Bayesian Analysis (2017) **12**, Number 3, pp. 855–877

Intrinsic Bayesian Analysis for Occupancy Models

Daniel Taylor-Rodríguez*, Andrew J. Womack†, Claudio Fuentes‡, and Nikolay Bliznyuk§

Abstract. Occupancy models are typically used to determine the probability of a species being present at a given site while accounting for imperfect detection. The survey data underlying these models often include information on several predictors that could potentially characterize habitat suitability and species detectability. Because these variables might not all be relevant, model selection techniques are necessary in this context. In practice, model selection is performed using the Akaike Information Criterion (AIC), as few other alternatives are available. This paper builds an objective Bayesian variable selection framework for occupancy models through the intrinsic prior methodology. The procedure incorporates priors on the model space that account for test multiplicity and respect the polynomial hierarchy of the predictors when higher-order terms are considered. The methodology is implemented using a stochastic search algorithm that is able to thoroughly explore large spaces of occupancy models. The proposed strategy is entirely automatic and provides control of false positives without sacrificing the discovery of truly meaningful covariates. The performance of the method is evaluated and compared to AIC through a simulation study. The method is illustrated on two datasets previously studied in the literature.

Keywords: imperfect detection, intrinsic priors, model priors, strong heredity, Bayesian variable selection, AIC.

1 Introduction

It is often the case that measurements recorded for a given response are, at best, a noisy version of the variable of interest. A particular case of this issue is known as imperfect detection, and constitutes a pervasive problem. For instance, in biological surveys subject to imperfect detection, "presence/absence" data for a given species actually become "detection/non-detection" data because a species may be present at a given site yet be undetected in a survey. Ignoring imperfect detection may lead to inaccurate measurement of the presences (Guillera-Arroita et al., 2014), which generally results in biased parameter estimates (MacKenzie et al., 2002).

*First coauthor, Postdoctoral Fellow, SAMSI/Duke University, Research Triangle Park, NC 27709, dt108@stat.duke.edu

†First coauthor, Assistant Professor, Department of Statistics, Indiana University, Bloomington, IN 47408, ajwomack@indiana.edu

‡Assistant Professor, Department of Statistics, Oregon State University, Corvallis, OR 97331, fuentesc@stat.oregonstate.edu

§Corresponding author. Assistant Professor, Departments of Agricultural and Biological Engineering, Biostatistics and Statistics, University of Florida, Gainesville, FL 32611, nbliznyuk@ufl.edu

 DOI: 10.1214/16-BA1014

As defined in the ecological literature, *occupancy* is the proportion of sites where a target species is present, constituting a state variable instrumental to assess the distribution of species (MacKenzie et al., 2002). Over the past decade, site occupancy models have been the main tool used by ecologists to estimate occupancy while accounting for imperfect detection. Occupancy models describe the observed data by linking two processes: presence and detection. Occupancy models adapt the conventional binary regression model to produce separate estimates for presence and detection probabilities (Dorazio and Taylor-Rodríguez, 2012). This separation is possible by surveying repeatedly the sampling locations, which provides additional information to better assess if non-detection of the species truly corresponds to its absence. Conveniently, these models can be fitted even if the number of surveys is unbalanced across sites. The core of the occupancy model is characterized by the hierarchy

$$\begin{aligned} y_{ij}|z_i &\sim \text{Bern}(z_i p_{ij}) \\ z_i &\sim \text{Bern}(\psi_i), \end{aligned} \tag{1}$$

where y_{ij} is the binary detection indicator at the ith site ($i = 1, \ldots, N$) during the jth survey ($j = 1, \ldots, J_i$). The detection probability for event $\{y_{ij} = 1\}$ is p_{ij} whenever the species is present; and z_i is the presence indicator at the ith site with success probability ψ_i. Note that the z_i are imperfectly observed. At site i, whenever the vector of detections $\mathbf{y}_i \neq \mathbf{0}$, then we know that $z_i = 1$, but $\mathbf{y}_i = \mathbf{0}$ does not imply that $z_i = 0$. To produce estimates of ψ_i and p_{ij}, site occupancy surveys collect information on several predictors with the potential to influence habitat suitability (characterizing ψ_i) and species detectability (characterizing p_{ij}). Given that some of the collected predictors may be uninformative or redundant, variable selection techniques are instrumental in identifying good models.

In this paper, we propose an objective Bayesian variable selection procedure for occupancy models. Our approach is based on intrinsic, objective priors for the model parameters. Additionally, we build priors over the model space that simultaneously account for test multiplicity, and, if interactions and/or polynomial terms are considered, respect the polynomial hierarchical structure among predictors.

Currently, variable selection procedures for occupancy models implemented in statistical software are mainly based on the Akaike Information Criterion (AIC) (Akaike, 1983). As a consequence, these procedures do not allow for valid post-selection inference and uncertainty quantification, and typically require enumerating and fitting every possible model in the space of models under consideration (e.g., Mazerolle and Mazerolle, 2013; Fiske and Chandler, 2011). In practice, this enumeration is feasible only if the model space is small enough, either because substantial knowledge about the underlying ecological processes is available to constrain the model space, or because only a few variables are considered to begin with. Nevertheless, many site occupancy surveys collect large amounts of covariate information about the sampled sites, and since the total number of candidate models grows exponentially in the number of predictors, choosing a reduced set of models based on ecological intuition becomes increasingly difficult.

The AIC is designed to find the model that is the closest to the true (unknown) model with respect to Kullback–Leibler divergence, identifying as good models those with

smaller AIC values. It has been shown, however, that the AIC has certain limitations as a model selection criterion. For instance, if nested models are being considered, the AIC will not necessarily select the true model (Wasserman, 2000). In fact, the AIC generally shows a weak signal-to-noise ratio and tends to prefer more complex models, even if the true model is available (Rao and Wu, 2001). Other versions of the AIC address this issue by including a bias correction factor that enhances the signal-to-noise ratio (see Hurvich and Tsai, 1989; McQuarrie et al., 1997); however, these modified versions cannot be used with occupancy models, as they depend on the effective sample size, which is unknown for these models.

In this context, Bayesian methods are more appealing. Under regularity conditions, when the true model is contained in a fixed model space, its posterior probability converges to one as the number of sites and surveys per site both increase. In addition, if the true model is not contained in the model space, the posterior probability of the most parsimonious model closest to the true data generating process tends asymptotically to one. In the finite sample context, Bayesian methods allow for full and faithful error propagation. Furthermore, the Bayesian machinery provides the means to conduct valid inference accounting for model uncertainty.

A Bayesian selection procedure for occupancy models was described in Hooten and Hobbs (2015). However, their implementation uses informative prior distributions on the model parameters, tailored specifically to the example discussed in the paper, which prevents the approach from being applicable to occupancy problems in general. It is often the case that subjective elicitation of parameter and model prior distributions is not possible, since neither the relationship between the response and the predictors, nor the advantages of one model over another, are clearly understood. In addition, the use of seemingly innocuous subjective priors may drastically affect outcomes. This has been a recurring argument in favor of objective Bayesian procedures, which appeal to the use of formal rules to build parameter priors that incorporate the structural information inside the likelihood while utilizing some objective criterion (Kass and Wasserman, 1996).

To the best of our knowledge, the method proposed in this article is the first general Bayesian selection procedure for occupancy models, that

1. bypasses the need for hyper-parameter tuning,
2. uses priors specifically designed for testing,
3. controls for test multiplicity, and
4. accounts for the hierarchical polynomial structure in the predictors.

In building our approach, we first derive intrinsic priors (Berger and Pericchi, 1996; Moreno et al., 1998) for the model parameters in both the presence and detection components of the single-season occupancy model. For the model priors, we consider the ones proposed in Taylor-Rodriguez et al. (2016). These priors, in addition to controlling for test multiplicity, allow restricting the model space to the set of models that respect

(weakly or strongly) the polynomial hierarchy among the predictors whenever interactions and higher-order terms are considered. As discussed in Peixoto (1987, 1990) when covariate interactions and polynomial terms are present, failure to restrict the class of models to those respecting strong heredity may result in incoherent variable selection. This is because the model design matrices are not invariant to linear transformations of order-one predictors (e.g., recentering of the main effect variables). Using the derived intrinsic priors on the parameter space and the multiplicity correction priors on the model space, we build a fast stochastic search algorithm that allows us to thoroughly explore large spaces for the single-season occupancy model framework. This strategy is completely automatic, avoiding the need for both tuning parameters in the sampling algorithm and subjective elicitation of parameter prior distributions. Furthermore, as any other Bayesian approach, it naturally enables parameter and model uncertainty quantification.

The outline of the paper is as follows: in Section 2, we provide background on occupancy models and set notation. In Section 3, we introduce our objective Bayesian model selection method and develop the Gibbs sampler. In Section 4, we present results from a simulation study and a comparison with selection using AIC. In Section 5, we illustrate our methodology on two datasets, which have been previously examined in the ecological literature (Kéry et al., 2005; Kery et al., 2010; Dorazio and Taylor-Rodríguez, 2012). We conclude the paper with a brief discussion. Code for all the tools proposed is available in the R package OccOBayes. A description of the stochastic search algorithm is included in the Supplementary Appendix (Taylor-Rodríguez et al., 2016).

