Bayesian Modeling and MCMC Computation in Linear Logistic Regression for Presence-only Data

Fabio Divino[1],Natalia Golini[2],Giovanna Jona Lasinio[2]
Antti Penttinen[3]

1-Division of Physics, Computer Science and Mathematics, University of Molise

2-Department of Statistical Sciences, University of Rome *"La Sapienza"*

3-Department of Mathematics and Statistics, University of Jyväskylä

April 29, 2013

Abstract

Presence-only data are referred to situations in which, given a censoring mechanism, a binary response can be observed only with respect to on outcome, usually called *presence*. In this work we present a Bayesian approach to the problem of presence-only data based on a two levels scheme. A probability law and a case-control design are combined to handle the double source of uncertainty: one due to the censoring and one due to the sampling. We propose a new formalization for the logistic model with presence-only data that allows further insight into inferential issues related to the model. We concentrate on the case of the linear logistic regression and, in order to make inference on the parameters of interest, we present a Markov Chain Monte Carlo algorithm with data augmentation that does not require the a priori knowledge of the population prevalence. A simulation study concerning 24,000 simulated datasets related to different scenarios is presented comparing our proposal to optimal benchmarks.

Keywords: Bayesian modeling, case-control design, data augmentation, logistic regression, Markov Chain Monte Carlo, population prevalence, presence-only data, simulation.

1   Introduction

There is a significant body of literature in statistics, econometrics and ecology dealing with the modeling of discrete responses under biased or preferential sampling designs. They are particularly popular in the natural sciences when species distributions are studied. Such sample schemes may reduce the survey cost especially when one of the responses is

rare. A large part of statistical literature concerns the case-control design, retrospective, choice-based or response-based sampling (Lancaster and Imbens, 1996). In the simplest case a sample of cases and a sample of controls are available and for each observation a set of "attributes/covariates" is observed in both samples. Then inference is carried out following standard statistical procedures (Armenian, 2009).

A case that has received increasing attention in the literature is the situation where the sample of controls is a random sample from the whole population with information only on the attributes and not on the response (Lancaster and Imbens, 1996). This situation is fairly common in ecological studies where only species' presence is recorded when field surveys are carried out. In the ecological literature, since the 1990's such data are called *presence only data* (see Araùjo and Williams, 2000, and references therein). Pearce and Boyce (2006) define presence-only data as "consisting only of observations of the organism but with no reliable data where the species was not found". Atlases, museum and herbarium records, species lists, incidental observation databases and radio-tracking studies are examples of such data.

In recent years we find a considerably growing literature describing approaches to the modeling of this type of data, among the many ecological papers we recall Keating and Cherry (2004), Pearce and Boyce (2006), Elith *et al.* (2006), Elith and Leathwick (2009), Franklin (2010) and, most notably, in the statistical literature Ward *et al.* (2009), Warton and Shepherd (2010), Chakraborty *et al.* (2011), Di Lorenzo *et al.* (2011) and Dorazio (2012). While in Warton and Shepherd (2010) and Chakraborty *et al.* (2011) to model the presence-only data Poisson point processes are considered in the likelihood and Bayesian framework respectively, in Ward *et al.* (2009) and Di Lorenzo *et al.* (2011) a modified case-control logistic model is adopted in the likelihood and Bayesian perspective respectively, in both papers there is no account for possible dependence structure in the observations. In Dorazio (2012) the asymptotic relations between the two approaches are discussed.

A different approach, MaxEnt, is based on the maximum entropy principle (Jaynes, 1957). In MaxEnt (Phillips *et al.*, 2006; Elith *et al.*, 2011) the relative entropy between the distribution of covariates at locations where the species is present and the unconditional background distribution of covariates is maximized subject to some constrains concerning empirical statistics (see Philips et al., 2006, for details). As pointed by Dorazio (2012) "the MaxEnt method requires knowledge of species' prevalence for its estimator of occurrence to be consistent".

In what follows we are going to use the name *presence-only data* when referring to the above sketched general problem of having information on the presence and covariates jointly on a sample from a population, while information on only the covariates is available on any sample from the same population. This work is developed in the same discrete setting as in Ward *et al.* (2009) and Di Lorenzo *et al.* (2011), i.e., we have a population of independent units, no dependence structure, such as spatial correlation, is anticipated. We defer the treatment of this extension to a subsequent work.

The main contribution of the paper is a new rigorous formalization of the logistic regression model with presence-only data that allows further insight into the inferential issues. This leads us to an algorithmic procedure that, among other results, returns a MCMC approximation of the response prevalence under general knowledge of the process generating

the data. We also present a large simulation study involving 24,000 simulated datasets and comparing our approach to other two models representing optimal benchmarks.

The paper is organized as follows. Section 2 introduces a general framework for the presence-only data problems, Section 3 presents our Bayesian approach, Section 4 describes the MCMC algorithm while results related to the simulation study are reported in Section 5. Finally in Section 6 some conclusions are drawn and future developments briefly described.

2 Linear logistic regression for presence-only data.

The analysis of a binary response related to a set of explicative covariates is usually carried out through the use of the logistic regression where the logit of the conditional probability of occurrence is modeled as a function of covariates. In this section, we first introduce a general framework for the modeling of presence-only data and then consider the case of the linear logistic regression. The approach proposed is built on two levels and we partially follow the formulation introduced by Ward *et al.* (2009) but adopting a Bayesian scheme as in Divino *et al.* (2011).

2.1 A two level approach.

Let Y be a binary variable informing on the presence ($Y = 1$) or absence ($Y = 0$) of a population's attribute and let $X = (X_1, ..., X_k)$ denote a set of highly informative, on the same attribute, covariates which are available on the same population. Then, the presence-only problem can be formalized by considering a censorship mechanism that acts when observing the response Y, so that part of the population units are not reachable. In particular, we refer to the situation in which we are able to detect only a partial set of units on which the attribute of interest is present while the information on the covariates X is available on the entire population. In this situation, we have to consider two types of uncertainty: the uncertainty due to the mechanism of censorship and the uncertainty due to the sampling procedure. Moreover, since we are not able to collect a random sample of observable data, we need to adjust for the sampling mechanism through the use of a case-control scheme (Breslow and Dey, 1980; Breslow, 2005; Armenian, 2009) .

In order to build a Bayesian model, in this framework we adopt the following conceptual scheme in two levels.

Level 1. Given the population of interest $\mathcal{U}$ of size N, the binary responses $\mathbf{y} = (y_1, ..., y_N)$ are generated independently by a probability law $\mathcal{M}$.

Level 2. Let $\mathcal{U}_p$ be the subset of $\mathcal{U}$ where we observe $Y = 1$. A modified case-control design is applied so that a sample of presences, considered as cases, is selected from $\mathcal{U}_p$

and a sample of "contaminated" controls (Lancaster and Imbens, 1996) is selected from the whole population $\mathcal{U}$, with all the covariates but no information on Y.

Here, we cannot approach the model construction using only a finite population approach (Särndal, 1978) because of the censoring mechanism that "masks" distributional information on Y already at the population level. By the introduction of Level 1 we can describe the censored observations as random quantities generated by the model $\mathcal{M}$. Hence, the problem of presence-only data can be formalized as a problem of missing data (Rubin, 1976; Little and Rubin, 1987).

2.2 The model generating population data.

At the first level, we assume that the law $\mathcal{M}$ is defined in terms of the conditional probability of occurrence $Pr(Y = 1|x)$, denoted by $\pi^*(x)$, when the covariates are $X = x$. Moreover, we consider that the relation between Y and X is formalized through a regression function $\phi(x)$ on the logit scale

$$\phi(x) = \operatorname{logit} \pi^*(x), \tag{1}$$

that is

$$\pi^*(x) = \frac{e^{\phi(x)}}{1 + e^{\phi(x)}}. \tag{2}$$

When the data $\mathbf{y} = (y_1, ..., y_N)$ are independently generated from $\mathcal{M}$, we denote by π the empirical prevalence of the binary response Y in $\mathcal{U}$, expressed as the ratio of the number of presences N_1 to the size of the population, that is

$$\pi = \frac{N_1}{N}.$$

2.3 The modified case-control design.

At the second level, we adopt a case-control design modified for presence-only data (Lancaster and Imbens, 1996) in order to account for the specific sampling procedure considered. The use of the case-control scheme is necessary at all times when it is appropriate to select observations in fixed proportions with respect to the values of the response variable. This can occur when the attribute of interest represents a phenomenon that is rare among the units of the population as for example a rare disease or a rare exposure in epidemiological studies (Woodward, 2005).

