Ecology, 97(7), 2016, pp. 1746–1758

Bayesian cross-validation for model evaluation and selection, with application to the North American Breeding Bird Survey

William A. Link[1] and John R. Sauer

USGS Patuxent Wildlife Research Center, Laurel, Maryland 20708 USA

Abstract. The analysis of ecological data has changed in two important ways over the last 15 years. The development and easy availability of Bayesian computational methods has allowed and encouraged the fitting of complex hierarchical models. At the same time, there has been increasing emphasis on acknowledging and accounting for model uncertainty. Unfortunately, the ability to fit complex models has outstripped the development of tools for model selection and model evaluation: familiar model selection tools such as Akaike's information criterion and the deviance information criterion are widely known to be inadequate for hierarchical models. In addition, little attention has been paid to the evaluation of model adequacy in context of hierarchical modeling, i.e., to the evaluation of fit for a single model. In this paper, we describe Bayesian cross-validation, which provides tools for model selection and evaluation. We describe the Bayesian predictive information criterion and a Bayesian approximation to the *BPIC* known as the Watanabe-Akaike information criterion. We illustrate the use of these tools for model selection, and the use of Bayesian cross-validation as a tool for model evaluation, using three large data sets from the North American Breeding Bird Survey.

Key words: Bayesian analysis; Bayesian predictive information criterion; cross-validation; hierarchical models; North American Breeding Bird Survey.

Introduction

It would be nice if there were a simple, intuitive and reliable means for comparing and selecting among complex hierarchical models. Akaike's information criterion (*AIC*) has been widely used in wildlife studies and is easily computed in context of maximum likelihood estimation, but in its simplest form is not appropriate for comparing models with random effects. Modifications of *AIC* for hierarchical models have been proposed "but these are seldom used due to stability problems and computational difficulties" (Gelman et al. 2014). The deviance information criterion (*DIC*) is easily computed in Bayesian analysis, but faces similar difficulties in extension to random effects models (Celeux et al. 2006, Plummer 2008, Gelman et al. 2014).

Hooten and Hobbs (2015) note that "despite the seeming consensus among ecologists and wildlife biologists in how to perform model selection and multimodel inference, it is far from settled among statisticians." We agree with this assessment, and would add that these tasks involve a fair bit of art, that they are not accomplished by proscriptive means. Hooten and Hobbs (2015) provide a comprehensive summary of relevant methods, and guidelines on choosing among them.

In this note, we describe Bayesian cross-validation, one of the most useful tools for model selection and multimodel inference. Bayesian cross-validation can be used for assessing goodness of fit; it is also used to define a Bayesian analog of *AIC* called the Bayesian predictive information criterion (*BPIC*; Gelman et al. 2014). We also describe a convenient approximation to *BPIC*, called the Watanabe-Akaike information criterion (*WAIC*; Watanabe 2009, 2010, 2013, Gelman et al. 2014, Hooten and Hobbs 2015).

Cross-validation is appealing in its conceptual simplicity, but computationally intensive. The idea is to evaluate how well models predict new data. We set aside part of our data, pretending that it has not yet been observed, fit the models to the rest of the data, and use the results to predict the data we set aside; the data that were set aside are then compared with the predictions. The procedure is performed repeatedly, each time setting aside a different subset of the data.

Our focus in this paper is on the North American Breeding Bird Survey (BBS), a large and highly structured data set, consisting of annual counts since 1966 of >400 species of birds on >4,700 routes. A standard analysis has been presented annually based on a model developed by Link and Sauer (2002; results *available online*).[2] The model was developed to be widely applicable, useful for many species and at varying geographic scales. It provides a basis for estimating long term population change for species with low abundances and sparse data, and for identifying more complex patterns of population change for species with better data.

Model selection in our view inevitably involves a number of more or less subjective choices; our experience with BBS data has taught us that generating species- and

Manuscript received 10 July 2015; revised 4 September 2015; accepted 11 September 2015. Corresponding Editor: E. G. Cooch.

[1]E-mail: wlink@usgs.gov

[2] www.mbr-pwrc.usgs.gov/bbs/

scale-specific analyses for each of the many species involved is neither feasible nor desirable. The omnibus analysis presently used serves its purpose well. Nevertheless, for in-depth studies of individual species, alternative models for BBS should be considered. The cross-validation methods presented here provide a basis for model assessment and selection that we believe is appropriate for the BBS and other complex data sets and models.

North American Breeding Bird Survey

North American Breeding Bird Survey data are total counts of birds seen or heard at 50 stops along fixed 24.5-mile (39.4-km) roadside routes. For an account of the history of the BBS, details on count protocols and data analyses, and comprehensive summary of results through 2011, see Sauer et al. (2013) and supplemental material *available online*[3]. Data analyses are structured by geographic strata, within which species' abundance and patterns of change are assumed to be homogeneous.

North American Breeding Bird Survey counts are conducted by skilled observers, but are not censuses (complete counts of closed populations). Variation in counts among observers is evident, and there is strong evidence of change in the observer pool through time (Sauer et al. 1994). The patterns of change are consistent with the assertion that new observers tend to be better birders than those they replace: they count more individuals of species that are difficult to observe and identify, and fewer individuals of common and easily identified species. Change among observers must be modeled; left unmodeled, it is a clear source of bias in estimating population change.

In addition to change among observers, there is also evidence of change within observers. For instance, all other things being held equal, an observer's first count tends to be 5–10% smaller than expected, perhaps due to unfamiliarity with the survey and route. Roughly a quarter of the observers provide counts over periods >12 years, and 10% for periods >23 years (Link and Sauer In press). It might be anticipated that an observer's counts change due to age-related effects, such as hearing loss. Some evidence exists that an observer's number of years of service can be a useful surrogate for age, to avoid biasing estimation of population change.

Most observers (~60%) count on only one BBS route; >90% count on three or fewer routes (Link and Sauer In press). For those that provide counts on more than one route, there is often little temporal overlap in the counts on different routes: the observer has switched routes. Consequently, observer effects are treated as nested within routes.

North American Breeding Bird Survey data for single species are usually analyzed with overdispersed Poisson regression models. Count C_{sry} (indexed by stratum, route, and year) has mean λ_{sry} modeled as

$$\log(\lambda_{sry}) = A_s + \boldsymbol{X}_{o(r,y),y}\boldsymbol{\eta} + \omega_{o(r,y)} + \gamma_{sy} + \epsilon_{sry} \qquad (1)$$

[3] www.fwspubs.org/doi/abs/10.3996/nafa.79.0001

where A_s describes baseline abundance for stratum s; subscript $o(r,y)$ indicates the observer that produced the count on route r in year y; $\mathbf{X}_{o,y}$ is a vector of covariates for observer o in year y, and η is a corresponding vector of fixed effects; ω_o is a random effect associated with observer o, $\omega_o \sim N(0,\tau^{\omega})$ (meaning that ω_o is a mean zero normal random variable with precision parameter τ^{ω}); γ_{sy} is the departure from stratum s baseline abundance in year y (with $\gamma_{sy_0} = 0$ for some year y_0, as an identifiability constraint); and $\epsilon_{sry} \sim N(0,\tau^{\epsilon})$ is a random effect inducing extra-Poisson variation to the counts.

Qoholeth said "Furthermore, my son, be admonished: of making many books there is no end; and much study is a weariness of the flesh," a sentiment that can be applied to model-based analysis as well. Of making many models there is no end. Consider the possibilities simply in context of Eq. 1:

1) Models for observer-specific covariates $\mathbf{X}_{o,y}$:
   a) First year of service effect.
   b) Year of service as factor.
   c) Year of service modeled as linear, quadratic, etc.
   d) Year of service modeled by smoother (first differences, spline models).
2) Variation in random effects (models for τ^{ϵ} and τ^{ω}) :
   a) Spatially homogeneous.
   b) Stratum specific, with fixed effect parameters τ^{ϵ}_s and τ^{ω}_s.
   c) Stratum specific, with hierarchically structured parameters: τ^{ϵ}_s and τ^{ω}_s treated as sampled from lognormal distributions.
   d) Structured by abundance, e.g., $\tau^{\epsilon}_s \propto 1/A_s$ or $\tau^{\epsilon}_s \propto 1/A_s^2$.
3) Pattern in population change ("trajectory parameters" γ_{sy}) :
   a) Unstructured model: γ_{sy} unknown constants, with $\gamma_{sy_0} \equiv 0$ for baseline year y_0.
   b) Temporally stationary random effects: $\gamma_{sy} \sim N(0,\tau^{\gamma})$, or $\gamma_{sy} \sim N(0,\tau^{\gamma}_s)$.
   c) Linearly trending random effects: $\gamma_{sy} \sim N(\beta_s(y-y_0),\tau^{\gamma})$.
   d) First difference random effects: $\gamma_{sy} - \gamma_{s,y-1} \sim N(0,\tau^{\gamma})$.

