



Profiting from pilot studies: Analysing mortality using Bayesian models with informative priors

William K. Morris*, Peter A. Vesk, Michael A. McCarthy

School of Botany, University of Melbourne, Parkville 3010, Australia

Received 10 May 2012; received in revised form 6 November 2012; accepted 6 November 2012

Abstract

Pilot studies are often used to help design ecological studies. Ideally the pilot data are incorporated into the full-scale study data, but if the pilot study's results indicate a need for major changes to experimental design, then pooling pilot and full-scale study data is difficult. The default position is to disregard the preliminary data. But ignoring pilot study data after a more comprehensive study has been completed forgoes statistical power or costs more by sampling additional data equivalent to the pilot study's sample size. With Bayesian methods, pilot study data can be used as an informative prior for a model built from the full-scale study dataset. We demonstrate a Bayesian method for recovering information from otherwise unusable pilot study data with a case study on eucalypt seedling mortality. A pilot study of eucalypt tree seedling mortality was conducted in southeastern Australia in 2005. A larger study with a modified design was conducted the following year. The two datasets differed substantially, so they could not easily be combined. Posterior estimates from pilot dataset model parameters were used to inform a model for the second larger dataset. Model checking indicated that incorporating prior information maintained the predictive capacity of the model with respect to the training data. Importantly, adding prior information improved model accuracy in predicting a validation dataset. Adding prior information increased the precision and the effective sample size for estimating the average mortality rate. We recommend that practitioners move away from the default position of discarding pilot study data when they are incompatible with the form of their full-scale studies. More generally, we recommend that ecologists should use informative priors more frequently to reap the benefits of the additional data.

Zusammenfassung

Pilotstudien werden oft genutzt, um das Design von ökologischen Untersuchungen zu bestimmen. Idealerweise werden die Daten aus der Pilotstudie in den Datensatz der Hauptstudie inkorporiert, aber wenn die Pilotstudie die Notwendigkeit größerer Veränderungen an der Versuchsanlage anzeigt, ist das Zusammenführen von Daten aus Pilot- und Hauptstudie schwierig. Die normale Entscheidung ist dann, die vorläufigen Daten nicht zu berücksichtigen. Aber die Ergebnisse aus der Pilotstudie zu ignorieren, nachdem die Hauptstudie abgeschlossen wurde, bedeutet, auf Teststärke zu verzichten, oder der Aufwand steigt durch das Sammeln zusätzlicher Daten, die den Probenumfang der Pilotstudie ausgleichen. Mit Bayesschen Methoden können Daten aus der Pilotstudie als informative a-priori-Verteilung ('informative prior') für ein Modell genutzt werden, das aus dem Datensatz der Hauptstudie hergestellt wird. Wir demonstrieren eine Bayessche Methode zur Gewinnung von Information aus anders nicht nutzbaren Pilotstudienanhand einer Fallstudie zur Mortalität von Eukalyptussetzlingen. Eine Pilotstudie zur Mortalität von Eukalyptussetzlingen wurde 2005 in SO-Australien durchgeführt. Eine größere Studie mit einem modifizierten Design wurde im Folgejahr durchgeführt. Die beiden Datensätze unterschieden sich erheblich, so dass sie nicht ohne weiteres zusammengeführt werden konnten. A-posteriori-Schätzwerte der Modellparameter für die Pilotstudie wurden einem Modell für den zweiten,

*Corresponding author. Tel.: +61 3 8344 8086; fax: +61 3 9347 5460.
E-mail address: wkmor1@gmail.com (W.K. Morris).

größeren Datensatz zugrundegelegt. Die Überprüfung des Modells zeigte, dass die Hinzunahme einer informativen a-priori-Verteilung die Vorhersagekraft des Modells in Bezug auf die Trainingsdaten erhielt. Die Hinzunahme einer informativen a-priori-Verteilung verbesserte die Genauigkeit des Modells für die Vorhersage eines Validierungsdatensatzes und steigerte Genauigkeit und effektive Probengröße für die Bestimmung der durchschnittlichen Mortalitätsrate. Wir empfehlen, dass Praktiker von der Standardpraxis abrücken sollten, Daten aus Pilotstudien zu verwerfen, wenn diese mit ihrer Hauptstudie inkompatibel sind. Ganz allgemein empfehlen wir, dass Ökologen informative a-priori-Verteilungen häufiger einsetzen sollten, um die Vorteile zusätzlicher Daten zu nutzen.

© 2012 Gesellschaft für Ökologie. Published by Elsevier GmbH. All rights reserved.

Keywords: Hierarchical models; Value of information; Transplant experiment

Introduction

The ability of Bayesian analyses to formally incorporate prior information has been little exploited in ecological research despite its distinct appeal in a world where data and resources for research are limited and the impetus for rapid learning is great. Many textbooks on Bayesian methods for ecologists introduce the concept of informative priors in the first few pages (e.g., Kéry 2010; McCarthy 2007), yet researchers typically use very vague priors. In effect, these researchers assert that they have no prior knowledge of model parameters. Informally and formally, researchers use prior information to determine questions, sampling regimes, and model structures, and to interpret results. Whereas use of informative priors is rare, perhaps because it is hard to express prior knowledge as a probability distribution (Clyde 1999) or because informative priors are perceived as overly subjective (Dennis 1996), though subjectivity is not a requirement of priors (Hobbs & Hilborn 2006). Another reason why informative priors are not used is a fear that they could reduce model accuracy. A prior not only affects the precision of estimates, but also the location of the posterior and therefore, potentially the predictive accuracy.

