain (Alonso et al., 2010): mean summer rainfall, mean annual rainfall, mean summer temperature, mean annual temperature, mean of maximum temperatures of the warmest month, mean of minimum temperatures of the coldest month, mean annual potential evapotranspiration, mean annual water surplus, and mean annual water deficit. All these data was obtained from the climatic data grids by applying the models for climatic estimation produced by Sánchez Palomares et al. 1999 to the Shuttle Radar Topography Mission (STRM) 3-arc-second (≈ 90 m) elevation dataset (Farr et al. 2009). These models interpolate monthly climate data from weather stations using latitude, longitude, and elevation as independent variables. Finally, we have used the European Soil Database (Panagos et al. 2012) to allocate each plot to a parent material class (calcareous or siliceous). The distribution of calcareous parent materials is a really useful predictor of plant species distribution in Mediterranean ecosystems (Gastón et al., 2009).

For the second project, titled “Species Distribution Modelling through Bayesian hierarchical approach” (de Rivera et al, 2017), was obtained directly from the inventory. Dr. Martin Sanz developed the inventory capturing data of presence/absence of the species and also the ambient where the point was captured. Coordinates were obtained from GPS points captured during the inventory.

The data set come from an inventory developed in Las Tablas de Daimiel National Park during 2011 and 2012, comprising 234 sample points with coordinates. Each sample point has the presence or absence of each species, elevation in meters and information about the ambient (categorical variable with the following categories: Salt marsh, Reed bed, Islands, areas of Typha latifolia, Cladium mariscus and free of vegetation).

For the third project, titled “Assessing the spatial and spatio-temporal distribution of forest species via Bayesian hierarchical modelling” (de Rivera et al, 2017), we have analysed data from the II, III and IV Spanish National Forest Inventory. In this case we have analysed only data from the province of Galicia, and only four species

In this case we have analysed the environmental variables from Herrera et. al. (2012 and 2016) and are those typically considered in this type of studies: mean annual temperature, mean of maximum temperatures of the warmest month, mean of minimum temperatures of the coldest month and mean annual rainfall. We also considered the distribution of calcareous parent materials is a useful predictor of plant species distribution in our study area (Gastón et al., 2009). We used the European Soil Database (Van Liedekerke et al., 2006) to allocate each plot to a parent material class (calcareous or siliceous).

1.4 Software

As we have commented previously, we have approached the species distribution model problem from several approaches.

The main software used was R statistical software and some packages/extensions. Moreover, during the first project we have worked with several software in order to understand and evaluate different aspects of species distribution models.

During the first project, model selection was based on ease of working with each model and the possibility of repeating each process with the same characteristics using the species studied. We therefore constructed the models with one of the most widely used models (MaxEnt), and others based on simple software developed by Saldorf System (CART and MARS); finally we built an additive model using R software.

1.5 Papers

The projects developed have been presented in different international meetings and divulgated in different international journals.

“Development and Comparison of Species Distribution Models for Forest Inventories” (de Rivera and López-Quílez, 2017), focused on model comparison using the most common species distribution models and comparing these against an additive model with thin plate splines was presented and published on June 2017, on the International Journal of Geo-Information, 6 (6), 176.

“Assessing the spatial and spatio-temporal distribution of forest species via Bayesian hierarchical modelling”. (de Rivera et al, 2018a), is an approach to species distribution models from a Bayesian perspective. Using reptiles and amphibians data we have analysed the spatial distribution using environmental variables. This project was presented during the Autumn Meeting on Latent Gaussian Models (2015) in Trondheim (Norway) and published on Theoretical Ecology, pp.1-11. In press.

Finally, “Assessing the spatial and spatio-temporal distribution of forest species via Bayesian hierarchical modelling” (de Rivera et al, 2018b). In this case, results were presented during the Statistical Ecology Research Fest (2016) in Canterbury (United Kingdom) and published on Forests, 9(9), p.573.

2 Methodology

The steps followed during this research were from the application of known techniques applied in population's research to more complex approaches used successfully in other disciplines.

We are going to summarise the different methods used. This summary is going to be a first approach that is going to be properly explained in the different papers attached. Below are the ways used to approach the problem, from frequentist to the final goal, the spatio-temporal model.

2.1 Usual Species Distribution Models

a) MaxEnt (maximum entropy) (Elith et al., 2006, Phillips et al. 2008)

MaxEnt is an artificial intelligence method based on the statistical principle of maximum entropy. Models are limited by the value of the variables used to develop the problem. For example, the expected value (mean value predicted by the model) of each independent variable must match its empirical average (mean value observed when sampling with an independent variable occurrence data item).

MaxEnt obtains the maximum entropy probability of distribution, in other words, the distribution nearest to the uniform distribution, with all the conditions.

MaxEnt is based on the following points: a) the presence of a species is represented by a likelihood function on a set of points in the study zone. The likelihood function gives a positive value everywhere so that the sum is the unity; b) building a model of the function with a group of constraints obtained from empirical data of presence; c) the restrictions are expressed as a simple function of known environmental variables; d) in the MaxEnt method, the average forces of each function of each variable are close to the actual average of the variable zones of presence; e) of the possible options available, a specific combination of features is selected to minimize the entropy function (measured by the Shannon index). The entropy function allows optimal selection of variables and functions based on their significance, and eliminates restrictions that do not provide the model with significance.

b) MARS (multivariate adaptive regression splines)

MARS is a statistical method developed by (Friedman, 1991). It involves designing flexible models in which the data are adjusted to partial regressions. When models are nonlinear, they are approximated by partial linear regression, where the grade of the equation changes from one step to another, establishing a node between the end of one linear regression and the beginning of the next.

A node indicates the end of one partial regression and the beginning of another. Between two consecutive nodes, logically the model is defined by a linear regression. The nodes are selected with the aid of a search procedure that generates a stepper algorithm. The model generated is overfitted, so the less relevant nodes are subsequently removed using a statistical approach known as generalized cross validation. Finally, we only considered the most significant nodes.

c) CART (classification and regression trees)

This method was established by Breiman et al. (1984) and generates binary trees (parent nodes are divided into two child nodes) by iterative partitions, in a process that can be repeated to attempt to turn each child node into a parent node. The algorithm searches for the optimal cutoff values among all the independent variables to obtain an optimal set of binary divisions, so as to minimize the variance within each node and maximize it between different nodes; it is therefore possible that some variables will be unused. Once the tree that best classifies the cases has been identified, with no limits on complexity, the algorithm 'prunes' or simplifies to avoid overfitting of the data. The result is a tree that establishes yes/no questions. Depending on the kind of dependent variable can be two types of trees: regression (continuous dependent variable) and classification (discrete variable).

The most important advantages of classification and/or regression trees are (Schiattino & Silva, 2008): a) structured knowledge is obtained in the form of classification rules or the values of a variable interval. This knowledge is easy to interpret, and in simple language characterizes the classes or values of a variable interval; b) as it is a nonparametric analysis (distribution free procedure), it requires no distributional assumptions to validate probability; c) it allows working with all types of predictor variables: binary, nominal, ordinal and interval or ratio; d) it allows unknown values for the predictor variables in the individuals, both in the construction phase and in the tree prediction; e) in the case of classification probability, it can be set to a priori classes; f) the observations can be weighed using an ad-hoc variable.

2.2 Generalized additive model with thin plate splines

a) Model structure

A generalized additive model is a generalized linear model in which the linear predictor be determined by linearly on unidentified smooth expressions of some variables, and interest focuses on inference about these smooth expressions. Additive models were originally built by (Schiattino & Silva, 2008) to combine properties of linear models with additive models.

A smoother is an instrument for summarizing the tendency of a dependent variable as an expression of one or more independent variable. It generates an estimate of the tendency that is less mutable than Y itself; therefore the name ‘smoother’.

The most significant characteristic of a smoother is its non-parametric nature, so the smooth function is also known as non-parametric function. Its biggest difference from the Generalized Linear Model is that it does not undertake an inflexible form for the function’s dependence on variables. It allows an approach with the addition of expressions (expressions that have separated input estimates), not just with one indefinite expression only. For this reason it is the building block of the generalized additive model algorithm (Liu, 2008)

Testing the different types of splines reveals that the best model helped with the AIC value is the Additive model with thin plate regression splines. The thin plate spline is the two-dimensional equivalent of the cubic spline in one dimension. It is the essential resolution to the biharmonic equation.

Assumed a dataset of points, a weighted mixture of thin plate splines concentrated about each point gives the interpolation expression that passes through the points precisely while reducing the so-called ‘bending energy.’ Bending energy is defined here as the integral over R2 of the squares of the second derivatives. Regularization should be used to decrease the necessity that the interpolant pass through the data points exactly.

The designation of ‘thin plate spline’ is a physical analogy referring to the flexible of a thin sheet of metal. In the physical situation, the deflection is in the z direction, at right angle to the plane. In order to apply this impression to the problem of coordinate conversion, the lifting of the plate is interpreted as a dislocation of the x or y coordinates within the plane (Donato & Belongie, 2002)

These splines are short rank isotropic smoothers of any number of variables. The splines are isotropic because any variation of the covariate co-ordinate system will not modify the output of smoothing. The low rank means that they have rarer coefficients than there are data to smooth. They are the default smooth for ‘s’ terms due to there is a clear logic in which they are the ideal smoother of any given basis measurement/rank (Wood, 2003).

In this case, as we are building the model with R we used the mgcv package (Wood 2000, 2003, 2004, 2006, 2011) to construct the additive model.

b) Model comparison

The area under the receiver operating characteristic (ROC) function (AUC) is taken to be an important index because it provides a single measure of overall accuracy that is independent upon a particular threshold (Deleo, 1995). If the objective is to rank the classifiers, comparisons using ROC plots are more robust since they are not dependent of the values in a confusion matrix (Fielding & Bell, 1997). An ROC graph is a method for visualizing, establishing, and selecting classifiers based on their presentation. ROC curve analysis was developed during World War II as a tool in signal processing, and is now used in many branches of science. Standard references for ROC curve analysis are (Fielding & Bell, 1997; Metz, 1978; Hanley & McNeil, 1982; Murphy & Winkler, 1992; Pearce & Ferrier, 2000; Marzban, 2004).

Although ROC graphs are conceptually simple, their application in research contexts gives rise to some complexities that are not obvious and their practical use entails some common misconceptions and pitfalls (Fawcett, 2005).

ROC graphs are two-dimensional graphs where the true positive rate is presented on the Y axis and the false positive rate is presented on the X axis. An ROC graph represents relative adjustments between profits (true positives) and expenses (false positives). Figure 2 shows the area under two ROC curves, A and B. Classifier A has a greater area and, therefore, better average performance.

Finally, it is probable for a low-AUC classifier to perform better in a specific region of the ROC space than a high-AUC classifier. Figure 2 shows an example of this: classifier B is generally worse than A, except at an fp rate > 0.6 where B has an insignificant advantage. However, in practice the AUC performs very well and is often used when a general measure of predictiveness is desired.

In order to analyse suitability of the different models we have used the ROCR package (Sing et al. 2012) to obtain the validations, and AUC values and graphics.

2.3 Geostatistical model

In several fields of research, researchers analyse data geographically referenced. These data are called spatial data and we can identify three areas: lattice data, point-reference (or geostatistical data) and spatial point patterns (Blangiardo and Cameletti 2015; Giraldo, 2002).

Area or Lattice data: locations belong to a discrete set and are selected by the researcher. These can be regular or irregularly spaced. Usually area is typically irregular and based on administrative boundaries and the second one is regular (grid). In these cases we are interested in mapping an outcome over the area analysed.

Spatial Point Patterns: locations belong to a set that can be discrete or continuous and their selection does not depend on the researcher. For example, we might be interested in the locations of trees of a species in a forest.

Geostatistical data: locations come from a continuous set and are selected according to the researcher's judgment. In this case we are looking in predicting the outcome at unobserved locations.

These models can be specified in a Bayesian framework extending the concept of hierarchical structure, letting us to account for connexions based on distance or relationships between neighbours.

a) Model structure

Spatial data are defined as realisations of a stochastic process indexed by space:

$$Y(s) \equiv \{y(s),\, s \epsilon D\}$$

where D is a (fixed) subset of R^d (here we consider d = 2). The actual data can be then represented by a collection of observations y = {y(s_1), ..., y(s_n)}, where the set (s_1, ..., s_n) indicates the spatial units where the measurements are taken. Depending on D being a continuous surface or a countable collection of d-dimensional spatial units, the problem can be specified as a spatially continuous or discrete random process, respectively (Gelfand et al., 2010). In our case, we can consider a collection of data points with presence/absence obtained from the inventory and the sampled points are the set (s_1, ..., s_n) of n points; y_s is the presence of each specie in each point and it is specified as

$$y_s \sim Bernoulli(\pi_s)$$

where π_s is the probability of the species being present.

Then on the logit(π_s) a linear model is specified including the different covariates, x_{ms} (Temperatures, precipitation, soil and elevation) and a spatial field ξ_s

$$\text{logit}\ (\pi_s) = b_0 + \sum_{m=1}^{M} \beta_m\, x_{ms} + \xi_s$$

where a discretely indexed spatial random process (see Lindgren et al. 2011) is included to approximate the continuous process. The key idea of the SPDE approach consists in defining the continuously indexed Mat´ern GF ξ(s) as a discrete indexed GMRF by means of a basis function representation defined on a triangulation of the domain D

$$\xi_s = \sum_{g=1}^{G} \varphi_g(s)\, \tilde{\xi}_g$$

Here G is the total number of vertices in the triangulation, $\{\varphi_g\}$ is the set of basis functions and $\{\tilde{\xi}_g\}$ are zero-mean Gaussian distributed weights. The basis functions are chosen to be piecewise linear on each triangle, i.e. ϕg is 1 at vertex g and 0 elsewhere. Notice that we use the formal notation ξ_s in the left-hand side of the expression since SPDE provides a representation of the whole spatial process (defined for any point s) that varies continuously in the considered domain D^4 (Blangiardo and Cameletti 2015).

b) Implementation

We have used the Integrated Nested Laplace Approximation (INLA) implemented in the R-INLA package to be used from within R statistical software. In R-INLA the first step required to run the geostatistical spatial model with only one covariates (M = 1 represented by elevation or vegetation), is the triangulation of the considered spatial domain. We use the inla.mesh.create specifying the spatial coordinates used for estimation. The inla.mesh.create performs a constrained refined Delaunay triangulation for a set of spatial locations: firstly the triangle vertices are placed at the observation locations and then further vertices are added in order to satisfy triangulation quality constraints (Lindgren et al., 2011). Depending on the values chosen for inla.mesh.create arguments, the total number of vertices changes with a trade-o between the accuracy of the GMRF representation and the computational and time costs. We can summarise the process as follows, with a similar approach than explained in Blangiardo et al. 2013.

With the setting used above we obtain a mesh with 3328 vertices, which can be obtained in the R terminal by typing `mesh$n`. Given the mesh, we have created the spde model object, to be used later in the f() term in the R-INLA formula, with the expression

```
spde=inla.spde2.matern(mesh=mesh)
```

We have used now the helper function `inla.stack` which builds the necessary matrices required by the SPDE approach and of combining the data, the observation matrix A and the linear predictor η; some details about the usage of the inla.stack function can be found also in Cameletti et al. (2011). Before employing `inla.stack`, we create the object `A.est` which corresponds to ~A

```
A.est = inla.spde.make.A(mesh, loc=loc)
```

and is a 2000 3328 sparse matrix that extracts the values of the latent spatial field at the observation locations. Moreover, we generate the required vectors of indices

```
field.indices=inla.spde.make.index("field",n.mesh=mesh$n)
```

with `field.indices` being a list whose first component is called field and contains the spatial vertex indices (i.e, the sequence of integers from 1 to 3328). Finally, we call the `inla.stack` function that takes in input the data (`data`), an identification string (`tag)` and the components of the observation matrix (`A`) and of the linear predictor (`effects`), combined together in list-type objects:

```
stack.est <- inla.stack(data=list(presence=species),
A=list(A.est,1,1,1,1,1,1,1), tag='est',
effects=list(field=field.indices, altII=inventory$Z,
soilII=inventory$A, PreII=inventory$Ppr,
TasII=inventory$Tas, tasMAXII=inventory$tasMAX,
tasMINII=inventory$tasMIN,intercept=rep(1,length(species)
)))
```

Note that each term in `A` has its own linear predictor component in the effects object so that, for example, `A.est` is paired with the list composed by `field.indices` and Intercept=1 (this may seem a little strange but it is due to how the SPDE related functions are internally coded). Similarly, we create the corresponding objects inla.val and stack.val for the validation stations with the only difference that, since we are interested in prediction, we have specified `data=list(presence=species)` in the `inla. stack` function.

```
stack.pred<- inla.stack(data=list(presence=species),
A=list(A.est,1,1,1,1,1,1,1), tag='pred',
effects=list(field=field.indices,altII=inventory$Z,soilII
=inventory$A,PreII=inventory$Ppr,TasII=inventory$Tas,tasM
AXII=inventory$tasMAX,tasMINII=inventory$tasMIN,intercept
=rep(1,length(species))))
```

Finally, we combine all the data, effects and observation matrices using the command

```
stack=inla.stack(stack.est, stack.pred)
```

In the `R-INLA formula` we include the `spde` model object named `field`; moreover, note that, due to the way `inla.stack` works, we need to specify an explicit `Intercept` term and remove the automatic intercept with `-1`.

```
Formula  = presence ~ -1 + intercept + alt + soil + Pre +
TasII + tasMAX + tasMIN + f(spatial.field, model=spde)
```

Finally, we can run the specified model calling the `inla` function as follows:

```
mod=inla(formula, data=inla.stack.data(stack,spde=spde),
family="binomial",
control.predictor=list(A=inla.stack.A(stack),
compute=TRUE), control.compute=list(dic=TRUE,waic=TRUE))
```

the functions `inla.stack.data` and `inla.stack.A` simply extract the required data and the observation matrix from the stack object. The option `compute=TRUE` is required to obtain the marginal distributions for the linear predictor. We retrieve the posterior summary statistics of the fixed effects a and b from the object `mod$summary.fixed`, while the posterior marginal of the precision $\tau_e = 1/\sigma_e^2$ is included in the list mod$marginals. hyperpar. If we are interested in the variance σ_e^2, we employ the function `inla.emarginal` for computing the expected value of the (reciprocal) transformation of the posterior marginal distribution. The results on the parameters of the Matèrn spatial covariance function can be obtained typing

```
mod.field=inla.spde2.result(mod, name="spatial.field",
spde)
```

where the string name refers to the name of the spde effect used in the inla formula.

Applying the suitable transformations through the inla.emarginal function as described in Cameletti et al. (2011), we obtain the posterior estimates for the spatial variance σ_e^2 C and for the range r. Then, we extract the linear predictor values on the mesh

```
index.pred=inla.stack.index(stack,"pred")$ data
lp.mean.pred=mod$summary.linear.predictor[index.pred,
"mean"]
lp.sd.pred=mod$summary.linear.predictor[index.pred, "sd"]
```

c) Model evaluation

A natural way to estimate out-of-sample prediction error is cross-validation (see Geisser and Eddy, 1979, and Vehtari and Lampinen, 2002, for a Bayesian perspective), but researchers have always sought alternative measures, as cross-validation requires repeated model and can run into trouble with sparse data (Gelman et al. 2013). In a comparative perspective (e.g. to evaluate which model the data best) the most used index is the DIC (Spiegelhalter et al., 2002, van der Linde, 2005) which consists of two components, a term that measures goodness of t and a penalty term for increasing model complexity.

More recently the WAIC (Watanabe, 2010) this approach has been proposed as a suitable alternative for estimating the out-of-sample expectation is a fully Bayesian approach. This approach starts with the computed log pointwise posterior predictive density and then adds a correction for the efective number of parameters to adjust for overfitting (Gelman et al. 2013). WAIC operates on predictive probability density of observed variables rather than on model parameter, hence it can be applied in singular statistical models (i.e models with non-identifiable parameterization (Li et al. 2015).

2.4 Extension spatio-temporal

Analysing only the spatial patter or ecological processes does not allow saying anything about their temporal variation which could even more interesting than only the spatial distribution. The objective is to understand and estimate a spatial varying phenomenon. If we consider the data aggregated over the time we can only model the spatial pattern, but if we disaggregate the data by time, we can now investigate a temporal trend.

The concept of spatial process can be extended to the spatio-temporal case including a time dimension. The data are then defined by a process

$$Y(s,t) \ \equiv \ \{y(s,t),(s,t) \,\epsilon\, D \,\epsilon\, \mathbb{R}^2 \times \mathbb{R}\,\}$$

As we define in the spatial model, we can consider a collection of data points with presence/absence obtained from the inventory and the sampled points are the set (s_1, ..., s_n) of n points; y_{st} is the species presence at each point in space and time, specified as

$$y_{st} \sim Bernoulli(\pi_{st})$$

where π_{st} is the probability of the species being present.

Then on the logit(π_{st}) a linear model is specified including the different covariates, x_{ms} (Temperatures, precipitation, soil and elevation) and a spatio-temporal field ω_{st}

$$\text{logit}(\pi_{st}) = b_0 + \sum_{m=1}^{M} \beta_m \, x_{ms} + \omega_{st}$$

where ω_{st} refers to the latent spatio-temporal process that changes in time with autoregressive dynamics and spatial correlation innovations, which we model as follows:

$$\omega_{st} = a\omega_{s(t-1)} + \xi_{st}$$

with t=2, …T, $|a|$<1 and $\omega_{s1} \sim Normal(0, \frac{\sigma^2}{1-a^2})$. ξ_{st}, is a zero-mean Gaussian field temporally independent with the following spatio-temporal covariance:

$$\mathrm{Cov}(\xi_{st}, \xi_{ju}) = \begin{cases} 0, & t \neq u \\ \mathrm{Cov}(\xi_{st}, \xi_{ju}), & t = u \end{cases}$$

for $i \neq j$, where $\mathrm{Cov}(\xi_{it}, \xi_{ju})$ is modeled through the Matern spatial covariance function.

To implement this model in R-INLA, we need to define a similar process to the spatial one, including the time in the expressions. To obtain the temporal differentiation we define a number of groups based in the dates of the inventories.

3 Results

In order to summarize results in this chapter we are going to show results from the process or technical perspective, avoiding, as far as possible, explanations about ecological findings not related with the models per se. As we said in the beginning we have approached similar problems from different perspectives.

From a general perspective and comparing the different models applied during the process, from MaxEnt to spatio-temporal models with INLA, we can affirm that the models that we have developed show better results that the already built. Also, it is difficult to compare between the different approaches, but the Bayesian approach shows more flexibility and also the inclusion of spatial field or the latent spatio-temporal process give in a way the possibility to understand that variables that are not possible to evaluate, can be included in this expression as a residual.

Results obtained with INLA in both projects show interesting information about how the different species are related to different environmental variables. As we said before, in this summary we are not looking to explain how the species are related to these variables, our main objective is to show the results related to the different models used.

Below we are going to summarise results of individual papers to identify the goals obtained during the different projects.

3.1 Development and Comparison of Species Distribution Models for Forest Inventories (Rivera and López-Quílez, 2017)

For the first study we performed with a large number of species revealing some important results.

A two-way ANOVA shows that all the environmental variables included in the models are significant for all measures of performance ($P < 0.05$). All models designed have good predictions and obtain high AUC values.

Analyzing predictability, based in AUC we have obtained that: for all the species analyzed, MARS and MaxEnt are the models with the lowest predictability and consequently with the lowest AUC average; however, CART and GAM generally have the highest AUC values.

The following ecological modelling methods are compared: MARS and MaxEnt, CART and GAM.TP. The scatterplot graph shows the different models' behavior, demonstrating that all have good predictability based on their AUC value.

For the different species, all the statistical models show similar behaviour and they performed in the same way. Moreover, comparing model predictability, the AUC values in GAM.TP have better results on average than the others.

Analysing the results, species with the highest number of presences have lower values in the predictions due to a wide range of the environmental variables. In contrast, the species with the more absences have the highest AUC values, perhaps due to the representative environmental characteristics that give rise to the presence of these species.

If we analyse presence-absence from the dataset and compare with the AUC average, we find that the relationship between AUC and percentage of presence is negative (based on the correlation index), with a value of -0.75. Species with the highest percentage of presence have lower values of AUC than other less represented species in the area of study

In summary, every AUC value obtained with those models is significant and all the models could be useful to represent the distribution of each species.

Overall, the Additive model with thin plate splines gave the best results. MaxEnt, CART and GAM.TP with thin plates splines obtained similar AUC values.

The worst capability was obtained with MARS. This model's performance was below average for several species.

The models we developed obtained better results because they allowed for changes and calibrations. In this case we were aware of all the processes that occurred during the modelling. By contrast, models obtained using specific software in general performs like "hermetic machines", because it could sometimes be impossible to understand the stages followed toward the final results.

3.2 Species Distribution Modelling through Bayesian hierarchical approach (de Rivera et al, 2018a)

During the second project we have analysed species with low number of presences, comparing the different models using WAIC (Watanabe, 2010). In this case, there are not big differences between models, only we can affirm that the model with a smaller number of environmental variables has better fit.

We have also calculated the conditional predictive ordinate (CPO) (Pettit, 1990) to evaluate model assessment. The conditional predictive ordinate (CPO) is based on leave-one-out-cross-validation. CPO estimates the probability of observing a value after having already observed the others. The mean logarithmic score (LCPO) was calculated as a measure of the predictive quality of the model (Gneiting and Raftery, 2007; Roos and Held, 2011). High LCPO values suggest possible outliers, high-leverage and influential observations.

