A Bayesian model for presence-only semicontinuous data, with application to prediction of abundance of *Taxus Baccata* in two Italian regions.

B. DI LORENZO, A. FARCOMENI and N. GOLINI

In studies about the potential distribution of ecological niches only the presence of the species of interest is usually recorded. Pseudo-absences are sampled from the study area in order to avoid biased estimates and predictions. For cases in which, instead of the mere presence, a continuous abundance index is recorded we derive a two-part model for semicontinuous (i.e., positive with excess zeros) data which explicitly takes into account uncertainty about the sampled zeros. Our model is a direct extension of the one in Ward et al. (2009). It is fit in a Bayesian framework, which has many advantages over the maximum likelihood approach of Ward et al. (2009), the most important of which is that the prevalence of the species does not need to be known in advance. We illustrate our approach with real data arising from an original study aiming at the prediction of the potential distribution of the *Taxus baccata* in two central Italian regions.

Key words: Bayesian methods; Ecological niche; Excess zeros; Potential distribution; Presence-only data; Semicontinuous data; Two-part model.

1 INTRODUCTION

The elaboration of appropriate conservation actions of a forest habitat should be based on the knowledge of its dynamics, and of the potential spread of the species in the habitat. Decisions are supported by definition of environmental niches, description of the use of habitats, and prediction of geographic distributions of species. Accurate maps of geographic distributions can also help understand the effects of climate change.

A potential distribution map provides a spatial prediction of the suitable areas for a species. The main aim of these analysis is related to prediction rather than to interpretation, so that often machine learning approaches are used (Prasad, Iverson, and Liaw 2006; Scarnati et al. 2009b). Chaubert-Pereira, Guédon, Lavergne, and Trottie (2009) propose an approach more focused on interpretation, in a related context.

B. Di Lorenzo is PhD Student, Department of Statistic, Probability and Applied Statistics, Sapienza - University of Rome, Rome, 00185, Italy. A. Farcomeni is Assistant Professor, Department of Infectious Diseases and Public Health, Sapienza - University of Rome, Rome, 00185, Italy. N. Golini is PhD Student, Department of Statistic, Probability and Applied Statistics, Sapienza - University of Rome, Rome, 00185, Italy (e-mail natalia.golini@uniroma1.it).

In recent years new data sources (atlases, museum and herbarium records, species lists, incidental observation databases and radio-tracking studies) and GIS tools have been increasingly used in ecological studies. A peculiar characteristic of these data sources is that, often, only information about locations where the species is present are available. Methodologies targeted especially for these data need to be developed.

Motivation of this work arises from a study about the spatial distribution of *Taxus baccata* in two central Italian regions. This original study was performed in order to support decisions for species management, in particular about conservation strategies for *Taxus baccata*.

In our study, not only we recorded the presence of the species, but also an index measuring its abundance, the so called Importance Value (IV). Nevertheless, we have measures only about locations with a positive abundance. As any other study in which only presences are observed, direct use of the data may lead to predictions of the potential distribution in which a presence is predicted too frequently, and abundance is over estimated.

This problem is usually tackled by sampling locations of *pseudo-absence*, that is, locations in which the species of interest is assumed to be absent, and using them as if they were truly observed zeros. See Pearce and Boyce (2006) for a brief discussion. Elith et al. (2006) illustrate that using sampled zeros can considerably improve predictions. See also Zaniewski, Lehmann, and Overton (2002) and Engler, Guisan, and Rechsteiner (2004).

A limitation of most common approaches is that they treat pseudo-absences as if they were truly observed zeros, without taking into account that some of the sampled zeros could actually be locations in which the species is present. There are few exceptions: Phillips, Anderson, and Schapire (2006) introduce the method of Maximum Entropy (Maxent) for modelling species geographic distributions; Ward et al. (2009) explicitly take into account bias due to presence-only sampling. The model in Ward et al. (2009) can only be used when the outcome of interest is a dichotomous variable simply measuring whether the species is present or absent in a given location. In this work we extend their model to abundance data, that is, to an outcome which is either zero or a positive real number. These data are usually referred to as semicontinuous data or data with excess zeros, and analysis can be carried out by means of a two-part model which combines modelling the probability of a positive outcome with the density of the outcome conditionally on the fact that it is positive. Two-part models similar to our proposal, but without uncertainty related to the zeros, have been recently discussed and fit in the classical framework by Li, Elashoff, Robbins, and Xun (2009) and by Leathwick et al. (2008) among others. Zhou and Tu (1999) derive an ANOVA type test for two-part models, and Lachenbruch (2002) compares different testing strategies. The main innovation in our work is that we have uncertainty related to the zeros, and explicitly take it into account. Our resulting model can hence also be thought to as a two-part model with partial possible measurement error. Further, we adjust for the case-control type sampling performed.

We derive inference for our model in a Bayesian framework, which has many advantages over the maximum likelihood approach of Ward et al. (2009), the most important being that the prevalence of the species does not need to be known in advance.

The paper is structured as follows. In Section 2 we present our model; in Section 3 we discuss computational aspects of the Bayesian approach for parameters estimation, derive the predictive distribution and show how to compute predictions minimizing the posterior expected loss. In Section 3.4 we describe a strategy for validating the predictive ability of our model. In section 4 we conduct a simulation study in order to illustrate the performance of our proposed approach, in Section 5 we outline the results of the application of our proposal to the motivating example. We conclude with a brief discussion in Section 6.

A sample `R` (R Development Core Team 2009) code for fitting the proposed model is available at `http://afarcome.interfree.it/mcmcTaxus.r`.

2 MODELLING METHOD

Let Y be a random variable measuring the outcome of interest, where $Y \geq 0$. We only observe locations in which $Y > 0$, and then sample a certain number of pseudo-absences from the remaining locations. The resulting random variable is denoted with Z. We observe $Z_1, \ldots, Z_n$ at n locations; where $Z_i > 0$ implies $Y_i = Z_i > 0$; but when $Z_i = 0$, we only know that $Y_i \geq 0$. We let $\xi = \Pr(Y > 0)$ denote the prevalence of the species of interest.

As is done in Ward et al. (2009), we assume that the observed locations with $Y > 0$ are sampled at random from all locations where the species is present and pseudo-absences are generated randomly from the remaining locations in the study area in a case-control fashion (i.e., the number of pseudo-absences is fixed and could be based on the number of available presences).

We must keep in mind that there may be a selection bias for the observed abundance, since for instance more accessible locations may be more often sampled. We will ignore for the time being this bias, which is sometimes counterbalanced by sampling zeros with a similar bias. This combination of the samples of observed presences and pseudo-absence is what we will refer to as *presence-only data*.

