SAPIENZA
UNIVERSITÀ DI ROMA

Dottorato di Ricerca in Statistica Metodologica

Tesi di Dottorato XXIV Ciclo – 2009/2011

Dipartimento di Scienze Statistiche

Bayesian Modeling of Presence-only Data

Natalia Golini

natalia.golini@uniroma1.it

Abstract

This thesis develops models and methods for statistical analysis of presence-only data. Besides constructing new models, the emphasis is on the theoretical characteristics of new models and on Bayesian prediction. Monte Carlo Markov chains algorithms are developed for the new presence-only data models in order to be able to simulate the posterior distribution of the unknowns and the predictive distribution of variable of interest. The new methods are applied to simulated data. One application in ecologic science have been a driving force behind the work.

Keywords: Bayesian models, Data augmentation, MCMC algorithm, Presence-only data, Potential distribution, Pseudo-absence approach, Semicontinuous data, Spatial statistics, Two-part model.

Contents

Chapter 1

Introduction

The aim of this first chapter is to provide both an introduction to presence-only data problem and an overview of modeling of presence-only data. In Section 1.1 the presence-only data problem is introduced. A first abstract definition of issues and features of interest in this setting is given in Section 1.2. In Section 1.3 a review of models proposed in the literature for presence-only data is considered. In Section 1.4 the structure of the thesis is explained.

1.1 Introduction to presence-only problem

The presence-only data problem represents an important issue in ecological studies. Here the researches are interested in prediction of the potential spatial extent of a species in suitable areas. Maps of species distributions or habitat suitability are required for many aspects of environmental research, resource management, and conservation planning (Scott & Csuti (1997)). These include biodiversity assessment, reserve design, habitat management, and restoration, species and habitat conservation plans and predicting the effects of environmental change on species and ecosystems. Presence-only data concerns both mobile (animal) and immobile (plant) species studies. In this thesis the focus is on plant ecology analysis.

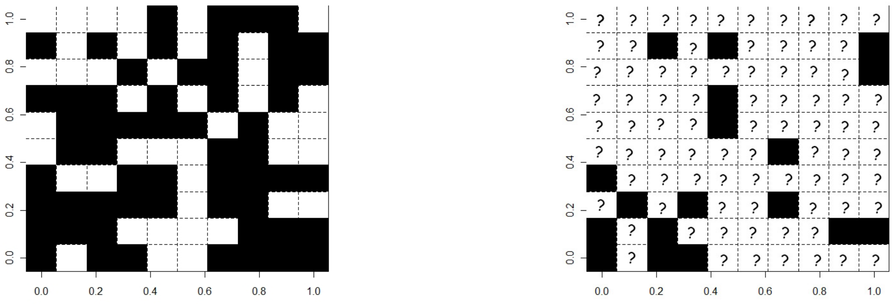
Given presence-absence data for a species, the logistic regression model is a very popular tool to model species distribution. In Keating & Cherry (2004) the use of such model for wildlife habitat-selection studies is promoted. Standard analysis methods for presence-absence data have been used in species distribution modeling since a long time (Austin (1985)). The currently dominant approaches for modeling species distribution are represented by the generalized linear and generalized additive models (GLM and GAM), and climate (environmental) envelope models. These last models use current distributions of species to build a scenario of the climatic conditions that may best suits the considered species. This envelope can then be used to forecast where species could live under predictions of future climate changes (Cressey (2008)). See Guisan & Zimmermann (2000) and Guisan et al. (2002) for a review.

However data availability is often a major constraint in modeling of specie distribution (Osborne et al. (2001) and Kaschner et al. (2006)). Collecting presence-absence data, in fact, could be expensive and/or difficult, see e.g. Guisan & Zimmermann (2000). "The vast majority of data that is available today consist of presence-only data sets" (Zaniewski et al. (2002)), defined by Pearce & Boyce (2006) as "consisting only of observations of the organism but with no reliable data where the species was not found". In practice, the only information that are available on the species is, very often, the true presence of the species in a given number of locations in the study area and the environmental covariates for the entire area. Atlases, museum and herbarium records, species lists, incidental observation databases and radio-tracking studies are example of such data.

Although presence-only data have always been used in ecology to model the species distributions, the term "presence-only data" was rarely used before the 1990's. Google scholar reports 13 papers containing the word "presence-only data" from 1990 to 2000, 141 from 2001 to 2005, 447 from 2006 to 2008 and 807 form 2009 to 2011. The literature describing and comparing methods of modeling presence-only data is growing too, see Keating & Cherry (2004), Pearce & Boyce (2006), Elith et al. (2006), Elith & Leathwick (2009) and Franklin (2010). The use of presence-only data in the last years has been supported by the increase of data availability. In some areas of ecological research the quantity and quality of available data is increased thanks to electronics collection, e.g. through remote sensing of environmental measurements (Lillesand et al. (2004)) and Geographic Information Systems (Austin (2002)) or GPS tracking of mobile species (Weimerskirch et al. (2002) . Additionally, such data is becoming increasingly available online (Stein & Wieczorek (2004)). Unfortunately, in other areas of research, especially in the study of rare species (Engler et al. (2004)), the data collection is timeconsuming and expensive yet.

1.2 Presence-only data problem: an abstract definition

The presence-only data problem can be seen as a censured (or missing) data problem. Let $\mathcal{D}$ be a regular (or irregular) lattice representing the study area, dived into units (or pixels), which are squares of equal size. Suppose that a presence-absence process $\boldsymbol{Y}$ on $\mathcal{D}$ (Figure 1.1(a)) "lives" on the lattice. In black are indicated the units where a presence is given ($y = 1$) and in white the units where an absence is given ($y = 0$). Imagine now that, for some unfathomable reason, one is able to observe only some units where the presence occurred in the study area and the covariates for the entire area. These data are refereed as "presence-only data", see Figure 1.1(b).



(a) black = presence, white = absence

(b) the question mark denotes lack of information on not observed squares of the lattice.

Figure 1.1: Presence-only data problem

In this setting three aspects are of primary interest: the species prevalence (the proportion of presences in the study area), the potential distribution of the species itself and the bias due to presence-only sampling. The first quantity answers to the question "how many presences", the second to the question "where they are located" and the third to the question "how to take into account the sampling bias".

1.3 An overview of the modeling of presence-only data

Given the nature of the available data, the desired presence-absence analysis is precluded. Then, different approaches to the modeling of species distribution based on presence-only data have been proposed in the literature. The modeling of presence-only data in ecology is reviewed in Keating & Cherry (2004), where it is referred to as use-availability data, and Pearce & Boyce (2006).

It is possible to distinguish, substantially, four different model-based group to model species presence, in relation to given environmental covariates when presence-only data are available. The type of method used depends on the nature of the data investigated (presence-only or abundance-only data) and on the assumption made on these (dependent or independent data).

Presence-only approach

In the first group the modeling focus is on the building of habitat suitability maps. Some techniques use the environmental covariates to model species presence and the results is a map returning *habitat suitability levels* for a given species in a given area. The idea is to summarize the suite of environmental attributes of species site occurrences and to extrapolate presence in other sites with similar attributes. These techniques, based directly on the environmental envelope associated with observed occurrences, represent the simplest approach to predicting species distributions based on presence-only data. The approach of defining limits for each of the environmental variables captures the sense of a niche as understood by ecologists: that the occurrence of species should be limited by a range of environmental factors, and that an envelope around those ranges would have predictive utility (Stockwell (2007), Chapter 4). A variety of models and softwares are proposed in this context (e.g. Busby (1986), Caughley et al. (1987), Lindemayer et al. (1991), Law (1994), Pearce & Lindermayer (1998), Walther et al. (2004)). However, the most widely applied profile techniques have been BIOCLIM (Busby (1986), Busby (1991)) and HABITAT (Walker & Cocks (1991)).
The BIOCLIM procedure identifies locations where all climatic indices fall within the extreme values determined from a set of observation records. Multiple levels of classification are achieved by identifying locations with climatic values contained within fractional parts of the study area. Thus BIOCLIM defines the environmental envelop for a target taxon as a rectilinear volume in a Euclidean space. Instead of a rectilinear volume in environmental space, HABITAT uses the convex hull of the training sites to more tightly constrain the environmental envelope.
Support vector machines (SVM) for one-class problems represents a variation of the previous approaches (e.g. Guo et al. (2005)). SVM seek to identify an environmental envelope (or hyperspace) containing the data points, in which the envelope is optimized with respect to the number of points in the envelope and to the number of outliers. The distance between the point and the center of the environmental

envelope determines membership to the envelope. One advantage of this approach over BIOCLIM is that the SVM hyperspace can have any shape, whereas BIOCLIM uses hyperboxes to enclose the presence data (Guo et al. (2005)). HABITAT also is more flexible than BIOCLIM. It defines the environmental envelop using a convex hull and the relative density of observations within the environmental space. In this sense, therefore, SVM may be considered a refinement of the HABITAT approach.

Other techniques that require only presence data are referred to multivariate association methods. Chief among these methods has been DOMAIN (Carpenter et al. (1993)). This model uses a point-to-point similarity metric to assign a classification value to a candidate site based on the proximity in environmental space of the most similar recorded site. DOMAIN offers significant advantages over spatial models that rely on rectilinear or convex hull environmental envelopes. It can be used to determine either environmental envelops or a continuous map of similarity, and is particularly well suited to applications where available site location records or environmental data are limited.

Profile techniques use different classification algorithms but often provide similar results. Arbitrary thresholds are typically used in identifying environmentally similar location and no uncertainty is associated with such predictions. Predictions are generally coarse. Also, the profile techniques summarize environmental characteristics at present locations and, typically, each record has equal weight within the model. Then, these techniques are highly dependent on biases in the presence records. A discussion of the pros and cons of geographical and climate envelope-based techniques is provided in Elith & Burgaman (2002). However, these techniques can be most useful when the data quality (i.e. species records, environmental predictor, biological information, etc.) is scarce.

Thus far deterministic methods based on geometric or machine learning techniques useful to determine the environmental envelope have been considered. Hereinafter two interesting not deterministic methods are illustrated. The first one is proposed in Heikkinen & Högmander (1994) where the observational process is directly modelled and the only covariate exploited in the model is the “square-specific coverage”. In this work the authors develop procedures for estimating biogeographical ranges as restoration of atlas maps, applying statistic methods of image analysis.

Also, the work of Phillips et al. (2006) is stressed because, thanks to its frequently superior predictive performance and availability of a user-friendly interface, MaxEnt is now becoming one of the standard approach for modeling presence-only data. In the paper, the authors propose the use of a maximum entropy method (MaxEnt) for modeling species geographic distributions with presence-only data. MaxEnt is a general-purpose machine learning method with a simple mathematical formulation for making predictions or inferences from incomplete information. Its origins lie in statistical mechanics (Jaynes (1957)) The core of MaxEnt model output is the estimate of ratio of conditional density of covariates on the presence sites and the marginal density of covariates across the entire study area. That gives insight on which features are important and allows to estimate the relative suitability of one location versus another. The aim is to estimate a potential species distribution by finding the probability distribution of maximum entropy, subject to a set of constraints that represent incomplete information about the target distribution. The information available about this distribution often presents itself as a set of real-valued covariates, and the constraints are that the expected value of each covariate should match its empirical average. The resultant surface is interpreted as providing the relative probability of observing a species at a given location compared to other location in the study area. However, MaxEnt is unable to provide an absolute intensity. It is not possible to determine the number of observations in a specified area. Also, MaxEnt is unable to attach any uncertainty to the resulting optimized estimates. A version of MaxEnt that handles incomplete presence-only data is proposed in Huang & Salleb-Aouissi (2009). The authors provide a formulation that is able to learn from known values of incomplete data without having to imputed values, which can be inaccurate. Also, a statistic explanation of MaxEnt for ecologists is in Elith et al. (2011).

Pseudo-absence approach

The second group consists of such techniques for presence-only data that are adaptations of existing models for presence-absence data. They require the generation of so called "pseudo-absences", a random sample of locations in the study area with known environmental covariates, to be used in place of the missing absences. The interpretation of the meaning of "pseudo-absence data" (or background data) varies. These locations may be selected without replacement from the entire study area either randomly (Hirzel & Guisan (2002)), or randomly with case-weighting to reduce the effective sample size of pseudo-absences (Ferrier & Watson (2009); Ferrier et al. (2002)), or by using environmentally weighted random sampling (Zaniewski et al. (2002)).

At the moment, most of pseudo-absence techniques are based on generalized linear models (GLM) and generalized additive models (GAM) (Ferrier et al. (2002)), artificial neural networks (Lek et al. (1996)), tree-based methods (Ferrier & Watson (1996), Elith et al. (2008)) and genetic algorithm (e.g. GARP) (Stockwell & Peters (1999)). GARP is based upon an artificial intelligence framework to produce a set of positive and negative rules that together give a binary (presence-absence) prediction. However, in Ferrier & Watson (1996), the authors show that regression model performs better than tree-based methods or genetic algorithm in predicting species presence. In particular, GLMs enabled pioneering regression-based species distribution models (SDMs; Elith & Leathwick (2009)). See Austin's work in 1970s and 1980s, cited in Austin (1985). Moreover their structural features (non-normal error distribution, additive terms, nonlinear fitted functions) continue to be useful and are part of many current methods including resource selection function (RSF; Manly et al. (2002)) and maximum entropy models (MaxEnt; Phillips et al. (2006)). In this group of modeling approach the work of Ward et al. (2009) is stressed because it represents the starting point of this thesis. The authors propose an application of the EM algorithm that provides a flexible method of estimating an underlying logistic regression model from presence-only data. However, this requires the knowledge of the overall population prevalence of the species. Without an independent estimate of overall species prevalence, estimation of the logistic model is unstable due to over-reliance on the logistic form of the model itself. Instead, in Divino et al. (2011*a*) is proposed a Bayesian model that allows to overcome the need of knowing a priori the population prevalence. This work represents an original contribute of the thesis. See Chapter 4 for details.

Point pattern analysis approach

In the third group of modeling approaches Poisson point process models for the analysis of presence-only data in likelihood (Warton & Shepherd (2010)) and Bayesian approach (Chakraborty et al. (2011)) are considered. In both papers the presence-only data are viewed as a point pattern. The idea is to model the intensity of a point process in terms of available information on the environments across the study area through regression modeling. In other words, with data consisting only of presences it is possible only to build a point pattern model to learn about the intensity of the process that drives this pattern. In these works the modeling prospective changes. Because to infer about the probability of presence at a location is not possible with presence-only data, to infer on the presences density becomes the focus. The latter is equivalent to infer on the distribution of presence locations over the study area.

In Warton & Shepherd (2010) the authors show that their method is approximately equivalent to a logistic regression, when a suitable number of regularly or randomly spaced pseudo-absences are used. However, they argue that the pseudo-approach has problems with model specification, interpretation and implementation and that each of these difficulties can be resolved using a point process modeling framework.

For the authors the pseudo-absence approach as it is usually applied appears to involve coercing the data to fit the model (presence-absence model) rather than choosing a model that fits the original data (point-events). Also, in the pseudo-absence approach is modelled the probability that a given point event is a presence not a pseudo-absences. In contrast the intensity of the process at a point has a natural interpretation as the limiting expected number of presences per units area. Yet, point process models offer a framework for choosing the "quadrature" (or pseudo-absence) points. Instead no such framework for the choice of pseudo-absences is offered in the pseudo-absence approach. Here how many pseudo-absences to choose and where to put them are tricky issues that have, instead, natural solution given a point process model specification of the problem.

In Chakraborty et al. (2011) the authors consider a hierarchical model to enable uncertainty in the inference with regard to the intensity surface. In Warton & Shepherd (2010), instead, the intensity of the process is modelled as a function of k explanatory variables with no additional uncertainty sources. Also, in Chakraborty et al. (2011) a spatial structure into their model is introduced for the intensity surface through spatial random effects. Then, a spatial model that model anticipated spatial dependence in presence-absence probabilities is defined. The authors argue how much of the works shown in the previous pages are "non-spatial in the sense that, though it includes spatial covariate information, they don't model anticipated spatial dependence in presence-absence probabilities. Accounting for the latter seems critical since causal ecological explanations such as [...] suggest that, at sufficiently high resolution, occurrence of a species at one location will be associated with its occurrence at neighboring locations (Ver Hoef et al. (2001))". Although point process models are a logical and elegant solution to the presence only problem, they rely on asymptotic results to obtain models estimates, than requiring large amount of data to be employed. When rare species are considered a little amount of data is available and a pseudo-absence approach is most likely the only feasible. However the logical problem pointed out by Chakraborty and coauthors can be bypassed by applying a simple trick. A (possibly) regular lattice is superimposed to the study area, a presence (1) is then associated to each cell where at least one presence has been recorded, an absence (0) otherwise. In this way countably many 1s and 0s can be observed. In this setting it would interesting to compare the results obtained in a simulation study using the model proposed in Chakraborty et al. (2011) with the one proposed in Divino et al. (2011*b*). Here a spatial extension of the model proposed in Divino et al. (2011*a*) is presented. This model represents an original contribution of this thesis. See Chapter 5 for details.

Abundance given presence only approach

Often, in ecological studies a measure of the relative abundance is made at locations where the species has been recorded. Examples are counts of individuals, indices of abundance and density measurements. However, few models have been proposed in the literature in order to treat this kind of data. Modified zero-inflated Poisson or negative binomial (ZIP or ZINB) regression models to model abundance given availability, where available locations are assigned a value zero, are proposed in Welsh et al. (1996), Barry & Welsh (2002), Dirnböck & Dullinger (2004) and Nielsen et al. (2005). At the moment seems that the only application explicitly modeling species abundance given presence-only data is the one proposed in Di Lorenzo et al. (2011). In this work, that represents an original contribution of this thesis, the analysis has been carried out by means of a two-part model See Chapter 3 for details.

1.4 Organization of the thesis

In this thesis models and methods for the statistical analysis of presence-only data have been developed in a Bayesian framework. Here the pseudo-absence approach is considered. Although conceptually wrong, in Warton & Shepherd (2010) the authors have demonstrated that the pseudo-absence approach is equivalent to the point process modeling for a large numbers of pseudo-absences regularly spaced or uniformly located at random over the study area. Then, under these assumptions, the potential distributions of a species obtained thought the pseudo-absence and the point pattern analysis approach, respectively, are equivalent. Also, the logistic regression model represents the most manageable tool used in ecological studies.

The pseudo-absence approach is based on a complete sample composed by two distinct samples: a sample composed by the locations where the true presences are observed and a sample of pseudo-absence or background data. In this thesis it is assumed that the pseudo-absence sample is randomly drawn from the entire area of study.

In Chapter 2 the methodological core of the thesis is developed. A logistic regression model adapted to three sampling designs (simple random sampling, case-control and censured case-control design) is introduced, and the analogy between the censured case-control sample design and the pseudo-absence approach is shown. Then a detailed analysis of a correction factor introduced to adapt the model to the various design is reported. This factor is based on the ratio of the sampling rates for the presences and absences in the complete sample.

In the pseudo-absence approach the management of the correction factor is a key point in the modeling of presence-only data and represents one of the main contributions of the present work. Remark that the presence-only data can be seen as "censured data" (or samples with not missing at random data). In this framework, because the censoring effect is acting on the pseudo-absence data, some of the quantities that define the correction factor can be considered random quantities. In Ward et al. (2009) an adjusted logistic model is proposed where the ratio of the sampling rates for the presences and absences in the complete sample is approximated by the ratio of the expected values of the sampling rates. Also, the authors argue that this quantity can be identified only if the prevalence of the population is a known quantity. Parameter estimates of the adjusted logistic regression is in likelihood framework and the maximization conducted via EM algorithm.

An original contribution of this thesis is represented by the results obtained in

> Di Lorenzo B., Farcomeni A. and Golini N. (2011). A Bayesian Model for Presence-Only Semicontinuous Data, With Application to Prediction of Abundance of Taxus Baccata in Two Italian Regions. *Journal of Agricolture, Biological, and Environmental Statistics*, **16**, 339 – 356.

Here the same approximation of the correction factor introduced by Ward et al. (2009) is used but the need of knowing a priori the prevalence of the population is partially overcome. The prevalence of the population is considered as a parameter of the model and the uncertainty about it is modeled by an informative prior distribution elicited by experts. From this consideration, Di Lorenzo et al. (2011) propose a Bayesian version of the model in Ward et al. (2009) extended to abundance data, that is, to an outcome which is either zero or a positive real number. These data are usually refereed to as semicontinuous data or data with excess zeros. The analysis can be carried out by means of a two-part model which combines a logistic model for the probability that the response is positive, and a regression model for the log-response conditionally on it being positive. Details of the methodology are shown in Chapter 3.

An other original contribution of this thesis is published as:

> Divino, F., Golini, N., Jona Lasinio, G. and Penttinen A. (2011). Data Augmentation Approach in Bayesian Modeling of Presence-only Data. *Procedia Environmental Sciences*, **7**, 38 – 43.

In this work a random approximation of the correction factor in the adjusted logistic model allows to overcome the need to acquire strong information on the population prevalence. The model is based on the assumption that the pseudo-absence sample is randomly drawn from the entire study area and that the observed environmental covariates are the only determinants of species distributions. Because of the censoring effect acting on the pseudo-absence sample, the proportion of presences in this sample can be represented by a random quantity, i.e. the random sample prevalence. Then, an unbiased estimate of unknown value of the population prevalence is given by the proportion of presences calculated in the pseudo-absence sample. Details of the methodology are illustrated in Chapter 4.

Chapter 3 and 4 illustrate models based on the assumption that the observed environmental covariates are the only determinants of species distributions. This assumption may not be adequate or sufficient to account for a species distribution. Those models may fail to provide adequate predictive power or may underestimate the degree of uncertainty of predictions. Then, in

> Divino, F., Golini, N., Jona Lasinio, G. and Penttinen A. (2011). Spatial Bayesian Modeling of Presence-only Data. Proceedings of the 17th EYSM, Lisbon, Portugal, 2011

a spatial extension of the model proposed in Divino et al. (2011*a*) is presented. This work represents the last contribution of this thesis. It is based on the assumption that the presence-absence data are spatially dependent or autocorrelated, i.e. the degree of correlation among observations depends on their relative locations. Spatial dependence in the data is incorporated into the regression model through a spatially structured random effect. Details of the methodology are illustrated in Chapter 5.

Chapter 2

Presence-only Data Model

In this Chapter the methodological core of the thesis is developed. A logistic regression model adapted to three sampling designs (simple random sampling, case-control and censured case-control design) is introduced, and the analogy between the censured case-control sample design and the pseudo-absence approach is shown. Then a detailed analysis of a correction factor introduced to adapt the model to the various design is reported. In the pseudo-absence approach the management of the correction factor is a key point in the modeling of presence-only data and represents one of the main contributions of the present work. In the last section the choice of the Bayesian estimation approach is motivated.

2.1 Notation

For the convenience of the reader some of the symbols appearing in this chapter are listed here:

- Y is a binary response variable;
- y is a realization of Y, $y = 0, 1$;
- $\boldsymbol{Y}$ is the presence-absence process;
- $\tilde{Y}$ is a Bernoulli random variable with probability of occurrence π;
- Z is a naive representation of Y;
- z is a realization of Z;
- $\boldsymbol{Z}$ is a naive representation of $\boldsymbol{Y}$;

- $\boldsymbol{X}$ is the matrix of explanatory variables or covariates;
- $\boldsymbol{x}$ is a generic row of $\boldsymbol{X}$;
- $\boldsymbol{x}_i$ is the $i-th$ row of $\boldsymbol{X}$, i.e. the vector of covariates available for the unit i;
- $\eta(\cdot)$ is a generic regression function;
- $\mathcal{U} = \{Y_1, \dots, Y_N\}$ is the target population of finite size N, with $y_i = 0, 1$;
- $\mathcal{U}_D$ is the population from which the sample or the samples are drawn;
- $\mathcal{U}_P$ is the subpopulation of $\mathcal{U}$ of size N_1 corresponding to $y = 1$;
- $\mathcal{U}_0$ is the subpopulation of $\mathcal{U}$ of size N_0 corresponding to $y = 0$;
- S_p is the sample of size n_p drawn from $\mathcal{U}_P$;
- S_a is the sample of size n_a drawn from $\mathcal{U}_0$;
- S_u is the sample of size n_u drawn from $\mathcal{U}$;
- S is the complete sample of size n;
- n_1 is the total number of $y = 1$ in the complete sample S;
- n_0 is the total number of $y = 0$ in the complete sample S;
- $\mathcal{S}$ is a indicator variable that indicates if a unit is enclosed $(s = 1)$ or not $(s = 0)$ in the complete sample S;
- γ_1 is the sampling rate for the units with $y = 1$;
- γ_0 is the sampling rate for the units with $y = 0$;
- n_{0u} is the unknown number of unobserved absences in S_u;
- n_{1u} is the unknown number of unobserved presences in S_u;
- n_{1p} is the number of observed presences in S_p;
- π is the prevalence in the target population $\mathcal{U}$;
- $\pi(\boldsymbol{x})$ is the conditional probability of observing a presence in $\mathcal{U}$, given $\boldsymbol{x}$;
- π_D is the prevalence in the design population $\mathcal{U}_D$;
- $\pi_D(\boldsymbol{x})$ is the conditional probability of observing a presence in $\mathcal{U}_D$, given $\boldsymbol{x}$;
- π^* is the probability of occurrence of a Bernoulli data-generating model.

