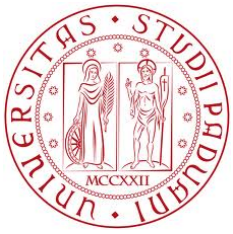

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE STATISTICHE
CICLO XXV

STATISTICAL MODELS IN BIOGEOGRAPHY

Direttore della Scuola: Prof. MONICA CHIOGNA

Supervisore: Prof. CARLO GAETAN

Co-supervisore: Dr. LUIGI SPEZIA

Dottorando: RICARDO ALVARADO-BARRANTES

Data consegna tesi 30 gennaio 2013

A Jorge

Cuando se viaja en pos de un objetivo,
es muy importante prestar atención al Camino.
Paulo Coelho

Acknowledgements

I am very grateful for working with Carlo Gaetan as a supervisor. Each time I visited him at the Universitá Ca'Foscari di Venezia, I had the impression of being in front of a great master in the field of Spatial Statistics. He always gave patient and detailed explanations of the different topics that were new for me. I definitely learnt a lot from him. Thanks to Luigi Spezia for his work as a co-supervisor and for letting me visit him at The James Hutton Institute in Aberdeen, United Kingdom.

A very important component of this work was the application to real data. I had very useful conversations about behavioural ecology of birds with Cecilia Soldatini in Venice that enlightened the results that I was obtaining from months of work.

During the time of doing this research, writing programs took a big portion of my time and efforts. Two smart persons helped me a lot in letting my programs work: Luca Sartore in Padua and Colin Cotinat in Aberdeen. I really appreciate their help. I also want to express my sincere gratitude to Susan Katz for reading the whole thesis and correcting many language mistakes, and to Clovis Kenne for giving advice all the time as a companion on the same road.

I thank the support from the Universidad de Costa Rica and from the Department of Statistics of that university, especially from the director Edgar Gutiérrez who encouraged me to get my PhD.

I want to leave here a big *gracias* to all the people that believed in this adventure. It was always very encouraging to read or listen words of appreciation from my students in Costa Rica, friends from all over, and my family.

Padua, January 29, 2013 *Ricardo Alvarado Barrantes*

Sommario

Ci concentriamo sui metodi statistici utilizzati in Biogeografia per modellare la distribuzione spaziale delle specie di uccelli. A causa della difficoltà nello specificare una struttura multivariata congiunta della covarianza spaziale nei processi ambientali, fattorizziamo tale distribuzione congiunta in una serie di modelli condizionati connessi asieme in un modello gerarchico. Abbiamo un processo che corrisponde ad una mappa non osservabile con le informazioni effettive su una specie di uccelli, ed i dati corrispondono alle osservazioni che sono collegate a tale processo. Vengono utilizzati gli approcci di simulazione *Markov chain Monte Carlo* (MCMC) per i modelli a piú livelli che incorporano strutture di dipendenza. Usiamo un algoritmo Bayesiano per estrarre campioni dalla distribuzione a posteriori al fine di ottenere stime dei parametri e ricostruire la vera immagine basata sui dati. Presentiamo diversi metodi per superare il problema del calcolo della distribuzione del campo aleatorio markoviano che viene utilizzato nell' algoritmo MCMC. Durante l'analisi, é opportuno eliminare alcuni predittori dal modello e utilizzare solo un sottoinsieme di covariate nella procedura di stima. Usiamo il metodo di Kuo & Mallick (1998) (KM) per la selezione delle variabili che, combinato all'uso dei piú catene independenti, incrementa con successo il *mixing* delle catene. Negli studi di simulazione, presentiamo le migliori prestazioni della pseudo-verosimiglianza rispetto agli altri metodi di approssimazione e le buone prestazioni del metodo KM per queso tipo di dati. Illustriamo l'applicazione dei metodi con l'analisi completa della distribuzione spaziale di due specie di uccelli (*Sturnella magna* e *Anas rubripes*), basandoci su di un insieme di dati reale. Dimostriamo i vantaggi nell'uso della struttura latente e del parametro di interazione spaziale nel modello spaziale markoviano latente rispetto agli altri modelli piú semplici, come l'ordinario modello logistico o il modello autologistico senza errori di osservazione.

Abstract

We concentrate on the statistical methods used in Biogeography for modelling the spatial distribution of bird species. Due to the difficulty of specifying a joint multivariate spatial covariance structure in environmental processes, we factor such a joint distribution into a series of conditional models linked together in a hierarchical framework. We have a process that corresponds to an unobservable map with the actual information about a bird species, and the data correspond to the observations that are connected to that process. Markov chain Monte Carlo (MCMC) simulation approaches are used for models involving multiple levels incorporating dependence structures. We use a Bayesian algorithm for drawing samples from the posterior distribution in order to obtain estimates of the parameters and reconstruct the true map based on data. We present different methods to overcome the problem of calculating the distribution of the Markov random field that is used in the MCMC algorithm. During the analysis it is desirable to delete some of the predictors from the model and only use a subset of covariates in the estimation procedure. We use the method by Kuo & Mallick (1998) (KM) for variable selection and combine it with multiple independent chains which successfully improves the mixing behaviour. In simulation studies we show the better performance of the pseudo-likelihood over other likelihood approximation methods, and the good performance of the KM method with this type of data. We illustrate the application of the methods with the complete analysis of the spatial distribution of two bird species (*Sturnella magna* and *Anas rubripes*) based on a real data set. We show the advantages of using the hidden structure and the spatial interaction parameter in the spatial hidden Markov model over other simpler models, like the ordinary logistic model or the autologistic model without observation errors.

Contents

List of Tables

List of Figures

Acronyms

APL	adjusted PL.
EAM	expanded autologistic model.
EMC	evolutionary Monte Carlo.
ERM	expanded regression model.
GLM	generalized linear model.
GVS	Gibbs variable selection.
HMRF	hidden MRF.
IQR	inter-quartile range.
KM	Kuo & Mallick (1998).
MCMC	Markov chain Monte Carlo.
MCMF	Markov chain of Markov field.
MIC	multiple independent chains.
MKMK	Metropolized-Kuo-Mallick.
MRF	Markov random field.
PL	pseudo-likelihood.
PS	path sampling.
RA	ratio approximation.

SHMM	spatial hidden Markov model.
SSVS	stochastic search variable selection.
WSEV	window sub-sampling empirical variance.

List of Symbols

N	Number of sites.
$\mathcal{S}$	Set of sites.
q	Number of covariates in the estimation procedure.
p	Number of covariates included in the variable selection procedure.
$\boldsymbol{y}$	Vector of observations.
$\boldsymbol{x}$	Vector of hidden values.
$j \sim i$	j is a neighbour of i.
$\boldsymbol{x}_{-i}$	Set of sites different from the i^{th} site.
$\boldsymbol{x}_{(i)}$	Set of neighbours of the i^{th} site.
$\boldsymbol{z}_i$	Vector of covariates for the i^{th} site.
$\boldsymbol{y}^r$	Vector of replicated data.
β	Spatial interaction.
ϕ_0	External field.
$\boldsymbol{\phi}$	Vector of regression parameters.
$\theta_{0,i}$	Probability of false observation at the i^{th} site.
$\theta_{1,i}$	Probability of true observation at the i^{th} site.
$\boldsymbol{\theta}$	Vector of probabilities associated with observation errors.
α_k	Hyperparameter for $\theta_{0,i}$ and $\theta_{1,i}$.
$\boldsymbol{\psi}$	Vector of unknown parameters.
γ_k	Auxiliary indicator variable for variable selection.

δ_k	Auxiliary variable for the effect size of each covariate.
p_m	Mutation rate.
$\mathcal{H}$	Data model.
$C(\beta, \boldsymbol{\phi}, z)$	Normalising constant.
$f(\boldsymbol{y}\|.)$	Distribution of the data.
$f(\boldsymbol{x}\|.)$	Distribution of the MRF.
$\pi(.)$	Prior distribution.
$p(\boldsymbol{x}, \psi\|.)$	Posterior distribution.
$q(\beta^*\|\beta^{(t-1)})$	Transition distribution of β.
ρ_β	Acceptance ratio of β.
J_β	Jacobian of β.
Ω	Multidimensional grid of points for PS.
s_1	Number of the original covariates erroneously excluded.
s_2	Number of the additional covariates erroneously included.
$\delta_{\boldsymbol{x}}(\boldsymbol{x}^*)$	Misclassification rate.

Chapter 1

Introduction

1.1 Overview

Biogeographical studies intend to give a description of the spatial distribution of animal or vegetal species. The information is obtained via atlas surveys in which the study area is divided into a grid of sites, which are typically squares of equal size. The purpose is to record the presence of the target species at each site. In this type of study, the resulting map with the observations tend to underestimate the true presence because of coverage problems in the fieldwork producing "non-detected presences"(Heikkinen & Högmander, 1997). On the contrary, presence could be recorded in sites that are actually not inhabited by the target species, referred to as "false observations."These two types of observation errors require the inclusion of two maps in the model, the actual map of real presence/absence of the species, and the observed map.

In the formulation of a statistical model, it is expected that adjacent sites tend to have the same condition of presence or absence of the species. The force of this association is included in the model as a parameter that is considered constant for all the sites in the study area; thus the probability of presence of the species in a specific site is determined by its presence in the neighbouring sites, weighted by this spatial interaction parameter. This simplification of the dependence of a specific site on the rest of the configuration to a local dependence of that site on its neighbourhood, creates a Markov random field (MRF), and the fact that the series of values are arranged in a two dimensional way produces a spatial MRF. Moreover, since the observed map could differ from the true one, the latter is said to be hidden and the model is referred as a spatial hidden Markov model (SHMM).

When relevant environmental information is available for each study site, corresponding to climatic and land use covariates, it can be used to refine the model that explains the probability of presence of the target species. The inclusion of this information produces a non-homogeneous model since the probability of presence of the species becomes dependent on the site. Besag (1974) proposed the autologistic model for spatial data with binary responses. In the first formulation of this model, no explanatory variables were included; however, Hughes et al. (2010) defined the autologistic model using covariates as part of a term that captures the non-homogeneity of the model. When a high number of covariates is available, it is important to define a procedure to select a suitable subset of covariates that ensures the most reliable predictions. Many methods for variable selection have been proposed but their limitations increase with the number of covariates. Starting with the method presented by George & McCulloch (1993), who used a hierarchical normal mixture model, Kuo & Mallick (1998) proposed another method that includes indicator variables embedded in the regression equation in such a way that all possible submodels are considered. Their method was modified later by Paroli & Spezia (2008) who proposed the acceptance of the vector of the covariate coefficients and the vector of the indicator variables in a single Metropolis step. They demonstrated that the algorithm has a better performance in the case of non-homogeneous hidden Markov models and Markov switching auto-regressive models when the explanatory covariates are strongly correlated.

In the following chapters we develop the ideas of hierarchical models in a Bayesian approach for the construction of statistical models in Biogeography. In Chapter 2 , we present a review of the ideas and methods of Biogeography and explain the details of the data that are analysed in Chapter 6. Chapters 3 and 4 are devoted to giving the methodological bases for the estimation procedures used in hierarchical models and variable selection. In Chapter 5 we present several simulation studies with the aim of testing the methodology exposed in Chapters 3 and 4. Finally, in Chapter 6 we perform the analysis of the spatial distribution of two selected species based on a real data set.

1.2 Main contributions of the thesis

The main contributions that we make with this thesis can be divided in two categories. In the first category we consider the contributions made to the statistical methodology. These results are obtained in Chapter 5 using simulation studies:

- We show the better performance of the pseudo-likelihood approximation over other methods of approximation to the likelihood function. In particular, we study the path sampling and the ratio of two normalising constants.

- We prove the good performance of the method by Kuo & Mallick (1998) for variable selection in the case of presence/absence data with high observation errors. We consider the special case when these errors are influenced by the amount of effort in the data collection stage. This method has been applied to time series and multiple regression data. With our simulations we show that it can be successfully applied when the data have a spatial structure and there is a hidden image in addition to the observed data.

The second type of contributions are in the applicative area. We present a complete analysis in Chapter 6 where we use a real data set:

- We show the advantages of using the hidden structure and the spatial interaction parameter in the spatial hidden Markov model over other simpler models like the ordinary logistic model or the autologistic model without observation errors. We present the analysis of the spatial distribution of two species of birds: *Sturnella magna* and *Anas rubripes*, commonly known as the Eastern Meadowlark and American Black Duck, respectively.

- A useful application of a SHMM is the creation of maps based on the posterior probabilities of presence rather than the reconstructed map as proposed by Heikkinen & Högmander (1994). Furthermore, this type of map could be created for new scenarios that reflect climate change and intensive land use, showing their effect on the distribution of the target species.

Chapter 2

Biogeography

In this chapter we present a review of general ideas related to Biogeography and introduce the problem of spatial autocorrelation and related sources. We concentrate on the statistical methods used in modelling the spatial distribution of bird species as a particular problem which has been considered by many ecologists and approached by various means. The interest in this problem is motivated by a particular data set which we describe. We explain the details of how these data are transformed to be used for the analysis presented in Chapter 6.

2.1 Biogeographical studies

Biographers observe, record and explain the geographic ranges of living things (Pielou, 1979). A geographic range of a wildlife population at some point in time can be defined as the collection of the locations of the individuals in the population at that moment (Gaston, 1994). Obtaining the complete collection of these locations becomes an impossible task in most cases, thus a solution is mapping the geographic ranges to coarsen the geographic accuracy which is used in the recording or reducing of the temporal resolution (duration of the study). The use of statistics in Biogeography is important because of the stochastic nature of the models used to explain the observed phenomena. One of the branches of Biogeography is the analysis of data from atlas surveys which is a popular method for mapping geographic ranges (Heikkinen & Högmander, 1997).

Spatial autocorrelation in data is a common phenomenon in ecology. Tobler (1970) formulates his Law of Geography to explain this concept in a natural

way stating that "everything is related to everything else, but near things are more related than distant things." Sources of spatial autocorrelation can be divided into exogenous and endogenous factors. Habitat preferences for spatially structured environmental gradients are exogenous factors (e.g. climate, soil type, stochastic disturbances) and may lead to a similar probability of occurrence in neighbouring sites, simply because the external factors show a specific autocorrelation pattern. Endogenous factors are due to the biology of the species under consideration (e.g. dispersal, colonial breeding, home-range size, competition, host availability, predation, parasitization risk) (Dormann, 2007). Exogenous factors can ideally be included into the statistical model as environmental covariates, reducing and even removing the residual spatial autocorrelation, while endogenous causes of spatial autocorrelation are usually much more difficult to quantify. In the analysis of species distribution data, ignoring spatial autocorrelation may lead to two kinds of possible errors: biased parameter estimates and standard errors that are more optimistic than they should be (Dormann, 2007). Although the consideration of spatial autocorrelation is fundamental, models may be wrongly specified because they contain the wrong explanatory variables which may lead to far worse models than ignoring spatial autocorrelation (Haining, 2003).

The problem of spatial autocorrelation in species distribution data has been addressed since Augustin et al. (1996), when they estimated the geographical distribution of plant and animal species from incomplete field survey data. They were inspired by the use of generalized linear models (GLMs) for modelling wildlife distributions, and noticed that in GLMs, spatial autocorrelation in the residuals was ignored. The approach they proposed used the autologistic model already introduced by Besag (1974), and was based on the extension of the logistic model to include an extra covariate derived from the responses at neighbouring sites. Several applications since then have been presented in modelling the distribution of plant species, insects, amphibians, birds, and mammals (e.g. Wu & Huffer (1997), Gumpertz et al. (2000), Knapp et al. (2003), Osborne et al. (2001), Teterukovsky & Edenius (2003)). Dormann et al. (2007) presented a review of different frequentist methods to account for spatial autocorrelation which are divided in three categories: 1) autocovariate regression and spatial eigenvector mapping; 2) generalized least squares methods; and 3) generalized estimating equations.

On the Bayesian side, the use of Bayesian image analysis in estimating biogeographical distributions was introduced by Högmander & Møller (1993), while Heikkinen & Högmander (1994) developed an alternative fully Bayesian

approach to decide whether the blank squares in a map of observations of common toads (*Bufo bufo*) indicate true absence or merely a lack of study there. This is essentially an image restoration problem, but it has properties that make the common empirical Bayesian procedures inadequate. Maps derived from observed occurrences of wildlife distributions surveys are profoundly affected by the duration and intensity of the observation and by methods used to delineate distributions, especially when detection is uncertain (Sargeant et al., 2005); thus these maps are imperfect and incomplete images. In this sense, maps of observations are analogous to degraded digital images, which commonly are restored via Markov chain Monte Carlo (MCMC) methods (Green, 1995a). Sargeant et al. (2005) produced an estimate of the underlying distribution of the Swift Fox (*Vulpes velox*) in western Kansas, rather than a map of observed occurrences, that reflected the uncertainty associated with estimates of model parameters. They used MRFs without including habitat covariates in their model. The problem of modelling the distribution of plant species in terms of variables such as temperature and rainfall was studied by Wu & Huffer (1997) also with an approach using MCMC methods, although not under the Bayesian framework.

The concept of MRF is explained in detail in Chapter 2 as the basis of the approach that we develop in this thesis. The inclusion of the aspects that we have introduced in this review, when modelling distributions of animal or vegetal species is crucial to obtaining a better explanation of the phenomenon that we study, which is the spatial distribution of the target species. These considerations include the important spatial autocorrelation, the exogenous factors as covariates that can also reduce or even remove the residual spatial autocorrelation, and the fact that the observed map of observations are imperfect representations of the actual map.

2.2 Data

In subsequent chapters, we will explain different approaches to the analysis of the spatial distribution of two species of birds in the Northeastern part of the United States. The data come from the eBird Reference Data set run by the National Audubon Society and the Cornell Lab of Ornithology (Munson et al., 2011). We select a region composed by the neighbouring states of New York, New Jersey, Pennsylvania, Rhode Island, Connecticut, New Hampshire, Massachusetts, Vermont, Delaware, Washington D.C., Maryland, Virginia, and West Virginia. We analyse the presence of the species *Sturnella magna* and *Anas*

rubripes (Fig. 2.1), commonly known as the Eastern Meadowlark and American Black Duck, respectively. These two species have the characteristic that the first one is more generalist while the second one is more specialist. A generalist species is able to thrive in a wide variety of environmental conditions and can make use of a variety of different resources, while a specialist species can only thrive in a narrow range of environmental conditions or have a limited diet. We analyse the impact of some covariates on the distribution of each type of species.

Figure 2.1: *An individual of the species Sturnella magna (Eastern Meadowlark) perched on a post (Salaroli, 2011) and a male preening showing speculum of the species Anas rubripes (American Black Duck)(Boland, 2007) .*

We use data that correspond to a total of 205,304 sampling events for this area in 2010. In some cases, the same location was visited more than once during the year; in those cases, we consider the species as observed in that location if it was observed at least in one of the events. We get a total of 35,878 different locations. In an atlas survey the study area is divided into a grid of sites, which are typically squares of equal size. The aim is to confirm at each site whether it is inhabited by the target species or not (Heikkinen & Högmander, 1997). Since the available data are not in a grid format, as it occurs in an atlas survey, we artificially create a grid of 15km by 15km squares or sites. This size of square is close to the size used in atlas surveys and at the same time avoids excess number of sites with missing information. Atlas surveys are usually vast projects and a typical size of an atlas square is 10km by 10km. The size of the site determines the spatial resolution (Heikkinen & Högmander, 1997).

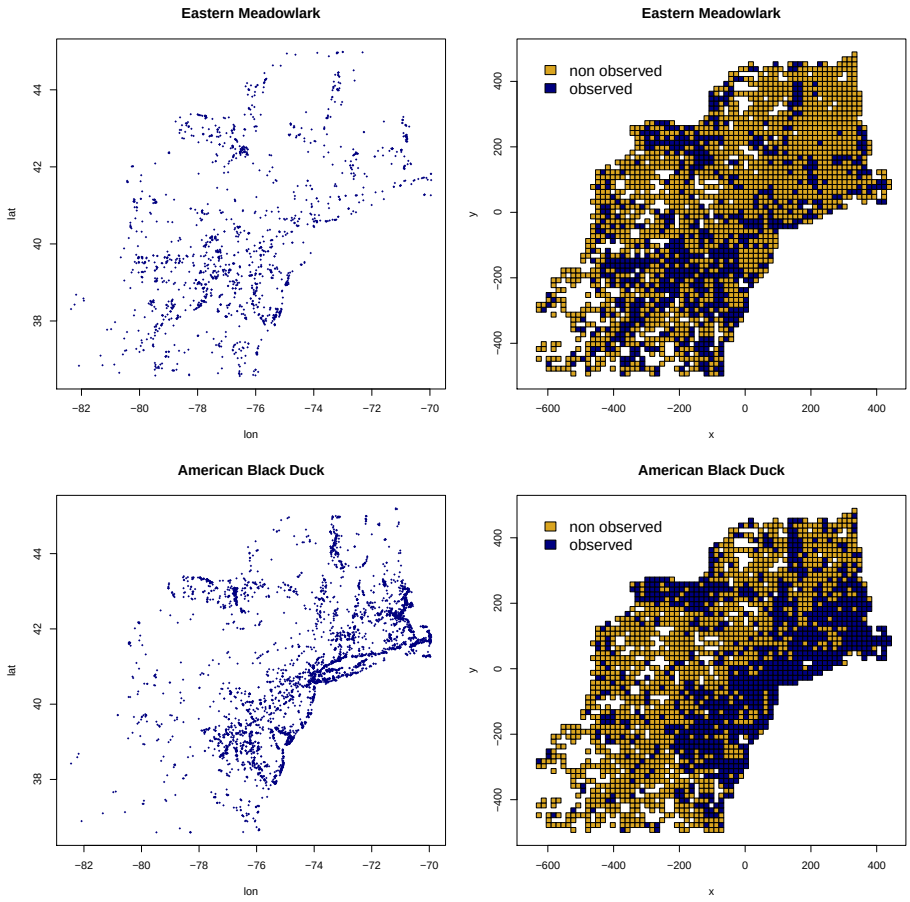

Figure 2.2: *Presence of the Eastern Meadowlark (top) and the American Black Duck (bottom) in a selected area of Northeastern U.S. Left: original sites where the species was observed (blue dots) at corresponding longitude and latitude. Right: 15km by 15km sites at x-y coordinate system with two colours to differentiate sites where the species was observed and not observed.*

The latitude and longitude of each location is projected in a x-y coordinate system using the *spectralGP* library (Paciorek, 2007) in R. The projection calculates (for all points) the great circle distance in the x direction to the mean longitude and in the y direction to the mean latitude, and uses these distances as the x-y coordinates of the location. In each square we consider the species as observed if it is observed in at least one of the locations sampled inside that site. The result is an irregular grid of 2,195 sites with information about the presence or absence of the species, and several squares in the interior with missing information (see Fig. 2.2).

The chance of observing the species of interest in a specific site can be influenced by many factors that are external to the covariates which we want to model, and can be considered as noise induced either by the observers or by the different conditions during the sampling activities. Osborne & Tigar (1992) and Heikkinen & Högmander (1994) include a measurement of the observational effort as a covariate to account for uneven coverage. In our data set, the number of sample events varies a lot from location to location, 55% of the locations were sampled only once, but there are extreme cases as high as 1,384 samples (a case where samples were taken several times a day during for many days). Also the number of locations visited in each site varies over a wide range with a median of 18 locations per site, 52 sites visited in only one location and a few sites visited in more than 400 locations. Thereby, in addition to the information provided by the static environment covariates and the habitat statistics, the total effort hours devoted to each site (sum of the effort hours in all the locations visited in a site) should be taken into account. The duration of the observation in hours is available for 88.4% of the 205,304 locations. The missing values are imputed using the average of the effort hours for the locations in the same site that do have that information. If all the locations in a site have missing values, we use the median of all the available locations. Once we have imputed the missing values, we add the number of effort hours for all the locations in each site. In Fig. 2.3 we observe that in the middle and upper section towards the right of the study area, the observation effort tends to be stronger than in the rest.

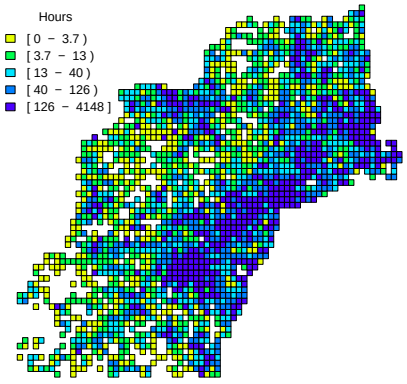

Figure 2.3: *Effort hours per site.*

We use the available information for each location to create covariates at square level using kriging interpolation (Cressie, 1993). Kriging is a weighted linear combination of the known sample values around the point to be estimated. We use the *fields* library (Furrer et al., 2012) in R, with a maximum of 200 known values for each interpolation, chosen in such a way that they are the closest points to the square for which we do the interpolation. Each square is divided into a smaller grid with 36 points separated 5km from each other; we obtain the estimated values for these points and take their mean as an estimate of the variable in the square. In Fig. 2.4, we present the values of the variable Patch density of deciduous forest for the points in the square formed by the set of vertices $\{(-150, -325), (-125, -325), (-125, -300), (-150, -300)\}$. The square has 10 points and we use a bigger region for the interpolation that contains 301 points. The values of the variable inside the small square have a range from 2.9 to 25.9. The average of these values is 17.4, while the estimated values for the 36 points in the grid (crosses in red) produce an average of 14.2. Fig. 2.5 shows the results of the interpolation procedure for the variable Elevation categorised into 5 groups according to quantiles (for illustration purposes).

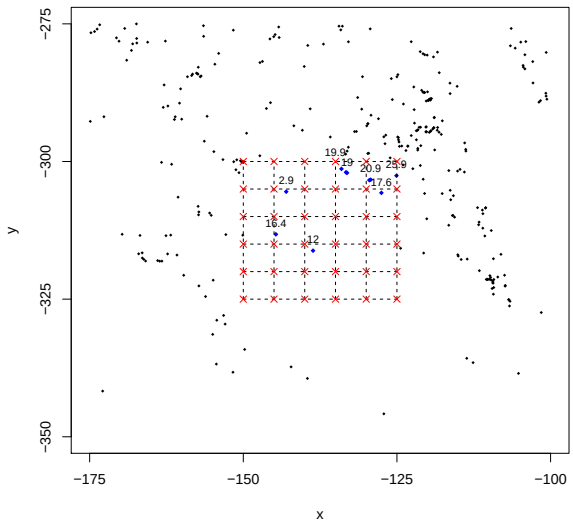

Figure 2.4: *Interpolation of the values in the grid (red crosses) using all 301 points in the big square. Original points within the small square in blue with the corresponding values of the variable Patch density of deciduous forest.*

We have some measurements that are static in the sense that they are derived from environmental snapshots tied to a wide time frame independent of exactly when the observations are made. We select 9 static environment covariates (Table 2.1) based on a preliminary descriptive analysis to make sure that they are spread in a range that allows some variability for an interesting future predictive analysis. Two of the covariates (CANOPY_MEAN and IMPERV_MEAN) are recorded at three different spatial extents to cover local ecological processes at small, medium and large ranges. The scale of the covariate is indicated by the "radius"of the neighbourhood (in meters). The neighbourhood border is twice the radius since it represents the radius of a circle inscribed within the neighbourhood square.

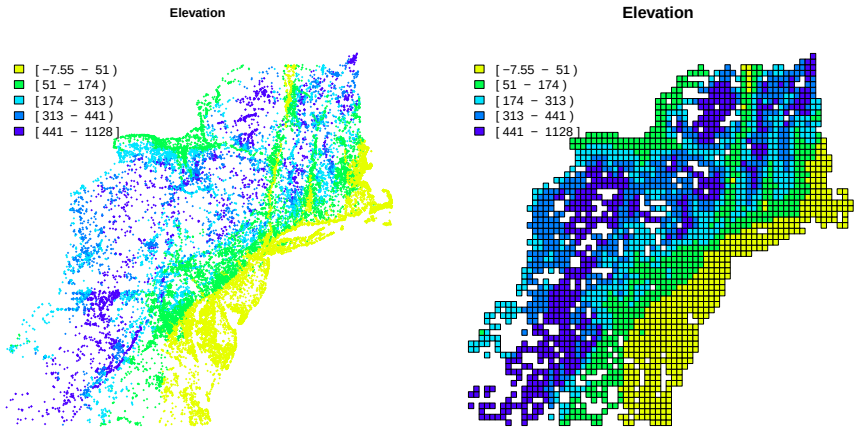

Figure 2.5: *Plot of a variable in the original locations (left) and the interpolation in 15km by 15km sites (right).*

Habitat statistics are computed from the National Land Cover Data from source data for the year 2006 (Fry et al., 2011). They include a landscape statistic (LPI), and land cover statistics (PD) by class using 14 categories described in Table 2.2. These statistics are calculated for the three different scales (75, 750 and 7500), but we use only the 750 radius (see Table 2.2). We end up with 9 static environment covariates, one landscape statistic, and 14 class level statistics, which results in a total of 24 covariates to start the analysis.

Variable Name	Description
	Static Environment Covariates
POP	Population per square mile (2000 census) for the census block group containing the location.
ELEV	Elevation (in meters). Horizontal resolution is roughly 1km by 1km.
T_AVG	Mean daily average temperature(1).
PREC	Mean total precipitation(1).
CANOPY_MEAN	Mean canopy cover (in meters) in square neighbourhood around location with radius 750m.
IMPERV_MEAN	Mean imperviousness to water (in meters) in square neighbourhood around location with radius 750m.
DIST_FLOW_FW	Distance (in meters) from flowing fresh water.
DIST_STD_FW	Distance (in meters) from standing fresh water.
DIST_WETVEG_FW	Distance (in meters) from wet vegetation, fresh water.
	Habitat Statistics
LPI	Percentage of landscape area occupied by the largest patch (any habitat class) around location with radius 750m.
CC_PD(2)	Patch density. Number of patches of habitat class CC(3) per 100 hectares in surrounding landscape with radius 750m.

(1) Annual aggregate statistics averaged over 30 years (1961-1990).

(2) NA occurs if habitat class is not in landscape, recoded as 0.

(3) Habitat classes are described in Table 2.2.

Table 2.1: *Covariate list.*

Class	Description	Class	Description
C11	Open water.	C43	Mixed forest.
C21	Developed, open space.	C52	Shrub/scrub.
C22	Developed, low intensity.	C71	Grassland/herbaceous.
C23	Developed, medium intensity.	C81	Pasture/hay.
C24	Developed, high intensity.	C82	Cultivated crops.
C41	Deciduous forest.	C90	Woody wetlands.
C42	Evergreen forest.	C95	Herbaceous wetlands.

Table 2.2: *Habitat classes.*

Chapter 3

Hierarchical modelling

Due to the difficulty of specifying a joint multivariate spatial covariance structure in environmental processes, it may be much easier to factor such joint distributions into a series of conditional models linked together in a hierarchical framework (Wikle, 2003). For complicated processes in the presence of data, the idea is to approach the problem by breaking it into three primary stages:

Stage 1. Data model: [data | process, parameters].

Stage 2. Process model: [process|parameters].

Stage 3. Parameter model: [parameters].

In biogeographical studies, the process corresponds to an unobservable map with the actual information about an animal or vegetal species, and the data correspond to the observations that are connected to that process. The first stage is concerned with the observations or "data model", which specifies the distribution of the data given the process of interest and some parameters. The second stage describes the process, conditional on other parameters. Finally, the last stage accounts for the uncertainty in the parameters. In applications, each of these stages may have multiple sub-stages.

Modelling the parameters as random instead of fixed effects allows the introduction of spatial autocorrelation structures among them, hence among the observed data as well (Banerjee et al., 2003). Hierarchical Bayesian methods enjoy broad application in the analysis of spatial data. When prior information is included in the model, the level of complexity increases and the Bayesian paradigm becomes necessary. Markov chain Monte Carlo (MCMC) simulation approaches like the Gibbs sampler and Metropolis-Hastings algorithm are

popular for almost any model involving multiple levels incorporating dependence structures (Gelman et al., 2003).

In this chapter we explain the three stages of a spatial hidden Markov model and how they are connected in a hierarchical way. Then, we introduce the Bayesian algorithm for drawing samples from the posterior distribution in order to obtain estimations of the parameters and reconstruct the true map based on data. Subsequently, we present different methods to overcome the problem of calculating the distribution of the Markov random field that is used in the MCMC algorithm.

3.1 Spatial hidden Markov models

Spatial hidden Markov models (SHMM) are pairs of stochastic processes $(\{\boldsymbol{X}\}, \{\boldsymbol{Y}\})$, where $\boldsymbol{X}$ is a collection of N variables from a latent unobserved process, and $\boldsymbol{Y}$ is a vector of N observed variables normally called data or observations that masks or hides the process $\boldsymbol{X}$. This process $\{\boldsymbol{X}\}$ is referred to as a hidden MRF (HMRF) which is a map degraded by (conditionally) independent noise. We emphasise in the term HMRF its latent unobserved nature and the Markov random field (MRF) assumption that is explained later. One context in which such models have been much used is image analysis, going back to Besag (1986) and beyond; another is disease mapping in epidemiology (e.g., Elliott et al. (2000)).

We deal with a special type of the SHMMs where each $\boldsymbol{X}$ and $\boldsymbol{Y}$ are binary vectors with state space $\{0, 1\}$. In this section we explain each stage of a SHMM in order to model the data and construct the posterior distribution of our interest, which is the joint distribution of the latent process and the parameters, given the data.

3.1.1 Data model

In biogeographical studies, the information is obtained via atlas surveys in which the study area is divided into a grid of sites, which are typically squares of equal size. The purpose is to record the presence of the target species at each site. In this type of study, the resulting map with the observations tend to underestimate the true presence because of coverage problems in the fieldwork producing non-detected presences (Heikkinen & Högmander, 1997). In the case of bird surveys, many factors influence the probability of detecting birds during auditory point counts. There are "measurement error" factors

associated with observer skill in identifying and localising individual birds, and hearing ability. There are other factors that include the spectral qualities of songs, song volume, singing rate, time of day, the orientation of singing birds (toward or away from observers), the number of species and number of individuals singing during a count, pairing status, stage of nesting cycle, mobility of bird species (mobile species may move in and out of the count area), vegetation structure, topography, weather, temperature, humidity, and ambient noise. Systematic variation in any of these factors will impart a systematic bias in count data (Simons et al., 2009).

As opposed to non-detected presences, the target species could be registered as present in sites that are actually not inhabited by that species. These type of records are referred to as false observations. These two types of observation errors require the inclusion of two maps in the model, the actual map of real presence/absence of the species, and the observed map. Thus, we assume that there is a true map of sites that we do not observe perfectly, but we collect data instead that are connected to that map.

In the first stage of the hierarchical model, we want to obtain the distribution of the data given the latent process. We define $\mathcal{S}= \{1, \ldots, N\}$, the set of sites. The true map is represented by the vector of hidden values $\boldsymbol{x}= (x_1, \ldots, x_N)'$, while $\boldsymbol{y}= (y_1, \ldots, y_N)'$ is the vector that corresponds to the observed map.

In our biogeographical application, when a species is observed at the i^{th} site we assign value 1 to Y_i, $Y_i = 1$, and 0 when it is not observed, $Y_i = 0$. The X_i values indicate if a species is actually present, $X_i = 1$, or absent, $X_i = 0$. The observed value y_i could differ from the corresponding true value x_i (at the same site) due to a false observation or because a real presence is not observed.

We want to connect the data with the actual values. We express this connection through the conditional distribution of the data given the hidden process and the parameters that are associated to the probability of observation errors. Let $\boldsymbol{\theta}= (\theta_0, \theta_1)'$, where θ_0 is the probability of a false observation (i.e. $\theta_0 = \Pr(Y_i = 1|X_i = 0)\,(i \in \mathcal{S})$), while θ_1 is the probability of a true observation (i.e. $\theta_1 = \Pr(Y_i = 1|X_i = 1)\,(i \in \mathcal{S})$).

We assume that $\boldsymbol{Y}$, given $\boldsymbol{X}$, is a vector of conditionally independent random variables, and that the conditional distributions of Y_i $(i \in \mathcal{S})$ are independent of the true values in other pixels, hence the conditional distribution can be factorised as

$$f(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i \in \mathcal{S}} f(y_i|\boldsymbol{x}) = \prod_{i \in \mathcal{S}} f(y_i|x_i)\,.$$

We assume that the conditional distribution of Y_i, given X_i, is Bernoulli

with the corresponding probability θ_0 or θ_1, thus we model the data given the latent process as

$$f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) = \prod_{i\in\mathcal{S}} \theta_{x_i}^{y_i}(1-\theta_{x_i})^{1-y_i}\,, \tag{3.1}$$

where we use the abbreviated notation θ_{x_i} to represent θ_0 when $X_i = 0$, and θ_1 when $X_i = 1$.

We could also consider the case when the observation errors are also non-homogeneous (i.e. they are site dependent). We could take into consideration covariate information, e.g. we use effort hours instead of research activity used by Heikkinen & Högmander (1994). Thereby, we can model $\theta_{0,i}$ (probability of false observation at the i^{th} site) and $\theta_{1,i}$ (probability of true observation at the i^{th} site) as a function of effort hours. The vector of parameters used in this case is $\boldsymbol{\theta} = (\theta_{0,1},\ldots,\theta_{0,N},\theta_{1,1},\ldots,\theta_{1,N})'$, and the conditional distribution of the data given the latent process is expressed as

$$f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) = \prod_{i\in\mathcal{S}} \theta_{x_i,i}^{y_i}(1-\theta_{x_i,i})^{1-y_i}\,, \tag{3.2}$$

where we use the abbreviated notation $\theta_{x_i,i}$ to represent $\theta_{0,i}$ when $X_i = 0$, and $\theta_{1,i}$ when $X_i = 1$.

3.1.2 Process model

Regardless of the data that we obtain, there exists a true map that, when there is observation error, is not observed. This map is obtained from a distribution of maps with some characteristics. In this stage we want to explain how these maps are produced.

We base the construction of a spatial model on the conditional probabilities. A special type of spatial model is the MRF (Besag, 1974) which is based on the assumption that the dependence of a specific site on the rest of the configuration is reduced to the local dependence of that site on its neighbourhood. The MRF assumption can be expressed in the following way:

$$\Pr[x_i|\boldsymbol{x}_{-i}] = \Pr[x_i|\boldsymbol{x}_{(i)}]\ \forall i \in \mathcal{S}\,,$$

where $\boldsymbol{x}_{-i}= \{x_j : j \in \mathcal{S}, j \neq i\}$ and $\boldsymbol{x}_{(i)}$ is the set of neighbours of the i^{th} site. When the map corresponds to a lattice of regular sites, as is the case of our application, the neighbouring set of the i^{th} site is defined as the set of nearby sites within a radius of r. In the first order neighbourhood system, every (interior) site has four neighbours (horizontally or vertically adjacent), while in a second-order neighbourhood there are eight neighbours for every (interior) site (additionally including diagonal adjacencies).

Under the MRF assumption, we need an expression for the conditional distribution of X_i ($i \in \mathcal{S}$), given its neighbours. For the case of a binary MRF, Besag (1974) introduced the autologistic model. The idea of the autologistic model is to add an extra explanatory variable to the logistic model that captures the effect of other response values in the spatial neighbourhood. The conditional probability of a single X_i in the autologistic isotropic model is defined as

$$f(x_i|\boldsymbol{x}_{(i)}, \phi_0, \beta) = \frac{\exp\left[x_i\left(\phi_0 + \beta\sum_{j\sim i} x_j\right)\right]}{1 + \exp\left(\phi_0 + \beta\sum_{j\sim i} x_j\right)}, \tag{3.3}$$

where $j \sim i$ indicates that j is a neighbour of i, ϕ_0 is called the external field and β is called the spatial interaction parameter. The coefficient ϕ_0 controls the relative abundance of 0's and 1's, while β represents the strength of the spatial interaction amongst sites, large values of β lead to realisations of $\{\boldsymbol{X}\}$ having patches of 0's and 1's. In the simpler case when there is a single spatial interaction parameter β regardless of the direction of the location of the neighbour, we are in front of an isotropic process (Li, 1995); otherwise, when there is more than one parameter that is direction dependent we have anisotropy.

