UNIVERSITAT DE VALÈNCIA

Programa de Doctorat en Estadística i Optimització

Recent statistical advances and applications of species distribution modeling

by

Joaquín Martínez Minaya

PhD Thesis
in Statistics and Optimization

Supervised by
David Valentín Conesa Guillén,
and Antonio Vicent Civera

in the
Faculty of Mathematics
Department of Statistics and Operations Research

April 2019

This Thesis has been supported by grant ACIF/2016/455 from the Generalitat Valenciana and the European Social Fund (ESF). ESF invests in your future. The last part of this research has been carried out during the two visits to Dr. Finn Lindgren at the University of Edinburgh.

Acknowledgements

Todavía me cuesta hacerme a la idea, pero parece, que tras 5 años, esta etapa llega a su fin ...

En este arduo camino, a veces desesperante pero muchas otras muy gratificante, han sido much@s l@s compañer@s de viaje, personas que me han tendido la mano en momentos difíciles y que sin ellas este proyecto no sería posible. Así, en estas líneas quiero manifestar mi más sincera gratitud a todas ellas.

En primer lugar, me gustaría destacar el papel de mis directores de tesis David y Antonio, sin ellos esta aventura no tendría sentido. Gracias a los dos por darme la oportunidad de emprender este viaje, por escucharme, por ilusionarme, por tener siempre un buen consejo para mi, pero sobre todo, gracias por cuidarme, por creer en mi desde el minuto 0. Gracias a los dos de corazón, porque más allá del ámbito científico, sois grandes maestros de la vida.

Por supuesto, no me puedo olvidar de Antonio L., Xavi, Anabel, María, Iosu, Rufo y Héctor, que siempre han tenido un consejo que compartir y unas palabras de optimismo que ofrecer, además de su inmesurable aporte científico. Mi más sincero agradecimiento a los colaboradores de los artículos que mostramos en esta tesis, en especial a Arnald Marcer y Xavier Picó, dos excelentes profesionales que siempre me han ofrecido unas palabras de ánimo y aliento.

También quiero expresar mi gratitud a todos mis compañer@s tanto del IVIA (Elena, Ana, Pablo, Neus, José, Félix, Daniel, Vidal, Ana (CIDE), Martina, etc.) como de la Universidad (Blanca, Consuelo, Elena, Miguel, Danilo, Jessica, Irene, Paula, Abel, Gabriel, Raúl, Priscila, Carolina, etc.) por alegrar cada uno de mis días. Todos los momentos vividos a vuestro lado han sido únicos e irrepetibles, tod@s y cada uno de vosotr@s sois un ejemplo de valentía y superación. Es difícil expresar mis sentimientos hacia vosotr@s, habéis dejado huella en mi, espero y deseo que nuestras vidas se vuelvan a cruzar. Quiero hacer mención especial a Blanca y a su pareja

Mar, porque su apoyo y consejo han sido claves y lo siguen siendo en mi vida.

Muchísimas gracias al resto de miembros del departamento de Protección Vegetal y Biotecnología y al departamento de Estadísica e Investigación Operativa por su cálida acogida desde que llegué y por supuesto, a mis secretari@s favorit@s (Trini, Tere, Juana y Alberto) por vuestra ayuda con la gestión y por recibirme siempre con una sonrisa.

Quiero extender mis agradecimientos a aquell@s amig@s que Valencia me ha brindado: Pedro, Benito, Chicote, Paco, Diego, Fer, Diana, Silvia, Mónica, Miguel, Ana Maria, Raquel, Dani, Josema, Ana, Inma, Adri, Daniela, Damián, Bea, Elvira, etc. Pero en especial a Pablo, José, Marta, Ángela y Paqui porque siempre he sentido vuestra presencia cuando más os he necesitado, me habéis demostrado que nuestra amistad no tiene límites.

I would like to thank to people that I met during my two visits in Edinburgh: Álvaro, Nausika, Miguel, Per, Finn, Ning, Michelle, Lorraine, Manu, Cristina, Serena, etc. thank you for making me spend unbelievable moments. Your courage and sympathy make you terrific.

En estas líneas, quiero recordar también a aquell@s profesores/as que me ayudaron a crecer no solo a nivel académico tambien a nivel personal: Ana María Huerta Mazcuñan, José Luis Serrano Ruiz, y José Luis Bueno Pareja.

Y como no, a mi familia, por su apoyo incondicional y por su inmedible confianza en mi. En particular:

A aquellos que ya no están, mi abuela Luisa, mi abuelo Enrique, mi abuelo Pablo, mi bisabuela Joaquina y mi bisabuelo Eufrasio, gracias porque desde niño siempre me habéis animado a luchar por aquello que quiero. Allá donde estéis este trabajo va por vosotr@s.

A mis ti@s Jesús Ángel, Amparo, Mercedes, José, Joaquín, mis prim@s Ángel, David, José Antonio, Mercedes, Mercedes, Joan, Pablo, Vero, José Luis, Victor, gracias porque a pesar de que no nos veamos con mucha frecuencia, siempre he sentido vuestro apoyo.

A mi tío Juanito, gracias porque eres el hermano mayor que nunca he tenido; y a mi prima Olivia, gracias por transmitirme siempre esa energía, optimismo y alegría.

A mi abuela Eufrasia, gracias por ser mi viviente ángel que me guía, siempre tienes un consejo, un gesto, una sonrisa, una caricia que cambia mi día. Eres mi ejemplo a seguir.

Y a vosotros, mis padres, Joaquín y Esperanza, gracias por ser las dos razones de mi existencia, mis fuentes de inspiración, por darme la vida y ser los pilares que la sustentan. Nadie sabe lo que habéis sacrificado y trabajado para que un día como hoy yo pudiese estar escribiendo estas líneas. Millones de gracias por mostrarme siempre que con trabajo y esfuerzo todo se consigue. Sois las personas más grandes que existen.

A tod@s vosotr@s gracias de corazón, habéis marcado mi camino.

"There is a driving force more powerful than steam, electricity and nuclear power: the will"

Albert Einstein

UNIVERSITAT DE VALÈNCIA

Resumen

Facultad de Ciencias Matemáticas
Departamento de Estadística e Investigación Operativa

Programa de Doctorado en Estadística y Optimización

En el mundo en que vivimos, producimos aproximadamente 2.5 quintillones de bytes de datos por día. Esta enorme cantidad de datos proviene de las redes sociales, Internet, satélites, etc. Todos estos datos, que se pueden registrar en el tiempo o en el espacio, son información que puede ayudarnos a comprender la propagación de una enfermedad, el movimiento de especies o el cambio climático. El uso de modelos estadísticos complejos ha aumentado recientemente en el contexto del estudio de la distribución de especies. Esta complejidad ha hecho que los procesos inferenciales y predictivos sean difíciles de realizar. El enfoque bayesiano se ha convertido en una buena opción para lidiar con estos modelos, debido a la facilidad con la que se puede incorporar la información previa, junto con el hecho de que proporciona una estimación de la incertidumbre más realista y precisa.

En esta tesis, mostramos una visión actualizada del uso de las últimas herramientas estadísticas que han surgido en la aplicación de modelos de distribución de especies (SDMs) en contextos reales desde una perspectiva bayesiana, y desarrollamos nuevas herramientas metodológicas para resolver algunos problemas estadísticos que aparecieron en ese proceso.

Con respecto a la aplicación de las últimas herramientas estadísticas en el contexto de los SDMs, los objetivos específicos han sido modelizar la producción de ascosporas *Plurivorosphaerella nawae* en la hojarasca de caqui; estudiar los factores espaciales y climáticos asociados con la distribución de la mancha negra de los cítricos causada por el hongo *Phyllosticta citricarpa*; analizar los efectos de la estructura genética y la autocorrelación espacial en los cambios de rango de distribución de las especies; y estudiar la distribución del delfín mular (*Tursiops truncatus*). Dos objetivos han marcado la parte más metodológica de la tesis: una

revisión centrada en los problemas estadísticos en SDMs y la implementación de la regresión de Dirichlet bayesiana en el contexto de la aproximación de Laplace anidada integrada (INLA).

La tesis que aquí presentamos es un compendio de ocho artículos y a continuación mostramos su estructura. En los cuatro primeros capítulos presentamos una introducción general que incluye una descripción de los objetivos (Capítulo 1), la base de la metodología empleada (Capítulos 2 y 3) y una descripción de los resultados obtenidos (Capítulo 4). En los ocho capítulos siguientes, mostramos todos los artículos que componen este compendio. Y por último, incluimos el Capítulo 13, donde se presentan algunas conclusiones y líneas futuras de investigación, seguido de una bibliografía genérica correspondiente a los capítulos introductorios.

Contents

List of Figures

List of Tables

Introduction

In the world that we live, we produce approximately 2.5 quintillion bytes of data per day. This huge amount of data comes from social media, internet, satellites, etc. All these data, which can be recorded in time or in space, are information that can help us to understand the spread of a disease, the movement of species or the climate change.

The use of complex statistical models has recently increased in the context of species distribution behavior. This complexity has made the inferential and predictive processes challenging to perform. The Bayesian approach has become a good option to deal with these models due to the ease with which prior information can be incorporated along with the fact that it provides a more realistic and accurate estimation of uncertainty.

This Thesis is devoted to provide an updated vision of the use of the latest statistical tools that have been emerging in the application of species distribution models (SDMs) in real contexts from a Bayesian perspective, and to develop new methodological tools to solve some statistical problems appeared in that process.

With regard to the application of the latest statistical tools in the context of SDMs, the particular objectives have been to model the production of *Plurivorosphaerella nawae* ascospores in persimmon leaf litter; to study the spatial and climatic factors associated with the distribution of the citrus

black spot disease caused by *Phyllosticta citricarpa*; to analyze the effects of geographic genetic structure and spatial autocorrelation on species distribution range shifts; and to study the bottlenose dolphin (*Tursiops truncatus*) distribution.

Two goals have guided the most methodological part of the Thesis: a review with the focus in the statistical issues in Species Distribution modeling, and the implementation of Bayesian Dirichlet regression in the context of the integrated nested Laplace approximation (INLA).

These two main objectives provide the following structure to the Thesis, which is a compendium of eight papers. The first four chapters are devoted to present a general introduction including a description of the objectives (Chapter 1), the basis of the methodology employed (Chapters 2 and 3) and a description of the results obtained (Chapter 4).

The next eigth chapters are dedicated to display all the papers which compose this compendium. In particular, in Chapter 5, we present a paper where a hierarchical Bayesian beta regression is constructed to fit the dynamics of *Plurivorosphaerella nawae* ascospore production in the leaf litter. Chapters 6, 7 and 8 use geostatistical tools and hierarchical Bayesian logistic regression models to study the spatial and climatic factors associated with the distribution of the citrus black spot disease. In Chapter 9, we develop spatial hierarchical Bayesian beta regression models to analyze the effects of geographic genetic structure and spatial autocorrelation on species distribution range shifts. In Chapter 10, a non-stationary hierarchical Bayesian logistic model is employed to study the bottlenose dolphin (*Tursiops truncatus*) distribution. Chapters 11 and 12 are devoted to cover the most methodological part of this Thesis. We present a review with the focus in the statistical issues in Species Distribution modeling (Chapter 11), and a way to implement the Bayesian Dirichlet regression in the context of the integrated nested Laplace approximation (Chapter 12).

The final part of the Thesis includes Chapter 13, where some conclusions and future lines of research are presented, and a generic bibliography corresponding to the introductory chapters.

Chapter 1

Modeling the distribution of species

1.1 Motivation

In many applied fields, the information on the presence/absence, abundance or proportion of a species is usually linked to environmental variables with the final objective of predicting where and how much of a species is likely to be present in unsampled locations or time periods. The models that do that are widely known as Species Distribution Models (SDMs).

This kind of models, where the spatio-temporal dynamics insight of species or diseases is a key issue, has been widely used in research areas such as ecology or plant epidemiology to study the risk associated with invasive species (Fitzpatrick et al., 2007; Luo and Opaluch, 2011), the potential effects of climate change (Iverson et al., 2004; Araújo et al., 2005; Brown et al., 2016), the design of protected areas, the protection of threatened species (Roos et al., 2015) or the potential distribution of infectious diseases (Peterson et al., 2002; Fatima et al., 2016; Juan et al., 2017; Martínez-Minaya et al., 2018), among many others.

In the last years, the complexity of the methods used in this context has increased (see for example, Guisan and Thuiller, 2005; Elith and Leathwick,

2009). They are mainly based on the assumption that the observations are conditionally-independent, while species distribution data often depict residual spatial autocorrelation (Kneib et al., 2008; Beale et al., 2010). Although the sampling is random, this spatial autocorrelation should be taken into account since the observations are often close and subject to similar environmental features (Banerjee et al., 2014). For this reason, the inclusion of spatial and spatio-temporal structures have grown enormously allowing the model to deal with other components to model variability not explained by the covariates.

However, the intricacy of the model is not only due to the effects included in the linear predictor, but also because of the likelihood. SDMs that correlate the occurrence or abundance of a species with abiotic variables (environmental variables) are typically used to investigate species-environment relationships. However, just a few cases the important influence of biotic interactions on especies is considered (Dormann et al., 2012). Pollock et al. (2014), which model the co-ocurrence of different frog species using a multivariate normal likelihood, or Wolf et al. (2017), which uses a multinomial likelihood for the study of the attack rates in a New Zealand intertidal whelk predator, are examples about how the interactions between especies are taken into account with a multivariate response. In particular, it is common to have measures of a multivariate phenomenon lying in a bounded interval that sum up to one. These data which mainly consist of proportions or percentages of disjoint categories are widely known as compositional data (Aitchison, 1982).

The combination of non-Gaussian data, in some cases multivariate data, a linear predictor and unobserved latent variables usually makes estimation and prediction computationally difficult. In the last years, Bayesian inference has become a good tool to deal with these complex models, because it allows both the observed data and model parameters to be random variables, resulting in a more realistic and accurate estimation of uncertainty. However, as usual in highly structured models, numerical approaches are needed to estimate and predict. Markov Chain Monte Carlo methods (MCMC) are so popular, but, the integrated nested Laplace approximation (INLA)

methodology (Rue et al., 2009) has become an alternative to MCMC, guaranteeing a higher computational speed for a particular case of models, the Latent Gaussian models (LGMs).

1.2 Main objectives

In what follows, we present the objectives of this Thesis. All these objectives have been developed in collaboration with different researchers from different institutions. Around these objectives, five papers have been published or accepted in indexed journals in the Journal Citation Reports (JCR, Journal Citation Reports Social Sciences Edition, 2017), as well as a letter to the editor of one of these journals, and two more are about to be sent to other two JCR indexed journals.

This Thesis arose from a collaboration with a research institute interested in plant disease epidemiology issues. We started by applying SDMs in this context, posteriorly, we extended it to the marine and vegetal species area. During this time, new statistical problems related to the application of SDMs in real contexts needing methodological development appeared. As a consequence, this Thesis can be structured around two main objectives:

- **Objective 1:** applying SDMs in plant disease epidemiology, and marine and vegetal species distribution. This objective can be split in four specific goals:
  - **Objective 1.1:** modeling the production of *Plurivorosphaerella nawae* ascospores in persimmon leaf litter. We considered a Bayesian beta regression to solve the problem. The results are about to be sent to a JCR indexed journal. As a subproduct from this model, we have also constructed a warning system that is about to be implemented in the Valencian Institute for Agricultural Research (IVIA) in order to help farmers to take prompt decisions on fungicide applications.
  - **Objective 1.2:** study of the spatial and climatic factors associated with the distribution of the citrus black spot disease. The

results obtained from this study (two papers and a letter to the Editor) have been published in the ***European Journal of Plant Pathology***.

- **Objective 1.3:** analysis of the effects of geographic genetic structure and spatial autocorrelation on species distribution range shifts. A paper has been accepted in the journal ***Molecular Ecology Resources***.
- **Objective 1.4:** the study of the bottlenose dolphin (*Tursiops truncatus*) distribution. Regarding to this topic, a paper has been accepted in the journal ***Ecological Modelling***.

- **Objective 2:** developing new methodological tools to solve statistical problems appeared in the application of SDMs in real contexts after checking the state of the art of the statistical analysis of SDMs. As a result, this objective has two specific aims:
  - **Objective 2.1:** a review with the focus in the statistical issues in species distribution modeling. This review has been published in the journal ***Stochastic Environmental Research and Risk Assessment***.
  - **Objective 2.2:** implementing Bayesian Dirichlet regression in the context of the integrated nested Laplace approximation. The methodological results along with their implementation in an `R`-package are about to be sent to a JCR indexed journal.

1.3 Methods in SDMs

Environmental niche modeling, habitat modeling or Species distribution modeling are just three ways to name an area whose mainly aim is to predict the distribution of a species across geographic space and time using environmental data. Environmental data are most often climate data (e.g. temperature, precipitation), but can include other variables such as soil type, water depth or land cover.

The aim of these models is to characterize the distribution of species. Usually, they deal with the ocurrence, the abundance or the relative abundance of a species. As we have pointed out before, they study how environmental conditions are related with the phenomenon. This is useful to predict the presence or the abundance of a species under climate change scenarios, or to predict the introduction of invasive species.

In the literature, there are different algorithms applied to classify species distribution as a function of a set of environmental variables. There is a group of methods which deal with presence only datasets, including maximum entropy algorithm, environmental distance, or envelope methods, such as Maxent (Phillips et al., 2006).

The main idea of Maxent consists of expressing a probability distribution where each grid cell has a predicted suitability of conditions for the species, drawn from a set of environmental variables and georeferenced ocurrence locations. With this method, we obtain maps with the representation of the probability of presence. The environmental layers and a set of grid cells jointly with a set of locations where the species has been observed determine the model. The suitability of each grid cell for the presence of the especies as a function of the environmental variables is expressed by the model. The higher the value of the function at a particular grid cell, the more suitable the conditions in the grid cells for that species. The resulting function is a probability distribution over all the grid cells. This distribution is subject to some constraints and it is selected as the one which has maximun entropy.

Another group of methods include machine-learning algorithms such as Boosting Regression Trees, Classification Trees or Random Forests (Cutler et al., 2007; Evans et al., 2011; Elith and Leathwick, 2017; Walker et al., 2017). The last group relates to traditional Generalized Linear Models (GLM) and this is our starting point.

1.4 Generalized linear models (GLM)

Generalized Linear Models (GLM) arose as a consequence of modeling difficulties when data came from non-normal distributions such as binomial,

gamma or Poisson distributions. With GLMs, phenomenons as animal counts, abundances of species, biomass data or presence/absence of diseases can be modeled. The main features of GLMs are:

- Response variables are independent $Y_1, \ldots, Y_n$, and they have the same distribution (parametric) in the exponential family.
- A parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ and a design matrix $\boldsymbol{V}$ which is formed by the covariates values are required. The product of the parameter vector with the design matrix comprises the linear predictor η_i.
- A diferenciable and monotonous function entitled link function $g(\cdot)$, that relates the mean $\mu_i = \mathrm{E}(y_i)$ with the linear predictor $\eta_i = \boldsymbol{V}_i\boldsymbol{\beta}$, i.e., $g(\mu_i) = \eta_i = \boldsymbol{V}_i\boldsymbol{\beta}$.

Then, a GLM is usually represented by means of:

$$\begin{aligned} Y_i \mid \boldsymbol{\beta} &\sim p(Y_i \mid \eta_i) \\ g(\mu_i) = \eta_i &= \boldsymbol{V}_i\boldsymbol{\beta}\,, \end{aligned} \tag{1.1}$$

where $p()$ is the distribution of the response variable, and $p(Y_i \mid \eta_i)$ the likelihood conditioned to the linear predictor.

It is worth noting that in GLMs there is a linear relationship between the linear predictor and the covariates. However, in some cases the relationships are not linear (Guisan et al., 2002). Generalized additive models (GAMs) can be used to model so. GAMs are a semi-parametric extension of GLMs, where in addition to linear functions, smooth terms are included in the model. These smooth terms are frequently modeled through different types of smoothing-splines, that allow fitting non-linear effects to the covariates.

It is also worth noting that, even though it does not belong to the exponential family, there exist another probability distribution that is useful for practitioners in the context of SDMs, the beta distribution. Its domain is defined in the open interval $(0, 1)$, and can be used to model the proportion of a species in an area, or just to model some indexes which take values in

this interval. The model is constructed in the same way as presented for GLMs (Equation 1.1), and it is widely known as Beta regression (Ferrari and Cribari-Neto, 2004).

All these methods are based on the assumption that the observations are conditionally-independent. But this is not always the case. It is often that model residuals display non-independent patterns or structures that covariates can not explain. The presence of correlated model residuals compromises the fit of the whole model and its quantification of uncertainty.

Depending on the process under study, the unobserved component can take several correlation structures. For example, we may expect correlated residuals within site if we have repeated measurements of a process at each sampling site. In this case, an independent random effect for each site can solve the problem. But, if this sampling is done over time, residuals can be temporary correlated. Time series analysis can be useful to deal with these processes.

The same can happen with space, model residuals are also prone to spatial correlation (Kneib et al., 2008). In that case, we rely on Tobler's principle "near things are more related than distant things" (Tobler, 1970). But this spatial structure can vary depending on the nature of the data and its spatial domain: if it is areal data, correlation structures are often specified using conditional autoregressive models with a given order of neighbouring regions (Besag et al., 1991). For the case where we deal with a continuous space, correlation functions need also to be continuous over distance.

In this Thesis, we deal with spatial data in a continuous space, which is usually known as point-referenced data or Geostatistal data. We discuss more about continuous fields and continuous autocorrelation functions in the following section.

1.5 Geostatistical data

Geostatistical or point-referenced data consist on a collection of data in a fixed set locations over a continuous spatial field. A GLM can be constructed

using the coordinates of these data as covariates to describe the spatial variation of the variable of interest. Alternatively, the coordinates can be incorporated by means of a GAM in order to describe the effect of the location.

However, it is more natural to formulate mixed-effects regression models in which the linear predictor is made of a trend plus a spatial variation (Haining, 2003). Usually the trend is composed of fixed effects or smooth terms on covariates, meanwhile, the spatial variation is usually modeled using correlated random effects. To construct the general model with spatial random effects, let's define first what is the structure of this random effect.

A random spatial effect $W(\boldsymbol{s})$ at a location $\boldsymbol{s} \in \mathcal{D}$ can be considered as a stochastic process characterized by a spatial index $\boldsymbol{s}$ which varies continuously in the fixed domain $\mathcal{D}$, where $\mathcal{D}$ is a fixed subset of r-dimensional Euclidean space. If $r = 1$, these kind of processes have a rich presence in the time series literature. In the spatial context, usually r is encountered to be 2 (northings and eastings) or 3 (northings, eastings, and altitude). In this Thesis, we assume $r = 2$ for the spatial processes.

Let $w(\boldsymbol{s}_i)$, $i = 1, 2, \ldots, n$ be a realization of $W(\boldsymbol{s})$ at n locations. Let's suppose that data $y(\boldsymbol{s}_i)$ have been observed at locations $\boldsymbol{s}_i$, $i = 1, \ldots, n$. $y(\boldsymbol{s})$ may represent the presence/absence of a species or a disease, the abundance of a species, or just the relative abundance of a species. If an underlying spatial process generate this data, the parameters of this process can be fitted by considering $y(\boldsymbol{s}_i) = w(\boldsymbol{s}_i)$. Observe that it is conceptually sensible to assume that the phenomenon can be measured at all possible sites in the domain, and so, in practice, the data are only a partial realization of that spatial process. That is, we only have measurements at a finite number of locations out of an infinite number of possible locations.

The main problem when we deal with this kind of spatial process $w(\boldsymbol{s})$ is the inference and prediction at new locations, based upon this partial realization. The main idea is to infer about the surface at an uncountable number of locations despite only seeing the process in a finite number of locations. In other words, we need to infer a distance based covariance function that best represents the underlying spatial proccess of our data, and then predict at unsampled locations using kriging (Cressie, 1990).

The spatial process $w(\boldsymbol{s})$ is Gaussian if for any $n \geq 1$ and any set of sites $\boldsymbol{s} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$, $\boldsymbol{w} = \{w(\boldsymbol{s}_1), \ldots, w(\boldsymbol{s}_n)\}$ has a multivariate normal distribution with mean $\boldsymbol{\mu} = \mathrm{E}(w(\boldsymbol{s}))$ and a structured covariance matrix $\boldsymbol{\Sigma}$. Usually $\boldsymbol{\mu}$ is assumed to be $\mathbf{0}$. In the literature, this process is widely known as a Gaussian field (GF; Rue and Held, 2005). From now on, we assume that the spatial process is a GF.

To model spatial dependence, it is usual to assume a probability distribution for the data conditional on an unobserved random effect, which is a GF. Then the model presented in Equation (1.1) with a spatial effect can be rewritten as:

$$\begin{aligned} y_i \mid \boldsymbol{\beta}, w_i &\sim p(y_i \mid \eta_i)\,, \\ \eta_i &= \boldsymbol{V}_i\boldsymbol{\beta} + w_i\,, \\ \boldsymbol{w} &\sim GF(\mathbf{0}, \boldsymbol{\Sigma})\,, \end{aligned} \tag{1.2}$$

where $\boldsymbol{w} = \{w(\boldsymbol{s}_1), \ldots, w(\boldsymbol{s}_n)\} = \{w_1, \ldots, w_n\}$ is the spatial random term.

The key issue in spatial statistics is the covariance function $\mathcal{C}$, which determines the covariance between random variables in two different points, and not only allows us to know how two points in space are correlated, but also defines the covariance matrix $\boldsymbol{\Sigma}$ of the GF. If $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ are two locations in space, then the covariance function is defined as

$$\mathcal{C}(w(\boldsymbol{s}_i), w(\boldsymbol{s}_j)) = Cov(w(\boldsymbol{s}_i), w(\boldsymbol{s}_j))\,, \tag{1.3}$$

and each element of the matrix $\boldsymbol{\Sigma}_{ij}$ is defined as

$$\boldsymbol{\Sigma}_{ij} = \mathcal{C}(w(\boldsymbol{s}_i), w(\boldsymbol{s}_j))\,. \tag{1.4}$$

Despite we know how to define the covariance function, it is not always easy to deal with it. This function depends on multiple characteristics. For instance, how the species spreads out. If we want to model the distribution of marine species in the sea, the spread of these species can be conditioned for the ocean currents what can make the spatial effect dependent on the direction. This leads us to define some properties which we have to take into account when we deal with GFs.

The first property is called **weak stationarity or second-order stationarity**. We say that the GF is second-order stationary if $\mu(\boldsymbol{s}) \equiv \mu$ and $Cov(w(\boldsymbol{s}), w(\boldsymbol{s}+\boldsymbol{h})) = \mathcal{C}(\boldsymbol{h})$ for all $\boldsymbol{h} \in \mathbb{R}^r$ such that $\boldsymbol{s}$ and $\boldsymbol{s}+\boldsymbol{h}$ lie within $\mathcal{D}$. In other words, the covariance function in two different locations depends on the distance vector between these two locations. An example could be the spread of a pathogen in plants. If there is a road close to the crop, maybe this pathogen could spread faster by the road in cars or trucks than in the crop, it would depend on the direction. When this property does not fulfill, we call the process non-stationary.

The second property is called **isotropy**. We say that the GF is isotropic if the covariance function depends only on the Euclidean distance between points, i.e., $Cov(w(\boldsymbol{s}), w(\boldsymbol{s}+\boldsymbol{h})) = \mathcal{C}(\|\boldsymbol{h}\|)$. For instance, if we think again in the spread of a pathogen in a crop, it would mean that the spread does not depend on the direction, just on the distance. If this property does not fulfill, we call the process anisotropic.

In most of the usual spatial data analyses we assum that the process is second-order stationary and isotropic. However, in the case of objective 1.4. where there exist barriers in the space, and then the space is not continuous, it is necessary to use another kind of processes. We will treat this case in the Chapter 3.

When the process is stationary and isotropic, different covariance functions have been proposed (Banerjee et al., 2014). However the Matérn class of covariance functions seems to be the most flexible because it embraces a number of covariance functions depending on the value of its smoothing parameter. For two locations $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ in $\mathcal{D}$, the stationary and isotropic Matérn covariance function is defined as:

$$\begin{aligned} \mathrm{Cov}(w(\boldsymbol{s}_i), w(\boldsymbol{s}_j)) &= \mathcal{C}(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|), \\ &= \frac{\sigma^2_{\boldsymbol{w}}}{2^{\nu-1}\Gamma(\nu)} (\kappa \|\boldsymbol{s}_i - \boldsymbol{s}_j\|)^{\nu} K_{\nu}(\kappa \|\boldsymbol{s}_i - \boldsymbol{s}_j\|), \end{aligned} \tag{1.5}$$

being $\|\boldsymbol{s}_i - \boldsymbol{s}_j\|$ the Euclidean distance between the two locations $\boldsymbol{s}_i, \boldsymbol{s}_j$ and $\sigma^2_{\boldsymbol{w}}$ the marginal variance. Moreover, K_ν is the modified Bessel function of the second kind and order $\nu > 0$, which measures the degree of smoothness of the process. Conversely, $\kappa > 0$ is a scaling parameter related to the

distance at which the spatial correlation becomes almost null, i.e., the range (for more information on the Matérn covariance model see Handcock and Stein, 1993; Stein, 1999). In Chapter 3, we show the relationship between κ and the range. If $\nu = \frac{1}{2}$ then the covariance function is the exponential covariance function, and if $\nu \to \infty$, then we obtain the Gaussian covariance function.

The covariance matrix of the GFs depends on two parameters κ and $\sigma^2_{\boldsymbol{w}}$. This can be easily embraced in a hierarchical structure. Taking into account that the covariance matrix of the GFs depends in these two parameters, κ and $\sigma^2_{\boldsymbol{w}}$, Equation (1.2) becomes:

$$\begin{aligned} y_i \mid \boldsymbol{\beta}, w_i &\sim p(y_i \mid \eta_i)\,, \\ \eta_i &= \boldsymbol{V}_i\boldsymbol{\beta} + \boldsymbol{w}_i\,, \\ \boldsymbol{w} &\sim GF(\mathbf{0}, \boldsymbol{\Sigma}(\kappa, \sigma^2_{\boldsymbol{w}}))\,. \end{aligned} \tag{1.6}$$

1.6 STAR and hierarchical modeling

Until now, we have focused above all in spatial patterns, however, the temporal variation could be equally important. The spread of a species can vary in space and also in time. Then, spatial models can be extended to spatio-temporal models including a time dimension. A more general structure for modeling species which comprises all kind of models until now presented can be constructed.

If $\boldsymbol{y} = (y_1, \ldots, y_n)$ represents the observed values of the corresponding response variable Y with mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$, each μ_i can be easily linked to a structured additive predictor η_i through a link function $g(\cdot)$, so that $g(\boldsymbol{\mu}) = \boldsymbol{\eta}$. The structured additive predictor η_i accounts for the effect of various covariates in an additive way:

$$\eta_i = \beta_0 + \sum_{m=1}^{M} \beta_m v_{mi} + \sum_{l=1}^{L} f_l(z_{li})\ , \tag{1.7}$$

where β_0 corresponds to the intercept; the coefficients $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_M\}$ quantify the (linear) effect of some covariates $\boldsymbol{v} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_M)$ on the response; and $\boldsymbol{f} = \{f_1(\cdot), \ldots, f_L(\cdot)\}$ are unknown functions of the covariates $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_L)$, and can assume different forms such as smooth nonlinear effects of covariates, time trends and seasonal effects, random intercept and slopes as well as temporal or spatial random effects. These models in which usually the mean of the response variable y_i is linked to a structured predictor that accounts for the effects of various covariates in an additive way are known as Structured Additive Regression (STAR) models (Fahrmeir and Tutz, 2001).

Note also that, as we have previously introduced in Equation (1.7), the model involves multiple parameters or random effects that can be connected by the structure of the problem. This kind of models can be considered as hierarchical models. Here, observable outcomes are modeled conditionally on certain parameters, which in turn are given a probabilistic specification in terms of further parameters and adding various levels of the modeling. Hierarchical models provide a generalization of all the models presented in this Thesis and allow us to deal with any kind of data that we can find when we deal with SDMs.

Once we know the structure of the models is time to make inference. Although other approaches can be used such as maximun likelihood or restricted maximum likelihood, in this Thesis we focus on the Bayesian approach. The next chapter describes some of the basis of this approach and how we can perform in practice the inferential and predictive task.

Chapter 2

Species distribution modeling using INLA

Analyzing the distribution of a species has the same concerns that one has to face when modeling practical real problems: model specification, estimation and prediction. The Bayesian approach provides a framework for combining complex data models (such as SDMs) and expert opinion, and it easily addresses model specification, and therefore inference and prediction (Banerjee et al., 2014). This chapter is devoted to present a recent but widely known tool to make computational Bayesian inference: the integrated nested Laplace approximation (INLA).

2.1 Bayesian methodology to statistical inference and prediction

The Bayesian approach to inference dates from the eighteenth century, when a British clergymen, Thomas Bayes, and a French scientist presented a simple but powerful mathematical treatment of the non-trivial problem of statistical data analysis, the Bayes theorem (Bayes, 1764; Laplace, 1812):

Given two events A and B, this theorem states that

$$P(B \mid A) = \frac{P(A \mid B) \times P(B)}{P(A)}, \tag{2.1}$$

being $P(B \mid A)$ the conditional probability of B given A, $P(A \mid B)$ the conditional probability of A given B, $P(A)$ the probability of A, and $P(B)$, the probability of B.

The interest lies in the probability of the event B given that A occurs. $P(B)$ is calculated before A is observed, then the probability of observing A given B is used to update the original $P(B)$ so that $P(B \mid A)$ is obtained. The main idea is update the probability of B ($P(B \mid A)$) using the information that the researcher has about B expressed in $P(B)$ and combining it with the result of an experiment $P(A \mid B)$. With this theorem a new philosophy was born. Initial beliefs could be evaluated, updated and modified with new information.

However, in the context of SDMs, as we presented in the previous chapter, we do not talk about events, we try to model some phenomenons, for example the abundance or the presence/absence of a species, in terms of some covariates (environmental covariates for instance) and random effects (spatial, temporal, etc.). We assign a probability distribution to the response variable that could be a Binomial, Poisson, Beta, etc. We denote this probability distribution as $p(\boldsymbol{y} \mid \boldsymbol{x})$ being $\boldsymbol{y}$ a vector of realizations of the response variable, and $\boldsymbol{x}$ as defined in previous chapter, the parameters and random effects which the response variable is depending on. Moreover, $p(\boldsymbol{y} \mid \boldsymbol{x})$ is called likelihood and reflects the information given by the data under the model defined by $\boldsymbol{x}$.

Besides the information given by the data (as frequentist approach does), additional information such as expert knowledge or previous studies can be included (Clark and Gelfand, 2006). This additional information is added in the model giving probability distributions to $\boldsymbol{x}$, in other words, parameters are considered random variables. We denote priors as $p(\boldsymbol{x})$. If there is not previous knowledge about a parameter, prior distribution should be as less informative as possible.

Lastly, using the Bayes theorem, prior distributions are combined with the likelihood to get posterior distributions updating the information of the parameters.

$$p(\boldsymbol{x} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{x}) \cdot p(\boldsymbol{x})}{p(x)} \propto p(\boldsymbol{y} \mid \boldsymbol{x}) \cdot p(\boldsymbol{x}) \,. \tag{2.2}$$

This process is well known as **Bayesian inference** and it has been a revolution in many applied fields because, in addition to the information given by the data (as frequentist approach does), other information such as expert knowledge or previous studies can be included (Clark and Gelfand, 2006).

2.2 Hierarchical Bayesian models

Bayesian inference allows us to make inference for different kind of models, from a simple linear regression to a spatio-temporal model. Nevertheless, when we deal with complex models as we can find in the SDMs context, new parameters are needed in the model.

For instance, the study phenomenon might be the presence/absence of a particular species in some locations. We could try to explain this phenomenon in terms of environmental variables and an spatial effect such as the one introduced in Chapter 1. This spatial effect depends on new parameters controlling the spatial similarity accross locations. These new parameters are called hyperparameters and prior probability distributions need to be assigned to make inference using the Bayesian paradigm.

The model is presented in terms of three entities, all of which have stochastic elements: data, process (formed by parameters and random effects) and the hyperparameters. As stochasticity is relevant for each stage of the process, we can think in terms of a joint distribution:

$$\begin{aligned} p(\text{data, process, hyperparameters}) \propto \\ \propto \;\; & p(\text{data} \mid \text{process, hyperparameters}) \\ & \times p(\text{process} \mid \text{hyperparameters}) \\ & \times p(\text{hyperparameters}) \,. \end{aligned}$$

The joint distribution on the left side is provided in terms of three pieces on the right hand side. These pieces are usually easier to consider individually rather than thinking about the entire joint distribution. These pieces define each of the levels of a Hierarchical Bayesian model:

LEVEL 1	Likelihood	$p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta_1})$
LEVEL 2	Prior distributions and random effects	$p(\boldsymbol{x} \mid \boldsymbol{\theta}_2)$
LEVEL 3	Hyperprior distributions	$p(\boldsymbol{\theta})$

In the first level, the likelihood is depicted $p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta_1})$. But likelihood is conditioned to a random vector composed by the parameters and the random effects of the model $p(\boldsymbol{x} \mid \boldsymbol{\theta}_2)$. In the second level, prior distributions for this random vector are asigned. In the last level, priors for the hyperparameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ are given $p(\boldsymbol{\theta})$.

In order to show how an SDM can be seen as a hierarchical Bayesian model, we rewrite the model presented in Equation (1.6) as a hierarchical Bayesian model:

- **LEVEL 1: Likelihood**.

$$\begin{aligned} y_i \mid \boldsymbol{\beta}, w_i &\sim p(y_i \mid \eta_i) \\ \eta_i &= \boldsymbol{V}_i\boldsymbol{\beta} + w_i \end{aligned} \tag{2.3}$$

- **LEVEL 2: Prior distribution for the parameters and random effects**

$$\begin{aligned} \boldsymbol{\beta} &\sim p(\boldsymbol{\beta}) \\ \boldsymbol{w} &\sim GF(\mathbf{0}, \boldsymbol{\Sigma}(\kappa, \sigma^2_{\boldsymbol{w}})) \end{aligned} \tag{2.4}$$

- **LEVEL 3: Priors for the hyperparameters**

$$\begin{aligned} \kappa &\sim p(\kappa) \\ \sigma^2_{\boldsymbol{w}} &\sim p(\sigma^2_{\boldsymbol{w}}) \end{aligned} \tag{2.5}$$

In Figure 2.1 we can appreciate in a more clearly way this hierarchy.

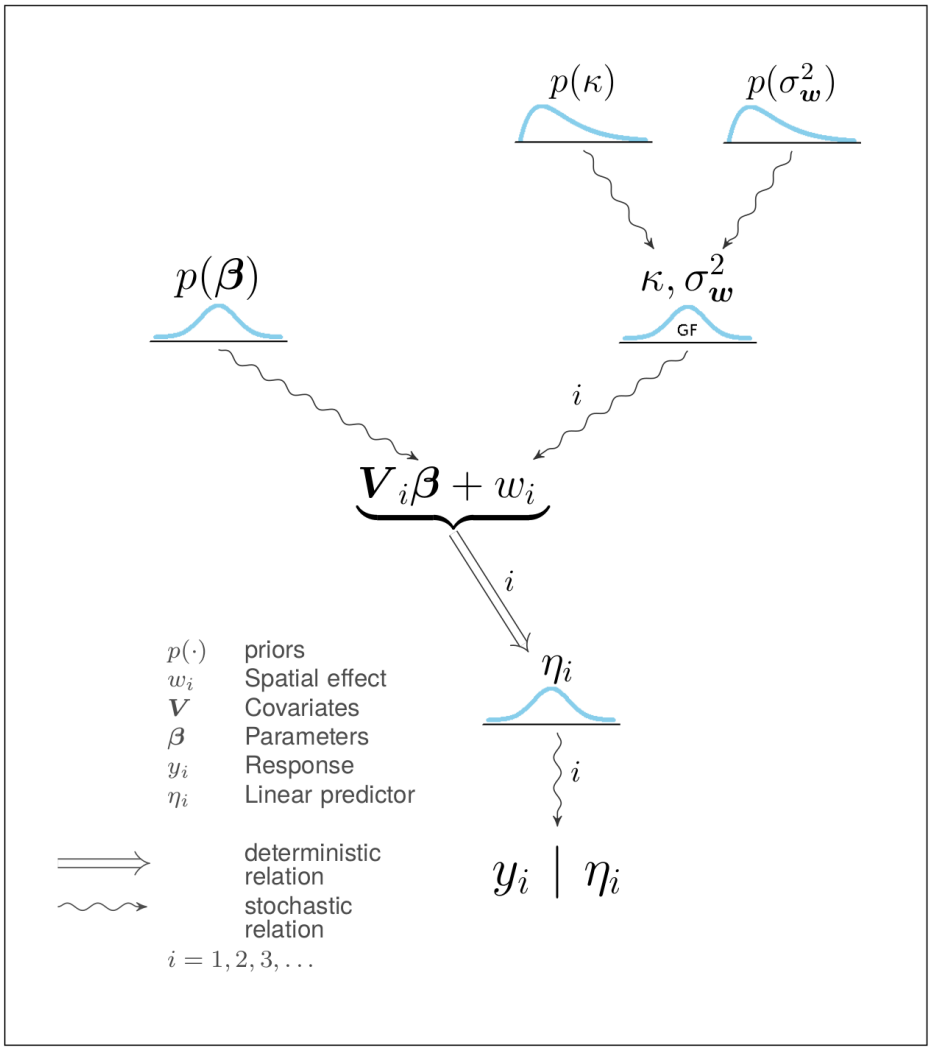

FIGURE 2.1: Example of a hierarchical Bayesian model with fixed effects and a continuous spatial effect.

2.3 Computational Bayes

The main interest in the Bayesian approach is to do inference in the parameters and the hyperparameters, i.e., calculate $p(\boldsymbol{x} \mid \boldsymbol{y})$ and $p(\boldsymbol{\theta} \mid \boldsymbol{y})$. From a mathematical point of view, the inference step is easy, prior beliefs about the unknown parameters are updated with available information in observed data, but this simplicity becomes an ardous task when computing the resulting posterior distributions (Rue et al., 2017).

The rise of new technologies has caused an explosion in the use of Bayesian statistics in the last 20 years, being now its peak. The computational power attained has made possible the development of computing tools in Bayesian practice. One of the most popular are the Markov Chain Monte Carlo methods (MCMC), which cleverly construct a Markov chain whose stationary distribution converges to the parameter posterior distribution. MCMC has been widely implemented in general Bayesian software/packages, such as the `BUGS` language (`Win/Open BUGS` (Lunn et al., 2000) and `JAGS` (Plummer et al., 2003)), `Stan` (Hoffman and Gelman, 2014) or `BayesX` (Umlauf et al., 2015). However, the dependence between simulated values due to the Markovian processes, and the computational cost that requires (above all in the context of spatial statistics) can make this methodology not appropiate in some situations.

On the other hand, the integrated nested Laplace approximation (INLA) methodology (Rue et al., 2009), whose main idea is to approximate the posterior distribution using the Laplace integration method, has become an alternative to MCMC, guaranteeing a higher computational speed for a particular case of models, the Latent Gaussian models (LGMs). The rest of the chapter is devoted to explain how the INLA methodology works.

2.4 Latent Gaussian Models (LGMs)

The reason underneath the possibility of using INLA is based on the fact that SDMs can also be seen as LGMs (Rue and Held, 2005), the class of models which INLA is designed for (Rue et al., 2009). LGMs can be also

expressed using a three-stage hierarchical Bayesian model formulation, in which observations $\boldsymbol{y}$ can be assumed to be conditionally independent, given a latent Gaussian random field $\boldsymbol{x}$ and hyperparameters $\boldsymbol{\theta_1}$,

$$\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta_1} \sim \prod_{i=1}^{n} p(y_i \mid x_i, \boldsymbol{\theta_1}) \,.$$

The versatility of the model class relates to the specification of the latent Gaussian field

$$\boldsymbol{x} \mid \boldsymbol{\theta_2} \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{\theta_2}), \boldsymbol{Q}^{-1}(\boldsymbol{\theta_2})) \,,$$

which includes all the latent (nonobservable) components of interest such as fixed effects and random terms describing the underlying process of the data. The hyperparameters $\boldsymbol{\theta} = (\boldsymbol{\theta_1}, \boldsymbol{\theta_2})$ control the Gaussian latent field and/or the likelihood for the data.

The LGMs are a class generalising the large number of related variants of additive and generalized models. If the likelihood $p(y_i \mid x_i, \boldsymbol{\theta})$ such that y_i only depends on its linear predictor η_i yields the generalized linear model setup, the set $\{x_i, i = 1, \ldots, n\}$ can be interpreted as η_i, being η_i the linear predictor which is additive with respect to other effects

$$\eta_i \;=\; \beta_0 + \sum_{m=1}^{M} \beta_m v_{mi} + \sum_{l=1}^{L} f_l(z_{li}) \;, \tag{2.6}$$

where β_0 corresponds to the intercept, the coefficients $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_M\}$ quantify the (linear) effect of some covariates $\boldsymbol{v} = (v_1, \ldots, v_M)$ on the response, and $\boldsymbol{f} = \{f_1(\cdot), \ldots, f_L(\cdot)\}$ are unknown functions of the covariates $\boldsymbol{z} = (z_1, \ldots, z_L)$ which are represented by Gaussian processes. If Gaussian prior is assumed for the intercept and the parameters of the fixed effects, the joint distribution of $\boldsymbol{x} = (\boldsymbol{\eta}, \beta_0, \boldsymbol{\beta}, \boldsymbol{f}_1, \boldsymbol{f}_2, \ldots)$ is then Gaussian. This yields the latent field $\boldsymbol{x}$ in the hierarchical LGM formulation. Regarding to the set of hyperparameters $\boldsymbol{\theta}$, it contains the parameters corresponding to the variance, scale or correlation of the likelihood and the model components (Martins et al., 2013).

In the INLA context, it is common to work with precision matrices $\boldsymbol{Q}$ that are usually sparse and make the latent field not only be Gaussian, but

also a sparse Gaussian Markov random field (GMRF; Rue and Held, 2005). A GMRF is just a GF with additional conditional independence properties: x_j and x'_j are conditionally independent given the remaining elements. This provides the INLA methodology with nice computational properties.

2.5 The core of INLA: the Laplace method

The underpinnings of INLA are the Laplace approximations to the marginal distributions of the parameters and hyperparameters of LGMs. As a result, understanding the Laplace method is of relevant interest, which is what this section is devoted.

2.5.1 Approximating integrals in general

The Laplace approximation is a technique used for the approximation of integrals (Barndorff-Nielsen and Cox, 1989) of the form

$$\int_a^b \exp\{Mf(x)\}dx \ , \tag{2.7}$$

being $f(x)$ some twice-differentiable function, M a large number, and the integral endpoints a and b could possibly be infinite.

Let x_0 be a global maximun of f, which it is not an endpoint of the interval of integration. Let assume the second derivate is less than 0, i.e., $f''(x) < 0$. Using the Taylor series expansion of order 2, $f(x)$ can be expanded around x_0,

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2 + R\,, \tag{2.8}$$

where $R = O\left((x - x_0)^3\right)$. As f has a global maximun at x_0, and since x_0 is not an endpoint, it is a stationary point, i.e., $f'(x_0) = 0$. The approximation of $f(x)$ around x_0 can be rewritten as

$$f(x) \approx f(x_0) - \frac{1}{2}|f''(x_0)|(x - x_0)^2 \ , \tag{2.9}$$

emphasising that the second derivative is negative at the global maximun $f(x_0)$. Replacing this in Equation (2.7)

$$\begin{aligned}\int_a^b \exp\{Mf(x)\}dx \approx \\ \approx \exp\{Mf(x_0)\}\int_a^b \exp\left\{-\frac{1}{2}M|f''(x_0)|(x-x_0)^2\right\}dx\,. \end{aligned} \quad (2.10)$$

Note that the previous integrand is a Gaussian kernel, and so, if endpoints go from $-\infty$ to $+\infty$, it can be easily integrated. Note also that endpoints can be assumed to be $-\infty$ and $+\infty$ because of the fast decay of f far away from x_0. By simplicity, let $C = \exp\{Mf(x_0)\}$, then:

$$\begin{aligned}&\int_a^b \exp\{Mf(x)\}dx \approx \\ &\approx C\int_a^b \exp\{-M|f''(x_0)|(x-x_0)^2/2\}dx \\ &= C\int_a^b \exp\left\{-\frac{1}{2}\left(\frac{x-x_0}{\frac{1}{\sqrt{M|f''(x_0)|}}}\right)^2\right\}dx \\ &= C\frac{\sqrt{2\pi}}{\sqrt{M|f''(x_0)|}}\int_a^b \frac{\sqrt{M|f''(x_0)|}}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{x-x_0}{\frac{1}{\sqrt{M|f''(x_0)|}}}\right)^2\right\}dx \\ &\approx \exp\{Mf(x_0)\}\sqrt{\frac{2\pi}{M|f''(x_0)|}}\,,\ \text{as } M\to\infty\,. \end{aligned} \quad (2.11)$$

This approximation can be extented to the multivariate case, i.e., if $\boldsymbol{x}$ is a n dimension vector, then

$$\int \exp\{Mf(\boldsymbol{x})\}d\boldsymbol{x} \approx \sqrt{\frac{(2\pi)^n}{M|-\boldsymbol{H}|}}\exp\{Mf(\boldsymbol{x_0})\}\,,\ \text{as } M\to\infty\,,$$

being

$$H_{ij} = \frac{\partial^2}{\partial x_i \partial x_j} f(\boldsymbol{x})\Big|_{\boldsymbol{x}=\boldsymbol{x_0}}\,.$$

In next section, we see how this method is employed with probability density functions.

2.5.2 Approximating density functions

In this section, we present how the Laplace method can be used to approximate probability density functions. The idea is simple but powerful: approximate the target density with a Gaussian by matching the mode and the curvature at the mode.

Let g an unnormalized probability density function and the integral $\int_{\mathbb{R}} g(x)dx$, the normalizing constant. We want to approximate the normalized function

$$p(x) = \frac{g(x)}{\int_{\mathbb{R}} g(x)dx} \ . \tag{2.12}$$

To do so, we also suppose that g has a stationary point in x_0, and we work using $\log(g)$, to look for an expression similar to (2.7):

$$\int_{\mathbb{R}} g(x)dx \leftrightarrow \int_{\mathbb{R}} \exp\{\log(g(x))\}dx \ . \tag{2.13}$$

Once the integral is defined, Taylor's theorem of second order in the mode x_0 of the function $\log(g)$, and the fact that $\frac{d\log(g(x_0))}{dx} = 0$ are used. We denote $\hat{\sigma} = \frac{-1}{\frac{d^2\log(g(x_0))}{dx^2}}$, then $\log(g(x))$ can be written as follows:

$$\log(g(x)) \approx \log(g(x_0)) - \frac{1}{2\hat{\sigma}^2}(x - x_0)^2 \ . \tag{2.14}$$

As a consequence, $g(x)$ can be approximated as:

$$\begin{aligned} g(x) &\approx g(x_0)\exp\left\{-\frac{1}{2\hat{\sigma}^2}(x-x_0)^2\right\} \\ &= g(x_0)\exp\left\{-\frac{1}{2}\left(\frac{x-x_0}{\hat{\sigma}}\right)^2\right\} \ . \end{aligned} \tag{2.15}$$

Then, the integral in Equation (2.13) can be easily computed.

$$\begin{aligned}
&\int_{\mathbb{R}} g(x)dx = \\
&= \int_{\mathbb{R}} \exp\{\log(g(x))\}dx \\
&\approx \int_{\mathbb{R}} \exp\{\log(g(x_0)) - \frac{1}{2\hat{\sigma}^2}(x-x_0)^2\}dx \\
&= \int_{\mathbb{R}} \exp\{\log(g(x_0))\} \exp\left\{-\frac{1}{2\hat{\sigma}^2}(x-x_0)^2\right\} dx \\
&= g(x_0) \int_{\mathbb{R}} \exp\left\{-\frac{1}{2\hat{\sigma}^2}(x-x_0)^2\right\} dx \\
&= g(x_0)\sqrt{2\pi}\hat{\sigma} \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}\hat{\sigma}} \exp\left\{-\frac{1}{2}\left(\frac{x-x_0}{\hat{\sigma}}\right)^2\right\} dx \\
&= \sqrt{2\pi}\hat{\sigma} g(x_0) \ .
\end{aligned}$$

Once the value of the integral is calculated, the normalized version of $g(x)$ is:

$$\begin{aligned}
p(x) &= \frac{g(x)}{\int_{\mathbb{R}} g(x)dx} \\
&\approx \frac{g(x_0)\exp\left\{-\frac{1}{2}\left(\frac{x-x_0}{\hat{\sigma}}\right)^2\right\}}{\sqrt{2\pi}\hat{\sigma}g(x_0)} = \frac{1}{\sqrt{2\pi}\hat{\sigma}} \exp\left\{-\frac{1}{2}\left(\frac{x-x_0}{\hat{\sigma}}\right)^2\right\} \ .
\end{aligned}$$

We conclude that $p(x)$ can be approximated using a Gaussian distribution with mean the mode x_0 of the function p, and variance the Fisher's information $\frac{-1}{\frac{d^2\log(g(x_0))}{dx^2}}$, i.e.,

$$p(x) \approx \mathcal{N}\left(x_0, \frac{-1}{\frac{d^2\log(g(x_0))}{dx^2}}\right) \ . \tag{2.16}$$

To show how this approximation works, we present an example using a beta distribution target. Our goal here is to approximate the beta density through the normal approximation obtained using the Laplace method. Different parameters for the beta distribution are computed. In Figure 2.2, we

depict the different approximations that we get.

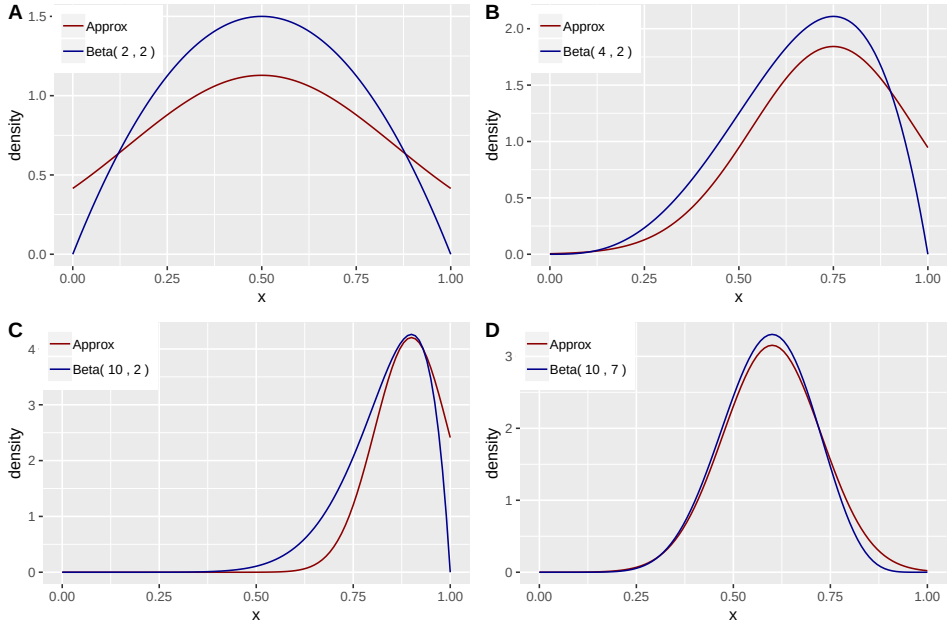

FIGURE 2.2: Gaussian approximation to the beta densities obtained using the Laplace method.

As this approximation does not work in a proper way when we deal with non simetric distributions, we present how this approximation is useful in the INLA context.

2.5.3 Laplace method in the INLA context

In this subsection, we present the way in which INLA makes use of the Laplace method. As above mentioned, the underpinnings of INLA are the approximations of marginal distributions of the parameters. In particular, if we are interested in computing a marginal distribution $p(\gamma_1)$ from a joint distribution $p(\boldsymbol{\gamma})$, this can be approximated by means of the definition of conditional probability and then making a Gaussian approximation of the

denominator obtained using the Laplace method $p(\boldsymbol{\gamma}_{-1} \mid \gamma_1)$, that is:

$$\begin{aligned} p(\gamma_1) &= \frac{p(\boldsymbol{\gamma})}{p(\boldsymbol{\gamma}_{-1} \mid \gamma_1)} \\ &\approx \left. \frac{p(\boldsymbol{\gamma})}{p_G(\boldsymbol{\gamma}_{-1}; \boldsymbol{\mu}(\gamma_1), \boldsymbol{Q}(\gamma_1))} \right|_{\boldsymbol{\gamma}^*_{-1} = \boldsymbol{\mu}(\gamma_1)} . \end{aligned} \tag{2.17}$$

If the distribution $p(\boldsymbol{\gamma}_{-1} \mid \gamma_1)$ is close to a Gaussian density, the results will be more accurate compared to a density that is very different from a Gaussian. Resultantly, unimodality will be necessary to accomplish an accurate approximation.

2.6 The integrated nested Laplace approximation (INLA)

The main idea behind the INLA approach is to approximate the posterior distribution of interest making use of the tools introduced in the previous sections. In particular, the interest is to approximate the marginal posteriors for the latent field $p(x_i \mid \boldsymbol{y})$ and the marginal posteriors for the hyperparameters $p(\theta_j \mid \boldsymbol{y})$:

$$\begin{aligned} p(x_i \mid \boldsymbol{y}) &= \int p(x_i, \boldsymbol{\theta}, \boldsymbol{y}) d\boldsymbol{\theta} \,, \\ &= \int p(x_i \mid \boldsymbol{\theta}, \boldsymbol{y}) p(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta} \,, \end{aligned} \tag{2.18}$$

$$p(\theta_j \mid \boldsymbol{y}) = \int p(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta}_{-j} \,. \tag{2.19}$$

Observe that in Equation (2.18) Bayesian theorem is used in order to be able to compute $p(x_i \mid \boldsymbol{y})$ in terms of conditional distributions. The key is to construct nested approximations. So, first, approximations to $p(x_i \mid \boldsymbol{\theta}, \boldsymbol{y})$ and $p(\boldsymbol{\theta} \mid \boldsymbol{y})$, i.e. $\tilde{p}(x_i \mid \boldsymbol{\theta}, \boldsymbol{y})$ and $\tilde{p}(\boldsymbol{\theta} \mid \boldsymbol{y})$, are computed. Posteriorly, integrating out in some integration points, the marginal posterior distributions

are obtained:

$$\tilde{p}(x_i \mid \boldsymbol{y}) = \int \tilde{p}(x_i \mid \boldsymbol{\theta}, \boldsymbol{y}) \tilde{p}(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta}\,, \tag{2.20}$$

$$\tilde{p}(\theta_j \mid \boldsymbol{y}) = \int \tilde{p}(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta}_{-j}\,. \tag{2.21}$$

The remaining of the Section deals with these two approximations.

2.6.1 Approximating the joint posterior of the hyperparameters

The first step is to compute the approximation of the joint posterior of hyperparameters $\tilde{p}(\boldsymbol{\theta} \mid \boldsymbol{y})$. Posterior distribution of hyperparameters are not usually Gaussian, reason why Rue et al. (2009) do not use directly the Laplace method, and construct an approximation to

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \frac{p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\theta})}{p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})}\,. \tag{2.22}$$

The approach requires the Gaussian approximation of the denominator, i.e.,

$$\tilde{p}(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \left.\frac{p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\theta})}{\tilde{p}_G(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})}\right|_{\boldsymbol{x}=\boldsymbol{x}(\boldsymbol{\theta})^*}, \tag{2.23}$$

being $\boldsymbol{x}(\boldsymbol{\theta})^*$ the mode of the posterior distribution $p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})$.

Observe that we are approximating $p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})$ using a Gaussian approximation. This approximation makes sense because in most cases $p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})$

is Gaussian or almost Gaussian:

$$\begin{aligned} p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y}) = & \\ &= \frac{p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) p(\boldsymbol{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\boldsymbol{y}, \boldsymbol{\theta})} \\ &\propto p(\boldsymbol{x} \mid \boldsymbol{\theta}) p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) \\ &\propto \exp\left\{-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{x}\right\} \cdot \prod_{i=1}^{n} p(y_i | x_i, \boldsymbol{\theta}) \\ &= \exp\left\{-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{x} + \sum_{i=1}^{n} \log p(y_i | x_i, \boldsymbol{\theta})\right\} \\ &\approx (2\pi)^{-\frac{n}{2}} |\boldsymbol{P}(\boldsymbol{\theta})|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta}))^T \boldsymbol{P}(\boldsymbol{\theta})(\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta}))\right\} \quad (2.24) \end{aligned}$$

where $\boldsymbol{P}(\boldsymbol{\theta}) = \boldsymbol{Q}(\boldsymbol{\theta}) + \text{diag}(\boldsymbol{c}(\boldsymbol{\theta}))$, $\boldsymbol{\mu}(\boldsymbol{\theta})$ is the location of the mode, and the vector $\boldsymbol{c}(\boldsymbol{\theta})$ contains the negative second derivatives of the log-likelihood at the mode with respect to x_i.

2.6.2 Approximating $p(x_i \mid \boldsymbol{\theta}, \boldsymbol{y})$

In this subsection, we explain how the approximations of the posterior density functions of the latent field are computed $\tilde{p}(x_i \mid \boldsymbol{\theta}, \boldsymbol{y})$. Rue et al. (2009) proposed three differentes alternatives:

- **Gaussian approximations.** This is the fastest way to do so. It consists on using the previous joint posterior distribution approximation $\tilde{p}_G(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})$ to compute the marginals,

  $$\tilde{p}(x_i \mid \boldsymbol{\theta}, \boldsymbol{y}) \approx \mathcal{N}(\mu_i(\boldsymbol{\theta}), \sigma_i^2(\boldsymbol{\theta}))\,,$$

  being $\mu_i(\boldsymbol{\theta})$ the marginals means, and $\sigma_i^2(\boldsymbol{\theta})$ the marginals variances. This approximation often gives reasonable results, but there can be error in the location and error due to the lack of skewness.

- **Laplace approximations**. Rue et al. (2009) proposed the use of the Laplace method (subsection 2.5.3) in order to get a better accuracy.

In particular,

$$\tilde{p}(x_i \mid \boldsymbol{\theta}, \boldsymbol{y}) \approx \left. \frac{p(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})}{\tilde{p}_{GG}(\boldsymbol{x}_{-i} \mid x_i, \boldsymbol{\theta}, \boldsymbol{y})} \right|_{\boldsymbol{x}_{-i} = \boldsymbol{x}^*_{-i}(x_i, \boldsymbol{\theta})}, \tag{2.25}$$

being $\boldsymbol{x}_{-i}$ the vector $\boldsymbol{x}$ with its i-th element excluded, $p_{GG}(\boldsymbol{x}_{-i} \mid x_i, \boldsymbol{\theta}, \boldsymbol{y})$ the Gaussian approximation to $\boldsymbol{x}_{-i} \mid x_i, \boldsymbol{\theta}, \boldsymbol{y}$ and $\boldsymbol{x}^*_{-i}(x_i, \boldsymbol{\theta})$ the mode configuration. Observe that expression (2.25) implies that $\tilde{p}_{GG}$ must be recomputed for each value of x_i and $\boldsymbol{\theta}$, since its precision matrix depends on x_i and $\boldsymbol{\theta}$. This makes the algorithm more and more expensive.

- **Simplified Laplace approximation**. In this approximation the numerator and the denominator of the expression (2.25) are expanded up to third order. This provides a correction for skewness in the approximation. The simplified Laplace gives the better trade of between accuracy and computational speed.

2.6.3 Joining all the pieces together

Once $\tilde{p}(x_i \mid \boldsymbol{\theta}, \boldsymbol{y})$ and $\tilde{p}(\boldsymbol{\theta} \mid \boldsymbol{y})$ are computed, the marginal posterior distributions $p(x_i \mid \boldsymbol{y})$ and $p(\theta_i \mid \boldsymbol{y})$ are approximated as it is pointed out in expressions (2.20) and (2.21).

Each marginal posterior $\tilde{p}(\theta_i \mid \boldsymbol{y})$ can be obtained using an interpolation algorithm based on the values of the density $\tilde{p}(\boldsymbol{\theta} \mid \boldsymbol{y})$ evaluated in a set of integrations points $\{\boldsymbol{\theta}^{(j)}\}$. These points are obtained after a grid exploration of $\tilde{p}(\boldsymbol{\theta} \mid \boldsymbol{y})$ (See Rue et al., 2009, for a more detailed description of the method).

Regarding to the marginal posterior of $p(x_i \mid \boldsymbol{y})$, it can be easily obtained through a finite weighted sum:

$$\tilde{p}(x_i \mid \boldsymbol{y}) \approx \sum_j \tilde{p}(x_i \mid \boldsymbol{\theta}^{(j)}, \boldsymbol{y}) \tilde{p}(\boldsymbol{\theta}^{(j)} \mid \boldsymbol{y}) \Delta_j \,, \tag{2.26}$$

for the same relevant integration points $\{\boldsymbol{\theta}^{(j)}\}$ with a corresponding set of weights $\{\Delta_j\}$.

Until now, we have explained the basic ideas concerning the INLA methodology and how to deal with GMRFs makes this methodology extremely fast. However, continuous spatial models can not be treated directly in this methodology. An extension was needed to fit those models. This extension was proposed by Lindgren et al. (2011), and the next chapter is devoted to explain it.

2.7 Model selection

In the Bayesian context, there are different ways to evaluate and compare models in order to get that one that best represent the phenomenon of interest (Schwarz et al., 1978; Geisser, 2013; Berger and Pericchi, 1996; Spiegelhalter et al., 2002; Watanabe, 2010; Vehtari and Ojanen, 2012; Gelman et al., 2014). But moreover, practitioners also want their models to have good predictive properties. Then, it is necessary to evaluate their fitting and predictive accuracy, compare them and select the most appropiate model for our particular data.

In this Thesis, we have usually had the problem of dealing with many environmental covariates and we have proceeded as follows:

1. As we are in the Bayesian context, expert knowledge can be applied to select the more relevant variables. This is the first stage in this process.

2. Once some variables are removed of the analysis, if there are still more than 7 or 8 possible covariates, Pearson correlations are calculated. If the correlation between two variables is greater than 0.7, one of those covariates is taken off the analysis (Dormann et al., 2012). Another alternative in this step is to conduct a principal component analysis. It is worth noting that the number of 7 or 8 appears from the fact that the final number of models analyzed is computationally reasonable.

3. When the number of covariates is reduced and following the method "Best subset selection" proposed by Heinze et al. (2018), all possible models with the different covariates are fitted to posteriorly choose the

best model according to an information criterion such as Deviance Information Criteria (DIC; Spiegelhalter et al., 2002), a log score of the conditional predictive ordinate (CPO; Geisser, 2013) or the Watanabe Akaike Information Criterion (WAIC; Watanabe, 2010). If M represents the number of covariates among which we want to select the best model, then there are 2^M possible models. As above mentioned we have been working with $M = 7$, 8 to keep reasonable computational cost.

4. Posteriorly, if there are more than one model with similar information criterions, the parsimony criterion is applied and models with less amount of covariates are selected.

5. After selecting the best model, the importance of the covariates selected are checked.

Chapter 3

Continuous spatial processes

In Chapter 1 we have pointed out how a model with a continuous spatial component can be formulated as a hierarchichal model. In Chapter 2 we have presented how to perform Bayesian inference in Latent Gaussian Models (LGMs) using the integrated nested Laplace approximation (INLA). This chapter is devoted to describe the common way to deal with Gaussian Fields (GFs) in a continuous space with the INLA methodology. First at all, the main problem in geostatistics is presented, followed by the existing troubles to implement it in INLA. Secondly, the Stochastic Partial Differential Equation (SPDE; Lindgren et al., 2011) approach to solve this problem is depicted, and finally, barrier models as proposed by Bakka et al. (2019) are explained.

3.1 The big n problem

There is no doubt that continuous GFs play an important role in the context of geostatistics (Cressie, 1990). As we have pointed out in Chapter 1, in the context of stationary and isotropic processes, the covariance function is only a function of the Euclidean distance between the locations. Despite of that, a computation of the inverse covariance matrix $\mathbf{\Sigma}^{-1}$ and the determinant of $\mathbf{\Sigma}$ is required to make inference and prediction. Due to the general cost of

$\mathcal{O}(n^3)$ to factorize dense $n \times n$ covariance matrices, these calculations can become very expensive when the number of locations increases. In some cases, it could be also unstable, due to the enormous number of operations required. This problem in the literature is well known as "the big n problem" (Banerjee et al., 2014).

The increasing popularity of the hierarchical Bayesian models has made this issue more important. There have been different approaches to try to solve this problem, but, with the INLA's birth a new age begun. In Chapter 2, it has been hightlighted that INLA exploits the computation properties of Gaussian Markov Random Fields (GMRFs) to fit a wide spectrum of LGMs. Nevertheless, as spatial GFs are continuous and geostatistical data are usually distributed irregularly, INLA can not deal with them directly.

3.2 The SPDE approach for stationary and isotropic processes

In 2011, Lindgren et al. (2011) proposed an alternative approach by using an approximate stochastic weak solution to a SPDE as a GMRF approximation to a GF with Matérn covariance structure. This allowed to reduce computation cost from a magnitude of $\mathcal{O}(n^3)$ to $\mathcal{O}(n^{3/2})$.

In Chapter 2 where the INLA basis were depicted, we saw that the nice properties of INLA are due to the fact that it works with precision matrices which usually are sparse. In line with this, Lindgren et al. (2011) showed how to formulate continuously indexed spatial models with Matérn covariance structure with sparse precision matrices as a weak solution to the following linear fractional SPDE

$$(\kappa^2 - \Delta)^{\alpha/2}(\tau w(\boldsymbol{s})) = \mathcal{U}(\boldsymbol{s}),\ \boldsymbol{s} \in \mathbb{R}^d,\ \alpha = \nu + \delta/2,\ \kappa > 0,\ \nu > 0, \quad (3.1)$$

where Δ is the Laplacian, α controls the smoothness, κ is the scale parameter, τ controls the variance, and $\mathcal{U}(\boldsymbol{s})$ is a Gaussian spatial white noise process. The exact and stationary solution to this SPDE is the stationary

GF $w(\boldsymbol{s})$ with Matérn covariance introduced in Chapter 1

$$\mathcal{C}(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|) = \frac{\sigma^2_{\boldsymbol{w}}}{2^{\nu-1}\Gamma(\nu)}(\kappa\,\|\boldsymbol{s}_i - \boldsymbol{s}_j\|)^{\nu} K_{\nu}(\kappa\,\|\boldsymbol{s}_i - \boldsymbol{s}_j\|). \tag{3.2}$$

According to Lindgren et al. (2011), the empirically derived definition for the range is

$$r = \frac{\sqrt{8\nu}}{\kappa}\,, \tag{3.3}$$

with r corresponding to the distance at which the spatial correlation is close to 0.1, for each $\nu \geq \frac{1}{2}$.

The link between Equations (3.1) and (3.2) is given by the expressions

$$\begin{cases} \nu &= \alpha - \frac{\delta}{2}\,, \\ \\ \sigma^2_{\boldsymbol{w}} &= \frac{\Gamma(\nu)}{\Gamma(\alpha)(4\pi)^{\delta/2}\kappa^{2\nu}\tau^2}\,. \end{cases}$$

But, as people usually work in a two dimensional space, dimension is assumed to be 2, i.e., $\delta = 2$, it follows that

$$\begin{cases} \nu &= \alpha - 1\,, \\ \\ \sigma^2_{\boldsymbol{w}} &= \frac{\Gamma(\nu)}{\Gamma(\alpha)(4\pi)\kappa^{2\nu}\tau^2}\,. \end{cases}$$

In `R-INLA` (`www.r-inla.org`), by default the smoothness parameter α is fixed to 2, corresponding with $\nu = 1$ (Blangiardo and Cameletti, 2015; Lindgren and Rue, 2015). With this assumption, the range is given by

$$r = \frac{\sqrt{8}}{\kappa}\,, \tag{3.4}$$

while the variance is given by

$$\sigma^2_{\boldsymbol{w}} = \frac{1}{4\pi\kappa^2\tau^2}\,. \tag{3.5}$$

With the definition of these new parameters, the Matérn covariance function presented in Chapter 1 and defined in Equation (3.2), can be expressed as

follows:

$$\mathcal{C}(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|) \quad = \quad \sigma^2_{\boldsymbol{w}} \left(\frac{\sqrt{8}}{r} \|\boldsymbol{s}_i - \boldsymbol{s}_j\| \right) K_1 \left(\frac{\sqrt{8}}{r} \|\boldsymbol{s}_i - \boldsymbol{s}_j\| \right). \tag{3.6}$$

In Figure 3.1 we can observe how the Matérn correlation function varies depending on the parameter range.

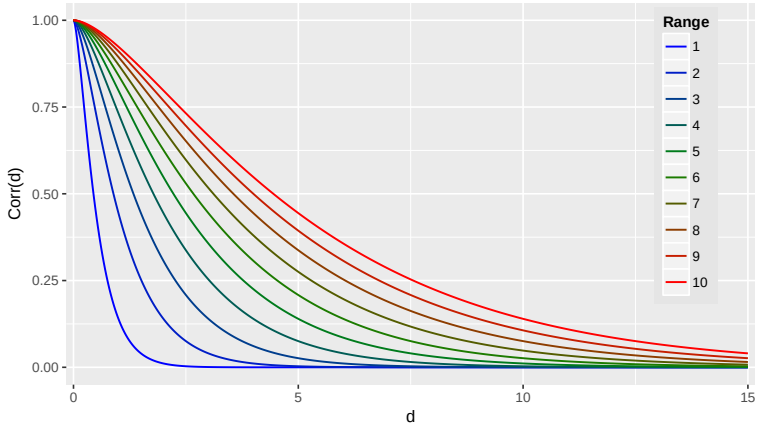

FIGURE 3.1: Representation of the Matérn correlation function for different values of the range.

The solution to the SPDE is approximated using the finite element method (Bathe, 2006) through a basis function representation defined on a Delaunay triangulation (Hjelle and Dæhlen, 2006) of the domain $\mathcal{D}$ (Figure 10.1):

$$w(\boldsymbol{s}) = \sum_{k=1}^{K} \phi_k(\boldsymbol{s}) \tilde{w}_k \,, \tag{3.7}$$

where K is the total number of vertices of the triangulation, $\{\phi_k\}$ is the set of basis functions, and $\{\tilde{w}_k\}$ are zero mean Gaussian distributed weights. The basis functions are defined to take values 1 at vertex k and 0 at all other vertices. This is done to get a Markov structure. With Neumann boundary conditions, the precision matrix obtained $\boldsymbol{Q}$ for the Gaussian weight vector

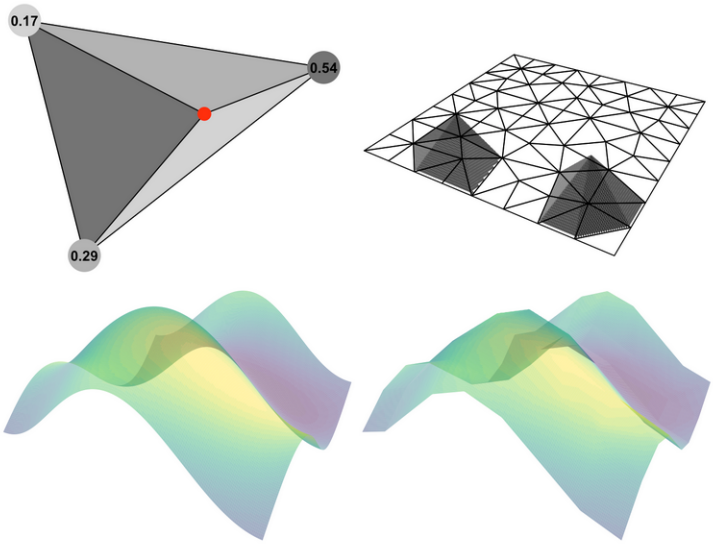

FIGURE 3.2: Image extracted from Krainski et al. (2018): two dimensional approximation illustration. A triangle and the areal coordinates for the point in red (top left). All the triangles and the basis function for two of them (top right). A true field for illustration (bottom left) and its approximated version (bottom right).

$\tilde{w} = \{\tilde{w}_1, \ldots, \tilde{w}_K\}$ is

$$\boldsymbol{Q} = \tau^2(\kappa^4\boldsymbol{C} + 2\kappa^2\boldsymbol{G} + \boldsymbol{G}\boldsymbol{C}^{-1}\boldsymbol{G}), \tag{3.8}$$

being $\boldsymbol{C}$ a diagonal matrix with values $C_{ii} = \int \phi_i(\boldsymbol{s})d\boldsymbol{s}$. $\boldsymbol{G}$ is a sparse matrix with elements $G_{ij} = \int \nabla\phi_i(\boldsymbol{s})\nabla\phi_j(\boldsymbol{s})d\boldsymbol{s}$, being ∇ the gradient.

The resulting precision matrix $\boldsymbol{Q}$ is sparse, and its elements depend on τ and κ. Thus $w(\boldsymbol{s})$ represents the approximated solution in a weak sense to the SPDE, and it is a GMRF with mean $\mathbf{0}$ and precision matrix $\boldsymbol{Q}$ (for a detailled description of the solution, see Bakka et al., 2018; Krainski et al., 2018).

3.3 Non-stationary Gaussian processes

Until now, all the methodology presented regarding to spatial statistics has focused above all in stationary and isotropic Gaussian processes. This is an assumption so common in the context of SDMs (Pennino et al., 2013; Paradinas et al., 2015; Rufener et al., 2017). However, there are some cases where barriers exist, for example islands in the sea, mountains on land or buildings in a city. These barriers block the spread of a species or a disease. They not only break the stationarity of a GF, but also the continuity. In this section, we focus on the approximation by Bakka et al. (2019), where a model that takes into account the non-stationarity situations is presented.

