Ecological Applications, 30(6), 2020, e02137
© 2020 by the Ecological Society of America

Model selection for the North American Breeding Bird Survey

William A. Link,[1] John R. Sauer, and Daniel K. Niven

U.S. Geological Survey Patuxent Wildlife Research Center, Laurel, Maryland 20708 USA

Citation: Link, W. A., J. R. Sauer, and D. K. Niven. 2020. Model selection for the North American Breeding Bird Survey. Ecological Applications 30(6):e02137. 10.1002/eap.2137

Abstract. The North American Breeding Bird Survey (BBS) provides data that can be used in complex, multiscale analyses of population change, while controlling for scale-specific nuisance factors. Many alternative models can be fit to the data, but most model selection procedures are not appropriate for hierarchical models. Leave-one-out cross-validation (LOOCV), in which relative model fit is assessed by omitting an observation and assessing the prediction of a model fit using the remainder of the data, provides a reasonable approach for assessing models, but is time consuming and not feasible to apply for all observations in large data sets. We report the first large-scale formal model selection for BBS data, applying LOOCV to stratified random samples of observations from BBS data. Our results are for 548 species of North American birds, comparing the fit of four alternative models that differ in year effect structures and in descriptions of extra-Poisson overdispersion. We use a hierarchical model among species to evaluate posterior probabilities that models are best for individual species. Models in which differences in year effects are conditionally independent (D models) were generally favored over models in which year effects are modeled by a slope parameter and a random year effect (S models), and models in which extra-Poisson overdispersion effects are independent and *t*-distributed (H models) tended to be favored over models where overdispersion was independent and normally distributed. Our conclusions lead us to recommend a change from the conventional S model to D and H models for the vast majority of species (544/548). Comparison of estimated population trends based on the favored model relative to the S model currently used for BBS summaries indicates no consistent differences in estimated trends. Of the 18 species that showed large differences in estimated trends between models, estimated trends from the default S model were more extreme, reflecting the influence of the slope parameter in that model for species that are undergoing large population changes. WAIC, a computationally simpler alternative to LOOCV, does not appear to be a reliable alternative to LOOCV.

Key words: Bayesian analysis; bayesian predictive information criterion; hierarchical models; leave-one-out cross-validation; model selection; North American Breeding Bird Survey; Watanabe/Akaike information criterion.

Introduction

The North American Breeding Bird Survey (BBS) provides count data used to monitor population change for over 500 species of birds (Pardieck et al. 2019). The survey, in which data are collected annually by skilled observers on preestablished survey routes, began in 1966 in the eastern United States and has expanded in size and geographic extent (Sauer et al. 2013). By 2018, the BBS database contained information from 5,745 routes across the contiguous United States, Alaska, and Canada, of which ~3,000 are surveyed annually. Additional survey routes have been established in Mexico, but are not routinely analyzed due to limited data.

The BBS is the only source of information on population change for most North American bird species and is the primary data source for conservation status assessments (Rosenberg et al. 2016) and State of the Birds Reports (North American Bird Conservation Initiative 2016). The United States Geological Survey (USGS) publishes estimates of population change at a variety of spatial scales, with results published via yearly website updates (e.g., Sauer et al. 2014). The wide use of these results is evidenced by over 3,400 published citations of the USGS website in the last 20 yr.

BBS data present a challenge for analysts. The survey spans the continent, varies greatly in consistency of coverage over space and time, and is conducted by thousands of observers that vary in skills (Sauer et al. 2013). Change in observers is an important feature of the data: the typical BBS observer provides four counts over 5 yr of service on a BBS route. Observers' participation duration on a route has 25th, 50th, 75th, and 90th percentiles of 1, 5, 12 and 23 yr (Link and Sauer 2016*a*). Bird species differ greatly in life history attributes and geographic distributions, and these differences also have consequences for statistical analysis.

To accommodate these complexities in the analysis, many analysts of BBS data use overdispersed Poisson

Manuscript received 24 January 2020; accepted 24 February 2020. Corresponding Editor: Paul B. Conn.

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

regression models, analyzed as Bayesian hierarchical models (e.g., Sauer and Link 2011). Expected counts are functions of observer effects and spatially stratified year effects, which we now describe.

Observer effects are of two sorts, “among-observer” and “within-observer.” Among-observer effects describe variation in count rates for different observers under identical circumstances. Temporal changes in counts attributable to changes in the pool of observers are well documented (Link and Sauer 1998); overlooking these effects biases estimates of population change, typically leading to positive biases in trends. Within-observer effects reflect temporal changes in an individual’s counts that are not attributable to changes in population size. For example, counts tend to be slightly lower than expected in an observer’s first year of service (Kendall et al. 1996) perhaps due to unfamiliarity with the route and survey. Within-observer effects have been used to investigate age-related change in counts (Link and Sauer 1998) and experimental changes in count protocols (Sauer et al. 2019). Exploration of nuisance factors such as observer differences on counts is a critical component of BBS analyses, as the field methods of the survey do not provide for the collection of covariates that could be used to directly model observer effects on counts (Sauer et al. 2017*b*).

Year effects reflect temporal and spatial changes in population change, the biologically relevant signals in noisy count data. Year effects are stratified by the intersection of states or provinces and Bird Conservation Regions (BCRs); the BCRs are physiographic regions that define major habitats relevant for birds across North America (Sauer et al. 2013). While it is possible to estimate year effects at the stratum level without further modeling, estimation is greatly enhanced by the use of a hierarchical model in which year effects are treated as random effects. The model currently used by the USGS (Model S) treats year effects as normally distributed with linearly trending means on the log scale; the slopes and intercepts of the regression coefficients, and the residual variance parameter are allowed to vary among strata, themselves as random effects.

Model S has the benefit of flexibility, being applicable to data of widely different quality. 493 of the 548 species considered in this paper have data for at least 47 of the 50 yr from 1966–2015. Among these, sample sizes range from 113 to 99,695, means from 0.05 to 83.5, coefficients of variation from 0.07 to 1.08, and numbers of strata ranging from 1 to 163. For many species, the data overwhelm the prior: the fitted trajectory (pattern of population of change) is distinctly nonlinear. For species with weak data, Model S is appropriate for estimation of a long term pattern of population trend.