And so on. With such a welter of possibilities, how should models be chosen for BBS data? We propose that the choice be made using Bayesian cross-validation tools, as we now describe.

Cross-Validation

Cross-validation refers to a family of techniques for evaluating the predictive performance of a model. Suppose that we fit a model to a set of data available today. One month from now, new data will be obtained, generated by the same basic mechanisms as the data presently available. We might have to wait to see covariate values associated with the new data, but we will be able

to use the present analysis, along with the covariate values once available, to predict the new data. Predictions disagreeing with realized values would indicate a lack of fit for our model; alternative models could be compared based on their typical performance. The new data allow an out-of-sample predictive check.

Cross-validation is precisely in this spirit, but based on the data presently available. The data are split: part for fitting the model and part for evaluating the model; the process is usually repeated for a prescribed set of data splits.

The simplest form of cross-validation is leave-one-out cross-validation (LOOCV). Suppose we have observations Y_i with associated covariate vectors X_i, for $i = 1,2, \ldots, n$, and a model M. Let $\boldsymbol{D}$ denote the full data set, $\boldsymbol{D} = \{(Y_i,X_i); i = 1,2, \ldots, n\}$ and let $\boldsymbol{D}_{-j}$ denote the data set remaining after removing observation Y_j. We fit M to $\boldsymbol{D}_{-j}$ and, pretending we do not know the value Y_j, use the fitted model to predict the observation Y_j based on its covariate vector X_j. Comparing the prediction Y_j^* to the actual value Y_j provides us with an indication of how well model M predicts observations, at least those with covariate vector X_j. To get an overall assessment of model M's predictive ability across covariate values, we can repeat the process across all values $j = 1,2,\ldots, n$ and compute a statistic like

$$R_M = \frac{1}{n}\sum_{j=1}^{n}(Y_j^* - Y_j)^2 \tag{2}$$

the average squared prediction error. Given models M_1 and M_2, $R_{M_1} < R_{M_2}$ can be taken as evidence that M_1 provides better predictions across the range of covariates considered.

Cross-validation allows an assessment of out-of-sample predictive performance. Predictions Y_j^* of values Y_j with covariate vectors X_j are made without allowing the actual observation to inform the model, thus avoiding the pitfall of overfitting.

Bayesian Cross-Validation

Prediction is very naturally handled under the Bayesian framework, making cross-validation an appealing technique for use in Bayesian analysis. Prediction is equivalent to evaluating missing data, and easily implemented in the BUGS packages (OpenBUGS, WinBUGS, or JAGS) as we describe subsequently.

Formally, we begin with priors on parameters $[\boldsymbol{\theta}]$ and sampling distributions defining $[\boldsymbol{D}|\boldsymbol{\theta}]$. Posterior distributions of unknown quantities are proportional to the product $[\boldsymbol{D}|\boldsymbol{\theta}][\boldsymbol{\theta}]$; treating Y_j as unobserved, we have

$$[Y_j,\boldsymbol{\theta}|\boldsymbol{D}_{-j},X_j] \propto [\boldsymbol{D}|\boldsymbol{\theta}][\boldsymbol{\theta}].$$

Marginalizing over $\boldsymbol{\theta}$, we obtain the predictive distribution $[Y_j|\boldsymbol{D}_{-j},X_j]$. It will be convenient to describe this distribution by the corresponding predictive distribution function $p_j(y|\boldsymbol{D}_{-j},X_j)$.

A cross-validation run omitting observation Y_j using the BUGS programs is accomplished by an easy modification of the analysis based on the full data $\boldsymbol{D}$: simply set Y_j to missing ("NA") in the data file. Adding Y_j as a node to be monitored provides a sample of values from the predictive distribution $p_j(y|\boldsymbol{D}_{-j},X_j)$.

Bayesian LOOCV consists of comparing observations Y_j with their predictive distributions. The predictive distribution tells us what we would expect for the observation based on the model, the covariate, the other observations, and the priors on parameters. In considering a single model, predictive distributions can be used to assess goodness of fit; with multiple models, they can be used for model selection.

Understanding and calculating the predictive distribution

It is helpful to think of $p_j(y|\boldsymbol{D}_{-j},X_j)$ as our best guess at the modeled probability distribution of observation Y_j, namely $p(y|\boldsymbol{\theta},X_j)$. If the parameters $\boldsymbol{\theta}$ were known, goodness of fit and model selection would be conducted by comparing observations with modeled distributions $p(y|\boldsymbol{\theta},X_j)$. Indeed, if all parameters were known for two models, the likelihood ratio for model selection would be

$$\mathrm{LR}^{(1,2)} = \prod_{j=1}^{n}\frac{p^{(1)}(Y_j|\boldsymbol{\theta}^{(1)},X_j)}{p^{(2)}(Y_j|\boldsymbol{\theta}^{(2)},X_j)} \tag{3}$$

here, we have used superscripts (m), $m = 1,2$ for models. Note that Eq. 3 evaluates the distribution functions $p^{(m)}(y|\boldsymbol{\theta}^{(m)},X_j)$ at the specific value $y = Y_j$.

But θ is unknown. Some goodness of fit and model selection procedures are based on substituting an estimate of θ in the formula for the modeled distribution: *AIC*, for instance, involves calculations based on $p(y|\hat{\boldsymbol{\theta}},X_j)$, where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator.

Instead of simply plugging in an estimate, Bayesian LOOCV accounts for uncertainty about $\boldsymbol{\theta}$ by averaging $p(y|\boldsymbol{\theta},X_j)$ over reasonable values of $\boldsymbol{\theta}$. Specifically, we average $p(y|\boldsymbol{\theta},X_j)$ against the posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{D}_{-j}$, obtaining

$$p_j(y|\boldsymbol{D}_{-j},X_j) = \int p(y|\boldsymbol{\theta},X_j)\,\pi(\boldsymbol{\theta}|\boldsymbol{D}_{-j},X_j)\,\mathrm{d}\boldsymbol{\theta} \tag{4}$$

$p_j(y|\boldsymbol{D}_{-j},X_j)$ estimates $p(y|\boldsymbol{\theta},X_j)$, the distribution function for observation Y_j. Averaging against $\pi(\boldsymbol{\theta}|\boldsymbol{D}_{-j},X_j)$ means that we are using all of the available information to inform us about the distribution of Y_j, but not cheating by using the observation Y_j itself.

It is rarely the case that one can calculate $\pi(\boldsymbol{\theta}|\boldsymbol{D}_{-j},X_j)$, much less that one can calculate the integral in Eq. 4. However, as noted before, a cross-validation MCMC run omitting Y_j can be conducted in BUGS by setting Y_j to "NA" in the data file, and adding Y_j as a node to be monitored. We will use notation Y_j° to distinguish a sampled value of this node, as distinct from the omitted value Y_j. The mean, standard deviation and percentiles

of $p_j(y|\boldsymbol{D}_{-j},X_j)$ can then be approximated by corresponding features of the sampled values Y_j°; for goodness of fit, we might consider tail values $\Pr(Y_j^\circ \geq Y_j|\boldsymbol{D}_{-j},X_j)$ calculated as the proportion of values $Y_j^\circ \geq Y_j$.

With discrete data, one can estimate $p_j(y|\boldsymbol{D}_{-j},X_j)$ by the sample proportion of values $Y_j^\circ = y$. However, we can do better: the predictive distribution can be approximated with greater efficiency using the functional form of $p(y|\boldsymbol{\theta},X_j)$. For the BBS models we consider

$$p_j(y|\boldsymbol{\theta},X_j) = \frac{\lambda(\boldsymbol{\theta},X_j)^y \exp\left(-\lambda(\boldsymbol{\theta},X_j)\right)}{y!}$$

is the Poisson distribution function, with parameter $\lambda(\boldsymbol{\theta},X_j)$. If we monitor node $\lambda(\boldsymbol{\theta},X_j)$ in our cross-validation run, we can later calculate $p_j(y|\boldsymbol{\theta},X_j)$ for any y based on stored MCMC output; the mean of these calculated values is $p_j(y|\boldsymbol{D}_{-j},X_j)$. The mean of the predictive distribution is efficiently approximated by the posterior mean of $\lambda(\boldsymbol{\theta},X_j)$; the variance of the predictive distribution by the sum of posterior mean and variance of $\lambda(\boldsymbol{\theta},X_j)$.

