THE CONDOR
Ornithological Applications
AmericanOrnithology.org

Volume 119, 2017, pp. 546–556
DOI: 10.1650/CONDOR-17-1.1

RESEARCH ARTICLE

Model selection for the North American Breeding Bird Survey: A comparison of methods

William A. Link,* John R. Sauer, and Daniel K. Niven

U.S. Geological Survey, Patuxent Wildlife Research Center, Laurel, Maryland, USA
* Corresponding author: wlink@usgs.gov

Submitted January 3, 2017; Accepted May 11, 2017; Published July 26, 2017

ABSTRACT

The North American Breeding Bird Survey (BBS) provides data for >420 bird species at multiple geographic scales over 5 decades. Modern computational methods have facilitated the fitting of complex hierarchical models to these data. It is easy to propose and fit new models, but little attention has been given to model selection. Here, we discuss and illustrate model selection using leave-one-out cross validation, and the Bayesian Predictive Information Criterion (BPIC). Cross-validation is enormously computationally intensive; we thus evaluate the performance of the Watanabe-Akaike Information Criterion (WAIC) as a computationally efficient approximation to the BPIC. Our evaluation is based on analyses of 4 models as applied to 20 species covered by the BBS. Model selection based on BPIC provided no strong evidence of one model being consistently superior to the others; for 14/20 species, none of the models emerged as superior. For the remaining 6 species, a first-difference model of population trajectory was always among the best fitting. Our results show that WAIC is not reliable as a surrogate for BPIC. Development of appropriate model sets and their evaluation using BPIC is an important innovation for the analysis of BBS data.

Keywords: Bayesian predictive information criterion, cross-validation, hierarchical models, model selection, North American Breeding Bird Survey, Watanabe-Akaike information criterion

Análisis del Conteo de Aves Reproductivas: comparación de modelos y criterios de selección de métodos

RESUMEN

El Conteo de Aves Reproductivas (BBS, por sus siglas en inglés) provee datos para más de 420 especies de aves, en múltiples estalas geográficas, por más de cinco décadas. Los métodos de computación modernos han facilitado el ajuste de modelos jerárquicos complejos a estos datos. Es fácil proponer y ajustar nuevos modelos, pero se ha prestado poca atención a la selección de los modelos. En este estudio comparamos 4 modelos aplicados a 20 especies que varían en abundancia y patrones en sus tendencias. También consideramos la selección de modelos usando validación cruzada dejando uno de los modelos por fuera, y el criterio bayesiano de información predictiva (CBIP). La validación cruzada es computacionalmente muy intensiva; por eso evaluamos el desempeño del criterio de información de Watanabe-Akaike (CIWA) como una aproximación computacionalmente más eficiente que el CBIP. La selección de los modelos basada en el CBIP no presenta evidencia fuerte de que uno de los modelos sea consistentemente superior a los otros; para 14 de 20 especies, ninguno de los modelos se reconoció como superior. Para las 6 especies restantes, un modelo de trayectoria poblacional de primera diferencia siempre estuvo entre los que presentaba mejor ajuste. Nuestros resultados muestran que el CIWA no es confiable como sustituto del CBIP. El desarrollo de los conjuntos apropiados de modelos y de su evaluación usando el CBIP es una innovación importante en el análisis de los datos del BBS.

Palabras clave: Conteo de Aves Reproductivas, Criterio Bayesiano de Información Predictiva, modelos jerárquicos, Criterio de Información Watanabe-Akaike. selección de modelos, validación cruzada

INTRODUCTION

The North American Breeding Bird Survey (BBS) provides 50 yr of data for bird populations in the United States and Canada, monitoring population change for >420 species at multiple geographic scales (Sauer et al. 2013, 2017). The data are counts of all birds seen within 400 m, or heard from any distance, by a skilled observer at each of 50 roadside stops along designated routes.

BBS counts are analyzed as overdispersed Poisson random variables with a log-linear model for the effects of explanatory variables. Explanatory variables include population effects, observer effects, and overdispersion effects ("noise" in Figure 1). Population effects account for spatial and temporal variability in the counts. Observer effects account for differences among survey participants, and for temporal change in individuals' count rates. The latter may include effects due to increased familiarity with

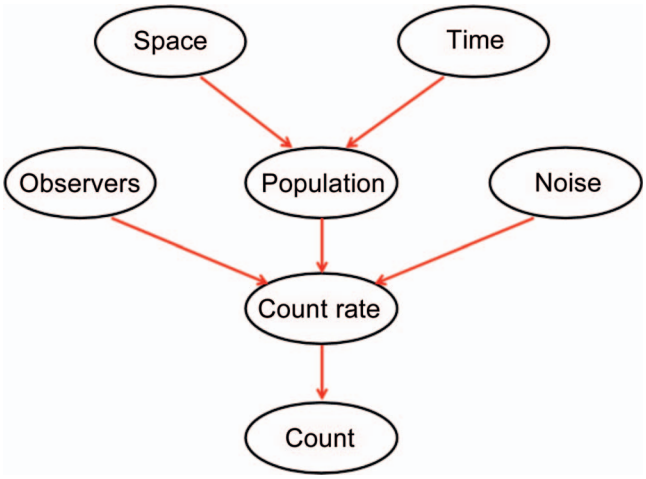

FIGURE 1. Factors influencing North American Breeding Bird Survey data. Counts reflect an underlying population that is changing through space and time, but counting is influenced by observers and random environmental noise.

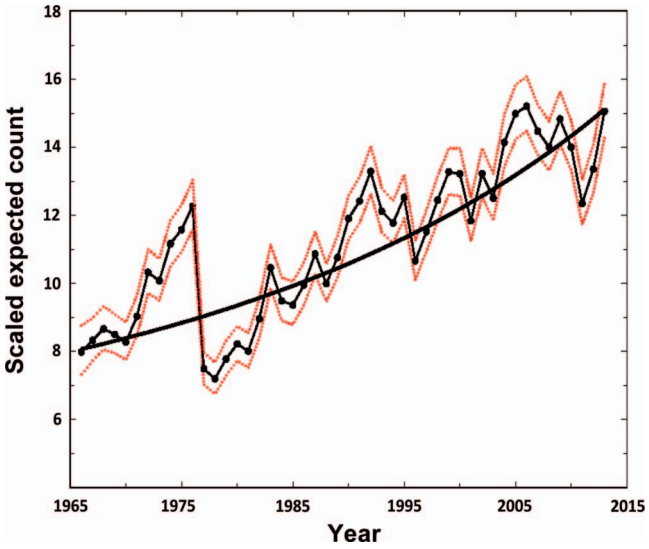

FIGURE 2. Estimated composite population trajectory for the Carolina Wren from the full North American Breeding Bird Survey dataset, with 95% credible intervals. Point estimates are posterior medians, and the solid line is the log-linear component of the trajectory.

survey methods and route locations, or changes relating to the aging of observers.