Now, let C be a binary indicator of inclusion into the sample ($C = 1$ denotes that a unit is in the sample), let $\rho_0 = Pr(C = 1|Y = 0)$ and $\rho_1 = Pr(C = 1|Y = 1)$ be the inclusion probability of the absences and the presences, respectively. Under the assumption that, given Y, the sampling mechanism is independent from the covariates X, the conditional probability of occurrence is modified through the Bayes rule as

$$Pr(Y = 1|C = 1, x) = \frac{\rho_1 e^{\phi(x)}}{\rho_0 + \rho_1 e^{\phi(x)}}. \tag{3}$$

4

Hence, the corresponding case-control regression function $\phi_{cc}(x)$ defined as the logit of (3) is given by

$$\phi_{cc}(x) = \phi(x) + \log \frac{\rho_1}{\rho_0}. \tag{4}$$

In particular, if the selection of cases (n_1) and controls (n_0) is made independently without replacement, the inclusion probabilities are given in terms of the empirical prevalence π by

$$\rho_0 = \frac{n_0}{(1 - \pi)N}$$

and

$$\rho_1 = \frac{n_1}{\pi N},$$

so that the equation (4) becomes

$$\phi_{cc}(x) = \phi(x) + \log \frac{n_1}{n_0} - \log \frac{\pi}{1 - \pi}. \tag{5}$$

In our framework, since the response variable Y is already censored at the population level, the standard case-control design cannot be adopted but it should be modified in such a way that a sample of presences is matched with an independent sample drawn from the entire population, named the *background sample* (Zaniewski *et al.*, 2002; Ward *et al.*, 2009). Remark that in this sample the response variable is unobserved and only the covariates are available.

In this way, the complete sample S is composed by a set S_u of n_u independent background data, where the response Y is not observed, drawn from the entire $\mathcal{U}$ and by a set S_p of n_p independent observations selected from the sub-population of presences $\mathcal{U}_p$. This procedure implies that the reference population $\mathcal{U}$ is augmented with its subset $\mathcal{U}_p$ so that the total number of observations considered in the sampling scheme becomes $N + N_1$. To illustrate the sampling framework we are going to adopt here, let us consider the following situation: we can label population units of type $y = 1$ only when they are isolated from units of type $y = 0$. This can be formalized by introducing a binary stratum variable Z such that $Z = 0$ indicates when an observation is drawn from the entire population $\mathcal{U}$ while $Z = 1$ denotes the sampling from the sub-population $\mathcal{U}_p$. Remark that $Z = 1$ implies $Y = 1$ whilst $Z = 0$ implies that Y is an unknown value $y \in \{0, 1\}$. Moreover, by construction Z is independent from the covariates X, given the response Y. The introduction of the variable Z allows us to define the structure of the data at the population level and at the sample level in terms of presences/absences (Y) and known/unknown data (Z), as reported in Table 1 and Table 2.

Y/Z	$Z = 0$	$Z = 1$	Total
$Y = 0$	N_0	0	N_0
$Y = 1$	N_1	N_1	$2N_1$
Total	N	N_1	$N + N_1$

Table 1: Data structure at the population level.

Y/Z	$Z = 0$	$Z = 1$	Total
$Y = 0$	n_{0u}	0	n_0
$Y = 1$	n_{1u}	n_p	n_1
Total	n_u	n_p	n

Table 2: Data structure at the sample level.

In Table 1, N_0 is the number of absences in the population $\mathcal{U}$ while in Table 2, n_{0u} and n_{1u} respectively denote the unknown frequencies of absences and presences in the sub-sample S_u. Remark that, in the above described situation, the inclusion probability of units with or without the mentioned attribute changes. In fact, while an absence can be drawn only when sampling from $\mathcal{U}$, a presence can be selected when sampling both from $\mathcal{U}$ and from $\mathcal{U}_p$. Thus, one has

$$\rho_0 = \frac{n_0}{N_0} = \frac{n_{0u}}{(1-\pi)N}, \tag{6}$$

and

$$\rho_1 = \frac{n_1}{2N_1} = \frac{n_{1u} + n_p}{2\pi N}. \tag{7}$$

The introduction of the stratum variable Z allows us also to exactly derive the logistic regression model under the case-control design modified for presence-only data. In fact, when we consider the population $\mathcal{U}$ augmented with its subset $\mathcal{U}_p$, the model $\pi^*(x)$ represents the conditional probability to mark a presence only when $Z = 0$, that is $Pr(Y = 1|Z = 0, x) = \pi^*(x)$. On the other hand, when $Z = 1$, we simply have $Pr(Y = 1|Z = 1, x) = 1$. We can prove the following result.

Proposition 1. Under the assumption that Z is independent from X given Y, one has

$$Pr(Y = 1|x) = \frac{2\pi^*(x)}{1 + \pi^*(x)}. \tag{8}$$

Proof. From the hypothesis of conditional independence it results

$$Pr(Z|Y, x) = Pr(Z|Y),$$

that can be express also as

$$\frac{Pr(Y|Z, x)Pr(Z|x)}{Pr(Y|x)} = \frac{Pr(Y|Z)Pr(Z)}{Pr(Y)}.$$

Let consider the case with $Y = 1$ and $Z = 0$, one has

$$\frac{Pr(Y = 1|Z = 0, x)Pr(Z = 0|x)}{Pr(Y = 1|x)} = \frac{Pr(Y = 1|Z = 0)Pr(Z = 0)}{Pr(Y = 1)}.$$

The probabilities enclosed in the second term can be derived from Table 1 and one has

$$\frac{\pi^*(x)Pr(Z = 0|x)}{Pr(Y = 1|x)} = \frac{\frac{N_1}{N}\frac{N}{N+N_1}}{\frac{2N_1}{N+N_1}} = \frac{1}{2}. \tag{9}$$

In the case $Y = 1$ and $Z = 1$ one similarly obtains

$$\frac{Pr(Z = 1|x)}{Pr(Y = 1|x)} = \frac{\frac{N_1}{N_1}\frac{N_1}{N+N_1}}{\frac{2N_1}{N+N_1}} = \frac{1}{2}. \tag{10}$$

From (10) it results $Pr(Y = 1|x) = 2\,Pr(Z = 1|x)$ and by substituting in (9), one can derive that $Pr(Z = 0|x) = \dfrac{1}{1 + \pi^*(x)}$ and hence $Pr(Z = 1|x) = \dfrac{\pi^*(x)}{1 + \pi^*(x)}$. Now, it is simple to obtain that

$$Pr(Y = 1|x) = \frac{2\pi^*(x)}{1 + \pi^*(x)}.$$

□

If we assume that, given Y, the inclusion into the sample $(C = 1)$ is independent from the covariates X, one has [1]

$$Pr(Y = 0|C = 1, x)\,Pr(C = 1|x) = \frac{1 - \pi^*(x)}{1 + \pi^*(x)}\rho_0 \tag{11}$$

and

$$Pr(Y = 1|C = 1, x)\,Pr(C = 1|x) = \frac{2\pi^*(x)}{1 + \pi^*(x)}\rho_1. \tag{12}$$

Then, from the ratio of (12) to (11), it results

$$\frac{Pr(Y = 1|C = 1, x)}{Pr(Y = 0|C = 1, x)} = \frac{2\pi^*(x)}{1 - \pi^*(x)}\frac{\rho_1}{\rho_0},$$

and by plugging the quantities ρ_0 and ρ_1, as defined in (6) and in (7), into the logit of $Pr(Y = 1|C = 1, x)$, one obtains the following relation

$$
\begin{aligned}
\text{logit}\,Pr(Y = 1|C = 1, x) \;&=\; \log\left[\frac{2\pi^*(x)}{1 - \pi^*(x)}\frac{\rho_1}{\rho_0}\right] \\
&=\; \log\left[\frac{2\pi^*(x)}{1 - \pi^*(x)}\frac{n_{1u} + n_p}{n_{0u}}\frac{1 - \pi}{2\pi}\right] \\
&=\; \log\left[\frac{\pi^*(x)}{1 - \pi(x)}\frac{n_{1u} + n_p}{n_{0u}}\frac{1 - \pi}{\pi}\right] \\
&=\; \text{logit}\,\pi^*(x) + \log\frac{n_{1u} + n_p}{n_{0u}} - \log\frac{\pi}{1 - \pi} \\
&=\; \phi(x) + \log\frac{n_{1u} + n_p}{n_{0u}} - \log\frac{\pi}{1 - \pi}, \tag{13}
\end{aligned}
$$

[1]see Appendix for the detailed proof.

that represents the logistic regression model under the case-control design for presence-only data. As well, we can now formalize the presence-only data regression function $\phi_{pod}(x)$ as

$$\phi_{pod}(x) = \phi(x) + \log \frac{n_{1u} + n_p}{n_{0u}} - \log \frac{\pi}{1 - \pi}. \tag{14}$$

Although the derivation is substantially different, we end with the same formulation as in Ward et $al.$ (2009). Now, in order to make parameter estimation possible, we need to handle the ratio

$$\frac{\rho_1}{\rho_0} = \frac{n_{1u} + n_p}{n_{0u}} \frac{1 - \pi}{2\pi}, \tag{15}$$

where the quantities π and n_{1u} are unknown ($n_{0u} = n_u - n_{1u}$).