Conditional predictive ordinates

One particular value of the predictive distribution function is of special interest: it is the value of $p_j(y|\boldsymbol{D}_{-j},X_j)$ evaluated at the omitted value Y_j. This value, $p_j(Y_j|\boldsymbol{D}_{-j},X_j)$, is known as the conditional predictive ordinate (Geisser 1980, 1993, Pettit 1990), CPO_j; it is fundamental to the Bayesian LOOCV estimate of out of sample predictive fit (Gelman et al. 2014, Hooten and Hobbs 2015), a Bayesian alternative to *AIC* which we describe subsequently.

In writing BUGS code for a cross-validation run, care must be taken to distinguish the fixed value Y_j (the omitted datum) from the values Y_j° sampled from $p_j(y|\boldsymbol{D}_{-j},X_j)$. A simple solution is to include equal vectors **Y** and **y** in the data statement, and to use upper case **Y**[j]'s for modeled data, and lower case **Y**[j]'s in calculating CPO_j. Vector **y** is ignored in the full analysis. In the jth cross-validation run, **Y**[j] is set to missing, **Y**[j] is is used for calculating a node that equals $p(Y_j|\boldsymbol{\theta},X_j)$; the posterior mean of this node is CPO_j.

Using the predictive distribution for goodness of fit

Model selection and model checking are distinct tasks, the former involving collections of models, the latter involving single models (Hobbs and Hooten 2015). While model selection might help in ruling out poor models, reliable inference requires that at least one of the models in the model set reasonably approximate the data generating mechanisms (Barker and Link 2015). Our emphasis in this paper is on using the predictive distribution as a means of model selection, but we note in passing its use for testing goodness of fit. The use of LOOCV for testing might remedy the well-known conservativeness of conventional Bayesian P-values (Gelman 2013).

Given a sample of a continuous random variable R with cumulative distribution $F(t) = \Pr(R \leq t)$, $F(R)$ (the quantile corresponding to R) has a uniform distribution on (0,1). Given a sample $R_1,R_2,\ldots,R_n$ and a fully specified candidate cumulative distribution function, simple tests of goodness of fit can be constructed by evaluating whether the values $U_i = F(R_i)$ appear to be a sample from a uniform distribution.

Similar tests can be constructed for discrete random variables (Dunn and Smyth 1996). For integer valued data as considered here, given a candidate cumulative distribution function $F(t)$ and its distribution function

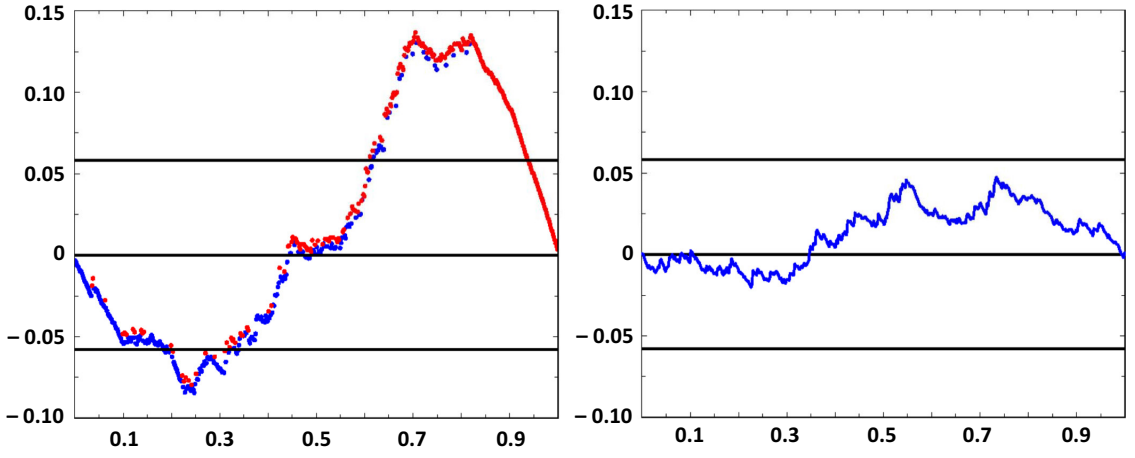

FIG. 1. In both panels, the x-axis indicates expected quantiles of a sample of n uniform random variables; the y-axis is observed minus expected quantiles for a given data set of size n and a candidate distribution. Solid lines indicate Kolmogorov-Smirnov 95% bounds for $n = 520$. For both panels, data were generated from a Poisson distribution with mean 2, Pois (2). In the right-hand panel, the candidate distribution was Pois (2); in the left-hand panel, half of the observation were compared to Pois (0.8) (red dots) and half to Pois (3.2) distributions (blue dots).

$f(t) = \Pr(R = t)$, we set $U_i = F(R_i - 1) + A_i f(R_i)$, where A_i is uniformly distributed on (0,1), $A_i \sim U(0,1)$.

Fig. 1 provides an example. We generated a sample of 520 Poisson random variables with mean 2; we denote their distribution as Pois (2). We calculated the values U_i as described, using Pois (2) as the candidate distribution. The difference between observed and expected values of the empirical distribution of the U_i's is given in the left panel. All values fall within the range ±0.0582, the 95% Kolmogorov-Smirnov bounds. We repeated the experiment, this time assigning Pois (0.8) for half of the candidate distributions, and Pois (3.2) for the remainder. Values U_i corresponding to Pois (0.8) candidates (red dots) tend to be too large, and values corresponding to Pois (3.2) candidates (blue dots) tend to be too small.

In the Bayesian framework considered here, we evaluate goodness of fit by comparing values Y_j to their predictive distributions. We calculate

$$U_j = \sum_{y<Y_j} p_j(y|\boldsymbol{D}_{-j},X_j) + A_j p_j(Y_j|\boldsymbol{D}_{-j},X_j) \tag{5}$$

where $A_j \sim U(0,1)$, and see whether the collection of values U_j appears similar to a sample from a uniform distribution. (Note that in context of MCMC fitting of a model, this is done after the Markov chain simulation, once the values $p_j(y|\boldsymbol{D}_{-j},X_j)$ have been calculated.)

The test procedure described here has the somewhat disconcerting property of depending, in part, on random values A_j. Given a different set of A_j's, the graphics in Fig. 1 will be changed. In practice, we suggest repeating the test with new values A_j, and seeing whether there is consistency among results. The appeal of the test is that it allows us to simultaneously evaluate n samples of size 1, each from a distinct discrete distribution.

Using the predictive distribution for model selection

We turn our attention to the problem of model selection, using the predictive distribution as our tool.

One could summarize the predictive distribution by a point prediction Y_j^* (e.g., the mean, median, or mode), calculate the squared error $(Y_j - Y_j^*)^2$ and R_M as at Eq. 2, and select models on the basis of smallest total squared error. However, the Bayesian analysis allows many more options. The predictive distribution $p_j(y|\boldsymbol{D}_{-j},X_j)$ describes the entire range of values we might have anticipated for Y_j, based on the model and information provided by the rest of the data set, and we can base our evaluation on any relevant feature.

For instance, we can imagine an infinite collection of predictions Y_j°, sampled from the predictive distribution, all predicting the same fixed outcome Y_j. To compare models' predictive ability, we can consider the mean squared error of predictions Y_j°, which is

$$\begin{aligned} MSE_j &= E\left((Y_j^\circ - Y_j)^2|\boldsymbol{D}_{-j},X_j\right) \\ &= Var(Y_j^\circ|\boldsymbol{D}_{-j},X_j) + \left(Y_j - E(Y_j^\circ|\boldsymbol{D}_{-j},X_j)\right)^2. \end{aligned} \tag{6}$$

Using the mean of the predictive distribution as a point prediction, the second term on the right hand side of Eq. 6 is the squared error $(Y_j - Y_j^*)^2$. MSE_j seems a more appropriate evaluation than the squared error: two models might predict the same value for Y_j, thus yielding the same value for $(Y_j - Y_j^*)^2$; MSE_j will be larger for the model with lesser predictive precision.

More commonly, model selection is based on ratios of conditional predictive ordinates. (For a discussion of optimality criteria ["locality" and "propriety"]) leading to the use of conditional predictive ordinates, see Gelman et al. (2014:998).

Using superscript m to denote models m = 1,2, the ratio

$$B_j^{(1,2)} = \frac{CPO_j^{(1)}}{CPO_j^{(2)}} = \frac{p_j^{(1)}(Y_j|\boldsymbol{D}_{-j},X_j)}{p_j^{(2)}(Y_j|\boldsymbol{D}_{-j},X_j)}$$

is the Bayes factor in favor of model 1 over model 2, if we assign priors on parameters based on the original priors [$\boldsymbol{\theta}$] as informed by $\boldsymbol{D}_{-j}$, and consider observation Y_j as the only data.

