**Where the birds at? Predicting Acadian Flycatcher (*Empidonax virescens*) detections in Iowan forests**

# 1. Introduction

Breeding site selection for forest birds is driven by habitat, leading to associations of different birds with different habitat characteristics (MacArthur & MacArthur, 1961). Forest bird-habitat associations encompass a range of spatial scale, from landscape scale forest cover to vegetation structure and florisitics at a site (Bakermans & Rodewald, 2006; Reidy et al., 2014; Rodewald & Abrams, 2002). Using habitat metrics at a variety of spatial scales, it is possible to predict the presence of different forest bird species at a site (McDermott et al., 2011; Mitchell et al., 2001). In North American, forest bird populations have declined by almost 20% in recent decades (Rosenberg et al., 2019). These declines make identification and protection of breeding territories of declining forest bird species an important population management priority. This project focuses on Acadian Flycatchers (*Empidonax virescens*) in south central Iowa, USA. The Acadian Flycatcher is a Species of Greatest To support the goal of predicting sites with Acadian Flycatcher territories, I assessed the ability of three model types to predict Acadian Flycatcher territory presence from habitat data.

# 2. Methods

## 2.1. Bird and Habitat Data

The bird and habitat data methods presented here are a summary; see West (2020) for full details. For all field surveys, I employed a grid of 493 points with 300 m spacing, distributed across public lands within three forested Bird Conservation Areas in south-central Iowa. The public lands included Sand Creek Wildlife Management Area and multiple units of Stephens State Forest. Over the course of the 2019 breeding season (late May to early August), observers conducted 10-minute bird surveys at each point; they visited each point twice. During each survey, singing Acadian Flycatchers were recorded as present or absent within 100 meters of each point. If at least one singing Acadian Flycatcher was recorded during at least one survey, an Acadian Flycatcher territory considered “present” at that site. I derived various habitat characteristics from point-scale field vegetation surveys conducted from mid July to late August 2019 (Table 1). Landscape-scale habitat characteristics were derived using Esri® ArcGIS Pro® to analyze the 2019 National Cropland Data Layer (Table 1; USDA 2019).

Table 1. Habitat variables from south central Iowa forests used to predict presence of Acadian Flycatcher territories. \*Standard deviation of midstory foliage density calculated from six measures of midstory foliage, each calculated as proportion of coverboard covered.

|  |  |
| --- | --- |
| **Variable name** | **Definition** |
| fprop\_1km | Forest landcover within 1 km (proportion of area) |
| fprop\_10km | Forest landcover within 10 km (proportion of area) |
| spp\_rich | Tree spp. richness in a variable radius 1-m factor forestry prism plot |
| live\_basal | Live tree basal area (m2/ha) |
| dead\_basal | Dead tree basal area (m2/ha) |
| total\_basal | Total tree basal area (m2/ha) |
| oak\_prop | Proportion of basal area composed of oak (*Quercus* spp.) |
| can\_clos\_prop | Canopy closure (proportion) |
| grass\_prop | Proportion of ground covered by grass or sedges |
| green\_prop | Proportion of ground covered by herbaceous or small woody plants |
| litter\_prop | Proportion of ground covered by leaf litter |
| shrub\_dens | Shrub density (stems/m2) |
| dist\_edge | Distance of point to forest edge (m; 0 m for points outside forest) |
| nearest\_patch\_size | Size of the closest forest patch (m2) |
| mid\_sd | Standard deviation of midstory foliage density\* |
| mid\_dens\_2.5m | Foliage density at 2.5 m height (mean proportion covered of three coverboards) |
| mid\_dens\_5m | Foliage density at 5 m height (mean proportion covered of three coverboards) |

## 2.2. Models

When fitting each model, I used a random 80% subsample of the original data set as a training dataset, and I reserved the remaining 20% as a hold-out set. Due to class imbalances, I performed resampling with replacement on the positive cases (points with Acadian Flycatcher territories) within the training dataset such that the number of resampled positive cases in the training dataset was equal to the number of negative cases in the training dataset. I tuned and compared three classification models. To tune hyperparameters of each model type, I used 5-fold cross-validation grid searches with predetermined hyperparameter combinations from Mahoney (2021). A cross-entropy loss function was used to compare predictive power between models within a grid search; the model with the lowest loss value was selected.

The first model was a classification decision tree, fit using the ‘rpart’ package in R Version 4.1.1 (R Core Team, 2021; Therneau & Atkinson, 2019). The minimum number of observations per leaf node was the only hyperparameter I tuned for the decision tree; a minimum of 31 observations per leaf node minimized cross-entropy loss. The second model evaluated was a random forest, fit using the ranger R package (Wright & Ziegler, 2017). The set of hyperparameters that minimized cross-entropy loss included 800 trees, 5 variables considered per split, a minimum of 3 observations per leaf node, sampling with replacement when fitting trees, and a proportion of 0.63 of the original data resampled for each tree. The third model evaluated was a stochastic gradient boosting machine fit with the lightgbm package (Shi et al., 2021). The grid search for this model was iterative and had three stages. At the first stage, only learning rate and the number of trees were tuned, leaving other hyperparameters with their default values. At the second stage, the best values from the prior stage were used, and both maximum tree depth and minimum data in bin were tuned. For the third stage, best values were only stochastic hyperparameters were tuned, comprising percent of observations sampled for each tree, bootstrapping resampling frequency, and the percentage of variables available to each tree.

