Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2015 doi: 10.1111/2041-210X.12368

Penalized likelihood methods improve parameter estimates in occupancy models

Rebecca A. Hutchinson[1,2]*, Jonathon J. Valente[2], Sarah C. Emerson[3], Matthew G. Betts[2] and Thomas G. Dietterich[1]

[1]School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR 97331, USA; [2]Department of Forest Ecosystems and Society, Oregon State University, Corvallis, OR 97331, USA; and [3]Statistics Department, Oregon State University, Corvallis, OR 97331, USA

Summary

1. Occupancy models are employed in species distribution modelling to account for imperfect detection during field surveys. While this approach is popular in the literature, problems can occur when estimating the model parameters. In particular, the maximum likelihood estimates can exhibit bias and large variance for data sets with small sample sizes, which can result in estimated occupancy probabilities near 0 and 1 ('boundary estimates').

2. In this paper, we explore strategies for estimating parameters based on maximizing a penalized likelihood. Penalized likelihood methods augment the usual likelihood with a penalty function that encodes information about what parameter values are undesirable. We introduce penalties for occupancy models that have analogues in ridge regression and Bayesian approaches, and we compare them to a penalty developed for occupancy models in prior work.

3. We examine the bias, variance and mean squared error of parameter estimates obtained from each method on synthetic data. Across all of the synthetic data sets, the penalized estimation methods had lower mean squared error than the maximum likelihood estimates. We also provide an example of the application of these methods to point counts of avian species. Penalized likelihood methods show similar improvements when tested using empirical bird point count data.

4. We discuss considerations for choosing among these methods when modelling occupancy. We conclude that penalized methods may be of practical utility for fitting occupancy models with small sample sizes, and we are releasing R code that implements these methods.

Key-words: boundary estimates, detection probability, maximum likelihood, occupancy modelling, parameter estimation, penalized likelihood

Introduction

Occupancy models, which explicitly model observation processes in addition to biological processes (MacKenzie *et al.* 2002, 2006), have seen widespread application in ecological modelling (Bailey, MacKenzie & Nichols 2013). Recent work has highlighted controversy over occupancy models, and the literature contains arguments in favour of modelling imperfect detection (Kéry 2011; Guillera-Arroita *et al.* 2014) and warnings about difficulties with fitting these models (Welsh, Lindenmayer & Donnelly 2013). A central argument for occupancy modelling states that when imperfect detection is not addressed, occupancy probabilities are systematically underestimated, and the relationships to habitat variables can be estimated incorrectly (MacKenzie *et al.* 2006). However, there is also evidence that occupancy models can produce poor solutions when detection probabilities are low or sample sizes are small (MacKenzie *et al.* 2002; Guillera-Arroita, Ridout & Morgan 2010; Welsh, Lindenmayer & Donnelly 2013).

Estimates of model parameters and corresponding probabilities are commonly evaluated using two optimality criteria: bias (how the average value of the estimate across repeat experiments would differ from the true value) and variance (how variable the estimate would be across repeat experiments). These criteria jointly determine an estimate's mean squared error (MSE), which is the average squared distance between the estimate and the true value. The MSE of an estimate is equal to its variance plus its squared bias; estimates with smaller MSE are preferred. Problems with parameter fitting in occupancy models occur in the context of maximum likelihood estimation. Maximum likelihood estimates (MLEs) have useful asymptotic (large-sample) properties that facilitate performance of statistical hypothesis tests and construction of confidence intervals. In small samples, however, MLEs can exhibit considerable bias and high MSE, and inference based on their asymptotic properties may be unreliable.

*Correspondence author. E-mail: rah@eecs.oregonstate.edu

Issues with MLEs for occupancy model parameters were first discussed in the introductory paper on occupancy modelling (MacKenzie *et al.* 2002), which noted that using MLEs for model parameters gives biased estimates when detection probabilities are low, and sometimes produces *boundary estimates*, assigning occupancy probabilities at all sites to be 1. Boundary estimates are values very close to the extremes of the set of possible parameter values, which are often unreliable; for probability parameters, boundary estimates are estimates near 0 or 1. Synthetic data experiments have demonstrated boundary estimate issues in occupancy modelling and indicated that uncertainty in these estimates is sometimes underestimated as well (Welsh, Lindenmayer & Donnelly 2013). Uncertainty around boundary estimates of probabilities can be underestimated when MLE parameter estimates are also used to estimate that uncertainty (i.e. a Wald interval is used), as is commonly done for probabilities (Agresti 2013).

Guillera-Arroita, Ridout & Morgan (2010) discussed issues with small sample sizes in occupancy modelling, provided advice for survey design, and developed a criterion to identify data sets in which the MLEs from an occupancy model without covariates include boundary estimates. Welsh, Lindenmayer & Donnelly (2013) showed that boundary estimates also occur in models with covariates. In a response, Guillera-Arroita *et al.* (2014) argued that boundary estimates are a relatively rare problem but agreed that they can occur in small data sets. Small data sets are common in ecology due to logistical constraints and small budgets for field studies (Nakagawa 2004). Since occupancy models are employed to investigate a wide variety of basic scientific and applied questions (Bailey, MacKenzie & Nichols 2013), these fitting problems have troubling implications for the conclusions that may be drawn from occupancy studies. Consequently, the focus of this paper is on practical approaches to improve parameter estimates and reduce the incidence of boundary estimates in small data sets.

To address this problem, we generate synthetic data to investigate the conditions that produce poor parameter estimates in occupancy models and explore three methods to improve these estimates. We examine the distribution of the MLEs across replicate data sets in a variety of settings. Our results are consistent with other studies in revealing boundary estimates and bias in the MLEs, particularly for small data sets. The MLEs improve as sample sizes increase, but in practice, collecting more data may be infeasible. For situations in which correcting for imperfect detection is crucial but limited data are available, we explore three parameter estimation approaches based on penalized likelihoods.

Penalized likelihood, or *regularization*, methods are used in machine learning and statistics to control model complexity and reduce the variance of parameter estimates (Murphy 2012). Penalized likelihood methods augment the likelihood with a penalty function, which can be chosen to encode prior knowledge about the parameters or discourage undesirable estimates (*e.g.* very large values). Parameter estimates are computed by maximizing the new objective function, combining the original likelihood and the penalty function. The first penalty we consider is an occupancy modelling analogue to ridge regression (Hoerl & Kennard 1970; Hastie, Tibshirani & Friedman 2009), which penalizes parameter estimates far from zero. The second method we consider is a penalty that corresponds to placing zero-mean Gaussian priors on the parameters. The third penalty we consider was developed specifically for occupancy models (Moreno & Lele 2010). This method shrinks parameter estimates on the occupancy side of the model towards their corresponding estimates from logistic regression.