Finally, we have used an AUC (Area Under operating Curve score) approach to calculate the predictive accuracy of each method by comparing the validation data with the predicted presence value. AUC represents a commonly used and adequately performing measure of predictive accuracy (Huang and Ling, 2005) and works by calculating the relative numbers of correctly and incorrectly identified predictions across all possible classification threshold values of the binomial response, with an AUC value equal to or below 0.5 indicating a predictive ability equal to random expectation and 1 a perfect predictive ability (Qiao et al. 2015).

Analysing the different results obtained, we can affirm that the first model obtained using only the elevation has a better fit than the model with elevation and ambient. However, based on LCPO the model with ambient has fewer outliers. Also, we have compared performance of the different models based in AUC, these analysis shows similar results than LCPO, obtaining better values in models without ambient.

Looking at the WAIC, most of the species have a better fit for the model with vegetation (except *Bufo bufo* and *Pelobates cultripes*), but looking at LCPO values Ambient seems to increase the number of outliers. Also, looking at AUC values, models with ambient have lower predictability. However, results are really similar across both models.

Finally, we can see that hierarchical models are particularly useful when data are sparse or species are similar. In our case Amphibia model, has different relationship with the environmental variables than the individual species included in the class model.

3.3 Assessing the spatial and spatio-temporal distribution of forest species via Bayesian hierarchical modelling (Rodriguez de Rivera et al, 2018)

Thinking in the second project as a test to apply this approach with a large number of presences and also in a bigger area, we applied the knowledge acquired to analyse the presence of different species of forest trees in Galicia (Spain). In this case we compare spatial and spatio-temporal models in order to understand which one could be the best approach. From general perspective we can affirm that the models, comparing the same species using WAIC, have similar fit, and also the outputs obtained are really similar. Moreover, we have seen interesting results when comparing the relationship between variables and presence of the species. With all the species, the inalterable variables show similar relationship in spatial and spatio-temporal models, but this is not the case with variables that change along the time.

Finally, as we said, spatial and spatio-temporal models show similar output. Moreover, the problems that we are approaching are dynamic situations with several changes along the time, so the difference between use a spatio-temporal model instead of a spatial model is based in the continuity of the process, avoiding understand each period of the analysis (in our case each inventory), as an independent process.

Most of the species show different relationships with environmental and climatic variables between spatial and spatio-temporal models. Also, if we generalise, species with more presences show larger differences between models. Also if we analyse the results from spatial to spatio-temporal models typically variables not showing a clear effect become positively or negatively associated with presence, depending of the species and the variable.

There are interesting differences between spatial and spatio-temporal models for the different species. As we have shown, not always the same variables have the same weight in the different models.

As we can see the behavior of the species is really different according to the characteristics and traditional uses and strategies, probably do to the impact of unobserved variables.

Finally, analysing the models we can affirm that the use of spatio-temporal models is an advantage for the understanding of the different ecological dynamics, giving the temporal perspective, not really frequent in environmental research projects.

4 Conclusions

Looking at the experience obtained during this research as learning process, we can conclude the following points:

If we analyse the models separately, from a frequentist approach, additive models with thin plate splines may be considered one of the greatest methods to analyse species distribution models working with presence-absence data, comparable to MaxEnt, CART and MARS. Our results show a better fit and more flexibility in the design.

Looking at the quality of the data and the possibility to work with presence/absence values and also with a systematic survey, we can confirm, looking our results, that the information obtained from the absences could be more important than the presences. Analysing this result from an ecological perspective, absences deliver of the species due to the combination of several environmental predictors.

Finally, we understand that there are more advanced approaches to apply in species distribution model, most of them through Bayesian approach (i.e R-INLA can be compiled with the stochastic partial differential equations (SPDE) approach (Lindgren et al. 2011) which through a discretisation of a continuous Gaussian field can cope efficiently with variables characterised by a complex spatial structure), but our objective along the first project was show the interesting opportunities that offer these explanatory techniques seek to assess the relationship between environmental variables.

As we said in the beginning of this conclusions this research period has worked as a learning process, and as a natural process we have realised that the Bayesian approach could be a better solution or at least a different approach for consideration.

The main advantage of the Bayesian model formulation is the computational ease in model fit and prediction compared to classical geostatistical methods. The main goal of this study has been to predict the occurrence of species with a relatively small number of data points, but the data was useful to show the power of this kind of process and the options of the model construction. To do so, instead of MCMC we have used the novel integrated nested Laplace approximation approach. More precisely, we have applied the work of Lindgren et al. (2011), which provides a link between Gaussian Fields and Gaussian Markov Random Fields through the Stochastic Partial Differential Equation (SPDE) approach. The SPDE approach can be easily implemented providing results in reasonable computing time (comparing with MCMC). We showed how SPDE is as useful tool in the analysis of species distribution. This modelling could be expanded to the spatio-temporal domain by incorporating an extra term for the temporal effect, using parametric or semiparametric constructions to reflect linear, nonlinear, autoregressive or more complex behaviours.

We conclude that SPDE and INLA are promising tools to work with species distribution model as they save in computational times and are easy to specify and to implement also for non-statistician when we work with a large data set.

There are interesting differences between spatial and spatio-temporal models for the different species. As we have shown, not always the same variables have the same weight in the different models.

As we can see the behaviour of the species is really different according to the characteristics and traditional uses and strategies. Moreover, other aspects as forest fires can be the reason of these changes in the distribution.

Comparing spatial and spatio-temporal models we can affirm that the use of spatio-temporal models is an advantage for the compression of the different ecological processes, giving the temporal perspective, not really common in environmental research projects.

An interesting point is that we have analysed the credible interval of the different variables from a frequentist point of view in order to understand the relationship between environmental variables and species presence. We can see that some variables change their “weight” depending of the inventory and also, several variables have the same behaviour in all the inventories and also along the spatio-temporal model.

Summarizing we can generalize that permanent and theoretical inalterable variables have similar performance in spatial and spatiotemporal models, showing similar relationship between presence of species and this variables along the time. Moreover, not always species presence has similar relationship with “non static” variables. This relationship is changing not only due to changes in environmental factors, but also based on species management and possible human disturbances.

Finally, we can conclude that Bayesian approach and particularly spatio-temporal models are really interesting approaches for the understanding of environmental dynamics, not only because of the possibility to develop and solve more complex problems but also for the easy understanding of the implementation processes.

5 Future lines of research

Spatial and Spatio-Temporal models through Integrated Nested Laplace are statistical techniques widely used in several fields of knowledge. The application of this approach to try to understand ecological problems is emerging and from this point of view, several “concerns” could be analysed through this Bayesian approach.

Development of populations of species and their behavior against the different problems caused by the Climate change can be a starting and generic point.

Particularly, understand important species behavior, i.e. pollinators, could be an interesting problem to analyse looking for a correct management of species and disturbances. In order to understand the problem we can approach that insect pollination is vitally important to terrestrial ecosystems and to crop production. The oft-quoted statistics are that 75% of our crop species benefit from insect pollinators (Klein et al. 2003), which provide a global service worth $215 billion to food production (Gallai et al. 2008). Hence, the chance that we may be facing a “pollination crisis” (Holden, 2006), in which crop yields begin to fall because of inadequate pollination, has generated understandable debate and concern and stimulated much research in recent decades. Nonetheless, knowledge gaps remain substantial with regard to both the extent and causes of pollinator declines.

Recent researches highlighted by Goulson et al. 2015 show the principal drivers of bee declines are: habitat loss, parasites and disease, pesticides, monotonous diets, competition and climate change. For most of them there are data base available to analyse and quantify the weight of each factor in bees decline.

Apart of this an as a part of pending work “discovered” during the last project another interesting approach is understand through this approach ecological hazards that can affect valuables resources. In this case, understand causality of forest fires, pest and diseases or flooding risk could be a challenging project to create interesting tools in order to improve management of resources. In this case forest fires could be an interesting subject to analyse from a causality perspective through the time.

Every year, over 60.000 forest fires are affecting Europe, mainly in the southern countries with a Mediterranean climate, burning more than 0.6 million hectares of vegetation. Most forest fires in these countries are human caused and burn virtually everywhere across their geography, through any type of vegetation, causing important damages to human and environmental assets (Spano et al. 2014).

Several models have been created to create fire danger mapping, most of them with a GIS approach (Chuvieco and Salas, 1996). Also, GIS has been applied to other fire management topics. The most promising are the location of look-out towers (Pawlina et al. 1990), dispatch planning (Salazar and Power 1988), ecological evolution after fire (Lowell and Astroch 1989), and fire growth simulation (Davis and Burrows 1990, Vasconcelos and Guertin 1992).

However, none of them where focused in understand the forest fire and his causality. An ideal approach could be to try to understand the forest as a disease, applying epidemiological concepts. Forest fire causality includes social, economic and ecological factors that affect rural and urban landscapes of the Mediterranean and reveal that forest fires are not the cause but the consequence of matters that go much beyond forests, and the smoke produced by flames. Understanding well the entire phenomenon must let us to identify potential solutions and even how we can take part in these solutions as citizens (Plana et al. 2016).

6 References

1. Alonso Ponce, R., López Senespleda, E., Sánchez Palomares, O., 2010. A novel application of the ecological field theory to the definition of physiographic and climatic potential areas of forest species. European Journal of Forest Research 129, 119-131.
2. Blangiardo M, Cameletti M., 2015. Spatial and Spatio-temporal Bayesian Models with R-INLA. WILEY
3. Blangiardo, M., Cameletti, M., Baio, G. and Rue, H., 2013. Spatial and spatio-temporal models with R-INLA. Spatial and spatio-temporal epidemiology, 7, pp.39-55.
4. Breiman, L., Friedman, F.H., Olshen, R. A. & Stone, C. J., 1984. Classification and regression trees. Wadsworth and Brooks. Pacific Grove, CA, USA.
5. Busby JR., 1991. BIOCLIM: A bioclimatic analysis and predictive system. In: Margules C, Austin M (eds) Nature conservation: cost effective biological surveys and data analysis, CSIRO, Canberra , 64-68
6. Cameletti, M., Ignaccolo, R. and Bande, S., 2011. Comparing spatio-temporal models for particulate matter in Piemonte. Environmetrics, 22(8), pp.985-996.
7. Chuvieco, E. and Salas, J., 1996. Mapping the spatial distribution of forest fire danger using GIS. International Journal of Geographical Information Science, 10(3), pp.333-345.
8. Cosandey-Godin A, Teixeira Krainski E, Worm B, Mills Flemming J., 2015. Applying Bayesian spatiotemporal models to fisheries bycatch in the Canadian Arctic. Can. J. Fish. Aquat. Sci. 72: 1–12
9. Cressie N, Calder CA, Clark JS, Hoef JMV, Wikle CK., 2009. Accounting for uncertainty in ecological analysis: The strengths and limitations of hierarchical statistical modeling. Ecol Appl 19, 553-5701
10. Davis, F.W. and Burrows, D.A., 1994. Spatial simulation of fire regime in Mediterranean-climate landscapes. The role of fire in Mediterranean-type ecosystems, pp.117-139.
11. de Rivera, O.R., Blangiardo, M., López-Quílez, A. and Martín-Sanz, I., 2018. Species distribution modelling through Bayesian hierarchical approach. Theoretical Ecology, pp.1-11.
12. Deleo, J.M., 1993. Receiver operating characteristic laboratory (ROCLAB): software for developing decision strategies that account for uncertainty. In:

Proceedings of the Second International Symposium on Uncertainty Modelling and Analysis, pp. 318–25. College Park, MD: IEEE Computer Society Press.

13. Donato G., Belongie S., 2002. Approximate Thin Plate Spline Mappings, ECCV, vol. 2, Copenhagen, Denmark, Springer Verlag, pp. 531-542.
14. Elith, J. and Burgman, M.A., 2002. Predictions and their validation: rare plants in the Central Highlands, Victoria, Australia. Predicting species occurrences: issues of accuracy and scale, pp.303-314.
15. Elith, J., H. Graham, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., J. Hijmans, R., Huettmann, F., R. Leathwick, J., Lehmann, A., Li, J., G. Lohmann, L., A. Loiselle, B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC. M. Overton, J., Townsend Peterson, A., J. Phillips, S., Richardson, K., Scachetti-Pereira, R., E. Schapire, R., Soberón, J., Williams, S., S. Wisz, M. and E. Zimmermann, N., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography, 29: 129–151. doi:10.1111/j.2006.0906-7590.04596.x
16. Elith, J. & Leathwick, J.R., 2009. Species distribution models: ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution and Systematics, 40, 677-697.
17. Farr, T.G., Rosen, P.A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., Alsdorf, D., 2007. The Shuttle Radar Topography Mission. Reviews of Geophysics 45, RG2004.
18. Fawcett T., 2005. An introduction to ROC analysis. Pattern Recognition Letters 27 (2006) 861–874
19. Fielding, A. H. and Bell, J. F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv. 24: 38-49.
20. Friedman, J.H., 1991. Multivariate adaptive regression splines. Annals of Statistics 19: 1-141
21. Gallai, N., Salles, J.M., Settele, J. and Vaissière, B.E., 2009. Economic valuation of the vulnerability of world agriculture confronted with pollinator decline. Ecological economics, 68(3), pp.810-821.
22. Gastón, A., Soriano, C., Gómez-Miguel, V., 2009. Lithologic data improve plant species distribution models based on coarse-grained occurrence data. Forest Systems 18, 42-49.

23. Geisser S, Eddy W., 1979. A predictive approach to model selection. Journal of the American Statistical Association 74, 153-160
24. Gelfand AE, Diggle P, Fuentes M, Guttorp P, (eds)., 2010. Handbook of spatial statistics. Chapman & Hall. Boca-Raton
25. Gelman A, Shalizi C., 2013. Philosophy and the practice of Bayesian statistics (with discussion). British Journal of Mathematical and Statistical Psychology 66, 8-80
26. Giraldo, R., 2002. Introducción a la geoestadística: Teoría y aplicación. Bogotá: Universidad Nacional de Colombia.
27. Goulson, D., Nicholls, E., Botías, C. and Rotheray, E.L., 2015. Bee declines driven by combined stress from parasites, pesticides, and lack of flowers. Science, 347(6229), p.1255957.
28. Gray, L.K. and Hamann, A., 2011. Strategies for reforestation under uncertain future climates: guidelines for Alberta, Canada. PLoS One, 6(8), p.e22977.
29. Gray, L.K. and Hamann, A., 2013. Tracking suitable habitat for tree populations under climate change in western North America. Climatic Change, 117(1-2), pp.289-303.
30. Guisan A, Thuiller W., 2005. Predicting species distribution: offering more than simple habitat models. Ecol Lett 8, 993-1009
31. Guisan A, Edwards TC, Hastie T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecol model 157, 89-100
32. Hamann, A. and Aitken, S.N., 2013. Conservation planning under climate change: accounting for adaptive potential and migration capacity in species distribution models. Diversity and Distributions, 19(3), pp.268-280.
33. Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29–36.
34. Herrera, S., Gutiérrez, J.M., Ancell, R., Pons, M.R., Frías, M.D. and Fernández, J., 2012. Development and analysis of a 50-year high-resolution daily gridded precipitation dataset over Spain (Spain02). International Journal of Climatology, 32(1), pp.74-85.
35. Herrera, S., Fernández, J. and Gutiérrez, J.M., 2016. Update of the Spain02 gridded observational dataset for EURO-CORDEX evaluation: assessing the effect of the interpolation methodology. International Journal of Climatology, 36(2), pp.900-908.

36. Hijman R, Graham C., 2006. The ability of climate envelope models to predict the effect of climate change on species distributions. Global Change Biol 12(12), 2272-2281

37. Hijmans, R. J., Elith J., 2015. Species distribution modelling with R. – http://cran.r-project.org/ web/packages/dismo/vignettes/sdm.pdf, The R foundation for statistical computing.

38. Illian, JB, Martino, S, Sørbye, SH, Gallego-Fernández, JB, Zunzunegui, M, Esquivias, MP, Travis, JMJ., 2013. Fitting complex ecological point process models with integrated nested Laplace approximation. Methods Ecol Evol, 4: 305–315. doi:10.1111/2041-210x.12017

39. Klein, A.M., Steffan–Dewenter, I. and Tscharntke, T., 2003. Fruit set of highland coffee increases with the diversity of pollinating bees. Proceedings of the Royal Society of London B: Biological Sciences, 270(1518), pp.955-961.

40. Leathwick JR, Rowe D, Richardson J, Elith J, Hastie T., 2005. Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. Freshwater Biol 50, 2034-2052

41. Li L, Qiu S, Zhang B, Feng CX., 2015. Approximating cross-validatory predictive evaluation in Bayesian latent variable model with integrated IS and WAIC. Stat Comput DOI 10.1007/s11222-015-9577-2

42. Lindgren F, Rue H, Lindström J, 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach [with discussion]. J R Stat Soc B 73(4), 423-98

43. Liu H., 2008. Generalized Additive Model. Department of Mathematics and Statistics University of Minnesota Duluth, Duluth.

44. Loarie SR, Carter BE, Hayhoe K, McMahon S, Moe R, Knight CA, Ackerly DD., 2008. Climate change and the future of Californias endemic flora. PLoS ONE 3(6):e2502

45. Lowell, K. E., and Astroh, J. H., 1989, Vegetative succession and controlled fire in a glades ecosystem. International Journal of Geographic Information Systems, 3, 69-81.

46. Martínez-Minaya, J., Cameletti, M., Conesa, D. and Pennino, M.G., 2018. Species distribution modeling: a statistical review with focus in spatio-temporal issues. Stochastic Environmental Research and Risk Assessment, pp.1-18.

47. Marzban, C., 2004. The ROC curve and the area under it as performance measures. Weather and Forecasting, 19, 1106–1114

48. Metz, C.E., 1978. Basic principles of ROC analysis. Seminars in Nuclear Medicine 8,283-298
49. Midgley GF, Thuiller W., 2007. Potential vulnerability of Namaqualand plant diversity to anthropogenic climate change. Journal of Arid Environments 70, 615-628
50. Muñoz F, Pennino MG, Conesa D, López-Quílez A, Bellido JM., 2013. Estimation and prediction of the spatial occurrence of fish species using Bayesian latent Gaussian models Stoch Environ Res Risk Assess 27, 1171-1180
51. Murphy, A. H., and Winkler, R. L., 1992: Diagnostic verification of probability forecasts. Int. J. Forecasting, 7, 435-455.
52. Panagos, P., Van Liedekerke, M., Jones, A. and Montanarella, L., 2012. European Soil Data Centre: Response to European policy support and public data requirements. Land Use Policy, 29(2), pp.329-338.
53. Pawlina, M. W., Buckley, D. J., and Strickland, R., 1990, Automation of visible area mapping for fire detection lookouts. In Proceedings of the GIS'90 Symposium, Vancouver, pp. 29-46.
54. Pearce, J. and Ferrier, S., 2000. Evaluating the predictive performance of habitat models developed using logistic regression. Ecological Modelling, 133, 225–245.
55. Phillips, S.J., Anderson, R.P. & Schapire, R.P., 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190(3/4): 231-259.
56. Plana, E; Font, M; Serra, M., Borràs, M., Vilalta, O. 2016. Fire and forest fires in the Mediterranean; a relationship story between forest and society. Five myths and realities to learn more. eFIREcom project. CTFC editions. 36pp
57. Pressey RL, Cabeza M, Watts EM, Cowling RM, Wilson KA., 2007. Conservation planning in a changing world. Trends in Ecology and Evolution 22, 583-592
58. Rivera, Ó.R.D.; López-Quílez, A.,2017. Development and Comparison of Species Distribution Models for Forest Inventories. ISPRS Int. J. Geo-Inf. 2017, 6 (6), 176.
59. Rodríguez de Rivera, Ó., López-Quílez, A. and Blangiardo, M., 2018. Assessing the Spatial and Spatio-Temporal Distribution of Forest Species via Bayesian Hierarchical Modeling. Forests, 9(9), p.573.

60. Rue, H. and Martino, S., 2007. Approximate Bayesian inference for hierarchical Gaussian Markov random field models. Journal of statistical planning and inference, 137(10), pp.3177-3192.
61. Rue H., Martino S., Chopin N., 2009. Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations (with discussion). Journal of the Royal Statistical Society, Series B, 71, 319-392
62. Sánchez Palomares, O., Sánchez Serrano, F., Carretero, M.P., 1999. Modelos y cartografía de estimaciones climáticas termopluviométricas para la España peninsular. Instituto Nacional de Investigaciones Agrarias, Madrid, Spain, 192 pp.
63. Salazar, L. A., and Power, J. D., 1988, Three-dimensional representations for fire management planning: a demonstration. In Proceedings of GIS'88, San Antonio, TX. Vol. 2, pp. 948-960.
64. Schelhaas, M.J., Nabuurs, G.J., Hengeveld, G., Reyer, C., Hanewinkel, M., Zimmermann, N.E. and Cullmann, D., 2015. Alternative forest management strategies to account for climate change-induced productivity and species suitability changes in Europe. Regional Environmental Change, 15(8), pp.1581-1594.
65. Schiattino I, Silva C, 2008. Árboles de Clasificación y Regresión: Modelos Cart. Cienc Trab. Oct-Dic; 10 (30): 161-166.
66. Simpson D, Lindgren F, Rue H., 2011. Fast approximate inference with INLA: the past, the present and the future. Technical report at arxiv.org
67. Sing, T., Sander, O., Beerenwinkel, B., Lenaguer, T., 2012. ROCR: Visualizing the performance of scoring classifiers. R package version 1.0-4. http://CRAN.R-project.org/package=ROCR
68. Spano, D., Camia, A., Bacciu, V., Masala, F., Duguy, B., Trigo, R., Sousa, P., Venäläinen, A., Mouillot, F., Curt, T. and Moreno, J.M., 2014. Recent trends in forest fires in Mediterranean areas and associated changes in fire regimes.
69. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A., 2002. Bayesian measures of model complexity and fit (with discussion). Journal of the Royal Statistical Society, Series B, 64 (4), 583-616
70. van der Linde, A., 2005. DIC in variable selection. Statistica Neerlandica 1, 45-56
71. Van Liedekerke, M., Jones, A., Panagos, P., 2006. ESDBv2 Raster Library, a set of rasters derived from the European Soil Database distribution v2.0.

European Commission and the European Soil Bureau Network, CDROM, EUR 19945 EN.

72. Vasconcelos, M., and Guertin, D. P., 1992, FIREMAP. Simulation of fire growth with a Geographic Information System. International Journal of Wildland Fire, 2, 87-96
73. Vehtari A, Lampinen J., 2002. Bayesian model assessment and comparison using cross validation predictive densities. Neural Computation 14, 2439-2468.
74. Wade, P R., 2000. Bayesian methods in conservation biology. Conserv. Biol. 14(5): 1308–1316. doi:10.1046/j.1523-1739.2000.99415.x.
75. Watanabe S., 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. Journal of Machine Learning Research 11, 3571-3594
76. Wintle B A, McCarthy M A, Volinsky C T , Kavanagh R P., 2003. The use of Bayesian model averaging to better represent uncertainty in ecological mod- els. Conserv. Biol. 17(6): 1579–1590. doi:10.1111/j.1523-1739.2003.00614.x.
77. Wisz MS, Hijmans RJ, Li J, Peterson AT, Graham CH, Guisan A., 2008. Effects of sample size on the performance of species distribution models. Divers Distrib 14, 763-773
78. Wood, S.N., 2000. Modelling and smoothing parameter estimation with multiple quadratic penalties. Journal of the Royal Statistical Society (B) 62(2):413-428.
79. Wood, S.N., 2003. Thin-plate regression splines. Journal of the Royal Statistical Society (B) 65(1):95-114.
80. Wood, S.N., 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models.Journal of the American Statistical Association. 99:673-686.
81. Wood, S.N., 2006. Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC, Boca Raton.
82. Wood, S.N., 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society (B) 73(1):3-36

Annexes

Annex I

Development and Comparison of Species Distribution Models for Forest Inventories

Óscar Rodríguez de Rivera * and Antonio López-Quílez

Department of Statistics and Operational Research, Faculty of Mathematics, University of Valencia, 46100 Burjassot (València), Spain; antonio.lopez@uv.es

* Correspondence: osroderi@alumni.uv.es; Tel.: +44-(0)7858-714047

Academic Editors: Duccio Rocchini and Wolfgang Kainz

Received: 19 January 2017; Accepted: 14 June 2017; Published: 16 June 2017

Abstract: A comparison of several statistical techniques common in species distribution modeling was developed during this study to evaluate and obtain the statistical model most accurate to predict the distribution of different forest tree species (in our case presence/absence data) according environmental variables. During the process we have developed maximum entropy (MaxEnt), classification and regression trees (CART), multivariate adaptive regression splines (MARS), showing the statistical basis of each model and, at the same time, we have developed a specific additive model to compare and validate their capability. To compare different results, the area under the receiver operating characteristic (ROC) function (AUC) was used. Every AUC value obtained with those models is significant and all of the models could be useful to represent the distribution of each species. Moreover, the additive model with thin plate splines gave the best results. The worst capability was obtained with MARS. This model's performance was below average for several species. The additive model developed obtained better results because it allowed for changes and calibrations. In this case we were aware of all of the processes that occurred during the modeling. By contrast, models obtained using specific software, in general, perform like “hermetic machines”, because it could sometimes be impossible to understand the stages that led to the final results.