Overdispersion effects are modeled as observation-specific, mean-zero random variables, additive on the log scale of the expected count. These are included to relax the restrictive relation between the mean and variance of the Poisson distribution, allowing for temporally stable variation in counts not explained by other components of the model. Having modeled the biologically irrelevant effects of observers and accounted for temporally stable overdispersion effects, the remaining variation in counts can reasonably be assumed to describe patterns of population change.

BBS counts are not intended to provide actual population numbers, and naive use of BBS counts to estimate population size or compare populations through space and time are misguided, misleading, and pointless (Sauer and Link 2004). Rather, the usefulness of BBS data lies in model-based estimation of an index to relative population size and its patterns through time and space. Focusing on a temporal pattern, the modeled pattern of variation in this index is referred to as the population trajectory. In combination with population size surveys, BBS trajectories can be used to predict population sizes from areas of overlap and areas covered only by BBS data (Runge et al. 2009, Millsap et al. 2013, Zimmerman et al. 2015, Runge and Sauer in press).

Selection of a model for estimating population change is a crucial component of the analysis of BBS data: the mathematical model used shapes our perception of population change. The model must be flexible enough to capture genuine signals in the data, but simple enough to filter irrelevant variation.

The widely cited and comprehensive analysis conducted by the U.S. Geological Survey (USGS; Sauer et al. 2014, www.mbr-pwrc.usgs.gov/bbs/) describes the population trajectory for an individual species as an additive effect on the log scale of the expected count, having the form:

$$\gamma_s(t) = \alpha_s + \beta_s(t - t^*) + \varepsilon_{st}. \quad (1)$$

Here, t denotes the year of the count and t^* is a reference year; α_S and β_S are spatially varying parameters reflecting baseline abundance and long-term pattern of change, indexed by geographic strata s; ε_{st} is a mean-zero normal random variable. The strata are defined by the intersections of U.S. states or Canadian provinces with Bird Conservation Regions (Sauer et al. 2013), and we focus only on the trajectory component of the model. The model contains additional parameters that describe other effects on counts. Model (1) is referred to as the linearly trending random year effects model.

It is important to recognize that model (1) does not impose a strictly log-linear pattern on the population trajectory. The log-linear component of the model, $\alpha_S + \beta_S\ (t - t^*)$, provides a baseline pattern from which fitted values may depart. Given sufficient data, the latent variables, ε_{st}, are precisely estimated, allowing estimates of $\gamma_S(t)$ to depart from the strictly linear form. This is illustrated in Figure 2, which displays survey-wide summary estimates of $\gamma_S(t)$ and its linear component for the Carolina Wren (*Thryothorus ludovicianus*), converted back to the original count scale. Note, for instance, that a population decline of ~40% associated with the severe winter of 1976–1977 (Link and Sauer 2007) is very much in evidence; this pattern has not been obscured by the use of the log-linear model.

On the other hand, model (1) allows for efficient estimation of a long-term trend, even with relatively weak data. Consequently, model (1) has served well as an omnibus model for population trajectories in the multi-species USGS analysis.

Nevertheless, alternative models might be considered, especially when intensively studying individual species. For instance, a model with the form:

$$\gamma_s(t) = \gamma_s(t-1) + \varepsilon_{st}, \quad (2)$$

referred to as the "first-difference model," might more faithfully reproduce nonlinear patterns in trend. Here, once again, ε_{st} are independent, mean-zero normal random variables; the model is typically parameterized with $\alpha_S = \gamma_S(t^*)$, the baseline abundance for stratum s in reference year t^*.

Using the language of Bayesian modeling, model (1) shrinks year effects toward a long-term trend line, and model (2) shrinks year effects toward adjacent year values. Given a strong signal from the data, the amount of shrinkage is small and the same general estimated pattern is obtained by the 2 models. If we were to overlay the fitted trajectory from model (2) on the fitted trajectory from model (1) in Figure 2, the 2 would closely coincide (Link and Sauer 2016:1754). This is because of the high quality of BBS data for Carolina Wrens; the species is widespread, abundant, and easily detectable.

For other species, trajectory models (1) and (2) might not agree, and the question arises of how we should select among models. Other models might be considered, such as:

$$\gamma_s(t) = \alpha_s + \rho_s\gamma_s(t-1) + \varepsilon_{st}, \quad (3)$$

which might be considered a compromise between models (1) and (2). Models with higher-order autoregressive structure might also be considered. The question of model selection goes further, including choices for how we model observer and overdispersion effects. Even once the general form of a model has been decided upon, decisions are required with regard to hierarchical structure. For instance, in models (1) and (2), one might treat parameters α_S as unrelated, or as random effects sampled from common distributions.

How should we select among models? We have heard practitioners suggest that one model is superior to another, based on its narrower confidence intervals—but confidence intervals are only relevant under the assumption the associated model is correct. We have also heard practitioners suggest that excessively parameterized models might be preferred, on the grounds that less precisely estimated trajectories might somehow better suggest associations between population change and candidate explanatory variables.

What is needed is a formal mechanism for model selection (Hooten and Hobbs 2015, Link and Sauer 2016). Akaike's information criterion (AIC) has been widely used and energetically advocated (Burnham and Anderson 2002), but is known to favor unnecessarily complex models (Barker and Link 2015); furthermore, its extension to hierarchical models is difficult. The deviance information criterion (DIC) is similar to AIC and has been widely used in analysis of hierarchical models, but recent research has called its usefulness into question (Gelman et al. 2014). The gold standard for model selection remains cross-validation, but this is highly computationally intensive.

In this paper, we report on a comparison of 4 models applied to a suite of 20 species. We had 2 goals for conducting and reporting these analyses. First, we were interested in the model set per se. Three of the models were based on the linearly trending random year effects trajectory (equation (1)), but with varying levels of hierarchical structure. We were interested in whether adding structure to the usual USGS model would improve analyses. The fourth model used the first-difference trajectory (equation (2)); we were interested in knowing whether the more flexible structure offered by the first-difference model improved analysis. The suite of 20 species that we selected varied from rare, low-abundance species to very common species and included species with consistent population trends and species undergoing population fluctuations; our primary interest was in determining whether our model selection procedure selected different models as better estimating trends among this diverse set of species and population trajectories.

Naturally, we cannot draw general conclusions about the relative merits of these models for application to other species. At best, the results reported here may be suggestive of general tendencies. More important is the notion of a model set, and available tools for choosing among the models. We envision the omnibus model used in USGS analyses being replaced by an omnibus model set and practical model-selection criteria.