Our experiments show that these penalized likelihood methods reduce the MSE of parameter estimates and decrease the frequency of boundary estimates. We find that the variance reduction effect of penalization dominates the effect of bias in reducing mean squared error. Furthermore, bootstrapped confidence intervals for the penalized estimators have coverage similar to bootstrapped confidence intervals for MLEs, but the width of the intervals is smaller under penalization. We also present a case study on 25 avian species, comparing occupancy model parameter estimates obtained from a relatively large data set with estimates obtained from a small subset of the data, to simulate a scenario where fewer data were collected. In some cases, the MLEs on the small samples fail badly, and the penalized estimation methods provide an alternative to eliminating these species from the analysis. In other cases, estimates from the MLEs and the penalized methods are similar. We conclude with a discussion of factors to consider when choosing an estimation method.

Materials and methods

A REVIEW OF OCCUPANCY MODELS

Occupancy models describe a joint probability distribution over the latent occupancy status of each site in a study and the observations made on multiple visits to those sites during a period of population closure, when the occupancy status of the sites is constant. Below, we index sites by i and visits by t. The probability of occupancy at site i is denoted by $\psi_i \in [0,1]$, and the latent occupancy status of site i is $Z_i \in \{0,1\}$. The probability of the observations, denoted $Y_{it} \in \{0,1\}$, is conditional on the occupancy status. The probability of detecting the species on visit t given that site i is occupied is $p_{it} \in [0,1]$. The probability of detecting the species when the site is not occupied is 0.

The logistic function $\sigma(x) = (1 + \exp(-x))^{-1}$ can be used to link the occupancy and detection probabilities to covariates. Arbitrary numbers of covariates can be included, but in this paper, we will focus mainly on the case with one covariate each for occupancy and detection, to ease visualization and interpretation of the results. Letting x and w denote the occupancy and detection covariates, respectively, the probability distributions over Z_i and Y_{it} are

$$\begin{aligned} \psi_i &= \sigma(\alpha_0 + \alpha_1 x_i) \\ Z_i &\sim \text{Bernoulli}\ (\psi_i) \\ p_{it} &= \sigma(\beta_0 + \beta_1 w_{it}) \\ Y_{it}|Z_i &\sim\ \text{Bernoulli}\ (Z_i p_{it}). \end{aligned} \qquad \text{eqn 1}$$

We denote the parameters of this model $\Theta = \{\alpha_0, \alpha_1, \beta_0, \beta_1\}$. The likelihood function for a data set with M sites visited T times each is

$$L(\Theta) = \prod_{i=1}^{M}\left[\psi_i \prod_{t=1}^{T}[p_{it}^{Y_{it}}(1-p_{it})^{1-Y_{it}}] + (1-\psi_i)I\left(\left(\sum_{t=1}^{T} Y_{it}\right) = 0\right)\right], \quad \text{eqn 2}$$

where the indicator function $I(\cdot)$ returns 1 if its argument is true and 0 otherwise, and the dependence on the parameters Θ is through ψ and p.

SYNTHETIC DATA SETS

We generated synthetic data sets by sampling from the probability distributions entailed by the occupancy modelling framework, following Welsh, Lindenmayer & Donnelly (2013) to choose covariates and parameter values. Their synthetic data were inspired by studies of two avian species in Australia. Each data set includes one site-specific covariate, which influences both occupancy and detection and is constant across visits. The covariate represents stand age of pine plantations surrounding the sites, and it takes values in {1,2,3,4,5}, with each value occurring at an equal number of sites.

We generated data sets using four different settings, called *ideal*, *sparse*, *nonzero* and *altcov*. The *ideal* parameters are $\Theta_{ideal} = \{-0{\cdot}405, 0, -0{\cdot}533, 0{\cdot}22\}$, which produce occupancy probabilities around 0·4. The *sparse* parameters are $\Theta_{sparse} = \{-2{\cdot}197, 0, -0{\cdot}533, 0{\cdot}22\}$, which produce occupancy probabilities around 0·1. For both data sets, detection probabilities are between 0·42 and 0·64. Under the *ideal* parameterization, we expect occupancy models to do well because occupancy and detection probabilities are not small, whereas we expect that the low number of positive observations under the *sparse* parameterization may be problematic. Both settings were explored by Welsh, Lindenmayer & Donnelly (2013).

We generated data with two additional settings. In the *nonzero* setting, we used the same covariates but a nonzero occupancy slope ($\Theta_{nonzero} = \{-0{\cdot}405, -0{\cdot}25, -0{\cdot}533, 0{\cdot}22\}$). The motivation for this data set was that parameter values of 0 are a special case for two of the penalized likelihood methods described below. In the *altcov* setting, we used the parameters Θ_{ideal}, but we generated a different covariate for occupancy. This setting ensures that the identifiability of the covariate's effect on the two sides of the model is not a concern. The new covariate was drawn from a *Uniform*(−1,1) distribution, but since $\alpha_1 = 0$, it had no effect on the observations, so this setting also introduces an irrelevant covariate.

For each parameterization, we generated four sizes of data sets, also following previous work (Welsh, Lindenmayer & Donnelly 2013). We sometimes denote data set sizes below as pairs of values for M (sites) and T (visits); the four sizes are $(M,T) \in \{(55,2),(55,5),(165,2),(165,5)\}$. For each of the sixteen scenarios (four parameterizations times four sizes), we created 5000 simulated data sets.

The work on which our synthetic data design is based considers one other setting, which we do not include, called *abund* (Welsh, Lindenmayer & Donnelly 2013). In that setting, detection probabilities are not computed in accordance with the occupancy modelling framework; instead, they are driven by simulated abundance. On these data, occupancy models fit detection probabilities poorly, since they cannot capture the functional form used to generate the data. This mismatch, or model misspecification, is an important concern for ecological modellers, but it is outside the scope of this paper.

EMPIRICAL CASE STUDY

We also studied an empirical data set of avian point counts. The original data set consisted of 656 sites each visited three times. We fit species-specific occupancy models to this large data set. Next, we created a reduced data set by randomly selecting 55 sites and eliminating the third visit to each, simulating a scenario in which we had collected fewer observations. We fit occupancy models to this reduced data set and compared the parameter estimates produced by maximum likelihood and penalized likelihood methods to one another and to the MLEs from the full data set.

Field sampling was conducted in the spring of 2011 in southern Indiana. We conducted three ten-minute avian point counts at each of the 656 locations, following a common protocol (Betts *et al.* 2008). We also took vegetation measurements at the locations. We chose *a priori* to model distributions of only those species detected at greater than 100 sites in the full data set. We set this threshold to ensure that the MLEs on the large data set would be reliable, since our comparisons with the reduced data set treat the large-sample MLEs as a proxy for the true parameter values. We also chose to incorporate only one detection covariate (with or without a quadratic term) and one occupancy covariate in our analysis for each species to simplify evaluation of the results. We followed a standard model selection procedure to specify models for each of 25 species. Because we are not interested in the biological implications of these covariates, we did not distinguish between them in the remainder of our analyses. All were standardized prior to inclusion in the models. (See Appendix S2.)