Keywords: additive model; area under the curve; AUC; forest inventory; receiver operating characteristic; ROC; species distribution model

1. Introduction

Species distribution models (SDMs) are mathematical tools based on combination of observations of species occurrence or abundance with environmental variables. These tools are used to analyze species distributions across landscapes [1].

In SDM we usually follow the following processes: (1) compile the locations of the presence of the species; (2) from databases we obtain different values of environmental variables (precipitation, temperature, etc.) for the compiled locations; (3) these environmental variables fit the models to estimate the relationship between sites of occurrence or species richness; and (4) the models are tools to predict the variable of interest across the space or time of interest [2].

Species distribution models comprise three main components: an ecological model, data, and a statistical model [3]. The most pertinent point in statistical modeling is the selection of the mathematical model, because a wrong selection may reduce the predictive power. Ecological modeling experts have shown a keen interest in the effects of mathematical methods on the predictive capacity of distribution models (e.g., [4,5]). A group from California University’s National Center for Ecological Analysis and Synthesis (NCEAS) carried out the most comprehensive study of modeling techniques to date [6]. Their research evaluated the predictive ability of sixteen methods with presence-only or presence-pseudo-absence data on six regions with more than 200 species. Results showed that new methods, such as maximum entropy (MaxEnt), have greater predictive power than other methods, such as logistic regression (both adjusted generalized linear models, GLM, and adjusted generalized additive models, GAM). Subsequent studies have also obtained better predictive capacity for MaxEnt than for logistic regression [7–11].

Decision support tools for plant species selection for ecological/environmental management have been based on species distribution models (also called ecological niche models) that analyze the probability of the presence of the species as a function of environmental variables (e.g., precipitation, temperature, or soil properties). The idea of developing statistical models, with several variables, to predict the potential distribution of species could be a complex task requiring in-depth study of several statistical methods that provide fairly inconsequential results. Several tools are now available to facilitate this task. Several studies have compared the performance of different statistical approaches to predict species distributions, obtaining a variety of suggestions about model selection [5,6,12].

Some of the latest species distribution models only use the presence of the species in the modeling process. Other methods use presence/absence data or 'background' data. Logistic regression is the traditional approach to analyzing presence/absence data [2]. Our study uses a large dataset with presence-absence and, therefore, requires a method that can use these data; in other words, a method with presence-absence data. We understand that MaxEnt is not a presence-absence method; in fact, it uses the presence-only data and a user-defined number of randomly-selected points, combining these with the covariates to build an index of habitat suitability for each cell ranging from 0 (least suitable habitat) to 1 (most suitable habitat). Moreover, MaxEnt was included in this analysis because of is one of the most commonly used methods as a species distribution model, as we can see summarized in [13].

The aim of this paper is two-fold. The first aim is to present and explain, from a mathematical point of view, different common tools designed for species distribution modeling. Our main target is to show how these tools develop models based on regressions, and explore the advantages and disadvantages of each model. The second aim is to compare these models and decide which is the most accurate according to easily understood indicators.

This study uses real data on seventeen forest species obtained from the Spanish National Inventory, supported with environmental variables. These species, with presence-absence, were first located with geographical coordinates. We then generated distribution models with tools designed to create this kind of model. Finally, we developed an additive model with R and compared the results from it to evaluate the predictive capability of all of the models in an attempt to answer the following questions: Does anyone statistical technique have a regularly greater predictive ability than the others for all types of relationships between environmental variables and the presence of the species? [14] Are species with a higher presence easier to predict than others that are less represented?

The paper proceeds as follows: In Section2(Material and Methods), we review the principal properties of different models and the way each model is evaluated in order to compare the prediction capability. In Section3(Results) we summarize the results of each distribution model and evaluate them to conclude which one we consider to be the most accurate and which has the best prediction capability. Finally in Section4(Discussion), we analyze the potential of that model.

2. Materials and Methods

2.1. Species Occurrence Data

We have used the Spanish National Forest Inventory (NFI) dataset to elaborate our research project. NFI comprises a systematic grid with 91,889 plots, each of which is 0.2 ha in size. From this dataset, we started by choosing 17 forest species with presence/absence in each plot. The species analyzed with the percentage of presence are as follows: Abies alba Miller (<5%), Castanea sativa Miller (5%), Fagus sylvatica L. (5.5%), Pinus halepensis Miller (15%), Pinus nigra Arnold (9.5%), Pinus pinea L. (15%), Pinus pinaster Aiton (15%), Pinus sylvestris L. (12.3%), Pinus uncinata Turra (<5%), Quercus canariensis Willd. (<5%), Quercus faginea Lam. (11%), Quercus humilis Miller (<5%), Quercus ilex L. (36%), Quercus petraea (Matt.) Liebl (<5%), Quercus pyrenaica Willd. (8.2%), Quercus robur L. (8.6%), and Quercus suber L. (5%). In this paper we only report the most characteristic species to evaluate the statistical processes used in the study.

2.2. Environmental Predictors

We have obtained the climatic data grids by applying the models for climatic estimation produced by [15] to the Shuttle Radar Topography Mission (STRM) 3-arc-second (≈90 m) elevation dataset [16]. These models interpolate monthly climate data from weather stations using latitude, longitude, and elevation as independent variables. We have analyzed 10 climatic predictors commonly used in tree species autoecology in Spain [17]: mean summer rainfall (SR), mean annual rainfall (R), mean summer temperature (ST), mean annual temperature (T), mean of maximum temperatures of the warmest month (MTWM), mean of minimum temperatures of the coldest month (MTCM), mean annual potential evapotranspiration (ETP), mean annual water surplus (SUP), and mean annual water deficit (DEF). Moreover, we have used the European Soil Database [18] to allocate each plot to a parent material class (calcareous or siliceous) (C). The distribution of calcareous parent materials is a very useful predictor of plant species distribution in Mediterranean ecosystems [19].

Model selection was based on the ease of working with each model and the possibility of repeating each process with the same characteristics using the species studied. Therefore, we constructed the models with one of the most widely-used models (MaxEnt), and others based on simple software developed by Saldorf System (San Diego, CA, USA) (CART and MARS). Finally, we built an additive model using R software.

The statistical methods used in this study are summarized below.

2.3. MaxEnt (Maximum Entropy)

MaxEnt [20,21] is an artificial intelligence method based on the statistical principle of maximum entropy. Models are limited by the value of the variables used to develop the problem. For example, the expected value (mean value predicted by the model) of each independent variable must match its empirical average (the mean value observed when sampling with an independent variable occurrence data item). MaxEnt obtains the maximum entropy probability of the distribution; in other words, the distribution nearest to the uniform distribution, with all of the conditions. Additionally, MaxEnt is based on the following points: (a) the presence of a species is represented by a likelihood function P on a set x of points in the study zone. P gives a positive value x everywhere so that the sum of P(x) is unity; (b) building a model of P with a group of constraints obtained from the empirical data of presence; (c) the restrictions are expressed as a simple function of known environmental variables, f(v); (d) in the MaxEnt method, the average forces of each function of each variable are close to the actual average of the variable zones of presence; and e) of the possible options available, a specific combination of features is selected to minimize the entropy function (measured by the Shannon index). The entropy function allows optimal selection of variables and functions based on their significance, and eliminates restrictions that do not provide the model with significance.

The general form of the probability function is, with i environmental variables:

$$P(x)=e^{\lambda \cdot f(x)}/Z_{\lambda},$$

where lambda is a weighting coefficient vector and f is the vector corresponding to the functions. Z is a normalization constant used to ensure that P(x) is the unit. The values P(x) obtained should be interpreted as relative suitability values. These values are normally processed by a logistic function that is adjusted to a more comprehensible level in the range between 0 (incompatible) and 1 (ideal).

Hypothetically, MaxEnt is most similar to generalized linear models and additive models. In what follows, we use the expressions of [22]. A commonly-used linear model is the Gaussian logit model, in which the logit of the predicted probability of occurrence is:

$$\alpha + \beta_1 f_1 + \gamma_1 f_1(x)^2 + \cdots + \beta_n f_n + \gamma_n f_n(x)^2$$

where the fj are environmental predictors; α, βj, and γj are fixed coefficients; and the logit function is defined by logit(p) = ln(p/(1 − p)). The above expression is no different in form as the log (rather than logit) of the likelihood of the pixel x in a MaxEnt expression with linear and quadratic structures. A common method for modeling interactions between variables in a linear model is to create product predictors, which is equivalent to the use of it in MaxEnt [21].

From a similar point of view, if the probability of presence/absence is modeled with an additive model using a logit link function, the logit of the predicted likelihood has the form:

$$g_1(f_1(x)) + \cdot \cdot \cdot + g_n(f_n(x))$$

where fi are the environmental predictors. gi are smooth functions fitted by the expression, with the quantity of smoothing measured by a measurement factor. This is a similar method as the log probability in a MaxEnt model, for pixel x, with threshold structures, and regularization has an equivalent effect to smoothing on the otherwise random functions gi. In both cases, the form of the response curve to each environmental predictor is determined by the data.

During the process, MaxEnt generates different probability distributions, opening from a uniform scattering, and improves the fitting to the data. This improvement is defined as the average possibility of occurrence data, removing a constant, which means that the uniform distribution has a gain of zero. Regardless of these similarities, several differences exist between generalized models and MaxEnt, leading them to create different results. When GLM/GAMs are developed to model the probability of presence, absences are needed. When applied to presence-only data, background pixels should be used as an alternative of true nonappearances [12,23]. However, the interpretation of the output is less clear-cut—it must be taken as a relative guide of ambient suitability. Dissimilarly, MaxEnt models a probability of presence over the pixels in the area of study, and on no account are pixels without records interpreted as no presences. Additionally, MaxEnt is a generative method, although GLM/GAMs are discriminative, and generative methods may give better likelihoods when the quantity of training data is insignificant [24].

For all species we use the model with the same variables, obtaining the following results shown below for each of the species under study.

2.4. MARS (Multivariate Adaptive Regression Splines)

MARS is a statistical method developed by Friedman [25]. It involves designing flexible models in which the data are adjusted to partial regressions. When models are nonlinear, they are approximated by partial linear regression, where the grade of the equation changes from one step to another, establishing a node between the end of one linear regression and the beginning of the next.

A node indicates the end of one partial regression and the beginning of another. Between two consecutive nodes, logically, the model is defined by a linear regression. The nodes are selected with the aid of a search procedure that generates a stepper algorithm. The model generated is overfitted, so the less relevant nodes are subsequently removed using a statistical approach known as generalized cross-validation. Finally, we only consider the most significant nodes. The function is a parameter interceptor β0, and βi is the weighted sum of one or more basic functions FBi. Therefore, the model will consist of a weighted sum of selected basic expressions from a large number of basic expressions that link all of the values of the predictor. The model is generated as follows:

$$f(x)=\beta_0+\sum\beta_i\, f_i\, FB_i$$

$$FB_i=\max(0,V-N)$$

$$FB_{i+1}=\max(0,N-V)$$

where FB is a basic function and acts as a new variable, V is the variable and N is the node.

Going deeper into the MARS algorithm, note that the models are constructed from double-sided truncated functions of the form (see Figure1):

$$(x-t)+=(x-t;\ x>t/0;\ \text{other}$$

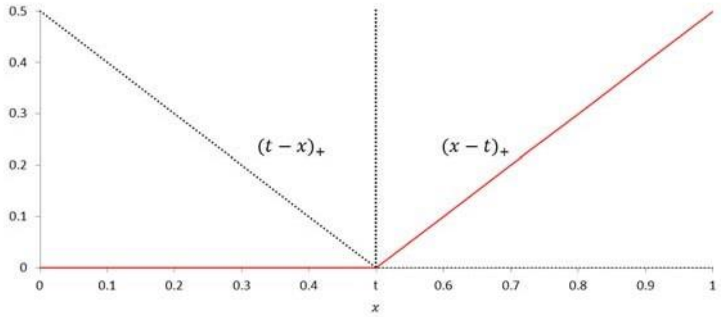


Figure 1. Representation of two basic functions, where parameter t is the knot. The (+) signs denotes that only positive values are considered.

Each expression is in linear pieces, with a lump in the value t, where each node is located at the end of one region of the data and starts at a different one [26]. The Salford Predictive Modeler Builder v6.6 (www.salford-systems.com) was used to generate these models.

2.5. *CART (Classification and Regression Trees)*

This method was established by Brieman et al. [27] and generates binary trees (the parent nodes are divided into two child nodes) by iterative partitions, in a process that can be repeated to attempt to turn each child node into a parent node. The algorithm searches for the optimal cutoff values among all of the independent variables to obtain an optimal set of binary divisions, so as to minimize the variance within each node and maximize it between different nodes; it is, therefore, possible that some variables will be unused. Once the tree that best classifies the cases has been identified, with no limits on complexity, the algorithm 'prunes', or simplifies, to avoid overfitting of the data. The result is a tree that establishes yes/no questions. Depending on the kind of dependent variable there can be two types of trees: regression (continuous dependent variable) and classification (discrete variable).

The aim of this method is to discriminate, estimate, or predict Y-based predictors X1, ..., Xp by successive partitions or by sets of individuals, maximizing a measure of information content with respect to the response variable. In the validation phase this same design, a training matrix, or a similar, but independent, matrix (validation or test sample) can be used; in this case we use the same matrix. The most important advantages of classification and/or regression trees are [28]: (a) structured knowledge is obtained in the form of classification rules or the values of a variable interval. This knowledge is easy to interpret and, in simple language, characterizes the classes or values of a variable interval; (b) as it is a nonparametric analysis (distribution-free procedure), it requires no distributional assumptions to validate probability; (c) it allows working with all types of predictor variables: binary, nominal, ordinal, and interval or ratio; (d) it allows unknown values for the predictor variables in the individuals, both in the construction phase and in the tree prediction; e) in the case of classification probability, it can be set to a priori classes; and (f) the observations can be weighed using an ad-hoc variable.

An expression known as recursive dividing is essential to the nonparametric statistical approach of classification and regression trees (CART) [27]. Supposing the data are given by D = {(Xi, Yi), i = 1, 2, ..., n}, where Yi are widths made on a uninterrupted response variable Y, and the Xi are measurements on an input r-vector X. We accept that Y is connected to X as in multiple regression, and the aim is to use a tree-based algorithm to predict Y from X.

Regression trees are built in a parallel way to classification trees, and the technique is generally stated as recursive-separating regression. In a classification tree, the class of a terminal knot is demarcated as the class that orders a plurality (generally in the two-class case) of all of the observations in that node, where ties are randomized. In a regression tree, the output is set to have the constant value Y(τ) at terminal node τ . Hence, the tree can be characterized as an r-dimensional histogram approximated of the regression surface, where r is the number of input variables, X1, X2, ..., Xr [29].

$$i(\tau)=\sum(Y_i -\bar{Y}_\tau)^2$$

where $\tilde{Y}_\tau$ is the average of the Y_i for all annotations assigned to node τ.

To determine the type of split in any node we take as our splitting strategy at node $\tau \in \check{T}$ the division that delivers the largest decrease in the value of i(τ).The reduction in i(τ) due to a division into τ_L and τ_R is expressed by

$$\Delta i(\tau)=i(\tau) - i(\tau_L) - i(\tau_R)$$

The left daughter node and right daughter node emanating from a (parent) node τ are denoted by τL and τR, respectively.

The best division at τ is the one that exploits Δi(τ). The consequence of employing such a splitting approach is that the best division will split up observations according to whether Y has a small or large value; in general, where divisions occur, we can see either y (τL) < y(τ) < y(τR) or its opposite with y(τL) and y(τR) interchanged.

We note that discovery τL and τR to exploit Δi(τ) is similar to reducing i(τL)+i(τR). Solving:

$$\min \tau_L,\tau_R\{p(\tau_L) s^2 (\tau_L)+p(\tau_R) s^2 (\tau_R)\}$$

where p(τL) and p(τR) are the proportions of observations in τ that divide to τL and τR, individually [28].

The Salford Predictive Modeler Builder v6.6 (www.salford-systems.com) was used to generate these models.

2.6. Generalized additive model with thin plate splines (GAM.TP)

A generalized additive model (GAM) is a generalized linear model in which the linear predictor be determined by linearly on unidentified smooth expressions of some variables, and interest focuses on inference about these smooth expressions. Additive models were originally built by [28] to combine properties of linear models with additive models.

The generalized additive model replaces

$$\sum\beta_j X_j$$

with

$$\sum f_j (x_j)$$

where fj is an unspecified (‘non-parametric’) function. It can be in a non-linear form:

$$E(Y \mid X_1 \ldots X_p) = f(X_1 \ldots X_p) = f_0 + f_1 (X_1) + \ldots + f_p (X_p)$$

The function fj(xj) is estimated in a flexible manner using a spline smoother [30].

A smoother is an instrument for summarizing the tendency of a dependent variable Y as an expression of one, or more, independent variables X1, . . . , Xp. It generates an estimate of the tendency that is less mutable than Y itself; therefore, the name ‘smoother’. The most significant characteristic of a smoother is its non-parametric nature, so the smooth function is also known as a non-parametric function. Its greatest difference from the GLM is that it does not undertake an inflexible form for the dependence of Y on X1, . . . , Xp. It allows an approach with the addition of expressions (expressions that have separated input estimates), not just with one indefinite expression only. For this reason it is the building block of the generalized additive model algorithm [31].

Testing the different types of splines reveals that the best model helped with the AIC value is the additive model with thin plate regression splines. The thin plate spline is the two-dimensional equivalent of the cubic spline in one dimension. It is the essential resolution to the biharmonic equation, and has the form:

$$U(r) = r^2 \ln(r)$$

Assuming a dataset of points, a weighted mixture of thin plate splines concentrated about each point gives the interpolation expression that passes through the points precisely while reducing the so-called ‘bending energy.’ Bending energy is defined here as the integral over R2 of the squares of the second derivatives:

$$I[f(x,y)] = \iint ((f_{xx})^2 + 2(f_{xy})^2 + (f_{yy})^2) \, dxdy$$

Regularization should be used to decrease the necessity that the interpolant pass through the data points exactly.

The designation of ‘thin plate spline’ is a physical analogy referring to the flexibility of a thin sheet of metal. In the physical situation, the deflection is in the z direction, at a right angle to the plane. In order to apply this impression to the problem of coordinate conversion, the lifting of the plate is interpreted as a dislocation of the x or y coordinates within the plane [32].

These splines are short rank isotropic smoothers of any number of variables. The splines are isotropic because any variation of the covariate co-ordinate system will not modify the output of smoothing. The low rank means that they have rarer coefficients than there are data to smooth. They are the default smoother for ‘s’ terms due to there being a clear logic in which they are the ideal smoother of any given basis measurement/rank [33].

In this case, as we are building the model with R we used the mgcv package [33–37] to construct the additive model and the ROCR package [38] to obtain the validations, AUC values, and graphics.

2.7. Evaluation

The area under the receiver operating characteristic (ROC) function (AUC) is taken to be an important index because it provides a single measure of overall accuracy that is independent upon a particular threshold [39]. If the objective is to rank the classifiers, comparisons using ROC plots are more robust since they are not dependent of the values in a confusion matrix [40]. An ROC graph is a method for visualizing, establishing, and selecting classifiers based on their presentation. ROC curve analysis was developed during World War II as a tool in signal processing, and is now used in many branches of science. Standard references for ROC curve analysis are [40–45].

Although ROC graphs are conceptually simple, their application in research contexts gives rise to some complexities that are not obvious and their practical use entails some common misconceptions and pitfalls [46].

Some formulae typical in ROC curves are

$$\text{tp rate} \sim (\text{Positives correctly classified})/(\text{Total positives})$$

$$\text{fp rate} \sim (\text{Negatives incorrectly classified})/(\text{Total negatives})$$

ROC graphs are two-dimensional graphs where the true positive rate is presented on the Y axis and the false positive rate is presented on the X axis. An ROC graph represents relative adjustments between profits (true positives) and expenses (false positives). Figure2shows the area under two ROC curves, A and B. Classifier A has a greater area and, therefore, better average performance.

Finally, it is probable for a low-AUC classifier to perform better in a specific region of the ROC space than a high-AUC classifier. Figure2shows an example of this: classifier B is generally worse than A, except at an fp rate > 0.6 where B has an insignificant advantage. However, in practice the AUC performs very well and is often used when a general measure of predictiveness is desired.

In order to analyze suitability of the different models, we have used 10% of the data available for each species.

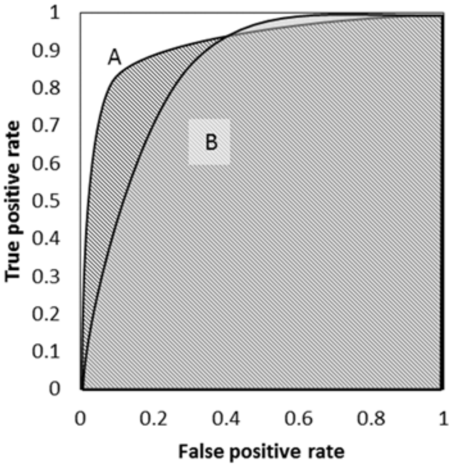


Figure 2. Example of two ROC curves and area under the curves (AUC).

3. Results

The study performed with the 17 species reveals some important results. For the sake of simplicity, we present here only the results summarizing the species using the two species with the highest and lowest results.

A two-way ANOVA shows that all the environmental variables included in the models are significant for all measures of performance ($p < 0.05$). All models designed have good predictions and obtain high AUC values.

Analyzing predictability, based on the AUC, we have obtained that for all the species analyzed, MARS and MaxEnt are the models with the lowest predictability and, consequently, with the lowest AUC average. However, CART and GAM generally have the highest AUC values.

Below are the AUC values obtained in the verification process for the species analyzed. The following ecological modeling methods are compared: MARS and MaxEnt, CART, and GAM.TP. The scatterplot graph shows the different models' behavior, demonstrating that all have good predictability based on their AUC values.

Figure3shows that MaxEnt, CART, and GAM.TP have AUC values near to 1, and that all of the models have good results for predicting species distribution. For the different species, all of the statistical models show similar behavior and performed in the same way. Comparing model predictability, the AUC values in GAM.TP have better results, on average, than the others. A comparison of one particular species with the highest AUC results and another with the lowest reveals no important differences between the models. As we can see in the table below, AUC values are very similar across all models.

In Table1, we can see that the average for the best species modeled was 0.986 with a deviation value of 0.029. Moreover, one of the species with the lowest AUC value was Pinus pinea with an average of 0.876; in this case the deviation was 0.019. In both cases MARS obtained the lowest AUC values. In contrast, the best values were obtained with GAM.TP, although the few differences between this model and the others were not very significant. Finally, we can see Quercus ilex, a species with greater presence in the dataset, and also the species with the lowest AUC for all the models.

Figure4a (Quercus canariensis) and 4b (Pinus pinea) show the results. Models are represented in different colors to facilitate understanding: MaxEnt (red), MARS (green), CART (blue), and GAM.TP (black). Figure4a shows that every model has very good results, but GAM.TP exceeds the others. However, Figure4b shows that MARS obtains worse results than the others and, for this reason, the line is well below the rest.

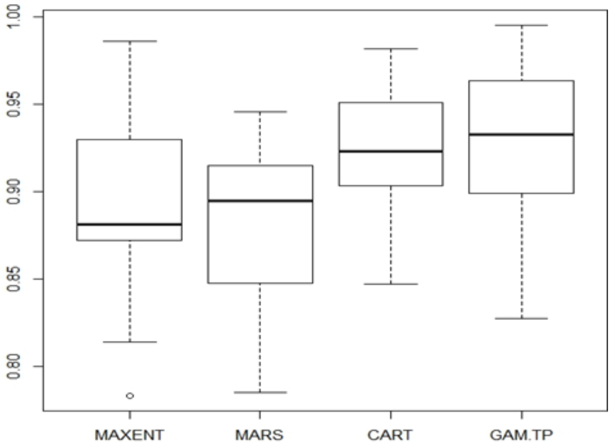


Figure 3. Average values of AUC for MaxEnt, MARS, CART, and GAM.TP models.

Table 1. Average of values of AUC for each model with *Quercus canariensis* and *Pinus pinea*.