2 Inference for a single model

This section briefly describes the estimation procedure for a single model. Assuming the probit link, the occupancy model can be characterized in terms of latent variables, which in turn allows one to relate the detection and occupancy probabilities to predictors. We build an objective prior distribution for the regression coefficients using the expected posterior prior framework (Pérez and Berger, 2002) where we condition on both the observed data as well as the unobserved latent variables (Leon-Novelo et al., 2012).

2.1 The occupancy model with Probit link

The occupancy model in (1) is completed in two ways. First, the probabilities for detection p_{ij} and for presence ψ_i are linked to vectors of predictors $\mathbf{q}_{ij}$ and $\mathbf{x}_i$, respectively, through appropriate link functions, $g_p(p_{ij}) = \mathbf{q}'_{ij}\boldsymbol{\lambda}$ and $g_\psi(\psi_i) = \mathbf{x}'_i\boldsymbol{\alpha}$. We assume that the link function is the inverse standard normal cdf, leading to probit models. Other binary regression models can be fit and lead to slightly more complicated computational algorithms. Second, the parameters of the underlying space, here $(\boldsymbol{\alpha}, \boldsymbol{\lambda})$, are given a prior distribution $\pi(\boldsymbol{\alpha}, \boldsymbol{\lambda})$. This paper proposes a prior distribution building on the expected posterior prior method of Leon-Novelo et al. (2012).

Letting $\mathbf{X}$ and $\mathbf{Q}$ be the matrices whose rows are, respectively, vectors $\mathbf{x}'_i$ and $\mathbf{q}'_{ij}$ for $i = 1, \ldots, N$ and $j = 1, \ldots, J_i$, the Bayesian probit occupancy model is specified as

$$\begin{aligned} y_{ij}|z_i, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \mathbf{Q}, \mathbf{X} &\sim \text{Bern}(z_i p_{ij}) \quad \text{with} \quad p_{ij} = \Phi\left(\mathbf{q}_{ij}'\boldsymbol{\lambda}\right) \\ z_i|\boldsymbol{\alpha}, \boldsymbol{\lambda}, \mathbf{Q}, \mathbf{X} &\sim \text{Bern}(\psi_i) \quad \text{with} \quad \psi_i = \Phi\left(\mathbf{x}_i'\boldsymbol{\alpha}\right) \\ \boldsymbol{\alpha}, \boldsymbol{\lambda}|\mathbf{Q}, \mathbf{X} &\sim \pi, \end{aligned} \tag{2}$$

where Φ is the standard normal cdf. As it will be made evident subsequently, we explicitly condition on $\mathbf{X}$ and $\mathbf{Q}$ since the priors devised for the model parameters depend on these design matrices. Again, note that the z_i are not perfectly observed. The sites with $\mathbf{y}_i = \mathbf{0}$ provide no detections but this does not necessarily imply a lack of presence. Thus, the model is a zero-inflated binary regression model where both lack of presence and individual instances of detection are predicted with covariates. The observed data vectors for the sites, $\mathbf{y}_1, \ldots, \mathbf{y}_n$, are independent given $(\boldsymbol{\alpha}, \boldsymbol{\lambda})$ and

$$\begin{aligned} p(\mathbf{y}_i|\boldsymbol{\alpha}, \boldsymbol{\lambda}, \mathbf{Q}, \mathbf{X}) = &\left(\Phi\left(\mathbf{x}_i'\boldsymbol{\alpha}\right) \prod_j \Phi\left(\mathbf{q}_{ij}'\boldsymbol{\lambda}\right)^{y_{ij}} \left(1 - \Phi\left(\mathbf{q}_{ij}'\boldsymbol{\lambda}\right)\right)^{1-y_{ij}} \right)^{\mathcal{I}_{\{\mathbf{y}_i \neq \mathbf{0}\}}} \\ &\times \left(\Phi\left(\mathbf{x}_i'\boldsymbol{\alpha}\right) \prod_j \left(1 - \Phi\left(\mathbf{q}_{ij}'\boldsymbol{\lambda}\right)\right) + \left(1 - \Phi\left(\mathbf{x}_i'\boldsymbol{\alpha}\right)\right) \right)^{\mathcal{I}_{\{\mathbf{y}_i = \mathbf{0}\}}}. \end{aligned}$$

The model can be expanded in the spirit of Albert and Chib (1993) by introducing latent variables at each level. Let v_i be the underlying continuous latent variable for presence at site i and w_{ij} be the underlying continuous latent variable for detection during survey j from site i. The hierarchical model in (2) becomes

$$\begin{aligned} y_{ij}|z_i, v_i, w_{ij}, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \mathbf{Q}, \mathbf{X} &= z_i \mathcal{I}_{\{w_{ij} > 0\}} \\ w_{ij}|z_i, v_i, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \mathbf{Q}, \mathbf{X} &\sim N\left(\mathbf{q}_{ij}'\boldsymbol{\lambda}, 1\right) \\ z_i|v_i, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \mathbf{Q}, \mathbf{X} &= \mathcal{I}_{\{v_i > 0\}} \\ v_i|\boldsymbol{\alpha}, \boldsymbol{\lambda}, \mathbf{Q}, \mathbf{X} &\sim N\left(\mathbf{x}_i'\boldsymbol{\alpha}, 1\right) \\ \boldsymbol{\alpha}, \boldsymbol{\lambda}|\mathbf{Q}, \mathbf{X} &\sim \pi. \end{aligned} \tag{3}$$

When one uses a multivariate normal prior for $(\boldsymbol{\alpha}, \boldsymbol{\lambda})$, the model in (3) can be fit using a Gibbs sampler. As described in Dorazio and Taylor-Rodríguez (2012), the only complication in using a Gibbs sampler is the fact that the sign of v_i determines the value of z_i and so the Gibbs sampler has to proceed in two blocks. The first block, which corresponds to a multivariate normal draw, is $(\boldsymbol{\alpha}, \boldsymbol{\lambda}|\mathbf{z}, \mathbf{v}, \mathbf{w}, \mathbf{y}, \mathbf{Q}, \mathbf{X})$. The second block is $(\mathbf{v}, \mathbf{w}, \mathbf{z}|\boldsymbol{\alpha}, \boldsymbol{\lambda}, \mathbf{y}, \mathbf{Q}, \mathbf{X})$. Each z_i is drawn from the distribution $[z_i|\boldsymbol{\alpha}, \boldsymbol{\lambda}, \mathbf{y}, \mathbf{Q}, \mathbf{X}]$, which is a Bernoulli distribution with probability of success

$$\xi_i = \mathcal{I}_{\{\mathbf{y}_i \neq \mathbf{0}\}} + \frac{\Phi\left(\mathbf{x}_i'\boldsymbol{\alpha}\right) \prod_j \left(1 - \Phi\left(\mathbf{q}_{ij}'\boldsymbol{\lambda}\right)\right)}{\Phi\left(\mathbf{x}_i'\boldsymbol{\alpha}\right) \prod_j \left(1 - \Phi\left(\mathbf{q}_{ij}'\boldsymbol{\lambda}\right)\right) + 1 - \Phi\left(\mathbf{x}_i'\boldsymbol{\alpha}\right)} \mathcal{I}_{\{\mathbf{y}_i = \mathbf{0}\}},$$

and the v_i and w_{ij} are sampled independently from their full conditionals. Each v_i has a truncated normal distribution with mean $\mathbf{x}_i'\boldsymbol{\alpha}$ and variance 1, restricted to the positive real line when $z_i = 1$ and to the negative real line when $z_i = 0$. Each w_{ij} has a truncated

normal distribution with mean $\mathbf{q}'_{ij}\boldsymbol{\lambda}$ and variance 1 that is supported on the positive real line when $z_i y_{ij} = 1$, the negative real line when $z_i(1 - y_{ij}) = 1$, and the whole real line when $z_i = 0$.

The marginal $p(\mathbf{y}|\mathbf{X}, \mathbf{Q})$ for the observed data can be estimated using the output from the Gibbs sampler (Chib, 1995). In this sampling scheme, one can also perform parameter expansions for both $\mathbf{v}$ and $\mathbf{w}$ (Liu and Wu, 1999). These dramatically decrease the autocorrelation between successive samples and reduce the asymptotic variance of estimators (Roy and Hobert, 2007).

Alternatively, one can perform inference for the model specified in (2) directly using a Metropolis-Hastings algorithm (e.g., an independence chain, a random walk, or Hamiltonian Monte Carlo). The output of the Metropolis Hastings algorithm can be used to estimate the marginal of the observed data using the method outlined in Chib and Jeliazkov (2001). When the sample size is large, an independence chain, using the Laplace approximation to the posterior as a proposal density, provides accurate numerical estimates of the posterior evaluated at its mode in a relatively small number of samples.