Here we extend the domain of using informative priors in ecological modelling (see Choy, O’Leary, & Mengersen 2009; Dupuis & Joachim 2006; McCarthy & Masters 2005; McCarthy, Citroen, & McCall 2008) with pilot study data. The primary goal of a pilot study is to inform the design of the subsequent full-scale study. Pilot studies are small studies aimed to help reduce important uncertainties, and reveal the sample size needed to detect particular effects. Or, they can reveal major drivers of system variation and help identify at what spatial and temporal scales the variation is propagated, or help refine a set of predictors. However, the data generated by a pilot study may not simply inform the design of the full-scale study, but may also help address the fundamental research question. Many ecologists’ default position is to disregard the preliminary data, a stance recommended by texts on data collection and analysis (e.g. Green 1979). However, treating the results of a pilot study as an informative prior using Bayesian methods provides a formal and transparent way to combine two otherwise incompatible data sources, improving the cost-effectiveness of the research.

We illustrate the use of Bayesian informative priors to recover the inferential and predictive power of otherwise unusable pilot study data with a case study on eucalypt tree seedling mortality. Mortality rate is a key demographic parameter to be estimated. Understanding how and why it varies, is key to describing and learning about population dynamics of all species (Zens & Peart 2003). However, the mortality events from which rates are calculated are often rare in absolute terms, and in many systems may also be episodic. Combined, these issues make mortality rates difficult to characterise, so large datasets that include many individuals and span large spatial and temporal ranges are often required.

Including prior information in a Bayesian model will increase the precision of relevant parameter estimates and posterior predictive distributions (McCarthy 2007). The effect on model accuracy is harder to define though no less important and can only be done via model validation. Effects of priors on accuracy have received limited attention in ecology. If a prior disagrees with the likelihood then the model will be less accurate with respect to the training data than it would without the prior information. But the cost to predictive accuracy specific to the training data may be outweighed by the increased generality of the model, as it can include information from a wider range of sources than the training data alone. In this paper, we demonstrate how to treat the knowledge learned during a pilot study on eucalypt seedling mortality as an informative Bayesian prior. We express this source of prior knowledge as the degree to which the full-scale study budget would need to increase to recover the loss of information by not including the prior. The particular demonstration of using informative priors we provide here highlights their general benefit.

Methods

Seedling survival experiment design and analyses

During 2005–2009 a pilot and full-scale transplant survival experiment were undertaken at 21 sites on 14 farming properties in the Goulburn–Broken Catchment, Victoria, Southeastern Australia. The pilot study began in October 2005 when 54 Grey Box eucalypt (*Eucalyptus microcarpa*)

seedlings were planted in each of four grazing exclosures in a split plot design with two treatments. For the first treatment plants were watered during the first six months at fortnightly intervals while the second involved removing a metre squared area of the top layer of soil and vegetation before planting (scalping). These two treatments were fully crossed. Seedling survival was then monitored for the next two years with eleven revisits at irregular intervals ranging from one to eight months. The preliminary findings of the pilot study showed that the major drivers of variation in mortality rate occurred at the between-site (landscape) level and not the plot level. Further, the interaction of watering and scalping treatments was sub-additive and therefore did not increase survival when combined. For the larger-scale, 17-site experiment, the split plot design and water treatment were abandoned. At each site, 35 seedlings were planted with individual mesh tree guards instead of using a single exclosure. Approximately half the seedlings at each site were randomly assigned the scalping treatment and all seedlings were watered daily for the first two weeks.

The major component of eucalypt seedling mortality in this system is summer water stress (Semple & Koen 1996; Stoneman, Dell, & Turner 1994; Vesik & Dorrough 2006) so related variables were sought as predictors in models of these data. We used topographic wetness index (TWI) and average daily maximum temperature.

TWI is derived for each pixel of a digital elevation model as:

$$TWI = \log \left(\frac{A \tan \theta}{L} \right), \quad (1)$$

where A is the area of the contributing catchment, θ is the slope (in degrees) and L is the length of the contour the point sits on. TWI approximates the relative plant-available water given a site's position in the landscape (Moore, Gessler, Nielsen, & Peterson 1993). Based on a 25 m resolution digital elevation model of the Goulburn–Broken catchment, TWI varied twofold across the 17 study sites in the large-scale experiment (7–16 units). TWI was not used for the pilot study as there were only four sites.

Average maximum daily temperature was used to account for the climatic component of plant water stress. For this system we assumed that plant water deficit is strongly influenced by temperature, especially during summer, when air temperature is a major driver of evapotranspiration and is negatively correlated with precipitation. Average maximum daily temperature was calculated for each site-intercensus period combination from daily observations at the closest weather station (Australian Government Bureau of Meteorology, <http://www.bom.gov.au>). The distance from each site to its nearest station was 1–20 km with some sites sharing the same station.