Species	MAXENT	MARS	CART	GAM.TP
Q. canariensis	0.986	0.920	0.970	0.995
P. pinea	0.884	0.847	0.874	0.899
Q. ilex	0.783	0.814	0.847	0.834

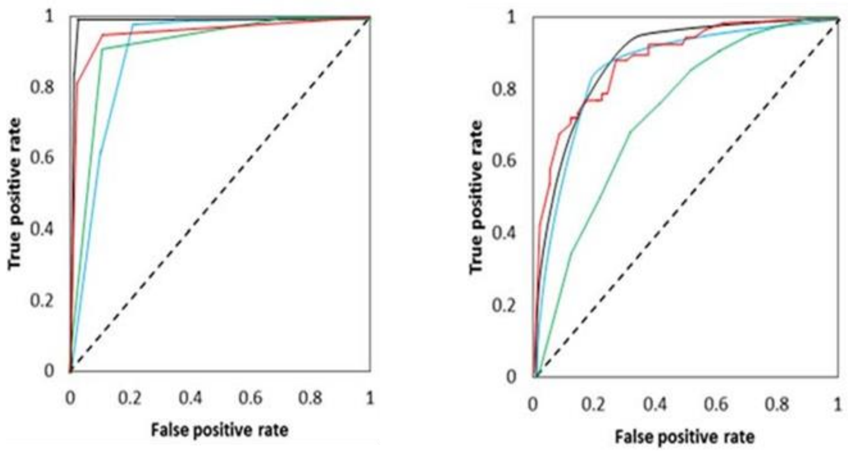
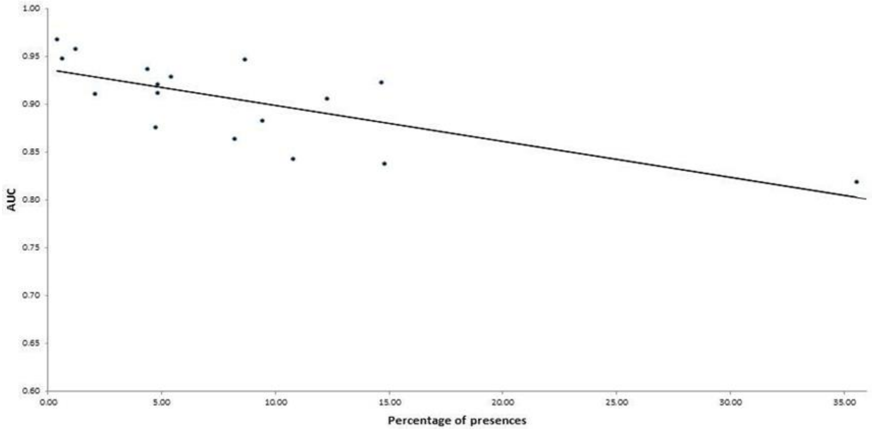


Figure 4. Curve ROC of different models with Quercus canariensis (right, a) and Pinus pinea (left, b). Models are represented in different colors to facilitate understanding: MaxEnt (red), MARS (green), CART (blue), and GAM.TP (black).

Analyzing the results, species with the highest number of presences have lower values in the predictions due to a wide range of the environmental variables. In contrast, the species with the most absences have the highest AUC values, perhaps due to the representative environmental characteristics that give rise to the presence of these species.

If we analyze presence-absence from the dataset and compare it with the AUC average, we find that the relationship between AUC and percentage of presence is negative (based on the correlation index), with a value of −0.75. Species with the highest percentage of presence have lower AUC values than other less-represented species in the area of study.

Figure5represents the relationship between the presence percentage and the AUC value, showing it to be negative; when the percentage of presence increases, the AUC value decreases. A simple trend line clearly shows the behavior between these values.



The trend line is represented with the expression y = −0.0038x + 0.9365, and an R2 value of 0.508.

Figure 5. Dispersion graph with the percentage of the number of presences (X axis) and AUC values (Y axis) with regression.

In summary, every AUC value obtained with those models is significant and all the models could be useful to represent the distribution of each species. Overall, the additive model with thin plate splines gave the best results. MaxEnt, CART, and GAM.TP with thin plates splines obtained similar AUC values. The worst capability was obtained with MARS. This model's performance was below the average for several species. The models we developed obtained better results because they allowed for changes and calibrations. In this case we were aware of all of the processes that occurred during the modeling. By contrast, models obtained using specific software, in general, perform like "hermetic machines" because it could sometimes be impossible to understand the stages leading toward the final results.

4. Discussion

Our modeling framework examines how applicable some of the most widely used models are to species, the presences of which are largely set by the physical environment.

As we can see, all of the techniques developed here proved capable of successfully predicting species distribution. Factually, all of the models obtain similar performance based in the AUC and, over all, all of the methods show good results for predicting species distribution. Thus, there are no important differences between the different techniques developed in this analysis. Moreover, we can highlight interesting points based in our results to try to clarify and support our model selection:

GAM.TP performed better overall than MaxEnt and MARS, even though these differences are not substantial when compared with regression trees. This result differed from other comparative analyses [5,14,47,48], where linear models and additive models performed better than classification trees. We establish that, despite the dissimilarities in model suppositions, all statistical techniques seem to provide the best predictions for additive models.

As we said in the introduction, MaxEnt was included in this analysis due to the popularity of the method. We can understand that, in some points, it is not comparable to presence/absence models. Moreover, the similarity in results gave us some chances to compare the models. As we said, MaxEnt is presence-only data; for this reason the interpretation of the output is less clear than the models for presence/absence data. MaxEnt output must be taken as a relative guide of environmental suitability. On the other hand, presence/absence models could be more reliable because these models use information from the real absences.

From the user perspective, if we compare the different modeling techniques, CART and MARS required the least amount of user guidance (probably because they were developed in tools designed in a friendly environment). Moreover these tools are less flexible than the other statistical techniques developed in this research and, also, there is a high complexity if the user tries to manipulate the default settings in order to improve the accuracy of the outputs.

GAM.TP requires some knowledge of statistical techniques due to the amount of possibilities that it offers for building one's own model. Moreover, this flexibility and the possibilities offered by this approach make this approach more attractive. Currently, several tutorials available for several packages in R make the possibility to elaborate advanced statistical models more affordable without master knowledge (although we recommend a deeper research in statistical techniques before applying any model to avoid misunderstanding and frustration).

In conclusion, additive models with thin plate splines may be considered one of the greatest methods to analyze species distribution models working with presence-absence data, comparable to MaxEnt, CART, and MARS. Our results show a better fit and more flexibility in the design.

Looking at the quality of the data and the possibility to work with presence/absence values, and also with a systematic survey, we can confirm, looking our results, that the information obtained from the absences could be more important than the presences. Analyzing this result from an ecological perspective, absences deliver more information about the species due to the combination of several environmental predictors.

From an ecological perspective, analyzing the variables used in all of the models, we can see some differences between the variables' importance, depending of the model used. Comparing the species used before, we can see in Table2that different models have different variables' weight. In our case, MARS and CART have the same set of the most important variables for both species (SUP, MTCM, C, SR). Moreover, MaxEnt uses different variables for the different species: MTCM, C, R, and ETP with Q. canariensis, and ST, SUP, MTCM, and SR in P. pinea model. Finally, with GAM. TP, the most important variables were MTCM, SR, R, and WD in Q. canariensis, and in P. pinea, all of the variables have similar weight, but the more important four were R, SR, MTWM, and ST.

Table 2. Summary of the most important variables for each model with Quercus canariensis and Pinus pinea.

Species	**MAXENT**	**MARS**	**CART**	**GAM.TP**
Q. canariensis	MTCM	SUP	SUP	MTCM
	C	MTCM	MTCM	SR
	R	C	C	R
	ETP	SR	SR	WD
P. pinea	ST	SUP	SUP	R
	SUP	MTCM	MTCM	SR
	MTCM	C	C	MTWM
	SR	SR	SR	ST

As we have seen in the results (Table1and Figure5), species that are less represented (i.e., with more absences), have better predictability than species with more presences. This situation shows us the importance of absences in predictive models. These absences give us several pieces of information about the suitability of species and defining absence areas. If we analyze these models as management tools, this information is essential regarding the species selection and, in our case, for forest management.

Finally, we understand that there are more advanced approaches that can be applied in species distribution models, most of them through the Bayesian approach (i.e., R-INLA (Integrated nested Laplace approximation) can be compiled with the stochastic partial differential equations (SPDE) approach [49] which, through a discretization of a continuous Gaussian field, can cope efficiently with variables characterized by a complex spatial structure). However, our objective was to show the interesting opportunities that these explanatory techniques offer and to assess the relationships between environmental variables.

Author Contributions: Óscar Rodríguez de Rivera and Antonio López-Quílez conceived and designed the experiments; Óscar Rodríguez de Rivera performed the experiments; Óscar Rodríguez de Rivera analyzed the data; Óscar Rodríguez de Rivera and Antonio López-Quílez wrote the paper.

References

1. Elith, J.; Leathwick, J.R. Species distribution models: Ecological explanation and prediction across space and time. Annu. Rev. Ecol. Evol. Syst. 2009, 40, 677–697.
2. Hijmans, R.J.; Elith, J. Species Distribution Modelling with R. The R Foundation for Statistical Computing, 2015. Available online:http://cran.r-project.org/web/packages/dismo/vignettes/sdm.pdf(accessed on 4 May 2016).
3. Austin, M.P. Spatial prediction of species distribution: An interface between ecological theory and statistical modeling. Ecol. Model. 2002, 157, 101–118.
4. Muñoz, J.; Felicísimo, A.M. Comparison of statistical methods commonly used in predictive modeling. J. Veg. Sci. 2004, 15, 285–292.
5. Segurado, P.; Araújo, M.B. An evaluation of methods for modeling species distributions. J. Biogeogr. 2004, 31, 1555–1568.
6. Elith, J.; Graham, H.C.; Anderson, P.R.; Dudík, M.; Ferrier, S.; Guisan, A.; Hijmans, J.R.; Huettmann, F.; Leathwick, R.J.; Lehmann, A.; et al. Novel methods improve prediction of species' distributions from occurrence data. Ecography 2006, 29, 129–151.
7. Gibson, L.; Barrett, B.; Burbidge, A. Dealing with uncertain absences in habitat modeling: A case study of a rare ground-dwelling parrot. Divers. Distrib. 2007, 13, 704–713.
8. Elith, J.; Graham, C.H. Do they? How do they? Why do they differ? On finding reasons for differing performances of species distribution models. Ecography 2009, 32, 66–77.
9. Roura-Pascual, N.; Brotons, L.; Peterson, A.; Thuiller, W. Consensual predictions of potential distributional areas for invasive species: A case study of Argentine ants in the Iberian Peninsula. Biol. Invasions 2009, 11, 1017–1031.
10. Tognelli, M.F.; Roig-Junent, S.A.; Marvaldi, A.E.; Flores, G.E.; Lobo, J.M. An evaluation of methods for modeling distribution of Patagonian insects. Rev. Chil. Hist. Nat. 2009, 82, 347–360.

11. Marini, M.; Barbet-Massin, M.; Lopes, L.; Jiguet, F. Predicting the occurrence of rare Brazilian birds with species distribution models. J. Ornithol. 2010, 151, 857–866.
12. Ferrier, S.; Watson, G.; Pearce, J.; Drielsma, M. Extended statistical approaches to modeling spatial pattern in biodiversity in northeast New South Wales. 1. Species-level modeling. Biodivers. Conserv. 2002, 11, 2275–2307.
13. Elith, J.; Phillips, S.J.; Hastie, T.; Dudík, M.; Chee, Y.E.; Yates, C.J. A statistical explanation of MaxEnt for ecologists. Divers. Distrib. 2011, 17, 43–57.
14. Meynard, C.N.; Quinn, J.F. Predicting species distributions: A critical comparison of the most common statistical models using artificial species. J. Biogeogr. 2007, 34, 1455–1469.
15. Sánchez Palomares, O.; Sánchez Serrano, F.; Carretero, M.P. Modelos y Cartografía de Estimaciones Climáticas Termopluviométricas Para la España Peninsular; Instituto Nacional de Investigaciones Agrarias: Madrid, Spain, 1999; p. 192.
16. Farr, T.G.; Rosen, P.A.; Caro, E.; Crippen, R.; Duren, R.; Hensley, S.; Kobrick, M.; Paller, M.; Rodriguez, E.; Roth, L.; et al. The Shuttle Radar Topography Mission. Rev. Geophys. 2007, 45, RG2004.
17. Alonso Ponce, R.; López Senespleda, E.; Sánchez Palomares, O. A novel application of the ecological field theory to the definition of physiographic and climatic potential areas of forest species. Eur. J. For. Res. 2010, 129, 119–131.
18. Panagos, P.; Van Liedekerke, M.; Jones, A.; Montanarella, L. European Soil Data Centre: Response to European policy support and public data requirements. Land Use Policy 2012, 29, 329–338.
19. Gastón, A.; Soriano, C.; Gómez-Miguel, V. Lithologic data improve plant species distribution models based on coarse-grained occurrence data. For. Syst. 2009, 18, 42–49.
20. Elith, J.; Burgman, M.A. Predictions and their validation: Rare plants in the Central Highlands, Victoria, Australia. In Predicting Species Occurrences: Issues of Accuracy and Scale; Scott, J.M., Heglund, P.J., Morrison, M.L., Raphael, M.G., Wall, W.A., Samson, F.B., Eds.; Island Press: Covelo, CA, USA, 2002; pp. 303–314.
21. Phillips, S.J.; Dudík, M. Modeling of species distributions with Maxent: New extensions and a comprehensive evaluation. Ecography 2008, 31, 161–175. [CrossRef]

22. Yee, T.W.; Mitchell, N.D. Generalized additive models in plant ecology. J. Veg. Sci. 1991, 2, 587–602.
23. Ferrier, S.; Watson, G. An Evaluation of the Effectiveness of Environmental Surrogates and Modelling Techniques in Predicting the Distribution of Biological Diversity; Environment Australia: Canberra, Australia, 1997.
24. Ng, A.Y.; Jordan, M.I. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. Adv. Neural Inform. Process. Syst. 2001, 14, 605–610.
25. Friedman, J.H. Multivariate adaptive regression splines. Ann. Stat. 1991, 19, 1–141.
26. Nedjah, N.; Luiza de Macedo, M. Fuzzy Systems Engineering: Theory and Practice; Springer: New York, NY, USA, 2005.
27. Breiman, L.; Friedman, F.H.; Olshen, R.A.; Stone, C.J. Classification and Regression Trees; Wadsworth and Brooks: Pacific Grove, CA, USA, 1984.
28. Schiattino, I.; Silva, C. Árboles de Clasificación y Regresión: Modelos Cart. Cienc. Trab. 2008, 10, 161–166.
29. Izenman, A. Modern Multivariate Statistical Techniques; Springer: New York, NY, USA, 2008.
30. Hastie, T.; Tibshirani, R.J. Generalized Additive Models; Chapman & Hall/CRC Press: London, UK, 1990.
31. Liu, H. Generalized Additive Model; Department of Mathematics and Statistics University of Minnesota Duluth: Duluth, MN, USA, 2008.
32. Donato, G.; Belongie, S. Approximate Thin Plate Spline Mappings. In Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Denmark, 28–31 May 2002; Springer: Copenhagen, Denmark, 2002; pp. 531–542.
33. Wood, S.N. Thin-plate regression splines. J. R. Stat. Soc. B 2003, 65, 95–114.
34. Wood, S.N. Modelling and smoothing parameter estimation with multiple quadratic penalties. J. R. Stat. Soc. B 2000, 62, 413–428.
35. Wood, S.N. Stable and efficient multiple smoothing parameter estimation for generalized additive models. J. Am. Stat. Assoc. 2004, 99, 673–686.
36. Wood, S.N. Generalized Additive Models: An Introduction with R; Chapman and Hall/CRC: Boca Raton, FL, USA, 2006.
37. Wood, S.N. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J. R. Stat. Soc. B 2011, 73, 3–36.

38. Sing, T.; Sander, O.; Beerenwinkel, B.; Lenaguer, T. ROCR: Visualizing the Performance of Scoring Classifiers. R Package Version 1.0-4. 2012. Available online:http://CRAN.R-project.org/package=ROCR(accessed on 27 April 2015).
39. Deleo, J.M. Receiver operating characteristic laboratory (ROCLAB): Software for developing decision strategies that account for uncertainty. In Proceedings of the Second International Symposium on Uncertainty Modelling and Analysis, College Park, MD, USA, 17–20 March 1993; IEEE Computer Society Press: Washington, DC, USA, 1995; pp. 318–325.
40. Fielding, A.H.; Bell, J.F. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv. 1997, 24, 38–49.
41. Metz, C.E. Basic principles of ROC analysis. Semin. Nucl. Med. 1978, 8, 283–298.
42. Hanley, J.A.; McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982, 143, 29–36. [CrossRef]
43. Murphy, A.H.; Winkler, R.L. Diagnostic verification of probability forecasts. Int. J. Forecast. 1992, 7, 435–455.
44. Pearce, J.; Ferrier, S. Evaluating the predictive performance of habitat models developed using logistic regression. Ecol. Model. 2000, 133, 225–245.
45. Marzban, C. The ROC curve and the area under it as performance measures. Weather Forecast. 2004, 19, 1106–1114.
46. Fawcett, T. An introduction to ROC analysis. Pattern Recognit. Lett. 2005, 27, 861–874.
47. Thuiller, W. BIOMOD—Optimizing predictions of species distributions and projecting potential future shifts under global change. Glob. Chang. Biol. 2003, 9, 1353–1362.
48. Phillips, S.J.; Anderson, R.P.; Schapire, R.P. Maximum entropy modeling of species geographic distributions. Ecol. Model. 2006, 190, 231–259.
49. Lindgren, F.; Rue, H.; Lindström, J. An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. J. R. Stat. Soc. B 2011, 73, 423–498.

Annex II

Species distribution modelling through Bayesian hierarchical approach

Oscar Rodríguez de Rivera[1,2] & Marta Blangiardo[2] & Antonio López-Quílez[1] & Ignacio Martín-Sanz[3]

* Correspondence: osroderi@alumni.uv.es; Tel.: +44-(0)7858-714047

Springer Nature B.V. 2018

Received: 30 June 2017 /Accepted: 21 June 2018

[1] Departament d’Estadística i I.O, Universitat de Valencia, Dr. Moliner 50, 46100 Burjassot, Spain

[2] MRC Centre for Environment and Health, Department of Epidemiology and Biostatistics, Imperial College London, London, UK

[3] Departament Sistemas y Recursos Naturales, ETSI de Montes, Forestal y del Medio Natural, Universidad Politécnica de Madrid, Ciudad Universitaria s/n, Madrid, Spain

Abstract

Usually in Ecology, the availability and quality of the data is not as good as we would like. For some species, the typical environmental study focuses on presence/absence data, and particularly with small animals as amphibians and reptiles, the number of presences can be rather small. The aim of this study is to develop a spatial model for studying animal data with a low level of presences; we specify a Gaussian Markov Random Field for modelling the spatial component and evaluate the inclusion of environmental covariates. To assess the model suitability, we useWatanabe-Akaike information criteria (WAIC) and the conditional predictive ordinate (CPO). We apply this framework to model each species of amphibian and reptiles present in the Las Tablas de Daimiel National Park (Spain).

Keywords Stochastic partial differential equation . Integrated nested Laplace approximation. Species distribution . Spatial model.

Introduction

Species distribution models (SDM) commonly used in ecology consist of numerical tools that combine observations of species occurrence or abundance with environmental covariates. They are used to gain ecological and example of evolutionary insight from SDM and to predict distributions across landscapes, sometimes requiring extrapolation in space and time (Elith and Leathwick 2009). In SDM, the following steps are usually taken: (1) locations of occurrence of a species (or other phenomenon) are compiled; (2) values of environmental predictor variables (such as climate) at these locations are extracted from spatial databases; (3) the environmental values are used to fit a model to estimate similarity to the sites of occurrence, or another measure such as abundance of the species; (4) the model is used to predict the variable of interest across the study region (and perhaps for a future or past climate) (Hijmans and Elith 2015).

Currently, the statistical understanding of applied scientists is increasing and new techniques can cope with larger, more complex data sets, so applied statisticians are faced with the need to specify sophisticated models. Logically, as the complexity of these models increase, it becomes harder to perform inference. The Bayesian approach is particularly appropriate as it is flexible and can deal with complex models, for instance including hierarchical structure or including missing data. Undoubtedly, the most popular family of approximate inference methods in Bayesian statistics is the class of Markov Chain Monte Carlo (MCMC) methods. These methods, which exploded into popularity in the mid-1980s have remained at the forefront of Bayesian statistics ever since, with the basic framework being extended to cope with increasingly more complex problems (Simpson et al. 2011).

Modelling patterns of the presence/absence of the species using local environmental factors have been a growing problem in Ecology in the last few years (Chakraborty et al. 2010). This kind of modelling has been extensively used to address several issues, including the identification of essential fauna habitats in order to classify and manage conservation areas (Pressey et al. 2007), and predicting the response of species to environmental features (Midgley and Thuiller 2007; Loarie et al. 2008). Different approaches and methodologies have been proposed in this perspective (see for instance Guisan and Thuiller 2005; Hijman and Graham 2006; Wisz et al. 2008), with Maximum Entropy modelling (MaxEnt) (Elith and Burgman 2002), more flexible models as generalised linear and additive ones (GLM and GAM) (Guisan et al. 2002), species envelope models such as BIOCLIM (Busby 1991) and the multivariate adaptive regression splines (MARS) (Leathwick et al. 2005) being some of the most commonly used (Muñoz et al. 2013).

Several projects have focused on comparing these different methods (see for instance Rivera and López-Quílez 2017); also a summary this comparison has been developed recently by Lecours 2017. Most of these applications consist of explanatory models that seek to assess the relationship between environmental variables (e.g. precipitation, bathymetry, etc.) (Guisan et al. 2002). Moreover, the theory of these methods is based on the fact that the observations are independent, while spatial autocorrelation is common in georeferenced ecological data (Crase et al. 2012). Spatial autocorrelation should be taken into account in the species distributionmodels, even if the data were collected through a standardised sampling scheme, since the observations are often close and subject to similar environmental features (Underwood 1981; Hurlbert 1984). In addition observer error (Royle et al. 2007; Cressie et al. 2009), gaps in the sampling, missing data, and spatial mobility of the species (Gelfand et al. 2006) can also affect the models.

The value of both reptiles and amphibians has been recognised as an integral part of natural ecosystems and as heralds of environmental quality (Gibbons and Stangel 1999). In recent years, as overall environmental awareness among the public has increased, concerns have raised on the ecological state of reptile and amphibian species as well as of their habitats (Gibbons et al. 2000). Habitats of many amphibians populations are small, temporary ponds and the surrounding forested area, which are usually affected by many stressors such as UV-radiation (Cummins 2003; Hatch and Blaustein 2003), the use of pesticides (Gendron et al. 2006; Fellers et al. 2004), industrial chemicals (Bishop and Gendron 1998; Sower et al. 2000) and climate change (Corn 2005). Since amphibians are sensitive to the alterations of their environment, they could be used as bioindicator organisms to follow changes in their habitats and in ecotoxicological studies (Henry 2000). As their population usually contains high numbers of individuals and they are good representatives of freshwater environments, they are ideal model organisms for pollution studies (Burger and

Snodgrass 1998). Gibbons et al. (2000), consider the vulnerability of reptiles within the context of the factors known or suspected to be associated with amphibian declines, using six categories of concern established by Partners in Amphibian and Reptile Conservation (PARC; Gibbons and Stangel 1999): habitat loss and degradation, introduced invasive species, environmental pollution, disease and parasitism, unsustainable use and global climate change.

The aim of this paper is to build a spatial model to predict the spatial distribution of several species characterised by a low level of presences, which leads to data sparsity. We will use real data on five species of amphibians obtained from inventories developed in Las Tablas de Daimiel National Park (TDNP-Spain) in 2011–2012 supported with environmental variables. On these species, we have presence/absence at geographical coordinates and we will generate distribution models for each species aswell as combine these into the corresponding the class (Amphibia). Our approach is to specify a Bayesian hierarchical geostatistical modelling framework accounting for spatial dependency. Hierarchical models can simplify complex interactions by allowing parameters to vary at more than one level via the introduction of random effects. The expected value of the response is then expressed conditional on these random effects (Cosandey-Godin et al. 2015). The advantages of using hierarchical Bayesian models emerge more so as complexity increases, when, for example, spatio-temporal variability needs to be modelled explicitly (Cressie et al. 2009). The Bayesian framework also offers the advantage of providing full inference, such that model parameters and uncertainty can be quantified, which has great utility in applied conservation (Wade 2000; Wintle et al. 2003).

Several authors have used a Bayesian approach to analyse species distribution. For instance Golding and Purse 2016, compared Gaussian processes against more traditional techniques obtaining better performance, while Gelfand et al. 2006, tried to illustrate spatial patterns applying hierarchical logistic regression trough a Bayesian framework. Also, Latimer et al. 2006, developed Bayesian regression models for species presence/absence and Royle 2004, estimated abundance of birds applying N-mixture model in a Bayesian perspective. Another interesting work developed by Mackenzie et al. 2002, focused on understanding site occupancy of some amphibians species when detection probabilities are below 1, again in a Bayesian approach.

Hierarchical Bayesian models have traditionally relied on MCMCsimulation techniques, which are computationally expensive and technically challenging, consequently limiting their use. However, a new statistical approach is now readily available, namely integrated nested Laplace approximations (INLA) via the R-INLA package (http://www.r-inla.org) (Cosandey-Godin et al. 2015). INLA methodology and its powerful application to modelling complex datasets has recently been introduced to a wider non-technical audience (Illian et al. 2013). As opposed to MCMC simulations, INLA uses an approximation for inference and hence avoids the intense computational demands, convergence, and mixing problems sometimes encountered by MCMC algorithms (Rue and Martino 2007). It can only be used for Gaussian models but this includes the class of models which we consider here for species distribution. Moreover, R-INLA can be compiled with the stochastic partial differential equations (SPDE) approach (Lindgren et al. 2011) which through a discretisation of a continuous Gaussian field can cope efficiently with variables characterised by a complex spatial structure. This is the case of this environmental inventory, since environmentalists or field workers start the inventory to target particular species, resulting in clustered spatial patterns and large regions without any values. Together, these new statistical methods and their implementation in R (R Core Team 2016) allow scientists to fit complex spatio-temporal models considerably faster and more reliably (Rue et al. 2009).

