Spatial Bayesian Modeling of Presence-only Data

Fabio Divino
S.T.A.T., University of Molise, fabio.divino@unimol.it

Natalia Golini, Giovanna Jona Lasinio
Department of Statistical Sciences, "Sapienza" University of Rome

Antti Penttinen
Department of Statistics, University of Jyväskylä

Abstract: When the only available information is the true presence of a species at few locations of a study area we refer to the data as *presence-only data*. Presence-only data problem can be seen as a missing data problem with asymmetric and partial information on a presence-absence process. This problem often characterizes ecological studies requiring the prediction of potential spatial extent of a species in suitable areas. Here we propose a Bayesian logistic spatial model adapted to presence-only data with environmental covariates available over the entire area. The spatial dependence among the observations is modelled indirectly as a latent Gaussian Markov field over the landscape, through a data augmentation MCMC algorithm we are able to estimate regression parameters jointly with the prevalence.

Keywords: Bayesian model, Data augmentation, MCMC, Presence-only data, Spatial distribution.

1 Introduction

In the environmental sciences, the evaluation of spatial distribution of species and its interaction with ecological variables is of primary interest *i.e.* to better plan and manage strategies in habitat conservation. When presence/absence information on a species is available in a given area together with environmental covariates, the logistic regression model represents the natural approach to estimate the prevalence of such species. Unfortunately, in many ecological studies, the collection of definitive absences can be expensive or difficult. In those cases the information available is not complete, we can observe only presences (Pierce and Boyce 2006) of the species at few locations jointly with the environmental covariates referred to the whole study area. In this work we propose a hierarchical Bayesian model to handle presence-only data, based on an adjusted logistic regression model (Ward et al. 2009). Following Divino et al. (2011) we introduce a random approximation of the correction factor

in the model that allows us to overcome the need to know *a priori* the prevalence of the species. We can estimate regression parameters jointly with prevalence through a data augmentation MCMC algorithm (Divino et al. 2011). We account for spatial variation adding a spatial random effect in the regression function.

2 Materials and Methods

With respect to a population $\mathcal{P}$ of spatially referenced sites i, let Y be a binary presence/absence process, X a set of covariates and $\mathcal{P}_p$ the subset of $\mathcal{P}$ where the species is present ($Y = 1$). When only presences are observed, samples (S_p) from the process Y can be drawn only from the population $\mathcal{P}_p$ and the usual case-control approach in logistic regression cannot be adopted as absences ($Y = 0$) are not directly observed. Lancaster and Imbens (1996) and Ward et al. (2009) proposed to overcome this problem by considering a completed sample composed by S_p and a second sample S_u, independent of S_p, ideally taken from the whole population $\mathcal{P}$. In this way the complete data sample S is composed by n_p presences (observed in S_p) and n_u unobserved values (S_u). Let Z be a stratum variable such that $Z_i = 0$ if $i \in S_u$ and $Z_i = 1$ if $i \in S_p$. Notice that $Z_i = 1$ implies $Y_i = 1$ while $Z_i = 0$ implies that Y_i can assume value in $\{0, 1\}$. Hence we can identify the following quantities: $(Z = 0, Y = 0)$ n_{0u} is the unknown number of absences in the subsample S_u, $(Z = 0, Y = 1)$ n_{1u} is the unknown number of presences in the subsample S_u, $(Z = 1, Y = 1)$ n_{1p} is the number of observed presences in the subsample S_p, n_0 is the unknown total number of absences in S, n_1 is the unknown total number of presences in S and $n = n_1 + n_0$ is the complete sample size. All the unknowns are random quantities induced by a censoring effect acting on the complete sample S. In particular we can write n_{1u} as $\tilde{n}_{1u} = \sum_{i \in S_u} Y_i$, where the $\sim$ represents the random nature of the quantity. Now let $\pi = P(Y = 1)$ be the prevalence of the species in the area, under the assumption that S_u is a random sample from the population $\mathcal{P}$ we have that $E[\tilde{n}_{1u}] = \pi n_u$. If we assume that the covariates X, concerning the environmental information on the process Y, are available for all sites in the population, we can use the approach introduced by Ward et al. (2009) and developed in a Bayesian framework by Divino et al. (2011). For a generic site in the sample with covariates x, starting from the usual case-control logistic model the conditional probability that a species of interest is present is given by

$$P(Y = 1|s = 1, \eta; x) = \frac{\exp\{\eta(x) + \log(\frac{\gamma_1}{\gamma_0})\}}{1 + \exp\{\eta(x) + \log(\frac{\gamma_1}{\gamma_0})\}} \tag{1}$$

where $s = 1$ denotes that the site is included in S, $\eta(x)$ is the regression function, $\gamma_0 = P(s = 1|Y = 0)$ and $\gamma_1 = P(s = 1|Y = 1)$ are the unknown probabilities of sampling from the absences and from the presences respectively. The ratio $\frac{\gamma_1}{\gamma_0}$ adjusts the logistic model under the case-control design. Following Ward et al. (2009), we can manage the presence-only data problem by considering the joint probability