We denote with X_i a vector of environmental covariates for the i-th location, which are known for the entire study area. We model the semicontinuous response through a two-part model. The two parts are usually made of a logistic model for the probability that the response is positive, and a regression model for the log-response conditionally on the fact that it is positive. We here extend the classical two-part model for taking into account uncertainty related to pseudo-absences. We use an indicator function s so that $s_i = 1$ indicates that the i-th observation is either an observed presence or a pseudo-absence.

The logit model can be specified as follows:

$$P(Y>0|\mathbf{x}) = \frac{e^{\eta(\mathbf{x})}}{1+e^{\eta(\mathbf{x})}} \Rightarrow \text{logit}[P(Y>0|\mathbf{x})] = \eta(\mathbf{x}), \tag{2.1}$$

where $\eta(\mathbf{x}) = \mathbf{x}'\beta$ is linear in $\mathbf{x}$.

For the continuous part of our model it is assumed that

$$E[\log(Y)|Y>0,\tilde{\mathbf{x}}] = \eta_{\prime}(\tilde{\mathbf{x}}), \tag{2.2}$$

where $\eta_{\prime}(\tilde{\mathbf{x}}) = \tilde{\mathbf{x}}'\alpha$ and $\log(Y)$ is conditionally distributed as a normal variate with standard deviation σ.

The notation used underlines that the same or a different set of covariates can be used on each part of the model. We bound our attention to linear models. Other choices are grounds for further work. Furthermore, in Section 5 we will illustrate that the above assumptions of linearity are adequate for the data at hand.

We now note that the case-control scheme used for sampling pseudo absences ignores prevalence. A naive model which ignores this fact may be seriously biased (Keating and Cherry 2004), especially when the species is not rare. Along the lines of Ward et al. (2009) we then adjust the two-part model through a "case-control style adjustment" (McCullagh and Nelder 1989, p. 111). In Appendix A we show that the logistic model (2.1) needs to be adjusted according to

$$\text{logit}(P(Y>0|\mathbf{x},s=1)) = \eta(\mathbf{x}) + \log\left(\frac{n_p+\xi n_u}{\xi n_u}\right), \tag{2.3}$$

where n_p and n_u are respectively the number of observed abundance $(z>0)$ and the number of pseudo-absence locations $(z=0)$; while the regression model (2.2) needs no adjustment conditionally on $s=1$.

In the case control framework, one usually uses a ratio of 1:1, 2:1 or 3:1 for the controls. In our case, controls are pseudo-absences. A different strategy is used in Elith et al. (2006), where a large number of downweighted pseudo-absences are used. In this paper we prefer setting $n_p = n_u$, and repeatedly sampling pseudo-absences and re-fitting the model in order to check for spurious effects which may be present among a large number of pseudo-absences, even if downweighted.

Secondly, since Y is not observed, we need to link parameters to the observed Z in order to perform inference. To do so, we derive the *observed likelihood* $L(\theta|\mathbf{z},X)$ for the presence-only data, which can be expressed as:

$$\prod_{i=1}^{n}\left[\frac{1+e^{\eta(\mathbf{x}_i)}}{1+\left(1+\frac{n_p}{\xi n_u}\right)e^{\eta(\mathbf{x}_i)}}\right]^{1_{\{z_i=0\}}}\left[\frac{\frac{n_p}{\xi n_u}\exp\left\{\eta(\mathbf{x}_i)-\frac{(\log(z_i)-\eta_{\prime}(\tilde{\mathbf{x}}))^2}{2\sigma^2}\right\}}{1+\left(1+\frac{n_p}{\xi n_u}\right)e^{\eta(\mathbf{x}_i)}}\frac{1}{z_i\sqrt{2\pi}\sigma}\right]^{1_{\{z_i>0\}}} \tag{2.4}$$

where θ is a short-hand notation for the parameters at stake and 1_C is the indicator function for condition C.

The algebra and rationale behind (2.4) are developed in full in Appendix B.

3 INFERENCE

3.1 PRIORS

In order to derive inference in the Bayesian framework we need to specify prior distributions. We use the following specification:

$$\pi(\beta,\alpha,\sigma,\xi) = \pi(\beta|\Sigma_\beta)\pi(\alpha|\Sigma_\alpha)\pi(\sigma)\pi(\xi), \tag{3.1}$$

where $\pi(\beta|\Sigma_\beta)$ and $\pi(\alpha|\Sigma_\alpha)$, following a standard practice in Bayesian regression, denote zero-centered multivariate normals.

We complete prior specification letting $\pi(\sigma)$ be an inverse Gamma and $\pi(\xi)$ summarize available information on prevalence.

A special role is played by the informative prior on ξ, due to the fact that data contain very little information on ξ (see Ward et al. (2009) for the identifiability issues related to ξ). In practice, inference and predictions are based on the integrated likelihood (with respect to the prior on ξ). We stress that nonidentification makes inference arbitrarily sensitive to the prior. The proposed model considers a parametrization with a simple contextual meaning, so that it is possible to elicit an informative prior for ξ. For methods in prior elicitation, see for instance Kadane et al. (1980); Kadane and Wolfson (1998); Garthwaite, Kadane, and O'Hagan (2005), and references therein.

3.2 MODEL FIT

Since not all Y_i are observed, we have a missing data model.

In order to approximate the posterior distribution, we make use of an MCMC sampling scheme adapted from Diebolt and Robert (1994). The MCMC sampling scheme we propose is a Bayesian counterpart of the EM algorithm, which gives samples from the posterior distribution after burn-in. The sampling scheme is based on alternating a data augmentation/imputation step, in which the latent observations Y are sampled from their full conditional, with posterior sampling steps.

We hence augment the data making use of the latent observations Y, and derive consequently the so called *complete likelihood*. In our context, the complete likelihood $L(\theta|\mathbf{z},\mathbf{y},X)$ for the presence-only data, in terms of both z and y, is given by:

$$\prod_i \left[\frac{1}{1+\left(1+\frac{n_p}{\xi n_u}\right)\exp\{\boldsymbol{\beta}\mathbf{x}_i\}}\right]^{1_{\{y_i=0\}}} \tag{3.2}$$
$$\left[\frac{\exp\{\boldsymbol{\beta}\mathbf{x}_i\}\left(1+\frac{n_p}{\xi n_u}\right)}{1+\left(1+\frac{n_p}{\xi n_u}\right)\exp\{\boldsymbol{\beta}\mathbf{x}_i\}}\frac{1}{y_i\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}\left(\log(y_i)-\boldsymbol{\alpha}\tilde{\mathbf{x}}_i\right)^2\right\}\right]^{1_{\{y_i>0\}}}$$

where $\boldsymbol{\beta}$ represents the vector of k covariates are used for the logistic part and $\boldsymbol{\alpha}$ represents the vector of h covariates are used for the linear regression part. So that $\eta(\mathbf{x}) = \boldsymbol{\beta}\mathbf{x}_i$ and $\eta'(\tilde{\mathbf{x}}) = \boldsymbol{\alpha}\mathbf{x}_i$. A detailed derivation of (3.2) is given in Appendix C.