Nevertheless, Model S is only one of many, infinitely many, models that can be fit to BBS data. The ease of fitting complex models via Markov chain Monte Carlo (MCMC) means that many models can be considered. Software for MCMC (e.g., JAGS [Plummer 2003], R2Jags, [Su and Yajima 2015]) requires nothing more of the analyst than the specification of models and priors; existing models are easily tweaked and new results produced. The merits of alternative models cannot be addressed on purely subjective grounds, such as interesting patterns in results, nor can a model be viewed as preferable because of its smaller standard errors, since the validity of the standard errors depends on the appropriateness of the model. Clearly, there is a need for objective model selection, based on sensibly defined criteria (Chatfield 1995, Burnham and Anderson 2002, Link and Barker 2006, Hooten and Hobbs 2015).

This paper presents results of the first large-scale formal model selection exercise for the BBS. We compare four models for 548 species, using data for the period from 1966 to 2015. The four models are a 2 × 2 cross-classification of models for year effects and models for overdispersion. We compared models using the Bayesian predictive information criterion (BPIC) and evaluated the Watanabe/Akaike information criterion (WAIC) as a convenient, computationally fast alternative to BPIC (Gelman et al. 2014, Link and Sauer 2016*b*).

We begin by describing a set of four candidate models for BBS data. These include the model presently used in the USGS analyses, models with alternative structures for population change, and models with alternative patterns of overdispersion relative to the Poisson distribution. Next, we describe the BPIC and WAIC model selection criteria. Both relate to the posterior predictive distribution, which is a well-known basis for evaluating model fit. We provide details on the calculations involved in our model selection, then describe the methods we use to compare results from the selected model with the model presently used by the USGS (Model S). We report results of model selection for 548 North American bird species based on BPIC and comment on the usefulness of WAIC as a surrogate for BPIC (comparisons of results based on BPIC and WAIC are provided in an appendix). Finally, we compare results from the selected model with those from Model S.

Model Set For BBS Data

Features shared among all models

The models we considered for BBS data are all overdispersed Poisson regressions of the form $Y_i|\lambda_i \sim P(\lambda_i)$ where Y_i’s are conditionally independent given their means, and

$$\log(\lambda_i) = \Gamma_i + \Omega_i + \varepsilon_i; \tag{1}$$

here Γ_i, Ω_i, and ε_i are year, observer, and overdispersion effects, respectively.

Year effects reflect the influence of bird abundance on counts. No attempt is made to model absolute abundance because of the questionable assumptions needed to do so (Barker et al. 2018, Link et al. 2018); instead, Γ_i

is a measure of relative abundance, a function of parameters describing spatial and temporal patterns in bird populations. In all of the models we consider, $\Gamma_i \equiv \gamma_{s(i),y(i)}$; here, indirect indexes $s(i)$ and $y(i)$ denote the stratum and year of the ith observation. Thus year effects are described by a set of parameters $\gamma_{s,y}$, modeled as functions of stratum s and year y.

Observer effects reflect biologically irrelevant variation in counts related to differences among observers, and differences within observers through time. In the models we consider, these include a fixed effect η for an observer's first year of service on a route, and a mean-zero normal random effect ω_i for each combination of observer and route. Thus $\Omega_i = \omega_{o(i)} + \eta\, f(i)$; here, $o(i)$ denotes the observer that produced count i and $f(i)$ is an indicator variable for whether count i is the observer's first count. Observer effects ω_o are mean zero normal random variables with precision (1/variance) τ^{ω}.

Conditional on the mean parameter λ, the mean and variance of Poisson random variables are identical. BBS counts are substantially more variable than their means would indicate, hence overdispersion effects are modeled through the addition of a mean zero random effect ε_i in the linear predictor (Eq. 1).

Slope, difference, and heavy-tailed models

The four models we consider are labeled D, DH, S, and SH. Model S is a slightly modified version of the model that has been used for most BBS analyses since 2011 (e.g., Sauer and Link 2011, Sauer et al. 2014). Labels S (for "slope"), D (for "difference") and H (for "heavy tails") correspond to features of the models, which we now describe.

The S models (S and SH) assume that year effects $\gamma_{s,y}$ are conditionally independent and normally distributed with precision τ_s^{γ}; the precision is allowed to vary among strata. The expected value of $\gamma_{s,y}$ is

$$E\left(\gamma_{s,y}\right) = S_s + \beta_s(y - y_0)$$

here, the intercept S_s is a baseline abundance parameter for stratum s, β_s is a trend parameter, and y_0 is a baseline year to center the regression. Baseline abundance and trend parameters are allowed to vary by stratum; S_s and β_s are modeled as random effects, across strata. The "slope" designation for models S and SH relates to the linear component of the model described. This linear component is solely a prior expectation, in the absence of data: actual year effects vary, and the model is capable of detecting distinctly nonlinear population trajectories.

The D models (D and DH) replace the assumption that year effects $\gamma_{s,y}$ are conditionally independent, with the assumption that *differences* in year effects are conditionally independent. Thus, $\gamma_{s,y}$ is normally distributed with mean $\gamma_{s,y-1}$ and precision τ_s^{γ}. As with the slope models, we fix a baseline year y_0, and set $E\left(\gamma_{s,y_0}\right) = S_s$, at the stratum mean. Given that population sizes are likely to be temporally autocorrelated, the D models present an appealing alternative to the S models.

The D models have constant prior expectation, but like the S model, are capable of detecting distinctly nonlinear population trajectories. With good data, estimated patterns of population change for S and D models can be nearly identical (Link and Sauer 2016*b*). Differences between fits relate to Bayesian shrinkage: in S models, year effect estimates are influenced by long term trends, while in D models, they are more heavily influenced by values in adjacent years.