The product of values $B_j^{(1,2)}$ is known as the pseudo-Bayes factor

$$pBF^{(1,2)} = \prod_{j=1}^{n} \frac{CPO_j^{(1)}}{CPO_j^{(2)}} = \prod_{j=1}^{n} \frac{p_j^{(1)}(Y_j|\boldsymbol{D}_{-j},X_j)}{p_j^{(2)}(Y_j|\boldsymbol{D}_{-j},X_j)} \tag{7}$$

(Geisser and Eddy 1979, Gelfand 1996). Like Bayes factors generally, it is interpreted as a change in relative support for the two models provided by the data, as measured by an odds ratio. The nth root of $pBF^{(1,2)}$ (the geometric mean of the values $B_j^{(1,2)}$) can be thought of as a change in support per observation (Vehtari and Lampinen 2002). Recalling that CPO_j is an estimate of the unknown $p(Y_j|\theta,X_j)$, it is clear that $pBF^{(1,2)}$ estimates the likelihood ratio for models m = 1,2 (Eq. 3).

The pseudo-Bayes factor is closely related to a Bayesian analog of *AIC*, the Bayesian predictive information criterion (Gelman et al. 2014). For a given model m, this is

$$BPIC^{(m)} = \sum_j \log(CPO_j^{(m)}) = \sum_j \log\left(p_j^{(m)}(Y_j|\boldsymbol{D}_{-j},X_j)\right).$$

The *BPIC* is a sum of logs of estimates of $p(Y_j|\boldsymbol{\theta},X_j)$ and thus except for a factor of −2 is similar to the leading term in *AIC*, $-2\Sigma_j \log(p(Y_j|\hat{\boldsymbol{\theta}},X_j))$. The usual penalty term in *AIC*, typically interpreted as a penalty for more highly parameterized models, is in fact a bias correction for use of the MLE $\hat{\boldsymbol{\theta}}$ as a plug-in estimator; no such correction is needed for *BPIC*, unless n is small (Gelman et al. 2014).

Selection of the model with largest *BPIC* operates like selection of the model with smallest *AIC*, providing a convenient summary of cross-validation analyses. *BPIC* differences are logarithms of pseudo-Bayes factors, $\Delta BPIC = BPIC^{(1)} - BPIC^{(2)} = \log(pBF^{(1,2)})$.

Cross-Validation Analyses for BBS Bald Eagle Data

We illustrate the use of cross-validation for comparing two models of population change for Bald Eagle (BAEA, *Haliaeetus leucocephalus*). The BBS data we consider consist of 11 134 counts made by 1647 observers in 68 physiographic strata from 1994 to 2013. Counts are typically small (76.9% zeros, 13.7% ones, and only 1.1% >10).

The models we considered are elaborations of the basic overdispersed Poisson regression model given at Eq. 1. The following features were common to both models:

- Random effects ω and ε were treated as spatially homogeneous (mean-zero normal with constant variance across strata).
- Stratum effects were treated as sampled from a common normal distribution.
- The only fixed effect associated with observers was an effect for first year of service.
- Precision parameters were assigned vague Gamma priors, $\Gamma(0.001,0.001)$; all other parameters were assigned vague normal priors, $\mathcal{N}(0,10^{-6})$.

The first model we considered assumed independent linearly trending random year effects $\gamma_{sy} \sim N(\beta_s(y-y_0),\tau^\gamma)$, for which we will use the abbreviation LTRE. Parameters β_s were treated as sampled from a common normal distribution, τ^γ, τ^ω, and τ^ϵ were constant across strata. This is the standard model reported in annual summaries of the BBS (Sauer et al. 2014), except for the hierarchical structure we have specified for parameters β_s and the use of a spatially constant value for τ^γ. These relatively minor changes were made in light of the low abundances of BAEA in the BBS.

The second model was identical to the first, except that first differences in year effects were independent and normally distributed: $\gamma_{sy} - \gamma_{s,y-1} \sim N(0,\tau^\gamma)$. We will use the abbreviation FDRE for this model structure. Year effects in base year y_0 were modeled as $\gamma_{sy_0} \sim N(0,0.001 \times \tau^\gamma)$ to ensure a noninformative prior of conjugate form (see Link and Barker 2010:292 for details)."

Before describing model selection, it is worth describing *why* model selection matters for this particular data set. The two models differ in process priors, hierarchical structures describing stochastic relations among the year effects γ_{sy}. As is almost always the case in Bayesian analysis, sufficient data will overwhelm the prior, and fitted values from the two analyses will be nearly identical (an example is presented later in this paper; see Fig. 5).

Most of the time, however, we are not dwellers in Asymptopia, and the process prior informs the pattern of change among the parameters. Under the LTRE model, the parameters γ_{sy} are random deviations from a line; under the FDRE model, they are a random walk. In the LTRE model, estimates of γ_{sy} are informed by the overall pattern of population change, while in the FDRE model they are influenced more by the values in adjacent years. The LTRE model highlights long term pattern of change, while the FDRE model exhibits a smoothly changing pattern.

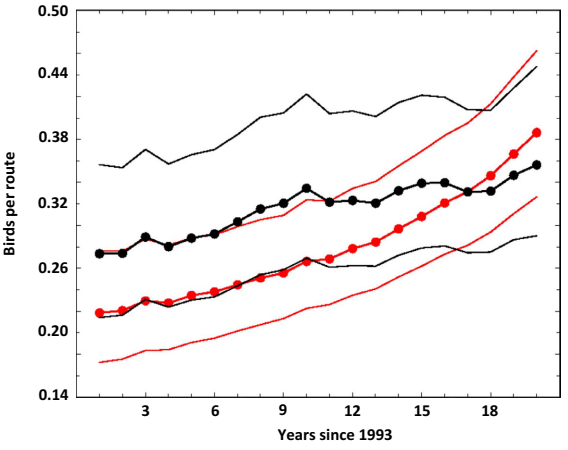

Fig. 2. Composite population trajectories (lines with circles) and 95% credible intervals (lines without circles) for Bald Eagle by first difference (black) and linearly trending (red) year effects models.

The consequences of these features to analysis of BBS Bald Eagle data are predictable. Due to low abundances on most BBS routes, estimates of γ_{sy} will be sensitive to the choice of process prior. Bald Eagle populations having grown substantially over the past 20 years, the underlying loglinear pattern in the LTRE model is one of rapid growth. However, recent counts suggest a slowing rate of growth; the FDRE model is less influenced by long term changes and leads to smaller estimates of γ_{sy} in recent years (see Fig. 2). Consequently the geometric mean rate of population change (the trend) was estimated as 3.1% (1.8%, 4.3%) [posterior median (95% CI)] under the LTRE model; the estimate was 1.4% (0.0%, 2.8%) under the FDRE model. Which model is to be favored, and which trend to be reported? We have noted a distressing tendency to assume that, the shorter the interval estimate, the better the model, but of course the reliability of the interval estimate depends on the adequacy of the model. An objective model selection criterion is needed.

We conducted analyses of the BAEA data using program *JAGS*. Analyses of the full data set were based on Markov chains of length 50,000, after burn-in of length 5,000, and took approximately 1 hour of run time on a workstation equipped with an i7-980x processor (Intel, Santa Clara, California, USA). There being 11,134 observations, a full leave-one-out cross-validation to compare the two models would thus take approximately 2.5 years, after which we would no doubt realize that there was a typo in our original code, and have to start over again. It is intriguing to note the enduring relevance of IJ Good's (1952) observation:

> "We must weigh up the expected time for doing the mathematical and statistical calculations against the expected utility of these calculations. Apparently less good methods may therefore sometimes be preferred."

We cut the time required for cross-validations in four ways. First, recognizing that the posterior distributions for the full analysis are similar to those for the cross-validation analyses, we stored the final state of the original sampler as starting values for subsequent cross-validation analyses, thus avoiding the need for burn-ins. Second, recognizing that slight imprecision in estimation of MSE_j and CPO_j would be of little consequence in comparing models, we reduced the Markov chain lengths to 10,000 for the cross-validation runs.

Third, we chose to limit cross-validations to a subset of the $n = 11{,}134$ observations (note Gelman et al. (2014) suggestion of "a random subset [chosen for] efficient computations if n is large"). We performed three cross-validations to evaluate distinct features of the models. First, we chose 26 observations at random from each of the 20 years of data, to evaluate the overall fit of the two models; the total number of cross-validation runs was thus reduced to $n^* = 520$. We refer to this analysis as our "cross-validation for fit."

Noting that our interest focuses on the capacity of the models to project into the future, our second cross-validation was restricted to the $n^* = 525$ observations in the final year of the data (2013). This analysis is precisely in the spirit of a full cross-validation based only on the last year's data, using priors informed by the previous 19 years of data. We also conducted a third cross-validation restricting attention to the $n^* = 540$ observations in the first year (1994). We refer to the second and third analyses as "targeted cross-validations," because they are aimed at specific features of the models.