It is possible to incorporate covariate information into the model structure. When covariates are included in the model, the conditional probability of presence is related to the values of the covariates in the site where the probability is evaluated. In this case, the model becomes non-homogeneous. The homogeneity is a property of independence of the relative position of the site where the probability is calculated (Li, 1995). For q spatial covariates we define the vector of covariates for the i^{th} site as $\boldsymbol{z}_i = (1, z_{i1}, \ldots, z_{iq})'$ ($i \in \mathcal{S}$). The conditional probability (3.3) can be modified in the following way:

$$f(x_i|\boldsymbol{x}_{(i)}, \beta, \boldsymbol{\phi}, \boldsymbol{z}_i) = \frac{\exp\left[x_i\left(\beta\sum_{j\sim i} x_j + \boldsymbol{z}_i'\boldsymbol{\phi}\right)\right]}{1 + \exp\left(\beta\sum_{j\sim i} x_j + \boldsymbol{z}_i'\boldsymbol{\phi}\right)}, \tag{3.4}$$

where $\boldsymbol{\phi} = (\phi_0, \ldots, \phi_q)'$ is the $(q+1)$ vector of regression parameters including the external field ϕ_0. This is just one way to introduce covariates in the autologistic model: by modifying the marginal distributions.

Finally, we are able to write the joint probability distribution of the MRF, which is given by

$$f(\boldsymbol{x}|\beta, \boldsymbol{\phi}, \boldsymbol{z}) = \frac{1}{C(\beta, \boldsymbol{\phi}, \boldsymbol{z})} \exp\left[\sum_{i\in\mathcal{S}} x_i\left(\frac{\beta}{2}\sum_{j\sim i} x_j + \boldsymbol{z}_i'\boldsymbol{\phi}\right)\right], \tag{3.5}$$

where $C(\beta, \boldsymbol{\phi}, z)$ is a normalising constant calculated over all possible vectors

$\boldsymbol{x}$:

$$C(\beta, \boldsymbol{\phi}, \boldsymbol{z}) = \sum_{\boldsymbol{x}} \exp\left[\sum_{i \in \mathcal{S}} x_i \left(\frac{\beta}{2} \sum_{j \sim i} x_j + \boldsymbol{z}_i' \boldsymbol{\phi}\right)\right]. \tag{3.6}$$

In general, the number of possible vectors can be extremely high; in the case for which we want to apply the procedure, the length of $\boldsymbol{x}$ is $N = 2,195$, thus we have a total of 2^{2195} vectors and the calculation becomes computationally prohibitive.

3.1.3 Parameter model

In the third stage of the hierarchy we have the parameters. Due to their uncertainty, in the Bayesian approach to statistical analysis, the vector of unknown parameters $\boldsymbol{\psi} = (\beta, \boldsymbol{\phi}', \boldsymbol{\theta}')'$ is supposed to be a random quantity (Banerjee et al., 2003). The distribution of $\boldsymbol{\psi}$ is denominated the prior distribution $\pi(\boldsymbol{\psi}|\boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ is in turn a vector of hyperparameters. We assume that the distributions of β, $\boldsymbol{\phi}$, and $\boldsymbol{\theta}$ are a priori mutually independent, i.e. $\pi(\boldsymbol{\psi}|\boldsymbol{\lambda}) = \pi(\beta|\boldsymbol{\lambda})\,\pi(\boldsymbol{\phi}|\boldsymbol{\lambda})\,\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})$.

We choose a uniform distribution on $[0, B]$ $(B > 0)$ as the prior for β. The spatial interaction parameter can be assumed positive for the kind of data that we want to analyse where we expect that neighbouring sites tend to have a similar condition of either presence or absence of the species. Negative spatial autocorrelation can occur as a function of processes such as competition and allelopathy, typically at finer scales than positive autocorrelation (Miller, 2012). The larger the β, the more probability exists that large patches with the same condition appear. When this parameter is extremely high (usually more than 3.5, based on simulations), vectors with only 0's or only 1's are produced, thus $B = 3.5$.

To avoid vectors with a composition of only 0's or only 1's, we use a compact parameter space for the coefficients ϕ_k $(k = 0, \ldots, q)$. We choose independent normal priors with mean 0 and high variance σ^2, truncated in the range $[-A, A]$. Since the variables in our database are all standardised, we expect these coefficients ϕ_k $(k = 0, \ldots, q)$ to be no greater than 5 in absolute value, thus $A = 5$. The variance is selected to obtain a relatively diffuse distribution. The expression for the prior of $\boldsymbol{\phi}$ is

$$\pi(\boldsymbol{\phi}|\sigma^2, A) \propto \prod_{k=0}^{q} \exp\left(-\frac{1}{2\sigma^2}\phi_k^2\right) I_{[-A,A]}(\phi_k)\,.$$

For the probabilities associated to observation errors, since θ_1 corresponds to a correct event while θ_0 corresponds to an error, it is natural to expect that

$\theta_0 < \theta_1$. The priors for these parameters are not invariant which produces non-identifiability of these two components, which would lead to so called label switching in the MCMC output. Following the recommendation by Frühwirth-Schnatter (2001) to impose an identifiability constraint based on expert judgement when the parameters have a physical meaning, in order to avoid label switching, it seems reasonable to require that $\theta_0 < \theta_1$. A simple prior could be a uniform prior over the triangle; $\pi(\boldsymbol{\theta}) \propto I_{(0,1)}(\theta_0)\, I_{(0,\theta_0)}(\theta_1)$ holds the previous constraint. In the case when we take into consideration effort hours with non-homogeneous observation errors, we model $\theta_{0,i}$ and $\theta_{1,i}$ in the following way:

$$\begin{aligned} \theta_{0,i} &= \frac{\exp(\alpha_1 + \alpha_2 w_i)}{1 + \exp(\alpha_1 + \alpha_2 w_i)} \\ \theta_{1,i} &= \frac{\exp(\alpha_3 + \alpha_4 w_i)}{1 + \exp(\alpha_3 + \alpha_4 w_i)}\, (i \in \mathcal{S})\,, \end{aligned} \tag{3.7}$$

where α_k $(k = 1, \ldots, 4)$ are hyperparameters with independent diffuse Gaussian priors with a mean of 0 and high variance, and w_i is the effort hours at the i^{th} site. The vector of parameters to be estimated becomes $\boldsymbol{\psi} = (\beta, \boldsymbol{\phi}', \boldsymbol{\alpha}')'$ and the joint prior of the parameters $\pi(\boldsymbol{\psi}) = \pi(\beta)\, \pi(\boldsymbol{\phi})\, \pi(\boldsymbol{\alpha})$.

The posterior distribution, $p(\boldsymbol{x}, \psi|.)$, summarises the current state of knowledge about all the uncertain quantities (including unobservable parameters and also missing, latent, and unobserved potential data) under the Bayesian approach (Gelman, 2002). We can construct the posterior distribution of interest, which is the joint distribution of the process (MRF) and parameters $(\boldsymbol{\psi})$, updated by the data $(\boldsymbol{Y})$ with some fixed explanatory covariates $(\boldsymbol{Z})$ and effort hours $(\boldsymbol{W})$. Although $\boldsymbol{W}$ is a covariate, we do not include it in the set of explanatory covariates since it is used to model the observation errors. We assume that the true map $(\boldsymbol{X})$ comes from a distribution that is independent of the observation errors in $\boldsymbol{\theta}$. We consider that the parameters in β and $\boldsymbol{\phi}$, as well as the covariates in $\boldsymbol{Z}$, affect directly only the MRF $(\boldsymbol{X})$, while $\boldsymbol{Y}$ is independent from all these elements, given $\boldsymbol{X}$. The posterior distribution is given by

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{\psi}\,|\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{w}) &\propto f(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\psi}, \boldsymbol{z}, \boldsymbol{w})\, f(\boldsymbol{x}|\boldsymbol{\psi}, \boldsymbol{z}, \boldsymbol{w})\, \pi(\boldsymbol{\psi}) \\ &= f(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{w})\, f(\boldsymbol{x}|\beta, \boldsymbol{\phi}, \boldsymbol{z})\, \pi(\beta)\, \pi(\boldsymbol{\phi})\, \pi(\boldsymbol{\alpha})\,, \end{aligned} \tag{3.8}$$

where $f(\boldsymbol{y}|.)$ corresponds to the distribution of the data model given in (3.2) with the definition of $\theta_{0,i}$ and $\theta_{1,i}$ given in (3.7), $f(\boldsymbol{x}|.)$ is the distribution of the process model given in (3.5), and $\pi(.)$ are the prior distributions of the parameter model.

3.2 Markov chain Monte Carlo algorithm for SHMMs

In this section we explain the algorithm to draw samples from the posterior distribution of $\boldsymbol{X}$ and ψ given in (3.8). Due to the factorised form of this distribution, we can proceed in a sequence of steps within the framework of MCMC methods. This is a general method based on drawing values of a parameter from approximate distributions and then correcting those draws to better approximate the target posterior distribution (Gelman et al., 2003). Two particular Markov chain simulation methods are used, the Gibbs sampler and the Metropolis-Hastings. The Gibbs sampler is used to draw from the distribution of the vector $\boldsymbol{X}$, and each X_i is considered a parameter to be updated. We use Metropolis-Hastings to draw from the marginal posterior distributions of β, $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. The whole MCMC algorithm can be split into four steps at each iteration. At the t^{th} iteration we draw $\boldsymbol{x}^{(t)}$, $\beta^{(t)}$, $\boldsymbol{\phi}^{(t)}$, and $\boldsymbol{\alpha}^{(t)}$ that depend on the previous draw $\boldsymbol{x}^{(t-1)}$, $\beta^{(t-1)}$, $\boldsymbol{\phi}^{(t-1)}$, and $\boldsymbol{\alpha}^{(t-1)}$. We explain each of these steps to draw the t^{th} sample.

Step 1. Given the previous vector $\boldsymbol{x}^{(t-1)}$, we draw $\boldsymbol{x}^{(t)}$ using Gibbs sampler. This vector has N parameters, thus we draw sequentially the N components $x_i^{(t)}$ $(i \in \mathcal{S})$ from the conditional distribution

$$f(x_i^{(t)} \,|\, \boldsymbol{x}_{(i)}, \beta, \boldsymbol{\phi}, \boldsymbol{z}_i, y_i, \boldsymbol{\theta}) \propto f(x_i^{(t)} \,|\, \boldsymbol{x}_{(i)}, \beta, \boldsymbol{\phi}, \boldsymbol{z}_i) f(y_i | x_i^{(t)}, \boldsymbol{\theta}), \tag{3.9}$$

which involves the conditional probability of the autologistic model given in (3.4) and the conditional probabilities of Y_i given X_i. Thus, to obtain each component in $\boldsymbol{x}^{(t)}$ we use the conditional probability

$$f(X_i^{(t)} = 1 \,|\, \boldsymbol{x}_{(i)}, \beta, \boldsymbol{\phi}, \boldsymbol{z}_i, y_i, \boldsymbol{\theta}) = \frac{\theta_{1,i} \exp\left[\left(\beta \sum_{j\sim i} x_j + \boldsymbol{z}_i'\boldsymbol{\phi}\right)\right]}{\theta_{0,i} + \theta_{1,i} \exp\left(\beta \sum_{j\sim i} x_j + \boldsymbol{z}_i'\boldsymbol{\phi}\right)}\,. \tag{3.10}$$

At each step of the Gibbs sampler we use the values of the components that have already been updated. The other given parameters $(\beta, \boldsymbol{\phi}, \boldsymbol{\theta})$ correspond to the draws in the previous step $(\beta^{(t-1)}, \boldsymbol{\phi}^{(t-1)}, \boldsymbol{\theta}^{(t-1)})$, but to simplify the notation we leave them without the index $^{(t-1)}$.

Step 2. We use Metropolis-Hastings to update the parameter β. We draw β^* from a transition distribution that depends on the previous draw $\beta^{(t-1)}$.

We describe the procedure of drawing a generic parameter λ^* from the transition distribution $q(\lambda^*|\lambda^{(t-1)})$, in two steps. First we have a random variable

U distributed as

$$\mathcal{N}\left(\log\frac{\lambda^{(t-1)}-A}{B-\lambda^{(t-1)}},\sigma^2_{\epsilon_\lambda}\right),$$

where the values A and B are the extremes of the prior parameter space for λ. Then we draw the proposal λ^* from the distribution of the transformed U, as follows:

$$\lambda^* = \frac{A+B\exp(U)}{1+\exp(U)}.$$

This selection for the transition distribution $q(\lambda^*|\lambda^{(t-1)})$ ensures that the chain will move within the interval (A, B), which is the specified range for λ a priori (Spezia, 2010).

We accept β^* with probability $\min\{1,\rho_\beta\}$, and acceptance ratio

$$\begin{aligned}\rho_\beta &= \frac{p(\beta^*|\boldsymbol{x},\boldsymbol{\phi},\boldsymbol{z})}{p(\beta^{(t-1)}|\boldsymbol{x},\boldsymbol{\phi},\boldsymbol{z})}\frac{q(\beta^{(t-1)}|\beta^*)}{q(\beta^*|\beta^{(t-1)})}\\ &= \frac{f(\boldsymbol{x}|\beta^*,\boldsymbol{\phi},\boldsymbol{z})\pi(\beta^*)}{f(\boldsymbol{x}|\beta^{(t-1)},\boldsymbol{\phi},\boldsymbol{z})\pi(\beta^{(t-1)})}\frac{J_{\beta^{(t-1)}}}{J_{\beta^*}},\end{aligned} \tag{3.11}$$

where the proposal ratio $q(\beta^{(t-1)}|\beta^*)/q(\beta^*|\beta^{(t-1)})$ is simplified to the ratio $J_{\beta^{(t-1)}}/J_{\beta^*}$ due to the symmetry of the Gaussian distribution, with J_β being the Jacobian of the transformation, i.e. $J_\beta = \frac{B-A}{(\beta-A)(B-\beta)}$ (see explanation of this simplification in Appendix A). If β^* is accepted then $\beta^{(t)} = \beta^*$, otherwise $\beta^{(t)} = \beta^{(t-1)}$.

Step 3. We use Metropolis-Hastings to update the vector of coefficients $\boldsymbol{\phi}$. We draw independently each proposal ϕ_k^* $(k = 0,\ldots,q)$ from a transition distribution $q(\phi_k^*|\phi_k^{(t-1)})$ as explained in the previous step. We accept in block the vector $\boldsymbol{\phi}^* = (\phi_0^*,\ldots,\phi_q^*)'$ with probability $\min\{1,\rho_{\boldsymbol{\phi}}\}$, and acceptance ratio

$$\begin{aligned}\rho_{\boldsymbol{\phi}} &= \frac{p(\boldsymbol{\phi}^*|\boldsymbol{x},\beta,\boldsymbol{z})}{p(\boldsymbol{\phi}^{(t-1)}|\boldsymbol{x},\beta,\boldsymbol{z})}\frac{q(\boldsymbol{\phi}^{(t-1)}|\boldsymbol{\phi}^*)}{q(\boldsymbol{\phi}^*|\boldsymbol{\phi}^{(t-1)})}\\ &= \frac{f(\boldsymbol{x}|\beta,\boldsymbol{\phi}^*,\boldsymbol{z})\,\pi(\boldsymbol{\phi}^*)}{f(\boldsymbol{x}|\beta,\boldsymbol{\phi}^{(t-1)},\boldsymbol{z})\,\pi(\boldsymbol{\phi}^{(t-1)})}\frac{J_{\boldsymbol{\phi}}^{(t-1)}}{J_{\boldsymbol{\phi}}^*},\end{aligned} \tag{3.12}$$

where

$$J_{\boldsymbol{\phi}} = \prod_{k=0}^{q}\frac{B_k-A_k}{(\phi_k-A_k)(B_k-\phi_k)},$$

and the ratio $q(\boldsymbol{\phi}^{(t-1)}|\boldsymbol{\phi}^*)/q(\boldsymbol{\phi}^*|\boldsymbol{\phi}^{(t-1)})$ is simplified to the ratio $J_{\boldsymbol{\phi}^{(t-1)}}/J_{\boldsymbol{\phi}^*}$. If $\boldsymbol{\phi}^*$ is accepted then $\phi_k^{(t)} = \phi_k^*$, otherwise $\phi_k^{(t)} = \phi_k^{(t-1)}\ \forall\ k = 0,\ldots,q$.

Step 4. We use Metropolis-Hastings to update $\boldsymbol{\alpha}$. We draw independently α_k^* $(k = 1,\ldots,4)$ from a Gaussian transition distribution $q(\alpha_k^*|\alpha_k^{(t-1)})$ with

mean $\alpha_k^{(t-1)}$ and variance $\sigma^2_{\epsilon_\alpha}$. We proceed to the acceptance in block of $\boldsymbol{\alpha}^*$ with probability $\min\{1, \rho_{\boldsymbol{\alpha}}\}$, and acceptance ratio

$$\begin{aligned}\rho_{\boldsymbol{\alpha}} &= \frac{p(\boldsymbol{\alpha}^*|\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{\alpha}^{(t-1)}|\boldsymbol{x},\boldsymbol{y})}\frac{q(\boldsymbol{\alpha}^{(t-1)}|\boldsymbol{\alpha}^*)}{q(\boldsymbol{\alpha}^*|\boldsymbol{\alpha}^{(t-1)})} \\ &= \frac{f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\alpha}^*,\boldsymbol{w})\pi(\boldsymbol{\alpha}^*)}{f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\alpha}^{(t-1)},\boldsymbol{w})\pi(\boldsymbol{\alpha}^{(t-1)})}\,,\end{aligned} \tag{3.13}$$

where the ratio $q(\boldsymbol{\alpha}^{(t-1)}|\boldsymbol{\alpha}^*)/q(\boldsymbol{\alpha}^*|\boldsymbol{\alpha}^{(t-1)}) = 1$ due to the symmetry of the Gaussian distribution. With the previous acceptance rule we obtain $\boldsymbol{\alpha}^{(t)}$ and apply (3.7) with each value w_i $(i \in \mathcal{S})$ to obtain $\boldsymbol{\theta}^{(t)}$.

Estimation. The algorithm is run for a high number of iterations. Usually a part of the sequence is discarded to diminish the effect of the starting distribution and ensure convergence of the sequence. The discarded early iterations are referred as the burn-in period (Gelman et al., 2003). Also there is the practice of thinning the sequence, i.e. once approximate convergence has been reached, whereby every k^{th} simulation draw from the sequence is kept and the rest is discarded. The remaining draws from the sequence (after burn-in and thinning) are used directly for inferences about the parameters. We also obtain a reconstruction $\hat{\boldsymbol{x}}$ of the true map $\boldsymbol{x}$ using the empirical posterior mode of each cell in the vector $\boldsymbol{x}$ by counting the number of 0s and 1s in the vector for each x_i $(i \in \mathcal{S})$ and assigning the value with the highest frequency.

3.3 Likelihood of the hidden map

In steps 2 and 3 of the algorithm of the previous section we need to calculate the distribution of $\boldsymbol{x}$ given in (3.5). This distribution requires the calculation of the normalising constant $C(\beta, \boldsymbol{\phi}, \boldsymbol{z})$ which is a sum over all possible vectors $\boldsymbol{x}$, and, for most cases, the calculation becomes computationally prohibitive. We use a simplified notation to refer to this normalising constant as $C(\boldsymbol{\psi})$, defining the p-parameter vector as $\boldsymbol{\psi} = (\beta, \phi_0, \phi_1, \ldots, \phi_q)'$ and omitting the dependence of this constant on $\boldsymbol{z}$.

A wide range of approximative techniques and stochastic approximations have been proposed to circumvent the problem of intractable normalising constants. The fact that $C(\boldsymbol{\psi})$ also depends on the values of the covariates imposes a computational difficulty that has not been resolved. Pettitt et al. (2003) proposed a method using a computational analytical procedure where the lattice is wrapped onto the cylinder, while Reeves & Pettitt (2004) presented an exact method, termed the recursion method, used to calculate $C(\boldsymbol{\psi})$ for a normalised

distribution expressible as a product of factors. Since this proposal was restricted to small lattices, Friel et al. (2009) extended the recursion method to arbitrary sized lattices; however, it is limited in that the study area must be regular and the process homogeneous. Bartolucci & Besag (2002) also presented a recursive algorithm for directly computing the likelihood of a MRF, in the form of a product of conditional probabilities. Hardouin & Guyon (2009) presented a recursive algorithm for the calculation of the marginal of a Gibbs distribution, and as a direct consequence the calculation of the normalising constant.

Several approaches are based on Monte Carlo methods since Geyer & Thompson (1992). The stochastic approximation expectation algorithm by Gu & Zhu (2001) and Zhu et al. (2007) is used to compute the maximum likelihood estimator for HMRF models. An alternative approach was presented by Liang (2007) in which a kernel density estimate of $C(\boldsymbol{\psi})$ is formulated based on Monte Carlo draws from $f(\boldsymbol{x}|\boldsymbol{\psi})$.

New approaches avoid direct approximations of $C(\boldsymbol{\psi})$. One method is presented by Møller et al. (2006) in which they introduce an auxiliary variable into a Metropolis-Hastings algorithm for the posterior of $\boldsymbol{\psi}$. The normalising constants are cancelled out from the Metropolis-Hastings ratio. This method is applicable whenever samples may be drawn from the likelihood without approximation, by perfect sampling for example. Another method in this line is the exchange algorithm and developments by Murray (2007) which also makes use of an auxiliary variable on the support of the data. The auxiliary variable in this case depends only on an additional auxiliary variable on the support of the parameter which in turn depends on the parameter.

With numerous antecedents, Lindsay (1988) crystallized the notion of a composite likelihood, defining it as a combination of valid likelihood entities. We use one particular composite likelihood which is the pseudo-likelihood (PL) introduced by Besag (1975). Then we explore an adjustment to the PL inspired by the work by Cooley et al. (2012) for another type of composite likelihood, the pairwise likelihood. Smith & Stephenson (2009) employed a pairwise likelihood within a Bayesian analysis to estimate the parameters in a different problem of an extended Gaussian max-stable process model for spatial extremes. As an alternative to using the PL we also apply two other methods of approximation, the first is the path sampling (Gelman & Meng, 1998), which approximates $C(\boldsymbol{\psi})$ directly, and the second is an approximation of $C(\boldsymbol{\psi}^{(t-1)})/C(\boldsymbol{\psi}^{(t)})$, the ratio of two normalising constants (Zhou & Schmidler, 2009), which are needed in (3.11) and (3.12) (Metropolis steps 2 and 3 of the MCMC algorithm).

3.3.1 Pseudo-likelihood approximation

The most common approach to overcome the problem of calculating the normalising constant in the likelihood function is to use the PL approximation, first presented by Besag (1975). PL estimation has been employed in a wide variety of settings. In particular, it has been used in the current context of HMRF by, e.g. Besag et al. (1991), Ryden & Titterington (1998), and Heikkinen & Högmander (1994).

The likelihood is approximated by the PL which is the product of likelihood objects, where each likelihood contribution is based on a conditional event with probability defined in (3.4). Formally the PL is defined as

$$f_P(\boldsymbol{x}|\beta, \boldsymbol{\phi}, \boldsymbol{z}) = \prod_{i \in \mathcal{S}} \frac{\exp\left[x_i \left(\beta \sum_{j \sim i} x_j + \boldsymbol{z}_i' \boldsymbol{\phi}\right)\right]}{1 + \exp\left(\beta \sum_{j \sim i} x_j + \boldsymbol{z}_i' \boldsymbol{\phi}\right)} . \tag{3.14}$$

The posterior distributions based on the PL tend to have higher variances than those that could be obtained based on the full likelihood. We apply an adjustment method (Cooley et al., 2012) modifying the curvature of the PL in order to get a more concentrated posterior with the practical aim of using the PL to provide valid inferences. The strategy consists of modifying the curvature of the PL around its maximum $\hat{\boldsymbol{\psi}}_c$. For a given parameter $\boldsymbol{\psi}$, the PL is calculated at the adjusted vector

$$\boldsymbol{\psi}^* = (\beta^*, \boldsymbol{\phi}^{*\prime})' = \hat{\boldsymbol{\psi}}_c + \boldsymbol{D}(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}_c) ,$$

where $\boldsymbol{D}$ is a $p \times p$ matrix. We write the logarithm of the PL function as

$$p\ell(\boldsymbol{\psi}; \boldsymbol{x}) = \sum_{i \in \mathcal{S}} \left[x_i \boldsymbol{w}_i' \boldsymbol{\psi} - \log(1 + \exp(\boldsymbol{w}_i' \boldsymbol{\psi}))\right] , \tag{3.15}$$

where $\boldsymbol{w}_i = (\sum_{j \sim i} x_j, 1, z_{i1}, \ldots, z_{iq})'$, we omit the conditioning on $\boldsymbol{z}$ and use the notation $p\ell(\boldsymbol{\psi}; \boldsymbol{x})$ to emphasise that we treat $\boldsymbol{\psi}$ as a variable while $\boldsymbol{x}$ is fixed. The logarithm of the adjusted PL (APL) function becomes

$$p\ell_A(\boldsymbol{\psi}; \boldsymbol{x}) = p\ell(\boldsymbol{\psi}^*; \boldsymbol{x}) = \sum_{i \in \mathcal{S}} \left[x_i \boldsymbol{w}_i' \boldsymbol{\psi}^* - \log(1 + \exp(\boldsymbol{w}_i' \boldsymbol{\psi}^*))\right] . \tag{3.16}$$

There is not a unique way to select the matrix $\boldsymbol{D}$, but it must be semi-definite negative such that

$$\boldsymbol{D}' \boldsymbol{H}(\boldsymbol{\psi}_0) \boldsymbol{D} = \boldsymbol{H}(\boldsymbol{\psi}_0) \boldsymbol{J}(\boldsymbol{\psi}_0)^{-1} \boldsymbol{H}(\boldsymbol{\psi}_0)' ,$$

where $\boldsymbol{\psi}_0$ is the true parameter, $\boldsymbol{H}(\boldsymbol{\psi}_0) = -\mathbb{E}[\nabla^2 p\ell(\boldsymbol{\psi}_0; \boldsymbol{x})]$, and $\boldsymbol{J}(\boldsymbol{\psi}_0) = \mathrm{Var}[\nabla p\ell(\boldsymbol{\psi}_0; \boldsymbol{x})]$. The PL function for a vector is the sum of the values of the

function calculated at each element of the vector, i.e. $p\ell(\boldsymbol{\psi}_0;\boldsymbol{x}) = \sum_{i\in\mathcal{S}} p\ell_i(\boldsymbol{\psi}_0;x_i)$, with $p\ell_i(\boldsymbol{\psi};x_i) = x_i\boldsymbol{w}_i'\boldsymbol{\psi} - \log(1+\exp(\boldsymbol{w}_i'\boldsymbol{\psi}))$. We take the first and second derivatives of $p\ell_i(\boldsymbol{\psi};x_i)$:

$$\begin{aligned}\frac{\partial p\ell_i(\boldsymbol{\psi};x_i)}{\partial\psi_j} &= w_{ij}\left(x_i - \frac{\exp(\boldsymbol{w}_i'\boldsymbol{\psi})}{1+\exp(\boldsymbol{w}_i'\boldsymbol{\psi})}\right),\\ \frac{\partial^2 p\ell_i(\boldsymbol{\psi};x_i)}{\partial\psi_j\psi_k} &= -w_{ij}w_{ik}\frac{\exp(\boldsymbol{w}_i'\boldsymbol{\psi})}{(1+\exp(\boldsymbol{w}_i'\boldsymbol{\psi}))^2}.\end{aligned} \tag{3.17}$$

We get an estimate of the matrix $\boldsymbol{H}$ evaluating it at the maximum PL estimate $\hat{\boldsymbol{\psi}}_c$ instead of the unknown $\boldsymbol{\psi}_0$. The estimate of $\boldsymbol{H}$ is the negative of the sum of the Fisher information for each element of the vector,

$$\hat{\boldsymbol{H}}(\hat{\boldsymbol{\psi}}_c) = -\sum_{i\in\mathcal{S}}\mathbb{E}[\nabla^2 p\ell_i(\hat{\boldsymbol{\psi}}_c;x_i)];$$

however, we use the observed information instead of the Fisher information, such that

$$\hat{\boldsymbol{H}}(\hat{\boldsymbol{\psi}}_c) = -\sum_{i\in\mathcal{S}}\nabla^2 p\ell_i(\hat{\boldsymbol{\psi}}_c;x_i). \tag{3.18}$$

The estimation of $\boldsymbol{J}$ is more complicated since it involves the second moment of the first derivative of $p\ell(\boldsymbol{\psi};\boldsymbol{x})$. We know that $\mathbb{E}[\nabla p\ell(\boldsymbol{\psi}_0;\boldsymbol{x})] = 0$ since $\nabla p\ell(\boldsymbol{\psi};\boldsymbol{x})$ is an unbiased estimating function, then we get

$$\begin{aligned}\boldsymbol{J}(\boldsymbol{\psi}_0) &= \mathbb{E}[\nabla p\ell(\boldsymbol{\psi}_0;\boldsymbol{x})\nabla p\ell(\boldsymbol{\psi}_0;\boldsymbol{x})'] - \mathbb{E}[\nabla p\ell(\boldsymbol{\psi}_0;\boldsymbol{x})]^2\\ &= \mathbb{E}[\nabla p\ell(\boldsymbol{\psi}_0;\boldsymbol{x})\nabla p\ell(\boldsymbol{\psi}_0;\boldsymbol{x})'].\end{aligned}$$

In this case we cannot use the the maximum PL estimate $\hat{\boldsymbol{\psi}}_c$ instead of the unknown $\boldsymbol{\psi}_0$ because $\nabla p\ell(\hat{\boldsymbol{\psi}}_c;\boldsymbol{x}) = 0$ and the estimate would be always $\hat{\boldsymbol{J}}(\hat{\boldsymbol{\psi}}_c) = 0$. Heagerty & Lumley (2000) propose the use of sub-sampling with overlapping windows to estimate the variance of estimating functions. Windows are contiguous regions usually with a fixed size for which the estimated function is calculated. To show the construction process of the windows, we use matrix notation to identify the cells in the grid, i.e. the $(i,j)^{th}$ cell is located in the i^{th} row and the j^{th} column. As it is illustrated in Fig. 3.1, we start with a squared window whose upper left corner is the $(1,1)^{th}$ cell, then we choose the next window by moving the upper left corner of the window s cells to the right from the previous one, and continue in this way until we finish that row. We proceed by moving the upper left corner s cells down to complete the next row of windows and continue until we finish the entire grid. At the end we get a collection of m windows denoted $\mathcal{D}_1,\ldots,\mathcal{D}_m$.

We denote $n = |\mathcal{S}|$, where $|\mathcal{A}|$ denotes the cardinality of $\mathcal{A}$, and use $\boldsymbol{J}_n$ instead of $\boldsymbol{J}$ to make evident the influence of the size of the vector in the

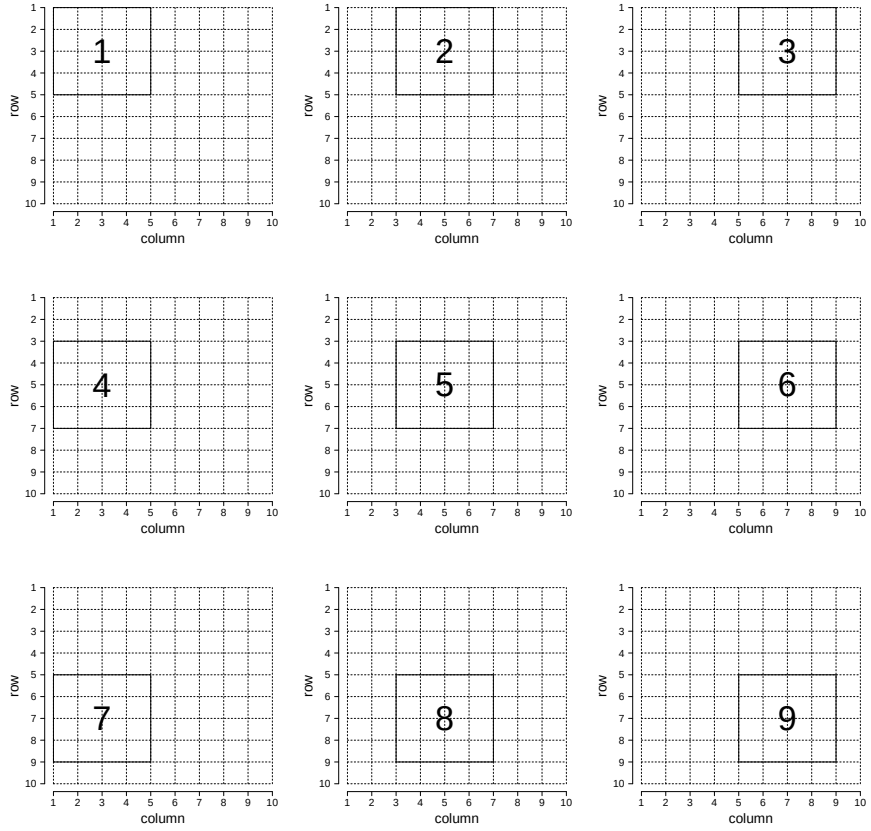

Figure 3.1: *Windows of size* 4×4 *in a* 10×10 *grid with* $s = 2$.

calculation. We assume that a limit is obtained in probability, such that

$$\frac{\boldsymbol{J}_n(\boldsymbol{\psi}_0)}{n} \xrightarrow{n\to\infty} \boldsymbol{J}_\infty(\boldsymbol{\psi}_0)\,,$$

thus we obtain first an estimate of $\boldsymbol{J}_\infty$ and then we get an estimate of $\boldsymbol{J}_n$. We define the window sub-sampling empirical variance (WSEV) estimator of $\boldsymbol{J}_\infty$ in two steps, first we get the second moment at each window as a local estimate of the variance, and next we average all these estimates resulting

$$\hat{\boldsymbol{J}}_\infty(\hat{\boldsymbol{\psi}}_c) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{|\mathcal{D}_i|}\sum_{j,k\in\mathcal{D}_i}\nabla p\ell_j(\hat{\boldsymbol{\psi}}_c; x_j)\nabla p\ell_k(\hat{\boldsymbol{\psi}}_c; x_k)'\,. \tag{3.19}$$

Finally, the estimate for $\boldsymbol{J}$ is given by

$$\hat{\boldsymbol{J}}(\hat{\boldsymbol{\psi}}_c) = n\,\hat{\boldsymbol{J}}_\infty(\hat{\boldsymbol{\psi}}_c) = |\mathcal{S}|\,\frac{1}{m}\sum_{i=1}^{m}\frac{1}{|\mathcal{D}_i|}\sum_{j,k\in\mathcal{D}_i}\nabla p\ell_j(\hat{\boldsymbol{\psi}}_c; x_j)\nabla p\ell_k(\hat{\boldsymbol{\psi}}_c; x_k)'\,. \tag{3.20}$$

3.3.2 Path sampling

Instead of using an approximation for the whole likelihood function, as in the case of the PL, we present a method here that approximates only the normalising constant. Gelman & Meng (1998) propose a method called path sampling (PS) based on the ratio of the normalising constant of our interest, denominated by $C(\boldsymbol{\psi}_1)$, and another normalising constant calculated exactly for a convenient selection of $\boldsymbol{\psi}_0$, i.e. they intend to calculate $C(\boldsymbol{\psi}_1)/C(\boldsymbol{\psi}_0)$. PS has been employed by authors (e.g. Green & Richardson (2002) and Dryden et al. (2003)), where they extend the problem to one of model selection with the number of hidden states as a parameter. They use an off-line approach to calculate the normalising constant.

In the selection of $\boldsymbol{\psi}_0$ it is very convenient to set $\beta = 0$, independently from the values of $\boldsymbol{\phi}$. In that case the normalising constant becomes

$$C(\boldsymbol{\psi}_0) = \sum_{\boldsymbol{x}} \exp\left(\sum_{i\in\mathcal{S}} x_i \boldsymbol{z}_i'\boldsymbol{\phi}\right) .$$

Although this expression seems difficult to calculate due to the inclusion of all possible vectors $\boldsymbol{x}$, it can be simplified to

$$C(\boldsymbol{\psi}_0) = \prod_{i\in\mathcal{S}} \left[1 + \exp(\boldsymbol{z}_i'\boldsymbol{\phi})\right] , \tag{3.21}$$

which does not consider $\boldsymbol{x}$ vectors (see proof in Appendix A).

We need to select a continuous path that links $\boldsymbol{\psi}_0$ and $\boldsymbol{\psi}_1$ indexed by $t \in [0, 1]$. The path is defined as $\boldsymbol{\psi}(t) = (\beta(t), \phi_0(t), \phi_1(t), \ldots, \phi_q(t))'$, with $\beta(t) = t\beta$, $\phi_k(t) = \phi_k\,(k = 0, \ldots, q)$, $\boldsymbol{\psi}(0) = \boldsymbol{\psi}_0 = (0, \boldsymbol{\phi}')'$, and $\boldsymbol{\psi}(1) = \boldsymbol{\psi}_1 = (\beta, \boldsymbol{\phi}')'$. Selecting $\boldsymbol{\psi}_0$ in this way we get the closest distance between $\boldsymbol{\psi}_0$ and $\boldsymbol{\psi}_1$ with the condition $\beta = 0$.