In the recent literature, two main approaches have been proposed. The first one by Ward et $al.$ (2009) replace the ratio $\frac{n_{1u} + n_p}{n_{0u}}$ with the ratio of the expected numbers of presences and absences in the sample, that is

$$\frac{\rho_1}{\rho_0} \approx \frac{E[n_{1u} + n_p]}{E[n_{0u}]} \frac{1 - \pi}{2\pi} = \frac{\pi n_u + n_p}{(1 - \pi)n_u} \frac{1 - \pi}{2\pi} = \frac{\pi n_u + n_p}{2\pi n_u}. \tag{16}$$

These authors adopt a likelihood approach and computation is carried out via the EM algorithm. As they underline, this approximation can be easily implemented if the empirical population prevalence π is known a priori. They discuss also the possibility to estimate π jointly with the regression function when the prevalence is identifiable, as for example in the linear logistic regression, and with respect to this case they present a simulation example. The difficulty in obtaining efficient joint estimates because of the correlation between π and the intercept of the linear regression term is discussed. Notice that Ward et $al.$ (2009) considers a slightly different representation of the ratio (16), omitting the multiplier "2" in the denominator.

Di Lorenzo et $al.$ (2011), dealing with a problem of abundance data, use the approximation (16), but they adopt a Bayesian approach and consider the population prevalence π as a further parameter in the model. They choose an informative $Beta$ prior for π, but their MCMC algorithm contains an unusual weakness since the simulation of π is performed from its prior and not from the posterior that can be derived through the interaction between the parameter π and the regression function $\phi(x)$.

A different approximation of the ratio (15) can be obtained by considering the sample prevalence in S_u (the background sample)

$$\pi_u = \frac{n_{1u}}{n_u},$$

where

$$n_{1u} = \sum_{i \in S_u} y_i.$$

Due to the censorship process, this quantity is unknown but it would be the maximum likelihood estimator for π if the data $\mathbf{y}_u = \{y_i, i \in S_u\}$ could be observed. Now, replacing π by π_u in (15) one obtains

$$\frac{\rho_1}{\rho_0} \approx \frac{n_{1u} + n_p}{n_{0u}} \frac{1 - \pi_u}{2\pi_u} = \frac{n_{1u} + n_p}{n_u - n_{1u}} \frac{n_u - n_{1u}}{2n_{1u}} = \frac{n_{1u} + n_p}{2n_{1u}}, \tag{17}$$

8

that allows to formulate a computable version of the regression function for presence-only data as

$$\phi_{pod}(x) \approx \phi(x) + \log \frac{n_{1u} + n_p}{n_{1u}}. \tag{18}$$

This function depends on the data $\mathbf{y}_u$ in S_u which are not directly observable, but if $\mathbf{y}_u$ is treated as missing data one can enclose it into the estimation process and then obtain a consistent approximation for $\phi_{pod}(x)$. In particular, in a Bayesian framework, this idea can be performed by using a Markov Chain Monte Carlo computation with data augmentation. Moreover from the use of MCMC simulations we can also obtain an approximation of π_u and therefore an estimate of the empirical population prevalence π. Details are given in Section 4.

The approximation (17) can, in principle, be always adopted, but some care must be used as identifiability issues are present. We follow the recommendation in Ward *et al.* (2009) to approach jointly estimates of $\phi(x)$ and π only when the latter is identifiable with respect to the regression function, as for example in the linear regression case (see Ward *et al.*, 2009, for mathematical details).

2.4 The linear logistic regression.

If we consider a linear regression function $\phi(x) = x\beta$, where $\beta = (\beta_1, ..., \beta_k)$ is the vector of the regression parameters, a computable model for presence-only data can be defined through the following approximation

$$\phi_{pod}(x) \approx x\beta + \log \frac{n_{1u} + n_p}{n_{1u}}, \tag{19}$$

or equivalently through the approximation of the conditional probability of occurrence at the sample level

$$
\begin{aligned}
Pr(Y = 1 | C = 1, x) &\approx \frac{\exp\{x\beta + \log \frac{n_{1u}+n_p}{n_{1u}}\}}{1 + \exp\{x\beta + \log \frac{n_{1u}+n_p}{n_{1u}}\}} \\
&= \frac{\left(1 + \frac{n_p}{n_{1u}}\right) \exp\{x\beta\}}{1 + \left(1 + \frac{n_p}{n_{1u}}\right) \exp\{x\beta\}}.
\end{aligned} \tag{20}
$$

In this particular case, all the unknowns of the model are the linear parameters vector β and the missing data $\mathbf{y}_u$ in the background sample S_u.

3 The hierarchical Bayesian model.

Due to the censorship process affecting the data, we can acquire complete information only on the stratum variable Z and not on the binary response Y. Then, it seems natural to model Z as the observable variable. If we consider the conditional joint distribution of Z and Y

$$Pr(Z, Y | C = 1, x) = Pr(Z | Y, C = 1, x) Pr(Y | C = 1, x), \tag{21}$$

through the marginalization over Y, the probability $Pr(Z|C = 1, x)$ can be obtained and we can express the relation between presences and covariates in terms of regression of Z respect to X. Notice that, while $Pr(Y|C = 1, x)$ can be obtained from (20), the term $Pr(Z|Y, C = 1, x)$, due to the conditional independence between Z and X given Y, simply reduce to be equal to $Pr(Z|Y, C = 1)$ that can be derived from Table 2.

We point out that, even if the response Y does not play an explicit role after the marginalization step, we need to keep it in the model as a hidden variable in order to obtain the approximation for the quantity $n_{1u} = \sum_{i \in S_u} y_i$, necessary to correct the linear regression function for presence-only data.

Now, we can formalize the hierarchical Bayesian model to estimate the parameters of a linear logistic regression under the case-control scheme adjusted for presence-only data. In order to better explain the conditional relationship underlying the hierarchy, we introduce the graph in Figure 1. The dashed node indicates a variable hidden with respect to the conditional relationships.
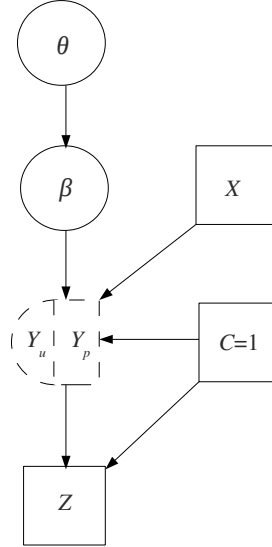

Figure 1: Graphical representation of the hierarchical Bayesian model.

The priors. At the top of the hierarchy, we assume the hyper parameter θ distributed as $p(\theta)$. At the second level, we consider the prior probability distribution on β depending

on the hyper parameter θ, that is $\beta|\theta \sim p(\beta|\theta)$. At the third level, the unobserved data $\mathbf{y}_u$ in S_u are considered latent parameters with prior distribution *Bernoulli* (denoted by *Be*) with probability of occurrence given by the approximation in (20), that is

$$y_i|C_i = 1, x_i, \beta \sim Be\left(\frac{\left(1 + \frac{n_p}{n_{1u}}\right)\exp\{x\beta\}}{1 + \left(1 + \frac{n_p}{n_{1u}}\right)\exp\{x\beta\}}\right), \; i \in S_u.$$

This point is important for deriving the predictive distribution of the unobserved data $\mathbf{y}_u$ necessary in the estimation algorithm.

The likelihood. At the lowest level of the hierarchy, we have the likelihood, defined with respect to the observable stratum variable Z. Recalling that from the Table 2 we have $Pr(Z = 1|Y = 0, C = 1) = 0$ and $Pr(Z = 1|Y = 1, C = 1) = \dfrac{n_p}{n_{1u} + n_p}$, when (21) is marginalized over Y, one obtains the approximation

$$
\begin{aligned}
Pr(Z = 1|C = 1, x, \beta) &\approx \frac{n_p}{n_{1u} + n_p} \frac{\exp\{x\beta + \log\frac{n_{1u}+n_p}{n_{1u}}\}}{1 + \exp\{x\beta + \log\frac{n_{1u}+n_p}{n_{1u}}\}} \\
&= \frac{\frac{n_p}{n_{1u}}\exp\{x\beta\}}{1 + \left(1 + \frac{n_p}{n_{1u}}\right)\exp\{x\beta\}}
\end{aligned}
\tag{22}
$$

and hence

$$
\begin{aligned}
Pr(Z = 0|C = 1, x, \beta) &= 1 - Pr(Z = 1|C = 1, x, \beta) \\
&\approx \frac{1 + \exp\{x\beta\}}{1 + \left(1 + \frac{n_p}{n_{1u}}\right)\exp\{x\beta\}}.
\end{aligned}
\tag{23}
$$

Thus, we can assume that for all $i \in S$ the conditional distribution of Z_i is *Bernoulli* with probability of occurence given by (22), that is

$$Z_i|C_i = 1, x_i, \beta \sim Be\left(\frac{\frac{n_p}{n_{1u}}\exp\{x\beta\}}{1 + \left(1 + \frac{n_p}{n_{1u}}\right)\exp\{x\beta\}}\right), \; i \in S.$$

Recalling that $Z_i = 0$ for all $i \in S_u$ while $Z_i = 1$ for all $i \in S_p$, the likelihood function can be written as

$$L(\beta; \mathbf{z}, \mathbf{x}) = \prod_{i \in S_u} \frac{1 + \exp\{x_i\beta\}}{1 + \left(1 + \frac{n_p}{n_{1u}}\right)\exp\{x_i\beta\}} \times \prod_{i \in S_p} \frac{\frac{n_p}{n_{1u}}\exp\{x_i\beta\}}{1 + \left(1 + \frac{n_p}{n_{1u}}\right)\exp\{x_i\beta\}}.$$

Ward *et al.* (2009) defines this function as the *observed likelihood* versus the *full likelihood* that, instead, considers the distribution of the stratum variable Z jointly with the response Y.

