Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Environmental Sciences 7 (2011) 38–43

Procedia
Environmental Sciences

Spatial Statistics 2011

Data Augmentation Approach in Bayesian Modelling of Presence-only Data

F. Divino[a], N. Golini[b*], G. Jona Lasinio[b], A. Penttinen[c]

[a]University of Molise, Contrada Fonte Lappone, Pesche (IS) 86090, Italy
[b]University of Rome, Sapienza, p.le Piazzale Aldo Moro 5, Roma 00185, Italy.
[c]University of Jyväskylä,, P.O.Box 35 FIN-40014, Finland.

Abstract

Ecologists are interested in prediction of potential distribution of species in suitable areas, essential for planning conservation and management strategies. Unfortunately, often the only available information in such studies is the true presence of the species at few locations of the study area and the associated environmental covariates over the entire area, referred as presence-only data. We propose a Bayesian approach to estimate logistic linear regressions adapted to presence-only data through the introduction of a random approximation of the correction factor in the adjusted logistic model that allows us to overcome the need to know a priori the prevalence of the species.

Keywords: Bayesian model; Data augmentation; MCMC algorithm; Presence-only data; Pseudo-absence approach; Potential distribution.

1. Introduction

An important issue in ecological studies is the estimation of the potential spatial extent of an ecological niche. The prediction of geographical distribution of species in suitable areas is essential for planning

* Natalia Golini. Tel.: +39-06-49910499; fax: +39-06-4959341.
E-mail address: natalia.golini@uniroma1.it

doi:10.1016/j.proenv.2011.07.008

conservation and management strategies, and it may concern studies on animal and plant species. Given a presence-absence process for a species, the logistic model represents a natural approach to estimate the species' distribution given the environmental covariates. In ecological studies, however, the collecting of presence-absence data could be expensive and/or difficult. Often the observable information in such studies is not complete and we can collect the true presence of the species only at a few locations of the study area while the associated environmental covariates are available over the entire area. In order to handle that situation an interesting approach is based on the combination of two samples: the first one is composed by sites with true presences while the second one is a sample from the whole population area (referred as background or pseudo-absence sample [2]). Such data are known in literature as presence-only data, see [1] for details.

Following the above approach, it is possible to adapt and fit standard case-control models also in the setting of presence-only data. In this paper we propose a Bayesian approach to estimate logistic linear regressions adapted to presence-only data of rare species through the introduction of a random approximation of the correction factor in the adjusted logistic model that allows us to overcome the need to know the population prevalence of the species a priori. Under the assumption that the pseudo-absence locations are randomly sampled from the entire study area, we can estimate regression parameters jointly with prevalence through a data augmentation MCMC algorithm [3].

The paper is organized as follows. In Section 2 we introduce the model and then we describe our algorithm. In Section 3 we conduct a simulation study in order to evaluate the performance of the proposed algorithm.

2. Model and Algorithm

Let $\boldsymbol{Y}$ be the true presence/absence process, where Y assumes value 1 if the species is present and value 0 if the species is absent. When presence-only data are considered we do not observe the $\boldsymbol{Y}$ process. We are able to asses information on a naive approximation $\boldsymbol{Z}$ of $\boldsymbol{Y}$. A first relation between $\boldsymbol{Y}$ and $\boldsymbol{Z}$ can be formalized as follow: Z=0 implies $Y\in\{0,1\}$ and Z=1 implies Y=1. The introduction of the naive variable $\boldsymbol{Z}$ allows us to consider the full sample S as composed by two subset: S_u of size n_u that includes pseudo-absences (with Z=0) and S_p of size n_p that includes observed presences (with Z=1). For each observation i belonging to S_u, we introduce a Bernoulli random variable $\tilde{Y}_i$ representing the missing-data process. Then, we can further formalize the relation between $\boldsymbol{Y}$ and $\boldsymbol{Z}$ as follow: $Z_i = 0$ implies the i-th observation belongs to S_u and the missing value Y_i can be represented by the Bernoulli random variable $\tilde{Y}_i$; $Z_i = 1$ implies the i-th observation belong to S_p and the observed Y_i is equal to 1.