The normalised joint distribution of a vector is labelled $q(\boldsymbol{x}|\boldsymbol{\psi})$, while $\log q(\boldsymbol{x}|\boldsymbol{\psi})$ is called energy function and can be expressed as a linear combination of the parameters and the corresponding potentials $U_k(\boldsymbol{x})$, in the following way:

$$\begin{aligned} \log q(\boldsymbol{x}|\boldsymbol{\psi}) &= \sum_{k=1}^{p} \psi_k U_k(\boldsymbol{x}) \\ \Rightarrow U_k(\boldsymbol{x}) &= \frac{\partial \log q(\boldsymbol{x}|\boldsymbol{\psi})}{\partial \psi_k} \,;\, (k = 1, \ldots, p). \end{aligned}$$

We express the normalising constant at a given point t of the path as

$$C(\boldsymbol{\psi}(t)) = \int q(\boldsymbol{x}|\boldsymbol{\psi}(t))\mu(d\boldsymbol{x}) .$$

Taking logarithms and then differentiating both sides with respect to t, and assuming the legitimacy of interchange of integration with differentiation yields

$$\begin{aligned}
\frac{d}{dt}\log C(\boldsymbol{\psi}(t)) &= \frac{d}{dt}\log \int q(\boldsymbol{x}|\boldsymbol{\psi}(t))\,\mu(d\boldsymbol{x}) \\
&= \frac{\int \frac{d}{dt} q(\boldsymbol{x}|\boldsymbol{\psi}(t))\,\mu(d\boldsymbol{x})}{\int q(\boldsymbol{y}|\boldsymbol{\psi}(t))\,\mu(d\boldsymbol{y})} \\
&= \int \frac{\frac{d}{dt} q(\boldsymbol{x}|\boldsymbol{\psi}(t))}{q(\boldsymbol{x}|\boldsymbol{\psi}(t))} \frac{q(\boldsymbol{x}|\boldsymbol{\psi}(t))}{C(\boldsymbol{\psi}(t))}\,\mu(d\boldsymbol{x}) \\
&= \int \frac{d}{dt}\log q(\boldsymbol{x}|\boldsymbol{\psi}(t)) f(\boldsymbol{x}|\boldsymbol{\psi}(t))\,\mu(d\boldsymbol{x}) \\
&= E_{\boldsymbol{\psi}(t)}\left[\frac{d}{dt}\log q(\boldsymbol{x}|\boldsymbol{\psi}(t))\right].
\end{aligned}$$

Defining the derivative of $\psi_k(t)$ with respect to t as $\dot{\psi}_k(t)\,(k = 1, \ldots, p)$, the derivative inside the expected value can be expressed as

$$\begin{aligned}
\frac{d}{dt}\log q(\boldsymbol{x}|\boldsymbol{\psi}(t)) &= \frac{\partial \log q(\boldsymbol{x}|\boldsymbol{\psi})'}{\partial \boldsymbol{\psi}} \frac{d\boldsymbol{\psi}(t)}{dt} \\
&= \left(\frac{\partial \log q(\boldsymbol{x}|\boldsymbol{\psi})}{\partial \psi_1}, \ldots, \frac{\partial \log q(\boldsymbol{x}|\boldsymbol{\psi})}{\partial \psi_p}\right)\left(\frac{d\psi_1(t)}{dt}, \ldots, \frac{d\psi_p(t)}{dt}\right)' \\
&= (U_1(\boldsymbol{x}), \ldots, U_p(\boldsymbol{x}))\left(\dot{\psi}_1(t), \ldots, \dot{\psi}_k(t)\right)' \\
&= \sum_{k=1}^{p} \dot{\psi}_k(t) U_k(\boldsymbol{x}),
\end{aligned}$$

Thus

$$\frac{d}{dt}\log C(\boldsymbol{\psi}(t)) = E_{\boldsymbol{\psi}(t)}\left[\sum_{k=1}^{p} \dot{\psi}_k(t) U_k(\boldsymbol{x})\right].$$

Integrating both sides of the previous expression from 0 to 1 yields

$$\lambda = \log \frac{C(\boldsymbol{\psi}_1)}{C(\boldsymbol{\psi}_0)} = \int_0^1 E_{\boldsymbol{\psi}(t)}\left[\sum_{k=1}^{p} \dot{\psi}_k(t) U_k(\boldsymbol{x})\right] dt\,. \tag{3.22}$$

The PS estimator for λ is obtained using MCMC methods as

$$\hat{\lambda} = \frac{1}{m}\sum_{h=1}^{m}\left[\sum_{k=1}^{p} \dot{\psi}_k(t_h) U_k(\boldsymbol{x}^{(h)})\right], \tag{3.23}$$

where m is the number of MCMC iterations, the t_h's are drawn uniformly from $[0, 1]$, and $\boldsymbol{x}^{(h)}$ is obtained from the distribution $f(\boldsymbol{x}|\boldsymbol{\psi}(t_h))$. From (3.5) we get the energy function

$$\log(q(\boldsymbol{x}|\boldsymbol{\psi})) = \sum_{i\in\mathcal{S}} x_i \left(\frac{\beta}{2}\sum_{j\sim i} x_j + \boldsymbol{z}_i'\boldsymbol{\phi}\right),$$

with the following potentials:

$$U_1(\boldsymbol{x}) = \sum_{i\in\mathcal{S}} x_i \sum_{j\sim i} x_j \,,$$
$$U_2(\boldsymbol{x}) = \sum_{i\in\mathcal{S}} x_i \,,$$
$$U_k(\boldsymbol{x}) = \sum_{i\in\mathcal{S}} x_i z_{i,k-2} \,; (k = 3,\ldots,p) \,,$$

and the derivatives $\dot{\psi}_1(t) = \beta$, and $\dot{\psi}_k(t) = 0\,(k = 2,\ldots,p)$, which simplifies the calculation of (3.23) to

$$\hat{\lambda} = \frac{1}{m}\sum_{h=1}^{m} \frac{\beta}{2} \sum_{i\in\mathcal{S}} x_i^{(h)} \sum_{j\sim i} x_j^{(h)} \,, \tag{3.24}$$

with $\boldsymbol{x}^{(h)}$ vectors generated by Gibbs sampler with the full conditional

$$f(x_i|\boldsymbol{x}_{(i)}, \boldsymbol{\psi}(t_h)) = \frac{\exp\left[x_i\left(t_h\beta\sum_{j\sim i} x_j + \boldsymbol{z}_i'\boldsymbol{\phi}\right)\right]}{1+\exp\left(t_h\beta\sum_{j\sim i} x_j + \boldsymbol{z}_i'\boldsymbol{\phi}\right)} \,; (i\in\mathcal{S}) \,. \tag{3.25}$$

The number of iterations in the Gibbs sampler to generate each vector was defined to ensure sampling from the stationary distribution of $\boldsymbol{x}$. That number can be as low as 200 iterations. On the other hand, the number of iterations m in the MCMC was selected at a level where the estimator of $\log C(\boldsymbol{\psi}_1)$ converges. That number can be from 200 to 400 iterations (see justification below).

Finally, we get the approximation of the desired normalising constant by

$$\log C(\boldsymbol{\psi}_1) = \hat{\lambda} + \log C(\boldsymbol{\psi}_0) \,, \tag{3.26}$$

using (3.21) to calculate $C(\boldsymbol{\psi}_0)$.

Although PS provides a good approximation of the normalising constant, its direct use at each step of an MCMC algorithm is computationally expensive. We propose instead to perform the calculation of $\log C(\boldsymbol{\psi}_i)$ for a selected set of r vectors $\boldsymbol{\psi}_i\ (i = 1,\ldots,r)$ and interpolate using a regression model. The first proposal is a non-parametric regression model that uses a B-spline (BS) (Hastie & Tibshirani, 1990). The second proposal is a parametric polynomial regression (PR) defined as

$$\hat{\omega}_i = \delta_0 + \boldsymbol{\psi}_i{}'\boldsymbol{\delta} + \boldsymbol{\psi}_i{}'\boldsymbol{\Delta}\boldsymbol{\psi}_i + \epsilon_i \,; (i = 1,\ldots,r) \,, \tag{3.27}$$

where $\omega_i = \log C(\boldsymbol{\psi}_i)$, $\boldsymbol{\delta} = (\delta_1,\ldots,\delta_p)'$, $\boldsymbol{\Delta} = (\delta_{jk}), (j,k = 1,\ldots,p)$ is symmetric, and ϵ_i is a random error. There are $1 + p + \frac{1}{2}p(p+1)$ parameters to be estimated that are obtained using the least squares method.

The whole estimation algorithm is the following:

Step 1. Apply the estimation algorithm using the PL approximation and obtain estimations of β and $\boldsymbol{\phi}$ ($\hat{\beta}$ and $\hat{\boldsymbol{\phi}}$, respectively) with the corresponding posterior means.

Step 2. Create a multidimensional grid of points centered on $\boldsymbol{\psi}^* = (\hat{\beta}, \hat{\phi_0}, \hat{\phi_1}, \ldots, \hat{\phi_q})'$. The resulting grid is $\Omega = \xi_1 \times \ldots \times \xi_p$, with $\xi_k = \{\psi_k^* - d_k, \psi_k^* - c_k, \psi_k^*, \psi_k^* + c_k, \psi_k^* + d_k\}$ $(k = 1, \ldots, p)$. We choose $c_1 = \hat{\beta}/4$ and $d_1 = 3\hat{\beta}/4$ to cover a good range for β, while $c_k = 0.5, d_k = 2.5$ if $|\phi_k| \leq 5$, and $c_k = 2, d_k = 10$ if $|\phi_k| > 5$ $(k = 2, \ldots, p)$.

Step 3. For each point in Ω apply PS to obtain an approximation of the normalising constant.

Step 4. Construct a regression model to allow interpolation within the p-dimensional parameter space.

Estimation MCMC. The estimation MCMC is modified in steps 2 and 3. At each iteration, when the likelihood function needs to be calculated, we use the interpolation from the regression model as an approximation of the corresponding normalising constant.

3.3.3 Ratio approximation

Instead of the direct approximation of the normalising constant like those presented in the previous subsection, Zhou & Schmidler (2009) propose the use of importance sampling to estimate the ratio of the two normalising constants involved in the Metropolis-Hastings algorithm. We referred to it as to the ratio approximation (RA).

We start by defining the normalising constant as

$$C(\boldsymbol{\psi}) = \int q(\boldsymbol{x}|\boldsymbol{\psi})\mu(d\boldsymbol{x})\,.$$

Importance sampling requires choosing a good distribution $g(\boldsymbol{x})$ that takes values in $\{0,1\}^{\mathcal{S}}$. We multiply and divide the integrand by this distribution to yield an expectation of a certain quantity with respect to the density $g(\boldsymbol{x})$, such that

$$C(\boldsymbol{\psi}) = \int \frac{q(\boldsymbol{x}|\boldsymbol{\psi})}{g(\boldsymbol{x})} g(\boldsymbol{x})\mu(d\boldsymbol{x}) = \mathbb{E}_g\left[\frac{q(\boldsymbol{x}|\boldsymbol{\psi})}{g(\boldsymbol{x})}\right]\,. \tag{3.28}$$

We obtain m vectors $\boldsymbol{x}^{(h)}$ $(h = 1, \ldots, m)$ from the distribution $g(\boldsymbol{x})$, and calculate the estimate of the normalising constant by

$$\hat{C}(\boldsymbol{\psi}) = \frac{1}{m} \sum_{h=1}^{m} \frac{q(\boldsymbol{x}^{(h)}|\boldsymbol{\psi})}{g(\boldsymbol{x}^{(h)})} \,. \tag{3.29}$$

In the calculation of the acceptance probabilities of the Metropolis algorithm we have the likelihood ratio

$$\frac{f(\boldsymbol{x}|\boldsymbol{\psi})}{f(\boldsymbol{x}|\boldsymbol{\psi}^{(t-1)})} = \frac{q(\boldsymbol{x}|\boldsymbol{\psi})/C(\boldsymbol{\psi})}{q(\boldsymbol{x}|\boldsymbol{\psi}^{(t-1)})/C(\boldsymbol{\psi}^{(t-1)})} = \frac{q(\boldsymbol{x}|\boldsymbol{\psi})}{q(\boldsymbol{x}|\boldsymbol{\psi}^{(t-1)})} \frac{C(\boldsymbol{\psi}^{(t-1)})}{C(\boldsymbol{\psi})} \,,$$

thus we can approximate the ratio of the two normalising constants instead of calculating each one separately. The approximation becomes

$$\frac{\hat{C}(\boldsymbol{\psi}^{(t-1)})}{\hat{C}(\boldsymbol{\psi})} = \frac{1}{m} \sum_{h=1}^{m} \frac{q(\boldsymbol{x}^{(h)}|\boldsymbol{\psi}^{(t-1)})}{g(\boldsymbol{x}^{(h)})} \Bigg/ \frac{1}{m'} \sum_{h=1}^{m'} \frac{q(\boldsymbol{x}^{(h)}|\boldsymbol{\psi})}{g(\boldsymbol{x}^{(h)})} \,.$$

Zhou & Schmidler (2009) propose the use of $g(\boldsymbol{x}) = f(\boldsymbol{x}|\boldsymbol{\psi})$ which simplifies the calculation to

$$\begin{aligned} \frac{\hat{C}(\boldsymbol{\psi}^{(t-1)})}{\hat{C}(\boldsymbol{\psi})} &= \frac{1}{m} \sum_{h=1}^{m} \frac{q(\boldsymbol{x}^{(h)}|\boldsymbol{\psi}^{(t-1)})}{f(\boldsymbol{x}^{(h)}|\boldsymbol{\psi})} \Bigg/ \frac{1}{m'} \sum_{h=1}^{m'} \frac{q(\boldsymbol{x}^{(h)}|\boldsymbol{\psi})}{f(\boldsymbol{x}^{(h)}|\boldsymbol{\psi})} \\ &= \frac{1}{m} \sum_{h=1}^{m} \frac{q(\boldsymbol{x}^{(h)}|\boldsymbol{\psi}^{(t-1)})}{q(\boldsymbol{x}^{(h)}|\boldsymbol{\psi})/C(\boldsymbol{\psi})} \Bigg/ \frac{1}{m'} \sum_{h=1}^{m'} \frac{q(\boldsymbol{x}^{(h)}|\boldsymbol{\psi})}{q(\boldsymbol{x}^{(h)}|\boldsymbol{\psi})/C(\boldsymbol{\psi})} \\ &= \frac{1}{m} \sum_{h=1}^{m} \frac{q(\boldsymbol{x}^{(h)}|\boldsymbol{\psi}^{(t-1)})}{q(\boldsymbol{x}^{(h)}|\boldsymbol{\psi})} \,. \end{aligned} \tag{3.30}$$

MCMC. The MCMC is modified by using (3.30) in steps 2 and 3 instead of the likelihood ratio in equations (3.11) and (3.12). In each of these steps we update either β or $\boldsymbol{\phi}$, thus the vector $\boldsymbol{\psi}$ represents $(\beta, \boldsymbol{\phi}^{(t-1)\prime})'$ in Step 2, and $(\beta^{(t)}, \boldsymbol{\phi}')'$ in Step 3.

Chapter 4

Model selection

In the formulation of the process model in Chapter 3 (Section 3.1) we include a fix number q of covariates or predictors . In many cases, the original set includes a large number p ($p \gg q$) of covariates that can be classified into three groups: covariates known to be important in the field of application, covariates thought to be important, and covariates included as pure speculation (Kuo & Mallick, 1998). During the analysis it is desirable to delete some of the predictors from the model and use only the selected q covariates in the estimation procedure. Mitchell & Beauchamp (1988) refer to the predictors that are candidates for deletion as vulnerable predictors and discuss some reasons for undertaking the search for a subset of covariates: (a) to express the relationship between the response and the predictors as simply as possible; (b) to reduce future cost of prediction; (c) to identify important and negligible predictors; and (d) to increase the precision of statistical estimates and predictions.

The classical Bayesian methods for variable selection that include selection criteria (Bayesian information criterion, asymptotic information criterion, Bayes factor, pseudo-Bayes factor) have to consider an extremely high number of possible sub-models (2^p), which is an overwhelming task of calculation. George & McCulloch (1993) proposed the stochastic search variable selection (SSVS) method for identifying promising subsets of predictors, but avoiding the calculation of the posterior probabilities of all 2^p models. They derive the subset from a hierarchical normal mixture model with the specification of a mixture prior which uses the data to assign larger posterior probability to the more promising models. Gibbs sampler is used to search for promising models rather than compute the entire posterior. In order to obtain convergence, the algorithm requires tuning - specification of fixed prior parameters which are data dependent (O'Hara & Sillanpää, 2009).

Using the SSVS method as a motivating starting point, Kuo & Mallick (1998) proposed a simpler method that includes indicator variables embedded in the regression equation in such a way that all possible sub-models are considered. They assume that the indicator variables and the regression coefficients are independent a priori. This formulation is referred to as the expanded regression model (ERM) and is not limited only to the regular regression model, but is also extended to the generalized linear model. Dellaportas et al. (2002) suggested an alternative model formulation called Gibbs variable selection (GVS), which is an hybrid of the SSVS and the approach of Kuo & Mallick (1998).

A different approach is the adaptive shrinkage in which indicator variables are not used in the model, but instead a prior is specified directly on the regression parameters. The prior should work by shrinking values towards zero if there is no evidence in the data for non-zero values, and there should be practically no shrinkage for data-supported values of covariate that are non-zero. The method is adaptive in the sense that a degree of sparseness is defined by the data, through the way it shrinks the covariates effects towards zero (O'Hara & Sillanpää, 2009).

An alternative approach to placing priors on the individual covariate coefficients is to view the model as a whole and place priors on q (the number of covariates selected in the model and their corresponding coefficients). The choice of which covariates are in the model becomes a secondary problem (O'Hara & Sillanpää, 2009). One of the model selection techniques under this approach is the reversible jump MCMC introduced by Green (1995b). This flexible method lets the Markov chain explore spaces of different dimensions. Another technique is the composite model space introduced by Godsill (2001), in which the maximum q is fixed to something less than p, and indicator variables are used to allow covariates to enter or leave the model.

We use the ERM as a basis for variable selection with the autologistic model which we call expanded autologistic model (EAM). We include some concepts of the evolutionary Monte Carlo (EMC) algorithm (Liang & Wong, 2000); in particular, we use the mutation operator to increase the mixing behaviour of the chain. Another technique in the attempt to improve the mixing is the use of multiple independent chains (MIC) (Drugan & Thierens, 2010). At the end of the chapter we present a method for the assessment of model fitness developed by Gelman et al. (1996) in the Bayesian framework.

4.1 Expanded Autologistic Model

The variable selection procedure can be seen as a method to choose the regression parameters ϕ_k $(k = 1, \ldots, p)$ that are equal to zero. We use the concept of a slab and spike prior distribution for each ϕ_k (Mitchell & Beauchamp, 1988). The prior should therefore have a probability mass (the spike) either exactly at or around zero, and a flat slab elsewhere (the slab). We use an auxiliary indicator variable γ_k $(k = 1, \ldots, p)$ that indicates inclusion or exclusion of the covariate Z_k in the model. When the covariate is included in the model $(\gamma_k = 1)$, it is in the slab part of the prior, whereas when it is excluded $(\gamma_k = 0)$, the covariate corresponds to the spike part of the prior. In the ERM, Kuo & Mallick (1998) use a second auxiliary variable δ_k, which represents the effect size of each covariate, and is used as an intermediate value to obtain the coefficient ϕ_k, such that $\phi_k = \gamma_k \, \delta_k$ $(k = 1, \ldots, p)$.

In the expanded autologistic model the indicator variables γ_k and the auxiliary variable δ_k are easily embedded in the model. The conditional probability of the autologistic model in (3.5) is expressed in the expanded form as

$$f(x_i|\boldsymbol{x}_{(i)}, \beta, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{z}_i) = \frac{\exp\left[x_i \left(\beta \sum_{j\sim i} x_j + \boldsymbol{z}_i' \boldsymbol{\Gamma} \boldsymbol{\delta}\right)\right]}{1 + \exp\left(\beta \sum_{j\sim i} x_j + \boldsymbol{z}_i' \boldsymbol{\Gamma} \boldsymbol{\delta}\right)}, \tag{4.1}$$

where $\boldsymbol{\gamma} = (1, \gamma_1, \ldots, \gamma_p)'$, and $\boldsymbol{\Gamma} = \mathrm{diag}(\boldsymbol{\gamma})$. We assume the intercept term is always included in building the model, thus $\gamma_0 \equiv 1$. The vector of indicator variables $\boldsymbol{\gamma}$ dictates which predictors are included. When the indicator $\gamma_k = 1$, the k^{th} value in $\boldsymbol{z}_i$ is included in the calculation of the conditional probability in (4.1), whereas when $\gamma_k = 0$, that value is excluded. Consequently, the joint distribution becomes

$$f(\boldsymbol{x}|\beta, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{z}) = C^{-1}(\beta, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{z}) \exp\left[\sum_{i\in\mathcal{S}} x_i \left(\frac{\beta}{2} \sum_{j\sim i} x_j + \boldsymbol{z}_i' \boldsymbol{\Gamma} \boldsymbol{\delta}\right)\right]. \tag{4.2}$$

Although $\boldsymbol{\gamma}$ itself can be modelled as a realisation from any (nontrivial) prior $\pi(\boldsymbol{\gamma})$ on the 2^p possible values of $\boldsymbol{\gamma}$, priors with independent individual components γ_k are easy to specify, substantially reduce computational requirements, and often yield sensible results (George & McCulloch, 1997). These priors take the form

$$\pi(\boldsymbol{\gamma}) = \prod_{k=1}^{p} v_k^{\gamma_k} \, (1 - v_k)^{(1-\gamma_k)}.$$

By setting v_k small, the prior can also be used to put increased weight on parsimonious models; however, in absence of any prior preference for the predictors

we assume an indifference prior for each γ_k which are independent Bernoulli with probability $v_k = 1/2$ $(k = 1, \ldots, p)$; thus $\pi(\boldsymbol{\gamma}) = (1/2)^p$.

The joint posterior distribution of the process MRF and the parameter vector $\boldsymbol{\psi} = (\beta, \boldsymbol{\delta}', \boldsymbol{\gamma}', \boldsymbol{\vartheta}')'$ becomes

$$p(\boldsymbol{x}, \boldsymbol{\psi} \,|\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{w}) \propto f(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\vartheta}, \boldsymbol{w})\, f(\boldsymbol{x}|\beta, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{z})\, \pi(\beta)\, \pi(\boldsymbol{\delta})\, \pi(\boldsymbol{\gamma})\, \pi(\boldsymbol{\vartheta})\,, \tag{4.3}$$

where $f(\boldsymbol{y}|.)$ is the conditional probability given in (3.3) and $f(\boldsymbol{x}|.)$ is the joint distribution given in (4.2).

4.2 MCMC algorithm for model selection

We explain in this section how the algorithm from Section 3.2 is modified to draw samples from the posterior distribution of $\boldsymbol{X}$ and $\boldsymbol{\psi}$ given in (4.3). We include an extra step to generate $\boldsymbol{\gamma}$ after the update of $\boldsymbol{\delta}$. In the original method proposed by Kuo & Mallick (1998) (KM) to generate and accept the vectors $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$ separately. As a variation, Paroli & Spezia (2008) proposed the Metropolized-Kuo-Mallick (MKMK) method, in which the vector $(\boldsymbol{\delta}', \boldsymbol{\gamma}')'$ is accepted in a single Metropolis step. They demonstrated in the case of non-homogeneous hidden Markov models and Markov switching auto-regressive models that this method performs well when the explanatory covariates are strongly correlated.

In addition to the regular update of $\boldsymbol{\gamma}$ we include some ideas from the EMC algorithm proposed by Liang & Wong (2000). This method has incorporated many attractive features of simulated annealing and genetic algorithms into a framework of MCMC. EMC works by simulating a population of Markov chains in parallel, where a different temperature is attached to each chain. The vector of parameters is updated by mutation, crossover and exchange operators. They presented numerical results showing that EMC offers an improvement over the traditional MCMC algorithms. In the MCMC algorithm that we propose here, we only apply the mutation operator to update $\boldsymbol{\gamma}$ with probability p_m (mutation rate) while we apply the KM update for $\boldsymbol{\gamma}$ with probability $1 - p_m$.

The whole MCMC algorithm is executed in five sequential steps at each iteration. At the t^{th} iteration we draw $\boldsymbol{x}^{(t)}$, $\beta^{(t)}$, $\boldsymbol{\delta}^{(t)}$, $\boldsymbol{\gamma}^{(t)}$, and $\boldsymbol{\vartheta}^{(t)}$. We explain each of these steps to draw the t^{th} sample.

Step 1. The vector $\boldsymbol{x}$ of states is generated by Gibbs sampler in the same

way as in Section 3.2. The full conditional is

$$f(x_i^{(t)} \mid \boldsymbol{x}_{(i)}, \beta, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{z}_i, y_i, \boldsymbol{\theta}) \propto f(x_i^{(t)} \mid \boldsymbol{x}_{(i)}, \beta, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{z}_i) f(y_i | x_i^{(t)}, \boldsymbol{\theta}) \, ; \, (i \in \mathcal{S}) \, , \tag{4.4}$$

which involves the conditional probability of the EAM given in (4.1).

Step 2. We use Metropolis-Hastings to update the parameter β with the procedure described in Section 3.2. We accept β^* with probability $\min\{1, \rho_\beta\}$, and acceptance ratio

$$\rho_\beta = \frac{f(\boldsymbol{x}|\beta^*, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{z}) \pi(\beta^*)}{f(\boldsymbol{x}|\beta^{(t-1)}, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{z}) \pi(\beta^{(t-1)})} \frac{J_{\beta^{(t-1)}}}{J_{\beta^*}} \, . \tag{4.5}$$

Step 3. We use Metropolis-Hastings to update the parameter $\boldsymbol{\delta}$ with the procedure described in Section 3.2. We accept in block the vector $\boldsymbol{\delta}^*$ with probability $\min\{1, \rho_{\boldsymbol{\delta}}\}$, and acceptance ratio

$$\rho_{\boldsymbol{\delta}} = \frac{f(\boldsymbol{x}|\beta, \boldsymbol{\delta}^*, \boldsymbol{\gamma}, \boldsymbol{z}) \, \pi(\boldsymbol{\delta}^*)}{f(\boldsymbol{x}|\beta, \boldsymbol{\delta}^{(t-1)}, \boldsymbol{\gamma}, \boldsymbol{z}) \, \pi(\boldsymbol{\delta}^{(t-1)})} \frac{J_{\boldsymbol{\delta}}^{(t-1)}}{J_{\boldsymbol{\delta}^*}} \, . \tag{4.6}$$

Step 4. We update the auxiliary vector $\boldsymbol{\gamma}$ using one of the following alternatives: the mutation operator from EMC with probability p_m or the KM update with probability $1 - p_m$.

KM update. For the generation of the indicator γ_k $(k = 1, \ldots, p)$ we create the vectors $\boldsymbol{\gamma}_{k,0}$ and $\boldsymbol{\gamma}_{k,1}$ which include the already updated values for $\boldsymbol{\gamma}$ (before the k^{th}), and the previous values for the non-updated (after the k^{th}), while we force $\gamma_k = 0$ in $\boldsymbol{\gamma}_{k,0}$ and $\gamma_k = 1$ in $\boldsymbol{\gamma}_{k,1}$. The resulting vectors are

$$\begin{aligned} \boldsymbol{\gamma}_{k,1} &= (1, \gamma_1^{(t)}, \ldots, \gamma_{k-1}^{(t)}, 1, \gamma_{k+1}^{(t-1)}, \ldots, \gamma_q^{(t-1)})' \, , \\ \boldsymbol{\gamma}_{k,0} &= (1, \gamma_1^{(t)}, \ldots, \gamma_{k-1}^{(t)}, 0, \gamma_{k+1}^{(t-1)}, \ldots, \gamma_q^{(t-1)})' \, . \end{aligned} \tag{4.7}$$

The posterior probability of $\gamma_k = 0$ involves the density of $\boldsymbol{x}$ calculated with $\boldsymbol{\gamma}_{k,0}$, while the posterior of $\gamma_k = 1$ involves that with $\boldsymbol{\gamma}_{k,1}$, as follows

$$\begin{aligned} p_{k,0} &= p(\gamma_k = 0) = f(\boldsymbol{x}|\beta, \boldsymbol{\delta}, \boldsymbol{\gamma}_{k,0}, \boldsymbol{z}) \, (1 - \pi(\gamma_k)) \, , \\ p_{k,1} &= p(\gamma_k = 1) = f(\boldsymbol{x}|\beta, \boldsymbol{\delta}, \boldsymbol{\gamma}_{k,1}, \boldsymbol{z}) \, \pi(\gamma_k) \, , \end{aligned}$$

where we omit all the conditioning on $\boldsymbol{x}$, β and $\boldsymbol{z}$. Each $\gamma_k^{(t)}$ $(k = 1, \ldots, p)$ is generated (in sequential order) from a Bernoulli distribution

$$q_k(\gamma_k^{(t)} \mid \boldsymbol{\delta}, \boldsymbol{\gamma}_{k,0}, \boldsymbol{\gamma}_{k,1}) = \left(\frac{p_{k,1}}{p_{k,1} + p_{k,0}} \right)^{\gamma_k^{(t)}} \left(\frac{p_{k,0}}{p_{k,1} + p_{k,0}} \right)^{1 - \gamma_k^{(t)}} \, . \tag{4.8}$$

We use Gibbs sampler to sequentially draw the value of $\gamma_k^{(t)}$ with the probability

$$q_k(\gamma_k^{(t)} = 1 \,|\, \boldsymbol{\delta}, \boldsymbol{\gamma}_{k,0}, \boldsymbol{\gamma}_{k,1}) = \frac{p_{k,1}}{p_{k,1} + p_{k,0}} \,. \tag{4.9}$$

Mutation. We randomly select an index between 1 and p, say k. We flip the value of $\gamma_k^{(t-1)}$ such that $\gamma_k^* = 1 - \gamma_k^{(t-1)}$, and keep the other values without change, i.e. $\gamma_j^* = \gamma_j^{(t)}\ \forall j \neq k$. This is a 1-point mutation, and we can also use 2-point mutations where two indexes are randomly selected and the corresponding values are reversed. One extreme case is the uniform mutation in which each element of $\boldsymbol{\gamma}^{(t-1)}$ has a nonzero probability of mutating. We denote the transition probability between $\boldsymbol{\gamma}$ vectors with $q(\cdot\,|\,\cdot)$. All these mutation operators are symmetric. We accept in block the vector $\boldsymbol{\gamma}^*$ with probability $\min\{1, \rho_{\boldsymbol{\gamma}}\}$, and acceptance ratio

$$\begin{aligned} \rho_{\boldsymbol{\gamma}} &= \frac{p(\boldsymbol{\gamma}^*|\boldsymbol{x}, \beta, \boldsymbol{\delta}, \boldsymbol{z})}{p(\boldsymbol{\gamma}^{(t-1)}|\boldsymbol{x}, \beta, \boldsymbol{\delta}, \boldsymbol{z})} \frac{q(\boldsymbol{\gamma}^{(t-1)}|\boldsymbol{\gamma}^*)}{q(\boldsymbol{\gamma}^*|\boldsymbol{\gamma}^{(t-1)})} \\ &= \frac{f(\boldsymbol{x}^*|\beta, \boldsymbol{\delta}, \boldsymbol{\gamma}^*, \boldsymbol{z})\ \pi(\boldsymbol{\gamma}^*)}{f(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}, \beta, \boldsymbol{\delta}, \boldsymbol{\gamma}^{(t-1)}, \boldsymbol{z})\ \pi(\boldsymbol{\gamma}^{(t-1)})} \,, \end{aligned} \tag{4.10}$$

since $q(\boldsymbol{\gamma}^{(t-1)}|\boldsymbol{\gamma}^*)/q(\boldsymbol{\gamma}^*|\boldsymbol{\gamma}^{(t-1)}) = 1$.

Step 5. We update $\boldsymbol{\vartheta}$ in the same way as in Section 3.2.

Model selection. In an attempt to improve the mixing behaviour of the MCMC, one could make use of multiple chains that run independently (MIC). The chains are started at different initial states and their output is observed simultaneously at each iteration. It is hoped that this way a more reliable sampling of the target distribution is obtained (Drugan & Thierens, 2010). It is important to note that no information exchange between the chains takes place. A vector $\boldsymbol{\gamma} = (1, \gamma_1, \ldots, \gamma_p)'$ is obtained for each chain at each iteration of the previous MCMC for variable selection. The marginal posterior distribution of $\boldsymbol{\gamma}$ is tabulated from the frequencies of all these vectors (combining all the chains) and the vector $\boldsymbol{\gamma}^*$ with the highest posterior probability is selected.

4.3 Posterior predictive assessment of model fitness

If the proposed model fails to provide a reasonable summary of the data at hand, it should be excluded; thus, any meaningful Bayesian analysis should at least include a check of the conformance of the model with the data. Gelman

et al. (1996) extend the essence of the classical approach of a goodness-of-fit test to the Bayesian framework, with the aim of providing pragmatic methods of assessing the fitness of a single model. They use the posterior predictive distribution for a discrepancy between data and the proposed model. This method is simple, both conceptually and computationally, and connects well to the classical goodness-of-fit methods that most researchers are familiar with. It is also very general, applicable for comparing observations with model predictions in any form. Crespi & Boscardin (2009) applied this method for evaluating the fit of Bayesian models for multivariate data, while Yano et al. (2001) used it to evaluate pharmacokinetic/pharmacodynamic models.

The focus of the method is to measure discrepancies between a model and the data, not to test whether a model is true. The data come from a disturbed map which, in the presence of high levels of observation error, tend to be very different from the the true values. Thus, when we estimate the hidden map we may obtain many values that differ from the data. This is not contradictory. We define $\boldsymbol{y}^r$ as the replicated data it would appear as if the experiment that produced $\boldsymbol{y}$ today were replicated tomorrow with the same model and the same values of the parameters in $\boldsymbol{\psi}$ that produced $\boldsymbol{y}$. We expect $\boldsymbol{y}^r$ to be close to $\boldsymbol{y}$. We refer to the data model as $\mathcal{H}$ and the distribution of the replicated data as $p(\boldsymbol{y}^r|\mathcal{H}, \boldsymbol{\psi})$. In the Bayesian framework, the inference for $\boldsymbol{\psi}$ is provided by its posterior distribution, $p(\boldsymbol{\psi}|\mathcal{H}, \boldsymbol{y})$, where the model $\mathcal{H}$ includes the prior distribution $\pi(\boldsymbol{\psi})$. Correspondingly, the reference distribution of the future observation $\boldsymbol{y}^r$ is its posterior predictive distribution,

$$p(\boldsymbol{y}^r|\mathcal{H}, \boldsymbol{y}) = \int p(\boldsymbol{y}^r|\mathcal{H}, \boldsymbol{\psi}) p(\boldsymbol{\psi}|\mathcal{H}, \boldsymbol{y}) d\boldsymbol{\psi} \,.$$

We select a discrepancy measure $D(\boldsymbol{y}; \boldsymbol{\psi})$ with a reference distribution derived from the joint posterior distribution of $\boldsymbol{y}^r$ and $\boldsymbol{\psi}$,

$$p(\boldsymbol{y}^r, \boldsymbol{\psi}|\mathcal{H}, \boldsymbol{y}) = p(\boldsymbol{y}^r|\mathcal{H}, \boldsymbol{\psi}) p(\boldsymbol{\psi}|\mathcal{H}, \boldsymbol{y}) \,. \tag{4.11}$$

We measure the location of the realised value $D(\boldsymbol{y}; \boldsymbol{\psi})$ within its reference distribution with the classical tail-area approach. We can formally define a tail-area probability of D under its posterior reference distribution as

$$p_D(\boldsymbol{y}) = \Pr(D(\boldsymbol{y}^r; \boldsymbol{\psi}) \geq D(\boldsymbol{y}; \boldsymbol{\psi})|\mathcal{H}, \boldsymbol{y}) \,. \tag{4.12}$$

Computation of the reference distribution of discrepancies and the corresponding tail-area probability is easily accomplished via Monte Carlo simulation. We compare the realised discrepancy $D(\boldsymbol{y}; \boldsymbol{\psi})$ to its reference distribution

under (4.11) by drawing a set of $\boldsymbol{\psi}^{(j)}$ $(j = 1, \ldots, J)$ and performing the following two steps for each j:

Step 1. Given $\boldsymbol{\psi}^{(j)}$, draw a simulated replicated data set $\boldsymbol{y}^{r(j)}$ from the sampling distribution $p(\boldsymbol{y}^{r(j)}|\mathcal{H}, \boldsymbol{\psi}^{(j)})$. We define this distribution based on (3.2) as

$$p(\boldsymbol{y}^{r(j)}|\boldsymbol{x}, \boldsymbol{\theta}^{(j)}) = \prod_{i \in \mathcal{S}} (\theta^{(j)}_{x_i,i})^{y_i^{r(j)}} (1 - \theta^{(j)}_{x_i,i})^{1 - y_i^{r(j)}} . \tag{4.13}$$

Step 2. Calculate discrepancies for each replicated data set, $D(\boldsymbol{y}^{r(j)}; \boldsymbol{\psi}^{(j)})$, and for the observed data, $D(\boldsymbol{y}; \boldsymbol{\psi}^{(j)})$.

The proportion of pairs for which $D(\boldsymbol{y}^{r(j)}; \boldsymbol{\psi}^{(j)})$ exceeds $D(\boldsymbol{y}; \boldsymbol{\psi}^{(j)})$ is an estimate $p_D(\boldsymbol{y})$ in (4.12). Once the replicates have been drawn in Step 1, the same draws can be used for as many realised discrepancy measures as one wishes. We use as a discrepancy measure the sum of squares of residuals of the data with respect to their expectations under a posited model. We use the Pearson residuals r_i which are the standardised differences between the observed and the expected values, $r_i = (y_i - \mathrm{E}(y_i|\boldsymbol{\theta}))/\sqrt{\mathrm{Var}(y_i|\boldsymbol{\theta})}$. The discrepancy measure is

$$D(\boldsymbol{y}; \boldsymbol{\theta}) = \sum_{i \in \mathcal{S}} \frac{(y_i - \mathrm{E}(y_i|\boldsymbol{\theta}))^2}{\mathrm{Var}(y_i|\boldsymbol{\theta})} = \sum_{i \in \mathcal{S}} \frac{(y_i - \theta_{x_i,i})^2}{\theta_{x_i,i}(1 - \theta_{x_i,i})} . \tag{4.14}$$

Chapter 5

Simulation studies

This chapter is devoted to several simulation studies with the aim of testing the methodology exposed in Chapter 3 (Hierarchical modelling) and Chapter 4 (Model selection). The first evaluation corresponds to the approximation methods of the likelihood of the hidden map presented in Section 3.3: pseudo-likelihood (PL) approximation, path sampling (PS) and ratio approximation (RA). We assess the performance of these methods in three stages. The first part is based on a small vector of values for which we can calculate the exact likelihood function. It is used to evaluate the performance of the PS approximation. For the second and third parts, we generate maps that present the same spatial configuration as our data set from Section 2.2 and include a covariate taken from that data set. We generate three maps that correspond to a population of maps not so different from a real situation. These maps are generated with three different levels of spatial autocorrelation. In the second part, we compare the posterior distributions of each parameter, obtained using the different methods of approximation. In the third part, we evaluate the complete estimation MCMC using observed maps obtained by disturbing the corresponding true maps.

We also perform an evaluation of the Kuo & Mallick (1998) method (KM) for variable selection. We use three maps generated with one covariate and three maps generated with four covariates taken from the data set. In each case, the three maps correspond to three different levels of spatial autocorrelation. We observe the 95% credible intervals for each parameter and the misclassification rates obtained with the different methods of approximation of the likelihood function.

5.1 Performance of the approximations

In this section, we first evaluate the PS approximation using exact values of the likelihood function calculated on a small vector. Then we use the PS as a good reference for the true likelihood function when we evaluate the marginal posterior distributions of each parameter, obtained using the different methods of approximation. Finally, we make an evaluation of the approximations in more complex situations where the whole MCMC is run to estimate all parameters of the model.

5.1.1 Exact calculation

We start with a simple evaluation of the performance of the PS approximation using a small vector of 16 values organised in a 4×4 rectangle. We include a covariate with the values in Table 5.1.

-1	-1	0.5	0.5
-1	-1	0.5	0.5
-0.5	-0.5	1	1
-0.5	-0.5	1	1

Table 5.1: *Image that corresponds to the vector* $\boldsymbol{x} = (1,1,1,1,1,0,1,1,\ 1,1,0,\ 0,0,0,0,0)'$. *Blue colour indicates presence and brown colour absence. Values of the covariate are inside the cells.*

For this evaluation we use 125 vectors $\boldsymbol{\psi} = (\beta, \phi_0, \phi_1)'$ on the grid defined by $\{0, 0.4, 0.8, 1.2, 1.6\} \times \{-4, -3, -2, -1, 0\} \times \{0.5, 1, 1.5, 2, 2.5\}$. We calculate the exact value of the normalising constant using (3.6) for each $\boldsymbol{\psi}$, and then apply the PS approximation to the same vectors using (3.21), (3.24), and (3.26). In Fig. 5.1 we plot the ratio of the path approximation and the exact $C(\boldsymbol{\psi})$. In general, the approximations are close enough to the exact values (within a 5% distance). The bigger the value of β, the bigger the distance between the approximation and the exact value. Also these distances decrease as ϕ_0 increases, but we cannot conclude that they are related to big or small values of ϕ_1.

We construct a curve of the exact $\log C(\boldsymbol{\psi})$ for different values of ϕ_1, while β and ϕ_0 are fixed ($\beta = 1.6$, $\phi_0 = -4$). We choose these values because the produce the highest differences between the approximation and the exact nor-

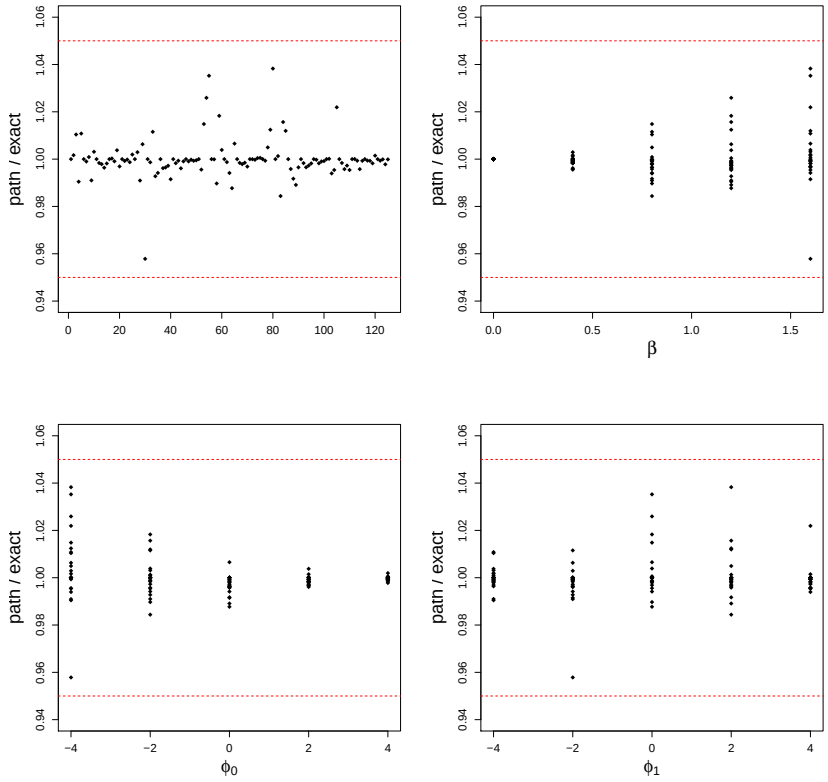

Figure 5.1: *Ratio of PS approximation and exact $C(\boldsymbol{\psi}_1)$ for 125 $\boldsymbol{\psi}$ vectors. Dashed lines represent the limits where the approximation is 5% higher or lower than the exact normalising constant.*

malising constant. In Fig. 5.2 we plot this curve and the PS approximations of $\log C(\boldsymbol{\psi})$; we observe that the approximations are around the exact curve.