The structure of the paper is as follows. In Section 2, we introduce our motivating problem regarding spatial distribution for amphibian species in Las Tablas de Daimiel National Park (TDNP) (Ciudad Real, Spain). Then, after discussing the available data, we describe the geostatistical spatial model. In Section 3, we explain the model evaluation and the comparison. In Section 4, we present the results of the analysis of the spatial distributions and show how the environmental variables could affect the presence of the species. Finally, in Section 5, we resume the conclusions of this work.

Motivating example

The data set come from an inventory developed in Las Tablas de Daimiel National Park (TDNP) during 2011 and 2012, comprising 234 sample points with coordinates. Each sample point has the presence or absence of each species, elevation inmeters and information about the ambient (categorical variable with the following categories: Salt marsh, Reed bed, Islands, areas of Typha latifolia, Cladium mariscus and free of vegetation).

The following species are included: Bufo bufo, Bufo calamita, Pelobates cultripes, Pelodytes punctatus and Triturus pygmaeus (see Table 1). As the aim of this analysis is not to study the distribution of each species, we are not going to explain the biology or characteristics of each species. The Tablas de Daimiel National Park is a floodplain wetland located in the Upper Guadiana Basin, central Spain (see Fig. 1). The landscape of Las Tablas de Daimiel is characterised by the horizontality of the terrain, with a range of altitude between the 599 m above sea level in the confluence of the rivers Guadiana and Cigüela, and the 623 m above sea level in the Pochela hill.

Table 1: Species presence (in percentage) by ambient (*C.m.*=*Cladium mariscus*, F.v.=Free of vegetation, I.=Islands, R.b.=Reed bed, S.m.=Salt marsh and *T.l.*=*Typha latifolia*) and by number of presences

	C.m.	**F.v.**	**I.**	**R.b.**	**S.m.**	***T.l.***	**Presences**
Bufo bufo	0%	0%	0%	100%	0%	0%	3%
Bufo calamita	5%	5%	48%	19%	0%	24%	54%
Pelobates cultripes	0%	0%	0%	100%	0%	0%	3%
Pelodytes punctatus	0%	21%	14%	50%	0%	14%	36%
Triturus pygmaeus	0%	0%	0%	50%	0%	50%	5%
Amphibia	3%	10%	31%	36%	0%	21%	100%

The Guadiana River is one of the three main drainage units of the Iberian Peninsula, having its source in central Spain before flowing into Portugal and then, in its lower reaches, acting as a natural border between the two countries. The TDNP is one of the core areas of the Mancha Húmeda, declared a Biosphere Reserve in 1980 by UNESCO. The wetland is the result of the mixture of inputs from Cigüela and Guadiana rivers, together with groundwater discharge from the West Mancha aquifer. The peripheral surface of the wetland is 1928 ha, but at present, the potentially flooded area is 1587 ha (Sánchez-Carrillo et al. 2010). The climate of the Upper

Guadiana Basin is semiarid, with an average temperature of 14.1 °C (1955–2009) and an average precipitation of 448 mm (1945–2009) (Yustres et al. 2013). The spatiotemporal distribution of rain is irregular (Acreman et al. 2000), and the high temperature values in the summer cause the potential evapotranspiration to be notably high. Subject to this semiarid climate regime, the water balance is particularly fragile in the river basin, with water shortage considered a structural characteristic of the system (Cots et al. 2007).

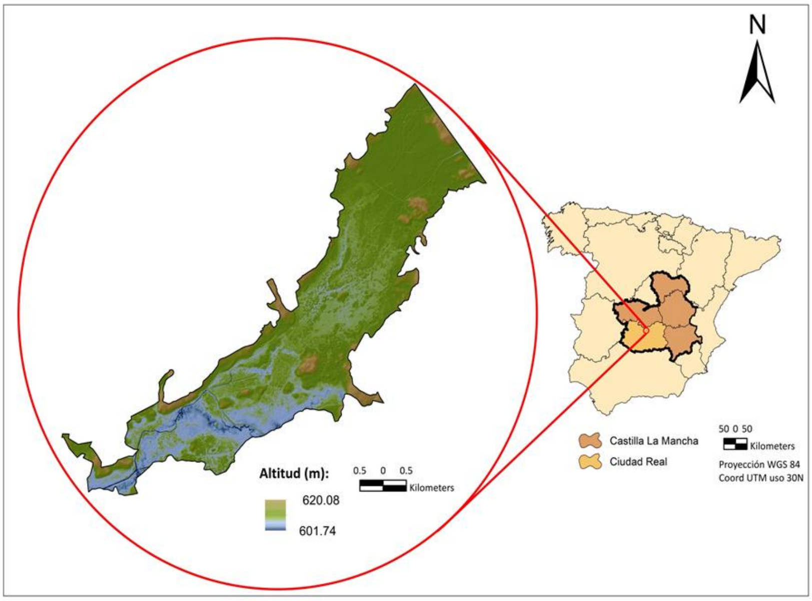


Fig. 1 Map of the situation of the Las Tablas de Daimiel National Park

Currently, the TDNP suffers a reduction in water inputs mainly because the groundwater discharge to the wetland is decreasing, since the aquifer has been suffering intensive groundwater overexploitation since the late 1970s (Navarro et al. 2011). The water quality has been also affected, since one of its peculiarities is the variety of hydrochemical and hydrological processes, i.e. the two main sources of water and salts and at the same time TDNP is an area where saline sulfate-rich surface waters flowing from the Cigüela River mix with calcium carbonate freshwater coming from groundwater (Coronado et al. 1974; Sánchez-Ramos et al. 2015).

We can see in Fig. 2 the distribution of the ambients in our data set, summarised as follows: Salt marsh (2% of the samples), Reed bed (31%), Islands (17%), areas of Typha latifolia (34%), Cladium mariscus (4%) and free of vegetation (13%).

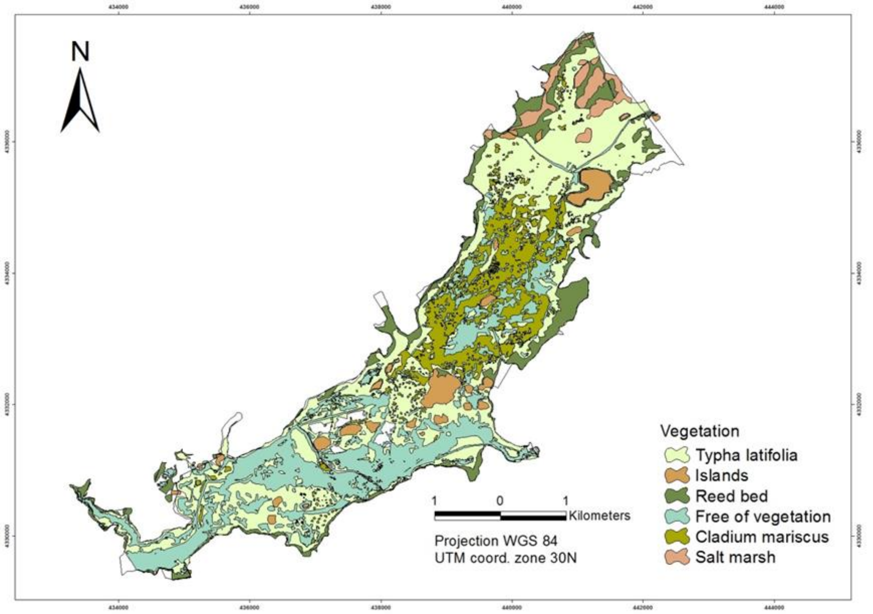


Fig. 2 Map of the distribution of the ambients in the Las Tablas de Daimiel National Park

Species distribution model

Model

Spatial data are defined as realisations of a stochastic process indexed by space:

$$Y(s) \equiv \{y(s), s\epsilon D\}$$

where D is a (fixed) subset of R^d (here we consider $d = 2$). The actual data can be then represented by a collection of observations y = $\{y(s_1), ..., y(s_n)\}$, where the set $(s_1, ..., s_n)$ indicates the spatial units at which the measurements are taken. Depending on D being a continuous surface or a countable collection of d-dimensional spatial units, the problem can be specified as a spatially continuous or discrete random process, respectively (Gelfand et al., 2010).

In our case, we can consider a collection of data points with presence/absence obtained from the inventory; the sampled points are the set $(s_1, ..., s_n)$ of n points; $y_s$is the presence of each specie in each point and it is specified as

$$y_s \sim Bernoulli(\pi_s)$$

where π_s is the probability of the species being present.

Then on the logit(π_s) a linear model is specified including covariates $\mathbf{x_1}$ (elevation), $\mathbf{x_2}$(ambient) and a spatial field ξ_s

$$\text{logit}(\pi_s)= b_s + x_{1S}\, \beta + x_{2S}\, \gamma + \xi_s$$

where a discretely indexed spatial random process (see Lindgren et al. 2011) is included to approximate the continuous process:

$$\xi_s = \sum \varphi_g(s)\, \widetilde{\xi_g}$$

In practice the discretisation is done dividing the study region in triangles and writing ξ_s as a linear combination of basic functions φ_g weighted by some zero mean terms $\widetilde{\xi_g}$ (for more details see Blangiardo and Cameletti 2015).

Implementation, inference and evaluation

The statistical inference has been carried out through the integrated nested Laplace approximation (INLA) implemented in the R-INLAwithin R statistical software. In R-INLA, the first step required to run the geostatistical spatial model through SPDE is the triangulation of the considered spatial domain. We use the inla.mesh.create specifying the spatial coordinates used for estimation. This function performs a constrained refined Delaunay triangulation for a set of spatial locations: firstly the triangle vertices are placed at the observation locations and then further vertices are added in order to satisfy triangulation quality constraints (Lindgren et al. 2011).

A natural way to estimate out-of-sample prediction error is cross-validation (see Geisser and Eddy 1979, and Vehtari and Lampinen 2002, for a Bayesian perspective), but researchers have always sought alternative measures, as cross-validation requires repeated model fits and can run into trouble with sparse data (Gelman and Shalizi 2013). In a comparative perspective (e.g. to evaluate which model fits the data best), the most used index is the DIC (Spiegelhalter et al. 2002; van der Linde 2005) which similarly to AIC consists of two components, a term that measures goodness of fit and a penalty term for increasing model complexity. More recently, the WAIC (Watanabe 2010) has been proposed as a suitable alternative for estimating the out-of-sample expectation in a fully Bayesian approach. This approach starts with the computed log pointwise posterior predictive density and then adds a correction for the effective number of parameters to adjust for overfitting (Gelman and Shalizi 2013). WAIC operates on predictive probability density of observed variables rather than on model parameter; hence, it can be applied in singular statistical models (i.e. models with nonidentifiable parameterization, see Li et al. 2015).

We have also calculated the conditional predictive ordinate (CPO) (Pettit 1990) to evaluate model assessment. The conditional predictive ordinate (CPO) is based on leave-one-outcross- validation. CPO estimates the probability of observing a value after having already observed the others. The mean logarithmic score (LCPO) was calculated as a measure of the predictive quality of the model (Gneiting and Raftery 2007; Roos and Held 2011). High LCPO values suggest possible outliers, high-leverage and influential observations.

Finally, we have used an AUC (Area Under operating Curve score) approach to calculate the predictive accuracy of each method by comparing the validation data with the predicted presence value. AUC represents a commonly used and adequately performing measure of predictive accuracy (Huang and Ling 2005) and works by calculating the relative numbers of correctly and incorrectly identified predictions across all possible classification threshold values of the binomial response, with an AUC value equal to or below 0.5 indicating a predictive ability equal to random expectation and 1 a perfect predictive ability (Qiao et al. 2015).

Results

Table 2 presents the main results of the analyses for the Amphibia, characterised by more data sparsity (a fewer presences). As we can see, the first model obtained using only the elevation has a better fit (WAIC = 21.17/LCPO = 1.759/ AUC = 0.762) than the model with elevation and ambient (WAIC = 24.53/LCPO = 1.898/AUC = 0.735). However, based on LCPO, the model without ambient has fewer outliers. Also, we have compared performance of the different models based in AUC, these analysis shows similar results than LCPO, obtaining better values in models without ambient.

Table 2: Posterior estimates for the models with elevation and with elevation and ambient

	Parameter	Mean	St. Dev.	2.5%	50%	97.5%
Model with elevation						
Amphibia **WAIC=21.17**	Intercept	-4.69	30.36	-64.30	-4.69	54.86
LCPO=1.759 **AUC=0.762**	Elevation	0.01	0.05	-0.10	0.01	0.10
	Model with elevation and ambient					
Amphibia **WAIC=24.53** **LCPO=1.898** **AUC=0.735**	Intercept	-4.70	30.51	-64.62	-4.70	55.16
	Elevation	-0.03	0.05	-0.13	-0.03	0.08
	Ambient. Reedbed	0.30	6.34	-21.97	0.75	13.04
	Ambient. Islands	0.71	8.16	-15.96	0.62	15.32
	Ambient. Freeofveg	0.82	6.74	-16.49	0.75	20.86
	Ambient. C.mariscus	-0.72	8.37	-25.06	0.69	7.13
	Ambient. Saltmarsh	-1.17	13.09	-33.29	-2.19	45.36

In Table 3, we can see the summary of theWAIC, LCPO and AUC values obtained in the different models for each species (model E with only Elevation and model E&A with Elevation and Ambient); this shows that, looking at theWAIC, most of the species have a better fit for the model with vegetation (except Bufo bufo and Pelobates cultripes), but looking at LCPO values Ambient seems to increase the number of outliers.. Also, looking at AUC values, models with ambient have lower predictability.

Table 3: WAIC, LCPO and AUC comparison in species models

	E			E&A		
Amphibia	WAIC	LCPO	AUC	WAIC	LCPO	AUC
Bufo bufo	1.83	1.965	0.726	13.03	2.260	0.712
Bufo calamita	14.04	1.521	0.783	13.03	1.531	0.752
Pelobates cultripes	1.103	1.894	0.83	3.27	2.265	0.758
Pelodytes punctatus	10.79	2.523	0.693	5.43	3.914	0.656
Triturus pygmaeus	4.61	2.327	0.725	3.38	2.986	0.718

From an ecological perspective, amphibian as group (Table 2) does not have any preference between the different ambients analysed, with small negative point estimates, but with a credibility intervals including zero. Analysing the different species (Table 4),we can summarise as follows: Bufo calamita has all the variables have point estimates positives and excluding zero except areas with Cladium mariscus where the point estimate includes zero; Pelodytes punctatus and Triturus pygmaeus have a different relationship with the variables, with all the point estimates including zero; however, Bufo bufo has predictability for areas with Reed bed as is the only variable with point estimate positive and the zero is not included; finally, Pelobates cultripes has better relationship with Reed bed, Islands and Free of vegetation areas.

In Fig. 3, we can see the maps of posterior mean for class Amphibia of models developed based on the covariates used. In both cases, we can see that the distribution are complementary: due to the nature of amphibians and reptiles, there are no points with both classes present at the same time. On the bottom, the map shows the distribution of the model with elevation and ambient. As we can see, the maps obtained with elevation and with elevation and ambient are very similar. Also, the model with only elevation as covariate has more details than the model with Ambient.

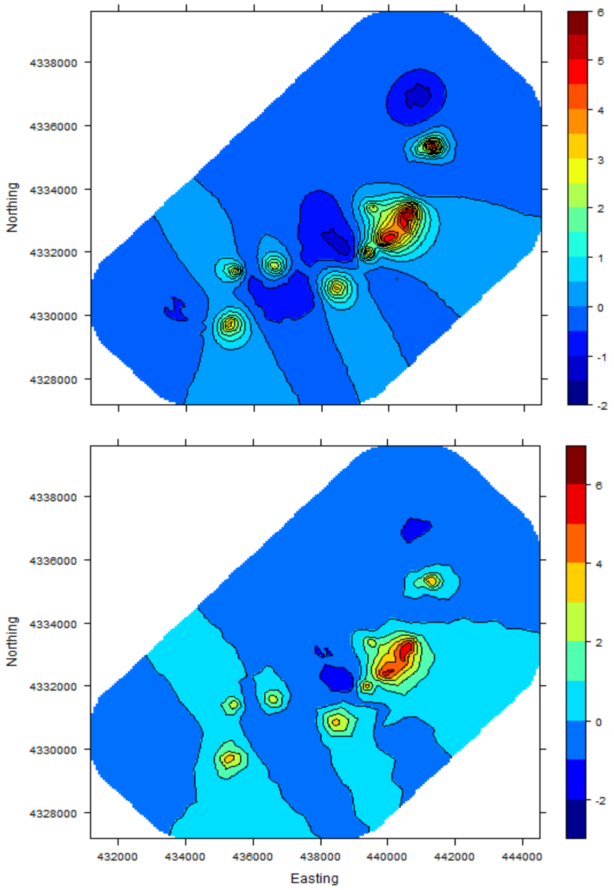


Fig. 3 Maps of posterior mean for the models with only Elevation as covariate (top); Elevation and Ambient (bottom)

Conclusions

Hierarchical models are commonly used in ecology (Clark 2005; Cressie et al. 2009). The hierarchical modelling framework has been useful when implementing process models that include ecological theory or when modelling the data collection process (Hooten and Wikle 2008; Hooten et al. 2007, Wikle 2003). The hierarchical modelling framework has also been useful and commonly implemented for count and presence/absence data (MacKenzie et al. 2006; Royle and Dorazio 2008), but has not been readily used for presenceonly data (Dorazio 2014; Fithian et al. 2015). However, we can affirm that absences are giving important information to define species distribution. Also, the hierarchical species distribution modelling approach can be readily extended to include multiple species, as we have done with amphibians, and possibly interacting species (Hui 2016; Ovaskainen and Soininen 2011; Warton et al. 2015). The spatio-temporal Poisson point process model is currently used for the analysis of species movement data captured using telemetry devices (Brost et al. 2015; Johnson et al. 2013; Russell et al. 2016).

Data models developed for telemetry data will have a similar use for species distribution models (Brost et al. 2015). The methodology used to account for repeated measurements (i.e., locations) of the same individual(s) developed for telemetry data will have analogous use for species distribution models that are used to model count, presence-absence, and presence-only data that includes multiple observations of the same individuals (Hefley and Hooten 2016).

In this work, we have specified a Bayesian spatial model for studying species distribution. We have evaluated the inclusion of two variables (elevation and ambient).

The main advantage of the Bayesian model formulation is the computational ease in model fit and prediction compared to classical geostatistical methods. The main goal of this study has been to predict the occurrence of species with a relatively small number of data points, but the data was useful to show the power of this kind of process and the options of the model construction. To do so, instead of MCMC, we have used the novel integrated nested Laplace approximation approach. More precisely, we have applied the work of Lindgren et al. (2011), which provides a link between Gaussian Fields and Gaussian Markov Random Fields through the

Stochastic Partial Differential Equation (SPDE) approach. The SPDE approach can be easily implemented providing results in reasonable computing time (comparing with MCMC). We showed how SPDE is as useful tool in the analysis of species distribution. This modelling could be expanded to the spatiotemporal domain by incorporating an extra term for the temporal effect, using parametric or semiparametric constructions to reflect linear, non-linear, autoregressive or more complex behavior.

On the other hand, we have concluded some interesting points from an ecological perspective. Amphibians are easy species to model: due to their dependence to the water, most of the species live in freshwater aquatic ecosystems. The relationship between the amphibians and the water is stronger than the relationship with the ambient, probably because most of the amphibians use the water as ambient for the reproductive habits. At the same time, vegetation and environment are less important than the elevation (as a distance of the water surface). Wetlands are essential breeding habitats for many amphibian species. Pond-breeding amphibians require aquatic habitats for breeding, and embryonic and larval development, whereas terrestrial habitats are used for foraging and aestivation, and as migration and dispersal routes. Both aquatic and terrestrial habitats are used for hibernation. Hence, pondbreeding amphibians can be susceptible to changes in the availability and quality of both local-scale (aquatic habitats) and landscape-scale habitat characteristics (Piha et al. 2007). Currently, habitat loss is considered one of the greatest threats to the world’s amphibian species, one-third of which are threatened (Stuart et al. 2004). Pond-breeding amphibians may be particularly influenced by the loss and increased isolation of important habitat types caused by agricultural intensification in our case due to the water lost from the agricultural ponds.

We understand that extending this framework to situations characterised by environmental changes, there is the possibility to experience climatic changes between points. And this framework could benefit by the inclusion of meteorological variables. However, Las Tablas de Daimiel National Park has a really small extension with only one meteorological station (IGME 2017), so we have avoided the use of climatic features in order to not include estimated data into the model selection.

Also, we can see that hierarchical models are particularly useful when data are sparse or species are similar. In our case, Amphibia model has different relationship with the environmental variables than the individual species included in the class model.

As we have explained about the study area, the National Park is a water area of about 1600 ha. As we are working with amphibians that develop most of their cycle in water and also, as there is no way to introduce the distance to the water due to the variation in water level, we have introduced elevation as a proxy. We assume that elevation can be affected by spatial autocorrelation. However, all methods assume spatial stationarity, i.e. spatial autocorrelation does only depend on distance between point locations, and there are very few methods to deal with non-stationarity in this context (Osborne et al. 2007).

Finally, we can conclude that due to the low level of observations, CPO is more robust than WAIC due to the presence of influential observations.

We conclude that SPDE and INLA are promising tools to work with species distribution model as they save in computational times and are easy to specify and to implement also for non-statisticians when we work with a large data set.

Summarising, R-INLA can be a complementary tool for ecologists. The major strength of R-INLA is that it allows to perform Bayesian inference, based on highly accurate approximations of posterior distributions,where models are specified using a syntax that should be familiar to R users, and where data are formatted in a straightforward way with relatively few lines of code. The straightforwardmodel syntax and data format could help remove barriers to the adoption of N-mixture models for biologists. The substantial decrease in computation time should also facilitate the use of a wider variety of model and variable selection techniques (e.g. cross-validation and model averaging) that are not commonly used in an MCMC context due to practical issues related to computing time (Kery and Schaub 2011).

Limitations of R-INLA are mostly related to the more restricted set of N-mixture models that can be specified. R-INLA does not handle site survey covariates, employs only Poisson-Binomial and Negative Binomial-Binomial mixtures, and handles random effects (exchangeable, spatially and temporally structured) for p only. In cases where site survey covariates are particularly significant and not otherwise controlled in the sampling design, R-INLA will not be the suitable tool (Meehan et al. 2017).

