Stochastic Environmental Research and Risk Assessment

Bayesian Logistic Regression for Presence-only Data

--Manuscript Draft--

Manuscript Number:	
Full Title:	Bayesian Logistic Regression for Presence-only Data
Article Type:	Original research
Keywords:	case-control design, censored data, data augmentation, Markov Chain Monte Carlo, presence-only data, stratified sampling, two-level scheme
Corresponding Author:	Giovanna Jona Lasinio, Ph.D. Sapienza University of Rome Rome, Rome ITALY
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Sapienza University of Rome
Corresponding Author's Secondary Institution:	
First Author:	Fabio Divino, Ph.D.
First Author Secondary Information:	
Order of Authors:	Fabio Divino, Ph.D.
	Giovanna Jona Lasinio, Ph.D.
	Natalia Golini, Ph.D.
	Antti Penttinen, Ph.D.
Order of Authors Secondary Information:	
Abstract:	Presence-only data are referred to situations in which a censoring mechanism acts on a binary response that can be partially observed only with respect to one outcome, usually called presence. In this work a Bayesian approach to the analysis of presence-only data based on a two-level scheme is presented. A probability law and a case-control design are combined to handle the double source of uncertainty: one due to censoring and the other one due to sampling. In the paper, through the use of a stratified sampling design with non-overlapping strata, a new formalization of the logistic model for presence-only data is proposed. In particular, the logistic regression with linear predictor is considered and a Markov Chain Monte Carlo algorithm with data augmentation, that does not require the a priori knowledge of the population prevalence, is presented. The performance of the new algorithm is validated by means of extensive simulation experiments using three scenarios and comparison with optimal benchmarks. An application to literature data is reported to discuss the model behavior in real world situations.
Suggested Reviewers:	Daniela Cocchi, PhD Full professor, University of Bologna daniela.cocchi@unibo.it She is a well known environmental statistician and she might be interested in the paper's topic
	Alessio Pollice, PhD Associate professor, University of Bari alessio.pollice@uniba.it He is a well known environmental and ecological statistician and he may be interested in revieweing the paper
	Robert Dorazio, PhD Researcher, U.S. Geological Survey's Southeast Ecological Science Center. bob_dorazio@usgs.gov He has published several papers on the topic and it maybe interested in reading this

	work.

cover letter

Dear Sirs,
we ask our paper “Bayesian Logistic Regression for Presence-only Data” to be considered for publication in Stochastic Environmental Research and Risk Assessment.

We suggest that Dr Emilio Porcu can act as Associate Editor to handle this paper.

Best Regards
Fabio Divino, Giovanna Jona Lasinio, Natalia Golini, Antti Penttinen

Noname manuscript No.
(will be inserted by the editor)

Bayesian Logistic Regression for Presence-only Data

Fabio Divino · Natalia Golini · Giovanna Jona Lasinio · and Antti Penttinen

Received: date / Accepted: date

Abstract Presence-only data are referred to situations in which a censoring mechanism acts on a binary response that can be partially observed only with respect to one outcome, usually called *presence*. In this work a Bayesian approach to the analysis of presence-only data based on a two-level scheme is presented. A probability law and a case-control design are combined to handle the double source of uncertainty: one due to censoring and the other one due to sampling. In the paper, through the use of a stratified sampling design with non-overlapping strata, a new formalization of the logistic model for presence-only data is proposed. In particular, the logistic regression with linear predictor is considered and a Markov Chain Monte Carlo algorithm with data augmentation, that does not require the a priori knowledge of the population prevalence, is presented. The performance of the new algorithm is validated by means of extensive simulation experiments using three scenarios and comparison with optimal benchmarks. An application to literature data is reported to discuss the model behavior in real world situations.

Keywords case-control design, censored data, data augmentation, Markov Chain Monte Carlo, presence-only data, stratified sampling, two-level scheme

Fabio Divino
Division of Physics, Computer Science and Mathematics, University of Molise,
Contrada Fonte Lappone, 86090 - Pesche (IS), Italy.
E-mail: fabio.divino@unimol.it

Natalia Golini
Department of Statistical Sciences, University of Rome *"La Sapienza"*.

Giovanna Jona Lasinio
Department of Statistical Sciences, University of Rome *"La Sapienza"*.

Antti Penttinen
Department of Mathematics and Statistics, University of Jyväskylä.

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 design 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 situation a sample of cases and an additional 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 situation 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) the presence-only data are modelled using Poisson point processes in the likelihood and Bayesian frames 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 Dorazio (2012) the asymptotic relations between the two approaches are discussed.
A different approach, MaxEnt (Phillips *et al.*, 2006; Elith *et al.*, 2011), is based on the maximum entropy principle (Jaynes, 1957). In MaxEnt the relative entropy between the distribution of covariates at locations where presences are observed (species presence in the related papers) and the unconditional background distribution of covariates is maximized subject to some constrains (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". Recently, further attention has been given to the connection existing between the point process and MaxEnt approaches (Fithian and Hastie, 2013; Renner and Warton, 2013). In particular in Fithian and Hastie (2013) the finite sample equivalence between point processes and MaxEnt is analyzed and the different behavior of standard logistic regression in the same setting highlighted.
n 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). In particular, we have a population of conditionally independent units given covariates and we imagine that if some type of dependence is detected, information on it is brought in by covariates. In this work, the presence of spatial dependence in the data is discussed in the application to real data, enhancing the issues arising when only censored information on the response variable is available.

The main contribution of the paper is a new rigorous formalization of the logistic regression for presence-only data based on a stratified sampling design with non-overlapping strata 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 simplified mechanism generating the data. We present an extensive simulation study comparing our approach to two models representing optimal benchmarks and an application to real data.

The paper is organized as follows. Section 2 introduces our general framework for the presence-only data problem, Section 3 presents the Bayesian approach, Section 4 describes the MCMC algorithm while results related to the simulation study are reported in Section 5. In Section 6, an application to real data is presented to show the relevance and flexibility of our proposal. Finally, in Section 7 some conclusions are drawn and future developments briefly described.

2 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 logistic modeling of presence-only data and then consider the case of the logistic regression with linear predictor. The proposed approach is built on two levels and we partially follow the formulation introduced by Ward *et al.* (2009) but adopting a stratified sampling design and a Bayesian scheme as in Divino *et al.* (2011).

2.1 A two-level scheme

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 informative, on the same attribute, covariates which are available on the same population and with domain $\mathcal{X}$. 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 the response Y is not completely observable, we need to adjust for the sampling mechanism through the use of a case-control design (Breslow and Day, 1980; Schlesselman, 1982; Breslow, 2005; Armenian, 2009) .

In order to build a statistical 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 conditional independently generated by a probability law $\mathbb{P}$.

Level 2 Let $\mathcal{U}_p$ be the subset of $\mathcal{U}$ where $Y = 1$. 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; Valliant *et al.*, 2000) 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 probability law $\mathbb{P}$. Hence, the problem of presence-only data can be formalized as a problem of missing data (Rubin, 1976; Little and Rubin, 1987). In particular, we consider a missing completely at random framework (Little and Rubin, 1987).