In Table 1, we represent the relation between $\boldsymbol{Y}$ and $\boldsymbol{Z}$**:**

Table 1. Sampling design for presence-only data

		Z		
		0	1	*Total*
Y	0	n_{0u}	0	n_0
	1	n_{1u}	n_{1p}	n_1
	Total	n_u	n_p	n

where n_{0u} is the unknown number of unobserved absences in the subsample S_u, n_{1u} is the unknown number of unobserved presences in the subsample S_p, n_{1p} is the number of observed presences in the subsample S_p, n_0 is the unknown total number of absences in S and n_1 is the unknown total number of presences in S. We assume that all the unknowns in Table 1 are random quantities from the effect of the missing data process in the subsample S_u. In particular we can consider n_{1u} as:

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

Moreover the following relationships hold:

$$\tilde{n}_{0u} = n_u - \tilde{n}_{1u} \qquad \tilde{n}_0 = \tilde{n}_{0u} \qquad \tilde{n}_1 = \tilde{n}_{1u} + n_p \tag{2}$$

Let π be the prevalence of the $\boldsymbol{Y}$ process, under the assumption that the subsample S_u is randomly drawn from the entire target population, the expected number of true presences in the full sample S is given by:

$$E[\tilde{n}_1] = E[\tilde{n}_{1u}] + n_p = \pi n_u + n_p \tag{3}$$

A typical presence-only data set consists of environmental information in the form of covariates covering the study area and of observed presences. That is why we cannot employ standard statistical approaches in the modeling of a species' presence covering the whole study area. In this work we propose the use an adjusted logistic model extending the approach in [2]. Denoting by s=1 that an observation is in the sample, the probability that a species of interest is present at a location with covariates $\boldsymbol{x}$ can be formalized as follow:

$$\Pr(y=1 \mid s=1,\eta;x) = \frac{\gamma_1 \exp\{\eta(x)\}}{\gamma_0 + \gamma_1 \exp\{\eta(x)\}} = \frac{\exp\left\{\eta(x) + \log\left(\frac{\gamma_1}{\gamma_0}\right)\right\}}{1+\exp\left\{\eta(x) + \log\left(\frac{\gamma_1}{\gamma_0}\right)\right\}}\ , \tag{4}$$

where $\eta(x)$ is a regression function, γ_0 =Pr(s=1|Y=0) and γ_1=Pr(s=1|Y=1) are the probabilities of sampling respectively from the absences and from the presences.

As introduced in [2] by Ward et al., we can handle the presence-only model following two approaches. The first one considers the full likelihood, i.e. the joint probability of the $\boldsymbol{Y}$ and $\boldsymbol{Z}$ processes, and the second one is based on the observed likelihood defined only with respect to the naive variable $\boldsymbol{Z}$.

First we analyse the full likelihood:

$$\begin{aligned} L(\eta; y,z,x) = &\prod_{i \in S_u} \left\{ \left[\Pr(y_i = 0 \mid s_i = 1,\eta; x_i)\right]^{1-y_i} \left[\frac{n_{1u}}{n_1} \Pr(y_i = 1 \mid s_i = 1, \eta; x_i)\right]^{y_i} \right\} \\ &\times \prod_{i \in S_p} \left\{ \left[\frac{n_{1p}}{n_1} \Pr(y_i = 1 \mid s_i = 1, \eta; x_i)\right]^{y_i} \right\} \end{aligned} \ . \tag{5}$$

In (6) it is introduced the observed likelihood with respect only to $\boldsymbol{Z}$:

$$L(\eta; z,x) = \prod_{i \in S_u} \Pr(z_i = 0 \mid s_i = 1, \eta; x_i) \times \prod_{i \in S_p} \Pr(z_i = 1 \mid s_i = 1, \eta; x_i)\ , \tag{6}$$

where

$$\Pr(z_i = 1 \mid s_i = 1, \eta; x_i) = \frac{n_p}{n_1} \frac{\exp\left\{\eta(x) + \log \frac{\gamma_1}{\gamma_0}\right\}}{1 + \exp\left\{\eta(x) + \log \frac{\gamma_1}{\gamma_0}\right\}} . \tag{7}$$

We propose an estimation procedure that works for both approaches. In order to manage computational aspects we need to discuss in details the role of the ratio $\frac{\gamma_1}{\gamma_0}$. From the case-control sampling, absences and presences are drawn separately at different and unknown rates, respectively γ_0 and γ_1, that we assume independent from the covariates $\boldsymbol{x}$. Given the size N of the target population we have that $\gamma_0 = \frac{n_0}{(1-\pi)N}$ and $\gamma_1 = \frac{n_1}{\pi N}$.