Our second goal was to investigate the performance of a new model-selection criterion, the Watanabe-Akaike information criterion (WAIC; Watanabe 2010, 2013, Gelman et al. 2014). The virtue of WAIC lies in its asymptotic equivalence to leave-one-out cross-validation (LOOCV), despite being much more easily calculated. We thus performed LOOCV analyses for our primary assessment of the model set, and then compared results based on WAIC.

METHODS

Leave-one-out Cross-validation

Leave-one-out cross-validation (LOOCV) of a model M involves setting aside a single observation Y_i from a data

set $\boldsymbol{Y}$, fitting M based on the reduced data set $\boldsymbol{Y}_{-i}$, and predicting the value Y_i based on its covariate vector $\boldsymbol{X}_i$ and the fitted model. We can repeat the process for all observations i, or for some subset of the observations, and compare the predictions with the observed values. Basing the predictions on reduced data sets of $\boldsymbol{Y}_{-i}$ avoids problems of overfitting.

Let $\boldsymbol{\theta}_M$ denote the unknown parameter vector for model M and denote the distribution function of Y_i under model M as $f_M(y|\boldsymbol{\theta}_M,\boldsymbol{X}_i)$. Let $[\boldsymbol{\theta}_M|\boldsymbol{Y}_{-i}]$ denote the posterior distribution for $\boldsymbol{\theta}_M$ from a Bayesian analysis using the reduced data set $\boldsymbol{Y}_{-i}$. Then, LOOCV prediction of Y_i is based on:

$$f_M(y|\ \boldsymbol{Y}_{-i},\boldsymbol{X}_i) = E_{[\boldsymbol{\theta}_M|\ \boldsymbol{Y}_{-i}]} f(y|\ \boldsymbol{\theta}_M,\boldsymbol{X}_i), \tag{4}$$

the expected value of $f_M(Y|\boldsymbol{\theta}_M,\boldsymbol{X}_i)$, averaged against the posterior distribution. Note that we are careful to distinguish $f_M(y|\boldsymbol{Y}_{-i},\boldsymbol{X}_i)$, the estimated distribution function (a function of y), from $f_M(Y_i|\boldsymbol{Y}_{-i},\boldsymbol{X}_i)$, its calculated value at the observed value Y_i. The quantity $f_M(Y_i|\boldsymbol{Y}_{-i},\boldsymbol{X}_i)$ is sometimes referred to as the conditional predictive ordinate of Y_i under model M.

For the discrete data of the BBS, $f_M(Y_i|\boldsymbol{Y}_{-i},\boldsymbol{X}_i)$ is the estimated probability of observation Y_i, based on the model M, the covariate vector $\boldsymbol{X}_i$, and all of the other observations. Calculation of this estimate in a Markov chain Monte Carlo (MCMC) analysis of $\boldsymbol{Y}_{-i}$ is straightforward; one need only calculate the expected value of Y_i (say, λ_i) and monitor the Poisson probability of Y_i (i.e. $e^{-\lambda_i}\lambda_i^{Y_i}/Y_i!$) as a derived parameter. The posterior mean of this derived parameter is $f_M(Y_i|\boldsymbol{Y}_{-i},\boldsymbol{X}_i)$.

The Bayesian predictive information criterion is defined as:

$$\mathrm{BPIC}^M = -2\sum_i \log(f_M(Y_i|\boldsymbol{Y}_{-i},\boldsymbol{X}_i)). \tag{5}$$

(Note that the multiplicative factor −2 is not always included; it puts BPIC on the deviance scale, meaning that smaller values are favored.) BPIC provides a convenient summary of LOOCV and an objective basis for model comparison.

A bare ranking of values BPIC^M across models isn't entirely satisfying, as one is left with the question of whether the differences in values are of practical or even statistical significance. Link and Sauer (2016) proposed a statistic, Z_B, as a measure of statistical significance. Noting that:

$$\Delta\mathrm{BPIC} = \mathrm{BPIC}^1 - \mathrm{BPIC}^2 = \sum_i \Delta_i^B, \tag{6}$$

where:

$$\Delta_i^B = \{-2\log(f_1(Y_i|\boldsymbol{Y}_{-i},\boldsymbol{X}_i))\} - \{-2\log(f_2(Y_i|\boldsymbol{Y}_{-i},\boldsymbol{X}_i))\}, \tag{7}$$

they defined Z_B as the sample mean value of Δ_i^B divided by its estimated standard error. Under a null hypothesis that models (1) and (2) are equal as representations of the data-generating model (i.e. they have equal Kullback-Leibler divergences), Z_B is treated as approximately standard normal.

BPIC is computationally intensive, requiring a complete reanalysis of the dataset for each set-aside value (Y_i). For the Mourning Dove (*Zenaida macroura*), the BBS dataset through 2014 had 95,394 counts by 15,381 observers in 154 geographic strata (Pardieck et al. 2015), for which a full analysis using MCMC took ∼3 hr. Repeating this 95,394 times would take more than 30 years.

It is possible to save time on LOOCV by restricting attention to a subset of the Y_i, by strategically reducing the length of the Markov chains used in MCMC analysis, and by using parallel processing on multicore computers. Doing so for 500 randomly selected Mourning Dove observations, we were able to reduce the time needed for calculation of BPIC to 10 days. This is still too long a time to be practical, prompting our interest in the Watanabe-Akaike information criterion (WAIC), which is easily and quickly calculated and is asymptotically equivalent to BPIC. WAIC has been applied to complex model selection studies using BBS data (e.g., Gorzo et al. 2016).

The Watanabe-Akaike Information Criterion

The Watanabe-Akaike information criterion, like the BPIC, is based on an estimate of the predictive distribution of individual observations. Using the same notation as used at equation (4), let:

$$f_M(y|\ \boldsymbol{Y},\boldsymbol{X}_i) = \mathrm{E}_{[\boldsymbol{\theta}_M|\ \boldsymbol{Y}]} f(y|\ \boldsymbol{\theta}_M,\boldsymbol{X}_i), \tag{8}$$

and let:

$$b_M(y|\ \boldsymbol{Y},\boldsymbol{X}_i) = \mathrm{Var}_{[\boldsymbol{\theta}_M|\ \boldsymbol{Y}]} \log\Big(f(y|\ \boldsymbol{\theta}_M,\boldsymbol{X}_i)\Big). \tag{9}$$