The models for both the pilot and full-scale data were of the form:

$$\text{cloglog}(\text{Pr}(y_{ijk} = 1)) = \log(\lambda_{ijk} T_k), \quad (2)$$

where

$$\lambda_{ijk} = \exp \left(\eta_{ijk} + \sum_{m=0}^{k-1} \text{Pr}(y_{ij(m)} = 1) \right). \quad (3)$$

Here the probability that the i th seedling at the j th site dies during the k th time interval, $\text{Pr}(y_{ijk} = 1)$, is modelled with a complementary log-log link function (where, $\text{cloglog}(\text{Pr}(y_{ijk} = 1)) = \log(-\log(1 - (\text{Pr}(y_{ijk} = 1))))$) and is the log of the product of the instantaneous mortality rate, λ_{ijk} and the intercensus length T . For each case, λ_{ijk} is modelled using an equation with two components. The first component, η_{ijk} , includes the environmental, climatic and treatment effects and the second component accounts for the change in hazard of seedlings over time.

To accommodate varying time intervals we used the complementary log-log link function to relate finite mortality rate to intercensus length and the other components that predict mortality (Allison 1982). The probability that an individual at a site, dies in a given time period, $\text{Pr}(y_{ijk} = 1)$, is equivalent to the proportion of individuals with the same hazard that die or, M , the finite mortality rate. M is related to the time interval, T , according to exponential decay, i.e.,

$$M = 1 - e^{-\lambda T}. \quad (4)$$

Taking the complementary log-log of both sides of Eq. (4) yields Eq. (2), which accounts for variable intercensus lengths and makes the regression model equivalent to that used in a standard survival analysis.

Bias can be introduced into models of death data because hazard varies over the length of monitoring (Zens & Peart 2003). This is because susceptible individuals die first and as the monitoring continues the probability of dying for the remaining individuals is lower on average than for the original population. We included the second term in Eq. (3) to overcome this potential source of bias. The term modulates the expected mortality rate at a given site by accounting for the predicted probability of mortality in the previous census intervals. It does so by summing all of the probabilities of death during all the previous time periods (where at $k=0$, $\text{Pr}(y_{ij} = 1) = 0$). Therefore, if there was a high probability of mortality in the past, expected future mortality would be low. The influence of this term is fixed, effectively having a coefficient of -1 . The effect must be negative as mortality rate decreases according to the assumptions of the process. Choosing a fixed coefficient for the variable hazard term was necessary to make the model identifiable. The problem of identifiability arises because the term includes previous mortality as a function of the same variables used to estimate current mortality. If not fixed, such an effect would be highly correlated with the other coefficients, and therefore unidentifiable.

In the model for λ (Eq. (3)) the site, environment and treatment effects were modelled as

$$\begin{aligned} \eta'_{ijk} &= \mu' + \beta'_1 \cdot temp'_{jk} + \beta'_2 \cdot water_{ij} \\ &\quad + \beta'_3 \cdot water_{ij} \cdot scalp'_{ij} + (\beta'_4 + v'_j) \cdot scalp'_{ij} + \epsilon'_j \quad (5) \\ v'_j &\sim N(0, \sigma'_v) \\ \epsilon'_j &\sim N(0, \sigma'_\epsilon), \end{aligned}$$

for the pilot study, and as

$$\begin{aligned} \eta''_{ijk} &= \mu'' + \beta''_1 \cdot temp''_{jk} + \beta''_2 \cdot twi''_{ij} \\ &\quad + \beta''_3 \cdot temp''_{jk} \cdot twi''_{ij} + (\beta''_4 + v''_j) \cdot scalp''_{ij} + \epsilon''_j \quad (6) \\ \begin{pmatrix} v''_j \\ \epsilon''_j \end{pmatrix} &\sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma''_v \sigma''_v & \rho \sigma''_v \sigma''_\epsilon \\ \rho \sigma''_v \sigma''_\epsilon & \sigma''_\epsilon \sigma''_\epsilon \end{pmatrix} \right), \end{aligned}$$

for the full-scale study, where μ is the average mortality rate and the β 's are regression coefficients. In both models the effect of average maximum temperature (*temp*) and of the soil scalping treatment (*scalp*) were included. The effect of watering (*water*) was absent from the full-scale study as this treatment was not continued after the pilot study. The effect of topographic wetness was only modelled for the full-scale study as it was a site level covariate and the four sites used in the pilot study were too few to warrant its inclusion. The full-scale study also included a term for the interaction between topographic wetness and average maximum temperature. Each model also had two normally distributed, site-level error terms with a mean of zero, one for the effect of scalping, v_j , with a standard deviation of σ_v , and one for the average mortality rate, ϵ_j , with standard deviation σ_ϵ . For the full-scale study the correlation between these errors was modelled with a variance–covariance matrix and correlation coefficient, ρ .

Model fitting and comparison

To evaluate using pilot study data as an informative prior we fit two models to the full-scale study dataset. The first model had vague priors for all parameters, and the second model was the same except for a prior distribution for the average mortality, μ'' (intercept), derived from the posterior distribution generated by the pilot study model. The models were fit using the free software package OpenBUGS version 3.2.1 (Lunn, Spiegelhalter, & Thomas 2009). OpenBUGS uses a graphical modelling approach to fit Bayesian models with Markov Chain Monte Carlo (MCMC) methods (see Appendix A). All data preparation, post-processing of posterior samples and graphics production was done using R version 2.14 (R Development Core Team 2010).