2.2 An objective prior for $(\boldsymbol{\alpha}, \boldsymbol{\lambda})$

Intrinsic priors, as defined by Moreno et al. (1998), are an example of expected posterior priors (Pérez and Berger, 2002). Concisely, an expected posterior prior for parameter $\boldsymbol{\theta}$ with prior π_M under a model M is given by

$$\pi_M^E(\boldsymbol{\theta}|\pi_M, m_0) = \int p_M(\boldsymbol{\theta}|D, \pi_M) m_0(D) \mathrm{d}D,$$

where D is some imaginary data that is integrated out, $p_M(\boldsymbol{\theta}|D, \pi_M)$ is the posterior of $\boldsymbol{\theta}$ given data D under the model M with parameter prior π_M, and m_0 is a fixed distribution for generating the data D. The properties of the data D are determined by the investigator. For regression problems, this amounts to determining the number of samples in the response and the associated design matrix. The generating model m_0 for the data D is usually taken to be a simple model, for instance an intercept-only model. Thus, the expected posterior prior under model M is calibrated to the distribution m_0.

Consider the context of multiple models, $M_0, M_1, \ldots, , M_K$, where M_0 is nested in M_k for all k and model M_k has parameter $\boldsymbol{\theta}_k$ with non-informative (often improper) prior π_k^N. In this context, M_0 is referred to as the base model. The intrinsic prior for each model is computed as

$$\pi_{M_k}^{IP}(\boldsymbol{\theta}_k|\pi_k^N, m_0^N) = \int p_{M_k}^N(\boldsymbol{\theta}_k|D_k, \pi_k^N) m_0^N(D_k) \mathrm{d}D_k,$$

where D_k is a training sample for model M_k and m_0^N is the marginal density for D_k under the base model. For the intrinsic prior, D_k is taken to be a minimal training sample for model M_k under the prior $\pi_{M_k}^N$, which is a dataset of the smallest possible size that provides a proper posterior for $p_{M_k}^N(\boldsymbol{\theta}_k|D_k, \pi_k^N)$. Of course, the intrinsic prior

for the base model is just its original non-informative prior. When the prior for model M_k is improper and only defined up to a multiplicative constant c_k, the intrinsic prior framework removes the ambiguity of these constants and each intrinsic prior is defined up to a common multiplicative constant c_0.

An extension of the intrinsic prior framework is to have the datasets D_k include both observable and unobservable latent variables. Leon-Novelo et al. (2012) used this approach in computing an objective prior for standard probit regression. There, the authors conditioned on both the observed binary data as well as the unobserved continuous latent variables. Following their development, we form an objective prior for the occupancy model by conditioning on the unobserved latent presence variables ($\mathbf{z}$) as well as the unobserved continuous latent variables for both presence and detection ($\mathbf{v}, \mathbf{w}$). We refer to this objective prior as an intrinsic prior though its derivation differs from that in Moreno et al. (1998) and Berger and Pericchi (1996).

Specifically, let $\mathbf{X}_0$ and $\mathbf{Q}_0$ be design matrices for presence and detection in the model M_0 and let $\mathbf{X}$ and $\mathbf{Q}$ be design matrices for a model M that nests M_0. Let $(\boldsymbol{\alpha}, \boldsymbol{\lambda})$ and $(\boldsymbol{\alpha}_0, \boldsymbol{\lambda}_0)$ be the parameters of M and M_0, respectively. Further, assume that the prior distributions for the parameters under each model are constant, $\pi_M^N = c_M$ and $\pi_0^N = c_0$. The intrinsic prior for $(\boldsymbol{\alpha}, \boldsymbol{\lambda})$ is given by

$$\pi_M^{IP}(\boldsymbol{\alpha}, \boldsymbol{\lambda} | \tilde{\mathbf{Q}}, \tilde{\mathbf{X}}) = \sum_{\tilde{\mathbf{z}}, \tilde{\mathbf{y}}} \iint p_M^N(\boldsymbol{\alpha}, \boldsymbol{\lambda} | \tilde{\mathbf{z}}, \tilde{\mathbf{v}}, \tilde{\mathbf{w}}, \tilde{\mathbf{y}}, \tilde{\mathbf{Q}}, \tilde{\mathbf{X}}) m_0^N(\tilde{\mathbf{z}}, \tilde{\mathbf{v}}, \tilde{\mathbf{w}}, \tilde{\mathbf{y}} | \tilde{\mathbf{Q}}_0, \tilde{\mathbf{X}}_0) \mathrm{d}\tilde{\mathbf{v}}\, \mathrm{d}\tilde{\mathbf{w}}, \quad (4)$$

where the "$\sim$" over variables indicates that these correspond to the training sample that is to be integrated out. The formula in (4) is greatly simplified by the fact that, under the non-informative prior, $(\boldsymbol{\alpha}, \boldsymbol{\lambda})$ are conditionally independent of $(\tilde{\mathbf{z}}, \tilde{\mathbf{y}})$ given the continuous latents $(\tilde{\mathbf{v}}, \tilde{\mathbf{w}})$. Moreover, the $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$ are conditionally independent of each other given the continuous latents. Thus, (4) simplifies to

$$\begin{aligned} \pi_M^{IP}(\boldsymbol{\alpha}, \boldsymbol{\lambda} | \tilde{\mathbf{Q}}, \tilde{\mathbf{X}}) &= \iint p_M^N(\boldsymbol{\alpha} | \tilde{\mathbf{v}}, \tilde{\mathbf{w}}, \tilde{\mathbf{Q}}, \tilde{\mathbf{X}}) p_M^N(\boldsymbol{\lambda} | \tilde{\mathbf{v}}, \tilde{\mathbf{w}}, \tilde{\mathbf{Q}}, \tilde{\mathbf{X}}) m_0^N(\tilde{\mathbf{v}}, \tilde{\mathbf{w}} | \tilde{\mathbf{Q}}_0, \tilde{\mathbf{X}}_0) \mathrm{d}\tilde{\mathbf{v}}\, \mathrm{d}\tilde{\mathbf{w}} \\ &= \int p_M^N(\boldsymbol{\alpha} | \tilde{\mathbf{v}}, \tilde{\mathbf{X}}) m_0^N(\tilde{\mathbf{v}} | \tilde{\mathbf{X}}_0) \mathrm{d}\tilde{\mathbf{v}} \times \int p_M^N(\boldsymbol{\lambda} | \tilde{\mathbf{w}}, \tilde{\mathbf{Q}}) m_0^N(\tilde{\mathbf{w}} | \tilde{\mathbf{Q}}_0) \mathrm{d}\tilde{\mathbf{w}}, \quad (5) \end{aligned}$$

where the last equality follows from the assumptions of (3) and the prior independence of $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$ under π_M^N. Both of the integrals in (5) are of the form of the integrals in Leon-Novelo et al. (2012). Thus, the intrinsic prior is given by a product of singular normal distributions.

The explication of these priors is greatly aided by the introduction of additional notation. Because M_0 is nested in M, we can write $\mathbf{X} = (\mathbf{X}_0 \quad \mathbf{X}_A)$ and $\mathbf{Q} = (\mathbf{Q}_0 \quad \mathbf{Q}_A)$ and can do the same for the design matrices for the minimal training sample. Similarly, we can write $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_0', \boldsymbol{\alpha}_A')'$ and $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_0', \boldsymbol{\lambda}_A')'$. The intrinsic prior is given by

$$\boldsymbol{\alpha}_A | \boldsymbol{\alpha}_0, \tilde{\mathbf{X}} \sim \mathcal{N}\left(\mathbf{0}, 2\left(\tilde{\mathbf{X}}_A'\left(\mathbf{I} - \tilde{\mathbf{H}}_{0_z}\right)\tilde{\mathbf{X}}_A\right)^{-1}\right) \quad (6)$$

$$\boldsymbol{\lambda}_A | \boldsymbol{\lambda}_0, \tilde{\mathbf{Q}} \sim \mathcal{N}\left(\mathbf{0}, 2\left(\tilde{\mathbf{Q}}_A'\left(\mathbf{I} - \tilde{\mathbf{H}}_{0_y}\right)\tilde{\mathbf{Q}}_A\right)^{-1}\right) \quad (7)$$

$$\boldsymbol{\lambda}_0, \boldsymbol{\alpha}_0 | \tilde{\mathbf{X}}, \tilde{\mathbf{Q}} \quad \sim \quad c_0 \times d_0 \tag{8}$$

where $\tilde{\mathbf{H}}_{0_z}$ and $\tilde{\mathbf{H}}_{0_y}$ are the hat matrices associated to $\tilde{\mathbf{X}}_0$ and $\tilde{\mathbf{Q}}_0$, respectively. Here we include two undefined constants c_0 and d_0 for the reference prior of the base model, corresponding to the flat priors for $\boldsymbol{\alpha}_0$ and $\boldsymbol{\lambda}_0$, respectively.

