Model selection for the North American Breeding Bird Survey, with observations on BPIC and WAIC model selection criteria

William A. Link[1,2] , John R. Sauer[1], and Daniel K. Niven[1]

[1] *USGS Patuxent Wildlife Research Center, Laurel, Maryland 20708, USA*

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

Abstract. The North American Breeding Bird Survey (BBS) provides data that can be used in complex, multi-scale 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 readily applied to 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 4 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 overdisperson effects are independent and t-distributed (H models) tended to be favored over models where overdisperson 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; hierarchical models; leave-one-out cross validation; model selection; North American Breeding Bird Survey; WAIC

Introduction

The North American Breeding Bird Survey (BBS) provides count data used to monitor population change for over 500 species of birds. 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 5745 routes across the contiguous United States, Alaska, and Canada, of which $\sim$ 3000 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 1300 published citations of the USGS website in the last 20 years.

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 4 counts over 5 years of service on a BBS route. Observers'

participation duration on a route have 25th, 50th, 75th, and 90th percentiles of 1, 5, 12 and 23 years (Link and Sauer 2016a). 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 regression models, analyzed as Bayesian hierarchical models (e.g., Sauer et al. 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).

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 years from 1966-2015. Among these, sample sizes range from 113 to 99695, 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 can 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 4 models for 548 species, using data for the period from 1966-2015.

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 2016b).

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 our primary results on model selection for 548 North American bird species, and comment on the usefulness of WAIC as a surrogate for BPIC. 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

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

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 heroic 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 i^{th} 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 (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 et al. 2014, Sauer and Link 2011). Labels S (for "slope"), D (for "difference") and H (for "heavy tails") describe 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

$$\mathrm{E}(\gamma_{s,y}) = S_s + \beta_s\,(y - y_0) \quad ;$$

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 absence of data: actual year effects vary, and the model is capable of detecting distinctly nonlinear population trajectories, or patterns of population change.

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 $\mathrm{E}(\gamma_{s,y_0}) = S_s$, at the stratum mean.

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 2016b). Differences between fits relate to Bayesian shrinkage. Thinking of a graph of year effects against years, S models shrink year effects vertically (toward a long term trend curve) while D models shrink horizontally (towards 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, and scale parameters are assigned flat gamma priors. 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 have been based on expected counts in the region. These were of the form $n_{s,y} = \exp(\gamma_{s,y} + \frac{1}{2}(\sigma_s^\gamma)^2 + \frac{1}{2}(\sigma_s^\epsilon)^2)$, where $\sigma_s^\gamma = 1/\sqrt{\tau^\gamma}$ and $\sigma_s^\epsilon = 1/\sqrt{\tau^\epsilon}$ are the standard deviations of the random effects distributions of years effects and observer effects, respectively. These expected counts were derived under the assumption that that ϵ_i's and ω_o's follow normal distributions (Sauer and Link 2011); for the H models, these weights must be modified. We replace σ_s^ϵ with 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 years).

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 $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_N)$ and $\boldsymbol{X} = (X_1, X_2, \ldots, X_N)$, respectively. We assume a model M for the data, in which the observations Y_i are conditionally independent, given $\boldsymbol{X}$, with probabilities $\Pr(Y_i|X_i, M, \theta^M)$.

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

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

The distribution function $\text{ppd}(y|X, \boldsymbol{D}, M)$ 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 asssessing 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 $\boldsymbol{D}$ should produce

ppd$(y|X, \boldsymbol{D}, M)$ 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, it turns out that we seek the model $M \in \mathcal{M}$ that maximizes

$$\psi^M = \mathrm{E}_f(\log(\mathrm{ppd}(Y|X, \boldsymbol{D}, M))) \quad . \tag{2}$$

How are we to estimate this quantity? Given a hypothetical new observation $\tilde{Y}$ sampled from $f(y|\tilde{X})$,

$$\log(\mathrm{ppd}(\tilde{Y}|\tilde{X}, \boldsymbol{D}, M))$$

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

$$\log(\mathrm{ppd}(Y_i|X_i, \boldsymbol{D}, M)) \quad , \tag{3}$$

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

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

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

logarithm of the average.

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

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

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

Calculating BPIC is computationally intensive. For each of the N observations, calculating (4) involves calculation of $\pi(\theta^M|\boldsymbol{D}_{\text{-}i}, M)$, the posterior distribution for θ^M based on the data set $\boldsymbol{D}_{\text{-}i}$. While importance sampling can be used to obtain samples of $\pi(\theta^M|\boldsymbol{D}_{\text{-}i}, M)$ based on $\pi(\theta^M|\boldsymbol{D}, M)$, 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 7389, collected on 24 strata. On a workstation equipped with Xeon E5-2630 v3 processors, MCMC with chain length 10000 takes approximately 12 minutes running on a single core. Running 7389 LOO analyses, even taking advantage of multicore processing on the system's 16 cores, would take over 3 days. And that's for a single model. We estimate that the corresponding analysis for the widespread Mourning Dove (*Zenaida macroura*) with 98372 observations on 158 strata would take something over 3 months.

Computational expense can be reduced by computing BPIC on a subset of $n \ll N$ observations (Link and Sauer, 2016b). We chose to conduct approximately $n = 100$ analyses per species, rather than the average $N = 18778$ analyses that would have been required for full computation

of BPIC.

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

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

where Var_{post} denotes the posterior variance of θ^M based on the full data set $\boldsymbol{D}$. 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 $\boldsymbol{D}$. 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

$$\bar{B}^M = \frac{1}{n} \sum_{j=1}^{n} \log(\text{ppd}(Y_{i_j}|X_{i_j}, \boldsymbol{D}_{\text{-}i_j}, M)) \quad , \tag{7}$$

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

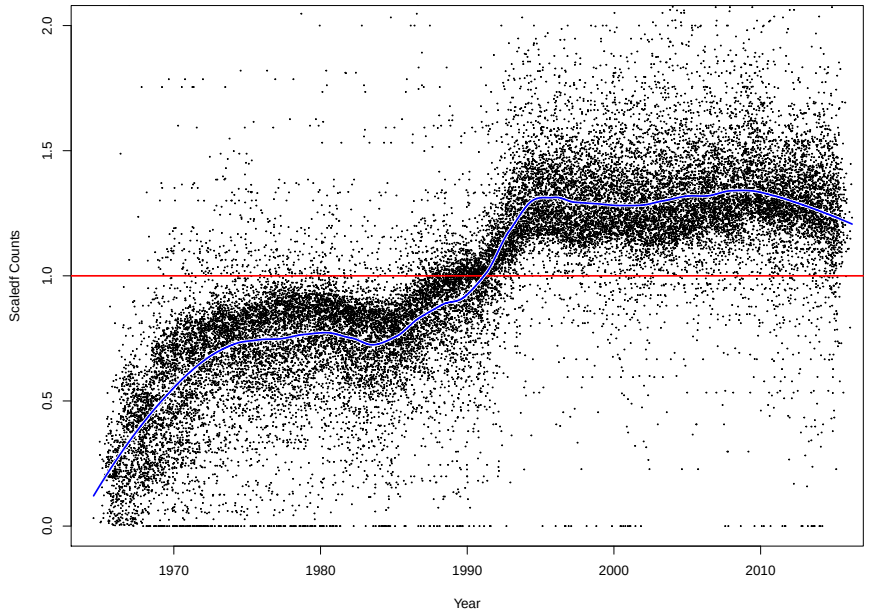


Figure 1: Scaled number of counts for the North American Breeding Bird Survey, each point corresponding to a species and a year. Each number has been scaled by the mean number of counts for the species over years subsequent to the species' first appearance in the data set. Blue curve is LOESS smooth.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

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 (Figure 1). We chose to sample 2 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 10000, 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 a multicore processor.

Having selected n indices $i_1, i_2, \ldots, i_n$, the j^{th} 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 10000 Markov chain samples. The log of the posterior mean for this node is the i_j^{th} summand in the definition of BPIC, equation (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) 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 2500 were satisfactory for stable and precise estimates of the summands of WAIC (equation (6)); we denote the i^{th} value by W_i.

Analysis of selected model

BBS trends are typically presented separately for "core" and "non-core" species (Sauer et al. 2017). For 426 core species we present trends for the interval 1970-2015. Because BBS data covered limited geographic areas in the early years of the survey, we use 1970 as the first year for trend estimation. The 122 non-core species have limited BBS data prior to 1993; for these species we present trend results for the interval 1993-2015.