In the case of the stationary GFs, the shortest Euclidean distance between two locations is the measure of interest. Nevertheless, when barriers exist, it does not make sense. Bakka et al. (2019) propose an approximation where in place of thinking about these shortest Euclidean distances, they select a collection of all possible paths from one location to another. Then, the dependency with paths crossing barriers are removed. In order to do so, they used the SPDE approach presented by Lindgren et al. (2011) explained in the previous section.

Bakka et al. (2019) present a system of two SPDEs where the first differential equation is devoted to model the spatial process as it is a stationary and isotropic GF, and the second one presents the same Matérn structure, but depending on other range, the range of the barrier area r_b, which is usually fixed.

In order to present the barrier model, a reparametrization of the previous SPDE in terms of the parameters r and $\sigma_{\boldsymbol{w}}$ assuming that $\alpha = 2$, $\nu = 1$ and $\delta = 2$ was conducted to have a better interpretability of the system of SPDEs:

$$w(s) - \nabla \cdot \frac{r^2}{8} \nabla w(s) \;=\; r\sqrt{\frac{\pi}{2}}\sigma_{\boldsymbol{w}}\, \mathcal{U}(s), \text{ for } s \in \Omega_n\,, \tag{3.9}$$

being $\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right)$ the gradient. Thus, the barrier GF is a solution to the system

$$\begin{aligned} w(s) - \nabla \cdot \frac{r^2}{8} \nabla w(s) &= r\sqrt{\frac{\pi}{2}} \sigma_{\boldsymbol{w}}\, \mathcal{U}(s), \text{ for } s \in \mathcal{D}_n\,, \\ w(s) - \nabla \cdot \frac{r_b^2}{8} \nabla u(s) &= r_b\sqrt{\frac{\pi}{2}} \sigma_{\boldsymbol{w}}\, \mathcal{U}(s), \text{ for } s \in \mathcal{D}_b\,, \end{aligned} \tag{3.10}$$

where $\mathcal{D}_n$ is the normal area and $\mathcal{D}_b$ is the barrier area. Their joint union gives the whole study area $\mathcal{D}$. This system represents a local averaging of nearby values. If there are two points separated by a barrier, the very small range stops the local averaging on the barrier. It forces the dependency to focus on moving around the barrier, via local averages in the non barrier area.

The system is again solved by constructing a Delaunay triangulation of the study area and applying the finite element method. To do it, Bakka et al. (2019) reformulate Equations (3.10) as:

$$\begin{aligned} \left(1 - \nabla \frac{r(\boldsymbol{s})^2}{8} \nabla\right) w(s) &= r(\boldsymbol{s}) \sqrt{\frac{\pi}{2}}\, \mathcal{U}(s)\,, \\ r(\boldsymbol{s}) &= r_q \text{ on } \mathcal{D}_q\,,\ q = 1, 2\,, \end{aligned} \tag{3.11}$$

with $q = 1$ representing the normal area and $q = 2$ the barrier area. The domain $\mathcal{D}$ is a disjoint union of $\mathcal{D}_1$ and $\mathcal{D}_2$ with Neumann boundary condition on $\partial\mathcal{D}$. Again, the spatial field approximation can be written as

$$w(\boldsymbol{s}) = \sum_{k=1}^{K} \phi_k(\boldsymbol{s}) \tilde{w}_k\,, \tag{3.12}$$

and defining

- $\boldsymbol{J}$ a matrix whose elements are $J_{ij} = \int \nabla\phi_i(\boldsymbol{s}) \nabla\phi_j(\boldsymbol{s}) d\boldsymbol{s}$,
- $\boldsymbol{P}_q$ with elements $(P_q)_{ij} = \int_{\mathcal{D}_q} \nabla\phi_i(\boldsymbol{s}) \nabla\phi_j(\boldsymbol{s}) d\boldsymbol{s}$,
- $\tilde{\boldsymbol{C}}_q$ a diagonal matrix $(\tilde{C}_q)_{i,i} = \int_{\mathcal{D}_q} \phi_i(\boldsymbol{s}) d\boldsymbol{s}$,
- $\boldsymbol{A} = \boldsymbol{J} - \frac{1}{8}\left(r_1^2 \boldsymbol{D}_1 + r_2^2 \boldsymbol{D}_2\right)$, and

- $\tilde{\boldsymbol{C}} = \frac{\pi}{2}\left(r_1^2\tilde{\boldsymbol{C}}_1 + r_2^2\tilde{\boldsymbol{C}}_2\right)$.

the resulting precision matrix $\boldsymbol{Q}$ depending on the r and $\sigma_{\boldsymbol{w}}^2$ is obtained:

$$\boldsymbol{Q} = \boldsymbol{A}\tilde{\boldsymbol{C}}^{-1}\boldsymbol{A}\,. \tag{3.13}$$

3.4 SDMs as LGMs with a continuous GF

This section is devoted to depict how we can specify a SDM as a LGM with a continuous GF. As we know by the previous chapter, a LGM is a hierarchical Bayesian model whose latent field is a GMRF. We specify the likelihood in the first stage, followed by the latent variables and random components, and finishing with the specification of the hyperparameters. Then the model with a continuous GF can be formulated as follows:

Likelihood

$$\begin{aligned} y_i \mid \boldsymbol{\beta}, w_i &\sim p(y_i \mid \eta_i) \\ \eta_i &= \boldsymbol{V}_i\boldsymbol{\beta} + \boldsymbol{w}_i \end{aligned}$$

GMRF

$$\begin{aligned} \beta_0, \ldots, \beta_M &\sim \mathrm{N}(0, \tau_0^{-1}) \\ \boldsymbol{w} &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}^{-1}(r, \sigma_{\boldsymbol{w}}^2)) \end{aligned}$$

Hyperparameters

$$r,\ \sigma_{\boldsymbol{w}}^2\,.$$

Observe that the matrix represented by $\boldsymbol{Q}(r, \sigma_{\boldsymbol{w}}^2)$ is the approximated precision matrix of the continuous GF that we have previously explained for the stationary and isotropic case, and the barrier case.

Lastly, hyperpriors (priors for the hyperparameters) of r and $\sigma_{\boldsymbol{w}}^2$ need to be assigned. In this Thesis, we specify them in two different ways: using basis functions as presented in Lindgren and Rue (2015), and in the most recent works, we use the penalize complexity priors (Fuglstad et al., 2018).

Chapter 4

Goals developed and Results

Up to this point, all the tools required to develop this Thesis have been described. This chapter is devoted to give a more detailed explanation of each of the objectives of this Thesis mentioned in Chapter 1, doing an introduction to the problem, describing concisely the methodology employed, and finally depicting the results.

4.1 Objective 1: SDMs in plant disease epidemiology, and marine and vegetal species distribution

As previously presented, the first goal of this Thesis was to apply SDMs to analyze practical problems of plant disease epidemiology, and marine and vegetal species distribution. Methodology described in Chapters 2 and 3 has been used by considering all the models as LGMs and applying INLA along with the SPDE when necessary. The objective 1.1 was modeled by means of a beta likelihood, some covariates and an unestructured random effect, while objective 1.2 required a Bernoulli likelihood and a continuous GF jointly with some covariates. In objective 1.3, a Bayesian spatial beta regression was conducted, while objective 1.4 required a LGM with Bernoulli

likelihood and a barrier spatial effect. In what follows we present a more detailed description of these four objectives.

4.1.1 Objective 1.1: modeling the production of *Plurivorosphaerella nawae* ascospores in persimmon leaf litter

Circular leaf spot (CLS), caused by *Plurivorosphaerella nawae*, is a serious disease of persimmon (Diospyros kaki) inducing necrotic lesions on leaves, defoliation and fruit drop. The disease was detected in semi-arid areas in Spain in 2008. Under Mediterranean conditions, *P. nawae* forms pseudothecia in leaf litter during winter and ascospores are released in spring infecting susceptible leaves. Persimmon growers in Spain are advised to apply fungicides for the control of circular leaf spot during the period of inoculum availability, which was defined based on ascospore counts under the microscope. Fungicide programs are effective for CLS control only when they coincide with the infection period, with the presence of ascospores, adequate environmental conditions and susceptible leaves. Then, in order to assist growers in scheduling fungicide sprays, a model of potential inoculum availability of *P. nawae* was developed and evaluated.

Samples of leaf litter were collected weekly in L'Alcudia (Valencia, Spain) from 2010 to 2015. Leaves were soaked, placed in a wind tunnel, and released ascospores of *P. nawae* were counted under the microscope. Proportions of released ascospores per year were computed. Environmental data were monitored hourly in each orchard with an automated meteorological station including sensors for temperature, relative humidity and rainfall. Following Rossi et al. (2009) time was expressed in physiological units. To do so, three different variables were calculated: accumulated degree days (ADD), ADD taking into account the vapor pressure deficit ($ADDvpd$) and ADD taking into account both the vapor pressure deficit and the rainfall ($ADDwet$).

Hierarchical Bayesian beta regression methods were used to fit the dynamics of ascospore production in the leaf litter. As we explained in Chapter 1, the beta likelihood does not belong to the exponential family, but it belongs to the LGM class of models. In addition to the physiological

variables, a random effect year was included in the models. Following the steps explained in Chapter 3 about the selection process, models covering all possible combinations of climatic explanatory variables and the random effect were fitted using the INLA methodology. The best was selected based on the WAIC and LCPO.

Results showed that *ADD* and *ADDvpd* jointly with the random effect year best described the dynamics of production of *P. nawae* ascospores. The resulting best model is about to be implemented in a disease warning system to schedule fungicide sprays for the control of circular leaf spot in Spain.

This work has been presented in the following paper which will be submitted to an indexed journal as soon as the warning system is finished. The paper is fully presented in Chapter 5 of this Thesis.

- **J. Martínez-Minaya, D. Conesa, A. López-Quílez, J.L. Mira and A. Vicent (2019). Bayesian Beta regression for modelling potential inoculum availability of *Plurivorosphaerella nawae* in persimmon leaf litter.**

4.1.2 Objective 1.2: study of the spatial and climatic factors associated with the distribution of citrus black spot disease

Citrus black spot (CBS) is the main fungal disease affecting citrus crops, and it is caused by the fungus *Phyllosticta citricarpa*. The Mediterranean Basin is free of the disease and thus phytosanitary measures are in place to avoid the entry of *P. citricarpa* in the EU territory. However, the suitability of the climates present in the Mediterranean Basin for CBS establishment and spread is debated.

Two different georeferenced datasets of CBS presence/absence in citrus areas were assembled for the stages of the epidemic 1950 and 2014 in South Africa. Climatic variables were obtained from the WorldClim database.

In order to fulfill the objective, two studies were conducted:

4.1.2.1 A historical analysis of the disease spread in South Africa

In this study, a historical analysis of disease spread in South Africa was done. Köppen-Geiger climate classification system (Köppen and Geiger, 1936) based on the updated version from Peel et al. (2007), and the Aschmann Mediterranean-type climate (Aschmann, 1973) using the gridded data from WorldClim were implemented.

To test the hypothesis that CBS presence occurs at random among grid cells, which should be considered before carrying out further advanced modelling studies, Moran's I and Geary's C analyses of spatial autocorrelation were used. For both indices, contiguity-based neighbours were defined in grid cells sharing edges or vertices.

Results showed that in 1950, CBS was still confined to areas of temperate climates with summer rainfall (Cw, Cf), but spread afterwards to neighbouring regions with markedly drier conditions. The hot arid steppe (BSh) is the predominant climate where CBS develops in South Africa. The disease was not detected in the Mediterranean-type climates Csa and Csb as defined by the Köppen-Geiger system and the more restrictive Aschmann's classification criteria. However, arid steppe (BS) climates, where CBS is prevalent in South Africa, are common in important citrus areas in the Mediterranean Basin.

The most noticeable change in the environmental range occupied by CBS in South Africa was the amount and seasonality of rainfall. Due to the spread of the disease to dryer regions, the minimum annual precipitation in CBS-affected areas declined from 663 mm in 1950 to 339 mm at present. The minimum value precipitation of warmest quarter also declined from 290 to 96 mm.

Strong spatial autocorrelation in CBS distribution data was detected being the Moran's I equal to 1 with $p - value < 0.0001$, and the Geary's C equal to 0 with $p - value < 0.0001$.

Three conclusions were extracted from this work:

- CBS in South Africa has expanded from its original geographic range in summer rainfall areas to adjacent, more arid regions.
- The results contradict statements indicating that CBS occurs exclusively in climates with summer rainfall.
- Further modeling studies were required to integrate the relative contribution of environmental variables and the spatial structure of the data.

This work was published in the following paper and is presented in Chapter 6:

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

After this publication, Fourie et al. (2017) published a Scientific critique to this paper. They concluded that the Martínez-Minaya et al. (2015) study relied on an approach that grossly overestimates the extent of the geographical area that could support *P. citricarpa*.

In 2017, we published a new letter refuting all the arguments presented in Fourie et al. (2017). The resulting following paper is presented in Chapter 7:

- **J. Martínez-Minaya, D. Conesa, A. López-Quílez and A. Vicent. Response to the letter on "Climatic distribution of citrus black spot caused by *Phyllosticta citricarpa*. A historical analysis of disease spread in South Africa" by Fourie et al. (2017). *European Journal of Plant Pathology*, 148, 503–508.**

After publishing this letter, the discussion concluded and the results obtained in the first paper were reafirmed. Indeed, the former paper (Martínez-Minaya et al., 2015) was selected for the European Food Safety Authority

(EFSA) Panel on Plant Health (PLH) in order to update the EFSA Scientific Opinion on the risk of *P. citricarpa* (EFSA, European Food Safety Authority, 2016).

4.1.2.2 A Bayesian latent Gaussian model approach to the distribution of CBS

The role of climate as a limiting factor for the establishment and spread of CBS to new areas had been widely debated, but previous studies did not address the effects of spatial factors in the geographic distribution of the disease.

In this work Moran's I and Geary's C were computed using different distances for the neighbor relationships. LGMs with Bernoulli response were fitted to analyze CBS presence/absence in 1950 or 2014 with the INLA methodology. Due to the high collinearity of the environmental covariates, principal components (PCs) or pre-selection of climatic variables based on their correlation coefficients were used. A continuous GF as explained in Chapter 3 was incorporated in the model.

In order to select the best model representing the phenomenon and again following the steps in Chapter 2, models including a selection of climatic explanatory variables with Pearson correlation (in absolute values) smaller than 0.7 or PCs were fitted to the response variable (CBS presence/absence). Models covering all possible combinations of these climatic explanatory variables or PCs were compared using the WAIC. Moreover, the geostatistical spatial term was incorporated into these models and the corresponding WAIC was calculated.

A validation dataset with CBS-present and CBS-absent grid cells was assembled by random sampling without replacement from the 2014 dataset, but excluding those grid cells used for model development in 1950. ROC curve analysis was used to evaluate the predictive ability of the models selected for the 1950 dataset.

Moran's I and Geary's C analyses indicated the presence of significant spatial autocorrelation in CBS distribution data in 1950 and 2014. Both

indices showed that spatial autocorrelation was stronger in 2014 than in 1950. Regarding to the fitted models, they indicated a positive relationship between CBS presence and climatic variables or PCs associated with warm temperatures and high precipitation. Nevertheless, in 1950, models that also included a spatial effect outperformed those with climatic variables only. Problems of model convergence were detected in 2014 due to the strong spatial structure of CBS distribution data.

As a conclusion, although climate was advocated as the main factor limiting the establishment and spread of CBS into new areas, our study indicates that spatial proximity to affected areas was also relevant in the geographic distribution of the CBS.

This work was published in the following paper and constitutes Chapter 8 of this Thesis:

- **J. Martínez-Minaya, D. Conesa, A. López-Quílez and A. Vicent (2018). Spatial and climatic factors associated with the geographical distribution of citrus black spot disease in South Africa. A Bayesian latent Gaussian model approach. *European Journal of Plant Pathology*, 151, 991–1007.**

4.1.3 Objective 1.3: analysis of the effects of geographic genetic structure and spatial autocorrelation on species distribution range shifts

Global climate change (GCC) is a change in the pattern of weather, and related changes in oceans, land surfaces and ice sheets, occurring over time scales of decades or longer. It is clear that it is dramatically affecting the distribution of many terrestrial, aquatic and marine organisms. Therefore, multiple efforts are currently focused on the development of models to better predict distribution range shifts due to GCC. In this paper, we proposed a different way to model range shifts by including intraspecific genetic structure and spatial autocorrelation (SAC) of data in distribution range models.

The Iberian Peninsula was used as a study area. A collection of 301 geo-referenced populations of the annual plant *Arabidopsis thaliana* during

12 years between 2000 and 2011 were employed. One representative indivual (accesion) for population was used to analyse the genetic structure. In previous studies, four genetic clusters had been inferred (Marcer et al., 2016). Finally, for each accesion, the membership proportion to each of the four genetic clusters were obtained. With these data, we wanted to depict current and future distribution ranges for the four genetic clusters. Again the WorldClim database was employed to get climatic variables.

With regard to the methodology, Bayesian spatial and non spatial beta regression were constructed as we have previously explained. Also, in order to compare with other methods, Maxent method was also applied. However, as Maxent is only able to model presences, response variable had to be transformed resulting in a loss of information.

In order to select the best model representing the phenomenon, we proceeded in a similar way to CBS: models including a selection of climatic explanatory variables with Pearson correlation (in absolute values) smaller than 0.7 were fitted to the response variable (membership proportion) for each genetic cluster. Models covering all possible combinations of these selected climatic explanatory variables and including the spatial effect were compared using the WAIC and LCPO. Moreover, mean absolute error MAE and root mean square error (RSME) of the best models were computed.

In order to predict, we used the Representative Concentration Pathways (RCPs) which are four greenhouse gas concentration trajectories adopted by the Intergovernmental Panel on Climate Change (IPCC). They describe four possible climate futures, all of which are considered possible depending on how much greenhouse gases are emitted in the years to come (van Vuuren et al., 2011). They describe four possible climate futures, although we selected the two most extremes in the year 2070:

- **RCP 2.6**. It assumes that global annual emissions peak between 2010-2020, with emissions declining substantially thereafter.
- **RCP 8.5**. Emissions continue to rise throughout the 21st century.

We obtained that spatial Beta regression models selected less bioclimatic predictors than non-spatial models and Maxent models to define the distribution range of the four genetic clusters. In addition, the MAE and RMSE were lower for spatial than for non-spatial models for all genetic clusters in which te comparison was possible. This indicated that spatial models had lower average model prediction errors in the response variable.

With regard to the distribution range shifts with GCC, Maxent models and Beta regression models were also used to quantify distribution range shifts of all *A. thaliana*'s genetic clusters with different GCC models and scenarios. The three modeling approaches yielded different GCC predictions for each genetic cluster based on suitability shifts in distribution range projections.

For genetic cluster C1, important reductions in distribution range were predicted for the two GCC scenarios with Maxent and non-spatial beta regression models, whereas spatial beta regression models predicted slight increases in distribution range. For genetic cluster C2, Maxent predicted increasing and decreasing distribution ranges with RCP 2.6 and RCP 8.5, respectively, whereas both beta regression models predicted slight fluctuations in distribution range in both GCC scenarios. For genetic cluster C3, Maxent showed very high increases in distribution range, particularly for the RCP 8.5 scenario, whilst non-spatial beta regression models also predicted slight fluctuations in distribution range in both GCC scenarios. Finally, for genetic cluster C4, all approaches predicted increases in distribution range in both GCC scenarios. Maxent gave higher increases in RCP 2.6 than in RCP 8.5 and vice-versa for both beta regression models.

To conclude, Maxent and non-spatial beta regression models presented some drawbacks, such as the loss of accessions with high genetic admixture in the case of Maxent, and the presence of residual SAC for both. Spatial beta regression models removed residual SAC, showed higher accuracy than non-spatial beta regression models, and handled the spatial effect on model outcomes. We concluded that these Hierarchical beta regression models enrich the toolbox of software available to evaluate GCC-induced distribution range shifts considering both geographic genetic heterogeneity and SAC.

This work was depicted in the following paper which has been accepted in the journal *Molecular Ecology Resources*. The paper is fully presented in Chapter 9 of this Thesis.

- **J. Martínez-Minaya, D. Conesa, C. Alonso-Blanco, M.J. Fortin, X. Picó and A. Marcer (2019). A hierarchical Bayesian Beta regression approach to study the effects of geographic genetic structure and spatial autocorrelation on species distribution range shifts.**

4.1.4 Objective 1.4: study of the bottlenose dolphin (*Tursiops truncatus*) distribution

Worldwide, cetacean species have started to be protected, but they are still very vulnerable to accidental damage from an expanding range of human activities at sea. To properly manage these potential threats, a detailed understanding of the seasonal distributions of these highly mobile populations is neccessary. To achieve this goal, a growing effort has been underway to develop species distribution models (SDMs) that correctly describe and predict preferred species areas.

However, accuracy is not always easy to achieve when physical barriers, such as islands, are present. Indeed, as we have already explained, SDMs assume, if only implicitly, that the spatial effect is stationary, and that correlation is only dependent on the distance between observations and not on the direction or a spatial coordinates. The application of stationary SDMs in these cases could lead to incorrect predictions and, consequently, to uninformed decision making. In this study, we identified vulnerable habitats for the bottlenose dolphin in the Archipelago de La Maddalena, Northern Sardinia (Italy) using Bayesian hierarchical SDMs that account for the physical barriers issue as explained in Chapter 3 and provide a full specification of the associated uncertainty.

The study was conducted in waters withing 3 miles off the coast of Archipielago de La Maddalena, Northern Sardinia (Italy). Random transects were performed from October 2007 to September 2008. Surveys were

conducted by experts during light hours from 6.00 A.M. to 8.00 P.M., and to identify species, observers scanned with both the naked eye and binoculars. Geographical information were collected, jointly with presence/absence of the species. Environmental variables were collected from the aqua-MODIS sensor with a resolution of $2km$ (`https://modis.gsfc.nasa.gov/`).

The methodology here applied to estimate and predict overall occurrence of bottlenose dolphins with respect to environmental is the one proposed by Bakka et al. (2019) and previously explained in depth in the Chapter 4. It basically consists on a spatial hierarchical Bayesian model where the spatial component is estimated as a solution of two differential equations which take into account the non-stationarity of the effect. Again, different models were fitted, and in order to find the best the WAIC and LCPO were used.

Results showed that dolphin occurrence in the Archipelago de La Maddalena was influenced by a seasonal effect in the area. In addition, it showed that estimated dolphing occurence was higher during the winter season. On the other hand, the spatial component seemed to reflect disturbance from pleasure boating. Thus, an effective conservation programme should take into account these findings: favourable areas for bottlenose dolphins should be identified and protected as SACs (Special Areas of Conservation). Protection measures should be devoted to limiting the disturbance from recreational boats, which is probably the main threat for this species in the area.

In conclusion, we proposed an approach which constitutes a major step forward in the understanding of cetacean species in many ecosystems where physical, geographical and topographical barriers are present.

This work was described in the following paper which has been accepted in the journal *Ecological Modelling*. The paper is fully presented in Chapter 10 of this Thesis.

- **J. Martínez-Minaya, D. Conesa, H. Bakka and M.G. Pennino (2019). Dealing with physical barriers in bottlenose dolphin *Tursiops truncatus* distribution.**

4.2 Objective 2: developing new methodological tools to solve statistical problems appeared in the application of SDMs

In the second part of this Thesis we have focused in the development of new statistical tools required in SDMs after performing an extensive study of the state of the art of revelant statistical issues for SDMs. In particular, we provide some advances in the compositional data context.

4.2.1 Objective 2.1: a review with the focus in the statistical issues in Species Distribution modeling

As we have brought to light during this Thesis, the use of complex statistical models has recently increased substantially in the context of SDMs. In line with this, an important objective of this Thesis has been to review some of the statistical challenges that can arise when the distribution of the species is modeled using geostatistical or point-referenced data. In particular, we reviewed the different sources of information and different approaches (frequentist and Bayesian) to model the distribution of a species. In the context of Bayesian inference, we discussed the importance of the INLA methodology under the assumption of Gaussian fields and hierachical modeling, in order to compute the marginal posterior distributions of the parameters involved in these kind of models. We finally discussed some important statistical issues that arise when researchers use species data: the presence of temporal autocorrelation in the model presenting different spatial and spatio-temporal structures, the problem of collecting data in SDMs using preferential sampling, the spatial misaligment, the non-stationarity, the imperfect detection and the excess of zeros.

We conclude that INLA is a powerful tool to deal with SDMs making it possible to perform complex models with a minimum computational effort while obtaining accurate estimates.

This work was published in the following paper and constitutes Chapter 11 of this Thesis:

- **J. Martínez-Minaya, M. Cameletti, D. Conesa and M.G. Pennino (2018). Species distribution modeling: a statistical review with focus in spatio-temporal issues. Stochastic Environmental Research and Risk Assessment, 32, 3227–3244**.

4.2.2 Objective 2.2: implementing Bayesian Dirichlet regression in the context of the integrated nested Laplace approximation

Compositional data (Aitchison and Egozcue, 2005), consisting of proportions or percentages of disjoint categories adding to one, play an important role in many fields such as ecology, geology, biology, etc. Dirichlet regression models are commonly used to analyse this kind of data relating them with covariates of interest. From a Bayesian perspective, it has been implemented in methodologies such as `Jags` or `BayesX`. However, in the context of `R-INLA` it has not been yet, as, in general, `R-INLA` can not deal with multivariate likelihoods. In this work, we propose an expansion of the INLA method for fitting Dirichlet linear regression, giving a theoretical foundation and describing the implementation as well as the use of the method. All these findings have been implemented in an `R` package called `dirinla`, which will be available soon.

The main idea of this approach is to approximate the effect of the log likelihood on the posterior using the Laplace method. After applying it, the multivariate initial observations are turned into independent Gaussian pseudo-observations which `R-INLA` can deal with. This method can be used for each multivariate likelihood whose second derivatives exist.

Different simulation studies and an application to a real example were conducted in order to show the reliability of the method. Results were compared with the software `R-jags` and they showed a good performance not only in accuracy, but also in terms of speed of calculations.

To summ up, we have applied the Laplace method in order to find a method which is able to deal with multivariate response, in particular with a Dirichlet likelihood. Apparently, it works fine, but there is still work to

do, adding the chance to deal with random effects and trying to apply it to other multivariate likelihoods.

This completed work has been depicted in the following paper which will be submitted to an indexed journal as soon as the package is finished. The paper is fully presented in Chapter 12 of this Thesis.

- **J. Martínez-Minaya, F. Lindgren, A. López-Quílez, D. Simpson and D. Conesa (2019). Modeling Dirichlet likelihoods using the integrated nested Laplace approximation (INLA).**

Chapter 5

Bayesian Beta regression for modelling potential inoculum availability of *Plurivorosphaerella nawae* in persimmon leaf litter

In this chapter, we present the actual version of our paper "Bayesian Beta Regression for Modelling Potential Inoculum Availability of *Plurivorosphaerella nawae* in Persimmon Leaf Litter" by Joaquín Martínez-Minaya (University of Valencia), David Conesa (University of Valencia), Antonio López-Quílez (University of Valencia), José Luis Mira (Valencian Institute for Agricultural Research) and Antonio Vicent (Valencian Institute for Agricultural Research). In order to keep the same structure of the chapters with published papers, this chapter ends with the references used in this work.

Abstract

Circular leaf spot, caused by *Plurivorosphaerella nawae*, is a serious disease of persimmon (*Diospyros kaki*) inducing necrotic lesions on leaves, defoliation and fruit drop. The disease was initially restricted to humid regions in Japan and Korea and in 2008 it was detected in semi-arid areas in Spain.

Under Mediterranean conditions, *P. nawae* forms pseudothecia in leaf litter during winter and ascospores are released in spring infecting susceptible leaves. Persimmon growers in Spain are advised to apply fungicides for the control of circular leaf spot during the period of inoculum availability, which was defined based on ascospore counts under the microscope. In order to assist growers in scheduling fungicide sprays, a model of potential inoculum availability of *P. nawae* was developed and evaluated. Samples of leaf litter were collected weekly in L'Alcudia from 2010 to 2015. Leaves were soaked, placed in a wind tunnel, and released ascospores of *P. nawae* were counted. Hierarchical Bayesian beta regression methods were used to fit the dynamics of ascospore production in the leaf litter. Results showed that accumulated degree days and accumulated degree days taking into account the vapor pressure deficit best described the dynamics of *P. nawae* ascospores. The resulting best model is being implemented in a disease warning system to schedule fungicide sprays for the control of circular leaf spot in Spain.

Keywords

Mycosphaerella nawae, INLA, warning system, accumulated degree days

5.1 Introduction

Circular leaf spot (CLF) disease of persimmon (Diospyros kaki Thunb.), caused by *Plurivorosphaerella nawae* (= *Mycosphaerella nawae*), induces necrotic lesions on leaves, chlorosis and defoliation. The presence of foliar lesions and premature leaf drop induce early fruit maturation and abscission, resulting in serious economic losses (Bassimba et al., 2017). The disease was first described in humid areas in Japan and later in Korea (Ikata and Hitomi, 1929; Kang et al., 1993). The detection of CLS in Easter Spain was the first report of the disease in a semi-arid area (Vicent et al., 2012).

The fungus forms pseudothecia in leaf litter during winter and ascospores are produced as temperatures increase in spring (Kang et al., 1993). Ascospores are wind-dispersed and infect persimmon leaves in the presence of

a film of water and adequate temperatures. The main infection period in Korea was from mid-May to the end of July (Kang et al., 1993; Kwon and Park, 2004) and from the beginning of April to early July in Spain (Vicent et al., 2012). The asexual stage of *P. nawae* was identified in Korea as belonging to the genus *Ramularia*, but its role in field epidemics is not fully understood (Kwon et al., 1998; Kwon and Park, 2004). In Spain, this secondary inoculum has not been observed (Vicent et al., 2012). The disease is characterised by a long incubation period of about 4 months (Kwon and Park, 2004; Vicent et al., 2012).

Fungicide schedules for the control of CLS in Korea consist of three to four foliar applications during the critical infection period between mid-July to early August. Although the efficacy of fungicide programs may differ depending on the year, nearly complete disease control was obtained under experimental conditions (Kwon et al., 1998; Kwon and Park, 2004). In Spain, two to four fungicide applications during the infection period in spring showed also good efficacy for the control of CLS, whereas post-infection sprays were ineffective (Bassimba et al., 2017; Berbegal et al., 2013). Cultural practices, such as leaf litter removal and moving from flood to drip irrigation systems, are also recommended to growers but their efficacy has not been quantified so far (Vicent et al., 2011, 2012).

Fungicide programs are effective for CLS control only when they coincide with the infection period, with the presence of ascospores, adequate environmental conditions and susceptible leaves. The presence of airborne ascospores is typically monitored using spore traps, either active volumetric or passive (West and Kimber, 2015). Nevertheless, the predictive ability of spore traps is somehow limited because they only detect the ascospores when already released in the orchard air. In the case of *P. nawae* in Spain, monitoring ascospore production in the leaf litter allowed to predict ascospore release 1–2 weeks in advance, so this method is routinely used to schedule fungicide sprays for CLS control (Vicent et al., 2012). Samples of leaf litter are collected weekly in affected persimmon orchards and soaked in distilled water. Immediately after soaking, leaves are placed in a wind tunnel until they dry. Ascospores released from the leaf litter are collected on glass microscope slides and counted under the microscope (Vicent et al., 2011). Although this method proved to be useful, it is time and resource

consuming, requires specific laboratory equipment as well as trained personnel. Consequently, the extent of the area to be monitored and the density of sampling network are rather limited.

Models for inoculum maturation in the leaf litter have been developed for some ascomycetes, as a more efficient alternative to direct observations of ascospore production and release (De Wolf and Isard, 2007). For instance, Gadoury et al. (1982) proposed a linear regression with a probit previous transformation to the proportion of ascospore discharge, Villalta et al. (2001) depicted a linear regression with a logit previous transformation; Rossi et al. (2009) and Eikemo et al. (2011) compared linear regressions with asymptotic, monomolecular, logistic and Gompertz transformations. Nevertheless, most of these models rely on previous transformations of the response variable and then fitted as a linear regression (Luley and McNabb Jr, 1991; Spotts et al., 1994) or fitting directly a nonlinear regression (Navas-Cortés et al., 1998b; Rossi et al., 1999; Cooley et al., 2007; Legler et al., 2014). However, as the proportion of discharge ascospores is being modeled, there are other modelling methods available such as the beta regression model, firstly introduced by Ferrari and Cribari-Neto (2004). Basically, this methodology consists on assuming that the response variable conditioned to the linear predictor follows a beta distribution which is depending on two parameters: a mean and a precision.

On the other hand, Bayesian hierarchical methods are becoming popular in many fields as they may better address the intrinsic complexity typical in many natural systems (Clark, 2005). In Bayesian inference, parameters are treated as random variables and data are related to model parameters using a likelihood function, getting the posterior distribution by combining the prior distribution and the likelihood function. However, getting the posterior distribution is not always straightforward and numerical algorithms are usually required. Markov Chain Monte Carlo (MCMC) methods (Gilks et al., 1996) are widely used to obtain posterior distributions but they involve computationally and time intensive simulations. The Integrated Nested Laplace Approximation (INLA) approach was developed as a computationally efficient alternative to MCMC in latent Gaussian models (Tierney and Kadane, 1986; Rue et al., 2009).

In this work, we propose the use of hierarchical Bayesian beta regression models with random effects to estimate the production *P. nawae* ascospores in persimmon leaf litter using the INLA methodology. These models will assist to predict the dynamics of *P. nawae* inoculum in the orchards based on environmental covariates, without the direct quantification of ascospores in the leaf litter. This will facilitate a wider implementation of a decision support system to optimize the fungicide programs for CLS control in Spain.

5.2 Materials and Methods

5.2.1 Field data

The study was conducted from 2010 to 2015 in a persimmon cv. Rojo Brillante orchard severely affected by CLS at L'Alcúdia in Valencia Province, Spain. Orchard was 11 yr old at the beginning of the study. Trees were grafted on *D. lotus* L. rootstock. Orchard was drip irrigated and with a 5 m across-row spacing and 4 m in-row spacing with a north-south row orientation. Plot size was 0.83 ha at L'Alcúdia. In the center of each orchard, an experimental area of 0.2 hectares (10×10 trees) remained untreated during the 6-yr period of study.

Environmental data was monitored hourly with an automated meteorological station (Hobo U30, Onset Computer Corp.) including sensors for temperature and relative humidity (Hobo S-THB, accuracies $\pm$ 0.2ºC, $\pm$ 2.5%), and rainfall (7852, Davis Instruments Corp, resolution 0.2 mm). Environmental monitors were located at 1.5 m above the soil surface within the row in the center of the experimental area.

Following Rossi et al. (2009) time was expressed in physiological units calculated by three different methods, all of them based on sums of the daily temperatures exceeding 0°C.

In particular, Accumulated Degree Days (ADD) were calculated as:

$$ADD_i = \sum_{j=biofix}^{N(i)} T_j \,, \tag{5.1}$$

where i and j are the subscripts for observations and days respectively, while T_j is the air temperature in each day (calculated as a mean of 24 hourly values). The biofix takes positive values when starting to count in the current year, and negatives when calculated for the previous year.

In second place, Accumulated Degree Days that taking into account the Vapor Pressure Deficit ($ADDvpd$) were calculated:

$$ADDvpd_i = \sum_{j=biofix}^{N(i)} T_j \cdot VPD_j \,, \tag{5.2}$$

being i the observation and j the subscript for days, with $j = biofix$ to $N(i)$, and T_j is the air temperature in each day (daily T was calculated as a mean of 24 hourly values) if $T_j > 0$, elsewhere $T_j = 0$. VPD_j is a dichotomic variable calculated as follows: when vapor pressure deficit $(vpd)_j \leq 4hPa$, $VPD_j = 1$, elsewhere $VPD_j = 0$, being vpd calculated from temperature an relative humidity (rh, %) as follows:

$$(vpd)_j = \left(1 - \frac{rh_j}{100}\right) \cdot 6.11 \cdot \exp\left(\frac{17.47 \cdot T_j}{239 + T_j}\right) . \tag{5.3}$$

Finally, a variable that incorporates information about rainfall was also considered. In particular, it was denoted as $ADDwet$ and as it can be seen by its definition corresponds to the Accumulated Degree Days but taking into account both the vpd and rainfall (R):

$$ADDwet_i = \sum_{j=biofix}^{N(i)} T_j \cdot WET_j \,, \tag{5.4}$$

being i the observation and j the subscript with $j = biofix$ to $N(i)$, and T_j is the air temperature in each day (daily T_j was calculated as a mean of 24 hourly values) if $T_j > 0$, elsewhere $T_j = 0$. WET_j is a dichotomic variable calculated as follows: when $R_j \geq 0.2$mm and $vpd_j \leq 4hPa$, $WET_j = 1$, elsewhere $WET_j = 0$.

The dynamics of *P. nawae* ascospore production in the leaf litter was studied from 2010 to 2015 in the persimmon orchard at L'Alcúdia described

above. Dry leaves on the orchard floor were covered with a plastic mesh (2×2 m^2, 5-by-5-mm openings) fixed with four stainless-steel pins. Plastic nets were located in the center of the experimental area in each orchard without overlying the soil area wetted by the drip irrigation system. Leaf litter density under the plastic nets was adjusted to 350 g of dry leaves m^2 (Vicent et al., 2011). A sample of 20 dry leaves was collected weekly in each orchard and soaked for 15 min in distilled water. Immediately after soaking, leaves were placed with the abaxial surface facing upward in a wind tunnel for 30 min until they were visibly dry (Whiteside, 1974; Vicent et al., 2011). During the process, air and water temperature was maintained at about 21°C.

Discharged ascospores were collected on a glass microscope slide (26×76 mm) coated with silicone oil (Merck). Spores were stained with lactophenol-acid cotton blue and examined at 400X magnification. All ascospores showing the morphological characteristics of *P. nawae*; spindle-shaped, $10-13 \times 3-4$ μm, hyaline, 2-celled with a medium or slightly supramedian septum (Kwon et al., 1998), were counted in four microscope field transects. Isolations were arbitrarily performed each year using additional leaf litter samples and collecting the ejected ascospores in Potato Dextrose Agar (PDA) amended with 0.5 g L^{-1} streptomycin sulphate (PDAS). Identification of the resulting fungal colonies was confirmed using a specific molecular method for *P. nawae* (Berbegal et al., 2013). For each week, the cumulative proportion of ascospores discharged was calculated based on the total collected in each orchard and year.

5.2.2 Beta regression

Beta regression is commonly used for variables that assume values in the unit interval (0,1) (Ferrari and Cribari-Neto, 2004). Beta distribution depends on two scaling parameters Be(p, q). Beta distribution can also be parametrized in terms of its mean $\frac{p}{p+q}$, a dispersion parameter $p + q$, and the variance $\sigma^2 = \frac{\mu(1-\mu)}{1+\phi}$. This reparametrization supports the truncated nature of the beta distribution, where the variance depends on the mean and maximun variance is observed at the centre of the distribution whereas it is minimun

at the edges. In addition, the dispersion of the distribution, for fixed μ, decreases as ϕ. The density function is

$$\pi(y \mid \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma(\phi(1-\mu)}y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}\,,\ 0<y<1\,, \quad (5.5)$$

where Γ is the gamma function.

Let $y_1, \ldots, y_n$ be independent beta variables, where each y_i, $i = 1, \ldots, n$, with mean μ and unknown precision ϕ. These variables, representing proportions (in our particular case, cumulative proportion of ascospores discharged), can be linked to the linear predictor using a similar approach to the generalized linear models (GLM) with the logit function.

$$\eta_i = g(\mu_i) = \beta_0 + \sum_{j=1}^{N_\beta}\beta_j x_{ji} + \sum_{k=1}^{N_f} f_k(z_{ki}) + v_i\,,\ i = 1, \ldots, n \quad (5.6)$$

where η_i enters the likelihood through a logit link, β_0 is the intercept of the model, β_j are the fixed effects of the model, f_k denote any smooth effects, and v_i represents unstructured error terms (random variables). The models which we deal with in this work include only fixed effects and in some cases an unstructured term corresponding to independent random effect year, but they could also incorporate spatial or spatio-temporal effects (Paradinas et al., 2018).

However, one of the main drawback of the beta distribution is its incapability to provide a satisfactory description of the data at the extremes, i.e. 0 and 1. Several solutions have been presented in the literature, like adding a small error value to the observations to satisfy this criterion (Warton and Hui, 2011) or using zero and one inflated models (Liu and Kong, 2015). In this study we adopt the approach by Ferrari and Cribari-Neto (2004), who proposed a transformation which compresses the data symmetrically around 0.5, so, extreme values are affected more than values lying close to 0.5. In particular, the transformed values are obtained as

$$y_i^* = \frac{y_i \cdot (n-1) + \frac{1}{2}}{n} \quad (5.7)$$

5.2.3 Bayesian inference using the INLA approach

Once the model was determined, the next step was to estimate its parameters. A Bayesian hierarchical approach was used to approximate the variation in the proportion of discharged ascospores with INLA (Rue et al., 2009). This methodology uses Laplace approximations (Tierney and Kadane, 1986) to get the posterior distributions in Latent Gaussian models (LGMs) (Rue et al., 2009). LGMs are a particular case of the Structured Additive Regression (STAR) models, where the mean of the response variable is linked to a structured predictor that accounts for the effects of various covariates in an additive way. The prior knowledge of the additive predictor is expressed using Gaussian prior distributions. In this context, all the latent Gaussian variables can be seen as components of a vector known as the latent Gaussian Field.

Vague Gaussian distributions were used here for the parameters involved in the fixed effects $\beta_j \sim \mathcal{N}(0, 10^{-5})$, and a multivariate independent Gaussian distribution for the random effect year, depending of a precision parameter $v_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q}(\tau))$. Precision of the Beta distribution (ϕ) was reparametrized as $\phi = \exp(\alpha)$ to assure that ϕ was a positive parameter. We assumed, following Simpson et al. (2017), pc-priors on the logprecision for both parameters.

The computational implementation R-INLA for R was used to perform approximate Bayesian inference (R Core Team, 2018). Model selection was conducted based on choosing the best subset of covariates (see, for instance, Heinze et al. (2018) for adetailed revision of model selection procedures). This method evaluates all 2^k (where k represents the number of components of the model: covariates and the random effect in our case) possible models and choose the best model according to an information criterion. In this work we have used the Deviance Information Criterion (DIC), which is a generalization of the Akaike Iinformation Ccriterion (AIC) developed for Bayesian model comparison (Spiegelhalter et al., 2002), and the Watanable-Akaike Information Criteria (WAIC) (Watanabe, 2010). The DIC and WAIC are the sum of two components, one quantifying model fit and other evaluating model complexity. The predictive ability of the models

was evaluated by cross validation using the Logarithmic Conditional Predictive Ordinate (LCPO) (Roos et al., 2011). Models with the lowest values of DIC, WAIC and LCPO were selected. Lastly, the marginal posterior densities for the parameters and predictive distributions for new observations were obtained with the best model.

Best model was evaluated plotting observed values against the mean of the poterior predictive distribution of μ (predicted). Linear regression was fitted and R^2 was computed. The Root Mean Square Error was also calculated (RMSE).

5.3 Results

The model that included the variables *ADD* and *ADDvpd* as fixed effects, and the variable year as a random effect showed lower DIC, WAIC and LCPO (Table 5.1).

As expected, in the best model, *ADD* and *ADDvpd* were relevant (Figure 5.1), and they had a positive effect on the expected cumulative proportion of ascospores discharged, being 0.285 the mean posterior distribution and [0.271, 0.299] a 95% credible interval for the parameter corresponding to the fixed effect *ADD*; and, 0.425 the mean posterior distribution and [0.360, 0.491] a 95% credible interval for the parameter corresponding to the fixed effect *ADDvpd*, i.e., the cumulative proportion of ascospores increases when *ADD* y *ADDvpd* are incremented (Table 5.2).

The posterior distribution of the hyperparameters was also computed (Table 5.2 and Figure 5.1), showing that the random effect has a high precision, which means that does not have so much variance, but enough to be important in our model. Mean of the posterior predictive distribution for μ was also plotted (Figure 5.2).

In Figure 5.3, observed values against predicted values were represented showing that the linear regression of predicted versus observed data accounted for more than 90% of the total variance ($R^2 = 0.97$). The RMSE was also calculated resulting in a value equal to 0.0375.

TABLE 5.1: Models for the cumulative proportion of *Plurivorosphaerella nawae* ascospores discharged from persimmon leaf litter based on accumulated degree-days (ADD), ADD taking into account vapor pressure deficit ($ADDvpd$), ADD taking into account vapor pressure deficit and rain ($ADDwet$), and a year random effect ($\boldsymbol{v}$).

MODEL	DIC[2]	WAIC[3]	LCPO[4]
$1 + ADD^{1} + ADDvpd + \boldsymbol{v}$	-1166.204	-1163.821	-1.847
$1 + ADD + ADDwet + ADDvpd + \boldsymbol{v}$	-1165.509	-1163.271	-1.846
$1 + ADD + ADDwet + \boldsymbol{v}$	-1151.959	-1150.240	-1.825
$1 + ADD + \boldsymbol{v}$	-1135.662	-1132.912	-1.798
$1 + ADD + ADDvpd$	-1119.148	-1117.161	-1.773

[1] $biofix$ = January 1, $T_{base} = 0^{o}$C
[2]Deviance Information Criterion
[3]Watanabe-Akaike Information Criteria
[4]Logarithmic Conditional Predictive Ordinate

TABLE 5.2: Mean, standard deviation (sd), quantiles (Q) and mode for the paramaters and hyperparameters (ϕ, τ) of the best model for the cumulative proportion of *Plurivorosphaerella nawae* ascospores discharged from persimmon leaf litter based on accumulated degree-days (ADD) and ADD taking into account vapor pressure deficit ($ADDvpd$). ϕ is the precision parameter of the likelihood and τ the precision of the random effect year.

Parameters	mean	sd	$Q_{0.025}$	$Q_{0.5}$	$Q_{0.975}$	mode
Intercept	-7.850	0.228	-8.239	-7.851	-7.400	-7.854
ADD	0.286	0.007	0.272	0.286	0.299	0.286
$ADDvpd$	0.425	0.033	0.360	0.424	0.491	0.424
Hyperpars	**mean**	**sd**	$Q_{0.025}$	$Q_{0.5}$	$Q_{0.975}$	**mode**
ϕ	23.477	2.446	18.942	23.392	28.532	23.272
τ	218.780	110.468	86.444	191.914	505.313	151.993

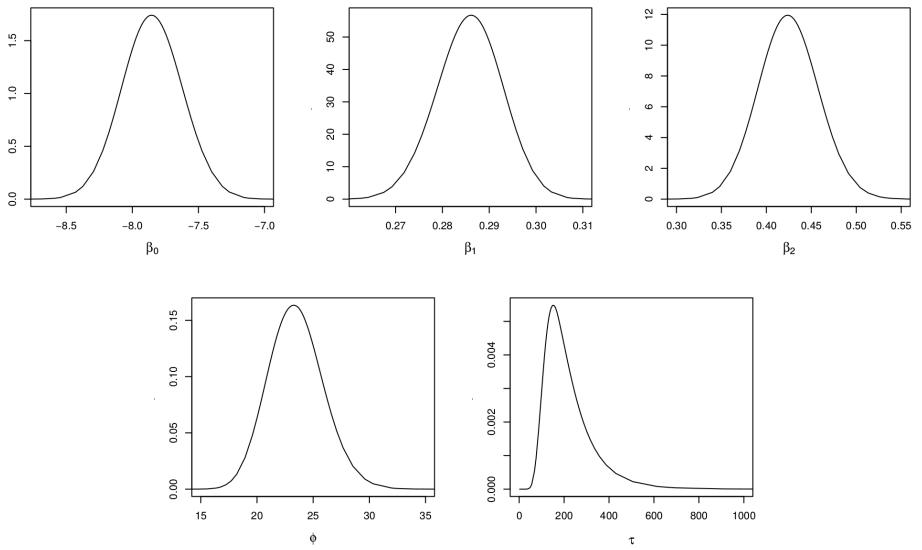

FIGURE 5.1: Posterior distribution of the paramaters and hyperparameters of the best model for the cumulative proportion of *Plurivorosphaerella nawae* ascospores discharged from persimmon leaf litter based on accumulated degree-days (*ADD*) and *ADD* taking into account vapor pressure deficit (*ADDvpd*); ϕ is the precision parameter of the likelihood and τ the precision of the random effect year.

5.4 Discussion

In the model selected, *ADDvpd* and *ADD* were the covariates driving the maturation of *P. nawae* ascospores. There are many examples in the literature indicating that models for ascospore maturation should be corrected for dry periods, by accumulating degree-days only when enough moisture was available in leaf litter. Navas-Cortés et al. (1998b) considered only ADD on rainy days ($\geq$ 1mm) to predict the maturation of *Mycosphaerella rabiei* pseudothecia in chickpea in Spain. Actually, they indicated that rain was essential for the synchronization between *M. rabiei* ascospore availability and vegetative growth of the host. In Norway, Stensvand et al. (2005) improved model accuracy for *V. inaequalis* ascospore maturity in dry years by halting degree-day accumulation if 7 consecutive days without rain occurred. When

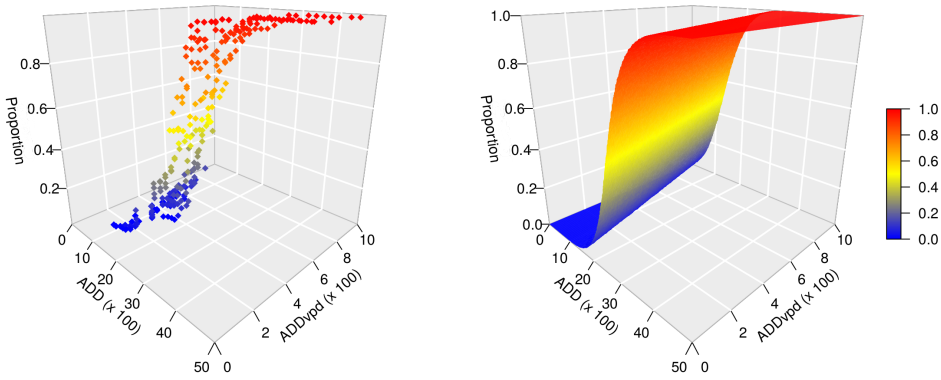

FIGURE 5.2: Representation of accumulated degree days (ADD) and the ADD taking into account the vapor preasure deficit ($ADDvpd$) against the cumulative proportion of *Plurivorosphaerella nawae* ascospores discharged from persimmon leaf litter. Left: data. Right: mean of the posterior predictive distribution for μ.

comparing different models for *V. inaequalis* ascospore maturation in different areas, Eikemo et al. (2011) indicated that those adjusted for dry periods were consistently the most accurate predictors of ascospore depletion.

During the periods of study, dews were much more frequent than rains. In the case of *P. nawae*, wetness induced by dew was not sufficient for ascospore discharge (Vicent et al., 2011), but in absence of rain it may favor pseudothecial development and subsequent ascospore maturation. This was described by Rossi et al. (1999) for *V. inaequalis* in Italy, where models accounting for leaf litter wetness significantly improved estimates of airborne ascospores. Furthermore, Mondal and Timmer (2002) demonstrated that alternate wetting and drying of the leaf litter was necessary for the formation of pseudothecia of *Zasmidium citrigriseum*.

The selection of the date from when degree-days are accumulated (i.e. biofix), has been pointed out as a critical factor in the models for ascospore maturation and release. In some cases, a date was chosen based on a specific phenological stage of the host, such as bud break or green tip (MacHardy and Gadoury, 1985; Eikemo et al., 2011). However, the synchrony between host and fungal phenology may differ among regions. Often, the date of detection

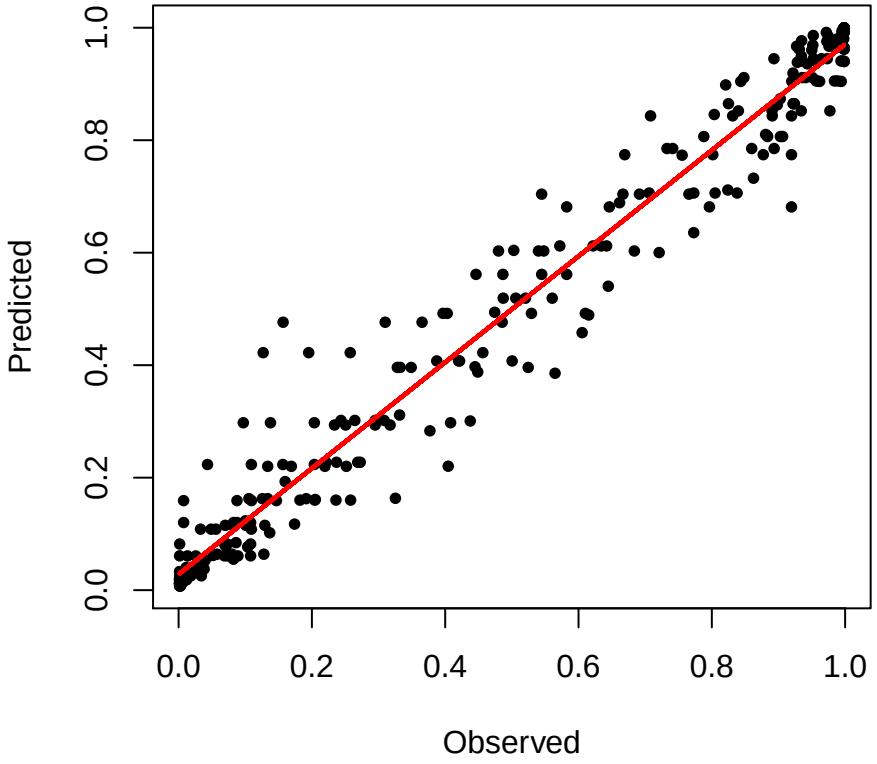

FIGURE 5.3: Observed values against mean of the posterior predictive distribution for μ (predicted) for the best model for the cumulative proportion of *Plurivorosphaerella nawae* ascospores discharged from persimmon leaf litter based on accumulated degree-days (ADD) and ADD taking into account vapor pressure deficit ($ADDvpd$). Red line is the regression line.

of the first mature pseudothecia or the first ascospore trapped has been used as the biofix (Spotts et al., 1994; Eikemo et al., 2011). Nevertheless, this approach relies upon the sensitivity of the detection methods used and, more importantly, requires leaf litter sampling or deployment of spore traps. Both methods are time and resource consuming, limiting the extent and

density of the monitoring network. The most convenient approach to set the biofix is to use a fixed calendar date (James and Sutton, 1982a), but it was argued that it does not take into account the climatic differences between regions (Llorente and Montesinos, 2004). Nevertheless, in our case, we demonstrated analytically that the selection of the biofix was not relevant in the beta regression models for the maturation of *P. nawae* ascospores (Appendix 5.5). Hence, January 1 was chosen as the biofix because, in our conditions, persimmon trees attain complete leaf fall always before this date and so all the leaves are on the orchard floor.

In our model for *P. nawae*, like those for other ascomycetes, temperature and moisture covariates were considered having a continuous positive effect on ascospore development. However, the process resulting in ascospore formation in the leaf litter can be divided in different phases, which may have distinct temperature and moisture requirements. For *M. rabiei*, Gamliel-Atinsky et al. (2005) defined pseudothecium ontogeny followed by initiation of asci and ascospores, and finally ascospore maturation. Navas-Cortés et al. (1998a) indicated that moisture was essential for pseudothecium ontogeny in *M. rabiei* whereas cool temperatures were required for the initiation of asci and ascospores. Actually, low temperatures and relatively long incubation periods are generally needed for the onset of sexual reproduction in many ascomycetes (Trapero-Casas et al., 1992). James and Sutton (1982b) indicated that the development of asci and ascospores in *V. inaequalis* was initiated in spring, after a dormant period which was not influenced by temperature or moisture levels. Then, rapid maturation of ascospores was favored by moisture and increasing temperatures. Knowledge about the temperature and moisture requirements for each phase of ascospore formation in *P. nawae* may help to develop models with improved performance for extrapolation to other areas.

Our models also corroborated previous studies in Spain indicating that *P. nawae* adapted to semi-arid conditions by advancing the period of ascospore production to escape from the typical Mediterranean dry summer. Consequently, ascospore production coincides with spring rains, from March to June, under more favorable conditions for infection. On the other hand,

lower winter temperatures in Korea delayed ascospore release to June-August, concurring with the abundant summer rains typical of this area (Kang et al., 1993; Kwon et al., 1995; Kwon and Park, 2004).

When comparing different methods to estimate the maturity and release of *V. inaequalis* ascospores, Gadoury et al. (2004) found that cumulative ascospore release in discharge tests from the leaf litter lagged behind that measured by spore traps. This was mainly attributed to leaf litter decay, which progressively reduced the overall ascospore population in the orchard air. However, in those discharge tests, a fixed area of nearly 8 cm^2 was sampled from overwintered leaf litter, thus not accounting for leaf litter attrition. In our previous studies, discharge tests allowed detection of mature ascospores of *P. nawae* in the leaf litter before they were released to air in the orchard (Vicent et al., 2012). In contrast to apple leaves, persimmon leaves are typically coriaceous and no substantial degradation of the leaf litter was observed under the conditions of our study. Moreover, in our case a fixed number of leaves instead of a predefined leaf area were sampled and placed into the wind tunnel to extract the available ascospores. Actually, discharge tests from the leaf litter are effectively used in Spain to predict inoculum availability and schedule fungicide sprays for *P. nawae* management.

Models for ascospore maturation in *V. inaequalis* are mainly aimed to predict the duration of the period for primary inoculum, when fungicide applications need to be intensified. Thus, practical performance of the models for *V. inaequalis* relies on their ability to accurately predict ascospore depletion more that the exponential phase of ascospore production (Gadoury et al., 2004; Eikemo et al., 2011). In the case of *P. nawae* in Spain, no secondary conidia have been observed and infections were caused by ascospores formed in the leaf litter (Vicent et al., 2012). Therefore, accurate predictions for the onset and conclusion of the primary inoculum in the leaf litter are paramount for designing efficient fungicide spray programs. Interestingly, beta regression models for *P. nawae* showed the highest accuracy in these two events of interest, the start and the end of ascospore production period.

References

Bassimba, D., Mira, J., Sedano, M., and Vicent, A. (2017). Control and yield loss modelling of circular leaf spot of persimmon caused by *Mycosphaerella nawae*. *Annals of Applied Biology*, 170(3):391–404.

Berbegal, M., Mora-Sala, B., and García-Jiménez, J. (2013). A nested-polymerase chain reaction protocol for the detection of *Mycosphaerella nawae* in persimmon. *European Journal of Plant Pathology*, 137(2):273–281.

Clark, J. S. (2005). Why environmental scientists are becoming Bayesians. *Ecology letters*, 8(1):2–14.

Cooley, D. R., Lerner, S. M., and Tuttle, A. F. (2007). Maturation of thyriothecia of *Schizothyrium pomi* on the reservoir host *Rubus allegheniensis*. *Plant Disease*, 91(2):136–141.

De Wolf, E. D. and Isard, S. A. (2007). Disease cycle approach to plant disease prediction. *Annu. Rev. Phytopathol.*, 45:203–220.

Eikemo, H., Gadoury, D., Spotts, R., Villalta, O., Creemers, P., Seem, R., and Stensvand, A. (2011). Evaluation of six models to estimate ascospore maturation in *Venturia pyrina*. *Plant disease*, 95(3):279–284.

Ferrari, S. L. P. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799 – 815.

Gadoury, D. M., MacHardy, W. E., et al. (1982). A model to estimate the maturity of ascospores of *Venturia inaequalis*. *Phytopathology*, 72(7):901–904.

Gadoury, D. M., Seem, R. C., MacHardy, W. E., Wilcox, W. F., Rosenberger, D. A., and Stensvand, A. (2004). A comparison of methods used to estimate the maturity and release of ascospores of *Venturia inaequalis*. *Plant Disease*, 88(8):869–874.

Gamliel-Atinsky, E., Shtienberg, D., Vintal, H., Nitzni, Y., and Dinoor, A. (2005). Production of *Didymella rabiei* pseudothecia and dispersal of ascospores in a Mediterranean climate. *Phytopathology*, 95(11):1279–1286.

Gilks, W. R., Richardson, S., and DJ, S. (1996). Introducing Markov Chain Monte Carlo. Markov Chain Monte Carlo in Practice.

Heinze, G., Wallisch, C., and Dunkler, D. (2018). Variable selection–a review and recommendations for the practicing statistician. *Biometrical Journal*, 60(3):431–449.

Ikata, S. and Hitomi, T. (1929). Studies on circular leaf spot of persimmon caused by *Mycosphaerella nawae*. *Special Bulletin of the Okayama Prefecture Agricultural Experiment Station*, 33:1–36. In Japanese.

James, J. and Sutton, T. (1982a). A model for predicting ascospore maturation of *Venturia inaequalis*. *Phytopathology*, 72(8):1081–1085.

James, J. and Sutton, T. (1982b). Environmental factors influencing pseudothecial development and ascospore maturation of *Venturia inaequalis*. *Phytopathology*, 72(8):1073–1080.

Kang, S., Kwon, J., Lee, Y., and Park, C. (1993). Effects of meteorological factors on perithecial formation and release of ascospores of *Mycosphaerella nawae* from the overwintered persimmon. *Rural Development Administration Journal of Agricultural Science*, 35:337–343. In Korean, abstract in English.

Kwon, J., Kang, S., Chung, B., and Park, C. (1995). Environmental factors affecting ascospore release of *Mycosphaerella nawae*, the causal organism of the spotted leaf casting of persimmon. *Korean Journal of Plant Pathology (Korea Republic)*, 11:344–347.

Kwon, J. H., Kang, S. W., Park, C. S., and Kim, H. K. (1998). Microscopic observation of the pseudothecial development of *Mycosphaerella nawae* on persimmon leares infected by ascospore and conidia. *Korean Journal of Plant Pathology*, 14:408–412.

Kwon, J.-H. and Park, C.-S. (2004). Ecology of disease outbreak of circular leaf spot of persimmon and inoculum dynamics of *Mycosphaerella nawae*. *Research in Plant Disease*, 10(4):209–216. In Korean, abstract in English.

Legler, S. E., Caffi, T., and Rossi, V. (2014). A model for the development of *Erysiphe necator* chasmothecia in vineyards. *Plant pathology*, 63(4):911–921.

Liu, F. and Kong, Y. (2015). zoib: an R package for Bayesian inference for beta regression and zero/one inflated beta regression. `http://CRAN.R-project.org/package=zoib`.

Llorente, I. and Montesinos, E. (2004). Development and field evaluation of a model to estimate the maturity of pseudothecia of *Pleospora allii* on pear. *Plant disease*, 88(2):215–219.

Luley, C. J. and McNabb Jr, H. S. (1991). Estimation of seasonal ascospore production of *Mycosphaerella populorum*. *Canadian Journal of Forest Research*, 21(9):1349–1353.

MacHardy, W. E. and Gadoury, D. M. (1985). Forecasting the seasonal maturation of ascospores of *Venturia inaequalis*. *Phytopathology*, 75(4):381–385.

Mondal, S. and Timmer, L. (2002). Environmental factors affecting pseudothecial development and ascospore production of *Mycosphaerella citri*, the cause of citrus greasy spot. *Phytopathology*, 92(12):1267–1275.

Navas-Cortés, J., Trapero-Casas, A., and Jiménez-Díaz, R. (1998a). Influence of relative humidity and temperature on development of *Didymella rabiei* on chickpea debris. *Plant Pathology*, 47(1):57–66.

Navas-Cortés, J., Trapero-Casas, A., and Jimenez-Diaz, R. (1998b). Phenology of *Didymella rabiei* development on chickpea debris under field conditions in spain. *Phytopathology*, 88(9):983–991.

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

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Roos, M., Held, L., et al. (2011). Sensitivity analysis in bayesian generalized linear mixed models for binary data. *Bayesian Analysis*, 6(2):259–278.

Rossi, V., Ponti, I., Marinelli, M., Giosue, S., and Bugiani, R. (1999). Field evaluation of some models estimating the seasonal pattern of airborne ascospores of *Venturia inaequalis*. *Journal of Phytopathology*, 147(10):567–575.

Rossi, V., Salinari, F., Pattori, E., Giosuè, S., and Bugiani, R. (2009). Predicting the dynamics of ascospore maturation of *Venturia pirina* based on environmental factors. *Phytopathology*, 99(4):453–461.

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

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

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Spotts, R., Cervantes, L., and Cervantes, L. (1994). Factors affecting maturation and release of ascospores of *Venturia pirina* in Oregon. *Phytopathology*, 84(3):260–263.

Stensvand, A., Eikemo, H., Gadoury, D. M., and Seem, R. C. (2005). Use of a rainfall frequency threshold to adjust a degree-day model of ascospore maturity of *Venturia inaequalis*. *Plant Disease*, 89(2):198–202.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Satistical Association*, 81(393):82–86.

Trapero-Casas, A., Kaiser, W., et al. (1992). Development of *Didymella rabiei*, the teleomorph of *Ascochyta rabiei*, on chickpea straw. *Phytopathology*, 82(11):1261–1266.

Vicent, A., Bassimba, D., and Intrigliolo, D. (2011). Effects of temperature, water regime and irrigation system on the release of ascospores of *Mycosphaerella nawae*, causal agent of circular leaf spot of persimmon. *Plant Pathology*, 60(5):890–908.

Vicent, A., Bassimba, D. D., Hinarejos, C., and Mira, J. L. (2012). Inoculum and disease dynamics of circular leaf spot of persimmon caused by *Mycosphaerella nawae* under semi-arid conditions. *European journal of plant pathology*, 134(2):289–299.

Villalta, O., Washington, W., Rimmington, G. M., and MacHardy, W. (2001). Environmental factors influencing maturation and release of ascospores of *Venturia pirina* in Victoria, Australia. *Australian Journal of Agricultural Research*, 52(8):825–837.

Warton, D. I. and Hui, F. K. (2011). The arcsine is asinine: the analysis of proportions in ecology. *Ecology*, 92(1):3–10.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec):3571–3594.

Whiteside, J. (1974). possibilities of using ground sprays to control citrus greasy spot. In *Proceedings of the Florida State Horticultural Society*, volume 86, pages 19–23.

5.5 Appendix

We evaluated analytically whether, for the same location and year, the selection of the biofix is relevant or not in beta regression models. Let X_{1i} and X_{2i} be two variables expressed in physiological units like ADD_i, $ADDvpd_i$ or $ADDwet_i$, with different biofix: $biofix_1$ and $biofix_2$ respectively, being $biofix_1 < biofix_2 <$date of the first observation. As X_{1i} and X_{2i} are two variables expressed in physiological units, then, both of them can be expressed as a sum of variables, i.e.,

$$X_{1i} = \sum_{j=biofix_1}^{N(i)} Z_j \,, \tag{5.8}$$

$$X_{2i} = \sum_{j=biofix_2}^{N(i)} Z_j \,, \tag{5.9}$$

being Z_j the variable ADD, $ADDvpd$ or $ADDwet$. As $biofix_1 < biofix_2$, X_{1i} can be rewritten as

$$\begin{aligned} X_{1i} &= \sum_{j=biofix_1}^{N(i)} Z_j \\ &= \sum_{j=biofix_1}^{biofix_2} Z_j + \sum_{j=biofix_2}^{N(i)} Z_j \\ &= X_{2i} + \sum_{j=biofix_1}^{biofix_2} Z_j \,. \end{aligned} \tag{5.10}$$

With two beta regression models having both the same response variable, but different covariates X_{1i} or X_{2i}, i.e.

$$\eta_i = \beta_0 + \beta_1 X_{1i} \,, \tag{5.11}$$

$$\eta_i = \beta_0' + \beta_1' X_{2i} \,. \tag{5.12}$$

the selection of the biofix is not relevant when $\beta_1 = \beta_1'$. From Equation (5.11) and Equation (5.12), using the definition of X_{1i} we obtain:

$$\eta_i = \beta_0 + \beta_1 X_{1i} = \beta_0 + \beta_1 \sum_{j=biofix_1}^{biofix_2} Z_j + \beta_1 X_{2i} \,. \tag{5.13}$$

$$\beta_0 + \beta_1 \sum_{j=biofix_1}^{biofix_2} Z_j = \beta_0'$$

$$\beta_1 = \beta_1' \,.$$

Chapter 6

Climatic distribution of citrus black spot caused by *Phyllosticta citricarpa*. A historical analysis of disease spread in South Africa

In this chapter, we present a version of our paper "Climatic distribution of citrus black spot caused by *Phyllosticta citricarpa*. A historical analysis of disease spread in South Africa" by Joaquín Martínez-Minaya (University of Valencia), David Conesa (University of Valencia), Antonio López-Quílez (University of Valencia) and Antonio Vicent (Valencian Institute for Agricultural Research) published in *European Journal of Plant Pathology*, 143, 69–83. The chapter contains at the end the references used in this work.

Abstract

Citrus black spot (CBS), caused by *Phyllosticta citricarpa*, is one of the main fungal diseases of citrus worldwide. The Mediterranean Basin is free of the disease and thus phytosanitary measures are in place to avoid the entry of *P. citricarpa* in the EU territory. However, the suitability of the climates present in the Mediterranean Basin for CBS establishment and

spread is debated. As a case study, an analysis of climate types and environmental variables in South Africa was conducted to identify potential associations with CBS distribution. The spread of the disease was traced and georeferenced datasets of CBS distribution and environmental variables were assembled. In 1950 CBS was still confined to areas of temperate climates with summer rainfall (Cw, Cf), but spread afterwards to neighbouring regions with markedly drier conditions. Actually, the hot arid steppe (Bsh) is the predominant climate where CBS develops in South Africa nowadays. The disease was not detected in the Mediterranean-type climates Csa and Csb as defined by the Köppen-Geiger system and the more restrictive Aschmann's classification criteria. However, arid steppe (Bs) climates, where CBS is prevalent in South Africa, are common in important citrus areas in the Mediterranean Basin. The most noticeable change in the environmental range occupied by CBS in South Africa was the amount and seasonality of rainfall. Due to the spread of the disease to dryer regions, the minimum annual precipitation in CBS-affected areas declined from 663 mm in 1950 to 339 mm at present. The minimum value precipitation of warmest quarter also declined from 290 mm to 96 mm. Strong spatial autocorrelation in CBS distribution data was detected, so further modelling efforts should consider the relative contribution of environmental variables and spatial effects to estimate the potential geographical range of CBS.

Keywords

Guignardia citricarpa, risk assessment, species distribution, biogeography, plant health

6.1 Introduction

Citrus black spot (CBS) is a serious disease caused by the fungus *Phyllosticta citricarpa* (McAlpine) Van der Aa (syn. *Guignardia citricarpa* Kiely). The pathogen was first reported in Australia and is currently present in the main citrus-growing regions of southern and central Africa, South America and

Asia (Kiely, 1948; Kotzé, 2000). In 2010 CBS was reported in Florida (USA) and was the first detection in North America (Schubert et al., 2012). The disease causes external blemishes on the rind which make the fruit unsuitable for the fresh market. In some cases, CBS also induces premature fruit drop resulting in severe crop losses (Araújo et al., 2013). Leaves are infected by *P. citricarpa* but lesions are visible only on highly susceptible varieties, such as lemons, or stressed trees. All commercial varieties of sweet orange, mandarin, lemon and grapefruit are susceptible to the disease (Kotzé, 2000).

The pathogen reproduces through sexual ascospores formed in pseudothecia in the leaf litter, but after completing a maturation process driven by temperature and moisture (Fourie et al., 2013; Lee et al., 1973). Mature ascopores are released from pseudothecia mainly by the effect of rain and disseminated by air currents (McOnie, 1964b). Ascospores infect susceptible fruit and leaves in the presence of moisture and adequate temperature, but quantitative information on the environmental requirements for infection are not known. The pathogen also reproduces asexually by conidia formed in pycnidia on fruit lesions and twigs, which are disseminated by rain splash (Spósito et al., 2011; Whiteside, 1967).

Cultural practices such as leaf litter management, irrigation and early fruit harvesting are used for CBS management. However, fungicide sprays are generally necessary for the economic control of the disease. Recent meta-analysis studies indicated that highly effective fungicide spray programs for CBS control are available (EFSA, European Food Safety Authority, 2014; Makowski et al., 2014), but their implementation increases production costs (Gebrehiwet et al., 2007).

Citrus-growing areas in the European Union (EU) are still free of CBS, thus phytosanitary measures are in place to avoid the entry of *P. citricarpa* (Anonymous, 2000). The import of citrus propagating material is banned in the EU and elsewhere. The import of citrus fruit from CBS-affected regions/orchards into the EU is allowed, but only under specific phytosanitary requirements. Orchards should be subjected to appropriate treatments against *P. citricarpa* and harvested fruit should be free of CBS symptoms. These measures are similar to those imposed by Japan (DAFF, Department of Agriculture Forestry and Fisheries South Africa, 2014) and less stringent

than those by USA, which prohibits the import of citrus fruits from CBS-affected areas (Anonymous, 2014b). However, a long-standing dispute is taking place about the appropriateness of EU phytosanitary regulations for CBS.

One of the key issues debated is the suitability of the climates in the EU citrus-growing areas for CBS establishment and spread. Two studies conducted at global scale using the software CLIMEX indicated that the climates in the Mediterranean Basin were not conducive for CBS development (Paul et al., 2005; Yonow et al., 2013). However, a recent CLIMEX study in the USA indicated that Mediterranean-type climate areas in California would be favourable for CBS (Er et al., 2013). Mechanistic (process-based) models were also used to estimate potential geographical range of CBS. Since the specific environmental requirements for *P. citricarpa* infection are not known, a generic model for foliar fungal pathogens was used (Magarey et al., 2005). One study did not consider the climates of the EU as unsuitable for the establishment of *P. citricarpa* (EFSA, European Food Safety Authority, 2008) but another indicated that CBS was not expected to have an impact in areas with commercial citrus production in Europe (Magarey et al., 2011). Recently, models for *Phyllosticta* spp. ascospore maturation and release were developed (Fourie et al., 2013). These models of inoculum availability were combined with the generic infection model, indicating that environmental conditions in many EU citrus-growing areas were suitable for CBS, though with a high degree of uncertainty (EFSA, European Food Safety Authority, 2014).

This present study develops a historical analysis of CBS spread in South Africa across geographic regions, climate types and selected environmental variables to identify potential associations with disease distribution. South Africa was selected as a case study due to its climate diversity, with citrus regions covering up to ten different climate types. Moreover, good quality datasets of CBS distribution were available for both the initial stages of the epidemics and the current status. The objectives of this study were: (i) to describe the climatic and environmental ranges of CBS in South Africa at the beginning of the epidemic and at the present time, and (ii) to study the

presence of spatial autocorrelation in CBS distribution data. This preparatory work was part of a larger modelling project where the potential geographical range of CBS will be estimated based on relevant environmental variables and spatial effects.

6.2 Materials and methods

6.2.1 CBS spread in South Africa

Scientific and regulatory references on CBS distribution in South Africa were searched. A systematic literature review was performed on July 31 2014 with Web of Knowledge, CAB Abstracts and Google Scholar (all years) combining the terms "citrus black spot", "citricarpa" and "south africa". In the relevant papers retrieved, cited references and citing articles were also reviewed. Phytosanitary regulations published by the Government Gazette from South Africa, the Code of Federal Regulations from USA and the Official Journal of the European Union were reviewed and relevant information on CBS was compiled. Personal communications without supporting verifiable documentation were not considered in the present study.

Locations and dates ($n = 54$) where CBS was detected in South Africa from 1940 to 1950 were extracted from the appendix 2 of Wager (1952) and georeferenced. Since the coexistence of pathogenic and non-pathogenic species of *Phyllosticta* in citrus was not discovered until a decade later (McOnie, 1964c), reports of the pathogen in absence of CBS symptoms were excluded from Wager (1952). A raster layer (299 × 259 pixels) of CBS distribution in South Africa georeferenced to the coordinate system WGS84 was generated from the original map published by Paul (2005) and its subsequent updates (Yonow et al., 2013; Anonymous, 2014a). Paul (2005) indicated that areas of CBS presence and absence in commercial orchards and backyard trees were mapped by six field specialists with extensive knowledge of the disease onto a map of South Africa at a scale 1:106 (2 × 2 m). Disease presence records, based on either identification of *P. citricarpa* or on observation of CBS symptoms, were transcribed to a 29.7 × 45-cm map and scanned. Data on CBS distribution were confirmed by 200 citrus growers

and researchers from South Africa at a citrus meeting in 2002. A map of the CBS distribution in Australia was also available (Paul, 2005), but without details and resolution of the original data, so it was not considered in the present study.

6.2.2 Spatial autocorrelation

To test the hypothesis that CBS presence occur at random among grid cells, which should be considered before carrying out further advanced modelling studies, Moran's Index (Moran's I) and Geary's C analyses of spatial autocorrelation were used (Plant, 2012). Moran's I values range from -1 indicating perfect dispersion to 1 indicating perfect correlation (i.e. clustering). The expected value of I in the absence of significant spatial autocorrelation is around 0. The value of Geary's C is 1 in the absence of spatial autocorrelation and approaches zero for strong autocorrelation. For both indices, contiguity-based neighbours were defined in grid cells sharing edges or vertices.

6.2.3 Climate types and environmental variables

Environmental data from South Africa were acquired from the WorldClim database (Hijmans et al., 2005), which reports gridded mean values from the 1950-2000 period. A resolution of 5' (arc min) was used in all datasets. In addition to average monthly mean temperature and precipitation, a set of derivative metrics available in WorldClim were used: minimum temperature of coldest month (BIO_6), mean temperature of wettest quarter (BIO_8), mean temperature of the coldest quarter (BIO_{11}), annual precipitation (BIO_{12}) and precipitation of warmest quarter (BIO_{18}). A derived variable was created with precipitation from October to January (spring-summer in the southern hemisphere).