2.2 The model generating population data

At the first level, we assume that the law $\mathbb{P}$ is defined in terms of the conditional probability of occurrence $\mathbb{P}(Y = 1|x)$, denoted by $\pi^*(x)$, when the covariates are $X = x$, and that the binary responses are conditionally independent given $X = x$. Moreover, we consider that the relation between Y and X is formalized through a regression function on the logistic scale, $\operatorname{logit} \pi^*(x) = \phi(x)$, that is $\pi^*(x) = \frac{\exp\{\phi(x)\}}{1+\exp\{\phi(x)\}}$. When the data $\mathbf{y} = (y_1, ..., y_N)$ are conditional independently generated from the law $\mathbb{P}$, we denote by π the empirical prevalence of the binary response Y in the population $\mathcal{U}$, 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 adjusted for presence-only data (Lancaster and Imbens, 1996) in order to account for the specific structure of the data. In general, the use of the case-control design is always necessary 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 exposure in epidemiological studies (Woodward, 2005).

The case-control design can be seen as a stratified sampling design (Schlesselman, 1982; Valliant *et al.*, 2000; Levy and Lemershow, 2008) with non-overlapping strata, the strata being the controls ($Y = 0$) and the cases ($Y = 1$). Observations are generated from each stratum using a random sampling design, i.e n_0 controls and n_1 cases are obtained. In our framework, we consider simple random sampling mechanisms without replacement and with discrete uniform probability of selection. Then an element of a

population of size N has probability $\frac{n}{N}$ to be selected in a sample of size n.
Let us introduce some notation. In our framework, we consider two levels of probabilistic formalization: the first one is represented by the probability law $\mathbb{P}$ generating the data, while the second one concerns the probality distribution of the response Y at the population level. For each value $x \in \mathcal{X}$, let us indicate by $\mathcal{U}(x)$ the subset of $\mathcal{U}$ where units have covariate X equal to x, that is $\mathcal{U}(x) = \{i \in \mathcal{U} | X(i) = x\}$, where $X(i)$ is the covariate X at unit i. Now, we denote by $P(Y = 1|x)$ the conditional probabilty of drawing a presence from $\mathcal{U}(x)$. Under the assumption that each unit has the same chance to be drawn one has

$$P(Y = 1|x) = \frac{N_1(x)}{N(x)} \tag{1}$$

where $N(x)$ is the size of the subset $\mathcal{U}(x)$ and $N_1(x) = \sum_{i \in \mathcal{U}(x)} y_i$. If the response Y is censored or only partially observed in $\mathcal{U}$, the probability in (1) cannot be determined because the quantity $N_1(x)$ is unknown. But under the assumption that, given $X(i) = x$, Y_i are conditionally independent binary random variables with identical conditional probability of occurrence $\mathbb{P}(Y = 1|x) = \pi^*(x)$, we can derive by simple algebra the expectation of the ratio in (1)

$$\mathbb{E}\left[\frac{N_1(x)}{N(x)}\right] = \mathbb{P}(Y = 1|x).$$

Therefore, in order to derive a computable model for presence-only data, throughout the paper we use the following approximation and its consequences

$$P(Y = 1|x) \approx \mathbb{P}(Y = 1|x) = \pi^*(x).$$

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

$$P(Y = 1|C = 1, x) = \frac{\rho_1 \exp\{\phi(x)\}}{\rho_0 + \rho_1 \exp\{\phi(x)\}}. \tag{2}$$

Hence, the corresponding case-control regression function, denoted by $\phi_{cc}(x)$ and defined as the logit of (2), is given by

$$\phi_{cc}(x) = \phi(x) + \log \frac{\rho_1}{\rho_0}. \tag{3}$$

In particular, if the selection of cases and controls 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 (3) becomes

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

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 population $\mathcal{U}$ and by a set S_p of n_p independent observations selected from the sub-population of presences $\mathcal{U}_p$.

To adopt the stratified sampling design formalization we need to clarify the role of the sub-population of presences $\mathcal{U}_p \subset \mathcal{U}$. We need disjoint strata to ensure the coherence of the probabilistic structure of the model. Hence we introduce a design population $\mathcal{U}_D$ defined as the reference population $\mathcal{U}$ augmented with the sub-population of presences $\mathcal{U}_p$, that is $\mathcal{U}_D = (\mathcal{U}; \mathcal{U}_p)$, and then of size $N_D = N + N_1$ or equivalently $N_D = N_0 + 2N_1$.

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 stratum of the entire population $\mathcal{U}$ while $Z = 1$ denotes the sampling from the stratum of the sub-population $\mathcal{U}_p$. Remark that $Z = 1$ implies $Y = 1$ while $Z = 0$ implies that Y is an unknown value $y \in \{0, 1\}$ because in this stratum we cannot observe the response Y. Moreover, by construction the stratum Z is conditionally independent of the covariates X, given the response Y. The introduction of the stratum 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 (response Y) and known/unknown data (stratum Z), as reported in Table 1 and Table 2. In Table 1, N_0 represents the number of absences in the population $\mathcal{U}$ while in Table 2, n_{0u} and n_{1u} respectively denote the

Table 1: Data structure at the design population level by stratum (Z) and response (Y).

	stratum		
response	$Z = 0$	$Z = 1$	Total
$Y = 0$	N_0	0	N_0
$Y = 1$	N_1	N_1	$2N_1$
Total	N	N_1	N_D

Table 2: Data structure at the sample level by stratum (Z) and response (Y).

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

unknown frequencies of absences and presences in the sub-sample S_u. From Table 1, it is easy to derive the probabilistic weight of each stratum, $P(Z=0)=\frac{N}{N+N_1}$ and $P(Z=1)=\frac{N_1}{N+N_1}$ and the probability distribution of the response Y in the design population $\mathcal{U}_D$, respectively $P(Y=0)=\frac{N_0}{N+N1}$ and $P(Y=1)=\frac{2N_1}{N+N1}$. Remark that, in the above described situation, the inclusion probabilities in the sample change. In fact, while an absence can be drawn only when sampling from the stratum $\mathcal{U}$, a presence can be selected when sampling from both strata $\mathcal{U}$ and $\mathcal{U}_p$. Hence, one has the following property(see Appendix for the detailed proof).

Property 1 Under a stratified random sampling design adjusted for presence-only data, with non-overlapping strata $\mathcal{U}$ and $\mathcal{U}_p$, the inclusion probabilities in the sample are given by

$$\rho_0=\frac{n_{0u}}{(1-\pi)N} \tag{4}$$

for the stratum of cases and by

$$\rho_1=\frac{n_{1u}+n_p}{2\pi N}. \tag{5}$$

for the stratum of controls.