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 (i.e., Figure 2). For each model, we present the median estimated trend, 2.5%CI, 97.5%CI, 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 sample size, the number of survey routes used in the analysis for the species. For these trend estimates, we summarize differences in results between model S to 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 are 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 default model and the 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 annual indices for default models and the selected models. We conducted summary analyses separately for core and non-core species.

Results: Model selection using BPIC

Basic 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 equation (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.

	D	**DH**	**S**	**SH**
Ranked Best	141	203	81	123
	25.7%	37%	14.8%	22.4%
Ranked Worst	138	70	213	127
	25.2%	12.8%	38.9%	23.2%

Table 1: Rank frequencies (based on sampled BPIC values) across species

Associated with each vector $\hat{\boldsymbol{F}}_s = (\bar{B}_s^D, \bar{B}_s^{DH}, \bar{B}_s^S, \bar{B}_s^{SH})'$ is an estimated covariance matrix $\hat{\boldsymbol{\Sigma}}_s$ (the sample covariance matrix for the n sampled values) from which one can calculate standard errors of differences, such as $(\bar{B}_s^D - \bar{B}_s^{DH})$. Using these, we calculate z-statistics to test null hypotheses of equal support for models (see Discussion in Link and Sauer 2016b, pg 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 6 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

	< -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

Table 2: Paired z-test statistics for BPIC differences.

SH over S. The other 4 comparisons have outcome frequencies shifted to the right, favoring D over S and SH, and DH over SH.

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 3 species for which DH is favored over all, and only 1 where D is favored over all. S and SH are never favored over all.

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 3 cases, and D over DH in 1. On the other hand, there are only 3 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 3 of these, DH is favored over SH; in none is SH favored over DH. On the other hand, there are only 3 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

Given enough data, the BPIC criterion will distinguish between the fits of alternative models. Thus the relatively small number of species for which clear preferences exist is due to the limited subsampling of counts. The tendency for first difference models D and DH to be favored over models S and SH, and for heavy-tailed models DH and SH to be favored over D and S is appropriately conducted using a hierarchical model, across species. Hierarchical modeling also enhances model selection for individual species by considering them collectively, "borrowing strength from the ensemble" (Morris 1983, Louis 1984, Rubin 1984).

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

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

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

The latent vector $\boldsymbol{\Delta}_s$ consists of differences $\psi_s^D - \psi_s^{DH}, \psi_s^S - \psi_s^{DH}$ and $\psi_s^{SH} - \psi_s^{DH}$, with ψ^M defined at equation (2) and subscript s added for species. Recalling that larger values

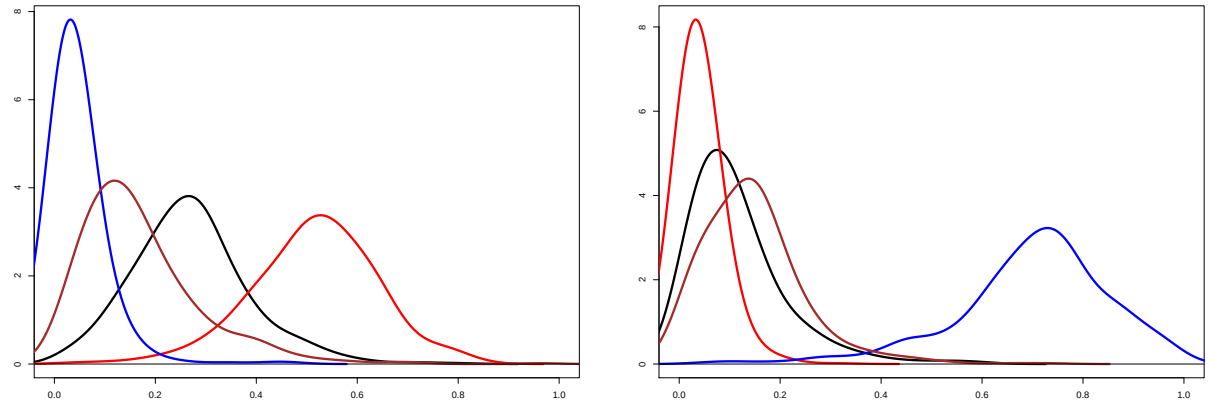


Figure 2: Smoothed density of posterior probability model is best (left panel, π_s^M) and worst (right panel, ρ_s^M) across 548 BBS species, for models M = D (black), DH (red), S (blue) and SH (brown).

of ψ^M indicate better expected predictive value for the combination of data and model, $\Delta_{s,1} > 0$ means that model D is favored over model DH, $\Delta_{s,2} > 0$ means that model S is favored over model DH, and $\Delta_{s,3} > 0$ means that model SH is favored over model D. The four models can be ranked based on $\boldsymbol{\Delta}_s$; for example, $\boldsymbol{\Delta}_s = (-1.2, 3.8, -2.5)$ implies the ordering S > DH > D > SH. Thus we obtained values $\pi_s^M = \Pr(\text{M is best for } s|\hat{\boldsymbol{\Delta}}.)$ and $\rho_s^M = \Pr(\text{M is worst for } s|\hat{\boldsymbol{\Delta}}.)$. Results are summarized in Table 3 and Figure 2; results by species are given in Appendix S4.

Results: WAIC as a surrogate for BPIC

	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%

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

In an earlier paper investigating a set of 4 models in application to 20 BBS species (Link et al. 2017) we noted that the components of WAIC (W_i) were poor surrogates for the components of BPIC (B_i). We confirmed that finding in the present study. Figure 3 plots values W_i against corresponding values B_i for model D, using data for all 548 species. The pattern is evident at the species level, and across models. Gray lines indicate the 1^{st}, 5^{th}, 25^{th}, and 50^{th} percentiles of BPIC values. Note that large values of B_i (suggesting good fit of the model) tend to be matched well by large values of W_i. However, small values of B_i (e.g. below the 5^{th} percentile) have corresponding values of W_i that are too large. The corresponding observations are those corresponding to poorest model fit, and which make the largest contributions to evaluating the model. Thus WAIC appears to overstate the predictive value of the model and data.

We considered using the n calculated values of B_i with the corresponding values of W_i in order to establish a regression-based prediction of B_i for the $N - n$ values not directly calculated, based on the easily obtained corresponding values of W_i. Unfortunately, the resulting prediction variances were large enough to render the predictions unreliable. Additionally, we could not establish a consistent regression relation between W_i's and B_i's across species.

Nevertheless, the relative ease with which WAIC can be calculated, and the potential advantages of using all N observations rather than the subset of n used in our BPIC analyses, invite further comparison. For comparability with our BPIC sampling of equal numbers of observations per year, we computed a weighted version of WAIC. This weighted WAIC is of the same form as (6), but includes weights $w_i \propto 1/n_{y(i)}$, where n_y is the number of counts in year y, and $y(i)$ is the year in which the i^{th} count was conducted.

We compare model rankings for the weighted WAIC to the rankings by posterior probability

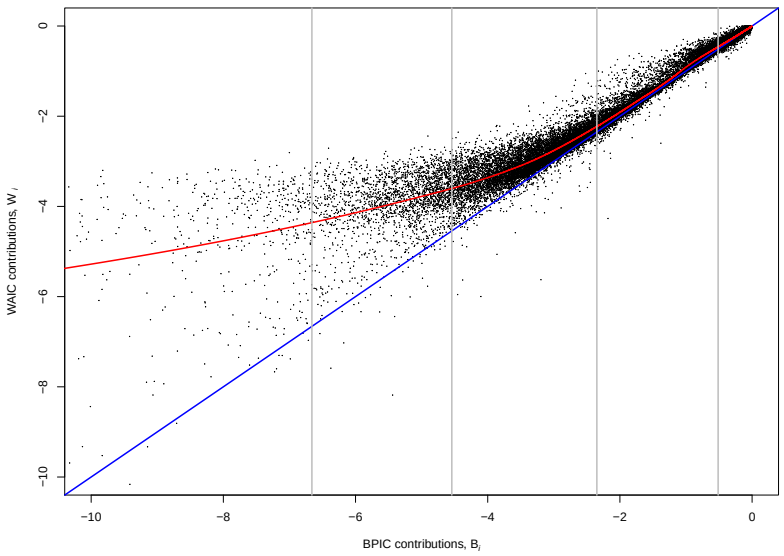


Figure 3: Components of WAIC for Model D (W_i^D) plotted against corresponding components of BPIC (B_i^D), 548 species. Blue line is identity; red is LOESS curve. Gray lines indicate 1^{st}, 5^{th}, 25^{th}, and 50^{th} percentiles of B_i^D values.

in the hierarchical analysis of sampled BPIC values. Across the 548 species, the weighted WAIC assigns rank 1,2,3,4 to the BPIC best model at rates of 46.2%, 24.8%, 15.0%, and 14.0%, respectively.