The posterior. Now, through the Bayes rule we derive the full posterior

$$p(\beta, \theta | \mathbf{z}, \mathbf{x}) \quad \propto \quad p(\theta) p(\beta | \theta) L(\beta; \mathbf{z}, \mathbf{x}) \tag{24}$$

that can be used to make inference on the quantities of interest.

4 The MCMC computation.

Samples from (24) can be obtained via Markov Chain Monte Carlo simulation (Robert and Casella, 2004; Liu, 2008). While it seems quite standard to implement a direct sampler for the vector β and the hyper parameter θ, we need to sample also the latent $\mathbf{y}_u$. For this reason we introduce a step of data augmentation (Tanner and Wong, 1987; Tanner, 1996) in the estimation procedure. The basic idea of the data augmentation technique is to augment the set of observed data to a set of completed data that follow a simpler distribution (Liu and Wu, 1999). In our framework, we need to augment the observations of the stratum variable $\mathbf{z}$ with the missing values $\mathbf{y}_u$ in order to have, at each iteration, a consistent value of the quantity n_{1u}, necessary to adjust the regression function $\phi_{pod}(x) \approx x\beta + \log \frac{n_{1u} + n_p}{n_{1u}}$ for presence-only data. The following result allows for an easy implementation of the data augmentation step.

Proposition 2. Using the approximation (17) of the ratio (15), the posterior predictive probability of occurrence for an unobserved response y in the sub-sample S_u is approximated by the model $\mathcal{M}$ that generates the data at the population level, that is

$$Pr(Y = 1 | Z = 0, C = 1, x) \approx \pi^*(x). \tag{25}$$

Proof. From the conditional independence between Z and X given Y, the predictive probability of occurrence in S_u is given by

$$Pr(Y = 1 | Z = 0, C = 1, x) = \frac{Pr(Z = 0 | Y = 1, C = 1) Pr(Y = 1 | C = 1, x)}{Pr(Z = 0 | C = 1, x)}.$$

From Table 2 we have that $Pr(Z = 0 | Y = 1, C = 1) = \frac{n_{1u}}{n_p + n_{1u}}$ and hence

$$Pr(Y = 1 | Z = 0, C = 1, x) = \frac{n_{1u}}{n_p + n_{1u}} \frac{Pr(Y = 1 | C = 1, x)}{Pr(Z = 0 | C = 1, x)}. \tag{26}$$

Now, recalling that in the general case one has

$$Pr(Y = 1 | C = 1, x) \approx \frac{\left(1 + \frac{n_p}{n_{1u}}\right) \exp\{\phi(x)\}}{1 + \left(1 + \frac{n_p}{n_{1u}}\right) \exp\{\phi(x)\}} \tag{27}$$

and

$$Pr(Z = 0|C = 1, x) \approx \frac{1 + \exp\{\phi(x)\}}{1 + \left(1 + \frac{n_p}{n_{1u}}\right)\exp\{\phi(x)\}},\qquad(28)$$

by substituting (27) and (28) in (26), one obtains

$$\begin{aligned}
Pr(Y = 1|Z = 0, C = 1, x) &\approx \frac{n_{1u}}{n_p + n_{1u}}\frac{\left(1 + \frac{n_p}{n_{1u}}\right)\exp\{\phi(x)\}}{1 + \exp\{\phi(x)\}}\\
&= \frac{\exp\{\phi(x)\}}{1 + \exp\{\phi(x)\}}\\
&= \pi^*(x).
\end{aligned}$$

$\square$

4.1 The data augmentation algorithm.

A general MCMC scheme to perform inference on a linear regression model for presence-only data can be defined as follow.

Step 0. Initialize θ, β and $\mathbf{y}_u$

Step 1. Set $n_{1u} = \sum_{i \in S_u} y_i$

Step 2. Sample θ from $p(\theta|\mathbf{z}, \mathbf{x}, \beta)$

Step 3. Sample β from $p(\beta|\mathbf{z}, \mathbf{x}, \theta)$

Step 4. Sample y_i from $p(y_i|Z_i = 0, C_i = 1, x_i, \beta)$ for all $i \in S_u$

Goto Step 1

After the initialization of all the arrays (Step 0), Step 1 sets a current value for the quantity n_{1u} to adjust the regression function $\phi_{pod}(x)$. Step 2 and Step 3 consider the sampling from the posterior of the hyper parameter θ and the regression parameter β, respectively, and they can be performed by Metropolis-Hasting schemes (Robert and Casella, 2004). Step 4 concerns the data augmentation for the unobserved $\mathbf{y}_u$ in order to update consistently the quantity n_{1u} at the following iteration. From the result (25), this simulation can be obtained by a Gibbs sampler (Robert and Casella, 2004) since the posterior predictive distribution for all $i \in S_u$ is approximated by *Bernoulli* with parameter of occurrence $\pi(x_i) = \frac{\exp\{x_i\beta\}}{1 + \exp\{x_i\beta\}}$.

4.2 The estimation of the prevalence π.

From the data augmentation algorithm we can obtain a MCMC estimate of the population prevalence π. In fact, if at each iteration t, after the Markov chain has reached the equilibrium, we save the current value $n_{1u}^{(t)}$, we can obtain a consistent MCMC approximation of the sample prevalence π_u in S_u by

$$\hat{\pi}_{mcmc} = \frac{\bar{n}_{1u}}{n_u} \tag{29}$$

where $\bar{n}_{1u}$ is the ergodic mean of the augmentations $n_{1u}^{(t)}$ over the Markov chain, that is

$$\bar{n}_{1u} = \frac{\sum_{t=1}^{T} n_{1u}^{(t)}}{T}.$$

Therefore, since π_u would be a consistent estimator for π, $\hat{\pi}_{mcmc}$ represents also a consistent estimation of the empirical population prevalence.

5 A comparative simulation study.

We present a simulation experiment to evaluate the performances of the model (20). To this aim we generate several datasets in the way described below and we compare our proposal with respect to two models acting in two different situations: (a) the censorship process does not act on the population $\mathcal{U}$ so that the data $\mathbf{y}$ are completely observed; (b) the censorship is present, but we assume known the population prevalence so that approximation (16) can be used. In (a) we are able to estimate a linear logistic model (denoted by M_0), no correction is required and $\phi_0(x) = x\beta$. In (b) we consider a linear logistic model for presence-only data, denoted by M_1, with regression function $\phi_1(x) = x\beta + \frac{\pi n_u + n_p}{\pi n_u}$. Model (20) (denoted by M_2) is estimated when the censorship process acts on the data and no information is available on the population prevalence. In this case, the regression function is given by $\phi_2(x) = x\beta + \frac{n_{1u} + n_p}{n_{1u}}$. Remark that model M_2 can be estimated when the least amount of information is available, M_1 requires less information than M_0 but more than M_2 and M_0 can be used only in the ideal situation of complete information. We assume M_1 as benchmark model in the case of presence-only data.

The generation of data. In order to set the simulation study, we need to generate the covariates X and the binary response Y. In particular, we consider two covariates: X_1, giving strong information on the distribution of the response Y, and X_2, representing a term of noise, not available in the estimation step. We assume X_1 distributed as a mixture of two *Gaussian* densities (denoted by N), centred in $\mu_a = 4.0$ and $\mu_b = -4.0$ respectively, and with equal variances $\sigma^2 = 4.0$, that is

$$X_1 \sim wN_a(\mu_a; \sigma^2) + (1 - w)N_b(\mu_b; \sigma^2).$$

The weight w is a realization of a *Bernoulli* random variable with probability of occurrence fixed to $p = 0.165$. X_2 has standard *Gaussian* distribution $N(0, 1)$. Finally, the binary

response Y, given the covariates X_1 and X_2, is *Bernoulli* distributed with probability of occurrence

$$\pi(x) = \frac{exp\{\beta_0 + \beta_1 x_1 + \beta_2 x_2\}}{1 + exp\{\beta_0 + \beta_1 x_1 + \beta_2 x_2\}}.$$

We generate covariates and binary response with respect to a population $\mathcal{U}$ of size $N = 10000$. Three general scenarios with different level of complexity have been considered:

(i) $\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = 0$: only the informative covariate X_1 generates the data;

(ii) $\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = 1$: a term of noise X_2 is added to the informative covariate;

(iii) $\beta_0 = 1$, $\beta_1 = 1$, $\beta_2 = 1$: X_1, X_2 and a constant effect generate the data.