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

Proposition 1 *Let us consider the population $\mathcal{U}$ augmented with its subset $\mathcal{U}_p$. Then, under the assumption that the stratum variable Z is conditionally independent of X given Y, one has that the conditional probability of presence in the design population $\mathcal{U}_D$ is given by*

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

From the above result, we obtain the following Corollary (see Appendix for the detailed proof).

Corollary 1 *Under the assumption that, given Y, the inclusion into the sample ($C=1$) is conditionally independent of the covariates X, one has*

$$P(Y=0|C=1,x)\,P(C=1|x)=\frac{1-\pi^*(x)}{1+\pi^*(x)}\,\rho_0 \tag{6}$$

and

$$P(Y=1|C=1,x)\,P(C=1|x)=\frac{2\pi^*(x)}{1+\pi^*(x)}\,\rho_1. \tag{7}$$

Then, from the ratio of (7) to (6), we can obtain that

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

and by plugging the quantities ρ_0 and ρ_1, as defined in (4) and in (5), into the logit of $P(Y=1|C=1,x)$, one obtains the following relation (see Property 2 in the Appendix for details)

$$\begin{aligned}\operatorname{logit} P(Y=1|C=1,x) = \operatorname{logit}\pi^*(x)+\log\frac{n_{1u}+n_p}{n_{0u}}\\ -\log\frac{\pi}{1-\pi}\end{aligned}$$

that represents the logistic regression model under the case-control design adjusted for presence-only data. Now, recalling that $\text{logit}\pi^*(x)=\phi(x)$, we can formalize the presence-only data regression function, denoted by $\phi_{pod}(x)$, as

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

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

$$\frac{\rho_1}{\rho_0}=\frac{n_{1u}+n_p}{n_{0u}}\,\frac{1-\pi}{2\pi}, \tag{8}$$

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) replaces 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

$$\begin{aligned}\frac{\rho_1}{\rho_0}\approx\frac{E[n_{1u}+n_p]}{E[n_{0u}]}\,\frac{1-\pi}{2\pi}&=\frac{\pi n_u+n_p}{(1-\pi)n_u}\,\frac{1-\pi}{2\pi}\\ &=\frac{\pi n_u+n_p}{2\pi n_u}.\end{aligned} \tag{9}$$

The authors adopt a maximum 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 logistic regression with linear predictor, and with respect to this case they present a simulation example. The difficulty in obtaining efficient joint estimates due to the correlation between π and the intercept of the regression term is discussed as well. Notice that Ward *et al.* (2009) considers a slightly different representation of the ratio (9), omitting the multiplier "2" in the denominator. This point shows the difference between our approach and the model by Ward *et al.* (2009). In our framework, we assume and formalize in details a stratified sampling design with non-overlapping strata as sampling scheme. In fact that design induces that the presences can be sampled from two disjoint strata, $\mathcal{U}$ and $\mathcal{U}_p$, hence from the results in Property 1 and Proposition 1 follows our representation of the ratio (8) .
Di Lorenzo *et al.* (2011), dealing with a problem of abundance data, use the approximation (9), 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

and in the MCMC algorithm they include an approximating step 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 (8) can be obtained by considering the sample prevalence $\pi_u = \frac{n_{1u}}{n_u}$ in S_u (the background sample).

Due to the censorship process, this quantity is unknown but it would be the maximum likelihood estimator for π if the data $\mathbf{y}_u = \{y_i, i \in S_u\}$ were observed. Now, replacing π by π_u in (8) one obtains

$$\begin{aligned}\frac{\rho_1}{\rho_0} \approx \frac{n_{1u}+n_p}{n_{0u}}\frac{1-\pi_u}{2\pi_u} &= \frac{n_{1u}+n_p}{n_u-n_{1u}}\frac{n_u-n_{1u}}{2n_{1u}} \\ &= \frac{n_{1u}+n_p}{2n_{1u}}, \end{aligned} \tag{10}$$

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

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

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 they can be included into the estimation process and then obtain an 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 (Tanner and Wong, 1987; Tanner, 1996) . 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 (10) 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 estimate $\phi(x)$ and π jointly when the latter is identifiable with respect to the regression function, as for example in the case of logistic regression with linear predictor (see Ward *et al.*, 2009, for mathematical details).

2.4 The logistic regression with linear predictor

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

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

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

$$\begin{aligned}P(Y=1|C=1,x,\beta) &\approx \frac{\exp\{x\beta + \log\frac{n_{1u}+n_p}{n_{1u}}\}}{1+\exp\{x\beta + \log\frac{n_{1u}+n_p}{n_{1u}}\}} \\ &= \frac{\left(1+\frac{n_p}{n_{1u}}\right)\exp\{x\beta\}}{1+\left(1+\frac{n_p}{n_{1u}}\right)\exp\{x\beta\}}. \end{aligned} \tag{11}$$

In this particular case, all the unknowns of the model are the vector of regression coefficients 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

$$\begin{aligned} P(Z,Y|C=1,x) &= P(Z|y,C=1,x) \\ &\quad \times P(Y|C=1,x), \end{aligned} \tag{12}$$

through the marginalization over Y, the probability $P(Z|C=1,x)$ can be obtained and we can express the relation between presences and covariates in terms of regression of Z with respect to X. Notice that, while $P(Y|C=1,x)$ is given by (11), the term $P(Z|y,C=1,x)$, due to the conditional independence between Z and X given Y, simplifies to $P(Z|y,C=1)$ 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, we need to keep it in the model as a hidden variable in order to obtain an approximation for the quantity $n_{1u} = \sum_{i \in S_u} y_i$, necessary to adjust the logistic regression model for presence-only data.

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

The likelihood In Figure 1, at the lowest level of the hierarchy, we have the likelihood, defined with respect to the observable stratum variable Z. Recalling that from Table 2 we have $P(Z=1|Y=0,C=1)=0$ and $P(Z=1|Y=1,C=1)=\dfrac{n_p}{n_{1u}+n_p}$, when (12) is marginalized over Y, one obtains the approximation

$$\begin{aligned} P(Z=1 \mid C=1,x,\beta) &\approx \\ &\approx \frac{n_p}{n_{1u}+n_p}\,\frac{\exp\{x\beta+\log\frac{n_{1u}+n_p}{n_{1u}}\}}{1+\exp\{x\beta+\log\frac{n_{1u}+n_p}{n_{1u}}\}} \\ &= \frac{\frac{n_p}{n_{1u}}\exp\{x\beta\}}{1+\left(1+\frac{n_p}{n_{1u}}\right)\exp\{x\beta\}} \end{aligned} \tag{13}$$

and hence

$$P(Z=0|C=1,x,\beta) \approx \frac{1+\exp\{x\beta\}}{1+\left(1+\frac{n_p}{n_{1u}}\right)\exp\{x\beta\}}.$$

Thus, we can assume that for all $i \in S$ the conditional distribution of Z_i is *Bernoulli*, denoted by $\mathcal{B}(\cdot)$, with probability of occurence given by (13), that is

$$Z_i|C_i=1,x_i,\beta \sim \mathcal{B}\left(\frac{\frac{n_p}{n_{1u}}\exp\{x_i\beta\}}{1+\left(1+\frac{n_p}{n_{1u}}\right)\exp\{x_i\beta\}}\right).$$

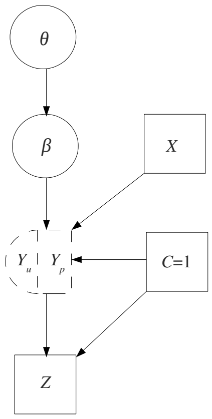


Fig. 1: Graphical representation of the hierarchical Bayesian model.