Figure 4 plots the results of a logistic regression of the success rate of the weighted WAIC in identifying the BPIC model, conducted to examine how the success rate increases with evidence in favor of the best model. That is, we let S_i be the indicator for species $i = 1, 2, \ldots, 548$, and modeled

$$\text{logit}(\Pr(S_i = 1)) = \alpha + \beta \text{ logit}(\pi_i) \quad ,$$

where π_i is the posterior probability of the best supported model for species i (note that with 4 models in the model set, this must be at least 0.25).

Finally, we considered the relation between complete set of rankings from the weighted WAIC

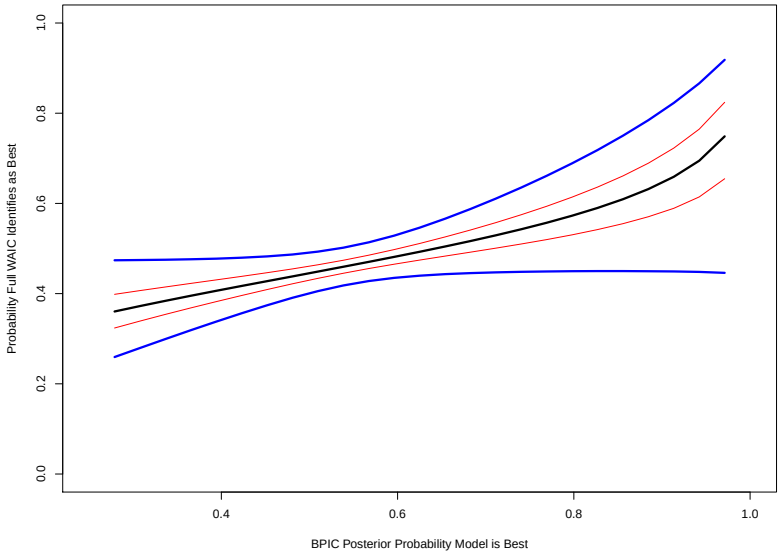


Figure 4: WAIC success rate in identifying BPIC best model, as function of posterior probability of BPIC best model. Red curves are 25^{th} and 75^{th} percentiles of estimation uncertainty in logistic regression; blue curves are corresponding 2.5^{th} and 97.5^{th} percentiles.

and the BPIC. Given BPIC ranks {1,2,3,4}, the corresponding WAIC rankings were {1,2,3,4}, {2,1,3,4}, {1,3,2,4}, {1,2,4,3}, or {2,1,4,3} for 50.2% of the species. Simply guessing, the chance of doing that well is 20.8%.

Taken together, these results show that the WAIC criterion is a weak surrogate for BPIC.

Results: Selected model versus traditional model

Results for the 548 species are presented in Appendix S4. We note that 4 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 4423 routes (median=341), and in estimated abundance from 0.003 to 20073 (median=48.9). Because D models were selected for the majority of species, we expected to see systematic differences in results between default model and the 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 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 trend estimates for which the 95% CIs did not overlap between default model and the selected model. 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. Choice of appropriate models is complicated by the size of the dataset, 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 2 alternative parameterizations of year effects and 2 overdispersion distributions, and apply a hierarchical model to the model selection results to compute posterior probabilities models are best for species. Results suggest that models that use the D year effects parameterization and t-distribution based overdispersion are favored for most species.

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 that it is the worst model for 524 species and the highest posterior probability that it is the best model for only 4 species. Model DH seems a much better candidate as a default model, with a highest posterior probability that it is 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 (equation 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 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). 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 2 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. Particularly in periods with limited data, predictions of annual indices based on the slope parameter are not particularly informative as the 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 dataset. In our experience, BBS data tend of 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. However, the H models introduce a complication in that the mathematical expectation of an exponentiated t-value is infinite; we must use an alternative characterization of a typical count. In Appendix S2, we provide both a clear quantitative rationale for estimation of expected counts in the case of normal overdispersion and suggest an analogue for use in the case of t-distributed overdispersion. To our knowledge this is the first published rationale for these abundance indices.

Acknowledgments

We thank the thousands of volunteers who have contributed to the North American Breeding Bird Survey. 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(1):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. The 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., J.R. Sauer, and D.K. Niven. 2017. Model selection for the North American Breeding Bird Survey: A comparison of methods. Condor 119(3):546–556.

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

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

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

559 Link, W.A., and J.R. Sauer. 2016b. Bayesian Cross-Validation for Model Evaluation and
560 Selection, with Application to the North American Breeding Survey. Ecology 97:1746–
561 1758.

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

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

566 North American Bird Conservation Initiative. 2016. The State of North America's
567 Birds 2016. Environment and Climate Change Canada: Ottawa, Ontario. 8 pages.
568 www.stateofthebirds.org.

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

573 Rubin, D.B. 1984. Bayesianly justifiable and relevant frequency calculations for the applied

574 statistician. Annals of Statistics 12:1151–1172.

575 Rosenberg, K.V., J.A. Kennedy, R. Dettmers, R.P. Ford, D. Reynolds, J.D. Alexander, C.J.
576 Beardmore, P.J. Blancher, R.E. Bogart, G.S. Butcher, A.F. Camfield, A. Couturier, D.
577 W. Demarest, W.E. Easton, J.J. Giocomo, R.H. Keller, A.E. Mini, A.O. Panjabi, D.N.
578 Pashley, T.D. Rich, J.M. Ruth, H. Stabins, J. Stanton, T. Will. 2016. Partners in Flight
579 Landbird Conservation Plan: 2016 Revision for Canada and Continental United States.
580 Partners in Flight Science Committee. 119 pp.

581 Sauer, J.R., and W.A. Link. 2011. Analysis of The North American Breeding Bird Survey
582 Using Hierarchical Models. The Auk 128:87–98. +Supplemental Online Material.

583 Sauer J.R., W.A. Link, J.E. Fallon, K.L. Pardieck, and D.J. Ziolkowski Jr. 2013. The North
584 American Breeding Bird Survey 1966–2011: Summary Analysis and Species Accounts.
585 North American Fauna 79:1–32. + 900pp Supplemental material.

586 Sauer, J.R., J.E. Hines, J.E. Fallon, K.L. Pardieck, D.J. Ziolkowski Jr., and W.A. Link. 2014.
587 The North American Breeding Bird Survey, Results and Analysis 1966 – 2013. Version
588 01.30.2015 USGS Patuxent Wildlife Research Center, Laurel, MD.

589 Sauer, J.R., D.K. Niven, K.L. Pardieck, D.J. Ziolkowski Jr, and W.A. Link. 2017. Expanding
590 the North American Breeding Bird Survey analysis to include additional species and
591 regions. Journal of Fish and Wildlife Management 8(1):154–172.

592 Sauer, J.R., W.A. Link, D.J. Ziolkowski, K.L. Pardieck, and D.J. Twedt. 2019. Consistency
593 counts: Modeling the effects of a change in protocol on Breeding Bird Survey counts. The
594 Condor 121: duz009.

595 Su, Y.-S., and Yajima M. 2015. R2jags: Using R to run 'JAGS.' Version 0.5-7.

https://CRAN.R-project.org/package/R2jags

Supporting Information: William A. Link, John R. Sauer, Daniel K. Niven. Model Selection for the North American Breeding Bird Survey with observations on the WAIC and BPIC model selection criteria. Ecology.

Appendix S1: JAGS code for models

The four models we consider are labeled D, DH, S, and SH. Labels S (for "slope"), D (for "difference") and H (for "heavy tails") describe features of the models. A JAGS model statement for S is given in Figure 5. We use this to describe common features of the models, then in Figure 6 present changes to the year effects needed for D models. Changes needed for the H models are described in the text of this Appendix.