An algorithm was developed to implement the Köppen-Geiger climate classification system (Köppen, 1936) based on the updated version from Peel et al. (2007). This system considers the following parameters based on temperature (^{o}C) and precipitation (mm): MAP = mean annual precipitation,

MAT = mean annual temperature, T_{hot} = mean temperature of the hottest month, T_{cold} = mean temperature of the coldest month, T_{mon10} = number of months where the mean temperature is above 10, P_{dry} = mean precipitation of the driest month, P_{sdry} = mean precipitation of the driest month in summer, P_{wdry} = mean precipitation of the driest month in winter, P_{swet} = mean precipitation of the wettest month in summer, P_{wwet} = mean precipitation of the wettest month in winter, $P_{threshold}$ varies according to the following rules: if 70% of MAP occurs in winter then $P_{threshold} = 2 \times MAT$, if 70% of MAP occurs in summer then $P_{threshold} = 2 \times MAT + 28$, otherwise $P_{threshold} = 2 \times MAT + 14$. Summer and winter are defined as the warmer and cooler, respectively, six-month period from October to March and April to September (Table 6.1).

The definition of a Mediterranean-type climate developed by Aschmann (1973) was also mapped applying an algorithm to the gridded data from WorldClim. This classification considers the following parameters based on temperature (ºC) and precipitation (mm): MAP = mean annual precipitation, MWP = mean winter precipitation, MAT = mean annual temperature, T_{cold} = mean temperature of the coldest month, T_{range} = range of mean monthly temperature. Winter was November to April in the northern hemisphere and May to October in the southern hemisphere. The Mediterranean-type climate should meet all the following criteria: $MWP \geq 0.65 \times MAP$, $275 \leq MAP \leq 900$, $T_{cold} < 15$ and $MAT \geq 0.7 \times T_{range} + 2.76$. This last condition was set originally by Aschmann (1973) as no more than 3% of the annual hours below 0°C. WorldClim does not include hourly temperature data, thus the relationship between MAT and T_{range} developed by Klausmeyer and Shaw (2009) based on a figure by Aschmann (1973) was used here. Although the present study was focused in South Africa, climatic maps of the Mediterranean Basin were also obtained to discuss the boundaries and geographic extent of Mediterranean-type climates.

Raster layers with maps of CBS presence in 1950 and current CBS presence, CBS absence and low pest (disease) prevalence were overlapped onto

TABLE 6.1: Description of Köppen-Geiger symbols and defining criteria for arid and temperate climates (Peel et al., 2007).

Climate type			Criteria[1]
B		Arid	$MAP < 10 \times P_{threshold}$
W		Desert	$MAP < 5 \times P_{threshold}$
S		Steppe	$MAP \geq 5 \times P_{threshold}$
	h	Hot	$MAT \geq 18$
	k	Cold	$MAT < 18$
C		Temperate	$T_{hot} > 10$ & $0 < T_{cold} < 18$
s		Dry Summer	$P_{sdry} < 40$ & $P_{sdry} < P_{wwet}/3$
w		Dry Winter	$P_{wdry} < P_{swet}/10$
f		Fully humid	Not (Cs) or (Cw)
	a	Hot Summer	$T_{hot} \geq 22$
	b	Warm Summer	Not (a) & $T_{mon10} \geq 4$
	c	Cold Summer	Not (a or b) & $1 \leq T_{mon10} < 4$

[1] MAP = mean annual precipitation, MAT = mean annual temperature, T_{hot} = mean temperature of the hottest month, T_{cold} = mean temperature of the coldest month, T_{mon10} = number of months where the mean temperature is above 10, $P_{dry=}$ mean precipitation of the driest month, P_{sdry} = mean precipitation of the driest month in summer, P_{wdry} = mean precipitation of the driest month in winter, P_{swet} = mean precipitation of the wettest month in summer, P_{wwet} = mean precipitation of the wettest month in winter, $P_{threshold}$ = if 70% of MAP occurs in winter then $P_{threshold} = 2 \times MAT$, if 70% of MAP occurs in summer then $P_{threshold} = 2 \times MAT + 28$, otherwise $P_{threshold} = 2 \times MAT + 14$. Summer (winter) is defined as the warmer (cooler) six-month period from October to March and April to September. In all cases, temperature in C and precipitation in mm.

raster layers with climate types and environmental variables. The proportion of grid cells in each climate type and CBS status was calculated. Median, minimum and maximum values of the environmental variables indicated above were calculated for each CBS status and for each combination of CBS status and climate type. The R software v.3.1.2 (R Core Team, 2013) with the packages spdep, rgdal, raster, and sp was used in all analysis

(Bivand, 2014; Bivand et al., 2014; Hijmans, 2014; Pebesma and Bivand, 2005). When necessary, the presence of citrus orchards in some specific grid cells was corroborated using the package RgoogleMaps (Loecher, 2014).

6.3 Results

6.3.1 CBS spread in South Africa

CBS was first described in South Africa in 1929 in citrus orchards near to Pietermaritzburg, KwaZulu-Natal (Figure 6.1A). The disease was confined to this location and it was considered of minor importance at that time (Doidge, 1929). During the next ten years, CBS spread slowly and in 1940 it was causing considerable damage in this area (Wager, 1952). The appendix 2 of (Wager, 1952) included details of an extensive survey conducted from 1940 to 1950. In 1945 the disease was first reported in Limpopo province (Figure 6.1B) and in 1946, it was detected in Mpumalanga and North West provinces (Figure 6.1C). Citrus-growing areas in Western Cape, Eastern Cape and Gauteng provinces were surveyed and no symptoms of CBS were observed. The Eastern Cape province was again surveyed in 1962 and 1963 by (McOnie, 1964a) and no signs of CBS were found.

The disease was cited by Kotzé (1981) as a major crop destroyer in the provinces of KwaZulu-Natal, Mpumalangaa and Limpopo. *P. citricarpa* was not among the list of regulated plant pathogens when the Agricultural Pests Act was implemented in 1984, but the introduction of citrus plants into the Western Cape, Eastern Cape and Northern Cape provinces was banned by this phytosanitary regulation (Anonymous, 1984). No specific data of CBS introduction in the Eastern Cape province was found, but Korf (1998) indicated that lemon orchards in the Eastern Cape were continuously protected against CBS with fungicides at that time.

In 2002 *P. citricarpa* was included on the list of regulated plant pathogens in South Africa. The movement of citrus plants from KwaZulu-Natal, Mpumalanga, Gauteng, Limpopo, North West and Eastern Cape to the Western Cape, Northern Cape and Free State was banned due to CBS.

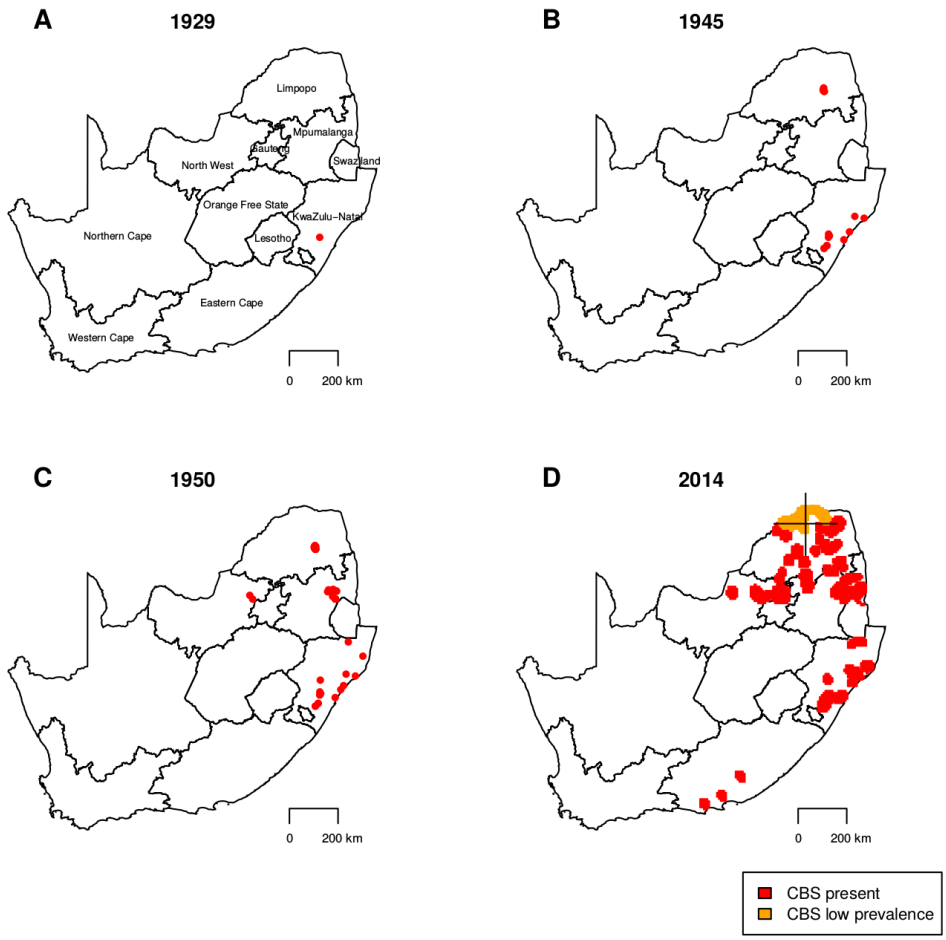

FIGURE 6.1: Geographic distribution of citrus black spot (CBS) caused by *Phyllosticta citricarpa* in South Africa (Anonymous, 2014a; Doidge, 1929; Paul, 2005; Paul et al., 2005; Wager, 1952; Yonow et al., 2013). Data for Lesotho and Swaziland were not available.

Within the Western Cape, the movement of citrus plants was also banned from the easternmost to the westernmost magisterial districts due to CBS (Anonymous, 2002, 2005a,b; DAFF, Department of Agriculture Forestry and Fisheries South Africa, 2009).

A map of CBS distribution in South Africa (Figure 6.1D) was published by Paul (2005) and Paul et al. (2005). CBS-affected areas were located in

the same provinces indicated above and the Western Cape, Northern Cape and Free State provinces were considered CBS-free areas. According to internationally adopted standards, pest (disease) free status is recognized in areas in which a specific pest (disease) does not occur as demonstrated by scientific evidence and in which this condition is officially maintained (IPPC, International Plant Protection Convention, 1995, 2007). The EU considers the entire Western Cape province as a CBS-free area (Anonymous, 2006), whilst the USA only recognizes disease freedom in the westernmost districts of the province (APHIS, Animal and Plant Health Inspection Service USA, 2012).

In 2008, the magisterial districts of Christiana and Taung in the North West province were considered CBS-free and Musina and Soutpansberg in Limpopo, north of the 22º 50'S latitude or west of 29º 20' E longitude, were considered areas of low pest (disease) prevalence for CBS (Anonymous, 2008; DAFF, Department of Agriculture Forestry and Fisheries South Africa, 2009). Low pest (disease) prevalence status is recognized in areas in which a specific pest (disease) occurs at low levels and which is subjected to effective surveillance, control or eradication measures (IPPC, International Plant Protection Convention, 2005, 2007). The CBS distribution map of Paul (2005) and Paul et al. (2005) was updated accordingly by Yonow et al. (2013) (Figure 6.1D).

The CBS-free status of Western Cape, Northern Cape and Free State provinces was documented by recent surveys (Carstens et al., 2012) and limitations for the movement of citrus plants within the Western Cape province due to CBS were lifted in 2014 (Anonymous, 2014a).

Data of CBS distribution in South Africa from the consolidated version of the map (Anonymous, 2014a; Paul, 2005; Yonow et al., 2013) showed a strong spatial autocorrelation (Moran's I = 1, $P < 0.0001$; Geary's C = 0, $P < 0.0001$).

6.3.2 Climate types

Current citrus areas in South Africa were present in all ten climate types in the country. According to the Köppen-Geiger system, arid desert climates

(Bw) were present in citrus areas in Limpopo and Northern Cape provinces (Figure 6.2a). Arid steppe climates (Bs) were present across citrus areas in all provinces. Temperate climates with dry summer (Cs) were present only in the Western Cape. Temperate climates with dry winter (Cw) were present in citrus areas in Gauteng, KwaZulu-Natal, Limpopo, Mpumulanga and North West provinces. Temperate climates without a dry season (Cf) were present in citrus areas in the Eastern Cape, KwaZulu-Natal and Western Cape provinces. Aschmann's Mediterranean-type climate was restricted to the Western Cape (Figure 6.2B).

In the Mediterranean Basin, arid steppe (Bs) climates were present in Spain, Greece, Turkey, Cyprus, Syria, Israel, Libya, Tunisia, Algeria and Morocco (Figure 6.3A). Climates of Mediterranean-type (Cs) climates were present in Portugal, Spain, France including Corsica, Italy including Sicily and Sardinia, Albania, Greece, Turkey, Syria, Israel, Cyprus, Malta, Libya, Tunisia, Algeria and Morocco (Figure 6.3B). Aschmann's Mediterranean-type climate was present in all of the same countries with Cs climates except Albania (Figure 6.3C).

The disease was first detected in South Africa in 1929 in a location with a temperate climate with a dry winter and warm summer (Cwb). In 1950, CBS was restricted to temperate climates with a dry winter (Cw) and fully humid (Cf), with 79.6% of the locations of the Cw climates (hot summer Cwa 57.4%; warm summer Cwb 22.2%) and 20.4% of the Cf climates (hot summer Cfa 16.7%; warm summer Cfb 3.7%).

Considering the grid cells of current citrus-growing areas in South Africa, 55.9% were affected by CBS, 9.2% were of low prevalence, and 34.9% were CBS-free (Figures 6.2A and 6.4). The hot arid steppe climate (Bsh) was the predominant climate where CBS develops, with 20.7% of the grid cells with disease present, 6.5% of low prevalence, and 1.4% CBS-free. The cold arid steppe climate (Bsk) comprised 1.8% of grid cells with CBS present and 5.7% CBS-free. The hot arid desert (Bwh) consisted of 2.3% grid cell of low prevalence and 12.8% CBS-free.

Climates of Cw type covered 21.5% of grid cells with CBS present (Cwa 11.9% and Cwb 9.6%) and 0.4% with low prevalence (Figures 6.2A and 6.4). Climates of Cf type encompassed 11.9% of grid cells with CBS present (Cfa

7.1% and Cfb 4.8%) and 2.1% disease-free (Cfa 0.6% and Cfb 1.5%). The disease was not detected in the cold arid desert (Bwk), Csa and Csb climates with 1.9%, 3.2% and 7.9% of the grid cells, respectively. All grid cells with Aschmann's Mediterranean-type climate (11.9%) were CBS-free (Fig 6.1B).

6.3.3 Environmental variables

Minimum temperature of the coldest month in grid cells with CBS present ranged from 2.3-11.3ºC in 1950 to 0.4-12.9ºC at present (Figure 6.5A). In CBS-free areas it ranged from -0.7ºC to 9.5ºC. Mean temperature of the coldest quarter ranged from 11.7ºC to 17.8ºC in grid cells where CBS was present in South Africa in 1950, from 9.8ºC to 18.8 ºC in current areas of CBS distribution, and from 6.2ºC to 15ºC in CBS-free areas (Figure 6.5C). Mean temperature of the wettest quarter in grid cells with CBS present varied from 20.3-25.1ºC in 1950 to 13.5-27.1ºC at present, with the maximum in areas of low prevalence (Figure 6.5E). The range for this climate variable in CBS-free areas was 6.8-27ºC. The range of annual precipitation in CBS-affected areas was 663-1199 mm in 1950 and 317-1397 mm at present (Figure 6.5B). The lowest mean annual precipitation was 317 mm in areas of low prevalence and 339 mm in areas of CBS presence. In CBS-free areas, the range of annual precipitation was 47-1033 mm. The precipitation of warmest quarter in grid cells with CBS present varied from 290-656 mm in 1950 to 96-756 mm at present, with a range of 6-232 mm in CBS-free areas (Figure 6.5D). The cumulative precipitation from October to January was 372-625 mm in CBS-affected locations in 1950, 121-728 mm in current areas of CBS-distribution, and 9-320 mm in CBS-free areas (Figure 6.2F). When not otherwise stated, values for areas of low prevalence where always higher than the minimum and lower than the maximum indicated for current CBS presence.

When climatic variables were analyzed along with climate types in the current areas of CBS distribution, minimum temperature of coldest month ranged from 0.4ºC in the Cwb climate to 12.9ºC in the Cfa climate (Table 6.2). Mean temperature of the coldest quarter ranged from 9.8ºC to 18.8ºC in the Cfb and BSh climates, respectively. Mean temperature of wettest

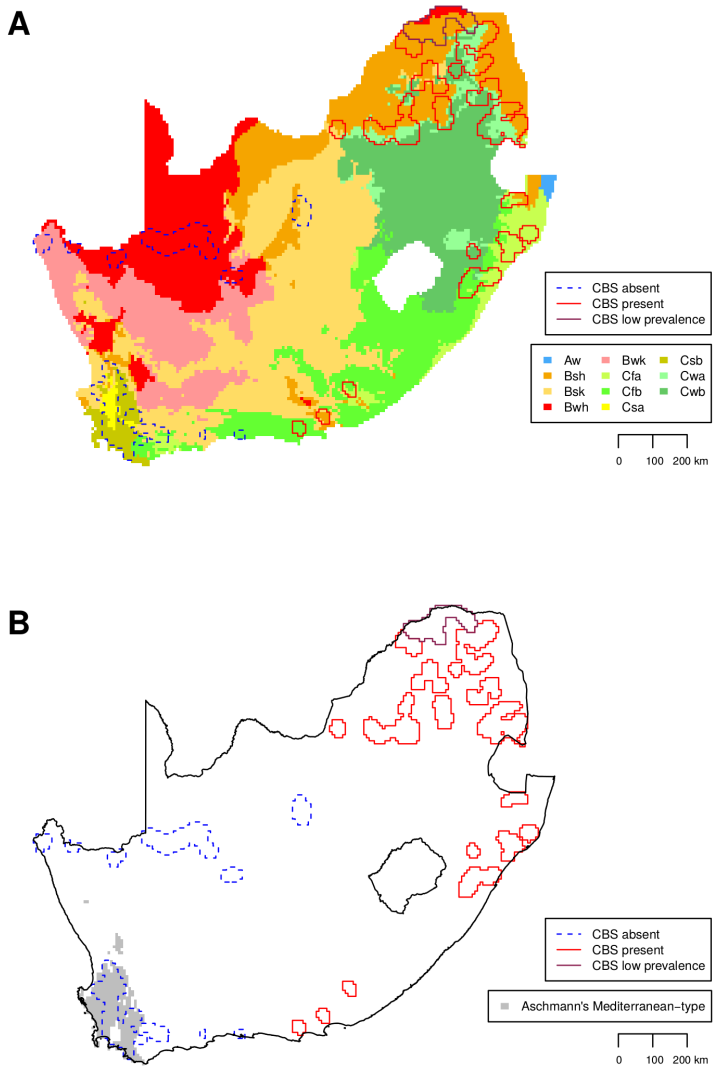

FIGURE 6.2: Climate types and citrus areas in relation to current distribution of citrus black spot (CBS) caused by *Phyllosticta citricarpa* in South Africa. **A** Köppen-Geiger system. **B** Mediterranean-type climate according to Aschmann (1973).

quarter varied from 13.5ºC in the Cfb climate to 27.1ºC in the BWh climate. The lowest annual precipitation was 317 mm in the BWh climate and the highest was 1397 mm in the Cwb climate. Precipitation of the warmest quarter ranged from 96 mm in the BSh climate to 756 mm in the Cwb climate. The minimum and maximum values of precipitation from October to January were 121 mm and 728 mm in the BSh and Cwb climates, respectively.

In CBS-free areas, minimum temperature of coldest month ranged from -0.7ºC in the BSk climate to 9.5ºC in the Csb climate (Table 6.2). Mean temperature of coldest quarter ranged from 6.2ºC to 14.9ºC and mean temperature of wettest quarter from 6.8ºC to 27ºC in the Csb and BWh climates, respectively. The lowest annual precipitation was 47 mm in the BWk climate and the maximum was 1034 mm in the Csb climate. Precipitation of warmest quarter ranged from 6 mm to 232 mm and precipitation from October to January ranged from 9 mm to 320 mm in the BWk and Cfb climates, respectively.

6.4 Discussion

Differences in the two datasets of CBS distribution should be taken into account to interpret the spread of CBS in South Africa. The 1950 dataset was comprised of point coordinates obtained at the beginning of the epidemic with a relative small sample size ($n = 54$). On the other hand, most recent data were gridded areas with a relatively large sample size ($n = 2065$). Furthermore, citrus areas in South Africa increased from 28.900 ha in 1961 to 73.900 ha in 2012 (FAO, Food and Agriculture Organization of the United Nations., 2014) and regions in the Northern Cape province were not even cropped with citrus in 1950 (Reuther et al., 1967). A resolution of 5' was selected for the present study, but similar results (not shown for the sake of simplicity) were obtained with the 30' resolution used in other studies (Paul, 2005; Paul et al., 2005; Yonow et al., 2013)

Historical data on CBS distribution in South Africa illustrated the slow epidemic development characteristic of this disease (Kotzé, 1981). It took

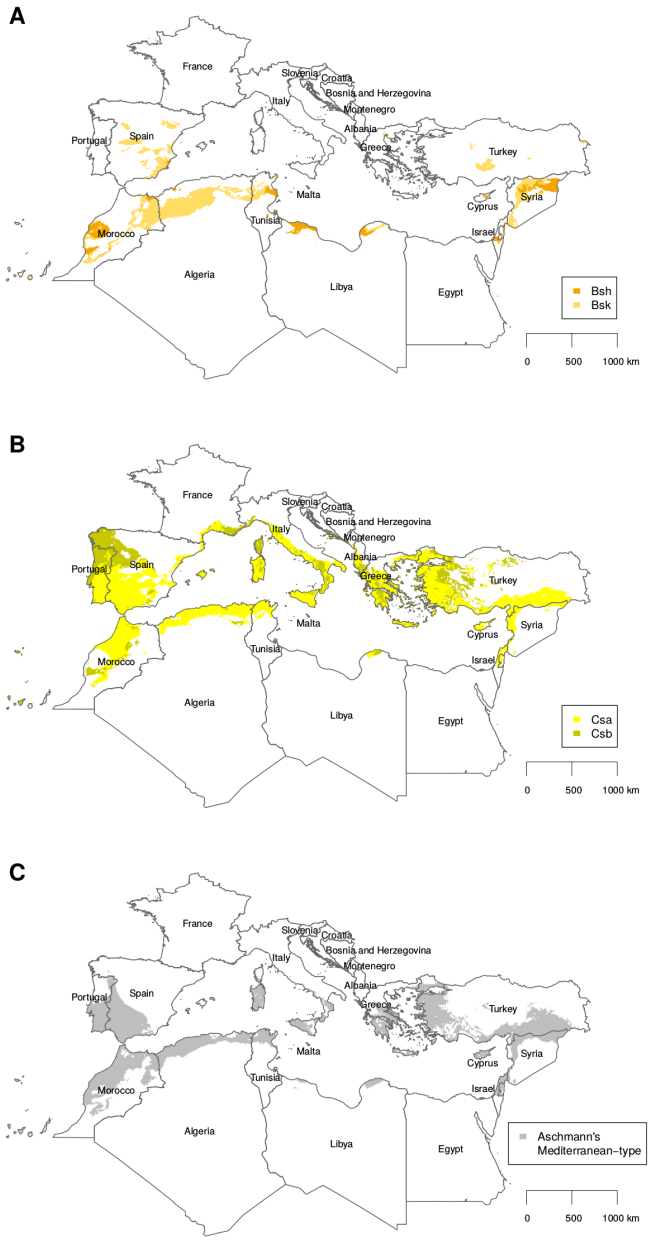

FIGURE 6.3: Climate types in the Mediterranean Basin. BSk and BSh (**A**) Csa and Csb (**B**) climate types of Köppen-Geiger system. **C** Mediterranean-type climate according to Aschmann (1973).

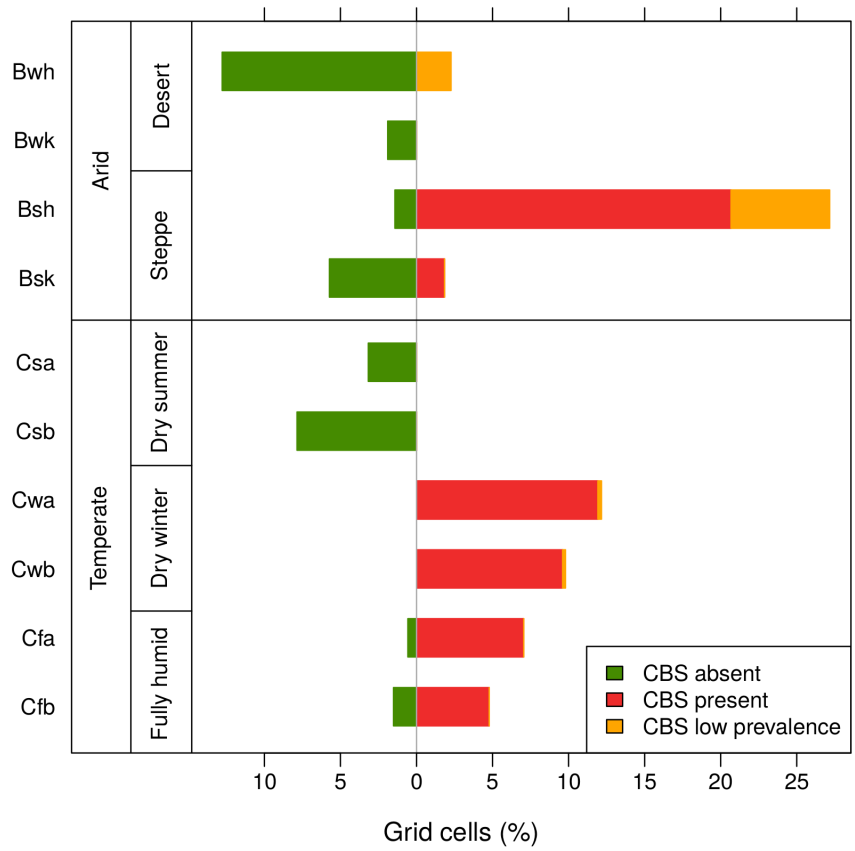

FIGURE 6.4: Proportion of grid cells according to the current status of citrus black spot (CBS) caused by *Phyllosticta citricarpa* in South Africa by Köppen-Geiger climate types (Anonymous, 2014a; Paul, 2005; Paul et al., 2005; Yonow et al., 2013).

several decades from the detection of the first CBS focus in the country to reach a relatively large geographic and climatic range (Figure 6.1). Data also showed that CBS emerged in areas of climates with summer rainfall (Cw, Cf) and later spread to neighbouring regions of arid steppe climate (Bs) with markedly drier conditions. Currently, these arid climates represent the major proportion of CBS-affected areas in the country (Figures 6.2A and 6.4).

In general, the potential for natural spread of CBS by *P. citricarpa* ascospores and conidia is poorly understood. Spatial aggregation of CBS in citrus orchards in Brazil indicated disease dispersion at short distances, below 24.7 m, but neither ascospores nor conidia were monitored in this study (Spósito et al., 2007). Under simulated wind-driven rain conditions, conidia from inoculated citrus fruit were splashed 0.6 m high and 8 m distant (Perryman et al., 2014). No information on the maximum distance movement by airborne *P. citricarpa* ascospores or the minimum concentration needed to initiate an epidemic was found. In other ascomycetes, it was reported that most of the ascospores originated from an infectious source remained within 50-90 m (Chandelier et al., 2014; Mondal et al., 2003). However, the relatively low proportion of ascospores at the tail of the dispersal kernel might contribute to disease spread over longer distances (Rieux et al., 2014).

Although the origin of CBS introductions remains generally unknown, human-assisted movement of infected plant material is considered the most important means of disease spread. The movement of citrus material in South Africa was not regulated until 1984, but quantitative trade data among provinces was not found. In any case, it seems conceivable that larger amounts of plant material were moved from CBS-affected areas to nearby regions than to distant provinces. Consequently, the potential for introduction might have been higher in regions adjacent to CBS-affected areas (Simberloff, 2009). The strong spatial autocorrelation detected in the current CBS distribution data seem to support this hypothesis and suggest that climate itself might not be the main factor limiting the spread of CBS in South Africa. However, further modelling studies are necessary to weigh the relative contribution of environmental variables and spatial effects in disease distribution (Latimer et al., 2006).

6.4.1 Environmental variables

Among the ten climates present in citrus-growing areas in South Africa, the only ones where CBS was not detected were the Mediterranean-type Csa and Csb as well as the BWk arid cold dessert (Figures 6.2A and 6.4). However, these three climates together represented only about 13% of the citrus area in the country and are restricted to locations in the Western Cape and

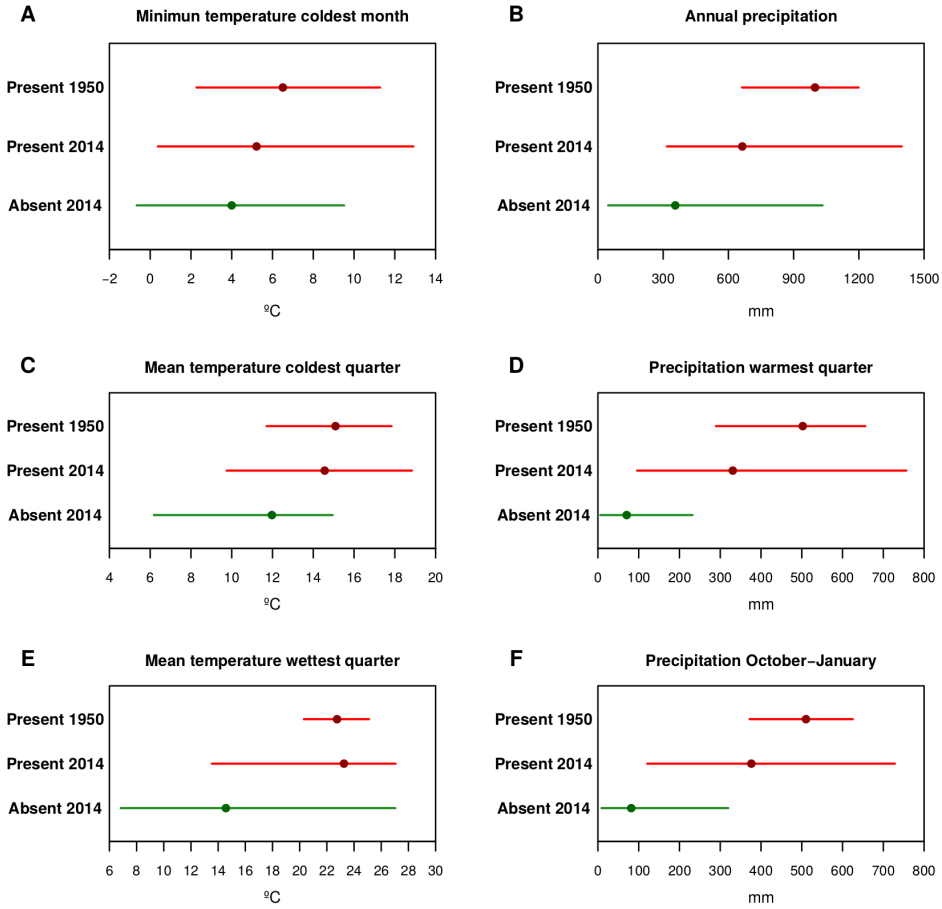

FIGURE 6.5: Median, minimum and maximum values of selected environmental variables in areas of South Africa according to the status of citrus black spot (CBS) caused by *Phyllosticta citricarpa* in 1950 and 2014. CBS presence in 2014 includes areas of low prevalence (Anonymous, 2014a; Paul, 2005; Paul et al., 2005; Wager, 1952; Yonow et al., 2013).

Northern Cape furthest from CBS-affected areas (¿ 450 km). Based on the data of Yonow et al. (2013), a similar pattern was present also in Australia. Areas with Cs climates represented only around 12% of the citrus area in this country and were located about 2500 km from CBS-affected areas (results not shown). It was stated that CBS does not occur in Mediterranean

climates (Yonow et al., 2013), which may be correct when considering only the Mediterranean-type climates Csa and Csb defined by the Köppen-Geiger system (Köppen, 1936; Peel et al., 2007) or the more restrictive Aschmann's classification (Aschmann, 1973; Klausmeyer and Shaw, 2009). However, this assertion is inaccurate when considering the BSh and BSk types, where CBS is most prevalent in South Africa currently (Figures 6.2A and 6.4). Climates of the BS type are also common in the Mediterranean Basin (Figure 6.3A), covering important citrus areas such as Souss, Haouz and Oriental regions in Morocco, Cap Bon peninsula in Tunisia, and the provinces of Castellón, Valencia, Alicante, Murcia and Almería in Spain with more than 70% of the total citrus area in this country (MAGRAMA, Ministerio de Agricultura, Alimentación y Medio Ambiente, 2013).

Studies with CLIMEX indicated that the potential distribution of CBS was mainly limited by cold conditions (Paul et al., 2005; Yonow et al., 2013), though these modelling approaches and their parameterization were questioned (EFSA, European Food Safety Authority, 2008, 2014; Vicent and García-Jiménez, 2008). A non-species-specific degree-day model also predicted a delay in *Phyllosticta* spp. pseudothecium maturation in climates with colder winters and springs (Fourie et al., 2013). Nevertheless, this model is empirically based and so its performance outside the environmental range of development is uncertain (EFSA, European Food Safety Authority, 2014). The minimum value of mean temperature of coldest quarter in South Africa was 3.5ºC lower in the CBS-free than in the CBS-affected areas, but with a wide range of overlap (Figure 6.5). When considering the minimum temperature of coldest month, the difference between CBS-free and CBS-affected areas was only 1ºC. The values for these two environmental variables were 1.9ºC higher in 1950 than at present. In 1950 the disease had a narrow range of mean temperature in wettest quarter between 20.3 and 25.1ºC, but progressively expanded to cooler areas with a range of 13.5-27.1ºC.

The most noticeable change in the environmental range occupied by CBS in South Africa since 1950 was the amount and seasonality of rainfall. Minimum values for the three precipitation variables analyzed were always lower in CBS-free areas, but differences were strongly reduced when CBS expanded to drier regions (Figure 6.5). Due to the spread of the disease from the original foci to neighbouring dry areas, the minimum annual

precipitation in CBS-affected areas was about 50% lower; 663 mm in 1950 and 339 mm at present. Average annual rainfall in areas of low prevalence with BWh climate in north of Limpopo province was 317-367 mm. Annual rainfall values of 339-400 mm were recorded in areas where CBS is endemic under BSh climate in the Eastern Cape and some regions in Limpopo (Figure 6.2A, Table 6.2). This shift in the rainfall pattern associated with the geographical range of CBS was particularly illustrated by the precipitation in the warmest quarter, which moved from a minimum value of 290 mm in 1950 to 96 mm at present. A similar trend was observed also in the precipitation from October to January (spring-summer), which is considered the critical infection period of *P. citricarpa* in some regions of South Africa (Kotzé, 1981; McOnie, 1964b).

The lowest values of summer rainfall in CBS-affected areas were observed in the Eastern Cape province under BSk and BSh climates. Quantitative data on CBS incidence and fungicide spray programs applied in this area were not found. It was pointed out that CBS has a low impact in this region (Fourie et al., 2013; Yonow et al., 2013), though according to international standards, it is not officially considered among the areas of low pest (disease) prevalence in South Africa (Anonymous, 2014a). In any case, as the data from South Africa and other countries indicated, CBS is characterized by slow epidemic development and past experiences warned that future impacts cannot be directly inferred from its present status.

In conclusion, these results clearly demonstrated that CBS expanded in South Africa from its original geographic range in summer rainfall areas to arid regions in the nearby provinces of Limpopo and the Eastern Cape. These results contradict overall statements indicating that CBS occurs exclusively in climates with summer rainfall (Graham et al., 2014; Kotzé, 2000). Further modelling studies should integrate the relative contribution of environmental variables together with the spatial structure of the data to better estimate the potential geographical range of CBS.

TABLE 6.2: Median, minimum and maximum values (in parentheses) of selected climatic variables by Köppen-Geiger climate types in grid cells with presence or absence of citrus black spot caused by *Phyllosticta citricarpa* in South Africa (Anonymous, 2014b; Paul, 2005; Yonow et al., 2013)

Climate type			Min. temp. coldest month (ºC)	Mean temp. coldest quarter (ºC)	Mean temp. wettest quarter (ºC)	Annual precipitation (mm)	Precipitation warmest quarter (mm)	Precipitation October to January (mm)
CBS present[1]								
Arid	Desert	BWh	5.3 (4.7, 9.6)	16 (15.5, 18.2)	26.3 (25.3, 27.1)	339.9 (317.1, 366.9)	187.6 (175.7, 208.4)	207.5 (195, 225.9)
	Steppe	BSh	5.1 (1.5, 10.8)	15 (11.7, 18.8)	24.4 (18.6, 26.6)	554.2 (339.8, 719.1)	281.9 (96.2, 408.4)	317.1 (121.1, 391.1)
		BSk	3.4 (1, 5.7)	12.3 (11.1, 13.8)	21.4 (17.9, 22.9)	582.5 (401.3, 630.9)	295 (113.1, 326.1)	342.4 (149, 383.8)
Temperate	Dry winter	Cwa	7.1 (1.5, 10.7)	15.2 (11.1, 18.1)	22.8 (21.2, 25.6)	776.8 (625.2, 1218.9)	395.7 (304.2, 666.7)	422.7 (341, 634.8)
		Cwb	3.7 (0.4, 7.8)	12.2 (9.9, 14.9)	20.6 (16.9, 21.8)	887.8 (624.2, 1396.6)	437.3 (319.1, 756.3)	487.8 (376.3, 728.2)
	Fully humid	Cfa	10.5 (3.9, 12.9)	16.9 (12.4, 18.3)	23.1 (20.9, 25.2)	948.4 (492.4, 1131.4)	356 (167.7, 417.4)	447.1 (205.9, 520.4)
		Cfb	5.1 (1, 9.1)	12.7 (9.8, 15.5)	20 (13.5, 21.4)	859.5 (501.6, 937.8)	365 (110.6, 427.8)	452.9 (169, 490.4)
CBS absent								
Arid	Desert	BWh	2.9 (0.5, 6.5)	12.2 (10.1, 14.9)	24.9 (17.7, 27)	188.7 (55.4, 275.9)	79.4 (10, 106.7)	72.9 (10, 100.7)
		BWk	7.4 (0.3, 8.6)	13.3 (9.7, 14.8)	14 (12.9, 23.1)	64.6 (47.3, 291.2)	9.5 (6, 115.2)	11.8 (8.9, 106.3)
	Steppe	BSh	1.3 (0.1, 7)	11.4 (10.6, 13.1)	23.5 (13, 24)	429.4 (242.2, 476.5)	199.8 (18.4, 223.7)	202.3 (32.9, 231.4)
		BSk	4.2 (-0.7, 7.4)	11.4 (7.9, 13.9)	13.8 (7.9, 23.1)	399.3 (270.2, 499)	66.6 (22.5, 222.9)	102.5 (38.6, 236.9)
Temperate	Dry summer	Csa	5.9 (3.7, 6.8)	12.2 (10.7, 12.8)	13 (11.4, 13.5)	448.9 (354.3, 916.7)	39.9 (28.7, 77.3)	74.7 (52.2, 143.5)
		Csb	5.1 (0.1, 9.5)	10.9 (6.2, 13.5)	11.1 (6.8, 13.6)	602 (288.6, 1033.5)	65 (31.6, 102.2)	119.5 (53.8, 187.5)
	Fully humid	Cfa	5.9 (4.9, 6.4)	13.1 (12.1, 13.5)	18 (14.2, 18.9)	545.6 (495.1, 559.8)	113.9 (92.8, 118.5)	156.6 (134.6, 158.6)
		Cfb	5.7 (2.9, 7.9)	11.8 (9.8, 13.7)	13.2 (9.8, 17.1)	520.4 (441.1, 920.2)	95.1 (69.9, 232)	134.2 (115.4, 319.8)

6.5 Acknowledgments

We thank V. Monzó (Plug Dayhe S.L.) for georeferencing disease distribution data, J.V. Castelló (IVIA) for retrieving historical references, and L.W Timmer (CREC-IFAS/University of Florida) and M. Pautasso (EFSA) for commenting the manuscript.

References

Anonymous (1984). R.110 Agricultural pest act, 1983 (Act 36 of 1983). Control measures. *Government Gazette*, 9047:6–11.

Anonymous (2000). Council Directive 2000/29/EC of 8 May 2000 on protective measures against the introduction into the Community of organisms harmful to plants or plant products and against their spread within the Community. *Official Journal of the European Communities*, L 169:1–112.

Anonymous (2002). R.831 Agricultural pest act, 1983 (Act 36 of 1983). Control measures: Amendment. *Government Gazette*, 23517:15–17.

Anonymous (2005a). R.457 Agricultural pest act, 1983 (Act 36 of 1983). Control measures: Amendment. *Government Gazette*, 27580:3–4.

Anonymous (2005b). R.563 Correction notice. Agricultural pest act, 1983 (Act 36 of 1983). Control measures: Amendment. *Government Gazette*, 27665:5–6.

Anonymous (2006). Commission Decision of 5 July 2006 recognising certain third countries and certain areas of third countries as being free from *Xanthomonas campestris* (all strains pathogenic to *Citrus*), *Cercospora angolensis* Carv. *et* Mendes and *Guignardia citricarpa* Kiely (all strains pathogenic to *Citrus*). *Official Journal of the European Union*, L 187:35–36.

Anonymous (2008). R.461 agricultural pest act, 1983 (act 36 of 1983). control measures: Amendment. *Government Gazette*, 30988:4–5.

Anonymous (2014a). R.442 agricultural pest act, 1983 (act 36 of 1983). control measures: Amendment. *Government Gazette*, 37702:4–11.

Anonymous (2014b). Title7: Agriculture. part 319 foreign quarantine notices. subpart 56 fruits and vegetables. u.s. *U.S. Government Printing Office. Code of Federal Regulations*, pages 304–373.

APHIS, Animal and Plant Health Inspection Service USA (2012). Pest-free areas. `http://www.aphis.usda.gov/import_export/plants/manuals/ports/downloads/DesignatedPestFreeAreas.pdf`. Accessed on 5 December 2014.

Araújo, D. D., Raetano, C. G., Ramos, H. H., Spósito, M. B., and Prado, E. P. (2013). Interferência da redução no volume de aplicação sobre o controle da mancha preta (*Guignardia citricarpa* Kiely) em frutos de laranja'Valência'. *Summa Phytopathologica*, 39:172–179.

Aschmann, H. (1973). Distribution and peculiarity of Mediterranean ecosystems. In Castri, F. D. and Mooney, H. A., editors, *Mediterranean type ecosystems*, pages 11–19. New York: Springer-Verlag.

Bivand, R. (2014). spdep: Spatial dependence: weighting schemes, statistics and models. R package version 0.5-57. `http://CRAN.R-project.org/package=spdep`.

Bivand, R., Keitt, T., and Rowlingson, B. (2014). rgdal: bindings for the geospatial data abstraction library. R package version 0.8-16. `http://CRAN.R-project.org/package=rgdal`.

Carstens, E., Roux, H. F. l., Holtzhausen, M. A., Rooyen, L. v., Coetzee, J., Wentzel, R., Laubscher, W., Dawood, Z., Venter, E., Schutte, G. C., Fourie, P. H., and Hattingh, V. (2012). Citrus black spot is absent in the Western Cape, Northern Cape and Free State Provinces. *South African Journal of Science*, 108:56–61.

Chandelier, A., Helson, M., Dvorak, M., and Gischer, F. (2014). Detection and quantification of airborne inoculum of *Hymenoscyphus pseudoalbidus* using real-time PCR assays. *Plant Pathology*, 63(6):1296–1305.

DAFF, Department of Agriculture Forestry and Fisheries South Africa (2009). Movement of citrus plants and other citrus related plants. citrus maps poster 2009. `http://www.nda.agric.za/doaDev/sideMenu/plantHealth/docs/CitrusMapsPoster2009.pdf`. Accessed on 21 October 2014.

DAFF, Department of Agriculture Forestry and Fisheries South Africa (2014). The standards of plant quarantine on fresh orange grapefruit and lemon produced in the Republic of South Africa and on fresh orange and grapefruit produced in the Kingdom of Swaziland. `http://www.nda.agric.za/doaDev/sideMenu/plantHealth/Japancitrusprotocol.htm`. Accessed on 5 December 2014.

Doidge, E. M. (1929). Some diseases of citrus prevalent in South Africa. *South African Journal of Science*, 26:320–325.

EFSA, European Food Safety Authority (2008). Scientific opinion of the panel on plant health (PLH) on a request from the European Commission on *Guignardia citricarpa* Kiely. *The EFSA Journal*, 925:1–108.

EFSA, European Food Safety Authority (2014). Scientific opinion on the risk of *Phyllosticta citricarpa* (*Guignardia citricarpa*) for the EU territory with identification and evaluation of risk reduction options. *EFSA Journal*, 12:3557.

Er, H., Roberts, P., Marois, J., and van Bruggen, A. (2013). Potential distribution of citrus black spot in the United States based on climatic conditions. *European Journal of Plant Pathology*, 137(3):635–647.

FAO, Food and Agriculture Organization of the United Nations. (2014). Crop production database FAOSTAT. `http://faostat.fao.org/default.aspx`. Accessed on 12 December 2014.

Fourie, P., Schutte, T., Serfontein, S., and Swart, F. (2013). Modeling the effect of temperature and wetness on *Guignardia* pseudothecium maturation and ascospore release in citrus orchards. *Phytopathology*, 103(3):281–292.

Gebrehiwet, Y., Ngqangweni, S., and Kirsten, J. F. (2007). Quantifying the trade effect of sanitary and phytosanitary regulations of OECD countries on South African food exports. *Agrekon*, 46(1):23–39.

Graham, J. H., Gottwald, T. R., Timmer, L. W., Bergamin Filho, A., Van Den Bosch, F., Irey, M. S., Taylor, E., Magarey, R. D., and Takeuchi, Y. (2014). Response to "Potential distribution of citrus black spot in the United States based on climatic conditions", Er et al. 2013. *European Journal of Plant Pathology*, 139(2):231–234.

Hijmans, R. J. (2014). raster: geographic data analysis and modeling. R package version 2.2-31. `http://CRAN.R-project.org/package=raster`.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25:1965–1978.

IPPC, International Plant Protection Convention (1995). Requirements for the establishment of pest free areas. International Standards for Phytosanitary Measures, ISPM 4. Rome: IPPC.

IPPC, International Plant Protection Convention (2005). Requirements for the establishment of areas of low pest prevalence. International Standards for Phytosanitary Measures, ISPM 22. Rome: IPPC.

IPPC, International Plant Protection Convention (2007). Recognition of pest free areas and areas of low pest prevalence. International Standards for Phytosanitary Measures, ISPM 29. Rome: IPPC.

Kiely, T. B. (1948). Preliminary studies on *Guignardia citricarpa*, n. sp.: The ascigenous stage of *Phoma citricarpa* McAlp. and its relation to black spot of citrus. *Proceedings of the Linnean Society of New South Wales*, 68:249–292.

Klausmeyer, K. R. and Shaw, M. R. (2009). Climate change, habitat loss, protected areas and the climate adaptation potential of species in Mediterranean ecosystems worldwide. *PloS One*, 4(7):e6392.

Korf, H. J. G. (1998). Survival of *Phyllostica citricarpa*, anamorph of the citrus black spot pathogen. Master's thesis, Pretoria: University of Pretoria.

Kotzé, J. (1981). Epidemiology and control of citrus black spot in South Africa. *Plant Disease*, 65:945–950.

Kotzé, J. M. (2000). Black spot. In L. W. Timmer, S. M. G. and Graham, J. H., editors, *Compendium of Citrus Diseases 2nd ed.*, pages 10–12. St. Paul, MN: APS Press.

Köppen, W. (1936). Das geographisca system der klimate. In Köppen, W. and Geiger, G., editors, *Handbuch der Klimatologie*, page 44. Berlin: Gebrüder Borntraeger.

Latimer, A. M., Wu, S., Gelfand, A. E., and Silander Jr, J. A. (2006). Building statistical models to analyze species distributions. *Ecological Applications*, 16(1):33–50.

Lee, Y., Huang, C., et al. (1973). Effect of climatic factors on the development and discharge of ascospores of the citrus black spot fungus. *Journal of Taiwan Agricultural Research*, 22(2):135–144.

Loecher, M. (2014). RgoogleMaps: overlays on Google map tiles in R. R package version 1.2.0.6. `http://CRAN.R-project.org/package=RgoogleMaps`.

Magarey, R., Chanelli, S., and Holtz, T. (2011). Validation study and risk assessment: *Guignardia citricarpa*, (citrus black spot). Technical report, USDA-APHIS-PPQ-CPHST-PERAL/NCSU.

Magarey, R., Sutton, T., and Thayer, C. (2005). A simple generic infection model for foliar fungal plant pathogens. *Phytopathology*, 95(1):92–100.

MAGRAMA, Ministerio de Agricultura, Alimentación y Medio Ambiente (2013). Anuario de estadística 2013. (pp. 1095). Madrid: MAGRAMA, Secretaría General Técnica, Centro de Publicaciones.

Makowski, D., Vicent, A., Pautasso, M., Stancanelli, G., and Rafoss, T. (2014). Comparison of statistical models in a meta-analysis of fungicide treatments for the control of citrus black spot caused by *Phyllosticta citricarpa*. *European Journal of Plant Pathology*, 139:79–94.

McOnie, K. (1964a). Apparent absence of *Guignardia citricarpa* Kiely from localities where citrus black spot is absent. *South African Journal of Agricultural Science*, 7:347–354.

McOnie, K. (1964b). Orchard development and discharge of ascospores of *Guignardia citricarpa* and onset of infection in relation to control of citrus black spot. *Phytopathology*, 54(12):1448–1454.

McOnie, K. C. (1964c). The latent occurrence in citrus and other hosts of *Guignardia* easily confused with *G. citricarpa*, the citrus black spot pathogen. *Phytopathology*, 54:40–43.

Mondal, S., Gottwald, T., and Timmer, L. (2003). Environmental factors affecting the release and dispersal of ascospores of *Mycosphaerella citri*. *Phytopathology*, 93(8):1031–1036.

Paul, I. (2005). *Modelling the distribution of citrus black spot caused by Guignardia citricarpa Kiely*. PhD thesis, Pretoria: University of Pretoria.

Paul, I., Van Jaarsveld, A., Korsten, L., and Hattingh, V. (2005). The potential global geographical distribution of Citrus Black Spot caused by *Guignardia citricarpa* (Kiely): likelihood of disease establishment in the European Union. *Crop Protection*, 24:297–308.

Pebesma, E. J. and Bivand, R. S. (2005). Classes and methods for spatial data in R. *R News*, 5:9–13.

Peel, M. C., Finlayson, B. L., and McMahon, T. A. (2007). Updated world map of the köppen-geiger climate classification. *Hydrology and Earth System Sciences*, 11:1633–1644.

Perryman, S., Clark, S., and West, J. (2014). Splash dispersal of *Phyllosticta citricarpa* conidia from infected citrus fruit. *Scientific Reports*, 4:6568.

Plant, R. E. (2012). *Spatial data analysis in ecology and agriculture using R*. Boca Raton, FL:CRC Press.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reuther, W., Webber, H. J., Batchelor, L. D., et al. (1967). *The citrus industry. Volume 1. History, world distribution, botany and varieties.* Berkeley, CA: University of California Press.

Rieux, A., Soubeyrand, S., Bonnot, F., Klein, E. K., Ngando, J. E., Mehl, A., Ravigne, V., Carlier, J., and de Bellaire, L. d. L. (2014). Long-distance wind-dispersal of spores in a fungal plant pathogen: estimation of anisotropic dispersal kernels from an extensive field experiment. *PLoS One*, 9(8):e103225.

Schubert, T., Dewdney, M., Peres, N., Palm, M., Jeyaprakash, A., Sutton, B., Mondal, S., Wang, N., Rascoe, J., and Picton, D. (2012). First report of citrus black spot caused by *Guignardia citricarpa* on sweet orange [*Citrus sinensis* (L.) Osbeck] in North America. *Plant Disease*, 96(8):1225.

Simberloff, D. (2009). The role of propagule pressure in biological invasions. *Annual Review of Ecology, Evolution, and Systematics*, 40:81–102.

Spósito, M., Amorim, L., Ribeiro Jr, P., Bassanezi, R., and Krainski, E. (2007). Spatial pattern of trees affected by black spot in citrus groves in brazil. *Plant Disease*, 91:36–40.

Spósito, M. B., Amorim, L., Bassanezi, R. B., Yamamoto, P. T., Felippe, M. R., and Czermainski, A. B. (2011). Relative importance of inoculum sources of *Guignardia citricarpa* on the citrus black spot epidemic in Brazil. *Crop Protection*, 30:1546–1552.

Vicent, A. and García-Jiménez, J. (2008). Risk of establishment of non-indigenous diseases of citrus fruit and foliage in Spain: an approach using meteorological databases and tree canopy climate data. *Phytoparasitica*, 36(1):7–19.

Wager, V. A. (1952). The black spot disease of citrus in South Africa. *Science Bulletin of the Department of Agriculture of the Union of South Africa*, 303:1–52.

Whiteside, J. (1967). Sources of inoculum of the black spot fungus, *Guignardia citricarpa*, in infected rhodesian citrus orchards. *Rhodesia, Zambia and Malawi Journal of Agricultural Research*, 5:171–177.

Yonow, T., Hattingh, V., and de Villiers, M. (2013). CLIMEX modelling of the potential global distribution of the citrus black spot disease caused by *Guignardia citricarpa* and the risk posed to Europe. *Crop Protection*, 44:18–28.

Chapter 7

Response to the letter on "Climatic distribution of citrus black spot caused by *Phyllosticta citricarpa*. A historical analysis of disease spread in South Africa" by Fourie et al. (2017)

In this chapter, we present a version of our paper "Response to the letter on "Climatic distribution of citrus black spot caused by *Phyllosticta citricarpa*. A historical analysis of disease spread in South Africa" by Fourie et al. (2017)" by Joaquín Martínez-Minaya (University of Valencia), David Conesa (University of Valencia), Antonio López Quílez (University of Valencia) and Antonio Vicent (Valencian Institute for Agricultural Research) published in *European Journal of Plant Pathology*, 148, 503–508. The chapter contains at the end the references used in this work.

Abstract

In a previous study, Martínez-Minaya et al. (2015) performed an analysis of climate-based distribution of citrus black spot (CBS) in South Africa. It was found that CBS was initially confined to humid areas with summer rainfall, but later spread to arid steppe and even desert climates. A strong spatial autocorrelation of CBS distribution was found. Fourie et al. (2017) take a critical view of our study, but without presenting any analysis of results to refute our findings. Furthermore, Fourie et al. (2017) appear to have misunderstood our work, since many of their criticisms relate to the potential distribution of CBS in Europe, which is beyond the scope of our original study. Fourie et al. (2017) highlight the limitations of climate classifications in species distribution modelling. However, this was made explicit in our study, indicating that it was a preparatory work and further advanced modelling studies, including spatial effects, will be needed. Fourie et al. (2017) incorrectly assume that we used all of South Africa as the background in the spatial autocorrelation analysis. However, only citrus areas were used and a strong spatial autocorrelation was detected at all distances evaluated. Contrary to what Fourie et al. (2017) suggest, similar climate distributions of CBS were obtained at $5'$ and $30'$ resolution, and also with the national land-cover map of South Africa. The figure comparison presented by Fourie et al. (2017) appears to ignore the fact that the maps we used were grid cells of 10×10 km and not the line polygons they suggest. Therefore, we consider the conclusions from the Martínez-Minaya et al. (2015) remain entirely valid.

Keywords

Guignardia citricarpa, spatial autocorrelation, mapping

Fourie et al. (2017) devote the greater proportion of their letter discussing the potential global distribution of citrus black spot (CBS), caused by *Phyllosticta citricarpa* (McAlpine) van der Aa, with a particular emphasis in European citrus-producing regions. However, it was clearly stated in the title and the objectives of our study (Martínez-Minaya et al., 2015) that it was limited to South Africa. Martínez-Minaya et al. (2015) stated that "maps of the Mediterranean Basin were also obtained to discuss the boundaries and geographic extent of Mediterranean-type climates". However, climatic suitability of the Mediterranean Basin for CBS was not analysed nor discussed in our study. Therefore, our response will not address those comments of Fourie et al. (2017) relating to the potential distribution of CBS in Europe. For a detailed discussion on this interesting topic, we recommend a recent report by EFSA, European Food Safety Authority (2016), where our study and others were thoroughly assessed by an independent panel of scientists.

We focus our response primarily on the methodological issues raised by Fourie et al. (2017), as they might affect the conclusions of Martínez-Minaya et al. (2015). To complement the results obtained by Martínez-Minaya et al. (2015), an additional raster layer was assembled with the map published by Paul (2005) and its subsequent updates (Yonow et al., 2013; Anonymous, 2014), but including only those grid cells of the class "cultivated commercial permanent orchards" in the 2013-2014 South African national land-cover (NLC) dataset (DEA, Department of Environmental Affairs South Africa, 2015). As in Martínez-Minaya et al. (2015), a resolution of $5'$ and the coordinate system WGS84 were used with the raster package for R (Hijmans, 2014).

Fourie et al. (2017) indicate that CBS distribution patterns in South Africa depended on the point of introduction and the movement of infected plant material, which we agree is self-evident. Furthermore, Fourie et al. (2017) point out that "*P. citricarpa* has had abundant opportunity over many years for range expansion, including the recorded movement of citrus trees from CBS-endemic areas (Powell, 1930; Kiely, 1948; Ramón-Laca, 2003)". Only the reference by Powell (1930) relates to South Africa, and

CBS was not mentioned in the publication, as the disease was only reported in South Africa in 1929 (Doidge, 1929). Powell (1930) indicated that citrus was first introduced in the Western Cape in 1654, from where the crop progressively expanded east. Recent phytosanitary regulations in South Africa still consider the Western Cape to be a CBS-free area, whereas most citrus regions in the east are CBS-affected (Anonymous, 2014). Therefore, Powell (1930) cannot be considered by any means as a valid reference for the movement of citrus trees from CBS-endemic areas in South Africa. As indicated by Martínez-Minaya et al. (2015), "the movement of citrus material in South Africa was not regulated until 1984, but quantitative trade data among provinces was not found". Fourie et al. (2017) do not provide any additional data or reference on this subject.

Fourie et al. (2017) indicate that the use of climate classifications is the most simplistic of all the species distribution models available, and so the biological relevance of climate zones should be carefully considered. We were fully aware of this point, as EFSA, European Food Safety Authority (2014) already indicated that global climate zones are based on factors and thresholds that are broad and not necessarily representative of those that are critical for the pathogen and its host. This point was stated explicitly by Martínez-Minaya et al. (2015) in the objectives: "This preparatory work was part of a larger modelling project where the potential geographical range of CBS will be estimated based on relevant environmental variables and spatial effects", and in the conclusions: "Further modelling studies should integrate the relative contribution of environmental variables together with the spatial structure of the data to better estimate the potential geographical range of CBS". Fourie et al. (2017) do not acknowledge these statements in their letter.

Martínez-Minaya et al. (2015) indicated that "A map of the CBS distribution in Australia was also available (Paul, 2005), but without details and resolution of the original data, so it was not considered in the present study". However, Fourie et al. (2017) criticize our study for not considering CBS data from Australia and claim that both CBS distribution maps, Australia and South Africa, had a similar level of detail. For South Africa, Paul (2005) indicated that "areas of CBS presence and absence in commercial

orchards and backyard trees were mapped by six field specialists with extensive knowledge of the disease onto a map of South Africa at a scale 1:106 (2×2 m). Disease presence records (...) were transcribed to a 29.7×45 cm map and scanned. Data on CBS distribution were confirmed by 200 citrus growers and researchers from South Africa at a citrus meeting in 2002". However, for Australia, Paul (2005) only indicated that "Information on the presence of CBS in Australia was obtained from the Australian Plant Pest Database" and "A map of the known occurrence of CBS in Australia was drawn up from these data". Hence, the paucity of details in Paul (2005) on the original data from Australia when compared with those from South Africa is self-evident.

With regard to our analysis of spatial autocorrelation of CBS distribution in South Africa, Fourie et al. (2017) point out that "Martínez-Minaya et al. (2015) used all of South Africa as the background for the analysis (...). Had Martínez-Minaya et al. (2015) used citrus production regions as the background for the autocorrelation analysis, the apparent levels of spatial autocorrelation can be expected to decrease significantly". Fourie et al. (2017) make these serious assertions without presenting any analysis of spatial autocorrelation of the data. Furthermore, it is an incorrect assumption of Fourie et al. (2017) that we used all of South Africa as the background. Moran's I and Geary's C analyses were performed with the 2014 dataset considering only grid cells in citrus areas, assigning a value of 0 for CBS absence and 1 for CBS presence or low prevalence. Indeed, Moran's I and Geary's C were not calculated for the 1950 dataset because only CBS presence, and not CBS absence, was available for that year.

In addition to the Moran's I and Geary's C calculated with contiguity-based neighbours by Martínez-Minaya et al. (2015), we present all the values for these indices at increasing distances (Figure 7.1). The presence of strong spatial autocorrelation in the current CBS distribution data in citrus areas in South Africa was evident in both, the dataset used by Martínez-Minaya et al. (2015) and the one assembled based on the NLC map. As pointed out by Martínez-Minaya et al. (2015), further modelling efforts should consider not only environmental variables, but also the spatial dependence of CBS distribution data in South Africa. Ignoring this dependence may lead to

inaccurate model parameterization and inadequate quantification of uncertainty (Banerjee et al., 2014).

In Martínez-Minaya et al. (2015), a raster layer of CBS distribution in South Africa was generated from the map published by Paul (2005) and its subsequent updates (Yonow et al., 2013; Anonymous, 2014). As indicated in the Material and Methods of Martínez-Minaya et al. (2015), and also noted in the acknowledgments, all data were georeferenced to the coordinate system WGS84 by a mapping specialist. In their letter, Fourie et al. (2017) point out that CBS distribution maps in Paul et al. (2005) are adequate for modelling at $30'$ scale, but not at a $5'$ scale used in our study. Martínez-Minaya et al. (2015) stated that "similar results (not shown for the sake of simplicity) were obtained with the $30'$ resolution". To demonstrate this, we now present the proportion of grid cells according to CBS status by Köppen-Geiger climate types (Köppen, 1936) at both the $30'$ and $5'$ resolution obtained by Martínez-Minaya et al. (2015) (Figure 7.2). As previously indicated and as can now be seen, similar results were obtained at both scales. Furthermore, comparable results were derived from the dataset assembled using the NLC map at $5'$ resolution, where only grid cells with "cultivated commercial permanent orchards" were considered (Figure 7.2c). In this dataset, no commercial citrus areas were located under the arid cold desert (BWk) climate, which were indeed considered as CBS-free by Martínez-Minaya et al. (2015).

Fourie et al. (2017) plot side by side Figure 7.1d of Martínez-Minaya et al. (2015) and Figure 7.1a of Yonow et al. (2013), stating that "polygons depicting CBS distribution were clearly coarser that those of Yonow et al. (2013)". Although Martínez-Minaya et al. (2015) made it explicit in Material and Methods, Fourie et al. (2017) apparently misunderstood that our maps were in fact grid cells and not line polygons as in Yonow et al. (2013). Each grid cell represented a $5'$ square of about 10×10 km, which was evident from the scale bars in both Figure 7.1 and Figure 7.2 of Martínez-Minaya et al. (2015). Values of CBS status and environmental variables were for the grid centroids, always contained within the polygons of the map published by Paul (2005) and its subsequent updates (Yonow et al., 2013; Anonymous, 2014). Fourie et al. (2017) persist in this misapprehension stating that "CBS-present polygons also extended into neighbouring countries and

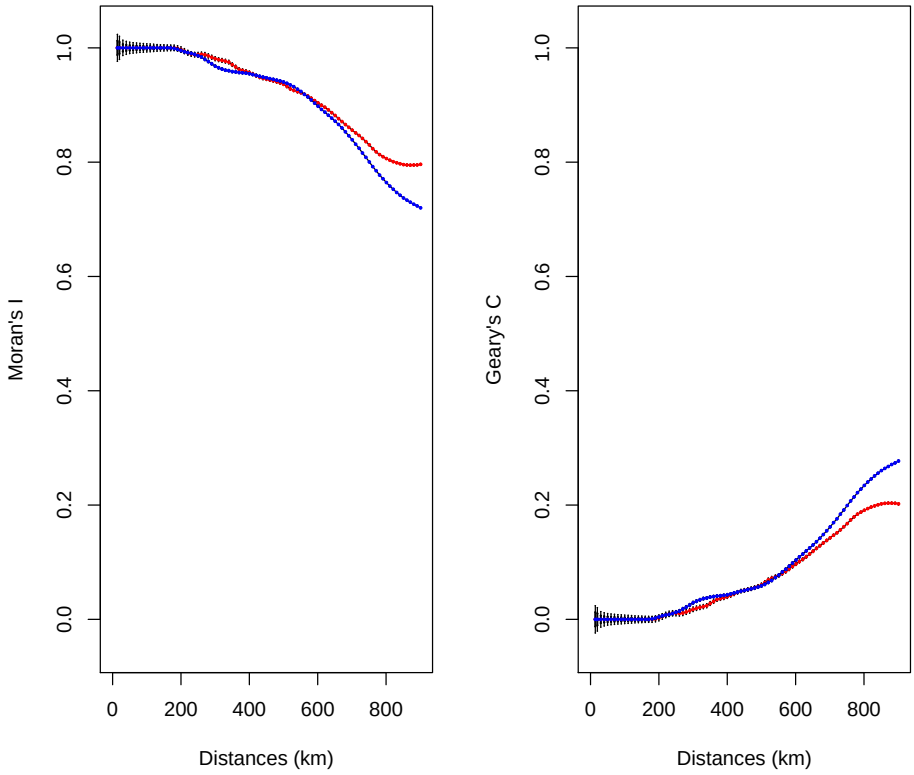

FIGURE 7.1: Moran's I and Geary's C values at increasing distances. The blue lines represent the dataset used by Martínez-Minaya et al. (2015) from Paul (2005) and its subsequent updates (Yonow et al., 2013; Anonymous, 2014). The red lines represent the same dataset but including only grid cells of the class "cultivated commercial permanent orchards" from the 2013-2014 South African national land-cover map (DEA, Department of Environmental Affairs South Africa, 2015).

even into the ocean". Again, it should be noted that grid cells and not polygons were represented by Martínez-Minaya et al. (2015). Moreover, the WorldClim database includes only land areas and not oceans (Hijmans et al., 2005). Also, as clearly indicated in Martínez-Minaya et al. (2015), our study was limited to South Africa. For more in-depth information on this point, the functions 'getData' and 'crop' in the raster package for R as used in our study should be examined (Hijmans, 2014).

Fourie et al. (2017) state that "it is clear that CBS occurs in the BSh climate zone in South Africa, and is essentially absent from BSk climates". Again, Fourie et al. (2017) make this strong assertion without presenting any analysis of the data. Presence of CBS under the cold arid steppe climate (BSk) was obtained by Martínez-Minaya et al. (2015) at 5′ and 30′ resolutions, as well as here with the dataset assembled using the NLC map (Figure 7.2). Fourie et al. (2017) focus more directly on the BSk climate, but they appear to overlook that the hot arid steppe climate (BSh), which is the predominant climate under which CBS is becoming established in South Africa (Figure 7.2), is also a climate type occurring in important citrus-producing areas in the Mediterranean Basin (Fig. 3a of Martínez-Minaya et al. (2015)).

From a biogeographical perspective, it is more remarkable that *P. citricarpa* thrives under the hot arid desert climate (BWh) found in parts of South Africa (2.3% of grid cells), although Fourie et al. (2017) made no comments on this fact. The arid desert areas where CBS is present in South Africa are located in northern Limpopo province (Figure 7.2a of Martínez-Minaya et al. (2015)), as clearly specified by phytosanitary regulations "The Limpopo province, towns of Musina and Soutpansberg - north of 22º 50'S or west of 29° 20' E " (Anonymous 2014). These areas have low pest (disease) prevalence for CBS (Anonymous 2014), so they are subject to effective surveillance, control or eradication measures (IPPC, International Plant Protection Convention, 2005, 2007). The presence of CBS in desert areas demonstrates that *P. citricarpa* is able to complete its disease cycle under arid conditions typical of the BWh climate classification. Indeed, with the dataset assembled using the NLC map, annual rainfall as low as 340 mm was recorded in areas were CBS is endemic, similar to the values reported by Martínez-Minaya et al. (2015).

The incorrect assumptions made by Fourie et al. (2017) in respect to our methods, their apparent misinterpretations of our results, and taking into account the methodological clarifications and additional evidence we present here, we consider that the conclusions of Martínez-Minaya et al. (2015) are indeed correct, valid, and stand, further demonstrating the facts that: i) CBS in South Africa has expanded from its original geographic range in summer rainfall areas to adjacent, more arid regions; ii) the results contradict statements indicating that CBS occurs exclusively in climates with

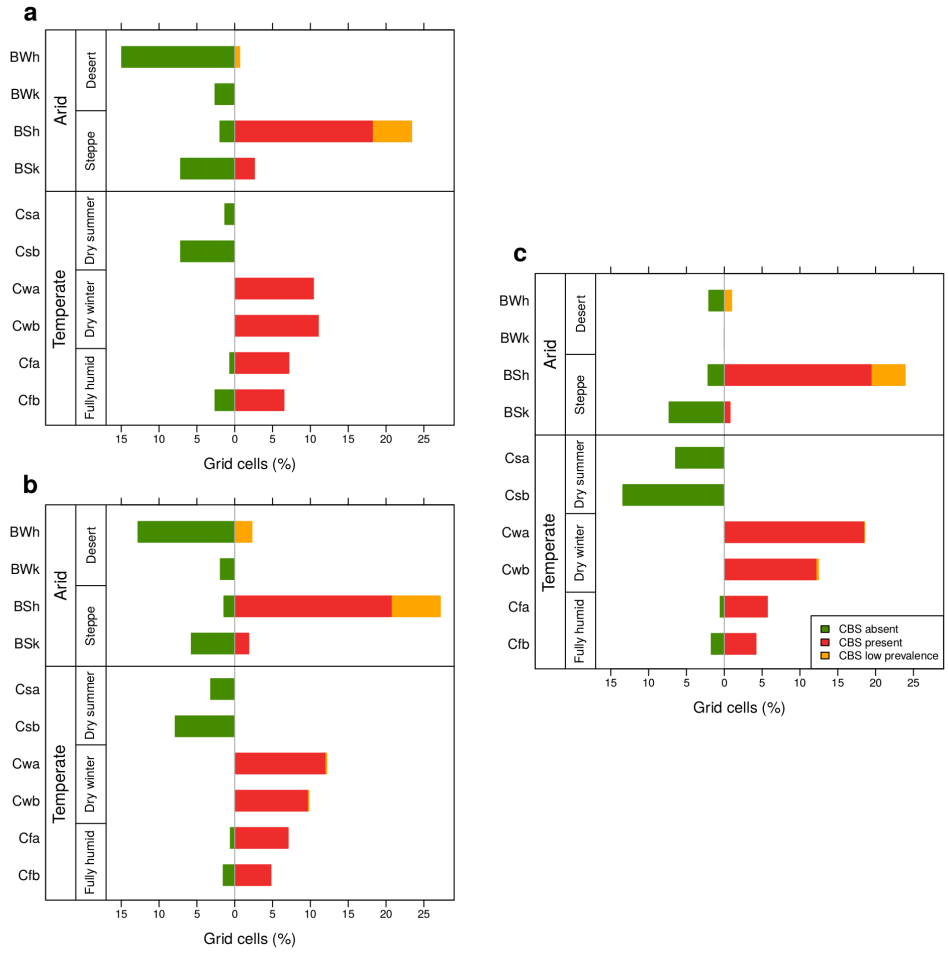

FIGURE 7.2: Köppen-Geiger climate types and citrus-growing areas in relation to the distribution of citrus black spot (CBS) caused by *Phyllosticta citricarpa* in South Africa. The dataset used by Martínez-Minaya et al. (2015) from Paul (2005) and its subsequent updates (Yonow et al., 2013; Anonymous, 2014) shown at 30′ (**a**) and 5′ (**b**) resolution, and (**c**) the same dataset at 5′ resolution but including only grid cells of the class "cultivated commercial permanent orchards" in the 2013-2014 South African national land-cover map (DEA, Department of Environmental Affairs South Africa, 2015).

summer rainfall (Fourie et al., 2017; Graham et al., 2014; Kotzé, 2000); and iii) further modelling studies are required to integrate the relative contribution of environmental variables and the spatial structure of the data.

References

Anonymous (2014). R.442 agricultural pest act, 1983 (act 36 of 1983). control measures: Amendment. *Government Gazette*, 37702:4–11.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. Crc Press.

DEA, Department of Environmental Affairs South Africa (2015). South African national land-cover dataset 2013-2014. Geoterraimage. `http://egis.environment.gov.za`. Accessed on 8 February 2016.

Doidge, E. M. (1929). Some diseases of citrus prevalent in South Africa. *South African Journal of Science*, 26(12):320–325.

EFSA, European Food Safety Authority (2014). Scientific opinion on the risk of *Phyllosticta citricarpa* (*Guignardia citricarpa*) for the EU territory with identification and evaluation of risk reduction options. *EFSA Journal*, 12:3557.

EFSA, European Food Safety Authority (2016). Evaluation of new scientific information on *Phyllosticta citricarpa* in relation to the EFSA PLH Panel (2014). Scientific Opinion on the plant health risk to the EU. *EFSA Journal*, 14:4513.

Fourie, P. H., Schutte, G. C., Carstens, E., Hattingh, V., Paul, I., Magarey, R. D., Gottwald, T. R., Yonow, T., and Kriticos, D. J. (2017). Scientific critique of the paper "Climatic distribution of citrus black spot caused by *Phyllosticta citricarpa*. A historical analysis of disease spread in South Africa" by Martínez-Minaya et al.(2015). *European Journal of Plant Pathology*, 148(3):497–502.

Graham, J. H., Gottwald, T. R., Timmer, L. W., Bergamin Filho, A., Van Den Bosch, F., Irey, M. S., Taylor, E., Magarey, R. D., and Takeuchi,

Y. (2014). Response to "Potential distribution of citrus black spot in the United States based on climatic conditions", Er et al. 2013. *European Journal of Plant Pathology*, 139(2):231–234.

Hijmans, R. J. (2014). raster: geographic data analysis and modeling. R package version 2.2-31. `http://CRAN.R-project.org/package=raster`.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25:1965–1978.

IPPC, International Plant Protection Convention (2005). Requirements for the establishment of areas of low pest prevalence. International Standards for Phytosanitary Measures, ISPM 22. Rome: IPPC.

IPPC, International Plant Protection Convention (2007). Recognition of pest free areas and areas of low pest prevalence. International Standards for Phytosanitary Measures, ISPM 29. Rome: IPPC.

Kiely, T. B. (1948). Preliminary studies on *Guignardia citricarpa*, n. sp.: The ascigenous stage of *Phoma citricarpa* McAlp. and its relation to black spot of citrus. *Proceedings of the Linnean Society of New South Wales*, 68:249–292.

Kotzé, J. M. (2000). Black spot. In L. W. Timmer, S. M. G. and Graham, J. H., editors, *Compendium of Citrus Diseases 2nd ed.*, pages 10–12. St. Paul, MN: APS Press.

Köppen, W. (1936). Das geographisca system der klimate. In Köppen, W. and Geiger, G., editors, *Handbuch der klimatologie*, page 44. Berlin: Gebrüder Borntraeger.

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

Paul, I. (2005). *Modelling the distribution of citrus black spot caused by Guignardia citricarpa Kiely*. PhD thesis, Pretoria: University of Pretoria.

Paul, I., Van Jaarsveld, A., Korsten, L., and Hattingh, V. (2005). The potential global geographical distribution of Citrus Black Spot caused by *Guignardia citricarpa* (Kiely): likelihood of disease establishment in the European Union. *Crop Protection*, 24:297–308.

Powell, H. C. (1930). The culture of the orange and allied fruits. South African agricultural series No. 8. Johannesburg: Central News Agency.

Ramón-Laca, L. (2003). The introduction of cultivated citrus to Europe via Northern Africa and the Iberian Peninsula. *Economic Botany*, 57(4):502–514.

Yonow, T., Hattingh, V., and de Villiers, M. (2013). CLIMEX modelling of the potential global distribution of the citrus black spot disease caused by *Guignardia citricarpa* and the risk posed to Europe. *Crop Protection*, 44:18–28.

Chapter 8

Spatial and climatic factors associated with the geographical distribution of citrus black spot disease in South Africa. A Bayesian latent Gaussian model approach

In this chapter, we present a version of our paper "Spatial and climatic factors associated with the geographical distribution of citrus black spot disease in South Africa. A Bayesian latent Gaussian model approach" by Fourie et al. (2017)" by Joaquín Martínez-Minaya (University of Valencia), David Conesa (University of Valencia), Antonio López-Quílez (University of Valencia) and Antonio Vicent (Valencian Institute for Agricultural Research) published in *European Journal of Plant Pathology*, 143, 69–83. The chapter contains at the end the references used in this work.