The fourth way in which we reduced computation time was by running analyses in parallel on multiple processor cores (see comments on parallel processing in Hooten and Hobbs 2015:11). The computational savings resulting from these limited cross-validations, plus the benefits of running analyses in parallel, resulted in analyses which could be conducted over a weekend.

Results for Bald Eagle Analyses

The *BPIC* values for models 1 (LTRE model) and 2 (FDRE model) were −335.04 and −344.28, indicating superior performance for model 1.

But how does one go about comparing these values? Is the difference real, or the result of sampling variation? Vehtari and Lampinen (2002) suggest that (for fixed m) LOOCV values $CPO_j^{(m)}$ are nearly independent if n is large. Differences in logarithms of CPO_j for models $m = 1,2$ could then be evaluated as if performing a paired two-sample z-test; the z-statistic provides a rough test of equal predictive value for the two models. The standard deviation of $\log(CPO_j^{(1)}) - \log(CPO_j^{(2)})$ was 0.174, so assuming independence, the associated standard deviation of $BPIC^{(1)} - BPIC^{(2)}$ would be $3.97 = \sqrt{520} \times 0.174$. The realized difference of 9.24 translates to a z score of 2.33, for a two-tailed P-value of just under 2%.

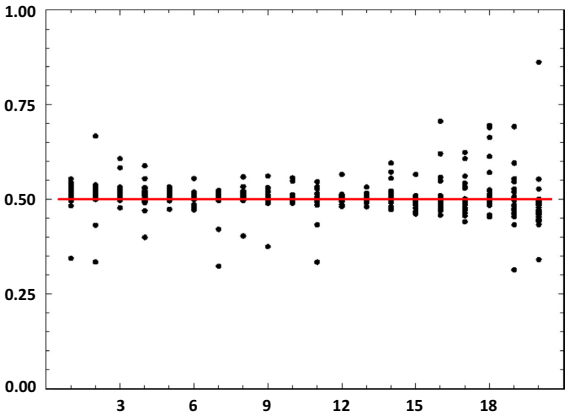

FIG. 3. Values $V_j^{(1,2)}$ plotted against year of observation j for Bald Eagle analyses. Values greater than 0.50 indicate that the corresponding observation was better predicted by model 1, the linearly trending year effects model.

In Fig. 3 we examine components of the *BPIC* values for these data. Recall that *BPIC* differences are logarithms of pseudo-Bayes factors

$$exp(\Delta BPIC) = pBF^{(1,2)} = \prod_{j=1}^{n^*} B_j^{(1,2)}$$

where $B_j^{(1,2)} = CPO_j^{(1)}/CPO_j^{(2)}$ can be interpreted as a Bayes factor in favor of model 1 over model 2 based on data Y_j. Then $V_j^{(1,2)} = B_j^{(1,2)}/(1 + B_j^{(1,2)})$ can similarly be interpreted as a posterior probability for model 1, given equal prior weights on the two models. Because $V_j^{(1,2)}$ is a monotone transformation (inverse logit) of the difference in values $\log(CPO_j^{(m)})$, inference based on these leads to the same conclusions in summary; the advantage of using $V_j^{(1,2)}$s lies in their intuitive interpretation as model probabilities associated with individual observations. We might think of each observation Y_j being allowed to vote for a model, and $V_j^{(1,2)}$ as its voting preference for model 1.

While 300 of the 520 values (57.7%) exceeded 0.50, their average was only 0.5042. In 520 flips of a fair coin, the probability of 300 or more heads is only 0.00026. A clear majority of "voters" favor model 1, but the average margin of their preference is slight.

An interesting pattern emerges on plotting $V_j^{(1,2)}$ against the year of observation j (Fig. 3). Model 1 appears to predict better in early years (up to year 11); model 2 in later years, especially the last. We thus turn our attention to the targeted cross-validations, to more carefully examine the performance of the models for the last year's data.

In the targeted cross-validation for 2013, only 177 of 525 final year observations (33.7%) favored model 1. Nevertheless, the *BPIC* values for models 1 and 2 were

TABLE 1. Summary of cross-validations for BAEA data.

	Number of eagles				
Year	0	1	2	≥3	All
1994					
Ratio	0.611	0.999	0.772	0.705	0.703
Fraction	439/457	9/49	6/14	11/20	465/540
Mean	1.083	0.693	0.564	0.835	1.013
2013					
Ratio	1.391	1.316	1.168	0.870	0.960
Fraction	17/349	84/93	43/45	33/38	177/525
Mean	0.880	1.309	1.713	2.160	1.066

Notes: Ratio is the ratio of $\Sigma_j \mathrm{MSE}_j$ values; fraction is proportion with $B_j^{(1,2)} > 1$ (i.e., favoring LTRE over FDRE model) over the total number of observations; mean is the geometric mean of ratios $B_j^{(1,2)}$. For instance in 1994, two eagles were reported on 14 counts; for these counts the total mean-squared error prediction under model 1 was 0.772 times that for model 2; the LTRE model gave the better prediction for 6 of the counts, and the geometric mean of the 14 ratios $B_j^{(1,2)}$ was 0.564.

−541.4 and −575.2, respectively, favoring model 1. The reason for this discrepancy is evident on examination of Table 1. 160 of the 177 observations favoring model 1 were for observations ≥1 whereas 332 of the 348 observations favoring model 2 were for observations = 0. For 2013 data, model 2 predicts the prevailing zero counts slightly better, but does substantially worse in predicting nonzero counts.

At the other end of the time-series, the situation is almost reversed: model 1 does a better job of predicting the prevailing zero observations, and model 2 does a better job predicting non-zero counts. However, the larger number of zero observations leads to the result that model 1 (LTRE) performs better overall; also, the total of MSE_j is uniformly smaller under model 1.

We conducted informal goodness of fit evaluations for both models, applying the Kolmogorov-Smirnov type test statistic previously described to the 520 predictive distributions calculated in our cross-validation for overall fit. We repeated the tests 100 times for each model; the results are shown in Fig. 4. While some variation existed among the replicate test values, the Kolmogorov-Smirnov 95% bounds were breached in only 1 of 100 cases for the LTRE model, and never for the FDRE model. These results suggest that both models provide adequate fits to the data.

Our expectation, prior to data analysis, was that in the presence of a long term pattern of increasing trend, and with weak data, the linear year effect model (model 1) would indicate larger populations in the last year, and smaller populations in the first year, than the first difference model. This was in fact what happened (Fig. 2). The question remaining was whether that phenomenon is genuine or an artifact of the models; our cross-validation results indicate that the linear year effect model is better supported by the data.

BBS Carolina Wren Data

Our analyses of BAEA data were restricted to the period 1994–2013, omitting earlier years in which abundance was very low; even so, 77.0% of the BAEA counts were zeros and 90.6% counts were ≤1. By contrast, Carolina Wren (CARW, *Thryothorus ludovicianus*) counts for the period 1966–2013 had median of 5, and 90th percentile of 30. The longer time period and slightly larger geographic range (74 strata for CARW vs. 68 for BAEA) lead to a larger data set (n = 42,039 for CARW vs. 11,134 for BAEA). CARW populations exhibit periods of rapid growth, punctuated by large declines associated with severe winters (Sauer et al. 1996, Link and Sauer 2007). These features

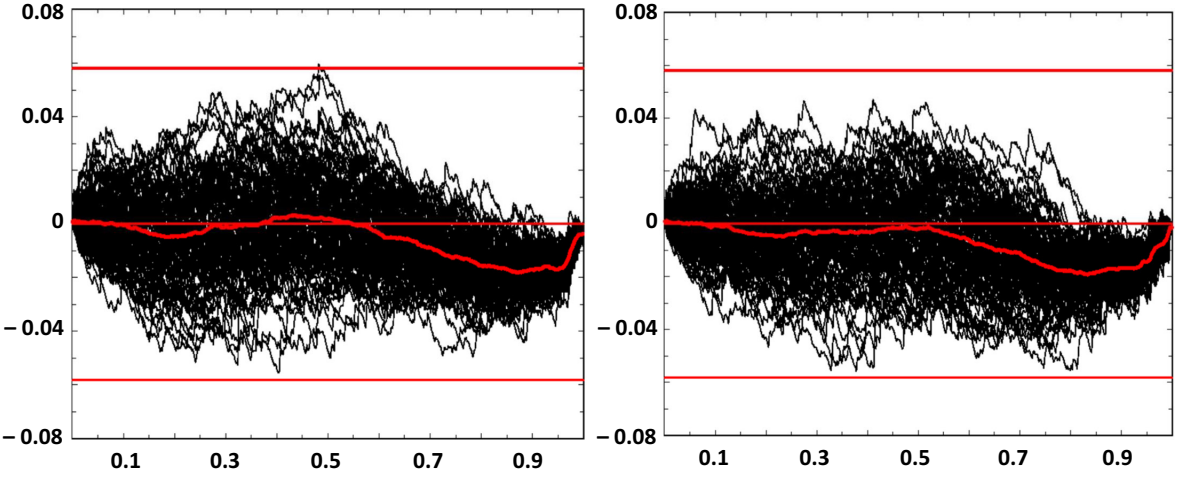

FIG. 4. In both panels, the x-axis plots expected quantiles of a sample of n uniform random variables, the y-axis is observed minus expected quantiles of data U_j, based on predictive distributions as at Eq. 5. Red lines at ±0.0582 indicate Kolmogorov-Smirnov 95% bounds for n = 520; 100 curves are plotted for first difference (right) and linearly trending (left) year effects models; in each case, the mean is plotted in red.

suggest that BBS data for CARW may be more appropriately described by the first difference model. We thus compared first difference and linear year effects models in application to BBS data for CARW.