ESTIMATION METHODS

Maximum likelihood estimation

To obtain MLEs, we employed the *optim()* R function and supplied our own implementations of the likelihood and gradient functions. The *unmarked* R package also relies on *optim()* for fitting occupancy and related models (Fiske & Chandler 2011; R Core Team 2013). For a subset of data sets, we confirmed that estimates from our implementation were identical to those from *unmarked*. We refer to this method as *MLE* below.

A ridge penalty

Maximum likelihood estimation in the standard regression context uses the data likelihood as the objective function for choosing parameters. In ridge regression, the objective function is modified, or *regularized*, by subtracting a penalty term from the data likelihood (Hoerl & Kennard 1970; Hastie, Tibshirani & Friedman 2009). The ridge penalty term is $\lambda\sum_i \theta_i^2$, where i indexes the non-intercept coefficients in the model and λ is a tuning parameter that modulates the trade-off between the likelihood term and the penalty term. This penalty shrinks the estimates towards zero, since the penalty is high for coefficients with large magnitude.

We apply ridge regularization to occupancy models by augmenting the objective function (the occupancy model likelihood) with an analogous penalty term on the non-intercept coefficients of both the occupancy and detection sides of the model. This can be interpreted as shrinking towards an occupancy model with constant occupancy and detection probabilities. For the synthetic data discussed above, this is $\lambda(\alpha_1^2 + \beta_1^2)$.

The new objective function is:

$$\log(L(\Theta)) - \lambda\frac{1}{2}(\alpha_1^2 + \beta_1^2), \qquad \text{eqn 3}$$

where the 1/2 is introduced to simplify derivatives. We refer to this method as *Ridge* below.

To choose a value for λ, we performed fivefold cross-validation on each data set (i.e. each of the 5000 for a given scenario), where λ was chosen to maximize the likelihood of the held out data. The choices for λ were {0·02, 0·1, 0·2, 0·33, 1, 2}. Once a value was chosen for λ, the estimation method was run on the entire data set using that value to obtain the final parameter estimates.

A bayes-inspired penalty

The ridge penalty term excluded the intercept coefficients. Next, we consider the same form of penalty expanded to include the intercept parameters. Intercept parameters are not commonly included in penalty functions for regression, but we explored this case since the MLEs for these parameters had high variance in our experiments, as discussed below. Specifically, we use $\lambda\sum_i \theta_i^2$, where i now indexes all of the parameters in the model and λ again is a tuning parameter trading off the likelihood and penalty terms. This can be interpreted as shrinking towards an occupancy model with constant occupancy and detection probabilities of 0·5. The new objective function is:

$$\log(L(\Theta)) - \lambda\frac{1}{2}\sum_i \theta_i^2. \qquad \text{eqn 4}$$

We use the same fivefold cross-validation procedure to choose a value for λ, again from among the values {0·02, 0·1, 0·2, 0·33, 1, 2}.

This penalty term has a connection to Bayesian methods. If we place an identical zero-mean Gaussian prior distribution on *all* of the θ_i values, then the log of this prior distribution is equal to $\frac{1}{(2\sigma^2)}\sum_i \theta_i^2$. This expression is identical to the penalty in Equation 4 with $\lambda = \frac{1}{(\sigma^2)}$ (so the λ values correspond to variances of $\sigma^2 \in \{50, 10, 5, 3{\cdot}03, 1, 0{\cdot}5\}$). Hence, the value of Θ that maximizes Equation 4 is equivalent to a maximum a posteriori probability (MAP) estimate computed from the Bayesian model. Consequently, we refer to this penalty as *Bayes*.

In a full Bayesian treatment, we would introduce a hyperprior over σ and then apply Markov chain Monte Carlo methods to sample from the posterior distribution [*e.g.* using WinBUGS; Lunn *et al.* (2009); Kéry (2010)]. For a limited number of cases, we verified that WinBUGS gives results consistent with our methods. Our approach of setting λ using cross-validation is essentially an empirical Bayes (or Type-II Likelihood) approach (Morris 1983). The key advantage of MAP estimation is that it can be solved using standard optimization software (*e.g. optim()* in R) instead of requiring Markov chain Monte Carlo methods.

A logistic regression penalty

The regularization approaches in the previous two sections penalize coefficient estimates far from zero. An alternative approach developed specifically for occupancy models regularizes towards the logistic regression solution obtained when assuming perfect detection (Moreno & Lele 2010). The rationale behind the approach is to leverage the numerical stability of the parameter estimates obtained from logistic regression.

The method first estimates the parameters of a logistic regression that predicts whether a positive observation was made on at least one visit using the occupancy covariates. Then, when fitting the occupancy model, the likelihood is augmented with a penalty term equal to the absolute difference between the occupancy coefficients and the logistic regression parameters: $\lambda\sum_i |\theta_i - \theta_i^{(lr)}|$, where i indexes the parameters on the occupancy side of the model and $\theta^{(lr)}$ denotes the logistic regression parameters. The new objective function is

$$\log(L(\Theta)) - \lambda\sum_i |\theta_i - \theta_i^{(lr)}|. \qquad \text{eqn 5}$$

Instead of using cross-validation to choose a weight λ for the penalty term, this approach computes it as

$$\lambda = \sqrt{\sum_j \text{var}\ (\theta_j) \times (1 - (1 - \bar{\hat{p}}_0)^T)(1 - \bar{\hat{\psi}}_{\text{naive}})}. \qquad \text{eqn 6}$$

Here, j indexes the detection parameters, $\bar{\hat{p}}_0$ is the average detection probability computed for an unconstrained occupancy model, T is the number of visits to the sites, and $\bar{\hat{\psi}}_{\text{naive}}$ is the average occupancy probability from the logistic regression. Using this setting, the strength of the regularization decreases as the number of sites or visits increases. Additionally, λ increases with large detection probabilities (since the logistic regression estimates are correct when detection is perfect), and λ decreases with large occupancy probabilities (since occupancy models are stable in this case). Following Moreno & Lele (2010), we refer to this method as *MPLE* (for maximum penalized likelihood estimation).

Note that Equation 6 depends on the variances of the detection parameters. These variances can typically be obtained from the covariance matrix returned by *MLE*. However, in some cases, the covariance matrix cannot be computed (i.e. the *unmarked* package fails to return standard errors on the parameters). When these cases occurred in the data sets generated for this paper, we removed those data sets from the MPLE and uncertainty analyses.[1]

Ignoring imperfect detection

We also fit logistic regression models that ignored imperfect detection to the same data sets. These models had an intercept term and a slope term on the covariate that affected both occupancy and detection. We cannot compare the two parameters estimated in this misspecified model to the four parameters that truly generated the data, but we can compare the occupancy estimates computed from each.

COMPARING PARAMETER ESTIMATION APPROACHES

The mean squared error of an estimator is equal to the sum of its squared bias and its variance. For synthetic data, we used the true values of the parameters to compute the MSE, bias and variance of the estimates from each method. The bias is computed as the estimate minus the truth, averaged over the 5000 data sets. The variance is computed over the parameter estimates from the 5000 replicate data sets. For empirical data, we compared the estimated parameters from each method with the MLEs from the full data set for each species.