Note that equation (8) is the same as equation (4), except that the complete data set $\boldsymbol{Y}$ is used instead of the reduced data set $\boldsymbol{Y}_{-i}$. This means that values $f_M(y|\boldsymbol{\theta}_M,\boldsymbol{X}_i)$ and $b_M(y|\ \boldsymbol{Y},\boldsymbol{X}_i)$ can be computed for all i with a single analysis of the dataset; there's no need to repeat the Mourning Dove analysis 95,394 times, for example. The term $b_M(y|\boldsymbol{Y},\boldsymbol{X}_i)$ can be thought of as a bias correction in estimating log $(f_M(y|\boldsymbol{\theta}_M,\boldsymbol{X}_i))$ based on $\boldsymbol{Y}$ rather than $\boldsymbol{Y}_{-i}$. WAIC is defined as:

$$\mathrm{WAIC}^M = -2\sum_i \Big\{\log\Big(f_M(Y_i|\ \boldsymbol{Y},\boldsymbol{X}_i)\Big) - b_M(Y_i|\ \boldsymbol{Y},\boldsymbol{X}_i)\Big\}. \tag{10}$$

A statistic, Z_W (analogous to Z_B), is based on differences in values $W_i^M = \log\,(f_M(Y_i|\boldsymbol{Y},\boldsymbol{X}_i)) - b_M(Y_i|\boldsymbol{Y},\boldsymbol{X}_i)$ between

TABLE 1. Species used to compare Watanabe-Akaike Information Criterion (WAIC) and Bayesian Predictive Information Criterion (BPIC) for model selection from North American Breeding Bird Survey data for the contiguous United States and southern Canada, 1966–2014. Trend is the estimated geometric mean rate of population change (% per year) from U.S. Geological Survey analyses (model *T*, linearly trending random year effects); 95% CI is the credible interval defined by the 2.5th and 97.5th percentiles of the posterior distribution.

Species name	Scientific name	Routes (*n*)	Strata (*n*)	Trend	95% CI
Wild Turkey	*Meleagris gallopavo*	2,165	122	7.86	6.83, 8.64
Eurasian Collared-Dove	*Streptopelia decaocto*	1,326	84	30.42	25.87, 33.65
Mourning Dove	*Zenaida macroura*	4,296	154	−0.33	−0.45, −0.21
Sora	*Porzana carolina*	982	55	0.79	−1.11, 1.99
Least Bittern	*Ixobrychus exilis*	113	13	−0.03	−2.18, 2.01
Bank Swallow	*Riparia riparia*	1,807	107	−5.46	−7.08, −4.14
Cliff Swallow	*Petrochelidon pyrrhonota*	2,976	138	0.44	−1.22, 1.07
Red-breasted Nuthatch	*Sitta canadensis*	1,682	65	0.65	0.03, 1.19
Carolina Wren	*Thryothorus ludovicianus*	1,739	76	1.31	1.11, 1.51
Eastern Bluebird	*Sialia sialis*	2,509	108	1.72	1.50, 1.93
Western Bluebird	*Sialia mexicana*	443	21	0.60	−0.61, 1.34
Wood Thrush	*Hylocichla mustelina*	2,109	86	−1.94	−2.12, −1.78
Pine Siskin	*Spinus pinus*	1,398	65	−4.61	−6.42, −3.38
McCown's Longspur	*Rhynchophanes mccownii*	113	8	−6.24	−8.91, −3.57
Lark Bunting	*Calamospiza melanocorys*	527	29	−5.42	−8.74, −3.44
Henslow's Sparrow	*Ammodramus henslowii*	346	27	−1.23	−2.49, 0.04
Lincoln's Sparrow	*Melospiza lincolnii*	846	42	−0.78	−2.07, 0.38
Tricolored Blackbird	*Agelaius tricolor*	78	3	−0.75	−5.05, 4.88
Eastern Meadowlark	*Sturnella magna*	2,513	108	−3.31	−3.70, −3.06
Western Meadowlark	*Sturnella neglecta*	2,045	82	−1.40	−1.63, −1.16

models (Link and Sauer 2016). Z_W is the mean difference divided by its estimated standard error.

Data, Models, and Analyses

We compared the fit of 4 models to BBS survey-wide data for 20 species using BPIC, and compared the performance of WAIC with BPIC in making these assessments. The 20 species (Table 1) were selected to represent a cross-section of species in regard to breeding distribution, abundance along BBS routes, patterns of population change, and data quality (Sauer et al 2014); BBS data from 1966 to 2014 were included in the analysis (Pardieck et al. 2015). The intent of this selection was to provide case studies of species for which we believe model selection might provide meaningful results. In particular, models with linearly trending year effects may be better suited for species undergoing consistent population change, while first-difference year effect models might better fit species undergoing population fluctuations. To examine this, we selected the Wild Turkey (*Meleagris gallopavo*), Eurasian Collared-Dove (*Streptopelia decaocto*), Bank Swallow (*Riparia riparia*), McCown's Longspur (*Rhynchophanes mccownii*), and Lark Bunting (*Calamospiza melanocorys*) as species that have experienced consistent and dramatic population changes, and chose the Sora (*Porzana carolina*), Cliff Swallow (*Petrochelidon pyrrhonota*), Red-breasted Nuthatch (*Sitta canadensis*), Carolina Wren, Eastern Bluebird (*Sialia sialis*), Western Bluebird (*Sialia mexicana*), Wood Thrush (*Hylocycla mustilina*) Pine Siskin (*Spinus pinus*), and Henslow's Sparrow (*Ammodramus henslowii*) as representative of species that have experienced population fluctuations. We also examined the Lincoln's Sparrow (*Melospiza lincolnii*), Tricolored Blackbird (*Agelaius tricolor*), Eastern Meadowlark (*Sturnella magna*), and Western Meadowlark (*Sturnella neglecta*), because these species have historically proven difficult to analyze due to large variances and apparent lack of model fit. Finally, we included the Least Bittern (*Ixobrychus exilis*), a wetland species that is rarely encountered on BBS routes, and the Mourning Dove, one of the most frequently encountered birds on BBS routes.

All of the models considered treated counts, conditional on their expected values, as independent Poisson random variables, with explanatory variables additive on the log scale of the expected count. All included an effect, η, for the observer's first year of service on a route. Observer effects were modeled as mean-zero normal random variables, with a precision parameter, denoted τ^{ω}. All of the models included count-specific overdispersion effects, modeled as mean zero normal random variables with precision τ^{ε}. Mean-zero normal random variables associated with temporal change were allowed to vary by stratum and denoted τ_s^{γ}. We performed Bayesian analysis, assigning gamma priors to precision parameters and vague normal priors to all others.