The case-control sampling. For each scenario, we sample under the case-control design with a ratio of presence/unobserved equal to $1 : 4$ and with respect to eight different sample sizes

$$n = 50, 100, 200, 500, 1000, 1500, 2000, 3000.$$

For example, if the sample size is equal to $n = 500$, we build the corresponding simulated experiment by extracting a random sample S_p from $\mathcal{U}_p$ of $n_p = 100$ presences and a random sample S_u from $\mathcal{U}$ of $n_u = 400$ unobserved values, covariates are available for the whole sample S. We consider $k = 1000$ independent replications of each experiment. In summary, we generate a database of 24,000 datasets (8 sample sizes, 3 scenarios and 1000 replications). With respect to the generating of the data, we considered a quite general framework since the contribution of an informative covariate was combined with a constant effect and a white *Gaussian* noise. With respect to the three scenarios, we obtained empirical population prevalences respectively $\pi_{(i)} = 0.215$, $\pi_{(ii)} = 0.223$ and $\pi_{(iii)} = 0.286$.

The MCMC estimation. The estimation is performed in a Bayesian framework for all the models M_0, M_1 and M_2. The likelihood function we use in the estimation is based on a model that does not always replicate the model used to generate data. More precisely for all experiments (i), (ii) and (iii) the estimation model is:

$$\text{logit}(Pr(Y = 1 | X_1 = x_1)) = \beta_0 + \beta_1 x_1 \tag{30}$$

than with scenario (i) the model that generates the data and the one defining the likelihood are the same, whilst for scenarios (ii) and (iii) the likelihood model becomes increasingly different from the one that generates the data. Notice that we consider a simpler structure than the one shown in Figure (1) as we choose a *Gaussian* prior $N(0, 25)$ for all regression parameters (β_0, β_1) and no hyper parameter is considered. Then, MCMC estimates are computed using 5000 runs after 10000 iterations of burn-in, no thinning is applied as samples autocorrelation is negligible.

Results. In what follows we report Figures and Tables built on scenario (iii) as it represents the most complex of the three alternatives and it is our "worst" case. In each replicate of an experiment, point estimates are computed as posterior means over 5000 iterations. In Figure 2 boxplots describing point estimates behaviour are reported, horizontal lines corresponding to the "true" values are drawn. The first box corresponds to procedure M_0, the second to M_1 and the third to our proposal M_2. In M_0 the prevalence π is estimated as the ratio of the observed presences in S_u to the sample size n_u. In M_1, although π is assumed known a priori, we consider its posterior prediction in S_u. Finally in M_2, the prevalence is obtained at each MCMC step as described in section 4.2 and then the mean over 5000 runs is taken. In Table 3 further details of the point estimates are reported: the median and in parenthesis the first and third quartiles. From the Figures and the values we can see that the three procedure lead to "comparable" values with the obvious reduction of variability when n increases. Remark that the estimates for M_2, although affected by a larger variability with small sample sizes, rapidly approaches M_0 and M_1 behaviour with increasing sample size. This can be seen more clearly in Figure 3 where rooted mean square errors (rmse) are reported. As far as β_1 is concerned the lack of knowledge on X_2 leads to biased point estimates regardless the estimation procedure. Tables 5 and 6 in Appendix report point estimates for scenarios (i) and (ii). For scenarios (i) unbiased estimates are obtained while (ii) is affected by the same distortion as (iii) but with smaller variability.

n	Model	β_0	β_1	π
	M_0	1.42 (0.68 ; 2.33)	1.15 (0.88 ; 1.55)	0.28 (0.25 ; 0.35)
50	M_1	3.13 (1.78 ; 4.46)	1.69 (1.17 ; 2.28)	0.31 (0.26 ; 0.35)
	M_2	1.79 (-3.38 ; 4.26)	1.44 (0.76 ; 2.12)	0.24 (0.13 ; 0.34)
	M_0	1.14 (0.72 ; 1.62)	1.00 (0.86 ; 1.22)	0.29 (0.25 ; 0.33)
100	M_1	2.12 (1.11 ; 3.51)	1.30 (0.97 ; 1.80)	0.30 (0.26 ; 0.34)
	M_2	1.92 (0.16 ; 3.87)	1.24 (0.89 ; 1.78)	0.28 (0.21 ; 0.36)
	M_0	1.01 (0.72 ; 1.36)	0.94 (0.83 : 1.06)	0.29 (0.26 ; 0.31)
200	M_1	1.53 (0.89 ; 2.39)	1.08 (0.87 ; 1.37)	0.29 (0.27 ; 0.32)
	M_2	1.49 (0.59 ; 2.62)	1.07 (0.83 ; 1.37)	0.29 (0.24 ; 0.34)
	M_0	0.94 (0.75 ; 1.15)	0.89 (0.82 ; 0.96)	0.29 (0.27 ; 0.30)
500	M_1	1.12 (0.78 ; 1.57)	0.94 (0.82 ; 1.10)	0.29 (0.28 ; 0.31)
	M_2	1.17 (0.62 ; 1.82)	0.94 (0.80 ; 1.12)	0.29 (0.26 ; 0.32)
	M_0	0.91 (0.78 ; 1.04)	0.88 (0.83 ; 0.92)	0.28 (0.28 ; 0.30)
1000	M_1	1.03 (0.79 ; 1.34)	0.91 (0.83 ; 1.01)	0.29 (0.28 ; 0.30)
	M_2	1.05 (0.68 ; 1.49)	0.91 (0.82 ; 1.03)	0.29 (0.27 ; 0.31)
	M_0	0.89 (0.80 ; 1.00)	0.86 (0.83 ; 0.91)	0.29 (0.28 ; 0.29)
1500	M_1	1.00 (0.78 ; 1.24)	0.89 (0.82 ; 0.98)	0.29 (0.28 ; 0.30)
	M_2	1.01 (0.71 ; 1.35)	0.90 (0.82 ; 0.99)	0.29 (0.27 ; 0.31)
	M_0	0.89 (0.82 ; 0.98)	0.87 (0.84 ; 0.90)	0.29 (0.28 ; 0.29)
2000	M_1	0.96 (0.79 ; 1.15)	0.89 (0.83 ; 0.95)	0.29 (0.28 ; 0.29)
	M_2	0.96 (0.71 ; 1.23)	0.88 (0.82 ; 0.96)	0.29 (0.27 ; 0.30)
	M_0	0.90 (0.83 ; 0.97)	0.87 (0.84 ; 0.89)	0.29 (0.28 ; 0.29)
3000	M_1	0.94 (0.82 ; 1.09)	0.88 (0.84 ; 0.93)	0.29 (0.28 ; 0.29)
	M_2	0.95 (0.76 ; 1.17)	0.88 (0.83 ; 0.94)	0.29 (0.28 ; 0.30)