In models S and D, extra-Poisson overdispersion effects ε_i are assumed to be independent and normally distributed with precision τ^{ε}. The normal model might not adequately account for extreme counts. Our experience is that extreme counts are a regular feature of BBS data, suggesting the need for a heavy-tailed alternative to the normal distribution. Thus the H models (SH and DH) specify a central t distribution for ε_i in modeling extra-Poisson variation, in place of a normal distribution. The t distribution has scale parameter τ^{ε} and degrees of freedom parameter ν. Following Juárez and Steel (2010) we use a Gamma distribution with mean 20 and variance 200 as an objective prior for ν. In all four models, mean parameters are assigned flat normal priors (mean zero with standard deviation = 1,000), and precision parameters are assigned vague gamma priors (shape parameter = rate parameter = 0.001). JAGS code for MCMC analysis is provided in Appendix S1.

Composite trends

Analyses of BBS data are typically summarized by estimates of trend and annual indices of abundance, computed for states, physiographic regions, countries, or for the entire survey area.

At the stratum level, annual indices produced by the USGS under model S are based on expected counts in the region. These are of the form $n_{s,y} = \exp(\gamma_{s,y} + \frac{1}{2}(\sigma^{\omega})^2 + \frac{1}{2}\left(\sigma^{\epsilon})^2\right)$, where $\sigma^{\omega} = 1/\sqrt{\tau^{\omega}}$ and $\sigma^{\epsilon} = 1/\sqrt{\tau^{\varepsilon}}$ are the standard deviations of the random effects distributions for observer effects and overdispersion, respectively. These expected counts were derived under the assumption that the ε_i's and ω_o's follow normal distributions (Sauer and Link 2011). For the H models, these weights must be modified; σ^{ϵ} is replaced by a multiple of the t distribution's scale parameter. Details are given in Appendix S2.

Regional trends and annual indices are derived statistics: the models are fit for all data for a species among the survey strata, statistics are computed at the level of survey strata and then aggregated among the survey strata to form regional estimates (e.g., Sauer and Link 2011). Composite annual indices for groups of strata are area-weighted stratum-level, annual indices. Trend, defined as an interval-specific estimate of geometric mean yearly change, is computed as a ratio of annual indices for the last and first years of the interval, taken to the power $1/(y_{\text{last}} - y_{\text{first}})$ (i.e., 1 over the length of the time interval, in yr).

Model Selection Criteria

We begin by describing two model selection criteria, the BPIC and WAIC. Both will be seen to relate to the posterior predictive distribution, commonly used in model checking.

We denote the complete set of BBS counts for a given species, and associated covariates, by $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)$ and $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, respectively. We assume a model M for the data, in which the observations Y_i are conditionally independent, given $\mathbf{X}$, with probabilities $\Pr(Y_i|X_i, M, \theta^M)$.

Given a prior distribution for θ^M, we can compute a posterior distribution $\pi(\theta^M|\mathbf{D}, M)$, where $\mathbf{D} = (\mathbf{Y}, \mathbf{X})$. Using this, the posterior predictive distribution is calculated as

$$\text{ppd}(y|X, \mathbf{D}, M) = \int \Pr(y|X, M, \theta^M)\pi(\theta^M|\mathbf{D}, M)d\theta^M.$$

The distribution function $\text{ppd}(y|X, \mathbf{D}, M)$ predicts the probability that a new observation Y with covariate X will take the value y, given that Y is generated by model M, taking into account the uncertainty associated with the parameter θ^M. The posterior predictive distribution is familiar as the basis of posterior predictive checks, used in assessing goodness of fit for Bayesian models.

Given a set of models $\mathcal{M}$, we seek the model $M \in \mathcal{M}$ that is best able to predict new data based on the data at hand. A good predictive model M based on data $\mathbf{D}$ should produce $\text{ppd}(y|X, \mathbf{D}, M)$ that is close to the true but unknown (data generating) distribution, which we denote by $f(y|X)$. Using the Kullback-Liebler divergence as a measure of closeness, the model $M \in \mathcal{M}$ that is best in this sense maximizes

$$\psi^M = E_f(\log(\text{ppd}(Y|X, \mathbf{D}, M))). \tag{2}$$

Given a hypothetical new observation $\tilde{Y}$ sampled from $f(y|\tilde{X})$,

$$\log(\text{ppd}(\tilde{Y}|\tilde{X}, \mathbf{D}, M))$$

is an unbiased estimate of ψ^M. However, if we evaluate the log of the posterior predictive distribution at the value of an observed datum (Y_i, X_i), the result

$$\log(\text{ppd}(Y_i|X_i, \mathbf{D}, M)) \tag{3}$$

is an overestimate of ψ^M; this happens because we have used observation Y_i in predicting its own outcome. Leave-one-out cross-validation (LOOCV) mitigates this bias by instead computing

$$B_i^M = \log(\text{ppd}(Y_i|X_i, \mathbf{D}_{-i}, M)) \tag{4}$$

here, $\mathbf{D}_{-i}$ is the data set $\mathbf{D}$, but with observation Y_i omitted. The necessity and benefit of LOOCV are illustrated in Appendix S3. If using MCMC, calculation of Eq. 4 is accomplished by averaging $\Pr(Y_i|X_i, M, \theta^M)$ against draws of θ_M from $\pi(\theta^M|\mathbf{D}_{-i}, M)$, then taking the logarithm of the average.

The Bayesian predictive information criterion (BPIC) is defined as the sum across all observations of Eq. 4, viz.,

$$\text{BPIC} = \sum_{i=1}^{N} B_i^M = \sum_{i=1}^{N} \log(\text{ppd}(Y_i|X_i, \mathbf{D}_{-i}, M)). \tag{5}$$

BPIC is regarded as the most reliable option currently available for selection among hierarchical models (Gelman et al. 2014, Link and Sauer 2016*b*).

Calculating BPIC is computationally intensive. For each of the N observations, calculating Eq. 4 involves calculation of $\pi(\theta^M|\mathbf{D}_{-i}, M)$, the posterior distribution for θ^M based on the data set $\mathbf{D}_{-i}$. While importance sampling can be used to obtain samples of $\pi(\theta^M|\mathbf{D}_{-i}, M)$ based on $\pi(\theta^M|\mathbf{D}, M)$, the procedure is unstable, and it is safest to perform a new MCMC analysis. The computations required are substantial.