Now we analyse the performance of the interpolations from the regression models. We take $\boldsymbol{x} = (1,1,1,1,1,0,1,1,1,1,0,0,0,\ 0,0,0)'$, which is represented in the map in Table 5.1, and fix $\boldsymbol{\psi}^* = (1,-1,-1.5)'$. We choose this particular $\boldsymbol{\psi}^*$ in the example because it is close to the maximum pseudo-likelihood estimator, and the estimates around this maximum are of special interest in the MCMC algorithm. We randomly select 20 vectors in the interior of the hypercube

$$\{\boldsymbol{\psi} : \psi_k = \psi_k^* \pm d_k,\ k = 1,\ldots,3\}\,,$$

where d_k is chosen as in the definition of Ω (multidimensional grid of points in Step 2, Subsection 3.3.2), i.e. $d_1 = 0.75$ and $d_k = 2.5$ $(k = 2,3)$. For each vector we calculate the exact log-likelihood, the PS approximation, and the

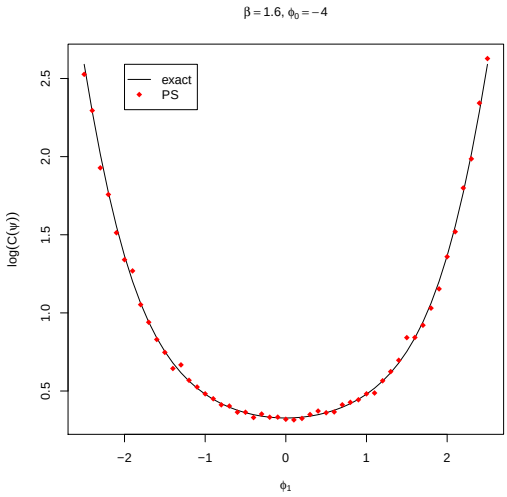

Figure 5.2: *Exact and PS approximated values of* $\log C(\boldsymbol{\psi})$ *against* ϕ_1*, keeping* β *and* ϕ_0 *fixed.*

interpolations (BS and PR) based on the estimations of the normalising constant obtained from the PS algorithm with Ω centered on $\boldsymbol{\psi}^*$. In Fig. 5.3 we plot the ratio of each approximation to the exact likelihood and observe that PS approximations are relatively accurate, out of the 20 approximations, 19 are within a 3:2 ratio to the exact values, and one is within a 2:1 ratio to the exact values. On the other hand, the interpolations based on the regression models (BS and PR) show in general an inappropriate accuracy level, most of them are outside of the limits of the 2:1 ratio to the exact values.

5.1.2 Posterior distributions

In this evaluation of the performance of the approximations, we generate maps on a grid with $N = 2195$ values arranged in the same way as the study data set presented in Section 2.2, with a single covariate taken from that data set. We use the Temperature as an example of one of the covariates that is expected to influence the presence of the species that we are studying. We generate three maps with values of β that produce different levels of spatial interaction ($\beta = 0, 1, 1.5$), and use the same regression coefficients for the Temperature in the three maps ($\phi_0 = -1.5, \phi_1 = 2$) (see Fig B.1 in Appendix B for the generated maps and the associated covariate). These coefficients are selected because they produce reasonable maps with proportions of sites with presence between 0.30 and 0.70. We avoid maps with too low or too high proportions of presence. We

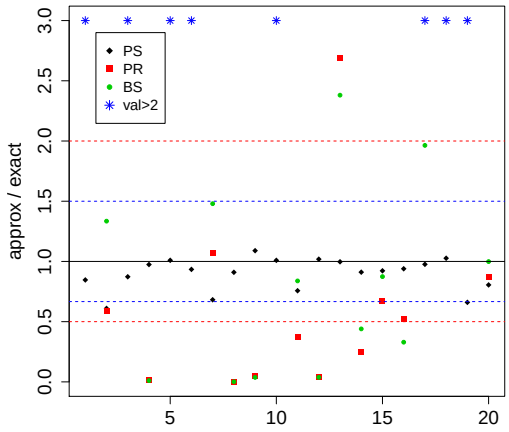

Figure 5.3: *Ratio of PS approximation and interpolations from PS (BS and PR) to exact likelihood for 20 vectors around* $\boldsymbol{\psi}^* = (1, -1, -1.5)'$ *with a fix vector* $\boldsymbol{x}$*. Dotted lines indicate limits for approximations in the ratio 2:1 or 1:2 (red), and 3:2 or 2:3 (blue) respect to the exact value. Values that are greater than three are indicated with a symbol *.*

notice how the influence of the covariate on the map increases as the spatial dependence increases.

In all cases when we estimate the posterior distribution of β we assume that it is uniformly distributed in the interval $[0, 3.5]$. For the coefficients ϕ_k $(k = 0, \ldots, q)$, we choose independent normal priors with mean 0 and high variance σ^2, truncated in the range $[-A, A]$. The tuning parameters $\sigma^2_{\epsilon_\beta}$ and $\sigma^2_{\epsilon_\phi}$ of Step 2 and Step 3 of the MCMC, respectively, are chosen to reach an acceptance rate between 0.20 and 0.50.

Path sampling and pseudo-likelihood approximations

For each parameter, we approximate the posterior distribution on a vector of m values on the interval defined for the prior distribution of that parameter (ψ^*_k), while the other two parameters are fixed at their corresponding true values $(\boldsymbol{\psi}^*_{-k})$. We create the vectors $\boldsymbol{\psi}_k = (\psi_{1,k}, \ldots, \psi_{m,k})'$ $(k = 1, 2, 3)$ and calculate the approximation to the likelihood function $f_*(\boldsymbol{x}|\psi_{j,k}, \boldsymbol{\psi}^*_{-k}, \boldsymbol{z})$ for each point $\psi_{j,k}$ $(j = 1, \ldots, m;\ k = 1, 2, 3)$ using the approximation with the PL. We also approximate the likelihood by using the PS approximation calculated directly on each point. For each calculation we run the algorithm for a total of 500

iterations, from which the first 200 draws are discarded in the burn-in and the remaining 300 draws are used to estimate the posterior distribution. In Appendix A we show that, in general for this algorithm, the chain stabilises at around 200 iterations and it converges at around 200 to 400 iterations. Next we obtain the interpolations (BS and PR) based on the estimations of the normalising constant obtained from the PS algorithm with a multidimensional grid of points Ω centered on $\boldsymbol{\psi}^*$. We calculate the posterior function for each $\psi_{j,k}$ using the following approximation:

$$\begin{aligned} p(\psi_{j,k}|\boldsymbol{x}, \boldsymbol{\psi}^*_{-k}, \boldsymbol{z}) &\approx \frac{f_*(\boldsymbol{x}|\psi_{j,k}, \boldsymbol{\psi}^*_{-k}, \boldsymbol{z})\, \pi(\psi_{j,k})}{\int f_*(\boldsymbol{x}|t, \boldsymbol{\psi}^*_{-k}, \boldsymbol{z})\, \pi(\psi_t) dt} \\ &\approx \frac{f_*(\boldsymbol{x}|\psi_{j,k}, \boldsymbol{\psi}^*_{-k}, \boldsymbol{z})\, \pi(\psi_{j,k})}{\sum_{i=1}^{m-1} \mu(\psi_{i,k})\, \Delta(\psi_{i,k})}, \end{aligned}$$

where $\mu(\psi_{i,k}) = \frac{1}{2}\{f_*(\boldsymbol{x}|\psi_{i+1,k}, \boldsymbol{\psi}^*_{-k}, \boldsymbol{z})\, \pi(\psi_{i+1,k}) + f_*(\boldsymbol{x}|\psi_{i,k}, \boldsymbol{\psi}^*_{-k}, \boldsymbol{z})\, \pi(\psi_{i,k})\}$, and $\Delta(\psi_{i,k}) = \psi_{i+1,k} - \psi_{i,k}$.

We take the PS approximation calculated directly on each point as a reference since it was the most accurate approximation in the previous simulations (Fig. 5.2 and Fig. 5.3). In Fig. 5.4 we plot the posterior distribution of β for the three maps. When $\beta = 0$ the PL is no longer an approximation but is equal to the true distribution (from 3.5, 3.14 and 3.21); this is confirmed with the closeness of the PL and the PS. We observe the poor performance of the interpolations based on PR when β is high. In Fig. 5.5 we plot the posterior distributions of ϕ_0 and ϕ_1. The interpolations based on PR are the worst for ϕ_1, and also the interpolations based on BS are poor when $\beta > 0$, the bigger the value of β the worse the performance. In all cases, the PL approximation is close to the reference PS.

We evaluate the adjustment to the PL directly with the same maps (generated with one covariate). We obtain the approximation to the posterior distribution as it was explained before. For the calculation of the window subsampling empirical variance (WSEV) estimate we take windows of size 10×10 and move $s = 3$ cells to the right and down. Due to the irregular border of the maps, in some cases the selected rectangle has a low number of cells with information, then we consider in the calculation only those rectangles with information in at least 50% of the cells. We plot the posterior distribution of β, ϕ_0 and ϕ_1 for the three maps (see Fig. B.6 and Fig. B.7 in Appendix B). In all cases the adjustment does not produce any important difference compared to the PL.

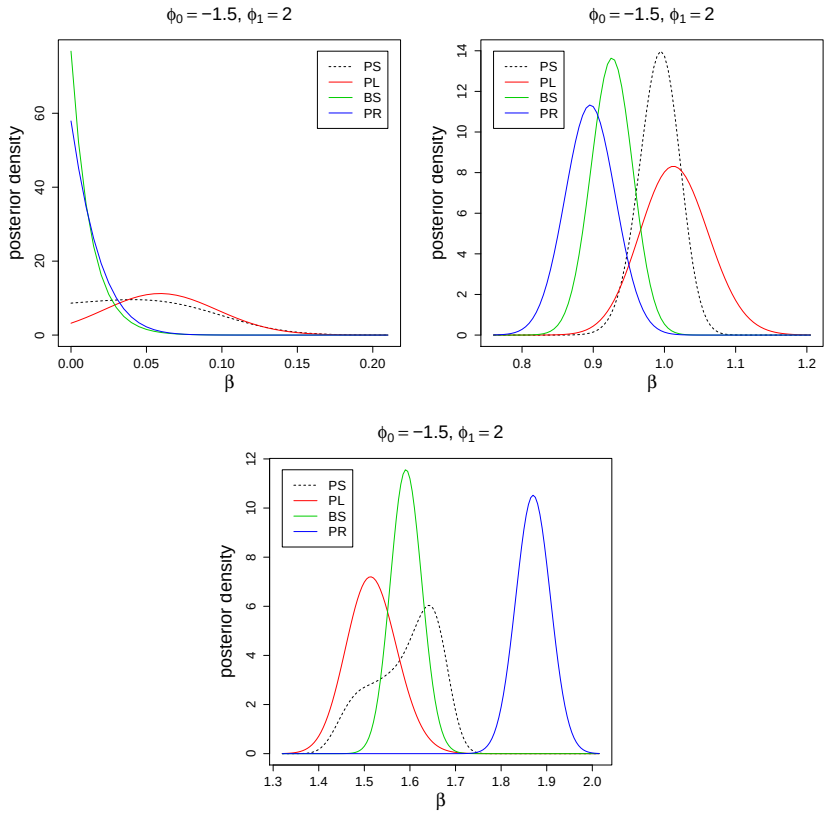

Figure 5.4: *Posterior distribution of β using approximations for the likelihood (PS, PL, BS, PR), and keeping ϕ fixed at the true value. Vertical lines indicate true values of β.*

Ratio approximation

Since the RA is used directly in the MCMC algorithm, we evaluate its performance with draws from the marginal posterior distribution of each parameter obtained from the estimation MCMC (Section 3.2). We only run one step of the MCMC to update the parameter for which we want to obtain the posterior, and keep the other parameters fixed at their true values.

The MCMC is run for a total of 10,000 iterations, from which the first 5,000 draws are discarded in the burn-in and the remaining 5,000 draws are used to estimate the posterior distribution. This number of iterations is enough to ensure convergence and good mixing of the chains, as well as no problems of strong autocorrelation (in Fig. B.4 and Fig. B.5, Appendix B, we show the trace, ergodic mean and ACF of the samples for one of the maps using both methods of approximation).

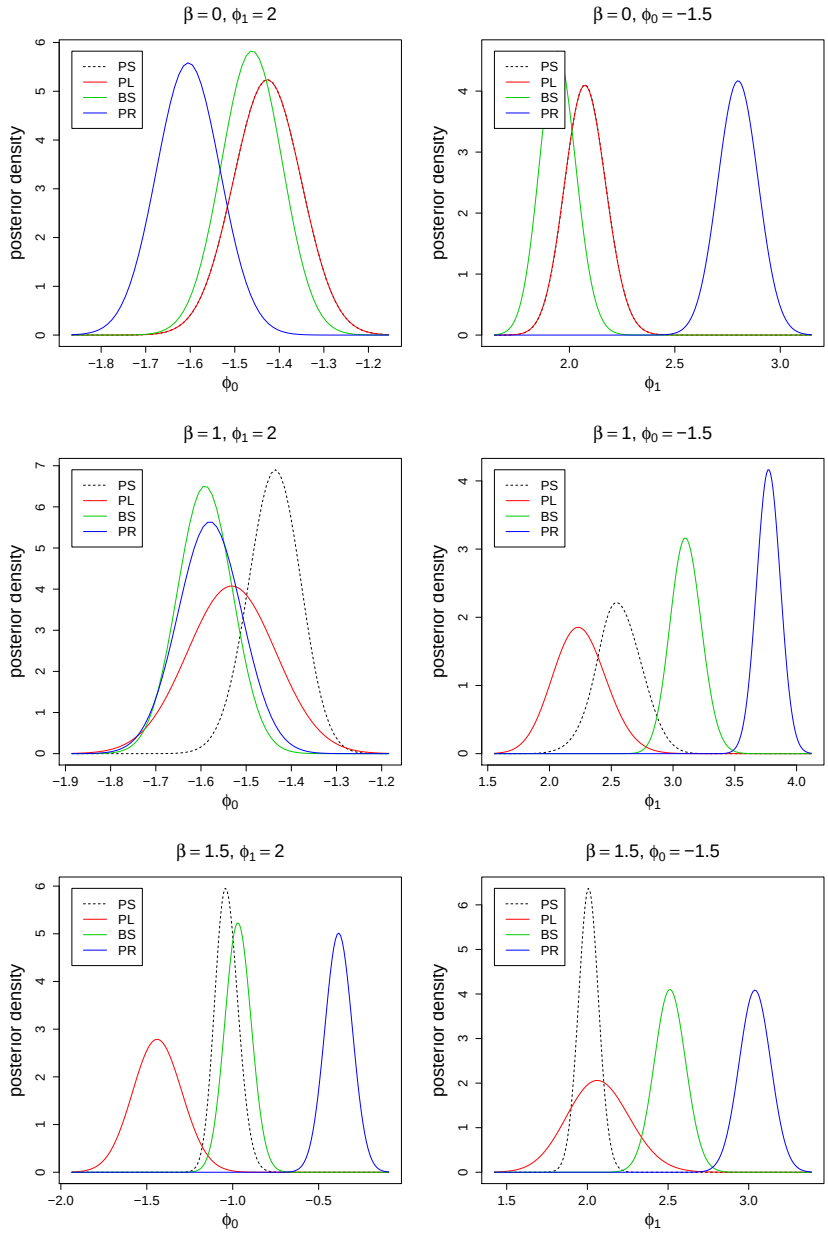

Figure 5.5: *Posterior distribution of ϕ_0 (left) and ϕ_1 (right) using approximations for the likelihood (PS, PL, BS, PR), and keeping the other parameters fixed at their true values. Vertical lines indicate true values of ϕ_0 or ϕ_1.*

For each calculation of the ratio in (3.30) we generate 200 vectors $\boldsymbol{x}^{(h)}$, from which the first 100 draws are discarded in the burn-in; we thin the chain by taking a draw every two vectors to reduce autocorrelation, resulting on 50 draws

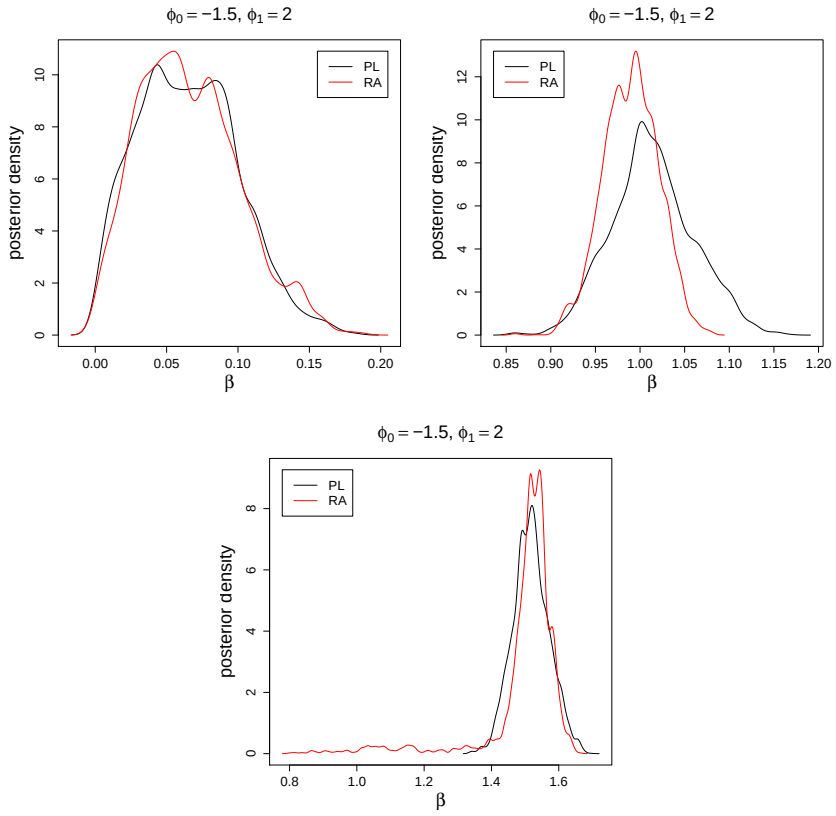

Figure 5.6: *Posterior distribution of β using the RA in the MCMC, and keeping $\boldsymbol{\phi}$ fixed at the true value.*

for the calculation of the ratio.

In Fig. 5.6 and Fig. 5.7 we plot the posterior distribution of β, ϕ_0 and ϕ_1 obtained from the estimation MCMC for the three maps. The distributions using RA are very close to the PL. In general, the variability of the posteriors obtained with RA are similar to or slightly smaller than the variability of those obtained with the PL.

5.1.3 MCMC

In this subsection we make an evaluation of the procedure in more complex situations. We apply the MCMC with observed maps obtained from the true maps by disturbing them with $\alpha = (-2.5, -0.5, 1.2, 2)'$, which produces values of $\theta_{0,i}$ and $\theta_{1,i}$ with means $\bar{\theta}_0 = 0.08$ and $\bar{\theta}_1 = 0.68$, respectively. We first perform this evaluation for the three maps generated with Temperature, and second with three more maps generated using more than one covariate. In the

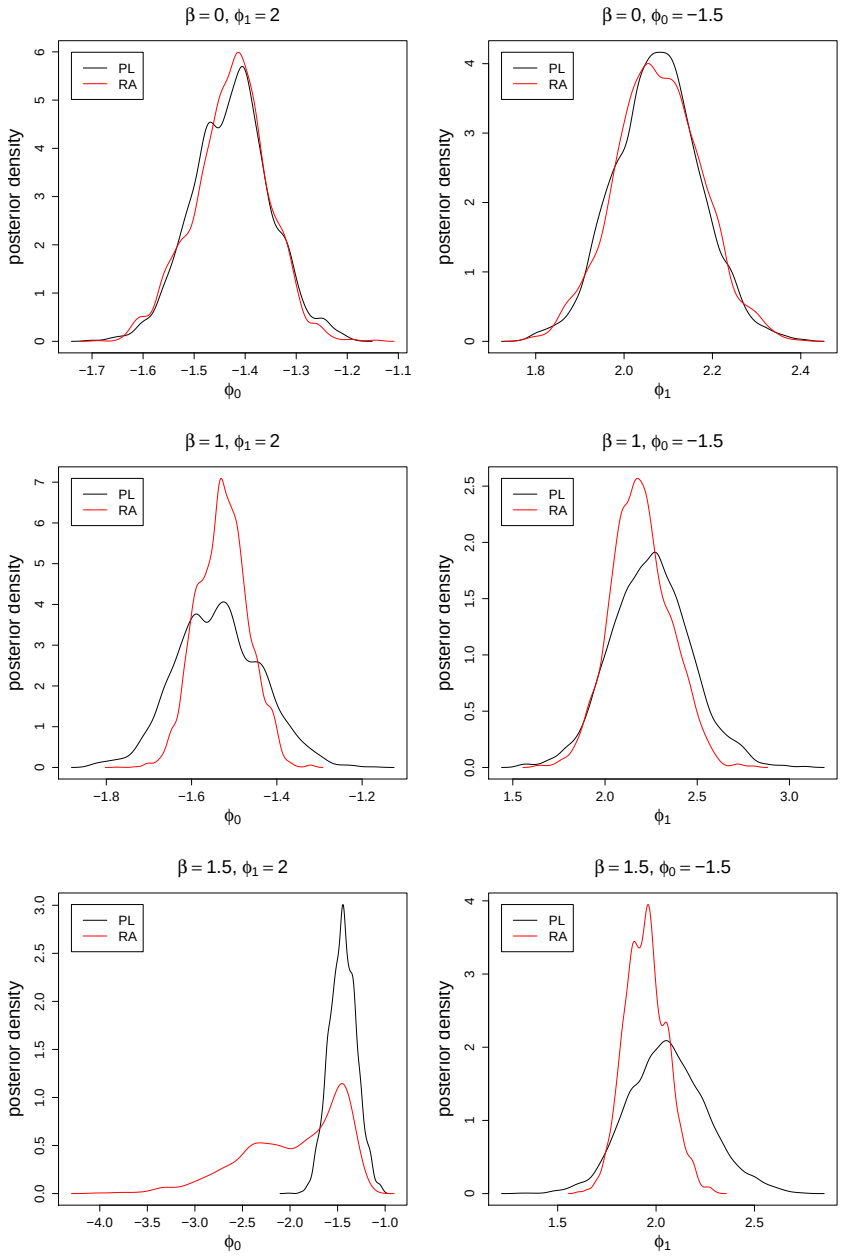

Figure 5.7: *Posterior distribution of ϕ_0 (left) and ϕ_1 (right) using the RA in the MCMC, and keeping the other parameters fixed at their true values.*

second case, we generate three new maps with the same values of β as before, which produces different levels of spatial interaction ($\beta = 0, 1, 1.5$). We use the following four covariates: Temperature, Open water, Distance from standing fresh water, and Grassland/herbaceous/pasture/crops (see Fig. C.1, Fig. C.2

and Fig. C.3 in Appendix C). We use the same coefficients $\boldsymbol{\phi} = (-2, -1.5, 2, 2, 1)'$ to generate the three maps. These coefficients are selected in such a way that we do not get maps with too low or too high proportions of presence. We also obtain the corresponding observed maps by disturbing the true maps with $\alpha = (-2.5, -0.5, 1.2, 2)'$ (see Fig. B.3 in Appendix B for the generated true and observed maps).

The MCMCs are run for a total of 210,000 iterations, from which the first 10,000 draws are discarded in the burn-in; we thin the chain by taking a draw every 20 iterations to reduce autocorrelation. The remaining 10,000 draws are used to estimate the parameters with convergence and good mixing of the chains (in Fig. B.8, Appendix B, we show the trace, ergodic mean and ACF of the samples for one of the maps using PL approximation). In addition, for the calculation of the ratio in the RA method we use 100 draws obtained from the generation of 400 vectors $\boldsymbol{x}^{(h)}$ (the first 200 draws are discarded in the burn-in and we take a draw every two vectors). The CPU time cost per 1000 runs on a 2.2GHz computer is: 2s for PL, 19s for BS, 17s for PR, and 173s for RA.

In Fig. 5.8 we have 95% credible intervals for β, ϕ_0 and ϕ_1 for the three maps generated with one covariate. The results with RA and PL are very similar to each other. They are concentrated around the true values of β and not far from the true values of ϕ_0 and ϕ_1. The variability of the estimates with these two methods increases with the size of β, i.e. a map generated with a higher β corresponds to estimates with more variability. This is more noticeable on the PL. We confirm what we already observed in the marginal posterior distributions of Subsection 5.1.2, that the distributions for BS and PR tend to be away from the true values of the parameters, in particular when the map is generated with $\beta > 0$.

We now extend the analysis to the case of maps generated with more than one covariate. Since the methods based on PS do not perform well in the case of one variable, we make the comparisons only for the other two methods (PL and RA). Credible intervals for the parameters are narrower in this case with four covariates (Fig. 5.9) than in the previous one with only one covariate (Fig. 5.8). We basically get a confirmation of what have been noticing, that the RA and PL produce similar results for all the estimates. All the parameters have distributions near the true values.

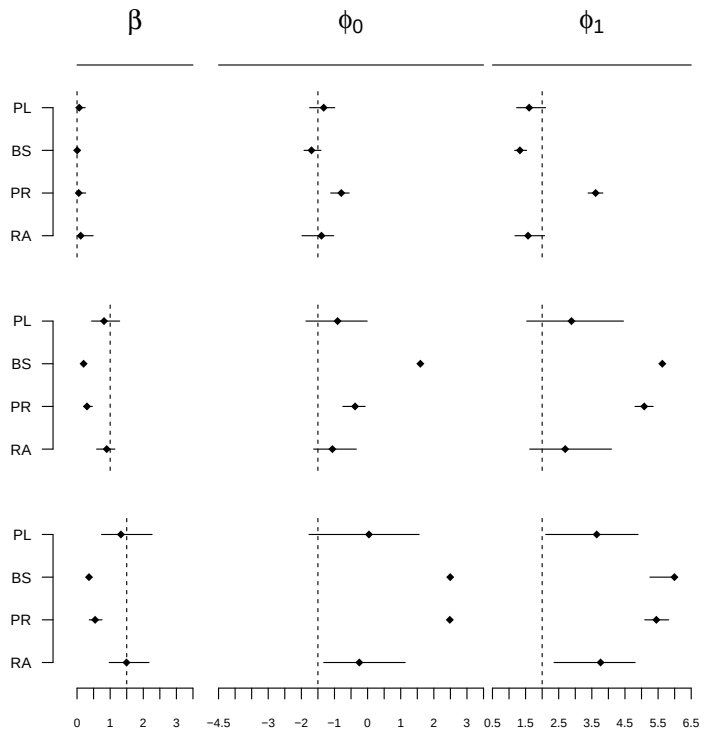

Figure 5.8: *95% credible intervals for β (left), ϕ_0 (center), and ϕ_1 (right) from the complete estimation MCMC, using different methods of approximation (PL, BS, PR and RA). Vertical lines indicate true values of the parameter.*

In a simulation study we can compare the reproduced map $\hat{\boldsymbol{x}}$ with the true map $\boldsymbol{x}$. We calculate the misclassification rate $\delta_{\boldsymbol{x}}(\boldsymbol{x}^*)$ as the proportion of sites that are different between a particular map $\boldsymbol{x}^*$ and the true map:

$$\delta_{\boldsymbol{x}}(\boldsymbol{x}^*) = 1 - \frac{1}{N}\sum_{i\in\mathcal{S}} I_{\{x_i\}}(x_i^*)\,.$$

Furthermore, we take as a baseline $\delta_{\boldsymbol{x}}(\boldsymbol{y})$, the proportion of misclassified sites between the observed map $\boldsymbol{y}$ and the true map.

In Table 5.2 we present the misclassification rates for the three maps generated with one covariate and the three maps generated with four covariates. In the case of one covariate we observe that for the two methods based on interpolations from the PS (BS and RA), the reproduced map presents a higher dissimilarity than the baseline. We observe the opposite situation with the PL and RA: as β increases, the misclassification rates obtained with the reproduced map are always lower than the baseline. The misclassification rates are very similar with both methods and also they are similar between maps generated

with one and four covariates; thus, the size of the interaction parameter (β) is the only factor that has an influence on the misclassification rate.

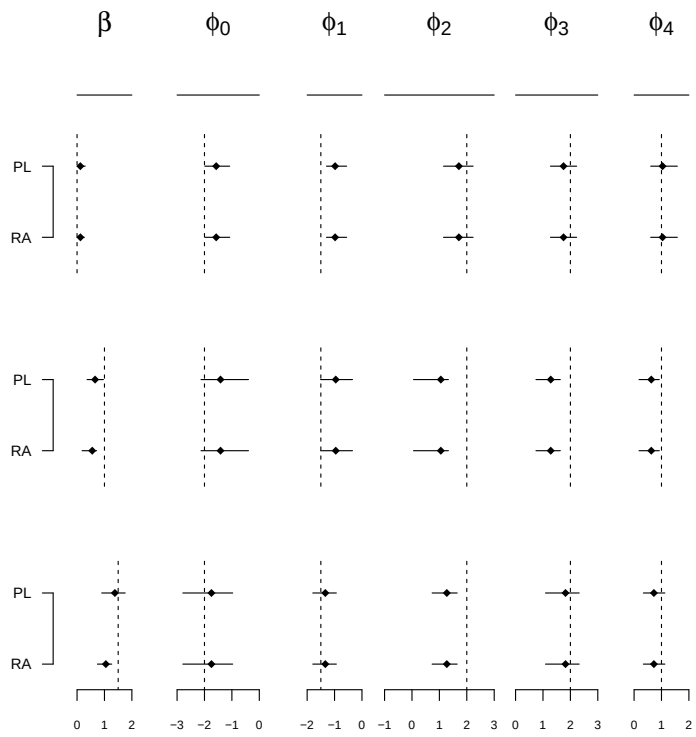

Figure 5.9: *95% credible intervals for β and ϕ_k $(k = 0, \ldots, 4)$ from the complete estimation MCMC, using different methods of approximation (PL and RA). Vertical lines indicate true values of the parameter.*

$\delta_{\boldsymbol{x}}(\boldsymbol{x}^*)$	*One covariate*			*Four covariates*		
	$\beta = 0$	$\beta = 1$	$\beta = 1.5$	$\beta = 0$	$\beta = 1$	$\beta = 1.5$
PL	0.10	0.08	0.04	0.08	0.10	0.06
BS	0.20	0.44	0.64	-	-	-
PR	0.28	0.29	0.09	-	-	-
RA	0.10	0.08	0.04	0.11	0.09	0.07
$\delta_{\boldsymbol{x}}(\boldsymbol{y})$	0.12	0.20	0.24	0.12	0.18	0.22

Table 5.2: *Misclassification rates from MCMC using different methods of approximation (PL, BS, PR, RA) for three maps generated with one covariate and three maps generated with four covariates.*

5.2 Evaluation of variable selection algorithm

We perform an evaluation of the KM method using the same maps as in the previous section, i.e. three maps generated with one covariate and three maps generated with four covariates, using in both cases three different levels of spatial autocorrelation ($\beta = 0, 1, 1.5$) . In the first case $q = 1$, while in the second case $q = 4$. We extend the number of covariates to be used in the variable selection MCMC to p ($p > q$) and consider the following three scenarios:

- **SC1.** A small number of vulnerable predictors with $p = q + 1$.
- **SC2.** A medium number of vulnerable predictors with $p = q + 7$.
- **SC3.** The case with all the available covariates in our study ($p = 17$) (see Table 6.1).

For each map we consider the three scenarios and run the MCMC several times to make sure of the consistency of the results. Each time we add $p - q$ new covariates randomly selected from the database described in Section 2.2, which are not necessarily the same in each repetition (except for the last scenario when we have always all 17 covariates). For example, the maps in the second case were generated using the set of covariates $\{Z_3, Z_8, Z_{11}, Z_{16}\}$; in the first scenario we use $p = 5$, thus, we need to add one new covariate and obtain a set like $\{Z_3, Z_6, Z_8, Z_{11}, Z_{16}\}$ which is used in the variable selection MCMC. The goal of the simulation is to run the MCMC and observe how many of the original covariates are erroneously excluded (s_1) and how many of the additional covariates are erroneously included (s_2). The ideal result would be to get $s_1 = 0$ and $s_2 = 0$, which would indicate that at the end of the procedure we obtain exactly the same covariates that were used in the generation of the map.

We obtain the correlation matrix for the database (Fig 5.10) and observe that the variables used to generate the maps (marked with an asterisk) do not show high correlations (more than 0.6) with any of the other covariates. At most they present medium correlations with some covariates (between 0.4 and 0.6). The problem could be more complicated if there are very high correlations among the covariates. In order to ensure that a variety of possibilities are considered, we select the covariates to be added in each repetition in a random way.

The MCMCs are run for a total of 30,000 iterations, from which the first 10,000 draws are discarded in the burn-in and the remaining 20,000 draws are used to estimate the marginal posterior distribution of γ using each chain separately without combining them. The vector γ^* with the highest frequency is

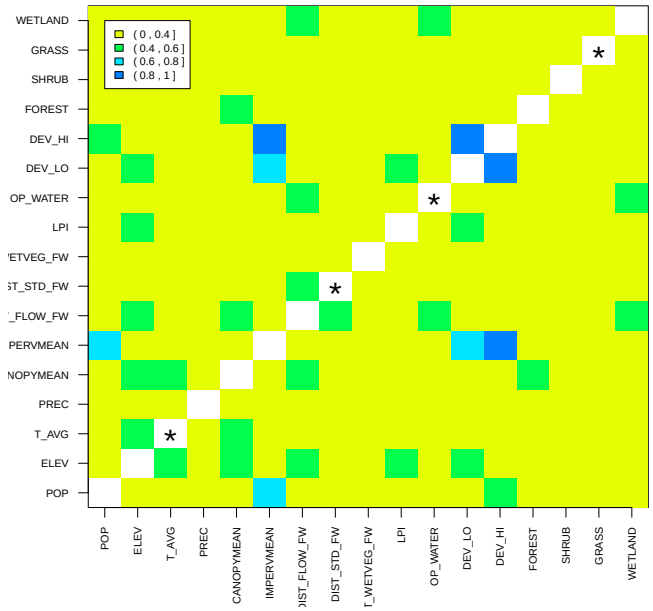

Figure 5.10: *Absolute correlation coefficients among covariates in four categories. Covariates with an asterisk (*) are used to generate the maps.*

selected. We repeat the selection algorithm five times for each combination of β, number of original covariates and number of additional covariates. We want to see the performance of the KM method without using mutation or the principle of MICs; the purpose of running repeated chains is to determine if the final model is consistently selected in the repetitions. The CPU time for running this number of iterations on a 2.2GHz computer is 315 seconds (5.25 minutes).

The model with the highest frequency was very consistent in the five repetitions; in only very few cases the selected model was not the true model. In Table 5.3 we present the results of one repetition for the maps generated with one covariate, while in Table 5.4 we present the results for the maps generated with four covariates. In these tables we include the three models with the highest frequencies. In all cases the statistic s_1 is equal to zero, which indicates that the algorithm always includes all the original covariates. In addition, for the model with the highest frequency, in most of the cases s_2 is equal to zero, and in the only case where it is different from zero, it is equal to one, i.e. the algorithm selects erroneously at most one additional covariate. The frequencies for the modal models (with the highest frequency) are very high when there

is just one additional covariate regardless of the number of original covariates. These probabilities decrease as the number of additional covariates increases, reaching the lowest values when the map is generated with four covariates, and there are 13 additional covariates. The mode of the posterior distribution is the desired vector even when there are more competing models and the distribution is flatter.

Scenario	Image	Covariates	s_1	s_2	Frequency
SC1: $+1$	$\beta = 0$	3	0	0	0.96
		3,11	0	1	0.04
		-	-	-	-
	$\beta = 1$	3	0	0	0.99
		3,11	0	1	0.01
		-	-	-	-
	$\beta = 1.5$	3	0	0	0.99
		3,11	0	1	0.01
		-	-	-	-
SC2: $+7$	$\beta = 0$	3	0	0	0.50
		1,3	0	1	0.23
		3,9	0	1	0.08
	$\beta = 1$	1,3	0	1	0.43
		3	0	0	0.30
		3,9	0	1	0.05
	$\beta = 1.5$	3	0	0	0.59
		1,3	0	1	0.21
		4,3	0	1	0.04
SC3: $17(+13)$	$\beta = 0$	3	0	0	0.39
		3,14	0	1	0.09
		3,1	0	1	0.09
	$\beta = 1$	3	0	0	0.37
		3,1	0	1	0.17
		3,14	0	1	0.07
	$\beta = 1.5$	3	0	0	0.30
		3,1	0	1	0.21
		3,6	0	1	0.03

Table 5.3: *Number of the original covariates erroneously excluded* (s_1)*, additional covariates erroneously included* (s_2)*, and posterior frequency for the three models with the highest frequencies for maps generated with one covariate.*

Scenario	Image	Covariates	s_1	s_2	Frequency
SC1: $+1$	$\beta = 0$	3,8,11,16	0	0	0.96
		1,3,8,11,16	0	1	0.04
		-	-	-	-
	$\beta = 1$	3,8,11,16	0	0	0.97
		1,3,8,11,16	0	1	0.03
		-	-	-	-
	$\beta = 1.5$	3,8,11,16	0	0	0.93
		1,3,8,11,16	0	1	0.07
		-	-	-	-
SC2: $+7$	$\beta = 0$	3,8,11,16	0	0	0.47
		3,8,11,16,17	0	1	0.11
		3,8,11,12,16	0	1	0.09
	$\beta = 1$	3,8,11,16	0	0	0.64
		3,8,9,11,16	0	1	0.07
		2,3,8,11,16	0	1	0.06
	$\beta = 1.5$	3,8,11,16	0	0	0.64
		2,3,8,11,16	0	1	0.08
		3,8,10,11,16	0	1	0.06
SC3: $17(+13)$	$\beta = 0$	3,8,11,16	0	0	0.26
		3,8,11,12,16	0	1	0.07
		3,8,11,16,17	0	1	0.07
	$\beta = 1$	3,8,11,16	0	0	0.32
		3,4,8,11,16	0	1	0.12
		3,8,9,11,16	0	1	0.05
	$\beta = 1.5$	3,8,11,16	0	0	0.25
		3,4,8,11,16	0	1	0.15
		2,3,8,11,16	0	1	0.08

Table 5.4: *Number of the original covariates erroneously excluded* (s_1)*, additional covariates erroneously included* (s_2)*, and frequency for the three models with the highest frequencies for maps generated with four covariates.*

5.3 Summary of evaluations

The motivation to use the interpolations (BS and PR) instead of the accurate PS approximations is the computer time, since the PS is extremely expensive in terms of computation effort. In the simulation with a small vector where we can calculate the exact likelihood, the results of the PS approximations are relatively accurate. However, the interpolations that use the PS approximations as a base, show in general an inappropriate accuracy level. The poor performance

of these interpolations is confirmed in the marginal posterior distributions of the parameters and in the results of the MCMC.

In the examples that we present, the performance of the MCMC is more satisfactory when the original map is generated with more covariates, which produces a higher reduction of the misclassification rate than in the case with only one covariate. Two methods (PL and RA) have a similar performance, either in the size of the credible intervals or in the misclassification rate; however, the PL is much faster than the RA. In a simulation run of the MCMC, the CPU time with the RA is 80 times longer than with the PL. This enormous difference makes the PL approximation more appealing when the MCMC requires long chains.