The general iteration of the (Metropolis within) Gibbs sampling scheme we propose is detailed in Algorithm 1.

Algorithm 1 Gibbs sampling scheme

1. Sample the latent variables Y_i from $f(Y_i|Z_i,X_i,s_i=1)$, $i=1,\ldots,n$; where $f(Y_i|Z_i,X_i) = 1_{Y_i=Z_i} * 1_{Z_i>0} + 1_{Z_i=0} f(Y_i|Z_i=0,X_i,s_i=1)$. That is, we simply set $Y_i = Z_i$ when $Z_i > 0$ and when $Z_i = 0$ note that

$$f(Y_i|Z_i=0,X_i,s_i=1) = f(Y_i|X_i,s_i=1),$$

since we assumed that data are sampled uniformly at random from the study area. Sampling of Y_i when $Z_i = 0$ must then be performed in two steps, since $f(Y_i|X_i,s_i=1)$ is a mixed measure. First, we shall sample $1_{Y_i>0}$ from $P(Y_i>0|X_i,s_i=1)$, and then set $Y_i = 1_{Y_i>0} y$, where y is sampled from $f(y|Y_i>0,X_i)$.

2. Sample β (the regression parameters for the logistic part) from

$$\pi(\beta|Y,X) \propto \pi(\beta) \frac{\exp\{\sum_i^n 1_{Y_i>0} \boldsymbol{\beta} \tilde{\mathbf{x}}_i\}}{1+\left(1+\frac{n_p}{\xi n_u}\right)\exp\{\sum_i^n \boldsymbol{\beta} \tilde{\mathbf{x}}_i\}}$$

3. Sample the remaining parameters α and σ simultaneously as

$$\pi((\alpha,\sigma)|Y,X) \propto \pi(\alpha,\sigma) L(\theta|\mathbf{z},\mathbf{y},X)$$

4. Sample ξ from its prior.

At Step 1 we sample latent variables from their full conditionals. The innovation with the sampling scheme in Diebolt and Robert (1994) is that latent variables are not discrete, and actually known when $Z_i > 0$. We then augment generating Y_i from its semicontinuous full conditional distribution when $Z_i = 0$. When $Z_i > 0$, Y_i is not sampled since its full conditional is a point mass on Z_i. Convergence of the chain is guaranteed from the fact that $f(Y_i|Z_i,X_i,s_i=1)$, albeit arising from a unusual semicontinuous distribution, is exactly the full conditional for Y_i. All needed regularity conditions are consequently implied by model assumptions.

After we have sampled or set values for Y, Step 2 arises from straightforward conditional independence conditions, (2.4) and (2.3).

In Step 3, we face several difficulties associated with setting up Metropolis Hastings (MH) steps for the parameters in (α,σ). Key to success for MH is linked to a clever choice for the candidate transition kernel, which does not seem readily available here. Furthermore,

the last full conditional distribution is also potentially multi-modal, and even if a good candidate transition kernel were available, tuning of MH would be made harder by volatility in the latent indicators $1_{Y_i>0}$. In order to avoid difficulties linked with setting up Metropolis Hastings, we sample (α, σ) simultaneously with Adaptive Rejection Metropolis Sampling (Gilks, Best, and Tan 1995). In our experience the ARMS approach works nicely, and needs no tuning.

For sampling the logistic regression parameters at Step 2 we still use an ARMS for simplicity, even if there are many different alternative approaches for this standard problem.

We sample ξ from its prior since presence-only data do not return information on this parameter. It is straightforward to check that the full conditional for ξ coincides with the prior. As noted before, posterior summaries will be then model averaged (Hoeting, Madigan, Raftery, and Volinsky 1999) with respect to prior knowledge about ξ. This approach is a powerful and simple tool for explicitly incorporating uncertainty related to the prevalence; and when the prevalence is known, it is straightforward to modify the approach by simply setting the prior for ξ as a point mass.

3.3 PREDICTION

The distribution of predictions is the main target of our analysis. Prediction involves obtaining the posterior distribution for the study area, which we can denote by $\pi(Y_{new}|Z, X_{new}, X)$, where (Y_{new}, X_{new}) denotes the response for a new location with the associated (known) covariates. Bayes theorem can be used to show that

$$\pi(Y_{new}|Z, X_{new}, X) = \int_{\theta} f(Y_{new}|X_{new}, \theta)\pi(\theta|Z, X)\, d\theta. \tag{3.3}$$

After MCMC sampling, having obtained a sample $\theta_1, \ldots, \theta_B$ from the posterior, computation of (3.3) is straightforward through Monte Carlo integration:

$$\pi(Y_{new}|Y, X_{new}, X) \cong 1/B \sum_j f(Y_{new}|X_{new}, \theta_j).$$

The use of the predictive distribution (3.3) goes well beyond building predictions for each location of interest. In pratice, a posterior probability distribution is available for each location. Hence, the predictive distribution can be used to build predictions by minimizing the posterior expected loss, but also it can be plotted for certain locations of interest, it can be used to predict more sophisticated quantities (e.g., the probability that an IV exceeds a certain threshold in a given location, the distribution of the number of locations in which the IV is above a certain threshold, and so on).

Many of these summaries are of interest to the biologists, but the most important is probably the map with predictions minimizing the posterior expected loss. The posterior expected loss is minimized if we predict a presence whenever

$$P(Y_{new} > 0|X_{new}, Z, X) \cong 1/B \sum_j \frac{e^{x'_{new}\beta_j}}{1 + e^{x'_{new}\beta_j}} > \rho,$$

where $\rho = 0.5$ under the 0-1 loss, and $\rho = K_1/(K_0 + K_1)$ if we set as K_1 the loss of a false positive and as K_0 the loss for a false negative (see Berger (1985), Bernardo and Smith (1994)). When a presence is predicted, the predicted abundance minimizing the quadratic loss is the posterior expected IV, that is,

$$E[\log(Y_{new})|X_{new}, Z, X, Y_{new} > 0] \cong 1/B \sum_j \tilde{x}' \alpha_j.$$

3.4 VALIDATION THROUGH CROSS-VALIDATION PREDICTIVE DENSITIES

Since prediction is the main objective of our analysis, it is important to use appropriate devices for validation of the predictive ability of the model. When the number of observed presences is large enough, the observations may be randomly split in a training and a test set, and the predictive accuracy estimated on the test set (this is the approach of Ward et al. (2009), see also Hastie, Tibshirani, and Friedman (2001)). When the number of observed presences is small, using a smaller sample for estimation may result in a sensibly poorer performance of the model. A cross-validation approach is probably more suitable in such cases.

In our Bayesian framework we hence derive cross-validation predictive densities (Gelfand, Dey, and Chang 1992; Gelfand 1996) to check whether the model is adequate.