The uncertainty in MLEs can be quantified with confidence intervals based on asymptotic properties of the estimators. As the data set size goes to infinity, the estimators follow a Gaussian distribution centred at the true parameter values, with variance that is asymptotically arbitrarily well approximated by the sample variance divided by the size of

[1] For data sets with M = 55 and T = 2, we removed 5 *altcov* data sets, 10 *ideal* data sets, 41 *nonzero* data sets and 241 *sparse* data sets. From the *sparse* data sets, we also removed 88 data sets with M = 5 and T = 5 and 7 data sets with M = 165 and T = 2.

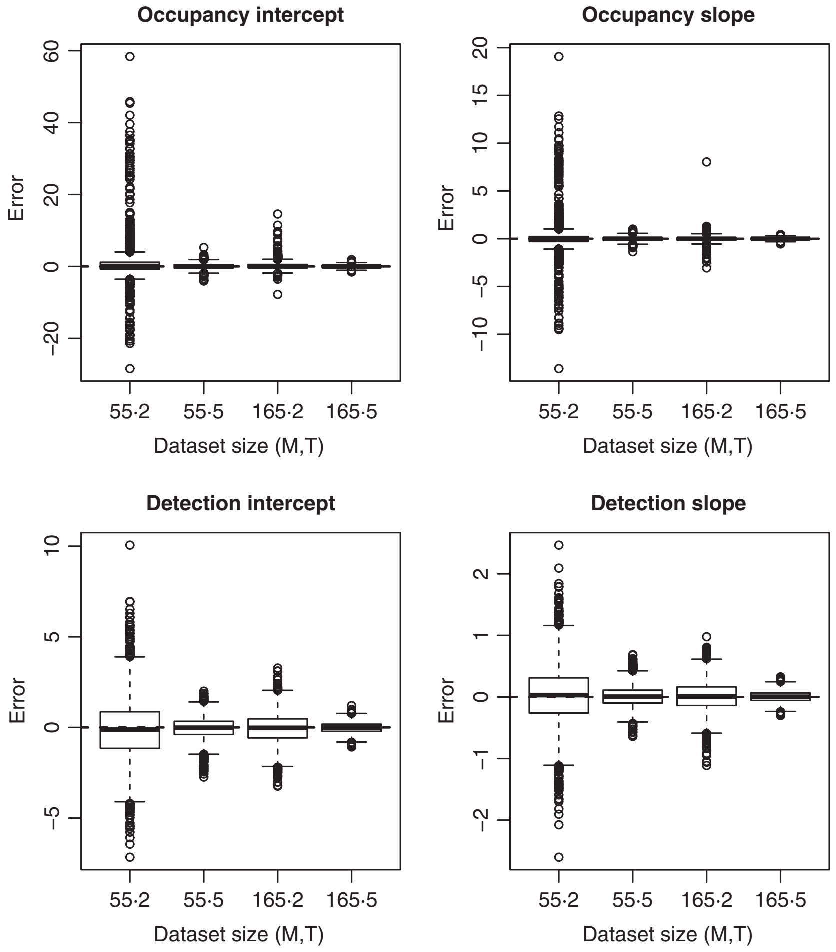

Fig. 1. Error in parameter estimates (estimate minus truth) produced by *MLE* on synthetic data sets of different sizes (*M* sites and *T* visits) for the *ideal* parameter settings (occupancy probabilities near 0·4). Each plot shows error in the estimates for 5000 data sets. The dashed line indicates zero error. Error decreases as data set size increases.

the data set. When penalties are introduced, these confidence intervals do not hold. Instead, we use the bootstrap to assess uncertainty in parameter estimates. For each of the 5000 synthetic data sets and for each species in the case study, we drew 200 bootstrap replicates and computed the parameter estimates for each. Following Moreno & Lele (2010), each bootstrap replicate was sampled with replacement from the set of sites, and the set of visits associated with each site remained unchanged. For the penalized methods, we ran the λ selection process on each replicate. We compared confidence intervals across methods based on their coverages and widths.

Results

Here, we present experimental results for a subset of synthetic data sets to illustrate our key findings. Analogous results for other cases are available in Appendices S1 and S2.

PROBLEMS WITH MAXIMUM LIKELIHOOD ESTIMATION

First, we examine the empirical distributions of the maximum likelihood estimates for the *ideal* parameterization across data sets of different sizes. The variability of the estimates is large relative to the magnitude of the parameters for the smallest data sets (Fig. 1). The variability in the occupancy intercept tends to be larger than the variability in the other parameter estimates. As expected, increasing the size of the data set (number of sites or number of visits) decreases the variability of the estimates. Error in the parameter estimates propagates to ψ and p as well, which can result in boundary estimates. Following Welsh, Lindenmayer & Donnelly (2013), we define a boundary estimate as a probability <0·0001 or >0·9999. For data sets with size (M,T) = (55,2), boundary estimates for ψ or p occurred in 2% of the data sets with the *ideal* parameterization and 36% of the data sets with the *sparse* parameterization. None of the data sets with (M,T) = (165,5) produced boundary estimates.

COMPARISON WITH PENALIZED LIKELIHOOD ESTIMATION

Since the problems described above are most prevalent for small sample sizes, we focus here on the smallest data sets (55 sites and 2 visits). The distributions of penalized likelihood estimates show less variability than the MLEs (Fig. 2). Lower variability in the raw parameter estimates translates into lower variability in the estimated occupancy and detection probabilities (Fig. 3) and fewer boundary estimates than the MLEs (Table 1). The estimated probabilities can also be compared against the alternative of ignoring imperfect detection. Figure 3 illustrates the bias introduced by assuming perfect detection for these data sets. The medians of the logistic regression estimates of the occupancy probabilities are further from the true value than the other methods. The influence of the logistic