The KM method for variable selection performs very well in the simulations. Although the frequency of the modal model tends to decrease in the situations when we have a high number of covariates, as is the case of our application, the procedure selects the original covariates in most cases.

Chapter 6

Real data analysis

In this chapter, we perform the analysis of the spatial distribution of two selected species (*Sturnella magna* and *Anas rubripes*), where the first one is generalist and the second one, specialist. A generalist species is able to thrive in a wide variety of environmental conditions and can make use of a variety of different resources, while a specialist species can only thrive in a narrow range of environmental conditions or have a limited diet. We start by giving a description of the species under study; then we perform the analysis of the data by using three models that are interrelated and are presented in increasing complexity: the logistic model, the autologistic model and the spatial hidden Markov model (SHMM). For each model we apply the model selection procedures from Chapter 4, estimate the parameters of the selected model with the algorithm from Chapter 3, and produce maps of the predictions. For the SHMM we provide ecological interpretations to the results and produce maps of the posterior probabilities of presence and maps of observation errors. Finally we perform a posterior predictive assessment of the model fitness and a sensitivity analysis.

6.1 Species description

6.1.1 Eastern Meadowlark

The Eastern Meadowlark (*Sturnella magna*) lives in the eastern part of Canada and the United States, Mexico, Central America and Cuba, and migrates south to several South American countries. These species lives year-round throughout most of its range (Lanyon, 1995) (see Fig. 6.1).

Although Eastern Meadowlarks commonly inhabit native grasslands, pastures and savannas, it also uses a wide variety of anthropogenic grassland habitats (COSEWIC, 2011). No consistent pattern of preference between the two grassland types has been documented (Walk & Warner, 1999). The Eastern Meadowlark closely resembles the Western Meadowlark (*S. neglecta*). Where the range of both species overlaps, Eastern Meadowlarks tend to use wetter, lower-lying grasslands (Lanyon, 1995).

Herkert (1991) states that Eastern Meadowlarks prefer large tracks over smaller fragments and also that breeding densities are associated with grassland area. Territory sizes have been reported in a range of 1.2 to 6 ha (Lanyon, 1957; Francq, 1972). The area required for Eastern Meadowlarks was estimated at 5 ha, in which the area required was defined as the "area at which a species' probability of occurrence equals 50% of its maximum" (Herkert, 1994). Habitat fragmentation is not consistently reported as a determinant factor for breeding density; while a study in Illinois considered the species to be moderately sensitive to grassland habitat fragmentation attributes (Hull, 2000, revised 2002), in Missouri and New York breeding density was not influenced by patch size and the species was not found to be affected by attributes such as edge density, distance to another patch of grassland/forest, cover, patch size, or core area of grassland/forest (Winter, 1998; Horn et al., 2000).

6.1.2 American Black Duck

The American Black Duck (*Anas rubripes*) breeds across northeastern North America from the eastern edge of the Great Plains east to the Atlantic coast and from the Hudson Bay south to the Mid-Atlantic region (see Fig. 6.1). They hide in plain sight in shallow wetlands and often flock with the ubiquitous Mallard (*Anas platyrhynchos*), as they look quite similar to female Mallards. American Black Ducks nest in eastern wetlands including freshwater and saltmarshes (Longcore et al., 2000). They prefer protected bodies of water such as saltmarshes and ponds. During migration and winter, they rest and forage in protected ponds, marshes, and bays. Land-use changes (including drainage associated with agriculture, deforestation, and urbanisation) have altered historical breeding habitats.

Black Ducks frequently mix with other species of ducks, especially Mallards. It has been proposed that in the past, Black Ducks and Mallards were formerly separated by habitat preference, with the dark-plumage Black Ducks having a selective advantage in shaded forest pools in eastern North America, and the lighter plumage Mallards in the brighter, more open prairie and

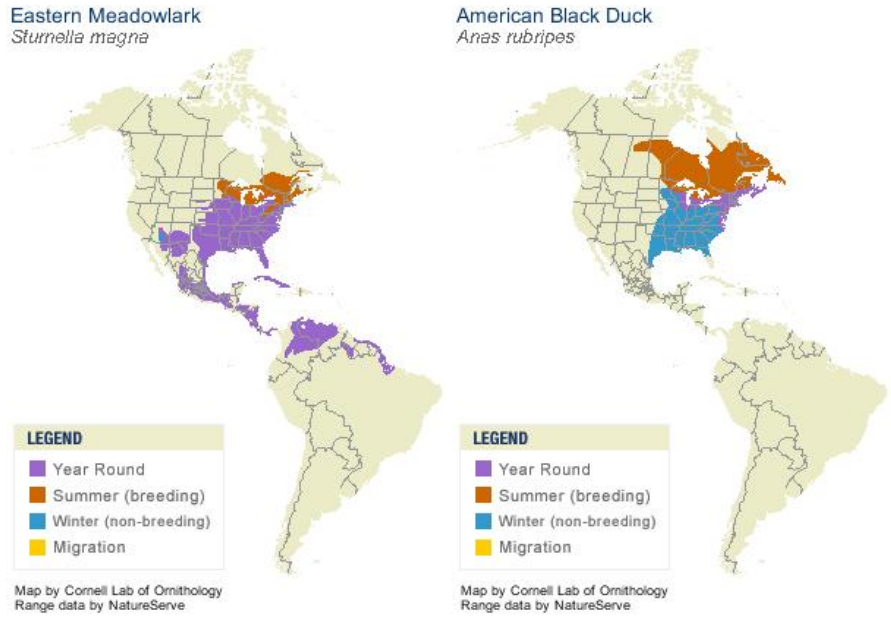

Figure 6.1: *Current global range of Eastern Meadowlark and American Black Duck (NatureServe, 2012).*

lakes in the plains. Deforestation in the East and tree planting on the plains are pointed as causes of the breakdown of habitat separation which, in turn, lead to the high levels of hybridisation now observed (Johnsgard, 1967). Mallard abundance has increased in eastern North America with a concomitant decline in the number of Black Duck in some parts of their range. It has been speculated that hybridisation or competitive exclusion have been the mechanisms for the opposing population trajectories of these species (Seymour & Mitchell, 2006).

6.2 Models for species distribution

We start the analysis with the 24 selected covariates described at the end of Section 2.2. We make an additional reduction in the number of covariates by grouping the habitat classes according to their similarity. We refer to the covariates as $Z_1, \ldots, Z_{17}$ according to Table 6.1, where the new habitat classes are $Z_{11}, \ldots, Z_{17}$. We illustrate these covariates using a two-dimensional representation and quantile categorisation in Appendix C (see Fig C.1, Fig C.2 and Fig C.3). We observe that some of them are spatially correlated and exhibit spatial patterns, as it would be expected for Temperature and Precipitation.

In this section we present the analysis of three models that are interrelated. They are presented in increasing complexity:

- Logistic model: the basic model where we assume that the observed map corresponds to the true map, and the N observed values are independent.
- Autologistic model: a model with a spatial interaction parameter but no observation errors, i.e. we also assume that the observed map corresponds to the true map.
- Spatial hidden Markov model: a model that considers a map degraded by (conditionally) independent noise and the spatial component is modelled under the Markov random field (MRF) assumption.

Code	Name	Short description
Z1	POP	Population
Z2	ELEV	Elevation
Z3	T_AVG	Average temperature
Z4	PREC	Precipitation
Z5	CANOPYMEAN	Mean canopy cover (radius 750)
Z6	IMPERVMEAN	Mean imperviousness to water (radius 750)
Z7	DIST_FLOW_FW	Distance from flowing fresh water
Z8	DIST_STD_FW	Distance from standing fresh water
Z9	DIST_WETVEG_FW	Distance from wet vegetation
Z10	LPI	Percentage area occupied by largest patch (radius 750)
Z11	OP_WATER	Patch density: Open water
Z12	DEV_LO	Patch density: Developed, open space and low intensity
Z13	DEV_HI	Patch density: Developed, medium and high intensity
Z14	FOREST	Patch density: Deciduous, evergreen and mixed forest
Z15	SHRUB	Patch density: Shrub/scrub
Z16	GRASS&CROPS	Patch density: Grassland/herbaceous/pasture/crops
Z17	WETLAND	Patch density: Woody and herbaceous wetlands

Table 6.1: *Codes and names of the covariates included in the model selection procedure.*

6.2.1 Logistic model

The first model for data analysis consists of an ordinary logistic regression on the presence of each species separately, assuming no observation errors. This is a natural and popular statistical model that has been used to explain the observed wildlife distributions by environmental factors (Heikkinen & Högmander, 1997). This approach has been applied, among others, by Walker (1990)

to model presence of kangaroos in Australia and by Osborne & Tigar (1992) to predict the presence of bird species across Lesotho. The logistic model is equivalent to the autologistic model in (3.5) when no spatial autocorrelation is assumed in the process model. The prior of β has all the mass concentrated on $\beta = 0$, i.e. $\pi(\beta) = I_{\{0\}}(\beta)$.

Model selection

We select a smaller subset of covariates using the variable selection algorithm with $\pi(\beta) = I_{\{0\}}(\beta)$. In the ordinary logistic model we assume no observation errors, i.e. the observed map corresponds to the true map ($\boldsymbol{x} = \boldsymbol{y}$). Thus, we only perform Steps 3 and 4 in the MCMC algorithm of Section 4.2 to update the parameters $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}$, respectively. We run the MCMC for a total of 100,000 iterations, with a burn-in of 50,000 samples; the remaining 50,000 samples are used to estimate the marginal posterior distribution of $\boldsymbol{\gamma}$ and select the posterior mode $\boldsymbol{\gamma}^*$. We repeat the algorithm several times with different initial values for $\boldsymbol{\gamma}^{(0)}$; however, the posterior mode that we get in different repetitions is not always the same. We run 10 independent chains (MIC) to obtain the posterior distribution with a total of 500,000 samples. We repeat the MIC and obtain the same posterior mode of $\boldsymbol{\gamma}$. The CPU time for running this number of iterations on a 2.2GHz computer is 950 seconds (15.8 minutes).

Eastern Meadowlark		*American Black Duck*	
Covariates	Frequency	Covariates	Frequency
3,4,5,6,11,13	0.19	2,3,6,7,14	0.15
3,5,6,11,13	0.16	2,3,6,9,14,17	0.11
3,4,5,6,13	0.09	2,3,6,7,14,17	0.10

Table 6.2: *Highest frequency models with the logistic model for the Eastern Meadowlark and the American Black Duck.*

In Table 6.2 we present the three models with the highest frequencies for each species. The proposed subset includes the following covariates: Z_3, Z_4, Z_5, Z_6, Z_{11}, Z_{13} for the Eastern Meadowlark with a frequency of 0.19, and Z_2, Z_3, Z_6, Z_7, Z_{14} for the American Black Duck with a frequency of 0.15. The models with the second highest frequency have covariates very similar to the modal models. Logistic regression for this type of data poses the problem that it tends to produce non-parsimonious models (Wu & Huffer, 1997). By using models that allow for spatial autocorrelation, we hope to require fewer covariates in an

empirical model for the distribution of plant or animal species (Augustin et al., 1996).

Parameter estimation

Once we have selected the variables for each species, we run the MCMC to estimate the parameters of the model for a total of 260,000 iterations, with a burn-in of 10,000 samples and taking a sample every 25 iterations. A total of 10,000 samples are used for estimation. The CPU time for running this number of iterations on a 2.2GHz computer is 131 seconds (2.2 minutes). Diagnostic plots for all the parameters are presented in Appendix C (Fig C.4 and Fig C.5), showing convergence of the chains. Autocorrelation plots show a steep decrease which indicates no strong autocorrelation among samples. In Table 6.3 we present the 95% credible intervals for the regression coefficients of the logistic model for both species.

The fitted values, $p_i = \Pr(X_i = 1)$, in the logistic model are obtained in a straightforward way by using the median value of the posterior distributions of the parameters from Table 6.3 in the following expression:

$$p_i = \frac{\exp(\boldsymbol{z}_i'\boldsymbol{\phi})}{1 + \exp(\boldsymbol{z}_i'\boldsymbol{\phi})}, \tag{6.1}$$

Eastern Meadowlark				*American Black Duck*			
Coeff.	$q_{0.025}$	$q_{0.50}$	$q_{0.975}$	Coeff.	$q_{0.025}$	$q_{0.50}$	$q_{0.975}$
ϕ_0	-1.07	-0.96	-0.86	ϕ_0	-0.58	-0.47	-0.35
ϕ_3	0.31	0.43	0.54	ϕ_2	-1.32	-1.14	-0.95
ϕ_4	-0.30	-0.19	-0.08	ϕ_3	-0.64	-0.51	-0.39
ϕ_5	-0.81	-0.68	-0.55	ϕ_6	0.62	0.80	1.00
ϕ_6	0.19	0.37	0.56	ϕ_7	0.14	0.27	0.40
ϕ_{11}	-0.28	-0.18	-0.08	ϕ_{14}	-0.35	-0.23	-0.12
ϕ_{13}	-0.70	-0.51	0.32				

Table 6.3: *Quantiles* 2.5%, 50% *and* 97.5% *for the parameters of the logistic model for the Eastern Meadowlark and the American Black Duck.*

We use the Pearson residuals r_i which are the standardised differences between the observed values and the fitted values, $r_i = (y_i - p_i)/\sqrt{p_i(1-p_i)}$. In Fig 6.2 we show maps of the residuals, it is clear the spatial dependence of the residuals for both species, with adjacent squares that tend to have similar values or colours in the map.

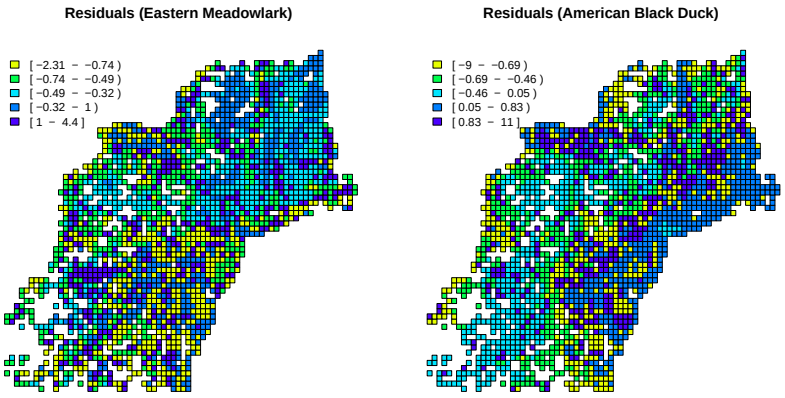

Figure 6.2: *Map of the residuals from the logistic regression model for the Eastern Meadowlark and the American Black Duck.*

Posterior maps

The fitted values are probabilities, thus they are values between 0 and 1. Presence for each species is accepted at a threshold probability. We define this threshold in the most intuitive and common way of a fixed cut-off of $p_i = 0.5$. Manel et al. (2002) have investigated the effects of varying this probability threshold on the performance of presence/absence modelling in ecology using receiver operating characteristic (ROC) curves.

Predicted	Observed			
	Eastern Meadowlark		*American Black Duck*	
	Absent	Present	Absent	Present
Absent	1355	456	1127	337
Present	164	220	180	551
Total	1519	676	1307	888

Table 6.4: *Classification of sites according to the observed and predicted values with the logistic model for the Eastern Meadowlark and the American Black Duck.*

A good model should produce predicted values very close to the observed ones. From Table 6.4 we obtain the misclassification rates for observed values as present (predicted as absent), which are 67% and 38% for the Eastern Meadowlark and for the American Black Duck, respectively. These percentages seem to be very high and are reflected in Fig 6.3, where we plot the observed map

and also the predicted map. We observe how the patches with goldenrod colour (absence or non-observed) dominate the predicted maps. For the other type of misclassification (absent predicted as present) we obtain 11% and 14%.

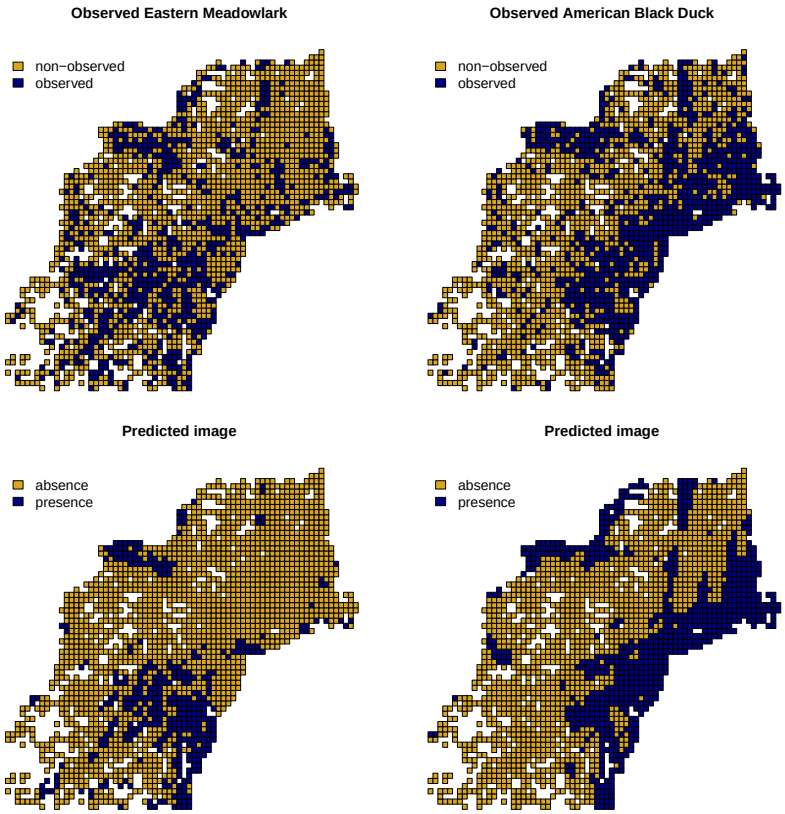

Figure 6.3: *Observed and predicted maps obtained with the logistic model for the Eastern Meadowlark (left) and the American Black Duck (right).*

6.2.2 Autologistic model

The second model that we use to analyse the data is the autologistic model introduced by Besag (1974), in which we add an extra explanatory variable to the logistic model that captures the effect of other response values in the spatial neighbourhood. The conditional probability of a single X_i in the autologistic isotropic model with covariates is defined in (3.4), and the joint probability distribution of the whole vector X is defined in (3.5). In this case we also assume that the observed map corresponds to the true map, i.e. the data are observed with no error. The strength of the spatial autocorrelation is assumed to be positive ($\pi(\beta) = I_{[0,3.5]}(\beta)$).

Model selection

We apply the variable selection algorithm with ($\pi(\beta) = I_{[0,3.5]}(\beta)$), assuming that the observed map corresponds to the true map ($\boldsymbol{x} = \boldsymbol{y}$). In this case we only perform Steps 2, 3 and 4 in the MCMC algorithm of Section 4.2 to update the parameters β, $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}$, respectively. We run 10 independent chains (MIC) to obtain the posterior distribution of the indicator variables $\boldsymbol{\gamma}$ with a total of 500,000 samples, where each chain has 100,000 samples with a burn-in of 50,000 samples. The CPU time for running this number of iterations on a 2.2GHz computer is 985 seconds (16.4 minutes).

In Table 6.5 we present the three models with the highest frequencies for each species. The proposed subset includes three covariates for the Eastern Meadowlark (Z_3, Z_5, Z_{13}) with a frequency of 0.11, and five covariates for the American Black Duck ($Z_2, Z_3, Z_5, Z_6, Z_{11}$) with a frequency of 0.37. We notice how the inclusion of the interaction parameter reduces the number of covariates for the Eastern Meadowlark with respect to the logistic model. In the case of the American Black Duck we get the same number of covariates as before, but some of them are the same as and some are different from those of the logistic model.

Eastern Meadowlark		*American Black Duck*	
Covariates	Frequency	Covariates	Frequency
3,5,13	0.11	2,3,5,6,11	0.37
3,5,12	0.09	2,3,6,11,14	0.07
3,5	0.08	2,3,5,6,11,16	0.05

Table 6.5: *Highest frequency models with the autologistic model for the Eastern Meadowlark and the American Black Duck.*

Parameter estimation

We estimate the parameters of the autologistic model with the covariates previously selected. We use the same number of iterations as in the logistic model. The CPU time for running this number of iterations on a 2.2GHz computer is 245 seconds (4.1 minutes). Plots of the trace and ergodic mean of the samples for all the parameters are presented in Appendix C (Fig C.6 and Fig C.7), showing convergence of the chains. Autocorrelation plots show a steep decrease which indicates no strong autocorrelation among samples. In Table 6.6

we present the 95% credible intervals for the regression coefficients of the autologistic model for both species.

Eastern Meadowlark				*American Black Duck*			
Coeff.	$q_{0.025}$	$q_{0.50}$	$q_{0.975}$	Coeff.	$q_{0.025}$	$q_{0.50}$	$q_{0.975}$
β	0.55	0.65	0.75	β	0.64	0.74	0.85
ϕ_0	-1.89	-1.72	-1.55	ϕ_0	-1.80	-1.60	-1.41
ϕ_3	0.11	0.23	0.35	ϕ_2	-0.68	-0.49	-0.30
ϕ_5	-0.60	-0.48	-0.35	ϕ_3	-0.42	-0.29	-0.15
ϕ_{13}	-0.26	-0.16	-0.06	ϕ_5	-0.39	-0.25	-0.12
				ϕ_6	0.33	0.51	0.70
				ϕ_{11}	0.12	0.24	0.37

Table 6.6: *Quantiles* 2.5%, 50% *and* 97.5% *for the parameters of the autologistic model for the Eastern Meadowlark and the American Black Duck.*

We obtain the fitted values using the median value of the posterior distributions of the parameters from Table 6.6 and the expression for $p_i = \Pr(X_i = 1)$ derived from (3.4):

$$p_i = \frac{\exp\left(\beta \sum_{j\sim i} x_j + \boldsymbol{z}_i'\boldsymbol{\phi}\right)}{1 + \exp\left(\beta \sum_{j\sim i} x_j + \boldsymbol{z}_i'\boldsymbol{\phi}\right)}, \tag{6.2}$$

We obtain the Pearson residuals which are shown in Fig 6.4. Although we expect no spatial dependence among the residuals in this case due to the inclusion of the interaction parameter β, we observe in the maps that the spatial dependence of the residuals has not disappeared.

Predicted	Observed			
	Eastern Meadowlark		*American Black Duck*	
	Absent	Present	Absent	Present
Absent	1353	399	1130	272
Present	166	277	177	616
Total	1519	676	1307	888

Table 6.7: *Classification of sites according to the observed and predicted values with the autologistic model for the Eastern Meadowlark and the American Black Duck.*

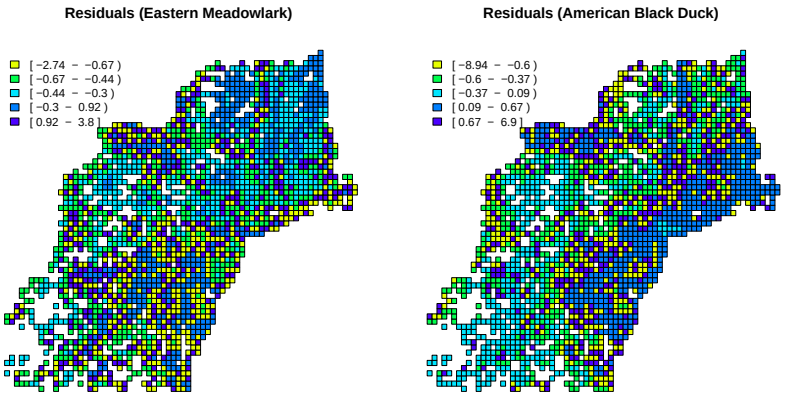

Figure 6.4: *Map of the residuals from the autologistic regression model for the Eastern Meadowlark and the American Black Duck.*

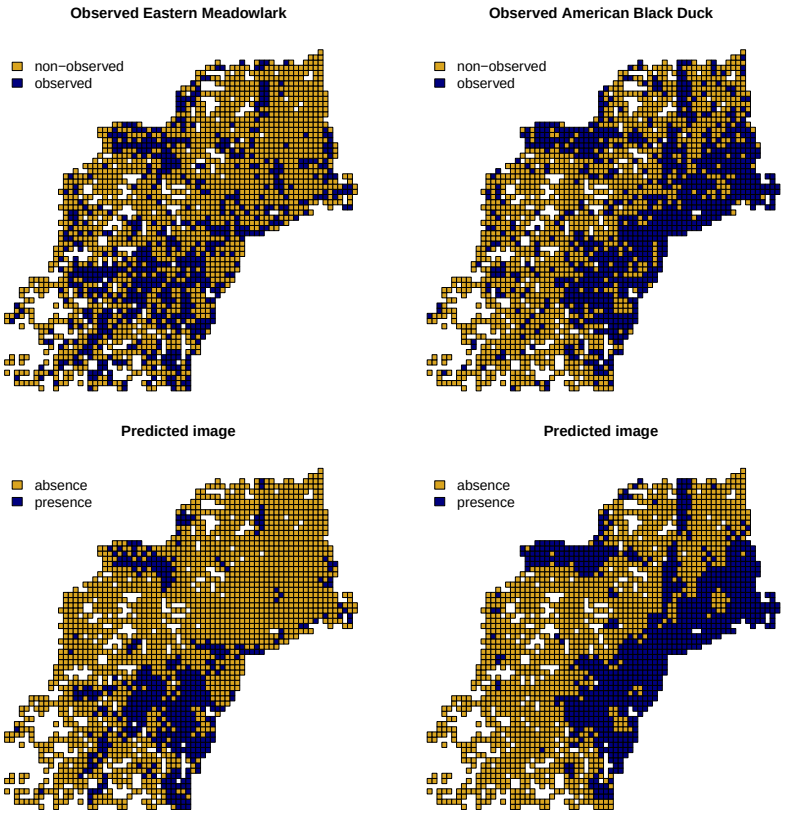

Figure 6.5: *Observed and predicted maps obtained with the autologistic model for the Eastern Meadowlark (left) and the American Black Duck (right).*

Posterior maps

We define the presence of the species in the predicted maps in the same way as the logistic model, with a threshold probability of 0.5. From Table 6.7 we obtain the misclassification rates, which are slightly smaller than in the logistic model: for observed values as present they are 59% and 31% for the Eastern Meadowlark and for the American Black Duck, respectively. The predicted maps in Fig 6.5 do not differ much from those of the logistic model (Fig 6.3).

6.2.3 Spatial hidden Markov model

Model selection

We include the 17 covariates to select a subset for each species using the SHMM. We assume that there is positive spatial autocorrelation ($\pi(\beta) = I_{[0,3.5]}(\beta)$). The spatial interaction parameter can be assumed positive for the kind of data that we want to analyse where we expect that neighbouring sites tend to have a similar condition of either presence or absence of the species. We also include the effort hours to model the probabilities of observation errors. We run the MCMC using 10 independent chains, each one for a total of 100,000 iterations, with a burn-in of 50,000 samples. The model with the highest frequency (modal model) obtained in these chains is not always the same. In Table 6.8 we show the selected variables in each chain and observe that some variables are selected in most of the chains while other variables are selected in only a small number of them. In order to have a model that is more consistently selected, we obtain the posterior distribution combining all 10 independent chains, that is with a total of 500,000 samples. We perform this procedure three times an consistently obtain the same modal models each time. The CPU time for running this number of iterations on a 2.2GHz computer is 1,110 seconds (18.5 minutes).

In Step 4 of the variable selection MCMC, we use different mutation probabilities, $p_m = 0, 0.05, 0.20, 0.50, 0.70$. In the first case, when $p_m = 0$, we have the original KM method, while in the other cases, we have the combination of KM and mutation. Low values for the mutation probability are usually recommended in order to avoid the disruption of good schemata. The mutation operator does not produce any improvement in the mixing of the chains. The results are very similar with either mutation probability.

In Table 6.9 we present the three models with the highest frequencies for each species. The proposed subset includes $Z_3, Z_5, Z_{11}, Z_{12}, Z_{15}, Z_{16}$ for the Eastern Meadowlark with a frequency of 0.15, and $Z_3, Z_5, Z_7, Z_{11}, Z_{13}$ for the

Cov.	Eastern Meadowlark										American Black Duck									
Z_1										×							×		×	
Z_2						×				×		×	×			×		×	×	
Z_3	×	×	×	×	×	×	×	×	×	×		×	×	×	×	×	×	×	×	×
Z_4									×											
Z_5	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×		×	×	×	×
Z_6														×	×					
Z_7												×		×		×	×	×		×
Z_8									×							×		×		
Z_9																				
Z_{10}																				
Z_{11}	×	×	×	×	×		×	×	×		×		×	×	×	×	×		×	×
Z_{12}	×	×	×	×	×		×	×	×											
Z_{13}						×					×	×	×			×		×		×
Z_{14}																×				
Z_{15}	×	×	×		×		×	×	×											
Z_{16}	×	×	×				×	×					×					×	×	
Z_{17}			×								×				×					

Table 6.8: *Modal models obtained in 10 chains for each species with the SHMM. Each column corresponds to the posterior mode of one chain.*

American Black Duck with a frequency of 0.07. The proposed model for the American Black Duck does not include the covariate Z_{17}, which corresponds to patch density of woody and herbaceous wetlands. In these areas, forest or shrubland vegetation accounts for greater than 20% of vegetative cover, or perennial herbaceous vegetation accounts for greater than 80% of vegetative cover. The soil or substrate is periodically saturated with or covered with water. American Black Ducks breed mostly in freshwater wetlands including beaver ponds, brooks lined by speckled alder, shallow lakes with reeds and sedges, bogs in boreal forests, and wooded swamps (Longcore et al., 2000). Wetlands are important to reproduction for two general purposes: cover (availability of protective structure), and abundance and accessibility of forage organisms. Palustrine emergent, scrub-shrub and deciduous forested wetlands provide optimal cover and forage conditions for brood-rearing (Ringelman et al., 1982). Thus, we consider important to include this covariate in the model.

We run again the MCMC for variable selection with $\gamma_{17} = 1$ fixed. We follow the same procedure explained above and obtain a new model with the following covariates: $Z_3, Z_5, Z_{11}, Z_{13}, Z_{17}$, with a frequency of 0.08 (see model (II) in Table 6.9). This model excludes the covariate Z_7, which corresponds to

Eastern Meadowlark		*American Black Duck*			
Covariates	Freq.	Covariates (I)	Freq.	Covariates (II)	Freq.
3,5,11,12,15,16	0.15	3,5,7,11,13	0.07	3,5,11,13,17	0.08
3,5,11,12,15,16,17	0.07	2,3,5,6,7,11	0.04	3,5,7,11,13,17	0.04
3,4,5,11,12,15,16	0.03	3,5,6,7,11	0.04	5,11,13,17	0.04

Table 6.9: *Highest frequency models with the SHMM for the Eastern Meadowlark and the American Black Duck.*

distance from flowing fresh water. In the case of the American Black Duck, it is less important to make depend the presence of the species on a variable that accounts for an habitat not frequented by this species.

When we observe the posterior distributions of the indicators $(\gamma_k;\ k = 1,\ldots,p)$ in the individual chains, there are some covariates that are never selected in the modal models corresponding to those distributions. For the Eastern Meadowlark, five covariates are never selected: mean imperviousness to water (Z_6), distance from flowing fresh water (Z_7), distance from wet vegetation (Z_9), percentage of area occupied by largest patch (Z_{10}), and patch density of deciduous, evergreen and mixed forest (Z_{14}). For the American Black Duck, four covariates are never selected in the modal models corresponding to individual posterior distributions of independent chains: precipitation (Z_4), distance from wet vegetation (Z_9), percentage of area occupied by largest patch (Z_{10}), and patch density of shrub/scrub (Z_{15}).

Parameter estimation

We estimate the posterior distribution of the parameters of the selected models using 10,000 samples (burn-in of 10,000 samples and taking a sample every 25 iterations). The CPU time for running this number of iterations on a 2.2GHz computer is 585 seconds (9.8 minutes). Plots of the trace and ergodic mean of the samples for all the parameters are presented in Appendix C (Fig C.8 and Fig C.9), showing convergence of the chains. Autocorrelation plots indicate no strong autocorrelation among the samples for all the parameters.

In Table 6.10 we present the median and the quantiles $(2.5\%$ and $97.5\%)$ that define 95% credible intervals for the parameters of the model for the Eastern Meadowlark. In Fig 6.6 we present the marginal posterior distributions of these parameters which are unimodal and symmetric. The parameter space for the prior distribution of β is $[0, 3.5]$, and for ϕ_k it is $[-5, 5]$ $(k = 0,\ldots,q)$.

Coeff.	$q_{0.025}$	$q_{0.50}$	$q_{0.975}$
β	0.60	0.78	0.96
ϕ_0	-1.71	-1.32	-0.94
ϕ_3	0.13	0.31	0.50
ϕ_5	-0.92	-0.69	-0.48
ϕ_{11}	-0.41	-0.23	-0.05
ϕ_{12}	-0.62	-0.46	-0.30
ϕ_{15}	-0.38	-0.22	-0.08
ϕ_{16}	0.06	0.23	0.43

Table 6.10: *Quantiles* 2.5%, 50% *and* 97.5% *for the parameters of the SHMM for the Eastern Meadowlark.*

We observe that most of the distribution of each parameter is concentrated in a small range of the parameter space, which is an indication that the data are very informative. Annual average temperature (Z_3) is an important covariate because Eastern Meadowlarks eat crops and seeds; sites with higher average temperatures facilitate the conditions for a longer period of crop production. During the summer months most of their food consists of insects and closely allied forms. They eat practically all of the principal pests of the fields and are particularly destructive to the dreaded cutworms, caterpillars, beetles, and grasshoppers. In the autumn, and especially in winter, when insect life is scarce, they resort in a large measure to seeds. They do feed on certain grains useful to man, such as corn, wheat, rye, and oats; but most of these are waste left behind at harvest time (Bent, 1989). Soon after his arrival to a new territory, the resident male leaves his companions and selects preferably a grassland or meadow, because of the great abundance of food as well as his decided liking for this type of habitat. Although, the coefficient of patch density of grassland, herbaceous, pasture and crops (Z_{16}) is not as high as expected, this variable is definitely an important determinant of the presence of the Eastern Meadowlark.

There are four covariates with negative coefficients: mean canopy cover (Z_5), patch density of open water (Z_{11}), patch density of developed, open space and low intensity (Z_{12}), and patch density of shrub and scrub (Z_{15}). The first one (Z_5) indicates the clear absence of interest of Eastern Meadowlarks on visiting forests, mainly because they are not adequate for foraging or nesting. These birds usually nest directly on the ground in litter or under dense, overhanging grasses (Lanyon, 1995); thus, they build their nests in grasslands,

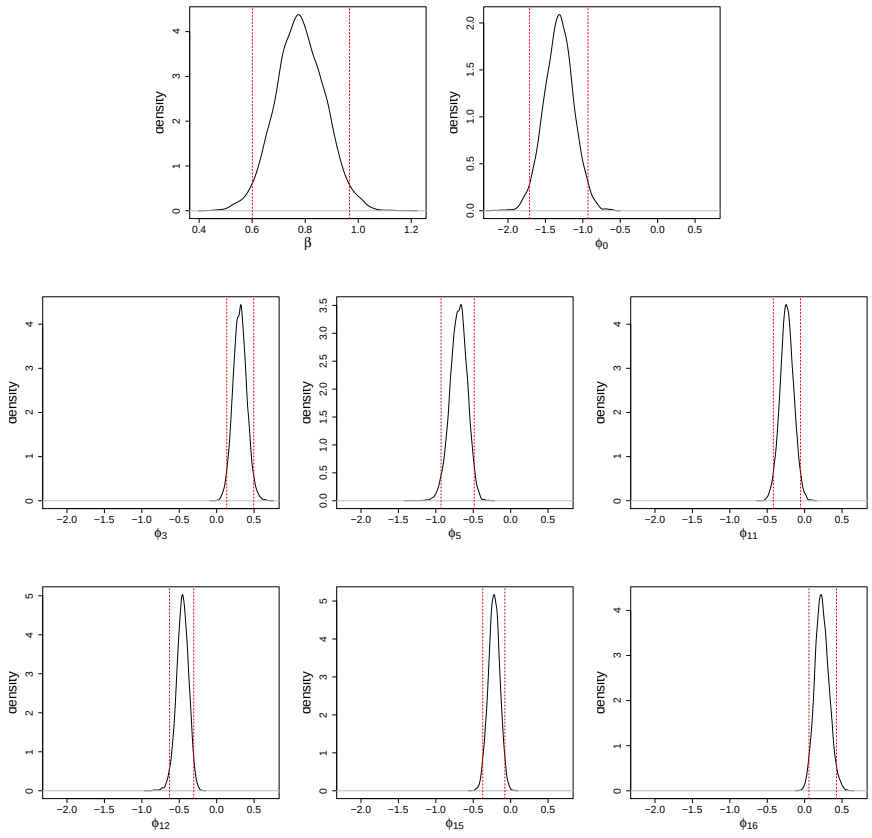

Figure 6.6: *Posterior distribution of the parameters of the SHMM for the Eastern Meadowlark. Vertical red lines indicate the quantiles* (2.5% *and* 97.5%*) that define* 95% *credible intervals for the parameters.*

meadows, and pastures, but never on the trees. For the second one (Z_{11}), we consider the fact that passeriformes are species that in general avoid open waters. The third one (Z_{12}) could be related to habitat fragmentation due to low intensity development. This variable accounts for the number of patches of areas with a mixture of some constructed materials, vegetation in the form of lawn grasses, and areas where impervious surfaces account for less than 50% of total cover. These areas most commonly include large-lot single-family housing units, parks, golf courses, and vegetation planted in developed settings for recreation, erosion control, or aesthetic purposes (Fry et al., 2011). Eastern Meadowlarks prefer large grassland areas over small areas for breeding (Herkert, 1994); thus, their presence is diminished when the land is fragmented by human presence. The last one (Z_{15}) confirms the preference of the

Eastern Meadowlarks of grasslands over shrub or scrub. They find the main sources of food in the grass, but when these sources are scarce, they move to this sub-optimal habitat to get insects or small fruits.