Abstract

Citrus black spot (CBS), caused by *Phyllosticta citricarpa*, is the main fungal disease affecting this crop and quarantine measures were have been implemented. The role of climate as a limiting factor for the establishment and spread of CBS to new areas has been debated, but previous studies did not address the effects of spatial factors in the geographic distribution of the disease. The effects of climatic and spatial factors were studied using South Africa as a case study, due to its diversity of climates within citrus-growing regions. Georeferenced datasets of CBS presence/absence in citrus areas were assembled for two stages of the epidemicsepidemic: 1950 and 2014. Climatic variables were obtained from the WorldClim database. Moran's I and Geary's C analyses indicated that CBS distribution data presented significant spatial autocorrelation, particularly in 2014. Collinearity was detected among climatic variables. Spatial logistic regressions, particular case of latent Gaussian models, were fitted to CBS presence/absence in 1950 or 2014 with Integrated Nested Laplace Approximation methodology. Principal components (PCs) or pre-selection of climatic variables based on their correlation coefficients were used to cope with collinearity. Spatial effects were incorporated with a geostatistical term. In general, the models indicated a positive relationship between CBS presence and climatic variables or PCs associated with warm temperatures and high precipitation. Nevertheless, in 1950, models that also included a spatial effect outperformed those with climatic variables only. Problems of model convergence were detected in 2014 due to the strong spatial structure of CBS distribution data. The consequences of ignoring spatial dependence to estimate the potential geographical range of CBS are discussed.

Keywords

Guignardia citricarpa, geostatistics, INLA, biogeography, risk assessment

8.1 Introduction

Citrus black spot (CBS) disease, caused by the fungus *Phyllosticta citricarpa* (McAlpine) van der Aa (synonym *Guignardia citricarpa* Kiely), is the most important fungal disease affecting this crop worldwide. The pathogen causes external blemishes on the fruit rind and induces premature fruit drop, resulting in serious quality and yield losses (Martínez-Minaya et al., 2015). When complementary mating types are present, the pathogen reproduces through sexual spores (ascospores) formed in the leaf litter after a maturation process driven mainly by temperature and moisture (McOnie, 1964b; Tran et al., 2017). Once mature, ascospores are discharged from the leaf litter and disseminated by air currents (McOnie, 1964b). Ascospores infect susceptible fruit, twigs and leaves in the presence of moisture and conducive temperatures. The pathogen also reproduces by asexual spores (conidia), which are rain-splashed mainly from lesions in citrus fruit and twigs (Perryman et al., 2014). Ascospores have been deemed as the main source of inoculum in South Africa (McOnie, 1964b), but studies in Brazil and Florida (USA) have suggested that conidia are epidemiologically important under certain conditions (Hendricks et al., 2017; Spósito et al., 2008; Wang et al., 2016). The application of chemical fungicides is generally needed for CBS control (Makowski et al., 2014), resulting in increased environmental and economic costs of citrus production.

The CBS disease is currently present in important citrus-growing regions of Australia, Asia, Africa and America. Quarantine measures have been established by several countries, such as South Africa, USA, Japan and the European Union (EU), to prevent the entry of *P. citricarpa* into areas that are still free of the pathogen (Martínez-Minaya et al., 2015). According to the International Plant Protection Convention (IPPC) and the World Trade Organization (WTO), phytosanitary regulations should be based on a scientific pest risk analysis (PRA). PRAs are based on standardized protocols aimed at estimating the likelihood of disease introduction (i.e. entry and establishment) and subsequent spread in order to devise the most efficient options as regards risk reduction. Maps describing host availability and climatic suitability for disease development are a key component of PRAs to set bounds on potential introductions into new areas (Venette et al., 2010).

Paul et al. (2005) estimated the potential global geographical range of CBS using CLIMEX. They concluded that climates in the Mediterranean Basin were not suitable for CBS and, therefore, phytosanitary measures for *P. citricarpa* in the EU were not necessary. However, Paul et al. (2005) were not able to predict the presence of the disease in the arid citrus-growing areas of the Eastern Cape province in South Africa, where CBS is endemic. In a subsequent study, Yonow et al. (2013) modified the parameters of Paul et al. (2005) allowing CLIMEX to predict the presence of CBS in this region. Using a new set of CLIMEX parameters, Er et al. (2013) predicted climatic suitability for CBS in arid areas of Mediterranean-type climates in California (USA).

A mechanistic (process-based) generic infection model (Magarey et al., 2005) was used to obtain maps of climate suitability for CBS. This generic infection model consisted of parameters for temperature and wetness duration, and it was specifically developed for exotic pathogens, like *P. citricarpa*, on which there is little biological information. One study using this model concluded that the climates of the EU cannot be considered as unsuitable for the establishment of *P. citricarpa* (EFSA, European Food Safety Authority, 2008), whereas another study indicated that CBS was not expected to have an impact in areas with commercial citrus production in Europe (Magarey et al., 2011). Nevertheless, due to the paucity of biological information available for *P. citricarpa*, the results obtained with process-based models were highly uncertain (EFSA, European Food Safety Authority, 2014).

Empirical models for *Phyllosticta* spp. ascospore maturation and release (Fourie et al., 2013) have been combined with the generic infection model so as to restrict predictions only to the periods of potential ascospore availability. One study using these models indicated that climatic conditions in many EU citrus-growing areas were suitable for CBS (EFSA, European Food Safety Authority, 2014), whereas another suggested that only a few isolated locations in Europe have a low to marginal risk of *P. citricarpa* establishment (Magarey et al., 2015). However, in this latter study, infection events were dramatically diminished, only those associated with rains being considered. Nevertheless, as indicated above, large uncertainties have been associated with these models and their applications (EFSA, European Food Safety Authority, 2016).

Correlative statistical models are widely used in different areas of biogeography, such as conservation, resource management, global warming and biological invasions (Franklin, 2009). However, their use for risk assessment in plant pathology is still limited and few studies are available for diseases caused by fungi or oomycetes (Elith et al., 2013; Meentemeyer et al., 2008; Narouei-Khandan et al., 2017). Typically, correlative species distribution models explore the relationships between species occurrences and climatic variables to produce maps of predicted distributions of the target organisms. Without enough biological and epidemiological information for process-based models, correlative species distribution models may help to identify climatic variables that are associated with CBS and, therefore, demarcate locations that would allow disease establishment.

Several statistical methods are used for species distribution modelling based on presence/absence data, such as generalized linear models (GLM) and generalized additive models (GAM) (Franklin, 2009). In many cases, species distribution models are used without considering the spatial dependence of the data, assuming that the geographical range is only driven by climate and the disease is in equilibrium with these factors. However, this assumption is often violated when disease spread is constrained due to dispersal barriers and/or absence of susceptible host plants. Moreover, ignoring spatial autocorrelation may lead to inaccurate parameter estimates, inadequate quantification of uncertainty, and thus poor predictive power. With spatially explicit hierarchical Bayesian models it is possible to introduce the effect of spatial dependence (Latimer et al., 2006). These complex models have usually been fitted with Markov chain Monte Carlo (MCMC) methods that are computationally costly, especially for large spatial datasets. In the specific case of latent Gaussian models, Approximate Bayesian inference with integrated nested Laplace approximations (INLA) is a much faster and computationally efficient alternative to MCMC (Lindgren et al., 2011; Rue et al., 2009).

The main objective of this study was to analyse the spatial and climate effects that influence the probability of CBS occurrence in South Africa. South Africa was used here as a case study because it is the only country with commercial citrus production under ten climate types, covering a wide range of environmental conditions. This information will help risk managers

to better understand the factors associated with the potential establishment and spread of CBS into new areas.

8.2 Materials and Methods

8.2.1 Datasets

Spatially gridded datasets including presence and absence of CBS in citrus-growing areas in South Africa were assembled for 1950 and 2014. A raster layer (5′ arc min resolution) of citrus distribution in South Africa was generated by a mapping specialist from the map of citrus trees in South Africa published by Powell (1930), based on the census carried out in 1927 and restricted to the boundaries of South Africa (Figure 8.1a). Grid cells with citrus were classified as CBS-present ($n = 28$) and CBS-absent ($n = 776$) based on the survey included in Appendix 2 of Wager (1952), which was conducted from 1940 to 1950. This latter year was used to denote the dataset. Since the coexistence of pathogenic and non-pathogenic species of *Phyllosticta* in citrus was not discovered until a decade later (McOnie, 1964c), reports of the pathogen in absence of CBS symptoms were excluded from Wager (1952).

For 2014, the spatially gridded South African national land-cover (NLC) dataset 2013-2014 was used (DEA, Department of Environmental Affairs South Africa, 2015), but including only those grid cells of the class "cultivated commercial permanent orchards" enclosed within the citrus areas of the map by Paul (2005) and its subsequent updates (Anonymous, 2014; Yonow et al., 2013). A raster layer of 5′ arc min resolution was assembled with CBS-present ($n = 620$) and CBS-absent ($n = 313$) grid cells (Figure 8.1b) (Anonymous, 2014; DEA, Department of Environmental Affairs South Africa, 2015; Paul, 2005; Yonow et al., 2013). Phytosanitary barriers for the internal movement of citrus plants in South Africa (Figure 8.1b) were gathered from official governmental regulations (Anonymous, 1984, 2002), as reviewed by (Martínez-Minaya et al., 2015).

Spatially gridded climatic data ($5'$ arc min resolution) from South Africa were acquired from the WorldClim database, which included mean monthly values for the period 1950-2000 (Hijmans et al., 2005). In addition to the 19 climatic variables available in WorldClim, precipitation from October to January and accumulated degrees (ADD) from July to October (*i.e.* average of T_{max} and T_{min} for each month with $T_{base} = 10$ ^{o}C) were also calculated (Table 8.1). The coordinate system WGS84 was used in all spatially gridded datasets with the raster package for R 3.2.5 (Hijmans, 2014; R Core Team, 2016).

8.2.2 Spatial autocorrelation, collinearity and PCA

To test the hypothesis that the response variable (i.e. CBS presence/absence) occurs at random among citrus grid cells, Moran's I and Geary's C analyses of spatial autocorrelation were used (Plant, 2012). For each dataset, 1950 and 2014, CBS-present citrus grid cells were coded with a 1 and CBS-absent citrus grid cells were coded with a 0. Grid cells without citrus were not considered in the analyses. Both indices were calculated by contigüity and at increasing distances from 20 to 900 km. Moran's I values range from -1, indicating perfect dispersion, to 1 indicating perfect correlation (i.e. clustering). The expected value of Moran's I in the absence of significant spatial autocorrelation is around 0. The value of Geary's C is 1 in the absence of spatial autocorrelation and approaches 0 for strong autocorrelation.

Pearson's correlation coefficient was used to detect collinearity among the 21 climatic explanatory variables included in the analysis. Pairwise correlations were classified as $|r| \leq 0.7$ or $|r| > 0.7$ according to Dormann et al. (2012a), who proposed this threshold of correlation as an appropriate indicator for when collinearity begins to severely distort model estimation and subsequent prediction.

Principal component analysis (PCA) was used to obtain independent linear combinations of 20 climatic variables in order to summarize most of the variation in each dataset. The climatic variable temperature annual range was not taken into account in the PCA, because it is a linear combination of

the variables maximum temperature of the warmest month and minimum temperature of the coldest month (Table 8.1). Principal components (PCs) were extracted sequentially according to the amount of variation explained. The relationship between the individual variables and the extracted PCs was expressed by a Varimax rotated component matrix with Kaiser's normalization. Values approaching unity indicated a greater contribution of the variable to the component (Chatfied and Collins, 2013). In addition, two 95% confidence ellipses were plotted for pairwise PCs, one for CBS-present grid cells and another for CBS-absent grid cells (Johnson and Wichern, 2002).

8.2.3 Models

A Bayesian hierarchical spatial approach was used to model the variation in the proportion of the presence. This approach can be considered as a spatial extension of a generalized linear model in the sense that a stochastic spatial effect is added to the linear predictor. Note also that this approach is highly suitable for situations in which data are observed at continuous locations occurring within a defined spatial domain. Nevertheless, the main interest when dealing with this kind of model is to predict the response in unsampled locations, usually known as kriging in honour of Krige's (Krige, 1951) seminal work. From the Bayesian point of view, this prediction can be performed via predictive distributions that easily allow the incorporation of uncertainty within the model parameters.

FIGURE 8.1: Citrus-growing areas and distribution of citrus black spot (CBS) in South Africa in **a** 1950 and **b** 2014, with lines indicating the prohibition boundary for the east-west movement of citrus plants in 1984 (dashed line) and 2002 (solid line) (Anonymous, 1984, 2002; DEA, Department of Environmental Affairs South Africa, 2015; Martínez-Minaya et al., 2015; Paul, 2005; Powell, 1930; Wager, 1952; Yonow et al., 2013). Data for Lesotho and Swaziland were not available. **c** Solid lines indicate province boundaries.

TABLE 8.1: Climatic variables (BIO) and three linear combinations (PC) extracted with principal component analysis (PCA) in the 1950 and 2014 datasets and their explained variability.

	Climatic variables[1]	**1950**			**2014**		
		$PC1$	$PC2$	$PC3$	$PC1$	$PC2$	$PC3$
BIO_1	Annual mean temperature	0.180	0	0.976	0.133	0.133	0.978
BIO_2	Mean diurnal range (mean of monthly (max temp - min temp)	0.932	-0.203	0	-0.307	0.900	0
BIO_3	Isothermality (BIO2/BIO7)	0	0.433	0	0.661	0.106	0.328
BIO_4	Temperature seasonality (standard deviation * 100)	0.887	-0.313	-0.150	-0.594	0.667	-0.258
BIO_5	Max temperature of warmest month	0.558	-0.361	0.695	-0.526	0.453	0.651
BIO_6	Min temperature of coldest month	-0.666	0	0.726	0.139	-0.638	0.716
BIO_8	Mean temperature of wettest quarter	0.447	0.408	0.647	0.451	0.367	0.741
BIO_9	Mean temperature of driest quarter	-0.323	-0.448	0.518	-0.635	-0.533	0
BIO_{10}	Mean temperature of warmest quarter	0.365	-0.106	0.909	-0.159	0.298	0.915
BIO_{11}	Mean temperature of coldest quarter	-0.236	0.136	0.958	0.273	-0.172	0.943
BIO_{12}	Annual precipitation	-0.287	0.896	-0.137	0.873	-0.400	0
BIO_{13}	Precipitation of wettest month	0.156	0.925	0	0.927	0	0
BIO_{14}	Precipitation of driest month	-0.876	-0.242	-0.230	0	-0.887	-0.184
BIO_{15}	Precipitation seasonality (Coefficient of variation)	0.719	0.522	0.250	0.408	0.664	0.287
BIO_{16}	Precipitation of wettest quarter	0.129	0.936	0	0.930	-0.130	0
BIO_{17}	Precipitation of driest quarter	-0.879	-0.238	-0.259	0	-0.889	-0.216
BIO_{18}	Precipitation of warmest quarter	0.166	0.970	0	0.929	0.102	0.302
BIO_{19}	Precipitation of coldest quarter	-0.606	-0.434	-0.289	-0.384	-0.599	-0.511
AP	Precipitation from October to January	0.125	0.974	0	0.948	0	0.231
ADD	Accumulated degrees from July to October with $T_{base} = 10$ ^{o}C	0.177	0.200	0.928	0.312	0.104	0.912
	% variability:	37.4	25.9	22.1	40.3	29.2	17.1

[1]Temperature variables in ^{o}C and precipitation variables in mm. The variable temperature annual range (BIO_7) was not included in the PCA because it is a linear combination of BIO_5 and BIO_6

In line with Muñoz et al. (2013), let Z_i be the binary response variable representing the presence (1) or absence (0) of CBS at location i. Then, its conditional distribution is $Z_i \mid \pi_i \sim \text{Ber}(\pi_i)$, π_i being the probability of CBS presence at location i. As usual with GLMs, the mean of the response variable was linked to the linear predictor and to the stochastic spatial effect by means of the logit link function defined as $\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$. In particular,

$$\text{logit}(\pi_i) = \boldsymbol{X}_i\boldsymbol{\beta} + W_i\,, \tag{8.1}$$

$\boldsymbol{\beta}$ being the regression coefficientes vector, $\boldsymbol{X}$ the covariates matrix and W_i the spatially structured random effect. The geostatistical term $\boldsymbol{W}$ was assumed to be a multivariate Gaussian distribution whose covariance matrix $\sigma^2_{\boldsymbol{W}} H(\phi)$ depends on the distance between locations, and with hyperparameters $\sigma^2_{\boldsymbol{W}}$ and ϕ representing the variance and range of the geostatistical term, respectively. Once the model had been determined, posterior distributions had to be obtained. As in the Bayesian framework, parameters were treated as random variables and prior knowledge had to be incorporated using the corresponding prior distributions. These priors were specified jointly with random effects, the final level of the Bayesian hierarchical model being the expression of the prior knowledge about the hyperparameters.

When dealing with Bayesian hierarchical models, posterior distributions for the parameters and hyperparameters do not usually have any analytic expression, therefore numerical approximations are needed. In the particular case of latent Gaussian models, INLA is a computationally efficient alternative to MCMC. Latent Gaussian models are a particular case of the Structured Additive Regression (STAR) models, where the mean of the response variable is linked to a structured predictor that accounts for the effects of various covariates in an additive way. The prior knowledge of the additive predictor is expressed using Gaussian prior distributions. In this context, all the latent Gaussian variables can be seen as components of a vector known as the latent Gaussian Field (Rue et al., 2009).

To fit and predict the particular case of continuously indexed Gaussian Fields with INLA, as in our case, $\boldsymbol{W}$, an additional module is required. Lindgren et al. (2011) proposed an explicit link between Gaussian Markov Random Fields (Rue and Held, 2005) and continuous Gaussian Fields with

a Matérn covariance structure via a weak solution to a stochastic partial differential equation (SPDE). Under this approximation, the geostatistical spatial term is reparameterized as follows, $\boldsymbol{W} \sim \mathcal{N}(0, \boldsymbol{Q}(\kappa, \tau))$, depending on two different parameters, κ and τ, determining the range of the effect and the total variance, respectively. More precisely, the range is approximately $\phi = \sqrt{\frac{8}{\kappa}}$ and the variance is $\sigma^2_{\boldsymbol{w}} = \frac{1}{4\pi\kappa^2\tau^2}$ (Lindgren et al., 2011).

As mentioned above, the final step is to specify the prior distributions for the parameters and hyperparameters. Normal vague priors with mean 0 and precision 10^{-4} were used for the regression coefficients vector. Although internally INLA works with κ and τ, priors for the geostatistical term were specified in terms of ϕ and $\sigma_{\boldsymbol{W}}$ using the reparameterizations $\log(\phi)$ and $\log(\sigma_{\boldsymbol{W}}$ as independent Gaussian distributions (Lindgren and Rue, 2015).

To conclude, the full model was stated as follows:

$$\begin{aligned} Z_i \mid \pi_i &\sim \text{Ber}(\pi_i) \\ \text{logit}(\pi_i) &= \boldsymbol{X}_i\boldsymbol{\beta} + W_i\,, \\ \beta_j &\sim \mathcal{N}(0, 10^{-4})\,,\ \boldsymbol{W} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}(\phi, \sigma_{\boldsymbol{W}})\,, \\ \log(\phi) &\sim \mathcal{N}(m_\phi, q_\phi)\,, \\ \log(\sigma_{\boldsymbol{W}}) &\sim \mathcal{N}(m_{\sigma_{\boldsymbol{W}}}, q_{\sigma_{\boldsymbol{W}}})\,, \end{aligned}$$

where m_ϕ was chosen automatically such that the range of the field was about 20% of the diameter of the region, and $m_{\sigma_{\boldsymbol{W}}}$ was chosen so that the corresponding variance of the field was 1 (in particular, $m_\phi = 1.476$ and $m_{\sigma_{\boldsymbol{W}}} = 0$). Finally, the precisions of the prior distributions for $\log(\phi)$ and $\log(\sigma_{\boldsymbol{W}})$ were $q_\phi = 0.1$ and $q_{\sigma_{\boldsymbol{W}}} = 0.1$.

Models including a selection of climatic explanatory variables with $|r| \leq 0.7$ or PCs were fitted to the response variable (CBS presence/absence). Models covering all possible combinations of climatic explanatory variables with $|r| \leq 0.7$ or PCs were compared using the Watanabe Akaike Information Criterion (WAIC), which uses the posterior densities more effectively than the traditional Deviance Information Criterion (Gelman et al., 2014; Watanabe, 2010). The models including climatic explanatory variables or PCs displaying the lowest WAIC were selected. The geostatistical

spatial term was incorporated into these models as described above and the corresponding WAIC was calculated.

A validation dataset with CBS-present ($n = 385$) and CBS-absent ($n = 259$) grid cells (Figure 8.7) was assembled by random sampling without replacement from the 2014 dataset, but excluding those grid cells used for model development in 1950. Receiver operating characteristic (ROC) curve analysis was used to evaluate the predictive ability of the models selected for the 1950 dataset. CBS presence/absence was considered as the binary classification variable. The mean of the predictive posterior distribution of π_i obtained with each model was evaluated as a continuous estimator of this binary classification variable. ROC curves showed the proportion of correctly classified absences (specificity) in the x-axis and the proportion of correctly classified presences (sensitivity) in the y-axis as the continuous variable moved over its range of values (i.e. from 0 to 1). The area under del ROC curve (AUC) was calculated by trapezoids using the pROC package for `R` (Robin et al., 2011).

8.3 Results

8.3.1 Spatial autocorrelation, collinearity and PCA

Moran's I and Geary's C analyses indicated the presence of significant spatial autocorrelation ($P < 0.0001$) in CBS distribution data in 1950 and 2014 (Figure 8.2). Both indices showed that spatial autocorrelation was stronger in 2014 than in 1950. In 1950, Moran's I was highest from contigüity to 50 km, with a maximum of 0.33. Spatial autocorrelation decreased with distance and values of Moran's I close to zero, approaching a random spatial pattern, were obtained from 600 km onwards. In 2014, values of Moran's I equal to one (indicating perfect correlation) were obtained from contigüity and distances between 20 and 180 km, with values higher than 0.79 from 190 to 900 km. In 1950, the lowest value of Geary's C was 0.69 for contigüity and values close to one, indicating an absence of spatial autocorrelation, were obtained with distances greater than 600 km. In 2014, values of Geary's C were lower than 0.21 in all cases.

A high degree of collinearity was detected among the climatic variables, with 189 out of a total of 210 pairwise correlations being significant ($P < 0.05$) in 1950 (Figure 8.8). Pairwise correlations with $|r| > 0.7$ were detected: 17 among the temperature variables, 13 among the precipitation variables and only 3 among the temperature and precipitation variables. In 2014, 193 pairwise correlations were significant ($P < 0.05$). Those with $|r| > 0.7$ were 17 among the temperature variables, 13 among the precipitation variables and only 3 among the temperature and precipitation variables (Figure 8.9).

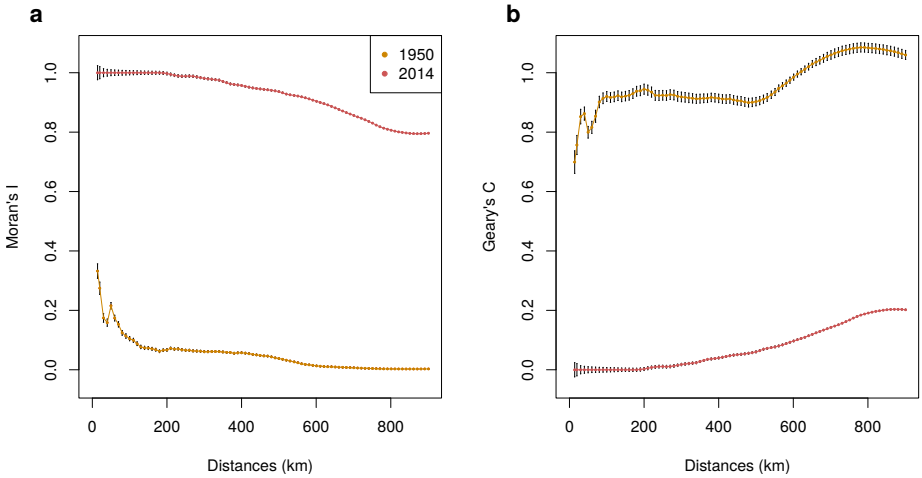

FIGURE 8.2: Moran's I (**a**) and Geary's C (**b**) values for contigüity and at increasing distances, with orange lines for 1950 and red lines for 2014.

Three PCs were extracted from the 1950 dataset, explaining 85.4% of the variability, with $PC1 = 37.4\%$, $PC2 = 25.9\%$ and $PC3 = 22.1\%$. The temperature variables with the most influence in $PC1$ were the mean diurnal range and temperature seasonality with positive coefficients of 0.932 and 0.887, respectively (Table 8.1). The precipitation of the driest month and the driest quarter made a negative contribution to $PC1$, with coefficients of -0.876 and -0.879, respectively. Precipitation seasonality made a positive contribution to $PC1$, with a coefficient of 0.719. When plotted onto the map of South Africa, the lowest values of PC1 coincided mainly with the Indian Ocean coastal areas (Figure 8.3a). Temperature variables did not contribute much to $PC2$ (Table 8.1). Annual precipitation, precipitation of

the wettest month and quarter made a strong positive contribution to $PC2$, with coefficients greater than 0.89. Precipitation in the warmest quarter and from October to January also made a strong positive contribution to $PC2$, with coefficients greater than 0.97. The highest values of $PC2$ were obtained in the eastern half of South Africa (Figure 8.3b). Precipitation variables were not very influential in $PC3$ (Table 8.1). Annual mean temperature, of the warmest and the coldest quarters, as well as ADD from July to October made a strong positive contribution to $PC3$, with coefficients greater than 0.90. Lower values of $PC3$ were obtained with increasing altitudes (Figures 8.3c and 8.10a). CBS presences and absences were not clearly separated when plotting the values of the PCs for each citrus grid cell (Figure 8.11).

Three PCs were extracted from the 2014 dataset, explaining 86.6% of the variability, with $PC1 = 40.3\%$, $PC2 = 29.2\%$ and $PC3 = 17.1\%$. Precipitation variables made the greatest contributions to $PC1$ (Table 8.1). The coefficient for annual precipitation was 0.87 and those for precipitation of the wettest month, wettest quarter, warmest quarter, and from October to January were greater than 0.92. Like $PC2$ in 1950, the highest values of $PC1$ in 2014 were obtained in the eastern half of South Africa (Figure 8.3d). In $PC2$, mean diurnal range made a strong positive contribution, whereas precipitation in the driest month and quarter had a strong negative influence (Table 8.1). The geographic representation of $PC2$ in 2014 was similar to that of $PC1$ in 1950, with the lowest values along the Indian Ocean coast (Figure 8.3e). Precipitation variables had little influence on $PC3$ (Table 8.1). Annual mean temperature, ADD from July to October, as well as mean temperature of the warmest and coldest quarters made a strong positive contribution to $PC3$, with coefficients greater than 0.91. Similarly to 1950, $PC3$ had lower values at higher altitudes (Figures 8.3f and 8.10a). CBS presences and absences were clearly discriminated when plotting the values of $PC1$ and $PC3$ for each citrus grid cell (Figure 8.11), with a small area of overlap corresponding to some citrus areas in the Eastern Cape, Western Cape, the North West and Northern Cape (Figure 8.12).

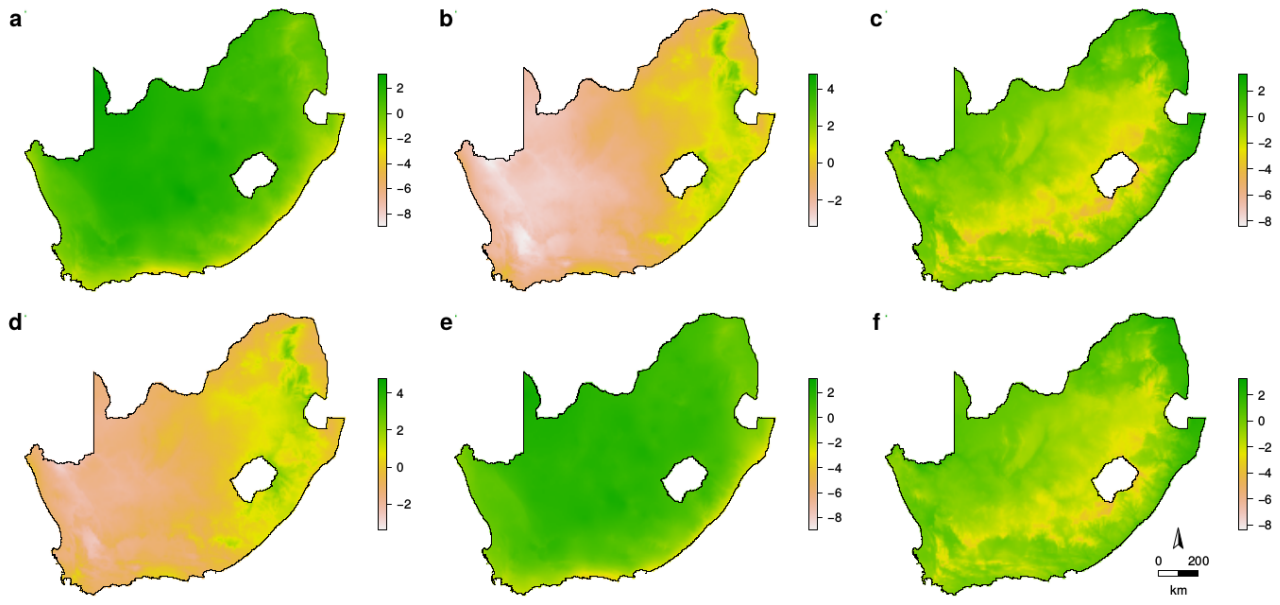

FIGURE 8.3: Geographic representation of the rotated principal components $PC1$ (**a**), $PC2$ (**b**), $PC3$ (**c**) for 1950 and $PC1$ (**d**), $PC2$ (**e**) and $PC3$ (**f**) for 2014.

8.3.2 Model fit and evaluation

In 1950, six climatic variables were selected with $|r| \leq 0.7$: maximum temperature of the warmest month, minimum temperature of the coldest month, mean temperature of the driest quarter, ADD from July to October, annual precipitation, and precipitation of the coldest quarter (Figure 8.10). The model that included the maximum temperature of the warmest month and annual precipitation showed the lowest WAIC with a value of 177.51 (Table 8.3). When a geostatistical term was included in this model, the WAIC was reduced to 126.14 (Table 8.2). Both climatic variables had positive estimates of their parameters. When PCs for 1950 were considered, the model retaining all three PCs had the lowest WAIC of 198.19 (Table 8.3). When a geostatistical component was included in this model, the WAIC was reduced to 131.26 (Table 8.2) and all three PCs had positive estimates of their parameters.

In 2014, the same six climatic variables with $|r| \leq 0.7$ were selected (Figure 8.10). The model that included the maximum temperature of the warmest month, precipitation of the coldest quarter, and ADD from July to October had the lowest WAIC with a value of 49.57 (Table 8.3). In this model, the maximum temperature of the warmest month and precipitation of the coldest quarter had negative estimates of their parameters, whereas that of ADD from July to October was positive (Table 8.2). When PCs were included, the lowest WAIC of 100.70 was obtained with the model retaining only $PC1$ and $PC3$, but similar to the WAIC of 101.98 with the model including all three PCs (Table 8.3). Both $PC1$ and $PC3$ had positive estimates of their parameters (Table 8.2). In 2014 it was not possible to include the geostatistical term in the models due to the fact that CBS presences and absences were completely separated on the map.

Similar predictive distributions were obtained with the models for 1950 including the maximum temperature of the warmest month and annual precipitation or three PCs (Figures 8.4ac). The highest probabilities were obtained along the coast of Kwazulu-Natal and part of the Eastern Cape, as well as in inland areas of Mpumalanga and Limpopo, with values of up to 0.93. The standard deviation associated with the predictive distributions of

TABLE 8.2: Best models for 1950 and 2014 with climatic variables (BIO), principal components (PC) and geostatistical term ($\boldsymbol{W}$).

Models[a]			WAIC[b]
	Mean	Sd	
1950			
$1 + BIO_5 + BIO_{12}$			177.51
Intercept	-26.593	5.401	
BIO_5	0.478	0.146	
BIO_{12}	0.012	0.002	
$1 + PC_1 + PC_2 + PC_3$			198.19
Intercept	-4.481	0.391	
PC_1	-0.64	0.24	
PC_2	1.325	0.231	
PC_3	0.515	0.209	
$1 + BIO_5 + BIO_{12} + \boldsymbol{W}$			126.14
Intercept	-95.237	35.005	
BIO_5	2.151	0.859	
BIO_{12}	0.031	0.009	
$1 + PC_1 + PC_2 + PC_3 + \boldsymbol{W}$			131.26
Intercept	-5.933	4.137	
$PC1$	2.037	2.732	
$PC2$	5.539	2.588	
$PC3$	4.62	2.257	
2014			
$1 + BIO_5 + BIO_{19} + ADD$			49.57
Intercept	48.91	11.716	
BIO_5	-2.886	0.616	
BIO_{19}	-0.126	0.028	
ADD	0.77	0.159	
$1 + PC_1 + PC_3$			100.70
Intercept	7.934	1.211	
$PC1$	9.145	1.259	
$PC2$	8.242	1.156	

[a]Maximum temperature of warmest month (BIO_5), minimum temperature of coldest month (BIO_6), mean temperature of driest quarter (BIO_9), accumulated degrees (ADD) from July to October with T_{base} = 10 °C, annual precipitation (BIO_{12}) and precipitation of coldest quarter (BIO_{19}).
[b]Watanabe Akaike Information Criterion. Lower values of WAIC reflect a better model fit balanced with model complexity.

these models was lower than 0.168, with the highest uncertainty in the areas of higher probability (Figures 8.4bd). The predictive distribution of the model with the two climatic variables and a geostatistical term was similar to those of the previous two models, but with a much higher probability in Kwazulu-Natal (Figure 8.4e). Larger standard deviation was associated with this model, with values of up to 0.41 around the areas of high probability and in the central regions of the country (Figure 8.4f). The model including three PCs and a geostatistical term predicted larger areas with a high probability of 0.99, entirely covering Kwazulu-Natal and regions in Mpumalanga, Limpopo and North West provinces (Figure 8.4g). Areas of high uncertainty were also much larger with this model, with values of standard deviation up to 0.44, particularly in the eastern half of the country (Figure 8.4h).

In 2014, similar predictive distributions were obtained with the models including $PC1$ and $PC3$ or the maximum temperature of the warmest month, precipitation of the coldest quarter, and ADD from July to October (Figures 8.5ac). High probabilities up to unity were obtained in Kwazulu-Natal, Mpumalanga, Limpopo, Gauteng, parts of the Eastern Cape, North West and Free State, as well as in coastal areas in the Western Cape and Northern Cape. The standard deviation associated with these predictive distributions was lower than 0.34, with the highest uncertainty around the areas of higher probability (Figures 8.5bd).

When the predictive distributions of the models for 1950 were evaluated against the distribution of CBS in 2014, excluding those grid cells used for model development, the highest AUC of 0.986 was obtained with the model including three PCs and a geostatistical term (Figure 8.6). The model with only three PCs had an AUC of 0.929. The model including the maximum temperature of the warmest month and annual precipitation had an AUC of 0.839, which was reduced to 0.821 when a geostatistical term was incorporated.

8.4 Discussion

Correlative species distribution models rely on the assumption that the organism modelled is in equilibrium with its environment within the region of study. Hence, the species occurs in all suitable environmental conditions (*i.e.* throughout the suitable environmental space), although not necessarily occupying the geographic space completely. This assumption is often violated in the case of biological invasions, where potentially suitable habitats were not yet reached by the species because of colonization time lag and/or dispersal constraints (Barve et al., 2011; Elith and Leathwick, 2009). It has been stated that CBS probably attained its full potential distribution in South Africa because the disease had many opportunities to invade citrus areas throughout the country (Yonow et al., 2013). CBS is certainly much more widespread nowadays in South Africa than in 1950 (Figure 8.1), but the assumption that CBS is in equilibrium with the environmental conditions and occurs in all suitable habitats in the country is questionable (EFSA, European Food Safety Authority, 2008; EFSA, European Food Safety Authority, 2014, 2016). In fact the movement of citrus material in South Africa was not regulated until 1984, although quantitative trade data among provinces was not found (Martínez-Minaya et al., 2015). Since then, internal phytosanitary barriers have been in place to impede the movement of citrus material and avoid the spread of *P. citricarpa* to other regions in the country (Figure 8.1b). The presence of dispersal constraints for more than three decades cannot be overlooked when interpreting CBS distribution in South Africa and the resulting model outcomes.

Process-based models comprising the entire environmental space of the species are thought to be more adequate for non-equilibrium scenarios (Dormann et al., 2012b), but they still rely on disease prevalence data to interpret model outputs and define thresholds for climate suitability. For instance, (Magarey et al., 2015) defined a specific output threshold to be able to consider a location suitable for CBS based on the values for sites with moderate disease prevalence, which was the Eastern Cape in their study. However, the crucial role of an accessible area fully applies to process-based models as well. Different thresholds for moderate disease prevalence might be chosen considering past, present or future disease distribution data. Moreover,

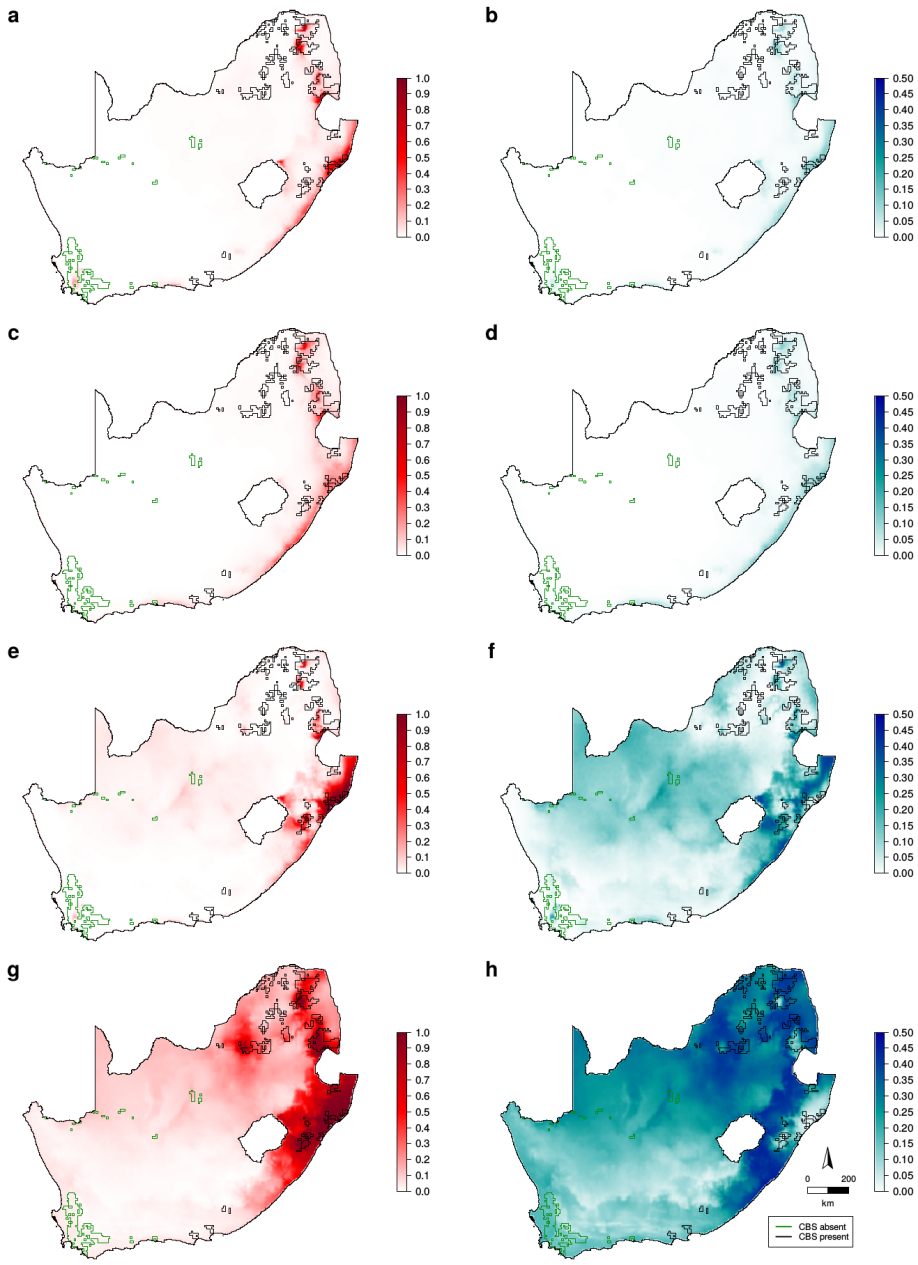

FIGURE 8.4: Mean (red) and standard deviation (blue) of the predictive posterior distribution for the probability of citrus black spot (CBS) presence with the best models of 1950 including climatic variables (**a,b**), principal components (**c,d**), climatic variables + geostatistical term (**e, f**) and principal components + geostatistical term (**g, h**).

CBS is characterized by slow epidemic development and thus future impacts cannot be directly inferred from its present status (Kotzé, 1981). Besides, mechanistic models for CBS are seriously affected by large uncertainties due to the lack of biological and epidemiological data (EFSA, European Food Safety Authority, 2014, 2016).

The consideration of true absences is also a controversial issue in species distribution models. In many cases, only presence data are available and models such as Maxent are preferred, which generate random pseudo-absences from an area around presence records. When the species being modelled is in its early stages of invasion, presence-only models are sometimes preferred because absences may not be associated with climatic unsuitability (Dupin et al., 2011). However, with pseudo-absences the accuracy of the model can be overestimated and reliable absence data are considered more appropriate for model validation (Václavík and Meentemeyer, 2012). In the case of the logistic regression used here, true absences are required for both model development and evaluation.

Disease presences in the 1950 dataset were obtained from Wager (1952), who surveyed the citrus-growing areas in South Africa for CBS. Nevertheless, molecular techniques for pathogen detection were not available at that time and, therefore, the possibility of missing CBS presences in a latent asymptomatic stage cannot be excluded. The map of citrus distribution in South Africa in 1927 (Powell, 1930) had a reasonable level of detail. However, it looks as if some citrus areas in the Eastern Cape might be overrepresented (Figure 8.1a), potentially increasing the number of CBS absences. More recent surveys for CBS in South Africa should comply with international standards (IPPC, International Plant Protection Convention, 1995, 2005, 2007), which ideally reduce the risk of imperfect detections and sampling bias (Guillera-Arroita et al., 2015). Although we restricted our data to only citrus areas (Anonymous, 2014; Paul, 2005; Yonow et al., 2013), the NLS class "cultivated commercial permanent orchards" also comprises other crops (DEA, Department of Environmental Affairs South Africa, 2015), which potentially increases CBS presences and/or absences in the 2014 dataset. On the other hand, the NLS dataset does not consider ornamental or back-yard citrus trees, thereby potentially reducing CBS presences and/or absences. In any case, since no other contrasting data sources

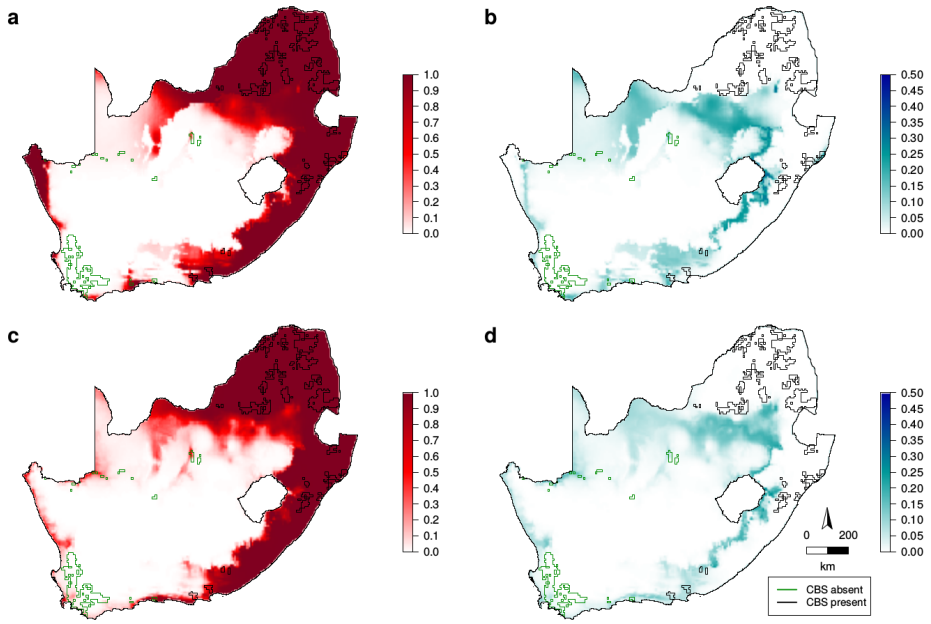

FIGURE 8.5: Mean (red) and standard deviation (blue) of the predictive posterior distribution for the probability of citrus black spot (CBS) presence with the best models of 2014 including climatic variables (**a,b**) or principal components (**c,d**).

were found, we consider that our analyses were based on the best information available. Further refinements of our models could be possible if more accurate datasets of CBS distribution in South Africa become accessible. Likewise, recent updates of the WorldClim database could also be used (Fick and Hijmans, 2017).

Significant spatial autocorrelation of CBS distribution was detected in 1950 and 2014 (Figure 8.2). Furthermore, the geostatistical term was relevant in the regression models for 1950, climatic variables or PCs also being included as explanatory variables (Table 8.2). Spatial autocorrelation occurs when disease observations in different locations are not independent from each other. Dispersal barriers, spatially structured gradients or intrinsic spread processes usually lead to spatial autocorrelation in species distribution data (Franklin, 2009). The strong spatial autocorrelation detected in CBS distribution data both in 1950 and 2014 suggests that climate itself

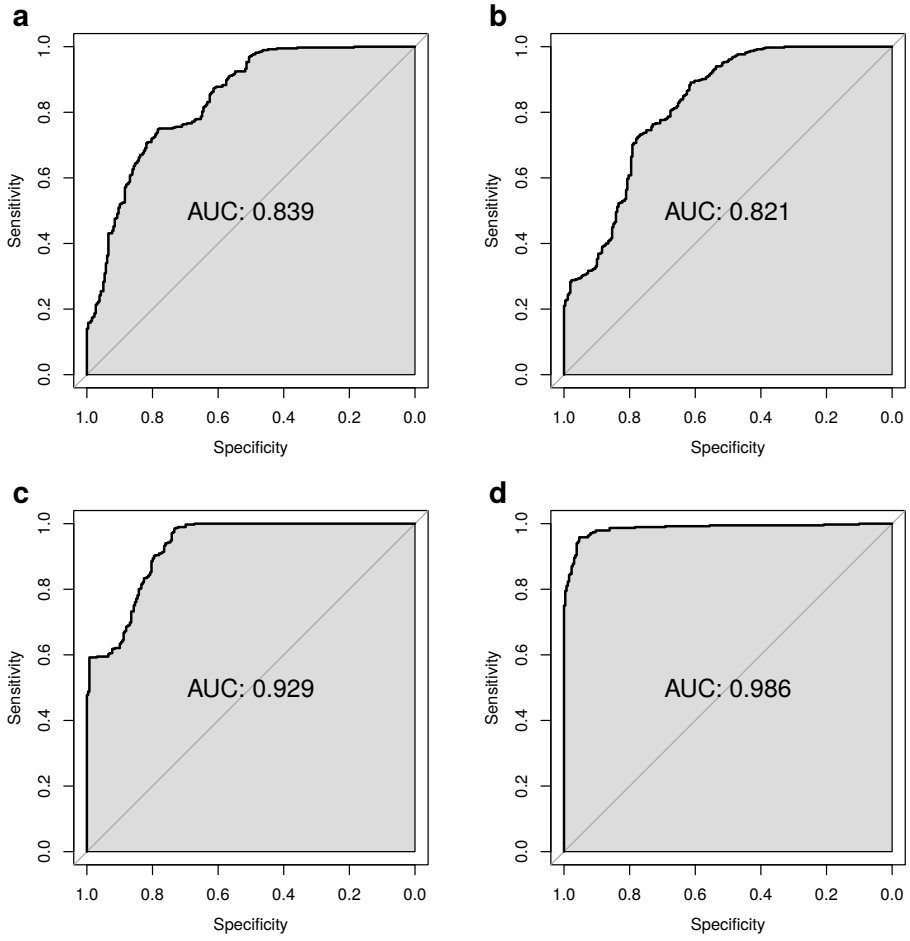

FIGURE 8.6: Receiver operating characteristic (ROC) curves and area under the curve (AUC) obtained with the 2014 validation dataset with the best models for the probability of citrus black spot presence in 1950 including climatic variables (**a**), principal components (**b**), climatic variables + geostatistical term (**c**) and principal components + geostatistical term (**d**).

might not be the main factor limiting the spread of CBS in South Africa. The natural spread of CBS through *P. citricarpa* spores is poorly understood. Under laboratory conditions, conidia from inoculated citrus fruit

were splashed 0.6 m high and to a distance of at least 8 m by simulated wind-driven rain (Perryman et al., 2014). No information is available on the distances airborne ascospores of *P. citricarpa* can spread. The drivers of CBS invasions worldwide remain generally unknown, but human-assisted movement of infected citrus material is considered the most important means of disease introduction and spread. Regardless of the mechanisms involved in the invasion process, the presence of significant spatial aggregation indicated a higher probability of CBS presence in grid cells near affected areas.

Ignoring spatial autocorrelation in the models can decrease the precision of parameter estimates and falsely reject the null hypothesis of no effect. In addition, the selection of explanatory variables may be biased towards those that are more autocorrelated, such as climatic gradients. Consequently, certain variables as well as more variables in general are likely to be retained, thereby making the resulting model potentially misleading (Chapman, 2010; F. Dormann et al., 2007; Franklin, 2009). Previous studies with CLIMEX to estimate the potential geographic distribution of CBS did not consider spatial autocorrelation (Er et al., 2013; Paul et al., 2005; Yonow et al., 2013). These studies were conducted on a much broader geographic scale and so consequences of ignoring spatial dependence are believed to be less problematic (Franklin, 2009). However, the presence of dispersal constraints like phytosanitary barriers (Figure 8.1b) and other local range-confining processes may have limited the area and environments accessible to CBS anyway. Moreover, none of these CLIMEX studies involved formal statistical inference for parameter estimation and so they are difficult to compare with our models.

Collinearity arises when two or more explanatory variables in a model are linearly related, which is common when climatic variables are considered. With collinearity, parameter estimates may be unstable with inflated standard errors and thus inference may be biased and select the wrong explanatory variables. Moreover, the effects of two collinear explanatory variables cannot be separated and model extrapolation may be seriously flawed (Dormann et al., 2012a). PCA is one of the most common ways to manage collinearity among explanatory variables in correlative species distribution models (Dupin et al., 2011; Manel et al., 2001; Kriticos et al., 2014; Petitpierre et al., 2012). In our case, the use of PCs as explanatory variables

in the models allowed us to integrate the contribution of a relatively large set of climatic variables with serious collinearity problems (Figures 8.8 and 8.9). Nevertheless, better model fit (i.e. lower WAIC) was obtained including a threshold-based pre-selection of climatic explanatory variables with pairwise correlations $|r| \leq 0.7$ (Dormann et al., 2012a).

In general, the regression analyses performed in our study indicated a positive relationship between CBS presence and climatic variables or PCs associated with warm temperatures and high precipitation (Tables 8.1 and 8.2). Indeed, it has been stated that CBS thrives mainly in warm wet climates (Yonow et al., 2013), although the disease is also present in arid desert conditions (Martínez-Minaya et al., 2015). Some of the previous studies with CLIMEX suggested that the potential distribution of CBS could be limited by cold conditions (Paul et al., 2005; Yonow et al., 2013), although these modelling approaches and their parameterization have been subject to debate (EFSA, European Food Safety Authority, 2008; EFSA, European Food Safety Authority, 2014). In our models, degrees accumulated during the period of ascocarp formation and ascospore maturation in South Africa (i.e. July to October) were positively related with CBS presence (Table 8.2). The empirical *ADD* model by (Fourie et al., 2013) predicted an earlier release of *Phyllosticta* spp. ascospores with warmer winters and springs, which might be associated with more favourable climate conditions for CBS establishment. However, this empirical model included both *P. citricarpa* and the non-pathogenic species *P. capitalensis* Henn., which is also widely established in relatively cold regions (Wikee et al., 2013).

Several studies have demonstrated that models for species in the early stages of invasion are more likely to underpredict potential distribution than models in advanced stages of invasion, where the equilibrium assumption is more plausible (Dupin et al., 2011; Václavík and Meentemeyer, 2012). In our case, relatively high accuracy was obtained with the models for the 1950 dataset, representing the early stages of CBS epidemics in South Africa. An AUC of 0.929 was obtained with the model including PCs and an AUC of 0.986 resulted when a geostatistical term was also incorporated (Figure 8.6). According to the criteria put forward by Swets (1988), these AUC values are indicative of rather high accuracy. However, despite their good accuracy, none of our models were able to predict subsequent CBS invasions in citrus

areas in the Eastern Cape and north of Limpopo (Figure 8.4), where the disease thrives under more arid conditions (Martínez-Minaya et al., 2015).

Citrus areas in the north of Limpopo are considered of low pest (disease) prevalence (Anonymous, 2014), which implies that CBS occurs at low levels and is subjected to effective surveillance, control or eradication measures (IPPC, International Plant Protection Convention, 2005, 2007). It has been claimed that CBS has low or moderate prevalence in the Eastern Cape (Fourie et al., 2013; Magarey et al., 2015), but this region is not officially considered an area of low CBS prevalence (Anonymous, 2014). Moreover, Schutte (1995) indicated that in the Eastern Cape lemons were sprayed with fungicides for CBS control. This has been confirmed by more recent reports, indicating that fungicide applications for CBS control have increased in the Eastern Cape and lemons must be frequently sprayed (Grout, 2015).

Citrus-growing areas in the Eastern Cape are the only ones affected by CBS nowadays that were left outside the phytosanitary barrier established in 1984 (Figure 8.1b). Considering the long lag phase of CBS epidemics (Kotzé, 1981) and that fungicides were applied for its control in the Eastern Cape in the 1990s (Schutte, 1995), it is conceivable that this region was already affected several years before, but perhaps at very low levels or still in an asymptomatic stage. Indeed, Kotzé (1981) indicated that *P. citricarpa* may be present for many years in a particular area before symptoms appear. Consequently, citrus areas in the Eastern Cape might have been inadvertently considered as CBS-free when designing the phytosanitary barrier in 1984. Interestingly, when representing $PC1$ and $PC3$ associated with precipitation and warm temperatures in 2014, citrus areas in the Eastern Cape currently affected by CBS overlapped with some CBS-free areas in the Western Cape, Northern Cape and North West provinces (Figure 8.12). Although the models for 2014 did not incorporate a geostatistical term, those including climatic variables or PCs displayed relatively high probabilities of CBS occurrence in these particular areas (Figure 8.5). Therefore, intensive surveys would be recommended to keep them free from disease.

Although climate has been advocated as the main factor limiting the establishment and spread of CBS into new areas (Magarey et al., 2015; Paul et al., 2005; Yonow et al., 2013), our study indicates that spatial proximity

to affected areas is also relevant in the geographic distribution of the disease in South Africa. Indeed, some historical evidence illustrated that too much hope had been pinned on climate as a limiting factor for CBS (Kotzé, 1981). In his detailed study, Wager (1952) indicated that CBS was first reported in South Africa by Doidge (1929) in a relatively cool mist-belt area with high rainfall. As at that time it was assumed that CBS required this type of conditions, no concern was therefore felt for its possible spread to other parts of South Africa (Wager, 1952). However, from 1940 to 1950 the disease spread to neighbouring citrus regions under much drier conditions. Based on this, Wager (1952) concluded that the old concept of CBS requiring cool, moist, or mist-belt conditions for its development was wrong. McOnie (1964a) surveyed the citrus areas in the Eastern Cape and concluded that *P. citricarpa* was absent due to unfavourable climatic conditions. However, the pathogen was later reported in the Eastern Cape and fungicide sprays are currently applied for CBS control (Grout, 2015; Schutte, 1995). In Zimbabwe, Whiteside (1965) stated that CBS may not become really serious under local climatic conditions. Nevertheless, the disease reached epidemic proportions in 1978 (Kotzé, 1981). More recently, Guarnaccia et al. (2017) reported, for the first time, the presence of *P. citricarpa* in the Mediterranean Basin, under dry-summer climate conditions. The future will determine whether current models for the potential geographical distribution of CBS can stand the test of time.

8.5 Acknowledgements

Authors would like to thank Iosu Paradinas from UV and Xavier Barber from MHU for support with the use of R and R-INLA, and the INLA-project team, in particular, Elias T. Krainski from UF Paraná, for their prompt support with technical aspects of the usage of R-INLA, and to V. Monzó (MON Topografía y Cartografía) for digitizing and georeferencing maps.

References

Anonymous (1984). R.110 Agricultural pest act, 1983 (Act 36 of 1983). Control measures. *Government Gazette*, 9047:6–11.

Anonymous (2002). R.831 Agricultural pest act, 1983 (Act 36 of 1983). Control measures: Amendment. *Government Gazette*, 23517:15–17.

Anonymous (2014). R.442 agricultural pest act, 1983 (act 36 of 1983). control measures: Amendment. *Government Gazette*, 37702:4–11.

Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S. P., Peterson, A. T., Soberón, J., and Villalobos, F. (2011). The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling*, 222(11):1810–1819.

Chapman, D. S. (2010). Weak climatic associations among British plant distributions. *Global Ecology and Biogeography*, 19(6):831–841.

Chatfied, C. and Collins, A. J. (2013). *Introduction to multivariate analysis*. Berlin: Springer.

DEA, Department of Environmental Affairs South Africa (2015). South African national land-cover dataset 2013-2014. Geoterraimage. `http://egis.environment.gov.za`. Accessed on 8 February 2016.

Doidge, E. M. (1929). Some diseases of citrus prevalent in South Africa. *South African Journal of Science*, 26(12):320–325.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., et al. (2012a). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1):27–46.

Dormann, C. F., Schymanski, S. J., Cabral, J., Chuine, I., Graham, C., Hartig, F., Kearney, M., Morin, X., Römermann, C., Schröder, B., et al. (2012b). Correlation and process in species distribution models: bridging a dichotomy. *Journal of Biogeography*, 39(12):2119–2131.

Dupin, M., Reynaud, P., Jarošík, V., Baker, R., Brunel, S., Eyre, D., Pergl, J., and Makowski, D. (2011). Effects of the training dataset characteristics on the performance of nine species distribution models: application to *Diabrotica virgifera virgifera*. *PLoS One*, 6(6):e20957.

EFSA, European Food Safety Authority (2008). Pest risk assessment and additional evidence provided by South Africa on *Guignardia citricarpa* Kiely, citrus black spot fungus–CBS. Scientific Opinion of the PLH Panel. *EFSA Journal*, 925:1–108.

EFSA, European Food Safety Authority (2014). Scientific opinion on the risk of *Phyllosticta citricarpa* (*Guignardia citricarpa*) for the EU territory with identification and evaluation of risk reduction options. *EFSA Journal*, 12:3557.

EFSA, European Food Safety Authority (2016). Evaluation of new scientific information on *Phyllosticta citricarpa* in relation to the EFSA PLH Panel (2014). Scientific Opinion on the plant health risk to the EU. *EFSA Journal*, 14:4513.

Elith, J. and 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.

Elith, J., Simpson, J., Hirsch, M., and Burgman, M. (2013). Taxonomic uncertainty and decision making for biosecurity: spatial models for myrtle/guava rust. *Australasian Plant Pathology*, 42(1):43–51.

Er, H., Roberts, P., Marois, J., and van Bruggen, A. (2013). Potential distribution of citrus black spot in the United States based on climatic conditions. *European Journal of Plant Pathology*, 137(3):635–647.

F. Dormann, C., M. McPherson, J., B. Araújo, M., Bivand, R., Bolliger, J., Carl, G., G. Davies, R., Hirzel, A., Jetz, W., Daniel Kissling, W., et al. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, 30(5):609–628.

Fick, S. E. and Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37(12):4302–4315.

Fourie, P., Schutte, T., Serfontein, S., and Swart, F. (2013). Modeling the effect of temperature and wetness on *Guignardia* pseudothecium maturation and ascospore release in citrus orchards. *Phytopathology*, 103(3):281–292.

Franklin, J. (2009). *Mapping species distributions: spatial inference and prediction*. New York: Cambridge University Press.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016.

Grout, T. G. (2015). The status of citrus IPM in South Africa. *Acta Horticulturae*, 1065:1091–1095.

Guarnaccia, V., Groenewald, J., Li, H., Glienke, C., Carstens, E., Hattingh, V., Fourie, P., and Crous, P. W. (2017). First report of *Phyllosticta citricarpa* and description of two new species, *P. paracapitalensis* and *P. paracitricarpa*, from citrus in Europe. *Studies in Mycology*, 87:161–185.

Guillera-Arroita, G., Lahoz-Monfort, J. J., Elith, J., Gordon, A., Kujala, H., Lentini, P. E., McCarthy, M. A., Tingley, R., and Wintle, B. A. (2015). Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography*, 24(3):276–292.

Hendricks, K. E., Christman, M., and Roberts, P. D. (2017). Spatial and temporal patterns of commercial citrus trees affected by *Phyllosticta citricarpa* in Florida. *Scientific Reports*, 7(1):1641.

Hijmans, R. J. (2014). raster: geographic data analysis and modeling. R package version 2.2-31. `http://CRAN.R-project.org/package=raster`.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25:1965–1978.

IPPC, International Plant Protection Convention (1995). Requirements for the establishment of pest free areas. International Standards for Phytosanitary Measures, ISPM 4. Rome: IPPC.

IPPC, International Plant Protection Convention (2005). Requirements for the establishment of areas of low pest prevalence. International Standards for Phytosanitary Measures, ISPM 22. Rome: IPPC.

IPPC, International Plant Protection Convention (2007). Recognition of pest free areas and areas of low pest prevalence. International Standards for Phytosanitary Measures, ISPM 29. Rome: IPPC.

Johnson, R. A. and Wichern, D. W. (2002). *Applied multivariate statistical analysis, 5ed.* ew Jersey: Prentice Hall.

Kotzé, J. (1981). Epidemiology and control of citrus black spot in South Africa. *Plant Disease*, 65:945–950.

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6):119–139.

Kriticos, D. J., Jarošik, V., and Ota, N. (2014). Extending the suite of bioclim variables: a proposed registry system and case study using principal components analysis. *Methods in Ecology and Evolution*, 5(9):956–960.

Latimer, A. M., Wu, S., Gelfand, A. E., and Silander Jr, J. A. (2006). Building statistical models to analyze species distributions. *Ecological Applications*, 16(1):33–50.

Lindgren, F. and Rue, H. (2015). Bayesian spatial modelling with `R-INLA`. *Journal of Statistical Software, Articles*, 63(19):1–25.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.

Magarey, R., Chanelli, S., and Holtz, T. (2011). Validation study and risk assessment: *Guignardia citricarpa*, (citrus black spot). Technical report, USDA-APHIS-PPQ-CPHST-PERAL/NCSU.

Magarey, R., Sutton, T., and Thayer, C. (2005). A simple generic infection model for foliar fungal plant pathogens. *Phytopathology*, 95(1):92–100.

Magarey, R. D., Hong, S. C., Fourie, P. H., Christie, D. N., Miles, A. K., Schutte, G. C., and Gottwald, T. R. (2015). Prediction of *Phyllosticta citricarpa* using an hourly infection model and validation with prevalence data from South Africa and Australia. *Crop Protection*, 75:104–114.

Makowski, D., Vicent, A., Pautasso, M., Stancanelli, G., and Rafoss, T. (2014). Comparison of statistical models in a meta-analysis of fungicide treatments for the control of citrus black spot caused by *Phyllosticta citricarpa*. *European Journal of Plant Pathology*, 139:79–94.

Manel, S., Williams, H. C., and Ormerod, S. J. (2001). Evaluating presence–absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, 38(5):921–931.

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

McOnie, K. (1964a). Apparent absence of *Guignardia citricarpa* Kiely from localities where citrus black spot is absent. *South African Journal of Agricultural Science*, 7:347–354.

McOnie, K. (1964b). Orchard development and discharge of ascospores of *Guignardia citricarpa* and onset of infection in relation to control of citrus black spot. *Phytopathology*, 54(12):1448–1454.

McOnie, K. C. (1964c). The latent occurrence in citrus and other hosts of *Guignardia* easily confused with *G. citricarpa*, the citrus black spot pathogen. *Phytopathology*, 54:40–43.

Meentemeyer, R. K., Anacker, B. L., Mark, W., and Rizzo, D. M. (2008). Early detection of emerging forest disease using dispersal estimation and ecological niche modeling. *Ecological Applications*, 18(2):377–390.

Muñoz, F., Pennino, M. G., Conesa, D., López-Quílez, A., and Bellido, J. M. (2013). Estimation and prediction of the spatial occurrence of fish species using Bayesian latent Gaussian models. *Stochastic Environmental Research and Risk Assessment*, 27(5):1171–1180.

Narouei-Khandan, H., Harmon, C., Harmon, P., Olmstead, J., Zelenev, V., van der Werf, W., Worner, S. P., Senay, S., and van Bruggen, A. (2017). Potential global and regional geographic distribution of *Phomopsis vaccinii* on *Vaccinium* species projected by two species distribution models. *European Journal of Plant Pathology*, 148(4):919–930.

Paul, I. (2005). *Modelling the distribution of citrus black spot caused by Guignardia citricarpa Kiely*. PhD thesis, Pretoria: University of Pretoria.

Paul, I., Van Jaarsveld, A., Korsten, L., and Hattingh, V. (2005). The potential global geographical distribution of Citrus Black Spot caused by *Guignardia citricarpa* (Kiely): likelihood of disease establishment in the European Union. *Crop Protection*, 24:297–308.

Perryman, S., Clark, S., and West, J. (2014). Splash dispersal of *Phyllosticta citricarpa* conidia from infected citrus fruit. *Scientific Reports*, 4:6568.

Petitpierre, B., Kueffer, C., Broennimann, O., Randin, C., Daehler, C., and Guisan, A. (2012). Climatic niche shifts are rare among terrestrial plant invaders. *Science*, 335(6074):1344–1348.

Plant, R. E. (2012). *Spatial data analysis in ecology and agriculture using R*. Boca Raton, FL:CRC Press.

Powell, H. C. (1930). The culture of the orange and allied fruits. South African agricultural series No. 8. Johannesburg: Central News Agency.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J., and Muller, M. (2011). pROC: an open-source package for R and S plus to analyze and compare ROC curves. *BMC Bioinformatics*, 12:77.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall.

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

Schutte, G. C. (1995). *Evaluation of control strategies for citrus black spot in Southern Africa*. PhD thesis, Pretoria: University of Pretoria.

Spósito, M., Amorim, L., Bassanezi, R., Filho, A. B., and Hau, B. (2008). Spatial pattern of black spot incidence within citrus trees related to disease severity and pathogen dispersal. *Plant Pathology*, 57(1):103–108.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, 240(4857):1285–1293.

Tran, N. T., Miles, A. K., Dietzgen, R. G., Dewdney, M. M., Zhang, K., Rollins, J. A., and Drenth, A. (2017). Sexual reproduction in the citrus black spot pathogen, *Phyllosticta citricarpa*. *Phytopathology*, 107(6):732–739.

Václavík, T. and Meentemeyer, R. K. (2012). Equilibrium or not? modelling potential distribution of invasive species in different stages of invasion. *Diversity and Distributions*, 18(1):73–83.

Venette, R. C., Kriticos, D. J., Magarey, R. D., Koch, F. H., Baker, R. H., Worner, S. P., Gómez Raboteaux, N. N., McKenney, D. W., Dobesberger, E. J., Yemshanov, D., De Barro, P. J., Hutchison, W. D., Fowler, G., Kalaris, T. M., and Pedlar, J. (2010). Pest risk maps for invasive alien species: a roadmap for improvement. *BioScience*, 60(5):349–362.

Wager, V. A. (1952). The black spot disease of citrus in South Africa. *Science Bulletin of the Department of Agriculture of the Union of South Africa*, 303:1–52.

Wang, N.-Y., Zhang, K., Huguet-Tapia, J. C., Rollins, J. A., and Dewdney, M. M. (2016). Mating type and simple sequence repeat markers indicate a clonal population of *Phyllosticta citricarpa* in Florida. *Phytopathology*, 106(11):1300–1310.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec):3571–3594.

Whiteside, J. (1965). Black spot disease in rhodesia. *Rhodesian Agricultural Journal*, 62:87–91.

Wikee, S., Lombard, L., Crous, P. W., Nakashima, C., Motohashi, K., Chukeatirote, E., Alias, S. A., McKenzie, E. H., and Hyde, K. D. (2013). *Phyllosticta capitalensis*, a widespread endophyte of plants. *Fungal Diversity*, 60(1):91–105.

Yonow, T., Hattingh, V., and de Villiers, M. (2013). CLIMEX modelling of the potential global distribution of the citrus black spot disease caused by *Guignardia citricarpa* and the risk posed to Europe. *Crop Protection*, 44:18–28.

8.6 Supplementary material

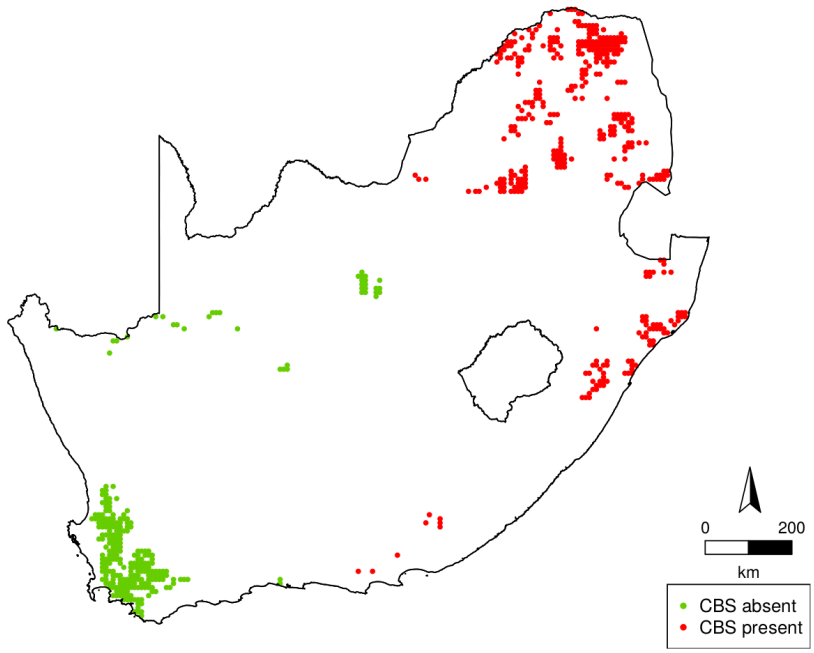

FIGURE 8.7: Validation dataset with citrus black spot (CBS) presences ($n = 385$) and absences ($n = 259$) in 2014, excluding those grid cells used for model development in 1950.

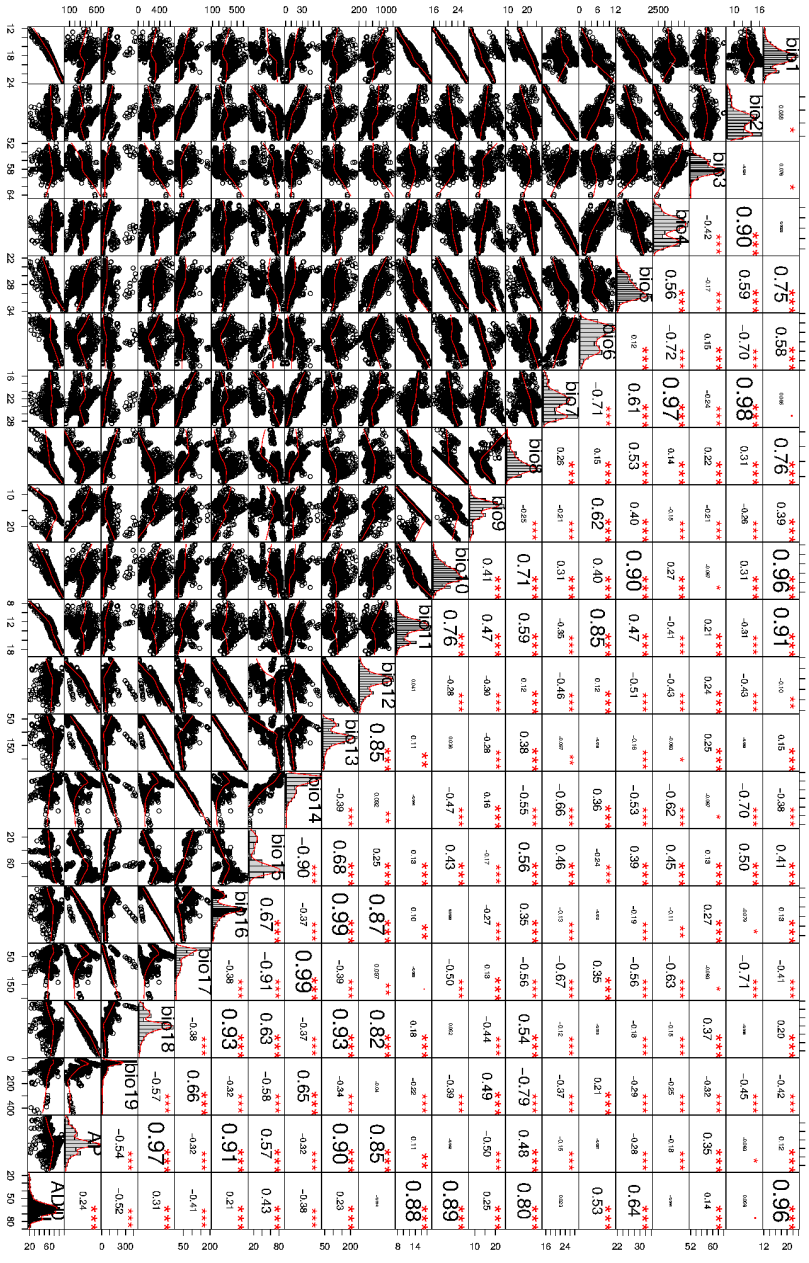

FIGURE 8.8: Correlation matrix for the climatic variables of the 1950 dataset of citrus black spot (CBS) distribution in South Africa.

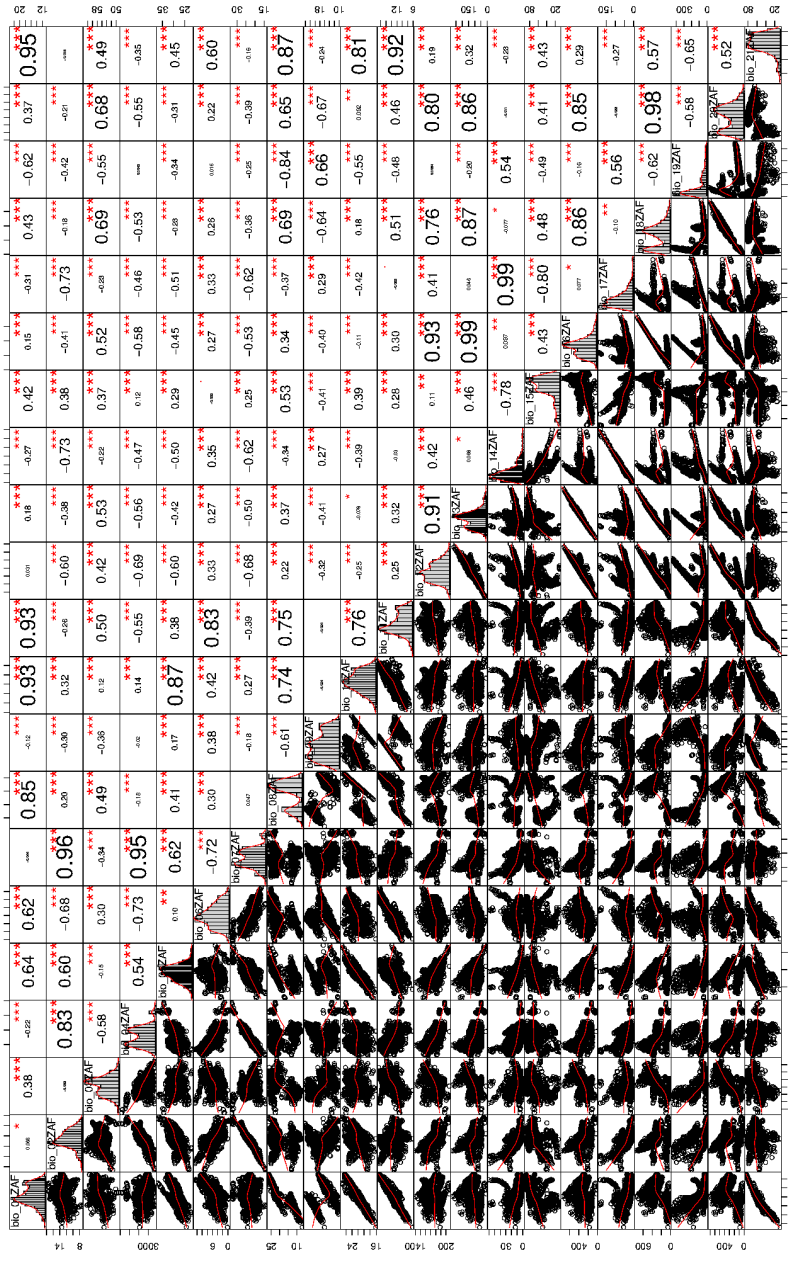

FIGURE 8.9: Correlation matrix for the climatic variables of the 2014 dataset of citrus black spot (CBS) distribution in South Africa.