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 approximated as

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

where $\mathbf{z} = \{z_i, i \in S\}$ and $\mathbf{x} = \{x_i, i \in S\}$. 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 hyperpriors, priors and missing data At the top of the hierarchy, we assume the hyperparameter θ distributed as $p(\theta)$. At the second level, we consider the prior probability distribution on β depending on the hyperparameter θ, that is $\beta|\theta \sim p(\beta|\theta)$. At the third level, the unobserved data $\mathbf{y}_u = \{y_i, i \in S_u\}$ are considered latent parameters with prior distribution *Bernoulli* with probability of occurrence given by the approximation in (11), that is for $i \in S_u$

$$y_i | C_i = 1, x_i, \beta \sim \mathcal{B} \left(\frac{\left(1 + \frac{n_p}{n_{1u}}\right) \exp\{x_i \beta\}}{1 + \left(1 + \frac{n_p}{n_{1u}}\right) \exp\{x_i \beta\}} \right).$$

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

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

$$p(\beta, \theta|\mathbf{z}, \mathbf{x}) \propto p(\theta)p(\beta|\theta)L(\beta; \mathbf{z}, \mathbf{x}) \tag{14}$$

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

4 The MCMC computation

Samples from (14) 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 hyperparameter 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 complete data that follow a simpler distribution (Liu and Wu, 1999). In our framework, we need to augment the observations $\mathbf{z}$ of the stratum variable 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 $\phi_{pod}(x) \approx x\beta + \log \frac{n_{1u}+n_p}{n_{1u}}$. The following result allows for an easy implementation of the data augmentation step (see the Appendix for the detailed proof).

Proposition 2 *Using the approximation* (10) *of the ratio* (8)*, the posterior predictive probability of occurrence for an unobserved response $Y = y$ in the sub-sample S_u is approximated by the probability law $\mathbb{P}$ that generates the data at the population level, that is*