Coeff.	$q_{0.025}$	$q_{0.50}$	$q_{0.975}$
β	0.77	0.97	1.16
ϕ_0	-1.66	-1.19	-0.62
ϕ_3	-0.47	-0.26	-0.06
ϕ_5	-0.79	-0.56	-0.34
ϕ_{11}	0.28	0.61	1.00
ϕ_{13}	0.19	0.46	0.76
ϕ_{17}	0.03	0.26	0.55

Table 6.11: *Quantiles* 2.5%, 50% *and* 97.5% *for the parameters of the SHMM for the American Black Duck.*

In Table 6.11 we have the median and the quantiles (2.5% and 97.5%) that define 95% credible intervals for the parameters of the model for the American Black Duck. In Fig 6.7 we present the marginal posterior distributions of these parameters, which are unimodal and symmetric. We observe that the estimate of the coefficient of patch density of woody and herbaceous wetlands (Z_{17}) is positive, as we expected. Other positive coefficients correspond to the covariates patch density of open water (Z_{11}), and patch density of developed, medium and high intensity (Z_{13}). The first one (Z_{11}) includes areas of open water, generally with less than 25% cover of vegetation or soil. Food availability, freedom from disturbance, protection from severe weather, and presence of large bodies of open water are interrelated factors that appear to affect habitat use by American Black Ducks in winter (Lewis & Garrison, 1984). The second one (Z_{13}) is an interesting covariate from the fact that urban areas tend to have ponds. Some ponds are created specifically for habitat restoration, including water treatment. Others, like water gardens, water features and koi ponds are designed for aesthetic ornamentation as landscape or architectural features (Clegg & Mansell, 1986). American Black Ducks are so ever present in all of the inner city ponds. In the ponds, they have almost become domesticated because so many people in the city feed them. When hunting season starts, wild ducks move into city ponds where there is no shooting allowed (de Leon, 2010). Covariates Z_{11}, Z_{13} and Z_{17} altogether are complementary in the sense that wild individuals are found in the open waters and wetlands, while domestic ones

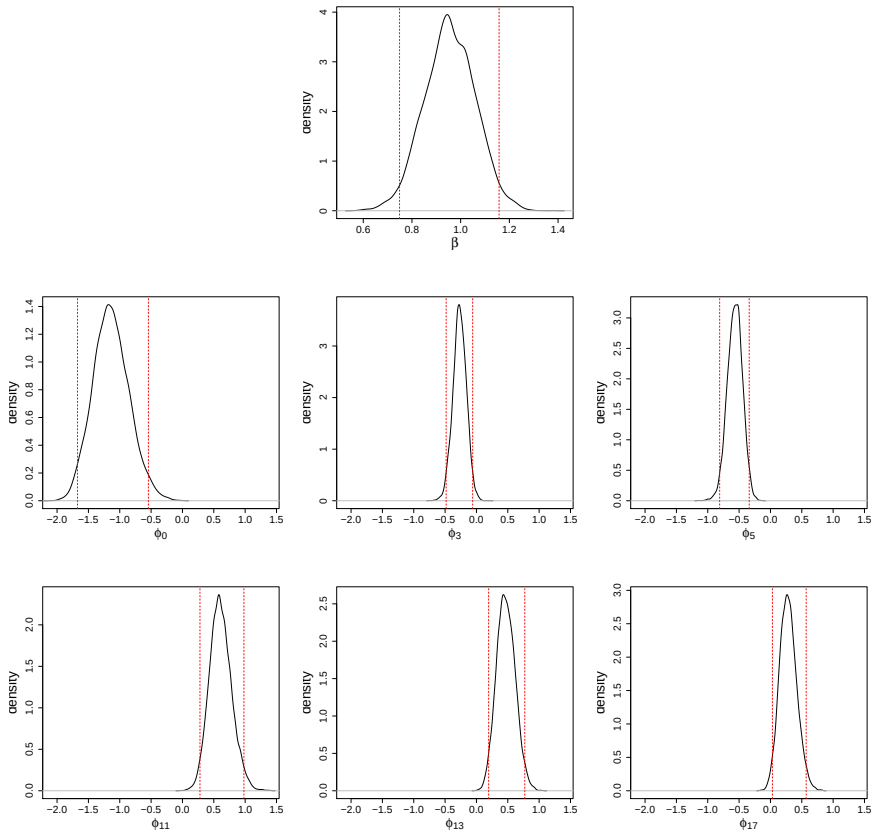

Figure 6.7: *Posterior distribution of the parameters of the SHMM for the American Black Duck. Vertical red lines indicate the quantiles (*2.5% *and* 97.5%*) that define* 95% *credible intervals for the parameters.*

live in the ponds of urban areas.

Two covariates have negative coefficients: annual average temperature (Z_3) and mean canopy cover (Z_5). The fact that Z_5 has a negative coefficient indicates that American Black Ducks do not look for trees in covered areas, they just need them for isolated nests and always surrounded by sources of water. The female selects the nest site, usually in a clump of grass, under a shrub or tree, or in a hole or fork in a tree, near the ground (Wright & Wyndham, 2005).

Weather and temperature (Z_3) are very important for migration. As temperatures drop and the feeding areas freeze over one by one, the southward migration starts. Cold, windy, and snowy fronts that lock up food and water stimulate major movements. Birds go where they find food, water, and safety. They stay as long as these habitat qualities satisfy their needs. Migration

is an adaptive strategy hinged on finding suitable habitats. Waterfowl may remain north of traditional terminuses until harsh environmental conditions force them southward. The reason to do that is because migration is physically costly, and the real rewards are survival and fitness; hence, some waterfowl remain as close to the breeding grounds as their bodies and resources permit (Kaminski, 2007). All ducks tend to return in fall and winter to the same marshes that they visited the previous year, but this trait is most pronounced in the American Black Duck. When tidal feeding areas have become frozen in New England, some American Black Ducks have starved rather than migrate farther south to unfamiliar ground (Wright & Wyndham, 2005). Although, cold temperatures condition southward movements, it is also clear that American Black Duck's original breeding grounds correspond to places with lower temperatures.

Posterior maps

The median of the spatial interaction parameter β in the SHMM is expected to be between 0.60 and 0.96 for the Eastern Meadowlark, while it is expected to be between 0.77 and 1.16 for the American Black Duck (with 95% chance) (see Table 6.10 and Table 6.11). Thus, we expect medium size patches of sites with presence of the Eastern Meadowlark and even larger patches of sites with presence of the American Black Duck.

As a result of the estimation procedure of the parameters, we get the posterior distribution of X. Now, we are interested on the conditional probability of presence for each X_i ($i \in \mathcal{S}$), given Y. An estimate of this probability can be obtained as the proportion of times that the species of interest is present at the i^{th} site, that is the number of times that $X_i = 1$, divided by the number of samples used to estimating the posterior distribution. These probabilities are used to create posterior maps of presence of the species. One way to create a map is by categorising the probabilities by levels of evidence of presence or absence of the species at each site. We use a total of 6 categories for the following levels:

- Strong evidence of absence: $\Pr(X_i = 1|\boldsymbol{y}) < 0.1$.
- Medium evidence of absence: $0.1 \leq \Pr(X_i = 1|\boldsymbol{y}) < 0.3$.
- Weak evidence of absence: $0.3 \leq \Pr(X_i = 1|\boldsymbol{y}) < 0.5$.
- Weak evidence of presence: $0.5 \leq \Pr(X_i = 1|\boldsymbol{y}) < 0.7$.
- Medium evidence of presence: $0.7 \leq \Pr(X_i = 1|\boldsymbol{y}) < 0.9$.
- Strong evidence of presence: $\Pr(X_i = 1|\boldsymbol{y}) \geq 0.9$.

In the maps (Fig. 6.8), we use shades of blue for the three categories that correspond to evidence of presence, where darker blue indicates stronger evidence of presence. On the other hand, we use shades of goldenrod for evidence of absence, where darker goldenrod indicates stronger evidence of absence.

In addition to the maps of probability of presence, we can obtain a reconstruction of the true map. These maps are obtained by using the posterior mode of each site. We assign a value of 1 to sites where the unconditional probability of presence is equal or greater than 0.5; otherwise, we assign a value of 0.

Differences between observed and reconstructed maps arise due to non-observed presences rather than false observations. We use the median of the posterior distribution of $\theta_{0,i}$ and $1 - \theta_{1,i}$, as an estimate of the probabilities of error at each site. The estimated probabilities of false observation are small for both species: $\theta_{0,i} < 0.001$ $(\forall\, i \in \mathcal{S})$ for the American Black Duck, and 95% of the sites have $\theta_{0,i} < 0.007$ for the Eastern Meadowlark. The probabilities of non-observed presence have a higher variability among sites $(0 \leq 1-\theta_{1,i} < 0.71)$ for both species. We use the inter-quartile range (IQR) of the posterior distribution of $\theta_{0,i}$ and $1 - \theta_{1,i}$ as a measure of the uncertainty of these estimates. The IQR for $\theta_{1,i}$ is always lower than 0.04 (see Fig. C.12 in Appendix C), which is an indication of the high precision of the estimates of these probabilities.

Because of the influence of effort hours on determining the probabilities of error, these probabilities are very similar for both species, as we can observe in Fig. 6.9 (the graphs in the center panel look very similar). The higher the effort hours, the lower the probability of error. Nonetheless, the sites that are classified with presence in the reconstructed maps even when the species was not observed (non-observed presence), vary from one species to the other. These sites are coloured with dark blue in the top panel of Fig. 6.9. The categories with lighter colours correspond to the sites that match between the observed

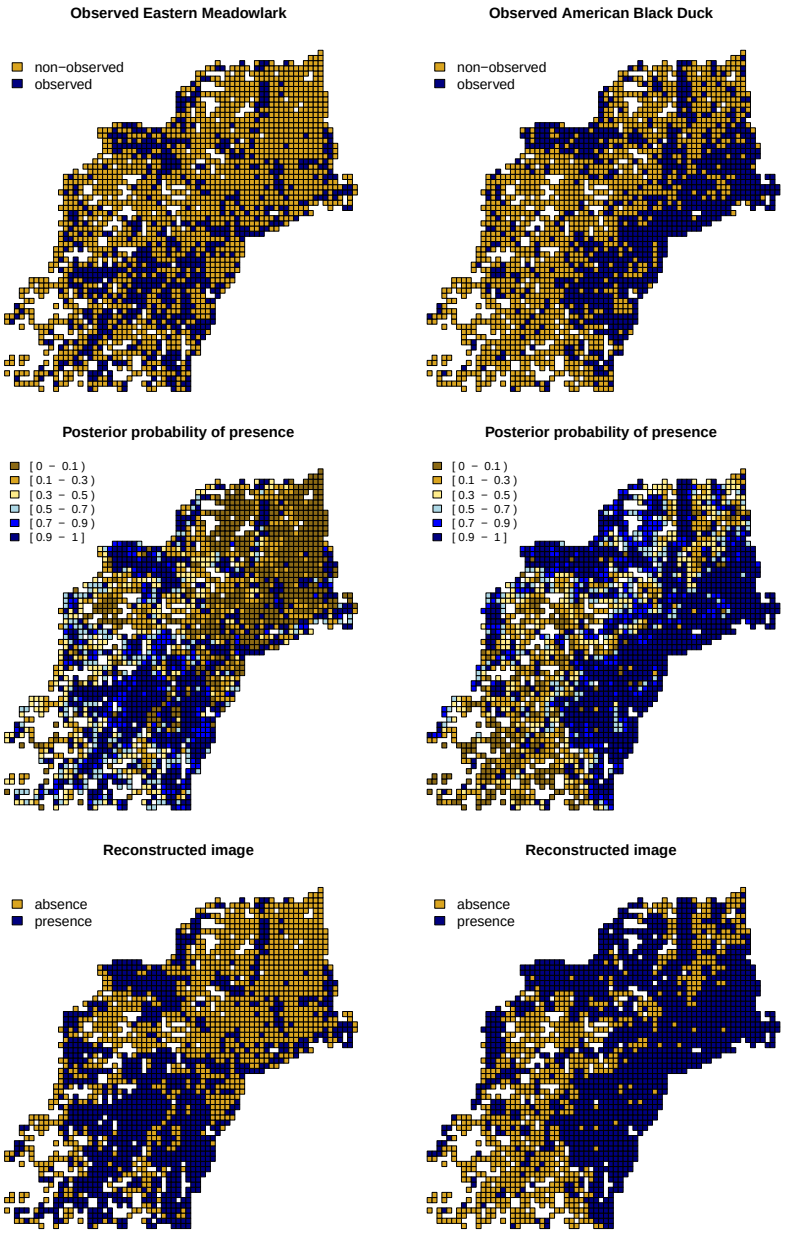

Figure 6.8: *Observed map, map of posterior probability of presence, and reconstructed map obtained with the SHMM for the Eastern Meadowlark (left) and the American Black Duck (right).*

map and the reconstructed map (the species was observed in the site and it was classified as present in the reconstructed map, or the species was not observed and it was classified as absent). Out of the total sites with reported

Reconstructed map	Non-observed (%)	Observed (%)	Total (%)
Eastern Meadowlark			
Absent	1139	0	1139
	(65)	(0)	(52)
Present	380	676	1056
	(35)	(100)	(48)
Total	1519	676	2195
(%)	(69)	(31)	(100)
American Black Duck			
Absent	822	0	822
	(63)	(0)	(37)
Present	485	888	1373
	(37)	(100)	(63)
Total	1307	888	2195
(%)	(60)	(40)	(100)

Table 6.12: *Classification of sites according to the observed and reconstructed maps for the Eastern Meadowlark and the American Black Duck with the SHMM. Numbers in parenthesis correspond to column percentages.*

absence of the species, those classified with presence (non-observed presence) represent 35% for the Eastern Meadowlark and 37% for the American Black Duck. On the other hand, none of the sites where the species was observed, was classified with absence in the reconstructed maps (false observations) (see Table 6.12). Thus, the reconstruction of the map could be seen as filling gaps on the observed map, by changing some goldenrod cells to blue ones and creating in this way bigger blue patches (see top and bottom panels in Fig. 6.8). We obtain a map for the Eastern Meadowlark with 48% of blue sites from an observed map with 31% of them; while the number of blue sites raises from 40% to 63% for the American Black Duck.

The fact that the data correspond to a whole year introduces noise; different sites were visited in different dates. Some sites may have been visited when the American Black Duck was not present due to migration; however, the map intends to represent the distribution over the year. Thus, American Black Ducks could be non-observed in sites where they actually live during a period of the year. On the other hand, Eastern Meadowlarks are less affected by this source of error, instead, their presence is underestimated mainly by other factors like

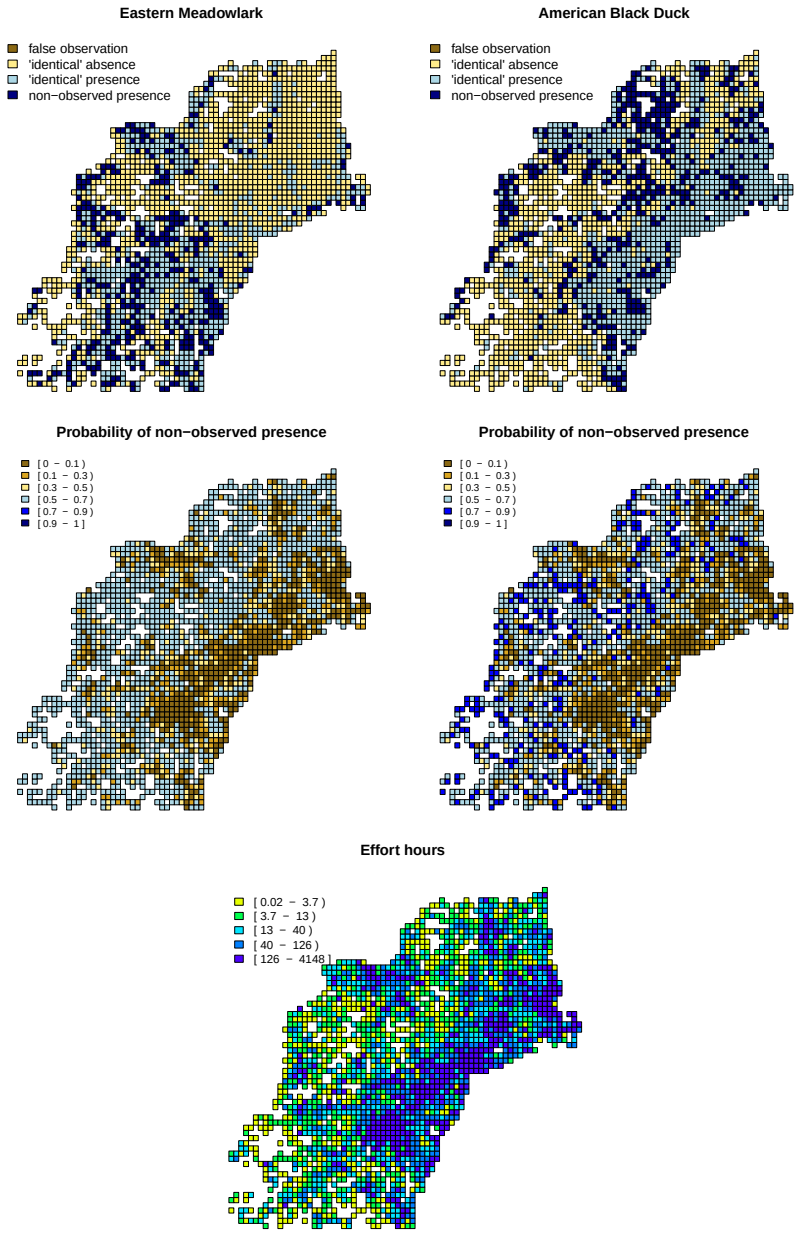

Figure 6.9: *Top: classification of sites according to matching between observed and reconstructed maps. Middle: map of the probability of non-observed presence obtained with the SHMM for the Eastern Meadowlark (left) and the American Black Duck (right). Bottom: Effort hours per site.*

song volume, time of day when the observation was done, pairing status, and stage of nesting cycle.

Hyperparameters

We use four hyperparameters for modelling the observation error probabilities $\theta_{0,i}$ and $1 - \theta_{1,i}$. These hyperparameters are α_k $(k = 1, \ldots, 4)$ and they are assumed to have independent diffuse Gaussian distributions with a mean of 0 and high variance. The first two hyperparameters α_1 and α_2 are used to determine $\theta_{0,i}$, while α_3 and α_4 are used to determine $\theta_{1,i}$ (see equation 3.7).

In Fig. C.10 and Fig. C.11 (Appendix C) we observe that it remains some autocorrelation only for the two hyperparameters related to $\theta_{0,i}$ (α_1 and α_2). However, plots of the ergodic means show convergence for these hyperparameters. We take the median of the posterior distribution of $\theta_{0,i}$ and $1 - \theta_{1,i}$ as an estimate of the probabilities of error at each site. The estimated probabilities of false observation $(\theta_{0,i})$ are very small for both species, and we could even hypothesise that they are equal to zero; thus, the hyperparameters α_1 and α_2 could be eliminated from the model.

6.3 Posterior predictive assessment

In the context of regression models, we expect the model to provide a good fit of the data, i.e. the differences between observed and predicted values should be as small as possible. When we use a SHMM, the closeness between the observed and reconstructed map depends on the amount of error present in the data. If the data are recorded with small error we would expect a reconstructed map very similar to the observed one; however, when the data are subject to high levels of observation error, they tend to be very different from the the true values. Thus, when we reconstruct the map we may obtain many values that differ from the data. Differences between observed and reconstructed maps are not necessarily an indication of a wrong model. We use the discrepancy measure defined in (4.14) with 1,000 vectors $\boldsymbol{\theta}^{(j)} = (\theta_{0,1}^{(j)}, \ldots, \theta_{0,N}^{(j)}, \theta_{1,1}^{(j)}, \ldots, \theta_{1,N}^{(j)})'$, and the corresponding maps $\boldsymbol{x}^{(j)}$ $(j = 1, \ldots, 1,000)$. For each pair $(\boldsymbol{x}^{(j)}, \boldsymbol{\theta}^{(j)})$ we generate a replicated data vector $\boldsymbol{y}^{r(j)}$ using (4.13). Next, we calculate discrepancies for each replicated data set, $D(\boldsymbol{y}^{r(j)}; \boldsymbol{\theta}^{(j)})$, and for the observed data, $D(\boldsymbol{y}; \boldsymbol{\theta}^{(j)})$.

In Fig. 6.10 we observe that the predictive discrepancies are higher than the realised discrepancies over half of the times for the SHMM (for both species). The tail-area probabilities deduced from this plots are 0.59 for the Eastern Meadowlark and 0.62 for the American Black Duck (see Table 6.13), which are the proportion of points above the 45°red line in the figures. This is not saying that the models are correct, but only that the values of the discrepancy measures we

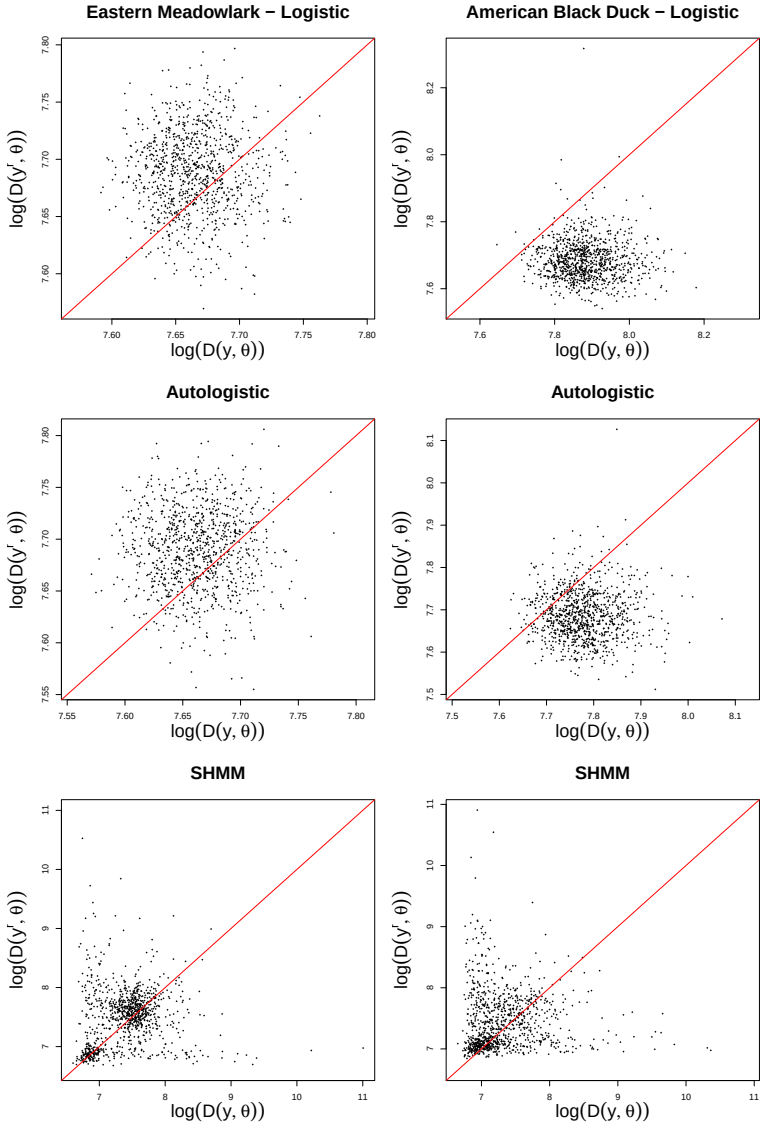

Figure 6.10: *Scatterplot of predictive* $(D(\boldsymbol{y}^r;\boldsymbol{\theta}))$ *vs. realised* $(D(\boldsymbol{y};\boldsymbol{\theta}))$ *discrepancies for the logistic model(top), the autologistic model (middle) and the SHMM (bottom) for the Eastern Meadowlark (left) and the American Black Duck (right).*

have examined are reasonable under the posited models. We know that the logistic and autologistic models show spatially correlated residuals (see Fig. 6.2 and Fig. 6.4). This problem is due to the absence of a spatial component in the logistic model and the strong influence of effort hours that generates observation errors not considered in these two models. However, the tail-area probabilities for the Eastern Meadowlark are 0.72 and 0.73 for the logistic and the

autologistic models, respectively. These values are even higher than the corresponding value for the SHMM. We complement these results with the fact that the reconstructed map obtained from the SHMM for the American Black Duck differs from the predicted maps obtained with the other two models. The logistic and autologistic models assume that the observed map corresponds to the true map which is clearly contradicted with the high levels of observation errors obtained with the SHMM. The result reveals the low power of the measure being used in this case. Other measures as Bayes factor and deviance information criteria (DIC) for hidden Markov models (McGrory & Titterington, 2009) should be used to clarify this issue. For the American Black Duck we get evidence against the logistic model since the tail-area probability is 0.01, and also a fairly low value for the autologistic model (0.13) compared to the value for the SHMM (0.62).

Model	*Eastern Meadowlark*	*American Black Duck*
Logistic	0.72	0.01
Autologistic	0.73	0.13
SHMM	0.59	0.62

Table 6.13: *Tail-area probabilities for discrepancies under the logistic model, the autologistic model and the SHMM for the Eastern Meadowlark and the American Black Duck.*

In Table 6.14 we summarise the covariates selected in each model and show the median of the posterior distributions of the corresponding parameters. It is noticeable the differences in the selected covariates in the three models. The logistic and the autologistic miss important variables for the Eastern Meadowlark, e.g. patch density of shrub/scrub (Z_{15}) and patch density of grassland and crops (Z_{16}), or they include unimportant variables, e.g. patch density of developed, medium and high intensity (Z_{13}). For the Easten Meadowlark the covariates selected by the logistic model are almost totally different from those selected by the SHMM. We notice the concordance with the signs of the coefficient among the three models when a covariate is included in more than one of them, even if the magnitude is not always similar (for both species). For example, the magnitude of ϕ_5 for the SHMM (-0.56) is as double as that for the autologistic model (-0.25) in the case of the American Black Duck.

Coeff.	*Eastern Meadowlark*			*American Black Duck*		
	Logistic	Autologistic	SHMM	Logistic	Autologistic	SHMM
β	-	0.65	0.78	-	0.74	1.16
ϕ_0	-0.96	-1.72	-1.32	-0.47	-1.60	-1.19
ϕ_1	-	-	-	-	-	-
ϕ_2	-	-	-	-1.14	-0.49	-
ϕ_3	0.43	0.23	0.31	-0.51	-0.29	-0.26
ϕ_4	-0.19	-	-	-	-	-
ϕ_5	-0.68	-0.48	-0.69	-	-0.25	-0.56
ϕ_6	0.37	-	-	0.80	0.51	-
ϕ_7	-	-	-	0.27	-	-
ϕ_8	-	-	-	-	-	-
ϕ_9	-	-	-	-	-	-
ϕ_{10}	-	-	-	-	-	-
ϕ_{11}	-0.18	-	-0.23	-	0.24	0.61
ϕ_{12}	-	-	-0.46	-	-	-
ϕ_{13}	-0.51	-0.16	-	-	-	0.46
ϕ_{14}	-	-	-	-0.23	-	-
ϕ_{15}	-	-	-0.22	-	-	-
ϕ_{16}	-	-	0.23	-	-	-
ϕ_{17}	-	-	-	-	-	0.26
p-discr.	0.72	0.73	0.59	0.01	0.13	0.62
Freq.	0.19	0.11	0.15	0.15	0.37	0.08

Table 6.14: *Median values of the parameters for the logistic, autologistic and SHMM for the Eastern Meadowlark and the American Black Duck. Tail-area probabilities (p-discr.) and frequencies of the modal models are included.*

6.4 Sensitivity analysis

Finally, we perform a sensitivity analysis on the assumptions for β. In Section 3.1. We estimate the vector of parameters $\boldsymbol{\phi}$ under four priors, and observe if the estimates change:

- (A1) We assume the spatial interaction parameter to be positive since we expect that neighbouring sites tend to have a similar condition of either presence or absence of the species: $\pi(\beta) = I_{[0,3.5]}(\beta)$.

- (A2) We relax assumption (A1) by allowing β to take negative values: $\pi(\beta) = I_{[-3.5,3.5]}(\beta)$. Negative values of β would be expected under competition and allelopathy. Although, we think this is not the case of the

species under study, we want to assess if the procedure is able to select a reasonable value of β even under such a wide range of possibilities.

- (A3) We assume $\pi(\beta) = I_{[-3.5,3.5]}(\beta)$ as in the case of the logistic regression, but in this case we still have a hidden map and an observed map.
- (A4) We use the same prior for β as in (A1), but we suppress the hidden structure by assuming no observation errors: $\pi(\beta) = I_{[0,3.5]}(\beta)$ and $\boldsymbol{x} = \boldsymbol{y}$.

We include the same set of covariates for the four priors ($Z_3, Z_5, Z_{11}, Z_{12}, Z_{15}, Z_{16}$ for Eastern Meadowlark and $Z_3, Z_5, Z_{11}, Z_{13}, Z_{17}$ for American Black Duck). In Fig. 6.11 and Fig. 6.12 we present 95% credible intervals for the parameters estimated using the previous assumptions. Results under (A1) and (A2) are almost identical, which shows robustness of the algorithm under the extension of the parameter space for β. We confirm that it is reasonable to expect the formation of medium size patches of sites inhabited by each of these two species.

The results under (A3) are not very different from (A1) for most of the coefficients. The only one that changes dramatically is ϕ_0. In this case, all the information of the neighbourhood is cancelled (since $\beta = 0$), and the external field (ϕ_0) intends to substitute it in the model. Thus, it is not surprising that ϕ_0 changes so much. Another noticeable change is that of ϕ_{17} for the American Black Duck, whose credible interval in (A3) is moved to the right from (A1). Furthermore, the variability of the posterior distribution of all the coefficients increase.

The last prior (A4) produces estimates with less variability than (A1). This is reasonable because the noise induced by the hidden structure is eliminated and the model assumes that the data correspond to the real values. Although, this is not a good assumption, we observe how the estimates change in the expected direction when we impose a strong condition as this one.

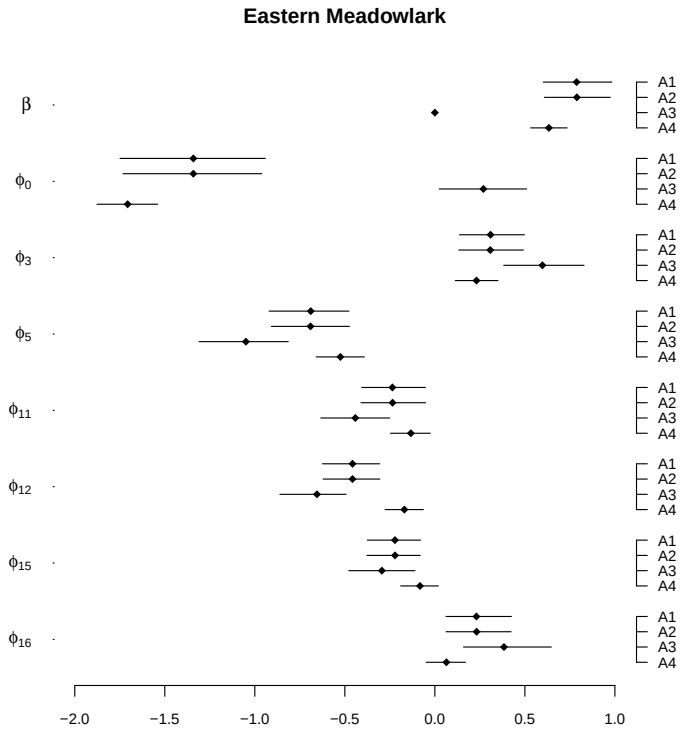

Figure 6.11: *95% credible intervals for β and ϕ_k $(k = 0, \ldots, q)$ with different assumptions: (A1) $\pi(\beta) = I_{[0,3.5]}(\beta)$; (A2) $\pi(\beta) = I_{[-3.5,3.5]}(\beta)$; (A3) $\pi(\beta) = I_{[-3.5,3.5]}(\beta)$; and (A4) $\pi(\beta) = I_{[0,3.5]}(\beta)$ and $\boldsymbol{x} = \boldsymbol{y}$. Models for the Eastern Meadowlark.*

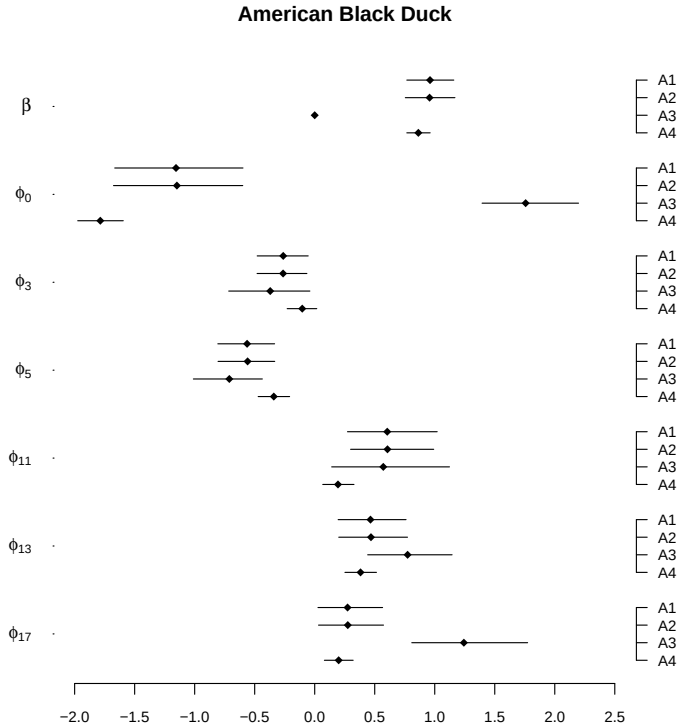

Figure 6.12: *95% credible intervals for β and ϕ_k $(k = 0, \ldots, q)$ with different assumptions: (A1) $\pi(\beta) = I_{[0,3.5]}(\beta)$; (A2) $\pi(\beta) = I_{[-3.5,3.5]}(\beta)$; (A3) $\pi(\beta) = I_{[-3.5,3.5]}(\beta)$; and (A4) $\pi(\beta) = I_{[0,3.5]}(\beta)$ and $\boldsymbol{x} = \boldsymbol{y}$. Models for the American Black Duck.*

Concluding remarks

The direct application of path sampling (PS) in simulation studies provided accurate approximations to the likelihood of the Markov random field (MRF); however, it is not computationally feasible. Furthermore, we obtained a poor performance of the predictions based on the PS. On the other hand, the ratio approximation performed very similar to the pseudo-likelihood (PL) approximation in terms of the size of the credible intervals and in the misclassification rate for the reconstruction of the true image. Nevertheless, the enormous difference in computer time makes the PL approximation more appealing when the MCMC requires long chains. The performance of the MCMC was more satisfactory when the original image was generated with more covariates than in the case with only one covariate, with a higher reduction of the misclassification rate.

The method by Kuo & Mallick (1998) (KM) for variable selection performed very well in the simulations. Although the frequency of the modal model was lower in the situations when we included a high number of covariates, as is the case of our application, the procedure selected the original covariates in most cases.

When we used real data, repetitions of the KM procedure for variable selection produced different models. We run multiple independent chains (MIC) to improve the mixing and obtained consistent models for both species. The mutation operator from EMC did not improve the results. The final model selected for the Eastern Meadowlark included a set of covariates with a reasonable ecological interpretation, while the model for the American Black Duck missed an important covariate. A second proposal for this species included that missing covariate, leading to a satisfactory set of interpretable covariates.

Although the logistic regression is widely used for modelling species distributions, it has the drawback that it does not take into account possible spatial dependence. In addition, it does not allow to recognise that the actual states of presence/absence of a bird species do not exactly correspond to the observed

map. The autologistic model provided a solution to account for the spatial dependence but it still did not resolve the problem of observation errors. Using a hierarchical approach made possible to incorporate the hidden image in the model and considerate observation errors.

We showed that there is a high probability of non-observed presence for the two species analysed in this thesis. Differences between observed and reconstructed images arose due to non-observed presences rather than false observations. The inclusion of effort hours was crucial when modelling observation errors.

Many new roads could be taken after what is presented in this thesis. In the methodological area, we could extend the study of pseudo-likelihood approximations and explore the possibility of building small blocks for which we can calculate the exact likelihood, conditioning on the border of the block. We calculated the conditional probability of a single site given the values of its neighbours; instead we could calculate this conditional probability for the block given the values of the neighbours of the whole block. We could also consider non-regular grids when more natural arrangements of data are available, without restricting the analysis to the type of data that we used here. There is also the relevant question of the level of the spatial dependence that affects the conditional probability of presence in the MRF. During our work we used a first order neighbourhood system, where every (interior) site had four neighbours (horizontally or vertically adjacent). We could extend the analysis to a second-order neighbourhood with eight neighbours for every (interior) site (additionally including diagonal adjacencies). We encountered a limitation when we analysed annual data, particularly for ducks that in the space of one year experience the full spectrum of seasonal changes. Thus, comparative analyses for different shorter periods during the year could give better information. In addition, trend studies in the spatial lattice could be performed since we have repeated measures over discrete time points, although these measures are not genuine repeated measures in the same locations and following the same conditions each time. This type of analysis can be motivated by the work presented by Zhu et al. (2008) on outbreaks of mountain pine beetle, and the concept of Markov chain of Markov field (MCMF) introduced by Guyon & Hardouin (2002) who define the MCMF on instantaneous interaction potentials and time-delay potentials.

Appendix A

Results

Proof of (3.21)

We want to prove (3.21):

$$C(\boldsymbol{\psi}_0) = \prod_{i \in \mathcal{S}} \left[1 + \exp(\boldsymbol{z}_i' \boldsymbol{\phi})\right] .$$

Proof. Let us use S_N instead of S to make evident the cardinality of S. We need to prove that

$$\sum_{\boldsymbol{x} \in \zeta_N} \exp\left(\sum_{i \in S_N} x_i \boldsymbol{z}_i' \boldsymbol{\phi}\right) = \prod_{i \in S_N} \left[1 + \exp(\boldsymbol{z}_i' \boldsymbol{\phi})\right] ,$$

where $\zeta_N = \{0, 1\}^N$.

We make the proof by induction. We notice that the expression is valid when $N = 1$ since

$$\sum_{\boldsymbol{x} \in \zeta_1} \exp\left(\sum_{i \in S_1} x_i \boldsymbol{z}_i' \boldsymbol{\phi}\right) = 1 + \exp(\boldsymbol{z}_1' \boldsymbol{\phi}) = \prod_{i \in S_1} \left[1 + \exp(\boldsymbol{z}_i' \boldsymbol{\phi})\right] .$$

Now we assume that the expression is valid when $N = n$ and prove its validity for $N = n + 1$. Thus, we assume that

$$\sum_{\boldsymbol{x} \in \zeta_n} \exp\left(\sum_{i \in S_n} x_i \boldsymbol{z}_i' \boldsymbol{\phi}\right) = \prod_{i \in S_n} \left[1 + \exp(\boldsymbol{z}_i' \boldsymbol{\phi})\right] .$$

We notice that the set ζ_{n+1} can be split into two subsets as $\zeta_{n+1} = \zeta_{n+1}^{(0)} \cup \zeta_{n+1}^{(1)}$,

with $\zeta_{n+1}^{(k)} = \{\boldsymbol{x} : \boldsymbol{x} \in \zeta_{n+1},\, x_{n+1} = k\}$ $(k = 0, 1)$. Now,

$$\begin{aligned}
\sum_{\boldsymbol{x}\in\zeta_{n+1}} \exp\left(\sum_{i\in S_{n+1}} x_i \boldsymbol{z}_i'\boldsymbol{\phi}\right) &= \sum_{\boldsymbol{x}\in\zeta_{n+1}^{(0)}} \exp\left(\sum_{i\in S_{n+1}} x_i \boldsymbol{z}_i'\boldsymbol{\phi}\right) + \\
&\quad \sum_{\boldsymbol{x}\in\zeta_{n+1}^{(1)}} \exp\left(\sum_{i\in S_{n+1}} x_i \boldsymbol{z}_i'\boldsymbol{\phi}\right) \\
&= \sum_{\boldsymbol{x}\in\zeta_{n}} \exp\left(\sum_{i\in S_{n}} x_i \boldsymbol{z}_i'\boldsymbol{\phi}\right) + \\
&\quad \sum_{\boldsymbol{x}\in\zeta_{n}} \exp\left(\sum_{i\in S_{n}} x_i \boldsymbol{z}_i'\boldsymbol{\phi} + \boldsymbol{z}_{n+1}'\boldsymbol{\phi}\right) \\
&= \sum_{\boldsymbol{x}\in\zeta_{n}} \exp\left(\sum_{i\in S_{n}} x_i \boldsymbol{z}_i'\boldsymbol{\phi}\right) [1 + \exp(\boldsymbol{z}_{n+1}'\boldsymbol{\phi})] \\
&= \prod_{i\in S_n} [1 + \exp(\boldsymbol{z}_i'\boldsymbol{\phi})]\, [1 + \exp(\boldsymbol{z}_{n+1}'\boldsymbol{\phi})] \\
&= \prod_{i\in S_{n+1}} [1 + \exp(\boldsymbol{z}_i'\boldsymbol{\phi})]\ .
\end{aligned}$$

Number of iterations to reach stationarity

We start from the fact that in the stationary distribution, every cell of any sequence has a fix probability of taking the value 1. We analyse the impact of the number of iterations on the generation of a sequence $\boldsymbol{x}$. For a given vector of parameters $\boldsymbol{\psi}$ and values of two selected covariates from the data set, we generate iteratively sequences $\boldsymbol{x}$ using Gibbs sampler, and estimate the probability of a 1 for each cell. At each iteration the estimated probability is calculated as the average of the values obtained in that cell during the iterations performed up to that one. In Fig. A.1 we plot the probabilities for 20 cells randomly selected and observe that they stabilise at around 200 iterations.