The only remaining task for this intrinsic prior is to define the design matrices for the minimal training samples. Letting $p_{\boldsymbol{\alpha}} = \dim(\boldsymbol{\alpha})$ and $p_{\boldsymbol{\lambda}} = \dim(\boldsymbol{\lambda})$, the minimal training samples for $\mathbf{v}$ and $\mathbf{w}$ contain $p_{\boldsymbol{\alpha}}$ and $p_{\boldsymbol{\lambda}}$ samples, respectively. Following Leon-Novelo et al. (2012) and Casella and Moreno (2006), we define $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Q}}$ to be any design matrices of dimensions $p_{\boldsymbol{\alpha}} \times p_{\boldsymbol{\alpha}}$ and $p_{\boldsymbol{\lambda}} \times p_{\boldsymbol{\lambda}}$ satisfying

$$\tilde{\mathbf{X}}'\tilde{\mathbf{X}} = \frac{p_{\boldsymbol{\alpha}}}{N}\mathbf{X}'\mathbf{X} \qquad \text{and} \qquad \tilde{\mathbf{Q}}'\tilde{\mathbf{Q}} = \frac{p_{\boldsymbol{\lambda}}}{J_\bullet}\mathbf{Q}'\mathbf{Q}, \tag{9}$$

where N is the number of sites and $J_\bullet = \sum_{i=1}^{N} J_i$ is the total number of surveys. Note that the covariance matrices in (6) and (7) are thus completely determined by $\mathbf{X}'\mathbf{X}$ and $\mathbf{Q}'\mathbf{Q}$.

3 The variable selection problem

The hierarchy in Equation (3) is given for a specific model with a fixed set of predictors. This section develops the model selection problem for occupancy models. Each model contains two components, one for presence and one for detection. Thus, model M is decomposed as $M = (M_y, M_z)$, where M_y is a component model for detection and M_z is a component model for presence. The base model is $M_0 = \left(M_{0_y}, M_{0_z}\right)$, where the component base model design matrices contain at least a column of ones for the intercept. Each model M is assumed to nest M_0 and the prior for model M is taken to be the intrinsic prior defined in (6)–(8). The largest model is denoted by $M_F = \left(M_{F_y}, M_{F_z}\right)$ and contains the largest possible component models for detection and presence. The design matrices for these full components are $\mathbf{X}_F$ and $\mathbf{Q}_F$.

Let $K = (K_y, K_z)$, where K_y and K_z denote the sets of column indices for $\mathbf{Q}_F$ and $\mathbf{X}_F$ that are not in $\mathbf{Q}_0$ and $\mathbf{X}_0$, respectively. The model space can then be represented by the Cartesian product $\mathcal{P}(K_y) \times \mathcal{P}(K_z)$, where $\mathcal{P}(B)$ is the powerset of B. A specific model is represented by $A = (A_y, A_z)$ with $A_y \subseteq K_y$ and $A_z \subseteq K_z$. Thus, the entire model space $\mathcal{M}$ is populated by models of the form $M_A = \left(M_{A_y}, M_{A_z}\right)$, where M_{A_y} and M_{A_z} are the corresponding component models for detection and presence determined by the base covariates as well as covariates with indices in A_y and A_z, respectively. It follows that for the presence process $\mathbf{z}$, the design matrix for the model M_A is of the form $\mathbf{X}_{M_A} = (\mathbf{X}_0 \quad \mathbf{X}_A)$, where $\mathbf{X}_0$ is the design matrix of the base model M_{0_z} and $\mathbf{X}_A$ is the matrix containing the covariates indexed by A_z (and similarly for $\mathbf{Q}_{M_A} = (\mathbf{Q}_0 \quad \mathbf{Q}_A)$). Denote the regression coefficients of the model M_A by $\boldsymbol{\alpha}_{M_A} = (\boldsymbol{\alpha}_0', \boldsymbol{\alpha}_A')'$ and $\boldsymbol{\lambda}_{M_A} = (\boldsymbol{\lambda}_0', \boldsymbol{\lambda}_A')'$ for presence and detection, respectively.

It is important to note that this construction using the Cartesian product provides the largest possible model space for the occupancy model given the structures of the base and full models. Investigators may wish to impose additional model space restrictions based upon their (subjective) judgment. One means of achieving this restriction is to

form two sets of models, $\mathcal{M}_y$ for detection and $\mathcal{M}_z$ for presence. The model space can then be defined by the Cartesian product, $\mathcal{M} = \mathcal{M}_y \times \mathcal{M}_z$. One particular example of such a restriction arises when higher-order terms are included in the detection or presence models. Heredity conditions (Chipman, 1996) can be imposed on either model space and appropriate priors defined (see Section 3.1).

3.1 Priors over the space of models

Here we outline the construction of prior distributions over the model space. To allow for flexible modeling, it is assumed that the sets of covariates can potentially include interaction effects, higher-order polynomial terms, and factor variables. The priors for either the presence or the detection component have the same structure, and the joint prior is the product of marginal priors of the two model components.

The priors placed on the model space for the presence and detection models respect the hierarchy of the terms that could be included in a given model. Aspects of the prior construction are described here and full details on such priors can be found in Taylor-Rodriguez et al. (2016). The full model for either the presence or the detection component is represented as a directed acyclic graph (DAG) with nodes representing polynomial terms (powers or interactions; e.g., x_1 or x_1^2 or $x_1x_2^2$) and with edges specifying inheritance relationships. For example, $x_1x_2^2$ has edges (inherits) from its parent nodes x_1x_2 and x_2^2, also x_1^2 inherits from its parent x_1 but not from x_2. Feasible models, also known as models obeying weak heredity, correspond to a special kind of connected subgraph of the full model DAG. First, they must include the base model DAG. Second, a node η can only be included in a model's subgraph only if there is a directed path from a node in the base model to η. The priors considered here focus on models satisfying strong heredity (also known as well-formulated models), which amounts to requiring that for each node η in a model's subgraph, all parents of η included in the model's subgraph.

Model prior probabilities are specified recursively via conditional node inclusion probabilities (given the parent DAG) using a type of Markov condition reflected in the principles of conditional independence and immediate inheritance (Chipman, 1996). Conditional node inclusion is identified with a latent Bernoulli random variable and a beta prior is placed on the inclusion probabilities (Taylor-Rodriguez et al., 2016). The model space prior is obtained by integrating out these probabilities. In the simplest case, all of conditional inclusion probabilities are assumed to be equal and the prior is called the hierarchical uniform prior (HUP). The amount of penalization of complex models can be adjusted (typically, increased relative to the purely combinatorial penalization of the HUP) using node-specific inclusion probabilities and stronger shrinkage via the beta hyper-priors on the inclusion probabilities; this results in the hierarchical independence (HIP) and order priors (HOP) that group nodes of similar complexity together.

3.2 Model posterior probabilities

In order to compute the posterior probabilities of interest, we take advantage of the model representation making use of the latent variables introduced for the presence and

detection processes. Specifically, a conditional independence argument provides

$$\begin{aligned} p(M_A|\mathbf{y},\mathbf{z},\mathbf{w},\mathbf{v}) &= \frac{m(\mathbf{y},\mathbf{z},\mathbf{w},\mathbf{v}|M_A)\pi(M_A)}{m(\mathbf{y},\mathbf{z},\mathbf{w},\mathbf{v})} \\ &= \frac{f_{\mathbf{y},\mathbf{z}}(\mathbf{y},\mathbf{z}|\mathbf{w},\mathbf{v})m(\mathbf{w},\mathbf{v}|M_A)\,\pi(M_A)}{f_{\mathbf{y},\mathbf{z}}(\mathbf{y},\mathbf{z}|\mathbf{w},\mathbf{v})\sum_{M^*\in\mathcal{M}} m(\mathbf{w},\mathbf{v}|M^*)\pi(M^*)} \\ &= \frac{m(\mathbf{w},\mathbf{v}|M_A)\pi(M_A)}{m(\mathbf{w},\mathbf{v})}, \end{aligned} \tag{10}$$

because $\mathbf{z}$ is independent of M_A once $\mathbf{v}$ is known and $\mathbf{y}$ is independent of M_A once $\mathbf{z}$ and $\mathbf{w}$ are known. In (10),

$$\begin{aligned} f_{\mathbf{y},\mathbf{z}}(\mathbf{y},\mathbf{z}|\mathbf{w},\mathbf{v}) &= \prod_{i=1}^{N} \mathcal{I}^{z_i}_{\{v_i>0\}}\mathcal{I}^{(1-z_i)}_{\{v_i\leq 0\}}\prod_{j=1}^{J}(z_i\mathcal{I}_{\{w_{ij}>0\}})^{y_{ij}}(1-z_i\mathcal{I}_{\{w_{ij}>0\}})^{1-y_{ij}}, \\ m(\mathbf{w},\mathbf{v}|M_A) &= m(\mathbf{v}|M_{A_z})m(\mathbf{w}|M_{A_y}) \\ &= \int\int\left(\prod_{i=1}^{N}\phi(v_i|\mathbf{x}_i'\boldsymbol{\alpha},1;M_{A_z})\right)\left(\prod_{i=1}^{N}\prod_{j=1}^{J_i}\phi(w_{ij}|\mathbf{q}_{ij}'\boldsymbol{\lambda},1;M_{A_y})\right)\times \\ &\qquad \pi^{IP}_{M_A}(\boldsymbol{\alpha},\boldsymbol{\lambda}|\tilde{\mathbf{Q}},\tilde{\mathbf{X}})\mathrm{d}\boldsymbol{\alpha}\mathrm{d}\boldsymbol{\lambda}, \end{aligned} \tag{11}$$

with $\phi(\cdot|\boldsymbol{\mu},\sigma^2;M)$ denoting the normal pdf with mean $\boldsymbol{\mu}$, variance σ^2 conditional on model M, and $\pi^{IP}_{M_A}(\boldsymbol{\alpha},\boldsymbol{\lambda}|\tilde{\mathbf{Q}},\tilde{\mathbf{X}})$ as defined in (5).