$$P(Y = 1|Z = 0, C = 1, x) \approx \pi^*(x). \tag{15}$$

4.1 The data augmentation algorithm

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

Algorithm 1 MCMC with Data Augmentation for presence-only data.

```
Step 0. Initialize $\theta$, $\beta$ and $\mathbf{y}_u$
Step 1. Set $n_{1u} = \sum_{i \in S_u} y_i$
Step 2. Sample $\theta$ from $p(\theta|\mathbf{z}, \mathbf{x}, \beta)$
Step 3. Sample $\beta$ 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 $i \in S_u$
Repeat from 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 hyperparameter 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 (15), 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 a *Bernoulli* random variable with parameter of occurrence $\pi(x_i) = \frac{\exp\{x_i\beta\}}{1+\exp\{x_i\beta\}}$.

4.2 The estimation of prevalence π

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

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

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 an estimator for π, $\hat{\pi}_{mcmc}$ represents also an estimate of the empirical population prevalence.

5 A comparative simulation study

We present a simulation experiment to evaluate the performances of the model (11). To this aim we generate several datasets in the way described below and we compare our proposal 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 (9) can be used. In (a) we are able to estimate a logistic model with linear predictor (denoted by M_0), no correction is required and $\phi_0(x) = x\beta$. In (b) we consider a logistic model with linear predictor for presence-only data, denoted by M_1, with regression function $\phi_1(x) = x\beta + \log\frac{\pi n_u + n_p}{\pi n_u}$. Model (11) (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 $\phi_2(x) = x\beta + \log\frac{n_{1u}+n_p}{n_{1u}}$. 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.

Generation of data In order to perform 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 in the generation of data, not available in the estimation step. We assume

X_1 distributed as a mixture of two *Gaussian* densities, denoted by $\mathcal{N}(.,.)$, centred in $\mu_a = 4.0$ and $\mu_b = -4.0$ respectively, and with equal variances $\sigma^2 = 4.0$, that is

$$X_1 \sim w\mathcal{N}_a(\mu_a, \sigma^2) + (1-w)\mathcal{N}_b(\mu_b, \sigma^2).$$

The weight w is fixed to 0.165 that implies presences are rare events but not "too rare". X_2 is assumed to follow standard *Gaussian* distribution $\mathcal{N}(0,1)$. 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 are 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 the stratum $\mathcal{U}_p$ of $n_p = 100$ presences and a random sample S_u from the stratum $\mathcal{U}$ of $n_u = 400$ unobserved values, covariates are available for the whole sample S. We consider $h = 1000$ independent replications of each experiment. In summary, we generate a database of 24,000 datasets (8 sample sizes, 3 scenarios and $h = 1000$ replications). The data generation framework is quite general since the contribution of an informative covariate is combined with a constant effect and a white *Gaussian* noise. With respect to the three scenarios, we obtain empirical population prevalences respectively $\pi_{(i)} = 0.215$, $\pi_{(ii)} = 0.223$ and $\pi_{(iii)} = 0.286$.

Posterior computation through MCMC 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:

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

than with scenario (i) the model that generates the data and the one defining the likelihood are the same, while 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 $\mathcal{N}(0,25)$ for all regression parameters (β_0 and β_1) and no hyperparameter is considered. Then, MCMC estimates are computed using 5000 runs after 10000 iterations of burn-in, no thinning is applied as the sample 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 scatter plot of point estimates together with their 95% credibility interval versus sample sizes are reported, horizontal lines correspond to the "true". Dots corresponds to M_0, squares to M_1 and triangles 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 on 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 procedures lead to "comparable" results 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 squared 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.

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 $h = 1000$ replications, 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 is found 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 differences are present between M_1 and M_2. For M_2 this correlation remains stable with changing sample size, than M_2 produces the largest positive correlation between point estimates.

To verify the predictive performance we consider 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 4 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.

6 Real data example: the North Carolina wren data

In the same spirit as the simulation study we compare the performance of our proposal (M_2) on real data to logistic model with linear predictor (M_0) and the already introduced model M_1. The data we consider in this example are well known in the literature (see for example Royle *et al.*, 2012; Merow and Silander Jr., 2014), they are taken from the North American Breeding Bird Survey (BBS). In particular we focus on the Carolina wren (*Thryothorus ludovicianus*) using four land cover variables (per cent cover of mixed forest (*pcMix*), deciduous forest (*pcDec*), coniferous forest (*pcCon*) and grasslands (*pcGr*)) and latitude (*Lat*) and longitude (*Lon*), these data are available in the `maxlike` R's package. Unlike Royle *et al.* (2012) and Merow and Silander Jr. (2014) we consider only simple covariates effects, we fit the simplest logistic regression to avoid overfitting and possible multicollinearity. The data include full information