Post-tuning, area under an ROC curve (AUC), overall accuracy, sensitivity, and specificity were used to compare the different model types. AUC was calculated using R package ‘pROC’ (), and other metrics were generated from confusion matrices produced by ‘caret’ ().

The third and final model evaluated was a stochastic gradient boosting machine, fit using the lightgbm package (Ke et al. 2021). Model tuning was done using iterative grid searches, evaluating progressively narrower ranges of multiple hyperparameters with each iteration. The set of hyperparameters which maximized AUC were selected, with the final model being fit using 3,000 trees allowed to grow to arbitrary depths, with a learning rate of 0.1 and a minimum of 13 observations per leaf node. Each tree was fit to a bootstrap sample that was 90% the size of the original training data, with a new sample taken every 5 trees, and each tree was fit using a randomly selected 30% of all features. All models were fit using methods to predict class probabilities, in order to produce ROC curves and AUC estimates. Classifications were then made using the thresholds which optimized both sensitivity and specificity, as identified using the training data. Models were assessed using their overall accuracy, sensitivity, specificity, and AUC, calculated against the 20% holdout set. ROC curves and AUC were calculated using the pROC package (Robin et al. 2011). Data wrangling used the dplyr, tidyr, and recipes packages (Wickham et al. 2021; Wickham 2021; Kuhn and Wickham 2021). Manuscript preparation used the ggplot2 and kableExtra packages (Wickham 2016; Zhu 2021). All analyses used the R statistical modeling software (R Core Team 2021). 3. Results Model accuracies are reported in Table 1. Logistic regression produced the model with the highest AUC, overall accuracy, and sensitivity, while the stochastic GBM fit through LightGBM provided the highest specificity. The random forest model performed the worst on all accuracy measured. ROC curves for each model are presented in Figure 1.

# Results

The proportion of sites occupied by singing Acadian Flycatchers was 0.263.

# References

Bakermans, M. H., & Rodewald, A. D. (2006). Scale-dependent habitat use of Acadian Flycatcher (Empidonax virescens) in central Ohio. *Auk*, *123*(2). https://doi.org/10.1642/0004-8038(2006)123[368:SHUOAF]2.0.CO;2

Layer, U. N. A. S. S. C. D. (2019). *Published crop-specific data layer [Online]*. https://nassgeodata.gmu.edu/CropScape/ (accessed 02.24.20)

MacArthur, R. H., & MacArthur, J. W. (1961). On Bird Species Diversity. *Ecology*, *42*(3). https://doi.org/10.2307/1932254

Mahoney, M. J. (2021). *Machine Learning Concepts and Applications*. https://mlca.mm218.dev/2021

McDermott, M. E., Wood, P. B., Miller, G. W., & Simpson, B. T. (2011). Predicting breeding bird occurrence by stand- and microhabitat-scale features in even-aged stands in the Central Appalachians. *Forest Ecology and Management*, *261*(3). https://doi.org/10.1016/j.foreco.2010.10.020

Mitchell, M. S., Lancia, R. A., & Gerwin, J. A. (2001). Using landscape-level data to predict the distribution of birds on a managed forest: Effects of scale. *Ecological Applications*, *11*(6). https://doi.org/10.1890/1051-0761(2001)011[1692:ULLDTP]2.0.CO;2

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. https://www.r-project.org/

Reidy, J. L., Thompson, F. R., & Kendrick, S. W. (2014). Breeding bird response to habitat and landscape factors across a gradient of savanna, woodland, and forest in the Missouri Ozarks. *Forest Ecology and Management*, *313*. https://doi.org/10.1016/j.foreco.2013.10.042

Rodewald, A. D., & Abrams, M. D. (2002). Floristics and avian community structure: Implications for regional changes in eastern forest composition. *Forest Science*, *48*(2). https://doi.org/10.1093/forestscience/48.2.267

Rosenberg, K. V., Dokter, A. M., Blancher, P. J., Sauer, J. R., Smith, A. C., Smith, P. A., Stanton, J. C., Panjabi, A., Helft, L., Parr, M., & Marra, P. P. (2019). Decline of the North American avifauna. *Science*, *366*(6461). https://doi.org/10.1126/science.aaw1313

Shi, Y., Ke, G., Soukhavong, D., Lamb, J., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., & Titov, N. (2021). *lightgbm: Light Gradient Boosting Machine*. https://cran.r-project.org/package=lightgbm

Therneau, T., & Atkinson, B. (2019). *rpart: Recursive Partitioning and Regression Trees*. https://cran.r-project.org/package=rpart

West, B. M. (2020). *Bird-habitat relationships in southern Iowa forests*. Iowa State University, Ames, IA.

Wright, M. N., & Ziegler, A. (2017). {ranger}: A Fast Implementation of Random Forests for High Dimensional Data in {C++} and {R}. *Journal of Statistical Software*, *77*(1), 1–17. https://doi.org/10.18637/jss.v077.i01