Under the intrinsic priors above, the closed-form expression for the marginal $m(\mathbf{v}|M_{A_z})$ is

$$\begin{aligned} m(\mathbf{v}|M_A) &= c_0\,(2\pi)^{-(n-p_{0_z})/2}\left(\frac{p_{A_z}}{2N+p_{A_z}}\right)^{\frac{(p_{A_z}-p_{0_z})}{2}}|\mathbf{X}_0'\mathbf{X}_0|^{-\frac{1}{2}}\times \\ &\qquad \exp\left[-\frac{1}{2}\mathbf{v}'\left(\mathbf{I}-\mathbf{H}_{0_z}-\left(\frac{2N}{2N+p_{A_z}}\right)\mathbf{H}^{\perp}_{A_z}\right)\mathbf{v}\right], \end{aligned} \tag{12}$$

where $\mathbf{H}^{\perp}_{A_z}$ is the hat matrix associated with $(\mathbf{I}-\mathbf{H}_{0_z})\mathbf{X}_A$. Similarly, the marginal distribution for $\mathbf{w}$ under model M_A is

$$\begin{aligned} m(\mathbf{w}|M_A) &= d_0\,(2\pi)^{-(J_\bullet-p_{0_y})/2}\left(\frac{p_{A_y}}{2J_\bullet+p_{A_y}}\right)^{\frac{(p_{A_y}-p_{0_y})}{2}}|\mathbf{Q}_0'\mathbf{Q}_0|^{-\frac{1}{2}}\times \\ &\qquad \exp\left[-\frac{1}{2}\mathbf{w}'\left(\mathbf{I}-\mathbf{H}_{0_y}-\left(\frac{2J_\bullet}{2J_\bullet+p_{A_y}}\right)\mathbf{H}^{\perp}_{A_y}\right)\mathbf{w}\right], \end{aligned} \tag{13}$$

where $\mathbf{H}^{\perp}_{A_y}$ is the hat matrix associated with $(\mathbf{I}-\mathbf{H}_{0_y})\mathbf{Q}_A$ and $J_\bullet=\sum_{i=1}^{N}J_i$ is the total number of surveys. Finally, the marginals for the base model $M_0=\left(M_{0_y},M_{0_z}\right)$ are

$$m(\mathbf{v}|M_0) = \int c_0\,\mathcal{N}\left(\mathbf{v}|X_0\boldsymbol{\alpha}_0,\mathbf{I}\right)\mathrm{d}\boldsymbol{\alpha}_0$$

$$= c_0(2\pi)^{-\frac{(n-p_{0_z})}{2}} |\mathbf{X}_0'\mathbf{X}_0|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left(\mathbf{v}'\left(\mathbf{I}-\mathbf{H}_{0_z}\right)\mathbf{v}\right)\right] \quad (14)$$

and

$$m(\mathbf{w}|M_0) = d_0(2\pi)^{-\frac{(J_\bullet-p_{0_y})}{2}} |\mathbf{Q}_0'\mathbf{Q}_0|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left(\mathbf{w}'\left(\mathbf{I}-\mathbf{H}_{0_y}\right)\mathbf{w}\right)\right]. \quad (15)$$

The specification of the model posteriors in Equation (10) is completed using the construction of the priors $\pi(M_A)$ over the model space; see Section 3.1.

The advantage of (10) is that the posterior of model M_A can be represented as

$$p(M_A|\mathbf{y}) = \iiint p(M_A|\mathbf{y},\mathbf{z},\mathbf{w},\mathbf{v}) f(\mathbf{z},\mathbf{w},\mathbf{v}|\mathbf{y}) \mathrm{d}\mathbf{z}\mathrm{d}\mathbf{w}\mathrm{d}\mathbf{v}, \quad (16)$$

which provides for straightforward ergodic estimation of $p(M_A|\mathbf{y})$ if samples can be drawn from $f(\mathbf{z},\mathbf{w},\mathbf{v}|\mathbf{y})$. If S such samples are obtained, then (16) can be approximated by

$$S^{-1}\sum_{\ell} p(M_A|\mathbf{y},\mathbf{z}^{(\ell)},\mathbf{w}^{(\ell)},\mathbf{v}^{(\ell)}). \quad (17)$$

Such draws can be obtained using reversible jump Markov Chain Monte Carlo (RJMCMC) (Green, 1995), as described in the Supplementary Appendix. One subtle point of difficulty is the calculation of $m(\mathbf{w},\mathbf{v}) = \sum_{M_A} m(\mathbf{w},\mathbf{v}|M_A)\pi(M_A)$ in the denominator of (10) when the space of models is too large to be enumerated (or if the necessary calculations for each model and each draw of $(\mathbf{w},\mathbf{v})$ are too arduous). In such a case, the sum may be approximated by $T^{-1}\sum_t m(\mathbf{w},\mathbf{v}|M^{(t)})\pi(M^{(t)})$, where t indexes a set of T models. For instance, t could index the set of models visited during the RJMCMC sampler or a larger set of models could be used (the posterior of a model M_A not in this set can be estimated using (17)).

4 Simulation experiments

This section considers nine different scenarios where we explore a range of detectability and prevalence regimes to assess the behavior of the proposed algorithm. For each model component, the base model is taken to be the intercept-only model, and the full models considered for the presence and the detection have, respectively, five and three predictors. Therefore, the model space contains $2^5 \times 2^3 = 256$ candidate models. The assumed true models are $M_{Tz} = \{1, x_1, x_2, x_5\}$ for the presence and $M_{Ty} = \{1, q_2, q_3\}$ for the detection, where 1 represents the intercept. This small model space is considered so that comparisons with selection using AIC (which generally requires complete enumeration of the model space) can be made.

The simulation scenarios we consider vary depending on where the distributions for the detection and presence probabilities are centered. That is, we set the average probability for detection and presence to predefined values $\bar{p}$ and $\bar{\psi}$, respectively. If the detection probabilities are centered near one, a non-detection commonly implies a non-presence since the detection is almost perfect. On the contrary, if the

detection probabilities are centered close to zero (as with cryptic species), then the uncertainty surrounding an observed zero is greater, making it more difficult to determine if this also corresponds to a true zero in the presence. Now, combining the different values for $\bar{p}$ with different values for the center of the distribution for the presence probabilities $\bar{\psi}$, we can account for a variety of possibilities observed in real data, ranging from cryptic but highly prevalent species, to easy to detect but very rare species.

The mean probability values for detection and presence that determine our scenarios correspond to the pairs $(\bar{p}, \bar{\psi}) \in \{0.2, 0.5, 0.8\}^{\times 2}$. To match the target values $(\bar{p}, \bar{\psi})$, 15 independent sets of $\{\mathbf{X}_F, \mathbf{Q}_F\}$ were drawn from the standard normal distribution, and for each of them the true model parameters were chosen to solve for $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$ the equations $\hat{\psi}(\boldsymbol{\alpha}) = \bar{\psi}$ and $\hat{p}(\boldsymbol{\lambda}) = \bar{p}$, where

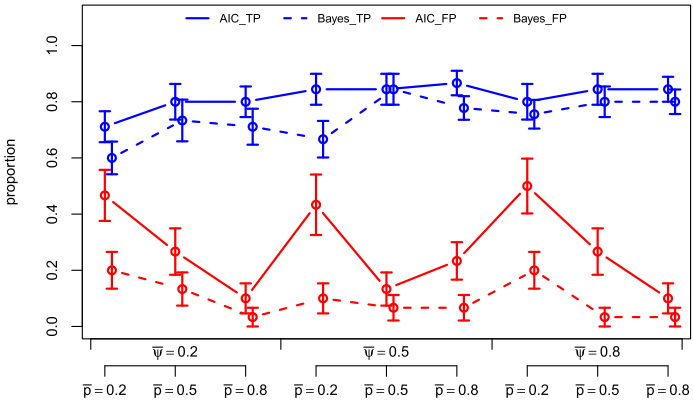
$$\begin{aligned} \hat{\psi}(\boldsymbol{\alpha}) &= \frac{1}{N}\sum_{i=1}^{N} \Phi(\mathbf{x}_i'\boldsymbol{\alpha}) \text{ and} \\ \hat{p}(\boldsymbol{\lambda}) &= \frac{1}{\sum_{i=1}^{N} J_i}\sum_{i=1}^{N}\sum_{j=1}^{J_i} \Phi(\mathbf{q}_{ij}'\boldsymbol{\lambda}). \end{aligned}$$

For each scenario and dataset combination, we used the best solution from ten runs of a gradient-based (quasi-Newton) algorithm initialized from independent standard normal draws. Finally, having determined the regression parameters corresponding to the different scenarios and conditioning on M_{Tz} and M_{Ty}, the true presence and detection indicators were drawn from the probit model described by (2) for each dataset.