For example, the BBS White-winged Dove (*Zenaida asiatica*) data set has roughly the median number of observations, with 7,389, collected on 24 strata. On a workstation equipped with Xeon E5-2630 v3 processors, MCMC with chain length 10,000 takes approximately 12 minutes running on a single core. Running 7,389 LOOCV analyses, even taking advantage of multicore processing on the system's 16 cores, would take over 3 d. And that's for a single model. We estimate that the corresponding analysis for the widespread Mourning Dove (*Zenaida macroura*) with 98,372 observations on 158 strata would take something over 3 months.

Computational expense can be reduced by computing BPIC on a subset of $n = N$ observations (Link and Sauer 2016*b*). We chose to conduct approximately $n = 100$ analyses per species, rather than the average $N = 18{,}778$ analyses that would have been required for full computation of BPIC.

The Watanabe/Akaike Information (WAIC) avoids the use of LOOCV, using Eq. 3 instead of Eq. 4 in estimating ψ^M. A bias correction term is added, so that WAIC is defined as

$$\begin{aligned}\text{WAIC} = \sum_{i=1}^{N} \{&\log(\text{ppd}(Y_i|X_i, \mathbf{D}, M)) \\ &- \text{Var}_{\text{post}}(\Pr(Y_i|X_i, M, \theta^M))\},\end{aligned} \tag{6}$$

where Var_{post} denotes the posterior variance of θ^M based on the full data set $\mathbf{D}$. WAIC is asymptotically equivalent to BPIC (Gelman et al. 2014) and is much more easily computed than BPIC, because it requires only one analysis of the data. We calculated WAIC on the full sets of N observations, as well as on the subsets of n observations, to evaluate its use in approximating BPIC.

The BPIC and WAIC definitions given are totals across observations, with larger values indicating better predictive value of the data, based on the data **D**. Some authors multiply by −2 to put these criteria on the same (positive) scale as the Akaike, Bayes, and deviance information criteria, in which case smaller values are favored. The only scaling used in this paper will be by $1/n$ to treat the subset BPIC criterion as an estimate of ψ^M. That is, we will write

$$\bar{B}^M = \frac{1}{n}\sum_{j-1}^{n} \log(\mathrm{ppd}(Y_{i_j}|X_{i_j}, \mathbf{D}_{-i_j}, M)), \qquad (7)$$

where $\{i_1, i_2, \ldots i_n\}$ is a randomly selected subset of n indices from $\{1, 2, \ldots, N\}$.

Model Selection Calculations For BBS Data

Our goal was to select models for 548 species of birds surveyed by the BBS, with particular interest in the trajectory component of the models. Our primary tool for model selection was BPIC, calculated for a subset of roughly 100 counts per species. The geographic range of the BBS and the density of routes have increased with time, meaning that species' data sets are more heavily weighted toward recent years (Fig. 1). We chose to sample two counts per year for each species, so that years were equally represented in evaluating population trajectories. Variation in the geographical extent of the survey resulted in different sample sizes, ranging from 34 to 100, with 457 species having $n \geq 96$.

For each species and model, we began by running a Markov chain of length 10,000, saving the final state of the chain as a burned-in starting value for the subsequent n leave-one-out analyses. The n LOO analyses are readily performed in parallel on computers with multicore processors. Having selected n indices $i_1, i_2, \ldots, i_n$, the jth analysis uses the same MCMC code as the full analysis, but with observation i_j omitted, and a node calculating $\Pr(Y_{i_j}|X_{i_j}, M, \theta^M)$ monitored for a further 10,000 Markov chain samples. The log of the posterior mean for this node is the i_jth summand in the definition of BPIC, Eq. 5, which we denote by B_{i_j}.

Finally, for each species and model, we performed one more analysis of the full data set via MCMC (using the burned-in starting values previously obtained) in order to compute WAIC values. Memory requirements were substantial because of the need to monitor N nodes, one per observation. We found that chains of length 2,500 were satisfactory for stable and precise estimates of the summands of WAIC (Eq. 6); we denote the ith value by W_i.

Analysis of Selected Model

By 1970, the BBS was well established in the contiguous United States and southern Canada. We refer to this area as the core BBS survey area. For 426 species that were encountered in this area we present trends for the interval 1970–2015. Post 1970, BBS routes were established in Alaska and parts of Canada outside the core survey area. In these non-core areas, sporadic data exist before 1993, and Sauer et al. (2017*a*) chose 1993 as the first year when sufficient data existed for analysis of an additional 122 species not found in the core area. Here, we present trends from 1993 to 2015 for 122 species found only in the non-core area.

For the 548 species, we present trend results for Model S (our base model) and for the selected model based on the posterior probability that the model is best from the BPIC hierarchical model. For each model, we present an estimate of trend (the posterior median), a 95% credible interval (percentiles 2.5 and 97.5 of the posterior distribution), and the annual index for the midyear. Trend estimates represent yearly percentage change for the interval 1970–2015 (for core species) or 1993–2015 (for non-core species). We also present the number of survey routes used in the analysis for the species. For these trend estimates, we summarize differences in results between model S and the selected model.

If the model selection procedure tends to generally favor D models or S models, the model selection procedure might result in systematic differences in trends between results from the default model and selected model. Prior work has shown that the differences in year effects parameterizations of D vs. S models can lead to systematic differences in trends. S models model a slope parameter, on which year effects are deviations, while the D models directly model year to year changes. Because the BBS had very limited data in the early years of the survey, we expect that these differences between the models lead to more extreme trends for the S models (as the trajectory in the early years of the survey is dominated by the slope parameter in the model). We also predict that D models will be less precise for years with weak data, for similar reasons, i.e., the slope-based model is dominated by a slope parameter while the D models are dominated by poorly estimated year effects.