Table 3: Scenario (iii): point estimates of regression parameters and prevalence computed as medians over 1000 replicates with increasing sample sizes and different models (M_0, M_1 and M_2). In parenthesis distributions quartiles are reported.
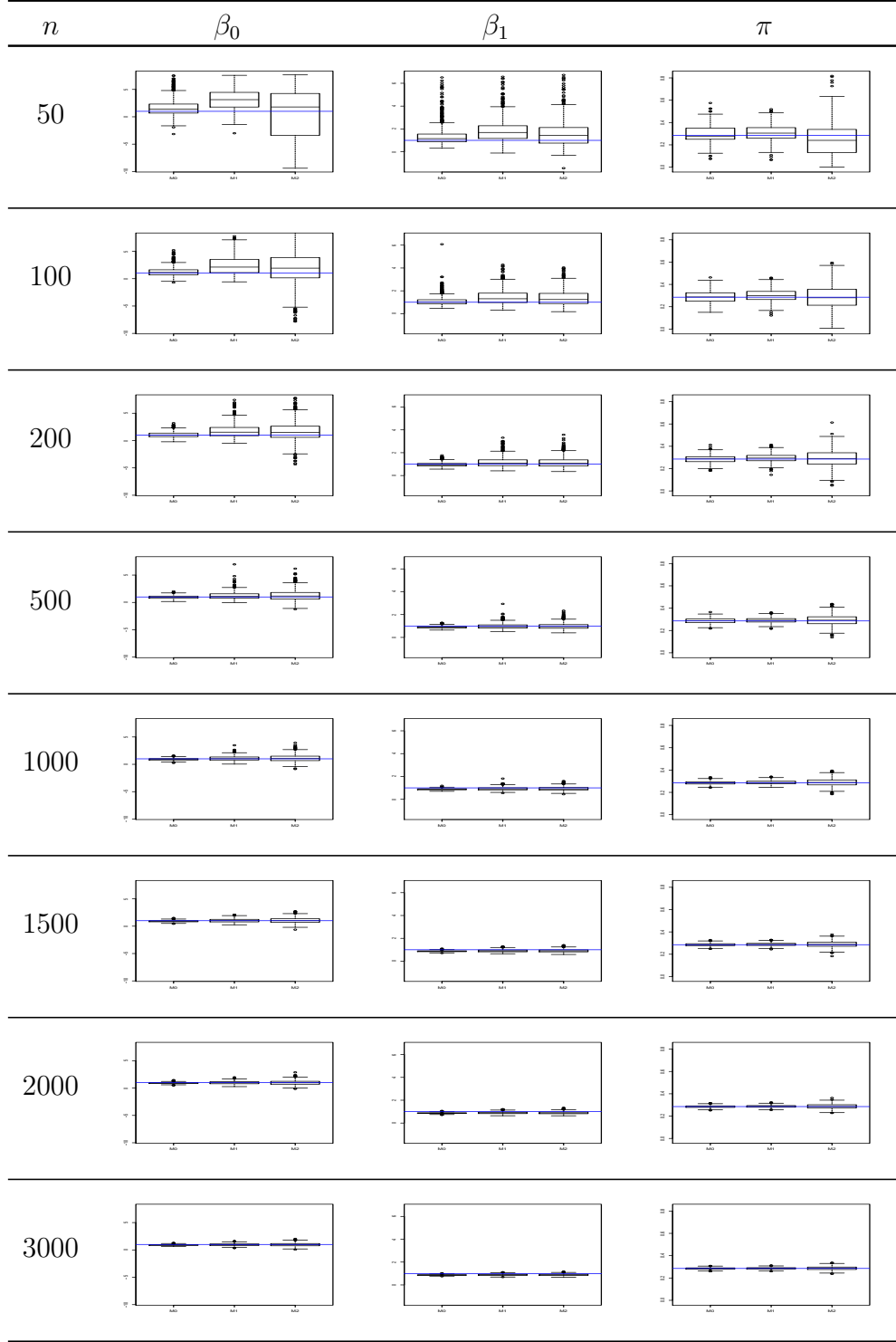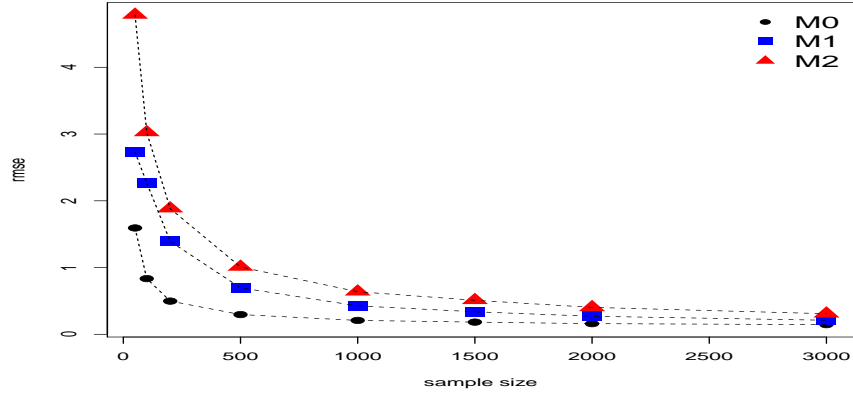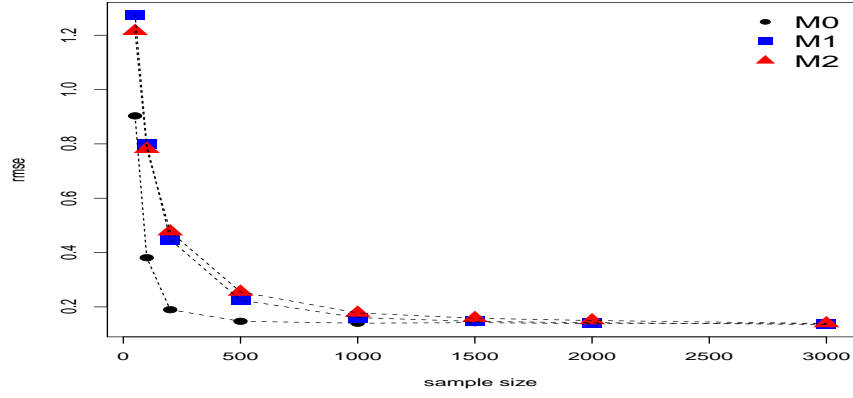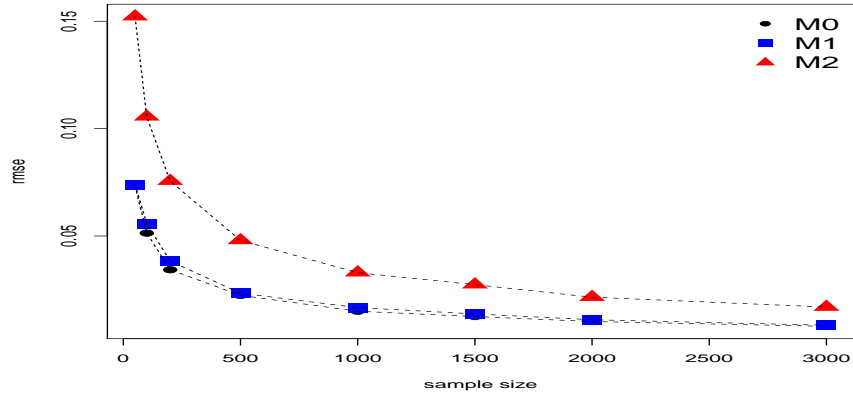
Figure 2: Scenario (iii): boxplots of simulations with increasing sample sizes and different models(M_0,M_1 and M_2).

Figure 3: Scenario (iii): root mean squared errors for different models (M_0, M_1 and M_2) over the 1000 replications, plots with increasing sample sizes for β_0 (a), β_1 (b) and π (c). Dashed trajectories are reported to show the patterns.

From Ward *et al.* (2009) we know that pairwise correlation between parameters is present. In Table 4 we report the empirical pairwise correlation measures, obtained as the averages with respect to the 1000 samples, with increasing sample sizes across the different models. No significant differences in the pattern of correlation $(\beta_0; \beta_1)$ between the models M_1 and M_2 while the correlation (β_1, π) has a general weaker pattern in M_1 than M_2. With respect to the correlation (β_0, π) more significant difference are present between M_1 and M_2. In Figure 4 scatterplots of β_0 versus π in the 1000 replicates are plotted with equal axis across estimation procedures and sample sizes. These pictures help us to understand how this correlation evolves with increasing sample sizes. M_2 produces the most positive correlated point estimates this being an advantage whenever the model is properly specified.

Model	M_0			M_1			M_2		
n	$\beta_0; \beta_1$	$\beta_0; \pi$	$\beta_1; \pi$	$\beta_0; \beta_1$	$\beta_0; \pi$	$\beta_1; \pi$	$\beta_0; \beta_1$	$\beta_0; \pi$	$\beta_1; \pi$
50	0.65	0.26	-0.09	0.59	0.10	-0.30	0.68	0.81	0.31
100	0.75	0.24	-0.12	0.89	0.29	0.02	0.82	0.76	0.37
200	0.78	0.34	-0.04	0.94	0.39	0.18	0.90	0.78	0.48
500	0.79	0.38	0.00	0.95	0.41	0.24	0.92	0.77	0.51
1000	0.77	0.38	-0.01	0.94	0.46	0.27	0.91	0.81	0.54
1500	0.78	0.42	0.00	0.95	0.48	0.28	0.92	0.81	0.55
2000	0.77	0.35	-0.06	0.94	0.49	0.30	0.92	0.81	0.55
3000	0.81	0.37	-0.01	0.95	0.43	0.23	0.91	0.80	0.52