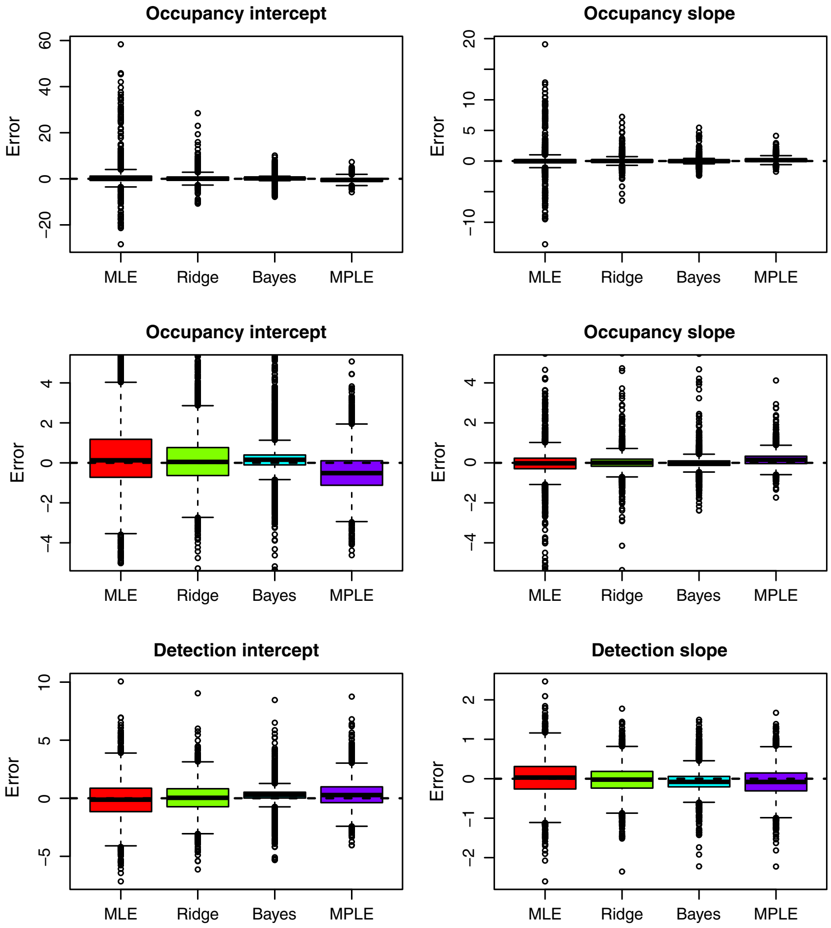

Fig. 2. Error in the parameter estimates generated by the four methods on 5000 data sets with $M = 55$ sites and $T = 2$ visits generated from the *ideal* parameterization (occupancy probabilities near 0·4). Dashed lines indicate zero error. The first two rows both depict occupancy parameters; the first row shows the full range of the estimates, and the second row is zoomed in to the region between −5 and 5. The variance of the error in penalized estimates is lower than the variance of the MLEs.

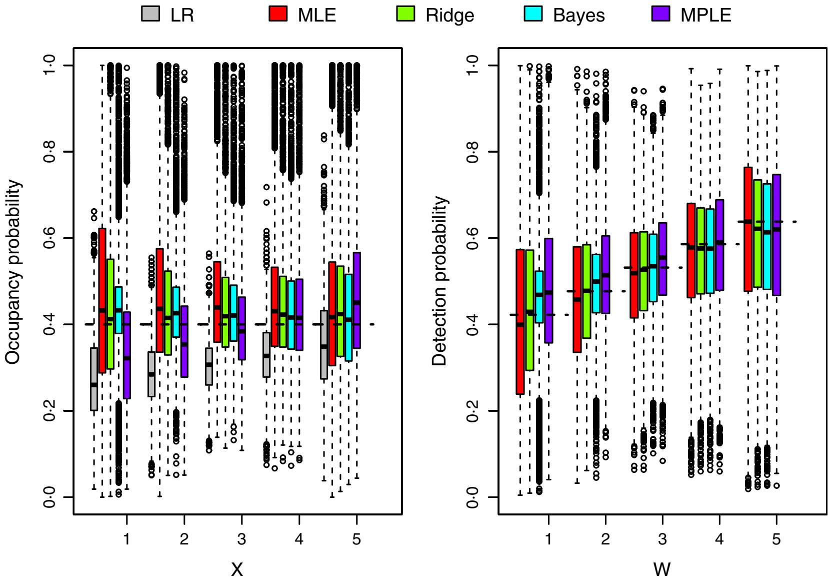

Fig. 3. Estimated occupancy (left) and detection (right) probabilities for five methods on 5000 data sets with 55 sites and 2 visits generated from the *ideal* parameters, as a function of the values of the covariates X and W. The dashed horizontal lines indicate the true values. Logistic regression (LR) is included on the left plot (which assumes perfect detection). LR is systematically biased below the true occupancy probability, MLE shows the largest variability, and the penalized methods have decreased variability. The influence of the LR estimates on the *MPLE* estimates is apparent. While the *Bayes* estimates are lower variance, the *Ridge* estimates are more centred on the true values.

regression solution on the *MPLE* results is also visible. While the logistic regression estimates are centred farther from the truth than the other methods, they also have lower variance and produce no boundary estimates. This is the numerical stability that *MPLE* attempts to exploit.

BIAS AND VARIANCE OF THE ESTIMATES

Here, we present the *nonzero* parameterization results in addition to the *ideal* results. The *ideal* parameterization has 0 as the true value for the occupancy slope parameter. Since the *Ridge*

Table 1. Percentage of data sets producing boundary estimates for the occupancy or detection probabilities and for combinations of parameter settings, data set sizes and methods. Boundary estimates are more prevalent for small sample sizes and sparse data sets

	M = 55		M=165	
	T = 2	T = 5	T = 2	T = 5
ideal				
MLE	2·32	0	0·04	0
Ridge	0·36	0	0·02	0
Bayes	0·20	0	0	0
MPLE	0·08	0	0·02	0
nonzero				
MLE	6·26	0·02	0·16	0
Ridge	1·06	0	0·04	0
Bayes	0·50	0	0·02	0
MPLE	1·00	0	0·02	0
sparse				
MLE	36·2	6·20	1·02	0
Ridge	11·4	2·60	0·34	0
Bayes	1·20	0·42	0·14	0
MPLE	15·6	1·96	0·34	0

and *Bayes* methods penalize parameter values far from 0, these methods have an advantage when 0 happens to be the truth. We present the *nonzero* results to compare against a data set in which the penalty function does not coincide with the true parameter values.

Our results show that the variance component of the MSE for the MLEs often dominates (Table 2). Additionally, for the small data sets, the MLEs exhibit bias of a magnitude similar to the true value of the parameter in some cases. In contrast, the penalized likelihood estimates have drastically reduced variance. In some cases, they are less biased than the MLEs; in other cases, the MLEs are less biased.

The methods differ in the rate at which bias decreases as data set size increases (Fig. 4). With a good choice of λ (via cross-validation) and as the number of samples grows, the effect of the penalty term shrinks until its influence becomes negligible. This behaviour is more apparent in *Ridge* than *Bayes* and *MPLE*. The bias of the *Ridge* estimates is nearly zero for the largest data sets and is lower than the bias of the MLEs.

Overall, the penalized methods usually outperform the MLEs. Of the 64 combinations of parameter, data set size and settings, *Bayes* has the lowest MSE in 58 cases, and *Ridge* has the lowest MSE in 2 cases, and *MPLE* has the lowest MSE for the other four cases. The MLEs do not achieve the lowest MSE in any of the cases considered in this study. *MLE* is least biased in 12 cases, *Ridge* is least biased in 40 cases, *Bayes* is least biased in six cases, and *MPLE* is least biased in six cases.