2.2 Logistic regression model

Logistic regression model is the most important model for binary response, see Agresti (2002). It is suitable when the response variable for each study unit can be viewed as the success or failure of a single trial.

Let Y be a random variable taking values $y = 1$ in case of success and $y = 0$ in case of failure, and let $\boldsymbol{x}$ be a set of explanatory variables available for a generic unit. Then, a logistic regression model describing the probability of observing a success given the explanatory variables is assumed to be:

$$\Pr(y = 1 \mid \boldsymbol{x}) = \frac{\exp\{\eta(\boldsymbol{x})\}}{1 + \exp\{\eta(\boldsymbol{x})\}} \tag{2.1}$$

where $\eta(\cdot)$ is the regression function used in the data-generating model. Linear, nonlinear and multimodal relationships can be accounted by modeling $\eta(\cdot)$ as a simple linear model or a generalized additive model (GAM) or a boosted tree.

One important feature of the logistic regression model is its applicability to data collected via various sampling schemes. Let $\boldsymbol{X}$ be a matrix of explanatory variables, a standard cross-section study involves simultaneous measurements of Y and $\boldsymbol{X}$ for a random sample of units from a target population. In this situation, a logistic regression analysis provides an estimate of the conditional distribution of the response variable given the explanatory variables in that population, see (2.1). It is also possible to learn about the relationship between Y and $\boldsymbol{X}$ via *retrospective* or *case-control* sampling. Then two independent samples of predeterminate size are drawn from the two subpopulations of the target population, corresponding to $y = 1$ and $y = 0$ respectively. In both samples a set of explanatory variables are observed for all observations. However, in this case a standard logistic regression analysis does not provide valid estimates of the logistic model in (2.1), see Section 2.1.2.

In the following Sections the use of logistic regression is addressed, distinguishing among three sample designs: simple random, case-control and censured case-control.

2.2.1 Simple Random design

Let $\mathcal{U} = \{Y_1, \ldots, Y_N\}$ be the target population of finite size N, with $y = 0, 1$. Also, let $\boldsymbol{x}$ be a generic row of the matrix of covariates $\boldsymbol{X}$ corresponding to the vector of covariates available for a generic unit. In a random sampling the design population $\mathcal{U}_D$, the population from which the sample or the samples are drawn, coincides with the target population $\mathcal{U}$. In detail, that sampling scheme represents the simplest sampling design in which n observations are drawn randomly from the N available units, and the response and the vector of covariates, $\boldsymbol{x}$, are observed for each observations. The conditional probability model for Y in the sample is the same as the conditional probability model for Y in the population and both are properly described by (2.1).

2.2.2 Case-control design

Often, more complex sampling schemes are useful in reducing costs in the survey, in particular when the response variable represents a rare event. The case-control design represents one of the most common choices in the literature (see Schlesselman (1982), Woodward (1999), Hosmer & Lemeshow (2000)).

Let $\mathcal{U}$ be again the target population of finite size N, and let $\mathcal{U}_P$ and $\mathcal{U}_0$ the two subpopulations of $\mathcal{U}$ corresponding to $y = 1$ and $y = 0$ respectively. Let N_1 and N_0 be the dimensions of the subpopulations, with $N = N_1 + N_0$. In a case-control sampling the design population $\mathcal{U}_D$ coincides with the target population $\mathcal{U}$. In particular, that sampling scheme involves the observation of units from two independent samples of predetermined size: S_p containing n_p observations randomly drawn, without replacement, from the subpopulation $\mathcal{U}_P$ with response $y = 1$ (the cases) and S_a containing n_a observations randomly drawn, without replacement, from $\mathcal{U}_0$ with response $y = 0$ (the controls). In both samples a set of covariates are observed for all observations. Then, the complete sample S, composed by the two samples S_p and S_a, is no longer described by (2.1) because conditional probability model of observing a success in S now differs from conditional probability model of observing a success in $\mathcal{U}$. To devise an appropriate model when this sample scheme is employed, an indicator variable $\mathcal{S}$ denoting whether an observation appears in the sample S is needed (Hosmer & Lemeshow (2000)). Then, let $s = 1$ for each observation that is sampled and $s = 0$ otherwise. Also, let $\gamma_1 = \Pr(s = 1 \mid y = 1)$ and $\gamma_0 = \Pr(s = 1 \mid y = 0)$ be the sampling rates for the cases and controls respectively, both being independent of the covariates $\boldsymbol{x}$. Let n_1 and n_0 be the total number of observations with $y = 1$ and $y = 0$ in the complete sample S, respectively. Remark that in this sampling design $n_p = n_1$ and $n_a = n_0$. When sampling from $\mathcal{U}$, a finite population, the sampling rates for the cases and controls are defined as:

$$\gamma_1 = \frac{n_1}{N_1} \qquad \text{and} \qquad \gamma_0 = \frac{n_0}{N_0}$$

and, the conditional probability model describing the case-control design is given by:

$$\Pr(y=1 \mid s=1,\boldsymbol{x}) \;=\; \frac{\exp\left\{\eta(\boldsymbol{x})+\ln\left(\frac{\gamma_1}{\gamma_0}\right)\right\}}{1+\exp\left\{\eta(\boldsymbol{x})+\ln\left(\frac{\gamma_1}{\gamma_0}\right)\right\}} \tag{2.2}$$

The using the modified logistic regression model (2.2) with case-control data allows us to estimate $\eta(\boldsymbol{x})$. Also, because the values of the quantities γ_0 and γ_1 are fixed a priori, correct estimates of the events probabilities can be obtained. In the following this statement is proved.

Let $\pi(\boldsymbol{x})$ be the conditional probability model of observing a success in the target population $\mathcal{U}$:

$$\pi(\boldsymbol{x}) = \Pr(y=1 \mid \boldsymbol{x}) \;=\; \frac{\exp\{\eta(\boldsymbol{x})\}}{1+\exp\{\eta(\boldsymbol{x})\}}. \tag{2.3}$$

Then the Bayes rule can be used to derive the case-control model for the observations enclosed in the sample. Because $\mathcal{S}$ and $\boldsymbol{x}$ are independent given Y, the probability model describing the case-control design in (2.2) can be derived as follows

$$\begin{aligned}
\Pr(y=1 \mid s=1,\boldsymbol{x}) &= \frac{\Pr(s=1 \mid y=1,\boldsymbol{x})\Pr(y=1 \mid \boldsymbol{x})}{\Pr(s=1 \mid y=0,\boldsymbol{x})\Pr(y=0 \mid \boldsymbol{x})+\Pr(s=1 \mid y=1,\boldsymbol{x})\Pr(y=1 \mid \boldsymbol{x})}\\
&= \frac{\Pr(s=1 \mid y=1)\Pr(y=1 \mid \boldsymbol{x})}{\Pr(s=1 \mid y=0)\Pr(y=0 \mid \boldsymbol{x})+\Pr(s=1 \mid y=1)\Pr(y=1 \mid \boldsymbol{x})}\\
&= \frac{\gamma_1\frac{\exp\{\eta(\boldsymbol{x})\}}{1+\exp\{\eta(\boldsymbol{x})\}}}{\gamma_0\frac{1}{1+\exp\{\eta(\boldsymbol{x})\}}+\gamma_1\frac{\exp\{\eta(\boldsymbol{x})\}}{1+\exp\{\eta(\boldsymbol{x})\}}}\\
&= \frac{\gamma_1\exp\{\eta(\boldsymbol{x})\}}{\gamma_0+\gamma_1\exp\{\eta(\boldsymbol{x})\}}\\
&= \frac{\exp\left\{\eta(\boldsymbol{x})+\ln\left(\frac{\gamma_1}{\gamma_0}\right)\right\}}{1+\exp\left\{\eta(\boldsymbol{x})+\ln\left(\frac{\gamma_1}{\gamma_0}\right)\right\}},
\end{aligned}$$

that proves the statement formalized in (2.2).

2.2.3 Censured case-control design

In a censured case-control design the data consist of two distinct samples. The first is a random sample from the units with a particular response in which the covariates are completely observed. The second is a sample from the target population $\mathcal{U}$ with information only on the covariates, and where no responses are observed. Such a situation might occur if one obtains a sample of observations with a particular response and wishes, for reasons of time and economy, to compare them with a random sample from a possibly different survey in which the particular response was not measured (see Lancaster & Imbens (1996)).

Let $\mathcal{U}$ be the target population and $\mathcal{U}_P$ be the population of success defined as in Section 2.2.2. In the censured case-control design the target population $\mathcal{U}$ and the design population $\mathcal{U}_D$ does not coincide. $\mathcal{U}_D$ consists now of the target population augmented with the population of successes and has size $N + N_1$: $\mathcal{U}_D = \{\mathcal{U}, \mathcal{U}_P\}$. That sampling scheme involves the random sampling of n_p observations with $y = 1$ from the population of successes $\mathcal{U}_P$ and n_u observations from the target population $\mathcal{U}$. Then, the complete sample S is composed by two distinct samples: S_p containing n_p observations with response $y = 1$ and S_u containing n_u observations with only information on the covariates, and no responses are observed. In practice S_u, being the control group in the case-control sampling, consists of a unknown mixture of observations with response $y = 1$ and $y = 0$.

Remark that, if the successes are rare the censured case-control and case-control designs are approximately equivalent. In fact S_u will consist almost entirely of failures.

Let π be the probability of observing a success in the target population, $\pi = \Pr(y = 1)$, then it is natural to expect that the sample S_u will contain, on average, $(1 - \pi)n_u$ failures and πn_u successes. Then, π represents the expected rate of cases in S_u, as defined in Lancaster & Imbens (1996).

Remark that $S_a \subseteq S_u$ and that the proportion of successes within the complete sample S is a biased estimator of the proportion of success in the target population. Then, the probability model describing the censured case-control design is give by:

$$\Pr(y = 1 \mid s = 1, \boldsymbol{x}) \quad = \quad \frac{2\exp\left\{\eta(\boldsymbol{x}) + \ln\left(\frac{\gamma_1}{\gamma_0}\right)\right\}}{1 + 2\exp\left\{\eta(\boldsymbol{x}) + \ln\left(\frac{\gamma_1}{\gamma_0}\right)\right\}} \tag{2.4}$$

where $\gamma_1 = \Pr(s = 1 \mid y = 1)$ and $\gamma_0 = \Pr(s = 1 \mid y = 0)$ are by definition:

$$\gamma_1 = \frac{n_1}{2N_1} \qquad \text{and} \qquad \gamma_0 = \frac{n_0}{N_0}. \tag{2.5}$$

Again, the independence of γ_1 and γ_0 from the covariates $\boldsymbol{x}$ is assumed. Let's prove the (2.4). Firstly it is necessary to prove that the conditional probability of observing a presence in the design population $\mathcal{U}_D$ is the following:

$$\pi_D(\boldsymbol{x}) = \Pr(y = 1 \mid \boldsymbol{x}) = \frac{2\exp\{\eta(\boldsymbol{x})\}}{1 + 2\exp\{\eta(\boldsymbol{x})\}}. \tag{2.6}$$

By definition the conditional probability of observing a presence in $\mathcal{U}$ is equal to the ratio of the number of units that assume value 1 with observed covariates $\boldsymbol{x}$ in $\mathcal{U}$ on the total number of units with observed covariates $\boldsymbol{x}$:

$$\pi(\boldsymbol{x}) = \frac{N_1(\boldsymbol{x})}{N(\boldsymbol{x})},$$

and it is described by the model defined in (2.3).

Starting from this statement, the conditional probability of observing a success for the generic unit i in the design population $\mathcal{U}_D$ can be derived as follows:

$$\begin{aligned}
\Pr(y_i = 1 \mid \boldsymbol{x}_i) &= \Pr(y_i = 1 \mid i \in \mathcal{U}, \boldsymbol{x}_i)\Pr(i \in \mathcal{U} \mid \boldsymbol{x}_i) \\
&+ \Pr(y_i = 1 \mid i \in \mathcal{U}_P, \boldsymbol{x}_i)\Pr(i \in \mathcal{U}_P \mid \boldsymbol{x}_i) \\
&= \pi(\boldsymbol{x}_i)\frac{N(\boldsymbol{x}_i)}{N(\boldsymbol{x}_i) + N_1(\boldsymbol{x}_i)} + 1\frac{N_1(\boldsymbol{x}_i)}{N(\boldsymbol{x}_i) + N_1(\boldsymbol{x}_i)} \\
&= \frac{\pi(\boldsymbol{x}_i)N(\boldsymbol{x}_i) + N_1(\boldsymbol{x}_i)}{N(\boldsymbol{x}_i) + N_1(\boldsymbol{x}_i)} \\
&= \frac{\pi(\boldsymbol{x}_i) + \frac{N_1(\boldsymbol{x}_i)}{N(\boldsymbol{x}_i)}}{1 + \frac{N_1(\boldsymbol{x}_i)}{N(\boldsymbol{x}_i)}} \\
&= \frac{2\pi(\boldsymbol{x}_i)}{1 + \pi(\boldsymbol{x}_i)} \\
&= \frac{2\exp\{\eta(\boldsymbol{x}_i)\}}{1 + 2\exp\{\eta(\boldsymbol{x}_i)\}}
\end{aligned}$$

for each $i \in \mathcal{U}_D$.

Then the Bayes rule can be used to derive the censured case-control model conditional on the event that an observation from $\mathcal{U}_D$ is in the sample. Because $\mathcal{S}$ and $\boldsymbol{x}$ are independent given Y, the probability model describing the censured case-control design in (2.4) can be derived as follows

$$
\begin{aligned}
\Pr(y=1 \mid s=1,\boldsymbol{x}) &= \frac{\Pr(s=1 \mid y=1,\boldsymbol{x})\Pr(y=1 \mid \boldsymbol{x})}{\Pr(s=1 \mid y=0,\boldsymbol{x})\Pr(y=0 \mid \boldsymbol{x})+\Pr(s=1 \mid y=1,\boldsymbol{x})\Pr(y=1 \mid \boldsymbol{x})}\\
&= \frac{\Pr(s=1 \mid y=1)\Pr(y=1 \mid \boldsymbol{x})}{\Pr(s=1 \mid y=0)\Pr(y=0 \mid \boldsymbol{x})+\Pr(s=1 \mid y=1)\Pr(y=1 \mid \boldsymbol{x})}\\
&= \frac{\gamma_1\frac{2\exp\{\eta(\boldsymbol{x})\}}{1+2\exp\{\eta(\boldsymbol{x})\}}}{\gamma_0\left(1-\frac{2\exp\{\eta(\boldsymbol{x})\}}{1+2\exp\{\eta(\boldsymbol{x})\}}\right)+\gamma_1\frac{2\exp\{\eta(\boldsymbol{x})\}}{1+2\exp\{\eta(\boldsymbol{x})\}}}\\
&= \frac{\gamma_1\frac{2\exp\{\eta(\boldsymbol{x})\}}{1+2\exp\{\eta(\boldsymbol{x})\}}}{\gamma_0\left(\frac{1+2\exp\{\eta(\boldsymbol{x})\}-2\exp\{\eta(\boldsymbol{x})\}}{1+2\exp\{\eta(\boldsymbol{x})\}}\right)+\gamma_1\frac{2\exp\{\eta(\boldsymbol{x})\}}{1+2\exp\{\eta(\boldsymbol{x})\}}}\\
&= \frac{\gamma_1\frac{2\exp\{\eta(\boldsymbol{x})\}}{1+2\exp\{\eta(\boldsymbol{x})\}}}{\gamma_0\left(\frac{1}{1+2\exp\{\eta(\boldsymbol{x})\}}\right)+\gamma_1\frac{2\exp\{\eta(\boldsymbol{x})\}}{1+2\exp\{\eta(\boldsymbol{x})\}}}\\
&= \frac{\gamma_1 2\exp\{\eta(\boldsymbol{x})\}}{\gamma_0+\gamma_1 2\exp\{\eta(\boldsymbol{x})\}}\\
&= \frac{\frac{\gamma_1}{\gamma_0}2\exp\{\eta(\boldsymbol{x})\}}{1+\frac{\gamma_1}{\gamma_0}2\exp\{\eta(\boldsymbol{x})\}}\\
&= \frac{2\exp\left\{\eta(\boldsymbol{x})+\ln\left(\frac{\gamma_1}{\gamma_0}\right)\right\}}{1+2\exp\left\{\eta(\boldsymbol{x})+\ln\left(\frac{\gamma_1}{\gamma_0}\right)\right\}}.
\end{aligned}
$$

2.3 Pseudo-absence approach

As discussed in Section 2.1, given success-failure data, logistic regression approach and its generalizations are typically used to model dichotomous variable. As discussed in Chapter 1, these models represent an important tool for ecological studies suitable for species distribution description.

Let Y be a binary random variable measuring the presence-absence of a given species, such that $y = 1$ if the species is observed at a location and $y = 0$ if not. Let $\boldsymbol{X}$ be a matrix of environmental covariates available over the entire study area, and $\boldsymbol{x}$ is a row of $\boldsymbol{X}$, i.e. the vector of environmental covariates available at a generic location.

A tricky point in ecological studies is the definition of the response variable. Often the response variable and the environmental covariates come from different data sources that are not, necessarily, aligned in space or time (spatial misalignment, Gelfand (2010)). In general, species occurrences data are referenced to point locations while environmental data are refereed to grid cells. It is possible to work at the scale of the responses, i.e., the sample sites, and assign to each sample site the values of the environmental covariates available on the grid cell where the site falls. Alternatively, it is possible to work at grid cell level, assigning a presence to a grid cell if any sample site in that cell showed a presence, or an absence can be assigned to the cell if no presences are recorded there (Latimer et al. (2006)). In this thesis the second option is considered. In particular, a presence is assigned to a grid cell if at least one sample site in that cell showed a presence, or an absence if no presences are recorded there. Remark that under this assumption the sampling intensity, as well as how many presences occurs in the cell, is not considered. For this reason the response variable can be seen as an approximation of the true presence-absence process.

Given the nature of presence-only data it is generally not possible to calculate probabilities of presence and, then, to provide the likelihood of species presence. As discussed in Chapter 1, following the pseudo-absence approach it is possible to prove that an adjusted case-control model is a suitable model for presence-only data. That approach is based on the generation of pseudo-absence (or background) data, a random sample of locations taken from the landscape of interest. In this locations the presence or absence of a species is unknown. However, they provide a summary against which the observed presences are compared. Ward et al. (2009) argue that if one treats the observed presences and background data as if they were the true presences and absences this is wrong for two reasons. Firstly, the observed presences and the pseudo-absence data are selected with unknown sampling rates. Secondly, pseudo-absence data contain unknown number of presences. This second problem is dealt with adapting the case-control model to presence-only data.

Now, let $\mathcal{U}$ be the target population of observed presences $y = 1$ and absences $y = 0$, of size N. Also, let $\mathcal{U}_P$ be the population of size N_1 of presences from $\mathcal{U}$. Then, the design population $\mathcal{U}_D$ consists of the target population augmented with the population of presences and has size $N + N_1$: $\mathcal{U}_D = \{\mathcal{U}, \mathcal{U}_P\}$. In the pseudo-absence approach the complete sample S is composed by two samples: S_u, of size n_u, randomly sampled from the target population $\mathcal{U}$ and S_p, of size n_p and independent of S_u, randomly drawn from the population of observed presences $\mathcal{U}_P$. Under this assumption the pseudo-absence approach can be seen as a censured case-control model. Given the sampling rates for the cases and controls, that define the correction factor (2.5) used in the adjust logistic model (2.4), then the probability model that can be used to describe presence-only data is:

$$\Pr(y = 1 \mid s = 1, \boldsymbol{x}) \quad = \quad \frac{2 \exp\left\{\eta(\boldsymbol{x}) + \ln\left(\frac{\gamma_1}{\gamma_0}\right)\right\}}{1 + 2\exp\left\{\eta(\boldsymbol{x}) + \ln\left(\frac{\gamma_1}{\gamma_0}\right)\right\}}$$

where $\gamma_1 = \Pr(s = 1 \mid y = 1)$ and $\gamma_0 = \Pr(s = 1 \mid y = 0)$ that are by definition equal to:

$$\gamma_1 = \frac{n_1}{2N_1} \qquad \text{and} \qquad \gamma_0 = \frac{n_0}{N_0}.$$

2.4 The role of the correction factor

In the pseudo-absence approach the management of the correction factor, based on the ratio

$$\frac{\gamma_1}{\gamma_0} = \frac{\Pr(s = 1 \mid y = 1)}{\Pr(s = 1 \mid y = 0)}, \tag{2.7}$$

represents a crucial point. In fact, it depends on unknown quantities.

Let π be the prevalence of the finite target population $\mathcal{U}$. Then, by definition, π is the ratio of the number of presences on the population size:

$$\pi \overset{\text{def.}}{=} \frac{N_1}{N} \qquad \Rightarrow \qquad N_1 = \pi N, \tag{2.8}$$

and:

$$1 - \pi \overset{\text{def.}}{=} 1 - \frac{N_1}{N} = \frac{N_0}{N} \qquad \Rightarrow \qquad N_0 = (1 - \pi)N. \tag{2.9}$$

Then, from the pseudo-absence model the ratio (2.7) can be expressed as follows:

$$\begin{aligned} \frac{\gamma_1}{\gamma_0} &= \frac{n_1}{2N_1} : \frac{n_0}{N_0} \\ &= \frac{n_1}{2\pi N} : \frac{n_0}{(1-\pi)N} \\ &= \frac{n_1}{n_0}\frac{1-\pi}{2\pi} \end{aligned}$$

where n_1 is the total number of presences in the complete sample S, n_0 is the total number of absences in the complete sample S. All the quantities that appear in this ratio are unknown.

When presence-only data are considered one cannot observe the true process $\boldsymbol{Y}$, but is able to assess information on a naive representation $\boldsymbol{Z}$ of $\boldsymbol{Y}$. Then the pseudo-absence problem can be reformalized in terms of the variable Z, where $z = 1$ implies $y = 1$ and $z = 0$ implies $y = 1$ or $y = 0$. This first relation between the true and observed process is summarized in the following scheme:

$$\begin{cases} z = 1 \Rightarrow & y = 1; \\ z = 0 \Rightarrow & y = 1 \text{ or } y = 0 \end{cases}$$

and represented as in Table 2.1

	$z = 0$	$z = 1$	Total
$y = 0$	N_0	0	N_0
$y = 1$	N_1	N_1	$2N_1$
Total	N	N_1	$N + N_1$

Table 2.1: Population scheme for presence-only data.

The naive variable Z can be seen as a stratum variable that indicates if the observation i belongs to S_u or S_p. The complete sample S is then composed by two samples: S_u, the sample of the pseudo-absences of size n_u, corresponding to $z = 0$ and S_p, the sample of observed presences of size n_p, corresponding to $z = 1$. The presence-only data can be graphically sketched as follows:

Z	$0, \dots, 0, \dots, 0$	$1, \dots, 1$
Y	$\tilde{Y}_1, \dots, \tilde{Y}_i, \dots, \tilde{Y}_{n_u}$	$1, \dots, 1$
	S_u	S_p

where $\tilde{Y}_i$ is a Bernoulli random variable associated to each observation i belonging to S_u that represents the unobserved response variable Y_i. Because the censoring effect acting on the sample S_u, the sequence Y_i of dependent and identically Bernoulli random variables, each drawn without replacement with probability of occurrence described in (2.4), is unknown in S_u. Then, a sequence $\tilde{Y}_i$ of independent an identically Bernoulli random variables, each drawn with replacement with probability of occurrence π, is associated to the sequence of Y_i. This step underlines the presence of a double source of randomness to be considered in this framework: the first cames from the sampling scheme (case-control design) and the second from the censoring effect acting on the sample S_u.

Then, the relation between Y and Z can be re-formalized as follows:

$$\begin{cases} z_i = 1 \Rightarrow & i \in S_p, \qquad Y_i \rightarrow y_i \quad \text{is an observed value equal to 1;} \\ z_i = 0 \Rightarrow & i \in S_u, \qquad Y_i \rightarrow \tilde{Y}_i \quad \text{is a Bernoulli random variable.} \end{cases}$$

and represented as in Table 2.2

	$z = 0$	$z = 1$	Total
$y = 0$	n_{0u}	0	n_0
$y = 1$	n_{1u}	n_{1p}	n_1
Total	n_u	n_p	n

Table 2.2: Z and Y relation at sample level.

where

n_{0u} is the unknown number of unobserved absences in the sample S_u;

n_{1u} is the unknown number of unobserved presences in the sample S_u;

n_{1p} is the number of observed presences in the sample S_p;

n_0 is the unknown total number of absences in the complete sample S;

n_1 is the unknown total number of presences in the complete sample S.