Three of the models that we considered used linearly trending random year effects to model the population trajectory (equation (1)). These models were distinguished

by the amount of hierarchical structure imposed on α_S, β_S, and τ_s^{ω}. In model *T*, these were all assigned independent vague priors. Model *S* was like model *T*, except that α_S and β_S were treated as independent normal random effects, varying by stratum. Model *V* was like model *S*, except that τ_s^{ω} were modeled as lognormal random effects, varying by stratum. Model *T* has been used since 2010 as the omnibus model for BBS analyses (Sauer and Link 2011); models *S* and *V* add hierarchical structure, similarly to the suggestions of Smith et al. (2014). The fourth model, model *D*, was identical to model *T*, except that the population trajectory was described by the first-difference year effect model (equation (2)).

We fitted the models using MCMC as implemented in program JAGS (Plummer 2003) through package R2jags (Su and Yajima 2015) in R 3.3.2 (R Core Team 2016). A preliminary analysis for each species and model was performed using Markov chains of length 10,000. The final states of these analyses were saved as starting values for subsequent LOOCV analyses, obviating the need for burn-in for those analyses.

The number of observations in most BBS datasets is prohibitively large for full LOOCV analyses. It suffices to compare models based on their fit to a large subset of observations. We could have sampled these observations completely at random, had we been interested in comparing the overall fit of the models. However, our interest was in the fit of the trajectory portion of the model. Therefore, for each species, we randomly selected 12 observations per year for each of the 47–49 years of data analyzed. We computed conditional predictive ordinates $f_M(Y_i|\boldsymbol{Y}_{-i},\boldsymbol{X}_i)$ from each of the 4 models for each of these observations, and generated values of BPIC^M and statistic Z_B in summary. Calculation of the conditional predictive ordinates thus typically involved $49 \times 12 \times 4 = 2{,}352$ MCMC analyses. In the interest of time, we reduced the Markov chain length to 10,000 for these analyses, noting that there was little difference between the values computed based on the first 5,000 vs. the second 5,000. For 4 species (Western Bluebird, McCown's Longspur, Lark Bunting, and Tricolored Blackbird), we omitted the first 1–2 years of data from the LOOCV analysis due to low data quality (the BBS was not fully implemented over most of the survey area until 1968; Sauer et al. 2013). Next, we performed a single analysis for each species and model, computing $f_M(Y_i|\boldsymbol{Y}_{-i},\boldsymbol{X}_i)$ and $b_M(Y_i|\boldsymbol{Y}_{-i},\boldsymbol{X}_i)$ for the same indices (i) as used for calculating BPIC^M, thus obtaining comparable observations of WAIC^M and Z_W.

The key advantage of computing both BPIC and WAIC is that both can be decomposed to evaluate the fit of single observations. Examination of single observations permits in-depth comparison of the approaches; it also provides complete flexibility in selecting the components of model fit to consider when comparing models in complex time series of data (Link and Sauer 2016).

RESULTS

Comparisons of Models by BPIC

Models *T* and *D* have nearly identical hierarchical structure, differing only in the form of the trajectory (model *T* having linearly trending random year effects, and model *D* allowing greater flexibility in pattern through random first differences in year effects). For the 20 species considered, there was a nearly even split between these models alone (Table 2): *T* was favored over *D* 11/20 times. In considering the full model set, model *D* was top-ranked in 6/20 cases.

Models *T*, *S*, and *V* share the same trajectory, but, among these, model *T* has the least and model *V* the most hierarchical structure. Restricting attention to these 3 models, *T* and *S* were each preferred in 5/20 cases, and *V* in 10/20 cases.

As noted previously, the raw ranking of BPIC values is not completely satisfactory; one might reasonably ask whether differences in BPIC values are of practical or statistical significance. We addressed the question of statistical significance using the statistic Z_B, judging the fit of a model to be superior to another if the Z_B value was less than −1.96 (Table 3). For 14/20 species, none of the models emerged as superior. In the 6 remaining cases, the Z_B statistic tended to define 2 groups, in which models *D* and *V* were indistinguishable but were preferred over the other 2 models.

The benefit of formal model-selection procedures becomes clear upon inspection of Figure 3. Over the last 30 years, fitted population trajectories for Lincoln's Sparrows using models *D* and *T* are in reasonably close agreement, but there is some disagreement in the indices for the early years, with model *D* suggesting an increase and model *T* a decrease in population. The long-term trend (geometric mean rate of annual change over 49 years) is −0.39% (95% CI: −1.58% to 0.71%) under model *T*, and 0.28% (95% CI: −0.77% to 1.07%) under model *D*. Model *D* assigns a 73% posterior probability to a larger population in the final year than in the first year; model *T* assigns only a 24% probability. While there is considerable overlap in the credible intervals of both the trajectories and the trend estimates, such that we would not assert that the 2 analyses are contradictory, it is surely desirable to be able to distinguish between the models on the basis of an objective evaluation. The results presented in Tables 2 and 3 indicate that model *D* is to be favored.

WAIC as an Approximation to BPIC

WAIC is asymptotically equivalent to BPIC, but can be implemented with a single MCMC analysis, and hence has

TABLE 2. Bayesian Predictive Information Criterion (BPIC) values for 4 models (*T*, *D*, *S*, and *V*) applied to North American Breeding Bird Survey data for 20 species. See Figure 3 for descriptions of models *T* and *D*. Model *S* was like model *T*, except that slopes and intercepts were treated as independent normal random effects, varying by strata. Model *V* was like model *S*, except that stratum-specific variances in observer effects were treated as lognormally distributed random effects. Values are based on 12 randomly selected observations per survey year for each species; the smallest BPIC value (best model fit) is underlined for each species.

	Model			
Species	*T*	*D*	*S*	*V*
Wild Turkey	778.35	779.29	783.82	783.45
Eurasian Collared-Dove	652.28	633.75	645.66	624.88
Mourning Dove	4083.63	4070.43	4084.43	4066.52
Sora	1170.59	1174.83	1168.48	1174.35
Least Bittern	475.82	478.76	475.29	482.06
Bank Swallow	1595.24	1583.23	1595.50	1581.53
Cliff Swallow	2517.63	2519.85	2518.97	2512.21
Red-breasted Nuthatch	1753.56	1763.72	1752.61	1753.45
Carolina Wren	2477.08	2462.56	2469.12	2475.79
Eastern Bluebird	2221.58	2211.56	2223.42	2226.25
Western Bluebird	1541.72	1543.37	1541.63	1540.58
Wood Thrush	2525.33	2529.77	2526.85	2520.78
Pine Siskin	2072.12	2073.48	2069.17	2063.54
McCown's Longspur	1861.07	1919.78	1866.72	1887.13
Lark Bunting	3167.02	3150.60	3164.39	3156.51
Henslow's Sparrow	486.01	482.19	484.79	484.12
Lincoln's Sparrow	1756.80	1736.72	1756.57	1763.76
Tricolored Blackbird	2398.67	2411.19	2399.78	2407.35
Eastern Meadowlark	3336.27	3374.28	3337.75	3383.62
Western Meadowlark	3684.06	3677.26	3684.82	3682.77

a greatly reduced computational burden. We were interested in evaluating the value of WAIC as a surrogate for BPIC in analyses of BBS data. We thus computed WAIC values for each of the 4 models and 20 species (Table 4).