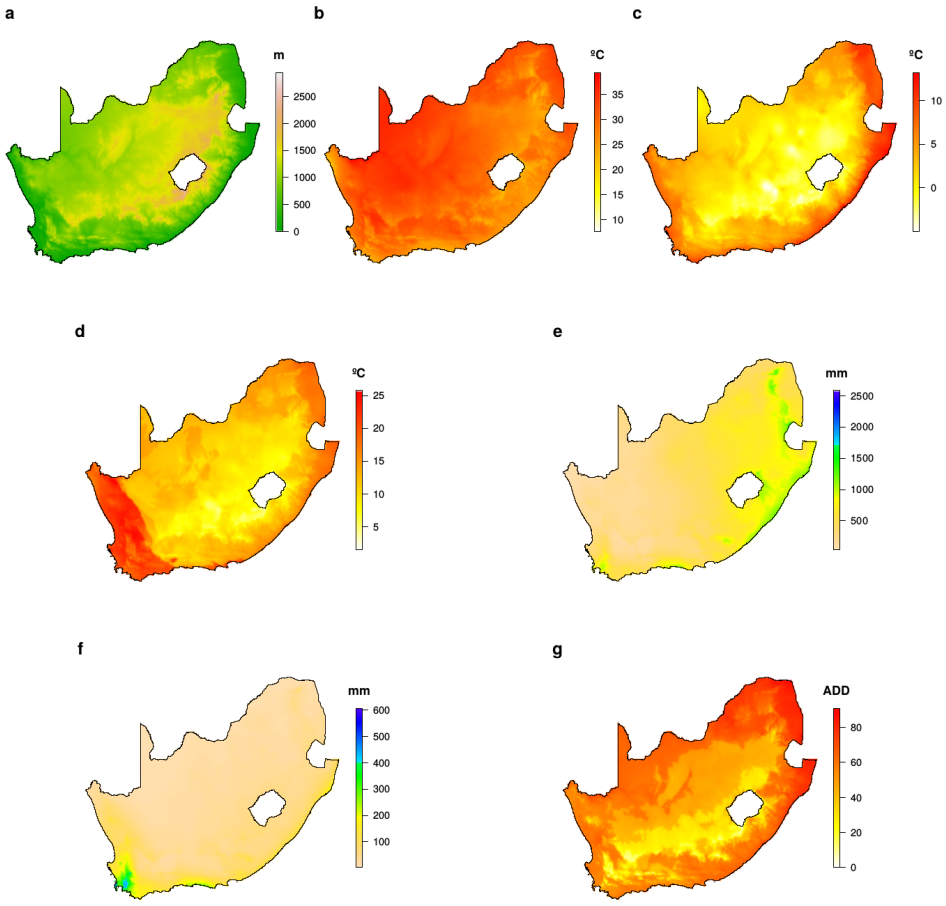

FIGURE 8.10: Maps of **a** altitude; **b** maximum temperature of the warmest month (BIO_5); **c** minimum temperature of the coldest month ($BIO6$); **d** mean temperature of the driest quarter (BIO_9); **e** annual precipitation (BIO_{12}); **f** precipitation of the coldest quarter (BIO_{19}); and **g** accumulated degrees (ADD) from July to October with $T_{base} = 10$ ºC for South Africa obtained from the WorldClim database (Hijmans et al., 2005).

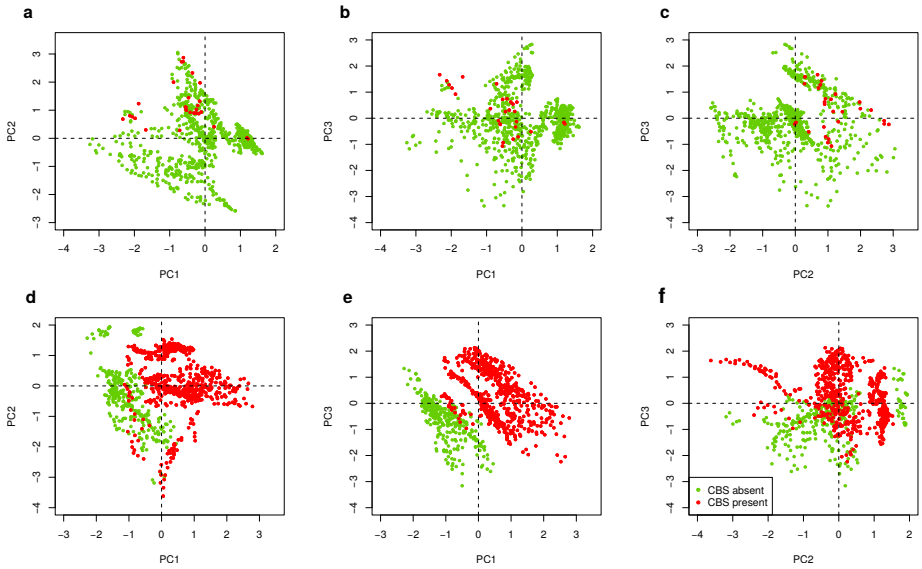

FIGURE 8.11: Scatterplots of the principal components for 1950 (**a,b,c**) and 2014 (**d,e,f**). Red dots are grid cells with citrus black spot (CBS) presence and green dots denote those with CBS absence.

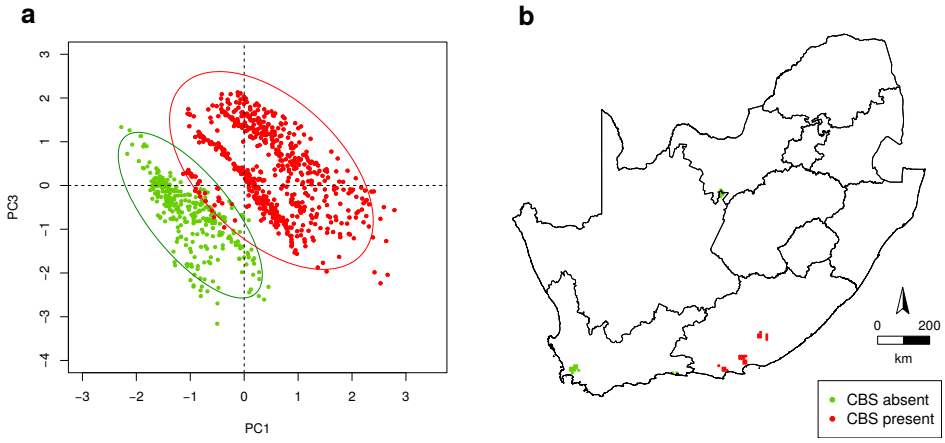

FIGURE 8.12: Scatterplot of the principal components $PC1$ and $PC3$ in 2014 with their corresponding 95% confidence ellipses (**a**), and a map representing the grid cells within the area of overlap of the two ellipses (**b**). Red dots are grid cells with citrus black spot (CBS) presence and green dots denote those with CBS absence.

TABLE 8.3: Best models for 1950 and 2014 with climatic variables (BIO), principal components (PC) and geostatistical term ($\boldsymbol{W}$)

Models[1]		WAIC[2]
1950		
Clim + spatial	$1 + BIO5 + BIO12 + \boldsymbol{W}$	126.14
Climatic	$1 + BIO5 + BIO12$	177.51
	$1 + BIO5 + BIO12 + BIO9 + BIO19 + ADD$	178.07
	$1 + BIO12 + BIO9 + BIO19 + ADD$	178.26
	$1 + BIO5 + BIO12 + BIO19$	178.85
	$1 + BIO6 + BIO12 + BIO9 + BIO19 + ADD$	178.93
	$1 + BIO5 + BIO12 + BIO9$	178.95
	$1 + BIO5 + BIO12 + ADD$	179.01
	$1 + BIO5 + BIO6 + BIO12 + BIO9 + BIO19 + ADD$	179.57
	$1 + BIO5 + BIO6 + BIO12$	179.61
	$1 + BIO12 + ADD$	179.73
PC + spatial	$1 + PC1 + PC2 + PC3 + \boldsymbol{W}$	131.26
PC	$1 + PC1 + PC2 + PC3$	198.19
	$1 + PC1 + PC2$	202.2
	$1 + PC2 + PC3$	202.44
	$1 + PC2$	208.52
	$1 + PC1 + PC3$	235.49
	$1 + PC1$	236.56
	$1 + PC3$	244.15
	1	245.06
2014		
Climatic	$1 + BIO5 + BIO19 + ADD$	49.57
	$1 + BIO5 + BIO6 + BIO19 + ADD$	52.23
	$1 + BIO6 + BIO9 + BIO12 + BIO19 + ADD$	63.77
	$1 + BIO6 + BIO12 + BIO19 + ADD$	65.68
	$1 + BIO5 + BIO6 + BIO19$	70.11
	$1 + BIO6 + BIO12 + BIO19$	70.52
	$1 + BIO6 + BIO9 + BIO12 + BIO19$	72.1
	$1 + BIO5 + BIO6 + BIO9 + ADD$	72.79
	$1 + BIO5 + BIO6 + BIO9 + BIO12 + ADD$	73.67
	$1 + BIO5 + BIO6 + ADD$	77.68
PC	$1 + PC1 + PC3$	100.7
	$1 + PC1 + PC2 + PC3$	101.98
	$1 + PC1 + PC2$	439.13
	$1 + PC1$	520.41
	$1 + PC2 + PC3$	892.93
	$1 + PC3$	946.62
	$1 + PC2$	1144.74
	1	1192.49

[1]Maximum temperature of warmest month (BIO_5), minimum temperature of coldest month (BIO_6), mean temperature of driest quarter (BIO_9), accumulated degrees (ADD) from July to October with T_{base} = 10ºC, annual precipitation (BIO_{12}) and precipitation of coldest quarter (BIO_{19}). [2]Watanabe Akaike Information Criterion.

Chapter 9

A hierarchical Bayesian Beta regression approach to study the effects of geographic genetic structure and spatial autocorrelation on species distribution range shifts

In this chapter, we present a preliminary version of our paper "A hierarchical Bayesian Beta regression approach to study the effects of geographic genetic structure and spatial autocorrelation on species distribution range shifts" by Joaquín Martínez-Minaya (University of Valencia), David Conesa (University of Valencia), Marie-Josée Fortin (University of Toronto), Carlos Alonso-Blanco (CSIC), F. Xavier Picó (CSIC) and Arnald Marcer (CREAF and Autonomous University of Barcelona) which has been accepted in the journal *Molecular Ecology Resources*. The chapter contains at the end the references used in this work.

Abstract

Global climate change (GCC) may be imposing distribution range shifts in many organisms worldwide. Multiple efforts are currently focused on the development of models to better predict distribution range shifts due to GCC. We addressed this issue by including intra-specific genetic structure and spatial autocorrelation (SAC) of data in distribution range models. Both factors reflect the joint effect of eco-evolutionary processes on the geographic heterogeneity of populations. We used a collection of 301 georeferenced accessions of the annual plant *Arabidopsis thaliana* in its Iberian Peninsula range, where the species shows a strong geographic genetic structure. We developed spatial and non-spatial hierarchical Bayesian models (HBMs) to depict current and future distribution ranges for the four genetic clusters detected. We also compared the performance of HBMs with Maxent (a presence-only model). Maxent and non-spatial HBM presented some shortcomings, such as the loss of accessions with high genetic admixture in the case of Maxent and the presence of residual SAC for both. As spatial HBMs removed residual SAC, these models showed higher accuracy than non-spatial HBMs and handled the spatial effect on model outcomes. The ease of modelling and the consistency among model outputs for each genetic cluster was conditioned by the sparseness of the populations across the distribution range. Our HBMs enrich the toolbox of software available to evaluate GCC-induced distribution range shifts considering both genetic heterogeneity and SAC, two inherent properties of any organism that should not be overlooked.

Keywords

Arabidopsis thaliana, geographic genetic structure, global climate change, hierarchical Bayesian models, Maxent, spatial autocorrelation.

9.1 Introduction

Climate and land-use changes recorded in practically all Earth's bioclimatic zones are dramatically affecting the distribution of many terrestrial, aquatic and marine organisms. Since the turn of the century, various global meta-analyses have quantified the fingerprint that global climate change (GCC) has already left on distribution ranges (Parmesan and Yohe, 2003; Perry et al., 2005; Parmesan, 2006; Chen et al., 2011; MacLean and Beissinger, 2017) and extinction rates (Urban, 2015; Wiens, 2016). At present, models are including some improvements to better predict changes in species distribution ranges due to GCC. Such improvements are aimed at considering all the possible organisms' responses to GCC, such as shifts in dispersal ability, phenology and physiology of life-history traits (Bellard et al., 2012; Lenoir and Svenning, 2015). However, precise data on these responses are lacking for many organisms because of the intensive amount of labour and data n eeded to estimate them properly for multiple populations across large areas of the distribution range. Nonetheless, modelling approaches clearly have to go beyond presence-background models and related approaches (e.g. presence-absence models, random pseudo-absence point models) using current and future climatic conditions to increase their accuracy and reliability (Guisan et al., 2017). In this study, we address this issue by considering two important biological aspects that should be considered when modelling current distribution range of terrestrial organisms and their GCC-induced shifts.

Firstly, demographic processes (i.e. extinction/colonisation dynamics and dispersal ability) and adaptation to local environmental conditions determine the extent of population stratification, that is, geographically distributed allele frequencies depicting subpopulations or clusters at different spatial scales (Anderson et al., 2010). Molecular data are commonly used to infer the number of genetically differentiated clusters and their degree of admixture (Pritchard et al., 2000; Falush et al., 2003). In addition, it is very informative to determine whether such genetic clusters are geographically distributed across the species distribution range (Rosenberg et al., 2005; Novembre et al., 2008). This is because we can interpret the number of genetic clusters, their geographic distribution and their degree of admixture

as the result of all demographic processes and adaptive forces acting on populations. This paradigm has steadily gained ground in studies estimating future distribution range shifts due to GCC by means of species distribution models (SDMs, Bálint et al., 2011; Jay et al., 2012; D'Amen et al., 2013; Yannic et al., 2014; Gotelli and Stanton-Geddes, 2015; Diniz-Filho et al., 2016; Marcer et al., 2016; Ikeda et al., 2017; Lima et al., 2017; Milanesi et al., 2018), stressing the need to consider the genetic heterogeneity inherent in organisms.

From a methodological viewpoint, working with intra-specific patterns of genetic diversity implies the combination of presence-only data, which commonly feed SDMs such as Maxent, with genetic structure data, which are mostly expressed as genetic cluster membership proportions (ranging between 0 and 1) that inform on the degree of genetic admixture (Serra-Varela et al., 2017). The problem arises when admixture information is lost because individuals have to be assigned to a single cluster, generally the one with the highest membership proportion according to some arbitrary threshold in order to run presence-only SDMs (Gotelli and Stanton-Geddes, 2015; Marcer et al., 2016; Ikeda et al., 2017). In doing that, the amount of data and information lost depends on the genetic structure of the study organism. For instance, species with a pronounced genetic structure will likely have individuals with high genetic cluster membership proportions, which facilitates the assignment of individuals to single genetic clusters. In contrast, individual assignment to single genetic clusters will exhibit higher uncertainty for weakly genetically structured organisms (e.g. high levels of individual genetic admixture), posing problems for the development of SDMs to study the effects of GCC on their patterns of geographic genetic structure. Either way, we lose valuable information that may reduce the value and impact of GCC model outcomes and therefore our understanding of the GCC effects on biodiversity.

The second biological aspect worth considering when modelling distribution ranges is the presence of spatial autocorrelation (SAC) in data and the problems that SAC poses for statistical and ecological interpretation. SAC can be defined as the dependence between close observations in space (Legendre and Legendre, 2012), and it may be caused by exogenous factors (e.g.

historical processes and autocorrelated environmental variables) and/or endogenous factors (e.g. dispersion) (Dale and Fortin, 2014). While variables representing exogenous factors may be readily available to researchers, variables describing endogenous factors representing important biological processes are more difficult to obtain (Belmaker et al., 2015). Overall, SAC is recognised as a major challenge when predicting species distributions (Dirnböck and Dullinger, 2004; de Oliveira et al., 2014) because it results in several modelling flaws, such as violation of the assumption of independent error, inflated estimations of model performance, bias in model selection, or inferential problems (Legendre, 1993; Dale and Fortin, 2002; Dormann, 2007; F. Dormann et al., 2007; Fortin and Dale, 2009; Beale et al., 2010; Swanson et al., 2013; Wagner and Fortin, 2013). To a certain extent, SAC can be dealt with data filtering, although often at a high cost of data loss. For these reasons, taking SAC into account in GCC models is considered as mandatory (Latimer et al., 2006; Beguin et al., 2012; Record et al., 2013; Swanson et al., 2013; Crase et al., 2014).

Here, we use hierarchical Bayesian models (HBMs), which account for the geographic distribution of intra-specific genetic diversity and the presence of SAC, to analyse current distribution range as well as the effect of GCC on its shifts. To that end, we use a collection of 301 natural populations of the annual plant *Arabidopsis thaliana* occurring in the Iberian Peninsula, the region of the distribution range with the highest genomic diversity (The 1001 Genomes Consortium, 2016). Genome-wide markers are used to infer Iberian *A. thaliana*'s geographic genetic structure by estimating genetic cluster membership proportions. In order to better understand the potential of our model, we compare three approaches, Maxent and two Bayesian, representing a gradient of complexity in the treatment of intra-specific genetic data and SAC. In particular, (1) presence-only SDMs (Maxent) that do not take SAC into account and based on binary data for genetic cluster membership proportions, (2) non-spatial hierarchical Bayesian models (HBMs) not accounting explicitly for SAC and based on continuous data for genetic cluster membership proportions, and (3) spatially-explicit HBMs accounting for SAC and based on continuous data for genetic cluster membership proportions. Although HBMs represent well-established methods for statistical inference in several research fields, the application

of Beta regression to climate-driven shifts in species distribution range is not common (see Martínez-Minaya et al., 2018). In particular, we promote the use of Beta regressions where data fitting can be achieved using integrated nested Laplace approximation (INLA) rather than Markov chain Monte Carlo (MCMC) methods. We discuss our results in terms of the relevance of intra-specific genetic variation and SAC to better interpret and contextualise the implications of GCC on species distribution range shifts, but also identifying the limitations and caveats of our approach.

9.2 Materials and Methods

9.2.1 Source populations and genetic structure

We used a collection of 301 natural populations of the annual plant *Arabidopsis thaliana* distributed across the Iberian Peninsula (ca. 800×700 km^2; 36.00º N – 43.48º N, 3.19º E - 9.30º W; Figure 9.1a). This set of populations belongs to a long-term project pursuing a permanent collection of natural populations from the western Mediterranean Basin (Spain, Portugal and North Africa) intended to unravel *A. thaliana*'s evolutionary ecology, functional genetics, and response to GCC (see Marcer et al., 2018, and references therein). Distance among populations and altitudes had a range of 1 – 1,038 km (mean ± SD = 360.9 ± 200.2 km) and 1 - 2,662 m.a.s.l (mean ± SD = 786.5 ± 391.3 m.a.s.l.), respectively, including a wide array of wild and humanised environments (Picó et al., 2008; Méndez-Vigo et al., 2011; Manzano-Piedras et al., 2014).

Populations included in this study come from field surveys that spanned 12 years (2000 - 2011). We sampled seed from several individuals from each population. Every year and a few months after every survey, field-collected seed was multiplied by the single seed descent method in a glasshouse at the Centro Nacional de Biotecnología (CNB-CSIC) in Madrid. Multiplied seeds were stored in dry, dark conditions in cellophane bags at room temperature, storing conditions that can preserve *A. thaliana* seeds for years. In this study, we employed one representative individual (accession hereafter) per population to analyse the genetic structure of *A. thaliana* in the

Iberian Peninsula. Importantly, accessions exhibited common phenotypes within their populations based on flowering time and/or the vernalization requirement for flowering, which are traits under strong selection in Iberian *A. thaliana* (Méndez-Vigo et al., 2013) that appear to be mediated by variation in temperature (Méndez-Vigo et al., 2011; Vidigal et al., 2016). This procedure increased the odds of using accessions best suited to their local environments and, therefore, common genotypes in the populations.

Nuclear genetic data were obtained from 250 genome-wide single nucleotide polymorphisms (SNPs) previously used to genetically characterise Iberian *A. thaliana* (Picó et al., 2008; Gomaa et al., 2011; Méndez-Vigo et al., 2011; Manzano-Piedras et al., 2014; Marcer et al., 2016). In short, SNPs were selected based on their polymorphism shown in Central Europe, the Iberian Peninsula and in a worldwide collection of accessions, and genotyped using the SNPlex technique (Applied Biosystems, Foster City, CA, USA). These SNPs are located across the genome at putatively neutral regions spaced at 0.5 Mb average intervals (range = 0.11 Kb – 1.82 Mb). All accessions were genetically different from each other.

Genetic structure was assessed using the Bayesian model-based clustering algorithm implemented in STRUCTURE v.2.3.3 (Falush et al., 2003), as previously described (Méndez-Vigo et al., 2011, 2013). In brief, model settings included haploid multilocus genotypes, correlated allele frequencies between populations and a linkage model. Each run consisted of 50,000 burn-in MCMC iterations and 100,000 MCMC after-burning repetitions for parameter estimation. To determine the K number of ancestral populations and the ancestry membership proportions of each accession in each population, the algorithm was run 20 times for each defined number of groups (K value) from 1 to 10. The number of distinct genetic groups was determined by testing the differences between the data likelihood for successive K values using Wilcoxon tests for two related samples. The final K number was estimated as the largest K value with significantly higher likelihood than that of K-1 runs (two-sided $P < 0.005$). This was supported by a high similarity among the ancestry membership matrices from different runs of the same K value (H'=0.99). The average symmetric similarity coefficient H' among runs and the average matrix of ancestry membership proportions, derived from the 20 runs, were calculated with CLUMPP v.1 (Jakobsson and

Rosenberg, 2007). This analysis inferred four genetic clusters in the Iberian Peninsula (Figure 9.1), in agreement with previous studies on *A. thaliana*'s genetic structure (Picó et al., 2008; Gomaa et al., 2011; Méndez-Vigo et al., 2011; Manzano-Piedras et al., 2014; Marcer et al., 2016).

9.2.2 Climatic variables and GCC scenarios

We selected a total of eight bioclimatic predictors to define the climatic space: BIO_1 (annual mean temperature), BIO_2 (mean diurnal range), BIO_3 (isothermality), BIO_4 (temperature seasonality), BIO_8 (mean temperature of the wettest quarter), BIO_{12} (annual precipitation), BIO_{15} (precipitation seasonality) and BIO_{18} (precipitation of the warmest quarter). These bioclimatic predictors were selected because their pairwise correlation coefficients were less than 0.7, a threshold value commonly used to avoid unacceptable co-linearity among independent variables (Pino et al., 2005). We used the dismo R package (Hijmans et al., 2017) to retrieve these climate layers from the Digital Atlas of the Iberian Peninsula (`http://www.opengis.uab.cat/wms/iberia/`), which provides interpolated surface layers of mean monthly data obtained from 2,285 weather stations for the period 1951 - 1999. We refer to this time period as year 2000.

We chose the year 2070 as the scenario to evaluate *A. thaliana*'s distribution range shifts due to GCC. In order to use the most and least conservative GCC scenarios, we selected the representative concentration pathways (RCP) 2.6 and 8.5 (Van Vuuren et al., 2011), respectively. In addition, we averaged four climate models: HadGEM2-ES (Met Office Hadley Centre, UK), MRI-CGCM3 (Meteorological Research Institute, Japan), MIROC-ESM (Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute, The University of Tokyo and National Institute for Environmental Studies, Japan), and NorESM1-M (Norwegian Climate Centre, Norway). Data for 2070 were downloaded from the WorldClim Global Climate Database v.1.4 (Hijmans et al., 2005). The resolution of the climatic spaces for the years 2000 and 2070 was 1 km.

9.2.3 Climatic variables and GCC scenarios

SDMs link information on the presence/absence or abundance of a species to environmental variables to predict where it is likely to be present in unsampled locations or time periods (Guisan and Thuiller, 2005; Elith and Leathwick, 2009). In the last years, the quantity and the quality of the datasets have substantially increased, resulting in a higher complexity of the statistical issues that have to be addressed when an SDM is created. As a result of this increasing complexity, the performance of the SDM inferential and predictive processes are becoming more challenging, forcing researchers to develop new sophisticated statistical techniques (see a review in Martínez-Minaya et al., 2018). Given the flood of methodologies developed to address this issue, we compared SDMs obtained with two contrasting approaches: a presence-only model (Maxent) and the hierarchical Bayesian Beta regressions (with and without a spatial term).

9.2.4 Maxent

We used Maxent v.3.3.3k (Phillips et al., 2006; Elith et al., 2011) to model the current and future distribution of *A. thaliana*'s genetic clusters. As Maxent uses presence-only data, we assigned each of the 301 accessions to its predominant genetic cluster using a cut-off value of 0.5 to each genetic cluster membership proportion given by STRUCTURE (as in Marcer et al., 2016). As a result, the number of accessions per genetic cluster was reduced resulting in a low of 35 for genetic cluster C4 to a high of 103 accessions for genetic cluster C1 (Figure 9.1b). The mean ($\pm$ SE) membership proportions to each genetic cluster were 0.66 $\pm$ 0.01 (range = 0.51 - 0.92) for genetic cluster C1, 0.74 $\pm$ 0.02 (range = 0.52 - 0.96) for genetic cluster C2, 0.89 $\pm$ 0.02 (range = 0.56 - 0.97) for genetic cluster C3, and 0.77 $\pm$ 0.02 (range = 0.51 - 0.94) for genetic cluster C4. Eighty-two accessions (27.2%) did not have any genetic cluster membership proportion higher than 0.5 and could not be included in the Maxent models, stressing one of the limitations of this approach when dealing with accessions with high genetic admixture. We fitted all possible models determined by the set of combinations between the eight climatic predictors without considering interactions. We then ranked

these models according to the five-fold cross-validated area-under-the-curve (AUC) metric and chose the most parsimonious one among the best five (Table 9.4). Then, we ran again the chosen model with all data points to obtain the final model. Maxent was used with default parameters with the exception of features, which were limited to the hinge type, making it similar to a Generalized Additive Model (Elith et al., 2011).

9.2.5 Hierarchical Bayesian Beta regression

Spatial and non-spatial HBMs were also used to model the current and future distribution of *A. thaliana*'s genetic clusters. In particular, spatial and non-spatial Beta regressions were conducted to estimate the genetic cluster membership probability, which in this particular context, can be thought of as the habitat suitability for each genetic cluster. In contrast to Maxent, Beta regressions allowed us to model each genetic cluster separately using all genetic information available, that is, the genetic cluster membership proportions of all 301 accessions. In other words, no data were excluded.

The class of Beta regression models is commonly used to model variables that assume values in the unit interval (between 0 and 1; Ferrari and Cribari-Neto, 2004), such as the case of membership probabilities of genetic clusters. A Beta distribution depends on two scaling parameters, Beta(a, b), which can be parameterised in terms of its mean, μ, a dispersion parameter, $\phi = a + b$ and the variance, $\sigma^2 = \frac{\mu(1-\mu)}{1+\phi}$. This parameterisation better supports the truncated nature of the Beta distribution because the variance depends on the mean, which translates into maximum variance at the centre of the distribution and minimum variance at the edges. In addition, the dispersion of the distribution for a fixed μ decreases as ϕ increases. We did not transform the data to avoid the problems posed by extreme values, as proposed elsewhere (Cribari-Neto and Zeileis, 2010), because data fell far enough from the extremes of the Beta distribution.

As we were interested in depicting the relationship between the genetic cluster membership probabilities and the bioclimatic predictors, we linked the mean and the precision of the response variable to the linear bioclimatic predictors via suitable link functions. In particular, if Y_i represents the

genetic cluster membership probability at location i, then its conditional distribution is $Y_i \mid \mu_i, \phi_i \sim \text{Beta}(\mu_i, \phi_i)$, where μ_i and ϕ_i are the Beta distribution parameters at location i. We used the logit and log links for μ_i and ϕ_i, respectively. The mean was linked to climatic covariates (non-spatial term) and, in the case of spatial models, to a stochastic spatial effect (spatial term). The precision was assumed to be not dependent of any effect. The resulting model with a spatial term is known as a point-referenced spatial Beta regression (Paradinas et al., 2016, 2018). It is highly suitable for situations in which data are observed at continuous locations occurring within a defined spatial domain:

$$\begin{aligned} \text{logit}(\mu_i) &= \boldsymbol{X}_i\boldsymbol{\beta} + W_i\,, \\ \log(\phi_i) &= \theta\,. \end{aligned} \tag{9.1}$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients $(\beta_0, \beta_1, \ldots, \beta_c)$, $\boldsymbol{X}_i$ is the vector corresponding to the ith row of the design matrix whose first element is 1 (the one multiplying the intercept β_0), the covariate values at location i being the remaining elements, and W_i is the spatially structured random effect at each location i. $\boldsymbol{W}$ is assumed to be a multivariate Gaussian distribution whose covariance matrix, $\sigma^2_{\boldsymbol{W}} H(\varphi)$, depends on the distance between locations, and its parameters, $\sigma^2_{\boldsymbol{W}}$ and φ, represent the variance and range of the spatial effect, respectively.

In the context of HBMs, parameters were treated as random variables and prior knowledge was incorporated using the corresponding prior distributions. These priors were specified in the second stage jointly with random effects. In the third and final level of the hierarchy, prior knowledge about the hyper-parameters was expressed. This hierarchical structure can also be considered as a latent Gaussian model (Rue and Held, 2005).

As posterior distributions for the parameters and hyper-parameters do not have an analytic expression, numerical approximations are usually needed. In the case of latent Gaussian models, integrated nested Laplace approximation (INLA; Rue et al., 2009) is a computationally efficient alternative to the MCMC method. However, to fit and predict the particular case of continuously indexed Gaussian fields with INLA, $\boldsymbol{W}$ in our case,

an additional module is required. Lindgren et al. (2011) proposed an approach using an approximate stochastic weak solution to a stochastic partial differential equation (SPDE) as a Gaussian Markov random field (GMRF) approximation to continuous Gaussian fields with Matérn covariance structure, a highly flexible and general family of functions in spatial statistics (Rue and Held, 2005). The Markov property allowed the use of a precision sparse matrix, enabling efficient numerical algorithms. Under this approximation, the spatial effect is re-parameterised as follows:

$$\boldsymbol{W} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q}(\kappa, \tau)) \,. \tag{9.2}$$

Here, $\boldsymbol{W}$ depends on two different parameters, κ and τ, which determine the range of the effect and the total variance, respectively. More precisely, the range is approximately $\varphi = \sqrt{\frac{8}{\kappa}}$ and the variance is $\sigma^2_{\boldsymbol{W}} = \frac{1}{4\pi\kappa^2\tau^2}$ (Lindgren et al., 2011).

We specified prior distributions for the parameters and hyperparameters. In particular, normal vague priors with mean 0 and precision 10^{-4} were used for the vector of regression coefficients. Although internally INLA works with parameters κ and τ, we specified the spatial effect in terms of φ and $\sigma_{\boldsymbol{W}}$ using the re-parameterisations $\log(\varphi)$ and $\log(\sigma_{\boldsymbol{W}}$ as independent normal vague distributions (Lindgren et al., 2015).

Overall, the full model was stated as follows:

$$\begin{aligned}
Y_i \mid \mu_i, \phi_i &\sim \text{Beta}(\mu_i, \phi_i) \\
\text{logit}(\mu_i) &= \boldsymbol{X}_i\boldsymbol{\beta} + W_i \\
\log(\phi_i) &= \theta \\
\\
\beta_0, \beta_1, \ldots, \beta_c &\sim \mathcal{N}(0, 10^{-4}) \\
\boldsymbol{W} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q}(\varphi, \sigma_{\boldsymbol{W}}) \\
\\
\log(\varphi) &\sim \mathcal{N}(m_\varphi, q_\varphi) \\
\log(\sigma_{\boldsymbol{W}}) &\sim \mathcal{N}(m_{\sigma_{\boldsymbol{W}}}, q_{\sigma_{\boldsymbol{W}}}) \\
\theta &\sim \text{logGamma}(0, 0.1)
\end{aligned} \tag{9.3}$$

where m_φ was automatically chosen so that the prior mean of φ was about 50% the diameter of the study geographic region, while $m_{\sigma_{\boldsymbol{W}}}$ was chosen in a way that the corresponding variance of the field was 1. For our analysis, this resulted in $m_\varphi = 13.517$ and $m_{\sigma_{\boldsymbol{W}}} = 0$. Finally, the default a priori precisions for $\log(\varphi)$ and $\log(\sigma_{\boldsymbol{W}})$ distributions were $q_\varphi = 0.25$ and $q_{\sigma_{\boldsymbol{W}}} = 0.25$, respectively.

These latter values, q_φ and $q_{\sigma_{\boldsymbol{W}}}$, express the large uncertainty about the parameters before the analysis, resulting in quite non-informative hyper-priors. This is important because it allows the range to take values between 0 and the total diameter of the Iberian Peninsula. In contrast to Maxent, HBMs can take space into account when modelling distribution ranges, which gives the possibility to evaluate its mean effects and uncertainty. As mentioned above, once the inference is done, the main interest becomes to predict the response in un-sampled locations. To do that, we applied the SPDE by constructing a Delaunay triangulation (Hjelle and Dæhlen, 2006) covering the whole Iberian Peninsula (Figure 9.5).

9.2.6 Model selection, distribution range shifts and residual SAC

HBMs were run with and without the spatial component with the `R-INLA` package (Lindgren et al., 2015) in order to quantify its effects on distribution range shifts with GCC and to be compared with Maxent outcomes. We fitted all possible models given by the set of combinations among the eight climatic predictors without interactions. To select the best model, we used $LCPO = \frac{1}{N}\sum_{i=1}^{N}\log(\text{CPO}_i)$ as a summary statistic of the conditional predictive ordinate (CPO; Geisser, 1993), which gives an overall measure of predictive performance (Hooten and Hobbs, 2015). CPO is defined as the cross-validated predictive density at a given observation. CPO can be used to compute predictive measures, such as the logarithmic score (Gneiting and Raftery, 2007) or the cross-validated mean Brier score (Schmid and Griffith, 2005). Among the best five models for each genetic cluster we selected the most parsimonious one, that is, the one with the least number of predictors. Model quality estimators, such as the deviance information criterion (DIC; Spiegelhalter et al., 2002) and the Watanabe–Akaike information criterion

(WAIC; Watanabe, 2010) were also computed. We also measured accuracy of spatial and non-spatial HBMs by means of mean absolute error (MAE) and root mean squared error (RMSE). Lower values of MAE and RMSE indicate better accuracy. The comparison between Maxent and HBMs in terms of accuracy can be misleading because Maxent used sub-samples of data for each genetic cluster whereas HBMs were always based on the entire data set (Figure 9.1).

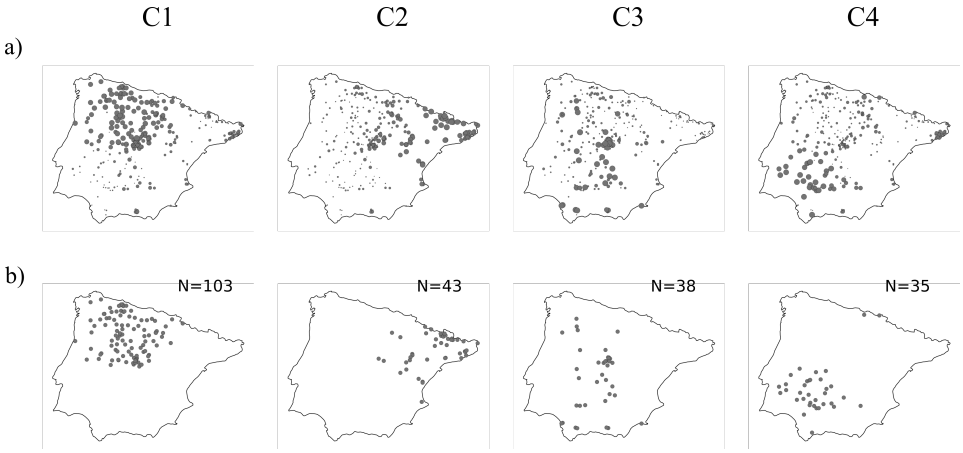

FIGURE 9.1: (a) Geographic position of the 301 *A. thaliana* accessions of study for the four genetic clusters detected in the Iberian Peninsula. Dot size is proportional to the genetic cluster membership proportion. For each accession, the four genetic cluster membership proportions sum to 1. (b) Geographic position of selected accessions after applying the membership proportion threshold of 0.5. The number of accessions included per genetic cluster is also indicated.

We compared distribution range shifts due to GCC when taking the spatial component into account (spatial HBM), when excluding the spatial component (non-spatial HBM), and Maxent. We added the probabilities calculated across the whole study area by each model for each time frame and GCC scenario to quantify the geographic extent of the suitability of each genetic cluster for each methodology. These probabilities were used to calculate the percentage loss or gain of suitability for each model and GCC scenario.

All models mentioned above were checked for residual SAC. We calculated the residuals for model predictions between observed and predicted

values and tested for residual SAC using the `spdep` `R` package (Bivand et al., 2013; Bivand and Piras, 2015). In order to calculate residual SAC in Maxent, we followed the methodology used elsewhere (De Marco Jr et al., 2008; Václavík and Meentemeyer, 2009). Basically, we estimated the Moran's I coefficient of autocorrelation with 10,000 MCMC iterations. Models with P-values > 0.05 were considered as SAC free. As expected, spatial HBM did not show residual SAC, while non-spatial HBM did. All Maxent models, except for genetic cluster C4, also retained residual SAC (Table 9.5).

9.3 Results

9.3.1 Current distribution range

We compared the performance of three modelling approaches, Maxent and HBMs with and without the spatial component, to depict the current distribution range of four *A. thaliana*'s genetic clusters using eight selected bioclimatic predictors. Maxent models included between four and seven bioclimatic predictors per genetic cluster (Table 9.1). With these predictors, Maxent yielded a clear geographic distribution of genetic cluster ranges in the Iberian Peninsula (Figure 9.2a), as found in earlier studies using the same approach but with different data (Marcer et al., 2016).

In the case of Bayesian models, spatial and non-spatial HBMs produced broadly similar geographic distributions of genetic clusters to those generated by Maxent models, particularly for genetic clusters C1 and C4 (Figure 9.2a). In the case of genetic cluster C2, the spatial HBM depicted a rather continuous distribution in NE Spain, which clearly differed from those given by Maxent and non-spatial HBM showing the truncated distribution that this genetic cluster actually has in the wild (Figure 9.2a). In general, spatial HBMs and Maxent models showed more compact distribution ranges than non-spatial HBMs, the former with more dramatic transitions between low and high probability values in all genetic clusters (Figures 9.2a and 9.6). The exception was genetic cluster C3, whose predicted distribution range with non-spatial HBMs was rather blurred in comparison with that obtained with Maxent (Figure 9.2a). In fact, it was not possible to fit spatial HBMs

for genetic cluster C3 because the results were inconsistent. When using vague hyper-priors for the range of the spatial effect, the resulting mean of the posterior distribution of the range was larger than the whole study area. On the contrary, when using more informative priors, results were different and very much conditioned by prior selections. Thus, the spatial effect for genetic cluster C3 did not provide further explanation than what can already be explained by the bioclimatic predictors.

For all genetic clusters, the uncertainty of the predictive mean for non-spatial HBMs was lower and more evenly distributed across space than that for spatial HBMs (Figure 9.2b). The main reason for this apparent reduction in uncertainty is that spatial models are reflecting the intrinsic variability of the Beta-distributed data, variability that is not reflected by non-spatial models. As a result, the distribution of means and standard deviations for spatial HBMs was more pronounced than those for non-spatial HBMs (Figure 9.2 and Table 9.6). Nonetheless, the values of the mean of the posterior distribution of the precision parameter, which are inversely proportional to the variance of the data, were larger in the spatial HBMs, reflecting their acceptable behaviour (Table 9.7). In the case of spatial HBMs, these models allowed the visualisation of the spatial effects, which clearly were more intense at the centre of the genetic cluster distribution ranges (Figure 9.3). Uncertainty of the mean of the spatial effect was greater for genetic cluster C4 than for the other two genetic clusters (Figure 9.3). Overall, spatial HBMs selected less bioclimatic predictors than non-spatial HBMs and Maxent models to define the distribution range of the four genetic clusters (Tables 1 and S5), particularly for genetic cluster C4. Finally, the combination of bioclimatic predictors used by Maxent models and spatial and non-spatial HBMs was quite different (Table 9.1).

Mean absolute error (MAE) and root mean squared error (RMSE) were lower for spatial compared to non-spatial HBMs for all genetic clusters in which the comparison was possible (Table 9.2). This indicated that spatial HBMs had lower average model prediction errors in the response variable.

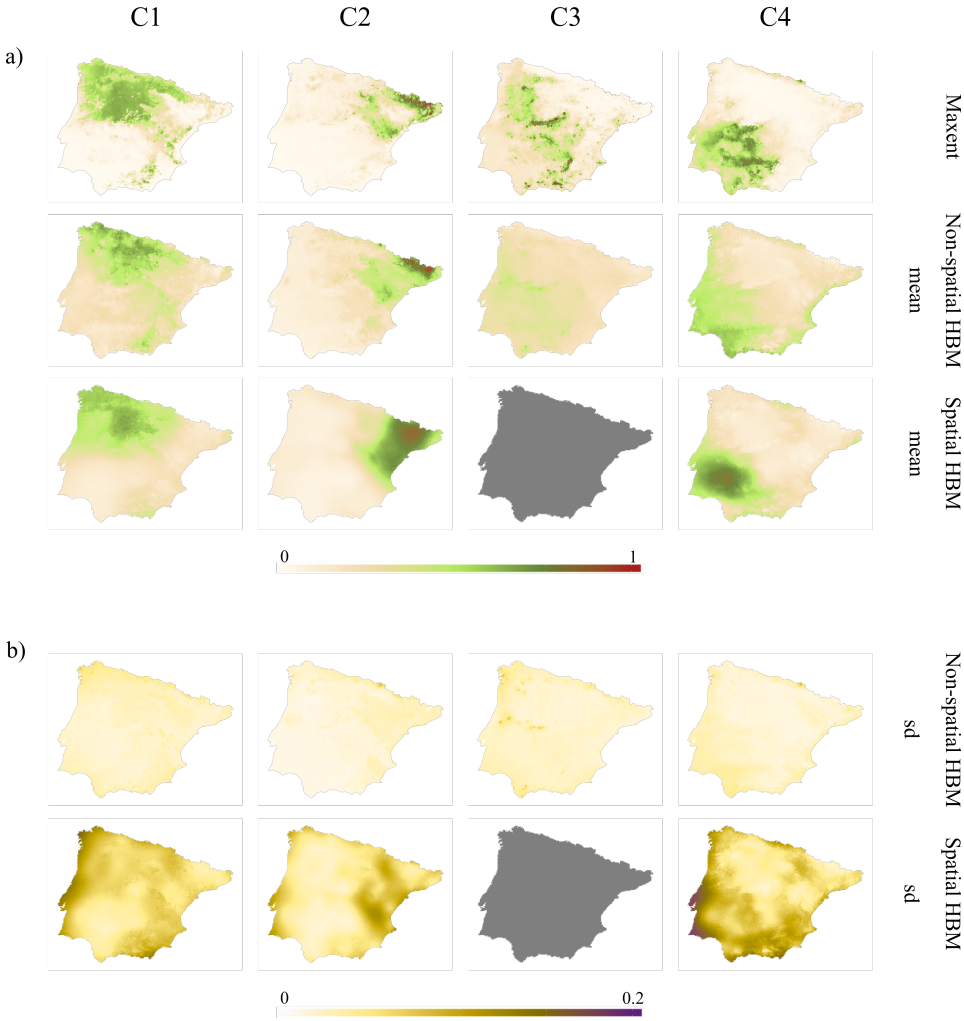

FIGURE 9.2: (a) Predicted current distributions (year 2000) for each *A. thaliana*'s genetic cluster and methodology (Maxent, non-spatial and spatial HBMs). Darker and lighter intensities indicate higher and lower suitability, respectively. (b) Uncertainty of non-spatial and spatial HBMs. Darker and lighter intensities indicate higher and lower uncertainty, respectively. Grey maps for genetic cluster C3 indicate that the spatial HBMs were not acceptable.

FIGURE 9.3: Mean and standard deviation of the spatial effects included in the spatial HBMs. Darker and lighter intensities (logit scale) indicate higher and lower spatial effects, respectively. Grey maps for genetic cluster C3 indicate that the spatial HBMs were not acceptable.

TABLE 9.1: Bioclimatic variable percentage contributions to the fit of the best Maxent models and β coefficients of the best non-spatial and spatial HBMs for the distribution range of each genetic cluster of *A. thaliana* in the Iberian Peninsula. Bioclimatic variables: BIO_1; Annual mean temperature, BIO_2; Mean diurnal range, BIO_3; Isothermality, BIO_4; Temperature seasonality, BIO_8; Mean temperature of the wettest quarter, BIO_{12}; Annual precipitation, BIO_{15}; Precipitation seasonality, and BIO_{18}; Precipitation of the warmest quarter. For Maxent, the number of occurrence points was 103, 43, 38, and 35 for genetic clusters C1, C2, C3 and C4, respectively. For non-spatial and spatial HBMs, models included all 301 occurrence points.

Cluster	**Model**	**Bioclimatic predictors**							
		BIO_1	BIO_2	BIO_3	BIO_4	BIO_8	BIO_{12}	BIO_{15}	BIO_{18}
C1	Maxent	64.07	–	–	9.73	17.30	–	1.74	7.16
	Non-spatial HBMs	–	–	5.773	-0.565	-0.104	–	-0.050	-0.009
	Spatial HBMs	0.147	–	–	–	-0.071	–	–	–
C2	Maxent	–	3.01	–	20.06	7.50	–	–	69.43
	Non-spatial HBMs	-0.112	–	–	0.373	0.104	-0.001	–	0.014
	Spatial HBMs	–	–	–	–	-0.044	-0.002	0.025	0.010
C3	Maxent	19.76	–	33.11	–	–	–	38.42	8.70
	Non-spatial HBMs	–	–	–	0.288	–	0.001	–	-0.006
	Spatial HBMs	–	–	–	–	–	–	–	–
C4	Maxent	22.65	2.52	–	11.54	–	20.18	–	43.11
	Non-spatial HBMs	0.197	–	–	–	–	–	0.021	0.004
	Spatial HBMs	0.164	–	–	–	–	–	–	–

TABLE 9.2: Mean absolute error (MAE) and root mean squared error (RMSE) for spatial and non-spatial HBMs applied to each genetic cluster of *A. thaliana* in the Iberian Peninsula. The spatial effect term ($\boldsymbol{W}$) is also indicated in spatial HBMs.

Cluster	**HBM**	**Model**	**MAE**	**RMSE**
C1	Non-spatial	$Y \sim \beta_0 + BIO_3 + BIO_4 + BIO_8 + BIO_{15} + BIO_{18}$	0.174	0.210
	Spatial	$Y \sim \beta_0 + BIO_1 + BIO_8 + \boldsymbol{W}$	0.134	0.171
C2	Non-spatial	$Y \sim \beta_0 + BIO_1 + BIO_4 + BIO_8 + BIO_{12} + BIO_{18}$	0.116	0.153
	Spatial	$Y \sim \beta_0 + BIO_8 + BIO_{12} + BIO_{15} + BIO_{18} + \boldsymbol{W}$	0.070	0.095
C3	Non-spatial	$Y \sim \beta_0 + BIO_4 + BIO_{12} + BIO_{18}$	0.217	0.268
	Spatial	–	–	–
C4	Non-spatial	$Y \sim \beta_0 + BIO_1 + BIO_{15} + BIO_{18}$	0.148	0.189
	Spatial	$Y \sim \beta_0 + BIO_1 + \boldsymbol{W}$	0.096	0.128

TABLE 9.3: Predicted cumulative probabilities for the entire Iberian Peninsula and percentage change, with respect to values in 2000, per genetic cluster, GCC scenario and modelling approach for each of the genetic clusters of *A. thaliana* in the Iberian Peninsula.

Cluster	**GCC**	**Maxent**		**Non-spatial HBMs**		**Spatial HBMs**	
		Cum. Prob.	% Change	Cum. Prob.	% Change	Cum. Prob.	% Change
C1	2000	5255.60	–	7277.59	–	7106.33	–
	RCP 2.6	3155.53	-39.96	6063.70	-16.68	7505.95	5.62
	RCP 8.5	1250.86	-76.20	4003.50	-44.99	8468.49	19.17
C2	2000	3020.12	–	4917.93	–	5318.48	–
	RCP 2.6	3746.01	24.04	5072.67	3.15	4953.02	-6.87
	RCP 8.5	2726.87	-9.71	4961.37	0.88	5194.97	-2.32
C3	2000	4540.23	–	6092.25	–	–	–
	RCP 2.6	6167.06	35.83	5984.34	-1.77	–	–
	RCP 8.5	8372.54	84.41	6343.11	4.12	–	–
C4	2000	4473.78	–	6698.21	–	6428.20	–
	RCP 2.6	5939.25	32.76	7894.27	17.86	7527.49	17.10
	RCP 8.5	4916.06	9.89	10378.67	54.95	8975.59	39.63

9.3.2 Distribution range shifts with GCC

Maxent models and HBMs were also used to quantify distribution range shifts of *A. thaliana*'s genetic clusters under different GCC models and scenarios. The three modelling approaches yielded different GCC predictions for each genetic cluster based on suitability shifts in distribution range projections (Table 9.3, Figure 9.4 and S3). Overall, Maxent showed a trend of predicting more dramatic changes in distribution range due to GCC for all genetic clusters compared to spatial and non-spatial HBMs (Table 9.3 and Figure 9.4). For genetic cluster C1, important reductions in distribution range were predicted for the two GCC scenarios with Maxent and non-spatial HBMs, whereas spatial HBMs predicted slight increases (Table 9.3 and Figure 9.4). For genetic cluster C2, Maxent predicted increasing and decreasing distribution ranges with RCP 2.6 and RCP 8.5, respectively, whereas both HBMs predicted small fluctuations in distribution range in both GCC scenarios (Table 9.3 and Figure 9.4). For genetic cluster C3, Maxent showed very large increases in distribution range, particularly for the RCP 8.5 scenario, whilst non-spatial HBMs predicted slight fluctuations in distribution range in both GCC scenarios (Table 9.3 and Figure 9.4). Finally, for genetic cluster C4, all approaches predicted increases in distribution range in both GCC scenarios. Maxent gave higher increases in RCP 2.6 than in RCP 8.5 and vice-versa for both HBMs (Table 9.3 and Figure 9.4).

9.4 Discussion

Distribution range shifts represent the most important effect of GCC on biodiversity because of their ecological implications and the potentially detrimental socio-economic impact on society. GCC models for distribution range shifts have to increase their sophistication by adding realism to the model outcomes, yet without losing interpretability or increasing uncertainty. In this study, we address this issue by developing hierarchical Bayesian models (HBMs) for the annual plant *Arabidopsis thaliana* incorporating two of these elements, which are inherent to practically all organisms: the geographic distribution of intra-specific genetic variation and the spatial

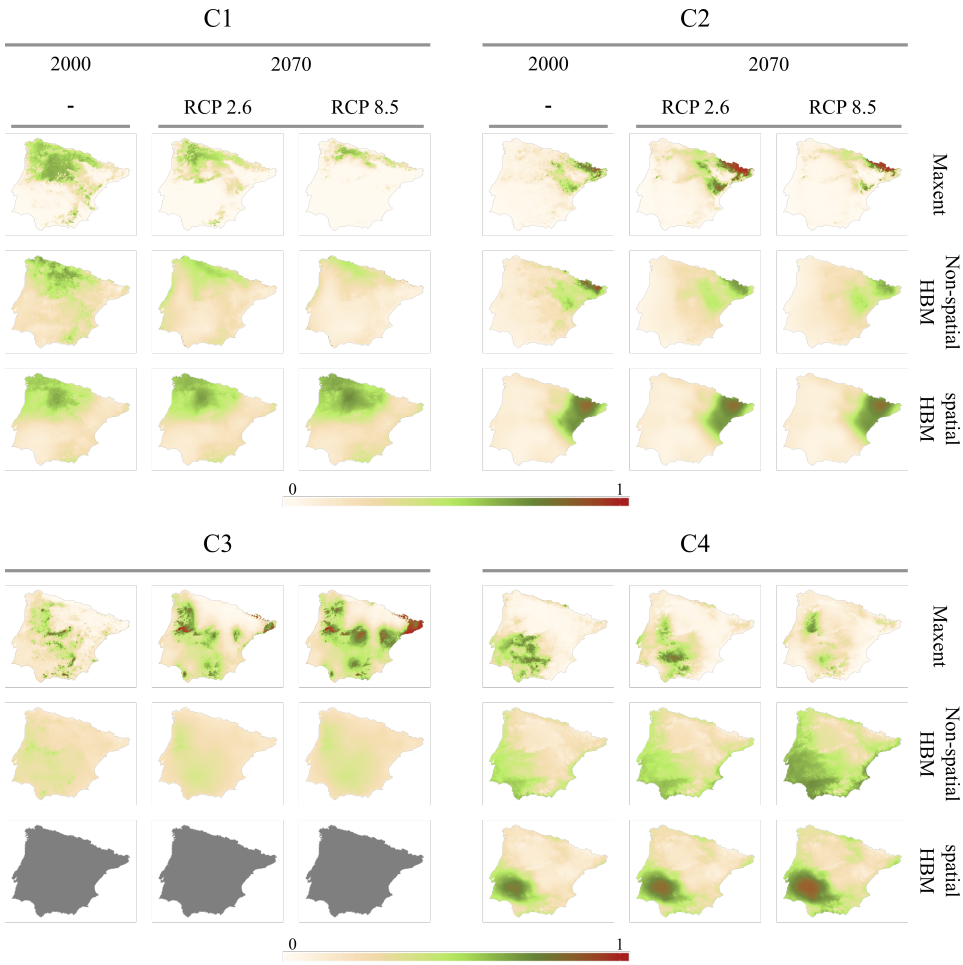

FIGURE 9.4: Predicted distributions in year 2070 for each *A. thaliana*'s genetic cluster and methodology (Maxent, non-spatial and spatial HBMs) under the two GCC scenarios (RCP 2.6 and RCP 8.5). For the sake of completeness, predicted current distributions in year 2000 given in Fig. 2 are also shown. Darker and lighter intensities indicate higher and lower suitability, respectively. Grey maps for genetic cluster C3 indicate that the spatial HBMs were not acceptable.

autocorrelation in data. Importantly, both geographic genetic structure and spatial autocorrelation can be considered as indicators of eco-evolutionary forces shaping species' distribution range, such as colonization/extinction dynamics, dispersal ability, local adaptation and historical factors.

9.4.1 Current distribution range

The selection of the modelling approach may have significant repercussions when considering a species as a genetically heterogeneous organism, whose geographic distribution of its genetic variation may have major implications for understanding the effects of GCC on its distribution range. In the case of Maxent, and of any modelling technique dealing with binary response variables, a major problem is the loss of data resulting from the conversion of a continuous variable (i.e. the genetic cluster membership proportions characterising each individual) into a binary variable (i.e. the assignment of each individual to the genetic cluster with the highest membership proportion) (Gotelli and Stanton-Geddes, 2015; Marcer et al., 2016). In our study, binarization of genetic data had an important cost in terms of data loss as 82 of 301 accessions did not reach the minimum membership proportion of 0.5 required to be assigned to a genetic cluster. As a result, the number of accessions per genetic cluster used in Maxent was reduced (Figure 9.1b). In contrast, HBMs did not have this limitation and used the whole set of 301 accessions also including genetic admixture among the four genetic clusters detected in the Iberian Peninsula. It must be emphasised that accessions exhibiting genetic admixture are quite relevant in biological terms. For example, they may indicate the existence of contact zones between genetic lineages where important processes affecting the distribution range may take place, such as the balance between selection and dispersal against hybrids (Barton and Hewitt, 1985). For this reason, HBMs represent a better choice to model distribution ranges of intra-specific genetic lineages if it is undesirable to discard accessions with too much genetic admixture.

Broadly speaking, Maxent and the Bayesian modelling approaches were consistent in depicting the current geographic distribution of the four genetic clusters of Iberian *A. thaliana* (Figs 1 and 2). The exception was the genetic cluster C3, in which non-spatial HBM blurred the distribution range and spatial HBM was not able to produce results due to unacceptable outcomes in a Bayesian framework. Interestingly, the genetic cluster C3 is strongly differentiated from the rest of clusters found in the Iberian Peninsula, as well as across the whole species' distribution range. In fact, the genetic cluster C3 is considered as the relict cluster with a long evolutionary history (Picó

et al., 2008; Brennan et al., 2014; The 1001 Genomes Consortium, 2016; Durvasula et al., 2017). The relict nature of the genetic cluster C3 is also supported by its scattered distribution across the Iberian Peninsula, a geographic distribution that is interpreted as the result of Iberian glacial refugia (Picó et al., 2008; Brennan et al., 2014; Marcer et al., 2016), whereas the rest of the genetic clusters exhibit geographically marked distributions, likely as a result of more recent demographic histories. Overall, this result indicates that modelling the distribution range of genetic clusters or species with scattered distributions may be difficult no matter what modelling approach is applied. For the particular case of genetic cluster C3, characterised by the high genetic membership proportions of their accessions and the relatively low admixture with other genetic clusters, Maxent predicts its distribution best.

For the rest of genetic clusters with marked geographic distributions (NW, NE and SW Iberian Peninsula for genetic clusters C1, C2 and C4, respectively), Maxent and Bayesian approaches were able to model their current distribution ranges. However, they exhibited some differences among genetic clusters. For example, genetic cluster C2 exhibits a disjunct distribution due to a major geographic barrier (i.e. the Ebro river valley in NE Spain), which was clearly depicted by Maxent (Figure 9.2). It is worth noting that such disjunt distribution is not a sampling problem, but the result of the low occurrence of the species confirmed after repeated field campaigns in the region. Bayesian Beta regression approaches, particularly the spatial HBM, blurred, albeit not totally erasing, the disjunct distribution of this genetic cluster. In contrast, Maxent and Bayesian Beta regression approaches were more consistent for genetic clusters C1 and C4, which exhibited more compact distributions. Overall, we conclude that the continuity of clusters' distribution range increases its suitability to be modelled by alternative means.

Spatial HBMs, along with Maxent for genetic cluster C4, did not show residual spatial autocorrelation, which is a desirable property to avoid inaccurate parameter estimates and inadequate quantification of uncertainty (Latimer et al., 2006; Beguin et al., 2012; Record et al., 2013; Crase et al., 2014). In addition, spatial HBMs exhibited lower average model prediction

errors than non-spatial HBMs. Hence, and from a purely statistical viewpoint, the higher rigour of spatial HBMs, in terms of higher accuracy and efficient removal of residual spatial autocorrelation, confers them a clear advantage (Swanson et al., 2013; Crase et al., 2014). Spatial HBMs also allowed the assessment of the spatial effects on distribution range, which were quite compact and with high intensities at the distribution range centre (Figure 9.3). Such patterns may account for the lower number of bioclimatic predictors selected by spatial HBMs in comparison with non-spatial HBMs and Maxent. As a matter of fact, the reduction of predictors represents a common shortcoming of spatial distribution models that, in some cases, may jeopardise the biological interpretation of the environmental factors underlying current distribution ranges (Beale et al., 2010; Swanson et al., 2013). We want to emphasise, however, that the five best spatial HBMs for each genetic cluster included additional variables compared to the best model, and all models were quite similar in terms of DIC, WAIC and LCPO values (Table 9.8). Thus, we have different options to identify environmental drivers of current distribution ranges. Furthermore, the reduction of predictors in spatial models may not reduce the models' interpretability.

9.4.2 Distribution range shifts with GCC

Taking spatial effects into account had a profound effect on the predictions of distribution range shifts due to GCC for *A. thaliana*'s genetic clusters in the Iberian Peninsula. In general, spatial HBMs exhibited more conservative patterns of change compared to Maxent and non-spatial HBMs (Figure 9.4). This result is in agreement with other research suggesting that environment-only models forecast substantially greater range shifts compared to models incorporating spatial effects (Swanson et al., 2013; Crase et al., 2014). The rationale is that organisms exhibiting a high spatial autocorrelation in the environmental drivers accounting for their distribution ranges will have larger areas with similar climates, which will also make GCC effects more predictable and homogeneous across space (Nadeau et al., 2017). For this reason, genetic clusters or species with continuous distributions not only increase the ease of modelling, but also facilitate the assessment of the spatial autocorrelation on distribution range shifts. Overall, there is no denying

that considering spatial autocorrelation adds realistic biological elements for understanding the long-term effects of GCC on biodiversity (De Marco Jr et al., 2008; Swanson et al., 2013; Crase et al., 2014; Cardador et al., 2014). Nevertheless, it must be noted that we are assuming that spatial autocorrelation and its underlying forces remain relatively constant during climate change. Clearly, this assumption, although beyond the scope of this study, will need to be addressed in the future.

In general, the outcomes generated by the three modelling approaches for GCC scenarios were quite different. Such a disparity in model outcomes may indicate the differential effects of environmental drivers and the sources of spatial autocorrelation on the GCC response of genetic clusters, but also the effect of geographic distribution of each genetic cluster on model performance. In fact, the problems affecting the modelling of current distribution ranges, namely the disjunct and scattered geographic distributions of genetic clusters C2 and C3, respectively, also affected the predictions of distribution range shifts with GCC. For example, in the case of genetic cluster C2, spatial HBM predicted a rather continuous distribution in NE Spain when it is more reasonable to expect that the barrier separating the two major nuclei of populations both sides of the Ebro river valley will be expanded with warming, as predicted by Maxent. Furthermore, the GCC effects on genetic cluster C3 are the tougher to predict. Although Maxent increased the potential distribution range of this genetic cluster over the Iberian Peninsula, as relict organisms exhibiting scattered distributions, the future of the C3 cluster may simply depend on the effect of GCC on the preservation of its habitats as they are today. In fact, populations of genetic cluster C3 may exhibit strong local adaptation (Méndez-Vigo et al., 2013), constraining the ability of relict genotypes to colonise novel habitats.

In contrast, interpreting the problems of the GCC effects on distribution range shifts for the other two genetic clusters with continuous distributions was totally different. For example, genetic cluster C1 has a continuous distribution across the NE Iberian Peninsula, which is characterised by Atlantic and continental climates. GCC models excluding spatial effects, i.e. Maxent and non-spatial HBMs, indicate that GCC will restrict *A. thaliana* to northern and mountainous areas, which is a typical scenario of migration towards environments that will probably retain similar characteristics in the future.

In contrast, spatial HBMs yielded a totally different outcome, indicating that genetic cluster C1 will maintain and even increase its current distribution range (Figure 9.4). The strong spatial effects detected by spatial HBMs for genetic cluster C1 may account for this result, as the response of *A. thaliana* to GCC is also expected to be more homogeneous. Recent experimental data from transplant experiments using accessions from this genetic cluster indicate that this scenario may be plausible, as *A. thaliana* performed well in warmer environments, highlighting the potential of this genetic cluster to cope with warming (Exposito-Alonso et al., 2018). The same applies to genetic cluster C4, which is also distributed continuously in the typically Mediterranean SW Iberian Peninsula. In this case, however, all modelling approaches predicted its expansions with GCC, although some discrepancies among modelling approaches were recorded (Figure 9.4). Although we lack experimental evidence of the effects of warming on performance of C4 A.thaliana accessions, we believe that such expansion with GCC is highly probable, as the genetic cluster C4 mostly occupies the warmest Iberian region.

9.4.3 Conclusions

We developed hierarchical Bayesian Beta regression models to explore the current distribution range and its GCC-induced shifts for an organism with a marked geographic genetic structure, which represents the outcome of historical, ecological and evolutionary forces probably acting in concert. For this reason, the effects of GCC have to be understood as a mosaic of responses varying in extent and intensity determined by the complexity of the geographic genetic structure exhibited by study organisms. Rather than predicting mere contractions or expansions for a single organism, we should expect a reshuffling of the genetic diversity and its geographic structure with GCC, which is obviously more difficult to predict. The HBMs developed here enrich the toolbox of software available to deal with such expectation.

From a statistical viewpoint, our HBMs allow the modelling of each genetic cluster avoids the binarization of genetic cluster membership proportions, required by Maxent, which may imply an important data loss.

This has the advantage that populations with high admixture can be included in HBMs. In addition, our HBMs can take the spatial autocorrelation of data into account, which not only improves the statistical properties of the model (i.e. removal of residual SAC) but also adds realism, as spatial autocorrelation may represent the result of eco-evolutionary processes shaping distribution ranges. Despite such desirable properties, our simulations of current and future distribution ranges of the four genetic clusters of Iberian *A. thaliana* indicated that the ease of modelling is strongly related to the continuity of their distributions. Furthermore, the biological knowledge of the study organism, namely, the identification of relict genetic lineages based on whole-genome sequencing, the detection of void areas after years of extensive field sampling, and the experimental quantification of plant performance with warming, emerges as an essential element in the understanding of model outputs. Finally, we believe that further work should also be conducted to validate model outputs by independent means (i.e. assignment of new *A. thaliana* populations to genetic clusters based on model predictions).

We conclude by stressing the importance of developing better models to forecast the effects of GCC on organisms' distribution range worldwide. Such predictive tools, and the comparison thereof, may lead to the mitigation of the inevitable impact of GCC on biodiversity. However, we have to keep increasing our comprehension of the evolutionary (e.g. physiological adaptive responses) and demographic (e.g. extinction/colonization dynamics and dispersal ability) factors accounting for the response of organisms to environmental changes imposed by GCC.

References

Anderson, C. D., Epperson, B. K., FORTIN, M.-J., Holderegger, R., James, P. M., Rosenberg, M. S., Scribner, K. T., and Spear, S. (2010). Considering spatial and temporal scale in landscape-genetic studies of gene flow. *Molecular Ecology*, 19(17):3565–3575.

Bálint, M., Domisch, S., Engelhardt, C., Haase, P., Lehrian, S., Sauer, J., Theissinger, K., Pauls, S., and Nowak, C. (2011). Cryptic biodiversity loss linked to global climate change. *Nature Climate Change*, 1(6):313–318.

Barton, N. H. and Hewitt, G. M. (1985). Analysis of hybrid zones. *Annual review of Ecology and Systematics*, 16(1):113–148.

Beale, C. M., Lennon, J. J., Yearsley, J. M., Brewer, M. J., and Elston, D. A. (2010). Regression analysis of spatial data. *Ecology letters*, 13(2):246–264.

Beguin, J., Martino, S., Rue, H., and Cumming, S. G. (2012). Hierarchical analysis of spatially autocorrelated ecological data using integrated nested Laplace approximation. *Methods in Ecology and Evolution*, 3(5):921–929.

Bellard, C., Bertelsmeier, C., Leadley, P., Thuiller, W., and Courchamp, F. (2012). Impacts of climate change on the future of biodiversity. *Ecology letters*, 15(4):365–377.

Belmaker, J., Zarnetske, P., Tuanmu, M.-N., Zonneveld, S., Record, S., Strecker, A., and Beaudrot, L. (2015). Empirical evidence for the scale dependence of biotic interactions. *Global Ecology and Biogeography*, 24(7):750–761.

Bivand, R., Hauke, J., and Kossowski, T. (2013). Computing the J acobian in G aussian Spatial Autoregressive Models: An Illustrated Comparison of Available Methods. *Geographical Analysis*, 45(2):150–179.

Bivand, R. and Piras, G. (2015). Comparing implementations of estimation methods for spatial econometrics. *Journal of Statistical Software*, 63:18.

Brennan, A. C., Méndez-Vigo, B., Haddioui, A., Martínez-Zapater, J. M., Picó, F. X., and Alonso-Blanco, C. (2014). The genetic structure of *Arabidopsis thaliana* in the south-western mediterranean range reveals a shared history between north africa and southern europe. *BMC plant biology*, 14(1):17.

Cardador, L., Sardà-Palomera, F., Carrete, M., and Mañosa, S. (2014). Incorporating spatial constraints in different periods of the annual cycle improves species distribution model performance for a highly mobile bird species. *Diversity and distributions*, 20(5):515–528.

Chen, I.-C., Hill, J. K., Ohlemüller, R., Roy, D. B., and Thomas, C. D. (2011). Rapid range shifts of species associated with high levels of climate warming. *Science*, 333(6045):1024–1026.

Crase, B., Liedloff, A., Vesk, P. A., Fukuda, Y., and Wintle, B. A. (2014). Incorporating spatial autocorrelation into species distribution models alters forecasts of climate-mediated range shifts. *Global Change Biology*, 20(8):2566–2579.

Cribari-Neto, F. and Zeileis, A. (2010). Beta regression in R. 34:1–24.

Dale, M. R. and Fortin, M.-J. (2002). Spatial autocorrelation and statistical tests in ecology. *Ecoscience*, 9(2):162–167.

Dale, M. R. and Fortin, M.-J. (2014). *Spatial analysis: a guide for ecologists*. Cambridge University Press.

De Marco Jr, P., Diniz-Filho, J. A. F., and Bini, L. M. (2008). Spatial analysis improves species distribution modelling during range expansion. *Biology letters*, 4(5):577–580.

de Oliveira, G., Rangel, T. F., Lima-Ribeiro, M. S., Terribile, L. C., and Diniz-Filho, J. A. F. (2014). Evaluating, partitioning, and mapping the spatial autocorrelation component in ecological niche modeling: a new approach based on environmentally equidistant records. *Ecography*, 37(7):637–647.

Diniz-Filho, J. A. F., Barbosa, A. C. O., Collevatti, R. G., Chaves, L. J., Terribile, L. C., Lima-Ribeiro, M. S., and Telles, M. P. (2016). Spatial autocorrelation analysis and ecological niche modelling allows inference of range dynamics driving the population genetic structure of a Neotropical savanna tree. *Journal of Biogeography*, 43(1):167–177.

Dirnböck, T. and Dullinger, S. (2004). Habitat distribution models, spatial autocorrelation, functional traits and dispersal capacity of alpine plant species. *Journal of Vegetation Science*, 15(1):77–84.

Dormann, C. F. (2007). Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global ecology and biogeography*, 16(2):129–138.

Durvasula, A., Fulgione, A., Gutaker, R. M., Alacakaptan, S. I., Flood, P. J., Neto, C., Tsuchimatsu, T., Burbano, H. A., Picó, F. X., Alonso-Blanco, C., et al. (2017). African genomes illuminate the early history and transition to selfing in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences*, 114(20):5213–5218.

D'Amen, M., Zimmermann, N. E., and Pearman, P. B. (2013). Conservation of phylogeographic lineages under climate change. *Global Ecology and Biogeography*, 22(1):93–104.

Elith, J. and 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.

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., and Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, 17(1):43–57.

Exposito-Alonso, M., Brennan, A. C., Alonso-Blanco, C., and Picó, F. X. (2018). Spatio-temporal variation in fitness responses to contrasting environments in *Arabidopsis thaliana*. *Evolution*, 72(8):1570–1586.

F. Dormann, C., M. McPherson, J., B. Araújo, M., Bivand, R., Bolliger, J., Carl, G., G. Davies, R., Hirzel, A., Jetz, W., Daniel Kissling, W., et al. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, 30(5):609–628.

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587.

Ferrari, S. L. P. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799 – 815.

Fortin, M.-J. and Dale, M. R. (2009). Spatial autocorrelation in ecological studies: a legacy of solutions and myths. *Geographical Analysis*, 41(4):392–397.

Geisser, S. (1993). *Predictive inference (1st ed.)*. London: Chapman and Hall.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gomaa, N. H., Montesinos-Navarro, A., Alonso-Blanco, C., and Pico, F. X. (2011). Temporal variation in genetic diversity and effective population size of Mediterranean and subalpine *Arabidopsis thaliana* populations. *Molecular ecology*, 20(17):3540–3554.

Gotelli, N. J. and Stanton-Geddes, J. (2015). Climate change, genetic markers and species distribution modelling. *Journal of Biogeography*, 42(9):1577–1585.

Guisan, A. and Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, 8(9):993–1009.

Guisan, A., Thuiller, W., and Zimmermann, N. E. (2017). *Habitat suitability and distribution models: with applications in R*. Cambridge University Press.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25(15):1965–1978.

Hijmans, R. J., Phillips, S., Leathwick, J., Elith, J., and Hijmans, M. R. J. (2017). Package 'dismo'. *Circles*, 9:1.

Hjelle, Ø. and Dæhlen, M. (2006). *Triangulations and applications*. Springer Science & Business Media.

Hooten, M. B. and Hobbs, N. T. (2015). A guide to Bayesian model selection for ecologists. *Ecological Monographs*, 85(1):3–28.

Ikeda, D. H., Max, T. L., Allan, G. J., Lau, M. K., Shuster, S. M., and Whitham, T. G. (2017). Genetically informed ecological niche models improve climate change predictions. *Global Change Biology*, 23(1):164–176.

Jakobsson, M. and Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, 23(14):1801–1806.

Jay, F., Manel, S., Alvarez, N., Durand, E. Y., Thuiller, W., Holderegger, R., Taberlet, P., and François, O. (2012). Forecasting changes in population genetic structure of alpine plants in response to global warming. *Molecular Ecology*, 21(10):2354–2368.

Latimer, A. M., Wu, S., Gelfand, A. E., and Silander, J. A. (2006). Building statistical models to analyze species distributions. *Ecological applications*, 16(1):33–50.

Legendre, P. (1993). Spatial autocorrelation: trouble or new paradigm? *Ecology*, 74(6):1659–1673.

Legendre, P. and Legendre, L. F. (2012). *Numerical ecology*. Toronto: Elsevier Science.

Lenoir, J. and Svenning, J.-C. (2015). Climate-related range shifts–a global multidimensional synthesis and new research directions. *Ecography*, 38(1):15–28.

Lima, J. S., Ballesteros-Mejia, L., Lima-Ribeiro, M. S., and Collevatti, R. G. (2017). Climatic changes can drive the loss of genetic diversity in a neotropical savanna tree species. *Global Change Biology*, 23(11):4639–4650.

Lindgren, F., Rue, H., et al. (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, 63(19):1–25.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.

MacLean, S. A. and Beissinger, S. R. (2017). Species' traits as predictors of range shifts under contemporary climate change: A review and meta-analysis. *Global Change Biology*, 23(10):4094–4105.

Manzano-Piedras, E., Marcer, A., Alonso-Blanco, C., and Picó, F. X. (2014). Deciphering the adjustment between environment and life history in annuals: lessons from a geographically-explicit approach in *Arabidopsis thaliana*. *PLoS One*, 9(2):e87836.

Marcer, A., Méndez-Vigo, B., Alonso-Blanco, C., and Picó, F. X. (2016). Tackling intraspecific genetic structure in distribution models better reflects species geographical range. *Ecology and Evolution*, 6(7):2084–2097.

Marcer, A., Vidigal, D. S., James, P., Fortin, M.-J., Méndez-Vigo, B., Hilhorst, H., Bentsink, L., Alonso-Blanco, C., and Picó, F. X. (2018). Temperature fine-tunes Mediterranean *Arabidopsis thaliana* life-cycle phenology geographically. *Plant Biology*, 20:148–156.

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*, 32(11):3227—-3244.

Méndez-Vigo, B., Gomaa, N. H., Alonso-Blanco, C., and Xavier Pico, F. (2013). Among-and within-population variation in flowering time of Iberian *Arabidopsis thaliana* estimated in field and glasshouse conditions. *New Phytologist*, 197(4):1332–1343.

Méndez-Vigo, B., Picó, F. X., Ramiro, M., Martínez-Zapater, J. M., and Alonso-Blanco, C. (2011). Altitudinal and climatic adaptation is mediated by flowering traits and FRI, FLC, and PHYC genes in *Arabidopsis*. *Plant Physiology*, 157(4):1942–1955.

Milanesi, P., Caniglia, R., Fabbri, E., Puopolo, F., Galaverni, M., and Holderegger, R. (2018). Combining Bayesian genetic clustering and ecological niche modeling: Insights into wolf intraspecific genetic structure. *Ecology and Evolution*, 8(22):11224–11234.

Nadeau, C. P., Urban, M. C., and Bridle, J. R. (2017). Climates past, present, and yet-to-come shape climate change vulnerabilities. *Trends in Ecology & Evolution*, 32(10):786–800.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M.,

and Bustamante, C. D. (2008). Genes mirror geography within Europe. *Nature*, 456(7218):98.

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

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

Parmesan, C. (2006). Ecological and evolutionary responses to recent climate change. *Review of Ecology, Evolution and Systematics*, 37:637–669.

Parmesan, C. and Yohe, G. (2003). A globally coherent fingerprint of climate change impacts across natural systems. *Nature*, 421:37–42.

Perry, A. L., Low, P. J., Ellis, J. R., and Reynolds, J. D. (2005). Climate change and distribution shifts in marine fishes. *science*, 308(5730):1912–1915.

Phillips, S. J., Anderson, R. P., and Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3):231–259.

Picó, F. X., Méndez-Vigo, B., Martínez-Zapater, J. M., and Alonso-Blanco, C. (2008). Natural genetic variation of arabidopsis thaliana is geographically structured in the iberian peninsula. *Genetics*, 180(2):1009–1021.

Pino, J., Font, X., Carbo, J., Jové, M., and Pallares, L. (2005). Large-scale correlates of alien plant invasion in Catalonia (NE of Spain). *Biological Conservation*, 122(2):339–350.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

Record, S., Fitzpatrick, M. C., Finley, A. O., Veloz, S., and Ellison, A. M. (2013). Should species distribution models account for spatial autocorrelation? a test of model projections across eight millennia of climate change. *Global Ecology and Biogeography*, 22(6):760–771.

Rosenberg, N. A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J. K., and Feldman, M. W. (2005). Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genetics*, 1(6):e70.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall.

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

Schmid, C. H. and Griffith, J. L. (2005). Multivariate classification rules: calibration and discrimination. *Encyclopedia of Biostatistics*.

Serra-Varela, M. J., Alía, R., Daniels, R. R., Zimmermann, N. E., Gonzalo-Jiménez, J., and Grivet, D. (2017). Assessing vulnerability of two mediterranean conifers to support genetic conservation management in the face of climate change. *Diversity and Distributions*, 23(5):507–516.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Swanson, A. K., Dobrowski, S. Z., Finley, A. O., Thorne, J. H., and Schwartz, M. K. (2013). Spatial regression methods capture prediction uncertainty in species distribution model projections through time. *Global Ecology and Biogeography*, 22(2):242–251.