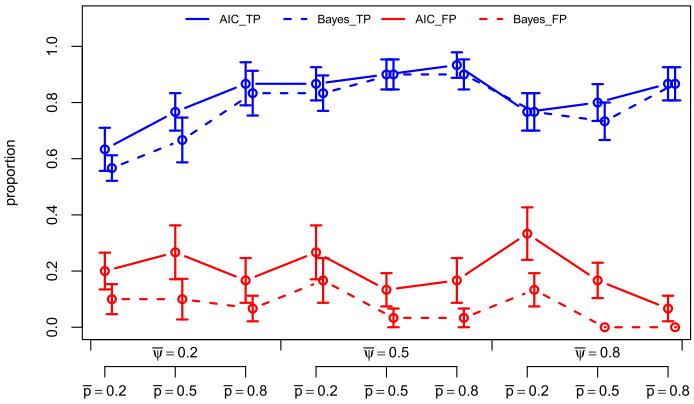
The results are shown in Figure 1, which depicts the average proportion of true positive (TP) and false positive (FP) predictors included in the selected models under each scenario. The TP predictors are those in the true model that are also in the selected model, and the FP predictors correspond to those absent from the true model but included in the chosen model. The selected models are the lowest AIC model and the median probability model (MPM) under the objective Bayes methodology. The MPM is the model that includes all predictors whose marginal posterior inclusion probability (MPIP) is greater than or equal to 0.5, where the MPIP for a given predictor is defined as

$$p(\text{predictor is included}|\mathbf{y}) = \sum_{M \in \mathcal{M}} p(M|\mathbf{y}, \mathcal{M}) \mathcal{I}_{\{\text{predictor} \in M\}}. \tag{18}$$

The TP and FP rates for both detection and presence components lead to the same conclusions. In terms of the TPs, the AIC selects a slightly higher number of true positive terms, especially for the component of the model associated to the presence indicators. Nonetheless, these differences are modest at most. Conversely, the resulting proportions of false positive terms (FP) tend to be strinkingly lower using our method, especially for the presence component in those scenarios where there is poor detection (i.e., $\bar{p} = 0.2$). Remarkably, whenever the species is highly prevalent ($\bar{\psi} = 0.8$) and detection ranges between moderate and high ($\bar{p} = 0.5, 0.8$), the number of false positive terms under our



(a) Presence



(b) Detection

Figure 1: Proportion of true positives (TP) and false positives (FP) using the proposed approach and AIC for the detection and the presence components of the model.

approach is very close to zero in both model components. Also, with $(\bar{p}, \bar{\psi}) = (0.2, 0.8)$ our method substantially outperforms AIC in filtering out the false positive terms both in the presence and detection components.

These results are very encouraging: the proposed method not only reduces the inclusion of false positive terms in comparison to AIC but also has comparable performance finding true predictors.

5 Case studies

In this section, we analyze two datasets. First, we consider presence–absence data for mallard wild ducks (*Anas platyrhynchos*), collected as part of the 2002 Swiss breeding bird monitoring program. For our second example, we consider the blue hawker dragonfly data, which had been previously studied using AIC as the variable selection strategy in Kery et al. (2010). The mallard data is extremely clean, with sufficient sites being surveyed, which for the most part are visited the same number of times. On the other hand, the blue hawker dataset was collected through a large scale citizen science effort. As such, although the number of sites visited is large for this type of data, it displays large asymmetries in the surveying effort, posing a more challenging problem for this type of analysis.

Both data sets contain a sufficiently small number of predictors so that enumeration of the entire model space is feasible. Therefore, for these data analyses, we present estimators of posterior probabilities from enumeration (EPE), renormalization (RPE), and visit frequency (FPE). While all estimates exhibit Monte Carlo error, we treat the enumeration estimators as a gold standard estimator because the Monte Carlo error can be easily controlled. We implement the method of Chib and Jeliazkov (2001) for estimation of the marginal and use a relative magnitude stopping rule to determine the length of sampling (Flegal and Gong, 2013). In particular, we require that the 95% confidence interval for the estimator for the log posterior evaluated at its mode be less than 1% of the size of the estimate.

To obtain the EPE for each model, we run the MCMC algorithm defined by (3) using the priors given in (6)–(8). These yield draws of the regression coefficients conditional on each model, which are then used to calculate the marginal density of the response. We calculate the EPEs using the marginals obtained under each model. Once the EPEs are in place, we then compare them to their corresponding MCMC estimates using either FPE or RPE. Expression (17) enables direct calculation of the RPEs for a specified set $\mathcal{M}_A$ of models, which may even include models that were not sampled. Given the moderate size of the model space for these examples, in both cases we set $\mathcal{M}_A$ to be the entire model space. In contrast, as a general rule the FPEs are only available for the set of the visited models in the RJMCMC. Finally, to compare our results to the traditional approach using AIC, we use the "Akaike weights" (see Burnham and Anderson, 2003; Burnham, 2004, for a definition and further information). These are obtained using functions occu and dredge from the R packages unmarked and MuMin, respectively. The AIC weights allow us to make direct comparison of the results provided by either method, as they can be seen as posterior probabilities obtained from a specific prior on the model parameters. However, as AIC is minimax-rate optimal for estimating the regression function, it cannot be a consistent model selector, as demonstrated in Yang (2005), making these priors ill-suited for variable selection.

5.1 The mallard data

As Switzerland is a small and mountainous country, it provides for large variation in its topography and physio-geography. As such, elevation is a good candidate to predict

species occurrence at a large spatial scale. It can serve as a proxy for habitat type, intensity of land use, temperature, as well as some other biotic factors (Kery et al., 2010). The data used in the illustration was collected by the Swiss breeding bird survey, and had been previously used to derive abundance estimates in Kéry et al. (2005).

The monitoring program for common breeding bird species comprises more than 250 1-km^2 quadrats distributed in a grid sample across Switzerland. Throughout the breeding season, each quadrat is surveyed two or three times annually by an experienced surveyor along a route, recording the date and whether visual or acoustic contact was made. Elevation (elev) and forest cover (forest) were matched for the studied locations from the Swiss Federal Statistical Office (Kéry and Schmid, 2004). Given that the route length (length) across quadrats was not homogeneous, route length (within a quadrat) was considered to account for variation in effective sample area. To model the detection probabilities, survey duration divided by route length (ivel) was used as a measure of effort. Also the date (date) was considered for the detection component since the surveys were collected over a three month period, and behavioral changes that might affect detection could be expected. Using the built-in feature of our algorithm to account for the polynomial structure in the predictors, we considered a full quadratic surface for the predictors, both in the presence as well as in the detection component. The dataset contains 235 quadrats, of which two were surveyed once, 42 twice, and 191 were visited three times.

Results

As mentioned above, given that this dataset contains only a few covariates, even when considering the full quadratic surfaces, it is possible to perform complete enumeration of the model space (which has 1,235 models). The results from our analyses are summarized in terms of the MPIPs (calculated using (18)), the top ranked models (in terms of their posterior probabilities), and the Median Probability Model (MPM), which is the model containing only terms whose MPIPs are greater than 0.5. These measures were all obtained for each method using the posterior probabilities from the joint model for presence and detection.

Table 1 displays the MPIPs calculated with EPEs, RPEs, FPEs and AIC_w. Although the MPIPs obtained from EPE are lower than those from the two other estimates (RPE and FPE), for the most part all three share the same ordering, with the exception of the length^2 term in the presence component. It is worth noting that, although the MPIPs are comparable for the three alternatives, for the detection component those from RPE are considerably closer to the ones from EPE than those from FPE. The MPIPs from AIC_w are considerably higher for most predictors than any of their Bayesian counterparts, implying that good models resulting from AIC selection are more complex, as expected.

Using each of the first three columns displayed in Table 1 one can extract the median probability models (MPM). Following the same approach, with the last column in Table 1, we obtain the 50% threshold model using the AIC weights. These models are displayed in Table 2. The MPM matches for RPE and FPE, and this model in turn is

	EPE	RPE	FPE	AIC_w
elev	0.9966	1.0000	1.0000	1.0000
forest	0.9446	0.9525	0.9489	0.9987
length	0.4305	0.5998	0.5983	0.9625
length*forest	0.2153	0.3803	0.4090	0.8737
elev*length	0.2069	0.3336	0.3491	0.7561
elev*forest	0.1297	0.1448	0.1732	0.3577
$elev^2$	0.1110	0.1293	0.1620	0.3347
$forest^2$	0.1067	0.1229	0.1504	0.3077
$length^2$	0.0734	0.1440	0.1639	0.5333

	EPE	RPE	FPE	AIC_w
date	0.1315	0.1982	0.3846	0.5573
ivel	0.0538	0.1476	0.3568	0.3086
$date^2$	0.0258	0.0560	0.1119	0.3645
$ivel^2$	0.0133	0.0540	0.0980	0.3220
ivel*date	0.0012	0.0250	0.0645	0.0527

Table 1: MPIPs from joint model for the presence (top) and the detection (bottom) components for the mallard dataset.

similar to that from EPE, but the latter excludes the forest term in the presence component. In spite of this discrepancy, it is noteworthy that the MPIP using EPE for this term is 0.4305, being relatively close to the 0.5 threshold for the MPM. The comparable model obtained using AIC weights is considerably larger than all the MPMs resulting with EPE, RPE and FPE, all of which are nested within it.

	Detection	Presence
EPE	{**1**}	{**1**, elev, forest}
RPE	{**1**}	{**1**, elev, forest, length}
FPE	{**1**}	{**1**, elev, forest, length}
AIC_w	{**1**, date}	{**1**, elev, forest, length, length*forest, elev*length}

Table 2: MPMs obtained from MPIPs using EPE, RPE and FPE and pseudo-MPM with AIC weights for the mallard dataset.