distribution of Y and Z and write the full likelihood model (see Ward et al. 2009 for details). We can also consider the observed likelihood, built only with respect to the stratum variable Z that results in an average over the process Y. In both likelihood models, the unknown ratio $\frac{\gamma_1}{\gamma_0}$ can be approximated as follow:

$$\frac{\tilde{\gamma_1}}{\tilde{\gamma_0}} \approx \frac{\tilde{n}_{1u} + n_p}{\tilde{n}_{1u}} \tag{2}$$

the above expression can be handled by a data augmentation step in the estimation procedure. The regression function adopted in this work is linear with a spatially structured random effect u accounting for latent factors introducing geographical dependence into species distribution. We can now write the hierarchical Bayesian model. Let δ be the vector of hyperparameters with hyperprior $p(\delta)$. Conditioned on δ, the regression parameters, β, are Gaussian random variables and the random effect u is a Gaussian Markov random field. Given β, u and the covariate x, the process Y is set of Bernoulli random variable with probability of occurence $\pi_s(x) = P(Y = 1|s = 1, \eta; x)$. At the lowest level of the hierarchy, the conditional distribution of Z given Y can be easily derived from the above described relations between the two processes. Then, the hierarchical Bayesian model is given by: (i) $\delta \sim p(\delta)$; (ii) $\beta|\delta \sim MN(\delta)$ and $u|\delta \sim GMRF(\delta)$; (iii) $Y_i|s_i = 1, \beta, u_i, x_i \sim Be[\pi_s(x_i)]$; (iv) $Z_i|Y_i, s_i \sim P(Z_i|Y_i, s_i = 1)$. Notice that the spatial structure of the random effect u is given by the geographical neighborhood system among all sites in the population $\mathcal{P}$. In the following scheme we describe the MCMC algorithm implementing the estimation of our model:

Step 0: initialize δ, β, u and Y over $\mathcal{P}$;
Step 1: set $n_{1u} = \sum_{i \in S_u} Y_i$;
Step 2: sample $\delta \sim P(\delta|Y, Z, \beta, u)$;
Step 3: sample $\beta \sim P(\beta|Y, Z, \delta)$;
Step 4: sample $u \sim P(u|Y, Z, \delta)$ over $\mathcal{P}$;
Step 5: sample $Y_i \sim P(Y_i|Z, \beta, u_i, x_i)$ over $\mathcal{P}$.

Remark that we need to perform data augmentation (Step 4 and Step 5) over the entire population $\mathcal{P}$ for both u and Y processes in order to consider the spatial structure of the sites enclosed in both samples S_u and S_p. The only requirement to perform the augmentation is that the covariates X are available for all sites in $\mathcal{P}$. A nice feature of this estimation procedure is that we can easily obtain the prevalence estimate $\hat{\pi}_u = \frac{\bar{n}_{1u}}{n_u}$, where $\bar{n}_{1u}$ is the MCMC average of samples drawn in Step 1.

3 Results

In this section we report preliminary results from a small simulation study aiming at investigating the behaviour of our proposal in a very simple situation. We generate a population of 100 observations on a regular 10×10 lattice from the above described model. In this example Y is obtained from the logistic model $\eta(\boldsymbol{X}) = \beta x + u$, where

$\beta = -2$, the covariate X is generated from a mixture distribution with two Gaussian components with standard deviation $\sigma_1 = \sigma_2 = 0.5$ and mean $\mu_1 = -2$ and $\mu_2 = 2$, u is a zero mean intrinsic first order Gaussian Markov random field with precision $k = 1.5$ and prevalence $\pi = 0.1$. From this population we obtain 100 samples by randomly thinning 30% of the available presences. We compare the performance of our model (M1), with unknown prevalence, with the same model but with known prevalence in the logistic correction (M2) and with the non spatial model prosed in Divino et al. (2011) (M3). The three models are fitted with the same prior settings: $\beta \sim N(0, 100)$ and k fixed (for M1 and M2). We run 20000 iterations of the MCMC procedure with a burn-in of 10000. To evaluate models performances we compute 95% credibility intervals (CI) for β in each simulation using the 10000 samples from the posterior distribution, the same intervals for the prevalence are computed from the 100 simulations and the misclassification error is computed for each model by setting to 1 grid cells with occurrence probability larger then 0.5 and compare results with the "true" population. Results are as expected: the "best" model in terms of point estimates accuracy is M2 with smaller CI for $\hat{\beta}$ and $\hat{\pi}$, followed by M1; all models have a tendency to overfit with empirical coverage around 99%. In terms of predictive capacity the average misclassification error is around 3% for all models, as expected M1 and M2 better perform as far as the localization of presences is concerned.

4 Concluding remarks

The above preliminary results are encouraging, especially in terms of predictive capacity of the proposed model. Several issues will be object of further work, such as identifiability problems related to a not zero intercept. Extensive simulation studies will be carried on too.

References

Divino F., Jona Lasinio G., Golini N., Pettinen A. (2011) Data Augmentation Approach in Bayesian Modelling of Presence-only Data, *Procedia Environmental Sciences* to appear.

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

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

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