Vague, normally distributed priors (mean = 0, standard deviation = 0.0001) were used for μ' in the pilot study and the full-scale model without an informative prior. The same vague prior was used for the regression coefficients for treatment and environment effects (β_1 , β_2 , β_3 and β_4) in both models. Weakly informative half-Cauchy priors with scale

parameters of 25 were used for σ'_v and σ'_ϵ to prevent over-inflated estimates of variance due to the small number of sites (Gelman 2006). Vague uniform priors in the interval 0–100 were used for the standard deviations of the site-level errors in the full-scale model. Similarly a vague uniform prior with the interval -1 to 1 was used for the correlation coefficient ρ .

When fitting each of the models, all continuous predictors were centred on their means and scaled by dividing by two times their standard deviations. Centring avoids correlation between parameters leading to better mixing of MCMC chains. Dividing by two standard deviations ensures that all variables are on the same scale as an equal probability binary variable. Scaling in this way aids the interpretation the effect sizes (Gelman 2008). MCMC chains were run for 5×10^8 iterations with a 1.5×10^8 burn-in with three independent chains for each model and assessed for convergence with visual inspections (see Appendix A).

We used multiple methods to evaluate and compare models built with and without informative prior. Some methods assess model precision, some assess model accuracy and some both. We compared the location and precision of the posterior distributions of model parameters expecting that informative priors should increase the precision of estimating average mortality with the mode shifted towards the mode of the informative prior. We also compared posterior predictive distributions of mortality to unobserved new sites with average environmental and climatic conditions. We produced posterior predictive distributions for new sites by simulating from the posterior distributions of the intercept and site-level variance terms. Where the posterior predictive distribution of a new site $\bar{\phi}$ for a model is:

$$\bar{\phi} \sim N(\mu'', \sigma''_\epsilon). \quad (7)$$

We also compared models via measures of overall performance. As the models predict a binary response we used the area under the receiver operator curve (or ROC statistic), a measure of model accuracy, to compare the two models' abilities to distinguish between death and survival. We used deviance (\bar{D}) and the deviance information criterion (DIC) to compare model predictive accuracy Spiegelhalter et al. (2002).

Posterior predictive checking also assesses model fit by combining precision and accuracy measurement. The premise of posterior predictive checking is that if the model fits the data then simulated data generated from the posteriors should look similar to the observed data (Gelman, Carlin, Stern, & Rubin 2004). Here we take the extra step of applying predictive checking not only to the training data but to an external validation dataset. We used both numerical and graphical posterior predictive checks.

For numerical posterior predictive checking, we used Bayesian p -values. To produce the simulated data for the posterior predictive checks, we first sampled from the joint posterior distribution of the model, θ , for each validation data

point, y_{ijk}^{val} a vector of death probabilities, ϕ_{ijk} given the input variables observed for each data point, i.e.,

$$\phi_{ijk} \sim p(\theta | y_{ijk}^{val}, X_{ijk}^{val}). \quad (8)$$

Then for each data point we simulated a set of death or survival events by taking random Bernoulli samples with the probabilities given by the elements of ϕ_{ijk} such that,

$$y_{ijk}^{sim} \sim \text{Bernoulli}(\phi_{ijk}), \quad (9)$$

to obtain a vector of simulated death data, y_{ijk}^{sim} . We then repeated this for $n = 1000$ simulations and compared the distribution of the n simulations of y_{ijk}^{sim} to the validation set y_{ijk}^{val} . Bayesian p -values were calculated as the tail-area probability that the observed proportion of individuals dying at each site was less than or equal to the simulated proportion of dead individuals after the six month simulations. Bayesian p -values close to zero or one indicate model predictions that are different from the observations while a p -value of 0.5 indicates an ideal model. A satisfactory model would include the observed proportions within the 95% credible intervals of the simulations.

The validation dataset of 159 seedlings at eight sites in three regions was collected in 2009–2010 in the north of the Goulburn–Broken catchment. Eucalypt seedlings at the eight sites were planted between Autumn and Spring of 2009 and living seedlings were tagged in October 2009. Their status was recorded during a revisit in April 2010. The seedlings included eight different species of eucalypts including Grey Box. Sites were within the range of topographic wetness index of the training data and the average maximum temperatures during the monitoring period were also similar. For the simulation we assumed that there were three revisits two months apart.

We used graphical posterior predictive checking to assess the fit of each model to the training data. For a subset of sites that span the range of site level mortality we compared the distribution of seedling deaths through time, to repeated simulations of seedling death conditional on the posterior distributions of the models and the observed weather conditions, wetness index and seedling treatments.