The cross-validation predictive densities are denoted by $\{\pi(y_r|Z_{(r)}), r = 1, \ldots, n\}$, where $Z_{(r)}$ is the set of observed presences and sampled pseudo-absences from which the r-th observation was deleted.

For each observation in the sample we use a leave-one-out principle and obtain a corresponding predictive density $\pi(y_r|Z_{(r)})$, which suggests what values of y_r are likely when the model is fit to all the observations except the r-th. The actual z_r can be compared to this density to see whether it is likely under the model, and consequently the model can be deemed to be valid if a large majority of the observations are likely with respect to their cross-validation predictive density.

In order to approximate the cross-validation predictive densities we use the so-called composition method (Gelfand 1996; Tanner 1996), which is computationally efficient and leads to reliable estimates. The cross-validation predictive densities are approximated, in this approach, with an importance sampling scheme which we now describe.

Let once again $\theta_1, \ldots, \theta_B$ be a sample from the posterior distribution $\pi(\theta|Z)$. Fix $r \in \{1, \ldots, n\}$, and compute the ratio

$$\frac{\pi(\theta_j|Z_{(r)})}{\pi(\theta_j|Z)} \propto w_j,$$

for $j = 1, \ldots, B$ for weighting. The weights $(w_1, \ldots, w_B)$ are then used for obtaining a sample from $\pi(\theta|Z_{(r)})$, which we denote with $(\theta_1^*, \ldots, \theta_B^*)$. The latter vector is obtained by resampling with replacement from the collection $(\theta_1, \ldots, \theta_B)$ with probabilities proportional to $(w_1, \ldots, w_B)$. A sample $y_1^*, \ldots, y_B^*$ from $\pi(y_r|Z_{(r)})$ is now obtained along the lines

of Section 3.3. The composition method shall be repeated for each $r = 1,\ldots,n$; finally obtaining n samples of size B from each cross-validation posterior predictive density.

Since the sampled pseudo-absences could be unobserved presences, we suggest to perform validation only using the predictive densities corresponding to the observed presences. We then perform model validation as follows (see Gelfand et al. (1992)): for each observed presence, we build a $(1-\alpha)$ credibility interval from the corresponding cross-validation predictive density. The $(1-\alpha)$ credibility interval is the shortest interval containing $(1-\alpha)$ posterior probability mass. Then, we define the *prediction coverage probability* as the proportion of predictive intervals covering the corresponding observed presences, and declare the model to be valid if the prediction coverage probability is close to the nominal level $(1-\alpha)$.

4 SIMULATION STUDY

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

We generate a semicontinuous response Y from the following two-part model

$$\text{logit}[P(Y > 0 \mid x_1, x_2)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

and, conditionally on $Y > 0$ and x_3

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

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

In Table 1 we report the average Mean Squared Error (MSE) of the parameter estimates of Bayesian and EM model for different values of N and λ.

[Table 1 about here.]

For our Bayesian method it can be seen that, as expected, the MSE decreases with N. On the other hand, there does not seem to be a strong dependence on λ, indicating that it does not really matter how many presences are obtained, as long as these are sampled independently and uniformly from the study area and the final sample size is large enough. The EM algorithm seems to be dependent both on N and λ, and it is sometimes outperformed by the Bayesian approach even if we gave it the unfair advantage of assuming a known, and correct, prevalence. The MSE for the regression coefficients are in general comparable, but the Bayesian approach seems to work much better than the frequentist method in estimating σ. This is due to a negative bias in the estimate of σ obtained with the EM algorithm, which could be explained by the optimism in assuming a known prevalence. The same assumption seems to lead to a smaller MSE in the coefficients of the logistic part when N is large, and larger when N is small. In Table 1 we also show the mean length of the 95% credibility intervals and their frequentist coverage for ξ. It can be seen that the frequentist coverage is very large, and that the mean length is large too and reflects prior inputs (i.e., a much smaller mean length could be obtained with a more concentrated prior).

A further comparison between the Bayesian and frequentist method is given in Table 2, where we compare the predictive performance of the methods. At each replication we compute the MSE for positive predictions, and summarize the performance of the presence-absence part of the model computing sensitivity and specificity.

[Table 2 about here.]

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

Finally, we provide a small study evaluating sensitivity of the parameter estimates to the choice of priors. In Table 3 we show the MSE of the parameters when $N = 10000$ and $\lambda = 0.1$ for additional sets of priors. Prior set (a) is the set used for the previous simulations and described at the beginning of the section. In prior set (b) we have added a bias to our priors. In fact, we have centered the priors for logistic and regression coefficients on -0.5, and further we have used a Gamma with parameters 1.5 and 1 for the precision parameter. Prior set (c) is equivalent to (a), with the exception of the prior for the prevalence parameter, where we have used a Beta distribution with parameters 0.46 and 2.64; and finally in prior set (d) we use zero centered Student's T distributions with three degrees of freedom for the β and α parameters.

[Table 3 about here.]

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

5 PREDICTING THE ABUNDANCE OF *TAXUS BACCATA*

Taxus baccata is a relict of the Cenozoic flora, characterized by warm-humid climatic conditions. It survived glaciations in refugia areas, and may have followed *Fagus* in successive postglacial expansions. This process has determined the current fragmented presence and reduced consistency. *Taxus baccata* has low resistance to intense cold and it probably survived mainly thanks to the ability of asexual reproduction and sex variations of adults in case of need. The data that we use was recorded in a study area located in central Italy, with specific reference to Abruzzo and Lazio regions. The area of interest extends for about 28000 Km^2, with an heterogenous morphology, which includes sandy coasts and the summits of the Apennines (the highest peak being the Gran Sasso, 2912 m of altitude). The forest habitat of *Taxus baccata* in these two regions is of high conservation priority in Europe (Scarnati et al. 2009a).

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

Climatic maps in GRID format, with a spatial resolution of 500 m, were built. These maps were obtained by interpolating precipitation and temperature data recorded in 300 meteorological stations and calculating the average data for the 1960–1990 period (see Attorre et al. (2007a) for technical details).

The environmental covariates considered were:

MIN_T_1 Minimum temperature of the coldest month (January)

MAX_T_7 Maximum temperature of the hottest month (July)

T_MED Average temperature in twelve consecutive months

$TOTAL\_P$ Total annual precipitation

$SUMM\_P$ Precipitation during summer

$WINT\_P$ Precipitation during winter

$MOISTURE$ Moisture index

ALT Altitude

Descriptive analysis of these are summarized in Table 4 for the entire grid, and in Table 5 for the plots in which a presence was recorded. Note that in Table 4 we report only on suitable locations for proliferation of *Taxus Baccata* (see below).

[Table 4 about here.]

[Table 5 about here.]