The number of the true presences n_1 is obtained as the sum of the number of observed presences in the sample S_p (n_{1p}) and the unknown number of unobserved presences in the sample S_u (n_{1u}). While n_0 is equal to the unknown number of unobserved absences in the sample S_u. In formula:

$$\begin{aligned} n_1 &= n_{1u} + n_{1p}, \\ n_0 &= n_{0u}, \end{aligned}$$

and

$$\begin{aligned} n_p &= n_{1p}, \\ n_u &= n_{0u} + n_{1u}. \end{aligned}$$

Because the censoring effect acting on sample S_u, all the unknown quantities (n_1, n_0, n_{1u}, n_{0u}) can be considered random quantities and are indicate with the symbol $\sim$:

$$\begin{aligned}\tilde{n}_{0u} &= n_u - \tilde{n}_{1u}, \\ \tilde{n}_0 &= \tilde{n}_{0u}, \\ \tilde{n}_1 &= \tilde{n}_{1u} + n_p.\end{aligned}$$

In particular n_{1u} can be written as:

$$\tilde{n}_{1u} = \sum_{i \in S_u} \tilde{Y}_i.$$

As direct consequence, the correction factor is based on a ratio of random quantities:

$$\frac{\tilde{\gamma}_1}{\tilde{\gamma}_0} = \frac{\tilde{n}_{1u} + n_p}{n_u - \tilde{n}_{1u}} \frac{1-\pi}{2\pi}. \tag{2.10}$$

In order to handle the correction factor, Ward et al. (2009) approximate the ratio in (2.10) by the ratio of expected values of the sampling rates. Being π the expected value of each $\tilde{Y}_i$ in S_u, the expected number of the true presences in S_u is given by:

$$\begin{aligned}\mathbb{E}[\tilde{n}_{1u}] &= \mathbb{E}\left[\sum_{i \in S_u} \tilde{Y}_i\right] \qquad (2.11) \\ &= \sum_{i \in S_u} \mathbb{E}\left[\tilde{Y}_i\right] \\ &= \pi n_u\end{aligned}$$

and, consequently, the expected number of the true presences in the complete sample S is:

$$\begin{aligned}\mathbb{E}[\tilde{n}_1] &= \mathbb{E}[\tilde{n}_{1u} + n_p] \qquad (2.12) \\ &= \mathbb{E}[\tilde{n}_{1u}] + n_p \\ &= \pi n_u + n_p.\end{aligned}$$

Then, given (2.12), the ratio in (2.10) can be approximated as follows:

$$\begin{aligned}\frac{\tilde{\gamma}_1}{\tilde{\gamma}_0} &\approx \frac{\mathbb{E}[\tilde{\gamma}_1]}{\mathbb{E}[\tilde{\gamma}_0]} \\ &= \frac{\mathbb{E}[\tilde{n}_{1u}] + n_p}{n_u - \mathbb{E}[\tilde{n}_{1u}]}\frac{1-\pi}{2\pi} \\ &= \frac{\pi n_u + n_p}{(1-\pi)n_u}\frac{1-\pi}{2\pi} \\ &= \frac{\pi n_u + n_p}{2\pi n_u}.\end{aligned} \tag{2.13}$$

In Ward et al. (2009), the authors argue that this quantity can be identified only if the prevalence of the population π is a known quantity.

Di Lorenzo et al. (2011) use the same approximation introduced by Ward et al. (2009) and model the information about the prevalence by an informative prior distribution (expert elicitation). Details are shown in Chapter 3.

While in Divino et al. (2011*a*) a random approximation of the correction factor allows to overcome the need to acquire strong information on the population prevalence. Details are shown in Chapter 4.

2.5 The prevalence: a remark

Now assume that a population of size N has been generated from the following model:

$$\mathcal{Y} \sim Ber(\pi^*) \tag{2.14}$$

Then

$$\mathcal{U} = \{Y_1, \ldots, Y_N\}$$

is a target population with empirical prevalence π, such that $\mathbb{E}(\pi) = \pi^*$.

Let $\mathcal{U}_P$ be the subpopulation of $\mathcal{U}$ of size N_1 corresponding to $y = 1$:

$$\mathcal{U}_P = \{1, \ldots, 1\}$$

called population of successes.

As discussed in Section 2.2.3, in the censured case-control sampling the target population $\mathcal{U}$ and the design population $\mathcal{U}_D$ not coincide. The design population $\mathcal{U}_D$, in fact, consists of the target population augmented with the population of known presences

$$\mathcal{U}_D = \{\mathcal{U}, \mathcal{U}_P\}$$

and it has size $N + N_1$.
Consequently the proportion of presences in $\mathcal{U}$ is different from the one in $\mathcal{U}_D$:

$$\pi \neq \pi_D.$$

This statement is henceforward proved. From Table 2.1

$$\begin{aligned} \pi &= \Pr(y = 1 \mid z = 0) \\ &= \frac{N_1}{N} \end{aligned}$$

and

$$\begin{aligned} \pi_D &= \Pr(y = 1) \\ &= \Pr(y = 1 \mid z = 0)\Pr(z = 0) \\ &+ \Pr(y = 1 \mid z = 1)\Pr(z = 1) \\ &= \frac{N_1}{N}\frac{N}{N + N_1} \\ &+ \frac{N_1}{N_1}\frac{N_1}{N + N_1} \\ &= \frac{2N_1}{N + N_1} \\ &= \frac{2\pi N}{(1 + \pi)N} \\ &= \frac{2\pi}{1 + \pi}. \end{aligned}$$

In practice, three different prevalences appear:

π^* the prevalence of the generating model,

π the prevalence of the target population $\mathcal{U}$,

π_D the prevalence of the design population $\mathcal{U}_D$,

linked by the following properties:

$$\pi^* = \mathbb{E}[\pi] \neq \mathbb{E}[\pi_D] = \mathbb{E}\left[\frac{2\pi}{1+\pi}\right].$$

2.6 Bayesian framework

In this thesis Bayesian estimates, inference and prediction are considered. This choice being motivated mostly by the possibility, in this framework, to handle the prevalence as a parameter of the model.

2.6.1 Bayesian Logistic regression

Methods for Bayesian estimation of the logistic regression model, whether with univariate or multivariate outcome, are well established (Congdon (2007), Ntzoufras (2009), Kruschke (2010)).

Consider the logistic regression model described in 2.1:

$$\pi(\boldsymbol{x}) = \Pr(y = 1 \mid \boldsymbol{x}) = \frac{\exp\{\eta(\boldsymbol{x})\}}{1 + \exp\{\eta(\boldsymbol{x})\}},$$

where $\eta(\cdot)$ is a generic parametric function and $\boldsymbol{\beta}$ is the vector of parameters of $\eta(\cdot)$.

In order to derive inference in a Bayesian framework it is necessary to specify prior distributions, $p(\cdot)$, for all the parameters of interest and a likelihood model for the observed data, $L(\cdot; data)$. Then, the Bayesian logistic regression model is expressed as follows:

$$Y \mid \boldsymbol{x} \sim Ber(\pi(\boldsymbol{x}))$$

$$\boldsymbol{\beta} \sim p(\boldsymbol{\beta}).$$

Using the Bayes theorem, the posterior distribution of the parameters under this model is given by:

$$p(\boldsymbol{\beta} \mid \boldsymbol{y}; \boldsymbol{X}) \propto L(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{X}) p(\boldsymbol{\beta})$$

Bayesian estimates and credibility intervals of the parameters of interest are obtained simulating the marginal profiles of that posterior distribution through the use of MCMC techniques.

Chapter 3

Bayesian Modeling of Presence-only Data: a first model

In this Chapter a first Bayesian model for presence-only data is proposed. The methodology here shown is published the paper

> Di Lorenzo B., Farcomeni A. and Golini N. (2011). A Bayesian Model for Presence-Only Semicontinuous Data, With Application to Prediction of Abundance of Taxus Baccata in Two Italian Regions. *Journal of Agricolture, Biological, and Environmental Statistics*, **16**, $339 - 356$.

The model developed in this work represents a Bayesian version of the one in Ward et al. (2009), where the authors proposed a model which explicitly takes into account bias due to presence-only sampling. Ward's model can be used when the outcome of interest is a dichotomous variable measuring whether the species is present or absent in a given location.

Di Lorenzo et al. (2011) extend Ward's model to abundance data, that is, to an outcome which is either zero or a positive real number. These data are usually refereed to as semicontinuous data or data with excess zeros, and analysis can be carried out by means of a two-part model which combines a logistic model for the probability that the response is positive, and a regression model for the log-response conditionally on it being positive.

The main innovation in the work is that the uncertainty related to the zeros is explicitly take into account. The resulting model can hence also be thought to as a two-part model with partial possible measurement error. Further, the adjustment for the case-control type sampling is performed.

The inference for the model is derived in a Bayesian framework because it allows to handle the prevalence of species as a parameter of the model. Then, the information about it can be modeled by an informative prior distribution (expert elicitation).

3.1 Modeling of presence-only data

Given (2.4) and (2.13), the conditional probability model describing the adjusted logistic model in logit form can be expressed as follows:

$$\begin{aligned} \operatorname{logit} \Pr(y = 1 \mid s = 1, \boldsymbol{x}) &= \eta(\boldsymbol{x}) + \ln(2) + \ln\left(\frac{\pi n_u + n_p}{2\pi n_u}\right) \\ &= \eta(\boldsymbol{x}) + \ln\left(\frac{\pi n_u + n_p}{\pi n_u}\right) \end{aligned} \tag{3.1}$$

where the regression function $\eta(\cdot)$ is linear in $\boldsymbol{x}$: $\eta(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \ldots + x_k$. Here the attention is focused on a linear form for $\eta(\boldsymbol{x})$ being the choice motivated by the adequacy of this model for the data at hand. Obviously this is not the only choice and other more flexible choice for $\eta(\cdot)$ could be used (e.g., generalized additive models, Hastie & Tibshirani (1990)). Modification of this approach for these different choices is often straightforward and does not usually lead to major modification of the inferential strategies described in Section 3.2.1.

As discussed in Section 2.3, when presence-only data are considered one cannot observe the true process $\boldsymbol{Y}$, but is able to assess information on a naive representation $\boldsymbol{Z}$ of $\boldsymbol{Y}$. Then it is necessary to link parameters to the observed process $\boldsymbol{Z}$ in order to perform inference and, to do so, is necessary to derive the *observed likelihood* $L(\boldsymbol{\theta}; \boldsymbol{z}, \boldsymbol{X})$ for the presence-only data. Using a short-hand notation $\boldsymbol{\theta}$ to denote the parameters at stake, the observed likelihood for presence-only data is given by:

$$\begin{aligned} L(\boldsymbol{\theta}; \boldsymbol{z}, \boldsymbol{X}) &= \prod_{i \in S_u} \Pr(z_i = 0 \mid s_i = 1, \boldsymbol{x}_i) \\ &\times \prod_{i \in S_p} \Pr(z_i = 1 \mid s_i = 1, \boldsymbol{x}_i). \end{aligned}$$

An explicit expression of the latter is given first using a probability argument across $y = 1$ and $y = 0$:

$$\begin{aligned}
\Pr(z=1 \mid s=1,\boldsymbol{x}) &= \Pr(z=1 \mid y=1, s=1, \boldsymbol{x}) \Pr(y=1 \mid s=1,\boldsymbol{x}) \\
&+ \Pr(z=1 \mid y=0, s=1, \boldsymbol{x}) \Pr(y=0 \mid s=1,\boldsymbol{x}) \\
&= \Pr(z=1 \mid y=1, s=1) \Pr(y=1 \mid s=1,\boldsymbol{x}) \\
&+ \Pr(z=1 \mid y=0, s=1) \Pr(y=0 \mid s=1,\boldsymbol{x})
\end{aligned} \tag{3.2}$$

as $Z \mid Y, \boldsymbol{x} \sim Z \mid Y$. Then, from the definition of conditional probability follows that

$$\Pr(z=1 \mid y=1, s=1) = \frac{\Pr(z=1, y=1 \mid s=1)}{\Pr(y=1 \mid s=1)}.$$

Following the pseudo-absence approach the expected number of true presences in the complete sample S is $n_p + \pi n_u$, see (2.12). Hence, $\Pr(y=1 \mid s=1) = (n_p + \pi n_u)/(n_p + n_u)$.

Also, from Table 2.2, $\Pr(z=1, y=1 \mid s=1) = n_p/(n_p+n_u)$, and consequently:

$$\Pr(z=1 \mid y=1, s=1) = \frac{n_p}{n_p + \pi n_u}. \tag{3.3}$$

Further, $\Pr(z=1 \mid y=0, s=1) = 0$ because all $z=1$ in the data must occur for $y=1$.

Combining (3.1) with (3.3), after some manipulations, it is obtained:

$$\Pr(z=1 \mid s=1,\boldsymbol{x}) = 0 + \frac{\frac{n_p}{\pi n_u}\exp\{\boldsymbol{x}\boldsymbol{\beta}\}}{1+\left(1+\frac{n_p}{\pi n_u}\right)\exp\{\boldsymbol{x}\boldsymbol{\beta}\}}.$$

Then, the explicit form of the *observed likelihood* for presence-only data is given by:

$$\begin{aligned}
L(\boldsymbol{\theta}; \boldsymbol{z}, \boldsymbol{X}) &= \prod_{i \in S_u} \Pr(z_i = 0 \mid s_i = 1, \boldsymbol{x}_i) \prod_{i \in S_p} \Pr(z_i = 1 \mid s_i = 1, \boldsymbol{x}_i) \\
&= \prod_{i \in S_u}\left[\frac{1+\exp\{\boldsymbol{x}_i\boldsymbol{\beta}\}}{1+\left(1+\frac{n_p}{\pi n_u}\right)\exp\{\boldsymbol{x}_i\boldsymbol{\beta})\}}\right] \prod_{i \in S_p}\left[\frac{\frac{n_p}{\pi n_u}\exp\{\boldsymbol{x}_i\boldsymbol{\beta}\}}{1+\left(1+\frac{n_p}{\pi n_u}\right)\exp\{\boldsymbol{x}_i\boldsymbol{\beta})\}}\right]
\end{aligned} \tag{3.4}$$

where $\boldsymbol{\theta}$ is a short-hand notion for the parameters at stake.

3.1.1 Inference

Priors

In order to derive inference in the Bayesian framework, the following prior distributions are assumed:

$$p(\boldsymbol{\beta}, \pi) = p(\boldsymbol{\beta})p(\pi). \tag{3.5}$$

where $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\boldsymbol{\beta}})$, with $\Sigma_{\boldsymbol{\beta}}$ fixed, and $p(\pi) \sim Beta(a_\pi, b_\pi)$, with a_π and b_π chosen in a such way to summarize the uncertainty on the true value of the prevalence.

The informative prior on π plays a special rule in this framework as the data contain very little information on π (see Ward et al. (2009) for the identifiability issues related to π).

In practice, inference and predictions are based on the integrated likelihood (with respect to the prior on π). The nonidentification makes inference arbitrarily sensitive to the prior. The proposed model considers a parametrization with a simple contextual meaning, so that it is possible to elicit an informative prior for π. For methods in prior elicitation, see for instance Kadane et al. (1980), Kadane & Wolfson (1998), Garthwaite et al. (2005), and references therein.

Model fit

As discussed in Section 1.1, since not all Y_i are observed, the presence-only data model can be seen a missing (or censured) data model. The posterior distribution is not known in closed form, then a MCMC algorithm is required to obtain samples from such distribution. Here a MCMC sampling scheme is proposed defining a Bayesian counterpart of the EM algorithm. The sampling scheme is based on alternating a data augmentation/imputation step, in which the latent observations Y_i are sampled from their full conditional, with Metropolis Hastings (MH) steps.

Hence, the data augmentation scheme based on the generation of latent observations Y_i, allows to derive the *complete likelihood*, $L(\boldsymbol{\theta}; \boldsymbol{z}, \mathbf{y}, \boldsymbol{X})$.
Using a conditioning argument, one gets that

$$\begin{aligned}
L(\boldsymbol{\theta};\boldsymbol{z},\boldsymbol{y},\boldsymbol{X}) &= \prod_{i\in S}\Pr(Y_i, Z_i \mid s_i=1,\boldsymbol{x}_i) \qquad (3.6)\\
&= \prod_{i\in S}\Pr(Z_i \mid Y_i, s_i=1,\boldsymbol{x}_i)\Pr(Y_i \mid s_i=1,\boldsymbol{x}_i)\\
&= \prod_{i\in S}\Pr(Y_i \mid s_i=1,\boldsymbol{x}_i)\\
&= \prod_{i\in S}\Pr(y_i=0 \mid s_i=1,\boldsymbol{x}_i)^{1_{\{y_i=0\}}}\Pr(y_i=1 \mid s_i=1,\boldsymbol{x}_i)^{1_{\{y_i=1\}}}
\end{aligned}$$

where $1_{\{C\}}$ is the indicator function for condition C.

$\Pr(Y_i \mid s_i=1,\boldsymbol{x}_i)$ follows directly from (3.1):

$$\begin{aligned}
\Pr(y=1 \mid s=1,\boldsymbol{x}) &= \frac{\exp\left\{\eta(\boldsymbol{x})+\ln\left(\frac{n_p+\pi n_u}{\pi n_u}\right)\right\}}{1+\exp\left\{\eta(\boldsymbol{x})+\ln\left(\frac{n_p+\pi n_u}{\pi n_u}\right)\right\}} \qquad (3.7)\\
&= \frac{\exp\{\eta(\boldsymbol{x})\}\frac{n_p+\pi n_u}{\pi n_u}}{1+\frac{n_p+\pi n_u}{\pi n_u}\exp\{\eta(\boldsymbol{x})\}}\\
&= \frac{\left(1+\frac{n_p}{\pi n_u}\right)\exp\{\eta(\boldsymbol{x})\}}{1+\left(1+\frac{n_p}{\pi n_u}\right)\exp\{\eta(\boldsymbol{x})\}}\\
&= \frac{\left(1+\frac{n_p}{\pi n_u}\right)\exp\{\boldsymbol{x}\boldsymbol{\beta}\}}{1+\left(1+\frac{n_p}{\pi n_u}\right)\exp\{\boldsymbol{x}\boldsymbol{\beta}\}}
\end{aligned}$$

and

$$\begin{aligned}
\Pr(y=0 \mid s=1,\boldsymbol{x}) &= 1-\Pr(y=1 \mid s=1,\boldsymbol{x}) \qquad (3.8)\\
&= \frac{1}{1+\left(1+\frac{n_p}{\pi n_u}\right)\exp\{\boldsymbol{x}\boldsymbol{\beta}\}}.
\end{aligned}$$

Then, the *complete likelihood* for presence-only data, in terms of both $\boldsymbol{Z}$ and $\boldsymbol{Y}$, can be obtained by substituting 3.7 and 3.8 into 3.6

$$\begin{aligned}
L(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{z},\boldsymbol{X}) &= \prod_{i\in S}\Pr(y_i=0 \mid s_i=1,\boldsymbol{x}_i)^{1_{\{y_i=0\}}}\Pr(y_i=1 \mid s_i,\boldsymbol{x}_i)^{1_{\{y_i=1\}}} \qquad (3.9)\\
&= \prod_{i\in S}\left[\frac{1}{1+\left(1+\frac{n_p}{\pi n_u}\right)\exp\{\boldsymbol{x}_i\boldsymbol{\beta}\}}\right]^{1_{\{y_i=0\}}}\left[\frac{\left(1+\frac{n_p}{\pi n_u}\right)\exp\{\boldsymbol{x}_i\boldsymbol{\beta}\}}{1+\left(1+\frac{n_p}{\pi n_u}\right)\exp\{\boldsymbol{x}_i\boldsymbol{\beta}\}}\right]^{1_{\{y_i=1\}}}
\end{aligned}$$

where $\boldsymbol{\beta}$ represents the k-dimensional parameters vector introduce to model the k covariates contribution in the logistic model so that $\eta(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta}$.

The proposed algorithm, a Metropolis within Gibbs sampling scheme, is detailed below (Algorithm 1).

Algorithm 1 Gibbs sampling scheme

1. Sample the latent variables Y_i from $p(Y_i \mid Z_i, s_i = 1, \boldsymbol{x}_i)$, $i = 1, \ldots, n$; where $p(Y_i \mid Z_i, s_i = 1, \boldsymbol{x}_i) = 1_{y_i = z_i} * 1_{z_i=1} + 1_{z_i=0} p(Y_i \mid z_i = 0, s_i = 1, \boldsymbol{x}_i)$. That is, set $y_i = z_i$ when $z_i = 1$ and when $z_i = 0$ note that

$$p(Y_i \mid z_i = 0, s_i = 1, \boldsymbol{x}_i) = p(Y_i \mid s_i = 1, \boldsymbol{x}_i),$$

since is assumed that data are sampled uniformly at random from the study area. Sampling of Y_i when $z_i = 0$ implies to sample Y_i form the Bernoulli distribution given in (3.1).

2. Sample the regression parameters ($\boldsymbol{\beta}$) from

$$p(\boldsymbol{\beta} \mid \boldsymbol{y}; \boldsymbol{X}) \propto p(\boldsymbol{\beta}) \frac{\exp\{\sum_i^n 1_{y_i=1} \boldsymbol{x}_i \boldsymbol{\beta}\}}{1 + \left(1 + \frac{n_p}{\pi n_u}\right) \exp\{\sum_i^n 1_{y_i=1} \boldsymbol{x}_i \boldsymbol{\beta}\}}.$$

3. Sample π from its prior.

Here the sampling scheme of Diebolt & Robert (1994) for mixture (missing data) models is adapted to the presence-only data. At Step 1 the latent variables are sampled form their full conditionals. One then augments generating Y_i from its discrete full conditional distribution when $z_i = 0$. When $z_i = 1$, Y_i is not sampled since its full conditional is a point mass on Z_i.

At Step 2 the logistic regression parameters for simplicity are sampled through an Adaptive Rejection Metropolis Sampling (Gilks et al. (1995)), even if there are many different alternative approaches for this standard problem.

For simplicity in this chapter the prevalence π is sampled from its prior.

3.2 Modeling of presence-only semicontinuous data

Suppose now that the outcome of interest is not only the presence of species, but also a continuous quantity measuring its abundance. In other words, the response variable Y is either zero or positive real number, $Y \geq 0$. These data are usually refereed to as semicontinuous data or data with excess zeros.

As for the presence-only data, when presence-only semicontinuous data are considered the true process $\boldsymbol{Y}$ it is not observed, but information on the naive representation $\boldsymbol{Z}$ of $\boldsymbol{Y}$ is available. Then $z > 0$ implies $y = z > 0$, $z = 0$ implies $y \geq 0$. This relation between the true and observed process can be summarized in the following scheme:

$$\begin{cases} z > 0 \Rightarrow & \text{y>0;} \\ z = 0 \Rightarrow & \text{y}\geq\text{0.} \end{cases}$$

Let $\pi = \Pr(y > 0)$ be the prevalence of the true presence-absence process $\boldsymbol{Y}$. Following the pseudo-absence approach shown in Chapter 2, the semicontinuous response Y is modelled through a two-part model. The two parts are usually made of a logistic model for the conditional probability that the response is positive, and a regression model for the log-response conditionally on the fact that is positive. In this section the classical two-part model is extended for taking into account uncertainty related to pseudo-absences.

The adjusted logit model can be specified as follows:

$$\operatorname{logit}\Pr(y > 0 \mid s = 1, \boldsymbol{x}) \quad = \quad \eta(\boldsymbol{x}) + \ln\left(\frac{\gamma_1}{\gamma_0}\right) \tag{3.10}$$

where the regression function $\eta(\cdot)$ is linear in $\boldsymbol{x}$ ($\eta(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \ldots + x_k$); n_p and n_u are the number of observed abundances ($z > 0$) and the number of pseudo-absence data ($z = 0$), respectively.

For the continuous part of the model is assumed that:

$$\mathbb{E}[\ln(Y) \mid y > 0, \boldsymbol{x}] \quad = \quad \eta(\boldsymbol{x}) \tag{3.11}$$

where $\eta(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\alpha}$ and $\ln(Y)$ is conditionally distributed as a normal variate with standard deviation σ.

Note that the same or a different set of covariates can be used on each part of the models. Again, the attention is restricted to linear models for the the probability of observing a presence as this model was adequate for the data at hand.

Conditionally on $s = 1$ the the regression model (3.11) needs no adjustment. Using the Bayes rule, one can see as:

$$\begin{aligned} f(Y \mid y > 0, s = 1, \boldsymbol{x}) &= \frac{\Pr(s = 1 \mid Y, y > 0, \boldsymbol{x}) f(Y \mid y > 0, \boldsymbol{x})}{\Pr(s = 1 \mid y > 0, \boldsymbol{x})} \\ &= \frac{\Pr(s = 1 \mid Y, y > 0) f(Y \mid y > 0, \boldsymbol{x})}{\Pr(s = 1 \mid y > 0)} \\ &= f(Y \mid y > 0, \boldsymbol{x}) \end{aligned} \tag{3.12}$$

so, the sampling rate depends only on the presence and not on the actual value of the abundance.

Derive now an explicit expression for the ratio $\frac{\gamma_1}{\gamma_0}$ in (3.10). Following the pseudo-absence approach, the complete sample S is composed by two samples: S_u, the sample of the pseudo-absence locations of size n_u, corresponding to $z = 0$ and S_p, the sample of observed abundances of size n_p, corresponding to $z > 0$. The presence-only data can be graphically sketched as follows:

Z	$0, \dots, 0, \dots, 0$	$> 0, \dots\dots, > 0$
Y	$Y_1^*, \dots, Y_i^*, \dots, Y_{n_u}^*$	$> 0, \dots\dots, > 0$
	S_u	S_p

where Y_i^* is a semicontinuous random variable associated to each observation i belonging to S_u that represents the unobserved semicontinuous response variable Y_i. Because the censoring effect acting on sample S_u to the unknown sequence Y_i of semicontinuous random variables can be associated a sequence of Y_i^* semicontinuous random variables defined by the following two-part model:

1. Adjusted case-control model:

$$\tilde{Y}_i \sim Ber(\pi) \qquad i.i.d.$$

drawn with replacement.