Surprisingly, rankings of models by WAIC were not consistent with those by BPIC (Tables 2 and 4, Figure 4). While the slope of the regression line was positive ($\hat{\beta} = 0.299$, 95% CI: 0.087 to 0.511, $P = 0.007$), the R^2 value was only 0.089.

Similar results were obtained when evaluating statistic Z_W as a surrogate for Z_B (Figure 5). Once again, the slope of the regression line was positive ($\hat{\beta} = 0.200$, 95% CI: 0.054 to 0.346, $P = 0.008$), but the R^2 value was only 0.057.

BPIC and WAIC are both based on estimates of $-2\log(f_M(Y_i|\boldsymbol{\theta}^M,\boldsymbol{X}_i))$. Denote these by $B_i = -2\log(f_M(Y_i|\boldsymbol{Y}_{-i},\boldsymbol{X}_i))$ and $W_i = -2\log(f_M(Y_i|\boldsymbol{Y},\boldsymbol{X}_i)) - b_M(Y_i|\boldsymbol{Y},\boldsymbol{X}_i)$, respectively. In an effort to understand the discrepancy between conclusions based on BPIC $= \sum_i B_i$ and WAIC $= \sum_i W_i$, we plotted observation-specific values of W_i against B_i for each of the 20 species and 4 models considered. Figure 6 displays the results for the Wood Thrush BBS data evaluated with model *T*; this pattern is typical of what was observed for all species and models. From its definition, we see that B_i decreases as a function of the probability of Y_i. For the highest-probability observations (~60% had $B_i < 5$), the agreement between B_i and W_i is good. However, for unlikely observations under the model, W_i is consistently too small relative to B_i, and increasingly so as the probability of the observation decreases. The bias correction term in W_i is too small, especially for unlikely observations, which are the ones that contribute most to the totals.

TABLE 3. Groupings of models (*T*, *D*, *S*, and *V*; see Figure 3 and Table 2 for model descriptions) based on statistic Z_B (see Figure 5 for definition) used to analyze North American Breeding Bird Survey data for 6 species. For a given species, models occurring in the same column had values $|Z_B| < 1.96$, and rows follow the ranking of models by BPIC from best (top) to worst (bottom). For the remaining 14 species in Tables 1 and 2, $|Z_B| < 1.96$ for all comparisons among the 4 models. * Note that for comparing models *T* and D for the Lincoln' Sparrow, $Z_B = 1.95$.

Wild Turkey		Eurasian Collared-Dove		Mourning Dove		Bank Swallow		Carolina Wren		Lincoln's Sparrow	
T		*V*		*V*		*V*		*D*		*D**	
D	*D*	*D*		*D*	*D*	*D*	*D*	*S*	*S*		*S*
V	*V*		*S*		*T*		*T*	*V*	*V*	*T**	*T*
	S		*T*		*S*		*D*		*T*		*V*

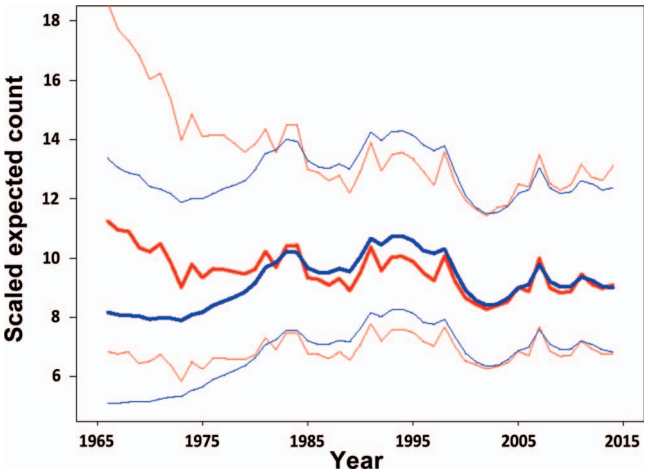

FIGURE 3. Estimated composite population trajectories for the Lincoln's Sparrow from the full North American Breeding Bird Survey dataset, with 95% credible intervals. Point estimates are posterior medians, and the solid line is the log-linear component of the trajectory. The red lines describe the population trajectory generated from model *T*, which uses linearly trending random year effects to model the population trajectory. The blue lines describe results for model *D*, which is identical to model T except that the population trajectory is described by a first-difference year effect model.

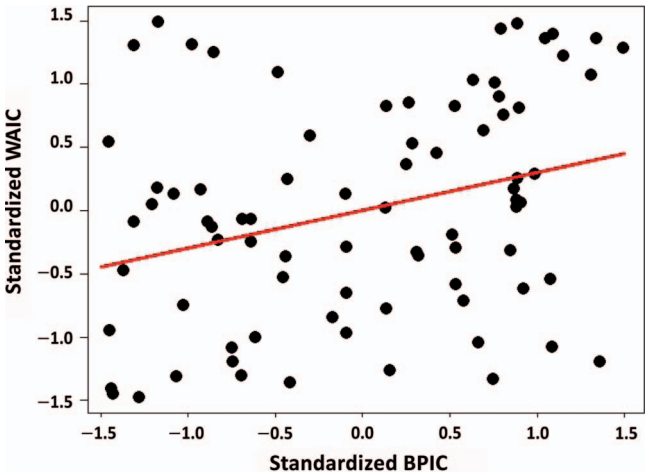

FIGURE 4. Standardized Watanabe-Akaike Information Criterion (WAIC) vs. Bayesian Predictive Information Criterion (BPIC) for 4 models applied to North American Breeding Bird Survey data for 20 species, based on the values from Tables 2 and 4 standardized across models by species. Rankings of models by WAIC were not consistent with those by BPIC.

DISCUSSION

The North American Breeding Bird Survey is unparalleled as a source of spatial and temporal information on population change for most North American bird species. Statistical innovations in its analysis and the care taken in maintaining careful protocols for data collection and curation enhance its value to science and management. However, the complex structure of its data dictates a need for caution in model development and assessment. The development of MCMC and its easy implementation in

TABLE 4. Watanabe-Akaike Information Criterion (WAIC) values for 4 models (*T*, *D*, *S*, and *V*; see Figure 3 and Table 2 for model descriptions) applied to North American Breeding Bird Survey data for 20 species. Values are based on the same 12 randomly selected observations per survey year for each species as used in analysis for Table 2; the smallest WAIC value (best model fit) is underlined for each species.