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

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

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

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

In order to reduce spurious effects, we repeat the pseudo-absence generation 40 times, and fit our model separately on each data set. At each repetition, we sample 97 pseudo-absences from the suitable sites with the case-control approach of Attorre et al. (2007b), select at random starting values for the parameters, and run Algorithm 1 for a total of 100000 sweeps. We allow a burn-in of 50000 iterations, and we use one each twentieth of the 50000 remaining iterations for posterior estimation. We perform model choice according to the structured stochastic search variable section (SSSVS) approach of Farcomeni (2010), to which we point the reader for details. We consider the possibility of including

any of the available covariates, plus all two-way interactions, in each part of the model. We use hierarchical constraints so that an interaction is not included in a model without both covariates contributing to it. SSSVS allows to estimate a probability of inclusion for each coefficient. As proved by Farcomeni (2010), consistency in model choice is achieved as long as covariates with a probability of inclusion larger than 50% are used in the model, and the other covariates are discarded.

Our 40 repetitions did not provide conflicting conclusions, so that probably we have observed no spurious effects in the sampled pseudo-absences. We provide results related to a single (randomly chosen) repetition.

In Table 6 we show the posterior means of each covariate included in the final model chosen with SSSVS, and individual probabilities of inclusion. All other covariates, including the interactions, have an estimated probability of inclusion smaller than 50%, and therefore are omitted from the final model. We remind that the parameter estimates should not be directly interpreted due to collinearity. In Table 7 we report the correlation matrix between the covariates included in the final model.

We compare our estimates with the maximum likelihood estimates obtained with an EM algorithm along the lines of Ward et al. (2009), assuming a known prevalence of 3%. It can be seen from Table 6 that final parameter estimates are comparable, especially with respect to the regression part of the model. The logistic part of the frequentist model is dependent on the assumptions related to the prevalence, which must be assumed known with the EM approach. Note further that the variance estimated with the maximum likelihood approach is slightly smaller than the posterior mean for the variance, and it has been observed in the simulation study that EM with known prevalence tends in fact to under estimate the variance of the continuous part of the model. Note finally that the posterior summaries for the prevalence parameters are essentially equivalent to the prior summaries, as the data contain very little information on prevalence.

[Table 6 about here.]

[Table 7 about here.]

We validate the predictive performance of our model by building $1-\alpha = 0.95$ predictive intervals for the observed presences. We finally obtain a prediction coverage probability of 0.948, so that we can claim the model valid from a predictive point of view. We have not experienced a strong prior sensitivity, and have obtained results equivalent for practical purposes by varying the prior assumptions in a reasonable range.

In Figure 1 we present a map of the potential distribution of the abundance of *Taxus baccata* built using GIS tools. The predictions in Figure 1 minimize the posterior expected loss.

[Figure 1 about here.]

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

The estimated potential distribution in Figure 1 leads us to conclude that *Taxus* is potentially situated at both a higher and lower altitude than expected. The first behavior (higher altitudes) is likely due to a retreating process to areas less accessible by livestock (for instance, cows). The second behavior (lower altitudes) has been seen in areas with a high moisture index (e.g., close to lakes in the Northwestern and Southwestern Lazio), which makes the area more suitable for a presence of *Taxus*.

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

We now focus on the locations corresponding to protected area (Special Protection Zone) ZPS12, in Monti Lepini, Lazio; established by European Community directive 79/409/CEE. We select the 363 locations corresponding to area ZPS12 and compatible with a presence of *Taxus*, and consider the posterior probability of observing a presence ($\Pr(IV > 0)$) and a moderately large abundance ($\Pr(IV > 2)$). Descriptive statistics for these probabilities computed at the 363 locations of special interest are reported in Table 8.

[Table 8 about here.]

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

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

6 DISCUSSION

In studies with presence-only data, a direct use of the observations may lead to biased predictions. A common solution is to add pseudo-absences to the sample but, especially if the species of interest is not rare, blindly using pseudo-absences as if they were true absences may obviously lead to pessimistic, and even biased, estimates. In this work we have introduced a model which takes into account uncertainty related to pseudo-absence sampling with abundance data.

We have restricted our attention to linear models for both the probability of observing a presence and for the density conditional on a positive abundance. Even if linear models have proved adequate for the data at hand, we point out that these are not the only choices, and that other more flexible choices for $\eta(\cdot)$ and $\eta'(\cdot)$ could be used (e.g., generalized additive models, Hastie and Tibshirani (1990)). The same reasoning applies to the parametric assumption on the continuous part of the model. In this paper we have in practice used a log-normal model, but other assumptions could be used (for instance, a Weibull model).

Modification of our approach for these different choices is often straightforward and does not usually lead to major modification of the inferential strategies described in Section 3.

ACKNOWLEDGEMENTS

Authors are grateful to an associate editor and a referee for detailed comments which substantially improved the paper. The authors are also grateful to Plant Biology Department of Sapienza - University of Rome, for permission to use the *Taxus baccata* data and to Prof. Giovanna Jona Lasinio for support.

A CASE-CONTROL ADJUSTMENT

Proposition 1. *Given the usual logistic model* $\text{logit}(\Pr(Y>0|X)=\eta(\boldsymbol{x})$ *we can use a case-control style adjustment to show:*

$$\text{logit}(\Pr(Y>0|X,s=1))=\eta(\boldsymbol{x})+\log\left(\frac{n_p+\xi n_u}{\xi n_u}\right) \tag{A.1}$$

where n_p *and* n_u *are respectively the number of observed abundance* $(Z>0)$ *and the number of pseudo-absence locations* $(Z=0)$.

Further, there is no need for adjustment for the linear regression part:

$$f(Y|Y>0,X,s=1)=f(Y|Y>0,X). \tag{A.2}$$

Proof. Following the usual case-control calculations presented in (McCullagh and Nelder 1989, p. 111) we define $\gamma_1=\Pr(s=1|Y>0)$ and $\gamma_0=\Pr(s=1|Y=0)$ the sampling rates of

our data from all true abundance and absences respectively. We assume that these sampling rates are independent of X, so in particular $\Pr(s=1|Y>0,X)=\Pr(s=1|Y>0)$.
An application of Bayes rule to $\Pr(Y>0|X,s=1)$ gives

$$\begin{aligned}\Pr(Y>0|X,s=1) &= \frac{\Pr(s=1|Y>0,X)\Pr(Y>0|X)}{\Pr(s=1|Y=0,X)\Pr(Y=0|X)+\Pr(s=1|Y>0,X)\Pr(Y>0|X)} = \\ &= \frac{\gamma_1 e^{\eta(\mathbf{x})}}{\gamma_0+\gamma_1 e^{\eta(\mathbf{x})}}.\end{aligned}$$

Hence,

$$\operatorname{logit}(\Pr(Y>0|X,s=1)) = \eta(\mathbf{x})+\log\left(\frac{\gamma_1}{\gamma_0}\right). \tag{A.3}$$