The 1001 Genomes Consortium (2016). 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell*, 166(2):481–491.

Urban, M. C. (2015). Accelerating extinction risk from climate change. *Science*, 348(6234):571–573.

Václavík, T. and Meentemeyer, R. K. (2009). Invasive species distribution modeling (iSDM): Are absence data and dispersal constraints needed to predict actual distributions? *Ecological Modelling*, 220(23):3248–3258.

Van Vuuren, D. P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., Hibbard, K., Hurtt, G. C., Kram, T., Krey, V., Lamarque, J.-F., Masui, T., Meinshausen, M., Nakicenovic, N., Smith, S. J., and Rose, S. K. (2011). The representative concentration pathways: an overview. *Climatic Change*, 109:5.

Vidigal, D. S., Marques, A. C., Willems, L. A., Buijs, G., Méndez-Vigo, B., Hilhorst, H. W., Bentsink, L., Picó, F. X., and Alonso-Blanco, C. (2016). Altitudinal and climatic associations of seed dormancy and flowering traits evidence adaptation of annual life cycle timing in *Arabidopsis thaliana*. *Plant, Cell & Environment*, 39(8):1737–1748.

Wagner, H. H. and Fortin, M.-J. (2013). A conceptual framework for the spatial analysis of landscape genetic data. *Conservation Genetics*, 14(2):253–261.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec):3571–3594.

Wiens, J. J. (2016). Climate-related local extinctions are already widespread among plant and animal species. *PLoS Biology*, 14(12):e2001104.

Yannic, G., Pellissier, L., Ortego, J., Lecomte, N., Couturier, S., Cuyler, C., Dussault, C., Hundertmark, K. J., Irvine, R. J., Jenkins, D. A., et al. (2014). Genetic diversity in caribou linked to past and future climate change. *Nature Climate Change*, 4(2):132–137.

9.5 Supplemental Information

TABLE 9.4: The best five Maxent models for each genetic cluster according to five-fold cross-validated AUC. We provide the model formula, the mean area under the curve (AUC) and its standard deviation (SD). The best model among the best five according to parsimony is indicated. The number of occurrences after applying a threshold cut value of 0.5 was 103, 43, 38 and 35 for genetic cluster C1, C2, C3 and C4, respectively.

Cluster	**Model**	**AUC**	**SD**
C1	$Y \sim \beta_0 + BIO_1 + BIO_3 + BIO_4 + BIO_8 + BIO_{15} + BIO_{18}$	0.812	0.028
C1	$Y \sim \beta_0 + BIO_1 + BIO_4 + BIO_8 + BIO_{15} + BIO_{18}$	0.812	0.029
C1	$Y \sim \beta_0 + BIO_1 + BIO_3 + BIO_4 + BIO_8 + BIO_{12} + BIO_{15}$	0.811	0.029
C1	$Y \sim \beta_0 + BIO_1 + BIO_2 + BIO_4 + BIO_8 + BIO_{12} + BIO_{15}$	0.811	0.029
C1	$Y \sim \beta_0 + BIO_1 + BIO_3 + BIO_4 + BIO_8 + BIO_{15}$	0.811	0.029
C2	$Y \sim \beta_0 + BIO_1 + BIO_2 + BIO_4 + BIO_8 + BIO_{18}$	0.910	0.039
C2	$Y \sim \beta_0 + BIO_1 + BIO_2 + BIO_4 + BIO_8 + BIO_{12} + BIO_{18}$	0.909	0.039
C2	$Y \sim \beta_0 + BIO_1 + BIO_2 + BIO_3 + BIO_4 + BIO_8 + BIO_{18}$	0.908	0.040
C2	$Y \sim \beta_0 + BIO_2 + BIO_3 + BIO_4 + BIO_8 + BIO_{18}$	0.908	0.039
C2	$Y \sim \beta_0 + BIO_2 + BIO_4 + BIO_8 + BIO_{18}$	0.908	0.039
C3	$Y \sim \beta_0 + BIO_1 + BIO_2 + BIO_3 + BIO_8 + BIO_{15} + BIO_{18}$	0.809	0.073
C3	$Y \sim \beta_0 + BIO_1 + BIO_2 + BIO_3 + BIO_{15} + BIO_{18}$	0.808	0.073
C3	$Y \sim \beta_0 + BIO_1 + BIO_3 + BIO_8 + BIO_{15} + BIO_{18}$	0.808	0.073
C3	$Y \sim \beta_0 + BIO_1 + BIO_3 + BIO_{15} + BIO_{18}$	0.808	0.072
C3	$Y \sim \beta_0 + BIO_1 + BIO_2 + BIO_3 + BIO_8 + BIO_{12} + BIO_{15} + BIO_{18}$	0.808	0.073
C4	$Y \sim \beta_0 + BIO_1 + BIO_2 + BIO_4 + BIO_{12} + BIO_{18}$	0.864	0.047
C4	$Y \sim \beta_0 + BIO_1 + BIO_2 + BIO_4 + BIO_8 + BIO_{12}$	0.863	0.044
C4	$Y \sim \beta_0 + BIO_1 + BIO_2 + BIO_3 + BIO_8 + BIO_{12}$	0.863	0.049
C4	$Y \sim \beta_0 + BIO_1 + BIO_2 + BIO_4 + BIO_8 + BIO_{12} + BIO_{18}$	0.863	0.047
C4	$Y \sim \beta_0 + BIO_1 + BIO_2 + BIO_3 + BIO_4 + BIO_{12} + BIO_{18}$	0.863	0.048

TABLE 9.5: Results for the Moran's I test on residual spatial autocorrelation (SAC) for each modelling approach and genetic cluster. Models with P-value > 0.05 are considered as residual SAC free. Spatial HBM for C3 is indicated by dashes because it did not produce acceptable results.

Method	**Cluster**	**Moran's I**	**P-value**	**SAC free**
Maxent	C1	0.0747	0.0158	No
Maxent	C2	0.3553	0.0001	No
Maxent	C3	0.7953	0.0001	No
Maxent	C4	0.0544	0.0626	Yes
Non-spatial HBM	C1	0.0775	0.0001	No
Non-spatial HBM	C2	0.0795	0.0001	No
Non-spatial HBM	C3	0.0417	0.0094	No
Non-spatial HBM	C4	0.1021	0.0001	No
Spatial HBM	C1	0.0147	0.1011	Yes
Spatial HBM	C2	0.0141	0.0982	Yes
Spatial HBM	C3	–	–	–
Spatial HBM	C4	-0.0015	0.4148	Yes

TABLE 9.6: The best five non-spatial (A) and spatial HBMs (B) for each genetic cluster according to LCPO. We provide the model formula, the deviance information criterion (DIC), the Watanabe-Akaike information criterion (WAIC), and the logarithmic conditional predictive ordinates (LCPO). For each genetic cluster, the best model among the best five according to parsimony is indicated. For spatial models, the spatial effect term ($\boldsymbol{W}$) is also indicated in the formula. Spatial HBM for C3 is indicated by dashes because it did not produce acceptable results.

(A) Non-spatial HBMs

Cluster	Model	DIC	WAIC	LCPO
C1	$Y \sim \beta_0 + BIO_3 + BIO_4 + BIO_8 + BIO_{15} + BIO_{18}$	-212.81	-213.39	-0.354
C1	$Y \sim \beta_0 + BIO_3 + BIO_4 + BIO_8 + BIO_{12} + BIO_{15} + BIO_{18}$	-213.01	-213.38	-0.354
C1	$Y \sim \beta_0 + BIO_2 + BIO_4 + BIO_8 + BIO_{12} + BIO_{15} + BIO_{18}$	-212.73	-212.87	-0.353
C1	$Y \sim \beta_0 + BIO_2 + BIO_4 + BIO_8 + BIO_{15} + BIO_{18}$	-212.43	-212.80	-0.353
C1	$Y \sim \beta_0 + BIO_2 + BIO_3 + BIO_8 + BIO_{15} + BIO_{18}$	-212.02	-212.60	-0.353
C2	$Y \sim \beta_0 + BIO_1 + BIO_4 + BIO_8 + BIO_{12} + BIO_{15} + BIO_{18}$	-456.02	-454.53	-0.755
C2	$Y \sim \beta_0 + BIO_1 + BIO_4 + BIO_8 + BIO_{12} + BIO_{18}$	-454.12	-452.62	-0.752
C2	$Y \sim \beta_0 + BIO_1 + BIO_3 + BIO_4 + BIO_8 + BIO_{12} + BIO_{15} + BIO_{18}$	-454.01	-452.61	-0.752
C2	$Y \sim \beta_0 + BIO_1 + BIO_2 + BIO_4 + BIO_8 + BIO_{12} + BIO_{15} + BIO_{18}$	-454.04	-452.59	-0.752
C2	$Y \sim \beta_0 + BIO_1 + BIO_2 + BIO_3 + BIO_8 + BIO_{12} + BIO_{15} + BIO_{18}$	-453.03	-451.67	-0.750
C3	$Y \sim \beta_0 + BIO_4 + BIO_{12} + BIO_{18}$	-356.48	-356.84	-0.593
C3	$Y \sim \beta_0 + BIO_1 + BIO_4 + BIO_{12} + BIO_{18}$	-356.06	-356.72	-0.593
C3	$Y \sim \beta_0 + BIO_1 + BIO_2 + BIO_4 + BIO_{12} + BIO_{18}$	-355.14	-355.93	-0.591
C3	$Y \sim \beta_0 + BIO_1 + BIO_3 + BIO_{12} + BIO_{18}$	-355.27	-355.80	-0.591
C3	$Y \sim \beta_0 + BIO_1 + BIO_3 + BIO_4 + BIO_{12} + BIO_{18}$	-354.78	-355.56	-0.591
C4	$Y \sim \beta_0 + BIO_1 + BIO_{15} + BIO_{18}$	-379.14	-377.35	-0.627
C4	$Y \sim \beta_0 + BIO_1 + BIO_8 + BIO_{15} + BIO_{18}$	-378.19	-376.57	-0.625
C4	$Y \sim \beta_0 + BIO_1 + BIO_2 + BIO_3 + BIO_4 + BIO_{15} + BIO_{18}$	-378.23	-376.15	-0.625
C4	$Y \sim \beta_0 + BIO_1 + BIO_4 + BIO_{15} + BIO_{18}$	-377.62	-375.81	-0.624
C4	$Y \sim \beta_0 + BIO_1 + BIO_3 + BIO_{15} + BIO_{18}$	-377.42	-375.71	-0.624

(B) Spatial HBMs

Cluster	Model	DIC	WAIC	LCPO
C1	$Y \sim \beta_0 + BIO_1 + BIO_3 + BIO_8 + \boldsymbol{W}$	-262.07	-262.46	-0.434
C1	$Y \sim \beta_0 + BIO_1 + BIO_8 + \boldsymbol{W}$	-261.30	-261.71	-0.434
C1	$Y \sim \beta_0 + BIO_1 + BIO_2 + BIO_8 + \boldsymbol{W}$	-261.25	-261.52	-0.433
C1	$Y \sim \beta_0 + BIO_1 + BIO_3 + BIO_4 + BIO_8 + \boldsymbol{W}$	-261.09	-261.47	-0.433
C1	$Y \sim \beta_0 + BIO_1 + BIO_2 + BIO_3 + BIO_8 + \boldsymbol{W}$	-261.07	-261.32	-0.432
C2	$Y \sim \beta_0 + BIO_2 + BIO_8 + BIO_{12} + BIO_{15} + BIO_{18} + \boldsymbol{W}$	-616.66	-616.29	-1.014
C2	$Y \sim \beta_0 + BIO_8 + BIO_{12} + BIO_{15} + BIO_{18} + \boldsymbol{W}$	-616.07	-615.55	-1.012
C2	$Y \sim \beta_0 + BIO_1 + BIO_2 + BIO_8 + BIO_{12} + BIO_{15} + BIO_{18} + \boldsymbol{W}$	-615.45	-614.96	-1.011
C2	$Y \sim \beta_0 + BIO_4 + BIO_8 + BIO_{12} + BIO_{15} + BIO_{18} + \boldsymbol{W}$	-615.11	-614.83	-1.010
C2	$Y \sim \beta_0 + BIO_3 + BIO_8 + BIO_{12} + BIO_{15} + BIO_{18} + \boldsymbol{W}$	-615.62	-613.85	-1.010
C3	–	–	–	–
C3	–	–	–	–
C3	–	–	–	–
C3	–	–	–	–
C3	–	–	–	–
C4	$Y \sim \beta_0 + BIO_1 + \boldsymbol{W}$	-492.78	-475.08	-0.779
C4	$Y \sim \beta_0 + BIO_1 + BIO_2 + \boldsymbol{W}$	-493.30	-475.13	-0.778
C4	$Y \sim \beta_0 + BIO_1 + BIO_3 + \boldsymbol{W}$	-492.88	-474.65	-0.777
C4	$Y \sim \beta_0 + BIO_1 + BIO_4 + \boldsymbol{W}$	-492.30	-474.26	-0.777
C4	$Y \sim \beta_0 + BIO_1 + BIO_2 + BIO_{15} + \boldsymbol{W}$	-491.88	-474.07	-0.777

TABLE 9.7: Summary of posterior distributions for the best non-spatial (A) and spatial HBMs (B) for each genetic cluster according to the logarithmic conditional predictive ordinates (LCPO). The mean, standard deviation (SD), quantiles (0.025, 0.5 and 0.975) and the mode are given. Results of spatial HBM for C3 is not given as it did not produce acceptable results.

(A) Non-spatial HBMs

C1	**Mean**	**SD**	**0.025 q**	**0.5 q**	**0.975 q**	**Mode**
β_0	4.3073	1.8921	0.5817	4.3108	8.0106	4.3178
BIO_3	5.7731	2.8855	0.1521	5.7570	11.4781	5.7247
BIO_4	-0.5653	0.1232	-0.8077	-0.5652	-0.3240	-0.5648
BIO_8	-0.1040	0.0156	-0.1348	-0.1040	-0.0733	-0.1039
BIO_{15}	-0.0501	0.0060	-0.0620	-0.0501	-0.0383	-0.0500
BIO_{18}	-0.0091	0.0015	-0.0120	-0.0091	-0.0063	-0.0091
C2	**Mean**	**SD**	**0.025 q**	**0.5 q**	**0.975 q**	**Mode**
β_0	-3.6781	0.8329	-5.3233	-3.6750	-2.0516	-3.6687
BIO_1	-0.1117	0.0328	-0.1757	-0.1118	-0.0469	-0.1121
BIO_4	0.3729	0.0916	0.1944	0.3724	0.5540	0.3713
BIO_8	0.1042	0.0244	0.0563	0.1042	0.1519	0.1042
BIO_{12}	-0.0011	0.0003	-0.0018	-0.0011	-0.0004	-0.0011
BIO_{18}	0.0137	0.0019	0.0100	0.0136	0.0174	0.0136
C3	**Mean**	**SD**	**0.025 q**	**0.5 q**	**0.975 q**	**Mode**
β_0	-3.1345	0.8826	-4.8767	-3.1314	-1.4107	-3.1252
BIO_4	0.2875	0.1184	0.0568	0.2869	0.5213	0.2857
BIO_{12}	0.0012	0.0003	0.0006	0.0012	0.0018	0.0012
BIO_{18}	-0.0062	0.0014	-0.0089	-0.0062	-0.0035	-0.0061
C4	**Mean**	**SD**	**0.025 q**	**0.5 q**	**0.975 q**	**Mode**
β_0	-4.9265	0.4764	-5.8607	-4.9270	-3.9904	-4.9281
BIO_1	0.1972	0.0290	0.1405	0.1971	0.2543	0.1970
BIO_{15}	0.0212	0.0059	0.0096	0.0212	0.0327	0.0212
BIO_{18}	0.0036	0.0012	0.0012	0.0036	0.0059	0.0036

(B) Spatial HBMs

C1	**Mean**	**SD**	**0.025 q**	**0.5 q**	**0.975 q**	**Mode**
β_0	-2.2260	1.5085	-5.4395	-2.2065	0.9344	-2.1654
BIO_1	0.1472	0.0507	0.0501	0.1464	0.2490	0.1447
BIO_8	-0.0713	0.0317	-0.1343	-0.0710	-0.0100	-0.0704
C2	**Mean**	**SD**	**0.025 q**	**0.5 q**	**0.975 q**	**Mode**
β_0	-1.9800	1.0121	-4.1034	-1.9534	-0.0309	-1.9139
BIO_8	-0.0442	0.0209	-0.0854	-0.0441	-0.0033	-0.0440
BIO_{12}	-0.0015	0.0005	-0.0025	-0.0015	-0.0005	-0.0015
BIO_{15}	0.0255	0.0111	0.0039	0.0254	0.0474	0.0253
BIO_{18}	0.0095	0.0032	0.0033	0.0095	0.0160	0.0094
C4	**Mean**	**SD**	**0.025 q**	**0.5 q**	**0.975 q**	**Mode**
β_0	-3.4532	0.6466	-4.7086	-3.4608	-2.1490	-3.4711
BIO_1	0.1635	0.0378	0.0910	0.1630	0.2393	0.1618

TABLE 9.8: Mean posterior distribution for the hyper-parameter $\phi_i = \exp\theta$ for each of the best non-spatial and spatial HBMs for each genetic cluster (C1, C2, C3 and C4).

Model	**C1**	**C2**	**C3**	**C4**
Non-spatial HBMs	3.645	6.232	2.171	3.879
Spatial HBMs	4.651	13.228	–	6.891

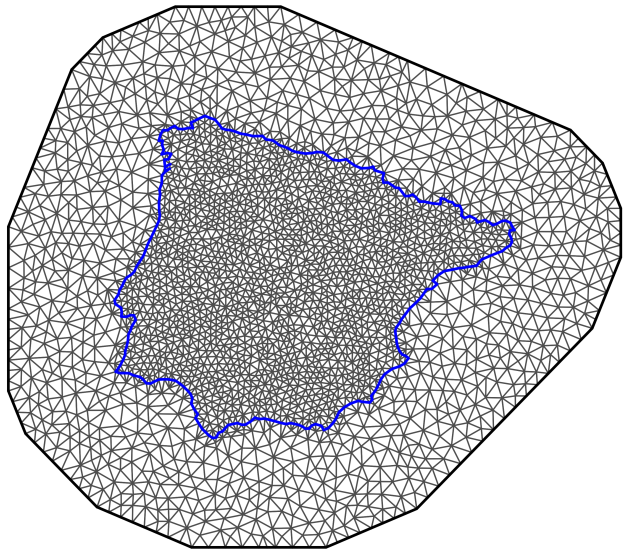

FIGURE 9.5: Delaunay triangulation used in HBMs to predict the response variable in un-sampled locations.

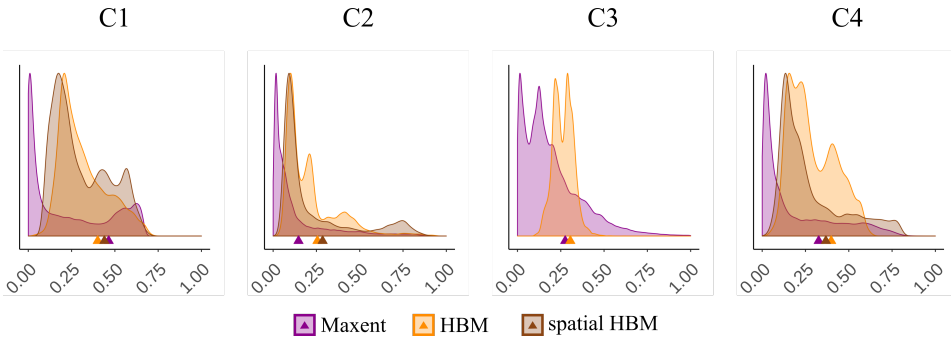

FIGURE 9.6: Density distributions of predicted genetic cluster membership probabilities for the whole of the study area for each modelling approach. Maxent densities must be interpreted as suitability for populations with a higher than 0.5 cluster coefficient. Small coloured triangles indicate the 0.75 percentile of the corresponding coloured distribution.

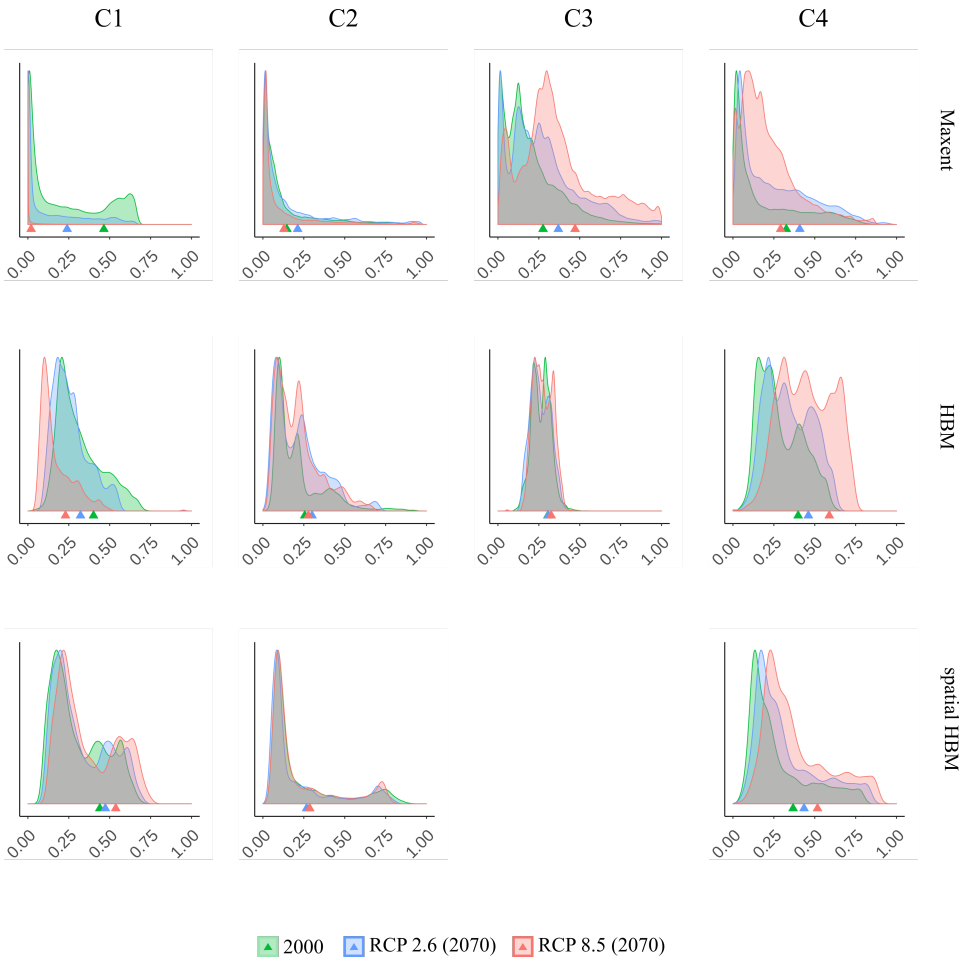

FIGURE 9.7: Density plots of predicted current (year 2000) and future distributions of suitability values across the Iberian Peninsula for each genetic cluster and modelling approach. Small coloured triangles indicate the 0.75 percentile of the corresponding coloured distribution.

Chapter 10

Dealing with physical barriers in bottlenose dolphin (*Tursiops truncatus*) distribution

In this chapter, we present the actual version of our paper "Dealing with physical barriers in bottlenose dolphin *Tursiops truncatus* distribution" by Joaquín Martínez-Minaya (University of Valencia), David Conesa (University of Valencia), Haakon Bakka (King Abdullah University of Science and Technology) and Maria Grazia Pennino (Spanish Institute of Oceanography), which has been accepted in the Journal *Ecological Modelling*. In order to keep the same structure of the chapters with published papers, this chapter ends with the references used in this work.

Abstract

Worldwide, cetacean species have started to be protected, but they are still very vulnerable to accidental damage from an expanding range of human activities at sea. To properly manage these potential threats we need a detailed understanding of the seasonal distributions of these highly mobile populations. To achieve this goal, a growing effort has been underway to develop species distribution models (SDMs) that correctly describe and predict preferred species areas. However, accuracy is not always easy to achieve

when physical barriers, such as islands, are present. Indeed, SDMs assume, if only implicitly, that the spatial effect is stationary, and that correlation is only dependent on the distance between observations and not on the direction or a spatial coordinates. The application of stationary SDMs in these cases could lead to incorrect predictions and, consequently, to uninformed decision making. In this study, we identify vulnerable habitats for the bottlenose dolphin in the Archipelago de La Maddalena, Northern Sardinia (Italy) using Bayesian hierarchical SDMs that account for the physical barriers issue and provide a full specification of the associated uncertainty. The approach we propose constitutes a major step forward in the understanding of cetacean species in many ecosystems where physical, geographical and topographical barriers are present.

Keywords

Archipelago de La Maddalena, cetaceans, hierarchical Bayesian spatial models, INLA, SPDE

10.1 Introduction

Globally, the importance of cetaceans as keystone and umbrella species is being increasingly recognized as protected areas designed on top predator distributions have been demonstrated to be highly efficient, leading to higher biodiversity levels and more ecosystem benefits (Sergio et al., 2008). However, cetacean populations have been facing various threats including depletion of resources, habitat loss, interactions with commercial fisheries, diseases produced by pollution and physical and acoustic disturbances caused by vessel traffic (Pennino et al., 2017).

Among cetaceans the bottlenose dolphin (*Tursiops truncatus*) is a vulnerable species Bearzi et al. (2012) that is more susceptible to anthropogenic activities due to its occurrence in coastal waters where most threats occur. This species is protected by the EU Habitats Directive 92/43/EEC and its

coastal ecotype is present in the ACCOBAMS (Agreement on the Conservation of Cetaceans in the Black Sea, Mediterranean Sea and contiguous Atlantic area) region Notarbartolo di Sciara (2002).

The protection of cetacean habitats, particularly those of bottlenose dolphins, should be a priority issue for marine conservation, given that protecting these areas constitutes an indirect measure toward global sea management (Pennino et al., 2016a). In order to achieve this goal, it is essential to have a solid understanding of the relationship that the species has with its habitat and apply robust analyses of existing information and databases to identify Special Areas of Conservation (SAC) (Pennino et al., 2016a). SACs should be designed around specific sensitive areas, where local bottlenose dolphins are known to have their centers of distribution (Gnone et al., 2011).

In this context, Species Distribution Models (SDMs) can be a useful tool to achieve these objectives given that they link spatial occurrence or species abundance data with multivariate environmental data that can estimate the relationship between the species and its habitat, and subsequently predict spatial occurrence or species abundance in un-sampled locations or time-periods Martínez-Minaya et al. (2018). Nonetheless, environmental conditions alone may not sufficiently explain species distribution as spatially intrinsic ecological processes, such as competition or predation, can also contribute.

Therefore, SDMs that incorporate spatial random effects to account for unexplained spatial dependence in data have been gaining increasing interest in marine ecology. Indeed, spatial random components could account for the spatial correlation driven by unmeasured covariates. SDMs usually assume, if only implicitly, that the spatial random effect is stationarity, i.e. that correlation is only dependent on the distance between the points, and not on the direction or the spatial coordinates Bakka et al. (2019). However, this assumption could be erroneous in areas where there are physical barriers such islands, thus leading to potentially biased predictions of species distributions.

Consequently, biased estimations and predictions of species distribution can lead to both uninformed decision making and inefficient management of natural resources Bakka et al. (2019). This is a fundamental issue in marine

ecology, where identification of vulnerable habitats (e.g., protected marine areas, nurseries, etc.) is one of the most common conservation management tools used to sustain the long-term viability of species populations.

In this paper we identify sensitive habitats for the bottlenose dolphin in the Archipelago de La Maddalena, Northern Sardinia (Italy) using a hierarchical Bayesian approach for spatial SDMs that account for physical barriers. As a tool to approximate the posterior distributions, we use the integrated nested Laplace approximation (INLA) Rue et al. (2009). The spatial effect that accounts for the physical barriers is included and measured by the approximation to a system of stochastic partial differential equations proposed by Bakka et al. (2019).

The Maddalena Archipelago is included within the Pelagos Cetacean Sanctuary, which is the only pelagic Marine Protected Area (MPA) for marine animals in the Mediterranean Sea. The bottlenose dolphin is one of the most common cetacean species in this area (Notarbartolo di Sciara, 2002), with a population of 71 photo-identified individuals (Pennino et al., 2013), of which 22 have been defined as residents (individuals sighted in all seasons during that one year and at least five times).

In line with all these, an improved understanding of the spatial distribution of the bottlenose dolphin in this area could contribute to management of this vulnerable species.

10.2 Materials and methods

10.2.1 Study area

This study was conducted in waters within 3 miles off the coast of Archipelago de La Maddalena, Northern Sardinia (Italy). This area is within a National Park located in the strait of Bonifacio, between the islands of Sardinia and Corsica, and is part of the Pelagos Cetacean Sanctuary that was established by Italy, France and Monaco in 1999.

The type of seabed of the inner shelf (from 0 to 70 m in depth) is mainly composed of rocky or sandy substrata covered with Posidonia seagrass (*Posidonia oceanica*, Delile, 1813) beds. A high hydrodynamism characterizes this area that, together with shallow depth and limited tidal range, generate very clean waters. The general aspect of the coast is indented, with small promontories, bays and narrow channels. The topography of the bottom is variable with large cracks, reefs and small islands.

10.2.2 Field and Study Methods

In order to equally monitor the area, random transects were performed from October 2007 to September 2008 on-board a zodiac boat with a speed of 8-10 knots. Surveys were conducted by experts during light hours from 6.00 A.M. to 8.00 P.M. To identify species, observers scanned with both the naked eye and binoculars (7 x 50 and 8 x 42). To ensure the same visibility across the study area, surveys were only performed when the sea-state was less than 3 (Douglas sea force scale) and in clear conditions with no precipitation. Data collected included sighting occurrence, date and geographical location. Geographical information were collected every minute using a GPS, logged to a computer equipped with "Mapsource" software (Garmin GPS device, 2010).

To avoid harassing the dolphins, sightings were performed from a respectful distance (no closer than 30 meters), with binoculars or telephoto lenses to get a good view of the animals. If the dolphins approached the boat, the course was maintained to avoid sudden changes in direction or speed that could injure the animals.

10.2.3 Environmental variables

Bottlenose dolphin distribution was modeled using five environmental variables selected for being known to affect their habitat: three oceanographic variables —Sea Surface Temperature (SST in C), Sea Surface Salinity (SSS in PSU) and Chlorophyll-a concentration (CHL in mg/m^3)— and two topographic covariates —depth (in meters) and slope (in degrees)—. SST, SSS

and CHL are strongly related to marine system productivity as they can affect nutrient availability, metabolic rates and water stratification. All these variables were derived from the aqua-MODIS sensor, as monthly values with a resolution of 2 km (https://modis.gsfc.nasa.gov/).

With respect to the importance of these selected variables, it is worth noting that these topographic covariates have frequently been used as predictors of cetacean species distribution (Panigada et al., 2008; Mannocci et al., 2014). Also, bathymetry-derived terrain variables, such as the slope of the seabed, are indicative of seabed morphology and have been widely used as predictors of cetacean distribution (Lauria et al., 2015; Fonseca et al., 2017; Pennino et al., 2016a). Usually low slope values correspond to a flat ocean bottom (areas of sediment deposition), while higher values indicate consolidate substrata (i.e., rocky substrate) Fonseca et al. (2017).

Bathymetric variables were derived from the MARSPEC database (available at http://www.marspec.org) with a spatial resolution of 1 km (Sbrocco and Barber, 2013). To maintain the same spatial resolution, all environmental data were gridded at 2 km using the raster package (Hijmans and van Etten, 2015) in the R software (R Core Team, 2018).

Collinearity between explanatory environmental variables was checked using a Draftsman's plot and the Pearson correlation index. The variables were not highly correlated ($r = 0.6$), and thus were considered in further analyses. Finally, all explanatory variables were centered and standardized following the approach of Gelman (2008).

10.2.4 Statistical model

The recently published hierarchical Bayesian spatial model that accounts for barriers (Bakka et al., 2019) was used to estimate and predict overall occurrence of bottlenose dolphins with respect to environmental predictors. Given that our data are composed of the presences and absences of bottlenose dolphins, the response variable Y_i can be assumed to follow a Bernoulli distribution with a mean of π_i that can take on values of 1 or 0 depending on whether the habitat is suitable ($Y_i = 1$) or not ($Y_i = 0$) for

the species. As usual in Generalized Linear Models, each π_i can be easily linked to a structured additive predictor η_i through a link function $g(\cdot)$, so that $g(\boldsymbol{\pi}) = \boldsymbol{\eta}$. The structured additive predictor $\boldsymbol{\eta}$ accounts for the effect of various covariates and the spatial effect in an additive way:

$$\eta_i \quad = \quad \beta_0 + \sum_{m=1}^{M} \beta_m x_{mi} + u(s_i) \ , \tag{10.1}$$

where β_0 corresponds to the intercept; the coefficients $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_M\}$ quantify the effect of the possible factors and covariates $\boldsymbol{x} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_M})$ on the response; and $u(s_i)$ denotes the spatial random effect.

SDMs usually assume stationarity in the spatial random effect $u(s_i)$. In other words, the spatial autocorrelation only depends on the distance between points (see, for instance Pennino et al., 2013; Paradinas et al., 2015; Rufener et al., 2017). Nevertheless, if there are physical barriers, such as islands, this assumption can be erroneous, thus prompting biased predictions. As a consequence, we suppose that the spatial random effect in the model depends also on the direction and the geographic coordinates, i.e., $u(\boldsymbol{s})$ is a non-stationary spatial random effect. Using the approximation of Bakka et al. (2019), it can be estimated as the continuous weak solution to the following system of stochastic differential equations:

$$\begin{aligned} u(s) - \nabla \cdot \frac{r^2}{8} \nabla u(s) &= r\sqrt{\frac{\pi}{2}}\sigma_u \mathcal{W}(s), \text{ for } s \in \Omega_n \, , \\ u(s) - \nabla \cdot \frac{r_b^2}{8} \nabla u(s) &= r_b\sqrt{\frac{\pi}{2}}\sigma_u \mathcal{W}(s), \text{ for } s \in \Omega_b \, , \end{aligned} \tag{10.2}$$

Where r and r_b are the ranges for the normal and barrier areas respectively, σ_u is the marginal standard deviation of u, $\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right)$, $\mathcal{W}(s)$ denotes the white noise, Ω_n is the normal terrain, Ω_b is the barrier, and their disjoint union gives the whole study area.

Unlike stationary spatial effects, the underlying idea is to construct a Gaussian Markov Random Field (GMRF) locally, with one governing equation for the normal area (sea), and another for the barrier area (earth). The prior spatial effect only depends on two unknown hyperparameters, the standard deviation (σ_u) and the range in the normal area (r), because the

range in the barrier area (r_b) is fixed at close to zero. As a result, the system in (10.2) represents a local averaging of nearby values. If there are two points separated by a landmass, the very small range stops the local averaging on the barrier. It forces the dependency to focus on moving around the barrier, via local averages in the water area. The system of differential equations in (10.2) can be solved by constructing a Delaunay triangulation of the study area (Figure 10.1) and then applying the finite element method as explained in Bakka et al. (2019).

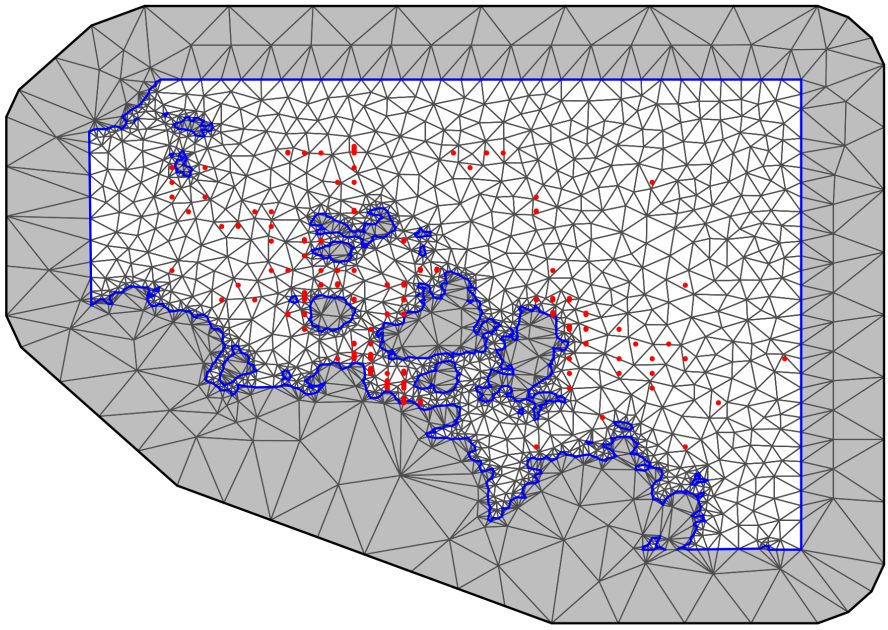

FIGURE 10.1: Map of the study area with sightings locations (red dots). Triangulation used to calculate the GMRF for the SPDE approach.

In addition to the environmental variables and the spatial effect, a factor representing the actual season was included in the model to account for temporal variability. Default priors were assigned for all fixed-effect parameters, which are approximations of non-informative priors designed to have little influence on the posterior distribution. PC priors Simpson et al. (2017) that followed the parametrization depicted in Fuglstad et al. (2018) were allocated for the only two hyperparameters in the model and define the

covariance structure of $u(s)$: σ_u and r. We set the median of the prior range to 0.3 (the extent of the area in geographic coordinates) and the median for the marginal standard deviation to 1.

10.2.5 Bayesian inference with INLA

All the models were fitted, i.e. posteriors of the model parameters computed, using the Integrated Nested Laplace Approximation (INLA) methodology (Rue et al., 2009), implemented in the R package INLA (https://www.r-inla.org). In INLA, the Bernoulli likelihood is approximated by a Laplace approximation, and the posterior for all parameters, conditionally on the two hyperparameters σ, r for the spatial field, can be computed quickly by sparse matrix algorithms. The posterior for the hyperparameters are found by exploring this two dimensional space, and is fast due to its low dimension. After representative values of the hyper-parameters have been chosen, these are integrated out to give a full posterior distribution for all the parameters and the spatial effect in the model.

Bayesian posterior distributions, unlike the mean and confidence intervals produced by classical analyses, enable simple probability statements about the unknown parameters. Thus, the region bounded by the 0.025 and 0.975 quantiles of the posterior distribution has an intuitive interpretation: for a specific model, the unknown parameter has a 95% chance of falling within this range of values.

As the interest was to analyze the probability of finding a dolphin in all the study area, a grid of prediction locations were included in the model fitting. At each grid, the posterior predictive distribution of the probability of observing the dolphin was obtained.

10.2.6 Model selection

Model selection was conducted based on choosing the best subset of covariates (see, for instance, Heinze et al. (2018) for a detailed revision of model selection procedures). This method evaluates all 2^k (k is the number

of components of the model: covariates and random effects, such as the spatial effect) possible models and choose the best model according to an information criterion, in our case, the Watanabe Akaike Information Criterion (WAIC) (Watanabe, 2010) and the mean logarithmic of the approximated Conditional Predictive Ordinate (LCPO) (Gneiting and Raftery, 2007). While WAIC values indicate the goodness of fit of the models, the LCPO evaluates the predictive capacity. Lower values for both WAIC and LCPO represent the best compromise between fit and parsimony. If the models are similar in terms of WAIC and LCPO, following the parsimony criterion, the model with less amount of covariates is selected.

10.3 Results

Between October 2007 and September 2008, bottlenose dolphins were sighted in 93 of the 206 surveys of the study area. More specifically, 34 sightings occurred in winter, 29 in spring, 8 in summer and 22 in autumn. The total sighting rate was about 0.45, 0.97 for the winter season, 0.32 in spring, 0.19 in summer and 0.57 in autumn (Table 10.1). Due mainly to atmospherically reasons the survey effort was not homogeneous in all the seasons, recording is maximum during the spring and summer period. Nevertheless almost 20% of the effort was distributed in every season, except in spring when the 44% of the effort took place (Table 10.1).

Table 10.1: Numerical summary of the survey effort and sighting rate by season.

Season	Sightings	N. surveys	Sighting rate (%)	Seasonal effort (%)
Winter	34	35	97.14	16.99
Spring	29	91	31.86	44.17
Summer	8	42	19.04	20.39
Autumn	22	38	57.89	18.45
Total	93	206	45.14	100

Regarding the hierarchical Bayesian SDMs, in addition to the five environmental variables, the season factor and the non-stationary spatial effect

were considered to select the best model. A total of 128 models were fitted. Table 10.2 displays the best 20 models and their WAIC and LCPO ordered by LCPO. As noticed, the presented 20 models were very similar in terms of WAIC and LCPO, and so these models can be considered equivalent. Thus, the parsimony criterion was employed in order to select the best model among those having equivalent values of WAIC and LCPO. The final selected model was the one with only one covariate, the seasonal effect.

After selecting this model we also investigated the importance of the covariates not selected. In particular, Bayesian estimation of the regression coefficients associated to the covariates not selected was negligible, in the sense that all the posterior distributions of the regression parameters were centered around zero and with variances smaller that the ones provided in the priors. This was a clear proof that those covariates should not be part of the final selected model.

Results in Table 10.3 showed that winter is the season with the highest estimated dolphin occurrence (posterior mean = 4.46; 95% CI = [2.32, 7.25]) with respect to the reference level (autumn season). Conversely, summer and spring seasons show lower estimated dolphin occurrence than the reference level (respectively, posterior mean = -2.37; 95% CI = [-3.71, -1.18] and posterior mean = -0.79; 95% CI = [-1.74, 0.14]).

The median for the posterior predictive distribution of the probability of occurrence showed higher values in the whole area during the winter season (Figure 10.2d). Conversely, in autumn and spring, a higher probability of occurrence (close to 1 in line with the high sighting rate observed) was found in the Northwest area (Figures 10.2a and 10.2c). Similarly, in summer, the most frequented area was the Northwest, but with probabilities of presence close to 0.5 (Figure 10.2b).

The spatial effect that indicates the intrinsic variability of the distribution of bottlenose dolphins after excluding environmental variables was consistent with the probability maps (Figure 10.3). Moreover, the mean of the range of the spatial effect of the normal area was about 0.157 geographical degrees, that are equivalent to 17.48 km. The physical meaning of this value is that sightings of dolphins that are this distance or greater apart are not spatially correlated.

TABLE 10.2: Model comparison. The acronyms are: Seasonal factor (S), Sea Surface Temperature (SST in C), Sea Surface Salinity (SSS in PSU) and Chlorophyll-a concentration (CHL in mg/m-3), two topographic covariates - depth (in meters) and slope (in degrees) and the non-stationary spatial effect (u). Models are ordered by LCPO.

	Models	WAIC	LCPO
1	1 + S + SST + u	185.33	0.453
2	1 + S + SSS + SST + u	185.57	0.454
3	**1 + S +** u	**186.43**	**0.455**
4	1 + S + SSS + u	186.37	0.456
5	1 + S + CHL + SSS + u	185.71	0.456
6	1 + S + CHL + SSS + SST + u	185.11	0.456
7	1 + S + SST + slope + u	186.57	0.456
8	1 + S + SSS + SST + slope + u	185.68	0.456
9	1 + S + SST + depth + u	186.59	0.456
10	1 + S + SSS + SST + depth + u	186.49	0.457
11	1 + S + CHL + SSS + slope + u	185.97	0.458
12	1 + S + SSS + slope + u	186.87	0.458
13	1 + S + CHL + SSS + SST + slope + u	184.52	0.459
14	1 + S + CHL + u	185.75	0.459
15	1 + S + slope + u	187.86	0.459
16	1 + S + CHL + SSS + SST + depth + u	186.05	0.459
17	1 + S + SSS + depth + u	187.91	0.459
18	1 + S + CHL + SSS + depth + u	187.13	0.459
19	1 + S + SSS + SST + depth + slope + u	187.28	0.460
20	1 + S + SST + depth + slope + u	188.04	0.460

10.4 Discussion

Seasonal sensitive habitats for the bottlenose dolphin in the Archipelago de La Maddalena were identified using hierarchical Bayesian SDMs that account for physical barriers. The proposed model showed that dolphin occurrence in the Archipelago de La Maddalena is influenced by a seasonal effect in the area. Our findings agree with those obtained by Brotons et al. (2008)

TABLE 10.3: Mean, standard deviation, quantiles and mode for the parameters and hyperparameters of the best model. Summer, Spring and Winter are the three levels of the factor Season (the remaining one being the reference level Autumn). $\sigma_{\boldsymbol{u}}$ represents the standard deviation of the spatial effect and r the range of the normal (non-barrier) area.

Parameters	mean	sd	$Q_{0.025}$	$Q_{0.5}$	$Q_{0.975}$	mode
Intercept	0.455	2.237	-3.942	0.424	5.049	0.397
Summer	-2.375	0.643	-3.708	-2.351	-1.182	-2.304
Spring	-0.794	0.480	-1.744	-0.792	0.141	-0.788
Winter	4.460	1.263	2.315	4.342	7.253	4.098
Hyperparameters	mean	sd	$Q_{0.025}$	$Q_{0.5}$	$Q_{0.975}$	mode
$\sigma_{\boldsymbol{u}}$	2.254	1.408	1.165	2.242	4.470	2.199
r	0.157	1.624	0.065	0.152	0.434	0.137

in the Balearic Islands, Campana et al. (2015) in the Western Mediterranean Sea, and Pennino et al. (2015) and (Pennino et al., 2016a) in our study area. Indeed, estimated dolphin occurrence is higher during the winter season and especially compared to spring and summer. Several possible reasons, either isolated or combined, could explain this seasonal variation. Natural seasonal movement of dolphins could be related to prey availability or reproduction patterns. Moreover, the intense nautical traffic in summer that characterizes this area could encourage these animals to move to areas where there are fewer pleasure boats and where the risk of collision and the noise is lower (Pennino et al., 2016b).

Another important factor driving dolphin occurrence is the spatial component, which is highest in the western zone. In this area, bottlenose dolphins show a residential attitude with their center of distribution in the identified favourable areas. The spatial effect usually captures the impact of important missing predictors and accounts for ecological processes (e.g., predation or competition) that may affect the spatial arrangement of a species (Roos et al., 2015). In our case, the spatial effect was not directly related to any environmental variable included in the final model but it could be reflecting disturbance from pleasure boating.

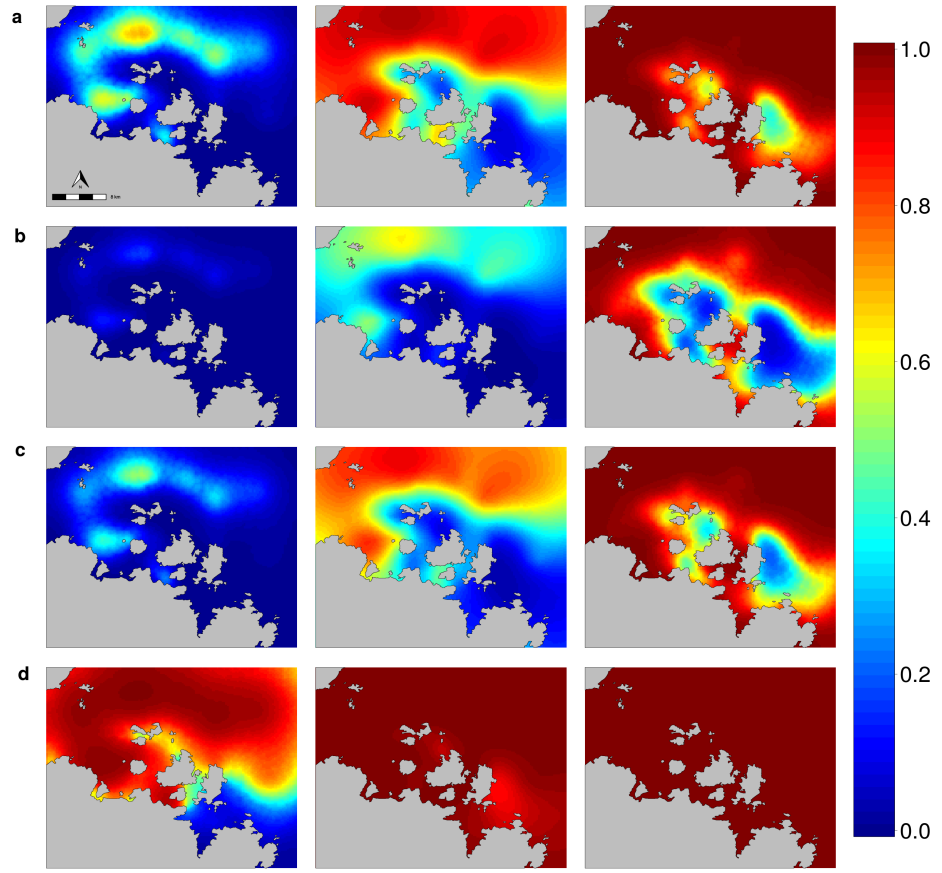

FIGURE 10.2: Posterior predictive distribution of the probability of presence: 95% credible intervals (First and third panel respectively) and the median (central panel) for the different seasons. a: autumn, b: summer, c: spring and d: winter.

An effective conservation programme should take into account these findings: favourable areas for bottlenose dolphins should be identified and protected as SACs (Special Areas of Conservation). Indeed, bottlenose dolphins are listed in Annex II of the Habitats Directive that specifically requires the identification of the SACs (Cañadas et al., 2005). SACs should be designed around special sensitive areas, such as the ones identified in this study. Protection measures should be devoted to limiting the disturbance from recreational boats, which is probably the main threat for this species in the area.

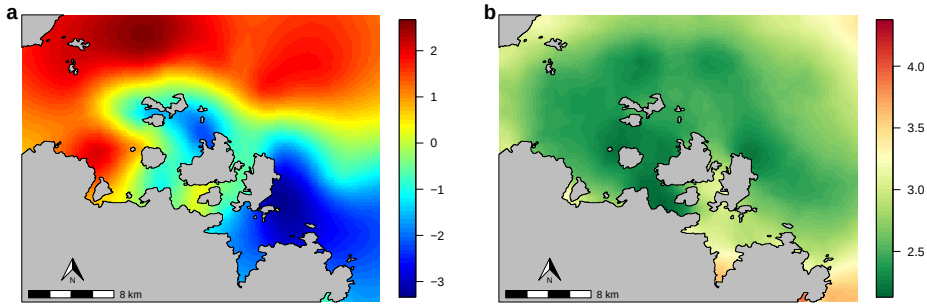

FIGURE 10.3: Mean and standard deviation for posterior distribution of the spatial effect u.

Spatial ecology has a direct applied relevance to cetaceans management, but it also has a broad ecological significance. Although it may be complicated to define the boundaries of habitats of these highly mobile species, it represents the first step towards facilitating effective spatial management. However, using a non-accurate approach could culminate in misidentification in both the posterior distributions of the fixed and random effects and in species habitat predictions, therefore leading to inappropriate management measures that can sometimes be irreversible Bakka et al. (2019).

In line with this, we have used here a hierarchical Bayesian spatial model that simultaneously deals with the presence of physical, geographical and topographical barriers, spatial autocorrelation issues and different sources of uncertainties. Our modeling is based on the novel approach by Bakka et al. Bakka et al. (2019), and allows us to analyze sparsely binary spatial data. Some advantages result from using our proposal. The first is a result of the Bayesian methodology itself, that is, that all multiple sources of uncertainty associated with both the observed data and ecological process can be included in the analysis, thus resulting in more robust statistical inference. Moreover, the posterior predictive distribution of the probability of finding the species turns out to be a very suitable tool that allows us to express our uncertainties associated with the entire species habitat prediction phenomenon and to explicitly describe the associated spatio-temporal

variability. The second advantage is that the proposal provides an accuracy that would not be easy to achieve when physical barriers are present. The application of stationary models in these cases could lead to uncertain predictions, and consequently to uninformed decision making. The third advantage is that we can present a map of the spatial effect along with its corresponding uncertainty. The final advantage is the computational gain from the use of the INLA approach, which allows us to easily make inferences and predictions within a highly structured model.

Finally, regarding the database used in this study, it worth to be mentioned that it has some flaws, especially due to the non-standardized sampling effort and limited field quantitative information (i.e. total and seasonal nautical mileage traveled). This can probably have affected the sighting rate per season, and so, the resulting predictive maps. Nevertheless, it is well known that collecting data at sea presents many logistic and financial challenges in particular due to find suitable seagoing vessels for data collection and atmospherically and oceanographic reasons. However, determining cetacean distribution is essential for proposing conservation policies and any advance in this sense is an improvement of the management and conservation of their populations. In conclusion, this approach constitutes a major step forward in the understanding of species in many aquatic ecosystems where physical, geographical and topographical barriers are present.

References

Bakka, H., Vanhatalo, J., B.Illian, J., Simpson, D., and Rue, H. (2019). Non-stationary Gaussian models with physical barriers. *Spatial Statistics*, 29:268 – 288.

Bearzi, G., Fortuna, C., and Reeves, R. (2012). Tursiops truncatus (Mediterranean subpopulation). The IUCN Red List of Threatened Species. version 2014.3.

Brotons, J. M., Grau, A. M., and Rendell, L. (2008). Estimating the impact of interactions between bottlenose dolphins and artisanal fisheries around the balearic islands. *Marine Mammal Science*, 24(1):112–127.

Campana, I., Crosti, R., Angeletti, D., Carosso, L., David, L., Di-Méglio, N., Moulins, A., Rosso, M., Tepsich, P., and Arcangeli, A. (2015). Cetacean response to summer maritime traffic in the Western Mediterranean Sea. *Marine environmental research*, 109:1–8.

Cañadas, A., Sagarminaga, R., De Stephanis, R., Urquiola, E., and Hammond, P. (2005). Habitat preference modelling as a conservation tool: proposals for marine protected areas for cetaceans in southern Spanish waters. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 15(5):495–521.

Fonseca, V. P., Pennino, M. G., de Nóbrega, M. F., Oliveira, J. E. L., and de Figueiredo Mendes, L. (2017). Identifying fish diversity hot-spots in data-poor situations. *Marine environmental research*, 129:365–373.

Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2018). Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association*, pages 1–8.

Garmin GPS device (2010). MapSource 6.16.3.

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in medicine*, 27(15):2865–2873.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gnone, G., Bellingeri, M., Dhermain, F., Dupraz, F., Nuti, S., Bedocchi, D., Moulins, A., Rosso, M., Alessi, J., McCrea, R. S., et al. (2011). Distribution, abundance, and movements of the bottlenose dolphin (Tursiops truncatus) in the Pelagos Sanctuary MPA (north-west Mediterranean Sea). *Aquatic Conservation: Marine and Freshwater Ecosystems*, 21(4):372–388.

Heinze, G., Wallisch, C., and Dunkler, D. (2018). Variable selection–a review and recommendations for the practicing statistician. *Biometrical Journal*, 60(3):431–449.

Hijmans, R. and van Etten, J. (2015). raster: Geographic data analysis and modeling. R package version 2.1-49. 2013.

Lauria, V., Power, A. M., Lordan, C., Weetman, A., and Johnson, M. P. (2015). Spatial transferability of habitat suitability models of Nephrops norvegicus among fished areas in the Northeast Atlantic: sufficiently stable for marine resource conservation? *PloS one*, 10(2):e0117006.

Mannocci, L., Catalogna, M., Dorémus, G., Laran, S., Lehodey, P., Massart, W., Monestiez, P., Van Canneyt, O., Watremez, P., and Ridoux, V. (2014). Predicting cetacean and seabird habitats across a productivity gradient in the South Pacific gyre. *Progress in Oceanography*, 120:383–398.

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*, 32(11):3227—-3244.

Notarbartolo di Sciara, G. (2002). Cetacean species occurring in the Mediterranean and Black Seas. In Notarbartolo di Sciara, G., editor, *Cetaceans of the Mediterranean and Black Seas: State of Knowledge and Conservation Strategies. A Report to the ACCOBAMS Secretariat.* ACCOBAMS.

Panigada, S., Zanardelli, M., MacKenzie, M., Donovan, C., Mélin, F., and Hammond, P. S. (2008). Modelling habitat preferences for fin whales and striped dolphins in the Pelagos Sanctuary (Western Mediterranean Sea) with physiographic and remote sensing variables. *Remote Sensing of Environment*, 112(8):3400–3412.

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

Pennino, M. G., Arcangeli, A., Fonseca, V. P., Campana, I., Pierce, G. J., Rotta, A., and Bellido, J. M. (2017). A spatially explicit risk assessment approach: Cetaceans and marine traffic in the Pelagos Sanctuary (Mediterranean Sea). *PloS one*, 12(6):e0179686.

Pennino, M. G., Mendoza, M., Pira, A., Floris, A., and Rotta, A. (2013). Assessing foraging tradition in wild bottlenose dolphins (Tursiops truncatus). *Aquatic Mammals*, 39(3):282.

Pennino, M. G., Mérigot, B., Fonseca, V. P., Monni, V., and Rotta, A. (2016a). Habitat modeling for cetacean management: Spatial distribution in the southern Pelagos Sanctuary (Mediterranean Sea). *Deep Sea Research Part II: Topical Studies in Oceanography*.

Pennino, M. G., Roda, M. A. P., Pierce, G. J., and Rotta, A. (2016b). Effects of vessel traffic on relative abundance and behaviour of cetaceans: the case of the bottlenose dolphins in the Archipelago de La Maddalena, north-western Mediterranean sea. *Hydrobiologia*, 776(1):237–248.

Pennino, M. G., Rotta, A., Pierce, G. J., and Bellido, J. M. (2015). Interaction between bottlenose dolphin (Tursiops truncatus) and trammel nets in the Archipelago de La Maddalena, Italy. *Hydrobiologia*, 747(1):69–82.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Roos, N. C., Carvalho, A. R., Lopes, P. F., and Pennino, M. G. (2015). Modeling sensitive parrotfish (Labridae: Scarini) habitats along the Brazilian coast. *Marine environmental research*, 110:92–100.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392.

Rufener, M.-C., Kinas, P. G., Nóbrega, M. F., and Oliveira, J. E. L. (2017). Bayesian spatial predictive models for data-poor fisheries. *Ecological Modelling*, 348:125–134.

Sbrocco, E. J. and Barber, P. H. (2013). MARSPEC: ocean climate layers for marine spatial ecology. *Ecology*, 94(4):979–979.

Sergio, F., Caro, T., Brown, D., Clucas, B., Hunter, J., Ketchum, J., McHugh, K., and Hiraldo, F. (2008). Top predators as conservation tools:

ecological rationale, assumptions, and efficacy. *Annual review of ecology, evolution, and systematics*, 39:1–19.

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

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec):3571–3594.

Chapter 11

Species distribution modeling: a statistical review with focus in spatio-temporal issues

In this chapter, we present a version of our paper "Species distribution modeling: a statistical review with focus in spatio-temporal issues" by Joaquín Martínez-Minaya (University of Valencia), Michela Cameletti, David Conesa (University of Valencia) and Maria Grazia Pennino (Spanish Institute of Oceanography) which has been published in *Stochastic Environmental Research and Risk Assessment*, 32, 3227–3244. The chapter contains at the end the references used in this work.

Abstract

The use of complex statistical models has recently increased substantially in the context of species distribution behavior. This complexity has made the inferential and predictive processes challenging to perform. The Bayesian approach has become a good option to deal with these models due to the ease with which prior information can be incorporated along with the fact that it provides a more realistic and accurate estimation of uncertainty. In this paper, we first review the sources of information and different approaches (frequentist and Bayesian) to model the distribution of a species. We also

discuss the Integrated Nested Laplace approximation as a tool with which to obtain marginal posterior distributions of the parameters involved in these models. We finally discuss some important statistical issues that arise when researchers use species data: the presence of a temporal effect (presenting different spatial and spatio-temporal structures), preferential sampling, spatial misalignment, non-stationarity, imperfect detection, and the excess of zeros.

Keywords

Geostatistics, hierarchical Bayesian models, INLA, point processes, preferential sampling, SPDE

11.1 Introduction

Understanding spatio-temporal dynamics of species or diseases is a key issue in many research areas such as ecology or epidemiology. Indeed, the so-called Species Distribution Models (SDMs), which link information on the presence/absence or abundance of a species to environmental variables to predict where (and how much of) a species is likely to be present in unsampled locations or time periods, are important tools in many applied fields.

In the particular case of ecology, SDMs have been implemented in different theoretical and practical cases, including the identification of critical habitats (Zhang, 2007; Zhang W, 2008; Paradinas et al., 2015; Rufener et al., 2017; Sadykova et al., 2017), the study of the risk associated with invasive species (Fitzpatrick et al., 2007; Luo and Opaluch, 2011), the potential effects of climate change (Iverson et al., 2004; Araújo et al., 2005; Brown et al., 2016), the design of protected areas, the protection of threatened species (Parviainen et al., 2008; Roos et al., 2015), the distribution of bioclimatic indices (Barber et al., 2017), the reintroduction of vulnerable species (Danks and Klein, 2002; Martinez-Meyer et al., 2006; Hendricks et al., 2016), the delineation of hot spots of biodiversity and species richness

(Jiménez-Valverde and Lobo, 2007; Gotelli et al., 2009; Goetz et al., 2014), the potential distribution of infectious diseases(Peterson et al., 2002; Fatima et al., 2016; Juan et al., 2017; Martínez-Bello et al., 2017; Martínez-Minaya et al., 2018), among many others.

SDMs have also been used in many other contexts, for instance evolutionary biology, where they have been applied to topics such as speciation or hybrid zones (Kozak et al., 2008); in humans epidemiology, to predict the spread of diseases in humans (Gosoniu et al., 2006), in veterinary epidemiology (González-Warleta et al., 2013; Barber et al., 2016), in plants epidemiology (Meentemeyer et al., 2011; Václavík and Meentemeyer, 2009; Neri et al., 2014; White et al., 2017), etc.

Several review papers on SDMs already exist (see for example, Guisan and Thuiller, 2005; Elith and Leathwick, 2009), but most of them are focused on the modeling of species data, maintaining a more general overview of the statistical critical issues. Our intention in this review is to describe in more detail some of the statistical issues that arise when dealing with SDMs.

In addition, the quantity and the quality of available datasets has substantially increased over the past ten years, resulting in a higher complexity of the statistical issues that have to be addressed when a SDM is performed. Moreover, a detailed spatial and temporal description of the modeled phenomenon is becoming mandatory in many research fields. As a consequence of this increasing complexity, the performance of the SDM inferential and predictive processes are becoming more challenging, forcing researchers to develop new sophisticated statistical techniques. Accordingly, new modeling approaches continue to be developed because using only geographic information systems (GIS) tools is not totally satisfactory because of the type of spatial data usually available. Indeed, over time model complexity has generally increased over time from the use of simple environmental matching (two good examples are BIOCLIM, Busby, 1991, and DOMAIN, Carpenter et al., 1993) to the use of models incorporating more complex non-linear relationships between species presence and the environment, such as generalized additive models (Guisan et al., 2002), neural networks (Park et al., 2003), or multivariate adaptive regression splines (Leathwick et al., 2005).

But more importantly, although most of the methods described in previous reviews (see for example, Guisan and Thuiller, 2005; Elith and Leathwick, 2009) have increased in their complexity, they are based on the assumption that the observations are conditionally-independent, while species distribution data often depict residual spatial autocorrelation (Kneib et al., 2008; Beale et al., 2010). In this review, we will focus on the fact that the spatial autocorrelation should be taken into account in species distribution models, even if the data were collected in a standardized sampling, since the observations are often close and subject to similar environmental features (Muñoz et al., 2013). Other complications also arise in the modeling of the species due to imperfect survey data such as observer error, gaps in the sampling, missing data, the spatial mobility of the species (Latimer et al., 2006) and the fact that data have been collected over long periods of time. As a consequence, ignoring these issues in this type of analysis could lead to misleading results.

As a consequence, the use of spatial and spatio-temporal models has grown enormously, allowing the incorporation of all these issues into the modeling process (Banerjee et al., 2014). Although there are other types of spatial data that could describe the behavior of a species (see for instance, Gelfand et al., 2010, for a detailed description of the three types of spatial data), we will focus in this review on geostatistical or point-referenced data that derive from those situations where the concern is to analyze spatially continuous phenomena. Bearing in mind that we want to include the effect of possible covariates in the modeling or to apply it to situations in which the stochastic variation in the data is known to be non-Gaussian, we will deal with the model-based geostatistics approach (Diggle and Ribeiro, 2007).

This combination of non-Gaussian data, a linear predictor and unobserved latent variables usually makes estimation and prediction computationally difficult. Bayesian inference proves to be a good option to deal with spatial hierarchical models because it allows both the observed data and model parameters to be random variables (Banerjee et al., 2014), resulting in a more realistic and accurate estimation of uncertainty. Another advantage of the Bayesian approach is the ease with which prior information can be incorporated. Note that prior information can usually be very helpful in discriminating spatial autocorrelation effects from ordinary non-spatial

linear effects (Gaudard et al., 1999). But, as is usual in Bayesian complex models, inference needs numerical approaches. Among them, in this review we will emphasize on the use of the integrated nested Laplace approximation (INLA) methodology (Rue et al., 2009) and software (http://www.r-inla.org) as an alternative to Markov chain Monte Carlo (MCMC) methods, the main reason being the speed of calculation.

To summarize, our intention in this review is to describe in more detail the main statistical issues that arise when dealing with these models. In particular, in Section 11.2 we focus on the statistical aspects of the available data, while Section 11.3 discusses the basic structure of these models and how to perform inference. In particular, we provide a critical review of the Bayesian approach along with a detailed description of INLA. Our review also includes a discussion on some of the particularities appearing when dealing with them, including temporal correlation, preferential sampling, spatial misalignment, non-stationarity, imperfect detection and excess of zeros in Section 11.4. Finally Section 11.5 concludes. To be noted is that we have tried to be simple in the notation so that the paper is readable by a large community of scientists.

11.2 Sources of information in SDMs

SDMs require basically two types of data input: data describing the observed species' distribution, and data describing the landscape and the environmental characteristics in which the species can be found. In this Section we first present biological data, i.e. the observed species distribution, and then the environmental data and the usual covariates that characterize the species distribution.

11.2.1 Biological data

The first type of data, which usually represent the response variable, can be either presence-only (i.e. records of localities where the species has been observed), presence/absence (i.e. records of presence and absence of the

sampling localities), abundance data (i.e. the quantity of the species at the sampling locations), or proportional data (i.e. the proportion of the species at the sampling locations). Consequently, biological data can be measured at nominal (e.g. presence/absence type), ordinal (e.g. ranked abundance), ratio (e.g. frequency of detection) or continuous (e.g. abundance, richness) levels, which impacts on the selection of the appropriate types of modeling algorithms to use, and subsequently the measurement level of model of this kind (e.g. probability or suitability of occurrence, type, expected mean).

Presence-only data lack absence observations, so that this type of dataset is unsuitable for many of the commonly used species distribution algorithms, unless $pseudo - absences$ are assigned to unsampled portions of the study area. Inclusion of $pseudo - absences$ records can seriously bias analyses. Indeed, methods used to generate pseudo-absences and their effects on model performance are an open research field in the species distribution context (Barbet-Massin et al., 2012; Iturbide et al., 2015).

With respect to abundance, this could be expressed as a continuous variable (biomass of the species) or as count data (number of individuals). Abundance data reflect the quantitative spatial distribution of the species within the area of interest, while presence/absence information can be used to measure the relative occurrence of species, thereby giving a different approximation. Although abundance data provide greater information for conservation and management purposes, they are less common, because occurrence data are easier and less expensive to be collected. Indeed, abundance estimations are sensitive to detectability, and sampling methods seldom detect all individuals present in an area. Consequently, many research studies rely on approximations of species abundance from species occurrence, although whether abundance can be inferred from such information has been questioned, because detection is not perfect and occurrence probability may not be linearly related to density (Nielsen et al., 2005; Joseph et al., 2006).

Proportional data are also widely used in many ecological processes. The traditional approach in ecology is based on Gaussian linear models with previous transformation in the proportions. However, model parameters cannot be easily interpreted in terms of the original response, and measures of proportions typically display asymmetry: hence, inferences based on the

normality assumption can be misleading (Ferrari and Cribari-Neto, 2004). Beta regression has recently appeared as a good alternative to deal with data of this type, allowing bounded estimates and intervals with model parameters that are directly interpretable in terms of the mean of the response (Paradinas et al., 2016, 2017b).

Also to be noted is that different species do not behave independently. There are several species whose abundance (or presence) is constrained by competition: a large increase in one is unavoidably linked to declines in others. In these cases, the response variable should be considered by using a joint distribution. The models used for data of this type are known as joint species distribution models (Clark et al., 2014; Pollock et al., 2014; Hui, 2017; Taylor-Rodriguez et al., 2017).

All these types of biological data describing the observed species' distribution can be obtained in a variety of ways, such as museum collection, designed field surveys, related activities (i.e. fisheries) or on-line resources.

11.2.2 Environmental data

With respect to the explanatory variables that could help to describe the species behavior, a wide range of environmental variables have been usually incorporated in SDMs. These variables are commonly related to climate (e.g. temperature, precipitation), topography (e.g., elevation, aspect, bathymetry, slope of the seabed), land cover type or seabed type in marine ecosystems. Variables tend to describe primarily the abiotic environment, although there is potential to include biotic interactions within the modeling.

These variables can be collected in situ, but they are usually derived from remoted sensing data. CRU (New et al., 2002), WorldClim (Hijmans et al., 2005), and MARSPEC (Sbrocco and Barber, 2013) are all examples of spatially explicit datasets of climatic remote sensing conditions. These datasets encompass climatic information based on interpolations from global weather stations. However, interpolations are only as good as the underlying data, and uneven geographical coverage leads to high model uncertainty,

especially in developing countries where few weather stations are in place (Daly, 2006; He et al., 2015). When uncertainty in spatial climate variables is not accounted for, coefficient estimates tend to be biased, and this leads to poor performances of the SDMs, as recently shown with simulations by Stoklosa et al. (2015). This problem, also known as misalignment, is treated in this review in section 11.4.3.

11.3 Inference

In what follows, after presenting the traditional methods that have been used to perform inference in SDMs, we first discuss the hierarchical modeling as one of the most flexible and encompassing approaches to deal with them. The second subsection presents the Bayesian framework as a good option for dealing with hierarchical models. The final subsection deals with the INLA approach to approximating the marginal posterior distributions of the parameters involved in the SDMs.

11.3.1 Gaussian Fields and hierarchical modeling

A number of alternative modeling algorithms have been applied to classify species distribution as a function of a set of environmental variables. A first group of methods developed to deal with presence-only datasets includes maximum entropy algorithm, environmental distance, similarity, and envelope methods such as MAXENT (Phillips et al., 2006), Gower metric, Mahalanobis distance, and ecological niche factor analysis, all of which describe some measure of habitat suitability.

A second group involves machine-learning algorithms that are iterative in nature, such as artificial neural networks. These 'ensemble' methods (e.g. Boosting Regression Trees, Classification Trees and Random Forests) generally involve developing multiple models on different subsets of the data, the results of which are averaged (Franklin, 2010).

A third group of methods relates to traditional regression and includes generalized linear models (GLM) and their non-parametric extension, generalized additive models (GAM), both of which can handle several measurement levels of the response variable by using a different link function (e.g. logistic for presence/absence or log for counts). GAM and a related method, multivariate adaptive regression splines (MARS), are more flexible than GLM as they are fitted using smoothing and piecewise linear splines, respectively, and are particularly useful for identifying the shape of species responses (Leathwick et al., 2005). MARS is computationally faster than GAM and the results are more easily converted to map predictions in a GIS; however, the currently used algorithms require normally distributed error terms. This makes MARS unsuitable for use with presence/absence data unless the basis functions are extracted and used to parameterize a GLM (Leathwick et al., 2005). Rodríguez de Rivera and López-Quílez (2017) present a comparison of these three groups of methodologies stating that GAM models gave the best results.

However, most of the above mentioned methods are based on the assumption that the observations are conditionally-independent. But this is not always the case beacause data of species distribution usually present residual spatial autocorrelation (Kneib et al., 2008). GAMs and MARS can model spatial and temporal autocorrelation using smoothing splines. A very powerful and flexible alternative is to incorporate this spatial relationship by considering the species distribution data as point-referenced or geostatistical data. Data of this type appear in those situations where the interest is to analyze spatially continuous phenomena. The most basic format for data of this kind is a pair composed by the spatial location coordinates defined throughout a continuous study region and the measurement value observed in the location. Geostatistical data require methods that make it possible to relate the species data with potential related covariates by quantifying the spatial dependence. However, one of the main interests in geostatistics concerns predicting the underlying process on those non-observed locations (Cressie and Wikle, 2011; Banerjee et al., 2014).

Geostatistical or point-referenced data can be seen as realizations of a spatial process (random field) $\{y(\boldsymbol{s}), \boldsymbol{s} \in \mathcal{D}\}$ characterized by a spatial index $\boldsymbol{s}$ which varies continuously in the fixed domain $\mathcal{D}$. This process is called a

Gaussian field (GF) if for any $n \geq 1$ and for each set of locations $(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n)$, the vector $(y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n))$ follows a multivariate Normal distribution with mean $\boldsymbol{\mu} = (\mu(\boldsymbol{s}_1), \ldots, \mu(\boldsymbol{s}_n))$ and with covariance matrix $\boldsymbol{\Sigma}$ defined by a covariance function $\mathcal{C}(\cdot, \cdot)$, such that $\Sigma_{ij} = Cov(y(\boldsymbol{s}_i), y(\boldsymbol{s}_j)) = \mathcal{C}(y(\boldsymbol{s}_i), y(\boldsymbol{s}_j))$. If the mean is constant in space, i.e. $\mu(\boldsymbol{s}_i) = \mu$ for each i, and the generic spatial covariance matrix element depends only on the difference vector $(\boldsymbol{s}_i - \boldsymbol{s}_j) \in \mathbb{R}^2$, the spatial process is second-order stationary. In addition, if the covariance function only depends on the Euclidean distance $\|\boldsymbol{s}_i - \boldsymbol{s}_j\|$, the process is said to be isotropic.

In a hierarchical framework, the first step in defining a model for a random field is to identify a probability distribution for the observations available at n spatial locations and represented by the vector $\boldsymbol{y} = (y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n)) = (y_1, \ldots, y_n)$ (the notation is simplified and the index i is used for denoting the generic spatial points $\boldsymbol{s}_i$). At the first level of the hierarchy, we usually select a distribution from the exponential family, characterized by a set of parameters. These parameters are linked with a linear predictor which also includes a latent GF denoted by $\xi(\boldsymbol{s})$ whose covariance function $\boldsymbol{\Sigma}$ depends on two parameters: σ^2 which represents the variance (partial sill in kriging terminology) and the range ϕ of the spatial effect.

Computational costs required to estimate these parameters are high when we deal with the spatial covariance function because the generated matrices are dense. This problem is known as the "big n problem" (Banerjee et al., 2014; Jona Lasinio et al., 2012) and despite computational power today, it is still a computational bottleneck in many situations. A computationally effective alternative is given by the stochastic partial differential equation (SPDE) approach proposed by Lindgren et al. (2011) (see Section 11.3.3).

In addition to the spatial pattern, the temporal variation could be equally important because the phenomenon can vary not only in space but also in time (see Hefley and Hooten, 2016, for a comprehensive overview of modeling species distribution with a spatio-temporal perspective). Then, extending the spatial case to the spatio-temporal case including a time dimension, the process indexed by space and time can be defined as

$\{y(\boldsymbol{s},t),(\boldsymbol{s},t) \in \mathcal{D} \subset \mathbb{R} \times \mathbb{R}\}$, and is observed at n spatial locations and at T time points.

The general structure for modeling the spatial distribution of species is given by the following formulation and notation. If $\boldsymbol{y} = (y_1, \ldots, y_n)$ represents the observed values of the corresponding response variable Y with mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$, each μ_i can be easily linked to a structured additive predictor η_i through a link function $g(\cdot)$, so that $g(\boldsymbol{\mu}) = \boldsymbol{\eta}$. The structured additive predictor $\boldsymbol{\eta}$ accounts for the effect of various covariates in an additive way:

$$\eta_i \quad = \quad \beta_0 + \sum_{m=1}^{M} \beta_m x_{mi} + \sum_{l=1}^{L} f_l(z_{li}) \ , \tag{11.1}$$

where β_0 corresponds to the intercept; the coefficients $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_M\}$ quantify the (linear) effect of some covariates $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M)$ on the response; and $\boldsymbol{f} = \{f_1(\cdot), \ldots, f_L(\cdot)\}$ are unknown functions of the covariates $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_L)$, and can assume different forms such as smooth nonlinear effects of covariates, time trends and seasonal effects, random intercept and slopes as well as temporal or spatial random effects. Note that this general structure can also be seen as a Generalized Additive Mixed Model (GAMM). Also to be noted is that here it is assumed that covariates are observed at the same locations of the response variable. The situation where covariates are observed in locations different from those of the response variable (misalignment) will be discussed in Section 11.4.3.

In many statistical applications, in particular, in SDMs, the model involves multiple parameters that can be regarded as related or connected in some way by the structure of the problem, implying that a joint probability model for these parameters should reflect their dependence (Gelman et al., 2014). It is common to model such a problem hierarchically, with observable outcomes modeled conditionally on certain parameters, which in turn are given a probabilistic specification in terms of further parameters, adding various levels of the modeling and thus defining a hierarchical model (HM). Note that Hierarchical models provide a generalization of all the models presented here; and moreover that they are able to deal with all the types of the data that we can be found when dealing with SDMs. Table 11.1

TABLE 11.1: Matching of models presented and data types. LM: linear models. LMM: linear mixed models. GLM: Generalized linear models. GLMM: Generalized linear mixed models. AM: additive models. AMM: additive mixed models. GAM: Generalized additive models. GAMM: Generalized additive mixed models. HM: Hierarchical models. By construction, these models are nested: LM < GLM < GAM < GAMM < HM.