	Model			
Species	*T*	*D*	*S*	*V*
Wild Turkey	635.77	641.28	636.84	653.39
Eurasian Collared-Dove	521.77	523.91	521.32	546.11
Mourning Dove	3606.91	3602.11	3607.17	3597.02
Sora	1076.43	1074.87	1075.05	1072.77
Least Bittern	441.29	449.19	446.85	452.58
Bank Swallow	1124.36	1123.16	1124.98	1126.89
Cliff Swallow	1768.66	1773.07	1772.87	1768.21
Red-breasted Nuthatch	1641.39	1643.86	1638.44	1639.55
Carolina Wren	2323.26	2311.16	2325.57	2325.89
Eastern Bluebird	2095.63	2077.63	2092.71	2090.47
Western Bluebird	1372.82	1374.72	1372.64	1373.56
Wood Thrush	2407.26	2414.46	2409.21	2409.95
Pine Siskin	1567.59	1569.91	1568.04	1567.84
McCown's Longspur	1446.56	1443.78	1445.06	1445.16
Lark Bunting	2193.98	2192.47	2198.63	2183.67
Henslow's Sparrow	454.93	457.41	455.68	456.19
Lincoln's Sparrow	1609.29	1594.74	1611.72	1605.55
Tricolored Blackbird	1303.44	1304.03	1302.68	1303.24
Eastern Meadowlark	3060.63	3053.53	3059.55	3071.43
Western Meadowlark	3151.83	3152.93	3153.54	3150.56

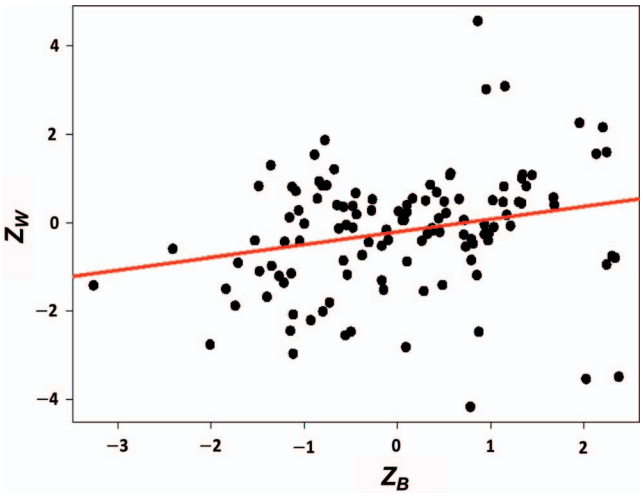

FIGURE 5. Scatterplot of Z_W against Z_B using 6 comparisons (among 4 models) for 20 species surveyed by the North American Breeding Bird Survey. The statistics Z_W and Z_B test a null hypothesis of equal Kullback-Leibler divergences, and are based on the Watanabe-Akaike Information Criterion and the Bayesian Predictive Information Criterion, respectively.

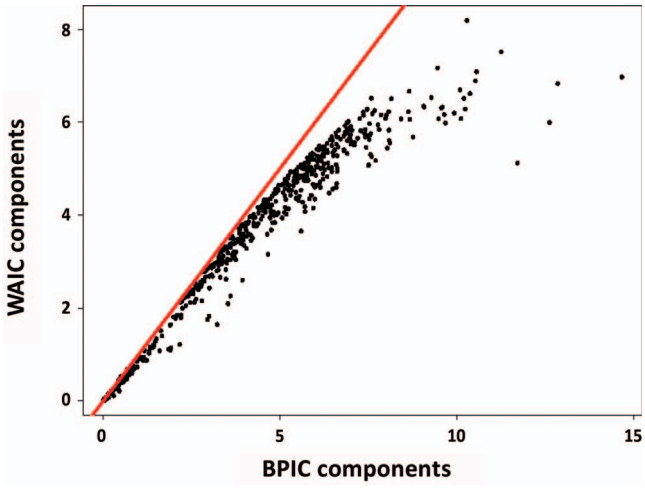

FIGURE 6. Observation-specific components W_i of the Watanabe-Akaike Information Criterion (WAIC) plotted against observation-specific components B_i of the Bayesian Predictive Information Criterion (BPIC) for BBS Wood Thrush data fitted under model *T* (linearly trending random year effects). The red line is the identity function (i.e. points falling on the line satisfy $W_i = B_i$).

programs such as JAGS have made it very easy to suggest new hierarchical models and to tweak existing models. Because our ability to produce and fit models has outstripped our ability (or desire) to formally critique and compare them, we need to consider how best to provide credible and defensible estimates of population change while still being open to modeling innovations. Development and comparison of model sets that provide reasonable alternatives for model structure are important steps in implementing a diversity of models while still permitting comparisons among them. The results presented here are a first step in exploring alternative approaches in model selection within such a model set.

Our model set was chosen to represent the current candidate models that a BBS analyst would be likely to employ, with the *T*, *S*, and *V* models sharing the same linearly trending random year effects trajectory but differing in the hierarchical nature of other model components. This comparison is timely; although model *T* is the one used in the analysis on the USGS Results and Analysis Website (Sauer et al. 2014), Sauer et al. (2017) recommend the use of model *S* because it facilitates expansion of the BBS analysis to additional regions where very limited data exist from the early years of the survey, and Smith et al. (2014) recommend use of model *V* for the BBS in Canada. Link and Sauer (2016) note that the first-difference model (model *D*) might provide more realistic results in cases where models using the slope–year effect parameterization might lead to overprediction of population change. The 20 species that we selected for our comparative analysis were chosen to provide a variety of life-history, distribution, and population change situations to compare the models.

Analyses using BPIC indicated that, for the majority of the 20 species that we thought would be likely to be sensitive to choice of model, no model was clearly superior. Although the ranking process inevitably produces a model with the smallest BPIC, the quantitative comparison using the Z_B statistic found a significant difference in BPIC value for only 25% of the species. Of these 6 species, Wild Turkeys and Eurasian Collared-Doves dramatically increased in population size and Bank Swallows steeply declined over the 1966–2014 interval. The population trajectory for Carolina Wrens was notably jagged, with large declines related to severe winters (Figure 2). Mourning Doves are one of the most commonly encountered species on BBS routes, hence the ability to discriminate among models for this species was not surprising. The final species, the Lincoln's Sparrow, showed dramatic differences in population trajectory in the early years depending on the model chosen (Figure 3). Although our limited selection of species does not allow us to make generalizations about appropriate models for the analysis of data for the 420 species presently monitored by the BBS, we suggest that these results have implications both for the present analysis of BBS data and for future directions in evaluations of appropriate models for BBS data analysis. First, our results suggest that, within this model set, the current BBS analysis model (model *T*) generally performs quite well across the large variety of sampling situations associated with the 20 analyzed species. Second, in the cases where model *T* was not preferred, model *D* or model

V was preferred. Our inclination is to suggest that model *D*, which presents an alternative formulation of change over time, is likely to be a superior model to model *V*, which retains the slope–year effect structure of models *T* and *S* but adds additional hierarchical structure. In the case of the Lincoln's Sparrow, it is likely that both models can accommodate limited data in certain strata in different ways, with model *D* providing more realistic estimates of change during periods with limited data.