References

1. Acreman M, Almagro J, Alvarez J, Bouraoui F, Bradford R, Bromley J, Croke B, Crooks S, Cruces J, Dolz J, Dunbar M, Estrela T, Fernandez-Carrasco P, Fornes J, Gustard G, Haverkamp R, De La Hera A, Hernández-Mora N, Llamas R, Martinez CL, Papamasorakis J, Ragab R, Sánchez M, Vardavas I, Webb T (2000) Groundwater and river resources programme on a European scale (GRAPES). Technical report to the European Union ENV4–CT95-0186. Institute of Hydrology, Wallingford
2. Bishop CA, Gendron AD (1998) Reptiles and amphibians: shy and sensitive vertebrates of the Great Lakes basin and St. Lawrence river. Environ Monit Assess 53:225–244
3. Blangiardo M, CamelettiM (2015) Spatial and spatio-temporal Bayesian models with R-INLA. John Wiley & Sons, Chichester
4. Brost BM, Hooten MB, Hanks EM, Small RJ (2015) Animal movement constraints improve resource selection inference in the presence of telemetry error. Ecology 96:2590–2597
5. Burger J, Snodgrass J (1998)Heavymetals in bullfrog (Rana catesbeiana) tadpoles: effects of depuration before analysis. Environ Toxicol Chem 17:2203–2209
6. Busby JR (1991) BIOCLIM: a bioclimatic analysis and predictive system. In: Margules C, Austin M (eds) Nature conservation: cost effective biological surveys and data analysis. CSIRO, Canberra, pp 64–68
7. Chakraborty A, Gelfand AE, Wilson AM, Latimer AM, Silander JA (2010) Modeling large scale species abundance with latent spatial processes. Ann Appl Stat 4(3):1403–1429
8. Clark JS (2005) Why environmental scientists are becoming Bayesians. Ecol Lett 8(1):2–14
9. Corn SP (2005) Climate change and amphibians. Anim Biodivers Conserv 28:59–67
10. Coronado R, Del Portillo F, Sáez-Royuela R (1974) Tablas de Daimiel National Park Guide. ICONA, Madrid
11. Cosandey-Godin A, Teixeira Krainski E, Worm B, Mills Flemming J (2015) Applying Bayesian spatiotemporal models to fisheries bycatch in the Canadian Arctic. Can J Fish Aquat Sci 72:1–12

12. Cots F, David Tàbara J, Werners S, McEvoy D (2007) Climate change and water adaptive management through transboundary cooperation. The case of the Guadiana river basin. Paper presented to the first International Conference on Adaptive and Integrative Water Management (CAIWA), Basel, Switzerland, November 2007
13. Crase B, Liedloff AC,Wintle BA (2012) A new method for dealing with residual spatial autocorrelation in species distribution models. Ecography 35(10):879–888
14. Cressie N, Calder CA, Clark JS,Hoef JMV,Wikle CK(2009) Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. Ecol Appl 19:553–5701
15. Cummins CP (2003) UV-B radiation, climate change and frogs—the importance of phenology. Ann Zool Fenn 40:61–67
16. Dorazio RM (2014) Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. Glob Ecol Biogeogr 23(12):1472–1484
17. Elith J, Burgman MA (2002) Predictions and their validation: rare plants in the Central Highlands, Victoria, Australia. In: Scott JM, Heglund PJ, Morrison ML, Raphael MG, Wall WA, Samson FB (eds) Predicting Species Occurrences: Issues of Accuracy and Scale. Island Press, Covelo, pp 303–314
18. Elith J, Leathwick JR (2009) Species distribution models: ecological explanation and prediction across space and time. Annu Rev Ecol Evol Syst 40:677–697
19. Fellers GM, Mcconnell LL, Pratt D, Datta S (2004) Pesticides in mountain yellow legged frogs (Rana muscosa) from the Sierra Nevada Mountains of California, USA. Environ Toxicol Chem 23:2170–2177
20. Fithian W, Elith J, Hastie T, Keith DA (2015) Bias correction in species distribution models: pooling survey and collection data for multiple species. Methods Ecol Evol 6(4):424–438
21. Geisser S, Eddy W (1979) A predictive approach to model selection. J Am Stat Assoc 74:153–160
22. Gelfand AE, Silander JA,Wu SJ, Latimer AM, Rebelo PLAG, HolderM (2006) Explaining species distribution patterns through hierarchical modeling. Bayesian Anal 1(1):41–92

23. Gelfand AE, Diggle P, Fuentes M, Guttorp P (eds) (2010) Handbook of spatial statistics. Chapman & Hall, Boca-Raton
24. Gelman A, Shalizi C (2013) Philosophy and the practice of Bayesian statistics (with discussion). Br J Math Stat Psychol 66:8–80
25. Gendron AD, Marcogliese DJ, Barbeau S, Christin MS, Brousseau P, Ruby S, Cyr D, Fournier M (2006) Exposure of leopard frogs to a pesticide mixture affects life history characteristics of the lungworm Rhabdias ranae. Oecologia 135:469–476
26. Gibbons JW, Stangel PW (eds) (1999) Conserving amphibians and reptiles in the new millenium. Proceedings of the Partners in Amphibian and Reptile Conservation (PARC). Conference; 2–4 June 1999; Atlanta (GA).Aiken (SC): Savannah River Ecology Laboratory. Herp Outreach Publication #2
27. Gibbons JW, Scott DE, Ryan TJ, Buhlmann KA, Tuberville TD,Metts BS, Greene JL, Mills T, Leiden Y, Poppy S,Winne CT (2000) The global decline of reptiles, Déjà vu amphibians. BioScience 50(8):653–666
28. Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 102(477):359–378
29. Golding N, Purse BV (2016) Fast and flexible Bayesian species distribution modelling using Gaussian processes. Methods Ecol Evol 7(5):598–608
30. Guisan A, Thuiller W (2005) Predicting species distribution: offering more than simple habitat models. Ecol Lett 8:993–1009
31. Guisan A, Edwards TC, Hastie T (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecol Model 157:89–100
32. Hatch AC, Blaustein AR (2003) Combined effects of UV-B radiation and nitrate fertilizer on larval amphibians. Ecol Appl 13:1083–1093
33. Hefley TJ, Hooten MB (2016) Hierarchical species distribution models. Curr Landscape Ecol Rep 1(2):87–97
34. Henry PFP (2000) Aspects of amphibian anatomy and physiology. In: Sparling DW, Linder G, Bishop CA (eds) Ecotoxicology of amphibians and reptiles. SETAC Press, Pensacola, pp 71-110
35. Hijman R, Graham C (2006) The ability of climate envelope models to predict the effect of climate change on species distributions. Glob Chang Biol 12(12):2272–2281

36. Hijmans RJ, Elith J (2015) Species distribution modelling with R. http://cran.r-project.org/web/packages/dismo/vignettes/sdm.pdf, The R foundation for statistical computing
37. Hooten MB, Wikle CK (2008) A hierarchical Bayesian non-linear spatiotemporal model for the spread of invasive species with application to the Eurasian collared-dove. Environ Ecol Stat 15(1):59–70
38. Hooten MB, Wikle CK, Dorazio RM, Royle JA (2007) Hierarchical spatiotemporal matrix models for characterizing invasions. Biometrics 63(2):558–567
39. Huang J, Ling CX (2005) Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowl Data Eng 17(3):299–310
40. Hui FK (2016) Boral–Bayesian ordination and regression analysis of multivariate abundance data in R.Methods Ecol Evol 7(5):744–750
41. Hurlbert SH (1984) Pseudoreplication and the design of ecological field experiments. Ecol Monogr 54:187–211
42. IGME (Instituto Geológico y Minero de España) (2017) http://www.igme.es/zonas_humedas/daimiel/medio_fisico/clima.htm. Date visited: 01/03/2017
43. Illian JB, Martino S, Sørbye SH, Gallego-Fernández JB, Zunzunegui M, Esquivias MP, Travis JMJ (2013) Fitting complex ecological point process models with integrated nested Laplace approximation. Methods Ecol Evol 4:305–315. https://doi.org/10.1111/2041-210x.12017
44. Johnson DS, Hooten MB, Kuhn CE (2013) Estimating animal resource selection from telemetry data using point process models. J Anim Ecol 82(6):1155–1164
45. Kery M, Schaub M (2011) Bayesian population analysis using WinBUGS: a hierarchical perspective. Academic Press, Burlington
46. LatimerAM,Wu SS, Gelfand AE, Silander JA (2006) Building statistical models to analyze species distributions. Ecol Appl 16(1):33–50
47. Leathwick JR, Rowe D, Richardson J, Elith J, Hastie T (2005) Using multivariate adaptive regression splines to predict the distributions of New Zealand’s freshwater diadromous fish. Freshw Biol 50: 2034–2052
48. Lecours V (2017) On the use of maps and models in conservation and resource management (warning: results may vary). Front Mar Sci 4:288
49. Li L, Qiu S, Zhang B, Feng CX (2015) Approximating cross-validatory predictive evaluation in Bayesian latent variable model with integrated IS

and WAIC. Stat Comput 26:881–897. https://doi.org/10. 1007/s11222-015-9577-2

50. Lindgren F, Rue H, Lindström J (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach [with discussion]. J R Stat Soc B 73(4):423–498
51. Loarie SR, Carter BE, Hayhoe K, McMahon S, Moe R, Knight CA, Ackerly DD (2008) Climate change and the future of Californias endemic flora. PLoS One 3(6):e2502
52. MacKenzie DI, Nichols JD, Lachman GB, Droege S, Royle JA, Langtimm CA (2002) Estimating site occupancy rates when detection probabilities are less than one. Ecology 83:2248–2255
53. MacKenzie DI, Nichols JD, Royle JA, Pollock KH, Bailey LL, Hines JE (2006) Occupancy estimation and modeling inferring patterns and dynamics of species occurrence. Academic Press, Burlington
54. Meehan TD, Michel NL, Rue H (2017) Estimating animal abundance with N-mixture models using the R-INLA package for R arXiv preprint arXiv:1705.01581
55. Midgley GF, ThuillerW(2007) Potential vulnerability of Namaqualand plant diversity to anthropogenic climate change. J Arid Environ 70:615–628
56. Muñoz F, Pennino MG, Conesa D, López-Quílez A, Bellido JM (2013) Estimation and prediction of the spatial occurrence of fish species using Bayesian latent Gaussian models. Stoch Environ Res Risk Assess 27:1171–1180
57. Navarro V, García B, Sánchez D, Asensio L (2011) An evaluation of the application of treated sewage effluents in Las Tablas de Daimiel National Park, Central Spain. J Hydrol 401:53–64
58. Osborne PE, Foody GM, Suárez-Seoane S (2007) Non-stationarity and local approaches to modelling the distributions of wildlife. Divers Distrib 13(3):313–323
59. Ovaskainen O, Soininen J (2011) Making more out of sparse data: hierarchical modeling of species communities. Ecology 92(2):289–295
60. Pettit LI (1990) The conditional predictive ordinate for the normal distribution. J R Stat Soc Ser B 52(1):175–184
61. Piha H, Luoto M, Merilä J (2007) Amphibian occurrence is influenced by current and historic landscape characteristics. Ecol Appl 17(8):2298–2309

62. Pressey RL, Cabeza M, Watts EM, Cowling RM, Wilson KA (2007) Conservation planning in a changing world. Trends Ecol Evol 22: 583–592
63. Qiao H, Soberón J, Peterson AT (2015) No silver bullets in correlative ecological niche modelling: insights from testing among many potential algorithms for niche estimation. Methods Ecol Evol 6(10): 1126–1136. https://doi.org/10.1111/2041-210X.12397
64. R Core Team (2016) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/
65. Rivera ÓRD, López-Quílez A (2017) Development and comparison of species distribution models for forest inventories. ISPRS Int J Geo- Inf 6(6):176
66. RoosM, Held L (2011) Sensitivity analysis in Bayesian generalized linear mixed models for binary data. Bayesian Anal 6(2):259–278
67. Royle JA (2004) N-mixture models for estimating population size from spatially replicated counts. Biometrics 60:108–115
68. Royle JA, Dorazio RM (2008) Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities. Elsevier, Amsterdam
69. Royle JA, Kery M, Gautier R, Schmidt H (2007) Hierarchical spatial models of abundance and occurrence from imperfect survey data. Ecol Monogr 77:465–481
70. Rue H, Martino S (2007) Approximate Bayesian inference for hierarchical Gaussian Markov random field models. J Stat Plan Inference 137(10):3177–3192
71. Rue H,Martino S, Chopin N (2009) Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). J R Stat Soc Ser B 71:319–392
72. Russell JC, Hanks EM, Haran M (2016) Dynamic models of animal movement with spatial point process interactions. J Agric Biol Environ Stat 21(1):22–40
73. Sánchez-Carrillo S, Angeler D G, Álvarez-Cobelas M, Sánchez-Andrés R (2010) Freshwater wetland eutrophication. In Eutrophication: causes, consequences and control. Springer, Dordrecht, pp 195–210
74. Sánchez-Ramos D, Sánchez-Emeterio G, Florín Beltrán M (2015) Changes in water quality of treated sewage effluents by their receiving environments

in Tablas de Daimiel National Park, Spain. Environ Sci Pollut Res 23:6082–6090. https://doi.org/10.1007/s11356-015-4660-y

75. Simpson D, Lindgren F, Rue H (2011) Fast approximate inference with INLA: the past, the present and the future. Technical report at arxiv.org
76. Sower SA, Reed KL, Babbitt KJ (2000) Limb malformations and abnormal sex hormone concentrations in frogs. Environ Health Perspect 108:1085–1090
77. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A (2002) Bayesian measures of model complexity and fit (with discussion). J R Stat Soc Ser B 64(4):583–616
78. Stuart S, Chanson JS, Cox NA, Young BE, Rodrigues ASL, Fishman DL, Waller RW (2004) Status and trends of amphibian declines and extinctions worldwide. Science 306:1783–1786
79. Underwood AJ (1981) Techniques of analysis of variance in marine biology and ecology. Oceanogr Mar Biol Annu Rev 19:513–605
80. van der Linde A (2005) DIC in variable selection. Statistica Neerlandica 59(1):45–56
81. Vehtari A, Lampinen J (2002) Bayesian model assessment and comparison using cross validation predictive densities. Neural Comput 14:2439–2468
82. Wade PR (2000) Bayesian methods in conservation biology. Conserv Biol 14(5):1308–1316. https://doi.org/10.1046/j.1523-1739.2000.99415.x
83. Warton DI, Blanchet FG, O’Hara RB, Ovaskainen O, Taskinen S,Walker SC, Hui FK (2015) Somany variables: jointmodeling in community ecology. Trends Ecol Evol 30(12):766–779
84. Watanabe S (2010) Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J Mach Learn Res 11:3571–3594
85. Wikle CK (2003) Hierarchical Bayesian models for predicting the spread of ecological processes. Ecology 84(6):1382–1394
86. Wintle BA, McCarthy MA, Volinsky CT, Kavanagh RP (2003) The use of Bayesian model averaging to better represent uncertainty in ecological models. Conserv Biol 17(6):1579–1590. https://doi.org/10.1111/j.1523-1739.2003.00614.x
87. Wisz MS, Hijmans RJ, Li J, Peterson AT, Graham CH, Guisan A (2008) Effects of sample size on the performance of species distribution models. Divers Distrib 14:763–773

88. Yustres A, Navarro V, Asensio L, Candel M, García B (2013) Groundwater resources in the Upper Guadiana Basin (Spain): a regional modelling analysis. Hydrogeol J 21(5):1129–1146

Annex III

Assessing the Spatial and Spatio-Temporal Distribution of Forest Species via Bayesian Hierarchical Modeling

Óscar Rodríguez de Rivera [1,2,]*, Antonio López-Quílez [1] and Marta Blangiardo [2]

[1] Department of Statistics and Operational Research. University of Valencia. Faculty of Mathematics. C/Dr. Moliner, 50. Burjassot, 46100 Valencia, Spain; antonio.lopez@uv.es

[2] MRC Centre for Environment and Health, Department of Epidemiology and Biostatistics, St Mary's Campus, Imperial College London, London SW7 2AZ, UK; m.blangiardo@imperial.ac.uk

* Correspondence: osroderi@alumni.uv.es; Tel.: +44-7858-714047

Received: 2 August 2018; Accepted: 11 September 2018; Published: date

Abstract: Climatic change is expected to affect forest development in the short term, as well as the spatial distribution of species in the long term. Species distribution models are potentially useful tools for guiding species choices in reforestation and forest management prescriptions to address climate change. The aim of this study is to build spatial and spatio-temporal models to predict the distribution of four different species present in the Spanish Forest Inventory. We have compared the different models and showed how accounting for dependencies in space and time affect the relationship between species and environmental variables.

Keywords: hierarchical Bayesian models; stochastic partial differential equation; integrated nested laplace approximation; species distribution; spatial model; spatio-temporal model

1. Introduction

Tree species distributions are undoubtedly associated with climatic factors through the direct effects of climate circumstances on tree biological processes (Running at al. 2004). Consequently, climate change is likely to affect forest development in the short term (Boisvenue & Running, 2006), as well as spatial distribution of species in the long term, through demographic processes (Davis & Shaw, 2001, Parmesan & Yohe, 2003). For temperature-limited boreal woods, the main expectation is a deep northward shift of appropriate tree species environment, while the situation in temperate areas tends to be more complex and is different between Mediterranean, continental, and maritime climates (Bonan, 2008). Climate impacts, such as water deficit and the elevated risk of forest fires, will threaten Mediterranean forests (Adams et al. 2010, Schröter et al. 2005), while forest development might benefit in continental and Atlantic forests, but only at sites where an increased evaporative demand can be satisfied by enough water availability (Lindner et al. 2010, Spathelf et al. 2014, Rivera et al. 2018).

Although the predicted impacts of global warming, there are uncertainties around the magnitude of the effects. Among forest researchers, awareness has increased that the global warming poses a huge impact to the management and environmental value of woodland areas (Lindner et al. 2014). In Europe, it was projected that financial losses may come to several billion euros by the end of this century if policies for the forest sector do not change in response to the predicted climate changes (Hanewinkel et al. 2013).

To help species selection in reforestation and forest management treatments to manage climate change, species distribution models (SDMs) are a valuable tool (Maaten et al. 2017). Species distribution models (SDMs) can be defined as a mathematical approach built on combination of observations of species presence or abundance with environmental factors. These models are treated to evaluate species distributions across sceneries (Elith & Leathwick, 2009). Although there are exceptions (e.g., O'Neill et al. 2008), SDMs usually predict the appropriate niches of species (Maaten et al. 2017).

Even though the limits of SDMs for global warming impact assessments on complex ecological structures, it has been recognized that species distribution models are theoretically sufficiently appropriate for simpler practical tasks: for example in leading global warming adaptation policies that include habitat restoration or species selection for reforestation or forest management (Gray & Hamann, 2011, Gray & Hamann, 2013, Hamann & Aitken, 2013, Schelhaas et al. 2015). For such management treatments, the key is to match the source and the ambient target. However, it is uncertain whether subsequent long-term forest developments are correctly described by species distribution models that can be used to influence early decisions on species selection for a geographic area (Maaten et al. 2017).

SDM typically consists of the following process: (1) compilation of the sites of occurrence of species; (2) collection of environmental variables from databases (pluviometry, soil composition, etc.) for the registered location; (3) regression algorithms to understand the connection between sites of presence or species abundance and the environmental variables collected in (2); (4) prediction of the outcome variable (occurrence or species richness) through the space/time of interest, based on the models in (3) (Hijmans & Elith, 2013).

Some of the newest SDMs only use the presences of the groups in the modeling process. Other approaches use presence/absence data or pseudo-absences. Logistic regression is the most common approach to studying presence/absence data (Hijmans & Elith, 2013). Currently the statistical knowledge of applied researchers is growing, and new approaches can handle bigger, more complex datasets, so that applied statisticians are faced with the necessity to specify sophisticated statistical approaches. Logically, as the difficulty of these models grows, it becomes more difficult to perform inferences. The Bayesian approach is mostly suitable as it is flexible and can deal with complex models, for instance, naturally accounting for a hierarchical structure, which could describe the data well, or deal with missing data imputation. Unquestionably, the most popular family of approximate inference methods in Bayesian statistics is the class of Markov Chain Monte Carlo (MCMC) approaches. These approaches, which exploded into popularity in the mid-1980s, have continued at the vanguard of Bayesian statistics ever since, with the basic structure being expanded to cope with progressively more difficult problems (Simpson et al. 2011).

The modeling patterns of the presence/absence of species using local ecological variables has been a rising problem in the field of ecology over the last few years (Chakraborty et al. 2010). This type of modeling has been highly used to address numerous questions, as the identification of essential wildlife habitats with the purpose of classifying and managing conservation regions (Pressey et al. 2007), and predicting the reaction of species to environmental structures (Midgley & Thuiller, 2007, Loarie et al. 2008). Several methodologies and approaches have been presented in this perspective (see for instance (Guisan & Thuiller, 2005, Hijmans & Graham, 2006, Wisz et al. 2008, Rivera, López-Quílez, 2017), with generalized linear models and additive models (GLM and GAM) (Guisan et al. 2002), species envelope models such as BIOCLIM (Busby, 1991), and the multivariate adaptive regression splines (MARS) (Leathwick et al. 2005) being some of the most commonly used models (Munoz et al. 2013).

Most of these approaches consist of regression models to assess the role of environmental factors (e.g., precipitation, bathymetry, etc.) in explaining the species presence (Guisan et al. 2002). However, some difficulties appear: for example, spatial autocorrelation must be taken into account, even if the data were captured through a consistent sampling scheme, as the observations are often adjacent and exposed to similar environmental characteristics (Underwood, 1981, Hurlbert, 1984). Furthermore observer error (Royle et al. 2007, Cressie et al. 2009), gaps in the sampling, missing data, and the mobility of the species (Gelfan et al. 2010) can also influence the models.

Even though traditionally climatic variables have been believed the principal factor in the spatial distribution of the European forest species (Svenning et al. 2008), several paleobotanic studies have shown that the Iberian forest structure has been influenced by the activity of the first agricultural societies from the Neolithic and Chalcolithic period (López Sáez et al. 2006a, López Sáez et al. 2006b, López Sáez et al. 2008, Carrión et al. 2007). Hence, more complex analysis of forest ecosystems are required to comprehend how land-cover variations can affect vegetation dynamics and spatial distribution (Matejicek et al. 2011). Unfortunately, some of these activities are not possible for inclusion in our spatio-temporal framework, not only because of the absence of geographical references, but also because of the difference of timestamps (i.e., Forest Inventory is developed each 10 years).

The forest ecosystems in Europe were affected by several disturbances during the last years. One of the most important issues is degradation due to stress caused by anthropogenic processes. Variation in forest circumstances were not connected to a particular issue but rather to a combination of stress factors that intensified one another. In order to understand the evolution of these ecosystems, the pertinent processes need to be approached by spatio-temporal modeling on detailed spatial and temporal scales (Matejicek et al. 2011). Spatio-temporal processes include the development of spatial patterns over time, thus providing a connection between pattern and process in ecological communities, and having a crucial role in understanding the ecosystem processes (Gratzer et al. 2004). Most of the analyses developed in forest communities were motivated by forest growth (i.e., (O'Rourke & Kelly, 2015, Diggle, 2003, Stoyan & Penttinen, 2000, Illian et al. 2008) for general surveys, and (Grabarnik & Särkkä, 2009) for a specific study), while in this paper, we take a slightly different perspective, and our interest is to show that trees communities are dynamic systems that are affected by environmental disturbances, and that these can also cause changes in the species distribution and dispersion in short periods of time.

Hierarchical models can manage complex interactions by specifying parameters varying on several levels via the introduction of random effects. The predicted value of the response is then articulated to be conditional on these random effects (Cosandey-Godin et al. 2014). The benefits of applying hierarchical Bayesian models arises moreso as complexity rises, when, for instance, spatio-temporal change needs to be modeled explicitly (Cressie et al. 2009). The Bayesian structure similarly offers the benefit of supplying the full posterior probability of the set of parameters of interest, so that point estimates and measures of uncertainty can be easily computed, but with the added benefit that any other function of the parameters can be obtained with no additional effort (Wade, 2000, Wintle et al. 2003).

Hierarchical Bayesian models have commonly relied on MCMC simulation techniques, which are challenging from a technical perspective and are computationally intensive, consequently limiting their use. However, a new statistical method is now available, namely integrated nested Laplace approximations (INLA) via the R-INLA package (http://www.r-inla.org) (Cosandey-Godin et al. 2014). The INLA approach and its potent application to handle complex datasets has been introduced to a wider nontechnical researchers (Illian et al. 2013). Differently from MCMC simulations, INLA applies an approximation for inference, and hence prevents the intense computational requests, convergence, and combining issues sometimes faced by MCMC algorithms (Rue & Martino, 2007). It is only implemented for latent Gaussian models, but this includes the class of models that we consider here for species distribution (for example, logistic regression). Moreover, when the interest lies in a continuous spatial phenomenon, for which realizations are obtained at discrete locations, R-INLA can be coupled with the stochastic partial differential equations (SPDE) approach (Lindgren et al. 2011) which performs discretization of the underlying continuous Gaussian field. This is the case of environmental inventories, which are typically characterized by clustered spatial patterns, and at the same time record large regions with absences. Jointly, these statistical approaches and their implementation in R allow researchers to fit intricated spatio-temporal models considerably faster and more reliably (Rue et al. 2009), due to the characteristics of this approach.

The aim of this paper is to build spatial and spatio-temporal models to predict the distribution of four different species present in the Spanish Forest Inventory. We want to compare the different models and show how accounting for dependencies in space and time affect the relationship between species and environmental variables. We will work with real data on four species of trees obtained from forest inventories developed in Galicia as part of the National Inventories 1970–2010, supported with environmental variables. In particular we consider the II (1980's), III (1990's), and IV (2000's) inventories. For these species we have their presence/absence at specific geographical coordinates, and we generate SDMs for each species. We specify a Bayesian hierarchical geostatistical modeling structure accounting for the spatial dependency.

Other studies have been developed to understand species distribution through a Bayesian approach using INLA, i.e., (Beguin et al. 2012) analyzing the spatial distribution of caribou, or (Dutra et al. 2017) analyzing the presence of invasive species and shrubs in Azores. One of the most interesting differences for our case is the temporal approach, as we work with data from three different inventories.

2. Materials and Methods

Our main data source was the Spanish National Forest Inventory (NFI) dataset, which comprises a systematic grid with 91,889 plots, each of which is 0.2 ha in size, collecting data every 10 years. In our case, we worked only with the Galician dataset, which has had three completed Forest Inventories since 1970. The following tree species are present in the different National Inventories: *Pinus sylvestris* L., *P. uncinate* Ram., *P. pinea* L., *P. halepensis* Mill., *P. nigra* Arn., *P. pinaster* Ait., *P. canariensis* C. Sm., *P. radiata* D. Don, *Abies alba* Mill., *Quercus robur* L./*Q. petraea* (Matt.) Liebl, *Q. pyrenaica* Chips./*Q. pubescens* Willd./*Q. humilis* Mill., *Q. faginea* Lam./*Q. canariensis* Willd., *Q. ilex* L., *Q. suber* L., *Alnus glutinosa* (L.) Gaertn., *Fraxinus* spp., *Populus nigra* L./*P. x Canadensis* Moench, *Eucalyptus globulus* Labill., *E. camaldulensis* Dehnh, *Olea europaea* L., *Ceratonia siliqua* L., *Castanea sativa* Mill., *Betula* spp., *Myrica faya* Ait./*Erica arborea* L., *Fagus sylvatica* L., and *Juniperus* spp. The data provided by the National Inventory included the presence/absence of species.

We chose the following four species from the Spanish National Inventory according to their characteristics, usage, and distribution in the Spanish Peninsula, and more specifically in Galicia:

2.1. Abies alba Mill.

Silver fir (*A. alba*) is a huge evergreen tree located in central Europe, and in some parts of southern and eastern Europe. It is one of the largest tree species of the genus *Abies* in Europe. This species is considered to be a significant ecological and efficient balancer of European forests, and an essential species for preserving high biodiversity in forest ecosystems. Its future distribution is subject to debate between palaeoecologists and modelers, with contrasting climate-response forecasts (San-Miguel-Ayanz et al. 2016).

2.2. *Castanea sativa* Mill.