We used paired t tests to determine whether systematic differences in trends existed between trend results from the default model and the selected model at the continental (survey-wide) scale of summary. We also evaluated whether half-widths of 95% credible intervals of trend were consistently different between the default and selected models. We also identify species for which the 95% CIs of trend estimates did not overlap between the default model and the selected model. Although this criterion is of limited quantitative significance due to lack of independence, it does highlight species for which the change in model selected can have large effects on the trend estimates. For these species, we provide graphs of

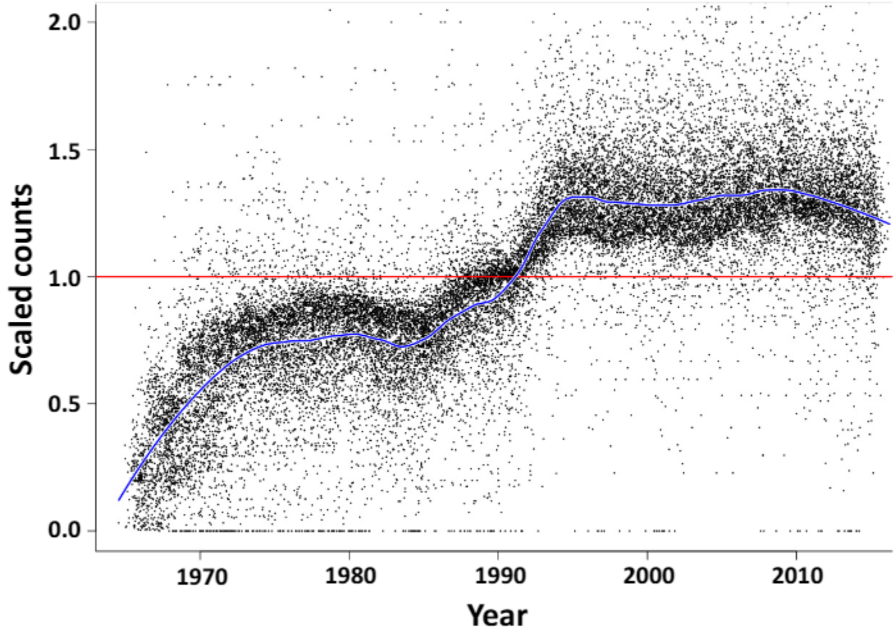

FIG. 1. Scaled number of counts for the North American Breeding Bird Survey, each point corresponding to a species and a year. Each number has been scaled by the mean number of counts for the species over years subsequent to the species' first appearance in the data set. Blue curve is LOESS smooth. Year values have been jittered to prevent overplotting.

annual indices for default models and the selected models. We conducted summary analyses separately for core and non-core species.

RESULTS: MODEL SELECTION

In an earlier paper investigating a set of four models in application to 20 BBS species (Link et al. 2017) we noted that the components of WAIC (W_i^M) were poor surrogates for the components of BPIC (B_i^M). We confirmed that observation in the present study, also finding that model selection by WAIC corresponds only weakly to model selection using BPIC. Details are included in Appendix S4. The results presented in this section are based on BPIC.

BPIC results

Let $\bar{B}_s^M$ denote the BPIC value based on the subsample of size $n \ll N$ for species $s = 1, 2, \ldots, 548$, as defined at Eq. 7. Ranking models based simply on raw values $\bar{B}_s^M$, model D is ranked as best 141/548 times (25.7%), model DH 203 times (37%), model S 81 times (14.8%), and model SH 123 times (22.4%). On the other hand, model D is ranked as worst 138 times (25.2%), model DH 70 times (12.8%), model S 213/548 times (38.9%), and model SH 127 times (23.2%). This quick appraisal (summarized in Table 1) suggests that across species, there is a tendency for DH to be favored and S to be disfavored.

Associated with each vector $\hat{\mathbf{F}}_s = \left(\bar{B}_s^D, \bar{B}_s^{\mathrm{DH}}, \bar{B}_s^S, \bar{B}_s^{\mathrm{SH}}\right)'$ is an estimated covariance matrix $\hat{\Sigma}_s$ (the sample covariance matrix for the n sampled values) from which one can calculate standard errors of differences, such as $\left(\bar{B}_s^D - \bar{B}_s^{\mathrm{DH}}\right)$. Using these, we calculate z statistics to test null hypotheses of equal support for models (see Discussion in Link and Sauer 2016*b*:1756). Most of the differences are not precisely estimated, with the result that relatively few of the model comparisons are significant at $\alpha = 0.05$ (274 out of 548 × 6, or 8.3%).

Nevertheless, inspection of Table 2 suggests that general tendencies emerge in comparisons of models, aggregated across species. The last row of Table 2 (labeled "Null") gives the expected frequencies for a sample of 548 normal random variables; these are the expected frequencies under a null hypothesis of no difference in model fits. For each of the six model comparisons, the distribution of results is shifted to the right or left relative to this baseline. Comparisons D vs. DH and S vs. SH have outcome frequencies shifted to the left, favoring DH over D, and SH over S. The other four comparisons have outcome frequencies shifted to the right, favoring D over S and SH, and DH over SH.

TABLE 1. Rank frequencies (freq; based on sampled Bayesian predicted information criterion [BPIC] values) across species.

Rank	D		DH		S		SH	
	Freq	%	Freq	%	Freq	%	Freq	%
Best	141	25.7	203	37.0	81	14.8	123	22.4
Worst	138	25.2	70	12.8	213	38.9	127	23.2

These results are consistent with an ordering DH > D > SH > S of general tendencies in preferences among models (here " > " means "has better predictive value than"). Heavy-tailed models and difference models are preferred, and of the two innovations, difference models offer the greater gain.

We will say that model M_1 is favored over model M_2 if the z statistic is greater than 1.96. There are only three species for which DH is favored overall, and only one where D is favored overall. S and SH are never favored overall.

For 33 species, model D or DH is favored over both S and SH. Of these, there is no clear preference for D or DH in 29 cases, with DH being favored over D in three cases, and D over DH in 1. On the other hand, there are only three species for which S or SH is favored over both D and DH; in none of these is either S or SH favored.

For 15 species, heavy-tailed models DH or SH are favored over the non-heavy-tailed alternatives D and S. For three of these, DH is favored over SH; in none is SH favored over DH. On the other hand, there are only three species for which D or S is favored over both DH and SH. D is preferred over S in one of these; neither S nor D is favored in the others.