Finally, Table 3 displays the five highest probability models (HPMs) under the three calculation alternatives, as well as those resulting from AIC based ranking. Remarkably, the highest probability model is the same under the true posterior probabilities and the two estimation methods considered. Among the set of top models resulting from EPE, four are among the top five from RPE, and three are among those from FPE. Additionally, the model ranked fifth using EPE, which does not match with any of the top five HPMs from RPE or FPE, is ranked eighth and ninth with RPE and FPE, respectively. Also, models ranked fifth under RPE (which coincides with model four with FPE) and fifth under FPE, which are not among the top five with EPE, are respectively ranked eighth and seventh with EPE. Again, more complex top models result from AIC selection in the presence component, and notably the model posterior probabilities are

highly diluted across the model space, with the five top models concentrating only about 8% of the posterior mass. This contrasts markedly with the mass harnessed by the top five models with the other three methods, which are approximately 26% with FPE, 43% for RPE and 55% with EPE.

EPE

	Detection	Presence	$p(M_y, M_z\|\mathbf{y})$
1	{**1**}	{**1**,elev,forest}	0.3101
2	{**1**}	{**1**,elev,length,forest}	0.0954
3	{**1**}	{**1**,elev,length,forest,elev*length,length*forest}	0.0634
4	{**1**}	{**1**,elev,length,forest,elev*length}	0.0420
5	{**1**}	{**1**,elev,forest,elev*forest}	0.0373

RPE

	Detection	Presence	$p(M_y, M_z\|\mathbf{y})$
1	{**1**}	{**1**,elev,forest}	0.1821
2	{**1**}	{**1**,elev,length,forest,elev*length,length*forest}	0.0933
3	{**1**}	{**1**,elev,length,forest,elev*length}	0.0576
4	{**1**}	{**1**,elev,length,forest}	0.0572
5	{**1**}	{**1**,elev,length,forest,length*forest}	0.0431

FPE

	Detection	Presence	$p(M_y, M_z\|\mathbf{y})$
1	{**1**}	{**1**,elev,forest}	0.1063
2	{**1**}	{**1**,elev,length,forest,elev*length,length*forest}	0.0600
3	{**1**}	{**1**,elev,length,forest,elev*length}	0.0354
4	{**1**}	{**1**,elev,length,forest,length*forest}	0.0300
5	{**1**,date}	{**1**,elev,forest}	0.0284

AIC_w

	Detection	Presence	$\text{AIC}_w(M_y, M_z\|\mathbf{y})$
1	{**1**,date}	{**1**,elev,forest,length,elev*length,forest*length}	0.0192
2	{**1**,date}	{**1**,elev,forest,length,length^2,elev*length,forest*length}	0.0190
3	{**1**}	{**1**,elev,forest,length,length^2,elev*length,forest*length}	0.0136
4	{**1**}	{**1**,elev,forest,length,elev*length,forest*length}	0.0136
5	{**1**,date^2}	{**1**,elev,forest,length,elev*length,forest*length}	0.0121

Table 3: Top five models with EPE, RPE, FPE and AIC for the mallard dataset.

The results in Tables 1–3 indicate that estimating the model posterior probabilities using either RPE or FPE yield reasonable approximations to the actual posterior probabilities. In particular, all methods rank models similarly, and if model averaging was to be performed, these would all produce comparable results, as the derived MPIPs resemble each other under the three alternatives. Nonetheless, following the results from Table 1 we prefer RPEs, as these appear to be converging faster towards the benchmark posterior values (EPEs). These results are consistent with the findings from exhaustive simulation experiments conducted in Taylor-Rodriguez et al. (2016),

where overwhelming evidence was found in favor of renormalized model posterior estimates when compared to the frequency-based ones in the multiple linear regression problem. For occupancy models, this behavior is more conspicuous in the detection component than in the presence one, possibly due to the additional uncertainty arising from only partially observing the presence indicators. In addition to the observation that the renormalized posteriors are closer to those from enumeration, in larger model spaces where not all models are visited by the stochastic search, it is possible to calculate renormalized posteriors for a larger set of models than those visited, while with frequency-based estimates this is not possible.

5.2 Blue hawker data

During 1999 and 2000, an intensive volunteer surveying effort coordinated by the Centre Suisse de Cartographie de la Faune (CSCF) was conducted to analyze the distribution of the blue hawker, *Ashna cyanea* (Odonata, Aeshnidae), a common dragonfly in Switzerland. Repeated visits to 1-ha pixels took place to obtain the corresponding detection history. In addition to the survey outcome, the x- and y-coordinates, thermal level, the date of the survey, and the elevation were recorded. Surveys were restricted to the known flight period of the blue hawker, which occurs between May 1 and October 10. In total, 2,572 sites were surveyed at least once during the surveying period. The number of surveys per site ranges from 1 to 22 times within each survey year, with as many as 67% of the sites being surveyed only once, and only 5% of the sites being surveyed more than 3 times. As such, the analysis of this data set is an illustration of a considerably more challenging problem.

Kery et al. (2010) summarize the results of this effort using AIC-based model comparisons. To select the predictors in the detection component, the authors follow a backwards elimination approach while keeping the presence component fixed at the most complex model. To select the presence model, they choose among a group of three models while using the chosen detection model. The full models considered in this study are

$$\begin{aligned}\Phi^{-1}(p) &= \lambda_0 + \lambda_1 \mathsf{year} + \lambda_2 \mathsf{elev} + \lambda_3 \mathsf{elev}^2 + \lambda_4 \mathsf{elev}^3 + \lambda_5 \mathsf{date} + \lambda_6 \mathsf{date}^2 \\ \Phi^{-1}(\psi) &= \alpha_0 + \alpha_1\, \mathsf{year} + \alpha_2\, \mathsf{elev} + \alpha_3\, \mathsf{elev}^2 + \alpha_4\, \mathsf{elev}^3,\end{aligned}$$

where the term year denotes $\mathcal{I}_{\{\mathsf{year}=2000\}}$.

Assuming these full models and intercept only base models (and disregarding the polynomial hierarchy among predictors), the model space for this problem contains $2^{6+4} = 1,024$ models in the joint model space. However, if the polynomial structure is respected, without considering interactions (for compatibility with the analysis in Kery et al. (2010)), the size of the model space for the detection component reduces to 24 models, and to eight models for the presence. This corresponds to a total of 192 models in the combined space. In the exercise below, when using the proposed approach we enforce the strong heredity condition through the priors over the model space.

As in the analysis of the Mallard dataset, we obtain the EPEs, the RPEs, and the FPEs. The model ranks obtained with the posterior probabilities (or their estimates)

are compared to those resulting from AIC selection. The functions used to conduct selection with AIC did not constrain the model space to respect strong heredity, hence for the AIC selection all 1024 models were considered. All results are compared to the models ultimately recommended by Kery et al. (2010), given by

$$\begin{aligned} \text{Detection:} \quad & \{\mathbf{1}, \mathsf{elev}, \mathsf{elev}^2, \mathsf{date}, \mathsf{date}^2\} \\ \text{Presence:} \quad & \{\mathbf{1}, \mathsf{elev}, \mathsf{elev}^2, \mathsf{elev}^3\}. \end{aligned}$$

Results

Table 4 shows the MPMs from either of the approaches considered obtained with the MPIPs found in Table B.1 of Supplementary Appendix B. The MPMs obtained with RPE and FPE coincide, and are similar to that from EPE, with the latter additionally including the elev^2 term. The pseudo-MPM that results when using AIC weights contains all the term included in the MPMs from RPE and FPE, but adds the elev^3 and year terms in the detection component. Note that this model does not respect the polynomial hierarchy, including elev^3 but not elev^2.

	Detection	Presence
EPE	{**1**,date,date^2,elev,elev^2}	{**1**,elev,elev^2}
RPE	{**1**,date,date^2,elev}	{**1**,elev,elev^2}
FPE	{**1**,date,date^2,elev}	{**1**,elev,elev^2}
AIC$_w$	{**1**,date,date^2,elev,elev^3,year}	{**1**,elev,elev^3}

Table 4: MPMs obtained from MPIPs using EPE, RPE and FPE and pseudo-MPM with AIC weights for the blue hawker dataset.

The top ranked models in terms of the true (EPE) and estimated posterior probabilities (RPE and FPE), and from AIC-based selection are displayed in Table 5. The top model obtained with EPE, RPE and FPE are the same for both the presence and detection components, with the top AIC model not respecting the polynomial hierarchy in the detection component (including the elev^3 but not elev^2) and having only the year term in the presence component. Interestingly, four out of the top five models found by EPE coincide with those from RPE, whereas only two from EPE are among the top 5 discovered with FPE, indicating again faster convergence of the renormalized estimates when compared to the frequency based ones. Again, it is worth emphasizing that the probability mass with AIC weights is much more diluted across the model space than with any of its Bayesian counterparts.