UNCERTAINTY IN THE ESTIMATES

We computed confidence intervals for the MLEs based on their asymptotic properties for *ideal* data sets of all sizes and evaluated their coverage and width. On the smallest data sets, the empirical coverage of the 95% confidence intervals is higher than 95% for the occupancy parameters and lower than 95% for the detection parameters (see Appendix S1). The confidence intervals, especially for the occupancy parameters, are quite wide compared to the values of the true parameters. For the largest data sets, the coverage of the intervals is close to the nominal 95% on average, and the intervals are narrower. The substantial deviations in coverage from the target of 95% on the smallest data sets indicates that these sample sizes are insufficient for asymptotically justified confidence intervals.

Table 2. Bias, variance and mean squared error (MSE) for each parameter, on data sets with M = 55 sites and T = 2 visits, generated from the *ideal* and *nonzero* parameterizations (lowest values in bold). The MLEs have the highest MSE, and the variance component of the MSE dominates the squared bias component

		MLE	Ridge	Bayes	MPLE
ideal					
Occupancy	Bias	0·515	0·183	**0·171**	−0·469
Intercept	Var	12·3	2·46	**0·844**	1·03
	MSE	12·6	2·50	**0·873**	1·25
Occupancy	Bias	**−0·00214**	0·00816	−0·00961	0·156
Slope	Var	1·32	0·188	**0·104**	0·106
	MSE	1·32	0·188	**0·104**	0·130
Detection	Bias	−0·132	**0·0433**	0·168	0·322
Intercept	Var	2·39	1·50	**0·736**	1·30
	MSE	2·41	1·50	**0·764**	1·40
Detection	Bias	0·0187	**−0·0274**	−0·0571	−0·0797
Slope	Var	0·211	0·122	**0·0787**	0·134
	MSE	0·211	0·122	**0·0819**	0·140
nonzero					
Occupancy	Bias	0·779	0·206	**0·137**	−0·520
Intercept	Var	25·6	4·36	1·12	**1·05**
	MSE	26·2	4·40	**1·14**	1·32
Occupancy	Bias	**0·0190**	0·0632	0·0314	0·178
Slope	Var	3·06	0·329	0·199	**0·139**
	MSE	3·06	0·333	0·200	**0·171**
Detection	Bias	**−0·0752**	0·172	0·228	0·455
Intercept	Var	8·870	2·77	**1·22**	3·83
	MSE	8·88	2·80	**1·27**	4·03
Detection	Bias	**−0·00932**	−0·0999	−0·117	−0·107
Slope	Var	1·16	0·329	**0·210**	0·582
	MSE	1·16	0·339	**0·224**	0·593

Confidence intervals computed using the bootstrap are applicable to all the methods we evaluated. For the smallest *ideal* data sets, the coverages are similar across methods, and they do not achieve exactly 95% (Table 3). The widths of the intervals are narrower for the penalized methods than the MLEs.

EMPIRICAL DATA CASE STUDY

Our analysis of the large-sample and reduced-sample empirical data sets provides insight into the estimates that may have been produced with fewer data. The MLE distribution on the reduced data set contains large outliers for a few species (Fig. 5). The penalized methods reduce the variance of the estimate distributions substantially. For the *Bayes* penalty, this variance reduction is accompanied by an increase in the difference from the large-sample estimates, whereas for *Ridge*, the estimates are close to the large-sample MLEs.

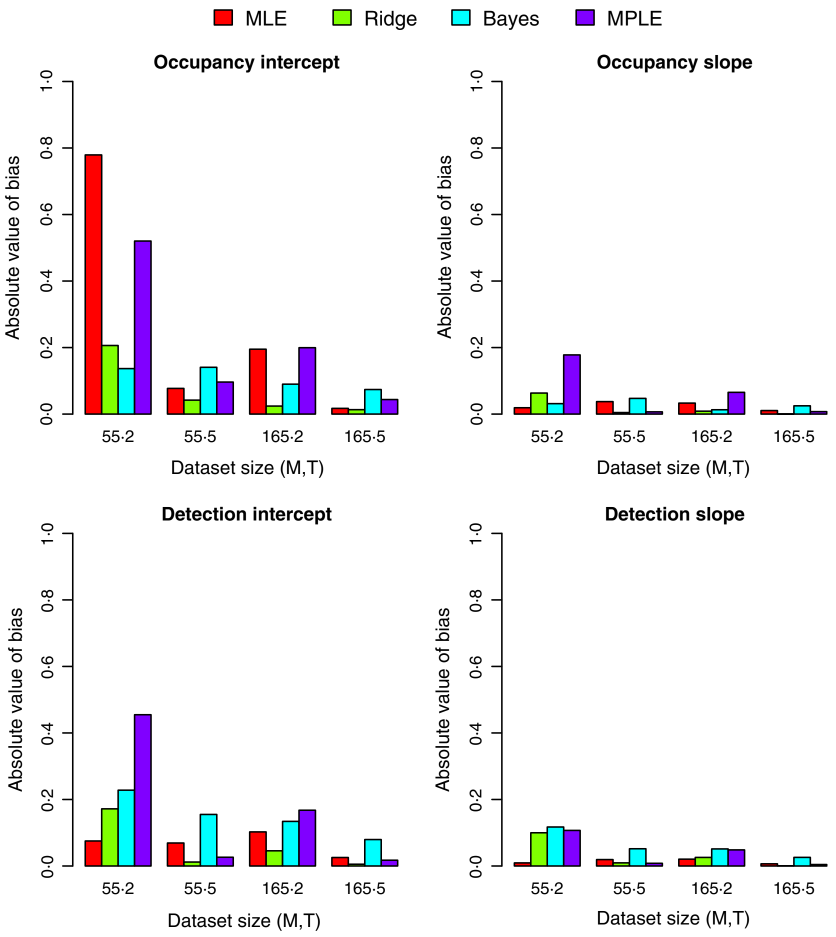

Fig. 4. Each plot shows the absolute value of the bias in the estimates for one parameter from each of the four methods on data sets of different sizes (M sites and T visits) generated from the *nonzero* parameterization. Bias in the penalized methods decreases as more data becomes available. Bias in *Ridge* decreases faster than *Bayes* as data set size increases.

Table 3. Coverage and width of 95% bootstrapped confidence intervals (200 replicates) for the *ideal* parameterization with M = 55 sites and T = 2 visits. Widths are averaged across data sets. Coverages are comparable across methods, and widths are lower for the penalized methods

		MLE	Ridge	Bayes	MPLE
Occupancy	Coverage	97·2%	97·4%	96·6%	92·3%
Intercept	Width	19·0	13·1	2·60	4·70
Occupancy	Coverage	97·2%	97·4%	97·7%	92·9%
Slope	width	6·60	4·04	1·01	1·66
Detection	Coverage	92·0%	92·4%	94·7%	92·1%
Intercept	Width	7·30	6·88	2·61	5·55
Detection	Coverage	93·6%	93·8%	93·0%	92·3%
Slope	Width	2·29	2·12	0·98	1·83

We used the large-sample MLEs as a proxy for the true values of the parameters to approximate the coverage of the bootstrapped confidence intervals for each method. While the confidence intervals do not contain the large-sample MLE exactly 95% of the time, the coverages are similar across methods (see Appendix S2). The *Ridge* and *Bayes* penalties have narrower confidence intervals than the *MLE* on the reduced data sets. The *MPLE* method also reduces confidence interval width for the occupancy parameters, but its intervals are wider than those of *MLE* for the detection parameters.