Number of iterations for convergence of $\log C(\boldsymbol{\psi}_1)$

We select 3 covariates from the data set, fix the parameter vector $\boldsymbol{\psi}^*$ and construct the grid Ω around $\boldsymbol{\psi}^*$ which has 65 points. For each point we apply the MCMC algorithm and plot the ergodic mean of $\log C(\boldsymbol{\psi}_1)$. Here we present the plots for 9 of these points (Fig. A.2) with less stable patterns and observe that they converge at around 200 to 400 iterations.

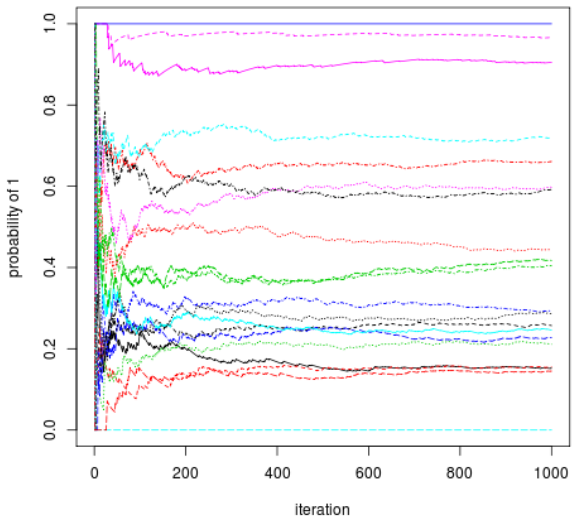

Figure A.1: *Convergence of the probability of presence for 20 cells randomly selected (fix parameter ψ).*

Transition distribution for a parameter λ with Gaussian noise

Given the value of λ at the $(t-1)^{th}$ iteration, $\lambda^{(t-1)}$, we want to find the conditional distribution of λ, $q(\lambda|\lambda^{(t-1)})$, that follows the random walk

$$h(\lambda) = h(\lambda^{(t-1)}) + \epsilon\,;\ \epsilon \sim \mathcal{N}(0, s)\,,$$

for a known monotone function $h(.)$. The distribution from where we generate values for a new λ, $q(\lambda|\lambda^{(t-1)})$, is called the transition distribution. Since we know the function $h(.)$ and the value $\lambda^{(t-1)}$, we have completely specified the distribution of $h(\lambda)$ which is also normal with mean $h(\lambda^{(t-1)})$ and variance of the noise s, i.e.:

$$f(h(\lambda)) = (2\pi s)^{-1/2} \exp\left[-(2s)^{-1}(h(\lambda) - h(\lambda^{(t-1)}))^2\right]\,.$$

Let $x = h(\lambda)$, $y = h^{-1}(x) = \lambda$, and $g(.) = h^{-1}(.)$, thus $y = g(x)$ and $x = g^{-1}(y)$. We get the distribution of the transformation of the random variable Y:

$$f_Y(y) = f_X(g^{-1}(y)) \left|\frac{\partial X}{\partial Y}\right|$$
$$\Rightarrow f(\lambda) = f(h(\lambda)) \left|\frac{\partial h(\lambda)}{\partial \lambda}\right|\,.$$

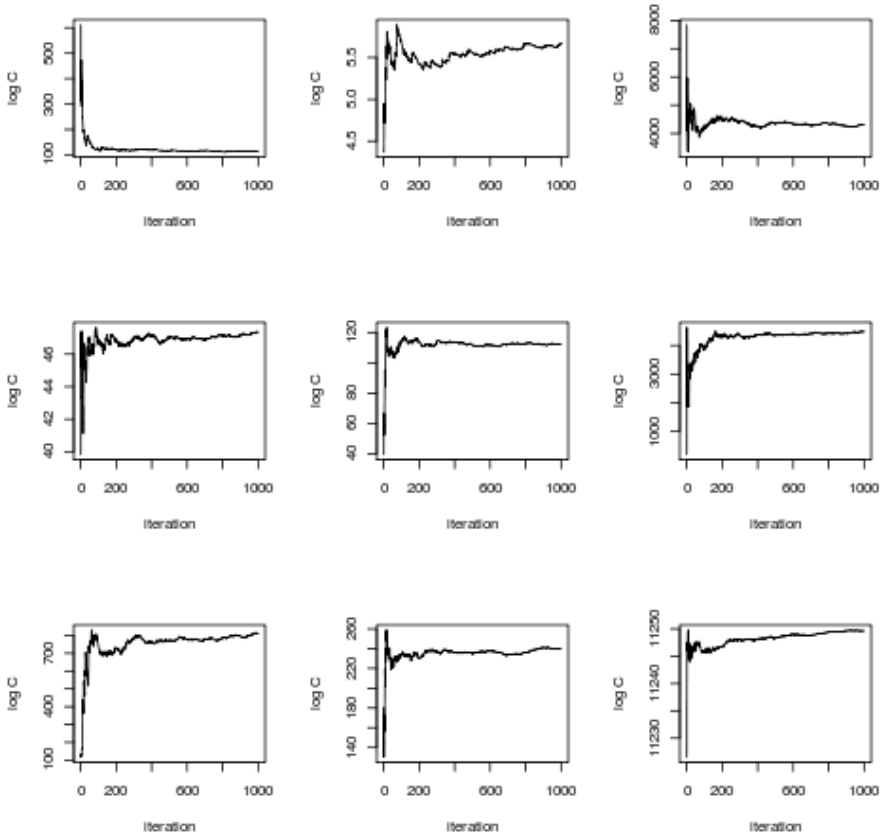

Figure A.2: *Convergence of* $\log C(\boldsymbol{\psi}_1)$ *for 9 points in* Ω *(fix parameter* $\boldsymbol{\psi}^*$*).*

Since $|\partial h(\lambda)/\partial\lambda|$ is a function of λ we call it J_λ. We substitute to get the conditional distribution of λ given $\lambda^{(t-1)}$:

$$q(\lambda|\lambda^{(t-1)}) = f(h(\lambda))\, J_\lambda = (2\pi s)^{-1/2} \exp\left[-(2s)^{-1}(h(\lambda) - h(\lambda^{(t-1)}))^2\right] J_\lambda\,.$$

Similarly we can express the distribution of $\lambda^{(t-1)}$ for a given λ:

$$\begin{aligned} q(\lambda^{(t-1)}|\lambda) &= f(h(\lambda^{(t-1)}))\, J_\lambda^{(t-1)} \\ &= (2\pi s)^{-1/2} \exp\left[-(2s)^{-1}(h(\lambda^{(t-1)}) - h(\lambda))^2\right] J_\lambda^{(t-1)}\,. \end{aligned}$$

The ratio of the two previous distributions is called the proposal ratio:

$$\frac{q(\lambda^{(t-1)}|\lambda)}{q(\lambda|\lambda^{(t-1)})} = \frac{(2\pi s)^{-1/2} \exp\left[-(2s)^{-1}(h(\lambda^{(t-1)}) - h(\lambda))^2\right] J_\lambda^{(t-1)}}{(2\pi s)^{-1/2} \exp\left[-(2s)^{-1}(h(\lambda) - h(\lambda^{(t-1)}))^2\right] J_\lambda} = \frac{J_\lambda^{(t-1)}}{J_\lambda}\,.$$

The last simplification is due to the symmetry of the random noise.

Appendix B

Figures from Chapter 5

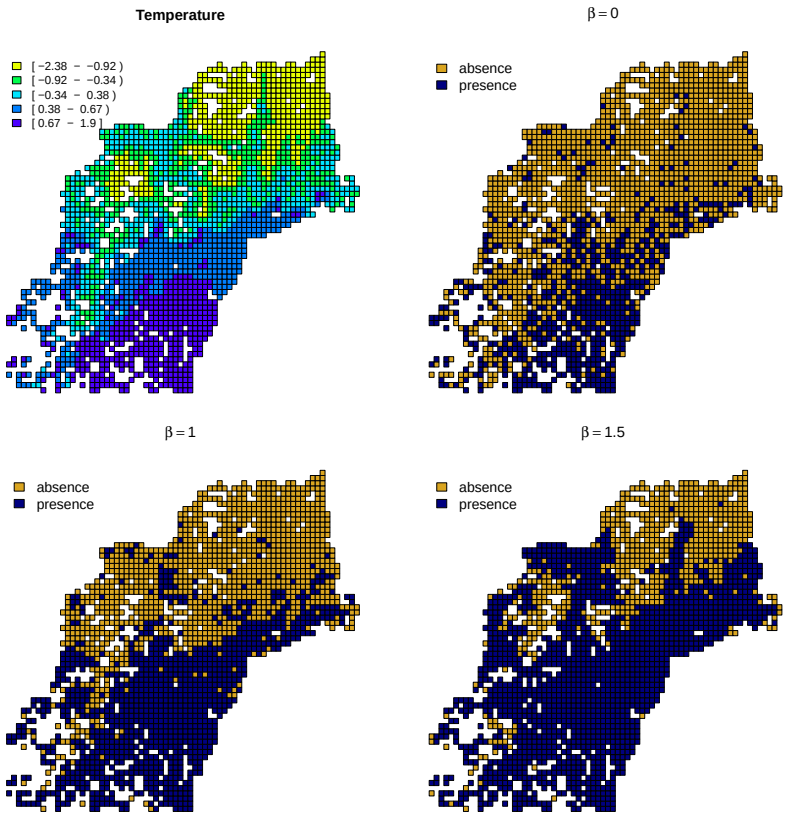

Figure B.1: *Images generated on a grid of* $N = 2195$ *values using the covariate Average temperature with* $\boldsymbol{\phi} = (-1.5, 2)'$*, and* $\beta = 0$ *(top right),* $\beta = 1$ *(bottom left),* $\beta = 1.5$ *(bottom right).*

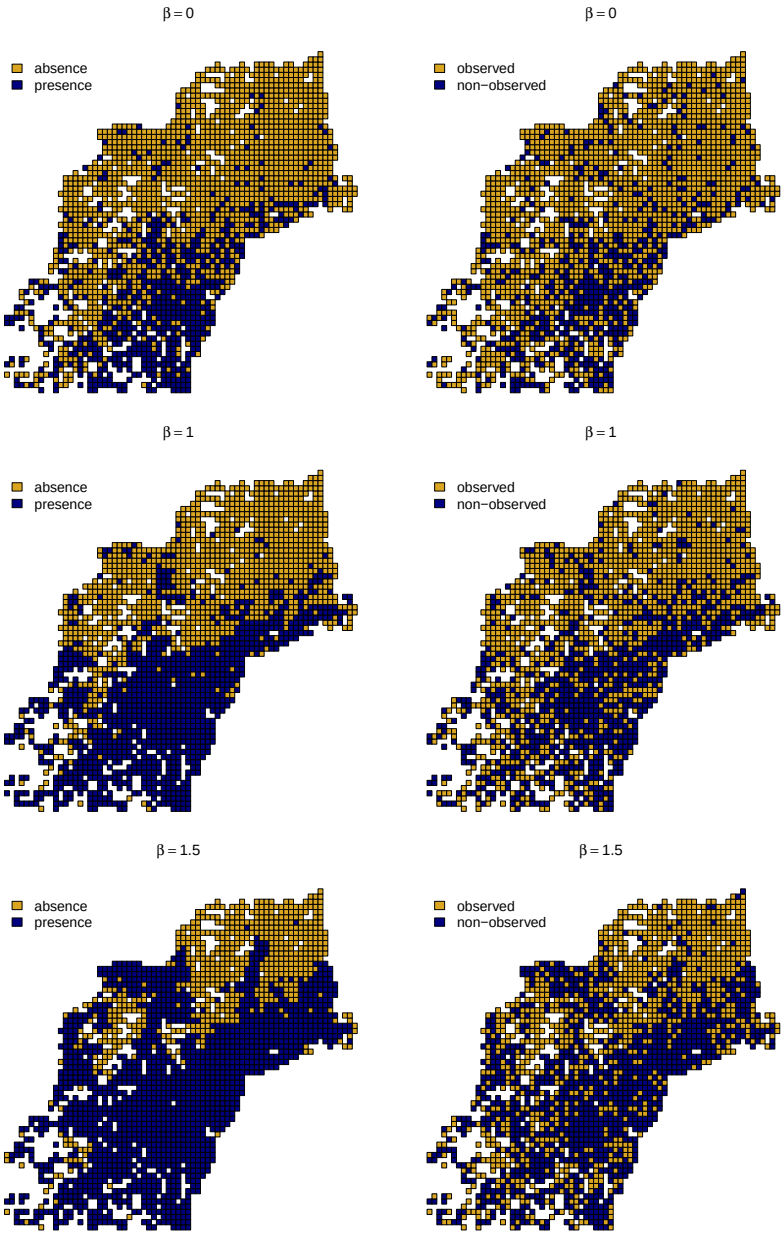

Figure B.2: *Observed maps (right) obtained by disturbing the true maps (left) with* $\alpha = (-2.5, -0.5, 1.2, 2)'$. *True maps are generated with Average temperature.*

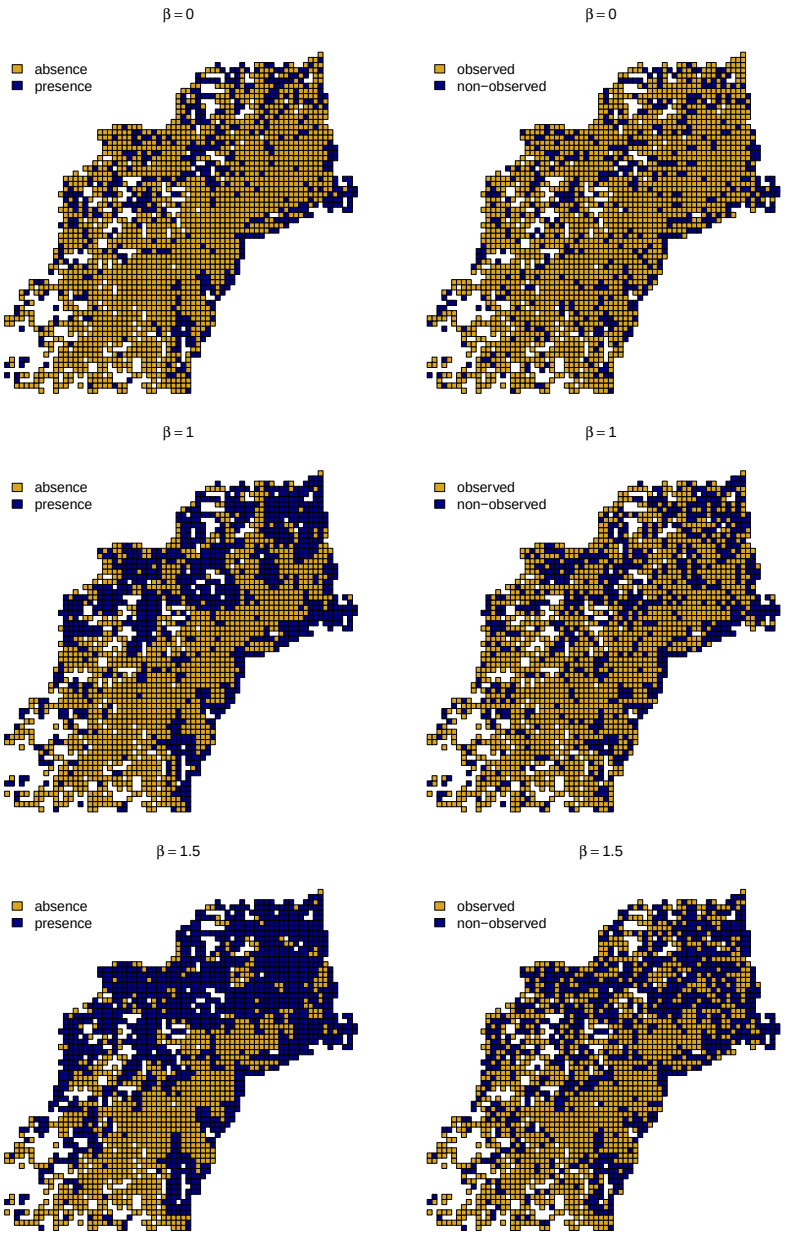

Figure B.3: *Observed maps (right) obtained by disturbing the true maps (left) with* $\alpha = (-2.5, -0.5, 1.2, 2)'$*. True maps are generated with Average temperature, Open water, Distance from standing fresh water, and Grassland/herbaceous/pasture/crops, with* $\boldsymbol{\phi} = (-2, -1.5, 2, 2, 1)'$*, and different values of* β*.*

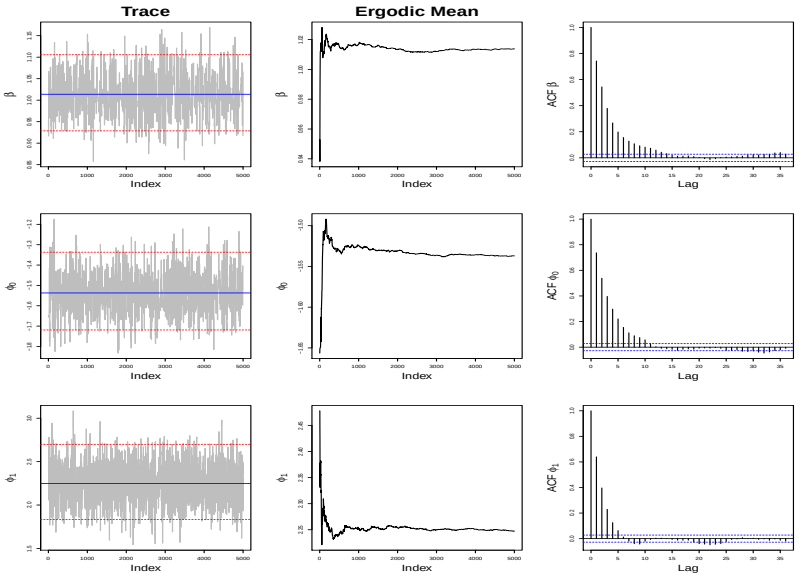

Figure B.4: *Diagnostic plots for β, ϕ_0, ϕ_1 with PL approximation.*

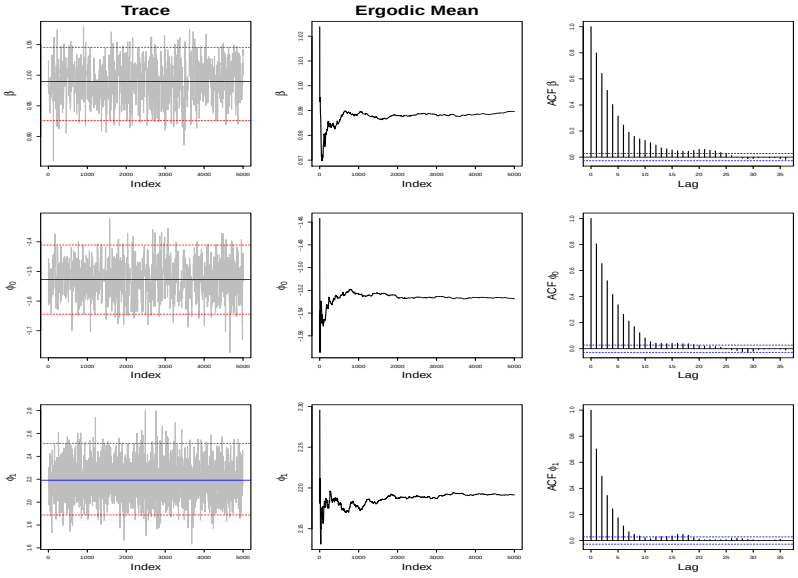

Figure B.5: *Diagnostic plots for β, ϕ_0, ϕ_1 with ratio approximation.*

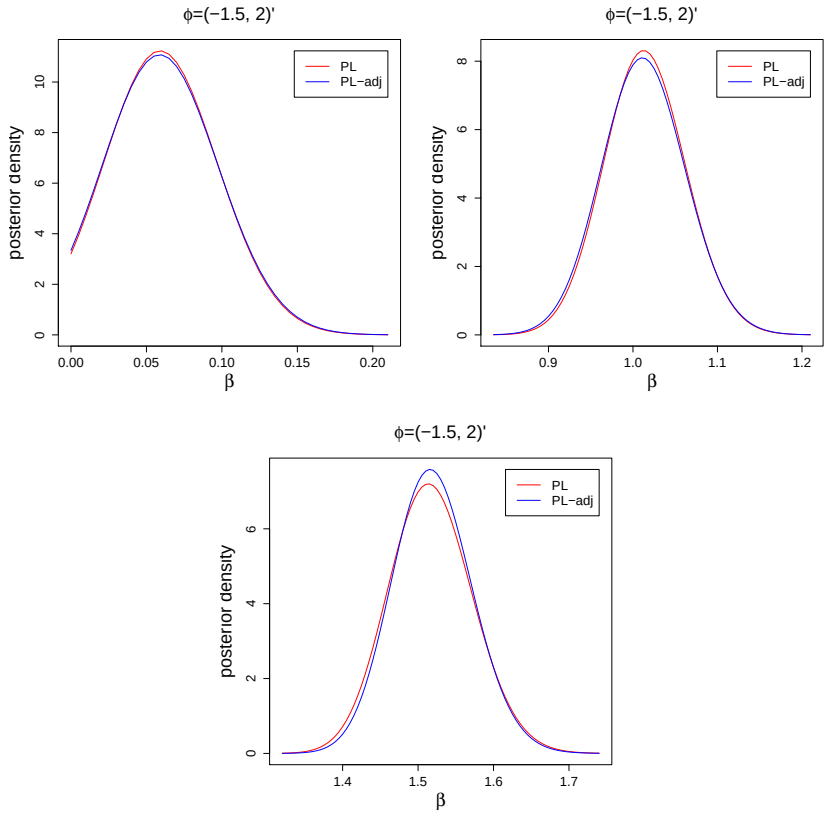

Figure B.6: *Posterior distribution of β using the PL approximation for the likelihood and the adjusted PL, and keeping ϕ fixed at the true value.*

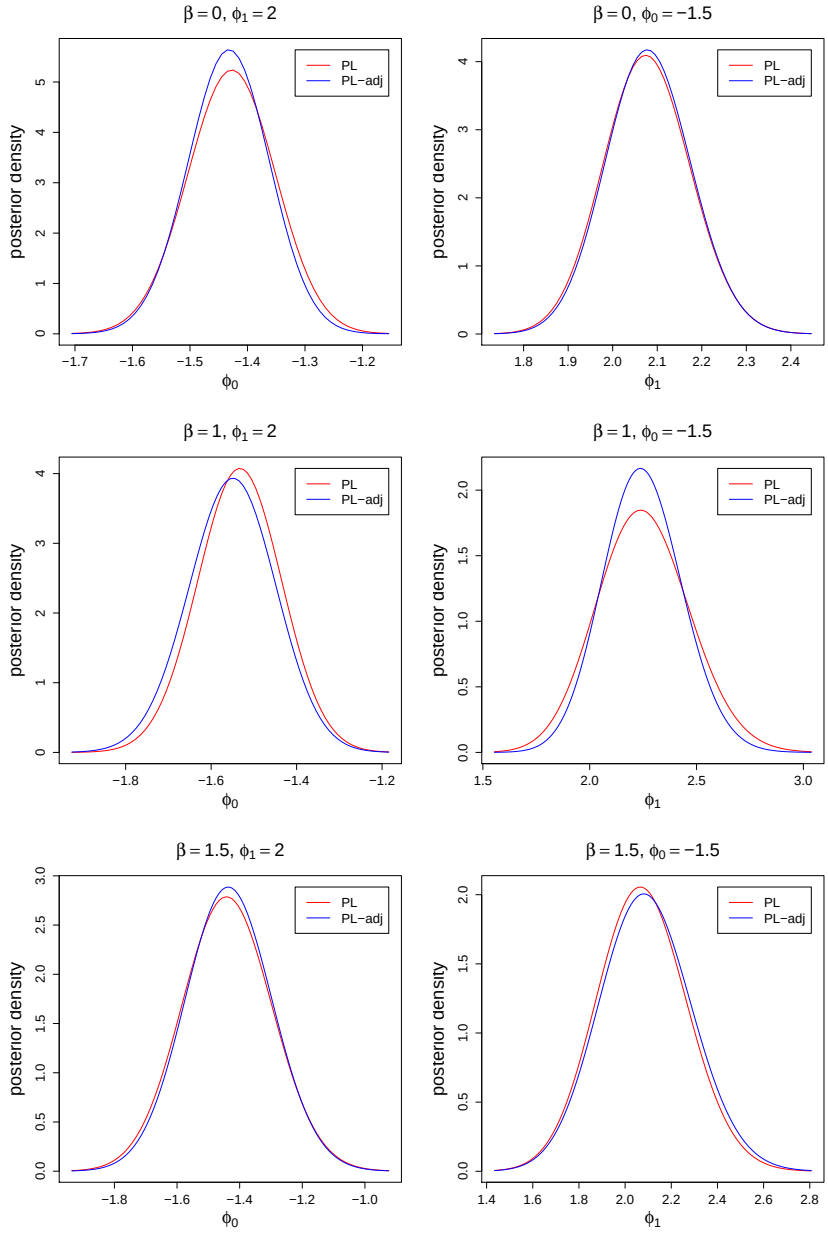

Figure B.7: *Posterior distribution of ϕ_0 (left) and ϕ_1 (left) using the PL approximation for the likelihood and the adjusted PL, and keeping the other parameters fixed at their true values.*

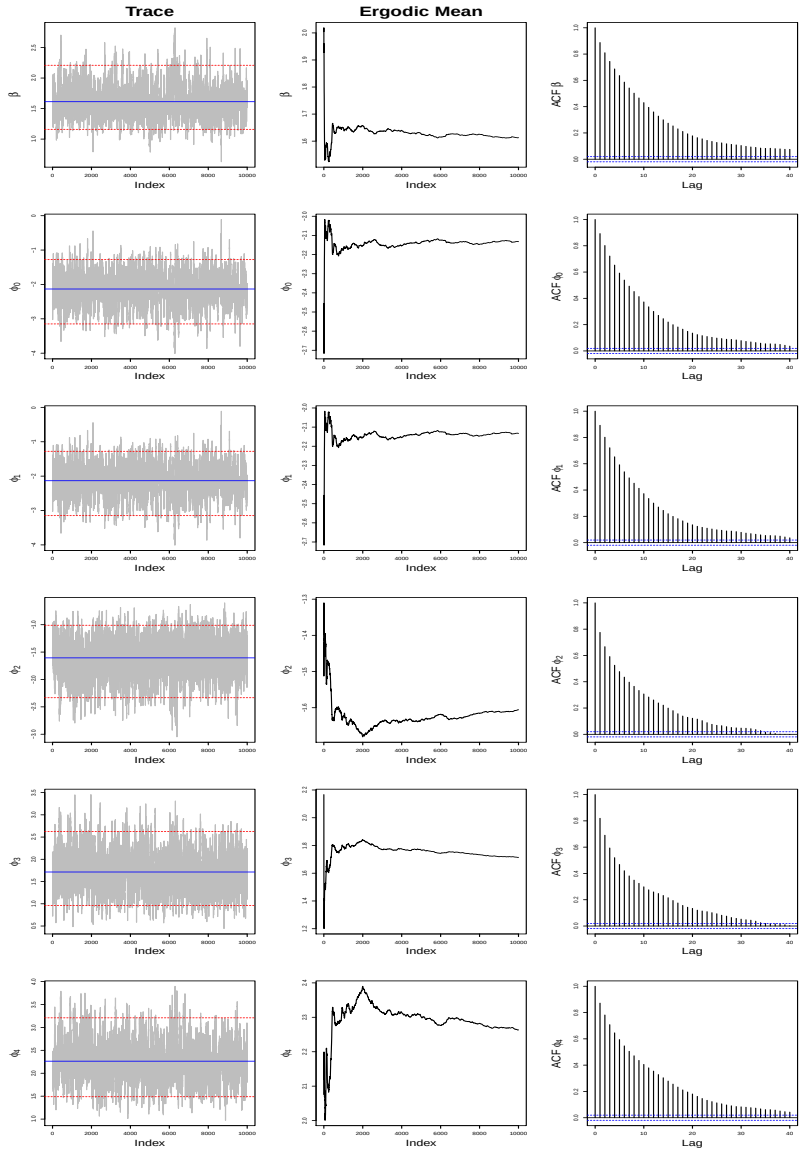

Figure B.8: *Diagnostic plots for β and ϕ_k $(k = 0, \ldots, 4)$ with a simulated map using PL approximation.*

Appendix C

Figures from Chapter 6

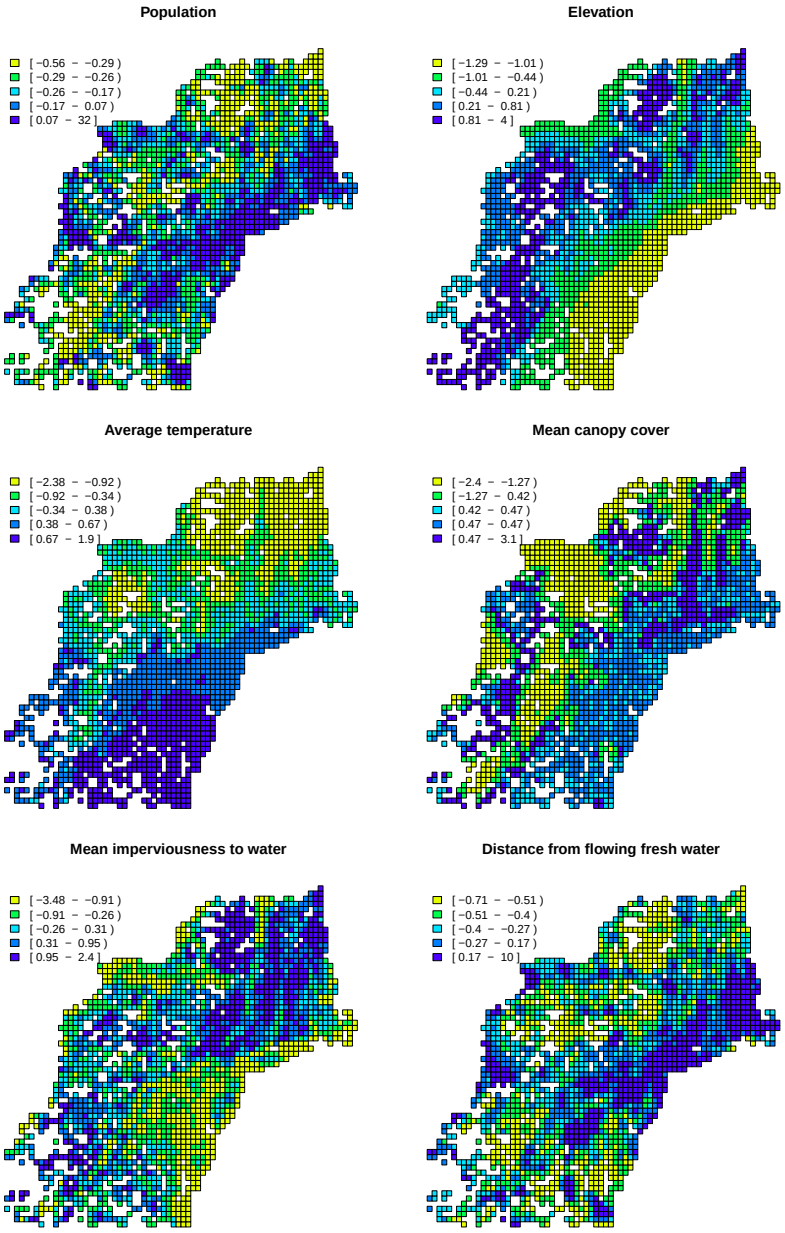

Figure C.1: *Two-dimensional representation of selected covariates that are categorised according to quantiles.*

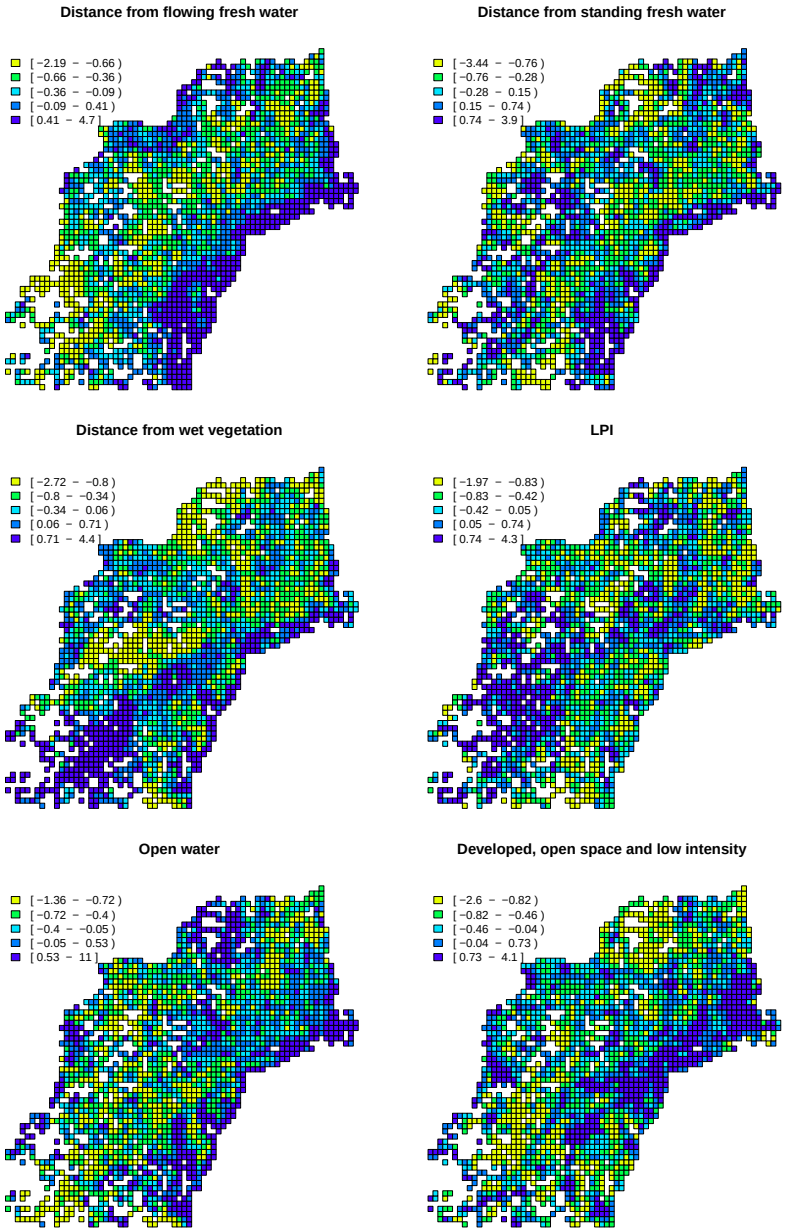

Figure C.2: *Two-dimensional representation of selected covariates that are categorised according to quantiles.*

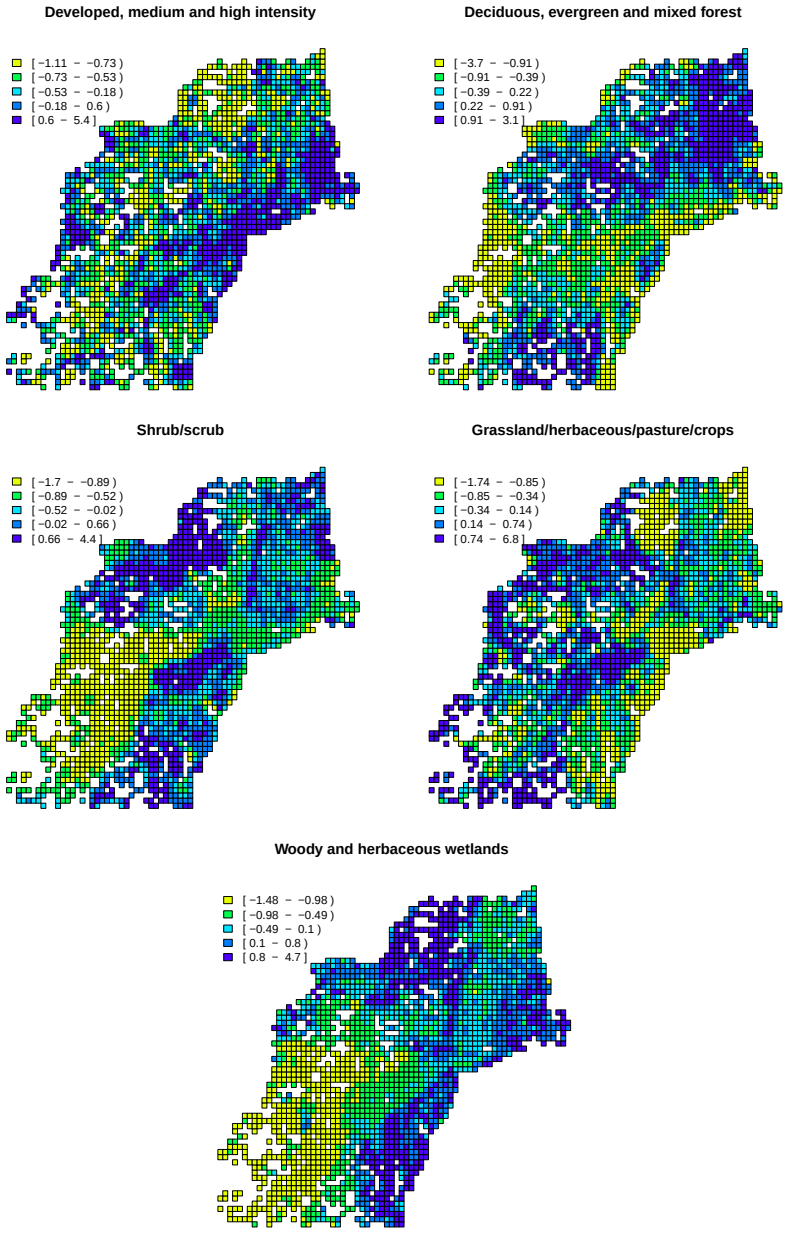

Figure C.3: *Two-dimensional representation of selected covariates that are categorised according to quantiles.*

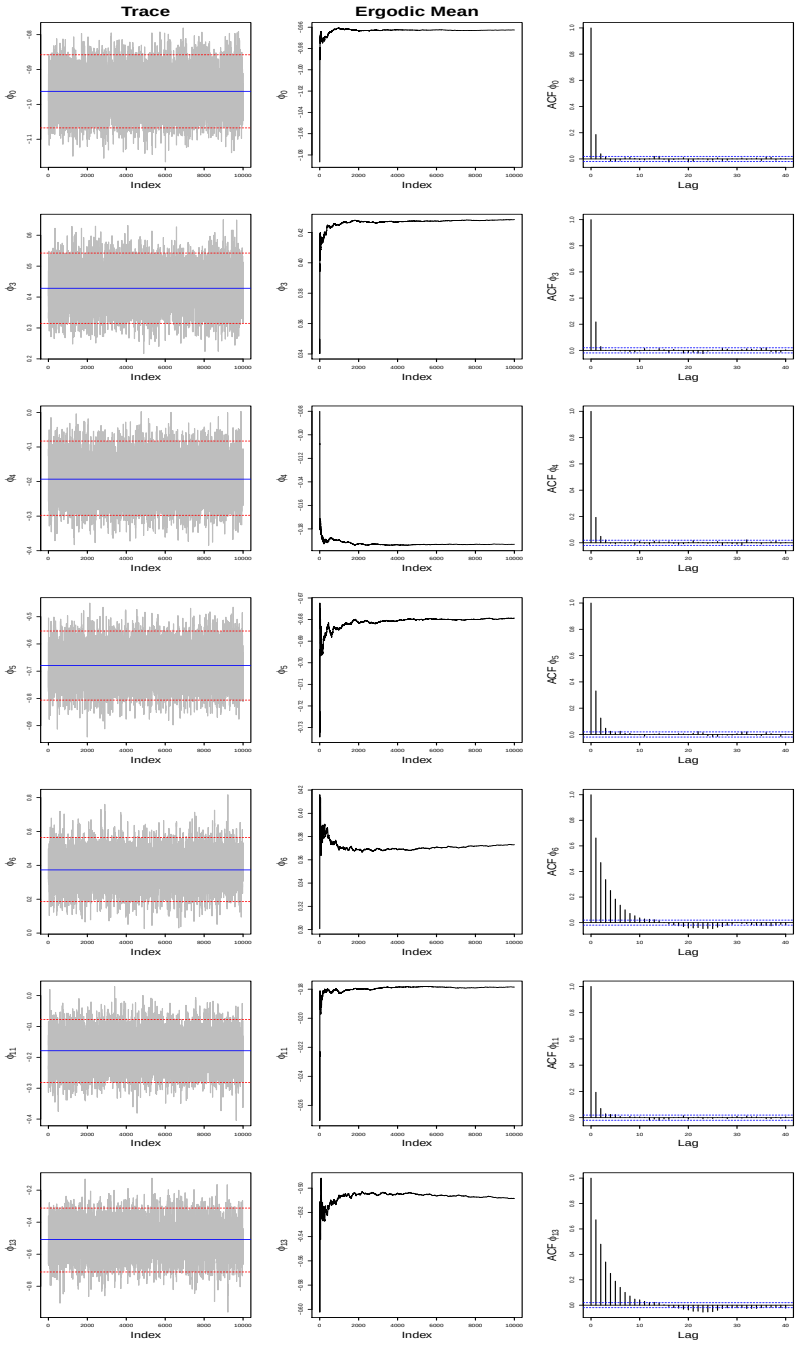

Figure C.4: *Diagnostic plots for* $\phi_0, \phi_3, \phi_4, \phi_5, \phi_6, \phi_{11}, \phi_{13}$ *with the logistic model for the Eastern Meadowlark.*