The `Data model` portion of the JAGS code is the same for all 4 models. Count i is modeled as a Poisson random variable with mean λ_i. We suppose that $\log(\lambda_i)$ is a random variable with mean

$$\mathrm{E}(\log(\lambda_i)) = \gamma_{s(i),y(i)} + \omega_{o(i)} + f(i)\,\eta \quad ;$$

here $\gamma_{s,y}$ is the expected count in stratum s at year y, ω_o is the effect of observer o, and η is an offset for the effect of an observer's first year of service (indicated by a zero-one variable $f(i)$). Note the use of indirect indexing: $s(i), y(i)$, and $o(i)$ index the stratum, year, and observer corresponding to the i^{th} observation. Observer effects are treated as a sample of a mean zero normal distribution with precision τ_ω.

In models S and D, $\log(\lambda_i)$ is assumed to be normally distributed with precision (1/variance) τ_ϵ, reflecting extra-Poisson variation. Heavy-tailed alternative models SH and DH assume a t-distribution in place of a normal distribution in modeling extra-Poisson variation. This entails two changes to the code in Figure 5. First, we change the definition of `loglambda[i]`

```
# JAGS model statement for BBS data #
model=function(){

  # Data model #
  for(i in 1:ncounts){
      E.loglambda[i] <- gamma[year[i],strat[i]]+omega[obser[i]]+eta*firstyr[i]
      loglambda[i] ~ dnorm(E.loglambda[i],tau.epsilon)
      log(lambda[i]) <- loglambda[i]
      count[i] ~ dpois(lambda[i])
  }
  tau.epsilon ~ dgamma(0.001,0.001)

  # Observer effects #
  for(o in 1:nobservers){
      omega[o] ~ dnorm(0.0,tau.omega)
  }
  tau.omega ~ dgamma(0.001,0.001)
  eta ~ dnorm(0.0,1.0E-6)

  # Year effects #
  for(s in 1:nstrata){
      for(y in 1:nyears){
          E.gamma[s,y] <- stratum[s]+beta[s]*(y-y.baseline)
          gamma[s,y] ~ dnorm(E.gamma[s,y],tau.gamma[s])
      }
      beta[s] ~ dnorm(mu.beta,tau.beta)
      stratum[s] ~ dnorm(mu.strata,tau.strata)
      tau.gamma[s] ~ dgamma(0.001,0.001)
  }
  mu.beta ~ dnorm(0.0,1.0E-6)
  tau.beta ~ dgamma(0.001,0.001)
  mu.strata ~ dnorm(0.0,1.0E-6)
  tau.strata ~ dgamma(0.001,0.001)
}
```

Figure 5: JAGS model statement for model S.

to `loglambda[i]` $\sim$ `dt(E.loglambda[i],tau.noise,nu)` and second, we add a prior `nu` $\sim$ `dgamma(2,0.10)` for the degrees of freedom parameter (Juárez and Steel, 2010).

In models S and SH, we 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

```
# Year effects #
for(s in 1:nstrata){
      tau.gamma.tiny[s] <- tau.gamma[s]*0.0001
      gamma[s,y.baseline] ~ dnorm(stratum[s],tau.gamma.tiny[s])
      for(y in 1:(y.baseline-1))
          gamma[s,y.baseline-y] ~ dnorm(gamma[s,y.baseline-y+1],tau.gamma[s])
      for(y in (y.baseline+1):nyears){
          gamma[s,y] ~ dnorm(gamma[s,y-1],tau.gamma[s])
      }
      stratum[s] ~ dnorm(mu.strata,tau.strata)
      tau.gamma[s] ~ dgamma(0.001,0.001)
}
mu.strata ~ dnorm(0.0,1.0E-6)
tau.strata ~ dgamma(0.001,0.001)
```

Figure 6: Changes to slope models (S and SH) to produce difference models (D and DH).

expected value of $\gamma_{s,y}$ is

$$\mathrm{E}(\gamma_{s,y}) = S_s + \beta_s \, (y - y_0) \quad ;$$

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.

In models D and DH, we 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 fix $\mathrm{E}(\gamma_{s,y_0}) = S_s$, at the stratum mean. Changes needed for the JAGS code are illustrated in Figure 6.

In all four models, prior distributions for means are assigned flat normal priors, and prior distributions for precisions are assigned flat gamma priors.

Appendix S2: Abundance indices for models with t-distributed overdispersion effects

Background

Let Y_i denote a count obtained from the North American Breeding Survey. We suppose that $Y_i \sim Pois(\lambda_i)$, with

$$\log(\lambda_i) = A_{s(i)} + \beta_{s(i)}(y(i) - y^*) + \gamma_{s(i),y(i)} + \omega_{o(i)} + \epsilon_i \quad ;$$

here, A_s and β_s are parameters for stratum s; $s(i), y(i)$ and $o(i)$ are the stratum, year, and observer associated with count i; and ω_o and ϵ_i are random effects.

As an index to abundance for stratum s in year y, we use the expected value of Y_i given $s(i) = s$ and $y(i) = y$. Without specific assumptions on the distributions of ω_o and ϵ_i, the double expectation theorem yields

$$\mathrm{E}(Y_i|s(i) = s, y(i) = y) = \exp(A_s + \beta_s(y - y^*) + \gamma_{sy}) \times \mathrm{E}(\exp(\omega_{o(i)} + \epsilon_i)) \quad , \qquad (8)$$

provided that $\mathrm{E}(\exp(\omega_{o(i)} + \epsilon_i))$ exists. Assume that ω_o and ϵ_i are mean zero normal random variables with variances σ^2_ω and σ^2_ϵ. Then $\mathrm{E}(\exp(\omega_{o(i)} + \epsilon_i)) = \exp(\frac{1}{2}\sigma^2_\omega + \frac{1}{2}\sigma^2_\epsilon)$. The right-hand side of equation (8) is

$$\theta_{sy} = \exp(A_s + \beta_s(y - y^*) + \gamma_{sy} + \frac{1}{2}\sigma^2_\omega + \frac{1}{2}\sigma^2_\epsilon) \quad , \qquad (9)$$

which can be calculated as a derived parameter and monitored in fitting the model via

MCMC. Parameter θ_{sy} is the expected value of a new observation Y_i^*, a Rao-Blackwellized version of a posterior prediction (Casella and Robert 1996), and is thus an appropriate index to abundance. It should be noted that θ_{sy} is a parameter, rather than a statistic; point and interval estimates for the index are based on its posterior distribution.

Suppose now that ϵ_i has a t-distribution with ν degrees of freedom. Remarkably, $\mathrm{E}(\exp(\epsilon_i)) = \infty$, for any finite value of ν, no matter how large. Thus an expected count (like equation (8)) is inappropriate as a representation of typical counts. How then are we to define an index to abundance for BBS counts, if we choose to model ϵ_i as having a t-distribution?

We seek an index that is comparable to the expected count index used under assumptions of normality. Our solution is to regard the right-hand side of equation (9) as an exponentiated percentile of the normal distribution. Such a definition for the index θ_{sy} can then be generalized to t-distributions, as we show in these notes.

Before doing that, it is worth examining alternative indices to abundance. One would be to simply use the median value of λ_i given $s(i) = s$ and $y(i) = y$. The index could be called the "median expected count"; it would only involve the first term on the right-hand side of (8). However, the median expected count can be substantially smaller than the expected count. It seems more appropriate to exponentiate a percentile other than the median, and to be guided by the existing indices obtained under assumptions of normality.

Matching T and Normal distributions

If $\log(\lambda) \sim N(\mu, \sigma^2)$, then $\mathrm{E}(\lambda) = \exp(\mu + \frac{1}{2}\sigma^2)$. Writing $X = \log(\lambda)$, we can think of this quantity as $\mathrm{E}(e^X)$ or as $\exp(X^*)$, where $X^* = \mu + \frac{1}{2}\sigma^2$. Furthermore, we can think of μ and σ as the mean and variance, or as the median and the scaling parameter, relative to the

standard normal.

T-distributions with ν degrees of freedom have density $\pi_{\nu,k}(t) \propto (1 + ((t-m)/k)^2/\nu)^{-(\nu+1)/2}$; here k is a scaling parameter, and m is the median of the distribution. We have noted that if X has a t-distribution with ν degrees of freedom, $\mathrm{E}(\exp(X))$ does not exist. However, if $\nu > 2$, $\mathrm{E}(X)$ and $\mathrm{Var}(X)$ exist: $\mathrm{E}(X) = m$ and

$$\mathrm{Var}(X) = \frac{k^2\,\nu}{\nu - 2} \quad , \tag{10}$$

where $k > 0$ is a scale parameter. We could calculate a value $X^* = \text{mean} + \frac{1}{2}\text{variance}$, and define our index of abundance by $\exp(X^*)$, even though $\mathrm{E}(\exp(X))$ is not defined. However, the non-existence of a variance for $\nu \leq 2$, and the potential instability of (10) for ν only slightly larger than 2, argue against this approach. Indeed, if $\nu \leq 1$, the t-distribution has neither variance nor mean.