The sweet chestnut is the single natural species of the genus in Europe. Extensive dispersion and active management caused the establishment of the species at the boundaries of its prospective ecological range. For this reason, it is difficult to trace its original natural area. In Europe, chestnut forests are mainly concentrated in a few countries such as Italy, France, Spain, and Portugal. This species has an extraordinary multipurpose nature, and can be managed for timber production, for chestnut production, and also for a broad range of secondary products and ecosystem services (San-Miguel-Ayanz et al. 2016).

2.3. *Pinus pinaster* Ait.

The maritime pine is a widespread medium-size tree native to the western Mediterranean basin. This pine dwells well in temperate-warm locations, from coasts to high mountains. It does not tolerate shade. Due to its undemanding behavior, salt spray tolerance, and fast growth, it has been used for soil protection, reforestation of degraded areas, and dune stabilization as shelterbelts and also in intensive plantations. The maritime pine has been also traditionally utilized for the extraction of resin for turpentine and rosin. In the Southern Hemisphere, where maritime pine has been introduced for environmental and economical purposes, it has been considered as a highly invasive species (San-Miguel-Ayanz et al. 2016).

2.4. *Quercus robur* L.

Pedunculate oak is a common deciduous tree species in Europe, found from the north (Scandinavia) to the southwest (Spain and Portugal). This genus has cultural importance for people through Europe, and the trees or leaves are commonly used in national or regional emblems. This genus can live several centuries and grow to about 40 m in height. The wood from oaks is strong and robust, and has been valued for centuries. It is preferred for structures, and also for barrels (to contain wine and spirits); overall, it was a main source of ship timbers. Currently, acute oak decline is one of the biggest concerns faced by this genus (San-Miguel-Ayanz et al. 2016).

2.5. Environmental Variables

We have used the following variables to elaborate our models: mean annual temperature, mean of the maximum temperatures of the warmest month, mean of the minimum temperatures of the coldest month, and mean annual rainfall, calcareous soil and elevation.

The environmental variables used in this analysis were obtained from [Herrera et al. 2012, Herrera et al. 2016), and are those typically considered in this type of studies: mean annual temperature, mean of the maximum temperatures of the warmest month, the mean of minimum temperatures of the coldest month, and mean annual rainfall. We also considered the distribution of the calcareous parent materials as a useful predictor of plant species distribution in our study area (Gastón et al. 2009). We used the European Soil Database (Van Liedekerke et al. 2006) to assign each plot to a parent material class. All of the values were related to the data points. In each data point, we have obtained all the environmental variables apart from the presence/absence of the species and coordinates (X, Y, Z).

We can see the summary of the different meteorological variables (mean annual temperature, mean of maximum temperatures of the warmest month, mean of minimum temperatures of the coldest month, and mean annual rainfall) below (Figure 1). As we can see, there are no large fluctuations between the three inventories.

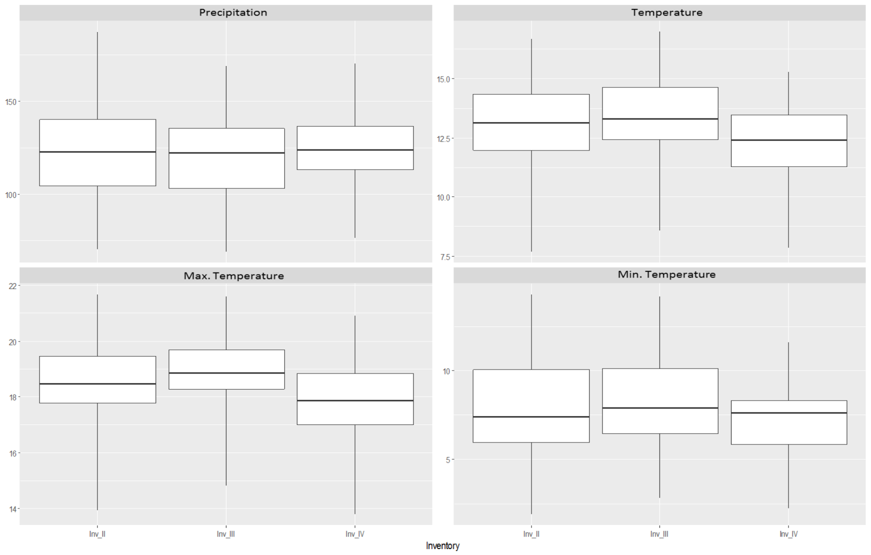


Figure 1. Boxplot summary of meteorological variables during the different inventories. From top to bottom: mean annual rainfall (Ppr) in mm; mean annual temperature (Tas) in °C, mean of the maximum temperatures of the warmest month (tas MAX) in °C, mean of minimum temperatures of the coldest month (tas MIN) in °C.

On the other hand, the presence of the different species varied among the different inventories: three of the species analyzed showed a decrease in the presence in the III inventory, while *A. alba* was almost constant across time (Table 1).

Table 1. Percentage of presences of the different species in Galicia during the three inventories.

Inventory	*Abies alba* **Mill.**	*Castanea sativa* **Mill.**	*Pinus pinaster* **Ait.**	*Quercus robur* **L.**
II (1980s)	1%	26%	51%	51%
III (1990s)	2%	15%	38%	40%
IV (2000s)	1%	35%	46%	59%

2.6. Spatial Model

Spatial data are described as realizations of a stochastic process indexed by space:

$$Y(s) \equiv \{y(s), s \epsilon D\}, \tag{1}$$

where D is a (fixed) subset of R^d (here we consider d = 2). The actual data can be then represented by a collection of observations y = {y(s$_1$), ..., y(s$_n$)}, where the set (s$_1$, ..., s$_n$) indicates the spatial units where the measurements are taken. Depending on D being a continuous surface or a countable collection of d-dimensional spatial units, the problem can be specified as a spatially continuous or discrete random process, respectively (Gelfand et al. 2006). In our case, we can consider a collection of data points with their presence/absence obtained from the inventory, and the sampled points being the set (s$_1$, ..., s$_n$) of n points; y_s is the presence of each species in each point, and it is specified as:

$$y_s \sim \text{Bernoulli}(\pi_s) \tag{2}$$

where π_s is the probability of the species being present.

Then, on the logit(π_s) a linear model is specified including the different covariates, x_{ms} (Temperatures, precipitation, soil and elevation) and a spatial field ξ_s:

$$\mathbf{logit}\,(\boldsymbol{\pi}_s) = \sum_{(\mathbf{m=1})}^{M} \boldsymbol{\beta}_m\, \mathbf{x_{ms}} + \boldsymbol{\xi}_s, \tag{3}$$

where *M* is the number of parameters, and a discretely indexed spatial random process (see Lindgren et al. 2011) is included to approximate the continuous process:

$$\boldsymbol{\xi}_s = \sum_{(\mathbf{g}=\mathbf{1})}^{G} \boldsymbol{\varphi}_{\mathbf{g}}\,(\mathbf{s})\boldsymbol{\xi}_{\mathbf{g}}, \tag{4}$$

where *G* is the total number of vertices of the triangulation.

In practice, the discretization is done dividing the study region in triangles, and writing ξ_s as a linear combination of basis functions φ_g weighted by some zero means terms ξ_g (for more details see Blangiardo & Cameletti, 2015).

The vector $\tilde{\boldsymbol{\xi}} = \{\xi_1, ..., \xi_G\}$ can be modeled as a Gaussian Markov Random Field with a structured covariance function of the distance.

2.7. Spatio-Temporal Model

The concept of the spatial process can be extended to the spatio-temporal case, including a time dimension. The data are then defined by a process:

$$Y(s,t) \equiv \{y(s,t), (s,t) \epsilon D \subset R^2 \times R \}, \tag{5}$$

As we define in the spatial model, we can consider a collection of data points with presence/absence obtained from the inventory and the sampled points are the set $(s_1, ..., s_n)$ of n points; y_{st} is the species presence at each point in space and time, specified as:

$$y_{st} \sim \text{Bernoulli}(\pi_{st}), \tag{6}$$

where π_{st} is the probability of the species being present.

Then, on the logit(π_{st}) a linear model is specified including the different covariates, x_{ms} (Temperatures, precipitation, soil and elevation) and a spatio-temporal field ω_{st}:

$$\textbf{logit}\;(\boldsymbol{\pi}_{st}) = \sum_{(\mathbf{m}=\mathbf{1})}^{M} \boldsymbol{\beta}_m\, \mathbf{x}_{\mathbf{ms}} + \boldsymbol{\omega}_{\mathbf{st}}, \tag{7}$$

where ω_{st} refers to the latent spatio-temporal process that changes in time with autoregressive dynamics and spatial correlation innovations, which we model as follows:

$$\omega_{st} = a\omega_{s(t-1)} + \xi_{st}, \tag{8}$$

with t = 2, …T, $|a| < 1$ and ω_{s1}~Normal $(0, \sigma^2/(1 - a^2))$. ξ_{st}, is a zero-mean Gaussian field that is temporally independent with the following spatio-temporal covariance:

$$\mathrm{Cov}(\xi_{st},\xi_{ju}) = \{(0, t \neq u; \mathrm{Cov}(\xi_s,\xi_j), t = u), \tag{9}$$

for $s \neq j$, where $\mathrm{Cov}(\xi_s,\xi_j)$ is modeled through the Matern spatial covariance function (Lindgren et al. 2011).

2.8. Implementation

We have used the Integrated Nested Laplace Approximation (INLA) implemented in R-INLA within the R statistical software.

The R-INLA package solves models using INLA, which is an approach to statistical inference for latent Gaussian Markov random field (GMRF). The approximation is divided in three stages. The first stage approximates the posterior marginal of θ using the Laplace approximation. The second stage calculates the Laplace approximation, or the simplified Laplace approximation, of $\pi(x_i|y,\theta)$, for selected values of θ, in order to improve on the Gaussian approximation. The third process combines the previous two using numerical integration (Rue et al. 2009).

In R-INLA, the first step needed to process the geostatistical spatial model through SPDE, is the triangulation of the spatial domain of the study. We have used inla.mesh.create providing the spatial coordinates used for estimation. This function executes a constrained refined Delaunay triangulation for a set of spatial locations: firstly the vertices of the triangles are placed at the observation coordinates, and then additional vertices are added, in order to satisfy triangulation quality constraints (Lindgren et al. 2011). Depending on the values selected for the arguments of the function, the total number of vertices changes, with a trade-off between the accuracy of the spatial field representation and the computational and time costs.

Given the mesh, we create the SPDE model object, to be used later in the specification of the final expression in our case, the different spatial and spatio-temporal models.

We consider now the triangulation of National Inventories using the inla.mesh.create(.)function. The function subdivides the region in the triangles, placing the initial vertices at the 2000 station locations and adding 1328 additional vertices (see Figure 2).

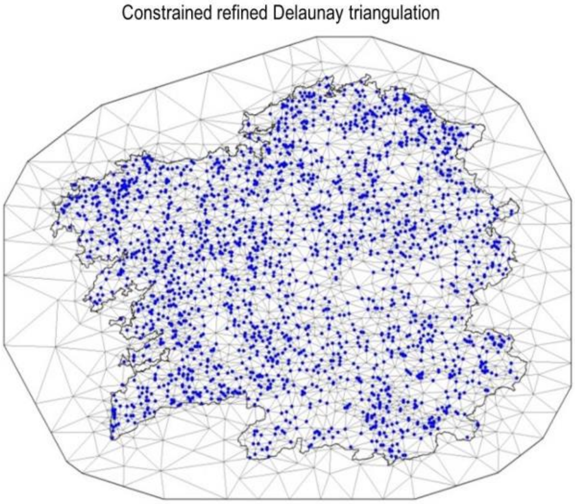


Figure 2. The National Inventory region triangulation with the extended boundary. Blue dots are the monitoring stations and the black bold line represents the region border.

3. Results

In this section we show how each of the four species previously described has evolved in different ways during the last 30 years. Note that we present the maps of the posterior mean of the spatial field from the spatial model, as this model represents better the evolution of the different species.

3.1. Abies alba

As we can see in Figure 3, *A. alba* started being located in the western and southwestern area of Galicia. As time passes, it moves to the northern part and expands in the later years to occupy the western and northern part of the region. This pattern could be explained by the fact that this species does not tolerate high temperatures, and the southeast of Galicia is characterized by high temperatures during the summer, so the introduction of this species in this area should be avoided.

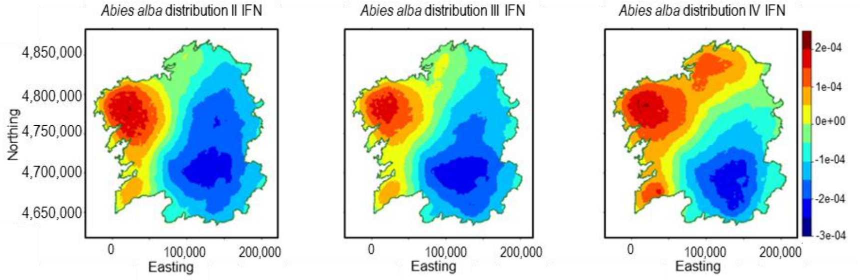


Figure 3. Maps of the posterior mean of the spatial field from the spatial model of *A. alba*, II, III, and IV forest inventories (coordinates in m.).

As we can see in Figure 4, *A. alba* did seem to be affected by the environmental variables in a different way, depending on the time (inventory) and model (spatial vs spatio-temporal). In the first instance, we could see that all the variables, except the soil characteristics, have the same behavior in all the different models, with small negative point estimates, but with credibility intervals excluding zero. On the other hand, Soil showed a larger point estimate, but with an interval including zero, which was narrower in the spatio-temporal model.

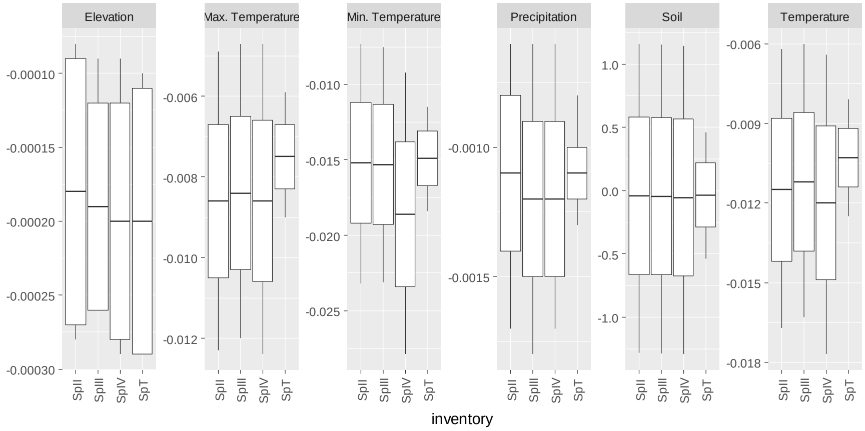


Figure 4. Boxplot diagram for the posterior estimates of the covariates for *A. alba* models: Spatial model II inventory (SpII); Spatial model III inventory (SpIII); Spatial model IV inventory (SpIV); Spatio-temporal model (SpT).

3.2. Castanea sativa

This species is less clustered than the previous one, being mostly present in the central and northern parts of the region at the beginning of the period considered; as time passes, its presence becomes more pronounced (III inventory), but it becomes scattered across the whole Galicia during the last inventory (see Figure 5).

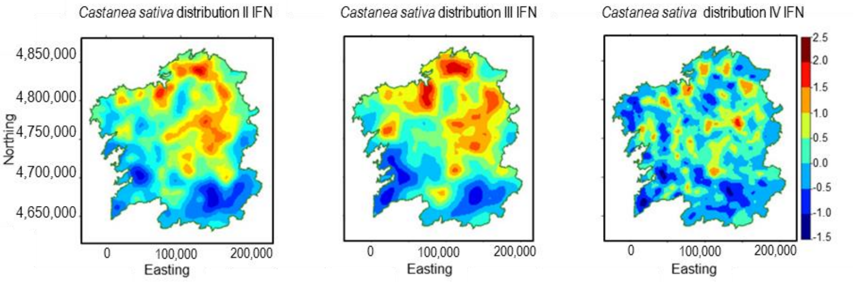

Castanea sativa distribution II IFN
Castanea sativa distribution III IFN
Castanea sativa distribution IV IFN
4,850,000
4,800,000
4,750,000
4,700,000
4,650,000
Northing
0
100,000
200,000
Easting
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5

Figure 5. Maps of the posterior means of the spatial field from the spatial model of *C. sativa*, II, III, and IV forest inventories (coordinates in m.).

This species showed a different relationship with the environmental variables if we analyzed the different models (see Figure 6 below). In the spatio-temporal model, the average annual temperature was the most important variable, followed by the temperatures of the coldest and warmest months. Also, in the spatial models, the type of substrate was an important variable followed by the temperatures. This could be explained by the adaptability of this species to the substrate: it generally preferred siliceous substrate, but it could be present also in certain calcareous soils if there were optimum conditions. Also, in this case, we could see different variables behavior between spatial and spatio-temporal models. Looking at the Elevation, in the II and III inventories, all the values in the credible interval were positives, but then in the IV inventory and in the spatio-temporal model, there were negative and positive values. If we compared the credible intervals for the remaining variables (Soil, Precipitation, Temperature, Maximum emperature, and Minimum Temperature), we can see similar behaviors in the spatial models with negative and positive values. Moreover, the spatio-temporal model showed that Soil has a similar performance than the spatial models, while the other variables show differences. Precipitation and Temperature show positive values in the credible interval, with values close to zero in Precipitation; Max and Min Temperature showed negative values in this interval. These results suggest that this species is affected positively by the temperature, but extreme temperatures can also affect its presence.

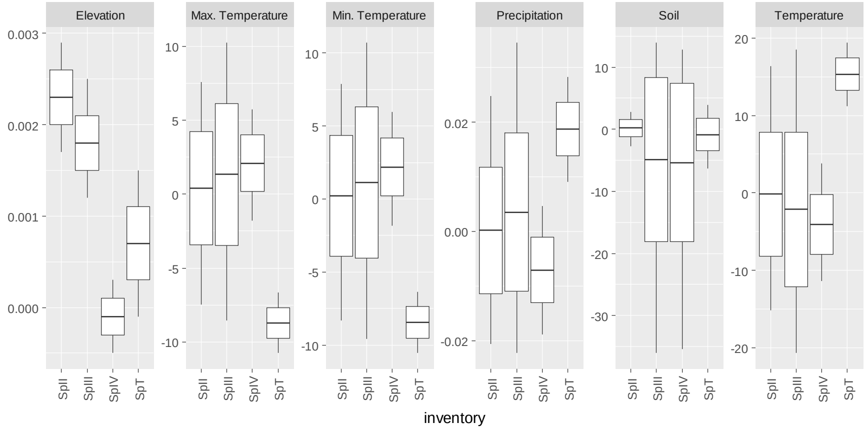


Figure 6. Boxplot diagram for posterior estimates of the covariates for *C. sativa* models: Spatial model II inventory (SpII); Spatial model III inventory (SpIII); Spatial model IV inventory (SpIV); Spatio-temporal model (SpT).

3.3. Pinus Pinaster

This species shows similarities with *C. sativa*: it starts with a low presence, then it becomes present in the whole area of analysis, but in this case, the posterior mean in inventory IV shows that this species becomes concentrated in the southern area, with low presences in the interior (see Figure 7).

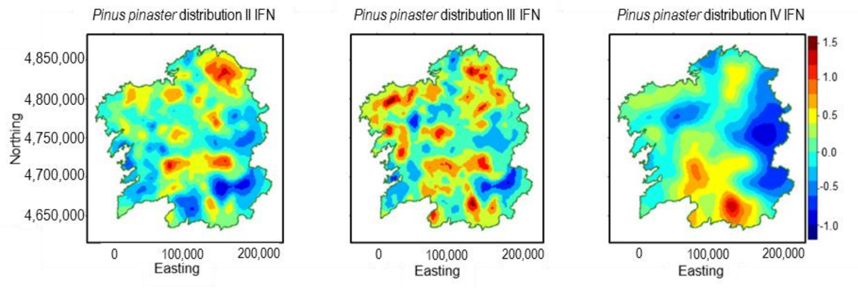

Pinus pinaster distribution II IFN
Pinus pinaster distribution III IFN
Pinus pinaster distribution IV IFN
Northing
4,850,000
4,800,000
4,750,000
4,700,000
4,650,000
0
100,000
200,000
Easting
1.5
1.0
0.5
0.0
-0.5
-1.0

Figure 7. Maps of the posterior mean of the spatial field from the spatial model of *P. pinaster*, II, III, and IV forest inventories (coordinates in m.).

Moreover, as we can see in Figure 8 below, not only temperatures, but also soil characteristics (calcareous) are very important in the different models for this species; we could see interesting differences between the spatial and the spatio-temporal models: in the latter, Temperature (average) positively affected the presence of the species, and extreme values (Maximum and Minimum), participated negatively. On the other hand, in spatial models, the most important variable was that the type of soil, followed by Temperatures in III and IV inventory. This species had a similar behavior to the previous one. Elevation and Soil had similar credible intervals in spatial and spatio-temporal models, while there are some differences in the other variables. Looking at Precipitation, in spatial models, the credible interval had positive and negative values, but only positive values were represented in the spatio-temporal model. Finally, the Temperature variables (average, maximum and minimum), had positive and negative values in the credible interval for the II and III inventories, but in the IV inventory and spatio-temporal model, all the values were positive in the average, and negatives in Max. and Min.

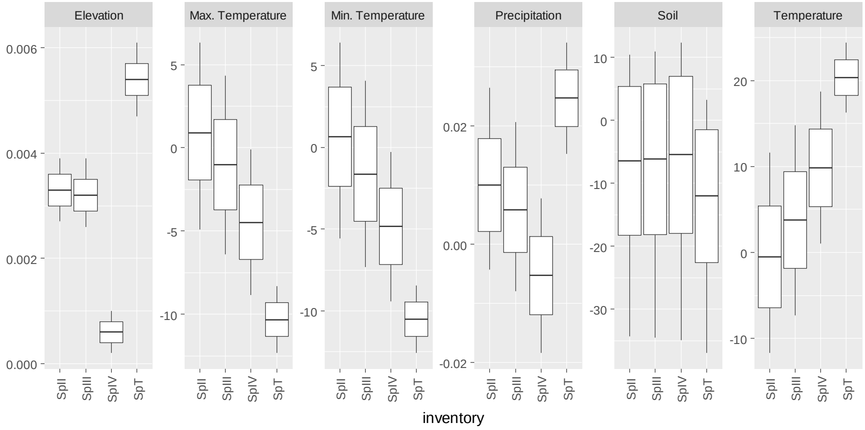


Figure 8. Boxplot diagram for the posterior estimates of the covariates for *P. pinaster* models: Spatial model II inventory (SpII); Spatial model III inventory (SpIII); Spatial model IV inventory (SpIV); Spatio-temporal model (SpT).

3.4. Quercus robur

This species showed an increasing presence in Galicia. During the third inventory, it was present in almost all of the areas, except the southeastern area (see Figure 9).

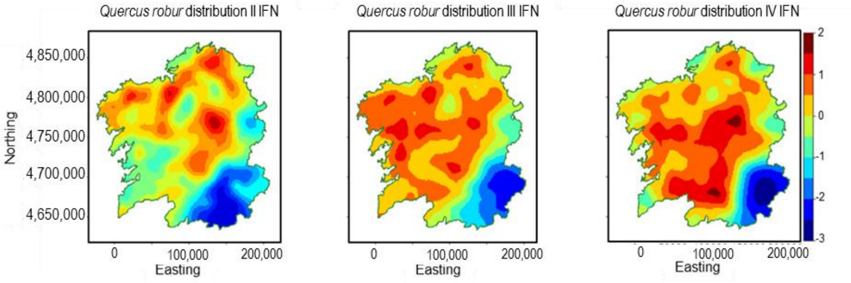

Quercus robur distribution II IFN
Quercus robur distribution III IFN
Quercus robur distribution IV IFN
Northing
4,850,000
4,800,000
4,750,000
4,700,000
4,650,000
0
100,000
200,000
Easting
2
1
0
-1
-2
-3

Figure 9. Maps of the posterior mean of the spatial field from the spatial model of *Quercus robur*, II, III, and IV forest inventories (coordinates in m.).

As can see in Figure 10, the variables in *Q. robur* have a similar behavior to *C. sativa*, except in the IV inventory model, where the calcareous soil had a very important part in the species distribution. As we have seen in most of the other species, Elevation had similar credible intervals in all the different models, in this case, always showing positive values that were close to zero. As we have seen in *C. sativa*, the rest of the variables had similar performances in the spatial models and different performances in the spatio-temporal model (except the Max. Temperature, with similar credible intervals in all the models). All the variables had 95% credible intervals, including zeros in the spatial models, while the spatio-temporal model showed positive values in Min. Temperature and negative values in the Soil, Precipitation, and Temperature.