We now need only to derive an explicit expression for the second summand in (A.3).
The true number of positive abundances $(Y>0)$ in our presence-only sample is not known, but it is straightforward to see that the expected number is $n_p+\xi n_u$, that is, the number of samples for which $Z>0$ plus a proportion ξ of the number of samples for which $Z=0$. Hence,

$$\Pr(Y>0|s=1) = \frac{n_p+\xi n_u}{n_p+n_u}$$

and similarly

$$\Pr(Y=0|s=1) = \frac{(1-\xi)n_u}{n_p+n_u}.$$

An application of Bayes rule gives

$$\begin{aligned}\gamma_1 &= \Pr(s=1|Y>0) = \frac{\Pr(Y>0|s=1)\Pr(s=1)}{\Pr(Y>0)} \\ &= \frac{n_p+\xi n_u}{\xi(n_p+n_u)}\Pr(s=1)\end{aligned}$$

and

$$\begin{aligned}\gamma_0 &= \Pr(s=1|Y=0) = \frac{\Pr(Y=0|s=1)\Pr(s=1)}{\Pr(Y=0)} \\ &= \frac{(1-\xi)n_u}{(1-\xi)(n_p+n_u)}\Pr(s=1);\end{aligned}$$

which can be combined to see that

$$\begin{aligned}\log\left(\frac{\gamma_1}{\gamma_0}\right) &= \log\left(\frac{n_p+\xi n_u}{(1-\xi)n_u}\right)-\log\left(\frac{\xi}{1-\xi}\right) \\ &= \log\left(\frac{n_p+\xi n_u}{\xi n_u}\right).\end{aligned}$$

To see (A.2) once again we can use Bayes rule to see

$$\begin{aligned} f(Y|Y>0,s=1,X) &= \frac{\Pr(s=1|Y,Y>0,X)f(Y|Y>0,X)}{\Pr(s=1|Y>0,X)} = \\ &= \frac{\Pr(s=1|Y>0)f(Y|Y>0,X)}{\Pr(s=1|Y>0)} = f(Y|Y>0,X), \end{aligned}$$

since we assumed the sampling rates to be dependent only on the presence, but not on the actual value for the abundance. □

B DERIVATION OF OBSERVED LIKELIHOOD

Proposition 2. *Using a short-hand notation θ to denote the parameters at stake, the* observed likelihood $L(\theta|\mathbf{z},X)$ *for the presence-only data is given by:*

$$\prod_i \left[\frac{1+e^{\eta(\boldsymbol{x}_i)}}{1+\left(1+\frac{n_p}{\xi n_u}\right)e^{\eta(\boldsymbol{x}_i)}}\right]^{1_{\{z_i=0\}}} \left[\frac{\frac{n_p}{\xi n_u}e^{\eta(\boldsymbol{x}_i)}}{1+\left(1+\frac{n_p}{\xi n_u}\right)e^{\eta(\boldsymbol{x}_i)}}\frac{1}{z_i\sqrt{2\pi}\sigma}\exp\left\{-\frac{(\log(z_i)-\eta'(\tilde{\boldsymbol{x}}))^2}{2\sigma^2}\right\}\right]^{1_{\{z_i>0\}}} \tag{B.1}$$

where 1_C is the indicator function for condition C.

Proof. The observed likelihood for presence-only data is given by:

$$\begin{aligned} L(\theta|\mathbf{z},X) &= \prod_i \Pr(z_i=0|X_i,s_i=1)^{1_{\{z_i=0\}}} \\ &\quad [\Pr(z_i>0|X,s_i=1)f(z_i|z_i>0,X_i,s_i=1)]^{1_{\{z_i>0\}}} \end{aligned}$$

We derive an explicit expression first using a total probability argument across $Y>0$ and $Y=0$:

$$\begin{aligned} \Pr(Z>0|X,s=1) &= \Pr(Z>0|Y>0,X,s=1)\Pr(Y>0|X,s=1) \\ &+ \Pr(Z>0|Y=0,X,s=1)\Pr(Y=0|X,s=1). \end{aligned}$$

An application of Bayes rule and the fact that $Z>0$ is independent of X gives:

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

We can now mimick computations used for deriving the case control adjustment to see that the expected number of true presences $(Y>0)$ in our data is $n_p+\xi n_u$. Hence, $\Pr(Y>0|s=1)=(n_p+\xi n_u)/(n_p+n_u)$. Also, by definition of Z and Y, $\Pr(Z>0,Y>0|s=1)=n_p/(n_p+n_u)$. Consequently:

$$\Pr(Z>0|Y>0,X,s=1) = \frac{n_p}{n_p+\xi n_u}$$

Further, $\Pr(Z>0|Y=0,s=1)=0$ because all $Z>0$ in the data must occur for $Y>0$. Combining (A.1) with the latter expressions, after some manipulation, we get that

$$\Pr(Z>0|X,s=1)=0+\frac{\frac{n_p}{\xi n_u}e^{\eta(\mathbf{x})}}{1+\left(1+\frac{n_p}{\xi n_u}\right)e^{\eta(\mathbf{x})}}.$$

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

$$\begin{aligned}
L(\theta|\mathbf{z},X) &= \prod_i \Pr(z_i=0|X,s_i=1)^{1_{\{z_i=0\}}} \\
&\quad [\Pr(z_i>0|X,s_i=1)f(z_i|z_i>0,X,s_i=1)]^{1_{\{z_i>0\}}} \\
&= \prod_i \left[\frac{1+e^{\eta(\mathbf{x}_i)}}{1+\left(1+\frac{n_p}{\xi n_u}\right)e^{\eta(\mathbf{x}_i)}}\right]^{1_{\{z_i=0\}}} \\
&\quad \left[\frac{\frac{n_p}{\xi n_u}e^{\eta(\mathbf{x}_i)}}{1+\left(1+\frac{n_p}{\xi n_u}\right)e^{\eta(\mathbf{x}_i)})}f(z_i|z_i>0,X,s_i=1)\right]^{1_{\{z_i>0\}}},
\end{aligned}$$

and one can simply substitute $f(z_i|z_i>0,X,s_i=1)$ with its expression to see (2.4).