2. Regression model:

$$Y_i^* \mid \tilde{y}_i = 1 \sim \mathcal{LN}(\mu, \sigma^2)$$

where $Y_i^* \mid \tilde{y}_i = 1$ is a log normal distribution.

As discussed in Section 2.4, in this framework a double source of randomness is considered.

Then, the relation between Y and Z can be re-formalized as follows:

$$\begin{cases} z_i > 0 \Rightarrow \quad i \in S_p, \qquad Y_i \to y_i \quad \text{is an observed semicnontinuous value;} \\ z_i = 0 \Rightarrow \quad i \in S_u, \qquad Y_i \to Y_i^* \quad \text{is a semicontinuous random variable.} \end{cases}$$

and represented as in Table 2.2

	$z = 0$	$z > 0$	Total
$y = 0$	n_{0u}	0	n_0
$y > 0$	n_{1u}	n_{1p}	n_1
Total	n_u	n_p	n

Table 3.1: Z and Y relation at sample level.

where, again,

n_{0u} is the unknown number of unobserved abundances ($y > 0$) in the sample S_u;

n_{1u} is the unknown number of unobserved abundances in the sample S_u;

n_{1p} is the number of observed abundances in the sample S_p;

n_0 is the unknown total number of absences in the complete sample S;

n_1 is the unknown total number of abundances in the complete sample S.

Then, the number of the true abundances n_1 is obtained as the sum of the number of observed abundances in the sample S_p (n_{1p}) and the unknown number of unobserved abundances in the sample S_u (n_{1u}). While n_0 is equal to the unknown number of unobserved absences in the sample S_u. In formula:

$$\begin{aligned} n_1 &= n_{1u} + n_{1p}, \\ n_0 &= n_{0u}, \end{aligned}$$

and

$$\begin{aligned} n_p &= n_{1p}, \\ n_u &= n_{0u} + n_{1u}. \end{aligned}$$

Yet, because of the effect induced by the censoring acting on the sample S_u, all the unknown quantities $(n_1, n_0, n_{1u}, n_{0u})$ can be considered random quantities and are indicate with the symbol $\sim$:

$$\begin{aligned} \tilde{n}_{0u} &= n_u - \tilde{n}_{1u}, \\ \tilde{n}_0 &= \tilde{n}_{0u}, \\ \tilde{n}_1 &= \tilde{n}_{1u} + n_p. \end{aligned}$$

In particular n_{1u} can be written as:

$$\tilde{n}_{1u} = \sum_{i \in S_u} \tilde{Y}_i.$$

As direct consequence, the correction factor becomes a ratio of random quantities:

$$\frac{\tilde{\gamma}_1}{\tilde{\gamma}_0} = \frac{\tilde{n}_{1u} + n_p}{n_u - \tilde{n}_{1u}} \frac{1 - \pi}{2\pi} \tag{3.13}$$

and as discussed in Section 2.4 can be approximate by the ratio of expected values of the sampling rates. Being π the expected value of each $\tilde{Y}_i$ in S_u, the expected number of the true abundances in S_u is given by:

$$\begin{aligned} \mathbb{E}[\tilde{n}_{1u}] &= \mathbb{E}\left[\sum_{i \in S_u} \tilde{Y}_i\right] \\ &= \sum_{i \in S_u} \mathbb{E}\left[\tilde{Y}_i\right] \\ &= \pi n_u \end{aligned} \tag{3.14}$$

and, consequently, the expected number of the true abundances in the complete sample S is:

$$\begin{aligned}\mathbb{E}[\tilde{n}_1] &= \mathbb{E}[\tilde{n}_{1u} + n_p] \\ &= \mathbb{E}[\tilde{n}_{1u}] + n_p \\ &= \pi n_u + n_p.\end{aligned} \tag{3.15}$$

Then, given (3.15), the random ratio can be approximated as follows:

$$\begin{aligned}\frac{\tilde{\gamma}_1}{\tilde{\gamma}_0} &\approx \frac{\mathbb{E}[\tilde{\gamma}_1]}{\mathbb{E}[\tilde{\gamma}_0]} \\ &= \frac{\mathbb{E}[\tilde{n}_{1u}] + n_p}{n_u - \mathbb{E}[\tilde{n}_{1u}]}\frac{1-\pi}{2\pi} \\ &= \frac{\pi n_u + n_p}{(1-\pi)n_u}\frac{1-\pi}{2\pi} \\ &= \frac{\pi n_u + n_p}{2\pi n_u}\end{aligned} \tag{3.16}$$

and the explicit form of (3.10) follows:

$$\operatorname{logit}\Pr(y > 0 \mid s = 1, \boldsymbol{x}) \;=\; \eta(\boldsymbol{x}) + \ln\left(\frac{\pi n_u + n_p}{\pi n_u}\right) \tag{3.17}$$

Then, the *observed likelihood* for presence-only semicontinuous data is given by:

$$\begin{aligned}L(\boldsymbol{\theta}; \boldsymbol{z}, \boldsymbol{X}) &= \prod_{i \in S_u} \Pr(z_i = 0 \mid s_i = 1, \boldsymbol{x}_i) \\ &\times \prod_{i \in S_p} \left[\Pr(z_i > 0 \mid s_i = 1, \boldsymbol{x}_i) f(z_i \mid z_i > 0, s_i = 1, \boldsymbol{x}_i)\right].\end{aligned}$$

where $\boldsymbol{\theta}$ is a short-hand notation to denote the parameters at stake.

Again, an explicit expression of the latter is given first using a probability argument across $y > 0$ and $y = 0$:

$$\begin{aligned}\Pr(z > 0 \mid s = 1, \boldsymbol{x}) &= \Pr(z > 0 \mid y > 0, s = 1, \boldsymbol{x}) \Pr(y > 0 \mid s = 1, \boldsymbol{x}) \\ &+ \Pr(z > 0 \mid y = 0, s = 1, \boldsymbol{x}) \Pr(y = 0 \mid s = 1, \boldsymbol{x}) \\ &= \Pr(z > 0 \mid y > 0, s = 1) \Pr(y > 0 \mid s = 1, \boldsymbol{x}) \\ &+ \Pr(z > 0 \mid y = 0, s = 1) \Pr(y = 0 \mid s = 1, \boldsymbol{x})\end{aligned}$$

as $Z \mid Y, \boldsymbol{x} \sim Z \mid Y$. Then, from the definition of conditional probability follows that:

$$\Pr(z > 0 \mid y > 0, s = 1) = \frac{\Pr(z > 0, y > 0 \mid s = 1)}{\Pr(y > 0 \mid s = 1)}.$$

As for the presence-only data, the expected number of true absences in the complete sample S is $n_p + \pi n_u$. Hence, $\Pr(y > 0 \mid s = 1) = (n_p + \pi n_u)/(n_p + n_u)$.

Yet, given the relation between Y and Z, $\Pr(z > 0, y > 0 \mid s = 1) = n_p/(n_p + n_u)$. Consequently:

$$\Pr(z > 0 \mid y > 0, s = 1, \boldsymbol{x}) = \frac{n_p}{n_p + \pi n_u}. \tag{3.18}$$

Further, $\Pr(z > 0 \mid y > 0, s = 1) = 0$ because all $z > 0$ in the data must occur for $y > 0$.

Combining (3.10) with (3.18), after some manipulations:

$$\Pr(z > 0 \mid s = 1, \boldsymbol{x}) = 0 + \frac{\frac{n_p}{\pi n_u} \exp\{\boldsymbol{x\beta}\}}{1 + \left(1 + \frac{n_p}{\pi n_u}\right) \exp\{\boldsymbol{x\beta}\}}.$$

Then, the explicit form of the *observed likelihood* for the presence-only data is given by:

$$\begin{aligned}
L(\boldsymbol{\theta}; \boldsymbol{z}, \boldsymbol{X}) &= \prod_{i \in S_u} \Pr(z_i = 0 \mid s_i = 1, \boldsymbol{x}_i) \qquad (3.19)\\
&\times \prod_{i \in S_p} \left[\Pr(z_i > 0 \mid s_i = 1, \boldsymbol{x}_i) f(Z_i \mid z_i > 0, s_i = 1, \boldsymbol{x}_i)\right]\\
&= \prod_{i \in S_u} \left[\frac{1 + \exp\{\boldsymbol{x}_i\boldsymbol{\beta}\}}{1 + \left(1 + \frac{n_p}{\pi n_u}\right) \exp\{\boldsymbol{x}_i\boldsymbol{\beta}\}}\right]\\
&\times \prod_{i \in S_p} \left[\frac{\frac{n_p}{\pi n_u} \exp\{\boldsymbol{x}_i\boldsymbol{\beta}\}}{1 + \left(1 + \frac{n_p}{\pi n_u}\right) \exp\{\boldsymbol{x}_i\boldsymbol{\beta}\}} f(z_i \mid z_i > 0, s_i = 1, \boldsymbol{x}_i)\right]\\
&= \prod_{i \in S_u} \left[\frac{1 + \exp\{\boldsymbol{x}_i\boldsymbol{\beta}\}}{1 + \left(1 + \frac{n_p}{\pi n_u}\right) \exp\{\boldsymbol{x}_i\boldsymbol{\beta}\}}\right]\\
&\times \prod_{i \in S_p} \left[\frac{\frac{n_p}{\pi n_u} \exp\{\boldsymbol{x}_i\boldsymbol{\beta}\}}{1 + \left(1 + \frac{n_p}{\pi n_u}\right) \exp\{\boldsymbol{x}_i\boldsymbol{\beta}\}} \frac{1}{z_i\sqrt{2\pi}\sigma} \exp\left\{\frac{(\ln(z_i) - \boldsymbol{x}_i\boldsymbol{\alpha})^2}{2\sigma^2}\right\}\right].
\end{aligned}$$

where $\boldsymbol{\theta}$ is a short-hand notion for the parameters at stake and 1_C is the indicator function for condition C.

3.2.1 Inference

Priors

The prior distributions of the model of presence-only semicontinuous data are specified below:

$$p(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma, \pi) = p(\boldsymbol{\beta}) p(\boldsymbol{\alpha}) p(\sigma) p(\pi), \tag{3.20}$$

where $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\boldsymbol{\beta}})$ and $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\boldsymbol{\alpha}})$ with $\Sigma_{\boldsymbol{\beta}}$ and $\Sigma_{\boldsymbol{\alpha}}$ fixed. Prior specification is completed letting $p(\sigma)$ be a $\mathcal{IG}(a_\sigma, b_\sigma)$ and let $p(\pi)$ be a $Beta(a_\pi, b_\pi)$ that summarizes available information on the uncertainty on the true value of the prevalence (see Section 3.1.1).

Model fit

In order to approximate the posterior distribution, an ad-hoc MCMC sampling scheme adapted from Diebolt & Robert (1994) is used. Hence, the data are augmented making use of the latent observations y_i, and the *complete likelihood*, $L(\boldsymbol{\theta}; \boldsymbol{z}, \boldsymbol{y}, \boldsymbol{X})$ is consequently derived.

Using a conditioning argument

$$\begin{aligned} L(\boldsymbol{\theta}; \boldsymbol{z}, \boldsymbol{y}, \boldsymbol{X}) &= \prod_{i \in S} \Pr(Y_i, Z_i \mid s_i = 1, \boldsymbol{x}_i) \\ &= \prod_{i \in S} \Pr(Z_i \mid Y_i, s_i = 1, \boldsymbol{x}_i) \Pr(Y_i \mid s_i = 1, \boldsymbol{x}_i) \\ &= \prod_{i \in S} \Pr(Y_i \mid s_i = 1, \boldsymbol{x}_i) \\ &= \prod_{i \in S} \Pr(y_i = 0 \mid s_i = 1, \boldsymbol{x}_i)^{1\{y_i=0\}} \\ &\times \prod_{i \in S} \Pr(y_i > 0 \mid s_i = 1, \boldsymbol{x}_i) f(Y_i \mid y_i > 0, s_i = 1, \boldsymbol{x}_i)^{1\{y_i>0\}}. \end{aligned} \tag{3.21}$$

The form of $\Pr(Y_i \mid s_i = 1, \boldsymbol{x}_i)$ follows directly from (3.10). Also, $\Pr(y > 0 \mid \boldsymbol{x}, s = 1)$ and $\Pr(y = 0 \mid s = 1, \boldsymbol{x})$ are the same expressed in (3.7)) and (3.8). Then:

$$\Pr(y > 0 \mid s = 1, \boldsymbol{x}) = \frac{\exp\{\boldsymbol{x}_i\boldsymbol{\beta}\}\left(1 + \frac{n_p}{\pi n_u}\right)}{1 + \left(1 + \frac{n_p}{\pi n_u}\right)\exp\{\boldsymbol{x}_i\boldsymbol{\beta}\}} \tag{3.22}$$

and

$$\begin{aligned} \Pr(y = 0 \mid s = 1, \boldsymbol{x}) &= 1 - \Pr(y > 0 \mid s = 1, \boldsymbol{x}) \\ &= \frac{1}{1 + \left(1 + \frac{n_p}{\pi n_u}\right)\exp\{\boldsymbol{x}_i\boldsymbol{\beta}\}}. \end{aligned} \tag{3.23}$$

The density $f(Y \mid y > 0, s = 1, \boldsymbol{x})$, due to 3.12, is given by:

$$f(Y \mid y > 0, s = 1, \boldsymbol{x}) = \frac{1}{y\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma^2}\left(\ln(y) - \boldsymbol{x}\boldsymbol{\alpha}\right)^2\right\} \tag{3.24}$$

Then, the explicit form of the *complete likelihood* for the presence-only semicontinuous data, in terms of both $\boldsymbol{Z}$ and $\boldsymbol{Y}$, can be obtained by substituting (3.22) and (3.23) into (3.21):

$$\begin{aligned} L(\boldsymbol{\theta}; \boldsymbol{z}, \boldsymbol{y}, \boldsymbol{X}) &= \prod_{i \in S} \Pr(y_i = 0 \mid s_i = 1, \boldsymbol{x}_i)^{1_{\{y_i=0\}}} \\ &\times \prod_{i \in S} \Pr(y_i > 0 \mid s_i = 1, \boldsymbol{x}_i) f(Y_i \mid y_i > 0, s_i = 1, \boldsymbol{x}_i)^{1_{\{y_i>0\}}} \\ &= \prod_{i \in S}\left[\frac{1}{1 + \left(1 + \frac{n_p}{\pi n_u}\right)\exp\{\boldsymbol{x}_i\boldsymbol{\beta})\}}\right]^{1_{\{y_i=0\}}} \\ &\times \prod_{i \in S}\left[\frac{\left(1 + \frac{n_p}{\pi n_u}\right)\exp\{\boldsymbol{x}_i\boldsymbol{\beta}\}}{1 + \left(1 + \frac{n_p}{\pi n_u}\right)\exp\{\boldsymbol{x}_i\boldsymbol{\beta}\}}\frac{1}{y_i\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma^2}\left(\ln(y_i) - \boldsymbol{x}_i\boldsymbol{\alpha}\right)^2\right\}\right]^{\{y_i>0\}} \end{aligned} \tag{3.25}$$

where $\boldsymbol{\beta}$ represents the vector of k parameters used for the logistic part and $\boldsymbol{\alpha}$ represents the vector of h parameters used for the linear regression part.

The general iteration of the (Metropolis within) Gibbs sampling scheme is detailed in Algorithm 2.

Algorithm 2 Gibbs sampling scheme

1. Sample the latent variables Y_i from $p(Y_i \mid Z_i, s_i = 1, \boldsymbol{x}_i)$, $i = 1, \dots, n$; where $p(Y_i \mid Z_i, \boldsymbol{x}_i) = 1_{y_i = z_i} * 1_{z_i > 0} + 1_{z_i = 0} p(Y_i \mid z_i = 0, s_i = 1, \boldsymbol{x}_i)$. That is, simply one sets $y_i = z_i$ when $z_i > 0$ and when $z_i = 0$ note that
$$p(Y_i \mid z_i = 0, s_i = 1, \boldsymbol{x}_i) = p(Y_i \mid s_i = 1, \boldsymbol{x}_i),$$
since is assumed that data are sampled uniformly at random from the study area. Sampling of Y_i when $z_i = 0$ must then be performed in two steps, since $p(Y_i \mid s_i = 1, \boldsymbol{x}_i)$ is a mixed measure. First, one shall sample a presence from $p(y_i > 0 \mid s_i = 1, \boldsymbol{x}_i)$, and then set $Y_i = y_i$, where y_i is sampled from $p(Y_i \mid y_i > 0, \boldsymbol{x}_i)$.

2. Sample the regression parameters ($\boldsymbol{\beta}$) for the logistic part from
$$p(\boldsymbol{\beta} \mid \boldsymbol{y}; \boldsymbol{X}) \propto p(\boldsymbol{\beta}) \frac{\exp\{\sum_i^n 1_{y_i > 0} \boldsymbol{x}_i \boldsymbol{\beta}\}}{1 + \left(1 + \frac{n_p}{\pi n_u}\right) \exp\{\sum_i^n 1_{y_i > 0} \boldsymbol{x}_i \boldsymbol{\beta}\}}.$$

3. Sample the remaining parameters $\boldsymbol{\alpha}$ and σ simultaneously as:
$$p((\boldsymbol{\alpha}, \sigma) \mid \boldsymbol{y}; \boldsymbol{X}) \propto p(\boldsymbol{\alpha}, \sigma) L(\boldsymbol{\theta} \mid \boldsymbol{z}, \boldsymbol{y}; \boldsymbol{X}).$$

4. Sample π from its prior.

At Step 1 of the MCMC algorithm, latent indicators are sampled. Here, a latent variable, Y_i, which is not discrete is sampled. Then, it is necessary to augment generating Y_i from its semicontinuous full conditional distribution when $z_i = 0$. When $z_i > 0$, Y_i is not sampled since its full conditional is a point mass on y_i. Convergence of the chain is guaranteed from the fact that $f(Y_i \mid Z_i, s_i = 1, \boldsymbol{x}_i)$, albeit arising from a unusual semicontinuous distribution, is exactly the full conditional for Y_i. All required regularity are consequently implied by model assumption.

After to have sampled or set values for Y, Step 2 arises from straightforward conditional independence conditions, (3.10) and (3.19).

At Step 3, several difficulties are associated with setting up Metropolis Hastings (MH) steps for the parameters $(\boldsymbol{\alpha}, \sigma)$. Key to success for MH is linked to a clever choice for the candidate transition kernel, which does not seem readily available here. Furthermore, the last full conditional distribution is also potentially multimodal, and even if a good candidate transition kernel were available, tuning of MH would be made harder by volatility in the latent indicators $1_{y_i>0}$. In order to avoid difficulties linked with setting up MH, it is possible to sample $(\boldsymbol{\alpha}, \sigma)$ simultaneously with Adaptive Rejection Metropolis Sampling (Gilks et al. (1995)). In this contest the ARMS works nicely, and needs essentially no tuning.

For sampling the logistic regression parameters at Step 2 an Adaptive Rejection Metropolis Sampling (Gilks et al. (1995)) is again used for simplicity, even if, as discussed in Section 3.1.1 there are many different alternative approaches for this standard problem.

Yet, for simplicity the prevalence π is sampled from its prior.

3.3 Simulation Study

The performance of the proposed model is investigated in this section on simulated data. The EM algorithm of Ward et al. (2009) is extended to abundance data, in order to obtain maximum likelihood estimates for comparison.

A semicontinuous response Y is generated from the following two-part model

$$\operatorname{logit}\Pr(y > 0 \mid x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

and, conditionally on $y > 0$ and x_3

$$\ln(Y) = \alpha_0 + \alpha_1 x_3 + \varepsilon,$$

where $\beta_0 = -4.5$, $\beta_1 = 3$, $\beta_2 = 2$, $\alpha_0 = 0.3$, $\alpha_1 = 1$, and ε is sampled from a standard normal. The covariates are generated independently as follows: x_1 is sampled from a Bernoulli with parameter 0.2, mimicking a categorical predictor, and the other two covariates are generated from standard normals. At each replication we generate a study area of N observations, and randomly select a proportion λ of the observed presences for the sample of presences used for model fitting. Then, the pseudo-absences sample is drawn from the remaining data, and fit the Bayesian and the oracle classical model in which the prevalence is correctly known. Remark that this sampling procedure is in according to the sampling assumption set out in this thesis only when N is large and π is small. For the Bayesian approach, the following priors are used: for the logistic and regression coefficients, normal zero centered priors with variance equal to 25 and 9 respectively; an exponential for the precision parameter (i.e., the inverse of σ), and a Beta with parameters 0.6 and 5 for prevalence. Data generation and model fitting are replicated 1000 times, and the average results are reported over the 1000 replications.

In Table 3.2 the average Relative Root Mean Squared Error (RRMSE) of the parameter estimates of Bayesian and EM algorithm for different values of N and λ is reported. RRMSE is calculated as ratio of RMSE to true parameter value.

	Bayesian Model				EM algorithm			
	$N = 10^4$	$N = 10^4$	$N = 900$	$N = 900$	$N = 10^4$	$N = 10^4$	$N = 900$	$N = 900$
	$\lambda = 10\%$	$\lambda = 30\%$	$\lambda = 10\%$	$\lambda = 30\%$	$\lambda = 10\%$	$\lambda = 30\%$	$\lambda = 10\%$	$\lambda = 30\%$
Parameters								
β_1	0.16	0.17	0.22	0.21	0.16	0.13	0.30	0.46
β_2	0.23	0.23	0.27	0.37	0.15	0.12	0.29	0.47
α_0	0.09	0.10	0.30	0.34	0.09	0.10	0.30	0.35
α_1	0.03	0.03	0.09	0.10	0.03	0.03	0.09	0.11
σ	0.02	0.02	0.07	0.11	0.29	0.29	0.30	0.28
π	0.07	0.07	0.07	0.07	-	-	-	-
95% CI π, L	0.1367	0.1361	0.1373	0.1424	-	-	-	-
95% CI π, C	1.0000	1.0000	1.0000	1.0000	-	-	-	-

Table 3.2: RRMSE of the parameter estimates of Bayesian model and EM model or different values of N and λ in simulated data. We omit β_0 since it is summarized in the final prevalence estimate. The last two lines report the mean length (L) and coverage (C) of the 95% CI for the prevalence parameter π. The number of replications is 1000.

For the proposed Bayesian method it can be seen that, as expected, the RRMSE decreases with N. On the other hand, there does not seem to be a strong dependence on λ, indicating that it does not really matter how many presences are obtained, as long as these are sampled independently and uniformly from the study area and the final sample size is large enough. The EM algorithm seems to be dependent both on N and λ, and it is sometimes outperformed by the Bayesian approach even if it has the unfair advantage of assuming a known, and correct, prevalence. The RRMSE for the regression coefficients are in general comparable, but the Bayesian approach seems to work much better than the frequentist method in estimating σ. This is due to a negative bias in the estimate of σ obtained with the EM algorithm, which could be explained by the optimism in assuming a known prevalence. The same assumption seems to lead to a smaller RRMSE in the coefficients of the logistic part when N is large, and larger when N is small. In Table 3.2 we also show the mean length of the 95% credibility intervals and their frequentist coverage for π. It can be seen that the frequentist coverage is very large, and that the mean length is large too and reflects prior inputs (i.e., a much smaller mean length could be obtained with a more concentrated prior).

A further comparison between the Bayesian and frequentist method is given in Table 3.3, where the predictive performance of the methods are compared. At each replication the RRMSE for positive predictions is calculated, and the predictive performance of the presence-absence part of the model computing sensitivity and specificity is summarized.

Test	N	λ	Bayesian Model	EM algorithm
RRMSE	10^4	10%	0.01	0.00
	10^4	30%	0.00	0.00
	900	10%	0.05	0.00
	900	30%	0.00	0.00
Sensitivity	10^4	10%	0.84	0.84
	10^4	30%	0.84	0.84
	900	10%	0.84	0.84
	900	30%	0.83	0.83
Specificity	10^4	10%	0.16	0.16
	10^4	30%	0.16	0.16
	900	10%	0.16	0.16
	900	30%	0.17	0.17

Table 3.3: RRMSE for positive predictions, sensitivity and specificity of the predicted presence/absence for different values of N and λ in simulated data. The results are averaged over 1000 replications.

The appears to be no difference between (oracle) frequentist and the Bayesian approach, in both cases the predicted values are very close to the observed values, sensitivity is rather large and specificity is rather small.

Finally, a small study is provide to evaluate sensitivity of the parameter estimates to the choice of priors. In Table 3.4 the RRMSE of the parameters when $N = 10000$ and $\lambda = 0.1$ for additional sets of priors is shown. Prior set (a) is the set used for the previous simulations and described at the beginning of the section. In prior set (b) a bias to the prior set (a) is added. The priors for logistic and regression coefficients are centered on -0.5, and further a Gamma with parameters 1.5 and 1 for the precision parameter is used. Prior set (c) is equivalent to (a), with the exception of the prior for the prevalence parameter, where a Beta distribution with parameters 0.46 and 2.64 is used; and finally in prior set (d) a zero centered Student's T distributions with three degrees of freedom for the β and α parameters is used.