The basic models considered were identical to those used with BAEA data, except that variation in year effects γ_{sy} was modeled with hierarchical structure across strata: rather than assume $\tau_s^\gamma \equiv \tau^\gamma$, we treated τ_s^γ as sampled from a lognormal distribution.

Survey-wide composite population trajectories for CARW are presented in Fig. 5, with red curves corresponding to the first difference model, and black to the linear year effects model. The smooth black curve is an estimate (posterior median) of the composite linear component of population change under the linear year effects model. Roughly speaking, the linear year effects model shrinks annual estimates toward the smooth curve, while the first difference model shrinks annual estimates toward adjacent values.

We make two observations about the performance of the models before comparing their fit by cross-validation. First, the general pattern of population change is nearly identical under the two analyses: we won't be led too far astray by use of one if the other is correct. The two models differ in specification of a process prior, and the data overwhelm the prior.

Second, it is to be anticipated that the first difference model will indicate a smoother pattern of population change than the linear year effects mode, and thus will estimate smaller changes in population size between years. Indeed, 39 of 47 differences in point estimates between years i and $i + 1$ are smaller under the first difference model; the average ratio of absolute change is 18% smaller. In particular, the linear year effects model indicates a 39% decline between years 10 and 11 (the severe winter of 1975–1976), while the first difference model indicates only a 33% change.

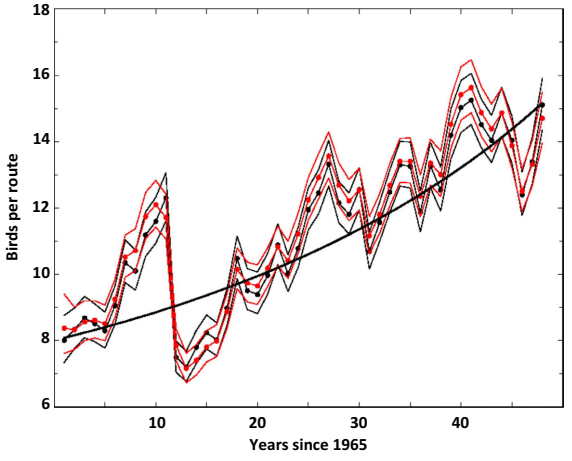

FIG. 5. Composite population trajectories and 95% credible intervals for Carolina Wren by first difference model (red) and linear year effects model (black); smooth curve (black) is estimate of linear component of trajectory in linear year effects model.

It is not enough to report such differences, and to assert that they represent a failing of one model or the other. Given that such differences can be anticipated a priori, as features of the models, it is necessary to compare how well the models fit the data.

We thus conducted 576 LOOCV analyses, one for each of 12 randomly selected observations for each of the 48 years. Values CPO_j were smaller for the linear year effects model in 333 of the 576 cases. The *BPIC* values were −1,294.2 and −1,286.4, for linear and first difference year effects models, respectively. Thus the mean difference in $log(CPO_j^{(m)})$ values was −0.0135; the standard deviation of these differences was 0.1504, leading to a z score of −2.15. These results favor the first difference model over the linear year effects model.

BBS Wood Thrush Data

Our analyses of BAEA and CARW data focused on model selection for patterns of temporal change; the CARW data set with larger counts and a highly dynamic population trajectory was more appropriately described by the FDRE model. We turn our attention now to a problem of model selection for nuisance parameters.

The term "nuisance parameters" is often used to describe parameters that are not of primary interest per se, but which are included out of concern that their omission would cause misleading inferences about parameters of primary interest. The analyst attempts to include enough nuisance parameters to avoid bias, while not including too many and unnecessarily sacrificing precision.

We were interested in assessing the necessity of spatial variation in parameters τ^ω and τ^ϵ. These parameters describe variation in the log of expected counts; this is (essentially) the coefficient of variation on the original scale. Parameter τ^ω describes the effect of differences in observers/routes on the CV; parameter τ^ϵ summarizes overdispersion relative to the Poisson model. Our intuition is that these are not the easiest parameters to estimate, requiring substantial data resources; attempts to model spatial variation in τ^ω and τ^ϵ would seem challenging on a priori grounds, and perhaps unnecessary. Our interest in evaluating the necessity of these structures arises from recent suggestions that spatial variation in τ^ω be modeled in analysis of BBS data (Smith et al. 2014).

We chose to use Wood Thrush (WOTH, *Hylocichla mustelina*) data to evaluate the necessity of spatial structure in these nuisance parameters. With 52,125 counts by 7,643 observers on 86 strata over 48 years (1966–2013) having mean of 7.3, standard deviation of 9.5, and quartiles of 1, 4, and 11,the WOTH data set is one of the best in the BBS database, and one of the most promising for distinguishing such fine-grained differences in nuisance parameters.

As with the CARW data, we performed $n^* = 576$ LOOCV analyses, omitting 12 randomly selected observations in each of the 48 years. The analyses were

conducted using four models. These began with the LTRE model used for the preceding CARW analyses, and included or excluded spatial variation in parameters τ^{ω} and τ^{ϵ}. Models with spatial variation in precision parameters treated these as following lognormal distributions. Using self-explanatory notation for the models, *BPIC* values obtained were as follows:

1) Model: $\{\tau^{\omega},\tau^{\epsilon}\}$, −1288.74
2) Model: $\{\tau^{\omega}_s,\tau^{\epsilon}\}$, −1287.43
3) Model: $\{\tau^{\omega},\tau^{\epsilon}_s\}$, −1288.94
4) Model: $\{\tau^{\omega}_s,\tau^{\epsilon}_s\}$, −1285.91

With four models, there are $\binom{4}{2} = 6$ pairwise z statistics for comparing model-specific values CPO_j. The largest value was $z = 1.20$ in favor of model 4 over model 2; all of the other z-statistics were <0.7 in absolute value. We conclude that there is little support for the additional model complexity required in allowing τ^{ω} and τ^{ϵ} to vary spatially.

The Watanabe-Akaike Information Criterion

Cross-validation provides a conceptually satisfying means for evaluating model fit and for model selection. Its drawback is its computational burden. The Bayesian predictive information criterion based on leave-one-out cross-validation can be written as

$$BPIC = \sum_j \log\left\{\mathrm{E}_{\mathrm{post}_{-j}}\left[p(Y_j|\boldsymbol{\theta},X_j)\right]\right\} \tag{8}$$

where the symbol $\mathrm{E}_{\mathrm{post}_{-j}}(\cdot)$ denotes computation of an expectation against the posterior distribution $[\boldsymbol{\theta}|\boldsymbol{D}_{-j}]$, and data set $\boldsymbol{D}_{-j}$ is obtained by omitting Y_j from the full data set $\boldsymbol{D}$. Each summand in Eq. 8 requires a reanalysis of (essentially) the full data set to evaluate $[\boldsymbol{\theta}|\boldsymbol{D}_{-j}]$.

The Watanabe-Akaike information criterion provides a convenient approximation to $(-2) \times$ *BPIC*, without the computational burden. Like the *DIC*, *WAIC* is based on posterior distributions obtained from a single analysis of the full data set. The two quantities are similar in their definitions

$$DIC = -2\sum_j \log p(Y_j|\hat{\boldsymbol{\theta}}_{\mathrm{Bayes}},X_j) + 2p_D^{DIC} \tag{9}$$

and

$$WAIC = -2\sum_j \log\left\{\mathrm{E}_{\mathrm{post}}\left[\,\mathrm{p}(Y_j|\boldsymbol{\theta},X_j)\right]\right\} + 2p_D^{WAIC} \tag{10}$$

here, $\hat{\boldsymbol{\theta}}_{\mathrm{Bayes}}$ is a Bayesian point estimate of $\boldsymbol{\theta}$ (typically the posterior mean) and

$$p_D^{DIC} = 2\mathrm{Var}_{\mathrm{post}}\sum_j\left(\log(p(Y_j|\boldsymbol{\theta},X_j))\right)$$

and

$$p_D^{WAIC} = \sum_j \mathrm{Var}_{\mathrm{post}}\left(\log(p(Y_j|\boldsymbol{\theta},X_j))\right)$$

are "data-based bias corrections" (Gelman et al. 2014). $\mathrm{E}_{\mathrm{post}}(\cdot)$ and $\mathrm{Var}_{\mathrm{post}}(\cdot)$ denote mean and variances against the posterior distribution $[\boldsymbol{\theta}|\boldsymbol{D}]$. *DIC* and *WAIC* are based on the full posterior distribution $[\boldsymbol{\theta}|\boldsymbol{D}]$, hence do not require the multiple analyses involved in calculating *BPIC*.