Table 3: Scenario (iii): summary of the marginal posterior distribution for the regression parameters and the prevalence expressed as median and upper and lower quartiles over $h = 1000$ replications with increasing sample sizes (n) and different models (M_0, M_1 and M_2).

n	Model	β_0	β_1	π
50	M_0	1.42 (0.68 ; 2.33)	1.15 (0.88 ; 1.55)	0.28 (0.25 ; 0.35)
	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)
100	M_0	1.14 (0.72 ; 1.62)	1.00 (0.86 ; 1.22)	0.29 (0.25 ; 0.33)
	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)
200	M_0	1.01 (0.72 ; 1.36)	0.94 (0.83 : 1.06)	0.29 (0.26 ; 0.31)
	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)
500	M_0	0.94 (0.75 ; 1.15)	0.89 (0.82 ; 0.96)	0.29 (0.27 ; 0.30)
	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)
1000	M_0	0.91 (0.78 ; 1.04)	0.88 (0.83 ; 0.92)	0.28 (0.28 ; 0.30)
	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)
1500	M_0	0.89 (0.80 ; 1.00)	0.86 (0.83 ; 0.91)	0.29 (0.28 ; 0.29)
	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)
2000	M_0	0.89 (0.82 ; 0.98)	0.87 (0.84 ; 0.90)	0.29 (0.28 ; 0.29)
	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)
3000	M_0	0.90 (0.83 ; 0.97)	0.87 (0.84 ; 0.89)	0.29 (0.28 ; 0.29)
	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 4: Scenario (iii): pairwise correlation (average over the $h = 1000$ replications) with increasing sample sizes (n) and different models (M_0, M_1 and M_2).

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

on both presence and absence over 4607 locations of which 1506 record the presence of a wren. We fit M_0 over samples of size 800 taken from the entire dataset, while to fit M_1 and M_2 we build a sample of 200 presences and background samples of size 800, the procedure is repeated 100 times. In Figure 5 we report the estimated models parameters and their 95% credibility intervals (see Table 7 in the appendix for details).

We compute, again following Fawcett (2006), the true positive rate that shows how M_2 is the most accurate model in detecting presences (true positive rate or sensitivity: $M_0 = 0.84$, $M_1 = 0.91$, $M_2 = 0.93$), while in terms of specificity the three models are very similar ($M_0 = 0.91$, $M_1 = 0.89$, $M_2 = 0.87$). The population prevalence is 0.33 and the M_2 estimated prevalence is 0.39 with 95% credibility interval (0.32,0.44) that includes the population value, while the M_1 predicted prevalence is 0.37 with 95% credibility interval (0.33, 0.40) where the population value is at the very edge of the interval. In terms of point estimates M_2 performs as well as the other two models, paying the price of a larger estimates variability due to the smaller amount of information it uses. The geographical variables play a strong role in the fitted models (see Figure 5). The data show a clear and strong spatial dependence as the presences are clustered on the eastern part of the United States. A spatial joint-count autocorrelation test (Cliff and Ord, 1981, page 20) with a neighborhood structure with 4 neighbors returned a p-value smaller than 2.216×10^{-16} confirming the strong spatial dependence present in the data. However all the three models assume spatial independence and devolve to covariates to account for such dependence. Being the covariates spatially informative estimates are perfectly coherent.

Notice that the spatial dependence is usually included in linear and non linear models according to two main approaches: (i) introducing an autoregressive component as predictor (Besag, 1974), **(ii) adding a latent, often Gaussian, component spatially structured (Gelfand, 2010; Banerjee *et al.*, 2004). Both approaches seem almost unfeasible with presence-only data. The first one it is definitely not feasible as we have incomplete information on the response variable and than any autoregressive component cannot be computed. The second one seems conceptually acceptable however, given the severe identifiability issues the logistic regression shows, it would be really difficult to learn anything on any latent component in the model. An attempt to build a model including a random effect spatially structured can be found in Aarts *et al.* (2008). The authors, propose a logistic, mixed-effects approach that uses generalized additive transformations of the environmental covariates and is fitted to a response data-set comprising telemetry data (presence-only data) integrated with simulated observations, under a case-control design. Estimation is carried out in the likelihood framework. This work, although very interesting, do not proposes any solution for the estimation of the case-control correction and prevalence and it does not solve in general terms the issue of how to simulate absences. Our proposal seems then a more feasible solution including spatial information using informative covariates, such as the locations coordinates.**

7 Conclusions

In this work, we presented a Bayesian procedure to estimate the parameters of logistic regressions for presence-only data. The approach is based on a two-level scheme where a generating probability law is combined with a case-control design adjusted for presence-only data. The new formalization using the stratified sample design with non-overlapping strata, allows to consider rigorously all the mathematical details of the model as for instance the approximation of the ratio (8) that represents the crucial point when modeling presence-only data in a finite population setting. From a computational point of view we propose an effective estimation procedure that does not require a priori knowledge of the population prevalence. The procedure is implemented as an efficient MCMC algorithm that compensates for the reduced knowledge on the response variable trough a data augmentation step. We concentrated on the case of the logistic regression with linear predictor 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 terms 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 point estimates, that resulted smaller for our model than for the benchmark. This difference was significant only when the sample size was smaller than $n = 1000$, i.e. when the number of observed presences was smaller than $n_p = 200$. 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 $n_p = 40$. The pairwise correlation between β_0 and π, that represents an important issue as pointed by Ward *et al.* (2009), remains stable with increasing sample sizes.