By comparing the uncertainty of the posterior distributions from the model with and the model without an informative prior we can assign a value to the inclusion of prior information. The uncertainty around parameters expressed as a proportion (i.e., bound between 0 and 1) can be summarised by a beta distribution. The beta distribution has the convenient property of being able to calculate the effective number of binomial samples, \bar{n} , from its mean, m and variance s^2 :

$$\bar{n} = \frac{m(1-m)}{s^2}. \quad (10)$$

By comparing \bar{n} for the models with and without informative priors, we calculated the proportional increase in sample size and applied this proportion to the sample-size-dependent costs of the full-scale study. This value is equivalent to the

amount of money that would have been required to be spent to equal the precision achieved by including the prior information.

Results

The average mortality rate was lower for the seedlings planted during the pilot study (13% per month, assuming an average maximum temperature of 28 °C), than for those planted for the full-scale study (39% per month, assuming an average maximum temperature of 28 °C). But in both cases there was much variation between sites. In the full-scale study topographically wetter sites tended to have lower rates of mortality. In both experiments, periods and places with higher average maximum temperatures had greater rates of mortality. Scalping the soil improved survival in both experiments and tended to have a greater effect at hotter and drier sites.

Including prior information from the pilot study increased the precision of the full-scale model but, more importantly, also increased the accuracy. Bayesian p -values were calculated to assess both models against the validation dataset. For models with and without informative priors the expected mortality rate was higher at all sites than was observed during the survey period, so all Bayesian p -values were less than 0.5 (Fig. 1). Note that a Bayesian p -value of 0.5 means that the median prediction and observation are the same. Higher values indicate underprediction and lower values indicate overprediction. The 95% credible bounds of the model simulation encapsulated the observed mortality in all cases with p -values in the range 0.09–0.31 for the uninformed model indicating that the observed mortality could be reasonably expected. But at all sites the posterior predictive check indicated that the model using informative priors gave a marginally more accurate estimate of mortality rate with the Bayesian p -values in the range 0.10–0.34.

Including prior information from the pilot study in the model of the full-scale study shifted the model parameter posterior distributions and posterior predictive distributions, and in some cases, as expected, increased their precision (Figs. 2 and 3). Considering the expected mortality rate of the seedlings in the full-scale dataset (the intercept term) the advantage gained by including prior information amounts to an increase in the effective sample size of 51%, from 5.1 to 7.7 binomial samples. The actual number of planted seedlings was 595 so adding the prior information was equivalent to planting an extra 303 seedlings. The cost of planting a seedling is approximately \$8 (Australian), a quarter of which is the cost of the seedling and tree guard and the remaining cost includes employing someone to plant them. So to increase the sample size to the level equivalent to the full-scale study plus the pilot study prior would cost \$2428 more than the actual expenditure of \$4760.

Including prior information from the pilot study did not modify the estimated effects of scalping treatment, climate and topographic wetness. The precision for these parameters

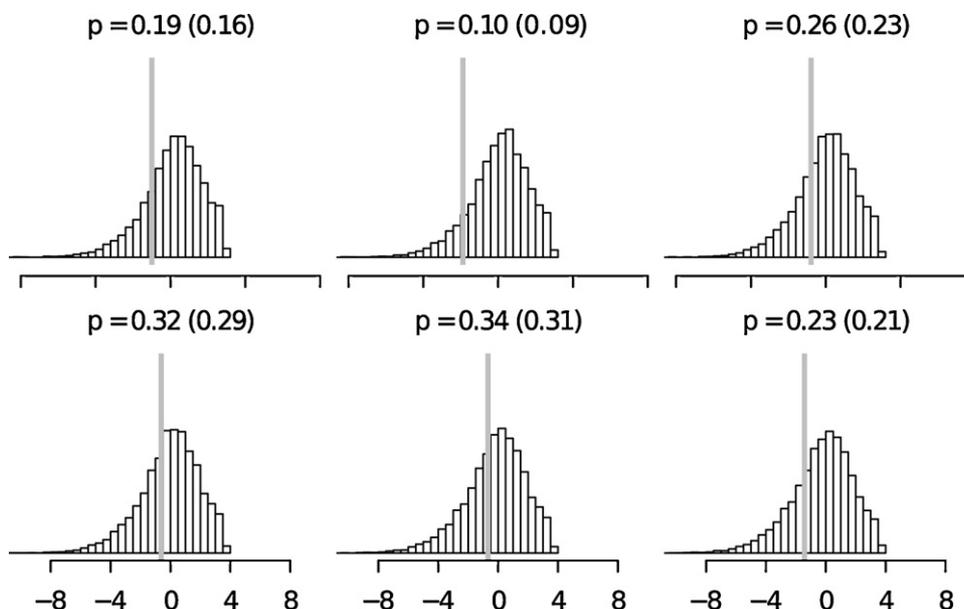


Fig. 1. Posterior predictive checks of the model against six validation sites. Grey lines show the observed mortality rate on the complementary log-log scale while histograms represent simulated probabilities from the posterior predictive distribution of the model using informative priors. Above each figure is the Bayesian p -value and in brackets the p -value for the site compared to equivalent simulations from the uninformed model.

also did not change. For the expected mortality rate and the predicted mortality rate at a new site were slightly lower when prior information from the pilot study was included (Fig. 2). However, in both cases the uncertainty in the informed and uninformed posteriors overlapped considerably.