	Prior settings			
Parameters	(a)	(b)	(c)	(d)
β_1	0.16	0.15	0.15	0.16
β_2	0.23	0.23	0.24	0.23
α_0	0.09	0.09	0.09	0.09
α_1	0.03	0.03	0.03	0.03
σ	0.02	0.02	0.02	0.02
π	0.07	0.07	0.07	0.07

Table 3.4: Sensitivity analysis: RRMSE obtained with (a) default priors, (b) biased priors, (c) biased prior on the prevalence parameter, (d) flat priors. The results are based on 1000 replications.

It can to see that there does not seem to be prior sensitivity with the sample sizes common encountered in real data applications.

3.4 Application to real data

Taxus baccata is a relict of the Cenozoic flora, characterized by warm-humid climatic conditions. It survived glaciations in refugia areas, and may have followed *Fagus* in successive postglacial expansions. This process has determined the current fragmented presence and reduced consistency. *Taxus baccata* has low resistance to intense cold and it probably survived mainly thanks to the ability of asexual reproduction and sex variations of adults in case of need.

The data used was recorded in a study area located in central Italy, with specific reference to Abruzzo and Lazio regions. The area of interest extends for about 28000 Km^2, with an heterogenous morphology, which includes sandy coasts and the summits of the Apennines (the highest peak being the Gran Sasso, 2912 m of altitude). The forest habitat of *Taxus baccata* in these two regions is of high conservation priority in Europe (Scarnati et al. 2009).

The aim of the analysis is to obtain a map of the potential distribution of *Taxus baccata*, through climatic, topographic, structural and environmental parameters. This map is then used for elaborating conservation strategies (Guisan & Zimmermann 2000).

Climatic maps in GRID format, with a spatial resolution of 500 m, were built. These maps were obtained by interpolating precipitation and temperature data recorded in 300 meteorological stations and calculating the average data for the $1960 - 1990$ period (see Attorre, Alfó, De Sanctis, Francesconi & Bruno (2007) for technical details).

The environmental covariates considered were:

MIN_T_1 Minimum temperature of the coldest month (January)

MAX_T_7 Maximum temperature of the hottest month (July)

T_MED Average temperature in twelve consecutive months

$TOTAL\_P$ Total annual precipitation

$SUMM\_P$ Precipitation during summer

$WINT\_P$ Precipitation during winter

$MOISTURE$ Moisture index

ALT Altitude

Descriptive analysis of these are summarized in Table 3.5(a) for the entire grid, and in Table 3.5(b) for the plots in which a presence was recorded. Note that in Table 3.5(a) only on suitable locations for proliferation of *Taxus Baccata* are reported (see below).

Temperatures are expressed in degree Celsius ($°C$), precipitations are expressed in millimetres (mm), moisture is expressed in $Mi = TOTAL\_P/ETp$, where ETp is the potential evapotraspiration and altitude is expressed in metres (m). As measure of abundance for Taxus the Importance Value (IV) expressed on relative basal area and the number of stems contained within each plot (Basal area/ha) is used. See Attorre, Alfó, De Sanctis, Francesconi & Bruno (2007) for technical details.

(a)

	Min	$1st$ Quartile	Median	Mean	$3rd$ Quartile	Max	Std. Dev.
MIN_T_1 $(°C)$	-6	-3	-2	-2	-1	4	1
MAX_T_7 $(°C)$	18	22	24	23	25	28	2
T_MED $(°C)$	5	8	9	9	10	13	1
$TOTAL\_P$ (mm)	629	1029	1189	1211	1403	1894	245
$SUMM\_P$ (mm)	91	145	165	170	191	292	33
$WINT\_P$ (mm)	153	304	363	375	447	706	95
$MOISTURE$ (Mi)	0.9	1.2	1.3	1.3	1.4	2.3	0.2
ALT (m)	900	1035	1217	1244	1424	1750	235

(b)

	Min	$1st$ Quartile	Median	Mean	$3rd$ Quartile	Max	Std. Dev.
MIN_T_1 $(°C)$	-4	-3	-2	-2	-1	2	1
MAX_T_7 $(°C)$	20	21	22	22	24	25	1
T_MED $(°C)$	6	8	8	8	9	12	1
$TOTAL\_P$ (mm)	22	1251	1405	1414	1620	1696	206
$SUMM\_P$ (mm)	141	161	203	196	219	254	31
$WINT\_P$ (mm)	256	336	440	429	507	560	86
$MOISTURE$ (Mi)	1.1	1.3	1.3	1.4	1.4	1.7	0.2
ALT (m)	969	1278	1430	1392	1503	1715	157
$ABUNDANCE$	1	7	12	20	30	78	18

Table 3.5: Descriptive statistics for the environmental covariates on the whole data (a) and for locations where abundance is positive (b). MIN_T_1: Minimum temperature of the coldest month (January). MAX_T_7: Maximum temperature of the hottest month (July). T_MED: Average temperature in twelve consecutive months. $TOTAL\_P$: Total annual precipitation. $SUMM\_P$: Precipitation during summer. $WINT\_P$: Precipitation during winter. $MOISTURE$: Moisture index. ALT: Altitude.

Locations with presence of *Taxus baccata* were identified by GPS coordinates, and selected through bibliographical information and indications of the staff of the protected areas. There are many indices of abundance which could be used. In this research the Importance Value (IV) is used, for a definition of which the reader is pointed to Scarnati et al. (2009). In each selected location the IV of *Taxus baccata* was measured based equally on relative basal area and the number of stems contained within it.

In this study 97 presences are been observed, and need to build predictions for a total of 111882 locations. A few of these 111882 locations are excluded from the analysis because they almost surely correspond to locations in which the species is absent: GIS tools are used to discard completely unsuitable locations due for instance to presence of lakes, cities, roads, and so on. Also, sites where one or more of habitat characteristics assume values that do not allow the plant growth are discarded.

In order to obtain information on prevalence ecologists and experts are been independently consulted, asking them a rough estimate of their expected prevalence, a minimum and a maximum. Also, the estimates of prevalence of *Taxus Baccata* are recorded and similar species obtained in previous studies dedicated at least to part of the area under consideration. A consensus was obtained on a prevalence between 2% and 6%. Consequently it is decided to conservatively center the prior on 0.03. Since the majority of the consulted sources indicated a prevalence of at most 5%, it is also decided to let the third quartile of the prior be slightly smaller than 0.05; and to have a 0.95 upper quantile of approximately 10%, an upper limit common to many of the considered sources. Given these information, a Beta prior was elicited with parameters 0.6 and 19.4, which has a mean of 0.03, a third quartile slightly larger than 0.04, and a .95 upper quantile of around 0.10.

For regression coefficients there are two default prior choices in practice: a zero-centered Gaussian with diagonal covariance matrix, and a zero-centered Gaussian with covariance matrix $\gamma X'X$, where X is the matrix of covariates used in the model. Then the priors for the remaining parameters was set by fixing $\Sigma_\beta = \Sigma_\alpha = \sigma^2 I$, where I denotes a diagonal matrix of the appropriate size; and center the prior for σ on its maximum likelihood estimate. A diagonal covariance matrix was preferred since it attenuates the final correlation between estimates, i.e., collinearity; plus, it was also avoid the arbitrary choice of the hyperparameter parameter γ.

In order to reduce spurious effects, the pseudo-absence generation was repeated 40 times, and the model fitted separately on each data set. At each repetition, 97 pseudo-absences was sampled from the suitable sites with the case-control approach of Attorre, Francesconi, Taleb, Scholte, Saed, Alfó & Bruno (2007), select at random starting values for the parameters, and run Algorithm 1 for a total of 100000 sweeps. A burn-in of 50000 iterations was allowed, and one each twentieth of the 50000 remaining iterations for posterior estimation was used. The perform model is chosen according to the structured stochastic search variable section (SSSVS) approach of Farcomeni (2010), to which the reader is pointed for details. The possibility of including any of the available covariates is considered, plus all two-way interactions, in each part of the model. The hierarchical constraints was used so that an interaction is not included in a model without both covariates contributing to it. SSSVS allows to estimate a probability of inclusion for each coefficient. As proved by Farcomeni (2010), consistency in model choice is achieved as long as covariates with a probability of inclusion larger than 50% are used in the model, and the other covariates are discarded. The 40 repetitions did not provide conflicting conclusions, so that probably no spurious effects was observed in the sampled pseudo-absences. The results related to a single (randomly chosen) repetition are provided.

In Table 3.6 the posterior means of each covariate included in the final model chosen with SSSVS and individual probabilities of inclusion is shown. All other covariates, including the interactions, have an estimated probability of inclusion smaller than 50%, and therefore are omitted from the final model.

Model	Parameters	Posterior Mean	Std. Err.	Prob. Inclusion	EM
Logistic	intercept	1.82	0.112	-	-0.80
	TOTAL_P	0.94	0.100	0.85	0.86
	MIN_T_1	-0.52	0.113	0.79	-0.24
	MAX_T_7	0.65	0.105	0.80	0.53
	ALT	0.66	0.126	0.81	0.82
Regression	intercept	2.88	0.013	-	2.98
	TOTAL_P	0.17	0.008	0.99	0.14
	MIN_T_1	-0.29	0.010	0.75	-0.34
	MOISTURE	-0.08	0.009	0.65	-0.09
	ALT	-0.65	0.012	0.98	-0.76
	σ	0.93	0.003	-	0.66
	π	0.03	0.005	$95\%CI : (0.000 - 0.136)$	-

Table 3.6: Posterior mean, estimated standard error and probability of inclusion for each covariate included in the final model after SSSVS; plus maximum likelihood estimates obtained with EM algorithm for comparison.

Note that the parameter estimates should not be directly interpreted due to collinearity. The correlation matrix between the covariates included in the final model is reported in Table 3.7:

	TOTAL_P	MAX_T_7	MIN_T_1	ALT	MOISTURE
TOTAL_P	1.00	-0.09	0.14	0.12	0.14
MAX_T_7	-0.09	1.00	0.42	-0.90	0.42
MIN_T_1	0.14	0.42	1.00	-0.57	1.00
ALT	0.12	-0.90	-0.57	1.00	-0.57
MOISTURE	0.14	0.42	1.00	-0.57	1.00

Table 3.7: Correlation between covariates used in the final model.

The estimates obtained with the proposed model are compared with the maximum likelihood estimates obtained with an EM algorithm along the lines of Ward et al. (2009), assuming a known prevalence of 3%. It can be seen from Table 3.6 that final parameter estimates are comparable, especially with respect to the regression part of the model. The logistic part of the frequentist model is dependent on the assumptions related to the prevalence, which must be assumed known with the EM approach. Note further that the variance estimated with the maximum likelihood approach is slightly smaller than the posterior mean for the variance, and it has been observed in the simulation study that EM with known prevalence tends in fact to under estimate the variance of the continuous part of the model. Note finally that the posterior summaries for the prevalence parameters are essentially equivalent to the prior summaries, as the data contain very little information on prevalence.

The predictive performance of the Bayesian model is validated by building $1-\alpha = 0.95$ predictive intervals for the observed presences. Finally a prediction coverage probability of 0.948 is obtained, so that it is possible to claim the model valid from a predictive point of view. A strong prior sensitivity has not experienced, and results equivalent for practical purposes have obtained by varying the prior assumptions in a reasonable range.

In Figure 3.1 a map of the potential distribution of the abundance of *Taxus baccata* built using GIS tools is shown. The predictions in Figure 3.1 minimize the posterior expected loss.

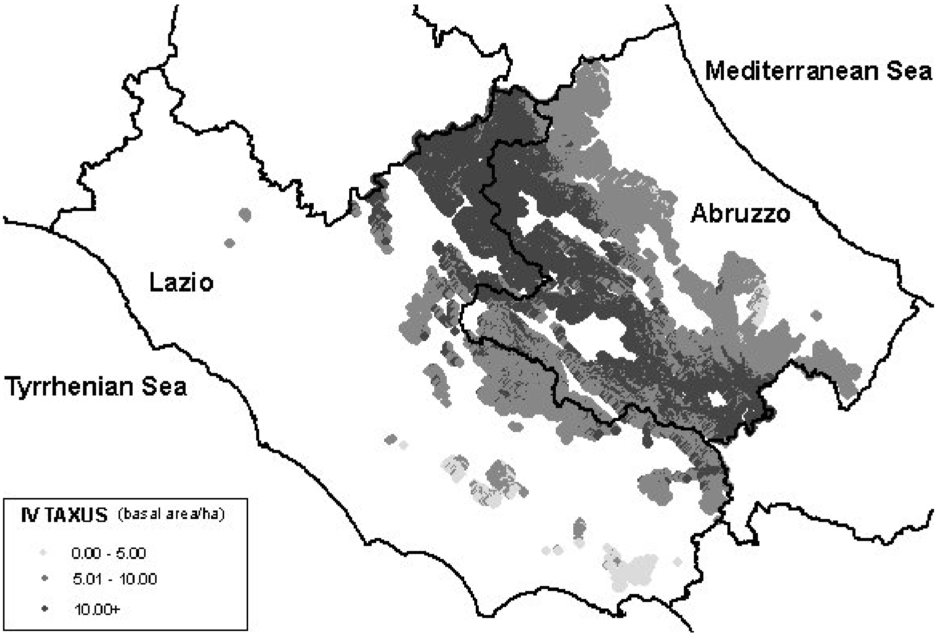


Figure 3.1: Potential distribution of the abundance of *Taxus baccata.* $R^2 = 0.18$, False Negative Rate=0

As an additional measure of goodness of fit the R^2 and the False Negative Rate (FNR) was calculated. These are equal to 0.18 and 0, respectively. The same measures are calculated for the maximum likelihood estimates, and $R^2 = 0.20$ and FNR= 0.16 is obtained. For comparison the same measures have computed with different distributions for the continuous component, obtaining similar results. It seems like peaks of large abundance are not captured well by the model, with a strong regression to the mean effect. Consequently distributions allowing for larger skewness are trained, but these did not seem to fit the data well overall. The peaks of abundance actually correspond to areas in which *Taxus baccata* was planted and is currently nurtured and protected by human intervention, and it would not have been so abundant otherwise. Then the R^2 is not large due to the fact that important covariates were not measured, rather than because of the log-normal distribution not approximating well the data. However in this application the aim is not to obtain a correct prediction of the actual abundance, but only of its potential distribution. Recall finally that the log-normal distribution is validated by the prediction coverage probability.

The estimated potential distribution in Figure 3.1 leads us to conclude that *Taxus* is potentially situated at both a higher and lower altitude than expected. The first behavior (higher altitudes) is likely due to a retreating process to areas less accessible by livestock (for instance, cows). The second behavior (lower altitudes) has been seen in areas with a high moisture index (e.g., close to lakes in the Northwestern and

Southwestern Lazio), which makes the area more suitable for a presence of *Taxus*.

Further, *Taxus* is more likely to be common on the western Tyrrhenian side, where the temperatures are higher (with respect to the eastern Adriatic side of the area). The same reasoning applies to the regions of the area in the central part of the map, which are facing South.

The focus is now on the locations corresponding to protected area (Special Protection Zone) ZPS12, in Monti Lepini, Lazio; established by European Community directive 79/409/CEE. 363 locations corresponding to area ZPS12 and compatible with a presence of *Taxus* are selected, and the posterior probability of observing a presence ($\Pr(IV > 0)$) and a moderately large abundance ($\Pr(IV > 2)$) are considered. Descriptive statistics for these probabilities computed at the 363 locations of special interest are reported in Table 3.8.

	Min	$1st$ Quartile	Median	Mean	$3rd$ Quartile	Max	Std. Dev.
$\Pr(IV > 0)$	0.86	0.93	0.95	0.94	0.96	0.98	0.02
$\Pr(IV > 2)$	0.23	0.41	0.58	0.56	0.73	0.83	0.17

Table 3.8: Descriptive statistics for the 363 posterior estimated probabilities of a positive and of a moderately large abundance in the Special Protection Zone ZPS12.

It can be argued that *Taxus* is very likely to be present in the entire area, but only in few locations an high IV is expected. About one quarter of locations with $\Pr(IV > 2) > 0.7$ are estimated, indicating that these locations are highly suitable for *Taxus*.

These results were used to select locations for conservation actions. In areas were a high suitability for *Taxus* was predicted two projects aimed at the construction of fences to protect its regeneration from livestock have recently started.

Chapter 4

Data Augmentation Approach in Bayesian Modeling of Presence-only Data

In this Chapter a second Bayesian model to estimate logistic linear regressions adapted to presence-only data is proposed. This work is published as:

> Divino, F., Golini, N., Jona Lasinio, G. and Penttinen A. (2011). Data Augmentation Approach in Bayesian Modeling of Presence-only Data. *Procedia Environmental Sciences*, **7**, 38 – 43.

Here a random approximation of the correction factor in the adjusted model (4.3) allows to overcome the need to acquire strong information on the population prevalence. The model is based on the assumption (not always adequate) that the environmental covariates are the only determinants of species distributions.

4.1 Modeling method

Let Y be a binary random variable measuring the presence-absence of a given species, such that $y = 1$ if the species is observed at location and $y = 0$ if not, and let the ratio defined in (2.10):

$$\frac{\tilde{\gamma}_1}{\tilde{\gamma}_0} \quad = \quad \frac{\tilde{n}_{1u} + n_p}{n_u - \tilde{n}_{1u}} \frac{1 - \pi}{2\pi}.$$

Let π_u be the proportion of presences in S_u. Because of the censoring effect acting on S_u, π_u is not observable and it can be represented by a random quantity, i.e. the random sample prevalence in S_u:

$$\tilde{\pi}_u = \frac{\tilde{n}_{1u}}{n_u}. \tag{4.1}$$

Also, being S_u a random sample from $\mathcal{U}$, when S_u reflects the composition of $\mathcal{U}$ or when n_u (the size of S_u) tends to N (the size of $\mathcal{U}$), π_u tends to π and $\tilde{\pi}_u$ represents an unbiased estimator of π.

Given (4.1), the ratio in (2.10) can be rewritten as follows

$$\begin{aligned}\frac{\tilde{\gamma}_1}{\tilde{\gamma}_0} &= \frac{\tilde{n}_{1u} + n_p}{n_u - \tilde{n}_{1u}} \frac{1-\pi}{2\pi} \\ &\approx \frac{\tilde{n}_{1u} + n_p}{n_u - \tilde{n}_{1u}} \frac{1-\tilde{\pi}_u}{2\tilde{\pi}_u} \\ &= \frac{\tilde{n}_{1u} + n_p}{n_u - \tilde{n}_{1u}} \frac{1-\frac{\tilde{n}_{1u}}{n_u}}{2\frac{\tilde{n}_{1u}}{n_u}} \\ &= \frac{\tilde{n}_{1u} + n_p}{n_u - \tilde{n}_{1u}} \frac{\frac{n_u-\tilde{n}_{1u}}{n_u}}{2\frac{\tilde{n}_{1u}}{n_u}} \\ &= \frac{\tilde{n}_{1u} + n_p}{2\tilde{n}_{1u}}.\end{aligned} \tag{4.2}$$

Remark that in (4.2) the direct effect of the population prevalence vanishes and that, calculated $\tilde{n}_{1u}$, the ratio is identified. Then a Bayesian model can be defined and $\tilde{\pi}_u$ can be introduced, indirectly, in a MCMC algorithm a step of data augmentation.

4.1.1 Observed Likelihood

As introduced in Chapter 3, to handle the presence-only model two approaches are available. The first one is based on the observed likelihood defined on the naive process $\boldsymbol{Z}$, and the second one considers the complete likelihood, i.e. the joint probability of the $\boldsymbol{Y}$ and $\boldsymbol{Z}$ processes.

The estimation procedure works for both approaches and then the Bayesian model will be represented using either the observed or complete likelihood. Let start to write the analytic expression of $\Pr(Y \mid s=1, \boldsymbol{x})$ and $\Pr(Z \mid s=1, \boldsymbol{x})$. Given the ratio in (4.2), the likelihood functions are obtained following the same procedure stated in Section 2.1.

First the observed likelihood $L(\boldsymbol{\theta}; \boldsymbol{z}, \boldsymbol{X})$ is considered:

$$\begin{aligned} L(\boldsymbol{\theta}; \boldsymbol{z}, \boldsymbol{X}) &= \prod_{i \in S} \Pr(Z_i \mid s_i = 1, \boldsymbol{x}_i) \\ &= \prod_{i \in S_u} \Pr(z_i = 0 \mid s_i = 1, \boldsymbol{x}_i) \prod_{i \in S_p} \Pr(z_i = 1 \mid s_i = 1, \boldsymbol{x}_i) \end{aligned}$$

where $\boldsymbol{\theta}$ is a short-hand notation to denote the parameters at stake. The expression of $\Pr(z = 1 \mid s = 1, \boldsymbol{x})$ is given in (3.2) according to the ratio in (4.2):

$$\begin{aligned} \Pr(z = 1 \mid s = 1, \boldsymbol{x}) &= \Pr(z = 1 \mid y = 1, s = 1, \boldsymbol{x}) \Pr(y = 1 \mid s = 1, \boldsymbol{x}) \\ &+ \Pr(z = 1 \mid y = 0, s = 1, \boldsymbol{x}) \Pr(y = 0 \mid s = 1, \boldsymbol{x}) \\ &= \Pr(z = 1 \mid y = 1, s = 1) \Pr(y = 1 \mid s = 1, \boldsymbol{x}) \\ &+ \Pr(z = 1 \mid y = 0, s = 1) \Pr(y = 0 \mid s = 1, \boldsymbol{x}) \end{aligned}$$

because $Z \mid Y, \boldsymbol{x} \sim Z \mid Y$, and $\Pr(y = 1 \mid s = 1, \boldsymbol{x})$, according to the ratio in (4.2), can be rewritten as

$$\begin{aligned} \Pr(y = 1 \mid s = 1, \boldsymbol{x}) &= \frac{2 \exp\left\{\eta(\boldsymbol{x}) + \ln\left(\frac{\tilde{\gamma}_1}{\tilde{\gamma}_0}\right)\right\}}{1 + 2 \exp\left\{\eta(\boldsymbol{x}) + \ln\left(\frac{\tilde{\gamma}_1}{\tilde{\gamma}_0}\right)\right\}} \qquad (4.3) \\ &\approx \frac{2 \frac{\tilde{n}_{1u} + n_p}{2\tilde{n}_{1u}} \exp\{\eta(\boldsymbol{x})\}}{1 + 2 \frac{\tilde{n}_{1u} + n_p}{2\tilde{n}_{1u}} \exp\{\eta(\boldsymbol{x})\}} \\ &= \frac{\left(1 + \frac{n_p}{\tilde{n}_{1u}}\right) \exp\{\boldsymbol{x}\boldsymbol{\beta}\}}{1 + \left(1 + \frac{n_p}{\tilde{n}_{1u}}\right) \exp\{\boldsymbol{x}\boldsymbol{\beta}\}}. \end{aligned}$$

Also, given the relation between $\boldsymbol{Y}$ and $\boldsymbol{Z}$, showed in Table 2.2, it is straightforward to derive the conditional probabilities $\Pr(Z \mid Y, s = 1)$:

	$z = 0$	$z = 1$
$y = 0$	0	1
$y = 1$	$\frac{\tilde{n}_{1u}}{\tilde{n}_1}$	$\frac{n_{1p}}{\tilde{n}_1}$

Table 4.1: Conditional probabilities of $Z \mid Y, s = 1$.