Instead, we will think of X^* as the median value of the t-distribution plus one half of a squared scaling parameter, with scaling relative to the standard normal distribution. We will accomplish the scaling by choosing the value $k = k(\nu)$ that leads to the closest approximation to a standard normal distribution by a central t-distribution with ν degrees of freedom and scale parameter k. The best approximating value of $k(\nu)$ will be that which minimizes Kullback-Leibler (KL) divergence from the standard normal distribution.

For each of 250 values of ν, evenly spaced on the interval [0.5, 40], we minimized the KL

686 divergence

$$\int_{-\infty}^{\infty} \log\left(\frac{\phi(t)}{\pi_{\nu,k}(t)}\right) \phi(t)\, dt \quad , \tag{11}$$

687 as a function of k; here $\phi(\cdot)$ is the density of a standard normal distribution. We approximated

688 the integral in (11) by a sum on 10^6 grid points over $[-10,10]$.

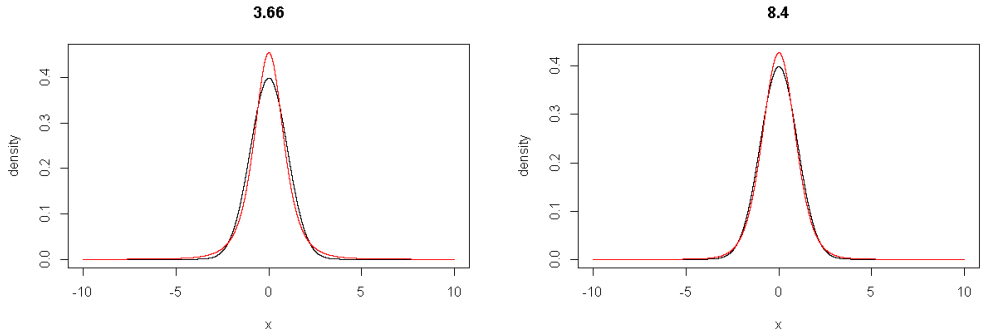


Figure 7: Standard normal (black) and best approximating $t_{\nu,k}$ distributions (red), for $\nu =$ 3.66, 8.40.

689 Two examples of the best matching densities are given in Figure (7). The relation between

690 the best scale parameter and the degrees of freedom is very close to linear on the log-logistic

691 scale ($R^2 = 0.9989$); the optimal values of k and fitted regression are given in Figure (8). The

692 regression relation is

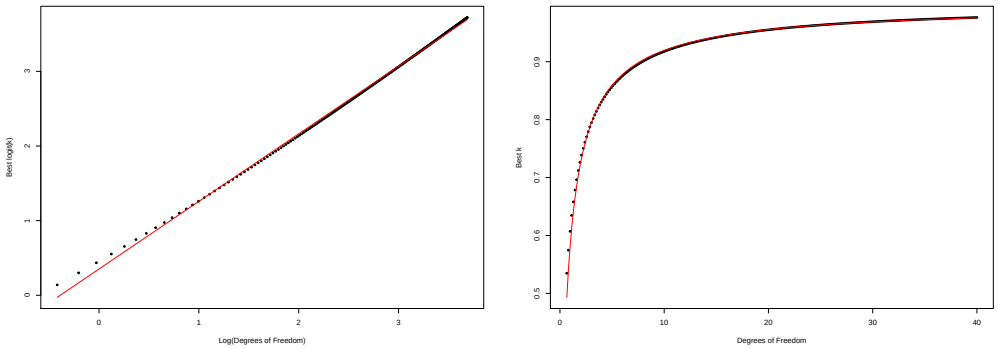


Figure 8: Fitted relation (red curves) between best k (black dots) and degrees of freedom.

$$k(\nu) = \frac{1.422\,\nu^{0.906}}{1\,+\,1.422\,\nu^{0.906}} \quad . \tag{12}$$

The value $k(\nu)$ is the scale parameter that most closely matches a $t_{\nu,k}$ distribution to a standard normal distribution, in terms of K-L divergence (equation (11)). From (10) it is anticipated that $k(\nu) \approx \sqrt{\frac{\nu-2}{\nu}}$; indeed, the ratio of the two is 1.04889 for $\nu = 7.5$, decreasing to a minimum of 0.9992 at $\nu = 151$, then tending to 1 as $\nu \to \infty$.

In JAGS code, we write `X` $\sim$ `dt(m,tau,nu)`. The parameter `tau` is a squared reciprocal of the basic scaling parameter, and must be adjusted by $k(\nu)$: the desired scaling value is $\left(1/\sqrt{\texttt{tau}}\right)/k(\texttt{nu})$.

Appendix S3: The need for and usefulness of leave-one-out calculations

Suppose data are a sample of $n = 20$ independent Poisson variables with expected value λ. Our data set happens to have an observation $Y_i = 11$, but the sample mean is a good bit lower, say $\bar{Y} = 5.3$. We want to calculate $\Pr(Y = 11|\lambda)$ to see whether the observation of 11 is consistent with the Poisson model. The problem is, we don't know λ, and the fact that one of our observations was $Y_i = 11$ suggests that the sample mean might be a bit high, with the result that the calculated value $\Pr(Y = 11|\lambda = 5.3) = 0.0116$ is too large. Omitting the value $Y_i = 11$ from the data set, the sample mean calculated from the remaining 19 observations is $(20 \times 5.3 - 11)/19 = 5.0$, and $\Pr(Y = 11|\lambda = 5.0) = 0.0082$. The original value, based on the full data set, was 40% larger. Not surprisingly, the inclusion of Y_i in the estimation of $\Pr(Y_i)$ makes the observation look more consistent with the model.

The very nature of maximum likelihood estimation guarantees such behavior, in evaluating

the fit of a model to a data set. The likelihood for θ^M is the product $\prod_i \Pr(Y_i|\theta^M, M)$, considered as a function of θ^M. We would like to evaluate this function at the true value of θ^M. By definition the MLE provides the largest possible value of the product, a value which is of necessity greater than or equal to what we'd get using the true value.

The same effect occurs in a Bayesian analysis. We continue with the example of estimating probabilities for a Poisson random variable with mean λ. We will write $f_k(\lambda) = \Pr(Y = k|\lambda) = \exp(-\lambda)\,\lambda^k/k!$, our notation emphasizing that the quantity of interest is a derived parameter, i.e., a function of λ.

Given a sample of size n, denoted by $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$, and a prior distribution for λ, inference will be based on the posterior distribution $[f_k(\lambda)|\boldsymbol{Y}]$, which can be obtained by transformation of the posterior distribution $[\lambda|\boldsymbol{Y}]$. In particular, we will suppose that a flat gamma prior has been chosen for λ, so that $[\lambda|\boldsymbol{Y}] = \text{Gamma}(\lambda; S, n)$, where $S = \sum_i Y_i$. Sampling λ's from this posterior distribution and calculating $f_k(\lambda)$ produces a sample of $[f_k(\lambda)|\boldsymbol{Y}]$.

Though our estimation procedure is Bayesian, we evaluate its frequency properties, i.e., the typical values of $[f_k(\lambda)|\boldsymbol{Y}]$ across replicate samples $\boldsymbol{Y}$ with a common (generating) value of λ. Frequentist evaluations of Bayesian procedures were advocated by Rubin (1984); Bayesian procedures with good frequentist operating properties are described as well-calibrated.