Species treated as replicates in hierarchical model for BPIC

The components B_i^M of BPIC vary among models and among observations Y_i. The variance among observations is considerably larger than the variance among models: for the 548 data sets, the median ratio of these variances was 256.6, and 95% exceeded 44.4. Consequently, our subsampling of $n \ll N$ LOO calculations introduced noise to the estimation of ψ^M.

Nevertheless, general tendencies toward favoring certain models emerged, across species. First difference models D and DH tended to be favored over models S and SH, and heavy-tailed models DH and SH tended to be favored over D and S. These general tendencies are reasonably interpreted as reflecting fundamental differences in the fit of the models and are appropriately evaluated using a hierarchical model. Hierarchical modeling also enhances model selection for individual species by considering them collectively, "borrowing strength from the ensemble" (Morris 1983, Louis 1984).

For species s, let $\hat{\Delta}_s = C\hat{\mathbf{F}}_s$, where C is the 3×4 contrast matrix

$$C = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix}.$$

That is, $\hat{\Delta}_s = \left(\bar{B}_s^D - \bar{B}_s^{\mathrm{DH}}, \bar{B}_s^S - \bar{B}_s^{\mathrm{DH}}, \bar{B}_s^{\mathrm{SH}} - \bar{B}_s^{\mathrm{DH}}\right)'$. We treat $\hat{\Delta}_s$ as having a trivariate normal distribution with mean vector Δ_s and known covariance matrix $\mathbf{V}_s = C\hat{\Sigma}_s C'$. We assume a trivariate normal distribution for the means, i.e., $\Delta_s \sim \mathcal{N}_3(\boldsymbol{\mu}, \mathbf{W})$, assign a flat normal prior to $\boldsymbol{\mu}$, and a vague inverse Wishart prior to $\mathbf{W}$ (3 degrees of freedom, with 3×3 identity matrix as the scale matrix). We analyzed this hierarchical model for BPIC values, obtaining posterior distributions for Δ_s given the complete collection of estimates, $\hat{\Delta}_\bullet = \left(\hat{\Delta}_1, \hat{\Delta}_2, \ldots, \hat{\Delta}_{548}\right)$ as data.

The latent vector Δ_s consists of differences $\psi_s^D - \psi_s^{\mathrm{DH}}, \psi_s^S - \psi_s^{\mathrm{DH}}$ and $\psi_s^{\mathrm{SH}} - \psi_s^{\mathrm{DH}}$, with ψ^M defined at Eq. 2 and subscript s added for species. Recalling that larger values of ψ^M indicate better expected predictive value for the combination of data and model, $\Delta_{s,1} > 0$ means that model D is to be preferred over model DH, $\Delta_{s,2} > 0$ means that model S is preferred over model DH, and $\Delta_{s,3} > 0$ means that model SH is preferred over model D. The four models can be ranked based on Δ_s; for example, $\Delta_s = (-1.2, 3.8, -2.5)$ implies the ordering S > DH > D > SH. Thus we obtained values $\pi_s^M = \Pr(M$ is best for $s|\hat{\Delta}_\bullet)$ and $\rho_s^M = \Pr(M$ is worst for $s|\hat{\Delta}_\bullet)$. Results are summarized in Table 3 and Fig. 2; results by species are given in Data S1: Table of Trend Results 1970–2015.csv.

Results: Selected Model vs. Traditional Model

Results for the 548 species are presented in Data S1: Table of Trend Results 1970–2015.csv. We note that four species had Model S as the preferred model, and we include those results in summary statistics. These species cover a large array of sampling and life history situations, varying in samples from 3 to 4,423 routes (median = 341), and in estimated abundance from 0.003 to

Table 2. Paired z test statistics for BPIC differences.

Model	< -1.96	$[-1.96, 0)$	$[0, 1.96)$	≥ 1.96	Mean	SD
D vs. DH	21	315	203	9	−0.28	1.03
D vs. S	6	205	289	48	0.37	1.13
D vs. SH	17	240	263	28	0.07	1.12
DH vs. S	5	176	314	53	0.51	1.06
DH vs. SH	4	196	309	39	0.39	1.08
S vs. SH	36	299	205	8	−0.32	1.09
Null	14	260	260	14	0	1.00

Notes: Labels S (for "slope"), D (for "difference") and H (for "heavy tails") describe features of the models. Models are described in *Model set for BBS data: Slope, difference, and heavy-tailed models.*

20,073 (median = 48.9). Because D models were selected for the majority of species, we expected to see systematic differences in results between the default and selected models.

For 426 core species with trends estimated for 1970–2015, we see no consistent differences between trends from the default model and the selected model (mean difference, default – selected = 0.04%/yr, 95% CI −0.073, 0.016); for the 122 non-core species, trends from the default models were consistently higher (mean difference = 0.99%/yr, 95% CI 0.024, 1.740) presumably reflecting the effects of the differences between S and D models in regions with limited data.

Comparisons of half-widths of credible intervals indicated that trends tended to be less precise in the selected model than in the default model for the core species results (mean difference = −0.40, 95% CI −0.558, −0.236), but for the non-core species results the half-widths of the CIs were greater for the default model than for the selected models (mean difference = 2.96, 95% CI 2.180, 3.733).

Eighteen of the core species had annual indices for which the 95% CIs under the default and selected models did not overlap. We present the annual indices for these species in Appendix S5.

Discussion

Choosing among alternative statistical models for analysis of BBS data is an important means of advancing our understanding of population change in North American birds.

BBS analyses have the dual role of controlling for factors that influence counts (e.g., observer effects) and modeling population change. Fortuitously, advances in statistical modeling and scientific computing have provided a means for reasonable analyses; controlling for observer effects and modeling population change at the continental scale can be accomplished with hierarchical models (Sauer et al 2017*b*). But this modeling has a cost: choice of appropriate models is complicated by the size of the data set, the unruly distributional nature of the data, the complexity of the models, and the temporal and spatial extent of the survey. Leave-one-out cross-validation provides a reasonable approach for model selection, and the BPIC statistic, as an observation-based metric of fit, provides greats flexibility for evaluating temporal and spatial patterns of model fit. The challenge has been to apply the approach in a computationally feasible manner to the 548 BBS species for which analyses are conducted. Here, we have sampled observations for BPIC in a temporally balanced design to assess model fit for a model set that contrasts two alternative parameterizations of year effects and two overdispersion distributions, and applied a hierarchical model to the model selection results to compute posterior probabilities that models are best, by species. Results suggest that models that use the D year effects parameterization and *t*-distribution-based overdispersion are favored for most species. The superiority of the D models and *t*-distribution-based overdispersion conformed to our expectations on biological grounds, as noted in our description of the models.