□

C DERIVATION OF COMPLETE LIKELIHOOD

Proposition 3. *The* complete likelihood $L(\theta|\mathbf{z},\mathbf{y},X)$ *in terms of both z and y, and letting* $\eta(\mathbf{x})=\beta' x$ *and* $\eta'(\tilde{\mathbf{x}})=\alpha'\tilde{x}$*, is given by:*

$$\begin{aligned}
&\prod_i \left[\frac{1}{1+\left(1+\frac{n_p}{\xi n_u}\right)\exp\{\beta' x_i\}}\right]^{1_{\{y_i=0\}}} \\
&\left[\frac{\exp\{\beta' x_i\}\left(1+\frac{n_p}{\xi n_u}\right)}{1+\left(1+\frac{n_p}{\xi n_u}\right)\exp\{\beta' x_i\}}\frac{1}{y_i\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}\left(\log(y_i)-\alpha'\tilde{x}_i\right)^2\right\}\right]^{1_{\{y_i>0\}}}
\end{aligned} \tag{C.1}$$

Proof. Using a conditioning argument, we get that

$$\begin{aligned}
L(\theta|\mathbf{z},\mathbf{y},X) &= \prod_i \Pr(y_i,z_i|X_i,s_i=1) \\
&= \prod_i \Pr(z_i|y_i,s_i=1,X_i)\Pr(y_i|X_i,s_i=1) \\
&= \prod_i \Pr(y_i|X_i,s_i=1) \\
&= \prod_i \Pr(y_i=0|X_i,s_i=1)^{1_{y_i=0}}[\Pr(y_i>0|X_i,s_i=1)f(y_i|y_i>0,X_i,s_i=1)]^{1_{y_i>0}}
\end{aligned} \tag{C.2}$$

The form of $\Pr(y_i|X,s_i=1)$ follows directly from (A.1):

$$\begin{aligned}\Pr(Y>0|X,s=1) &= \frac{e^{\eta(\mathbf{x})+\log\left(\frac{n_p+\xi n_u}{\xi n_u}\right)}}{1+e^{\eta(\mathbf{x})+\log\left(\frac{n_p+\xi n_u}{\xi n_u}\right)}} \\ &= \frac{e^{\eta(\mathbf{x})}\frac{n_p+\xi n_u}{\xi n_u}}{1+e^{\eta(\mathbf{x})}\frac{n_p+\xi n_u}{\xi n_u}} \\ &= \frac{e^{\eta(\mathbf{x})}\left(1+\frac{n_p}{\xi n_u}\right)}{1+\left(1+\frac{n_p}{\xi n_u}\right)e^{\eta(\mathbf{x})}} \\ &= \frac{\exp\{\beta' x_i\}\left(1+\frac{n_p}{\xi n_u}\right)}{1+\left(1+\frac{n_p}{\xi n_u}\right)\exp\{\beta' x_i\}} \end{aligned} \tag{C.3}$$

and

$$\begin{aligned}\Pr(Y=0|X,s=1) &= 1-\Pr(Y>0|X,s=1) \\ &= \frac{1}{1+\left(1+\frac{n_p}{\xi n_u}\right)\exp\{\beta' x_i\}} \end{aligned} \tag{C.4}$$

The density $f(y|y>0,X,s_i=1)$, due to (A.2), is given by:

$$f(y|y>0,X,s=1) = \frac{1}{y\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}\left(\log(y)-\alpha'\tilde{x}_i\right)^2\right\} \tag{C.5}$$

Then, the explicit form of the *complete likelihood* for the presence-only data, in terms of both z and y, can be obtained by substituting (C.3), (C.4) and (C.5) into (C.2).

□

REFERENCES

Attorre, F., Alfó, M., De Sanctis, M., Francesconi, F., and Bruno, F. (2007a), "Comparison of interpolation methods for mapping climatic and bioclimatic variables at regional scale," *Int J Climatol*, 27, 1825–1843.

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

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis*, Springer.

Bernardo, J. M. and Smith, A. F. M. (1994), *Bayesian Theory*, Chichester: Wiley.

Chaubert-Pereira, F., Guédon, Y., Lavergne, C., and Trottie, C. (2009), "Markov and semi-markov switching linear mixed models used to identify forest tree growth components," *Biometrics*, available online.

Diebolt, J. and Robert, C. (1994), "Estimation of finite mixture distributions through bayesian sampling," *Journal of the Royal Statistical Society, (Ser. B)*, 56, 363–375.

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

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

Farcomeni, A. (2010), "Bayesian constrained variable selection," *Statistica Sinica*, 20, 1043–1062.

Garthwaite, P., Kadane, J., and O'Hagan, A. (2005), "Statistical methods for eliciting probability distributions," Technical Report 808, Carnegie Mellon University.

Gelfand, A. E. (1996), "Model determination using sampling-based methods," in *Markov Chain Monte Carlo in Practice*, eds. W. Gilks, S. Richardson, and D. Spiegelhalter, London: Chapman & Hall, pp. 145–161.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992), "Model determination using predictive distributions with implementation via sampling-based methods," *Bayesian Statistics*, 4, 147–167.

Gilks, W. R., Best, N. G., and Tan, K. K. C. (1995), "Adaptive rejection Metropolis sampling within Gibbs sampling (corr: 97v46 p541-542 with R. M. Neal)," *Applied Statistics*, 44, 455–472.

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

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning*, New York: Springer.

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

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian model averaging: a tutorial," *Statistical Science*, 14, 382–417.

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

Kadane, J. B. and Wolfson, L. (1998), "Experiences in elicitation," *The statistician*, 47, 3–19.

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

Lachenbruch, P. (2002), "Analysis of data with excess zeros," *Statistical Methods in Medical Research*, 11, 297–302.

Leathwick, J., Moilanen, A., Francis, M., Elith, J., Taylor, P., Julian, K., Hastie, T., and Duffy, C. (2008), "Novel methods for the design and evaluation of marine protected areas in offshore waters," *Conservation Letters*, 1, 91–102.

Li, N., Elashoff, D. A., Robbins, W. A., and Xun, L. (2009), "A hierarchical zero-inflated log-normal model for skewed responses," *Statistical Methods in Medical Research*, available online.

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

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

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

Prasad, A. M., Iverson, L. R., and Liaw, A. (2006), "Newer classification and regression tree techniques: Bagging and random forests for ecological prediction," *Ecosystems*, 9, 181–199.

R Development Core Team (2009), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

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

Scarnati, L., Attorre, F., Farcomeni, A., Francesconi, F., and De Santis, M. (2009b), "Modelling the spatial distribution of tree species with fragmented populations from abundance data," *Community Ecology*, 10, 215–224.

Tanner, M. A. (1996), *Tools for Statistical Inference*, New York: Springer.

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

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

Zhou, X. and Tu, W. (1999), "Comparison of several different population means when their samples contain log-normal and possibly zero observations," *Biometrics*, 55, 645–651.