In what follows, we present 3 calculations. First, we average $[f_k(\lambda)|\boldsymbol{Y}]$ over an unconstrained set of replicate data sets $\boldsymbol{Y}$. We denote the resulting out-of-sample mean posterior distribution by $P^k_{os}(\lambda)$. Next, we average $[f_k(\lambda)|\boldsymbol{Y}]$ over a constrained set of replicate data sets $\boldsymbol{Y}$, chosen to include an observation $Y_i = k$. We denote the resulting within-sample mean posterior distribution by $P^k_{ws}(\lambda)$. Finally, we average a leave-one-out distribution over the same

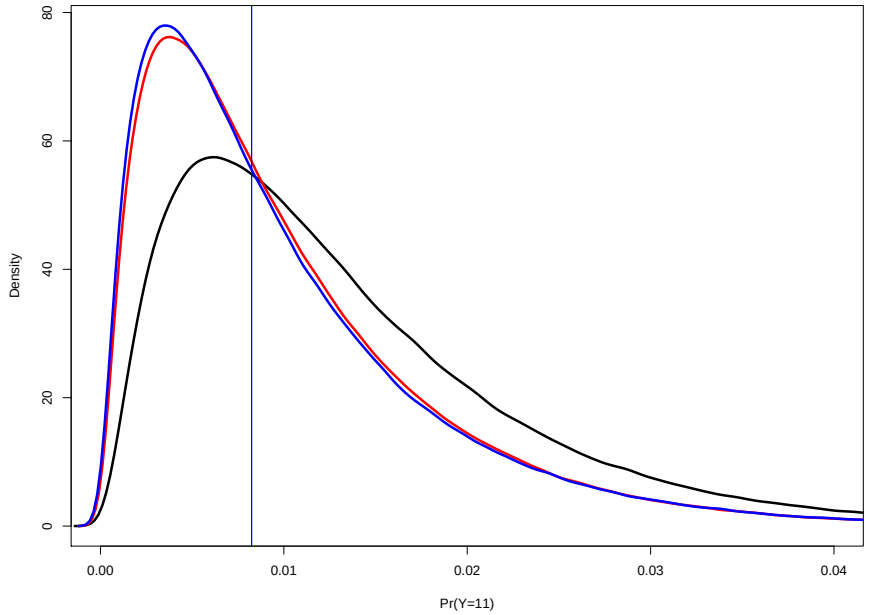


Figure 9: Average values of $\Pr(Y = 11|\boldsymbol{Y})$ taken over random samples $\boldsymbol{Y}$ of 20 Poisson(5) random variables. Using notation defined in text, red curve is density for $P_{os}^{11}(\lambda)$, black curve is for $P_{ws}^{11}(\lambda)$, and blue curve is for $P_{Loo}^{11}(\lambda)$.

constrained set of replicate data; we denote the resulting within-sample mean leave-one-out distribution by $P_{Loo}^{k}(\lambda)$.

Figure 9 displays the mean value of $[f_k(\lambda)|\boldsymbol{Y}]$ for $k = 11$, averaged over 10^6 sampled data sets with $n = 20$ and $\lambda = 5$ (red curve). This density can be thought of as the typical basis of inference about $\Pr(Y = 11|\lambda)$ for a sample of size 20, given $\lambda = 5$. We refer to this as an out-of-sample mean, because (in contrast to subsequent calculations) the sampling of replicate data sets was done without reference to any outcomes in a given data set. Inference based on $P_{os}^{k}(\lambda)$ appears reasonably well calibrated, with 51.4% of the mass in the density lying to the left of the true value 0.00824 (indicated by vertical blue line).

Now suppose that the analyst's interest in $\Pr(Y = 11|\lambda)$ is sparked by the occurrence of an observation $Y_i = 11$ in his dataset. If $\lambda = 5$, an observation $Y_i = 11$ stands out as large in

a sample of size 20. The sample itself is atypical; its sample mean is likely larger than the true mean, and estimates of $\Pr(Y = 11|\lambda)$ will be too large. We need to evaluate the average value of $[f_{11}(\lambda)|\boldsymbol{Y}]$ over replicate data sets including an observation equal to 11. So once again we generated 10^6 sampled data sets with $n = 20$ and $\lambda = 5$; this time, however, we used rejection sampling to require that at least one observation in the data set was equal to 11. The result is the within-sample mean posterior distribution $P^k_{ws}(\lambda)$, plotted in black in Figure 9. Only 35.8% of the mass of $P^k_{ws}(\lambda)$ lies to the left of the true value: the estimates tend to be too large, as anticipated. The mean value is 30.9% larger than the mean for $P^k_{ws}(\lambda)$. Using sampled values to compute their own probabilities leads to poorly-calibrated inferences.

The solution is to omit the value $Y_i = 11$ from the data set and treat what's left (denoted $\boldsymbol{Y}_{-i} = \boldsymbol{Y}\backslash Y_i$) as though it were a random sample of size $n - 1$. We thus sample $\lambda \sim \text{Gamma}(\lambda; S - Y_i, n - 1)$ and compute $f_{11}(\lambda)$. With a slight abuse of notation, we describe the collection of sampled values as a sample of $[f_{11}(\lambda)|\boldsymbol{Y}_{-i}]$. To evaluate the calibration of this procedure, we once again use rejection sampling to generate 10^6 data sets with at least one observation equal to 11. The average value of $[f_{11}(\lambda)|\boldsymbol{Y}_{-i}]$ for these data sets is $P^k_{Loo}(\lambda)$, plotted in blue in Figure 9; it closely approximates the density for $P^k_{os}(\lambda)$.

While use of the posterior distribution $[f_k(\lambda)|\boldsymbol{Y}]$ has good out-of-sample properties, its within-sample tendency is to overestimate. Using an observation to predict the probability of one just like it isn't a good idea. The solution to this problem is to use leave-one-out calculations. Even when evaluated using within-sample simulations, the leave-one-out procedure has frequency properties approximating the out-of-sample properties of the full posterior inference.

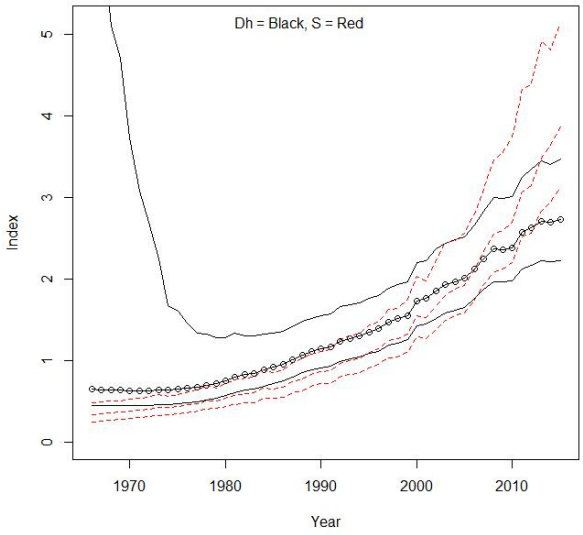
Appendix S4: Results for selected vs traditional model

Excel spreadsheet `Table of Trend Results 1970-2015.xlsx` summarizes population trend (%/yr) and 95% credible interval limits (2.5^{th} and 97.5^{th} percentiles) estimated at the survey-wide scale from BBS data for 548 species. We present results for the default model (S) and the selected model. We also present numbers of survey routes used in the analysis, and Relative Abundances (Rel. Abun., defined as the estimated annual index (birds/route) for the species in the year 1990) based on the default and selected model. Core species results are based on the interval 1970-2015; non-core species results are based on the interval 1993-2015.

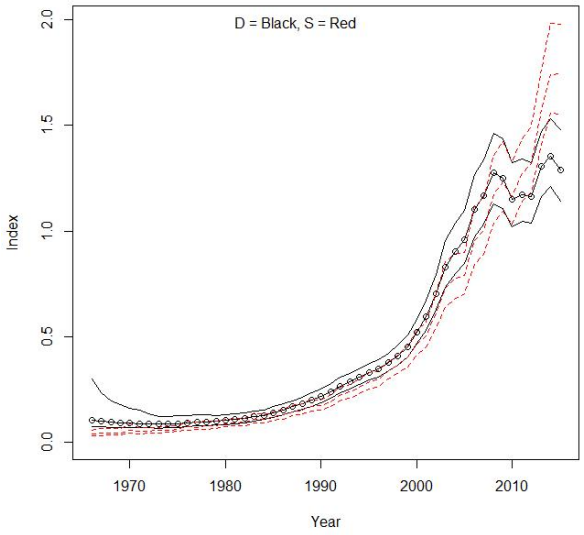
Appendix S5: Annual indices for 18 (of 426) core species with discrepant results between selected model and traditional model

Annual indices for 18 species for which trend results differed between the dafault model (S) and the selected model. Annual indices and 95% credible intervals for the default model are portrayed in red stipples; indices for selected model are portrayed in solid black. For 3 species (Wild Turkey (*Meleagris gallopavo*), Chimney Swift (*Chaetura pelagica*), and Eurasian Collared-Dove (*Streptopelia decaocto*)) model D was selected; model DH was selected for the remaining 15 species (Sandhill Crane (*Antigone canadensis*), Turkey Vulture (*Cathartes aura*), Black Vulture (*Coragyps atratus*), Cooper's Hawk (*Accipiter cooperii*), Merlin (*Falco columbarius*), Peregrine Falcon (*Falco peregrinus*), Common Nighthawk (*Chordeiles minor*), Ruby-throated Hummingbird (*Archilochus colubris*), Horned Lark (*Eremophila alpestris*), Eastern Meadowlark (*Sturnella magna*), Brewer's Blackbird (*Euphagus cyanocephalus*), Chestnut-collared Longspur (*Calcarius ornatus*), Western Tanager (*Piranga ludoviciana*), Barn Swallow (*Hirundo rustica*), and House Sparrow (*Passer domesticus*).)