Discussion

In the frequentist setting, the task of an estimation procedure is to provide parameter estimates close to their true values. Mean squared error between the estimates and truth is a straightforward way to measure performance on this task. Our experiments confirm other reports that maximum likelihood estimation of occupancy model parameters can produce estimates with large error for small sample sizes. We find that while the MLEs can exhibit some bias, the variance component of the MSE dominates. As expected, both bias and variance decrease with larger data sets. When possible, sample sizes should be increased to alleviate concerns about the quality of the MLEs. Ideally, literature on survey design can be consulted [*e.g.* MacKenzie & Royle (2005); Bailey *et al.* (2007); Guillera-Arroita & Lahoz-Monfort (2012)] before the data collection process begins to ensure that sufficient data will be collected to address the scientific question of interest. When collecting more data is impossible, we recommend penalized likelihood estimation. In every setting we tested, the penalized estimation methods had lower MSE than the MLEs.

Bias and variance analysis reveals general trends regarding the source of the error reduction for the penalized likelihood methods. While all the penalized methods reduce the variance in the estimates, *Bayes* does so most aggressively, which often results in the lowest MSE. On the other hand, *Ridge* often

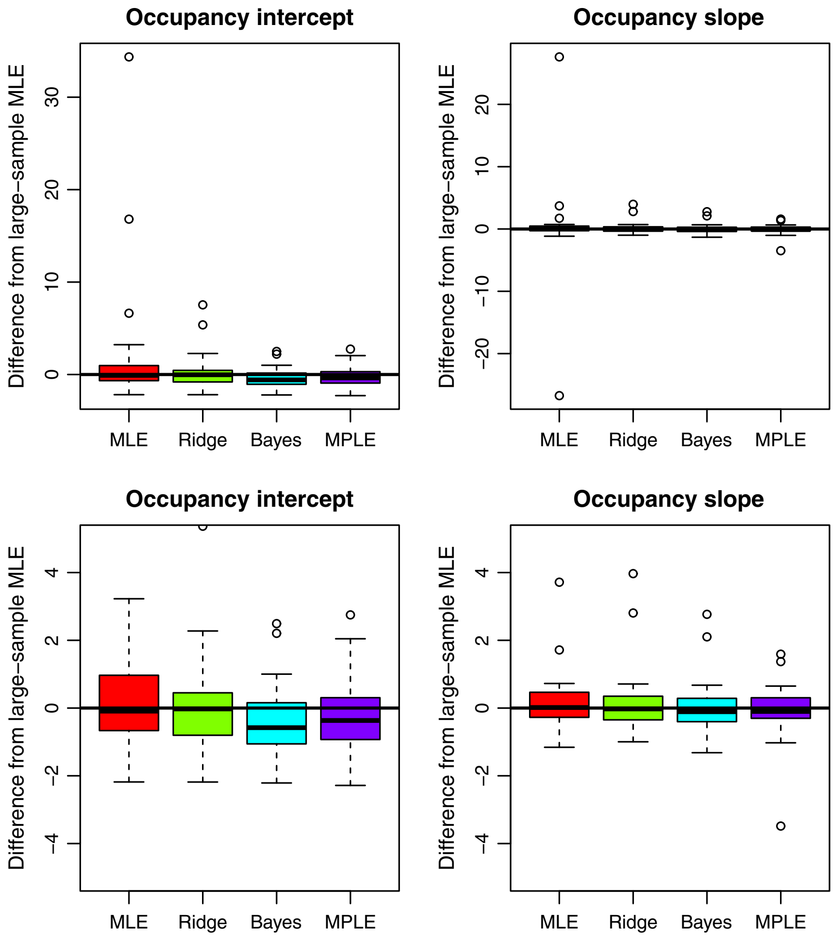

Fig. 5. Each plot shows differences between the large-sample MLEs and the reduced-sample estimates, across the four methods for the 25 species in the empirical case study. The bottom plots are the same as the top plots, zoomed in to [−5,5]. A few species produce extreme outlying MLEs. Penalized estimates on the reduced data set have lower variability than the MLEs and remain close to the large-sample estimates.

exhibits the least bias on larger data sets, whereas the bias incurred by *Bayes* decreases more slowly with increasing sample size in our experiments. The performance of *MPLE* varies, with few cases in which it is clearly the best method and several cases with relatively high error.

We see two potential concerns regarding the *MPLE* method (Moreno & Lele 2010). First, it requires fitting an occupancy model to compute the penalization weight λ. In extreme cases, the standard errors of the detection parameters required for this computation cannot be obtained. Secondly, *MPLE* regularizes towards the logistic regression parameters. Since a central motivation for occupancy modelling is to correct for the bias incurred by ignoring imperfect detection, using the biased estimates as a target for the penalty term is counter-intuitive. None of the penalization methods are guaranteed to regularize towards the true parameter values, but penalizing towards zero may be seen as more agnostic than using the logistic regression estimates.

In addition to the expected error of the estimates, several other considerations may arise in selecting an estimation method for occupancy modelling. In particular, confidence intervals are often computed for the parameter estimates. Penalized methods do not satisfy the assumptions necessary to compute confidence intervals based on standard asymptotic theory. Instead, uncertainty in the parameter estimates can be assessed using bootstrap methods. The bootstrapped confidence intervals for *MPLE* encapsulate the variation in the computed λ across data sets, and for *Ridge* and *Bayes*, the intervals account for the variation that results from using cross-validation to choose among the specified λs. Recent work has addressed the question of how to compute confidence intervals and *P* values in some penalized regression settings (Javanmard & Montanari 2013; Lockhart *et al.* 2013) without the bootstrap, and it may be possible to extend that work to occupancy models.

Ease of use is another consideration for evaluating estimation methods. The *Ridge* and *Bayes* methods require tuning the value of λ (i.e. cross-validation). The *MPLE* method requires some computation to set up (*e.g.* fitting a logistic regression), but it avoids the tuning process. All three penalized likelihood methods, including cross-validation for λ, will be included in the next release of the *unmarked* R package (Fiske & Chandler 2011).

We note that since *Bayes* frequently had the lowest error in our experiments, one might expect that a fully Bayesian treatment would also be competitive. For example, one could use WinBUGS to implement Gaussian priors, and it may be more natural in this setting to choose a value for σ^2 *a priori* or place a hyperprior on it. WinBUGS admits other forms of priors [i.e. non-Gaussian; Kéry (2010)], but the fast implementation used for the experiments above is specific to the Gaussian prior case.