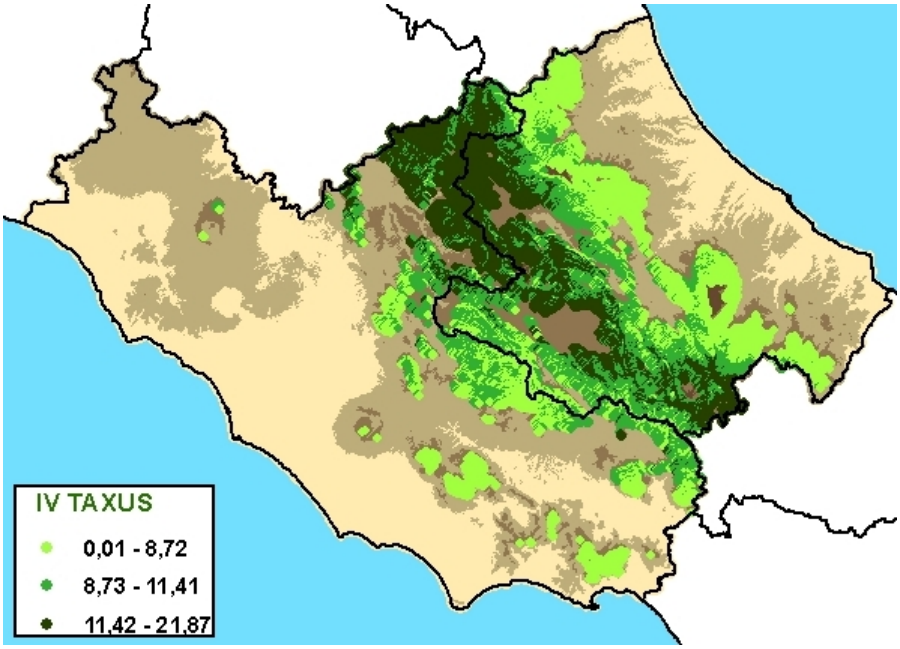

Figure 1. Potential distribution of the abundance of Taxus baccata*.* $R^2 = 0.18$*, False Negative Rate=0*

	Bayesian Model				EM algorithm			
	$N = 10^4$ $\lambda = 10\%$	$N = 10^4$ $\lambda = 30\%$	$N = 900$ $\lambda = 10\%$	$N = 900$ $\lambda = 30\%$	$N = 10^4$ $\lambda = 10\%$	$N = 10^4$ $\lambda = 30\%$	$N = 900$ $\lambda = 10\%$	$N = 900$ $\lambda = 30\%$
Parameters								
β_1	0.3668	0.3847	0.6028	0.6000	0.2910	0.2055	1.4512	1.4103
β_2	0.4056	0.4018	0.4772	0.4338	0.1078	0.0782	0.6281	0.5891
α_0	0.0012	0.0015	0.0129	0.0146	0.0012	0.0015	0.0130	0.0149
α_1	0.0012	0.0016	0.0136	0.0160	0.0012	0.0016	0.0136	0.0169
σ	0.0001	0.0001	0.0001	0.0001	0.0860	0.0871	0.0927	0.0859
ξ	0.0006	0.0008	0.0073	0.0169	-	-	-	-
95% CI ξ, L	0.1367	0.1361	0.1373	0.1424	-	-	-	-
95% CI ξ, C	1.0000	1.0000	1.0000	1.0000	-	-	-	-

Table 1. MSE of the parameter estimates of Bayesian model and EM model for different values of N and λ in simulated data. We omit β_0 since it is summarized in the final prevalence estimate. The last two lines report the mean length (L) and coverage (C) of the 95% CI for the prevalence parameter ξ. The number of replications is 1000.

Test	N	λ	Bayesian Model	EM model
MSE	10^4	10%	0.0000	0.0000
	10^4	30%	0.0000	0.0000
	900	10%	0.0002	0.0000
	900	30%	0.0002	0.0000
Sensitivity	10^4	10%	0.8401	0.8401
	10^4	30%	0.8402	0.8402
	900	10%	0.8408	0.8409
	900	30%	0.8303	0.8271
Specificity	10^4	10%	0.1600	0.1600
	10^4	30%	0.1603	0.1603
	900	10%	0.1597	0.1601
	900	30%	0.1673	0.1681

Table 2. MSE for positive predictions, sensitivity and specificity of the predicted presence/absence for different values of N and λ in simulated data. The results are averaged over 1000 replications.

	Prior settings			
Parameters	(a)	(b)	(c)	(d)
β_1	0.3668	0.3918	0.3974	0.3811
β_2	0.4056	0.4109	0.3289	0.3966
α_0	0.0012	0.0012	0.0012	0.0012
α_1	0.0012	0.0014	0.0010	0.0011
σ	0.0001	0.0001	0.0024	0.0001
ξ	0.0006	0.0006	0.0006	0.0006

Table 3. Sensitivity analysis: MSE obtained with (a) default priors, (b) biased priors, (c) biased prior on the prevalence parameter, (d) flat priors. The results are based on 1000 replications.

	Min	1*st* Quartile	Median	Mean	3*rd* Quartile	Max	Std. Dev.
MIN_T_1	-5.56	-3.00	-2.17	-1.99	-1.05	4.09	1.42
MAX_T_7	17.92	22.04	23.51	23.35	24.76	27.99	1.79
T_MED	5.09	7.82	8.94	8.87	9.95	13.34	1.42
TOTAL_P	629.00	1029.00	1189.00	1210.98	1403.00	1894.00	244.59
SUMM_P	91.00	145.00	165.00	170.16	191.00	292.00	33.45
WINT_P	153.00	304.00	363.00	374.99	447.00	706.00	95.06
MOISTURE	0.94	1.20	1.29	1.32	1.38	2.29	0.18
ALT	900.00	1035.00	1217.00	1243.68	1424.00	1750.00	234.75

Table 4. Descriptive statistics for the environmental covariates on the whole data

	Min	1*st* Quartile	Median	Mean	3*rd* Quartile	Max	Std. Dev.
MIN_T_1	-3.99	-2.57	-1.90	-1.79	-1.25	2.02	1.22
MAX_T_7	19.70	21.05	22.41	22.27	23.52	24.84	1.41
T_MED	5.90	7.56	8.38	8.45	9.04	11.61	1.16
TOTAL_P	22.00	1251.00	1405.00	1413.74	1620.00	1696.00	205.68
SUMM_P	141.00	161.00	203.00	196.09	219.00	254.00	31.19
WINT_P	256.00	336.00	440.00	429.27	507.00	560.00	86.17
MOISTURE	1.10	1.30	1.35	1.38	1.40	1.74	0.16
ALT	969.00	1278.00	1430.00	1392.15	1503.00	1715.00	157.19
ABUNDANCE	1.16	7.00	12.00	20.11	30.00	78.00	17.82

Table 5. Descriptive statistics for the environmental covariates for locations where abundance is positive.

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

Table 6. Posterior mean, estimated standard error and probability of inclusion for each covariate included in the final model after SSSVS; plus maximum likelihood estimates obtained with EM algorithm for comparison.

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

Table 7. Correlation between covariates used in the final model.

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

Table 8. Descriptive statistics for the 363 posterior estimated probabilities of a positive and of a moderately large abundance in the Special Protection Zone ZPS12.