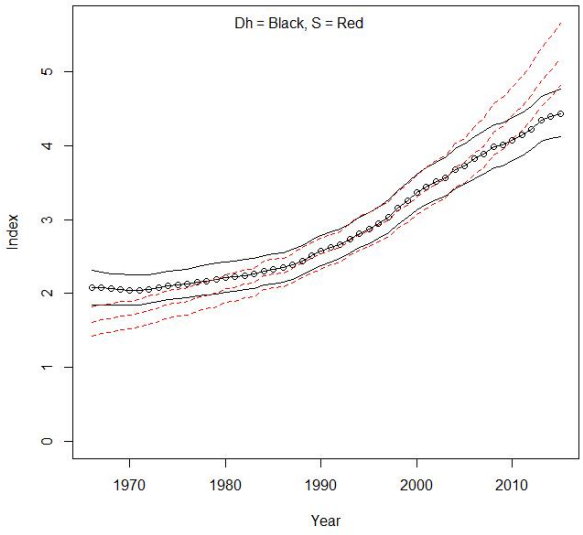
Sandhill Crane



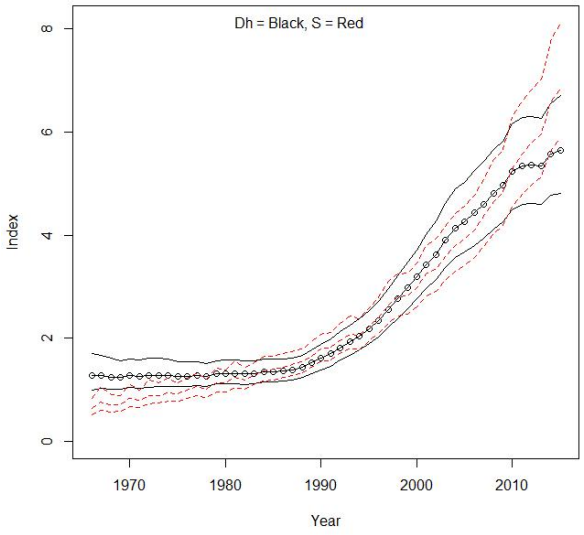
Wild Turkey



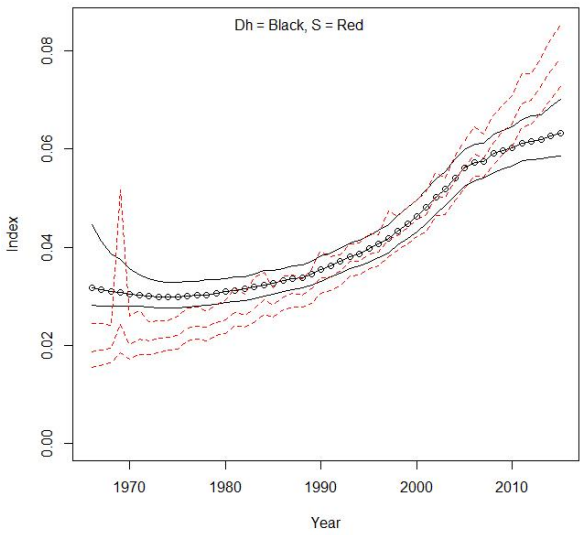
Turkey Vulture



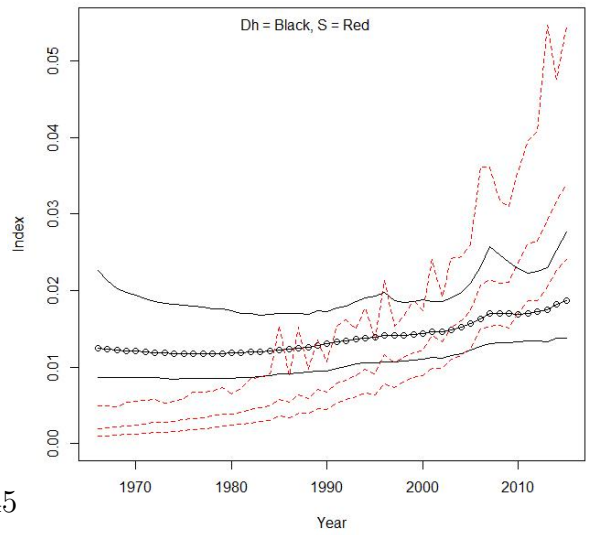
Black Vulture



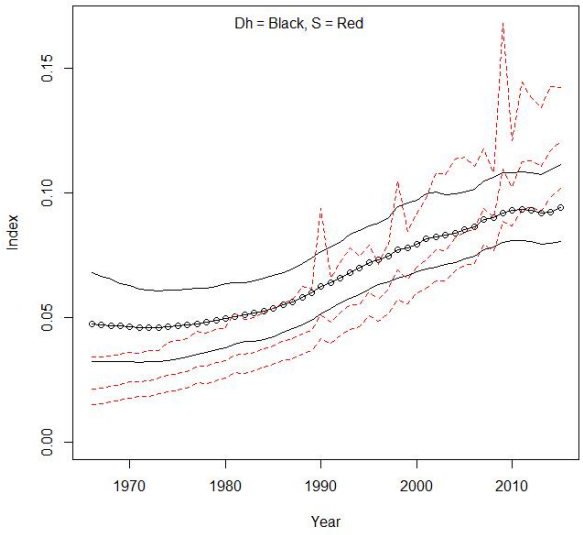
Cooper's Hawk



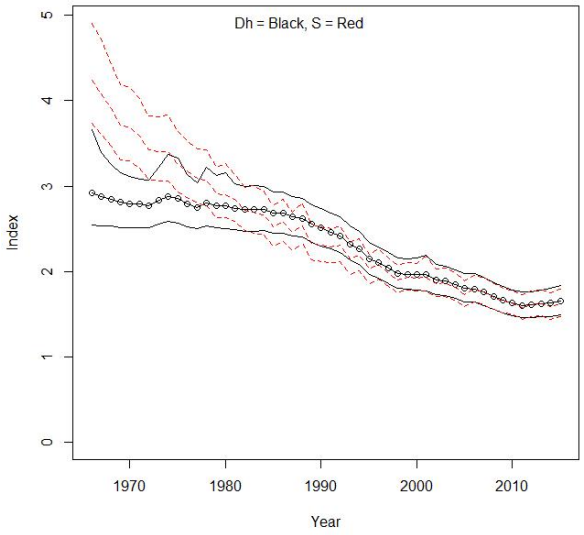
Peregrine Falcon



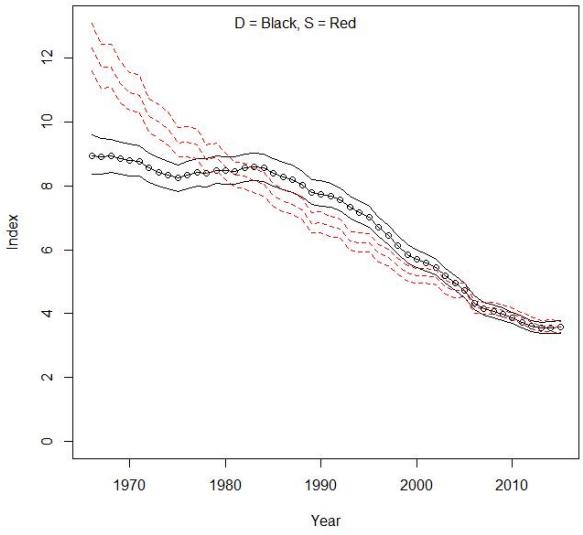
Merlin



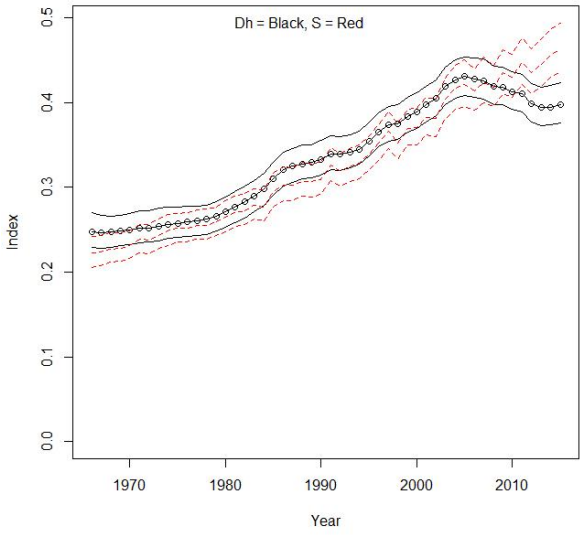
Common Nighthawk