Table 3. Posterior probabilities that model is best (π_s^M) or worst (ρ_s^M), summarized across species.

Parameter	D	DH	S	SH
Highest π_s^M	53	458	4	33
Mean π_s^M	27%	51.5%	4.6%	17%
Highest ρ_s^M	13	0	524	11
Mean ρ_s^M	11.5%	4.1%	69.9%	14.6%

Our results begin with simple comparisons of sampled BPIC values, then follow with statistical tests comparing fits between models, ending with a hierarchical model that allows us to obtain posterior probabilities that each model is best for a given species. Each successive refinement of analysis enhanced the view of DH and D as the top-ranked models. Based on these results, it seems sensible to abandon the S model as our default model as it has the highest posterior probability of being the worst model for 524 species and the highest posterior probability of being the best model for only four species. Model DH seems a much better candidate as a default model, with highest posterior probability of being the worst and best model for 0 and 458 species, respectively.

A concern with use of BPIC is the challenge to fully implement it for most BBS species. Our subsample analysis, based on $n \ll N$ observations, was necessitated by the computational burden of computing values B_i^M (Eq. 4), the components of BPIC. As it turns out, the variation in B_i^M across models is small relative to the variation across observation Y_i. Sampling a larger portion of the observations would clearly lead to increased power of statistical tests among models and also provide more species-specific information to inform the hierarchical modeling. For species where estimates of population change differ among models, it would seem prudent to conduct larger sampling of observations to better estimate total BPIC.

The numbers of BBS counts have changed over time, as has the spatial extent of the survey. These features of the BBS need to be accommodated in model selection. In the interest of optimizing selection of temporal pattern, we chose to balance years in our sampling of BPICs. We suspect that if we had not included the temporal balancing of sampling, the S models would have been more favored, as presumably a completely random sample of observations would have placed greater emphasis on years in the middle and later years where more routes were sampled.

Larger samples of BPIC components B_i^M would be useful in evaluating model fit in specific areas and periods of interest. For example, species such as Barn Swallow (*Hirundo rustica*) show large differences in population trajectories in the early years of the survey (Appendix S5) between S and D models, and additional

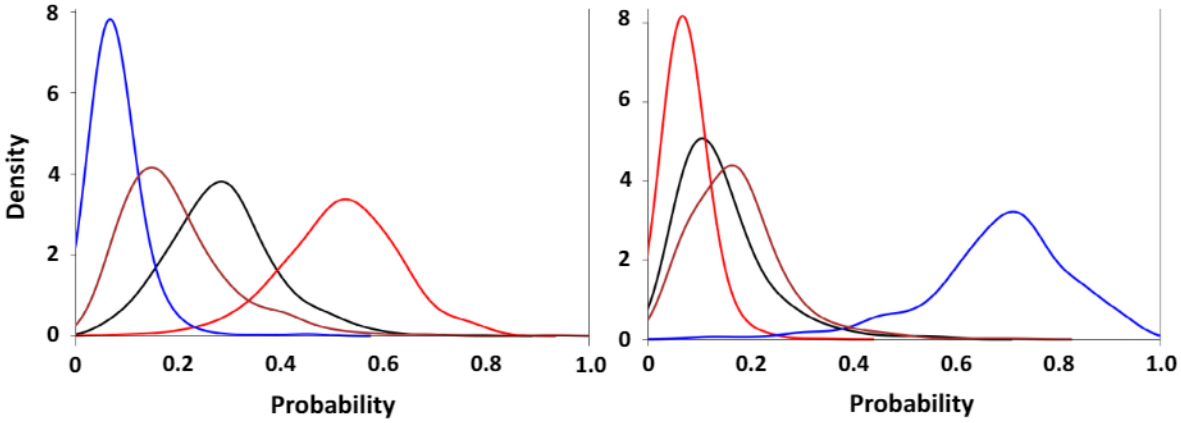

FIG. 2. Smoothed density of posterior probability model is best (left panel, π_s^M) and worst (right panel, ρ_s^M) across 548 BBS species, for models M = D (black), DH (red), S (blue), and SH (brown). Labels S (for "slope"), D (for "difference") and H (for "heavy tails") describe features of the models. Models are described in *Model set for BBS data: Slope, difference, and heavy-tailed models.*

evidence of the superiority of a D model for these years could be provided by sampling additional observations from these years. Barn Swallows also show striking regional differences in population trajectories, with increases in the southern United States countered by strong declines in the central and northern United States and Canada (Sauer et al. 2017*a*). For this species, it would be of great interest to evaluate both spatial and temporal patterns of model fit by mapping observation-specific values of B_i^M.

WAIC offers a potential solution to these concerns about small sample sizes, as it is easily computed for all observations. Unfortunately, results presented here confirm concerns about WAIC as a poor surrogate for BPIC. Matched against the yardstick of BPIC, WAIC results appear to be somewhat better than chance, but only match the BPIC rankings 21% of the time. In particular, the differential bias in WAIC with magnitude of BPIC further limits the value of WAIC in evaluating temporal and spatial patterns of model fit.

The BPIC values presented here have informed our determination of the relative merits of the models in our model set, but were computed at significant cost in terms of computational effort, data storage, and personnel time. Hopefully, they also have some value in helping to design future model selection activities with different model sets. In the short term, for studies that propose alternative models (such as semiparametric smooths to model year effects, or the value of covariates to examine phenology effects on counts) it would be possible to use the data structure of years and sampled observations from the current study and estimate BPIC components B_i^M for the new models, based on the same samples of observations. BPIC for the new models could then be compared directly with the existing data.