The tendency for no clear best model to be determined for analysis of our selection of species is reflected in summary results for the species. For the 20 species, credible intervals for long-term trends at the survey-wide scale showed general overlap among the 4 models, except those for Wild Turkeys and Eurasian Collared-Doves, for which trends tended to be lower for the *D* model and higher for the *V* model than for the *T* and *S* models.

Comparison of BPIC results with WAIC results indicated that WAIC is not a reasonable surrogate for BPIC in model selection, and we caution against its use. Similar findings have been reported by Vehtari et al. (2017). This is a disappointing result, as WAIC is much easier and less time consuming to produce, and appears to be similar to BPIC conceptually. We investigated the use of WAIC as a screening surrogate for BPIC, using it to identify species for which models might differ, which could then be subjected to confirmatory BPIC analysis. However, that approach does not seem feasible without a better understanding of when WAIC fails. Because observation-specific values of both BPIC and WAIC can be computed, it may be possible to empirically model the relationship between BPIC and WAIC and "correct" WAIC. Although such an approach may not seem worth the effort, the current difference in cost between the 2 approaches makes further investigation of a means of improving WAIC a tempting idea. The development of computationally efficient approximations to BPIC remains an area of active research (see Vehtari et al. 2017).

We believe that cross-validation remains the best tool for objective examination of models. Much additional work is needed to establish how best to implement a BPIC-based comparison of models. Because it is not feasible, or perhaps even desirable, to compute BPIC for all samples, the question of what component of model fit is of interest must be explicitly addressed and a design established to select samples to meet this goal. Because the BBS has been constantly expanding in area and consistency of coverage, a primary concern in summary analysis is to not allow long-term results to be determined primarily by extrapolation from intervals with the most data, that is through estimation of a slope in the slope–year effects models. To address that concern, we balanced the random sample of observations for computing BPIC by year to limit the influence of the greater numbers of observations in recent years. However, the choice of observations to sample for BPIC can also address other goals. For example, a primary emphasis for modeling might be a subinterval of the larger time series, such as a time period when a species underwent an unusual fluctuation or when an important environmental change occurred, and thus the primary interest is in appropriately modeling the effect of the intervention on the population. More discussion on this issue can be found in Link and Sauer (2016).

Finally, we view the model set for BBS analyses as not being well established. The model set in this paper is a logical one, containing the commonly used BBS models. It is realistic, in that there is real interest in determining which of these models is in some sense best. Although we observed general similarity in the results from these models, controversy does still exist with regard to what model is best; even small differences in trend can have management implications, and practitioners tend to focus on point estimates that can appear quite different even though credible intervals may overlap. The results presented here should reassure practitioners that there are consistencies in current BBS modeling activities, but should also provide a cautionary note that alternative models will always be in development, and that future BBS analyses will almost certainly be based on model selection in the context of a set of candidate models.

ACKNOWLEDGMENTS

The BBS is the product of thousands of volunteers and a hierarchy of organizers, coordinators, and data managers. We thank all of them. We also thank Keith Pardieck and Adam Smith for thoughtful reviews of the manuscript. Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Funding statement: This work was partially funded by the U.S. Geological Survey (USGS) Status and Trends Program. Keith Pardieck was the project manager. This paper has been peer reviewed and approved for publication consistent with USGS Fundamental Science Practices (http://pubs.usgs.gov/circ/1367).

Ethics statement: This research was conducted in compliance with all applicable USGS study design and review protocols.

Author contributions: W.A.L. formulated the questions, developed analysis approaches, conducted analyses, and wrote the paper. J.R.S. collaborated in the development of analysis approaches, conducted analyses, and cowrote the paper. D.K.N. conducted analyses and cowrote the paper.

Data deposits: BBS data are deposited at ftp://ftpext.usgs.gov/pub/er/md/laurel/BBS/Archivefiles/Version2015v0/

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.

Burnham, K. P., and D. R. Anderson (2002). Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, second edition. Springer, New York, NY, USA.

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

Gorzo, J. M., A. M. Pidgeon, W. E. Thogmartin, A. J. Allstadt, V. C. Radeloff, P. J. Heglund, and S. J. Varvus (2016). Using the North American Breeding Bird Survey to assess broad-scale response of the continent's most imperiled avian community, grassland birds, to weather variability. The Condor: Ornithological Applications 118:502–512.

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

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

Millsap, B. A., G. S. Zimmerman, J. R. Sauer, R. M. Nielson, M. Otto, E. Bjerre, and R. Murphy (2013). Golden Eagle population trends in the western United States: 1968–2010. The Journal of Wildlife Management 77:1436–1448.

Pardieck, K. L., D. J. Ziolkowski, Jr., and M.-A. R. Hudson (2015). North American Breeding Bird Survey Dataset 1966–2014, version 2014.0. U.S. Geological Survey, Patuxent Wildlife Research Center, Laurel, MD, USA. www.pwrc.usgs.gov/BBS/RawData/

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

R Core Team (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Runge, M. C., and J. R. Sauer (In press). Allowable take of Red-winged Blackbirds in the northern Great Plains. In Ecology and Management of Blackbirds (Icteridae) in the United States (G. M. Linz, M. L. Avery, and R. A. Dolbeer, Editors). Taylor and Francis, Milton Park, Abingdon, UK.

Runge, M. C., J. R. Sauer, M. L. Avery, B. F. Blackwell, and M. D. Koneff (2009). Assessing allowable take of migratory birds. The Journal of Wildlife Management 73:556–565.

Sauer, J. R., and W. A. Link (2004). Some consequences of using counts of birds banded as indices to populations. In Monitoring Bird Populations with Mist Nets (C. J. Ralph and E. H. Dunn, Editors). Studies in Avian Biology 29:168–172.

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

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., 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). The first 50 years of the North American Breeding Bird Survey. The Condor: Ornithological Applications 119:576–593.

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.

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

Vehtari, A., A. Gelman, and J. Gabry (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing 27:1413–1432.

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.

Zimmerman, G. S., J. R. Sauer, K. Fleming, W. A. Link., and P. R. Garrettson (2015). Combining waterfowl and breeding bird survey data to estimate Wood Duck breeding population size in the Atlantic Flyway. The Journal of Wildlife Management 79:1051–1061.