In our setting, being n_{1u} interpreted as a random quantity, we have that also γ_0 and γ_1 are random and we can write:

$$\frac{\tilde{\gamma}_1}{\tilde{\gamma}_0} = \frac{n_1}{n_0} \frac{1-\pi}{\pi} = \frac{\tilde{n}_{1u} + n_p}{n_u - \tilde{n}_{1u}} \frac{1-\pi}{\pi} . \tag{8}$$

In order to handle (8), Ward et al. [2] approximate $\tilde{\gamma}_1$ and $\tilde{\gamma}_0$ by their expected values:

$$\frac{\tilde{\gamma}_1}{\tilde{\gamma}_0} \approx \frac{E[\tilde{\gamma}_1]}{E[\tilde{\gamma}_0]} = \frac{\pi n_u + n_p}{\pi n_u} . \tag{9}$$

That ratio can be identified only if the prevalence π of population is known a priori. Our approach, which is based on MCMC simulations, does not require such knowledge. Under the assumption that S_u is a random sample from S, we can approximate π through the empirical random prevalence $\tilde{\pi}_u = \frac{\tilde{n}_{1u}}{n_u}$ which can be introduced in the MCMC algorithm through an augmentation step. In this way we have:

$$\frac{\tilde{\gamma}_1}{\tilde{\gamma}_0} \approx \frac{\tilde{n}_{1u} + n_p}{\tilde{n}_{1u}} . \tag{10}$$

From (10) we derive $\Pr(y_i \mid s_i = 1, \eta; x_i)$ and $\Pr(z_i \mid s_i = 1, \eta; x_i)$ that can be plugged into the full likelihood (5) and into the observed likelihood (6). We can now write the Bayesian model using either (5) or (6).

2.3 Bayesian model and MCMC algorithm.

Let $p(\eta)$ be the prior distribution of the regression function, e.g. if we consider a linear regression we can choose that as a multivariate Gaussian distribution over the support of the parameters. Given η and the covariate information $\boldsymbol{x}$, the $\boldsymbol{Y}$ process in the sample S can be described by independent Bernoulli random variables, each one with parameter $\theta_s(x) = \Pr(y = 1 \mid s = 1, \eta; x)$.
Then, the hierarchical Bayesian model is given by:

- $\eta \square p(\eta)$
- $Y_i \mid s_i = 1, \eta; x_i \square Be[\theta_s(x_i)]$
- $Z_i \mid y_i, s_i = 1 \square \Pr(z_i \mid y_i, s_i = 1)$ that can be derived from Table 1.

Then the joint posterior distribution of the model can be derived with respect to the full likelihood or to the observed likelihood. In the following scheme we present the MCMC algorithm which can be applied to both approaches in order to result estimates of the quantities of interest.

```
    Step 0: initialize $y_i \square Be(0.5)\ \forall i \in S_u$
Repeat:
    Step 1: set $n_{1u} = \sum_{i \in S_u} y_i$
    Step 2: sample $\eta \square \Pr(\eta \mid y, z)$ ;
    Step 3: sample $y_i \square \Pr(y_i \mid z_i, s_i = 1, \eta, x_i)\ \forall i \in S_u$ .
```

In Step 3 we draw from the Bernoulli distribution with parameter $\theta(x_i) = \frac{\exp\{\eta(x_i)\}}{1 + \exp\{\eta(x_i)\}}$. An estimation of the prevalence can be obtained by $\hat{\pi}_u = \frac{\overline{n}_{1u}}{n_u}$, where $\overline{n}_{1u}$ is the average of the simulations in Step 1.

3. Simulation Study

In this section we investigate the performance of our proposal on simulated data. We generate a binary response *Y* from the logistic model $\operatorname{logit}\theta(x_i) = \beta x_i$, where β= -1 while the covariate *x* is generated from a mixture of two Gaussian components with common variance and central values respectively $\mu_1 = -2$ and $\mu_2 = 2$.

In Figure 1 we show the distribution of ***X*** and ***Y*** for different level of the dispersion in the covariate.