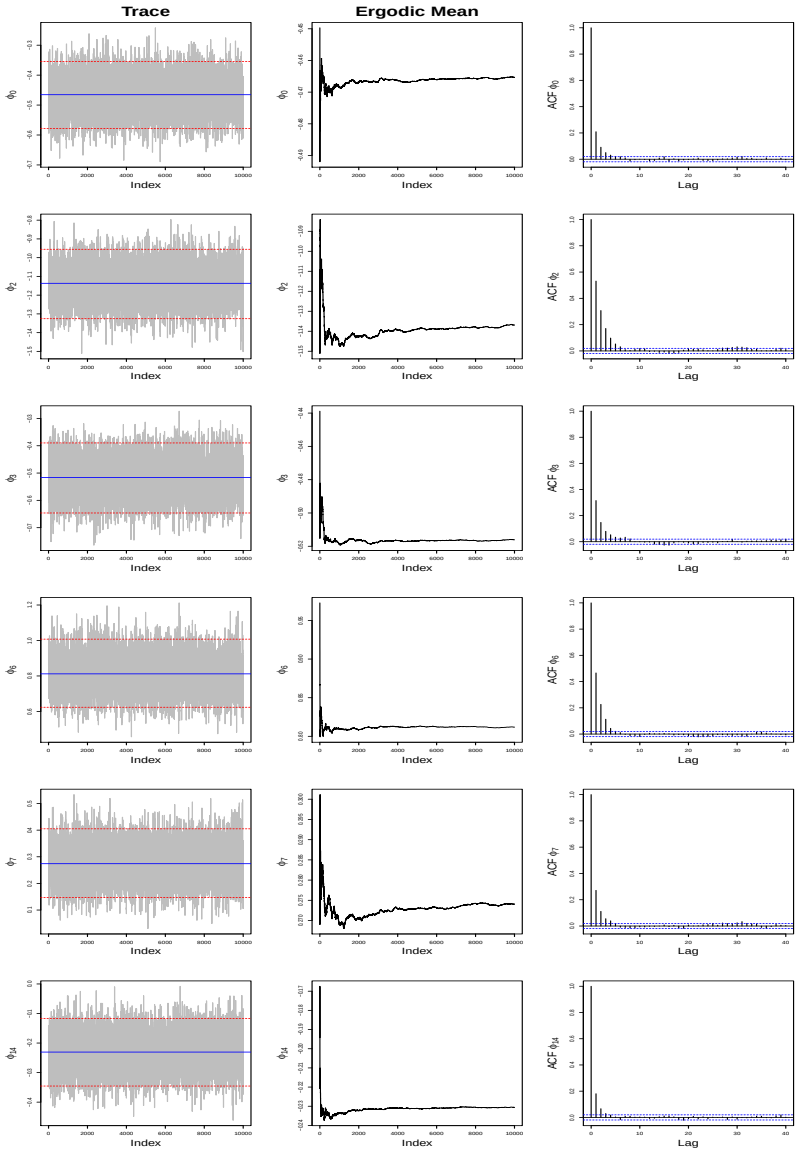

Figure C.5: *Diagnostic plots for $\phi_0, \phi_2, \phi_3, \phi_6, \phi_7, \phi_{14}$ with the logistic model for the American Black Duck.*

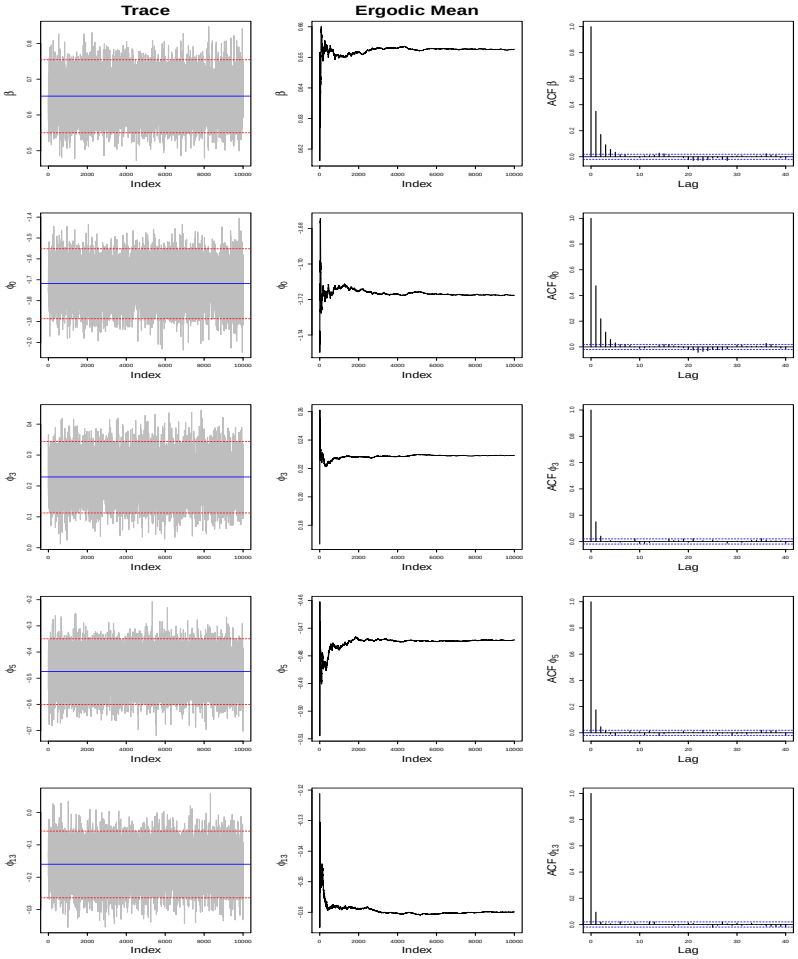

Figure C.6: *Diagnostic plots for β, $\phi_0, \phi_3, \phi_5, \phi_{13}$ with the autologistic model for the Eastern Meadowlark.*

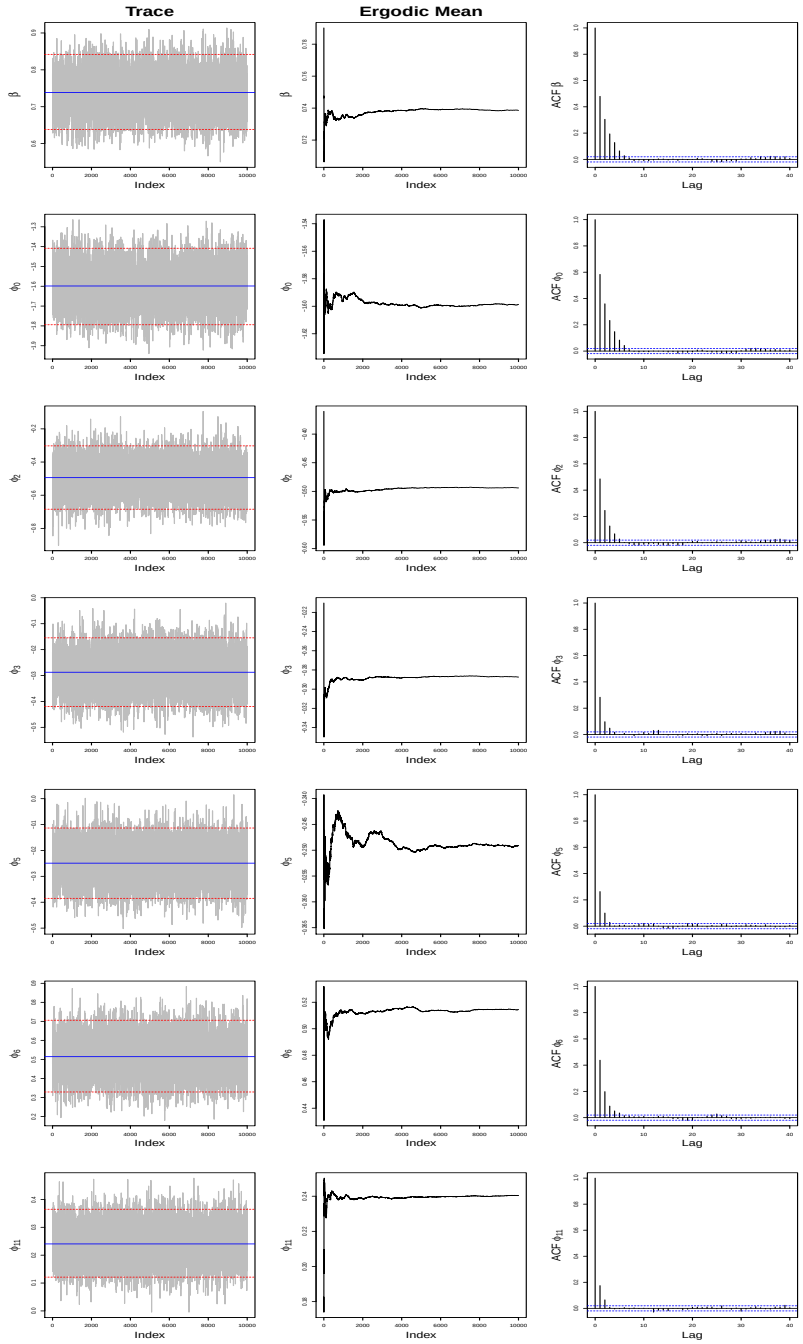

Figure C.7: *Diagnostic plots for β, $\phi_0, \phi_2, \phi_3, \phi_5, \phi_6, \phi_{11}$ with the autologistic model for the American Black Duck.*

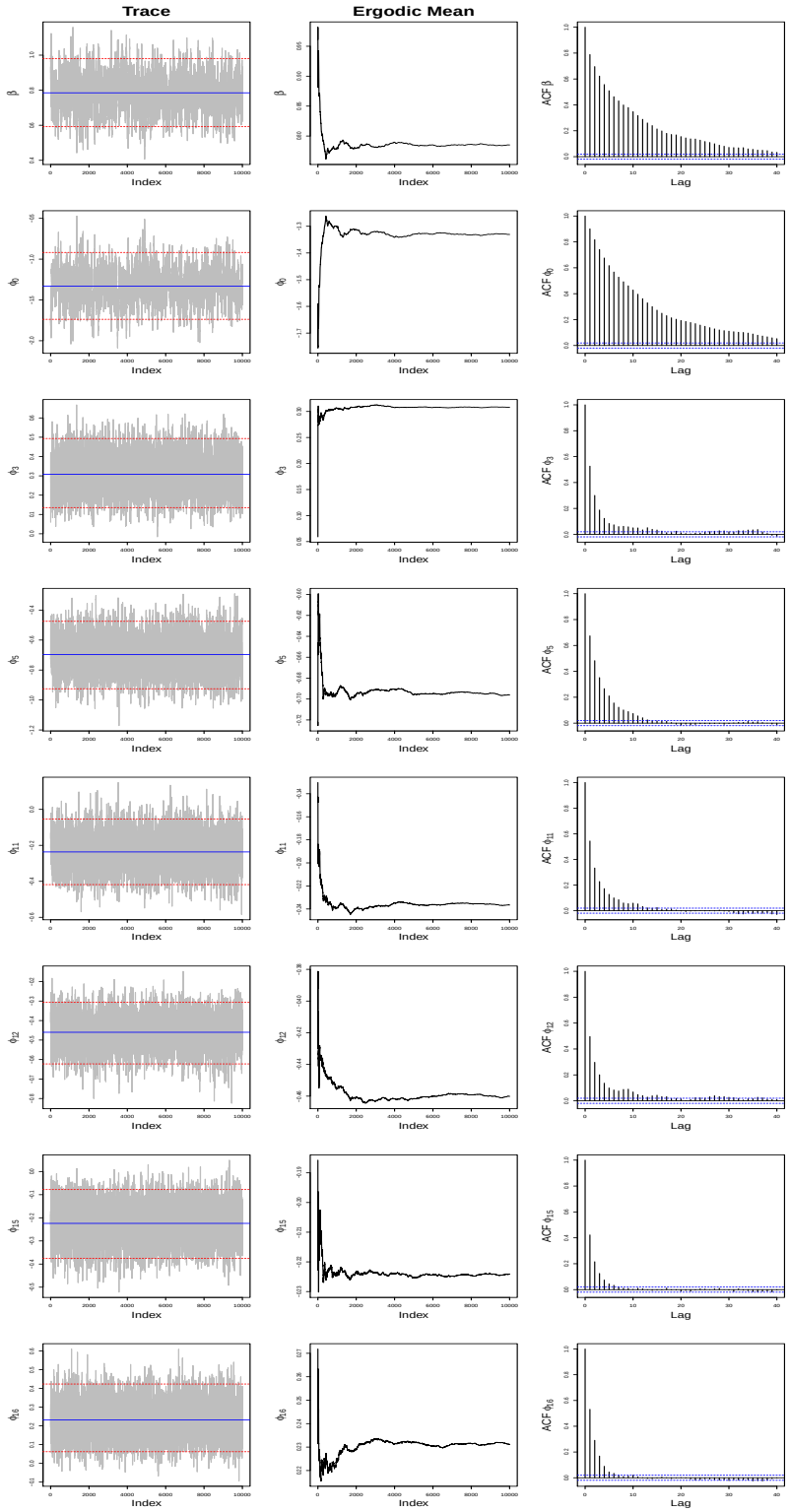

Figure C.8: *Diagnostic plots for* $\beta, \phi_0, \phi_3, \phi_5, \phi_{11}, \phi_{12}, \phi_{15}, \phi_{16}$ *with the SHMM for the Eastern Meadowlark.*

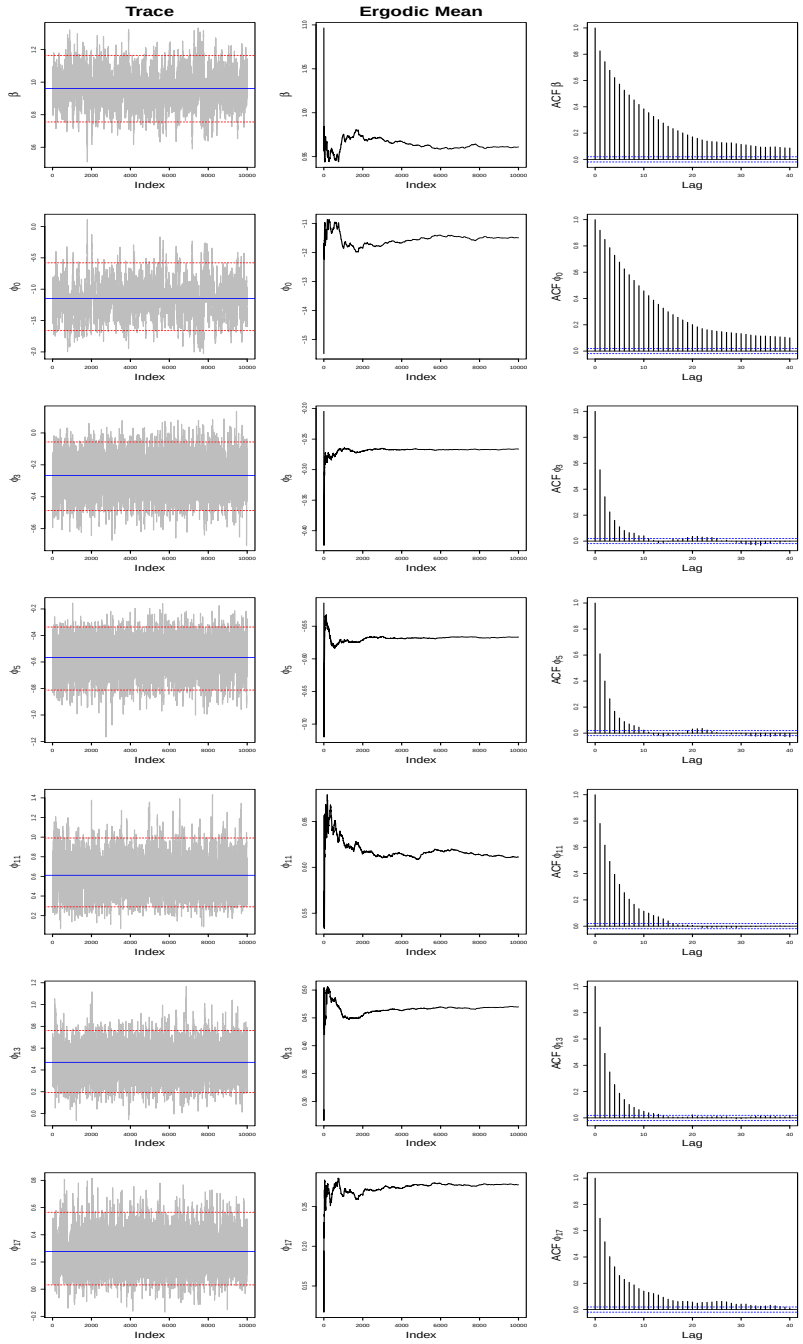

Figure C.9: *Diagnostic plots for* $\beta, \phi_0, \phi_3, \phi_5, \phi_{11}, \phi_{13}, \phi_{17}$ *with the SHMM for the American Black Duck.*

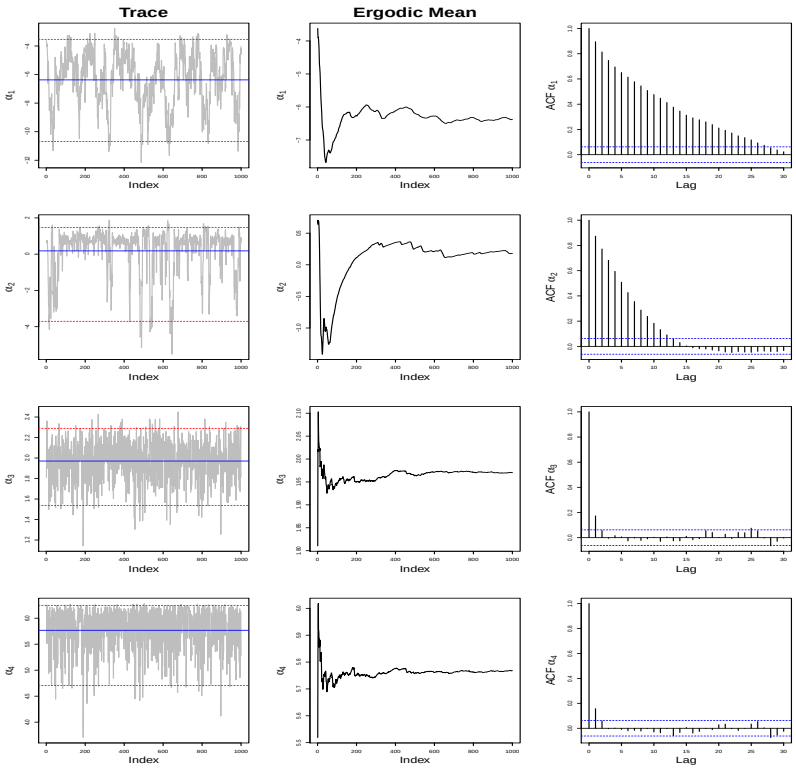

Figure C.10: *Diagnostic plots for $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ with the SHMM for the Eastern Meadowlark.*

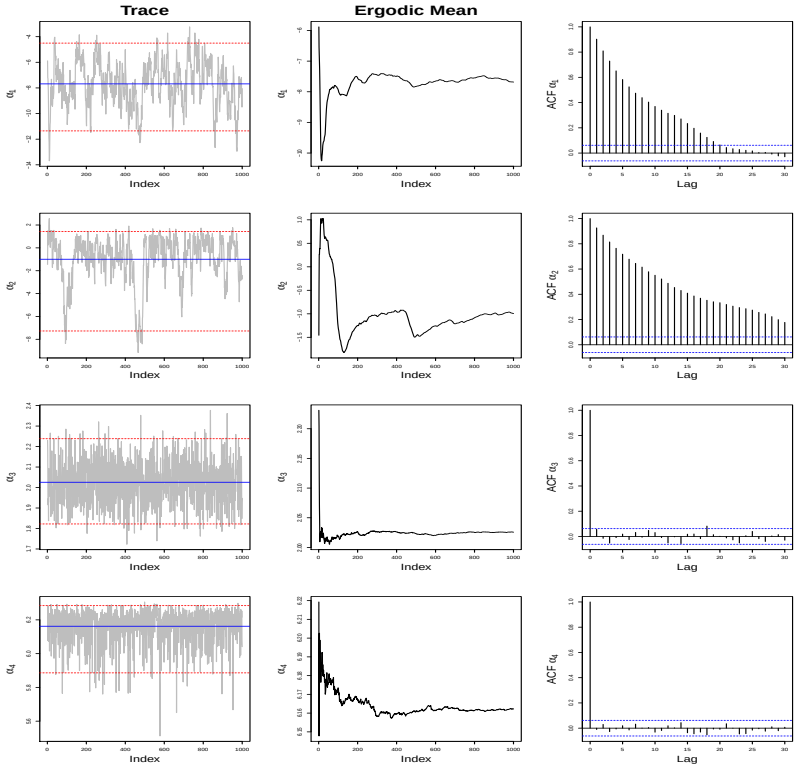

Figure C.11: *Diagnostic plots for $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ with the SHMM for the American Black Duck.*

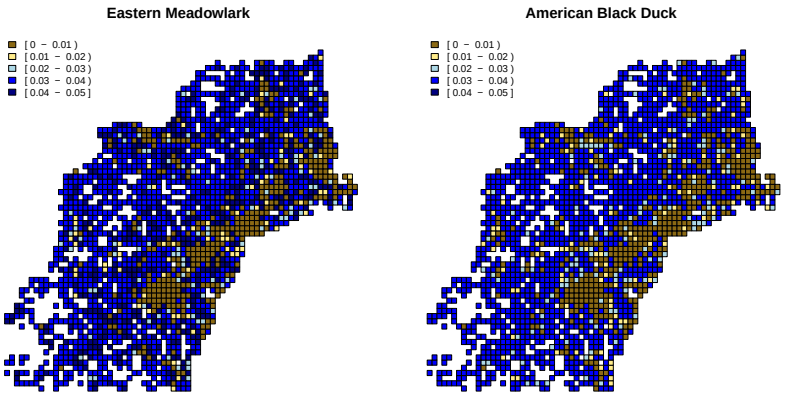

Figure C.12: *Uncertainty (inter-quartile range) for the probabilities of non-observed presence from the SHMM for the Eastern Meadowlark and the American Black Duck.*

Bibliography

AUGUSTIN, N. H., MUGGLESTONE, M. A. & BUCKLAND, S. T. (1996). An autologistic model for the spatial distribution of wildlife. *Journal of Applied Ecology* 36 339–347.

BANERJEE, S., CARLIN, B. P. & GELFAND, A. E. (2003). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC.

BARTOLUCCI, F. & BESAG, J. E. (2002). A recursive algorithm for Markov random fields. *Biometrika* 89 724–730.

BENT, A. C. (1989). *Life histories of North American flycatchers, larks, swallows, and their allies*. Dove Publications.

BESAG, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Ser B* 36 192–236.

BESAG, J. E. (1975). Statistical analysis of non-lattice data. *The Statistician* 24 179–195.

BESAG, J. E. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society, Ser B* 48 259–302.

BESAG, J. E., YORK, J. & MOLLIE, A. (1991). Bayesian image restauration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* 43 1–59.

BOLAND, T. (2007). The Internet Bird Collection. `http://ibc.lynxeds.com/photo/american-black-duck-anas-rubripes/male-preening-showing-speculum`. Accessed: 13/11/2012.

CLEGG, J. & MANSELL, E. (1986). *Observer's Book of Pond Life*. Frederick Warne.

COOLEY, D., DAVISON, A. C. & RIBARET, M. (2012). Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statistica Sinica* 22 813–845.

COSEWIC (2011). COSEWIC assessment and status report on the Eastern Meadowlark Sturnella magna in Canada. Committee on the status of endangered wildlife in Canada. http://publications.gc.ca/collections/collection_2012/ec/CW69-14-624-2011-eng.pdf. Accessed: 07/11/2012.

CRESPI, C. M. & BOSCARDIN, W. J. (2009). Bayesian model checking for multivariate outcome data. *Computational Statistics and Data Analysis* 53 3765–3772.

CRESSIE, N. (1993). *Statistics for Spatial Data*. Wiley, New York.

DE LEON, L. (2010). Retiring with Lisa de Leon: retiring to birdwatching and photography. http://retiringwithlisadeleon.blogspot.it/2010/10/black-duck.html. Accessed: 30/12/2012.

DELLAPORTAS, P., FORSTER, J. J. & NTZOUFRAS, I. (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing* 12 27–36.

DORMANN, C. F. (2007). Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global Ecology and Biogeography* 16 129–138.

DORMANN, C. F., MCPHERSON, J. M., ARAUJO, M. B., BIVAND, R., BOLLIGER, J., CARL, G., DAVIES, R. G., HIRZEL, A., JETZ, W., KISSLING, W. D., KUHN, I., OHLEMULLER, R., PERES-NETO, P. R., REINEKING, B., SHRODER, B., SHURR, F. M. & WILSON, R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography* 30 609–628.

DRUGAN, M. M. & THIERENS, D. (2010). Recombination operators and selection strategies for evolutionary Markov chain Monte Carlo algorithms. *Evolutionary Inteligence* 3 79–101.

DRYDEN, I. L., SCARR, M. R. & TAYLOR, C. C. (2003). Bayesian texture segmentation of weed and crop images using reversible jump Markov chain Monte Carlo methods. *Applied Statistics* 52 31–50.

ELLIOTT, P., WAKEFIELD, J. C., BEST, N. G. & BRIGGS, D. J. (2000). *Spatial Epidemiology: Methods and Applications*. Oxford University Press.

FRANCQ, G. E. (1972). *Parental care of the Eastern Meadowlark (Sturnella magna)*. Master's thesis, Kansas State Teachers College, Emporia, Kansas.

FRIEL, N., PETTITT, A. N., REEVES, R. & WIT, E. (2009). Bayesian inference in hidden Markov random fields for binary data defined on large lattices. *Journal of Computational and Graphical Statistics* 18 243–261.

FRÜHWIRTH-SCHNATTER (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statisistical Association* 96 194–209.

FRY, J., XIAN, G., JIN, J., S. ADN DEWITZ, HOMER, C., YANG, L., BARNES, C., HEROLD, N. & WICKHAM, J. (2011). Completion of the 2006 National Land Cover Database for the Conterminous United States. *Photogrammetric Engineering and Remote Sensing* 77 858–864.

FURRER, R., NYCHKA, D. & SAIN, S. (2012). fields: Tools for spatial data. R package version 6.6.3. `http://CRAN.R-project.org/package=fields`. Accessed: 07/10/2012.

GASTON, K. J. (1994). *Rarity*. Springer.

GELMAN, A. (2002). Posterior distribution. *Encyclopedia of Environmetrics* 3 1627–1628.

GELMAN, A., CARLIN, J., STERN, H. & RUBIN, D. (2003). *Bayesian Data Analysis*. Chapman & Hall/CRC.

GELMAN, A. & MENG, X. L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science* 13 163–185.

GELMAN, A., MENG, X. L. & STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* 6 733–807.

GEORGE, E. L. & MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statisistical Association* 88 881–889.

GEORGE, E. L. & MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statisistica Sinica* 7 339–373.

GEYER, C. J. & THOMPSON, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society, Ser. B* 54 657–699.

GODSILL, S. J. (2001). On the relationship between MCMC model uncertainty methods. *Journal of Computational and Graphical Statistics* 10 230–248.

GREEN, P. (1995a). MCMC in image analysis. *Markov chain Monte Carlo in practice* 381–399.

GREEN, P. J. (1995b). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82 711–732.

GREEN, P. J. & RICHARDSON, S. (2002). Hidden Markov models and disease mapping. *Journal of the American Statisistical Association* 97 1055–1070.

GU, M. G. & ZHU, H. T. (2001). Maximum likelihood estimation for spatial models by Markov chain Monte Carlo stochastic approximation. *Journal of the Royal Statistical Society, Ser. B* 63 339–355.

GUMPERTZ, M., WU, C.-T. & PYE, J. (2000). Logistic regression for southern pine beetle outbreaks with spatial and temporal autocorrelation. *Forest Science* 46 95–107.

GUYON, X. & HARDOUIN, C. (2002). Markov chain markov field dynamics: models and statistics. *Statistics: A Journal of Theoretical and Applied Statistics* 36 339–363.

HAINING, R. (2003). *Spatial Data Analysis - Theory and Practice*. Cambridge University Press.

HARDOUIN, C. & GUYON, X. (2009). Exact marginals and normalizing constant for gibbs distributions. *C. R. Acad. Sci. Paris, Ser. I* 348 199–201.

HASTIE, T. J. & TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. Chapman & Hall/CRC.

HEAGERTY, P. J. & LUMLEY, T. (2000). Window subsampling of estimating functions with application to regression models. *Journal of the American Statisistical Association* 95 197–211.

HEIKKINEN, J. & HÖGMANDER, H. (1994). Fully Bayesian approach to image restoration with an application in Biogeography. *Applied Statistics* 43 569–582.

HEIKKINEN, J. & HÖGMANDER, H. (1997). Statistics in Biogeography. In: Kotz, S., Read, C. B. & Banks, D. L. (eds.). Encyclopedia of Statistical Sciences, Update volume 1. John Wiley & Sons, Inc., New York. p. 56-61.

HERKERT, J. R. (1991). Prairie birds of Illinois: population response to two centuries of habitat change. *Illinois Natural History Survey Bulletin* 34 393–399.

HERKERT, J. R. (1994). The effects of habitat fragmentation on midwestern grassland bird communities. *Ecological Applications* 4 461–471.

HÖGMANDER, H. & MØLLER, J. (1993). Classification of atlas maps using statistical methods of image analysis. Research Report 263. Department of Theoretical Statistics, University of Aarhus, Aarhus.

HORN, D. J., FLETCHER, R. J. & KOFORD, R. R. (2000). Detecting area sensitivity: a comment on previous studies. *American Midland Naturalist* 144 28–35.

HUGHES, J., HARAN, M. & CARAGEA, P. (2010). Autologistic models for binary data on a lattice. *Environmetrics* 22 857–871.

HULL, S. D. (2000, revised 2002). Effects of management practices on grassland birds: Eastern Meadowlark. Northern Prairie Wildlife Research Center, Jamestown, ND.

JOHNSGARD, P. A. (1967). Sympatry changes and hybridization incidence in Mallards and Black Ducks. *American Midland Naturalist* 77 51–63.

KAMINSKI, R. M. (2007). Waterfowl fall migration. Why, when, and where? `http://www.ducks.org/hunting/migration/waterfowl-fall-migration`. Typeset. Accessed: 30/12/2012.

KNAPP, R., MATTHEWS, K., PREISLER, H. & JELLISON, R. (2003). Developing probabilistic models to predict amphibian site occupancy in a patchy landscape. *Ecological Applications* 13 1069–1082.

KUO, L. & MALLICK, B. (1998). Variable selection for regression models. *Sankhya* 60 65–81.

LANYON, W. E. (1957). The comparative biology of the meadowlarks (Sturnella) in Wisconsin. *Publications of the Nuttall Ornithological Club* 1. Cambridge, Massachusetts.

LANYON, W. E. (1995). Eastern Meadowlark (Sturnella magna). *A. Poole and F. Gill, editors. The Birds of North America* 160. The Academy of Natural Sciences, Philadelphia, PA.

LEWIS, J. C. & GARRISON, R. L. (1984). Habitat suitability index models: American black duck (wintering). USFWS. FWS/OBS-82/10.68. 16 pp.

LI, S. Z. (1995). *Markov Random Field Modeling in Computer Vision*. Springer-Verlag New York, Inc.

LIANG, F. (2007). Continuous contour Monte Carlo for marginal density estimation with an application to a spatial statistical model. *Journal of Computational and Graphical Statistics* 16 608–632.

LIANG, F. & WONG, W. H. (2000). Evolutionary Monte Carlo: applications to C_p model sampling and change point problem. *Statistics Sinica* 10 317–342.

LINDSAY, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics* 80 221–239.

LONGCORE, J. R., MCAULEY, D. G., HEPP, G. R. & RHYMER, J. M. (2000). American Black Duck (Anas rubripes), The birds of North America online (A. Poole, ed.). Ithaca: Cornell Lab of Ornitology. `http://bna.birds.cornell.edu/bna/species/481`. Accessed: 23/11/2012.

MANEL, S., WILLIAMS, H. C. & ORMEROD, S. J. (2002). Evaluating presence-absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology* 38 921–931.

MCGRORY, C. & TITTERINGTON, D. (2009). Variational bayesian analysis for hidden markov models. *Australian & New Zealand Journal of Statistics* 51 227–244.

MILLER, J. A. (2012). Species distribution models: spatial autocorrelation and non-stationarity. *Progress in Physical Geography* 36 681–692.

MITCHELL, T. J. & BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression (with discussion). *Journal of the American Statistical Association* 83 1023–1032.

MØLLER, J., PETTITT, A., REEVES, R. & BERTHELSEN, K. (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normilizing constants. *Biometrika* 93 451–458.

MUNSON, M. A., WEBB, K., SHELDON, D., FINK, D., HOCHACHKA, W. M., ILIFF, M., RIEDWALD, M., SOROKINA, D., SULLIVAN, B., WOOD, C. & KELLING, S. (2011). The eBird reference dataset, version 3.0. Ithaca: Cornell Lab of Ornitology and National Audubon Society. `http://www.ebird.org/`. Accessed: 30/07/2012.

MURRAY, I. (2007). *Advances in Markov chain Monte Carlo methods*. Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London, U.K.

NATURESERVE (2012). All About Birds. http://www.allaboutbirds.org/. Accessed: 12/11/2012.

O'HARA, R. B. & SILLANPÄÄ, M. J. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis* 4 85–118.

OSBORNE, P. E., ALONSO, J. & BRYANT, R. (2001). Modelling landscape-scale habitat use using GIS and remote sensing: a case study with great bustards. *Journal of Applied Ecology* 38 458–471.

OSBORNE, P. E. & TIGAR, B. J. (1992). Interpreting bird atlas data using logistic models: an example from Lesotho, Southern Africa. *Journal of Applied Ecology* 29 55–62.

PACIOREK, C. J. (2007). Bayesian smoothing with Gaussian processes using Fourier basis functions in the spectralGP library. *Journal of Statistical Software* 19 nihpa22751.

PAROLI, R. & SPEZIA, L. (2008). Bayesian variable selection in Markov mixture models. *Communications in Statistics-Simulation and Computation* 37 25–47.

PETTITT, A. N., N, F. & REEVES, R. (2003). Efficient calculation of the normalising constant of the autologistic and related models on the cylinder and lattice. *Journal of the Royal Statistical Society, Ser. B* 65 235–247.

PIELOU, E. (1979). Biogeography. 351 pp. *New York: Wiley4nterscience* .

REEVES, R. & PETTITT, A. N. (2004). Efficient recursions for general factorisable models. *Biometrika* 91 751–757.

RINGELMAN, J. K., LONGCORE, J. R. & OWEN, R. B. J. (1982). Breeding habitat selection and home range of radio-marked black ducks (Anas rubripes) in Maine. *Canadian Journal of Zoology* 60 241–248.

RYDEN, T. & TITTERINGTON, D. M. (1998). Computational Bayesian analysis of hidden Markov models. *Journal of Computational and Graphical Statistics* 7 194–211.

SALAROLI, B. (2011). The Internet Bird Collection. http://ibc.lynxeds.com/photo/eastern-meadowlark-sturnella-magna/individual-adult-male/. Accessed: 12/11/2012.

SARGEANT, G. A., SOVADA, M. A., SLIVINSKI, C. C. & JOHNSON, D. H. (2005). Markov chain Monte Carlo estimation of species distributions: a case

study of the Swift Fox in western Kansas. *Journal of Wildlife Management* 69 483–497.

SEYMOUR, N. R. & MITCHELL, S. C. (2006). Mallard A. platyrhynchos abundance, occurrence of heterospecific pairing and wetland use between 1976 and 2003 in Northeastern Nova Scotia, Canada. *Wildfowl* 56 79–93.

SIMONS, T., POLLOCK, K., WETTROTH, J., ALLDREDGE, M., PACIFICI, K. & BREWSTER, J. (2009). Sources of measurement error, misclassification error, and bias in auditory avian point count data. *Modeling demographic processes in marked populations* 237–254.

SMITH, E. & STEPHENSON, A. (2009). An extended gaussian max-stable process model for spatial extremes. *Journal of Statistical Planning and Inference* 139 1266–1275.

SPEZIA, L. (2010). Bayesian analysis of multivariate Gaussian hidden Markov models with an unknown number of regimes. *Journal of Time Series Analysis* 31 1–11.

TETERUKOVSKY, A. & EDENIUS, L. (2003). Effective field sampling for predicting the spatial distribution of Reindeer (Rangifer tarandus) with help of the Gibbs sampler. *Ambio* 32 568–572.

TOBLER, W. R. (1970). A computer movie simultaing urban growth in the Detroit region. *Economic Geography* 46 234–240.

WALK, J. W. & WARNER, R. E. (1999). Effects of habitat area on the occurrence of grassland birds in Illinois. *American Midland Naturalist* 141 339–344.

WALKER, P. A. (1990). Modelling wildlife distributions using a geographic information system: kangaroos in relation to climate. *Journal of Biogeography* 17 279–289.

WIKLE, C. K. (2003). Hierarchical models in environmental science. *International Statistical Review* 71 181–199.

WINTER, M. (1998). *Effect of habitat fragmentation on grassland-nesting birds in southwestern Missouri.* Ph.D. thesis, University of Missouri, Columbia, MI. 215 pp.

WRIGHT, B. S. & WYNDHAM, M. (2005). American Black Duck. Hinterland Who's Who bird fact sheets. Canadian Wildlife Service and Canadian Wildlife Federation. `http://www.hww.ca/en/species/birds/american-black-duck-1-2.html`. Accessed: 30/12/2012.

WU, H. & HUFFER, F. W. (1997). Modelling the distribution of plant species using the autologistic regression model. *Environmental and Ecological Statistics* 4 49–64.

YANO, Y., BEAL, S. & SHEINER, L. (2001). Evaluating pharmacokinetic/pharmacodynamic models using the posterior predictive check. *Journal of pharmacokinetics and pharmacodynamics* 28 171–192.

ZHOU, X. & SCHMIDLER, S. C. (2009). Bayesian parameter estimation in Ising and Potts models: A comparative study with applications to protein modeling. Technical Report. Duke University.(a).

ZHU, H. T., GU, M. G. & PETERSON, B. (2007). Maximum likelihood estimation for spatial models by Markov chain Monte Carlo stochastic approximation. Technical Report. Department of Statistical Science, Duke University, Durham.

ZHU, J., ZHENG, Y., CARROLL, A. & AUKEMA, B. (2008). Autologistic regression analysis of spatial-temporal binary data via monte carlo maximum likelihood. *Journal of agricultural, biological, and environmental statistics* 13 84–98.