Hence, $\Pr(Z \mid s=1,\boldsymbol{x})$ can be expressed as follows:

$$\begin{aligned}
\Pr(z=1 \mid s=1,\boldsymbol{x}) &= \Pr(z=1 \mid y=1, s=1)\Pr(y=1 \mid s=1,\boldsymbol{x}) \\
&+ \Pr(z=1 \mid y=0, s=1)\Pr(y=0 \mid s=1,\boldsymbol{x}) \\
&= \frac{\tilde{n}_p}{\tilde{n}_1}\Pr(y=1 \mid s=1,\boldsymbol{x}) + 0 \\
&= \frac{\tilde{n}_p}{\tilde{n}_1}\frac{2\exp\left\{\eta(\boldsymbol{x}) + \ln\left(\frac{\tilde{\gamma}_1}{\tilde{\gamma}_0}\right)\right\}}{1+2\exp\left\{\eta(\boldsymbol{x}) + \ln\left(\frac{\tilde{\gamma}_1}{\tilde{\gamma}_0}\right)\right\}} \\
&\approx \frac{\tilde{n}_p}{\tilde{n}_1}\frac{2\frac{\tilde{n}_{1u}+n_p}{2\tilde{n}_{1u}}\exp\left\{\eta(\boldsymbol{x})\right\}}{1+2\frac{\tilde{n}_{1u}+n_p}{2\tilde{n}_{1u}}\exp\left\{\eta(\boldsymbol{x})\right\}} \\
&= \frac{\frac{n_p}{\tilde{n}_{1u}}\exp\{\boldsymbol{x}\boldsymbol{\beta}\}}{1+\left(1+\frac{n_p}{\tilde{n}_{1u}}\right)\exp\{\boldsymbol{x}\boldsymbol{\beta}\}}
\end{aligned}$$

and

$$\begin{aligned}
\Pr(z=0 \mid s=1,\boldsymbol{x}) &= 1-\Pr(z=1 \mid s=1,\boldsymbol{x}) \\
&\approx \frac{1+\exp\{\boldsymbol{x}\boldsymbol{\beta}\}}{1+\left(1+\frac{n_p}{\tilde{n}_{1u}}\right)\exp\{\boldsymbol{x}\boldsymbol{\beta}\}}.
\end{aligned}$$

Then, the observed likelihood $L(\boldsymbol{\theta};\boldsymbol{z},\boldsymbol{X})$ can be written as follows:

$$\begin{aligned}
L(\boldsymbol{\theta};\boldsymbol{z},\boldsymbol{X}) &= \prod_{i\in S}\Pr(Z_i \mid s_i=1,\boldsymbol{x}_i) \qquad (4.4)\\
&= \prod_{i\in S_u}\Pr(z_i=0 \mid s_i=1,\boldsymbol{x}_i)\prod_{i\in S_p}\Pr(z_i=1 \mid s_i=1,\boldsymbol{x}_i) \\
&\approx \prod_{i\in S_u}\left[\frac{1+\exp\{\boldsymbol{x}_i\boldsymbol{\beta}\}}{1+\left(1+\frac{n_p}{\tilde{n}_{1u}}\right)\exp\{\boldsymbol{x}_i\boldsymbol{\beta}\}}\right]\prod_{i\in S_p}\left[\frac{\frac{n_p}{\tilde{n}_{1u}}\exp\{\boldsymbol{x}\boldsymbol{\beta}_i\}}{1+\left(1+\frac{n_p}{\tilde{n}_{1u}}\right)\exp\{\boldsymbol{x}\boldsymbol{\beta}_i\}}\right].
\end{aligned}$$

4.1.2 Complete Likelihood

As anticipated in the previous section, another approach to handle the presence-only model is based on the complete likelihood $L(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{z},\boldsymbol{X})$, i.e. the joint probability of the $\boldsymbol{Y}$ and $\boldsymbol{Z}$ processes:

$$
\begin{aligned}
L(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{X}) &= \prod_{i \in S} \Pr(Y_i, Z_i \mid s_i = 1, \boldsymbol{x}_i) \\
&= \prod_{i \in S} \Pr(Z_i \mid Y_i, s_i = 1, \boldsymbol{x}_i) \Pr(Y_i \mid s_i = 1, \boldsymbol{x}_i) \\
&= \prod_{i \in S} \Pr(Z_i \mid Y_i, s_i = 1) \Pr(Y_i \mid s_i = 1, \boldsymbol{x}_i) \\
&= \prod_{i \in S_u} \Pr(Z_i \mid Y_i, s_i = 1) \Pr(Y_i \mid s_i = 1, \boldsymbol{x}_i) \\
&\times \prod_{i \in S_p} \Pr(Z_i \mid Y_i, s_i = 1) \Pr(Y_i \mid s_i = 1, \boldsymbol{x}_i).
\end{aligned}
$$

Then, given the conditional probabilities $\Pr(Z \mid Y, s = 1, \boldsymbol{x})$ defined in Table 4.1 and given the $\Pr(Y \mid s = 1, \boldsymbol{x})$ in (4.3), the complete likelihood is obtained as follows:

$$
\begin{aligned}
L(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{X}) &= \prod_{i \in S_u} \Pr(Z_i \mid Y_i, s_i = 1) \Pr(Y_i \mid s_i = 1, \boldsymbol{x}_i) \\
&\times \prod_{i \in S_p} \Pr(Z_i \mid Y_i, s_i = 1) \Pr(Y_i \mid s_i = 1, \boldsymbol{x}_i) \\
&= \prod_{i \in S_u} \left\{ [\Pr(y_i = 0 \mid s_i = 1, \boldsymbol{x}_i)]^{1-y_i} \left[\frac{\tilde{n}_{1u}}{\tilde{n}_1} \Pr(y_i = 1 \mid s_i = 1, \boldsymbol{x}_i) \right]^{y_i} \right\} \\
&\times \prod_{i \in S_p} \left\{ \frac{n_{1p}}{\tilde{n}_1} \Pr(y_i = 1 \mid s_i = 1, \boldsymbol{x}_i) \right\} \\
&\approx \prod_{i \in S_u} \left\{ \left[\frac{1}{1 + \left(1 + \frac{n_p}{\tilde{n}_{1u}}\right) \exp\{\boldsymbol{x}_i \boldsymbol{\beta}\}} \right]^{1-y_i} \left[\frac{\tilde{n}_{1u}}{\tilde{n}_1} \frac{\left(1 + \frac{n_p}{\tilde{n}_{1u}}\right) \exp\{\boldsymbol{x}_i \boldsymbol{\beta}\}}{1 + \left(1 + \frac{n_p}{\tilde{n}_{1u}}\right) \exp\{\boldsymbol{x}_i \boldsymbol{\beta}\}} \right]^{y_i} \right\} \\
&\times \prod_{i \in S_p} \left\{ \frac{n_{1p}}{\tilde{n}_1} \frac{\left(1 + \frac{n_p}{\tilde{n}_{1u}}\right) \exp\{\boldsymbol{x}_i \boldsymbol{\beta}\}}{1 + \left(1 + \frac{n_p}{\tilde{n}_{1u}}\right) \exp\{\boldsymbol{x}_i \boldsymbol{\beta}\}} \right\} \\
&= \prod_{i \in S_u} \left\{ \left[\frac{1}{1 + \left(1 + \frac{n_p}{\tilde{n}_{1u}}\right) \exp\{\boldsymbol{x}_i \boldsymbol{\beta}\}} \right]^{1-y_i} \left[\frac{\exp\{\boldsymbol{x}_i \boldsymbol{\beta}\}}{1 + \left(1 + \frac{n_p}{\tilde{n}_{1u}}\right) \exp\{\boldsymbol{x}_i \boldsymbol{\beta}\}} \right]^{y_i} \right\} \\
&\times \prod_{i \in S_p} \left\{ \frac{\frac{n_p}{\tilde{n}_{1u}} \exp\{\boldsymbol{x}_i \boldsymbol{\beta}\}}{1 + \left(1 + \frac{n_p}{\tilde{n}_{1u}}\right) \exp\{\boldsymbol{x}_i \boldsymbol{\beta}\}} \right\}.
\end{aligned} \tag{4.5}
$$

4.2 Bayesian model

Let $\boldsymbol{\delta}$ be a vector of hyperparameters with hyperprior $p(\boldsymbol{\delta})$. Conditioned on $\boldsymbol{\delta}$, the regression parameters, βs, are Gaussian random variables. Given $\boldsymbol{\beta}$ and $\boldsymbol{x}$, the process $\boldsymbol{Y}$ is a set of independent Bernoulli random variables with probability of occurrence $\pi_S(\boldsymbol{x})$ given in (4.3). At the lowest level of the hierarchy, the conditional distribution of Z given Y can be easily derived from Table 4.1. Then, the hierarchical Bayesian model is:

level 1. $\boldsymbol{\delta} \sim p(\boldsymbol{\delta})$;

level 2. $\boldsymbol{\beta} \mid \boldsymbol{\delta} \sim p(\boldsymbol{\beta} \mid \boldsymbol{\delta})$;

level 3. $Y \mid s = 1, \boldsymbol{x} \sim Ber(\pi_S(\boldsymbol{x}))$;

level 4. $Z \mid Y, s = 1 \sim p(Z \mid Y, s = 1)$.

Given the prior distributions, the joint posterior distribution of $\boldsymbol{\theta}$ can be derived with respect to the complete likelihood or alternatively with respect to the observed likelihood.

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{x}) \propto L(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{x}) p(\boldsymbol{\beta} \mid \boldsymbol{\delta}) p(\boldsymbol{\delta})$$

or

$$p(\boldsymbol{\theta} \mid \boldsymbol{z}, \boldsymbol{x}) \propto L(\boldsymbol{\theta}; \boldsymbol{z}, \boldsymbol{x}) p(\boldsymbol{\beta} \mid \boldsymbol{\delta}) p(\boldsymbol{\delta}).$$

4.3 MCMC algorithm

In the following scheme a MCMC computation that can be applied to the complete and observed likelihood is shown.

Algorithm 3 Data Augmentation MCMC

```
    Step 0: initialize $\boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{Y}$;
Repeat:
    Step 1: set $n_{1u} = \sum_{i \in S_u} y_i$;
    Step 2: $\boldsymbol{\delta} \sim p(\boldsymbol{\delta} \mid \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{X})$;
    Step 3: $\boldsymbol{\beta} \sim p(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\delta}, \boldsymbol{X})$;
    Step 4: $y_i \sim p(Y_i \mid Z_i, \boldsymbol{\beta}, \boldsymbol{\delta}, s_i = 1, \boldsymbol{x}_i)$
```

At Step 0 let assign arbitrary values to $\boldsymbol{\delta}, \boldsymbol{\beta}$ and $\boldsymbol{Y}$. In particular, to Y_i is assigned a realization of a Bernoulli distribution with probability of occurrence equal to 0.5 if the observation belongs to S_u, while $y_i = 1$ if the observation belongs to S_p.

After to have set or sampled values for Y, at Step 1 the number of presences in the sample S_u is obtained as sum of the Y_i simulated at the previous iteration. Then the ratio in (4.2) is identified and the proposed algorithm can be computed.

At Step 2 the hyperparameters are sampled from their conditional distributions.

At Step 3 the regression parameters ($\boldsymbol{\beta}$) are sampled from the following conditional probability:

$$p(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\delta}; \boldsymbol{X}) \propto p(\boldsymbol{\beta}) \frac{\exp\left\{\sum_i^n 1_{y_i=1} \boldsymbol{x}_i \boldsymbol{\beta}\right\}}{1 + \left(1 + \frac{n_p}{\tilde{n}_{1u}}\right) \exp\left\{\sum_i^n 1_{y_i=1} \boldsymbol{x}_i \boldsymbol{\beta}\right\}}$$

At Step 4 the unobserved Y_i is simulated from its conditional distribution, if the complete likelihood is used, or from its predictive distribution, if instead the observed likelihood is considered.

The only requirement to perform the augmentation is that the covariates $\boldsymbol{X}$ are available for each unit belonging to the population $\mathcal{U}$.

Note that an estimate of π_u can be obtained as

$$\hat{\pi}_u = \frac{\bar{n}_{1u}}{n_u}$$

where $\bar{n}_{1u}$ is the MCMC average of the simulations of n_{1u} saved after the burn-in period.

At computational level, the MCMC algorithm is simplified thanks as given $\boldsymbol{Z}$, the posterior distribution of $\boldsymbol{Y}$ does not depend on the censured case-control design. This means that for each unit $i \in S_u$, the unobserved Y_i is simulated from a Bernoulli distribution with probability of occurrence

$$\pi(\boldsymbol{x}_i) = \frac{\exp\{\eta(\boldsymbol{x}_i)\}}{1 + \exp\{\eta(\boldsymbol{x}_i)\}}$$

where

$$\eta(\boldsymbol{x}_i) = \boldsymbol{x}_i \boldsymbol{\beta}.$$

This is because

$$p(Y_i \mid Z_i, s_i = 1, \boldsymbol{x}_i) \sim Ber(\pi(\boldsymbol{x}_i)), \forall i \in S_u$$

and

$$p(Y_i \mid Z_i, s_i = 1, \boldsymbol{x}_i) \sim Dirac(1), \forall i \in S_p.$$

This statement is henceforward proved:

$\forall i \in S_u$

$$\begin{aligned}
\Pr(y_i = 1 \mid Z_i, s_i = 1, \boldsymbol{x}_i) &= \Pr(y_i = 1 \mid z_i = 0, s_i = 1, \boldsymbol{x}_i) \\
&= \frac{Pr(z_i = 0 \mid y_i = 1, s_i = 1, \boldsymbol{x}_i) Pr(y_i = 1 \mid s_i = 1, \boldsymbol{x}_i)}{Pr(z_i = 0 \mid s_i = 1, \boldsymbol{x}_i)} \\
&= \frac{\frac{\tilde{n}_{1u}}{\tilde{n}_1} \frac{\left(1+\frac{\tilde{n}_p}{\tilde{n}_{1u}}\right) \exp\{\eta(\boldsymbol{x}_i)\}}{1+\left(1+\frac{\tilde{n}_p}{\tilde{n}_{1u}}\right) \exp\{\eta(\boldsymbol{x}_i)\}}}{1 - \frac{\frac{n_p}{\tilde{n}_1 u} \exp\{\eta(\boldsymbol{x}_i)\}}{1+\left(1+\frac{\tilde{n}_p}{\tilde{n}_{1u}}\right) \exp\{\eta(\boldsymbol{x}_i)\}}} \\
&= \frac{\frac{\exp\{\eta(\boldsymbol{x}_i)\}}{1+\left(1+\frac{\tilde{n}_p}{\tilde{n}_{1u}}\right) \exp\{\eta(\boldsymbol{x}_i)\}}}{\frac{1+\exp\{\eta(\boldsymbol{x}_i)\}}{1+\left(1+\frac{\tilde{n}_p}{\tilde{n}_{1u}}\right) \exp\{\eta(\boldsymbol{x}_i)\}}} \\
&= \frac{\exp\{\eta(\boldsymbol{x}_i)\}}{1 + \exp\{\eta(\boldsymbol{x}_i)\}}
\end{aligned}$$

and

$$\begin{aligned}
\Pr(y_i = 0 \mid Z_i, s_i = 1, \boldsymbol{x}_i) &= 1 - \Pr(y_i = 1 \mid z_i = 0, s_i = 1, \boldsymbol{x}_i) \\
&= 1 - \frac{\exp\{\eta(\boldsymbol{x}_i)\}}{1 + \exp\{\eta(\boldsymbol{x}_i)\}} \\
&= \frac{1}{1 + \exp\{\eta(\boldsymbol{x}_i)\}};
\end{aligned}$$

$\forall i \in S_p$

$$\begin{aligned}
\Pr(y_i = 1 \mid Z_i, s_i = 1, \boldsymbol{x}_i) &= \Pr(y_i = 1 \mid z_i = 1, s_i = 1, \boldsymbol{x}_i) \\
&= \frac{\Pr(z_i = 1 \mid y_i = 1, s_i = 1, \boldsymbol{x}_i) \Pr(y_i = 1 \mid s_i = 1, \boldsymbol{x}_i)}{\Pr(z_i = 1 \mid s_i = 1, \boldsymbol{x}_i)} \\
&= \frac{\Pr(z_i = 1 \mid y_i = 1, s_i = 1, \boldsymbol{x}_i) \Pr(y_i = 1 \mid s_i = 1, \boldsymbol{x}_i)}{\Pr(z_i = 1 \mid s_i = 1, \boldsymbol{x}_i)} \\
&= \frac{\frac{n_p}{\tilde{n}_1} \frac{\left(1+\frac{\tilde{n}_p}{\tilde{n}_{1u}}\right) \exp\{\eta(\boldsymbol{x}_i)\}}{1+\left(1+\frac{\tilde{n}_p}{\tilde{n}_{1u}}\right) \exp\{\eta(\boldsymbol{x}_i)\}}}{\frac{\frac{n_p}{\tilde{n}_1 u} \exp\{\eta(\boldsymbol{x}_i)\}}{1+\left(1+\frac{\tilde{n}_p}{\tilde{n}_{1u}}\right) \exp\{\eta(\boldsymbol{x}_i)\}}} \\
&= 1
\end{aligned}$$

and

$$\begin{aligned}
\Pr(y_i = 0 \mid Z_i, s_i = 1, \boldsymbol{x}_i) &= 1 - \Pr(y_i = 1 \mid z_i = 0, s_i = 1, \boldsymbol{x}_i) \\
&= 0.
\end{aligned}$$

4.4 Simulation Study

In this section the performance of the proposed model is investigated on simulated data.

A population of 10000 units on a regular grid 100×100 from the following model is generated:

$$logit\,\Pr(y = 1 \mid X) = \eta(X)$$

$$\eta(X) = \beta X$$

where $\beta = 3$ and the covariate X is sampled from a Gaussian random field with mean -2, variance 3 and range 15.

The resulting simulated data are reported in Figure 4.1:

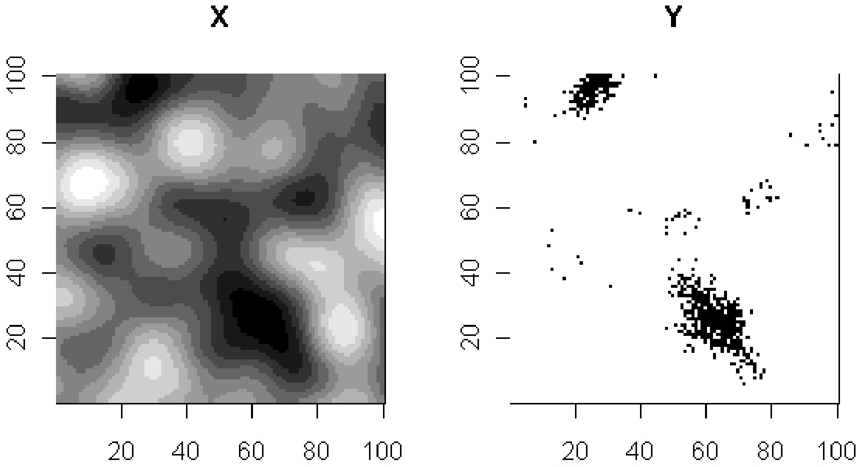


Figure 4.1: Simulated data.

From this population 1000 samples of size n are drawn randomly selecting the 70% of observed presences for S_p and pseudo-absences for S_u in a rate $1 : 5$. The prevalence of the population is equal to 0.044.
A Bayesian model, in the observed likelihood version, for two different situations is fitted: with unknown π (M1) and assuming the population prevalence to be known in the correction factor (M2). The second situation represents the benchmark of the model proposed in this chapter and it can be considered the Bayesian version of the model developed in Ward et al. (2009). Both models are fitted assuming the standard Gaussian $\mathcal{N}(0, 100)$ as the prior for β. 20000 iterations are considered and 10000 are discarded as burn-in.

In Table 4.2 the average Relative Root Mean Squared Error (RRMSE) of the parameter estimates of model M1 and M2 for $n = 100$ is reported. The mean of samples prevalence in S_u over the 1000 samples is 0.04435. Also, in Table 4.2 the predictive performance of the models are compared. Misclassification Error (ME), sensibility and specifity of predicted presence-absence data are summarized in same table:

$n = 100$	RRMSE of $\hat{\beta}$	RRMSE of $\hat{\pi}_u$	ME	Sensitivity	Specificity
M1	1.9185	0.4556	0.027	0.6023	0.9901
M2	0.2590	0.3818	0.027	0.6023	0.9902

Table 4.2: RRMSE of $\hat{\beta}$ and $\hat{\pi}_u$. Misclassification Error, Sensitivity and Specificity of the predicted presence-absence data. The results are averaged over 1000 samples for each model.

For the proposed Bayesian method (M1) it can be seen that, as expected, the RRMSE is large than one of M2. Instead, the appears to be no difference between M1 and M2, in both cases the predicted values are very close to the observed values, sensitivity is large and specificity is very large. The results related to the predictive performance of the model (sensitivity and specificity) shown in Table 4.2 would be in contrast to those obtained in Table 3.3. Remark that, despite of the simulation study shown in Section 3.3, here the data are generated from a model with zero intercept. In the pseudo-absence approach the intercept estimate is an issue because it is highly correlated to the prevalence estimate (see Ward et al. (2009)). In simulation study shown in Section 3.3 the intercept parameter is overestimated and then more 1s than those observed are predicted (large sensitivity and small specificity).

In Figure 4.2 and 4.3 are reported the box plots of the parameter estimates of β and π over 1000 random samples, respectively, for each model M1 and M2 and for different values of n.

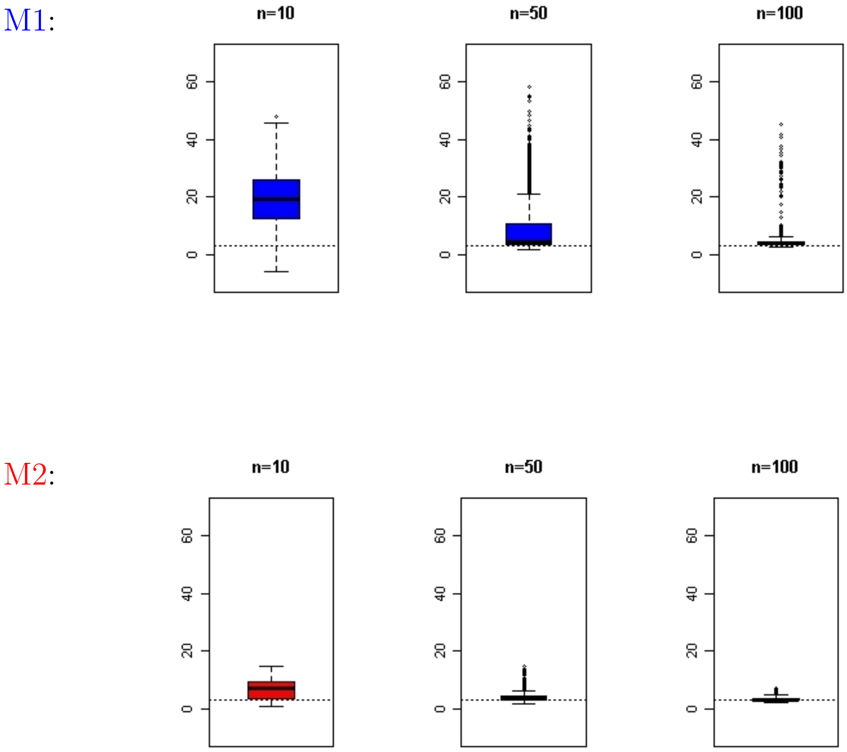


Figure 4.2: $\hat{\beta}$ estimated over 1000 random samples for each model.

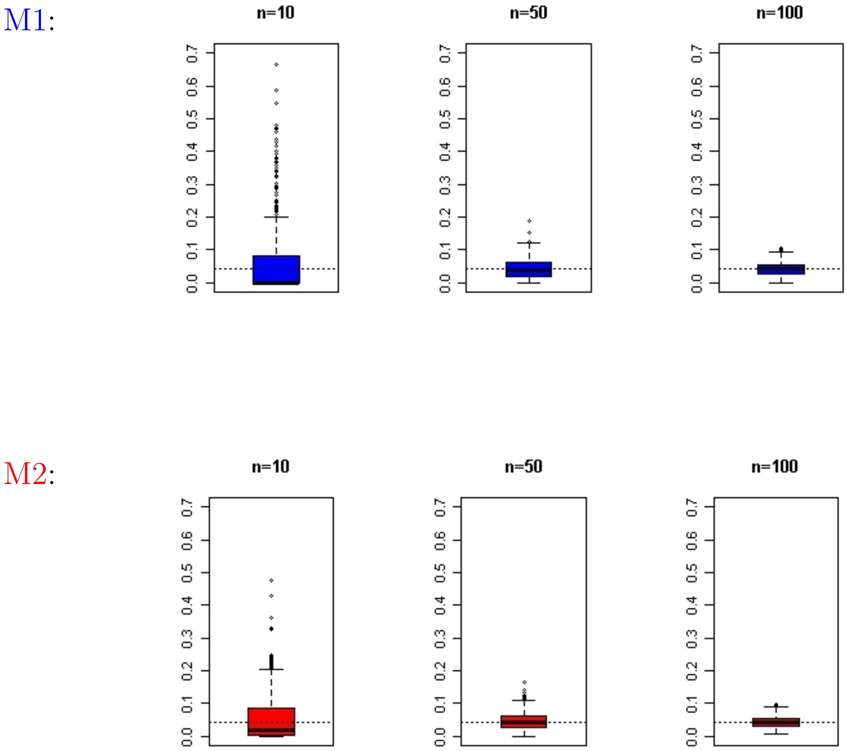


Figure 4.3: $\hat{\pi}_u$ estimated over 1000 random samples for each model.

It is possible to note that when n increases the estimates of β and π become closer to the true values of the parameters. Obviously, the parameter estimates with respect M2 tend more quickly to the true value of parameters then ones obtained by M1.

In Table 4.3 is reported the posterior mean and the 95% credibility interval for β and the posterior mean for π for 3 samples of size 100 for each model. This samples are chosen in order to show the performance of two models where particular samples are drawn.

		$M1$			$M2$		
$n = 100$	π_u	$\hat{\beta}$	95%CI	$\hat{\pi}_u$	$\hat{\beta}$	95%CI	$\hat{\pi}_u$
$s1$	0.0125	5.04	$(3.25; 7.46)$	0.017	3.86	$(2.60; 5.78)$	0.020
$s2$	0.0500	3.75	$(2.22; 5.63)$	0.042	3.39	$(2.41; 4.78)$	0.043
$s3$	0.1125	2.59	$(1.25; 4.88)$	0.100	2.70	$(1.98; 3.63)$	0.095

Table 4.3: Posterior mean and credibility interval for β and posterior mean for π_u.

Note that when the sample prevalence observed in S_u is similar to the true prevalence of the population ($\pi = 0.044$) the estimates of π_u, $\hat{\pi}_u$, represents a good approximation of π. Also, the estimates of β are closer to the true value of the parameter.