Model selection within the model set described here allows for two generalizations with regard to model structure: (1) the D models, in which year effects are defined in terms of changes from adjacent years, seems generally preferable to the S models, in which year effects are defined as deviations from a consistent underlying slope parameter, and (2) the models that allow for *t*-distributed overdispersion are generally to be preferred over models that allow for normally distributed overdispersion. However, empirical results indicate that choice of model does not result in dramatically different views of population change within species. This is comforting, as a strong dependence on models (that are admittedly sometimes difficult to discriminate using our model selection tools) tends to undermine the credibility of results. Also, the distinction between D and S models can be thought of as a difference in process priors for model trajectories. For almost all species the large sample sizes appear adequate for the data to overwhelm the prior.

However, as illustrated by our evaluation of species with non-overlapping credible intervals of selected and default trend estimates, differences tend to occur when a pair of conditions exists: (1) the species is undergoing dramatic and consistent population changes, and (2) the species has very limited data in the early years of the survey. This combination of circumstances leads to cases in which the slope parameter in the S models is the dominant expression of population change and the year effects are poorly estimated. In periods with limited data, predictions of annual indices based on the slope parameter are not particularly informative as they only lead to a linear (on a log scale) prediction of change. On the other hand, the D model does not have the constraint of a consistent prediction of change but instead only predicts change based on the information in the interval and a prior expectation of no change. Thus, the absence of information in the S model predicts the yearly change based on the prior defined by the slope parameter while the D model has a prior expectation of zero change.

The preeminence of the H models also conforms to our expectations regarding the BBS data set. In our experience, BBS data tend to have many extreme observations. These extreme observations lead to lack of fit; the *t* distributions modeling overdispersion in the H models accommodate these extreme observations. The H models introduce a complication in that the mathematical expectation of an exponentiated *t* value is infinite. In previous implementations of models S and D, the expected count has been used as an index of abundance. However, for heavy-tailed overdispersion models (SH and DH) an alternative characterization of a typical count is needed (Appendix S2).

Acknowledgments

We thank the thousands of volunteers who have contributed to the North American Breeding Bird Survey. We also thank Evan Cooch, Allan O'Connell, and an anonymous referee whose comments improved the manuscript. Any use of trade, product, or firm names in this publication is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Literature Cited

Barker, R. J., M. R. Schofield, W. A. Link, and J. R. Sauer. 2018. On the reliability of N-mixture models for count data. Biometrics 74:369–377.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference: a practical information-theoretic approach. Second edition. Springer-Verlag, New York, New York, USA.

Chatfield, C. 1995. Model uncertainty, data mining and statistical inference (with discussion). Journal of the Royal Statistical Society (London), Series A 158:419–466.

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

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

Juárez, M. A., and M. F. Steel. 2010. Model-based clustering of non-Gaussian panel data based on skew-t distributions. Journal of Business and Economic Statistics 28:52–66.

Kendall, W. L., B. G. Peterjohn, and J. R. Sauer. 1996. First-time observer effects in the North American Breeding Bird Survey. Auk 113:823–829.

Link, W. A., and R. J. Barker. 2006. Model weights and the foundations of multimodel inference. Ecology 87:2626–2635.

Link, W. A., and J. R. Sauer. 1998. Estimating relative abundance from count data. Austrian Journal of Statistics 27:83–97.

Link, W. A., and J. R. Sauer. 2016*a*. Modeling participation duration, with application to the North American Breeding Bird Survey. Communications in Statistics—Theory and Methods 45:6311–6320.

Link, W. A., and J. R. Sauer. 2016*b*. Bayesian cross-validation for model evaluation and selection, with application to the North American Breeding Bird Survey. Ecology 97:1746–1758.

Link, W. A., J. R. Sauer, and D. K. Niven. 2017. Model selection for the North American Breeding Bird Survey: a comparison of methods. Condor 119:546–556.

Link, W. A., M. R. Schofield, R. J. Barker, and J. R. Sauer. 2018. On the robustness of N-mixture models. Ecology 99:1547–1551.

Louis, T. A. 1984. Estimating a population of parameter values using Bayes and empirical Bayes methods. Journal of the American Statistical Association 79:393–398.

Morris, C. N. 1983. Parametric empirical Bayes inference: theory and applications. Journal of the American Statistical Association 78:47–55.

North American Bird Conservation Initiative. 2016. The state of North America's birds 2016. Environment and Climate Change Canada, Ottawa, Ontario, Canada.

Pardieck, K. L., D. J. Ziolkowski Jr., M. Lutmerding, V. Aponte, and M-A.R. Hudson. 2019. North American Breeding Bird Survey dataset 1966–2018. Version 2018.0. U.S. Geological Survey, Patuxent Wildlife Research Center, Laurel, Maryland, USA. https://doi.org/10.5066/P9HE8XYJ

Plummer, M. 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. Pages 1–10 *in* K. Hornik, F. Leisch, and A. Zeileis, editors. Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003) March 20–22, Vienna, Austria.

Rosenberg, K. V., et al. 2016. Partners in Flight Landbird Conservation Plan: 2016 Revision for Canada and Continental United States. Partners in Flight Science Committee 119 pp.

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.

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

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., W. A. Link, D. J. Ziolkowski, K. L. Pardieck, and D. J. Twedt. 2019. Consistency counts: modeling the effects of a change in protocol on Breeding Bird Survey counts. Condor 121:duz009.

Sauer, J. R., D. K. Niven, K. L. Pardieck, D. J. Ziolkowski Jr., and W. A. Link. 2017*a*. Expanding the North American Breeding Bird Survey analysis to include additional species and regions. Journal of Fish and Wildlife Management 8:154–172.

Sauer, J. R., K. L. Pardieck, D. J. Ziolkowski Jr., A. C. Smith, M-A.R. Hudson, V. Rodriguez, H. Berlanga, D. K. Niven, and W. A. Link. 2017*b*. The first 50 years of the North American Breeding Bird Survey. Condor 119:576–593.

Su, Y.-S., and M. Yajima. 2015. R2jags: Using R to run 'JAGS'. Version 0.5-7. https://CRAN.R-project.org/package/R2jags

Supporting Information

Additional supporting information may be found online at: http://onlinelibrary.wiley.com/doi/10.1002/eap.2137/full