The principal difference between *DIC* and *WAIC* is evident in Eqs. 9 and 10: the *DIC* is calculated with a "plug-in" estimator of $\boldsymbol{\theta}$, whereas *WAIC* averages the loglikelihood against the posterior distribution of $\boldsymbol{\theta}$.

Deviance information criterion, Watanabe-Akaike information criterion, and Bayesian predictive information criterion are fundamentally similar: all are based on estimation of $\log(\prod_i p(Y_i|\boldsymbol{\theta},\mathbf{X}_i))$. DIC-differences, WAIC-differences, and $-2\times$ *BPIC* differences are essentially changes in deviances between models, or −2 times the logarithm of a likelihood ratio for models. The difference is in how uncertainty in parameter values is handled; *BPIC* is the most appealing mathematically, followed by *WAIC* (as an approximation to *BPIC*).

The *BPIC* is much more computationally expensive, but when one considers the mountains of work that go into collecting most data sets (like the BBS!) it seems silly to cut corners on analysis. Yet IJ Good's dictum that "we must weigh up the expected time for doing the mathematical and statistical calculations against the expected utility of these calculations" rings true. It is interesting, therefore, to compare results for *DIC* and *WAIC* relative to those for *BPIC* in the examples we have considered.

Table 2. Summary of model selection criteria for North American Breeding Bird Survey data for Bald Eagle (BAEA), Carolina Wren (CARW), and Wood Thrush (WOTH)

	$-2(n/n^*) \times BPIC$	*WAIC*	*DIC*
BAEA			
LTRE	**14347.4**	**15093.5**	**17793.2**
FDRE	14743.1	15206.0	18491.6
CARW			
LTRE	188912.8	182206.9	**197426.6**
FDRE	**187774.2**	**181906.4**	198804.7
WOTH			
$\tau^{\omega},\tau^{\epsilon}$	233248.5	217423.9	239101.5
$\tau^{\omega}_s,\tau^{\epsilon}$	233011.4	217150.0	**236963.9**
$\tau^{\omega},\tau^{\epsilon}_s$	233284.7	215746.0	239591.9
$\tau^{\omega}_s,\tau^{\epsilon}_s$	**232736.3**	**215438.1**	239512.9

Notes: BAEA and CARW data were analyzed with linearly trending random year effects (LTRE) and first difference random year effects (FDRE) models. WOTH data were analyzed with and without spatial variation in precision parameters τ_{ω} (for observer effects) and τ_{ϵ} (for overdispersion); subscript s in the model description indicates the inclusion of spatial variation. Bayesian predictive information criterion (*BPIC*) values were calculated for a subset of size $n^* < n$; multiplication by a factor $-2(n/n^*)$ presents the values on the same scale. For each data set, the most favorable (smallest) value of each criterion is shown in boldface type. *WAIC*, Watanabe-Akaike information criterion; *DIC*, deviance information criterion.

Table 2 summarizes the selection criteria for the three sets of analyses we performed. We note that for the three applications presented here, *WAIC* agrees with *BPIC* in selection of a best model, but not in the overall ratings for the WOTH data. *DIC* did not agree with *BPIC* in two of three cases. For further comparisons of the performance of these criteria in application to specific data sets, see Hooten and Hobbs (2015).

Discussion

Leave-one-out cross-validation fills a significant need, providing a conceptually simple if computationally intensive means of evaluating fit and selecting among complex hierarchical models. The capacity to compare individual values CPO_j allows us to inspect the relative effects of covariate $\mathbf{X}_j$ on the fits of multiple models. Model specific values of CPO_j can also be used for assessing model-specific goodness of fit.

Sums of logarithms of CPO_j values constitute the Bayesian predictive information criterion (*BPIC*) which is comparable to the better known *AIC* and *DIC*, but does not share their failings in application to hierarchical models.

Differences in *BPIC* values can be evaluated by an approximate *z*-test. Using superscript (*m*) to denote models, we have

$$BPIC^{(1)} - BPIC^{(2)} = \sum_j \log(CPO_j^{(1)}) - \sum_j \log(CPO_j^{(2)}) = \sum_j \Delta_j^{(1,2)}$$

where $\Delta_j^{(1,2)} = \log(CPO_j^{(1)}) - \log(CPO_j^{(2)})$. Let

$$Z = \frac{\bar{X}(\Delta)}{\sqrt{S^2(\Delta)/n^*}}$$

with $\bar{X}(\Delta)$ and $S^2(\Delta)$ denoting the sample mean and variance of the values $\Delta_j^{(1,2)}$, and n^* being the number of values $\Delta_j^{(1,2)}$ in the calculation. Letting $f^{(1)}(y)$ and $f^{(2)}(y)$ denote the modeled distributions of observations under the two models, and letting $t(y)$ denote the data generating distribution, $\bar{X}(\Delta)$ estimates

$$\boldsymbol{\theta} = E_t\left(\log\left(\frac{f^{(1)}(Y)}{f^{(2)}(Y)}\right)\right)$$

the expected value of the log-likelihood ratio for the candidate models, based on a value Y sampled from the data generating distribution. Thus

$$\boldsymbol{\theta} = E_t\left(\log\left(\frac{t(Y)}{f^{(2)}(Y)}\right)\right) - E_t\left(\log\left(\frac{t(Y)}{f^{(1)}(Y)}\right)\right)$$

the difference in Kullback-Leibler divergences from $t(y)$. The statistic Z is approximately standard normal under the assumption that $\boldsymbol{\theta} = 0$; values >0 favor model 1, values <0 favor model 2. Consequently, we need not select models merely on the basis of their sorted *BPIC* values, but can evaluate the magnitude of differences in the values, by an approximate test of H_0:$\boldsymbol{\theta} = 0$. In our considerations of models for BBS WOTH data, a bare ranking of *BPIC* values would seem to favor the most highly parameterized models, but the difference in *BPIC* values among the four models considered is slight, suggesting that the less complex models generally used are adequate.

As a matter of interest, note that if $t(y) = f^{(1)}(y)$, $\boldsymbol{\theta}$ is the Kullback-Leibler divergence of $f^{(2)}$ from $f^{(1)}$, hence $\boldsymbol{\theta} > 0$; but if $t(y) = f^{(2)}(y)$, $\boldsymbol{\theta}$ is -1 times the Kullback-Leibler divergence of $f^{(1)}$ from $f^{(2)}$, so $\boldsymbol{\theta} < 0$. Thus there exists a mixture of $f^{(1)}$ and $f^{(2)}$ such that $\boldsymbol{\theta} = 0$; two distinct models can diverge equally from the data generating model.

Akaike's information criterion, Bayesian predictive information criterion, Deviance information criterion, and Watanabe-Akaike information criterion differences can be thought of as estimates of loglikelihood ratios for models; the difference is in how parameter uncertainty is accounted for. *BPIC* is the most mathematically appealing, but computationally complex; *WAIC* is asymptotically equivalent to *BPIC*. We find ourselves in agreement with Gelman et al. (2014:1015), who wrote

> "… our preferred choice is cross-validation [*BPIC*], with *WAIC* as a fast and computationally-convenient alternative. *WAIC* is fully Bayesian (using the posterior distribution rather than a point estimate), gives reasonable results in the examples we have considered here, and has a more-or-less explicit connection to cross-validation …"

Models for the North American Breeding Bird Survey

The BBS is a challenging data set for analysis because the survey encompasses many geographic scales (individual survey point to continental), and each scale requires covariates to model both population processes and observational processes. Hierarchical models enable investigators to comprehensively explore features influencing both counts and actual populations at multiple scales, and to incorporate latent attributes of population change (e.g., Hostetler and Chandler 2015). These explorations allow us to address concerns, such as observer influences or seasonal effects on counts, that have dogged the BBS from its inception. Unfortunately, our inability to judge the merits of these models in terms of goodness of fit or relative support for alternative models can lead to chaos. The cross-validation procedures we describe provide a reasonable means of judging value of models and choosing between alternative models.