Table 4: Scenario (iii): pairwise parameters correlation (average over the 1000 replicates) with increasing sample sizes and different models (M_0, M_1 and M_2).
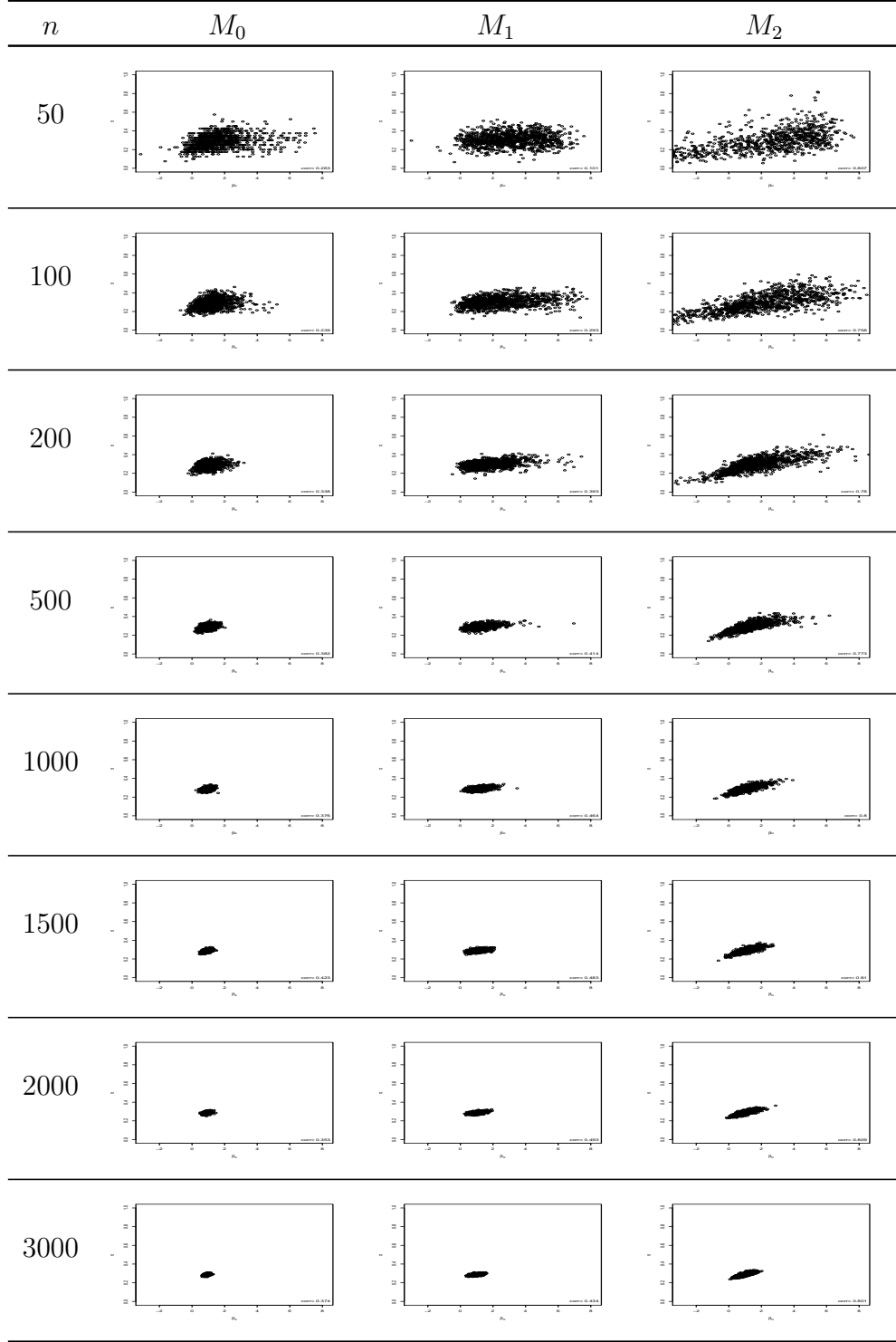
n	M_0	M_1	M_2
50			
100			
200			
500			
1000			
1500			
2000			
3000			

Figure 4: Scenario (iii): scatterplot of π versus β_0 with increasing sample sizes and different models (M_0,M_1 and M_2).

To verify the predictive performance we considered relative measures of specificity and

sensitivity (Fawcett, 2006) build as the ratio of the same measures for M_2 (numerator) and for M_1 (denominator) respectively. In Figure 5 the obtained values are reported versus sample sizes. Remark that M_2 rapidly reaches the same level of performance as M_1 with increasing sample size.
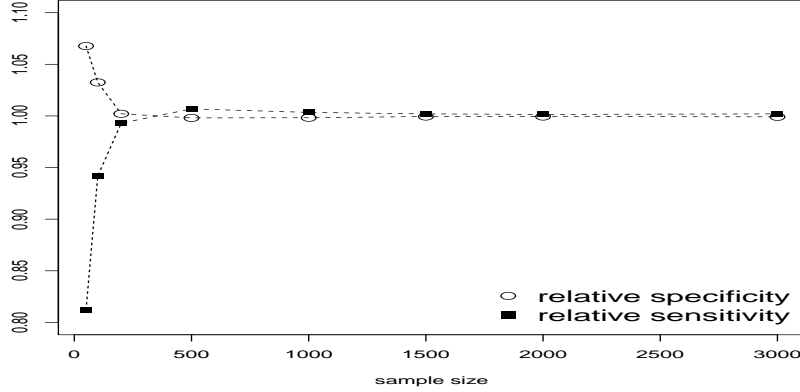

Figure 5: Scenario (iii): relative specificity and sensitivity computed as ratios between M_2 and M_1 specificity and sensitivity measures with increasing sample sizes. Dashed trajectories are reported to show the patterns.

6   Conclusions

In this work, we presented a Bayesian procedure to estimate the parameters of logistic regressions for presence-only data. The approach we proposed is based on a two levels scheme where a generating probability law is combined with a case-control design adjusted for presence-only data. The new formalization allows to consider rigorously all the mathematical details of the model as for instance the approximation of the ratio (15) that represents the crucial point when modeling presence-only data in the finite population setting. We want to point out that our formalization is substantially different from the work by Ward *et al.* (2009), although we end with the same statistical model. We concentrated on the case of the linear logistic regression because we were aware that some care is necessary to handle the identifiability issues present in the model.

The comparative simulation study considered three scenarios with different levels of complexity across increasing sample sizes. We presented detailed results with respect to the most difficult case where the contribution of an informative covariate was mixed with a constant effect and a white *Gaussian* noise. In term of point estimation, the estimates based on our model were comparable to those obtained under the presence-only data benchmark in which the empirical population prevalence was assumed to be known. On the other hand, this lack of information on the population prevalence affected the efficiency of the estimates, that resulted smaller for our model than for the benchmark. This difference was significant only when the sample size n was smaller than 1000, i.e. when the number of observed presences n_p was smaller than 200. From the predictive point

of view, our model performed as well as the benchmark already for sample sizes about $n = 200$, i.e. for a number of observed presences at least $n_p = 40$. Also the pairwise correlation between β_0 and π, that represents an important issue as pointed by Ward *et al.* (2009), became negligible with increasing sample sizes.

From the computational point of view, the procedure were carried out through a MCMC scheme with data augmentation and implemented in Fortran codes.

Future work will investigate the possibility of adding dependence structures among the population units into the model as, for instance, through the use of regression functions with structured random effects.

Appendix

Proposition 3. Under the assumption that, given Y, the inclusion into the sample $(C = 1)$ is independent from the covariates X, it results

$$Pr(Y = 0|C = 1, x) \, Pr(C = 1|x) = \frac{1 - \pi^*(x)}{1 + \pi^*(x)} \rho_0$$

and

$$Pr(Y = 1|C = 1, x) \, Pr(C = 1|x) = \frac{2\pi^*(x)}{1 + \pi^*(x)} \rho_1.$$

Proof. In general we have that

$$Pr(Y|C = 1, x) = \frac{Pr(C = 1|Y, x) \, Pr(Y|x)}{Pr(C = 1|x)} \tag{31}$$

From the conditional independence between $C = 1$ and X given Y, the (31) becomes

$$Pr(Y|C = 1, x) = \frac{Pr(C = 1|Y) \, Pr(Y|x)}{Pr(C = 1|x)}.$$

Recalling that $Pr(Y = 1|x) = \frac{2\pi^*(x)}{1+\pi^*(x)}$ and the definitions of $\rho_0 = Pr(C = 1|Y = 0)$ and $\rho_1 = Pr(C = 1|Y = 1)$ the proofs for $Y = 0$ and $Y = 1$ can be derive by simple algebra.