Chapter 5

Spatial Bayesian Modeling of Presence-only Data

Chapter 3 and 4 illustrate models based on the assumption that the observed environmental covariates are the only determinants of species distributions. This assumption may not be adequate or sufficient to account for a species distribution. Those models may fail to provide adequate predictive power or may underestimate the degree of uncertainty of predictions. The methodology shown in this chapter has been published in

> Divino, F., Golini, N., Jona Lasinio, G. and Penttinen A. (2011). Spatial Bayesian Modeling of Presence-only Data. Proceedings of the 17th EYSM, Lisbon, Portugal, 2011.

Here a spatial extension of the model proposed in Chapter 4 is presented. It is based on the assumption that the presence-absence data are spatially dependent or autocorrelated, i.e. the degree of correlation among observations depends on their relative locations. Spatial dependence in the data is incorporated into the regression model through a spatially structured random effect.

5.1 Why a spatial model?

In ecology some processes (i.e. reproduction or dispersal) may affect the spatial arrangement of species distributions causing spatial dependence (or spatial autocorrelation) in species occurrences (or abundances), see Gaston (2003). Spatial dependence leads to a dependence among locations that decays with distance: pairs of locations that are closer together often tend to have measures of species occurrences (or abundances) more similar than pairs of locations that are farther apart. Also, spatial autocorrelation invalidates the assumption of independence among sample locations on which the traditional statistical models employed in distribution modeling (i.e. regression models) are based. Then, using models that ignore spatial dependence can lead to inaccurate parameters estimates and inadequate quantification of uncertainty (Ver Hoef et al. (2001)). Yet, to ignore the spatial dependence implies not consider additional information, such as the values at neighboring locations, that can help to improve the predictive power of the model (Wikle (2003)). In Latimer et al. (2006) the authors affirm that "making distribution models spatially explicit can be essential for accurately characterizing the environmental responde species, predicting their probability of occurrence, and assessing uncertainty in the model results". These statements are also reasserted in Bahn et al. (2006), Dormann et al. (2007) and Dormann (2007).

Since ecological data are spatial data, they contain information about both the attribute of interest as well as its location. Following Cressie (1993), spatial data can be categorized into three distinct types: "geostatistical or point-level data", "lattice or areal (regionally aggregated) data" or "point process data". In this chapter species occurrences are considered as areal data. In literature there are two approaches to model binary areal data: logistic spatial generalized linear mixed model and the autologistic model. Each approach is characterized by its modeling of spatial dependence. The first, when a random effect is added, models dependence indirectly, by way of a latent Gaussian Markov random field over the lattice of interest (Banerjee et al. (2004)). The autologistic model, formulated by Besag in Besag (1974), models dependence directly, through the so-called autocovariate, which is a function of the observations themselves. It has since found may applications in several fields, in particular ecology and epidemiology, see (Augustin et al. (1997), Gumpertz et al. (1997), Huffer & Wu (1998), Hoeting et al. (2000), He et al. (2003), Caragea & Kaiser (2009), Hughes et al. (2011). In this chapter the focus will be on the logistic spatial generalized linear mixed model.

5.2 Gaussian Markov random field models

Let Λ be a regular lattice of N knots and $\boldsymbol{u} = (u_1, \ldots, u_N)'$ a real values Gaussian Markov random field on Λ. Let $\boldsymbol{u}_{-i}$ denote the vector $\boldsymbol{u}$ excluding u_i. A possible way to define a Gaussian Markov random filed (GMRF) is to specify it implicitly thought the full conditionals $\{p(u_i \mid \boldsymbol{u}_i)\}$. This approach was pioneered by Besag (1974) and Besag (1975). For each location i, u_i is considered in term of its conditional distribution given the remaining random variables, $\boldsymbol{u}_{-i}$:

$$u_i \mid \boldsymbol{u}_{-i} \sim \mathcal{N}\left(\mu_i - \sum_{j:j\sim i} c_{ij}(u_j - \mu_j), \frac{1}{\kappa_i}\right), \qquad i = 1, \ldots, N \tag{5.1}$$

where μ_i is the marginal mean of u_i, $\kappa_i > 0$ is the precision parameter and c_{ij} describes the effect of the neighborhood structure. The notation $j \sim i$ implies that i and j are neighbors. Note that c_{ij} is non-zero only if i and j are neighbors. Being the joint distribution of the Gaussian process, $\boldsymbol{u}$, specified trough the set of conditional distributions given in (5.1), some conditions on the parameters must be added to ensure that the resulting joint distribution is "well" defined. Since the neighborhood relationship ($\sim$) is symmetric, the following requirement immediately is given:

$$\text{if} \quad c_{ij} \neq 0 \quad \text{then} \quad c_{ji} \neq 0.$$

Let Q be the precision matrix of $\boldsymbol{u}$. If Q is positive defined, i.e.,

$$Q_{ii} = \kappa_i, \quad \text{and} \quad Q_{ij} = \kappa_i c_{ij}$$

and also symmetric, i.e.,

$$\kappa_i c_{ij} = \kappa_j c_{ji}$$

then a valid joint distribution for $\boldsymbol{u}$

$$\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{\mu}, Q^{-1}) \tag{5.2}$$

can be obtained by the set of conditional distributions given in (5.1), see Besag (1974).

In order to avoid identifiability problems, it is often assumed a constant precision parameter, i.e. $\kappa_i = \kappa$ for all i. Hence, $Q = \kappa(I + C)$ where C is an $N \times N$ matrix with zero diagonal entries and $c_{ij} \neq 0$ only when $i \sim j$, $i \neq j$. Bayesian inference for the linear GMRF model specified by (5.2) can therefore proceed after assuming a prior distribution for κ.

5.3 Spatial generalized linear models

Spatial generalized linear models (SGLMs) are linear models (McCullagh & Nelder (1989)) for spatially associated data. In particular here the SGLMs model are refereed to spatial generalized "mixed" models since the spatial dependence is introduced by adding an error term modeled via a Gaussian Markov random field.

5.3.1 Spatial logistic model

Let $\boldsymbol{Y}$ be the random field of interest, where $y_i \in \{0, 1\}$ represents the observation at the ith lattice point for $i = 1, \dots, N$. An SGLM for binary data may be specified as follows

$$logit \, \Pr(y = 1 \mid \boldsymbol{x}) = \eta(\boldsymbol{x}) \tag{5.3}$$

$$\eta(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{u} \tag{5.4}$$

where $\eta(\cdot)$ is a linear regression function with a spatially structured random effect $\boldsymbol{u}$ modeled via GMRF.

5.4 Modeling method

Given the ratio in (4.2) and according to the model specified in (5.3) and (5.4), the adjusted case-control model can be expressed as follows:

$$logit \, \Pr(y = 1 \mid s = 1, \boldsymbol{x}) = \eta(\boldsymbol{x}) + \ln\left(\frac{\tilde{n}_{1u} + n_p}{\tilde{n}_{1u}}\right) \tag{5.5}$$

$$\eta(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{u} \tag{5.6}$$

where, as in Section 5.3.1, $\eta(\cdot)$ is a linear regression function with a spatially structured random effect $\boldsymbol{u}$ modeled via GMRF. In this work an improper version of GMRF, the so called "Intrinsic Gaussian Markov random field" (IGMRF) of first order is considered (see Besag et al. (1991), Besag & Kooperberg (1995)). In particular here a zero mean IGMRF is used. Then $\boldsymbol{u}$ has joint distribution:

$$p(\boldsymbol{u}) \propto \kappa^{\frac{N-1}{2}} \exp\{-\boldsymbol{u}' Q \boldsymbol{u}\}, \tag{5.7}$$

where Q has off-diagonals entries equal to $-\kappa c_{ij}$ and ith diagonal element $\kappa \sum_j c_{ij}$. Because Q is not of full rank, the distribution defined in (5.7) is improper. This form is a very popular prior distribution for spatially structured random effects in generalized linear models (Banerjee et al. (2004); Rue & Held (2005)) and it is used in various applications, especially in disease mapping (see MacNab (2003), Paciorek (2007), Wakefield (2007)). Generally GMRFs are a convenient models from both a computational and theoretical point of view: they have the Markov property and they are jointly Gaussian. The Markov property is also important for models relying on inference based on MCMC sampling as it ensures rapid computation of the conditional density. A first order zero mean IGMRF represents the easiest and computational faster way to implement a GMRF. Also, the use of an intrinsic GMRF is not an issue as long as the posterior is proper (a detailed discussion on the conditions under which posterior property is guaranteed for various GMRF models is given in Sun et al. (1999)). Yet, in ecological studies the use of an intrinsic GMRF is justified from a conceptual point of view. This model reflects the idea that some ecological processes may effect the spatial arrangement of the species distributions causing spatial dependence in species occurrences, as discussed in Section 5.1. To consider an intrinsic GMRF as model to describe $\boldsymbol{u}$ implies to affirm that the measure of species occurrences at a location also will depend on the average of values at neighboring locations.

When $c_{ij} = 1$ if $j \sim i$ and 0 otherwise, (5.7) simplifies to the “pairwise-difference form”:

$$p(\boldsymbol{u}) \propto \kappa^{\frac{N-1}{2}} \exp\left(-\frac{\kappa}{2}\sum_{i\sim j}(u_i - u_j)^2\right), \tag{5.8}$$

where $i \sim j$ denotes the set of all unordered pairs of neighbors. The requirement for the pair to be unordered prevents from double counting as $i \sim j \Leftrightarrow j \sim i$.

Let n_i denote the number of neighbors of location i. If $c_{ij} = 1$ when i and j are neighbors and 0 otherwise, the precision matrix Q in (5.7) has elements

$$Q_{ij} = \kappa \begin{cases} n_i, & i = j; \\ -1, & i \sim j; \\ 0, & \text{otherwise.} \end{cases}$$

from which it follows an intuitive conditional specification:

$$u_i \mid \boldsymbol{u}_{-i}, \kappa \sim \mathcal{N}\left(\frac{\sum_{j:j\sim i} u_j}{n_i}, \frac{1}{\kappa n_i}\right). \tag{5.9}$$

Hence, the distribution of u_i is normal with mean given by the average of its neighbors and its variance decreases as the number of neighbors increases. This means that the measure of species occurrences at location i also will depend on the average of values at neighboring locations.
See Rue & Held (2005) for a discussion of the related theory of IGMRF models.

5.4.1 Observed and complete likelihood

The procedure to obtain the observed and complete likelihood is the same used in Chapter 4. It is straightforward to prove that the observed likelihood can be specified as

$$\begin{aligned} L(\boldsymbol{\theta}; \boldsymbol{z}, \boldsymbol{X}) &= \prod_{i \in S} \Pr(Z_i \mid s_i = 1, \boldsymbol{x}_i) \qquad (5.10) \\ &= \prod_{i \in S_u} \Pr(z_i = 0 \mid s_i = 1, \boldsymbol{x}_i) \prod_{i \in S_p} \Pr(z_i = 1 \mid s_i = 1, \boldsymbol{x}_i) \\ &\approx \prod_{i \in S_u} \left[\frac{1 + \exp\{\boldsymbol{x}_i \boldsymbol{\beta} + u_i\}}{1 + \left(1 + \frac{n_p}{\tilde{n}_{1u}}\right) \exp\{\boldsymbol{x}_i \boldsymbol{\beta} + u_i\}} \right] \prod_{i \in S_p} \left[\frac{\frac{n_p}{\tilde{n}_{1u}} \exp\{\boldsymbol{x}_i \boldsymbol{\beta} + u_i\}}{1 + \left(1 + \frac{n_p}{\tilde{n}_{1u}}\right) \exp\{\boldsymbol{x}_i \boldsymbol{\beta} + u_i\}} \right] \end{aligned}$$

while the complete likelihood as

$$\begin{aligned} L(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{X}) &= \prod_{i \in S} \Pr(Z_i \mid Y_i, s_i = 1) \Pr(Y_i \mid s_i = 1, \boldsymbol{x}_i) \qquad (5.11) \\ &= \prod_{i \in S_u} \left\{ [\Pr(y_i = 0 \mid s_i = 1, \boldsymbol{x}_i)]^{1-y_i} \left[\frac{\tilde{n}_{1u}}{\tilde{n}_1} \Pr(y_i = 1 \mid s_i = 1, \boldsymbol{x}_i) \right]^{y_i} \right\} \\ &\times \prod_{i \in S_p} \left\{ \frac{n_{1p}}{\tilde{n}_1} \Pr(y_i = 1 \mid s_i = 1, \boldsymbol{x}_i) \right\} \\ &\approx \prod_{i \in S_u} \left\{ \left[\frac{1}{1 + \left(1 + \frac{n_p}{\tilde{n}_{1u}}\right) \exp\{\boldsymbol{x}_i \boldsymbol{\beta} + u_i\}} \right]^{1-y_i} \left[\frac{\exp\{\boldsymbol{x}_i \boldsymbol{\beta} + u_i\}}{1 + \left(1 + \frac{n_p}{\tilde{n}_{1u}}\right) \exp\{\boldsymbol{x}_i \boldsymbol{\beta} + u_i\}} \right]^{y_i} \right\} \\ &\times \prod_{i \in S_p} \left\{ \frac{\frac{n_p}{\tilde{n}_{1u}} \exp\{\boldsymbol{x}_i \boldsymbol{\beta} + u_i\}}{1 + \left(1 + \frac{n_p}{\tilde{n}_{1u}}\right) \exp\{\boldsymbol{x}_i \boldsymbol{\beta} + u_i\}} \right\} \end{aligned}$$

where $\boldsymbol{\theta}$ is a short-hand notation to denote the parameters at stake.

5.5 Bayesian model

Let $\boldsymbol{\delta}$ be a vector of hyperparameters with hyperprior $p(\boldsymbol{\delta})$. Conditioned on $\boldsymbol{\delta}$, the regression parameters, βs, are Gaussian random variables and the random effect $\boldsymbol{u}$ is an intrinsic Gaussian Markov random field. Given $\boldsymbol{\beta}$, $\boldsymbol{u}$ and the set of covariates $\boldsymbol{X}$, the process $\boldsymbol{Y}$ is a set of Bernoulli random variables with probability of occurrence $\pi_c(\boldsymbol{x})$ defined in (5.5) and (5.6). At the lowest level of the hierarchical specification of the model, the conditional distribution of Z given Y can be derived from the relations between the two process described in Section 2.4.

Then, the hierarchical Bayesian model is given by:

level 1. $\boldsymbol{\delta} \sim p(\boldsymbol{\delta});$

level 2. $\boldsymbol{\beta} \mid \boldsymbol{\delta} \sim p(\boldsymbol{\beta} \mid \boldsymbol{\delta})$ and $\boldsymbol{u} \mid \boldsymbol{\delta} \sim IGMRF(\boldsymbol{\delta});$

level 3. $Y \mid s = 1, \boldsymbol{x} \sim Ber(\pi_c(\boldsymbol{x}));$

level 4. $Z \mid Y, s = 1 \sim p(Z \mid Y, s = 1).$

Given the prior distributions, the joint posterior distribution of $\boldsymbol{\theta}$ can be derived only with respect to the complete likelihood:

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{x}) \propto L(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{x}) p(\boldsymbol{\beta} \mid \boldsymbol{\delta}) p(\boldsymbol{u} \mid \boldsymbol{\delta}) p(\boldsymbol{\delta}).$$

Note, in fact, that the spatial structure of the random effect $\boldsymbol{u}$ is given by the neighborhood system defined among all sites in the target population $\mathcal{U}$.

5.6 MCMC algorithm

In the following scheme the MCMC algorithm used to estimate the proposed model is illustrated, it uses the complete likelihood specification given above.

Algorithm 4 Data Augmentation MCMC

```
    Step 0: initialize $\boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{Y}$;
Repeat:
    Step 1: set $n_{1u} = \sum_{i \in S_u} y_i$;
    Step 2: $\boldsymbol{\delta} \sim p(\boldsymbol{\delta} \mid \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{X})$;
    Step 3: $\boldsymbol{\beta} \sim p(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{u}, \boldsymbol{\delta}, \boldsymbol{X})$;
    Step 4: $\boldsymbol{u} \sim p(\boldsymbol{u} \mid \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{X})$ over $\mathcal{U}$;
    Step 5: $y_i \sim p(Y_i \mid Z_i, \boldsymbol{\beta}, u_i, \boldsymbol{\delta}, s_i = 1, \boldsymbol{x}_i)$ over $\mathcal{U}$.
```

At Step 0 let assign arbitrary values to $\boldsymbol{\delta}, \boldsymbol{\beta}$, $\boldsymbol{u}$ and $\boldsymbol{Y}$. In particular, to Y_i is assigned a realization of a Bernoulli distribution with probability of occurrence equal to 0.5 if the observation belongs to S_u, while $y_i = 1$ if the observation belongs to S_p. To u_i is assigned value 0 for each observation $i \in \mathcal{U}$.

As in Chapter 3, after to have set or sampled values for $\boldsymbol{Y}$, at Step 1 the number of presences in the sample S_u is obtained as sum of the Y_i simulated at the previous iteration. Then the ratio in (4.2) identified and the proposed algorithm can be computed.

At Step 2 the hyperparameters are sampled from their conditional distributions.

At Step 3 the regression parameters ($\boldsymbol{\beta}$) are sampled from the following conditional probability:

$$\Pr(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{u}, \boldsymbol{\delta}, \boldsymbol{X}) \propto p(\boldsymbol{\beta}) \frac{\exp\left\{\sum_i^n 1_{y_i=1} \boldsymbol{x}_i \boldsymbol{\beta} + u_i\right\}}{1 + \left(1 + \frac{n_p}{\tilde{n}_{1u}}\right) \exp\left\{\sum_i^n 1_{y_i=1} \boldsymbol{x}_i \boldsymbol{\beta} + u_i\right\}}$$

At Step 4 the spatially structured random effect $\boldsymbol{u}$ is simulated. For each location i, u_i is generated from the following conditional probability:

$$\Pr(u_i \mid y_i, z_i, \boldsymbol{\beta}, \boldsymbol{u}_{-i}, \boldsymbol{\delta}, \boldsymbol{x}) \propto p(u_i \mid \boldsymbol{u}_{-i}, \kappa) \frac{\exp\left\{\sum_i^n 1_{y_i=1} \boldsymbol{x}_i \boldsymbol{\beta} + u_i\right\}}{1 + \left(1 + \frac{n_p}{\tilde{n}_{1u}}\right) \exp\left\{\sum_i^n 1_{y_i=1} \boldsymbol{x}_i \boldsymbol{\beta} + u_i\right\}}$$

At Step 5 the unobserved Y_i is simulated from its conditional distribution.

At computational level, the MCMC algorithm is more complicated than one proposed in Chapter 4. Remark that it is necessary to perform data augmentation (Step 4 and Step 5) over the target population $\mathcal{U}$ for both $\boldsymbol{u}$ and $\boldsymbol{Y}$ processes in order to consider the spatial structure of the sites enclosed in both samples S_u and S_p. The only requirement to perform the augmentation is that the covariates $\boldsymbol{X}$ are available for each unit belonging to the population $\mathcal{U}$.

Again, an estimate of π_u is easily obtained as

$$\hat{\pi}_u = \frac{\bar{n}_{1u}}{n_u}$$

where $\bar{n}_{1u}$ is the MCMC average of the simulations of n_{1u} saved after the burn-in period.

As discussed in Section 4.3, given $\boldsymbol{Z}$ the posterior distribution of $\boldsymbol{Y}$ does not depend on the censured case-control design. But now, for each unit $i \in S_u$, the unobserved Y_i is simulated from a Bernoulli distribution with probability of occurrence

$$\pi_{sp}(\boldsymbol{x}_i) = \frac{\exp\{\eta(\boldsymbol{x}_i)\}}{1 + \exp\{\eta(\boldsymbol{x}_i)\}}$$

where

$$\eta(\boldsymbol{x}_i) = \boldsymbol{x}_i \boldsymbol{\beta} + u_i.$$

This is because

$$p(Y_i \mid Z_i, \beta, u_i, \boldsymbol{\delta}, s_i = 1, \boldsymbol{x}_i) \sim Ber(\pi_{sp}(\boldsymbol{x}_i)), \forall i \in S_u$$

and

$$p(Y_i \mid Z_i, \beta, u_i, \boldsymbol{\delta}, s_i = 1, \boldsymbol{x}_i) \sim Dirac(1), \forall i \in S_p.$$

The above statements are straightforward to prove, see Section 4.3.

5.7 Simulation Study

In this section preliminary results from a small simulation study are reported, with the aim to investigate the performance of the proposed model in a very simple situation.

A population of 10000 units on a regular 100 × 100 grid from the following model is generated from a spatial logistic model with linear regression function

$$logit\,\Pr(y = 1 \mid X) = \eta(X)$$

$$\eta(X) = \beta X + U$$

where $\beta = -1$, the covariate X is generated from a mixture of two Gaussian components with standard deviation $\sigma_1 = \sigma_2 = 0.5$ and mean $\mu_1 = -2$ and $\mu_2 = 2$, U is a first order zero mean IGMRF with precision $\kappa = 1$ as defined in (5.8). The prevalence of the generated population, π, is equal to 0.2720.

The resulting simulated data are reported in Figure 5.1:

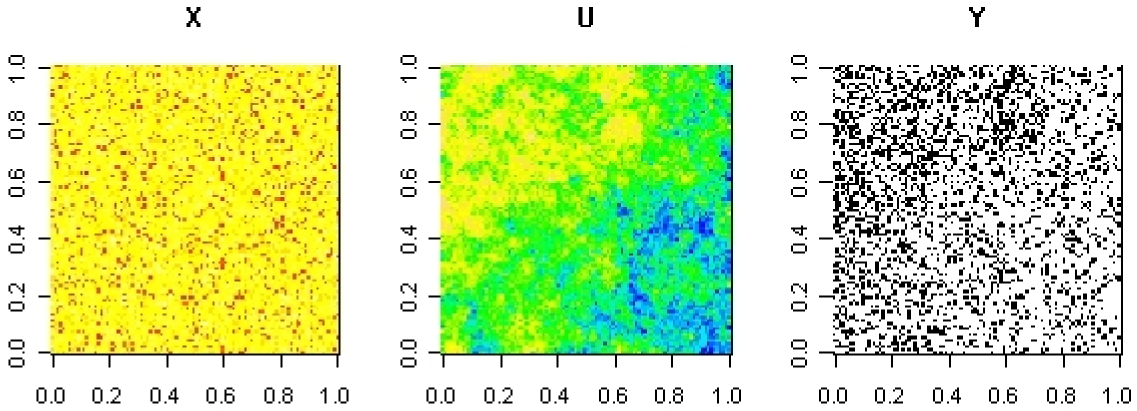


Figure 5.1: Simulated data.

From the Figure 5.1 it is possible to note that the generated data ($\boldsymbol{Y}$) are not strongly spatial dependent.
Let S_u be the pseudo-absence sample that coincides to $\mathcal{U}$ with $z_i = 0$, $\forall i \in \mathcal{U}$, such that $n_u = 10000$. Also, let S_p be the sample of presences of size $n_p = 1904$ obtained by randomly selecting the 70% of true presences from the original population. Hence at each iteration of the algorithm the complete sample will have size $n = 11904$. Note that the sample prevalence of S_u for each of the 1000 samples is an unbiased

estimator of the prevalence of the target population, $\pi_u = \pi = 0.2720$.

A Bayesian model for three different situations is fitted: spatial model with unknown π (M1) proposed in this Chapter, spatial model assuming the population prevalence to be known in the correction factor (M2), non-spatial model proposed in Chapter 4 with unknown prevalence (M3). Remark that the model M2 can be considered as the spatial Bayesian version of the model developed in Ward et al. (2009).

The three models are fitted with the same prior settings: standard Gaussian $\mathcal{N}(0, 100)$ as the prior for β, first order zero mean intrinsic Gaussian Markov random field with precision κ as defined in (5.8) as prior for $\boldsymbol{u}$ and a Gamma distribution $\mathcal{G}(0.0001, 0.0001)$ as prior for κ. In literature a Gamma distribution with mean 1 and large (or infinite) variance as prior for the precision parameter represents a standard choice, see Best et al. (1999), Spiegelhalter et al. (2003), Gelfand et al. (2005), Latimer et al. (2006). 20000 iterations are considered and 10000 are discarded as burn-in. In Table 5.1 the average Relative Root Mean Squared Error (RRMSE) of the parameter estimates of model M1 and M2 for 1000 samples is reported. RRMSE is calculated as ratio of Root Mean Squared Error (RMSE) to true parameter value. Also, the predictive performance of the models are compared. Misclassification Error (ME), sensibility and specifity of predicted presence-absence data are summarized in the same table:

	RRMSE of $\hat{\beta}$	RRMSE of $\hat{\pi}_u$	ME	Sensitivity	Specificity
M1	0.2201	0.2327	0.1925	0.3301	0.9859
M2	0.7322	0.5920	0.1925	0.3301	0.9859
M3	0.1675	0.1065	0.1925	0.3301	0.9859

Table 5.1: RRMSE of $\hat{\beta}$ and $\hat{\pi}_u$. Misclassification Error, Sensitivity and Specificity of the predicted presence-absence. The results are averaged over 1000 samples for each model.

For the model M2 it can be seen that the RRMSE of the parameter estimates is larger than one of M1 and M3. The "best" model in terms of point estimates accuracy is M3 followed by M1. Instead, the appears to be no difference between M1, M2 and M3: in all three cases the predicted values are close to the observed values, sensitivity is enough large and specificity is very large.

In Figure 5.2 is reported the box plots of the parameter estimates of β and π over 1000 random samples, respectively, for each model M1, M2 and M3.