Choice of parameterization of population change as LTRE or FDRE provides a good example. We have fit versions of both models (following the basic model framework described in Sauer and Link 2011); results indicate clearly that trends (composite population change

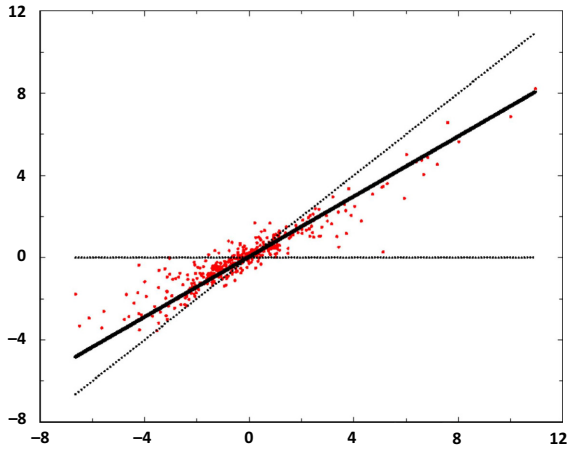

FIG. 6. Trend estimates (% change per year) from first difference year effects model (y-axis) plotted against estimates from linearly trending year effects model for 415 species from the North American Breeding Bird Survey, 1966–2013. Solid line is regression relation, accounting for sampling variation in estimates.

from 1966–2013) from FDRE results tend to be more conservative (340/415 raw estimates were closer to 0) than trends from LTRE results (Fig. 6). This has important consequences for managers and conservationists, who often use magnitude of trend estimates as a criterion for judging priorities for conservation activities (Panjabi et al. 2012). For BEAE, these differences in results among analysis methods can have major consequences, as monitoring data are used to judge species' recovery from endangered status; if BAEA populations decline, endangered status can be reinstated. Although both LTRE and FDRE results clearly indicate increasing populations for BAEA, the differences in magnitude between FDRE and LTRE results are of both ecological and political consequence. Our cross-validation results for BEAE suggest that the LTRE analysis is likely a more reasonable model for estimation of population change for the species.

Choice of model often has important consequences for precision of results. For example, including spatial variation in τ^{ω} leads to less precise results (as defined by half-widths of credible intervals) for both FDRE and LTRE analyses (J. R. Sauer, unpublished analysis of survey-wide trend estimates of BBS data for 417 species, 1966–2013). However, Smith et al. (2014) recommend incorporating spatial variation in τ^{ω} in BBS analyses to accommodate regional differences in counts on routes. Our cross-validation results for WOTH indicate that incorporation of regional variation τ^{ω} provides no substantial advantage in fit. Single-species results do not lead to generalizations, but we do note that WOTH was a species identified by Smith et al. (2014) as one for which incorporation of regional variation in τ^{ω} was of particular interest because of its influence on regional trend estimates. In our analyses of WOTH data, we also investigated regional variation in τ^{e}; failure to consider the relationship between regional variation in these variance components might lead to misinterpretation of sources of regional variation in counts.

The development of reliable tools for model comparison and goodness of fit is a boon for BBS analysts, as they will for the first time have reasonable means of comparing models and assessing the consequences of critical covariates related to climate change and other environmental drivers of population change in birds. Implementing these approaches for a variety of model structures and species will inform our understanding of the validity of the current generic model for BBS analyses and define cases in which alternative models provide superior estimates of population change. The suite of model structures defined in the *Introduction* form the outer bounds of the current model set for BBS data. One reasonable means of retaining the spirit of a generic analysis is by reducing this model set down to a small collection of candidate models and using them as an operational generic model set for BBS data, along with prescribed methods of model selection. The tools described in this paper provide a step in that direction.

ACKNOWLEDGMENTS

Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government. We thank Mevin Hooten, Jim Nichols, an anonymous reviewer, and the associate editor for helpful reviews.

LITERATURE CITED

Barker, R. J., and W. A. Link. 2015. Truth, models, model sets, AIC, and multimodel inference: A Bayesian perspective. The Journal of Wildlife Management 79:730–738.

Celeux, G., F. Forbes, C. P. Robert, and D. M. Titterington. 2006. Deviance information criteria for missing data models. Bayesian Analysis 1:651–673.

Dunn, P. K., and G. K. Smyth. 1996. Randomized quantile residuals. Journal of Computational and Graphical Statistics 5:236–244.

Geisser, S. 1980. Predictive sample reuse techniques for censored data. Trabajos de Estadística y de Investigación Operativa 31:433–468.

Geisser, S., and W. F. Eddy. 1979. A predictive approach to model selection. Journal of the American Statistical Association, 74:153–160.

Geisser, S. 1993. Predictive Inference: An Introduction. Chapman and Hall, New York.

Gelfand, A. E. 1996. Model determination using sampling-based methods. Pages 145–161 *in* W. Gilks, S. Richardson, and D. Spiegelhalter, editors. Markov Chain Monte Carlo in practice. Chapman and Hall, London, UK.

Gelman, A. 2013. Two simple examples for understanding posterior p-values whose distributions are far from uniform. Electronic Journal of Statistics 7:2595–2602.

Gelman, A., J. Hwang, and A. Vehtari. 2014. Understanding predictive information criteria for Bayesian models. Statistics and Computing 24:997–1016.

Good, I. J. 1952. Rational decisions. Journal of the Royal Statistical Society. Series B (Methodological) 14:107–114.

Hobbs, N. T., and M. B. Hooten 2015. Bayesian models: a statistical primer for ecologists. Princeton University Press, Princeton, New Jersey, USA.

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

Hostetler, J. A., and R. B. Chandler. 2015. Improved state-space models for inference about spatial and temporal variation in abundance from count data. Ecology 96:1713–1723.

Link, W. A., and J. R. Sauer. 2002. A hierarchical analysis of population change with application to Cerulean Warblers. Ecology 83:2832–2840.

Link, W. A., and J. R. Sauer. 2007. Seasonal components of avian population change: joint analysis of two large-scale monitoring programs. Ecology, 88:49–55.

Link, W. A., and R. J. Barker. 2010. Bayesian inference with ecological applications. Academic Press, London, UK.

Link, W. A., and J. R. Sauer. In press. Modeling participation duration, with application to the North American Breeding Bird Survey. Communications in Statistics.

Panjabi, A. O., P. J. Blancher, R. Dettmers, and K. V. Rosenberg. 2012. Partners in Flight Technical Series No. 3. Rocky Mountain Bird Observatory: www.rmbo.org/pubs/downloads/Handbook2012.pdf.

Pettit, L. I. 1990. The conditional predictive ordinate for the normal distribution. Journal of the Royal Statistical Society. Series B (Methodological) 52:175–184.

Plummer, M. 2008. Penalized loss functions for Bayesian model comparison. Biostatistics, 9:523–539.

Sauer, J. R., and W. A. Link. 2011. Analysis of the North American Breeding Bird Survey using hierarchical models. The Auk 128:87–98.

Sauer, J. R., B. G. Peterjohn, and W. A. Link 1994. Observer differences in the North American Breeding Bird Survey. The Auk 111:50–62.

Sauer, J. R., G. W. Pendleton, and B. G. Peterjohn 1996. Evaluating causes of population change in North American insectivorous songbirds. Conservation Biology 10:465–478.

Sauer, J. R., W. A. Link, J. E. Fallon, K. L. Pardieck, and D. J. Ziolkowski Jr. 2013. The North American Breeding Bird Survey 1966–2011: Summary analysis and species accounts. North American Fauna 79:1–32.

Sauer, J. R., J. E. Hines, J. E. Fallon, K. L. Pardieck, D. J. Ziolkowski Jr., and W. A. Link. 2014. The North American Breeding Bird Survey, Results and Analysis 1966–2013. Version 01.30.2015, USGS Patuxent Wildlife Research Center, Laurel, Maryland, USA. http://www.mbr-pwrc.usgs.gov/bbs/

Smith, A. C., M. A. R. Hudson, C. Downes, and C. M. Francis 2014. Estimating breeding bird survey trends and annual indices for Canada: how do the new hierarchical Bayesian estimates differ from previous estimates? The Canadian Field-Naturalist, 128:119–134.

Vehtari, A., and J. Lampinen. 2002. Bayesian model assessment and comparison using cross-validation predictive densities. Neural Computation, 14:2439–2468.

Watanabe, S. 2009. Algebraic geometry and statistical learning theory. Cambridge University Press, Cambridge, UK.

Watanabe, S. 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. The Journal of Machine Learning Research 11:3571–3594.

Watanabe, S. (2013). A widely applicable Bayesian information criterion. The Journal of Machine Learning Research 14:867–897.