Informed and uninformed models can be examined by comparing specific parameters and posterior predictive distributions, but also by comparing the overall fit of the model to the training data. Both models had a deviance of 1045. But comparing the models using DIC the model with informative priors had a lower score, 1084 compared to 1086, and is therefore the model with marginally greater support.

ROC scores measure the rate at which model predictions discriminate between observed binary outcomes. In essence the ROC score is the proportion of allocations of higher

probabilities assigned to successes when a success and a failure are randomly chosen from the dataset (Hanley & McNeil 1982). ROC scores for both the informed and uninformed models were high, 0.91, meaning that both models had more or less equal power to discriminate between deaths and survivals.

We employed a graphical posterior predictive check (Fig. 4) to assess the fit of the informed and uninformed models to the training data. Comparing the observed data to the replicated datasets using either model appears to conform broadly to reality. Both models produce simulations that

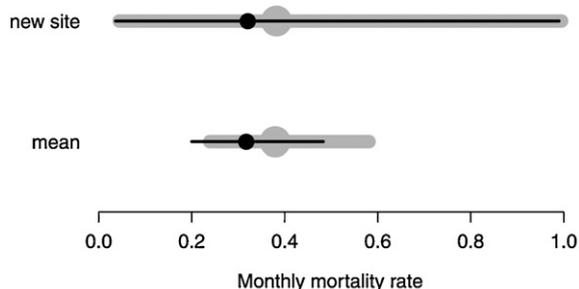


Fig. 2. Comparison of the posterior distribution of the parameter μ'' (mean) and the posterior predictive distribution of the expected monthly mortality rate at a new site for the uninformed (grey background) and informed (dark foreground) models of the full-scale study. Dots indicate the posterior means while bars span one posterior standard deviation either side.

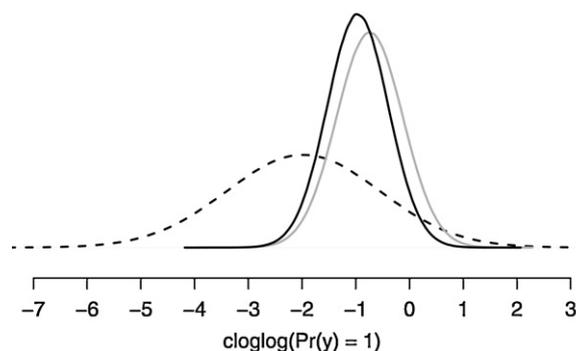


Fig. 3. Prior (dashed line), likelihood (grey line) and posterior (thick line) distributions of the average monthly mortality rate of Grey Box seedlings on the complementary log-log scale, given average modelled climatic and site conditions. The prior is derived from the posterior predictive distribution of the pilot study model, while the likelihood and posterior correspond to the posterior distributions of the parameter μ from the uninformed and informed models of the full-scale study, respectively.