The avian case study started with a relatively large data set for which maximum likelihood estimates of occupancy model parameters are unlikely to be problematic and compared against parameter estimates fit to a reduced sample. For some species, the MLEs produced from the reduced data set included extreme values, corresponding to boundary estimates for occupancy probabilities. Large parameter estimates, wide confidence intervals and boundary estimates could indicate problems with the fitted models for these species. A likely scenario in a maximum likelihood approach might then be to remove these species from the analysis altogether. Penalized methods provide an alternative to excluding these species from the study, since the penalized estimates are closer to the large-sample estimates (which are closer to the truth in a frequentist analysis) than the MLEs. One approach to a multispecies analysis would be to fit MLEs to all species, examine the fits, and try penalized estimators for problematic species instead of removing them. Alternatively, one could use a penalized method chosen *a priori* for all species. For species with outlying MLEs and boundary estimates on the reduced data set, penalized methods provide dramatic improvement over the MLEs. For species on which the MLEs perform well, the penalized estimates are similar to the MLEs. That is, for extreme cases, the penalized methods help substantially, and for average cases, the penalized methods do not hurt substantially. Cross-validation automatically produces this behaviour, because it selects small values of λ when less regularization is needed.

On synthetic data, the MLEs included boundary estimates infrequently when the occupancy rate was near 40% and more frequently when occupancy was closer to 10%. Rare species are common in ecology generally (MacArthur 1972), and species of conservation concern are particularly likely to be rare. The improved results produced by the penalized likelihood methods for the *sparse* data set suggest that use of these methods could expand the scope of problems to which occupancy modelling can be applied. Of course, penalized methods are not a substitute for careful sampling design and data collection, and they still require some minimum amount of information in the data in order to draw reasonable conclusions. We leave more detailed characterization of the limits of these methods to future work.

Ecological modellers face many challenges, including study design, sampling bias, model misspecification, and estimation error. This paper considers the problem of estimating model parameters as accurately as possible when the model is specified correctly and its assumptions are met in the data. As discussed in the literature, maximum likelihood estimation in occupancy models can be difficult even in this relatively ideal situation when sample sizes are small. For situations in which augmenting a given data set is infeasible, we have proposed three parameter estimation techniques based on penalized likelihoods that show improvement over the maximum likelihood method on a variety of data sets. We conclude that penalized likelihood methods are a useful tool for maximizing the utility of small data sets in occupancy studies.

Acknowledgements

We thank Marc Kéry for helpful comments. This research was funded in part by the National Science Foundation under Grant numbers CCF-1215950, ARC-0941748 and DEB-1050954. Additional funding was provided by the Department of Defense Strategic Environmental Research and Development Program (SERDP) Grant number RC-2121.

Data accessibility

Details of the synthetic data generation process are provided in the article. Data for the empirical case study are deposited in the Dryad repository: http://datadryad.org/resource/doi:10.5061/dryad.t40f2 (Hutchinson *et al.* 2015).

References

Agresti, A. (2013) *Categorical Data Analysis*, 3rd edn. John Wiley and Sons, pp. 14–16.

Bailey, L.L., Hines, J.E., Nichols, J.D. & MacKenzie, D.I. (2007) Sampling design tradeoffs in occupancy studies with imperfect detection: examples and software. *Ecological Applications*, **17**, 281–290.

Bailey, L.L., MacKenzie, D.I. & Nichols, J.D. (2013) Advances and applications of occupancy models. *Methods in Ecology and Evolution*, **5**, 1269–1279.

Betts, M.G., Rodenhouse, N.L., Sillett, T.S., Doran, P.J. & Holmes, R.T. (2008) Dynamic occupancy models reveal within-breeding season movement up a habitat quality gradient by a migratory songbird. *Ecography*, **31**, 592–600.

Fiske, I. & Chandler, R. (2011) unmarked: An R package for fitting hierarchical models of wildlife occurrence and abundance. *Journal of Statistical Software*, **43**, 1–23.

Guillera-Arroita, G. & Lahoz-Monfort, J.J. (2012) Designing studies to detect differences in species occupancy: power analysis under imperfect detection. *Methods in Ecology and Evolution*, **3**, 860–869.

Guillera-Arroita, G., Ridout, M.S. & Morgan, B.J. (2010) Design of occupancy studies with imperfect detection. *Methods in Ecology and Evolution*, **1**, 131–139.

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

Hastie, T., Tibshirani, R. & Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science, New York, USA.

Hoerl, A. & Kennard, R. (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Hutchinson, R.A., Valente, J.J., Emerson, S.C., Betts, M.G. & Dietterich, T.G. (2015) Data from: Penalized likelihood methods improve parameter estimates in occupancy models. *Methods in Ecology and Evolution*, doi:10.5061/dryad.t40f2.

Javanmard, A. & Montanari, A. (2013) Confidence intervals and hypothesis testing for high dimensional regression. *The Journal of Machine Learning Research*, **15**, 2869–2909.

Kéry, M. (2010) *Introduction to WinBUGS for Ecologists*. Academic Press, Burlington.

Kéry, M. (2011) Toward the modelling of true species distributions. *Journal of Biogeography*, **38**, 617–618.

Lockhart, R., Taylor, J., Tibshirani, R.J. & Tibshirani, R. (2013) A significance test for the lasso. arxiv.org/abs/1301.7161.

Lunn, D., Spiegelhalter, D., Thomas, A. & Best, N. (2009) The bugs project: Evolution, critique and future directions. *Statistics in Medicine*, **28**, 3049–3067.

MacArthur, R.H. (1972) *Geographical Ecology*. Harper and Row, New York, USA.

MacKenzie, D.I. & Royle, J.A. (2005) Designing occupancy studies: general advice and allocating survey effort. *Journal of Applied Ecology*, **42**, 1105–1114.

MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Royle, J.A. & Langtimm, C.A. (2002) Estimating Site Occupancy Rates When Detection Probabilities Are Less Than One. *Ecology*, **83**, 2248–2255.

MacKenzie, D.I., Nichols, J.D., Royle, J.A., Pollock, K.H., Bailey, L.L. & Hines, J.E. (2006) *Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence*. Elsevier, San Diego, CA, USA.

Moreno, M. & Lele, S.R. (2010) Improved estimation of site occupancy using penalized likelihood. *Ecology*, **91**, 341–346.

Morris, C. (1983) Parametric empirical bayes inference: Theory and applications. *Journal of the American Statistical Association*, **78**, 47–59.

Murphy, K.P. (2012) *Machine Learning: A Probabilistic Perspective*. The MIT Press.

Nakagawa, S. (2004) A farewell to Bonferroni: the problem of low statistical power and publication bias. *Behavioral Ecology*, **15**, 1044–1045.

R Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Welsh, A.H., Lindenmayer, D.B. & Donnelly, C.F. (2013) Fitting and interpreting occupancy models. *PLOS ONE*, **8**, e52015.

Received 18 December 2014; accepted 23 February 2015
Handling Editor: Robert Freckleton

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1 Synthetic data analyses.

Appendix S2 Empirical case study.