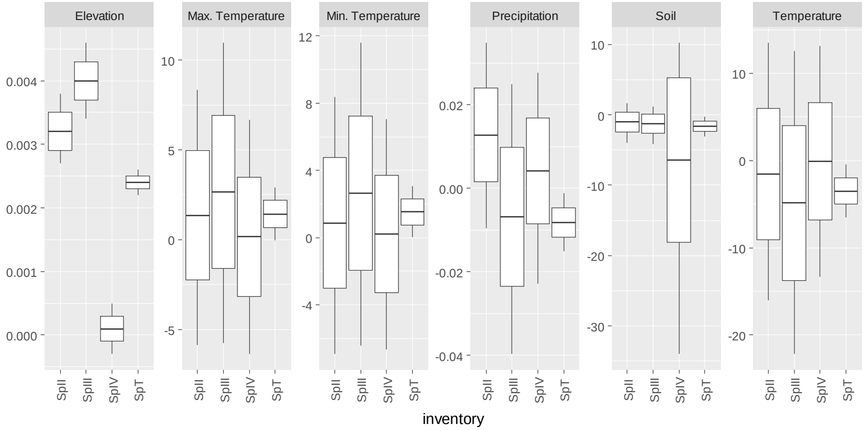


Figure 10. Boxplot diagram for posterior estimates of the covariates for *Q. robur* models: Spatial model II inventory (SpII); Spatial model III inventory (SpIII); Spatial model IV inventory (SpIV); Spatio-temporal model (SpT).

3.5. Summary

As we can see in Table 2, most of the species show different relationships with environmental and climatic variables between the spatial and spatio-temporal models. The positive symbol (+) summarizes a positive relationship between the variable and the presence; the negative symbol (−) represents the opposite; Rn does not show a clear relationship with a credible interval with the positives and negatives values. Also, if we generalize, species with more presences (see Table 1) show larger differences between models. Also, if we analyse the results from spatial to spatio-temporal models, typically variables not showing a clear effect become positively or negatively associated with presence, depending on the species and the variable.

Table 2. Posterior estimates summary. Comparison between variables and presence of different species in spatial (Sp) and spatio-temporal (Sp-T) models (+) represents a positive relationship, (−) a negative relationship, and (Rn) not a clear relationship.

	Abies alba		*Castanea sativa*		*Pinus pinaster*		*Quercus robur*	
	Sp	Sp-T	Sp	Sp-T	Sp	Sp-T	Sp	Sp-T
Elevation	−	−	Rn	Rn	+	+	+	+
Soil	Rn	Rn	Rn	Rn	Rn	Rn	Rn	−
Precipitation	−	−	Rn	Rn	Rn	+	Rn	−
Temperature	−	−	Rn	Rn	Rn	+	Rn	−
Max. Temperature	−	−	Rn	−	Rn	−	Rn	Rn
Min. Temperature	−	−	Rn	−	Rn	−	Rn	−

An usual way to estimate out-of-sample prediction error is cross-validation (see (Geisser & Eddy, 1979, Vehtari & Lampinen, 2002)) for a Bayesian approach), but scientists have always looked for alternative methods, as cross-validation involves repeated model fits and it can run into trouble with sparse data (Gelman at al. 2013). When the aim is model comparison, the most common index is the DIC (Spiegelhalter et al. 2003, Van del Linde, 2005), which, in the same way to the Akaike information criterion AIC involves two components, a term that measures the goodness of fit, and a penalty term for growing model complexity. More recently, the Watanabe-Akaike information criterion WAIC (Watanabe, 2010) has been suggested as an appropriate alternative for estimating the out-of-sample expectation in a fully Bayesian approach. This method starts with the calculated log pointwise posterior predictive density, and then adds a correction for the effective number of parameters to adjust for overfitting (Gelman at al. 2013). WAIC works on predictive probability density of observed variables rather than on model parameter; hence, it can be applied in singular statistical models (i.e., models with non-identifiable parameterization, see (Li et al. 2016).

We have also considered the conditional predictive ordinate (CPO) [74] to perform model evaluation. The conditional predictive ordinate (CPO) is established on leave-one-out cross-validation. CPO estimates the probability of observing a value, after having already observed the others. The mean logarithmic score (LCPO) was calculated as a measure of the predictive quality of the model (Gneiting & Raftery, 2007, Roos & Held, 2011). High LCPO values indicate possible outliers, high-leverage, and influential observations.

In Table 3, we can see the summary of the WAIC and LCPO values obtained in the different models for each species (spatial and spatio-temporal models); this shows that, looking at the WAIC, most of the species have a better fit for the spatio-temporal model with vegetation (except *C. sativa*), also looking at LCPO spatial model has more outliers than spatio-temporal models.

Table 3. Watanabe-Akaike information criterion **(**WAIC) and logarithmic score of conditional predictive ordinate (LCPO) comparison in spatial (Sp) and spatio-temporal (Sp-T) models.

	Abies alba		*Castanea sativa*		*Pinus pinaster*		*Quercus robur*	
	Sp	Sp-T	Sp	Sp-T	Sp	Sp-T	Sp	Sp-T
WAIC	15.54	12.73	5.435	10.629	4.621	3.385	9.676	1.837
LCPO	1.531	1.251	2.327	2.986	3.725	2.327	2.382	1.965

Summarizing the computational costs of performing the different models, all the models were executed from the same terminal (laptop Core i7 with 12 GB RAM). Spatial models need between 10 and 30 min to obtain the results. However, spatio-temporal models need between 5 and 12 hr to finish the process.

4. Conclusions

We have built spatial and spatio-temporal models to predict the distributions of four different species present in the Spanish Forest Inventory. We have compared the different models and show the relationship between species and environmental variables. We have shown that this relationship changes between spatial and spatio-temporal models. Most of the spatial models show a vague relationship with the environmental variables, which becomes more clear when we analyzed all of the time series when developing the spatio-temporal model. Also, we have shown how the species evolve in space along time, changing their distributions between the II to the IV inventory.

Initially our aim in this project was to apply these models to the whole of Spain, assuming that the spatial continuity is essential to understand species distribution. However due to technical issues, we were not able to finalize this. Currently not all the data from the last inventory are available for all the provinces, and also some of the areas have different reference systems to locate the parcels. Another problem was the computational cost; the available resources were not powerful enough to work with this data volume (more than 90,000 points per inventory).

There are interesting differences between spatial and spatio-temporal models for the different species. As we have shown, not all the same variables have the same weight in the different models.

Several factors can affect spatial distribution of species. Environmental factors are not the only variables that can affect this distribution, but socioeconomic factors, policies, and management criteria can also be important agents that have different impacts in the species presence.

Analysing the models, we can affirm that the use of spatio-temporal models is an advantage for the understanding of the different ecological dynamics, given the the temporal perspective is not very frequent in environmental research projects.

We have analyzed the credible interval of the different variables in order to understand the relationship between the environmental variables and species presence. We can see that some variables change their “weight” depending of the inventory, and several variables also have the same behavior in all the inventories, and also along the spatio-temporal model.

Summarizing, we can generalize that permanent and theoretical inalterable variables have similar performances in spatial and spatio-temporal models, showing a similar relationship between the presence of species and these variables along time. Moreover, species presence does not always have a similar relationship with “non static” variables. This relationship is changing, not only due to changes in environmental factors, but also based on species management and possible human disturbances.

Spatio-temporal models and the R-INLA package appear to offer additional benefits beyond the common SDM or spatially-explicit modeling. The combination of using a complex spatial latent field to capture spatial processes and an underlying simple additive regression model for the response variables relationship to environmental factors, means that the fixed effects are potentially more straightforward to interpret (Goldin & Purse, 2016). Another benefit of a Bayesian approach is the capture of uncertainty for each predicted value, with predictive uncertainty being an often ignored aspect of SDM modeling and prediction. R-INLA models are extremely flexible in their specifications, with spatial autocorrelation and observer bias being straightforwardly incorporated as random effects, while standard error distributions, such as Gaussian, Poisson, binomial, and a variety of zero-inflated models, can be used interchangeably (Rue et al. 2009). This method, therefore, has a built-in potential for extending SDM analysis away from simple binomial models by, for example, incorporating two or more types of data (Warton et al. 2015), hierarchical seasonal models (Redding et al. 2016), or fitting point-process models (Renner & Warton, 2013). We hope that our research will aid in the uptake of such fast spatial Bayesian methods, as this approach shows great promise for other analyses in ecology.

Author Contributions: Conceptualization, O.R.d.R., A.L-Q. and M.B.; Methodology, O.R.d.R., A.L-Q and M.B.; Formal Analysis, O.R.d.R.; Investigation, O.R.d.R.; Data Curation, O.R.d.R.; Writing-Original Draft Preparation, O.R.d.R.; Writing-Review & Editing, A.L-Q. and M.B.; Supervision, A.L-Q. and M.B.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Adams, H.D.; Macalady, A.K.; Breshears, D.D.; Allen, C.D.; Stephenson, N.L.; Saleska, S.R.; Huxman, T.E.; McDowell, N.G. Climate-induced tree mortality: Earth system consequences. Eos 2010, 91, 153–154.
2. Beguin, J.; Martino, S.; Rue, H.; Cumming, S.G. Hierarchical analysis of spatially autocorrelated ecological data using integrated nested Laplace approximation. Methods Ecol. Evol. 2012, 3, 921–929.
3. Blangiardo, M.; Cameletti, M. Spatial and Spatio-Temporal Bayesian Models with R-INLA; John Wiley & Sons: Hoboken, NJ, USA, 2015.
4. Boisvenue, C.; Running, S.W. Impacts of climate change on natural forest productivity–evidence since the middle of the 20th century. Glob. Chang. Biol. 2006, 12, 862–882.
5. Bonan, G.B. Forests and climate change: Forcings, feedbacks, and the climate benefits of forests. Science 2008, 320, 1444–1449.
6. Busby, J. BIOCLIM—A bioclimate analysis and prediction system. Plant Prot. Q. (Aust.) 1991, 6, 64–68.
7. Carrión, J.S.; Fuentes, N.; González-Sampériz, P.; Quirante, L.S.; Finlayson, J.C.; Fernández, S.; Andrade, A. Holocene environmental change in a montane region of southern Europe with a long history of human settlement. Quat. Sci. Rev. 2007, 26, 1455–1475.
8. Chakraborty, A.; Gelfand, A.E.; Wilson, A.M.; Latimer, A.M.; Silander, J.A., Jr. Modeling large scale species abundance with latent spatial processes. Ann. Appl. Stat. 2010, 4, 1403–1429.
9. Cosandey-Godin, A.; Krainski, E.T.; Worm, B.; Flemming, J.M. Applying Bayesian spatiotemporal models to fisheries bycatch in the Canadian Arctic. Can. J. Fish. Aquat. Sci. 2014, 72, 186–197.
10. Cressie, N.; Calder, C.A.; Clark, J.S.; Hoef, J.M.V.; Wikle, C.K. Accounting for uncertainty in ecological analysis: The strengths and limitations of hierarchical statistical modeling. Ecol. Appl. 2009, 19, 553–570.
11. Davis, M.B.; Shaw, R.G. Range shifts and adaptive responses to Quaternary climate change. Science 2001, 292, 673–679.

12. Diggle, P.J. Statistical Analysis of Spatial Point Patterns; Arnold: London, UK, 2003.
13. Dutra Silva, L.; Brito de Azevedo, E.; Bento Elias, R.; Silva, L. Species Distribution Modeling: Comparison of Fixed and Mixed Effects Models Using INLA. ISPRS Int. J. Geo-Inf. 2017, 6, 391.
14. Elith, J.; Leathwick, J.R. Species distribution models: Ecological explanation and prediction across space and time. Ann. Rev. Ecol. Evol. System. 2009, 40, 677–697.
15. Gastón, A.; Soriano, C.; Gómez-Miguel, V. Lithologic data improve plant species distribution models based on coarse-grained occurrence data. For. Syst. 2009, 18, 42–49.
16. Geisser, S.; Eddy, W.F. A predictive approach to model selection. J. Am. Stat. Assoc. 1979, 74, 153–160.
17. Gelfand, A.E.; Diggle, P.J.; Fuentes, M.; Guttorp, P. (Eds.) Handbook of Spatial Statistics; CRC Press: Boca Raton, FL, USA, 2010.
18. Gelfand, A.E.; Silander, J.A.; Wu, S.; Latimer, A.; Lewis, P.O.; Rebelo, A.G.; Holder, M. Explaining species distribution patterns through hierarchical modeling. Bayesian Anal. 2006, 1, 41–92.
19. Gelman, A.; Shalizi, C.R. Philosophy and the practice of Bayesian statistics. Br. J. Math. Stat. Psychol. 2013, 66, 8–38.
20. Gneiting, T.; Raftery, A.E. Strictly proper scoring rules, prediction, and estimation. J. Am. Stat. Assoc. 2007, 102, 359–378.
21. Golding, N.; Purse, B.V. Fast and flexible Bayesian species distribution modelling using Gaussian processes. Methods Ecol. Evol. 2016, 7, 598–608.
22. Grabarnik, P.; Särkkä, A. Modelling the spatial structure of forest stands by multivariate point processes with hierarchical interactions. Ecol. Model. 2009, 220, 1232–1240.
23. Gratzer, G.; Canham, C.; Dieckmann, U.; Fischer, A.; Iwasa, Y.; Law, R.; Lexer, M.J.; Sandmann, H.; Spies, T.A.; Splechtna, B.E.; et al. Spatio-temporal development of forests–current trends in field methods and models. Oikos 2004, 107, 3–15.

24. Gray, L.K.; Hamann, A. Strategies for reforestation under uncertain future climates: Guidelines for Alberta, Canada. PLoS ONE 2011, 6, e22977.
25. Gray, L.K.; Hamann, A. Tracking suitable habitat for tree populations under climate change in western North America. Clim. Chang. 2013, 117, 289–303.
26. Guisan, A.; Edwards, T.C.; Hastie, T. Generalized linear and generalized additive models in studies of species distributions: Setting the scene. Ecol. Model. 2002, 157, 89–100.
27. Guisan, A.; Thuiller, W. Predicting species distribution: Offering more than simple habitat models. Ecol. Lett. 2005, 8, 993–1009.
28. Hamann, A.; Aitken, S.N. Conservation planning under climate change: Accounting for adaptive potential and migration capacity in species distribution models. Divers. Distrib. 2013, 19, 268–280.
29. Hanewinkel, M.; Cullmann, D.A.; Schelhaas, M.J.; Nabuurs, G.J.; Zimmermann, N.E. Climate change may cause severe loss in the economic value of European forest land. Nat. Clim. Chang. 2013, 3, 203–207.
30. Herrera, S.; Fernández, J.; Gutiérrez, J.M. Update of the Spain02 gridded observational dataset for EURO-CORDEX evaluation: Assessing the effect of the interpolation methodology. Int. J. Climatol. 2016, 36, 900–908.
31. Herrera, S.; Gutiérrez, J.M.; Ancell, R.; Pons, M.R.; Frías, M.D.; Fernández, J. Development and analysis of a 50-year high-resolution daily gridded precipitation dataset over Spain (Spain02). Int. J. Climatol. 2012, 32, 74–85.
32. Hijmans, R.J.; Elith, J. Species Distribution Modelling with R. The R Foundation for Statistical Computing. Available online: http://cran.r-project.org/ web/packages/dismo/vignettes/sdm.pdf (accessed on 05 May 2018).
33. Hijmans, R.J.; Graham, C.H. The ability of climate envelope models to predict the effect of climate change on species distributions. Glob. Chang. Biol. 2006, 12, 2272–2281.
34. Hurlbert, S.H. Pseudoreplication and the design of ecological field experiments. Ecol. Monogr. 1984, 54, 187–211.
35. Illian, J.; Penttinen, A.; Stoyan, H.; Stoyan, D. Statistical Analysis and Modelling of Spatial Point Patterns; John Wiley & Sons: Hoboken, NJ, USA, 2008; Volume 70.

36. Illian, J.B.; Martino, S.; Sørbye, S.H.; Gallego-Fernández, J.B.; Zunzunegui, M.; Esquivias, M.P.; Travis, J.M. Fitting complex ecological point process models with integrated nested Laplace approximation. Methods Ecol. Evol. 2013, 4, 305–315.
37. Leathwick, J.R.; Rowe, D.; Richardson, J.; Elith, J.; Hastie, T. Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. Freshw. Biol. 2005, 50, 2034–2052.
38. Li, L.; Qiu, S.; Zhang, B.; Feng, C.X. Approximating cross-validatory predictive evaluation in Bayesian latent variable models with integrated IS and WAIC. Stat. Comput. 2016, 26, 881–897.
39. Lindgren, F.; Rue, H.; Lindström, J. An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. J. R. Stat. Soc. Ser. B 2011, 73, 423–498.
40. Lindner, M.; Fitzgerald, J.B.; Zimmermann, N.E.; Reyer, C.; Delzon, S.; Maaten, E.; Schelhaas, M.J.; Lasch, P.; Eggers, J.; Maaten-Theunissen, M.; et al. Climate change and European forests: What do we know, what are the uncertainties, and what are the implications for forest management? J. Environ. Manag. 2014, 146, 69–83.
41. Lindner, M.; Maroschek, M.; Netherer, S.; Kremer, A.; Barbati, A.; Garcia-Gonzalo, J.; Seidl, R.; Delzon, S.; Corona, P.; Kolström, M.; et al. Climate change impacts, adaptive capacity, and vulnerability of European forest ecosystems. For. Ecol. Manag. 2010, 259, 698–709.
42. Loarie, S.R.; Carter, B.E.; Hayhoe, K.; McMahon, S.; Moe, R.; Knight, C.A.; Ackerly, D.D. Climate change and the future of California's endemic flora. PLoS ONE 2008, 3, e2502.
43. López Sáez, J.A.; Galop, D.; Iriarte Chiapusso, M.J.; López Merino, L. Paleoambiente y antropización en los Pirineos de Navarra durante el Holoceno medio (VI–IV milenios cal. BC): Una perspectiva palinológica. Veleia 2008, 24–25, 645–653.
44. López Sáez, J.A.; López García, P.; López Merino, L. El impacto humano en la Cordillera Cantábrica: Estudios palinológicos durante el Holoceno medio. Zona Arqueol. 2006, 7, 122–131.

45. López Sáez, J.A.; López García, P.; López Merino, L. La transición Mesolítico-Neolítico en el Valle Medio del Ebro y en el Prepirineo aragonés desde una perspectiva paleoambiental: Dinámica de la antropización y origen de la agricultura. Rev. Iberoam. Hist. 2006, 1, 4–11.
46. Maaten, E.; Hamann, A.; Maaten-Theunissen, M.; Bergsma, A.; Hengeveld, G.; Lammeren, R.; Mohren, F.; Nabuurs, G.J.; Terhürne, R.; Sterck, F. Species distribution models predict temporal but not spatial variation in forest growth. Ecol. Evol. 2017, 7, 2585–2594.
47. Matejicek, L.; Vavrova, E.; Cudlin, P. Spatio-temporal modelling of ground vegetation development in mountain spruce forests. Ecol. Model. 2011, 222, 2584–2592.
48. Midgley, G.F.; Thuiller, W. Potential vulnerability of Namaqualand plant diversity to anthropogenic climate change. J. Arid Environ. 2007, 70, 615–628.
49. Munoz, F.; Pennino, M.G.; Conesa, D.; López-Quílez, A.; Bellido, J.M. Estimation and prediction of the spatial occurrence of fish species using Bayesian latent Gaussian models. Stoch. Environ. Res. Risk Assess. 2013, 27, 1171–1180.
50. O'Neill, G.A.; Hamann, A.; Wang, T. Accounting for population variation improves estimates of the impact of climate change on species' growth and distribution. J. Appl. Ecol. 2008, 45, 1040–1049.
51. O'Rourke, S.; Kelly, G.E. Spatio-temporal modelling of forest growth spanning 50 years—The effects of different thinning strategies. Procedia Environ. Sci. 2015, 26, 101–104.
52. Parmesan, C.; Yohe, G. A globally coherent fingerprint of climate change impacts across natural systems. Nature 2003, 421, 37–42.
53. Pettit, L.I. The conditional predictive ordinate for the normal distribution. J. R. Stat. Soc. Ser. B 1990, 52, 175–184.
54. Pressey, R.L.; Cabeza, M.; Watts, M.E.; Cowling, R.M.; Wilson, K.A. Conservation planning in a changing world. Trends Ecol. Evol. 2007, 22, 583–592.

55. Redding, D.W.; Cunningham, A.A.; Woods, J.; Jones, K.E. Spatial and seasonal predictive models of Rift Valley Fever disease. Philos. Trans. R. Soc. Lond. B 2016, 372, 20160165.
56. Renner, I.W.; Warton, D.I. Equivalence of MAXENT and Poisson Point Process Models for Species Distribution Modeling in Ecology. Biometrics 2013, 69, 274–281.
57. Rivera, O.R.; Blangiardo, M.; López-Quílez, A.; Martín-Sanz, I. Species distribution modelling through Bayesian hierarchical approach. Theor. Ecol. 2018, doi:10.1007/s12080-018-0387-y.
58. Rivera, Ó.R.; López-Quílez, A. Development and Comparison of Species Distribution Models for Forest Inventories. ISPRS Int. J. Geo-Inf. 2017, 6, 176.
59. Roos, M.; Held, L. Sensitivity analysis in Bayesian generalized linear mixed models for binary data. Bayesian Anal. 2011, 6, 259–278.
60. Royle, J.A.; Kéry, M.; Gautier, R.; Schmid, H. Hierarchical spatial models of abundance and occurrence from imperfect survey data. Ecol. Monogr. 2007, 77, 465–481.
61. Rue, H.; Martino, S. Approximate Bayesian inference for hierarchical Gaussian Markov random field models. J. Stat. Plan. Inference 2007, 137, 3177–3192.
62. Rue, H.; Martino, S.; Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J. R. Stat. Soc. Ser. B 2009, 71, 319–392.
63. Running, S.W.; Nemani, R.R.; Heinsch, F.A.; Zhao, M.; Reeves, M.; Hashimoto, H. A continuous satellite-derived measure of global terrestrial primary production. Bioscience 2004, 54, 547–560.
64. San-Miguel-Ayanz, J.; Rigo, D.D.; Caudullo, G.; Houston Durrant, T.; Mauri, A. European Atlas of Forest Tree Species; European Commission, Joint Research Centre: Brussels, Belgium, 2016.
65. Schelhaas, M.J.; Nabuurs, G.J.; Hengeveld, G.; Reyer, C.; Hanewinkel, M.; Zimmermann, N.E.; Cullmann, D. Alternative forest management strategies to account for climate change-induced productivity and species suitability changes in Europe. Reg. Environ. Chang. 2015, 15, 1581–1594.

66. Schröter, D.; Cramer, W.; Leemans, R.; Prentice, I.C.; Araújo, M.B.; Arnell, N.W.; Bondeau, A.; Bugmann, H.; Carter, T.R.; Gracia, C.A.; et al. Ecosystem service supply and vulnerability to global change in Europe. Science 2005, 310, 1333–1337.
67. Simpson, D.; Lindgren, F.; Rue, H. Fast approximate inference with INLA: The past, the present and the future. arXiv 2011, arXiv:1105.2982.
68. Spathelf, P.; van der Maaten, E.; van der Maaten-Theunissen, M.; Campioli, M.; Dobrowolska, D. Climate change impacts in European forests: The expert views of local observers. Ann. For. Sci. 2014, 71, 131–137.
69. Spiegelhalter, D.; Best, N.G.; Carlin, B.P.; Van der Linde, A. Bayesian measures of model complexity and fit. Qual. Control Appl. Stat. 2003, 48, 431–432.
70. Stoyan, D.; Penttinen, A. Recent applications of point process methods in forestry statistics. Stat. Sci. 2000, 15, 61–78.
71. Svenning, J.C.; Normand, S.; Kageyama, M. Glacial refugia of temperate trees in Europe: Insights from species distribution modelling. J. Ecol. 2008, 96, 1117–1127.
72. Underwood, A.J. Techniques of analysis of variance in experimental marine biology and ecology. Oceanography and marine biology: An annual review. Ann. Rev. Oceanogr. Mar. Biol. 1981, 19, 513–605.
73. Van Der Linde, A. DIC in variable selection. Stat. Neerl. 2005, 59, 45–56.
74. Van Liedekerke, M.; Jones, A.; Panagos, P. ESDBv2 Raster Library—A Set of Rasters Derived from the European Soil Database Distribution v2.0; European Commission and the European Soil Bureau Network, CDROM, EUR, 19945; European Commission: Brussels, Belgium, 2006.
75. Vehtari, A.; Lampinen, J. Bayesian model assessment and comparison using cross-validation predictive densities. Neural Comput. 2002, 14, 2439–2468.
76. Wade, P.R. Bayesian methods in conservation biology. Conserv. Biol. 2000, 14, 1308–1316.
77. Warton, D.I.; Blanchet, F.G.; O'Hara, R.B.; Ovaskainen, O.; Taskinen, S.; Walker, S.C.; Hui, F.K. So Many Variables: Joint Modeling in Community Ecology. Trends Ecol. Evol. 2015, 30, 766–779.

78. Watanabe, S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J. Mach. Learn. Res. 2010, 11, 3571–3594.
79. Wintle, B.A.; McCarthy, M.A.; Volinsky, C.T.; Kavanagh, R.P. The use of Bayesian model averaging to better represent uncertainty in ecological models. Conserv. Biol. 2003, 17, 1579–1590.
80. Wisz, M.S.; Hijmans, R.J.; Li, J.; Peterson, A.T.; Graham, C.H.; Guisan, A. Effects of sample size on the performance of species distribution models. Divers. Distrib. 2008, 14, 763–773.Running, S.W.; Nemani, R.R.; Heinsch, F.A.; Zhao, M.; Reeves, M.; Hashimoto, H. A continuous satellite-derived measure of global terrestrial primary production. *Bioscience* **2004**, *54*, 547–560.