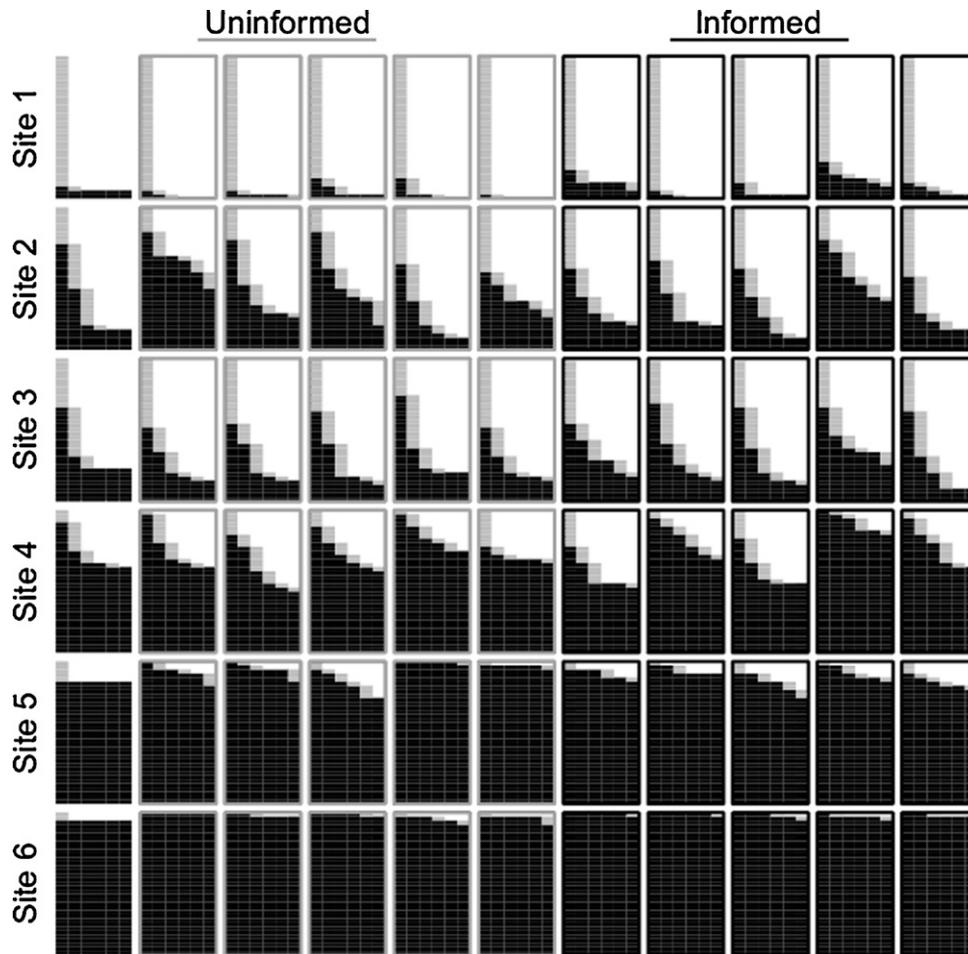


Fig. 4. Graphical posterior predictive check comparing repeated simulations based on the models with and without an informative prior to the observed mortality during monitoring of full-scale study at a subset of the site. Each row of matrices represents a different site of initially 35 seedlings. The first column of matrices is the observed mortality at each site and the remaining ten are simulations based on the joint posterior distribution of the model. The five columns with grey outlines are simulations from the model without the pilot study informative prior included while the remaining five with black outlines are simulations from the models which have the informative prior. Each matrix has 35×6 cells representing the status of each seedling during each monitoring period where the matrix rows represent seedlings and columns successive monitoring periods. Black cells indicate the seedling survived the monitoring period while grey cells indicate that the seedling died sometime during the interval.

have similar patterns to the real dataset across the range of sites with different site-level mortality rates.

Discussion

The effect of priors on model precision is well known and documented (e.g. McCarthy & Masters 2005). However, the effect on accuracy is rarely examined, if at all. We know of no such examples in the ecological literature. Here, the observed small improvement in accuracy may be due to the increase in model generality and scope achieved by including the prior derived from the pilot study. This represents a general benefit of including prior information, and in particular from a pilot study, as it expands the temporal coverage of the study. Unlike valuing improvements in precision of given parameters, which can be quantified in terms of changes

in effective sample size, it is hard to quantify the value of increased generality, though it is no less important. However, shifting posterior modes may also represent a caveat of using priors. If the data used for the prior were biased with respect to variation through time then shifting the posterior towards it would reduce the generality of the model. The shift in expected mortality rate when including pilot study information was due to the lower mortality rate observed in the pilot study. This may have been in part because conditions during the pilot study were less harsh and were not completely summarised by modelling the effect of temperature or due to the characteristics of sites chosen for the pilot study.

Others have noted the value of including prior information in a Bayesian analyses. McCarthy and Masters (2005) showed that they could reduce the length of a study of European dipper mortality by three years and achieve the same accuracy and precision by including prior information

gleaned from the relationship between body-mass and mortality for a suite of other birds. Similarly, Garrard et al. (2012) demonstrate increases in effective sample size when including prior information in models of bird natal dispersal. The value of including prior information in a model depends on what parameters or posterior predictive distributions are in focus. Here we have shown that the estimate of expected mortality rate of seedlings increases precision when including the pilot prior. Including the prior information for the expected mortality rate of seedlings at a new site did not increase precision (Figs. 2 and 3). This was because sites vary greatly and little information was learned about this variability from the pilot study that could then be used to inform the model of the full-scale study. Thus, the value of any given prior varies depending on the focus of the predictions or inferences.

Bayesian statistics has sometimes been criticised as not suitable for ecology (e.g. Dennis 1996; Lele & Dennis 2009). Much of this sentiment stems from the idea that Bayesian priors are overly subjective. Subjective priors can be used in Bayesian models and have been demonstrated to work effectively when data are scarce and often when there are experts from which to elicit information (e.g. Choy et al. 2009; Martin, Kuhnert, Mengersen, & Possingham 2005). But a prior does not have to be subjective. There are examples of ecological studies where the same level of objectivity used to collect the model training data were used to construct the prior (e.g. Dupuis & Joachim 2006; McCarthy & Masters 2005; McCarthy et al. 2008). Using pilot study data to form a prior is another example of an objective prior.

Comparing DIC for the model with an informative prior to the model with an informative prior showed greater support for including the informative prior. However, the difference in DIC was within the range normally treated as equally supportive of both models. DIC is affected by the prior in two ways. The deviance component is a measure of the predictive error of a model with respect to its training data. The prior can affect the predictive error of a model by changing the posterior distribution. In this case both models had equivalent predictive error. But DIC also accounts for the number of parameters in a model—the complexity. More complex models have more parameters and greater DIC when predictive error is equal (Spiegelhalter et al. 2002). In a Bayesian model the effective number of parameters depends on the priors. The more information in the joint prior distribution, the less parameters effectively need to be estimated and the lower the DIC will be. Here the informative prior accounted for about two parameters explaining a DIC.

Conclusion

We have demonstrated how preliminary data can be used as an informative Bayesian prior. A key finding of this study is that including the prior information increased the precision of some parameters at the same time as improving or at least not compromising the model's predictive accuracy.