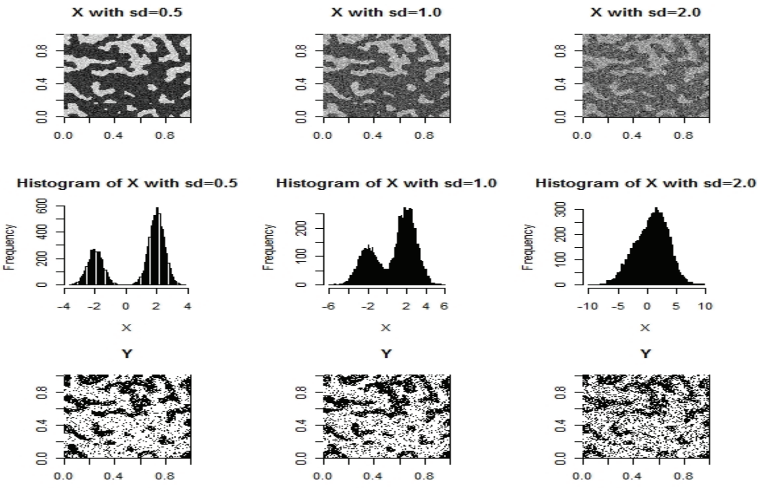

Fig. 1. Distribution of ***X*** and ***Y*** for different level of σ_X .

We generate a study area of N=10000 observations and randomly we select observed presences for S_p and pseudo-absences for S_u in a rate of 1:4. Then we fit the Bayesian model, in the observed likelihood version, for two different situations: with unknown π (M1) and assuming the population prevalence to be known (M2). The second situation represents our benchmark model and it can be considered similar to the model introduced in [2] by Ward et al.. Both models are fitted assuming the standard Gaussian $N(0,1)$ as the prior for β. We run 20000 iterations and discard the first 10000 as burn-in. In Table 2 we report the MCMC posterior mean and the 95% credibility interval for β and the MCMC estimation for the population prevalence π with respect to different levels of σ_X and sizes of S.

Table 2. Posterior mean and credibility interval for β and posterior mean for the prevalence (π^* represents the true population prevalence).

		σ_X =0.5; π^* = 0,3681		σ_X =1.0; π^* = 0,3809		σ_X =2.0; π^* = 0,4088	
		$\hat{\beta}$ (95% CI)	$\hat{\pi}_u$	$\hat{\beta}$ (95% CI)	$\hat{\pi}_u$	$\hat{\beta}$ (95% CI)	$\hat{\pi}_u$
n=50	M1	-0.14 (-0.75; 0.45)	0.487	-0.50 (-1.21; 0.12)	0.467	-0.89 (-1.87; -0.12)	0.463
	M2	-0.18 (-0.80; 0.46)	0.484	-0.53 (-1.27; 0.10)	0.466	-0.90 (-1.90; -0.12)	0.462
n=500	M1	-0.85 (-1.12; -0.61)	0.431	-0.93 (-1.21; -0.69)	0.4007	-0.80 (-1.10; -0.55)	0.419
	M2	-0.90 (-0.80; 0.46)	0.428	-0.96 (-1.23; -0.69)	0.3996	-0.82 (-1.11; -0.58)	0.418
n=5000	M1	-1.030 (-1.11; -0.95)	0.364	-1.07 (-1.16; -0.99)	0.3756	-1.02 (-1.13; -0.92)	0.4008
	M2	-1.029 (-1.11; -0.94)	0.364	-1.06 (-1.16 -0.97)	0.3759	-1.01 (-1.11; -0.91)	0.4011

We can see that when n increases, then the prevalence estimates become closer to the true population one with respect to all the different levels of dispersion in $\boldsymbol{X}$. Also the performance with respect to the benchmark model seems quite promising; in fact the estimates concerning the two models M1 and M2 are very close in all the several scenarios.

4. Conclusions

The model proposed aims at estimating the parameters of a logistic linear regression adapted to presence-only data. The introduction of a random approximation of the correction factor in the adjusted logistic model allows us to overcome the need to know the population prevalence a priori. Under the assumption that the pseudo-absence locations are randomly sampled from the entire study area, we can estimate regression parameters jointly with the prevalence through a data augmentation step in the MCMC algorithm.

The simulation study, described in Section 3, allowed us to evaluate of the performance of the method in practice. The results are good and encouraging. Further works envisage real data applications and the development of an explicitly spatial model for presence-only data.

References

[1] Pearce JL, Boyce, MS. Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology* 2006; **43**: 405-412.

[2]Ward G, Hastie T, Barry S, Elith J, Leathwick A. Presence-only data and the EM algorithm. *Biometrics* 2009; **65**: 554-563.

[3] Liu SJ. *Monte Carlo Strategies in Scientific Computing*. Springer:New York; 2004.