In the real data example we showed that our model is perfectly capable of capturing the "true" population prevalence and returning prediction that are as good as those of the benchmark models.

We want to stress again that with our proposal a perfectly reasonable estimate of the population prevalence is obtained in all scenarios and in the real data example. It is somehow obvious that if the covariates available on the entire population do not provide "enough" information on the population prevalence no model would be able to return "plausible" estimates. Informative covariates allows to account for latent dependence structures in the data, as we showed in the real data example.

Further developments may include the exploration and adaptation to presence-only data of tools to "remove" spatial dependence from the data when information is complete (presence/absence data). For example as proposed by Carl and Kuhn (2008)), the authors suggest to adopt binary wavelets to this end. Or in Liao and Wei (2014) where spatially expanded coefficients are adopted in logistic regression.

Appendix

Property 1 Under a stratified random sampling design adjusted for presence-only data, with non-overlapping strata $\mathcal{U}$ and $\mathcal{U}_p$, the inclusion probabilities in the sample are given by

$$\rho_0 = \frac{n_{0u}}{(1-\pi)N}$$

for the stratum of cases and by

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

for the stratum of controls.

Proof By definition one has

$$\begin{aligned} P(C = 1 \mid y) = & \\ &= P(C = 1 \mid y, Z = 0)P(Z = 0 \mid y) + \\ &+ P(C = 1 \mid y, Z = 1)P(Z = 1 \mid y). \end{aligned}$$

The term $P(Z = z \mid y) = \frac{P(Z,Y)}{P(Y)}$ can be computed easily from Table 1. In particular, one has

$$\begin{aligned} P(Z = 0 \mid Y = 0) &= \frac{N_0}{N_0} = 1, \\ P(Z = 1 \mid Y = 0) &= \frac{0}{N_0} = 0, \\ P(Z = 0 \mid Y = 1) &= \frac{N_1}{2N_1} = \frac{1}{2}, \\ P(Z = 1 \mid Y = 1) &= \frac{N_1}{2N_1} = \frac{1}{2}. \end{aligned}$$

Now it is easy to derive

$$\begin{aligned} \rho_0 &= P(C = 1 \mid Y = 0) \\ &= \frac{n_{0u}}{N_0} \times 1 + 0 \\ &= \frac{n_{0u}}{(1 - \pi)N} \end{aligned}$$

and

$$\begin{aligned} \rho_1 &= P(C = 1 \mid Y = 1) \\ &= \frac{n_{1u}}{N_1} \times \frac{1}{2} + \frac{n_p}{N_1} \times \frac{1}{2} \\ &= \frac{n_{1u} + n_p}{2\pi N} \end{aligned}$$

□

Proposition 1 *Let us consider the population $\mathcal{U}$ augmented with its subset $\mathcal{U}_p$. Then, under the assumption that the stratum variable Z is conditionally independent of X given Y, one has that the conditional probability of presence in the design population $\mathcal{U}_D$ is given by*

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

Proof From the hypothesis of conditional independence it results

$$P(Z|y, x) = P(Z|y),$$

which can be express also as

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

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

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

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

$$\frac{\pi^*(x)P(Z=0|x)}{P(Y=1|x)} = \frac{\frac{N_1}{N}\frac{N}{N+N_1}}{\frac{2N_1}{N+N_1}} = \frac{1}{2}. \tag{16}$$

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

$$\frac{P(Z=1|x)}{P(Y=1|x)} = \frac{\frac{N_1}{N_1}\frac{N_1}{N+N_1}}{\frac{2N_1}{N+N_1}} = \frac{1}{2}. \tag{17}$$

From (17) it is obtained $P(Y=1|x) = 2\,P(Z=1|x)$ and by substituting into (16), one can derive that $P(Z=0|x) = \dfrac{1}{1+\pi^*(x)}$ and hence $P(Z=1|x) = \dfrac{\pi^*(x)}{1+\pi^*(x)}$. Now, it is easy to obtain that

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

□

Corollary 1 *Under the assumption that, given Y, the inclusion into the sample ($C=1$) is conditionally independent of the covariates X, one has*

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

and

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

Proof In general we have that

$$P(Y|C=1,x) = \frac{P(C=1|y,x)\,P(Y|x)}{P(C=1|x)} \tag{18}$$

From the conditional independence between $C=1$ and X given Y, the (18) becomes

$$P(Y|C=1,x) = \frac{P(C=1|y)\,P(Y|x)}{P(C=1|x)},$$

hence

$$P(Y|C=1,x)P(C=1|x) = P(C=1|y)\,P(Y|x).$$

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

Property 2 Under the case-control design adjusted for presence-only data the logistic regression function $\phi_{pod}(x)$ is given by

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

Proof By the definition of logistic regression function for presence-only data and by simple algebra one has

$$\begin{aligned}
\phi_{pod}(x) &= \text{logit}P(Y=1|C=1,x) \\
&= \log\frac{P(Y=1|C=1,x)}{P(Y=0|C=1,x)} \\
&= \log\frac{P(Y=1|C=1,x)P(C=1 \mid x)}{P(Y=0|C=1,x)P(C=1 \mid x)} \\
&= \log\frac{\frac{2\pi^*(x)}{1+\pi^*(x)}\rho_1}{\frac{1-\pi^*(x)}{1+\pi^*(x)}\rho_0} \\
&= \log\frac{2\pi^*(x)\frac{n_{1u}+n_p}{2\pi N}}{[1-\pi^*(x)]\frac{n_{0u}}{(1-\pi)N}} \\
&= \log\frac{\pi^*(x)\frac{n_{1u}+n_p}{\pi}}{[1-\pi^*(x)]\frac{n_{0u}}{1-\pi}} \\
&= \log\frac{\pi^*(x)}{1-\pi^*(x)} + \log\frac{n_{1u}+n_p}{n_{0u}} - \log\frac{\pi}{1-\pi}
\end{aligned}$$

□

Proposition 2 *Using the approximation* (10) *of the ratio* (8)*, the posterior predictive probability of occurrence for an unobserved response $Y = y$ in the sub-sample S_u is approximated by the probability law $\mathbb{P}$ that generates the data at the population level, that is*

$$P(Y=1|Z=0,C=1,x) \approx \pi^*(x).$$

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

$$\begin{aligned}
&P(Y=1 \mid Z=0,C=1,x) = \\
&\qquad = \frac{P(Z=0|Y=1,C=1)P(Y=1|C=1,x)}{P(Z=0|C=1,x)}.
\end{aligned}$$