Explanatory	Response variable distribution	
Variable(s)	**NORMAL**	**OTHER DIST. EXP. FAMILY**
LP	LM	GLM
R. effects	LMM	GLMM
Non-Lin. effects	AM	GAM
R. effects + Non-Lin. effects	AMM	GAMM

describes all the models mentioned in this subsection along with a diagram emphasizing their nested nature.

Although other approaches can be used such as maximum likelihood (MLE; Le Cam, 1990), restricted maximum likelihood (RMLE; Bartlett, 1937), quasi-maximum likelihood (QMLE; Cox and Reid, 2004), the method of moments (Bowman and Shenton, 2006), the generalized method of moments (GMM; Hansen, 1982), M-estimators (Shapiro, 2000), the maximum spacing estimation (MSE; Anatolyev and Kosenok, 2005), etc., here we will focus on the Bayesian approach to making inference for hierarchical models with a linear predictor of the form (11.1).

11.3.2 Bayesian approach

The use of the Bayesian framework as a way to make inference has increased in the past 50 years and it has been applied in different areas, such as social sciences (Jackman, 2009), medicine and public health (Berry and Stangl, 1999), finance (Rachev et al., 2008), ecology (McCarthy, 2007), bioinformatics (Mallick et al., 2009), health economics (Baio, 2012), physical sciences (Andreon and Weaver, 2015) and econometrics (Gómez-Rubio et al., 2014).

Bayesian reasoning is based on the assumption that parameters are random variables, and prior knowledge has to be incorporated via the corresponding prior distributions of the said parameters. Bayes' theorem is the tool that combines prior information with the likelihood yielding the posterior distributions. To be noted is that the Bayesian approach is perfectly suited for complex spatial models such as SDMs because it allows model parameters to be random variables, resulting in a more realistic and accurate estimation of uncertainty.

SDMs are a very good example of a hierarchical structure that can be expressed as a hierarchical Bayesian model (Wikle and Hooten, 2010; Hefley and Hooten, 2016). They can be structured in three levels: the first one refers to the data and is conditioned on the process and parameters in whatever aspects of the process are appropriate. The second level contains the latent components, which can be spatial and/or dynamic and the stochastic form can be univariate or multivariate. Finally, the third stage defines the priors for the parameters on which the latent processes depend. The parameters in this level are also known as hyperparameters.

The approach most commonly used to perform Bayesian inference for spatial species distribution models is based on MCMC methods (Gelfand et al., 2006); they are flexible computational tools which can be easily adapted to any kind of inferential problem. The software most frequently used to implement MCMC algorithms are `WinBUGS` (Lunn et al., 2000; Brooks et al., 2011), `OpenBUGS` (Lunn et al., 2009) and `JAGS` (Plummer, 2016), which can also be run within other programs like `R` (through the `R2OpenBUGS`, `R2WinBUGS`, `BRugs` and `rjags` packages), Stata and SAS. Alternatively other `R` packages are `BayesX` (Brezger et al., 2003), `CARBayes` (Lee, 2013), `stocc` (for binary data only), `spatcounts` (for count data only), `CARramps` (for Gaussian data only), and `spdep` (for Gaussian data only). Several hierarchical models including ecological processes (habitat suitability, spatial dependence and anthropogenic disturbance) and observation processes (species detectability) can also be performed using the `hSDM` package of `R` developed by Vieilledent et al. (2014). Functions in this `R` package use an adaptive Metropolis algorithm (Robert and Casella, 2011) in a Gibbs sampler (Gelfand and Smith, 1990) to obtain the posterior distribution of model parameters. The Gibbs sampler is written in `C` code and compiled

to optimize computation efficiency. Thus, the `hSDM` package can be used for very large data-sets while drastically reducing the computation time. However, with `hSDM` it is not possible at present to model spatio-temporal or proportion response variables.

Despite their generalized use, to be noted is that MCMC methods still have many challenges to deal with (like the so-called "big n problem" mentioned above; see Banerjee et al. 2014; Jona Lasinio et al. 2012). Indeed, they can be extremely slow and even computationally unfeasible especially when the models are extremely complex (with many random effects or hierarchical levels) or when big datasets are considered in the space-time setting.

As a result, other options have appeared to make inference in SDMs. Taking advantage of the hierarchical structure of SDMs, Golding and Purse (2016) propose the use of an empirical Bayesian approach. In particular, they maximize the marginal posterior density of the model, which, in their words, enables the incorporation of prior knowledge over hyperparameters whilst being much less computationally intensive than fully Bayesian inference.

Here, we will focus on the integrated nested Laplace approximation (INLA) methodology (Rue et al., 2009), as a computational effective alternative to MCMC. Our choice is due to two considerations: the speed of calculation, and the ease with which model comparison can be performed.

11.3.3 INLA and SPDE framework

The INLA methodology is now a well-established tool for Bayesian inference in several research fields, including ecology, epidemiology, econometrics and environmental science (Rue et al., 2017). It can be used through `R` with the `R-INLA` package. For more details on INLA for spatial and spatio-temporal models we refer the reader to Blangiardo et al. (2013) and Blangiardo and Cameletti (2015), where practical examples and code guidelines are also provided.

The reason why INLA can be used is that SDMs can be seen as latent Gaussian models (Rue and Held, 2005), for which the class of models INLA

is designed. After identifying the distribution for the observed data, we can link its corresponding mean to the linear predictor as in Eq.(11.1). If conditional independence is assumed, the distribution of the n observations is given by the likelihood

$$p(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\psi}) = \prod_{i=1}^{n} p(y_i \mid \theta_i, \boldsymbol{\psi}) \ , \tag{11.2}$$

where $\boldsymbol{\theta}$ represents the set of latent (nonobservable) components of interest $\boldsymbol{\theta} = \{\beta_0, \boldsymbol{\beta}, \boldsymbol{f}\}$, also known as the latent field, and $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_K)$ denotes the vector of K hyperparameters. As we can observe in Eq. (11.2), each data point y_i is connected to one element θ_i in the latent field. This assumption can be relaxed, and each observation can be connected with a linear combination of elements in $\boldsymbol{\theta}$ (Martins et al., 2013). In addition, the multiple likelihood case can also be taken into account.

In the context of latent Gaussian models, assumed is a multivariate Normal prior distribution on $\boldsymbol{\theta}$ with mean $\mathbf{0}$ and precision matrix $\boldsymbol{Q}(\boldsymbol{\psi})$, i.e, $\boldsymbol{\theta} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{Q}^{-1}(\boldsymbol{\psi}))$ with density function given by

$$p(\boldsymbol{\theta} \mid \boldsymbol{\psi}) = (2\pi)^{-n/2} |\boldsymbol{Q}(\boldsymbol{\psi})|^{1/2} \exp\left(-\frac{1}{2}\boldsymbol{\theta}'\boldsymbol{Q}(\boldsymbol{\psi})\boldsymbol{\theta}\right) \ , \tag{11.3}$$

being $|\cdot|$ the matrix determinant and $'$ the transpose operation. When the precision matrix $\boldsymbol{Q}(\boldsymbol{\psi})$ is sparse a GF becomes a Gaussian Markov random field (GMRF, Rue and Held, 2005). Interestingly, when making inference with GMRFs, linear algebra operations are performed using numerical methods for sparse matrices, and this yields computational benefits.

In spite of the wide acceptance of INLA, its precision and its computational efficiency in many latent Gaussian models (see for instance, Martino et al., 2011; Schrödle et al., 2011; Ruiz-Cárdenas et al., 2012, for a description of how to use INLA in spatio-temporal disease mapping, in state-space models and in survival models, respectively), INLA cannot be directly applied when dealing with models that incorporate geostatistical data (that is, continuously indexed Gaussian Fields). The underlying reason is that a parametric covariance function needs to be specified and fitted based on the data, which determines the covariance matrix $\boldsymbol{\Sigma}$ and enables prediction in

unsampled locations. But from the computational perspective, the cost of factorizing the dense covariance matrix $\mathbf{\Sigma}$ is cubic in its dimension. Despite current computational power, in many situations it is still challenging to factorize it for computing the inverse and the determinant.

Lindgren et al. (2011) proposed an alternative approach by using an approximate stochastic weak solution to a Stochastic Partial Differential Equation (SPDE) as a GMRF approximation to a continuous Gaussian Field (GF) with Matérn covariance structure. Specifically, they used the fact that a Gaussian Field $\xi(\boldsymbol{s})$ with Matérn covariance is a solution to the linear fractional SPDE

$$(\kappa^2 - \Delta)^{\alpha/2}(\tau\xi(\boldsymbol{s})) = \mathcal{W}(\boldsymbol{s}),\ \boldsymbol{s} \in \mathbb{R}^d,\ \alpha = \nu + \delta/2,\ \kappa > 0,\ \nu > 0, \quad (11.4)$$

where Δ is the Laplacian, α controls the smoothness, κ is the scale parameter, τ controls the variance, and $\mathcal{W}(\boldsymbol{s})$ is a Gaussian spatial white noise process. The exact and stationary solution to this SPDE is the stationary GF $\xi(\boldsymbol{s})$ with Matérn covariance function given by:

$$\begin{aligned} \text{Cov}(\xi(\boldsymbol{s}_i), \xi(\boldsymbol{s}_j)) &= \mathcal{C}(\xi_i, \xi_j) \quad (11.5) \\ &= \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\kappa||\boldsymbol{s}_i - \boldsymbol{s}_j||)^\nu K_\nu(\kappa||\boldsymbol{s}_i - \boldsymbol{s}_j||), \end{aligned}$$

being $||\boldsymbol{s}_i - \boldsymbol{s}_j||$ the Euclidean distance between two locations $\boldsymbol{s}_i, \boldsymbol{s}_j \in \mathbb{R}^d$, and σ^2 the marginal variance. Moreover, K_ν is the modified Bessel function of the second kind and order $\nu > 0$, which measures the degree of smoothness of the process. This parameter is usually kept fixed due to its poor identifiability. Conversely, $\kappa > 0$ is a scaling parameter related to the distance at which the spatial correlation becomes almost null, i.e., the range (for more information on the Matérn covariance model see Handcock and Stein, 1993; Stein, 1999). Typically, as pointed out in Lindgren et al. (2011), the empirically derived definition for the range is $r = \frac{\sqrt{8\nu}}{\kappa}$, with r corresponding to the distance at which the spatial correlation is close to 0.1, for each $\nu \geq \frac{1}{2}$.

The link between equations (11.4) and (11.6) is given by the expressions $\nu = \alpha - \frac{\delta}{2}$, and $\sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\alpha)(4\pi)^{\delta/2}\kappa^{2\nu}\tau^2}$. In the particular case where the

dimension is 2, i.e., $\delta = 2$, it follows that $\nu = \alpha - 1$ and $\sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\alpha)(4\pi)\kappa^{2\nu}\tau^2}$.

Finally, in `R-INLA`, the Gaussian field $\xi(\boldsymbol{s})$ is found numerically as a weak solution to the SPDE in (11.4), and by default the smoothness parameter α is fixed to 2, corresponding with $\nu = 1$. With this assumption, the range is given by $\phi \approx r = \sqrt{8}/\kappa$, while the variance is given by $\sigma^2 = 1/(4\pi\kappa^2\tau^2)$.

Bayesian geostatistical analysis using `R-INLA` has already been applied in various contexts. Along with introducing the `geostatsinla` package for performing geostatistics with INLA in an easy way, Brown (2015) applies it in the context of mapping the *Loa loa* filiarasis disease (a dataset previously cited in Diggle and Ribeiro, 2007). Moreover, Karagiannis-Voules et al. (2013) have used Bayesian geostatistical negative binomial models to analyze reported incidence data of cutaneous and visceral leishmaniasis in Brazil covering a 10-year period, while González-Warleta et al. (2013) have used Bayesian geostatistical binomial models to predict the probability of infection of paramphistomosis in Galicia (NW Spain). In the context of fisheries, Bayesian geostatistical analysis using `R-INLA` has also been used to predict the presence/absence, the abundance, or the proportion of fish species (Muñoz et al., 2013; Pennino et al., 2013, 2014, 2016a,b; Paradinas et al., 2015, 2016; Cosandey-Godin et al., 2015; Quiroz et al., 2015; Roos et al., 2015; Rufener et al., 2017).

11.4 Extending statistical modeling of species distribution

There are a number of additional potential sources of bias and error that should be carefully considered when analyzing and modeling species distribution data. Errors may arise through the incorrect identification of species, or inaccurate spatial referencing of samples. Biases can also be introduced because collectors tend to sample in easily accessible locations. Here we discuss some of these issues.

11.4.1 Temporal autocorrelation

As mentioned above, in addition to the spatial pattern, the temporal variation could be equally important because the phenomenon may vary not only in space but also in time. This happens in problems such as the evolution of epidemics (Stein et al., 1994; Hefley et al., 2017b), the spatio-temporal evolution of temperature (Hengl et al., 2012) or the understanding of the spatial dynamism of species over time (Wikle, 2003; Hooten et al., 2007; Hooten and Wikle, 2008; Paradinas et al., 2015, 2017a; Williams et al., 2017).

As pointed out by Cressie and Wikle (2011), temporal correlation depends on the same principle as spatial correlation: temporally close observations tend to be more related than temporally distant ones. Consequently, model fitting and predictions improve when a temporal term is added. However, temporal and spatial scales are different and the spatio-temporal analysis is more complicated than the simple addition of an extra dimension to the continuous spatial domain.

In the context of species distribution modeling, most studies (surveys, plant coverage surveys, air pollution surveys, etc.) have been repeated periodically for long periods of time (Gitzen, 2012; Aizpurua et al., 2015). Although the main interest is the spatial evolution of the system under study, it must be considered that it varies not only in space but also in time. Here we focus on this most common situation of discrete and regular time observations. For situations in which data are collected in irregular time-lags - that is, when the issue is handling continuous-time data - a good option is to consider 1D SPDE models with a second order B-Spline basis representation (Lindgren and Rue, 2008, 2015).

The spatio-temporal behavior of the data can vary depending on the nature of the process under study and the available sampling resolution. In particular, the basic model in (11.1) can be rewritten by splitting the f term into two terms, one indicating different possible spatio-temporal structures, and the other indicating any other latent model or non-linear effect. If y_{it} represents the response variable analyzed at location $\boldsymbol{s}_i$ $(i = 1, \ldots, n)$ at time t $(t = 1, \ldots, T)$, then the mean of the response variable μ_{it} is linked to

the linear predictor with a link function $g(\cdot)$, as

$$\eta_{it} \quad = \quad g(\mu_{it}) = \beta_0 + \sum_{m=1}^{M} \beta_m x_{mit} + \sum_{k=1}^{K} f_k(z_{kit}) + u_{it} \ , \qquad (11.6)$$

where β_0 corresponds to the intercept; the coefficients $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_M\}$ quantify the linear effect of some covariates on the response; u_{it} represents the spatio-temporal structure of the model; z_{kit} is the value of the k-th explanatory variable at a given location $\boldsymbol{s}_i$ and time t; and f represents any latent model applied to the covariates.

Among other structures, and following Paradinas et al. (2017a), we comment here on four basic structures for u_{it}, each one allowing for different degrees of flexibility in the temporal domain of the spatio-temporal model. Paradinas et al. (2017a) provide a figure that schematically illustrates all these structures:

- **Opportunistic spatial distribution**: this flexible structure consists in expressing u_{it} as different spatial realizations $\boldsymbol{w}_t = \{w_{1t}, \ldots, w_{it}, \ldots, w_{nt}\}$ of the same spatial field for each time unit t, while sharing a common covariance function (same κ and τ) to avoid overfitting:

$$\begin{aligned} &u_{it} = w_{it} \ , \\ &\boldsymbol{w}_t \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{Q}^{-1}(\kappa, \tau)) \ . \end{aligned} \qquad (11.7)$$

  This structure is a good approximation for processes where the spatial distribution varies considerably among different time units and unrelatedly among neighboring times. This structure has been used in Cosandey-Godin et al. (2015) and in Paradinas et al. (2015).

- **Persistent spatial distribution with random intensity changes over time**: when the pattern of spatial variation persists over time, but with possibly varying scales of intensity, a time structure is introduced into the model using a zero mean Gaussian random noise effect v_t. In this case, u_{it} is decomposed in a common spatial realization w_{it} along with an independent random noise effect v_t that absorbs the

different mean intensities at each time t:

$$\begin{aligned} u_{it} &= w_{it} + v_t \ , \\ \boldsymbol{w}_t &\sim \mathrm{N}(\mathbf{0}, \boldsymbol{Q}^{-1}(\kappa, \tau)) \ , \\ v_t &\sim \mathrm{N}(0, \tau_v^{-1}) \, . \end{aligned} \tag{11.8}$$

For processes where the spatial component persists in time, this structure may be the most suitable. It has been used by Pennino et al. (2014) and in Paradinas et al. (2015).

- **Persistent spatial distribution with temporal intensity trend**: the process could show a temporal progression in its mean. To model that, a temporal trend effect $h(t)$ can be added to the linear predictor. In this case, u_{it} is decomposed into a common spatial realization w_i and an independent temporal structured trend $h(t)$ to absorb the temporal progression of the process:

$$\begin{aligned} u_{it} &= w_i + h(t) \ , \\ \boldsymbol{w} &\sim \mathrm{N}(\mathbf{0}, \boldsymbol{Q}^{-1}(\kappa, \tau)) \ . \end{aligned} \tag{11.9}$$

This structure is highly recommended in situations where a temporal tendency is present. It was proposed by Paradinas et al. (2016) to identify intra-annual trends in fishery discards.

- **Progressive spatio-temporal distribution**: this structure incorporates both spatial and temporal correlation of the data to accommodate those cases where the spatial realizations change in a related manner over time. Here, u_{it} is decomposed into a common spatial realization w_{it} and an autoregressive temporal term r_{it} expressing the correlation among temporal neighbors of order K:

$$\begin{aligned} u_{it} &= w_{it} + r_{it} \ , \\ \boldsymbol{w}_t &\sim \mathrm{N}(\mathbf{0}, \boldsymbol{Q}^{-1}(\kappa, \tau)) \ , \\ r_{it} &\sim \mathrm{N}\left(\sum_{k=1}^{K} \rho_k r_{i(t-k)}, \tau_r^{-1} \right) . \end{aligned} \tag{11.10}$$

This structure is preferred when the spatial realization varies between different times but not as much as in (11.7). Indeed, the structure has been used by Cameletti et al. (2011, 2013) and also by Cosandey-Godin et al. (2015).

Note that this list is only an overview of the different spatio-temporal structures which allow us to discern the nature of the general spatial behavior of the process over time. Unfortunately, the temporal resolution of spatio-temporal datasets is typically too low to fit most of the highly structured models.

11.4.2 Preferential sampling

In studies on species distributions, collecting data on the species of interest is not a trivial problem. With the exception of a few studies, species distribution models rely on opportunistic data collection due to the high cost and time-consuming nature of collecting data in the field, especially on a large spatial scale. As an example, studies on bird monitoring data are often collected by volunteers who concentrate the sampling process on areas where they expect to find species of interest. These types of opportunistically collected data tend to suffer from a specific complication: the sampling process that determines the data locations and the species observations are not independent (Diggle et al., 2010). Statistical models used for species distribution usually assume, if only implicitly, that sampling is non-preferential and that the selection of the sampling locations does not depend on the values of the spatial variable. However, opportunistic data are a clear example of preferential sampling, that occurs because sampling locations are deliberately chosen in areas where the values of the species of interest are thought likely to be particularly high or low (Diggle et al., 2010).

Hence, applying standard geostatistical methods to preferentially sampled data potentially yields biased results if the choice of monitoring locations is not accounted for in the modeling process. A possible approach to correct this issue is to interpret the data as a marked point pattern (Fortin and Dale, 2005; Diggle, 2013) where the sampling locations form a point

pattern and the observations taken in those locations are the marks. By assuming that the intensity of the point process depends on the amount of species of interest, the marks and the pattern become not independent.

A preferential sampling model can be considered as a two-part model that share information. Firstly, it is supposed that the observed locations $(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n)$ come from a non-homogeneous Poisson process with intensity $\Lambda_i = \exp\{\alpha_1 + w_i\}$, i.e., a log-Gaussian Cox process (LGCP; Fortin and Dale, 2005; Diggle, 2013) is assumed, being α_1 the intercept of the LGCP and w_i the spatial effect of the model and $i = 1, \ldots, n$ the index corresponding to the $\boldsymbol{s}_i$ location. Secondly, the species characteristic (usually the abundance) y_i is assumed to follow an exponential family distribution (such as a Normal or a Gamma distribution when dealing with abundances, although other options such as exponential, lognormal, etc., could clearly be possible), whose mean is related with the spatial term using a link function $g(\cdot)$, $g(\mu_i) = \alpha_2 + \beta w_i$, being α_2 the intercept of the model and w_i the spatial term shared with the LGCP, but scaled by β to allow for the differences in scale between the abundances and the LGCP. More formally, the model can be expressed as follows:

$$\begin{aligned} y_i &\sim \mathrm{F}(\mu_i, \gamma^2) \\ \mathrm{g}(\mu_i) &= \alpha_2 + \beta w_i \\ \boldsymbol{w} &\sim \mathrm{N}(\mathbf{0}, \boldsymbol{Q}^{-\mathbf{1}}(\kappa, \tau)) \end{aligned} \tag{11.11}$$

where $\boldsymbol{w} = \{w_1, \ldots, w_n\}$, the precision matrix $\boldsymbol{Q}(\kappa, \tau)$ is computed internally by the SPDE approach and represents the GMRF approximation to the continuous GF (see Illian et al., 2012; Krainski et al., 2017; Pennino et al., 2017, for details about how to implement these models within INLA), and $F(\mu, \gamma)$ represents a distribution coming from the Exponential family with mean μ and variance γ^2.

11.4.3 Spatial misalignment

A crucial issue in studying the effect of environmental physical factors on species distribution concerns spatial misalignment (Clark and Gelfand, 2006; Gelfand et al., 2010) (Foster et al., 2012; Miller, 2012).

This occurs when the response biological variable (e.g. presence/absence of the species) is observed in locations which are different from the spatial points where covariate data are available. Additionally, it can happen that covariates have a different spatial scale if they are defined at the area or cell grid level (as in the case of remote sensing data).

The naïve solution for spatial misalignment is a two-stage approach: the first step consists in the prediction of the covariate in the spatial locations where the response variable is observed (through a geostatistical model by means of kriging or inverse-distance weighting) or in the downscaling of the gridded covariate to the point-level resolution (usually considered is the value of the cell where the spatial point is located). Then, at the second stage, these predicted values are plugged into the linear predictor (11.1) as known constants. The problem with this approach is that it does not take account of the uncertainty related to the covariate spatial estimation of the first stage, with the consequence of erroneous inference of the statistical model and a potential biased estimate of the environmental variable effect on the response variable (Foster et al., 2012).

A solution to incorporate the spatial prediction uncertainty in SDMs consists in implementing one of the so-called *errors-in-variables models* (Carroll et al., 2016) which can be estimated in a frequentist (by means of the EM-algorithm) or Bayesian framework (with MCMC or INLA). If we assume for example that the predicted covariate is a noisy version of the true one, a classic measurement error model can be adopted (Stoklosa et al., 2015). Otherwise, a Berkson-error model can be considered if the predicted covariate is a smoothed (i.e. less variable) version of the true variable (Foster et al., 2012). As reported in Stoklosa et al. (2015) *"Which of these two types of error models to consider will depend on what the analyst believes to be the 'true underlying explanatory variable', and how the data were collected/measured. The analyst must take into account: how and whether the species responds to a particular climate observation (Berkson); or that it might respond to an average, such that relatively minor deviations from this are immaterial (classical)".*

Another alternative to the two-stage approach is the joint modeling strategy implemented in Barber et al. (2016) to evaluate the presence of the

Fasciola hepatica in Galicia (Spain) using the annual mean temperature as covariate. In this case a spatial geostatistical model is specified for the covariate and is estimated jointly with the species distribution models in a Bayesian context. The joint model is specified as follows

$$\begin{aligned} y_i &\sim \text{Bernoulli}(\pi_i) \\ \text{logit}(\pi_i) &= \beta_0 + \beta_1 \phi_i + w_i \\ \boldsymbol{w} &\sim \text{N}(\mathbf{0}, \boldsymbol{Q}^{-\mathbf{1}}(\kappa, \tau)) \\ x_i &\overset{iid}{\sim} \text{N}(\phi_i, \sigma_x^2) \\ \boldsymbol{\phi} &\sim \text{N}(\mathbf{0}, \boldsymbol{Q}^{-\mathbf{1}}(\gamma, \delta)) \end{aligned} \tag{11.12}$$

where π_i is the probability of occurrence at site $\boldsymbol{s}_i$, x_i is the covariate of interest whose spatial distribution is specified through its mean (a realization of the Matérn Gaussian process $\boldsymbol{\phi}$ depending on the parameters γ and δ), and through its variance σ_x^2, which is introduced to express any possible measurement error. The model also includes another spatial process for the response represented by $\boldsymbol{w}$. This kind of model pertains to the latent Gaussian model family and can be estimated using the SPDE-INLA approach (see Blangiardo and Cameletti, 2015, Chap. 8 and Muff et al., 2015). The advantage is that this joint model allows to properly propagate all the uncertainty related to the covariate prediction; on the other it can be extremely computationally expensive especially when there is more than one explanatory variable.

Finally, another alternative is the one proposed by Gómez-Rubio and Rue (2017) that, using a more general approach, deals with missing values in the covariates, based on fitting conditional latent Gaussian models where covariates are imputed using a Metropolis-Hastings algorithm.

11.4.4 Non-stationarity

The Matérn spatial covariance function $\mathcal{C}(\cdot, \cdot)$ specified by Eq. (11.6) enjoys the second-order stationarity and isotropy property, i.e. it depends only on the distance between the spatial locations and not on the direction or the coordinates. In some situations, this stationarity assumption, which is very

convenient to simplify the inferential procedures, may not be suitable. For example, for some applications it is not realistic to assume that the spatial dependence structure is the same throughout the domain considered, especially when geographical elements or physical barriers (river, lakes, islands, etc.) exist. In such situations characterized by spatial heterogeneity and barriers, it may be more reasonable to adopt a non-stationary Gaussian field (see Gelfand et al. 2010, Chapter 9 and Risser 2016 for a review).

In ecological applications, heterogeneity in space (i.e. non-stationarity) occurs when a latent *global* process is also affected by some underlying local processes (Miller, 2012). A local modeling technique to include this heterogeneity in SDMs is given by the geographically weighted regression (GWR) characterized by covariate coefficients which vary spatially and are specific for each spatial location; a spatial kernel function is used to define spatial neighborhoods (see e.g. Brunsdon et al. 1998; Windle et al. 2010; Holloway and Miller 2015; Liu et al. 2017). Some authors do not completely agree with the use of these models due to the large degree of multicollinearity that their coefficients tend to exhibit, as well as strong positive spatial autocorrelation. As an alternative, spatial filtering provides a methodology for dealing better with multicollinearity, while accounting for spatial autocorrelation (see e.g. Griffith 2008). The Bayesian counterpart of GWR models, which are usually estimated by weighted least squares, is given by spatially-varying coefficients models (Gelfand et al., 2003; Finley, 2011).

In the SPDE framework non-stationarity is achieved by allowing the Matérn covariance function parameters to vary smoothly over space according to a log-linear function: thus, we will have $\sigma^2(\boldsymbol{s})$ for the marginal variance in (11.6) and $r(\boldsymbol{s})$ for the spatial range (Ingebrigtsen et al., 2014; Lindgren and Rue, 2015). Bakka et al. (2016) extend this approach to solve specifically the barrier problem for SDMs. In particular, they force the spatial correlation to go around the barriers (and not through them) by means of a partition of the considered spatial field - in a normal and in a barrier area - and in the specification of two corresponding non-stationary processes with different range parameters (in particular for the barrier region the range parameter is almost zero). The application considered in Bakka et al. (2016) regards fish larvae data in the Finnish archipelago.

11.4.5 Imperfect detection

Studies on species abundance and distribution are often imperfect due to observer error (Nichols et al., 2000), species rarity (Dettmers et al., 1999) or because detection varies with confounding variables such as environmental conditions (Gu and Swihart, 2004; Pennino et al., 2016b). When detection is imperfect, additional steps are usually needed to improve inference. Indeed, failure to do so could result in biased estimation and erroneous conclusions.

In recent years, new models called site-occupancy (Hoeting et al., 2000; MacKenzie et al., 2002) for presence-absence data and N-mixture models (Royle, 2004) for abundance data have been developed to solve this problem. These models combine two processes: an ecological process to describe habitat suitability and an observation process to take imperfect detection into account. To estimate detectability, these models use information from repeated observations at several sites. Detectability may vary with site characteristics such as habitat variables, or survey characteristics such as weather conditions, since suitability relates only to site characteristics. Various studies showing the advantages of site occupancy and N-mixture models over classical models that do not consider the problem of detectability can be found in the literature: Royle (2004); Dorazio et al. (2006) for birds, MacKenzie et al. (2002) for amphibians or Pennino et al. (2016b) for cetaceans. In addition to the detectability problem, a variety of methods have been developed to correct for the effects of spatial autocorrelation (Latimer et al., 2006; Johnson et al., 2013; Hefley et al., 2017a).

A Bayesian version for site-occupancy spatial models and N-mixture spatial models could also be implemented to take simultaneously account of both imperfect detection and spatial autocorrelation. To describe Bayesian site-occupancy spatial models, let z_i be a random variable describing habitat suitability at site $\boldsymbol{s}_i$. It can take the value 1 or 0 depending on the habitat suitability, i.e. $z_i = 1$ or $z_i = 0$, thus a Bernoulli distribution is assumed with parameter π_i. Several visits at time $t = 1, \ldots, T$ can happen at site i. Let y_{it} be a random variable representing the presence of the species at site i and time t. The species is observed at site i ($\sum_t y_{it} \geq 1$) only if the habitat is suitable ($z_i = 1$). The species is unobserved at site i ($\sum_t y_{it} = 0$) if the habitat is not suitable ($z_i = 0$), or if the habitat is suitable ($z_i = 1$)

but the probability α_{it} of detecting the species at site $\boldsymbol{s}_i$ and time t is lower than 1. Then, y_{it} follows a Bernoulli distribution of parameter $z_i\alpha_{it}$, and the model is expressed as follows

Ecological process:

$$\begin{aligned} z_i &\sim \text{Bernoulli}(\pi_i) \ , \\ \text{logit}(\pi_i) &= \beta_0 + \sum_{m=1}^{M_1} \beta_m x_{mi}^{(1)} + w_i \ , \end{aligned} \tag{11.13}$$

Detection process:

$$\begin{aligned} y_{it} &\sim \text{Bernoulli}(z_i\,\alpha_{it}) \ , \\ \text{logit}(\alpha_{it}) &= \gamma_0 + \sum_{m=1}^{M_2} \gamma_m x_{mit}^{(2)} \ , \end{aligned} \tag{11.14}$$

where $\{\beta_0, \ldots, \beta_{M_1}\}$ and $\{\gamma_0, \ldots, \gamma_{M_2}\}$ are the parameters that quantify the linear effects of some covariates $(\boldsymbol{x}_1^{(1)}, \ldots, \boldsymbol{x}_{M_1}^{(1)})$ and $(\boldsymbol{x}_1^{(2)}, \ldots, \boldsymbol{x}_{M_2}^{(2)})$ in the ecological and observation process respectively. These covariates are usually variables refereed to site characteristics such as habitat variables or survey characteristics such as weather conditions. $\boldsymbol{w} = (w_1, \ldots, w_n)$ represents the spatial effect in the ecological process. Normally, this spatial effect is a Gaussian process that can be incorporated as geostatistical terms (in the way already introduced in Section 11.3), but other options are possible (such as CAR Normal distributions, as in Pennino et al. (2016b)). The `R`-package `hSDM`, which make inference using `MCMC`, can be used easily to fit some of these models. In addition, the `inlabru` package also handle the problem of detectability (Yuan et al., 2016).

With respect to N-mixture models, which are used for count data with imperfect detection, they implement a Poisson distribution for the ecological process, while using a Binomial distribution for the observability process (Royle and Nichols, 2003; Dodd Jr and Dorazio, 2004; Royle, 2004). The structure of the model is similar to the site-occupancy model, in particular:

Ecological process:

$$\begin{aligned} N_i &\sim \text{Poisson}(\lambda_i) \ , \\ \log(\lambda_i) &= \beta_0 + \sum_{m=1}^{M_1} \beta_m x_{mi}^{(1)} + w_i \ , \end{aligned} \tag{11.15}$$

Detection process:

$$\begin{aligned} y_{it} &\sim \text{Bernoulli}(N_i \, \alpha_{it}) \ , \\ \text{logit}(\alpha_{it}) &= \gamma_0 + \sum_{m=1}^{M_2} \gamma_m x_{mit}^{(2)} \ . \end{aligned} \tag{11.16}$$

The `R`-package `hSDM` allow us to fit some of these models. In addition, the INLA group is developing some methods to fit N-mixture models (Meehan et al., 2017).

11.4.6 Excess of zeros

The study of datasets with zero excess has an important role in the literature, particularly, in species distribution modeling (Agarwal et al., 2002; Ver Hoef and Jansen, 2007; Neelon et al., 2013), becoming highly relevant in recent years especially. Bayesian softwares like INLA already contain different functions to handle situations with zero excess. Generally, these situations are a source of overdispersion caused by a disagreement between the data and the distribution assumed: there are more zeros in the dataset than the proposed distribution could reasonably explain.

Zero-inflated models are a widely known tool for dealing with this problem. These models assume that the data follow a finite mixture of a degenerate distribution with all its mass at zero with a discrete distribution with support in $\mathbb{Z}^+ \cup \{0\}$ (Yau et al., 2003). If $1-\pi_i$ represents the probability of species presence, π_i the probability of the species absence, i.e., $p(y_i|\pi_i) = \pi_i$ and $p(y_i > 0) = 1 - \pi_i$, and h a probability mass function (pmf) of some parametric discrete distribution with support on $\mathbb{Z}^+ \cup \{0\}$, the distribution

of y_i has the following mixture density:

$$p(y_i|\pi_i, \mu_i, \boldsymbol{\psi}_1) = \pi_i \delta_0 + (1 - \pi_i) h(y_i|\mu_i, \boldsymbol{\psi}_1) \ , \tag{11.17}$$

being δ_0 the Dirac delta function, μ_i and $\boldsymbol{\psi}_1$ hyperparameters depending on h, and h is a pmf coming from a Poisson, binomial or negative-binomial (note that this latter distribution is one of those considered to account for overdispersion). The model is completed when linking π_i and μ_i with the linear predictors by means of:

$$\begin{aligned} \text{logit}(\pi_i) = \eta_i^{(1)} &= \alpha^{(1)} + \sum_{m=1}^{M^{(1)}} \beta_m^{(1)} x_{mi}^{(1)} + \sum_{l=1}^{L^{(1)}} f_l^{(1)}(z_{li}^{(1)}) \ , \\ g(\mu_i) = \eta_i^{(2)} &= \alpha^{(2)} + \sum_{m=1}^{M^{(2)}} \beta_m^{(2)} x_{mi}^{(2)} + \sum_{l=1}^{L^{(2)}} f_l^{(2)}(z_{li}^{(2)}) \ , \end{aligned} \tag{11.18}$$

where logit denotes the link function between the linear predictor $\eta_i^{(1)}$ and the probability of absence π_i, and $g(\cdot)$ is an appropriate link for the mean of h.

An alternative to these models is given by hurdle models (Mullahy, 1986; Cameron and Trivedi, 1998), where data are assumed to follow a finite mixture of a degenerate distribution with all its mass at zero and a zero truncated discrete distribution. That is, unlike the zero inflated models, in hurdle models, all observed zeros come from the zero-degenerate distribution. Following the same notation of Eq. (11.17), a hurdle model can be expressed as follows:

$$p(y_i|\pi_i, \mu_i, \boldsymbol{\psi}_1) \quad = \quad \pi_i \delta_0 + (1 - \pi_i) h(y_i|\mu_i, \boldsymbol{\psi}_1) I_{[y_i > 0]} \, . \tag{11.19}$$

As in (11.18), the hurdle model is completed when linking π_i and μ_i with their corresponding linear predictors.

However, the response variable is not always a discrete variable. Semi-continuous processes like rain, plant coverage, chemical concentrations, etc., are measured in the $[0, \infty)$ interval having high proportions of zero values, and there are neither an appropriate probability distribution nor a transformation available to fit them adequately. To model processes of this type,

an extension of hurdle models for continuous data is required (Aitchison, 1955; Quiroz et al., 2015). Again, data are modeled as two independent sub-processes: one determines whether the response is zero, and the other determines the intensity when the response is non-zero using a continuous well known distribution like the log-Normal or the Gamma (Stefánsson, 1996; Brynjarsdóttir and Stefánsson, 2004; Paradinas et al., 2017b). In this case, hurdle models are defined as a finite mixture of a degenerate distribution with point mass at zero and a distribution with support on $\mathbb{R}^+$. If h is a pdf of some parametric continuous distribution with support on $\mathbb{R}^+$ (e.g. Gamma, log-Normal or log-logistic), the hurdle model for y_i (now assumed to be a continuous distribution) has the same mixture density as in (11.19). Although there exist an extensive list of zero-inflated or hurdle models dealing with correlated discrete data in many fields (Agarwal et al., 2002; Ver Hoef and Jansen, 2007), this approach has not been widely used with continuous responses.

It is worth noting that all the models commented upon in this section are a mixture of two processes, and in almost all cases, they are modeled independently (Neelon et al., 2013; Balderama et al., 2016). However, generally both sub-processes are related: low intensities are linked to low probabilities of presence and vice versa. Shared component modeling (SCM) is a good tool to deal with it by combining information both from the two subprocesses (Paradinas et al., 2017b).

11.5 Discussion

This paper has reviewed some of the statistical challenges that can arise when the distribution of the species is modeled using geostatistical or point-referenced data. In particular, after describing in detail data and methods commonly used to model species distribution, we have focused on complex issues and we have discussed how they can be solved using Bayesian hierarchical spatio-temporal models. Specifically, in this review we have focused on the Bayesian approach and the INLA methodology (Rue et al., 2009) because they have several benefits with respect to the classic geostatistical methods. INLA makes it possible to perform complex models with

a minimum computational effort while obtaining accurate estimates. Its importance in the context of SDMs can be even more appreciated with the appearance of the recent project `inlabru` which has been created to develop and implement innovative methods to model spatial distribution and change from ecological survey data (`https://sites.google.com/inlabru3.org/inlabru`). In addition, classic geostatistical methods typically overestimate their predictive accuracy by using plug-in estimations of parameters in their predictive equations. (Diggle and Ribeiro, 2007). On the contrary, inference about uncertainty, based on the observations and models, is a byproduct of the model predictions when the Bayesian framework is employed.

However, some limitations can arise when the INLA approach is used. For example, INLA can not handle missing values in spatially structured covariates. This issue can be framed in the misalignment problem discussed in Section 11.4.3; this means that it could be overcome by applying a two-stage or joint modeling approach that allows prediction of the covariate values in the locations where they were not measured. As mentioned above, an alternative is the one proposed by Gómez-Rubio and Rue (2017) that, using a more general approach, deals with missing values in the covariates, based on fitting conditional latent Gaussian models where covariates are imputed using a Metropolis-Hastings algorithm.

We would like to remark that, due to space limitations, we have not fully reviewed the several complications that can derive from the sampling process. Indeed, we have only focused on the preferential sampling problem (Diggle et al., 2010), which, as previously mentioned, refers to the possibility that the sample design is stochastically dependent on the studied process. Nevertheless, other types of sampling procedures could produce different issues that should be taken into account in the statistical analysis. For example, one of the most popular methods used in ecology to estimate an animal population's size is the capture-recapture method that involves capturing, marking and releasing an initial sample of individuals (Otis et al., 1978; McInerny and Purves, 2011). Subsequently, a second sample of animal individuals is obtained independently and it is noted how many of them in that sample were marked. To model data of this type, a feasible solution could be the implementation of Bayesian hierarchical N-mixture models

described in Section 11.4.5, which are currently being developed in INLA (Meehan et al., 2017).

Finally, an important point to consider is that INLA is not the only computational approach to making inference for Bayesian spatio-temporal models. In recent years, other approaches that also make it possible to achieve accurate species distribution models results, such as `stan` (Stan Development Team, 2015; Monnahan et al., 2017), have been widely used.

References

Agarwal, D. K., Gelfand, A. E., and Citron-Pousty, S. (2002). Zero-inflated models with application to spatial count data. *Environmental and Ecological Statistics*, 9(4):341–355.

Aitchison, J. (1955). On the distribution of a positive random variable having a discrete probability mass at the origin. *Journal of the American Statistical Association*, 50(271):901–908.

Aizpurua, O., Paquet, J.-Y., Brotons, L., and Titeux, N. (2015). Optimising long-term monitoring projects for species distribution modelling: how atlas data may help. *Ecography*, 38(1):29–40.

Anatolyev, S. and Kosenok, G. (2005). An alternative to maximum likelihood based on spacings. *Econometric Theory*, 21(2):472–476.

Andreon, S. and Weaver, B. (2015). *Bayesian Methods for the Physical Sciences: Learning from Examples in Astronomy and Physics*, volume 4. Springer Series in Astrostatistics.

Araújo, M. B., Pearson, R. G., Thuiller, W., and Erhard, M. (2005). Validation of species-climate impact models under climate change. *Global Change Biology*, 11(9):1504–1513.

Baio, G. (2012). *Bayesian Methods in Health Economics*. CRC Chapman and Hall.

Bakka, H., Vanhatalo, J., Illian, J., Simpson, D., and Rue, H. (2016). Accounting for physical barriers in species distribution modeling with non-stationary spatial random effects. *ArXiv e-prints*.

Balderama, E., Gardner, B., and Reich, B. J. (2016). A spatial-temporal double-hurdle model for extremely over-dispersed avian count data. *Spatial Statistics*, 18:263–275.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*. CRC.

Barber, X., Conesa, D., Lladosa, S., and López-Quílez, A. (2016). Modelling the presence of disease under spatial misalignment using Bayesian latent Gaussian models. *Geospatial Health*, 11:415.

Barber, X., Conesa, D., López-Quílez, A., Mayoral, A., Morales, J., and Barber, A. (2017). Bayesian hierarchical models for analysing the spatial distribution of bioclimatic indices. *SORT-Statistics and Operations Research Transactions*, 1(2):277–296.

Barbet-Massin, M., Jiguet, F., Albert, C. H., and Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution*, 3(2):327–338.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 160(901):268–282.

Beale, C. M., Lennon, J. J., Yearsley, J. M., Brewer, M. J., and Elston, D. A. (2010). Regression analysis of spatial data. *Ecology letters*, 13(2):246–264.

Berry, D. A. and Stangl, D. (1999). *Bayesian Biostatistics*. Marcel Dekker.

Blangiardo, M. and Cameletti, M. (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons.

Blangiardo, M., Cameletti, M., Baio, G., and Rue, H. (2013). Spatial and spatio-temporal models with R-INLA. *Spatial and Spatio-temporal Epidemiology*, 7:39–55.

Bowman, K. and Shenton, L. (2006). Estimation: Method of moments. *Encyclopedia of statistical sciences*.

Brezger, A., Kneib, T., and Lang, S. (2003). BayesX: Analysing Bayesian structured additive regression models. Technical report, Discussion paper//Sonderforschungsbereich 386 der Ludwig-Maximilians-Universität München.

Brooks, S., Gelman, A., Jones, G. L., and Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. CRC Press, Taylor & Francis Group.

Brown, C. J., O'connor, M. I., Poloczanska, E. S., Schoeman, D. S., Buckley, L. B., Burrows, M. T., Duarte, C. M., Halpern, B. S., Pandolfi, J. M., Parmesan, C., and Richardson, A. J. (2016). Ecological and methodological drivers of species distribution and phenology responses to climate change. *Global Change Biology*, 22:1548–1560.

Brown, P. (2015). Model-based geostatistics the easy way. *Journal of Statistical Software*, 63:1–24.

Brunsdon, C., Fotheringham, S., and Charlton, M. (1998). Geographically weighted regression. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(3):431–443.

Brynjarsdóttir, J. and Stefánsson, G. (2004). Analysis of cod catch data from Icelandic groundfish surveys using generalized linear models. *Fisheries Research*, 70(2):195–208.

Busby, J. (1991). Bioclim-a bioclimate analysis and prediction system. *Plant protection quarterly (Australia)*.

Cameletti, M., Ignaccolo, R., and Bande, S. (2011). Comparing spatio-temporal models for particulate matter in Piemonte. *Environmetrics*, 22(8):985–996.

Cameletti, M., Lindgren, F., Simpson, D., and Rue, H. (2013). Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Advances in Statistical Analysis*, 97(2):109–131.

Cameron, C. A. and Trivedi, P. K. (1998). *Regression Analysis of Count Data*. Cambridge University Press, New York.

Carpenter, G., Gillison, A., and Winter, J. (1993). DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity and Conservation*, 2(6):667–680.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, L. A. (2016). *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*. Chapman and Hall/CRC.

Clark, J. and Gelfand, A. (2006). *Hierarchical Modeling for the Environmental Sciences. Statistical Methods and Applications*. Oxford University Press, New York.

Clark, J. S., Gelfand, A. E., Woodall, C. W., and Zhu, K. (2014). More than the sum of the parts: forest climate response from joint species distribution models. *Ecological Applications*, 24(5):990–999.

Cosandey-Godin, A., Krainski, E. T., Worm, B., and Flemming, J. M. (2015). Applying Bayesian spatio-temporal models to fisheries bycatch in the Canadian Arctic. *Canadian Journal of Fisheries and Aquatic Sciences*, 72(2):186–197.

Cox, D. R. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91(3):729–737.

Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. Wiley.

Daly, C. (2006). Guidelines for assessing the suitability of spatial climate data sets. *International Journal of Climatology*, 26(6):707–721.

Danks, F. and Klein, D. (2002). Using GIS to predict potential wildlife habitat: a case study of muskoxen in northern Alaska. *International Journal of Remote Sensing*, 23(21):4611–4632.

Dettmers, R., Buehler, D. A., Bartlett, J. G., and Klaus, N. A. (1999). Influence of point count length and repeated visits on habitat model performance. *The Journal of wildlife management*, pages 815–823.

Diggle, P. J. (2013). *Statistical Analysis of Spatial and Spatio-temporal Point Patterns*. CRC.

Diggle, P. J., Menezes, R., and Su, T.-l. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2):191–232.

Diggle, P. J. and Ribeiro, P. J. (2007). *Model-based Geostatistics*. Springer.

Dodd Jr, C. K. and Dorazio, R. M. (2004). Using counts to simultaneously estimate abundance and detection probabilities in a salamander community. *Herpetologica*, 60(4):468–478.

Dorazio, R. M., Royle, J. A., Söderström, B., and Glimskär, A. (2006). Estimating species richness and accumulation by modeling species occurrence and detectability. *Ecology*, 87(4):842–854.

Elith, J. and 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.

Fatima, S. H., Atif, S., Rasheed, S. B., Zaidi, F., and Hussain, E. (2016). Species distribution modelling of *Aedes aegypti* in two dengue-endemic regions of Pakistan. *Tropical Medicine & International Health*.

Ferrari, S. L. P. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799 – 815.

Finley, A. O. (2011). Comparing spatially-varying coefficients models for analysis of ecological data with non-stationary and anisotropic residual dependence. *Methods in Ecology and Evolution*, 2(2):143–154.

Fitzpatrick, M. C., Weltzin, J. F., Sanders, N. J., and Dunn, R. R. (2007). The biogeography of prediction error: why does the introduced range of the fire ant over-predict its native range? *Global Ecology and Biogeography*, 16(1):24–33.

Fortin, M.-J. and Dale, M. R. (2005). *Spatial Analysis: A Guide for Ecologists*. Cambridge University Press.

Foster, S. D., Shimadzu, H., and Darnell, R. (2012). Uncertainty in spatially predicted covariates: is it ignorable? *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(4):637–652.

Franklin, J. (2010). *Mapping Species Distributions: Spatial Inference and Prediction*. Cambridge University Press.

Gaudard, M., Karson, M., Linder, E., and Sinha, D. (1999). Bayesian spatial prediction. *Environmental and Ecological Statistics*, 6(2):147–171.

Gelfand, A. E., Diggle, P. J., Fuentes, M., and Guttorp, P. (2010). *Handbook of Spatial Statistics*. Chapman & Hall.

Gelfand, A. E., Kim, H.-J., Sirmans, C. F., and Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98(462):387–396.

Gelfand, A. E., Silander, J. A., Wu, S., Latimer, A., Lewis, P. O., Rebelo, A. G., Holder, M., et al. (2006). Explaining species distribution patterns through hierarchical modeling. *Bayesian Analysis*, 1(1):41–92.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian Data Analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA.

Gitzen, R. A. (2012). *Design and Analysis of Long-term Ecological Monitoring Studies*. Cambridge University Press.

Goetz, S. J., Sun, M., Zolkos, S., Hansen, A., and Dubayah, R. (2014). The relative importance of climate and vegetation properties on patterns of North American breeding bird species richness. *Environmental Research Letters*, 9(3):034013.

Golding, N. and Purse, B. V. (2016). Fast and flexible bayesian species distribution modelling using gaussian processes. *Methods in Ecology and Evolution*, 7(5):598–608.

Gómez-Rubio, V., Bivand, R. S., and Rue, H. (2014). Spatial models using Laplace approximation methods. In *Handbook of Regional Science*, pages 1401–1417. Springer.

Gómez-Rubio, V. and Rue, H. (2017). Markov Chain Monte Carlo with the Integrated Nested Laplace Approximation. *ArXiv e-prints*.

González-Warleta, M., Lladosa, S., Castro-Hermida, J. A., Martínez-Ibeas, A. M., Conesa, D., Muñoz, F., López-Quílez, A., Manga-González, Y., and Mezo, M. (2013). Bovine paramphistomosis in Galicia (Spain): Prevalence, intensity, aetiology and geospatial distribution of the infection. *Veterinary parasitology*, 191(3):252–263.

Gosoniu, L., Vounatsou, P., Sogoba, N., and Smith, T. (2006). Bayesian modelling of geostatistical malaria risk data. *Geospatial health*, 1(1):127–139.

Gotelli, N. J., Anderson, M. J., Arita, H. T., Chao, A., Colwell, R. K., Connolly, S. R., Currie, D. J., Dunn, R. R., Graves, G. R., Green, J. L., et al. (2009). Patterns and causes of species richness: a general simulation model for macroecology. *Ecology Letters*, 12(9):873–886.

Griffith, D. A. (2008). Spatial-filtering-based contributions to a critique of geographically weighted regression (GWR). *Environment and Planning A*, 40(11):2751–2769.

Gu, W. and Swihart, R. K. (2004). Absent or undetected? effects of non-detection of species occurrence on wildlife–habitat models. *Biological Conservation*, 116(2):195–203.

Guisan, A., Edwards, T. C., and Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling*, 157(2):89–100.

Guisan, A. and Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, 8(9):993–1009.

Handcock, M. S. and Stein, M. L. (1993). A Bayesian analysis of kriging. *Technometrics*, 35(4):403–410.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, 50(4):1029–1054.

He, K. S., Bradley, B. A., Cord, A. F., Rocchini, D., Tuanmu, M.-N., Schmidtlein, S., Turner, W., Wegmann, M., and Pettorelli, N. (2015). Will remote sensing shape the next generation of species distribution models? *Remote Sensing in Ecology and Conservation*, 1(1):4–18.

Hefley, T. J., Broms, K. M., Brost, B. M., Buderman, F. E., Kay, S. L., Scharf, H. R., Tipton, J. R., Williams, P. J., and Hooten, M. B. (2017a). The basis function approach for modeling autocorrelation in ecological data. *Ecology*, 98(3):632–646.

Hefley, T. J. and Hooten, M. B. (2016). Hierarchical species distribution models. *Current Landscape Ecology Reports*, 1(2):87–97.

Hefley, T. J., Hooten, M. B., Hanks, E. M., Russell, R. E., and Walsh, D. P. (2017b). Dynamic spatio-temporal models for spatial data. *Spatial Statistics*, 20:206–220.

Hendricks, S. A., Clee, P. R. S., Harrigan, R. J., Pollinger, J. P., Freedman, A. H., Callas, R., Figura, P. J., and Wayne, R. K. (2016). Re-defining historical geographic range in species with sparse records: implications for the Mexican wolf reintroduction program. *Biological Conservation*, 194:48–57.

Hengl, T., Heuvelink, G. B., Tadić, M. P., and Pebesma, E. J. (2012). Spatio-temporal prediction of daily temperatures using time-series of MODIS LST images. *Theoretical and applied climatology*, 107(1-2):265–277.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25(15):1965–1978.

Hoeting, J. A., Leecaster, M., and Bowden, D. (2000). An improved model for spatially correlated binary responses. *Journal of agricultural, biological, and environmental statistics*, pages 102–114.

Holloway, P. and Miller, J. A. (2015). Exploring spatial scale, autocorrelation and nonstationarity of bird species richness patterns. *ISPRS International Journal of Geo-Information*, 4(2):783–798.

Hooten, M. B. and Wikle, C. K. (2008). A hierarchical Bayesian non-linear spatio-temporal model for the spread of invasive species with application to the Eurasian Collared-Dove. *Environmental and Ecological Statistics*, 15(1):59–70.

Hooten, M. B., Wikle, C. K., Dorazio, R. M., and Royle, J. A. (2007). Hierarchical spatiotemporal matrix models for characterizing invasions. *Biometrics*, 63(2):558–567.

Hui, F. K. (2017). Model-based simultaneous clustering and ordination of multivariate abundance data in ecology. *Computational Statistics & Data Analysis*, 105:1–10.

Illian, J. B., Sørbye, S. H., and Rue, H. (2012). A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (`INLA`). *The Annals of Applied Statistics*, pages 1499–1530.

Ingebrigtsen, R., Lindgren, F., and Steinsland, I. (2014). Spatial models with explanatory variables in the dependence structure. *Spatial Statistics*, 8:20–38.

Iturbide, M., Bedia, J., Herrera, S., del Hierro, O., Pinto, M., and Gutiérrez, J. M. (2015). A framework for species distribution modelling with improved pseudo-absence generation. *Ecological Modelling*, 312:166–174.

Iverson, L. R., Schwartz, M. W., and Prasad, A. M. (2004). How fast and far might tree species migrate in the eastern united states due to climate change? *Global Ecology and Biogeography*, 13(3):209–219.

Jackman, S. (2009). *Bayesian Analysis for the Social Sciences*. John Wiley & Sons.

Jiménez-Valverde, A. and Lobo, J. M. (2007). Determinants of local spider (*Araneidae* and *Thomisidae*) species richness on a regional scale: climate and altitude vs. habitat structure. *Ecological Entomology*, 32(1):113–122.

Johnson, D. S., Conn, P. B., Hooten, M. B., Ray, J. C., and Pond, B. A. (2013). Spatial occupancy models for large data sets. *Ecology*, 94(4):801–808.

Jona Lasinio, G., Mastrantonio, G., and Pollice, A. (2012). Discussing the "big n problem". *Statistical Methods & Applications*, pages 1–16.

Joseph, L. N., Field, S. A., Wilcox, C., and Possingham, H. P. (2006). Presence-absence versus abundance data for monitoring threatened species. *Conservation Biology*, 20(6):1679–1687.

Juan, P., Díaz-Avalos, C., Mejía-Domínguez, N. R., and Mateu, J. (2017). Hierarchical spatial modeling of the presence of chagas disease insect vectors in argentina. a comparative approach. *Stochastic environmental research and risk assessment*, 31(2):461–479.

Karagiannis-Voules, D.-A., Scholte, R. G., Guimarães, L. H., Utzinger, J., and Vounatsou, P. (2013). Bayesian geostatistical modeling of leishmaniasis incidence in Brazil. *PLOS Neglected Tropical Diseases*, 7(5):e2213.

Kneib, T., Müller, J., and Hothorn, T. (2008). Spatial smoothing techniques for the assessment of habitat suitability. *Environmental and Ecological Statistics*, 15(3):343–364.

Kozak, K. H., Graham, C. H., and Wiens, J. J. (2008). Integrating GIS-based environmental data into evolutionary biology. *Trends in Ecology & Evolution*, 23(3):141–148.

Krainski, E. T., Lindgren, F., Simpson, D., and Rue, H. (2017). The `R-INLA` tutorial: SPDE models.

Latimer, A. M., Wu, S., Gelfand, A. E., and Silander, J. A. (2006). Building statistical models to analyze species distributions. *Ecological applications*, 16(1):33–50.

Le Cam, L. (1990). Maximum likelihood: an introduction. *International Statistical Review/Revue Internationale de Statistique*, pages 153–171.

Leathwick, J., Rowe, D., Richardson, J., Elith, J., and Hastie, T. (2005). Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. *Freshwater Biology*, 50(12):2034–2052.

Lee, D. (2013). CARBayes: an R package for Bayesian spatial modeling with conditional autoregressive priors. *Journal of Statistical Software*, 55(13):1–24.

Lindgren, F. and Rue, H. (2008). On the second-order random walk model for irregular locations. *Scandinavian journal of statistics*, 35(4):691–700.

Lindgren, F. and Rue, H. (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software, Articles*, 63(19):1–25.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.

Liu, C., Wan, R., Jiao, Y., and Reid, K. B. (2017). Exploring non-stationary and scale-dependent relationships between walleye (*Sander vitreus*) distribution and habitat variables in lake erie. *Marine and Freshwater Research*, 68(2):270–281.

Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). The BUGS project: evolution, critique and future directions. *Statistics in Medicine*, 28(25):3049–3067.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, 10(4):325–337.

Luo, M. and Opaluch, J. J. (2011). Analyze the risks of biological invasion. *Stochastic environmental research and risk assessment*, 25(3):377–388.

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Andrew Royle, J., and Langtimm, C. A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83(8):2248–2255.

Mallick, B. K., Gold, D., and Baladandayuthapani, V. (2009). *Bayesian Analysis of Gene Expression Data*. Wiley.

Martínez-Bello, D., López-Quílez, A., and Prieto, A. T. (2017). Spatiotemporal modeling of relative risk of dengue disease in colombia. *Stochastic Environmental Research and Risk Assessment*.

Martinez-Meyer, E., Peterson, A. T., Servín, J. I., and Kiff, L. F. (2006). Ecological niche modelling and prioritizing areas for species reintroductions. *Oryx*, 40(4):411–418.

Martínez-Minaya, J., Conesa, D., López-Quílez, A., and Vicent, A. (2018). Spatial and climatic factors associated with the geographical distribution of citrus black spot disease in south africa. a bayesian latent gaussian model approach. *European Journal of Plant Pathology*, 151:991—-1007.

Martino, S., Akerkar, R., and Rue, H. (2011). Approximate Bayesian inference for survival models. *Scandinavian Journal of Statistics*, 38(3):514–528.

Martins, T. G., Simpson, D., Lindgren, F., and Rue, H. (2013). Bayesian computing with `inla`: new features. *Computational Statistics & Data Analysis*, 67:68–83.

McCarthy, M. A. (2007). *Bayesian Methods for Ecology*. John Wiley & Sons.

McInerny, G. J. and Purves, D. W. (2011). Fine-scale environmental variation in species distribution modelling: regression dilution, latent variables and neighbourly advice. *Methods in Ecology and Evolution*, 2(3):248–257.

Meehan, T. D., Michel, N. L., and Rue, H. (2017). Estimating animal abundance with N-mixture models using the `R-INLA` package for `R`. *ArXiv e-prints*.

Meentemeyer, R. K., Cunniffe, N. J., Cook, A. R., Filipe, J. A., Hunter, R. D., Rizzo, D. M., and Gilligan, C. A. (2011). Epidemiological modeling of invasion in heterogeneous landscapes: spread of sudden oak death in California (1990–2030). *Ecosphere*, 2(2):1–24.

Miller, J. A. (2012). Species distribution models. *Progress in Physical Geography*, 36(5):681–692.

Monnahan, C. C., Thorson, J. T., and Branch, T. A. (2017). Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in Ecology and Evolution*, 8(3):339–348.

Muñoz, F., Pennino, M. G., Conesa, D., López-Quílez, A., and Bellido, J. M. (2013). Estimation and prediction of the spatial occurrence of fish species using Bayesian latent Gaussian models. *Stochastic Environmental Research and Risk Assessment*, 27(5):1171–1180.

Muff, S., Riebler, A., Held, L., Rue, H., and Saner, P. (2015). Bayesian analysis of measurement error models using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(2):231–252.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3):341–365.

Neelon, B., Ghosh, P., and Loebs, P. F. (2013). A spatial Poisson hurdle model for exploring geographic variation in emergency department visits. *Journal of the Royal Statistical Society: Series A*, 176(2):389–413.

Neri, F. M., Cook, A. R., Gibson, G. J., Gottwald, T. R., and Gilligan, C. A. (2014). Bayesian analysis for inference of an emerging epidemic: citrus canker in urban landscapes. *PLOS Computational Biology*, 10(4).

New, M., Lister, D., Hulme, M., and Makin, I. (2002). A high-resolution data set of surface climate over global land areas. *Climate research*, 21(1):1–25.

Nichols, J. D., Hines, J. E., Sauer, J. R., Fallon, F. W., Fallon, J. E., and Heglund, P. J. (2000). A double-observer approach for estimating detection probability and abundance from point counts. *The Auk*, 117(2):393–408.

Nielsen, S. E., Johnson, C. J., Heard, D. C., and Boyce, M. S. (2005). Can models of presence-absence be used to scale abundance? two case studies considering extremes in life history. *Ecography*, 28(2):197–208.

Otis, D. L., Burnham, K. P., White, G. C., and Anderson, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife monographs*, (62):3–135.

Paradinas, I., Conesa, D., López-Quílez, A., and Bellido, J. M. (2017a). Spatio-Temporal model structures with shared components for semi-continuous species distribution modelling. *Spatial Statistics*, page in press.

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

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

Paradinas, I., Pennino, M. G., López-Quílez, A., Marín, M., Bellido, J. M., and Conesa, D. (2017b). Modelling spatially sampled proportion processes. *RevStat*, page in press.

Park, Y.-S., Céréghino, R., Compin, A., and Lek, S. (2003). Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters. *Ecological Modelling*, 160(3):265–280.

Parviainen, M., Luoto, M., Ryttäri, T., and Heikkinen, R. K. (2008). Modelling the occurrence of threatened plant species in taiga landscapes: methodological and ecological perspectives. *Journal of Biogeography*, 35(10):1888–1905.

Pennino, M. G., Conesa, D., López-Quílez, A., Muñoz, F., Fernández, A., and Bellido, J. M. (2016a). Fishery-dependent and-independent data lead to consistent estimations of essential habitats. *ICES Journal of Marine Science: Journal du Conseil*, 73(9):2302–2310.

Pennino, M. G., Mérigot, B., Fonseca, V. P., Monni, V., and Rotta, A. (2016b). Habitat modeling for cetacean management: Spatial distribution in the southern Pelagos Sanctuary (Mediterranean sea). *Deep Sea Research Part II: Topical Studies in Oceanography*.

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

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

Pennino, M. G., Paradinas, I., Illian, J. B., Muñoz, F., Bellido, J. M., López-Quílez, A., and Conesa, D. (2017). Accounting for preferential sampling in species distribution models. submitted.

Peterson, A. T., Sánchez-Cordero, V., Beard, C. B., and Ramsey, J. M. (2002). Ecologic niche modeling and potential reservoirs for chagas disease, Mexico. *Emerging Infectious Diseases*, 8(7):662–667.

Phillips, S. J., Anderson, R. P., and Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3):231–259.

Plummer, M. (2016). `Rjags`: Bayesian graphical models using MCMC. `R` *Software Package for Graphical Models*.

Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., Vesk, P. A., and McCarthy, M. A. (2014). Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution*, 5(5):397–406.

Quiroz, Z. C., Prates, M. O., and Rue, H. (2015). A Bayesian approach to estimate the biomass of anchovies off the coast of Perú. *Biometrics*, 71(1):208–217.

Rachev, S. T., Hsu, J. S., Bagasheva, B. S., and Fabozzi, F. J. (2008). *Bayesian methods in finance*, volume 153. John Wiley & Sons.

Risser, M. D. (2016). Review: Nonstationary Spatial Modeling, with Emphasis on Process Convolution and Covariate-Driven Approaches. *ArXiv e-prints*.

Robert, C. and Casella, G. (2011). A short history of Markov Chain Monte Carlo: subjective recollections from incomplete data. *Statistical Science*, 26(1):102–115.

Rodríguez de Rivera, Ó. and López-Quílez, A. (2017). Development and comparison of species distribution models for forest inventories. *ISPRS International Journal of Geo-Information*, 6(6):176.

Roos, N. C., Carvalho, A. R., Lopes, P. F., and Pennino, M. G. (2015). Modeling sensitive parrotfish (Labridae: Scarini) habitats along the Brazilian coast. *Marine Environmental Research*, 110:92–100.

Royle, J. A. (2004). N-mixture models for estimating population size from spatially replicated counts. *Biometrics*, 60(1):108–115.

Royle, J. A. and Nichols, J. D. (2003). Estimating abundance from repeated presence-absence data or point counts. *Ecology*, 84(3):777–790.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall.

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

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2017). Bayesian computing with `INLA`: a review. *Annual Review of Statistics and Its Application*, 4:395–421.

Rufener, M.-C., Kinas, P. G., Nóbrega, M. F., and Oliveira, J. E. L. (2017). Bayesian spatial predictive models for data-poor fisheries. *Ecological Modelling*, 348:125–134.

Ruiz-Cárdenas, R., Krainski, E. T., and Rue, H. (2012). Direct fitting of dynamic models using integrated nested Laplace approximations-`INLA`. *Computational Statistics & Data Analysis*, 56(6):1808–1828.

Sadykova, D., Scott, B. E., De Dominicis, M., Wakelin, S. L., Sadykov, A., and Wolf, J. (2017). Bayesian joint models with inla exploring marine mobile predator–prey and competitor species habitat overlap. *Ecology and evolution*, 7(14):5212–5226.

Sbrocco, E. J. and Barber, P. H. (2013). MARSPEC: ocean climate layers for marine spatial ecology. *Ecology*, 94(4):979–979.

Schrödle, B., Held, L., Riebler, A., and Danuser, J. (2011). Using integrated nested Laplace approximations for the evaluation of veterinary

surveillance data from Switzerland: a case-study. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(2):261–279.

Shapiro, A. (2000). On the asymptotics of constrained local M-estimators. *Annals of Statistics*, 28(3):948–960.

Stan Development Team (2015). Stan Modeling Language: User's Guide and Reference Manual.

Stefánsson, G. (1996). Analysis of groundfish survey abundance data: combining the GLM and delta approaches. *ICES Journal of Marine Science*, 53(3):577–588.

Stein, A., Kocks, C., Zadoks, J., Frinking, H., Ruissen, M., and Myers, D. (1994). A geostatistical analysis of the spatio-temporal development of downy mildew epidemics in cabbage. *Phytopathology*, 84(10):1227–1238.

Stein, M. (1999). *Interpolation of Spatial Data. Some Theory for Kriging.* Springer.

Stoklosa, J., Daly, C., Foster, S. D., Ashcroft, M. B., and Warton, D. I. (2015). A climate of uncertainty: accounting for error in climate variables for species distribution models. *Methods in Ecology and Evolution*, 6(4):412–423.

Taylor-Rodriguez, D., Kaufeld, K., Schliep, E. M., Clark, J. S., Gelfand, A. E., et al. (2017). Joint species distribution modeling: dimension reduction using Dirichlet processes. *Bayesian Analysis*, 12(4):939–967.

Václavík, T. and Meentemeyer, R. K. (2009). Invasive species distribution modeling (iSDM): Are absence data and dispersal constraints needed to predict actual distributions? *Ecological Modelling*, 220(23):3248–3258.

Ver Hoef, J. M. and Jansen, J. K. (2007). Space-time zero-inflated count models of Harbor seals. *Environmetrics*, 18(7):697–712.

Vieilledent, G., Latimer, A., Gelfand, A., Merow, C., Wilson, A., Mortier, F., and Silander Jr, J. (2014). hSDM: hierarchical Bayesian species distribution models. `R` *package version*, 1.

White, S. M., Bullock, J. M., Hooftman, D. A., and Chapman, D. S. (2017). Modelling the spread and control of *Xylella fastidiosa* in the early stages of invasion in Apulia, Italy. *Biological Invasions*, pages 1–13.

Wikle, C. K. (2003). Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology*, 84(6):1382–1394.

Wikle, C. K. and Hooten, M. B. (2010). A general science-based framework for dynamical spatio-temporal models. *Test*, 19(3):417–451.

Williams, P. J., Hooten, M. B., Womble, J. N., Esslinger, G. G., Bower, M. R., and Hefley, T. J. (2017). An integrated data model to estimate spatiotemporal occupancy, abundance, and colonization dynamics. *Ecology*, 98(2):328–336.

Windle, M. J. S., Rose, G. A., Devillers, R., and Fortin, M.-J. (2010). Exploring spatial non-stationarity of fisheries survey data using geographically weighted regression (GWR): an example from the Northwest Atlantic. *ICES Journal of Marine Science*, 67(1):145.

Yau, K. K., Wang, K., and Lee, A. H. (2003). Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal*, 45(4):437–452.

Yuan, Y., Bachl, F., Lindgren, F., Brochers, D., Illian, J., Buckland, S., Rue, H., and Gerrodette, T. (2016). Point process models for spatio-temporal distance sampling data. *arXiv:1604.06013*.

Zhang, W. (2007). Supervised neural network recognition of habitat zones of rice invertebrates. *Stochastic Environmental Research and Risk Assessment*, 21(6):729–735.

Zhang W, Zhong X, L. G. . (2008). Recognizing spatial distribution patterns of grassland insects: neural network approaches. *Stochastic Environmental Research and Risk Assessment*, 22(2):207–216.

Chapter 12

Modeling Dirichlet likelihoods using the integrated nested Laplace approximation (INLA)

In this chapter, we present the actual version of our paper "Modeling Dirichlet likelihoods using the integrated nested Laplace approximation (INLA)" by Joaquín Martínez-Minaya (University of Valencia), Finn Lindgren (University of Edinburgh), Antonio López-Quílez (University of Valencia), Daniel Simpson (University of Toronto) and David Conesa (University of Valencia). In order to keep the same structure of the chapters with published papers, this chapter ends with the references used in this work.

Abstract

Dirichlet regression models can be used to analyze a set of variables lying in a bounded interval that sum up to one exhibiting skewness and heteroscedasticity, without having to transform the data. These data which mainly consist of proportions or percentages of disjoint categories are widely known as compositional data and are common in areas such as ecology, geology, and psychology. Bayesian inference has become a popular tool to deal with complex models, but numerical approaches for this are needed. Markov chain Monte Carlo (MCMC) methods have seen widespread use, however,

the integrated nested Laplace approximation (INLA) is an alternative to MCMC, that for a large class of models provides higher accuracy for a limited computational budget. However, the implemented `R-INLA` package can not deal with multivariate likelihoods, such as, in particular, the Dirichlet likelihood. In this work, we propose an expansion of the INLA method for Dirichlet regression, giving a theoretical foundation and describing the implementation as well as the application of Dirichlet regression. This method is being implemented in the package `dirinla` in the R-language.

Keywords

Hierarchical Bayesian models, INLA, Dirichlet regression, multivariate likelihood

12.1 Introduction

Compositional data (Aitchison and Egozcue, 2005), consisting of proportions or percentages of disjoint categories adding to one, play an important role in many fields such as ecology, geology, etc. Due to the complexity in some cases, some form of statistical analysis is essential for the adequate investigation and interpretation of the data.

In the literature, different approaches have been presented to deal with them, but the first clear and unified approximation was presented by Aitchison (1986), whose methods were based in the idea that "information in compositional vectors is concerned with relative, not absolute magnitudes", and the use of logratios emerged as the preferred method to deal with the unit-sum constraint (Aitchison, 1981, 1982, 1983, 1984).

However, logratio analysis is not the only methodology to deal with compositional data, the Dirichlet regression was presented as an alternative to deal with them (Hijazi and Jernigan, 2009). If the number of categories are two, the model used is well known as beta regression. Both of them have been used in many papers yielding reasonable and interpretable results (Ferrari and Cribari-Neto, 2004; Hijazi and Jernigan, 2009). In addition,

different packages have been made in R (R Core Team, 2018) in order to analyze compositional data using beta regression and Dirichlet regression (Cribari-Neto and Zeileis, 2010; Maier, 2014). All of these packages have been constructed using frequentist methods.

But, in the last years, Bayesian inference has become a good tool to deal with complex models, but as usual in Bayesian inference, numerical approaches are needed. Markov chain Monte Carlo (MCMC) methods have been so popular, however, the integrated nested Laplace approximation (INLA) methodology (Rue et al., 2009) and software `R-INLA` (http://www.r-inla.org) has become an alternative to MCMC, guaranteeing a higher computing speed for the particular case of latent Gaussian models (LGMs).

Until now, different Bayesian approaches have been developed in order to deal with compositional data using Dirichlet regression models. These methods have been implemented in different R-packages: `BayesX` (Klein et al., 2015), `Stan` (Sennhenn-Reulen, 2018), `Bugs` (van der Merwe, 2018) or `R-JAGS` (Plummer, 2016). However, `R-INLA` does not allow to deal with compositional data when the number of categories is bigger than 2. In this work, we present a way to deal with this kind of data using the INLA methodology which has been implemented in the R-package `dirinla`.

The remaining of the paper is structured as follows. Section 12.2 introduces basics of the Dirichlet regression. Section 12.3 gives a basic understanding about LGMs and the INLA methodology. In Section 12.4, the new approach is depicted followed by an introduction to the package `dirinla` in Section 12.5. In Sections 12.6 and 12.7 simulations and real data examples are conducted to compare with `R-JAGS`. Finally, Section 12.8 concludes.

12.2 Hierarchical Dirichlet regression

12.2.1 Dirichlet distribution

The Dirichlet distribution is the generalization of the widely known beta distribution, and it is defined by the following probability density

$$p(\boldsymbol{y} \mid \boldsymbol{\alpha}) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{c=1}^{C} y_c^{\alpha_c - 1}, \tag{12.1}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_C)$ is known as the vector of shape parameters for each category, $\alpha_c > 0 \ \forall c$, $y_c \in (0, 1)$, $\sum_{c=1}^{C} y_c = 1$, and $\mathrm{B}(\boldsymbol{\alpha})$ is the multinomial beta function, which serves as the normalizing constant. The multinomial beta function is defined as $\prod_{c=1}^{C} \Gamma(\alpha_c) / \Gamma(\sum_{c=1}^{C} \alpha_c)$. The sum of all αs, i.e., $\alpha_0 = \sum_{c=1}^{C} \alpha_c$ is usually interpreted as a precision parameter. Beta distribution is the particular case when $C = 2$. In addition, each variable is marginally beta-distributed with $\alpha = \alpha_c$ and $\beta = \alpha_0 - \alpha_c$.

Let $\boldsymbol{y} \sim \mathcal{D}(\boldsymbol{\alpha})$ denote a realization of a variable which is Dirichlet-distributed. The expected values are $\mathrm{E}(y_c) = \alpha_c / \alpha_0$, the variances are $\mathrm{Var}(y_c) = [\alpha_c(\alpha_0 - \alpha_c)] / [\alpha_0^2(\alpha_0 + 1)]$ and the covariances are $\mathrm{Cov}(y_c, y_{c'}) = (-\alpha_c \alpha_{c'}) / [\alpha_0^2(\alpha_0 + 1)]$.

12.2.1.1 Dealing with zeros and ones

The Dirichlet variable is defined in the open interval $(0, 1)$, nevertheless, data may come from the interval $[0, 1]$. In order to deal with this issue, a transformation was proposed in Smithson and Verkuilen (2006) to deal with zeros and ones in beta distributions, and posteriorly extended to Dirichlet distributions in Maier (2014).

$$\boldsymbol{y}^* = \frac{\boldsymbol{y}(N - 1) + 1/C}{N}. \tag{12.2}$$

This transformation compresses the data symmetrically around 0.5 from a range of $m = 1$ to $(N - 1)/N$, so extreme values are affected more than values lying close to $1/2$. Additionally, as it is pointed out in Maier (2014), if $N \to \infty$ the compression vanishes, that is, larger data sets are less affected by this transformation.

From now on, we suppose that our response variable take values in the open interval $(0, 1)$. If not, this transformation is done.

12.2.2 Dirichlet regression

Let $\boldsymbol{Y}$ be a matrix with C rows and N columns denoting N observations for the different categories C of the C dimensional response variables $\boldsymbol{Y}_{\bullet n} \sim \mathcal{D}(\boldsymbol{\alpha}_n)$. Let η_{cn} be the value of the linear predictor for the ith observation in the cth category, so $\boldsymbol{\eta}$ is a matrix with C rows and N columns. Let $\boldsymbol{V}^{(c)}$, $c = 1, \ldots, C$, represent a matrix with dimension $N \times J_c$ which contains the covariates values for each individual and each category, so $\boldsymbol{V}_{n\bullet}^{(c)}$ shows the covariates values for the nth observation and the cth category. Let $\boldsymbol{\beta}$ be a matrix with J_c rows and C columns representing the regression coefficients in each dimension, then the model is set up as:

$$g(\alpha_{cn}) = \eta_{cn} = \boldsymbol{V}_{n\bullet}^{(c)} \boldsymbol{\beta}_{\bullet c} \,, \tag{12.3}$$

where $g(\cdot)$ is the link-function, in this case as $\alpha_c > 0\,, \forall c = 1, \ldots, C$, the $\log(\cdot)$ is employed. The regression coefficients $\boldsymbol{\beta}_{\bullet c}$ are a column vector with J_c elements.

The previous equation (12.3) can be rewritten in a vectorized form. Let

$$\tilde{\boldsymbol{\eta}} = \underbrace{\begin{bmatrix} \boldsymbol{\eta}_{\bullet 1} \\ \vdots \\ \boldsymbol{\eta}_{\bullet N} \end{bmatrix}}_{CN \times 1}$$

denotes a restructured linear predictor. Then, the model in matrix notation can be set up as:

$$\tilde{\boldsymbol{\eta}} = \boldsymbol{A}\boldsymbol{x} \,, \tag{12.4}$$

being $\boldsymbol{A}$ the matrix with covariates properly constructed with CN rows and $\sum_{c=1}^{C} J_c$ columns, $\boldsymbol{x}$ is a vector formed by the regression coefficientes with $\sum_{c=1}^{C} J^c$ rows and 1 column.