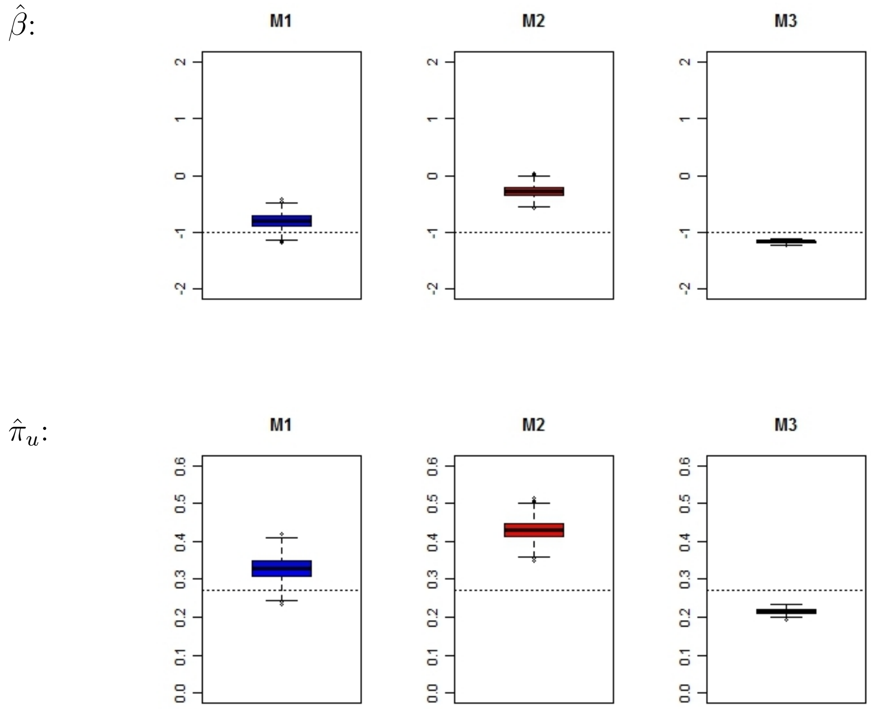


Figure 5.2: $\hat{\beta}$ estimated over 1000 random samples for each model.

It is possible to note that for the model M1 some estimates of β and π become closer to the true values of the parameters.

5.8 Discussion

Several other simulation studies have been performed with the aim to give heuristic guidelines whereby the spatial model would be the best choice in order to predict species distributions. This model would be more adequate than the "independence" model when strong a priori information on the spatial interaction parameter is available, when the latent spatial structure of the phenomenon is known from other studies, when the data generating model includes strongly spatially structured covariates and the estimation model uses an appropriate Gaussian field to model the spatial dependence in the species occurrences.

Conclusions and Further Developments

The present work has involved mainly the study and development of presence-only data models in a Bayesian framework. Among many approaches to the modeling of species distribution with presence-only data that can be found in the ecological literature, the pseudo-absence approach, although it raises some conceptual concerns (see Section 1.3), has been chosen in the present research. The main issue that has to be addressed in this approach relates to the need to know a priori the population prevalence in order to implement logistic-type models for the study on populations distribution. Here three models, trying to address the above mentioned problem, have been developed.

In Chapter 3 a first Bayesian model allowing to partially overcome the need to know a priori the prevalence of the population (π) is presented. The model can be used in situations where π is known with some uncertainty. In the proposed model π has been considered as a parameter of the model and the strength of knowledge about it has been summarized by an informative prior distribution elicited by experts. The model is based on the assumption that the pseudo-absence sample is randomly drawn from the entire study area and that the observed environmental covariates are the only determinants of species distribution.

In Chapter 4 a second Bayesian model allowing to overcome the need to acquire strong information on π is developed. Although here π is not added as a parameter to the model, it can be indirectly estimated in the proposed MCMC algorithm. Also this model is based on the assumption that the pseudo-absence sample is randomly drawn from the entire study area and that the observed environmental covariates are the only determinants of species distribution. In the applications the accuracy of the parameter estimates (and, indirectly, of the prevalence) and the predictive power are related with the significance of available environmental covariates. It could be interesting in future to apply the model to rare species and improve it using auxiliaries data: presence-only data for common species. Frequently, in ecological survey the collected data are common and rare species which are related with habitat selection and do not belong essentially to the same family. Then, given the natural relationship between the common and rare species, is usual to observe that in the locations where the rare species is present also the presence of common species is recorded. The contrary is not true. Hence, conditional on the distribution of presence-only data of the common species, the rare species data can be considered presence-absence data. Then, a simple logistic model can be used to model the potential distribution of the rare species. Furthermore, this model can be easily extended to abundance data.

The models illustrated in Chapter 3 and 4, based on the idea that the observed environmental covariates are the only determinants of species distribution, may fail to provide accurate parameter estimates and adequate predictive power or may underestimate the degree of uncertainty of predictions when species occurrences are spatially dependent. Then in Chapter 5 a spatial extension of the model proposed in Chapter 4 has been reported. The model is based on the assumption that the presence-absence data are spatially dependent, i.e. the degree of correlation among the observations depends on their relative locations. Simulation examples return interesting results suggesting heuristic guidelines whereby the spatial model would be the best choice in order to predict species distributions. The identifiability problems related to a not zero intercept should be investigated. A possible solution could be to consider the value of the intercept at each iteration of the MCMC algorithm as a constrained residual from the fitting of the logistic model estimated at the previous step. Yet, it could be interesting to compare the proposed spatial model with the one proposed in Chakraborty et al. (2011), where the presence-only data are viewed as a point pattern, by means of a simulation study. Note that in Warton & Shepherd (2010) the authors have demonstrated that the pseudo-absence approach is equivalent to the point process approach when a large number of pseudo-absences regularly spaced or uniformly located at random over the study area is taken.

Remark that in each proposed models the attention has been restricted to linear models for both the conditional probability that the response is positive (logistic model) and for the response conditionally on it being positive (regression model). Other more flexible choices for the logistic and regression model could be used (e.g., generalized additive models, non-parametric function and others).

Acknowledgements

I would like to express my gratitude to Professors Giovanna Jona Lasinio and Fabio Divino for their supervision and encouragement during my work.

I am also very grateful to Professor Antti Penttinen for his guidance and for his support, for providing an interesting and gratifying environment and for introducing me to some special people (Harri Högmander, Salme Kärkkäinen, Mari Myllymäki, Tuomas Rajala (Grande) and Tuomas Piccolo).

It has been a pleasure to collaborate with Alessio Farcomeni and Bianca Di Lorenzo: thank you guys.

I especially would like to thank all the colleagues at the Department of Statistical Sciences for providing a pleasant working environment.

My thanks also belong to my parents for their encouragement and support. Moreover, I am very grateful to my brother and friends for being present in my life.

Rome, February 2012

Natalia Golini

Bibliography

Agresti, A. (2002), *Categorical Data Analysis*, Wiley-Interscience, New York, USA.

Attorre, F., Alfó, M., De Sanctis, M., Francesconi, F. & Bruno, F. (2007), 'Comparison of interpolation methods for mapping climatic and bioclimatic variables at regional scale', *International Journal of Climatology* **27**, 1825–1843.

Attorre, F., Francesconi, F., Taleb, N., Scholte, P., Saed, A., Alfó, M. & Bruno, F. (2007), 'Will dragonblood survive the next period of climate change? Current and future potential distribution of *Dracaena cinnabari*', *Biological Conservation* **138**, 430–439.

Augustin, N., Mugglestone, M. & Buckland, S. (1997), 'An autologistic model for the spatial distribution of wildlife', *Journal of Applied Ecology* **33(2)**, 339–347.

Austin, M. (2002), 'Spatial prediction of species distribution: An interface between ecological theory and statistical modelling', *Ecological Modelling* **157**, 101–118.

Austin, M. P. (1985), 'Continuum concept, ordination methods and niche theory', *Annual Review of Ecology and Systematics* **16**, 39–61.

Bahn, V., O'Connors, R. & Kro, W. (2006), 'Importance of spatial autocorrelation in modeling bird distributions at a continental scale', *Ecography* **29**, 835–844.

Banerjee, S., Carlin, B. & Gelfand, A. (2004), *Hierarchical modeling and analysis for spatial data*, Chapman & Hall Ltd.

Barry, S. C. & Welsh, A. H. (2002), 'Generalised additive modelling and zero inflated count data', *Ecological Modelling* **157**, 179–188.

Besag, J. (1974), 'Spatial Interaction and the Statistical Analysis of Lattice Systems (with discussions)', *Journal of the Royal Statistical Society, Series B* **36**, 192–236.

Besag, J. (1975), 'Statistical Analysis of Non-Lattice Data', *The Statistician* **24**, 179–195.

Besag, J. & Kooperberg, C. (1995), ‘On conditional and intrinsic autoregressions’, *Biometrika* **82**, 733–746.

Besag, J., York, J. & Mollié, A. (1991), ‘Bayesian image restoration with two applications in spatial statistics(with discussion)’, *Ann. Inst. Statist. Math.* **43**, 1–59.

Best, N., Waller, L., Thomas, A., Conlon, E. & Arnold, R. (1999), *Bayesian Statistics 6*, New York, USA, chapter Bayesian models for spatially correleted diseas and exposure data (with discussion), pp. 131–156.

Busby, J. (1986), ‘A biogeographical analysis of *Nothofagus cunninghamii* (Hook.) Oerst. in southeastern Australia’, *Australian Journal of Ecology* **11**, 1–7.

Busby, J. R. (1991), ‘BIOCLIM - a bioclimatic analysis and prediction system’, *in "Nature Conservation: cost effective biological surveys and data analysis"* pp. 64–68. eds. C.R. Margules & M.P. Austin. CSIRO, Australia.

Caragea, P. & Kaiser, M. (2009), ‘Autologistic models with interpretable parameteres’, *Journal of Agricolture, Biological and Environmental Statistics* **14(3)**, 281–300.

Carpenter, G., Gillison, A. N. & Winter, J. (1993), ‘ DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals ’, *Biodiversity and Conservation* **2**, 667–680.

Caughley, G., Short, J., Grigg, G. & Nix, H. (1987), ‘Kangoors and climate: an anlysis of distribution’, *Journal of Animal Ecology* **56**, 751–761.

Chakraborty, A., Gelfand, A., Wilson, A., Latimer, A. & Silander, J. (2011), ‘Point pattern modelling for degraded presence-only data over large regions’, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **5**, 757–776.

Congdon, P. (2007), *Baysian Statistical Modelling*, John Wiley & Sons Ltd, 2 edition, England.

Cressey, D. (2008), ‘Pushing the modelling envelope’, *Nature* p. available online.

Cressie, N. (1993), *Statistics for Spatial Data*, John Wiley & Sons, New York, 2nd. edition.

Di Lorenzo, B., Farcomeni, A. & Golini, N. (2011), ‘A Bayesian model for presence-only semicontinuous data, with application to prediction of abundance of *Taxus Baccata* in two Italian regions’, *Journal of Agricultural, Biological, and Environmental Statistics* **16**, 339–356.

Diebolt, J. & Robert, C. (1994), 'Estimation of Finite Mixture distributions through Bayesian Sampling', *Journal of the Royal Statistical Society, (Ser. B)* **56**, 363–375.

Dirnböck, T. & Dullinger, S. (2004), 'Habitat distribution models, spatial correlation, functional traits and disperal capacity of alpine-oriented approaches', *Journal of Vegetation Science* **15**, 77–84.

Divino, F., Golini, N., Jona Lasinio, G. & Penttinen, A. (2011*a*), 'Data augmentation approach in Bayesian modeling of presence-only data', *Procedia Environmental Sciences* **7**, 38–43.

Divino, F., Golini, N., Jona Lasinio, G. & Penttinen, A. (2011*b*), 'Spatial Bayesian modeling of presence-only data', *Proceedings of the 17th EYSM, Lisbon, Portugal, 2011* .

Dormann, C., McPherson, J., Araujo, M., Bivand, R., Bolliger, J., Carl, G., Davies, R., Hirzel, A., Jetz, W., Kissling, W., Kühn, I., Ohlemüller, R., Peres-Neto, P., Reineking, B., Schrüder, B., Schurr, F. & Wilson, R. (2007), 'Methods to account for spatial autocorrelation in the analysis of species distributional data: a review', *Ecography* **30**, 609–628.

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

Elith, J. & Burgaman, M. (2002), *Population Viability in Plants*, chapter Habitat models for PVA.

Elith, J., Graham, C. H., Anderson, R. P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M., Peterson, A. T., Phillips, S. J., Richardson, K. S., Scachetti-Pereira, R., Schapire, R. E., Soberon, J., Williams, S., Wisz, M. S. & Zimmermann, N. E. (2006), 'Novel methods improve prediction of species'distribution from occurence data', *Ecography* **29**, 129–151.

Elith, J., Leathwick, J. & Hastie, T. (2008), 'A working guide to boosted regression trees', *Journal of Animal Ecology* **77**, 802–813.

Elith, J. & Leathwick, J. R. (2009), 'Species distribution models: ecological explanation and prediction across space and time', *Annual Review of Ecology, Evolution and Systematics* **40**, 677–697.

Elith, J., Phillips, S., Hastie, T., Dudík, M., Chee, Y. & Yates, C. (2011), 'A statistical explanation of MaxEnt for ecologists', *Diversity and Distributions* **17**, 43–57.

Engler, R., Guisan, A. & Rechsteiner, L. (2004), ‘An improved approach for predicting the distribution of rare and endagered species from occurence and pseudo-absence data’, *Journal of Applied Ecology* **41**, 263–274.

Farcomeni, A. (2010), ‘Bayesian Constrained Variable Selection’, *Statistica Sinica* **20**, 1043–1062.

Ferrier, S. & Watson, G. (1996), ‘An evaluation of the Effectiveness of Environmental Surrogates and Modelling Techniques in Predicting the Distribution of Biological Diversity’, *Consultancy report prepared by the New South Wales National Parks and Wildlife Service for the Department of Environmental, Sport and Territories* .

Ferrier, S. & Watson, G. (2009), ‘Species distribution models: ecological explanation and prediction across space and time’, *Environmenatal* **40**, 677–697.

Ferrier, S., Watson, G., Pearce, J. & Drielsma, M. (2002), ‘Extended statistical approaches to modelling spatial pattern in biodiversity in northest New South Wales. I. Species-level modelling’, *Biodiversity and Conservation* **11**, 2275–2307.

Franklin, J. (2010), *Mapping Species Distributions: Spatial Inference and Prediction*, Cambridge University Press, Cambridge, UK.

Garthwaite, P., Kadane, J. & O’Hagan, A. (2005), Statistical Methods for eliciting probability distributions, Technical Report 808, Carnegie Mellon University.

Gaston, K. (2003), *The structure and dynamics of geographic ranges*, Oxford University Press, New York, USA.

Gelfand, A. (2010), *Handbook of spatial statistics*, chapter Misaligned spatial data: The change of support problem, pp. 517–539.

Gelfand, A., Schmidt, A., Wu, S., Silander, J. & A., L. (2005), ‘Modelling species diversity through species level hierarchical modelling’, *Applied Statistics* **54(1)**, 1–20.

Gilks, W. R., Best, N. G. & Tan, K. K. C. (1995), ‘Adaptive Rejection Metropolis Sampling Within Gibbs Sampling (Corr: 97V46 P541-542 With R. M. Neal)’, *Applied Statistics* **44**, 455–472.

Guisan, A., Edwards, J. & Hastie, T. (2002), ‘Generalized linear and generalized additive models in studies of species distributions: setting the scene’, *Ecological Modelling* **157**, 89–100.

Guisan, A. & Zimmermann, N. E. (2000), 'Predictive habitat distribution models in ecology', *Ecological Modelling* **135**, 147–186.

Gumpertz, M., Graham, J. & Reistaino, J. (1997), 'Autologistic model of spatial pattern of Phytophora epidemic in bell pepper: Effects of soil variables on sisease presence', *Journal of Agricolture, Biological and Environmental Statistics* **2**, 131–156.

Guo, Q., Kelly, M. & Graham, C. (2005), 'Support vector machines for predicting distribution of Sudden Oak Death in California', *Ecological Modelling* **182**, 75–90.

Hastie, T. & Tibshirani, R. J. (1990), *Generalized Additive Models*, Chapman & Hall, London.

He, F., Zhou, J. & Zhu, H. (2003), 'Autologistic regression model for the distribution vegetation', *Journal of Agricolture, Biological and Environmental Statistics* **8(2)**, 205–222.

Heikkinen, J. & Högmander, H. (1994), 'Fully Bayesian approach to image restoration with an application in biogeography', *Applied Statistics - Journal of the Royal Statistical Society, Series C* **43**, 569–582.

Hirzel, A. & Guisan, A. (2002), 'Which is the optimal sampling strategies for habitat suitability modelling?', *Ecological Modelling* **157**, 331–341.

Hoeting, J., Leecaster, M. & Bowed, D. (2000), 'An improved model for spatially correlated binary responses', *Journal of Agricolture, Biological and Environmental Statistics* **5(1)**, 102–114.

Hosmer, D. & Lemeshow, S. (2000), *Applied logistic regression analysis. Second Edition*, John Wiley & Sons, New York, New York, USA.

Huang, B. & Salleb-Aouissi, A. (2009), Maximum Entropy Density Estimation with Incomplete Presence-only Data, *in* 'JMLR: W&CP 5', Vol. 5, Clearwater Beach, Florida, USA.

Huffer, F. & Wu, H. (1998), 'Markov chain Monte Carlo for autologistic regression models with application to the distribution of plant species', *Biometrics* **54(2)**, 509–524.

Hughes, J., Haran, M. & Caragea, P. (2011), 'Autologistic models for binary data on a lattice', *Envirometrics* **22(7)**, 857–871.

Jaynes, E. (1957), 'Information theory and statistical mechanism', *Phys. Rev.* **106**, 620–630.

Kadane, J., Dickey, J., Winkler, R., Smith, W. & Peters, S. (1980), 'Interactive elicitation of opinion for a normal linear model', *Journal of the American Statistical Association* **75**, 845–854.

Kadane, J. & Wolfson, L. (1998), 'Experiences in elicitation', *The statistician* **47**, 3–19.

Kaschner, K., Watson, R., Trites, A. W. & Pauly, D. (2006), 'Mapping word-wide distributions of marine species using a relative environmental suitability (RES) model', *Marine Ecology Progress Series* **316**, 285–310.

Keating, K. A. & Cherry, S. (2004), 'Use and interpretation of logistic regression in habitat-selection studies', *Journal of Wildlife Management* **68**, 774–789.

Kruschke, J. (2010), *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*, Academic Press, USA.

Lancaster, T. & Imbens, G. (1996), 'Case-control studies with contaminated control', *Journal of Econometrics* **71**, 145–160.

Latimer, A., Wu, S., Gelfand, A. & Silander Jr., A. (2006), 'Building statistical models to analize species distributions', *Ecological Applications* **16(1)**, 33–50.

Law, B. (1994), 'Climatic limation of the southern distribution of the common blossom bat *Syconycteris autralis*, New South Wales', *Australian Journal of Ecology* **19**, 366–374.

Lek, S., Delacoste, M., Baran, P., Dimopoulus, I., Lauga, J. & Aulagnier, S. (1996), 'Application of neural networks to modelling nonlimear relationships in ecology', *Ecological Modelling* **90**, 39–52.

Lillesand, T. M., Kiefer, R. W. & Chipman, J. W. (2004), *Remote sensing and image restoration*, John Wiley & Sons Ltd.

Lindemayer, D., Nix, H., McMahon, J., M.F., H. & Tanton, M. (1991), 'The conservation of leadbeather's possum, *Gymobelideus leadbeateri* (McCoy): a case study of the use of bioclimatic modelling', *Journal of Biogegraphy* **18**, 371–383.

MacNab, Y. (2003), 'Hierarchical Bayesian Modeling of Spatially Correlated Health Service Outcome and Utilization Rates', *Biometrics* **59**, 305–316.

Manly, B. F. J., McDonald, L. L., Thomas, D. L., McDonald, T. L. & Erickson, W. P. (2002), *Resource Selection by Animals*, 2nd edn. Kluwer Academic Publisher, Dordrecht, the Netherlands.

McCullagh, P. & Nelder, J. A. (1989), *Generalized Linear Models*, Chapman & Hall, CRC, London.

Nielsen, S. E., Johnson, C. J., Heard, D. C. & Boyce, M. S. (2005), 'Can models of presence-absence be used to scale abundance? Two case studies considering extremes in life history', *Ecography* **28**, 1–12.

Ntzoufras, I. (2009), *Bayesian Modeling Using WinBUGS*, John Wiley & Sons, Inc., New Jersey.

Osborne, P. E., Alonso, J. C. & G., B. R. (2001), 'Modelling landscape-scale habitat use using GIS and remote sensing', *Journal of Applied Ecology* **38**, 458–471.

Paciorek, C. (2007), 'Computational Techniques for Spatial Logistic Regression with Large Datasets', *Computational Statistics and Data Analysis* **51(8)**, 3631–3653.

Pearce, J. L. & Boyce, M. S. (2006), 'Modelling distribution and abundance with presence-only', *Journal of Applied Ecology* **43**, 405–412.

Pearce, J. & Lindermayer, D. (1998), 'Bioclimatic ananlysis to enhance reintroduction biology of the endangered Helmeted Honeyeater (*Lichenostomus melanops cassidix*) in southearnest Australia', *Restoration Ecology* **6**, 238–243.

Phillips, S. L., Anderson, R. P. & Schapire, R. E. (2006), 'Maximum entropy modeling of species geographic distributions', *Ecological Modelling* **190**, 231–259.

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

Scarnati, L., Attorre, F., De Sanctis, M., Farcomeni, A., Francesconi, F., Mancini, M. & Bruno, F. (2009), 'A multiple approach for the evaluation of the spatial distribution and dynamics of a forest habitat: the case of Apennine beech forests with *Taxus baccata* and *Ilex aquifolium*', *Biodiversity and conservation* **18**, 3099–3113.

Schlesselman, J. (1982), *Case-control studies: design, conduct, analysis*, Oxford University Press, Inc, New York, USA.

Scott, J. M. & Csuti, B. (1997), *Biodiversity, II. Understanding and protecting our biological resources*, Washington, DC, USA, chapter Gap analysis for biodiversity survey and maintenance, pp. 321–340.

Spiegelhalter, D., Thomas, N. & Lunn, D. (2003), *WinBUGS user manual, Version 1.4.1*, Medical Research Council Biostatistics Unit, Istitute of Public Health, Cambridge, UK.

Stein, B. R. & Wieczorek, J. (2004), 'Mammals of the world: Manis as an example of data integration in a distributed network environmental', *Biodiversity Informatics* **1**, 14–22.

Stockwell, D. (2007), *Niche modeling predictions from statistical distributions*, Chapman & Hall.

Stockwell, D. & Peters, D. (1999), 'The GARP modelling system: problems and solutions to automed spatial prediction', *International Journal of Geographyraphic Information Science* **2**, 143–158.

Sun, D., Tsutakawa, R. & Speckman, P. (1999), 'Posterior distribution of hierarchical models using CAR(1) distributions', *Biometrics* **86**, 341–350.

Ver Hoef, J., Cressie, N., Fisher, R. & Case, T. (2001), *Spatial uncertainty in ecology*, chapter Uncertainty and spatial linear models for ecological data, pp. 214–237.

Wakefield, J. (2007), 'Disease mapping and spatial regression with count data ', *Biostatistics* **8(2)**, 158–183.

Walker, P. A. & Cocks, K. D. (1991), 'HABITAT: A Procedure for Modelling a Disjoint Environmental Envelope for a Plant or Animal Species', *Global Ecology and Biogeography Letters* **2**, 108–118.

Walther, B., Wisz, M. & Rahbek, C. (2004), 'Known and predicted African winter distributions and habitat use of the endareged Basra reed warbler (*Acrocephalus griseldis*) and the near-threatened cinereous bunting (*Emberiza cineracea*)', *Journal of Ornithology* **88**, 287–299.

Ward, G., Hastie, T., Barry, S., Elith, J. & Leathwick, A. (2009), 'Presence-only data and the EM algorithm', *Biometrics* **65**, 554–563.

Warton, D. I. & Shepherd, L. (2010), 'Poisson point porcess models solve the "pseudo-absence problem" for presence-only data in ecology', *Annals of Applied Statistics* **4(3)**, 1383–1402.

Weimerskirch, H., Bonadonna, F., Bailleul, F., Mabille, G., Dell'Olmo, G. & Lipp, H. (2002), 'Gps tracking of foraging albatrosses', *Science* pp. 295–1259.

Welsh, A. H., Cunningham, R. B., Donnelly, C. F. & Lindenmayer, D. B. (1996), 'Modelling the abundance of rare species: statistical models for aounts with extra zeros', *Ecological Modelling* **88**, 297–308.

Wikle, C. (2003), 'Hierarchical Bayesian models for predicting the spread of ecological processes.', *Ecology* **84**, 1382–1394.

Woodward, M. (1999), *Epidemology: Study design and Data Analysis*, Chapman & Hall/CRC, USA.

Zaniewski, A. E., Lehmann, A. & Overton, J. (2002), 'Predicting species spatial distribution using presence only data, a case study of native New Zeland ferns', *Ecological Modelling* **157**, 261–280.