As well as changing the precision, including prior information also changed the location of some posterior distributions. Changing the location could be beneficial or undesirable depending on what is driving the disagreement between prior and likelihood—beneficial if the shift is due to the prior capturing variability associated with the increased scale of the study, but undesirable if it is driven by a biased and therefore irrelevant prior. We recommend that practitioners move away from the default position of discarding pilot study data when it is incompatible with the form of their full-scale studies. Instead, in many cases preliminary data should be used as an informative prior in a Bayesian model. Using pilot study data as an informative prior highlights the benefit of informative priors more generally. In many instances, prior information will be available that is as objective and at the same time less costly to acquire than new data. If such information is incorporated into ecological models as informative priors, the cost of doing ecology will decrease.

Acknowledgements

We thank Libby Rumpff, Megan Watson, James Camac, Chris Jones, Rhiannon Apted, Warwick McCallum and Alex Thompson for help in the field. We also acknowledge the assistance of Carla Miles (Goulburn Broken Catchment Management Authority), Kate Hill (Department of Sustainability and Environment) and the land owners who allowed us to undertake this study on their properties. We also thank Bob O'hara, Rod Fensham, and anonymous reviewers for helpful comments on earlier versions of this manuscript. William K. Morris was funded by an Australian Postgraduate Award. This research was supported by the Australian Research Council (DP0985600) and the Goulburn Broken Catchment Management Authority.

Appendix A. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.baae.2012.11.003>.

References

- Allison, P. (1982). Discrete-time methods for the analysis of event histories. *Sociological Methodology*, 13, 61–98.
- Choy, S., O'Leary, R., & Mengersen, K. (2009). Elicitation by design in ecology: Using expert opinion to inform priors for Bayesian statistical models. *Ecology*, 90, 265–277.
- Clyde, M. (1999). Bayesian model averaging: A tutorial: Comment. *Statistical Science*, 14, 401–404.
- Dennis, B. (1996). Discussion: Should ecologists become Bayesians? *Ecological Applications*, 6, 1095–1103.

- Dupuis, J. A., & Joachim, J. (2006). Bayesian estimation of species richness from quadrat sampling data in the presence of prior information. *Biometrics*, *62*, 706–712.
- Garrard, G. E., McCarthy, M. A., Vesk, P. A., Radford, J. Q., & Bennett, A. F. (2012). A predictive model of avian natal dispersal distance provides prior information for investigating response to landscape change. *Journal of Animal Ecology*, *81*, 14–23.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*, 515–533.
- Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, *27*, 2865–2873.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis. Texts in Statistical Science*. Boca Raton: Chapman and Hall/CRC.
- Green, R. H. (1979). *Sampling design and statistical methods for environmental biologists*. New York: John Wiley and Sons.
- Hanley, J. A., & McNeil, B. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*, 29.
- Hobbs, N. T., & Hilborn, R. (2006). Alternatives to statistical hypothesis testing in ecology: A guide to self teaching. *Ecological Applications*.
- Kéry, M. (2010). *Introduction to WinBUGS for ecologists*. Amsterdam: Academic Press.
- Lele, S., & Dennis, B. (2009). Bayesian methods for hierarchical models: Are ecologists making a Faustian bargain? *Ecological Applications*, *19*, 581–584.
- Lunn, D., Spiegelhalter, D., & Thomas, A. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, *28*, 3049–3067.
- Martin, T. G., Kuhnert, P. M., Mengersen, K., & Possingham, H. P. (2005). The power of expert opinion in ecological models using Bayesian methods: Impact of grazing on birds. *Ecological Applications*, *15*, 266–280.
- McCarthy, M. A. (2007). *Bayesian methods for ecology*. Cambridge: Cambridge University Press.
- McCarthy, M. A., Citroen, R., & McCall, S. C. (2008). Allometric scaling and Bayesian priors for annual survival of birds and mammals. *The American Naturalist*, *172*, 216–222.
- McCarthy, M. A., & Masters, P. (2005). Profiting from prior information in Bayesian analyses of ecological data. *Journal of Applied Ecology*, *42*, 1012–1019.
- Moore, I., Gessler, P., Nielsen, G., & Peterson, G. (1993). Soil attribute prediction using terrain analysis. *Soil Science Society of America Journal*, *57*, 443–443.
- R Development Core Team. (2010). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Semple, W. S., & Koen, T. B. (1996). Effect of seedbed on emergence and establishment from surface sown and direct drilled seed of *Eucalyptus* spp. and *Dodonaea viscosa*. *The Rangeland Journal*, *19*, 80–94.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. R., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal Of The Royal Statistical Society Series B-Statistical Methodology*, *64*, 583–616.
- Stoneman, G. L., Dell, B., & Turner, N. C. (1994). Mortality of *Eucalyptus marginata* (jarrah) seedlings in Mediterranean-climate forest in response to overstorey, site, seedbed, fertilizer application and grazing. *Australian Journal of Ecology*, *19*, 103–109.
- Vesk, P., & Dorrough, J. (2006). Getting trees on farms the easy way? Lessons from a model of eucalypt regeneration on pastures. *Australian Journal of Botany*, *54*, 509–519.
- Zens, M., & Peart, D. (2003). Dealing with death data: Individual hazards, mortality and bias. *Trends in Ecology & Evolution*, *18*, 366–373.

Available online at www.sciencedirect.com

SciVerse ScienceDirect