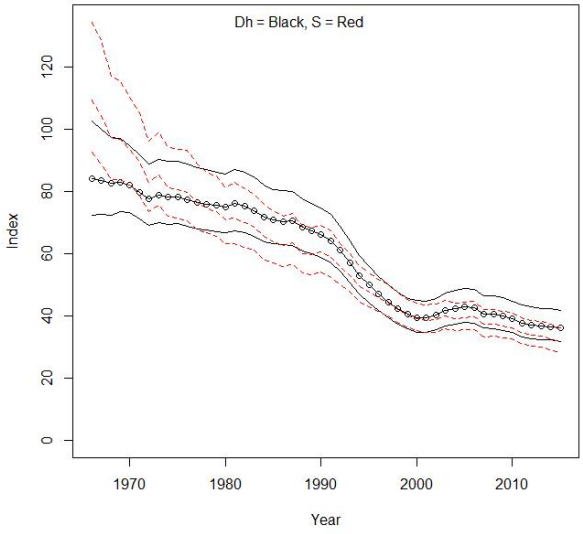
Chimney Swift



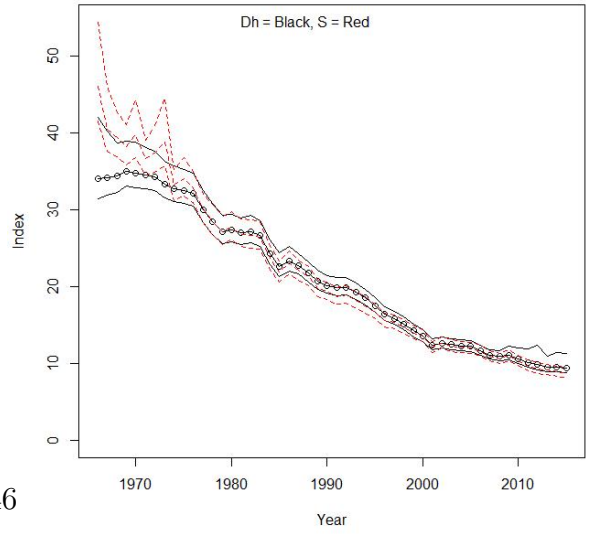
Ruby-throated Hummingbird



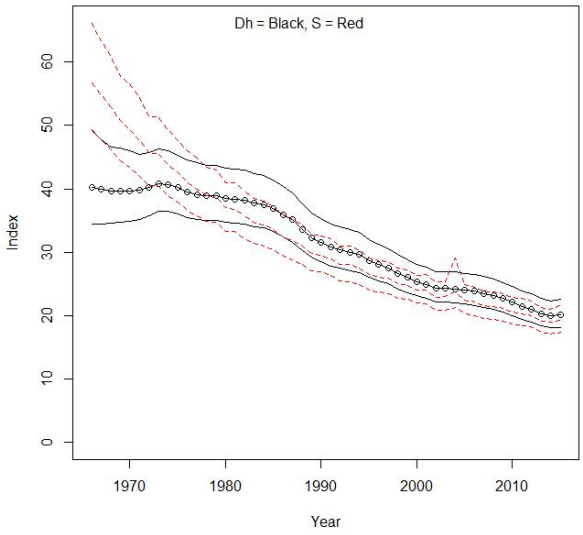
Horned Lark



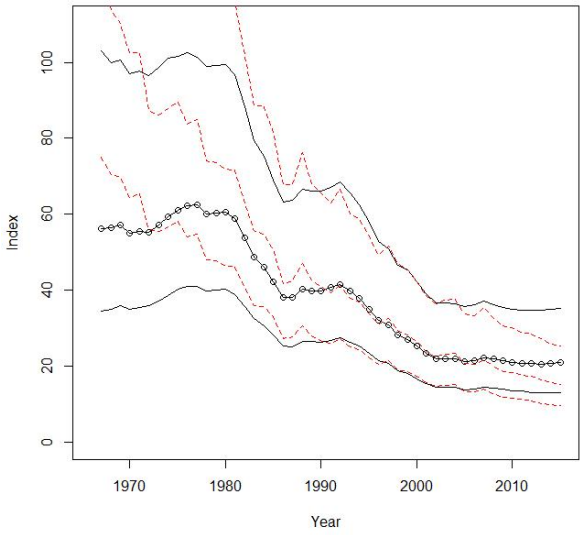
Eastern Meadowlark



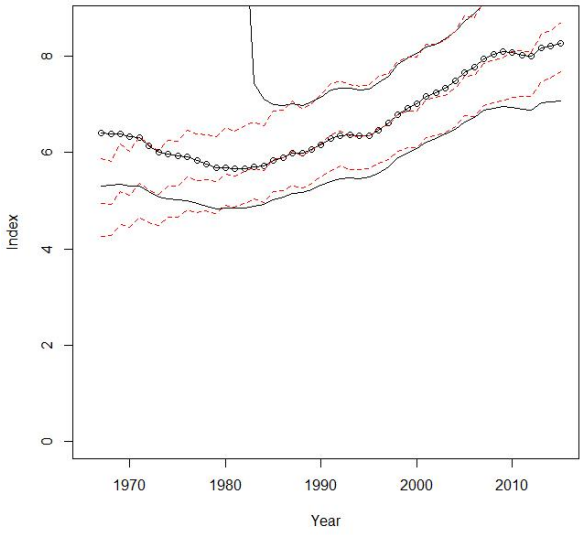
Brewer's Blackbird



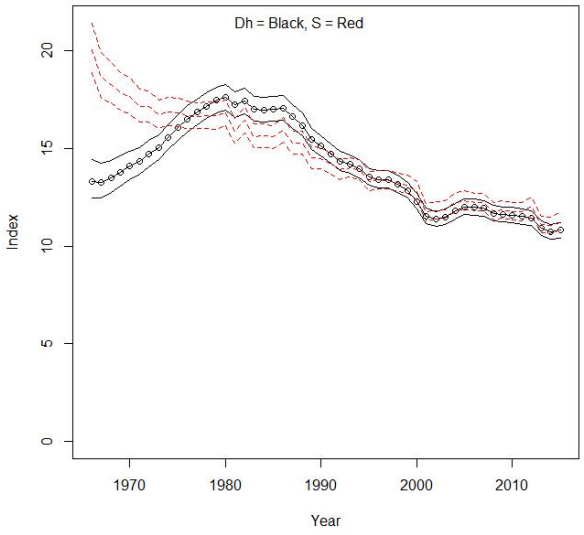
Chestnut-collared Longspur



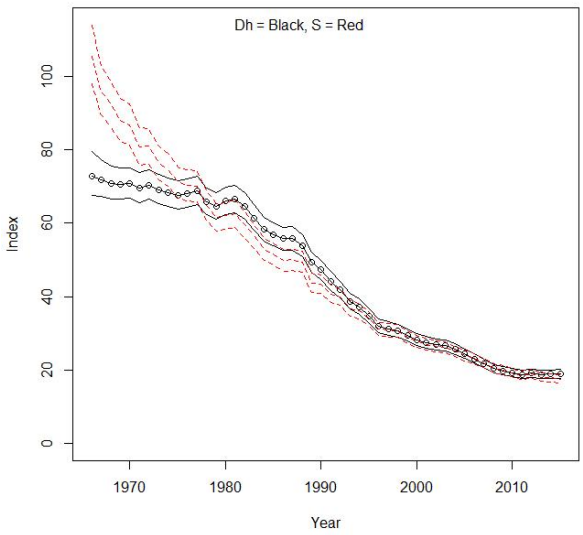
Western Tanager



Barn Swallow



House Sparrow



Eurasian Collared-Dove