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

$$\begin{aligned}
&P(Y=1 \mid Z=0,C=1,x) = \\
&\qquad = \frac{n_{1u}}{n_p+n_{1u}}\frac{P(Y=1|C=1,x)}{P(Z=0|C=1,x)}.
\end{aligned} \tag{19}$$

Now, recalling that in the general case one has

$$P(Y=1|C=1,x) \approx \frac{\left(1+\frac{n_p}{n_{1u}}\right)\,\exp\{\phi(x)\}}{1+\left(1+\frac{n_p}{n_{1u}}\right)\exp\{\phi(x)\}} \tag{20}$$

and

$$P(Z=0|C=1,x) \approx \frac{1+\exp\{\phi(x)\}}{1+\left(1+\frac{n_p}{n_{1u}}\right)\exp\{\phi(x)\}}, \tag{21}$$

by substituting (20) and (21) in (19), one obtains

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

□

References

Aarts, G., MacKenzie, M., McConnell, B., Fedak, M., and Matthiopoulos, J. (2008). Estimating space-use and habitat preference from wildlife telemetry data. *Ecography*, **31**, 140–160.

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.

Banerjee, S., Carlin, B., and Gelfand, A. E. (2004). *Hierarchical Modeling And Analysis For Spatial Data*. Chapman & Hall Ltd, New York, USA.

Besag, J. (1974). Spatial interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society. Series B*, **36**(2), 192–236.

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

Breslow, N. E. and Day, 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.

Carl, G. and Kuhn, I. (2008). Analyzing spatial ecological data using linear regression and wavelet analysis. *Stochastic Environmental Research and Risk Assessment*, **22**(3), 315–324.

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.

Cliff, A. D. and Ord, J. K. (1981). *Spatial processes*. Pion, London, UK.

Table 5: Scenario (i): summary of the marginal posterior distribution for the regression parameters and the prevalence expressed as median and upper and lower quartiles over $h = 1000$ replications with increasing sample sizes (n) and different models (M_0, M_1 and M_2).

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)

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*,

Table 6: Scenario (ii): summary of the marginal posterior distribution for the regression parameters and the prevalence expressed as median and upper and lower quartiles over $h = 1000$ replications with increasing sample sizes (n) and different models (M_0, M_1 and M_2).

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)

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.

Fithian, W. and Hastie, T. (2013). Finite-sample Equivalence in Statistical Models for Presence-Only Data. *Annals of Applied Statistics*, **7**(4), 1917–1939.

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

Gelfand, A. E. (2010). *Handbook of Spatial Statistics*, chapter Misaligned spatial data: The change of support problem, pages 517–539. Chapman & Hall, New York, USA.

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.

Table 7: North Carolina Wren data, results of the marginal posterior distribution for the regression parameters and the prevalence expressed as posterior median and 95% credibility intervals computed over $h = 100$ replications with sample sizes ($n = 1000$) and different models (M_0,M_1 and M_2).

	Model		
Parameters	M_0	M_1	M_2
β_0	-2.12 (-2.50 ; -1.68)	-2.44 (-3.29 ; -1.84)	-2.18 (-3.30 ; -1.29)
β_{pcMix}	-0.30 (-0.45 ; -0.18)	-0.40 (-0.93 ; -0.02)	-0.48 (-1.01 ; 0.02)
β_{pcDec}	0.41 (0.23 ; 0.62)	0.57 (0.02 ; 1.37)	0.77 (0.17 ; 1.65)
β_{pcCon}	0.06 (-0.36 ; 0.31)	-0.52 (-1.57 ; 0.29)	-0.69 (-1.72 ; 0.25)
β_{pcGr}	-0.06 (-0.36 ; 0.19)	-0.04 (-1.10 ; 0.57)	-0.05 (-1.21 ; 0.65)
β_{Lat}	-3.43 (-4.28 ; -2.90)	-6.31 (-8.75 ; -4.72)	-7.40 (-10.86; -5.33)
β_{Lon}	4.24 (3.42 ; 5.33)	6.51 (5.25 ; 8.61)	7.71 (5.55 ; 9.61)
π	0.33 (0.30 ; 0.36)	0.37 (0.33 ; 0.40)	0.39 (0.32 ; 0.44)

Levy, P. S. and Lemershow, S. (2008). *Sampling Of Population: Methods And Applications*. John Wiley & Sons, New York, USA.

Liao, F. and Wei, Y. (2014). Modeling determinants of urban growth in dongguan, china: a spatial logistic approach. *Stochastic Environmental Research and Risk Assessment*, **28**(4), 801–816.

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.

Merow, C. and Silander Jr., J. A. (2014). A comparison of maxlike and maxent for modelling species distributions. *Methods in Ecology and Evolution*, pages DOI: 10.1111/2041–210X.12152.

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.

Renner, I. W. and Warton, D. I. (2013). Equivalence of MAXENT and Poisson Point Process Models for Species Distribution Modeling in Ecology. *Biometrics*, **69**, 274–281.

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

Royle, J. A., Chandler, R. B., Yackulic, C., and Nichols, J. D. (2012). Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. *Methods in Ecology and Evolution*, **3**, 545–554.

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.

Schlesselman, J. J. (1982). *Case-Control Studies*. Oxford University Press, New York, USA.

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.

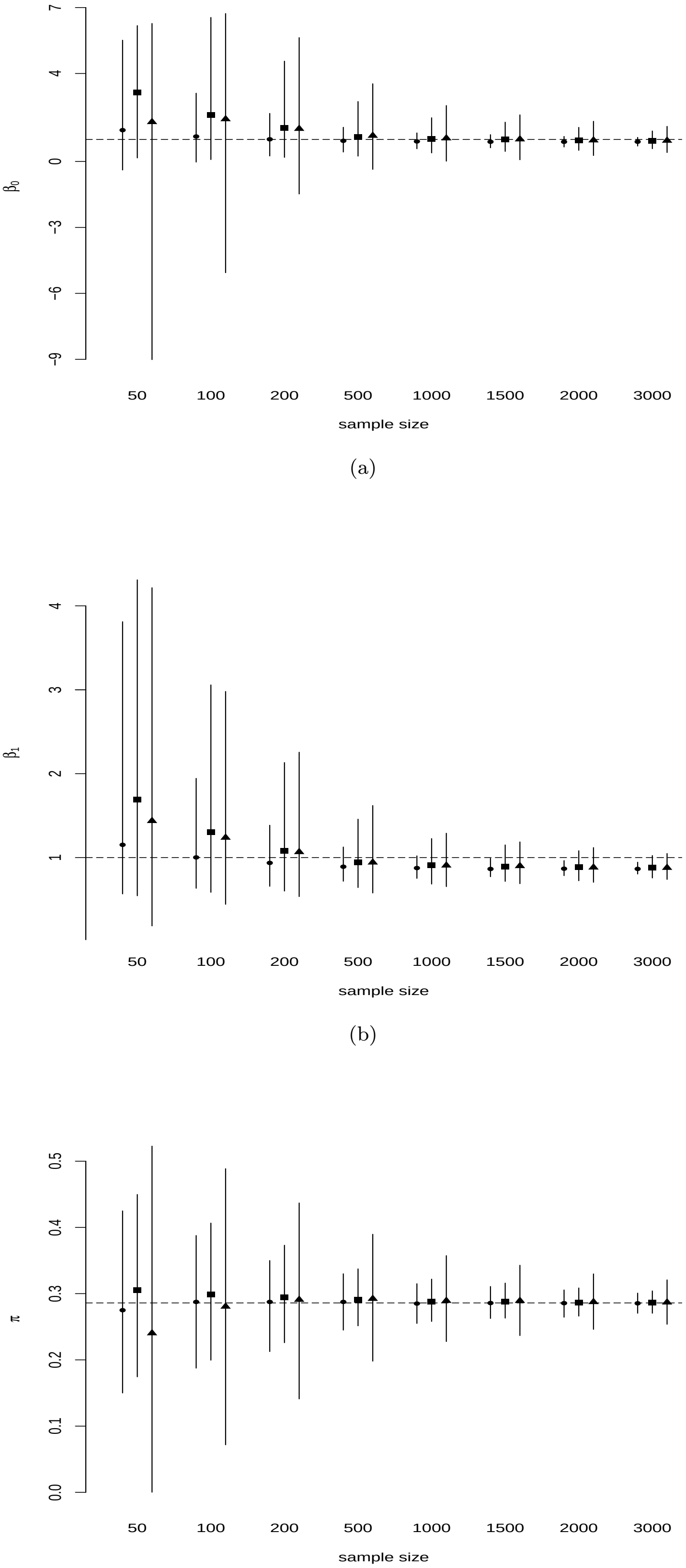
Valliant, R., Dorfman, A. H., and Royall, M. R. (2000). *Finite Population Sampling And Inference: A Prediction Approach*. John Wiley & Sons, New York, USA.

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.

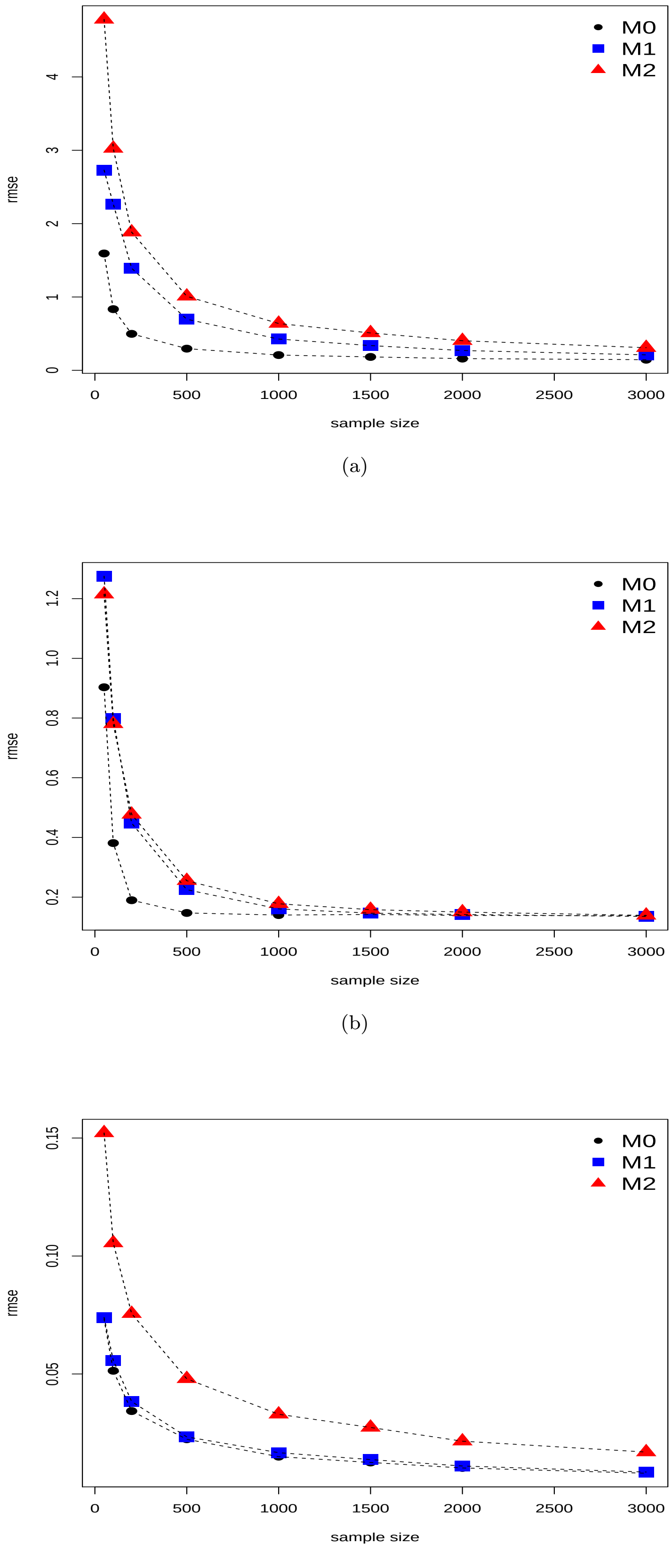


(a)

(b)

(c)

Fig. 2: Parameters estimates for β_0 (a), β_1 (b) and π (c), and their 95% credibility intervals under scenario (iii) for M_0 (dots), M_1 (squares) and M_2 (triangles). The dashed line indicates the "true" parameter's value.



(a)

(b)

(c)

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

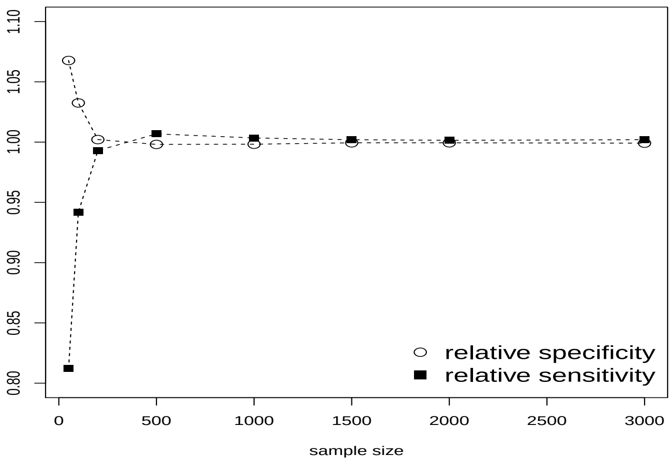


Fig. 4: Scenario (iii): relative specificity and sensitivity computed as ratios between M_2 and M_1 specificity and sensitivity measures with increasing sample sizes. Dashed trajectories are reported to show the patterns.

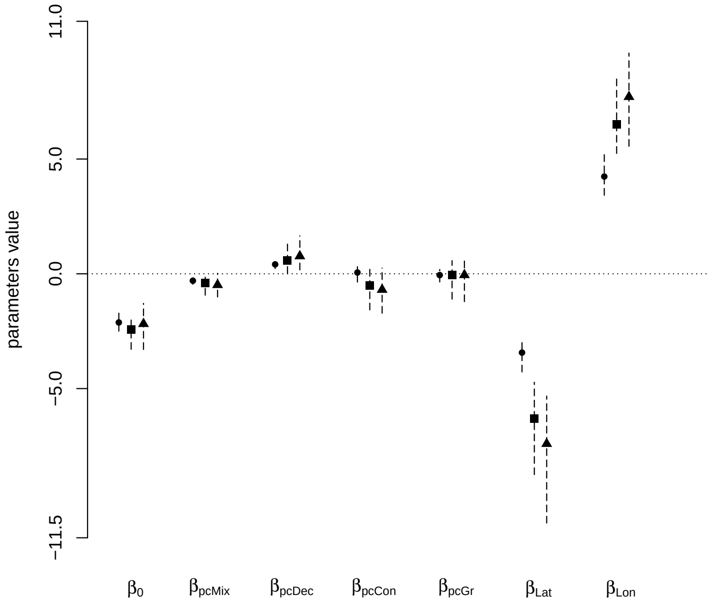


Fig. 5: North Carolina Wren data, parameters estimates and their 95% credibility intervals for for M_0 (dots), M_1 (squares) and M_2 (triangles).