n	Model	β_0	β_1	π
50	M_0	0.40 (-0.31 ; 1.45)	1.56 (1.08 ; 2.72)	0.20 (0.18 ; 0.25)
	M_1	2.19 (0.68 ; 3.57)	2.19 (1.37 ; 3.74)	0.23 (0.18 ; 0.27)
	M_2	1.03 (-2.51 ; 3.35)	2.00 (1.07 ; 3.44)	0.19 (0.12 ; 0.26)
100	M_0	0.31 (-0.18 ; 0.88)	1.23 (0.99 ; 1.61)	0.21 (0.19 ; 0.25)
	M_1	1.24 (0.29 ; 2.36)	1.55 (1.12 ; 2.22)	0.23 (0.19 ; 0.26)
	M_2	1.22 (-0.40 ; 2.69)	1.50 (1.07 ; 2.22)	0.16 (0.16 ; 0.27)
200	M_0	0.11 (-0.20 ; 0.46)	1.08 (0.95 ; 1.28)	0.22 (0.19 ; 0.24)
	M_1	0.46 (0.00 ; 1.27)	1.23 (0.99 ; 1.56)	0.22 (0.20 ; 0.24)
	M_2	0.48 (-0.24 ; 1.55)	1.23 (0.96 ; 1.59)	0.22 (0.19 ; 0.25)
500	M_0	0.06 (-0.10 ; 0.25)	1.02 (0.94 ; 1.12)	0.22 (0.20 ; 0.23)
	M_1	0.17 (-0.09 ; 0.47)	1.04 (0.92 ; 1.19)	0.22 (0.20 ; 0.23)
	M_2	0.14 (-0.26 ; 0.61)	1.03 (0.89 ; 1.20)	0.22 (0.19 ; 0.24)
1000	M_0	0.04 (-0.05 ; 0.17)	1.01 (0.95 ; 1.07)	0.22 (0.21 ; 0.22)
	M_1	0.04 (-0.13 ; 0.26)	0.99 (0.91 ; 1.08)	0.21 (0.21 ; 0.22)
	M_2	0.03 (-0.22 ; 0.34)	0.98 (0.90 ; 1.09)	0.21 (0.20 ; 0.23)
1500	M_0	0.05 (-0.04 ; 0.15)	0.99 (0.95 ; 1.04)	0.21 (0.21 ; 0.22)
	M_1	0.01 (-0.12 ; 0.18)	0.97 (0.91 ; 1.05)	0.21 (0.21 ; 0.22)
	M_2	0.00 (-0.24 ; 0.23)	0.97 (0.90 ; 1.05)	0.21 (0.20 ; 0.22)
2000	M_0	0.03 (-0.04 ; 0.10)	0.99 (0.95 ; 1.03)	0.21 (0.21 ; 0.22)
	M_1	0.00 (-0.12 ; 0.14)	0.97 (0.92 ; 1.03)	0.21 (0.21 ; 0.22)
	M_2	-0.02 (-0.22 ; 0.14)	0.96 (0.90 ; 1.02)	0.21 (0.20 ; 0.22)
3000	M_0	0.03 (-0.02 ; 0.10)	0.98 (0.96 ; 1.02)	0.21 (0.21 ; 0.22)
	M_1	0.00 (-0.10 ; 0.09)	0.96 (0.92 ; 1.00)	0.21 (0.21 ; 0.22)
	M_2	-0.03 (-0.18 ; 0.11)	0.95 (0.91 ; 1.00)	0.21 (0.20 ; 0.22)

Table 5: Scenario (i): point estimates of regression parameters and prevalence computed as medians over 1000 replicates with increasing sample sizes and different models (M_0, M_1 and M_2). In parenthesis distributions quartiles are reported.

n	Model	β_0	β_1	π
50	M_0	0.42 (-0.33 ; 1.42)	1.34 (0.94 ; 2.13)	0.23 (0.18 ; 0.28)
	M_1	2.12 (0.78 ; 3.39)	1.95 (1.25 ; 2.96)	0.24 (0.20 ; 0.28)
	M_2	1.26 (-2.99 ; 3.33)	1.75 (0.95 ; 2.79)	0.20 (0.13 ; 0.28)
100	M_0	0.19 (-0.20 ; 0.73)	1.07 (0.88 ; 1.35)	0.23 (0.19 ; 0.25)
	M_1	1.13 (0.36 ; 2.40)	1.39 (1.00 ; 1.96)	0.23 (0.21 ; 0.26)
	M_2	1.03 (-0.40 ; 2.65)	1.34 (0.96 ; 1.96)	0.22 (0.17 ; 0.28)
200	M_0	0.13 (-0.18 ; 0.45)	0.97 (0.84 ; 1.12)	0.23 (0.20 ; 0.25)
	M_1	0.48 (0.02 ; 1.17)	1.08 (0.88 ; 1.36)	0.23 (0.21 ; 0.25)
	M_2	0.48 (-0.38 ; 1.56)	1.07 (0.83 ; 1.41)	0.23 (0.18 ; 0.27)
500	M_0	0.09 (-0.07 ; 0.27)	0.92 (0.85 ; 1.00)	0.22 (0.21 ; 0.24)
	M_1	0.22 (-0.04 ; 0.53)	0.95 (0.84 ; 1.09)	0.23 (0.21 ; 0.24)
	M_2	0.23 (-0.24 ; 0.69)	0.95 (0.81 ; 1.09)	0.22 (0.20 ; 0.25)
1000	M_0	0.08 (-0.02 ; 0.20)	0.90 (0.86 ; 0.95)	0.22 (0.21 ; 0.23)
	M_1	0.09 (-0.07 ; 0.31)	0.90 (0.83 ; 0.99)	0.22 (0.21 ; 0.23)
	M_2	0.08 (-0.19 ; 0.38)	0.89 (0.82 ; 1.00)	0.22 (0.21 ; 0.24)
1500	M_0	0.08 (0.00 ; 0.18)	0.90 (0.86 ; 0.94)	0.22 (0.22 ; 0.23)
	M_1	0.08 (-0.05 ; 0.23)	0.89 (0.84 ; 0.96)	0.22 (0.22 ; 0.23)
	M_2	0.05 (-0.17 ; 0.30)	0.89 (0.82 ; 0.96)	0.22 (0.21 ; 0.23)
2000	M_0	0.07 (0.00 ; 0.15)	0.89 (0.86 ; 0.92)	0.22 (0.22 ; 0.23)
	M_1	0.06 (-0.06 ; 0.20)	0.89 (0.84 ; 0.95)	0.22 (0.22 ; 0.23)
	M_2	0.02 (-0.17 ; 0.24)	0.88 (0.82 ; 0.94)	0.22 (0.21 ; 0.23)
3000	M_0	0.07 (0.01 ; 0.13)	0.89 (0.87 ; 0.91)	0.22 (0.22 ; 0.23)
	M_1	0.04 (-0.04 ; 0.15)	0.88 (0.84 ; 0.92)	0.22 (0.22 ; 0.23)
	M_2	0.02 (-0.13 ; 0.17)	0.87 (0.83 ; 0.92)	0.22 (0.21 ; 0.23)

Table 6: Scenario (ii): point estimates of regression parameters and prevalence computed as medians over 1000 replicates with increasing sample sizes and different models (M_0,M_1 and M_2). In parenthesis distributions quartiles are reported.

References

Araùjo, M. and Williams, P. (2000). Selecting areas for species persistence using occurrence data. *Biological Conservation*, **96**, 331–345.

Armenian, H. (2009). *The Case-Control Method: Design And Applications*. Oxford University Press, New York, USA.

Breslow, N. E. (2005). *Handbook of Epidemiology*, chapter 6: Case-Control Studies, pages 287–319. Springer, New York, USA.

Breslow, N. E. and Dey, N. E. (1980). *Statistical Methods In Cancer Research, Volume 1 - The analysis of case-control studies*. WHO International Agency for Research on Cancer, Lyon, France.

Chakraborty, A., Gelfand, A. E., Wilson, A. M., Latimer, A. M., and Silander, J. A. (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.

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 Agricultural, Biological and Environmental Statistics*, **16**(3), 339–356.

Divino, F., Golini, N., Jona Lasinio, G., and Pettinen, A. (2011). Data augmentation approach in bayesian modelling of presence-only data. *Procedia Environmental Sciences*, **7**, 38–43.

Dorazio, R. M. (2012). Predicting the geographic distribution of a species from presence-only data subject to detection errors. *Biometrics*, **68**, 1303–1312.

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

Elith, J., 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., and Zimmermann, N. E. (2006). Novel methods improve prediction of species' distribution from occurence data. *Ecography*, **29**, 129–151.

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., and Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, **17**, 43–57.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letter*, **27**, 861–874.

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

Jaynes, E. T. (1957). Information theory and statistical mechanics. *The Physical Review*, **106**(4), 620–630.

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

Lancaster, T. and Imbens, G. (1996). Case-control studies with contaminated controls. *Journal of Econometrics*, **71**, 145–160.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis With Missing Data*. John Wiley & Sons, New York, USA.

Liu, J. S. (2008). *Monte Carlo Strategies In Scientific Computing*. Springer, New York, USA.

Liu, S. Y. and Wu, Y. N. (1999). Parameter expansion for data augmentation. *Journal of American Statistical Association*, **94**, 1264–1274.

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

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

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Metods*. Springer, New York, USA.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **63**(3), 581–592.

Särndal, C. E. (1978). Design-based and model-based inference in survey sampling. *Scandinavian Journal of Statistics*, **5**, 27–52.

Tanner, M. (1996). *Tools for Statistical Inference: Observed Data And Data Augmentation*. Springer, New York, USA.

Tanner, M. and Wong, W. (1987). The calculation of posterior distribution by data augmentation. *Journal of American Statistical Association*, **82**, 528–550.

Ward, G., Hastie, T., Barry, S., Elith, J., and Leathwick, A. (2009). Presence-only data and the EM algorithm. *Biometrics*, **65**, 554–563.

Warton, D. I. and 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.

Woodward, M. (2005). *Epidemiology: Study Design And Data Analysis*. Chapman & Hall, New York, USA.

Zaniewski, A. E., Lehmann, A., and Overton, J. M. (2002). Prediction species spatial distributions using presence-only data: a case study of native New Zeland ferns. *Ecological Modelling*, **157**, 261–280.