12.3 INLA for Latent Gaussian Models (LGMs)

In this section, we start with a brief explanation about the LGMs (subsection 12.3.1), followed by the main idea of the Laplace approximation (subsection 12.3.2) and finishing with the INLA methodology (subsection 12.3.3).

12.3.1 LGMs

The reason underneath the possibility of using INLA is based on the fact that the mostly of the models can also be seen as LGMs (Rue and Held, 2005), the class of models which INLA is designed for (Rue et al., 2009). The statistical inference is obtained using a three-stage hierarchical model formulation, in which observations $\boldsymbol{y}$ can be assumed to be conditionally independent, given a latent Gaussian random field $\boldsymbol{x}$ and hyperparameters $\boldsymbol{\theta_1}$,

$$\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta_1} \sim \prod_{n=1}^{N} p(y_n \mid x_n, \boldsymbol{\theta_1}) \,.$$

The versatility of the model class relates to the specification of the latent Gaussian field

$$\boldsymbol{x} \mid \boldsymbol{\theta_2} \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{\theta_2}), \boldsymbol{Q}^{-1}(\boldsymbol{\theta_2}))$$

which includes all the latent (nonobservable) components of interest such as fixed effects and random terms describing the underlying process of the data. The hyperparameters $\boldsymbol{\theta} = (\boldsymbol{\theta_1}, \boldsymbol{\theta_2})$ control the latent Gaussian field and/or the likelihood for the data.

The LGMs are a class generalising the large number of related variants of additive and generalized models. If the likelihood $p(y_n \mid x_n, \boldsymbol{\theta})$ such that "y_n only depends on its linear predictor η_n" yields the generalized linear model setup, the set $\{x_n, n = 1, \ldots, N\}$ can be interpreted as η_n, being η_n the linear predictor which is additive with respect to other effects,

$$\eta_n = \beta_0 + \sum_j v_{nj}\beta_j + \sum_k f_{k,n} \,, \tag{12.5}$$

where β_0 is the intercept, $\boldsymbol{v}$ represents the fixed covariates with linear effects $\{\beta_j\}$, and the terms $\{f_k\}$ represent specific Gaussian processes. Each $f_{k,n}$ is the contribution of the model components f_k to the nth linear predictor (Rue et al., 2017). If Gaussian prior is assumed for the intercept and the parameters of the fixed effects, the joint distribution of $\boldsymbol{x} = (\boldsymbol{\eta}, \beta_0, \boldsymbol{\beta}, \boldsymbol{f}_1, \boldsymbol{f}_2, \ldots)$ is then Gaussian. This yields the latent field $\boldsymbol{x}$ in the hierarchical LGM formulation. Regarding to the set of hyperparameters $\boldsymbol{\theta}$, it comprises the parameters of the likelihood and the model components. Usually, these parameters include some kind of variance, scale or correlation parameters.

In general, the latent field is not only Gaussian, but also it is a sparse Gaussian Markov random field (GMRF) (Rue and Held, 2005). A GMRF is just a Gaussian with additional conditional independence properties: x_j and x'_j are conditionally independent given the remaining elements. This provides the INLA methodology with nice computational properties.

12.3.2 Laplace Approximation

Laplace approximation (Barndorff-Nielsen and Cox, 1989) is a technique used to approximate integrals with the next form:

$$I_n = \int \exp(nf(x))dx\,, \text{ as } n \to \infty\,. \tag{12.6}$$

The main idea is to approximate the target with a Gaussian distribution, matching the mode and the curvature at the mode.

If x_0 is the point where $f(x)$ has its maximum, then

$$\begin{aligned} I_n &\approx \int \exp(n(f(x_0) + \frac{1}{2}(x - x_0)^2 f''(x_0)))dx \\ &= \exp(nf(x_0))\sqrt{\frac{2\pi}{-nf''(x_0)}} = \tilde{I}_n\,. \end{aligned} \tag{12.7}$$

If $nf(x)$ is interpreted as the sum of log-likelihoods and x as the unknown parameter, the Gaussian approximation will be exact as $n \to \infty$.

If we are interested in computing a marginal distribution $p(\gamma_1)$ from a joint distribution $p(\boldsymbol{\gamma})$, the approximation works as follows:

$$\begin{aligned} p(\gamma_1) &= \frac{p(\boldsymbol{\gamma})}{p(\boldsymbol{\gamma}_{-1} \mid \gamma_1)} \\ &\approx \frac{p(\boldsymbol{\gamma})}{p_G(\boldsymbol{\gamma}_{-1}; \boldsymbol{\mu}(\gamma_1), \boldsymbol{Q}(\gamma_1))}\Big|_{\boldsymbol{\gamma}^*_{-1}=\boldsymbol{\mu}(\gamma_1)}, \end{aligned} \tag{12.8}$$

where the fact that $p(\boldsymbol{\gamma}_{-1} \mid \gamma_1)$ is Gaussian is exploited. If the posterior is close to a Gaussian density, the results will be more accurate compared to a density that is very different from a Gaussian. In this context, uni-modality is necessary since the integrand is being approximated with a Gaussian.

12.3.3 INLA

The main idea of INLA approach is to approximate the posteriors of interest: the marginal posteriors for the latent field $p(x_m \mid \boldsymbol{y})$ and the marginal posteriors for the hyperparameters $p(\theta_k \mid \boldsymbol{y})$. These posteriors can be written as

$$p(x_m \mid \boldsymbol{y}) = \int p(x_m \mid \boldsymbol{\theta}, \boldsymbol{y}) p(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta}, \tag{12.9}$$

$$p(\theta_k \mid \boldsymbol{y}) = \int p(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta}_{-k}. \tag{12.10}$$

The nested formulation is used to compute $p(x_m \mid \boldsymbol{y})$ by approximating $p(x_m \mid \boldsymbol{\theta}, \boldsymbol{y})$ and $p(\boldsymbol{\theta} \mid \boldsymbol{y})$, and then using numerical integration to integrate out $\boldsymbol{\theta}$. Similarly, approximating $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ and integrating out $\boldsymbol{\theta}_{-k}$, $p(\theta_k \mid \boldsymbol{y})$ can be computed.

The marginal posterior distributions in equations (12.9) and (12.10) are computed using the Laplace approximation presented in subsection 12.3.2. In Rue et al. (2009), it is shown that the nested approach yields a very accurate approximation if applied to LGMs.

All this methodology can be used through R with the `R-INLA` package. For more details about `R-INLA` we refer the reader to Blangiardo and

Cameletti (2015); Zuur et al. (2017); Wang et al. (2018); Krainski et al. (2018), where practical examples and code guidelines are provided.

However, despite the advantages of `R-INLA` implementation, there are some limitations. For instance, `R-INLA` is not able to work with multivariate response variables, as it can not associate more than one data to only one individual. So, in order to fit the Dirichlet regression, we propose an extension of the paper Rue et al. (2009) for models with multivariate response variable and multiple linear predictors.

12.4 Inference in multivariate likelihoods

12.4.1 Motivation

The INLA methodology is a tool which allows deal with a widely range of LGMs, since the most simple linear regression models until models with multiple likelihoods and multiple linear predictors. However, when a multivariate response is required and several linear predictors are needed to explain it, the implemented `R-INLA` methodology has some limitations. Although, in this section we depict a general method for the case of the multivariate response, we focus on the particular case of the Dirichlet likelihood. In order to make this likehood handy, we propose:

1. Approximate the effect of the log likelihood on the posterior using the Laplace approximation.
2. Convert the multivariate initial observations in to independent Gaussian pseudo-observations which `R-INLA` can deal with.
3. As `R-INLA` can deal with independent Gaussian observations, we can call `R-INLA`.

12.4.2 The approximation

In this section, we present the theorical fundaments to approximate the log-likelihood function $\log p(\boldsymbol{Y} \mid \boldsymbol{x}, \boldsymbol{\theta})$ using the Laplace approximation getting conditioned independent Gaussian pseudo-observations.

In order to do so, let $\boldsymbol{\eta}_n := \boldsymbol{\eta}_{\bullet n}$ denote the linear predictor corresponding to the *nth* observation $\boldsymbol{y}_n := \boldsymbol{Y}_{\bullet n}$, we define $l(\boldsymbol{y} \mid \boldsymbol{x}) = -\log p(\boldsymbol{y} \mid \boldsymbol{x})$ for any $\boldsymbol{y}$ and $\boldsymbol{x}$. In particular, we denote $l(\boldsymbol{y}_n \mid \boldsymbol{\eta}_n) = -\log p(\boldsymbol{y}_n \mid \boldsymbol{\eta}_n)$ the log-likelihood function expressed for the n*th* observation, being $\boldsymbol{y}_n$ and $\boldsymbol{\eta}_n$ vectors with C components.

Let $\boldsymbol{\eta_0}_n$ be a vector. We define the gradient of l in $\boldsymbol{\eta_0}$ as $\boldsymbol{g_0}_{\boldsymbol{\eta}_n} = \nabla_{\boldsymbol{\eta}_n}(l)(\boldsymbol{\eta_0}_n, \boldsymbol{y}_n)$, and the Hessian of l in $\boldsymbol{\eta_0}$. $\boldsymbol{H_0}_{\boldsymbol{\eta}_n}$ can be either the true Hessian $(\nabla^2_{\boldsymbol{\eta}_n}(l)(\boldsymbol{\eta_0}_n, \boldsymbol{y}_n))$ or the expected Hessian $(\mathrm{E}_{\boldsymbol{y}_n|\boldsymbol{\eta}_n}(\nabla^2_{\boldsymbol{\eta}_n}(l)(\boldsymbol{\eta_0}_n, \boldsymbol{y}_n)))$ in $\boldsymbol{\eta_0}_n$. Let $\boldsymbol{L_0}_n$ be the result of applying the Cholesky factorization to $\boldsymbol{H_0}_{\boldsymbol{\eta}_n}$, $\boldsymbol{H_0}_{\boldsymbol{\eta}_n} = \boldsymbol{L_0}_n \boldsymbol{L_0}_n^T$.

Theorem 12.1. *If Laplace approximation (subsection 12.3.2) is applied for $l(\boldsymbol{y}_n \mid \boldsymbol{\eta}_n)$ in vector $\boldsymbol{\eta_0}_n$, then the vector*

$$\boldsymbol{z_0}_n := \boldsymbol{L_0}_n^T(\boldsymbol{\eta_0}_n - \boldsymbol{H_0}_{\boldsymbol{\eta}_n}^{-1}\boldsymbol{g_0}_{\boldsymbol{\eta}_n}) = \boldsymbol{L_0}_n^T\boldsymbol{\eta_0}_n - \boldsymbol{L_0}_n^{-1}\boldsymbol{g_0}_{\boldsymbol{\eta}_n}\,, \tag{12.11}$$

is conditionally independent Gaussian distributed (see Appendix 12.9 for the proof):

$$l(\boldsymbol{y}_n \mid \boldsymbol{\eta}_n) \;\approx\; C_1 + \frac{1}{2}[\boldsymbol{z_0}_n - \boldsymbol{L_0}_n^T\boldsymbol{\eta}_n]^T[\boldsymbol{z_0}_n - \boldsymbol{L_0}_n^T\boldsymbol{\eta}_n]\,, \tag{12.12}$$

i.e. $\boldsymbol{z_0}_n \mid \boldsymbol{\eta}_n \sim \mathcal{N}(\boldsymbol{L_0}_n^T\boldsymbol{\eta}_n, \boldsymbol{I_d})$ *and* $z_{0nk} \mid \boldsymbol{\eta}_n \sim \mathcal{N}([\boldsymbol{L_0}_n^T\boldsymbol{\eta}_n]_k, 1)$*, and the constant value of the expression is* $l(\boldsymbol{y}_n \mid \boldsymbol{\eta_0}_n) - \frac{1}{2}\boldsymbol{g_0}_{\boldsymbol{\eta}_n}^T(\boldsymbol{H_0}_{\boldsymbol{\eta}_n}^{-1})^T\boldsymbol{g_0}_{\boldsymbol{\eta}_n}$.

The observation vector $\boldsymbol{y}_n$ has been converted into Gaussian conditionally independent pseudo-observations $\boldsymbol{z_0}_n$. Note that this theorem can be expanded to multiple observations. In order to do so, we present the following notation.

Notation 12.2.

$$\tilde{\boldsymbol{\eta}}_{\mathbf{0}} = \underbrace{\begin{bmatrix} \boldsymbol{\eta}_{\mathbf{0}\bullet 1} \\ \vdots \\ \boldsymbol{\eta}_{\mathbf{0}\bullet N} \end{bmatrix}}_{CN\times 1}, \; \boldsymbol{g}_{\mathbf{0}\tilde{\boldsymbol{\eta}}} = \underbrace{\begin{bmatrix} \boldsymbol{g}_{\mathbf{0}1} \\ \vdots \\ \boldsymbol{g}_{\mathbf{0}N} \end{bmatrix}}_{CN\times 1}, \; \boldsymbol{L}_{\mathbf{0}} = \underbrace{\begin{bmatrix} \boldsymbol{L}_{\mathbf{0}1} & & 0 \\ & \ddots & \\ 0 & & \boldsymbol{L}_{\mathbf{0}N} \end{bmatrix}}_{CN\times dN},$$

$$\boldsymbol{H}_{\mathbf{0}\tilde{\boldsymbol{\eta}}} = \underbrace{\begin{bmatrix} \boldsymbol{H}_{\mathbf{0}\boldsymbol{\eta}_1} & & 0 \\ & \ddots & \\ 0 & & \boldsymbol{H}_{\mathbf{0}\boldsymbol{\eta}_N} \end{bmatrix}}_{CN\times CN}.$$

Proposition 12.3. *Using the previous notation, the matrix*

$$\tilde{\boldsymbol{z}}_{\mathbf{0}} := \boldsymbol{L}_{\mathbf{0}}{}^{T}\tilde{\boldsymbol{\eta}}_{\mathbf{0}} - \boldsymbol{L}_{\mathbf{0}}{}^{-1}\boldsymbol{g}_{\mathbf{0}\tilde{\boldsymbol{\eta}}} \tag{12.13}$$

is conditionally independent Gaussian distributed by columns,

$$\tilde{\boldsymbol{z}}_{\mathbf{0}} \mid \tilde{\boldsymbol{\eta}} \;\; \sim \;\; N(\boldsymbol{L}_{\mathbf{0}}{}^{T}\tilde{\boldsymbol{\eta}}, \boldsymbol{I}_{\boldsymbol{CN}}) \,. \tag{12.14}$$

Note that this approximation has been constructed for a generic $\tilde{\boldsymbol{\eta}}_{\mathbf{0}}$, but, as we are interested in building a Gaussian approximation of the effect of the likelihood on the posterior distribution, $\tilde{\boldsymbol{\eta}}_{\mathbf{0}}$ has been chosen as the posterior mode of $l(\boldsymbol{x} \mid \boldsymbol{Y})$ in $\tilde{\boldsymbol{\eta}}_{\mathbf{0}}$. In what follows, we depict the different steps in order to compute the marginals posterior distributions of the latent field.

12.4.3 The algorithm

In this section we present the computational approach of this method to compute the posterior marginals of the latent Gaussian field, $p(\boldsymbol{x}_n \mid \boldsymbol{Y})$, $n = 1, \ldots, N$. The approximation is computed in three steps. The first step computes the mode of the posterior distribution of the latent field $\boldsymbol{x}_{\mathbf{0}}$ and the mode of the posterior distribution of the linear predictor $\tilde{\boldsymbol{\eta}}_{\mathbf{0}}$. The second step computes the conditionally independent Gaussian observations in order to call `R-INLA`. Last step consists on calling `R-INLA` to get the posterior distributions of the latent field.

- **Computing the mode in $\boldsymbol{x}$ of $p(\boldsymbol{x} \mid \boldsymbol{Y})$.** The mode $\boldsymbol{x_0}$ in $\boldsymbol{x}$ of $-\log(\tilde{p}(\boldsymbol{x} \mid \boldsymbol{Y}))$ is computed using a quasi-Newton method with line search strategy. As we can see in previous sections, the likelihood function can be approximated with a quadratic expression being $\tilde{\boldsymbol{z}}_{\boldsymbol{0}}$ defined as in expression (12.13). On the other hand, as we are in the context of LGMs, prior distribution for $\boldsymbol{x}$ is multivariate Gaussian with precision matrix $\boldsymbol{Q_x}$. Thus, the minus log-posterior density approximation of $\boldsymbol{x}$ is computed as follows:

$$\begin{aligned} l(\boldsymbol{x} \mid \tilde{\boldsymbol{z}}_{\boldsymbol{0}}, \boldsymbol{\theta}) &= l(\boldsymbol{Y} \mid \tilde{\boldsymbol{\eta}}) + l(\boldsymbol{x} \mid \boldsymbol{\theta}) \\ &\approx C_1 + \frac{1}{2}\{[\tilde{\boldsymbol{z}}_{\boldsymbol{0}} - \boldsymbol{L_0}^T \boldsymbol{A}\boldsymbol{x}]^T[\tilde{\boldsymbol{z}}_{\boldsymbol{0}} - \boldsymbol{L_0}^T \boldsymbol{A}\boldsymbol{x}] + \boldsymbol{x}^T \boldsymbol{Q_x} \boldsymbol{x}\}\,. \end{aligned} \tag{12.15}$$

The target function to optimize is $l(\boldsymbol{x} \mid \tilde{\boldsymbol{z}}_{\boldsymbol{0}}, \boldsymbol{\theta})$, always keeping in mind that $\tilde{\boldsymbol{z}}_{\boldsymbol{0}}$ is depending on $\boldsymbol{x_0}$. To compute the quasi Newton-Raphson method, calculate the gradient and the Hessian of the expression (12.15) is needed. Note that this method works when first and second derivatives exist.

$$\begin{aligned} \frac{\partial l(\boldsymbol{x} \mid \tilde{\boldsymbol{z}}_{\boldsymbol{0}}, \boldsymbol{\theta})}{\partial \boldsymbol{x}} &= -\boldsymbol{A^T L_0}(\tilde{\boldsymbol{z}}_{\boldsymbol{0}} - \boldsymbol{L_0}^T \boldsymbol{A}\boldsymbol{x}) + \boldsymbol{Q_x}\boldsymbol{x}\,, \\ \frac{\partial^2 l(\boldsymbol{x} \mid \tilde{\boldsymbol{z}}_{\boldsymbol{0}}, \boldsymbol{\theta})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^T} &= +\boldsymbol{A}^T \boldsymbol{H}_{\boldsymbol{0}\tilde{\boldsymbol{\eta}}} \boldsymbol{A} + \boldsymbol{Q_x}\,. \end{aligned} \tag{12.16}$$

The Newton-Raphson algorithm with line search strategy and Armijo conditions is employed in order to find the mode $\boldsymbol{x_0}$. In each iteration of the algorithm, the Hessian, the gradient, the conditionally independent Gaussian quasi-observations has to be computed in each iterative point until the method converges. Once the method reachs the mode $\boldsymbol{x_0}$, $\tilde{\boldsymbol{\eta}}_{\boldsymbol{0}}$ can be easily calculated as $\tilde{\boldsymbol{\eta}} = \boldsymbol{A}\boldsymbol{x}$.

- **Calculating the conditionally independent Gaussian pseudo-observations $\tilde{\boldsymbol{z}}_{\boldsymbol{0}}$.** At the modal configuration, the Hessian matrix $\boldsymbol{H}_{\boldsymbol{0}\tilde{\boldsymbol{\eta}}}$ is computed. If the submatrix corresponding to the *nth* indivual $\boldsymbol{H}_{\boldsymbol{0}\boldsymbol{\eta}^n}$ is negative definite, the expected Hessian is required to guarantee $\boldsymbol{H}_{\boldsymbol{0}\tilde{\boldsymbol{\eta}}}$ to be positive definite. Following the approximation previously presented, the Cholesky factorization is computed in

$\boldsymbol{H_{0\tilde{\eta}}} = \boldsymbol{L_0}\boldsymbol{L_0}^T$. Gradient ($\boldsymbol{g_{0\tilde{\eta}}}$) is also calculated in $\tilde{\boldsymbol{\eta}}_\mathbf{0}$. According to the equation (12.13), the scale and rotation of the original observations are done to get the pseudo-observations $\tilde{\boldsymbol{z}}_\mathbf{0}$.

- **Call** `R-INLA`. Lastly, as we have conditionally independent Gaussian observations, we are able to call `R-INLA` and make inference.

After depicting the complete method, we focus on the `dirinla` package to fit Dirichlet regression models.

12.5 The R-package dirinla

This section is devoted to show how the approximation works for the case of the Dirichlet regression using the `dirinla` package. This is not available to install from Comprehensive R Archive Network (CRAN), but it will be available soon.

In order to illustrate how this package can be useful for practitioners, we present an example of a Dirichlet regression simulation with 50 data, four different categories and one different covariate per category. The model that we want to simulate to posteriorly fit is:

$$\begin{aligned} \boldsymbol{Y_n} &\sim \text{Dirichlet}(\alpha_{1n}, \ldots, \alpha_{4n}), n = 1, \ldots, 50 \\ \log(\alpha_{1n}) &= \beta_{01} + \beta_{11} v_{1n} \\ \log(\alpha_{2n}) &= \beta_{02} + \beta_{12} v_{2n} \\ \log(\alpha_{3n}) &= \beta_{03} + \beta_{13} v_{3n} \\ \log(\alpha_{4n}) &= \beta_{04} + \beta_{14} v_{4n} \end{aligned} \tag{12.17}$$

being -1.5, 1, -3, 1.5 the values for the intercepts β_{01}, β_{02}, β_{03}, β_{04}; and 2, -3 , -1, 5 for the slopes β_{11}, β_{12}, β_{13}, β_{14}. All these parameters compose the latent field. Covariates are simulated from a Uniform distribution (0,1). To posteriorly fit the model, Gaussian prior distributions are assigned with precision 10^{-4} to all the elements of the Gaussian field.

12.5.1 Data simulation

This subsection is devoted to present how the simulation of the data is conducted.

- First, we simulate the covariates from a Uniform(0,1).

```
R> set.seed(1000)
R> N <- 50
R> V <- as.data.frame(matrix(runif((10)*N, 0, 1), ncol=10))
R> names(V) <- paste0('v', 1:4)
```

- We define the formula that we want to fit in order to keep the values of the different categories in a list. Posteriorly this object is used to construct the $\boldsymbol{A}$ matrix. We use the function `formula_list()` from the package `dirinla`.

```
R> formula <- y ~ 1 + v1 | 1 + v2 | 1 + v3 | 1 + v4
R> (names_cat <- formula_list(formula))
```

The output is a list indicating the covariates in each category:

```
$`category 1`
[1] "intercept" "v1"
$`category 2`
[1] "intercept" "v2"
$`category 3`
[1] "intercept" "v3"
$`category 4`
[1] "intercept" "v4"
```

- We fix the values of the parameters which take part of the latent field, and that we want to fit. As we have previously depicted, these parameters are -1.5, 1, -3, 1.5 for the intercepts β_{01}, β_{02}, β_{03}, β_{04}; and 2, -3 , -1, 5 for the slopes β_{11}, β_{12}, β_{13}, β_{14} respectively.

```
R> x <- c(-1.5, 1, -3, 1.5, 2, -3 , -1, 5)
```

- We call the function `data_stack_dirich()` of the package `dirinla` to construct the $\boldsymbol{A}$ matrix presented in previous sections. This function uses the `inla.stack()` structure of the package `R-INLA`. As a consecuence the returning object is an `inla.stack` object. Observe that the arguments are the response variable `y` (in this case it has not been generated yet), the names of the categories `covariates`, a matrix with the values of the covariates `data`, the number of categories `d` and the number of data `N`.

```
R> C <- length(names_cat)
R> data_stack_construct <-
          data_stack_dirich(
                  y = as.vector(rep(NA, N*C)),
                        covariates = names_cat,
                        data       = V,
                        d          = C,
                        n          = N )
R> A_construct <- data_stack_construct\$A
R> A_construct[1:8,]
```

The sparse matrix $\boldsymbol{A}$ is then easily computed:

```
[1,] 1 . . . 0.3278787 .         .         .
[2,] . 1 . . .         0.7267993 .         .
[3,] . . 1 . .         .         0.5993679 .
[4,] . . . 1 .         .         .         0.3224284
[5,] 1 . . . 0.7588465 .         .         .
[6,] . 1 . . .         0.6820559 .         .
[7,] . . 1 . .         .         0.4516818 .
[8,] . . . 1 .         .         .         0.5613199
```

- The next step is calculate the linear predictor as $\tilde{\boldsymbol{\eta}} = \boldsymbol{A}\boldsymbol{x}$ using the parameters fixed in the latent field. Moreover, using the exponential transformation, it is easy to get α parameters of the Dirichlet distribution.

```
R> eta <- A_construct %*% x
R> alpha <- exp(eta)
R> alpha <- matrix(alpha,
```

```
                ncol  = C,
                byrow = TRUE)
```

- The last stage is generate the response variable using the function `rdirichlet()` of the package `DirichletReg`.

  ```
  R> y <- rdirichlet(n, alpha)
  R> colnames(y) <- paste0("y", 1:C)
  R> head(y)
  ```

  In the output, a matrix with the response variable, and we can see that each row is summing up to one.

```
               y1           y2           y3        y4
[1,] 0.0216425724 8.269851e-04 1.646888e-12 0.9775304
[2,] 0.0174529730 1.600423e-03 4.366696e-18 0.9809466
[3,] 0.0004755623 3.226895e-02 2.841956e-07 0.9672552
[4,] 0.0034944201 1.558691e-03 3.755472e-13 0.9949469
[5,] 0.0027955541 8.049758e-06 4.523045e-14 0.9971964
[6,] 0.0012871593 4.840909e-04 2.125868e-38 0.9982287
```

Once the data is simulated, it is time to show how to fit the model.

12.5.2 Fitting the model

In order to fit the model using the `dirinla` package, we just need to call the main function `dirinlareg`. This function is the core of the package and it carries out all the steps presented in Section 12.4.

```
R> model.inla <- dirinlareg(
        formula  = y ~ 1 + v1 | 1 + v2 | 1 + v3 | 1 + v4  ,
        y        = y,
        data.cov = V,
        prec     = 0.0001)
```

where we just need to specify the `formula`, the response variable $\boldsymbol{Y}$ in a matrix format (in that case ,with dimension 50×4), the `data.frame` with

the covariates `data.cov` and lastly, the precision of the Gaussian priors (`prec`) for the latent field $\boldsymbol{x}$. If we want to follow the process step by step, we can add the instruction `verbose = TRUE`.

Once the model is computed, we can see a summary of the posterior distribution of the fixed effects just employing the function `summary` to the object generated which is `dirinlareg` class. Some model selection criteria are also displayed: Deviance Information Criterion (Spiegelhalter et al., 2002), Watanabe-Akaike information criteria (Gelman et al., 2014), and the - mean of the logarithm of the conditional predictive ordinate (Gneiting and Raftery, 2007). Lastly, the number of observations and the number of categories are also depicted.

```
dirinlareg(formula = y ~ 1 + v1 | 1 + v2 | 1 + v3 | 1 + v4,
        y = y,
    data.cov = V, prec = 1e-04, verbose = TRUE)


---- FIXED EFFECTS ----
========================================================================
Category  1
------------------------------------------------------------------------
             mean     sd 0.025quant 0.5quant 0.975quant   mode
intercept -1.423 0.2864     -1.985   -1.423     -0.861 -1.423
v1         1.948 0.4711      1.023    1.948      2.872  1.948
========================================================================
Category  2
------------------------------------------------------------------------
             mean     sd 0.025quant 0.5quant 0.975quant   mode
intercept  0.829 0.2453     0.3473   0.8289      1.310  0.829
v2        -3.037 0.4363    -3.8932  -3.0365     -2.181 -3.037
========================================================================
Category  3
------------------------------------------------------------------------
             mean     sd 0.025quant 0.5quant 0.975quant   mode
intercept -3.044 0.3064     -3.645   -3.044    -2.4425 -3.044
v3        -1.054 0.5196     -2.074   -1.054    -0.0344 -1.054
========================================================================
Category  4
------------------------------------------------------------------------
            mean     sd 0.025quant 0.5quant 0.975quant  mode
intercept 1.507 0.3220     0.8749    1.507      2.139 1.507
v4        4.896 0.4442     4.0242    4.896      5.768 4.896
========================================================================

DIC = 4103.375 , WAIC = 3393.1158 , LCPO = 1695.9688
```

```
Number of observations: 50
Number of Categories: 4
```

With the command `model.inla$marginals_fixed` and `model.inla$summary_fixed`), a representation and a summary of the fixed effects marginals can be obtained (Figure 12.1).

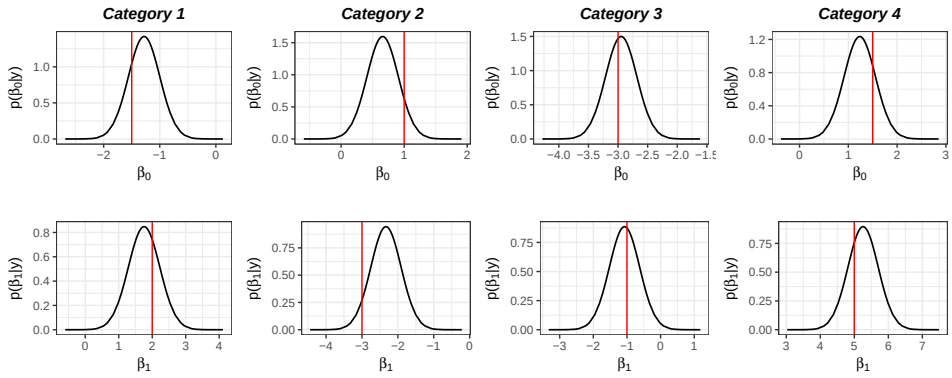

FIGURE 12.1: Marginal posterior distributions of the latent field for the different categories. Real values are indicated with a red line. The amount of data is 50.

Also, the posterior distribution for the scale parameters of the Dirichlet α are computed. With `model.inla$marginals_fixed` and `model.inla$summary_fixed` a representation and a summary for each category can be obtained. Parameter means and precision are also obtained doing `model.inla$marginals_means` or `model.inla$summary_means` for the means, and `model.inla$marginals_precision` or `model.inla$summary_precision` for the precision.

12.6 Simulation studies

This section provides examples of applications of the INLA approach for Dirichlet regression models using `dirinla` package, in comparison with a widely used method for Bayesian inference using MCMC algorithms, `R-JAGS`

(Plummer, 2016). For each example, we have obtained three approximations: firstly, we have employed `R-JAGS` with number of iterations enough to guarantee chains convergence; secondly, we have used INLA methodology through the R package `dirinla`; and lastly, we have employed a "long" `R-JAGS` with a big amount of iterations in order to get really good representation of the posterior distributions. The computations have been performed on a processor Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz. For each simulation, we propose three different datasets with the same parameters but with a different number of observations: 50, 100 and 500.

12.6.1 Simulation 1

We start by illustrating the posterior approximation comparison of the latent field $\boldsymbol{x}$ in a quite simple example. It is based on a Dirichlet regression with four categories and just one parameter per category, the intercept.

$$\begin{aligned} \boldsymbol{Y_n} &\sim \text{Dirichlet}(\alpha_{1n}, \ldots, \alpha_{4n}), n = 1, \ldots, N \\ \log(\alpha_{1n}) &= \beta_{01} \\ \log(\alpha_{2n}) &= \beta_{02} \\ \log(\alpha_{3n}) &= \beta_{03} \\ \log(\alpha_{4n}) &= \beta_{04} \end{aligned} \tag{12.18}$$

The different datasets with the structure in expression (12.18) have been simulated. The values β_{0c} with $c = 1, \ldots, 4$ have been -2.4, 1.2, -3.1, 1.3 respectively. In order to fit the model, some vague prior distributions for the latent field have been settled, in particular $p(x_m) \sim \mathcal{N}(0, \tau = 0.0001)$. As the response values are not so close to 0 and 1, no transformation has been needed.

For the purpose of fitting the model with MCMC algorithms, the number of iterations used with `R-JAGS` has been 1000 with a burning of 100, thin 5 and 3 chains for the three different simulated datasets. On the contrary, in order to have a really good representation of the posterior, `long R-JAGS` has been performed using 1000000 of iterations with a burning of 100000, thin 5 and 3 chains.

In Figures 12.2, 12.3 and 12.4, the marginal posterior distributions for the β_{0c} with $1, \ldots, 4$ are displayed showing that the approximation performance is almost perfect comparing with models fitted using `R-JAGS` and with the real value.

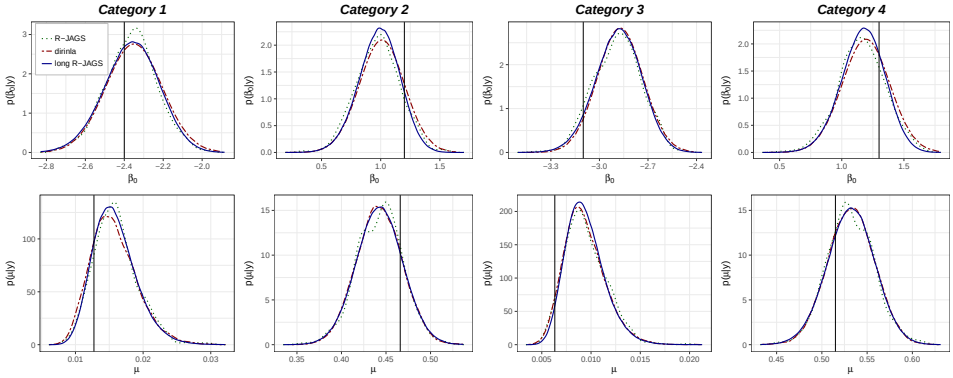

FIGURE 12.2: Marginal posterior distributions of the latent field for the different categories, and using different methodologies `R-JAGS`, `dirinla` and long `R-JAGS`, when the amount of data is 50.

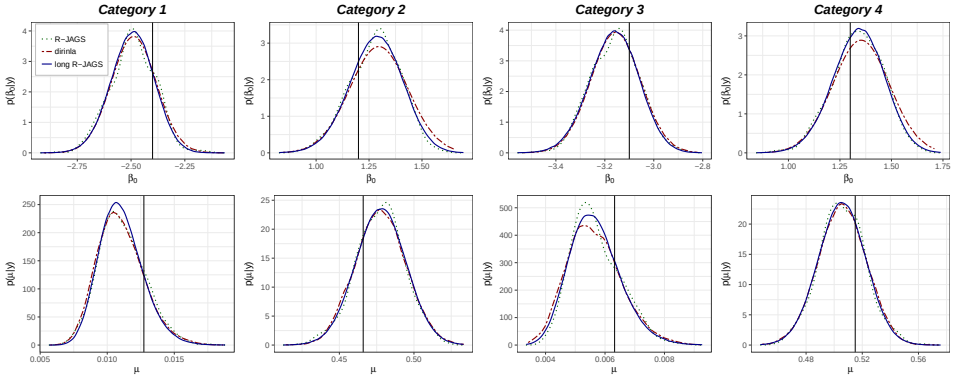

FIGURE 12.3: Marginal posterior distributions of the latent field for the different categories, and using different methodologies `R-JAGS`, `dirinla` and long `R-JAGS`, when the amount of data is 100.

As we know, one of the great advantage of the Laplace approximation is

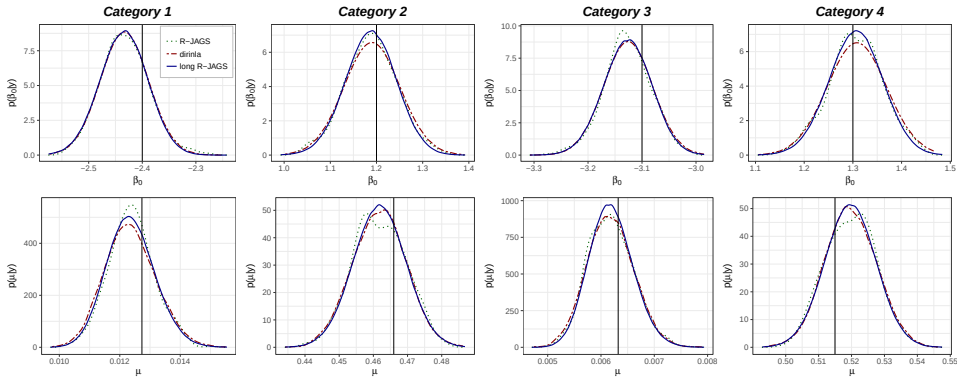

FIGURE 12.4: Marginal posterior distributions of the latent field for the different categories, and using different methodologies `R-JAGS`, `dirinla` and long `R-JAGS`, when the amount of data is 500.

the low computational cost required. Here, despite the previous process before to call `R-INLA` implemented in the R-package `dirinla`, the approximation presented has been faster than long `R-JAGS` in all the cases as depicted in Table 12.1. `R-JAGS` with an enough number of iterations to guarantee convergence is faster than `dirinla` in some cases.

N	`R-JAGS`	`dirinla`	`long R-JAGS`
50	2.945	4.336	2869.940
100	21.335	14.232	9839.647
500	28.454	41.082	30383.49

TABLE 12.1: Computational time in seconds for the different simulated data and with the different methodologies.

12.6.2 Simulation 2

In this second example, we illustrate the posterior approximations of the latent field $\boldsymbol{x}$ in a Dirichlet regression where a different covariate per category

is included.

$$\begin{aligned} \boldsymbol{Y_n} &\sim \text{Dirichlet}(\alpha_{1n}, \ldots, \alpha_{4n}), i = 1, \ldots, N \\ \log(\alpha_{1n}) &= \beta_{01} + \beta_{11} v_{1n} \\ \log(\alpha_{2n}) &= \beta_{02} + \beta_{12} v_{2n} \\ \log(\alpha_{3n}) &= \beta_{03} + \beta_{13} v_{3n} \\ \log(\alpha_{4n}) &= \beta_{04} + \beta_{14} v_{4n} \end{aligned} \tag{12.19}$$

Again, simulations for $N = 50, 100$ and 500 have been done with the previous structure. Values for β_{0c} and β_{1c} for $c = 1, \ldots, 4$ have been $-1.5, 1, -3, 1.5, 2, -3, -1, 5$ respectively, and covariates have been generated from a Uniform distribution with mean in the interval $(0, 1)$. Vague prior distributions for the latent field have been established $p(x_n) \sim \mathcal{N}(0, \tau = 0.0001)$. As the data generated did not present zeros and ones, any transformation has been needed.

When $N = 50$, the number of iterations used in `R-JAGS` has been 1000 with a burning of 100, thin 5 and 3 chains. Different conditions have been introduced when $N = 100$ and $N = 500$, the number of iterations has been 2000 with a burning of 200, thin 5 and 3 chains. On the contrary, in the case of `long R-JAGS`, the conditions employed have been 1000000 of iterations with a burning of 100000, thin 5 and 3 chains.

Marginal posterior distributions have been depicted in the Figures 12.5, 12.6, 12.7. all the posteriors capture the real value in a proper way. With regard to the comparison of `dirinla` with `R-JAGS` and `long R-JAGS`, posteriors have similar shape, which it is pointing out that our method performance is correct.

In Table 12.2 computational times are displayed, showing that in most cases, the INLA methodology guarantees a faster computational speed.

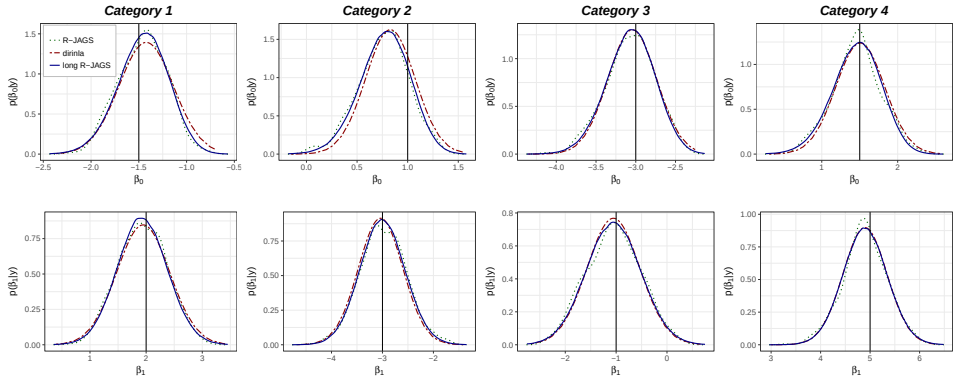

FIGURE 12.5: Marginal posterior distributions of the latent field for the different categories, and using different methodologies `R-JAGS`, `dirinla` and long `R-JAGS`, when the amount of data is 50.

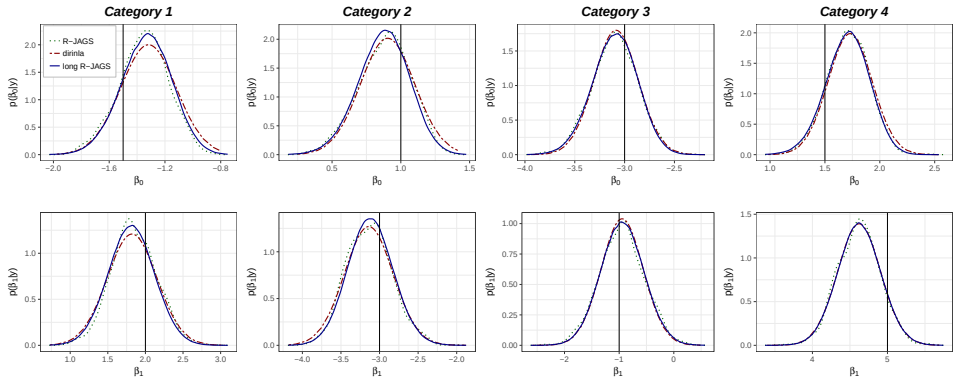

FIGURE 12.6: Marginal posterior distributions of the latent field for the different categories, and using different methodologies `R-JAGS`, `dirinla` and long `R-JAGS`, when the amount of data is 100.

12.7 Real example: Glacial tills

Similarly to the previous section, here we provide an example with real data using the INLA approach for Dirichlet regression models, and we compare it with `R-JAGS` and `long R-JAGS`.

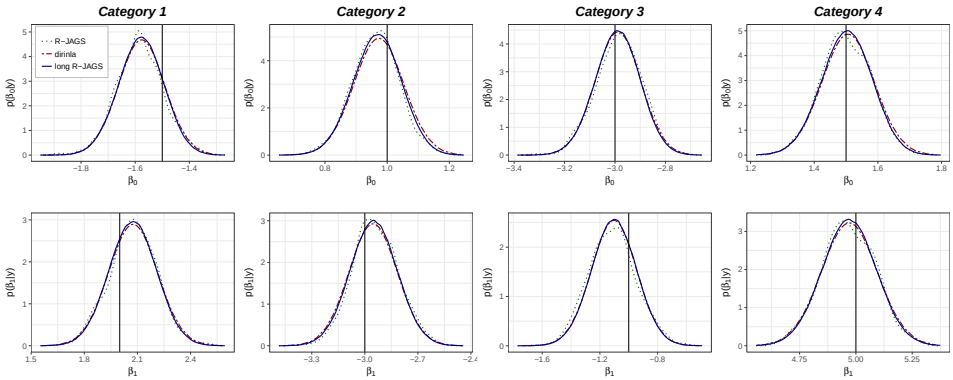

FIGURE 12.7: Marginal posterior distributions of the latent field for the different categories, and using different methodologies `R-JAGS`, `dirinla` and long `R-JAGS`, when the amount of data is 500.

N	`R-JAGS`	`dirinla`	`long R-JAGS`
50	5.715	6.664	4916.804
100	21.335	14.232	9839.647
500	133.973	54.873	57347.919

TABLE 12.2: Computational time in seconds for the different simulated data and with the different methodologies.

The data of this real example has been extracted from Aitchison (2003). The aim has been to do a pebble analysis of glacial tills. The total number of pebbles in each of 92 samples has been counted and the pebbles has been sorted into four categories,

- A: red sandstone,
- B: gray sandstone,
- C: crystalline,
- D: miscellaneous.

The percentages of these four categories and the total pebble counts have been recorded. The glaciologist has been interested in describing whether the compositions are in any way related to abundance.

With the purpose of analyzing this compositional data, a Dirichlet regression has been proposed. If $\boldsymbol{Y}_n$ represents a multivariate response with the proportion of red sandstone, gray sandstone, crystalline, and miscellaneous, then the model is written as follows

$$\begin{aligned} \boldsymbol{Y_n} &\sim \text{Dirichlet}(\alpha_{1n}, \ldots, \alpha_{4n}), n = 1, \ldots, N \\ \log(\alpha_{1n}) &= \beta_{01} + \beta_{11} Pcount_n \\ \log(\alpha_{2n}) &= \beta_{02} + \beta_{12} Pcount_n \\ \log(\alpha_{3n}) &= \beta_{03} + \beta_{13} Pcount_n \\ \log(\alpha_{4n}) &= \beta_{04} + \beta_{14} Pcount_n \end{aligned} \tag{12.20}$$

where $Pcount_n$ is the covariate pebble counts for the i individual divided by 100. Vague prior distributions for the latent field has been settled, in particular $p(x_n) \sim \mathcal{N}(0, \tau = 0.0001)$. As the data presented zeros and ones, the transformation introduced in 12.2.1.1 has been computed.

The number of iterations used in `R-JAGS` has been 1000 with a burning of 100, thin 5 and 3 chains. On the contrary, in order to have a really good representation of the posterior, `long R-JAGS` has been performed using 1000000 of iterations with a burning of 100000, thin 5 and 3 chains.

In the Figure 12.8, the marginal posterior distribution for β_{0c} and β_{1c}, $c = 1, \ldots, 4$ are displayed. We can notice that in most cases distributions obtained with `R-JAGS` match perfectly with the posteriors obtained using `R-INLA`. Regarding to the computational time, `R-JAGS` took 11.489 seconds, `dirinla` 7.030 and `long R-JAGS` 11588.554.

12.8 Discussion and Future Work

In this paper, the INLA methodology is extended to fit a model with a multivariate likelihood, the Dirichlet regression. The main idea is to approximate

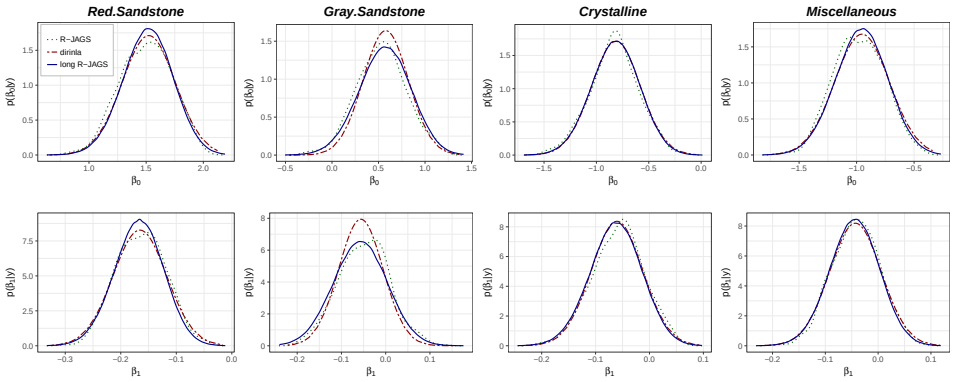

FIGURE 12.8: Posterior distributions of the latent field for the different categories.

it by likelihoods that can be fitted by `R-INLA`, in this particular case, using a Gaussian likelihood. In Simpson et al. (2016) a similar way was employed: they constructed a Poisson approximation to the true log-Gaussian Cox process likelihood to perform inference on a regular lattice over the observation window, counting the number of points in each cell. They implemented this technique in the R package `inlabru` (Bachl and Lindgren, 2018).

With regard to the computational aspect, here, we are presenting some results in order to fit models with just fixed effects. But there is still work to do. As we are converting original multivariate observations in conditionally independent Gaussian observations which only depends on the linear predictor, we expect to be able to incorporate random effects to the model, in particular, all the random effects which `R-INLA` can deal with allowing the user to fit spatial, temporal and spatio-temporal models.

But it is not the only challenge for the future. This approximation has been presented for a Dirichlet likelihood, but we expect to extend it for other multivariate likelihoods such as Multivariate normal regression (Anderson et al., 1958) or Multinomial regression (Menard, 2002).

References

Aitchison, J. (1981). A new approach to null correlations of proportions. *Mathematical Geology*, 13(2):175–189.

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 139–177.

Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika*, 70(1):57–65.

Aitchison, J. (1984). The statistical analysis of geochemical compositions. *Journal of the International Association for Mathematical Geology*, 16(6):531–564.

Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman and Hall London.

Aitchison, J. (2003). *The Statistical Analysis of Compositional Data.* The Blackburn Press, Caldwell, NJ.

Aitchison, J. and Egozcue, J. J. (2005). Compositional data analysis: where are we and where should we be heading? *Mathematical Geology*, 37(7):829–850.

Anderson, T. W., Anderson, T. W., Anderson, T. W., Anderson, T. W., and Mathématicien, E.-U. (1958). *An introduction to multivariate statistical analysis*, volume 2. Wiley New York.

Bachl, F. E. and Lindgren, F. (2018). *inlabru: Spatial Inference using Integrated Nested Laplace Approximation*. R package version 2.1.9.

Barndorff-Nielsen, O. and Cox, D. (1989). *Asymptotic Techniques for Use in Statistics*. Boca Raton, FL: Chapman and Hall/CRC.

Blangiardo, M. and Cameletti, M. (2015). *Spatial and spatio-temporal Bayesian models with `R-INLA`*. John Wiley & Sons.

Cribari-Neto, F. and Zeileis, A. (2010). Beta regression in r. *Journal of Statistical Software*, 34(2).

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and computing*, 24(6):997–1016.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Hijazi, R. H. and Jernigan, R. W. (2009). Modelling compositional data using Dirichlet regression models. *Journal of Applied Probability & Statistics*, 4(1):77–91.

Klein, N., Kneib, T., Klasen, S., and Lang, S. (2015). Bayesian structured additive distributional regression for multivariate responses. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(4):569–591.

Krainski, E. T., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., and Rue, H. (2018). *Advanced Spatial Modeling with Stochastic Partial Differential Equations using R and INLA*. CRC Press.

Maier, M. J. (2014). DirichletReg: Dirichlet regression for compositional data in R.

Menard, S. (2002). *Applied logistic regression analysis*, volume 106. Sage.

Plummer, M. (2016). *rjags: Bayesian Graphical Models using MCMC*. R package version 4-6.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace

approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2017). Bayesian computing with `INLA`: a review. *Annual Review of Statistics and Its Application*, 4:395–421.

Sennhenn-Reulen, H. (2018). Bayesian Regression for a Dirichlet Distributed Response using Stan. *arXiv preprint arXiv:1808.06399*.

Simpson, D., Illian, J. B., Lindgren, F., Sørbye, S. H., and Rue, H. (2016). Going off grid: Computationally efficient inference for log-Gaussian Cox processes. *Biometrika*, 103(1):49–70.

Smithson, M. and Verkuilen, J. (2006). A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods*, 11(1):54.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

van der Merwe, S. (2018). A method for Bayesian regression modelling of composition data. *arXiv preprint arXiv:1801.02954*.

Wang, X., Ryan, Y. Y., and Faraway, J. J. (2018). *Bayesian Regression Modeling with INLA*. Chapman and Hall/CRC.

Zuur, A. F., Ieno, E. N., and Saveliev, A. A. (2017). *Beginner's guide to spatial, temporal, and spatial-temporal ecological data analysis with R-INLA*.

12.9 Appendix

12.9.1 Each observation

Let $\boldsymbol{\eta}_n := \boldsymbol{\eta}_{\bullet n}$ denotes the linear predictor corresponding to the nth observation $\boldsymbol{y}_n := \boldsymbol{Y}_{\bullet n}$, we define $l(\boldsymbol{y} \mid \boldsymbol{x}) = -\log p(\boldsymbol{y} \mid \boldsymbol{x})$ for any $\boldsymbol{y}$ and $\boldsymbol{x}$. In particular, we denote $l(\boldsymbol{y}_n \mid \boldsymbol{\eta}_n) = -\log p(\boldsymbol{y}_n \mid \boldsymbol{\eta}_n)$ the log-likelihood function expressed for the nth observation, being $\boldsymbol{y}_n \in \mathbb{R}^C$ and $\boldsymbol{\eta}_n \in \mathbb{R}^C$. Using the Taylor series expansion in vector $\boldsymbol{\eta}_{\mathbf{0}n}$, we obtain the approximation.

$$
\begin{aligned}
& l(\boldsymbol{y}_n \mid \boldsymbol{\eta}_n) \approx \\
& \quad \approx \quad l(\boldsymbol{y}_n \mid \boldsymbol{\eta}_{\mathbf{0}n}) + [\nabla_{\boldsymbol{\eta}_n}(l)(\boldsymbol{\eta}_{\mathbf{0}n}, \boldsymbol{y}_n)]^T [\boldsymbol{\eta}_n - \boldsymbol{\eta}_{\mathbf{0}n}] \\
& \qquad + \frac{1}{2}[\boldsymbol{\eta}_n - \boldsymbol{\eta}_{\mathbf{0}n}]^T [\nabla^2_{\boldsymbol{\eta}_n}(l)(\boldsymbol{\eta}_{\mathbf{0}n}, \boldsymbol{y}_n)][\boldsymbol{\eta}_n - \boldsymbol{\eta}_{\mathbf{0}n}] \\
& \quad = \quad l(\boldsymbol{y}_n \mid \boldsymbol{\eta}_{\mathbf{0}n}) + \boldsymbol{g}_{\mathbf{0}\boldsymbol{\eta}_n}{}^T [\boldsymbol{\eta}_n - \boldsymbol{\eta}_{\mathbf{0}n}] + \frac{1}{2}[\boldsymbol{\eta}_n - \boldsymbol{\eta}_{\mathbf{0}n}]^T \boldsymbol{H}_{\mathbf{0}\boldsymbol{\eta}_n} [\boldsymbol{\eta}_n - \boldsymbol{\eta}_{\mathbf{0}n}] \quad (12.21) \\
& \quad = \quad C_1 + \frac{1}{2}[\boldsymbol{\eta}_n - (\boldsymbol{\eta}_{\mathbf{0}n} - \boldsymbol{H}_{\mathbf{0}\boldsymbol{\eta}_n}{}^{-1} \boldsymbol{g}_{\mathbf{0}\boldsymbol{\eta}_n})]^T \boldsymbol{H}_{\mathbf{0}\boldsymbol{\eta}_n} [\boldsymbol{\eta}_n - (\boldsymbol{\eta}_{\mathbf{0}n} - \boldsymbol{H}_{\mathbf{0}\boldsymbol{\eta}_n}{}^{-1} \boldsymbol{g}_{\mathbf{0}\boldsymbol{\eta}_n})] \,,
\end{aligned}
$$

where $\boldsymbol{g}_{\mathbf{0}\boldsymbol{\eta}_n} = \nabla_{\boldsymbol{\eta}_n}(l)(\boldsymbol{\eta}_{\mathbf{0}n}, \boldsymbol{y}_n)$ and $\boldsymbol{H}_{\mathbf{0}\boldsymbol{\eta}_n}$ is either the true Hessian $(\nabla^2_{\boldsymbol{\eta}_n}(l)(\boldsymbol{\eta}_{\mathbf{0}n}, \boldsymbol{y}_n))$ or the expected Hessian$(\mathrm{E}_{\boldsymbol{y}_n|\boldsymbol{\eta}_n}(\nabla^2_{\boldsymbol{\eta}_n}(l)(\boldsymbol{\eta}_{\mathbf{0}n}, \boldsymbol{y}_n)))$. C_1 is a constant whose value is $l(\boldsymbol{y}_n \mid \boldsymbol{\eta}_{\mathbf{0}n}) - \frac{1}{2}\boldsymbol{g}_{\mathbf{0}\boldsymbol{\eta}_n}{}^T (\boldsymbol{H}_{\mathbf{0}\boldsymbol{\eta}_n}{}^{-1})^T \boldsymbol{g}_{\mathbf{0}\boldsymbol{\eta}_n}$. Now we consider the Cholesky factorization of $\boldsymbol{H}_{\mathbf{0}\boldsymbol{\eta}_n}$, $\boldsymbol{H}_{\mathbf{0}\boldsymbol{\eta}_n} = \boldsymbol{L}_{\mathbf{0}n} \boldsymbol{L}_{\mathbf{0}n}^T$ and rewrite expression (12.21) as follows:

$$
\begin{aligned}
& l(\boldsymbol{y}_n \mid \boldsymbol{\eta}_n) \approx \qquad\qquad (12.22) \\
& \quad \approx \quad C_1 + \frac{1}{2}[\boldsymbol{L}_{\mathbf{0}n}^T \boldsymbol{\eta}_n - \boldsymbol{L}_{\mathbf{0}n}^T (\boldsymbol{\eta}_{\mathbf{0}n} - \boldsymbol{H}_{\mathbf{0}\boldsymbol{\eta}_n}{}^{-1} \boldsymbol{g}_{\mathbf{0}\boldsymbol{\eta}_n})]^T \\
& \qquad [\boldsymbol{L}_{\mathbf{0}n}^T \boldsymbol{\eta}_n - \boldsymbol{L}_{\mathbf{0}n}^T (\boldsymbol{\eta}_{\mathbf{0}n} - \boldsymbol{H}_{\mathbf{0}\boldsymbol{\eta}_n}{}^{-1} \boldsymbol{g}_{\mathbf{0}\boldsymbol{\eta}_n})] \,.
\end{aligned}
$$

Defining

$$
\boldsymbol{z}_{\mathbf{0}n} := \boldsymbol{L}_{\mathbf{0}n}^T (\boldsymbol{\eta}_{\mathbf{0}n} - \boldsymbol{H}_{\mathbf{0}\boldsymbol{\eta}_n}{}^{-1} \boldsymbol{g}_{\mathbf{0}\boldsymbol{\eta}_n}) = \boldsymbol{L}_{\mathbf{0}n}^T \boldsymbol{\eta}_{\mathbf{0}n} - \boldsymbol{L}_{\mathbf{0}n}{}^{-1} \boldsymbol{g}_{\mathbf{0}\boldsymbol{\eta}_n} \,, \qquad (12.23)
$$

a conditionally Gaussian approximation is constructed.

$$
\log p(\boldsymbol{y}_n \mid \boldsymbol{\eta}_i, \boldsymbol{\theta}) \quad \approx \quad -C_1 - \frac{1}{2}[\boldsymbol{z}_{\mathbf{0}n} - \boldsymbol{L}_{\mathbf{0}n}^T \boldsymbol{\eta}_n]^T [\boldsymbol{z}_{\mathbf{0}n} - \boldsymbol{L}_{\mathbf{0}n}^T \boldsymbol{\eta}_n] \,. (12.24)
$$

Thus, $\boldsymbol{z_{0}}_n \mid \boldsymbol{\eta}_n \sim \mathcal{N}(\boldsymbol{L_{0}}_n^T \boldsymbol{\eta}_n, \boldsymbol{I_d})$, *i.e.*, $z_{0ik} \mid \boldsymbol{\eta}_n \sim \mathcal{N}([\boldsymbol{L_{0}}_n^T \boldsymbol{\eta}_n]_k, 1)$. The observation vector $\boldsymbol{y}_n$ has been converted into conditionally independent Gaussian pseudo-observations $\boldsymbol{z_{0}}_n$. This approximation can be expanded to the whole dataset.

12.9.2 N observations

First at all, we rewrite $l(\boldsymbol{Y} \mid \boldsymbol{\eta})$ for all the observations N as in equation (12.22).

$$\begin{aligned} l(\boldsymbol{Y} \mid \boldsymbol{\eta}) &\approx \\ &\approx NC_1 + \frac{1}{2}\sum_{n=1}^{n}[\boldsymbol{L_{0}}_n^T \boldsymbol{\eta}_n - \boldsymbol{L_{0}}_n^T(\boldsymbol{\eta_{0}}_n - \boldsymbol{H_{0\eta}}_n^{-1}\boldsymbol{g_{0\eta}}_n)]^T \\ &\quad [\boldsymbol{L_{0}}_n^T \boldsymbol{\eta}_n - \boldsymbol{L_{0}}_n^T(\boldsymbol{\eta_{0}}_n - \boldsymbol{H_{0\eta}}_n^{-1}\boldsymbol{g_{0\eta}}_n)]\,. \end{aligned} \tag{12.25}$$

Using the notation

$$\tilde{\boldsymbol{\eta}}_{\boldsymbol{0}} = \underbrace{\begin{bmatrix} \boldsymbol{\eta_{0}}_{\bullet 1} \\ \vdots \\ \boldsymbol{\eta_{0}}_{\bullet N} \end{bmatrix}}_{CN\times 1}, \quad \boldsymbol{g}_{\boldsymbol{0}\tilde{\boldsymbol{\eta}}} = \underbrace{\begin{bmatrix} \boldsymbol{g_{0}}_1 \\ \vdots \\ \boldsymbol{g_{0}}_N \end{bmatrix}}_{CN\times 1}, \quad \boldsymbol{L_0} = \underbrace{\begin{bmatrix} \boldsymbol{L_{0}}_1 & & 0 \\ & \ddots & \\ 0 & & \boldsymbol{L_{0}}_N \end{bmatrix}}_{CN\times CN},$$

$$\boldsymbol{H}_{\boldsymbol{0}\tilde{\boldsymbol{\eta}}} = \underbrace{\begin{bmatrix} \boldsymbol{H_{0\eta}}_1 & & 0 \\ & \ddots & \\ 0 & & \boldsymbol{H_{0\eta}}_N \end{bmatrix}}_{CN\times CN},$$

and we rewrite equation (12.25) as follows:

$$\begin{aligned} l(\boldsymbol{Y} \mid \tilde{\boldsymbol{\eta}}) &\approx \\ &\approx NC_1 + \frac{1}{2}[\boldsymbol{L_0}^T \tilde{\boldsymbol{\eta}} - \boldsymbol{L_0}^T(\tilde{\boldsymbol{\eta}}_{\boldsymbol{0}} - \boldsymbol{H}_{\boldsymbol{0}\tilde{\boldsymbol{\eta}}}^{-1}\boldsymbol{g}_{\boldsymbol{0}\tilde{\boldsymbol{\eta}}})]^T \\ &\quad [\boldsymbol{L_0}^T \tilde{\boldsymbol{\eta}} - \boldsymbol{L_0}^T(\tilde{\boldsymbol{\eta}}_{\boldsymbol{0}} - \boldsymbol{H}_{\boldsymbol{0}\tilde{\boldsymbol{\eta}}}^{-1}\boldsymbol{g}_{\boldsymbol{0}\tilde{\boldsymbol{\eta}}})]\,. \end{aligned} \tag{12.26}$$

Defining

$$\tilde{\boldsymbol{z}}_{\mathbf{0}} := \boldsymbol{L_0}^T(\tilde{\boldsymbol{\eta}}_{\mathbf{0}} - \boldsymbol{H}_{\mathbf{0}\tilde{\boldsymbol{\eta}}}{}^{-1}\boldsymbol{g}_{\mathbf{0}\tilde{\boldsymbol{\eta}}}) = \boldsymbol{L_0}^T\tilde{\boldsymbol{\eta}}_{\mathbf{0}} - \boldsymbol{L_0}^{-1}\boldsymbol{g}_{\mathbf{0}\tilde{\boldsymbol{\eta}}}\,, \tag{12.27}$$

we obtain $p(\tilde{\boldsymbol{z}}_{\mathbf{0}} \mid \tilde{\boldsymbol{\eta}})$,

$$\tilde{\boldsymbol{z}}_{\mathbf{0}} \mid \tilde{\boldsymbol{\eta}} \;\sim\; \mathcal{N}(\boldsymbol{L_0}^T\tilde{\boldsymbol{\eta}}, \boldsymbol{I}_{\boldsymbol{CN}})\,, \tag{12.28}$$

and the observation vector $\boldsymbol{Y}_{\bullet n}$ has been turned into Gaussian conditionally independent pseudo-observations $\tilde{\boldsymbol{z}}_{\mathbf{0}}$, a likelihood which `R-INLA` can deal with.

Chapter 13

Final remarks and future work

In this Thesis we have sought to provide an updated vision of the use of the latest statistical tools that have been emerging in the application of species distribution models (SDMs) in real contexts from a Bayesian perspective. The applications presented have arisen from questions proposed by experts in various areas. Our second big aim has been to develop new methodological tools to solve some statistical problems that have appeared during the application of SDMs in those real problems. In particular:

1. In order to model the production of *Plurivorosphaerella nawae* ascospores in persimmon leaf litter, we have proposed a hierarchical Bayesian beta regression method, which is going to be implemented in a warning system to help farmers of the Valencian Comunity to optimize fungicide sprays for disease control.

2. In order to study the spatial and climatic factors associated with the geographic distribution of the citrus black spot disease, in addition to a climate descriptive analysis, we have developed a spatial hierarchical Bayesian logistic model which help experts to a better understanding of the phenomenon.

This study leads us to the conclusion that, although climate was advocated as the main factor limiting the establishment and spread of CBS into new areas, our study indicates that spatial proximity to affected areas was also relevant in the geographic distribution of the CBS.

3. In order to analyze the effects of geographic genetic structure and spatial autocorrelation on species distribution range shifts, we have developed spatial hierarchical Bayesian beta regression models .

   With this study, we have seen that, Maxent and non-spatial beta regression models presented some drawbacks, such as the loss of accessions with high genetic admixture in the case of Maxent, and the presence of residual spatial autocorrelation (SAC) for both. Spatial beta regression models removed residual SAC, showed higher accuracy than non-spatial beta regression models, and handled the spatial effect on model outcomes.

   We conclude that these hierarchical beta regression models enrich the toolbox of software available to evaluate GCC-induced distribution range shifts considering both geographic genetic heterogeneity and SAC.

4. In order to study the bottlenose dolphin (*Tursiops truncatus*) distribution, a non-stationary hierarchical Bayesian logistic models has been employed. This approach constitutes a major step forward in the understanding of cetacean species in many ecosystems where physical, geographical and topographical barriers are present, and it leads us to the conclusion that an effective conservation programme should take into account these findings: favourable areas for bottlenose dolphins should be identified and protected as SACs (Special Areas of Conservation). Protection measures should be devoted to limiting the disturbance from recreational boats, which is probably the main threat for this species in the area.

5. In order to learn about the statistical problems of interest in the SDMs context, we have performed a detailed review of some statistical issues in species distribution modeling. We conclude that INLA is a powerful tool to deal with SDMs making it possible to perform complex

models with a minimum computational effort while obtaining accurate estimates.

6. In order to implement the Bayesian Dirichlet regression in the INLA context, we have presented an approximation using the Laplace method, showing again that this deterministic tool to do Bayesian inference is extremely powerfull.

In overall, in this PhD, we have proposed a number of model structures that have quite effectively tackled some challenges in areas such as ecology or plant disease epidemiology. In addition, we have proposed a methodology to deal with compositional data in the INLA context, the Dirichlet regression. However, the scope of research is still extensive. Here is a list of topics that we consider of special interest in this context.

1. Apply the models proposed in this Thesis to different areas or other problems in ecology or plant disease epidemiology, for example, to study the factors associated with the distribution of *Xylella fastidiosa*, which is a lethal plant disease pathogen affecting olives and almonds in Spain, or, the study of the Mediterranea Sea biodiversity.

2. Extent the Dirichlet regression to models with random effects, where all the possibilities addressed in `R-INLA` package or `inlabru` (`https://sites.google.com/inlabru.org/inlabru`) package can be incorporated to the model.

3. Extent the approximation that we have done for the Dirichlet regression to other regressions with multivariate response: multinomial, multivariate normal, etc.

General references

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 139–177.

Aitchison, J. and Egozcue, J. J. (2005). Compositional data analysis: where are we and where should we be heading? *Mathematical Geology*, 37(7):829–850.

Araújo, M. B., Pearson, R. G., Thuiller, W., and Erhard, M. (2005). Validation of species-climate impact models under climate change. *Global Change Biology*, 11(9):1504–1513.

Aschmann, H. (1973). Distribution and peculiarity of Mediterranean ecosystems. In *Mediterranean type ecosystems*, pages 11–19. Springer.

Bakka, H., Rue, H., Fuglstad, G.-A., Riebler, A., Bolin, D., Illian, J., Krainski, E., Simpson, D., and Lindgren, F. (2018). Spatial modeling with r-inla: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(6):e1443.

Bakka, H., Vanhatalo, J., B.Illian, J., Simpson, D., and Rue, H. (2019). Non-stationary Gaussian models with physical barriers. *Spatial Statistics*, 29:268 – 288.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*. CRC.

Barndorff-Nielsen, O. and Cox, D. (1989). *Asymptotic Techniques for Use in Statistics*. Boca Raton, FL: Chapman and Hall/CRC.

Bathe, K.-J. (2006). *Finite element procedures*. Klaus-Jurgen Bathe.

Bayes, T. (1764). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418.

Beale, C. M., Lennon, J. J., Yearsley, J. M., Brewer, M. J., and Elston, D. A. (2010). Regression analysis of spatial data. *Ecology Letters*, 13(2):246–264.

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433):109–122.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1–20.

Blangiardo, M. and Cameletti, M. (2015). *Spatial and spatio-temporal Bayesian models with `R-INLA`*. John Wiley & Sons.

Brown, C. J., O'connor, M. I., Poloczanska, E. S., Schoeman, D. S., Buckley, L. B., Burrows, M. T., Duarte, C. M., Halpern, B. S., Pandolfi, J. M., Parmesan, C., and Richardson, A. J. (2016). Ecological and methodological drivers of species distribution and phenology responses to climate change. *Global Change Biology*, 22:1548–1560.

Clark, J. and Gelfand, A. (2006). *Hierarchical Modeling for the Environmental Sciences. Statistical Methods and Applications*. Oxford University Press, New York.

Cressie, N. (1990). The origins of kriging. *Mathematical Geology*, 22(3):239–252.

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., and Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, 88(11):2783–2792.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., et al. (2012a). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1):27–46.

Dormann, C. F., Schymanski, S. J., Cabral, J., Chuine, I., Graham, C., Hartig, F., Kearney, M., Morin, X., Römermann, C., Schröder, B., et al. (2012b). Correlation and process in species distribution models: bridging a dichotomy. *Journal of Biogeography*, 39(12):2119–2131.

EFSA, European Food Safety Authority (2016). Evaluation of new scientific information on *Phyllosticta citricarpa* in relation to the EFSA PLH Panel (2014). Scientific Opinion on the plant health risk to the EU. *EFSA Journal*, 14:4513.

Elith, J. and Leathwick, J. (2017). Boosted regression trees for ecological modeling. *R documentation. Available at https://cran. r-project. org/web/packages/dismo/vignettes/brt. pdf.*

Elith, J. and 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.

Evans, J. S., Murphy, M. A., Holden, Z. A., and Cushman, S. A. (2011). Modeling species distribution and change using random forest. In *Predictive species and habitat modeling in landscape ecology*, pages 139–159. Springer.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate statistical modelling based on generalized linear models*. Springer Science & Business Media.

Fatima, S. H., Atif, S., Rasheed, S. B., Zaidi, F., and Hussain, E. (2016). Species distribution modelling of *Aedes aegypti* in two dengue-endemic regions of Pakistan. *Tropical Medicine & International Health*.

Ferrari, S. L. P. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799 – 815.

Fitzpatrick, M. C., Weltzin, J. F., Sanders, N. J., and Dunn, R. R. (2007). The biogeography of prediction error: why does the introduced range of

the fire ant over-predict its native range? *Global Ecology and Biogeography*, 16(1):24–33.

Fourie, P. H., Schutte, G. C., Carstens, E., Hattingh, V., Paul, I., Magarey, R. D., Gottwald, T. R., Yonow, T., and Kriticos, D. J. (2017). Scientific critique of the paper "Climatic distribution of citrus black spot caused by Phyllosticta citricarpa. A historical analysis of disease spread in South Africa" by Martínez-Minaya et al.(2015). *European Journal of Plant Pathology*, 148(3):497–502.

Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2018). Constructing priors that penalize the complexity of gaussian random fields. *Journal of the American Statistical Association*.

Geisser, S. (2013). *Predictive inference*, volume 55. CRC Press.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian Data Analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA.

Guisan, A., Edwards, T. C., and Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling*, 157(2):89–100.

Guisan, A. and Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, 8(9):993–1009.

Haining, R. P. (2003). *Spatial data analysis: theory and practice*. Cambridge University Press.

Handcock, M. S. and Stein, M. L. (1993). A Bayesian analysis of kriging. *Technometrics*, 35(4):403–410.

Heinze, G., Wallisch, C., and Dunkler, D. (2018). Variable selection–a review and recommendations for the practicing statistician. *Biometrical Journal*, 60(3):431–449.

Hjelle, Ø. and Dæhlen, M. (2006). *Triangulations and applications*. Springer Science & Business Media.

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

Iverson, L. R., Schwartz, M. W., and Prasad, A. M. (2004). How fast and far might tree species migrate in the eastern united states due to climate change? *Global Ecology and Biogeography*, 13(3):209–219.

JCR, Journal Citation Reports Social Sciences Edition (2017). (clarivate analytics, 2018).

Juan, P., Díaz-Avalos, C., Mejía-Domínguez, N. R., and Mateu, J. (2017). Hierarchical spatial modeling of the presence of chagas disease insect vectors in argentina. a comparative approach. *Stochastic Environmental Research and Risk Assessment*, 31(2):461–479.

Kneib, T., Müller, J., and Hothorn, T. (2008). Spatial smoothing techniques for the assessment of habitat suitability. *Environmental and Ecological Statistics*, 15(3):343–364.

Krainski, E. T., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., and Rue, H. (2018). *Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA*. Chapman and Hall/CRC.

Köppen, W. and Geiger, G. (1936). Das geographische system der klimat. *Handbuch der klimatologie*, page 44.

Laplace, P. S. (1812). *Théorie analytique des probabilitérue*. Courcier.

Lindgren, F. and Rue, H. (2015). Bayesian spatial modelling with `R-INLA`. *Journal of Statistical Software, Articles*, 63(19):1–25.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10(4):325–337.

Luo, M. and Opaluch, J. J. (2011). Analyze the risks of biological invasion. *Stochastic environmental research and risk assessment*, 25(3):377–388.

Marcer, A., Méndez-Vigo, B., Alonso-Blanco, C., and Picó, F. X. (2016). Tackling intraspecific genetic structure in distribution models better reflects species geographical range. *Ecology and Evolution*, 6(7):2084–2097.

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

Martínez-Minaya, J., Conesa, D., López-Quílez, A., and Vicent, A. (2018). Spatial and climatic factors associated with the geographical distribution of citrus black spot disease in South Africa. A Bayesian latent Gaussian model approach. *European Journal of Plant Pathology*, 151(4):991–1007.

Martins, T. G., Simpson, D., Lindgren, F., and Rue, H. (2013). Bayesian computing with `inla`: new features. *Computational Statistics and Data Analysis*, 67:68–83.

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

Peel, M. C., Finlayson, B. L., and McMahon, T. A. (2007). Updated world map of the Köppen-Geiger climate classification. *Hydrology and Earth System Sciences Discussions*, 4(2):439–473.

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

Peterson, A. T., Sánchez-Cordero, V., Beard, C. B., and Ramsey, J. M. (2002). Ecologic niche modeling and potential reservoirs for chagas disease, Mexico. *Emerging Infectious Diseases*, 8(7):662–667.

Phillips, S. J., Anderson, R. P., and Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3):231–259.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*. Vienna, Austria.

Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., Vesk, P. A., and McCarthy, M. A. (2014). Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution*, 5(5):397–406.

Roos, N. C., Carvalho, A. R., Lopes, P. F., and Pennino, M. G. (2015). Modeling sensitive parrotfish (Labridae: Scarini) habitats along the Brazilian coast. *Marine Environmental Research*, 110:92–100.

Rossi, V., Salinari, F., Pattori, E., Giosuè, S., and Bugiani, R. (2009). Predicting the dynamics of ascospore maturation of *Venturia pirina* based on environmental factors. *Phytopathology*, 99(4):453–461.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall.

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

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2017). Bayesian computing with `INLA`: a review. *Annual Review of Statistics and Its Application*, 4:395–421.

Rufener, M.-C., Kinas, P. G., Nóbrega, M. F., and Oliveira, J. E. L. (2017). Bayesian spatial predictive models for data-poor fisheries. *Ecological Modelling*, 348:125–134.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the*

Royal Statistical Society: Series B (Statistical Methodology), 64(4):583–639.

Stein, M. (1999). *Interpolation of Spatial Data. Some Theory for Kriging.* Springer.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography*, 46:234–240.

Umlauf, N., Adler, D., Kneib, T., Lang, S., and Zeileis, A. (2015). Structured Additive Regression Models: An R Interface to BayesX. *Journal of Statistical Software*, 63(21):1–46.

Van Vuuren, D. P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., Hibbard, K., Hurtt, G. C., Kram, T., Krey, V., Lamarque, J.-F., Masui, T., Meinshausen, M., Nakicenovic, N., Smith, S. J., and Rose, S. K. (2011). The representative concentration pathways: an overview. *Climatic Change*, 109:5.

Vehtari, A. and Ojanen, J. (2012). A survey of bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228.

Walker, G. A., Robertson, M. P., Gaertner, M., Gallien, L., and Richardson, D. M. (2017). The potential range of *Ailanthus altissima* (tree of heaven) in South Africa: the roles of climate, land use and disturbance. *Biological Invasions*, 19(12):3675–3690.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec):3571–3594.

Wolf, C., Novak, M., and Gitelman, A. I. (2017). Bayesian characterization of uncertainty in species interaction strengths. *Oecologia*, 184(2):327–339.