	Detection	Presence	$p(M_y, M_z\|\mathbf{y})$
EPE	{**1**,date,date^2,elev}	{**1**,elev,elev^2}	0.2090
RPE	{**1**,date,date^2,elev}	{**1**,elev,elev^2}	0.3725
FPE	{**1**,,date,date^2,elev}	{**1**,elev,elev^2}	0.1974
AIC$_w$	{**1**,date,date^2,elev,elev^3,year}	{**1**,year}	0.0422

Table 5: Top ranked models using EPE, RPE, FPE and AIC weights for the blue hawker dataset.

6 Discussion

This paper developed the first objective Bayes methodology for variable selection using single-season site occupancy models, based on intrinsic priors derived from non-informative priors. This solution uses latent variables to data-augment the analysis, helping to seamlessly calculate the model posterior probabilities. Working on the latent scale additionally facilitates the construction of a straightforward MCMC sampler and posterior estimation using sample averages.

Because the intrinsic priors are built from non-informative priors, the need for hyper-parameter specification is avoided, making the method entirely automatic and widely applicable. Additionally, the types of prior distributions assumed on the model space (HIP, HOP and HUP) enforce the heredity constraints required when performing selection with interactions and higher-order polynomial predictors. These classes also allow for stronger penalization than the usual equal probability prior, further helping control the false positive rate. These have been shown to be particularly useful in problems with small and moderate sample sizes (for more details see Taylor-Rodriguez et al., 2016). An important advantage of our method, relative to the AIC-based selection, is that the resulting model posterior probabilities provide a measure of uncertainty associated with choosing a particular model.

The stochastic search algorithm can be used to thoroughly explore large model spaces by means of the renormalized posterior estimates (instead of the frequency-based ones). This tool will allow practitioners to explore the model space without having to enumerate it or preselect a subset of models, enabling its use with larger model spaces.

The simulation experiments confirmed the ability of the method to identify the predictors present in the true model when considering both the highest and median probability models. The objective Bayes method proved to be competitive with AIC in detecting true predictors, and greatly outperformed AIC in reducing the number of false positive predictors included in the models with high posterior probabilities.

The software used throughout the article was built into the R package OccOBayes available at request. This package includes functions to run the variable selection procedure, as well as some auxiliary functions to validate a set of "best" models using a held-out data set.

Supplementary Material

Supplementary Appendices of "Intrinsic Bayesian Analysis for Occupancy Models" (DOI: 10.1214/16-BA1014SUPP; .pdf).

References

Akaike, H. (1983). Information measures and model selection. *Bulletin of the International Statistical Institute*, 50:277–290. MR0820726. 856

Albert, J. H. and Chib, S. (1993). Bayesian-analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679. MR1224394. 859

Berger, J. and Pericchi, L. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433):109–122. MR1394065. doi: http://dx.doi.org/10.2307/2291387. 857, 861

Burnham, K. and Anderson, D. (2003). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer New York.* MR1919620. 868

Burnham, K. P. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, 33(2):261–304. MR2086350. doi: http://dx.doi.org/10.1177/0049124104268644. 868

Casella, G. and Moreno, E. (2006). Objective Bayesian variable selection. *Journal of the American Statistical Association*, 101(473):157–167. MR2268035. doi: http://dx.doi.org/10.1198/016214505000000646. 862

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321. MR1379473. 860

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the metropolis–hastings output. *Journal of the American Statistical Association*, 96(453):270–281. MR1952737. doi: http://dx.doi.org/10.1198/016214501750332848. 860, 868

Chipman, H. (1996). Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, 24(1):17–36. MR1394738. doi: http://dx.doi.org/10.2307/3315687. 863

Dorazio, R. M. and Rodríguez, D. T. (2012). A Gibbs sampler for Bayesian analysis of site-occupancy data. *Methods in Ecology and Evolution*, 3(6):1093–1098. 856, 858, 859

Fiske, I. and Chandler, R. (2011). unmarked: An R package for fitting hierarchical models of wildlife occurrence and abundance. *Journal of Statistical Software*, 43(10). 856

Flegal, J. M. and Gong, L. (2013). Relative fixed-width stopping rules for markov chain monte carlo simulations. arXiv:1303.0238. 868

Green, P. J. (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, 82(4):711–732. MR1380810. doi: http://dx.doi.org/10.1093/biomet/82.4.711. 865

Guillera-Arroita, G., Lahoz-Monfort, J. J., MacKenzie, D. I., Wintle, B. A., and McCarthy, M. A. (2014). Ignoring imperfect detection in biological surveys is dangerous: A response to: Fitting and interpreting occupancy models'. *PLoS ONE*, 9(7):e99571. 855

Hooten, M. B. and Hobbs, N. T. (2015). A guide to Bayesian model selection for ecologists. *Ecological Monographs*, 85(1):3–28. 857

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76:297–307. MR1016020. doi: http://dx.doi.org/10.1093/biomet/76.2.297. 857

Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343. MR1478684. doi: http://dx.doi.org/10.1214/lnms/1215453065. 857

Kery, M., Gardner, B., and Monnerat, C. (2010). Predicting species distributions from checklist data using site-occupancy models. *Journal of Biogeography*, 37(10):1851–1862. Kery, Marc Gardner, Beth Monnerat, Christian. 858, 868, 869, 872, 873

Kéry, M., Royle, J. A., and Schmid, H. (2005). Modeling avian abundance from replicated counts using binomial mixture models. *Ecological Applications*, 15(4):1450–1461. 858, 869

Kéry, M. and Schmid, H. (2004). Monitoring programs need to take into account imperfect species detectability. *Basic and Applied Ecology*, 5(1):65–73. 869

Leon-Novelo, L., Moreno, E., and Casella, G. (2012). Objective Bayes model selection in probit models. *Statistics in Medicine*, 31(4):353–65. MR2879809. doi: http://dx.doi.org/10.1002/sim.4406. 858, 861, 862

Liu, J. S. and Wu, Y. N. (1999). Parameter expansion for data augmentation. *Journal of the American Statistical Association*, 94(448):1264–1274. MR1731488. doi: http://dx.doi.org/10.2307/2669940. 860

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Royle, J. A., and Langtimm, C. A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83(8):2248–2255. 855, 856

Mazerolle, M. and Mazerolle, M. (2013). Package 'AICcmodavg'. (c). 856

McQuarrie, A., Shumway, R., and Tsai, C.-L. (1997). The model selection criterion AICu. *Statistics & Probability Letters*, 34(3):285–292. MR1458023. doi: http://dx.doi.org/10.1016/S0167-7152(96)00192-7. 857

Moreno, E., Bertolino, F., and Racugno, W. (1998). An intrinsic limiting procedure for model selection and hypotheses testing. *Journal of the American Statistical Association*, 93(444):1451–1460. MR1666640. doi: http://dx.doi.org/10.2307/2670059. 857, 860, 861

Peixoto, J. L. (1987). Hierarchical variable selection in polynomial regression models. *American Statistician*, 41(4):311–313. 858

Peixoto, J. L. (1990). A property of well-formulated polynomial regression-models. *American Statistician*, 44(1):26–30. MR1136106. doi: http://dx.doi.org/10.2307/2684952. 858

Pérez, J. and Berger, J. (2002). Expected-posterior prior distributions for model selection. *Biometrika*, 491–511. MR1929158. doi: http://dx.doi.org/10.1093/biomet/89.3.491. 858, 860

Rao, C. R. and Wu, Y. (2001). *On model selection*, volume 38 of *Lecture Notes–Monograph Series*, pages 1–57. Institute of Mathematical Statistics, Beachwood, OH. MR2000751. doi: http://dx.doi.org/10.1214/lnms/1215540960. 857

Roy, V. and Hobert, J. P. (2007). Convergence rates and asymptotic standard errors for Markov chain Monte Carlo algorithms for Bayesian probit regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):607–623. MR2370071. doi: http://dx.doi.org/10.1111/j.1467-9868.2007.00602.x. 860

Taylor-Rodriguez, D., Womack, A., and Bliznyuk, N. (2016). Bayesian Variable selection on model spaces constrained by heredity conditions. *Journal of Computational and Graphical Statistics* 25(2):515–535. MR3499692. doi: http://dx.doi.org/10.1080/10618600.2015.1056793. 857, 863, 871, 874

Taylor-Rodríguez, D., Womack, A. J., Fuentes, C., and Bliznyuk, N. (2016). "Supplementary Appendices of "Intrinsic Bayesian Analysis for Occupancy Models"." *Bayesian Analysis*. doi: http://dx.doi.org/10.1214/16-BA1014SUPP. 858

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44(1):92–107. MR1770003. doi: http://dx.doi.org/10.1006/jmps.1999.1278. 857

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? *Biometrika*, 92(4):937–950. MR2234196. doi: http://dx.doi.org/10.1093/biomet/92.4.937. 868

Acknowledgments

Taylor-Rodriguez, Womack and Bliznyuk were supported by the National Science Foundation grant DMS-1105127. Taylor-Rodriguez was additionally supported by the National Science Foundation under grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute. Bliznyuk was additionally supported by the National Institutes of Health grants U54GM111274 and R21AI119773. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the National Institutes of Health. The blue hawker dataset was kindly provided by Christian Monnerat and Marc Kéry and authorized for use by the Swiss Biodiversity Monitoring program of the Swiss Federal Office for the Environment (FOEN).