

# Predictions

Andrew Crosby

08 August, 2022

## Relative predictive ability for model comparison

We evaluated out-of-sample (OOS) predictive ability in two ways. The first method looked at survey-specific predictive ability. For each saved draw from the posterior distribution, we calculated the expected song count for each breeding status at each survey. We then calculated the probability of getting the observed song count, given the expected song count for each breeding status as

$$P(SC|BS_i, T_j),$$

where  $SC$  is the observed song count,  $BS_i$  is breeding status  $i$ , and  $T_j$  is time relative to sunrise at survey  $j$ . Next, we multiplied these probabilities by the probability of being in each breeding status, calculated as

$$\mathbf{P}(\mathbf{BS}_i|\mathbf{D}_j),$$

where  $D_j$  is the Julian day at survey  $j$ , and  $\mathbf{P}(\mathbf{BS}_i)$  is the vector of probabilities of being in each breeding status. These products represented the estimated probability of the bird being in each breeding status at survey  $j$ , conditional on song rate, Julian day, and time relative to sunrise (Equation 1 in manuscript). Finally, we calculated the log likelihood of the data given the models as the probability of the observed breeding status given the estimated probability of being in each breeding status, and summed these across all test surveys.

The model likelihood score for each cross validation fold, based on this method, was the sum of the survey likelihoods averaged over all draws from the posterior distribution, so that:

$$L_{log} = \frac{1}{s} \sum_{s=1}^S -2 \left( \sum_{j=1}^J \log(P(y_j|\mathbf{P}[\mathbf{BS}_j])) \right),$$

where  $y_j$  is the observed breeding status at survey  $j$ ,  $\mathbf{P}[\mathbf{BS}_j]$  is the vector of expected probabilities for each breeding status at survey  $j$ , and  $s$  is the individual draw from the posterior distribution.

In the second method, we calculated the likelihood of the observed proportion of bird-days in each breeding status, given the expected proportion of bird-days in each breeding status from the model at each draw from the posterior distribution. Because in most cases there were multiple surveys for the same bird-day, we calculated the vector of breeding status probabilities for bird-day  $d$  as

$$\mathbf{P}[\mathbf{BS}_d] = \frac{1}{\sum_{J_d}} \sum_{j=1}^J \mathbf{P}(\mathbf{BS}_{dj}),$$

where  $\mathbf{P}(\mathbf{BS}_{dj})$  is the vector of probabilities of being in each breeding status at survey  $j$  on bird-day  $d$  estimated by the model, and  $\sum_{J_d}$  is the number of surveys conducted on bird-day  $d$ . We then estimated the breeding status for bird-day  $d$  as

$$BS_d \sim \text{Multinomial}(\mathbf{P}[\mathbf{BS}_d]).$$

The Log-score for each cross validation fold based on proportion of bird-days in each breeding status,  $L_{log}$ , was the probability of the number of observed breeding statuses in each category,  $BS_{oos}$ , given the expectations from the model, averaged over all  $s$  draws from the posterior distribution, so that

$$L_{log} = \frac{1}{s} \sum_{s=1}^S -2\log(P(BS_{oos} | \mathbf{P}[\mathbf{BS}])),$$

where  $\mathbf{P}[\mathbf{BS}]$  is the vector of probabilities for each breeding status calculated as the estimated proportion of bird-days in each status from the model. We summed the Log-scores over all  $k$  folds from the cross validation to get the final score for each model.

## Actual predictive ability

While all well and good for comparing models, the likelihood methods outlined above gives relative measures of predictive ability. What is needed for determining the usefulness of the model is an absolute estimate of how accurately the model actually estimates breeding status, as well as the uncertainty around those estimates. We proceed with the theory that, while the model may not be able to predict breeding status very well at any given survey, it may be able to accurately estimate the proportion of individuals in each breeding status across a number of bird-days. Another good estimate is how well the model can estimate the proportion of birds in the population that made it to the feeding young (FY) stage, as this is an important component of breeding success.

First, we separated the 65-day breeding season covered by our samples into 13 five-day periods. We then estimated the accuracy and uncertainty of model predictions of the proportion of samples in each breeding status within each period. Our estimates for each period were derived from the model predicted breeding status at each draw from the posterior distribution for each bird-day in the sample, as we did in the second log-score method above, giving us the uncertainty distribution around the estimate. For each breeding status at each period, we calculated the mean and standard deviation of the predicted proportion of bird-days in that breeding status, and compared those values to the known proportion.

To test how well the model estimated the proportion of the population that made it to the FY stage, we extracted the predicted status at each bird-day at each iteration of the model. Then, for each bird-year at each iteration, we calculated the proportion of days it was predicted to be in each breeding status, beginning with the first day the bird was predicted to be in the FY stage. We then used the proportion of bird-days predicted as FY as the probability the bird had reached the FY stage. If the bird-year was never predicted to be in the FY stage, we set the probability of reaching the FY stage to zero for that iteration. Finally, we predicted whether each bird-year reached the FY stage as  $fy \sim \text{Bernoulli}(p_{fy})$ , where  $fy$  is an indicator of reaching the FY stage (1 or 0), and  $p_{fy}$  is the probability of reaching the FY stage derived from the model, and calculated the proportion of bird-years predicted to have reached the FY stage. Finally, we calculated the mean and standard deviation of the distribution of the predicted proportion over all iterations and compared these to the known proportion of individuals that reached the FY stage.